8-21-2020

Estimation of individual treatment effect via Gaussian mixture model

Juan Wang

HONG KONG BAPTIST UNIVERSITY

Master of Philosophy

THESIS ACCEPTANCE

DATE: August 21, 2020

STUDENT'S NAME: WANG Juan

THESIS TITLE: Estimation of Individual Treatment Effect via Gaussian Mixture Model

This is to certify that the above student's thesis has been examined by the following panel members and has received full approval for acceptance in partial fulfilment of the requirements for the degree of Master of Philosophy.

Chairman:	Dr Leung Ken C F Associate Professor, Department of Chemistry, HKBU (Designated by Dean of Faculty of Science)
Internal Members:	Dr Tong Tiejun Associate Professor, Department of Mathematics, HKBU (Designated by Head of Department of Mathematics)
	Prof Cheng Ming-Yen Professor, Department of Mathematics, HKBU
External Examiner:	Dr LIAN Heng Associate Professor Department of Mathematics City University of Hong Kong

Issued by Graduate School, HKBU

Estimation of Individual Treatment Effect via Gaussian Mixture Model

WANG Juan

A thesis submitted in partial fulfilment of the requirements
for the degree of
Master of Philosophy

Principal Supervisor:
Prof. CHENG Ming-yen(Hong Kong Baptist University)

August 2020

DECLARATION

I hereby declare that this thesis represents my own work which has been done after registration for the degree of MPhil at Hong Kong Baptist University, and has not been previously included in a thesis or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

I have read the University's current research ethics guidelines, and accept responsibility for the conduct of the procedures in accordance with the University's Research Ethics Committee (REC). I have attempted to identify all the risks related to this research that may arise in conducting this research, obtained the relevant ethical and/or safety approval (where applicable), and acknowledged my obligations and the rights of the participants.

Signature: 

Date: August 2020

-

Abstract

In this thesis, we investigate the estimation problem of treatment effect from Bayesian perspective through which one can first obtain the posterior distribution of unobserved potential outcome from observed data, and then obtain the posterior distribution of treatment effect. We mainly consider how to represent a joint distribution of two potential outcomes - one from treated group and another from control group, which can give us an indirect impression of correlation, since the estimation of treatment effect depends on correlation between two potential outcomes. The first part of this thesis illustrates the effectiveness of adapting Gaussian mixture models in solving the treatment effect problem. We apply the mixture models - Gaussian Mixture Regression (GMR) and Gaussian Mixture Linear Regression (GMLR)- as a potentially simple and powerful tool to investigate the joint distribution of two potential outcomes. For GMR, we consider a joint distribution of the covariate and two potential outcomes. For GMLR, we consider a joint distribution of two potential outcomes, which linearly depend on covariate. Through developing an EM algorithm for GMLR, we find that GMR and GMLR are effective in estimating means and variances, but they are not effective in capturing correlation between two potential outcomes. In the second part of this thesis, GMLR is modified to capture unobserved covariance structure (correlation between outcomes) that can be explained by latent variables introduced through making an important model assumption. We propose a much more efficient Pre-Post EM Algorithm to implement our proposed GMLR model with unobserved covariance structure in practice. Simulation studies show that Pre-Post EM Algorithm performs well not only in estimating means and variances, but also in estimating covariance.

Keywords: Potential outcome; Treatment effect; Unobserved Correlation; Gaussian Mixture Model; Gaussian Mixture Linear Model; Gaussian Mixture Linear Model with unobserved covariance structure; EM Algorithm; Pre-Post EM Algortihm.

Acknowledgements

I would like to express my deep gratitude to my supervisor Prof. CHENG Ming-yen for providing me this opportunity and her guidance on my study. I also would like to give my thanks to Prof. Zhu Lixing for his kind help. Many thanks to Dr. Peng Heng, my co-supervisor, for his guidance and encourage on my research. I learned a lot from his lectures.

It has been a great pleasure to study with my classmates and friends, who gave me great encouragement.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Treatment Effect

Treatment Effect is an important research problem that can be found and tackled in a range of applications such as researches on the impacts of drugs on health outcomes, evaluations of the effects of advertising or public policies. There are different evaluations of treatment effect in different occasions. For example, treatment effects of new drugs have been estimated by medical institutions applying "Treatment/Control tests" to assess the benefits of the drugs. Here, treatment effect is presented as the difference of health outcomes between treatment and control groups. Alternatively, researchers may completely randomly assign patients to the treatment and control groups that have normal distributions and same variances in two groups and evaluate the benefits of drugs by using Two-sample t-test. Researchers may also take the perspective of "Potential Outcomes Framework" to evaluate the benefits of drugs. In this regard, treatment effect is denoted as the difference of two potential outcomes, which assumed that each patient would have a particular outcome if he took the treatment, whereas he would have a different outcome if he had not take the assignment.

Regularly, let $\{(\mathbf{x}_i, T_i, y_i)\}_{i=1}^{n}$ represent the data. Here, $\mathbf{x}_i$ is the covariate, $\mathbf{X}$ is the set of individuals and y_i is the observed potential outcome, respectively. $T_i = 0$ represents individual i belongs to control group, and $T_i = 1$ intervention group(or treatment group). $Y_i(0)$ represents the potential outcome when $T_i = 0$ and $Y_i(1)$

represents the potential outcome when $T_i = 1$. If the individual i takes the treatment, we would observe the outcome $y_i = Y_i(1)$, otherwise we would observe the outcome $y_i = Y_i(0)$. Thus, $y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. Average treatment effect(ATE) is defined as the mean difference between average outcomes of two groups

$$\tau = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}\left[\mathbb{E}[Y(1)|\mathbf{x}] - \mathbb{E}[Y(0)|\mathbf{x}]\right]. \tag{1.1}$$

Another measure is Individual treatment effect (ITE), which is defined as mean difference between two potential outcomes conditional on observed covariates for an individual

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0)|\mathbf{x}\right] = \mathbb{E}\left[Y(1)|\mathbf{x}\right] - \mathbb{E}\left[Y(1)|\mathbf{x}\right]. \tag{1.2}$$

ITE reflects the effect of a treatment on each unit which is an individual-level measurement.

Treatment/control tests and two-sample t test are estimating average treatment effect, which is the benefits of treatment in the data. The promise of personalized, precision medicine has never been closer to a reality than it is today, which motivates the enthusiasm for individual treatment effect. Individual treatment effect evaluates the benefits of a specific treatment on each unit, which could provide a guideline for each unit. Therefore we focus on how to estimate individual treatment effect.

Under the potential outcome framework, there are some problems needs to be discussed. The first one is Random assignment to treatment, which requires that units there is no difference between the treatment and control groups. A more general case is that individuals in treatment and control groups are are different, that is, assignment mechanism is not random. Consider the situation:

$$\mathrm{P}(T = 1|\mathbf{x}, Y(0), Y(1)) = \mathrm{P}(T = 1|\mathbf{x}) = \pi(\mathbf{x}), \tag{1.3}$$

treatment assignment is determined by covariates, which causes unbalanced distributions of covariates in two groups. The distribution of covariate in treated group

is

$$f(\mathbf{x} \mid T=1) = \frac{f(\mathbf{x}, T=1)}{\int f(\mathbf{x}, T=1)d\mathbf{x}} = \frac{P(T=1 \mid \mathbf{x})f(\mathbf{x})}{\int P(T=1 \mid \mathbf{x})f(\mathbf{x})d\mathbf{x}} = \frac{\pi(\mathbf{x})f(\mathbf{x})}{\int \pi(\mathbf{x})f(\mathbf{x})d\mathbf{x}} \quad (1.4)$$

The distribution of covariate in control group is

$$f(\mathbf{x} \mid T=0) = \frac{f(\mathbf{x}, T=0)}{\int f(\mathbf{x}, T=0)d\mathbf{x}} = \frac{P(T=0 \mid \mathbf{x})f(\mathbf{x})}{\int P(T=0 \mid \mathbf{x})f(\mathbf{x})d\mathbf{x}} = \frac{(1-\pi(\mathbf{x}))f(\mathbf{x})}{1-\int \pi(\mathbf{x})f(\mathbf{x})d\mathbf{x}} \quad (1.5)$$

This assignment mechanism is very common in medical experiments and policy evaluation: doctors tend to give a limited amount of drugs on patients with better physical features and companies tends to target their advertisements at customers with a larger buying preference(e.g.Rubin [1974], Rubin [1977]). The other problem is unobserved potential outcomes, which means that the potential outcome $Y_i(0)$ is unobserved if $T_i = 1$ otherwise $Y_i(1)$ is unobserved(e.g.Robins et al. [1994]). Unobserved potential outcomes and assignment mechanism 1.3 bring about challenges for estimation of individual treatment effect.

1.2 Related work

Literatures related to individual treatment effect are classified into two parts in a board sense, one is Frequentist or classical inference, which focus on point and interval estimation of treatment effect, such as Propensity weighting, tree-based methods and Domain adaption; the other one is Bayesian paradigm, which offers inference for the distribution of missing data and treatment effect. The meaning of Bayesian is that: one can first obtain the posterior distribution of the unobserved potential outcome based on observed data and then obtain the posterior distribution of $\tau(\mathbf{x})$, which follows Bayesian causal inference.

1.2.1 Propensity weighting

Generally, it is necessary to put some assumptions on data generation such that we could estimate $\tau(\mathbf{x})$ from the observed data $\{(\mathbf{x}_i, T_i, y_i)\}_{i=1}^n$. Two fundamental assumptions in this section are Probabilistic Assignment and Unconfounded Assignment

(Imbens and Rubin [2015]). Probabilistic Assignment requires that the probability of assigning the individual to treatment group, is strictly between zero and one

$$0 < P(T = 1 \mid \mathbf{x}) < 1 \tag{1.6}$$

Unconfounded Assignment assumes that there is conditional independence between treatment assignment T and two potential outcomes.

$$\{Y(0), Y(1)\} \perp T|\mathbf{x}. \tag{1.7}$$

Let $\pi(\mathbf{x}) = \mathbb{E}[T|\mathbf{x}]$ denote the propensity of assigning to the treatment group at $\mathbf{x}$, and then we have

$$\begin{aligned} \mathbb{E}\left[\frac{Ty}{\pi(\mathbf{x})}|\mathbf{x}\right] &= \mathbb{E}[y|T=1,\mathbf{x}], \\ \mathbb{E}\left[\frac{(1-T)y}{1-\pi(\mathbf{x})}|\mathbf{x}\right] &= \mathbb{E}[y|T=0,\mathbf{x}]. \end{aligned} \tag{1.8}$$

the equation 1.8 leads to an unbiased estimator of potential outcomes. Thus Hirano et al. [2003] estimated the probability of assignment, many early literatures focused on estimation of $\pi(\mathbf{x})$ by using boosting and bagging methods, like random forest, or a neural network, or even random forests(e.g. McCaffrey et al. [2004]; Westreich et al. [2010]; Tan [2007]). After the estimation of $\pi(\mathbf{x})$, they estimated $\tau(\mathbf{x})$ by using

$$\hat{\tau}(x) = \frac{\frac{1}{nh^k}\sum_{i=1}^{n}\left(\frac{T_i y_i}{\hat{\pi}(\mathbf{x}_i)} - \frac{(1-T_i)y_i}{1-\hat{\pi}(\mathbf{x}_i)}\right)K\left(\frac{\mathbf{x}_i - x}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^{n}K\left(\frac{\mathbf{x}_i - x}{h}\right)} \tag{1.9}$$

Given individualistic assignment, the combination of probabilistic and unconfounded assignment is considered in Rosenbaum and Rubin [1983], i.e. strongly treatment assignment. An immediate consequence of strongly ignorable treatment assignment (SITA) is that

$$\mathrm{E}\left[y\left(\frac{T}{\pi(\mathbf{x})} - \frac{1-T}{1-\pi(\mathbf{x})}\right)|\mathbf{x}\right] = \tau(\mathbf{x}). \tag{1.10}$$

Another important method induced from probability weighting is doubly robust estimator (e.g. Lee et al. [2017]; Lunceford and Davidian [2004]; Tan [2007]). Let $\mu_j(\mathbf{x})$ denote the expectation of potential outcome, i.e. $\mu_j(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}, T = j]$ for

$j = 0, 1$ and

$$\psi_1(\mathbf{x}) \equiv \frac{Ty}{\pi(\mathbf{x})} - \frac{T - \pi(\mathbf{x})}{\pi(\mathbf{x})}\mu_1(\mathbf{x}),$$
$$\psi_0(\mathbf{x}) \equiv \frac{(1-T)y}{1-\pi(\mathbf{x})} + \frac{T - \pi(\mathbf{x})}{1-\pi(\mathbf{x})}\mu_0(\mathbf{x}),$$
$$\psi(\mathbf{x}) \equiv \psi_1(\mathbf{x}) - \psi_0(\mathbf{x}).$$

$\psi_1(\mathbf{x})$ and $\psi_0(\mathbf{x})$ correspond to inverse probability weighting. ITE is

$$\tau(\mathbf{x}) = \mathbb{E}[\psi(\mathbf{x})|\mathbf{x}]. \tag{1.11}$$

There are different models to estimate $\pi(\mathbf{x})$ and $\mu_j(\mathbf{x}), j = 0, 1$, like the parametric model, the semiparametric model or the nonparametric model(e.g. Abrevaya et al. [2015]Ding and Wang [2011]; Funk et al. [2011]; Guo et al. [2018]; Han [2018]). Some researches consider high dimensional cases, which brings about challenges for estimating model(e.g.). If we got $\hat{\pi}(\mathbf{x})$ and $\hat{\mu}_j(\mathbf{x}), j = 0, 1$

$$\hat{\tau}(x) = \frac{\frac{1}{nh^k}\sum_{i=1}^n \hat{\psi}_1(\mathbf{x}_i) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)} - \frac{\frac{1}{nh^k}\sum_{i=1}^n \hat{\psi}_0(\mathbf{x}_i) K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)}{\frac{1}{nh^k}\sum_{i=1}^n K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)} \tag{1.12}$$

Doubly robust estimate contains two estimate procedures, one is to estimate potential outcome, the other is to estimate propensity score. It performs well as long as either estimation of $\pi(\mathbf{x})$ or $\mu_j(\mathbf{x})$ is set correctly and quickly dominates in estimating treatment effect.

1.2.2 Domain Adaption

Propensity weighting balances covariates' distribution by using inverse Propensity weighting. A closely related approach is domain adaption, which committed to learn a representative distribution such that the difference between itreated and control distributions is the minimum. Shalit et al. [2017] gave a generalization-error bound for ITE based on the balanced representation. Zhao and Heffernan [2017] applied residual counterfactual networks to individual treatment effect and gave a balanced representation.

Assume that there is a function f which relates $\mathbf{X}$ and T to Y such that $f(\mathbf{x}_i, T_i) \approx$

y_i (Zhao and Heffernan [2017]).

$$\hat{\mathrm{ITE}}\left(\mathbf{x}_i\right) = \begin{cases} y_i - f\left(\mathbf{x}_i, 1 - T_i\right), & T_i = 1 \\ f\left(\mathbf{x}_i, 1 - T_i\right) - y_i, & T_i = 0 \end{cases} \tag{1.13}$$

The key idea of Zhao and Heffernan [2017] is to find a representative distribution to minimize distance between treated distribution $p_1 = p(\mathbf{x}|T = 1)$ and control distribution $p_0 = p(\mathbf{x}|T = 0)$ by using an Integral Probability Metric (IPM), which is defined as

$$\mathrm{IPM}_{\mathcal{G}}\left(p_0, p_1\right) := \sup_{g \in \mathcal{G}} \left| \int_S g dp_0 - \int_S g dp_1 \right|. \tag{1.14}$$

Where $\mathbf{x} \in \mathcal{S} \subset \mathcal{R}^d$, $g : \mathcal{S} \to \mathcal{R}$ and $\mathcal{G}$ is a class of real-valued bounded measurable functions on S and $f(\mathbf{x}, t) = h(g, t)$. Related work can be found in Yao et al. [2018], Du et al. [2019], Hartford et al. [2017] and Yoon et al. [2018].

Domain adaption and propensity weighting assumed that distributions of treated and control group are different because of treatment assignment. The goal of propensity weighting is to readjust potential outcomes by using propensity score, while the main contribution of domain adaption is to give a representative distribution which minimizes the distance between treated and control distributions. Both approaches have widespread applications.

1.2.3 Tree-based methods

Assume that the potential outcome takes the form as:

$$Y = m(\mathbf{x}, T) + \varepsilon. \tag{1.15}$$

ε is noise variable.

- Hill [2011] estimated the response planes for treatment group ($E[Y(1)|\mathbf{x}] = m(\mathbf{x}, 1)$) and control group ($E[Y(0)|\mathbf{x}] = m(\mathbf{x}, 0)$) by using Bayesian additive regression trees (BART) and gave the confidence interval for individual treatment effect via Backfitting MCMC algorithm.

- Lu et al. [2018] and Wager and Athey [2018] both estimated the treatment effect by using random forest. The former one mainly compared the random forest with BART and VT(virtual twins) while the latter one focused on heterogeneous individual treatment effect.

The main goal of BART algorithm and Random Forest is to use observed data to fit model, getting rid of unobserved potential outcomes. Individual treatment effect is

$$\begin{aligned} \tau(\mathbf{x}) &= \mathbb{E}[y|T=1,\mathbf{x}] - \mathbb{E}[y|T=0,\mathbf{x}] \\ &= m(\mathbf{x},1) - m(\mathbf{x},0). \end{aligned} \tag{1.16}$$

BART algorithm and Random Forest performs well by fitting many trees to give estimate of $\tau(\mathbf{x})$ (e.g. Hahn et al. [2020]; Alaa and van der Schaar [2018]; Athey et al. [2019] and Tan and Roy [2019]).

1.2.4 Bayesian paradigm

Ding et al. [2018] pointed out that we can't derive an exact correlation between two potential outcomes based on observed data because of only one observed potential outcome. We can vary it from 0 to 1 and obtain the poster distribution for fixed parameter ρ (correlation parameter between two potential outcomes), but the results are sensitive to the parameter ρ. van Klaveren et al. [2015] also emphasised the statistical interactions between potential outcomes. Then Ding et al. [2018] pointed out that the Bayesian framework offers a unified and flexible approach to inferring causal parameters in complex settings but there are no further discussions in his paper.

Take a normal linear example in Ding et al. [2018]

$$\begin{pmatrix} Y(1) \\ Y(0) \end{pmatrix} \Bigg| \mathbf{x}, \theta_{y|\mathbf{x}}, \rho \Bigg) \sim N\left(\begin{pmatrix} \mathbf{x}\boldsymbol{\beta}_1 \\ \mathbf{x}\boldsymbol{\beta}_0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_0 \\ \rho\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix} \right). \tag{1.17}$$

Where $\theta_{Y|X} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_0, \sigma_1^2, \sigma_0^2, \rho)$.

Let $\Delta = Y(1) - Y(0)$, the model implies

$$\Delta|\mathbf{x}, \theta_{y|x}, \rho \sim N\left(\mathbf{x}\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right), \sigma_1^2 + \sigma_0^2 - 2\rho\sigma_1\sigma_0\right). \tag{1.18}$$

We can see from the equation (1.18) that, given a specific $\mathbf{x}$, the distribution of Δ depends on the parameter ρ, ignoring this item will lead to misestimate variance. For a fixed ρ, we can get the conditional distribution of Δ, but the distribution is sensitive to the parameter ρ. The performance of interval estimate for treatment effect with $\rho \neq 0$ is inevitably sensitivity to the setting of this parameter.

Some Monte Carlo simulation algorithms have been applied to estimate confidence interval of treatment effect, like Bayesian additive regression tree (Hill [2011]) and random forest (Wager and Athey [2018]). These approaches estimated the regular surfaces for $Y(0)$ and $Y(1)$ and gave the confidence interval of $\tau(\mathbf{x})$ by backfitting MCMC algorithm. For unit i, confidence interval was obtained by computing the $\alpha/2\%$ and $(1-\alpha/2)\%$ quantiles of the set of response estimates. However, these methods only provide confidence intervals from the perspective of simulation. There is no theoretical guarantee and analysis of correlation between potential outcomes.

Heckman et al. [2014] discussed the correlation between two potential outcomes derived by containing some latent factors. They assumed that there are a set of measurements generated by the latent factors. The model of Heckman et al. [2014] is

$$\begin{aligned} T &= \mathbf{1}\left(T^* > 0\right), \\ T^* &= \mathbf{z}\boldsymbol{\gamma} + U_T, \\ Y(1) &= \mathbf{x}\boldsymbol{\beta}_1 + U_1, \\ Y(0) &= \mathbf{x}\boldsymbol{\beta}_0 + U_0. \end{aligned} \tag{1.19}$$

Where T is a binary treatment decision, depending on some observed characteristics $\mathbf{z}$. $\mathbf{1}(\cdot)$ is an indicator function that takes the value 1 if $T^* > 0$, and equals 0 otherwise. The difference between $\mathbf{x}$and $\mathbf{z}$ is that there are at least one additional covariate should be included in the set $\mathbf{z}$, which is the exclusion assumption.

The covariance structure of the model is expressed as

$$\text{Cov}\begin{pmatrix} U_T \\ U_1 \\ U_0 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{T1}\sigma_1 & \rho_{T0}\sigma_0 \\ \rho_{T1}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{T0}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix}. \tag{1.20}$$

Assume that there are K latent factors $\mathbf{F}$ to describe the unobserved correlation between three items U_T, U_1, U_0:

$$\begin{aligned} T &= \mathbf{1}\left(T^* > 0\right), \\ T^* &= \mathbf{z}\boldsymbol{\gamma} + \mathbf{f}\boldsymbol{\alpha}_T + \varepsilon_T, \\ Y_1 &= \mathbf{x}\boldsymbol{\beta}_1 + \mathbf{f}\boldsymbol{\alpha}_1 + \varepsilon_1, \\ Y_0 &= \mathbf{x}\boldsymbol{\beta}_0 + \mathbf{f}\boldsymbol{\alpha}_0 + \varepsilon_0. \end{aligned} \tag{1.21}$$

$\varepsilon_T, \varepsilon_1, \varepsilon_0$ are independent normal distribution.

Assuming that Q continuous variables $M = (M_1, \ldots, M_Q)'$ can measure the latent factors $\mathbf{f}$, the following equation may be included to the model:

$$M = \mu(\mathbf{x}) + \mathbf{f}\boldsymbol{\Lambda} + \varepsilon_M. \tag{1.22}$$

Where $\mu(\mathbf{x})$ describes the relationship between $\mathbf{x}$ and M, which is a deterministic function. $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_Q)'$ is the factor loading matrix whose dimension is $(Q \times K)$, and ε_M is a Q-dimensional vector of error terms whose components are independent.

For example (Carneiro et al. [2003]), we want to evaluate the effect of college education on the individual earning, we collect education status T, earnings Y, and co-variate X such as experience, age, distance, cognitive ability has a huge impact on earnings but we can't collect the variable, we can use test results to replace it such as arithmetic reasoning, word knowledge, paragraph composition, math knowledge and coding speed as measurements M.

Heckman et al. [2014] estimated this model by Gibbs sampling and analyzed the properties of individual treatment effect. The model in Heckman et al. [2014] describes a joint distribution for two potential outcomes and therefore the distribution of Δ can be obtained.

1.3 Thesis Outline

Bayesian paradigm is a promising direction for potential outcomes since it explains the correlation between two potential outcomes and then we can estimate the distribution of treatment effect. In this thesis, I want to present a distribution of treatment effect such that the expectation and variance of treatment benefit is easy to derive. Meanwhile, some further investigations like quantile and confidence interval can be conducted. From the perspective of Bayesian paradigm, we should firstly estimate the joint distribution between the potential outcome when the individual takes the treatment and the potential outcome when the individual doesn't take the treatment. Then the distribution of treatment benefit is obtained. I mainly consider three kinds of mixture models:

- Cohn et al. [1996] proposed Gaussian Mixture Regression to present a joint distribution over the space $\mathbf{X} \times \mathbf{Y}$. Applying it to treatment effect, we consider a joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{Y} = (Y(0), Y(1))$, then we can derive the distribution of $\mathbf{Y}$ conditional on X. EM algorithm for estimating GMR in Cohn et al. [1996] should be improved to deal with unobserved potential outcomes.

- The limitation of Gaussian Mixture Regression is that we cannot define a model like $y = f(\mathbf{x}) + \epsilon$ (ϵ is noise variable) to explain the effect of $\mathbf{X}$ on $\mathbf{Y}$. Faria and Soromenho [2010] gave the detail of Gaussian Mixture Linear Regression. We extend GMLR to two dimensional case and apply it to estimate treatment effect. EM algorithm for estimating GMLR in two dimensional case is presented in this thesis when there are unobserved potential outcomes.

- Heckman et al. [2014] proposed a model with latent factors to investigate unobserved correlation between two potential outcomes and presented the model Under the assumption of exogenous. Inspired by this, we combined the model in Heckman et al. [2014] with Gaussian Mixture Linear Regression and proposed Gaussian Mixture Linear Regression with unobserved covariance structure. Besides, we drop the exogenous assumption and adopt the unconfounded assignment, that is, given a specific $\mathbf{x}$, the treatment assignment is independent of two potential outcomes. Consequently, a joint distribution between two po-

tential outcomes is discussed. Gibbs sampling, which is adopted in Heckman et al. [2014] is inapplicable because of Mixture model and unobserved potential outcomes. We propose an efficient algorithm-Pre-Post EM Algorithm to realize our proposed Gaussian Mixture Linear regression with unobserved covariance structure in practice.

The rest of the thesis is organized as follows: Chapter 2 reviews Gaussian Mixture Regression and Gaussian Mixture Linear Regression and apply them to individual treatment effect with simulation; Chapter 3 presents Gaussian Mixture Linear Regression with unobserved covariance structure and we also proposed Pre-Post EM Algorithm for this model. The simulation shows that Pre-Post EM Algorithm is effective; Chapter 4 is a short conclusion.

Chapter 2

Gaussian Mixture Regression and Gaussian Mixture Linear Regression

Follow the Bayes theorem, the meaning of Bayesian framework here is that: One can first obtain the posterior distribution of unobserved potential outcome $\mathbf{y}^m$ from observed potential outcome $\mathbf{y}^o$ and covariate $\mathbf{x}$

$$\mathrm{P}\left(\mathbf{y}^{\mathrm{m}}, \theta | \mathbf{y}^{\mathrm{o}}, \mathbf{x}\right) \propto \mathrm{P}(\theta) \prod_{i=1}^{n} \mathrm{P}\left(Y_i(0), Y_i(1), \mathbf{x}_i | \theta\right). \tag{2.1}$$

and then obtain the posterior distribution of Δ, here $\Delta = Y(1) - Y(0)$). Under the Bayesian framework, the discussion of joint distribution is necessary when it comes to in-depth investigation of treatment effect. Gaussian Mixture model is a powerful way to describe flexible distribution, and we mainly consider three kinds of model based on mixture models in this thesis. We will present the first two models in this chapter and postpone the last one to the next chapter.

2.1 Gaussian Mixture Regression

The Gaussian mixture model(GMM) is an extension of a single Gaussian model, using multiple Gaussian distributions to quantify the distribution of a certain variable

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{2.2}$$

Suppose that $\mathbf{z}$ is a vector of K components, in which only one element z_k equals 1, which means that the data belongs to the particular Gaussian distribution. z_k therefore takes value 1 or 0 and $\sum_k z_k = 1$.

The distribution of $\mathbf{z}$ is specified

$$p\left(z_k = 1\right) = \pi_k. \tag{2.3}$$

We can represent this distribution as follows

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \tag{2.4}$$

Given a particular value for z like z_k, the distribution of $\mathbf{x}$ is a Gaussian distribution

$$p\left(\mathbf{x}|z_k = 1\right) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{2.5}$$

The distribution of $\mathbf{x}$ is

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right). \tag{2.6}$$

The log likelihood of n observations is

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}. \tag{2.7}$$

Where, $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ are parameters to be estimated.

EM Algorithm is a widespread way to estimate Gaussian mixture model, through maximizing the likelihood function with respect to the parameters (the means and

covariance and the mixing components), one can get the estimation of parameters. Specifically, in the E step, $p\left(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}\right)$ is the posterior distribution of unobserved components based on the current parameter values $\boldsymbol{\theta}^{old}(\boldsymbol{\theta}=\{\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\})$. In the M step, the expectation of the complete-data log likelihood, denoted as $\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)$, is updated by using the current density estimates

$$\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)=\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})]. \tag{2.8}$$

where

$$\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\left\{\ln\pi_k+\ln\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\right\}. \tag{2.9}$$

In the M step, we maximize the Q function to update parameter estimate θ^{new}

$$\boldsymbol{\theta}^{\text{new}}=\arg\max_{\theta}\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right). \tag{2.10}$$

Specifically, in the E step, take the derivation of equation (2.11) with respect to π_k

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right). \tag{2.11}$$

Thus

$$0=\sum_{i=1}^{n}\frac{\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)}{\sum_j\pi_j\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)}+\lambda. \tag{2.12}$$

$$\pi_k=\frac{N_k}{n}. \tag{2.13}$$

where we defined

$$\begin{aligned}\gamma\left(z_{ik}\right)&=\frac{\pi_k\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)}{\sum_j\pi_j\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)},\\ N_k&=\sum_{i=1}^{n}\gamma\left(z_{ik}\right).\end{aligned} \tag{2.14}$$

In the M step, take the derivation of equation (2.11) respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we obtain

$$0=-\sum_{i=1}^{n}\underbrace{\frac{\pi_k\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)}{\sum_j\pi_j\mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)}}_{\gamma(z_{nk})}\boldsymbol{\Sigma}_k\left(\mathbf{x}_i-\boldsymbol{\mu}_k\right). \tag{2.15}$$

set the derivation to 0

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma\left(z_{ik}\right) \mathbf{x}_i. \tag{2.16}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{i=1}^{n} \gamma\left(z_{ik}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^{\mathrm{T}}. \tag{2.17}$$

EM Algorithm is presented in Algorithm 1.

Algorithm 1 EM Algorithm for Gaussian Mixture Model

1. Initialize $\boldsymbol{\mu}_k$ $\boldsymbol{\Sigma}_k$ and π_k, $k = 1, 2, ...K$.

2. **E** step. Update the responsibilities based on the current parameter values

$$\gamma\left(z_{nk}\right) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)},$$

$$N_k = \sum_{i=1}^{n} \gamma\left(z_{ik}\right).$$

3. **M** step. Update the parameters based on the current responsibilities

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{i=1}^{n} \gamma\left(z_{ik}\right) \mathbf{x}_i, \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{n} \gamma\left(z_{ik}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}, \\
\pi_k^{\text{new}} &= \frac{N_k}{n}.
\end{aligned}$$

4. Record the log likelihood function

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right\}.$$

and then check the convergence of the parameters otherwise return to step 2.

2.1.1 Model assumption

Adjusting Gaussian mixture model to estimate the distribution, Cohn et al. [1996] proposed Gaussian Mixture Regression. Gaussian Mixture Regression present a joint density over the space of $\mathbf{X} \times \mathbf{Y}$.

For the Gaussian distribution k

$$p(\mathbf{x}, \mathbf{y}|z_k = 1) = \frac{1}{\sqrt{(2\pi)^p |\mathbf{\Sigma}_k|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}^+ - \boldsymbol{\mu}_k\right)^T \mathbf{\Sigma}_k^{-1}\left(\mathbf{x}^+ - \boldsymbol{\mu}_k\right)\right]. \tag{2.18}$$

where,

$$\mathbf{x}^+ = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \quad \boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{x,k} \\ \boldsymbol{\mu}_{y,k} \end{bmatrix}, \quad \mathbf{\Sigma}_k = \begin{bmatrix} \mathbf{\Sigma}_{xx,k} & \mathbf{\Sigma}_{xy,k} \\ \mathbf{\Sigma}_{xy,k} & \mathbf{\Sigma}_{yy,k} \end{bmatrix}.$$

Then the distribution of $\mathbf{y}$ conditional on $\mathbf{x}$ is multivariate normal $(\mathbf{y}|\mathbf{x}) \sim N(\overline{\boldsymbol{\mu}}_k, \overline{\mathbf{\Sigma}}_k)$

$$\begin{aligned} \overline{\boldsymbol{\mu}}_k &= \boldsymbol{\mu}_{y,k} + \mathbf{\Sigma}_{xy,k}\mathbf{\Sigma}_{xx,k}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_{x,k}\right), \\ \overline{\mathbf{\Sigma}}_k &= \mathbf{\Sigma}_{yy,k} - \mathbf{\Sigma}_{xy,k}\mathbf{\Sigma}_{xx,k}^{-1}\mathbf{\Sigma}_{xy,k}. \end{aligned} \tag{2.19}$$

Adjusting Gaussian Mixture Regression and replacing the component of $\mathbf{y}$ in the equation (2.18) by

$$\mathbf{y} = \begin{bmatrix} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{bmatrix}, \quad \mathbf{x}^+ = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \quad \boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{x,k} \\ \boldsymbol{\mu}_{y,k} \end{bmatrix}, \quad \mathbf{\Sigma}_k = \begin{bmatrix} \mathbf{\Sigma}_{xx,k} & \mathbf{\Sigma}_{xy,k} \\ \mathbf{\Sigma}_{xy,k} & \mathbf{\Sigma}_{yy,k} \end{bmatrix}.$$

Thus a joint distribution for the space $\mathbf{X} \times \mathbf{Y}(0) \times \mathbf{Y}(1)$ is

$$p(\mathbf{x}^+) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}^+|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}^+|\boldsymbol{\mu}_k, \mathbf{\Sigma}_k\right). \tag{2.20}$$

2.1.2 EM Algorithm

EM Algorithm has two important applications: one is to learn mixture models and another important application is to learn model from data sets with missing values (e.g. Little and Rubin [2019] ; Dempster et al. [1977]). Algorithm 1 is the first application, which estimates Gaussian mixture model by setting the derivation of $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \mathbf{\Sigma}) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$ with respect to $\boldsymbol{\mu}_k$ and $\sum_k$ to zero and updating parameters, the latter application can be find in a lot of statistics literatures to learn non-mixture densities. In this thesis we combine two applications of EM to estimate joint distribution of potential outcomes since there are hidden latent variables and unobserved potetnial outcomes. Some adjustments of the EM algorithm should be

made to estimate GMR. An potential solution is provided by Ghahramani and Jordan [1994], who updated responsibilities at E-step based on the observed dimensions and imputed missing values at M-step by corresponding first and second moments. That is,

- In the **E** step, update following equation based on observed data

$$\gamma\left(z_{ik}\right)=\frac{\pi_k\mathcal{N}\left(\mathbf{x}_i^{+o}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K}\pi_j\mathcal{N}\left(\mathbf{x}_i^{+o}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)}. \tag{2.21}$$

$$\pi_k=\frac{N_k}{n}. \tag{2.22}$$

  where o represents observed data and we have defined

$$N_k=\sum_{i=1}^{n}\gamma\left(z_{ik}\right). \tag{2.23}$$

  which means that

$$if \quad T_i=0, \quad \mathbf{x}_i^{+o}=(\mathbf{x}_i,Y_i(0)), \tag{2.24}$$

$$else, \quad \mathbf{x}_i^{+o}=(\mathbf{x}_i,Y_i(1)). \tag{2.25}$$

- In the **M** step, update following items

$$\text{a) } \mu_k^{new}=\frac{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\mathbf{x}_i^{+}}{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)}, \tag{2.26}$$

$$\text{b) } \Sigma_k^{new}=\frac{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\left(\mathbf{x}_i^{+}-\mu_k^{new}\right)\left(\mathbf{x}_i^{+}-\mu_k^{new}\right)^T}{\sum_{i=1}^{N}\gamma\left(z_{ik}\right)}. \tag{2.27}$$

  If $T_i=0$, impute missing values $Y_i(1)$ in the M step by

$$\begin{aligned}E\left[z_{ik}Y_i(1)|\mathbf{x}_i,Y_i(0),\theta_k\right]&=\lambda\left(z_{ik}\right)E\left[Y_i(1)|z_{ik}=1,\mathbf{x}_i,Y_i(0),\theta_k\right]\\&=\lambda\left(z_{ik}\right)\left(\boldsymbol{\mu}_k^{\mathrm{m}}+\Sigma_k^{\mathrm{mo}}\Sigma_k^{\mathrm{oo}^{-1}}\left((\mathbf{x}_i^{+o})^T-(\boldsymbol{\mu}_k^{\mathrm{o}})^T\right)\right).\end{aligned} \tag{2.28}$$

  where m,o represent missing and observed items. $\lambda\left(z_{ik}\right)=E\left[z_{ik}|\mathbf{x}_i,Y_i(0),\theta_k\right]$.

Define $\hat{Y}_{ik}(1) \equiv E\left[Y_i(1)|z_{ik}=1,\mathbf{x}_i,Y_i(0),\theta_k\right]$,

$$E\left[z_{ik}Y_{ik}^2(1)|Y_i(1),\theta_k\right] = \lambda(z_{ik})\left(\Sigma_k^{mm} - \Sigma_k^{mo}\Sigma_k^{oo^{-1}}\Sigma_k^{moT} + \hat{Y}_{ik}^2(1)\right). \tag{2.29}$$

Similar for those missing values $Y_i(0)$.

EM Algorithm for Gaussian Mixture Regression within treatment effect is presented in Algorithm 2.

Algorithm 2 EM Algorithm for GMS within treatment effect

1. **Initialize** $\boldsymbol{\pi}_k$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$

2. **E-step**. Update the responsibilities based on the current parameter values

$$\gamma\left(z_{ik}\right) = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_i^{+o}|\boldsymbol{\mu}_k^o,\boldsymbol{\Sigma}_k^{oo}\right)}{\sum_{j=1}^K \pi_j \mathcal{N}\left(\mathbf{x}_i^{+o}|\boldsymbol{\mu}_j^o,\boldsymbol{\Sigma}_j^{oo}\right)}.$$

3.**M-step**. Update the parameters using the current responsibilities

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k}\sum_{i=1}^n \gamma\left(z_{ik}\right)\mathbf{x}_i^+, \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k}\sum_{i=1}^n \gamma\left(z_{ik}\right)\left(\mathbf{x}_i^+ - \boldsymbol{\mu}_k^{\text{new}}\right)\left(\mathbf{x}_i^+ - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}}, \\
\pi_k^{\text{new}} &= \frac{N_k}{n}.
\end{aligned}$$

where

$$N_k = \sum_{i=1}^n \gamma\left(z_{ik}\right).$$

If $T_i = 0$, impute missing values $Y_i(1)$ in the M step by

$$\begin{aligned}
E\left[z_{ik}Y_i(1)|\mathbf{x}_i,Y_i(0),\theta_k\right] &= \lambda\left(z_{ik}\right)E\left[Y_i(1)|z_{ik}=1,\mathbf{x}_i,Y_i(0),\theta_k\right] \\
&= \lambda\left(z_{ik}\right)\left(\boldsymbol{\mu}_k^{\mathrm{m}} + \Sigma_k^{\mathrm{mo}}\Sigma_k^{\mathrm{oo}^{-1}}\left((\mathbf{x}_i^{+o})^T - (\boldsymbol{\mu}_k^{\mathrm{o}})^T\right)\right).
\end{aligned}$$

$$E\left[z_{ik}Y_{ik}^2(1)|y_i(1),\theta_k\right] = \lambda(z_{ik})\left(\Sigma_k^{mm} - \Sigma_k^{mo}\Sigma_k^{oo^{-1}}\Sigma_k^{moT} + \hat{Y}_{ik}^2(1)\right).$$

4. Record the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{i=1}^n \ln\left\{\sum_{k=1}^K \pi_k \mathcal{N}\left(\mathbf{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\right\}.$$

and check the convergence of the sum of parameters otherwise return to step 2.

A "complete-data" log likelihood function of Gaussian Mixture Regression is

$$\ln p(\mathbf{X}^+, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln P\left(\mathbf{x}_i^+|\mathbf{z}_i, \theta\right) P\left(\mathbf{z}_i|\theta\right) \tag{2.30}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln \mathcal{N}\left(\mathbf{x}_i^+|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln \pi_k. \tag{2.31}$$

and Q function takes the expectation to unknown component z and missing values:

$$\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = \mathbb{E}_{\mathbf{Z},\mathbf{Y}^m}[\ln p(\mathbf{X}^+, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \tag{2.32}$$

$$= \sum_{\mathbf{Z}} p\left(\mathbf{Z}|\mathbf{X}^+, \boldsymbol{\theta}^{\text{old}}\right) \mathbb{E}_{\mathbf{Y}^m}[\ln p(\mathbf{X}^+, \mathbf{Z}|\boldsymbol{\theta})] \tag{2.33}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \lambda\left(z_{ik}\right) \left[\ln \pi_k - \frac{p+2}{2}\ln 2\pi - \frac{\ln|\boldsymbol{\Sigma}_k|}{2} - \frac{\left(\mathbf{x}_i^\circ - \mu_k^\circ\right)^T \boldsymbol{\Sigma}_k^{-1,\text{oo}} \left(\mathbf{x}_i^\circ - \mu_k^\circ\right)}{2}\right. \tag{2.34}$$

$$\left. + \mathbb{E}_{\mathbf{y}_i^m}\left[-\left(\mathbf{x}_i^\circ - \mu_k^\circ\right)^T \boldsymbol{\Sigma}_k^{-1,\text{om}} \left(\mathbf{y}_i^{\text{m}} - \mu_k^{\text{m}}\right) - \frac{\left(\mathbf{y}_i^{\text{m}} - \mu_k^{\text{m}}\right)^T \boldsymbol{\Sigma}_k^{-1,\text{mm}} \left(\mathbf{y}_i^{\text{m}} - \mu_k^{\text{m}}\right)}{2}\right]\right]. \tag{2.35}$$

Since there are some missing values $y(0)$ and $y(1)$, we estimate $\lambda(z_{ik})$ and π_k based on observed data. In the **M** step, we take the expectation to unknown component z and missing values and then take derivation of Q function with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For those unobserved values $y(0)$ and $y(1)$, we use the equation 2.28 and 2.29 to replace those expectations of missing values in the Q function.

Therefore, Algorithm 2 can be used in the case where there are unknown components and unobserved potential outcomes in Gaussian Mixture Regression.

2.1.3 Simulation

In this section, we do some simulations.

- *Initial conditions*. In our simulation study, we generate random values from the set $\{1, 2\}$ as initial values of $\boldsymbol{\mu}$ and generate orthogonal matrices U and diagonal matrix D=$diag[1, 3, 1, 3]$ and $\boldsymbol{\Sigma} = U \times D \times U'$.

- *Data set*. We generate a datum $(\mathbf{x}_i, T_i, Y_i(0), Y_i(1))$ for each i by the following

steps.

i). z_i was generated from a binary distribution with $P(z_1 = 1) = 0.4$ and $P(z_2 = 1) = 0.6$.

ii). $(\mathbf{x}_i, Y_i(0), Y_i(1))$ was generated from multivariate normal distribution.

iii). T_i was generated from a logistic distribution with $P(T = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\alpha})}{1+\exp(\mathbf{x}^T\boldsymbol{\alpha})}$, $\boldsymbol{\alpha} = (1, -2)$.

iv). Reorganize datum. delete $Y_i(0)$ for those observations $T_i = 1$ and delete $Y_i(1)$ for those observations $T_i = 0$.

- *Number of samples*. The size of the sample was 1000.

- *Stopping rules*. The stopping criterion for EM Algorithm we consider was that the absolute sum of parameters is less than 10^{-2}. We also investigate how the Q function reacts to the increasing number of iterations.

The simulation process contains 2 steps as follows:

(1). Create a data set whose size is n.

(2). Perform EM algorithm 2 to estimate the Gaussian mixture model. Save Q function and the estimated parameters $\hat{\theta} = \left(\hat{\pi}_1, \hat{\pi}_2, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2\right)$

Simulation results are displaced in table 2.1 and 2.2.

Table 2.1: True means & Estimate for a two-component GMR.

	True values		Estimate	
	π	$\boldsymbol{\mu}$	$\hat{\pi}$	$\hat{\boldsymbol{\mu}}$
parallel	0.4	$(0, 0, 2, 4)$	0.3880	$(-0.1458, -0.0630, 2.0604, 4.0118)$
concurrent	0.6	$(0, 0, 3, -4)$	0.6120	$(0.1425, -0.0138, 2.7797, -4.0417)$

Conclusion

- Gaussian Mixture Regression performs well in estimating expectation of two potential outcomes but it fails to capture correlation between two potential outcomes.

Table 2.2: True covariance matrix & Estimate for a two-component GMR.

	True values $\boldsymbol{\Sigma}$	Estimate $\hat{\boldsymbol{\Sigma}}$
parallel	$\begin{bmatrix} 2.5095 & -0.6148 & -0.4800 & -0.4312 \\ -0.6148 & 2.9363 & -1.1977 & -0.0233 \\ -0.4800 & -1.1977 & 2.5834 & -0.1783 \\ -0.4312 & -0.0233 & -0.1783 & 1.9708 \end{bmatrix}$	$\begin{bmatrix} 2.7153 & -0.8002 & -1.2190 & -0.5065 \\ -0.8002 & 3.0615 & 0.1278 & -0.0513 \\ -1.2190 & 0.1278 & 3.5454 & 0.7572 \\ -0.5065 & -0.0513 & 0.7572 & 2.3952 \end{bmatrix}$
concurrent	$\begin{bmatrix} 2.4427 & 0.4064 & 0.1466 & 0.6836 \\ 0.4064 & 2.0969 & 0.1475 & 1.1183 \\ 0.1466 & 0.1475 & 3.2336 & 0.4294 \\ 0.6836 & 1.1183 & 0.4294 & 2.2268 \end{bmatrix}$	$\begin{bmatrix} 2.1280 & 0.4021 & 0.2534 & 0.6939 \\ 0.4021 & 1.9000 & -0.9010 & 0.9892 \\ 0.2534 & -0.9010 & 3.4972 & -0.5239 \\ 0.6939 & 0.9892 & -0.5239 & 2.0051 \end{bmatrix}$

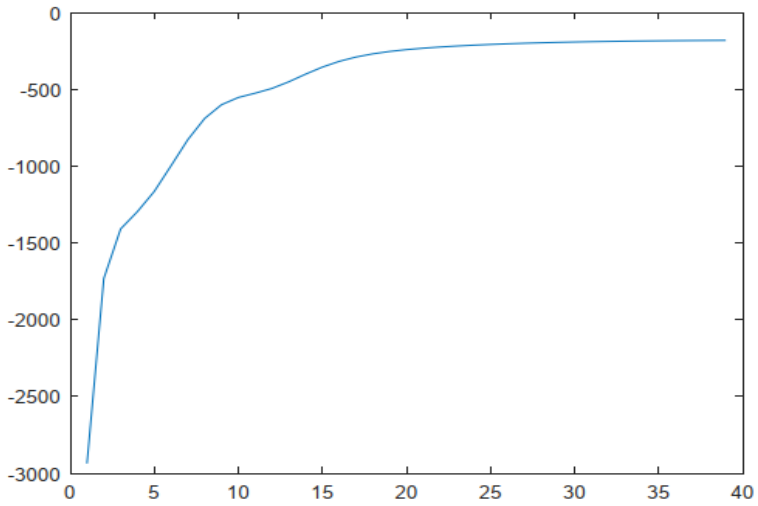

Figure 2.1: Q function for Gaussian Mixture Regression

- The general EM algorithm has the property that each cycle of the algorithm will increase the incomplete-data log likelihood until it is already at a local maximum). The Q function of GMR increases as the iteration proceeding, which proves the effectiveness of EM algorithm 2. That's because we impute missing values in the M step in order to get effective estimation.

2.2 Gaussian Mixture Linear Regression

Another adaptive model is to estimate the joint distribution of potential outcomes conditional on $\mathbf{x}$ directly via Gaussian Mixture Linear Regression(GMLR). Faria and Soromenho [2010] introduced the detail of Gaussian Mixture Linear Regression in one dimensional case. In this thesis, I expand it to two dimensional case from the

perspective of linear regression.

2.2.1 Model assumption

GMLR is presented as follows (Faria and Soromenho [2010])

$$y = \begin{cases} \mathbf{x}\beta_1 + \epsilon_1 & \text{with probability } \pi_1, \\ \mathbf{x}\beta_2 + \epsilon_2 & \text{with probability } \pi_2, \\ \vdots & \vdots \\ \mathbf{x}\beta_K + \epsilon_K & \text{with probability } \pi_K. \end{cases} \tag{2.36}$$

which can be presented as

$$p(y|z_k = 1) = \frac{1}{\sqrt{2\pi\,|\sigma_k^2|}} \exp\left[-\frac{(y - \mathbf{x}\boldsymbol{\beta}_k)^2}{2\sigma_k^2}\right]. \tag{2.37}$$

The distribution of $\mathbf{z}$ in GMLR is the same as the equation 2.3

$$p\left(z_k = 1\right) = \pi_k. \tag{2.38}$$

In this chapter, I extend Gaussian Mixture Linear Regression to two dimensional case. Specifically, for the Gaussian distribution k,

$$p(\mathbf{y}|z_k = 1) = \frac{1}{\sqrt{(2\pi)^2\,|\Sigma_k|}} \exp\left[-\frac{1}{2}\left(\mathbf{y} - A\boldsymbol{\beta}_k\right)^T \Sigma_k^{-1}\left(\mathbf{y} - A\boldsymbol{\beta}_k\right)\right]. \tag{2.39}$$

where,

$$\mathbf{y} = \begin{bmatrix} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{x} & \\ & \mathbf{x} \end{bmatrix}, \quad \boldsymbol{\beta}_k = \begin{bmatrix} \boldsymbol{\beta}_{k,0} \\ \boldsymbol{\beta}_{k,1} \end{bmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\sigma}_{k1}^2 & \rho_k \boldsymbol{\sigma}_{k1}\boldsymbol{\sigma}_{k0} \\ \rho_k \boldsymbol{\sigma}_{k1}\boldsymbol{\sigma}_{k0} & \boldsymbol{\sigma}_{k0}^2 \end{bmatrix}.$$

The log-likelihood function of n observations is

$$\ln p(\mathbf{Y}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}_i|\mathbf{A}_i\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k\right)\right\}. \tag{2.40}$$

Applying Gaussian Mixture Linear Regression into treatment effect, we present

the joint distribution of treated potential outcome and control potential outcome and the element on the off diagonal of the covariance matrix explains the correlation between two potential outcomes.

2.2.2 EM Algorithm

Faria and Soromenho [2010] discussed how to estimate this mixture regression for one dimensional response variable. In this section, I extend Gaussian Mixture Regression to two dimensional case and apply it to treatment effect, the obstacle is unobserved potential outcomes. Following the idea of Ghahramani and Jordan [1994], this section gives the detail of the algorithm, which can be find in Algorithm 3.

In the E step, take the derivation of $\ln p(\mathbf{Y}|\mathbf{Z},\mathbf{X},\boldsymbol{\theta})+\lambda(\sum_{k=1}^{K}\pi_k-1)$ with respect to π_k, we obtain

$$\pi_k^{new}=\frac{\sum_{i=1}^{n}\lambda(z_{ik})}{n}\quad(k=1,\ldots,K). \tag{2.41}$$

Where

$$\gamma\left(z_{ik}\right)=\frac{\pi_k\mathcal{N}\left(y_i^o|\mathbf{x}_i\boldsymbol{\beta}_{k,o},\sigma_{k,o}^2\right)}{\sum_{j=1}^{K}\pi_j\mathcal{N}\left(y_i^o|\mathbf{x}_i\boldsymbol{\beta}_{j,o},\sigma_{j,o}^2\right)}. \tag{2.42}$$

use the same method to deal with missing values:

If $T_n=1$

$$\gamma\left(z_{ik}\right)=\frac{\pi_k\mathcal{N}\left(Y_i(1)|\mathbf{x}_i\boldsymbol{\beta}_{k1},\sigma_{k1}^2\right)}{\sum_{j=1}^{K}\pi_j\mathcal{N}\left(Y_i(1)|\mathbf{x}_n\boldsymbol{\beta}_{j1},\sigma_{j1}^2\right)}, \tag{2.43}$$

If $T_n=0$

$$\gamma\left(z_{ik}\right)=\frac{\pi_k\mathcal{N}\left(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{k0},\sigma_{k0}^2\right)}{\sum_{j=1}^{K}\pi_j\mathcal{N}\left(Y_i(0)|\mathbf{x}_n\boldsymbol{\beta}_{j0},\sigma_{j0}^2\right)}. \tag{2.44}$$

In the M step, take the derivation of $\ln p(\mathbf{Y}|\mathbf{Z},\mathbf{X},\boldsymbol{\theta})+\lambda(\sum_{k=1}^{K}\pi_k-1)$ with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$,

$$\boldsymbol{\beta}_k^{\text{new}}=\left[\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\mathbf{A}_i\Sigma_k^{-1}\mathbf{A}_i^T\right]^{-1}\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\mathbf{A}_i^T\Sigma_k^{-1}\mathbf{y}_i, \tag{2.45}$$

$$\Sigma_k{}^{\text{new}}=\frac{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\left(y_i-\mathbf{A}_i\boldsymbol{\beta}_k^{\text{new}}\right)\left(y_i-\mathbf{A}_i\boldsymbol{\beta}_k^{\text{new}}\right)^T}{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)}. \tag{2.46}$$

If $T_i=0$, impute missing values $Y_i(1)$ in M step by

Algorithm 3 EM Algorithm for GMLR

1. **Initialize** $\boldsymbol{\pi}_k$, β_k, $\boldsymbol{\Sigma}_k$

2. **E-step**. Update the responsibilities based on the current parameter values
If $T_n = 1$

$$\gamma\left(z_{ik}\right) = \frac{\pi_k \mathcal{N}\left(Y_i(1)|\mathbf{x}_i\boldsymbol{\beta}_{k1}, \sigma_{k1}^2\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(Y_i(1)|\mathbf{x}_i\boldsymbol{\beta}_{j1}, \sigma_{j1}^2\right)}.$$

Otherwise $T_n = 0$

$$\gamma\left(z_{ik}\right) = \frac{\pi_k \mathcal{N}\left(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{k0}, \sigma_{k0}^2\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{j0}, \sigma_{j0}^2\right)}.$$

3. **M-step**. Update the parameters based on the current responsibilities

$$\boldsymbol{\beta}_k^{\text{new}} = \left[\sum_{i=1}^{n} \gamma\left(z_{ik}\right) \mathbf{A}_i \Sigma_k^{-1} \mathbf{A}_i^T\right]^{-1} \sum_{i=1}^{n} \gamma\left(z_{ik}\right) \mathbf{A}_i^T \Sigma_k^{-1} \mathbf{y}_i,$$

$$\Sigma_k{}^{\text{new}} = \frac{\sum_{i=1}^{n} \gamma\left(z_{ik}\right)\left(y_i - \mathbf{A}_i\boldsymbol{\beta}_k^{\text{new}}\right)\left(y_i - \mathbf{A}_i\boldsymbol{\beta}_k^{\text{new}}\right)^T}{\sum_{i=1}^{n} \gamma\left(z_{ik}\right)}.$$

where

$$N_k = \sum_{i=1}^{n} \gamma\left(z_{nk}\right).$$

If $T_i = 0$, impute missing values $Y_i(1)$ in M step by

$$\begin{aligned} E\left[\lambda\left(z_{ik}\right) Y_i(1)|\mathbf{A}_i, Y_i(0), \theta_k\right] &= \lambda\left(z_{ik}\right) E\left[Y_i(1)|z_{ik} = 1, \mathbf{x}_i, Y_i(0), \theta_k\right] \\ &= \lambda\left(z_{ik}\right)\left(\mathbf{x}_i\boldsymbol{\beta}_{k1} + \Sigma_k^{\text{mo}}\Sigma_k^{\text{oo}^{-1}}\left(Y_i(0) - \mathbf{x}_i\boldsymbol{\beta}_{k0}\right)\right). \end{aligned}$$

$$E\left[z_{ik}Y_i^2(1)|Y_i(0), \theta_k\right] = \lambda\left(z_{ik}\right)\left(\Sigma_k^{mm} - \Sigma_k^{mo}\Sigma_k^{oo^{-1}}\Sigma_k^{moT} + \hat{Y}_{ik}^2(0)\right).$$

Similar for those missing values $Y_i(0)$.

4. Record the log likelihood

$$\ln p(\mathbf{Y}|\mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}_i|\mathbf{A}_i\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k\right)\right\}.$$

and check the convergence of the sum of the parameters otherwise return to step 2.

$$E\left[\lambda\left(z_{ik}\right)Y_i(1)|\mathbf{A}_i,Y_i(0),\theta_k\right]=\lambda\left(z_{ik}\right)E\left[Y_i(1)|z_{ik}=1,\mathbf{x}_i,Y_i(0),\theta_k\right] \tag{2.47}$$

$$=\lambda\left(z_{ik}\right)\left(\mathbf{x}_i\boldsymbol{\beta}_{k1}+\Sigma_k^{\mathrm{mo}}\Sigma_k^{\mathrm{oo}^{-1}}\left(Y_i(0)-\mathbf{x}_i\boldsymbol{\beta}_{k0}\right)\right). \tag{2.48}$$

where m, o represent missing and observed items. $\lambda\left(z_{ik}\right)=E\left[z_{ik}|\mathbf{x}_i,Y_i(0),\theta_k\right]$. Define $\hat{Y}_{ik}(1)\equiv E\left[Y_i(1)|z_{nk}=1,\mathbf{x}_i,Y_i(0),\theta_k\right]$,

$$E\left[z_{ik}Y_i^2(1)|Y_i(0),\theta_k\right]=\lambda\left(z_{ik}\right)\left(\Sigma_k^{mm}-\Sigma_k^{mo}\Sigma_k^{oo^{-1}}\Sigma_k^{moT}+\hat{Y}_{ik}^2(0)\right). \tag{2.49}$$

Similar for those missing values $Y_i(0)$.

The log likelihood function of Gaussian Mixture Regression is

$$\begin{aligned}\ln p(\mathbf{Y},\mathbf{Z}|\mathbf{A},\boldsymbol{\beta},\boldsymbol{\Sigma},\boldsymbol{\pi})&=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\ln P\left(\mathbf{Y}_i|\mathbf{x},\mathbf{z}_i,\theta\right)P\left(\mathbf{z}_i|\theta\right)\\&=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\ln\mathcal{N}\left(\mathbf{y}_i|\mathbf{A}_i\boldsymbol{\beta}_k,\boldsymbol{\Sigma}_k\right)+\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\ln P\left(\mathbf{z}_i|\theta\right).\end{aligned} \tag{2.50}$$

and Q function takes the expectation to unknown component z and missing values:

$$\begin{aligned}\mathcal{Q}\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}\right)&=\sum_{\mathbf{Z}}p\left(\mathbf{Z}|\mathbf{Y},\mathbf{A},\boldsymbol{\theta}^{\mathrm{old}}\right)\mathbb{E}_{\mathbf{Y}^{\mathrm{m}}}\left[\ln p(\mathbf{Y},\mathbf{Z}|\mathbf{A},\boldsymbol{\theta})\right]\\&=\sum_{i=1}^{n}\sum_{k=1}^{K}\lambda\left(z_{ik}\right)\left[\ln\pi_k-\ln 2\pi-\frac{\ln|\boldsymbol{\Sigma}_k|-\left(\mathbf{y}_i^{\circ}-\mathbf{x}_i\beta_k^{\circ}\right)^T\boldsymbol{\Sigma}_k^{-1,\mathrm{oo}}\left(\mathbf{y}_i^{\circ}-\mathbf{x}_i\beta_k^{\circ}\right)}{2}\right.\\&\left.\mathbb{E}_{\mathbf{Y}^{\mathrm{m}}}\left[-\left(\mathbf{y}_i^{\circ}-\mathbf{x}_i\beta_k^{\circ}\right)^T\boldsymbol{\Sigma}_k^{-1,\mathrm{om}}\left(\mathbf{y}_i^{\mathrm{m}}-\mathbf{x}_i\beta_k^{\mathrm{m}}\right)-\frac{\left(\mathbf{y}_i^{\mathrm{m}}-\mathbf{x}_i\beta_k^{\mathrm{m}}\right)^T\boldsymbol{\Sigma}_k^{-1,\mathrm{mm}}\left(\mathbf{y}_i^{\mathrm{m}}-\mathbf{x}_i\beta_k^{\mathrm{m}}\right)}{2}\right]\right].\end{aligned} \tag{2.51}$$

Since there are some missing values $Y(0)$ and $Y(1)$, we estimate $\lambda(z_{ik})$ and π_k based on observed data in the E step. In the **M** step, we take the expectation to unknown component z and missing potential outcomes and then take derivation of Q function with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For those unobserved values $Y(0)$ and $Y(1)$, we use the equation 2.48 and 2.49 to replace those expectations of missing values in the Q function. Adaptive EM Algorithm is used to estimate GMLR when there are hidden latent variable and unobserved potential outcomes in Gaussian Mixture Linear Model.

2.2.3 Simulation

In this section, we do some simulations.

- *Initial conditions.* In our simulation study, we generate random values from the set $\{1, 2\}$ as initial values of $\boldsymbol{\beta}$ and generate orthogonal matrices U and diagonal matrix D=[1, 3] and $\boldsymbol{\Sigma} = U \times D \times U'$.

- *Data set.* We generate each datum $(\mathbf{x}_i, T_i, Y_i(0), Y_i(1))$ as follows.

  i). z_i was generated from a binary distribution with $P(z_1 = 1) = 0.4$ and $P(z_2 = 1) = 0.6$.

  ii). $\mathbf{x}_i$ was generated from multivariate normal distribution.

  iii). T_i was generated from a logistic distribution with $P(T = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\alpha})}{1+\exp(\mathbf{x}^T\boldsymbol{\alpha})}$, $\boldsymbol{\alpha} = (1, -2)$.

  iv). $\boldsymbol{\varepsilon}_i|z_{ik} = 1$ was generated from multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_k)$;

  v). $\mathbf{y}_i = \mathbf{A}_i\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_i$.

- *Number of samples.* The size of the sample was 1000.

- *Stopping rules.* A strict stopping criterion for EM Algorithm is that the absolute sum of parameters is less than 10^{-2}. We investigate how the Q function reacts to the increasing number of iterations.

The simulation process contains 2 steps:

(1). Create a data set whose size is n.

(2). Perform EM algorithm 3 to estimate the Gaussian Mixture Linear Regression. Save Q function and the estimated parameters $\hat{\theta} = \left(\hat{\pi}_1, \hat{\pi}_2, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2\right)$

Simulation results are displaced in table 2.3-2.6.

Conclusion

(1). Gaussian Mixture Linear Regression performs well in estimating expectation and variance of two potential outcomes. Estimations of distribution of $Y(0)$ and

Table 2.3: True parameter values with a two-component GMLR.

Configuration	Control		Treat	
	(β_1, β_2)	σ^2	(β_1, β_2)	σ^2
parallel	$(0, 2)$	2	$(2, 4)$	1
concurrent	$(4, 1)$	1	$(1, -1)$	3

Table 2.4: Estimate results with a two-component GMLR.

Configuration	Control		Treat	
	$(\hat{\beta}_1, \hat{\beta}_2)$	$\hat{\sigma}^2$	$(\hat{\beta}_1, \hat{\beta}_2)$	$\hat{\sigma}^2$
parallel	$(-0.0014, 1.9160)$	2.0060	$(2.1700, 4.340)$	0.9366
concurrent	$(3.9390, 1.0640)$	1.0320	$(1.2700, -1.1870)$	2.8000

Table 2.5: Probability Configuration with a two-component GMLR.

Configuration	control	treatment
parallel	0.4	0.7
concurrent	0.6	0.3

Table 2.6: Estimate for Probability Configuration with a two-component GMLR.

Configuration	control	treatment
parallel	0.4068	0.6782
concurrent	0.5932	0.3218

$Y(1)$ are pretty good, which means that we get a good estimation of expectation of $\tau(\mathbf{x})$. It fails to give an accurate estimation of variance of $\tau(\mathbf{x})$.

(2). The Q function increases as the iteration proceed, which proves the effectiveness of the algorithm. That's because we impute missing values in the M step.

(3). The limitation of Gaussian mixture regression and Gaussian mixture linear regression is that they fail to capture correlation between potential outcomes.We can explain this by using Q function, we always find the optimal solution in every step, but in M step we use the conditional distribution of missing values based on observed data to impute unobserved potential outcomes. Thus the correlation between $Y(0)$ and $Y(1)$ is static and it is the initial correlation. From these two examples, what we conclude is that Gaussian Mixture Model preforms well in estimating means and variances, last but most important, estimation of correlation based on observed data is infeasible. Additional information describing correlation should be included in model.

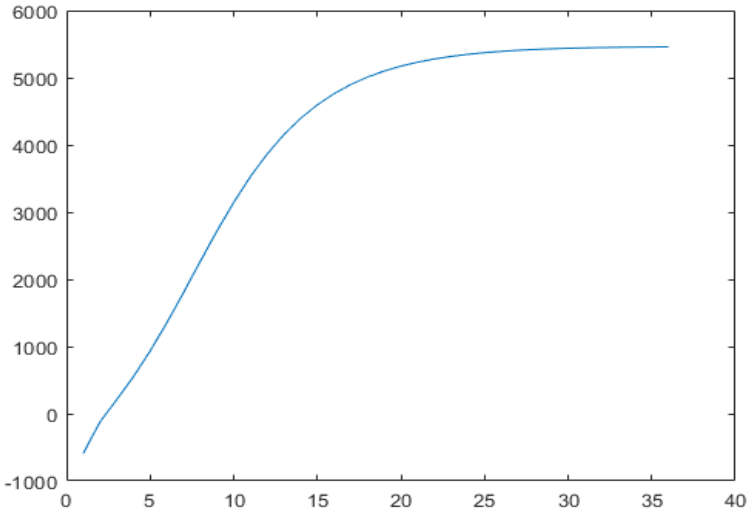

Figure 2.2: Q function for Gaussian Mixture Linear Model

Chapter 3

Gaussian Mixture Linear Regression with unobservable covariance structure

Estimation of individual treatment effect is the foundation of this thesis. The goal of this chapter is to present a joint distribution for two potential outcomes such that we can not only derive the expectation of benefit of treatment effect, but also its variance. Therefore we can get the distribution of treatment effect like some important quantiles and confidence interval, which will help us gain a better understanding of treatment effect. From Bayesian perspective, the joint distribution of two potential outcomes is necessary to derive the distribution of benefit of treatment effect. The first two models show their advantages in estimating means and variances but they also show their disadvantages in estimating covariance, i.e. correlation between two potential outcomes due to unobserved potential outcomes. What we can conclude is that additional information should be included into the model structure so the the correlation can be captured by the model.

Heckman et al. [2014] made inference for treatment effect from Bayesian perspective, in which they assumed that latent variables explain unobserved correlation

between potential outcomes. The model assumed in Heckman et al. [2014] is:

$$\begin{aligned} T &= 1\left(T^* > 0\right), \\ T^* &= \mathbf{z}\gamma + U_T, \\ Y(1) &= \mathbf{x}\beta_1 + U_1, \\ Y(0) &= \mathbf{x}\beta_0 + U_0. \end{aligned} \tag{3.1}$$

where $1(\cdot)$ takes the value 1 if $T^* > 0$ or 0 otherwise, which means that the individual weather takes the treatment. The treatment decision T is a threshold-crossing model and linearly depend on a set of observed characteristics $\mathbf{z}$. The two potential outcomes linearly depends on some covariate $\mathbf{x}$ through slope parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$. Exclusion restriction requires that $\mathbf{z}$ includes some variables that are not in $\mathbf{x}$, therefore $Y(0)$ and $Y(1)$ are not independent of the treatment decision $\mathbf{T}$. Unobserved correlation between two potential outcomes and treatment decision $\mathbf{T}$ are described by the error terms U_T, U_1, and U_0. The covariance structure of the model is

$$\text{Cov}\begin{pmatrix} U_T \\ U_1 \\ U_0 \end{pmatrix} = \begin{pmatrix} 1 & \rho_{T1}\sigma_1 & \rho_{T0}\sigma_0 \\ \rho_{T1}\sigma_1 & \sigma_1^2 & \rho_{10}\sigma_1\sigma_0 \\ \rho_{T0}\sigma_0 & \rho_{10}\sigma_1\sigma_0 & \sigma_0^2 \end{pmatrix}. \tag{3.2}$$

and zero means and finite variances are required. Heckman et al. [2014] assumed that latent factors $\mathbf{f}$ drives the unobserved correlation between two potential outcomes and treatment decision $\mathbf{T}$

$$\begin{aligned} T &= 1\left(T^* > 0\right), \\ T^* &= \mathbf{z}\gamma + \mathbf{f}\alpha_T + \varepsilon_T, \\ Y(1) &= \mathbf{x}\beta_1 + \mathbf{f}\alpha_1 + \varepsilon_1, \\ Y(0) &= \mathbf{x}\beta_0 + \mathbf{f}\alpha_0 + \varepsilon_0. \end{aligned} \tag{3.3}$$

They also assumed that there are some continuous variables $M = (M_1, \ldots, M_Q)'$ to describe latent factors $\mathbf{f}$, the following system can be included to the model:

$$M = \mu(\mathbf{x}) + \Lambda\mathbf{f} + \varepsilon_M. \tag{3.4}$$

Where ε_M are mutually independent Gaussian distributions with zero means, $\mathrm{E}[\varepsilon_T] = \mathrm{E}[\varepsilon_1] = \mathrm{E}[\varepsilon_0] = 0$, and finite variances, $\mathrm{V}[\varepsilon_T] = 1$, $\mathrm{V}[\varepsilon_1] = \sigma_1^2$, $\mathrm{V}[\varepsilon_0] = \sigma_0^2$. Latent factors $\mathbf{f}$ explains all of the unobservable correlation of the model.

$$\mathrm{Cov}(M|\mathbf{x}) = \Lambda\Sigma_f\Lambda^T + Y_\varepsilon. \tag{3.5}$$

Besides, let ε denote $\{\varepsilon_T, \varepsilon_1, \varepsilon_0, \varepsilon_M\}$, $\mathbf{f} \perp \varepsilon \perp (\mathbf{z}, \mathbf{x})$. The latent factors $\mathbf{f}$ are also assumed to be a Gaussian distribution with zero means and diagonal covariance matrix.

3.1 Model Assumptions

The adoption of instrumental variables in the treatment effect model is an innovative start to investigate correlation between two potential outcomes, Heckman et al. [2014] gave details of this method. Under the framework of Heckman et al. [2014], I replaced exogenous hypothesis with Unconfounded Assignment assumption and combine the model with Gaussian Mixture Linear Regression, therefore the model in this thesis is called Gaussian Mixture Linear Regression with unobservable covariance structure.

Formally, treatment assignment depends on observed covariate $\mathbf{x}$ and is independent of latent variables $\mathbf{f}$, which is different from the model structure in Heckman et al. [2014]. Heckman et al. [2014] assumed that treatment assignment depends on $\mathbf{z}$ and $\mathbf{f}$ and $\mathbf{T}$ is related to two potential outcomes. But here, we adopted a more general assumption: Unconfounded assignment assumption, which is $\{Y(0), Y(1)\} \perp T|\mathbf{x}$. Therefore $\mathbf{T}$ does not help to estimate correlation between two potential outcomes.

For the Gaussian distribution k, if we observed the latent factor $\mathbf{f}$,

$$\begin{aligned} Y(1) &= \mathbf{x}\boldsymbol{\beta}_{k1} + \mathbf{f}\boldsymbol{\alpha}_{k1} + \varepsilon_{k1}, \\ Y(0) &= \mathbf{x}\boldsymbol{\beta}_{k0} + \mathbf{f}\boldsymbol{\alpha}_{k0} + \varepsilon_{k0}. \end{aligned} \tag{3.6}$$

ε_{k1} and ε_{k0} are independent normal distribution with zero means

$$\text{Cov}\begin{pmatrix} \varepsilon_{k1} \\ \varepsilon_{k0} \end{pmatrix} = \begin{pmatrix} \sigma_{k1}^2 & 0 \\ o & \sigma_{k0}^2 \end{pmatrix}.$$

Which means

$$p\left(\mathbf{y}|\mathbf{x},\mathbf{f},z_k=1\right) = \frac{1}{\sqrt{(2\pi)^p\left|\mathbf{\Sigma}_k\right|}}\exp\left[-\frac{1}{2}\left(\mathbf{y}-\mathbf{A}\boldsymbol{\beta}_k-\mathbf{W}\boldsymbol{\alpha}_k\right)^T\mathbf{\Sigma}_k^{-1}\left(\mathbf{y}-\mathbf{A}\boldsymbol{\beta}_k-\mathbf{W}\boldsymbol{\alpha}_k\right)\right]. \tag{3.7}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{x} & \\ & \mathbf{x} \end{bmatrix}, \quad \boldsymbol{\beta}_k = \begin{bmatrix} \boldsymbol{\beta}_{k,0} \\ \boldsymbol{\beta}_{k,1} \end{bmatrix},$$
$$\mathbf{W} = \begin{bmatrix} \mathbf{f} & \\ & \mathbf{f} \end{bmatrix}, \quad \boldsymbol{\alpha}_k = \begin{bmatrix} \boldsymbol{\alpha}_{k,0} \\ \boldsymbol{\alpha}_{k,1} \end{bmatrix}, \quad \mathbf{\Sigma}_k = \begin{bmatrix} \sigma_{k1}^2 & \\ & \sigma_{k0}^2 \end{bmatrix}.$$

If we don't observe the latent factor,

$$\begin{aligned} Y(1) &= \mathbf{x}\boldsymbol{\beta}_{k1} + U_{k1}, \\ Y(0) &= \mathbf{x}\boldsymbol{\beta}_{k0} + U_{k0}. \end{aligned} \tag{3.8}$$

The covariance structure of the model is

$$\text{Cov}\begin{pmatrix} U_{k1} \\ U_{k0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_{k1}^T\boldsymbol{\alpha}_{k1}+\sigma_{k1}^2 & \boldsymbol{\alpha}_{k1}^T\boldsymbol{\alpha}_{k0} \\ \boldsymbol{\alpha}_{k1}^T\boldsymbol{\alpha}_{k0} & \boldsymbol{\alpha}_{k0}^T\boldsymbol{\alpha}_{k0}+\sigma_{k0}^2 \end{pmatrix}.$$

Then

$$p\left(\mathbf{y}|\mathbf{x},z_k=1\right) = \frac{1}{\sqrt{(2\pi)^p\left|\mathbf{\Sigma}_k\right|}}\exp\left[-\frac{1}{2}\left(\mathbf{y}-A\boldsymbol{\beta}_k\right)^T\mathbf{\Sigma}_k^{-1}\left(\mathbf{y}-A\boldsymbol{\beta}_k\right)\right]. \tag{3.9}$$

where

$$\mathbf{y} = \begin{bmatrix} \mathbf{Y}(0) \\ \mathbf{Y}(1) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \mathbf{x} & \\ & \mathbf{x} \end{bmatrix}, \quad \boldsymbol{\beta}_k = \begin{bmatrix} \boldsymbol{\beta}_{k,0} \\ \boldsymbol{\beta}_{k,1} \end{bmatrix},$$
$$\boldsymbol{\alpha}_k = \begin{bmatrix} \boldsymbol{\alpha}_{k,0} \\ \boldsymbol{\alpha}_{k,1} \end{bmatrix}, \quad \mathbf{\Sigma}_k = \begin{bmatrix} \boldsymbol{\alpha}_{k1}^T\boldsymbol{\alpha}_{k1}+\sigma_{k1}^2 & \boldsymbol{\alpha}_{k1}^T\boldsymbol{\alpha}_{k0} \\ \boldsymbol{\alpha}_{k1}^T\boldsymbol{\alpha}_{k0} & \boldsymbol{\alpha}_{k0}^T\boldsymbol{\alpha}_{k0}+\sigma_{k0}^2 \end{bmatrix}.$$

Equation (3.9) explains the correlation between two potential outcomes clearly.

The log-likelihood function of n observations is

$$\ln p(\mathbf{Y}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}_i | \mathbf{A}_i \boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k\right) \right\}. \tag{3.10}$$

Assuming that there are Q continuous variables $M = (M_1, \ldots, M_Q)'$ to describe latent factors $\mathbf{f}$, we consider:

$$M = \boldsymbol{\Lambda}\mathbf{f} + \varepsilon_M. \tag{3.11}$$

Here, we assume that $E(\mathbf{f}) = 1$ and $var(\mathbf{f}) = I_p$.

The differences between above model (denote as model A) and the model in Heckman et al. [2014] (denote as model B) are lies in:

1. The model B included Exogenous hypothesis, which means that at least one variable not included in $\mathbf{X}$ is required, which is different from the model A. Here we drop out Exogenous hypothesis and replaced it with a general assumption - Unobservable assignment assumption.

2. There are nonzero correlations between T and $Y(0), Y(1)$ in model B. In this thesis, independence between T and $Y(0), Y(1)$ given $\mathbf{X}$ is assumed, which is a common case.

3. Gaussian Mixture Model is an effective way to describe distribution. Combining Gaussian Mixture Model with instrumental variables is an innovative investigation of treatment effect. Gaussian Mixture Model captures flexible feature of distribution and instrumental variables explains unobserved correlation.

4. The correlation between two potential outcomes is reflected in the load of instrumental variables on the potential outcomes, which alleviate the dilemma of Gaussian mixture regression and Gaussian mixture linear regression in estimating covariance.

3.2 Deriving Treatment effects

Let $\Delta \equiv Y(1) - Y(0)$ denote the benefits of the treatment effect, the conditional distribution of Δ is written as:

$$p(\Delta|\mathbf{x}, z_k = 1) = \mathcal{N}\left(\Delta|\mathbf{x}(\boldsymbol{\beta}_{k1} - \boldsymbol{\beta}_{k0}), \sigma^2_{\Delta k}\right). \tag{3.12}$$

where $\sigma^2_{\Delta k} \equiv (\boldsymbol{\alpha}_{k1} - \boldsymbol{\alpha}_{k0})^T (\boldsymbol{\alpha}_{k1} - \boldsymbol{\alpha}_{k0}) + \sigma^2_{k1} + \sigma^2_{k0}$.

The joint distribution of two potential outcomes is a Gaussian Mixture Model. Thus, the distribution of Δ is also a Gaussian Mixture Model, which can be presented as:

$$p(\Delta) = \sum_z p(z)p(\Delta|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\Delta|\mathbf{x}(\boldsymbol{\beta}_{k1} - \boldsymbol{\beta}_{k0}), \sigma^2_{\Delta k}\right). \tag{3.13}$$

Individual treatment effect is

$$\text{ITE}(\mathbf{x}) \equiv \text{E}[\Delta|\mathbf{x}] = \sum_{k=1}^{K} \pi_k \mathbf{x}(\boldsymbol{\beta}_{k1} - \boldsymbol{\beta}_{k0}). \tag{3.14}$$

Besides, the probability of treatment benefit is

$$\text{P}(\Delta > 0|\mathbf{x}) = \sum_{k=1}^{K} \pi_k \Phi\left(\frac{\mathbf{x}\left(\boldsymbol{\beta}_{k1} - \boldsymbol{\beta}_{k0}\right)}{\sigma_{\Delta k}}\right). \tag{3.15}$$

where $\Phi\left(\frac{\mathbf{x}(\boldsymbol{\beta}_{k1} - \boldsymbol{\beta}_{k0})}{\sigma_{\Delta k}}\right)$ denotes the cumulative distribution function of multivariate Gaussian distribution with mean $\mathbf{x}\left(\boldsymbol{\beta}_{k1} - \boldsymbol{\beta}_{k0}\right)$ and variance $\sigma^2_{\Delta k}$.

3.3 Pre-Post EM Algorithm

Gibbs sampling, used in Heckman et al. [2014], is unapplicable because of Gaussian Mixture model. We propose an effective algorithm -Pre-Post EM Algorithm- to estimate Gaussian Mixture Linear Regression within unobserved correlation structure. Before EM Algorithm, there are some restrictions to be discussed. The covariance matrix of the factors $\Sigma_{\mathbf{f}}$ can be a very general matrix, like an off-diagonal matrix containing variable dependencies, similar for the covariance matrix $\Lambda\Sigma_{\mathbf{f}}\Lambda^T$ of measurements M. The covariance structure is unchanged if there is a nonsingular matrix

U such that $\Lambda^* = \Lambda U$ and $\mathbf{f}^* = U^{-1}\mathbf{f}$. Therefore the indeterminacy problem of $\Sigma_\mathbf{f}$ and $\Lambda\Sigma_\mathbf{f}\Lambda^T$ prevent us from getting an accurate estimation of the model. Some solutions can be: i) Assuming that the matrix of factors $\Sigma_\mathbf{f}$ are diagonal, which means that factors are uncorrelated. But this assumption only partially solves the indeterminacy problem, any orthogonal matrix such that $U^{-1} = U'$ will leads to the problem; ii) Anderson and Rubin [1956] put forward that some other restrictions should be made on the matrix $\mathbf{\Lambda}$ and/or $\Sigma_\mathbf{f}$, such as diagonal Σ_f ,block factor loading matrix or fixing loading, which means that a block of measurements on M are used to describe every component of $\mathbf{f}$.

In this section, I generalize Gaussian Mixture Regression to two dimensional case and apply it to treatment effect. Additional challenge is only one observed potential outcome. Following the idea of Ghahramani and Jordan [1994], this thesis gives the detail of Pre-Post EM Algorithm, which can be find in Algorithm 4.

The log-likelihood function of the model is

$$\begin{aligned}\ln p(\mathbf{Y}, \mathbf{M}|\mathbf{A}, \boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \mathbf{\Lambda}, \mathbf{\Sigma}_M) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}_i|\mathbf{A}_i\boldsymbol{\beta}_k + \boldsymbol{\alpha}_k f_i, \mathbf{\Sigma}_k\right) \right\} \\ + \ln \mathcal{N}\left(\mathbf{M}_i|\mathbf{\Lambda} f_i, \mathbf{\Sigma}_M\right).\end{aligned} \tag{3.16}$$

There is a step before E step, called "Pre step", which takes derivation of the equation 3.16 with respect to $\mathbf{\Lambda}$ and $\mathbf{\Sigma}_M$, we obtain

$$\begin{aligned}\Lambda^{new} &= \frac{\sum_{i=1}^{n} \mathbf{M}_i f_i}{\sum_{i=1}^{n} f_i}, \\ \mathbf{\Sigma}_M^{new} &= \frac{\sum_{i=1}^{n} \left(\mathbf{M}_I - \mathbf{\Lambda} f_i\right)\left(\mathbf{M}_i - \mathbf{\Lambda} f_i\right)^{\mathrm{T}}}{n}.\end{aligned} \tag{3.17}$$

In the E step, take the derivative of $\ln p(\mathbf{Y}|\mathbf{A}, \mathbf{Z}, \boldsymbol{\theta}) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$ with respect to π_k, we obtain

$$\pi_k^{new} = \frac{\sum_{i=1}^{n} \lambda(z_{ik})}{n} \quad (k = 1, \ldots, k). \tag{3.18}$$

where

$$\gamma\left(z_{ik}\right) = \frac{\pi_k \mathcal{N}\left(y_i^o|\mathbf{x}_i\boldsymbol{\beta}_{k,o} + \boldsymbol{\alpha}_{k,o} f_i, \sigma_{k,o}^2\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(y_i^o|\mathbf{x}_i\boldsymbol{\beta}_{j,o} + \boldsymbol{\alpha}_{j,o} f_i, \sigma_{j,o}^2\right)}. \tag{3.19}$$

similarly, update $\lambda(z_{ik})$ based on observed data.

If $T_n = 1$

$$\gamma\left(z_{ik}\right) = \frac{\pi_k \mathcal{N}\left(Y_i(1)|\mathbf{x}_n\boldsymbol{\beta}_{k1} + \alpha_{k1}f_i, \sigma_{k1}^2\right)}{\sum_{j=1}^{K}\pi_j \mathcal{N}\left(Y_i(1)|\mathbf{x}_i\boldsymbol{\beta}_{j1} + \alpha_{j1}f_i, \sigma_{j1}^2\right)}.$$

If $T_n = 0$

$$\gamma\left(z_{ik}\right) = \frac{\pi_k \mathcal{N}\left(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{k0} + \alpha_{k0}f_i, \sigma_{k0}^2\right)}{\sum_{j=1}^{K}\pi_j \mathcal{N}\left(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{j0} + \alpha_{j0}f_i, \sigma_{j0}^2\right)}.$$

In the M step, take the derivation of Q function with respect to $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_k$,

$$\boldsymbol{\beta}_k^{\text{new}} = \left[\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\mathbf{A}_i\Sigma_k^{-1}\mathbf{A}_i^T\right]^{-1}\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\mathbf{A}_i^T\Sigma_k^{-1}\left(\mathbf{y}_i - \boldsymbol{\alpha}f_i\right), \tag{3.20}$$

$$\Sigma_k{}^{\text{new}} = \frac{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)\left(y_i - \mathbf{A}_i\boldsymbol{\beta}_k^{\text{new}} - \boldsymbol{\alpha}f_i\right)\left(y_i - \mathbf{A}_i\boldsymbol{\beta}_k^{\text{new}} - \boldsymbol{\alpha}f_i\right)^T}{\sum_{i=1}^{n}\gamma\left(z_{ik}\right)}. \tag{3.21}$$

If $T_i = 0$, impute missing values $Y_i(1)$ in M step by

$$E\left[\lambda\left(z_{ik}\right)Y_i(1)|\mathbf{x}_i, Y_i(0), \theta_k\right] = \lambda\left(z_{ik}\right)E\left[Y_i(1)|z_{ik} = 1, \mathbf{x}_i, Y_i(0), \theta_k\right] \tag{3.22}$$

$$= \lambda\left(z_{ik}\right)\left(\mathbf{x}_i\boldsymbol{\beta}_{k1} + \boldsymbol{\alpha}_{k1}f_i + \Sigma_k^{\text{mo}}\Sigma_k^{\text{oo}^{-1}}\left(Y_i(0) - \mathbf{x}_i\boldsymbol{\beta}_{k0} - \boldsymbol{\alpha}_{k0}f_i\right)\right). \tag{3.23}$$

where m, o represent missing and observed items. $\lambda\left(z_{ik}\right) = E\left[z_{ik}|\mathbf{x}_i, Y_i(0), \theta_k\right]$. Define $\hat{Y}_{ik}(1) \equiv E\left[Y_i(1)|z_{nk} = 1, \mathbf{x}_i, Y_i(0), \theta_k\right]$,

$$E\left[z_{ik}Y_i^2(1)|Y_i(0), \theta_k\right] = h_{ik}\left(\Sigma_k^{mm} - \Sigma_k^{mo}\Sigma_k^{oo^{-1}}\Sigma_k^{moT} + \hat{Y}_{ik}^2(0)\right). \tag{3.24}$$

Similar for those missing values $Y_i(0)$.

After M step, there is also additional step, called "Post step". Variable f is a latent variable and we update values by taking derivation of the equation 3.16 with respect to f

If $T_i = 1$

$$\boldsymbol{f}_i^{new} = \frac{\sum_{k=1}^{K}\lambda(z_{ik})\frac{\alpha_{k1}(Y_i(1) - \mathbf{x}_i\beta_{k1})}{\sigma_{k1}^2} + \sum_{j=1}^{Q}\frac{M_{ij}\lambda_j}{\sigma_{Mj}^2}}{\sum_{k=1}^{K}\frac{\lambda(z_{ik})\alpha_{k1}^2}{\sigma_{k1}^2} + \sum_{j=1}^{Q}\frac{\lambda_j^2}{\sigma_{Mj}^2}}. \tag{3.25}$$

If $T_i = 0$

$$\boldsymbol{f}_i^{new} = \frac{\sum_{k=1}^{K}\lambda(z_{ik})\frac{\alpha_{k0}(Y_i(0) - \mathbf{x}_i\beta_{k0})}{\sigma_{k0}^2} + \sum_{j=1}^{Q}\frac{M_{ij}\lambda_j}{\sigma_{Mj}^2}}{\sum_{k=1}^{K}\frac{\lambda(z_{ik})\alpha_{k0}^2}{\sigma_{k0}^2} + \sum_{J=1}^{Q}\frac{\lambda_J^2}{\sigma_{Mj}^2}}. \tag{3.26}$$

Since we have assumed that $f \sim \mathcal{N}(0, I)$, it s necessary to normalize those values f.

The log likelihood function of Gaussian Mixture Linear Regression within unobserved correlation structure is

$$\begin{aligned}
\ln p(\mathbf{Y}, \mathbf{M}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln P\left(\mathbf{Y}_i|\mathbf{x}, \mathbf{z}_i, f_i\theta\right) P\left(\mathbf{z}_i|\theta\right) + \ln p(\mathbf{M}_i|f_i, \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln \mathcal{N}\left(\mathbf{y}_i|\mathbf{A}_i\boldsymbol{\beta}_k + f_i\boldsymbol{\alpha}_k, \boldsymbol{\Sigma}_k\right) + \sum_{i=1}^{n} \ln \mathcal{N}\left(\mathbf{M}_i|\boldsymbol{\Lambda} f_i, \boldsymbol{\Sigma}_M\right) \\
&\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \ln P\left(\mathbf{z}_i|\theta\right).
\end{aligned} \tag{3.27}$$

and Q function takes the expectation to unknown component z and missing values:

$$\begin{aligned}
\mathcal{Q}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) &= \sum_{\mathbf{Z}} p\left(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}^{\text{old}}\right) \mathbb{E}_{\mathbf{Y^m}}[\ln p(\mathbf{Y}, \mathbf{M}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \lambda\left(z_{ik}\right) \Bigg\{ -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_k| \\
&\quad - \frac{1}{2}\left(\mathbf{y}_i^{\circ} - \mathbf{x}_i\boldsymbol{\beta}_k^{\circ} - \mathbf{f}_i\boldsymbol{\alpha}_k^{\circ}\right)^T \Sigma_k^{-1,\text{oo}} \left(\mathbf{y}_i^{\circ} - \mathbf{x}_i\boldsymbol{\beta}_k^{\circ} - \mathbf{f}_i\boldsymbol{\alpha}_k^{\circ}\right) \\
&\quad \mathbb{E}_{\mathbf{Y^m}}\Big[- \left(\mathbf{y}_i^{\circ} - \mathbf{x}_i\boldsymbol{\beta}_k^{\circ} - \mathbf{f}_i\boldsymbol{\alpha}_k^{\circ}\right)^T \Sigma_k^{-1,\text{om}} \left(\mathbf{y}_i^{\text{m}} - \mathbf{x}_i\boldsymbol{\beta}_k^{\text{m}} - \mathbf{f}_i\boldsymbol{\alpha}_k^{\text{m}}\right) \\
&\quad - \left(\mathbf{y}_i^{\text{m}} - \mathbf{x}_i\boldsymbol{\beta}_k^{\text{m}} - \mathbf{f}_i\boldsymbol{\alpha}_k^{\text{m}}\right)^T \Sigma_k^{-1,\text{mm}} \left(\mathbf{y}_i^{\text{m}} - \mathbf{x}_i\boldsymbol{\beta}_k^{\text{m}} - \mathbf{f}_i\boldsymbol{\alpha}_k^{\text{m}}\right)\Big] \Bigg\} \\
&\quad + \left[-\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_M| - \left(\mathbf{M}_i - \boldsymbol{\Lambda} f_i\right)^T \Sigma_M^{-1} \left(\mathbf{M}_i - \boldsymbol{\Lambda} f_i\right)\right].
\end{aligned} \tag{3.28}$$

Since there are some missing values $y(0)$ and $y(1)$, we estimate $\lambda(z_{ik})$ and π_k based on observed data. In the **M** step, we take the expectations to unknown component z and missing values and then take derivation of Q function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. For those unobserved values $y(0)$ and $y(1)$, we use the equation 3.23 and 3.24 to replace those expectations of missing values in the Q function.

Pre-Post EM Algorithm 4 is used to estimate GMLR within unobserved correlation structure when there are hidden latent variable and unobserved potential outcomes in Gaussian Mixture Linear Model.

Algorithm 4 EM Algorithm for GMLR with unobserved correlation structure

1.**Initialize** $\boldsymbol{\beta}_{k1}$, $\boldsymbol{\beta}_{k0}$, $\boldsymbol{\alpha}_{k1}$, $\boldsymbol{\alpha}_{k0}$,$\boldsymbol{\sigma}_{k1}$, $\boldsymbol{\sigma}_{k0}$, $\boldsymbol{\Lambda}$, ε_M, $f_1, f_2, ..., f_n$

2. **Pre-step**

$$\lambda_i^{new} = \frac{\sum_{n=1}^{N} M_{ni} f_n}{\sum_{n=1}^{N} f_n^2}, \qquad \sigma_{mi}^{2\ new} = \frac{\sum_{n=1}^{N} (M_{ni} - \lambda_i f_n)^2}{n}.$$

3. **E-step**. Update the responsibilities based on the current parameter values
If $T_i = 1$

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(Y_i(1)|\mathbf{x}_i\boldsymbol{\beta}_{k1}, \sigma_{k1}^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(Y_i(1)|\mathbf{x}_i\boldsymbol{\beta}_{j1}, \sigma_{j1}^2\right)}.$$

Otherwise

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{k0}, \sigma_{k0}^2)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(Y_i(0)|\mathbf{x}_i\boldsymbol{\beta}_{j0}, \sigma_{j0}^2\right)}.$$

4. **M-step**. Update the parameters based on the current responsibilities

$$\begin{aligned}
\boldsymbol{\beta}_{k1}^{\text{new}} &= \left[\sum_{T_i=1} \gamma(z_{ik})\, \mathbf{x}_i \mathbf{x}_i^T\right]^{-1} \sum_{T_i=1} \gamma(z_{ik})\, \mathbf{x}_i^T (Y_i(1) - \alpha_{k1} f_i), \\
\boldsymbol{\alpha}_{k1}^{\text{new}} &= \frac{\sum_{T_i=1} \gamma(z_{ik})\, f_i (Y_i(1) - \mathbf{x}_i \beta_{k1})}{\sum_{T_i=1} \gamma(z_{ik})\, f_i^2}, \\
\sigma_{k1}^{2\ \text{new}} &= \frac{\sum_{T_i=1} \gamma(z_{ik}) \left(Y_i(1) - \mathbf{x}_i \boldsymbol{\beta}_{k1}^{\text{new}} - \alpha_{k1} f_i\right)^2}{\sum_{T_n=1} \gamma(z_{ik})}.
\end{aligned}$$

and $\pi_k^{\text{new}} = \frac{N_k}{n}, \quad N_k = \sum_{i=1}^{n} \gamma(z_{ik})$.
Impute missing values $Y_i(1)$ and $Y_i(0)$ in M step by the equations 3.23 and 3.24.

5. **Post-step** Update latent variables.
If $T_i = 1$

$$\boldsymbol{f}_i^{new} = \frac{\sum_{k=1}^{K} \lambda_{ik} \frac{\alpha_{k1}(Y_i(1) - \mathbf{x}_i\beta_{k1})}{\sigma_{k1}^2} + \sum_{j=1}^{Q} \frac{M_{ij}\lambda_j}{\sigma_{Mj}^2}}{\sum_{k=1}^{K} \frac{\lambda_{ik}\alpha_{k1}^2}{\sigma_{k1}^2} + \sum_{j=1}^{Q} \frac{\lambda_j^2}{\sigma_{Mj}^2}}.$$

If $T_i = 0$

$$\boldsymbol{f}_i^{new} = \frac{\sum_{k=1}^{K} \lambda_{ik} \frac{\alpha_{k0}(Y_i(0) - \mathbf{x}_i\beta_{k0})}{\sigma_{k0}^2} + \sum_{j=1}^{Q} \frac{M_{ij}\lambda_j}{\sigma_{Mj}^2}}{\sum_{k=1}^{K} \frac{\lambda_{ik}\alpha_{k0}^2}{\sigma_{k0}^2} + \sum_{j=1}^{Q} \frac{\lambda_j^2}{\sigma_{Mj}^2}}.$$

Normalize f.

6. Record the log likelihood

$$\ln p(\mathbf{Y}, \mathbf{M}|\mathbf{A}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}_i|\mathbf{A}_i\boldsymbol{\beta}_k + \mathbf{f}_i\boldsymbol{\alpha}_k, \boldsymbol{\Sigma}_k) \right\} + \ln p(\mathbf{M}_i|\boldsymbol{\Lambda}_M f_i, \boldsymbol{\Sigma}_M).$$

and check the convergence of sum of the parameters otherwise return to step 2.

3.4 Simulation

In this section, we do some simulations.

- *Initial conditions.*In our simulation study, we generate random values from the set {1,2} as initial values of $\boldsymbol{\beta}$ and generate orthogonal matrices U and diagonal matrix D = diag[1, 3] and $\boldsymbol{\Sigma} = U \times D \times U'$. Set $\boldsymbol{\alpha}_{k0} = 0.5, k = 1, 2$ $\alpha_{k1} = 0.5, k = 1, 2$ $\boldsymbol{\alpha}_{k1} = 0.5, k = 1, 2$ $\Lambda = [1, 1, 1, 1]$

- Data set. We generate each datum $(\mathbf{x}_i, T_i, Y_i(0), Y_i(1), \mathbf{M}_i)$ as follows.

  i). z_i was generated from a binary distribution with $P(z_1 = 1) = 0.4$ and $P(z_2 = 1) = 0.6$.

  ii). $\mathbf{x}_i$ was generated from multivariate normal distribution.

  iii). T_i was generated from a logistic distribution with $P(T = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\alpha})}{1+\exp(\mathbf{x}^T\boldsymbol{\alpha})}, \boldsymbol{\alpha} = (1, -2)$.

  iv). $\varepsilon_i|z_{ik} = 1$ was generated from multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_k)$;

  v). f_i was generated from standard normal distribution;

  vi). $(y_i|\mathbf{A}_i, z_{ik} = 1) = \mathbf{A}_i\boldsymbol{\beta}_k + \boldsymbol{\alpha}_k f_i + \boldsymbol{\varepsilon}_i$.

  vii). ε_M was generated from multivariate normal distribution, $\mathbf{M}_i = \boldsymbol{\Lambda} f_i + \varepsilon_{M_i}$.

- *Number of samples.* The size of the sample is 1000.

- *Stopping rules.* A stopping criterion for Pre-Post EM Algorithm we consider is that the absolute sum of parameters is less than 10^{-2}. We investigate how the Q function reacts to the increasing number of iterations.

The simulation process contains 2 steps:

(1). Create a data set whose size is n.

(2). Perform EM algorithm 4 to estimate the Gaussian Mixture Linear Regression within unobserved correlation structure. Save Q function and the estimated parameters $\hat{\theta} = \left(\hat{\pi}_1, \hat{\pi}_2, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2\right)$.

Simulation results are displaced in table 3.1-3.3.

Table 3.1: True parameter values with a two-component GMLMUC.

Configuration	Control			Treat		
	(β_1, β_2)	α_1	σ^2	(β_1, β_2)	α_0	σ^2
parallel(0.4)	$(-2.0, 0.5)$	1	1	$(0.5, 2.0)$	1.5	0.64
concurrent(0.6)	$(-1.5, 1.0)$	2	0.81	$(1.0, 1.5)$	1	1

Table 3.2: Estimate results with a two-component GMLM-UC.

Configuration	Control			Treat		
	(β_1, β_2)	α_1	σ^2	(β_1, β_2)	α_0	σ^2
parallel(0.39)	$(-2.23, 0.58)$	0.91	0.70	$(0.48, 2.03)$	1.27	0.66
concurrent(0.61)	$(-1.46, 0.90)$	1.98	0.89	$(1.02, 1.42)$	1.10	0.93

Table 3.3: True parameter values & Estimate results for GMLMUC.

	$(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$	$(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$
True values	$(2, 1, 0.5, 0.25)$	$(0.1, 0.2, 0.3, 0.5)$
Estimate results	$(2.04, 0.99, 0.52, 0.29)$	$(0.01, 0.22, 0.30, 0.50)$

Measures of algorithm performance

There are several measurements to compare among various models. But we presented the distribution of Δ and obtained estimate results of parameters in three models. A meaningful measurement is the precision in Estimation of Heterogeneous Effect(PEHE), which is reported in table 3.4

$$PEHE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(Y_i(1) - Y_i(0)) - (\hat{y}_i(1) - \hat{y}_i(0))\right]^2}. \tag{3.29}$$

It evaluates the ability of each method to capture treatment effect heterogeneity.

Table 3.4: Estimate of treatment effect for GMLR-UC.

	BART	CasualForest	GMLR-UC
PEHE	9.9918	5.0621	2.7076

Conclusion

1. Gaussian Mixture Linear Regression with Unobserved Covariance structure performs well not only in estimating expectation and variance of two potential outcomes, but also in estimating correlation. Estimation of the joint distribution of $Y(0)$ and $Y(1)$ are pretty good, which means that we get a good estimation

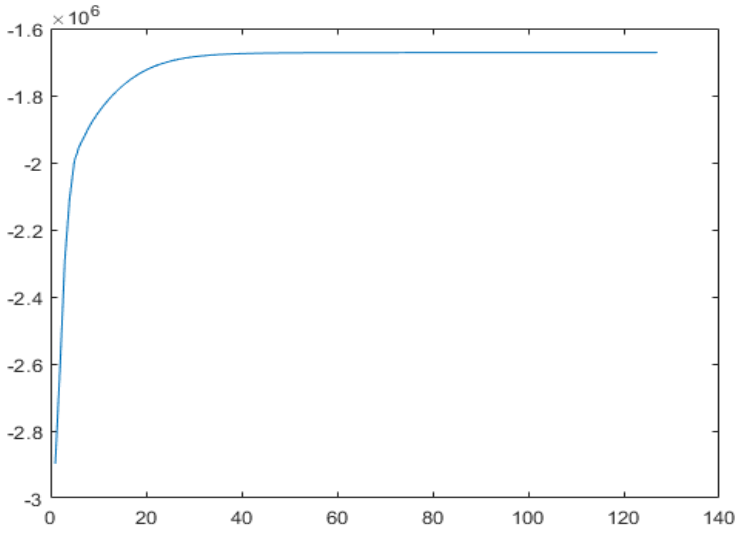

Figure 3.1: Q function for Gaussian Mixture Linear Regression with Covariance structure

for expectation of distribution of $\tau(\mathbf{x})$.

2. The Q function increases with the number of iterations, which proves the effectiveness of Pre-Post EM Algorithm. That's because we impute missing values in the M step in order to make sure the increasing of Q function.

3. The table 3.4 also shows the superiority of Gaussian Mixture Linear Regression with Unobservable covariance. Compared with other model like Bayesian Additive Regression Tree and Casual forest, Gaussian Mixture Linear Regression with unobserved covariance structure presents a distribution of treatment effect, which enables us to better understand the treatment effect.

Chapter 4

Conclusion

In this work, we explained the concept of treatment effect and reviewed literatures related to individual treatment effect, one is Frequentist or classical inference, such as methods based on propensity score, domain adaption and Tree-based methods, the other one is Bayesian paradigm, which offers inference for the distribution of missing data and treatment effect. In order to investigate the correlation between two potential outcomes from Bayesian perspective, we improved and expanded Gaussian Mixture Model to treatment effect in the Chapter 2. Besides, latent factors explain correlation between two potential outcomes and help us get rid of the dilemma, in which we failed to give right estimation of covariance matrix.

Firstly, we improved Gaussian Mixture Regression and applied it to treatment effect and replaced missing values by conditional expectation. Our simulation performs well in estimating means and variances. But we found that the estimation of covariance is terrible, that's because we update missing items by using same correlation when we update covariance metrix.

Then, we extended Gaussian Mixture Linear Regression to two dimensional case and applied it to estimate treatment effect. Because it is meaningful to describe the effect of X to Y. Replacing missing values with conditional expectation, we adjusted the general EM algorithm to estimate treatment effect via Gaussian Mixture Linear Regression. Same as Gaussian Mixture Regression, our simulation performs well in estimating means and variances. But we also found that the estimation of covariance is terrible. What we conclude from these two models is that the current

data is not enough to describe correlation between two potential outcomes, additional information should be included to capture correlation.

Last but not least, we improved Gaussian Mixture Linear Regression and combined it with latent variables. Latent factors explain correlation between two potential outcomes and Gaussian Mixture Linear Regression describe flexible joint distribution of potential outcomes. Besides, we dropped the Exogenous assumption and adopted a more general assumption- unconfounded treatment assignment, which means that the treatment assignment is independent of two potential outcomes if we have observed covariates. Similarly, we replaced missing values with conditional expectation and proposed an effective algorithm - Pre-Post EM Algorithm- to estimate Gaussian Mixture Linear Regression within unobserved correlation. Our simulation performs well in estimation process.

Gaussian Mixture Model is flexible to describe distribution and EM algorithm is an effective tool to estimate mixture model. Q function increases with the increasing of number of iteration-Q functions for three mixture model are plotted, which shows the effectiveness of alternated EM Algorithms.

The main contributions of this thesis are: (1) Gaussian Mixture Model is considered firstly in the treatment effect problems to motivate further investigation. (2) Altered versions of Gaussian Mixture Regression and Gaussian Mixture Linear Regression are proposed in this thesis to improve the performance. Meanwhile we adjusted the models for more general cases- unconfunded treatment assignment.(3)Altered EM algorithms are proposed to estimate proposed models and simulations are presented to justify the effectiveness of proposed algorithms.

We will consider future work from following two points: The first one is high dimensional case, how does GMLR-UC perfom if the dimension of covariate X is relatively high? Weather we can get a good estimate of parameters or not?

The second one is the Statistical guarantees for the EM algorithm and Pre-Post EM algorithm. Wainwrightt et al. [2017] discussed some statistical properties about standard EM algorithm, like this equation,

$$\left\|\theta^t - \theta^*\right\|_2 \leq \kappa^t \left\|\theta^0 - \theta^*\right\|_2 + \frac{1}{1-\kappa}\varepsilon_M\left(\frac{n}{T}, \frac{\delta}{T}\right).$$

we can find that the difference between new θ and the true θ is bounded. So we can investigate statistical properties of our adaptive EM Algorithm and Pre-Post EM algorithm.

Bibliography

J. Abrevaya, Y.-C. Hsu, and R. P. Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

A. M. Alaa and M. van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5):1031–1046, 2018.

T. W. Anderson and H. Rubin. Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pages 111–150, 1956.

S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

P. Carneiro, K. T. Hansen, and J. J. Heckman. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review*, 44(2): 361–422, 2003.

D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

P. Ding, F. Li, et al. Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237, 2018.

X. Ding and Q. Wang. Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association*, 106 (495):1193–1207, 2011.

X. Du, L. Sun, W. Duivesteijn, A. Nikolaev, and M. Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *arXiv preprint arXiv:1904.13335*, 2019.

S. Faria and G. Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

M. J. Funk, D. Westreich, C. Wiesen, T. Stürmer, M. A. Brookhart, and M. Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173 (7):761–767, 2011.

Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an em approach. In *Advances in neural information processing systems*, pages 120–127, 1994.

X. Guo, Y. Fang, X. Zhu, W. Xu, and L. Zhu. Semiparametric double robust and efficient estimation for mean functionals with response missing at random. *Computational Statistics & Data Analysis*, 128:325–339, 2018.

P. R. Hahn, J. S. Murray, C. M. Carvalho, et al. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

F. Han. Doubly robust estimation of the causal effects in the causal inference with missing outcome data. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–9, 2018.

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423, 2017.

J. J. Heckman, H. F. Lopes, and R. Piatek. Treatment effects: A bayesian perspective. *Econometric reviews*, 33(1-4):36–67, 2014.

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

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

S. Lee, R. Okui, and Y.-J. Whang. Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics*, 32 (7):1207–1225, 2017.

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

M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, 2018.

J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

D. F. McCaffrey, G. Ridgeway, and A. R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

D. B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

Y. V. Tan and J. Roy. Bayesian additive regression trees and the general bart model. *Statistics in medicine*, 38(25):5048–5069, 2019.

Z. Tan. Comment: Understanding or, ps and dr. *Statistical Science*, 22(4):560–568, 2007.

D. van Klaveren, Y. Vergouwe, V. Farooq, P. W. Serruys, and E. W. Steyerberg. Estimates of absolute treatment benefit for individual patients required careful modeling of statistical interactions. *Journal of clinical epidemiology*, 68(11):1366–1374, 2015.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018.

Wainwrightt, M. J., Bin, Balakrishnan, and Sivaraman. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics: An Official Journal of the Institute of Mathematical Statistics*, 2017.

D. Westreich, J. Lessler, and M. J. Funk. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, 63(8):826–833, 2010.

L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

J. Yoon, J. Jordon, and M. van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

S. Zhao and N. Heffernan. Estimating individual treatment effects from educational studies with residual counterfactual networks. In *10th International Conference on Educational Data Mining*, 2017.

CURRICULUM VITAE

Academic qualifications of the thesis author, Ms. WANG Juan:

- Received the degree of Bachelor of Economics from Shanxi University of Finance and Economics, June 2016.
- Received the degree of Master of Economics from University of International Business and Economics, June 2018.

August 2020