Optimization and Machine Learning Methods for Medical and Healthcare Applications

Cao Xiao

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

W. Art Chaovalitwongse, Chair

Shuai Huang, Chair

Sandor Toth

Shan Liu

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Optimization and Machine Learning Methods for Medical and Healthcare Applications

Cao Xiao

Chair of the Supervisory Committee:
Professor W. Art Chaovalitwongse
Industrial and Systems Engineering

The increasing amounts of data being gathered in healthcare and medical systems and the convergence of different domains are leading medical and healthcare research to a new direction of precision and personalized medicine. The trends bring a unique opportunity and good promise to solving various critical tasks in medical and healthcare research. However, such promise heavily relies on whether we could find useful patterns to characterize the target problems, discover informative mechanisms underlying the noisy and fragmented factual information, as well as transform these knowledge into intelligent decision making. Previously many efforts have been made via various approaches such as machine learning, optimization, statistical analysis, mathematical modeling, biomedical informatic, etc. This thesis will extend along the lines of machine learning and optimization methods, and to build novel models for solving the medical or healthcare challenges. Particularly, the thesis focus on the following topic domains, including medical image or recording based disease differentiation and diagnosis, image guided precision surgery, disease progression modeling, and personalized health behavior recommendation.

To summarize, the thesis mainly includes five models that are fit for a wide varieties of healthcare and medical applications.

The first model is an integrated feature ranking and selection framework that is capable of selecting a sparse model while preserving the most informative features. The framework combines information theoretic criteria and the least absolute shrinkage and selection operator (lasso) method

into a two-step feature selection process. It can be applied to biomarker selection problems when the number of subjects is small comparing with number of candidate biomarkers.

The second is a structure learning model capable of achieving personalized identification of surgery insertion location. I introduce a method to craft novel patient-specific features from their medical images. In addition, I propose a supervised structure learning and prediction model with special inter-dimensional and response structure regularization terms to capture spatial relations of features and responses.

The third is a systematic framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion to translate users' behavioral data into deeply personalized health planning. The framework is fit for longitudinal user behavior data, and can potentially be a backend algorithm for mobile health (mHealth) technologies.

The fourth model is an optimal expert knowledge elicitation strategy for identifying pairwise Bayesian network structure from observational data. It combines observational data and expert knowledge and iteratively elicit new expert knowledge that is optimally matched to the observational data to maximally reduce the uncertainty in the structure identification. The strategy can be applied to leverage expert knowledge in learning causal relation from data when the observational data is limited. In the thesis we use it to learn the underlying progression mechanism of Alzheimer's disease.

Last, the thesis also presents a decomposed version of the k nearest neighbor (DKNN) method for the classification of an unseen object based on the distances to the centroids of the k nearest neighbors. The DKNN has a training process to learn the local-optima-free distance metric by solving a convex optimization problem. The optimization problem not only learns a metric that minimizes classification errors and maximizes the margin between intra-class and inter-class distance. In addition, it also selects important features and removes irrelevant features when L1 regularization is incorporated in the optimization model. The DKNN algorithm is designed as a general method, but could be used in identifying discriminating features for a particular disease

(binary cases) or different patient cohorts (multi-class cases).

To demonstrate the utility of the proposed machine learning and optimization models, numerical studies are performed using simulation, public and/or real data. Real data including magnetic resonance imaging (MRI) data, electroencephalogram (EEG) recordings, positron emission tomography (PET) data, data from wearable devices, computed tomography (CT) scan, etc. Through extensive experiments and analysis, the proposed approaches outperform baseline methods, and demonstrate their utility and efficacy in medical and healthcare applications.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I wish to express my deep and sincere appreciation to my Ph.D. advisor Dr. Wanpracha Art Chaovalitwongse, for his insightful guidance, endless patience and encouragement, and generous support throughout my research at University of Washington. It was him, who introduced me to data mining, offered me lots of great opportunities and resources, and guided me when I got stuck. Without him, this dissertation would not have become possible.

I also want to thank my co-advisor Dr. Shuai Huang. It has been my honor to be advised by him. He has taught me how good methodological works are done. I appreciate all his contributions to make my Ph.D. experience productive and stimulating. The joy and enthusiasm he has in research was contagious and motivational for me. In addition, for this dissertation I would like to thank my reading committee members: Dr. Shan Liu and Dr. Sandor Toth, who provided valuable comments and suggestions for both my dissertation and my future career.

The members of the TeamArt and my friends in the Department of Industrial and Systems Engineering have contributed immensely to my personal and professional time at UW. I am especially grateful for Dr. Shouyi Wang, Dr. Joe Chou, Dr. Daehan Won, Fangfang Wang, Dr. Caroline Krejci, Dr. Pengbo Zhang, Dr. Yuqing Wu, Wei Guo, Yan Jin, Tianshu Feng, and Qiang Meng (ordered by the time of our first encounter).

During my Ph.D. studies, I have spent tremendous time interning at LinkedIn and IBM Research. I would like to thank people who mentored or helped me at LinkedIn, including: David Mandell Freeman, Ted Hwa, Paul Ogilvie, Joel Young, Emily Huang, etc. I also would like to thank my mentors and friends at IBM Research, Healthcare Analytics Research Group, including Fei Wang, Ping Zhang, Kenney Ng, and Jianying Hu. Thank you and I look forward to continuing working with you guys.

Last but not least, I would like to thank my family for their love and encouragement, especially my loving, supportive, encouraging, and patient husband Yongning, whose faithful support during all stages of my study is very much appreciated. And my kitties, Miaomiao and Feifei for their sweetness and loyal company during many working nights.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Motivation

The increasing amounts of data being gathered in healthcare and medical systems and the convergence of different domains are leading medical and healthcare research to a new direction of precision and personalized medicine. Precision medicine is an emerging approach for disease treatment and prevention that takes into account individual variability [2], while personalized medicine is a medical procedure that provides medical decisions, practices, interventions or products being tailored to the individual patient based on their predicted response [3]. These trends bring a unique opportunity and good promise to solving various critical tasks in medical and healthcare research. However, such promise heavily relies on whether we could find useful patterns to characterize the target problems, discover informative mechanisms underlying the noisy and fragmented factual information, as well as transform these knowledge into intelligent decision making.

Previously many efforts have been made via various approaches such as machine learning, optimization, statistical analysis, mathematical modeling, biomedical informatic, etc. This thesis will extend along the lines of machine learning and optimization methods, and to build novel models for solving the medical or healthcare challenges. Particularly, the following challenging problem domains are identified, and will be the main focus of the models proposed in this thesis. In the following we will give a brief overview of these problems, the state-of-the-art methods being proposed to address them, and our perspectives to solving the challenges.

1.1.1 Image or Recording Based Disease Biomarker Selection for Neurological Diseases

Technological development has paved the way for accelerated biomarker discovery and is bringing image based disease diagnosis into our attention. The goal for image based biomarker discovery and disease diagnosis are to uncover optimally and objectively targeted biomarkers tailored to an individual or a cohort. Neurological diseases are promisingly suited use cases for precision medicine due to the neurological change is not obvious in other easily observed phenotypic features during the early stage of disease onset, but could be discovered from neuroimages. Here in this thesis we design machine learning and optimization methods to bring greater clarity to the clinical and biological complexity of neurological diseases. The diseases include 1) neurodevelopmental disorders, e.g. attention deficit hyperactivity disorder (ADHD); 2) neurological disorder, e.g. epileptic seizure; and 3) neuro-degenerative disorder, e.g. Alzheimer's disease.

Neurodevelopmental disorder

Neurodevelopmental disorders are impairments in development of the brain functions, and would affect emotion, learning ability, self-control and memory and other functions that underpin human development. In this thesis, we mainly concern attention deficit hyperactivity disorder (ADHD), one of the most common child and adult neurodevelopmental disorders [4]. Current diagnosis of ADHD still remains a challenge, requiring long term and extended involvement from clinicians, parents and teachers, and also heavily relies on experiences and intuitions when conducting diagnostic interview and observational measures. These issue could cause delay or incorrect diagnosis, which would further result in a significant negative impact on a patient's social and emotional development, while an early and accurate detection of ADHD can strongly influence the course of the condition development by delivery of appropriate treatments to the patient. In addition to the traditional clinical diagnosis, there is a pressing need to find a set of more discriminative and objective features to characterize ADHD that can be used to facilitate a more precise and personalized ADHD diagnosis.

Previous studies on the etiology of ADHD are mostly based on structural or functional neu-

roimaging research to calculate group level (ADHD vs. control) differences. Some informative features extracted are blood oxygenation level dependent (BOLD) signals from functional magnetic resonance imaging (fMRI) data [5], wavelet synchronization likelihoods extracted from electroencephalography (EEG) data [6], rolandic spikes from EEG data [7], brain volume measure extracted from magnetic resonance imaging (MRI) data [8]. The pursuit of neuroanatomical biomarkers has a great potential to facilitate new discriminative methods that are etiologically informed and validated by neuropsychological theories. However, due to high cost of neuroimaging data acquisition, most current ADHD studies are based on relatively small sample sizes, which reduce the statistical power needed to validate meaningful discriminative variable from a very large number of features extracted from structural MRI [9]. A limited sample size with equivalent number of features raises new challenges to traditional machine learning algorithms, such as logistic regression or support vector machines (SVM), as they tend to overfit and lack a generalization power when training on a dataset containing the number of features far larger than the sample size ($p \gg n$ problem). In previous work, some models either use hundreds of features as an input or exhaustively search on a preselected smaller subset of features. SVM is mostly favored [10] and some variant of feed-forward neural networks [11] is also used. We believe that those methods are either susceptible to overfitting or too restrictive in the search space. The interpretation of the final models is very difficult to validate by existing neuropsychological theories.

My work in Chapter 3 is to develop an integrated feature ranking and selection framework that is capable of selecting a sparse model while preserving the most informative features. The framework can be applied to biomarker selection problems when the number of subjects is small comparing with number of candidate biomarkers.

Neurological disorders

Neurological disorders refer to the diseases of the nervous system. The neurological disorder in this thesis mainly includes epileptic seizure, which is the most common neurological brain disorders next to strokes, and about 1% of human population (40 million people) suffer from epilepsy [12]. To find an accurate diagnosis, the electroencephalogram (EEG) recording play a central role since it

could directly detect electrical activity in the brain. In practice, a prolonged (24-h) EEG monitoring is often necessary due to the epileptic diagnosis heavily relies on a tedious visual screening process by neurologists from lengthy EEG recordings that require the presence of seizure (ictal) activities.

In the past decades, there have been many quantitative analysis systems to help neurologists identify epileptic patterns from long-term EEG recordings for seizure detection and prediction. However, it is costly and often difficult to obtain long-term EEG data with seizure activities for epilepsy patients, especially in the areas that lack of medical resources and well-trained neurologists. There have been very few studies that using short-term interictal EEG for more convenient and affordable epilepsy diagnosis. There is a desperate need for a new medical diagnostic tool that is capable of providing quick and accurate epilepsy-screening using short-term interictal EEG signals.

In Section 3 our approach is to investigate the application of short-term interictal EEG signals for epilepsy diagnosis using machine learning techniques. In particular, we propose an information-theory-guided feature selection and prediction framework to identify epilepsy-specific EEG patterns in a fast screening process in a human-computer interaction task using visually-evoked potentials (VEP). The proposed method has a potential to be applied to determine whether a patient is epileptic or non-epileptic in a quick screening process.

Neuro-degenerative disorders

Neurodegenerative diseases are one type of neuro disorders that have incurable and debilitating conditions and show progressive degeneration. Disease Progression Modeling (DPM) [13], the modeling of the progression of a target disease with computational methods, is an important technique that can help with the early detection and management of neurodegenerative diseases. By characterizing the entire disease progression trajectory, DPM also facilitates disease prognosis improvement and clinical trial design.

There have been a few existing work on DPM that targets a specific domain. For example, [14] presented a general Hidden Markov Model for simultaneously estimating the transition rates and the probabilities of stage misclassification. [15] conducted a meta analysis to model the longitudi-

nal changes of patients with mild to moderate AlzheimerâĂŹs disease. [16] developed a dynamic Bayesian network based technique to model the progression of Coronary Atherosclerosis.

In Chapter 6 of this thesis, we will address the modeling challenges for Alzheimer's disease (AD), a neuro-degenerative disease causing mental condition degeneration and ultimately leading to dementia. Our task is to identify the cascade of hypermetabolism reduction events for Alzheimer's disease. Knowing the cascade of hypermetabolism reduction events will help us understand the progression stages of the disease, which is especially beneficial for early detection of the disease. To achieve this learning goal, we will learn the ordering of hypermetabolism reduction event occurring to brain regions of interest using positron emission tomographs (PET) image. Our model builds on the evidence from disease progression model developed in [17]. In [17], the authors make valuable hypothesis of the cascade of the abnormal clinical events along the AD progression process, however, such a model is still quite coarse and on the conceptual level, and is undergoing experimental validation. Thus, we propose to conduct quantitative analysis, and particularly focus on the FDG-PET imaging data that can measure the hypermetabolism reduction events, to enrich the disease progression model of AD and provide better resolution of the clinical events that happen along the progression process. The way we solve the modeling challenge is to develop an optimal expert knowledge elicitation strategy for identifying pairwise Bayesian network structure from observational data. The strategy can be applied to leverage expert knowledge in learning causal relation from data when the observational data is limited. Thus it is useful for learning the underlying progression mechanism of Alzheimer's disease and other neurodegenerative diseases.

1.1.2 Image guided precision surgery

Image guided precision surgery is another important area in healthcare and medical research, particularly for orthopedic surgeries. This is due to many surgical areas include complex anatomic situations with high-risk structures and high demands for functional and/or aesthetic results. Conventional surgery requires that the surgeon could transfer complex anatomic and surgical planning information derived from their past experiences to guide an unseen case. Such surgical procedure

highly depends on experience and the manual skills of the operator. Recent years, the development of image-guided surgery provides new revolutionary opportunities by integrating presurgical 3D imaging and intraoperative manipulation. In this thesis, our focus will be Computed Tomography (CT) image guided knee tibia surgery.

To be specific, our task is to perform intra-operative identification of the native cruciate ligament insertion sites, which is considered as a requisite for anatomical reconstruction. This task has long been a tremendous challenge since 1) not all surgeons can maintain an acute awareness of the anatomy: about 85% of ACL reconstructions are done by surgeons who perform fewer than 10 cases per year [18] and PCL reconstructions are even less frequently performed by most surgeons; 2) for those who can, factors including the arthroscopic distortion and disappearance of the ligament remnant (naturally or due to a notchplasty procedure) can still cause misidentification of the natural insertion or attachment sites. There is considerable variability of knee anatomy in terms of bone and soft tissue insertion morphology (position, size, and shape) [19]. Sample data from our preliminary study of tibial insertion site morphometry also suggest that simplistic cross-referencing or generalization from one patient to another is likely to lead to non-anatomical tunnel drilling and iatrogenic injury to adjacent tissue structures. Although it may be difficult to gauge the incidence and impact of these iatrogenic injuries as complications of ACL or PCL surgeries, the importance of minimizing the risk of such injuries is readily recognized [20, 21].

The key to anatomic cruciate ligament surgery with minimized risk of iatrogenic injury is an accurate, quantitative knowledge base of the tissue morphology, documenting inter-person variability and specificity vs. uncertainty associated with alternative ways to predict morphometrics. Studies have investigated the quantification of the insertion sites of the cruciate ligaments and other soft tissue components using statistical and quantitative approaches [22, 23, 1]. However, such quantitative analysis and measures generally cannot fully capture the accurate spatial arrangement of soft tissue insertions. The location and morphological measures cannot account for the inter-person variability of cruciate ligament and meniscus insertions, which are mostly characterized by qualitative measures [24, 25]. Advanced imaging techniques, such as 3D CT or MRI, which can be useful to visualize the major structure outline of the knee of a patient clearly, cannot be applied

directly to determine the insertion sites for cruciate ligaments reconstruction. This is because the imaging shows the structure of a knee with serious cruciate ligaments injuries and structure misalignment. While in the surgery, it is crucial to identify the native location of the cruciate ligaments to reconstruct the natural anatomy of the ligament structure. Therefore, one needs inference on the appropriate insertion sites of native cruciate ligaments. Due to the complex anatomy of the knee, the identification of insertion sites of cruciate ligaments in knee reconstruction surgery is still an unsolved problem.

In Chapter 4 we aim to develop a new quantitative analysis method to achieve personalized identification of cruciate ligament and meniscal insertions using patient-specific knee morphological features. In particular, the proposed framework first digitalized outlines of tibia from 3D CT images and aligned the outlines using Generalized Procrustes Analysis (GPA) techniques. It then extracted patient-specific features, trained a supervised structure learning and prediction model, and predicted the centroids of the sites of the cruciate ligament and meniscal insertions using the learned prediction model. The supervised structure learning and prediction model captured the relationship between the spatial arrangement of soft tissue insertions and the patient-specific features extracted from the tibia outlines, which can be easily and reliably measured from CT images. To the best of our knowledge, this is the first supervised machine learning algorithms in knee soft tissue site identification. The proposed learning and prediction framework provides a critical step to achieve the highly demanded personalized surgical planning in cruciate ligament reconstruction.

1.1.3 personalized health behavior recommendation for mobile health

Consumer health intelligence is increasing rapidly. The traditional concept of passive, qualitative "wellness" is being replaced by a proactive, quantitative approach called personal health management, with empowered consumers taking more responsibility for their healthcare. Many of these personalized health management are based on real-world user data and carried out via mobile devices.

Most existing methods in this area only exploit limited value from the data so feedback to individuals is often limited to either overall statistics [26, 27], visualization [28, 29], or generic

suggestions [30, 31]. One exception is [32] that proposed to use the multi-armed Bandit algorithm [33] to automate and personalize the behavioral change plan, which is, however, only able to select from a few pre-specified behavioral change options. Our method is also remotely related to tailored health communication method [30] that is customized for groups of similar users, and the PERSPeCT [34], another tailored health communication that elicits users' behaviors via collaborative filtering. However, accurate dynamic models that can automate the learning have been absent in the literature, not to mention the smart planning engine that can be built on top of the dynamic model.

To address the arising challenges, in Chapter 5, we develop a longitudinal planning framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion. The framework comprises of the following components: 1) A dynamic system learning method – SSMO, which can automatically remove the effects of potential outliers in the dataset, impute the missing values, and conduct model identification; 2) A dynamic planning system that can learn the optimal behavior change plan guided by the individual's own dynamic model. To enhance the quality of the planning, we further formalize users' preferences and needs as constraints, and constructing an action polyhedron for each user by adopting dictionary learning on the actions of peers that form a potential mentor group; and 3) An efficient algorithms to solve the learning and planning problems with specific optimization strategies to ensure the feasibility and robustness of the algorithms. Extensive numerical studies on both synthetic and real-world data demonstrate the utility and efficacy of our methods.

1.2 Thesis Organization and Overview of Methods and Contributions

This dissertation is presented in a multiple manuscript format. Each of the chapters, Chapter 2 to Chapter 6, is written as an individual research paper, including an abstract, a main body, and a conclusion. The relationships among these chapters are depicted in Figure 1.1.

Chapter 2 presents a decomposed version of the k nearest neighbor (DKNN) method for the classification of an unseen object based on the distances to the centroids of the k nearest neighbors [35]. The DKNN has a training process to learn the local-optima-free distance metric by solving

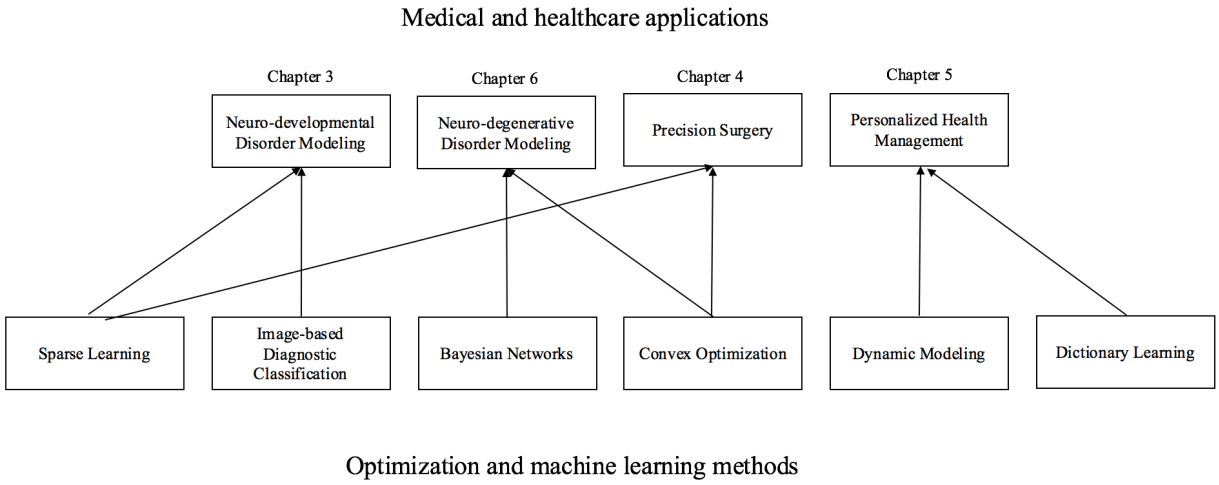


Figure 1.1: Overview of Applications and Methods

a convex optimization problem. The optimization problem not only learns a metric that minimizes classification errors and maximizes the margin between intra-class and inter-class distance. In addition, it also selects important features and removes irrelevant features when L1 regularization is incorporated in the optimization model. The feature selection component is extremely useful for high-dimensional classification problems. Tested on public datasets, the proposed DKNN is competitive and often outperforms traditional machine learning algorithms. The DKNN algorithm could be used in identifying discriminating features for a particular disease (binary cases) or different patient subgroups (multi-class cases).

In Chapter 3, we design a MI-Lasso framework to select discriminating biomarkers from high dimensional medical imaging data, with application in Attention-deficit/hyperactivity disorder (ADHD) classification and seizure prediction [36, 37]. For the diagnostic differentiation of ADHD, recent advances in neuroimaging studies have enabled us to search for both structural (e.g., cortical thickness, brain volume) and functional (functional connectivity) abnormalities that can potentially be used as new biomarkers of ADHD. However, structural and functional characteristics of neuroimaging data, especially magnetic resonance imaging (MRI), usually generate a large number of features. With a limited sample size, traditional machine learning techniques can be problematic

to discover the true characteristic features of ADHD due to the significant issues of overfitting, computational burden, and interpretability of the model. There is an urgent need of efficient approaches to identify meaningful discriminative variables from a higher dimensional feature space when sample size is small compared with the number of features. To tackle this problem, this paper proposes a novel integrated feature ranking and selection framework that utilizes normalized brain cortical thickness features extracted from MRI data to discriminate ADHD subjects against healthy controls. The proposed framework combines information theoretic criteria and the least absolute shrinkage and selection operator (Lasso) method into a two-step feature selection process which is capable of selecting a sparse model while preserving the most informative features. The experimental results showed that the proposed framework generated the highest/comparable ADHD prediction accuracy compared with the state-of-the-art feature selection approaches with minimum number of features in the final model. The selected regions of interest in our model were consistent with recent brain-behavior studies of ADHD development, and thus confirmed the validity of the selected features by the proposed approach. As for the task of seizure prediction, the proposed MI-LASSO method serves as a novel information-theory-guided spare feature selection framework to select the most important EEG features to discriminate epileptic or non-epileptic EEG patterns accurately. The proposed approach were tested on an EEG dataset with 11 patients and 11 normal subjects, achieved an impressive diagnostic accuracy of 90% based on visually-evoked potentials in a human-computer task. This preliminary study indicates that it is promising to provide fast, reliable, and affordable epilepsy diagnostic solutions using short-term interictal EEG signals.

In Chapter 4, we propose a patient-specific model for predicting tibia soft tissue insertions from bony outlines using a spatial structure supervised learning framework [38]. Recreating the natural anatomy in ligament reconstruction is crucial to fully restore the knee joint function and reduce impingement on iatrogenic injury to adjacent structures, yet is subject to the difficulties in locating ligament and other associated soft tissues insertion sites intra-operatively and the high inter-person morphological variability cross patients. In this study we present a new quantitative analysis method capable of achieving personalized identification of cruciate ligament and soft tissue insertions. We craft patient-specific features of tibia outline that can be accurately and reliably

measured from CT images. In addition, we propose a supervised structure learning and prediction model with special inter-dimensional and response structure regularization terms to capture relationship between the spatial arrangement of soft tissue insertions and the patient-specific features extracted from the tibia outlines. In experiment, the proposed model outperforms baseline models and provides an accurate and accessible approach that can be used as the first and the most critical step to achieve personalized surgical planning in cruciate ligament reconstruction.

Next in Chapter 5, we present an algorithm of learning longitudinal planning for personalized health management from daily behavioral data. Mitigating globally emerging health problems such as obesity needs scalable solutions that can promote healthier lifestyles outside of clinical settings. Such scalable solutions, while targeting general population, need to automatically provide personalized behavior change plans that fit an individual's preferences and needs. There has been fast-growing development of sensing devices and applications for continuous monitoring of human behavior (such as physical activity and food intake) and health status such as BMI. However, there are challenges to translate these noisy and dynamic behavioral data into personalized planning. To address both challenges, we develop a systematic framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion to translate users' behavioral data into deeply personalized health planning. We further apply our proposed model on a real-world user behavioral dataset, which demonstrates the promising utility and efficacy of our method.

In Chapter 6, we propose an optimal expert knowledge elicitation strategy for identifying Bayesian network structure from observational data. Bayesian network (BN) has been a popular tool for gaining mechanistic understanding of variables by revealing how the variables influence each other. It has been found very effective in a few studies in quality control and process monitoring. However, for complex problems where the structure of a BN is unknown, a common approach is to learn the BN structure from observational data. A fundamental bottleneck of this approach is that observational data can only be used to discover part of the influential relationships among variables. To overcome this problem, we propose to combine observational data and expert knowledge. To the best of our knowledge, our approach is the first of its kind that can automate the expert elicitation process and collect the most informative expert knowledge, optimally matched to

the observational data, to learn the BN structure.

Chapter 2

OPTIMIZATION MODELS FOR FEATURE SELECTION OF DECOMPOSED NEAREST NEIGHBOR

2.1 abstract

The traditional k-nearest neighbor (KNN) method classifies an object by the majority vote of its neighbors, and only one parameter k is to be optimized. We propose a decomposed version of the k nearest neighbor (DKNN) method, which classifies an unseen object based on the distances to the centroids of the k nearest neighbors. DKNN has a training process to learn the local-optima-free distance metric by solving a convex optimization problem. The optimization problem not only learns a metric that minimizes classification errors and maximizes the margin between intra-class and inter-class distance. In addition, it also selects important features and removes irrelevant features when L1 regularization is incorporated in the optimization model. The feature selection component is extremely useful for high-dimensional classification problems. Tested on public datasets, the proposed DKNN is competitive and often outperforms traditional machine learning algorithms.

2.2 Introduction

The k-nearest neighbor (KNN) method classifies an object by a majority vote of its neighbors, with the object being assigned to the most common class amongst its k nearest neighbors [39, 40]. When $k = 1$, the object is simply assigned to the class of its nearest neighbor and the decision boundaries are concatenated Voronoi tessellation. In this paper, we propose a new decomposed k nearest neighbor method (DKNN), of which the classification rule is based on the k nearest neighbors (called prototype) *from each class*. The prototype of a particular class can be determined from the k nearest neighbors of the same class through various linkage functions. A simple choice of

linkage function is to use the centroid of k nearest neighbors from the same class.

In the traditional KNN, all feature directions are equally considered in the classification rule [41]. The assumption of isotropy is often invalid and undesirable, as the neighborhood may extend into the direction of irrelevant features and include samples of different classes [42, 43]. To deal with such an issue, efforts have been put into selecting or scaling features via local discriminant analysis to improve classification [44]. Some other works try to solve a similar clustering problem from metric learning perspective [45, 46, 47]. Although there are many extensions of KNN developed in the literature [48, 49, 50, 51, 52, 53], the feature selection component in the KNN framework has not been proposed in a convex metric learning setting. Most methods proposed in this field are either not convex, or do not have a shrinkage method that produces a sparse model or cannot be generalized beyond 1NN.

Based on the DKNN rule, we develop a new convex optimization model for feature selection as well as introduce a regularization term (Frobenius [54, 55] or L1 [54]) to increase the robustness of our classification model (i.e., reducing prediction variance and improving prediction accuracy). In this paper, we use the centroid linkage function, which allows the feature selection model to be framed as a convex optimization problem. In this paper, we also show strong connections between our DKNN optimization model with two other state-of-the-art classification algorithms, linear discriminant analysis (LDA) [56] and support vector machine (SVM) [57].

The rest of the paper is organized as follows. In Section 2.3 we introduce the prototype based similarity measure for the DKNN algorithm. In Section 2.4 we present our DKNN convex optimization models for feature selection. Then in Section 2.5 we describe the connections of DKNN to other state-of-the-art algorithms including KNN, SVM and LDA. In section 2.6, we show the performance of our algorithm using synthetic datasets and 8 datasets from UCI Machine Learning Repository. Then we conclude our work in section 4.5.

2.3 Prototype Based Similarity Measure

For the DKNN algorithm, classification is made based on the modified distance to the centroid of k nearest neighbors in each class. Similar to k-means clustering [58, 59, 60], the centroid for

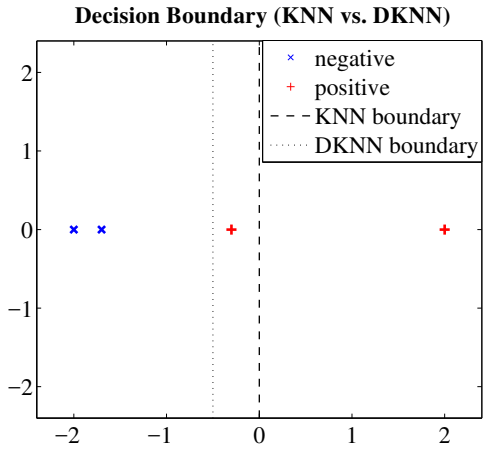


Figure 2.1: Decision Boundary (KNN vs. DKNN)

k nearest neighbors of data sample x in class c, denoted as $\mu_c(x)$, acts as the representative, aka prototype, for each class. DKNN classifies a new data point to the class of the closest centroid. Operationally, DKNN divides the baseline data I into positive and negative sample denoted as I^+ and I^- respectively. For test sample x, we find the k nearest neighbors from each class (positive and negative). The distance from x to every positive and negative sample is commonly measured by the Euclidean distance although other distance measures can be used. Then the centroids of the k samples of the same class are computed as the class prototypes. The distance from x to centroid $\mu_c(x)$ is denoted as $d(x, \mu_c(x))$.

Fig. 1 shows an example of 4 training samples on a straight line in a 2D plane with 2 positive samples on the right and 2 negative samples on the left. The decision boundary for 3NN is a perpendicular line midway between the leftmost negative sample and rightmost positive sample. It actually does not matter where the two samples in between locate between the interval, and the decision boundary for 3NN will not change. Our method, if using 2 nearest neighbors from each class, takes full considerations of equally sized samples from each class to make the classification decision. We believe that the new prototype based classification rule takes fair amount of information from each class, and therefore has better performance in many cases.

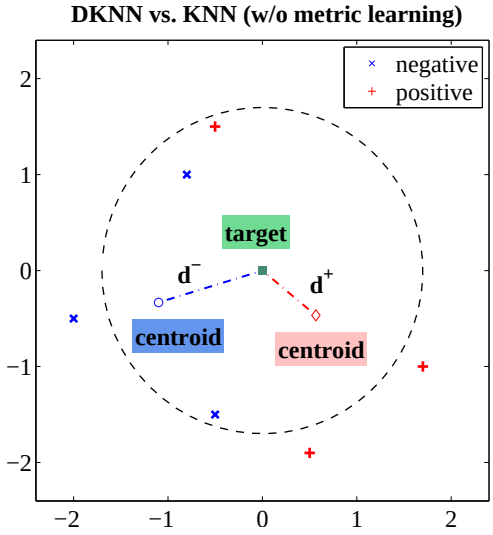


Figure 2.2: DKNN vs. KNN Without Metric Learning

Fig. 2. shows an example that KNN ($k = 3$) will classify the target point as negative since in the dotted circle there are two negative samples but only one positive sample. However, DKNN will calculate the centroids of three nearest neighbors from I^+, I^- respectively and then classify the target point as positive since the centroid of positive samples is closer.

2.3.1 Distance Measure

We proposed a modified distance measure for DKNN, defined by $f_c(x) = d_A(x, \mu_c(x))^2 - \log \pi_c$. The first term d_A is the squared Mahalanobis distance [61] and the second term π_c is a correction bias related to the class prior distribution. In the formulation, we take logarithm of π_c to have a better representation in Bayesian formula. The Mahalanobis distance uses a positive semi-definite matrix A to define the distance of two vectors u, v in $\mathcal{R}^p$ as $d_A(u, v) = ||u - v||_A = \sqrt{(u - v)^T A(u - v)}$. This metric is a generalized version of the Euclidean distance. It can be considered as a similarity measure between two random vectors u, v with the distributions that have the same covariance matrix A^{-1}. Also note that the Mahalanobis distance in our modeling is flexible to represent various feature scaling and selection choices. A flexible model can reduce

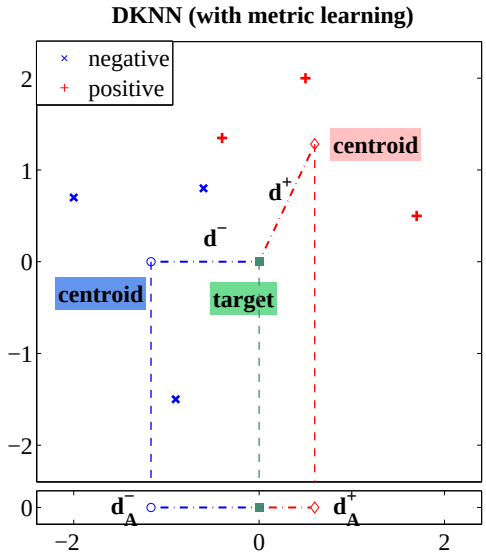


Figure 2.3: DKNN with Metric Learning

the bias imposed by the model constraints as well as decorrelate features using a covariance matrix $\Sigma = A^{-1}$.

Fig. 3. shows how the Mahalanobis distance can be used in reducing dimensions of feature space and correcting classification errors. DKNN would fail to classify the target point as positive if the distances to centroids are measured by the Euclidean distance (e.g. d^+ and d^-). The Mahalanobis distance can help correct the classification error by only selecting the projection on x-axis as the distance measure. Under the new distance measure, DKNN will classify the target point as positive since $d_A^+ < d_A^-$.

2.3.2 Remarks on Alternative Similarity Measures

Given a distance matrix, test sample x and its k nearest neighbors from each class, we may choose different linkage functions to make decision. After computing pairwise modified distances between x and its neighbors, we can measure the similarity by (1) the maximal distance from each class, (2) the minimal distance from each class, (3) the average distance from each class, or (4) the distance

to the centroid of k nearest neighbors from each class. Among them, (1), (2) do not result in a convex problem in our formulation unless $k = 1$. Only (3) and (4) result in convex formulations, but (3) may have higher computational cost than (4). In this paper, we choose (4) the distance to the centroid of k nearest neighbors from each class in our DKNN model. Later, we will demonstrate that (4) has closer connection to LDA.

2.4 Optimization Models for Feature Selection of DKNN

The inputs of our optimization models are the intra-class and inter-class distances and class prior distribution π_c in the baseline data. Given a training sample vector x with class label vector y, the intra-class distance is denoted by $f_y(x)$ and the inter-class distance is denoted by $f_{c\neq y}(x)$. During the training process, we try to find a distance metric A that correctly classifies the training samples and further maximizes the margin between the intra-class distance and the inter-class distance. The optimization model for DKNN feature selection is given by

$$\min \quad \alpha \sum_{\forall i \in I^+} \epsilon_i + (1-\alpha) \sum_{\forall i \in I^-} \epsilon_i \tag{2.1}$$

$$\text{s.t.} \quad f_{y_i}(x_i) + 1 - \epsilon_i \leq f_{c\neq y_i}(x_i),\ \forall i \in I \tag{2.2}$$

$$\epsilon_i \geq 0,\ \forall i \in I \tag{2.3}$$

$$f_c(x) = ||x - \mu_c(x)||_A^2 - \log \pi_c, \tag{2.4}$$

where $0 < \alpha < 1$ is the important weight of false positive, $\epsilon_i = 0$ if DKNN correctly classifies training sample i with sufficient margin. The objective function in Eq. (2.1) is to minimize the sum of weighted false positive and false negative fractions. Specifically, the model minimizes the weighted fractions of activated surplus variables for both positive and negative classes. There is a set of constraints in Eq. (2.2) to ensure that each training sample reserves enough margin between the intra-class distance and the inter-class distance. Eq. (2.4) represents the DKNN distance for class c (the square of Mahalanobis distance and the adjustment for class distribution). The elements of A are linear in Eq. (2.4) and therefore constraint function Eq. (2.2) is convex. Later we will show the optimization problem is a convex program.

2.4.1 Regularization

Without regularization, classification models are often sensitive to overfitting and have poor generalization power. A common approach is to introduce a regularization term, which represents the norm of the predictor variables, in the objective function. In DKNN, we regularize the Mahalanobis distance matrix in the objective function as in Eq. (2.5), where β is a penalty tuning parameter. For the choice of norm, we may consider the Frobenius norm or apply the L1 norm on the flattened matrix. The regularized formulation is given by:

$$\min \quad \alpha \sum_{\forall i \in I^+} \epsilon_i + (1-\alpha) \sum_{\forall i \in I^-} \epsilon_i + \beta ||A|| \tag{2.5}$$

$$\text{s.t.} \quad f_{y_i}(x_i) + 1 - \epsilon_i \leq f_{c \neq y_i}(x_i), \ \forall i \in I \tag{2.6}$$

$$\epsilon_i \geq 0, \ \forall i \in I \tag{2.7}$$

$$f_c(x) = ||x - \mu_c(x)||_A^2 - \log \pi_c. \tag{2.8}$$

Frobenius Norm

The Frobenius norm is a common L2-norm defined as the square root of the sum of the absolute squares of matrix elements [55]. The Frobenius norm, defined by $||A||_F = \left(\sum_{i=1}^{p} \sum_{j=1}^{p} |a_{ij}|^2 \right)^{1/2}$, is easy to calculate and invariant in rotation. It will be used in regularizing the DKNN models when $p < n$ to increase prediction stability. However, L2-norms usually allow predictor variables with many small nonzero values, and therefore do not have feature selection when $p \gg n$.

L1 Norm

The vector norm of matrix treats flattened matrix as a vector. The L1 vector norm on A, defined by $||A||_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |a_{ij}|$, is the sum of absolute values of all matrix elements. The L1 regularization has been extensively studied and used in recent years for the $p \gg n$ problem (i.e., the number of features is far greater than the number of training samples). Such problem has an insufficient degree of freedom to estimate the full model. The L1-norm has been used as a shrinkage penalty in LASSO [62]. It penalizes and forces some variables to be zero, hence effectively selects a sparse

model. In our case, the L1-norm reduces non-zero elements in the matrix A, selects and retains useful features in the distance calculation. Even when $p < n$, the L1-norm can be used to remove irrelevant features.

2.4.2 Multiclass Extension

Our DKNN feature selection model in Eq. (2.5)-(2.8) can be naturally extended to multi-class classification. The extension only requires that the left hand side of Eq. (2.6) is less than or equal to $f_c(x_i)$ of all other classes ($c \neq y_i$) as shown in Eq. (2.10). Given α_c is the weight of error for class c, the multiclass model of DKNN feature selection is given by

$$\min \quad \sum_{i=1}^{n} \alpha_{y_i} \epsilon_i + \beta ||A|| \tag{2.9}$$

$$\text{s.t.} \quad f_{y_i}(x_i) + 1 - \epsilon_i \leq \min_{c \neq y_i} f_c(x_i), \ \forall i \in I \tag{2.10}$$

$$\epsilon_i \geq 0, \ \forall i \in I \tag{2.11}$$

$$f_c(x) = ||x - \mu_c(x)||_A^2 - \log \pi_c. \tag{2.12}$$

2.4.3 Convex Optimization and Model Implementation

Distance metric learning with a convex objective function was first proposed to learn a Mahalanobis metric for clustering [45]. Distance metric learning over positive semidefinite metric space was also explored in previous work [63, 64, 46]. In this paper, the proposed DKNN feature selection models are formulated as a semidefinite program (SDP), which is based on the distance metric learning between intra-class and inter-class distances. Generally SDP minimizes a linear objective function over the intersection of the cone of positive semidefinite matrices with an affine space. SDP is convex, and its global optimum can be efficiently computed.

Proposition 1 *DKNN feature selection model is SDP and convex.*

Proof With decision variables A and ϵ, the convex optimization problem is of the form:

$$\min \quad g_0(A, \epsilon) \tag{2.13}$$

$$\text{s.t.} \quad g_i(A, \epsilon) \leq 0, \ \forall i \in 1, \ldots, m, \tag{2.14}$$

where $g_0, \ldots, g_m$ are convex functions. In our case,

$$g_0(A, \epsilon) = \sum_{i=1}^{n} \alpha_{y_i} \epsilon_i + \beta ||A||, \tag{2.15}$$

$$\begin{aligned} g_i(A, \epsilon) &= ||x_i - \mu_c(x_i)||_A^2 - \log \pi_c + 1 - \epsilon_i \\ &\quad -\min_{c \neq y_i} \left(||x_i - \mu_c(x_i)||_A^2 - \log \pi_c \right), \\ &\quad i \in 1, \ldots, |I|, \end{aligned} \tag{2.16}$$

$$g_i(A, \epsilon) = -\epsilon_i, i \in |I| + 1, \ldots, 2|I|. \tag{2.17}$$

Eq. (2.17) are affine functions so they are convex. $||x_i - \mu_c(x_i)||_A^2$ is also affine for each entry a_{kl} of matrix A. The minimum of several affine functions is concave but, with a minus sign, $-\min_{c \neq y_i} (||x_i - \mu_c(x_i)||_A^2 - \log \pi_c)$ is convex. $g_i, i \in 1, \ldots, |I|$ is an addition of 3 convex terms, and is thus convex. The element-wise norm of matrix A is convex, and $\sum_{i=1}^{n} \alpha_{y_i} \epsilon_i$ is affine; therefore, the objective function is convex. Since the Mahalanobis distance $||x - \mu_c(x)||_A$ imposes that A is real symmetric positive semi-definite, constraint Eq. (2.8) is linear, our model is SDP.

To solve our DKNN model, we used a convex program package called CVX [65, 66]. Since our model is SDP, we solve it in semidefinite programming mode and use linear matrix inequality (LMI) notation $A \geq 0$ to denote $A \in S_+^p$. The training procedure is summarized in Algorithm **??**. The evaluation procedure is summarized in Algorithm **??**.

2.4.4 Multi-pass Extension

For DKNN, the target neighbors must be priori specified. Without prior knowledge, we use Euclidean distances to determine nearest neighbors before learning the Mahalanobis distance metric A. Since the actual nearest neighbors may change as a result of the linear transformation of the input space[46], we extend DKNN to multi-pass (multi-iteration) and iteratively learn a series of Mahalanobis distance metric $A_i, i \in \mathcal{N}$. For $n-$pass DKNN, we start with A_0 as identity matrix, during i^{th} pass of DKNN, we select the nearest neighbors using Mahalanobis distance metric A_{i-1}, and learn new metric A_i by solving a convex optimization program (2.5)-(2.8). During evaluation,

we also use A_{i-1} to select the nearest neighbors and calculate $f_c(x)$ with A_i. We iterate the process until $i = n$.

In Section 2.6.1, we evaluate such iterative learning approach by performing multi-pass DKNN on the synthetic dataset. We will compare the results of $n-$pass DKNN (where $n = 1, 2, 3, 4$) and see multi-pass works well and generally improve testing accuracy.

2.5 Connection to Other State-of-the-Art Algorithms

2.5.1 Connection to KNN

As a variant of KNN, our DKNN model share many similarities with the original KNN. Some conditions of KNN and DKNN always provide identical classification results. To see their similarity, we first define the parameters $\hat{k}$ for traditional KNN and $\bar{k}$ for DKNN as the number of their k nearest neighbors. Without metric learning, we can establish the following results.

Proposition 2 *For $\hat{k} = \bar{k} = 1$, both KNN and DKNN yield an identical classification result.*

Proof Let d_1^+ be the distance of the test data to its nearest neighbor in the positive class and d_1^- is the distance of the test data to its nearest neighbor in the negative class. In other words, $d_1^+ \leq d_i^+,\ i \in I^+$ and $d_1^- \leq d_j^-,\ j \in I^-$, where d_i^+ is the distance of the test data to positive baseline sample i, and d_j^- is the distance of the test data to negative baseline sample j. If $d_1^+ < d_1^-$, DKNN classifies the test data to positive or vice versa. For KNN, the test data is classified to positive when d_1^+ is the smallest distance to all baseline samples. That is, $d_1^+ \leq d_i,\ i \in I$, which also satisfies $d_1^+ \leq d_i^+$ and $d_1^- \leq d_i^-$. We can conclude that if DKNN classifies the test data to positive, KNN must classify the test data to positive. It is clear that the reverse is also true and a similar argument can be established for the classification of the negative case.

Proposition 3 *For $\hat{k} = 3, \bar{k} = 2$, both KNN and DKNN yield an identical positive classification if and only if $d_1^+ \leq d_1^-$; an identical false classification if and only if $d_1^- \leq d_1^+$*

Proof Assume that the test data sample is classified to the positive class. Based on its definition, DKNN classifies the test sample to positive if and only if the following condition is satisfied:

$d_1^+ + d_2^+ \leq d_1^- + d_2^-$, where $d_1^+ \leq d_2^+ \leq d_i^+,\ i \in I^+$ and $d_1^- \leq d_2^- \leq d_i^-,\ i \in I^-$. We shall establish the equivalence proof of KNN in following two cases.

a) All $\hat{k} = 3$ nearest neighbors are from the positive class. This case implies that $d_1^+ \leq d_2^+ \leq d_3^+ \leq d_i,\ i \in I$. If $d_1^+ \leq d_2^+ \leq d_3^+ \leq d_i,\ i \in I$ then $d_1^+ + d_2^+ \leq d_i + d_j,\ i, j \in I$. Thus, the positive classification of DKNN is satisfied and it must classify the test sample to positive.

b) Two nearest neighbors are from the positive class and one is from the negative class. This case implies that $d_1^+ \leq d_2^+ \leq d_i,\ i \in I$ and $d_1^- \leq d_i,\ i \in I$. For DKNN to classify the test sample to positive, $d_1^+ + d_2^+ \leq d_1^- + d_2^-$ must be true. Given $d_1^+ \leq d_1^-$, we can say that if $d_2^+ \leq d_2^-$ is true, DKNN classifies the test sample as positive. As $d_2^+ \leq d_i,\ i \in I$, one can show that $d_2^+ \leq d_2^-$ is always true.

We can now conclude that, for both cases, if DKNN classifies the test data to positive, KNN must classify the test data to positive. It is obvious that the reverse is also true and a similar argument can be established for the classification of the negative case.

The difference between DKNN and KNN is that traditional KNN only counts majority votes in k nearest neighbors, while DKNN also considers the distance of neighbors through centroid linkage. DKNN has advantages over KNN in that DKNN also learns a metric matrix A that minimizes the classification error and maximizes the margin of decision boundary. KNN does not have the training process to find an optimal metric. With the L1-norm, DKNN also incorporates the feature selection functionality in an optimization model.

2.5.2 Connection to Gaussian Linear Discriminant Analysis (LDA)

LDA uses Bayes' theorem to estimate the class posteriors for classification. With the conditional density $f_c(x)$ and prior probability π_c for class c, Bayes' theorem yields $P(c|x) = \frac{\pi_c f_c(x)}{\sum \pi_k f_k(x)}$. LDA is an approximation of a Bayesian classifier with the assumption that data samples x are from Gaussian distributions with the same covariance matrix. Several research attempts have been made

to extend Gaussian assumption to mixtures of Gaussians [44] or nonparametric density estimates [67], and relax the same covariance matrix constraint [68, 69].

In Gaussian LDA, data points in each class are assumed to be Gaussian-distributed. The optimal decision boundary is determined by discriminant functions $\delta_c(x) = x^T\Sigma^{-1}\mu_c - \frac{1}{2}\mu_c^T\Sigma^{-1}\mu_c + \log\pi_c$, where μ_c is the mean for all samples in the cth class, π_c is the prior probability that an observation belongs to the cth class, and Σ^{-1} is the inverse of Gaussian covariance matrix.

Now we redefine $\mu_c(x)$ as the mean of the nearest neighbors of x in class c. We replace Σ^{-1} with $2A$, the Mahalanobis distance matrix, and replace $\delta_c(x)$ with $-f'_c(x)$ to reflect the difference from a Gaussian LDA discriminant function. $f'_c(x)$ is defined as follows:

$$f'_c(x) = -2x^T A\mu_c(x) + \mu_c(x)^T A\mu_c(x) - \log\pi_c. \tag{2.18}$$

Eq. (2.6) allows adding or removing a constant specific to x. $f_c(x) = f'_c(x) + x^T Ax = ||x - \mu_c(x)||_A^2 - \log\pi_c$.

The classification formulation is equivalent to

$$\min \quad \alpha \sum_{\forall i \in I^+} \epsilon_i + (1-\alpha) \sum_{\forall i \in I^-} \epsilon_i + \beta||A|| \tag{2.19}$$

$$\text{s.t.} \quad \max_{c \neq y_i} \delta_c(x_i) + 1 - \epsilon_i \leq \delta_c(x_i),\ \forall i \in I \tag{2.20}$$

$$\epsilon_i \geq 0,\ \forall i \in I. \tag{2.21}$$

$f_k(x)$ has the meaning of the DKNN distance for class c. For each training sample, we want to minimize the intra-class distance of c nearest neighbors by choosing a Mahalanobis distance matrix A. In connection to LDA, $f_c(x)$ has additional Bayesian interpretation:

$$P(c_i|x) = \frac{e^{-f_{c_i}(x)}}{\sum e^{-f_c(x)}} = \frac{\pi_{c_i} e^{-d_A^2(x,\mu_{c_i}(x))}}{\sum \pi_c e^{-d_A^2(x,\mu_c(x))}}.$$

Although the discriminant function of LDA may have a similar form with the distance function of DKNN, the two approaches differ in many ways. Firstly, LDA assumes a Gaussian distribution on all observations, while DKNN is non-parametric and only concerned about k nearest neighbors from each class. Secondly, LDA produces a linear decision boundary, while the decision boundary

of DKNN can be highly non-linear. Thirdly, although DKNN may use all training samples in each class to calculate the centroid, i.e., $\mu_c(x) = \mu_c$, DKNN and LDA are still different in the way A or Σ is estimated. Σ is estimated by Gaussian distribution whereas A is estimated by maximizing the margin of the decision boundary.

2.5.3 Connection to Support Vector Machine (SVM)

SVM algorithm is amongst the most widely studied machine learning algorithms [57]. Without kernel trick, it is called support vector classifier, which usually trains on previously known data to differentiate groups/classes of data samples by maximizing the decision boundary between the two classes.

Given feature vector x, support vector classifier predicts the label by $f(x) = \arg\max_c f_c(x)$, or $\hat{y} = \arg\max_c w_c^T x$. The two-class classifier may use 2 affine functions, $f_c(x) = w_c^T x$, $c = 1, 2$. To correctly classify the training samples, the classifier needs $f_y(x) > \max_{k \neq y} f_c(x)$. With a safe margin, we have the following non-strict inequality: $f_y(x) \geq 1 + \max_{c \neq y} f_c(x)$. This motivates the following formulation:

$$\min \quad L(W) = \sum_{i=1}^{m} \epsilon_i \tag{2.22}$$

$$\text{s.t.} \quad \max_{k \neq y_i} f_k(x_i) + 1 - \epsilon_i \leq f_{y_i}(x_i),\ \forall i \in I \tag{2.23}$$

$$\epsilon_i \geq 0,\ \forall i \in I \tag{2.24}$$

$L(W) = \sum_{i=1}^{m} \epsilon_i$ is the loss function. For affine functions $f_c(x) = w_c^T x$, this formulation is convex.

The above formulation resembles Eq. (2.19)-Eq. (2.21). The difference is that support vector classifier uses a fixed vector w_c per class, while DKNN uses more localized $A\mu_c(x)$ per class per x as in Eq. (2.18). DKNN would perform better for non-linear decision boundary.

The kernel trick was shown very useful in handling non-linear cases with SVM. It maps data into a higher dimensional feature space to perform classification. DKNN can also benefit from the kernel method by replacing x with $\phi(x)$.

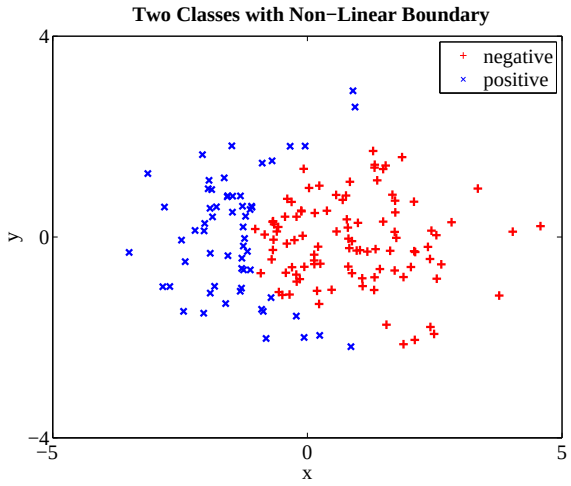


Figure 2.4: Two-Dimensional Gaussian Data Plot

2.6 Computational Results

2.6.1 Synthetic Data

To compare the classification performance of our DKNN with that of other state-of-the-art classification algorithms (i.e., LDA, SVM and CART), we randomly generated 3 synthetic datasets. In Dataset 1(Fig. 2.4), the data points are merged from two standardized Gaussian classes separated by two units in x-axis. The class boundary is roughly an arc splitting 160 points into 60 negative points on the left and 100 positive points on the right. Dataset 2 (Fig. 2.5) has the same data points as in Dataset 1, and is augmented with 1 additional dimension following an independent standard Gaussian distribution. Dataset 3 has the same data points as in Dataset 1, and is augmented with 2 additional dimensions also following an independent standard Gaussian distribution.

Table 2.1 compares classification accuracies obtained by DKNN with different pass parameters. The parameter k was optimized by iterating its value over $k = 1 \ldots 10$. We evaluate multi-pass (with pass parameter $= 1 \ldots 4$) DKNN on all synthetic datasets. The best values will be used in comparing performance of DKNN with other algorithms.

Table 2.2 compares classification accuracies obtained by DKNN and those by other state-of-

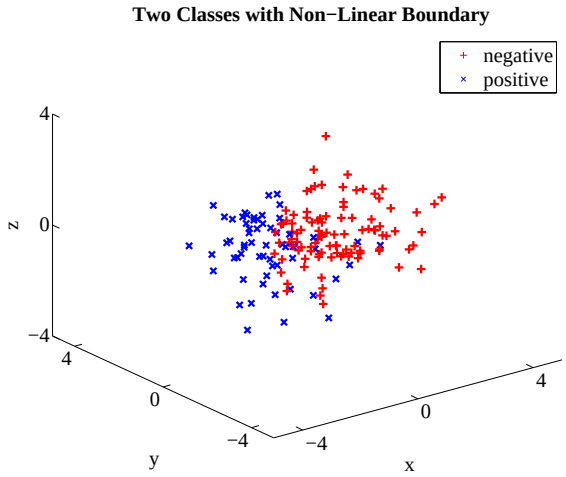


Figure 2.5: Two-Dimensional Gaussian with One Additional Noise Dimension

Table 2.1: LOOCV Accuracy of Multi-pass DKNN on Synthetic Datasets

Dataset	**Dimension**	**1 Pass**	**2 Pass**	**3 Pass**	**4 Pass**
1)	2	0.981	0.981	0.988	0.981
2)	3	0.963	0.981	0.988	0.981
3)	4	0.963	0.981	0.981	0.981

Table 2.2: LOOCV Accuracy on Synthetic Datasets

Dataset	**Dimension**	**DKNN**	**KNN**	**LDA**	**SVM**	**CART**
1)	2	0.988	0.981	0.831	0.831	0.869
2)	3	0.988	0.931	0.831	0.831	0.869
3)	4	0.981	0.900	0.813	0.819	0.869

the-art algorithms. For KNN, the parameter k was optimized by iterating its value over $k = 1 \ldots 10$. For other algorithms, standard (default) parameters were used. In all cases, the results suggest that there is a limitation of linear decision boundary that reflects on poor performance of LDA and SVM. The decision tree algorithm CART has better performance than LDA and SVM since its decision boundary can be non-linear. In Dataset 1 (2D case), DKNN yields the best leave-one-out cross-validation (LOOCV) result when $k = 1$. The benefit of non-parametric learning is clear in this scenario. In Dataset 2 (3D case) when we add an non-informative dimension with Gaussian noise, the performance of KNN deteriorates. The distance measure of KNN was negatively influenced by the variance of non-informative dimension whereas DKNN was able to maintain and even improve the LOOCV performance by learning the new metric. In Dataset 3 (4D case) two non-informative features were added, the performance of KNN further deteriorates. DKNN was able to zero-out two noisy dimensions so that the discriminative power is less impacted by the noisy dimensions.

2.6.2 UCI Machine Learning Repository Datasets

To assess the classification performance, we also tested our DKNN on eight additional datasets that were selected from the University of California Irvine (UCI) Machine Learning Repository [70]. The eight datasets are from real natural science, engineering, or medical domains. Table 2.3 provides the characteristics of the selected UCI datasets.

In this study, we evaluated the classification performance using accuracy as our metric (the percentage of correct predictions) based on 10 runs with a 70/30 split for the training and testing sets. Specifically, for a sample with n subjects, the procedure consists of 10 runs. In each run, we trained a classification model from $0.7n$ subjects, used the model to predict the remaining $0.3n$ subjects, and evaluated the prediction with the actual class label. Testing accuracy was calculated as the number of correctly predictions divided by the total number of subjects. For DKNN and KNN, the parameter k was optimized on the training data by iterating its value over $k = 1 \ldots 10$. Table 2.4 compares the classification performance of DKNN with KNN, SVM, and LDA. The results in the table show that DKNN yielded the best or second best testing accuracy amongst all

Table 2.3: Facts of UCI Datasets

Dataset	**#samples**	**#class**	**#features**	**Class distribution**
Ionosphere	351	2	34	0.36, 0.64
Sonar	208	2	60	0.47, 0.53
BUPA Liver	345	2	6	0.42, 0.58
Balance Scale	625	3	4	0.46, 0.08, 0.46
Wine	178	3	13	0.33, 0.40, 0.27
Iris	150	3	4	0.333, 0.333, 0.333
Seeds	210	3	7	0.333, 0.333, 0.333
Knowledge	258	4	5	0.093, 0.322, 0.341, 0.244

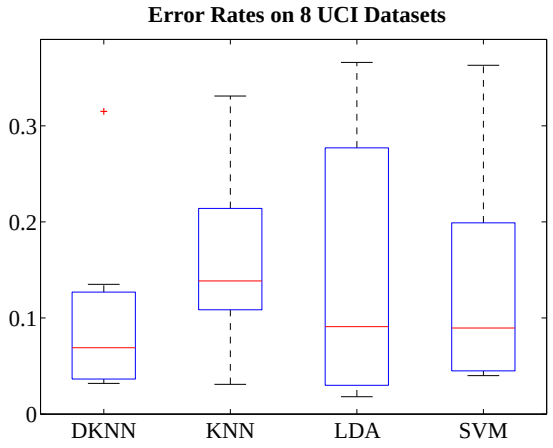


Figure 2.6: Error Rates on 8 UCI Datasets

classification techniques. Overall, the average error rate across all 8 datasets of DKNN is 0.101, which is a lot lower than that of KNN (0.161), LDA (0.148) and SVM (0.139). Boxplot of error rates are shown in Fig. 2.6.

Table 2.4: Comparison of Testing Accuracies Between DKNN and Other State-of-the-art Algorithms.

Dataset	**DKNN**	**KNN**	**LDA**	**SVM**
Ionosphere	0.881 $\pm$0.029	0.858 $\pm$0.036	0.865 $\pm$0.037	0.851 $\pm$0.032
Sonar	0.865 $\pm$0.049	0.865 $\pm$0.046	0.737 $\pm$0.052	0.751 $\pm$0.047
BUPA Liver	0.685 $\pm$0.039	0.669 $\pm$0.020	0.634 $\pm$0.052	0.637 $\pm$0.051
Balance Scale	0.910 $\pm$0.022	0.875 $\pm$0.033	0.709 $\pm$0.027	0.899 $\pm$0.022
Wine	0.952 $\pm$0.020	0.732 $\pm$0.053	0.976 $\pm$0.025	0.950 $\pm$0.029
Iris	0.967 $\pm$0.021	0.969 $\pm$0.016	0.982 $\pm$0.014	0.960 $\pm$0.025
Seeds	0.960 $\pm$0.028	0.908 $\pm$0.028	0.964 $\pm$0.017	0.960 $\pm$0.022
Knowledge	0.968 $\pm$0.022	0.840 $\pm$0.042	0.953 $\pm$0.022	0.922 $\pm$0.039

2.7 Conclusion

In this paper, we present a new classification algorithm with optimization models for feature selection. The algorithm is called the decomposed k nearest neighbor (DKNN) method, which takes into account k nearest neighbors from each class and makes the classification based on the distances to the centroids of the k nearest neighbors. DKNN has a training process to learn the local-optima-free distance metric by solving a convex optimization problem. The optimization problem not only learns a metric that minimizes classification errors and maximizes the margin between intra-class and inter-class distance, but also selects important features and removes irrelevant features when L1 regularization is used. We demonstrate that DKNN can be naturally extended to multi-class classification. We also provide new theoretical insights showing that DKNN is a generalized version of KNN, and has strong connections to SVM and LDA. Beyond the basic framework of DKNN, we also describe its multipass extension, which learns a series of Mahalanobis distance metrics during passes (iterations) and iteratively re-estimates the target neighbor assignments. Computa-

tional results also demonstrate that our DKNN algorithm provides accurate classification results (both binary and multi-class classification) when compared to other state-of-the-art classification algorithms.

Chapter 3

A MI-LASSO FRAMEWORK WITH APPLICATION IN ADHD CLASSIFICATION AND SEIZURE PREDICTION

3.1 Abstract

Today diagnosis of attention deficit hyperactivity disorder (ADHD) still primarily relies on a series of subjective evaluations that highly rely on a doctor's experiences and intuitions from diagnostic interviews and observed behavior measures. An accurate and objective diagnosis of ADHD is still a challenge and leaves much to be desired. Many children and adults are inappropriately labeled with ADHD conditions whereas many are left undiagnosed and untreated. Recent advances in neuroimaging studies have enabled us to search for both structural (e.g., cortical thickness, brain volume) and functional (functional connectivity) abnormalities that can potentially be used as new biomarkers of ADHD. However, structural and functional characteristics of neuroimaging data, especially magnetic resonance imaging (MRI), usually generate a large number of features. With a limited sample size, traditional machine learning techniques can be problematic to discover the true characteristic features of ADHD due to the significant issues of overfitting, computational burden, and interpretability of the model. There is an urgent need of efficient approaches to identify meaningful discriminative variables from a higher dimensional feature space when sample size is small compared with the number of features. To tackle this problem, this paper proposes a novel integrated feature ranking and selection framework that utilizes normalized brain cortical thickness features extracted from MRI data to discriminate ADHD subjects against healthy controls. The proposed framework combines information theoretic criteria and the least absolute shrinkage and selection operator (Lasso) method into a two-step feature selection process which is capable of selecting a sparse model while preserving the most informative features. The experimental results showed that the proposed framework generated the highest/comparable ADHD prediction accuracy

compared with the state-of-the-art feature selection approaches with minimum number of features in the final model. The selected regions of interest in our model were consistent with recent brain-behavior studies of ADHD development, and thus confirmed the validity of the selected features by the proposed approach.

3.2 Introduction

Attention deficit hyperactivity disorder (ADHD) is among the most common child and adult neurodevelopmental disorder. ADHD symptoms include inattention, hyperactivity, and impulsivity. It affects approximately $5 - 10\%$ of all school-age children and nearly 5% of adults on their motor, cognitive, and emotional development [4]. Diagnosis of ADHD still remains a challenge, requiring long term and extended involvement from clinicians, parents and teachers. Clinicians rely heavily on experiences and intuitions when conducting diagnostic interview and observational measures. A delay or incorrect diagnosis of ADHD could have a significant negative impact on a patient's social and emotional development, while an early and accurate detection of ADHD can strongly influence the course of the condition development by delivery of appropriate treatments to the patient. In addition to the traditional clinical diagnosis, there is a pressing need to find a set of more discriminative and objective features to characterize ADHD that can be used to facilitate ADHD diagnosis.

Previous studies on the etiology of ADHD are mostly based on structural or functional neuroimaging research of group level (ADHD vs. control) differences. Some informative features extracted are blood oxygenation level dependent (BOLD) signals from functional magnetic resonance imaging (fMRI) data [5], wavelet synchronization likelihoods extracted from electroencephalography (EEG) data [6], rolandic spikes from EEG data [7], brain volume measure extracted from magnetic resonance imaging (MRI) data [8]. The pursuit of neuroanatomical biomarkers has a great potential to facilitate new discriminative methods that are etiologically informed and validated by neuropsychological theories. However, due to high cost of neuroimaging data acquisition, most current ADHD studies are based on relatively small sample sizes, which reduce the statistical power needed to validate meaningful discriminative variable from a very large number of features

extracted from structural MRI [9]. A limited sample size with equivalent number of features raises new challenges to traditional machine learning algorithms, such as logistic regression or support vector machines (SVM), as they tend to overfit and lack a generalization power when training on a dataset containing the number of features far larger than the sample size ($p \gg n$ problem). In previous work, some models either use hundreds of features as an input or exhaustively search on a preselected smaller subset of features. SVM is mostly favored [10] and some variant of feed-forward neural networks [11] is also used. We believe that those methods are either susceptible to overfitting or too restrictive in the search space. The interpretation of the final models is very difficult to validate by existing neuropsychological theories.

In this study, we propose an integrated feature ranking and selection framework that uses brain cortical thickness, extracted from structural MRI data, as features and constructs a prediction model to identify ADHD subjects versus normal controls. The framework performs a two-step feature selection process based on both information theoretic criteria and regularization concept. To mitigate the inconsistent feature selection issue of regularization, especially the lasso method [71], the framework pre-analyzes all features to rank informative features based on mutual information scores [72]. In feature selection, it extends the lasso method [44] to construct a prediction model by fixing those pre-selected highly informative features when performing regression. Tested on both simulated and real datasets, our framework is shown to effectively preserve highly informative features identified in the feature ranking step and improve the model accuracy while searching in a full feature space and maintaining the sparsity in the feature selection step. With a prediction accuracy of 80.9%, our framework selects two sparse models, each with only 4 or 5 cortical thickness features. Previous neurodevelopmental studies of ADHD also consistently suggest that the features selected in our models have a deeper connection to the neurodevelopmental basis of ADHD, and thus making the models highly interpretable to clinicians. The proposed feature selection and prediction framework is a necessary first step to help clinicians find more features of characterizing ADHD using an objective measure with high discriminative accuracy.

The rest of the paper is organized as follows. In Section 3.3, we introduce the background of ADHD, including the brain cortical thickness and its connection to ADHD. We also review the

current feature ranking and selection algorithms. Section 4.3.4 presents the proposed two-step feature ranking and selection framework, including the model formulations and model validation using simulated datasets. Section 3.5 shows the experimental results of the proposed framework on ADHD characterization using a real MRI neuroimaging dataset. Finally, we conclude the study in Section 4.5.

3.3 Background

3.3.1 Feature Extraction of ADHD

ADHD is considered a neurodevelopmental disorder given the age-related differences in cortical maturation that characterize ADHD. Researchers suggest that the origins of attention can be observed in infants as young as three months when the young infant is able to selectively attend (i.e., recognize and orient towards) to their caregiverâĂŹs face [73]. According to these researchers, attention is composed of differential structures and circuits, called an organ system. Furthermore, as a child matures during preschool and early elementary school years, attention response grows into the ability to self-regulate (i.e., adjust oneâĂŹs emotional state/behavior depending on the demands of the environment) in a changing and dynamic environment. Those higher level attention abilities are often described with the term âĂIJexecutive functionsâĂİ. Such development not only relies on social demand, but also is due to the brain maturation of the prefrontal cortex. In Posner and FanâĂŹs (2008) model, self-regulation leads to the second stage in attention development, the executive network. During the ages of 5 to 9, children with deficits in self-regulation and attention are noticed by teachers and parents, as their behaviors deviate from what would be developmentally appropriate.

Choosing brain cortical thickness as the features in ADHD characterization is not only supported by theory, but also benefits from advances of neuroimaging techniques. Numerous theories have hypothesized the cause of ADHD [74] [75] [76] [77] [78] [79]. Those hypothesis are further supported by neuroimaging research, which provides an accurate way to measure the relationship between behaviors or symptoms and underlying brain morphology and brain functioning. As struc-

tural and functional neuroimaging techniques have improved vastly over the last thirty years, MRI provides excellent spatial resolution, uses no ionizing radiation (unlike computed tomography, CT), and thus can be used in pediatric samples of clinical and non-clinical typically developing controls. Cortical and surfaced-based neuroimaging techniques improve on conventional volumetric analysis by allowing for a direct measure of cortical thickness in millimeters, thus may present a more sensitive tool for understanding and measuring brain abnormalities in children with ADHD. So far, a large number of neuroimaging studies have observed that ADHD manifests via a general deficit in the dopaminergic system of the brain including prefrontal cortex [74][8] or abnormalities in brain structures rich in dopamine receptors in children and adults with ADHD [80] [81] [82] [83] [84] [8]. .

3.3.2 Feature Selection

Although recent advances in neuroimaging studies have enabled us to search for structural brain abnormalities caused by the disease that can potentially be used as new biomarkers of ADHD, characterization using traditional machine learning techniques can be difficult because structural characteristics of neuroimaging data, especially MRI data, usually result in large number of features. Even grouping raw features into region of interests (ROI), finding discriminative features for ADHD is still not easy due to relative small sample size with a limited number of patients and healthy participants. Learning from limited sample size with equivalent feature size raises significant issues of overfitting and interpretability of the final model. This study is motivated by the challenge and is aimed to develop efficient feature selection approaches that can construct a sparse model with the most clinical meaningful features preserved. In particular, this paper proposes a novel integrated feature ranking and selection framework which combines information theoretic criteria and the least absolute shrinkage and selection operator (Lasso) method into a two-step feature selection process. The current information theory-based and the Lasso-based feature selection approaches will be discussed in the following.

Feature Selection using Mutual Information

Mutual information [72] [85] is a measure of the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence. MI measures how much information a feature contains about the class without making any assumptions about the nature of their underlying relationships. It is formulated as

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right).$$

If the feature is a perfect indicator for the class membership, its MI reaches its maximum value. A basic intuition is that a stronger mutual information implies a greater predictive ability when using the feature. As an information theoretic criteria, MI have been applied in many feature selection problems [86]. To know whether a given candidate feature should be included, one must be able to evaluate the joint mutual information $I(X,Y)$. However, as feature matrix X is generally multi-dimensional with a continuous distribution, the joint mutual information $I(X,Y)$ is thus extremely difficult to reliably estimate. To solve the problem, one can assume each feature is independent of all other features, and rank the features in descending order according to their individual mutual information score $I(X_i,Y)$. The feature selection is simply picking the top K features, where K can be determined by a predefined certain number of features or some stopping criterion. The feature selection criterion based on mutual information score is commonly adopted in literature. It is often referred as Mutual Information Maximisation (MIM) approaches [87]. However, the performance of this approach is known to be suboptimal if features are interdependent, which is a general case in most studies. In addition, it is widely accepted that a useful set of features should not only be individually relevant to class label, but also should not be redundant with respect to each other, that is features should not be highly correlated in the selected subset. To consider both relevancy and redundancy, a number of approaches have been proposed. For example, Battiti [88] proposed the Mutual Information Feature Selection (MIFS) criterion, which introduces an inter-feature correlation term into the MIM criterion. A penalty parameter β is employed to control the tradeoff between relevancy and redundancy. If the penalty parameter β is set to 0, it is equivalent to the MIM criterion. Peng et. al. [89] presented the Maximum-Relevance-Minimum-Redundancy

(MRMR) criterion, which is in principle equivalent to MIFS with the $\beta = 1/(n-1)$, where n is the number of selected features in the current subset. Yang and Moody [90] used Joint Mutual Information (JMI) to focus on increasing complementary information between features. In particular, the mutual information between the class label and a joint random variable $X_k X_j$ is calculated. By pairing a candidate X_k with each previously selected feature. The principle idea is that if the candidate feature is âĂŸcomplementaryâĂŹ with the existing features, it should be included in the feature subset. Fleuret proposed the Conditional Mutual Information Maximization (CMIM) criterion [91], which examines the information between a feature and the class label, conditioned on each current feature. Instead of taking the mean of the redundancy term, CMIM takes the maximum value in the redundancy term and thus penalize more on feature redundancy.

Although mutual-information-based feature selection approaches gained wide popularity in the literature, there are still some significant issues unsolved. First, all these criteria rely on highly restrictive assumptions on the underlying data distributions. In particular, due to the computational difficulties in high-dimensional mutual information estimation, most approaches only consider pairwise and conditional pairwise interactions, and omit the higher-order interactions. Second, most current MI-based approaches perform feature selection sequentially starting from high-ranked features. As a result, by excluding low MI ranking features, such approaches deny the possibility that a set of low-ranked features combined together may generate strong predictive power (e.g. in the famous XOR problem [92]). We have the risk of missing that strong signal by only working on the preselected candidate set [93] [89] [94].

Feature Selection with Regularization

In medical research, due to high cost of data acquisition, researchers often run into the issue of insufficient samples to train and validate developed models. Instead of heuristic selection schemes (such as many MI-based approaches), objective optimization methods have received more attention since they can be conveniently formulated as convex optimization problems with global optimal solutions. A typical objective function consists of an error term and a regularization term. One of the most widely used such feature selection algorithms is the least absolute shrinkage and se-

lection operator (Lasso) [62]), which allows computationally efficient feature selection based on linear dependencies between input features and output values. The Lasso method as a shrinkage and selection method for linear regression gradually receives high recognition and a fast coordinate descent algorithm has been devised to solve the optimization problem. The optimization framework of lasso to minimize the sum of squared errors with a l_1-norm penalty (bound on the sum of the absolute values of the coefficients) is formulated as follows:

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda ||\beta||_1.$$

By penalizing and forcing some variables to be zero, lasso can effectively select a sparse model. However, it sacrifices unbiasedness to reduce the variance of the predicted value [95].

There are still some challenges for application of Lasso method in feature selection. The Lasso result is often subject to the scaling of features. Inappropriate scaling may cause imbalanced penalty on linear coefficients. The true underlining features with high coefficients may be suppressed to have smaller coefficients. As a result, the total explained variance is limited. Instead of rescaling all features, more generally one can employ adaptive Lasso [96] with penalty term $\lambda \sum w_i ||\beta_i||_1$. Even so, effects of strong signal will be diminished due to shrinkage.

3.4 New Integrated Feature Ranking and Selection Model

3.4.1 Model Formulation and Solution

The proposed Integrated Feature Ranking and Selection Framework is performed in two stages: mutual information-based feature ranking and Lasso-based feature selection. In the feature-ranking step, all features are ranked by their MI scores, and a subset of high-ranked features are selected and considered to have the best informative power. Among those features, a redundancy removal step is performed by checking pairwise correlation between the features. For a highly correlated feature pair (higher than a threshold), the feature with lower MI score is considered redundant and removed from the feature subset to prevent multicollinearity. In the feature selection step, we set the best informative features penalty-free in the generalized lasso method. We use Lasso to select

additional features from the full feature space, not restricted to the subset of high MI features. The additional features selected, although have lower MI scores individually, can improve model classification performance when combined together. Within the subset of high MI features, we start with setting the single top-ranked feature penalty-free, then all combinations of two top features, then all combinations of three top features, iteratively. The feature selection and classification model was validated by leave-one-out-cross-validation (LOOCV). The search process stops when validation accuracy cannot be further improved. The resulting model will be the best model for class prediction. Comparing with other MI-ranking based methods, the proposed framework can select from the full feature space while still creating a sparse model. Comparing with standard regression approaches with regularization, the proposed framework integrates the information theoretic criteria in the generalized Lasso model, and sets the most informative features penalty-free to improve prediction accuracy and enhance model interpretability. The flowchart of the proposed integrated feature ranking and selection framework is shown in Fig. 3.1.

Mathematically, our framework can be formulated as an optimization problem. Let M be the set of indexes of top MI features selected from the MI ranking step. We set indexes in S penalty-free, where S is a subset of M. For each S, we want to solve the following problem.

$$\min_{\beta} \quad \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{3.1}$$

$$\text{s.t.} \quad \beta_{j\in S} = 0. \tag{3.2}$$

The optimization model in our Integrated Feature Ranking and Selection Framework can be solved under generalized lasso framework [97], which is more flexible than lasso and is better in representing the intention to set coefficients of certain informative features penalty-free. Basically, it introduces an arbitrary matrix $D \in \mathbb{R}^{m\times p}, m \leq p$ to define the weights and relations of each element in β.

$$\min_{\beta\in\mathbb{R}^p} \qquad ||y - X\beta||_2^2 + \lambda||D\beta||_1$$

We can construct a proper D in the generalized lasso framework to adjust penalty levels for each feature. To find such a D, we propose and prove the following two propositions.

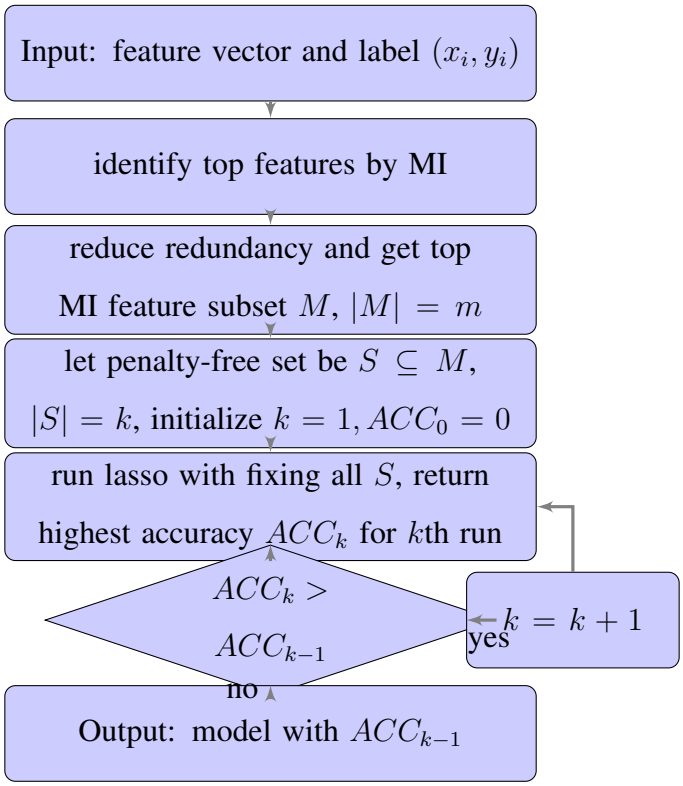


Figure 3.1: Flowchart of Integrated Feature Ranking and Selection Model

Proposition 4

$$\min_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 + ||(\lambda_1\beta_1, \lambda_2\beta_2, \ldots, \lambda_p\beta_p)||_1$$

The above problem of assigning weights λ_k for each feature is equivalent to the generalized lasso with diagonal matrix D and $\lambda_k = d_k\lambda$. (The above formula has also been previously presented as adaptive lasso [96].)

Proof Let D be diagonal matrix diag$(d_1, d_2, \ldots, d_p)$, we have

$$\lambda||D\beta||_1 = \lambda||(d_1\beta_1, d_2\beta_2, \ldots, d_p\beta_p)^T||_1 = ||(\lambda_1\beta_1, \ldots, \lambda_p\beta_p)||_1.$$

If D is $p \times p$ and invertible, β can be transformed into $\theta = D\beta$. The generalized form can be reduced to the standard lasso:

$$\min_{\theta \in \mathbb{R}^p} \quad ||y - XD^{-1}\theta||_2^2 + \lambda||\theta||_1.$$

Proposition 5 *Without loss of generality, to keeping features $X_{p-k+1}, X_{p-k+2}, \ldots, X_p$ penalty-free is equivalent to setting $d_{p-k+1} = 0, d_{p-k+2} = 0, \ldots, d_p = 0$.*

Proof In this case, D is a rank-deficient matrix

$$\text{diag}(d_1, d_2, \ldots, d_{p-k}, 0, \ldots, 0).$$

$$\lambda||D\beta||_1 = ||(\lambda_1\beta_1, \lambda_2\beta_2, \ldots, \lambda_{p-k}\beta_{p-k})||_1.$$

Following the construction procedures in [97], we can transform and reduce the problem to a standard lasso problem. First, we create a full rank matrix $\tilde{D}$ by removing the last k rows from D and adding $k \times p$ matrix A to the bottom, where $m = p - k < p$.

$$\tilde{D} = \begin{bmatrix} d_1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & & \\ 0 & \cdots & d_m & & & \vdots \\ 0 & & & 1 & \cdots & 0 \\ \vdots & & & \vdots & \ddots & \vdots \\ 0 & & \cdots & 0 & \cdots & 1 \end{bmatrix}_{p \times p}$$

In the above matrix $\tilde{D}$, A's rows are clearly orthogonal to those in D. Let $\theta = \tilde{D}\beta = (\theta_a, \beta_b)^T$, where θ_a is related to the coefficient vector β_a of the first m features that are not in the desired set. Now the objective function is

$$\min_{\theta \in \mathbb{R}^p} ||y - X_a\theta_a - X_b\beta_b||_2^2 + \lambda||\theta_a||_1,$$

where X_a is the rescaled first m columns of X, X_b is the original last k columns.

We optimize β_b, θ_a in a sequential way. First, fixing θ_a, the problem regarding β_b is a standard linear regression. The new objective function is to

$$\min_{\theta_a \in \mathbb{R}^m} ||(1-P)y - (1-P)X_a\theta_a||_2^2 + \lambda||\theta_a||_1,$$

where $P = X_b(X_b^T X_b)^{-1}X_b^T$. We get a standard lasso problem regarding θ_a. After solving θ_a, we can in turn determine β_b by $\hat{\beta}_b = (X_b^T X_b)^{-1}X_b^T(y - X_a\hat{\theta}_a)$ from the result of linear regression. The solution of the original generalized lasso solution is $\hat{\beta} = \tilde{D}^{-1}\hat{\theta} = \tilde{D}^{-1}\left[\hat{\theta}_a, \hat{\beta}_b\right]^T$.

Despite the formulation similarity between our model and adaptive lasso [96], adaptive lasso was previously proposed to include a data-dependent weight vector w. The weight vector is estimated as $\hat{w} = 1/|\hat{\beta}|^\gamma$ and no element is intended to be zero. From the formulation perspective, adaptive lasso is a special case of generalized lasso with a full-rank diagonal matrix. In our case, we construct D as a (0,1)-matrix that has exact one non-zero element in each row (i.e. $\sum_j d_{ij} = 1$) and at most one non-zero element in each column (i.e. $\sum_i d_{ij} \leq 1$). The column indices of non-zero elements are the features subject to l_1 penalty. The complement set of $p - m$ features are those, we believe, that are information rich and thus set penalty-free.

3.4.2 Performance Evaluation Using Simulated Dataset

To evaluate the performance of the proposed feature selection framework, we used a simulated dataset with binary response and contain $p = 45$ predictors and $n = 50$ samples. The dataset was generated in such a way that only two predictors were related to the response. Using LOOCV,

Table 3.1: Performance Comparison on Simulated Dataset

Method	**Validation Accuracy**	**Training Accuracy**	**# features Selected**
Our Model	0.92	0.94	5
LR+lasso	0.86	0.97	8

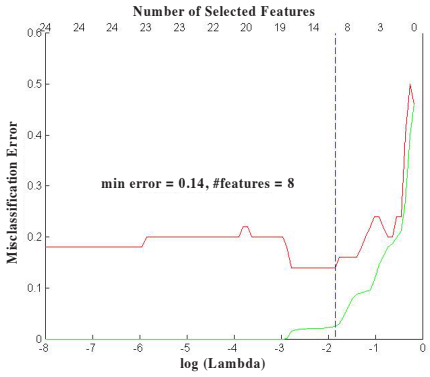


Figure 3.2: Best prediction error using LR+lasso (green curve as training error, red curve as testing error, dashed line cuts at min testing error.)

the proposed framework achieved a validation accuracy of 0.92 with five features selected. As a comparison, we also tested the logistic regression (LR) with lasso, which generated a validation accuracy of 0.86 with 8 features selected. The detailed comparison results are summarized in Table 3.1 as well as Figs. 3.1 and 3.2. From those results, one can see clearly that the proposed framework is capable of selecting a model with higher validation accuracy while with less selected features compared to lasso.

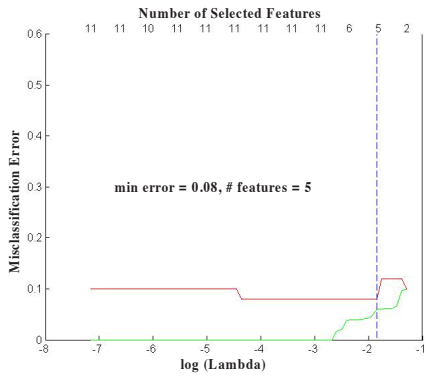


Figure 3.3: Best Prediction Error Using Our Framework (green curve as training error, red curve as testing error, dashed line cuts at min testing error.)

3.5 Application in the Diagnosis of ADHD

3.5.1 Dataset

This study used a dataset that was collected as part of a larger study from the University of Texas at Austin and the University of Texas Health Science Center in San Antonio by Dr. Margaret Semrud-Clikeman.

A total of 47 subjects matched on gender, SES, and ethnicity participated in the study. All subjects were right-handed. There were two groups: 32 ADHD-Combined participants and 15 healthy subjects in a control group. All ADHD participants had less than 15 standard score point differences between general conceptual ability (DAS-GCA) and all achievement measures. The ADHD subjects were matched on severity of symptoms as measured by ConnersâĂŹ Ratings Scale (Conners, 1998a). All ADHD subjects met DSM IV-TR criterion for ADHD CombinedâĂŞtype and no other psychiatric or psychological disorder including Learning Disorders, Anxiety Disorders, Mood Disorder, or Oppositional Defiant Disorder. Control participants did not meet any criteria for a psychiatric or learning diagnosis nor have a history of medication treatment. All participants were recruited from a diversity of socioeconomic and ethnic backgrounds in order to control for potential group differences.

MRI images are acquired at the University of Texas Health Science Center at San Antonio using three-dimensional gradient recalled acquisitions in the study state (3D-GRASS) with a repetition time (TR) = 33 msec, echo time (TE) = 12 msec, and a flip angle of 60 degrees to obtain a 256 $\times$ 192 $\times$ 192 volume of data with a spatial resolution of 1mm $\times$ 1mm $\times$ 1mm. Then all MRI images were processed and normalized using the FreeSurfer image analysis suite [98] [99] by Dr. Jesse Bledsoe on a Linux platform at MSU. All regions of interest (ROI) in the FreeSurfer suite (45 cortical ROIs) were developed using an automated labeling system based on gyral regions of the Desikan-Killiany Atlas [100]. We employed the brain cortical thickness of those ROIs as possible features for ADHD feature characterization and selection in this study.

3.5.2 Results of Feature Ranking and Decorrelation Step

The first step in our framework is to perform feature ranking using mutual information. The top ten features of cortical thickness with highest MI were picked first for further analysis. They are right rostral anterior cingulate (MI=0.124), total rostral anterior cingulate (0.122), left rostral anterior cingulate (0.078), left caudal middle frontal (0.071), right frontal pole (0.068), right lateral orbito frontal (0.063), left caudal anterior cingulate (0.062), total caudal middle frontal (0.051), left inferior parietal (0.051), and left pars orbitalis (0.05). In the next step, we calculated the correlation between each pair of the high-ranked 10 features. If the correlation of a pair of features is 0.6 or higher, we consider one feature in the pair to be redundant, and remove the feature with a lower mutual information value. In this way, the following two features were removed: total rostral anterior cingulate and total caudal middle frontal. The remaining eight features were used as feature candidates in the Lasso-based feature selection step.

3.5.3 Results of Feature Selection Step

Comparison of Testing Accuracy

In feature selection step, within top eight highest MI and uncorrelated feature set, we started with fixing the single top feature penalty-free, then all combinations of two features penalty-free, then

all combinations of three features, iteratively. We evaluated the selection and prediction model using the validation accuracies in a LOOCV procedure. The model search process stops at fixing four features penalty-free, as when fixing more features, the validation accuracy started to decrease. The resulting model is the best prediction model with the highest LOOCV validation accuracy. As shown in Table 3.2, the proposed framework achieved a testing accuracy of 0.81 with a sensitivity of 0.81 and a specificity of 0.80.

In addition, we also tested and compared the performance of the state-of-the-art feature selection algorithms, including the aforementioned information theoretic methods MRMR [89], MIFS [88], JMI [90], CMIM [91], MIM [87], as well as the popular Pudil's floating search method [101], and the principle component analysis (PCA)-based approach, for which we took the components that account for 95% of data variance as the selected features in prediction. The prediction results of these approaches are also summarized in Table 3.2. One can observe that the proposed method achieved higher validation accuracy (0.81) than all other compared feature selection approaches, while using the lowest number of features in the final prediction model. These experimental results confirmed that our model is efficient to select the most predictive features of ADHD given a small sample size.

Analysis of Features in Best Models

To investigate the model interpretability, we also checked the locations of the selected cortical thickness. All the features (regions of interest) selected by the best models were located in prefrontal cortex (PFC), anterior cingulate cortex, and parietal cortex. Structural and functional impairments are in accordance with current understanding of brain-behavior relationships in ADHD.

The prefrontal cortex (PFC) is connected with nearly every cortical structure of the central nervous system [102] and is involved in nearly all aspects of human personality and cognition. The PFC has received much attention in the ADHD literature given a large body of research on impairments in tests thought to tap PFC functioning [103] [104]. For example, the PFC has been implicated in complex behavior relevant to central impairments in ADHD such as inhibitory control [105, 106],attention, working memory, and planning [102, 107]. Furthermore, specific differences

within the frontal pole and orbital frontal cortex observed here may provide further evidence for impairments in frontal limbic structures and emotional disorders which often co-occur in children with ADHD [108].

The anterior cingulate cortex is a key structure implicated in attentional control [107]. It is implicated in a wide variety of cognitive operations including response inhibition, reward processing, behavioral motivation, target detection, and decision making [109]. Functional neuroimaging studies suggest hypoactivation of areas of the anterior cingulate in children and adults with ADHD [110] [111] [112]. Studies observed decreased activation of the anterior cingulate in tasks thought to require behavioral inhibition (e.g., counting Stroop task) in children with ADHD compared to controls [110, 112] also reported reduced activation of the anterior cingulate during tasks of behavioral inhibition (e.g., stop signal task) in children with ADHD-C. Further, cortical thinning of the anterior cingulate cortex has been demonstrated in adults with ADHD [113]. Moreover, the right rostral anterior cingulate cortex (ACC) contributed the most predictive variance in classifying those with ADHD from typically developing controls. This finding supports the hypothesis that abnormal development of the the right ACC, in particular, may be considered a biomarker for ADHD and inhibitory control [114]. The ACC is likely implicated in ADHD due to its involvement in complex behavior. However, the ACC itself, is unlikely to contribute to impaired attention. Rather, future work will need to address the complex networks and systems that involve the ACC in order to provide valid causal pathways for ADHD.

The left inferior parietal cortex also contributed to the classification of ADHD versus healthy children. This was a particularly interesting finding given recent work that has implicated abnormalities in parietal cortex during resting-state functional MRI [115]. Prior to this work, the posterior cortex was proposed to underlie the basis for arousal and vigilance which were considered precursors for targeted attention [107] [116]. And, more recent work has found the posterior parietal lobe to be important for shifting attention during dynamic attention tasks [117]. Structurally, reduced cortical thinning of the right pareital cortex has also been observed in adults with ADHD [113]. Taken together, the parietal cortex, likely due to its frontal projections, is another important area in the attention network that may undergo abnormal development in those with ADHD.

The prefrontal cortex, anterior cingulate cortex, and parietal cortex have all been implicated in attentional control and ADHD. Given these regions provided the best classification of ADHD from controls, the proposed model would appear to be theoretically valid. A significant advantage of the proposed approach is that we novelly integrate the information theoretic feature selection framework with the generalized lasso framework. Through adaptively manipulating penalty weights of each feature in regularization term, we are able to preserve the most informative features in the final model and eliminate less-informative and redundant features.

3.6 Conclusion

ADHD feature characterization and selection has never been an easy task. In this paper, the proposed integrated feature ranking and selection framework provides a sparse, accurate, and highly interpretable model to assist ADHD feature characterization. With the proposed two-step formulation, one can integrate information theory conveniently to supervise the feature selection process while the optimal solutions can be guaranteed due to the convex optimization formulations in a generalized lasso framework. The information-guided selection structure enforces the most useful discriminative predictors to be included in the final prediction model while eliminating less-informative and redundant variables to create an accurate sparse prediction model. In addition to mutual information, due to the flexible structure of the proposed framework, one can also conveniently integrate clinical prior knowledge into the feature selection model. For example, one can set clinician-identified potentially important features penalty-free and encourage them to be included in the final prediction model. The information theory-guided and clinical prior knowledge-guided feature selection framework will be greatly useful to construct prediction models that are more transparent and interpretable by medical and healthcare professionals. Such a supervised feature selection framework is highly demanded in making clinical decisions compared to the 'black box' predictive models generated by traditional machine learning algorithms. As this is a general feature selection approach, the proposed technique can also be applied to other decision-making problems that require interpretable prediction models. The research in this study also suggest that machine learning techniques can be useful tools for understanding and measuring brain abnormalities asso-

ciated with ADHD.

3.7 Epilepsy Diagnosis

3.7.1 Background

In this study, EEG was recorded from a 128-channel electrode array using a geodesic sensor net and Electrical Geodesics, Inc. (EGI; Eugene, OR) amplifier system with signal amplified at a gain of 1000 and bandpass filtered between 0.1 Hz and 100 Hz. During recording, EEG was referenced to the vertex electrode and digitized continuously at 500 Hz.

Visually evoked potentials (VEPs) are electrical potentials (usually EEG) recorded in presence of visual stimuli, and are distinct from spontaneous EEG potentials recorded without stimulation. In particular, the steady-state visually evoked potentials (SSVEPs) have been widely investigated in the past 40 years and have been shown to be useful to analyze many brain cognitive paradigms (visual attention, binocular rivalry, working memory, and brain rhythms) and clinical neuroscience (epilepsy, aging, schizophrenia, migraine, autism, depression, anxiety, and stress). SSVEPs are evoked responses induced by long stimulus trains with flickering visual stimuli. The steady-state potentials are periodic with a stationary distinct spectrum showing stable characteristic SSVEPs peaks over a long time period. It has been found that photosensitivity is found to be common in patients with epilepsy, and visual stimulation may engage the mechanism underlying hyperexcitability in the patients. A series of experiments by Wilkins et al. indicated that spatial properties of visual patterns can elicit epileptiform EEG abnormalities [118]. The epileptic response was reported to be sensitive to luminance, with higher luminance inducing a higher risk of epilepsy [119]. People with migraine or epilepsy are especially prone to symptoms of visual perceptual distortions and visual stress on viewing flicking striped patterns. In a recent study, Birca et al. showed that SSVEP harmonics in the gamma range (50-100 Hz) have significantly stronger amplitudes and greater phase alignment for patients with febrile seizures. In children with focal epilepsy, a similar effect in the gamma range was shown by [120]. As patient with epilepsy are prone to exhibit abnormal EEG responses to repetitive modulated flicking patterns, the resulting SSVEPs can

be employed to discriminate epileptic and non-epileptic patients in a short EEG test rather than a long-term EEG monitoring often around or longer than 24 h. The experimental design of this study is based on this observation. We make an attempt to test the hypothesis that epileptic and non-epileptic EEG recordings during steady state visual stimulation can be classified.

3.7.2 Experimental Design

Eleven patients with epilepsy and eleven healthy subjects were recruited in this experiment. The 11 patients had been diagnosed with idiopathic generalized epilepsy (IGE) at University of Washington (UW) Medicine Regional Epilepsy Center at Harborview. The patients with history of photic-induced seizure or photoparoxysmal responses (PPR) were exclude in order to minimize the risk of inducing seizures during the experiment. The 11 healthy subjects were selected from those who did not have a history of neurological or psychiatric diagnoses such as migraine or schizophrenia. All the patients and normal subjects had normal or corrected-to-normal visual acuity. Each subject underwent the same experimental protocol during EEG recording. Visual stimuli were consisted of a high contrast strip pattern presented on a 19-inch LaCie Electron Blue IV monitor at a resolution of 800×600 pixels, with a 72H vertical refresh rate and a mean luminance of $34cd/m2$. The strip contrast pattern flickering (condition 1) or switching (condition 2) at 7.5 Hz and the contrast level were temporally modulated by 10 levels from lowest contrast (level 1) to highest contrast (level 10) periodically. Each contrast level lasted for 1.067 s with 16 reversals of the flicker pattern. Thus, each stimulus of 10 contrast levels was 10.67 s. Each subject performed 20 trials for condition 1 and 20 trials for condition 2 with brief breaks between trials. A typical session of each subject is about 10âĂŞ15 min.

3.7.3 Signal Processing and Feature Extraction

The visual stimulation flicking at a constant frequency can evoke harmonic oscillations and the SSVEPs were found to have the same fundamental frequency (rst harmonic) as the visual stimulating frequency [121]. A recent study showed that the higher SSVEP harmonics can also play an

important role in studying brain functions [122]. In this study, we extracted frequency features of SSVEPs by Discrete Fourier Transform (DFT) with a 0.5 Hz resolution for each EEG channel of each trial with a time length of 1.067 s. The frequency components obtained from DFT are subject to signal variations. If signal strengths are different, the DFT coefficients are also different even the two time series signals share similar wave patterns. EEG signal is known to have significant inter-individual variability [123], and the signal amplitudes can vary considerable from one person to another. Thus, the extracted DFT frequency components can be problematic in feature selection and model construction across subjects. To tackle this problem, we introduced an normalization step based on ParsevalâĂŹs Theorem. Parsevals Theorem states that the power spectrum summed over all frequencies is equal to the variance of the signal. Based on this rule, we take standard deviation of a signal as a normalization factor and normalize the signal to unit variance before applying DFT.

From the normalized DFT frequency components, the components at stimulation frequency (7.5 Hz) and multiple of stimulation frequency (up to 9th harmonics) were selected as signal features. Then a segment of EEG signal is represented by nine features that include nine harmonic frequency components that may be informative. The feature extraction was applied to each EEG channel of each trial for each subject. For each subject, the features from trials with the same contrast level were averaged to be the features of the contrast level. In summary, there are 128 (channel) $\times$ 10 (contrast level) $\times$ 9 (frequency component) $=$ 11520 features for each subject. In the next, we will present a new feature selection approach to select the most informative features to discriminate epileptic patients from normal subjects.

3.7.4 Assessment and Validation

The feature subset assessment was based on leave-one-out cross-validation procedure. In order to reduce the bias of training and testing data, cross validation techniques have been extensively to assess a classification model. In this study, we employed a leave-onepatient- out cross-validation methodology in order to avoid the potential bias of having EEG samples from the same patients in both the training and testing data. We measured model classification accuracy by the average

of sensitivity and specificity. Sensitivity and specificity are widely used in the medical domain as classification performance measures.We labeled the EEG samples from epileptic patients as positive and those from non-epileptic patients as negative. The sensitivity measures the fraction of positive cases that are classified as positive; the specificity measures the fraction of negative cases classified as negative.

3.7.5 Computational Results

We performed our feature selection and classification approach for each of the 10 contrast level and each of the 9 harmonic frequencies independently. This experimental setup is specially designed to find out which contrast and which harmonic frequency are most prominent to discriminate epileptic patients from normal subjects. In the feature selection step, we selected the top ten highest MI feature set first, and performed Lasso to select additional features from the remaining features with relative-low MI values. Once we finalize the feature candidates (lasso-selected low-MI features and top 10 high-MI features), we enumerate feature subset starting from one feature. The feature combination with the highest cross-validation classification accuracy was selected as the as the optimal feature subset. The classification accuracies for each contrast level and harmonic frequency are shown in Table 1. We notice that the contrast level 7 and the 5th harmonic frequency generated the best validation accuracy of 90%. There were six selected channels: 53, 54, 56, 75, 114, 119. Using prior knowledge guided feature selection have very good interpretability to physicians and neurologist. We also compared three popular feature selection approaches, regular Lasso feature selection [124], stepwise feature selection using statistical significance test [125], PudilâĂŹs floating search [126]. Table 2 shows the classification performance comparisons of our method with the three popular feature selection methods. The feature subset picked up by our approach generated the highest cross-validation accuracy of 90%, followed by the PudilâĂŹs floating search with an accuracy of 85%. Both regular Lasso and stepwise selection got the validation accuracy of 80%. Also for the overall performance cross the 10 contrast levels and 6 harmonic

3.7.6 Conclusion and Discussion

A quick and accurate epilepsy-screening tool could enormously reduce associated healthcare costs and improve the current diagnosis procedure. To reliably recognize if a patient has epilepsy, we developed a novel mutual-information-guided sparse feature selection and classification framework to identify epilepsy-specific patterns from visually-evoked potentials in a human-computer task. The experimental results confirmed that the proposed method achieved the best diagnostic accuracy compared with several popular methods. The proposed method has a potential to help physicians to determine whether a patient is epileptic or non-epileptic in a quick screening process. More importantly, the proposed information-theory-guided sparse feature selection is an generally framework. It is also promising to help physicians and neurologists in recognizing abnormal brainwave patterns in huge medical dataset with different brain imaging techniques (such as EEG, MEG, and fMRI). The long-term goal of this study is to develop a fast, reliable, and affordable epilepsy diagnostic system using shortterm interictal EEG signals. Such a system can revolutionize the current epilepsy diagnosis practice with wide and convenient applications.

Table 3.2: Comparison of Testing Results (Leave One Out Cross Validation)

# Selected Features	**Testing Accuracy**	**Training Accuracy**	**Sensitivity**	**Specificity**	**Selection Method**
4	0.81	0.87	0.81	0.80	Proposed Method
5	0.76	0.78	0.75	0.80	MRMR (Peng et al, 2005)
7	0.66	0.76	0.66	0.67	Pudil's Floating Search (P. Pudil 1994)
14	0.70	0.74	0.72	0.67	PCA
5	0.74	0.75	0.81	0.60	MIM (Brown, 2009, Kwak & Choi, 2002, Lin & Tang 2006)
5	0.70	0.76	0.69	0.73	MIFS (Battiti, 1994)
5	0.72	0.78	0.72	0.73	JMI (Yang & Moody, 1999)
5	0.74	0.76	0.75	0.73	CMIM (Fleuret, 2004).

Chapter 4

A PATIENT-SPECIFIC MODEL FOR PREDICTING TIBIA SOFT TISSUE INSERTIONS FROM BONY OUTLINES USING A SPATIAL STRUCTURE SUPERVISED LEARNING FRAMEWORK

4.1 Abstract

Recreating the natural anatomy in ligament reconstruction is crucial to fully restore the knee joint function and reduce impingement on iatrogenic injury to adjacent structures, yet is subject to the difficulties in locating ligament and other associated soft tissues insertion sites intra-operatively and the high inter-person morphological variability cross patients. In this study we present a new quantitative analysis method capable of achieving personalized identification of cruciate ligament and soft tissue insertions. We craft patient-specific features of tibia outline that can be accurately and reliably measured from CT images. In addition, we propose a supervised structure learning and prediction model with special inter-dimensional and response structure regularization terms to capture relationship between the spatial arrangement of soft tissue insertions and the patient-specific features extracted from the tibia outlines. In experiment, the proposed model outperforms baseline models and provides an accurate and accessible approach that can be used as the first and the most critical step to achieve personalized surgical planning in cruciate ligament reconstruction.

4.2 Introduction

The knee is a complex joint that supports relatively large loads and great mobility, making it vulnerable to a variety of injuries. The anterior and posteior cruciate ligaments (ACL and PCL) are two primary stabilizers of the human knee. The ACL is commonly injured: an estimated 175,000 ACL reconstructions are performed annually, with a financial impact exceeding 2 billion dollars, in the United States alone. PCL injuries occur less frequently are believed to be under- diagnosed;

yet they affect about 3% of the general population and account for as many as 40% of patients with knee trauma seen in emergency rooms. .

The anatomic reconstruction procedure involves creating the bone tunnels and placing the substituting grafts in the exact anatomical positions as the native ligaments. An accurate replication of the natural anatomy in anatomic reconstruction is crucial to fully restore knee joint function and reduce impingement on or iatrogenic injury to adjacent structures [127, 128, 129, 130, 131]. However, current approaches to treating knee injuries are not quite good in terms of consistency and effectiveness in restoring knee function and preventing the development of osteoarthritis (OA). Analyses of long-term outcome after ACL reconstruction have revealed that only in 37% of the patients was normal restored in terms of knee structure and function [132], and 90% of the ACL-reconstructed knees exhibited radiographic evidence of OA 3-10 years after injury [133]. A growing body of evidence is suggesting that anatomic reconstruction, performed by creating the bone tunnels and placing the substituting graft at the native ligament insertion site, can better restore the joint function and deter the development of OA. A number of challenges are present in practical anatomic reconstruction of cruciate ligaments.

Intra-operative identification of the native cruciate ligament insertion sites, as a requisite for anatomical reconstruction, poses a tremendous challenge. Not all surgeons can maintain an acute awareness of the anatomy: about 85% of ACL reconstructions are done by surgeons who perform fewer than 10 cases per year [18] and PCL reconstructions are even less frequently performed by most surgeons; for those who can, factors including the arthroscopic distortion and disappearance of the ligament remnant (naturally or due to a notchplasty procedure) can still cause misidentification of the natural insertion or attachment sites. There is considerable variability of knee anatomy in terms of bone and soft tissue insertion morphology (position, size, and shape) [19]. Sample data from our preliminary study of tibial insertion site morphometry suggest that simplistic cross-referencing or generalization from one patient to another is likely to lead to non-anatomical tunnel drilling and iatrogenic injury to adjacent tissue structures. Although it may be difficult to gauge the incidence and impact of these iatrogenic injuries as complications of ACL or PCL surgeries, the importance of minimizing the risk of such injuries is readily recognized [20, 21].

The key to anatomic cruciate ligament surgery with minimized risk of iatrogenic injury is an accurate, quantitative knowledge base of the tissue morphology, documenting inter-person variability and specificity vs. uncertainty associated with alternative ways to predict morphometrics. Studies have investigated the quantification of the insertion sites of the cruciate ligaments and other soft tissue components using statistical and quantitative approaches [22, 23, 1]. However, such quantitative analysis and measures generally cannot fully capture the accurate spatial arrangement of soft tissue insertions. The location and morphological measures cannot account for the inter-person variability of cruciate ligament and meniscus insertions, which are mostly characterized by qualitative measures [24, 25]. Advanced imaging techniques, such as 3D CT or MRI, which can be useful to visualize the major structure outline of the knee of a patient clearly, cannot be applied directly to determine the insertion sites for cruciate ligaments reconstruction. This is because the imaging shows the structure of a knee with serious cruciate ligaments injuries and structure misalignment. While in the surgery, it is crucial to identify the native location of the cruciate ligaments to reconstruct the natural anatomy of the ligament structure. Therefore, one needs inference on the appropriate insertion sites of native cruciate ligaments. Due to the complex anatomy of the knee, the identification of insertion sites of cruciate ligaments in knee reconstruction surgery is still an unsolved problem.

This study aims to develop a new quantitative analysis method to achieve personalized identification of cruciate ligament and meniscal insertions using patient-specific knee morphological features. In particular, the proposed framework first digitalized outlines of tibia from 3D CT images and aligned the outlines using Generalized Procrustes Analysis (GPA) techniques. It then extracted patient-specific features, trained a supervised structure learning and prediction model, and predicted the centroids of the sites of the cruciate ligament and meniscal insertions using the learned prediction model. The supervised structure learning and prediction model captured the relationship between the spatial arrangement of soft tissue insertions and the patient-specific features extracted from the tibia outlines, which can be easily and reliably measured from CT images. To the best of our knowledge, this is the first supervised machine learning algorithms in knee soft tissue site identification. The proposed learning and prediction framework provides a critical step

to achieve the highly demanded personalized surgical planning in cruciate ligament reconstruction.

The rest of the paper is organized as follows. Section 5.3 presents the data acquisition and knee imaging data processing including the digitalization of tibia and soft tissues from 3D CT image, the image alignment and normalization using GPA, the innovative feature engineering via coordinates transformation, and the structure learning and prediction model for the soft tissue insertion centroids. Section 4.4 presents the prediction performance the proposed framework as well as the performance comparison with the available baseline models. Finally, we conclude this paper in Section 4.5.

4.3 Methods

4.3.1 Data Collection

Twenty tibia specimens (10 left and 10 right unpaired knees; 11 from men and 9 from women; mean age at death: 61 ± 5 years) were used to acquire the morphometric data [134]. All epithelial, subcutaneous, and muscular tissues were removed from the specimens. High-resolution CT scans of the tibias were taken with slice spacing of 0.625 mm and three-dimensional (3D) bone models of the tibias were created in Mimics (Materialise Inc., Belgium). A Polaris Spectra optical tracking system (Northern Digital Inc., Ontario, Canada), with a manufacturer-reported accuracy of ± 0.25 mm, was used to digitize the outlines of the ACL, PCL, the medial cartilage (mcart), the lateral cartilage (lcart), anterior and posterior medial meniscal root (ammr and pmmr), and anterior-lateral and posterior-lateral meniscal root (almr and plmr). The digitization was performed by the same experimenter with the repeatability, as assessed by intraclass correlation coefficients (ICCs), ranging from 0.94 to 0.99. The digitized outlines were mapped onto the CT-based 3D tibia models with a fiducial registration error smaller than 2% [1]. A closed spline was then fitted to each outline, resulting in 100 equidistant discrete points to represent the outline, as shown in Figure 4.1.

A 3D coordinate system was defined on each tibia based on its digitized and mapped cartilage outlines (Fig. 4.1). First, the origin of the coordinate system was determined as the midpoint of the medial and lateral cartilage centroids. The principal components analysis (PCA) was then

performed on the equidistant discrete points representing the cartilage outlines (200 points in total). The X-axis was the first principal component axis passing the origin and pointing laterally. The Y-axis was orthogonal to the X-axis, passing the origin and pointing anteriorly. To make the Z-axis point proximally, the coordinate system was designed as a right-handed system for the right tibia and a left-handed system for the left tibia.

4.3.2 Image Alignment and Normalization Using Generalized Procrustes Analysis (GPA)

Cartilage outlines for all 20 tibias were optimally aligned using GPA, which is an iterative process of applying Procrustes superimposition to all possible pairs of configurations, a configuration here refers to a set of cartilage outline landmark coordinates in a pre-defined order. For each cartilage configuration pair, one configuration served as the base and the other as the target. Procrustes Superimposition matches the target configuration onto the base, centering, rotating and uniformly scaling the target configuration to minimize the shape difference (Figure 4.2). For multiple (20 in this study) configurations, GPA identified the reference or overall base configuration as the one with the smallest overall Procrustes Distance (PD) to all others (i.e., the 19 remaining tibial cartilage configurations). The 19 remaining configurations were then Procrustes-superimposed onto this selected reference and their insertion sites transformed accordingly by the same translation, rotation, and scaling rules, without any shape distortion. Figure 4.3 shows the outlines of tibial cartilage and six insertion sites from 20 subjects before and after cartilage-based GPA.

With the ultimate goal of 3D prediction directly, we first need to investigate and evaluate the feasibility if the internal structure of cruciate ligaments can be possibly learned from the outline of tibia. Thus, we first limit our scope in a 2D plane and investigate if the structure relation can be learned. Conceptually, if certain 3D relationship between tibia outline and locations of cruciate ligaments exist, then the structure should also exhibit in the projected 2D plane. In other words, if the structural prediction performance on a 2D plane is promising, then it will confirm that the knee outline information is useful in prediction of native locations of soft tissues. Most importantly, it will also indicate structural relation between tibia outline and soft tissue locations may also be valid in the 3D space.

4.3.3 Feature Engineering using Coordinates Transformation

The outline of tibia can be easily and reliably measured, making it a feature candidate to predict the locations of intangible soft tissues. From examining the digitalized 3D data, we observed that the length and width dimensions did not change significantly along with the change of the depth dimension. Hence, the depth dimension can be considered a noise dimension to be filtered from digital images. After preprocessing, the digitalized image data in Cartesian coordinates can represent each point in two dimensional space (x, y). In particular, we employed the polar coordinates to reduce feature dimension. The feature dimension for Cartesian coordinates for each 2-D point on the tibia boundary is 2, however, for polar coordinates, since we take points from equal intervals and drop the angle, and the feature dimension becomes 1.

A 2-D point with Cartesian coordinates (x, y) can be transformed into polar coordinates (ρ, θ) following the rule: $x = \rho\cos\theta, y = \rho\sin\theta$. For the shape of tibia, we consider the aligned origin as its center, and transform all points (including all boundary and all inner points) into polar coordinates according to the above rule. To represent the outline of tibia with a limited number of features, we implemented a discretization and boundary detection algorithm. The algorithm first divides a complete cycle into $N = 36$ equal intervals, creating a series of 10^o intervals. Then the algorithm finds the maximal ρ in each interval and records it as ρ_t, where $t = 1, ..., 36$, resulting a functional data series of tibia shape (Fig. 4.4). In this way, each outline of tibia was represented by a 36-dimensional vector $(\rho_1, \rho_2, ..., \rho_{36})$.

Such transformation characterizes the shape of tibia by a functional data series for each individual patient. Such patient-specific tibia features will be used to predict the sites of ligament and other soft tissue insertions.

There is high variability in ligament and meniscus insertions. In a surgical procedure, the centroids of the insertions are of particular importance to tunnel locations. Thus, we focused on the prediction of the centroids of the ligament and meniscus insertions instead of the complete morphology of the insertions.

For each subject $i \in \{1, \ldots, 20\}$, the centroid of an insertion site j, (a_{ij}, b_{ij}) is defined as

$$(\frac{\max(x_{ij}) + \min(x_{ij})}{2}, \frac{\max(y_{ij}) + \min(y_{ij})}{2}) \tag{4.1}$$

.

4.3.4 Structure Supervised Learning Model for Soft Tissue Centroid Prediction

Regardless of inter-person variability existing in morphology, the spatial arrangement of the eight soft tissues basically follow an intrinsic pattern of inter-tissue structure. For example in 2D CT scan, ACL is above PCL, AMMR is always adjacent to ACL, no soft tissue can be out of the tibia outline. To correctly retain such structure when predicting the centroid of each soft tissue, we also need to consider the correlations among soft tissue centroids. Since centroids are the responses in our model, we adopt the following strategy to address such consideration.

In multivariate regression, we learn the relationship between m response variables $\{y^{(1)}, y^{(2)}, \ldots, y^{(m)}\}$ and p predictors $\{x^{(1)}, x^{(2)}, \ldots, x^{(p)}\}$. Each $y^{(j)}$ has its own regression model

$$y^{(j)} = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \ldots + \beta_{pj}x_p + \epsilon_j \tag{4.2}$$

The above problem can be generalized as sparse multi-task learning [135], which minimizes overall errors among m responses instead of the error of each individual response with a proper regularization on matrix β.

$$\min \sum_{j=1}^{m} ||y^{(j)} - X\beta^{(j)}|| + ||\beta|| \tag{4.3}$$

To capture the spatial arrangement of the soft tissues, based on our previous studies on prediction model development for medical problems[136, 137, 138], we modeled the spatial structure in a simplified linear relation as follows.

$$(y^{(1)}, \ldots, y^{(m)})A^{(j)} = 0, j \in \{1, 2, \ldots, m\}, a_{jj} \neq 0. \tag{4.4}$$

It shows that each response variable can be linearly represented by the other seven response variables. The optimization objective should not only consider how well the prediction $\hat{y}^{(j)}$ fit the

response, but also how well the predictions $\{\hat{y}^{(1)}, \hat{y}^{(2)}, \ldots, \hat{y}^{(m)}\}$ fit the structure that the responses would fit.

To predict the centroids of the eight soft tissues within the tibia, we have developed a structure supervised learning model to construct a regression model for the centroids of all eight soft tissues. In the model, the independent variables (features) are the functional data series of tibial shape and the response variables are the distance from soft tissue centroids to the tibia centroid $(0, 0)$.

Let x_i, $i = 1, ..., 20$ be features extracted from the tibia outline, where x_i is a vector of 36 dimensions that represent the tibia functional data series. Let y_i be response variables (soft tissue centroids), each is a vector of 16 dimensions (x-y coordinates off the centroids of all 8 soft tissues.) Let A be a matrix describing linear relations between each components of y_i, β be the matrix indicating linear relation between y_i and x_i. Let α_1, α_2 and α_3 be the tuning parameters. The objective function of our structure-based regression model is given by

$$\min_{A,\beta} \sum_{i=1}^{20} ||y_i - x_i\beta|| + \alpha_1 \sum_{i=1}^{20} ||y_i A|| \\ +\alpha_2 \sum_{i,j} |A_{i,j}| + \alpha_3 \sum_{i,j} |\beta_{i,j}|. \tag{4.5}$$

Where the first term is the sum of Euclidean norm of the error terms based on linear regression between y_i and x_i ("Euclidean norm error term"). The second term is the sum of Euclidean norm of the error terms based on linear regression between different dimensions of y_i ("Inter-dimensional Regularization Term"). The last two terms are l_1 regularization terms, where the third regularizes on the structures on response ("Response Structure Regularization Term"), and the last one is the l_1 regularization term on β ("$\beta - l_1$ term"). With the three regularization components, our framework can predict on response variables with learned spatial structures.

4.3.5 Soft Tissue Centroid Prediction

The trained model balances the fitness between feature vector x and response vector y and the spatial structure embedded in the response vector. With the learned model parameters β, A, the

prediction step is to find a $\hat{y}$ that minimizes the objective function as follows

$$\min_{y} ||y - x\beta|| + \alpha_1 ||yA||. \tag{4.6}$$

Given a feature vector of a new patient, the optimization framework ensures that the predicted centroid locations follows the learned spatial structure and also minimizes the fitting errors.

4.4 Evaluations

In this section, we test the proposed model on the knee data with 20 subjects and also compare the model with several alternative methods. Since the sample size is smaller than the feature size, we try lasso, population mean, k-nearest neighbor approaches.

4.4.1 Performance Measures

The leave-one-out-cross-validation (LOOCV) was employed to obtain an unbiased estimate of the generalization ability of the prediction models. For a dataset with n subjects, the procedure consists of n trials. In each trial, a model is trained from $n - 1$ subjects to predict the remaining one, and calculate the squared error against its actual value. The procedure repeats n times until all subjects have been tested once. We used the mean squared error (MSE) over the 20 subjects as the performance metric to evaluate the accuracy of centroid prediction to compare different approaches.

4.4.2 The Effect of Spatial Structure Learning

First, we examine the proposed formulations on the knee data of 20 subjects. In particular, we compared our proposed framework with the following three variations: 1) Model A: minimize Euclidean norm error term with only l_1 regularization term on β; 2) Model B: structural learning with linear predictor (no optimization); 3) Model C: minimize Euclidean norm error term with only inter-dimensional error term and l_1 regularization term on β, and 4) Model D: the population mean method. The comparison will be based on prediction error for each predictor variables and the total number of variables that outperform population mean method.

The comparison results are summarized in Figure 4.5. One can observe that with only l_1 regularization term on β (Model A), the prediction suffered from high prediction error due to ignorance of variance and structural properties in the response. only Inter-dimensional regularization term and l_1 regularization term on β (Model C), prediction still cannot outperform population mean method due to lack of consideration of structural properties in the response. Structural learning without optimization procedure (Model B) does not predict well either due to ignoring the correlation among the aforementioned terms. Our proposed model, with adding Inter-dimensional regularization term, response structure regularization term, l_1 regularization term on β, and optimization procedure, can outperform population mean method in terms of total residual sum of errors.

4.4.3 Comparison with Baseline Methods

To show the advantages of the proposed method, we also compared the model with several alternative methods, including lasso, population mean, k-nearest neighbor approach described as follows.

Linear Regression with Regularization: Since the number of features is greater than sample size, we apply lasso [62] to each coordinate of eight soft tissues. This approach does not consider the inter-relationship of the positions.

Population Mean: To predict soft tissue centroids using population-based prediction, in our case, we consider our 20 subjects as a target population. When we make prediction, we need to ensure that the predicted soft tissue centroids are consistent with the community's average characteristics.

K nearest neighbor Algorithm with Three Distance Measures: K nearest neighbor [139] can be used both in classification and in regression. In classification, we assign class label to a subject according to the majority vote of its k closest training examples. In regression, we compute the value of a subject by taking average of its k closest training examples. From its k most nearest training samples, we can learn the knowledge of a subject. The k nearest neighbor algorithm provides us with more patient-specific information than population based approach.

In the k-nearest-neighbor algorithm, distances between two subjects can be considered simi-

larity between them. In our analysis, we compare three different distance measures in calculating such similarities.

(1) Euclidean distance measure is the distance on Euclidean space. Intuitively, it is the length of a line segment connecting 2 data points.

(2) The two sample t-test measure. Lower p-value indicates bigger difference.

(3) In time series analysis, Dynamic Time Warping (DTW) [140] measures similarity between two temporal sequences which may vary in time or speed. When we compare two sequences, instead of making point-to-point comparison, we create time windows, and consider the minimal distance in a time window as the distance. We try two difference time windows. Window parameter is the unit of time shift we allow.

Comparison Results with Baseline Models: Table 4.1 summarizes the prediction performance of the eight soft tissue centroids by the proposed model and the baseline models aforementioned. For each soft tissue centroid, the average residual sum of squares of x and y axis over the 20 subjects are reported. The MSE of the proposed model is equal or better than that of population mean based method in 13 out of 16 axes (Table. 4.1), is the best (or tie) in 10 out of 16 axes among all the methods. Table 4.2 summarizes the Euclidean distances between the predicted and actual soft tissue centroids. Again, the proposed method achieved the best prediction performance in six out of the eight soft tissue centroids. These experimental results confirmed that the proposed supervised learning model was effective to learn the spatial structure of the eight soft tissues to improve prediction performance. Figure 4.6 illustrates the centroid predictions on one subject. The proposed method generated predictions closer to the actual centroids in six out of the eight soft tissues.

In addition, we also compared a simple linear model with the proposed model without regularization, and the prediction results are summarized in Table. 4.3. One can observe that the prediction performance of the linear model has a high variance and is among the worst compared

with other baseline models. Also without the structural learning term, the proposed method is reduced to the regular lasso model, and the prediction performance was significantly deteriorated. This observation also indicated that the structural learning term is indeed useful to improve the prediction performance by considering the learned spatial structure of the eight soft tissues.

4.5 Discussion and Conclusions

The motivation of this study is to provide an accurate quantitative approach to aid soft tissue insertion localization using patient-specific measures that can be reliably and accurately acquired from knee imaging data. An extensive quantitative analysis of the location and the inter-relationship of soft tissue insertions on the tibial plateau has been performed, including digitalization of tibia outlines from 3D CT images, the imaging alignment using generalized Procrustes analysis, patient-specific morphological feature extraction from tibia outlines, integration of a spatial-structure learning in a regularized regression-based model training framework, and a patient-specific prediction framework with the learned spatial structure of the soft tissue insertions. In particular, we demonstrated the possibility of using outlines of tibia to predict the centroids of eight soft tissue insertions that are crucial in anatomical reconstruction of cruciate ligaments. The proposed methodology yielded the best prediction accuracies compared with other baseline models for the eight soft tissue locations. The integration of the proposed spatial-structure learning framework has demonstrated to improve the prediction performance significantly.

With the limited data size of 20 subjects, the primary goal of this study was to investigate the feasibility of establishing a quantitative prediction methodology of soft tissue insertions. The proposed structure learning and prediction framework could potentially be strengthened or complemented if more data samples become available. In this study, we only restricted our analysis of tibia outlines and soft tissue centroids in a 2D space using the GPA aligned tibia outlines and a PCA-transformed coordinates. A natural extension of this work is to construct a full 3D prediction model using 3D outlines of tibia. Although the 3D knee models for the subject-specific geometry of bone and soft tissues are widely available from imaging data, such as CT and MRI [141, 142, 143], it is still challenging to obtain accurate location information of the ligament insertion sites and

meniscal root attachments conveniently. To obtain an accurate anatomical structure, specific MRI sequences and configurations may be required for specific tissue structures [144, 145, 146], or one has to acquire data in vitro [144, 147]. Such procedures are generally expensive and often impractical for broad clinical applications. The supervised structure learning and prediction method developed in this study has a potential to provide accurate information of soft tissue insertion sites only using the tibia outlines that can be reliably and easily acquired from imaging data of the knee. The efficient quantitative analysis framework facilitates establishing a clinical applicable tool to assist surgeons with identifying soft tissue insertion sites, based on which a close replication of the native anatomy can be created and the risk of iatrogenic injury to adjacent tissue structures can be minimized. The quantitative modeling and analysis methodology in this work is among the first efforts to facilitate the highly demanded personalized surgical planning and achieve more precise and better-navigated surgeries in anatomical reconstruction of cruciate ligaments.

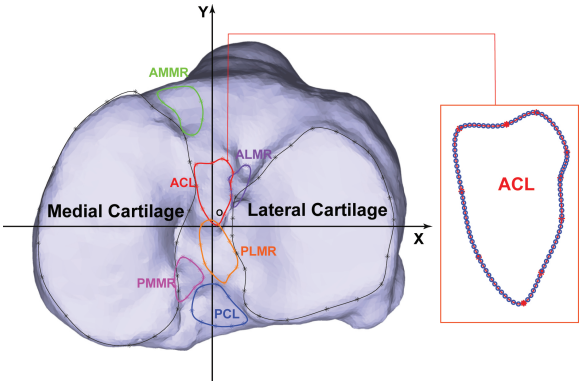


Figure 4.1: The digitized cartilage and insertion site outlines mapped onto the CT-based 3D tibia model. The digitized points (asterisks) were spline-fitted, generating 100 equidistant points (circles on the close-up view of ACL insertion outline) on the fitted outlines to facilitate the subsequent analyses [1].

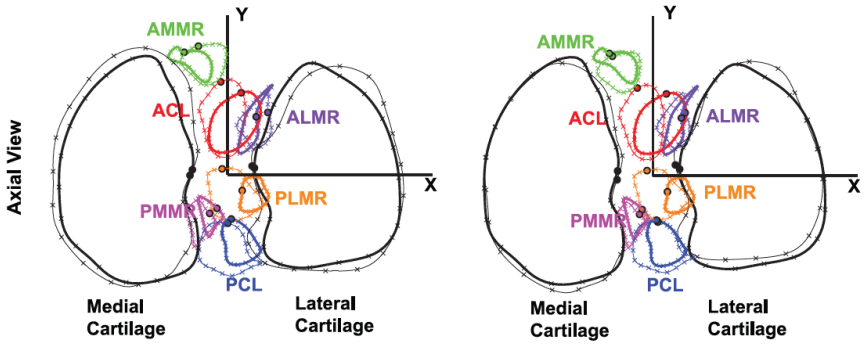


Figure 4.2: The effect of Procrustes Superimposition illustrated by one pair of tibias. One cartilage configuration served as the base (thick) and another as the target (thin). Six tissue structure insertions before (left column) and after (right column) superimposition are also shown in three orthogonal views [1].

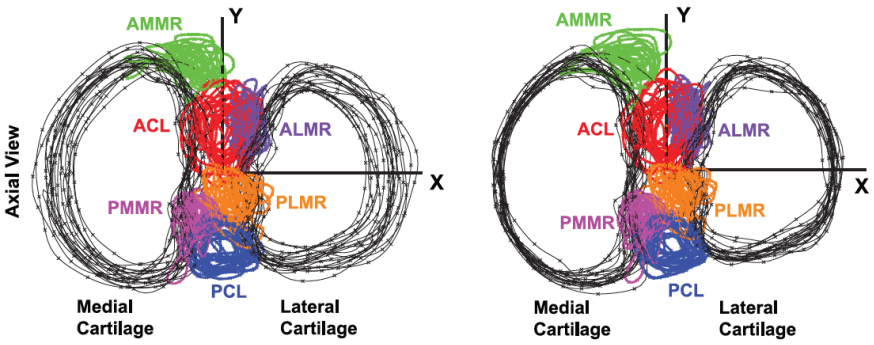


Figure 4.3: The outlines of tibial cartilage and six insertion sites from 20 subjects before (left plot) and after (right plot) cartilage-based Generalized Procrustes Analysis [1].

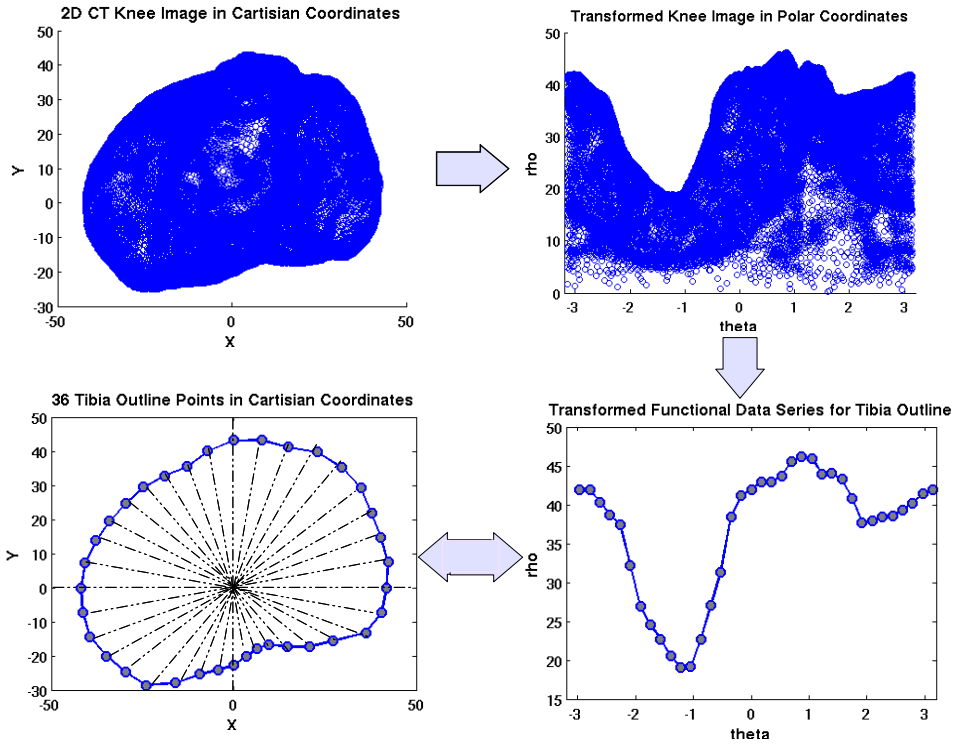


Figure 4.4: The patient specific feature extraction procedure from CT image of the Tibia. First, the tibia outline in Cartesian coordinates is converted into polar coordinates. We divide a complete cycle into 36 equal intervals, and then find the maximal magnitude in each interval to generate a 36-dimensional functional data series as the input predictive variable to train the prediction model of soft tissue insertion sites.

Table 4.1: Knee Soft Tissue Centroid Prediction MSE (mm) (Leave One Out Prediction) in Cartesian Coordinates

Soft Tissue	**Proposed Model**	**lasso Model**	**Population Mean**	**KNN (Euclidean)**	**KNN (t-stat)**	**KNN (DTW=1)**	**KNN (DTW=2)**
lcart (x,y)	(17.49, 21.65)	(28.44, 21.18)	(70.01, 21.73)	(19.64, 21.88)	(34.82, 22.36)	(19.19, 20.98)	(22.09, 21.47)
mcart (x,y)	(22.32, 11.63)	(25.60, 11.63)	(69.50, 11.63)	(23.84, 15.72)	(44.60, 12.90)	(27.63, 13.54)	(32.18, 12.53)
acl (x,y)	(73.60, 99.78)	(77.69, 108.33)	(70.77 101.50)	(51.41, 137.32)	(62.32, 103.00)	(56.72, 120.75)	(57.61, 126.24)
almr (x,y)	(25.58, 76.84)	(34.40, 107.26)	(28.93, 82.60)	(28.00, 98.17)	(28.09, 85.07)	(29.68, 87.77)	(29.98, 93.07)
ammr (x,y)	(224.71, 57.37)	(222.64, 40.95)	(214.33, 33.68)	(355.41, 41.10)	(228.54, 31.43)	(271.12, 31.34)	(261.33, 30.49)
pcl (x,y)	(56.70, 36.39)	(49.15, 49.79)	(58.60, 40.15)	(62.19, 34.91)	(71.78, 36.23)	(67.77, 35.39)	(63.49, 36.68)
plmr (x,y)	(66.20, 64.78)	(78.47, 82.59)	(68.85, 68.54)	(74.36, 67.75)	(65.78, 66.02)	(83.66, 69.17)	(76.27, 69.64)
pmmr (x,y)	(65.51, 88.69)	(68.42, 102.61)	(74.99, 102.85)	(73.18, 110.79)	(78.15, 95.08)	(68.63, 105.83)	(58.84, 106.00)

Table 4.2: Knee Soft Tissue Centroid Prediction MSE (mm)(Leave One Out Prediction) in Polar Coordinates

Soft Tissue	**Proposed Model**	**lasso Model**	**Population Mean**	**KNN (Euclidean)**	**KNN (t-stat)**	**KNN (DTW=1)**	**KNN (DTW=2)**
lcart (r)	39.14	49.62	91.74	41.52	57.18	40.17	43.56
mcart (r)	33.95	37.23	81.13	39.56	57.5	41.17	44.71
acl (r)	173.38	186.02	172.27	188.73	165.32	177.47	183.85
almr (r)	102.42	141.66	111.53	126.17	131.16	117.45	123.05
ammr (r)	282.08	263.59	248.01	396.51	259.97	302.46	291.82
pcl (r)	93.09	98.94	98.75	97.1	108.01	103.16	100.17
plmr (r)	130.98	161.06	137.39	142.11	131.8	152.83	145.91
pmmr (r)	154.2	171.03	177.83	183.97	173.23	174.46	164.84

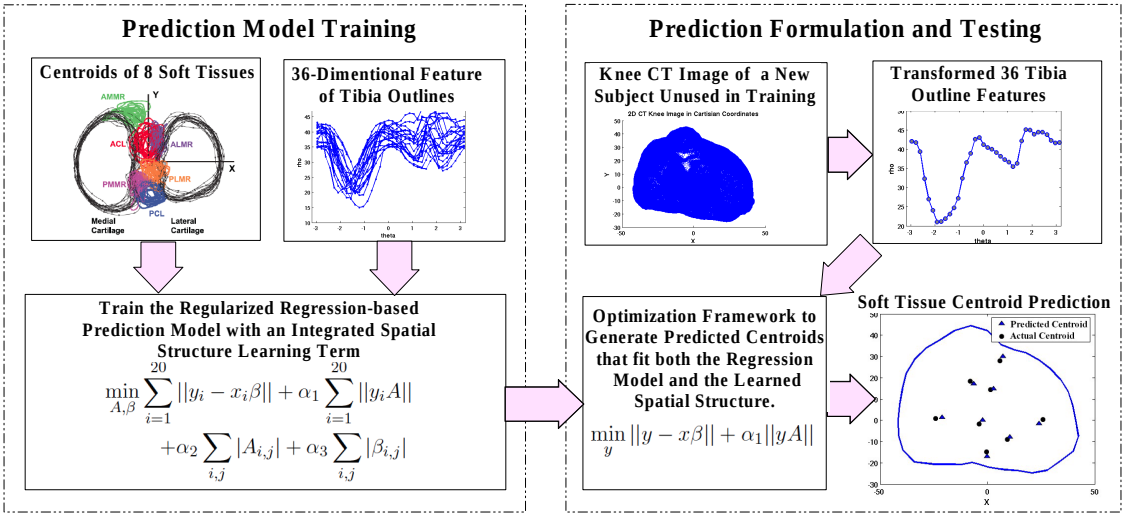


Figure 4.5: The flowchart of the training and testing of the proposed patient-specific prediction model of soft tissue insertion sites using the 36-dimensional features extracted from tibia outlines.

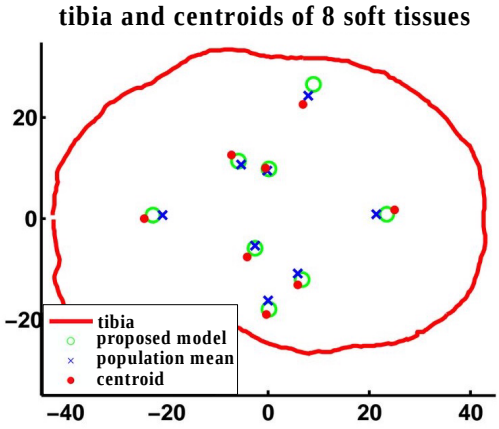


Figure 4.6: The outline of tibia and the centroids of soft tissues are in red. The predicted centroids (green) by our model are in general closer to the actual locations than the predicted centroids (blue) by population mean.

Table 4.3: Knee Soft Tissue Centroid Prediction MSE (mm) (Over-fitting)

Soft Tissue	**Linear Model**	**lasso Model**
lcart (x,y)	(107.84, 102.42)	(51.44, 34.55)
mcart (x,y)	(56.66, 112.55)	(32.08, 35.74)
acl (x,y)	(769.07, 321.92)	(192.03, 225.45)
almr (x,y)	(326.28, 727.48)	(75.46, 208.19)
ammr (x,y)	(670.52, 720.83)	(418.54, 88.61)
pcl (x,y)	(332.68, 109.64)	(148.62, 27.27)
plmr (x,y)	(570.84, 520.73)	(180.58, 159.46)
pmmr (x,y)	(459.02, 1219.24)	(286.87, 235.75)

Chapter 5

LEARNING LONGITUDINAL PLANNING FOR PERSONALIZED HEALTH MANAGEMENT FROM DAILY BEHAVIORAL DATA

5.1 Abstract

Mitigating globally emerging health problems such as obesity needs scalable solutions that can promote healthier lifestyles outside of clinical settings. Such scalable solutions, while targeting general population, need to automatically provide personalized behavior change plans that fit an individual's preferences and needs. There has been fast-growing development of sensing devices and applications for continuous monitoring of human behavior (such as physical activity and food intake) and health status such as BMI. However, there are challenges to translate these noisy and dynamic behavioral data into personalized planning. To address both challenges, we develop a systematic framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion to translate users' behavioral data into deeply personalized health planning. We further apply our proposed model on a real-world user behavioral dataset, which demonstrates the promising utility and efficacy of our method.

5.2 Introduction

Emerging technologies, such as smartphones and wearable sensors, have provided health professionals unprecedented monitoring and intervention capacity to materialize the envisioned personalized and preventative healthcare, particularly for those health management problems that need scalable solutions to promote healthier lifestyles outside of clinical settings. In this paper we study how to learn personalized and achievable behavior change plans from users' daily behavioral data. A typical application is obesity prevention, which relates to over one-third of the US adult population [148]. Addressing obesity-related problems is believed to be beyond the capacity of the

healthcare industry [149], calling for scalable solutions that can automate personalized decision-makings. Emerging tracking devices and applications can help monitor and regulate physical activity and food intake as well as collecting health data [150]. However, existing strategies only exploit limited value of these data so that feedback to individuals is often limited to either overall statistics [150], visualization of self-tracked data [29], or generic suggestions [30] without being personalized to a user's lifestyle.

We hypothesize that the fine-grained information contained in behavioral data monitored daily, or by hour or even minute, can be better exploited. For instance, the underlying dynamic model, which governs the relationships between the behavioral variables (e.g. activities and food intake) and health outcomes, can be learned and leveraged to generate personalized and actionable suggestions. A visionary paper [151] pointed out that "it is promising to apply control systems engineering principles [152] to design and implement a behavior change system that is optimized to deploy finely tailored, properly dosed, just-in-time bouts of intervention at the precise moment and context when they will be maximally effective at influencing your behavior." Thus, the ultimate objective is to translate users' behavior data into deeply personalized and achievable health recommendation. It requires a smart learning and planning engine that can first learn the dynamic model that governs the relationships between the users' physical activity, dietary behavior, and heath outcomes such as BMI; and then, strategically suggest changes to those behaviors for a healthier lifestyle. Personalized recommendation is not reinventing the wheel. It's applying what doctors have been doing on a larger scale via a smart and automatic engine, enabled by the emerging big personal behavior and health data.

To the best of our knowledge, there still lacks a systematic methodology that can learn from the non-existing fine-grained data for dynamic modeling, let alone personalized planning that should build on the learned dynamic model. To make the best use of the data, we recognize that there are both learning challenges and planning challenges as elaborated below:

- Challenges in learning from behavior data: On top of the underlying complex multivariate dynamics, the missing values and outliers, commonly found in users' behavioral data, present

difficulty in learning [153, 154]. The ability to automatically learn the dynamic model(s) for many individuals from noisy behavioral data is currently lacking in literature.

- Challenges in personalized longitudinal planning: How to formulate the optimal planning on the foundation of the dynamic model (that characterizes the hidden principles representing the physical or physiological constraints that any health planning has to comply with); how to characterize the user's preferences and incorporate them in planning; and how to learn from peers (that form a potential mentor group of the target user), are the planning challenges.

To address the arising challenges, we develop a longitudinal planning framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion. Our contributions include:

- A dynamic system learning method – SSMO, which can automatically remove the effects of potential outliers in the dataset, impute the missing values, and conduct model identification.

- A dynamic planning system that can learn the optimal behavior change plan guided by the individual's own dynamic model. To enhance the quality of the planning, we further formalize users' preferences and needs as constraints, and constructing an action polyhedron for each user by adopting dictionary learning on the actions of peers that form a potential mentor group.

- Efficient algorithms to solve the learning and planning problems with specific optimization strategies to ensure the feasibility and robustness of the algorithms. Extensive numerical studies on both synthetic and real-world data demonstrate the utility and efficacy of our methods.

5.3 Methodology

Our proposed learning pipeline comprises three major components: (1) a dynamic system identification engine that automatically takes care of missing values and outliers; (2) a dictionary learning

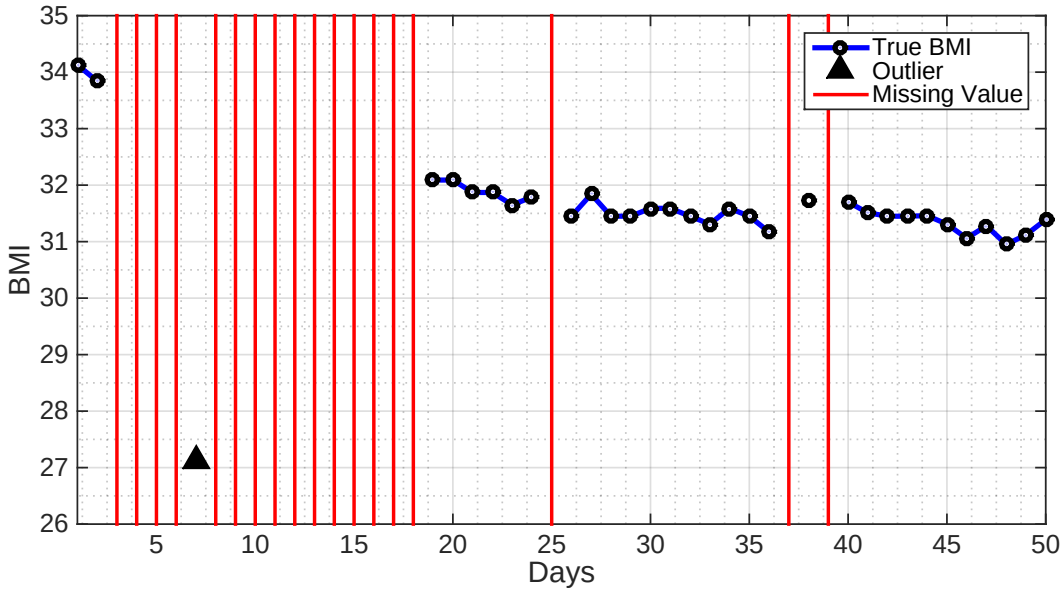


Figure 5.1: Outliers and missing values of BMI

approach for action polyhedron construction; and (3) a personalized longitudinal planning algorithm.

5.3.1 Dynamic System Identification with Simultaneous Missing Value Estimation and Outlier Detection (SSMO) Engine

We develop a dynamic system learning method, SSMO, which can automatically remove the effects of potential outliers in the dataset, impute the missing values, and conduct model identification at the same time. Given a user record with a series of observations, $\{\mathbf{x}_t, \mathbf{u}_t\}_{t=1}^{T}$, where $\mathbf{x}_t$ denotes the outcome at time t which can be extended to any dimension and $\mathbf{u}_t$ denotes the behavior profile as a column vector at time t, we propose to adopt a linear dynamic system as the underlying model to characterize the relationships between the behavioral and outcome variables. The linear dynamic system is a very flexible model that can characterize a wide range of dynamics. For instance, in what follows we illustrate the specific use of a linear dynamic system to model the 3rd-order

dynamics:

$$\begin{bmatrix} \mathbf{x}_{t+1} \\ \mathbf{x}_t \\ \mathbf{x}_{t-1} \end{bmatrix} = A \begin{bmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \mathbf{x}_{t-2} \end{bmatrix} + B \begin{bmatrix} \mathbf{u}_t \\ \mathbf{u}_{t-1} \\ \mathbf{u}_{t-2} \end{bmatrix} + C + \mathbf{w}_t, \tag{5.1}$$

where $\mathbf{w}_t$ is white noise and C is a bias term. Such a formulation can capture both spontaneous effect (i.e., from $\mathbf{u}_t$ to $\mathbf{x}_t$) as well as delayed effect (i.e., from $\mathbf{u}_{t-2}$ to $\mathbf{x}_{t+1}$). Apparently, this formulation is generic and can be further extended to capture higher-order dynamics. It can also be recognized as an equivalent form with the common continuous linear dynamic system that models $[\mathbf{x}(t); \dot{\mathbf{x}}(t); \ddot{\mathbf{x}}(t)]$ [155] if we rewrite $[\mathbf{x}_t; \mathbf{x}_{t-1}; \mathbf{x}_{t-2}]$ as $[\mathbf{x}_t; (\mathbf{x}_t - \mathbf{x}_{t-1}); (\mathbf{x}_t + \mathbf{x}_{t-2} - 2\mathbf{x}_{t-1})]$ (note that the parameters A, B, and C would be different then).

The main challenges for system identification to learn model parameters A, B, and C arise from the large number of missing values and frequent outlier points in $\{\mathbf{x}_t, \mathbf{u}_t\}_{t=1}^{T}$, clearly illustrated in a fragment of the real-world time series of BMI measurements in Figure 5.1. Inappropriate handling of missing values and outliers may lead to the computational difficulties from holes in the dataset, as well as the bias and loss of precision due to distortion of the data distribution [156]. For example, among the approaches that handle missing values [157], the "mean imputation" method ignores the context as it fails to utilize the underlying dynamics of the variables. The "last value carried forward" method takes a conservative approach, underestimating the changes over time. Thus, neither method is suitable for imputing missing values in the dynamic modeling context. Later both methods will be used as the baseline methods for performance comparison in Section 5.4.1. In addition to missing values, the existence of outliers would inflate the variable variances and cause large estimation errors, particularly for linear models.

The basic idea of SSMO for better dynamic system identification is to simultaneously impute missing values, delete outliers, and train the dynamic model, so that the imputation of missing values and detection of outliers are in a proper context defined by the learned dynamic model. Let $X = [\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T]$ be the state matrix and $U = [\mathbf{u}_0, \mathbf{u}_1, \cdots, \mathbf{u}_{T-1}]$ be the action matrix. Ω_x and Ω_u represent the observed elements in X and U, respectively, with their complement sets denoting

missing values. Define $\hat{X} = [\hat{\mathbf{x}}_0, \hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_T]$ and $\hat{U} = [\hat{\mathbf{u}}_0, \hat{\mathbf{u}}_1, \cdots, \hat{\mathbf{u}}_{T-1}]$ as the estimates of X and U. We propose to learn $\hat{X}$ and $\hat{U}$ in SSMO to be consistent with X and U on the observed sets Ω_x, Ω_u respectively. The identified outlier and missing elements in X are essentially free variables in $\hat{X}$.

$$\min_{\substack{A,B,C,\\ \hat{X},\hat{U}}} \frac{1}{2} \sum_{t=0}^{T-1} \left\| \begin{bmatrix} \hat{\mathbf{x}}_{t+1} \\ \hat{\mathbf{x}}_t \\ \hat{\mathbf{x}}_{t-1} \end{bmatrix} - \left(A \begin{bmatrix} \hat{\mathbf{x}}_t \\ \hat{\mathbf{x}}_{t-1} \\ \hat{\mathbf{x}}_{t-2} \end{bmatrix} + B \begin{bmatrix} \hat{\mathbf{u}}_t \\ \hat{\mathbf{u}}_{t-1} \\ \hat{\mathbf{u}}_{t-2} \end{bmatrix} + C \right) \right\|^2 \tag{5.2a}$$

$$\text{s.t.} \quad \|(\hat{X} - X)_{\Omega_x}\|_0 \leq a; \quad \|(\hat{U} - U)_{\Omega_u}\|_0 \leq b. \tag{5.2b}$$

The objective (5.2a) is a squared loss function to evaluate the goodness-of-fit of the system parameters A, B, C and the estimates $\hat{U}, \hat{X}$. The constraints (5.2b) are to control the number of different estimated elements from the observed elements in Ω_x and Ω_u, which essentially controls the number of outliers among the observed elements. The parameters a and b restrict the maximal number of outliers. When they are set to 0, (5.2) only handles missing values. The values for a and b actually are not hard to decide. For example, we could estimate the upper bound of the percentage of outliers that is easily accessible in many applications. This algorithm is indeed robust as long as a and b are greater than the actual number of outliers but not far away from it.

Solving SSMO by Block Coordinate Descent(BCD)

We apply Block Coordinate Descent (BCD) [158] to solve (5.2) by alternatively optimizing two groups of variables $\{A, B, C\}$ and $\{\hat{X}, \hat{U}\}$:

- To optimize $\{A, B, C\}$, it is a least squares optimization with a closed-form solution.

- To optimize $\{\hat{X}, \hat{U}\}$, we adopt the projected gradient descent method to iteratively update:

$$\hat{X}_{k+1} = \arg\min_{\hat{X}} \big\{ \|\hat{X} - (\hat{X}_k - \gamma g_{\hat{X}_k})\|_F^2; \\ \text{s.t. } \|(\hat{X} - X)_{\Omega_x}\|_0 \leq a \big\},$$

  where $g_{\hat{X}_k}$ is the partial derivative of the objective function (5.2a) w.r.t. $\hat{X}_k$; γ is the step size that could be chosen to be a sufficiently small constant; and $\|\cdot\|_F$ denotes Frobenius

norm. It actually also admits a closed-form solution that can be found by: First, selecting a elements in $(\hat{X}_k - \gamma g_{\hat{X}_k} - X)_{\Omega_x}$ with the largest magnitudes as the outliers at the current iteration and forming a set S; Second, setting the elements outside of Ω_x and in set S: $(\hat{X}_{k+1})_{\bar{\Omega}_x \cup S} = (\hat{X}_k - \gamma g_{\hat{X}_k})_{\bar{\Omega}_x \cup S}$; Third, setting the remaining elements in $\hat{X}_{k+1}$ to take the same values in $\hat{X}_k$. To update $\hat{U}_{k+1}$ from $\hat{U}_k$, one can follow a similar procedure. Due to the space limit, we omit the detailed derivation.

We summarize all the steps in Algorithm 1.

Algorithm 1 BCD for SSMO

Require: $X_{\Omega_x}, U_{\Omega_u}, a, b$

Ensure: $A, B, C, \hat{X}, \hat{U}$

1: **repeat**

2: Optimize A, B, C by minimizing the least squares problem (5.2a) without any constraint.

3: Optimize $\hat{X}$: Select top a largest elements in $(\hat{X} - \gamma g_{\hat{X}} - X)_{\Omega_x}$, which forms the index set S; Update elements of $\hat{X}$ in $\bar{\Omega}_x \cup S$ by

$$(\hat{X})_{\bar{\Omega}_x \cup S} \leftarrow (\hat{X} - \gamma g_{\hat{X}})_{\bar{\Omega}_x \cup S}.$$

4: Optimize $\hat{U}$: similar to the updates of $\hat{X}$;

5: **until** convergence.

6: **return** $A, B, C, \hat{X}, \hat{U}$;

5.3.2 Personalized Longitudinal Planning

We present our personalized longitudinal planning model to derive recommendations that accommodate an individual user's needs and preferences based on the user's dynamic model. Specifically, it is to identify an optimal sequence of actions, denoted by $\mathbf{u}_0$, $\mathbf{u}_1$, $\cdots$, $\mathbf{u}_{T-1}$, to drive the user's initial health status $\mathbf{x}_0$ to the target status $\mathbf{x}_T$ in T days. For example, with BMI as the health status, the user may want to reduce BMI from the current level $\mathbf{x}_0 = 30$ to the target 28 in 90 days. With

a dynamic model, any proposed planning can be evaluated with the predicted future health status. The challenge is how to utilize this capacity to derive the optimal planning. On the other hand, we should formalize the user's preferences as optimization constraints to enhance the quality of the generated optimal planning. This leads to the following formulation:

$$\min_{\mathbf{u}_0,\mathbf{u}_1,\cdots,\mathbf{u}_{T-1}} \quad \sum_{t=1}^{T} \mathbf{c}^\top \mathbf{u}_t + \lambda \sum_{t=1}^{T-1} \|\mathbf{u}_t - \mathbf{u}_{t-1}\|_1 \tag{5.3a}$$

$$\text{s.t.} \quad \hat{\mathbf{x}}_T \leq \text{target} \tag{5.3b}$$

$$\mathbf{u}_- \leq \mathbf{u}_t \leq \mathbf{u}_+ \tag{5.3c}$$

$$\mathbf{u}_0 = \mathbf{h} \tag{5.3d}$$

$$\mathbf{u}_t \in \text{conv}(D). \tag{5.3e}$$

The objective function (5.3a) consists of two terms: The first term is to measure the cost of the adopted action, as different users might have different preferences or difficulties in conducting the actions; and the second term is to measure the smoothness of actions across all time points, assuming that users do not like sudden changes between consecutive actions as what the low-effort theory implies [159]. The first constraint (5.3b) is to ensure that, by following the planning, the user will achieve the pre-specified goal, where $\hat{\mathbf{x}}_T$ is the estimated final health status using the dynamic model (5.1). Specifically, based on the initial status $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2$ and the action sequence $\mathbf{u}_0$, $\mathbf{u}_1$, $\cdots$, $\mathbf{u}_{T-1}$, $\hat{\mathbf{x}}_T$ can be estimated from $[\hat{\mathbf{x}}_T; \hat{\mathbf{x}}_{T-1}; \hat{\mathbf{x}}_{T-2}] = A^{T-2}[\mathbf{x}_2; \mathbf{x}_1; \mathbf{x}_0] + \sum_{t=2}^{T-1} A^{T-t-1} B[\mathbf{u}_t; \mathbf{u}_{t-1}; \mathbf{u}_{t-2}]$. The second box constraint (5.3c) is to avoid unwanted and unrealistic actions for a specific user. The third constraint (5.3d) ensures the recommendation to start from the habits $\mathbf{h}$ of the user. We further impose another constraint (5.3e) to enforce the recommended actions to be within a set $\text{conv}(D)$. This is named as the *action polyhedron* that specifies the action space, in which a realistic planning can be constructed. It provides us great flexibility to incorporate domain knowledge or any kind of prior knowledge to guide the personalized longitudinal planning. In Section 5.3.3, we will introduce a learning-from-peers approach to construct this action space by a novel dictionary learning approach.

To conclude this subsection, it is worth giving a summary of how to set up the longitudinal

planning formulation in practice: First, we can set the personalized lower and upper bounds for all action variables: $\mathbf{u}_-$ and $\mathbf{u}_+$; Then, we set the preference scores for extreme actions, $\mathbf{c}_-$ and $\mathbf{c}_+$ corresponding to $\mathbf{u}_-$ and $\mathbf{u}_+$; Based on the user's preferences, we can calculate the action weight vector as the preference score $\mathbf{c}$ in the objective function: $\mathbf{c} = (\mathbf{c}_+ - \mathbf{c}_-) \oslash (\mathbf{u}_+ - \mathbf{u}_-)$, where $\oslash$ is the element-wise division.

Optimization

The longitudinal planning formulation actually embodies a linear programming (LP) problem. Note that all $\mathbf{u}_t$'s are automatically restricted in the convex hull of D. Thus, for any $\mathbf{u}_t$, there exists an α_t such that $\mathbf{u}_t = D\alpha_t$, $\alpha_t \geq 0$ and $1^\top \alpha_t = 1$. With this transformation, the problem (5.3) turns out to be Eq. (5.4).

$$\begin{aligned} \min_{\alpha_0,\cdots,\alpha_{T-1}} \quad & \sum_{t=1}^{T} \mathbf{c}^\top D\alpha_t + \lambda \sum_{t=1}^{T-1} \|D(\alpha_t - \alpha_{t-1})\|_1 \\ \text{s.t.} \quad & \hat{\mathbf{x}}_T \leq \text{target}, \quad \mathbf{u}_- \leq D\alpha_t \leq \mathbf{u}_+, \\ & D\alpha_0 = \mathbf{h}, \quad \alpha_t \geq 0, \quad \mathbf{1}^\top \alpha_t = 1. \end{aligned} \tag{5.4}$$

We further adopt a standard technique in optimization [160] to replace the L_1-norm term by introducing two variables, $\beta_t^+ \geq 0$ and $\beta_t^- \geq 0$, to represent the positive and negative components of $D(\alpha_t - \alpha_{t-1})$, so that we derive the final LP formulation that can be solved using any LP solver:

$$\begin{aligned} \min_{\substack{\{\alpha\}_{t=0}^{T-1}, \\ \{\beta_t^+,\beta_t^-\}_{t=1}^{T}}} \quad & \sum_{t=1}^{T} \mathbf{c}^\top D\alpha_t + \lambda \sum_{t=1}^{T-1} \mathbf{1}^\top (\beta_t^+ + \beta_t^-) \\ \text{s.t.} \quad & \hat{\mathbf{x}}_T \leq \text{target}, \quad \mathbf{u}_- \leq D\alpha_t \leq \mathbf{u}_+, \\ & \alpha_t \geq 0, \quad D\alpha_0 = \mathbf{h}, \quad \mathbf{1}^\top \alpha_t = 1, \\ & D(\alpha_t - \alpha_{t-1}) = \beta_t^+ - \beta_t^-, \\ & \beta_t^+ \geq 0, \quad \beta_t^- \geq 0. \end{aligned} \tag{5.5}$$

5.3.3 Action Polyhedron Construction (APC) Engine

We now introduce the dictionary learning approach to construct the action polyhedron conv(D) as the feasible region of actions, ensuring the recommendations are realistic and reasonable in

practice. In particular, conv(D) can be viewed as a summary of the typical action patterns of other users that form a potential mentor group for the target user.

Let $U \in \mathbb{R}^{p\times n}$ be the action matrix that we could use to learn conv(D), i.e., each column represents an action vector that has been undertaken in real life by a certain user. U can be constructed by collecting n actions from different users whose behavioral patterns could inspire the planning for the target user. Then, the formulation of the dictionary learning can be written as (5.6).

$$\begin{aligned} \min_{W} \quad & \frac{1}{2}\|U - UW\|_F^2 + \gamma\|W\|_{2,1} \\ \text{s.t.} \quad & W \geq 0 \\ & \mathbf{1}^\top W = \mathbf{1}. \end{aligned} \tag{5.6}$$

The first term is to measure the approximation adequacy to represent all actions (vectors) in U using a set of basis actions (vectors). The second term is to enforce sparsity in the basis matrix W, i.e., the nonzero columns of the learned W indicate the selected actions. Figure 5.2 illustrates the goal of finding a polyhedron to represent the whole action set U using a convex hull as the action polyhedron D. Figure 5.2 is based on over 10,000 behavioral actions collected from over 30 users. It shows that the behavioral actions undertaken by this cohort exhibit a clear regularity, indicating that human behavioral actions follow certain principles and are not totally random. Thus, to generate personalized longitudinal planning, it is required that the planning should consist of reasonable actions that fit the "human patterns". Further, Figure 5.2 shows that the dictionary learning formulation provides a very effective approach to extract patterns and summarize the massive data matrix U.

Challenges in Optimization Note that the proposed dictionary learning method is different from the existing methods that have been used in pattern recognition such as [161, 162], and event detection [163]. To solve (5.6), we face two challenges: The first one is that (5.6) involves high dimensional $W \in \mathbb{R}^{n\times n}$; The other challenge lies on that (5.6) includes nonsmooth regularization term and constraints. It takes hours to solve it with $n = 1000$ if using general solvers, for example, CVX [65]. In the following, we will derive an efficient algorithm to solve it. We apply the general

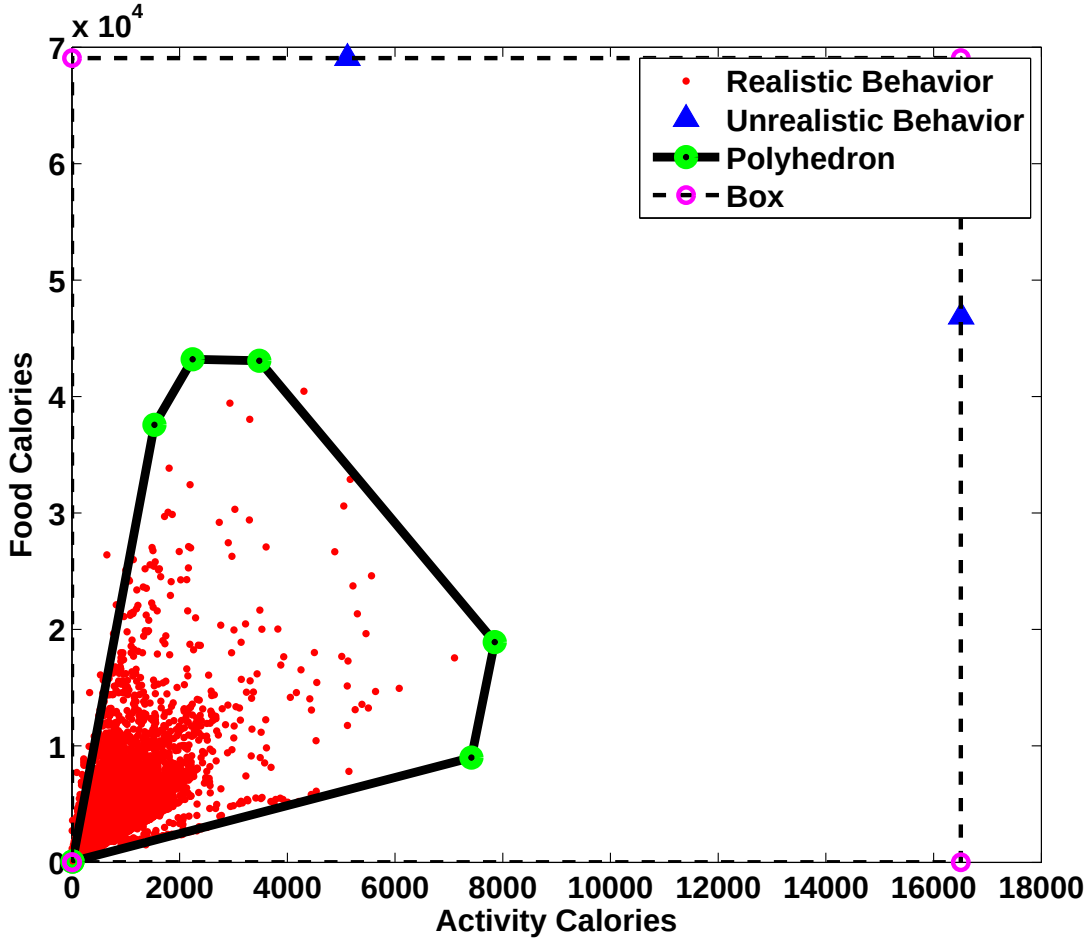


Figure 5.2: The constructed action polyhedron D learned from over 10,000 behavioral actions collected from over 30 users

ADMM optimization framework to decompose Eq. (5.6) into several nontrivial subproblems. We skip some derivation details due to the space limit, and directly present the computational procedure. First, to adopt the ADMM framework, we duplicate the variable W with another variable V, and rewrite the formulation Eq. (5.6) as follows:

$$\begin{aligned} \min_{W,V} \quad & \frac{1}{2}\|U - UW\|_F^2 + \gamma\|V\|_{2,1} \\ \text{s.t.} \quad & V \geq 0 \\ & \mathbf{1}^\top W = \mathbf{1}^\top \\ & W = V. \end{aligned} \tag{5.7}$$

Then, we define the augmented Lagrangian of Eq. (5.7) and further convert it to be:

$$\begin{aligned} L_\rho(W, V, \Lambda) = & \frac{1}{2}\|U - UW\|_F^2 + \gamma\|V\|_{2,1} + \frac{\rho}{2}\|W - V\|_F^2 \\ & + \langle \Lambda, W - V \rangle + \mathbb{I}_{V \geq 0}(V) + \mathbb{I}_{\mathbf{1}^\top W = \mathbf{1}^\top}(W), \end{aligned} \tag{5.8}$$

where $\mathbb{I}_{\text{condition}(\cdot)}$ is the Delta function: It gives zero if the condition is satisfied by the augment; Otherwise, it returns $+\infty$. ρ could be an arbitrary positive number.

Following the ADMM procedure, we then iterate over three steps:

Minimize $L_\rho(W, V, \lambda)$ w.r.t. W: It essentially solves the following optimization:

$$\min_{W} \quad \frac{1}{2}\|U - UW\|_F^2 + \frac{\rho}{2}\|W - V\|_F^2 + \langle \Lambda, W \rangle + \mathbb{I}_{\mathbf{1}^\top W = \mathbf{1}^\top}(W).$$

In general, one needs an iterative algorithm to solve it. We derive a closed form to solve this problem efficiently. By deriving the KKT conditions [164] of this problem, the optimal solution is equivalently defined as

$$\begin{bmatrix} U^\top U + \rho\mathbf{I} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} \begin{bmatrix} W \\ \Phi^\top \end{bmatrix} = \begin{bmatrix} U^\top U + \rho V - \Lambda \\ \mathbf{1}^\top \end{bmatrix}, \tag{5.9}$$

where Φ is the dual variable. By solve the equation (5.9), we can obtain the closed-form solution as:

$$W = (\mathbf{I} - N \cdot \mathbf{1}^\top)A^{-1}B + N \cdot \mathbf{1}^\top,$$

where

$$N = \frac{A^{-1}\mathbf{1}}{\mathbf{1}^\top A^{-1}\mathbf{1}}, A = U^\top U + \rho\mathbf{I},$$
$$B = U^\top U + \rho V - \Lambda, \tag{5.10}$$

and $\mathbf{I}$ is an identity matrix and by applying the Sherman-Morrison-Woodbury theorem with algebraic manipulations, A^{-1} can be calculated efficiently.

Minimize $L_\rho(W, V, \Lambda)$ w.r.t. V: It essentially solves the following optimization:

$$\min_V \frac{\rho}{2}\|W - V\|_F^2 - \langle \Lambda, V\rangle + \gamma\|V\|_{2,1} + \mathbb{I}_{V\geq 0}(V). \tag{5.11}$$

For readers who are familiar with group LASSO [165, 163], this is similar to the proximal step for $L_{2,1}$-norm regularization. The key difference here lies on the positive constraint $\mathbb{I}_{V\geq 0}(V)$. We can also derive a closed-form solution for V, by applying a projection operation onto the positive constraint and then a followed proximal step for $L_{2,1}$-norm regularization: $V_{i\cdot} = \max(0, 1 - \frac{\beta}{\|Y_{i\cdot}\|}) \cdot Y_{i\cdot}$, where $Y = \max(0, W + \frac{1}{\rho}\Lambda)$, $\beta = \frac{\lambda}{\rho}$ and $i\cdot$ indicates the ith row of the matrices V and Y.

Update Λ: This step mimics the dual gradient ascent: $\Lambda = \Lambda + \rho(W - V)$.

We finally summarize the algorithm in Algorithm 2.

Algorithm 2 ADMM for APC

Require: U, $\rho > 0$, and γ

Ensure: W

1: **repeat**
2: Minimize $L_\rho(W, V, \Lambda)$ in terms of W and update $W \leftarrow (I - N\mathbf{1}^\top)A^{-1}B + N\mathbf{1}^\top$, where A, B, N are defined in (5.10);
3: Minimize $L_\rho(W, V, \Lambda)$ in terms of V and update $V_{i\cdot} \leftarrow \max(0, 1 - \frac{\beta}{\|Y_{i\cdot}\|}) \cdot Y_{i\cdot}$;
4: $\Lambda \leftarrow \Lambda + \rho(W - V)$;
5: **until** convergence
6: **return** W;

5.4 Numerical Studies

The proposed framework includes three main components: the SSMO engine, the dictionary construction APC, and the personalized recommendation. In the following, we evaluate the effectiveness of each key component and the techniques being used. We show by numerical studies that removing any of them can lead to the decline in the overall performance.

The data used in experiment is collected in a longitudinal study that involves more than 1000 real-world users and each user has several yearsâĂŹ daily measurements (collected from wearable devices, including diet, sleep, exercise information, and BMI). We evaluated the framework based on this dataset and also evaluated the effect of personalized recommendation using 25 users whose data are preprocessed and ready for analysis. For the 25 users, in total we have more than 10000 daysâĂŹ measurements. We also would release a fitness data simulator upon the acceptance of this paper with a full dictionary of the variables.

5.4.1 Performance Evaluation of SSMO

We first evaluate the performance of SSMO and compare with the existing benchmark methods for imputing missing values and removing outliers, including the "mean imputation", the "last value carried forward" method, and the variants of them with a median filter to remove outliers, as mentioned in [157]. In addition, we also compared the SSMO with more missing data imputation methods including the ones based on Functional Data Analysis (FDA) with different bases such as B-spline, Haar wavelet [**?**], non-parametric Principal Analysis by Conditional Expectation (PACE)[**?**], MICE[**?**], Amelia[**?**], MissForest[**?**], and MI[**?**]. Results in Table 5.1 clearly show that SSMO consistently outperforms all benchmark methods with much lower errors and variances.

We further evaluate the performance of SSMO when both missing values and outliers present in the dataset. Here, we analyze a real-life fitness data with users' daily fitness behaviors collected from sensoring devices, including diet, sleep, exercise information, and BMI as health outcomes. There are 25 users' data in this dataset, while almost every user's data show significant missing values and outliers with similar patterns in Figure 5.5. Again SSMO achieves more accurate model

Table 5.1: Comparison of Estimation Error

	SSMO	Mean	Last	Mean+Med	Last+Med	FDA (B-Spline)
RMSE	**0.67 ± 0.49**	1.31± 2.13	1.06 ±1.06	0.74 ± 0.60	0.86 ± 0.57	0.69 ± 0.51
	FDA (Haar)	PACE	MICE	Amelia	MissForest	MI
RMSE	0.72± 0.46	0.71± 0.55	1.34 ± 0.46	1.83 ± 0.95	1.73 ± 0.86	1.30 ± 0.44

estimation, as reflected by the prediction errors in those users. Figure 5.5 shows the details of the prediction results by SSMO, and the other two imputation methods. For each user, the dynamic model has been built based on the data generated during the first half of days, and evaluated on the other half for prediction errors.

5.4.2 Performance Evaluation of Dictionary Construction with APC

Dictionary construction restricts the recommended actions within a space spanned by some existing users' action data. There is an implicit assumption of this method that hypothesizes that, although people are different and have heterogeneous behavioral patterns, there are some regularity or canonical structure governing the human behavior. Therefore, the effectiveness of the dictionary construction method APC depends on how valid this assumption holds true in reality. Here, we apply APC on the 25 users' behavioral data. Figure 5.3 provides the results regarding how many basis vectors we need to represent all the behavioral data of all the users. Apparently, the larger the dictionary size is, the better (lower) the error of representation by the squared Frobenius norm that it can provide. On the other hand, we can also observe that the error of representation drops quickly with the increasing number of basis vectors in the dictionary. With eleven basis vectors, the error of representation approaches 1.0.

A visualization of the five basis vectors are shown in Figure 5.4, representing five typical health management routines used in the cohort and three levels of combinations of the routines. For example, Pattern 1 highlights the decent amount of calories consumed from food with less activities,

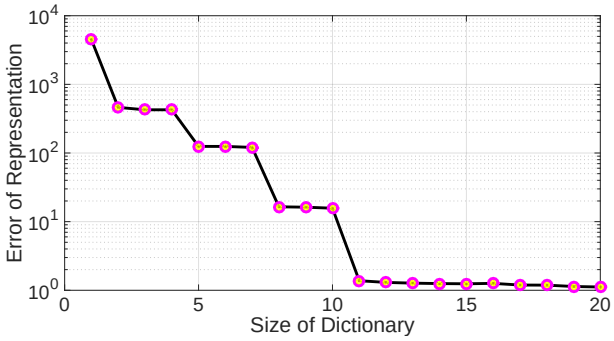


Figure 5.3: Errors of representation of U based on the size of dictionary

while Pattern 3 is a more balanced diet and exercise routine. Interestingly, the learned patterns can be mapped to the official guideline for obesity prevention[166].

Figure 5.4: Five Recommended Routines and Three Plans

5.4.3 Performance Evaluation of the Planning Method

While the ultimate evaluation for any healthcare planning strategy is to conduct clinical trials, it is expensive and often can only be realistic at the later stage of the intervention development. On the other hand, the literature shows that compliance to health recommendations is such a complex

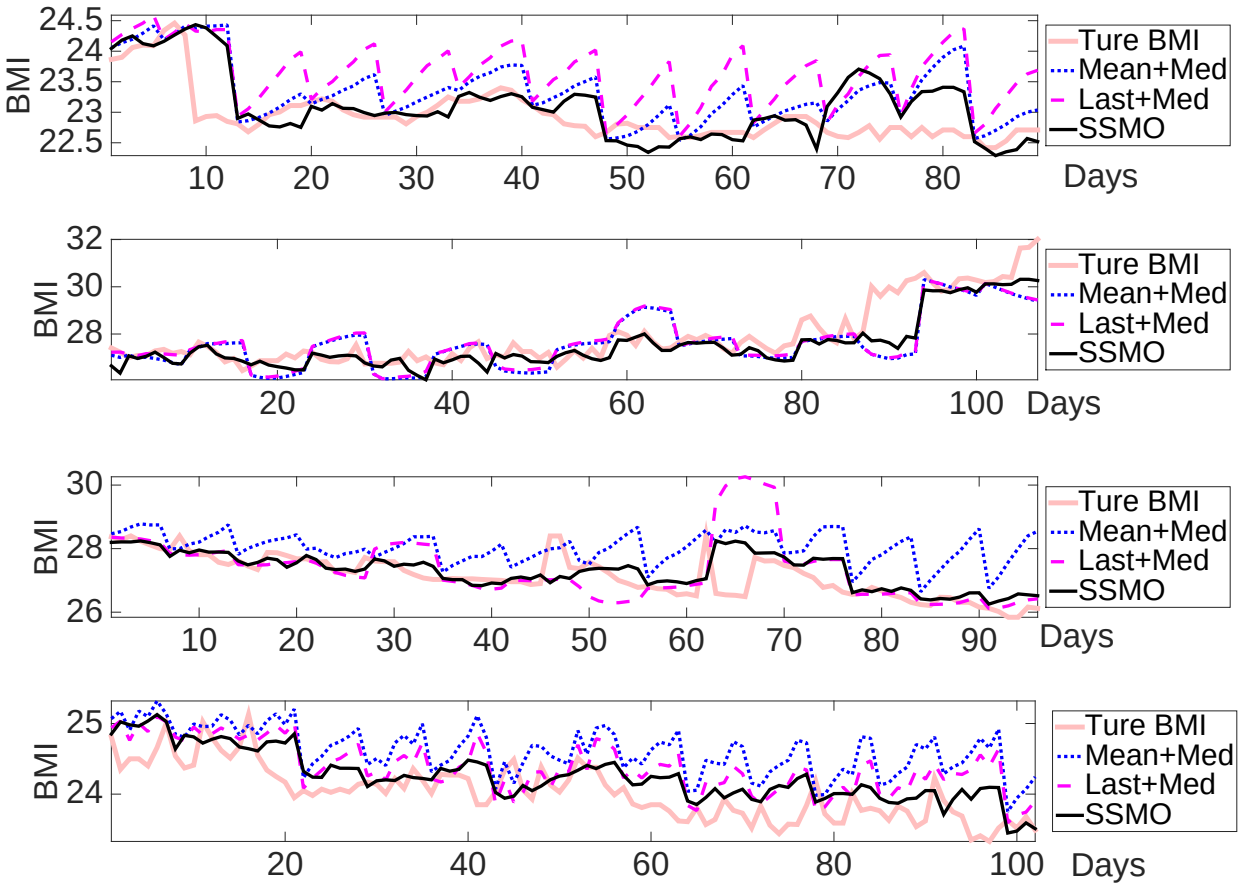


Figure 5.5: BMI Estimation with 3 Different Methods.

behavior that the compliance levels vary from user to user and even for the same user. Therefore, we take a pragmatic approach to evaluate the efficacy of our strategy via data-based simulations with scenarios that reflect different user compliance levels. We first simulate the dynamic change of the health outcome. Specifically, we randomly pick up four users with their behavioral data (with last M measurements held out for evaluation) to train the dynamic model using SSMO and further derive the optimal planning as a temporal action set U using our planning formulation. We investigate a range of compliance levels. For example, a 80% compliance level means 80% actions are randomly picked from the optimal plan, and the rest are from the originally observed behavior. We then predict the BMI change based on users' learned dynamic model.

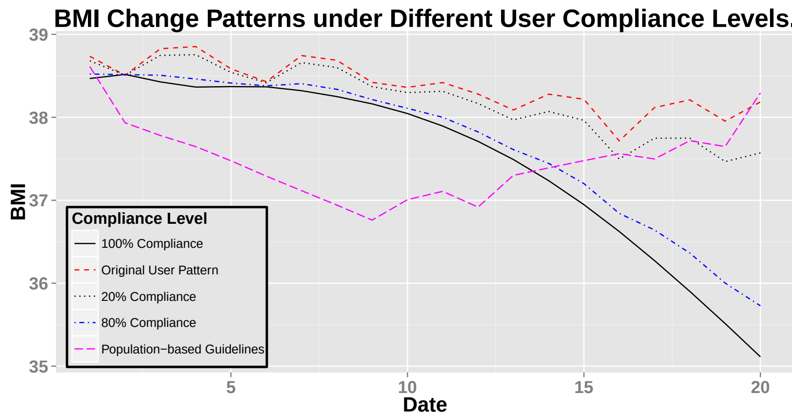
The results in Figure 5.6 indicate that outcomes from users' original routines will either fluctuate and then become worse, or stay in a plateau, or continue to become worse. While for all users, complete or partial compliance to the optimal plans always leads to better and steady results. Note that, we also compare our method with the population-based planning, i.e., recommending the user with the mean level of activities from the user's peers. The population-based planning might be reasonable for users with average conditions, but may not perform as well or even result in condition deterioration for the users with very high or low BMI profiles. Again, as shown in Figure 5.6, our method outperforms population-based planning by providing more effective personalized plans and interventions.

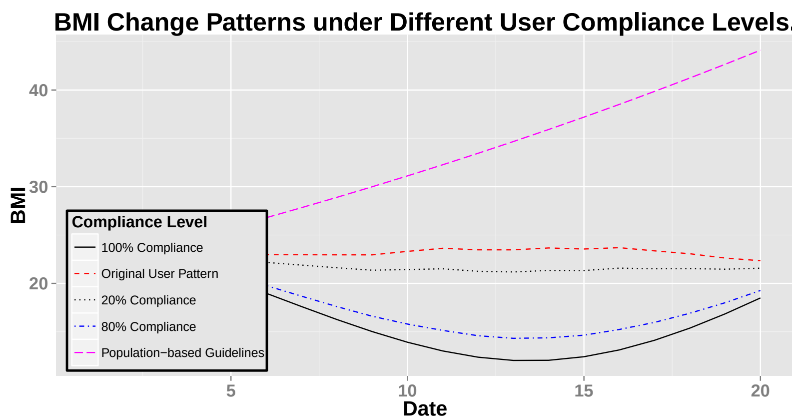
5.4.4 Effect of the Smoothness and Dictionary Constraints on Recommendation

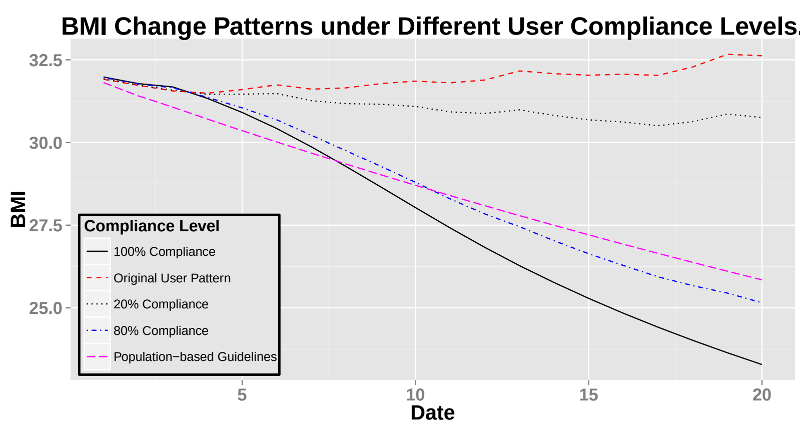
We also investigate the effectiveness of the smoothness constraint and the action polyhedron on the quality of the generated plannings. While the planning quality is a multi-facet concept, a simple criterion is that the derived plan should fit the existing behavioral patterns. Thus, our strategy here is to randomly pick up an user and generate three optimal plans: Plan A from the full model as depicted in Section 5.3.2; Plan B from a reduced model without the constraint (5.3e); and Plan C from a further reduced model by removing the constraint (5.3e) and setting λ in (5.3a) to 0. The three plans are drawn on top of the five typical patterns in Figure 5.4. It is obvious that the optimal plan derived from the full model fits the existing patterns better, which seems to be more realistic and has a higher likelihood to be adopted by the users than the other two plans that are quite different from the existing patterns. In addition, we quantitatively evaluated this conclusion by calculating the distance between the plans to the space defined by the basis vectors (as a score representing how similar the plans with real-world actions). We used cosine similarity that showed the scores for the Plan A, B, C are 0.901, 0.856, and 0.855, respectively.

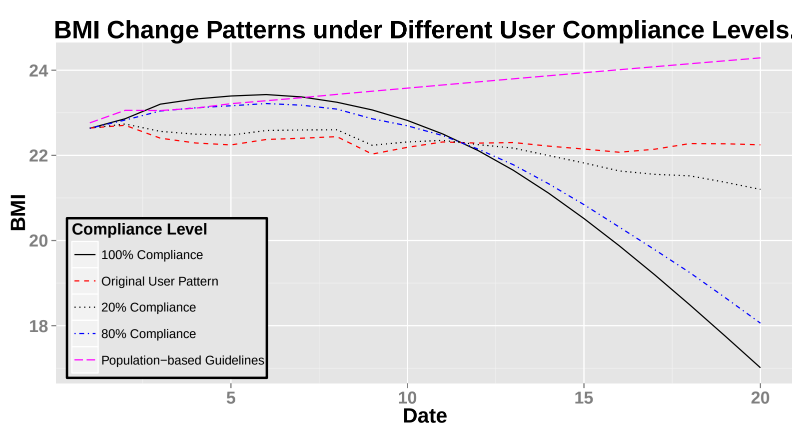
5.5 Conclusions

In this work, we have developed a systematic framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion, to automate the process of translating the user's behavior data into deeply personalized and achievable health planning. Our method is generic and can be applied to a wide range of health management problems such as obesity, fitness, diabetes, or any chronic conditions, where the disease process is a complex dynamic process that can be modified by exogenous variables such as environmental, behavioral, and clinical variables. Our framework holds great potential to provide scalable solutions for mitigating these health problems, which can promote healthier lifestyles outside of clinical settings. We further apply our proposed model on real-world daily behavioral data, which demonstrate promising utility and efficacy of our method.


BMI Change Patterns under Different User Compliance Levels.
BMI
Date
39
38
37
36
35
5
10
15
20
Compliance Level
100% Compliance
Original User Pattern
20% Compliance
80% Compliance
Population−based Guidelines


BMI Change Patterns under Different User Compliance Levels.
BMI
Date
40
30
20
5
10
15
20
Compliance Level
100% Compliance
Original User Pattern
20% Compliance
80% Compliance
Population−based Guidelines


BMI Change Patterns under Different User Compliance Levels.
BMI
Date
32.5
30.0
27.5
25.0
5
10
15
20
Compliance Level
100% Compliance
Original User Pattern
20% Compliance
80% Compliance
Population−based Guidelines


BMI Change Patterns under Different User Compliance Levels.
BMI
Date
24
22
20
18
5
10
15
20
Compliance Level
100% Compliance
Original User Pattern
20% Compliance
80% Compliance
Population−based Guidelines

Chapter 6

OPTIMAL EXPERT KNOWLEDGE ELICITATION FOR BAYESIAN NETWORK STRUCTURE IDENTIFICATION

6.1 Abstract

Bayesian network (BN) has been a popular tool for gaining mechanistic understanding of variables by revealing how the variables influence each other. It has been found very effective in a few studies in quality control and process monitoring. However, for complex problems where the structure of a BN is unknown, a common approach is to learn the BN structure from observational data. A fundamental bottleneck of this approach is that observational data can only be used to discover part of the influential relationships among variables. To overcome this problem, we propose to combine observational data and expert knowledge. To the best of our knowledge, our approach is the first of its kind that can automate the expert elicitation process and collect the most informative expert knowledge, optimally matched to the observational data, to learn the BN structure.

6.2 Introduction

A Bayesian network (BN) is a graphical model for representing influential relationships among variables. It is also interpreted as a causal model in some applications where some strong assumptions can be imposed to establish the equivalency between the statistical dependency among variables as causality ([167]). No matter whether or not the causality can be derived, BN models have been a popular tool for gaining mechanistic understanding of variables by revealing how the variables influence each other. Its popularity has been evidenced by its wide applications in many fields such as genetics in [168, 169, 170], ecology in [171, 172], social sciences in [173, 174], medical sciences in [175], brain sciences in [176, 177] and manufacturing in [178]. There has also been a growing awareness of the use of BN for a number of quality control and monitoring tasks,

as evidenced by the existing works in [179, 180, 181, 182, 183]. Figure 6.1 actually shows an example provided in [183], where the influential relationships among the variables in the hot forming process can be characterized as a DAG structure. Apparently, knowing the influential relationships among the variables could greatly facilitate the fault diagnosis. On the other hand, knowing the DAG structure actually simplifies the statistical modeling of a joint distribution. For instance, to model the joint distribution of the five variables in Figure 6.1, we only need to model a few conditional distributions since $P(X_1, ..., X_5) = P(X_1)P(X_2)P(X_3|X_1, X_2)P(X_4|X_2)P(X_5|X_3, X_4)$. Thus, using BN to model complex systems will facilitate many decision-making tasks to better manage and improve these systems.

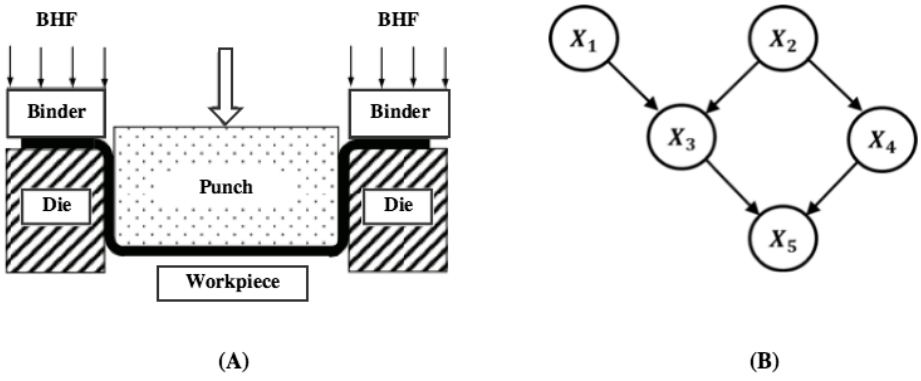


Figure 6.1: (A): 2-D illustration of the hot forming process; (B) the corresponding BN structure, here, X_1 represents the blank holding force; X_2 represents the temperature; X_3 represents the tension in workpiece; X_4 represents the material flow stress; X_5 represents the final dimension of workpiece.

However, learning the DAG structure of variables appears to be a very challenging task. As a BN essentially embodies a joint distribution, statistical estimation approaches have been developed to learn the DAG structure based on observational data, which are commonly assumed to be randomly sampled from the underlying joint distribution. This approach has been extensively studied

in [184, 169, 185, 186, 187, 188]. However, it has been found that the theoretical bottleneck of these methods is that observational data can be used to only discover part of the influential relationships, encoded in the so-called "essential graph" (or "equivalent class"). The essential graph of a BN is a mixed graph that includes both directed arcs and undirected arcs. A deeper reason of this limitation is that, merely from observational data, we could only identify statistical dependency relations between variables. Thus, the DAG structures that imply the same set of dependency relations between variables are not distinguished by observational data alone. For instance, the following three DAGs, $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$, cannot be distinguished by observational data alone, since all of them encode the same independence relations. Thus, to augment the observational data and identify all influential relationships among the variables, a common philosophy is to pursue experimental design strategies for intervention data collection. For instance, considering two variables, X and Y. If intervention can be imposed on X and it turns out that the distribution over Y does not change, while intervention on Y does change the distribution over X, then it implies that the DAG is $Y \rightarrow X$. Motivated by this observation, a line of research works have been spurred to develop optimal experimental design methods to maximize the likelihood of learning the DAG structure with minimized number of interventions. Exemplary works include [189, 190, 191, 192].

In our study, we pursue another line of philosophy to augment the observational data for DAG learning, which is the expert knowledge elicitation. We focus on a particular type of elicitation operation that acquires pairwise comparison between variables, i.e., ask an expert if a variable is likely to be upstream of another variable. This is the most common form of expert knowledge regarding the influential relationships between variables. While pairwise comparison owns the advantages such as ease to implement, on the other hand, apparently, the number of potential pairwise comparison grows exponentially with the number of variables. Thus, we aim to develop a computational framework that can generate an optimal set of operations for expert knowledge elicitation regarding the ordering of the variables. Note that, expert knowledge elicitation has been studied in the BN literature, mostly for parameter learning rather than structure learning. There is a few works such as [193, 194] that studied the use of expert knowledge elicitation for structure learn-

ing, however, these are heuristic procedures that are not scalable, neither automatically optimized. Also, due to their qualitative nature, how they can be optimally integrated with learning algorithms based on observational data is also lacking. To the best of our knowledge, our systematic optimization formulation to automate the expert elicitation process to learn the DAG structure of a BN, in combination of the observational data, is the first of its kind.

Comparing with the approaches that use interventions to perturb the system in order to learn the influential relationships [189, 190, 191, 192], our approach is more cost-effective and can be applied to some applications where intervention is physically hard to conduct. The following application provides such an example that is common in real-world applications, but is outside of the scope of existing methods. For instance, the Key Performance Indicator (KPI) has been a very important concept in business analytics and performance management, drawing increasingly attentions from many corporations to measure and monitor many KPIs of their interest on daily basis if not hourly or minutely. It has been pointed out in the literature such as [195, 196, 197, 198] that the key to analyze the KPIs, and to further convert them into valuable business decision-makings, is to study the influential relationships between the KPIs. In other words, it is important to understand which KPIs drive which KPIs, so management or investment strategies can be better informed and implemented. However, although an abundance of KPI measurements can be obtained, whether or not we could learn this "mechanistic understanding" of the KPIs from observational data is up to debate, and how to learn it is still an open question. On the other hand, expert knowledge has been found very useful to identify the influential relationships among the KPIs in [199]. Apparently, knowledge-based practice has the difficulty of being scaled up. Also, it lacks the flexibility to incorporate the objective information encoded in the observational data. Thus, there is a need to develop a computational framework that can learn the influential relationships among the KPIs by using both the observational data and expert knowledge automatically and cost-effectively elicited to augment the observational data.

Our general approach is to first develop a Bayesian learning framework that can combine the two types of data. This is plausible, since given a specific DAG structure, the likelihood of the observational data can be analytically derived based on the corresponding joint distribution the DAG

encodes ([200, 201, 184, 202, 203]). On the other hand, we further develop another probabilistic framework to model the expert pairwise comparison data. Then, within a Bayesian learning framework, both sources of data can be combined to obtain the probability distribution of the possible BN models. Based on this probability distribution, uncertainty of the ordering of the variables can be evaluated which will provide critical evidence for us to better collect new data via expert knowledge elicitation, that can maximally reduce the uncertainty of our estimation of the ordering of the variables.

The rest of the paper is organized as follows: We will introduce the basic concepts and background of BN, and some BN structure learning algorithms from observational data in Section 6.3. Then, we will present our proposed method in Section 6.4. Specifically, we will first present the Bayesian framework that can learn the probability distribution of the possible BN models using both observational data and expert comparison in Section 6.4.1, and then, introduce how we build a systematical optimization framework to automate expert knowledge elicitation in Section 6.4.2. We will conduct extensive numerical experiments in Section 6.5 to show that the proposed approach outperforms baseline approaches across a number of benchmark BN models and different levels of expert knowledge accuracy. We further implement the proposed method on two real-world applications, one in healthcare and another one in business analytics in Section 6.6. Finally we conclude our work and discuss future directions in Section 6.8.

6.3 Background and Related Works

6.3.1 Learning Bayesian Networks from Data

A Bayesian network (BN) over a set of random variables $X = \{X_1, ..., X_p\}$ is a set (G, θ_G) that represents a distribution over the joint space of X via chain rule: $P(X_1, ..., X_p) = \prod_i P(X_i|U_i)$, where U_i is the parent set of X_i according to the DAG structure of the BN, and θ_G is the corresponding parameter. The DAG structure that encodes the parent-child relation of the variables is commonly denoted as $G = \{V, E\}$, where V denotes for the p nodes (each node is a variable) and E is the edge set. Learning the BN structure from observational data refers to the challenge

to find the optimal DAG structure G that maximizes a certain score which evaluates the goodness-of-fit of the DAG structure to the observed data. For instance, a Bayesian Score was developed in [201] for discrete BNs while another score was developed in [184] for Gaussian BNs. It is commonly assumed that the observational data are randomly sampled from the joint distribution of the variables specified by the BN model, so most of the score functions are developed based on this probabilistic framework ([201, 184, 204]). There are some other information-theoretic score functions developed as well in [205, 206], but still the fact that a BN is a structured representation of a complex joint distribution of the variables lays the foundation of these score functions. With a score function, a search procedure is commonly used to search through the eligible DAG structures to identify the optimal DAG. Figure 6.2 shows such a score-function guided DAG search procedure which underlies most of the score-based BN learning algorithms.There is another line of BN learning algorithms, named as constrained-based algorithms ([207, 208]), which use hypothesis testing methods to identify the dependency structure of the variables. We omit the details of these algorithms here since our method is mostly related to the score-based algorithms.

6.3.2 Observational Equivalence

Despite the success of the BN structure learning algorithms in many applications, it has been found that the theoretical bottleneck of these methods is that observational data can be used to only discover part of the directional relationships, encoded in the so-called "essential graph" (or "equivalent class"). As a matter of fact, the validity of learning the DAG structure of a BN from observational data is established on two assumptions, namely that the BN and its encoded joint distribution obey the faithfulness assumption and causal sufficiency. The faithfulness assumption ensures that all independence relationships revealed by the data are results of the DAG structure. Causal sufficiency means that there are no latent variables. While these assumptions establish the theoretical foundation for learning the DAG structure from observational data, it has also been found that, only the essential graph can be discovered with observational data alone. A brief elaboration of the essential graph involves the concepts *skeleton* and *v-structures*. The skeleton of G is the undirected graph on V where every arc in E has been undirected. A *v-shape* in a graph

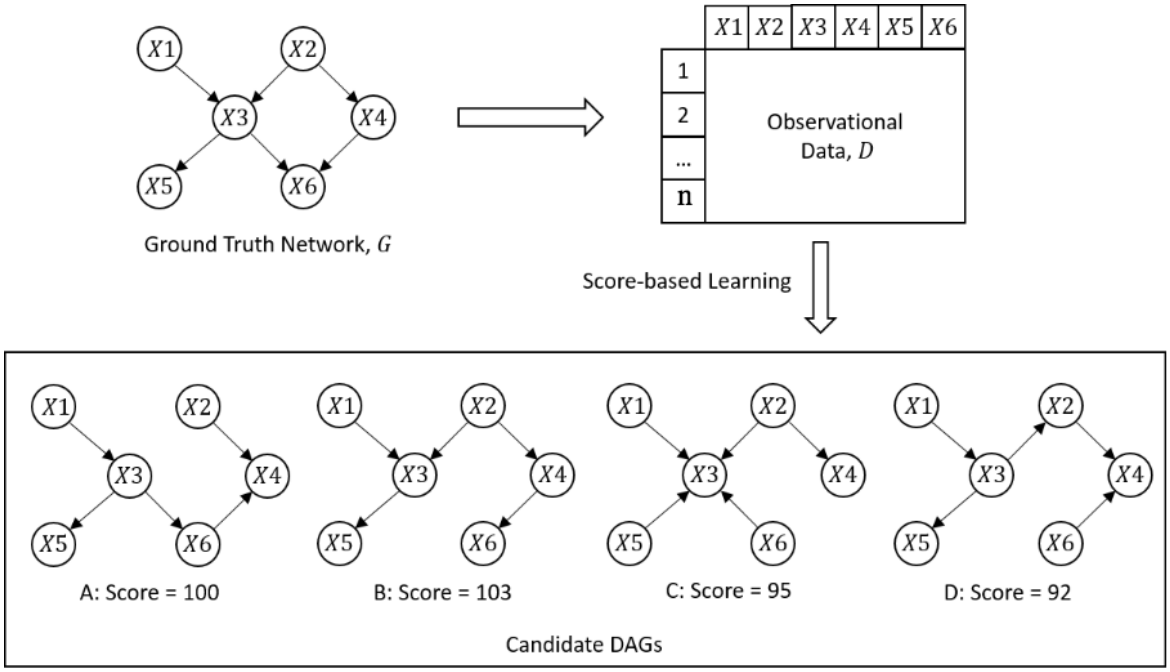


Figure 6.2: The typical score-function guided DAG search procedure which underlies most of the score-based BN learning algorithms.

is an ordered set (a, b, c) of three distinct nodes with exactly two arcs, both directed toward node b. The *v-structures* are the set of *v-shapes*. Then, it is known that the two DAGs, G_1 and G_2, are indistinguishable by observational data alone, if and only if they have the same *skeleton* and the same sets of *v-structures* ([209]). These DAGs that are indistinguishable by observational data are also said to belong to the same Markov equivalence class. Based on this result, the concept, "essential graph", was developed. The essential graph is a mixed graph, where an edge is oriented if and only if it has the same orientation in every DAGs in the equivalence class. The essential graph essentially encodes the maximum set of directed arcs that we could learn from the observational data under the faithfulness assumption and causal sufficiency.

6.3.3 Intervention Data and Expert Elicitation

Motivated by the limitation of learning the structure from data, taking experiments to collect intervention data has been held as a promising means for learning the full DAG structure. The common setting for intervention data collection assumed in the literature is, as illustrated in Section 6.2, to intervene on some variables and observe the influence on other variables. A few studies have been developed to optimize the intervention strategies, i.e., such as to minimize the number of interventions. It has also been studied regarding how many experiments are required to discover G. This line of research was initiated in a series of works by Eberhardt, Glymour, and Scheines ([189, 190, 191, 192]), while most of them focused on single-variable interventions, i.e., each time, an intervention is imposed on one variable only. Eberhardt considered multi-variable interventions to be âĂIJfar more complicatedâĂİ to analyze ([191]). [210] proposed another methodology that is not based on intervention, rather, it is more like a query operation to selectively sample from the BN. Obviously, our proposed expert elicitation method is fundamentally different from this line of works. We have pointed out in Section 6.2 that there are a few works such as [193, 194] that studied the use of expert knowledge elicitation for structure learning, however, these are heuristic procedures that are not scalable for large-scale applications. Also, due to their qualitative nature, the interface with the learning algorithms based on observational data is also lacking, and there is lack of systematic optimization formulation to automate the expert elicitation process.

6.3.4 Existing Works of BN in Quality Literature

It has been found that the BN can be used to improve a number of quality control and monitoring tasks, particularly for improving the root-cause diagnosis and sensor allocation if the underlying process model can be represented by a BN ([179, 180, 181, 182, 183]). A typical motivating example was provided in [183], as shown in Figure 6.1, where the BN was used to further improve the diagnosis procedure of the T^2 chart by characterizing the cascade relationships between the process variables in multivariate processes. Specifically, Figure 6.1 shows a hot forming process where the cascade relationships between the five process variables (X_1, blank holding force; X_2,

temperature; X_3, tension in workpiece; X_4, material flow stress; X_5, final dimension of workpiece) can be represented as a Bayesian Network (BN). If X_1 is out-of-control, its effect will propagate to X_3 and further impact X_5, resulting in out-of-control signals on all these variables. Without the knowledge of this cascade relationship, it would be hard for any root-cause diagnosis procedure to identify X_1 as the root of this chain reaction. More examples of how BN can be useful for quality control and monitoring can be found in [179, 180, 181, 182, 183]. However, a crucial assumption underlying those works is the availability of an accurate BN model to represent the process, which is actually very hard to obtain in many applications.

6.4 Methodology

Our proposed optimal expert elicitation framework, as illustrated in Figure 6.3, can be decomposed into the learning and sensing modules. In the learning module, we develop a Bayesian framework to combine both observational data and expert comparison data to obtain a posterior distribution over ordering of the variables. In the sensing module, the optimization model will identify the most informative new comparisons data that should be collected from the expert, which can maximally reduce the posterior uncertainty of the ordering of the variables.

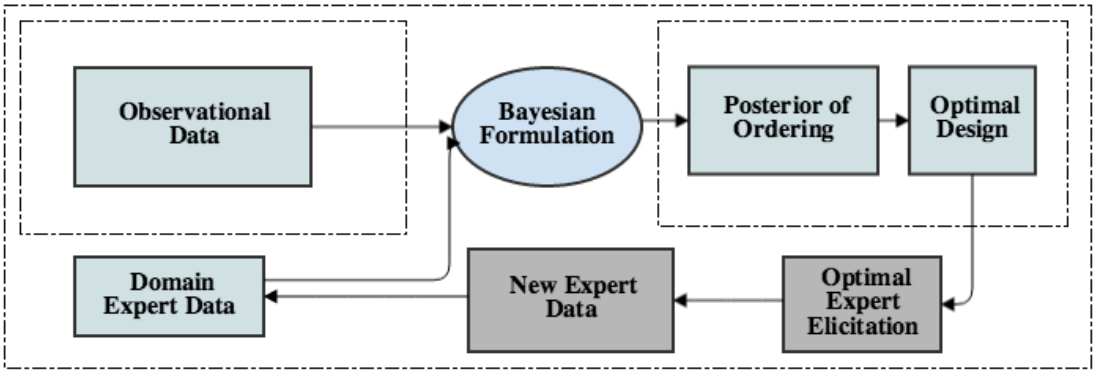


Figure 6.3: Flowchart of the Bayesian learning and sensing framework for optimal expert elicitation.

6.4.1 *A Bayesian Learning Framework for Combining Observational Data and Expert Elicitation Data*

Denote the observational data and expert comparison data as D^{obs} and D^{ex}, respectively. Here, D^{obs} consists of n random samples of the variables $X = \{X_1, ..., X_p\}$, which are i.i.d observations from the joint distribution of the variables represented by the underlying BN. D^{ex} consists of two parts. First, note that, for a networked system with p variables, there are $N = \binom{p}{2}$ possible pairwise comparisons. For each pairwise comparison, for example, considering the k^{th} comparison that involves variables X_i and X_j, the expert might be asked whether or not X_i is upstream of X_j (i.e., denoted as $X_i > X_j$). The expert's response will be denoted as y_k, i.e., a positive y_k indicates that the expert knowledge more tends to support that variable X_i is upstream of variable X_j, while a negative y_k indicates the opposite. Note that the larger the y_k, stronger the knowledge. Thus, the expert data D^{ex} consists of the set of pairwise comparisons that have been queried (denoted as a set S) and the corresponding expert response data (denoted as a vector y). Particularly, considering the sequential nature of our proposed method in data collection, we use subscript to distinguish the data collected in different stages. For instance, following this line, we denote the initial observational data set as D_0^{obs}, and denote the initial expert comparison data as $D_0^{ex} = (S_0, y_0)$. Then, the proposed Bayesian learning and sensing framework will learn the posterior distribution of the ordering of variables, building on which, an optimal expert knowledge elicitation plan could be derived to further collect new expert comparison data which will be denoted as $D_1^{ex} = (S_1, y_1)$. This new data will help update the posterior distribution of the ordering of variables, and the data collection process will continue if needed. We will provide more detailed discussion of this sequential process and the stopping criteria later.

In what follows, we introduce the details of how we could combine both the observational data and expert comparison data to derive the posterior distribution of the ordering of variables. Instead of deriving the posterior distribution of the DAG structure of the BN, here we focus on the learning of the ordering of variables because this is the essential task for BN structure learning, and it has been found that learning the ordering can significantly reduce the computational complexity

since the search space for ordering of variables is much smaller than the search space for DAGs ([169, 211]).

Development of the probabilistic formulation for expert data: It is reasonable to consider that the expert knowledge is always with uncertainty, and different experts may have different accuracy levels. Thus, a probabilistic model is needed to characterize not only the correspondence between the underlying ordering of variables with the expert comparison data, but also the expert's accuracy level. To develop this model, first, we invent a numerical vector to be a surrogate of the ordering of variables. This numerical vector, denoted as $\phi \in R^p$, encodes the same ordering information between variables, since an upstream variable will have larger value in ϕ than its downstream variables. Then, we could establish a probabilistic relationship between ϕ and the observed D^{ex}, i.e., for the k^{th} comparison that involves variables X_i and X_j, we could assume that $y_k \sim N(\phi_i - \phi_j, \sigma^2/w_k)$. This essentially assumes that if the variable X_i is upstream (or downstream) of the variable X_j, we will expect to see positive (or negative) values of y_k. This is consistent with the nature of the expert comparison data we are adopting in this study. Note that, σ^2 encodes the overall accuracy level of the expert knowledge, as more knowledgeable expert will tend to have smaller σ^2. Also, w_k encodes uncertainty in this particular comparison, acting as the local accuracy level of the expert knowledge. In practice, experts could also provide their confidence level, i.e., w_k, along with y_k. Alternatively, when this information is lacking, we could simply assume $w_k = 1$ for all the comparison data.

Following this line, we could further illustrate how we could represent the expert comparison data in a more compact matrix form. First, we invent a Boolean matrix, denoted as B, where $B_{k,j}$ is defined on the set $(k \in |S|, j \in n)$ as:

$$B_{k,j} = \begin{cases} 1 & \text{if } j = head(k) \\ -1 & \text{if } j = tail(k) \\ 0 & \text{otherwise} \end{cases} \tag{6.1}$$

Here, $j = tail(k)$ if the k^{th} pairwise comparison is asked in the form as "$X_i > X_j$"; otherwise,

$j = head(k)$ if "$X_j > X_i$". Then, it can be shown that

$$y \sim N(B\phi, \sigma^2 W^{-1}) \tag{6.2}$$

where W is the diagonal matrix of w. Thus, for the initial expert comparison data, we could derive that $y_0 \sim N(B_0\phi, \sigma^2 W_0^{-1})$ where B_0 is defined on the set S_0.

Characterization of the prior distribution of ϕ based on observational data: In what follows, we propose our method to derive the prior distribution of ϕ by exploiting the information encoded in D^{obs}. It is an analytically intractable task, so we propose a computational procedure that is inspired by existing literature of BN structural learning. Some authors such as [168] suggested to use data perturbation methods such as Bootstrap in [168] to repeatedly bootstrap D_0^{obs} and apply an appropriate BN structure learning method on the perturbed dataset to learn the optimal DAG structure or the ordering of variables. It has been found in [168] that such a bootstrap procedure is especially robust. Thus, we propose the computational procedure that is depicted in Algorithm 3 as a sampling procedure of the orderings of variables based on D^{obs}:

Algorithm 3 Bootstrap the Observational Data

procedure BOOTSTRAP(m) ▷ m is the number of bootstrapping times
  $i = 1$
  while $i \leq m$ **do**
    Re-sample the n instances from data set D_0^{obs} with replacement.
    Apply an appropriate BN structural learning algorithm (i.e., the DAGlearn algorithm in [212] can be used for continuous BNs while the K2 algorithm in [201] can be used for discrete BNs) on the re-sampled dataset to learn $\hat{\phi}_0^i$
    $i = i + 1$
  end while
  return $\{\hat{\phi}_0^i, i = 1, 2, \ldots, m\}$
end procedure

After we draw the samples of ordering of variables $\{\hat{\phi}_0^i, i = 1, 2, \ldots, m\}$ from observational

data, the next step is to translate this knowledge and create the prior distribution of ϕ. Assuming that the prior distribution takes the form as $\phi \sim N(\mu_0, \Lambda_0^{-1})$. Then, the $\{\hat{\phi}_0^i\}$ could be treated as random samples from the prior distribution and can be readily used to estimate the unknown parameters μ_0 and Λ_0^{-1} by maximum likelihood estimation.

The Bayesian learning framework: As a summary, based on the initial expert data D_0^{ex}, it has been known that we could derive that $y_0 \sim N(B_0\phi, \sigma^2 W_0^{-1})$ based on (6.2). And the prior distribution can be obtained from observational data as $\phi \sim N(\mu_0, \Lambda_0^{-1})$. Here, for the ease of derivation, in what follows we rewrite the prior as $\phi \sim N(\mu_0, \sigma^2\Lambda_0^{-1})$, which is numerically equivalent if we change the scale of Λ_0. Then, we could derive the posterior distribution of ϕ with posterior mean μ_1 and posterior variance Λ_1^{-1}, by learning from the Bayesian linear model literature such as [213]. While what we will derive in the following is largely borrowed from the literature, we present critical details in the derivation for completeness of our development. Specifically, the posterior distribution of ϕ can be derived as:

$$\begin{aligned} p(\phi, \sigma^2 \mid y_0, B_0) &\propto p(y_0 \mid B_0, \phi, \sigma^2) p(\phi \mid \sigma^2) p(\sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y_0 - B_0\phi)^T W_0 (y_0 - B_0\phi)\right) \\ &\quad \times (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2}(\phi - \mu_0)^T \Lambda_0 (\phi - \mu_0)\right) \\ &\quad \times (\sigma^2)^{-a_0+1} \exp\left(-\frac{b_0}{\sigma^2}\right) \end{aligned} \tag{6.3}$$

Note that, here, we follow the Bayesian linear model literature ([213]) and use the inverse-Gamma distribution as the prior distribution for σ^2, where a_0 and b_0 are two parameters that can be specified by prior knowledge. We follow the suggestions made in ([213]) to specify a_0 and b_0 based on the non-informative prior principle, since it could provide robust performance for many applications. Also, our main interest is on the learning of the ordering of variables. The exponential parts in

equation (6.3) could be combined as

$$\begin{aligned}
& (y_0 - B_0\phi)^T W_0 (y_0 - B_0\phi) + (\phi - \mu_0)^T \Lambda_0 (\phi - \mu_0) \\
= \;& y_0^T W_0 y_0 - \phi^T B_0^T W_0 y_0 - y_0^T W_0 B_0 \phi + \phi^T B_0^T W_0 B_0 \phi \\
& \quad + \phi^T \Lambda_0 \phi - \phi^T \Lambda_0 \mu_0 - \mu_0^T \Lambda_0 \phi + \mu_0^T \Lambda_0 \mu_0 \\
= \;& (\phi - \mu_1)^T \left(B_0^T W_0 B_0 + \Lambda_0 \right) (\phi - \mu_1) + y_0^T y_0 - \mu_1^T (B_0^T W_0 B_0 + \Lambda_0)\mu_1 + \mu_0^T \Lambda_0 \mu_0 \qquad (6.4)
\end{aligned}$$

where $\mu_1 = (B_0^T W_0 B_0 + \Lambda_0)^{-1}(B_0^T W_0 y_0 + \Lambda_0 \mu_0)$. Then, we could see that, the joint posterior distribution of ϕ and σ^2 is actually a product of a normal distribution and an inverse-gamma distribution,

$$\begin{aligned}
p(\phi, \sigma^2 \mid y_0, B_0) &\propto (\sigma^2)^{-k/2} \exp\left(-\frac{1}{2\sigma^2} (\phi - \mu_1)^T (B_0^T W_0 B_0 + \Lambda_0)(\phi - \mu_1) \right) \\
&\quad \times (\sigma^2)^{\frac{-n-2a_0+2}{2}} \exp\left(-\frac{2b_0 + y_0^T y_0 - \mu_1^T (B_0^T W_0 B_0 + \Lambda_0)\mu_1 + \mu_0^T \Lambda_0 \mu_0}{2\sigma^2} \right) \\
&\propto p(\phi \mid \sigma^2, y_0, B_0) p(\sigma^2 \mid y_0, B_0)
\end{aligned}$$

Essentially this suggests that the posterior distribution of ϕ and σ^2 are $N(\mu_1, \sigma^2 \Lambda_1^{-1})$ and Inv-Gamma, respectively. In summary, by combining the initial observational data D_0^{obs} and D_0^{ex}, we could derive the posterior mean of ϕ as $\mu_1 = (B_0^T W_0 B_0 + \Lambda_0)^{-1}(B_0^T W_0 y_0 + \Lambda_0 \mu_0)$, and the posterior variance as $\Lambda_1^{-1} = (B_0^T W_0 B_0 + \Lambda_0)^{-1}$.

6.4.2 A Semidefinite Programming (SDP) Formulation for Optimal Expert Comparison Elicitation

In this section, we will develop an automated process for expert knowledge elicitation with a systematic optimization formulation. The central question to ask is, given the available data D_0^{obs} and D_0^{ex}, what is the optimal set of new expert comparison data we should further collect? As we have been able to derive the posterior distribution of ϕ based on D_0^{obs} and D_0^{ex} via our method developed in 6.4.1, a natural idea is to identify the candidate expert comparisons that can maximally reduce the uncertainty of the estimated ordering ϕ. To see how this could be formulated, it is worthy of analyzing the structure of the posterior variance of ϕ, the $\Lambda_1^{-1} = (B_0^T W_0 B_0 + \Lambda_0)^{-1}$. Obviously,

it can be derived that, given new expert comparison data $D_1^{ex} = (S_1, y_1)$, the posterior variance will be $\Lambda_2^{-1} = (B_1^T W_1 B_1 + B_0^T W_0 B_0 + \Lambda_0)^{-1}$, where B_1 is defined on the set S_1. In order to more clearly identify the relationship between the candidate expert comparisons with Λ_2^{-1}, we denote a matrix B^* that is defined on the set S^*. Obviously, S^* includes all the candidate comparisons that have not been included in S_0. Then, we could further rewrite $B_1^T W_1 B_1$ as $B_1^T W_1 B_1 = \sum_{k=1}^{|B^*|} x_k a_k^T a_k$, where a_k is the k-th row of B^* and x is a Boolean vector $\in \{0,1\}^{|B^*|}$, while $x_k = 1$ if the k^{th} comparison is included in S_1. With this, we have almost formulated the optimization problem for new expert data elicitation, i.e., the goal is to identify the optimal solution of the decision variables x that can maximize the decrease of the posterior variance $(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \Lambda_0)^{-1}$, under the constraint that only a certain number of new expert comparisons, i.e., denoted as $\xi \in [0, |B^*|]$, can be elicited.

This resembles many of the optimal design problems that have been discussed in the context of linear models. While many existing optimal design methods assume general structure for the design matrix and thus are limited by optimization options, we recognize that in our problem there is a special structure that can be exploited, which will lead to more powerful optimization formulations such as the Semi-definite Programming (SDP) formulation. On the other hand, while a few optimality criterion have been developed in optimal design, here, we propose to study the E-optimal design criteria first due to its robust nature and the subsequent computational benefit on optimization. Particularly, the E-optimal design criteria proposes to identify the subset of "design points" (in our case, the candidate comparisons) that can maximize the smallest nonzero eigenvalue of the information matrix, i.e., the information matrix is the inverse of the variance matrix, which is $(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \Lambda_0)$ in our case. This leads to the following optimization framework:

$$
\begin{aligned}
\max_{x} \quad & \lambda_1 (B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l) \\
\text{subject to} \quad & \mathbf{1}^T x \leq \xi \\
& x \in \{0,1\}^{|B^*|}
\end{aligned}
$$

where $\lambda_1(A)$ denotes the smallest nonzero eigenvalue of matrix A. Since this problem is difficult to solve exactly, we propose to replace the Boolean constraint $x \in \{0,1\}^{|B^*|}$ by a relaxation, i.e., $x \in [0,1]^{|B^*|}$. Also the constraint $\mathbf{1}^T x \leq \xi$ is binding in Formula (5) as $\xi \leq |B^*|$, since $x \leq 1$, the optimal value given by $\xi > |B^*|$ is the same as that of $\xi = |B^*|$. Therefore we have the following relaxation form:

$$\begin{aligned} \max_{x} \quad & \lambda_1(B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l) \\ \text{subject to} \quad & \mathbf{1}^T x = \xi \\ & 0 \leq x \leq 1 \end{aligned} \tag{6.5}$$

Obviously, the optimal value from this relaxation form is an upper bound on the optimal value of the original form. Such a relaxation is a convex optimization problem since the constraints are linear functions of x, and we only need to show that the objective function is a concave function of x, which is shown in the following lemma.

Lemma 6 *The optimization in* (6.5) *is a convex optimization problem.*

Proof To show the convexity of (6.5), one needs to show that the constraint defines a convex set and the objective is a concave (convex) function if it is a maximization (minimization) problem. The linear constraint in (6.5) defines a polyhedron which is convex apparently. Therefore, we only need to prove the objective is a concave function, i.e., to prove that $\lambda_1(B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l)$ is a concave function. Denote $B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l$ as $L(x)$. If we could prove that $-\lambda_1(L)$ is a convex function of L, then it is known that $-\lambda_1(L(x))$ is convex, since $L(x)$ is a linear transformation in terms of x which does not affect the convexity (or concavity) of the objective function. Thus, to show that $-\lambda_1(L(x))$ is convex, first, note that we could write $-\lambda_1(L(x))$ using the definition for the minimal eigenvalue of a positive semi-definite matrix as

$$\begin{aligned} -\lambda_1(L(x)) &= -\inf_{\|z\|=1} z^T L(x) z \\ &= -\inf_{\|z\|=1} z^T L(x) z \\ &= \sup_{\|z\|=1} \langle -zz^T,\ L(x)\rangle. \end{aligned}$$

To prove a function is convex, it is equivalent to show that its epigraph is a convex set. The epigraph of $-\lambda_1(L(x))$ is

$$\begin{aligned} \text{epigraph}(-\lambda_1(L(x))) &= \text{epigraph}(\sup_{\|z\|=1} \langle -zz^T,\ L(x)\rangle) \\ &= \cap_{\|z\|=1}\ \text{epigraph}(\langle -zz^T, L(x)\rangle). \end{aligned}$$

Since $\langle -zz^T, L(x)\rangle$ is a linear function, its epigraph is a half space above the hyperplane defined by the linear function. So it is a convex set. As we know, the intersection of convex sets is still a convex set. Therefore, we have that epigraph$(-\lambda_1(L(x)))$ is a convex set, which proves the convexity of the function $-\lambda_1(L(x))$. It completes the proof.

We could further show that the convex relaxation above leads to a semidefinite programming (SDP) problem. To see that, note that the objective in optimization in (6.5) can be equivalently stated as maximizing a new variable s where it is required that $s \leq \lambda_1(B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l)$. This could be further restated as $B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l - sI$ should be positive semi-definite. This leads to the following SDP formulation:

$$\begin{aligned} \max \quad & s \\ \text{subject to} \quad & sI \preceq (B_0^T W_0 B_0 + \Lambda_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l) \\ & \mathbf{1}^T x = \xi \\ & 0 \leq x \leq 1 \end{aligned} \tag{6.6}$$

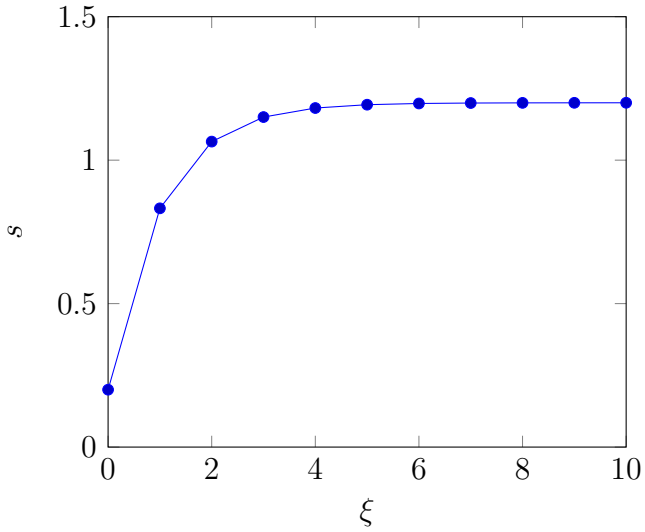


Figure 6.4: Diminishing marginal effectiveness of the objective function in optimization in (6.6)

This SDP can be solved using any standard SDP solver. In our experiments, we use the package "cvx" ([65]) to solve our problem which is shown to be fairly effective.

We could further show a useful property of the proposed formulation that indicates there is diminishing marginal effectiveness of the number of expert comparisons ξ on the objective function. This property is shown in Figure 6.4 that is according to our empirical study of the proposed SDP formulation.

From this Figure 6.4, it is clear that, due to the concave shape of the objective function on ξ, the objective function has a much greater increase at the beginning than later on. This indicates that the first few expert comparisons could lead to a much greater marginal effect in reduction of posterior variance. And since this effect will gradually diminish, it indicates that probably only a few expert comparisons are needed for accurate learning of ordering of variables. While this is an empirical observation, to show that, denote $f(x) = -\lambda_1(L(x))$ which has been known to be a convex function of x by Lemma 6. Denote $\omega(\xi) = \inf\{f(x)|\mathbf{1}^T x \leq \xi, 0 \leq x \leq 1\}$, which can also be shown to be convex (by learning the proof from page 216 in [214]).

Lemma 7 *The function $\omega(\xi)$ is convex.*

Proof To show the function $\omega(\xi)$ is convex. First, we could see that the domain $\Omega = \{\xi | \exists x, \mathbf{1}^T x \leq \xi, 0 \leq x \leq 1\}$ is a convex set. Then let $\xi_1, \xi_2 \in \Omega$, then we could find two vectors x_1, x_2 such that $\mathbf{1}^T x_1 = \xi_1, \mathbf{1}^T x_2 = \xi_2$ and $0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1$. By the definition of convexity of a function, for any $\alpha \in (0,1)$, $\alpha\xi_1 + (1-\alpha)\xi_2 \in \Omega$, we need to show that

$$\omega(\alpha\xi_1 + (1-\alpha)\xi_2) \leq \alpha\omega(\xi_1) + (1-\alpha)\omega(\xi_2)$$

This can be shown by the following induction:

$$\begin{aligned}
\omega(\alpha\xi_1 + (1-\alpha)\xi_2) &= \inf\{f(x) | \mathbf{1}^T x = \alpha\xi_1 + (1-\alpha)\xi_2, 0 \leq x \leq 1\} \\
&\leq \inf\{f(x = \alpha x_1 + (1-\alpha)x_2) | \mathbf{1}^T x = \alpha\xi_1 + (1-\alpha)\xi_2, 0 \leq x \leq 1\} \\
&\leq \alpha \inf\{f(x_1) | \mathbf{1}^T x_1 = \xi_1, 0 \leq x_1 \leq 1\} \\
&\quad + (1-\alpha) \inf\{f(x_2) | \mathbf{1}^T x_2 = \xi_2, 0 \leq x_2 \leq 1\} \\
&= \alpha\omega(\xi_1) + (1-\alpha)\omega(\xi_2)
\end{aligned}$$

Therefore we proved the convexity of $\omega(\xi)$, and we used the convexity of $f(x)$ in the last inequality.

On the other hand, it is easy to show that $-\omega(\xi)$ is monotonically increasing with the value of ξ, and is bounded on the domain of ξ. Since it is not likely that $-\omega(\xi)$ is linear due to its complicated functional form, only a function that takes the shape as shown in Figure 6.4 with diminishing marginal effectiveness can satisfy all these properties simultaneously.

6.4.3 Extension to Applications without Observational Data

Note that our proposed method can be easily extended to applications where observational data is not available. One approach is to assume non-informative prior distribution for ϕ, i.e., by assuming that $\Lambda_0^{-1} = \sigma_0^2 I$ where σ_0^2 is a very large number and I is the identify matrix. Then, we could still apply our method to these applications based on the Bayesian learning framework developed in Section 6.4.1 and the optimal expert elicitation formulation developed in Section 6.4.2. That said,

there is still a particular problem that we need to address. Note that the posterior variance of the ordering of variables is $(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I)^{-1}$. It can actually be shown that, the smallest eigenvalue of the corresponding information matrix, $(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I)$, is σ_0^{-2}.

Lemma 8 *The smallest eigenvalue of the matrix,* $(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I)$*, is* σ_0^{-2}*, and the corresponding eigenvector is* $\mathbf{1}$*.*

Proof It can be seen that $\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0$ is a singular positive semi-definite symmetric matrix. To see that, for any vector z, we could show that $z^T(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0)z = \sum_{k=1}^{|B^*|}(x_k + w_k)(a_k z)^2 \geq 0$, since $\forall k, x_k \geq 0, w_k \geq 0$. Since $B_0\mathbf{1} = \mathbf{0}$, the vector $\mathbf{1}$ is in the null space of $\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0$ and the corresponding eigenvalue is 0. On the other hand, it is known that for the matrix $\sigma_0^{-2} I$, it has σ_0^{-2} as an eigenvalue and $\mathbf{1}$ as an eigenvector. Then we could draw the bounds of the smallest eigenvalue of $\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I$ as $\lambda_1(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0) + \lambda_1(\sigma_0^{-2} I) \leq \lambda_1(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I) \leq \lambda_1(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0) + \lambda_p(\sigma_0^{-2} I)$, which is based on the Theorem 8.1.5 in [215]. This essentially imply that $\sigma_0^{-2} \leq \lambda_1(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I) \leq \sigma_0^{-2}$. Therefore, the smallest eigenvalue of the matrix $\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I$ is σ_0^{-2} with $\mathbf{1}$ as an eigenvector.

As the Lemma 8 indicates, we could not maximize the smallest eigenvalue of the information matrix. Thus, we propose to maximize the second smallest eigenvalue of $(\sum_{k=1}^{|B^*|} x_k a_k^T a_k + B_0^T W_0 B_0 + \sigma_0^{-2} I)$. Particularly, this will result in the following SDP formulation for the expert knowledge elicitation process

$$
\begin{aligned}
\max \quad & s \\
\text{subject to} \quad & s(I - \mathbf{1}\mathbf{1}^T/n) \preceq (B_0^T W_0 B_0 + \sum_{l=1}^{|B^*|} x_l a_l^T a_l + \sigma_0^{-2} I) \\
& \mathbf{1}^T x = \xi \\
& 0 \leq x \leq 1
\end{aligned}
$$

6.4.4 A Validation Procedure for the Utility of the Expert Data

Theoretically, the expert comparison data is useful to help the learning of influential relationships between variables. However, in practice, it is reasonable to have the concern that whether or not the expert comparison data could be trusted, since expert data is subjective and varies among experts. It is worthy of highlighting that our proposed method actually takes into consideration of this concern. We use a probabilistic framework to characterize the relationship between the expert data with the underlying ordering of variables, and the parameters, σ^2 and W, actually evaluate how accurate the expert data could be. For extreme cases where the expert is providing random guess information, our Bayesian learning framework will be able to estimate a large value of σ^2, and therefore, automatically assign more weight on the observational data for the final learning result of the BN. Besides the use of the parameters σ^2 and W to evaluate the utility of expert data, here, we also propose another approach, as a pseudo hypothesis testing procedure as a validation procedure. The basic rationale is that, if the expert data could be trusted, then it should be significantly different from the random guess. Thus, our pseudo hypothesis testing procedure consists of three steps: 1) In the first step, a certain number of DAG structures will be randomly generated, and each of the DAG structures will be used to obtain a likelihood value on the observational data; 2) In the second step, we will sample the same number of DAG structures from the distribution of the ordering of the variables, that is learned solely from expert data (this can be done in our proposed Bayesian framework as a special case where no observational data is used), and again, each of these DAG structures will be used to obtain a likelihood value on the observational data; 3) Then, in the final step, we could compare the two distributions of the likelihood values and see if they are significantly different, i.e., by the use of a t-test. It is expected to see that there is significant difference between the expert knowledge with random guess which could justify the use of expert knowledge. We will demonstrate the use of this validation procedure in our numerical studies.

6.5 Experiments on Simulated Data

6.5.1 Methodology

In this section, we conduct experiments to evaluate the performance of the proposed method using a range of benchmark BN models. Specifically, we select seven benchmark networks from the Bayesian Network Repository (BNR). These BNs have been widely used in the BN literature for performance evaluation since it provides a high quality representation of the diverse BN structures that we may encounter in real-world applications. As shown in Table 6.1, this cohort of BNs has the network sizes ranging from small to moderately large.

Table 6.1: Facts of Benchmark Networks from Bayesian Network Repository (BNR)

Data	**# Nodes**	**# Arc**	**Avg Degree**
Earthquake	5	4	1.6
Asia	8	8	2.00
Child	20	25	1.25
Insurance	27	52	3.85
Mildew	35	46	2.63
Alarm	37	46	2.49
Barley	48	84	3.50

We compare our proposed method with the random sampling method that elicits expert knowledge on a random basis. To implement the random sampling method, we could follow the framework as shown in Figure 6.3 and just replace the use of the SDP method with the random sampling method.

In each simulation study, first, we select a BN network from Table 6.1, then, randomly generate parameters by following conventions that have been defined in the BN structure learning literature

such as [212, 216, 188]. As a BN model is essentially a joint distribution of the variables, we simulate observation data with a sample size 1000, which is our D_0^{obs}. We then apply a benchmark BN learning method on D_0^{obs} to learn the orderings of variables, i.e., we use the L1MB method, proposed in [212], that can be implemented in the Matlab DAGLearn Toolbox (`http://www.cs.ubc.ca/~murphyk/Software/DAGlearn/index.html`). Expert knowledge will be generated by the probabilistic model as mentioned in Section 3.1. Different accuracy levels of expert knowledge (i.e., as encoded in the parameter σ^2) will be used in our study.

We use two metrics for performance evaluation and comparison. The first metric is the correlation between the learned ordering of variables with the true ordering to evaluate the learning accuracy of both methods. The second metric is to compare the variance reduction of both methods. Specifically, we define the variance reduction ratio as follows:

$$\text{Variance Reduction Ratio} = \frac{Var(RS)_i - Var(SDP)_i}{Var(RS)_{i-1} - Var(RS)_i}$$

where $(\cdot)_i$ refers to results from the i-th iteration during the sequential learning process, and $var(SDP)$ and $var(RS)$ indicate the posterior variance of ϕ obtained by the proposed method and random selection method, respectively. The variance reduction ratio essentially evaluates how much extra variance reduction the SDP method could provide on top of the random selection method. This metric facilitates the performance comparison since it is scale-invariant, and the value could be interpreted across different settings of other parameters. As a comparison, variance itself varies from case to case, which is hence not a good metric for performance evaluation and comparison.

6.5.2 Evaluation of Estimation Accuracy of the Ordering of Variables

In this section, we investigate the learning accuracy of our proposed Bayesian learning and sensing framework. The experimental results are presented in Figure 6.5, while all the 7 BN networks are investigated. A sequential procedure for expert knowledge elicitation is used, that has 10 iterations in total, while in each iteration, 4 comparison tasks are queried. Also, the accuracy level of the expert knowledge is set to be $\sigma^2 = 2$. Note that, experimental results for other settings

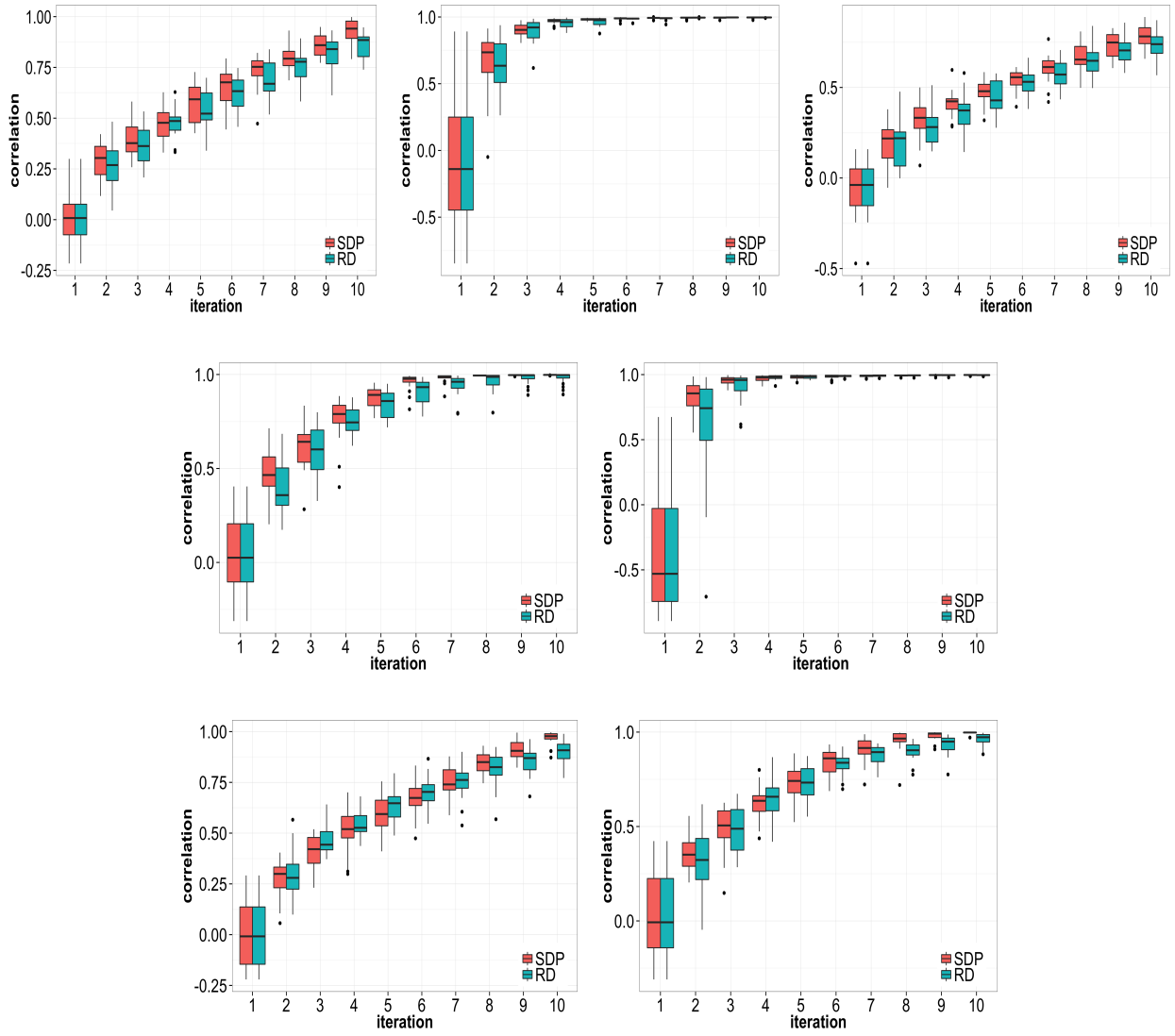


Figure 6.5: Evaluation of the proposed Bayesian learning and sensing framework, with either the SDP (red) method or the random selection method (blue), for estimating the ordering of the variables. Each figure corresponds to a network, which are earthquake, asia, child, insurance, mildew, alarm, barley, from left to right and top to down. The accuracy level of the expert knowledge is set to be $\sigma^2 = 2$.

of these parameters lead to similar conclusion, so we only present the results as shown in Figure 6.5. Since for each simulated scenario, we repeat the experiments 100 times, so boxplot is used to present the overall results. The results clearly show that: 1) the proposed framework, either being integrated with the SDP method or the random selection method, could lead to effective learning of the underlying ordering of the variables. This indicates that the proposed Bayesian learning and sensing framework is effective in general. 2) The SDP method is better than the random selection method as it can lead to quicker detection of the underlying ordering of the variables. Also, the performance of the SDP method is more robust as it often generates results with tighter error bounds across the 7 BN networks.

6.5.3 Evaluation of Variance Reduction Performance

In this section, we further evaluate the efficiency of the proposed Bayesian learning and sensing framework in generating the candidate expert comparison tasks via the SDP formulation. Using the same setting of the parameters (such as $\sigma^2 = 2$ and the number of expert comparison tasks queried in each iteration is set to be 4), the experimental results are shown in Figure 6.6. Similar conclusions can be drawn as: 1) the proposed framework, either being integrated with the SDP method or the random selection method, could lead to effective reduction of the estimation variance. 2) The SDP method is better than the random selection method as it can lead to quicker reduction in the estimation variance of the orderings. For small network, we could observe that the estimation variance of the ordering by the proposed method quickly converges to very low (~ 0), while the estimation variance of the ordering by the random selection method still stays at a significant nonzero level. For networks with larger sizes, such as the four networks in the middle (i.e. the "child","insurance","mildew" and "alarm"), the estimation variance of the ordering by the proposed method usually approaches zero after $6 \sim 10$ iterations, while the random selection method needs more iterations.

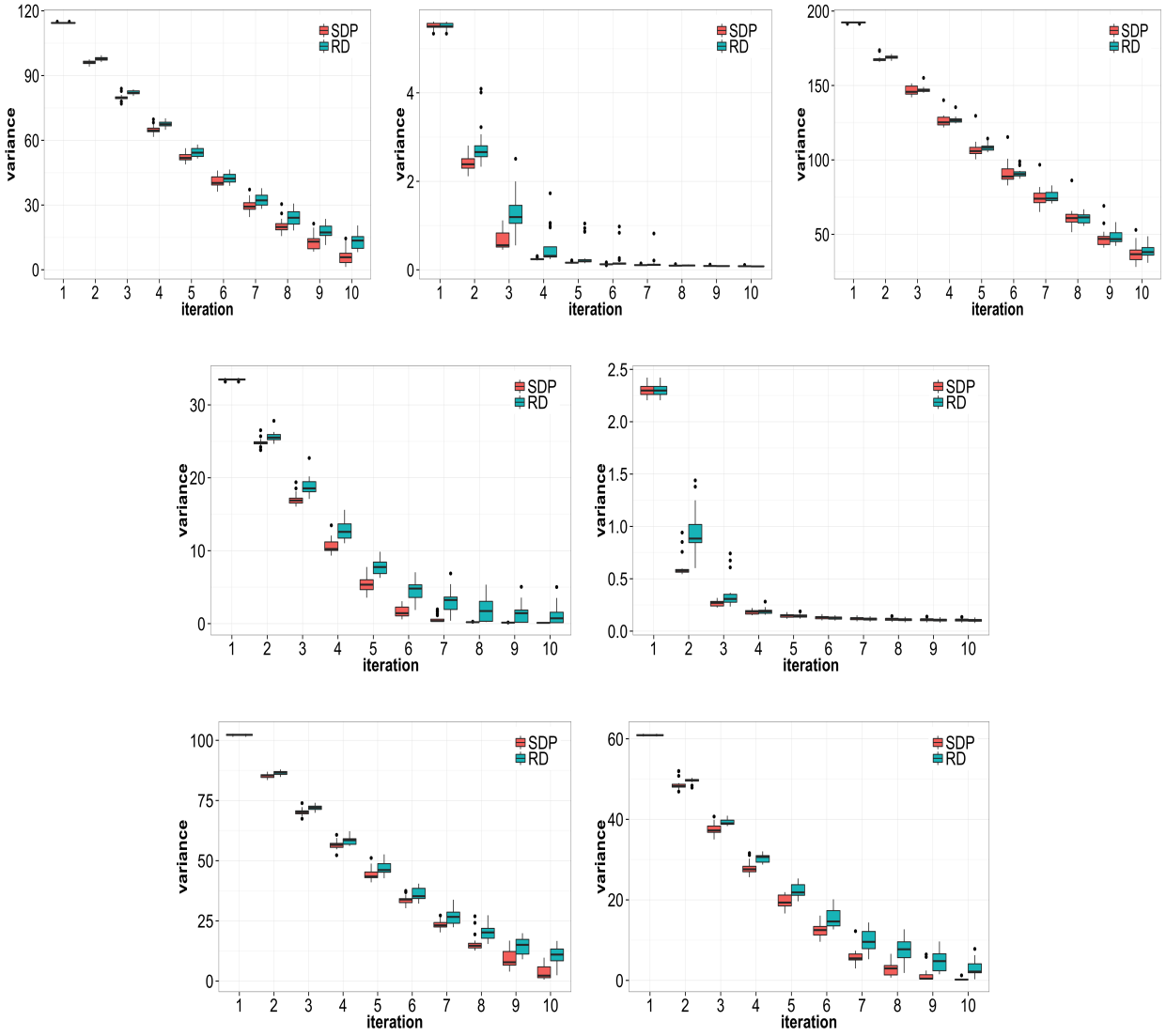


Figure 6.6: Evaluation of the proposed Bayesian learning and sensing framework, with either the SDP (red) method or the random selection method (blue), for reducing the variance of estimation of the ordering. Each figure corresponds to a network, which are earthquake, asia, child, insurance, mildew, alarm, barley, from left to right and top to down. The accuracy level of the expert knowledge is set to be $\sigma^2 = 2$.

Dataset	Scenario under $\sigma^2 = 1$	**Variance Reduction Ratio** $\pm s.d.$		
		@3th Iteration	**@6th Iteration**	**@9th Iteration**
Alarm	$nQuery = 1$	0.14 ± 0.10	0.15 ± 0.10	0.16 ± 0.11
	$nQuery = 2$	0.13 ± 0.11	0.11 ± 0.19	0.08 ± 0.20
	$nQuery = 4$	0.07 ± 0.08	0.08 ± 0.09	0.23 ± 0.32
Asia	$nQuery = 1$	0.29 ± 0.47	0.41 ± 0.63	1.92 ± 3.35
	$nQuery = 2$	0.20 ± 0.19	0.23 ± 0.40	0.50 ± 1.70
	$nQuery = 4$	0.11 ± 0.13	0.01 ± 0.02	-0.08 ± 0.09
Child	$nQuery = 1$	0.12 ± 0.09	0.16 ± 0.12	0.31 ± 0.53
	$nQuery = 2$	0.11 ± 0.09	0.19 ± 0.19	0.51 ± 0.50
	$nQuery = 4$	0.06 ± 0.07	0.21 ± 0.12	0.99 ± 2.06
Insurance	$nQuery = 1$	0.14 ± 0.09	0.15 ± 0.09	0.23 ± 0.18
	$nQuery = 2$	0.13 ± 0.09	0.07 ± 0.09	0.21 ± 0.26
	$nQuery = 4$	0.11 ± 0.07	0.19 ± 0.10	0.43 ± 0.38
Mildew	$nQuery = 1$	0.11 ± 0.11	0.16 ± 0.11	0.24 ± 0.23
	$nQuery = 2$	0.10 ± 0.06	0.09 ± 0.10	0.10 ± 0.15
	$nQuery = 4$	0.08 ± 0.07	0.08 ± 0.12	0.39 ± 0.29
Earthquake	$nQuery = 1$	0.11 ± 0.11	0.38 ± 0.45	0.64 ± 2.54
	$nQuery = 2$	0.11 ± 0.16	0.08 ± 0.18	-0.02 ± 0.09
	$nQuery = 4$	0.02 ± 0.02	0.03 ± 0.03	-0.19 ± 0.10
Barley	$nQuery = 1$	0.09 ± 0.10	0.12 ± 0.09	0.18 ± 0.19
	$nQuery = 2$	0.06 ± 0.06	0.03 ± 0.08	0.01 ± 0.11
	$nQuery = 4$	0.05 ± 0.07	0.02 ± 0.07	0.11 ± 0.16

Table 6.2: Summary of experimental results when $\sigma^2 = 1$.

Dataset	Scenario under $\sigma^2 = 2$	**Variance Reduction Ratio** $\pm s.d.$		
		@3th Iteration	**@6th Iteration**	**@9th Iteration**
Alarm	$nQuery = 1$	0.15 ± 0.10	0.15 ± 0.11	0.14 ± 0.12
	$nQuery = 2$	0.09 ± 0.11	0.08 ± 0.10	0.03 ± 0.15
	$nQuery = 4$	0.07 ± 0.06	0.04 ± 0.08	0.21 ± 0.24
Asia	$nQuery = 1$	0.18 ± 0.21	0.29 ± 0.26	1.26 ± 2.99
	$nQuery = 2$	0.20 ± 0.14	0.25 ± 0.25	0.06 ± 0.04
	$nQuery = 4$	0.16 ± 0.17	0.10 ± 0.27	-0.08 ± 0.12
Child	$nQuery = 1$	0.22 ± 0.30	0.25 ± 0.26	0.17 ± 0.23
	$nQuery = 2$	0.14 ± 0.12	0.15 ± 0.14	0.38 ± 0.35
	$nQuery = 4$	0.13 ± 0.13	0.22 ± 0.12	0.53 ± 0.68
Insurance	$nQuery = 1$	0.14 ± 0.12	0.19 ± 0.22	0.27 ± 0.31
	$nQuery = 2$	0.14 ± 0.13	0.16 ±0.18	0.27 ± 0.31
	$nQuery = 4$	0.08 ±0.08	0.13 ±0.12	0.41 ± 0.40
Mildew	$nQuery = 1$	0.12 ± 0.09	0.14 ± 0.08	0.22 ± 0.20
	$nQuery = 2$	0.08± 0.07	0.12 ±0.11	0.19 ± 0.21
	$nQuery = 4$	0.06 ± 0.05	0.07 ± 0.09	0.25 ± 0.22
Earthquake	$nQuery = 1$	0.22 ± 0.26	0.39 ± 0.64	0.46 ± 1.51
	$nQuery = 2$	0.14 ± 0.18	0.04 ± 0.06	-0.04 ± 0.14
	$nQuery = 4$	0.05 ± 0.09	-0.02 ± 0.03	ïij■0.16 ± 0.10
Barley	$nQuery = 1$	0.12 ± 0.09	0.16 ±0.10	0.18 ± 0.12
	$nQuery = 2$	0.06 ± 0.12	0.05 ± 0.09	-0.03 ± 0.15
	$nQuery = 4$	0.02 ± 0.07	0.01 ± 0.14	0.02 ± 0.18

Table 6.3: Summary of experimental results when $\sigma^2 = 2$.

Dataset	Scenario under $\sigma^2 = 4$	**Variance Reduction Ratio** $\pm s.d.$		
		@3th Iteration	**@6th Iteration**	**@9th Iteration**
Alarm	$nQuery = 1$	0.12 $\pm$ 0.10	0.21 $\pm$ 0.19	0.21 $\pm$ 0.12
	$nQuery = 2$	0.12 $\pm$ 0.14	0.05 $\pm$ 0.09	0.03 $\pm$ 0.21
	$nQuery = 4$	0.05 $\pm$ 0.08	0.10 $\pm$ 0.09	0.20 $\pm$ 0.27
Asia	$nQuery = 1$	0.16 $\pm$ 0.11	0.47 $\pm$ 0.54	1.14 $\pm$ 1.56
	$nQuery = 2$	0.19 $\pm$ 0.14	0.16 $\pm$ 0.24	0.40 $\pm$ 1.20
	$nQuery = 4$	0.15 $\pm$ 0.09	0.05 $\pm$ 0.16	-0.05 $\pm$0.10
Child	$nQuery = 1$	0.14 $\pm$ 0.11	0.20 $\pm$ 0.15	0.23 $\pm$ 0.26
	$nQuery = 2$	0.12 $\pm$ 0.07	0.24 $\pm$ 0.20	0.74 $\pm$ 1.10
	$nQuery = 4$	0.10 $\pm$ 0.11	0.20 $\pm$ 0.18	0.28 $\pm$ 0.45
Insurance	$nQuery = 1$	0.16 $\pm$ 0.14	0.21 $\pm$ 0.23	0.21 $\pm$ 0.18
	$nQuery = 2$	0.17 $\pm$ 0.16	0.12 $\pm$ 0.09	0.22 $\pm$ 0.19
	$nQuery = 4$	0.07 $\pm$ 0.05	0.12 $\pm$ 0.14	0.52 $\pm$ 0.39
Mildew	$nQuery = 1$	0.13 $\pm$ 0.08	0.13 $\pm$ 0.10	0.15 $\pm$ 0.15
	$nQuery = 2$	0.10 $\pm$ 0.10	0.11 $\pm$ 0.13	0.11 $\pm$ 0.21
	$nQuery = 4$	0.10 $\pm$ 0.06	0.06 $\pm$ 0.09	0.21 $\pm$ 0.25
Earthquake	$nQuery = 1$	0.21 $\pm$ 0.23	0.38 $\pm$ 0.40	1.23 $\pm$ 3.55
	$nQuery = 2$	0.11 $\pm$ 0.15	0.31 $\pm$ 0.91	-0.02 $\pm$ 0.13
	$nQuery = 4$	0.03 $\pm$ 0.07	-0.03 $\pm$ 0.03	-0.16 $\pm$ 0.09
Barley	$nQuery = 1$	0.11 $\pm$ 0.09	0.14 $\pm$ 0.11	0.16 $\pm$ 0.11
	$nQuery = 2$	0.09 $\pm$ 0.06	0.07 $\pm$ 0.11	0.05 $\pm$ 0.16
	$nQuery = 4$	0.02 $\pm$ 0.07	0.01 $\pm$ 0.07	0.01 $\pm$ 0.19

Table 6.4: Summary of experimental results when $\sigma^2 = 4$.

6.5.4 The Relationships between the Performance of the Proposed Method with Sample Sizes and Accuracy Levels of Expert Knowledge

In this section, we aim to provide a comprehensive evaluation of the relationships between the performance of the proposed method with sample sizes (i.e., $nQuery = 1, 2, 4$) and accuracy levels of expert knowledge (i.e., $\sigma^2 = 1$, 2 and 4). Here, $nQuery$ denotes the number of expert comparison tasks we inquiry during each iteration of the sequential expert knowledge solicitation process. Again, for each simulation scenario, we repeat the experiments 100 times and report both the average performance metric and its variance. Results are shown in Table 6.2,6.3 and 6.4, for $\sigma^2 = 1$, $\sigma^2 = 2$, and $\sigma^2 = 4$, respectively. Recall that, the metric, Variance Reduction Ratio, effectively evaluates how much extra variance reduction the SDP method could provide on top of the variance reduction performance by the random selection method. Overall, we could observe that, the SDP method is superior than the random selection method. Also, the proposed method is still effective even when the expert knowledge is fuzzy, i.e., when $\sigma^2 = 4$, as shown in Table 6.4. Note that, we further conduct simulation studies for evaluating our proposed method in Section 3.3 where no observational data is given. Due to page limit, we present the results in the Supplementary file.

6.6 Experiment on Real-World Applications

In what follows, we present our experimental results on two real-world applications, one is to learn the influential relationships among some critical KPIs for better understanding of human resource management in some manufacturing companies, and another one is to model the cascade of hypermetabolism reduction events for better understanding and staging of the Alzheimer's Disease (AD).

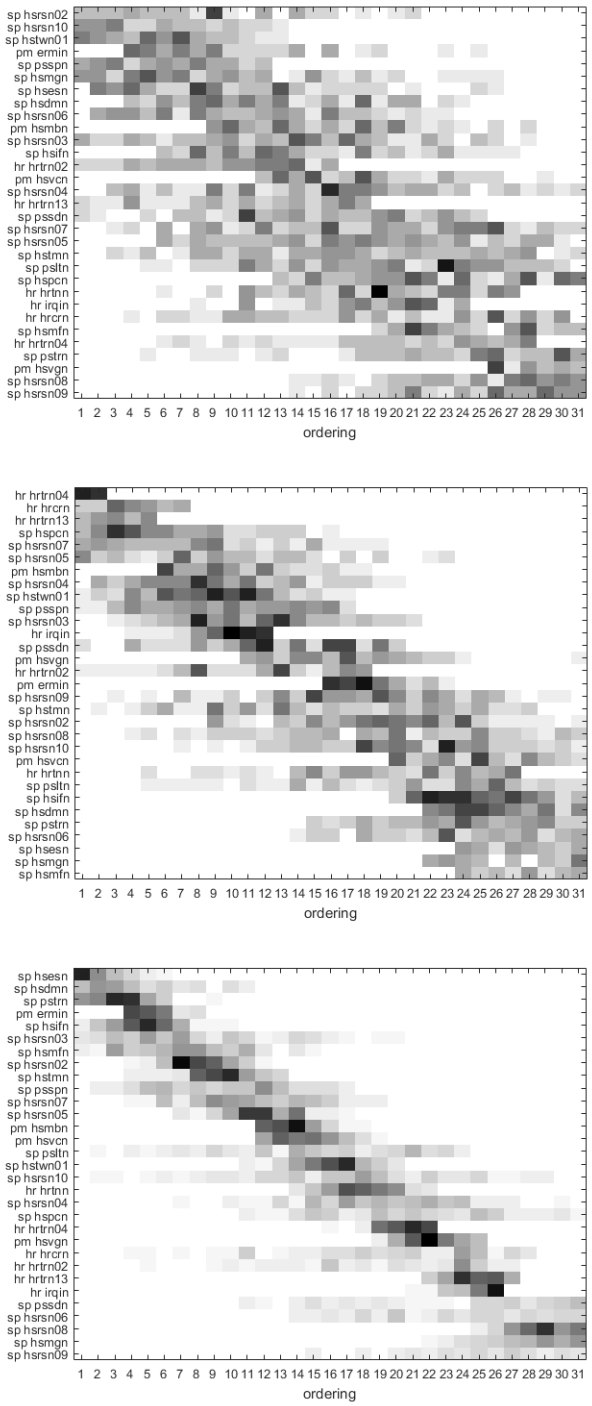


Figure 6.7: Uncertainty of ordering of the KPIs when only observational data is used (top), observation data and 10 expert comparisons are used (middle), observation data and 20 expert comparisons are used (bottom), respectively. Note that, the rows correspond to the KPIs while the numbers in the x-axis represent the ordering of the variables.

6.6.1 *Identification of Influential Relationships of Key Performance Indicators (KPI) for Human Resource Management*

There has been considerate interests in both academia and industry to develop methods to analyze the KPIs measurements, in order to gain a mechanistic understanding of the KPIs. It is because that a mechanistic understanding that can identify which KPIs drive which KPIs will be of great value for facilitating decision-makings. Thus, to demonstrate the utility of our method for KPI data analysis, we use a database that was collected from 31 KPIs of Human Resource Management from 197 manufacturing companies. Those KPIs cover a range of critical quality dimensions of the human resource management practices, which include the Employee Trait such as technical and problem solving skills, the reward policy in the company, the salaried or hourly employee turnover rates, the supply lead time and stability of demand, to name a few. There has been some prior knowledge regarding the influential relationships among some of the KPIs as mentioned in [217, 218], but a systematical study that can effectively combine both the observational data with expert knowledge has not been done yet to the best of our knowledge. Thus, we implement our proposed Bayesian learning and sensing framework on this dataset and interact with our expert (one of our co-authors) who is knowledgeable on this dataset to obtain the inquired expert comparison data. Note that, here, we limit our total number of expert comparison as 20 due to the limited capacity of the expert knowledge.

We show our results in Figure 6.7. Particularly, in Figure 6.7, the three figures from top to bottom show the uncertainty of ordering of the KPIs when only observational data is used, observation data and 10 expert comparisons are used, and observation data and 20 expert comparisons are used, respectively. Apparently, it can be observed that, with observational data alone, the uncertainty of ordering of the variables can only achieve limited accuracy on the ordering. On the other hand, the proposed Bayesian learning and sensing framework effectively elicit expert knowledge to reduce the estimation uncertainty, and with 20 expert comparisons, we could achieve fairly accurate estimation of the ordering of the KPIs.

Table 6.6.1 provides a summary of some KPIs that can be categorized into three groups: the

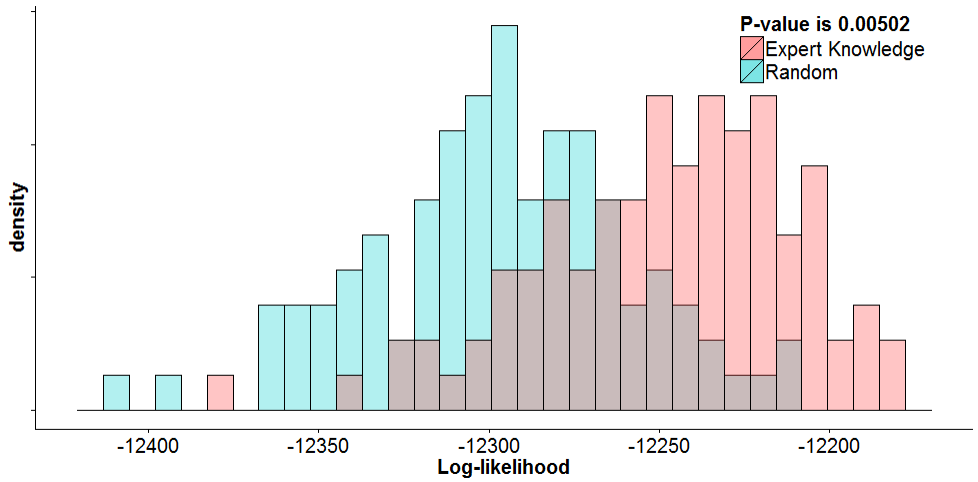


Figure 6.8: Validation of the utility of the expert comparison data in the KPI case study by following the approach mentioned in Section 6.4.4. It clearly shows that the expert comparison data is significantly different from random guess

driver, mediator, and downstream KPIs. A full list of the KPIs could be found in the Supplementary file. The driver KPIs that appear in the upstream positions in the learned ordering of the KPIs, correspond to solid recruiting criteria such as technical skills and work attitudes, indexes of internal resource and coordination, index of reward systems, and leadership index. It is reasonable for these KPIs to be drivers of the identified downstream KPIs. For instance, selection of employees who have sufficient technical skills, effective internal coordination, and reward system that motivate employees, all could have influence on the downstream indexes such as employee turnover rate. Among the downstream KPIs, what is worth noting is the "multi-functional employee". It is a qualification index of employees and can further lead to stability of business. The rationale behind this KPI to measure stability of business is based on the "personal successors system" mentioned in [217], when employees increase their qualification, they are able to occupy two positions, while at the same time, organization has personal successors on all positions such that it can run in normal mode when some employees suddenly quit the jobs. Knowing that this KPI could be driven by some upstream KPIs, quality improvement strategies targeting this KPI could be defined

by improving on the corresponding upstream KPIs to remove the root-causes as a more proactive and preventative strategy. It is also interesting to notice that, the KPIs that correspond to the "employee trait" and "hiring process" act as mediators that deliver the influence from the driver KPIs to downstream KPIs. This result seems also reasonable, as these KPIs relate to human capitals' growing potential, which will have impact on the growth (into multi-functional employees) as well as job turnover rate (due to either being overqualified or unqualified) ([218]).

In addition, the validation process of the utility of the expert data is also conducted via the method we proposed in Section 6.4.4. The result is reported in Figure 6.8. It clearly shows that the expert data is significantly different from random guess, demonstrating its utility for helping the learning of the influential relationships between the KPIs.

Ordering of Human Resource Management Key Performance Indicators (KPIs)		
	Index	**Meaning**
Driver KPI	hr-hrcrn	Compensation, Reward, and Incentive Systems
	pm-hsmbn	Management Breath of Experience
	sp-hspcn	Reward-Manufacturing Coordination
	sp-hsrsn04	Hiring Criteria: work values and attitudes
	sp-hsrsn10	Hiring Criteria: technical skills
	pm-hsvcn	Cooperation
Mediator KPI	sp-hsrsn02	Employee Trait: teamwork
	sp-hsrsn03	Employee Trait: problem solving
	sp-hsrsn07	Hiring Process: large candidate pool
	sp-hsrsn08	Hiring Process: effective interview
Downstream KPI	sp-hsmfn	Multi-functional Employees
	hr-hrtrn02	Turnover rate of Hourly Employees
	sp-psltn	Supply Lead Time

6.6.2 *Identification of Cascade of Hypermetabolism Reduction Events for Alzheimer's Disease (AD)*

In this section, we will implement our method to identify the cascade of hypermetabolism reduction events for Alzheimer's disease. Knowing the cascade of hypermetabolism reduction events will help us understand the progression stages of the disease, which is especially beneficial for early detection of the disease. Although disease progression model of AD has been developed in [17] that gives valuable hypothesis of the cascade of the abnormal clinical events along the AD progression process, such a model is still quite coarse and on the conceptual level, and is undergoing experimental validation. Thus, we propose to conduct quantitative analysis, and particularly focus on the FDG-PET imaging data that can measure the hypermetabolism reduction events, to enrich the disease progression model of AD and provide better resolution of the clinical events that happen along the progression process.

Our FDG-PET imaging data includes a diverse set of subjects that are on different progression stages. There are 49 AD subjects, 116 mild cognitive impairment (MCI) subjects, and 67 normal aging (NC) subjects. We have identified 42 regions of interest (ROIs) that have been found related to AD. We perform the following data preparation steps: the original PET scan data set could be written as $D \in R^{232\times 42}$, corresponding to FDG-PET measurements of 42 regions of the 232 subjects. In order to derive the hypermetabolism reduction events based on FDG-PET data, we build a $\bar{X}$ chart for each ROI to characterize the normal FDG-PET level for each region. If the FDG-PET measurement of the subject in this ROI is normal, then we label this region of this subject as 0; otherwise, we label it as 1. This is just like the Phase-I analysis of control chart. Through this procedure, we could derive the "event dataset", denoted as $E \in I^{232\times 42}$. We then implement our method on $E \in I^{232\times 42}$ and combine it with the "expert" based on an existing cascade model of hypermetabolism reduction events of AD reported in the literature. e.g. [219]. The final results are shown in Figure 6.9, which can be interpreted in the same way as Figure 6.7. Furthermore, the validation process of the utility of the expert data is also conducted and the result reported in Figure 6.11 shows that the expert data is significantly different from random guess,

demonstrating its utility for helping the learning of ordering of the hypermetabolism reduction events.

Again, from Figure 6.9 we could observe that the uncertainty of ordering is large, when only observational data is used. With addition of expert knowledge, the uncertainty can be effectively reduced. Figure 6.10 provides a visualization of the mean ordering of the variables extracted from the posterior distribution. It clearly shows where the neurodegeneration strikes first along the progression of AD. Based on the severity of the degeneration conditions from early to late, we use different colors to highlight these events from yellow (NC stage) through orange (MCI stage) to red (AD stage).

From Figure 6.10, we can see that the following regions may be involved in early stages of the AD progression. Those regions include the frontal mid orbitalis cortex (node 4), which is an important area involved in the cognitive processing of decision-making; the hippocampus cortex (node 20), located at the medial temporal lobe, is known to play key roles in the consolidation of information from short-term memory to long-term memory and spatial navigation; the middle temporal gyrus (node 16) serves in contemplating distance, recognition of known faces, and accessing word meaning while reading; the anterior cingulate cortex (ACC) (node 6) is involved in a wide range of cognitive functions such as rational cognitive functions, reward anticipation, decision-making, empathy, impulse control, and emotion; the inferior parietal lobule (node 8) is involved in the perception of emotions in facial stimuli and interpretation of sensory information; the Precuneus (node 9) is involved in episodic memory, visuospatial processing, reflections upon self, and aspects of consciousness. Apparently, many of these regions are related to the memory function, which is one of the early signs of AD. It is also exciting to see that the hippocampus region is involved in the AD progression in early stage, which is consistent with existing knowledge of AD in ([220, 221]) since the hippocampus has been a hallmark of AD. The fact that the Occipital Inferior region shows up at early stage is also interesting. It is known that the occipital inferior region belongs to the sensorimotor network (SMN), which relates to primary visual functions. A number of studies such as [222, 223, 224] have indicated that the functional changes in the visual system might precede the onset of AD, i.e., by impaired functional connectivity in visual

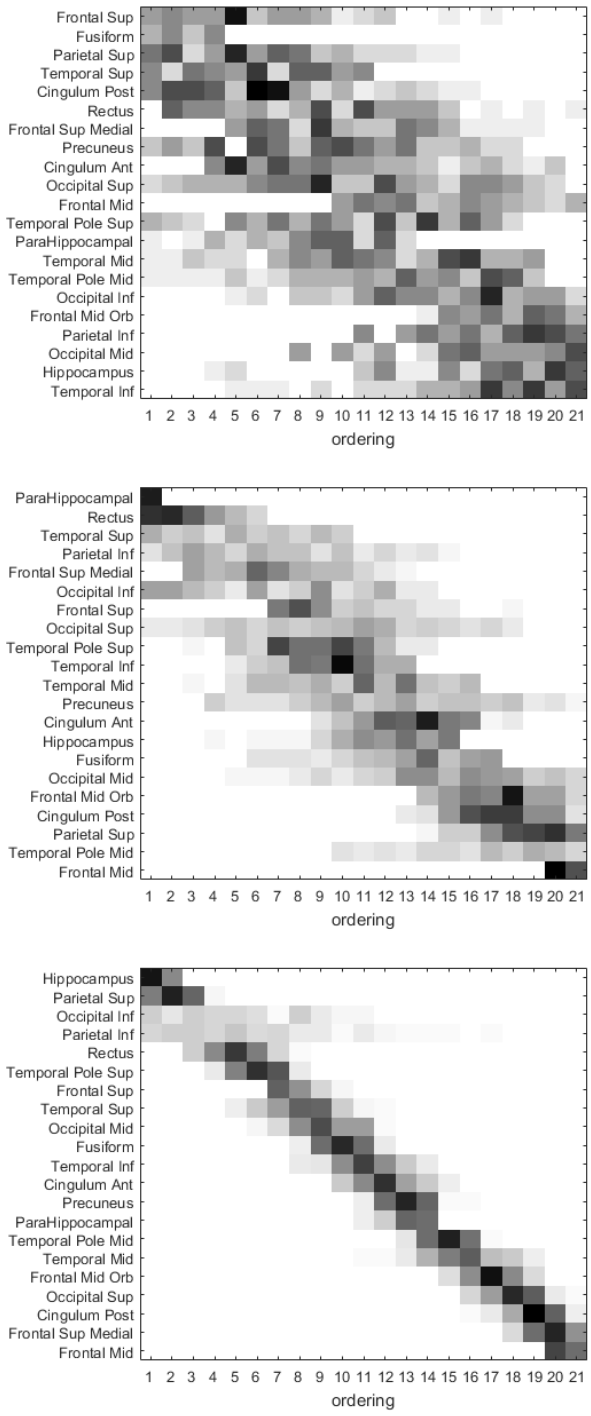
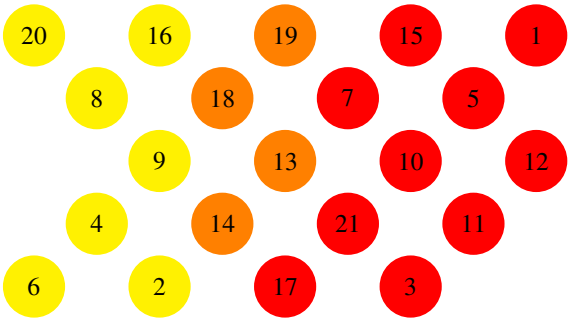


Figure 6.9: Uncertainty of ordering of the hypermetabolism reduction events when only observational data is used (top), observational data and 10 expert comparisons are used (middle), and observational data and 20 expert comparisons are used (bottom), respectively. Note that, the rows correspond to the hypermetabolism reduction events while the numbers in the x-axis represent the ordering of the hypermetabolism reduction events.



1	Frontal Sup	7	Parietal Sup	13	Occipital Inf	19	Fusiform
2	Frontal Mid	8	Parietal Inf	14	Temporal Sup	20	Hippocampus
3	Frontal Sup Medial	9	Precuneus	15	Temporal Pole Sup	21	ParaHippocampal
4	Frontal Mid Orb	10	Cingulum Post	16	Temporal Mid		
5	Rectus	11	Occipital Sup	17	Temporal Pole Mid		
6	Cingulum Ant	12	Occipital Mid	18	Temporal Inf		

Figure 6.10: The Hypermetabolism Reduction Event Cascade of AD progression

systems. Our discovery raises an interesting hypothesis that decline in visual functions may be an early noninvasive biomarker for the diagnosis of AD. Similar findings were previously discovered in [225].

The regions involved in the intermediate progression stages reveal valuable insights regarding the neurological underpinning of the MCI stage. From Figure 6.10, we can see that these regions include the frontal mid lobe (node 2), which contains most of the dopamine sensitive neurons in the cerebral cortex. Since the dopamine system plays a critical role in memory, planning, and motivation, the reduction in hypermetabolism in this area could result in poorer performance and inefficient functioning during memory tasks. On the other hand, the occipital inferior lobe (node 13) is a critical visual related cortex. Impairment of this region can cause visual hallucinations or blindness. Its relatedness to MCI has been pointed out in recent studies ([226, 227]). Such finding also provides evidence for the "sensory deprivation hypotheses" in [228], which proposes that sensory (including visual) underload may reduce opportunity for intellectual stimulating exchanges with the environment, and eventually reduce the general level of cognitive ability. In addition, the fusiform gyrus (node 19) is related to processing of color information, face, body, and word

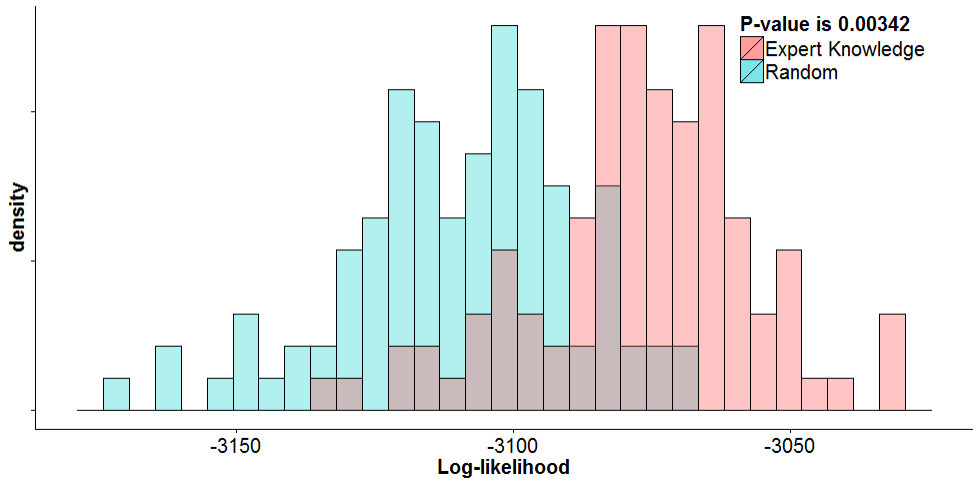


Figure 6.11: Validation of the utility of the expert comparison data in the AD case study by following the approach mentioned in Section 6.4.4. It clearly shows that the expert comparison data is significantly different from random guess

recognition. As many of these regions relate to the recognition ability, this is consistent with the scientific evidence from neuropsychological studies of AD that found that word recognition tasks are sensitive tests to identify the individuals who will develop AD soon. We also notice that, the regions such as the prefrontal lobe (node 1, 3,5), the parietal lobe(node 7, 10), the occipital lobe (node 11, 12), are involved in the late stages of AD progression. This is also consistent with the literature that has found these regions are usually less affected by AD in early stages ([229]).

6.7 Conclusion and Future Work

In this paper, we propose a first of its kind method that can systematically elicit expert knowledge, optimally matched to observational data that has been collected, to identify the influential relationships between variables. This work is motivated by the success of the BN models in characterizing a wide range of complex systems, which use DAG structure to represent how variables influence each other. The BN models have also been found very useful for a number of quality control and monitoring tasks, as evidenced by the existing works in [179, 180, 181, 182, 183]. However, as

the common approach for learning the DAG structure of a BN is merely using the observational data, it has been found that the theoretical bottleneck of this approach lies on the fact that only part of the influential relationships can be identified. On the other hand, in many applications, it is possible that knowledgeable expert could provide crucial information regarding the influential relationships among the variables. However, there has been a lack of systematical method that can automate the expert knowledge elicitation process, integrate it with existing learning methods from observational data, and further optimize it. Thus, we develop a Bayesian learning and sensing framework that can combine both observational data and expert elicitation data in the form as a posterior distribution of orderings of the variables. Such a Bayesian learning framework will further lead to a systematical optimization formulation to automate the expert elicitation process. We conduct extensive numerical experiments on simulated data and real-world data to demonstrate the utility of our method and show its superior performance than baseline approaches.

There are many future directions we could exploit. For example, one important direction is to extend this framework to other optimal design methods such as $A-$Optimal or $D-$Optimal to fit needs from a broad range of application domains. New optimization models will be constructed accordingly. Another direction is to develop a better probabilistic model to characterize the distribution of the ordering of the variables. In this study, we propose to use a surrogate vector of the ordering of variables, which is essentially a relaxation of the ordering which is a permutation set. As relaxation has been a common and effective tool for gaining numerical performance and computational feasibility, there is always a possibility that more statistical power could be gained if the relaxation could be tightened or not used at all. In addition, we may extend our method to be able to interact with multiple experts that have different accuracy levels of expert knowledge. Last but not least, it is worthy of pointing it out that the Bayesian learning framework demonstrated in the Figure 6.3 is generic. Thus, we could plug in any Bayesian network structure learning algorithm to learn the prior distribution of the orderings from observational data. Also, instead of Bootstrap, we could sample orderings of variables using Markov Chain Monte Carlo (MCMC) such as in [210, 169, 230]. In addition, there is an alternative approach to sample for DAGs rather than orderings of the variables in Algorithm 1. The output will be a collection of DAG structures then,

denoted as $\{\hat{G}_0^i\}_{i=1,\cdots,m}$, that provide a good sample-based representation of the posterior distribution of the DAG structure. To integrate this DAG-sampling framework with the proposed Bayesian learning framework, we need a procedure to derive the prior distribution of ϕ. Specifically, we recognize that there is a resemblance between the learned DAG $\hat{G}_0^i$ with the pairwise comparison data obtained from expert knowledge elicitation, i.e., a learned DAG $\hat{G}_0^i$ could be viewed as a collection of pairwise comparison data by defining that w_k = count of directed edges (i, j) that shows up in all samples $\hat{G}_0^i$, and $y_k = \frac{1}{w_k} \sum_{(i,j)\in\hat{G}_0^1} sgn(i \to j)\cdot$, where k corresponds to the directed edge (i, j). Here $sgn(y)$ is used to indicate the direction of arcs. Following this line, we could derive the prior distribution of ϕ. All these directions are worthy of exploiting to further enhance and enrich the proposed methodology, that can combine observational data and expert knowledge for better learning of the influential relationships between variables.

6.8 Conclusions and Future Work

In this work, we have developed a systematical framework that unifies dynamic modeling, sparse learning, dictionary learning, and matrix completion, to automate the process of translating the user's behavior data into deeply personalized and achievable health planning. Our method is generic and can be applied to a wide range of health management problems such as obesity, fitness, diabetes, or any chronic conditions, where the disease process is a complex dynamic process that can be modified by exogenous variables such as some behavioral and clinical variables. Our framework holds great potential to provide scalable solutions for mitigating these health problems, which can promote healthier lifestyles outside of clinical settings. We further apply our proposed models to extensive experiments on real-world daily behavioral data, which demonstrate promising utility and efficacy of our method. Future works include extensions of the model to other dynamic models that have different characteristic than linear models, e.g., some diseases, such as Depression, might follow a different dynamic process. While currently we assume sufficient data has been collected for each individual, it is also of interest to develop an adaptive learning and planning model that can be applied to scenarios where data come in sequentially.

BIBLIOGRAPHY

[1] K. Li, M. O'Farrell, D. Martin, S. Kopf, C. Harner, and X. Zhang, "Mapping ligament insertion sites onto bone surfaces in knee by co-registration of CT and digitization data," *Journal of Biomechanics*, vol. 42, no. 15, pp. 2624–2626, 2009.

[2] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[3] A. of Medical Sciences., "Stratified, personalised or p4 medicine: a new direction for placing the patient at the centre of healthcare and health education (technical report)." *Academy of Medical Sciences.*, 2015.

[4] J. M. Swanson, G. A. Sunohara, J. Kennedy, R. Regino, E. Fineberg, T. Wigal, M. Lerner, and L. W. et al, "Association of the dopamine receptor d4 (drd4) gene with a refined phenotype of attention deficit hyperactivity disorder (adhd): a family-based approach," *Mol. Psychiatry*, vol. 3, pp. 38–41, 1998.

[5] W. Cheng, X. Ji, and J. Feng, "Individual classification of adhd patients by integrating multiscale neuroimaging markers and advanced pattern recognition techniques," *Frontiers in System Neuroscience*, vol. 6, no. 58, 2012.

[6] M. Ahmadlou and H. Adeli, "Wavelet-synchronization methodology: A new approach for eeg-based diagnosis of adhd," *Clinical EEG and Neuroscience*, vol. 41, 2010.

[7] M. Holtmann, K. Becker, B. Kentner-Figura, and M. H. Schmidt, "Increased frequency of rolandic spikes in adhd children," *Epilepsia*, vol. 44, p. 1241Ã¢ÂĂÂŞ1244, 2003.

[8] G. W. Hynd, M. Semrud-Clikeman, A. R. Lorys, E. S. Novey, D. Eliopulos, and H. Lyytinen, "Corpus callosum morphology in attention deficit-hyperactivity disorder: Morphometric analysis of mri," *Journal of Learning Disability*, vol. 24, 1991.

[9] D. E. J. Linden, "The challenges and promise of neuroimaging in psychiatry," *Neuron*, vol. 73, p. 8Ã¢ÂĂÂŞ22, 2012.

[10] A. Tenev, S. Markovska-Simoska, L. Kocarev, J. Pop-Jordanov, A. Muller, and G. Candrian, "Machine learning approach for classification of adhd adults," *International journal of psychophysiology: official journal of the International Organization of Psychophysiology*, 2013.

[11] X. Peng, P. Lin, T. Zhang, and J. Wang, "Extreme learning machine-based classification of adhd using brain structural mri data," *PLoS ONE*, vol. 8(11), 2013.

[12] E. Foundation, "Epilepsy foundation: not another moment lost to seizures," 2006. [Online]. Available: {http://www.epilepsyfoundation.org}

[13] D. Mould, "Models for disease progression: new approaches and uses." *Clinical Pharmacology & Therapeutics*, vol. 92, no. 1, p. 125âĂŞ131, 2012.

[14] 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. [Online]. Available: http://dx.doi.org/10.1111/1467-9884.00351

[15] K. Ito, S. Ahadieh, B. Corrigan, J. French, T. Fullerton, and T. Tensfeldt, "Disease progression meta-analysis model in alzheimerâĂŹs disease." *AlzheimerâĂŹs and Dementia*, vol. 6(1), p. 39âĂŞ53, 2010.

[16] M. Exarchos, T. Exarchos, C. Bourantas, M. Papafaklis, K. Naka, L. Michalis, O. Parodi, and D. Fotiadis, "Prediction of coronary atherosclerosis progression using dynamic bayesian networks." *EMBC*, p. 3889âĂŞ3892, 2013.

[17] C. Jack, D. Knopman, W. Jagust, L. Shaw, P. Aisen, M. Weiner, R. Petersen, and J. Trojanowski, "Hypothetical model of dynamic biomarkers of the alzheimer's pathological cascade," *Lancet Neurol.*, vol. 9, no. 1, pp. 119–28, 2010.

[18] W. E. Garrett, M. F. Swiontkowski, J. N. Weinstein, J. Callaghan, R. N. Rosier, D. J. Berry, J. Harrast, and G. P. Derosa, "American board of orthopaedic surgery practice of the orthopaedic surgeon: Part-ii, certification examination case mix." *J Bone Joint Surg Am*, vol. 88, pp. 660–667, 2006.

[19] M. Ferretti, D. Doca, S. Ingham, M. Cohen, and F. Fu, "Bony and soft tissue landmarks of the acl tibial insertion site: an anatomical study," *Knee Surgery Sports Traumatology Arthroscopy*, vol. 20, pp. 62–68, 2012.

[20] H. Gadikota, J. Sim, A. Hosseini, T. Gill, and G. Li, "The relationship between femoral tunnels created by the transtibial, anteromedial portal, and outside-in techniques and the anterior cruciate ligament footprint," *The American Journal of Sports Medicine*, vol. 40, pp. 882–888, 2012.

[21] J. H. Lubowitz, "Anteromedial portal technique for the anterior cruciate ligament femoral socket: pitfalls and solutions." *Arthroscopy*, vol. 25, pp. 95–101, 2009.

[22] C. D. Harner, G. H. Baek, T. M. Vogrin, G. J. Carlin, S. Kashiwaguchi, and S. L.-Y. Woo, "Quantitative analysis of human cruciate ligament insertions," *Arthroscopy: The Journal of Arthroscopic & Related Surgery*, vol. 15, no. 7, pp. 741–749, 1999.

[23] R. Siebold, T. Ellert, S. Metz, and J. Metz, "Tibial insertions of the anteromedial and posterolateral bundles of the anterior cruciate ligament: morphometry, arthroscopic landmarks, and orientation model for bone tunnel placement," *Arthroscopy*, vol. 24, pp. 154–161, 2008.

[24] A. Edwards, A. Bull, and A. Amis, "The attachments of the anteromedial and posterolateral fibre bundles of the anterior cruciate ligament: Part 1: tibial attachment," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 15, pp. 1414–1421, 2007.

[25] G. Tajima, M. Nozaki, T. Iriuchishima, S. Ingham, W. Shen, P. Smolinski, and F. Fu, "Morphology of the tibial insertion of the posterior cruciate ligament," *Journal of Bone and Joint Surgery*, vol. 91, pp. 859–866, 2009.

[26] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, I. Smith, and J. A. Landay, "Activity sensing in the wild: A field trial of ubifit garden," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[27] N. D. Lane, M. Lin, M. Mohammod, X. Yang, H. Lu, G. Cardone, S. Ali, A. Doryab, E. Berke, A. T. Campbell, and T. Choudhury, "Bewell: Sensing sleep, physical activities and social interactions to promote wellbeing," *Mob. Netw. Appl.*, vol. 19, no. 3, Jun. 2014.

[28] L. MyFitnessPal, 2013. [Online]. Available: http://www.myfitnesspal.com/

[29] I. Fitbit, 2013. [Online]. Available: http://www.fitbit.com/

[30] R. Kukafka, "Tailored health communication." *Consumer Health Informatics: Informing Consumers and Improving Health Care*, pp. 22–33, 2005.

[31] J. Lester, C. Hartung, L. Pina, R. Libby, G. Borriello, and G. Duncan, "Validated caloric expenditure estimation using a single body-worn sensor," in *Proceedings of the 11th International Conference on Ubiquitous Computing*, ser. UbiComp '09, 2009.

[32] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury, "Mybehavior: Automatic personalized health feedback from user behaviors and preferences using smartphones," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15, 2015.

[33] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed Bandit Allocation Indices (2nd Edition)*, 2011.

[34] R. J. Adams, R. S. Sadasivam, K. Balakrishnan, R. L. Kinney, T. K. Houston, and B. M. Marlin, "Perspect: Collaborative filtering for tailored health communications," in *Proceedings of the 8th ACM Conference on Recommender Systems*, ser. RecSys '14, 2014.

[35] C. Xiao and W. Chaovalitwongse, "Optimization models for feature selection of decomposed nearest neighbor." *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46(2), pp. 177–184, 2016.

[36] C. Xiao, J. Bledsoe, S. Wang, S. Mehta, M. Semrud-Clikeman, T. Grabowski, and W. Chaovalitwongse, "An integrated feature ranking and selection framework for adhd diagnosis," *Brain Informatics*, pp. 1–11, Apr 2016.

[37] S. Wang, C. Xiao, J. Tsai, W. Chaovalitwongse, and T. Grabowski, "A novel mutual-information-guided sparse feature selection approach for epilepsy diagnosis using interictal eeg signals." in *Proceedings of the 2016 International Conference on Brain Informatics and Health*, ser. BIH 16', 2016.

[38] C. Xiao, S. Wang, L. Zheng, X. Zhang, and W. Chaovalitwongse, "A patient-specific model for predicting tibia soft tissue insertions from bony outlines using a spatial structure supervised learning framework." *IEEE Transactions on Human-Machine Systems*, vol. 46(5), 2016.

[39] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.

[40] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

[41] L. Torresani and K.-C. Lee, "Large margin component analysis," in *Advances in Neural Information Processing Systems 19*. MIT Press, 2007, pp. 1385–1392.

[42] C. Domeniconi, J. Peng, and D. Gunopulos, "An adaptive metric machine for pattern classification," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2001, pp. 458–464.

[43] C. Domeniconi and D. Gunopulos, "Efficient local flexible nearest neighbor classification," in *Proceedings of the 2002 SIAM International Conference on Data Mining*, ch. 21, pp. 353–369.

[44] T. Hastie and R. Tibshirani, "Discriminant analysis by gaussian mixtures," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 155–176, 1996.

[45] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 521–528.

[46] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[47] J. Ye, "Multiple closed-form local metric learning for k-nearest neighbor classifier," *Clinical Orthopaedics and Related Research*, 2013.

[48] M.-S. Yang and C.-H. Chen, "On the edited fuzzy k-nearest neighbor rule," *IEEE Trans. Syst., Man, Cybern. B*, vol. 28, no. 3, pp. 461–466, Jun 1998.

[49] F. Moreno-Seco, L. MicÃş, and J. Oncina, "A modification of the laesa algorithm for approximated k-nn classification," *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 47–53, 2003.

[50] M. Khan, Q. Ding, and W. Perrizo, "k-nearest neighbor classification on spatial data streams using p-trees," in *Advances in Knowledge Discovery and Data Mining, 6th Pacific-Asia Conference, PAKDD 2002*. Springer Berlin Heidelberg, 2002, vol. 2336, pp. 517–528.

[51] J. Gou, T. Xiong, and Y. Kuang, "A novel weighted voting for k-nearest neighbor rule," *Journal of Computers*, vol. 6, no. 5, 2011.

[52] D. Tarlow, K. Swersky, L. Charlin, I. Sutskever, and R. Zemel, "Stochastic k-neighborhood selection for supervised and unsupervised learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, vol. 28, no. 3, May 2013, pp. 199–207.

[53] N. Pal and S. Ghosh, "Some classification algorithms integrating dempster-shafer theory of evidence with the rank nearest neighbor rules," *IEEE Trans. Syst., Man, Cybern. A*, vol. 31, no. 1, pp. 59–66, Jan 2001.

[54] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. Cambridge University Press, 2012, ch. 3.

[55] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[56] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annu. Eugen.*, vol. 7, no. 2, pp. 179–188, 1936.

[57] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[58] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar 1982.

[59] E. W. Forgy, "Cluster analysis of multivariate data: efficiency versus interpretability of classifications," *Biometrics*, vol. 21, pp. 768–769, 1965.

[60] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C*, vol. 28, no. 1, pp. 100–108, 1979.

[61] P. C. Mahalanobis, "On the generalised distance in statistics," in *Proceedings National Institute of Science, India*, vol. 2, no. 1, Apr. 1936, pp. 49–55.

[62] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1996.

[63] S. Shalev-shwartz, Y. Singer, and A. Y. Ng, "Online and batch learning of pseudo-metrics," in *In ICML*. ACM Press, 2004, pp. 743–750.

[64] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *Proceedings of the 23rd International Conference on Machine Learning*, ser. ICML '06. New York, NY, USA: ACM, 2006, pp. 353–360.

[65] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[66] ——, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. Springer-Verlag Limited, 2008, pp. 95–110.

[67] T. Hastie, R. Tibshirani, and A. Buja, "Flexible discriminant analysis by optimal scoring," *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, vol. 89, pp. 1255–1270, 1993.

[68] J. H. Friedman, "Regularized discriminant analysis," *J. Amer. Statist. Assoc*, pp. 165–175, 1989.

[69] J. Ye, R. Janardan, V. Cherkassky, T. Xiong, J. Bi, and C. Kambhamettu, "Efficient model selection for regularized linear discriminant analysis," in *In Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 2006, pp. 532–539.

[70] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[71] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[72] W. Weaver, "Recent contributions to the mathematical theory of communication," *The Mathematical Theory of Communication*, 1949.

[73] M. Posner and J. Fan, *Attention as an organ system. In J. R. Pomerantz (Ed.), Topics in Integrative Neuroscience: From Cells to Cognition*. Cambridge University Press, 2008.

[74] R. Barkley, "Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of adhd," *Psychological Bulletin*, vol. 121(1), pp. 65–94, 1997.

[75] D. Dougherty, A. Bonab, T. Spencer, S. Rauch, B. Madras, and A. Fischman, "Dopamine transporter density in patients with attention deficit hyperactivity disorder," *The Lancet*, vol. 354(9196), pp. 2132–2133, 1999.

[76] J. Nigg, E. Willcutt, A. Doyle, and E. Sonuga-Barke, "Casual heterogeneity in attention-deficit hyperactivity disorder: Do we need neuropsychologically impaired subtypes," *Biological Psychiatry*, vol. 57, pp. 1224–1230, 2005.

[77] T. Sagvolden and J. Sergeant, "Attention deficit hyperactivity disorder: From brain dysfunctions to behaviour," *Behavioral Brain Research*, vol. 94(1), pp. 1–10, 1998.

[78] E. Sonuga-Barke, "Psychological heterogeneity in adhd - a dual pathway model of behaviour and cognition," *Behavioral Brain Research*, pp. 29–36, 2002.

[79] J. Swanson, G. Elliott, L. Greenhill, T. Wigal, E. Arnold, and B. et al., "Effects of stimulant medication on growth rates across 3 years in the mta follow-up," *Journal of the American Academy of Child and Adolescent Psychiatry, 46(8)*, pp. 1015–1027, 2007.

[80] F. Castellanos, J. Giedd, P. Berquin, J. Walter, W. Sharp, and T. Tran, "Quantification brain magnetic resonance imaging in girls with attention-deficit/hyperactivity disorder," *Archives of General Psychiatry*, vol. 58, pp. 289–295, 2001.

[81] F. Castellanos, J. Giedd, W. Eckburg, A. Marsh, D. Kaysen, and S. Hamburger, "Quantitative morphology of the caudate nucleus in attention deficit hyperactivity disorder," *American Journal of Psychiatry*, vol. 151(1212), pp. 1791–1796, 1994.

[82] F. Castellanos, J. Giedd, A. Marsh, S. Hamburger, A. Vaiturzis, and D. Dickstein, "Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder," *Archives of General Psychiatry*, vol. 53(7), pp. 607–616, 1996.

[83] F. Castellanos, P. Lee, W. Sharp, N. Jeffries, D. Greenstein, and L. C. et al., "Developmental trajectories of brain volume abnormalities in childrne with adolescents with attention-deficit/hyperactivity disorder," *The Journal of the American Medical Association*, vol. 28(4), pp. 1740–1749, 2002.

[84] S. Durston, Hulshoff, H. Schnack, J. Buitelaar, M. Steenhuis, and R. M. et al., "Magnetic resonance imaging of boys with attention-deficit/hyperactivity disorder and their unaffected siblings," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 43, no. 3, pp. 332–340, 2004.

[85] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2001.

[86] G. D. Tourassi, E. D. Frederick, M. K. Markey, and J. Carey E. Floyd, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *American Association of Physicists in Medicine*, 2001.

[87] D. Lewis, "Feature selection and feature extraction for text categorization," in *In Proceedings of Speech and Natural Language Workshop*. Morgan Kaufmann, 1992, pp. 212–217.

[88] R. Battiti, "Using the mutual information for selecting in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, pp. 537–550, 1994.

[89] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *The IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, p. 1226, 2005.

[90] H. Yang and J. Moody, "Feature selection based on joint mutual information," *In Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, 1999.

[91] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, 2004.

[92] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[93] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, p. 1667, 2002.

[94] R. Kohavi and G. H. John, "Wrapper for feature subset selection," *Artificial intellegence*, vol. 97, p. 273, 1997.

[95] Y. Kim, S. Kwon, H. Choi, and X. Shen, "Consistent model selection criteria on high dimensions," 2012.

[96] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, pp. 1418–1429, 2006.

[97] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[98] A. Dale, B. Fischl, and M. Sereno, "Cortical surface-based analysis. i. segmentation and surface reconstruction," *NeuroImage*, vol. 9, pp. 179–194, 1999.

[99] A. Dale and M. Sereno, "Improved localization of cortical activity by combining eeg and meg with mri cortical reconstruction: A linear approach," *Journal of Cognitive Neuroscience*, vol. 5, pp. 162–176, 1993.

[100] R. Desikan, F. Segonne, B. Fischl, B. Quinn, B. Dickerson, and D. B. et al., "An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest," *NeuroImage*, vol. 31, pp. 968–980, 2006.

[101] P. a. Pudil, "Floating search methods in feature selection," *Pattern Recognition Letters*, 1994.

[102] J. Fuster, *The Prefrontal Cortex (4th ed.)*. London: Academic Press, 2008.

[103] R. Martinussen, J. Hayden, S. Hogg-Johnson, and R. Tannock, "A meta-analysis of working memory impairments in children with attention-deficit hyperactivity disorder," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 44, pp. 377–384, 2005.

[104] J. Nigg, L. Blaskey, Huang-Pollock, and Rappley, "Neuropsychological executive functions and dsm-iv adhd subtypes," *Journal of the American Academy of Child and Adolescent Psychiatry*, vol. 41, pp. 59–66, 2002.

[105] A. Aron, J. Dowson, B. Sahakian, and T. Robbins, "Methylphenidate improves response inhibition in adults with attention-deficit hyperactivity disorder," *Biological Psychiatry*, vol. 54(12), pp. 1465–1468, 2003.

[106] R. et al, "Hypofrontality in attention deficit hyperactivity disorder during higher-order motor control: A study with functional mri," *American Journal of Psychiatry*, vol. 156(6), pp. 891–896, 1999.

[107] M. Posner and S. Petersen, "The attention system of the human brain," *Annual Review of Neuroscience*, vol. 13, pp. 25–42, 1990.

[108] F. Castellanos and E. Proal, "Large-scale brain systems in adhd: Beyond the prefrontal striatal model," pp. 17–26, 2006.

[109] G. Bush, P. Luu, and M. Posner, "Cognitive and emotional influences in anterior cingulate cortex," *Trends in Cognitive Sciences*, 2000.

[110] G. Bush, J. Frazier, S. Rauch, L. Seidman, P. Whalen, M. Jenike, B. Rosen, and J. Biederman, "Anterior cingulate cortex dysfunction in attention-deficit hyperactivity disorder revealed by fmri and the counting stroop," *Biological Psychiatry*, vol. 45, pp. 1542–1552, 1999.

[111] E. et al, "Neural substrates of decision making in adults with attention deficit hyperactivity disorder," *The American Journal of Psychiatry*, vol. 160, pp. 1061–1070, 2003.

[112] P. et al, "Neuroimaging of inhibitory control areas in children with attention deficit hyperactivity disorder who were treatment naive or in long-term treatment," *American Journal of Psychiatry*, vol. 163(6), pp. 1052–1060, 1999.

[113] N. Makris, J. Biederman, E. M. Valera, G. Bush, J. Kaiser, D. N. Kennedy, V. S. Caviness, S. V. Faraone, and L. J. Seidman, "Cortical thinning of the attention and executive networks in adults with attention-deficit hyperactivity disorder," *Cerebral Cortex*, vol. 17(6), pp. 1364–1375, 2007.

[114] J. C. Bledsoe, M. Semrud-Clikeman, and S. R. Pliszka, "Anterior cingulate cortex and symptom severity in attention-deficit/hyperactivity disorder," pp. 558–565, 2013.

[115] B. G, "Cingulate, frontal, and parietal cortical dysfunction in attention-deficit/hyperactivity disorder," pp. 1160–1167, 2011.

[116] e. a. Sturm W, "Functional anatomy of intrinsic alertness: evidence for a fronto-parietal-thalamic-brainstem network in the right hemisphere," pp. 797–805, 1999.

[117] A. Sapir, A. Hayes, A. Henik, S. Danziger, and R. Rafal, "Parietal lobe lesions disrupt saccadic remapping of inhibitory location tagging," *Journal of Cognitive Neuroscience*, vol. 16, pp. 503–509, 2004.

[118] A. Wilkins, I. Nimmo-Smith, A. Tait, C. McManus, S. Sala, and A. Tilley, "A neurological basis for visual discomfort." *Brain*, vol. 107, p. 989âĂŞ1017, 1984.

[119] F. Vialatte, M. Maurice, J. Dauwels, and A. Cichocki, "Steady-state visually evoked potentials: focus on essential paradigms and future perspectives." *Prog. Neurobiol.*, vol. 90, no. 4, p. 418âĂŞ438, 2010.

[120] E. Asano, M. Nishida, M. Fukuda, R. Rothermel, C. Juhasz, and S. Sood, "Differential visually-induced gamma-oscillations in human cerebral cortex." *NeuroImage*, vol. 45, p. 477âĂŞ489, 2009.

[121] D. Regan, "Human brain electrophysiology: Evoked potentials and evoked magnetic fields in science and medicine." 1989.

[122] G. Muller-Putz, R. Scherer, C. Brauneis, and G. Pfurtscheller, "Steady-state visual evoked potential (ssvep)-based communication: impact of harmonic frequency components." *J. Neural Eng.*, vol. 2(4), p. 123âĂŞ130, 2005.

[123] S. Smith, "Eeg in the diagnosis, classification, and management of patients with epilepsy." *J. Neurol. Neurosurg. Psychiatry*, vol. 76, p. ii2âĂŞii7, 2005.

[124] R. Tibshirani, "Regression shrinkage and selection via the lasso." *J. R. Stat. Soc.*, vol. B 58, p. 267âĂŞ288, 1994.

[125] N. Draper and H. Smith, "Applied regression analysis." 1998.

[126] P. Pudil, J. Novoviov, and J. Kittler, "Floating search methods in feature selection." *Pattern Recogn. Lett.*, vol. 15(11), p. 1119âĂŞ1125, 1994.

[127] S. Plaweski, D. Petek, and D. Saragaglia, "Morphometric analysis and functional correlation of tibial and femoral footprints in anatomical and single bundle reconstructions of the anterior cruciate ligament of the knee," *Orthopaedics & Traumatology: Surgery & Research*, vol. 97, pp. S75–79, 2011.

[128] T. Zantop, N. Diermann, T. Schumacher, S. Schanz, F. Fu, and W. Petersen, "Anatomical and nonanatomical double-bundle anterior cruciate ligament reconstruction: importance of femoral tunnel location on knee kinematics," *The American Journal of Sports Medicine*, vol. 36, pp. 678–685, 2008.

[129] R. Simmons, S. Howell, and M. Hull, "Effect of the angle of the femoral and tibial tunnels in the coronal plane and incremental excision of the posterior cruciate ligament on tension of an anterior cruciate ligament graft: an in vitro study," *Journal of Bone and Joint Surgery*, vol. 85-A, pp. 1018–1029, 2003.

[130] T. Iriuchishima, G. Tajima, S. Ingham, K. Shirakura, and F. Fu, "Pcl to graft impingement pressure after anatomical or non-anatomical single-bundle acl reconstruction," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 20, pp. 964–969, 2012.

[131] M. Galloway, E. Grood, J. Mehalik, M. Levy, S. Saddler, and F. Noyes, "Posterior cruciate ligament reconstruction. an in vitro study of femoral and tibial graft placement," *The American Journal of Sports Medicine*, vol. 24, pp. 437–445, 1996.

[132] D. Biau, C. Tournoux, S. Katsahian, P. Schranz, and R. Nizard, "Acl reconstruction: a meta-analysis of functional scores." *Clin Orthop Relat Res*, vol. 458, pp. 180–187, 2007.

[133] D. C. Fithian, E. W. Paxton, M. L. Stone, W. F. Luetzow, R. P. Csintalan, D. Phelan, and D. M. Daniel, "Prospective trial of a treatment algorithm for the management of the anterior cruciate ligament- injured knee." *Am J Sports Med*, vol. 33, pp. 335–346, 2005.

[134] L. Zheng, C. D. Harner, and X. Zhang, "The morphometry of soft tissue insertions on the tibial plateau: Data acquisition and statistical shape analysis," *PLoS ONE*, vol. 9, no. 5, p. e96515, 05 2014.

[135] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 41–48.

[136] F. YJ and W. Chaovalitwongse, "Optimizing feature selection to improve medical diagnosis," *Annals of Operations Research*, vol. 174(1), pp. 169–183.

[137] W. Chaovalitwongse, Y. Jeong, J. MK, D. SF, and W. S, "Pattern recognition approaches for identifying subcortical targets during deep brain stimulation surgery," *Intelligent Systems, IEEE*, vol. 26(5), pp. 54–63, 2011.

[138] O. Seref, Y. Fan, and W. Chaovalitwongse, "Mathematical programming formulations and algorithms for discrete k-median clustering of time-series data," *INFORMS Journal on Computing*, vol. 26(1), pp. 160–172, 2013.

[139] T. Cover and P. Hart, "Nearest neighbor pattern classification," *Information Theory, IEEE Transactions on*, vol. 13, no. 1, pp. 21–27, January 1967.

[140] B.-K. Yi, H. Jagadish, and C. Faloutsos, "Efficient retrieval of similar time sequences under time warping," in *Proceedings of 14th International Conference on Data Engineering*, Feb 1998, pp. 201–208.

[141] E. Pena, B. Calvo, M. Martinez, and M. Doblare, "A three-dimensional finite element analysis of the combined behavior of ligaments and menisci in the healthy human knee joint," *Journal of Biomechanics*, vol. 39, pp. 1686–1701, 2006.

[142] T. Guess, G. Thiagarajan, M. Kia, and M. Mishra, "A subject specific multibody model of the knee with menisci," *Medical Engineering & Physics*, vol. 32, pp. 505–515, 2010.

[143] A. Sandholm, C. Schwartz, N. Pronost, M. de Zee, M. Voigt, and D. Thalmann, "Evaluation of a geometry-based knee joint compared to a planar knee joint," *The Visual Computer*, vol. 27, no. 2, pp. 161–171, 2011.

[144] G. Li, J. Gi, A. Kanamori, and S. Woo, "A validated three-dimensional computational model of a human knee joint," *Journal of Biomechanical Engineering*, vol. 121, pp. 657–662, 1999.

[145] J. Brody, M. Hulstyn, B. Fleming, and G. Tung, "The meniscal roots: gross anatomic correlation with 3-t mri findings," *American Journal of Roentgenology*, vol. 188, pp. W446–450, 2007.

[146] P. Araujo, C. van Eck, M. Torabi, and F. Fu, "How to optimize the use of MRI in anatomic ACL reconstruction," *Knee Surgery, Sports Traumatology, Arthroscopy*, vol. 21, pp. 1495–1501, 2012.

[147] T. Donahue, M. Hull, M. Rashid, and C. Jacobs, "A finite element model of the human knee joint for the study of tibio-femoral contact," *Journal of Biomechanical Engineering*, vol. 124, pp. 273–280, 2002.

[148] C. Ogden, M. Carroll, B. Kit, and K. Flegal, "Prevalence of childhood and adult obesity in the united states, 2011-2012." *JAMA.*, vol. 311, no. 8, 2014.

[149] G. Rodgers and F. Collins, "The next generation of obesity research: No time to waste." *JAMA.*, vol. 308, no. 11, 2012.

[150] S. Consolvo, P. Klasnja, D. W. McDonald, and J. A. Landay, "Designing for healthy lifestyles: Design considerations for mobile technologies to encourage consumer health and wellness," *Found. Trends Hum.-Comput. Interact.*, vol. 6, 2014.

[151] B. Spring, M. Gotsis, A. Paiva, and D. Spruijt-Metz, "Healthy apps: mobile devices for continuous monitoring and intervention." *IEEE Pulse.*, pp. 34–40., 2013.

[152] D. Rivera, *Optimized behavioral interventions: What does system identification and control engineering have to offer?*, part 1 ed., 2012, vol. 16, pp. 882–893.

[153] D. Fung and C. Sheung, "Methods for the estimation of missing values in time series," Ph.D. dissertation, Edith Cowan University Perth, 2006.

[154] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 12, no. 2, 2010.

[155] P. J. Antsaklis and A. N. Michel, *A Linear Systems Primer*, 2007.

[156] R. Somasundaram and R. Nedunchezhian, "Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values," *International Journal of Computer Applications*, 2011.

[157] A. Gelman and J. Hill, *Data analysis using regression and multilevel and hierarchical models*. Cambridge University Press, 2007.

[158] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of optimization theory and applications*, vol. 109, no. 3, 2001.

[159] B. Fogg, "A behavior model for persuasive design," in *Proceedings of the 4th International Conference on Persuasive Technology*, ser. Persuasive '09, 2009.

[160] Y. Huang and J. Liu, "Exclusive sparsity norm minimization with random groups via cone projection," *arXiv preprint arXiv:1510.07925*, 2015.

[161] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009.

[162] C.-C. M. Yeh and Y.-H. Yang, "Supervised dictionary learning for music genre classification," in *Proceedings of the 2Nd ACM International Conference on Multimedia Retrieval*, ser. ICMR '12, 2012.

[163] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 3449–3456.

[164] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[165] J. Huang, T. Zhang *et al.*, "The benefit of group sparsity," *The Annals of Statistics*, vol. 38, no. 4, 2010.

[166] A. Fitch, C. Fox, and K. e. a. Bauerly, "Guideline of prevention and management of obesity for children and adolescents," 2013.

[167] J. Pearl, *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge University Press, 2000.

[168] N. Friedman, M. Linial, and I. Nachman, "Using bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, pp. 601–620, 2000.

[169] N. Friedman, I. Nachman, and D. Peer, "Learning Bayesian network structure from massive datasets," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '99, 1999, pp. 206–215.

[170] M. Zou and S. D. Conzen, "A new dynamic Bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data," *Bioinformatics*, vol. 21, no. 1, pp. 71–79, Jan. 2005.

[171] C. A. Pollino, O. Woodberry, A. Nicholson, K. Korb, and B. T. Hart, "Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment," *Environmental Modelling and Software*, vol. 22, no. 8, pp. 1140–1152, Aug. 2007.

[172] B. G. Marcot, J. D. Steventon, G. D. Sutherland, and R. K. McCann, "Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation," *Canadian Journal of Forest Research*, vol. 36, no. 12, pp. 3063–3074, 2006.

[173] A. Goldenberg, A. Zheng, S. Fienberg, and E. Airoldi, "A survey of statistical network models," *Foundations and Trends in Machine Learning*, vol. 2, no. 2, pp. 129–233, Feb. 2010.

[174] J. He, "A social network-based recommender system," Ph.D. dissertation, Los Angeles, CA, USA, 2010.

[175] D. Nikovski, "Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 4, pp. 509–516, Jul. 2000.

[176] S. Huang, J. Li, J. Ye, A. Fleisher, K. Chen, T. Wu, and E. Reiman, "Brain effective connectivity modeling for Alzheimer's disease by sparse Gaussian Bayesian network," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11, 2011, pp. 931–939.

[177] ——, "A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1328–1342, Jun. 2013.

[178] M. Hanafy and H. ElMaraghy, "Co-design of products and systems using a bayesian network," *Procedia CIRP*, vol. 17, pp. 284 – 289, 2014.

[179] J. Li and J. Shi, "Knowledge discovery from observational data for process control using causal Bayesian networks," *IIE Transactions*, vol. 39, no. 6, pp. 681 – 690, 2007.

[180] S. Verron, J. Li, and T. Tiplica, "Fault detection and isolation of faults in a multivariate process with Bayesian network," *Journal of Process Control*, vol. 20, no. 8, pp. 902 – 911, 2010.

[181] K. Liu and J. Shi, "Objective-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a Bayesian network," *IIE Transactions*, vol. 45, no. 6, pp. 630 – 643, 2013.

[182] K. Liu, X. Zhang, , and J. Shi, "Adaptive sensor allocation strategy for process monitoring and diagnosis in a Bayesian network," *IEEE Transactions on Automation Science and Engineering*, vol. 11, no. 2, pp. 452 – 462, 2014.

[183] J. Li, J. Jin, and J. Shi, "Causation-based T2 decomposition for multivariate process monitoring and diagnosis," *Journal of Quality Technology.*, vol. 40, no. 1, pp. 46 – 58, 2008.

[184] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[185] K. P. Murphy, "Active learning of causal Bayes net structure," Tech. Rep., 2001.

[186] D. M. Chickering, "Learning equivalence classes of Bayesian-network structures," *Journal of Machine Learning Research*, vol. 2, pp. 445–498, Mar. 2002.

[187] J. Cheng, D. A. Bell, and W. Liu, "Learning belief networks from data: an information theory based approach," in *Proceedings of the Sixth International Conference on Information and Knowledge Management*, ser. CIKM '97, 1997, pp. 325–331.

[188] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu, "Learning Bayesian networks from data: An information-theory based approach," *Artificial Intelligence*, vol. 137, no. 1âĂŞ2, pp. 43 – 90, 2002.

[189] R. Scheines, P. Spirtes, and C. Glymour, "A qualitative approach to causal modeling," in *Qualitative Simulation Modeling and Analysis*, P. A. Fishwick and P. A. Luker, Eds., 1991, pp. 72–97.

[190] F. Eberhardt, "Almost optimal intervention sets for causal discovery," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, ser. UAI '08, 2008.

[191] F. Eberhardt and R. Scheines, "Interventions and causal inference," *Philosophy of Science*, pp. 74–981, 2007.

[192] F. Eberhardt, C. Glymour, and R. Scheines, "On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables," in *Proceedings of the Fifty-seventh Annual Conference on Uncertainty in Artificial Intelligence*, ser. UAI '05, 2005, pp. 178–184.

[193] M. J. Flores, A. E. Nicholson, A. Brunskill, K. B. Korb, and S. Mascaro, "Incorporating expert knowledge when learning bayesian network structure: A medical case study," *Artificial Intelligence in Medicine*, vol. 53, no. 3, pp. 181 – 204, 2011.

[194] H. Langseth and T. D. Nielsen, "Fusion of domain knowledge with data for structural learning in object oriented domains," *Journal of Machine Learning Research*, vol. 4, pp. 339–368, Dec. 2003.

[195] F. T. Sheldon, R. K. Abercrombie, and A. Mili, "Evaluating security controls based on key performance indicators and stakeholder mission," in *Proceedings of the 4th Annual Workshop on Cyber Security and Information Intelligence Research: Developing Strategies to Meet the Cyber Security and Information Intelligence Challenges Ahead*, ser. CSIIRW '08, 2008, pp. 41:1–41:11.

[196] O. Junior and R. Rabelo, "A KPI model for logistics partners' search and suggestion to create virtual organisations," *International Journal of Network Virtual Organ*, vol. 12, no. 2, pp. 149–177, May 2013.

[197] A. Maté, J. Trujillo, and J. Mylopoulos, "Conceptualizing and specifying key performance indicators in business strategy models," in *Proceedings of the 2012 Conference of the Center for Advanced Studies on Collaborative Research*, 2012, pp. 102–115.

[198] J. Cai, X. Liu, Z. Xiao, and J. Liu, "Improving supply chain performance management: A systematic approach to analyzing iterative KPI accomplishment," *Decision Support Systems*, vol. 46, no. 2, pp. 512–521, Jan. 2009.

[199] T. Schulz, L. Radliński, T. Gorges, and W. Rosenstiel, "Defect cost flow model: A Bayesian network for predicting defect correction effort," in *Proceedings of the 6th International Conference on Predictive Models in Software Engineering*, 2010, pp. 16:1–16:11.

[200] D. Geiger and D. Heckerman, "Learning Gaussian networks," in *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, ser. UAI'94, 1994, pp. 235–243.

[201] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, pp. 309–347, 1992.

[202] N. Friedman and M. Goldszmidt, "Learning Bayesian networks with local structure," in *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, ser. UAI '96, 1996, pp. 252–262.

[203] D. M. Chickering, D. Heckerman, and C. Meek, "A Bayesian approach to learning Bayesian networks with local structure," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'97, 1997, pp. 80–89.

[204] W. Buntine, "Theory refinement on Bayesian networks," in *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI '91, 1991, pp. 52 – 60.

[205] L. M. De Campos, "A scoring function for learning Bayesian networks based on mutual information and conditional independence tests," *Journal of Machine Learning Research*, vol. 7, pp. 2149–2187, Dec. 2006.

[206] T. Roos, T. Silander, P. Kontkanen, and P. Myllymaki, "Bayesian network structure learning using factorized NML universal models," in *Information Theory and Applications Workshop*. IEEE, 2008.

[207] R. Mahdi and J. Mezey, "Sub-local constraint-based learning of Bayesian networks using a joint dependence criterion," *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1563–1603, Jan. 2013.

[208] C. P. De Campos, Z. Zeng, and Q. Ji, "Structure learning of Bayesian networks using constraints," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09, 2009, pp. 113–120.

[209] T. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, ser. UAI '90, 1991, pp. 255–270.

[210] S. Tong and D. Koller, "Active learning for structure in bayesian networks," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 863–869.

[211] M. Teyssier, "Ordering-based search: A simple and effective algorithm for learning Bayesian networks," in *Proceedings of the Fifty-seventh Annual Conference on Uncertainty in Artificial Intelligence*, ser. UAI '05, 2005, pp. 584–590.

[212] M. Schmidt, A. Niculescu-Mizil, and K. Murphy, "Learning graphical model structure using l1-regularization paths," in *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, ser. AAAI'07, 2007, pp. 1278–1283.

[213] A. Gelman, C. Robert, N. Chopin, and J. Rousseau, *Bayesian Data Analysis*. Chapman and Hall, 1995.

[214] D. Luenberger, *Optimization by Vector Space Methods*. New York, NY, USA: John Wiley & Sons, Inc., 1969.

[215] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[216] D. Heckerman, "A tutorial on learning with Bayesian networks," in *Learning in Graphical Models*, M. I. Jordan, Ed., 1999, pp. 301–354.

[217] G. Iveta, "Human resources key performance indicators," *Journal of Competitiveness*, vol. 4, no. 1, pp. 117 – 128, 2012.

[218] B. Becker, M. Huselid, and D. Ulrich, *The HR scorecard: linking people, strategy and performance*. Harvard Business Review Press, 2001.

[219] H. Fonteijn, M. Modat, M. Clarkson, J. Barnes, M. Lehmann, N. Hobbs, R. Scahill, S. Tabrizi, S. Ourselin, N. Fox, and D. Alexander, "An event-based model for disease progression and its application in familial Alzheimer's disease and huntington's disease," *Neuroimage*, vol. 60, no. 3, pp. 1880–9, 2012.

[220] M. West, P. Coleman, D. Flood, and J. Troncoso, "Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease," *The Lancet*, vol. 344, no. 8925, pp. 769 – 772, 1994.

[221] Y. Mu and F. Gage, "Adult hippocampal neurogenesis and its role in Alzheimer's disease," *Molecular Neurodegeneration*, vol. 6, no. 85, 2011.

[222] M. Albers, G. Gilmore, and J. Kaye, "At the interface of sensory and motor dysfunctions and Alzheimer's Disease." *Alzheimer's and dementiaâĂŕ: the journal of the Alzheimer's Association*, vol. 11, no. 1, pp. 70–98, 2015.

[223] D. Devanand, X. Liu, M. Tabert, G. Pradhaban, K. Cuasay, and K. Bell, "Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease." *Biological Psychiatry*, vol. 64, no. 10, pp. 871–879, 2008.

[224] K. Possin, "Visual spatial cognition in neurodegenerative disease," *Neurocase*, vol. 16, no. 6, pp. 466–487, 2010.

[225] P. Wang, B. Zhou, H. Yao, Y. Zhan, Z. Zhang, Y. Cui, K. Xu, J. Ma, L. Wang, N. An, X. Zhang, Y. Liu, and T. Jiang, "Aberrant intra- and inter-network connectivity architectures in alzheimerâĂŹs disease and mild cognitive impairment." *Scientific Reports*, vol. 5, no. 14824, 2015.

[226] V. Jelic, S. Johansson, O. Almkvist, P. Shigeta, M. Julin, A. Nordberg, B. Winblad, and L. Wahlund, "Quantitative electroencephalography in mild cognitive impairment: longitudinal changes and possible prediction of Alzheimer's disease." *Neurobiology of Aging*, vol. 21, no. 4, pp. 533 – 540, 2000.

[227] K. Vlcek and J. Laczo, "Neural correlates of spatial navigation changes in mild cognitive impairment and alzheimerâĂŹs disease." *Frontiers in Behavioral Neuroscience.*, vol. 8, no. 89, 2014.

[228] U. Lindenberger and P. Baltes, "Sensory functioning and intelligence in old age: a strong connection." *Psychology and Aging*, vol. 9, no. 3, pp. 339–55, 1994.

[229] R. Gould, B. Arroyo, R. Brown, A. Owen, E. Bullmore, and R. Howard, "Brain mechanisms of successful compensation during learning in alzheimer disease," *Neurology*, vol. 67, no. 1, 2006.

[230] N. Friedman and D. Koller, "Being Bayesian about network structure," in *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI '00, 2000, pp. 201–210.