EPFL

Thèse n° 10008

Personalizable intervention systems to promote healthy behavior change

Présentée le 24 janvier 2020

à la Faculté informatique et communications
Laboratoire d'intelligence artificielle
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

Igor KULEV

Acceptée sur proposition du jury

Prof. K. Aberer, président du jury
Prof. B. Faltings, Dr P. Pu Faltings, directeurs de thèse
Prof. Y. Lu, rapporteuse
Dr P. Viappiani, rapporteur
Prof. V. Cevher, rapporteur

École polytechnique fédérale de Lausanne

2020

I can accept failure,
everyone fails at something.
But I can't accept not trying.
— Michael Jordan

To my family...

Acknowledgements

I am deeply grateful to my advisors, Dr. Pearl Pu and Prof. Boi Faltings, for guiding me through the PhD process. I would also like to thank the members of my thesis committee: Prof. Karl Aberer, Prof. Yuan Lu, Dr. Paolo Viappiani and Prof. Volkan Cevher, for all of their insightful comments and suggestions. I thank my collaborators from the REACH project I have been involved in, especially Dr. Thomas Linner for his support and great project management. I thank Aleksei for his valuable feedback on this thesis. I thank Diego for providing and reviewing the French version of the abstract. I thank the current and former members of the AI lab and the HCI lab for helping me in various ways and making my PhD experience enjoyable: Goran, Florent, Claudiu, Valentina, Marina, Aleksei, Onur, Fei, Diego, Panayiotis, Naman, Aris, Yubo, Kavous, Anuradha, Ekaterina, Adam, Marija, Ljubomir. I thank the Macedonian community — including Gorica, Ana P., Kristina, Nikolche, Marjan, Emil, Darko, Vase, Vladimir, Elena, Dino, Maja, Mladen, Ana M., Bojan, Jovanche, Zarko, Aleksandar, Zhivka and many others — for spending great moments together. I thank the students I worked with on interesting projects, including Wenyuan Lv, whose work was partially included in this thesis. I am especially grateful to my girlfriend Elizabeta for her love and support during the whole PhD journey. Finally, I would like to thank my parents, Sonja and Krste, and my sister Ana for always being there for me and for their unconditional love and support.

Lausanne, January 10, 2020 I. K.

Abstract

Adopting healthy behaviors can prevent the onset of many adverse health conditions. However, behavior changes are difficult to make, and often, people who like to improve their behaviors do not know how to do that. Personalizable intervention systems could assist them to achieve healthy behavior change. These systems decide what would be the optimal intervention for the target user based on his or her characteristics, including current and past behavior patterns. In this thesis, we propose novel solutions that address the main challenges in building a personalizable intervention system to promote healthy behavior change. First, we propose a system based on a Bayesian mixture model to identify subpopulations with different behavior changes from longitudinal data. This system is especially suitable when the amount of data is limited, and when there are unobserved factors that might affect behavior change. Second, we propose CLINT, a system based on a latent-variable model, to discover and predict behavior change patterns from fine-grained sensor data. The novelty of this system is that it produces interpretable patterns that could be used to suggest successful behavior change strategies from the existing users similar to the target user. Third, we propose a personalizable intervention system to improve the physical activeness of senior adults. The main novelty of this system is that it uses historical time series fitness data to decide which intervention to recommend. Finally, we propose ACFR, an adversarial approach to reduce intervention bias in observational data. This approach learns a balanced representation of the covariates that allows personalizable intervention systems to make a better estimate of the intervention effect. Our solutions turn existing human behavior data into actionable insights for future users who may have unhealthy lifestyles.

Keywords: recommender systems, time series analysis, behavior change, mixture model, preventive healthcare, user modeling

Résumé

Adopter des comportements sains peut prévenir l'apparition de nombreux problèmes de santé. Cependant, les changements de comportement sont difficiles à entreprendre et souvent, les personnes qui souhaitent améliorer leur comportement ne savent pas comment procéder. Des systèmes d'intervention personnalisables pourraient les accompagner dans une démarche d'un changement de comportement sain. Ces systèmes décident quelle serait l'intervention optimale pour l'utilisateur cible en fonction de ses caractéristiques, notamment des comportements actuels et passés. Dans cette thèse, nous proposons de nouvelles solutions qui répondent aux principaux défis de la construction d'un système d'intervention personnalisable visant à promouvoir un changement de comportement sain. Premièrement, nous proposons un système basé sur un modèle de « mixture bayésienes » pour identifier les sous-populations présentant des changements de comportement différents à partir des données longitudinales. Ce système est particulièrement adapté lorsque la quantité de données est limitée et qu'il existe des facteurs non observés, susceptibles d'influer sur le changement de comportement. Deuxièmement, nous proposons CLINT, un système basé sur un modèle à variables latentes, pour découvrir et prédire les modèles de changement de comportement à partir de données de capteurs. La nouveauté de ce système est qu'il produit des modèles interprétables qui pourraient être utilisés pour suggérer aux utilisateurs existants des stratégies de changement de comportement réussies, similaires à l'utilisateur cible. Troisièmement, nous proposons un système d'intervention personnalisable pour améliorer l'activité physique des personnes âgées. La principale nouveauté de ce système réside dans le fait qu'il utilise les données temporelles de la condition physique de l'utilisateur, pour lui recommander les interventions. Finalement, nous proposons ACFR, une approche « adversariale » visant à réduire les biais d'intervention dans les données d'observation. Cette approche apprend une représentation équilibrée des covariables, permettant aux systèmes d'intervention personnalisables de mieux estimer l'effet de l'intervention. Nos solutions transforment les données existantes sur le comportement humain en informations exploitables pour les futurs utilisateurs susceptibles d'avoir un mode de vie malsain.

Mots-clés : systèmes de recommandation, analyse de série temporelle, changement de comportement, mixture de modèles, soins de santé préventifs, modélisation de l'utilisateur

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Challenges

With the advance of technology, the amount of *personal* (or *user-specific*) data increases rapidly. For example, fitness trackers continuously monitor users' physical activities, nutrition mobile apps keep track of users' food intake, web sites capture users' browsing behavior, massive online open courses track and quantify users' learning activities, etc. This data presents an opportunity for us to gain better insight into people's behavior patterns and train our machine learning algorithms to suggest improvements. This task is especially important in the health domain. Unhealthy behaviors, such as physical inactivity, increase the risk of many adverse health conditions, including major non-communicable diseases [75]. Even modest adjustments to lifestyle behaviors are likely to have considerable health benefits [71].

People often know what the requirements for healthy behaviors are (e.g., exercise most days, eat a varied and nutritious diet or get enough sleep) but do not know the practical ways of reaching them [56]. For example, getting up earlier and exercising before going to work may appear to be a good strategy, but translating this goal into a feasible regime may not be trivial. Even when people know the ways and have the necessary skills, it may be difficult to achieve behavior change and stick to it because it requires a lot of effort. In most cases, external help or support may be needed. These actions, which are called *interventions* (or *treatments*), are designed to foster or support behavior change [86]. Interventions may be delivered in different ways, such as through digital technology, and may be based on different motivational strategies, such as feedback, rewards, and social support [126]. Thus, a natural and important question is "how to suggest the most appropriate intervention given a particular user?". Consider the following scenario illustrated in Figure 1.1:

> Sophia is a senior adult in her 70's. She has been recently diagnosed of type 2 diabetes. According to her doctor, she needs to change her habits and become physically more active. Recently, Sophia's granddaughter bought a fitness tracker for her birthday. She was delighted and enthused to try it out. While using the

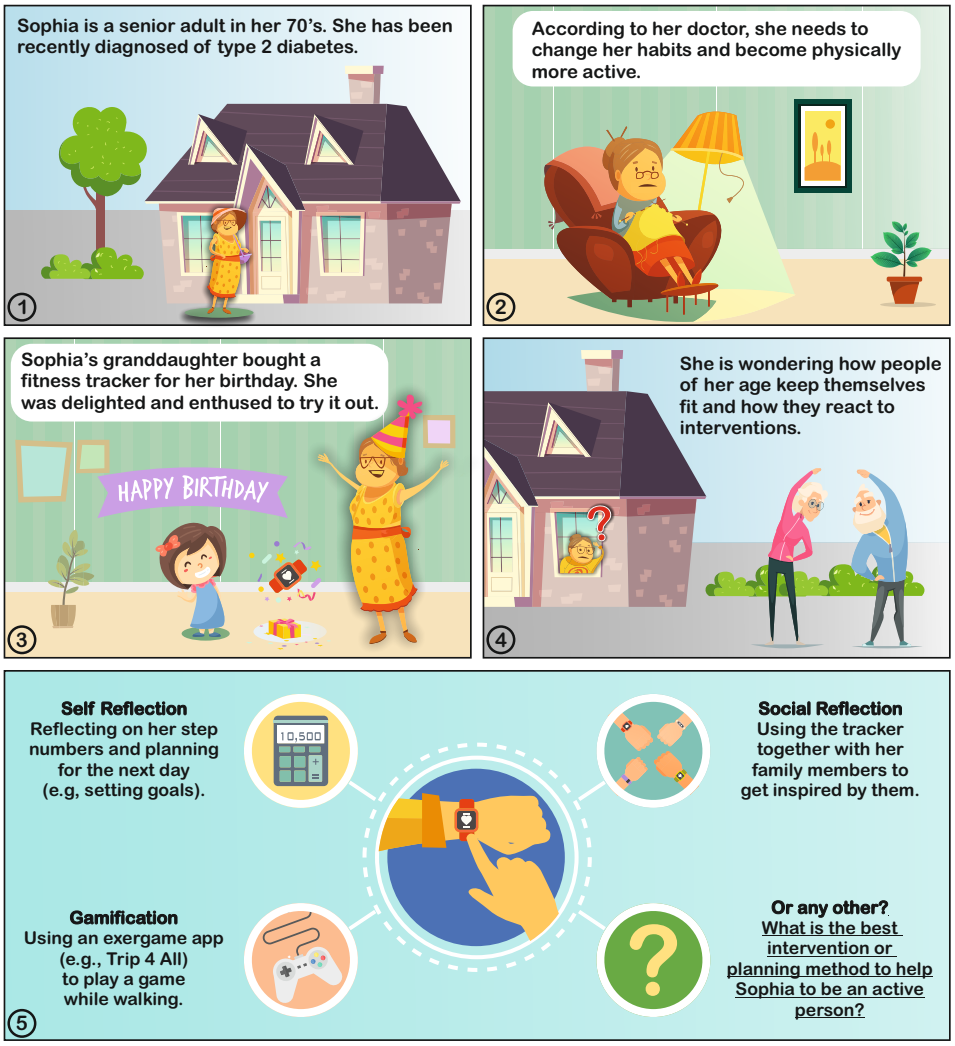


Figure 1.1 – A scenario that illustrates the need for a suitable personalized intervention to promote healthy behavior change.

> fitness tracker itself motivated Sophia to walk, she thinks how to adopt an efficient yet safe pattern over time. Sophia wonders what people of her age do to keep themselves them fit. She wants to choose a suitable intervention to improve her behavior: reflect on her daily step counts and set a goal for the next day, use the tracker together with her family members, or start using an exergame app (e.g., Trip 4 All [122]).

Sophia could try all possible ways to improve her behavior and then choose the most beneficial intervention. However, this strategy is time-consuming, inefficient, and might cause injuries to her. For example, if she decides to set personal goals, these goals might be similar to her

Illustration by Kavous Salehzadeh Niksirat.

past performances, resulting in no behavior change; if she decides to use an exergame app, the tasks given by the app might be too ambitious, resulting in potential injuries. Thus, we need a smart way to find and suggest the intervention most likely to work for Sophia. We should learn from past users, similar to Sophia, who received an intervention and improved their behavior. This requires machine learning methods to analyze fitness data and understand how different lifestyles may affect behavior change under different interventions. These methods could be integrated into a personalizable intervention system to suggest the intervention that is likely to work for Sophia, based on her fitness data.

However, it is challenging to model frequently-sampled time series data, such as fitness data, because human behavior is extremely complex. More specifically, human behavior is (1) time-varying, e.g., going to work in the morning; (2) interdependent, e.g., drinking water after workouts; (3) periodic, e.g., eating every few hours [70]. Thus, processing human behavior data requires taking into account its temporal nature. Another challenge is that data might be obtained from a limited number of users. This is often the case because data curation can be time-consuming and expensive. The goal of health intervention systems is to understand and predict the impact of interventions on behavior change, often using limited amount of frequently-sampled data. The conventional approach uses a randomized controlled trial (RCT) to measure the average intervention effect on the overall population, without fully utilizing the fine-grained information present in the sensor data. However, the best intervention for the whole population is not likely to be equally effective for each individual. Interventions should be tailored to the individual and his or her characteristics, such as his or her past behavior. It is less explored how the same intervention affects people with different behavior patterns. This holds a great opportunity to learn which behavior patterns are associated with healthy behavior change.

In this thesis we propose and develop personalizable intervention systems to promote healthy behavior change. Our work aligns with the goals of preventive healthcare [61]. These systems monitor the target user and decide *when* they need to act, but also *how* to intervene and *what* to suggest to the user. They consist of three main components: data collection, predictive modeling and generating recommendations. We give an overview of the system in Figure 1.2. The first component provides data containing evidence about intervention effectiveness (see Chapter 2). It collects data from people who received an intervention and whose behavior was tracked both before and after the intervention. Data should come from at least two groups of people receiving different interventions to allow the intervention system to compare and evaluate multiple interventions. These groups should have similar pre-intervention characteristics; otherwise, the data may contain intervention bias, meaning that some people are more likely to receive the intervention. Intervention bias makes it difficult for the system to infer causal relationships. However, it is possible, under some assumptions, to remove the intervention bias from the data and generate unbiased estimates of the intervention effect (see Chapter 6).

The second component, predictive modeling, learns to predict potential behavior changes

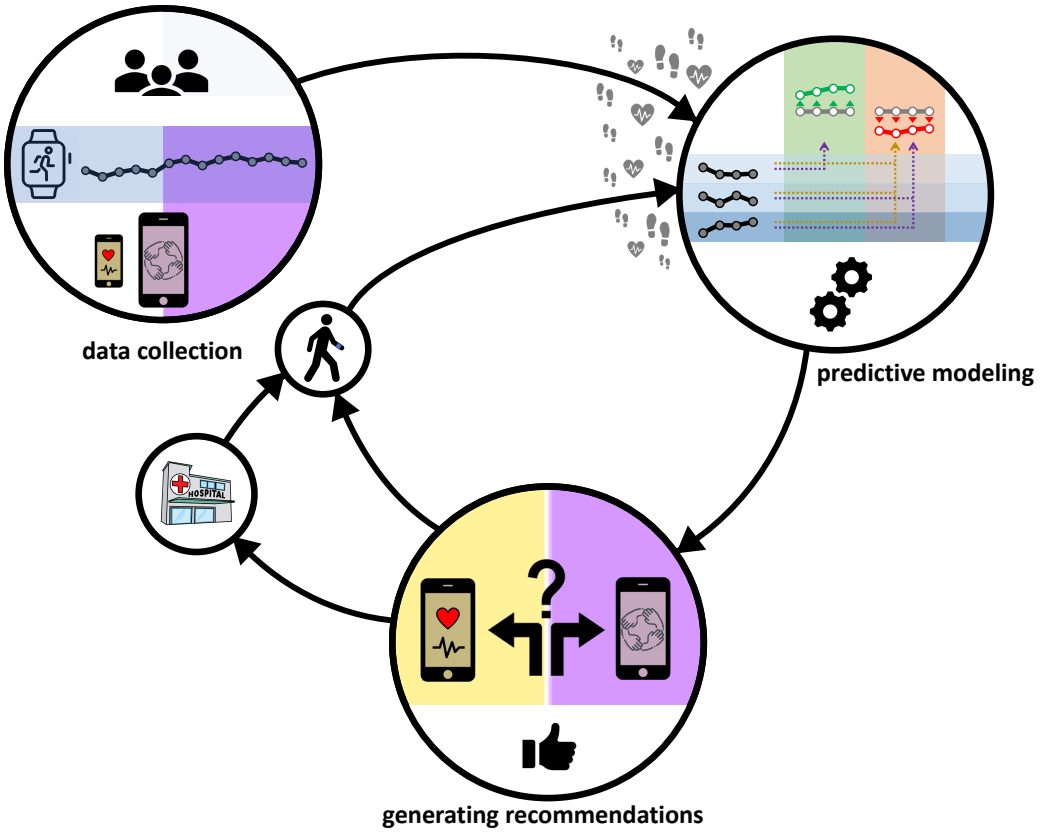


Figure 1.2 – An overview of the personalizable intervention system for healthy behavior change. Human behavior data is collected from people who received an intervention. This data is used to train a model that predicts potential behavior change for the target user under different interventions. Then, interventions are compared, and the optimal intervention is recommended to the user.

for new people under different interventions based on current and past observations. This problem has a few challenges. First, the set of observations may be large and represented as a frequently-sampled time series data. In this case, the component needs to learn to extract relevant features for the predictive task of interest in an automated way (see Chapters 4 and 5). These features should capture the complexity of human behavior. Second, there may be unobserved variables that affect the intervention effect. In this case, a standard regression model may not explain well the behavior change. Thus, it may be more desirable to use latent-variable models for this purpose (see Chapters 3 and 4). The latent class model discovers subpopulations that were affected by the intervention in different ways and estimates probabilities for new users to belong to each population based on their pre-intervention data. Third, most machine learning models are difficult to interpret. However, interpretability is important for any decision support system, especially in the medical domain [91]. By explaining how it works, the system becomes more transparent and increases users' confidence or trust [145]. Thus, we need to use predictive models that are both accurate and explainable.

Fourth, training data may be obtained from a limited amount of people. This makes it difficult for the predictive model to learn relevant patterns that generalize well to the target users (see Chapter 3).

The third component selects and recommends an optimal personalizable intervention for the target user based on the predictions generated by the second component. It compares and evaluates multiple interventions and recommends the intervention likely to produce the most desirable behavior change (see Chapter 5). Predictions about future human behavior may also be used to generate recommendation strategies that would assist the target user to achieve healthy behavior change. The system should learn these strategies from the existing users who have achieved positive behavior change (see Chapter 4). These strategies should be both *feasible* and *effective*. They aim to improve behavior change without introducing much risk for the target user — very ambitious strategies might be harmful, e.g., recommendations to extremely increase physical activity levels may cause injuries to the sensitive elderly population.

1.2 Research Agenda

In this thesis, we address the following research challenges in building a personalizable intervention system to promote healthy behavior change:

- **Intervention-Based Clustering.** Discovering subpopulations for which an intervention is most beneficial (or harmful) is an important goal of many clinical trials [60]. Often, these trials include only a limited number of participants. Different clustering (or *partitioning*) approaches could be used to discover subpopulations with differential intervention effects. Most of the existing methods found in the literature are based on trees. The main disadvantage of the tree-based methods is that they do not support complex decision boundaries between clusters. Can we discover more distinctive subpopulations using non-linear decision boundaries? How can we learn clusters that generalize well to new, unseen people from a limited amount of data?

- **Learning from Time Series Data.** Sensor devices often provide frequently-sampled time series data. This data may contain relevant information that could be used to predict better behavior change. Deep learning models, such as Long short-term memory (LSTM) networks, have gained much attention in recent years due to their application in time series modeling. However, these models require a large amount of data collected from many people to learn the variation in behavior change. Also, their predictions are hard to explain. How can we learn relevant predictive information from a limited amount of frequently-sampled time series data? Can we extract insights from data that explain the behavior change of different people?

- **Generating Recommendations.** The purpose of the personalizable intervention system in the focus of this thesis is to recommend an intervention that is likely to cause healthy behavior change. The system uses predictions about potential behavior change to

compare and evaluate multiple interventions. Conventional recommender systems suggest items that are similar to the items that the user liked in the past. However, personalizable intervention systems to promote healthy behavior change consider the user's personal goal, besides his or her past data, to generate recommendations. For example, if the user was not active in the past, but he or she wants to become more active, the system should not recommend actions similar to the ones that he or she performed in the past. Instead, the system should recommend actions that would improve his or her behavior, based on successful users similar to the target user. How can we adapt the existing recommendation approaches for the health domain? How can we learn from successful users and suggest feasible and efficient recommendation strategies for target users with unhealthy behaviors?

- **Removing Intervention Bias.** Observational studies allow monitoring the human behavior as well as the actions that lead to behavior changes in a natural setting. These studies are easier to conduct than experimental studies. However, data obtained from these studies may contain intervention bias. Neglecting this bias might lead to wrong conclusions about the intervention effect. How can we remove the intervention bias and estimate better the unbiased intervention effect? How can we improve existing methods that do not consider intervention bias so that they support data from observational studies besides data from experimental trials?

1.3 Main Contributions

In this thesis, we propose different solutions that address our research challenges. The principal contributions of this thesis are:

- We propose a system based on a Bayesian mixture model to identify subpopulations with different behavior changes from longitudinal data [68]. This system is useful when we are interested in the effect of various demographic, social, environmental, and behavioral factors on the long-term behavior change under an intervention. We show that our system can discover the subpopulations that respond to the intervention from limited amount of data.

- Frequently-sampled sensor data provides insight into people's low-level behavior patterns, e.g., daily routines, in contrast to longitudinal data. We propose a system to discover and predict behavior change patterns from this type of data [67]. The system learns both the pre-intervention behavior patterns and the post-intervention behavior change patterns. The system also estimates the transition probabilities between these patterns, allowing it to predict behavior change for new users. We demonstrate that the system produces explainable patterns that may be used to recommend strategies for healthy behavior change.

- Elderly population would benefit the most from healthy behavior change. Thus, we

propose a system that aims to promote physical activeness in senior adults [69]. The system takes minute-by-minute step count data provided by fitness trackers as input. It recommends a mobile app intervention that is most likely to work for the target user based on his or her activity patterns.

- Data from randomized controlled trials allow intervention systems to make less biased estimates of the intervention effect. However, observational studies are relatively quick, inexpensive, and easy to undertake, compared to randomized controlled trials [52]. We propose an adversarial approach to reduce bias when estimating the intervention effect from observational data. Our approach learns a balanced representation of the covariates across different treatment groups. This representation preserves the predictive information from the covariates as much as possible while reducing intervention bias. We show that our approach performs better than the existing approaches on a widely-used benchmark dataset. We demonstrate how to adapt existing personalizable intervention systems that do not consider intervention bias to support data from observational studies.

1.4 Thesis Structure

We organize this thesis as follows:

- in Chapter 2, we describe the different ways to collect data for personalizable intervention systems to promote healthy behavior change.
- in Chapter 3, we present a system based on a Bayesian mixture model that discovers subpopulations with different behavior changes from longitudinal data.
- in Chapter 4, we present CLINT, a system that discovers and predicts behavior change patterns from fine-grained sensor data.
- in Chapter 5, we present a system that recommends mobile app interventions promoting physical activeness to senior adults.
- in Chapter 6, we present an adversarial approach to reduce intervention bias from observational data.
- in Chapter 7, we review the contributions of our thesis and present directions for future work.

2 Data Collection

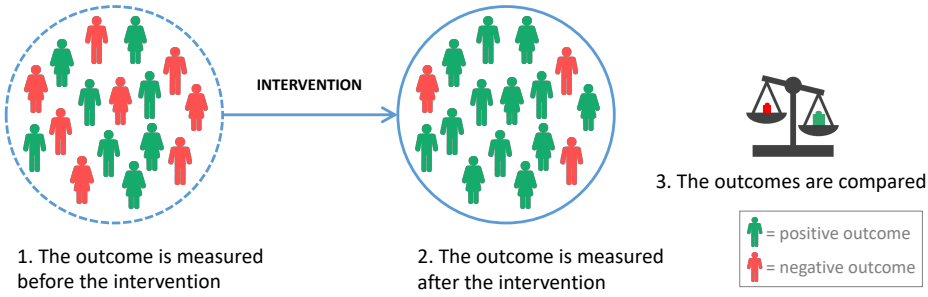
Personalizable intervention systems to promote healthy behavior change generate recommendations based on data containing evidence about intervention effectiveness. This data should come from a study with real people who are similar to the target users of the system. Within the study, some or all participants receive an intervention, and their behavior is measured both before and after this moment. Behavior change is defined as the difference between pre- and post-intervention behavior, e.g., the change of average daily step count. Personalizable intervention systems aim to predict the impact of different interventions on behavior change (an *outcome* of interest) based on pre-intervention data and recommend the intervention likely to achieve the most healthy behavior change. Three main study designs may be used to obtain relevant data to train the intervention systems: single-case study, randomized controlled trial, and observational study.

Single-case Study. This study design is particularly useful when a small number of participants are observed for a relatively long period of time [19]. In a single-case study, each participant receives an intervention and serves as his or her own control (see Figure 2.1a). Human behavior is repeatedly measured both before and after the intervention to obtain a stable estimate of the behavior change. All the variables that may affect behavior change should be recorded, if possible. These variables include time-invariant predictors (e.g., sex), time-variant predictors (e.g., stress), and contextual factors (e.g., weather). Current and past behavior patterns may be highly predictive of behavior change; thus, time series sensor data representing different aspects of human behavior should also be used as a predictor, if possible. A disadvantage of the single-case design is that behavior change may not be fully attributed to the intervention. For example, the change of weather after the intervention may better explain the decrease in activity levels of participants than the intervention itself. When conducting a single-case study, we need to ensure that the factors excluding the intervention that affect human behavior before and after the intervention are similar.

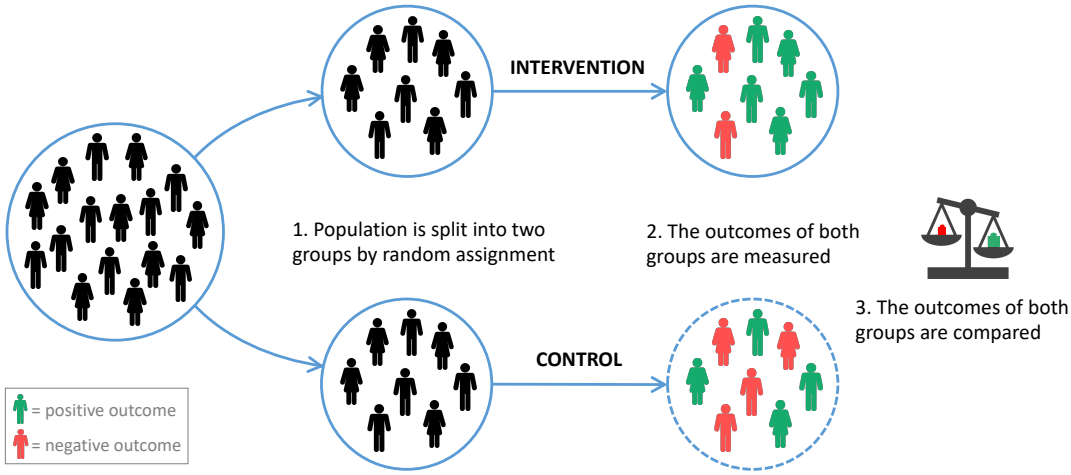
Randomized Controlled Trial. This study design aims to reduce bias when estimating the intervention effect. Participants are randomly assigned to an intervention group where they receive the intervention or a control group where they do not receive it (see Figure 2.1b). In our context, behavior change is measured for each participant in both groups, in a similar way as in single-case study. Then, behavior change in the intervention group is compared with the behavior change in the control group to estimate the intervention effect. Since the participants are randomly assigned to a group, the factors affecting behavior change in both groups are similar, except for the intervention. In this way, bias is reduced when estimating the intervention effect. But this benefit comes at a price: randomized controlled trials require more participants than single-case studies.

Observational Study. Compared to randomized controlled trials, observational studies are relatively quick, inexpensive, and easy to undertake [52]. Participants are observed in their natural environment, and they make their own decision whether to receive an intervention or not. Figure 2.1c illustrates this study design. The bias in intervention assignment does not allow a direct estimate of the intervention effect, in contrast to randomized controlled trials. Consider the case where a subpopulation (e.g., younger people) is more likely to receive an intervention (e.g., start using a specific mobile app that promotes physical activeness). Also, let us assume that the intervention is not effective for this subpopulation. Since most of the people who received the intervention belong to this subpopulation, by comparing the behavior change in people receiving and not receiving the intervention, we may wrongly conclude that the intervention is not effective — although the rest of the population may still benefit from the intervention. It is still possible to identify the intervention effect in observational studies if the collected data contains all the confounders: factors that affect both the intervention assignment and the outcome.

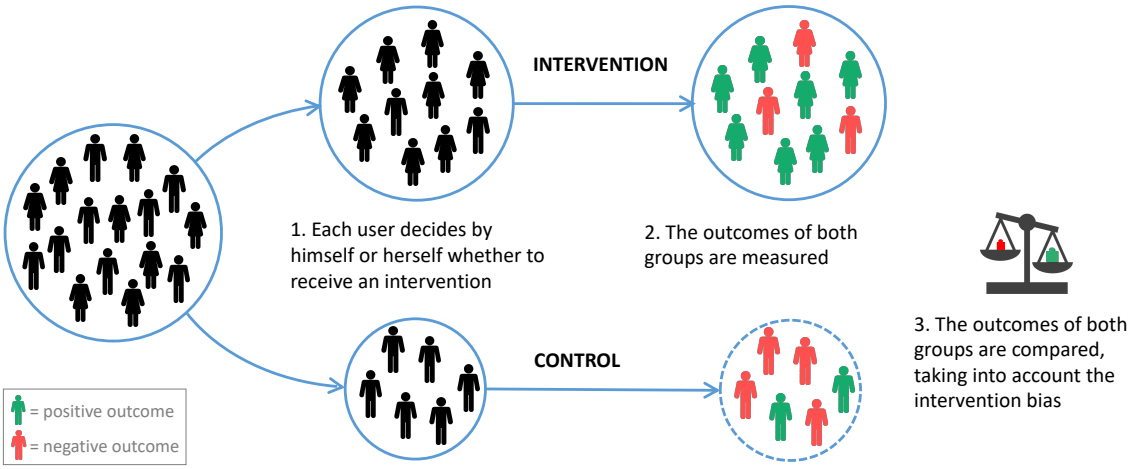
Sensor devices that are used to track people's behavior should be non-obtrusive, otherwise, they may act as an additional intervention — making it difficult to estimate the impact of the intervention of interest. Also, the number of recruited people should be high enough to capture the heterogeneity in the behavior change that exists in the population. A small sample size may produce models that would not generalize well to the target users.



(a) Single-case study



(b) Randomized controlled trial



(c) Observational study

Figure 2.1 – Comparison of different study designs.

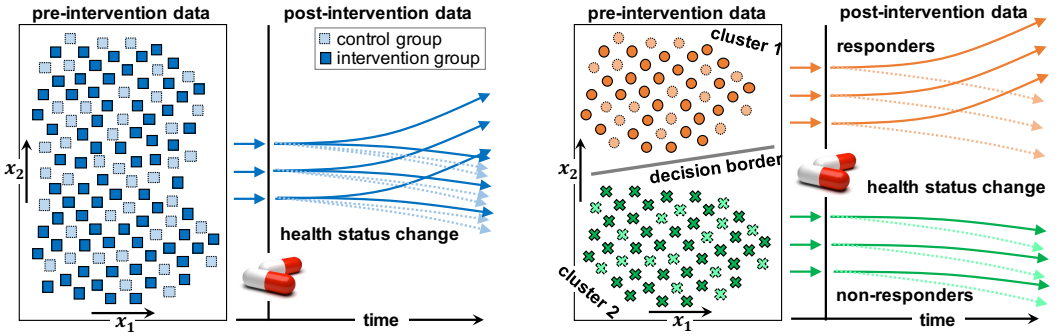
3 Intervention-Based Clustering

3.1 Introduction

Many situations require applying *interventions*, which are actions designed to bring about a change in a process or an individual. Examples of interventions are medical treatments, special offers in marketing, government policies, exergame apps, and exercises in teaching. In this chapter, we focus on the example of health interventions, but the techniques also apply to other domains.

The adoption of a new intervention requires scientific proof that it provides benefit. The conventional approach uses an RCT (randomized controlled trial) design to measure the intervention effect. Subjects are randomly assigned to a control group where they do not receive the intervention or a treatment group where they do. One or several variables, known as responses (e.g., a person's health status), are measured before and after the intervention. If the average response for the treatment group is better while it remains unchanged for the control group, it is likely the intervention worked. However, this method misses an important opportunity to examine the intervention effects at a more detailed level. Consider the case where a subpopulation (orange circles in Figure 3.1b) improves after taking medication while another group (green crosses in Figure 3.1b) does not. In both cases, effect changes are compared to the baseline (orange and blue dash lines). We may not find this difference if we used effect averages (Figure 3.1a). While some people improved (blue solid lines going up in Figure 3.1a), overall speaking the health status of a population did not change significantly. Our goal is to sub-divide the population into clusters taking into account their respective responses to the intervention (Figure 3.1b). In this manner, we will be able to decide whether or not to administer an intervention depending on the individual's characteristics. We believe this approach, which we call Intervention-Based Clustering (IBC), holds great promises in personalized medicine and preventive healthcare. Previous work in discovering the heterogeneity of the treatment effect (HTE) has addressed some of the challenges. Compared to these baseline

This chapter is based on the work of a paper published in the ACM Transactions on Intelligent Systems and Technology (TIST) [68].



(a) Conventional approach

(b) Intervention-based clustering approach

Figure 3.1 – Modeling the treatment effects of a given population: conventional vs. our approach

methods, we are providing the following advantages:

A more accurate model of the subpopulation. Subpopulations with differential treatment effects may be associated with complex membership functions, which cannot be modeled by traditional HTE methods. These membership functions might depend on both observed and unobserved variables and as a result, it is impossible to determine the cluster membership with full certainty. The ability of our method to model the cluster membership more precisely while taking into account its certainty could help in deciding who has the highest chances to respond positively to the intervention.

Modeling multiple objectives. Dealing with multivariate outcomes is also important when identifying subpopulations with differential treatment effects. This is because some interventions may affect multiple variables simultaneously, possibly causing desirable and adversarial outcomes at the same time. For example, energy drinks can give a person a strong boost while making him or her anxious. If we model both variables, we may identify a subpopulation who gets a boost without the anxiety side-effect. When there is more than one outcome variable of interest, we are able to make a trade-off between the beneficial and the harmful effects.

Bayesian approach. It can be challenging to identify the true subpopulations with differential treatment effect when the sample used in the analysis is small. This is because the treatment effect estimates become more variable and less stable as we decrease the size of the associated subpopulation and increase the number of parameters. Also, individuals with extreme responses (outliers) could significantly affect the estimates. For this reason, we have decided to use the Bayesian approach which allows us to include prior knowledge in the model.

Various methods that identify heterogeneous groups have been investigated in the literature [100]. The novelty of our approach is that it creates complex and more accurate decision

boundaries and allows reasoning about the trade-off of multivariate outcomes. It performs soft clustering and incorporates prior knowledge. The results of our approach can affect inclusion criteria in later clinical trials or can be used in deciding with higher confidence whether a person should receive a treatment based on how likely he or she is to respond to it.

This chapter is organized as follows. First, we review some of the existing methods used to identify HTE. Then we describe our approach in the third section. In the fourth and fifth section, we apply our approach to both synthetic and real data, and we compare it with two existing methods, QUINT (Qualitative INteraction Trees) and Growth Mixture Model. In the sixth section, we evaluate and discuss the relationship between the sample size and the quality of the HTE estimates. We conclude in the seventh section.

3.2 Related Work

Identifying the causal effect of a treatment on a patient is a difficult problem. To make an accurate causal inference, we need to observe the potential outcome if the subject received the treatment, the potential outcome if the subject received the alternative treatment and to compare the outcomes. This is not possible, because once treatment is applied to a patient, at most one potential outcome can be observed. Although we cannot simultaneously observe a single patient with and without the treatment, we can simultaneously observe a group with the treatment that is functionally identical to one without the treatment [97]. The causal effect of a treatment for that population can be estimated by comparing their average outcomes. The average treatment effect (ATE) can be easily estimated without bias in randomized experiments [60]. However, treatments might have different causal effects on each subject. Existing work is focused on either estimating the patient-level treatment effect [134] or searching for subgroups with differential treatment effects [80]. In both cases, we make use of the pre-treatment variables because they can be highly predictive of the potential outcomes. More concretely, we are interested in the conditional average treatment effect (CATE) which is an estimate of ATE for all possible combinations of values for the covariates.

3.2.1 Methods for Estimating CATE

To estimate CATE, we can use modern predictive modeling approaches such as boosting, random forest or support vector machines [143]. These methods essentially establish a relationship between attributes and outcomes, with a penalty parameter that penalizes model complexity [10]. Recently, Wager and Athley [134] developed a non-parametric causal forest that extends Breiman's widely used random forest algorithm [17]. The method utilizes the strength of the random forests to model interactions in high dimensions and provides asymptotically unbiased and normal estimates of CATE under the assumption of randomization conditional on the covariates or "unfoundedness" [108].

3.2.2 Methods for Identifying Subpopulations with Differential Treatment Effect

In practice, we are more likely to be interested in identifying subpopulations with differential treatment effects than simply estimating the patient-level CATE. For example, the existence of subgroups that appear to respond differently to treatment can affect inclusion criteria in later clinical trials or labeling decisions for approved drugs [57, 6]. Identifying subpopulations with differential treatment effect is a methodologically challenging task, especially when many characteristics are available that may interact with treatment and when no comprehensive a priori hypotheses on relevant subgroups are available [42]. The most popular methods for resolving this challenge found in the literature are based on trees. Trees produce a partition of the population according to covariates so that each subpopulation associated with a leaf has a distinct relationship between the covariates and the response. The most important feature of the trees is interpretability, enhanced by visualizations of the fitted decision trees [143]. Several different tree-based methods have been developed, including STIMA (simultaneous threshold interaction modeling) [41], Interaction Trees [125], Model-based recursive partitioning [143], Virtual Twins [50], SIDES (subgroup identification based on differential effect search) [78] and QUINT (Qualitative INteraction Trees) [42, 43].

Interaction Trees [125] follow the CART [18] convention, which consists of three major steps: (1) growing a large initial tree; (2) pruning; and (3) validation for determining the best tree size. Their splitting criterion is based on a measure for assessing the interaction that assigns high values when the squared difference between ATE in the left and right subtree is large and when the variance is small. Pruning is done using an interaction-complexity measure that penalizes trees with a large number of internal nodes. Each leaf represents one subpopulation and all the patients in a subpopulation receive the same estimate of CATE.

Model-based recursive partitioning [143] gives a tree where every leaf is associated with a fitted model such as a maximum likelihood model or a linear regression. The model in each leaf is fitted by minimizing some objective function e.g. the sum of squared errors or negative loglikelihood. Splitting is done if the parameter estimates are not stable with respect to at least one partitioning variable.

Virtual Twins [50] is based on the concept of potential outcomes [110]. The method consists of two steps. In the first step, a random forest is applied to data to estimate CATE for each patient. In the second step, a regression or classification tree is estimated with the patient-level treatment effect as the response variable. The algorithm outputs all the leaves in which the predicted differential treatment effect is larger than a threshold.

The goal of *QUINT* [42, 43] is to identify subgroups that are involved in optimal "qualitative" treatment-subgroup interactions where one treatment performs better than another in one subgroup and worse in another subgroup. The method outputs three groups, the first contains those patients for whom Treatment A is better than Treatment B, the second contains those for whom B is better than A, and the third (optional) contains those for whom it does not make any difference. The method builds a tree so that each leaf belongs to one of the three groups.

The partitioning criterion maximizes the absolute differential treatment effect in the first two groups and their sample sizes.

The main advantage of tree-based methods is that they do not require assumptions about the distribution of the dependent variable. Unfortunately, two main disadvantages remain. First, the splitting of each node is induced by a threshold on only one covariate, so space is always split using a hyperplane perpendicular to one of the axes and parallel to the other axes. This is why these methods may not fully identify the additive impact of multiple variables [137]. The second disadvantage is that they use a greedy approach to build the tree, which does not always result in an optimal tree.

The focus of the methods presented so far is to identify subpopulations with differential treatment effects measured at one instance after the intervention. However, when we work with longitudinal data [109], we may be interested in knowing how the treatment effect develops during the time after the intervention. Bauer and Curran [13] advocate the strong need for trajectory methods that are capable of discerning and testing hypotheses about the developmental growth of unobserved population subgroups called latent trajectory classes. Latent growth modeling approaches, such as *Latent Class Growth Analysis* (LCGA) [63] and *Growth Mixture Modeling* (GMM) [93] have been increasingly recognized for their usefulness for identifying homogeneous subpopulations within the larger heterogeneous population and for the identification of meaningful groups or classes of individuals [63]. Besides the pre-intervention variables, these methods include time-related variables and optionally, time-varying variables [93], which explain the development of the subpopulation over time. The main idea behind these methods is that they represent the trajectory as a latent variable and the propensity of a patient to belong to a particular trajectory depends on its baseline characteristics. The main difference between LCGA and GMM is that LCGA assumes no within-class variance on the growth factors, whereas GMM freely estimates the within-class variances [63]. These models mostly have been applied to non-interventional data [93, 48], however, they have also been successfully used to analyze interventions, for example, interventions aimed at reducing aggressive behavior [94]. An advantage of GMM over tree-based methods is that it can identify the additive impact of multiple variables on cluster membership. Another advantage is that it assigns soft cluster memberships to each patient. As we mentioned before, in this way we model the reality better. For example, it is unrealistic to expect that patients who are otherwise very similar, but belong to different leaves of the tree due to hard constraints for splitting the tree, would be affected by the intervention in a very different way (determined by ATE in the corresponding leaves). A limitation of GMM is that there is no clear criterion for determining the optimal number of subpopulations [137]. Another issue with this model is the existence of singularities. This can be especially important if we use GMM to analyze the treatment effect measured at one moment after the intervention.

3.2.3 Comparison with our Approach

In our work, we have developed a Bayesian mixture model which is suitable for identifying the subpopulations with differential treatment effect. We consider that each person responds to the treatment in a particular way which is unobserved but is partially explained by the pre-intervention data. Our goal is to discover the different ways subjects respond to the treatment and to estimate the propensity of a subject belonging to a subpopulation that responds in a particular way. Higher uncertainty of the cluster membership may suggest that there are important factors that are not measured but explain the treatment effect better. In contrast to GMM, our method utilizes prior information to avoid the singularity problem and stabilize the treatment effect estimates.

Recently a number of researchers began using Bayesian approaches. For example, a Bayesian tree-based approach was proposed by Berger et al. [14]. Unfortunately, this method is not able to discover clusters with complex (nonlinear) decision boundaries. In another recent work, Shahn and Madigan [118] proposed a Bayesian framework for modeling treatment effect heterogeneity. In comparison with our approach, their method is not able to model multi-dimensional responses, such as the combination of effects stated earlier. Tree-based methods have been proposed for subgroup discovery in data sets with multi-dimensional responses [129, 79], but they have less power to identify complex subpopulations. We aim to overcome the limitations of the existing methods to produce higher-quality clusters. The novelty of our approach is that it combines several desirable qualities in a single method to effectively identify subpopulations with differential treatment effects: complex decision boundaries, multi-dimensional continuous outcomes, soft cluster membership and the ability to stabilize the highly variable treatment effect estimates.

3.3 Model

Randomized controlled trials (RCT) are the most rigorous way of determining whether a cause-effect relation exists between treatment and outcome [121]. In RCT, N people are allocated at random to receive one of M different treatments. One of these treatments is the standard of comparison or control. There are three types of observed variables in RCT: pre-intervention variables, treatment variables, and outcome variables. The pre-intervention variables represent the baseline characteristics of each subject and its environment, for example, age, gender, education, medical condition, rainy weather, etc. The treatment variables represent the type of intervention the subject received, for example, drug, or a persuasion message delivered in a mobile phone app, etc. The outcome variables represent the outcome of interest, for example, well-being or health status change. In personalizable intervention systems to promote healthy behavior change, the outcome may be the long-term behavior change measured sometime after the intervention. After conducting RCT, the analysis is focused on estimating the size of the difference in predefined outcomes between intervention groups [121]. However, people with different baseline characteristics might respond to the same treatment differently. We propose

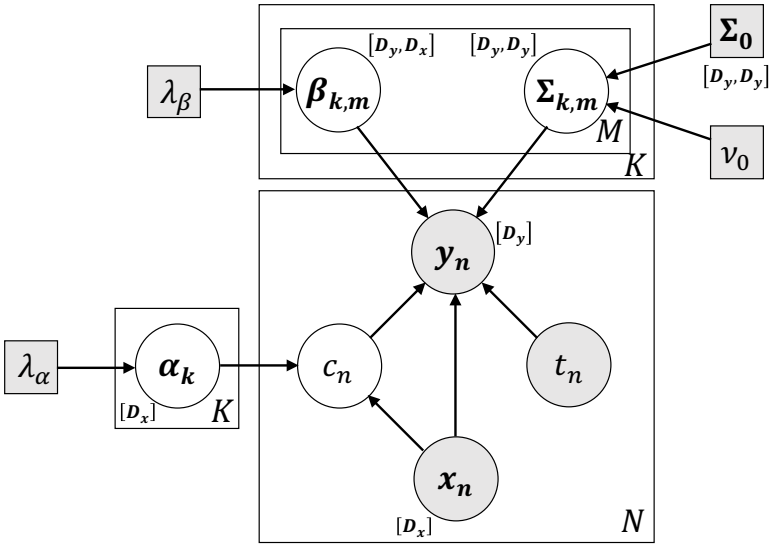


Figure 3.2 – Plate notation for our graphical model. The observed variables are displayed in gray circles, the unobserved variables are displayed in white circles and the hyper-parameters are displayed in gray squares. The dimensions of the multidimensional variables are displayed next to the variables' names. c_i represents the subpopulation (or the *cluster*) the user belongs to. α represents the logistic regression coefficients used to determine the probability of the user belonging to each cluster. β represents the regression coefficients used to estimate the potential outcome of the user if he or she received a particular intervention and belonged to a particular cluster.

a probabilistic graphical model to identify the homogeneous subpopulations (clusters) within the larger heterogeneous population. In the rest of this section, we will describe our model (Figure 3.2). Let's denote the pre-intervention, treatment and outcome variables associated with the n-th person by x_n, t_n and y_n correspondingly. x_n and y_n are multidimensional continuous variables whose dimensions are D_x and D_y, while t_n is a categorical variable with M levels. In RCT, y_n depends on both x_n and t_n, but x_n and t_n are independent because t_n is chosen randomly. We introduce a categorical variable $c_n \in \{1, 2, \ldots, K\}$ that is hidden and that identifies the type of response to the intervention (discrete heterogeneity of the treatment effect). There are K different types of responses and each person is associated with one of them. Observing the person's characteristics x_n we would like to determine the prior odds for him or her to respond according to each of the treatment effect types. This is why we set x_n to affect c_n in our model. If it was the opposite, c_n would represent both the treatment effect type and the type of person who receives the intervention. We say that people with $c_n = k$ belong to the k-th cluster, so a cluster represents a subpopulation with the same type of treatment effect. Besides c_n, x_n also directly affects y_n allowing for variation in subjects' individualized responses to treatment within the same cluster. This allows us to perform a more precise causal inference. The difference between the estimated individualized responses of the same person under two interventions, one of which is the control intervention, represents the causal effect of the experimental intervention on that person. We define $p(c_n = k|x_n, \alpha)$ to be

a softmax function:

$$p\left(c_n = k|x_n, \alpha\right) = \frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{i=1}^{K} \exp\left(\alpha_i^T x_n\right)} \tag{3.1}$$

where α represents the logistic regression coefficients used to determine the probability of a given user to belong to each cluster, given his or her pre-intervention data x_n. The motivations behind using softmax function are that its derivative is easy to calculate and it is simple to interpret i.e. an increase of the dot product $\alpha_k^T x_n$ increases the odds of the n-th person to belong to the k-th cluster (and vice versa). The first element of x_n should be set to one to interpret α_k^1 as the intercept. α_K should be a zero vector that is not affected in the learning process to make our model identifiable. In this way, we decrease the degrees of freedom without losing modeling power. We define y_n to be normally distributed with density:

$$\begin{aligned} p\left(y_n|c_n = k, x_n, t_n, \beta, \Sigma\right) &= \mathcal{N}\left(y_n|\beta_{k,t_n} x_n, \Sigma_{k,t_n}\right) \\ &= (2\pi)^{-\frac{D_y}{2}} \left|\Sigma_{k,t_n}\right|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(y_n - \beta_{k,t_n} x_n\right)^T \Sigma_{k,t_n}^{-1}\left(y_n - \beta_{k,t_n} x_n\right)\right] \end{aligned} \tag{3.2}$$

where $\beta_{k,t}$ represents the linear regression coefficients used to estimate the expected outcome if the user received the t-th intervention and belonged to the k-th cluster. $\Sigma_{k,t}$ represents the covariance matrix of the outcome variable associated to the t-th intervention and the k-th cluster. In our model, we associate Normal prior distributions to the logistic regression coefficients α and the linear regression coefficients β:

$$p\left(\alpha\right) = \prod_{k=1}^{K}\prod_{j=2}^{D_x} \mathcal{N}\left(\alpha_k^j|0, \frac{1}{\lambda_\alpha}\right) = \prod_{k=1}^{K}\prod_{j=2}^{D_x} \frac{1}{\sqrt{2\pi\frac{1}{\lambda_\alpha}}} \exp\left(-\frac{1}{2}\lambda_\alpha {\alpha_k^j}^2\right) \tag{3.3}$$

$$p\left(\beta\right) = \prod_{k=1}^{K}\prod_{m=1}^{M}\prod_{i=1}^{D_y}\prod_{j=2}^{D_x} \mathcal{N}\left(\beta_{k,m}^{i,j}|0, \frac{1}{\lambda_\beta}\right) = \prod_{k=1}^{K}\prod_{m=1}^{M}\prod_{i=1}^{D_y}\prod_{j=2}^{D_x} \frac{1}{\sqrt{2\pi\frac{1}{\lambda_\beta}}} \exp\left(-\frac{1}{2}\lambda_\beta {\beta_{k,m}^{i,j}}^2\right) \tag{3.4}$$

Also, we associate Wishart prior distribution to the covariance matrix:

$$\begin{aligned} p\left(\Sigma\right) &= \prod_{k=1}^{K}\prod_{m=1}^{M} W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right) \\ &= \prod_{k=1}^{K}\prod_{m=1}^{M} \frac{\left|\Sigma_{k,m}\right|^{\frac{\nu_0 - D_y - 1}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma_0^{-1}\Sigma_{k,m}\right)\right]}{2^{\frac{\nu_0 D_y}{2}} |\Sigma_0|^{\frac{\nu_0}{2}} \Gamma_{D_y}\left(\frac{\nu_0}{2}\right)} \end{aligned} \tag{3.5}$$

where Σ_0 and ν_0 are the scale matrix and the degrees of freedom of Wishart distribution $W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right)$.

3.3.1 Parameter Estimation

Using the method of maximum a posteriori estimation (MAP), we estimate model parameters as the mode of the posterior distribution of these random variables:

$$\arg\max_{\alpha,\beta,\Sigma} p\left(\alpha,\beta,\Sigma|Y,X,T\right) = \arg\max_{\alpha,\beta,\Sigma} \frac{p\left(Y|X,T,\alpha,\beta,\Sigma\right) p\left(\alpha,\beta,\Sigma\right)}{\int\int\int p\left(Y|X,T,\alpha,\beta,\Sigma\right) p\left(\alpha,\beta,\Sigma\right) d\alpha\, d\beta\, d\Sigma} \tag{3.6}$$

The denominator of the posterior distribution (so-called marginal likelihood) is always positive and does not depend on model parameters. Therefore, it plays no role in the optimization and we can rewrite the optimization objective as:

$$\arg\max_{\alpha,\beta,\Sigma} p\left(\alpha,\beta,\Sigma|Y,X,T\right) = \arg\max_{\alpha,\beta,\Sigma} p\left(Y|X,T,\alpha,\beta,\Sigma\right) p\left(\alpha,\beta,\Sigma\right) \tag{3.7}$$

$$= \arg\max_{\alpha,\beta,\Sigma} p\left(Y,\alpha,\beta,\Sigma|X,T\right) \tag{3.8}$$

When we logarithmize the product of probabilities we obtain the following function:

$$\log p\left(Y,\alpha,\beta,\Sigma|X,T\right) = \sum_{n=1}^{N} \log p\left(y_n|x_n,t_n,\alpha,\beta,\Sigma\right) + \log p\left(\alpha\right) + \log p\left(\beta\right) + \log p\left(\Sigma\right) \tag{3.9}$$

$$= \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} p\left(y_n|c_n=k,x_n,t_n,\beta,\Sigma\right) p\left(c_n=k|x_n,\alpha\right) \right]$$

$$+ \log p\left(\alpha\right) + \log p\left(\beta\right) + \log p\left(\Sigma\right) \tag{3.10}$$

After replacing Equations 3.1, 3.2, 3.3, 3.4 and 3.5 in 3.10, we obtain:

$$\log p\left(Y,\alpha,\beta,\Sigma|X,T\right) = \sum_{n=1}^{N} \log \left[\sum_{k=1}^{K} \mathcal{N}\left(y_n|\beta_{k,t_n} x_n, \Sigma_{k,t_n}\right) \frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{i=1}^{K} \exp\left(\alpha_i^T x_n\right)} \right]$$

$$+ \sum_{k=1}^{K} \sum_{i=2}^{D_x} \log \mathcal{N}\left(\alpha_k^i|0,\frac{1}{\lambda_\alpha}\right) + \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i=1}^{D_y} \sum_{j=2}^{D_x} \log \mathcal{N}\left(\beta_{k,m}^{i,j}|0,\frac{1}{\lambda_\beta}\right)$$

$$+ \sum_{k=1}^{K} \sum_{m=1}^{M} \log W\left(\Sigma_{k,m}|\Sigma_0,\nu_0\right) \tag{3.11}$$

This is our objective function and our goal in the learning procedure is to find the parameter values at the global maximum. Unfortunately, the function is not concave and it is difficult to find the global extreme point i.e. the most likely model parameters. However, we can find locally optimal parameter estimates using the Expectation-Maximization algorithm (EM). The algorithm starts with some initial parameter estimates $\alpha^{(0)}$, $\beta^{(0)}$, $\Sigma^{(0)}$, and iteratively updates and improves the estimates until convergence. Two steps are performed in each iteration: Expectation and Maximization. In the Expectation step, we use the current parameter values to find the posterior distribution of the latent variables. Given these probabilities, EM computes a tight lower bound to the true likelihood function. In the Maximization step, the lower bound is maximized, and the corresponding new estimate is guaranteed to lie closer to the

location of the nearest local maximum of the likelihood [35]. In the Expectation step of our learning procedure we calculate the posterior over c_n given the current estimates of the model parameters $\alpha^{(l)}$, $\beta^{(l)}$, $\Sigma^{(l)}$:

$$p_{n,k}^{(l)} = p\left(c_n = k|y_n, x_n, t_n, \alpha^{(l)}, \beta^{(l)}, \Sigma^{(l)}\right) = \frac{\mathcal{N}\left(y_n|\beta_{k,t_n}^{(l)} x_n, \Sigma_{k,t_n}^{(l)}\right) p\left(c_n = k|x_n, \alpha^{(l)}\right)}{\sum_{i=1}^{K} \mathcal{N}\left(y_n|\beta_{i,t_n}^{(l)} x_n, \Sigma_{n,t_n}^{(l)}\right) p\left(c_n = i|x_n, \alpha^{(l)}\right)} \tag{3.12}$$

We use the estimated posterior and Jensen's inequality to find the lower bound of Equation 3.11:

$$\begin{aligned}\log p\left(Y, \alpha, \beta, \Sigma|X, T\right) \geqslant & \sum_{n=1}^{N}\sum_{k=1}^{K} p_{n,k}^{(l)} \left[\log \mathcal{N}\left(y_n|\beta_{k,t_n} x_n, \Sigma_{k,t_n}\right) + \log \frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{i=1}^{K}\exp\left(\alpha_i^T x_n\right)}\right] \\ & + \sum_{k=1}^{K}\sum_{i=2}^{D_x} \log \mathcal{N}\left(\alpha_k^i|0, \frac{1}{\lambda_\alpha}\right) + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{D_y}\sum_{j=2}^{D_x} \log \mathcal{N}\left(\beta_{k,m}^{i,j}|0, \frac{1}{\lambda_\beta}\right) \\ & + \sum_{k=1}^{K}\sum_{m=1}^{M} \log W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right) = Q\left(\alpha, \beta, \Sigma|\alpha^{(l)}, \beta^{(l)}, \Sigma^{(l)}\right)\end{aligned} \tag{3.13}$$

The new parameter estimates are obtained by maximizing $Q\left(\alpha, \beta, \Sigma|\alpha^{(l)}, \beta^{(l)}, \Sigma^{(l)}\right)$. This is a concave function because it is represented as a sum of concave functions. It means that the function has only one global maximum. At this point, the derivatives of the function with respect to α, β, Σ are equal to zero, so by solving these equations we can find the optimal parameter estimates. The derivative of $Q(\cdot)$ with respect to α_k ($k < K$) is:

$$\begin{aligned}\frac{\partial Q(\cdot)}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N}\sum_{i=1}^{K} p_{n,i}^{(l)} \log \frac{\exp\left(\alpha_i^T x_n\right)}{\sum_{j=1}^{K}\exp\left(\alpha_j^T x_n\right)} + \sum_{i=2}^{D_x} \log \mathcal{N}\left(\alpha_k^i|0, \frac{1}{\lambda_\alpha}\right)\right] \\ &= \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N}\sum_{i=1}^{K} p_{n,i}^{(l)}\left(\alpha_i^T x_n - \log \sum_{j=1}^{K}\exp\left(\alpha_j^T x_n\right)\right) - \frac{\lambda_\alpha}{2}\sum_{i=2}^{D_x} {\alpha_k^i}^2\right]\end{aligned} \tag{3.14}$$

where $\bar{\alpha}_k^1 = 0$, and $\bar{\alpha}_k^i = \alpha_k^i$ for all $i > 1$. $\frac{\partial}{\partial \alpha_k}\alpha_i = 0$ for all $i \neq k$, thus:

$$\begin{aligned}\frac{\partial Q(\cdot)}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N}\left(p_{n,k}^{(l)} \alpha_k^T x_n - \log \sum_{j=1}^{K}\exp\left(\alpha_j^T x_n\right)\right) - \frac{\lambda_\alpha}{2}\sum_{i=2}^{D_x} {\alpha_k^i}^2\right] \\ &= \sum_{n=1}^{N}\left[p_{n,k}^{(l)} - \frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{j=1}^{K}\exp\left(\alpha_j^T x_n\right)}\right] x_n - \lambda_\alpha \bar{\alpha}_k = 0\end{aligned} \tag{3.15}$$

There is no closed-form solution to the equation above. This is why we use gradient ascent to find the optimal parameter values for logistic regression coefficients α_k. The derivative of $Q(\cdot)$

with respect to $\beta_{k,m}$ is:

$$\frac{\partial Q(\cdot)}{\partial \beta_{k,m}} = \frac{\partial}{\partial \beta_{k,m}} \left[\sum_{n=1}^{N} p_{n,k}^{(l)} \log \mathcal{N}\left(y_n | \beta_{k,t_n} x_n, \Sigma_{k,t_n}\right) + \sum_{i=1}^{D_y} \sum_{j=2}^{D_x} \log \mathcal{N}\left(\beta_{k,m}^{i,j} | 0, \frac{1}{\lambda_\beta}\right) \right] \tag{3.16}$$

Linear regression coefficients $\beta_{k,m}$ affect the distribution of the output variable y only in people who received intervention m. Also, $\frac{\partial}{\partial \beta_{k,m}} \beta_{i,m} = 0$ for all $i \neq k$. Thus, we can rewrite Equation 3.16 as:

$$\begin{aligned} \frac{\partial Q(\cdot)}{\partial \beta_{k,m}} &= \frac{\partial}{\partial \beta_{k,m}} \left[-\frac{1}{2} \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \left(y_n - \beta_{k,m} x_n\right)^T \Sigma_{k,m}^{-1} \left(y_n - \beta_{k,m} x_n\right) - \frac{\lambda_\beta}{2} \sum_{i=1}^{D_y} \sum_{j=2}^{D_x} {\beta_{k,m}^{i,j}}^2 \right] \\ &= \Sigma_{k,m}^{-1} \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \left(y_n - \beta_{k,m} x_n\right) x_n^T - \lambda_\beta \bar{\beta}_{k,m} = 0 \end{aligned} \tag{3.17}$$

where $\bar{\beta}_{k,m}^{i,1} = 0$ for all i, and $\bar{\beta}_{k,m}^{i,j} = \beta_{k,m}^{i,j}$ for all i and $j > 1$. $\mathbb{1}(t_n = m)$ is an indicator function that returns 1 if $t_n = m$ and 0 otherwise. There is no closed-form solution to this equation as well, so we can use gradient ascent to find the optimal parameter values for linear regression coefficients $\beta_{k,m}$ if the optimal covariance matrix $\Sigma_{k,n}$ is given. The derivative of $Q(\cdot)$ with respect to the covariance matrix $\Sigma_{k,m}$ is:

$$\begin{aligned} \frac{\partial Q(\cdot)}{\partial \Sigma_{k,m}} &= \frac{\partial}{\partial \Sigma_{k,m}} \left[\sum_{n=1}^{N} \sum_{i=1}^{K} p_{n,i}^{(l)} \log \mathcal{N}\left(y_n | \beta_{i,t_n} x_n, \Sigma_{i,t_n}\right) + \log W\left(\Sigma_{k,m} | \Sigma_0, \nu_0\right) \right] \\ &= \frac{\partial}{\partial \Sigma_{k,m}} \left[-\frac{1}{2} \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \left(\log \left|\Sigma_{k,m}\right| + \left(y_n - \beta_{k,m} x_n\right)^T \Sigma_{k,m}^{-1} \left(y_n - \beta_{k,m} x_n\right) \right) \right. \\ &\qquad \left. + \frac{1}{2} \left(\nu_0 - D_y - 1\right) \log \left|\Sigma_{k,m}\right| - \frac{1}{2} \mathrm{tr}\left(\Sigma_0^{-1} \Sigma_{k,m}\right) \right] \\ &= -\frac{1}{2} \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \left(\Sigma_{k,m}^{-1} - \Sigma_{k,m}^{-1} \left(y_n - \beta_{k,m} x_n\right) \left(y_n - \beta_{k,m} x_n\right)^T \Sigma_{k,m}^{-1} \right) \\ &\qquad + \frac{1}{2} \left(\nu_0 - D_y - 1\right) \Sigma_{k,m}^{-1} - \frac{1}{2} \Sigma_0^{-1} = 0 \end{aligned} \tag{3.18}$$

We transform Equation 3.18 by multiplying from left and right by the covariance matrix $\Sigma_{k,m}$ and after regrouping we get:

$$\begin{aligned} -\Sigma_{k,m} \Sigma_0^{-1} \Sigma_{k,m} + \left(\nu_0 - D_y - 1 - \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \right) \Sigma_{k,m} \\ + \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \left(y_n - \beta_{k,m} x_n\right) \left(y_n - \beta_{k,m} x_n\right)^T = 0 \end{aligned} \tag{3.19}$$

The solution $\Sigma_{k,m}$ to this equation for given linear regression coefficients $\beta_{k,m}$ is also a solution to the following Riccati equation [8]:

$$A^T X E + E^T X A - \left(E^T X B + S\right) R^{-1} \left(B^T X E + S^T\right) + Q = 0 \tag{3.20}$$

where

$$X = \Sigma_{k,m} \tag{3.21}$$

$$A = \frac{1}{2}\left(\nu_0 - D_y - 1 - \sum_{n=1}^{N} \mathbb{1}\left(t_n = m\right) p_{n,k}^{(l)}\right) I_{D_y} \tag{3.22}$$

$$B = E = I_{D_y} \tag{3.23}$$

$$S = 0_{D_y} \tag{3.24}$$

$$R = \Sigma_0 \tag{3.25}$$

$$Q = \sum_{n=1}^{N} \mathbb{1}\left(t_n = m\right) p_{n,k}^{(l)} \left(y_n - \beta_{k,m} x_n\right) \left(y_n - \beta_{k,m} x_n\right)^T \tag{3.26}$$

This equation has unique solution if:

$$\begin{bmatrix} Q & S \\ S_T & R \end{bmatrix} > 0 \tag{3.27}$$

It can be proven than this holds in our case using the definition of positive definiteness. We choose z to be non-zero vector of real numbers of size $2D_y$. Let's denote the first part of the vector of size D_y by z_1 and the second part of the vector of size D_y by z_2. Then:

$$z^T \begin{bmatrix} Q & S \\ S_T & R \end{bmatrix} z = \begin{bmatrix} z^T \begin{bmatrix} Q \\ S^T \end{bmatrix} & z^T \begin{bmatrix} S \\ R \end{bmatrix} \end{bmatrix} z = \begin{bmatrix} z_1^T Q & z_2^T R \end{bmatrix} z = z_1^T Q z_1 + z_2^T R z_2 > 0 \tag{3.28}$$

Therefore we can find the optimal $\Sigma_{k,m}$ by solving a Riccati equation if the optimal $\beta_{k,n}$ is given. We can find the optimal $\Sigma_{k,m}$ and $\beta_{k,m}$ in the Expectation step in an iterative process, by fixing $\Sigma_{k,m}$ to calculate new $\beta_{k,m}$, and by fixing $\beta_{k,m}$ to calculate new $\Sigma_{k,m}$, until convergence. The EM algorithm does not necessarily find the global extreme of the function. The quality of the solution depends a lot on the initial parameter values. We use a random restart approach for escaping a local maximum. Besides the parameters, our model has five hyper-parameters: λ_α, λ_β, ν_0, Σ_0 and K. They can be determined using grid search and cross-validation.

3.3.2 Alternative Approach to Estimate Model Parameters

It is possible to use a Markov Chain Monte Carlo (MCMC) approach instead of EM to estimate model parameters. Gibbs sampling is an MCMC approach where we iteratively replace the

value of one of the variables by a value drawn from the distribution of that variable conditioned on the values of the remaining variables until convergence. In our case, we should alternate between drawing samples from the conditional distributions $p\left(c_n = k|Y, X, T, C_{-n}, \alpha, \beta, \Sigma\right)$, $p\left(\alpha|Y, X, T, C, \beta, \Sigma\right)$, $p\left(\beta|Y, X, T, C, \alpha, \Sigma\right)$ and $p\left(\Sigma|Y, X, T, C, \alpha, \beta\right)$. The first conditional distribution is categorical and it is easy to draw samples from it. However, it is difficult to directly sample the model parameters because their conditional distributions are complex. For this purpose, we could use the importance sampling method [16]. The idea behind importance sampling is to simulate the conditional distribution using a different proposal distribution. L samples are generated from the proposal distribution and weights are assigned to each sample to correct the bias introduced by sampling from the wrong distribution. Then we use the discrete distribution defined by these samples and the normalized weights to simulate sampling from the complex conditional distribution. This results in generating a large number of samples.

The Gibbs sampling approach combined with importance sampling has two main advantages over EM. First, it is easier to implement because we don't need to maximize $Q\left(\alpha, \beta, \Sigma|\alpha^{(l)}, \beta^{(l)}, \Sigma^{(l)}\right)$ used in the Maximization step of EM. Second, given enough computational resources, it could converge to better parameter estimates than EM. However, this approach has three disadvantages. First, the convergence could be slow if the variables have strong dependencies. Second, in the importance sampling we need to choose the proposal distribution to be as similar as possible to the target distribution, so if this distribution is very biased, we will need a huge number of importance samples for this technique to achieve a sufficient confidence [92]. Third, importance sampling may not work well in high dimensions because in this case most of the samples carry no useful information [92], so an even larger number of samples need to be generated. This is an important limitation because in the learning algorithm we need to estimate the matrices $\Sigma_{k,m}$ which could be high dimensional, depending on the dataset. We cannot separately sample each value in the matrix because if we do this we might not obtain a positive definite matrix. Because of all these limitations, the application of Gibbs sampling on our problem would result in excessive time complexity. This is why we use the EM algorithm to estimate the model parameters. However, if the outcome variable is one-dimensional, Gibbs sampling may also be suitable.

3.4 Evaluation on Simulated studies

This section contains simulated experiments designed to evaluate the capability of our approach to capture the true underlying HTE present in the data. We defined several synthetic datasets to be used in the experiments. Each dataset involved two or three subpopulations with different treatment effects. A good model should recognize the true subpopulations. We validate our model (1) qualitatively, by comparing the true decision boundaries with the inferred decision boundaries and (2) quantitatively, by analyzing the prediction errors.

3.4.1 Complex Decision Boundaries

The goal of the first experiment was to evaluate the ability of our method to capture clusters with complex decision boundaries. For the purpose of this experiment we defined one simulated dataset that involves two continuous pre-intervention variables X_1 and X_2, and one continuous response Y. We generated 1,000 subjects so that for each subject x_n was randomly sampled from a Mixture of 20 Gaussians. We choose a distribution of X_1 and X_2 so that clusters cannot be clearly distinguished in this space. We randomly assign one of two treatments to each subject (control and intervention group) and we define three different types of responses to the treatments (c_n). We divide the subjects into three subpopulations and we associate one type of response to each subpopulation. The subpopulations were defined so that the boundaries between them are non-linear. The distribution of subjects and their true cluster memberships can be seen in Figure 3.3. The response of a subject $Y_{k,t}$ as a function of its subpopulation k and treatment group t was defined in the following way:

$$Y_{1,1} \sim 1+\varepsilon;\ Y_{1,2} \sim 0+\varepsilon;\ Y_{2,1} \sim 0+\varepsilon;\ Y_{2,2} \sim 0+\varepsilon;\ Y_{3,1} \sim 0+\varepsilon;\ Y_{3,2} \sim 1+\varepsilon \tag{3.29}$$

where ε comes from a Normal distribution with mean 0 and standard deviation 0.5.

We apply our approach to the simulated data to discover the HTE and to identify the subpopulations which respond to the intervention differently. We don't dismiss the possibility that there could be more complex non-linear decision boundaries between the subpopulations, so we include polynomial terms in the model up to degree P. We treat P as a hyper-parameter, besides the number of clusters K. In our approach, we set less informative prior on the model parameters ($\lambda_\alpha = 0.1$, $\lambda_\beta = 0.1$, $\nu_0 = \{4\}$, $\Sigma_0 = \mathrm{Cov}(Y)/\nu_0$). We built 20 different models, using different combinations of $P \in \{1,2,3,4,5\}$ and $K \in \{1,2,3,4\}$. We applied each model on an independent validation dataset of 10,000 subjects. The average log-likelihood on the validation dataset is given in Table 3.1. The model with the highest generalization power is the model with $K = 3$ and $P = 2$. Thus, both clustering and polynomial terms improve the log-likelihood on the validation dataset (see A.1 of the Appendix). We observe that the model correctly identified the true number of subpopulations. In Figure 3.3 we visualize the most likely cluster membership for different points in the pre-intervention variable space. We observe that the decision boundaries correctly discriminate between members of different true subpopulations. We also compared the discovered decision boundaries with the true decision boundaries and we observed that they are consistent. The results from the experiments suggest that our model can capture the true HTE and identify the subpopulations with differential HTE, even if they are separated with complex non-linear boundaries in the pre-intervention variable space. The root-mean-square error (RMSE) obtained by IBC is 0.5364 and is very close to the standard deviation of ε (0.5). This means that our model can identify the treatment effect associated to the subpopulations. We should note that the prediction error is lower than the error obtained by a linear regression (0.719198).

Table 3.1 – Average log-likelihood on the validation dataset for different number of clusters K and different number of polynomial terms P. We choose the model with the highest log-likelihood, indicated in bold.

	$P=1$	$P=2$	$P=3$	$P=4$	$P=5$
$K=1$	-1.0462	-1.0462	-1.0462	-1.0462	-1.0462
$K=2$	-0.8913	-0.8827	-0.8772	-0.8731	-0.8728
$K=3$	-0.8158	**-0.7872**	-0.7885	-0.7896	-0.8004
$K=4$	-0.7924	-0.7937	-0.7978	-0.8071	-0.8110

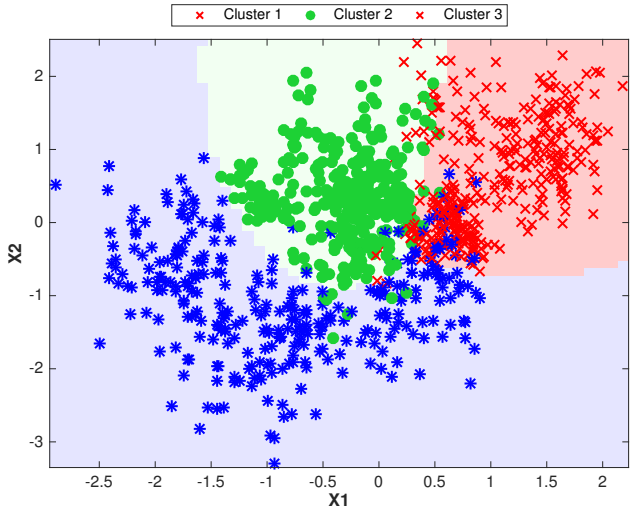


Figure 3.3 – Distribution of the data points in the first experiment. The color and the symbol associated with each patient indicate its true cluster membership. The background color and the decision borders indicate the most likely prior cluster membership generated by our model.

3.4.2 Comparison with Tree-Based Methods

In the second experiment, we compare our method with the tree-based method QUINT [42, 43]. We chose this method for comparison because its recovery performance is generally better than that of STIMA and as good as Interaction Trees, for true models comparable in complexity and size of the interaction effect [43]. For the purpose of our experiment, we defined five simulated datasets, each involving two continuous pre-intervention variables X_1 and X_2 and one continuous response Y (see Figure 3.4). Each subject in the datasets has equal chances to receive one of two treatments (control and intervention group). In these datasets, we defined linear decision boundaries to separate the subpopulations, in contrast to the previously used dataset. This was done to accommodate the QUINT model that cannot handle non-linear decision boundaries. We applied our approach and QUINT[1] on the

[1] R package quint with default parameters

simulated datasets, and we compared the results. For our approach, we set less informative prior on the model parameters ($\lambda_\alpha = 0.1$, $\lambda_\beta = 0.1$, $\nu_0 \in \{5, 10\}$, $\Sigma_0 = \text{Cov}(Y)/\nu_0$) and we run cross-validation to find the optimal number of clusters ($K \in \{1,2,3,4,5\}$).

In the first simulated dataset, we generated 1,000 subjects. Each subject was assigned to one of three subpopulations by sampling from a categorical distribution that is a function of X_1 and X_2. We defined this function so that there was high uncertainty in the process of sampling the subpopulation assignments, as shown in Figure 3.4a. The darker color means higher certainty that a subject belonging to that region will belong to the subpopulation associated with that color. ATEs in the first, the second and the third subpopulation were 0, -1 and 1, respectively. Our method recognized the true subpopulations in this dataset and assigned soft cluster memberships which could further be used to identify the most prominent responders. However, QUINT produced some heterogeneous subpopulations. This means that its members did not generally belong to one true subpopulation. As a consequence, the prediction error produced by QUINT was higher than IBC (Table. 3.2).

In the second simulated dataset, we generated 1,000 subjects belonging to two subpopulations so that the ATEs in the first and the second subpopulation were -1 and 1, respectively. In this setting, the subpopulation membership was fully determined by the additive impact of X_1 and X_2 and as a result, the subpopulations were separated by a straight line under the 45-degree angle (Figure 3.4b). Our approach was accurate to identify the true subpopulations, but QUINT identified only half of the true positive and true negative responders (lower left and higher right region). The other subpopulations were a mixture of true positive and true negative responders. As a result, the overall response in these subpopulations was neutral. This wrong estimate is used to predict the response for new subjects belonging to these regions and as a result, the RMSE for QUINT is much higher than the RMSE for IBC (Table. 3.2). We should note that increasing the size of the data should enable QUINT to produce smaller homogeneous regions. This, in general, holds true if the decision boundaries are more complex. In this case, each true subpopulation is distributed in a larger number of leaves.

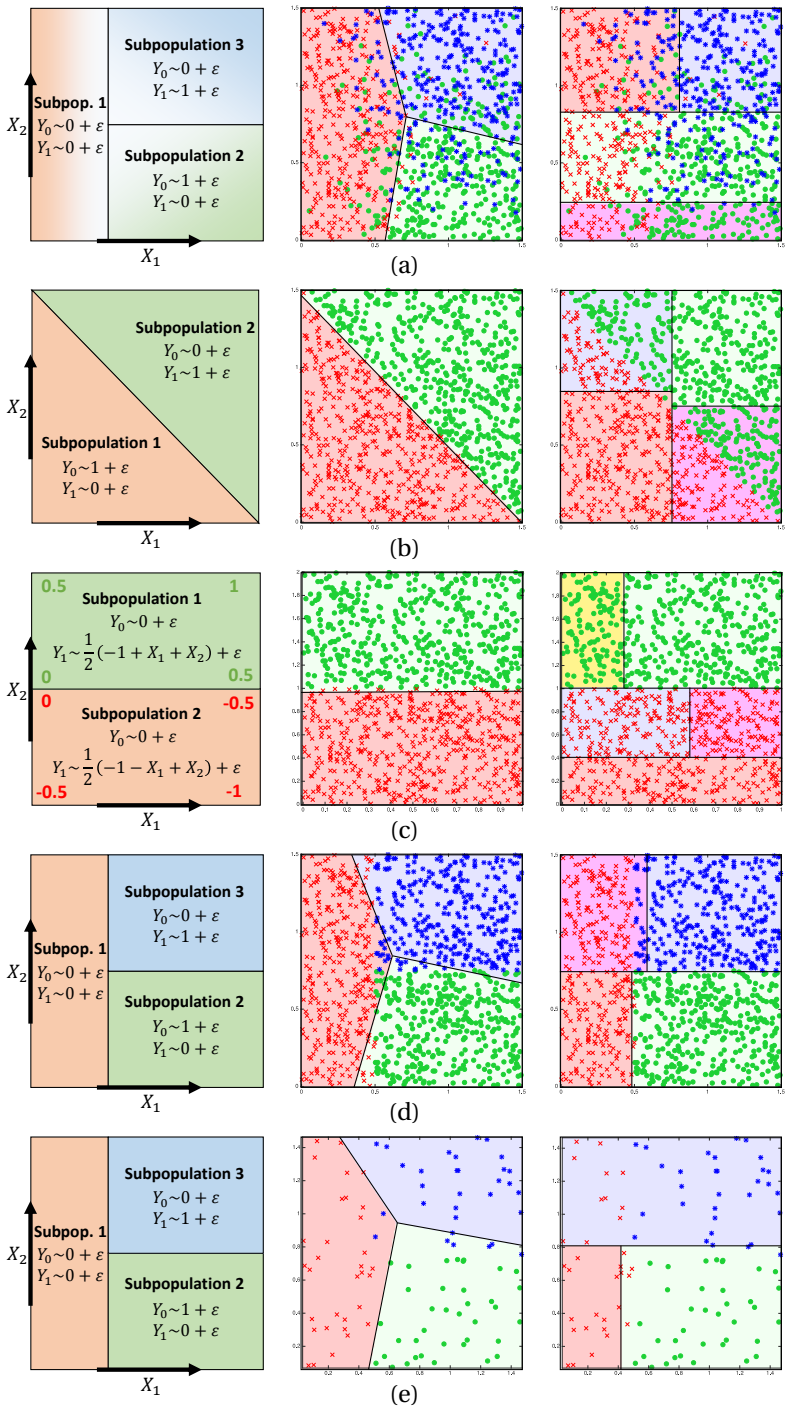


Figure 3.4 – The true subpopulations and the subpopulations discovered by our approach and QUINT in five different simulated datasets. The true ground truth model is shown on the left, and the results of our approach and QUNIT are shown in the middle and the right, correspondingly. The background color corresponds to the regions discovered by our method and QUINT, and the color associated with each subject corresponds to its true subpopulation membership.

Table 3.2 – RMSE produced by IBC, QUINT and linear regression on five synthetic datasets. The best performer on each dataset is indicated in bold.

	IBC	QUINT	LinReg
Dataset 1	**0.591910**	0.632236	0.661348
Dataset 2	**0.524258**	0.613043	0.845543
Dataset 3	**0.531897**	0.538983	0.693013
Dataset 4	0.564238	**0.542187**	0.763315
Dataset 5	**0.585064**	0.613337	0.831486

In the third simulated dataset, we generated 1,000 subjects coming from two equally sized subpopulations. In this setting, the response of the subjects in each subpopulation was defined to be a linear function of X_1 and X_2, not just a constant, as shown in Figure 3.4c. QUINT cannot identify responses that are more complex functions of x_n unless there is a large amount of data. In this case, QUINT decomposes the complex function into a union of simpler constant functions, each associated with one leaf. In this dataset, QUINT identified five subpopulations that differed in the direction and magnitude of the ATE. We should emphasize that the subjects from the intervention group respond in opposite (symmetric) ways in the lower and the upper half of the space. What is surprising is that the lower and the upper regions discovered by QUINT are not symmetric. This means that QUINT does not produce stable results even though the sample size is relatively large. In contrast, our model correctly identified the true subpopulations, as expected. However, the prediction errors were similar for both models and very close to the lowest possible RMSE (Table. 3.2).

In the fourth simulated dataset, we generated 1,000 subjects in a similar way as in the first dataset, except that we removed the uncertainty in the cluster membership. This means that a given x_n belongs to exactly one subpopulation uniquely determined by x_n. Our approach selected a model with three subpopulations that corresponded to the true subpopulations, as can be seen in Figure 3.4d. QUINT produced 4 instead of 3 subpopulations. The reason behind this is that the method is greedy and does not consider splitting on X_1 in the root node (any initial split on X_1 produces two sets of subjects with equal average treatment responses). However, all the subpopulations were homogeneous i.e. their members belonged mostly to one true subpopulation. This resulted in low RMSE, even lower than the RMSE produced by our method (Table. 3.2). We explain this by the fact that the decision boundaries discovered by IBC are not straight lines parallel to the axis. To perfectly reconstruct the true subpopulations, some of the model parameters need to have very extreme values. This is not likely to happen in our experiment because of the prior we imposed on the model parameters.

The fifth simulated dataset had the same underlying model as the fourth dataset but contained a smaller number of subjects (100). Our approach was robust enough to recognize the three true subpopulations. QUINT also produced three subpopulations, but not all of them were homogeneous. This is because QUINT did not have enough data to differentiate between different types of responses.

We conclude that IBC is better than QUINT in reconstructing the true subpopulations with differential treatment effects and produces smaller prediction errors. We also applied a linear regression model on the five datasets and its predictions were worse than both IBC and QUINT. This indicates that heterogeneity in the treatment effect shouldn't be neglected in the prediction task. It is interesting that if we just estimated the overall average responses for each treatment group and compared them, there would be no difference between the groups. So we might wrongly conclude that the intervention is not effective. However, when we apply the IBC approach we can identify the correct subpopulations with differential treatment effects.

3.5 Evaluation on Acupuncture Data

3.5.1 Dataset

We evaluated our algorithm on a randomized trial data where patients were randomly allocated to receive up to 12 acupuncture treatments over three months, in addition to standard care, or to a control intervention offering usual care [133, 132]. Headache score, SF-36 health status [136], and use of medication were assessed at baseline, three, and 12 months. The analysis of this data set showed that acupuncture leads to persisting, clinically relevant benefits for primary care patients with chronic headache, particularly migraine [133]. We applied our method on the acupuncture data to discover homogeneous subpopulations that were affected by the intervention in different ways.

We chose two output measures in our analysis: energy and emotional well-being. Higher scores indicate a better condition. These scores are estimated as a weighted sum of a particular subset of questions in the SF-36 questionnaire [136]. This questionnaire is a 36-item, patient-reported survey of patient health. Patients filled in the questionnaire before and after the intervention. We are interested in the long-term effect of the intervention, so the difference between energy and emotional-well being, assessed at 0 months and 12 months, is the outcome variable in our model. In the original analysis, the difference between the control and the intervention group reached significance for energy, but not for emotional well-being. Baseline energy, baseline emotional well-being, and age were included as pre-intervention variables in our model. There were 262 participants in the trial. They gave full responses on the SF-36 and were split into 121 for control and 141 for the intervention group respectively.

3.5.2 Model

Since our model involves hyper-parameters, we need to perform model selection. We use grid search for this purpose. We use 10-fold cross-validation to estimate the generalization performance. Before we build each model, we standardize the pre-intervention and the outcome variables in the training data set. In the model building process, we run 100 random restarts and choose the model parameters that maximize the likelihood function. To reduce the search space, we decided to define Σ_0 to be a function of ν_0 so that $\Sigma_0 = \mathrm{Cov}(Y)/\nu_0$. This is

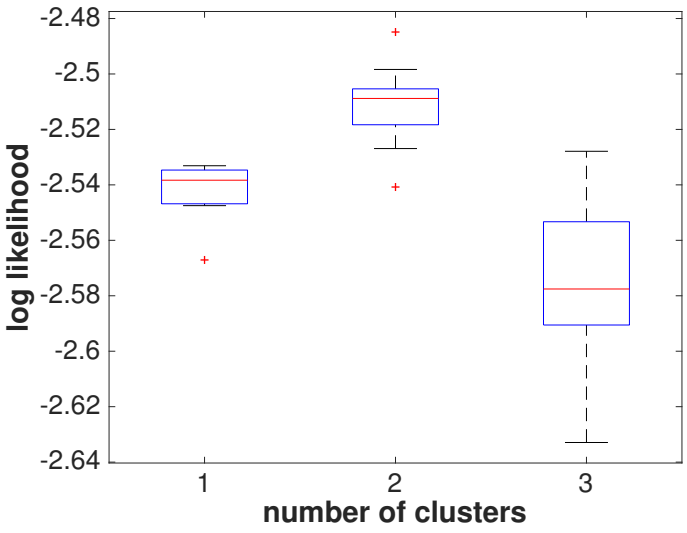


Figure 3.5 – Boxplot of the log-likelihood on the validation dataset for different number of clusters $K \in \{1,2,3\}$. For each K, we repeated all cross-validations 10 times. We show the boxplot of the log-likelihood on the validation dataset for the model with the optimal remaining hyper-parameters. The model with two clusters is suggested (statistically significant result with p-value < 0.00001).

how we ensure that the expected value of a Wishart random matrix is equal to the covariance matrix of the outcome variable. We choose 4 different values for each λ_α, λ_β and ν_0, i.e. $\lambda_\alpha \in \{0, 0.1, 1, 10\}$, $\lambda_\beta \in \{0, 0.1, 1, 10\}$ and $\nu_0 \in \{4, 8, 12, 16\}$. We varied the number of clusters K from 1 to 3. We repeated all cross-validations 10 times, each time with different random partitions to obtain higher relevance of the results. Our model selection procedure suggested a model with two clusters. This can be seen in Figure 3.5. The log-likelihood on the validation dataset for $K = 2$ is significantly higher than the log-likelihood on the validation dataset for $K = 1$ or $K = 3$.

After we select the optimal model, for each patient we could estimate the prior or the posterior odds for cluster membership. We use only pre-intervention variables to estimate the prior odds, and all variables to estimate the posterior odds. We are more interested in the first case because our goal is to predict the future behavior of the patient by only using the pre-intervention data. If we know that the patient is likely to belong to a cluster of people who respond to the intervention, then it is more likely that we should recommend the intervention to him or her. The most likely prior cluster membership for given baseline energy and baseline emotional well-being, with age fixed to zero, is shown in Figure 3.6. In the figure we can also see the prior cluster memberships for all the patients. The clustering that includes age as a predictor is not very different than the clustering that does not use age as a predictor. This indicates that age does not play a significant role in determining the prior odds for cluster membership, as it can be seen in Table 3.3. On the other hand, emotional well-being (EW) is the most important variable in determining the prior odds for cluster membership, because,

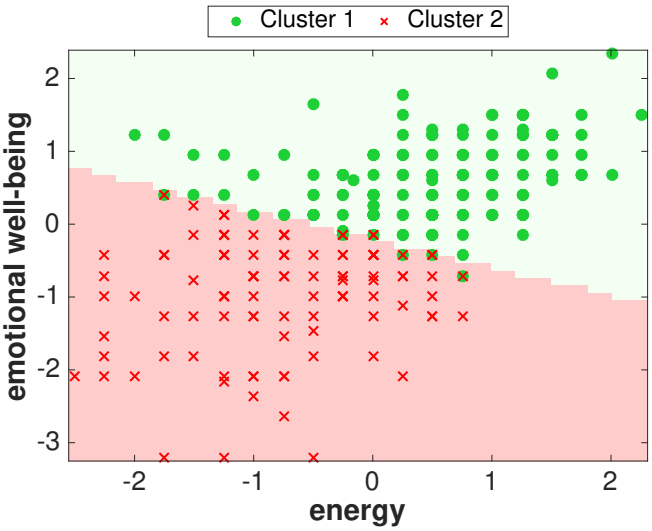


Figure 3.6 – The most likely prior cluster membership in the pre-intervention variable space (age is fixed to zero) and the most likely prior cluster membership for all patients.

Table 3.3 – Estimated model parameters.

Feature	α_1	$\beta_{1,1}^{1,:}$	$\beta_{1,2}^{1,:}$	$\beta_{2,1}^{1,:}$	$\beta_{2,2}^{1,:}$	$\beta_{1,1}^{2,:}$	$\beta_{1,2}^{2,:}$	$\beta_{2,1}^{2,:}$	$\beta_{2,2}^{2,:}$
intercept	0.12	-0.03	0.79	-0.38	-0.35	0.05	0.54	0.00	-0.45
age	-0.05	-0.32	0.00	0.21	0.04	-0.29	0.15	0.44	0.20
energy	0.20	-0.13	-0.76	-0.70	-0.22	0.51	0.16	-0.24	0.03
EW	0.52	-0.23	-0.25	0.00	0.02	-1.08	-0.85	-0.05	-0.55

for each increase of EW by one unit (standard deviation), the odds of belonging to the first cluster increase by 0.52. The average prior odds for the most likely cluster are not very high (0.625), suggesting that there are other unobserved variables that might improve the prediction of the treatment effect. The first cluster consists of healthier people, having better emotional well-being and higher energy than the people in the second cluster. There are 161 people in the first cluster (78 in the control and 83 in the intervention group), and 101 people in the second cluster (43 in the control and 58 in the intervention group).

In the rest of this section, we analyze how people from different clusters change their energy and EW after the intervention. In Figure 3.7, we see the mean relative change of energy and EW after the randomization for each cluster. The relative change at 12 months after the randomization represents the long-term Average Treatment Effect (ATE). Although we don't use the measurements of energy and EW performed 3 months after the randomization, we show them in the figure to better observe people's behavior in the post-intervention period. People from the intervention group in the first cluster increased their energy significantly more than the people from the control group (p-value < 0.01). However, there was no change in emotional well-being for both groups in the first cluster. Also, there were no significant

differences between the outcomes for both groups in the second cluster. Interestingly, these people improved both their energy and their emotional well-being regardless of whether there were or they were not under intervention.

We can use the obtained results to generate recommendations for better health (improved energy and/or EW). If people already have higher energy and EW (they belong to the first cluster), then recommend them with acupuncture treatment in addition to standard care. We expect that this would result in higher energy, but no change of EW. If people have low energy and EW (they belong to the second cluster), then recommend them only standard care. We expect that this would result in higher energy and higher EW. Giving acupuncture to these people in addition to standard care would not make a significant difference and would be more costly. The recommendation strategy that is based on our model is more cost-effective than the strategy that gives both standard treatment and acupuncture to everyone. However, we must emphasize that acupuncture is not a good intervention because it doesn't treat the people who need it more i.e. those having low energy and low EW.

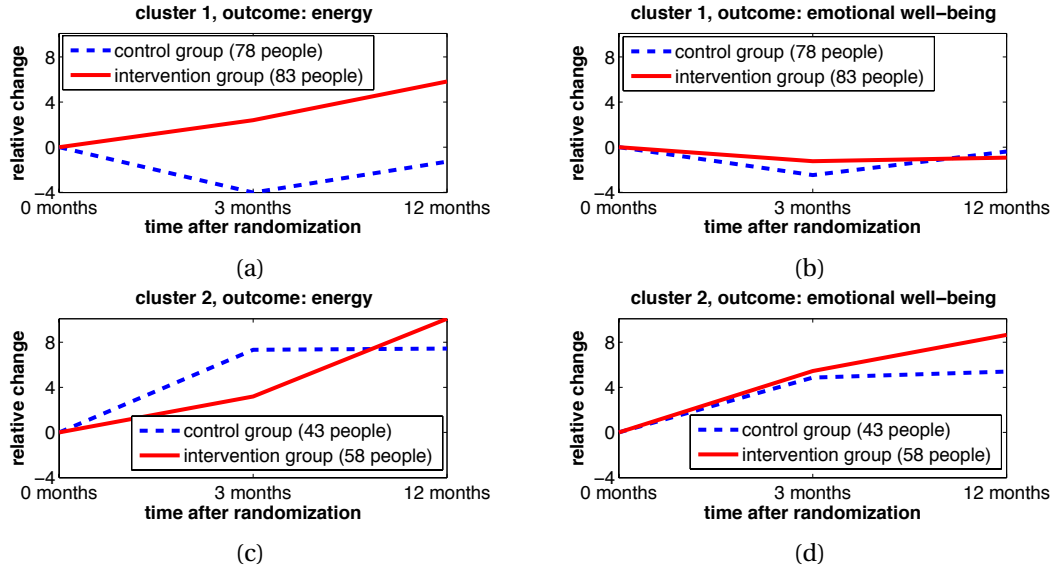


Figure 3.7 – Mean relative change of energy and emotional well-being in each cluster after the randomization. Each individual was assigned in the most likely cluster according to the prior odds for cluster membership.

3.5.3 Comparisons

In this section, we analyze the performance of our model and compare it with existing methods. We will perform standard cross-validation on the acupuncture dataset and we will use the log-likelihood and the root mean squared error (RMSE) on the held-out data as our performance measures.

In the first analysis, we compare three versions of our model: IBC-SIMPLE, IBC-COMPLEX,

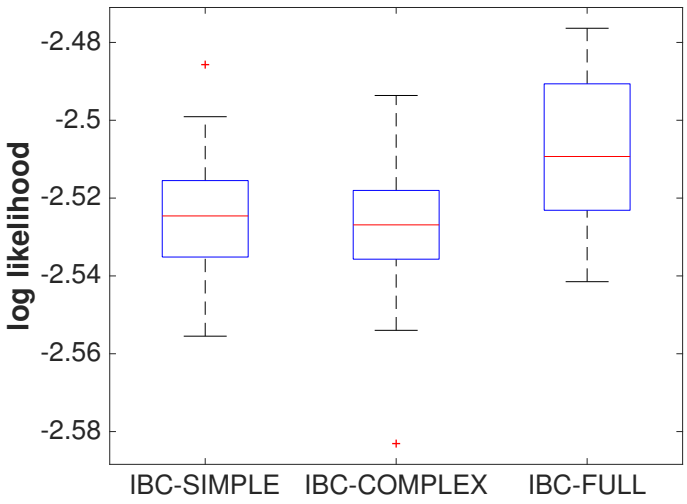


Figure 3.8 – Average log-likelihood on the validation dataset produced by three versions of IBC which differ in their power to represent the impact of the intervention (from left to right, lowest to highest). The unconstrained model IBC-FULL has the best performance on the validation dataset (p-value < 0.01).

and IBC-FULL. IBC-SIMPLE is a constrained version of our model where we set $\Sigma_{k,1}$=$\Sigma_{k,2}$ and $\beta^{i,j}_{k,1} = \beta^{i,j}_{k,2}$ for all i and $j > 1$ (the superscript denotes the position of the element in the matrix). IBC-COMPLEX is another constrained version of our model where we set just $\Sigma_{k,1}$=$\Sigma_{k,2}$. IBC-FULL is the unconstrained version of the model. We decided to use constrained versions of our model in the analysis because they have a smaller number of parameters and might generalize better on a small dataset like ours. After we trained the three versions of the IBC model, we observed that IBC-FULL performs the best and produces the highest log-likelihood on the held-out data (Figure 3.8). This demonstrates that although the unconstrained version of IBC has the highest degrees of freedom, the priors on the model parameters enable it to generalize well on small datasets.

In the second analysis, we compare our method with other existing methods. In this case, we use RMSE on the held-out data as our performance measure. The simplest model we compare with is the one that for a new user predicts that her response would be equal to the average response in the treatment group she belongs in (BASELINE-mean-treatment). A second baseline is a linear regression. We should note that linear regression can be considered as a constrained version of IBC with $K = 1$ and uninformative priors. Another model we compare with is QUINT. This model was separately applied to both responses, energy, and EW, with different critical minimum values [42] and the best model was chosen in each case. The last model we compare with is GMM[2]. We defined two variants of GMM, GMM-COMPLEX, and GMM-SIMPLE according to whether we allow the intervention indicator to interact with

[2] R package lcmm with default parameters

Table 3.4 – RMSE on the validation dataset for ten different models. Our model produces the smallest RMSE.

Model	energy	EW
BASELINE-mean-treatment	0.9930	1.0030
BASELINE-lin-reg	0.9337	0.9293
QUINT	0.9803	1.0021
GMM-SIMPLE-2-outputs	1.0182	0.9361
GMM-COMPLEX-2-outputs	1.0085	0.9544
GMM-SIMPLE-1-output	0.9576	0.9525
GMM-COMPLEX-1-output	0.9609	0.9525
IBC-SIMPLE	0.9318	0.9069
IBC-COMPLEX	0.9318	0.9325
IBC-FULL	**0.9265**	**0.9060**

the pre-intervention variables or not. These models were separately applied to both responses, energy, and EW, and we used random restarts to escape local maxima (GMM-SIMPLE-1-output and GMM-COMPLEX-1-output). The implementation of GMM that we used allowed modeling multivariate outcomes through a link function, so we additionally defined two more variants of GMM that use a linear link function to model both outcomes at the same time (GMM-SIMPLE-2-outputs and GMM-COMPLEX-2-outputs). We applied all these models on the acupuncture dataset and we compared their RMSE with the RMSE produced by our model. In Table 3.4 we can see that IBC-FULL produces the smallest RMSE on the held-out dataset. Linear regression also performs well on this dataset, in contrast to the other synthetic datasets on which it performed poorly. This might be because this dataset is of much smaller size, so simple methods still generalize well. QUINT performed very poorly and its RMSE was very close to the RMSE of our simplest baseline method.

3.6 Relationship Between the Sample Size and the Quality of the Results

The acupuncture dataset is small, so our method might not have enough information to approximate the true underlying model that generated the data. In this section, we try to get more insight into the amount of data needed to reconstruct the true underlying model with our method under the assumption that the true underlying model is the optimal model that we obtained from the acupuncture dataset. For the purpose of our analysis, we generated 10 synthetic datasets simulating RCT with 100 to 1,000 subjects based on this model. The pre-intervention data was sampled according to the distribution of the pre-intervention variables in the original dataset. Also, each subject was randomly assigned to one of two treatment groups. The outcome data were generated using the model that we trained on the original acupuncture dataset whose parameters are given in Table 3.3.

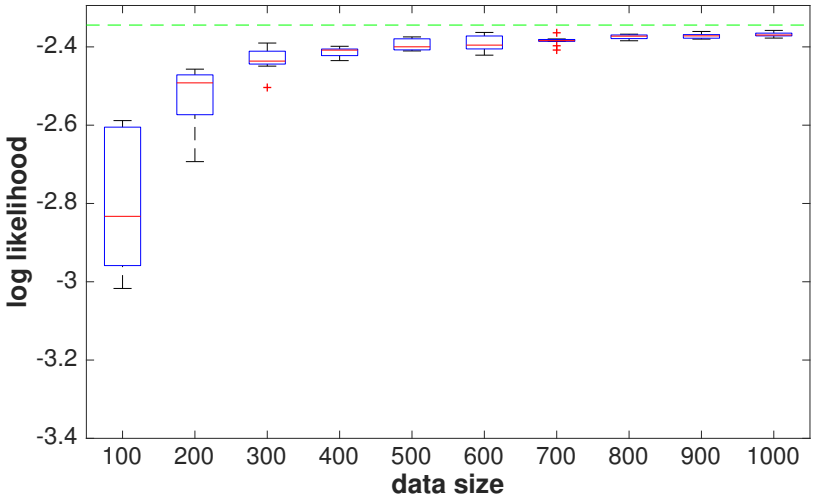


Figure 3.9 – Average log-likelihood on a large independent test dataset obtained using models trained on data from 100 to 1,000 subjects. The dashed green line shows the optimal log-likelihood obtained with the true model used to generate the data. If the trained model approximates well the true underlying model, then the average log-likelihood associated to this model should be close to the optimal log-likelihood. A sample size of 300 or more is required to obtain a good approximation of the true underlying model.

We performed an experiment to test how well the model training procedure can learn the true model parameters for different data sample sizes. This is shown in Figure 3.9. We generated a large independent dataset with 10,000 subjects to evaluate the models trained on the smaller synthetic datasets. If the learned model parameters are correct, they would result in the maximal average log-likelihood on the test data (dashed green line). We can see that the log-likelihood converges and becomes relatively stable at the point when the sample size is 300 or more. This means that 300 subjects would be enough to have a good approximation of the true underlying model, under the assumption that our model was powerful enough to describe the true data-generating process, as in this case. Otherwise, a good approximation may not be achievable. For example, if there was a non-linear relationship between the pre-intervention and the outcome variables, a linear model wouldn't be able to discover the true model. In this case, we should have used polynomial features to increase modeling power. We need to take into account that capturing more complex models with nonlinear decision borders, nonlinear response or larger number of clusters requires more data. For example, if we define our true underlying model simulating the acupuncture dataset to consist of more than two clusters, we would need more than 300 subjects to discover these clusters. Probably the acupuncture dataset we worked with is not large enough for us to discover more than two clusters if they were present.

3.7 Chapter Summary

The best intervention for the general population is not likely to be equally effective for each individual. We aim to provide interventions that are tailored to the individual, taking into account his or her characteristics and the characteristics of his or her environment. For this purpose, it is important to classify individuals into subpopulations that respond differently to the same intervention. Identifying subpopulations with differential treatment effect is a methodologically challenging task, especially when many characteristics may interact with treatment and no comprehensive a priori hypotheses on relevant subgroups are available [42]. The most popular approaches for this purpose are based on trees since trees provide features that are easy to interpret. However, many limitations remain as we have analyzed in the related section. We propose a Bayesian mixture model that combines four useful features to overcome the disadvantages of the tree-based approaches: generates soft cluster memberships for each subject, supports more complex decision boundaries, handles multivariate outcomes, and utilizes the strength of the Bayesian approaches to model better subpopulations with small sample sizes. Our method has two disadvantages: it does not guarantee that it can identify the optimal partition, and it has a higher computational cost than tree-based methods.

We applied our method on both simulated and real data and compared it with existing methods. Our model was able to capture the true HTE present in the simulated data, while QUINT, the tree-based method we were comparing with, had difficulties when there was uncertainty in the cluster membership (unobserved factors affecting the cluster membership), when the subpopulations were separated with decision boundaries at an angle, and when the response was a complex function of the pre-intervention covariates. We also demonstrated that if we look just at the overall treatment effect we might wrongly conclude that the intervention is not effective. However, when our method is applied to the data, it reveals subpopulations that respond differently than the overall response (if they exist). We also evaluated our algorithm on real-world randomized trial data. We were able to discover two distinct clusters of people. The intervention was effective in one of the clusters, suggesting that acupuncture significantly increases the energy levels of the people with high emotional well-being. We compared our method with QUINT and GMM, a mixture model that is mostly used to model longitudinal study data. Our method was able to predict the long-term treatment effect in the acupuncture dataset more accurately than the baseline methods. From our experiments and the qualitative and quantitative analysis of the results, we conclude that in comparison with the existing clustering methods (QUINT and GMM), our method produces more stable clusters (is more robust), reconstructs the true subpopulations better and has higher predictive power. In summary, the Bayesian approach to intervention-based clustering proposed in this chapter provides a better insight into the way different people respond to the same intervention. This allows for generating more suitable tailored interventions for healthy behavior change from longitudinal data.

4 Discovering Intervention Profiles From Time Series Data

4.1 Introduction

There is a growing number of wearable devices and mobile apps on the market that are able to track different human behaviors such as fitness and sleep [90, 53]. This data presents an opportunity for us to gain better insight into people's behavior patterns and to understand how they change over time. This has the potential to help users maintain and improve their personal well-being. Unhealthy behaviors are key risk factors for non-communicable diseases, including cardiovascular disease, cancer, and diabetes [124].

Different interventions, such as exergaming apps [7, 104] or SMS messages [64], can affect human behavior. They use different behavior change strategies, such as self-monitoring, goal-setting, feedback, prompts/cues, and gamification [135]. Althoff et al. showed that Pokemon Go — a mobile app that combines gameplay with physical activity — leads to substantial short-term activity increases [7]. However, as we saw in Chapter 3, the same intervention does not necessarily affect all people in the same way [72]. Only a subset of people might improve their behavior after the intervention (Figure 4.1). Understanding the correlation between pre-intervention behavior and behavior change is useful for personalizable intervention systems promoting physical activeness to decide *when* they need to act, but also to decide *how* to intervene and *what* to recommend to the user according to his or her personal goals.

Modeling behavior change is challenging, mainly because human behavior is complex. For example, people perform different activities depending on the time of the day and the context [70]. Accurately modeling human behavior allows predicting future human actions, e.g., biking, based on a sequence of past actions [70]. These models capture the regularities in human behavior, such as daily habits. Under intervention, there could be a change in the regular human behavior. This makes the prediction problem even more difficult because predictive machine learning models need to consider both the regularity in human behavior and the intervention effect to generate predictions. Existing techniques for identifying the intervention

This chapter is based on the work of a paper published at the 2016 NIPS Time Series Workshop [67].

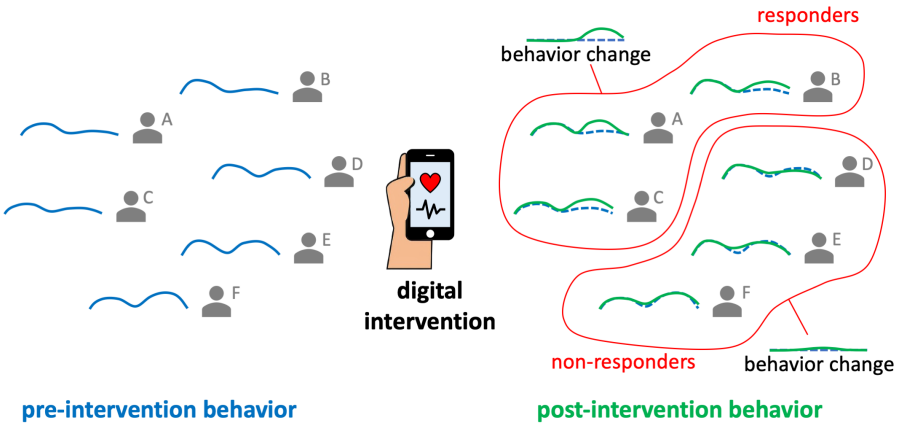


Figure 4.1 – The same health intervention does not affect people in the same manner. Our method discovers the distinctive patterns of behavior change in a population of users.

effect include Growth Mixture Model (GMM) [94], Interaction trees [125], Qualitative interaction trees [43] and Intervention-Based Clustering (IBC) (see Chapter 3). Most of these methods are not able to identify the impact of the intervention on a behavior that is represented as a time series [125, 43]. Other methods that work with time series data focus only on modeling the longitudinal developmental trajectories of individuals [94], e.g., trajectories of functional decline in older adults over a few years' time [112]. In contrast, in this chapter we are focused on modeling and predicting the post-intervention change of the *regular* (or *periodic*) human behavior represented as a time series, e.g., the typical daily activities during 24 hours.

Proper utilization of this data could lead to more informative insights into human behavior and its change. For example, the total number of steps on a daily basis (*coarse-grained* data) tells us whether some person is active or not, but the minute-by-minute measurements of his or her step counts (*fine-grained* data) additionally tell us when he or she is the most active during the day [141]. Behavior change modeling allows us to predict behavior change for new people without actually administering the intervention. A personalizable intervention system could use these predictions to generate feasible and effective recommendations for healthy behavior change in two different ways. First, it could recommend an intervention that caused positive behavior changes in existing people that are similar to the target user. Fine-grained behavior change predictions could help the intervention system to select the optimal intervention that would improve the target user while considering his or her personal preferences and constraints. For example, a physical activity intervention that has been most effective during the afternoon for the existing users would not be very useful to be recommended to the target user who is at work during that time. Second, a personalizable intervention system could support the target user to achieve his or her goal by recommending strategies that proved successful in responder users similar to him or her. Berndsen et al. demonstrated that predictions about future behavior can be used to generate explainable, adaptable pacing recommendations to marathon runners [15].

We present CLINT, a novel system for the task of discovering and predicting behavior change patterns after a given intervention from time series data. As illustrated in Figure 4.1, we are interested in finding subpopulations — subsets of users with similar pre-intervention behavior — that changed their post-intervention behavior in different ways. For example, those who responded positively to the intervention, vs. those who responded negatively. In contrast to existing methods, CLINT uses fine-grained time series data both as a predictor and an outcome of the intervention. This allows us to have interpretable and more-fine grained predictions about the behavior change than the existing approaches. CLINT models both the pre-intervention behavior patterns and post-intervention behavior change patterns, and estimates the transition probabilities between them, allowing to predict behavior change for new users. CLINT is based on a polynomial regression mixture model. This model is suitable when the observations are curves or time series [23]. Polynomial regression mixture model has been successfully used to discover: clusters of patients with dominant alcohol use patterns [34], customer habits in terms of water consumption [25], clusters of drivers' trajectories [139], etc. We extend the polynomial regression mixture model to discover changes in human behavior. There are three main reasons why we chose to use mixture model: (1) it models the unobserved factors that might affect behavior change; (2) it produces interpretable and informative patterns; and (3) it encodes the uncertainty of the behavior change. Our method uses time series data from people who already received the intervention and have been observed both before and after the intervention. This data could be provided from a single-case experiment in which the participant(s) acts as their own control [45].

We tested CLINT with a real-world dataset that was curated using a single-case experimental design. We discovered subpopulations with distinct and interpretable behavior change patterns. We used these patterns to predict fine fine-grained behavior change and we showed that our method is more accurate than the existing methods. We also demonstrated how the behavior change patterns could be used to generate recommendations that support the target user to improve his or her activity levels.

4.2 Related Work

Predicting human behavior. Much work has focused on modeling human behavior to predict future human actions [70, 102, 32, 147, 29]. These methods capture the regularities of human behavior, i.e., human actions depend on time and past actions. Kurashima et al. recently developed TIPAS, a mixture model that can model human behavior to predict which action (e.g. going for a run, going to sleep) will happen next and when [70]. Although our work is similar to [70] in terms of modeling behavior based on temporal features, it differs in one important aspect: we are interested in modeling and predicting behavior change after the intervention. Another difference is that TIPAS works with sequences of labeled user's actions, but CLINT works with sequences of continuous observations that could be generated by wearable devices such as fitness trackers. Predictions about the future human behavior may be used to generate recommendations that would assist the target user to achieve the desired behavior change.

Berndsen et al. used marathon performance data to provide personal guidance to runners who are predicted to slow down [15]. They generated recommendations based on successful runners who were similar to the target user. We also aim to learn from the responder users and recommend successful behavior change strategies to the target user. In our work, we directly use the behavior change patterns discovered by CLINT to generate recommendations. These recommendations suggest the target user *when* and *how* he or she should change his or her behavior, e.g., to increase his or her activity levels during the evening.

Discovering dominant behavior patterns. Another line of work has focused on discovering the dominant behavior patterns from raw mobile data [147, 44, 47]. Different methods such as Principal Component Analysis, Latent Dirichlet Allocation, and Gaussian Mixture Model, have been used for this purpose. Eagle and Pentland approximated an individual's behavior over a specific day by a weighted sum of the principal components of the complete behavioral dataset [44]. They demonstrated that when these weights are calculated halfway through a day, they can be used to predict the day's remaining behaviors with 79% accuracy. We are also interested to discover the dominant behavior patterns from time series data, but in addition to this, we want to find the correlation between these patterns and post-intervention behavior, allowing us to predict behavior change for new users.

Polynomial regression mixture model. Polynomial regression (PR) models human behavior by using a curve. Polynomial regression mixture (PRM) models assume that each curve is drawn from one of K clusters of curves with mixing proportions [22]. Each cluster of curves is modeled by a polynomial regression model. Polynomial regression is useful because it allows for nonlinear dependencies in the mixture components. Chamroukhi et al. proposed MixHMMR, an extension of PRM that can cluster time series with regime changes [24]. MixHMMR incorporates a hidden Markov chain allowing for transitions between different polynomial regressions over time [24]. The model was applied on real time series of railway switch operations to discover a cluster of time series corresponding to an operating state with a defect. CLINT also aims to model time series data with regime changes. The main difference between CLINT and MixHMMR is that our method is tailored to the problem of modeling behavior change, thus it discovers clusters of behavior change curves in addition to clusters of normal behavior curves.

4.3 Proposed Model

This section elaborates on the design of our model that can be used to discover behavior changes. Here we focus on the example of daily activities (more specifically, *calorie expenditure*), but our model can be applied to other types of behavior as well. Our goal is to understand how the intervention affects the calories burned at different times of a typical day. The number of calories burned indicates the amount of physical activity performed by the individual at a

given time.

4.3.1 Design Principles

We assume that each user behaves according to a specific activity pattern before the intervention and changes his or her behavior according to a specific activity change pattern after the intervention. This information is hidden and we want to infer it using the observed data: minute-by-minute measurements of the calories burned that could be obtained using fitness trackers. The user's *activity pattern* relates each minute of the day to the average number of calories burned in that minute during a typical day before the intervention. The user's *activity change pattern* relates each minute of the day to the change of the average number of calories burned in that minute during a typical day after the intervention. Our model is able to simultaneously discover the activity patterns and the activity change patterns, as well as to discover the conditional probability of observing an activity change pattern given an observed activity pattern (which we refer to as *intervention profiles*). This insight could be used to predict how a new user would change his or her daily activities after the intervention.

We use the Polynomial regression mixture (PRM) framework to model the conditional distribution of the measurements as a function of time. The main advantage of a mixture model is that it provides probabilities that a given user's behavior belongs to each of the possible patterns. In this way, we model reality better. For example, part of the time the user might behave according to one activity pattern, and the rest of the time he or she might behave according to another activity pattern. In the PRM framework, polynomial regression is used to represent the patterns. An alternative way to represent the patterns is by modeling the joint distribution of measurements and time using a Gaussian Mixture Model (GMM). The main advantage of polynomial regression over GMM is that it has a closed-form solution in contrast to GMM. Also, CLINT already represents human behavior as a mixture of patterns, so representing each pattern as another nested mixture model would make it more difficult for the learning algorithm to find an optimal solution.

4.3.2 Model Definition

Suppose there are N users in the dataset who received an intervention and each user's fitness tracker reports sequences of time-stamped measurements before and after the intervention. In our case, each measurement represents the number of calories burned in one minute. The l-th observation made before the intervention for the n-th user is defined by a tuple $\langle x^0_{n,l}, y^0_{n,l} \rangle$. $x^0_{n,l}$ stores the temporal information and $y^0_{n,l}$ stores the calorie expenditure information. Since we are interested in the typical daily behavior, we define $x^0_{n,l} = \left[1, t, t^2, ..., t^D\right]$ so that t represents the time of the measurement relative to the beginning of the day (in minutes) and D represents the number of polynomial terms. Higher D allows for more complex nonlinear modeling of the calorie expenditure. We treat $x^0_{n,l}$ as a predictor of $y^0_{n,l}$. In a similar way we define $\langle x^1_{n,l}, y^1_{n,l} \rangle$ to be the l-th observation made after the intervention for the n-th user. For

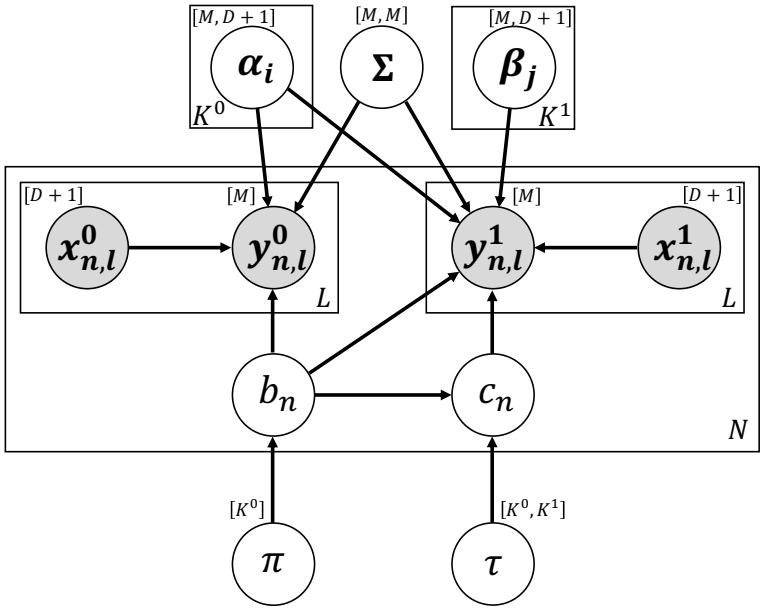


Figure 4.2 – Plate notation for CLINT. The observed variables are displayed in gray circles and the unobserved variables are displayed in white circles. The dimensions of the multidimensional variables are displayed next to the variables' names. b_n represents the type of user behavior before the intervention. c_n represents the type of user behavior change after the intervention.

simplicity we assume that for each user there are L observations made before the intervention and L observations made after the intervention, although the model can be easily adapted when the amount of data before and after the intervention is different.

In our model there are two categorical latent variables b_n and c_n associated to every user. The first latent variable indicates the n-th user's activity pattern, and the second latent variable indicates the n-th user's activity change pattern. There are K^0 different daily activity patterns and K^1 different activity change patterns. The relationship between the latent and observed variables is given in Figure 4.2. In the model we define that the pre-intervention observations are generated from the following probability distribution:

$$p\left(y^0_{n,l}|x^0_{n,l}, b_n, \alpha, \Sigma\right) = \mathcal{N}\left(y^0_{n,l}|\alpha_{b_n} x^0_{n,l}, \Sigma\right) \tag{4.1}$$

where α_i is $M \times (D+1)$ matrix representing the regression coefficients associated to the i-th activity pattern. M is the dimensionality of the observation, in our case $M = 1$ (we measure only calorie expenditure). The post-intervention observations are generated from the following probability distribution:

$$p\left(y^1_{n,l}|x^1_{n,l}, b_n, c_n, \alpha, \beta, \Sigma\right) = \mathcal{N}\left(y^1_{n,l}|\left(\alpha_{b_n} + \beta_{c_n}\right) x^1_{n,l}, \Sigma\right) \tag{4.2}$$

where β_j is $M \times (D+1)$ matrix representing the regression coefficients associated to the j-th

activity change pattern. We can interpret $\alpha_{b_n} x^0_{n,l}$ and $\left(\alpha_{b_n} + \beta_{c_n}\right) x^1_{n,l}$ as the user's typical pre- and post-intervention behavior, respectively. For simplicity, the covariance matrix Σ is common for both distributions. The prior probability distribution over b_n is defined by the model parameter π, $p(b_n = i|\pi) = \pi_i$, so that $\sum_{i=1}^{K^0} \pi_i = 1$. The prior distribution over c_n given b_n is defined by the model parameter τ, $p\left(c_n = j|b_n = i, \tau\right) = \tau_{i,j}$, so that for each i, $\sum_{j=1}^{K^1} \tau_{i,j} = 1$. The marginal log-likelihood of data is:

$$\mathscr{L}(\Theta; X, Y) = \log p\left(Y^0, Y^1 | X^0, X^1, \alpha, \beta, \Sigma, \pi, \tau\right) = \sum_{n=1}^{N} \log \sum_{i=1}^{K^0} \pi_i \gamma_{n,i} \sum_{j=1}^{K^1} \tau_{i,j} \xi_{n,i,j} \tag{4.3}$$

where

$$\gamma_{n,i} = \prod_{l=1}^{L} \mathscr{N}\left(y^0_{n,l} | \alpha_i x^0_{n,l}, \Sigma\right) \tag{4.4}$$

$$\xi_{n,i,j} = \prod_{l=1}^{L} \mathscr{N}\left(y^1_{n,l} | \left(\alpha_i + \beta_j\right) x^1_{n,l}, \Sigma\right) \tag{4.5}$$

4.3.3 Model Inference

In the learning phase we want to find optimal parameter values for α, β, Σ, π and τ so that the marginal log-likelihood is maximized. There is no closed form solution for this optimization problem. We use the Expectation Maximization (EM) algorithm to find (local) maximum likelihood parameters [36]. EM is an iterative method that alternates between performing expectation and maximization step. In the expectation step, the learning algorithm creates a function for the expectation of the log-likelihood using the current estimate for the parameters. In the maximization step, the learning algorithm computes parameters maximizing the expected log-likelihood found in the expectation step. In the expectation step of our algorithm we calculate the posteriors over the latent variables s and t:

$$\mathrm{v}_{n,i,j} = p\left(c_n = j | b_n = i, x^1_n, y^1_n, \alpha, \beta, \Sigma, \tau\right) = \frac{\tau_{i,j} \xi_{n,i,j}}{\sum_{q=1}^{K^1} \tau_{i,q} \xi_{n,i,q}} \tag{4.6}$$

$$\mathrm{u}_{n,i} = p\left(b_n = i | x^0_n, y^0_n, x^1_n, y^1_n, \alpha, \beta, \Sigma, \pi, \tau\right) = \frac{\pi_i \gamma_{n,i} \sum_{q=1}^{K^1} \tau_{i,q} \xi_{n,i,q}}{\sum_{p=1}^{K^0} \pi_p \gamma_{n,p} \sum_{q=1}^{K^1} \tau_{p,q} \xi_{n,p,q}} \tag{4.7}$$

We use the estimated posterior and Jensen's inequality to find the lower bound of Equation 4.3:

$$\begin{aligned}\mathscr{L}(\Theta; X, Y) &\geq \sum_{n=1}^{N}\sum_{i=1}^{K^0} \mathrm{u}_{n,i}\left[\log\pi_i + \log\gamma_{n,i} + \sum_{j=1}^{K^1}\mathrm{v}_{n,i,j}\left(\log\tau_{i,j} + \log\xi_{n,i,j}\right)\right] \\ &= Q\left(\alpha,\beta,\Sigma,\pi,\tau|\alpha^{(k)},\beta^{(k)},\Sigma^{(k)},\pi^{(k)},\tau^{(k)}\right) \end{aligned} \tag{4.8}$$

The new parameter estimates are obtained by maximizing $Q\left(\alpha,\beta,\Sigma,\pi,\tau|\alpha^{(k)},\beta^{(k)},\Sigma^{(k)},\pi^{(k)},\tau^{(k)}\right)$ with respect to α, β, Σ, π and τ. This optimization problem can be solved in closed form:

$$\begin{aligned}\alpha_i^{(k+1)} = &\left[\sum_{n=1}^{N}\mathrm{u}_{n,i}^{(k)}\left[\left[\sum_{l=1}^{L} y_{n,l}^0\left[x_{n,l}^0\right]^T\right] + \sum_{j=1}^{K^1}\mathrm{v}_{n,i,j}^{(k)}\sum_{l=1}^{L}\left(y_{n,l}^1 - \beta_j^{(k)}x_{n,l}^1\right)\left[x_{n,l}^1\right]^T\right]\right] \\ &\times\left[\sum_{n=1}^{N}\mathrm{u}_{n,i}^{(k)}\left[\left[\sum_{l=1}^{L} x_{n,l}^0\left[x_{n,l}^0\right]^T\right] + \sum_{j=1}^{K^1}\mathrm{v}_{n,i,j}^{(k)}\sum_{l=1}^{L} x_{n,l}^1\left[x_{n,l}^1\right]^T\right]\right]^{-1}\end{aligned} \tag{4.9}$$

$$\begin{aligned}\beta_j^{(k+1)} = &\left[\sum_{n=1}^{N}\sum_{i=1}^{K^0}\mathrm{u}_{n,i}^{(k)}\mathrm{v}_{n,i,j}^{(k)}\sum_{l=1}^{L}\left(y_n^1 - \alpha_i^{(k)}x_{n,l}^1\right)\left[x_{n,l}^1\right]^T\right] \\ &\times\left[\sum_{n=1}^{N}\sum_{i=1}^{K^0}\mathrm{u}_{n,i}^{(k)}\mathrm{v}_{n,i,j}^{(k)}\sum_{l=1}^{L} x_{n,l}^1\left[x_{n,l}^1\right]^T\right]^{-1}\end{aligned} \tag{4.10}$$

$$\Sigma^{(k+1)} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{K^0}\mathrm{u}_{n,i}^{(k)}\left[\eta_{n,i}^{(k)} + \sum_{j=1}^{K^1}\mathrm{v}_{n,i,j}^{(k)}\varepsilon_{n,i,j}^{(k)}\right]}{2NL} \tag{4.11}$$

$$\pi_i^{(k+1)} = \frac{\sum_{n=1}^{N}\mathrm{u}_{n,i}^{(k)}}{\sum_{n=1}^{N}\sum_{j=1}^{K_0}\mathrm{u}_{n,j}^{(k)}} \tag{4.12}$$

$$\tau_{i,j}^{(k+1)} = \frac{\sum_{n=1}^{N}\mathrm{u}_{n,i}^{(k)}\mathrm{v}_{n,i,j}^{(k)}}{\sum_{n=1}^{N}\sum_{p=1}^{K^1}\mathrm{u}_{n,i}^{(k)}\mathrm{v}_{n,i,p}^{(k)}} \tag{4.13}$$

where

$$\eta_{n,i}^{(k)} = \sum_{l=1}^{L}\left(y_{n,l}^0 - \alpha_i^{(k)}x_{n,l}^0\right)\left(y_{n,l}^0 - \alpha_i^{(k)}x_{n,l}^0\right)^T \tag{4.14}$$

$$\varepsilon_{n,i,j}^{(k)} = \sum_{l=1}^{L}\left(y_{n,l}^1 - \left(\alpha_i^{(k)} + \beta_j^{(k)}\right)x_{n,l}^1\right)\left(y_{n,l}^1 - \left(\alpha_i^{(k)} + \beta_j^{(k)}\right)x_{n,l}^1\right)^T \tag{4.15}$$

The quality of the solution depends a lot on the initial parameter values. We use the random restart approach for escaping a local maximum. The full algorithm for training CLINT is presented in Algorithm 1. A disadvantage of CLINT is that we need to define the hyper-parameters K^0 and K^1 in advance. To find their optimal values, we use grid search and k-fold cross-validation. We choose the model that produces the highest log-likelihood on the held-out data [120].

Algorithm 1 Optimization Algorithm

Input: X^0, Y^0, X^1, Y^1
Hyperparameters: K^0, K^1, tolerance
Output: $\Theta = \{\alpha, \beta, \Sigma, \pi, \tau\}$

$k \leftarrow 0$
Random initialization of $\Theta^{(0)} = \{\alpha^{(0)}, \beta^{(0)}, \Sigma^{(0)}, \pi^{(0)}, \tau^{(0)}\}$
do
  Calculate $\mathrm{v}_{n,i,j}^{(k)}$, for each n, i, j, using Equation 4.6
  Calculate $\mathrm{u}_{n,i}^{(k)}$, for each n, i, using Equation 4.7
  Calculate $\alpha_i^{(k+1)}$, for each i, using Equation 4.9
  Calculate $\beta_j^{(k+1)}$, for each j, using Equation 4.10
  Calculate $\Sigma^{(k+1)}$ using Equation 4.11
  Calculate $\pi_i^{(k+1)}$, for each i, using Equation 4.12
  Calculate $\tau_{i,j}^{(k+1)}$, for each i, j, using Equation 4.13
  $k \leftarrow k+1$
while $\mathscr{L}\left(\Theta^{(k)}; X, Y\right) - \mathscr{L}\left(\Theta^{(k-1)}; X, Y\right) >$ tolerance
return $\Theta^{(k)}$

4.3.4 Algorithm Complexity Analysis

In each step of the EM algorithm, we need to calculate Equation 4.6, Equation 4.7, Equation 4.9, Equation 4.10, Equation 4.11, Equation 4.12 and Equation 4.13 from data and current parameter estimates. The dominant operations in these equations are: determinant, matrix inversion and matrix multiplication. The time complexity of each of the first two operations is $O\left(n^{2.373}\right)$, where n is the dimension of the square matrix. The time complexity of multiplying one $[n, m]$ matrix and one $[m, p]$ matrix is $O(nmp)$. The time complexity of each equation performed in a single step of the EM algorithm is given below:

- all $\gamma_{n,i}$: $O\left(M^{2.373} + K^0 NL\left(MD + M^2\right)\right)$
- all $\xi_{n,i,j}$: $O\left(M^{2.373} + K^0 K^1 NL\left(MD + M^2\right)\right)$
- all $\mathrm{v}_{n,i,j}$ (given all $\xi_{n,i,j}$): $O\left(K^0 K^1 N\right)$
- all $\mathrm{u}_{n,i}$ (given all $\gamma_{n,i}$ and all $\xi_{n,i,j}$): $O\left(K^0 K^1 N\right)$
- all α_i: $O\left(K^0\left(D^{2.373} + MD^2 + K^1 NLMD + K^1 NLD^2\right)\right)$

- all β_j: $O\left(K^1\left(D^{2.373} + MD^2 + K^0NLMD + K^0NLD^2\right)\right)$
- Σ: $O\left(K^0K^1NL\left(MD + M^2\right)\right)$
- all π_i: $O\left(K^0N\right)$
- all $\tau_{i,j}$: $O\left(K^0K^1N\right)$

The total time complexity of one step of the EM algorithm is:

$$O\left(M^{2.373} + \left(K^0 + K^1\right)\left(D^{2.373} + MD^2\right) + K^0K^1NL\left(MD + M^2 + D^2\right)\right) \tag{4.16}$$

In our case, $M = 1$, thus we have

$$O\left(\left(K^0 + K^1\right)D^{2.373} + K^0K^1NLD^2\right) \tag{4.17}$$

If we further assume that in practice $D^{0.373} \ll NL$, then the final time complexity is dominated by the second term:

$$O\left(K^0K^1NLD^2\right) \tag{4.18}$$

We conclude that the time complexity of a single step of the EM algorithm is linear in the number of users (N), the number of observations for each user (L), the number of daily behavior patterns (K^0) and the number of activity change patterns (K^1), but quadratic in the number of polynomial terms used in the model.

4.4 Experiments

Dataset. The HealthyTogether dataset contains the calorie expenditure data of 45 users wearing Fitbit (a wearable accelerometer) for 12 consecutive days, starting on Monday. There is one time series for each person and day. Each time series contains 1,440 minute-by-minute measurements of the number of calories burned. The minimum calorie expenditure value per minute is 0.77 (resting metabolic rate). During the weekend the users received an intervention, more concretely, they started using a mobile application that enabled them to participate in physical activities together with a partner, send each other messages, and earn badges. We consider the measurements obtained in the 5 working days before and the 5 working days after the intervention as pre- and post-intervention data, respectively.

Research objective. The goals of our analysis are to (1) understand how the regular daily behavior of different people changed after the intervention, (2) evaluate CLINT's ability to predict the regular daily post-intervention behavior of new people and (3) validate strategies to generate recommendations based on insights obtained from CLINT.

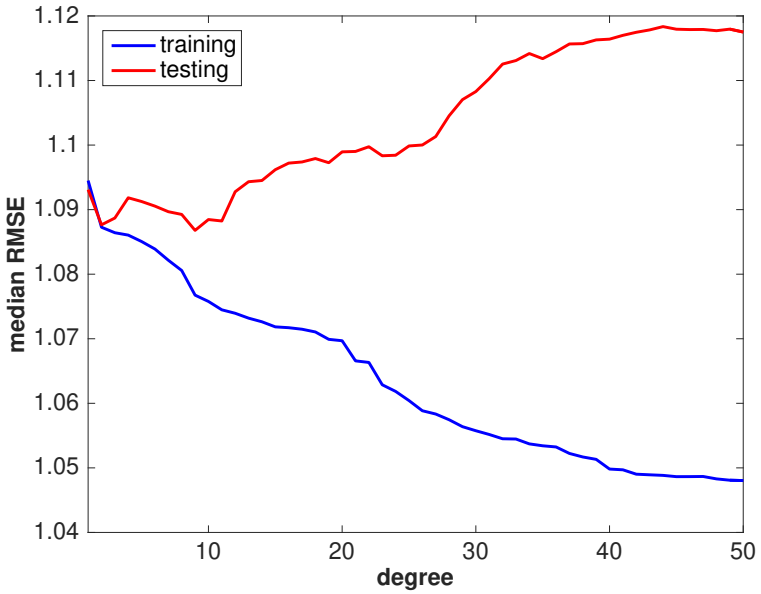


Figure 4.3 – Median RMSE for polynomial regression models with different degrees applied on the post-intervention data.

4.4.1 Evaluation Metrics

Our main task is to predict the regular daily post-intervention behavior (represented as a time series of length 1440) given the raw pre-intervention measurements (represented as a time series of length 5×1440). The n-th user's regular daily post-intervention behavior can be defined as a sequence $\text{post}_n = \{\mathbb{E}[V_{n,t}]\}_{t=1}^{1440}$ where $\mathbb{E}[V_{n,t}]$ is the expected calorie expenditure at minute t relative to the beginning of the day. The expectation is actually the mean calorie expenditure at time t over many days. In our task the amount of data is limited (we have just 5 measurements for each $V_{n,t}$), thus sample averaging would not produce a precise estimate of $\mathbb{E}[V_t]$. As the number of time series increases, their average becomes smoother, contains less extreme changes from one minute to another and approximates the expectation better. We try to approximate the expectation using polynomial regression (a smoothing method) and we learn the optimal degree of the polynomial (smoothness level) using a cross-validation.

We used 5-fold cross-validation to determine the optimal degree. In each round, we trained a separate polynomial regression for each user using data from four post-intervention days and we evaluated the model on the data from the held-out post-intervention day. We calculated the median root-mean-square error (RMSE) on the held-out data so that half of the users had higher RMSE and half of them had lower RMSE than the median. We used the median instead of the mean because the median is more robust against outliers. Figure 4.3 shows the median RMSE on the training and the held-out post-intervention data for polynomial regression models with different degrees. The analysis suggested using 10 polynomial terms to approximate the regular daily post-intervention behavior. Thus, we fit a separate polynomial regression of degree 10 on each user's post-intervention measurements and we considered

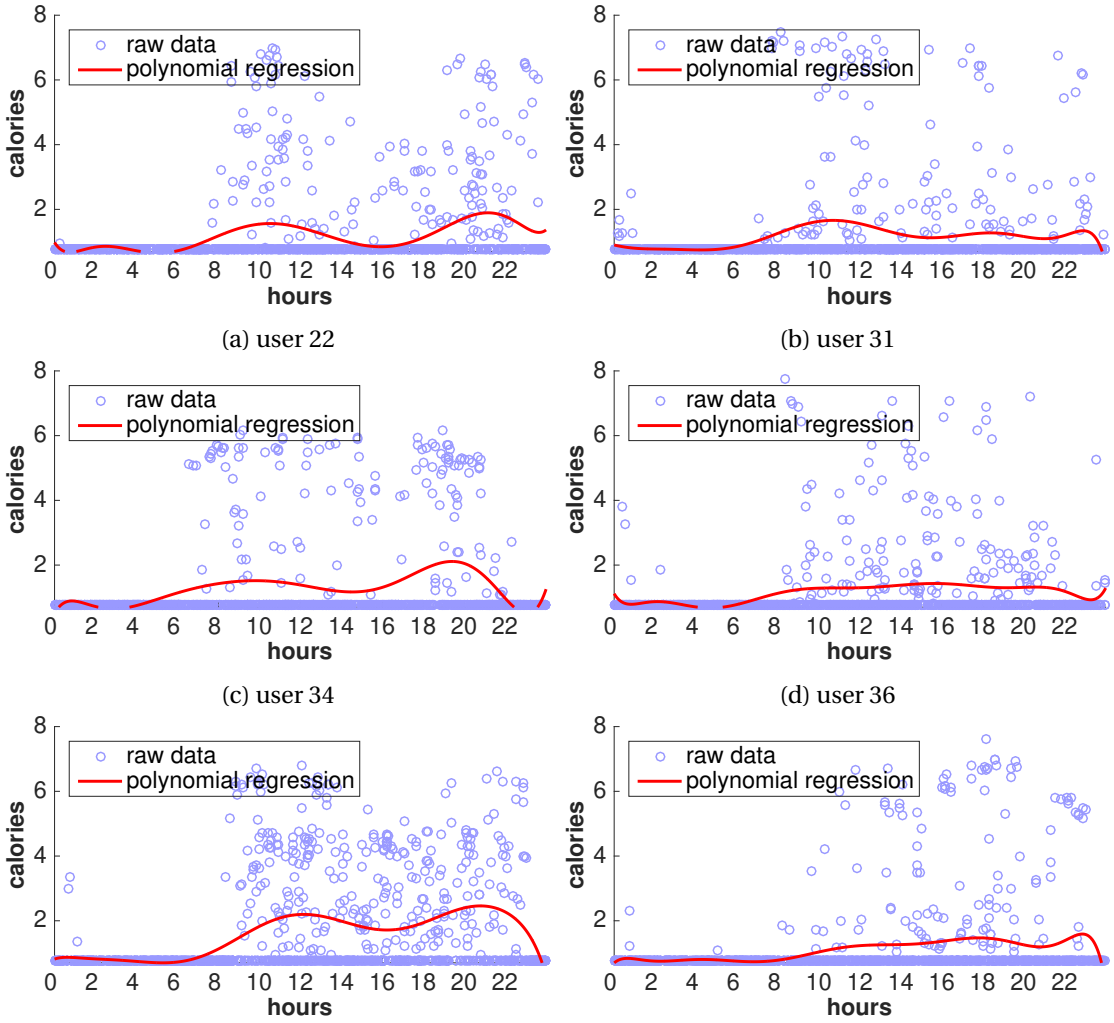
this curve as the user's regular daily post-intervention behavior $\text{post}_n = \{\text{post}_{n,t}\}_{t=1}^{1440}$ (see Figure 4.4). CLINT generates predictions about the regular daily post-intervention behavior of each user $\text{pred}_n = \{\text{pred}_{n,t}\}_{t=1}^{1440}$. We measure the accuracy of the predictions using the mean absolute error (MAE):

$$\text{MAE} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1440}\sum_{t=1}^{1440}\left|\text{pred}_{n,t} - \text{post}_{n,t}\right| \tag{4.19}$$

MAE is conceptually simpler and more interpretable than RMSE [146]. Another relevant metric is the mean bias error (MBE):

$$\text{MBE} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1440}\sum_{t=1}^{1440}\left(\text{pred}_{n,t} - \text{post}_{n,t}\right) \tag{4.20}$$

It represents the tendency of the model to produce higher or lower predictions than the actual observations.



(a) user 22

(b) user 31

(c) user 34

(d) user 36

(e) user 40

(f) user 44

Figure 4.4 – The regular daily post-intervention behaviors of different users represented as separate polynomial regressions. The figures also show random samples of 1,000 raw post-intervention measurements associated to each user.

4.4.2 Discovering Intervention Profiles

Model selection. We applied CLINT on the HealthyTogether dataset to discover the patterns of behavior change for the given intervention. We used grid search with 9-fold cross-validation (so that each fold had equal number of users) to determine the optimal K^0, K^1 and D. We evaluated models with up to five daily activity patterns and up to five activity change patterns. For each fold, we trained 100 models with different initial parameter values using Algorithm 1 and we chose the model with the highest log-likelihood. The results suggested to chose the model with three daily activity patterns, three activity change patterns and ten polynomial terms. The average log-likelihood per observation for the optimal model was -1.5062.

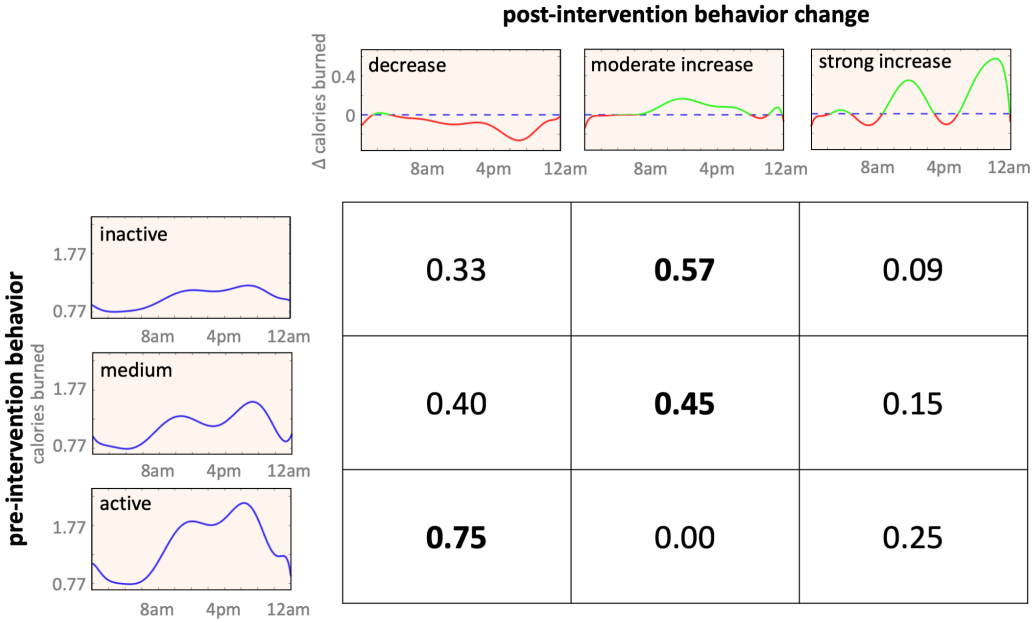


Figure 4.5 – The probability a pre-intervention behavior (left) changes to a post-intervention behavior (up).

The extracted daily activity patterns are shown in Figure 4.5. All of them have two peaks: during the morning and during the evening. However, the intensity of their activity levels is different. The activity patterns represent three different subsets of people: "inactive" (47% of the population), "moderately active" (44% of the population), and "highly active" (9% of the population). We also discovered three different activity change patterns (Figure 4.5). The first pattern represents people who decreased their activity levels through the day, mostly during the evening ("negative responders", 40% of the population). The second pattern represents people who moderately increased their activities through the day, mostly during the morning ("moderate responders", 47% of the population). The third pattern represents people who strongly increased their activities through the day, both in the morning and in the evening ("strong responders", 13% of the population). Chow test [30] indicated that the activity changes after the intervention are significant. The results show that most people increased their activity levels after the intervention. From the transition probabilities in Figure 4.5 we conclude that less active people benefit more from the intervention: 66% of inactive people improved their activity levels after the intervention in contrast to only 25% of the active people. This suggests that the intervention is suitable for the people who need it more.

The people that provided the data were mostly young adults. In Section A.2 of the Appendix, we use CLINT to show that the activity patterns of these people are different than the activity patterns of senior adults. It is difficult to determine whether other patterns may describe the behavior of the population better than the patterns discovered by CLINT because we don't know the ground truth. This is probably less likely because, in another experiment, we showed

that CLINT is able to extract the true behavior patterns from an artificial dataset with a known underlying data-generating mechanism (see Section A.3 of the Appendix).

4.4.3 Predicting Post-Intervention Behavior

We continued the analysis by evaluating CLINT's ability to estimate the regular daily post-intervention behavior for new users from raw sensor data. Estimates are generated in the following way. First, we determine how likely is that the new user behaves according to each activity pattern given his or her pre-intervention data:

$$p\left(b_n = i \mid x_n^0, y_n^0\right) = \frac{\pi_i \gamma_{n,i}}{\sum_{p=1}^{K^0} \pi_p \gamma_{n,p}} \tag{4.21}$$

Then, we use these weights to estimate the regular daily post-intervention behavior at each minute l ($1 \leq l \leq 1440$):

$$\text{pred}_{n,l} = \sum_{i=1}^{K^0} p\left(b_n = i \mid x_n^0, y_n^0\right) \sum_{j=1}^{K^1} \tau_{i,j} \left(\alpha_i + \beta_j\right) x_{n,l}^1 \tag{4.22}$$

We used 9-fold cross-validation and mean absolute error (MAE) between the estimates and the regular daily post-intervention behavior as an evaluation metric. We compared our method with six other methods described below:

- **CLINT-Random**. A variant of CLINT that ignores the pre-intervention data for new users to generate estimates. It assigns an activity pattern by randomly sampling from π (does not use personalization at all) and determines the expected post-intervention behavior conditional on the assigned activity pattern.
- **k-nearest neighbours**. First fits a Gaussian process regression (GPR) model to each user's pre- and post-intervention data to obtain smoothed representations of his or her pre- and post-intervention behavior. Then estimates the Euclidean distance between the smoothed pre-intervention behaviors of the new user and the existing users, and selects the k nearest neighbours. Predicts that after the intervention the new user will behave in the same way as the average smoothed post-intervention behavior of his or her closest neighbours.
- **k-means**. First fits a Gaussian process regression (GPR) model to each user's pre- and post-intervention data. Then uses k-means algorithm with Euclidean distance to obtain clusters of people with similar smoothed pre-intervention behavior. Predicts that the new user will behave in the same way as the average smoothed post-intervention behavior of the people belonging to the same cluster.
- **PolyReg**. Predict that each user will behave in the same way after the intervention. The prediction is represented as a polynomial regression curve that was trained on all raw post-intervention data.

- **LSTM-ED**. Transforms the pre-intervention data into a set of individualized features and then attempts to reconstruct the post-intervention data based on these features. This model uses LSTM [59] for the encoder and the decoder.

- **Clockwork-ED**. Transforms the pre-intervention data into a set of individualized features and then attempts to reconstruct the post-intervention data based on these features. This model uses Clockwork RNN [66] for the encoder and the decoder.

Our method had a better prediction performance (lower MAE) than the baseline methods (Table 4.1). The lower performance of CLINT-Random indicates that the intervention profiles extracted from the training data are meaningful, i.e., CLINT is able to extract useful predictive information from the pre-intervention data. 1-NN did not perform very well, although it used personalization. This is probably because the method is too individualized and does not use insights from the whole dataset to generate better estimates. When we increased the number of neighbours (k-NN), the results were significantly improved over 1-NN. Surprisingly, PolyReg performed very well, although it did not use personalization at all. This can be explained by the fact that it produces more stable estimates of the behavior change using data from all people. k-means does not use post-intervention data in the clustering procedure, and this is probably the reason why it performs worse than our model. Although LSTM-ED is a powerful method, it performs poorly because (1) it requires much more data in the learning process and (2) it is not able to deal with very long sequences. Clockwork-ED has lower RMSE than LSTM-ED because it is able to better model long-term dependencies in the sequences, however, it still requires much more training data to produce accurate estimates. It is both because of the personalization and because of the behavior change modeling that our model performs better than the baseline models. In addition, it doesn't require so much data as the RNN-based methods and it offers interpretable insights helpful to understand better the behavior change of the population after the intervention.

Table 4.1 – Prediction performance of different methods on the X dataset (smaller MAE is better) and the improvement score of CLINT over the baselines.

Method	**MAE**	**CLINT improvement**
Non-parametric		
1-NN	0.247	54.4%
k-means	0.172	7.5%
k-NN	0.171	6.9%
Parametric		
LSTM-ED	0.298	86.3%
Clockwork-ED	0.220	37.5%
CLINT-Random	0.199	24.4%
PolyReg	0.171	6.9%
CLINT	**0.160**	-

4.4.4 Generating Recommendations

Standard recommender systems generate recommendations based on existing users that are similar to the target user. However, these recommendations may not be useful in the health domain. For example, if similar users tend to decrease their physical activity levels over time, the target user would not benefit from their behavioral strategies. This is why personalizable intervention systems for healthy behavior change should base their recommendations on existing users that are not only similar to the target user but also succeeded to improve their health (we call them *responders*).

In our domain, we are interested to recommend strategies that will help target users to improve their physical activity levels. These strategies suggest *when* and *how* the user should change his or her behavior during the day. For example, a recommendation may suggest that the user should be slightly more active in the mornings. An intervention system could base its recommendations on the insights obtained from CLINT. For example, the "inactive" people who improved their behavior are most likely to moderately increase their activity levels. Thus, a recommender system should recommend this behavior change strategy to the "inactive" target users. This recommendation would assist the "inactive" subpopulation in positively changing their behavior. We expect that 33% of this subpopulation would benefit from the recommendations because the same fraction of existing "inactive" users decreased their activity levels without support of a recommender system. We suggest the following way to generate recommendations for new users based on CLINT. First, we monitor the target user and we calculate how likely is that he or she behaves according to each activity pattern given his or her pre-intervention data. For this purpose we use Equation 4.21. Second, we estimate the total benefit when a user changes his or her behavior according to each behavior change pattern j ($1 \leq j \leq K^1$):

$$g(j) = \sum_{l=1}^{1440} \beta_j x^1_{n,l} \tag{4.23}$$

Then, we use the weights and the benefit scores to generate a recommendation about the target user's post-intervention behavior at each minute l ($1 \leq l \leq 1440$):

$$\text{rec}_{n,l} = \left[\sum_{i=1}^{K^0} p\left(b_n = i \mid x^0_n, y^0_n\right) \alpha_i x^1_{n,l} \right] + \left[\sum_{i=1}^{K^0} p\left(b_n = i \mid x^0_n, y^0_n\right) \sum_{j:g(j)>0} \tau_{i,j} \beta_j x^1_{n,l} \right] \tag{4.24}$$

In the recommendation process, we ignore the behavior change patterns that could decrease the activity levels of the target user.

Our recommendations are based on existing users who improved their behavior, thus we expect that new users who improve their behavior are more likely to follow our strategies, in contrast to new users who decrease their behavior. To evaluate our recommendations, we calculate the mean absolute error (MAE) between the recommended post-intervention behavior and the observed regular daily post-intervention behavior separately for the users

who improved (*responders*) and the users who decreased their activity levels (*non-responders*). Formally, a responder user is the one who burned more calories after the intervention than before, i.e.

$$\sum_{l=1}^{L} y_{n,l}^{1} > \sum_{l=1}^{L} y_{n,l}^{0} \tag{4.25}$$

We used 9-fold cross-validation to estimate MAE so that the intervention system was trained only on data from users that are different than the target user for whom we generate recommendations. We compare our recommendations with recommendations from k-NN (this method corresponds to a *collaborative filtering* recommender system). Figure 4.6 shows that responders tend to follow more CLINT recommendations than k-NN recommendations (lower MAE). On the other hand, non-responders' behavior is described better by the k-NN predictions. The large MAE error is due to the tendency of the CLINT intervention system to suggest an increase in the physical activity levels for the non-responders. This explains the large positive bias for the non-responders in Figure 4.7. The amount of bias corresponds to the number of additional calories that would be burnt if the user was walking 30 minutes more during the day than before (168 calories). The bias of the k-NN recommendations for the non-responders is very close to zero, thus, they wouldn't be able to improve target users' physical activity levels. Ideally, a personalizable intervention system should recommend *feasible* and *effective* behavior change strategies. CLINT recommendations are both more feasible and more effective than k-NN recommendations because they describe better the behavior of the people who improve (feasibility) and encourage people who do not improve to increase their activity levels (effectiveness).

CLINT recommendations can be generated for each user, or only for users who are predicted to decrease their activity levels. In the second case, we predict that the n-th target user will decrease his or her activity levels if

$$\sum_{i=1}^{K^0} p\left(b_n = i \mid x_n^0, y_n^0\right) \sum_{j=1}^{K^1} \tau_{i,j} \beta_j x_{n,l}^1 < 0 \tag{4.26}$$

Based on the pre-intervention data, we predicted that 13 users would be non-responders. 6 out of them turned out to be responders. These responders were more likely to follow CLINT recommendations than k-NN recommendations (5/6) and the bias of the k-NN recommendations was positive for the non-responders (0.07). This confirms the previous results and demonstrates that the insights from CLINT are useful to generate feasible and efficient recommendations for healthy behavior change.

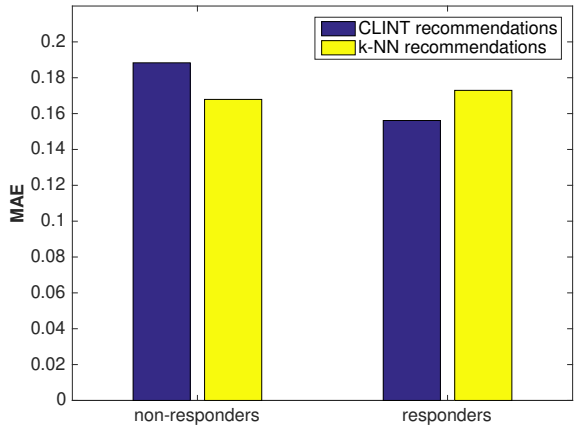


Figure 4.6 – Mean absolute error (MAE) between the suggested strategies and the observed behavior for users that received the recommendations.

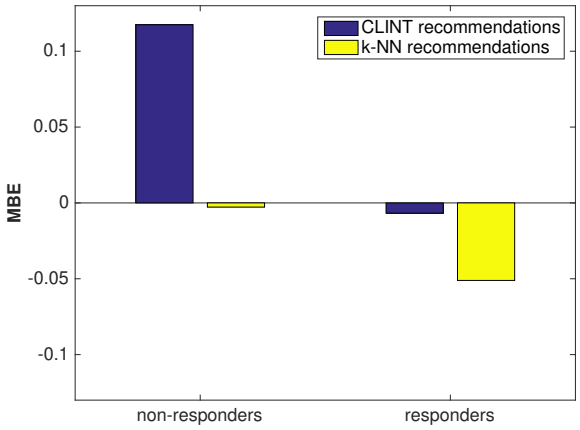


Figure 4.7 – Mean bias error (MBE) between the suggested strategies and the observed behavior for users that received the recommendations.

4.5 Chapter Summary

We presented CLINT — a system to discover the patterns of behavior change for a given intervention from time series data. In addition, our system learns the existing behaviors that explain the potential behavior change after the intervention. This allows us to predict how a new user would behave if he or she received an intervention and generate recommendations that assist the new user in positively changing his or her behavior.

We applied CLINT to calorie expenditure data obtained from 45 people. Although the number of people was limited, we were still able to extract meaningful behavior change patterns from the data. People with different intensity of their physical activities during the day responded differently to the same intervention. We also used the intervention profiles generated by our model to predict the post-intervention behavior of new people and we demonstrated that CLINT generates more accurate predictions than the baseline methods.

We proposed a method to recommend strategies for positive behavior change based on CLINT. Our method learns successful behavior change strategies from the existing users that improved their behavior. The recommended strategies suggest when and how the target user should change his or her behavior during the day. We validated our recommendations and demonstrated that they are feasible and effective.

CLINT discovers the persuasive power of a given intervention. Furthermore, it can predict post-intervention behavior change without actually administering the intervention. The most novel contribution is the discovery of periods when the intervention is the most effective (such as the morning time) due to the fine-grained approach. These insights could support the target user in achieving positive behavior change.

5 Personalizable Intervention System for Senior Adults

5.1 Introduction

The proportion of the global population aged 60 years or over increases rapidly: from 8.5% in 1980 to 12.7% in 2017 [130]. It is projected to rise to 21.3% in 2050 [130]. One of the main reasons for this is the increasing trend of the global average life expectancy: it increased by 5.5 years between 2000 and 2016 (as indicated by the World Health Organization), and is projected to increase by 4.4 years by 2040 [49]. The health of the elderly population is an enormous challenge for the health and social care services [128]. Increased longevity is associated with increases in the number of chronic diseases in the elderly population [98]. The leading contributors to disease burden in older people are cardiovascular diseases (30.3% of the total burden) [103].

Demographic, social, and environmental factors, including physical activity and dietary habits, play a major role in the health and functioning of older adults [39]. Increased physical activity is associated with a lower risk of developing cardiovascular disease when compared to less physical activity [20]. Moreover, it is associated with a clear decrease in the risk of mortality from any cause [76]. Physical activities that improve muscular strength, endurance, and flexibility also improve the ability to perform the tasks of daily living [39]. In this way, physical activity could enhance the quality of life of the older population.

There is evidence that cognitive and behavioral interventions designed to improve physical activity behavior are effective, both for the general [31] and the elderly population [99]. By intervention, we mean an explicit and pro-active recommendation for the purpose of changing the current behaviors of a user. Successful strategies include goal-setting, self-monitoring, feedback, rewards, social support, and coaching [126]. As technology advances, it is becoming easier to integrate new and emerging platforms, software, and devices into behavioral interventions to improve physical activity [77]. Wearable devices can measure heart rate, number of steps, distance, and sleep duration with very high accuracy [138]. Also, activity trackers are

This chapter is based on the work that is submitted to the Special Issue of the Journal of Population Ageing "Responsive engagement of older persons promoting activity and customized healthcare" [69].

becoming more comfortable to wear: this is an important factor for user acceptance. The number of health and fitness apps on the market is growing. Over 318,000 health apps are now available on top app stores worldwide with more than 200 health apps being added each day [2]. These apps help people to change their behavior using different behavior change strategies [38].

It is less investigated how behavioral interventions actually result in responses at the individual levels. We are mainly interested in knowing whether a given intervention can lead to positive and engaged responses and avoid giving harmful interventions. We believe it is possible to scientifically analyze and assess the intervention effect before it is given. As we said before in Chapters 3 and 4, the best intervention for the general population is not likely to be equally effective for each individual. Interventions should be tailored to the individual needs, account for personal levels of fitness, allow for personal control of the activity and its outcomes, and provide for social support by family, peers, and communities [117]. Tailoring variables include time-invariant predictors (e.g., sex), time-variant predictors (e.g., stress) and contextual factors (e.g., weather, day of the week) [101].

In this chapter, we describe our effort in creating a personalizable intervention system that predicts the personalized effects of different behavior interventions on the same user to select one intervention over another, based on historical fitness data. The elderly population would benefit the most from personalizable intervention systems to promote physical activities. We have identified seven main challenges for developing this system:

- **Interventions should be based on successful behavior change strategies.** For all types of interventions, the development process benefits from applying evidence-based theories and techniques because they indicate under which conditions the interventions are effective [89].
- **Predictive models should be trained on data from senior adults.** The aging process leads to a reduction in physical activity level and functional fitness [87]. Thus, the distribution of the fitness data collected from young people and senior adults is different. This is why models trained on data from the general population might not work for the elderly subpopulation.
- **Data collection should not interfere with the normal functioning of the elderly.** This happens when the sensor devices are obtrusive or when the data collection is performed in a laboratory environment. Sasaki et al. have demonstrated that the algorithms developed on free-living accelerometer data are more accurate in classifying activity type in free-living senior adults than the algorithms developed on laboratory accelerometer data [114].
- **Intervention bias in the data should be minimized.** Intervention bias exists when people who receive different interventions are not drawn from the same population. Predictive models trained on a dataset with large intervention bias could underestimate

or overestimate the true intervention effect on the target population. Ideally, training data should be collected from a randomized controlled trial (RCT). If this is not possible, bias-reducing techniques [119] should be applied before the predictive model is built.

- **Data should contain repeated time-varying measurements of the individual's behavior.** Most of the existing works do not consider temporal dynamics as a predictor of behavior change. Human actions vary over time, for example, based on time of day [70]. Kurashima et al. have demonstrated that the time-varying action propensities can be useful to predict the next user action and when this action will occur [70].

- **Predictive models should be able to learn features in an automated way.** Manual feature engineering is both difficult and expensive. This process could generate a large number of features out of which only some are relevant for the predictive task. Also, during manual feature engineering, some important information from the data could be missed, resulting in trained models that have poor prediction performance. This is why features should be learned in an automated way that takes into account the machine learning task of interest.

- **Recommendations should be feasible and effective**. Very ambitious recommendations might cause injuries to the sensitive elderly population. This is why the intervention system should provide evidence that the interventions are likely to cause positive behavior change in the individual before they are suggested to him or her. One of the ways to do this is by ensuring that the predictions for the target user are comparable with the actual responses of existing users that are similar to the target user. If similar people who already received the intervention did not respond to it, this intervention should not be recommended to the target user.

Existing work focuses only on a subset of these challenges. For example, Phatak et al. developed a system that generates recommendations based on the median value of steps/day from the baseline period [101]. Their system does not take into account the distribution of the physical activity throughout the day as a predictor of the behavior change. In contrast, we propose a novel personalizable intervention system that could be used to engage senior adults in daily activities while addressing all the challenges discussed above. The main novelty of our intervention system is that it uses time series fitness data to predict intervention effect.

In this chapter, we consider two different mobile app interventions that aim to promote physical activeness in senior adults. Each mobile application incorporates one of two motivational strategies: self-reflection and social reflection. Under the first intervention, users were able to see real-time step count information only about themselves. Under the second intervention, users were paired up and were able to see real-time step count information about each other. In previous studies, it has been shown that self-monitoring and social support are associated with increased physical activity [54, 99].

Our system requires pre- and post-intervention fitness data from real users. For this purpose,

we designed a randomized trial so that each participant received either a self-reflection mobile app or a social reflection mobile app. Participants were wearing a fitness tracker for three weeks before and five weeks after they received the intervention. We used this data to train the intervention system and perform an offline evaluation. More specifically, we built machine learning methods to predict the change of the physical activity levels after each intervention for new users. This allows the system to decide which intervention should be recommended. The quality of the recommendations could be estimated in online evaluation where one part of users are served by the derived intervention system, and another part of users are given an intervention without taking into account their current behavior. However, in this chapter we focus on the offline evaluation — in the future we plan to perform an online evaluation as well.

5.2 Related Work

Different machine learning methods have been used to gain meaningful insights from health data. Supervised methods have been focused on either detection of health conditions or prediction of health conditions. The former refers to the process of analyzing information to understand the health condition better. The latter refers to the process of analyzing information to predict a health outcome of interest. For example, deciding whether a patient has heart arrhythmia from electrocardiograms is a detection task, but guessing whether the patient will develop a heart arrhythmia in the next year from electrocardiograms and fitness data is a prediction (or *forecasting*) task. Unsupervised methods have generally been used to discover the dominant patterns in the data that explain people's behavior. For example, going to bed early vs. going to bed late.

5.2.1 Detection of Health Conditions

Existing work has focused on classifying activity types from acceleration[114] and body tags data [83], detecting falls from acceleration [4] and body tags data [83], detecting anxiety and depression from socio-demographic and health-related data [115], estimating physical activity levels from questionnaire data [99], estimating body fat from accelerometer data [127], estimating mental health burden from self-reported physical activity data [26], detecting heart arrhythmia from electrocardiograms [106], detecting multiple medical conditions, including diabetes, high cholesterol, high blood pressure, and sleep apnea, from heart rate sensor data [11], etc. Feature engineering was common in most of this works. For example, Sasaki et al. extracted time- and frequency- domain features from acceleration signals in 20-second windows. This process requires a lot of effort and may often result with features that are not predictive of the target variable. Andrew Ng stated that "coming up with features is difficult, time-consuming, requires expert knowledge" [96].

Deep learning techniques are able to automatically extract relevant features and they are especially useful when the data has a complex nature. Rajpurkar et al. used convolutional

neural networks to detect irregular heart rhythms from electrocardiograms better than a cardiologist [106]. In terms of data and feature extraction, our work is most similar to the work of Ballinger et al. [11]. They used sequence-to-sequence autoencoder to learn features from step count and heart rate time series data. These features were more useful than hand-engineered biomarkers derived from the medical literature to detect diabetes, high cholesterol, high blood pressure, and sleep apnea. In our task, we also use sequence-to-sequence autoencoder to learn features from step count time series data. Our work differs from [11] in that we use longer user history to make a prediction and we have a limited amount of data to train the model. Also, we try to predict (forecast) intervention effect. This makes the data collection phase more challenging.

5.2.2 Prediction of Health Conditions

Predicting health conditions is an important task because it can turn data into actionable insights. Much work has focused on early prediction of future health conditions that could be used in preventive healthcare. These works include: predicting future cognitive impairment in senior adults from variables, which are commonly collected in community health care institutions [95], predicting mortality in senior adults from medical history, diet, exercises and lifestyle activity [76], predicting mortality in older women from mean daily step counts [74], predicting changes in exercise behavior from historical physical activity data [65], predicting daily blood pressure levels from historical blood pressure and health behavior [27], predicting in-hospital mortality, readmission, prolonged length of stay and final discharge diagnoses from electronic health records data [105], predict future actions from past activities [70], etc. The early prediction information can be useful for a personalizable intervention system to decide *when* it needs to act, however, it is not sufficient to decide *how* to act. Expert knowledge is usually needed to choose a suitable preventive intervention based on these predictions.

Another line of work has applied data-driven approach to understand and predict the effect of interventions on people's health [123, 107, 142, 101]. The most traditional method estimates the population-level intervention effect from randomized controlled trial (RCT) data [123, 107]. The problem with this method is that not all people respond to the same intervention in the same way. The optimal intervention depends on the individual's characteristics — this aligns with the goals of personalized medicine [12]. Zeevi et al. demonstrated that people eating identical meals present high variability in post-meal blood glucose response [142]. They developed a machine learning algorithm that uses information about blood parameters, dietary habits, anthropometrics, physical activity, and gut microbiota to predict personalized blood glucose response to real-life meals. These predictions were used to design personalized diets composed of the meals predicted by the algorithm to have low post-meal blood glucose responses. They performed both offline and online evaluation and showed that their dietary interventions improve multiple aspects of glucose metabolism.

In another work, Phatak et al. developed a system to deliver personalized daily step goals

that aimed to improve people's physical activity levels [101]. For this purpose, they performed a data collection study with people wearing activity tracker for 14 weeks. Baseline physical activity measured in weeks 1–2 has been used to inform personalized daily step goals delivered in weeks 3–14. They learned a regression model that predicts daily step count given few different variables including people's baseline median daily step count and current daily step goal. In contrast to [101] and [142], we are interested to generate predictions based on much more fine-grained baseline data, such as the minute-by-minute step count measurements. This requires machine learning methods that are able to extract temporal patterns from time series data which are relevant to predict intervention effect. Also, our work differs from [101] in that our work focuses on improving physical activity levels in senior adults — there is no other group in our society that can benefit more from physical activities [46].

Building a personalizable intervention system for responsive engagement of senior adults in daily activities is a challenging problem. One of the main reasons is that it is very costly to perform a study providing evidence about intervention effectiveness. Another reason is that existing personalized models for predicting intervention effect cannot learn relevant insights when they are directly applied to frequently-sampled fitness data from a limited number of people. In our research, we provide a complete solution to this problem — from data collection to recommendation generation. We show the benefit of learning representations from time series data in predicting intervention effect. This allows the intervention system to select better one intervention over another.

5.3 Intervention System

Our intervention system pipeline consists of four phases: data collection, representation learning, predictive modeling and recommendation generation. The first phase provides time series sensor data containing evidence of the intervention effectiveness. The second phase reduces the dimensionality of each time series while preserving information about its underlying temporal dynamics. The third phase builds machine learning models that are able to predict how an elderly person would respond if he or she was given an intervention. These models take as input the representations learned in the previous phase. The fourth phase selects and recommends an optimal personalizable intervention based on the predictions generated in the third phase.

5.3.1 Data Collection

The first step towards building a personalizable intervention system was to collect sensor data from senior adults who received an intervention. We used data from an experiment conducted in Eindhoven that included 55 senior adults aged 65+ years wearing a Fitbit Flex 2 wristband for eight weeks. The Fitbit device recorded the number of steps performed in each minute. After the first three weeks, each participant received one of two mobile app

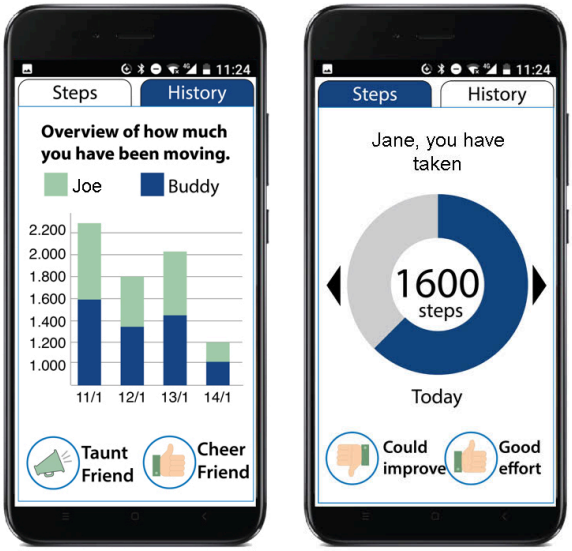


Figure 5.1 – Mobile app interface

interventions by random assignment[1]. Under the first intervention (*self-reflection*), users were able to see real-time step count information only about themselves. Under the second intervention (*peer-to-peer*), users were paired up and were able to see real-time step count information about each other. The app had been used for five weeks, until the end of the trial. The user interface[2] of this mobile app is given in Figure 5.1.

Participants were instructed to wear the Fitbit at all times during the trial. However, it was not possible to know for sure whether they were always wearing the device or not. Only the days with a positive number of steps were counted as valid days of data. We filtered out participants who didn't have at least one valid day of data for each different day of the week, both before and after the intervention. Also, we filtered out one participant who had an increase of his or her average daily step count by 145% after the intervention (an *outlier*). 49 participants remained and the data associated with them were included in our analysis. Out of these people, 14 received the first intervention, "self-reflection", and 35 received the second intervention, "peer-to-peer". The average daily step count per day in the trial is given in Figure 5.2. It can be observed that people manifested periodic weekly behavior and they were the least active during the weekends. In 18.5% of the minutes there was at least one step performed. There was no significant difference between the two intervention groups in terms of their pre-intervention average daily step counts (two-sample t-test for the null hypothesis that the two samples have identical average values, p-value=0.838245). Figure 5.3 shows that users varied a lot in terms of their average pre-intervention (post-intervention) daily step count: the least active ones performed 2,500 steps, and the most active ones performed 18,000 steps per

[1]The app was jointly developed with a research team at Eindhoven University of Technology. The trial was conducted by a research team at Eindhoven University of Technology.

[2]Photo courtesy of Carlijn Valk at Eindhoven University of Technology.

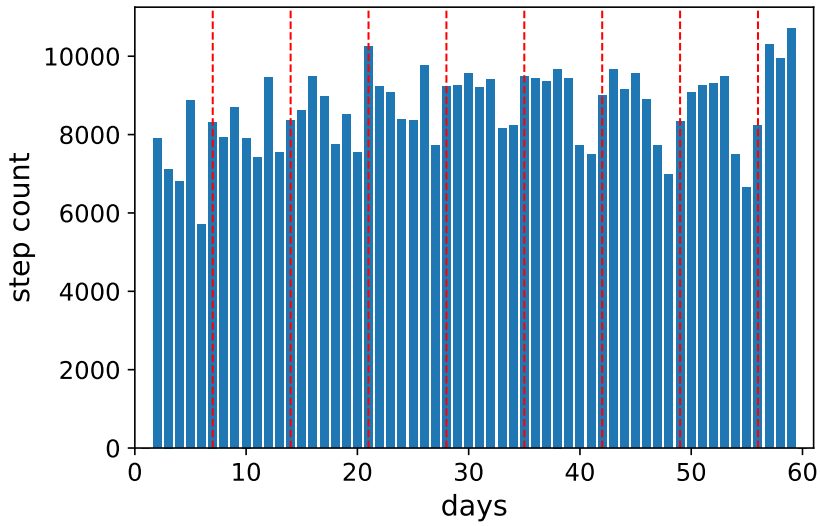


Figure 5.2 – Average daily step count per day in the trial (only valid days of data were included in the estimate). Red dashed lines indicate the beginning of each week (Monday).

day on average.

We define the *absolute improvement* of the user i as the difference between his or her post-intervention average daily step count post_i and his or her pre-intervention average daily step count pre_i. We define the *relative improvement* of the user i as the relative increase of his or her post-intervention average daily step count compared to his or her pre-intervention average daily step count:

$$\text{relative improvement}_i = \frac{\text{post}_i - \text{pre}_i}{\text{pre}_i} \tag{5.1}$$

Increase of 1,000 steps might not be much for someone who performs 18,000 steps per day, but might be for someone who performs 2,500 steps per day. This is why we are more interested in the relative improvement. We assume that the relative improvement is a proxy for the individual intervention effect. The peer-to-peer group improved more on average than the self-reflection group. The improvement was significant for the peer-to-peer group (8.1%, one-sample t-test for the null hypothesis that the mean is positive, p-value= 0.005409), but not for the self-reflection group (2.1%, one-sample t-test for the null hypothesis that the mean is positive, p-value= 0.277266). The main task of the intervention system is to predict the individual's relative improvement under each of the two interventions given his or her pre-intervention data. This allows the system to select and recommend the optimal intervention for each individual.

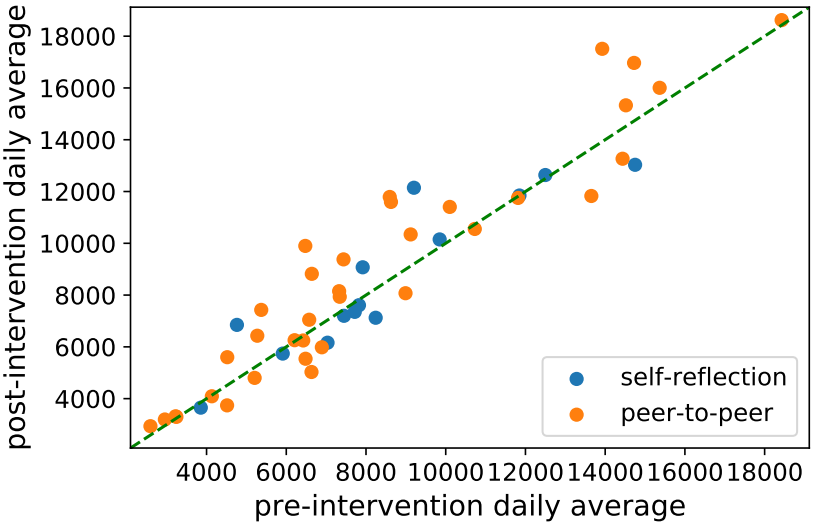


Figure 5.3 – Pre- vs post-intervention average daily step count per user (only valid days of data were included in the estimate).

5.3.2 Representation Learning

Fitbit provides 1,440 step count measurements per day. This granularity makes it difficult for predictive models to gather insights from data. Most prior works [127, 101, 123, 74] used aggregated step count data in their analysis. In contrast, we were interested to extract much more information explaining the time series dynamics that could be further used for the prediction task. We used RNN (Recurrent Neural Network) autoencoder to generate embeddings that preserve the low-level information contained in the daily time series as much as possible. These embeddings were used directly by the predictive models described in the next section.

The autoencoder consists of two parts: an *Encoder* and a *Decoder*. The Encoder processes the input time series and produces a low-dimensional embedding. The Decoder tries to reconstruct the original time series given its embedding as input. We model both the Encoder and the Decoder using RNN. The most popular kind of RNN is built using LSTM (Long short-term memory) units [51]. However, in our model, we use Clockwork RNN because LSTM RNN performs worse than Clockwork RNN in time series reconstruction [66]. Clockwork RNN is an architecture in which the hidden layer is partitioned into separate modules, each processing inputs at its own temporal granularity, making computations only at its prescribed clock rate [66]. As a consequence, long-term information propagates faster through the network.

853 time series collected before the intervention and associated to valid days of data were used to train the model. First, we pre-processed the data using aggregation and data transformation. The aggregation step included segmenting each sequence into 10-minute non-overlapping sliding windows and summing up the minute-level step counts belonging to the same window. In this way, we reduced the length of the time series from 1,440 to 144, without losing much

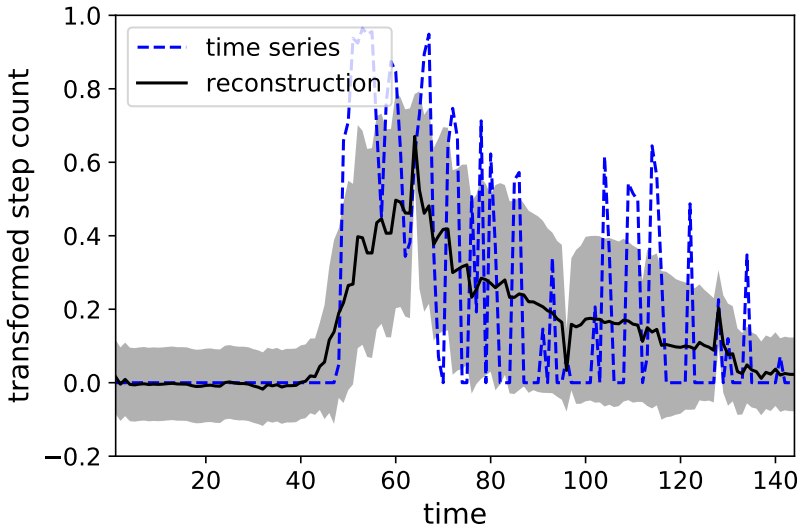


Figure 5.4 – Mean and standard deviation of a time series reconstruction generated by the RNN autoencoder. The dashed blue line represents a sample daily time series given as input.

information about the distribution of the physical activities during the day. Also, in this way, we improved the balance between the observations indicating no activity and the observations with a positive step count. The distribution of the observations with a positive step count was skewed to the right. Thus, we used box-cox transformation [113] to transform these data into a more normal distribution.

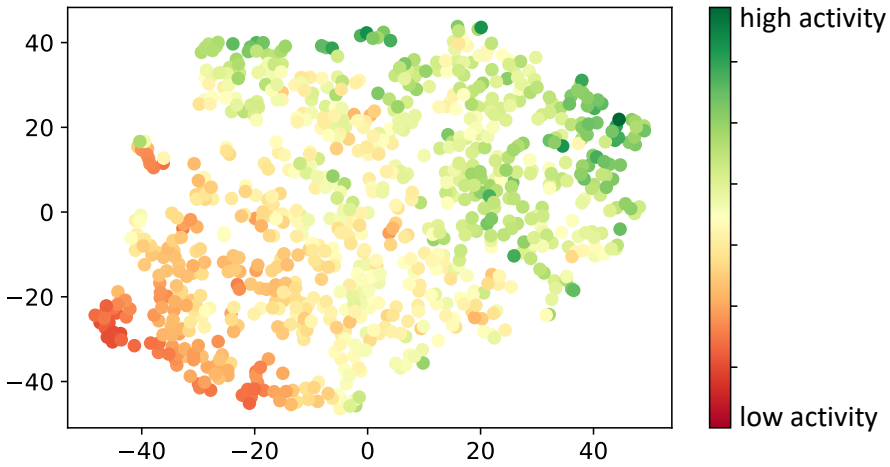


Figure 5.5 – 2D visualization of the time series embeddings generated by the Encoder. The embeddings are visualized using t-SNE. Each point represents a time series. Closer points indicate similar time series.

Our autoencoder tries to learn parameters of a Normal distribution (mean and standard deviation) for each time step of the Decoder so that it is more likely that the observations are generated from the associated distributions. We specified a minimum value for the

standard deviation and we used weight regularization to reduce overfitting. Figure 5.4 shows a reconstruction of a sample time series. It can be seen that the Decoder was able to capture the high activity in the middle and the low activity in the second part of the time series. We used an embedding space of 10 dimensions.

For visualization purposes, we further reduced these embeddings into two dimensions using t-SNE [85]. This allowed us to see each time series as a point in a two-dimensional space and to visually validate the embeddings generated by the autoencoder. In Figure 5.5, we see that embeddings that are closer to each other represent time series that have similar step counts. In addition, these time series have a similar distribution of the physical activities during the day. This demonstrates that the Encoder has learned to embed similar time series into similar vectors.

5.3.3 Predictive Modeling

The main component of our intervention system utilizes the pre-intervention sensor data to predict how a new senior person would respond if he or she was given the intervention. For this purpose, we trained supervised machine learning models separately on the data from the people that received the self-reflection intervention and the people that received the peer-to-peer intervention. Our models used the features extracted from the pre-intervention time series data to predict his or her *relative improvement*. The main idea behind predictive modeling is that people who behave similarly will respond to the same intervention in a similar way.

Before we apply predictive modeling, we needed to deal with missing data; 17% of the days of pre-intervention data were invalid i.e. did not contain any activity at all. Most machine learning models do not support missing data as input, thus we decided to replace the missing values with an estimate. There are correlations in the data that could help us choose a more relevant estimate. An important observation is that users differ in terms of their activity levels, but maintain consistent and periodic behavior from one day to another. We used an imputation method that replaces the missing data by a random time series from the valid days of data generated by the same user on the same day of the week. Alternative data imputation methods are to replace the missing time series with user average, or to generate time series using deep learning. User average is a simple, but unsuitable method because it produces smoothed time series whose distribution is different from the distribution of raw sensor time series data. Deep learning techniques are unsuitable as well because we don't have enough data to learn to generate realistic time series from the conditional distribution.

Using the whole pre-intervention data as a predictor in our machine learning models means that when we deploy the intervention system, it needs to observe a new user for 3 weeks before it decides which intervention is better for him or her. Ideally, the user should be observed for as short period as possible. Thus, we were interested in the minimum amount of pre-intervention data that we could use to predict an individual's relative improvement under

each intervention. In our experiments, we applied models that take as input either one day or one week of data. The number of data samples per user was 18 in the first case (one for each day) and 3 in the second case (one for each week).

We scaled our output variable (relative improvement) so that its variance was one. We used root-mean-square error (RMSE) to measure the error of our models in predicting the scaled relative improvement:

$$\text{RMSE} = \sqrt{\frac{\sum_i^N \left(\text{relative improvement}_i - \text{prediction}_i\right)^2}{N}} \tag{5.2}$$

where N is the number of data samples, $\text{relative improvement}_i$ is the scaled relative improvement of the user associated to the i-th data sample, and prediction_i is a prediction generated by our method, based on the pre-intervention data associated to the i-th data sample. The generalization ability of our methods was estimated using 10-fold cross-validation. We ensured that data from the same user belonged to the same fold. In this way, the model was tested on users whose data was not used in the training process. We repeated each cross-validation 10 times with a different random partition each time to obtain the mean and the variance of the test error.

We used six different models to predict the relative improvement under each intervention. Five of them were based on ridge regression. In ridge regression the prediction is a linear function of the feature vector x_i:

$$\text{prediction}_i = \mathrm{x}_i^T \mathrm{w} + b \tag{5.3}$$

where w denotes the regression coefficients and b is the intercept term. In the learning phase we minimize the following objective function:

$$\frac{1}{N}\sum_{i=1}^{N}\left(\text{relative improvement}_i - \text{prediction}_i\right)^2 + \lambda \mathrm{w}^T \mathrm{w} \tag{5.4}$$

where λ is a hyperparameter indicating the regularization strength. Larger λ forces weights to decay more towards zero, so $\lambda = 0$ means that we do not do any regularization at all. This hyperparameter is determined using cross-validation. Below we summarize the predictive models that we used to predict the relative improvement under each intervention:

- **Const.** The simplest model that could be used is to predict that new people would improve according to the mean improvement of the existing people who already received the intervention, i.e, a ridge regression where $\text{prediction}_i = b$. Thus, this model does not base its predictions on the pre-intervention data at all. It was expected that this model would score RMSE ≈ 1 because the standard deviation of the output variable is 1.
- **DayAgg.** This model is a simple ridge regression that uses one day of data to generate predictions. It does not use the whole time series as a predictor, but only the total (or

aggregated) daily step count, i.e., x_i is one-dimensional vector.

- **DayEmb.** This model also uses one day of data to generate predictions. Ridge regression is used to generate predictions. In contrast to model DayAgg, it uses the whole time series as a predictor. More specifically, the feature vector contains the time series embeddings provided by the RNN autoencoder in addition to the total step count, i.e, x_i is 11-dimensional vector: the first 10 features represent the time series embeddings and the last feature represents the total step count.

- **WeekAgg.** This model is a ridge regression that uses one week of data to generate predictions. It takes as input a set of features that represent the total daily step counts for each different day of the week, i.e, x_i is 7-dimensional vector.

- **WeekEmb.** This model is a ridge regression that also uses one week of data to generate predictions. It considers the input data as a set of embeddings (plus the total daily step count associated with each embedding). The total number of input features is 77: the first 70 correspond to the embeddings associated to each different day of the week and the last 7 indicate the total daily step count for each different day of the week.

- **WeekEmbRNN.** This model is a RNN that takes the same input as the model WeekEmb. It considers the input data as a sequence of 7 embeddings (plus the total daily step count associated with each embedding). In this way, it tries to extract relevant predictive information from both the temporal dynamics within a single day and the temporal dynamics from one day to another. In contrast to Clockwork RNN that we used as an autoencoder, here we used RNN with LSTM units because the sequences are very short in length (7 time steps, one for each different day of the week). In the learning procedure we minimized the RMSE.

The test errors for each model applied separately on the data from the self-reflection and the peer-to-peer group are given in Figure 5.6. It can be seen that personalized models (these are all except model Const) applied on the data from the peer-to-peer group performed the same as the sample mean estimator (model Const). This means the either (1) one week of minute-by-minute data does not contain enough information to explain the individual response to the peer-to-peer intervention, or (2) we don't have enough data to learn the individual response. On the other hand, two of the personalized models (model DayEmb and model WeekEmb) applied to the data from the self-reflection group performed much better than the sample mean estimator (model Const). Both models were applied to fine-grained time series data. This means that the low-level information contained in the daily time series data is an important predictor of the intervention effect. Model WeekEmb performs better than model DayEmb. This suggests that the information about the higher-level human behavior in different days of the week contains relevant predictive information as well. The same conclusion holds true for the youth population (see Section A.4 of the Appendix). It is interesting that model WeekEmbRNN does not perform better than model WeekEmb although it takes the same input as model WeekEmb, but is more complex. This can be explained by the

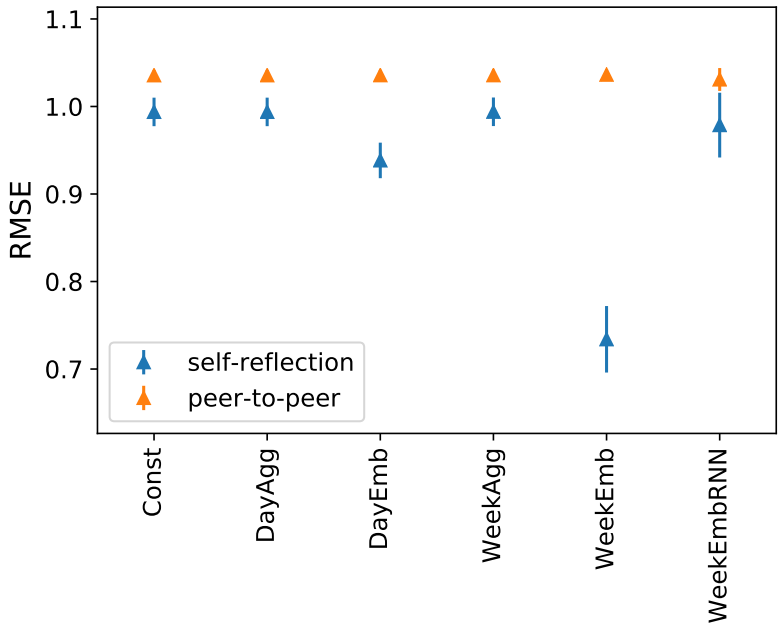


Figure 5.6 – Comparison of the test error of different models.

fact that model WeekEmbRNN has a large number of parameters, but there is a small amount of training data available. In Section A.5 of the Appendix, we demonstrate that prediction accuracy improves as we increase the amount of training data.

We continued the analysis by inspecting the predictions generated by the optimal model WeekEmb for the self-reflection group on the test set (see Figure 5.7). There is a significant positive correlation between the true relative improvements and the predictions (p-value< 0.000001). We were also interested to know how accurate would be the continuous predictions if we used just their sign (positive or negative) to predict whether the user will increase or decrease his physical activities after the intervention (a binary prediction). Since 64.29% of the participants who received the self-reflection intervention did not improve their physical activities, the simplest baseline method would predict that every user would not improve. In this way the accuracy of the method would be 64.29%. If we generated predictions about the improvement using our optimal model, but just care about the sign of the predictions, then we would obtain accuracy of 65.36% — not much different than the baseline method. However, if we define a threshold, and we predict the direction (or sign) of the behavior change only if the absolute value of the prediction is larger than this threshold, the accuracy improves. This can be seen in Figure 5.8. When the threshold is 0.6 and above, we obtain accuracy of more than 87% — much better than the baseline method. In other words, when the predictions have higher absolute value, we are more certain whether the user will increase or decrease his activities after the intervention. In practice, this means that we could choose a subset of people that are more likely to benefit from the intervention and give the intervention only to those people. However, there is a trade-off: larger subset means lower certainty in the sign of

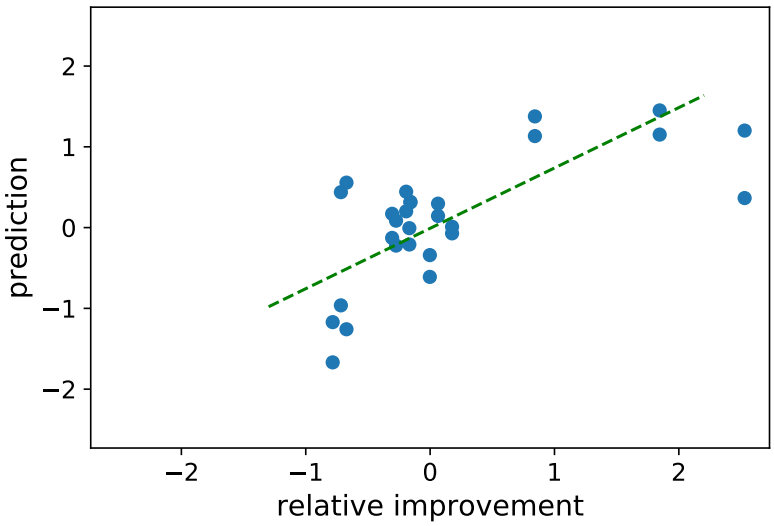


Figure 5.7 – True relative improvement vs. predicted improvement for the self-reflection group. There are two data points per each user. The dashed green line shows the trend.

the improvement.

5.3.4 Generating Recommendations

The best predictive model (model WeekEmb) is used to predict the potential improvement of a new user under each intervention. The personalizable intervention system chooses the intervention that is associated with a higher improvement and recommends it to the user. To evaluate the recommendations, we analyzed whether the improvement of a new user that received one of the two interventions is better predicted by the model that is trained on existing users that received the same intervention as the target user. We discovered that the RMSE is smaller when we apply the correct model. This observation strengthens the results from the previous section. Our intervention system would give the peer-to-peer intervention to 75% of the people that actually received the self-reflection intervention and the self-reflection intervention to 25.7% of the people that actually received the peer-to-peer intervention. This means that, although the peer-to-peer intervention is more beneficial for the general population, it is likely that 25% of the people would still benefit more from the self-reflection intervention. Directly comparing the predictions about the potential improvements of the same user under different interventions is analogous to comparing the causal effects of both interventions on the user's relative improvement.

The intervention system could be designed not to generate any recommendation at all if it predicted that the target user would worsen his or her activity levels under any available intervention. It could also be designed to consider only recommendations that are more likely to cause positive improvement i.e. interventions that are associated with predictions larger

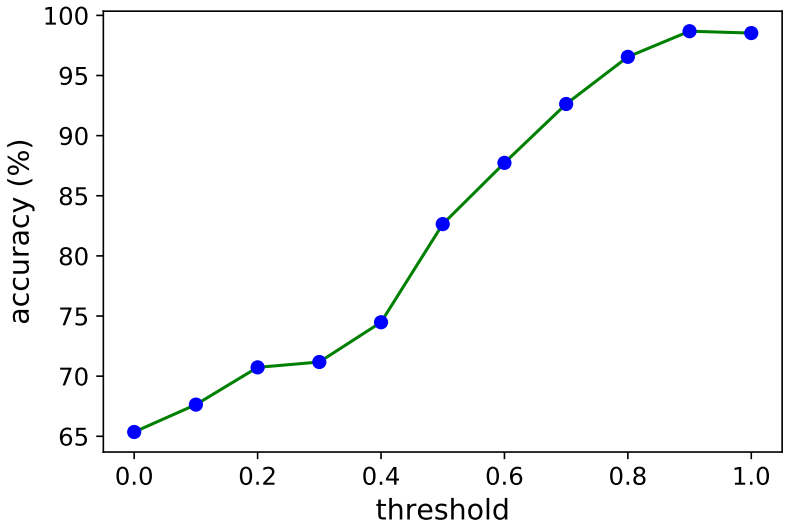


Figure 5.8 – Percentage of people in the self-reflection group for whom we would predict correctly the direction of the improvement (positive or negative) if we apply the predictions only when their absolute value is larger than a threshold.

than a threshold (Figure 5.8). This is especially important for the vulnerable elderly population. The proposed design could also be useful when there is a limited number of interventions and a large population, because in this case the interventions could be given only to the people that are most likely to benefit from them.

5.4 Chapter Summary

In this chapter, we proposed a novel personalizable intervention system that aims to promote physical activeness in senior adults. The main novelty of our intervention system is that it uses time series fitness data to predict the intervention effect. These predictions allow the system to select better one intervention over another. We trained the system using data from real senior adults. In our experiment, we used two different mobile app interventions. The first intervention (self-reflection) allowed the user to see a real-time step count information about himself or herself. The second intervention (peer-to-peer) allowed the user to see a real-time step count information about himself or herself and his or her partner. Although the peer-to-peer intervention was more beneficial for the general elderly population, we demonstrated that 25% of the senior adults are still likely to benefit more from the self-reflection intervention. Our personalized predictive models were able to discover who are these people based on their pre-intervention behavior. We showed that models that utilize fine-grained sensor data from a longer period (one week) perform better. This suggests that both the lower-level human behavior within a single day and the higher-level human behavior from one day of the week to another are important predictors of behavior change in senior adults.

The following use case demonstrates the use of our intervention system. Consider a senior adult who wants to become more physically active, but doesn't know how to achieve that. He or she installs our recommendation app that implements our personalizable intervention system and starts wearing an unobtrusive fitness tracker. After one week, the recommendation app learns his or her behavior patterns and recommends him or her to start using either the mobile app that incorporates self-reflection or the mobile app that incorporates social reflection to promote physical activeness. The target user installs the recommended app and improves his or her physical activity levels over time.

In the beginning of our research, we identified several main challenges for developing a personalizable intervention system whose main purpose is to select those interventions that are most likely to work for senior adults. In our intervention system design and implementation, we ensured that our system addresses all these challenges. First, we used mobile app interventions based on two motivational strategies that were shown to be successful in the scientific literature: self-reflection and social reflection. Second, we trained our predictive models using only data from senior adults. In this way, the models learned behavioral patterns that are characteristic of this vulnerable subpopulation. Third, we collected data from a trial in which participants were wearing a smart wristband that tracked their activities without interfering with their normal functioning. Fourth, our machine learning models utilized randomized trial data that allowed them to make a more relevant comparison between the different interventions. Fifth, the generated predictions are based on how existing people that are similar to the target user responded to the same intervention, thus, there is evidence about the effectiveness of the recommendations. Finally, our intervention system utilizes fully the frequently-sampled time series data and learns relevant predictive information from it in an automated way, without human interference.

Our intervention system is scalable and fast to train. It also supports multi-variate time series and multiple interventions. For example, the same physical activities may result in different heart rates in different people. Thus, heart rate time series could be an important predictor for behavior change besides step count time series. Other time-invariant predictors (e.g., sex) and contextual factors (e.g., weather) could also be used to explain the behavior change. Our intervention system supports both simpler and more complex predictive models, such as deep learning. The latter could generate more accurate predictions. However, deep learning methods require a large amount of data. We showed that LSTM performs much worse than ridge regression on our dataset from a limited number of users.

Personalized recommendations for increased physical activity are of great practical value for senior adults. We believe that our system for personalized recommendations is an important contribution in this field because it learns relevant predictive information from unlabeled time series sensor data that is easy to collect, in an automated way. In our future work, we plan to perform an online evaluation of our intervention system.

6 Reducing Intervention Bias using Adversarial Balancing

6.1 Introduction

Data from randomized controlled trials allow personalizable intervention systems to make less biased estimates of the intervention effect. However, Randomized Controlled Trials (RCTs) are often expensive or unfeasible to conduct [40]. Also, observational studies are relatively quick, inexpensive, and easy to undertake, compared to randomized controlled trials [52]. In this chapter, we are interested in estimating the individualized treatment effect (ITE) for a new patient from observational data.

With the advance of technology, the amount of observational data increases rapidly. For example, Electronic Health Records consist of longitudinal patient health data, including demographics, diagnoses, procedures, and medications [84]. Understanding whether a patient would benefit from a particular medical procedure or medication — based on historical EHR data — aligns with the goals of personalized medicine. Also, understanding whether a patient would improve his or her behavior after an intervention aligns with the goals of preventive healthcare.

This problem has two main challenges. First, only one outcome is observed. Each patient either received or did not receive the treatment, so we don't know his or her potential outcome in the opposite case. Second, in observational studies there could be a treatment bias, meaning that some patients are more likely to receive the treatment. For example, a doctor might prescribe medicine to a patient based on his or her laboratory tests. Treatment bias makes it difficult for traditional supervised learning methods to infer causal relationships because the distribution of covariates across different treatments is not the same. Thus, a predictive model trained on data from patients who received the treatment will not generalize well to patients who did not receive the treatment and vice versa.

It is possible to identify the ITE in observational studies if the collected data contains all the confounders: factors that affect both the treatment assignment and the outcome. This is a common assumption that we also make in this chapter. Similarly to [119], we formulate the

problem as a counterfactual regression (CFR). In other words, we try to estimate the counterfactual (unobserved) outcomes given the observed (and possibly *biased*) data. Johansson et al. perform CFR by simultaneously learning balanced representations of the covariates across different treatment groups and predicting the factual (observed) outcomes. Balancing is a form of regularization that provides more robust counterfactual predictions; it assigns less importance to features whose distribution is different across the treatment groups and it reduces the treatment bias. Although some features are imbalanced, they could contain useful predictive information. Thus, CFR makes a tradeoff between minimizing the predictive loss and the imbalance loss.

This chapter builds on [119]. We are also interested to simultaneously learn balanced representations and predict factual outcomes. The novelty of our method is that we learn balanced representations in an adversarial way. We call our method ACFR referring to Adversarial balancing for CounterFactual Regression. ACFR contains a Discriminator component that tries to distinguish between the patients who received the treatment and the patients who did not receive the treatment based on their latent representation. In this way we are able to generate more balanced representations, resulting in improved or matched performance on two benchmark datasets.

6.2 Problem setup

Suppose there are N patients in the observational dataset $\mathscr{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathscr{X}$ denotes the i-th patient's feature vector, $t_i \in \{0,1\}$ indicates whether he or she received a treatment, and $y_i \in \mathscr{Y}$ denotes the outcome of interest. Let y_i^1 and y_i^0 denote the potential outcome if the i-th patient received and not received a treatment, respectively. For each patient only one potential outcome is observed:

$$y_i = t_i\, y_i^1 + (1 - t_i)\, y_i^0 \tag{6.1}$$

This is the fundamental problem of causal inference. The setting is known as the Neyman–Rubin causal model [111]. We call y_i the *factual* outcome, and refer to the unobserved potential outcome as a *counterfactual* outcome. We are interested to learn the *individualized treatment effect* (ITE) for a new patient x:

$$\tau(x) = \mathbb{E}\left[Y^1 - Y^0 \mid X = x\right] \tag{6.2}$$

The following two assumptions about the treatment assignment are sufficient for the ITE function to be identifiable:

1. **Unconfoundedness**. Treatment assignment is independent of the outcomes:

$$\left(Y^0, Y^1\right) \perp T \mid X \tag{6.3}$$

2. **Overlap**. Each patient has non-zero probabilities to receive or not receive a treatment:

$$0 < \mathbb{P}(T = 1 \mid X = x) < 1 \tag{6.4}$$

The ITE estimate for the i-th patient is:

$$\hat{\tau}(x_i) = f(x_i, 1) - f(x_i, 0) \tag{6.5}$$

where

$$f(x, t) \approx \mathbb{E}\left[Y^t \mid X = x\right] \tag{6.6}$$

Our goal is to learn a function $f(x, t)$ using the observed sample $\mathscr{D} = \left\{\left(x_i, t_i, y_i\right)\right\}_{i=1}^{N}$ that can be used to predict the unobserved counterfactual outcome $y_i^{1-t_i}$. In a randomized study the treatment assignment does not depend on x, i.e., $\mathbb{P}(T = t \mid X = x) = \mathbb{P}(T = t)$, so the factual and the counterfactual distribution are the same. This allows supervised learning algorithms to generate unbiased estimates of the counterfactual outcome. However, in observational study, the treatment assignment depends on x, making the problem more difficult. Ignoring the difference between the factual and the counterfactual distribution would produce biased estimates of the counterfactual outcome.

If the outcome is a continuous variable, then the performance of the generator could be measured using two different metrics: the absolute error in estimated Average Treatment Effect (ATE) and the expected Precision in Estimation of Heterogeneous Effect (PEHE). The absolute error in ATE is calculated in the following way:

$$\epsilon_{ATE} = \left| \frac{1}{N}\left[\sum_{i=1}^{N} \hat{\tau}(x_i)\right] - \frac{1}{N}\left[\sum_{i=1}^{N} y_i^1 - y_i^0\right] \right| \tag{6.7}$$

The expected PEHE can be calculated in the following way:

$$\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}(x_i) - \left(y_i^1 - y_i^0\right)\right)^2} \tag{6.8}$$

Both metrics require information about potential outcomes. This information is not present in observational data. This is why it is necessary to evaluate the models on datasets with known data-generating mechanism, such as simulated and semi-simulated datasets.

6.3 Related work

Non-parametric methods. Traditional non-parametric methods for estimating ITE are based on matching. The potential outcome of a new patient with respect to t is defined using the observed outcome of the closest neighbour(s) who received t (closest in the covariate space).

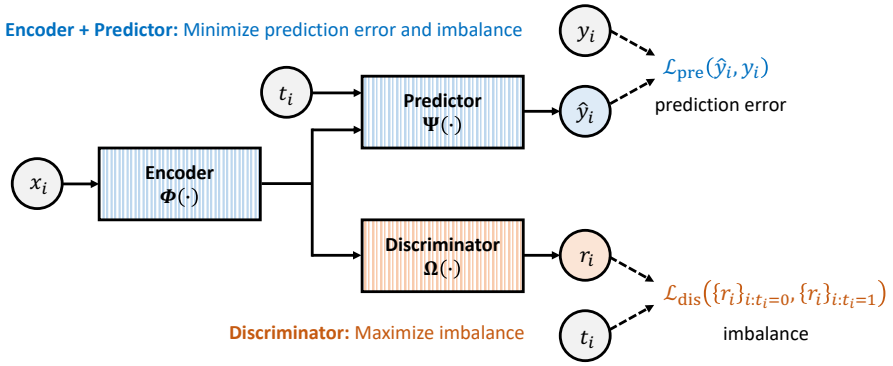


Figure 6.1 – The Predictor and the Discriminator play an adversarial game. As a result, the Encoder learns balanced representations of the covariates across different treatment groups. The Predictor ensures that the representations preserve the most important predictive information from the covariates.

Propensity score matching methods [108, 116] are used when the covariate space is high-dimensional. They match patients with similar probability to receive t. Another line of work uses tree-based methods to estimate ITE [28, 134]. BART (Bayesian Additive Regression Trees) adopts a Bayesian approach to an ensemble of trees [28]. Causal forest is an extension to the random forest algorithm that generates asymptotically unbiased ITE predictions [134].

Representation-learning methods. Representation-learning methods [62, 119, 9] try to find a new representation of the covariates that preserves the predictive information as much as possible and contains less treatment bias (the distribution across the different treatment groups is similar in the representation space). They use different distance metrics such as cross-entropy, Wasserstein distance and Maximum Mean Discrepancy distance to measure imbalance. DeepTreat first learns balanced representations of the covariates using bias-removing auto-encoder network and then generates predictions based on these representations [9]. In contrast, CFR simultaneously learns the balanced representations and generates predictions [119]. In this way, less important information from the covariates is not preserved in the representation.

Generative methods. Generative methods try to approximate the distribution of the observed and unobserved variables in the data to generate proxies of the counterfactual outcomes [82, 73, 140]. GANITE [140] uses a counterfactual generator that generates a potential outcome vector and a counterfactual discriminator that attempts to determine whether a given outcome came from the factual distribution or the counterfactual distribution. CEVAE [82] and CEGAN [73] are based on variational autoencoders and generative adversarial network, respectively. They try to generate robust ITE predictions when hidden confounders are present.

6.4 Method

Similarly to [62], we perform counterfactual inference by simultaneously learning balanced representations and generating predictions about the potential outcomes using these representations. The novelty of our approach is that we learn the balanced representations in an adversarial way. Our model consists of three components: Encoder, Discriminator and Predictor.

Encoder. The goal of the Encoder is to learn a new representation $\Phi : \mathscr{X} \rightarrow \mathbb{R}^d$. Φ is parameterized by a deep neural network that allows learning a complex non-linear representation of the covariates.

Discriminator. The Discriminator tries to distinguish between the patients who received the treatment ($t = 1$) and the patients who did not receive the treatment ($t = 0$) based only on their representation $\Phi(x)$. It is also parameterized by a deep neural network $\Omega : \mathbb{R}^d \rightarrow \mathbb{R}^m$. The Discriminator is associated to a function that returns low values when the distribution of the treated patients $\{\Omega(\Phi(x_i))\}_{i:t_i=1}$ is similar to the distribution of the untreated patients $\{\Omega(\Phi(x_i))\}_{i:t_i=0}$ over the last layer of the Discriminator. Different discrepancy functions can be used for this purpose: we focus on Wasserstein distance and cross-entropy. Wasserstein distance $W(p,q)$ between two distributions is informally defined as the minimum cost of transporting mass in order to transform the distribution q into the distribution p (where the cost is mass times transport distance) [55]. Computing $W(p,q)$ may be expensive. Thus, we are approximating Wasserstein distance in two different ways: (1) using the Sinkhorn-Knopp matrix scaling algorithm [33, 119] and (2) using gradient-penalty WGAN [55]. For the cross-entropy, the Discriminator first estimates the probability that the patient will receive a treatment:

$$\pi(x_i) = \frac{\exp(\Omega(\Phi(x_i)))}{1 + \exp(\Omega(\Phi(x_i)))} \tag{6.9}$$

where $m = 1$. Then, cross entropy is calculated in the following way:

$$\mathscr{L}_{\text{dis}}\left(\{\Omega(\Phi(x_i))\}_{i:t_i=0}, \{\Omega(\Phi(x_i))\}_{i:t_i=1}\right) = \sum_{i=1}^{N} t_i \log \pi(x_i) + (1 - t_i) \log(1 - \pi(x_i)) \tag{6.10}$$

Higher $\mathscr{L}_{\text{dis}}$ is desirable for the Discriminator.

Predictor. The goal of the Predictor is to estimate the factual outcomes for each patient. It bases its predictions on the transformed covariates $\Phi(x_i)$ and the treatment indicator t_i. The Predictor is parameterized by a deep neural network $\Psi : \mathbb{R}^d \times \{0,1\} \rightarrow \mathbb{R}^n$. Our deep neural network follows the TARNet architecture [119] and defines a separate prediction network for the data coming from each treatment group. This is more beneficial than having a single

network that takes as input a concatenation of $\Phi(x_i)$ and t_i because in the case when d is large, the treatment information might be lost during training. We use mean squared error as a loss function when the output is continuous, and cross-entropy loss function when the output is categorical. We denote the loss function by $\mathscr{L}_{\text{pre}}$.

In classical supervised machine learning we would minimize the prediction loss by jointly optimizing the parameters of the Encoder and the Predictor. In this case our model would be biased and would not generalize well for the entire population. This is why we need to obtain balanced representations of the covariates. Balancing can be considered as form of regularization. Perfect balancing can be achieved if the Discriminator is not able to distinguish between the patients who received the treatment ($t = 1$) and the patients who did not receive the treatment ($t = 0$). Achieving perfect balancing is not always desirable because in this case some important predictive information from the covariates might be lost. So a trade-off should be made between decreasing the prediction error and increasing the balance. The objective function of our model is:

$$\max_{\Theta_\Omega} \min_{\Theta_\Phi, \Theta_\Psi} \frac{1}{N} \left[\sum_{i=1}^{N} w_i \mathscr{L}_{\text{pre}} \left(\Psi\left(\Phi\left(x_i\right), t_i\right), y_i \right) \right] + \alpha \mathscr{L}_{\text{dis}} \left(\left\{ \Omega\left(\Phi\left(x_i\right)\right) \right\}_{i:t_i=0}, \left\{ \Omega\left(\Phi\left(x_i\right)\right) \right\}_{i:t_i=1} \right) + \\ + \lambda_{\text{gen}} \mathscr{R}_{\text{gen}}\left(\Theta_\Phi\right) + \lambda_{\text{pre}} \mathscr{R}_{\text{pre}}\left(\Theta_\Psi\right) - \lambda_{\text{dis}} \mathscr{R}_{\text{dis}}\left(\Theta_\Omega\right) \quad (6.11)$$

where

$$w_i = \frac{t_i}{2u} + \frac{1 - t_i}{2\left(1 - u\right)} \text{ and } u = \frac{1}{N} \sum_{i=1}^{N} t_i \quad (6.12)$$

$\mathscr{R}_{\text{gen}}$, $\mathscr{R}_{\text{pre}}$ and $\mathscr{R}_{\text{dis}}$ are terms that penalize complex functions for the Encoder, Predictor and Discriminator, respectively. The weights w_i compensate for the difference in treatment group sizes [119].

The Discriminator plays an adversarial game with the Predictor. The goal of the Predictor is to predict the factual outcomes accurately and to obtain balanced distribution across the different treatment groups over the last layer of the Discriminator. However, the Predictor cannot affect the parameters of the Discriminator to achieve balancing, it can only affect the parameters of the Encoder. Thus, in order to achieve its goal, the best strategy is to make the distribution of the treated patients $\{\Phi(x_i)\}_{i:t_i=0}$ and the distribution of the untreated patients $\{\Phi(x_i)\}_{i:t_i=1}$ over the representation layer as similar as possible.

The Discriminator tries to obtain imbalanced distributions in the last layer by only affecting its own parameters. Since it doesn't have access to the parameters of the Encoder, its goal cannot be achieved if the Encoder already produced balanced distributions in the representation layer. By playing an adversarial game, it is possible to generate perfectly balanced distributions over the representation layer while preserving predictive information as much as possible. In this way, ACFR has an increased ability to produce balanced representations over CFR.

In the training phase, we use Adam optimizer to update network weights and exponential decay learning rate [1]. We update the Discriminator once per each update of the Encoder and the Predictor. The number of updates of the Discriminator can also be treated as a hyper-parameter because a higher number of updates may be beneficial in some cases [88]. We use held-out validation dataset to determine the optimal model during the training process. We choose the model with the lowest Predictor and Discriminator loss (Equation 6.11 without the regularization terms) on the validation dataset.

6.5 Experiments

6.5.1 Experiments on Benchmark Datasets

It is difficult to evaluate causal inference algorithms using observational data because this data does not contain the counterfactual outcomes. Thus, simulated and semi-simulated datasets are used for this purpose. We applied ACFR on two commonly used benchmark datasets: IHDP and Twins. We measured PEHE and ATE as performance metrics in two different settings: within-sample and out-of-sample [119]. In the first case, PEHE and ATE are measured on the training and the validation set — meaning that the algorithm has access to the factual outcome. In the second case, PEHE and ATE are measured on the testing set and the goal is to estimate ITE for patients with no observed outcomes. The second setting is more important because it corresponds to a case when a new patient arrives and we need to recommend the best possible treatment for him or her.

We compared our method with the following baseline methods: Ordinary Least Squares with treatment as a feature (OLS-1), Ordinary Least Squares with separate regressors for each treatment (OLS-2), k-nearest neighbour (k-NN), Bayesian Additive Regression Trees (BART) [28], Random forest [17], Causal Forest [134], Causal Effect Variational Autoencoder (CEVAE) [82], Balancing Neural Networks (BNN) [62], Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) [140], Balancing Linear Regression (BLR) [62], TARNet [119], CFR-MMD [119], CFR-WASS [119], Logistic Regression with treatment as a feature (LR-1), Logistic Regression with separate regressors for each treatment (LR-2), multi-task Gaussian Process (CMGP) [3] and CEGAN [73].

We use three variations of our method according to the discrepancy function used by the Discriminator. ACFR-CE uses cross-entropy to measure imbalance. ACFR-WGAN and ACFR-WASS use an approximation of Wasserstein distance to measure imbalance. ACFR-WGAN implements gradient-penalty WGAN and ACFR-WASS uses the Sinkhorn-Knopp matrix scalling algorithm (in the same way as in CFR-WASS).

This section is based on the semester project "Causal Inference in Observational Data", which was completed by Wenyuan Lv, under the supervision of Igor Kulev and Boi Faltings.

IHDP dataset. This dataset is based on RCT data from the Infant Health and Development Program (IHDP) that aimed to raise the cognitive test scores of low-birth-weight, premature infants [58]. Part of the treated population has been removed to simulate treatment bias. The dataset contains 747 subjects (608 control and 139 treated) and 25 covariates associated to each subject. These covariates measure different properties of the child and his or her mother. The output is simulated and continuous. More details about the dataset can be found in [58] and [119]. We average over 1,000 realizations of the simulated outcome with 63/27/10 train/validation/test/splits.

Twins dataset. This dataset utilizes data from twin births in the USA between 1989-1991 [5, 82]. We define treatment $t = 1$ as being the heavier twin, and the outcome as the mortality after one year. Thus, the potential outcomes are present in the data. We focus only on same-sex twins that have birth weights below 2 kg. The dataset contains 10,286 twins with 49 features available before birth. We follow the procedure described in [73] to obtain a semi-simulated dataset that contains treatment bias. More concretely, we assign treatment to each twin by sampling from $t_i \sim \text{Bern}(\sigma(wz_i))$, where $w \sim \mathcal{N}\left(10, 0.1^2\right)$ and z is the min-max normalized value of the feature GESTAT which represents the gestation age in weeks. We average over 50 realizations of the simulated treatment assignment with 64/16/20 train/validation/test/splits.

Table 6.1 and Table 6.2 show the results of the IHDP and Twins experiments. The best baseline methods on the IHDP dataset are based on counterfactual regression. ACFR further improves CFR and demonstrates that adversarial balancing is useful to generate more robust predictions. ACFR is still better than CFR on the Twins dataset and is competitive with the best state-of-the-art methods. We should note that causal inference on the Twins dataset is more challenging because the treatment groups are highly imbalanced: the smaller group contains just 3% of the subjects. Causal Forest performs the best in the within-sample setting, however, its performance drops in the out-of-sample setting. In contrast, the performance of ACFR is similar in both settings and we estimate ITE for new patients more accurately than the baseline methods (lowest out-of-sample PEHE).

We performed an additional experiment on the IHDP dataset to analyze the impact of the imbalance penalty α on PEHE. These results are shown in Fig. 6.2. Both CFR and ACFR perform the best when $\alpha = 1$, however ACFR achieves lower PEHE error: the gap between the two methods is stable for different values of α. When $\alpha = 0$, both models perform the same, as expected.

Table 6.1 – Results on the IHDP dataset. Lower is better.

Method	in-sample		out-sample	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
OLS-1	5.8 ± 0.3	$.73 \pm 0.04$	5.8 ± 0.3	$.94 \pm 0.06$
OLS-2	2.4 ± 0.1	$.14 \pm 0.01$	2.5 ± 0.1	$.31 \pm 0.02$
BLR	5.8 ± 0.3	$.72 \pm 0.04$	5.8 ± 0.3	$.93 \pm 0.05$
k-NN	2.1 ± 0.1	$.14 \pm 0.01$	4.1 ± 0.2	$.79 \pm 0.05$
BART	2.1 ± 0.1	$.23 \pm 0.01$	2.3 ± 0.1	$.34 \pm 0.02$
Random Forest	4.2 ± 0.2	$.73 \pm 0.05$	6.6 ± 0.3	$.96 \pm 0.06$
Causal Forest	3.8 ± 0.2	$.18 \pm 0.01$	3.8 ± 0.2	$.40 \pm 0.03$
CEVAE	2.7 ± 0.1	$.34 \pm 0.01$	2.6 ± 0.1	$.46 \pm 0.02$
BNN	2.2 ± 0.1	$.37 \pm 0.03$	2.1 ± 0.1	$.42 \pm 0.03$
GANITE	1.9 ± 0.4	$.43 \pm 0.05$	2.4 ± 0.4	$.49 \pm 0.05$
TARNet	$.88 \pm 0.0$	$.26 \pm 0.01$	$.95 \pm 0.0$	$.28 \pm 0.01$
CFR-MMD	$.73 \pm 0.0$	$.30 \pm 0.01$	$.78 \pm 0.0$	$.31 \pm 0.01$
CFR-WASS	$.71 \pm 0.0$	$.25 \pm 0.01$	$.76 \pm 0.0$	$.27 \pm 0.01$
ACFR-CE	$.66 \pm 0.0$	$.16 \pm 0.01$	$.69 \pm 0.0$	$.18 \pm 0.01$
ACFR-WGAN	$.69 \pm 0.1$	$\mathbf{.15 \pm 0.01}$	$.76 \pm 0.0$	$.17 \pm 0.01$
ACFR-WASS	$\mathbf{.58 \pm 0.1}$	$\mathbf{.15 \pm 0.01}$	$\mathbf{.62 \pm 0.1}$	$\mathbf{.16 \pm 0.01}$

Table 6.2 – Results on the Twins dataset. Lower is better.

Method	in-sample		out-sample	
	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
LR-1	0.365 ± 0.00	0.045 ± 0.02	0.367 ± 0.00	0.186 ± 0.03
LR-2	0.404 ± 0.02	0.128 ± 0.03	0.411 ± 0.02	0.206 ± 0.04
k-NN	0.486 ± 0.02	0.254 ± 0.04	0.506 ± 0.02	0.264 ± 0.04
Causal Forest	$\mathbf{0.356 \pm 0.01}$	0.025 ± 0.02	0.372 ± 0.01	0.188 ± 0.03
BART	0.569 ± 0.06	0.432 ± 0.08	0.562 ± 0.06	0.429 ± 0.08
CMGP	0.367 ± 0.01	0.034 ± 0.03	0.365 ± 0.01	0.036 ± 0.04
CFR-WASS	0.371 ± 0.03	0.056 ± 0.06	0.371 ± 0.03	0.071 ± 0.06
CEVAE	0.363 ± 0.00	0.071 ± 0.01	0.364 ± 0.00	0.165 ± 0.01
CEGAN	0.363 ± 0.00	$\mathbf{0.018 \pm 0.01}$	0.362 ± 0.00	$\mathbf{0.017 \pm 0.01}$
ACFR-CE	0.363 ± 0.00	0.020 ± 0.01	$\mathbf{0.357 \pm 0.00}$	0.024 ± 0.01
ACFR-WGAN	0.360 ± 0.00	0.020 ± 0.01	0.360 ± 0.00	0.020 ± 0.01
ACFR-WASS	0.361 ± 0.00	0.020 ± 0.01	0.360 ± 0.00	0.021 ± 0.01

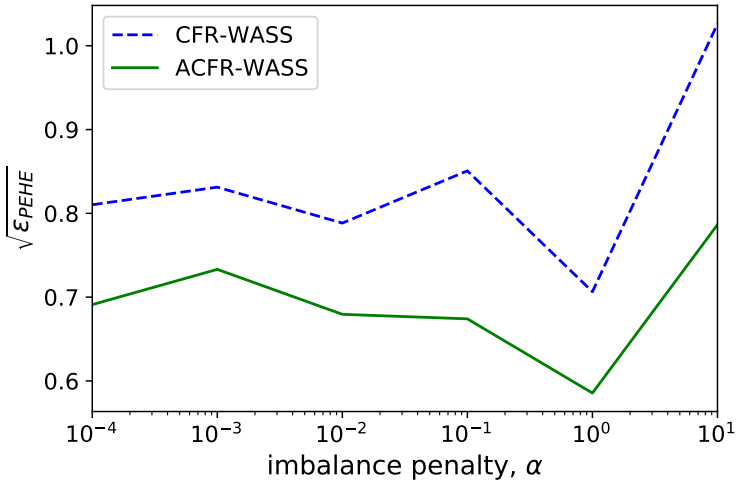


Figure 6.2 – PEHE as a function of the imbalance penalty α for CFR and ACFR.

6.5.2 Extension of the Intervention-Based Clustering Method

Our framework is also useful because it can be adapted to existing methods for predicting treatment effect that do not consider treatment bias. In this way, these methods can be applied to observational data besides randomized trial data. In this section, we demonstrate the usefulness of adversarial balancing to Intervention-Based Clustering (IBC) that we introduced in Chapter 3. IBC discovers subpopulations that were affected by the intervention in different ways i.e. the discrete heterogeneity in the treatment effect. Discrete heterogeneity is likely to be present if a treatment works through unobserved causal pathways that may be discretely open or closed [118]. For example, suppose a drug works better for patients with a specific genetic profile and this information is not recorded. Let's assume that weight is associated with the presence of this genetic profile and this information is recorded. Then, the output as a function of weight will be better approximated by a latent class model with weight as a predictor of latent class membership than by any smooth function of weight alone [118].

The disadvantage of IBC is that it could provide biased results when applied directly to observational data. However, IBC may still perform well on observational data when its inputs are the representations $\Phi(\mathscr{X})$ learned by ACFR instead of the original feature vectors $\mathscr{X}$. Integrating the IBC approach withing the ACFR framework allows us to discover clusters by simultaneously learning balanced representations of the covariates across different treatment groups and predicting the factual outcomes. We modified the Predictor and its loss function in our ACFR framework to support intervention-based clustering. We introduced a probabilistic loss function:

$$\mathscr{L}_{\text{ibc}}\left(\Psi\left(\Phi\left(x_i\right), t_i\right), y_i\right) = -\log \sum_{k=1}^{K} \pi_k\left(\Phi\left(x_i\right)\right) \mathscr{N}\left(y_i | \mu_k\left(\Phi\left(x_i\right), t_i\right), \Sigma\right) \tag{6.13}$$

where $\pi_k(\Phi(x_i))$ denotes the prior odds for the i-th user to belong to the k-th cluster and $\mu_{k,t_i}(\Phi(x_i))$ denotes the i-th user's outcome if he or she belonged to the k-th cluster. Both functions are parameterized by a deep neural network Ψ. However, the first function π_k does not depend on the treatment indicator t_i. This allows us to determine the most likely cluster memberships for new people before giving them any treatment. Similarly to Chapter 3, we set the covariance matrix Σ to be the same for all distributions. The regularizer term $\mathscr{R}_{\text{pre}}(\Theta_\Psi)$ in Equation 6.11 can be interpreted as the prior distribution of the IBC model parameters. We added two more regularizers in Equation 6.11 to better separate subpopulations with differential treatment effect: *sample-wise entropy* and *batch-wise entropy* [144]. Sample-wise entropy can be defined in the following way:

$$H_{\text{sam}}(\pi) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\pi_k(\Phi(x_i))\log(\pi_k(\Phi(x_i))) \tag{6.14}$$

Lower sample-wise entropy means that each patient receives one dominant cluster assignment. Batch-wise entropy can be defined in the following way:

$$H_{\text{bat}}(\pi) = -\sum_{k=1}^{K}\bar{p}_k\log\left(\bar{p}_k\right) \tag{6.15}$$

where

$$\bar{p}_k = \frac{1}{N}\sum_{i=1}^{N}\pi_k(\Phi(x_i)) \tag{6.16}$$

Higher batch-wise entropy means that all clusters contain a similar number of patients. Lower H_{sam} and higher H_{bat} are desirable because they characterize a clustering in which subpopulations are large and well-separated. Thus, our final objective is:

$$\begin{aligned}\max_{\Theta_\Omega}\min_{\Theta_\Phi,\Theta_\Psi}\frac{1}{N}\left[\sum_{i=1}^{N}w_i\mathscr{L}_{\text{ibc}}\left(\Psi(\Phi(x_i),t_i),y_i\right)\right] + \alpha\mathscr{L}_{\text{dis}}\left(\{\Omega(\Phi(x_i))\}_{i:t_i=0},\{\Omega(\Phi(x_i))\}_{i:t_i=1}\right) + \\ + \lambda_{\text{gen}}\mathscr{R}_{\text{gen}}(\Theta_\Phi) + \lambda_{\text{pre}}\mathscr{R}_{\text{pre}}(\Theta_\Psi) - \lambda_{\text{dis}}\mathscr{R}_{\text{dis}}(\Theta_\Omega) + \lambda_{\text{sam}}H_{\text{sam}}(\pi) - \lambda_{\text{bat}}H_{\text{bat}}(\pi)\end{aligned} \tag{6.17}$$

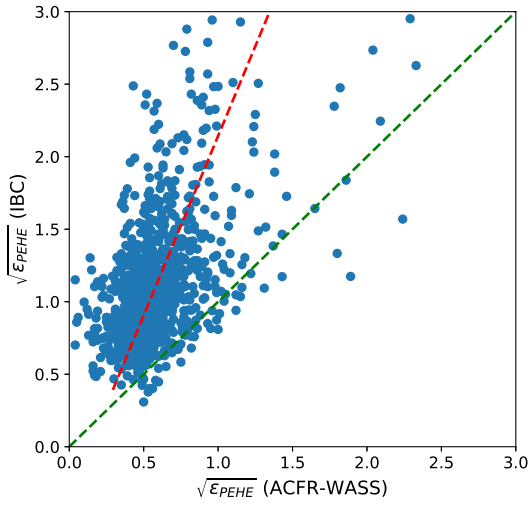
We trained and evaluated the IBC model on the IHDP dataset. We defined three different variants of the model: IBC is the model with $\alpha = 0$ (does not use balancing), IBC-BAL is the model with $\alpha = 1$ (uses balancing), and IBC-BAL-RR is the model with the lowest validation loss over two runs of IBC-BAL with different initial parameter values (uses balancing and random restarts). We set $K = 2$ for all variants of the model. The results are given in Table 6.3. It can be seen that balancing helps IBC to improve its treatment effect estimation on observational data. Also, running the method several times with random initializations provides better solutions. Random restarts decreased the probability to obtain extreme PEHE values. IBC-BAL-RR performs better than CFR-WASS, but still not better than ACFR-WASS. IBC-BAL-RR could perform at least as good as ACFR-WASS because ACFR-WASS is a special case of IBC-BAL-RR

Table 6.3 – Evaluation of different IBC models on the IHDP dataset. Lower is better.

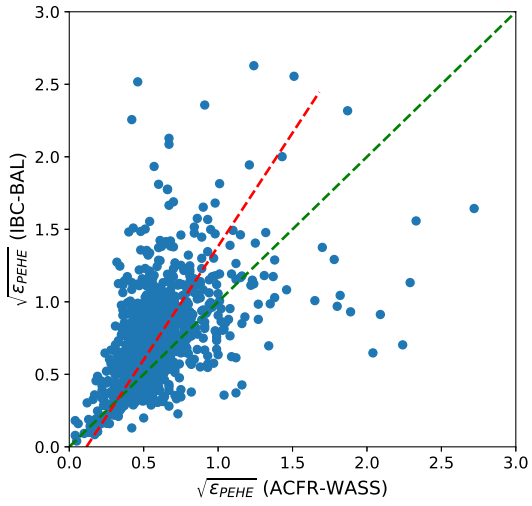
	in-sample		out-sample	
Method	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}	$\sqrt{\epsilon_{PEHE}}$	ϵ_{ATE}
IBC	1.1 ± 0.0	$.23 \pm 0.01$	1.2 ± 0.0	$.26 \pm 0.01$
IBC-BAL	$.74 \pm 0.1$	$.18 \pm 0.01$	$.81 \pm 0.0$	$.20 \pm 0.01$
IBC-BAL-RR	$.68 \pm 0.1$	$.18 \pm 0.01$	$.74 \pm 0.0$	$.19 \pm 0.01$

when $K = 1$. However, we did not train the model with different values of K because the training process was very time-consuming. PEHE for all realizations of the simulated outcome in the IHDP dataset is visualized in Figure 6.3. It can be seen that in almost all realizations IBC performs worse than ACFR-WASS (almost all points are above the green dashed line). In contrast, there is a high correlation between PEHEs for IBC-BAL-RR and ACFR-WASS.

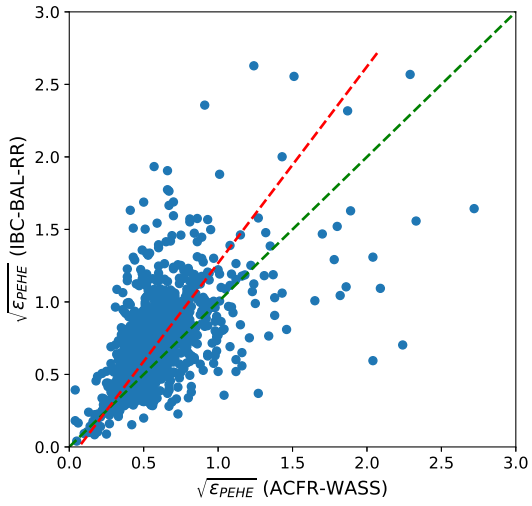
We continued the analysis by inspecting the clusterings obtained by IBC and IBC-BAL. IBC produced well-separated clusters more often than IBC-BAL. We explain this by the fact that the loss function in IBC-BAL has an additional term $\mathscr{L}_{\text{dis}}$ which decreases the importance of the entropy terms H_{sam} and H_{bat}. Figure 6.4 shows the clustering results for one realization of the simulated outcome. IBC-BAL discovered two subpopulations that differ in terms of the distribution of the output. People from the second cluster that did not receive the treatment have a lower response than the people from the first cluster that did not receive the treatment. In contrast, people from both clusters that received the treatment responded similarly. This means that the people from the second cluster are more likely to benefit from the treatment than the people from the first cluster (assuming that higher value of the output is better for the user).



(a)

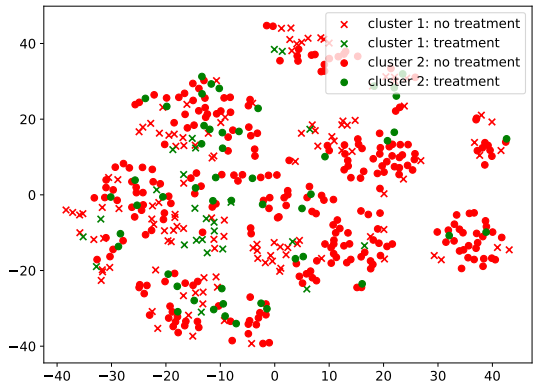


(b)

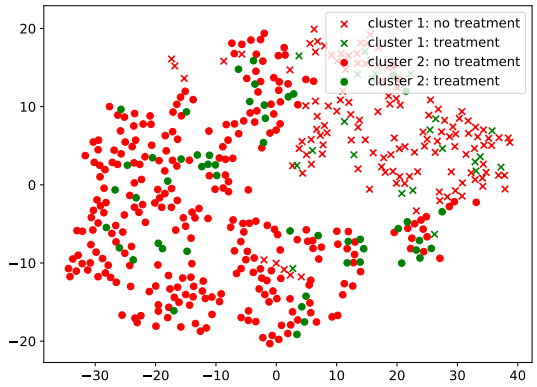


(c)

Figure 6.3 – Comparison of out-sample $\sqrt{\epsilon_{PEHE}}$ between different IBC models and ACFR.



(a) t-SNE visualization of the clustering in the input space.



(b) t-SNE visualization of the clustering in the embedded space.

(c) Average outputs per cluster and intervention.

Figure 6.4 – Visualization of the clustering results obtained with IBC-BAL on the IHDP dataset.

6.6 Chapter Summary

Estimating the individualized treatment effect is an essential task in many domains, especially in the health domain. Counterfactual regression solves this task by simultaneously learning balanced representations of the covariates across different treatment groups and predicting the factual outcomes. In this chapter, we presented a novel method based on counterfactual regression, ACFR, that performs the regression by learning balanced representations in an adversarial way. In this way, the model has a higher power to obtain more balanced representations while still preserving the important predictive information from the covariates.

We compared the performance of ACFR with state-of-the-art methods on two commonly used benchmark datasets, and we demonstrated that ACFR performs better than the baseline models on the IHDP dataset and is competitive on the Twins dataset. It is important to note that on both datasets, ACFR was the most accurate in estimating ITE for new patients (low out-of-sample PEHE). The results indicate that adversarial balancing is useful to generate predictions that generalize better on the counterfactual distribution.

We demonstrated that our balancing approach could be used to apply intervention-based clustering (IBC) on observational data. Balancing helped IBC to improve its treatment effect estimation and to discover more relevant clusters from observational data. The same balancing approach can be used to extend the application of other existing intervention systems that predict treatment effect without considering treatment bias.

7 Conclusion

7.1 Summary

Unhealthy behaviors are associated with increased cardiovascular morbidity and mortality [81]. However, behavior changes are difficult to make. Also, often, people who like to improve their behavior do not know how to do that. Personalizable intervention systems could assist them to achieve healthy behavior change. These systems decide what would be the optimal intervention for the target user based on his or her characteristics, including current and past behavior patterns. They consist of three main components: data collection, predictive modeling and generating recommendations. The first component provides data from past users: each of them either received one of the interventions of interest or did not receive anything. This data is used by the second component to predict how a new user would change his or her behavior if he or she was given an intervention. The third component generates a recommendation strategy based on these predictions.

In this thesis, we proposed novel solutions that address the main challenges in building a personalizable intervention system to promote healthy behavior change. First, we proposed a system based on a Bayesian mixture model to identify subpopulations with different behavior changes from longitudinal data. This system is especially suitable when the amount of data is limited, and when there are unobserved factors that might affect behavior change. We applied it on a dataset obtained from a randomized controlled trial where patients were randomly allocated to receive up to 12 acupuncture treatments over three months or to a control intervention offering usual care. We discovered two distinct subpopulations that were separated by non-linear decision boundaries. The intervention was effective only for one subpopulation, suggesting that acupuncture significantly increases the energy levels of the people with high emotional well-being.

Second, we proposed CLINT, a system based on a latent-variable model, to discover and predict behavior change patterns from fine-grained sensor data. The novelty of this system is that it learns interpretable patterns that describe pre-intervention behavior and post-intervention behavior change in an unsupervised way. The fine-grained approach enables the discovery of

periods when the intervention is the most effective, e.g., the morning time. We applied CLINT to calorie expenditure data obtained from 45 people who received a mobile app intervention. We discovered that less active people benefit more from the intervention: they increased their physical activities mostly during the morning. We demonstrated that CLINT could use the discovered patterns to predict the post-intervention behavior of new people better than existing methods. We also proposed a way to recommend strategies for positive behavior change by learning from the existing users that improved their behavior. We validated our recommendations and demonstrated that they are feasible and effective.

Third, we proposed and developed a personalizable intervention system to improve the physical activeness of senior adults. The main novelty of our system is that it uses historical time series fitness data to decide which intervention to recommend. We trained the system using fitness data obtained from 55 senior adults: each of them received one of two mobile app interventions by random assignment. To deal with the high-dimensionality of fitness data, we used an autoencoder to transform each time series into a low-dimensional representation. Our system suggested that the intervention that worked better for the overall population was likely to be effective only on 75% of the people. Also, we showed that using fine-grained sensor data from a longer period leads to more accurate predictions.

Finally, we proposed ACFR, an adversarial approach to reduce intervention bias in observational data. Our approach converts the pre-intervention data into a representation that has a balanced distribution of the latent features across different treatment groups and preserves the predictive information from the covariates as much as possible. During the learning phase, the two objectives are simultaneously optimized in an adversarial way. We demonstrated that our approach estimates the intervention effect better than existing methods on a widely-used benchmark dataset. We also showed that our balancing approach could be used to adapt existing personalizable intervention systems designed for data from experimental trials to support data from observational data as well.

Personal intervention systems hold great potential to turn existing human behavior data into actionable insights for people that have unhealthy lifestyles. We believe that the work described in this thesis may lead to a better understanding of human behavior and the ways to promote healthy behavior change. The advance of wearable technology and the increased availability of human behavior data may further facilitate the adoption of personalizable intervention systems for healthy behavior change.

7.2 Future Directions

The accuracy of personalizable intervention systems may be improved if these systems have access to data from a larger number of people. However, conducting trials to provide additional training data is an expensive and time-consuming process. A possible solution to this problem is to add the data from the target users to the training dataset after they receive the recommendations. This would not be an additional burden for the target users because

personalizable intervention systems already track their behavior to generate personalized recommendations. The system could be re-trained when new (biased) data is available. This is an online learning problem and is related to the contextual bandits problem [37]. In the beginning, the intervention system does not know which intervention works the best for a given user, thus it recommends a random intervention — a setting similar to randomized trials. As there are more and more data available, the system learns which intervention works better for a new user. The system recommends this intervention to the new user with a higher probability.

In this thesis, we considered the case when the system decides *which* intervention to recommend out of a finite set of interventions. However, in some cases it may be desirable to determine the optimal *dosage* of a single intervention. For example, what daily step goal should be recommended to the target user, e.g., 10,000 steps. We may be interested in the optimal daily step goal that would motivate the user to improve himself or herself. Another example of intervention dosage is the frequency of sending suggestions to promote healthy behavior change to people with unhealthy behaviors. Adapting the systems and methods proposed in this thesis to support intervention dosages instead of binary interventions would be an interesting research problem. One approach to solve this problem is to discretize the intervention dosage and to treat each discrete value as a separate intervention. A limitation of this approach is that it requires data from many users receiving different dosages. Another approach to solving this problem is to adapt the adversarial balancing method introduced in Chapter 6. This would require defining a new loss function for the Discriminator so that the distribution across different intervention dosages is similar in the representation space.

Recurrent Neural Networks (RNN) are powerful methods to generate accurate behavior change predictions from time series data. These methods could be integrated into personalizable intervention systems when there is a large amount of training data available. However, it is difficult to explain how they work to the target users. A possible solution to this problem is to use RNN with attention mechanisms [131]. These mechanisms reveal which part of the data was the most predictive of behavior change, e.g., activities that happen during the morning. This feature could be useful to understand why the intervention system described in Chapter 5 chooses one intervention over another for a given user. The system could visualize and highlight the parts of the input it focuses on when predicting the behavior change. Showing information about existing users who had similar behavior patterns in these segments and improved their behavior leads to increased transparency and user trust in the system. Future work may investigate the potential use of attention mechanisms in personalizable intervention systems to promote healthy behavior change.

A Appendix

A.1 IBC: Ablation Study

In Section 3.4.1 we generated a simulated dataset and we tested the ability of Intervention-Based Clustering (IBC) to capture clusters with complex decision boundaries. The method discovers the complex decision boundaries by adding polynomial terms in the input data. Based on the results given in Table 3.1, we performed an ablation study to compare the impact of both clustering and polynomial terms on the log-likelihood. The model that performs clustering without including polynomial terms results with a higher log-likelihood than the model that performs no clustering at all (see Table A.1). Adding polynomial terms in the input data further improves the log-likelihood on the validation dataset.

Table A.1 – Ablation study comparing different IBC model components on the log-likelihood. Higher is better.

	log-likelihood
no clustering	-1.0462
clustering	-0.7924
clustering + polynomials	**-0.7872**

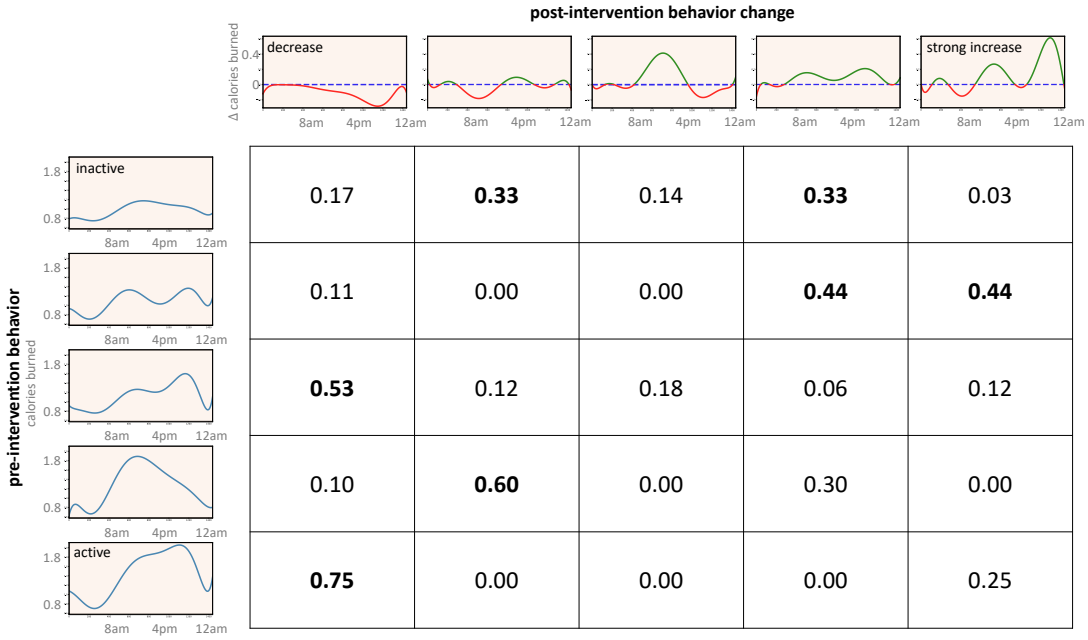
A.2 CLINT: External Validation

Young and senior adults have different physical activity patterns [21]. We performed an experiment to determine whether CLINT could discover these differences, based on fitness data obtained from young and senior adults. For this purpose, we merged the HealthyTogether dataset described in Chapter 4 that contains data collected from young adults, mostly students, and part of the dataset described in Chapter 5 that contains data collected from senior adults

who received a *peer-to-peer* intervention[1]. In this way, all the people in the merged dataset received very similar social interventions.

We ran 9-fold cross-validation to estimate the generalization ability of our model. In each round, we removed a group of users (fold), we trained the model using the remaining users, and we applied the model to the held-out group. In the learning process, we built 10 models with different initial parameter values, and we chose the model with the highest log-likelihood on the training set. We trained models with different combinations of values for K^0, K^1 and D, and we chose the model that generalizes the best. We defined the following search ranges: $1 \leq K^0 \leq 5$, $1 \leq K^1 \leq 5$ and $1 \leq D \leq 10$. The optimal model with $K^0 = 5$, $K^1 = 5$ and $D = 9$ is used further in our analysis.

The extracted daily activity patterns are shown in Figure A.1. The first daily activity pattern (46% of the population) represents inactive people. This pattern is the most common among users. The second activity pattern (12% of the population) represents moderately active people. It has two peaks: during the morning and during the evening. The third activity pattern (22% of the population) also represents moderately active people. It has two peaks: during the morning and during the evening. The peak in the evening is higher than the peak in the morning. The fourth activity pattern (13% of the population) represents highly-active people who are more active during the morning, and the fifth activity pattern (5% of the population) represents highly-active people who are more active during the evening.



0.17	**0.33**	0.14	**0.33**	0.03
0.11	0.00	0.00	**0.44**	**0.44**
0.53	0.12	0.18	0.06	0.12
0.10	**0.60**	0.00	0.30	0.00
0.75	0.00	0.00	0.00	0.25

Figure A.1 – The probability a pre-intervention behavior (left) changes to a post-intervention behavior (up) estimated on the extended HealthyTogether dataset.

[1]We used only data from the last week before, and the first week after the intervention, to be consistent with the structure of the HealthyTogether dataset. We removed 4 participants who had missing data during this period.

Table A.2 – The number of people who moved from each particular activity pattern (AP) to each activity change pattern (ACP). The first number in each bracket denotes the number of transitions by the young adults, and the second number denotes the number of transitions by the senior adults.

	ACP1	ACP2	ACP3	ACP4	ACP5
AP1	(0,6)	(6,6)	(1,4)	(7,5)	(1,0)
AP2	(1,0)	(0,0)	(0,0)	(4,0)	(3,1)
AP3	(9,0)	(2,0)	(3,0)	(1,0)	(2,0)
AP4	(0,1)	(1,5)	(0,0)	(0,3)	(0,0)
AP5	(3,0)	(0,0)	(0,0)	(0,0)	(1,0)

The extracted activity change patterns are also shown in Figure A.1. The first activity change pattern represents people who decreased their activity levels during the evening (-143 calories, 26% of the population). The second activity change pattern represents people who decreased their activity levels during the morning (-30 calories, 26% of the population). The third activity change pattern represents people who improved during the middle of the day (+82 calories, 11% of the population). The fourth activity change pattern represents people who moderately increased their activities both during the morning and during the evening (+119 calories, 26% of the population). The fifth activity change pattern represents people who strongly increased their activity levels both during the morning and during the evening (+170 calories, 26% of the population). The transition table in Figure A.1 suggests that people represented by the second activity pattern are most likely to improve their activity levels. The table also suggests that people that are very active during the evening (third and fifth activity pattern) tend to decrease their activities during this period and also, people that are very active during the morning (fourth activity pattern) tend to decrease their activities during this period.

The number of people who moved from each activity pattern to each activity change pattern is given in Table A.2. It can be seen that young and senior adults differ in terms of their activity patterns. Young adults are less likely to belong to the fourth activity pattern (early morning people) while senior adults belong to either the first (very inactive) or the fourth activity pattern (early morning people). Also, young adults are more likely to strongly increase their activity levels both during the morning and during the evening than senior adults. The cluster purity for the clusters determined by the pre-intervention behavior is 0.776. The cluster purity for the clusters determined by the post-intervention behavior change is 0.618. This suggests that young and senior adults differ more in terms of their regular behavior, but less in terms of their behavior change.

A.3 CLINT: Experiments on Artificial Dataset

Dataset. The underlying real-world-data-generating mechanism is complex and not directly observable — making it difficult to determine how well CLINT is able to extract the true behavior patterns that exist in the data. This is why we performed an additional experiment on an artificial dataset with known data-generating mechanism. We assumed that there are 200 people manifesting a periodic behavior before and after the intervention. This behavior has a period of length 50 and there are 5 periods before and after the intervention i.e., $M = 250$. We assumed that people's behavior can be explained using 3 behavior patterns and 2 behavior change patterns i.e., $K^0 = 3$ and $K^1 = 2$. Data was generated according to the model shown in Fig. 4.2. We used 4 polynomial terms up to degree 3 to represent x^0 and x^1. The entries in π and τ were generated using normalized samples from a uniform distribution. The regression coefficients in α and β were sampled from $\mathcal{N}(0, 0.8)$ and $\mathcal{N}(0, 0.4)$ respectively. Σ was sampled from $W_1(0.1/3, 3)$. The parameters were chosen arbitrarily.

In this experiment, we were interested to know whether CLINT is able to extract the true ground-truth patterns from the data. Initially, we assumed that the true number of patterns and the true complexity of the behavior are unknown. We used grid search to discover this information from the data. The number of behavior and behavior change patterns varied from 1 to 4 and the number of polynomial terms included in x^0 and x^1 varied from 1 to 7. The artificial dataset was split into training and validation sets of equal size. We used 30 random restarts to obtain the optimal model for each combination of hyperparameters.

The log-likelihood on the validation dataset was improving as we were reaching the true number of patterns and polynomial terms, as it can be seen in Fig. A.2 and Fig. A.3. After that, the log-likelihood did not change significantly. In this way, we demonstrated that our learning procedure is able to extract and reconstruct the true behavior patterns and behavior change patterns from the data.

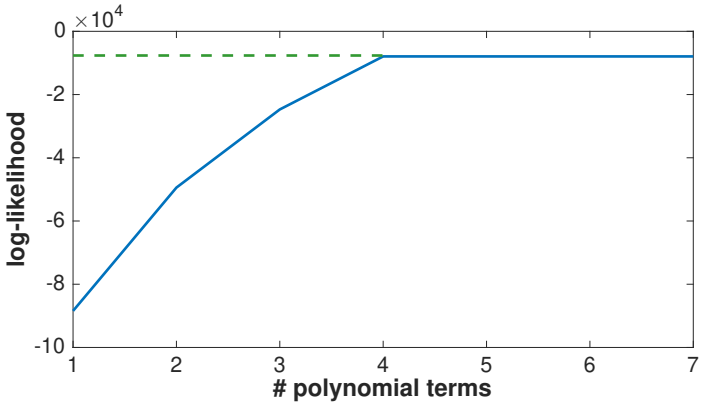


Figure A.2 – Log-likelihood of models with different number of polynomial terms, trained on the artificial dataset. The optimal log-likelihood obtained with the true model parameters is shown with dashed green line.

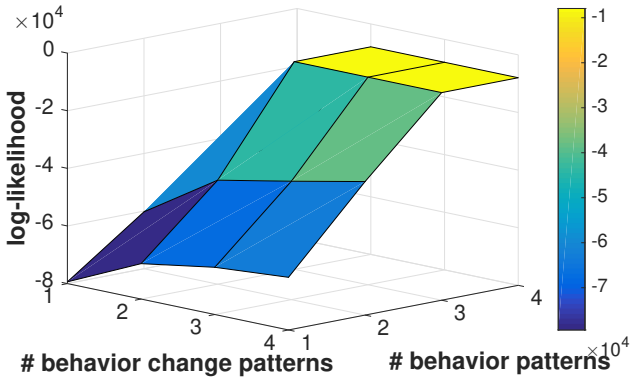


Figure A.3 – Log-likelihood of models with different K^0 and K^1, trained on the artificial dataset. There is no increase of the log-likelihood when $K^0 > 3$ and $K^1 > 2$.

A.4 Predictive Modeling: Additional Evaluation

We performed an additional analysis of the predictive models described in Chapter 5 using data from the HealthyTogether dataset described in Chapter 4. The test errors for each model are given in Figure A.4. These results suggest that both the lower-level human behavior within a single day and the higher-level human behavior from one day to another improve prediction accuracy.

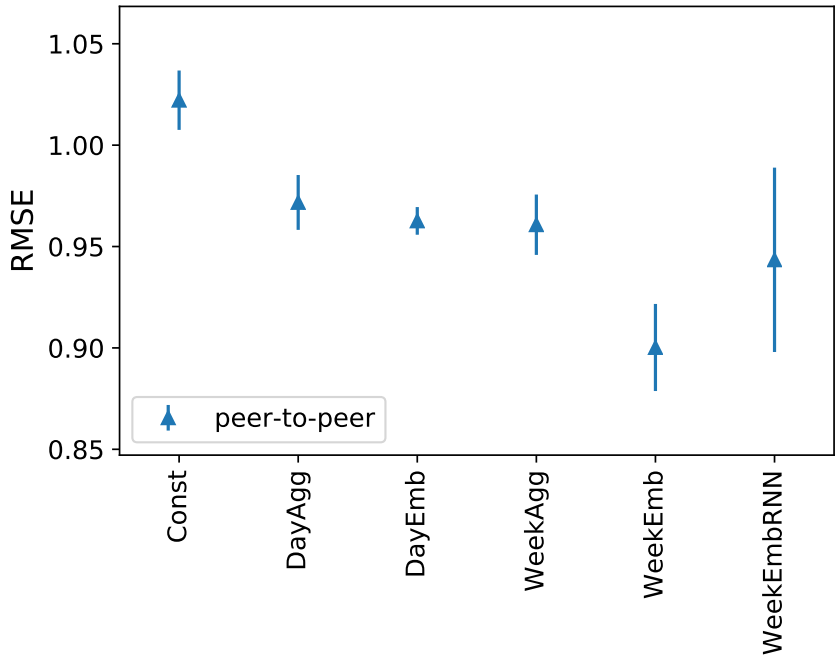


Figure A.4 – Comparison of the test error of different predictive models applied on the HealthyTogether dataset.

A.5 Prediction Accuracy and Sample Size

We performed an experiment to evaluate the accuracy of the predictive models described in Chapter 5 as a function of the amount of training data. For this purpose, we used the optimal model applied to the HealthyTogether data (see Figure A.4). We sampled different amounts of data to train the model. The prediction errors on the held-out data are given in Figure A.5. As we increase the amount of training data, the prediction error decreases. Since it is not evident that the RMSE reaches a plateau at 100%, it seems that the RMSE would continue decreasing if we had used more data.

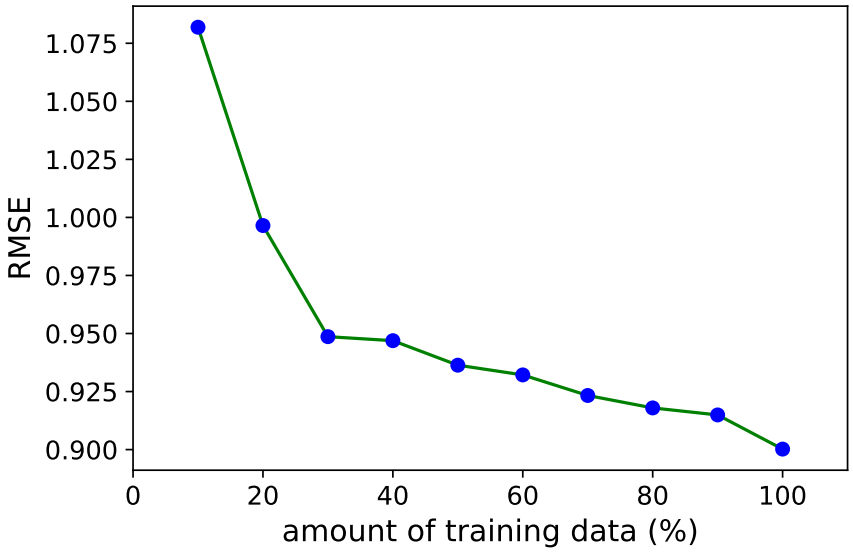


Figure A.5 – RMSE as a function of the amount of data used to train the model (smaller is better).

Bibliography

[1] ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G., ISARD, M., ET AL. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (2016), pp. 265–283.

[2] AITKEN, M., CLANCY, B., AND NASS, D. The growing value of digital health: evidence and impact on human health and the healthcare system. *IQVIA Institute for Human Data Science* (2017).

[3] ALAA, A. M., AND VAN DER SCHAAR, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems* (2017), pp. 3424–3432.

[4] ALBERT, M. V., KORDING, K., HERRMANN, M., AND JAYARAMAN, A. Fall classification by machine learning using mobile phones. *PloS one 7*, 5 (2012), e36556.

[5] ALMOND, D., CHAY, K. Y., AND LEE, D. S. The costs of low birth weight. *The Quarterly Journal of Economics 120*, 3 (2005), 1031–1083.

[6] ALOSH, M., FRITSCH, K., HUQUE, M., MAHJOOB, K., PENNELLO, G., ROTHMANN, M., RUSSEK-COHEN, E., SMITH, F., WILSON, S., AND YUE, L. Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research 7*, 4 (2015), 286–303.

[7] ALTHOFF, T., WHITE, R. W., AND HORVITZ, E. Influence of pokémon go on physical activity: Study and implications. *Journal of Medical Internet Research 18*, 12 (2016).

[8] ARNOLD, W. F., AND LAUB, A. J. Generalized eigenproblem algorithms and software for algebraic Riccati equations. *Proceedings of the IEEE 72*, 12 (1984), 1746–1754.

[9] ATAN, O., JORDON, J., AND VAN DER SCHAAR, M. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[10] ATHEY, S., AND IMBENS, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*, 27 (2016), 7353–7360.

[11] BALLINGER, B., HSIEH, J., SINGH, A., SOHONI, N., WANG, J., TISON, G. H., MARCUS, G. M., SANCHEZ, J. M., MAGUIRE, C., OLGIN, J. E., ET AL. Deepheart: semi-supervised sequence learning for cardiovascular risk prediction. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[12] BATES, S. Progress towards personalized medicine. *Drug discovery today 15*, 3-4 (2010), 115–120.

[13] BAUER, D. J., AND CURRAN, P. J. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological methods 8*, 3 (2003), 338.

[14] BERGER, J. O., WANG, X., AND SHEN, L. A bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics 24*, 1 (2014), 110–129.

[15] BERNDSEN, J., SMYTH, B., AND LAWLOR, A. Pace my race: recommendations for marathon running. In *Proceedings of the 13th ACM Conference on Recommender Systems* (2019), ACM, pp. 246–250.

[16] BISHOP, C. M. *Pattern recognition and machine learning*. springer, 2006.

[17] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[18] BREIMAN, L., FRIEDMAN, J., STONE, C. J., AND OLSHEN, R. A. *Classification and regression trees*. CRC press, 1984.

[19] BRYSON-BROCKMANN, W., AND ROLL, D. Single-case experimental designs in medical education: an innovative research method. *Academic Medicine* (1996).

[20] CARNETHON, M. R. Physical activity and cardiovascular disease: how much is enough? *American journal of lifestyle medicine 3*, 1_suppl (2009), 44S–49S.

[21] CASPERSEN, C. J., PEREIRA, M. A., AND CURRAN, K. M. Changes in physical activity patterns in the united states, by sex and cross-sectional age. *Medicine & Science in Sports & Exercise 32*, 9 (2000), 1601–1609.

[22] CHAMROUKHI, F. Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation. *Journal of Classification 33*, 3 (2016), 374–411.

[23] CHAMROUKHI, F. Unsupervised learning of regression mixture models with unknown number of components. *Journal of Statistical Computation and Simulation 86*, 12 (2016), 2308–2334.

[24] CHAMROUKHI, F., SAMÉ, A., AKNIN, P., AND GOVAERT, G. Model-based clustering with hidden markov model regression for time series with regime changes. In *The 2011 International Joint Conference on Neural Networks* (2011), IEEE, pp. 2814–2821.

[25] CHEIFETZ, N., NOUMIR, Z., SAMÉ, A., SANDRAZ, A.-C., FÉLIERS, C., AND HEIM, V. Modeling and clustering water demand patterns from real-world smart meter data. *Drinking Water Engineering and Science 2*, 10 (2017), pp–75.

[26] CHEKROUD, S. R., GUEORGUIEVA, R., ZHEUTLIN, A. B., PAULUS, M., KRUMHOLZ, H. M., KRYSTAL, J. H., AND CHEKROUD, A. M. Association between physical exercise and mental health in 1· 2 million individuals in the usa between 2011 and 2015: a cross-sectional study. *The lancet psychiatry 5*, 9 (2018), 739–746.

[27] CHIANG, P.-H., AND DEY, S. Personalized effect of health behavior on blood pressure: Machine learning based prediction and recommendation. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)* (2018), IEEE, pp. 1–6.

[28] CHIPMAN, H. A., GEORGE, E. I., MCCULLOCH, R. E., ET AL. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics 4*, 1 (2010), 266–298.

[29] CHOUJAA, D., AND DULAY, N. Predicting human behaviour from selected mobile phone data points. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing* (New York, NY, USA, 2010), UbiComp '10, ACM, pp. 105–108.

[30] CHOW, G. C. Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society* (1960), 591–605.

[31] CONN, V. S., HAFDAHL, A. R., AND MEHR, D. R. Interventions to increase physical activity among healthy adults: meta-analysis of outcomes. *American journal of public health 101*, 4 (2011), 751–758.

[32] CUI, P., LIU, H., AGGARWAL, C., AND WANG, F. Uncovering and predicting human behaviors. *IEEE Intelligent Systems 31*, 2 (2016), 77–88.

[33] CUTURI, M., AND DOUCET, A. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning* (2014), pp. 685–693.

[34] DAEPPEN, J.-B., FAOUZI, M., SANGLIER, T., SANCHEZ, N., COSTE, F., AND BERTHOLET, N. Drinking patterns and their predictive factors in control: A 12-month prospective study in a sample of alcohol-dependent patients initiating treatment. *Alcohol and alcoholism 48*, 2 (2012), 189–195.

[35] DELLAERT, F. The expectation maximization algorithm. Tech. rep., Georgia Institute of Technology, 2002.

[36] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)* (1977), 1–38.

[37] DIMAKOPOULOU, M., ATHEY, S., AND IMBENS, G. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077* (2017).

[38] DIREITO, A., DALE, L. P., SHIELDS, E., DOBSON, R., WHITTAKER, R., AND MADDISON, R. Do physical activity and dietary smartphone applications incorporate evidence-based behaviour change techniques? *BMC public health 14*, 1 (2014), 646.

[39] DREWNOWSKI, A., AND EVANS, W. J. Nutrition, physical activity, and quality of life in older adults: summary. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 56*, suppl_2 (2001), 89–94.

[40] DREYER, N. A., TUNIS, S. R., BERGER, M., OLLENDORF, D., MATTOX, P., AND GLIKLICH, R. Why observational studies should be among the tools used in comparative effectiveness research. *Health Affairs 29*, 10 (2010), 1818–1825.

[41] DUSSELDORP, E., CONVERSANO, C., AND VAN OS, B. J. Combining an additive and tree-based regression model simultaneously: Stima. *Journal of Computational and Graphical Statistics 19*, 3 (2010), 514–530.

[42] DUSSELDORP, E., DOOVE, L., AND VAN MECHELEN, I. Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. *Behavior research methods 48*, 2 (2016), 650–663.

[43] DUSSELDORP, E., AND VAN MECHELEN, I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine 33*, 2 (2014), 219–237.

[44] EAGLE, N., AND PENTLAND, A. S. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology 63*, 7 (May 2009), 1057–1066.

[45] EVANS, J. J., GAST, D. L., PERDICES, M., AND MANOLOV, R. Single case experimental designs: Introduction to a special issue of neuropsychological rehabilitation. *Neuropsychological rehabilitation 24*, 3-4 (2014), 305–314.

[46] EVANS, W. J. Exercise training guidelines for the elderly. *Medicine and science in sports and exercise 31*, 1 (1999), 12–17.

[47] FARRAHI, K., AND GATICA-PEREZ, D. What did you do today?: discovering daily routines from large-scale mobile data. In *Proceedings of the 16th ACM international conference on Multimedia* (2008), ACM, pp. 849–852.

[48] FLAHERTY, B. P. Assessing reliability of categorical substance use measures with latent class analysis. *Drug and alcohol dependence 68* (2002), 7–20.

[49] FOREMAN, K. J., MARQUEZ, N., DOLGERT, A., FUKUTAKI, K., FULLMAN, N., MCGAUGHEY, M., PLETCHER, M. A., SMITH, A. E., TANG, K., YUAN, C.-W., ET AL. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet 392*, 10159 (2018), 2052–2090.

[50] FOSTER, J. C., TAYLOR, J. M., AND RUBERG, S. J. Subgroup identification from randomized clinical trial data. *Statistics in medicine 30*, 24 (2011), 2867–2880.

[51] GERS, F. A., SCHMIDHUBER, J., AND CUMMINS, F. Learning to forget: continual prediction with LSTM. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)* (Sep. 1999), vol. 2, pp. 850–855 vol.2.

[52] GILMARTIN-THOMAS, J. F., LIEW, D., AND HOPPER, I. Observational studies and their utility for practice. *Australian prescriber 41*, 3 (2018), 82.

[53] GRAND VIEW RESEARCH (GVR). mhealth app market by type (fitness, lifestyle management, nutrition & diet, women's health, healthcare providers, disease management) and segment forecasts, 2018–2025. Tech. rep., Grand View Research (GVR), 2017.

[54] GREAVES, C. J., SHEPPARD, K. E., ABRAHAM, C., HARDEMAN, W., RODEN, M., EVANS, P. H., AND SCHWARZ, P. Systematic review of reviews of intervention components associated with increased effectiveness in dietary and physical activity interventions. *BMC public health 11*, 1 (2011), 119.

[55] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V., AND COURVILLE, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems* (2017), pp. 5767–5777.

[56] HARVARD MEDICAL SCHOOL. Why behavior change is hard - and why you should keep trying. https://www.health.harvard.edu/mind-and-mood/why-behavior-change-is-hard-and-why-you-should-keep-trying. Accessed: 2019-11-07.

[57] HENDERSON, N. C., LOUIS, T. A., WANG, C., AND VARADHAN, R. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services and Outcomes Research Methodology 16*, 4 (2016), 213–233.

[58] HILL, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics 20*, 1 (2011), 217–240.

[59] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[60] IMAI, K., RATKOVIC, M., ET AL. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*, 1 (2013), 443–470.

[61] JAYANTI, R. K., AND BURNS, A. C. The antecedents of preventive health care behavior: An empirical study. *Journal of the academy of marketing science 26*, 1 (1998), 6–15.

[62] JOHANSSON, F., SHALIT, U., AND SONTAG, D. Learning representations for counterfactual inference. In *International conference on machine learning* (2016), pp. 3020–3029.

[63] JUNG, T., AND WICKRAMA, K. An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass 2*, 1 (2008), 302–317.

[64] KAPTEIN, M., DE RUYTER, B., MARKOPOULOS, P., AND AARTS, E. Adaptive Persuasive Systems: A Study of Tailored Persuasive Text Messages to Reduce Snacking. *ACM Transactions on Interactive Intelligent Systems 2*, 2 (2012), 1–25.

[65] KOTSEV, G., NGUYEN, L. T., ZENG, M., AND ZHANG, J. User exercise pattern prediction through mobile sensing. In *6th International Conference on Mobile Computing, Applications and Services* (2014), IEEE, pp. 182–188.

[66] KOUTNÍK, J., GREFF, K., GOMEZ, F. J., AND SCHMIDHUBER, J. A clockwork RNN. *CoRR abs/1402.3511* (2014).

[67] KULEV, I., PU, P., AND FALTINGS, B. Discovering persuasion profiles using time series data. In *Proceedings of the 2nd NIPS Time Series Workshop* (2016).

[68] KULEV, I., PU, P., AND FALTINGS, B. A Bayesian approach to intervention-based clustering. *ACM Trans. Intell. Syst. Technol. 9*, 4 (Jan. 2018), 44:1–44:23.

[69] KULEV, I., WALK, C., LU, Y., AND PU, P. Recommender system for responsive engagement of senior adults in daily activities. *Journal of Population Ageing* (under review).

[70] KURASHIMA, T., ALTHOFF, T., AND LESKOVEC, J. Modeling interdependent and periodic real-world action sequences. In *Proceedings of the... International World-Wide Web Conference. International WWW Conference* (2018), vol. 2018, NIH Public Access, p. 803.

[71] KVAAVIK, E., BATTY, G. D., URSIN, G., HUXLEY, R., AND GALE, C. R. Influence of individual and combined health behaviors on total and cause-specific mortality in men and women: the united kingdom health and lifestyle survey. *Archives of internal medicine 170*, 8 (2010), 711–718.

[72] LACROIX, J., SAINI, P., AND HOLMES, R. The relationship between goal difficulty and performance in the context of a physical activity intervention program. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services* (2008), ACM, pp. 415–418.

[73] LEE, C., MASTRONARDE, N., AND VAN DER SCHAAR, M. Estimation of individual treatment effect in latent confounder models via adversarial learning. *arXiv preprint arXiv:1811.08943* (2018).

[74] LEE, I.-M., SHIROMA, E. J., KAMADA, M., BASSETT, D. R., MATTHEWS, C. E., AND BURING, J. E. Association of step volume and intensity with all-cause mortality in older women. *JAMA internal medicine* (2019).

[75] LEE, I.-M., SHIROMA, E. J., LOBELO, F., PUSKA, P., BLAIR, S. N., KATZMARZYK, P. T., GROUP, L. P. A. S. W., ET AL. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The lancet 380*, 9838 (2012), 219–229.

[76] LEITZMANN, M. F., PARK, Y., BLAIR, A., BALLARD-BARBASH, R., MOUW, T., HOLLENBECK, A. R., AND SCHATZKIN, A. Physical activity recommendations and decreased risk of mortality. *Archives of internal medicine 167*, 22 (2007), 2453–2460.

[77] LEWIS, B. A., NAPOLITANO, M. A., BUMAN, M. P., WILLIAMS, D. M., AND NIGG, C. R. Future directions in physical activity intervention research: expanding our focus to sedentary behaviors, technology, and dissemination. *Journal of behavioral medicine 40*, 1 (2017), 112–126.

[78] LIPKOVICH, I., DMITRIENKO, A., DENNE, J., AND ENAS, G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine 30*, 21 (2011), 2601–2621.

[79] LOH, W.-Y., FU, H., MAN, M., CHAMPION, V., AND YU, M. Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine 35*, 26 (2016), 4837–4855.

[80] LOH, W.-Y., HE, X., AND MAN, M. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine 34*, 11 (2015), 1818–1833.

[81] LÖNNBERG, L., EKBLOM-BAK, E., AND DAMBERG, M. Improved unhealthy lifestyle habits in patients with high cardiovascular risk: results from a structured lifestyle programme in primary care. *Upsala journal of medical sciences 124*, 2 (2019), 94–104.

[82] LOUIZOS, C., SHALIT, U., MOOIJ, J. M., SONTAG, D., ZEMEL, R., AND WELLING, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* (2017), pp. 6446–6456.

[83] LUŠTREK, M., AND KALUŽA, B. Fall detection and activity recognition with machine learning. *Informatica 33*, 2 (2009).

[84] MA, F., CHITTA, R., ZHOU, J., YOU, Q., SUN, T., AND GAO, J. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (2017), ACM, pp. 1903–1911.

[85] MAATEN, L. V. D., AND HINTON, G. Visualizing data using t-SNE. *Journal of machine learning research 9*, Nov (2008), 2579–2605.

[86] MICHIE, S., YARDLEY, L., WEST, R., PATRICK, K., AND GREAVES, F. Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop. *Journal of medical Internet research 19*, 6 (2017), e232.

[87] MILANOVIĆ, Z., PANTELIĆ, S., TRAJKOVIĆ, N., SPORIŠ, G., KOSTIĆ, R., AND JAMES, N. Age-related decrease in physical activity and functional fitness among elderly men and women. *Clinical interventions in aging 8* (2013), 549.

[88] MIYATO, T., AND KOYAMA, M. cGANs with projection discriminator. *arXiv preprint arXiv:1802.05637* (2018).

[89] MOLLER, A. C., MERCHANT, G., CONROY, D. E., WEST, R., HEKLER, E., KUGLER, K. C., AND MICHIE, S. Applying and advancing behavior change theories and techniques in the context of a digital health revolution: proposals for more effectively realizing untapped potential. *Journal of behavioral medicine 40*, 1 (2017), 85–98.

[90] MÜCK, J. E., ÜNAL, B., BUTT, H., AND YETISEN, A. K. Market and patent analyses of wearables in medicine. *Trends in biotechnology* (2019).

[91] MULLENBACH, J., WIEGREFFE, S., DUKE, J., SUN, J., AND EISENSTEIN, J. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695* (2018).

[92] MURPHY, K. P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[93] MUTHÉN, B. Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences 345* (2004), 368.

[94] MUTHÉN, B., BROWN, C. H., MASYN, K., JO, B., KHOO, S.-T., YANG, C.-C., WANG, C.-P., KELLAM, S. G., CARLIN, J. B., AND LIAO, J. General growth mixture modeling for randomized preventive interventions. *Biostatistics 3*, 4 (2002), 459–475.

[95] NA, K.-S. Prediction of future cognitive impairment among the community elderly: A machine-learning based approach. *Scientific reports 9*, 1 (2019), 3335.

[96] NG, A. Machine learning and ai via brain simulations. *Accessed: May 3* (2013), 2018.

[97] NICHOLS, A., ET AL. Causal inference with observational data. *Stata Journal 7*, 4 (2007), 507.

[98] ORY, M. G., AND COX, D. M. Forging ahead: Linking health and behavior to improve quality of life in older people. *Social Indicators Research 33*, 1-3 (1994), 89–120.

[99] PARK, C.-H., ELAVSKY, S., AND KOO, K.-M. Factors influencing physical activity in older adults. *Journal of exercise rehabilitation 10*, 1 (2014), 45.

[100] PEYSAKHOVICH, A., AND LADA, A. Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385* (2016).

[101] PHATAK, S. S., FREIGOUN, M. T., MARTÍN, C. A., RIVERA, D. E., KORINEK, E. V., ADAMS, M. A., BUMAN, M. P., KLASNJA, P., AND HEKLER, E. B. Modeling individual differences: A case study of the application of system identification for personalizing a physical activity intervention. *Journal of biomedical informatics 79* (2018), 82–97.

[102] PREUM, S. M., STANKOVIC, J. A., AND QI, Y. Maper: A multi-scale adaptive personalized model for temporal human behavior prediction. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (New York, NY, USA, 2015), CIKM '15, ACM, pp. 433–442.

[103] PRINCE, M. J., WU, F., GUO, Y., ROBLEDO, L. M. G., O'DONNELL, M., SULLIVAN, R., AND YUSUF, S. The burden of disease in older people and implications for health policy and practice. *The Lancet 385*, 9967 (2015), 549–562.

[104] RABBI, M., CHOUDHURY, T., PFAMMATTER, A., ZHANG, M., AND SPRING, B. Automated Personalized Feedback for Physical Activity and Dietary Behavior Change With Mobile Phones: A Randomized Controlled Trial on Adults. *JMIR mHealth and uHealth 3*, 2 (2015), e42.

[105] RAJKOMAR, A., OREN, E., CHEN, K., DAI, A. M., HAJAJ, N., HARDT, M., LIU, P. J., LIU, X., MARCUS, J., SUN, M., ET AL. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine 1*, 1 (2018), 18.

[106] RAJPURKAR, P., HANNUN, A. Y., HAGHPANAHI, M., BOURN, C., AND NG, A. Y. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836* (2017).

[107] ROBINSON, S. A., BISSON, A. N., HUGHES, M. L., EBERT, J., AND LACHMAN, M. E. Time for change: using implementation intentions to promote physical activity in a randomised pilot trial. *Psychology & health 34*, 2 (2019), 232–254.

[108] ROSENBAUM, P. R., AND RUBIN, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* (1983), 41–55.

[109] ROY, T., AND LLOYD, C. E. Epidemiology of depression and diabetes: a systematic review. *Journal of affective disorders 142* (2012), S8–S21.

[110] RUBIN, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*, 5 (1974), 688.

[111] RUBIN, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association 100*, 469 (2005), 322–331.

[112] SAITO, J., KONDO, N., SAITO, M., TAKAGI, D., TANI, Y., HASEDA, M., TABUCHI, T., AND KONDO, K. Exploring 2.5-year trajectories of functional decline in older adults by applying a growth mixture model and frequency of outings as a predictor: A 2010–2013 jages longitudinal study. *Journal of epidemiology* (2018), JE20170230.

[113] SAKIA, R. The box-cox transformation technique: a review. *Journal of the Royal Statistical Society: Series D (The Statistician) 41*, 2 (1992), 169–178.

[114] SASAKI, J. E., HICKEY, A., STAUDENMAYER, J., JOHN, D., KENT, J. A., AND FREEDSON, P. S. Performance of activity classification algorithms in free-living older adults. *Medicine and science in sports and exercise 48*, 5 (2016), 941.

[115] SAU, A., AND BHAKTA, I. Predicting anxiety and depression in elderly patients using machine learning technology. *Healthcare Technology Letters 4*, 6 (2017), 238–243.

[116] SCHWAB, P., LINHARDT, L., AND KARLEN, W. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656* (2018).

[117] SEEFELDT, V., MALINA, R. M., AND CLARK, M. A. Factors affecting levels of physical activity in adults. *Sports medicine 32*, 3 (2002), 143–168.

[118] SHAHN, Z., MADIGAN, D., ET AL. Latent class mixture models of treatment effect heterogeneity. *Bayesian Analysis* (2016).

[119] SHALIT, U., JOHANSSON, F. D., AND SONTAG, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 3076–3085.

[120] SHALIZI, C. *Advanced data analysis from an elementary point of view*. Citeseer, 2013.

[121] SIBBALD, B., AND ROLAND, M. Understanding controlled trials. why are randomised controlled trials important? *BMJ: British Medical Journal 316*, 7126 (1998), 201.

[122] SIGNORETTI, A., MARTINS, A. I., ALMEIDA, N., VIEIRA, D., ROSA, A. F., COSTA, C. M., AND TEXEIRA, A. Trip 4 all: A gamified app to provide a new way to elderly people to travel. *Procedia Computer Science 67* (2015), 301–311.

[123] SLOAN, R. A., KIM, Y., SAHASRANAMAN, A., MÜLLER-RIEMENSCHNEIDER, F., BIDDLE, S. J., AND FINKELSTEIN, E. A. The influence of a consumer-wearable activity tracker on sedentary time and prolonged sedentary bouts: secondary analysis of a randomized controlled trial. *BMC research notes 11*, 1 (2018), 189.

[124] SMITH, R., CORRIGAN, P., EXETER, C., GROUP, N.-C. D. W., ET AL. *Countering Non-communicable Disease Through Innovation: Report of the Non-Communicable Disease Working Group 2012*. Global Health Policy Summit, 2012.

[125] SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M., AND LI, B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research 10*, Feb (2009), 141–158.

[126] SULLIVAN, A. N., AND LACHMAN, M. E. Behavior change with fitness technology in sedentary adults: a review of the evidence for increasing physical activity. *Frontiers in public health 4* (2017), 289.

[127] SWARTZ, A. M., TARIMA, S., MILLER, N. E., HART, T. L., GRIMM, E. K., ROTE, A. E., AND STRATH, S. J. Prediction of body fat in older adults by time spent in sedentary behavior. *Journal of aging and physical activity 20*, 3 (2012), 332–344.

[128] THE, L. P. H. Ageing: a 21st century public health challenge? *The Lancet. Public health 2*, 7 (2017), e297.

[129] UMEK, L., AND ZUPAN, B. Subgroup discovery in data sets with multi-dimensional responses. *Intelligent Data Analysis 15*, 4 (2011), 533–549.

[130] UNITED NATIONS, D. O. E., AND AFFAIRS, S. World population ageing 2017: highlights, 2017.

[131] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.

[132] VICKERS, A. J. Whose data set is it anyway? sharing raw data from randomized trials. *Trials 7*, 1 (2006), 15.

[133] VICKERS, A. J., REES, R. W., ZOLLMAN, C. E., MCCARNEY, R., SMITH, C. M., ELLIS, N., FISHER, P., AND VAN HASELEN, R. Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *Bmj 328*, 7442 (2004), 744.

[134] WAGER, S., AND ATHEY, S. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342* (2015).

[135] WANG, Y., FADHIL, A., AND REITERER, H. Health behavior change in HCI: trends, patterns, and opportunities. *CoRR abs/1901.10449* (2019).

[136] WARE JR., J. E. *SF-36 health survey.* Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1999, ch. SF-36 Health Survey., pp. 1227–1246.

[137] WILLKE, R. J., ZHENG, Z., SUBEDI, P., ALTHIN, R., AND MULLINS, C. D. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC medical research methodology 12*, 1 (2012), 185.

[138] XIE, J., WEN, D., LIANG, L., JIA, Y., GAO, L., AND LEI, J. Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: comparative study. *JMIR mHealth and uHealth 6*, 4 (2018), e94.

[139] YI, D., SU, J., LIU, C., AND CHEN, W.-H. Trajectory clustering aided personalized driver intention prediction for intelligent vehicles. *IEEE Transactions on Industrial Informatics* (2018).

[140] YOON, J., JORDON, J., AND VAN DER SCHAAR, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations* (2018).

[141] YÜRÜTEN, O., ZHANG, J., AND PU, P. Decomposing activities of daily living to discover routine clusters. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), AAAI Press, pp. 1348–1354.

[142] ZEEVI, D., KOREM, T., ZMORA, N., ISRAELI, D., ROTHSCHILD, D., WEINBERGER, A., BEN-YACOV, O., LADOR, D., AVNIT-SAGI, T., LOTAN-POMPAN, M., ET AL. Personalized nutrition by prediction of glycemic responses. *Cell 163*, 5 (2015), 1079–1094.

[143] ZEILEIS, A., HOTHORN, T., AND HORNIK, K. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 492–514.

[144] ZHANG, D., SUN, Y., ERIKSSON, B., AND BALZANO, L. Deep unsupervised clustering using mixture of autoencoders. *arXiv preprint arXiv:1712.07788* (2017).

[145] ZHANG, Y., LAI, G., ZHANG, M., ZHANG, Y., LIU, Y., AND MA, S. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (2014), ACM, pp. 83–92.

[146] ZHANG, Y.-Z., IMOTO, S., MIYANO, S., AND YAMAGUCHI, R. Reconstruction of high read-depth signals from low-depth whole genome sequencing data using deep learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2017), IEEE, pp. 1227–1232.

[147] ZHENG, J., AND NI, L. M. An unsupervised framework for sensing individual and cluster behavior patterns from human mobile data. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (2012), ACM, pp. 153–162.

IGOR KULEV

INR 236, Station 14 ⋄ 1015 Lausanne, Switzerland
+41 21 693 66 8 ⋄ igor.kulev@epfl.ch

EDUCATION

École Polytechnique Fédérale de Lausanne, Switzerland *September 2014 - Present*
PhD degree in Computer Science
School of Computer and Communication Sciences

Ss. Cyril and Methodius University, Macedonia *November 2011 - April 2013*
Master degree in Electrical Engineering and Information Technologies
GPA 10.0/10.0
Faculty of Computer Science and Engineering

Ss. Cyril and Methodius University, Macedonia *September 2007 - June 2011*
Bachelor degree in Electrical Engineering and Information Technologies
GPA 10.0/10.0
Faculty of Electrical Engineering and Information Technologies

INTERNSHIPS AND WORK EXPERIENCE

École Polytechnique Fédérale de Lausanne, Switzerland *September 2014 - Present*
Doctoral Assistant

I have participated in Horizon 2020 project "Responsive Engagement of the Elderly promoting Activity and Customized Healthcare". My role has been to analyze health data and develop methods to provide personalized interventions. I have been a Teaching Assistant in courses: Information, Computation and Communication; Intelligent Agents; Human Computer Interaction. I have supervised several semester projects.

Ss. Cyril and Methodius University, Macedonia *September 2011 - August 2014*
Teaching and Research Assistant

· Teaching Assistant in courses: Algorithms and data structures; Advanced Algorithms; Intelligent User Interfaces; Intelligent Information systems; Multimedia Systems; Computer architecture and organization. I also prepared teaching materials for most of these courses.

United States Agency for International Development *August 2012 - September 2012*
IT Consultant

· I developed a web-based questionnaire that was used in the Interethnic Integration in Education Project financed by USAID. The clients needed an application that will implement their complex questionnaire flow (more than 300 questions in 4 languages).

The European Organization for Nuclear Research (CERN) *June 2010 - August 2010*
Summer Student Intern

· I was working on the experiment AEGIS and my project was "Simulation of antihydrogen atoms under static magnetic and temporary varying electric fields".

ACHIEVEMENTS AND AWARDS

Award "Engineering ring" – Engineering Institution of Macedonia. The best engineering student of generation 2010/2011.

Award "Gold coin" – Ss. Cyril and Methodius University, Macedonia. The best student of generation 2010/2011.

Graduated with Honours, BSc. Eng., Degree, Electrical Engineering and Information Technology, Ss. Cyril and Methodius University, Skopje, Macedonia.

4th place and winners of Region 8 (Europe, Middle East and Africa) - IEEEXtreme 6.0 programming competition (2012).

1st place – National ACM-ICPC contest for algorithmic programming (2009 and 2010)

Johnson Controls fellowship (2007-2011).

Gold medal and 1st place at MOI 2007 (Macedonian Olympiad in Informatics).

TECHNICAL STRENGTHS

Programming languages	Python, Java, C++, JavaScript, C, Matlab, SQL, C#
Software & Tools	Tensorflow, Flask, Eclipse, .NET Framework, Microsoft Visual Studio

LANGUAGES

Macedonian	native speaker
English	full professional
French	basic

SELECTED PUBLICATIONS

Kulev, I., Walk, C., Lu, Y., and Pu, P. Recommender system for responsive engagement of senior adults in daily activities. *Journal of Population Ageing* (under review).

Kulev, I., Pu, P., and Faltings, B. A bayesian approach to intervention-based clustering. *ACMTrans. Intell. Syst. Technol.* 9, 4 (Jan. 2018), 44:1–44:23.

Kulev, I., Pu, P., and Faltings, B. Discovering persuasion profiles using time series data. In *Proceedings of the 2nd NIPS Time Series Workshop* (2016).

Vlahu-Gjorgievska, E., Koceski, S., **Kulev, I.**, and Trajkovik, V. Connected-health algorithm: Development and evaluation. *Journal of Medical Systems* 40, 4 (Feb 2016), 109.

Trajkovik, V., Vlahu-Gjorgievska, E., Koceski, S., and **Kulev, I**. General assisted living system architecture model. In Mobile Networks and Management (Cham, 2015), Springer International Publishing, pp. 329–343.