Optimizing Personalized Treatment Selection for Partially Observable Chronic Conditions

Jue Gong

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Shan Liu, Chair

Shuai Huang

Archis Ghate

Program Authorized to Offer Degree:
Industrial & Systems Engineering

University of Washington

Abstract

Optimizing Personalized Treatment Selection for Partially Observable Chronic Conditions

Jue Gong

Chair of the Supervisory Committee:
Assistant Professor Shan Liu
Industrial & Systems Engineering

For many chronic diseases, an individual patient may experience a wide variety of progression pathways. Personalized medicine needs tools to predict the trajectory of an individual patient's disease progression, which can in turn enable clinicians to optimize the sequence of treatments. The objective of this thesis is to design artificial intelligence methods to support clinicians in making smart treatment selections in chronic disease care. To achieve this objective, we develop algorithms optimized for an individual patients' demographic profiles, past medical history, and response to current treatment by utilizing electronic health record (EHR) data of a large population. Developing a personalized treatment plan is a difficult sequential decision-making problem that seeks to improve overall health outcome, efficiency, and reduce unnecessary cost. One challenge is to understand the complex disease progression in a heterogeneous population. Another challenge is the lack of adaptive treatment strategies for diseases with partially observable health states.

We develop innovative methodologies for personalized treatment selection to mitigate these challenges. First, we model the heterogeneity in disease trajectories of a population by detecting the subtypes of a chronic disease from longitudinal treatment data using an artificial neural network. Then we propose a framework called the partially observable collaborative model (POCM), to learn the individual disease progression model under various treatment options when the true health state is hidden to the decision maker. Next, utilizing the

learned individual models, a personalized treatment plan can be derived by solving a partially observable Markov decision process (POMDP). We further extend this framework to mitigate the risk of reduced performance of POMDPs with uncertainty in transition dynamics by finding robust policies.

Mental health is an understudied disease area that may greatly benefit from optimization in personalized medicine. Using simulated data informed by the Mental Health Research Network's EHR, we apply the proposed methods to simulate the treatment of chronic depression. The contributions of this thesis include a novel framework for learning personalized disease progression model, a robust and adaptive treatment selection method, and an application on chronic depression treatment optimization. This thesis helps to advance the development of artificial intelligent decision support tools for chronic disease care.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I wish to express my deep and sincere appreciation to my advisor, Dr. Shan Liu, for her generous support, endless patience, and insightful guidance throughout my Ph.D study at University of Washington. She introduced me to medical decision-making, offered me a lot of great opportunities and resources, and guided me when I got stuck. I could not have imaged having a better advisor and mentor for my Ph.D. study.

In addition, for this dissertation I would like to thank my reading committee members: Dr. Archis Ghate, Dr. Shuai Huang, and my GSR Dr. Samual Burden, who provided valuable comments and suggestions for both my dissertation and my future career.

The members of the Liu's research group 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. Ying Lin, Dr. Yan Jin, Dr. Cao Xiao, Ting-Yu Ho, Qiang Meng, Tianshu Feng, Zhanlin Liu, and so on.

Thanks to my collaborator, Dr. Gregory E. Simon, from Kaiser Permanente Washington Health Research Institute. It is my great preseasure to work with him and have his advises and suggestions.

I would like to thank all other people who helped me at UW.

Last but not least, I would like to thank my family for their love and encouragement, my wife Shuning Tong, and parents Weidong Gong and Chunmei Zhang. Their faithful support and understanding during all stages of my study is very much appreciated.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Motivation

Selecting effective treatments for an individual patients in a heterogeneous population where their health conditions follow complex evolution is a challenging problem in many chronic disease applications. Chronic diseases require sequences of treatments over time to address the changing characteristics of the disease and the patient. The decision making in such medical problems must balance the potential benefit and harm of treatment. Take chronic depression for example, in the United States alone, an estimated 7% of all adults had at least one major depressive episode in the year 2015 [1]. Typical treatment options, including antidepressant medication or psychotherapy, have long durations, high cost, and risky side effects. Recent research demonstrated the effectiveness of antidepressants to treat depression, but it is still unclear when they should be used or who should get them [2]. This requires a treatment selection plan that selects the treatment options over time in order to improve a patient's depression symptoms and reduce side-effect risk and unnecessary cost.

Currently, treatments are often selected by physicians and other healthcare providers following guidelines and expert consensus documents informed at the population level. In medical decision-making research, researchers have developed several mathematical models for the optimal treatment selection problem. The Markov decision processes (MDP) are widely used to model sequential decision-making for many chronic diseases [3, 4, 5, 6, 7], in which the disease progression is represented using state transition probabilities estimated from population-level data. However, treatment strategies designed at the population level might not be optimal for an individual patient. Personalized treatment strategies are tailored to an individual patient based on his/her demographic characteristics, treatment history, and

response to treatment.

In recent years, widely implemented electronic health record (EHR) systems, expanded use of clinical decision-support tools, telemedicine, and mobile health apps are moving clinical care into an era of personalized medicine or precision medicine. However, mathematical models of decision support systems for personalized treatment using EHR are still lacking for many chronic diseases. We notice several challenges in developing such models. The first challenge is to understand the heterogeneity of disease progression, i.e., find a representation of the individual disease progression model. The second challenge is to develop an adaptive treatment strategy of selecting optimal treatments tailored to individual patients over time, in order to support treatment planning decisions of healthcare providers. Since the estimation of the disease progression is commonly inaccurate due to the noisy and sparse data available, we face a third challenge of finding robust treatment strategies in a treatment process parameterized with uncertain disease transition dynamics.

This thesis focuses on developing a decision-support framework to mitigate these challenges. We provide the objectives and organization of this thesis in the following sections.

1.2 Research Objectives

The objective of this research is to develop artificial intelligence methods to support clinicians in making smart treatment selections for chronic diseases, by designing algorithms optimized for individual patient's characteristics and utilizing treatment history data of a large population. This objective can be achieved by completing the following tasks:

1. Apply machine learning methods to characterize the heterogeneity of chronic depression progression through disease trajectory modeling and pattern discovery.

2. Develop optimization algorithms to learn latent parameters of individual disease progression models, and to make a sequence of optimal treatment decisions based on estimated individual disease dynamics over time.

3. Formulate a robust optimization model for adaptive sequential treatment to maximize health outcomes when the individual progression model is inaccurately estimated from the past treatment records.

This thesis achieves these goals to build a solid foundation for the future development of personalized treatment selection in chronic disease care.

1.3 Organization of Thesis

This thesis is outlined as follows: We first provide an analysis of the heterogeneity of disease progression in a population from the trajectories of chronic depression data in Chapter 2. We discovered several subtypes of depression in the trend of depression symptom progression. This finding becomes a fundamental assumption in later chapters, that there are several basis models of the disease progression in the population. Next, we developed a framework to represent the individual disease progression as a linear combination of these basis models under partially observability assumptions. We further developed a dynamic treatment selection algorithm based on the individual model learning in Chapter 3. Chapter 4 discusses some future directions of extending the model in the previous chapters to robust optimization.

Chapter 2 focuses on effectively extracting latent structures from the trajectory data of depressive symptoms and discovering subtypes of depression. We analyze a depression treatment population's EHR. We first estimate correlations between the depression symptom and the suicide ideation. We discover patterns in trajectories of depressive symptoms using artificial neural networks (ANN) and detect five patterns in the depression trajectories. These patterns can be interpreted as five subtypes of depression such that each individual patient in the same subtype has a similar progression trend. This work contributes to the emerging field of personalized medicine by providing a tool for predicting the measurements of depressive symptom from associative measurements or the average trend of the associated subtype.

Chapter 3 describes a framework of estimating the individualized chronic disease progression model, the Partially Observable Collaborative Model (POCM), to capture the individual

variations in a heterogeneous population. We provide the algorithm for learning the parameters of POCM, which is an iterative updating rule for solving a nonconvex optimization problem. We also discuss how the treatment strategy based on the Partially Observable Markov Decision Process (POMDP) can be tailored for individuals by utilizing the model learned from the POCM. The result shows the POCM can give a better estimation of personal disease progression model than the traditional approach of estimating the hidden Markov model (HMM) when there are strong subgroup structures in disease progression.

In Chapter 4, we extend the personalized treatment model to a robust optimization problem. We introduce two types of robust POMDPs which generalizes a standard POMDP by allowing uncertainties in the reward function and transition probabilities. The first model assumes parameters bounded with linear constraints, and the second assumes the transition probabilities follow Dirichlet distribution. We develop the value iteration algorithm for each type of the POMDP to find optimal solutions when the worse case (or nearly the worst case) of the uncertain transition probability is considered. Finally, we illustrate the effectiveness of our approach using a case study in designing personalized treatment plan for chronic depression in a heterogeneous population.

This thesis contributes to the methodologies of disease progression modeling by proposing an algorithm that combines the MDP and collaborative model in the case of unobservable disease states. Another contribution is proposing robust optimization in personalized treatment selection to the subgroup structure. We apply these methods to chronic depression care, which enhances the understanding of the subgroup structure of depression progression and the development of artificial intelligent systems to support personalized treatment of depression.

Chapter 2

MACHINE LEARNING DISCOVERY OF LONGITUDINAL PATTERNS OF DEPRESSION AND SUICIDAL IDEATION

This chapter focuses on the discovery of heterogeneous patterns of depression trajectories. We analyze a depression treatment population's electronic health record (EHR). Depression is often accompanied by thoughts of self-harm which are a strong predictor of subsequent suicide attempt and suicide death. Few empirical data are available regarding the temporal correlation between depression symptoms and suicidal ideation. We first estimate correlations between the depression symptom and the suicide ideation. We discover patterns in trajectories of depressive symptoms using artificial neural networks (ANN) and detect five patterns in the depression trajectories. These patterns can be interpreted as 5 subtypes of depression such that each individual patient in the same subtype has a similar progression trend.

2.1 Introduction

Depression is a complex and dynamic mental disorder that is among the leading causes of disability worldwide. In the United States alone, an estimated 7% of all adults had at least one major depressive episode in the year 2015 [1]. A common practice in the prevention and treatment of chronic depression is to identify risk factors, including various types of longitudinal predictors such as depression symptoms and suicidal ideation. Depression is often accompanied by the thought of self-harm, which is believed to be a predictor of subsequent suicide attempt and suicide death. Simon et al. [8] found that the risk of suicide attempt increases with a higher frequency of thoughts on self-harm. A meta-analysis of longitudinal studies on suicide risk factors, Ribeiro et al. [9] showed that the most common outcome

(48%) was suicide attempt after self-injurious thoughts and behavior, but the effect was considerably weaker than anticipated. Franklin et al. [10] pointed out that traditional methods of identifying risk factors of self-injurious thoughts are limited, and they highlighted the need for machine learning-based algorithm to examine suicidal ideation risk.

Furthermore, there is a claim that improvement in depression may be accompanied by an increase in suicidal ideation [11]. Few empirical studies are available regarding the temporal association between depression and suicidal ideation, which is further complicated by the heterogeneity in depression symptoms' trajectories of individuals under treatment (i.e., subtypes of depression trajectory patterns). Additionally, most available data regarding temporal associations between suicidal ideation and other symptoms of depression are derived from clinical trials rather than patients treated in community practices [11, 12, 13].

This chapter focuses on the temporal correlation between depression and suicidal ideation for a heterogeneous patient population, in which we assume there are individual patterns in the depression progression and suicide ideation trajectories. The objective of this study is to discover heterogeneous patterns of depression trajectories and investigate the traditional concern that suicidal ideation may increase during a period of depression improvement by using the electronic health record (EHR) data. This study does not try to predict suicide death or suicide attempts from depression severity measures using EHR.

Depression severity is measured by the Patient Health Questionnaire (PHQ)-9; a self-administered questionnaire that includes 9 multiple-choice questions to assess the frequency of depressive symptoms within the previous two weeks [14]. The total score without the 9th question, the PHQ-8 score, has a similar ability to predict major depressive disorder compared to the PHQ-9 score [14]. The 9th question (Item 9) which asks about suicidal ideation is analyzed as a separate symptom in this study. Previous research has shown that Item 9 identifies patients at increased risk of suicide attempt [8]. Mittal et al. [11] questioned the belief in the clinical community that patients with depression are at increased risk of suicide as they begin to recover and their motivation return. They found such claim remains to be substantiated. These studies have not investigated the correlation between

PHQ-8 and Item 9 scores over time, nor have they examined potential heterogeneity in the trajectory patterns of depressive symptoms among patients. The relationship between depression improvement and changes in suicidal ideation has not been supported by analysis of depression and suicidal ideation trajectory data.

EHR containing depression records of a large number of individuals over extended periods can be useful in discovering heterogeneity in depression trajectories. However, conventional statistical methods on modeling trajectories often fail to accommodate the statistical challenges of longitudinal EHR data, such as data sparsity and irregularity [15, 16]. Research on modeling depression trajectories based on longitudinal datasets have emerged since 2005 [17]. There are wide variations in the time intervals between consecutive symptom observations, with most studies reporting six months or longer. The long gap in observations is problematic because depression assessment questionnaire such as the PHQ-9 asks patients to report symptoms in the past two weeks. Therefore, sparse measurements may not represent the true underlying trajectories [13]. In this study, we address this issue by selecting patients with frequent records of observations. We then use the Gaussian process regression (GPR) method to transform the irregular and sparse data of longitudinal PHQ scores into a continuous function of time [18]. We develop a data-driven method based on Artificial Neural Networks (ANN) to extract subtypes of depression with respect to the local progression patterns from collected trajectory information and investigate the question about synchrony of change. One benefit of ANN is its ability to transform longitudinal data of large volume to latent variables of much smaller size as a high-level abstraction of the original data [19]. In addition, our model of latent structure learning discovers patterns in the trajectory data without prior knowledge on the subtypes; this is a form of unsupervised learning. In this model, we reconstruct the depression trajectory of individual patient by learning the patterns which are a set of latent basis trajectories similar to latent factor models; then, the subtypes of depression are discovered by classifying patients based on these learned latent patterns [20]. Therefore, we can identify the trajectories of overall depression severity and trajectories of suicidal ideation and examine their synchrony.

This study contributes to the understanding of depression progression heterogeneity by applying an ANN method for subtype detection and estimating the temporal correlation between depressive symptoms and suicide ideation trajectories using cross-correlation analysis. To the best of our knowledge, this chapter provides the first time-series study of trajectories of PHQ-8 and Item 9 for chronic depression patients. This work consists of three parts: the first is estimating the temporal correlation between depression symptoms; the second is learning latent patterns from trajectories of depressive symptoms in the population; and the third is discovering subtypes of depression patients from the latent structure such that patients in each subtype have similar trajectory patterns.

This chapter is organized as follows: Section 2.2 provides a brief description of the data, as well as introduces multiple time series analysis, the ANN for latent feature learning and the clustering methods for subtype discovery. Section 2.3 presents numerical results from a depression treatment population's EHR dataset. Section 2.4 provides further insights into the clinical utility of these methods and discusses limitations and future research directions.

2.2 Methods

2.2.1 Data description

Data are drawn from the EHR of four health systems participating in the Mental Health Research Network (Health Partners, and the Colorado, Washington, and Southern California regions of Kaiser Permanente), totaling 1.2 million observations [21, 22, 23, 24]. All four health systems recommend routine use of the PHQ-9 questionnaire at all specialty mental health visits and at all primary care visits including diagnosis or treatment of depression. The dataset includes individuals' longitudinal PHQ-9 measures between the years 2007 and 2012, and are linked to relative time between measurements, type of providers (primary care, specialist, mental health) where the questionnaire was conducted, individuals' age, sex, race/ethnicity, diagnosis and treatment status, and the Charlson Comorbidity Index (a standard indicator of chronic disease severity). In the EHR data, 9,306 individuals have

Table 2.1: Summary statistics of the 610 on-going treatment patients.

Age count (%)	
18-29	63 (10.3%)
30-44	169 (27.7%)
45-64	292 (47.9%)
65+	86 (14.1%)
Sex count (%)	
Female	414 (67.9%)
Male	196 (32.1%)
Average PHQ-9	
Min	0.10
25% Quantile	8.20
Median	12.17
75% Quantile	16.00
Max	25.22
Average PHQ-8	
Min	0
25% Quantile	8.01
Median	11.59
75% Quantile	15.36
Max	23.88
Average Item 9	
Min	0.00
25% Quantile	0.00
Median	0.27
75% Quantile	0.83
Max	2.75

received ongoing depression treatment (defined as either receiving antidepressant medication or making specialty mental health visits for depression during the past 6 months). We select

a subset containing 610 patients (62% age 45 and older; 68% female) with at least six PHQ-9 scores recorded during twenty consecutive two-week periods from the original sample. The summary statistics of this group are listed in Table 2.1. The PHQ-9 records do not have equal intervals between two consecutive data points; we transform the PHQ-9 scores of each patient into a smooth curve using Gaussian Process Regression (GPR) (details in Appendix A.1.2), and then we sample the data at equal time interval for further analysis. Individual scores for each PHQ-9 question are used to compute the PHQ-8 and Item 9 scores. We are interested in the correlation between the trajectory of overall depression severity (as indicated by the PHQ-8) and the trajectory of suicidal ideation (as indicated by Item 9), which can be described by the cross-correlation function (2.1) in Section 2.2.2.

2.2.2 Cross-correlation of multiple time series

We investigate the temporal correlations between PHQ-8 and Items 9 score using the cross-correlation function (CCF). CCF is a measure of similarity of two series as a function of a relative displacement (i.e. a lag of k time units). For two time series x_t and y_t which are observed from stationary stochastic processes, their sample CCF $\hat{\rho}(k)$ is defined as

$$\hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\sigma}_x \hat{\sigma}_y}, \tag{2.1}$$

where N is the number of observations, $\hat{\sigma}_x$, $\hat{\sigma}_y$ are the sample deviations of the processes, $\hat{\gamma}_{xy}(k)$ is the sample cross-covariance at lag k, defined as

$$\hat{\gamma}_{xy}(k) = \begin{cases} \frac{1}{N-k} \sum_{t=1}^{N-k} (x_{t+k} - \bar{x})(y_t - \bar{y}), & k \geq 0 \\ \frac{1}{N-|k|} \sum_{t=1}^{N-|k|} (x_t - \bar{x})(y_{t+|k|} - \bar{y}), & k < 0 \end{cases} \tag{2.2}$$

where $\bar{x}$, $\bar{y}$ are the estimated mean values of the time series [25]. The range of CCF is between -1 to 1, with values close to 1 representing high similarity, and values close to 0 representing low similarity between two time series. Incorporating the lag function, for example, if CCF is close to 1 at lag $k < 0$, then the shape of series x_t is similar to y_t at a time $|k|$ units ahead. See Figure 2.1. If the CCF is close to 1 at lag $k = 0$, then there is a high probability that the 2 time series evolve simultaneously.

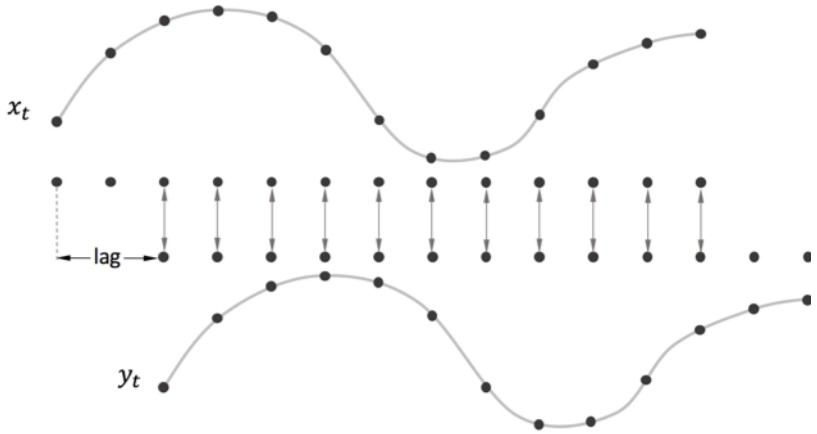


Figure 2.1: Illustration of CCF estimation, the correlation $\hat{\rho}_{xy}(k)$ is given by a mean over products of two zero-mean observation series with a lag of k units apart. Here the lag $k < 0$ and time series x_t shows high similarity to y_t at time $|k|$ units ahead, indicating that x_t leads y_t with $|k|$ units.

2.2.3 Artificial Neural Network for latent feature detection

In this section, we use autoencoder, a type of artificial neural network, to discover the latent structure of depression trajectories in a population. [26, 27] Neural network for subtype discovery has been proposed on other diseases including gout and leukemia Lasko et al. [26]. We assume that the trajectory of a patient can be decomposed as a linear combination of a set of basis trajectories. We treat these basis trajectories as latent patterns of the population. Each basis trajectory stands for a trend in the depression severity, such as an increasing trend in severity over time, or an increasing and then decreasing trend. The value of the coefficient of each pattern, called the activation, stands for how strongly the basis trajectory is represented in the patient's trajectory. We further use these activations as features to extract a measure of similarities among patients. The latent structure consists of the latent patterns of basis trajectories and the feature of activation of each patient in the population.

More specifically, an autoencoder is a particular neural network structure with three layers: (1) an input layer of M nodes representing the fitted trajectories; (2) an latent layer of H

nodes representing the hidden features by transforming the input (the encoder); and (3) the output layer with M nodes representing a reconstruction of the input data by transforming the latent layer (the decoder) [28, 29]. The encoder typically transforms an input trajectory represented as a vector $\mathbf{m} \in \mathbb{R}^M$ into the hidden or transformed representation $\mathbf{h} \in \mathbb{R}^H$ by

$$\mathbf{h} = u(\mathbf{W}\mathbf{m} + \mathbf{b}) \tag{2.3}$$

where M is the length of input vector $\mathbf{m}$, i.e., the number of observations, H is the size of latent features, i.e., basis trajectories, the matrix $\mathbf{W} \in \mathbb{R}^{H \times M}$ is a matrix of latent features, the vector $\mathbf{b} \in \mathbb{R}^H$ is a vector of learned bias offsets, and u is a pre-specified nonlinear function such as the logistic sigmoid function $u(x) = 1/(1+\exp(-x))$. The decoder computes a reconstruction $\hat{\mathbf{m}}$ with the function $\hat{\mathbf{m}} = \mathbf{W}'\mathbf{h} + \mathbf{b}'$, where the function v is the identity function for continuous input data. The values in $\mathbf{W}'$ can either be learned separately or tied such that $\mathbf{W}' = \mathbf{W}^\top$. The structure of the autoencoder is illustrated in Figure 2.2.

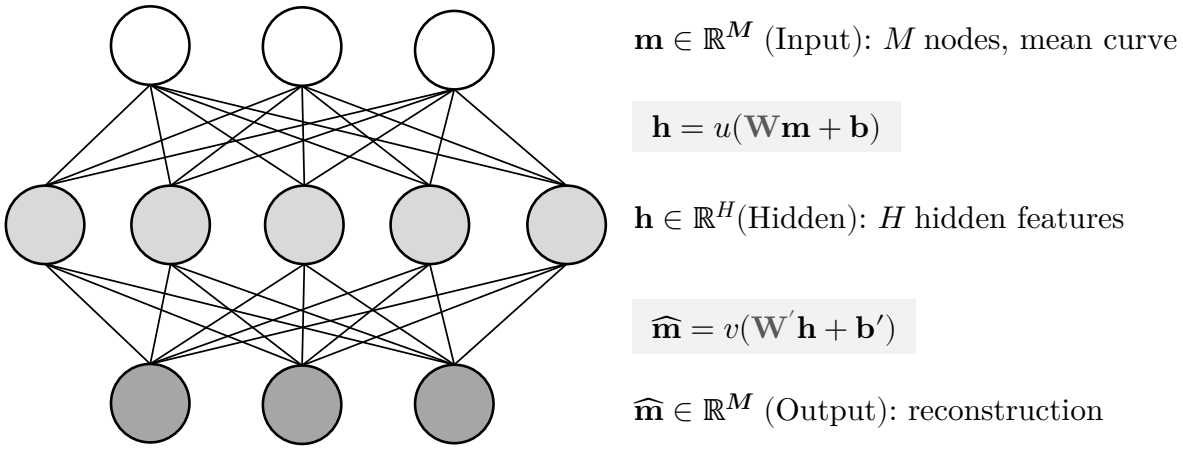


Figure 2.2: The structure of the autoencoder we used to detect the latent feature of the depression.

We can train the parameters of the autoencoder by the following optimization problem

$$J = \sum_{j=1}^{N} \sum_{i=1}^{M} (\hat{m}_i^j - m_i^j)^2 + \lambda \sum_{k,l} \left(W_{kl}^2 + W_{lk}'^2 \right) + \beta \sum_{i=1}^{H} D(\eta, \hat{\eta}_i) \tag{2.4}$$

The first term is the squared-error loss between the input data (the depressive symptom trajectory) and the reconstructed data (the set of basic trajectories), where the superscript j denotes the j-th input trajectory (out of N total inputs), and the subscript i denotes the i-th data point of each input trajectory. The second term is the regularization term to ensure the elements of $\mathbf{W}$ are small. The third term is the sparsity term to make the activation $\mathbf{h}$ sparse (i.e., most elements are near zero), so that the input trajectory is represented by only a small number of reconstructed trajectories. The sparsity measure $\hat{\eta}_i$ is the average activation of the i-th hidden node.

$$\hat{\eta}_i = \frac{1}{N} \sum_{j=1}^{N} h_i^j$$

The function D is the Kullback-Leibler divergence [30]

$$D(\eta, \hat{\eta}) = \eta \log \frac{\eta}{\hat{\eta}} + (1 - \eta) \log \frac{1 - \eta}{1 - \hat{\eta}} \tag{2.5}$$

that forces all $\hat{\eta}_i$ to be close to the sparsity target η, and λ and β are tuning parameters. This cost function (2.4) produces a sparse autoencoder [28]. Each row of $\mathbf{W}$ after training is a vector representing one of the learned patterns or basis trajectories that is the objective of this step. The value of $\mathbf{W}'$ are not used beyond training. The $\mathbf{W}$ forms a compact set of basis trajectories that can be linearly combined to represent the input sample trajectory. Given a set of learned patterns $\mathbf{W}_i$, any data vector $\mathbf{m}$, including a previously unseen one, can be represented in terms of these patterns using Eq. (2.3). The resulting element $\mathbf{h}_i$ is the activation of pattern $\mathbf{W}_i$ for the input $\mathbf{m}$, indicating how strongly $\mathbf{W}_i$ represented in $\mathbf{m}$.

There is a unique activation vector for each patient to represent the similarity of this patient to each basis group. We can cluster the patients using the activation vector to discover the subtypes of depression. We use the k-means clustering algorithm [31]. The number of clusters was decided by comparing the inertia, the sum of squared distances to the closest centroid for all observations.

2.3 Results

2.3.1 Cross-correlation between measurements

We examined the temporal similarity between two depressive symptom measurements for each patient, PHQ-8 and Item 9. We first transformed the irregular records of each patient's PHQ-8 and Item 9 scores into continuous trajectories using GPR, and sampled the series with a period of 2 weeks. We used grid-based methods to search for the optimal parameters in the regression with respect to maximizing the mean of marginal likelihood over all patients (the result is listed in Appendix A.1.2). Two sets of parameters are used for the PHQ-8 scores and Item 9.

We estimated the cross-correlation between the transformed trajectories of PHQ-8 and Item 9 for each patient. From Section 2.2.2, it is easy to see that if one trajectory consists of all zeros, then the cross-correlation between this trajectory and any other trajectory is zero. We provide three examples of patient's CCF between PHQ-8 and Item 9 in Figure 2.3. In Figure 2.3(a), this patient's CCF between PHQ-8 and Item 9 peaks at lag $k = 0$, which indicates that the two trajectories are most similar with no lag, thus her PHQ-8 and Item 9 follow the same trend at the same time. In Figure 2.3(b), this patient's CCF between PHQ-8 and Item 9 peaks at lag $k = 1$; this positive lag means that her PHQ-8 and Item 9 are the most similar if PHQ-8 moves 1 unit to the left; this can be interpreted as her Item 9 and PHQ-8 change with a most similar trend such that Item 9 leads 1 unit of time period (i.e., 2 weeks). In Figure 2.3(c), this patient's CCF between PHQ-8 and Item 9 peaks at lag $k = -2$; this negative lag means his PHQ-8 and Item 9 is the most similar if PHQ-8 moves 2 units to the right.

We then investigated the distribution of the lag at maximum CCF among the 394 patients with nonzero CCFs. Results are shown in Figure 2.4. For PHQ-8 vs. Item 9, the majority of the patients have their PHQ-8 and Item 9 scores moving with the same trend simultaneously; other patients have their PHQ-8 score leading Item 9 score with some time delay, or vice-versa; the period of delay is roughly uniformly distributed.

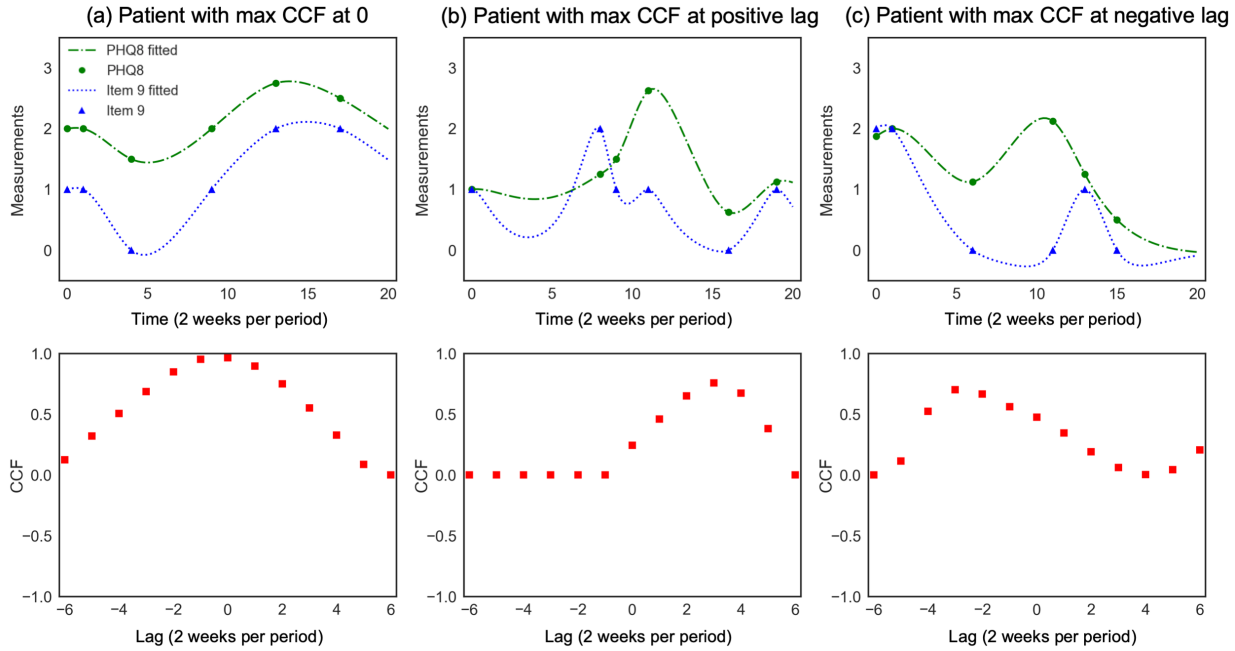


Figure 2.3: Three sample patients' CCFs between time series. The left column is the records of the two series (PHQ-8 and Item 9) for each patient and their fitted curve; the right column is the CCFs between PHQ-8 and Item 9. One unit of time is two weeks.

2.3.2 PHQ-8 and Item 9 changing patterns

To examine the belief that depression improvement and an increase in suicidal ideation may happen simultaneously as stated in Mittal et al. [11], we used Spearman's rank-order correlation to measure the strength and direction of association between the PHQ-8 and Item 9 changes. We considered both short-term effect from changes in PHQ-8 and Item 9 values between consecutive observations within 1 month, and long-term effects from changes in the two values with longer time windows. For the 610 subjects, we collected the following data for both PHQ-8 and Item 9 using unfitted observations in the EHR: (1) the rate of change of the PHQ-8 and the Item 9 scores within 1 month (i.e., the difference of the scores between two consecutive observations within 1 month divided by the period length); (2) the first observation (month 0); (3) the average of the records between month $t0.5$ and $t+0.5$ for

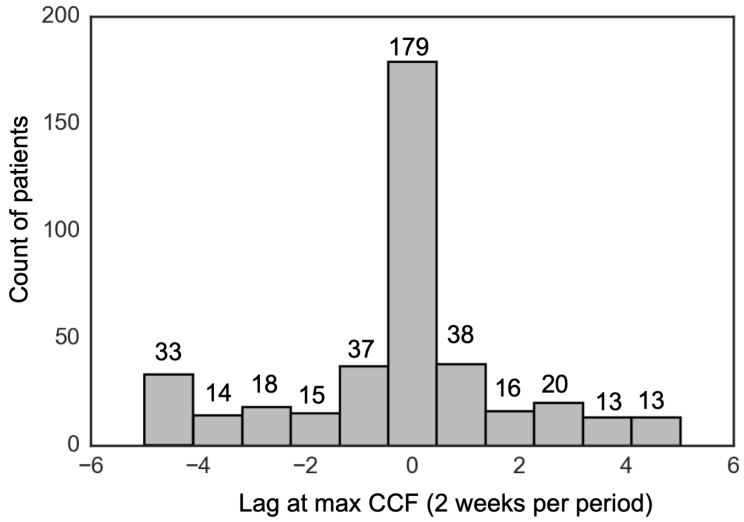


Figure 2.4: Histogram of the lag k at maximum CCF in the population, PHQ-8 vs. Item 9. We excluded 214 patients with zero CCF between PHQ-8 and Item 9 for all lags between -5 and 5 from this figure.

month $t = 3, 6, 9$. The PHQ-8 scores were converted to the average score of each question with a range of 0 to 3.

The short-term association is indicated by Spearman's rank-order correlation [32] calculated using the rate of change of the consecutive PHQ-8 and Item 9 values within 1 month. In Table 1, the result shows that the two scores have a positive monotonic relationship in the short term. We also conducted a linear regression using the same data, in which the result indicated a positive correlation (i.e., slope of the regression line is positive; see Figure 3). To examine the long-term effect, we computed the Spearman's rank-order correlation between the changes of PHQ-8 and Item 9 at month 3, 6, 9 from month 0, separately. The results also show strong positive correlations in all three cases (see Table 2.2 and Figure 2.5). Therefore, we found that in the majority of the cases in our EHR sample, patients' PHQ-8 and Item 9 scores tend to change in the same direction.

Furthermore, we investigated two patterns of PHQ-8 and Item 9 changing in opposite

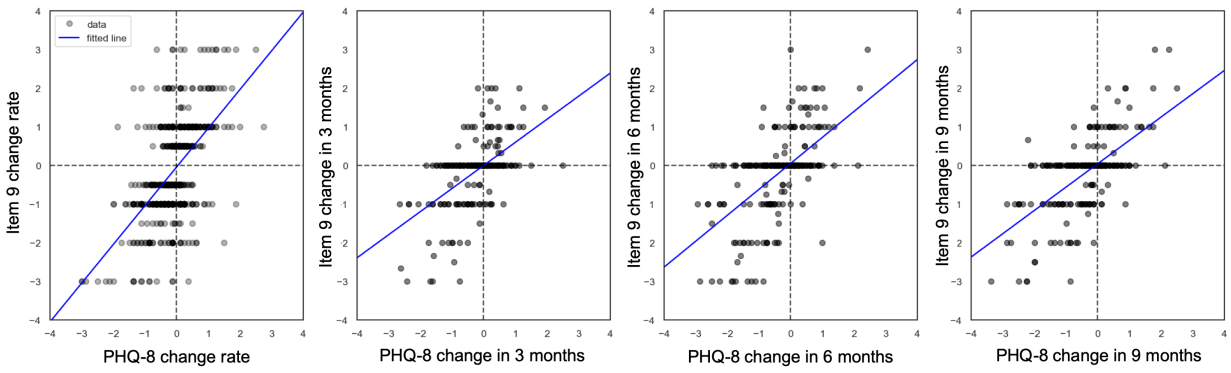


Figure 2.5: The rate of change of PHQ-8 score and Item 9 from all patients. The solid line is the result of linear regression.

		Spearman's rank-order correlation (p-value)	Slope of linear regression (R^2 value)
Short-term	within 1 month	0.52 ($p < 10^{-4}$)	1.00 ($R^2 = 0.31$)
Long-term	3 months	0.53 ($p < 10^{-4}$)	0.60 ($R^2 = 0.30$)
	6 months	0.57 ($p < 10^{-4}$)	0.67 ($R^2 = 0.34$)
	9 months	0.56 ($p < 10^{-4}$)	0.60 ($R^2 = 0.39$)

Table 2.2: Spearman's rank-order correlation and linear regression for PHQ-8 and Item 9.

directions using unfitted observations in the EHR dataset: (a) PHQ-8 increases and Item 9 decreases, and (b) PHQ-8 decreases and Item 9 increases (using unfitted observations in the EHR dataset). We define an "Item 9 change" as an increasing or decreasing score by at least 1 unit between observations, and a "PHQ-8 change" as a score increasing or decreasing by at least d units between observations, where d is the threshold. We tested $d = 2, 3, 4$ in this study. We split the population of 396 patients (Item 9 not all-zeros) into four mutually exclusive and collectively exhaustive subgroups (see Table 2.3). We recognized around 8% to 13% of the depression patients have experienced an increase in suicidal ideation during improvement of PHQ-8 scores, based on different thresholds of defining changes in PHQ-8

scores. Therefore, the claim in [11] is partially supported in our EHR study sample.

Table 2.3: Subgroup sizes for patients with various PHQ-8 and Item 9 changing patterns.

Threshold of PHQ-8 change	Number of patients in each subgroup			
	patients with pattern (a) only	patients with pattern (b) only	patients with both patterns	patients with no patterns
2	59 (14.9%)	53 (13.3%)	41 (10.4%)	243 (61.4%)
3	47 (11.9%)	41 (10.3%)	23 (5.8%)	285 (72.0%)
4	38 (9.6%)	31 (7.8%)	17 (4.3%)	310 (78.3%)

We perform the chi-square test of homogeneity on these 4 subgroups to test if they have significantly different distributions on certain categorical at a significance level of $\alpha = 0.05$, see Table 2.4. The features include:

(1) Age: 4 levels: $18 \sim 29, 30 \sim 44, 45 \sim 64, 65+$

(2) Sex: 2 levels: Male, Female

(3) Mean of Charlson Index: which is a measure of patients' comorbid conditions, 4 levels

0 : $0 \leq$ mean Charlson Index < 0.5

1 : $0.5 \leq$ mean Charlson Index < 1

2 : $1 \leq$ mean Charlson Index < 2

3 : $2 \leq$ mean Charlson Index

The test is repeated for different threshold for defining PHQ-8 changes (threshold $d = 2, 3, 4$). We found no significant differences in distributions for any categories at all thresholds. Par-

Table 2.4: The p-value of Chi-square Test on Homogeneity for various features.

Threshold for PHQ changes (d)	p-value of Chi-square Test on Homogeneity for features		
	Age	Sex	Mean of Charlson Index
2	0.6049	0.8085	0.4035
3	0.3997	0.5672	0.1657
4	0.0570	0.6500	0.1345

ticularly, we found that the distribution of age can be considered as different across the subgroups at a threshold of 4 units (a significance level of $\alpha = 0.06$).

2.3.3 Subtype discovery in depression trajectory patterns

Next, we discovered patterns in trajectories of depressive symptoms using artificial neural networks. The input of the AutoEncoder is the transformed observations of the PHQ-8 and Item 9 records of each patient (input size $M = 20$), which are further normalized to zero mean and unit variance. By performing cross validation in training the AutoEncoder, we found that the cost function of the autoencoder was minimized using the stochastic gradient descent method [28]. The hyperparameters η, and λ and β are determined using a set of cross validation analysis. We randomly partition the 610 patients into five subsamples with equal size. A single instance of cross-validation (CV) is to use one of the five subsamples as the validation data, and the rest four subsamples as training data. The cross-validation process is repeated 5 times with each of the five subsamples used exactly once as the validation data. The criterion of validation is the mean squared error (MSE) between the original trajectory and the reconstructed trajectory. For example, one patient from the validation set has the original trajectory $\mathbf{m} \in \mathbb{R}^{21}$, and the reconstructed trajectory is $\hat{\mathbf{m}} = v\left(\mathbf{W}^\top\left(u(\mathbf{W}\mathbf{m} + \mathbf{b})\right) + \mathbf{b}'\right)$, where $\mathbf{W}, \mathbf{b}, \mathbf{b}'$ are learned from the training set, then

the MSE is defined as

$$\varepsilon = \frac{1}{21}(\mathbf{m} - \hat{\mathbf{m}})^\top(\mathbf{m} - \hat{\mathbf{m}}). \tag{2.6}$$

The five results from the folds can then be averaged to produce a single estimation. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once.

For the PHQ-8 trajectories, we perform the cross-validation on different set of hyper-parameters, including the number of nodes in hidden layers (H), the three regularization parameters (λ, β, ρ). The models with the lower MSE in CV is preferred. The result shows the values of β and ρ have negligible effect. We use $\beta = 0.1$ and $\rho = 0.5$. Next, we focus on the selection of number of nodes in the hidden layers (H) and regularization parameters λ (on scale of $\mathbf{W}$). See Figure 2.6. Smaller H and larger λ generally gives better model. There are two other considerations before choosing the parameter: (1) Too small H does not give sufficient variety of hidden features in term of the shape of the hidden pattern. (2) For λ greater than 5, increasing λ dose not improve the performance greatly. Therefore, we choose $H = 25$ and $\lambda = 4$.

We found that the latent layer with very small size is not able to contain a sufficient variety of hidden features in term of the shape of the hidden pattern. On the other hand, a very large latent layer does result in large amount of repeated hidden features. We choose the number of latent nodes to be 25, based on both the results from cross-validation, and general rules of designing neural network structure design found in literature, such as "the number of latent units no more than twice of the inputs" in Swingler [33]. The cross-validation also prevents the issue of overfitting. Therefore, the result derived from our dataset can be used on other datasets.

We presented the results using PHQ-8 trajectories as input in this section (results on Item 9 as input can be found in Appendix A.2). The learned patterns of PHQ-8 are interpreted as a set of basis trajectories, which can be linearly combined to represent each patient's trajectory as their linear combination. Figure 2.7(a) showed 25 latent patterns: many latent

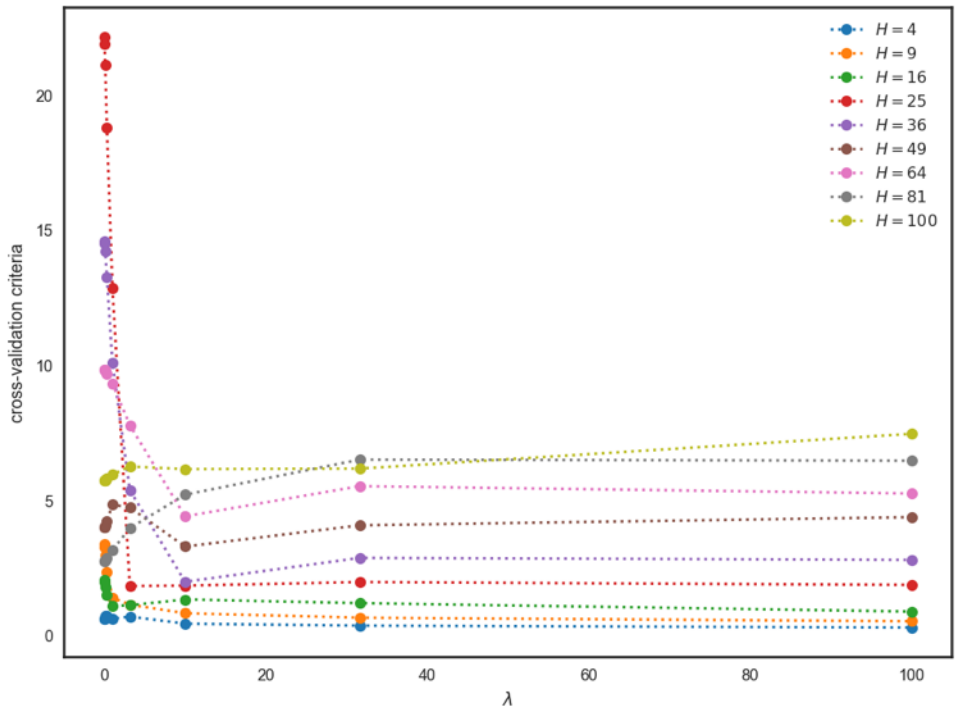


Figure 2.6: The result of cross-validation for model selection by the PHQ-8 trajectories, focused on number of nodes in hidden layers (H) and regularization parameters λ (on scale of W)

patterns are simple trend detectors (e.g., the increasing trend, the decreasing trend, and the stable trend, etc.). Some latent patterns contain multiple trends, such as increasing first and then becoming stable, or increasing first and then decreasing. A small portion of the latent patterns are sinusoidal functions, which implies that the trajectory has a periodic behavior within a short time.

Next, we investigated the similarity of each patient's activation ($\mathbf{h}$) feature on each latent pattern by embedding the activation in a two-dimensional space using t-Distributed Stochastic Neighbor Embedding (t-SNE) [34], which preserves clusters in the original data and reveals substructures within clusters. The activation of the population in the reduced dimensional space from t-SNE using PHQ-8 data as input is shown in Figure 2.7(b). We

grouped patients on the reduced dimensional space into clusters using the k-means clustering algorithm [31], such that the clusters represent the latent structure. A heuristic way to validate the clustering result is to reorder the rows of the matrix of activation values ($\mathbf{h}$) by grouping the patients within the same cluster. We plotted the value of the activation matrix after reordering by the clustering result learned; see Figure 2.7(c). It is evident that the features of activation in the same group are similar, and those in different groups have apparent dissimilarity.

Another approach to validate the clustering result is to show how distinguishable the mean trajectories are among different groups. The mean trajectory is the average measurement of the members in each group taken at each time point. Figure 2.7(d) showed the mean trajectories of PHQ-8 and Item 9 by the clustering result using the hidden patterns learned from the PHQ-8 trajectories. Five subgroups are discovered from the clustering. We can see that group 2, 4 and 5 had a trend of decreasing PHQ-8 over time, group 1 has a trend of PHQ-8 increasing first and then decreasing, and group 3 had a trend of relative stability. In Appendix A.2, we provide the results of the mean trajectories by clustering using latent patterns learned from the Item 9 trajectories. The result was similar to the above, that the mean trajectories were more distinguishable among groups when using the same measurement as input. The five subgroups were similar between PHQ-8 and Item 9, which means patients in each group followed the same average patterns in their PHQ-8 and Item 9 trajectories (e.g., comparing the right and left panels of Figure 2.7(d)).

2.4 Discussion and Summary

This study describe the longitudinal association between depression severity and suicidal ideation. In addition, the specific concern that suicidal ideation may increase during a period of depression improvement is investigated. We estimated the temporal correlation between multiple trajectories of depressive symptoms by first transforming depression records (PHQ-8 and Item 9) to continuous longitudinal trajectories, and then using the cross-correlation functions to uncover the temporal correlation between depression measures. It is worthwhile

to note that PHQ-8 and Item 9 have a strong temporal correlation; 45% of the patients with nonzero CCFs in our dataset have their PHQ-8 and Item 9 change with the most similar trend at the same time. The symmetric distribution of lag scores indicates that there exist subgroups of patients that their changes in suicidal ideation either precede or follow changes in overall depression severity. The cross-correlation between measurements of depressive symptoms can provide useful insights to the practice of depression monitoring and suicide prevention, in that if we can determine the lag at the highest CCF between two trajectories, the history of the leading series can provide substantial evidence in the prediction of the following series.

In addition, we used ANN to discover the latent structure of the depressive symptom trajectories of a treatment population, which can be interpreted as the set of basis trajectories by ANN. The latent structure provides insights on the basic patterns of the depression progression. We further exploited this structure to classify patients into subtypes and displayed the mean trajectories of the clustered groups. We found five subtypes by the local patterns from the trajectories of the PHQ-8 scores. This result is similar to recent research on subtype identification in depression [13, 22, 24, 35, 36, 37]. Our results showed that the trend of a measurement (PHQ-8 or Item 9) is more distinguishable when the classification is based on the features of latent patterns learned from the trajectories of the same measurement, but the overall patterns are similar between PHQ-8 and Item 9.

We note the following contributions. To the best of our knowledge, this research is the first time-series study on the temporal correlation between PHQ-8 and Item 9 scores of the PHQ depression questionnaire using EHR data. It is different from the time-series analysis in Snippe et al. [38] on daily changes in mindfulness, repetitive thinking, and depressive symptoms during a mindfulness-based treatment. Gunn et al. [13] has proposed that when the measurement of depressive symptoms is collected every three months, it is possible that some people may have experienced short-lived changes in their depression status between measurement periods. Our study has improved on existing literature by selecting 610 patients with more frequent records of PHQ-9 scores during a 40 weeks period. We found that 8% to

13% of the patients have experienced an increase in suicidal ideation during the improvement of PHQ-8 scores. This work is the first study to show some evidence that subgroups of depressive patients are at increased risk of suicide ideation during recovery in EHR data.

Our study is based on the assumption that thought of self-harm is an indicator to subsequent suicide attempt. There have been some debates on how significantly these two factors are associated. In the two meta-analyses of longitudinal studies over the past decades, Ribeiro et al. [9] and [10] concluded that the ability of self-injurious thought and behavior to predict suicidal attempt is weak. However, in two recent studies, Simon et al. [8] found that Item 9 of the PHQ-9 is a strong predictor of suicide attempt, after adjustment for age, sex, treatment history, and overall depression severity, by using an EHR dataset of over 80,000 patients.

There are several recent studies that aim to predict suicide and examine suicidal behaviors using large-scale EHR data ranging from thousands to millions of patients [22, 39, 36, 37]. For example, Simon et al. [22] designed a logistic regression model to predict suicide risk using EHR data of over 2.9 million patients. [40] examined 19 physical health conditions on over 2,000 individuals who died by suicide and found that traumatic brain injury had the strongest association with increased suicide risk. These new results based on health records of large number of depressive patients have supported our assumption that suicidal ideation is an important risk factor to examine in order to prevent suicide attempts. We note that our study sample of 610 patients appears to be small in comparison, this is because we focus on a chronic depression treatment population with frequent observations to study their temporal patterns, which has much more restrictive data requirement than these extensive suicide risk prediction studies.

There are several limitations to this work. First, the Gaussian process regression is appropriate in the transformation of PHQ-8 scores which range from 0 to 24, and the variance is small for an individual patient. However, for sparse data like Item 9 which has the majority of measurements being zeroes, the GPR does not have a significant advantage compared to other interpolation methods like the linear spline. Second, the results are based on a small

patient cohort that is frequently monitored. Selection of patients with frequent visits may be biased toward those with more severe illness, but this is the patient group of most significant concern regarding worsening of suicidal ideation. All records rely on self-report and the patients in the dataset are mostly older adults (middle-aged and older) in the western states. It may not be representative of young patients or those from different cultural backgrounds. Furthermore, our samples include only patients receiving treatment for depression; thus our findings may not be generalizable to people with unrecognized or untreated depression. Third, our data does not include sufficient information on demographics, education, income and other useful clinical factors to help us understand the findings. In addition, EHR data are sparse and irregular (e.g., we have no knowledge on whether there is a change of trend during a long gap between two consecutive PHQ records, which means that the changes calculated from a long period are not as reliable as the one from a shorter period). Finally, the latent structure is a local pattern, which means we do not specify the starting point of this latent trajectory in the whole disease history. This limits the application of this model to other chronic diseases for the following reason. Some clinical outcomes are periodic which are appropriate inputs to the latent structure learning model in this paper, while others are not periodic in which learning the local trend does not guarantee the trend in the long run. To apply this model to other disease applications, the periodicity of the clinical outcomes should be examined first.

In this chapter, we established a framework to extract subtypes of depression progression patterns from trajectories of depressive symptoms and examined the temporal relationship between PHQ-8 and Item 9 scores. This work contributes to the emerging field of personalized depression monitoring and suicide prevention by providing some insights on analyzing temporal measurements to understand the association between different depressive symptoms. Future work can be extended to include additional measurements and clinical notes [37], such as examining temporal relationships between alcohol and drug use with depression symptoms trajectories.

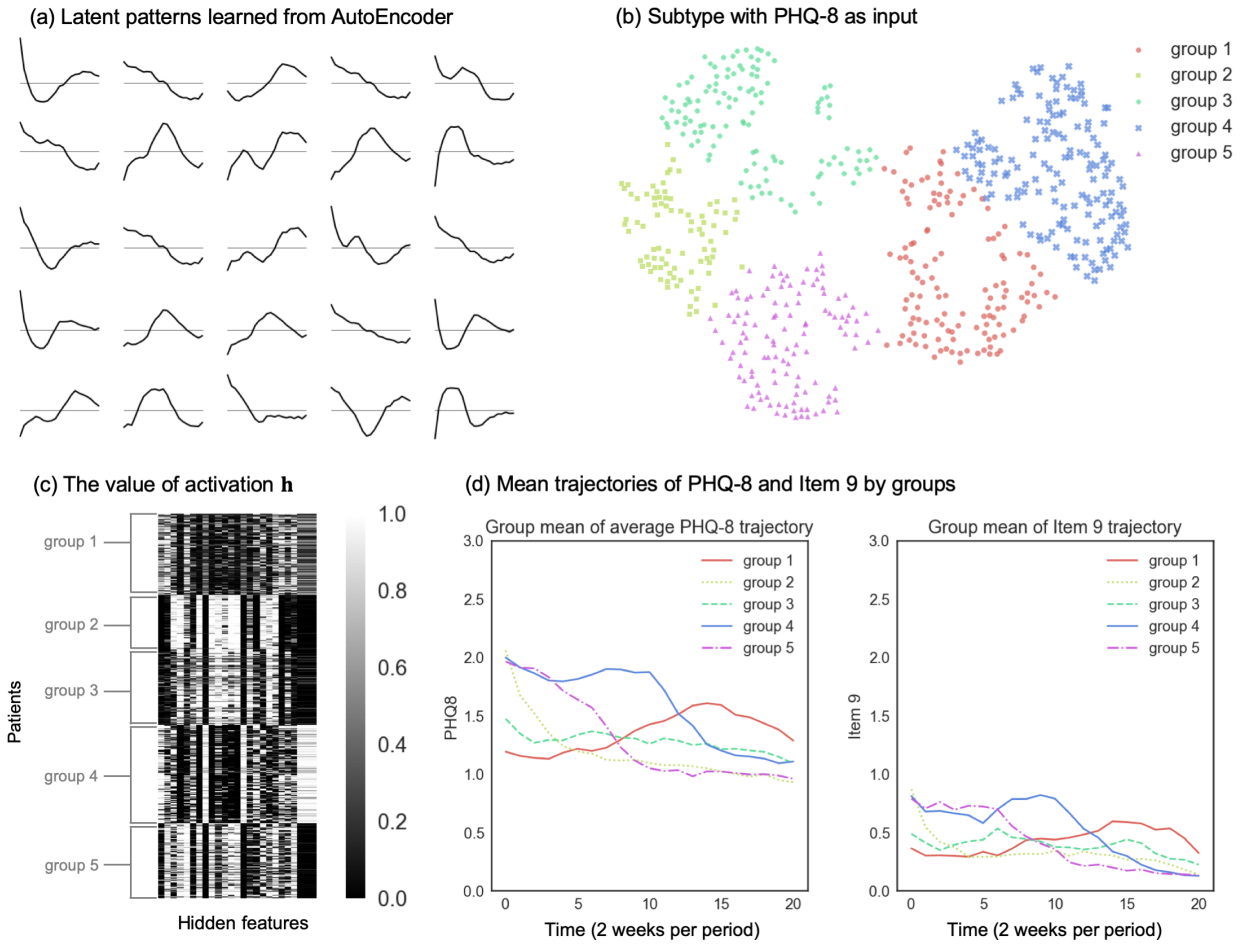


Figure 2.7: The subtype analysis result with hidden structures learned from the PHQ-8 trajectories. (a) Latent patterns learned from the PHQ-8 data. These patterns are visualized as the rows $\mathbf{W}_i$. In each panel, the x-axis is the time with a period of 2 weeks, and the y-axis represents the PHQ-8 scores of each basis trajectory. (b) Embed the activation $\mathbf{h}$ (25 dimensions) of each patient into 2-dimensional space with t-SNE and cluster them with the k-means algorithm. (c) The value of activation $\mathbf{h}$ on each latent pattern (25 columns) of each patient (610 rows) after reordering the rows by the clustering. (d) Mean trajectories of average PHQ-8 and Item 9 by groups, using the clustering results. One unit of time is two weeks. We use the average score of the first 8 questions in the PHQ to represent PHQ-8, which has the same range of 0 to 3 to Item 9.

Chapter 3

PARTIALLY OBSERVABLE COLLABORATIVE MODEL FOR OPTIMIZING PERSONALIZED TREATMENT SELECTION

In the previous chapter, we have developed methods to extract subtypes of chronic disease based on the symptom trajectories of a heterogeneous population. We want to develop a personalized treatment selection strategy by utilizing this structure of the heterogonous population. In particular, we propose a framework called the Partially Observable Collaborative Model (POCM), for the estimation of the individualized chronic disease progression. This chapter includes the formulation of the POCM and the algorithm for learning the parameters of POCM, which is an iterative updating rule for solving a nonconvex optimization problem. In addition, we discuss how to construct personalized treatment selection strategy based on the Partially Observable Markov Decision Process (POMDP) by utilizing the model learned from the POCM. A numerical experiment on the depression based on a simulated dataset is provided to illustrate the performance of the POCM powered strategy compared to conventional strategies.

3.1 Introduction

Personalized treatment of chronic disease is a sequence of treatments tailored to individual patient's characteristics, disease history, and treatment response. Developing a personalized treatment plan is a difficult sequential decision-making problem due to uncertainty in observing the patient's true health state, predicting disease progression, and estimating treatment response. The goal of treatment selection over time is to obtain the optimal health outcome for the patient within the resource limit. Main challenges include insufficient knowledge of personal disease progression dynamics and the learning of individual response to treatment in

real time. To address these challenges, we propose a mathematical framework for optimizing personalized treatment of chronic disease under partially observable health conditions and uncertain treatment outcomes.

For many chronic diseases, there are multiple treatment options that can be selected over a long time horizon. Treatment options may include medications, medical devices, behavioral therapies, etc., or no further treatment. Currently, treatment decisions are made by physicians based on their individual experience and expertise during outpatient visits. These decisions are generally consistent with published treatment guidelines and expert consensus documents established from clinical trial data at the population level [41, 42]. In recent years, widely implemented electronic health record (EHR) systems, expanded use of clinical decision-support tools, telemedicine, and mobile health apps are moving clinical care into an era of personalized medicine or precision medicine. The merging of big data analytics and artificial intelligence (AI) in medicine is generating a vibrant research area that foretells the next revolution in healthcare. Foreseeable benefits of personalized medicine include faster, safer, cheaper, more convenient, and higher quality of care for patients.

The first challenge in designing a personalized treatment framework is to model the individual disease progression dynamics. In a heterogeneous population, each patient's disease progression can be characterized as a unique dynamic process. Maximum likelihood estimation (MLE) of the progression parameters from observational data is commonly used. A naïve approach is to estimate a model separately for each individual. However, this approach is less effective because the information from other patients is not exploited. In addition, it is difficult to collect a large quantity of longitudinal data for individual patients in practice. Consequently, measurements from an individual patient may not be sufficient to develop an accurate disease progression model. A possible approach to mitigate this limitation is to use cohort data to identify homogeneous subgroups of patients in which the disease progression dynamics can be represented by basis models, and then design a specific treatment plan for each subgroup. The problem in this approach is that the disease progression of an individual patient may not be represented by the model in a single subgroup; in other words, it can be a

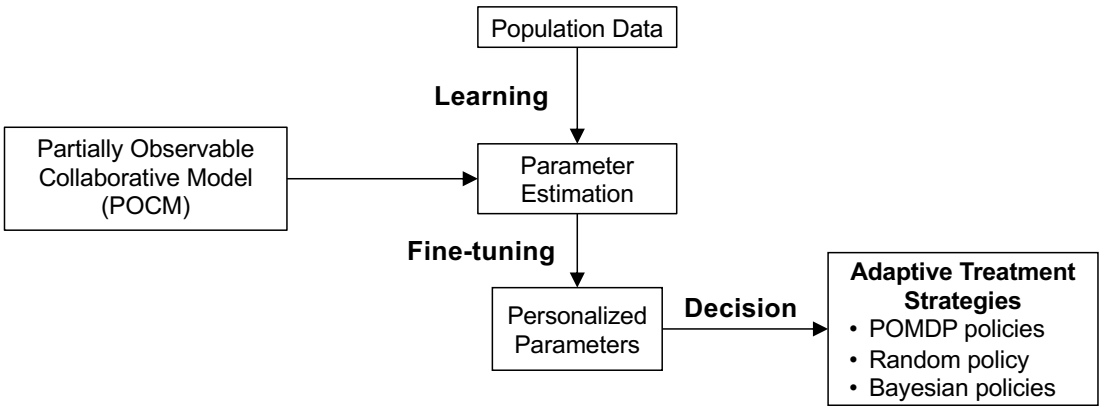
mixture of basis features from several subgroups. Methods such as K-mean clustering, hierarchical linear modeling, growth mixture modeling, latent class analysis, latent class growth analysis, and latent class growth mixture modeling are used to analyze longitudinal cohort data to discover subgroup structures [43]. A more advanced approach is the collaborative model (CM) [35, 44, 24, 45]. The collaborative model framework uses a set of basis models to represent patterns as subtypes of disease progression, based on which an individual patient's progression dynamic is a combination of multiple basis models. Suppose there are K basis models discovered in the population, and we define the membership of the individual patient i as $c_{ij} \in [0, 1]$, representing the degree to which the model of the individual patient i corresponds to the basis model j. Therefore, these basis models span the modeling space, which means the distinct progression model of any individual patient can be approximated by a weighted combination of the K basis models, where the weight is the membership.

The second challenge is designing the adaptive treatment selection strategy based on the partially-observed health state of the patient (e.g., current treatment outcome) and past treatment information. Among research in algorithmic based treatment planning, Markov decision process (MDP) [46], reinforcement learning [47], and multi-armed bandits [48] are among the most widely-used tools. For example, MDP models have been developed for medical treatment decisions in the literature, such as diabetes [5, 49], hepatitis C [50, 51], liver transplantation [52], etc. Disease progression is assumed to be a Markov model if the health state is fully observable. The treatment selection problem is then modeled by an MDP, with an objective of maximizing the total reward during treatment. However, the Markov model fails to capture the process in which the health state is observed only indirectly via a collection of incomplete or imperfect observations. A standard model for the hidden state is the Hidden Markov model (HMM). In order to formulate an individual model with hidden health states that utilizes the heterogeneous structure in the population data, we propose a model that combines CM and HMM, named the Partially Observable Collaborative Model (POCM). In the POCM, the basis model is an HMM for each disease progression subtype, and the individual progression model is a weighted combination of

the basis HMMs. We estimate the parameters of each basis model and the membership of each individual model in the POCM from longitudinal observations of a population. We develop an efficient algorithm to solve the parameter estimation problem and prove that the proposed algorithm can guarantee convergence to a stationary solution similar to CM. We do not assume a priori knowledge of the subgroup assignment of each patient before the estimation; instead, the basis model and the memberships are learned simultaneously. We also assume that the basis model learned from longitudinal observations of existing patients can be used to estimate the individual models for future patients. Therefore, for a new patient, we only need to learn the membership from a small number of treatment trial periods to determine the individual disease model. Our hypothesis is that when low-dimensional heterogeneity structure exists, POCM will be a better model for individualized progression than naïve HMM. An example of heterogeneity structure is that, assuming there exist several basis models, the individual membership on these models can be evenly spread or can be extremely focused on a particular model.

Next, we propose a partially-observable MDP (POMDP) model for making a sequence of adaptive treatment decisions based on the estimated individual disease dynamics. The hidden states of the POMDP are health states representing the disease severity, observations are test scores that are imperfect measurements of the health state, and decisions are treatment types. For example, the objective can be to maximize health outcome over time by optimally selecting between two treatments (e.g., Treatment I is a traditional treatment, and Treatment II is an experimental treatment) in each time period. Model parameters for the progression dynamics include transition probabilities between health states and emission probabilities of the observations given the true health states. POCM or an existing HMM algorithm are used to estimate individual transition and emission matrices, which represent the individual response to the treatments. The objective is to maximize discounted total rewards measured using Net Monetary Benefit (NMB), defined as total health benefits $\times$ willingness-to-pay $-$ total cost. Health benefits can be measured by quality-adjusted life years gained.

We are interested in the question of whether POCM is a better model than HMM in the sense that the parameters inferred from POCM can lead to a better treatment selection policy. The process of creating the adaptive treatment strategy includes three steps. (1) In the learning step, the basis models for each progression subtype and patients' memberships, are learned from an existing dataset of patients under Treatment I. The population average treatment effect for Treatment II is assumed to be known from clinical trials or observational studies, and such knowledge is used to estimate the basis model parameters for Treatment II. (2) In the fine-tuning step, the personal dynamic for a new patient is initialized using model parameters estimated from the learning step, and updated under both treatment options by running separate short trial periods; this is again accomplished by the POCM algorithm or the HMM algorithm. The membership is then solved for each new patient under either Treatment I or II. (3) In the decision step, the optimal treatment strategy is obtained by solving a rolling horizon POMDP and compared with other heuristic policies.



Overview of the POCM model and treatment strategy.

We demonstrate our framework using a simulated population of chronically depression patients. Depression is one of the most common mental disorders in the U.S., affecting more than 10% of the population [53]. Treatment for depression includes psychotherapy, antidepressants, or a combination of the two with supportive care. There are no clear evidence-based guidelines on when/how to switch treatment for depression. Furthermore,

staying on inefficient treatments may induce more cost without any benefits for patients, and lead to treatment-resistance or addiction to medications. In our simulation, Treatment I is the usual care using antidepressant medication, and Treatment II is an intensive outpatient program with additional behavioral counseling. We find that with the individual parameters learned from the POCM, the treatment selection policy can lead to higher NMBs and lower prediction error of the true health state over the course of treatment.

The main contributions of this chapter are twofold. First, we propose a mathematical framework to characterize the individual variations in chronic disease progression in which the health states are partially observable. Through a simulation study of chronic depression, we show that POCM is a better model than individual HMM estimation in the majority of the model settings. Second, we design an adaptive treatment selection algorithm using the individual disease progression model learned from a small set of treatment experiments, and compare the resulting policy's performance to a set of heuristic policies. The optimized treatment plan is tested in sensitivity analyses including uncertainties in the estimated disease progression by treatment types, cost of treatments, and health utilities of the health states, etc.

The remainder of the paper is structured as follows. Section 3.2 provides the relevant literature on optimal treatment models, applications and solution algorithms. Section 3.3 introduces the methods that are evaluated in this study. The numerical results of a simulation study on the performance of these methods are given in section 3.4, followed by concluding remarks in section 3.5.

3.2 Relevant Literature

Our work is related to the literature on optimal treatment selection for chronic diseases. MDP and POMDP have been used to determine the optimal treatment plan for a number of diseases with the assumption that the disease transition is Markovian. Shechter et al. [54] developed an MDP model to find the optimal time to initiate HIV therapy for U.S. veterans. Mason et al. [49] used an MDP model to optimize the treatment decision for patients with

type 2 diabetes. Maillart et al. [55] formulated a partially-observed Markov model to select efficient breast cancer screening policies. Faissol et al. [50] used a POMDP to determine the best timing of treatment decisions when the presence of the disease is not known in advance of hepatitis C screening. Saure et al. [56] formulated a discounted infinite-horizon MDP for scheduling cancer treatments in radiation therapy units. Other methods include the Kalman filter [57, 58], multi-arm bandit [59, 60], and mixed integer programming [61]. A key distinction between these research and our settings is that, this stream of work mainly focuses on finding the optimal treatment based on population characteristics, while our study focuses on the treatment selection optimized for an individual patients.

Our work also relates to the stream of research that focuses on the personalization of treatment selection. Wang et al. [62] proposed a personalized disease progression model by combining Markov jump process and Markov chains. Schulam and Saria [63] proposed a hierarchical latent variable model that individualizes predictions of disease trajectories, and provided the algorithm for learning population, subpopulation and individual parameters. Ayer et al. [64] designed a personalized mammography screening policy based on the prior screening history and personal risk characteristics of women. Lavieri et al. [65] proposed an individual disease progression of prostate cancer patients using dynamic Kalman filter model to estimate the individual parameters from the population characteristics. These papers focus on a target disease and require domain knowledge on the progression of the target disease. Our purpose is different: we propose a general-purpose model learning method for any chronic disease based on the longitudinal observations. Our method is based on a general disease progression model for fully-observable diseases–CM [35, 44, 24]. We extend CM with the ability to model the partially-observable disease conditions by adding latent variables to represent health states. MLE is the most common approach of parameter inference from observational data in disease models. However, inference of latent variables with MLE is usually difficult due to the nonconvexity of the likelihood function. The Expectation-Maximization (EM) algorithm can overcome this difficulty by iteratively estimating the intermediate states and maximizing the approximate likelihood based on the intermediate states [66, 67]. We

develop a variant of the EM algorithm for the inference of CM for diseases with latent health state.

Finally, our work relates to the stream of research that applies AI methods to medical decision-making (MDM) problems. A recent review of supervised and unsupervised learning applications in MDM is Jiang et al. [68]. Reinforcement learning (RL), which formulate the process of an agent (e.g., the decision maker) interacting with an environment (e.g., the disease progression model), is widely used for the sequential decision-making problems in healthcare. Ayer [69] proposed an inverse RL to identify the optimal screening strategies for breast cancer in the setting of a partially observable environment. In order to simultaneously utilize the biomedical dynamics across multiple patients, Lee et al. [70] designed three classes of RL policies for the screening of Hepatocellular Carcinoma. In addition, RL can be used to solve MDP and POMDP problems approximately, which is useful when the state space is large and computation resource is limited [71, 72]. Otherwise, exact solution via dynamic programming may be preferred. In addition, RL requires a large amount of iterations of interactions (typically larger than 10,000) with the environment to ensure convergence of the optimal policy, while dynamic programming does not require interaction with the environment before performing the policies. In MDM problems, the cost of performing the treatment is high and the interval between two treatments may be long, which limit the application of RL to such problems. In this paper, we used the incremental pruning method to find the optimal treatment policy of the POMDP [73].

3.3 Model Formulation

Our personalized treatment method is performed in 3 stages: (1) Learning stage: the basis models of the POCM are learned from existing treatment records of the population; (2) Fine-tuning stage: the individual disease progression model is fine-tuned from the basis models in separate short trial periods; (3) Decision stage: the type of treatment is selected for an individual patient in each period by solving the POMDP or using heuristic policies. We provide descriptions of the disease model in Section 3.3.1. The POCM formulation and the

solution algorithm for parameter estimation are provided in Section 3.3.2. In Section 3.3.3, we present the POMDP model for optimal treatment selection.

3.3.1 Disease model

We formulate the treatment decision process as a finite-state, finite horizon, discrete-time POMDP where the underlying Hidden Markov Model represents the progression dynamics of a patient's health state. Decision-makers such as patients and clinicians aim to maximize the total expected net monetary benefit (NMB) of the patient. We assume that the decision maker is risk neutral. The notation used in the model is as follows.

Decision epoch. Treatment decisions are made at a finite and discrete set of time periods $t = 1, 2, \ldots, T, T < \infty$. The time interval between decision epochs for chronic disease is usually the time between treatment decisions made by the clinician [74]. For instance, in the case of chronic depression, the decision on treatment type can be made monthly. For the treatment of chronic disease, we consider finite-horizon and exclude the possibility of death during treatment.

State and observation. Let s_t denote the health state at time t. The state space $\mathcal{S}$ is the set of all possible health states. We assume the true health state cannot be observed. Instead, at each time period, the patient's health condition is examined by some observable measures. The observations can be either continuous variables or categorical variables depending on the applications. In this paper, we focus on categorical observations. Let s_t denote the observation at time period t, and the observation space Ω is a finite set of all possible observation values, i.e., $o_t \in \Omega$. Since the true health state s_t is hidden, we can only maintain an estimation of the probability distribution of the state over the state space $\Delta(\mathcal{S}) = \left\{b(s_t) \in [0,1], s_t \in \mathcal{S} \middle| \sum_{s_t \in \mathcal{S}} b(s_t) = 1\right\}$, which is usually called the belief of states.

Action. Actions include the selection of treatment types. Let a_t denotes the action taken at time t. At the beginning of each time period, an action is selected with the current policy. We consider two types of treatment; Treatment II is more effective and more expensive than Treatment I. The action space $\mathcal{A}$ is the set of all possible actions, i.e. $a_t \in \mathcal{A} = \{\text{I, II}\}$. A

policy $\pi : \Delta(\mathcal{S}) \to \mathcal{A}$ is the probability of taking action a when the belief of states is $\Delta(\mathcal{S})$. At the beginning of each decision epoch, the treatment type is selected with respect to the policy. The action and the corresponding observation can provide the clinician with valuable information about the true health state of the patients, which in turn helps the clinician to evaluate the policy.

System Dynamic. The state transition probabilities in a POMDP model of chronic disease describes the disease progression under different types of treatment. It is defined as the probability that the patient will be in state $j \in \mathcal{S}$ at time $t+1$, given that she is in state i and action a is taken, denote as $\mathbf{A}^a(s, a, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$. The emission probability is the probability of making an observation $o \in \Omega$ at time t when the true health state is $s \in \mathcal{S}$, denote as $\mathbf{B}(s, o) = \Pr(o_t = o | s_t = s)$. We assume this probability is independent of the action. We name $\{\mathbf{A}(s, a, s')\}_{a \in \mathcal{A}, s \in \mathcal{S}, s' \in \mathcal{S}} \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{S}|}$ as the transition probability matrix, and $\{\mathbf{B}(s, o)\}_{s \in \mathcal{S}, o \in \Omega} \in \mathbb{R}^{|\mathcal{S}| \times |\Omega|}$ as the emission probability matrix.

Reward. The reward includes both health outcomes and economic costs associated with treatment. Health outcomes may include life expectancy gains, and quality-of-life decrements due to the treatment side effect. Costs may include medication or hospital expenses. Quality-adjusted life years (QALYs) is a common metric to quantify the quality-of-life gains from medical interventions. One QALY represents a patient living in perfect health for one year. The immediate reward $r(s_t, a_t)$ is the NMB of treatment in one period when the patient's true health state is y and the action selected is a,

$$r(s_t, a_t) = \lambda \kappa(s_t) - c(a_t) \tag{3.1}$$

where $\kappa(s_t)$ is the utility of being in state s_t measured in QALYs; $c(a_t)$ is the cost of treatment a_t; λ is the willingness to pay (WTP) which assigns a monetary value to a QALY; \$50,000 per QALY is commonly used for WTP in the literature.

3.3.2 Partially Observable Collaborative Model (POCM)

We model subgroup structure in the disease progression with the CM method [35, 44, 24]. In short, the CM assumes that a basis model represents a subtype of disease progression in the population. Each individual model is a weighted combination of the basis models. By assigning each individual a distinct weight vector (denoting c_{ik}, the weight of subgroup k for patient i) on the basis models, it captures the individual to individual variations. We assume there are K basis models and N patients. In POCM, the underlying disease dynamic is an HMM, with the basis transition matrix $\mathbf{A}_k$, the basis emission matrix $\mathbf{B}_k$, and the basis initial distribution of state $\boldsymbol{\pi}_k$ for group $k \in \{1, \ldots, K\}$. We denote $\theta = \{\mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\pi}_k, k = 1, \ldots, K\}$ as the basis parameters of POCM. Each patient's individual progression model, which is also an HMM, is assumed to be a linear combination of the basis models. The weight of the linear combination is called the membership vector $\mathbf{C}_i \in \mathbb{R}^K$ for patient i; we denote $\mathbf{C} = [\mathbf{C}_1, \ldots, \mathbf{C}_N] \in \mathbb{R}^{N \times K}$ as the membership matrix. The individual parameters for patient i can be described as (1) Initial distribution of state $\hat{\boldsymbol{\pi}}_i = \sum_{k=1}^{K} c_{ik} \boldsymbol{\pi}_k$, (2) Transition probability matrix $\hat{\mathbf{A}}_i = \sum_{k=1}^{K} c_{ik} \mathbf{A}_k$, and (3) Emission probability matrix $\hat{\mathbf{B}}_i = \sum_{k=1}^{K} c_{ik} \mathbf{B}_k$. We denote the observations of each patient $\mathbf{O}_i = [o_{i,1}, o_{i,2}, \ldots, o_{i,T}] \in \mathbb{R}^T$, and the observations of all patients as $\mathbf{O} = [\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_N] \in \mathbb{R}^{N \times T}$. We denote $s_{i,t}$ as the true health state for patient i at time period t, $\mathbf{S}_i = \{s_{i,1}, s_{i,2}, \ldots, |\mathcal{S}|_{i,T}\} \in \mathbb{R}^T$ as the series of true health states for patient i, and $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_N\} \in \mathbb{R}^{N \times T}$ is the set of latent states for all patients. We denote $\mathbb{S}$ as the set of all possible combinations of $\mathbf{S}$. The objective of the POCM is to optimize the maximum likelihood estimator of the observed sequence $\mathbf{O}$ with the following optimization problem.

$$\max_{\theta,\mathbf{C}} . \quad f(\theta, \mathbf{C}) = \log \Pr(\mathbf{O}|\theta, \mathbf{C}) - \frac{\mu}{2} \sum_{i,j} w_{ij} \|\mathbf{c}_i - \mathbf{c}_j\|^2 \tag{3.2a}$$

$$\text{s.t.} \quad \sum_{s'} \mathbf{A}_k(s, s') = 1,\ s = 1, \ldots, |\mathcal{S}|, k = 1, \ldots, K \tag{3.2b}$$

$$\sum_{o=1}^{|\Omega|} \mathbf{B}_k(s, o) = 1,\ s = 1, \ldots, |\mathcal{S}|, k = 1, \ldots, K \tag{3.2c}$$

$$\sum_{s=1}^{|\mathcal{S}|} \boldsymbol{\pi}_k(s) = 1,\ k = 1, \ldots, K \tag{3.2d}$$

$$\sum_{k=1}^{K} c_{i,k} = 1,\ i = 1, \ldots, N \tag{3.2e}$$

$$\mathbf{A}_k(s, s'), \mathbf{B}_k(s, o), \boldsymbol{\pi}_k(s), c_{i,k} \geq 0,\ s, s' = 1, \ldots, |\mathcal{S}|, k = 1, \ldots, K,\ i = 1, \ldots, N$$

The first four constraints guarantee that the transition probability, emission probability, initial state distribution, and membership vector sum up to 1. Note the last term of the objective function is the regularization term that incorporates the similarity between patients. The regularization coefficient μ is a tuning parameter. $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the similarity matrix. The similarity between two individuals can be quantified by comparing their profiles of the covariate, such as the demographics, social-economical, genetic and imaging information [35, 44, 24]. We can simplify the regularization term as

$$\frac{1}{2} \sum_{i,j} w_{ij} \|\mathbf{C}_i - \mathbf{C}_j\|^2 = \sum_i \left(\sum_j w_{ij} \right) \mathbf{C}_i \mathbf{C}_i^\top - \sum_{i,j} w_{ij} \mathbf{C}_j \mathbf{C}_i^\top = \mathrm{Tr}\left(\mathbf{C}^\top \mathbf{L} \mathbf{C} \right),$$

where $\mathbf{L}$ is the Laplacian matrix of w_{ij}, $\mathbf{L} = \mathbf{D} - \mathbf{W}$, and $\mathbf{D}$ is a diagonal matrix with elements $d_{ii} = \sum_j w_{ij}$.

The EM algorithm is a standard approach for the inference of the model with latent variables like the POCM. An example is the Baum-Welch algorithm for the HMM inference [75]. In the EM algorithm, in each iteration $m = 1, 2, \ldots$, we will estimate the latent states and maximizing the likelihood based on the latent states. Computing the likelihood of observed sequence with latent variables are computationally intractable. Instead, we can

replace the likelihood with an equivalent function Q, defined as

$$Q\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right) := \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{O}, \mathbf{S}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O}, \mathbf{S}|\theta, \mathbf{C}\right)], \tag{3.3}$$

where $\theta^{(m)}, \mathbf{C}^{(m)}$ is the estimation of the parameters $\theta, \mathbf{C}$ in after m iterations. [76, Chapter 9] We can solve the POCM by maximizing $Q\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right)$ through updating θ and $\mathbf{C}$.

Theorem 3.1. *The following two objective functions are equivalent*

$$\arg\max_{\theta,\mathbf{C}} \Pr\left(\mathbf{O}|\theta, \mathbf{C}\right) = \arg\max_{\theta,\mathbf{C}} \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{O}, \mathbf{S}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O}, \mathbf{S}|\theta, \mathbf{C}\right)].$$

The proof of Theorem 3.1 is provides in Appendix B.1. This optimization problem can be solved with with the Lagrangian multiplier method. First, the Lagrangian is

$$\begin{aligned} L\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right) = Q\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right) - \mu \operatorname{Tr}\left(\mathbf{C}^T\mathbf{L}\mathbf{C}\right) - \sum_{k=1}^{K} \lambda_k^{(\boldsymbol{\pi})} \left(\sum_{s=1}^{|\mathcal{S}|} \boldsymbol{\pi}_k(s) - 1\right) \\ - \sum_{k=1}^{K}\sum_{s=1}^{|\mathcal{S}|} \lambda_{s,k}^{(\mathbf{A})} \left(\sum_{s'=1}^{|\mathcal{S}|} \mathbf{A}_k(s, s') - 1\right) - \sum_{k=1}^{K}\sum_{s=1}^{|\mathcal{S}|} \lambda_{s,k}^{(\mathbf{B})} \left(\sum_{o=1}^{|\Omega|} \mathbf{B}_k(s, o) - 1\right) \\ - \sum_{i=1}^{N} \lambda_i^{(\mathbf{C})} \left(\sum_{k=1}^{K} c_{i,k} - 1\right), \end{aligned} \tag{3.4}$$

where $\lambda_k^{(\boldsymbol{\pi})}, \lambda_{s,k}^{(\mathbf{A})}, \lambda_{s,k}^{(\mathbf{B})}$ and $\lambda_i^{(\mathbf{C})}$ are dual variables. The optimization problem can be simplified as maximizing the Lagrangian L by repeating the following steps until convergence:

1. Fix $\mathbf{C}^{(m)}$, set $\theta^{(m+1)} = \arg\max_\theta Q\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)$

2. Fix $\theta^{(m+1)}$, set $\mathbf{C}^{(m+1)} = \arg\max_{\mathbf{C}} Q\left(\theta, \mathbf{C}|\theta^{(m+1)}, \mathbf{C}^{(m)}\right) - \mu \operatorname{Tr}(\mathbf{C}^\top\mathbf{L}\mathbf{C})$

where $\theta^{(m)}$ denotes the value after m iterations. We will describe the optimization in each step as follows.

Step 1: Update basis models with fixed membership

The partial derivatives of L with respect to $\boldsymbol{\pi}_k(s)$ and the corresponding dual variables are

$$\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \boldsymbol{\pi}_k(s)} = \sum_{i=1}^{N} \frac{c_{i,k}\gamma_{i,1}^{(m)}(s)}{\sum_{k'=1}^{K} c_{i,k'}\boldsymbol{\pi}_{k'}(s)} - \lambda_k^{(\boldsymbol{\pi})} = 0, \quad s = 1, \ldots, |\mathcal{S}|, k = 1, .., K;$$

$$\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \lambda_k^{(\boldsymbol{\pi})}} = -\left(\sum_{s=1}^{|\mathcal{S}|} \boldsymbol{\pi}_k(s) - 1\right) = 0, \quad k = 1, \ldots, K.$$

Substitute this into $\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \boldsymbol{\pi}_k(s)} \boldsymbol{\pi}_k(s) = 0$, we have

$$\sum_{i=1}^{N} \frac{c_{i,k}\gamma_{i,1}^{(m)}(s)\boldsymbol{\pi}_k(s)}{\sum_{k'=1}^{K} c_{i,k'}\boldsymbol{\pi}_{k'}(s)} - \sum_{s=1}^{|\mathcal{S}|}\sum_{i=1}^{N} \frac{c_{i,k}\gamma_{i,1}^{(m)}(s)\boldsymbol{\pi}_k(s)}{\sum_{k'=1}^{K} c_{i,k'}\boldsymbol{\pi}_{k'}(s)} \boldsymbol{\pi}_k(s) = 0$$

which leads to the updating rules

$$\boldsymbol{\pi}_k^{(m+1)}(s) = \frac{\sum_{i=1}^{N} \frac{c_{i,k}^{(m)}\gamma_{i,1}^{(m)}(s)\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'=1}^{K} c_{i,k'}\boldsymbol{\pi}_{k'}^{(m)}(s)}}{\sum_{s=1}^{|\mathcal{S}|}\sum_{i=1}^{N} \frac{c_{i,k}^{(m)}\gamma_{i,1}^{(m)}(s)\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'=1}^{K} c_{i,k'}^{(m)}\boldsymbol{\pi}_{k'}^{(m)}(s)}}, \tag{3.5}$$

where $\gamma_{i,t}^{(m)}(s) = \Pr\left(s_{i,t} = s \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right)$ is the probability that the system is at state s at time t, given the observation sequence $\mathbf{O}_i$ and the model $\theta^{(m)}, \mathbf{C}^{(m)}$.

Similarly, we have the updating rules for $\mathbf{A}$ and $\mathbf{B}$

$$\mathbf{A}_k^{(m+1)}(s, s') = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1} \frac{c_{i,k}^{(m)}\xi_{i,t}^{(m)}(s,s')\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'=1}^{K} c_{i,k'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}}{\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\sum_{t=1}^{T-1} \frac{c_{i,k}^{(m)}\xi_{i,t}^{(m)}(s,s')\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'=1}^{K} c_{i,k'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}}, \tag{3.6}$$

$$\mathbf{B}_k^{(m+1)}(s, o) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \frac{c_{i,k}^{(m)}\gamma_{i,t}^{(m)}(s)I(o_{i,t}=o)\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'=1}^{K} c_{i,k'}\mathbf{B}_{k'}(s,o)}}{\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\sum_{t=1}^{T} \frac{c_{i,k}^{(m)}\gamma_{i,t}^{(m)}(s)I(o_{i,t}=o)\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'=1}^{K} c_{i,k'}\mathbf{B}_{k'}(s,o)}}, \tag{3.7}$$

where $I(\cdot)$ is the identity function, $\xi_{i,t}^{(m)}(s, s') = \Pr\left(s_{i,t} = s, s_{i,t+1} = s' \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right)$ is the probability of being at state s at time t, and at state s' at time $t+1$, given the observation sequence $\mathbf{O}_i$ and the model $\theta^{(m)}, \mathbf{C}^{(m)}$. The intermediate states $\gamma_{i,t}^{(m)}(s)$ and $\xi_{i,t}^{(m)}(s, s')$ can be estimated from the forward-backward algorithms using the current estimation of $\boldsymbol{\pi}$, $\mathbf{A}$ and $\mathbf{B}$, see Appendix B.2.3 for details.

Step 2: Update membership with fixed basis models

Next, we fix $\theta^{(m)}$, let us focus on the $\mathbf{C}_{i,k}$'s

$$\frac{\partial L\left(\theta,\mathbf{C}|\theta^{(m+1)},\mathbf{C}^{(m)}\right)}{\partial c_{i,k}}=\sum_{s=1}^{|\mathcal{S}|}\frac{\boldsymbol{\pi}_k(s)\gamma_{i,1}^{(m)}(s)}{\sum_{k'=1}^{K}\boldsymbol{\pi}_{k'}(s)c_{i,k'}}+\sum_{t=1}^{T-1}\sum_{s=1}^{|\mathcal{S}|}\sum_{s'=1}^{|\mathcal{S}|}\frac{\mathbf{A}_k(s,s')\xi_{i,t}^{(m)}(s,s')}{\sum_{k'=1}^{K}\mathbf{A}_{k'}(s,s')c_{i,k'}}$$
$$+\sum_{t=1}^{T}\sum_{s=1}^{|\mathcal{S}|}\frac{\mathbf{B}_k\left(s,o_{i,t}\right)\gamma_{i,t}^{(m)}(s)}{\sum_{k'=1}^{K}\mathbf{B}_{k'}\left(s,o_{i,t}\right)c_{i,k'}}-\mu\left(\mathbf{LC}\right)_{i,k}-\lambda_i^{(\mathbf{C})}=0,$$
$$(i=1,..,N,k=1,..,K)$$
$$\frac{\partial L\left(\theta,\theta^{m}\right)}{\partial\lambda_i^{(\mathbf{C})}}=-\left(\sum_{k=1}^{K}c_{i,k}-1\right)=0,\ (i=1,\ldots,N)$$

Use $\sum_{k=1}^{K}\frac{\partial L\left(\theta,\mathbf{C}|\theta^{(m)},\mathbf{C}^{(m)}\right)}{\partial c_{i,k}}c_{i,k}=0$, and note $\sum_{k=1}^{K}c_{i,k}=1$, we have

$$\lambda_i^{(\mathbf{C})}=1+S(T-1)+T-\mu\left(\mathbf{LC}\right)_i\mathbf{C}_i^{\top}=S\left(T+\gamma_{i,1}^{(m)}(s)\right)-\mu\left(\mathbf{LC}\right)_i\mathbf{C}_i^{\top}\tag{3.8}$$

Substitute (3.8) into $\frac{\partial L\left(\theta,\mathbf{C}|\theta^{(m)},\mathbf{C}^{(m)}\right)}{\partial c_{i,k}}c_{i,k}=0$, we have

$$\sum_{s=1}^{|\mathcal{S}|}\frac{\boldsymbol{\pi}_k(s)\gamma_{i,1}^{(m)}(s)c_{i,k}}{\sum_{k'=1}^{K}\boldsymbol{\pi}_{k'}(s)c_{i,k'}}+\sum_{t=1}^{T-1}\sum_{s=1}^{|\mathcal{S}|}\sum_{s'=1}^{|\mathcal{S}|}\frac{\mathbf{A}_k(s,s')\xi_{i,t}^{(m)}(s,s')c_{i,k}}{\sum_{k'=1}^{K}\mathbf{A}_{k'}(s,s')c_{i,k'}}+\sum_{t=1}^{T}\sum_{s=1}^{|\mathcal{S}|}\frac{\mathbf{B}_k\left(s,o_{i,t}\right)\gamma_{i,t}^{(m)}(s)c_{i,k}}{\sum_{k'=1}^{K}\mathbf{B}_{k'}\left(s,o_{i,t}\right)c_{i,k'}}$$
$$-\left[S(T-1)+T+1\right]c_{i,k}+\mu\left[\left(\mathbf{D}-\mathbf{W}\right)\mathbf{C}\right]_i\mathbf{C}_i^{\top}c_{i,k}-\mu\left[\left(\mathbf{D}-\mathbf{W}\right)\mathbf{C}\right]_{i,k}c_{i,k}=0\tag{3.9}$$

Then we obtain the updating rule for $\mathbf{C}$

$$c_{i,k}^{(m+1)}=\frac{c_{i,k}^{(m)}}{\eta}\left(\sum_{s=1}^{|\mathcal{S}|}\frac{\boldsymbol{\pi}_k^{(m+1)}(s)\gamma_{i,1}^{(m)}(s)}{\sum_{k'=1}^{K}\boldsymbol{\pi}_{k'}^{(m+1)}(s)c_{i,k'}}+\sum_{t=1}^{T-1}\sum_{s=1}^{|\mathcal{S}|}\sum_{s'=1}^{|\mathcal{S}|}\frac{\mathbf{A}_k^{(m+1)}(s,s')\xi_{i,t}^{(m)}(s,s')}{\sum_{k'=1}^{K}\mathbf{A}_{k'}^{(m+1)}(s,s')c_{i,k'}}\right.$$
$$\left.+\sum_{t=1}^{T}\sum_{s=1}^{|\mathcal{S}|}\frac{\mathbf{B}_k^{(m+1)}(s,o_{i,t})\gamma_{i,t}^{(m)}(s)}{\sum_{k'=1}^{K}\mathbf{B}_{k'}^{(m+1)}(s,o_{i,t})c_{i,k'}}+\mu\left(\mathbf{DC}^{(m)}\right)_i\mathbf{C}_i^{(m)\top}+\mu\left(\mathbf{WC}^{(m)}\right)_{i,k}\right)\tag{3.10}$$

where η is a normalizing variable as

$$\eta=S(T-1)+T+1+\mu\left(\mathbf{WC}^{(m)}\right)_i\mathbf{C}_i^{(m)\top}+\mu\left(\mathbf{DC}^{(m)}\right)_{i,k}.$$

The inference of the POCM is an EM algorithm, where the E-step is to compute the intermediate states $\gamma_{i,t}^{(m)}(s)$ and $\xi_{i,t}^{(m)}(s,s')$ using the forward-backward algorithm, and the M-step is to maximize $Q\left(\theta,\mathbf{C}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)$ by updating the basis model and the memberships separately. The POCM algorithm is summarized in Algorithm 3.1.

Data: observations on N individuals $\mathbf{O}_1, \ldots, \mathbf{O}_N$; initial values for the parameters $\mathbf{C}^{(0)}$, $\theta^{(0)}$; similarity matrix $\mathbf{W}$; regularization coefficient μ; number of basis models, K; stopping criteria ϵ

Result: Estimator of basis model θ^* and membership $\mathbf{C}^*$

Initialize $\mathbf{C}^{(0)}$, $\theta^{(0)}$

for $m \leftarrow 1, \ldots$ *until converge* **do**

  E-step: Compute intermediate states $\gamma_{i,t}^{(m)}(s)$ and $\xi_{i,t}^{(m)}(s, s')$ using the forward-backward algorithm;

  M-step: Fix $\mathbf{C}^{(m)}$, set $\theta^{(m+1)} = \arg\max_\theta Q\left(\theta, \mathbf{C} | \theta^{(m)}, \mathbf{C}^{(m)}\right)$ using (3.5-3.7);

  M-step: Fix $\theta^{(m+1)}$, set $\mathbf{C}^{(m+1)} = \arg\max_{\mathbf{C}} Q\left(\theta, \mathbf{C} | \theta^{(m+1)}, \mathbf{C}^{(m)}\right) - \mu \operatorname{Tr}(\mathbf{C}^\top \mathbf{L} \mathbf{C})$ using (3.10);

  if *all elements of* $\left|\nu^{(m+1)} - \nu^{(m)}\right|$ *is less than* ϵ*, where* $\nu \in \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{B}, \mathbf{C}\}$ **then**

    break;

  end

end

Algorithm 3.1: The Partially-Observable Collaborative Model (POCM) parameter inference algorithm.

Theorem 3.2. *The basis parameters and the membership converge to an optimal solution under the iterative updating rule in Algorithm 3.1.*

We present the detailed derivation of the updating rule in POCM and the proof of Theorem 3.2 in Appendix B.2.5 and B.3, respectively.

The individual treatment plan is developed in the following 3 stages. In the learning stage, the transition and emission matrices for the basis models are learned from existing treatment record of large patient population under treatment I. In the fine-tuning stage, we estimate the personal dynamics for a new patient under both treatment types by running separate short trial periods. The membership of the new patient is learned while keeping the basis

parameters learned from stage 1 fixed, i.e., only Step 2 of the POCM inference algorithm will be performed. For example, the new patient takes both treatments in two separate periods; Treatment I for T_1 periods, and Treatment II for T_2 periods. Since Treatment II may be experimental and lack existing data on large treatment records, we assume the basis transition matrices of Treatment II can be estimated using the population average treatment effect from drug trials and the basis transition matrices of Treatment I (details refer to Eqn.(3.17) in Section 3.4.1).

3.3.3 Adaptive decisions

In the third stage, the decision stage, the clinician will select treatment for each individual patient based on the policy at each period. In the POMDP model, the policy is derived by maximizing the total expected reward

$$R = \sum_{t=1}^{T} \gamma^t r(s_t, a_t), \tag{3.11}$$

where γ is the discount factor. When new observation $o_{t+1} = o$ is obtained after taking action $a_{t+1} = a$, we can update the belief by the Bayes' rule,

$$b_{t+1}(s') = \frac{\mathbf{B}(s', o) \sum_{s \in \mathcal{S}} \mathbf{A}(s, a, s') b_t(s)}{\sum_{s' \in \mathcal{S}} \mathbf{B}(s', o) \sum_{s \in \mathcal{S}} \mathbf{A}(s, a, s') b_t(s)}, \forall s' \in \mathcal{S}. \tag{3.12}$$

Treatment is selected with respect to a policy $\pi : \Delta(\mathcal{S}) \rightarrow \mathcal{A}$, where $\Delta(\mathcal{S})$ denotes the continuous set of probability distributions over $\mathcal{S}$, i.e., $a_t = \pi(b_t)$. We define value function of a policy π as $V^\pi : \Delta(\mathcal{S}) \rightarrow \mathbb{R}$, which is the expected discounted reward when following policy π starting from belief b

$$V^\pi(b) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r(b_t, \pi(b_t)) \middle| b_0 = b \right]. \tag{3.13}$$

where $r(b_t, \pi(b_t)) = \sum_{\mathbf{S} \in \mathbb{S}} r(y, \pi(b_t)) b_t(y)$. The optimal value function $V^*(b) = V^\pi(b)$ is the best value function that can be achieved with an optimal policy π^*. The Bellman equation

describes the fundamental relation between V_t and V_{t+1}:

$$V_t(b) = \max_a \sum_{s_t \in \mathcal{S}} b(s_t) \left(r(s_t, a_t) + \gamma \sum_{o_{t+1} \in \Omega} \mathbf{A}(s_t, a, s_{t+1}) \sum_{s_{t+1} \in \mathcal{S}} \mathbf{B}(s_{t+1}, x_{t+1}) V(b_{t+1}(s_{t+1})) \right). \tag{3.14}$$

The value functions in finite-horizon POMDP are piecewise-linear and convex with respect to the belief, and can be represented as

$$V_t(b) = \sum_{s_t \in \mathcal{S}} r(s, a) b_t(s) = b_t \cdot \alpha_t^a \tag{3.15}$$

where $\alpha_t^a \in \mathbb{R}^{|\mathcal{S}|}$ is a set of support vectors and $\cdot$ denotes the inner product. At period t, when the belief is b_t, the optimal action is $a_t^* = \arg\max_{a \in \mathcal{A}} b_t \cdot \alpha_t^a$. We can construct the support vector set $\{\alpha_t^a\}_{t=1}^T$ backward from $t = T$ to 1. In this paper, we use incremental pruning to accelerate the support vector enumeration [73]. The details of the algorithms are in Appendix B.4.

3.4 Simulation Experiment

To illustrate the effectiveness of the POCM in estimating individual disease progression model in a heterogeneous population, we apply the POCM inference algorithm to a simulated patient population. We demonstrate the capabilities of our method using a chronic depression example. Currently there lacks available dataset that contains both of the longitudinal observations of depression severity and the ground truth depression states. Therefore, we use simulation data with known population subgroup structure to test the performance of the POCM. Our analyses include a comparison of the performance of POCM with traditional model learning method using HMM, and a comparison of the POMDP policies using the parameters estimated from the POCM and several heuristic policies for treatment selection.

Depression, as a complex and dynamic mental disorder, is among the leading causes of disability worldwide and 7% of adults is estimated to experience depression in America [53]. We assume three health states of depression, healthy (H), mild depression (M), and severe depression (S), in ascending severity of depression, i.e., $\mathcal{S} = \{\text{H,M,S}\}$. At each time period,

patients under ongoing depression treatment will take The Patient Health Questionnaire (PHQ-9), a self-administered questionnaire, to evaluate their depression status [77]. The result of PHQ-9 is a score with range from 0 to 27, with a higher score indicating more severe condition. The PHQ-9 score can be categorized into three levels, where scores $0 \sim 4$ (P1) stand for healthy to mild depression severity, scores $5 \sim 9$ (P2) stand for mild to moderate, and scores $10 \sim 27$ stand for major depression, i.e., $\Omega = \{\text{P1,P2,P3}\}$, see Figure 3.1. The utility of each true state is denoted $\kappa(\text{H}) = 1, \kappa(\text{M}) \in (0,1), \kappa(\text{S}) \in (0,1)$, and $\kappa(\text{M}) > \kappa(\text{S})$, by the assumption that more severe depression condition will lead to lower health utility for patient. The values of $\kappa(\text{M})$ and $\kappa(\text{S})$ can be estimated using health utility elicitation methods and may vary between studies from different regions and different populations. We assume the monitoring decision epoch is 1 month.

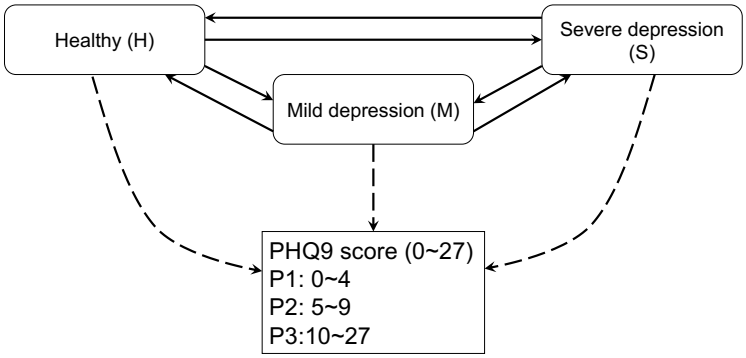


Figure 3.1: State transition and observation from states in chronic depression.

3.4.1 Model settings

We use a simulation approach to construct a set of patients with individual disease progression dynamics, which is characterized by the unique initial state belief ($\boldsymbol{\pi}_i^S$), the transition matrix ($\mathbf{A}_i^S$) and emission matrix ($\mathbf{B}_i^S$) associated with the patient $i, i = 1, \ldots, N$. Firstly, a number (K^S) of basis hidden Markov models are constructed. Without loss of generality, in what follows we use three basis models; each with 3 health states, i.e., $K^S = 3$ and

$|\mathcal{S}| = 3$. The three basis models correspond to high risk of depression progression, low risk of depression progression, and stable condition, respectively. The values of the basis parameters, i.e., the initial state belief $\left\{\boldsymbol{\pi}_k^S\right\}_{k=1}^K$, the transition probability matrices, $\left\{\mathbf{A}_k^S\right\}_{k=1}^K$, and the emission probability matrices, $\left\{\mathbf{B}_k^S\right\}_{k=1}^K$ are listed in Appendix B.5.1. We assume that the emission probability matrices and the initial distributions are the same across the basis models. Next, we generate the membership vector for each patient in a population with subgroup structure following the approach in Lin et al. [24] (see Appendix B.5.2). The subgroup structure is controlled by the parameter ζ^2 such that larger magnitude of ζ^2 corresponds to a more significant subgroup structure. Given the basis models and the membership vector, the individual parameters can be represented as: $\mathbf{A}_i^S = \sum_k c_{ik}^S \mathbf{A}_k^S$ for the transition probability matrix, $\mathbf{B}_i^S = \sum_k c_{ik}^S \mathbf{B}_k^S$ for emission probability matrix, and $\boldsymbol{\pi}_i^S = \sum_k c_{ik}^S \boldsymbol{\pi}_k^S$ for initial distribution, which we assume are the ground truth parameters. We simulate the membership for N^S subjects ($N^S = 1000$) and randomly select N^{train} subjects ($N^{\text{train}} = 800$) for learning the POCM basis models in the learning stage. The rest of the subjects ($N^{\text{test}} = N^S - N^{\text{train}}$) are treated as new patients, and we learn the individual memberships for these subjects in Step 2 (Fine-tuning). We generate observation data for training subjects for 10 periods under Treatment I, and generate observation data for testing subjects for 10 periods under Treatment I and 10 periods under Treatment II.

To obtain the initial value of membership and basis parameters in the POCM inference algorithm, we first classify the patients in the training set into $\tilde{K}$ groups. The classification is performed by the K-means methods [78], using each patient's simulated covariates as features, see Appendix B.5.3. The initial membership for each patient is the weight of the group that the patient is assigned to in the classification. We learn the basis parameters of the POCM on the observation data of the training patients, and use the population average treatment effect factor to transform the basis parameters for Treatment I to basis parameters for Treatment II Eqn. (3.17). Then the membership of the 200 testing patients can be learned by performing the fine-tuning step and using Step 2 of the POCM algorithm (i.e., update membership with fixed basis model), and thus we estimate their individual

parameters $\hat{\boldsymbol{\pi}}_i, \hat{\mathbf{A}}_i, \hat{\mathbf{B}}_i$, $i = 1, \ldots, N^{\text{test}}$. We compare the performance of POCM with HMM in modeling the individual disease progression. The learning of HMM parameters is a variant of Baum-Welch algorithm, see Appendix B.2.4. We evaluate model accuracy by examining the difference between the true transition and emission matrices and the estimated matrices of each testing patient. The difference is the Frobenius norm [79].

$$\delta_i^{\mathbf{A}} = \sqrt{\text{Tr}\left((\mathbf{A}_i^S - \hat{\mathbf{A}}_i)^\top (\mathbf{A}_i^S - \hat{\mathbf{A}}_i)\right)}, \delta_i^{\mathbf{B}} = \sqrt{\text{Tr}\left((\mathbf{B}_i^S - \hat{\mathbf{B}}_i)^\top (\mathbf{B}_i^S - \hat{\mathbf{B}}_i)\right)}, i = 1, \ldots, N. \tag{3.16}$$

We assume the patient's health condition will improve more on Treatment II than Treatment I in the long run. We describe this effect as the increased probability of healthier states in the stationary distribution. Particularly, in our model, the transition matrices of the two treatment actions are related with the following linear transformation

$$\mathbf{A}^{\text{II}} = \mathbf{U}\mathbf{A}^{\text{I}} + \mathbf{V}, \tag{3.17}$$

where

$$\mathbf{U} = \begin{bmatrix} \rho & 0 & \cdots & 0 \\ 0 & \rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, \ \mathbf{V} = \begin{bmatrix} 1-\rho & 0 & \cdots & 0 \\ 1-\rho & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}.$$

This transformation can be interpreted as moving a proportion $(1-\rho)$ of all the other elements to the first element in each row. Therefore, the treatment effect of Treatment II is parameterized by the treatment effect factor ρ, which can be estimated from comparative effectiveness trials with two treatment arms. See details in Appendix B.6.

3.4.2 Numerical results

Parameter Learning.

In the learning stage, we apply both Step 1 and Step 2 of Algorithm 1 to the observation data of the training patients to learn the basis parameters of the three subgroups. In the

fine-tuning stage, we keep the basis parameters from stage 1 and learn the memberships of each testing patient with their experimental data. The results show that the converged membership vectors for the testing patients contains only 1 or 0. Therefore, each new patient is assigned to one basis group after the fine-tuning stage, even though we assume the individual model could be a combination of the basis models. As comparison, we also apply the Baum-Welch algorithm to estimate the basis HMM model for each group. For the testing patients, we first assign each of them to the basis group that has the most similar average profile to the patient, then learn the individual HMM model by using the group HMM parameters as initial values.

We compare the performance of the POCM parameter inference and HMM inference using the population average $\delta_{\cdot}^{\mathbf{A}}$ and $\delta_{\cdot}^{\mathbf{B}}$ Eqn.(3.16) in both the learning and the fine-tuning stages. The results are based on various levels of significance of latent structure and the number of health states. In Figure 3.2(a), we can see that POCM performs well in very weak structure, and generally performs better than HMM Baum-Welch in estimating the transition matrix. In Figure 3.2(b), we can see that for both the transition matrix and the emission matrix, and in both the learning stage and the fine-tuning stage, the estimation error of POCM decreases when number of states increases, and its performance is better than the HMM on large number of states. Therefore, when choosing disease progression models for complex diseases with a large number of states, POCM may be better than HMM.

Treatment Policy Evaluations.

In the decision stage, we apply a set of treatment policies to each testing patient. At each decision epoch, the treatment type will be determined by the policy. We will consider the policy derived from POMDP and other heuristic policies. For POMDP policies, the transition and emission matrices are either estimated from the HMM with the Baum-Welch Algorithm, or from the POCM with the POCM inference algorithm. There are two evaluation criteria, the expected total reward (NMB) $\frac{1}{N}\sum_{t=1}^{T}\sum_{j=1}^{N}\gamma^t r(s_{jt}, a_{jt})$, and the prediction accuracy $\frac{1}{NT}\sum_{t=1}^{T}\sum_{j=1}^{N} b_{jt}(s_{jt})$, where N is the number of patients and T is the number of treatment

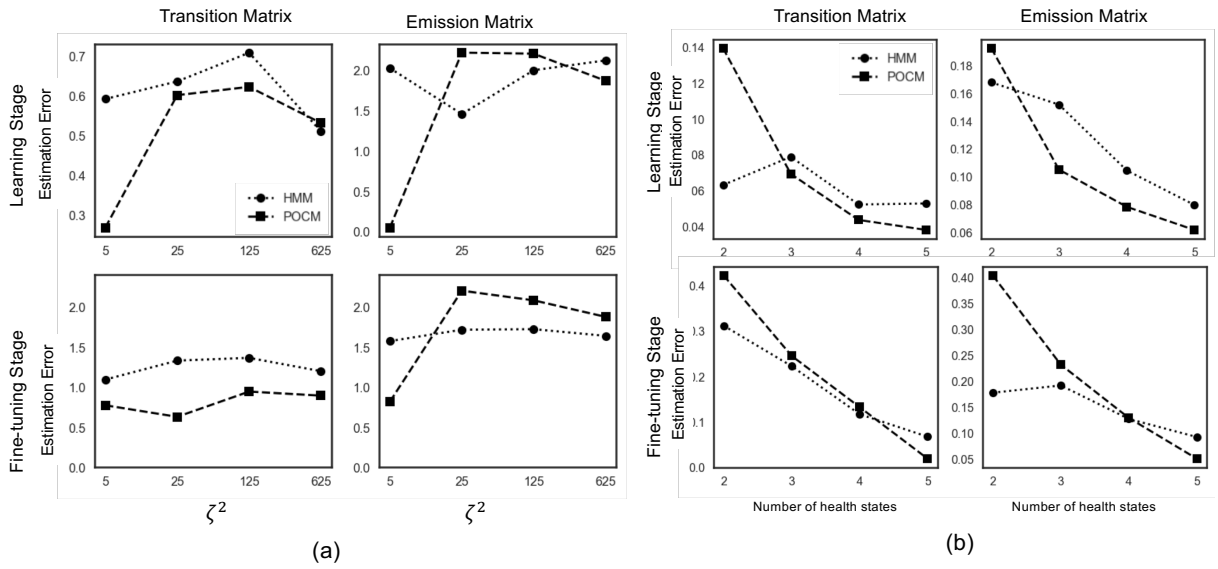


Figure 3.2: Sensitivity analysis on the converged estimation error between the POCM and the HMM for individualized depression progression modeling. (a) Effect of latent structure significance, $\zeta^2 = 5, 25, 125, 625$; larger value of ζ^2 indicates strong latent structure. (b) Effect of number of health states, including 2, 3, 4, 5; the average matrix distance is divided by the number of elements in each matrix.

periods (N=200, T=24). A good policy is defined as a policy with low prediction error and high expected total reward. The alternative policies include:

Random policy: At each period, choose the action randomly.

Observation policy: At each period, if the patient's PHQ-9 score is greater than a threshold, then Treatment II is selected; otherwise, Treatment I is selected.

Bayesian policy: At each period, after updating the belief state by the Bayesian rule in Eqn.(3.12), if the belief probability on the healthy state (H) is smaller than the threshold $\tilde{b}$(H), Treatment II is selected; otherwise, Treatment I is selected. The thresholds in our experiment are 0.5 and 0.9. Larger threshold stands for more aggressive treat-

ment policies, i.e., switch to better treatment at lower health risk. Another set of Bayesian policies are based on the belief of the severe depression state (S): if the belief of S is larger than the threshold $\tilde{b}(\mathrm{S})$, then choose Treatment II. The thresholds in our experiment are 0.1, 0.5. Smaller threshold stands for more aggressive treatment policies.

We apply 10 policies in the decision stage in total, which are listed in Table 3.1.

Table 3.1: The policies to be examined in the decision stage.

	Short name	Description
1	`random`	Select Treatment II with a probability of 0.3, independent of the actions and observation.
2	`pomdp_true`	POMDP policy using the ground-truth individual parameters.
3	`pomdp_pocm`	POMDP policy using the individual parameters estimated from POCM.
4	`pomdp_hmm`	POMDP policy using the individual parameters estimated from HMM.
5	`s_0.1`	Bayesian policy: select Treatment II when $b(\mathrm{S}) \geq 0.1$
6	`s_0.5`	Bayesian policy: select Treatment II when $b(\mathrm{S}) \geq 0.5$, less aggressive
7	`h_0.9`	Bayesian policy: select Treatment II when $b(\mathrm{H}) \leq 0.9$
8	`h_0.5`	Bayesian policy: select Treatment II when $b(\mathrm{H}) \leq 0.5$, less aggressive
9	`o_5`	Observation policy: select Treatment II when the PHQ-9 is greater than 5
10	`o_10`	Observation policy: select Treatment II when the PHQ-9 is greater than 10, less aggressive

Base-Case Treatment Outcomes.

We display the performance of the 10 policies in Figure 3.3. Each policy is applied to all the testing patients on 20 independent repeating runs. The parameters for the base case are: (1) strong latent structure, $\zeta^2 = 125$; (2) the initial state $s_0 = \text{H}$; (3) the utility structure $[\kappa(\text{H}), \kappa(\text{M}), \kappa(\text{S})] = [1.0, 0.4, 0.1]$ (i.e., QALY of living in each state for one year); (4) the cost structure $c(\text{I}) = \$1,000$, $c(\text{II}) = \$2,000$; (5) WTP $\lambda = \$50{,}000$/QALY; (6) strong intervention effect $\rho = 0.2$. A point on the upper right corner of Figure 3.3 indicates a better policy, with higher prediction accuracy and higher rewards. We can see a trend that the policy with larger reward tend to have higher prediction accuracy. The POMDP policies using the ground truth and estimation with the POCM have the highest rewards. In addition, the POMDP policy using the POCM estimation has the highest prediction accuracy. For the Bayesian policies, `h_0.9` and `s_0.5` are very close and show higher rewards, and `s_0.1` and `h_0.5` are very close. The more aggressive observation policy (i.e., `o_5`) also shows high reward, which means selecting Treatment II when the patient shows mild depression symptom can be an effective policy.

Sensitivity Analyses.

We include detailed one-way sensitivity analyses on model parameters in Appendix B.7, where each parameter is varied individually to examine the policies' performance in each setting. When the policy selects the more effective Treatment II, the health state is expected to improve more in the future, but the cost also increases. Since the reward is in the form of net monetary benefit which combines the QALY and cost, we want to investigate the relationship between Treatment II and the average reward. We plot the proportion of Treatment II in the treatment horizon vs. the average reward, and the average health state, respectively in Figure 3.4. We assign the number 0, 1, 2 to the health state H, M, S respectively. Each color stands for a specific model setting. The results in Figure 3.4 shows that when a policy tends to select more Treatment II, the average health state

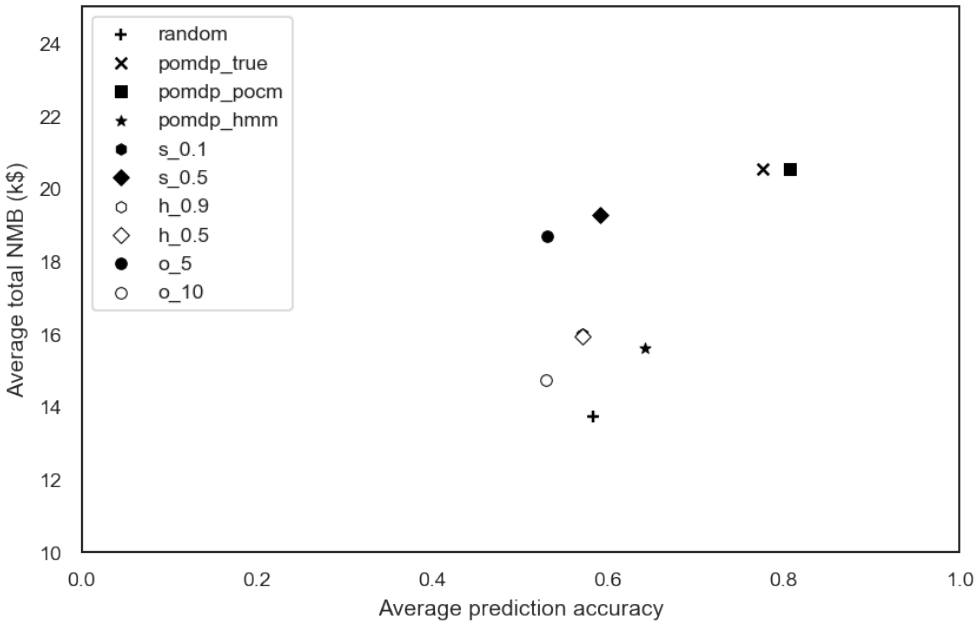


Figure 3.3: Comparison of different policies, average prediction error vs. average reward. A point to the upper left corner indicates a better policy. With the initial state $s_0 = \text{H}$, the utility structure $[\kappa(\text{H}), \kappa(\text{M}), \kappa(\text{S})] = [1.0, 0.4, 0.1]$, the cost structure $c(\text{I}) = \$1,000, c(\text{II}) = \$2,000$, WTP $\lambda = \$50,000/\text{QALY}$ and strong intervention effect $\rho = 0.2$.

is healthier, and the average total reward is higher. Among all the policies, the POMDP policies in general tend to select more Treatment II than all the other policies under all model settings. Therefore, when the objective is to maximize the NMB, the POMDP policies are good performers. This conclusion is robust to parameter uncertainty under the one-way sensitivity analyses. The POMDP policies produce higher proportions of Treatment II than other policies, except for the models with weak intervention effect and higher utility in depressive states. We note that in these two models, the improvement on health outcome under Treatment II is reduced. Therefore, the POMDP policies tend to avoid Treatment II to reduce the treatment cost, while the reduction in health outcome is not significant.

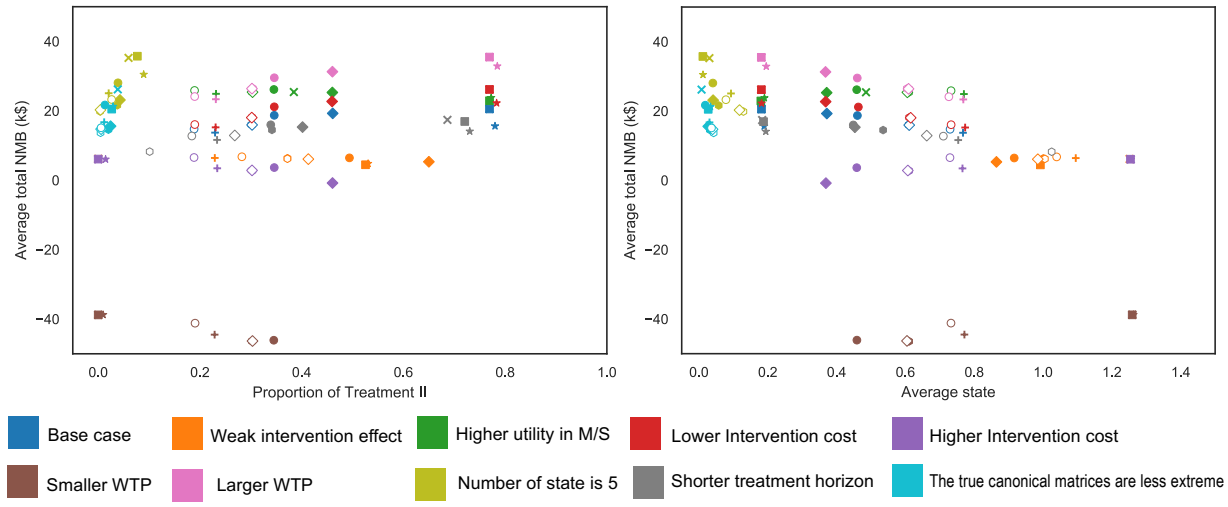


Figure 3.4: The relation between proportion of Treatment II, the average reward, and the average state. Each color stands for a specific model setting. Each dot stands for a policy, the symbols follow the same definition in Figure 3.3.

Treatment Switching.

In the decision stage, a policy switches between the two types of treatment based on the observations from previous periods and the belief on the current health state. The number of treatment switches for the 10 policies is displayed in Figure 3.5. We can see that all the POMDP policies have a mean of about 4 switches in 24 periods, which is less frequent than other policies. In practice, frequent switches between different types of treatment may not be feasible, since patients may need some time to adjust to a new treatment or maintain continuity of care. Therefore, from the perspective of reducing the frequency of switching between treatment types, POMDP policies may be preferred.

Subgroup Analysis.

We are also interested in investigating the policy performance difference by subgroup of the testing patients. We divide the 200 testing patients into four groups by their ground truth

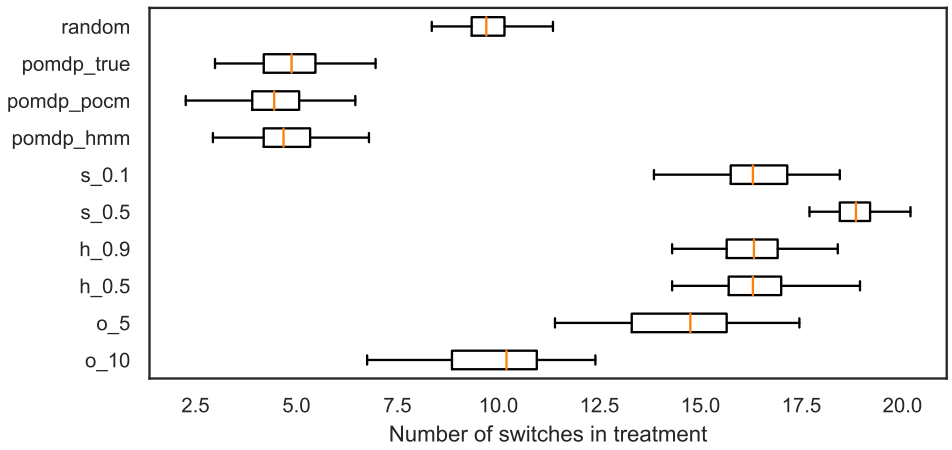


Figure 3.5: The boxplot for the number of switches in treatment type during the decision stage for different policies. The box plot shows the minimum, the 25 percentile, the median, the 75 percentile and the maximum of the number of switches for the testing patients.

memberships. The first three groups are those with membership close to 1 on one of the basis models: high risk, low risk, and stable. The fourth group are patients with no extreme membership on any basis model. In Figure 3.6, we show the performance outcomes by subgroups, including the group-averaged number of switches in treatment type, the proportion of Treatment II, and the total NMB. There are no significant differences on the number of switches between the 4 groups. Among the POMDP policies, the high risk group receives more Treatment II, and has a relatively smaller total NMB.

3.5 Conclusions and Future Work

In this paper, we proposed a quantitative framework to build individualized chronic disease progression models for optimal treatment selection in a heterogeneous population. The model, POCM, has the following features: (1) The health state is not fully observable, which is a common scenario in chronic diseases; (2) There is a set of basis models, each representing a unique pattern of disease progression; (3) Treatment is tailored to individual patient by

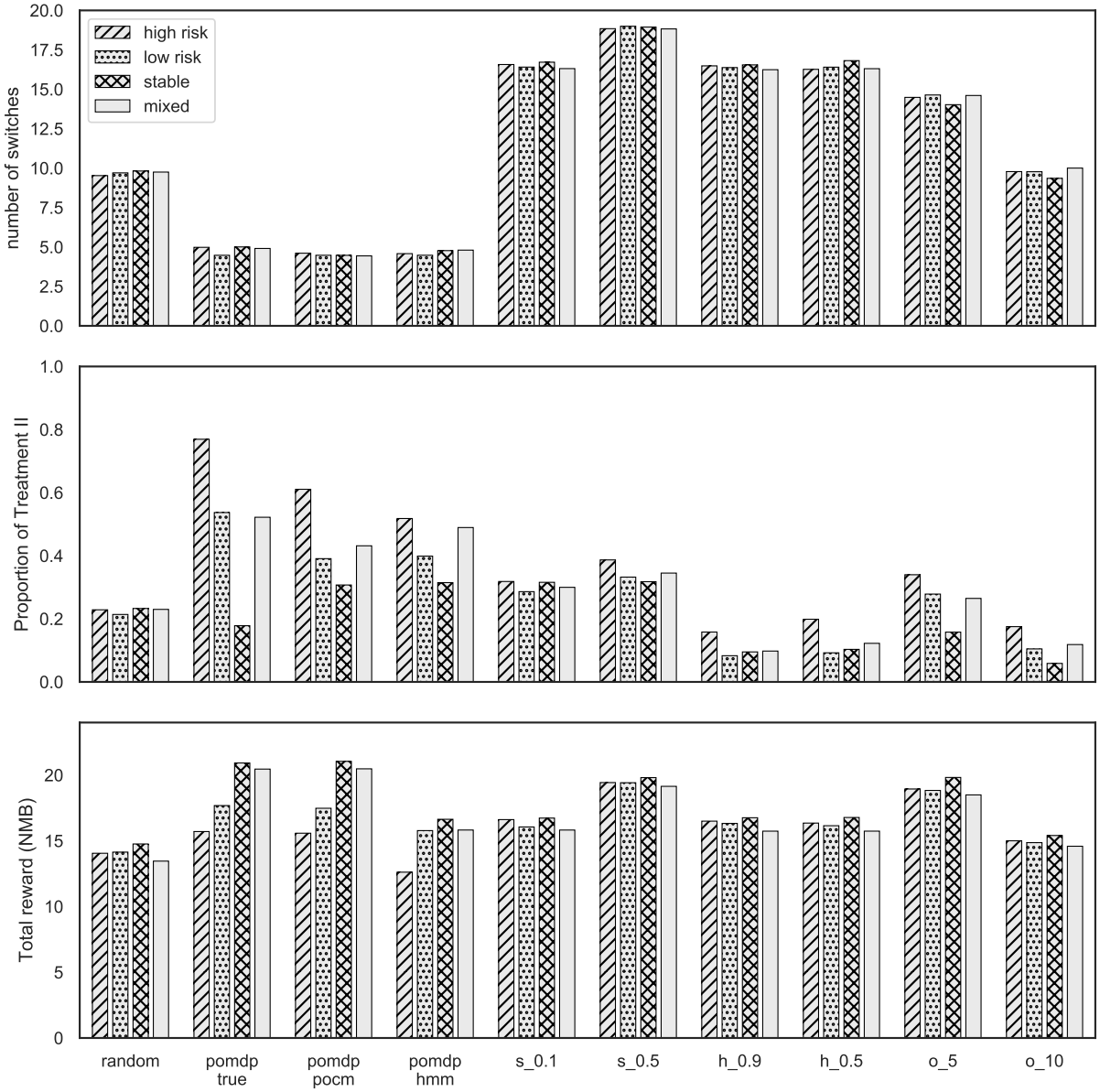


Figure 3.6: The difference of policy performances for each subgroup, including the group-averaged number of switches in treatment type, the proportion of Treatment II, and the NMB.

learning his/her personal disease progression model. We develop an efficient computational algorithm to estimate the parameters of the POCM model.

We designed a simulation study on chronic depression to demonstrate that the proposed POCM methodology can lead to better performance on parameter estimation over standard estimation method of the HMM. In addition, we evaluated the performance of several adaptive treatment policies (POMDP policies and Bayesian policies) and simple heuristic policies based on immediate past observations. The POMDP policies explicitly trade off the treatment benefit and cost by maximizing the NMB as the objective. Applying all the policies to the 200 testing patients in the decision stage, we assessed their performance in NMBs and prediction error. The three main conclusions are as follows: (1) The POMDP policies give higher rewards and lower prediction errors compared to other policies under various model settings. Among the POMDP policies, the policy with the disease progression model learned from POCM has the best performance; (2) The POMDP policies tend to select more effective treatment, which leads to a better average health state and higher reward compared to other policies; (3) The POMDP policies produce fewer switches between treatment types, which reduce the risk associated with treatment switching.

There are several directions to expand this research. First, although we only presented the possibility of applying the POCM to one disease application in chronic depression, POCM can be applied to a wider range of chronic diseases that meet similar assumptions on partial-observable health state, the disease progress is Markovian, and long treatment duration with treatment switching options. In addition, POCM is not limited to medical decision-making problem. Take machine maintenance as an example. The state of the machine can change over time, and the probability of state transition is different for each individual machine. Therefore, the health progression of an individual machine can be modeled with POCM, and machine maintenance policies can be tailored to an individual machine by using the basis model and membership learned by POCM. Another example is personalized health management by longitudinal planning from daily behavioral data [80, chap. 5]. The fast-growing development of sensing devices enables the continuous monitoring of human behavior

(such as physical activity and food intake), and health state measurements such as the body mass index (BMI). A personalized health management program such as obesity prevention can be achieved by learning the basis behavior models with POCM, which lead to individual behavior model, and then we can find the optimal plan of health activity via POMDP.

The POCM method in this paper is limited due to several strong assumptions, which can be relaxed to form new models. One limitation is that the transformation of the transition probability from Treatment I to Treatment II is linear, and it can be described by one parameter. It is possible to model the transformation using linear transformation with more parameters, or with non-linear transformation. In future research, we could also include more than two types of treatment, which would increase the complexity of both the parameter learning algorithm and the treatment selection process.

In summary, we developed a framework to represent the individual model of chronic disease progression and constructed a dynamic treatment plan based on learned individual models. We designed a simulation study on chronic depression, which shows that the proposed methodology can lead to better performance on parameter estimation over the traditional method. This framework is promising for modeling the chronic disease progression process and developing a personalized adaptive treatment plan for individuals in a heterogeneous population.

Chapter 4

ROBUST PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES WITH UNCERTAIN PARAMETERS

In the previous chapters, we have established an individual disease progression model and a personalized treatment selection plan for patients in a heterogeneous population. The disease progression model POCM requires the estimation of basis progression models, which represent the basic types of disease progression in the population, and the individual memberships, which represent the similarity that one patient resembles to each basis group. The basis models are estimated from large scale trajectories data of the population. However, the estimation of the individual membership is derived from small experiment data. Therefore, although we can be confident that the basis models are accurate, we are not as confident about the accuracy of the individual membership and hence the individual progression model. Inaccurate individual progression model will reduce the performance of the POMDP to derive optimal treatment selection plans for each individual. Therefore, we need a robust treatment selection strategy with the ability to find optimal policies for an individual patient when the disease progression model is not sufficiently accurate. In this chapter, we introduce two types of robust POMDPs which generalizes a standard POMDP by allowing uncertainties in the reward function and transition probabilities. The first model assumes parameters bounded with linear constraints, and the second assumes the transition probabilities follow Dirichlet distribution. We develop value iterations for each type of the POMDP to find optimal solutions when the worse case (or nearly the worst case) of the uncertain transition probability is considered. Finally, we illustrate the effectiveness of our approach using a case study in designing a personalized treatment plan for chronic diseases in a heterogeneous population.

4.1 Introduction

Partially observable Markov decision process (POMDP) is a general framework for sequential decision-making under uncertainty [81, 82, 83]. A general decision-making process involves an agent interacting with the environment by taking actions with a set of rules (or policies) and the observed states of the environment. The Markov decision process (MDP) assumes that the state transition behaves like a Markov chain and the observations represent the true states of the environment. The POMDP extends the MDP by relaxing the second assumption such that the true states are not observable. Instead, the states are modeled as latent variables, and most algorithms of finding the optimal policy of POMDP involve estimating the states of the environment. However, in many real-world situations, the exact values of POMDP parameters (including the rewards and transition dynamics) are usually inaccessible. In practice, these parameters are either estimated from the observational data or from domain knowledge which will include errors. If the policy derived from the POMDP with the estimated parameters is evaluated under the true model parameters, the performance can be much worse than expectation [84]. Therefore, robust POMDPs are needed to mitigate such uncertainty in model parameters.

Applications of POMDPs include medical decision-making [85, 86, 87, 88], inventory control [89, 90], revenue management [91], machine replacement [92, 93, 94, 95, 96, 97], and so on. However, having explicit knowledge of the POMDP parameters, such as the transition probabilities, is a questionable assumption for these applications. Since robust POMDP extends POMDPs by relaxing this assumption, they provide a useful framework to make more realistic and robust decisions in a variety of applications. We are motivated by the individualized treatment problem for chronic diseases (e.g., depression, obesity, Alzheimer's disease). In the literature of medical decision making, POMDPs are widely used for designing optimal treatment and screening strategies, with transition dynamics defined using longitudinal observational data and domain expertise. however, in the case of modeling individual disease progression model, the parameters are subject to errors due to the high cost of longitudinal

data acquisition.

Robust POMDP is a special case of robust optimization. Robust optimization is an approach to solve optimization problems under uncertainty. The key idea is to define an uncertainty set of possible realizations of the uncertain parameters and then optimize against worst-case realizations within this set [98, 99]. This is called the pessimistic criterion. However, the pessimistic criterion has been criticized as too conservative, that for every feasible point only the worst case in the uncertainty set is considered. The agent may alternatively choose the opposite criterion (the optimistic criterion), that the best-case model in the uncertainty set is used in optimization. The choice of criterion depends on the application. The uncertainty set can be viewed as a set of hard constraints for the POMDP parameters, such that constraint violation cannot be allowed for any realization of the data in the uncertainty set. The agent can only select between pure pessimistic or pure optimistic criterion. An alternative way of modeling the parameter uncertainty is to treat the uncertain parameters as random variables. This representation is a set of soft constraints, such that the region with small probability density represents the less realizable region, similar to the outside of the uncertainty set. The agent can find a policy through a chance-constrained problem that obtains a $1 - \epsilon$ guarantee that the policy will perform better than the objective [100]. This percentile criterion provides the agent with the flexibility to find robust policies with a mixture of optimistic view and pessimistic view.

The main question we are trying to answer in this chapter is: how can we improve the standard dynamic programming methods of finding the optimal policy of POMDPs to account for parameter uncertainty? We introduce two types of uncertainty for the transition probability: the first is a set of linear hard constraints such that the transition probability lies in the uncertainty set; the second is a distribution of the transition probability indicating the region where the transition probability is more likely to appear. We extend the classic method for solving the POMDP with dynamic programming, Value Iteration (VI), to include the uncertainty of transition parameters. We then test these methods on a simple problem instance. Finally, we establish the effectiveness of the robust POMDP using a case study

that addresses uncertainty in the context of personalized treatment of chronic depression in a heterogeneous population.

This paper contributes to the robust POMDP literature by comparing two formulations of robust POMDP: (1) the linear-constrained POMDP (LC-POMDP) with uncertainties in the transition probability in the form of a bounded region in the probability simplex, and (2) the chance-constrained POMDP (CC-POMDP) with uncertainties in the transition probability in the form of the Dirichlet distribution in the probability simplex. The LC-POMDP provides the agent with the choice of optimistic and pessimistic criterion and the CC-POMDP with the level of optimism available to select by the agent. We describe the method of constructing policy trees from the value function vectors to avoid the updating of beliefs which is infeasible without the exact value of the transition probability. We then develop the value iteration for LC-POMDP and CC-POMDP. We explore the special conditions that the LC-POMDP will degenerate to a classic POMDP.

The rest of the chapter is organized as follows. In Section 4.2, we briefly review the related studies. The classic POMDP and the existing solution approaches are discussed in Section 4.3. The LC-POMDP and CC-POMDP frameworks are presented in Section 4.4 and 4.5 respectively. Section 4.6 demonstrates the effectiveness of the methods on a set of random instances of POMDPs. Section 4.7 provides the applications of robust POMDP in simulated personalized treatment of chronic depression. Section 4.8 provides the concluding remarks.

4.2 Related work

The POMDP model was first introduced in Drake [101]. Sondik [102] solved the POMDP with dynamic programming by showing the property of the piecewise-linearity and convexity (PWLC) of the value function. This result was extended for finite and infinite-horizon problems by Smallwood and Sondik [103] and Sondik [104] respectively. In Lovejoy [105], Rieder [106] and Lovejoy [107] the structural results of the monotonicity of the value function in the belief state and the convexity of policy regions in POMDPs is discovered. One of our

main goals in this work is to extend such structural results from POMDPs to robust POMDPs when parameter uncertainty is unavoidable.

Another stream of research on MDP/POMDP focused on decision-making under uncertainty. There are two types of formulation of the parameter uncertainty. The first group of papers assume that parameters lie in a given uncertainty set, see Nilim and El Ghaoui [108], Iyengar [109], Givan et al. [110], Bagnell et al. [111]. The robust MDP is treated as a robust optimization problem, and the policy are generated through robust dynamic programming. Wiesemann et al. [112] derived a confidence region that contains the unknown transition probabilities with a pre-specified probability which is used to find a policy that attains the highest worst-case performance. A similar approach was shown in Delage and Mannor [113], while the author also provided the result on uncertain rewards. These solutions, however, can be overly conservative since they are based on worst-case realization. Variants of robust MDP formulations have been proposed to mitigate the conservativeness when additional information on parameter distribution [114, 115] or coupling among the parameters are available [116]. The first model in our work assumes that the upper and lower bounds of the parameters are known to the agent from prior knowledge. The problem with the parameter bound formulation is that the agent chooses the best strategy for the worst-case scenario, which is generally overly conservative strategies.

Although robust MDP has been studied for a long time, papers on robust POMDPs are sparse. Methods of robust POMDP are either introduced from robust MDP by including the partial observability assumption, or from the POMDP solution methods by allowing parameter uncertainty. Itoh and Nakamura [117] proposed a robust POMDP and obtained a belief-state MDP using the concept of second-order belief. However, they consider the initial belief as part of the model. In addition, the computational complexity of the belief-state MDP is exponential in the number of states to the original POMDP. Saghafian [118] proposed a different formulation of robust POMDP that consider both imperfect state information and ambiguity regrading the nominal probabilistic model, in which the robustness is achieved by using α-maximin expected utility (α-MEU). Ni and Liu [119] proposed a POMDP with

imprecise but bounded parameters and developed a modified value iteration to find robust policies. The LC-POMDP in Section 4.4 extends their model to allow any form of linear constraints in the transition dynamics.

An alternative way is treating the parameters as random variables and use the Bayesian approach. This approach does not require the assumption that parameters lie in a bounded uncertainty set. Mannor et al. [84] introduced the first MDP model with transition probability as a random variable. Such framework can lead to a performance measure called the percentile criterion, which exploits the trade-off between optimistic and pessimistic strategies when facing parameter uncertainty. Percentile optimization for MDP was first developed in Delage and Mannor [100] where the percentile criteria are studied under different forms of uncertainty for both the rewards and the transitions probability. Their model can also be viewed as an extension of the chance-constrained criterion for single-period optimization problems to multi-stage decision problems. The CC-POMDP in Section 4.5 applies the percentile criterion to POMDP as an extension of the model in Delage and Mannor [100] with latent states.

4.3 The classic POMDP and policy tree

POMDP is a common framework for modeling sequential decision-making while the state of the process is not observable. A discrete-time finite-horizon, discounted-reward POMDP with finite actions and states is defined by a tuple $M = \langle \mathcal{S}, \mathcal{A}, \Omega, \mathbf{A}, \mathbf{B}, R, g, \gamma \rangle$, where $\mathcal{S}$, $\mathcal{A}$ and Ω are respectively the sets of states, actions, and observations. The transition probability $\mathbf{A}(s, a, s') = \Pr(s'|s, a)$ gives the probability of ending in state s', given that the agent takes action a in state s. The emission probability $\mathbf{B}(s', a, o) = \Pr(o|s', a)$ gives the probability of making observation o, given that the agent takes action a and arrives at state s'. The reward function $R(s, a)$ denotes the expected reward of taking action a at state s. The terminating reward $g(s)$ denotes the reward received when the process ends in state s. Future rewards are discounted at a rate of $\gamma \in (0, 1]$. We only consider POMDPs with finite horizon $T < \infty$. The total reward is the realized value of the sequence of actions over the planning horizon:

$r(T) = \sum_{t=1}^{T} R_t(s_t, a_t)$. Since the true state is not observable, the agent maintains a belief of the system state $b(s)$, $\forall s \in \mathcal{S}$, which is a probability distribution over the state space. A policy is a rule of choosing the action based on the belief of the current state $\pi : \Delta(S) \rightarrow A$. The goal of solving a POMDP is to find a policy $\pi^\star$ that maximizes the expected future discounted reward $V^\pi(b)$ the agent can gather by following π starting from belief b

$$V^\pi(b) = E_\pi \left[\sum_{t=0}^{T} \gamma^t \tilde{R}(b_t, \pi(b_t)) \middle| b_0 = b \right], \tag{4.1}$$

where $\tilde{R}(b_t, \pi(b_t)) = \sum_{s \in \mathcal{S}} R(s, \pi(b_t)) b_t(s)$.

The most common approach to find the optimal policy of POMDPs is Value Iteration (VI), a dynamic programming routine that builds a sequence of value-function estimates which converge to the optimal value function [102]. The value of an optimal policy $\pi^\star$ is defined by the optimal value function $V^\star$ computes through iteration of several stages. At each stage, we only consider a step further into the future, as described in the Bellman equation

$$V_t = HV_{t+1}, \tag{4.2}$$

where H is the Bellman backup operator for POMDP, defined as

$$V_t(b) = \max_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} b(s) \left[R(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathbf{A}(s, a, s') \sum_{o \in \Omega} \mathbf{B}(s', a, o) V_{t+1}(b(s')) \right]. \tag{4.3}$$

Here $b(s')$ is a updated belief using the Bayes' rule, given the previous belief state as well as the previous action and observation,

$$b(s') = \frac{\mathbf{B}(s', a, o) \sum_{s \in \mathcal{S}} \mathbf{A}(s, a, s') b(s)}{\sum_{s' \in \mathcal{S}} \mathbf{B}(s', a, o) \sum_{s' \in \mathcal{S}} \mathbf{A}(s, a, s') b(s)}, \forall s' \in \mathcal{S}. \tag{4.4}$$

The value function has a particular structure that it can be parameterized by a finite number of vectors and has a piece-wise linear convex (PWLC) shape [102]. For stage t, we can represent the value function V_t by a finite set of vectors or hyperplanes $\Gamma_t = \{\alpha_t^k\}, k = 1, \ldots, |V_t|$. Given a set of vectors $\Gamma_t = \{\alpha_t^k\}_{k=1}^{|V_t|}$, at stage t, the value of a belief b is given by

$$V_t(b) = \max_{\Gamma_t} b \cdot \alpha_t^k \tag{4.5}$$

For each vector there is an action $a(\alpha_t^k) \in \mathcal{A}$ indicating the optimal action to take in the current step, where α_t^k is the maximizing vector. The vectors resulting from back-projecting α_t^k for a particular a and o is

$$g_{ao}(s) = \sum_s \mathbf{A}(s, a, s')\mathbf{B}(s', a, o)\alpha_t^k(s'). \tag{4.6}$$

Monahan [120] proposed the most straightforward way of vector backup by calculating all possible ways HV_t could be constructed, $\Gamma_t = \texttt{backup}(\Gamma_{t+1})$, by exploiting the PWLC structure of the value function:

$$HV_t = \bigcup_a \Gamma_a, \Gamma_a = \bigoplus_o \Gamma_a^o, \Gamma_a^o = \{R(a)/|\Omega| + \gamma g_{ao}^k\}_k, \tag{4.7}$$

where $\oplus$ denotes the cross-sum operator: $\bigoplus_k R_k = R_1 \oplus R_2 \oplus \cdots \oplus R_k$, with $P \oplus Q = \{p + q | p \in P, q \in Q\}$. This process generates all possible combinations of $|\Omega|$ vectors α_t from Γ_t, which is a finite but exponential number of vectors $|A||V_t|^{|\Omega|}$. The regions of many of the generated vectors will be empty, and these vectors are useless as they will not influence the agent's policy. Therefore, all value-iteration methods in the enumeration family employ some form of pruning. In particular, Monahan [120] prunes HV_t after computing it

$$V_t = \texttt{prune}(HV_{t+1}), \tag{4.8}$$

where HV_t is the Bellman backup operator defined in (4.3). The prune operator is implemented by solving a linear program [121]. Incremental Pruning methods [73, 122] save computation time by exploiting the fact that

$$\texttt{prune}(\Gamma \oplus \Gamma' \oplus \Gamma'') = \texttt{prune}(\texttt{prune}(\Gamma \oplus \Gamma') \oplus \Gamma''). \tag{4.9}$$

In this way the number of constraints in the linear program used for pruning grows slowly [73], leading to better performance. The basic Incremental Pruning algorithm exploits (4.9) when computing V_{t+1} as follows

$$HV_t = \texttt{prune}(\bigcup_a \Gamma_a), \text{with} \tag{4.10}$$

$$\Gamma_a = \texttt{prune}(\bigoplus_o \Gamma_a^o) = \texttt{prune}(\cdots \texttt{prune}(\texttt{prune}(\Gamma_a^1 \oplus \Gamma_a^2) \oplus \Gamma_a^3) \cdots \oplus \Gamma_a^{|\Omega|}), \tag{4.11}$$

We denote the set of vectors before pruning as Γ_t^+, and the set of vectors after pruning as $\Gamma_t^\star$. The process of one-step back up is $\Gamma_t^+ = \texttt{backup}(\Gamma_{t+1}^\star)$, and $\Gamma_t^\star = \texttt{prune}(\Gamma_t^+)$.

An alternative way to represent policies in POMDPs is based on the notion of policy tree, in which each node denotes an action and each arc denotes an observation [83]. Figure 4.1 shows a policy tree with T levels, in which the agent starts at the root node of the tree. Each node specifies an action which the agent executes at the particular node. Next, it receives an observation o, which determines what next node the agent transitions to. The depth of the tree depends on the planning horizon T. If we want the agent to consider taking T steps, the corresponding policy tree has depth T. Repeating the above operations forms a path from the root node to one leaf node, which specifies the history the agent experiences. When a policy tree is executed, for each t, each possible history with length t corresponds to one node at level t of the policy tree, and the history can be traced by the path from the root node to the corresponding node. Therefore, the policy tree representation for POMDP avoids estimating the belief of state when executing the policy.

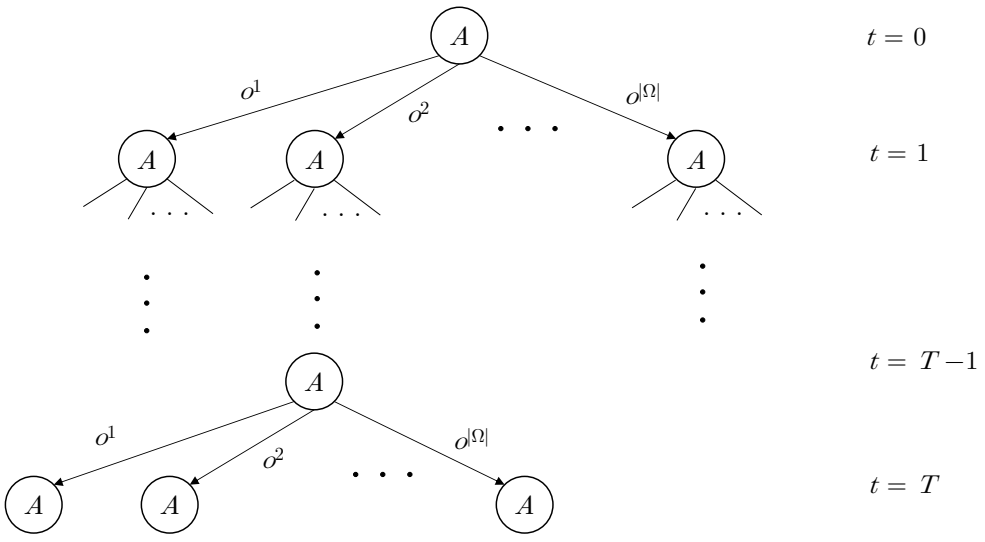


Figure 4.1: A T-level policy tree. The top level represents the initial actions. A policy tree is constructed from the bottom level to the top level.

The expected value of executing the t-step policy tree pt from state s is

$$V_{pt}(s) = R(s, a(pt)) + \gamma \sum_{s' \in \mathcal{S}} \mathbf{A}(s, a(pt), s') \sum_{o^i \in \Omega} \mathbf{B}(s', a(pt), o^i) V_{o^i(pt)}(s'), \tag{4.12}$$

where $a(pt)$ is the action at the root node of pt, $o^i(pt)$ is the $(t+1)$-step subtree following arc o^i [119]. The value of executing pt for initial belief b is $V_{pt}(b) = \sum_{s \in \mathcal{S}} b(s) V_{pt}(s)$, and the value function is $V_t(b) = \max_{pt \in PT_t} V_{pt}(b)$, where PT_t denotes the set of t-step policy trees. For each pt, V_{pt} can also be represented by the α vectors. Note that the value backup (4.6, 4.7) can be combined as

$$HV_t = \bigcup_a \Gamma_a, \qquad \Gamma_a = \left\{ \alpha_t(s, a; o^1, \ldots, o^{|\Omega|}) \middle| \forall o^i \in \Omega \right\}, \tag{4.13}$$

$$\alpha_t(s, a; o^1, \ldots, o^{|\Omega|}) = R(s, a) + \gamma \sum_{s'} \mathbf{A}(s, a, s') \sum_i \mathbf{B}(s', a, o^i) \alpha_{t+1}^{o^i}(s), \tag{4.14}$$

where the set of $|\Omega|$ vectors $\alpha_{t+1}^{o^i}$ can be any combinations of the $|\Gamma_{t+1}^{\star}|$ vectors in $\Gamma_{t+1}^{\star}$. Therefore, each vector after one-step backup (before pruning) corresponds to a set of vectors from $\Gamma_{t+1}^{\star}$ that maps to an observations. A policy tree pt can be represented as a tuple $\langle a, \alpha_{t+1}^{o_1}, \ldots, \alpha_{t+1}^{o_{|\Omega|}} \rangle$, where a is the action assigned to the root node and $\alpha_{t+1}^{o^i} \in \Gamma_{t+1}^{\star}$ represents the estimated value of executing the $(t+1)$-step subtree following o^i. In the rest of this paper, we denote a policy tree by its corresponding $(|\Omega|+1)$-tuple. Constructing a t-level policy tree from a $t+1$ level policy tree is equivalent to one-step backup.

4.4 Linear-Constrained POMDP (LC-POMDP)

A first robust model of POMDP is to assume that the parameters, including the transition probability, the emission probability, and the reward function, are bounded by a known upper and lower limit. Ni and Liu [119] proposed the first robust POMDP with imprecise but bounded parameters, bounded-parameter POMDP. However, they ignore the fact that the transition vector for each state sums to 1, which is vital to find special structures in the solution of the optimal policy. Therefore, we introduce the Linear-Constrained POMDP (LC-POMDP), an improved model of robust POMDP that is able to model the imprecise

parameters with any linear constraints. The LC-POMDP is defined as a set of POMDPs, and is denoted by a tuple $\tilde{M} = \langle \mathcal{S}, \mathcal{A}, \Omega, \tilde{\mathbf{A}}, \mathbf{B}, \tilde{R}, g, \delta, \gamma \rangle$, where

- $\mathcal{S}, \mathcal{A}, \Omega, \mathbf{B}, g$, and γ are the same as those defined in POMDPs.
- $\tilde{\mathbf{A}} = \left\langle \underline{\mathbf{A}}, \overline{\mathbf{A}} \right\rangle$ denotes the set of the state-transition probabilities of POMDPs in $\tilde{M}$, where $0 \leq \underline{\mathbf{A}}(s, a, s') \leq \overline{\mathbf{A}}(s, a, s') \leq 1, \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$. The uncertainty set of transition probabilities is defined as

$$u(\mathbf{A}(s,a)) = \{\mathbf{A}(s,a,\cdot) \,|\underline{\mathbf{A}}(s,a,s') \leq \mathbf{A}(s,a,s') \leq \overline{\mathbf{A}}(s,a,s'), \forall s, s' \in \mathcal{S}, a \in \mathcal{A} \text{ and } \sum_{s' \in \mathcal{S}} \mathbf{A}(s,a,s') = 1, \forall s \in \mathcal{S}, a \in \mathcal{A}\}. \tag{4.15}$$

- $\tilde{R} = \left\langle \underline{R}, \overline{R} \right\rangle$ denotes the set of the reward functions of POMDPs in $\tilde{M}$, where $\underline{R}(s,a) \leq \overline{R}(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}$. The uncertainty set of reward function is defined as

$$u(R(s,a)) = \left\{ R\left(s,a,\right) \middle| \underline{R}(s,a) \leq R(s,a) \leq \overline{R}(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A} \right\}. \tag{4.16}$$

Note that the LC-POMDP is a more general framework than POMDP, since an LC-POMDP degenerates into a POMDP when $\underline{\mathbf{A}} = \overline{\mathbf{A}}$, $\underline{\mathbf{B}} = \overline{\mathbf{B}}$, and $\underline{R} = \overline{R}$. In addition, it is assumed that $\underline{R}(s,a)$ and $\overline{R}(s,a)$ are both finite values, so the one-step reward is bounded. Thus, for any infinite-horizon problem, we can instead solve a finite-horizon problem with sufficiently long horizon. The agent may choose one of the two criteria: $\delta = 0$ indicates the pessimistic criterion such that the policy is optimized against the worst case in the uncertainty set; and $\delta = 1$ indicates the optimistic criterion such that the policy is optimized against the best case. In this paper, we focus on the finite-horizon LC-POMDP problems.

4.4.1 Modified Value Iteration

The value backup method is modified for LC-POMDP to handle parameter uncertainty. Because the precise values of the transition probability is not available, we cannot update the belief of state in each step using (4.4). We adopt the policy tree representation for

LC-POMDP to avoid belief updating. Based on PT^*_{t+1}, a set of t-step policy trees PT^+_t is constructed, in which each tree is produced by assigning an action to the root node and choosing the $(t+1)$-step subtrees from PT^*_{t+1}. According to (4.12), we can generate a vector set $\Gamma^\star_t$ corresponding to PT^+_t for each POMDP $M \in \tilde{M}$. In general, the number of POMDPs in $\tilde{M}$ is uncountable, so there are uncountably many such $\Gamma^\star_t$s. We denote the union of such $\Gamma^\star_t$ by $\tilde{\Gamma}^+_t$, i.e.,

$$\tilde{\Gamma}^+_t = \bigcup_{\substack{a\in\mathcal{A}, \\ \left\{\alpha^{o^i}_{t+1}\right\}_{i=1}^{|\Omega|} \text{ is any combination of } \Gamma^\star_{t+1}}} \bigcup_{M\in\tilde{M}} \{\alpha = [\alpha(1),\ldots,\alpha(S)] | \alpha(s) \text{ satisfies (4.12) for } M\}. \tag{4.17}$$

For each policy tree $pt \in PT^+_t$, we have a set of vectors,

$$\tilde{\alpha} = \left\{\alpha \middle| \forall s \in \mathcal{S}, \alpha(s) \text{ satisfies (4.12) for } M; M \in \tilde{M}\right\}. \tag{4.18}$$

$\tilde{\alpha}$ contains uncountably many vectors, each of which represents the value of pt updated from $\tilde{\Gamma}^*_{t+1}$ with some POMDP in $\tilde{M}$ being the underlying model. The following theorem states that $\tilde{\alpha}$ has a lower bound and a upper bound. This results extends Theorem 1 of Ni and Liu [119] that such lower and upper bound exists when the constraints of probability sum is added.

Theorem 4.1. *Given a policy tree* $pt = \left\langle a, \alpha^{o^1}_t, ..., \alpha^{o^{|\Omega|}}_t \right\rangle \in PT^+_t$, *where* $a \in \mathcal{A}$ *and* $\alpha^{o^i}_{t+1} \in \hat{\Gamma}^*_{t+1}$, $\forall o^i \in \Omega$. *The set of vectors* $\tilde{\alpha}$ *is constructed by (4.18). There exist two POMDPs* $\underline{M} \in \tilde{M}$ *and* $\overline{M} \in \tilde{M}$ *such that for any vector* $\alpha \in \tilde{\alpha}$ *and each* $s \in \mathcal{S}$,

$$\underline{\alpha}(s) \le \alpha(s) \le \overline{\alpha}(s), \tag{4.19}$$

where $\underline{\alpha}$ *and* $\overline{\alpha}$ *are computed by (4.14) with* $\underline{M}$ *and* $\overline{M}$ *as the underlying model, respectively.*

Proof. The upper bound vector of a policy tree $pt = \left\langle a, \alpha^{o^1}_t, ..., \alpha^{o^{|\Omega|}}_t \right\rangle \in PT^+_t$ can be found

by solving the following linear program (LP):

$$\underset{\mathbf{A}(s,a,s')}{\text{maximize}} \quad \alpha_t(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathbf{A}(s,a,s') \sum_{o^i \in \Omega} \mathbf{B}(s',a,o^i) \alpha_{t+1}^{o^i}(s') \tag{4.20a}$$

$$\text{subject to} \quad \underline{\mathbf{A}}(s,a,s') \leq \mathbf{A}(s,a,s') \leq \overline{\mathbf{A}}(s,a,s'), \forall s' \in \mathcal{S} \tag{4.20b}$$

$$\sum_{s' \in \mathcal{S}} \mathbf{A}(s,a,s') = 1, \tag{4.20c}$$

$$\underline{\mathbf{B}}(s',a,o) \leq \mathbf{B}(s',a,o) \leq \overline{\mathbf{B}}(s',a,o), \forall s' \in \mathcal{S}, o \in \Omega \tag{4.20d}$$

$$\sum_{o \in \Omega} \mathbf{B}(s',a,o) = 1, \forall s' \in \mathcal{S} \tag{4.20e}$$

$$\underline{R}(s,a) \leq R(s,a) \leq \overline{R}(s,a)$$

The lower bound can be found by solving for the minimum of the objective function in (4.20a) with the same constraints (4.20b-c). Since $R(s,a), \mathbf{B}(s,a,o^i), \alpha_{t+1}^{o^i}(s)$ are bounded and the feasible set defined by (4.20b-c) are bounded, then the optimal solution of (4.20a) is bounded for both the minimization problem and the maximization problem. Let the optimal variable of (4.20a) be $\mathbf{A}^\star(s,a) \in \mathbb{R}^{|\mathcal{S}|}$, then the POMDP $\overline{M}$ is constructed by $\mathbf{A}^\star(s,a)$ for all $a \in \mathcal{A}, s \in \mathcal{S}$. The POMDP $\underline{M}$ is constructed by $\mathbf{A}^\star(s,a)$ for all $a \in \mathcal{A}, s \in \mathcal{S}$ from the minimization problem. □

Such a set of vectors $\tilde{\alpha}$ is called a bounded vector set (BVS) [119]. $\underline{\alpha}$ ($\overline{\alpha}$) is called the lower (upper) bound vector of $\tilde{\alpha}$ and *pt*. Note that $\Gamma_t^\star$ can be viewed as the union of the BVSs corresponding to the policy trees in PT_t^+. The LP (4.20a) has a direct solution if the following condition holds.

Theorem 4.2. *If $\alpha_{t+1}^{o^i}(s) \geq 0$ for all $s \in \mathcal{S}$ and all $o^i \in \Omega$, then we can solve LP (4.20a) directly. Let $\theta(s,a,s') = \gamma \sum_{o^i \in \Omega} \mathbf{B}(s',a,o^i) \alpha_{t+1}^{o^i}(s')$ be the coefficient of $\mathbf{A}(s,a,s')$ in (4.20a). The solution to the linear program in (4.20a) is $\mathbf{A}^\star(s,a,s^{(i)}) = \overline{\mathbf{A}}(s,a,s^{(i)})$ for $i = 1, \ldots, \lfloor |\mathcal{S}|/2 \rfloor$, and $\mathbf{A}^\star(s,a,s^{(i)}) = \underline{\mathbf{A}}(s,a,s^{(i)})$ for $i = \lceil |\mathcal{S}|/2 \rceil + 1, \ldots, |\mathcal{S}|$, where (i) indicates the i-th largest element of $\{\mathbf{A}(s,a,s^i)\}_{i=1}^{|\mathcal{S}|}$.*

Theorem 4.2 is a special case of the following lemma.

Lemma 4.1. *Consider the linear program*

$$\underset{x_1,\ldots,x_N}{\textit{maximize}} \quad \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_N x_N \tag{4.21a}$$

$$\textit{subject to} \quad \underline{x}_i \leq x_i \leq \overline{x}_i, \forall i \in \{1, 2, \ldots, N\} \tag{4.21b}$$

$$\sum_{i=1}^{N} x_i = 1 \tag{4.21c}$$

The coefficient $\{\theta_i\}_{i=1}^N$ *of the objective function satisfies* $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_N > 0$. *Then: (1) The optimal solution* $\{x_i^\star\}_{i=1}^N$ *will keep the same when the order of* $\{\theta_i\}_{i=1}^N$ *does not change; (2) The optimal solution is* $x_i = \overline{x}_i$ *for* $i = 1, \ldots, \lfloor N/2 \rfloor$, *and* $x_i = \underline{x}_i$ *for* $i = \lceil N/2 \rceil + 1, \ldots, N$.

The proof of Lemma 4.1 is provided in Appendix C.1.1. Lemma 4.1 give the optimal solution and the condition of preserving the optimality for a set of special LP such that the decision variables form a vector in a probability simplex and have predefined upper and lower bounds. Note that the constraints (4.20b-c) are identical to (4.21b-c). If the coefficients of the objective function are not descending in the order of ascending index, we can reindex the decision variables such that the coefficients are descending, then the conclusion holds for the new index. If the objective is to minimize and the coefficients are all non-negative, then we can reindex the decision variables such that the coefficients are ascending, then the optimal solutions will be the lower bound for the first half of the decision variables, and upper bound for the second half of the decision variables.

In order to estimate the value of a policy tree pt, the band should be reduced to a line, i.e., the corresponding BVS $\tilde{\alpha}$ should be replaced by a single vector, which is corresponding to a POMDP in $\tilde{M}$. If the agent holds the pessimistic criterion, then the vector is the lower bound. Otherwise, the upper bound is selected under the optimistic criterion. By replacing BVSs with vectors, the infinite set $\tilde{\Gamma}_t^\star$ is replaced by a finite set $\hat{\Gamma}_t^\star$, whose volume is the same as that of PT_t^+. Note that Ni and Liu [119] pruned the BVS with the requirement that none of the upper bound is dominated by any lower bound, which is not consistent with the optimistic and pessimistic criteria in this paper.

The advantage of using policy tree representation is that belief is not inccured in the planning, but the incremental pruning in o cannot be applied, since the LP includes all $o \in \Omega$. However, we can still apply incremental pruning in a.

4.4.2 Connection to classic POMDP

In the previous section, we have seen that when the upper and lower bound of the LC-POMDP parameters coincide, then the LC-POMDP degenerates to a classic POMDP. In this section, we will discover the condition under which one specific POMDP $M_0 \in \tilde{M}$ achieve optimality across value iterations.

Theorem 4.3. *Consider the linear program in (4.20a). Let*

$$\theta(s, a, s') = \sum_{j=1,\dots,|\Omega|} \mathbf{B}(s', a, o^j)\alpha_{t+1}^{\ell(j)}(s'), \tag{4.22}$$

where $\ell(j) \in \{1, 2, \dots, |\Gamma|\}$ *is a permutation of* $\{1, 2, \dots, |\Gamma|\}$*. If we have*

$$y(i, a) = \sum_{j=1}^{p} \mathbf{B}(s^i, a, o^j) \tag{4.23}$$

is nonincreasing in i *for all* $a \in \mathcal{A}, p \in \{1, 2, \dots, |\Omega|\}$*, and* $g(s)$ *and* $\alpha_{t+1}^{\ell}(s)$ *is nonincreasing in* s*,* $R(s^i, a)$ *nonincreasing in* i *for all* $a \in \mathcal{A}$*, and* $\underline{\mathbf{A}}(s^i, a, s'), \overline{\mathbf{A}}(s^i, a, s')$ *nonincreasing in* i *for all* $a \in A$*,* $s' \in \mathcal{S}$*, then* $\theta(s, a, s^i)$ *is nonincreasing in* i*. If the condition holds for every iteration, then the order of* $\theta(s, a, s^i)$ *is the same in* i *across iterations. Therefore, the optimal model is the same across iterations.*

If such conditions are satisfied, the LC-POMDP degenerates to a classic POMDP parameterized in M_0. The proof of Theorem 4.3 is provided in Appendix C.1.2.

4.4.3 Computation complexity

In one step of backup, the complexities of the elementary operations are: (1) computing a lower/upper bound vector requires time of $O(|\mathcal{S}|(|\Omega| \log |\Omega| + |\mathcal{S}|))$; (2) pruning a vector

set $\Gamma^{\star}_{t+1}$ to a set $\Gamma^{\star}_{t}$ requires $|\Gamma^{\star}_{t+1}|$ LPs with $O(|\Gamma^{\star}_{t+1}||\Gamma^{\star}_{t}|)$ constraints in the worst case and $O(|\Gamma^{\star}_{t}|^2)$ constraints in the best case [73].

One of the drawbacks of the policy tree representation is that the incremental pruning cannot be applied. Incremental pruning requires the vectors to be generated separately for each observation (4.7). However, in LC-POMDP, the vectors are generated by solving the LP (4.20a) which involves a set of $|\Omega|$ vectors each coresponds to a unique observations. Therefore, we prefer to solve the LC-POMDP when it can be degenerated to a classic POMDP to take advantage of the incremental pruning procedure for the reduced computation complexity in pruning. The modified Value Iteration algorithm for LC-POMDP is summarized in Algorithm 4.1.

Data: LC-POMDP $\tilde{M} = \langle \mathcal{S}, \mathcal{A}, \Omega, \tilde{\mathbf{A}}, \mathbf{B}, \tilde{R}, g, \gamma \rangle$, decision horizon T, criterion δ
Result: T-level policy tree PT

```
Γ*_T ← g
for t = T − 1, . . . , 0 do
    if LC-POMDP satisfies Theorem 4.3 then
        Construct Γ*_t from Γ*_{t+1} using Monahan enumeration (4.7) and Incremental
         Pruning (4.9);
    else
        Construct Γ+_t from Γ*_{t+1} using enumeration and LP (4.20a) with criterion δ;
        Prune Γ+_t to Γ*_t;
    end
    Construct the sub-policy-tree for each node in Γ*_t;
end
```

Algorithm 4.1: Modified Value Iteration for LC-POMDP.

4.5 Chance-Constrained POMDP (CC-POMDP)

The linear constraints for the transition probabilities can be viewed as hard constraints, that the transition probabilities are within the uncertainty set, and will not fall out of it. However, in some applications, it is not possible to define the hard constraints. Instead, it is more convenient to define a probability distribution over the parameter space, to indicate which region has a higher probability. Since each row of the transition matrix in a POMDP is defined on a probability simplex, it is natural to assume that the stochastic transition probability is Dirichlet distributed. This section develops an algorithm to solve the POMDP with Dirichlet distributed transition probability. In this section, we extend the percentile optimization for MDP with the uncertain transition probability in Delage and Mannor [100] to POMDP.

For each state-action pair (s, a), we will use independent Dirichlet priors to model the uncertainty in the transition probability $\mathbf{A}(s, a, s')$. This assumption is very convenient for describing prior knowledge about transition parameters due to the fact that, after gathering new transition observations, one can easily evaluate a posterior distribution over these parameters. More specifically, for a vector of transition parameters $\mathbf{A}(s, a) = [\mathbf{A}(s, a, s^1), \ldots, \mathbf{A}(s, a, s^{|\mathcal{S}|})]$, the Dirichlet distribution over $\mathbf{A}(s, a)$ follows the density function

$$f(\mathbf{A}(s, a)) = \frac{1}{Z(\phi)} \prod_{j=1}^{|\mathcal{S}|} \mathbf{A}(s, a, s^j)^{\phi_j - 1}, \tag{4.24}$$

where $\phi \in \mathbb{R}^{|\mathcal{S}|}$ is modeling parameter for the Dirichlet prior and $Z(\phi)$ is a normalization factor. Suppose there is a set of transition of latent states $\{s^1, s^2, \ldots, s^{|\mathcal{S}|}\}$ from the multinomial distribution $\Pr(s^j|\mathbf{A}(s, a)) = \mathbf{A}(s, a, s^j)$, the posterior distribution over $\mathbf{A}(s, a)$ can be analytically derived, which takes the same Dirichlet form $\mathbf{A}(s, a) \sim \mathrm{Dir}(\phi)$ with probability density

$$f\left(\mathbf{A}(s, a)\middle|s^1, s^2, \ldots, s^{|\mathcal{S}|}\right) = \frac{1}{Z(\phi, N_1, \ldots, N_{|\mathcal{S}|})} \prod_{j=1}^{|\mathcal{S}|} \mathbf{A}(s, a, s^j)^{\phi_j + N_j - 1}, \tag{4.25}$$

where N_j is the number of times of a transition to s^j. Since the true state is not observable in POMDP, we could use the forward-backward algorithm to estimate the latent state from the observations.

Note the transition probability vector can be any point in the $|\mathcal{S}|$-dimensional probability simplex. The set of point with high density indicates a region that the model is more likely to realize. Therefore, we can specify a percentile criterion that the model will be realized within a region with a confidence level of $1-\epsilon$, where $\epsilon \in (0,1)$. This criterion is a more flexible criterion that balances the pessimistic criterion and the optimistic criterion. Particularly, $\epsilon = 0$ is equivalent to a pure pessimistic criterion, and $\epsilon = 1$ is equivalent to a pure optimistic criterion. It gives the agent the ability to choose how much the agent believe the best criterion is close to the pure pessimistic criterion parameterized by ϵ.

In each step of value iteration in solving POMDP, the backup procedure will be replaced with a chance-constrained problem (CCP) as follows. Suppose the minimal set of vectors $\Gamma_t^\star$ is constructed after t steps, for each state-action pair (s,a) we try to solve

$$\underset{\mathbf{A}(s,a)}{\text{maximize}} \quad y(s,a) \tag{4.26a}$$

$$\text{subject to} \quad \Pr_{\mathbf{A}(s,a)\sim \text{Dir}(\phi)}(\alpha(s,a) \geq y(s,a)) \geq 1-\epsilon \tag{4.26b}$$

where the α-vector $\alpha(s,a)$ is the same as (4.14), and $y(s,a)$ is a dummy decision variable. The solution $y^\star(s,a)$ will be used to represent $\alpha(s,a)$, which indicates the threshold of the α vector such that with $1-\epsilon$ chance the true value is higher than the threshold. The α-vector is constructed by solving CCP (4.26a) for every $s \in \mathcal{S}$. A set of vectors Γ_t^+ is constructed by solving the CCP (4.26a) for all combinations of $|\Omega|$ vectors from $\Gamma_{t+1}^\star$, which will be pruned to obtain the minimal set $\Gamma_t^\star = \texttt{prune}(\Gamma_t^+)$.

Next, we will briefly discuss the solution to the CCP (4.26a). Suppose $\alpha(s,a)$ follows a distribution with cumulative distribution function (CDF) $F(x)$, then $y^\star = F^{-1}(\epsilon)$. Therefore, the problem becomes finding the distribution of $\alpha(s,a)$ which is a linear combination

of the components of Dirichlet distribution. We can write (4.14) as

$$\alpha(s,a) = k + \sum_{j=1}^{|\mathcal{S}|} \theta_j x_j, \tag{4.27}$$

where $k = R(s,a)$ is a constant, $\theta_j = \gamma \sum_{o^i \in \Omega} \mathbf{B}\left(s^j, a, o^i\right) \alpha_{t+1}^{o^i}(s^j)$ is the coefficient, $x_i := \mathbf{A}(s,a,s^j)$ is the decision variable. Note that $x_j \sim \mathrm{Dir}(\phi)$, thus the range of $\alpha(s,a)$ is $[\theta_{\min}, \theta_{\max}]$, where $\theta_{\min}$ and $\theta_{\max}$ is respectively the minimum and maximum of the set of coefficients $\{\theta_j\}_{j=1}^{|\mathcal{S}|}$ in 4.27. For $|\mathcal{S}| = 2$, it is easy to show that [123] $\alpha(s,a)$ is a shifted Beta distribution [123]:

$$\frac{\alpha(s,a) - \theta_{\min}}{\theta_{\max} - \theta_{\min}} \sim \mathrm{Beta}(\phi_1, \phi_2). \tag{4.28}$$

For $|\mathcal{S}| \geq 3$, Provost and Cheong [123] used the results from Imhof [124] to express the CDF of $\alpha(s,a)$ as follows:

$$\Pr(\alpha(s,a) < x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin\left[\sum_{j=1}^{|\mathcal{S}|} \phi_j \tan^{-1} \theta_j - xu\right]}{u \prod_{j=1}^{|\mathcal{S}|} \{1 + (\theta_j - x^2 u^2)\}^{\phi_j/2}} \, \mathrm{d}u \tag{4.29}$$

for $\theta_1 < x < \theta_{|\mathcal{S}|}$. Then finding $y^\star$ is the problem of finding the root of the equation

$$\Pr(\alpha(s,a) < y) = \epsilon. \tag{4.30}$$

The analytical solution to (4.30) is complex. In practice, we find the numerical solution using numerical methods like the bisection methods, since the CDF is monotonically increasing in y.

The modified Value Iteration algorithm for CC-POMDP is summarized in Algorithm 4.2. Performing the robust policy for CC-POMDP is identical to that of LC-POMDP with the policy tree generated from the value iteration.

4.6 Computational experiment

In this section, we describe a set of computational experiments for comparing the performance of robust policies and classic policies for a POMDP with uncertain parameters. The

Data: CC-POMDP $\tilde{M} = \langle \mathcal{S}, \mathcal{A}, \Omega, \tilde{\mathbf{A}}, \mathbf{B}, \tilde{R}, g, \gamma \rangle$, decision horizon T, criterion ϵ

Result: T-level policy tree PT

$\Gamma_T^\star \leftarrow g$

for $t = T-1, \ldots, 0$ **do**

  Construct $\Gamma_t^\star$ from $\Gamma_{t+1}^\star$ using enumeration (4.14) and CCP (4.26a);

  Prune Γ_t^+ to $\Gamma_t^\star$;

  Construct the sub-policy-tree for each node in $\Gamma_t^\star$;

end

Algorithm 4.2: Modified Value Iteration for CC-POMDP.

experiments are based on a series of random instances of POMDPs. To generate the random test instances, first the number of states, actions, models, and decision epochs for the problem were defined. Then, model parameters were randomly sampled. The transition probabilities are generated by sampling from a uniform distribution so that $\tilde{\mathbf{A}}(s, a, s') \sim U(0, 1)$. Then, for every $(s, a, s) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, the transition probabilities were normalized so that the row of the transition probability matrix had elements that sum to one:

$$\mathbf{A}(s, a, s') = \frac{\tilde{\mathbf{A}}(s, a, s')}{\sum_{s'' \in \mathcal{S}} \tilde{\mathbf{A}}(s, a, s'')}. \tag{4.31}$$

The emission probabilities are generated similarly. The uncertainty set of the transition and emission probabilities are generated by relaxing the nominal values by 0.1, i.e., the upper bound is the nominal value plus 0.1, and the lower bound is the nominal value minus 0.1. The reward function are generated by sampling from a uniform distribution $\tilde{R}(s, a) \sim U(0, 4)$, and its uncertainty set are obtained by relaxing the nominal reward function by 0.5. For the CC-POMDP, the posterior Dirichlet distribution has the mode same as the true model for each row of the transition probability. Suppose $\mathbf{A}(s, a, s') \sim \mathrm{Dir}(\phi)$, then $\mathbf{A}_{\mathrm{true}}(s, a, s') = \frac{\phi(s')}{\sum_{s'' \in \mathcal{S}} \phi(s'')}$ for all $s' \in \mathcal{S}$.

We solve a set of 20 random instances of robust POMDPs with 2 states, 2 actions, for 5 decision epochs. For each instance of the robust POMDP, we generate 100 independent

classic POMDP within the uncertainty set of the transition probability and the reward function. Since the parameters of these models deviate from the true model, we name them the biased POMDP. The reward of the robust policy from the robust POMDPs and the policies from the biased and unbiased classic POMDPs are compared on 100 replications of running the policies on the true POMDP models. For each instance of the robust POMDP, we compare the total reward of the unbiased POMDP policy, the LC-POMDP with lower and upper bound, and the CC-POMDP with $\epsilon = 0.5, 0.1, 0.01$ to the set of biased POMDP policies. The relative reward of a policy to a set of the biased POMDP policies are defined as

$$\tilde{r} = \frac{r - r_{\min}}{r_{\max} - r_{\min}}, \tag{4.32}$$

where r is the reward of the target policy, and $r_{\min}, r_{\max}$ are the minimum and maximum of the set of the biased POMDP policies respectively. The result is illustrated using box plots in Figure 4.2. We can see that all robust policies beat 70% to 80% of the biased POMDP policies. LC-POMDP policies updated with lower bound (for pessimistic criterion) are better than with upper bound (for optimistic criterion). CC-POMDP policies with small ϵ (less optimistic) performs better than with large ϵ (more optimistic).

4.7 Case study: personalized treatment for chronic depression

We choose the personalized disease treatment problem as an application for our methods. The personalized treatment of disease with latent health state can be modeled with a POMDP if the disease progression model is a hidden Markov model. Since the parameters of the POMDP in disease treatment is commonly learned from past treatment history, the estimation of these parameters has unavoidable uncertainty. In a heterogeneous population, we assume that the disease progression model of each patient is unique, so the treatment policy should be tailored to each individual. In chapter 3 we proposed the partially observable collaborative model (POCM) for modeling the individual disease progression. The POCM uses a set of basis models to represent patterns as subtypes of disease progression, and an individual patient's progression dynamic is a combination of multiple basis models.

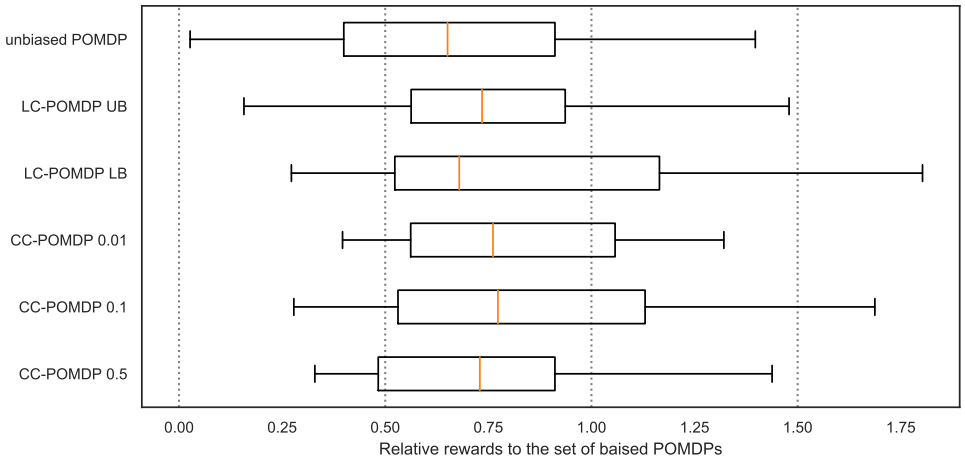


Figure 4.2: Comparing robust policies and the biased policies for 20 randomly generated instances of robust POMDPs. Each box indicates the distribution of the relative rewards to the set of 100 biased POMDPs of the six policies. The middle bar shows the median and the left and right bound of the box indicates the first and third quantile of the distribution.

We assume that there are K basis models discovered in the population, and the membership of the individual patient i is defined as $c_{ik} \in [0, 1]$, representing the degree to which the model of the individual patient i resembles the basis model k [125].

The individual treatment policies are developed in three steps: (1) The learning step, where the basis models for each progression subtype are learned from an existing dataset of treatment records containing a large amount of patients. (2) The fine-tuning step, where the personal dynamic for a new patient is initialized using model parameters estimated from the learning step, and updated by running separate short trial periods due to the cost and potential side effect of treatment; (3) The decision step, where the optimal treatment strategy is obtained by solving a rolling horizon POMDP. We assume the basis model is accurate since it is learned from samples of a large size, while the membership learning in the fine-tuning stage is inaccurate due to insufficient treatment records.

The objective of the numerical example is to compare the performance of the treatment

policy derived from the true disease model, the estimated model from POCM and the robust POMDP including LC-POMDP and CC-POMDP. We use chronic depression as an example. There are 3 states of depression: healthy (H), mild depression (M), and severe depression (S), in ascending severity of depression, i.e., the state space is $\mathcal{S} = \{\text{H,M,S}\}$. At each time period, patients under ongoing depression treatment will take the Patient Health Questionnaire (PHQ-9), a self-administered questionnaire, to evaluate their depression status, which is the observation of the POMDP [77]. The result of PHQ-9 is a score with a range from 0 to 27, with a higher score indicating more severe mental condition. The PHQ-9 score can be categorized into three levels, where scores $0 \sim 4$ (P1) stand for healthy to mild depression severity, scores $5 \sim 9$ (P2) stand for mild to moderate, and scores $10 \sim 27$ stand for major depression, i.e., the observation space is $\Omega = \{\text{P1,P2,P3}\}$. There are 2 types of treatments: Treatment I is the usual care using antidepressant medication, and Treatment II is an intensive outpatient program with additional behavioral counseling. The cost for taking each treatment per period is: \$1,000 for Treatment I and \$2,000 for Treatment II. The reward for one period is defined as

$$R(s_t, a_t) = \lambda\kappa(s_t) - c(a_t) \tag{4.33}$$

where $\kappa(s_t)$ is the utility of being in state s_t measured in quality-adjusted life-years (QALYs); $c(a_t)$ is the cost of treatment a_t; λ is the willingness to pay (WTP) which assigns a monetary value to a QALY; \$50,000 per QALY is commonly used for WTP in the literature. The reward function trades off the health utility based on the true states, WTP and the cost of the treatment, and is measured in Net Monetary Benefit (NMB). The utility of each state is $\kappa(\text{H}) = 1, \kappa(\text{M}) \in (0,1), \kappa(\text{S}) \in (0,1)$, and $\kappa(\text{M}) > \kappa(\text{S})$, assuming that more severe depression condition will lead to lower health utility for patient.

We use a simulation approach (similar to the approach in 3.4.2) to construct a set of patients with individual disease progression dynamics, which is characterized by the unique initial state belief (π_i^S), the state-transition function ($\mathbf{A}_i^S$) and emission probability ($\mathbf{B}_i^S$) associated with the patient i ($i = 1, \ldots, N$). Such model is the true model of each patient. In

this experiment, we assume that there are 3 basis hidden Markov models, each corresponds to a high risk of depression progression, low risk of depression progression, and stable condition. The values of the parameters of the basis models are listed in Appendix B.5.1. To construct the true model for an individual patient, we generate the membership vector $\mathbf{c}_i$ for each patient i in a population with subgroup structure following the approach in Lin et al. [24]. The membership of the testing patients can be learned by performing the fine-tuning step by updating membership with fixed basis models. In the fine-tuning stage, each patient is given a short course of trial treatments, typically 20 periods, which is generated using the true model with mixed treatment options, e.g., 10 periods of Treatment I and 10 periods of Treatment II.

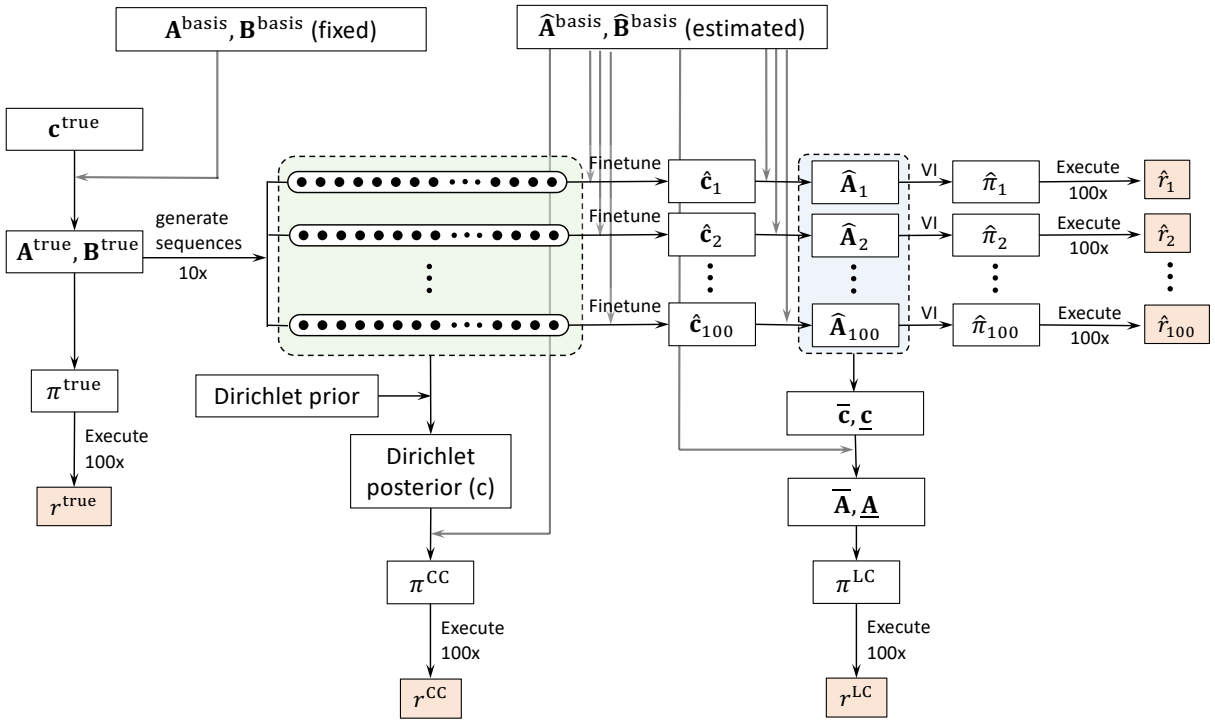


Figure 4.3: The process of experiment on personalized treatment of chronic depression.

To test the outcome of the robust POMDPs, we run policies on 4 types of patients: 3 patients each behaves very closely to each of the 3 basis models (e.g., a patient similar to

the first basis model has the true membership close to $[0.8, 0.1, 0.1]$), and a patient with a mixture of the 3 basis models (i.e., the membership are distributed evenly across the 3 basis models). In the fine-tuning stage, we simulate 100 independent observation data, and estimate the membership using the POCM. See Figure 4.3. For each estimated membership $\hat{\mathbf{c}}_i$, $i = 1, \ldots, 100$, we can construct a biased POMDP $\hat{\mathbf{A}}_i$ (since the values deviate from the true POMDP) and find the treatment policy $\hat{\pi}_i$ using the value iteration. We also find robust policies by modeling the treatment with the two types of robust POMDP we develop. For the LC-POMDP, we use the upper and lower bound of the memberships in the fine-tuning stage to construct the upper and lower bound of the transition probability. For the CC-POMDP, we update the posterior of the transition probability using the estimated transitions between states using the forward-backward method, see Appendix B.2.3. We also obtain the classic POMDP policy with the unbiased model. The four types of policies (the unbiased POMDP policy, the biased POMDP policies, the LC-POMDP policies, and the CC-POMDP policies) are evaluated by applying to the simulated patient for 24 periods (which equals 2 years in total). The disease progression follows the true model parameters and the initial health state is set as healthy. The evaluation is replicated independently for 100 times for each policy, and the discounted total reward is collected to indicate the performance of the policies. The result is shown in Figure 4.4, where each panel corresponds to one type of patient. The dotted curve shows the empirical cumulative distribution of the rewards of the set of the 100 biased POMDP policies $\{\hat{r}_i\}_{i=1}^{100}$, and a number of markers indicate the rewards of the other policies and the percentile of them in the set of the biased POMDP policies.

We can see that for all types of patients, the robust policy derived from LC-POMDP and CC-POMDP with $\epsilon = 0.1$ has similar performance to the policy from POMDP with ground truth. They are around 70 to 80 percentile of all the policies from estimated models, which gains about \$ 500 dollars of annual reward compared to the median of the biased POMDPs. For the CC-POMDP, using hyperparameter $\epsilon = 0.5$ will result in much worse performance to $\epsilon = 0.1$ or 0.01, indicating that optimistic criteria in CC-POMDP is not as good as pessimistic criteria. On the other hand, a very small ϵ such as 0.01 leads a slightly worse

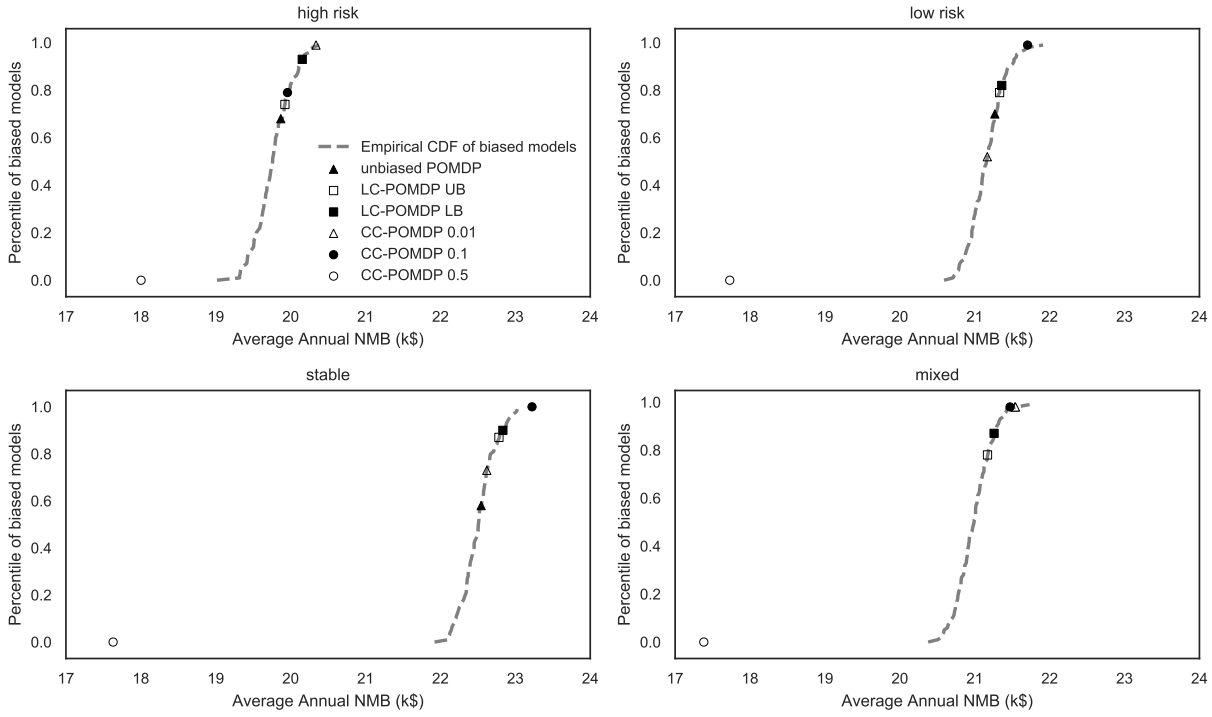


Figure 4.4: Comparing robust policies and the biased policies for 20 randomly generated instances of robust POMDPs. Each panel corresponds to a type of patient in terms of the response to depression. The unit of the reward is one thousand US dollars.

performance, which means an extremely conservative policy will not have extra benefits.

We discovered a set of conditions that the LC-POMDP can degenerate to a POMDP in Theorem 4.3. Such conditions in the treatment of chronic diseases can be intepereted as (1) The terminating reward of a better health state is higher than a worse health state; (2) The upper and lower bound of the transition probability to better health state is higher than a worse health state; (3) The emission probability of observing a high PHQ-9 score given better health state is greater than that given worse health state, and the emission probability of observing a low PHQ-9 score given better health state is smaller than that given a worse health state.

We performed one-way sensitivity analyses on model parameters, in which each parameter

Table 4.1: Sensitivity Analysis of the robust POMDP policies. The numbers indicate the percentile of the rewards of 6 policies in the set of rewards from 100 independently generated biased POMDPs.

patient type	unbiased POMDP	LC-POMCP upper bound	LC-POMCP lower bound	CC-POMCP $\epsilon = 0.01$	CC-POMCP $\epsilon = 0.1$	CC-POMCP $\epsilon = 0.5$
(1) Higher utility in M/S: u(H,M,S) = [1, 0.8, 0.6]						
high risk	1.00	0.99	0.95	0.67	0.99	0.00
low risk	1.00	0.99	0.93	0.90	0.85	0.00
stable	1.00	0.99	0.99	0.99	0.77	0.00
mixed	1.00	0.89	0.95	0.81	0.88	0.00
(2) Lower Treatment II cost: c(II) = \$1,500						
high risk	0.63	0.81	0.96	0.99	0.68	0.00
low risk	0.59	0.83	0.87	0.99	0.90	0.00
stable	0.52	0.90	0.74	0.99	0.89	0.00
mixed	0.68	0.84	0.91	0.93	0.94	0.00
(3) Higher Treatment II cost: c(II) = \$5,000						
high risk	0.56	0.77	0.99	0.87	0.98	0.00
low risk	0.71	0.66	0.59	0.99	0.72	0.00
stable	0.50	0.69	0.87	0.87	0.80	0.00
mixed	0.83	0.90	0.85	0.87	0.88	0.00
(4) Smaller WTP: λ = \$10,000/QALY						
high risk	1.00	0.43	0.88	0.96	1.00	0.00
low risk	0.96	0.46	0.72	1.00	0.93	0.00
stable	0.96	0.28	0.78	0.95	0.80	0.00
mixed	0.98	0.43	0.78	0.70	0.78	0.00
(5) Larger WTP: λ = \$150,000/QALY						
high risk	0.40	0.67	0.86	0.93	0.86	0.00
low risk	0.40	0.67	0.71	0.99	0.79	0.00
stable	0.29	0.75	0.49	0.93	0.69	0.00
mixed	0.46	0.68	0.79	0.63	0.70	0.00

(e.g., the cost of treatment, the utility and the willingness-to-pay) is varied individually to examine the policies' performance in each setting. We compared the rewards of the unbiased POMDP and 5 robust POMDPs to the rewards of a set of 100 independent biased POMDPs. The percentile of each policy to the set of biased POMDP policies for each scenario and each type of patient are listed in Table 4.1. The advantage of robust POMDP policies to biased POMDP policies are shown in the majority of model settings.

4.8 Conclusion

In this chapter, we propose two types of robust POMDP models to reduce the risk of inaccurate parameter estimation in performing the policy derived from a POMDP: the LC-POMDP assuming parameters bounded with linear constraints, and the CC-POMDP assuming that the transition probability follows a Dirichlet distribution. The main advantage of the robust POMDPs is mitigating the loss of policy performance under the risk of inaccurate parameter estimation. The choice of the form of robust POMDP depends on the prior knowledge of the application: LC-POMDP is prefered when parameters lie in known uncertainty sets, while CC-POMDP is prefered when there lacks knowledge of parameter bounds or the decision maker treats the parameters as random variables. We developed the value iteration algorithms for both models to find robust optimal policies in the form of policy tree. We discussed the special cases when LC-POMDP would degenerates to classic POMDP. The computation complexity of the value iteration algorithm is analyzed.

We evaluated the performance of our solution methods using a large set of randomly-generated test instances and also the personalized treatment of chronic depression as a case study. The robust policies generated from the value iteration performed very well across the randomly-generated test cases. For the personalized treatment problem, we used the simulated treatment responses of an individual patient to estimate the parameter bounds of the individual disease model for LC-POMDP modeling. The posterior of the individual transition probability is also updated form the treatment record. We showed improved rewards in terms of NMBs compared to classic POMDP with biased parameters.

There are several future directions of robust POMDPs. So far we only considered a set of simple linear constraints for the LC-POMDP, i.e., the upper and lower bounds plus the properties of probability vectors. However, the LC-POMDP proposed in this paper is capable of modeling the parameters with any linear constraint. Further work can explore the solution of the LC-POMDP when the uncertainty set takes any form of linear constraints. An example is the convex hull of all possible values or recorded values from history. The uncertainty set can even be relaxed to general convex sets. A second direction is to incorporate the uncertainty in emission probability, while currently only the uncertainty in the transition probability is included in both types of robust POMDPs. The third direction is to consider VI for robust POMDPs with infinite horizons, and additionally the solution via policy iteration.

In summary, we approach two ways of incorporating parameter uncertainty in POMDPs. These approaches allow the decision maker to specify the forms of parameter uncertainties with either a set of hard constraints or a set of soft constraints like a probability distribution and a confidence level. The robust POMDP can be valuable in many applications including inventory control, scheduling, finance, and healthcare, when explicit knowledge of the POMDP parameters are not available.

Chapter 5

CONCLUSION

In this thesis, we tackle the problem of personalized treatment of chronic diseases in three steps: (1) Detect the subgroup structure of disease progression in a heterogeneous population; (2) Learn the individual disease progression model from the records of a short course of treatment utilizing a pre-learned model from a large population; (3) Develope two robust POMDP models to find the optimal policy when the individual model learned in step 2 is subject to errors. Due to the fact that the availability of complete medical treatment data is limited, designing efficient medical decision tools is still a challenging problem. In addition, the high cost, risk and long duration of performing experiments/pragmatic trials using the treatment policies derived from mathematical models makes validation on real patients difficult in general.

This work contributes to the medical decision-making research community with the introduction of POCM for modeling the individual disease progression model, and two types of POMDPs to find robust policies for models with parameter uncertainty. We also provide several case studies on chronic depression with both real longitudinal data and simulated patients.

We hope this work inspires future researches to develop efficient medical decision-making tools for personalized treatment. There are several potential directions for improving the decision tools developed in this thesis. Firstly, we only considered diseases with a small number of discrete states. The computation complexity of the POCM learning algorithm and Value Iteration algorithm in LC-POMDPs and CC-POMDPs for diseases with a large number of states (i.e., more than 10) is unacceptable for real-time decision making. Therefore, approximate approaches to these algorithms are needed. In addition, the POCM needs modification

to apply to diseases with continuous states. Secondly, the POCM can be extended to include larger numbers of covariates from patients' EHR in addition to their treatment records as features to estimate the individual disease progression model. Thirdly, the robust POMDP can be extended by using different optimization criteria, such as the distributionally robust optimization (DRO) which assumes an ambiguity parameter that controls the size of the deviation from the nominal model in the worst case [126].

Appendix A

APPENDIX FOR CHAPTER 2

A.1 Details of Methods

A.1.1 Methods of Chi-Square Test on Homogeneity

Suppose the data were sampled from $r = 4$ subgroups and the categorical feature has c levels. The subgroups are

(1) patients with pattern (a) only

(2) patients with pattern (b) only

(3) patients with both patterns

(4) patients with no patterns

Recall from the main paper that the two patterns are (a) PHQ-8 increases and Item 9 decreases, and (b) PHQ-8 decreases and Item 9 increases (using unfitted observations in the EHR dataset). The categorical features are

(1) Age, with 4 levels

(2) Sex, with 2 levels

(3) Mean of Charlson Index, with 4 levels

At any specified level of the categorical variable, the null hypothesis states that each subgroup has the same proportion of observations. Therefore,

H_0: $P_{\text{level 1 of subgroup 1}} = P_{\text{level 1 of subgroup 2}} = \ldots = P_{\text{level 1 of subgroup 4}}$

H_0: $P_{\text{level 2 of subgroup 1}} = P_{\text{level 2 of subgroup 2}} = \ldots = P_{\text{level 2 of subgroup 4}}$

$\ldots$

H_0: $P_{\text{level c of subgroup 1}} = P_{\text{level c of subgroup 2}} = \ldots = P_{\text{level c of subgroup 4}}$

The degree of freedom is DF $= (r-1) \times (c-1)$. The expected frequency counts are computed separately for each subgroup at each level of the categorical variable, $E_{r,c} = n_r \times n_c/n$, where $E_{r,c}$ is the expected frequency count for subgroup r at level c of the categorical variable, n_r is the total number of observations from subgroup r, n_c is the total number of observations at level c, and n is the total sample size. The test statistic is a chi-square random variable defined by

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{r,c} - E_{r,c})^2}{E_{r,c}}, \tag{A.1}$$

where $O_{r,c}$ is the observed frequency count in subgroup r for level c of the categorical variable, and $E_{r,c}$ is the expected frequency count in subgroup r for level c of the categorical variable. Finally, the p-value is defined as $p = P\left(x > \chi^2_{\text{DF}}\right)$ [127].

A.1.2 Gaussian Process Regression

Gaussian process regression (GPR) is used to transform the irregular and sparse longitudinal data into a continuous function of time. We follow the method and notation as defined in Lasko et al. [29]. The GPR model assumes that there is an unobserved source function $f(t)$ that represents the true depression measurement trajectory (such as PHQ-8 and Item 9 scores) over time. The observed depression measurement sequence y is considered as a set of samples taken from the source function, with observation noise. The transformation is a parametric Bayesian inference of the source function from the sampled data. We define the probability of a given continuous function $f(t)$ in terms of an infinite-dimensional Gaussian process $\mathcal{GP}$ as

$$\Pr(f(t)) = \mathcal{GP}(m(t), C(t_1, t_2)). \tag{A.2}$$

The mean function $m(t)$ is a function of time, and the covariance function $C(t_1, t_2)$ is a function of the pair of times t_1 and t_2, which defines the dependence between two function values $f(t_1)$ and $f(t_2)$. The Gaussian process defined by C represents a prior probability density over all possible source functions for the given trajectory. GPR produces a second Gaussian process that represents a posterior probability density given the prior and the observations in the trajectory.

Although the Gaussian process represents the probability density of the continuous function $f(t)$, the density $P(f(t_i))$ can be calculated at a finite set of times t_i, and they have the same values of the entire continuous trajectory $P(f(t))$ with samples at the time points t_i. Given a vector of observations $\mathbf{S}^0 \in \mathbb{R}^n$ made at times $\mathbf{t}^0 \in \mathbb{R}^n$, we can compute the posterior probability $\Pr\left(f(t) = y|\mathbf{S}^0, \mathbf{t}^0\right)$ that the true source function f passes through the point (t, y), which also represents the probability that a new measurement made at time t would produce the value y. GPR assumes that at any time t, the posterior density is Gaussian,

$$\Pr\left(f(t) = y\middle|\mathbf{S}^0, \mathbf{t}^0\right) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{(y-\hat{y})^2}{2\hat{\sigma}^2}\right] \tag{A.3}$$

where $y = \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{S}^0$ is the posterior mean value, $\hat{\sigma}^2 = \kappa - \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{k}$ is the posterior variance, $\mathbf{K}$ is a matrix with elements $\mathbf{K}_{ij} = C(\mathbf{t}_i^0, \mathbf{t}_j^0)$, $\mathbf{k}$ is a vector with elements $\mathbf{k}_i = C(\mathbf{t}_i^0, t)$, and $\kappa = C(t, t)$ is a scalar. Equation (A.3) is used to compute the functions representing the best estimate $\hat{y}(t)$, uncertainty in the estimate $\hat{\sigma}^2$, and the probability density $\Pr(f(t_i) = y|\mathbf{S}^0, \mathbf{t}^0)$over values of y, all calculated at times t_i. This was the goal of the transformation step. The rational quadratic function

$$C_{\mathrm{RQ}}(t_1, t_2) = \sigma^2 \exp\left[1 + \frac{(t_1 - t_2)^2}{2\alpha\tau^2}\right]^{-\alpha} \tag{A.4}$$

is one covariance function to control the estimate of $\hat{y}$ and $\hat{\sigma}$. It is an infinite sum of squared exponential covariance functions, each with a different time scale t and their relative contribution defined by a gamma distribution over $\ell = \tau^{-2}$, parameterized by $\alpha > 0$ [18].

We can tune the hyperparameters of covariance functions for an optimal fitting using the exact marginal likelihood of the hyperparameters because it balances the fit against the

complexity of the model:

$$\log \Pr\left(\mathbf{S}^0 \middle| \mathbf{t}^0, h\right) = -\frac{1}{2}\mathbf{S}^{0\top}\mathbf{K}^{-1}\mathbf{S}^0 - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi, \tag{A.5}$$

The first term assesses how well the model fits the observed data, the second term is a penalty on $\mathbf{K}$, and the third term is a normalization constant. The parameters we used in this paper is listed in Table A.1.

Table A.1: The parameters for Gaussian process regression.

Measurement	σ^2	τ	α
PHQ-8	100	0.25	0.12
Item 9	100	0.5	0.1

A.1.3 Unsupervised classification with K-means algorithm

We treat the activation $\mathbf{h}$ as the features of each patient for further analysis of subtype detection. The k-means algorithm divides a set of N samples into K disjoint clusters C, each described by the mean μ_j of the samples in the cluster. The means are commonly called the cluster "centroids". The k-means algorithm aims to choose centroids that minimize the inertia, or within-cluster sum of squared criterion

$$\sum_{i=0}^{n}\left(\|x_j - \mu_i\|^2\right), \tag{A.6}$$

The correct choice of K is often ambiguous, with interpretations depending on the shape and scale of the distribution of points in a data set and the desired clustering resolution of the user. In addition, increasing K without penalty will always reduce the amount of error in the resulting clustering, to the extreme case of zero error if each data point is considered

its own. The optimal choice of K should keep a balance between maximum compression of the data using a single cluster, and maximum accuracy by assigning each data point to its own cluster [78]. As the number of clusters increases, we choose K as the number of clusters if the inertia decreases fast for number of clusters smaller than K, and decreases slowly for number of clusters greater than K. For the initialization of the centroids, we use the k-means++ algorithm that aims at that minimize the intra-class variance [128].

A.2 The results of Item 9 trajectories

We perform cross-validation analysis for the Item 9 trajectories. The result is $\beta = 10$ and $\rho = 0.5$, $H = 25$ and $\lambda = 3$.

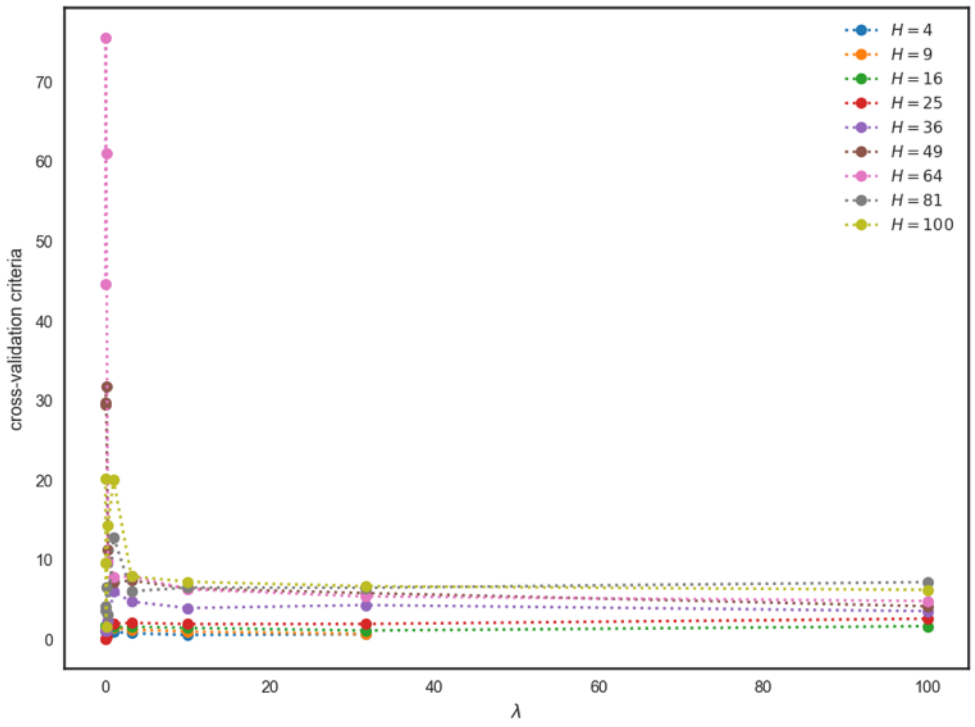


Figure A.1: The result of cross-validation for model selection by the Item 9 trajectories, focused on number of nodes in hidden layers (H) and regularization parameters λ (on scale of W)

The learned patterns contains the following shapes: (1) Simple trend detectors, such as the increasing trend, the decreasing trend, and the stable trend; (2) Multiple trends, such as increasing first and then becoming stable, or increasing first and then decreasing. (3) Sinusoidal functions, which implies that the trajectory has a periodic behavior within a short time.

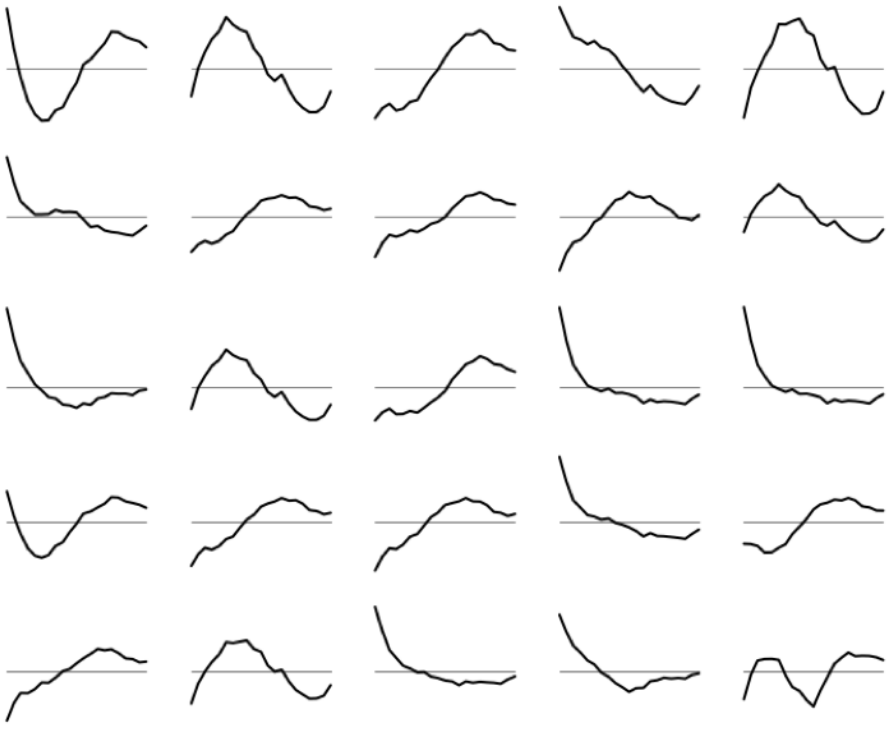


Figure A.2: Hidden patterns learned from the Item 9 trajectories. These patterns are visualized directly as the rows W_i.

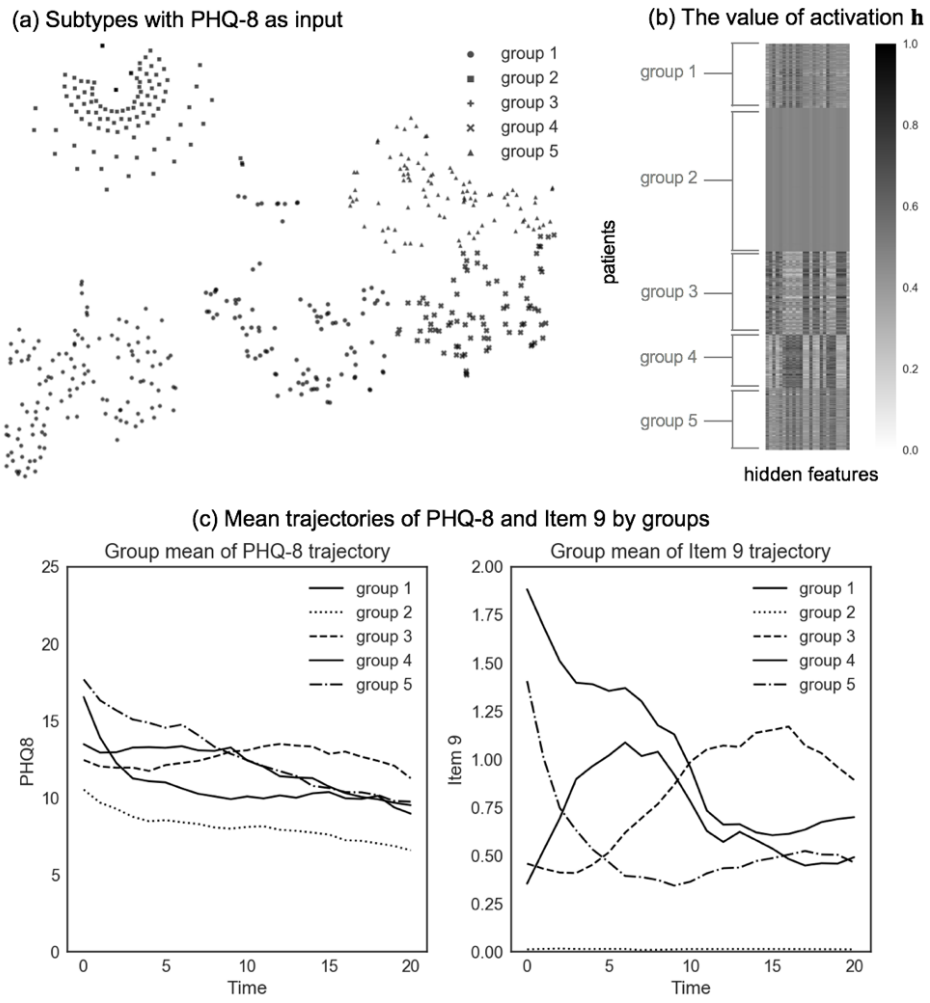


Figure A.3: The subtype analysis result with hidden structures learned from the Item 9 trajectories. (a) Embed the activation **h** (25 dimensions) of each patient into 2 dimensional space with t-SNE and cluster them with the k-means algorithm. (b) The value of activation **h** on each latent pattern (25 columns) of each patient (610 rows) after reordering the rows by the clustering. (c) Mean trajectories of PHQ-8 and Item 9 by groups, using the clustering results. One unit of time is two weeks. We can see that the group mean Item 9 trajectories are more distinguishable than that of PHQ-8, because the subtype classification is performed on the features learned from the Item 9 trajectories.

Appendix B

APPENDIX FOR CHAPTER 3

B.1 Proof of Theorem 3.1

The proof of Theorem 3.1 is based on the proof of the equivalent objective in the general EM algorithm in [76, Charter 9]. First we prove the following lemma.

Lemma B.1 (The EM Theorem for POCM)**.** *Denote* $\mathbf{O}$ *as the sequence of observations and* $\mathbf{S}$ *as the sequence of the latent states. Let* $\theta, \mathbf{C}$ *and* $\theta', \mathbf{C}'$ *be two different basis parameters, if*

$$\sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log[\Pr(\mathbf{O},\mathbf{S}|\theta',\mathbf{C}')] > \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log[\Pr(\mathbf{O},\mathbf{S}|\theta,\mathbf{C})] \tag{B.1}$$

then $\Pr(\mathbf{O}|\theta',\mathbf{C}') > \Pr(\mathbf{O}|\theta,\mathbf{C})$.

From Lemma B.1, if we can find basis parameters $\theta', \mathbf{C}'$ for which (B.1) holds, then the observation sequence $\mathbf{O}$ will be more probable under $\theta', \mathbf{C}'$ than under $\theta, \mathbf{C}$.

Proof of Lemma B.1. Since $\sum_{s\in\mathcal{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) = 1$, we have

$$\begin{aligned}
&\log[\Pr(\mathbf{O}|\theta',\mathbf{C}')] - \log[\Pr(\mathbf{O}|\theta,\mathbf{C})] \\
&= \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log[\Pr(\mathbf{O}|\theta',\mathbf{C}')] - \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log[\Pr(\mathbf{O}|\theta,\mathbf{C})] \\
&= \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log \frac{\Pr(\mathbf{O},\mathbf{S}|\theta',\mathbf{C}')}{\Pr(\mathbf{S}|\mathbf{O},\theta',\mathbf{C}')} - \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log \frac{\Pr(\mathbf{O},\mathbf{S}|\theta,\mathbf{C})}{\Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C})} \\
&= \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log \frac{\Pr(\mathbf{O},\mathbf{S}|\theta',\mathbf{C}')}{\Pr(\mathbf{S}|\mathbf{O},\theta',\mathbf{C}')} + \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log \frac{\Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C})}{\Pr(\mathbf{O},\mathbf{S}|\theta,\mathbf{C})} \\
&= \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log \frac{\Pr(\mathbf{O},\mathbf{S}|\theta',\mathbf{C}')}{\Pr(\mathbf{O},\mathbf{S}|\theta,\mathbf{C})} + \sum_{\mathbf{S}\in\mathbb{S}} \Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}) \log \frac{\Pr(\mathbf{S}|\mathbf{O},\theta,\mathbf{C})}{\Pr(\mathbf{S}|\mathbf{O},\theta',\mathbf{C}')}.
\end{aligned}$$

From Gibbs inequality [129, §2.6], the KL-divergence between any two random variables are nonnegative,

$$D_{\mathrm{KL}}\left(\mathbf{S}|\mathbf{O},\theta,\mathbf{C} \parallel \mathbf{S}|\mathbf{O},\theta',\mathbf{C}'\right) = \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}\right) \log \frac{\Pr\left(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}\right)}{\Pr\left(\mathbf{S}|\mathbf{O},\theta',\mathbf{C}'\right)} \geq 0. \tag{B.2}$$

Therefore

$$\log[\Pr\left(\mathbf{O}|\theta',\mathbf{C}'\right)] - \log[\Pr\left(\mathbf{O}|\theta,\mathbf{C}\right)] \geq \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}\right) \log \frac{\Pr\left(\mathbf{O},\mathbf{S}|\theta',\mathbf{C}'\right)}{\Pr\left(\mathbf{S}|\mathbf{O},\theta,\mathbf{C}\right)} > 0. \tag{B.3}$$

The last step is based on the given condition. Then we have $\Pr\left(\mathbf{O}|\theta',\mathbf{C}'\right) > \Pr\left(\mathbf{O}|\theta,\mathbf{C}\right)$. □

Proof of Theorem 3.1. Let $\theta^{(m)}, \mathbf{C}^{(m)}$ be the basis parameters after m iterations. From Lemma B.1, if

$$\sum_{s\in\mathcal{S}} \Pr\left(\mathbf{S}\middle|\mathbf{O},\theta^{(m)},\mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O},\mathbf{S}|\theta',\mathbf{C}'\right)] > \sum_{s\in\mathcal{S}} \Pr\left(\mathbf{S}\middle|\mathbf{O},\theta^{(m)},\mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O},\mathbf{S}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)],$$

then $\Pr\left(\mathbf{O}|\theta',\mathbf{C}'\right) > \Pr\left(\mathbf{O}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)$, which indicates

$$\arg\max_{\theta,\mathbf{C}} \Pr\left(\mathbf{O}|\theta,\mathbf{C}\right) = \arg\max_{\theta,\mathbf{C}} \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{S}\middle|\mathbf{O},\theta^{(m)},\mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O},\mathbf{S}|\theta,\mathbf{C}\right)]. \tag{B.4}$$

Noting that $\Pr\left(\mathbf{O},\mathbf{S}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right) = \Pr\left(\mathbf{O}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right) \Pr\left(\mathbf{S}\middle|\mathbf{O},\theta^{(m)},\mathbf{C}^{(m)}\right)$, and $\Pr\left(\mathbf{O}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)$ is constant because the observation sequence is not affected by the choice of basis parameters. Therefore,

$$\begin{aligned}
\arg\max_{\theta,\mathbf{C}} \Pr\left(\mathbf{O}|\theta,\mathbf{C}\right) &= \arg\max_{\theta,\mathbf{C}} \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{S}\middle|\mathbf{O},\theta^{(m)},\mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O},\mathbf{S}|\theta,\mathbf{C}\right)] \\
&= \arg\max_{\theta,\mathbf{C}} \sum_{\mathbf{S}\in\mathbb{S}} \frac{\Pr\left(\mathbf{O},\mathbf{S}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)}{\Pr\left(\mathbf{O}|\theta^{(m)},\mathbf{C}^{(m)}\right)} \log[\Pr\left(\mathbf{O},\mathbf{S}|\theta,\mathbf{C}\right)] \\
&= \arg\max_{\theta,\mathbf{C}} \sum_{\mathbf{S}\in\mathbb{S}} \Pr\left(\mathbf{O},\mathbf{S}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right) \log[\Pr\left(\mathbf{O},\mathbf{S}|\theta,\mathbf{C}\right)] \\
&= \arg\max_{\theta,\mathbf{C}} Q\left(\theta,\mathbf{C}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right).
\end{aligned}$$

□

B.2 Derivation of POCM updating rule

B.2.1 Objective function

From Theorem 3.1, we have

$$\arg\max_{\theta,\mathbf{C}} \Pr\left(\mathbf{O}|\theta,\mathbf{C}\right) = \arg\max_{\theta,\mathbf{C}} Q\left(\theta,\mathbf{C}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right).$$

Therefore,

$$\theta^{(m+1)} = \arg\max_{\theta} Q\left(\theta,\mathbf{C}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right), \tag{B.5}$$

$$\mathbf{C}^{(m+1)} = \arg\max_{\mathbf{C}} Q\left(\theta,\mathbf{C}\middle|\theta^{(m+1)},\mathbf{C}^{(m)}\right). \tag{B.6}$$

We can write $\Pr(\mathbf{O},\mathbf{S}|\theta,\mathbf{C})$ as

$$\Pr\left(\mathbf{O},\mathbf{S}|\theta,\mathbf{C}\right) = \prod_{i=1}^{N}\left(\hat{\boldsymbol{\pi}}_i\left(s_{i,1}\right)\hat{\mathbf{B}}_i\left(s_{i,1},o_{i,1}\right)\prod_{t=2}^{T}\hat{\mathbf{A}}_i\left(s_{i,t+1},s_{i,t}\right)\hat{\mathbf{B}}_i\left(s_{i,t},o_{i,t}\right)\right), \tag{B.7}$$

where $i = 1,\ldots,N$ is the set of individuals; $\hat{\boldsymbol{\pi}}_i$, $\hat{\mathbf{A}}_i$, $\hat{\mathbf{B}}_i$ is the estimated individual parameters, and $o_{i,t}$ and $s_{i,t}$ is the observation and latent state of patient i at period t respectively. Taking the log gives us

$$\log[\Pr\left(\mathbf{O},\mathbf{S}|\theta,\mathbf{C}\right)] = \sum_{i=1}^{N}\left[\log\hat{\boldsymbol{\pi}}_i\left(s_{i,1}\right) + \sum_{t=2}^{T}\log\hat{\mathbf{A}}_i\left(s_{i,t+1},s_{i,t}\right) + \sum_{t=1}^{T}\log\hat{\mathbf{B}}_i\left(s_{i,t},o_{i,t}\right)\right] \tag{B.8}$$

Plugging this into $Q\left(\theta, \mathbf{C} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right)$, we get

$$\begin{aligned}
Q\left(\theta, \mathbf{C} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right) &= \sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \log \hat{\boldsymbol{\pi}}_i(s_{i,1}) \Pr\left(\mathbf{O}, \mathbf{S} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right) \\
&+ \sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \hat{\mathbf{A}}_i(s_{i,t+1}, s_{i,t}) \Pr\left(\mathbf{O}, \mathbf{S} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right) \\
&+ \sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \hat{\mathbf{B}}_i(s_{i,t}, o_{i,t}) \Pr\left(\mathbf{O}, \mathbf{S} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right) \\
&\text{(expand the basis model)} \\
&= \sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \mathbf{c}_{ik} \boldsymbol{\pi}_k(s_{i,1}) \right) \Pr\left(\mathbf{O}, \mathbf{S} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right) \\
&+ \sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\sum_{k=2}^{K} \mathbf{c}_{ik} \mathbf{A}_{k(s_{i,t+1}, s_{i,t})} \right) \Pr\left(\mathbf{O}, \mathbf{S} | \theta^{(m)}, \mathbf{C}^{(m)}\right) \\
&+ \sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} \mathbf{c}_{ik} \mathbf{B}_k\left(s_{i,t}, o_{i,t}\right) \right) \Pr\left(\mathbf{O}, \mathbf{S} | \theta^{(m)}, \mathbf{C}^{(m)}\right)
\end{aligned}$$

This is a nice form which we can optimize analytically with Lagrange multipliers. The objective function then can be written in the equivalent form as

$$f'(\theta, \mathbf{C}) = Q\left(\theta, \mathbf{C} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right) - \frac{\mu}{2} \operatorname{Tr}\left(\mathbf{C}^\top \mathbf{L} \mathbf{C}\right)$$

B.2.2 Lagrange multiplier method

We need Lagrange multipliers because we have equality constraints which come from requiring that $\boldsymbol{\pi}$, $\mathbf{A}_i$ and $\mathbf{B}_i$ form valid probability distributions. Let $L(\theta, \mathbf{C} | \theta^{(m)}, \mathbf{C}^{(m)})$ be the

Lagrangian

$$L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right) = Q\left(\theta, \mathbf{C}|\theta^{(m)}, C^{(m)}\right) - \mu \operatorname{Tr}\left(\mathbf{C}^T\mathbf{L}\mathbf{C}\right) - \sum_{k=1}^{K} \lambda_k^{(\boldsymbol{\pi})} \left(\sum_{s=1}^{|\mathcal{S}|} \boldsymbol{\pi}_k(s) - 1\right)$$
$$- \sum_{k=1}^{K}\sum_{s=1}^{|\mathcal{S}|} \lambda_{s,k}^{(\mathbf{A})} \left(\sum_{s'=1}^{|\mathcal{S}|} \mathbf{A}_k(s,s') - 1\right) - \sum_{k=1}^{K}\sum_{s=1}^{|\mathcal{S}|} \lambda_{s,k}^{(\mathbf{B})} \left(\sum_{o=1}^{\Omega} \mathbf{B}_k(s,o) - 1\right)$$
$$- \sum_{i=1}^{N} \lambda_i^{(\mathbf{C})} \left(\sum_{k=1}^{K} c_{i,k} - 1\right)$$

where all the $\lambda_{\cdot}^{(\cdot)}$'s are Lagrangian multipliers. Next, we will solve this optimization problem with a variant of the Expectation-Maximization (EM) algorithm. The E-step is to estimate some intermediate state with current estimator of parameters, and then in the M-step the new parameters are updated using the estimated intermediate state from the E-step. We can show that the E-step in the inference algorithms of HMM and POCM are the same, which is called the Forward-backward algorithm. The M-step of HMM learning is a direct result of Forward-backward algorithm, while the M-step of POCM learning is itself a iterative procedure.

B.2.3 Forward-backward algorithm

At step m, with parameter estimator $\theta^{(m)}, \mathbf{C}^{(m)}$, we can estimate the following probabilities for each subject $i = 1, \ldots, N$: (1) $\gamma_{i,t}^{(m)}(s) = \Pr\left(s_{i,t} = s \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right)$, the probability that the system is at state s at period t given the observation sequence $\mathbf{O}_i$ and the model $\theta^{(m)}, \mathbf{C}^{(m)}$, and (2) $\xi_{i,t}^{(m)}(s,s') = \Pr\left(s_{i,t} = s, s_{i,t+1} = s' \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right)$, the probability of being at state s at time t, and at state s' at time $t+1$ given the observation sequence $\mathbf{O}_i$ and the model $\theta^{(m)}, \mathbf{C}^{(m)}$. First denote the initial distribution of state $\hat{\boldsymbol{\pi}}_i^{(m)} = \sum_{k=1}^{K} c_{i,k}^{(m)} \boldsymbol{\pi}_k^{(m)}$, the transition matrix $\hat{\mathbf{A}}_i^{(m)} = \sum_{k=1}^{K} c_{i,k}^{(m)} \mathbf{A}_k^{(m)}$, and the emission matrix $\hat{\mathbf{B}}_i^{(m)} = \sum_{k=1}^{K} c_{i,k}^{(m)} \mathbf{B}_k^{(m)}$.

Let $\alpha_{i,t}^{(m)}(s) = \Pr\left(\mathbf{O}_{i,(1:t)}, s_{i,t} = s|\theta^{(m)}, \mathbf{C}^{(m)}\right)$, where $\mathbf{O}_{i,(1:t)}$ is the partial sequence observations, up to time period t. Then we can compute the vectors $\alpha_{i,t}^{(m)}(s)$ iteratively

- Initialize $\alpha_{i,1}^{(m)}(s) = \hat{\boldsymbol{\pi}}_i^{(m)}(s)\hat{\mathbf{B}}_i^{(m)}(s, o_{i,1})$

- Repeat, for $t = 1, \ldots, T$, and for all s

$$\alpha_{i,t+1}^{(m)}(s) = \hat{\mathbf{B}}_i^{(m)}(s, o_{i,t+1}) \sum_{s'=1}^{|\mathcal{S}|} \alpha_{i,t}^{(m)}(s) \hat{\mathbf{A}}_i^{(m)}(s, s') \tag{B.9}$$

Let $\beta_{i,t}^{(m)}(s) = P\left(\mathbf{O}_{i,(t+1:T)} \middle| s_{i,t} = s, \theta^{(m)}, \mathbf{C}^{(m)}\right)$, where $\mathbf{O}_{i,(t+1,T)}$ is the partial sequence observations, from time $t+1$ to the final time T. Then we can compute the vectors $\beta_{i,t}^{(m)}(s)$ iteratively

- Initialize $\beta_{i,T}^{(m)}(s) = 1$ for $s = 1, \ldots, |\mathcal{S}|$

- Repeat, for $t = t+1, \ldots, 1$, and for all s

$$\beta_{i,t}^{(m)}(s) = \sum_{s'=1}^{|\mathcal{S}|} \beta_{i,t+1}^{(m)}(s') \hat{\mathbf{A}}_i^{(m)}(s, s') \hat{\mathbf{B}}_i^{(m)}(s', o_{i,t+1}) \tag{B.10}$$

Then

$$\gamma_{i,t}^{(m)}(s) = \frac{\alpha_{i,t}^{(m)}(s) \beta_{i,t}^{(m)}(s)}{P\left(\mathbf{O} | \theta^{(m)}, \mathbf{C}^{(m)}\right)} = \frac{\alpha_{i,t}^{(m)}(s) \cdot \beta_{i,t}^{(m)}(s)}{\sum_{s=1}^{|\mathcal{S}|} \alpha_{i,t}^{(m)}(s)}, \tag{B.11}$$

$$\xi_{i,t}^{(m)}(s, s') = \eta_{i,t}^{(m)} \cdot \alpha_{i,t}^{(m)}(s) \cdot \beta_{i,t+1}^{(m)}(s') \cdot \hat{\mathbf{A}}_i^{(m)}(s, s') \cdot \hat{\mathbf{B}}_i^{(m)}\left(s', o_{i,t+1}\right) \tag{B.12}$$

where $\eta_{i,t}^{(m)}$ is a normalization factor, such that $\sum_{s,s'} \xi_{i,t}^{(m)}(s, s') = 1$.

B.2.4 The Baum-Welch Algorithm

The algorithm of parameter inference of the hidden Markov model is known as the Baum-Welch Algorithm. The M-step is as follows. The updated initial distribution is

$$\pi^{(m+1)}(s) = \gamma_{1,t}^{(m)}(s). \tag{B.13}$$

The entries of the updated transition matrix are

$$\hat{\mathbf{A}}^{(m+1)}(s, s') = \frac{\sum_{t=1}^{T-1} \xi_{i,t}^{(m)}(s, s')}{\sum_{t=1}^{T-1} \gamma_{i,t}^{(m)}(s)}. \tag{B.14}$$

The entries of the updated emission matrix are

$$\hat{\mathbf{B}}^{(m+1)}(s, o) = \frac{\sum_{t=1}^{T} \gamma_{i,t}^{(m)}(s) \cdot I(o_{i,t} = o)}{\sum_{t=1}^{T} \gamma_{i,t}^{(m)}(s)} \tag{B.15}$$

B.2.5 POCM Learning First Stage: updating basis parameters

The M-step for POCM learning is divided into two stages. We first fix the membership vector $\mathbf{C}^{(m)}$, and obtain the basis parameters $\boldsymbol{\pi}_k^{(m+1)}, \mathbf{A}_k^{(m+1)}, \mathbf{B}_k^{(m+1)}$ by solving the optimization problem. Then we fix the basis parameters $\boldsymbol{\pi}_k^{(m+1)}, \mathbf{A}_k^{(m+1)}, \mathbf{B}_k^{(m+1)}$, and compute the optimal membership vector $\mathbf{C}^{(m+1)}$.

(1) First let us focus on the $\boldsymbol{\pi}_k$'s.

$$\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \boldsymbol{\pi}_k(s)} = \frac{\partial}{\partial \boldsymbol{\pi}_k(s)} \left(\sum_{\mathbf{S}\in\mathbb{S}} \sum_{i=1}^{N} \log \left(\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s_{i,1}) \right) P\left(\mathbf{O}, \mathbf{S}|\theta^{(m)}, \mathbf{C}^{(m)}\right) \right) - \lambda_k^{(\boldsymbol{\pi})}$$
$$= \frac{\partial}{\partial \boldsymbol{\pi}_k(s)} \left(\sum_{s'=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \log \left(\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s') \right) P\left(s_{i,1}=s' \middle| \mathbf{X}, \theta^{(m)}, \mathbf{C}^{(m)}\right) \right) - \lambda_k^{(\boldsymbol{\pi})}$$

(derivative is w.r.t. to $\boldsymbol{\pi}_k(s)$, so eliminate all $s' \neq s$, and use the fact: $\frac{\partial}{\partial x_k} \log\left(\sum_{k'=1}^{K} \alpha_{k'} x_{k'}\right) = \frac{\alpha_k}{\sum_{k'=1}^{K} \alpha_{k'} x_{k'}}$, we have)

$$= \sum_{i=1}^{N} \frac{c_{i,k} P\left(s_{i,1}=s \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s)} - \lambda_k^{(\boldsymbol{\pi})} = \sum_{i=1}^{N} \frac{c_{i,k} \gamma_{i,1}^{(m)}(s)}{\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s)} - \lambda_k^{(\boldsymbol{\pi})} = 0$$

$$(s = 1, \ldots, |\mathcal{S}|, k = 1, \ldots, K)$$

$$\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \lambda_k^{(\boldsymbol{\pi})}} = -\left(\sum_{s=1}^{|\mathcal{S}|} \boldsymbol{\pi}_k(s) - 1 \right) = 0, (k = 1, \ldots, K)$$

Use $\sum_{s=1}^{|\mathcal{S}|} \frac{\partial L\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \boldsymbol{\pi}_k(s)} \boldsymbol{\pi}_k(s) = 0$, and note $\sum_{s=1}^{|\mathcal{S}|} \boldsymbol{\pi}_k(s) = 1$, we have

$$\lambda_k^{(\boldsymbol{\pi})} = \sum_{s=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \frac{c_{i,k} \gamma_{i,1}^{(m)}(s) \boldsymbol{\pi}_k(s)}{\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s)}$$

Substitute this into $\frac{\partial L\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \boldsymbol{\pi}_k(s)} \boldsymbol{\pi}_k(s) = 0$, we have

$$\sum_{i=1}^{N} \frac{c_{i,k} \gamma_{i,1}^{(m)}(s) \boldsymbol{\pi}_k(s)}{\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s)} - \sum_{s=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \frac{c_{i,k} \gamma_{i,1}^{(m)}(s) \boldsymbol{\pi}_k(s)}{\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s)} \boldsymbol{\pi}_k(s) = 0$$

which leads to the updating rules

$$\boldsymbol{\pi}_k^{(m+1)}(s) = \frac{\sum_{i=1}^{N} \frac{c_{i,k}^{(m)} \gamma_{i,1}^{(m)}(s) \boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'=1}^{K} c_{i,k'}^{(m)} \boldsymbol{\pi}_{k'}^{(m)}(s)}}{\sum_{s=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \frac{c_{i,k}^{(m)} \gamma_{i,1}^{(m)}(s) \boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'=1}^{K} c_{i,k'}^{(m)} \boldsymbol{\pi}_{k'}^{(m)}(s)}}$$

(2) We now follow a similar process for the $\mathbf{A}_k$'s

$$\begin{aligned}
&\frac{\partial L\left(\theta, \mathbf{C} | \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \mathbf{A}_k(s, s')} \\
&= \frac{\partial}{\partial \mathbf{A}_k(s, s')} \left(\sum_{s \in \mathcal{S}} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \log \left(\sum_{k=1}^{K} c_{i,k} \mathbf{A}_k\left(s_{i,t}, s_{i,t+1}\right) \right) P\left(\mathbf{O}, \mathbf{S} | \theta^{(m)}, \mathbf{C}^{(m)}\right) \right) - \lambda_{s,k}^{(\mathbf{A})} \\
&= \frac{\partial}{\partial \mathbf{A}_k(s, s')} \left(\sum_{s'=1}^{|\mathcal{S}|} \sum_{s''=1}^{S} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \log \left(\sum_{k=1}^{K} c_{i,k} \mathbf{A}_k\left(s', s''\right) \right) P\left(s_{i,t} = s', s_{i,t+1} = s'', \mathbf{O} | \theta^{(m)}, \mathbf{C}^{(m)}\right) \right) - \lambda_{s,k}^{(\mathbf{A})} \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k} P\left(s_{i,t} = s', s_{i,t+1} = s'' \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{A}_{k'}(s, s')} - \lambda_{s,k}^{(\mathbf{A})} \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k} \xi_{i,t}^{(m)}(s, s')}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{A}_{k'}(s, s')} - \lambda_{s,k}^{(\mathbf{A})} = 0, \quad (s, s' = 1, \ldots, |\mathcal{S}|, k = 1, \ldots, K)
\end{aligned}$$

$$\frac{\partial L\left(\theta, \mathbf{C} | \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \lambda_{s,k}^{(\mathbf{A})}} = -\left(\sum_{s'=1}^{|\mathcal{S}|} \mathbf{A}_k(s, s') - 1 \right) = 0, (s = 1, \ldots, |\mathcal{S}|, k = 1, \ldots, K)$$

Use $\sum_{s'=1}^{|\mathcal{S}|} \frac{\partial L\left(\theta, \mathbf{C} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \mathbf{A}_k(s, s')} \mathbf{A}_k(s, s') = 0$, and note $\sum_{s'=1}^{|\mathcal{S}|} \mathbf{A}_k(s, s') = 1$, we have

$$\lambda_{s,k}^{(\mathbf{A})} = \sum_{s'=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k} \xi_{i,t}^{(m)}(s, s') \mathbf{A}_k(s, s')}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{A}_{k'}(s, s')}$$

Substitute this into $\frac{\partial L\left(\theta, \mathbf{C} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \mathbf{A}_k(s, s')} \mathbf{A}_k(s, s') = 0$, we have

$$\sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k} \xi_{i,t}^{(m)}(s, s') \mathbf{A}_k(s, s')}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{A}_{k'}(s, s')} - \sum_{s'=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k} \xi_{i,t}^{(m)}(s, s') \mathbf{A}_k(s, s')}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{A}_{k'}(s, s')} \mathbf{A}_k(s, s') = 0$$

which leads to the updating rules

$$\mathbf{A}_k^{(m+1)}(s, s') = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k}^{(m)} \xi_{i,t}^{(m)}(s,s') \mathbf{A}_k^{(m)}(s,s')}{\sum_{k'=1}^{K} c_{i,k'}^{(m)} \mathbf{A}_{k'}^{(m)}(s,s')}}{\sum_{s'=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{c_{i,k}^{(m)} \xi_{i,t}^{(m)}(s,s') \mathbf{A}_k^{(m)}(s,s')}{\sum_{k'=1}^{K} c_{i,k'}^{(m)} \mathbf{A}_{k'}^{(m)}(s,s')}}$$

(3) And similarly for $\mathbf{B}_k$'s

$$\begin{aligned}
&\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \mathbf{B}_k(s,o)} \\
&= \frac{\partial}{\partial \mathbf{B}_k(s,o)} \left(\sum_{s\in\mathcal{S}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} c_{i,k} \mathbf{B}_k\left(s_{i,t}, o_{i,t}\right) \right) P\left(\mathbf{O}, \mathbf{S}|\theta^{(m)}, \mathbf{C}^{(m)}\right) \right) - \lambda_{s,k}^{(\mathbf{B})} \\
&= \frac{\partial}{\partial \mathbf{B}_k(s,o)} \left(\sum_{s'=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} c_{i,k} \mathbf{B}_k(s',o) \right) P\left(s_{i,t}=s', \mathbf{O}|\theta^{(m)}, \mathbf{C}^{(m)}\right) \right) - \lambda_{s,k}^{(\mathbf{B})} \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k} P\left(s_{i,t}=s \middle| \mathbf{O}_i, \theta^{(m)}, \mathbf{C}^{(m)}\right) I(o_{i,t}=o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)} - \lambda_{s,k}^{(\mathbf{B})} \\
&= \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)} - \lambda_{s,k}^{(\mathbf{B})} = 0, \quad (s=1,\ldots,|\mathcal{S}|, k=1,\ldots,K)
\end{aligned}$$

$$\frac{\partial L\left(\theta, \mathbf{C}|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \lambda_{s,k}^{(\mathbf{B})}} = -\left(\sum_{o=1}^{|\Omega|} \mathbf{B}_k(s,o) - 1 \right) = 0, \quad (s=1,\ldots,|\mathcal{S}|, k=1,\ldots,K)$$

Use $\sum_{o=1}^{|\Omega|} \frac{\partial L\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \mathbf{B}_k(s,o)} \mathbf{B}_k(s,o) = 0$, and note $\sum_{o=1}^{|\Omega|} \mathbf{B}_k(s,o) = 1$, we have

$$\lambda_{s,k}^{(\mathbf{B})} = \sum_{o=1}^{|\Omega|} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o) \mathbf{B}_{k'}(s,o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)}$$

Substitute this into $\frac{\partial L\left(\theta, \mathbf{C}\middle|\theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial \mathbf{B}_k(s,o)} \mathbf{B}_k(s,o) = 0$, we have

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o) \mathbf{B}_{k'}(s,o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)} - \sum_{o=1}^{|\Omega|} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o) \mathbf{B}_{k'}(s,o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)} \mathbf{B}_k(s,o) = 0$$

which leads to the updating rules

$$\mathbf{B}_k^{(m+1)}(s,o) = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k}^{(m)} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o) \mathbf{B}_k^{(m)}(s,o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)}}{\sum_{o=1}^{|\Omega|} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{c_{i,k}^{(m)} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o) \mathbf{B}_k^{(m)}(s,o)}{\sum_{k'=1}^{K} c_{i,k'} \mathbf{B}_{k'}(s,o)}}$$

B.2.6 POCM Learning Second Stage: Updating membership vectors

Then we fix $\theta^{(m)}$, let us focus on the $\mathbf{c}_{i,k}$'s

$$
\begin{aligned}
&\frac{\partial L\left(\theta, \mathbf{C} \middle| \theta^{(m+1)}, \mathbf{C}^{(m)}\right)}{\partial c_{i,k}} \\
&= \frac{\partial}{\partial c_{i,k}} \left(\sum_{s'=1}^{|\mathcal{S}|} \sum_{i=1}^{N} \log \left(\sum_{k'=1}^{K} c_{i,k'} \boldsymbol{\pi}_{k'}(s') \right) P\left(s_{i,1} = s', \mathbf{O} \middle| \theta^{(m+1)}, \mathbf{C}^{(m)}\right) \right) \\
&+ \frac{\partial}{\partial c_{i,k}} \left(\sum_{s \in \mathcal{S}} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \log \left(\sum_{k=1}^{K} c_{i,k} \mathbf{A}_k\left(s_{i,t}, s_{i,t+1}\right) \right) P\left(\mathbf{O}, \mathbf{S} \middle| \theta^{(m+1)}, \mathbf{C}^{(m)}\right) \right) \\
&+ \frac{\partial}{\partial c_{i,k}} \left(\sum_{s \in \mathcal{S}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\sum_{k=1}^{K} c_{i,k} \mathbf{B}_k\left(s_{i,t}, o_{i,t}\right) \right) P\left(\mathbf{O}, \mathbf{S} \middle| \theta^{(m+1)}, \mathbf{C}^{(m)}\right) \right) \\
&- 2\mu \left(\mathbf{L}\mathbf{C}\right)_{i,k} - \lambda_i^{(\mathbf{c})} \\
&= \sum_{s=1}^{|\mathcal{S}|} \frac{\boldsymbol{\pi}_k(s) P\left(s_{i,1} = s \middle| \mathbf{O}_i, \theta^{(m+1)}, \mathbf{C}^{(m)}\right)}{\sum_{k'=1}^{K} \boldsymbol{\pi}_{k'}(s) c_{i,k'}} \\
&+ \sum_{t=1}^{T-1} \sum_{s=1}^{|\mathcal{S}|} \sum_{s'=1}^{|\mathcal{S}|} \frac{\mathbf{A}_k(s, s') P\left(s_{i,t} = s, s_{i,t+1} = s' \middle| \mathbf{O}_i, \theta^{(m+1)}, \mathbf{C}^{(m)}\right)}{\sum_{k'=1}^{K} \mathbf{A}_{k'}(s, s') c_{i,k'}} \\
&+ \sum_{t=1}^{T} \sum_{s=1}^{|\mathcal{S}|} \frac{\mathbf{B}_k\left(s, o_{i,t}\right) P\left(s_{i,t} = s \middle| \mathbf{O}_i, \theta^{(m+1)}, \mathbf{C}^{(m)}\right)}{\sum_{k'=1}^{K} \mathbf{B}_{k'}\left(s, o_{i,t}\right) c_{i,k'}} - 2\mu \left(\mathbf{L}\mathbf{C}\right)_{i,k} - \lambda_i^{(\mathbf{c})} \\
&= \sum_{s=1}^{|\mathcal{S}|} \frac{\boldsymbol{\pi}_k(s) \gamma_{i,1}^{(m)}(s)}{\sum_{k'=1}^{K} \boldsymbol{\pi}_{k'}(s) c_{i,k'}} + \sum_{t=1}^{T-1} \sum_{s=1}^{|\mathcal{S}|} \sum_{s'=1}^{|\mathcal{S}|} \frac{\mathbf{A}_k(s, s') \xi_{i,t}^{(m)}(s, s')}{\sum_{k'=1}^{K} \mathbf{A}_{k'}(s, s') c_{i,k'}} \\
&+ \sum_{t=1}^{T} \sum_{s=1}^{|\mathcal{S}|} \frac{\mathbf{B}_k\left(s, o_{i,t}\right) \gamma_{i,t}^{(m)}(s)}{\sum_{k'=1}^{K} \mathbf{B}_{k'}\left(s, o_{i,t}\right) c_{i,k'}} - 2\mu \left(\mathbf{L}\mathbf{C}\right)_{i,k} - \lambda_i^{(\mathbf{c})} = 0,
\end{aligned}
$$

$(i = 1, \ldots, N, k = 1, \ldots, K)$

$$
\frac{\partial L\left(\theta, \theta^{(m)}\right)}{\partial \lambda_i^{(\mathbf{C})}} = -\left(\sum_{k=1}^{K} c_{i,k} - 1 \right) = 0, \quad (i = 1, \ldots, N)
$$

Use $\sum_{k=1}^{K} \frac{\partial L\left(\theta, \mathbf{C} \middle| \theta^{(m)}, \mathbf{C}^{(m)}\right)}{\partial c_{i,k}} c_{i,k} = 0$, and note $\sum_{k=1}^{K} c_{i,k} = 1$, we have

$$
\lambda_i^{(\mathbf{C})} = 1 + S(T-1) + T - 2\mu \left(\mathbf{L}\mathbf{C}\right)_i \mathbf{C}_i^{\top}
$$

Substitute this into $\frac{\partial L\left(\theta,\mathbf{C}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)}{\partial c_{i,k}}c_{i,k}=0$, we have

$$\sum_{s=1}^{|\mathcal{S}|}\frac{\boldsymbol{\pi}_k(s)\gamma_{i,1}^{(m)}(s)c_{i,k}}{\sum_{k'=1}^{K}\boldsymbol{\pi}_{k'}(s)c_{i,k'}}+\sum_{t=1}^{T-1}\sum_{s=1}^{|\mathcal{S}|}\sum_{s'=1}^{|\mathcal{S}|}\frac{\mathbf{A}_k(s,s')\xi_{i,t}^{(m)}(s,s')c_{i,k}}{\sum_{k'=1}^{K}\mathbf{A}_{k'}(s,s')c_{i,k'}}+\sum_{t=1}^{T}\sum_{s=1}^{|\mathcal{S}|}\frac{\mathbf{B}_k\left(s,o_{i,t}\right)\gamma_{i,t}^{(m)}(s)c_{i,k}}{\sum_{k'=1}^{K}\mathbf{B}_{k'}\left(s,o_{i,t}\right)c_{i,k'}}$$
$$-\left[S(T-1)+T+1\right]c_{i,k}+2\mu\left[\left(\mathbf{D}-\mathbf{W}\right)\mathbf{C}\right]_i\mathbf{C}_i^{\top}c_{i,k}-2\mu\left[\left(\mathbf{D}-\mathbf{W}\right)\mathbf{C}\right]_{i,k}c_{i,k}=0$$

$$c_{i,k}^{(m+1)}=\frac{c_{i,k}^{(m)}}{\eta}\left(\sum_{s=1}^{|\mathcal{S}|}\frac{\boldsymbol{\pi}_k^{(m+1)}(s)\gamma_{i,1}^{(m)}(s)}{\sum_{k'=1}^{K}\boldsymbol{\pi}_{k'}^{(m+1)}(s)c_{ik'}}+\sum_{t=1}^{T-1}\sum_{s=1}^{|\mathcal{S}|}\sum_{s'=1}^{|\mathcal{S}|}\frac{\mathbf{A}_k^{(m+1)}(s,s')\xi_{i,t}^{(m)}(s,s')}{\sum_{k'=1}^{K}\mathbf{A}_{k'}^{(m+1)}(s,s')c_{ik'}}\right.$$
$$\left.+\sum_{t=1}^{T}\sum_{s=1}^{|\mathcal{S}|}\frac{\mathbf{B}_k^{(m+1)}(s,o_{i,t})\gamma_{i,t}^{(m)}(s)}{\sum_{k'=1}^{K}\mathbf{B}_{k'}^{(m+1)}(s,o_{i,t})c_{ik'}}+2\mu\left(\mathbf{D}\mathbf{C}^{(m)}\right)_i\mathbf{C}_i^{(m)\top}+2\mu\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{i,k}\right)$$

where

$$\eta=S(T-1)+T+1+2\mu\left(\mathbf{W}\mathbf{C}^{(m)}\right)_i\mathbf{C}_i^{(m)\top}+2\mu\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{i,k}$$

B.3 Proof of Theorem 3.2

We will prove the convergence of the POCM parameter inference algorithm using the methods similar in [35, 44, 24].

The Q function in is bounded from above by zero, so the Lagrangian L is also bounded from above. The Lagrangian L will converge if L is monotonically nonincreasing. We need to show that L is nonincreasing under the updating rules of the basis parameters in each step of the iterative algorithm. Since the updating rules are essentially element wise, it is sufficient to show that

$$L_{k,s}\left(\boldsymbol{\pi}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)=\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\log\left(\sum_{k'=1}^{K}c_{i,k'}\boldsymbol{\pi}_{k'}(s')\right)\gamma_{i,1}^{(m)}(s') \tag{B.16}$$

$$L_{k,s_1,s_2}\left(\mathbf{A}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)=\sum_{s'=1}^{|\mathcal{S}|}\sum_{s''=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{A}_k\left(s',s''\right)\right)\xi_{i,t}^{(m)}\left(s',s''\right) \tag{B.17}$$

$$L_{k,s,o}\left(\mathbf{B}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)=\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{B}_k(s',o)\right)\gamma_{i,t}^{(m)}(s')I(o_{i,t}=o) \tag{B.18}$$

$$\begin{aligned}L_{i,k}\left(\mathbf{C}\middle|\theta^{(m)},\mathbf{C}^{(m)}\right)&=\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\log\left(\sum_{k'=1}^{K}c_{i,k}\boldsymbol{\pi}_k(s')\right)\gamma_{i,1}^{(m)}(s')\\&+\sum_{s'=1}^{|\mathcal{S}|}\sum_{s''=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{A}_k\left(s',s''\right)\right)\xi_{i,t}^{(m)}\left(s',s''\right)\\&+\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{B}_k(s',o)\right)\gamma_{i,t}^{(m)}(s')I(o_{i,t}=o)\end{aligned} \tag{B.19}$$

are monotonically nonincreasing under the updating steps. The updating of each parameter is with respect to the following Lagrangian functions

$$F_1(\boldsymbol{\pi},\mathbf{C})=\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\log\left(\sum_{k=1}^{K}c_{i,k}\boldsymbol{\pi}_{k'}(s')\right)\gamma_{i,1}^{(m)}(s')-\sum_{k=1}^{K}\lambda_k^{(\boldsymbol{\pi})}\left(\sum_{s=1}^{|\mathcal{S}|}\boldsymbol{\pi}_k(s)-1\right)$$

$$\begin{aligned}F_2(\mathbf{A},\mathbf{C})&=\sum_{s'=1}^{|\mathcal{S}|}\sum_{s''=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{A}_k\left(s',s''\right)\right)\xi_{i,t}^{(m)}\left(s',s''\right)\\&-\sum_{k=1}^{K}\sum_{s=1}^{|\mathcal{S}|}\lambda_{s,k}^{(\mathbf{A})}\left(\sum_{s'=1}^{|\mathcal{S}|}\mathbf{A}_k(s,s')-1\right)\end{aligned}$$

$$\begin{aligned}F_3(\mathbf{B},\mathbf{C})&=\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{B}_k(s',o)\right)\gamma_{i,t}^{(m)}(s')I(o_{i,t}=o)\\&-\sum_{k=1}^{K}\sum_{s=1}^{|\mathcal{S}|}\lambda_{s,k}^{(\mathbf{B})}\left(\sum_{o=1}^{|\Omega|}\mathbf{B}_k(s,o)-1\right)\end{aligned}$$

$$\begin{aligned}F_4\left(\boldsymbol{\pi},\mathbf{A},\mathbf{B},\mathbf{C}\right)&=\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\log\left(\sum_{k'=1}^{K}c_{i,k'}\boldsymbol{\pi}_{k'}(s')\right)\gamma_{i,1}^{(m)}(s)\\&+\sum_{s'=1}^{|\mathcal{S}|}\sum_{s''=1}^{S}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{A}_k\left(s',s''\right)\right)\xi_{i,t}^{(m)}(s,s')\\&+\sum_{s'=1}^{|\mathcal{S}|}\sum_{i=1}^{N}\sum_{t=1}^{T}\log\left(\sum_{k=1}^{K}c_{i,k}\mathbf{B}_k(s',o)\right)\gamma_{i,t}^{(m)}(s)I(o_{i,t}=o)-\mu\,\mathrm{Tr}\left(\mathbf{C}^T\mathbf{L}\mathbf{C}\right)\end{aligned}$$

The goal is simplified to prove that $F_1(\boldsymbol{\pi},\mathbf{C})$, $F_2(\mathbf{A},\mathbf{C})$, $F_3(\mathbf{B},\mathbf{C})$, $F_4(\boldsymbol{\pi},\mathbf{A},\mathbf{B},\mathbf{C})$ are all monotonically nonincreasing under the updating steps.

A function $G(x', x)$ is an auxiliary function of $L(x)$ if $G(x', x) \leq L(x)$, $G(x, x) = L(x)$. ' By constructing $G(x', x)$, we define

$$x^{(m+1)} = \arg\max_{x} G\left(x, x^{(m)}\right) \tag{B.20}$$

Thus, we have $L\left(x^{(m)}\right) = G\left(x^{(m)}, x^{(m)}\right) \leq G\left(x^{(m+1)}, x^{(m)}\right) \leq L\left(x^{(m+1)}\right)$. This leads to the monotonicity of L under the iterative updating rules. We will introduce 2 lemmas first.

Lemma B.2. *For any positive variables x_k, $\log\left(\sum_k x_k\right) \geq \sum_k q_k \log(x_k/q_k)$, with $\sum_k q_k = 1, q_k \geq 0$.*

Lemma B.3. *For any $x > 0$, $\log(x) + 1 \leq x$.*

We can construct the auxiliary functions for $L_{ks}(\boldsymbol{\pi})$, $L_{k,s_1,s_2}(\mathbf{A})$ and $L_{k,s,o}(\mathbf{B})$. Using Lemma B.2, we have

$$\log\left[\sum_k c_{ik}\boldsymbol{\pi}_k(s)\right] \geq \sum_k \frac{c_{ik}^{(m)}\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)}\boldsymbol{\pi}_{k'}^{(m)}(s)}\left[\log\left(c_{ik}\boldsymbol{\pi}_k(s)\right) - \log\left(\frac{c_{ik}^{(m)}\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)}\boldsymbol{\pi}_{k'}^{(m)}(s)}\right)\right],$$
$$\forall i, \forall s$$

$$\log\left[\sum_k c_{ik}\mathbf{A}_k(s,s')\right] \geq \sum_k \frac{c_{ik}^{(m)}\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'} c_{ik'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}\left[\log\left(c_{ik}\mathbf{A}_k(s,s')\right) - \log\left(\frac{c_{ik}^{(m)}\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'} c_{ik'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}\right)\right],$$
$$\forall i, \forall s, \forall s'$$

$$\log\left[\sum_k c_{ik}\mathbf{B}_k(s,o)\right] \geq \sum_k \frac{c_{ik}^{(m)}\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'} c_{ik'}^{(m)}\mathbf{B}_{k'}^{(m)}(s,o)}\left[\log\left(c_{ik}\mathbf{B}_k(s,o)\right) - \log\left(\frac{c_{ik}^{(m)}\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'} c_{ik'}^{(m)}\mathbf{B}_{k'}^{(m)}(s,o)}\right)\right],$$
$$\forall i, \forall s, \forall o$$

The equalities are achieved if and only if $\boldsymbol{\pi}_k(s) = \boldsymbol{\pi}_k^{(m)}(s)$, $\mathbf{A}_k(s,s') = \mathbf{A}_k^{(m)}(s,s')$, $\mathbf{B}_k(s,o) = \mathbf{B}_k^{(m)}(s,o)$.

The auxiliary functions for F_1, F_2 and F_3 are

$$G_1\left(\boldsymbol{\pi}_k(s), \boldsymbol{\pi}_k^{(m)}(s), \mathbf{C}^{(m)}\right) = \sum_{i=1}^{N} \gamma_{i,1}^{(m)}(s) \frac{c_{ik}^{(m)}\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)}\boldsymbol{\pi}_{k'}^{(m)}(s)}\left[\log\left(c_{ik}\boldsymbol{\pi}_k(s)\right) - \log\left(\frac{c_{ik}^{(m)}\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)}\boldsymbol{\pi}_{k'}^{(m)}(s)}\right)\right] - \lambda_k^{(\boldsymbol{\pi})}\boldsymbol{\pi}_k(s) + H_1, \forall s, \forall k \tag{B.21}$$

$$G_2\left(\mathbf{A}_k(s,s'),\mathbf{A}_k^{(m)}(s,s'),\mathbf{C}^{(m)}\right)=\sum_{i=1}^{N}\sum_{t=1}^{T-1}\xi_{i,t}^{(m)}(s,s')\frac{c_{ik}^{(m)}\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'}c_{ik'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}\Bigg[\log(c_{ik}\mathbf{A}_k(s,s'))-\log\left(\frac{c_{ik}^{(m)}\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'}c_{ik'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}\right)\Bigg]-\lambda_k^{(\mathbf{A})}\mathbf{A}_k(s,s')+H_2,\forall s,\forall s',\forall k \tag{B.22}$$

$$G_3\left(\mathbf{B}_k(s,o),\mathbf{B}_k^{(m)}(s,o),\mathbf{C}^{(m)}\right)=\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{i,t}^{(m)}(s)I(o_{i,t}=o)\frac{c_{ik}^{(m)}\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'}c_{ik'}^{(m)}\mathbf{B}_{k'}^{(m)}(s,o)}\left[\log(c_{ik}\mathbf{B}_k(s,o))-\log\left(\frac{c_{ik}^{(m)}\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'}c_{ik'}^{(m)}\mathbf{B}_{k'}^{(m)}(s,o)}\right)\right]-\lambda_k^{(\mathbf{B})}\mathbf{B}_k(s,o)+H_3,\forall s,\forall o,\forall k \tag{B.23}$$

where H_1, H_2 and H_3 represent the parts unrelated to $\boldsymbol{\pi}_k(s)$, $\mathbf{A}_k(s,s')$ and $\mathbf{B}_k(s,o)$ in F_1, F_2 and F_3 respectively.

Letting the partial deviations of G_1, G_2 and G_3 with respect to $\pi_k(s)$, $A_k(s,s')$ and $B_k(s,o)$ to be zeros, we have

$$\frac{\partial G_1}{\partial\boldsymbol{\pi}_k(s)}=\frac{1}{\boldsymbol{\pi}_k(s)}\sum_{i=1}^{N}\gamma_{i,1}^{(m)}(s)\frac{c_{ik}^{(m)}\boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'}c_{ik'}^{(m)}\boldsymbol{\pi}_{k'}^{(m)}(s)}-\lambda_k^{(\boldsymbol{\pi})}=0$$

$$\frac{\partial G_2}{\partial\mathbf{A}_k(s,s')}=\frac{1}{\mathbf{A}_k(s,s')}\sum_{i=1}^{N}\sum_{t=1}^{T-1}\xi_{i,t}^{(m)}(s,s')\frac{c_{ik}^{(m)}\mathbf{A}_k^{(m)}(s,s')}{\sum_{k'}c_{ik'}^{(m)}\mathbf{A}_{k'}^{(m)}(s,s')}-\lambda_k^{(\mathbf{A})}=0$$

$$\frac{\partial G_3}{\partial\mathbf{B}_k(s,o)}=\frac{1}{\mathbf{B}_k(s,o)}\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma_{i,t}^{(m)}(s)I(o_{i,t}=o)\frac{c_{ik}^{(m)}\mathbf{B}_k^{(m)}(s,o)}{\sum_{k'}c_{ik'}^{(m)}\mathbf{B}_{k'}^{(m)}(s,o)}-\lambda_k^{(\mathbf{B})}=0$$

which lead to the solutions of $\boldsymbol{\pi}_k^{(m+1)}(s)$, $\mathbf{A}_k^{(m+1)}(s,s')$ and $\mathbf{B}_k^{(m+1)}(s,o)$ that maximize F_1, F_2 and F_3

$$\boldsymbol{\pi}_k^{(m+1)}(s) = \frac{\sum_{i=1}^{N} \gamma_{i,1}^{(m)}(s) \frac{c_{ik}^{(m)} \boldsymbol{\pi}_k^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)} \boldsymbol{\pi}_{k'}^{(m)}(s)}}{\lambda_k^{(\boldsymbol{\pi})}}, \tag{B.24}$$

$$\mathbf{A}_k^{(m+1)}(s,s') = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-1} \xi_{i,t}^{(m)}(s,s') \frac{c_{ik}^{(m)} \mathbf{A}_k^{(m)}(s,s')}{\sum_{k'} c_{ik'}^{(m)} \mathbf{A}_{k'}^{(m)}(s,s')}}{\lambda_k^{(\mathbf{A})}}, \tag{B.25}$$

$$\mathbf{B}_k^{(m+1)}(s,o) = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_{i,t}^{(m)}(s) I(o_{i,t}=o) \frac{c_{ik}^{(m)} \mathbf{B}_k^{(m)}(s,o)}{\sum_{k'} c_{ik'}^{(m)} \mathbf{B}_{k'}^{(m)}(s,o)}}{\lambda_k^{(\mathbf{B})}}. \tag{B.26}$$

By substituting the expression of $\lambda_k^{(\boldsymbol{\pi})}$, $\lambda_k^{(\mathbf{A})}$, $\lambda_k^{(\mathbf{B})}$, we recover the updating rules.

To find the auxiliary function for F_4 we approximate the $\mathrm{Tr}\left(\mathbf{C}^\top \mathbf{L}\mathbf{C}\right)$ term using Taylor expansion at $c_{ik}^{(m)}$

$$\begin{aligned}
\mathrm{Tr}(\mathbf{C}^\top \mathbf{L}\mathbf{C}) &= \mathrm{Tr}\left(\mathbf{C}^{(m)^T} \mathbf{L}\mathbf{C}^{(m)}\right) + 2\left(\mathbf{L}\mathbf{C}^{(m)}\right)_{ik}\left(c_{ik} - c_{ik}^{(m)}\right) + \mathbf{D}_{ii}\left(c_{ik} - c_{ik}^{(m)}\right)^2 \\
&= \mathrm{Tr}\left(\mathbf{C}^{(m)^T} \mathbf{L}\mathbf{C}^{(m)}\right) + 2\left[\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right]\left[\left(c - c_{ik}^{(m)}\right)\right. \\
&\quad \left. + \frac{\mathbf{D}_{ii}}{2\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{ik} + 2\left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}}\left(c_{ik} - c_{ik}^{(m)}\right)^2\right] \\
&\quad - 2\left[\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right] c_{ik}^{(m)}\left(\frac{c}{c_{ik}^{(m)}} - 1\right)
\end{aligned} \tag{B.27}$$

By using Lemma B.3, we have

$$\begin{aligned}
\mathrm{Tr}(\mathbf{C}^\top \mathbf{L}\mathbf{C}) &\leq \mathrm{Tr}\left(\mathbf{C}^{(m)^T} \mathbf{L}\mathbf{C}^{(m)}\right) + 2\left[\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right]\left[\left(c - c_{ik}^{(m)}\right)\right. \\
&\quad \left. + \frac{\mathbf{D}_{ii}}{2\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{ik} + 2\left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}}\left(c_{ik} - c_{ik}^{(m)}\right)^2\right] \\
&\quad - 2\left[\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right] c_{ik}^{(m)} \log\left(\frac{c}{c_{ik}^{(m)}}\right)
\end{aligned}$$

The equality is achieved if and only if $c = c_{ik}^{(m)}$. Thus we get the auxiliary function for F_4 as

$$
\begin{aligned}
G_4(c_{ik}, \boldsymbol{\pi}, \mathbf{A}, \mathbf{B}, \mathbf{C}) = G_1 + G_2 + G_3 - \mu \operatorname{Tr}\left(\mathbf{C}^{(m)^T}\mathbf{L}\mathbf{C}^{(m)}\right) \\
- 2\mu\left[\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right]\left[\left(c_{ik} - c_{ik}^{(m)}\right)\right. \\
\left. + \frac{\mathbf{D}_{ii}}{2\left(\mathbf{D}\mathbf{C}^{(m)}\right)_{ik} + 2\left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}}\left(c_{ik} - c_{ik}^{(m)}\right)^2\right] \\
+ 2\mu\left[\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right] c_{ik}^{(m)} \log\left(\frac{c_{ik}}{c_{ik}^{(m)}}\right) - \lambda_i^{(\mathbf{C})} c_{ik} + H_4, \\
\forall i, \forall k
\end{aligned}
$$

Letting the partial deviation of G_4 with respect to c_{ik} to be zero, we have

$$
\begin{aligned}
\frac{\partial G_4}{\partial c_{ik}} = \frac{1}{c_{ik}} \sum_{s=1}^{|\mathcal{S}|} \gamma_{i,1}^{(m)}(s) \frac{c_{ik}^{(m)} \boldsymbol{\pi}_k^{(m+1)}(s)}{\sum_{k'} c_{ik'}^{(m)} \boldsymbol{\pi}_{k'}^{(m+1)}(s)} + \frac{1}{c_{ik}} \sum_{s=1}^{|\mathcal{S}|} \sum_{s'=1}^{|\mathcal{S}|} \sum_{t=1}^{T-1} \xi_{i,t}^{(m)}(s, s') \frac{c_{ik}^{(m)} \mathbf{A}_k^{(m+1)}(s, s')}{\sum_{k'} c_{ik'}^{(m)} \mathbf{A}_{k'}^{(m+1)}(s, s')} \\
+ \frac{1}{c_{ik}} \sum_{s=1}^{|\mathcal{S}|} \sum_{t=1}^{T} \gamma_{i,t}^{(m)}(s) \frac{c_{ik}^{(m)} \mathbf{B}_k^{(m+1)}(s, o_{i,t})}{\sum_{k'} c_{ik'}^{(m)} \mathbf{B}_{k'}^{(m+1)}(s, o_{i,t})} + \frac{2\mu}{c_{ik}} c_{ik}^{(m)} \left[\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right] \\
- \lambda_i^{(\mathbf{C})} = 0 \qquad \text{(B.28)}
\end{aligned}
$$

We can obtain the solution of $c_{ik}^{(m+1)}$ that maximize the auxiliary function G_4 as

$$
\begin{aligned}
c_{ik}^{(m+1)} = \frac{c_{ik}^{(m)}}{\eta} \left(\sum_{s=1}^{|\mathcal{S}|} \frac{\boldsymbol{\pi}_k^{(m+1)}(s) \gamma_{i,1}^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)} \boldsymbol{\pi}_{k'}^{(m+1)}(s)} + \sum_{s=1}^{|\mathcal{S}|} \sum_{s'=1}^{|\mathcal{S}|} \sum_{t=1}^{T-1} \frac{\mathbf{A}_k^{(m+1)}(s, s') \xi_{i,t}^{(m)}(s, s')}{\sum_{k'} c_{ik'}^{(m)} \mathbf{A}_{k'}^{(m+1)}(s, s')} \right. \\
\left. + \sum_{s=1}^{|\mathcal{S}|} \sum_{t=1}^{T} \frac{\mathbf{B}_k^{(m+1)}(s, o_{i,t}) \gamma_{i,t}^{(m)}(s)}{\sum_{k'} c_{ik'}^{(m)} \mathbf{B}_{k'}^{(m+1)}(s, o_{i,t})} + 2\mu \left[\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)T}\right] \right) \qquad \text{(B.29)}
\end{aligned}
$$

where $\eta = \lambda_i^{(\mathbf{C})} + 2\mu \left[\left(\mathbf{W}\mathbf{C}^{(m)}\right)_{ik} + \left(\mathbf{D}\mathbf{C}^{(m)}\right)_i \mathbf{C}_i^{(m)\top}\right]$. By substituting the expression of $\lambda_i^{(\mathbf{C})}$, we recover the updating rules.

B.4 Incremental Pruning

In the POMDP, the agent maintains a belief b over states, which can be updated using the Bayes' rule when new observation is obtained. The value functions are piecewise linear and

convex in the finite-horizon setting, and can be defined using a set of vectors $\{\alpha^k\}$ [130]. The value of a belief b is given by $V_t(b) = \max_k b \cdot \alpha_t^k$, and $V_{t+1}(b)$ can be computed using the Bellman backup operator H_{POMDP},

$$H_{\text{POMDP}} V_t = \bigcup_{a \in \mathcal{A}} \Gamma_t(a), \text{ where } \Gamma_t(a) = \bigoplus_{o \in \Omega} \Gamma_t(a, o), \forall a \in \mathcal{A}, \text{ and} \tag{B.30}$$

$$\Gamma_t(a, o) = \left\{ \frac{r}{|\Omega|} + \gamma \sum_{s' \in \mathcal{S}} \mathbf{A}(s, a, s') \mathbf{B}(s', o) \alpha_t^k \middle| 1 \leq k \leq |V_t| \right\}$$

The operator $\oplus$ denotes the cross sum operator. For two sets U and V the operator can be defined as $U \oplus V = \{u + v | u \in U, v \in V\}$. When computing $H_{\text{POMDP}} V_t$, it contains more vectors than necessary if there are vectors which are never the value-maximizing vector for a given belief b. A pruning subroutine `prune` can be executed after computing each cross sum. The resulting algorithm is known as incremental pruning Cassandra et al. [73] and computes a Bellman backup as $H_{\text{POMDP}} V_t = \texttt{prune}\left(\bigcup_{a \in A} \Gamma_t(a)\right)$, where $\Gamma_t(a) = \texttt{prune}(\texttt{prune}(\tilde{\Gamma}_t(a, 1) \oplus \tilde{\Gamma}_t(a, 2)) \cdots \oplus \tilde{\Gamma}_t(a, |\Omega|))$ and $\tilde{\Gamma}_t(a, o) = \texttt{prune}(\Gamma_t(a, o))$. The pruning operator can be implemented using a series of linear programs (LPs). Algorithm B.1 shows a pruning algorithm proposed by White and Lark White [131]. The procedure `BestVector` returns the vector from W with the highest value in belief b [132]. The procedure `FindBeliefStd` uses an LP to find the belief in which the value function U improves the most when adding vector w. The procedure is shown in Algorithm B.2. The bottle neck of the procedure `FindBeliefStd` is the exponentially increasing number of constraints in the LP. Several traditional methods for solving large-scale LP, such as Bender's decomposition [133] and interior points method, can be applied to accelerated the procedure `FindBeliefStd` and the entire incremental pruning algorithm.

Result: pruned set D

$D \leftarrow \varnothing$ **while** $W \neq \varnothing$ **do**
  $w \leftarrow$ arbitrary element in W;
  if $w(s) \leq u(s), \exists u \in D, \forall s \in S$ **then**
    $W \leftarrow W \backslash \{w\}$
  else
    if $b = \phi$ **then**
      $W \leftarrow W \backslash \{w\}$
    else
      $w \leftarrow$ `BestVector`(b, W);
      $D \leftarrow D \cup \{w\}, W \leftarrow W \backslash \{w\}$
    end
  end
end
return D

Algorithm B.1: Lark's algorithm for pruning a set of vectors [131].

B.5 Simulation experiment details

B.5.1 Parameters in simulation experiment

The three basis models correspond to high risk of depression progression ($\mathbf{A}_1^S$, $\mathbf{B}_1^S$, $\boldsymbol{\pi}_1^S$), low risk of depression progression ($\mathbf{A}_2^S$, $\mathbf{B}_2^S$, $\boldsymbol{\pi}_2^S$), and stable condition ($\mathbf{A}_3^S$, $\mathbf{B}_2^S$, $\boldsymbol{\pi}_2^S$), respectively.

$$\mathbf{A}_1^S = \begin{bmatrix} 0.2 & 0.7 & 0.1 \\ 0 & 0.6 & 0.4 \\ 0 & 0.22 & 0.78 \end{bmatrix}, \mathbf{A}_2^S = \begin{bmatrix} 0.56 & 0.44 & 0 \\ 0.12 & 0.72 & 0.16 \\ 0 & 0.27 & 0.73 \end{bmatrix}, \mathbf{A}_3^S = \begin{bmatrix} 0.98 & 0.02 & 0 \\ 0.02 & 0.97 & 0.01 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

Data: vector set U, vector w

Result: belief state b or symbol ϕ

if $|U| = 0$ **then**

  return *arbitrary belief* b;

else

  max d

  s.t. $(w - u) \cdot b \geq d$, $\forall u \in U$;

  $\sum_{i=1}^{|S|} b_i = 1, b_i \geq 0$, $\forall i, d$ free;

  return b *if* $d > 0$ *and* ϕ *otherwise*;

end

Algorithm B.2: `FindBeliefStd` – computes the belief in which w improves U the most.

$$\mathbf{B}_1^S = \mathbf{B}_2^S = \mathbf{B}_3^S = \begin{bmatrix} 0.95 & 0.04 & 0.01 \\ 0.03 & 0.9 & 0.07 \\ 0.01 & 0.01 & 0.98 \end{bmatrix}, \boldsymbol{\pi}_1^S = \boldsymbol{\pi}_2^S = \boldsymbol{\pi}_3^S = \begin{bmatrix} 0.3 & 0.4 & 0.3 \end{bmatrix}$$

B.5.2 Membership generation for a heterogeneous population

In the simulation, we need to generate the ground truth memberships of the patients with certain group structures. For example, some patients might have heavy weight on one basis model, but some may have even weight over all the basis models. We assume there are 3 basis models in the following discussion. We design 3 multivarite normal distributions

$$F_1(\mathbf{c}) \sim N\left(0, \begin{bmatrix} \zeta^2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right), F_2(\mathbf{c}) \sim N\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & \zeta^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right), F_2(\mathbf{c}) \sim N\left(0, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \zeta^2 \end{bmatrix}\right). \tag{B.31}$$

For generating $\mathbf{c}_i$, we first randomly select one multivariate normal distribution among the three distributions and generate a random sample from the selected distribution. The result-

ing random sample is further normalized to obtain $\mathbf{c}_i$. Evidently, the larger the magnitude of ζ^2 in F_i, the more dominant the i-th element in $\mathbf{c}_i$. Note that, ζ^2 controls the significance of the subgroup structure of the population [23]. We randomly select F_i for each patient using the following procedure. First, we assume the distribution of the gender (male 50%, female 50%) and age (18-29: 20%, 30-44: 20%, 45-60: 30%, 60+: 30%) to sample patients from the population. Then for each types (age and gender) of patient, sample the index of normal distribution F in (B.31) using distributions in Table B.1. This is due to the fact that older females tend to be more depressed in the literature.

Table B.1: The distribution to sample F_i for the 8 types of patients in simulation.

	Male	Female
Age 18-29	[0.2, 0.3, 0.5]	[0.2, 0.4, 0.4]
Age 30-44	[0.3, 0.3, 0.4]	[0.3, 0.4, 0.3]
Age 45-60	[0.4, 0.4, 0.2]	[0.4, 0.5, 0.1]
Age 60+	[0.5, 0.4, 0.1]	[0.6, 0.3, 0.1]

B.5.3 Similarity matrix generation

The similarity w_{ij} between two patients i and j is represented by the correlation of the membership, $w_{ij} = \text{corr}(c_i, c_j) = \mathbf{\Sigma}_{ij}$. Following the approach in Lin et al. [35, 44, 24], we simulate a set of P covariates, $[z_1, \ldots, z_N], z_p \in \mathbb{R}^{P\times 1}$, that follow the similarity structure. We used a multivariate normal distribution with zero mean and the covariance matrix $\mathbf{\Sigma}$. Each data point simulated from the multivariate normal distribution corresponds to a covariate $z_p^S = \left[z_{1p}^S, \ldots, z_{Np}^S\right]$, that has N correlated observations. By simulating P data points from the multivariate normal distribution, we obtained a set of covariates that preserve the similarity structure between individuals. We also incorporated a certain degree of noise to

the covariance matrix and evaluate the sensitivity of proposed method under different level of noise. We finally used the simulated covariates for measuring the similarities between individuals by

$$w_{ij} = \exp\left(-\left\|z_i - z_j\right\|^2 / \sigma^2\right). \tag{B.32}$$

where σ^2 controls the level of similarity. We used $\sigma^2 = 10$ in the simulation experiment.

B.6 Treatment effect

In the treatment of chronic disease, we model the progression of the disease as a Markov chain for each treatment type. The difference in the parameters of the Markov chain represents the effect of treatment change. Consider a discrete-time Markov chain with discrete states. The state space is $S = \{s_1, s_2, \ldots s_n\}$ with n elements. The transition matrix is T with $T_{ij} = \Pr\{s_{t+1} = j | s_t = i\}$. Suppose the steady state distribution π (row vector) of this Markov chain exists, it satisfies

$$\pi T = \pi, \sum_{i=1}^{|S|} \pi_i = 1.$$

The change of steady state distribution is associated with the change in the transition matrix. In the disease progression model, we assume lower states stand for better health. Therefore a health improvement from treatment is equivalent to the steady state distribution putting higher weights on lower states.

Consider a row of transition matrix $p = T_{i\cdot} = (p_1, p_2, \ldots, p_n)$, it is the probability of transition from state s_i to some state, in other words, it is a distribution over the state space. We define the cumulative distribution as $c_i = \sum_{j=1}^{i} p_j$. We define a null state s_0 and define $c_0 = 0$. Thus c is a non-decreasing function from 0 at state s_0 to 1 at state s_n. We also have $p_i = c_i - c_{i-1}$ for valid state i.

Consider a transform function $f(s)$ for all $s \in S$, with the following properties: (1) f is convex on S; (2) $f(s_0) = 1, f(s_n) = 1, f(s) < 1$ for $s \in S \backslash \{s_0, s_n\}$. Then we transform the cumulative distribution c as

$$1 - c'(s) = f(s) \cdot (1 - c(s))$$

Then c' is associated with an improved treatment. We perform the same transition on each row of the transition matrix, see Algorithm B.3.

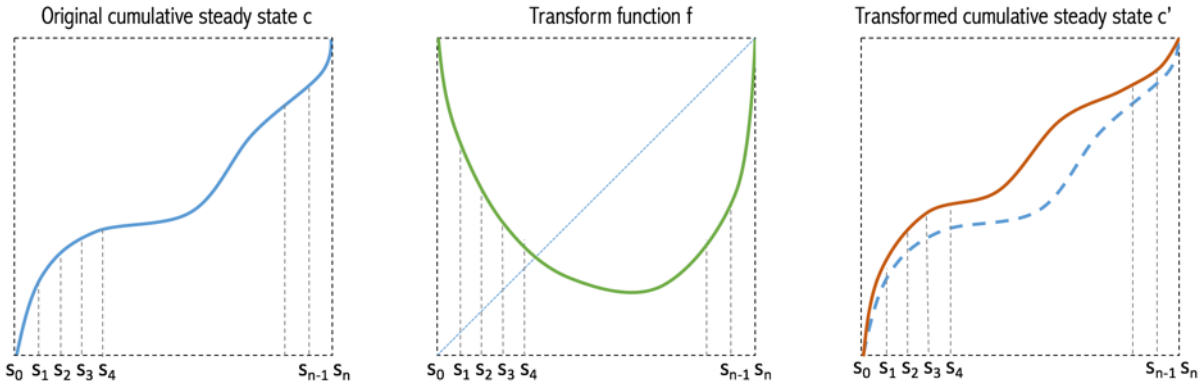


Figure B.1: Abstract examples of transformation in the transition matrix.

Data: original transition matrix T, transformation function f

for *each row of* T, *denoted* $p = (p_1, p_2, \ldots, p_n)$ **do**

  $c \leftarrow \{c_i : c_i = \sum_{j=1}^{i} p_i, i = 1, 2, \ldots, n\}$;

  $c' \leftarrow \{c'_i : c'_i = 1 - f_i(1 - c_i), i = 0, 1, \ldots, n\}$;

  $p' \leftarrow \{p' : p'_i = c'_i - c'_{i-1}, i = 0, 1, \ldots n\}$;

end

return $T' = (p'^{\top}, \ldots)^{\top}$;

Algorithm B.3: Transformation of the transition matrix with treatment effect.

It can be shown that

$$p'_1 = c'_1 = 1 - f_1 + f_1 p_1, \tag{B.33}$$

$$p'_i = c'_i - c'_{i-1} = f_{i-1} - f_i + (f_i - f_{i-1})(p_i + \ldots + p_{i-1}) + f_i p_i. \tag{B.34}$$

Therefore, for fixed f, T' is a linear transformation from T, $T' = TA + B$ with

$$A = \begin{bmatrix} f_1 & f_2 - f_1 & f_3 - f_2 & \cdots & f_n - f_{n-1} \\ 0 & f_2 & f_3 - f_2 & \cdots & f_n - f_{n-1} \\ 0 & 0 & f_3 & \cdots & f_n - f_{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & f_n \end{bmatrix}$$

and each row of B is the same,

$$B_{i\cdot} = [1 - f_1, f_1 - f_2, f_2 - f_3, \ldots, f_{n-1} - f_n]$$

The steady state distribution is the eigenvector of $T^\top$ associated with eigenvalue 1.

$$\pi = \text{eig}\left(T^\top, \lambda = 1\right) \tag{B.35}$$

$$\pi' = \text{eig}\left(T'^\top, \lambda = 1\right) = \text{eig}\left(A^\top T^\top + B^\top, \lambda = 1\right) \tag{B.36}$$

Note that an arbitrary f does not guarantee the transformed c' to be non-decreasing, or equivalently p' are non-negative.

A special case used in the disease treatment model in our paper is as follows. Consider the transform function

$$f(s_0) = 1, f(s_n) = 1, f(s) = \rho, \ \forall s \in S \backslash \{s_0, s_n\}$$

As illustrated in Figure B.2. Then we have $T' = TA + B$, with

$$A = \begin{bmatrix} \rho & \rho - 1 & 0 & \cdots & 0 & 1 - \rho \\ 0 & \rho & 0 & \cdots & 0 & 1 - \rho \\ 0 & 0 & \rho & \vdots & 0 & 1 - \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 - \rho & 0 & \cdots & 0 & \rho - 1 \\ 1 - \rho & 0 & \cdots & 0 & \rho - 1 \\ \vdots & \vdots & \ddots & 0 & \vdots \\ 1 - \rho & 0 & \cdots & 0 & \rho - 1 \end{bmatrix}$$

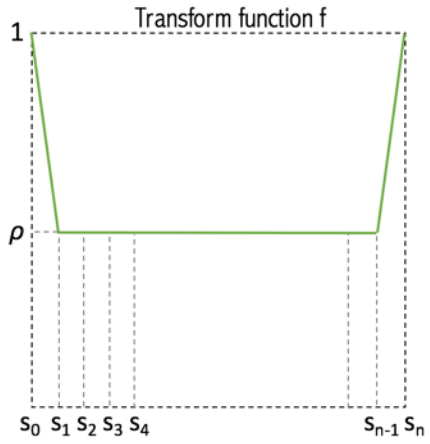


Figure B.2: The transformation of transition matrix in the disease treatment model.

We can show this is equivalent to

$$A = \begin{bmatrix} \rho & 0 & \cdots & 0 \\ 0 & \rho & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \rho \end{bmatrix}, B = \begin{bmatrix} 1-\rho & 0 & \cdots & 0 \\ 1-\rho & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho & 0 & \cdots & 0 \end{bmatrix}.$$

This transformation can be interpreted as moving a proportion $(1-\rho)$ of all the other elements to the first element in each row. This transformation guarantees p' to be nonnegative. The steady state distribution of T and T' still satisfies

$$\pi = \text{eig}(T^\top, \lambda = 1) \tag{B.37}$$

$$\pi' = \text{eig}(T'^\top, \lambda = 1) = \text{eig}(\rho T^\top + B^\top, \lambda = 1). \tag{B.38}$$

B.7 Result of Sensitivity Analysis

We perform one-way sensitivity analysis on the performance of the 10 policies. The results are in Figure 3. We use the annual discount rate of 3%, which is 0.25% for each decision epoch (1 month).

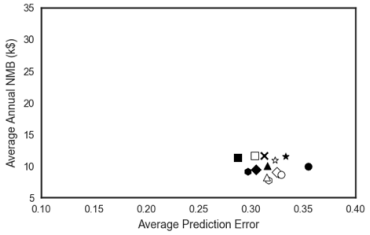
Intervention effect. In Figure B.3a, we show the policy performance under weak intervention effect $\rho = 0.8$. We can see that the reward differences between policies are much smaller. For the POMDP policy, small intervention effect results in smaller probability of choosing Treatment II and subsequently less time spent in healthier states.

Utility structure. In Figure B.3b, we show the change in the utility associated with health state, from $[\kappa(\text{H}),\kappa(\text{M}),\kappa(\text{S})] = [1.0,0.4,0.1]$ to $[1,0.8,0.6]$, i.e., the utilities of Mild and Severe Depression are higher, which indicate that we reduce the loss of utility caused by depression. We can see that the reward differences between policies are decreased. In the POMDP policies, higher utility in Mild and Severe Depression leads to lower probability of selecting Treatment II treatment, and the performance on average reward is close to that of Bayesian policies and Observation policies. The reason is that under such utility structure, the depression states do not have a significantly negative influence on the reward. Therefore, Treatment II is less attractive and treatment cost plays a bigger role in the calculation of NMB.

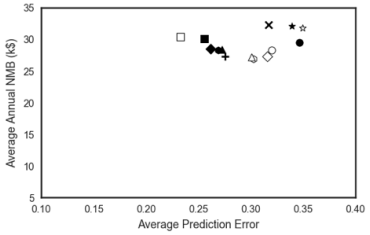
Treatment cost. In Figure B.3c, when the cost of Treatment II decreases from \$2,000 to \$1,500, there are no significant changes. In Figure B.3d the cost of Treatment II increases to \$5,000. The rewards of all policies are decreased due to increased cost in Treatment II. The gaps in rewards between different policies are also reduced.

Willingness to pay. In Figure B.3e, a smaller WTP $\lambda = \$10{,}000/\text{QALY}$ is used. We can see that the gaps in rewards between different policies are much smaller. Also, the rewards of all policies are decreased, because smaller WTP leads to smaller NMB. In Figure B.3f, a larger WTP $\lambda = \$150{,}000/\text{QALY}$ is used. In contrast to B.3e, the rewards of all policies are increased, because larger WTP leads to higher NMB.

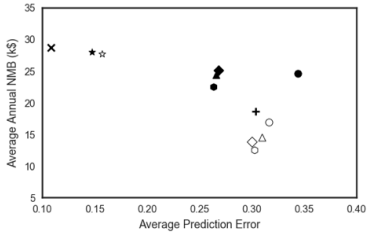
Treatment horizon in the fine-tuning stage. We perform 5 periods of Treatment I and 5 periods of Treatment II on each testing patient in the fine-tuning stage. We can see that for all policies, the rewards and the prediction accuracy are reduced. This is because the personalized models are not well trained in the fine-tuning stage due to fewer observation data.



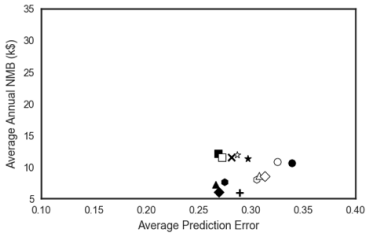
(a) Weak intervention effect: $\rho = 0.8$.



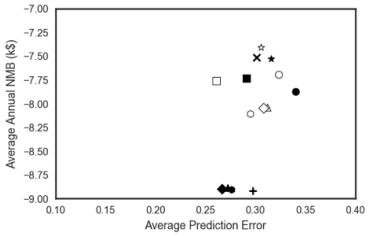
(b) Higher utility in M/S: u(H,M,S) = [1, 0.8, 0.6].



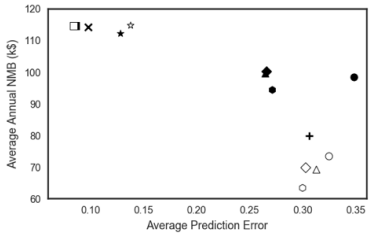
(c) Lower Treatment II cost: c(II) = \$1,500.



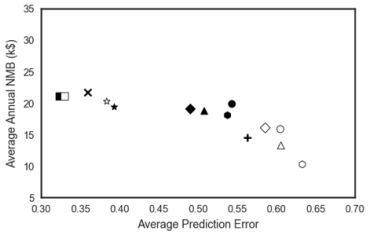
(d) Higher Treatment II cost: c(II) = \$5,000



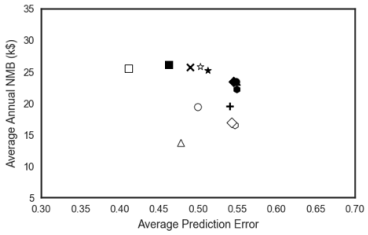
(e) Smaller WTP: λ = \$10,000/QALY.



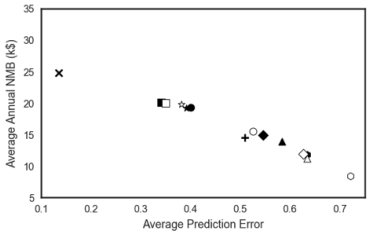
(f) Larger WTP: λ = \$150,000/QALY.



(g) Shorter treatment horizon in the fine-tuning stage.



(h) The number of health states is 5.



(i) The true canonical matrices are less extreme.

Figure B.3: Sensitivity analysis for policies comparison. Each subfigure has only one parameter changed.

Number of health states. We now assume depression has 5 severity states. We use the utility structure of κ(H,M1,M2,M3,S)= $[1, 0.5, 0.4, 0.3, 0.1]$. The rewards of all policies are increased, and their prediction accuracy is reduced.

A different set of true basis parameters. We adjust the true basis transition and emission matrices to make the treatment effect on the three basis groups less distinguishable. The results show that the prediction accuracy of POMDP policies is significantly better than others.

In the above experiments, the performance ranking of policies in the prediction accuracy changes under different settings, while the ranking in the rewards varies much less. This is because in the treatment plan, a higher NMB may appear even with ”bad” estimation since the action space is small.

Appendix C

APPENDIX FOR CHAPTER 4

C.1 Proofs

C.1.1 Proof of Lemma 4.1

Proof. First consider the case of $N = 2$. The objective function is $z = \theta_1 x_1 + \theta_2 x_2$ where $\theta_1 \geq \theta_2 > 0$. The feasible set is the line segement connecting the points $(\underline{x}_1, \overline{x}_2)$ and $(\overline{x}_1, \underline{x}_2)$ in the (x_1, x_2) space. The optimal solution is $z^\star = \theta_1 \overline{x}_1 + \theta_2 \underline{x}_2$ with $(x_1^\star, x_2^\star) = (\overline{x}_1, \underline{x}_2)$. For any other point in the feasible set $(x_1', x_2') = (1-\lambda)(\underline{x}_1, \overline{x}_2) + \lambda(\overline{x}_1, \underline{x}_2)$ where $0 \leq \lambda < 1$. Then

$$\begin{aligned} z' - z^\star &= \theta_1((1-\lambda)\underline{x}_1 + \lambda\overline{x}_1) + \theta_2((1-\lambda)\overline{x}_2 + \lambda\underline{x}_2) - (\theta_1\overline{x}_1 + \theta_2\underline{x}_2) \\ &= \theta_1(1-\lambda)(\underline{x}_1 - \overline{x}_1) + \theta_2(1-\lambda)(\overline{x}_2 - \underline{x}_2) \\ &= -(1-\lambda)[\theta_1(\overline{x}_1 - \underline{x}_1) - \theta_2(\overline{x}_2 - \underline{x}_2)] \\ &= -(1-\lambda)(\theta_1 - \theta_2)(\overline{x}_1 - \underline{x}_1) \leq 0. \end{aligned}$$

The last line holds since $\overline{x}_1 + \underline{x}_2 = \underline{x}_1 + \overline{x}_2$ and $\theta_1 \geq \theta_2 > 0$.

For $N = 3$, the optimal solution is $(x_1^\star, x_2^\star, x_3^\star) = (\overline{x}_1, 1 - \overline{x}_1 - \underline{x}_3, \underline{x}_3)$. For any other feasible point (x_1', x_2', x_3') with $x_1' < x_1^\star, x_3' > x_3^\star$, we have

$$\begin{aligned} z^\star - z' &= \theta_1\overline{x}_1 + \theta_2(1 - \overline{x}_1 - \underline{x}_3) + \theta_3\underline{x}_3 - (\theta_1 x_1' + \theta_2 x_2' + \theta_3 x_3') \\ &= \theta_1(\overline{x}_1 - x_1') + \theta_2(1 - x_2 - \overline{x}_1 - \underline{x}_3) + \theta_3(\underline{x}_3 - x_3') \\ &= (\theta_1 - \theta_2)(\overline{x}_1 - x_1') + (\theta_2 - \theta_3)(x_3' - \underline{x}_3) \geq 0, \end{aligned}$$

which uses the condition that $x_1' + x_2' + x_3' = 1$ and $\theta_1 \geq \theta_2 \geq \theta_3 > 0$. This proves the optimality of $x^\star$ when $N = 3$. The optimal $x^\star$ will not change as long as the order $\theta_1 \geq \theta_2 \geq \theta_3 > 0$ preserves.

For $N > 3$, first obtain the optimal solution for x_1, x_N: $(x_1^\star, x_N^\star) = \overline{x}_1, \underline{x}_N$, due to the fact that $\theta_1 \geq \theta_i \geq \theta_N$ for $i = 2, \ldots, N-1$, then the objective function (4.21a) becomes $z = \sum_{i=2}^{N-1} \theta_i x_i$ and the constraint (4.21c) becomes $\sum_{i=2}^{N-1} x_i = 1 - x_1^\star - x_N^\star$, which is the same type of LP. Therefore, by induction, we can conclude that the optimal solution to the LP is $x_i^\star = \overline{x}_i$ for $i = 1, \ldots, \lfloor N/2 \rfloor$, and $x_i^\star = \underline{x}_i$ for $i = \lceil N/2 \rceil + 1, \ldots, N$. T he The optimality preserves when the order of the coefficeints of the objective function does not change for $N \geq 2$. □

C.1.2 Proof of Theorem 4.3

Proof. In equation (4.23) let $p = 1$ we have $\mathbf{B}(s^1, a, o^1) \geq \mathbf{B}(s^2, a, o^1) \geq \ldots \geq \mathbf{B}(s^{|\mathcal{S}|}, a, o^1)$, since $\alpha_{t+1}^{\ell}(s)$ is nonincreasing in s. Then,

$$\mathbf{B}(s^1, a, o^1)\alpha_{t+1}^{\ell(1)}(s^1) \geq \mathbf{B}(s^2, a, o^1)\alpha_{t+1}^{\ell(1)}(s^2) \geq \ldots \geq \mathbf{B}(s^1, a, o^1)\alpha_{t+1}^{\ell(1)}\left(s^{|\mathcal{S}|}\right) \tag{C.1}$$

So the case of $p = 1$ is proved. Suppose we have

$$\sum_{j=1}^{p} \mathbf{B}(s^1, a, o^j)\alpha_{t+1}^{\ell(j)}(s^1) \geq \sum_{j=1}^{p} \mathbf{B}(s^2, a, o^j)\alpha_{t+1}^{\ell(j)}(s^2) \geq \ldots \geq \sum_{j=1}^{p} \mathbf{B}(s^S, a, o^j)\alpha_{t+1}^{\ell(j)}(s^S) \tag{C.2}$$

Then

$$\begin{aligned}
\sum_{j=1}^{p+1} \mathbf{B}(s^i, a, o^j)\alpha_{t+1}^{\ell(j)}(s^i) &= \sum_{j=1}^{p} \mathbf{B}(s^i, a, o^j)\alpha_{t+1}^{\ell(j)}(s^i) + \mathbf{B}(s^i, a, o^{p+1})\alpha_{t+1}^{\ell(p+1)}(s^i) \\
&\geq \sum_{j=1}^{p} \mathbf{B}(s^{i+1}, a, o^j)\alpha_{t+1}^{\ell(j)}(s^{i+1}) + \mathbf{B}(s^{i+1}, a, o^{p+1})\alpha_{t+1}^{\ell(p+1)}(s^{i+1}) \\
&= \sum_{j=1}^{p+1} \mathbf{B}(s^{i+1}, a, o^j)\alpha_{t+1}^{\ell(j)}(s^{i+1})
\end{aligned}$$

for all $i = 1, \ldots, |\mathcal{S}| - 1$. Then by induction,

$$\sum_{j=1}^{p} \mathbf{B}(s^1, a, o^j)\alpha_{t+1}^{\ell(j)}(s^1) \geq \sum_{j=1}^{p} \mathbf{B}(s^2, a, o^j)\alpha_{t+1}^{\ell(j)}(s^2) \geq \ldots \geq \sum_{j=1}^{p} \mathbf{B}(s^S, a, o^j)\alpha_{t+1}^{\ell(j)}(s^S) \tag{C.3}$$

for all $p = 1, \ldots, |\Omega|$. In other words, $\theta(s, a, s^i)$ is nonincreasing in i. For the first iteration, we have $\Gamma_T^\star = \langle g \rangle$ and $g(s^i)$ is nonincreasing in i. Therefore the initial condition (C.1) is satisfied and the conclusion is proved for the first iterations.

Since the optimal model of $\alpha(s,a)$ is the same for all s, therefore, we have

$$\sum_{s' \in \mathcal{S}} \theta(s,a,s') \mathbf{A}^{\star}(s,a,s') \tag{C.4}$$

is nonincreasing in s. Since $R(s,a)$ nonincreasing in s for all $a \in \mathcal{A}$,

$$\alpha(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \theta(s,a,s') \mathbf{A}(s,a,s') \tag{C.5}$$

is nonincreasing in s, for all a. □

BIBLIOGRAPHY

[1] NIMH. National institute of mental health: Depression. `https://www.nimh.nih.gov/health/topics/depression/index.shtml`. Accessed: 2016-04-02.

[2] Bruce Arroll, Weng-yee Chin, Waldron Martis, Felicity Goodyear-Smith, Vicki Mount, Douglas Kingsford, Stephen Humm, Grant Blashki, and Stephen MacGillivray. Antidepressants for treatment of depression in primary care: a systematic review and meta-analysis. *Journal of primary health care*, 8(4):325–334, 2016.

[3] Oguzhan Alagoz, Heather Hsu, Andrew J Schaefer, and Mark S Roberts. Markov decision processes: a tool for sequential decision making under uncertainty. *Medical Decision Making*, 30(4):474–483, 2010. ISSN 0272-989X.

[4] Jingyu Zhang, Brian T Denton, Hari Balasubramanian, Nilay D Shah, and Brant A Inman. Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 14(4):529–547, 2012. ISSN 1523-4614.

[5] Brian T Denton, Murat Kurt, Nilay D Shah, Sandra C Bryant, and Steven A Smith. Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making*, 29(3):351–367, 2009. ISSN 0272-989X.

[6] Brian T Denton, Murat Kurt, Nilay D Shah, Sandra C Bryant, and Steven A Smith. Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making*, 29(3):351–367, 2009. ISSN 0272-989X.

[7] Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. *Modeling medical treatment using Markov decision processes*, pages 593–612. Springer, 2005.

[8] Gregory E Simon, Carolyn M Rutter, Do Peterson, Malia Oliver, Ursula Whiteside, Belinda Operskalski, and Evette J Ludman. Does response on the phq-9 depression questionnaire predict subsequent suicide attempt or suicide death? *Psychiatric Services*, 64(12):1195–1202, 2013. ISSN 1075-2730.

[9] JD Ribeiro, JC Franklin, Kathryn Rebecca Fox, KH Bentley, Evan M Kleiman, BP Chang, and Matthew K Nock. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychological Medicine*, 46(2):225–236, 2016. ISSN 0033-2917.

[10] Joseph C Franklin, Jessica D Ribeiro, Kathryn R Fox, Kate H Bentley, Evan M Kleiman, Xieyining Huang, Katherine M Musacchio, Adam C Jaroszewski, Bernard P Chang, and Matthew K Nock. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143(2):187, 2017. ISSN 1939-1455.

[11] Vikrant Mittal, Walter A Brown, and Edward Shorter. Are patients with depression at heightened risk of suicide as they begin to recover? *Psychiatric services*, 60(3):384–386, 2009. ISSN 1075-2730.

[12] Thomas E. Ellis, Kelly L. Green, Jon G. Allen, David A. Jobes, and Michael R. Nadorff. Collaborative assessment and management of suicidality in an inpatient setting: Results of a pilot study. *Psychotherapy (Chicago, Ill.)*, 49(1):72–80, 2012. ISSN 0033-3204 1939-1536. doi: 10.1037/a0026746. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3752846/`.

[13] Jane Gunn, Peter Elliott, Konstancja Densley, Aves Middleton, Gilles Ambresin, Christopher Dowrick, Helen Herrman, Kelsey Hegarty, Gail Gilchrist, and Frances Griffiths. A trajectory-based approach to understand the factors associated with persistent depressive symptoms in primary care. *Journal of affective disorders*, 148(2):338–346, 2013. ISSN 0165-0327.

[14] Kurt Kroenke and Robert L Spitzer. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515, 2002. ISSN 0048-5713.

[15] K Bruce Bayley, Tom Belnap, Lucy Savitz, Andrew L Masica, Nilay Shah, and Neil S Fleming. Challenges in using electronic health record data for cer: experience of 4 learning organizations and solutions applied. *Medical care*, 51:S80–S86, 2013. ISSN 0025-7079.

[16] Michael G Kahn, Marsha A Raebel, Jason M Glanz, Karen Riedlinger, and John F Steiner. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Medical care*, 50, 2012.

[17] Katherine L Musliner, Trine Munk-Olsen, William W Eaton, and Peter P Zandi. Hetero-

geneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. *Journal of affective disorders*, 192:199–211, 2016. ISSN 0165-0327.

[18] Christopher K Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

[19] Olof Jacobson and Hercules Dalianis. Applying deep learning on electronic health records in swedish to predict healthcare-associated infections. *ACL 2016*, page 191, 2016.

[20] Hanna M Van Loo, Peter De Jonge, Jan-Willem Romeijn, Ronald C Kessler, and Robert A Schoevers. Data-driven subtypes of major depressive disorder: a systematic review. *BMC medicine*, 10(1):156, 2012. ISSN 1741-7015.

[21] MHRN. Mental health research network. `http://hcsrn.org/mhrn/en/`. Accessed: 2016-04-02.

[22] Gregory E Simon, Eric Johnson, Jean M Lawrence, Rebecca C Rossom, Brian Ahmedani, Frances L Lynch, Arne Beck, Beth Waitzfelder, Rebecca Ziebell, Robert B Penfold, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, 175(10):951–960, 2018.

[23] Ying Lin, Shuai Huang, Gregory E Simon, and Shan Liu. Analysis of depression trajectory patterns using collaborative learning. *Mathematical Biosciences*, 282:191–203, 2016. ISSN 0025-5564.

[24] Ying Lin, Shan Liu, and Shuai Huang. Selective sensing of a heterogeneous population of units with dynamic health conditions. *IISE Transactions*, 50(12):1076–1088, 2018.

[25] Kira Rehfeld, Norbert Marwan, Jobst Heitzig, and Jürgen Kurths. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, 18(3):389–404, 2011. ISSN 1023-5809.

[26] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PLOS ONE*, 8(6):1–13, 06 2013. doi: 10.1371/journal.pone.0066341. URL `https://doi.org/10.1371/journal.pone.0066341`.

[27] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

http://www.deeplearningbook.org.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016. ISBN 0262337371.

[29] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6):e66341, 2013. ISSN 1932-6203.

[30] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012. ISBN 1118585771.

[31] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. ISSN 0018-9448.

[32] Wayne W Daniel. The spearman rank correlation coefficient. *Biostatistics: A Foundation for Analysis in the Health Sciences*, 1987.

[33] Kevin Swingler. *Applying neural networks: a practical guide*. Morgan Kaufmann, 1996. ISBN 0126791708.

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[35] Ying Lin, Shuai Huang, Gregory E Simon, and Shan Liu. Analysis of depression trajectory patterns using collaborative learning. *Mathematical biosciences*, 282:191–203, 2016. ISSN 0025-5564.

[36] Robert J. Valuck, Heather O. Anderson, Anne M. Libby, Elias Brandt, Cathy Bryan, Richard R. Allen, Elizabeth W. Staton, David R. West, and Wilson D. Pace. Enhancing electronic health record measurement of depression severity and suicide ideation: A distributed ambulatory research in therapeutics network (dartnet) study. *The Journal of the American Board of Family Medicine*, 25(5):582–593, 2012. ISSN 1557-2625. doi: 10.3122/jabfm.2012.05.110053.

[37] Qiu-Yue Zhong, Elizabeth W. Karlson, Bizu Gelaye, Sean Finan, Paul Avillach, Jordan W. Smoller, Tianxi Cai, and Michelle A. Williams. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by nat-

ural language processing. *BMC Medical Informatics and Decision Making*, 18(1):30, May 2018. ISSN 1472-6947. doi: 10.1186/s12911-018-0617-7. URL https://doi.org/10.1186/s12911-018-0617-7.

[38] Evelien Snippe, Elisabeth H Bos, Karen M van der Ploeg, Robbert Sanderman, Joke Fleer, and Maya J Schroevers. Time-series analysis of daily changes in mindfulness, repetitive thinking, and depressive symptoms during mindfulness-based treatment. *Mindfulness*, 6(5): 1053–1062, 2015. ISSN 1868-8527.

[39] Heather D Anderson, Wilson D Pace, Elias Brandt, Rodney D Nielsen, Richard R Allen, Anne M Libby, David R West, and Robert J Valuck. Monitoring suicidal patients in primary care using electronic health records. *J Am Board Fam Med*, 28(1):65–71, 2015.

[40] Brian K Ahmedani, Edward L Peterson, Yong Hu, Rebecca C Rossom, Frances Lynch, Christine Y Lu, Beth E Waitzfelder, Ashli A Owen-Smith, Samuel Hubley, and Deepak Prabhakar. Major physical health conditions and risk of suicide. *American Journal of Preventive Medicine*, 2017. ISSN 0749-3797.

[41] Michelle Campbell, Ray Fitzpatrick, Andrew Haines, Ann Louise Kinmonth, Peter Sandercock, David Spiegelhalter, and Peter Tyrer. Framework for design and evaluation of complex interventions to improve health. *BMJ: British Medical Journal*, 321(7262):694, 2000.

[42] Kenneth Dickstein, Alain Cohen-Solal, Gerasimos Filippatos, John JV McMurray, Piotr Ponikowski, Philip Alexander Poole-Wilson, Anna Strömberg, Dirk J Veldhuisen, Dan Atar, and Arno W Hoes. Esc guidelines for the diagnosis and treatment of acute and chronic heart failure 2008. *European journal of heart failure*, 10(10):933–989, 2008. ISSN 1879-0844.

[43] Jos Twisk and Trynke Hoekstra. Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *Journal of clinical epidemiology*, 65(10):1078–1087, 2012.

[44] Ying Lin, Kaibo Liu, Eunshin Byon, Xiaoning Qian, Shan Liu, and Shuai Huang. A collaborative learning framework for estimating many individualized regression models in a heterogeneous population. *IEEE Transactions on Reliability*, 67(1):328–341, 2018. ISSN 0018-9529.

[45] Mingdi You, Bingjie Liu, Eunshin Byon, Shuai Huang, and Jionghua Judy Jin. Direction-

dependent power curve modeling for multiple interacting wind turbines. *IEEE Transactions on power systems*, 33(2):1725–1733, 2018. ISSN 0885-8950.

[46] Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. Modeling medical treatment using markov decision processes. In *Operations research and health care*, pages 593–612. Springer, 2005.

[47] Robert D Vincent, Joelle Pineau, Norma Ybarra, and Issam El Naqa. Chapter 16: Practical reinforcement learning in dynamic treatment regimes. In *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, pages 263–296. SIAM, 2015.

[48] Diana M Negoescu, Kostas Bimpikis, Margaret L Brandeau, and Dan A Iancu. Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science*, 2017.

[49] Jennifer E Mason, Darin A England, Brian T Denton, Steven A Smith, Murat Kurt, and Nilay D Shah. Optimizing statin treatment decisions for diabetes patients in the presence of uncertain future adherence. *Medical Decision Making*, 32(1):154–166, 2012. ISSN 0272-989X.

[50] Daniel M Faissol, Paul M Griffin, and Julie L Swann. Timing of testing and treatment of hepatitis c and other diseases. In *Proceedings*, page 11, 2007.

[51] S Liu, ML Brandeau, and JD Goldhaber-Fiebert. Optimizing patient treatment decisions in an era of rapid technological advances: the case of hepatitis c treatment. *Health Care Manag Sci*, 20(1):16–32, 2015.

[52] O Alagoz, AJ Schaefer, LM Maillart, and MS Roberts. Determining the optimal timing of living-donor liver transplantation using a markov decision process (mdp) model. *Medical Decision Making*, 22(6):558, 2002.

[53] LA Pratt and DJ Brody. *Depression in the US household population, 2009–2012*. NCHS data brief, Hyattsville, MD, 2014.

[54] SM Shechter, MD Bailey, AJ Schaefer, and MS Roberts. The optimal time to initiate hiv therapy under ordered health states. *Operations Research*, 56(1):20–33, 2008.

[55] Lisa M Maillart, Julie Simmons Ivy, Scott Ransom, and Kathleen Diehl. Assessing dynamic

breast cancer screening policies. *Operations Research*, 56(6):1411–1427, 2008. ISSN 0030-364X.

[56] Antoine Saure, Jonathan Patrick, Scott Tyldesley, and Martin L Puterman. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research*, 223(2):573–584, 2012. ISSN 0377-2217.

[57] Jonathan E Helm, Mariel S Lavieri, Mark P Van Oyen, Joshua D Stein, and David C Musch. Dynamic forecasting and control algorithms of glaucoma progression for clinician decision support. *Operations Research*, 63(5):979–999, 2015. ISSN 0030-364X.

[58] Pooyan Kazemian, Jonathan E Helm, Mariel S Lavieri, Joshua D Stein, and Mark P Van Oyen. Dynamic monitoring and control of irreversible chronic diseases with application to glaucoma. *Production and Operations Management*, 2016.

[59] Turgay Ayer, Can Zhang, Anthony Bonifonte, Anne C Spaulding, and Jagpreet Chhatwal. Prioritizing hepatitis c treatment in us prisons. *Operations Research*, 2019.

[60] Diana M Negoescu, Kostas Bimpikis, Margaret L Brandeau, and Dan A Iancu. Dynamic learning of patient response types: An application to treating chronic diseases. *Management Science*, 2017. ISSN 0025-1909.

[61] Qiushi Chen, Turgay Ayer, and Jagpreet Chhatwal. Optimal m-switch surveillance policies for liver cancer in a hepatitis c–infected population. *Operations Research*, 2018. ISSN 0030-364X.

[62] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2015. ISBN 145032956X.

[63] Peter Schulam and Suchi Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.

[64] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034, 2012. ISSN 0030-364X.

[65] Mariel S Lavieri, Martin L Puterman, Scott Tyldesley, and William J Morris. When to

treat prostate cancer patients based on their psa dynamics. *IIE Transactions on Healthcare Systems Engineering*, 2(1):62–77, 2012. ISSN 1948-8300.

[66] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. ISSN 0035-9246.

[67] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233, 1982. ISSN 0035-9246.

[68] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*, pages svn–2017–000101, 2017. ISSN 2059-8688.

[69] Turgay Ayer. Inverse optimization for assessing emerging technologies in breast cancer screening. *Annals of operations research*, 230(1):57–85, 2015. ISSN 0254-5330.

[70] Elliot Lee, Mariel S Lavieri, Michael L Volk, and Yongcai Xu. Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity. *Health care management science*, 18(3):363–375, 2015. ISSN 1386-9620.

[71] Tommi Jaakkola, Satinder P Singh, and Michael I Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *Advances in neural information processing systems*, pages 345–352, 1995.

[72] Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On improving deep reinforcement learning for pomdps. *arXiv preprint arXiv:1804.06309*, 2018.

[73] Anthony Cassandra, Michael L Littman, and Nevin L Zhang. Incremental pruning: A simple, fast, exact method for partially observable markov decision processes. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 54–61. Morgan Kaufmann Publishers Inc., 1997. ISBN 1558604855.

[74] Lauren N Steimle and Brian T Denton. Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*, pages 189–222. Springer, 2017.

[75] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique

occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970. ISSN 0003-4851.

[76] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[77] Kurt Kroenke and Robert L Spitzer. The phq-9: a new depression diagnostic and severity measure. *Psychiatric annals*, 32(9):509–515, 2002.

[78] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. ISSN 0018-9448.

[79] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 1990. ISBN 0521386322.

[80] Cao Xiao. *Optimization and Machine Learning Methods for Medical and Healthcare Applications*. PhD thesis, Ph.D. Dissertation, University of Waashington, 2017.

[81] Karl J Astrom. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

[82] Joelle Pineau, Geoffrey Gordon, and Sebastian Thrun. Anytime point-based approximations for large pomdps. *Journal of Artificial Intelligence Research*, 27:335–380, 2006.

[83] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[84] Shie Mannor, Duncan Simester, Peng Sun, and John N Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007. ISSN 0025-1909.

[85] Chuanpu Hu, William S. Lovejoy, and Steven L. Shafer. Comparison of some suboptimal control policies in medical drug therapy. *Operations Research*, 44(5):696–709, 1996.

[86] Turgay Ayer, Oguzhan Alagoz, and Natasha K Stout. Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034, 2012. ISSN 0030-364X.

[87] Lisa M Maillart, Julie Simmons Ivy, Scott Ransom, and Kathleen Diehl. Assessing dynamic

breast cancer screening policies. *Operations Research*, 56(6):1411–1427, 2008. ISSN 0030-364X.

[88] Jagpreet Chhatwal, Oguzhan Alagoz, and Elizabeth S. Burnside. Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations Research*, 58(6):1577–1591, 2010.

[89] Nicole DeHoratius, Adam Mersereau, and Linus Schrage. Retail inventory management when records are inaccurate. *Manufacturing & Service Operations Management*, 10(2):257–277, 2008.

[90] Soroush Saghafian and Brian Tomlin. The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research*, 64(1):167–185, 2016.

[91] Yossi Aviv and Amit Pazgal. A partially observed markov decision process for dynamic pricing. *Management Science*, 51(9):1400–1416, 2005.

[92] Chelsea C White. A markov quality control process subject to partial observation. *Management Science*, 23(8):843–852, 1977.

[93] Chelsea C White. Optimal inspection and repair of a production process subject to deterioration. *Journal of the Operational Research Society*, 29(3):235–243, 1978.

[94] Chelsea C White. Optimal control-limit strategies for a partially observed replacement problem. *International Journal of Systems Science*, 10(3):321–332, 1979.

[95] Robert C. Wang. Optimal replacement policy with unobservable states. *Journal of Applied Probability*, 14(2):340–348, 1977.

[96] Lisa M. Maillart. Maintenance policies for systems with condition monitoring and obvious failures. *IIE Transactions*, 38(6):463–475, 2006.

[97] Lu Jin, Kazuhiro Kumagai, and Kazuyuki Suzuki. Control limit policy for partially observable markov decision process based on stochastic increasing ordering. *Quality Technology & Quantitative Management*, 8(4):479–493, 2011.

[98] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

[99] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

[100] Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

[101] Alvin W Drake. *Observation of a Markov process through a noisy channel*. PhD thesis, Massachusetts Institute of Technology, 1962.

[102] Edward Jay Sondik. The optimal control of partially observable markov processes. Technical report, STANFORD UNIV CALIF STANFORD ELECTRONICS LABS, 1971.

[103] Richard D Smallwood and Edward J Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

[104] Edward J Sondik. The optimal control of partially observable markov processes over the infinite horizon: Discounted costs. *Operations research*, 26(2):282–304, 1978. ISSN 0030-364X.

[105] William S Lovejoy. Some monotonicity results for partially observed markov decision processes. *Operations Research*, 35(5):736–743, 1987.

[106] Ulrich Rieder. Structural results for partially observed control models. *Zeitschrift für Operations Research*, 35(6):473–490, 1991.

[107] William S Lovejoy. On the convexity of policy regions in partially observed systems. *Operations Research*, 35(4):619–621, 1987.

[108] Arnab Nilim and Laurent El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[109] Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30 (2):257–280, 2005.

[110] Robert Givan, Sonia Leach, and Thomas Dean. Bounded-parameter markov decision processes. *Artificial Intelligence*, 122(1-2):71–109, 2000.

[111] J. Andrew (Drew) Bagnell, Andrew Y. Ng, and Jeff Schneider. Solving uncertain markov decision problems. Technical Report CMU-RI-TR-01-25, Carnegie Mellon University, Pittsburgh, PA, August 2001.

[112] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013. ISSN 0364-765X.

[113] Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010. ISSN 0030-364X.

[114] Malcolm Strens. A bayesian framework for reinforcement learning. In *ICML*, pages 943–950, 2000.

[115] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2505–2513, 2010.

[116] Shie Mannor, Ofir Mebel, and Huan Xu. Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*, 2012.

[117] Hideaki Itoh and Kiyohiko Nakamura. Partially observable markov decision processes with imprecise parameters. *Artificial Intelligence*, 171(8-9):453–490, 2007. ISSN 0004-3702.

[118] Soroush Saghafian. Ambiguous partially observable markov decision processes: Structural results and applications. *working paper*, 2018.

[119] Yaodong Ni and Zhi-Qiang Liu. Bounded-parameter partially observable markov decision processes. In *ICAPS*, pages 240–247, 2008.

[120] George E. Monahan. A survey of partially observable markov decision processes: theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982. doi: 10.1287/mnsc.28.1.1.

[121] Chelsea C White. A survey of solution techniques for the partially observed markov decision process. *Annals of Operations Research*, 32(1):215–230, 1991. ISSN 0254-5330.

[122] Zhengzhu Feng and Shlomo Zilberstein. Region-based incremental pruning for pomdps. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 146–153. AUAI Press, 2004.

[123] Serge B. Provost and Young-Ho Cheong. On the distribution of linear combinations of the components of a dirichlet random vector. *Canadian Journal of Statistics*, 28(2):417–425, 2000.

[124] J. P. Imhof. Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48(3/4):419–426, 1961.

[125] Jue Gong and Shan Liu. Partially observable collaborative model foroptimizing personalized treatment selection. *submitted to INFORMS Journal of Computing*, 2019.

[126] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010. ISSN 0030-364X.

[127] Wilhelm Kirch, editor. *Test of Homogeneity, Chi-SquareTest of homogeneity, chi-square*, pages 1386–1386. Springer Netherlands, Dordrecht, 2008. ISBN 978-1-4020-5614-7. doi: 10.1007/978-1-4020-5614-7_3475. URL https://doi.org/10.1007/978-1-4020-5614-7_3475.

[128] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. ISBN 0898716241.

[129] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[130] EJ Sondik. *The optimal control of partially observable Markov processes*. PhD thesis, Ph.D. Dissertation, Stanford University, 1971.

[131] Chelsea C White. A survey of solution techniques for the partially observed markov decision process. *Annals of Operations Research*, 32(1):215–230, 1991. ISSN 0254-5330.

[132] Michael Lederman Littman. *Algorithms for sequential decision making*. PhD thesis, Ph.D. Dissertation, Brown University, 1996.

[133] Erwin Walraven and Matthijs TJ Spaan. Accelerated vector pruning for optimal pomdp solvers. In *AAAI*, pages 3672–3678, 2017.