A Bayesian Approach to Intervention-Based Clustering

IGOR KULEV, PEARL PU, and BOI FALTINGS, École Polytechnique Fédérale de Lausanne

An important task for intelligent healthcare systems is to predict the effect of a new intervention on individuals. This is especially true for medical treatments. For example, consider patients who do not respond well to a new drug or have adversary reactions. Predicting the likelihood of positive or negative response before trying the drug on the patient can potentially save his or her life. We are therefore interested in identifying distinctive subpopulations that respond differently to a given intervention. For this purpose, we have developed a novel technique, Intervention-based Clustering, based on a Bayesian mixture model. Compared to the baseline techniques, the novelty of our approach lies in its ability to model complex decision boundaries by using soft clustering, thus predicting the effect for individuals more accurately. It can also incorporate prior knowledge, making the method useful even for smaller datasets. We demonstrate how our method works by applying it to both simulated and real data. Results of our evaluation show that our model has strong predictive power and is capable of producing high-quality clusters compared to the baseline methods.

CCS Concepts: •**Information systems** →**Data mining;**

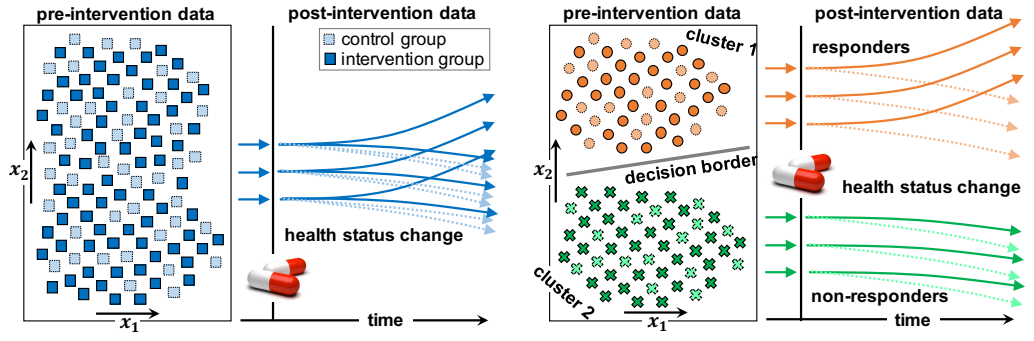
Additional Key Words and Phrases: Clustering, Bayesian analysis, mixture model, heterogeneous treatment effects, subgroup analysis, randomized controlled trial, cross-validation, personalized medicine

ACM Reference format:
Igor Kulev, Pearl Pu, and Boi Faltings. 2018. A Bayesian Approach to Intervention-Based Clustering. *ACM Trans. Intell. Syst. Technol.* 9, 4, Article 44 (January 2018), 23 pages.
DOI: 10.1145/3156683

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, and exercises in teaching. In this article, we focus on the example of medical treatments, 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 a randomized controlled trial (RCT) 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, then 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 Fig. 1b) improves after taking a medication while another group (green crosses in Fig. 1b) does not. In both cases, effect changes are compared to the baseline (orange and green dashed lines). We may not find this difference if we used effect averages (Fig. 1a). While some people improved (blue solid lines going up in Fig. 1a), overall the health status of a population did not change significantly. Our goal is to sub-divide the population into clusters taking into account their



(a) Conventional approach (b) Intervention-based clustering approach

Fig. 1. Modeling the treatment effects of a given population: conventional vs. our method

respective response to the intervention (Fig. 1b). In this manner, we will be able to decide whether 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. Previous work in discovering the heterogeneity of the treatment effect (HTE) has addressed some of the challenges. Compared to these baseline methods, we are providing the following advantages:

A More Accurate Model of Subpopulations. The subpopulations with differential treatment effect 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 effect. 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, then we can 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 tradeoff 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.

Various methods that identify heterogeneous groups have been investigated in the literature [24]. The novelty of our approach is that it creates complex and more accurate decision boundaries and allows reasoning about the tradeoff 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 she is to respond to it.

This article 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 sections, we apply our approach on both synthetic and real data, and we compare it with two existing methods, qualitative interaction trees (QUINT) and the 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.

2 RELATED WORK

Identifying the causal effect of a treatment on a patient is a difficult problem. To make 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 a treatment is applied on 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 [23]. The causal effect of a treatment for that population can be estimated by comparing their average outcomes. The ATE can be easily estimated without bias in randomized experiments [15]. However, treatments might have different causal effects on each subject. Existing work is focused on either estimating the patient-level treatment effect [34], or searching for subgroups with differential treatment effects [19]. 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.

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 [37]. These methods essentially establish a relationship between attributes and outcomes, with a penalty parameter that penalizes model complexity [2]. Recently, Wager and Athley [34] developed a non-parametric causal forest that extends Breiman's widely used random forest algorithm [6]. The method utilizes the strength of the random forests to model interactions in high dimensions and provides asympthotically unbiased and normal estimates of CATE under the assumption of randomization conditional on the covariates or "unfoundedness" [25].

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 effect than simply estimating the patient-level CATE. For example, existence of subgroups that appear to respond differently to treatment can affect inclusion criteria in later clinical trials or in labeling decisions for approved drugs [1, 14]. 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 [10]. 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 to 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 [37]. Several different tree-based methods have been developed, including simultaneous threshold interaction modeling (STIMA) [9], Interaction Trees

[30], Model-based recursive partitioning [37], Virtual Twins [13], subgroup identification based on differential effect search (SIDES) [17], and QUINT (qualitative interaction trees) [10, 11].

Interaction Trees[30] follow the CART [7] 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 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[37] 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., sum of squared errors and minus the loglikelihood. A splitting is done if the parameter estimates are not stable with respect to at least one partitioning variable.

Virtual Twins[13] is based on the concept of potential outcomes [27]. The method consists of two steps. In the first step, a random forest is applied on the data in order 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*[10, 11] is to identify subgroups that are involved in an 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 samples 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 the space is always splitted using a hyperplane perpendicular to one of the axes and parallel to the other axes. This is why these methods may not fully identify additive impact of multiple variables [36]. 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 effect measured at one instance after the intervention. However, when we are working with longitudinal data [26], we are interested in knowing how the treatment effect develops during the time after the intervention. Bauer and Curran [3] 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) [16] and *Growth Mixture Modeling* (GMM) [21] 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 [16]. In addition to the pre-intervention variables, these methods include time-related variables and, optionally, time-varying variables [21], 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 [16]. These models mostly have been applied to non-interventional data [12, 21], however, they have also been successfully used to analyze interventions, for example, interventions aimed at reducing aggressive behavior [22]. 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 of 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 [36]. 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.

2.3 Comparison with our Approach

In our work we have developed a Bayesian mixture model that is suitable for identifying the subpopulations with differential treatment effect. We consider that each person responds to the treatment in a particular way that 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 other researchers began using Bayesian approaches. For example, a Bayesian tree-based approach was proposed by Berger et al. [4]. Unfortunately, this method is not able to discover clusters with complex (nonlinear) decision boundaries. In another recent work, Shahn and Madigan [28] 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 datasets with multi-dimensional responses [18, 31], 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 effect: complex decision boundaries, multi-dimensional continuous outcomes, soft cluster membership, and the ability to stabilize the highly variable treatment effect estimates.

3 MODEL

RCT are the most rigorous way of determining whether a cause-effect relation exists between treatment and outcome [29]. 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 under treatment and its environment, for example: age, gender, education, medical condition, rainy weather, and so on. The treatment variables represent the type of intervention the subject received, for example: drug, or a persuasion message delivered in a mobile phone app, and so on. The outcome variables represent the outcome of interest, for example, well-being or health status change. The outcome is usually the change of a variable some time after the intervention.

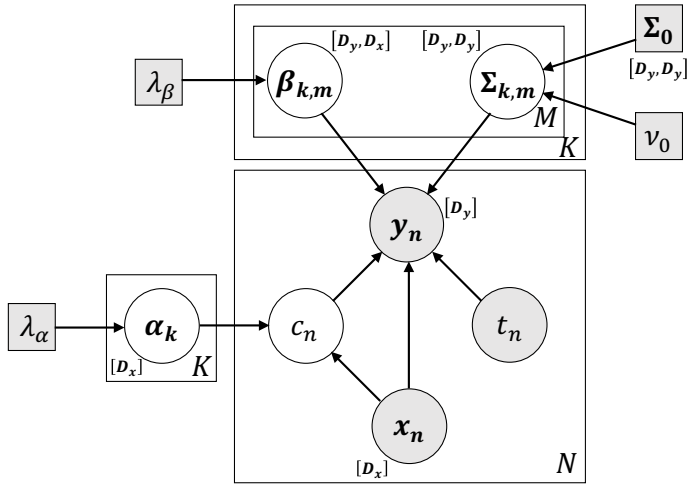


Fig. 2. Plate notation for our graphical model. The observed variables are displayed in gray circles, the unobserved variables are displayed in white circles and the hyper-parameters are displayed in gray squares. The dimensions of the multidimensional variables are displayed next to the variables' names.

After conducting RCT, the analysis is focused on estimating the size of the difference in predefined outcomes between intervention groups [29]. However, people with different baseline characteristics might respond to the same treatment in a different way. We propose 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 (Fig. 2). Let us denote the pre-intervention, treatment, and outcome variables associated to 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 discrete 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 discrete variable $c_n \in \{1, 2, \ldots, K\}$ that identifies the type of response to the intervention (discrete heterogeneity of the treatment effect) and that is hidden. There are K different types of responses and each person is associated to one of them. Observing the person's characteristics x_n we would like to determine the prior odds for 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, then 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' individual responses to treatment within the same cluster. We define $p(c_n = k|x_n, \alpha)$ to be a softmax function:

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

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 1 in order to interpret α_k^1 as the intercept. α_K should be a zero vector that is not affected in the learning process in order 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:

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

In our model, we associate the following prior distributions to the parameters:

$$p(\alpha) = \prod_{k=1}^{K}\prod_{i=2}^{D_x}\mathcal{N}\left(\alpha_k^i|0, \frac{1}{\lambda_\alpha}\right) = \prod_{k=1}^{K}\prod_{i=2}^{D_x}\frac{1}{\sqrt{2\pi\frac{1}{\lambda_\alpha}}}\exp\left(-\frac{1}{2}\lambda_\alpha{\alpha_k^i}^2\right) \quad (3)$$

$$p(\beta) = \prod_{k=1}^{K}\prod_{m=1}^{M}\prod_{i=1}^{D_y}\prod_{j=2}^{D_x}\mathcal{N}\left(\beta_{k,m}^{i,j}|0, \frac{1}{\lambda_\beta}\right) = \prod_{k=1}^{K}\prod_{m=1}^{M}\prod_{i=1}^{D_y}\prod_{j=2}^{D_x}\frac{1}{\sqrt{2\pi\frac{1}{\lambda_\beta}}}\exp\left(-\frac{1}{2}\lambda_\beta{\beta_{k,m}^{i,j}}^2\right) \quad (4)$$

$$p(\Sigma) = \prod_{k=1}^{K}\prod_{m=1}^{M}W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right)$$
$$= \prod_{k=1}^{K}\prod_{m=1}^{M}\frac{\left|\Sigma_{k,m}\right|^{\frac{\nu_0 - D_y - 1}{2}}\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma_0^{-1}\Sigma_{k,m}\right)\right]}{2^{\frac{\nu_0 D_y}{2}}\left|\Sigma_0\right|^{\frac{\nu_0}{2}}\Gamma_{D_y}\left(\frac{\nu_0}{2}\right)} \quad (5)$$

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

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:

$$\arg\max_{\alpha,\beta,\Sigma} p(\alpha, \beta, \Sigma|Y, X, T) = \arg\max_{\alpha,\beta,\Sigma}\frac{p(Y|X, T, \alpha, \beta, \Sigma)\, p(\alpha, \beta, \Sigma)}{\int\int\int p(Y|X, T, \alpha, \beta, \Sigma)\, p(\alpha, \beta, \Sigma)\, d\alpha\, d\beta\, d\Sigma} \quad (6)$$

$$= \arg\max_{\alpha,\beta,\Sigma} p(Y|X, T, \alpha, \beta, \Sigma)\, p(\alpha, \beta, \Sigma) \quad (7)$$

When we logarithmize the product of probabilities in Eq. 7 we obtain the following function:

$$\log p(Y, \alpha, \beta, \Sigma|X, T) = \sum_{n=1}^{N}\log\left[\sum_{k=1}^{K}\mathcal{N}\left(y_n|\beta_{k,t_n}x_n, \Sigma_{k,t_n}\right)\frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{i=1}^{K}\exp\left(\alpha_i^T x_n\right)}\right]$$
$$+ \sum_{k=1}^{K}\sum_{i=2}^{D_x}\log\mathcal{N}\left(\alpha_k^i|0, \frac{1}{\lambda_\alpha}\right) + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{D_y}\sum_{j=2}^{D_x}\log\mathcal{N}\left(\beta_{k,m}^{i,j}|0, \frac{1}{\lambda_\beta}\right)$$
$$+ \sum_{k=1}^{K}\sum_{m=1}^{M}\log W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right) \quad (8)$$

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

$$p_{n,k}^{(l)} = p\left(c_n = k|y_n, x_n, t_n, \alpha^{(l)}, \beta^{(l)}, \Sigma^{(l)}\right) = \frac{\mathcal{N}\left(y_n|\beta_{k,t_n}^{(l)} x_n, \Sigma_{k,t_n}^{(l)}\right) p\left(c_n = k|x_n, \alpha^{(l)}\right)}{\sum_{i=1}^{K} \mathcal{N}\left(y_n|\beta_{i,t_n}^{(l)} x_n, \Sigma_{n,t_n}^{(l)}\right) p\left(c_n = i|x_n, \alpha^{(l)}\right)} \tag{9}$$

We use the estimated posterior and Jensen's inequality to find the lower bound of Eq. 8:

$$\begin{aligned}
\log p\left(Y, \alpha, \beta, \Sigma|X, T\right) \geqslant & \sum_{n=1}^{N}\sum_{k=1}^{K} p_{n,k}^{(l)} \left[\log \mathcal{N}\left(y_n|\beta_{k,t_n} x_n, \Sigma_{k,t_n}\right) + \log \frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{i=1}^{K} \exp\left(\alpha_i^T x_n\right)}\right] \\
& + \sum_{k=1}^{K}\sum_{i=2}^{D_x} \log \mathcal{N}\left(\alpha_k^i|0, \frac{1}{\lambda_\alpha}\right) + \sum_{k=1}^{K}\sum_{m=1}^{M}\sum_{i=1}^{D_y}\sum_{j=2}^{D_x} \log \mathcal{N}\left(\beta_{k,m}^{i,j}|0, \frac{1}{\lambda_\beta}\right) \\
& + \sum_{k=1}^{K}\sum_{m=1}^{M} \log W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right) = Q\left(\alpha, \beta, \Sigma|\alpha^{(l)}, \beta^{(l)}, \Sigma^{(l)}\right)
\end{aligned} \tag{10}$$

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

$$\begin{aligned}
\frac{\partial Q(\cdot)}{\partial \alpha_k} &= \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N}\sum_{i=1}^{K} p_{n,i}^{(l)} \log \frac{\exp\left(\alpha_i^T x_n\right)}{\sum_{j=1}^{K} \exp\left(\alpha_j^T x_n\right)} + \sum_{i=2}^{D_x} \log \mathcal{N}\left(\alpha_k^i|0, \frac{1}{\lambda_\alpha}\right)\right] \\
&= \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N}\sum_{i=1}^{K} p_{n,i}^{(l)}\left(\alpha_i^T x_n - \log \sum_{j=1}^{K} \exp\left(\alpha_j^T x_n\right)\right) - \frac{\lambda_\alpha}{2}\sum_{i=2}^{D_x} {\alpha_k^i}^2\right] \\
&= \frac{\partial}{\partial \alpha_k}\left[\sum_{n=1}^{N}\left(p_{n,k}^{(l)} \alpha_k^T x_n - \log \sum_{j=1}^{K} \exp\left(\alpha_j^T x_n\right)\right) - \frac{\lambda_\alpha}{2}\sum_{i=2}^{D_x} {\alpha_k^i}^2\right] \\
&= \sum_{n=1}^{N}\left[p_{n,k}^{(l)} - \frac{\exp\left(\alpha_k^T x_n\right)}{\sum_{j=1}^{K} \exp\left(\alpha_j^T x_n\right)}\right] x_n - \lambda_\alpha \bar{\alpha}_k = 0
\end{aligned} \tag{11}$$

where $\bar{\alpha}_k^1 = 0$, and $\bar{\alpha}_k^i = \alpha_k^i$ for all $i > 1$. There is no closed-form solution to the equation above. This is why we use gradient ascent to find the optimal parameter values for α_k. The derivative of

$Q(\cdot)$ with respect to $\beta_{k,m}$ is:

$$\begin{aligned}\frac{\partial Q(\cdot)}{\partial \beta_{k,m}} &= \frac{\partial}{\partial \beta_{k,m}}\left[\sum_{n=1}^{N} p_{n,k}^{(l)} \log \mathcal{N}\left(y_n|\beta_{k,t_n}x_n, \Sigma_{k,t_n}\right) + \sum_{i=1}^{D_y}\sum_{j=2}^{D_x} \log \mathcal{N}\left(\beta_{k,m}^{i,j}|0, \frac{1}{\lambda_\beta}\right)\right] \\ &= \frac{\partial}{\partial \beta_{k,m}}\left[-\frac{1}{2}\sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)} \left(y_n - \beta_{k,m}x_n\right)^T \Sigma_{k,m}^{-1}\left(y_n - \beta_{k,m}x_n\right) - \frac{\lambda_\beta}{2}\sum_{i=1}^{D_y}\sum_{j=2}^{D_x} {\beta_{k,m}^{i,j}}^2\right] \\ &= \Sigma_{k,m}^{-1}\sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)}\left(y_n - \beta_{k,m}x_n\right)x_n^T - \lambda_\beta \bar{\beta}_{k,m} = 0 \end{aligned} \tag{12}$$

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 $j > 1$. There is no closed-form solution to this equation as well, so we can use gradient ascent to find the optimal parameter values for $\beta_{k,m}$ if the optimal $\Sigma_{k,n}$ is given. The derivative of $Q(\cdot)$ with respect to $\Sigma_{k,m}$ is:

$$\begin{aligned}\frac{\partial Q(\cdot)}{\partial \Sigma_{k,m}} &= \frac{\partial}{\partial \Sigma_{k,m}}\left[\sum_{n=1}^{N}\sum_{i=1}^{K} p_{n,i}^{(l)} \log \mathcal{N}\left(y_n|\beta_{i,t_n}x_n, \Sigma_{i,t_n}\right) + \log W\left(\Sigma_{k,m}|\Sigma_0, \nu_0\right)\right] \\ &= \frac{\partial}{\partial \Sigma_{k,m}}\Bigg[-\frac{1}{2}\sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)}\left(\log\left|\Sigma_{k,m}\right| + \left(y_n - \beta_{k,m}x_n\right)^T \Sigma_{k,m}^{-1}\left(y_n - \beta_{k,m}x_n\right)\right) \\ &\qquad + \frac{1}{2}\left(\nu_0 - D_y - 1\right)\log\left|\Sigma_{k,m}\right| - \frac{1}{2}\mathrm{tr}\left(\Sigma_0^{-1}\Sigma_{k,m}\right)\Bigg] \\ &= -\frac{1}{2}\sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)}\left(\Sigma_{k,m}^{-1} - \Sigma_{k,m}^{-1}\left(y_n - \beta_{k,m}x_n\right)\left(y_n - \beta_{k,m}x_n\right)^T \Sigma_{k,m}^{-1}\right) \\ &\qquad + \frac{1}{2}\left(\nu_0 - D_y - 1\right)\Sigma_{k,m}^{-1} - \frac{1}{2}\Sigma_0^{-1} = 0 \end{aligned} \tag{13}$$

We transform Eq. 13 by multiplying from left and right by $\Sigma_{k,m}$ and after regrouping we get:

$$\begin{aligned}-\Sigma_{k,m}\Sigma_0^{-1}\Sigma_{k,m} &+ \left(\nu_0 - D_y - 1 - \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)}\right)\Sigma_{k,m} \\ &+ \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)}\left(y_n - \beta_{k,m}x_n\right)\left(y_n - \beta_{k,m}x_n\right)^T = 0 \end{aligned} \tag{14}$$

The solution $\Sigma_{k,m}$ to this equation for given $\beta_{k,m}$ is also a solution to the following Riccati equation:

$$A^T X E + E^T X A - \left(E^T X B + S\right) R^{-1}\left(B^T X E + S^T\right) + Q = 0 \tag{15}$$

where

$$X = \Sigma_{k,m} \tag{16}$$

$$A = \frac{1}{2}\left(\nu_0 - D_y - 1 - \sum_{n=1}^{N} \mathbb{1}(t_n = m)\, p_{n,k}^{(l)}\right) I_{D_y} \tag{17}$$

$$B = E = I_{D_y} \tag{18}$$

$$S = 0_{D_y} \tag{19}$$

$$R = \Sigma_0 \tag{20}$$

$$Q = \sum_{n=1}^{N} \mathbb{1}\left(t_n = m\right) p_{n,k}^{(l)} \left(y_n - \beta_{k,m} x_n\right) \left(y_n - \beta_{k,m} x_n\right)^T \tag{21}$$

This equation has unique solution if

$$\begin{bmatrix} Q & S \\ S_T & R \end{bmatrix} > 0 \tag{22}$$

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 us 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:

$$z^T \begin{bmatrix} Q & S \\ S_T & R \end{bmatrix} z = \begin{bmatrix} z^T \begin{bmatrix} Q \\ S^T \end{bmatrix} & z^T \begin{bmatrix} S \\ R \end{bmatrix} \end{bmatrix} z = \begin{bmatrix} z_1^T Q & z_2^T R \end{bmatrix} z = z_1^T Q z_1 + z_2^T R z_2 > 0 \tag{23}$$

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 heavily depends 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.

3.2 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 a 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 [5]. 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 in order 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 large number of samples.

There are two main advantages of the Gibbs sampling approach combined with importance sampling over EM. First, it is easier to implement, because we do not 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, there are three disadvantages of this approach. 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 sufficient confidence [20]. Third, importance sampling may not work well in high dimensions, because in this case most of the samples carry no useful information [20], 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 $\Sigma_{k,m}$ that could be high dimensional, depending on the dataset. We cannot separately sample each value in the matrix, because if we do this, then 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, then Gibbs sampling may also be suitable.

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 involves two or three subpopulations with different treatment effects. A good model should recognize the true subpopulations. We validate this (1) qualitatively by comparing the true decision boundaries with the inferred decision boundaries and (2) quantitatively by analyzing the prediction errors.

4.1 Complex decision boundaries

The goal of the first experiment is 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 Fig. 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:

$$Y_{1,1} \sim 1+\varepsilon;\ Y_{1,2} \sim 0+\varepsilon;\ Y_{2,1} \sim 0+\varepsilon;\ Y_{2,2} \sim 0+\varepsilon;\ Y_{3,1} \sim 0+\varepsilon;\ Y_{3,2} \sim 1+\varepsilon \tag{24}$$

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

We apply our approach[1] on the simulated data to discover the HTE and to identify the subpopulations that respond to the intervention in a different way. We do not 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 independent validation dataset with 10,000 subjects. The average log likelihood on the validation dataset is given in Table 1. The model with the highest generalization power is the model with $K = 3$ and $P = 2$. We observe that the model correctly identified the true number of subpopulations. In Fig. 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 is able to capture the true HTE and to identify the subpopulations with differential HTE, even

[1] our own Matlab implementation

Table 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

if they are separated with complex non-linear boundaries in the preintervention 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 is able to identify the treatment effect associated to the subpopulations. We should note that the prediction error is lower than the error obtained by linear regression model[2] (0.719198).

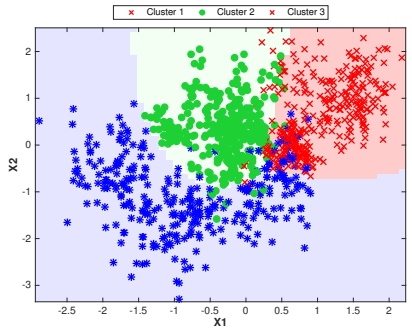


Fig. 3. Distribution of the data points in the first experiment. The color and the symbol associated to each patient indicate its true cluster membership. The background color and the decision borders indicate the most likely prior cluster membership generated by our model.

4.2 Comparison with tree-based methods

In the second experiment, we compare our method with the tree-based method QUINT[10, 11]. 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 [11]. 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 Fig. 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 in order to accommodate the QUINT model that cannot handle non-linear decision boundaries. We applied our approach and QUINT[3] on the simulated datasets, and we compared the results. For our approach, we set less informative prior on the model parameters ($\lambda_\alpha = 0.1$, $\lambda_\beta = 0.1$, $\nu_0 \in \{5, 10\}$, $\Sigma_0 = \mathrm{Cov}(Y)/\nu_0$) and we run cross-validation to find the optimal number of clusters ($K \in \{1, 2, 3, 4, 5\}$).

[2]polynomial regression with regularization
[3]R package quint with default parameters

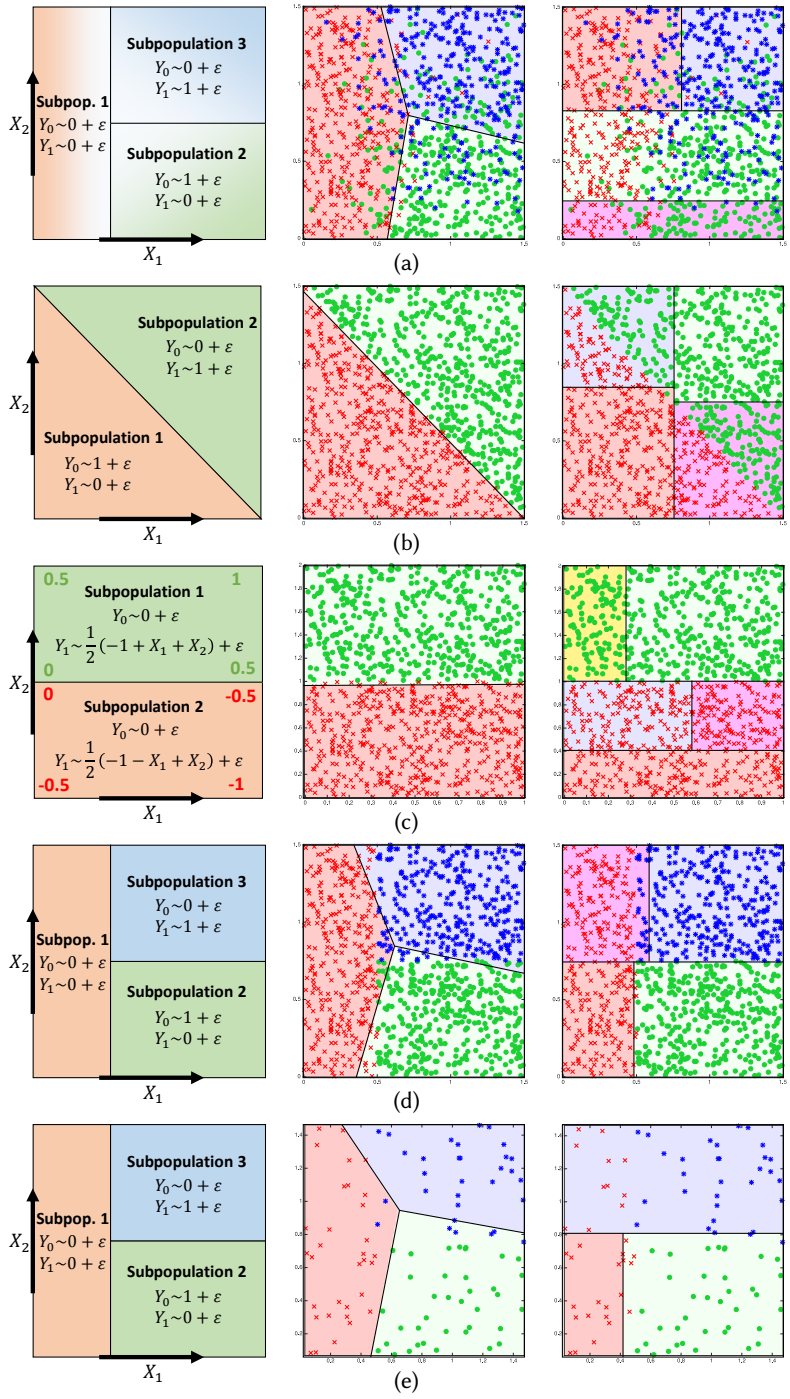


Fig. 4. The true subpopulations and the subpopulations discovered by our approach and QUINT in five different simulated datasets. The true ground truth model is shown on the left, and the results of our approach and QUNIT are shown in the middle and in the right, correspondingly. The background color corresponds to the regions discovered by our method and QUINT, and the color associated to each subject corresponds to its true subpopulation membership.

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 Fig. 4a. The darker color means higher certainty that a subject belonging to that region will belong to the subpopulation associated to 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 that 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. 2).

In the second simulated dataset, we generated 1000 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 (Fig. 4b). Our approach was accurate to identify the true subpopulations, but QUINT identified only half of the true positive and 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. 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.

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 Fig. 4c. QUINT does not have the ability to 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 to 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 discovered regions in the lower and the upper half of the space 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. 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 it can be seen in Fig. 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 with low RMSE, even lower than the RMSE produced by our method (Table. 2). We explain this by the fact that the decision boundaries discovered by IBC are not straight lines parallel to the axis. To

Table 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

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 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 effect and produces smaller prediction errors. We also applied 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 should not 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 effect.

5 EVALUATION ON ACUPUNCTURE DATA

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 3 months, in addition to standard care, or to a control intervention offering usual care [32, 33]. Headache score, SF-36 health status [35], and use of medication were assessed at baseline and at 3 and 12 months. The analysis of this dataset showed that acupuncture leads to persisting, clinically relevant benefits for primary care patients with chronic headache, particularly migraine [33]. We applied our method on the acupuncture data to discover homogeneous subpopulations who were affected by the intervention in similar 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 [35]. 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 differences between the energy and emotional well-being, assessed at 0 months and 12 months, are the outcome variables 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

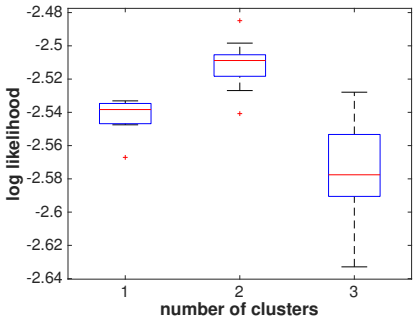


Fig. 5. Boxplot of the log likelihood on the validation dataset for different number of clusters $K \in \{1, 2, 3\}$. For each K, we repeated all cross-validations 10 times. We show the boxplot of the log likelihood on the validation dataset for the model with the optimal remaining hyper-parameters. The model with two clusters is suggested (statistically significant result with p-value < 0.00001).

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.

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 dataset. 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 how we ensure that the expected value of a Wishart random matrix is equal to the covariance matrix of the outcome variable. We choose four 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 in order to obtain higher relevance of the results. Our model selection procedure suggests a model with two clusters. This can be seen in Fig. 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 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 recommend the intervention to him or her. The most likely prior cluster membership for a given baseline energy and baseline emotional well-being, with age fixed to zero, is shown in Fig. 6. In the figure we can also see the prior cluster memberships for all the patients. Although we also use age information to determine the prior cluster membership for the patients, the clustering is not much different than the case when we set patient's age to zero. This indicates that age does not play a significant role in determining the prior odds for cluster membership, as can be seen in Table 3. On the other hand, emotional well-being (EW) is the most important variable in determining the prior odds for cluster membership, because, for each increase of EW by one unit (standard deviation), the odds of belonging to the first cluster increase by 0.52. The

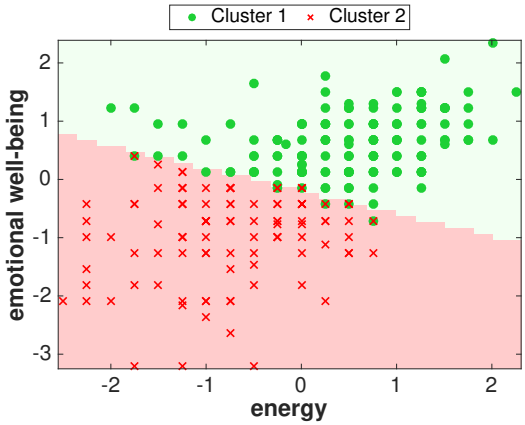


Fig. 6. The most likely prior cluster membership in the pre-intervention variable space (age is fixed to zero) and the most likely prior cluster membership for all patients.

Table 3. 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

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 Fig. 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 do not use the measurements of energy and EW 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. It is interesting that 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 acupuncture treatment in addition to standard care is recommended for them. We expect that

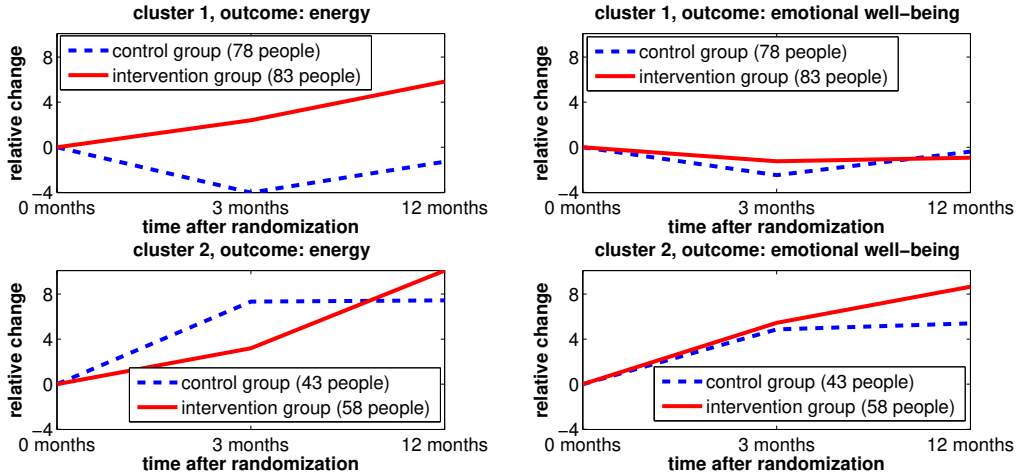


Fig. 7. Mean relative change of energy and emotional well-being in each cluster after the randomization. Each individual was assigned in the most likely cluster according to the prior odds for cluster membership.

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 does not treat the people who need it more, i.e., those having low energy and low EW.

5.3 Comparisons

In this section, we analyze the performance of our model and compare it with existing methods. We will perform a 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, 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 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 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 (Fig. 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 use for comparison is the one where a new user predicts that his or her response would be equal to the average response

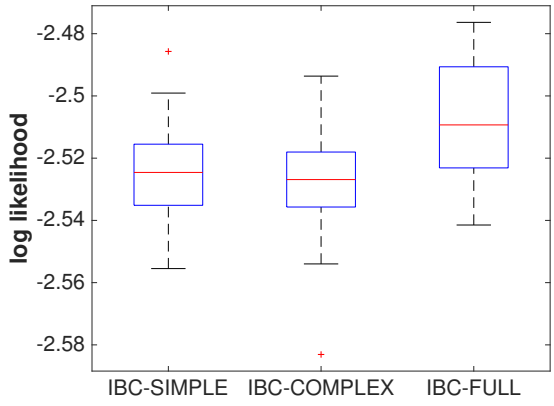


Fig. 8. Average log likelihood on the validation dataset produced by three versions of IBC which differ in their power to represent the impact of the intervention (from left to right, lowest to highest). The unconstrained model IBC-FULL has the best performance on the validation dataset (p-value < 0.01).

Table 4. RMSE on the validation dataset for nine 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**

in the treatment group he or she belongs in (BASELINE-mean-treatment). A second baseline is linear regression. We should note that a 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 on both responses, energy and EW, with different critical minimum values [10] and the best model was chosen in each case. The last model we compare with is GMM[4]. We defined two variants of GMM, GMM-COMPLEX and GMM-SIMPLE according to whether we allow the intervention indicator to interact with the pre-intervention variables or not. These models were separately applied on 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 in the same time (GMM-SIMPLE-2-outputs and GMM-COMPLEX-2-outputs). We applied all these models

[4]R package lcmm with default parameters

on the acupuncture dataset, and we compared their RMSE with the RMSE produced by our model. In Table 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.

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 will try to get more insight in 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 a RCT with 100 to 1,000 subjects based on this model. The pre-intervention data were 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.

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 Fig. 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, then 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 is powerful enough to describe the true data generating process.

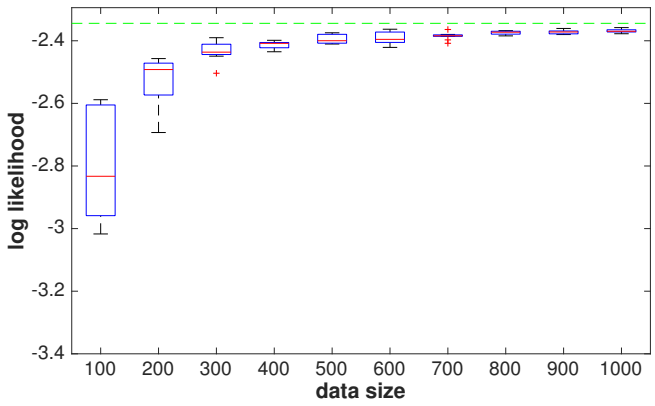


Fig. 9. Average log likelihood on a large independent test dataset obtained using models trained on data from 100 to 1,000 subjects. The dashed green line shows the optimal log likelihood obtained with the true model used to generate the data. If the trained model approximates well the true underlying model, then the average log likelihood associated to this model should be close to the optimal log likelihood. Sample size of 300 or more is required to obtain a good approximation of the true underlying model.

Otherwise, the results are no longer valid. For example, if there was a non-linear relationship between the pre-intervention and the outcome variables, then a linear model would not be able to discover the true model. In this case, we should have used polynomial features to increase the modeling power. We need to take into account that capturing more complex models with nonlinear decision borders, nonlinear responses, or larger numbers of clusters requires more data. For example, if we defined our true underlying model simulating the acupuncture dataset to consist of more than two clusters, then we would need more than 300 subjects to discover these clusters. It is likely that the acupuncture dataset we worked with is not large enough for us to discover more than two clusters if they were present.

7 CONCLUSION

The best treatment for the general population is not likely to be equally effective for each individual. Personalized medicine aims to provide treatments 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 who respond similarly to the same intervention. Identifying subpopulations with differential treatment effects is a methodologically challenging task, especially when many characteristics may interact with treatment and no comprehensive a priori hypotheses on relevant subgroups are available [10]. The most popular approaches for this purpose are based on trees, since trees provide features that are easily 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: It 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 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, then we might wrongly conclude that the intervention is not effective. However, when our method is applied on the data, it reveals subpopulations who respond differently than the overall response (if they exist). We also evaluated our algorithm on a 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 can 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 article provides a better insight into the way different people respond to the same intervention. This insight allows personalized medicine to provide more suitable tailored treatments. In the future, we plan to extend our method to time series and multiple interventions.

8 ACKNOWLEDGEMENTS

The work presented in this paper was made possible in part by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 690425.

REFERENCES

[1] Mohamed Alosh, Kathleen Fritsch, Mohammad Huque, Kooros Mahjoob, Gene Pennello, Mark Rothmann, Estelle Russek-Cohen, Fraser Smith, Stephen Wilson, and Lilly Yue. 2015. Statistical considerations on subgroup analysis in clinical trials. *Statistics in Biopharmaceutical Research* 7, 4 (2015), 286–303.

[2] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[3] Daniel J Bauer and Patrick J Curran. 2003. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological methods* 8, 3 (2003), 338.

[4] James O Berger, Xiaojing Wang, and Lei Shen. 2014. A Bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics* 24, 1 (2014), 110–129.

[5] Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[7] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.

[8] Frank Dellaert. 2002. *The expectation maximization algorithm*. Technical Report. Georgia Institute of Technology.

[9] Elise Dusseldorp, Claudio Conversano, and Bart Jan Van Os. 2010. Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical Statistics* 19, 3 (2010), 514–530.

[10] Elise Dusseldorp, Lisa Doove, and Iven Van Mechelen. 2016. 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.

[11] Elise Dusseldorp and Iven Van Mechelen. 2014. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine* 33, 2 (2014), 219–237.

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

[13] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. 2011. Subgroup identification from randomized clinical trial data. *Statistics in medicine* 30, 24 (2011), 2867–2880.

[14] Nicholas C Henderson, Thomas A Louis, Chenguang Wang, and Ravi Varadhan. 2016. Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research. *Health Services and Outcomes Research Methodology* 16, 4 (2016), 213–233.

[15] Kosuke Imai, Marc Ratkovic, and others. 2013. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7, 1 (2013), 443–470.

[16] Tony Jung and KAS Wickrama. 2008. An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass* 2, 1 (2008), 302–317.

[17] Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. 2011. 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.

[18] Wei-Yin Loh, Haoda Fu, Michael Man, Victoria Champion, and Menggang Yu. 2016. Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine* 35, 26 (2016), 4837–4855.

[19] Wei-Yin Loh, Xu He, and Michael Man. 2015. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine* 34, 11 (2015), 1818–1833.

[20] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[21] Bengt Muthén. 2004. Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences. Thousand Oaks, CA: Sage Publications* (2004), 345–68.

[22] Bengt Muthén, C Hendricks Brown, Katherine Masyn, Booil Jo, Siek-Toon Khoo, Chih-Chien Yang, Chen-Pin Wang, Sheppard G Kellam, John B Carlin, and Jason Liao. 2002. General growth mixture modeling for randomized preventive interventions. *Biostatistics* 3, 4 (2002), 459–475.

[23] Austin Nichols and others. 2007. Causal inference with observational data. *Stata Journal* 7, 4 (2007), 507.

[24] Alexander Peysakhovich and Akos Lada. 2016. Combining observational and experimental data to find heterogeneous treatment effects. *arXiv preprint arXiv:1611.02385* (2016).

[25] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* (1983), 41–55.

[26] Tapash Roy and Cathy E Lloyd. 2012. Epidemiology of depression and diabetes: a systematic review. *Journal of affective disorders* 142 (2012), S8–S21.
[27] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688.
[28] Zach Shahn, David Madigan, and others. 2016. Latent Class Mixture Models of Treatment Effect Heterogeneity. *Bayesian Analysis* (2016).
[29] Bonnie Sibbald and Martin Roland. 1998. Understanding controlled trials. Why are randomised controlled trials important? *BMJ: British Medical Journal* 316, 7126 (1998), 201.
[30] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10, Feb (2009), 141–158.
[31] Lan Umek and Blaz Zupan. 2011. Subgroup discovery in data sets with multi-dimensional responses. *Intelligent Data Analysis* 15, 4 (2011), 533–549.
[32] Andrew J Vickers. 2006. Whose data set is it anyway? Sharing raw data from randomized trials. *Trials* 7, 1 (2006), 15.
[33] Andrew J Vickers, Rebecca W Rees, Catherine E Zollman, Rob McCarney, Claire M Smith, Nadia Ellis, Peter Fisher, and Robbert Van Haselen. 2004. Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. *Bmj* 328, 7442 (2004), 744.
[34] Stefan Wager and Susan Athey. 2015. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342* (2015).
[35] John E Ware Jr. 1999. SF-36 health survey. (1999).
[36] Richard J Willke, Zhiyuan Zheng, Prasun Subedi, Rikard Althin, and C Daniel Mullins. 2012. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC medical research methodology* 12, 1 (2012), 185.
[37] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. 2008. Model-based recursive partitioning. *Journal of Computational and Graphical Statistics* 17, 2 (2008), 492–514.

Received May 2017; revised September 2017; accepted October 2017