I, Jinzhong Liu, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Mathematical Sciences.

It is entitled:
Bayesian Inference for Treatment Effect

Student's name: Jinzhong Liu

This work and its defense approved by:

Committee chair: Siva Sivaganesan, Ph.D.
Committee member: Bin Huang
Committee member: Seongho Song, Ph.D.
Committee member: Xia Wang, Ph.D.
Bayesian Inference for Treatment Effect

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Department of Mathematical Sciences

of the College of Arts and Sciences

by

Jinzhong Liu

July 2017

Committee:
Siva Sivaganesan, Ph.D., Chair
Bin Huang, Ph.D., Co-Advisor
Soongho Song, Ph.D.
Xia Wang, Ph.D.
Abstract

Evaluation of overall treatment effect and heterogeneity in treatment effect is of interest in both randomized clinical trials and in observational studies. In this thesis, we first develop a Bayesian approach to subgroup analysis using ANOVA models with multiple covariates. We assume a two-arm clinical trial with normally distributed response variable. The covariates are assumed categorical and a priori specified. The subgroups of interest are represented by a collection of models. And we use a model selection approach to find subgroups with heterogeneous effects. Then we propose a Bayesian semiparametric approach for estimating the population mean treatment effect with observational data using Gaussian process (GP), which accomplishes matching and modeling outcome mechanism in a single step. We demonstrate a close link between matching method and GP regression for estimating average treatment effect. The proposed method utilizes a distance similar to Mahalanobis distance but determines the range of matching automatically without imposing a caliper arbitrarily. Finally, we proposed a Bayesian semiparametric approach for predicting the heterogeneous treatment effect for new patients using two conditionally independent Gaussian processes (GP), one for response surface of control group, the other for treatment effect. The prediction can be used to visualize the treatment effect and help researchers investigate the pattern of the treatment effect for different patient baseline characteristics and hence decide whether the treatment is effective for patients with certain characteristics and possibly define
a subgroup that the treatment is significantly effective on. We also illustrate the proposed methods using real data obtained from different studies.
Acknowledgments

First and foremost, I would like to take this opportunity to express my sincere gratitude to my advisor Dr. Siva Sivaganesan and co-advisor Dr. Bin Huang, who introduced me this interesting topic and offered invaluable suggestions and constant support during the research, with their expertise, understanding and patience. They spent a lot of time discussing with me about the topic I am interested in, tackling the problems that I came front in the research and revising my manuscripts. Without their tremendous help and guidance, this thesis would not be possible. We acknowledge the funding support from Patient Centered Outcome Research Institute (PCORI ME-148-19894 PI Huang).

I also would like to thank my committee faculty members: Dr. Seongho Song and Dr. Xia Wang, for their reviews and comments on the research. The lectures that they offered in the earlier years of my PhD study helped me solidify my statistical foundation and inspired me in my future studies. And other faculty members, Dr. Emily Lei Kang, Dr. Bledar Alex Konomi and Dr. Hang J. Kim, also provided me a lot of helpful suggestions and discussed some potential directions for the project during the advanced exam.

Finally, I would like to thank my dear parents, brothers, sisters and beloved one, Miao, for their selfless support and love that encouraged me pursuing my dream and continuing my research with happiness.
# Table of Contents

List of Figures vi

List of Tables vii

1 Introduction 1
   1.1 Subgroup Analysis in Clinical Trial 2
   1.2 Causal Inference with Observational Data 4
   1.3 Organization of this Dissertation 6

2 A Bayesian subgroup analysis using collections of ANOVA models 8
   2.1 Introduction 9
   2.2 ANOVA Model, Subgroup Effects and Model Space 13
      2.2.1 Subgroup Effects and Models 15
      2.2.2 An index representation of the models 17
   2.3 The $2 \times 2$ Case 18
      2.3.1 The Model Space and the Index Space 19
      2.3.2 Extension to more than 2 covariates 24
   2.4 Prior distributions and the Posterior Probabilities 26
      2.4.1 Prior distribution on the model space 26
      2.4.2 Prior distribution of the model-specific parameters 30
      2.4.3 Posterior Probabilities of the Models 31
      2.4.4 Equal Variance 32
   2.5 A Stepwise Algorithm for Finding Subgroup Effects 34
      2.5.1 The stepwise algorithm 36
   2.6 Operating characteristics and a real data example 41
      2.6.1 Introduction to STI example 42
      2.6.2 Error Rates 43
      2.6.3 Unequal Variance 44
      2.6.4 Equal Variance 47
2.6.5 Result for STI example ........................................ 49
2.7 Summary and discussion ......................................... 50

3 Bayesian Causal Inference Using Gaussian Process 54
3.1 Introduction ....................................................... 55
3.2 Causal inference ................................................. 58
3.3 Gaussian process for causal inference .......................... 59
  3.3.1 GP for estimating mean treatment effect ..................... 61
  3.3.2 Prior ......................................................... 63
  3.3.3 Posterior inference ......................................... 65
3.4 Links between GP and matching ................................. 67
  3.4.1 Estimating potential outcome with GP ....................... 67
  3.4.2 GP and subclassification matching .......................... 69
  3.4.3 GP and Mahalanobis matching ............................... 71
  3.4.4 GP and propensity score matching .......................... 72
3.5 Simulation study ................................................. 74
3.6 Case study ........................................................ 78
3.7 Discussion ........................................................ 80

4 Estimating the Heterogeneous Treatment Effects Using Gaussian Process 83
4.1 Introduction ....................................................... 84
4.2 Notation and assumptions ....................................... 87
  4.2.1 Notation ..................................................... 87
  4.2.2 Assumptions ............................................... 88
4.3 Gaussian process for estimating HTE .......................... 89
  4.3.1 Model ....................................................... 89
  4.3.2 Prior ....................................................... 91
  4.3.3 Posterior inference ....................................... 93
  4.3.4 Prediction ............................................... 95
4.4 Numerical study .................................................. 97
  4.4.1 Smooth surface .......................................... 97
  4.4.2 Stepwise treatment effect ................................ 102
4.5 Case study ........................................................ 103
4.6 Discussion ........................................................ 106

5 Summary and Future Work ........................................ 108

Bibliography .......................................................... 111
APPENDICES

A Subgroup Analysis
   A.1 Computing the marginal likelihoods under different models

B Causal Inference
   B.1 Algorithm for sampling from the posterior distribution

C HTE
   C.1 Algorithm for sampling from the posterior distribution
List of Figures

2.1 Unequal variance: TIE vs $c_0$. 

2.2 Unequal variance: (a) Rates vs effect size under overall effect model $M_1$, (b) Rates vs effect size under subgroup effect model $M_{(1,2),(0,0),(0,0,0,0)}$. 

2.3 Equal variance: TIE vs $c_0$. 

2.4 Equal variance: (a) Rates vs effect size under overall effect model $M_1$, (b) Rates vs effect size under subgroup effect model $M_{(1,2),(0,0),(0,0,0,0)}$. 

3.1 Weight of observations vs $|v_i - v_0|$. 

3.2 Results of Mahalanobis matching with different calipers. The circles are the average biases of estimates of ATE using Mahalanobis matching with specified calipers. The corresponding vertical lines indicate the ranges between 5th and 95th percentiles of the biases. The horizontal lines are the 5th percentile, average and 95th percentile of the biases of the estimates from GP. 

4.1 True treatment effect. 

4.2 Scatter plot of $(x_1, x_2)$. 

4.3 Predictive treatment effect vs $(x_1, x_2)$. 

4.4 Predictive treatment effect vs $x_1$. 

4.5 Predictive treatment effect vs Time and LspBMD. 

4.6 Scatter plot of (Time, PcntFat). 

vi
## List of Tables

2.1 Means of $Y_0$ and $Y_1$ within each subgroup .................................................. 19  
2.2 Partitions of equivalent model indices and the models they represent ................. 22  
2.3 Models and their unique indices when interaction is included ............................... 23  
2.4 Prior distribution on the index space $W_γ$ ......................................................... 29  
2.5 Sample sizes for the four subgroups for the STI example .................................... 43  
3.1 Performance of GP ...................................................................................................... 77  
3.2 Performance of linear model ....................................................................................... 77  
3.3 Performance of IPTW ................................................................................................. 78  
3.4 Performance of BART ................................................................................................. 78  
3.5 Baseline characteristics of patients .......................................................................... 80
Chapter 1

Introduction

Traditionally, a randomized control trial (RCT) focuses on assessing the efficacy of a treatment for the whole population. More and more investigators are now interested in studying the differential effect of the treatment on patients with different characteristics. Often, randomized control experiment is not feasible due to ethical, logical or cost concern. We need to use causal inference to estimate the population average treatment effect and find out the heterogeneous effect with the observational data. In this chapter, we will present some related concepts and existing methods for subgroup analysis and causal inference, followed by the outline of the dissertation.
1.1 Subgroup Analysis in Clinical Trial

While the clinical trial data mainly serves for evaluating the overall effect of a treatment, i.e., new drug or therapy, it or subsets of it could also be utilized for subgroup analysis to investigate if and how patients would benefit differently from the treatment. Subgroup analysis would help us to answer questions like "Is the treatment more effective on female than male?". Especially when the treatment is not significantly effective for all the participant, it would be more appealing to find out if it is significant in certain subgroups, such as patients within a age range or those with certain biomarkers.

Subgroup analysis splits the participants in a clinical trial data into several subgroups according to their baseline characteristics, such as gender, race, age, biomarkers, so as to compare the treatment effect on them. When conducting subgroup analysis, several difficulties would arise according to Pocock et al. (2002). Firstly, the randomized control trial is designed to evaluate the overall main effect of a treatment. It is quite possible that there is not sufficient sample size for detecting the subgroup effect even if it does exist. Secondly, investigators need to avoid data dredging by having vague definitions of subgroups. If the subgroups are not defined before the study, post hoc data exploration could pick up an "interesting" subgroup falsely. Thirdly, in the trial reports, the most appropriate methods for subgroup analysis may not be used. Statistical tests for interaction is recommended by Pocock et al. (2002) despite its lack of statistical power. Fourthly, when interpreting the
result of subgroup analysis in a trial report, the claim that the differential effect exists in a small subgroup is not advocated.

Wang et al. (2007) pointed out that when conducting subgroup analysis for each baseline characteristics, multiple statistical testings are carried out, which will obviously increase the probability of false positive finding. Bonferroni adjustment could be used to address this problem. Or we can note the number of nominally significant tests that would be expected to happen by chance instead of adjusting the test itself.

In Sivaganesan et al. (2011), subgroups were defined by a single covariate, and subgroup effects were identified by the presence of heterogeneity in treatment effect across the different subgroups. For each covariate, various configurations of the treatment effect across the subgroups were used to define models for the response variable, forming a separate model space for each covariate. A zero-inflated Polya Urn scheme was used to construct prior probabilities of the models such that small prior probabilities are assigned to small subgroups. Bayesian model selection was then used to choose among the models using pre-determined thresholds for the posterior probabilities, and to find subgroups with heterogeneous effects.

In chapter 2, we will extend the approach in Sivaganesan et al. (2011) to ANOVA based models with subgroups defined by 2 or more covariates, assuming that the covariates are specified a priori by the investigators, and are not confounder. Please see Section 2.1 for more literature review and detail of our method.
1.2 Causal Inference with Observational Data

Causal inference studies the causation based on the condition that occurs accompanying the effect. Ideally, a randomized control trial could be carried out to evaluate the causal effect. However, due to ethical, logistic or cost reason, many studies cannot answer causal questions with randomized experiment. With observational data or non-experimental data, the differences of the patient characteristics between treatment group and control group are no longer random, which will lead to biased causal estimate.

When examine causal effect of a treatment and control group conditions, each patient has two potential outcomes, Rubin (1973, 1978). Depending on which group he/she is assigned to, his/her potential outcome in the other group could be considered as a missing value. A widely used method for causal inference with observational data is matching, which tries to construct a comparison group that is similar to the treatment group, such that the distributions of covariates in treated and control group are the same. Therefore, matching can eliminate confounding bias. Stuart (2010) provided a summary of the matching methods and guidance on their use. The major issue with matching is that it discards data that it could not find a match from points in the other group. Subsequently, it causes serious concern over the change of population definition and whether the study conclusion may be generalizable.

Other methods for causal inference include propensity score (PS) method and regression
model of outcome. The propensity score is the (estimated) probability of a subject being assigned to the treatment group. There are many different ways to incorporate PS when estimating the average treatment effect, such as a linear model with ps as one of the covariates, inverse probability of treatment weight and propensity score matching. The accuracy of PS is important for obtaining unbiased estimate when using those methods alone. The average treatment effect is estimated with the average difference of the (potential) outcomes in the treatment group and control group.

Bayesian regression with PS adjustment, Bayesian additive regression tree (BART) are Bayesian nonparametric methods used for causal inference. Hill (2012) advocated the use of BART for its simplicity to use and "requiring less guess work in model fitting". The Bayesian nonparametric modeling, including BART, can be used to evaluate the treatment effect by estimating the missing potential outcome(s) given certain covariates values and calculate the difference between two potential outcomes. Bayesian nonparametric method also can be used to estimate the heterogeneous treatment effect (HTE). One of the advantages of using Bayesian methodology is that the credible interval of treatment effect could be obtained naturally with all the uncertainty considered. Roy et al. (2017) proposed Bayesian nonparametric approach to causal inference in the point treatment setting using enriched Dirichlet proces.

Another Bayesian nonparametric method, Gaussian process, is widely used in spatial analysis and machine learning. Gaussian process assumes the responses occur in a continuous
domain and every finite collection of these points correspond to random variables following a multivariate normal distribution, whose covariance matrix is determined by a covariance function, a function of the observed covariates. We introduce a method using Gaussian process for estimating ATE and HTE in Chapter 3 and Chapter 4, respectively. The links between matching and Gaussian process approach are studied in Chapter 3.

We provide more information and literature review in Section 3.1 for estimating the average treatment effect (ATE) with observational data and in Section 4.1 for estimating the heterogeneous treatment effect (HTE) for patients with different baseline characteristics.

1.3 Organization of this Dissertation

In this dissertation, we focus on studying the treatment effect with data from randomized control experiment and observational data. Each section of this dissertation is organized as independent as possible from the rest so that readers could skip to the contents that are the most interesting to them if necessary.

Chapter 2 is about subgroup analysis for two or more covariates. In this chapter, we propose a Bayesian approach to subgroup analysis using a collection of ANOVA models with multiple covariates, extending an earlier work of Sivaganesan et al. (2011). In Chapter 3, we use Gaussian process to evaluate the average treatment effect with observational data. Bayesian methodology is utilized to get the uncertainty of the estimated average treatment...
effect and impose the prior preference. The links between matching and the proposed method are studied. Chapter 4 studies the heterogeneous treatment effect with observational data using two conditionally independent Gaussian processes.

Each chapter contains introduction for the background and literature review. The proposed approaches, prior for parameters and posterior inference are then followed by simulation (numerical) studies. We illustrate the proposed methods using real data before the discussions ending the chapters.
Chapter 2

A Bayesian subgroup analysis using collections of ANOVA models

We develop a Bayesian approach to subgroup analysis using ANOVA models with multiple covariates, extending an earlier work. We assume a two-arm clinical trial with normally distributed response variable. We also assume that the covariates for subgroup finding are categorical and are a priori specified, and parsimonious easy-to-interpret subgroups are preferable. We represent the subgroups of interest by a collection of models and use a model selection approach to finding subgroups with heterogeneous effects. We develop suitable priors for the model space and use an objective Bayesian approach that yields multiplicity adjusted posterior probabilities for the models. We use a structured algorithm based on the posterior probabilities of the models to determine which subgroup effects to report.
Frequentist operating characteristics of the approach are evaluated using simulation. While our approach is applicable in more general cases, we mainly focus on the $2 \times 2$ case of 2 covariates each at 2 levels for ease of presentation. The approach is illustrated using a real data example.

2.1 Introduction

While the primary goal of a randomized clinical trial is to assess the efficacy of a treatment for all subjects in the study population, investigators are often interested in finding subgroups where the efficacy is different, for example, where it is non-existent or elevated. As a result, subgroup analysis has been recognized as important and has received increased attention. In particular, authors have pointed out certain pitfalls and challenges in the use of subgroup analysis, such as a tendency to do unplanned and unlimited number of subgroup testing and the need to adjust for multiplicity. Guidelines have been for subgroup analysis and subgroup reporting by Pocock et al. (2002), Rothwell (2005), and Wang et al. (2007).

One of the early approaches to subgroup analysis using Bayesian paradigm was due to Dixon and Simon (1991), and later, Simon (2002), who used hierarchical priors for the treatment by covariate interaction coefficients and provided shrinkage estimates. Later, Hodges et al. (2007) and Jones et al. (2011) extended the approaches by considering more general random effects and provided more flexible shrinkage estimates for interaction coefficients.
Schnell et al. (2015) propose an approach that could be characterized as constructing a credible interval for the desired subgroups. The approach defines an inclusive and an exclusive subset of the covariate space, with posterior probability greater than or equals $1 - \alpha$ that the inclusive set contains all covariate values with differential treatment effect and similarly for the exclusive set to contain only covariates with differential treatment effect. Sivaganesan et al. (2017) used Bayesian additive regression trees to model the response and identified subgroups by maximizing posterior expected utility defined by a candidate subgroup and the predicted response for that subgroup.

Among the papers that use Bayesian model selection for subgroup analysis are Sivaganesan et al. (2011); Mueller et al. (2010); Laud et al. (2013) and Berger et al. (2014). Sivaganesan et al. (2011), Mueller et al. (2010) Laud et al. (2013) used a Bayesian model selection approach combined with a decision theoretic justification. They addressed the question of whether treatment effect differed across the subgroups defined by a set of predetermined covariates, taken one at a time. Subgroups were defined by the different values of a covariate, and subgroup effects were identified by the presence of heterogeneity in treatment effect across the different subgroups. For each covariate, various configurations of the treatment effect across the subgroups were used to define models for the response variable, forming a separate model space for each covariate. A zero-inflated Polya Urn scheme was used to construct prior probabilities of the models. Bayesian model selection was then used to choose among the models using pre-determined thresholds for the posterior probabilities,
and to find subgroups with heterogeneous effects. In an interesting approach using model selection, Berger et al. (2014) used a tree-based method to enumerate regression models with dichotomous covariates to represent possible subgroup effects. They reported posterior probabilities of the existence of subgroup effects as well as patient-specific probabilities and estimates for the treatment effects. In another interesting Bayesian approach useful in adaptive clinical trials, Xu et al. (2014) used a tree-based model and developed a subgroup-based adaptive (SUBA) design which allows for adaptive discovery of patient subgroups and allocation of patients to treatments, simultaneously.

In this chapter, we extend the approach used in Sivaganesan et al. (2011) to ANOVA-based models with 2 or more subgroup defining covariates. As recommended in the literature, e.g., Rothwell (2005), we assume that the covariates for which subgroup analysis of interest are specified a priori by the investigators. We also assume that these covariates are not confounders in the sense that they do not influence treatment assignment. This is a crucial assumption when doing subgroup analysis, and is a reasonable assumption in our context of randomized trials. In the interest of parsimony, sub-population size, and ease of reporting, we further assume that subgroup effects representing main treatment effects are of primary interest for reporting. These are represented by treatment-covariate interactions. Subgroup effects relating to interaction effects (between treatment and two or more covariates) are only of subsidiary interest. On this basis, we develop a method that aims to first determine if there is sufficiently more evidence to report a subgroup effect, as compared to reporting
an overall (constant) treatment effect or no treatment effect, for the whole study population. If it is the former, our approach proceeds to find a suitable parsimonious subgroup, among those specified a priori, for which there is evidence of heterogeneous treatment effect. We enumerate a collection of models to represent different subgroup effects, and find their posterior probabilities using a prior distribution on the model space developed as in traditional variable selection, and using a set of objective priors for the model specific parameters. We then use an algorithm based on a set of pre-set thresholds for the posterior probabilities to decide whether to report a subgroup effect and if so which one(s), or to report no subgroup effect and conclude either an overall treatment effect or no treatment effect. As is common in Bayesian approaches to clinical trials, we investigate frequentist operating characteristics of the proposed algorithm through simulation, and use it to set threshold values that correspond to a set of desired error rates. We illustrate our approach with data from a clinical trial. While our proposed approach can be easily adapted to other common types of outcomes and to the case of multiple covariates each with multiple levels, for the purpose of illustration, we assume a normally distributed outcome and focus mainly on the case of 2 covariates each at 2 levels.

The chapter is organized as follows. In Section 2, we give the sampling model and the parametrization used, introduce an index notation for the models, and specify the model space relating to the subgroups of interest. In Section 3, we focus on the $2 \times 2$ case providing further details of the parametrization, the index representation, and the model space. In
Section 4, we specify the prior probabilities of the models and the prior distribution of the model-specific parameters, and give the formulas for posterior probabilities of the models.

In Section 5, we give a stepwise algorithm to decide which subgroup effects, if any, to report, and show, as in Mueller et al. (2010), that the stepwise algorithm is an approximate Bayes rule corresponding to a specified utility function. In Section 6, we define certain frequentist operating characteristics of the stepwise algorithm, calculate them using simulation, and apply the proposed approach to a reported trial. We end with a discussion in Section 7.

2.2 ANOVA Model, Subgroup Effects and Model Space

We consider a two-arm randomized clinical trial comparing the means of a response under a treatment and control. Assume that the outcome is a continuous variable observed on a random sample of $N$ subjects, of which $N_0$ are assigned randomly to the control, $N_1 = N - N_0$ to the treatment. We use $y_{1j}, j = 1, 2, ..., N_1$ and $y_{0j}, j = 1, 2, ..., N_0$ to denote the observations under the treatment and control, respectively. We assume that $y_{ij}$'s are independent and normally distributed.

Suppose that $x_1$ and $x_2$ are two categorical covariates of a priori interest for subgroup analysis, among those for which measurements were recorded. If needed, continuous covariates could be re-coded as categorical by using appropriate cutoffs. Also, let $k = 1, ..., K$ and $\ell = 1, ..., L$ be the levels of the covariates $x_1$ and $x_2$, respectively. The subjects in the study
population can thus be classified into $K \times L$ subgroups, $S_{kl}$’s, based the two covariates.

We consider the following ANOVA models for the response variable,

\[ y_{0j} | x_1 = k, x_2 = \ell \sim N(\mu_{k\ell}, \sigma^2_{k\ell(0)}), \quad j = 1, ..., n_{k\ell(0)} \]  
\[ y_{1j} | x_1 = k, x_2 = \ell \sim N(\mu_{k\ell} + \theta_{k\ell}, \sigma^2_{k\ell(1)}), \quad j = 1, ..., n_{k\ell(1)} \]

where, for $k = 1, ..., K; \ell = 1, ...L; \mu_{k\ell}$ is the mean of the response under control, $\theta_{k\ell}$ is the subgroup-specific treatment effect for subgroup $S_{k\ell}$; and, $n_{k\ell(0)}$ and $n_{k\ell(1)}$ are the sample sizes, $\sigma^2_{k\ell(0)}$ and $\sigma^2_{k\ell(1)}$ are the variances under treatment and control, respectively, for subgroup $S_{k\ell}$. Our focus in this chapter is on the subgroup-specific treatment effects as measured by the differences in the means of the response, and not on their variances. We therefore use the ANOVA model only for the treatment effect, and regarded the variances as nuisance parameters.

Thus, we write

\[ \theta_{kl} = \delta_k + \eta_{\ell} + \lambda_{k\ell} \quad k = 1, 2, ..., K \quad \ell = 1, 2, ..., L, \]

where $\theta_{kl}$ is split into main (subgroup) effects, $\delta_k$ and $\eta_{\ell}$, for the subgroups defined by the $k$-th level of $X_1$ and $\ell$-th level of $X_2$, respectively, and an interaction (subgroup) effect $\lambda_{k\ell}$ for the subgroup defined by $k$th level of $X_1$ and $\ell$th level of $X_2$. While this is an over-parametrized version of the model with unidentifiable parameters, we find that it is more convenient for defining and interpreting the various subgroup effects of interest, and for specifying prior parameters.
distributions for the subgroup effect parameters. We will address the nonidentifiability by considering only one of the many equivalent parametrizations of a model. More details are given in Section 2.3.1.

We also considered the case where the variances are the same under different cells and under different treatment and control groups. We give the corresponding numerical result for the real data example for this case when discussing the example. The model and prior specification, and the computational details for this case are given in supplementary material.

### 2.2.1 Subgroup Effects and Models

Let \( \theta = (\theta_{11}, \theta_{12}, ..., \theta_{KL})' \), \( \delta = (\delta_1, \delta_2, ..., \delta_K)' \), \( \eta = (\eta_1, \eta_2, ..., \eta_L)' \) and \( \lambda = (\lambda_{11}, \lambda_{12}, ..., \lambda_{KL})' \) denote the vectors of parameters in (2.3). When \( \theta = 0 \), there is no treatment effect in any subgroup; we call the corresponding model for \( Y \), the overall null model, and denote it by \( M_0 \). Similarly, when \( \theta = 1\phi \), for some \( \phi \neq 0 \), there is a constant treatment effect across the whole study population; we call the corresponding model the overall treatment effect model, and denote it by \( M_1 \). These two models represent what we refer to as no-subgroup effects models.

When \( \theta_{k\ell}'s \) differ among the different subgroups, e.g., \( \theta_{11} = \theta_{12} \neq 0, \theta_{21} = \theta_{22} = 0 \) with \( K = L = 2 \), we call the corresponding model a subgroup effect model.

In the interest of parsimony, and for ease of interpretation, we assume that the subgroup
effects involving the main effects are preferable for reporting, and those involving further interactions are of only subsidiary interest. We further assume that subgroups involving at most one interaction are of interest. This assumption can be relaxed to allow more than one interaction, especially when there are more than two variables. We however do not pursue it here.

Thus we specify subsets of models corresponding to the subgroup reports that we will introduce later, as:

\[
S_1 = \{ (\delta, \eta, \lambda) : \delta \neq \phi \mathbf{1} \text{ and } \eta = 0 \}, \quad (2.4)
\]

\[
S_2 = \{ (\delta, \eta, \lambda) : \delta = 0 \text{ and } \eta \neq \phi \mathbf{1} \}, \quad (2.5)
\]

\[
S_{12} = \{ (\delta, \eta, \lambda) : \delta \neq \phi \mathbf{1} \text{ and } \eta \neq \phi \mathbf{1} \}, \quad (2.6)
\]

\[
S_\lambda = \{ (\delta, \eta, \lambda) : \delta = \phi \mathbf{1}, \ \eta = 0, \text{ and exactly one of } \lambda_{k\ell}'s \neq 0 \}, \quad (2.7)
\]

where \( \phi \) is a constant (which can be zero or non-zero) and \( \mathbf{1} \) is a vector of 1’s, and at most one of \( \lambda_{k\ell} \) is assumed non-zero in (2.4-2.6). These subsets can be regarded as a collection of ANOVA models, all of which are special cases of the model in (2.2) and (2.3).

Thus, the overall model space of interest for subgroup analysis is

\[
\mathcal{M} = \{ M_0, M_1 \} \cup S_1 \cup S_2 \cup S_{12} \cup S_\lambda. \quad (2.8)
\]

We will set up the subgroup analysis as a model selection problem, selecting models from these five subsets of the overall model space, and then adding flexibility to report more
than one subgroup. We therefore refer to $S_1, S_2, S_{12}, S_\lambda$ as model spaces or equivalently as subgroup reports, depending on the context. Specifically, (i) the subgroup reports in $S_1, S_2$ and $S_{12}$ refer only to the main effects for $X_1$ and $X_2$, without constraints on $\lambda_{kl}$, except at most one of its elements can be non-zero; and (ii) the subgroup reports in $S_\lambda$ refer to interaction effect only at one of the $K \times L$ levels, with no main effects ($\delta = 0$ and $\eta = 0$) or a common main effect of equal size (e.g., $\delta = 1\phi$ and $\eta = 0$) - the other case with $\delta = 0$ and $\eta = 1\phi$ is equivalent.

2.2.2 An index representation of the models

The subgroup effects of interest and the associated models described above involve specifying which of the subgroups corresponding to a covariate have different treatment effects. For example, a subgroup effect identified by a model in $S_1$ would be specified by which components of $\delta$ are zero, and which of the nonzero components, if any, are different. A convenient way to express such a configuration is to use vectors of non-negative integers, called index vectors, as in Sivaganesan et al. (2011). Here, we use index vectors $\gamma_1 = (\gamma_1^1, \gamma_1^2, ..., \gamma_1^K)$ and $\gamma_2 = (\gamma_2^1, \gamma_2^2, ..., \gamma_2^L)$ to represent the configuration of $\delta$ and $\eta$, respectively, and use $\gamma = (\gamma_1, \gamma_2)$ to denote the combined index.

Specifically, $\gamma_i^s = 0$ indicates no treatment effect in the subgroup defined by the $s$-th level of the $i$-th covariate. Any two identical non-zero integers, $\gamma_i^{s_1} = \gamma_i^{s_2} \neq 0$, indicate that there
is a non-zero treatment effect of equal size in the subgroups defined by the \(s_1\)-th and \(s_2\)-th levels of the \(i\)-th covariate. The non-zero integers in \(\gamma_i\) are non-descending by the order of their appearance in the index vector. Therefore, the appearance of a new integer value in the \(s\)-th entry of \(\gamma_1\) (or \(\gamma_2\)) indicates a subgroup effect that is different in magnitude from those in the previous \(s - 1\) subgroups. For example, for two covariates \(X_1\) and \(X_2\) each with 3 levels, the index vectors \(\gamma_1 = (0, 1, 1)\) and \(\gamma_2 = (1, 0, 2)\) describe a model for which \(\delta_1 = 0\), \(\delta_2 = \delta_3 \neq 0\) and \(\eta_2 = 0\), \(0 \neq \eta_1 \neq \eta_3 \neq 0\). Similar to \(\gamma_1\) and \(\gamma_2\), we use \(\beta = (\beta_{11}, \beta_{12}, ..., \beta_{KL})\) as the index vector for \(\lambda = (\lambda_{11}, \lambda_{12}, ..., \lambda_{KL})\).

In the following, we use the notation \(M_{\gamma, \beta}\) to denote the subgroup effect models. By setting \(\beta = 0\) and removing it from the notation, we also get the index representation \(\gamma = (\gamma_1, \gamma_2)\) for the configuration of the subgroup effect vectors \((\delta, \eta)\) corresponding to the models without interaction, which we denote by \(M_\gamma\), where \(\gamma = (\gamma_1, \gamma_2)\).

### 2.3 The \(2 \times 2\) Case

In practice, it would be commonly desirable to limit the number of levels of each covariate, and henceforth, we will assume each covariate is at two levels. Although our approach can be extended to cover more general cases, for illustrative purpose, we focus on the case of two covariates. In this case, the means of \(Y_0\) and \(Y_1\) for the subgroups can be written as in Table 2.1.
Table 2.1: Means of $Y_0$ and $Y_1$ within each subgroup

<table>
<thead>
<tr>
<th></th>
<th>$X_2 = 1$</th>
<th>$X_2 = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 = 1$</td>
<td>$E(Y_0)$</td>
<td>$E(Y_1)$</td>
</tr>
<tr>
<td></td>
<td>$\mu_{11}$</td>
<td>$\mu_{11} + \delta_1 + \eta_1 + \lambda_{11}$</td>
</tr>
<tr>
<td>$X_1 = 2$</td>
<td>$\mu_{21}$</td>
<td>$\mu_{21} + \delta_2 + \eta_1 + \lambda_{21}$</td>
</tr>
</tbody>
</table>

Here, the first two indexing vectors of subgroup effects are $\gamma_1 = (\gamma_1^1, \gamma_1^2)$ and $\gamma_2 = (\gamma_2^1, \gamma_2^2)$, each of which can take one of the five values: $(0,0), (1,0), (0,1), (1,2)$ and $(1,1)$. Since at most one interaction is assumed to be non-zero, $\|\beta\| = 0$ or $1$, depending on whether interaction is included or not, and the vector $\beta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})$ can take one of the five values: $(0,0,0,0), (1,0,0,0), (0,1,0,0), (0,0,1,0)$ and $(0,0,0,1)$.

### 2.3.1 The Model Space and the Index Space

Letting $\mathcal{W}_i$ be the index space associated with the main effects of $X_i$ alone, for $i = 1, 2$, and $\mathcal{W}_\lambda$ be the index space associated with interaction alone, we can consider the full index space as the product space $\mathcal{W}_1 \times \mathcal{W}_2 \times \mathcal{W}_\lambda$.

Specifically, for the $2 \times 2$ case, there are $5 \times 5 \times 5 = 125$ distinct model indices $(\gamma, \beta)$. However, due to the non-identifiability of the parametrization in Table 2.1, some of these indices actually represent the same subgroup effect model and correspond to a unique likelihood
function. For instance, the following three sets of indices,

$$
\gamma_1 = (0, 0), \gamma_2 = (1, 1), \beta = 0,
$$

$$
\gamma_1 = (1, 1), \gamma_2 = (0, 0), \beta = 0, \text{ and}
$$

$$
\gamma_1 = (1, 1), \gamma_2 = (1, 1), \beta = 0,
$$

all correspond to

$$
\theta_{11} = \theta_{12} = \theta_{21} = \theta_{22} \neq 0, \quad (2.9)
$$

which is the overall treatment effect model. Hence, we define the model indices which give
the same subgroup effect model as "equivalent".

In order to identify the equivalent indices, we only need to check if the columns cor-
responding to the subgroup effects generate the same column space. For example, the four
indices in the following set

$$
\{(1, 2), (0, 0), 0\}, \{(1, 2), (1, 1), 0\}, \{(1, 0), (1, 1), 0\}, \{(0, 1), (1, 1), 0\}\}
$$

all have the same column space, and represent the (same) subgroup effect model: different
treatment effects for the levels of $X_1$, no differential effects for the levels of $X_2$ and no
interaction effect. The corresponding design matrices for their treatment effect parts are

\[
\begin{bmatrix}
\delta_1 & \delta_2 \\
\theta_{11} & \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \\
\theta_{12} & \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}, \\
\theta_{21} & \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \\
\theta_{22} & \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}
\end{bmatrix}
\begin{bmatrix}
\delta_1 \\
\delta_2 \\
\eta_1 \\
\eta_1
\end{bmatrix}
\]

All of the above have the same column space, and hence represent the same subgroup effect model. Thus we group these four into one model to represent the same subgroup effect. In order to get a unique index representation of the subgroup effect models, we divide the index space into partitions of equivalent indices, and choose a unique index from each partition to represent the model corresponding to that partition. For this, we choose the index that has the most zeros, i.e., most parsimonious one. For example, we will use \( M_{((1,2),(0,0),0)} \) to represent the model in the previous example. The resulting index representation of the models, which is a subset of the full product index space, has a one-to-one correspondence with the model space, \( \mathcal{M} \), and is therefore used to represent \( \mathcal{M} \).

When interaction is not included, i.e., \( \beta = 0 \), the resulting sub-model space, say \( \mathcal{M}_{NI} \subset \mathcal{M} \), in the \( 2 \times 2 \) case has 13 models. The set of equivalent indices, and the models they represent with their respective unique index given in the the subscript, are given in Table 2.2.
Table 2.2: Partitions of equivalent model indices and the models they represent

<table>
<thead>
<tr>
<th>( \gamma = (\gamma_1, \gamma_2) )</th>
<th>( M \in \mathcal{M}_{NI} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>{((0,0),(0,0))}</td>
<td>( M_{(0,0),(0,0)} )</td>
</tr>
<tr>
<td>{((0,0),(1,1)), ((1,1),(0,0)), ((1,1),(1,1))}</td>
<td>( M_{(1,1),(0,0)} )</td>
</tr>
<tr>
<td>{((1,2),(0,0)), ((1,2),(1,1)), ((1,0),(1,1)), ((0,1),(1,1)), ((0,1),(1,1))}</td>
<td>( M_{(1,2),(0,0)} )</td>
</tr>
<tr>
<td>{((0,0),(1,2)), ((1,1),(1,2)), ((1,1),(1,0)), ((1,1),(0,1))}</td>
<td>( M_{(0,0),(1,2)} )</td>
</tr>
<tr>
<td>{((1,0),(0,0))}</td>
<td>( M_{(1,0),(0,0)} )</td>
</tr>
<tr>
<td>{((0,0),(1,0))}</td>
<td>( M_{(0,0),(1,0)} )</td>
</tr>
<tr>
<td>{((0,1),(0,0))}</td>
<td>( M_{(0,1),(0,0)} )</td>
</tr>
<tr>
<td>{((0,0),(0,1))}</td>
<td>( M_{(0,0),(0,1)} )</td>
</tr>
<tr>
<td>{((1,0),(1,0))}</td>
<td>( M_{(1,0),(1,0)} )</td>
</tr>
<tr>
<td>{((0,1),(1,0))}</td>
<td>( M_{(0,1),(1,0)} )</td>
</tr>
<tr>
<td>{((1,0),(0,1))}</td>
<td>( M_{(1,0),(0,1)} )</td>
</tr>
<tr>
<td>{((0,1),(0,1))}</td>
<td>( M_{(0,1),(0,1)} )</td>
</tr>
<tr>
<td>{((0,1),(1,2)), ((1,0),(1,2)), ((1,2),(0,1)), ((1,2),(1,0)), ((1,2),(1,2))}</td>
<td>( M_{(0,1),(1,2)} )</td>
</tr>
</tbody>
</table>

**Note:** The table gives the partitions of equivalent model indices and the models they represent for the \( 2 \times 2 \) case when interaction is not included. The first column gives the partitions of equivalent indices in each row and the second column gives the model they represent with the corresponding unique index as subscript. Thus, \( M_{(0,0),(0,0)} \) and \( M_{(1,1),(0,0)} \) here represent the overall null model \( M_0 \) and overall effect model \( M_1 \), respectively.

The following result, the proof of which is omitted, gives the size of (i.e., number of models in) \( \mathcal{M}_{NI} \) in terms of \( K \). For example, for \( K = 1, 2 \) and 3, they are 5, 13 and 35, respectively.

**Result:**

In a \( 2^K \) design, the size of the model space \( \mathcal{M}_{NI} \) when no interaction is included is given by

\[
2 + \sum_{i=1}^{K} \binom{K}{i} (1 + 2^i)
\]

When the interaction is included, the unique index representation of a model is obtained by
adding an interaction index \( \beta \) to the unique model index from the no-interaction case. The resulting model space \( \mathcal{M} \) consists of 46 models for the \( 2 \times 2 \) case, and are listed in Table 2.3.

Table 2.3: Models and their unique indices when interaction is included

<table>
<thead>
<tr>
<th>( M \in \mathcal{M} )</th>
<th>( \beta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_0 )</td>
<td>( \beta = 0 )</td>
</tr>
<tr>
<td>( M_1 )</td>
<td>( \beta = 0 )</td>
</tr>
<tr>
<td>( M_{(1,2),(0,0),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{12} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(0,0),(1,2),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{21} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(1,0),(0,0),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{12} = 1 ) or ( \beta_{21} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(0,0),(1,0),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{12} = 1 ) or ( \beta_{21} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(0,1),(0,0),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{11} = 1 ) or ( \beta_{12} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(0,1),(0,0),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{11} = 1 ) or ( \beta_{21} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(0,1),(1,0),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{12} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(1,0),(0,1),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{21} = 1 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(0,1),(1,1),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{11} = 1 ), or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(1,0),(1,2),\beta} )</td>
<td>( \beta = 0 ) or ( \beta_{22} = 1 )</td>
</tr>
<tr>
<td>( M_{(1,1),(0,0),\beta} )</td>
<td>one of ( \beta_{ij} ) equals one</td>
</tr>
<tr>
<td>( M_{(0,0),(0,0),\beta} )</td>
<td>one of ( \beta_{ij} ) equals one</td>
</tr>
</tbody>
</table>

*Note:* The table lists the models by their unique model indices when interaction is included. For each model, the model index in the first column consists of the unique representation \( (\gamma_1, \gamma_2) \) as in the no-interaction case, along with the interaction index vector \( \beta \) which is given in the second column.

We thus let \( \mathcal{W} = \{(\gamma, \beta) : M_{\gamma, \beta} \in \mathcal{M}\} \) be the index space consisting of the unique index representations of the models in Table 2.3. We also let \( \mathcal{W}_\gamma = \{\gamma = (\gamma_1, \gamma_2) : M_{\gamma, \beta} \in \mathcal{M}\} \), and \( \mathcal{M}_\gamma = \{M_{\gamma} : \gamma \in \mathcal{W}_\gamma\} \), which consist of the indices and the models that leave the interaction unspecified, and they correspond to the subgroup effects specified in (2.4) to (2.6). We will use \( \mathcal{W} \) and \( \mathcal{W}_\gamma \) to define the prior distribution on the model space in the next
2.3.2 Extension to more than 2 covariates

Here, we outline the steps that would be involved when there are more than two variables. For brevity, we use the case of three variables. Suppose we have covariates $X_1, X_1, X_3$ each at two or higher number of levels. Letting $\theta_{k\ell m}$ be the subgroup effect corresponding to $X_1 = k, X_2 = \ell, X_3 = m$ where the values of $k, \ell$ and $m$ span the levels of the covariates $X_1, X_2$ and $X_3$, respectively. Using similar notation as in (2.3), we can write

$$\theta_{k\ell m} = \delta_k + \eta_\ell + \psi_m + \sum \lambda^{(2)}_{uv} + \lambda^{(3)}_{k\ell m}, \quad (2.11)$$

where the summation is over $u, v = k, \ell, m$ with $u \neq v$, and $\lambda^{(2)}$ and $\lambda^{(3)}$ represent the second and third order interactions. As indicated earlier, we assume that the investigators can a priori specify a few variables and a few of the interactions among them as plausible and of interest for subgroup identification. This would be used to reduce the number of terms in (2.11) and hence the model space representing the subgroups. We can resolve the unidentifiability issue using the column space approach described earlier to determine the unique models representing the subgroups, and use index notation $(\gamma_1, \gamma_2, \gamma_3, \beta)$ to denote the subgroup models, and the model space of interest.

The coding needed for the computation with three (or more) covariates would involve separate enumeration of the models in each of the subsets of models (analogous to those
in Section 2.2.1). Besides the null effect model and overall effect model, subsets containing exactly one, two and three main effects are considered in separated subsets. And models without main effect but with two-way or three-way interaction effect are contained in other subsets. For each subset, we would list the models of interest using the subgroup effect parameters representing the different configurations (e.g., \( \delta \) for \( S_1 \)), and identify those with distinct column spaces using a stand-alone function, which could check the rank of the design matrix to see if a model is duplicate. An easy way to enumerate all the possible models without interaction is listing all the possible subgroup models for each covariate, then take the cartesian product. To eliminate the duplicate models, we would start with the models having the fewest (non-zero) parameters and delete all other models with equivalent column space to achieve parsimonious representations. Then the interaction terms could be added in based on these models. And we again delete all the duplicates introduced by adding the interaction terms. Once the distinct models are selected, they can be represented using the index notation based on which components of the vector parameters (e.g., \( \delta \) and \( \gamma \)'s for models in \( S_1 \)) are set to zero and which are distinct, for each model.

An easy way to enumerate all the possible models without interaction is listing all the possible subgroup models for each covariate, then take the cartesian product.

The prior distribution over the model space and the prior distributions for the model specific parameters can be specified as already described in the following sections for the general setting. The derivation and computing of the marginal likelihoods described in
Appendix A can be extended to the cases with three or more covariates. However, a more parsimonious assumption about the variances or a more efficient scheme would be needed for large number of covariates.

When the number of variables is large, the complexity of notation and book-keeping would be cumbersome. But, in such cases, it would also be natural, in view of ease of interpretation and parsimony, to reduce the number of levels to two, and focus on a few interactions of a priori interest, and hence keeping the model enumeration and representation within a scope that is amenable for computing posterior probabilities. Use of a suitable MCMC algorithm to calculate posterior probabilities of the models when the model space is not easy to enumerate would also help obtain approximate posterior probabilities.

2.4 Prior distributions and the Posterior Probabilities

2.4.1 Prior distribution on the model space

In this section we specify a prior distribution for the models in $M_{\gamma,\beta} \in \mathcal{M}$. We do this by specifying a prior on the equivalent index space $W$, and do it in three steps. In the first step, a prior distribution is assigned to $\gamma = (\gamma_1, \gamma_2) \in W_{\gamma}$, in the second step a prior distribution is assigned to $\|\beta\| (=0 \text{ or } 1)$, given $\gamma$, and in the third step a prior is assigned to $M_{\gamma,\beta} \in \mathcal{M}$, given $\gamma$ and $\|\beta\|$. Specification of the prior in the first step is motivated by an
a priori preference for subgroups effects associated with single covariates, and it also helps with writing down the prior for subgroups with interaction. Here, we describe the prior for the general case of $K$ covariates, and later illustrate it for the $2 \times 2$ case.

**Step 1:**

We propose to assign prior probabilities to $(\gamma_1, ..., \gamma_K)$ based on the set of covariate(s), $S \subset \{X_1, .., X_K\}$, to which subgroup effects are assigned. For example, in the case of $K = 2$, the possibilities are: no subgroup effect (neither $X_1$ nor $X_2$ is selected for subgroup effect), with $S = \emptyset$; one variable (either $X_1$ or $X_2$) is selected for subgroup effect, with $S = \{X_1\}$ or $S = \{X_2\}$; or, both selected, with $S = \{X_1, X_2\}$.

Following the standard approach to assigning probabilities in the context of variable selection, we first assign probability of $S$ given $p$ as follows,

$$P(S|p) = p^s(1 - p)^{K-s}; \quad S \in S, \tag{2.12}$$

where $p = P(a \text{ variable is included for defining subgroup})$ and $s = |S|$ is the the cardinality of $S$. In the $2 \times 2$ case, we have $K = 2$ and $s = 0, 1, \text{ or } 2$, and $S$ is as in ((2.4)-(2.6)). Conditional on $S$, we specify the probability of the index vector corresponding to each variable as follows.

We first let $q = P(\text{subgroup effect is not present at any specific level of a covariate})$. When $S = \emptyset$, there are only two options, which we denote by $w_0 \ w_1$; the first consisting of all zeros representing the overall null model, and the second consisting of all ones for one variable and
zeroes for all other., representing the overall effect model. For example, when \( K = 2 \), they are \( \mathbf{w}_0 = ((0,0),(0,0)) \) and \( \mathbf{w}_1 = ((1,1),(0,0)) \). Since this is determined solely by whether first indicator in the index is 0 or 1, we let

\[
P (\gamma = \mathbf{w}_0 | q, S = \emptyset) = q, \tag{2.13}
\]

\[
P (\gamma = \mathbf{w}_1 | q, S = \emptyset) = 1 - q. \tag{2.14}
\]

For \( \gamma \in \mathcal{W}_\gamma \setminus \{\mathbf{w}_0, \mathbf{w}_1\} \), let \( a_i \) denote the number of levels at which subgroup effect is not present for covariate \( X_i \in S \), i.e., \( a_i \) is the number of zeroes in the index \( \gamma_i \). For example, for the index \( ((1,0),(0,0)) \), \( a_1 = 1 \).

Given \( S \), the conditional probabilities of the indices in \( \mathcal{W}_\gamma \setminus \{\mathbf{w}_0, \mathbf{w}_1\} \) are defined based on \( a_i \)'s as follows:

\[
P(\gamma | S, S \neq \emptyset, q) = c_S(q) \prod_{i : X_i \in S} q^{a_i} (1 - q)^{2 - a_i}, \tag{2.15}
\]

where, \( c_S(q) \) is the normalizing constant specific to \( S \). For, example, for \( \gamma = ((1,1),(0,0)) \), \( S = \{X_1\}, s = 1, a_1 = 1 \) and \( S = S_1 \) as in (2.4). The two other models in \( S_1 \) have \( a_1 = 1 \) and \( a_2 = 2 \). So, \( 1/c_S(q) = 2q(1 - q) + (1 - q)^2 = (1 - q^2) \), and

\[
P(\gamma = ((1,1),(0,0)) | S = \{X_1\}, q) = q/(1 + q). \tag{2.16}
\]

Noticing that the index \( \gamma \) corresponds to a unique \( S \), the conditional probability \( P(\gamma | p, q) \) is then given by

\[
P(\gamma | p, q) = P(\gamma | S, q) P(S | p) \text{ for } S \text{ that matches } \gamma. \tag{2.17}
\]
For the 2x2 case, \( P(\gamma|p, q) = P(M_\gamma|p, q) \) are given in Table 2.4.

**Step 2:**

Using the previously defined probability inclusion of a variable, \( p \), for the interaction term, we have

\[
P(\|\beta\| | \gamma, p) = \begin{cases} 
p & \text{if } \|\beta\| = 1, \\
1 - p & \text{if } \|\beta\| = 0.
\end{cases}
\] (2.18)

**Table 2.4: Prior distribution on the index space \( W_\gamma \)**

| Model index \( \gamma = (\gamma_1, \gamma_2) \) | \( P(\gamma|p, q) \) | \( P(\gamma) \) |
|----------------------------------------------|------------------|------------------|
| \((0, 0), (0, 0)\)                           | \((1 - p)^2q\)   | 0.16667          |
| \((1, 1), (0, 0)\)                           | \((1 - p)^2(1 - q)\) | 0.16667          |
| \((1, 2), (0, 0)\)                           | \(p(1 - p)(1 - q)/(1 + q)\) | 0.06438          |
| \((0, 0), (1, 2)\)                           | \(p(1 - p)(1 - q)/(1 + q)\) | 0.06438          |
| \((1, 0), (0, 0)\)                           | \(p(1 - p)q/(1 + q)\) | 0.05114          |
| \((0, 0), (1, 0)\)                           | \(p(1 - p)q/(1 + q)\) | 0.05114          |
| \((0, 0), (0, 1)\)                           | \(p(1 - p)q/(1 + q)\) | 0.05114          |
| \((0, 1), (0, 0)\)                           | \(p(1 - p)q/(1 + q)\) | 0.05114          |
| \((0, 1), (1, 1)\)                           | \(p^2q/(1 + 3q)\) | 0.05373          |
| \((1, 0), (1, 0)\)                           | \(p^2q/(1 + 3q)\) | 0.05373          |
| \((0, 1), (1, 0)\)                           | \(p^2q/(1 + 3q)\) | 0.05373          |
| \((1, 0), (0, 1)\)                           | \(p^2q/(1 + 3q)\) | 0.05373          |
| \((0, 1), (1, 2)\)                           | \(p^2(1 - q)/(1 + 3q)\) | 0.11839          |

**Note:** The table gives conditional prior probabilities for the unique model index \( \gamma \), or equivalently for the models in \( M_\gamma \), given \( p \) and \( q \), and the unconditional probabilities when \( p \) and \( q \) assigned independent uniform prior distributions over \( (0, 1) \). These are also the marginal probabilities obtained from the probabilities of the models \( M_{\gamma, \beta} \in M \), marginalized over \( \beta \).

**Step 3:**

29
Given $\gamma$ and $\|\beta\| = 1$, there can be more than one possible model, see Table 2.3. Thus, we assign the uniform prior probability distribution

$$P(\gamma, \beta | \gamma, \|\beta\| = 1) = 1/n_{\gamma}, \quad (2.19)$$

where $n_{\gamma} =$number of models with interaction, given $\gamma$.

As described at the beginning of this subsection, we now have

$$P(M_{\gamma,\beta} | p, q) = P((\gamma, \beta) | p, q) = P((\gamma, \beta) | \gamma, \|\beta\|) P(\|\beta\| | \gamma, p) P(\gamma | p, q) \quad (2.20)$$

For example, given $p$ and $q$, the overall null Model $M_0 = M_{0,0}$ is assigned probability $P(M_{0,0} | p, q) = (1 - p)^3 q$. Then the prior distribution over the model space $\mathcal{M}$ is

$$P(M_{\gamma,\beta}) = \int P(M_{\gamma,\beta} | p, q) P(p, q) dq dp, \quad (2.21)$$

where $P(p, q)$ is the prior distribution of $p$ and $q$. Prior probabilities of $M_{\gamma}$ can be obtained by summing over the values of $\beta$. By assigning a prior distribution for $(p, q)$, the resulting posterior probabilities of models are adjusted for multiplicity, see Scott and Berger (2006) and Scott and Berger (2010).

### 2.4.2 Prior distribution of the model-specific parameters

For the $2 \times 2$ ANOVA model, the unknown parameters are $\mu_0 = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})^T$, $\theta = (\theta_{11}, ..., \theta_{22})'$ and the vector of variances $\sigma^2 = (\sigma^2_{11(0)}, \sigma^2_{12(0)}, ..., \sigma^2_{22(1)})'$, where, as in (2.3), $\theta_{k\ell}$ can be written in terms of covariate specific subgroup effects $\psi = (\delta_1, \delta_2, \eta_1, \eta_2, \lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})^T$. 

30
Under a subgroup effect model by $M_j$, $M_j \in \mathcal{M}$, we can write $\theta_{k\ell} = Z_{k\ell}^t \psi_j$ where $\psi_j$ is a subgroup-specific subset of $\psi$, and write $\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})' = Z_j \psi_j$, $\psi_j = (\psi_{j1}, \psi_{j2}, ..., \psi_{jp})'$, where $\psi_j$ and the design matrix $Z_j$ depend on the model $M_j$.

Objective priors for the parameters in the context of model selection in normal linear model has been studied extensively, e.g., Liang et al. (2008), Maruyama and George (2011) and Bayarri et al. (2012). Following (Liang et al. (2008)), we use a mixture $g$-priors for the parameters, suitable for comparison of the models $M_0$ and $M_j$, as follows:

$$p(\mu_0, \sigma^2) = \prod \sigma_{k\ell(0)}^{-2} \sigma_{k\ell(1)}^{-2},$$

$$\psi_j | V, g \sim N(0, (Z_j^t V^{-1} Z_j)^{-1}gn),$$

$$p(g) = \frac{1}{(1 + g)^2}, g \geq 0,$$

where $V = diag(\sigma_{11}^2, \sigma_{12}^2, \sigma_{21}^2, \sigma_{22}^2)$ and $\sigma_{k\ell}^2 = \sigma_{k\ell(1)}^2/n_{k\ell}^{(1)} + \sigma_{k\ell(0)}^2/n_{k\ell}^{(0)}$.

### 2.4.3 Posterior Probabilities of the Models

Using $\bar{P}(M_j)$ to denote the posterior probability $P(M_j|y)$ of a model $M_j$, we have, for $M_j \in \mathcal{M}\backslash M_0$

$$\bar{P}(M_j) = \frac{P(y|M_j)P(M_j)}{\sum_{M \in \mathcal{M}} P(y|M)P(M)} = \frac{B_{j0} P_{j0}}{1 + \sum_{M_j' \in \mathcal{M}\backslash M_0} B_{j'0} P_{j0}}, \tag{2.22}$$

and

$$\bar{P}(M_0) = \frac{P(y|M_0)P(M_0)}{\sum_{M \in \mathcal{M}} P(y|M)P(M)} = \frac{1}{1 + \sum_{M_{j'} \in \mathcal{M}\backslash M_0} B_{j'0} P_{j'0}}, \tag{2.23}$$
where \( P_{j0} = P(M_j)/P(M_0) \) is the prior odds, \( B_{j0} \) is the Bayes factor for comparing \( M_j \) and \( M_0 \) given by \( B_{j0} = P(y|M_j)/P(y|M_0) \), and \( P(y|M_j) \) and \( P(y|M_0) \) are the marginal likelihoods under the models \( M_j \) and \( M_0 \), respectively, given by

\[
P(y|M_0) = \int p(y|M_0, \mu_0, \sigma^2)p(\mu_0, \sigma^2)d\mu_0d\sigma^2. \tag{2.24}
\]

and

\[
P(y|M_j) = \int p(y|M_j, \psi_j, \mu_0, \sigma^2)p(\psi_j|V, g)p(g)p(\mu_0, \sigma^2)dg \psi_j d\mu_0 d\sigma^2, \tag{2.25}
\]

According to (2.22), we only need \( B_{j0} \)’s and \( P_{j0} \)’s to compute the posterior probabilities of the models in \( M \). \( P_{j0} \) can be calculated using (2.21). Calculation of \( B_{j0} \)’s require the marginal likelihoods in (2.24) and (2.25). The details their computation are given in Appendix A.1.

### 2.4.4 Equal Variance

If researchers are confident that the variances are equal, i.e., \( \sigma^2_{11(0)} = \sigma^2_{12(0)} = \ldots = \sigma^2_{22(1)} \). It will greatly reduce the computational complexity if we use the objective priors in Maruyama and George (2011) and Bayarri et al. (2012), which have closed form Bayes factors. We used the mixture \( g \)-priors recommended in these two papers, applied them to the simulated data, and found both performed similarly in terms of error rates reported later.
By (2.1) and (2.2), the sampling distribution under \( M_0 \) and \( M_j \) are

\[
y|M_0, \mu_0, \sigma^2 \sim N(X_0 \mu_0, \sigma^2 I),
\]

\[
y|M_j, \psi_j, \mu_0, \sigma^2 \sim N(X_0 \mu_0 + Z_j \psi_j, \sigma^2 I),
\]

where \( X_0 \) and \( Z_j \) are parts of the design matrix corresponding to \( \mu_0 \) and \( \psi_j \), respectively.

Here, we report the results for the robust prior recommended in Bayarri et al. (2012). The robust prior for the subgroup effect parameters in \( \psi_j \) and a standard non-informative priors for the rest of the parameters are given by

\[
\psi_j | g \sim N(0, g \Sigma_j),
\]

\[
\pi(\mu_0, \sigma) \propto \sigma^{-1},
\]

where \( \Sigma_j = \sigma^2 (V_j^T V_j)^{-1}, \) \( V_j = (I_n - X_0 (X_0^T X_0)^{-1} X_0^T) Z_j \), and the hyperparameter \( g \) is assigned a proper prior

\[
P_j(g) = \frac{a [\rho_j (b + n)]^a}{(g + b)^{(a+1)}} \cdot I_{\{g > \rho_j (b + n) - b\}}.
\]

Here, \( I_{\{g > \rho_j (b + n) - b\}} = 1 \) if \( g > \rho_j (b + n) - b \), and equals 0, otherwise, with constants \( a > 0, b > 0, \rho_j \geq b / (b + n) \). We take \( a = 0.5, b = 1 \) and \( \rho_j = 1 / (k_0 + k_j) \), as suggested in Bayarri et al. (2012), where \( k_0 = \text{rank}(X_0), k_j = \text{rank}(Z_j) \).

In summary, the prior distribution of the parameters of \( M_j \neq M_0 \) is factorized as

\[
P(\psi_j, \mu_0, \sigma, g) = P(\psi_j | g) P_j(g) \pi(\mu_0, \sigma).
\]
For $M_j = M_0$, we only need

$$P(\mu_0, \sigma) = \pi(\mu_0, \sigma). \quad (2.32)$$

As mentioned in the previous subsection, we only need to calculate $B_{j0}$ for models in $M$, which are given by, see Bayarri et al. (2011),

$$B_{j0} = Q_{j0}^{-(n-k_0)/2} \frac{2a}{k_j + 2a} [\rho_j(n+b)]^{-k_j/2} AP_j, \quad (2.33)$$

where $Q_{j0} = \frac{SSE_j}{SSE_0}$ is the ratio of the sum of squared errors of models $M_j$ and $M_0$, and $AP_j$ is the hypergeometric function of two variables

$$AP_j = \frac{2a + k_j}{2} \int_0^1 \lambda^{a+k_j/2} \left(1 - \frac{b-1}{\rho_j(b+n)} \lambda\right)^{n-k_j-k_0} \left(1 - \frac{b-Q_{j0}^{-1}}{\rho_j(b+n)} \lambda\right)^{-n-k_0} d\lambda \quad (2.34)$$

### 2.5 A Stepwise Algorithm for Finding Subgroup Effects

Our approach to finding subgroups is done in two steps. First, we decide whether to report an overall treatment effect $M_1$, report no treatment effect $M_0$, or report a subgroup effect. In the first two scenarios, we stop and report either $M_0$ or $M_1$. In the third scenario, we proceed to find and report a subgroup effect(s). We implement using a stepwise algorithm, consisting of two steps.

In the first step, posterior odds of a subgroup effect model $M_j$ versus the overall null model $M_0$ and the overall alternative model $M_1$ - the models corresponding to no-subgroup effect - are used to determine whether to report a subgroup effect or not. Here we require
that the posterior odds exceed certain pre-specified thresholds in order to proceed with reporting a subgroup effect. If it is determined that no subgroup effect is to be reported, the overall effect model is reported if its posterior odds against the overall null model exceeds a set threshold, otherwise the overall null models is reported. We call this step, the Model Selection Step.

The algorithm proceeds to the second step, Subgroup Reporting Step, when a determination is made in the first step to report a subgroup effect. Here, as discussed in Section 2.2.2, subgroups based on the main effect(s) only are first considered for reporting, i.e., from $M_\gamma = S_1 \cup S_2 \cup S_{12}$, as in (2.4)-(2.6). Failing that, a subgroup with an interaction effect is reported. We first determine the overall "best" subgroup effect model, $M^*$, in terms of the highest posterior probability, and compare it with the best subgroup effect models in each of $S_1$, $S_2$, and $S_{12}$. Any subgroup effect (model) that compares favorably against $M_0$ and $M_1$ and falls close to $M^*$, in terms of their posterior probability, are chosen for reporting, and, in which case, the algorithm stops without looking into subgroup effects based on interaction. If none found, $M^*$, which by default would be a subgroup effect based on a single interaction, is reported. Thus, we allow the possibility of reporting more than one subgroup, although the prior probability distribution on the model space postulates only one "true" model. In the following, the posterior probabilities of $M_\gamma \in S_1 \cup S_2 \cup S_{12}$ is taken as the corresponding marginal posterior probability over the possible values of $\beta$, i.e.,
\[
\bar{P}(M_{\gamma_1,\gamma_2}) = \sum_{M_{\gamma_1,\gamma_2} \in \mathcal{M}} \bar{P}(M_{\gamma_1,\gamma_2,\beta}).
\]

2.5.1 The stepwise algorithm

Here we provide the algorithm for finding subgroup effects in terms of the steps involved in its implementation, which uses pre-specified thresholds \(c_i\)'s for the ratio of the posterior probabilities.

A. Model Selection Step.

Let \( M^* = \arg\max_{M \in \mathcal{M} \setminus \{M_0, M_1\}} \bar{P}(M) \). Conclude subgroup effect exists and go to the subgroup reporting procedure if

\[
\frac{\bar{P}(M^*)}{\bar{P}(M_0)} > c_0 \quad \text{and} \quad \frac{\bar{P}(M^*)}{\bar{P}(M_1)} > c_1.
\]

Else, conclude no subgroup effect exists, and report \( M_1 \) if

\[
\frac{\bar{P}(M_1)}{\bar{P}(M_0)} > c_0
\]

Else, report \( M_0 \).

B. Subgroup Reporting Step.

B1. Let \( M^*_j = \arg\max_{M \in \mathcal{S}_i} \bar{P}(M) \), for \( j = 1, 2, 12 \). Make a subgroup report corresponding to each model \( M^*_j, j = 1, 2, 12 \) that satisfies
\[
\frac{P(M_j^*)}{P(M_0)} > c_0, \quad \frac{P(M_j^*)}{P(M_1)} > c_1 \quad \text{and} \quad \frac{P(M_j^*)}{P(M^*)} > c_2,
\]

where \(0 < c_2 < 1\).

B2. If no model \(M_j^\ast\) is selected in the previous step, make a subgroup report corresponding to \(M^\ast\), which is an interaction only subgroup effect model.

As commonly done in Bayesian approach to clinical trials, we choose the threshold values based on the frequentist operating characteristics of the proposed algorithm, the details of which are given in the next section. We also show that this algorithm approximates the Bayes rule corresponding to the utility functions given below in Theorem 1. The threshold values can alternatively be chosen based on the assessed values of the utility for correct and wrong decisions, which we do not pursue here.

The utility functions consist of two components used sequentially. The first one is used to decide whether to report a subgroup effect, and, if the first step leads to a decision to report a subgroup effect, the second one is used to decide which subgroup to report.

A. In the first step, we decide if \(M_0, M_1\) or a subgroup(s) should be reported. Hence, the action space is \(\mathcal{D}^{(1)} = \{M_0, M_1, A\}\), where \(A\) corresponds to deciding to report a subgroup effect. For convenience, we also write \(A\) as the set of possible subgroup effects to be reported, namely, \(A \subset S_1 \cup S_2 \cup S_{12} \cup S_\lambda\). We assume that at most one model should be reported for each \(S \in \{S_1, S_2, S_{12}, S_\lambda\}\). If \(A\) is selected, we use a suitable subset \(A^\ast\) of \(A\) as the action.
space in the second step to determine which subgroup(s) to report. We define the utility functions in two steps, as described above. In the first step it is given by

\[
u^{(1)}(\delta^{(1)}, M) = \begin{cases} 
u_0 I(M = M_0), & \text{if } \delta^{(1)} = M_0, \\ \nu_1 I(M = M_1), & \text{if } \delta^{(1)} = M_1, \\ \nu_2 I(M \in A) - \max \{0, (n_A - 1)\} \nu_j I(M = M_j), & \text{if } \delta^{(1)} = A, \end{cases}
\]  

(2.35)

where \(\delta^{(1)}\) is the action considered, \(M\) is the true model, \(n_A\) denotes the number of models in \(A\), and \(\nu_i > 0\) for \(i = 0, 1, 2\). \(\nu_j I(M = M_j)\) can be seen as a penalty for adding one more model into \(A\) when it is not empty. We let \(j = 0\) or \(1\).

B. In the second step we choose which subgroup(s) to report, choosing from a set \(A^*\) identified in Step 1. And we report either the models in \(A^* \setminus M_\lambda^*\) or the interaction only model \(M_\lambda^*\). So, the action space is \(\mathcal{D}^{(2)} = \{B : B \subset A^* \setminus M_\lambda^*\} \cup \{M_\lambda^*\}\). The utility function here is given by

\[
u^{(2)}(\delta^{(2)}, M) = \begin{cases} \nu_3 I(M \in B) - \max \{0, (n_B - 1)\} \nu_4, & \text{if } \delta^{(2)} = B, \\ I(M = M_\lambda^*), & \text{if } \delta^{(2)} = M_\lambda^*, \end{cases}
\]  

(2.36)

where \(\nu_i > 0\) for \(i = 3, 4\).

**Theorem 1.** The proposed two-step algorithm approximates the Bayes rule w.r.t. the utility functions in (2.35) and (2.36).
Proof. Step 1: The posterior expected utility is \( U^{(1)}(\delta^{(1)}, y) = \sum u^{(1)}(\delta^{(1)}, M) \bar{P}(M) \). It can be shown that if the optimal rule maximizing the posterior expected utility, denoted \( \delta^{(1)*} \), is equal to \( A \), then it must be given by,

\[
A^* = \{ M : \bar{P}(M) > \frac{u_i}{u_2} \bar{P}(M_0) \text{ and } M \in S^* \},
\]

(2.37)

where \( S^* = \{ M : M = \arg \max_{M' \in S} \bar{P}(M') \text{, for } S \in \{ S_1, S_2, S_{12}, S_{\lambda} \} \} \) and \( j = 0 \) or \( 1 \). As a conservative choice in reporting subgroups, we can replace \( u_j \bar{P}(M_j) \) with \( \max \{ u_0 \bar{P}(M_0), u_1 \bar{P}(M_1) \} \).

Then, \( A^* = \{ M : \bar{P}(M) > \frac{1}{u_2} \max \{ u_0 \bar{P}(M_0), u_1 \bar{P}(M_1) \} \text{ and } M \in S^* \}, \) and

\[
U^{(1)}(\delta^{(1)}, y) = \begin{cases} 
\ u_0 \bar{P}(M_0) & \text{, if } \delta^{(1)} = M_0, \\
\ u_1 \bar{P}(M_1) & \text{, if } \delta^{(1)} = M_1, \\
\ u_2 \sum_{M \in A^*} \bar{P}(M) - \max \{ 0, (n_{A^*} - 1) \} u_j \bar{P}(M_j) & \text{, if } \delta^{(1)} = A^*. 
\end{cases}
\]

(2.38)

Letting \( M^* = \arg \max_{M \in S^*} \bar{P}(M) \), we have

\[
\delta^{(1)*}(y) = \begin{cases} 
A^* & \text{, if } \frac{\bar{P}(M_1)}{\bar{P}(M_0)} < \frac{u_0}{u_1} + \sum_{M \in A^* \setminus M^*} \frac{u_2 \bar{P}(M) - u_j \bar{P}(M_j)}{u_i \bar{P}(M^*)} \text{ for } i = 0, 1, \\
M_1 & \text{, if } \frac{\bar{P}(M_1)}{\bar{P}(M_0)} > \frac{u_0}{u_1} \text{ and } \frac{\bar{P}(M_1)}{\bar{P}(M^*)} > \frac{u_2}{u_1} + \sum_{M \in A^* \setminus M^*} \frac{u_2 \bar{P}(M) - u_j \bar{P}(M_j)}{u_i \bar{P}(M^*)}, \\
M_0 & \text{, otherwise.} 
\end{cases}
\]

(2.39)

Using (2.37), we can simplify the above rule and make it more conservative in reporting subgroups by substituting \( u_2 \bar{P}(M) - u_j \bar{P}(M_j) \) with \( 0 \) for all \( M \in A^* \setminus M^* \). Let \( c_0 = \frac{u_0}{u_2} \),
This gives a simplified version of the Bayes rule, given by

\[
\delta^{(1)*}(y) = \begin{cases} 
A^*, & \text{if } \frac{P(M^*)}{P(M_i)} > c_i \text{ for } i = 0, 1, \\
M_1, & \text{if } \frac{P(M_1)}{P(M_0)} > \frac{c_0}{c_1} \text{ and } \frac{P(M^*)}{P(M_i)} < c_1, \\
M_0, & \text{otherwise}.
\end{cases}
\] (2.40)

This is identical to the first step of our algorithm if \(c_0/c_1\) in (2.40) is substituted with \(c_0\).

Note that this also makes TIE less sensitive to the value of \(c_1\), making it easier for us to determine \(c_0\) and \(c_1\) that yields specified values for TIE and FSR.

Step 2: In this step we determine which subgroup(s) to report when \(\delta^{(1)*}(y) = A^*\). More specifically, the action space is \(D^{(2)} = \{B : B \subset A^* \setminus M_\lambda^*\} \cup \{M_\lambda^*\}\), where \(M_\lambda^* = \arg \max_{M' \in S_\lambda} \bar{P}(M')\). Note that \(M_\lambda^*\) may not be in \(A^*\).

We denote the posterior expected utility by \(U^{(2)}(\delta^{(2)}, y)\) and the optimal rule by \(\delta^{(2)*}\). As before, if \(\delta^{(2)*}(y) = B\), then \(B\) should be

\[
B^* = \{M : \bar{P}(M) > \frac{u_4}{u_3} \text{ and } M \in A^* \setminus M_\lambda^*\}.
\] (2.41)

Letting \(u_4 = \bar{P}(M^*)\), \(B^* = \{M : \bar{P}(M) > \frac{1}{u_3} \bar{P}(M^*) \text{ and } M \in A^* \setminus M_\lambda^*\}\), and the posterior expected utility is

\[
U^{(2)}(\delta^{(2)}, y) = \begin{cases} 
\frac{u_3}{u_4} \sum_{M \in B^*} \bar{P}(M) - \max \{0, (n_{B^*} - 1)\} \bar{P}(M^*) & , \text{if } \delta^{(2)} = B^*, \\
\bar{P}(M_\lambda^*) & , \text{if } \delta^{(2)} = M_\lambda^*.
\end{cases}
\] (2.42)
and the corresponding optimal rule is

\[
\delta^{(2)*}(y) = \begin{cases} 
B^*, & \text{if } \bar{P}(M_B^*) > c_2 \bar{P}(M^*) \\
M^*_\lambda, & \text{otherwise.}
\end{cases} 
\] (2.43)

where \( M_B^* = \arg \max_{M' \in S^* \setminus M^*_\lambda} \bar{P}(M') \) and \( c_2 = 1/u_3 \).

This simplified version of the Bayes rule is equivalent to the subgroup reporting procedure in the second step of our algorithm.

\[ \square \]

2.6 Operating characteristics and a real data example

To evaluate the performance of the stepwise procedure, we simulated data from

\[
Y_0 | x_{1,k}, x_{2,l} \sim N(\mu_{kl}, \sigma_{k0}^2) \] (2.44)

\[
Y_1 | x_{1,k}, x_{2,l} \sim N(\mu_{kl} + \delta_k + \eta_l + \lambda_{kl}, \sigma_{k1}^2) \] (2.45)

with \( \mu_{kl} = 0, k, l = 1, 2 \). We simulated under different settings each corresponding to a different value for \( \{\delta_k, \eta_l\} \) and subgroup effect model. For all the simulations, the cell sizes and cell-specific variances were fixed according to the sample sizes and sample variances from the Serial Trial Intervention (STI) data described below. In all cases, we fixed the prior distribution for \( p \) and \( q \) as independent \( U(0,1) \), and fixed \( c_2 = 0.95 \).
2.6.1 Introduction to STI example

Kovach et al. (2006) reported on a double-blinded randomized clinical trial to study the effectiveness of Serial Trial Intervention (STI), the treatment, on comfort and behavior in patients with late-stage dementia. The study was conducted in 14 nursing homes on 112 subjects with late-stage dementia. STI is an innovative clinical protocol for assessment and management of unmet needs in people with late-stage dementia. The outcome variable of interest was the difference, pre- and post-intervention, in Discomfort-DAT (Discomfort-Dementia of the Alzheimer’s Type scale), a measure of discomfort felt by the subjects. The sample sizes for the treatment and control were 55 and 57, respectively. The investigators (personal communication) were interested in the subgroups defined by two covariates, Functional Assessment Staging of Dementia (FAST) score (covariate $X_1$) and Presence/Absence of Vocalization in Behavioral Symptoms (MVOCAL) initiating treatment (covariate $X_2$). Two subgroups were of interest based on $X_1$, defined by $X_1 = 1$ when FAST Score $\geq 7$, and, $X_1 = 0$ if FAST Score $< 7$. The presence ($X_2 = 1$) and absence ($X_2 = 0$) of vocalization in behavioral symptoms initiating treatment also define two subgroups of interest. Subgroup sample sizes after deleting the missing data are shown in Table 2.5. Thus we have 2 covariates, each with $K = 2$ levels. Together, they define four subgroups. There are 46 models, the overall null model $M_0$, the overall effect model $M_1$, 36 subgroup effect models with subgroup effect at levels of $X_1$ or $X_2$ and 8 subgroup effect models with the subgroup effect at interaction term.
Table 2.5: Sample sizes for the four subgroups for the STI example

<table>
<thead>
<tr>
<th></th>
<th>$X_2 = 0$</th>
<th>$X_2 = 1$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_1 = 0$</td>
<td>8</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>$X_1 = 1$</td>
<td>9</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>35</td>
<td>52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$X_2 = 0$</th>
<th>$X_2 = 1$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_1 = 0$</td>
<td>10</td>
<td>21</td>
<td>31</td>
</tr>
<tr>
<td>$X_1 = 1$</td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>30</td>
<td>47</td>
</tr>
</tbody>
</table>

only.

2.6.2 Error Rates

We define the following error/correct rates that are of most interest in subgroup analysis, where $P_f$ represents probability under repeated experiments.
TIE: Type-I Error \( P_f(M_0 \text{ not selected } | M_0) \)

FNR: False Negative Rate \( P_f(M_0 \text{ selected } | M_1) \)

TPR: True Positive Rate \( P_f(M_1 \text{ selected } | M_1) \)

FSR: False Subgroup Rate \( P_f(M \in \mathcal{M}\{M_0, M_1\} \text{ selected } | M_1) \)

TSR: True Subgroup Rate \( P_f(M_{\gamma_1, \gamma_2} \text{ selected } | M_h = M_{\gamma_1, \gamma_2, \beta}) \)

\quad \text{for } \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0

FPR: False (overall) Positive Rate \( P_f(M_1 \text{ selected } | M_h = M_{\gamma_1, \gamma_2, \beta}) \)

MSR: Misspecified Subgroup Rate \( P_f(M_{\gamma_1, \gamma_2} | M_{0,0,\beta}) \) \text{ for } \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0

We first simulated 1000 data sets from the null model \( M_0 \), and implemented the algorithm, with \( c_1 \) fixed at 1, and \( c_2 \) at 0.95. We estimated the TIE (Type-I Error) as the percentage of not reporting \( M_0 \) out of the 1000 simulations. A grid of values for \( c_0 \) was evaluated.

2.6.3 Unequal Variance

Using the unequal variance setup, the grid search for \( c_0 \) is shown in Figure 2.1. We selected the threshold of \( c_0 = 1.84 \) that fixed TIE at 0.05. Then we simulated another 1000 data sets from the overall effect model \( M_1 \) with an effect size of one standard deviation, and implemented the algorithm with \( c_0 = 1.84 \), and a grid of values for \( c_1 \). For FSR=0.05, we
found $c_1 = 2.4$ based on $c_0 = 1.84$. But for FSR=0.1, we found $c_1 = 1.2$ based on a different $c_0 = 0.76$. Note that TIE may change slightly when $c_1$ changes. Therefore, if TIE ≠ 0.05 with the fixed $c_0$ and $c_1$, we do another grid search around the previously found $c_0$ and $c_1$ to fix TIE and FSR at the levels we want. The resulting $c_0$ and $c_1$ are then plugged in the algorithm to analyze the simulated data and real data.

![Figure 2.1: Unequal variance: TIE vs $c_0$.](image)

For the simulation study, we only use $c_0 = 0.76$ and $c_1 = 2.62$ that fixed both TIE and FSR at 0.05. At these values, for data was simulated from the overall effect model $M_1$, TPR was close to 0.95, and FSR and FNR were 0.05 and 0.01, respectively. When the data was simulated from the subgroup effect model $M_{(1,2), (0,0), (0)}$, with effect size of 1.5, TSR and FPR were 0.7 and 0.25, respectively. When the data was simulated from the subgroup effect
model $M_{(0,0)(0,0)(0,0,0,1)}$ with effect size of 2, MSR was 0.313.

![Figure 2.2: Unequal variance: (a) Rates vs effect size under overall effect model $M_1$, (b) Rates vs effect size under subgroup effect model $M_{(1,2),(0,0),(0,0,0,0)}$.](image)

To study how effect size affects error rates, we simulated data sets from the same models with effect size ranging from 0 to 2. Here, we take the effect size of one as the estimate of the standard deviation of the difference in the response under treatment and control, which we estimate by the average of the estimated standard deviations of the difference over the four cells (in the $2 \times 2$ case). Similar to traditional power curve, the rates are plotted against the standardized effect sizes for data simulated from the overall effect model and the subgroup effect models $M_{(1,2),(0,0),(0,0,0,0)}$. We report the results in Figure 2.2(a) and (b), respectively, with the fixed thresholds $c_0$, $c_1$ and $c_2$ obtained previously. In Figure 2.2(a), the
TPR increases with the effect size, as expected, and FSR increases slightly and then decreases while remaining low. In Figure 2.2(b), TSR increases with the effect size, while FPR and FSR both increase first and then decrease.

2.6.4 Equal Variance

Using the model and prior setup in Section 2.4.4, the grid search for $c_0$ is shown in Figure 2.3. We selected the threshold of $c_0 = 0.91$ that fixed TIE at 0.05. Then we simulated another 1000 data sets from the overall effect model $M_1$ with an effect size of one standard deviation, and implemented the algorithm with $c_0 = 0.91$, and a grid of values for $c_1$. For
FSR=0.05, we found $c_1 = 1.13$ based on $c_0 = 0.91$. Then another grid search is done around $c_0 = 0.91$ and $c_1 = 1.13$. We find that TIE=FSR=0.05 when $c_0 = 0.87$ and $c_1 = 1.14$. But for TIE=FSR=0.1, we found $c_0 = 0.39$ and $c_1 = 0.6$. The resulting $c_0$ and $c_1$ are then plugged in the algorithm to analyze the simulated data and real data.

For the simulation study, we only use $c_0 = 0.87$ and $c_1 = 1.14$ that fixed both TIE and FSR at 0.05. At these values, for data was simulated from the overall effect model $M_1$, TPR was close to 0.95, and FSR and FNR were 0.05 and 0.01, respectively. When the data was simulated from the subgroup effect model $M_{(1,2),(0,0),(0,0)}$, with effect size of 1.5, TSR and FPR were 0.765 and 0.197, respectively. When the data was simulated from the subgroup effect model $M_{(0,0)(0,0)(0,0,0,1)}$ with effect size of 2, MSR was 0.313.

![Figure 2.4: Equal variance: (a) Rates vs effect size under overall effect model $M_1$, (b) Rates vs effect size under subgroup effect model $M_{(1,2),(0,0),(0,0,0)}$.](image)
To study how effect size affects error rates, we simulated data sets from the same models with effect size ranging from 0 to 2. Here, we take the effect size of one as the standard deviation. Similar to traditional power curve, the rates are plotted against the standardized effect sizes for data simulated from the overall effect model and the subgroup effect models $M_{(1,2),(0,0),(0,0,0,0)}$. We report the results in Figure 2.4(a) and (b), respectively, with the fixed thresholds $c_0$, $c_1$ and $c_2$ obtained previously. In Figure 2.4(a), the TPR increases with the effect size, as expected, and FSR increases slightly and then decreases while remaining low. In Figure 2.4(b), TSR increases with the effect size, while FPR and FSR both increase first and then decrease. Those patterns are exactly the same as the plots for unequal variance case.

2.6.5 Result for STI example

Using independent Uniform$(0,1)$ for $p$ and $q$, and thresholds $c_0 = 0.76$, $c_1 = 2.62$, and $c_2 = 0.95$, we applied the stepwise algorithm to the STI study. These thresholds correspond to both TIE and FSR of 0.05. The two models with the highest posterior probabilities were the overall effect model $M_1$ and the subgroup effect model $M_{(0,1),(0,1)}$, with the respective posterior probabilities 0.186 and 0.207, while all other models had much smaller posterior probabilities. This subgroup effect model represents the case that the treatment (STI) effect is present among those with FAST Score $\geq 7$ ($X_1 = 1$) and absent among those with FAST
Score< 7 (X_1 = 0), and, treatment effect exists among those with vocalization in behavioral symptoms (X_2 = 1), and does not exist when there is no vocalization in behavioral symptoms (X_2 = 0). The algorithm selected the overall effect model M_1 when FSR=0.05, which means the treatment is effective and no subgroup effect exists, and the conclusion did not change with FSR set at 0.1. The conclusion with equal variance assumption was slightly different in that the algorithm picked the overall effect model when FSR was set at 0.05 but picked the subgroup effect model \( M_{(0,1),(0,1)} \) when FSR was set at 0.1. As a comparison, we estimated the cell specific treatment effect using the standard frequentist approach, and tested the overall effect, and used a 2-way ANOVA and tested the three contrasts corresponding to the two main effects and the interaction. When each test was carried out at 5% level, the contrast corresponding to testing the difference between MVOCAL levels was significant, albeit slightly, but none except the overall effect was significant when tested using Bonferroni adjustment with a family wise level of 5%.

2.7 Summary and discussion

In this chapter, we proposed a Bayesian model selection approach to subgroup analysis when subgroups are defined based on 2 or more covariates. We used a collection of ANOVA models that included models with all the covariates of interest for subgroup analysis simultaneously, extending an earlier model selection based approach which considered models with one co-
variates at a time. Our approach is based on a premise that, investigators can a priori specify a set of covariates for which subgroup analysis is of interest, and that, in the interest of parsimony, interpretability and subpopulation size, the investigators’ primary interest and goal is to determine whether there is a subgroup effect associated with any single covariate. It is only in the scenario that there is no evidence of subgroup effect based on a single covariate, there is interest in finding whether there is subgroup effect based on an interaction effect involving more than one covariate.

We have defined a model space consisting of models that represent the subgroup effects of interest, and developed prior distribution on the model space by casting the existence of subgroup effects in terms of the familiar variable selection context and using ideas from the literature on constructing variable selection priors. The posterior probabilities of the models are used as measures of evidence for various subgroups of interest. But, in order to decide whether subgroup effect exists and if so, which subgroup effect to report, we proposed a stepwise algorithm, consistent with the stated goal, based on pre-specified thresholds for the posterior probabilities. The values of thresholds can be set based on a utility function specifying the relative utilities associated with different types of decisions. Instead, we have used the frequentist operating characteristics of the proposed algorithm to estimate the error rates and chosen thresholds that correspond to certain specified error rates. These error rates included the overall Type-I error rate and false discovery of a subgroup effect when there is an overall effect. Through the use of suitable prior distribution on the models to calculate
the posterior probabilities of the subgroup effect models, and the use of frequentist error rates to set the threshold for the posterior probabilities, our approach achieves multiplicity adjustment in both Bayesian and frequentist sense, and provides a way to decide on the existence of subgroup effects. We have also verified that the proposed algorithm can be justified as an approximate Bayes rule corresponding to a specific utility function.

In most of our presentation in the chapter, we have focused on the case of two covariates each at two levels, in order to present the proposed approach while keeping complexity of the notation at a minimum. We have also briefly outlined how our approach can be extended to higher number of covariates and the related challenges and possible remedies. We have also presented certain aspects of the approach such as the specification of the prior in a more general setting. However, in the interest of avoiding data "dredging" and keeping the search for subgroup effects to subgroups a priori deemed plausible, it would be useful to set out a search for parsimonious and most promising subgroups. Such an approach may include setting priority ordering on the covariates, their values (when there are more than two), and on the nature of the subgroup effects.

We have explored ways to avoid the use of unidentifiable parameters by using the constraints commonly used in ANOVA models, and thus avoid the need to consider equivalent models and model indices. But, we have run into difficulties in specifying priors for the resulting parameters that are suitable for testing for subgroup effects configurations such as equality of effects at certain levels of a covariate and unequal effects at some other levels.
Also, taking an approach of using a Bayesian variable selection and picking a best model to determine subgroup effects, we believe, would encounter difficulties in choosing a suitable prior for the parameters. It would also not accommodate the priority ordering among subgroups for reporting that we seek, such as giving priority to subgroup effects based on main effects over reporting an interaction effect. In this chapter we use the stepwise algorithm and posterior probabilities of the models to achieve this. We will continue to explore ways to simplify the approach while achieving the desired goals.

The work in this chapter is published in Liu et al. (2017).
Chapter 3

Bayesian Causal Inference Using

Gaussian Process

Bayesian approach to causal inference has traditionally been modeling the outcome mechanism and ignoring the fact that treatments are selectively assigned. On the other hand, frequentist approach has been relying on removing treatment selection bias via baseline covariates matching or propensity score methods, both of which are two-step methods.

This chapter proposes a fully Bayesian semiparametric approach for estimating the population mean treatment effect using Gaussian process (GP), which accomplishes matching and modeling outcome mechanism in a single step. We demonstrate a close relationship between matching method and GP regression for estimating average treatment effect. The proposed method utilizes a distance similar to Mahalanobis distance but determines the
range of matching automatically without imposing a caliper arbitrarily.

The results from our simulation study suggest that GP regression leads to an accurate and more efficient estimate than the linear regression modeling with adjustment for propensity score, inverse probability of treatment weighting and BART.

We also illustrate the proposed method using data obtained from the Pediatric Rheumatology Collaborative Outcome Improvement Network for estimating the effect of early aggressive biological disease modifying anti-rheumatic drugs.

### 3.1 Introduction

When dealing with observational data, we can no longer take advantage of the randomized assignment of treatments. Ignoring the treatment selection bias may lead to undesired bias in causal inference since only one of the potential outcomes can be observed for each individual. To solve this problem, causal inference methods have been developed predominately within frequentist framework. Most of them required two steps. The first step is the design stage without knowledge of the outcome, via estimation of propensity score (PS) or covariate matching. The second step is evaluating causal effect based on the matched sample, or the propensity score methods.

Depending on how the propensity score is constructed and used, many different propensity score methods have been proposed. For example, the inverse probability of treatment
weighting (IPTW) estimates the population average treatment effect by weighted mean difference, where weight is its probability being assigned to the corresponding treatment group. However when propensity score is not correctly specified, the validity of the causal inference is in question. There are also many different matching methods, such as K nearest neighbor matching, subclassification matching, exact matching, full matching, genetic matching, etc (Stuart, 2010). A critical limitation of the matching methods is discarding data points that are not being matched on, which may lead to a sample that cannot represent the population. Moreover, some matching methods require matching within a caliper, where the caliper is determined arbitrarily.

Bayesian inference to causal treatment effect can be approached by predicting missing potential outcomes through directly modeling observed outcomes (Rubin, 1978). However, such approach may suffer from bias when the outcome model is misspecified. Much recent work has been searching for ways to include propensity score into the Bayesian causal inference. This includes including PS as a covariate into the outcome regression model (McCandless et al., 2009). However, joint modeling of outcome model with treatment selection model leads to a feedback issue, and two-stage approaches were suggested in Zigler et al. (2013). Others investigated Bayesian methods for variable selection and model averaging (Zigler and Dominici, 2014). Saarela et al. (2016) proposed to incorporate inverse treatment assignment probabilities as importance sampling weights in Monte Carlo (MC) integration, which offers a Bayesian version of IPTW. As the authors point out, it has been difficult to
incorporate treatment selection process within a formally likelihood-based framework.

An issue under debate about the two-stage methods is whether the uncertainty of the first step should be taken into account when obtaining the final result in the second step. (Hill and Reiter, 2006) suggested that researchers can use bootstrap to account for the uncertainty from the matching procedure if necessary. But Rubin and Thomas (1996); Rubin and Stuart (2006), implied that when matching method using estimated propensity scores, not including the uncertainty in exposure model is already conservative for calculating the confidence limits.

To avoid these problems, (Hill, 2012) suggested using Bayesian Additive Regression Tree (BART) for causal inference. BART can model the response surface very flexibly and provide the estimate of average treatment effect from the posterior sample. However, it cannot utilize the prior knowledge of the treatment selection and outcome mechanism.

In this chapter, we propose a Bayesian semi-parametric model to obtain the average treatment effect directly. The proposed approach can integrate the benefits of the regression model and matching method. In our approach, we use Gaussian process modeling, and include the outcome mechanism in the mean and the covariates for matching in the covariance function. The result of our simulation study shows that our approach produces a more accurate and efficient estimate of average treatment effect than IPTW, linear regression with adjustment for propensity score and BART.

The chapter is organized as follows. In Section 3.2, we introduce the problem of interest,
notations and assumptions. In Section 3.3, GP is introduced to estimate the mean treatment effect. Bayesian inference is utilized for imposing the prior belief and parameter estimate. The link between the proposed GP model and matching method is discussed in Section 3.4. In Section 3.5, we present a simulation study to compare the proposed approach with some well-known methods. In Section 3.6, the new approach is applied to a real data. We end with a discussion in Section 3.7.

### 3.2 Causal inference

This section describes notations, estimand of interest and assumptions in this chapter.

We consider the case of one treatment and one control group. Using $T_i$ as treatment indicator, $i^{th}$ patient either accepts the treatment, $T_i = 1$, or is in the control group, $T_i = 0$, $i = 1,2,...,n$. The potential outcomes under treatment and control are denoted by $Y_i^{(1)}$ and $Y_i^{(0)}$, respectively. We use $Y_i$ and $Y_i^c$ to denote the observed outcome and unobserved potential outcome. The following notations are used for the observed data and potential responses when describing the model, $T = (T_1, T_2, \ldots, T_n)'$, $Y = (Y_1, \ldots, Y_n)'$, $Y^c = (Y_1^c, \ldots, Y_n^c)'$ and $Y^{(1)} = (Y_1^{(1)}, \ldots, Y_n^{(1)})'$. Let $X$ be the collection of covariates that are related to the outcome mechanism and $V$ be composed of covariates sufficient to ensure unconfoundedness. It is of interest to find out the population average treatment effect, $\text{PATE} = E\left(Y^{(1)} - Y^{(0)}\right)$. However, when comparing with other methods, sample average
treatment effect, SATE = \( n^{-1} \sum_{i=1}^{n} \left( Y_i^{(1)} - Y_i^{(0)} \right) \), may be used if they cannot estimate PATE. For simplicity, we call both of them average treatment effect (ATE). In this chapter, we focus on continuous response and dichotomous treatment assignment. But the approach can be extended to other types of responds and applied to the continuous treatment case.

To estimate the ATE, we need the following assumptions. First, stable unit treatment value assumption (SUTVA) is necessary (Rubin, 1986), which requires the potential outcome of one unit should not be influenced by how the treatment is assigned to this unit and the treatment assignments of other units. Second, we assume that there is an observed covariate set \( V \), which is sufficient to ensure unconfoundedness, i.e., given \( V \), potential outcomes \( (Y^{(0)}, Y^{(1)}) \) and treatment assignment \( T \) are conditionally independent. Lastly, for all values of \( V \), we assume that the probability assigning an individual to a certain group should not be zero or one. The later two assumptions can be written as \( (Y^{(0)}, Y^{(1)}) \perp T | V \) and \( 0 < P(T = 1 | V) < 1 \).

### 3.3 Gaussian process for causal inference

In this section, we introduce a GP model for causal inference to estimate the average treatment effect. Gaussian process assumes that the observations occur in a continuous domain and every finite collection of these data points correspond to random variables following a multivariate normal distribution. It has been widely used in machine learning and spatial
analysis. Xu et al. (2016) proposed a dependent Dirichlet process prior with Gaussian process base measure for dynamic treatment regimes. Roy et al. (2017) proposed a combination of a dependent Dirichlet process and Gaussian Process (GP) to model the observed data, which did not require parametric modeling assumptions and accounted for uncertainty in the identifying assumption.

Our motivation of taking this approach is that it can predict the unobserved potential outcome much like the matching method. Matching methods usually impute the missing value with the nearest matched value or the (weighted) average of the responds within its neighborhood that defined by a set of variables, such as $V$, which could involve the propensity score. The imputed data is then reweighted and used in further analysis. In matching, the imputed value is correlated with the points nearby but not correlated with the points far from it. If we think of the response surface of a treatment group as a function following GP and construct the covariance function using a distance defined by $V$ such that the correlation \[ |\text{corr}(Y_i, Y_j)| \geq |\text{corr}(Y_{i'}, Y_{j'})| \] if and only if $d_{ij} \leq d_{i'j'}$, where $d_{ij}$ is the distance between the $i^{th}$ and $j^{th}$ individuals, then this GP can be seem as a “matching method” by setting the correlation according to the magnitude of their distance. The relationship between GP and matching is further discussed in section 3.4.
3.3.1 GP for estimating mean treatment effect

Consider that the response in control group as a Gaussian process $Y^{(0)}(\cdot)$ with mean $x_b \beta_b$ and covariance function $k_b(s, u) = \sigma^2_f \exp \left( -(d_{su})^2 \right) + \sigma^2_0 \delta_{su}$, where $x_b \beta_b$ is the linear trend, $x_b$ is a vector of the covariates in the linear trend, $\beta_b = (\beta_0, \beta_1, ..., \beta_p)'$; $d_{su}$ is the distance between data points $s$ and $u$, which will be defined later, $\delta_{su}$ is the Kronecker function, equals 1 if $s = u$, otherwise zero. The response in the treatment group is $Y^{(1)}(\cdot) = Y^{(0)}(\cdot) + \Delta$, where $\Delta$ is the treatment effect. Although $\Delta$ could be a random variable depending on some covariates, we consider it as a location parameter for estimating the average treatment effect.

Then the response in treatment group is a Gaussian process with mean $x_b \beta_b + \Delta$, covariance function $k_b(\cdot, \cdot)$. Thus given $T$, the observed response $Y(\cdot)$ is a Gaussian process with mean $x \beta = x_b \beta_b + \Delta \cdot T$ and covariance function $k_b(\cdot, \cdot)$, where $\beta = (\beta_0, \beta_1, ..., \beta_p, \Delta)'$. Then the observations follow a multivariate normal distribution

$$Y \sim \text{MVN}(X \beta, \Sigma_b),$$

where $X = \begin{pmatrix} X_b, & T \end{pmatrix}$ is the design matrix for the linear trend, $\Sigma_b = (\sigma_{ij})_{n \times n}$, $\sigma_{ij} = \sigma^2_f \exp \left( -(d^2_{ij}) \right) + \sigma^2_0 \delta_{ij}$, $i, j = 1, 2, ..., n$. Without loss of generality, $X$ is assumed to be full rank. Let $R_\phi = (r_{ij})_{n \times n}$, $r_{ij} = \exp \left( -(d^2_{ij}) \right)$, $G = (R_\phi + \nu I_n)$, where $\nu = \frac{\sigma^2_0}{\sigma^2_f} > 0$, then $\Sigma_b = \sigma^2_f G$.

We define the distance $d_{ij} = \sqrt{\sum_{k=1}^{q} \phi_k^{-1} s_k^{-2} (v_{ki} - v_{kj})^2}$, where $v_k = \{v_{k1}, v_{k2}, ..., v_{kn}\}$, $k = 1, 2, ..., q$, are the covariates in set $V$ and $s_k$ is the sample standard deviation of $v_k$. Note
that the covariates in the mean of GP and those for evaluating distance could be different. We suggest using a regression model reflecting the outcome mechanism in the mean and using variables in $V$ to evaluate distance. In the definition of distance, $\phi_k^{-1}$ indicates how important the corresponding standardized covariates $v_k/s_k$ is for “matching”, comparing with other covariates in $V$. The covariates in the distance are standardized so that the $\phi_k$'s are comparable and can be assigned the same prior distribution.

When $\phi_k^{-1}$ is small, $s_k^{-1}|v_{ki} - v_{kj}|$ contributes less to the distance $d_{ij}$ and thus has little effect on the covariance $\sigma_{ij}^2$. In other words, $\phi_k$'s define how large the neighborhood that we need to consider for imputing the missing potential outcome. For instance, any observations $Y_j$ with $s_k^{-1}|v_{ki} - v_{kj}| > \sqrt{2\phi_k}$ may have little effect on imputing the missing potential outcome $Y^c_i$, since $\text{corr}(Y_i, Y_j) = \frac{\sigma_i^2 \exp(-d_{ij}^2)}{\sigma_i^2 + \sigma_0^2} = (1 + \nu)^{-1} \exp(-d_{ij}^2) < \exp(-2) < 0.0184$ provides a small correlation between $Y_i$ and $Y_j$. When all the $\phi_k^{-1}$'s are very large, $d_{ij}^2$ is close to zero for all $i \neq j$. Then $\Sigma_b$ will be approximately diagonal and $v_k$'s are not useful in the covariance function.

The upper bound of the correlation between two different observations is $\frac{\sigma_i^2}{\sigma_i^2 + \sigma_0^2} = \frac{1}{1+\nu}$. Therefore, a large $\nu$ can result in an approximately diagonal covariance matrix too. In that case, $v_k$'s are not useful even if $\phi_k$'s are small. In particular, the GP model becomes a linear model $N(X\beta, \sigma_0^2 I_n)$ when $\nu \to \infty$. A small $\nu$ will lead to an upper bound of the correlation close to 1, which means the GP model will gather more information from the observations nearby when estimating the missing potential outcome. Therefore, the larger $\sigma_f^2$ is comparing
with $\sigma_0^2$, the more important $V$ is to match on.

### 3.3.2 Prior

The following priors are assigned to the parameters in GP.

$$\beta \sim \text{MVN}(\mu_\beta, \sigma_\beta^2 G_\beta),$$

$$\sigma_f^2 \sim \text{IG}(a_f, b_f),$$

$$\sigma_0^2 \sim \text{IG}(a_0, b_0),$$

$$\phi_k \sim \text{IG}(a_2, b_2), (k = 1, 2, ..., q),$$

where $\mu_\beta = 0$, $G_\beta = g (X'X)^{-1}$, $g = 10^6$, $a_2 = b_2 = 1$, $a_0 = a_f = 2$, $b_f = b_0 = \frac{\hat{\sigma}_{lm}^2}{2}$ and $\hat{\sigma}_{lm}^2$ is the MSE when fitting model $Y \sim N(X\beta, I\sigma_{lm}^2)$.

Since the GP model is attempting to explain the variation of $Y$, we assign the priors for $\sigma_0^2$ and $\sigma_f^2$ such that the expectation of $\sigma_0^2 + \sigma_f^2$ equals to $\hat{\sigma}_{lm}^2$ and substantial probability is assigned around $\hat{\sigma}_{lm}^2$. $a_2$ and $b_2$ are set to one such that $\phi_k^{-1}$ follows exponential distribution with mean and variance both one and smaller values are preferred for $\phi_k^{-1}$, indicating the preference to smoother response surfaces. $g$ is set to a large value such that a vague prior is assigned to $\beta$. We also tried other values for the hyper-parameters of $\phi_k$, $\sigma_0^2$ and $\sigma_f^2$ to investigate the sensitivity of changing priors. The estimates of those parameters are slightly influenced, but the estimate of average treatment effect is not significantly affected.
Prior knowledge could be easily incorporated. For example, if we think that \( v_{k_0} \) is twice as important as the other covariates to match on, then we can assign prior \( \phi_{k_0} \sim IG(a_2, b_2/2) \), such that the prior mean of \( \phi_{k_0}^{-1} \) is twice as large as the other scaling parameters \( \phi_k^{-1} \)'s. 

For the ease of computation, we transform the parameters as following,

\[
\nu = \frac{\sigma_2^2}{\sigma_f^2},
\]

\[
\omega = \log(\nu),
\]

\[
\rho_k = -\log(\phi_k).
\]

Let \( \Phi = (\phi_0, \phi_1, ..., \phi_q)' \), \( \rho = (\rho_1, \rho_2, ..., \rho_q)' \). Since the priors for \( (\phi, \sigma_0^2, \sigma_f^2) \) are proper, we can obtain the prior distribution for \( (\omega, \rho, \sigma_f^2) \).

\[
p(\omega, \rho, \sigma_f^2) = p(\rho) \cdot p(\omega, \sigma_f^2),
\]

\[
p(\rho) = \prod_{k=1}^{q} \left[ f_{\phi_k}(\exp(-\rho_k)) \cdot \exp(-\rho_k) \right],
\]

\[
p(\omega, \sigma_f^2) \propto (\sigma_f^2)^{-a_f-1} \exp \left\{ \frac{-b_f}{\sigma_f^2} \right\} \cdot (\sigma_f^2 \nu)^{-a_0-1} \exp \left\{ \frac{-b_0}{\sigma_f^2 \nu} \right\} \cdot \sigma_f^2 \exp(\omega)
\]

\[
\propto (\sigma_f^2)^{-(a_f+a_0)-1} \exp \left\{ -\sigma_f^{-2} (b_f + b_0 \exp(-\omega)) \right\} \cdot \exp(-a_0 \omega),
\]

\[
p(\omega, \sigma_f^2) = p(\sigma_f^2 | \omega) \cdot p(\omega),
\]

where \( f_{\phi_k} \) is the pdf of the prior for \( \phi_k \), \( p(\sigma_f^2 | \omega) \) is the pdf of \( IG(a_{fa}, b_{fa}) \), \( a_{fa} = a_f + a_0 \), \( b_{fa} = b_f + b_0 \exp(-\omega) \), and \( p(\omega) = \left( \frac{b_{fa}^{a_{fa}}}{\Gamma(a_{fa})} \right) \left( \frac{b_f^{a_f}}{\Gamma(a_f)} \right) \left( \frac{b_{fa}^{a_{fa}+a_0}}{\Gamma(a_{fa}+a_0)} \right)^{-1} \cdot \exp(-a_0 \omega) \).
3.3.3 Posterior inference

Multiplying likelihood by priors, we have

\[ L = p(Y, \beta, \sigma_f^2, \rho, \omega) \propto p(Y|\beta, \Sigma_b) \cdot p(\beta) \cdot p(\omega, \rho, \sigma_f^2), \]

where

\[ p(Y|\beta, \Sigma_b) \propto |\Sigma_b|^{-\frac{3}{2}} \exp \left\{ -\frac{1}{2} (Y - X\beta)' \Sigma_b^{-1} (Y - X\beta) \right\}. \]

Full conditional distribution of \( \beta \) is

\[
p(\beta|Y, \Sigma_b) \propto \exp \left\{ -\frac{1}{2} \left[ (Y - X\beta)' \Sigma_b^{-1} (Y - X\beta) + (\beta - \mu_{\beta p})' \sigma_f^{-2} G_{\beta p}^{-1} (\beta - \mu_{\beta p}) \right] \right\}
\]

\[
\propto \exp \left\{ -\frac{1}{2} (\beta - \mu_{\beta p})' \Sigma_{\beta p}^{-1} (\beta - \mu_{\beta p}) \right\},
\]

where \( \Sigma_{\beta p} = (X'\Sigma_b^{-1}X + \sigma_f^{-2} G_{\beta p}^{-1})^{-1} = \sigma_f^2 G_{\beta p} \) and \( \mu_{\beta p} = G_{\beta p} (X'G^{-1}Y + G_{\beta p}^{-1} \mu_{\beta p}), \)

Thus

\[ (\beta|Y, \Sigma_b) \sim MVN \left( \mu_{\beta p}, \Sigma_{\beta p} \right). \]

(3.1)
Marginal posterior distribution of \((\omega, \rho, \sigma_f^2)\) is

\[
p(\omega, \rho, \sigma_f^2|Y) \propto \int_{-\infty}^{+\infty} p(Y, \beta, \sigma_f^2, \rho, \omega)d\beta
\]

\[
\propto |\Sigma_b|^{-\frac{1}{2}} |\Sigma_{\beta p}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} Y'\Sigma_b^{-1}Y - \frac{1}{2} \sigma_f^{-2} \mu'_{\beta} G_{\beta}^{-1} \mu_{\beta} + \frac{1}{2} \mu'_{\beta p} \Sigma_{\beta p}^{-1} \mu_{\beta p} \right\}
\]

\[
\cdot p(\omega, \rho, \sigma_f^2)
\]

\[
\propto \sigma_f^{-\frac{1}{2}n} |G|^{-\frac{1}{2}} \sigma_{f p}^{\frac{1}{2}} |G_{\beta p}|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} \sigma_f^{-2} \left[ Y'G^{-1}Y + \mu'_{\beta p} G_{\beta}^{-1} \mu_{\beta p} \right] \right\}
\]

\[
\cdot \exp \left\{ -\frac{1}{2} \sigma_f^{-2} \left[ -\mu'_{\beta p} G_{\beta p}^{-1} \mu_{\beta p} \right] \right\}
\]

\[
\cdot p(\rho) \cdot (\sigma_f^2)^{-(a_f+a_0)-1} \exp \left\{ -\sigma_f^{-2} (b_f + b_0 \exp(-\omega)) \right\} \cdot \exp(-a_0\omega).
\]

Conditional distribution of \(\sigma_f^2\) is

\[
P(\sigma_f^2|Y, \rho, \omega) \propto (\sigma_f^2)^{-a_{fp}-1} \cdot \exp \left\{ -\frac{b_{fp}}{\sigma_f^2} \right\},
\]

where \(a_{fp} = \frac{1}{2}(n-p)+a_f+a_0\) and \(b_{fp} = b_f+b_0\exp(-\omega)+\frac{1}{2} Y'G^{-1}Y + \frac{1}{2} \left[ \mu'_{\beta} G_{\beta}^{-1} \mu_{\beta} - \mu'_{\beta p} G_{\beta p}^{-1} \mu_{\beta p} \right].\)

Thus

\[
(\sigma_f^2|Y, \rho, \omega) \sim IG(a_{fp}, b_{fp}).
\]

Integrating out \(\sigma_f^2\) in (3.2), we obtain the posterior distribution of \((\rho, \omega)\),

\[
p(\rho, \omega|Y) \propto b_{fp}^{-a_{fp}} |G|^{-\frac{1}{2}} |G_{\beta p}|^{\frac{1}{2}} \cdot \exp(-a_0\omega)
\]

\[
\cdot \prod_{k=1}^{q} \left[ f_{\phi} (\exp(-\rho_{bk})) \cdot \exp(-\rho_{bk}) \right].
\]

With (3.4), (3.3) and (3.1), we can sample from the posterior distribution using the algorithm shown in Appendix B.1.
3.4 Links between GP and matching

In this section, we will show i) imputing a missing value using GP results in a normal distribution with mean equal to a weighted sum of the observations; ii) how GP can naturally incorporate subclassification matching; iii) a comparison between results from GP and Mahalanobis matching with calipers; and iv) a comparison between results from GP and propensity score matching.

3.4.1 Estimating potential outcome with GP

Consider the responses in a certain group

\[ Y \sim \text{MVN}(0, \Sigma), \quad (3.5) \]

where \( Y = (Y_0, Y_1, \ldots, Y_n)' \), \( \Sigma = \begin{pmatrix} \sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{pmatrix} = (\sigma_{ij})_{(n+1) \times (n+1)}, \quad i, j = 0, 1, \ldots, n; \) \( Y_0 \) is the missing value to impute, with an observed covariate \( V_0 \); \( \{(Y_i, V_i), i = 1, 2, \ldots, n\} \) are the observations.

Let \( Y_{\text{obs}} = (Y_1, Y_2, \ldots, Y_n)' \), then we have

\[ Y_0 | Y_{\text{obs}} \sim N(\mu_{0p}, \sigma_{0p}^2), \quad (3.6) \]

where \( \mu_{0p} = \Sigma_{01} \Sigma_{11}^{-1} Y_{\text{obs}}, \sigma_{0p}^2 = \sigma_{00} - \Sigma_{01} \Sigma_{11}^{-1} \Sigma_{10} \) and \( \Sigma_{11}^{-1} = |\Sigma_{11}|^{-1} \text{adj}(\Sigma_{11}) \).

The element at \( i^{th} \) row, \( j^{th} \) column of \( \text{adj}(\Sigma_{11}) \) is

\[ \text{adj}(\Sigma_{11})_{ij} = (-1)^{i+j} M_{ji}, \]
where $M_{ji}$ is the determinant of the $(n-1) \times (n-1)$ matrix that results from deleting the $j^{th}$ row and $i^{th}$ column of $\Sigma_{11}$.

Thus $\mu_{0p} = \sum_{j=1}^{n} w_j Y_j$, where $w_j = \frac{1}{|\Sigma_{11}|} \sum_{i=1}^{n} (-1)^{i+j} \sigma_{0i} M_{ji} = \frac{|\Sigma_{11}|}{|\Sigma_{11}|}$, $\Sigma_{11j}$ is the matrix that results from replacing the $j^{th}$ row of $\Sigma_{11}$ with $\Sigma_{01}$.

To study the trend of weight $w_j$ under different conditions, we consider $V$ containing only one covariate, whose value is $V_j = j$ for the $j^{th}$ observation, take $n = 50$ and define the covariance function as in Section 3.3.1. In Figure 3.1, we take $\nu = 0.1, 1, 10, \phi = 0.5^2, 1, 2^2$, and fix $\sigma_f^2 = 1$ since it would not change the weight of the observations given $\nu$ and $\phi$. In each plot, with $\nu$ fixed, near $V_0$, smaller $\phi$ results in more rapid decrease of weight $w_j$ when distance increases. When $|V_j - V_0|$ is large enough, depending on $\phi$, weights of observations fluctuate around zero. That means inputing with GP would give more weights to the observations nearby and is hardly affected by the observations that are far away from it.

Comparing the three plots, we can see that as $\nu$ increases, the maximum of the weights decreases. This is due to the fact that the correlation of $Y$ decrease when $\nu$ increases, which causes more weight assigned to the prior value zero. Noticing that $\nu$ and $\phi$ are estimated in the proposed method, the decreasing rate and the range of observations with obvious nonzero weights are also automatically decided by the data.
3.4.2 GP and subclassification matching

We assume that the data are already subclassified into several partitions by subclassification matching. In subclassification matching, the imputed value is decided only by the observations in the same partition. Accordingly, we set the covariances between data points from different partitions to zero. If we sort the data points according to the partitions, the covariance matrix of the normal distribution will be a block diagonal matrix. Then the imputed value by GP will also only depend on the observations within the same partition.

From here on, we consider imputing one missing value within a certain partition.

After the subclassification matching, we usually impute the missing value by the average of observations in the corresponding treatment group,

$$\hat{Y}_{0(match)} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$ (3.7)
We can modify the covariance function such that \( \sigma_{ij} = \sigma^2 \) for \( i = j \) and \( \sigma_{ij} = \rho \sigma^2 \), \( 0 < \rho < 1 \) for \( i \neq j \). Then (3.5) becomes

\[
\Sigma = \sigma^2 [(1 - \rho) I_{n+1} + \rho J_{n+1}],
\]

and

\[
\Sigma_{01} = \rho \sigma^2 \mathbf{1}_n',
\]

where \( J_{n+1} \) is a \((n + 1) \times (n + 1)\) matrix with all the elements equal to one and \( \mathbf{1}_n \) is a vector with \( n \) elements equal to one.

In this specific case, the inverse of \( \Sigma_{11} \) is

\[
\Sigma_{11}^{-1} = \sigma^{-2} (1 - \rho)^{-1} \left[ I_n - \frac{\rho}{1 - \rho + \rho n} J_n \right].
\]

We then obtain the the mean in (3.6) for imputation

\[
\mu_{0p} = \rho \sigma^2 \mathbf{1}_n' \sigma^{-2} (1 - \rho)^{-1} \left[ I_n - \frac{\rho}{1 - \rho + \rho n} J_n \right] Y_{\text{obs}}
\]

\[
= \frac{\rho m}{1 - \rho + \rho n} \hat{Y}_{0(\text{match})} + \frac{1 - \rho}{1 - \rho + \rho n} 0.
\]

Therefore, using this specific covariance function for GP, the posterior mean of a missing value is a weighted average of \( \hat{Y}_{0(\text{match})} \) and its prior mean zero. When \( \rho \to 0^+ \), \( Y_0 \) becomes less linear correlated with \( Y_{\text{obs}} \) and \( \mu_{0p} \to 0 \). When \( \rho \to 1^- \), \( Y_0 \) becomes more linearly correlated with \( Y_{\text{obs}} \) and \( \mu_{0p} \to \bar{Y} \), i.e., \( \hat{Y}_{0(\text{match})} \). Since (3.7) uses intercept only model for each partition whereas we fit a GP model within each partition, our approach is much more flexible.
3.4.3 GP and Mahalanobis matching

The Mahalanobis matching with caliper defines distance between \( V_i = (v_{1i}, v_{2i}, ..., v_{qi}) \) and \( V_j = (v_{1j}, v_{2j}, ..., v_{qj}) \) as

\[
D_{ij} = \begin{cases} 
\sqrt{(V_i - V_j)' S^{-1} (V_i - V_j)} , & \text{if } |v_{ki} - v_{kj}| < c \cdot s_k \text{ for } k = 1, 2, ..., q, \\
\infty , & \text{otherwise},
\end{cases}
\]

where \( c \) is the caliper, \( V \) is the covariates considered for matching, \( S \) is the sample variance covariance matrix of \( X \), \( s_k \) is the sample standard deviation of covariate \( v_k \).

In section 3.3.1, we defined a similar distance for GP, which can be rewritten as

\[
d_{ij} = \sqrt{(V_i - V_j)' S_{GP}^{-1} (V_i - V_j)},
\]

where \( S_{GP} \) is a diagonal matrix with diagonal elements \( (\phi_1 s_1^2, \phi_2 s_2^2, ..., \phi_q s_q^2)' \). It differs from Mahalanobis distance in two aspects. First, instead of the covariance matrix \( S \), it introduces a scale parameter \( \sqrt{\phi_k} \) for each covariate. Second, it does not require a caliper. Instead, the scale parameter \( \sqrt{\phi_k} \) indicates the range to “match” for covariate \( v_k \), which we have discussed in section 3.3.1. Using this distance with Bayesian GP, the proposed approach decides the optimal range and estimates the treatment effect simultaneously to gain not only an accurate estimate but also more sensible uncertainty of the estimation.

In order to compare the result of Mahalanobis matching with that from the proposed method, we simulate 100 data sets from the following setup with sample size equal to 100.
$x_1, x_2$ are randomly generated from uniform distribution $U(-2, 2)$. Treatment assignment $T_i \sim Ber(\pi_i)$, where $\pi_i = \expit(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$, $\eta_i = -x_1 - x_2$. Outcome model is

$$y_i = 3 + \Delta T_i + x_3^3 + \epsilon_i,$$

where treatment effect $\Delta = 5$ and $\epsilon_i \sim i.i.d. N(0, 1)$. We estimate ATE using Mahalanobis matching with caliper from 0.125 to 1 and the GP approach, respectively, and plot the results in figure 3.2. For Mahalanobis matching, we use R function `Match` in package `Matching` by Sekhon (2008). Figure 3.2 shows that there is a trade-off between accuracy and the uncertainty of the estimates using Mahalanobis matching. It tends to underestimate the ATE when caliper is large. The difference between 95th and 5th percentiles of the estimates drops dramatically when caliper increase from 0.125 to 0.225. From the plot, calipers within $(0.2, 0.4)$ are more preferable in this study. However, comparing with Mahalanobis matching, GP does not need to preset a caliper but scales covariates automatically by estimating parameter $\phi$ to achieves small bias with less uncertainty, which are shown as reference lines. The minimum root mean squared error (RMSE) using matching is 0.353 while GP gets RMSE = 0.250.

### 3.4.4 GP and propensity score matching

To compare with the propensity score matching, we generate a sample with 1000 observations using the setup in section 3.4.3. The propensity score (ps) is estimated with the correct
Figure 3.2: Results of Mahalanobis matching with different calipers. The circles are the average biases of estimates of ATE using Mahalanobis matching with specified calipers. The corresponding vertical lines indicate the ranges between 5th and 95th percentiles of the biases. The horizontal lines are the 5th percentile, average and 95th percentile of the biases of the estimates from GP.

model. The mean of GP consists of intercept and the treatment indicator $T$. The estimate propensity score is the only variable for evaluating the distance in covariance function. Using the $Match$ function mentioned in section 3.4.3, we get $\hat{ATE} = 5.0090$ with standard error 0.2524. The calipers between 0.125 and 1 yield the same result since the PS is quite dense in this case and it is easy to find matched pairs within their neighbors. The proposed
method obtains $\hat{ATE} = 4.9971$ with standard error 0.1937, which are close to the result from propensity score matching but with more accurate estimate, smaller standard error and the true treatment effect is still within one standard error around the estimate.

From the simulation study in previous subsection and the numerical study here, we can see that GP is comparable with matching method in estimating the average treatment effect using either the covariates or the propensity score only.

### 3.5 Simulation study

We modified the simulation setup in Kang and Schafer (2007) to include the treatment effect. $z_1, z_2, z_3, z_4$ are randomly generated from standard normal distribution. Treatment assignment $T_i \sim Ber(\pi_i)$, where $\pi_i = expit(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, $\eta_i = -z_{i1} + 0.5z_{i2} - 0.25z_{i3} - 0.1z_{i4}$.

We take

$$y_i = 210 + \Delta T_i + 27.4z_{i1} + 13.7z_{i2} + 13.7z_{i3} + 13.7z_{i4} + \epsilon_i,$$

where $\epsilon_i \sim i.i.d.N(0, 1)$ and treatment effect $\Delta = 5$. To assess the performances of the methods under misspecification, we use transformations $x_1 = exp(z_{1}/2)$, $x_2 = z_2/(1+exp(z_{1}))+10$, $x_3 = ((z_{1}z_{3})/25 + 0.6)^3$, and $x_4 = (z_2 + z_4 + 20)^2$.

Using this setup, we generate 100 data sets with sample size equal to 100. Results of GP, linear model, IPTW and BART are shown in Table 3.1 to Table 3.4. We consider settings where the true PS can be correctly estimated and the more realistic setting when
PS is estimated using a misspecifed model. For GP and linear model, we consider models with and without adjustment of PS. The root mean squared error (RMSE), bias, rate of the 95% confident interval including the true value ($R_c$), average of the estimated standard error ($SE_{avg}$) and empirical standard error ($SE_{emp}$) for $\Delta$ are reported in each table.

In these tables, $S_{Mean}$ contains the covariates in the mean, which is a linear combination of those covariates. Intercept is always included in the mean and thus neglected in $S_{Mean}$. $S_{Cov}$ is consist of the covariates that included in the covariance function for GP. The correct exposure model uses $z_1, z_2, z_3, z_4$ as covariates in logistic model, while the incorrect exposure model uses $x_1, x_2, x_3, x_4$. “NA” means the corresponding model does not need to estimate propensity score $e$. The estimated coefficient of $T$ is the estimate of ATE. For BART, we use R function bart in package BayesTree to obtain posterior random sample of the potential outcomes and calculate the SATE as in Hill (2012).

Comparing Table 3.1 and Table 3.2, the proposed GP method has an overall advantage over linear model with adjustment for propensity score. When the exposure model is incorrect, the linear model 2.4 has bias -4.146 and RMSE 5.146. While the worst GP method has RMSE 2.405 and absolute bias 1.265 in model 1.4. By including $\{x_1, x_2, x_3, x_4\}$ in the mean, model 1.2 reduces the RMSE to 1.989 and absolute bias to 0.948. The coverage rate is also greatly improved by GP method. When the exposure model is correct, the linear model 2.3, compared to the GP model 1.6, still has larger RMSE and absolute bias.

In Table 3.1, model 1.1, 1.3 and 1.5 are built on $\{z_1, z_2, z_3, z_4\}$ without estimating $e$. 
Model 1.3 has only one covariate $T$ in the mean to estimate the treatment effect and put covariates \{$z_1, z_2, z_3, z_4\}$ in the covariance function. But model 1.5 has the same mean as the true model, the model generating the data, and an overly complicated covariance specified by \{$x_1, x_2, x_3, x_4\}$. They both have relatively good performance, similar to the true model 2.2. For model 1.5, the estimated covariance matrix is close to an identity matrix and the coefficients estimates of GP mean are similar to those of the corresponding linear model 2.2.

As shown in Table 3.3, although the result of IPTW has good coverage rate, but both of its average SE and absolute bias are large even if the exposure model is correctly specified, due to the small sample size.

Comparing our semiparametric method with the nonparametric method BART, model 1.2 vs model 4.2 and model 1.3 vs model 4.1, we can see that the GP method is superior with smaller absolute value of bias and RMSE, and better coverage rate when only \{$x_1, x_2, x_3, x_4\}$ is used.

Looking back at Table 3.1, we can see that the results of model 1.1, 1.3 and model 1.5 are close to the true model 2.2. In model 1.1 and 1.3, \{$z_1, z_2, z_3, z_4\}$ are in the covariance of GP, which results in a very smooth surface for a treatment group and gets great imputations for the unobserved potential outcomes. In model 1.5, the mean of GP is correctly specified in the sense that it has the same mean as the true model, which will lead to an estimate of the covariance matrix approximately diagonal. Thus the estimate of ATE is accurate but with coverage rate slightly less than the nominated level because of the overfitting in
the covariance part. Therefore, using GP with either a correctly specified mean or a good selection of covariates in the covariance function can achieve an accurate estimate of ATE and its confidence interval.

Table 3.1: Performance of GP

<table>
<thead>
<tr>
<th>Model</th>
<th>( S_{\text{Mean}} )</th>
<th>( S_{\text{Cov}} )</th>
<th>Exposure</th>
<th>RMSE</th>
<th>Bias</th>
<th>( R_c )</th>
<th>( SE_{\text{avg}} )</th>
<th>( SE_{\text{emp}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>{T, x_1, x_2, x_3, x_4}</td>
<td>{z_1, z_2, z_3, z_4}</td>
<td>NA</td>
<td>0.249</td>
<td>-0.005</td>
<td>1</td>
<td>0.496</td>
<td>0.25</td>
</tr>
<tr>
<td>1.2</td>
<td>{T, x_1, x_2, x_3, x_4}</td>
<td>{x_1, x_2, x_3, x_4}</td>
<td>NA</td>
<td>1.989</td>
<td>-0.948</td>
<td>0.91</td>
<td>1.674</td>
<td>1.757</td>
</tr>
<tr>
<td>1.3</td>
<td>{T}</td>
<td>{z_1, z_2, z_3, z_4}</td>
<td>NA</td>
<td>0.248</td>
<td>0.006</td>
<td>1</td>
<td>0.926</td>
<td>0.249</td>
</tr>
<tr>
<td>1.4</td>
<td>{T}</td>
<td>{x_1, x_2, x_3, x_4}</td>
<td>NA</td>
<td>2.405</td>
<td>-1.265</td>
<td>0.92</td>
<td>2.159</td>
<td>2.056</td>
</tr>
<tr>
<td>1.5</td>
<td>{T, z_1, z_2, z_3, z_4}</td>
<td>{x_1, x_2, x_3, x_4}</td>
<td>NA</td>
<td>0.23</td>
<td>-0.001</td>
<td>0.92</td>
<td>0.229</td>
<td>0.231</td>
</tr>
<tr>
<td>1.6</td>
<td>{T, e, x_1, x_2, x_3, x_4}</td>
<td>{x_1, x_2, x_3, x_4}</td>
<td>correct</td>
<td>0.861</td>
<td>-0.045</td>
<td>0.99</td>
<td>0.841</td>
<td>0.864</td>
</tr>
<tr>
<td>1.7</td>
<td>{T, e, x_1, x_2, x_3, x_4}</td>
<td>{x_1, x_2, x_3, x_4}</td>
<td>incorrect</td>
<td>1.982</td>
<td>-0.873</td>
<td>0.9</td>
<td>1.681</td>
<td>1.789</td>
</tr>
</tbody>
</table>

Note: \( S_{\text{Mean}} \) contains the covariates in the mean, which is a linear combination of those covariates and intercept. \( S_{\text{Cov}} \) is consist of the covariates that included in the covariance matrix for GP. The correct exposure model uses \( z_1, z_2, z_3, z_4 \) as covariates in logistic model, while the incorrect exposure model uses \( x_1, x_2, x_3, x_4 \). “NA” means the corresponding model does not need propensity score \( e \). On the right-hand side, the root mean squared error (RMSE), bias, rate of the 95% confidence interval including the true value (\( R_c \)), average of the estimated standard error (\( SE_{\text{avg}} \)) and empirical standard error (\( SE_{\text{emp}} \)) for \( \Delta \) are reported.

Table 3.2: Performance of linear model

<table>
<thead>
<tr>
<th>Model</th>
<th>( S_{\text{Mean}} )</th>
<th>Exposure</th>
<th>RMSE</th>
<th>Bias</th>
<th>( R_c )</th>
<th>( SE_{\text{avg}} )</th>
<th>( SE_{\text{emp}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>{T, x_1, x_2, x_3, x_4}</td>
<td>NA</td>
<td>6.442</td>
<td>-5.56</td>
<td>0.67</td>
<td>3.571</td>
<td>3.277</td>
</tr>
<tr>
<td>2.2</td>
<td>{T, z_1, z_2, z_3, z_4}</td>
<td>NA</td>
<td>0.224</td>
<td>0.011</td>
<td>0.96</td>
<td>0.225</td>
<td>0.225</td>
</tr>
<tr>
<td>2.3</td>
<td>{T, e, x_1, x_2, x_3, x_4}</td>
<td>correct</td>
<td>1.258</td>
<td>-0.134</td>
<td>1.00</td>
<td>2.651</td>
<td>1.257</td>
</tr>
<tr>
<td>2.4</td>
<td>{T, e, x_1, x_2, x_3, x_4}</td>
<td>incorrect</td>
<td>5.146</td>
<td>-4.146</td>
<td>0.76</td>
<td>3.433</td>
<td>3.065</td>
</tr>
</tbody>
</table>

Note: See Table 3.1
Table 3.3: Performance of IPTW

<table>
<thead>
<tr>
<th>Model</th>
<th>Exposure</th>
<th>RMSE</th>
<th>Bias</th>
<th>$R_e$</th>
<th>$SE_{avg}$</th>
<th>$SE_{emp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>correct</td>
<td>4.615</td>
<td>-1.296</td>
<td>1.00</td>
<td>7.189</td>
<td>4.452</td>
</tr>
<tr>
<td>3.2</td>
<td>incorrect</td>
<td>6.772</td>
<td>-4.720</td>
<td>0.96</td>
<td>7.160</td>
<td>4.872</td>
</tr>
</tbody>
</table>

Note: See Table 3.1

Table 3.4: Performance of BART

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariates used in BART</th>
<th>RMSE</th>
<th>Bias</th>
<th>$R_e$</th>
<th>$SE_{avg}$</th>
<th>$SE_{emp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>$T, z_1, z_2, z_3, z_4$</td>
<td>0.880</td>
<td>-0.388</td>
<td>0.98</td>
<td>1.069</td>
<td>0.794</td>
</tr>
<tr>
<td>4.2</td>
<td>$T, x_1, x_2, x_3, x_4$</td>
<td>3.097</td>
<td>-2.478</td>
<td>0.79</td>
<td>2.192</td>
<td>1.868</td>
</tr>
</tbody>
</table>

Note: See Table 3.1

### 3.6 Case study

This study utilizes data from a multi-center patient network registry study, the Pediatric Rheumatology Collaborative Outcome Improvement Network (PRCOIN). Patients from 13 geographically diverse participating centers who are receiving cares for Juvenile Idiopathic Arthritis (JIA) and willing to participate in the registry are enrolled in the PRCOIN. The study collected primarily clinical, disease characteristics, physiological, clinical and functional data from each patient encounters. For the purpose of evaluating the effectiveness of early aggressive biological disease modifying anti-rheumatic drugs (DMARDs) vs. non-biological DMARDs, our study is designed to include only those patients who were enrolled into the PRCOIN as new patients (less than 6 months since diagnosis of the disease), and was followed up for at least 6 months after the initial treatment. The initial treatment was determined based on the medication prescribed to the patients at the first visit captured in the registry.
The primary outcome is the clinical Juvenile Arthritis Disease Activity Score (cJADAS) after 6 months of treatment. cJADAS is a disease severity score calculated based on three core clinical measures: MD’s assessment of global disease activity, patient’s self-assessment of overall wellbeing, and number of active joint counts. It ranges between 0 to 30, with 0 indicating no disease activity and 30 maximum disease activity. Patients who were treated by other therapy are excluded from the study in order to ensure equipoise of the two treatment comparison groups. Patients who missed cJADAS at sixth month are also excluded from this analysis. Consequently, only 468 patients are included in this case study. The missing covariates are imputed by predictive mean matching method implemented in the R package MICE.

Four variables are considered based on clinical knowledge and previous experiences. Their baseline characteristics are described in Table 3.5, from which we can see that the baseline characteristics in the two groups are very different.

We put all of the selected variables in the covariance function of GP and assign the linear combination of them and treatment assignment with an intercept as the mean. Variable "JRA subtype" is recoded as two dummy variables. The estimated population average treatment effect is -1.924 with standard error equal to 0.570, and the 95% credible interval is (-3.059, -0.838). The result suggests that 6 month later after initial treatment, the early aggressive biological DMARDs significantly reduces cjadas by nearly 2 points more than non-biological DMARDs. However, in our model, $\hat{\sigma}_f^2 = 3.835$ is much less than $\hat{\sigma}_0^2 = 28.803$, ...
Table 3.5: Baseline characteristics of patients

<table>
<thead>
<tr>
<th></th>
<th>Control (n = 307)</th>
<th>Treated (n = 161)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at baseline, years</td>
<td>10.15±4.8</td>
<td>11.0±4.71</td>
</tr>
<tr>
<td>Positive for RF no. (%)</td>
<td>12.7±33.4</td>
<td>27.3±44.7 †</td>
</tr>
<tr>
<td>cJADAS at baseline</td>
<td>11.31±6.29</td>
<td>12.54±6.21 †</td>
</tr>
<tr>
<td>PPCI Parent Social Support</td>
<td>6.89±3.32</td>
<td>8.04±3.26 †</td>
</tr>
<tr>
<td>JRA subtype (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td>83 (27.0)</td>
<td>66 (41.0)</td>
</tr>
<tr>
<td>Group 2</td>
<td>176 (57.3)</td>
<td>85 (52.8)</td>
</tr>
<tr>
<td>Group 3</td>
<td>48 (15.6)</td>
<td>10 (6.2)</td>
</tr>
</tbody>
</table>

Note: Except where indicated otherwise, values are the mean ± SD; RF = Rheumatoid Factor. PPCI = Pediatric Pain Coping Inventory. JRA = Juvenile Rheumatoid Arthritis. † p-value < 0.01 when conducting a t-test for equal means ‡ p-value < 0.05 when conducting a t-test for equal means

which indicates that the covariates in the covariance part may not contribute much in this case. To verify that, we fit a linear model with the same covariates and obtain the estimate of treatment effect equal to -1.885 with standard error equal to 0.566 and p-value < 0.01, which is similar to the result from GP. BART obtains estimate equal to -1.692 with standard error 0.552. Mahalanobis matching without caliper obtains a similar result to GP, \( \hat{ATE} = -1.907 \) with standard error 0.649.

3.7 Discussion

In this chapter, we propose a Bayesian semiparametric method for estimating the population average treatment effect for non-experimental data, and demonstrate that GP regression
offers a natural way for Bayesian causal inference to address the treatment selection bias.
We demonstrated the close link between GP regression and matching method, then compared it with IPTW, linear model with adjustment for propensity score and BART in a simulation study.

The proposed method utilizes a modified Mahalanobis distance to evaluate the closeness between data points, estimates ATE and the scale parameter for each covariate in the distance simultaneously. The simulation study result suggests GP regression outperforms the other methods in the study. The proposed approach serves to guard against potential model misspecification in two ways: 1) by including the potential outcome mechanism in the mean; 2) by borrowing strength from matching method. Given the corresponding GP, the unobserved potential outcome could be imputed by the weighted average of the observed data. The weights of data points are determined by their proximity in the covariates space. Future work is needed to investigate further its robust properties and compare it with Saarela et al. (2016).

Researchers derived different kinds of priors for GP, for instance, the objective prior in Berger et al. (2001), Kazianka and Pilz (2012) and Ren et al. (2012). Gelfand Gelfand et al. (2005) suggested using uniform prior for $\frac{1}{\phi}$ in a spatial analysis. But we find that using a prior with preference to smooth surface is more suitable for our purpose. Researcher could also blend their knowledge in the prior to obtain a more efficient estimate. Here we only considered squared exponential covariance function, but other types of covariate function
could also be considered, e.g., using different powers for covariates.

Though we focused on a binary treatment assignment scenario, the model and inference in this chapter are also suitable for multivalued or continuous treatment variables by assuming that the treatment effects are constant or linearly correlated with the treatment dosage. A well estimated or known propensity score can also be included as a covariate in the model to improve the accuracy. A drawback of the proposed method is its computational cost, due to inverting a large $n$ by $n$ matrix. There are some literatures about applying the GP to large data, such as Snelson and Ghahramani (2005) and Banerjee et al. (2008), whose methods can be utilized to solve the potential computational problem. Further study is needed for conducting variable selection in the GP model such as selecting covariates that should be included in the distance evaluation.
Chapter 4

Estimating the Heterogeneous Treatment Effects Using Gaussian Process

Many literatures studied the mean treatment effect with observational data but the effect heterogeneity was often ignored, which could be an important question in reality. In this chapter, we propose a Bayesian semiparametric approach for predicting the heterogeneous treatment effect and response of new patient using two conditionally independent Gaussian processes (GP), one for response surface of control group, the other for treatment effect. The prediction can be used to visualize the treatment effect, help researchers investigate the pattern of the treatment effect for different patient characteristics, hence decide whether the treatment is effective for patients with certain characteristics and possibly define a subgroup that the treatment is significantly effective on.
Using the Gaussian process, prior knowledge of the outcome mechanism can be incorporated in the mean as a linear model. The linear model is then automatically corrected by the Gaussian process if necessary. We use Bayesian methodology in this chapter for parameters estimation and inference, which enables us to get the credible interval for mean treatment effect conditioning on covariates affecting the treatment effect.

The result from our numerical study suggests that the proposed approach outperforms BART and random forest in the sense that estimation of the heterogeneous treatment effect is more accurate and better in capturing the pattern of treatment effect when the treatment effect changes smoothly. We also illustrate the proposed method using data obtained from a clinical trial for estimating the effect of calcium modifying the change of total BMD.

4.1 Introduction

In casual inference, most researches focused on estimating the average treatment effect in a population. However, a treatment may not work equally well on each patient. Studying the differential treatment effectiveness caused by observed patient baseline characteristics, i.e., the heterogeneity of treatment effect (HTE), could help in developing personalized therapies for certain patients. A lot of Bayesian subgroup analysis methods were proposed to address this problem for randomized control trials, such as Sivaganesan et al. (2011), Mueller et al. (2010) Berger et al. (2014), Liu et al. (2017) and Sivaganesan et al. (2017). However,
those methods cannot be directly applied to the observational data. And for continuous covariates, subgroup analysis needs to have cutoff values to separate the patients into distinct groups without double checking its validity afterwards. The cutoff values are either pre-specified or determined by some searching algorithm like splitting a node in a tree. The difference between subgroups are checked or tested, but the within-group variations are usually neglected.

Brand and Thomas (2013) discussed causal effect heterogeneity and described key strategies for the study of heterogeneous treatment effects with observational data. A straightforward method is allowing the interaction of treatment indicator and covariates in the model to study how the treatment effects vary across different values of these covariates. Morgan and Todd (2008) used weighted regression to detect the consequential heterogeneity of causal effects. Morgan (2001) used propensity score matching with replacement to investigate the causal effect of Catholic schooling varying across those who attended the alternative school. Some other methods are developed by statistical modeling over the propensity score distribution, e.g., Brand (2010); Xie et al. (2012). Athey and Imbens (2016) proposed a data-driven method to recursively partition the data into subpopulations with different magnitude of treatment effects. They suggested splitting the data into two parts, one for constructing the partition and another for estimating treatment effects for each subpopulation to achieve the "honest estimation" of HTE.

Virtual twins method could also be helpful in investigating the potential treatment effect
heterogeneity. By including the covariates of interest and treatment assignment in random forest (RF), Bayesian additive regression model (BART) or other statistical model, we can impute the unobserved potential outcome and estimate the treatment effect for each individual, e.g., (Hill, 2012). But these tree based models tend to change dramatically in response when there are small changes in covariates, which may conflict with the nature that most treatment effect may change slightly when patients have similar characteristics.

In this chapter, we propose a full Bayesian approach for estimating the HTE with observational data. For subgroup constructed by one or two continuous covariates, the contour map of the predictive treatment effect is helpful in exploring the possible subgroups. We assume that the potential outcome for being in control group and the treatment effect are two independent Gaussian processes conditioning on the covariates constructing the subgroups. The observations are then a realization of the sum of this two Gaussian process. The conditional independence assumption is based on the correct selection of the covariates affecting the treatment effect, which indicates that the treatment effect is determined solely by them and thus independent of the potential outcome. We use priors that prefer smooth surface for Bayesian inference. By using Bayesian inference, we could incorporate prior knowledge to achieve more accurate estimate and obtain credible intervals for treatment effects. As discussed in Chapter 3, the GP for potential outcome have similar effect to matching method on reducing the affect of selection bias. The proposed approach outperforms the tree based methods in the sense that it is better in capturing the pattern of the treatment effect in a
numerical study when the treatment effect changes continuously.

The chapter is organized as follows. In Section 4.2, we introduce the problem of interest and assumptions. In Section 4.3, two conditionally independent Gaussian processes are constructed for the potential outcome for being in control group and the treatment effect, respectively. Prior for parameters, posterior inference and prediction are discussed. In Section 4.4, we present a numerical study to compare the proposed approach with random forest and BART. The proposed method is applied to a real data in Section 4.5. We conclude in Section 4.6 with discussion about the proposed method and future work.

4.2 Notation and assumptions

4.2.1 Notation

In this chapter, we consider an observational data with one treatment and one control group. Using $T_i$ as treatment indicator, $i^{th}$ patient either accepts the treatment, $T_i = 1$, or is in the control group, $T_i = 0$, $i = 1, 2, ..., n$. We denote the potential outcomes under treatment and control by $Y^{(1)}$ and $Y^{(0)}$, respectively, denote the observed outcome by $Y$, unobserved potential outcome by $Y^c$ and treatment effect by $h$. Let $V$ be the collection of covariates. It is of interest to estimate the mean treatment effect given $V$, i.e., $E(h|V) = E (Y^{(1)} - Y^{(0)}|V)$, and predict the treatment effects for new patients. We use
the following notations for the observed data and latent variables when describing models,

\[ T = (T_1, T_2, \ldots, T_n)' , \quad Y^{(0)} = \left( Y^{(0)}_1, \ldots, Y^{(0)}_n \right)' , \quad Y^{(1)} = \left( Y^{(1)}_1, \ldots, Y^{(1)}_n \right)' , \quad Y = (Y_1, \ldots, Y_n)' , \]

\[ Y^c = (Y^c_1, \ldots, Y^c_n)' , \quad h = (h_1, \ldots, h_n)' . \]

In this chapter, we assume that the response variable is continuous.

### 4.2.2 Assumptions

In this chapter, we assume that (i) \( Y^{(0)} = \mu_{Y^{(0)}} + \epsilon_0, \quad h = \mu_h + \epsilon_h, \) where \( \epsilon_0, \epsilon_h \) are random errors, \( \mu_{Y^{(0)}} \) and \( \mu_h \) are two unknown functions of covariates in \( V_b \subseteq V \) and \( V_h \subseteq V \), respectively, where \( V_b \) is the covariate set that insures the unconfoundedness (ideally, it could be covariates affecting outcome regardless of treatment assignment) and \( V_h \) is the covariate set that influences the treatment effects; (ii) Given \( V_h, Y^{(0)} \) and \( h \) are independent, then \( Y^{(1)} = Y^{(0)} + h \). And \( E(h|V) = E(Y^{(1)} - Y^{(0)}|V) = E(h|V_h) = \mu_h \); (iii) For all values of \( V \), the probability assigning an individual to the treatment group is neither zero or one, i.e., \( 0 < P(T = 1|V) < 1 \). For the sake of clarity, in the following sections we would not specify whether a covariate is in \( V_b \) or \( V_h \) when it is self-evident; (iv) Stable unit treatment value assumption (SUTVA) is necessary (Rubin, 1986), which requires the potential outcome of one unit should not be influenced by how the treatment is assigned to this unit and the treatment assignments of other units.
4.3 Gaussian process for estimating HTE

In this section, we introduce a GP model for estimating the heterogeneous treatment effect and predicting the treatment effect for new patients. Gaussian process assumes responses occur in a continuous domain and every finite collection of these points correspond to random variables following a multivariate normal distribution. By specifying the covariance function as a decreasing function of distance, it could borrow information from nearby observation according to their distances to the new data point when performing estimation and prediction.

4.3.1 Model

Consider that the response in control group $Y^{(0)}(\cdot)$ as a Gaussian process with mean $x_{bs} \beta_b$ and covariance function $k_b(s, u) = \sigma^2_b \exp\left(-\left(d_{su}^{(0)}\right)^2\right) + \sigma^2_0 \delta_{su}$, where $x_{bs} \beta_b$ is the linear trend, $x_{bs}$ is the covariates in the linear trend, $\beta_b = (\beta_{b_0}, \beta_{b_1}, ..., \beta_{b_p})'$; $d_{su}^{(0)}$ is the distance between data points $s$ and $u$, $\delta_{su}$ is the Kronecker function, equals 1 if $s = u$, otherwise zero.

And we assume that, given $V$, treatment effect $h(\cdot)$ is another Gaussian process independent of $Y^{(0)}(\cdot)$ with mean $x_{hs} \beta_h$ and covariance function $k_h(v, u) = \sigma^2_h \exp\left(-\left(d_{su}^{(1)}\right)^2\right) + \sigma^2_{h0} \delta_{su}$, where $x_{hs} \beta_h$ is the linear trend, $x_{hs}$ is the covariates in the linear trend, $\beta_h = (\beta_{p_0+1}, \beta_{p_0+2}, ..., \beta_{p_0+p_1})'$; $d_{su}^{(1)}$ is the distance between data points $s$ and $u$. Thus, given $T$, the observed response $Y(\cdot)$ is a Gaussian process with mean $x_s \beta = x_{bs} \beta_b + x_{hs} \beta_h \cdot T$ and
covariance function $k(s,u) = k_b(s,u) + k_h(s,u) \cdot T$, where $\beta = (\beta_0, \beta_1, ..., \beta_p)'$. Then we have

$$Y \sim N \left( X\beta, \Sigma \right), \quad (4.1)$$

where $X\beta = X_b\beta_b + D_zX_h\beta_h$ is the design matrix for the linear trend, $D_z = \text{diag}(T_1, T_2, ..., T_n)$, $\Sigma = \Sigma_b + D_z\Sigma_h D_z$, $\Sigma_b = (\sigma_{ij}^{(b)})_{n \times n}$, $\Sigma_h = (\sigma_{ij}^{(h)})_{n \times n}$, $\sigma_{ij}^{(b)} = \sigma_b^2 \exp \left(-\left(d_{ij}^{(0)}\right)^2\right) + \sigma_0^2 \delta_{ij}$, $\sigma_{ij}^{(h)} = \sigma_h^2 \exp \left(-\left(d_{ij}^{(1)}\right)^2\right) + \sigma_{h0}^2 \delta_{ij}$, $i,j = 1, 2, ..., n$. Without loss of generality, $X$ is assumed full rank.

We define the distances $d_{ij}^{(0)} = \sqrt{\sum_{k=1}^{q} \phi_{bk}^{-1} \left(s_k^{(b)}\right)^{-2} \left(v_{ki}^{(b)} - v_{kj}^{(b)}\right)^2}$ for $Y^{(0)}(\cdot)$ and $d_{ij}^{(1)} = \sqrt{\sum_{k=1}^{q} \phi_{hk}^{-1} \left(s_k^{(h)}\right)^{-2} \left(v_{ki}^{(h)} - v_{kj}^{(h)}\right)^2}$ for $h(\cdot)$, where $v_k^{(l)} = \{v_{k1}^{(l)}, v_{k2}^{(l)}, ..., v_{kn}^{(l)}\}$, $k = 1, 2, ..., q^{(l)}$, $l = b, h$, are the covariates in $V_l$, $s_k^{(l)}$ is the corresponding standard deviation. Note that the covariates in $V_b$ and $V_h$ can be different, indicating that the covariates affecting the potential outcome $Y^{(0)}$ and treatment effect $h$ are different. The covariates $v_k^{(l)}$'s in the distances are standardized so that the $\phi_{lk}$'s are comparable and can be assigned the same prior distribution.

Large $\phi_{bk}$ and $\phi_{bk}$ will result in a smooth change of response surface when the corresponding covariate changes. Otherwise, the response surface may become rough. Large ratio $\frac{\sigma_2^2}{\sigma_j^2}$ indicates that the random error or the variation unexplained by the model is large, the response surface will also be smooth. And we suggest using a regression model reflecting the outcome mechanism in the means of $Y^{(0)}(\cdot)$ and $h(\cdot)$ to take advantage of the prior
knowledge.

4.3.2 Prior

The following priors are assigned to the parameters in the model.

\[
\beta \sim MVN(\mu_\beta, \sigma^2_\beta G_\beta),
\]

\[
\sigma^2_f \sim IG(a_f, b_f),
\]

\[
\sigma^2 \sim IG(a_0, b_0),
\]

\[
\sigma^2_{hf} \sim IG(a_{hf}, b_{hf}),
\]

\[
\sigma^2_h \sim IG(a_{h0}, b_{h0}),
\]

\[
\phi_{bk} \sim \text{i.i.d. } f_\phi(\phi_{bk}), k = 1, ..., q,
\]

\[
\phi_{hk} \sim \text{i.i.d. } f_\phi(\phi_{hk}), k = 1, ..., q_h,
\]

where \( f_\phi \) is the probability density function of inverse exponential distribution with rate equal to one, i.e. \( IG(1, 1) \). We take \( G_\beta = g(X'X)^{-1}, g = 10^6, a_0 = a_f = a_{hf} = 2, b_f = b_0 = \frac{\hat{\sigma}^2_{lm0}}{2} \) and \( b_{hf} = b_{h0} = \frac{\hat{\sigma}^2_{lmh}}{4} \), where \( \hat{\sigma}^2_{lm0} \) is the MSE when fitting model \( Y \sim N(X\beta, I\sigma^2_{lm0}) \) with data in the control group and \( \hat{\sigma}^2_{lmh} \) is the MSE when fitting model \( Y \sim N(X\beta, I\sigma^2_{lmh}) \) with data in the treatment group.

For the ease of computation, we use the following transformations,

\[
\nu = \frac{\sigma^2}{\sigma^2_f}, \nu_h = \frac{\sigma^2_h}{\sigma^2_f}, \nu_{hf} = \frac{\sigma^2_{hf}}{\sigma^2_f},
\]

91
\[ \omega = \log(\nu), \omega_h = \log(\nu_h), \omega_f = \log(\nu_f), \]  \quad (4.10)

\[ \rho_k = -\log(\phi_{bk}), \rho_{hk} = -\log(\phi_{hk}). \]  \quad (4.11)

Let \( \phi = (\phi_{b1}, ..., \phi_{bq})', \phi_h = (\phi_{h1}, ..., \phi_{hq})', \rho_b = (\rho_{b1}, \rho_{b2}, ..., \rho_{bq})',\rho_h = (\rho_{h1}, \rho_{h2}, ..., \rho_{hq})' \), \( \rho = (\rho_b', \rho_h')' \) and \( \omega = (\omega, \omega_h, \omega_f)' \).

Since the priors for \( (\phi, \sigma_0^2, \sigma_f^2) \) and \( (\phi_h, \sigma_h^2, \sigma_{hf}^2) \) are proper, we can obtain the prior distribution for \( (\omega, \rho, \sigma_f^2) \) as follows.

\[
p(\omega, \rho, \sigma_f^2) \propto p(\rho) \cdot p(\omega, \sigma_f^2)
\]

\[
p(\rho) \propto \prod_{k=1}^{q_0} \left[ f_\phi(\exp(-\rho_{bk})) \cdot \exp(-\rho_{bk}) \right] \prod_{k=1}^{q_1} \left[ f_\phi(\exp(-\rho_{hk})) \cdot \exp(-\rho_{hk}) \right]
\]

\[
p(\omega, \sigma_f^2) \propto (\sigma_f^2)^{-a_f-1} \exp \left\{ -\frac{b_f}{\sigma_f^2} \right\} \cdot (\sigma_f^2 \nu_f)^{-a_0-1} \exp \left\{ -\frac{b_0}{\sigma_f^2 \nu_f} \right\}
\]

\[
\cdot (\sigma_f^2 \nu_h)^{-a_f-1} \exp \left\{ -\frac{b_f}{\sigma_f^2 \nu_h} \right\} \cdot (\sigma_f^2 \nu_f)^{-a_0-1} \exp \left\{ -\frac{b_0}{\sigma_f^2 \nu_f} \right\}
\]

\[
\cdot (\sigma_f^2)^3 \exp(\omega + \omega_h + \omega_f)
\]

\[
\propto p(\sigma_f^2 | \omega) \cdot p(\omega),
\]

where \( p(\sigma_f^2 | \omega) \) is the pdf of \( IG(a_{fa}, b_{fa}) \), \( a_{fa} = a_f + a_0 + a_f + a_0, b_{fa} = b_f + \frac{b_0}{\nu} + \frac{b_f}{\nu_f} + \frac{b_0}{\nu_h} \), and \( p(\omega) = \left( \frac{b_{fa}}{\Gamma(a_{fa})} \right) \left( \frac{b_f}{\Gamma(a_f)} \right) \left( \frac{b_{fa}}{\Gamma(a_{fa})} \right) \cdot \exp(-a_0\omega - a_0\omega_h - a_f\omega_f) \).
4.3.3 Posterior inference

Multiplying likelihood by priors, we have

\[ L = p(Y, \beta, \sigma_f^2, \rho, \omega) \propto p(Y|\beta, \Sigma_h, \Sigma_b) \cdot p(\beta) \cdot p(\omega, \rho, \sigma_f^2), \]  

(4.12)

where

\[ p(Y|\beta, \Sigma_h, \Sigma_b) \propto |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (Y - X\beta)'\Sigma^{-1}(Y - X\beta) \right\}. \]  

(4.13)

Full conditional distribution of \( \beta \) is

\[ p(\beta|Y, \Sigma) \propto \exp \left\{ -\frac{1}{2} \left[ (Y - X\beta)'\Sigma^{-1}(Y - X\beta) + (\beta - \mu_{\beta_p})'\sigma_f^{-2}G_{\beta}^{-1}(\beta - \mu_{\beta_p}) \right] \right\} \]
\[ \propto \exp \left\{ -\frac{1}{2} (\beta - \mu_{\beta_p})'\Sigma_{\beta_p}^{-1}(\beta - \mu_{\beta_p}) \right\}, \]

where \( \Sigma_{\beta_p} = (X'S^{-1}X + \sigma_f^{-2}G_{\beta}^{-1})^{-1} \) and \( \mu_{\beta_p} = \Sigma_{\beta_p}(X'S^{-1}Y + \sigma_f^{-2}G_{\beta}^{-1}\mu_\beta) \). Thus

\[ (\beta|Y, \Sigma) \sim MVN(\mu_{\beta_p}, \Sigma_{\beta_p}). \]  

(4.14)

Note that if we assign flat prior for \( \beta \), we would get \( \Sigma_{\beta_p} = (X'S^{-1}X)^{-1} \) and \( \mu_{\beta_p} = \Sigma_{\beta_p}X'S^{-1}Y \). That is setting \( \Sigma_{\beta}^{-1} = 0 \) and \( \mu_{\beta} \) finite.

Let \( G = \sigma_f^{-2}\Sigma \). Integrating out \( \beta \), we get the marginal posterior distribution of \((\omega, \rho, \sigma_f^2)\),

\[ p(\omega, \rho, \sigma_f^2|Y) \propto \int_{-\infty}^{+\infty} p(Y, \beta, \sigma_f^2, \rho, \omega) d\beta \]
\[ \propto |\Sigma|^{-\frac{1}{2}} |\Sigma_{\beta_p}|^{-\frac{3}{2}} \exp \left\{ -\frac{1}{2} Y'S^{-1}Y - \frac{1}{2} \sigma_f^{-2}\mu_{\beta_p}'\Sigma_{\beta_p}^{-1}\mu_{\beta_p} + \frac{1}{2} \mu_{\beta_p}'\Sigma_{\beta_p}^{-1}\mu_{\beta_p} \right\} \]
\[ \cdot p(\omega, \rho, \sigma_f^2), \]
\[ p(\omega, \rho, \sigma_f^2 | Y) \propto \sigma_f^{-\frac{1}{2}n} |G|^{-\frac{1}{2} \frac{1}{2}} |X'G^{-1}X + \Sigma_{\beta}^{-1}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2\sigma_f^2} \left[ Y'G^{-1}Y + \mu_{\beta}'\Sigma_{\beta}^{-1}\mu_{\beta} \right] \right\} \]

\[ \exp \left\{ -\frac{1}{2\sigma_f^2} \left[ - (X'G^{-1}Y + \Sigma_{\beta}^{-1}\mu_{\beta})' (X'G^{-1}X + \Sigma_{\beta}^{-1})^{-1} (X'G^{-1}Y + \Sigma_{\beta}^{-1}\mu_{\beta}) \right] \right\} \]

\[ \cdot p(\rho) \cdot (\sigma_f^2)^{-\left( a_f + a_0 + a_f + a_0 - 1 \right)} \exp \left\{ -\sigma_f^{-2} \left( b_f + \frac{b_0}{\nu} + \frac{b_f}{\nu_f} \right) \right\} \]

\[ \cdot \exp \left\{ -a_0\omega - a_0\omega_h - a_f\omega_f \right\}. \]

Thus, conditional distribution of \( \sigma_f^2 \),

\[ p(\sigma_f^2 | Y, \rho, \omega) \propto (\sigma_f^2)^{-a_{fp} - 1} \cdot \exp \left\{ -\frac{b_{fp}}{\sigma_f^2} \right\}, \]

where \( a_{fp} = \frac{1}{2}(n - p) + a_f + a_0 + a_f + a_0 \), \( b_{fp} = b_f + \frac{b_0}{\nu} + \frac{b_f}{\nu_f} + \frac{b_0}{\nu_h} + \frac{1}{2}A \) and \( A = \left[ Y'G^{-1}Y + \mu_{\beta}'\Sigma_{\beta}^{-1}\mu_{\beta} - (X'G^{-1}Y + \Sigma_{\beta}^{-1}\mu_{\beta})' (X'G^{-1}X + \Sigma_{\beta}^{-1})^{-1} (X'G^{-1}Y + \Sigma_{\beta}^{-1}\mu_{\beta}) \right] \).

Therefore

\[ (\sigma_f^2 | Y, \rho, \omega) \sim IG(a_{fp}, b_{fp}). \quad (4.15) \]

And posterior distribution of \((\rho, \omega)\),

\[ p(\rho, \omega | Y) \propto b_{fp}^{-a_{fp}} |G|^{-\frac{1}{2} \frac{1}{2}} |X'G^{-1}X + \Sigma_{\beta}^{-1}|^{-\frac{1}{2}} \]

\[ \cdot \exp \left\{ -a_0\omega - a_0\omega_h - a_f\omega_f \right\} \]

\[ \cdot \prod_{k=1}^{q_0} [f_\phi(\exp(-\rho_{bk}) \cdot \exp(-\rho_{bk})) \cdot \prod_{k=1}^{q_1} [f_\phi(\exp(-\rho_{hk}) \cdot \exp(-\rho_{hk})]. \]

With (4.16), (4.15) and (4.14), we can sample from the posterior distribution \( p(\beta, \omega, \rho, \sigma_f^2 | Y) \) using the algorithm shown in Appendix C.1.
4.3.4 Prediction

Given new input $x^*, v^*, T^*$ with $n^*$ observations, predictions of the treatment effect $h^*$ and the outcome $Y^*$ are of interest.

**Prediction of treatment effect $h^*$**

Since $h(\cdot)$ is a Gaussian process, we have

\[
\begin{pmatrix}
  h \\
  h^*
\end{pmatrix}
\sim N
\begin{pmatrix}
  \left( X_h \right) \\
  X^*_h
\end{pmatrix}
\begin{pmatrix}
  \beta_h, \Sigma_{hh^*}^{-1}
\end{pmatrix},
\]

where the covariance is calculated by the covariance function $k_h(\cdot, \cdot)$.

Given $X, Y$ and posterior sample of $(\beta, \sigma^2_f, \rho, \omega)$,

\[
p\left( \begin{pmatrix}
  h \\
  h^*
\end{pmatrix} \right) | Y, \beta, \sigma^2_f, \rho, \omega \propto \exp \left\{ -\frac{1}{2} \left[ (Y - X^*_b \beta_h - D_z h)^T \Sigma_{b}^{-1} (Y - X^*_b \beta_h - D_z h) \right] \right\}
\]

\[
\times \exp \left\{ -\frac{1}{2} \left[ \begin{pmatrix}
  h - X_h \beta_h \\
  h^* - X^*_h \beta_h
\end{pmatrix}^T \Sigma_{hh^*}^{-1} \begin{pmatrix}
  h - X_h \beta_h \\
  h^* - X^*_h \beta_h
\end{pmatrix} \right] \right\}
\]

\[
\times \exp \left\{ -\frac{1}{2} \left[ \begin{pmatrix}
  h \\
  h^*
\end{pmatrix} - \mu_{hp} \right]^T \Sigma_{hp}^{-1} \begin{pmatrix}
  h \\
  h^*
\end{pmatrix} - \mu_{hp} \right\},
\]

where $\Sigma_{hp} = \left( \Sigma_{hh^*}^{-1} + D_z \Sigma_{b}^{-1} D_z \begin{pmatrix} 0 & 0 \end{pmatrix} \right)^{-1}$ and $\mu_{hp} = \Sigma_{hp} \begin{pmatrix} D_z \Sigma_{b}^{-1} (Y - X \beta) \end{pmatrix} + \begin{pmatrix} X_h \beta_h \\
X^*_h \beta_h \end{pmatrix}$. Thus.
\[
\begin{pmatrix}
  h \\
  h^*
\end{pmatrix}
| Y, \beta, \sigma_f^2, \rho, \omega \sim N(\mu_{hp}, \Sigma_{hp}).
\] (4.17)

Partition matrix \(\Sigma_{hp}\) into
\[
\begin{pmatrix}
  \Sigma_{hp}^{(11)} & \Sigma_{hp}^{(12)} \\
  \Sigma_{hp}^{(21)} & \Sigma_{hp}^{(22)}
\end{pmatrix},
\]
where \(\Sigma_{hp}^{(11)}\) is a \(n\) by \(n\) matrix. Then we have
\[
\mu_{hp} = \begin{pmatrix}
  \Sigma_{hp}^{(11)} \\
  \Sigma_{hp}^{(21)}
\end{pmatrix} \begin{pmatrix}
  D_{z} \Sigma_{b}^{-1} (Y - X\beta) + X_h \beta_h \\
  X_{h*} \beta_h
\end{pmatrix}.
\]

Posterior distribution of the treatment effect is
\[
(h|Y, \beta, \sigma_f^2, \rho, \omega) \sim N(\mu_h, \Sigma_h),
\] (4.18)
where \(\Sigma_h = \Sigma_{hp}^{(11)}\) and \(\mu_h = \Sigma_{hp}^{(11)} D_{z} \Sigma_{b}^{-1} (Y - X\beta) + X_h \beta_h\).

And posterior predictive distribution of the treatment effect is
\[
(h^*|Y, \beta, \sigma_f^2, \rho, \omega) \sim N(\mu_{h*}, \Sigma_{h*}),
\] (4.19)
where \(\Sigma_{h*} = \Sigma_{hp}^{(22)}\), \(\mu_{h*} = \Sigma_{hp}^{(21)} D_{z} \Sigma_{b}^{-1} (Y - X\beta) + X_{h*} \beta_h\).

In the following sections, when we predict the treatment effect, we are actually conducting posterior inference on the mean treatment effect \(\mu_{h*}\).

**Prediction of outcome \(Y^*\)**

Although our main interest is \(h^*\), it is also possible to provide prediction of \(Y^*\) directly.
Since $Y(\cdot)$ is a Gaussian process, we have

$$
\begin{pmatrix}
Y \\
Y^* 
\end{pmatrix}
|\beta, \sigma_f^2, \rho, \omega \sim N
\begin{bmatrix}
\begin{pmatrix}
X\beta \\
X^*\beta
\end{pmatrix},
\begin{pmatrix}
\Sigma^{(11)} & \Sigma^{(12)} \\
\Sigma^{(21)} & \Sigma^{(22)}
\end{pmatrix}
\end{bmatrix},
$$

where the covariance is calculated by the covariance function $k(\cdot, \cdot)$ and $\Sigma^{(11)}$ is a $n$ by $n$ matrix.

Thus

$$
Y^*|Y, \beta, \sigma_f^2, \rho, \omega \sim N (\mu_{Y^*}, \Sigma_{Y^*}), \quad (4.20)
$$

where $\Sigma_{Y^*} = \Sigma^{(22)} - \Sigma^{(21)} \left(\Sigma^{(11)}\right)^{-1} \Sigma^{(12)}$ and $\mu_{Y^*} = X^*\beta + \Sigma^{(21)} \left(\Sigma^{(11)}\right)^{-1} (Y - X\beta)$.

### 4.4 Numerical study

In this section, we evaluate the performance of the proposed method using two different settings and compare the results with those from BART and random forest.

#### 4.4.1 Smooth surface

This simulation includes two covariates $x_1$ and $x_2$, which are randomly sampled from uniform distribution $U(-2, 2)$. Treatment assignment $T \sim Bernoulli(e)$, where $e = 1/ (1 + exp (-x_2))$.

Treatment effect $h = 0$ if $x_1 \leq 0$; $h = x_1 (4 - x_2^2)$ if $x_1 > 0$, such that there is no effect when $x_1 \leq 0$; the treatment effect is positive and becomes larger when $x_1$ increases above zero and
patient with $x_2 = 0$ obtains greatest treatment effect when $x_1$ is fixed. The treatment effect vs covariates $(x_1, x_2)$ is plotted in figure 4.1.

![Figure 4.1: True treatment effect](image)

And we construct a non-constant response surface for patients in control group, $Y^{(0)} = 50 + 4 \exp(x_1/4) + 3 x_2 + \epsilon$, where random error $\epsilon \sim N(0, 1)$. Thus the true outcome model is

$$Y = 50 + 4 \exp(x_1/4) + 3 x_2 + Tx_1(4 - x_2^2) \cdot I(x_1 > 0) + \epsilon. \quad (4.21)$$

We simulate a random sample with sample size $= 200$ from this model and predict the treatment effect within the range $-2 < x_k < 2, k = 1, 2$ using different methods. The scatter plot of $(x_1, x_2)$ is shown in figure 4.2. The triangles represent the patients assigned to the treatment group, while the circles are patients in the control group. The patients with larger $x_2$ are more likely to be assigned in the treatment group.
For the proposed method, we use only an intercept in each mean and put both $x_1$ and $x_2$ in each covariance functions for $Y^{(0)}(\cdot)$ and $h(\cdot)$. We utilizes the function `bartMachine` in R package bartMachine to implement the BART methods with $x_1$, $x_2$ and treatment assignment $T$ as predictors and outcome $Y$ as dependent variable, and then get the prediction of $Y^{(1)}$ and $Y^{(0)}$ by setting $T = 1$ and 0. The treatment effect is predicted by $h = Y^{(1)} - Y^{(0)}$. And similarly, we utilizes the function `randomForest` in R package randomForest to implement the third method. We plot the predictive treatment effect in figure 4.3. Although all three plots show a correct trend of the treatment effect, it is clear that the the proposed method recovers the true treatment effect in figure 4.1 best. Due to the nature of tree, BART and random forest both result in rough surfaces and cannot predict the smooth
treatment effect well. All the predictions from these three methods are not very accurate near the boundaries, i.e. areas near \(((x_1, x_2)|x_1 \in \{-2, 2\} \text{ or } x_2 \in \{-2, 2\}\)}, since there is less information around the boundaries than points in the center.
Figure 4.3: Predictive treatment effect vs \((x_1, x_2)\)
4.4.2 Stepwise treatment effect

We simulate a covariate $x_1$ randomly sampled from $Uniform(-2, 2)$. Treatment assignment $T \sim Bernoulli(e)$, where $e = 1/(1 + \exp(-x_1))$. Treatment effect is a step function of $x_1$:

- $h = 0$ if $x_1 < 0$;
- $h = 4$ if $0 \leq x_1 < 1$;
- $h = 8$ if $x_1 \geq 1$.

We construct a non-constant response surface for patients in control group, $Y^{(0)} = 50 + 4\exp(x_1/4) + \epsilon$, where random error $\epsilon \sim N(0, 1)$. Thus the true outcome model is

$$Y = 50 + 4\exp(x_1/4) + T \cdot (4I(x_1 \geq 0) + 4I(x_1 \geq 1)) + \epsilon. \quad (4.22)$$

We simulate a random sample with sample size = 200 from this model and predict the treatment effect within the range $-2 < x_1 < 2$ using different methods. To measure the error of the predictions, we use the following criteria.

$$err = \int_{-2}^{2} \left| \hat{h} - h \right| dx_1,$$

where $h$ is the true treatment effect and $\hat{h}$ is the estimate. We have $err_{GP} \approx 2.09$, $err_{BART} \approx 1.34$, $err_{RF} \approx 5.10$, which indicates that the BART has the best performance and GP is better than random forest. From the result in figure 4.4, we can see that the proposed model can still correctly predict the trend of the treatment effect but the step function is estimated as a smooth function. Both BART and random forest have better performance in finding out the change points, but the predicted lines are rough.
4.5 Case study

This section utilizes data from a clinical study for estimating the patient’s improvement of total BMD (Bone Mass Density) after taking calcium. After deleting the data with only one visit, there are 171 patients left in the study. We use $Y$, the difference of TotBMD (Total BMD) between last and first visit, as response variable. Variable $Time$ is the difference of Age between last and first visit. PcntFat is the percentage of fat in a patient’s body. LspBMD is the Lumbar Spine BMD. "wtpcnt" stands for weight percentile.

We use only an intercept in each mean of $Y^{(0)}(\cdot)$ and $h(\cdot)$, put covariates $Time$, LspBMD, Age, Heightcm, PcntFat, BMI, wtpcnt and TotBMD in the covariance function $k_0$, put Time
and PcntFat in the covariance function $k_h$ to evaluate how these two covariates influence the treatment effect. The prediction and 95% Upper limit of the mean treatment effect given Time and PcntFat are plotted in Figure 4.5. It clearly shows that the treatment is more possible to work for patients with $25 < \text{PcntFat} < 35$ and is significantly effective when $25 < \text{PcntFat} < 35$ and $Time \in (1.6, 2.5) \cup (0.5, 1)$. The mean treatment effect is maximized around $Time = 2$, $\text{PcntFat} = 30$. We further check the distribution of $(Time, \text{PcntFat})$ in Figure 4.6, find that most of the patients have Time around 2. Therefore, the above results indicate that patients with $\text{PcntFat} \in (25, 35)$ are benefited more from the treatment than others. But this conclusion is not very reliable when $Time < 1.6$ since there are little data in this region to support it.
(a) Prediction of the mean treatment effect given Time and PcntFat

(b) Upper limit of the mean treatment effect given Time and PcntFat

Figure 4.5: Predictive treatment effect vs Time and LspBMD
4.6 Discussion

In this chapter, we propose a Bayesian semiparametric method for estimating and predicting the heterogeneous treatment effect with observational data. The new method is compared with BART and random forest in numerical studies. And we also illustrate the method with data for studying the effect of calcium in improving the total BMD.

The proposed method utilizes a modified Mahalanobis distance to evaluate the closeness between data points and estimates conditional mean treatment effect for patients with certain characteristics. Intuitively, it impute the missing potential outcome $Y^{(0)}$ and $Y^{(1)}$ using Gaussian processes $Y^{(0)}(\cdot)$ and $Y^{(0)}(\cdot) + h(\cdot)$. It outperforms the other methods in the numerical study when the response surface is smooth, in the sense that the estimation of
the heterogeneous treatment effect is more accurate and its pattern is more similar to the true model than BART and random forest. When the response surface is not smooth, the proposed method is still able to obtain a reasonable result that capture the trend of treatment effect correctly, but not as good as BART or random forest at finding out the change points.

As mentioned in Chapter 3, there are several priors proposed for parameters in Gaussian process with a nugget term. We choose to use the prior in section 4.3.2 to indicate our preference to smooth surface. Prior knowledge could be used to obtain a more informative prior to achieve a better estimate of the treatment effect. And other types of covariance function could also be considered according to the data structure and research objective.

The model and inference in this chapter could be extended to binary outcome by introducing a link function. Continuous treatment is also possible by assuming that the treatment effects and response surface of control group are conditionally independent. A well estimated or known propensity score could be helpful in improving the accuracy. A common disadvantage of using GP is its high computational cost when inverting a $n$ by $n$ matrix. Readers could refer to Snelson and Ghahramani (2005) and Banerjee et al. (2008) for approximate methods.
Chapter 5

Summary and Future Work

In this thesis, we developed novel Bayesian methods for studying the treatment effect in clinical data and observational data.

In Chapter 2, we proposed a Bayesian model selection approach to subgroup analysis when subgroups are defined based on 2 or more covariates. We used a collection of ANOVA models that included models with all the covariates of interest for subgroup analysis simultaneously, extending an earlier model selection based approach which considered models with one covariate at a time. Our approach is based on a premise that, investigators can a priori specify a set of covariates for which subgroup analysis is of interest, and that, in the interest of parsimony, interpretability and subpopulation size, the investigators’ primary interest and goal is to determine whether there is a subgroup effect associated with any single covariate. It is only in the scenario that there is no evidence of subgroup effect based on a single covari-
ate, there is interest in finding whether there is subgroup effect based on an interaction effect involving more than one covariate. We have explored ways to avoid the use of unidentifiable parameters by using the constraints commonly used in ANOVA models, and thus avoid the need to consider equivalent models and model indices. But, we have run into difficulties in specifying priors for the resulting parameters that are suitable for testing for subgroup effects configurations such as equality of effects at certain levels of a covariate and unequal effects at some other levels. Also, taking an approach of using a Bayesian variable selection and picking a best model to determine subgroup effects, we believe, would encounter difficulties in choosing a suitable prior for the parameters. It would also not accommodate the priority ordering among subgroups for reporting that we seek, such as giving priority to subgroup effects based on main effects over reporting an interaction effect. In this chapter we use the stepwise algorithm and posterior probabilities of the models to achieve this. We will continue to explore ways to simplify the approach while achieving the desired goals.

In Chapter 3, we proposed a Bayesian semiparametric method for estimating the population average treatment effect for non-experimental data, and demonstrated that GP regression offers a natural way for Bayesian causal inference to address the treatment selection bias. The close link between GP regression and matching method is demonstrated. The proposed method utilizes a modified Mahalanobis distance to evaluate the closeness between data points and estimates ATE and the scale parameter for each covariate in the distance simultaneously. Other types of covariance function could be explored depending on the data
structure and the question of interest. We focused on a binary treatment assignment scenario in this chapter. We will continue to extend this model to multivalued and continuous treatment variables by assuming that the treatment effects are constant or linearly correlated with the dosage. A drawback of the proposed method is its expensive computation, especially for inverting a large $n \times n$ matrix, which could be improved by consulting literatures about applying the GP to large data, such as Snelson and Ghahramani (2005) and Banerjee et al. (2008). Further study is needed to do variable selection in the GP model.

In Chapter 4, we propose a Bayesian semiparametric method for estimating and predicting the heterogeneous treatment effect with observational data. The proposed method utilizes a modified Mahalanobis distance to evaluate the closeness between data points and estimates conditional mean treatment effect for patients with certain characteristics. Intuitively, it impute the missing potential outcomes using Gaussian processes $Y^{(0)}(\cdot)$ and $Y^{(0)}(\cdot) + h(\cdot)$. According to the simulation study result, it outperforms the other methods in the numerical study when the response surface is smooth, in the sense that the estimation of the heterogeneous treatment effect is more accurate and its pattern is more similar to the true model than BART and random forest. When the response surface is not smooth, the proposed method is still able to obtain a reasonable result that captures the trend of treatment effect, although not as good as BART or random forest at finding out the change points. Similar to Chapter 3, the model and inference in this chapter could also be extended to multi-valued treatment. And computational expense could be reduced.
Bibliography


Brand JE (2010) Civic returns to higher education: A note on heterogeneous effects. Social forces; a scientific medium of social study and interpretation 89(2):417


Rubin DB (1973) Matching to remove bias in observational studies. Biometrics pp 159–183


Appendix A

Subgroup Analysis

A.1 Computing the marginal likelihoods under different models

Here, we derive the formulas used to calculate the marginal likelihood under each model, as in (2.24) and (2.25), while skipping most of the algebraic details. We recall that $y_{k\ell m}$’s are iid $N(\mu_{k\ell}, \sigma_{k\ell}^2)$, and are independent of each other for $t = 0$ (control) and $t = 1$ (treatment), and $\theta_{k\ell} = \mu_{k\ell}^{(1)} - \mu_{k\ell}^{(0)}$. Let $y_{k\ell}^{(t)} = (y_{k\ell1},...,y_{k\ell n_{k\ell}}^{(t)})$ be the observations in cell $(k, \ell)$ under $t = 0, 1$. Using the priors in Section 2.4.2, and integrating the likelihood of $(y_{k\ell}^{(0)}, y_{k\ell}^{(1)})$ over $\mu_{k\ell}^{(0)}$, the corresponding marginal likelihood, denoted by $L_{k\ell}$, is given by

$$L_{k\ell} = S_{k\ell} \cdot exp \left\{ -\frac{(h_{k\ell} - \theta_{k\ell})^2}{2\sigma_{k\ell}^2} \right\}.$$
where \( S_{k\ell} = \left(2\pi\sigma^2_{k(0)} \right)^{-\frac{1}{2}} \left(n^{(0)}_{k\ell} - 1 \right) \left(2\pi\sigma^2_{k(1)} \right)^{-\frac{1}{2}} \left(n^{(1)}_{k\ell} - 1 \right) \left(n^{(0)}_{k\ell} n^{(1)}_{k\ell} \right)^{-\frac{1}{2}} \left(2\pi\sigma^2_{k\ell} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (n^{(0)}_{k\ell} - 1)^2 \sigma^2_{k(0)} - \frac{(n^{(1)}_{k\ell} - 1)^2 \sigma^2_{k(1)}}{2\sigma^2_{k\ell}} \right\}, \) \( h_{k\ell} = y_{k\ell}^{(1)} - y_{k\ell}^{(0)}, \) and \( \sigma^2_{k\ell} = \frac{\sigma^2_{k(1)}}{n^{(1)}_{k\ell}} + \frac{\sigma^2_{k(0)}}{n^{(0)}_{k\ell}}. \)

Marginal likelihood under a model \( M_j \neq M_0: \)

Letting \( h = (h_{11}, h_{12}, h_{21}, h_{22})' \), the marginal likelihood the full data \( Y \) under model \( M \) is

\[
P(Y|M_j) = \int \int \prod_{k,\ell} L_{k\ell} \cdot p(y_j|Y)p(g)p(\sigma^2_{(0)})p(\sigma^2_{(1)})d\psi_j d\sigma^2_{(0)} d\sigma^2_{(1)} dg
\]

Using the priors in Section 2.4.2, after some algebra, we get

\[
P(Y|M_j) = \int \int \left( \prod_{k,\ell} Q_{k\ell} \right) \cdot A \cdot B_j d\sigma^2_{(0)} d\sigma^2_{(1)} dg
\]

where \( Q_{k\ell} = IG(\sigma^2_{k(0)} : a_{k(0)}, b_{k(0)})IG(\sigma^2_{k(1)} : a_{k(1)}, b_{k(1)}) \cdot (\sigma^2_{k\ell})^{-\frac{1}{2}} \left[ \frac{b_{k(0)}^{a_{k(0)}}}{\Gamma(a_{k(0)})} \cdot \frac{b_{k(1)}^{a_{k(1)}}}{\Gamma(a_{k(1)})} \right]^{-1} b_{k\ell}^{a_{k\ell}} \left[ \Gamma(a_{k\ell}) \right]\),

\[
a_{k(0)} = \frac{1}{2} \left( n^{(0)}_{k\ell} - 1 \right), \quad b_{k(0)} = \frac{(n^{(0)}_{k\ell} - 1) \sigma^2_{k(0)}}{2}, \quad a_{k(1)} = \frac{1}{2} \left( n^{(1)}_{k\ell} - 1 \right) + a_{k\ell}, \quad b_{k(1)} = \frac{(n^{(1)}_{k\ell} - 1) \sigma^2_{k(1)}}{2} + b_{k\ell},
\]

\[
A = \prod_{k,\ell} \left[ (2\pi)^{-\frac{1}{2}} \left( n^{(0)}_{k\ell} + n^{(1)}_{k\ell} - 1 \right) \left( n^{(0)}_{k\ell} n^{(1)}_{k\ell} \right)^{-\frac{1}{2}} \right],
\]

\[
B_j = \frac{1}{(1 + g)^2} \cdot (1 + gn)^{-p/2} \cdot \exp \left\{ -\frac{1}{2} h'V^{-1}h + \frac{1}{2} \mu_j' \Sigma_j^{-1} \mu_j \right\},
\]

\[
\Sigma_j = (Z_j'V^{-1}Z_j)^{-1} \left( 1 + \frac{1}{gn} \right)^{-1}, \quad \mu_j = \Sigma_j Z_j'V^{-1}h, \quad \text{and} \quad IG(a, b) \text{ is the probability density function of the inverse-gamma distribution with parameters } a \text{ and } b. \text{ We calculate this inte-}
\]
As in the previous case, likelihood corresponding to observations in cell $(k, \ell)$, integrated over $\mu_{k\ell}^{(1)}$ is

$$L_{0k\ell} = \int_{-\infty}^{\infty} f(y_{k\ell}^{(0)}, y_{k\ell}^{(1)} | \mu_{k\ell}^{(0)}, \sigma_{k\ell(0)}^{2}, \sigma_{k\ell(1)}^{2}) d\mu_{k\ell}^{(0)}$$

$$= (2\pi\sigma_{k\ell(0)}^{2})^{-\frac{3}{2} n_{k\ell}^{(0)}} \left( 2\pi\sigma_{k\ell(1)}^{2} \right)^{-\frac{1}{2} n_{k\ell}^{(1)}} \exp \left\{ -\frac{1}{2} \frac{n_{k\ell}^{(0)}}{\sigma_{k\ell(0)}^{2}} (\bar{y}_{k\ell}^{(0)})^2 + \frac{n_{k\ell}^{(1)}}{\sigma_{k\ell(1)}^{2}} (\bar{y}_{k\ell}^{(1)})^2 - \sigma_{k\ell(0)}^{2} \left( \frac{n_{k\ell}^{(0)}}{\sigma_{k\ell(0)}^{2}} \bar{y}_{k\ell}^{(0)} + \frac{n_{k\ell}^{(1)}}{\sigma_{k\ell(1)}^{2}} \bar{y}_{k\ell}^{(1)} \right)^2 \right\},$$

where $\sigma_{k\ell(1)}^{2} = \left( \frac{n_{k\ell}^{(0)}}{\sigma_{k\ell(0)}^{2}} + \frac{n_{k\ell}^{(1)}}{\sigma_{k\ell(1)}^{2}} \right)^{-1}$.

We then have

$$P(Y|M_0) = \int \int \prod_{i,j} L_{0k\ell} \cdot \sigma_{k\ell(0)}^{2} \cdot \frac{b_{k\ell}^{a_{k\ell}}}{\Gamma(a_{k\ell})} \left( \sigma_{k\ell(1)}^{2} \right)^{-a_{k\ell}-1} \exp \left( -\frac{b_{k\ell}}{\sigma_{k\ell(1)}^{2}} \right) \frac{d\sigma_{k\ell(0)}^{2}}{d\sigma_{k\ell(1)}^{2}}$$

$$= \int \int \prod_{i,j} [Q_{0k\ell} \cdot A_{0k\ell}] d\sigma_{k\ell(0)}^{2} d\sigma_{k\ell(1)}^{2},$$

where

$$Q_{0k\ell} = IG(\sigma_{k\ell(0)}^{2} : a_{k\ell(0)}, b_{k\ell(0)}) IG(\sigma_{k\ell(1)}^{2} : a_{k\ell(1)}, b_{k\ell(1)}) \cdot \frac{b_{k\ell}^{a_{k\ell}}}{\Gamma(a_{k\ell})} \cdot \left( \frac{b_{k\ell(0)}^{a_{k\ell(0)}}}{\Gamma(a_{k\ell(0)})} \cdot \frac{b_{k\ell(1)}^{a_{k\ell(1)}}}{\Gamma(a_{k\ell(1)})} \right)^{-1},$$

$$a_{k\ell(0)} = \frac{1}{2} n_{k\ell}^{(0)}, b_{k\ell(0)} = \frac{\left( n_{k\ell}^{(0)} + n_{k\ell}^{(1)} \bar{y}_{k\ell}^{(0)} \right)^{2}}{2}, a_{k\ell(1)} = \frac{1}{2} n_{k\ell}^{(1)} + a_{k\ell}, b_{k\ell(1)} = \frac{\left( n_{k\ell}^{(1)} + n_{k\ell}^{(1)} \bar{y}_{k\ell}^{(1)} \right)^{2}}{2}.$$
and

\[
A_{0k\ell} = (2\pi)^{-\frac{1}{2}} n_{k\ell}^{(0)} n_{k\ell}^{(1)} + \frac{1}{2} \left( \sigma_{k\ell(01)}^2 \right)^{\frac{1}{2}} \cdot \exp \left\{ \frac{1}{2} \left[ \sigma_{k\ell(01)}^2 \left( \frac{n_{k\ell}^{(0)}}{\sigma_{k\ell(0)}^2} y_{k\ell}^{(0)} + \frac{n_{k\ell}^{(1)}}{\sigma_{k\ell(1)}^2} y_{k\ell}^{(1)} \right) \right]^2 \right\}.
\]

We use the following proposal distributions \( \sigma_{k\ell(0)}^2 \sim IG(a_{k\ell(0)}, b_{k\ell(0)}) \) and \( \sigma_{k\ell(1)}^2 \sim IG(a_{k\ell(1)}, b_{k\ell(1)}) \) to evaluate the integral using importance sampling.
Appendix B

Causal Inference

B.1 Algorithm for sampling from the posterior distribution

To obtain random sample from the posterior distribution $p(\beta, \sigma_f^2, \rho, \omega|Y)$, we use the following algorithm:

1. Get random sample from $p(\rho, \omega|Y)$ in (3.4) using Gibbs sampling;

2. Given $\rho$ and $\omega$ from step 1, obtain random sample from the conditional posterior distribution $p(\sigma_f^2|Y, \rho, \omega)$ in (3.3), which is an Inverse-Gamma distribution;

3. Given $\sigma_f^2$, $\rho$ and $\omega$ from step 2 and step 1, obtain random sample from the full conditional distribution $p(\beta|Y, \Sigma)$ in (3.1), which is a Normal distribution.
After that, we can also obtain random sample from $p ( \beta, \sigma^2_f, \phi, \sigma^2_0 | Y )$ and $p ( Y^c | Y )$, then do the corresponding inference.
Appendix C

HTE

C.1 Algorithm for sampling from the posterior distribution

Algorithm for sampling from the posterior distribution of parameters in GP.

Step 1. Use Gibbs sampling to sample from $p(\rho, \omega | Y)$ in (4.16).

Step 2. Given $(\rho, \omega)$ from step 1, sample $(\sigma_f^2 | Y, \rho, \omega)$ from (4.15).

Step 3. Given $(\rho, \omega, \sigma_f^2)$ from step 1&2, sample $(\beta | Y, \Sigma)$ from (4.14).

Then we get the posterior sample of $(\beta, \sigma_f^2, \rho, \omega)$. For prediction, we need the following steps to get the posterior sample given $(\beta, \sigma_f^2, \rho, \omega)$.

Step 4. Sample $(h^* | Y, \beta, \sigma_f^2, \rho, \omega)$ from (4.19).
Step 5. Sample \((Y^*|Y, \beta, \sigma_f^2, \rho, \omega)\) from (4.20).