Bayesian Modeling Using Latent Structures

by

Xiaojing Wang

Department of Statistical Science
Duke University

Date: _____
Approved:

_____
James O. Berger, Supervisor

_____
David Banks

_____
Merlise A. Clyde

_____
Donald S. Burdick

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2012

Abstract

Bayesian Modeling Using Latent Structures

by

Xiaojing Wang

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
James O. Berger, Supervisor

_____
David Banks

_____
Merlise A. Clyde

_____
Donald S. Burdick

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2012

Abstract

This dissertation is devoted to modeling complex data from the Bayesian perspective via constructing priors with latent structures. There are three major contexts in which this is done – strategies for the analysis of dynamic longitudinal data, estimating shape-constrained functions, and identifying subgroups. The methodology is illustrated in three different interdisciplinary contexts: (1) adaptive measurement testing in education; (2) emulation of computer models for vehicle crashworthiness; and (3) subgroup analyses based on biomarkers.

Chapter 1 presents an overview of the utilized latent structured priors and an overview of the remainder of the thesis. Chapter 2 is motivated by the problem of analyzing dichotomous longitudinal data observed at variable and irregular time points for adaptive measurement testing in education. One of its main contributions lies in developing a new class of Dynamic Item Response (DIR) models via specifying a novel dynamic structure on the prior of the latent trait. The Bayesian inference for DIR models is undertaken, which permits borrowing strength from different individuals, allows the retrospective analysis of an individual's changing ability, and allows for online prediction of one's ability changes. Proof of posterior propriety is presented, ensuring that the objective Bayesian analysis is rigorous.

Chapter 3 deals with nonparametric function estimation under shape constraints, such as monotonicity, convexity or concavity. A motivating illustration is to generate an emulator to approximate a computer model for vehicle crashworthiness. Although

Gaussian processes are very flexible and widely used in function estimation, they are not naturally amenable to incorporation of such constraints. Gaussian processes with the squared exponential correlation function have the interesting property that their derivative processes are also Gaussian processes and are jointly Gaussian processes with the original Gaussian process. This allows one to impose shape constraints through the derivative process. Two alternative ways of incorporating derivative information into Gaussian processes priors are proposed, with one focusing on scenarios (important in emulation of computer models) in which the function may have flat regions.

Chapter 4 introduces a Bayesian method to control for multiplicity in subgroup analyses through tree-based models that limit the subgroups under consideration to those that are a priori plausible. Once the prior modeling of the tree is accomplished, each tree will yield a statistical model; Bayesian model selection analyses then complete the statistical computation for any quantity of interest, resulting in multiplicity-controlled inferences. This research is motivated by a problem of biomarker and subgroup identification to develop tailored therapeutics. Chapter 5 presents conclusions and some directions for future research.

To Huimin Huang and Zixuan Wang

Contents

List of Tables

List of Figures

Acknowledgements

In the completion of my PhD program, I have acquired many debts. My eternal and heartfelt thanks should first and foremost go to my supervisor, Professor James O. Berger. His great insight and learning have inspired me step by step to bring forth these focuses in the very Bayesian view. His illuminated guidance, timely suggestion and sincere criticism in the process of writing my dissertation have led me to my full and smooth completion in the happiest circumstances. Without his great patience and generous help, my dissertation could never be well accomplished. And his way of thinking and teaching have left me an unforgettable impression and will light up my academic life in future, too.

My great appreciation should then be expressed to Dr. Donald S. Burdick from MeteMetric Inc., who has generously provided the firsthand dataset for me and his instruction always appears timely. I am also grateful to Professor Merlise A. Clyde, for her precious time to revise my dissertation. And thank Professor David Banks very much for his valuable suggestions and comments, who is the mentor that every student would like to turn to for help along the way.

Another great gratitude of mine should go to Professor Mike West for his instructive guidance to my first research project at Duke and his great encouragement for us to write interesting subjects even on the initial stage of our PhD study. Further thanks shall then go to Professor Alan Gelfand, Professor A. Ronald Gallant, Professor Dalene Stangl, Professor David B. Dunson, Professor Charles Becker, Dr. Fan

Li and Dr. Li Ma for their impressive lectures and helps during my stay at Duke.

Moreover, I am much obligated to Statistical and Applied Mathematical Sciences Institute, MetaMetrics Inc. and Eli Lilly and Company for their generosity in partially financing my graduate study and conference travels. And I would much like to thank all my insightful collaborators: Dr. Jack Stenner, Hal Burdick, Dr. Carl Swartz and Sean Hanlon at MetaMetrics Inc., for their great supports and guidance for my work in the field of Education Statistics.

In addition, I am quite lucky to meet many smart and kindhearted friends and classmates in Duke who make my life very significant and colorful. So I am grateful to Ioanna Manolopoulou, James Scott, Hao Wang, P. Richard Hahn, Matt Heaton, Anirban Bhattacharya and Chunlin Ji for many interesting discussion; to Min Huang, Lin Lin, Yajuan Si, Minghui Shi and Dan Shen for sharing their happiness and experience with me; to Hongxia Yang, Jianyu Wang, Yingbo Li, Debdeep Pati, Francesca Petralia, Fangpo Wang, Zhengzi Li, Kai Cui, Fernando V. Bonassi, Jouchi Nakajima, Thomas J. Leininger and many others for their support in need.

To my wonderful Da Huang: I shall express my special heartfelt gratitude to him for his love and great efforts for support on my way to a success. With his accompanying, life is more enjoyable here at Duke.

Finally, I am deeply indebted to my dear parents, Huimin Huang and Zixuan Wang for their boundless love and faith in me. Every step that I took from my childhood to my PhD program has witnessed their great love, tireless efforts and lots of sacrifices. I could not have been offered more from them on my way to success.

I believe that no dissertation for PhD degree is the sole domain of the author herself/himself, so does mine.

1

Introduction

Many statistical problems are concerned with interpreting, incorporating and identifying meaningful structure in data. For instance, with time series data, the core issue is often to model the data as a function of underlying latent states that have a specified structure.

In function estimation, the unknown function is often known to possess a latent structure, such as unimodality, monotonicity or convexity. As a practical example, Figure 1.1 (see Bayarri et al. (2009) for further details) describes the velocity pulses of a vehicle over a period of 0.065 seconds in reacting to a crash. In this situation, it is reasonable to assume that the velocity output of the vehicle decreases monotonically with time. Thus any model we build of the function should utilize this type of constrained structure.

In subgroup analyses, biological information often suggests that only certain latent subgroup structures are plausible, and basing the analysis on such structures can greatly improve their power.

The central purpose of this thesis is to exemplify how the Bayesian use of prior information allows one to naturally incorporate complex latent structures into the

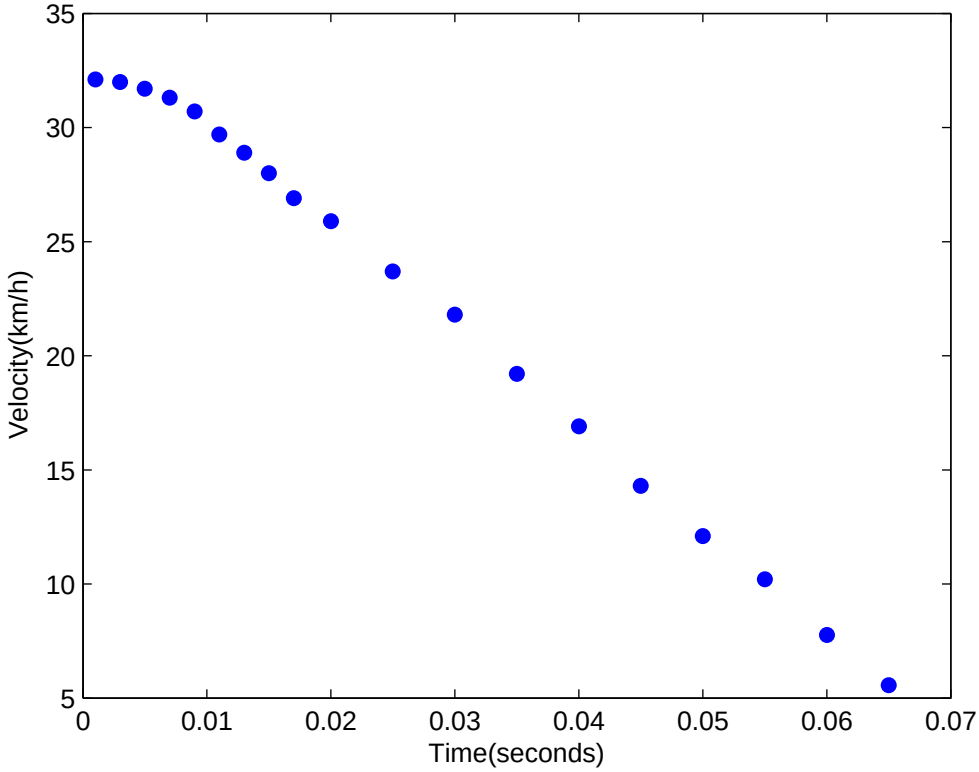


FIGURE 1.1: Velocity pulses in the occupant compartment from vehicle crash

analysis, through appropriate stochastic representations of the structures. Because the Bayesian approach is completely probabilistic, these prior structures can then be seamlessly integrated with the probabilistic data model to draw needed inferences.

This theme of prior modeling of latent structures is illustrated in three domains: 1) analyzing dynamic latent trajectories of growth in unequally-spaced longitudinal/time series data; 2) imposing shape constraints (such as monotonicity, convexity or concavity) on nonparametric function estimation with Gaussian process priors; 3) subgroup analyses with latent subgroup structures.

1.1 Three Latent Structure Prior Models

Three of the common approaches to modeling prior structure – state space modeling, Gaussian processes, and tree-based models – will be reviewed here. All three will be utilized in later chapters of the thesis.

1.1.1 State Space Models

Longitudinal data is frequently encountered in social and behavioral sciences, medical and public health sciences, and finance and economics. Longitudinal data can be regarded as a collection of many time series, each for one subject. The primary interest of longitudinal data analysis usually lies in exploring the mechanism of changes over time, including growth, aging, time effects of covariances and so on. State space models are a successful class of models for identifying and interpreting this latent processes underlying the observed time series. They allow a natural interpretation of time series as the combination of several components, such as trend, seasonal or regressive components. They also have an elegant and flexible probabilistic structure, which makes computation easy through recursive algorithms; this feature also makes them natural to consider in a Bayesian framework.

The idea of state space modeling was clearly stated in Kalman (1960), in connection with the theory of controls in linear system. Duncan and Horn (1972) showed that linear mixed models have a state space representation. Akaike (1974), Harrison and Stevens (1976) used state space models for analysis of time series. State space models for analyzing longitudinal data were described by Jones (1993). A comprehensive description of these models can be found in West and Harrison (1997), Petris et al. (2009) and Prado and West (2010).

A state space model consists of an $\mathbf{R}^p$-valued time series of the latent state process $\boldsymbol{\theta}_t$ and an $\mathbf{R}^d$-valued time series of the observation process $\{\mathbf{Y}_t\}$, for $t = 0, 1, \cdots, T$,

and is defined to satisfy the following assumptions (see Chapter 2 in Petris et al. (2009)):

1. $\boldsymbol{\theta}_t$, for $t = 0, 1, \cdots, T$ is a Markov chain;

2. Conditionally on $\boldsymbol{\theta}_t$, for $t = 0, 1, \cdots, T$, the $\mathbf{Y}_t$'s are independent and $\mathbf{Y}_t$ only depends on $\boldsymbol{\theta}_t$.

Thus, the state space model is completely specified by the initial distribution $\pi(\boldsymbol{\theta_0})$ and the conditional densities $\pi(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1})$ and $\pi(\mathbf{Y}_t \mid \boldsymbol{\theta}_t)$ for $t \geq 1$. In fact, for any $T \geq 1$,

$$\pi(\boldsymbol{\theta}_{0:T}, \mathbf{Y}_{1:T}) = \pi(\theta_0) \prod_{i=1}^{T} \pi(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{i-1}) \pi(\mathbf{Y}_i \mid \boldsymbol{\theta}_i).$$

The dependence structure of the state space model can be presented as a special case of a directed acyclic graph (cf. Cowell et al. (2003)) shown in Figure 1.2.

$$\begin{array}{ccccccccccccc}
\theta_0 & \rightarrow & \theta_1 & \rightarrow & \theta_2 & \rightarrow \cdots \rightarrow & \theta_{t-1} & \rightarrow & \theta_t & \rightarrow & \theta_{t+1} & \rightarrow \cdots \\
 & & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
 & & Y_1 & & Y_2 & & Y_{t\text{-}1} & & Y_t & & Y_{t+1} &
\end{array}$$

FIGURE 1.2: Dependence structure for a state space model

An important class of state space models is Dynamic linear models (DLM's), where the dependence of latent states and observations have a Gaussian linear structures. To be more specific, a DLM is specified by a normal prior distribution at the p-dimensional initial state, i.e.

$$\boldsymbol{\theta}_0 \sim \mathcal{N}_p(\mathbf{m}_0, \mathbf{C}_0),$$

together with the pair of observation and system equations as below for any $t \geq 1$,

$$\begin{aligned} \text{Observation Equation:} \quad \mathbf{Y}_t &= \mathbf{A}_t \boldsymbol{\theta}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{i.i.d.}{\sim} \mathcal{N}_d(\mathbf{0}, \mathbf{V}_t), \\ \text{System Equation:} \quad \boldsymbol{\theta}_t &= \mathbf{B}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \overset{i.i.d.}{\sim} \mathcal{N}_p(\mathbf{0}, \mathbf{W}_t), \end{aligned}$$

where $\mathbf{A}_t$ and $\mathbf{B}_t$ are $d \times p$ and $p \times p$ matrices, respectively, that can be known or estimated. Estimation and forecasting with DLM's can be easily done using the well-known Kalman filter and Kalman smoother (see Proposition 2.2 and Proposition 2.4 in Petris et al. (2009)).

The closed forms of the Kalman filter and Kalman smoother are attractions of DLM's, but the assumption of the linear Gaussian structure for states and observations may not hold in many situations, especially when DLM's are extended to analyze longitudinal data. For example, in the analysis of adaptive testing in education discussed in Chapter 1, a time series of binary observations is recorded for each individual; this clearly cannot be normally distributed. The literature generalizing state space models to nonlinear and non-Gaussian situations includes Kitagawa (1987), Carlin et al. (1992), Fahrmeir (1992), West and Harrison (1997), Jørgensen et al. (1999), and Song (2007).

Another limitation of standard state space models is clearly shown in Figure 1.2: there is no additional structure in the model to allow for uncertainty in the timings between different observations, an issue that arises frequently in studies in social or biomedical research. Huerta and West (1998) revised DLM's in order to consider a time series of oxygen isotope observations sampled at irregularly spaced intervals. Xu et al. (2007) extended state space models to analyze unequally spaced longitudinal count data from a frequentist perspective. One approach to overcoming this problem is to let the matrices $\mathbf{B}_t$ and $\mathbf{W}_t$ in system equations be functions of the time intervals between observations.

1.1.2 Gaussian Process models

Suppose the data consists of n pairs of covariates/inputs $\mathbf{t}_i \in \mathcal{T}$ and noisy responses/outputs $x_i \in \mathbb{R}$ where the responses are independently

$$x_i = f(\mathbf{t}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

with σ^2 being the variance of the noise. A natural question for a Bayesian is how to build a prior distribution for the function $f(\cdot) : \mathcal{T} \to \mathbb{R}$.

One of the most popular nonparametric prior distributions over functions is the Gaussian process (GP) (see Rasmussen and Williams (2006)). According to Quiñonero Candela and Rasmussen (2005),

Definition 1.1. *A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

This definition reveals that the main idea underlying the GP is to model $(f(\mathbf{t}_1), \cdots, f(\mathbf{t}_n))'$ for any $n \in \mathbb{N}_+$ and any $\mathbf{t}_i \in \mathcal{T}, \forall i \in \{1, \cdots, n\}$ jointly as a multivariate normal distribution with some mean vector $(m(\mathbf{t}_1), \cdots, m(\mathbf{t}_n))'$ and covariance matrix $\mathbf{K}$. Here $m(\cdot)$ is a mean function mapping $\mathcal{T}$ to $\mathbb{R}$ and $K(\mathbf{t}_i, \mathbf{t}_j)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, n$ indicates entries of covariance matrix $\mathbf{K}$, which are determined by a covariance function. (For examples, see Rasmussen and Williams (2006) Section 4.2 and Banerjee et al. (2004) Section 2.1.3).

Letting $r = \|\mathbf{t} - \mathbf{t}'\|$, where $\|\cdot\|$ denotes Euclidean distance, two of the most common covariance functions used in GP to estimate smooth functions are

- *Power Exponential*, (Oliveira et al. (1997))

$$K_{PE}(r) = \sigma^2 \exp\left(\frac{r}{\beta}\right)^{\gamma}, \quad \text{for} \quad \beta > 0, \gamma \in (0, 2].$$

  This family of covariance functions includes both the exponential ($\gamma = 1$) and squared exponential ($\gamma = 2$) covariance functions.

- *Matérn*, (Matérn (1986), Handcock and Stein (1993))

$$K_M(r) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r}{\beta} \right)^{\nu} \mathcal{K}_\nu \left(\frac{\sqrt{2\nu} r}{\beta} \right), \quad \text{for} \quad \beta > 0, \nu > 0,$$

  where $\Gamma(\cdot)$ is the gamma function, $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the order ν (Abramowitz and Stegun (1965), Section 9.6). This family contains the exponential covariance function ($\nu = 0.5$) and the squared exponential covariance function ($\nu \to \infty$).

If the joint multivariate normal distribution holds for any $\mathbf{t}_1, \cdots, \mathbf{t}_n$ with $n \in \mathbb{N}_+$, the prior distribution for $f(\cdot)$ is a Gaussian process. Hence the Gaussian process could be viewed as a generalization of the multivariate normal distribution to the infinite extreme.

Let $Z(\mathbf{t})$ be a real stochastic process and define mean function $m(\mathbf{t})$ and the covariance function $K(\mathbf{t}, \mathbf{t}')$ as

$$\begin{aligned} m(\mathbf{t}) &= \mathrm{E}[Z(\mathbf{t})] \\ K(\mathbf{t}, \mathbf{t}') &= \mathrm{E}[(Z(\mathbf{t}) - m(\mathbf{t}))(Z(\mathbf{t}') - m(\mathbf{t}'))]. \end{aligned}$$

Then, the GP is denoted as

$$Z(\mathbf{t}) \sim \mathcal{GP}(m(\mathbf{t}), K(\mathbf{t}, \mathbf{t}')),$$

which implies the GP is completely specified by its mean function and covariance function. Moreover, the mathematical properties of the realization of a GP, such as continuity and differentiability are also determined by its mean function and covariance function (see Adler (1981) for details).

One attraction of GP models is that, often, much of the computation will simply be parametric Gaussian computation. In addition, GP models are very flexible for modeling complex phenomena since they allow possible nonlinear effects and can

handle dependencies between covariates. There have been numerous uses of GP models in statistics, including the early papers Blight and Ott (1975) and O'Hagan and Kingman (1978). Rasmussen and Williams (2006), Kuss (2006), and references therein demonstrated the use of GP models in regression, classification and reinforcement learning in machine learning. Another important application of GP models is multivariate interpolation, for instance, in spatial statistics under the well-known name 'kriging' (see Cressie (1993), Banerjee et al. (2004) and reference therein), or in emulating deterministic computer experiments with Gaussian response surface approximations (see Santner et al. (2003) and Fang et al. (2005) for details). Shi and Choi (2011) summarized the use of GP models in the analysis of functional data.

GP models have not been widely used when there is prior knowledge concerning the shape of the function. This is mainly due to the normality property of a GP, which can not guarantee that realizations of the GP would be positive, monotonic, convex, etc. Chapter 3 in this thesis will be devoted to the study of the possibilities for introducing such constraints into Gaussian processes.

1.1.3 Tree-based Models

Although linear regressions and other parametric models provide a useful way for interpreting simple structures in data, such simple structures often do not hold uniformly over the entire dataset. Instead, such structures might hold locally, but with the structure changing over the range of the data. A popular tool for addressing such problems is tree-based modeling.

Tree-based models use a recursive method to partition the feature space. Usually, by recursively bisecting the predictor space, the hierarchical tree structure divides the dataset into homogeneous groups and handles interactions and nonlinearities between predictors and response in a implicit way. The subset of data in each terminal node of the tree will supposedly be a homogeneous population that can then be modeled

using traditional statistical structures.

To elaborate further, a tree-based model contains two important parts: 1) a binary tree T that partitions the predictor space $\mathcal{X}$ into b different disjoint subspaces, where b is the terminal nodes of the tree T; 2) a parameter space $\Theta = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_b)$, where the parameter $\boldsymbol{\theta}_i$ is associated with the ith terminal node to specify the submodel for observations and predictors within the terminal node. Notice that every observation and each value of the predictor are assigned to only one specific terminal node in the tree T. Each interior node of the binary tree T has two children, using a single predictor variable in the splitting rule. Whether a tree-based model is called a regression tree or a classification tree depends on whether the observation is quantitative or qualitative. A simple example of using a binary tree structure to partition the predictor space (X_1, X_2) is shown in Figure 1.3, where X_1 and X_2 are assumed to be continuous variables and Y_i's, for $i = 1, \cdots, 10$, are data. The tree divides the space of X_1 and X_2 at $X_1 = t_1$ and $X_2 = t_2$. The orange nodes indicate they are terminal nodes in this binary tree.

The tree partitioning, also known as recursive partitioning, was used early on in the analysis of survey data. Belson (1959) addressed a matching issue to do prediction. Morgan and Sonquist (1963) proposed the automatic interaction detector algorithm for growing a binary regression tree. The application of tree-based models has grown enormously since the development of Classification and Regression Trees (CART) by Breiman et al. (1984). For example, Bahl et al. (1989) introduced a tree-based language model for natural language speech recognition, Geman and Jedynak (1996) employed decision trees to form an active testing model for tracking roads in satellite images, De'ath and Fabricius (2000) summarized the application of CART to ecological data, Lee et al. (2006) demonstrated the effectiveness of credit scoring by using CART on a bank credit card dataset, and Roko and Gilli (2008) used the classification tree to do stock selection. Moreover, the binary tree representation as

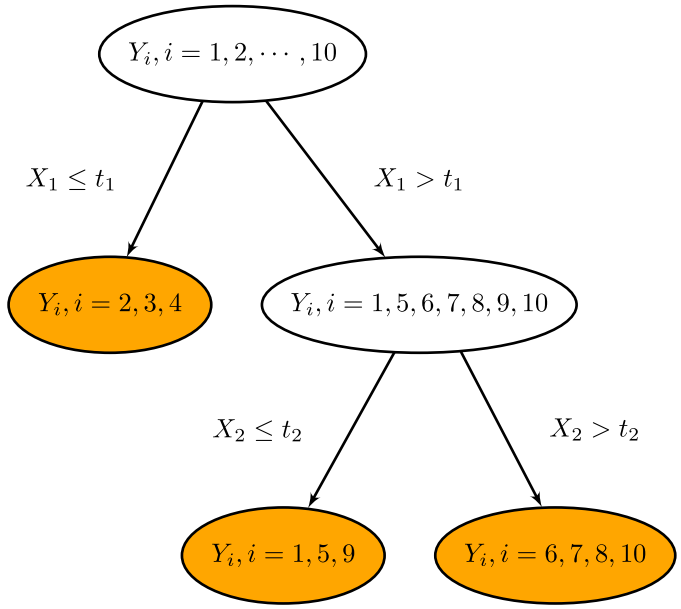


FIGURE 1.3: The binary tree representation of partitioning the predictor space

shown in Figure 1.3 is natural in stratifying a population into strata of high and low outcomes, on the basis of subjects' characteristics, and so has become popular in medical and clinical studies, for instance, in Marshall (2001), Lemon et al. (2003) and Ruberg et al. (2010). The latter is particularly relevant to this thesis, in that it considers subgroup analyses in clinical trials, the subject of Chapter 4.

The conventional approach to finding a 'good' binary tree is to use a greedy algorithm to grow a tree and then prune it back to avoid overfitting (Breiman et al. (1984) and Quinlan (1986)). In contrast, Chipman et al. (1998) and Denison et al. (1998) proposed a Bayesian approach which induces a posterior distribution to guide a stochastic search towards 'more promising' treed models. A key aspect of any Bayesian method is to introduce prior distributions on all unknowns. As mentioned earlier, there are two important components constituting a tree-based model, the tree structure T and terminal node models Θ. Thus, the unknowns could be identified as (Θ, T) and the question is how to specify a prior distribution $\Pr(\Theta, T)$. Since Θ

indexes the parametric model for each tree T, it is often more convenient to write $\Pr(\Theta, T)$ as

$$\Pr(\Theta, T) = \Pr(\Theta \mid T)\mathrm{P}(T),$$

and then specify $\Pr(\Theta \mid T)$ and $\mathrm{P}(T)$ separately. The advantage of this structure is that the choice of T does not depend on the form of the submodel distribution at the terminal node, which makes the specification of prior in the tree space very flexible. In Chapter 4, we specify $\Pr(T)$ implicitly by a random tree generating stochastic process. To be specific, each tree is a realization of such process, and can be considered as a random draw from this prior, as in Chipman et al. (1998) and Wu et al. (2007).

Typically, the stochastic process for drawing a tree from the prior $\Pr(T)$ is described in a recursive manner, where the tree starts with a single root node, and grows by randomly splitting nodes to form children, who can be further split. The two core issues in this growing process are the splitting probabilities and the probability of assigning some splitting rule to the node, which actually determine the complexity of the tree. Chipman et al. (1998) used the splitting probability $\alpha(1 + d_\eta)^{-\beta}$ to control the size and shape of the generated trees, where d_η here indicates the depth of the node η in the tree, $\alpha \in (0, 1)$ and $\beta \geq 0$ are prechosen control parameters. Wu et al. (2007) argued that the prior Chipman et al. (1998) used are controlling the number of nodes and shape of the tree implicitly. They, instead, proposed a "pinball prior" to allow for the combination of an explicit specification of a distribution for both the tree size and the tree shape. The random mechanism we adopt in Chapter 4 to specify the splitting and other properties of the tree is quite different from theirs, and is tailored to specific issues arising in subgroup analyses of clinical trails.

1.2 Outline and Contributions of the Thesis

In Chapter 2, a new class of state space models, called Dynamic Item Response (DIR) models, is developed, in the context of the adaptive measurement testing in psychometric and educational studies. Chapter 3 develops two alternative ways to incorporate shape constraints into Gaussian process priors. Chapter 4 considers the problem of subgroup analyses, where the focus is to develop a Bayesian method to control for multiplicity. Chapter 5 gives a summary of the thesis and indicates directions and topics for future research. In the following, the motivations and major statistical contributions of each chapter are elaborated.

Chapter 2

Item Response Theory (IRT) models have been widely used in measurement testing in social and behavioral sciences. In classical IRT, it is assumed that one's response to a particular item in the test is independent given that person's ability and the item difficulties in the test. However, the increased availability of large and complex longitudinal data in measurement testing especially for the advent of computer adaptive testing, has challenged this fundamental assumption in various ways. A case in point is the reading test dataset from MetaMetrics, Inc. The three major characteristics of MetaMetrics data are: 1) there are repeated observations available for individuals through time, and moreover, this longitudinal data is observed at variable and irregular time points; 2) besides the ability and the item difficulty, there are numerous other factors such as health status, the understanding of the background knowledge or context and etc. to influence one's response to a particular item in the test, thus going against the fundamental assumption of IRT models; 3) there is some uncertainty associated with each item difficulty, which implies the item is randomly drawn from a bank of items where all items have the same ensemble mean.

Thus, the main contribution of Chapter 2 lies in developing a new class of state space models, called Dynamic Item Response (DIR) models, in the context of item response theory, to deal with the three critically needed generalizations mentioned above. The dynamic structure to accommodate changes in the ability/trait is incorporated into the model via specifying the prior of the latent variable, i.e. the ability/trait. A random effect approach is utilized to study the potential dependencies between items in the test. The models can be applied either retrospectively to the full data or prospectively, in cases where real-time prediction is needed.

Chapter 3

Computer models are increasingly being used to represent complex physical processes. However, computer models are often very time consuming to run, and hence cannot be directly utilized for many important tasks such as designs or Bayesian inferences. A common solution is to build a statistical approximation – called an *emulator* – to the output of the computer model. Gaussian process models is the most common approximation method used to approximate the output of the computer models.

Often, the response of the computer model to certain inputs has a shape-constrained form. For example, as shown in Figure 1.1, the velocity of a relevant feature of a vehicle in reacting to a crash will be decreasing with time. The main contribution in Chapter 3 is to allow introduction of such shape constraints in the construction of Gaussian process emulators. This is done by imposing constraints on the derivatives of the Gaussian process at points in the input space. Two alternative ways of incorporating derivative information into Gaussian processes are proposed, with the second method being more suitable for functions that have flat areas. A computational implementation of this idea is developed and shown to be successful in estimating constrained functions.

Chapter 4

Patients in a clinical trial are often not a homogeneous sample; their responses to a treatment and the impacts of different treatments on them may vary. Indeed, it is often plausible that there are specific subgroups of patients who would respond more effectively than others to a treatment and the identification of such subgroups is of great interest. But, from a statistical perspective, searching for such subgroups is known to be problematical, because of the dangers of multiple testing.

Chapter 4 develops a new Bayesian approach to subgroup analyses using tree-based models, in which the population is partitioned into allowable subgroups arising from terminal nodes of trees based on population covariate splits, with possible zero treatment effects or zero baseline effects. The terminal nodes of each tree provide a latent partition of the population into different subgroups. Following this approach and insisting on certain rules for tree construction result in a dramatically reduced number of biologically reasonable subgroups that need to be considered. Determining how to embed prior information about subgroups into these tree-based models is one of the vital parts for this research and will stimulate further investigation.

After the completion of the specification of the tree prior, each tree results in a statistical model, and Bayesian model selection techniques are utilized to yield a multiplicity-controlled posterior subgroup analysis.

2

Dynamic Item Response Models

Is it possible that we monitor the changes of one's ability as easy as we do that of the height and weight? Item Response Theory (IRT) models have already provided a powerful way to measure the latent ability of individuals at one time and have been widely used in educational measurement testing. With the advent of adaptive measurement testing, how to measure dynamic changes for the individual ability throughout the time is the most critical issue that is badly in need of solution. However, the recent literature of IRT models keeps silent on this. This chapter is going to present a generalization of IRT models in the context of adaptive measurement testing on developing a new class of IRT models, which is motivated from a large collection of reading test data obtained from MetaMetrics, Inc.. It will focus on extending the IRT models in the situations: when there are repeated observations available for individuals through all the time; when there are various violations of the common assumption that test results are conditionally independent, given abilities and item difficulties; and when there are partial uncertainties associated with item difficulties. The Dynamic Item Response (DIR) models proposed in this chapter cover and extend the paper of Wang et al. (2012). The significant contributions for

DIR models rest on adding a novel dynamic structure for the latent trait of the ability to the framework of IRT models to accommodate changes in the ability as well as adding random components to it to account for correlated relationship between items within a test. Although the main concentration in this chapter is on IRT models, our idea is straightforward to be applied to discrete/continuous response with longitudinal data observed at variable and irregular time points in other kinds of model concerned.

2.1 Literature Review and Motivations

2.1.1 Background

Item response theory (IRT) models are frequently used in modeling dichotomous data from educational tests, since they allow separate assessment of the ability of examinees and effectiveness of the test items. A typical one-parameter IRT model is of the form

$$\Pr(X_{il} = 1 \mid \theta_i, d_l) = \mathrm{F}(\theta_i - d_l), \tag{2.1}$$

where θ_i indicates the ability of the i-th person; d_l indicates the difficulty of the l-th test item; the item response variable X_{il} could be either 0 or 1, corresponding to whether the l-th test item taken by the i-th person is answered correctly or not; and the item characteristic curve, $\mathrm{F}(\cdot)$, is a cdf from a continuous distribution. When $\mathrm{F}(\cdot)$ is the standard logistic cdf, the one-parameter IRT model (2.1) becomes the famous Rasch model

$$\Pr(X_{il} = 1 \mid \theta_i, d_l) = \frac{\exp(\theta_i - d_l)}{1 + \exp(\theta_i - d_l)}. \tag{2.2}$$

If $\mathrm{F} = \Phi$, where Φ is the standard normal cdf, then

$$\Pr(X_{il} = 1 \mid \theta_i, d_l) = \Phi(\theta_i - d_l) \tag{2.3}$$

defines the one-parameter Normal Ogive or Probit model. We will focus on the former model in this chapter, for reasons to be discussed later, although the analysis of the Probit model is actually easier and can be done with a simplified version of the methodology developed here.

The development of item response theory from the classical point of view owes much to the pioneering work of Lord (1953), Rasch (1961), and their colleagues. Among the many noteworthy contributions are Andersen (1970) and Darrell Bock and Lieberman (1970).

In classical IRT, it is assumed that the X_{il} are independent, given the person's ability θ_i and the difficulty levels d_l. This is often referred to as the *local independence* assumption. There are situations in which this assumption is violated. One such is Computer Adaptive Testing (CAT), wherein the selection of the next test item typically depends specifically on the previous questions and answers.

The situation is less clear with what is studied herein, MetaMetrics' educational assessment program called Computer Adaptive Instruction and Testing (CAIT) program. With the CAIT, a test pool of articles is selected for the student based on an estimate of their current ability; the student selects an article from this pool; and the test questions (described later) are then generated before reading commences. Thus, in the environment of the CAIT, the possible violation in the local independence would arise from sources such as article selection by the student, and the fact that the test questions relate to the same article so that overall understanding of the article could affect all answers; in this chapter, such possible effects will be called *test effects*. Other factors that could cause violation of the local independence include health status and emotional status of the student on a given day; these will be referred to as *daily effects*. In the MetaMetrics scenario, there had been no previous demonstration of the violation of the local independence through the presence of test effects or daily effects, and there was a considerable interest in establishing such

presence for possible enhancement of current models.

Pioneering papers that addressed the local dependence were Stout (1987) and Strout (1990), where an essential dimensionality of the collection of test items and an essential independence were introduced, and Gibbons and Hedeker (1992), which considered the conditional dependence within identified subsets of items by allowing random effects in the analysis. More recent work in this direction is the testlet response theory modeling, proposed by Bradlow et al. (1999). They defined a testlet as the subset of items; for example, a reading comprehensive section in the SAT is defined as the testlet. They then modified the classic IRT models to add a random effect term to represent the common factor affecting the responses in the testlet. Another approach to handle the local dependence is the introduction of the Markov structure, such as Jannarone (1986) which introduced the conjunctive IRT kernel. A recent paper concerned is Andrich and Kreiner (2010), where they modified the Rasch model by allowing the conditional probability of a response to an item to depend on the answer of a previous item.

For the modeling in this chapter, the random effect approach will be followed. Indeed, two levels of random effects will be introduced, to model the daily effects and test effects, respectively.

Another essential generalization of the IRT models lies in their applicability to longitudinal data, i.e., to scenarios in which an individual is tested repeatedly over time; then, the interest typically centers on the growth of an ability in the individual. Embretson (1991) and Marvelde et al. (2006) presented a multidimensional Rasch model to represent the change of an ability as an initial ability and one or more modifiabilities. Based on the belief that a person's ability growth would be increasing over time, Adler (1981), Tan et al. (1999) and Johnson and Raudenbush (2006) used linear or polynomial regression of the time variable to measure the growth of ability; their analysis required the same time span and testing points for all examinees.

Martin and Quinn (2002) modeled the transition of a voting preference as a first-order Markov process, where they assumed voting preference changes from the previous time point to a new point by a random shock; this work did not incorporate a time trend. Park (2011) supposed that changes in a voting preference were subject to discrete agent-specific regime changes and modeled the indicator of the preference regime changes as a first-order Markov process.

Our approach to the longitudinal issue is based on a new class of dynamic linear models (DLM's) (see the background of DLM's in Section 1.1.1of Chapter 1). The literature on DLM's or state space models, in the framework considered here of longitudinal binomial data is also reviewed in Section 1.1.1 of Chapter 1. Our models are distinguished from this literature by simultaneously allowing for the following features: (i) observations at variable and irregular time points; (ii) continuously changing ability, but with an incorporation of knowledge concerning trends (e.g., increasing ability over time) in a non-dogmatic way (thus accommodating, say, a drop in reading ability over a summer vacation); (iii) an analysis that is either individual or hierarchical across a group of individuals, the latter allowing for "borrowing strength" in estimates of certain overall parameters; (iv) either a retrospective analysis based on the full data, or a real-time analysis and prediction for an individual based on the data to date.

We consider the case in which the item difficulties are nominally specified, which is the situation in CAIT, the test items are often computer-generated and have theoretically determined difficulties. The actual item difficulties are quite uncertain, however, and this uncertainty is also accommodated in the analysis. Previous papers that introduced random effects for item parameters include Sinharay et al. (2003) and De Boeck (2008).

2.1.2 Testbed Application

The model proposed in this chapter is motivated by CAIT testing developed by MetaMetrics Inc. The main applied goals are as follows:

- The original goal is to assess the appropriateness of the local independence assumption for this type of data. This evolves into the goal of better understanding the nature of the daily and test effects.
- A second goal is to understand growth in the ability of students, by retrospectively producing their estimated growth trajectories for the study.
- A third goal is to enable on-line prediction of one's ability (based solely on data obtained up to that point), to enable a better assignment of reading materials to match his/her ability, and to enable teachers to better assist students.

The data considered is from a school district in Mississippi. The data consisted of 1983 students registered over two years in a CAIT reading test program conducted by MetaMetrics Inc. The students were in different grades and entered and left the program at different times between 2007 and 2009. Individuals took tests on different days and had different time lapses between tests. Because of the long periods of testing, a fully adaptive model accommodating continual changes in one's ability is needed.

The data was generated during sessions in which a student read an article selected from a large bank of available articles. The articles in this bank had been assigned to text complexity measured in Lexiles, using the Lexile Receptive Analyzer®, the software developed by MetaMetrics Inc. to evaluate the semantic and syntactic complexity of the text. The Lexile measure represents either an individual's reading ability or the complexity of a piece of text. The scale for Lexiles ranges from 0 to 1800, with 0 indicating no reading ability and with 1800 being the maximum.

A session begins like this: a student selects from a generated list of articles having Lexile complexities in a range targeted to the current estimate of the student's ability. For the selected article, a subset of words from the article are eligible to be *clozed*, i.e. removed and replaced by a blank. The computer, following a prescribed protocol, randomly selects a sample of the eligible words to be clozed and presents the article to the student with these words clozed. When a blank is encountered while reading the article, the student clicks it, which is then presented with the true removed word along with three incorrect options called foils. As with the target word, the foils are selected randomly according to a prescribed protocol. The student selects a word to fill in the blank from among the four choices and an immediate feedback is provided in the form of the correct answer.

The dichotomous items produced by this procedure are called "Auto-Generated-Cloze" items. They are single-use items generated at the time of an encounter between a student and an article. If another student selects the same article to read, a new set of target words and foils are selected. Although it is not strictly impossible for an individual item to be taken by more than one student, such an occurrence is highly improbable. As a consequence, it is not feasible to obtain data-based estimates of item calibration parameters.

Instead, the difficulties of the items generated for an encounter between a student and an article can be modeled as a sample from an ensemble of item difficulties associated with the article. The text complexity in Lexiles provides a theoretical value for the ensemble mean. An estimated student ability in combination with assumptions about the ensemble allows the calculation of a predicted success rate for the encounter. A comparison of the observed success rate with predicted, aggregated over many encounters, provides a basis for assessing the viability of the assumptions incorporated into the model. The predicted success rates in Table 1 in Stenner (2010) include the assumption that the mean of the ensemble of item difficulties for

an article is given by its theoretical text complexity. The agreement with observed success rates supports that assumption.

Although the MetaMetrics data is typically presented in Lexile units, there is a simple linear transformation from Lexiles to logit units. We will utilize the more common logit units for all data and results in this chapter. Note that this also motivates the use of the logistic IRT model in this chapter – to preserve compatibility with the MetaMetrics data.

2.1.3 Preview

Because of the complexity of the model considered (and of the testbed data set), as well as the need to incorporate prior information into the model, the analysis will be carried out using Bayesian methodologies and Markov chain Monte Carlo (MCMC) computational techniques. A side benefit of using these methodologies is that all uncertainties in all quantities are combined in the overall assessment of inferential uncertainties. The MCMC procedure utilizes a novel combination of a Gibbs sampling together with a block sampling scheme involving a forward filtering and backward sampling.

In Section 2.2, we formally describe the proposed models to capture the dynamic changes in a person's ability as well as the local dependence between item responses. Section 2.3 presents the MCMC strategy to carry out the statistical inference. Section 2.4 tests the methodology on some simulated example (where the truth is known). Section 2.5 applies the proposed models to the MetaMetrics dataset. Section 2.6 draws conclusions from both statistical and psychological sides.

2.2 Dynamic Item Response (DIR) Models Proposed

This section formally introduces the proposed one-parameter DIR model. Although the focus is on generalizing one-parameter IRT models, it would be straightforward

to similarly generalize two-parameter or three-parameter IRT models.

2.2.1 The Observation Equation in DIR Models

In a typical one-parameter IRT model (2.1), the index of the item response X_{il} indicates the correctness of the i-th person's answer to the l-th question in a single test. Consider the more involved situation in which the individual completes a series of tests within a given day and over different days. Thus, the item response variable is $X_{i,t,s,l}$, which corresponds to the correctness of the answer of the l-th item in the s-th test on the t-th day taken by the i-th person. Here $i = 1, \cdots, n$; $t = 1, \cdots, T_i$; $s = 1, \cdots, S_{i,t}$; and $l = 1, \cdots, K_{i,t,s}$.

Likewise, let $d_{i,t,s,l}$ represent the difficulty level of the l-th item in the s-th test at the t-th day taken by the i-th person. As described in Section 2.1, we model the item difficulties as being nominally specified, but with uncertainties. Thus we write

$$d_{i,t,s,l} = a_{i,t,s} + \epsilon_{i,t,s,l}, \tag{2.4}$$

where $a_{i,t,s}$ indicates the ensemble mean difficulty for the items in the s-th test taken by the i-th person on the t-th day, and $\epsilon_{i,t,s,l}$ is the random deviation from this ensemble mean difficulty for the l-th item within the s-th test. In the scenario we consider, the value of $a_{i,t,s}$ is assumed to be known, from the theoretical analysis of the text complexity, while it is assumed that $\epsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$ with σ^2 being specified from the test design in the CAIT testing.

As mentioned in Section 2.1, we will also incorporate a term of daily random effects, $\varphi_{i,t}$, as well as a term of test random effects, $\eta_{i,t,s}$, to account for the possible local dependent factors when the person i takes several tests during the day t. It is assumed that $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$ and, letting $\eta_{i,t} = (\eta_{i,t,1}, \ldots, \eta_{i,t,S_{i,t}})'$ denote the vector of test random effects on the day t for the individual i, that $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\boldsymbol{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, with differing and unknown precision parameters δ_i and τ_i for each

individual i. The normal distribution for $\eta_{i,t}$ is actually a singular normal distribution because it is conditioned on the sum of the day's test effects being zero, done to remove any possibility of confounding with the daily random effects. (In the analysis and computation, this singular distribution is replaced by the corresponding lower dimensional non-singular normal distribution.)

Finally, at the observation level, the dichotomous test data is modeled as

$$\begin{aligned} & \Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l}) \\ = \; & \mathrm{F}(\theta_{i,t} - d_{i,t,s,l} + \varphi_{i,t} + \eta_{i,t,s}) \\ = \; & \mathrm{F}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}), \end{aligned}$$

where $\theta_{i,t}$ represents the i-th person's ability on the day t; we are thus assuming that a person's ability is constant over a given day, although there could be random fluctuations captured by the $\varphi_{i,t}$ and $\eta_{i,t,s}$. Letting $\mathrm{F}(\cdot)$ be the logistic cdf, as previously discussed, results in

$$\begin{aligned} & \Pr(X_{i,t,s,l} = 1 \mid \theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l}) \\ = \; & \frac{\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}{1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}. \end{aligned} \tag{2.5}$$

2.2.2 The System Equation in DIR Models

As mentioned in Section 2.1, both parametric growth models and Markov chain models have been utilized in contexts similar to that of this chapter. Here we combine these ideas, through a generalization of dynamic linear models, to model an individual's ability growth trajectory over time. The proposed model is

$$\theta_{i,t} = \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t}. \tag{2.6}$$

which has three terms, modeling how the current ability, $\theta_{i,t}$ for the i-th person on the t-th day, relates to the past ability and other factors. The first term is simply the ability at the previous time point, $\theta_{i,t-1}$.

The second term is a parametric growth model. Here c_i can be thought of as the average growth rate of the i-th person's ability over time and $\Delta_{i,t}^{+}$ is the time lapse between the person's t-th test day and $(t-1)$-th test day but truncated by a pre-specified maximum time interval $\Delta_{T_{\max}}$, i.e. $\Delta_{i,t}^{+} = \min\{\Delta_{i,t}, \Delta_{T_{\max}}\}$; thus $c_i\Delta_{i,t}^{+}$ would reflect the ability growth over the given time interval if the growth was indeed linear. However, this growth is truncated at $\Delta_{T_{\max}}$ (chosen herein to be 14 days), reflecting the fact that, when on vacation, the student's ability may not be growing. Furthermore, the growth rate often declines as an ability increases (indeed an ability typically eventually plateaus), so that a linear growth model is often unsuitable when $\theta_{i,t}$ becomes large. The "correction factor," $-\rho\theta_{i,t-1}$ in (2.6), compensates for this effect, slowing down the linear growth as the ability level becomes larger. ρ is the parameter controlling the rate of this adjustment, and could be known or unknown. In our testbed example, ρ is known, based on experiments conducted at MetaMetrics (Hanlon et al. (2010)). In principle, ρ should be individual-specific, but it is distinguishable from c_i only as the individual's ability level is reaching its maturation; our investigation of an ability growth in the testbed data focuses on early age students, so only the c_i are made individual-specific.

As in all dynamic linear models, the third term, $w_{i,t}$ in (2.6), represents the random component of the change in the i-th person's ability on the t-th day. We assume it is $\mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, where ϕ is unknown. Note that this presumes that the random component of a person's ability change has a variance proportional to the time period between test days. Note, also, that we suppose that ϕ is common across individuals. The reason for this is clear from (2.5), in which $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$ has individual-specific δ_i; there would be a substantial risk of confounding in the likelihood between δ_i's and $\phi^{-1}\Delta_{i,t}$ if the time lapse between tests for the student were equally spaced.

It is possible to rewrite (2.6) as a first-order Markov process, and this is beneficial

for computational reasons. Indeed, letting $\lambda_{i,t} = \theta_{i,t} - \rho^{-1}$ and $g_{i,t} = 1 - c_i \rho \Delta_{i,t}^{+}$, the system equation (2.6) becomes

$$\lambda_{i,t} = g_{i,t}\lambda_{i,t-1} + w_{i,t}\,, \tag{2.7}$$

where $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, and this is in the form of a standard dynamic linear model. (Note that c_i and ϕ need to be known for this reduction.)

2.2.3 DIR Model Summary

To sum up, the one-parameter DIR model is constructed in two levels as follows:

$$\begin{aligned}
\text{System equation:} \qquad & \theta_{i,t} = \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta_{i,t}^{+} + w_{i,t}\,, \\
\text{Observation equation:} \qquad & \Pr(X_{i,t,s,l} = 1 \;\mid \theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l}) \\
& = \frac{\exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}{1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})}\,,
\end{aligned}$$

where $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\epsilon_{i,t,s,l} \sim \mathcal{N}(0, \sigma^2)$, $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\boldsymbol{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, and $\Delta_{i,t}^{+} = \min\{\Delta_{i,t}, \Delta_{T_{\max}}\}$, with the $a_{i,t,s}$, ρ, $\Delta_{i,t}$, $\Delta_{T_{\max}}$ and σ being known and $\theta_{i,t}$, c_i, ϕ, δ_i and τ_i being unknown.

2.3 Statistical Inferences for DIR Models

In this section, the Bayesian methods that will be used for statistical inferences in DIR models are described. Computations are based on a Gibbs sampling scheme, in conjunction with a forward filtering and backward sampling.

2.3.1 Prior Distributions for the Unknown Parameters

Prior distributions in a Bayesian analysis must be specified carefully, but they can be either evidence-based priors, reflecting scientific knowledge of the system under study, or they can be objective priors, reflecting a lack of such knowledge but possessing good overall properties – e.g., good frequentist properties (see, e.g., Berger (2006));

a mix of both will be used in the analysis herein. Specifications of evidence-based priors are, of course, context dependent and, here, will be done within the context of the MetaMetrics testbed application.

A natural choice of the prior distribution for an individual's initial latent ability, $\theta_{i,0}$, is

$$\theta_{i,0} \sim \mathcal{N}(\mu_{G_{j_i}}, V_{G_{j_i}}),$$

where $\mu_{G_{j_i}}$ and $V_{G_{j_i}}$ are the mean and the variance, on a logit scale, of the population (j) to which the individual i belongs – for instance, the individual's grade in school for the testbed application. For the average growth rate c_i in the system equation (2.6), the natural objective prior is a constant prior (since c_i is a linear parameter) but we constrain c_i to be positive, reflecting the belief that there is a positive learning rate; thus we choose the prior

$$\pi(c_i) \propto I(c_i > 0) \text{ for all } i.$$

Although ϕ is a scale parameter, it occurs at the system-level of the two-stage model and, hence, the usual scale objective prior ($1/\phi$) would result in an improper posterior; the computationally simplest adjustment is to use $\pi(\phi) = 1/\phi^{3/2}$, which does result in a proper posterior. Similarly, for the scale parameters δ_i and τ_i we utilize the objective priors $\pi(\delta_i) = 1/\delta_i^{3/2}$ and $\pi(\tau_i) = 1/\tau_i^{3/2}$. A natural alternative would be to try to "borrow information" across individuals, by utilizing gamma hyperpriors for δ_i's and τ_i's. This complicates the computation, however, and does not seem necessary for the testbed application.

2.3.2 Posterior Distribution

To facilitate the use of Gibbs sampling techniques in computation, we utilize a mixture of normals representation of the logistic distribution. From Andrews and Mallows (1974), if Y has a logistic distribution with location parameter 0 and scale $\pi^2/3$

($\mathcal{L}(0, \frac{\pi^2}{3})$), one can write the density as

$$f(y) = \frac{e^{-y}}{(1+e^{-y})^2} = \int_0^\infty \left[\frac{1}{\sqrt{2\pi}} \frac{1}{2\nu} \exp\left\{ -\frac{1}{2} (\frac{y}{2\nu})^2 \right\} \right] \pi(\nu) d\nu \,, \tag{2.8}$$

where ν has the Kolmogorov-Smirnov(K-S) density

$$\pi(\nu) = 8 \sum_{\alpha=1}^{\infty} (-1)^{(\alpha+1)} \alpha^2 \nu \exp\{-2\alpha^2 \nu^2\}, \quad \nu \geq 0 \,. \tag{2.9}$$

Note that the density in square brackets in (2.8) is $\mathcal{N}(0, 4\nu^2)$. By using the idea of data augmentation from Tanner and Wong (1987), we consider the latent variable $Y_{i,t,s,l}$ for each response variable $X_{i,t,s,l}$, where $Y_{i,t,s,l} \sim \mathcal{N}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}, 4\nu^2_{i,t,s,l})$ and define $X_{i,t,s,l} = 1$ if $Y_{i,t,s,l} > 0$ and $X_{i,t,s,l} = 0$ otherwise. It is then easy to show that $\Pr(X_{i,t,s,l} = 1 | \theta_{i,t}, a_{i,t,s}, \varphi_{i,t}, \eta_{i,t,s}, \epsilon_{i,t,s,l}) = \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}) / (1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l}))$, so that the introduction of the latent variables $Y_{i,t,s,l}$ will not alter the model (except that there are now formally many more unknown parameters).

As $\epsilon_{i,t,s,l} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, it can be marginalized out in the distribution of $Y_{i,t,s,l}$, resulting in $Y_{i,t,s,l} \sim \mathcal{N}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, 4\nu^2_{i,t,s,l} + \sigma^2)$. Therefore, the one-parameter DIR models (2.5) and (2.6) can be rewritten, with latent variables $\{Y_{i,t,s,l}\}$, as

$$\theta_{i,t} = \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t}, \tag{2.10}$$

$$Y_{i,t,s,l} = \theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \xi_{i,t,s,l} \tag{2.11}$$

$$\nu_{i,t,s,l} \sim \text{K-S } \textit{distribution} \tag{2.12}$$

where $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\varphi_{i,t} \sim \mathcal{N}(0, \delta_i^{-1})$, $\eta_{i,t} \sim \mathcal{N}_{S_{i,t}}(0, \tau_i^{-1}\boldsymbol{I} \mid \sum_{s=1}^{S_{i,t}} \eta_{i,t,s} = 0)$, and $\xi_{i,t,s,l} \sim \mathcal{N}(0, \psi^{-1}_{i,t,s,l})$ with $\psi^{-1}_{i,t,s,l} = 4\nu^2_{i,t,s,l} + \sigma^2$.

Define $\theta = (\theta_1, \cdots, \theta_n)'$, where $\theta_i = (\theta_{i,0}, \theta_{i,1}, \cdots, \theta_{i,T_i})'$, $c = (c_1, \cdots, c_n)'$ and $\tau = (\tau_1, \cdots, \tau_n)'$ for $i = 1, \cdots, n$; $Y = \{Y_{i,t,s,l}\}$, $\nu = \{\nu_{i,t,s,l}\}$ and $X = \{X_{i,t,s,l}\}$

for $l = 1, \cdots, K_{i,t,s}$, $s = 1, \cdots, S_{i,t}$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, n$; $\varphi = \{\varphi_{i,t}\}$ for $t = 1, \cdots, T_i$, $i = 1, \cdots, n$; $\eta = \{\eta_{i,t,s}\}$ for $s = 1, \cdots, S_{i,t}$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, n$ and $\eta^*_{i,t} = (\eta_{i,t,1}, \cdots, \eta_{i,t,S_{i,t}-1})'$. Then the joint posterior density of θ, Y, c, τ, φ, η, ν and ϕ given the data X, in the one-parameter DIR model, is proportional to

$$
\begin{aligned}
& \pi(\theta, Y, c, \tau, \varphi, \eta, \nu, \phi \mid X) \qquad (2.13) \\
\propto \; & \left\{ \prod_{i=1}^{n} \pi(\theta_{i,0}) \pi(c_i) \pi(\delta_i) \pi(\tau_i) \right\} \pi(\phi) \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} \pi(\nu_{i,t,s,l}) \right\} \\
\cdot \; & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} \left(I\{Y_{i,t,s,l} > 0\} I\{X_{i,t,s,l} = 1\} + I\{Y_{i,t,s,l} \le 0\} I\{X_{i,t,s,l} = 0\} \right) \right. \\
\cdot \; & \left. \sqrt{\frac{\psi_{i,t,s,l}}{2\pi}} \exp\left(-\frac{\psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2}\right) I\{\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s}\} \right\} \\
\cdot \; & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \left(\frac{\tau_i}{2\pi}\right)^{\frac{S_{i,t}-1}{2}} \exp\left(-\frac{\tau_i \eta^{*}_{i,t}{}' \Sigma^{-1}_{i,t} \eta^*_{i,t}}{2}\right) \right\} \left\{ \prod_{i-1}^{n} \prod_{t=1}^{T_i} \sqrt{\frac{\delta_i}{2\pi}} \exp\left(-\frac{\delta_i \varphi^2_{i,t}}{2}\right) \right\} \\
\cdot \; & \left\{ \prod_{i=1}^{n} \prod_{t=1}^{T_i} \sqrt{\frac{\phi}{2\pi\Delta_{i,t}}} \exp\left(-\frac{\phi\{\theta_{i,t} - \theta_{i,t-1} - c_i(1 - \rho\theta_{i,t-1})\Delta^+_{i,t}\}^2}{2\Delta_{i,t}}\right) \right\}
\end{aligned}
$$

where

$$
\Sigma^{-1}_{i,t} = \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & \cdots & 2 \end{pmatrix}_{(S_{i,t}-1)\times(S_{i,t}-1)},
$$

and $I(Z \in A)$ is the indicator function equal to 1 if the random varible Z is contained in the set A; $\pi(\theta_{i,0})$, $\pi(c_i)$, $\pi(\delta_i)$, $\pi(\tau_i)$, $\pi(\phi)$ are the priors specified in the previous subsection; and $\pi(\nu_{i,t,s,l})$ is the K-S density defined at the beginning of this subsection. This is a proper posterior under very mild conditions; see Appendix A.

2.3.3 Computation

The computation was done by a MCMC scheme that samples from the posterior (2.13) via a block Gibbs sampling scheme, utilizing the forward filtering and backward sampling algorithm at a key point. Because of the block Gibbs sampling scheme, we need only specify the conditional distributions of a block of variables given the data and other unknown variables. The steps of the algorithm are given as below:

- *Step 1*: **Sampling Y: Truncated Normal Distribution Sampling**

  Given θ, φ, η and ν, the latent variables $\{Y_{i,t,s,l}\}$ are sampled from

  $$Y_{i,t,s,l} \sim \mathcal{N}_+(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 1$$
  $$Y_{i,t,s,l} \sim \mathcal{N}_-(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s}, \psi^{-1}_{i,t,s,l}) \quad \text{if} \quad X_{i,t,s,l} = 0\,,$$

  where $\mathcal{N}_+$ means the normal distribution truncated at the left by zero while $\mathcal{N}_-$ is the normal distribution truncated at the right by zero and $\psi^{-1}_{i,t,s,l} = 4\nu^2_{i,t,s,l} + \sigma^2$. Sampling from truncated normals is fast and easy.

- *Step 2*: **Sampling θ: Forward Filtering and Backward Sampling**

  The latent ability vector $\theta_i = (\theta_{i,0}, \cdots, \theta_{i,T_i})$, for each individual, is typically high dimensional with highly correlated coordinates, so sampling of the variables would appear to be highly challenging. To overcome this roadblock, the proposed model was constructed so that θ_i could be block sampled – within a Gibbs sampling step conditional on the other parameters – by the highly efficient forward filtering and backward sampling algorithm.

  To see this, consider ϕ, c, Y, φ, η and ν as given (the Gibbs sampling step). Define $Z_{i,t,s,l} = Y_{i,t,s,l} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s} - \rho^{-1}$ and utilize the formulation of the model in (2.7). Then, the (conditional) one-parameter DIR model fits the

framework of dynamic linear models (West and Harrison (1997)), i.e.

$$\text{system equation:} \qquad \lambda_{i,t} = g_{i,t}\lambda_{i,t-1} + w_{i,t},$$

$$\text{observation equation:} \qquad Z_{i,t,s,l} = \lambda_{i,t} + \xi_{i,t,s,l},$$

where $w_{i,t} \sim \mathcal{N}(0, \phi^{-1}\Delta_{i,t})$, $\xi_{i,t,s,l} \sim \mathcal{N}(0, \psi^{-1}_{i,t,s,l})$ with $\psi^{-1}_{i,t,s,l} = 4\nu^2_{i,t,s,l} + \sigma^2$. As indicated in West and Harrison (1997), the forward filtering and backward sampling algorithm to block update each vector θ_i proceeds as follows.

Since $\lambda_{i,0} = \theta_{i,0} - \rho^{-1}$ and $\theta_{i,0} \sim \mathcal{N}(\mu_{G_j}, V_{G_j})$, the conditional prior for $\lambda_{i,0}$ is $\lambda_{i,0} \sim \mathcal{N}(\mu_{G_j} - \rho^{-1}, V_{G_j})$. Define information available on the t-th day for the i-th person as

$$D_{i,t} = \{g_{i,q}, \phi, \psi, \varphi, \eta, c, Z_{i,q,1,1}, \cdots, Z_{i,q,S_{i,q},K_{i,q,S_{i,q}}}\}_{q=1}^{t}.$$

We claim that the posterior distribution of $\lambda_{i,t}$ is then

$$\lambda_{i,t} \mid D_{i,t} \sim \mathcal{N}(\mu_{i,t}, V_{i,t}), \tag{2.14}$$

which can be verified by induction as follows. Assume that, on the $(t-1)$-th day, the posterior of $\lambda_{i,t-1}$, given $D_{i,t-1}$, is $\mathcal{N}(\mu_{i,t-1}, V_{i,t-1})$. And it is easy to see this assumption is true when $t = 1$. Then, from the system equation, it is easy to establish that $\lambda_{i,t} \mid D_{i,t-1} \sim \mathcal{N}(d_{i,t}, R_{i,t})$ is a prior for $\lambda_{i,t}$, where $d_{i,t} = g_{i,t}\mu_{i,t-1}$ and $R_{i,t} = g^2_{i,t}V_{i,t-1} + \phi^{-1}\Delta_{i,t}$. Therefore, we have

$$\begin{aligned}
\Pr(\lambda_{i,t} \mid D_{i,t}) &\propto \Pr(\lambda_{i,t} \mid D_{i,t-1}) \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} \Pr(Z_{i,t,s,l} \mid \lambda_{i,t}) \\
&\propto \exp\{-\frac{R^{-1}_{i,t}(\lambda_{i,t} - d_{i,t})^2}{2}\} \left\{ \prod_{s=1}^{S_{i,t}} \prod_{l=1}^{K_{i,t,s}} \exp\{-\frac{\psi_{i,t,s,l}(Z_{i,t,s,l} - \lambda_{i,t})^2}{2}\} \right\}.
\end{aligned}$$

Then, at the t-th day, the posterior distribution of $\lambda_{i,t}$ is as (2.14), where $\mu_{i,t} = V_{i,t}(R^{-1}_{i,t}d_{i,t} + \sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l}Z_{i,t,s,l})$ and $V_{i,t} = (\sum_{s=1}^{S_{i,t}} \sum_{s=1}^{K_{i,t,s}} \psi_{i,t,s,l} + R^{-1}_{i,t})^{-1}$.

The above updating procedure is called forward filtering and after it is complete and all quantities, i.e. $\mu_{i,t}$ and $V_{i,t}$ are saved, we can begin the backward sampling of $\lambda_{i,t}$. For the time $t = T_i$, we sample $\lambda_{i,t}$ directly from $\mathcal{N}(\mu_{i,T}, V_{i,T})$. As the time from $t = (T_i - 1)$ to 0, at each time we draw $\lambda_{i,t}$ from

$$\lambda_{i,t} \mid \lambda_{i,t+1}, D_{i,t} \sim \mathcal{N}(h_{i,t}, H_{i,t})$$

where $h_{i,t} = H_{i,t}(V_{i,t}^{-1}\mu_{i,t} + \phi g_{i,t+1}\Delta_{i,t+1}^{-1}\lambda_{i,t+1})$ and $H_{i,t} = (\phi g_{i,t+1}^2\Delta_{i,t+1}^{-1} + V_{i,t}^{-1})^{-1}$. This follows from

$$\begin{aligned}\Pr(\lambda_{i,t} \mid \lambda_{i,t+1}, D_{i,t}) &\propto \Pr(\lambda_{i,t} \mid D_{i,t})\Pr(\lambda_{i,t+1} \mid \lambda_{i,t}, D_{i,t}) \\ &\propto \exp\{-\frac{V_{i,t}^{-1}(\lambda_{i,t} - \mu_{i,t})^2}{2}\}\exp\{-\frac{\phi\Delta_{i,t+1}^{-1}(\lambda_{i,t+1} - g_{i,t+1}\lambda_{i,t})^2}{2}\}.\end{aligned}$$

Thus, for $t = 0, \cdots, T_i$, we set $\theta_{i,t} = \lambda_{i,t} + \rho^{-1}$ and each vector θ_i is sampled as a whole block, noticing that

$$\Pr(\theta_i \mid D_{i,T_i}) = \Pr(\theta_{i,T_i} \mid D_{i,T_i})\Pr(\theta_{i,T_i-1} \mid \theta_{i,T_i}, D_{i,T-1}) \cdots \Pr(\theta_{i,0} \mid \theta_{i,1}, D_{i,0}).$$

- *Step 3*: **Sampling c: Truncated Normal Distribution Sampling**

  When θ and ϕ are given, the full conditional distribution of c_i is the truncated normal distribution

$$c_i \sim \mathcal{N}_+\left(\frac{\sum_{t=1}^{T_i}(1-\rho\theta_{i,t-1})(\theta_{i,t}-\theta_{i,t-1})\Delta_{i,t}^{+}\Delta_{i,t}^{-1}}{\sum_{t=1}^{T_i}(\Delta_{i,t}^{+}(1-\rho\theta_{i,t-1}))^2\Delta_{i,t}^{-1}}, \frac{1}{\phi\sum_{t=1}^{T_i}(\Delta_{i,t}^{+}(1-\rho\theta_{i,t-1}))^2\Delta_{i,t}^{-1}}\right).$$

- *Step 4*: **Sampling η: Multivariate Normal Distribution Sampling**

  When θ, φ, τ, Y and ν are given, if $S_{i,t} > 1$, then the full conditional distribution of $\eta_{i,t}^*$ is the multivariate normal distribution

$$\eta_{i,t}^* \sim \mathcal{N}_{S_{i,t}-1}\left((A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}A_{i,t} + \tau_i\Sigma_{i,t}^{-1})^{-1}A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}Y_{i,t}^*, (A_{i,t}^T\Sigma_{\psi_{i,t}}^{-1}A_{i,t} + \tau_i\Sigma_{i,t}^{-1})^{-1}\right),$$

where $Y^*_{i,t} = (Y_{i,t,1,1} - \theta_{i,t} + a_{i,t,1} - \varphi_{i,t}, \cdots, Y_{i,t,1,K_{i,t,1}} - \theta_{i,t} + a_{i,t,K_{i,t,1}} - \varphi_{i,t}, \cdots, Y_{i,t,S_{i,t},K_{i,t,S_{i,t}}} - \theta_{i,t} + a_{i,t,K_{i,t,S_{i,t}}} - \varphi_{i,t})'$, $\Sigma^{-1}_{\psi_{i,t}} = \mathrm{diag}((\psi_{i,t,1,1}, \cdots, \psi_{i,t,S_{i,t},K_{i,t,S_{i,t}}})')$,

$$A_{i,t} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & -1 \end{pmatrix}_{(\sum_{s=1}^{S_{i,t}} K_{i,t,s}) \times (S_{i,t}-1)},$$

and $\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1} \eta_{i,t,s}$. When $S_{i,t} = 1$, $\eta_{i,t,S_{i,t}} = 0$.

- *Step 5*: **Sampling τ: Gamma Distribution Sampling**

  When η is given, the full conditional distribution of τ_i is the gamma distribution

$$\tau_i \sim \mathcal{G}a\left(\frac{\sum_{t=1}^{T_i} S_{i,t} - (T_i + 1)}{2}, \frac{\sum_{t=1}^{T_i} \eta^{*}_{i,t}{}' \Sigma^{-1}_{i,t} \eta^{*}_{i,t}}{2}\right).$$

- *Step 6*: **Sampling φ: Normal Distribution Sampling**

  When θ, η, Y and ν are given, the full conditional distribution of $\varphi_{i,t}$ is the normal distribution

$$\varphi_{i,t} \sim \mathcal{N}\left(\frac{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l}(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \eta_{i,t,s})}{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l} + \delta_i}, \frac{1}{\sum_{s=1}^{S_{i,t}} \sum_{l=1}^{K_{i,t,s}} \psi_{i,t,s,l} + \delta_i}\right).$$

- *Step 7*: **Sampling δ: Gamma Distribution Sampling**

  When φ is given, the full conditional distribution of δ_i is the gamma distribution

$$\delta_i \sim \mathcal{G}a\left(\frac{T_i - 1}{2}, \frac{\sum_{t=1}^{T_i} \varphi^2_{i,t}}{2}\right).$$

- *Step 8*: **Sampling ϕ: Gamma Distribution Sampling**

  When θ, c is given, the full conditional distribution of ϕ is the gamma distribution

$$\phi \sim \mathcal{G}a\left(\frac{\sum_{i=1}^{n} T_i - 1}{2}, \frac{\sum_{i=1}^{n}\sum_{t=1}^{T_i} \Delta_{i,t}^{-1}(\theta_{i,t} - \theta_{i,t-1} - c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^{+})^2}{2}\right).$$

- *Step 9*: **Sampling ν : Metropolis-Hastings Sampling**

  Given Y, θ, φ and η, the full conditional distribution of $\nu_{i,t,s,l}$ is proportional to

$$\pi(\nu_{i,t,s,l}|Y,\theta,\varphi,\eta) \propto \sqrt{\frac{1}{\sigma^2 + 4\nu_{i,t,s,l}^2}} \exp\left\{-\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2(\sigma^2 + 4\nu_{i,t,s,l}^2)}\right\},$$

  which is not in closed form. So we shall resort to a Metropolis-Hastings scheme to sample this distribution. A suitable proposal for sample ν is K-S distribution itself. Thus, we first sample ν from the K-S distribution whose density is defined in (2.9). Then, we let

$$\nu_{i,t,s,l}^{(M)} = \begin{cases} \nu^*, & \text{with probability} \quad \min(1, LR) \\ \nu_{i,t,s,l}^{(M-1)}, & \text{otherwise} \end{cases}$$

  where, given Y, θ, φ and η,

$$\begin{aligned} LR &= \sqrt{\frac{\sigma^2 + 4(\nu_{i,t,s,l}^{(M-1)})^2}{\sigma^2 + 4(\nu^*)^2}} \exp\left\{-\frac{(Y_{i,t,s,l} - \theta_{i,t} + a_{i,t,s} - \varphi_{i,t} - \eta_{i,t,s})^2}{2}\right. \\ &\cdot \left.\left(\frac{1}{\sigma^2 + 4(\nu^*)^2} - \frac{1}{\sigma^2 + 4(\nu_{i,t,s,l}^{(M-1)})^2}\right)\right\}, \end{aligned}$$

  and M indicates the M-th iteration step in MCMC.

Therefore, the Gibbs sampling starts at step 1, with initial values for $\theta^{(0)}$, $c^{(0)}$, $\phi^{(0)}$, $\varphi^{(0)}$, $\eta^{(0)}$, $\delta^{(0)}$, $\tau^{(0)}$ and $\nu^{(0)}$, and then loops through step 9 until the MCMC has converged. The initial values chosen in the applications were $\theta^{(0)} = \vec{0}$, $c^{(0)} = \vec{0}$, $\phi^{(0)} = 1$, $\varphi^{(0)} = \vec{0}$, $\eta^{(0)} = \vec{0}$, $\delta^{(0)} = \vec{1}$, $\tau^{(0)} = \vec{1}$ and $\nu^{(0)} = \vec{1}$. The convergence was evaluated informally by looking at trace plots, and was found to obtain at most after 30,000 to 50,000 iterations in the examples.

From the MCMC samples, statistical inferences are straightforward. For example, an estimate and 95% credible interval for the latent ability trait $\theta_{i,t}$ can be formed from the median, 2.5%, and 97.5% empirical quantiles of the corresponding MCMC realizations. In examples, these will be graphed as a function of t, so that the adaptive nature of the model is apparent.

2.4 A Simulated Example

In this section, a simulated example is used to illustrate the inferences from the proposed one-parameter DIR models, and to study their properties, primarily from a frequentist perspective.

The simulation examines the model's behavior for multiple individuals taking a series of tests that are scheduled during different time periods. In particular, suppose there are 10 individuals, each of whom has taken tests on 50 different days. Thus $n = 10$ and $T_i = 50$, for $i = 1, \cdots, 10$. During each distinctive test day, the individual takes four tests; thus $S_{i,t} = 4$ for $t = 1, \cdots, 50$, $i = 1, \cdots, 10$. Each test consists of 10 items, so that $K_{i,t,s} = 10$ for $s = 1, \cdots, 4$, $t = 1, \cdots, 50$ and $i = 1, \cdots, 10$. For the i-th person, the time lapse between two different tests is assumed to be a function of the t-th day, i.e., $\Delta_{i,t} = 10 + t$, for $i = 1, \cdots, 10$, $t = 1, \cdots, T_i/2$ and $\Delta_{i,t} = t - 10$, for $t = T_i/2, \cdots, T_i$. Finally, the unknown values of parameters in the models are chosen as follows:

- $\phi = 1/0.0218^2$, and the corresponding standard deviation of the random component $w_{i,t}$ in the system equation (2.6) is $0.0218\sqrt{\Delta_{i,t}}$.

- $c = (0.0055, 0.0065, 0.0026, 0.0037, 0.0061, 0.0047, 0.0035, 0.0043, 0.0039, 0.0015)'$, where each element in the vector c corresponds to i-th person's average growth rate, respectively, for $i = 1, \cdots, 10$.

- $\delta = (2.0408, 1.3333, 1.8182, 1.2346, 1.5873, 1, 2.2222, 1.0526, 1.1494, 2)'$, where each element in the vector δ corresponds to the precision parameter of daily random effects for the i-th person, respectively, $i = 1, \cdots, 10$.

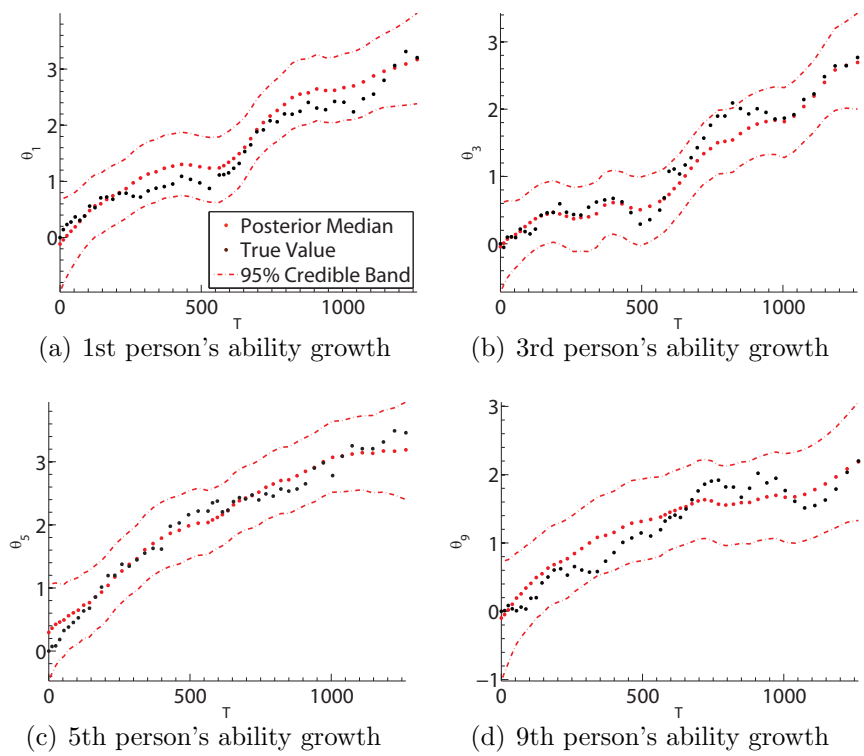
- $\tau = (4, 3.1250, 4.3478, 2.7027, 3.7037, 2.8571, 4, 2.2222, 9.0909, 4.5455)'$, where each element in the vector τ corresponds to the precision parameter of test random effects for the i-th person, respectively, $i = 1, \cdots, 10$.

According to the observation equation (2.5), we then simulated values for the unknown variables and set the test difficulties, $a_{i,t,s}$, to be $\theta_{i,t} + \zeta$, where ζ is a random variable with the uniform distribution on (-0.1,0.1). The values of $\epsilon_{i,t,s,l}$ were drawn from $\mathcal{N}(0, 0.7333^2)$ and the value of 0.7333 is used in the test design for MetaMetrics. Finally, we chose $\rho = 0.1180$, which is the value estimated by MetaMetrics in their studies (Hanlon et al. (2010)).

From dichotomous data obtained from the simulation, the Bayesian machinery from Section 2.3.3 was used in estimating the model parameters in (2.5) and (2.6). Figure 2.1 below shows estimates of the ability trajectory for the 1st, 3rd, 5th and 9th individuals. The red dots in the figures correspond to the estimated posterior median of the ability $\theta_{i,t}$ at the t-th day for the i-th person, and the red dashed lines give the 2.5% and 97.5% quantile trajectories of $\theta_{i,t}$, for $t = 1, \cdots, 50$. The black dots are the real abilities at the t-th day for the i-th person in the simulation. The third trajectory is typical of what is expected in terms of an increasing ability,

and is smoothly handled by the Bayesian machinery. The other three trajectories are highly non-monotonic; the Bayesian estimates err in trying to be increasing (as they are designed to do), but do adapt to the non-monotonicity when the evidence becomes strong enough.

One method of evaluating the success of the inferential scheme is to evaluate the percentage of time that the true ability, $\theta_{i,t}$, is contained in the 95% credible interval of the estimated ability for each individual. For the ten individuals, these estimated coverages were 100%, 100%, 99%, 99%, 100%, 100%, 94%, 100%, 100%, and 91%, which produce an overall estimated coverage of 98.3%. Thus, while the inferential method is Bayesian, it seems to be yielding sets that have good frequentist coverage.



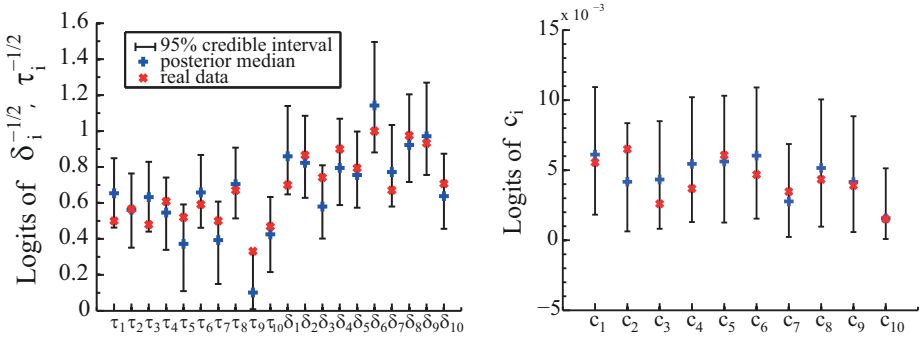
(a) 1st person's ability growth (b) 3rd person's ability growth

(c) 5th person's ability growth (d) 9th person's ability growth

FIGURE 2.1: Estimated and actual ability trajectories of 4 individuals from the simulated data.

To summarize the results for the c_i's, $\tau_i^{-1/2}$'s and $\delta_i^{-1/2}$'s , we compare their true

values with the corresponding estimated values in Figure 2.2. In these plots, the black bar represents the 95% credible interval of the posterior distribution. The blue plus stands for the estimated posterior median and the red cross is the true value in the simulation. Moreover, the estimated posterior median of $\phi^{-1/2}$ is 0.0315 and its 95% credible interval is $[0.0148, 0.0484]$.



(a) 95% credible intervals of $\frac{1}{\sqrt{\delta_i}}$ and $\frac{1}{\sqrt{\tau_i}}$ (b) 95% credible intervals of c_i

FIGURE 2.2: 95% credible intervals of c_i, $\frac{1}{\sqrt{\tau_i}}$ and $\frac{1}{\sqrt{\delta_i}}$, for $i = 1, \cdots 10$ with the simulated data.

Note that the true values of the c_i's, $\tau_i^{-1/2}$'s, $\delta_i^{-1/2}$'s and ϕ are all contained in the 95% credible intervals except $\tau_9^{-1/2}$; thus the empirical coverage for those parameters is 96.77%.

2.5 MetaMetrics Testbed

In this section, we apply the DIR model to the testbed MetaMetrics data. A sample of 25 individuals from the data base of students in certain elementary schools in Mississippi is considered here; the differing characteristics of the students are described in Table 2.1. The primary focus is the goals of this study mentioned in Section 2.1.2.

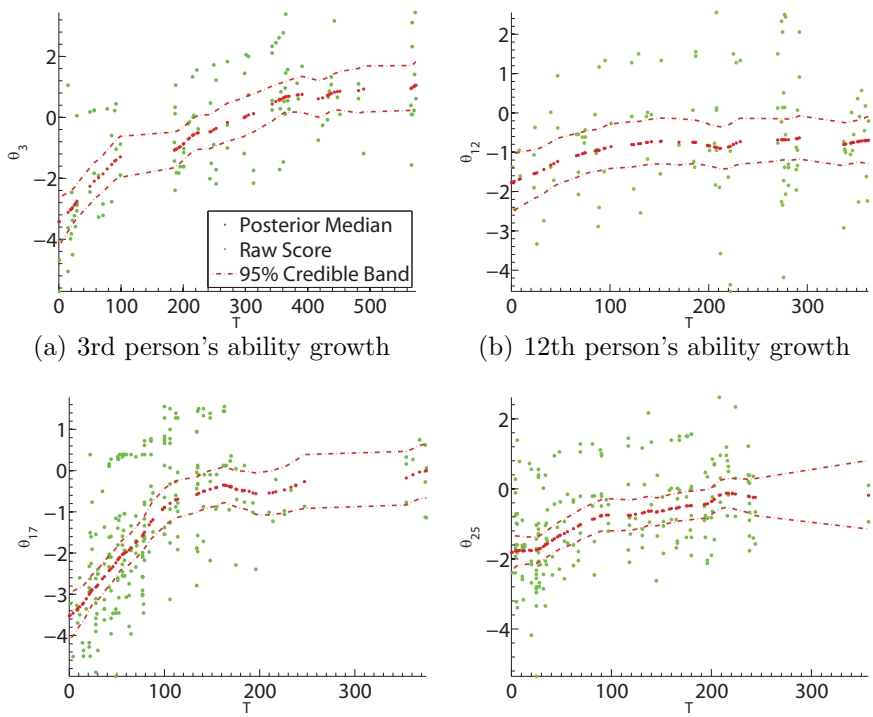
Table 2.1: Characteristics of the 25 considered individuals from the dataset collected by the MetaMetrics

	Total Tests	Days	Max. Tests/Days	Range of Items/Test	Max. Gap	Grade
No.1	147	73	8	4-25	105	4
No.2	162	64	9	3-17	102	2
No.3	118	77	4	3-21	87	2
No.4	93	53	4	5-25	147	2
No.5	114	89	3	6-25	109	2
No.6	157	57	29	4-20	116	2
No.7	153	63	7	4-20	97	2
No.8	60	50	5	3-24	168	6
No.9	135	53	7	4-24	93	2
No.10	137	54	6	4-17	219	1
No.11	214	100	11	3-18	108	2
No.12	113	76	4	4-16	45	2
No.13	95	65	4	4-14	113	2
No.14	116	57	6	5-17	107	2
No.15	155	71	9	4-20	107	1
No.16	247	76	13	3-19	113	2
No.17	254	76	12	3-18	107	2
No.18	304	53	31	3-12	49	2
No.19	167	83	5	3-23	58	2
No.20	101	68	9	4-23	117	2
No.21	88	58	9	3-23	110	2
No.22	220	96	8	2-23	104	3
No.23	80	66	6	2-25	93	6
No.24	105	60	6	6-24	62	3
No.25	218	74	12	3-25	113	2

2.5.1 Retrospective Estimation of an Ability Growth

First consider retrospective estimation of the reading ability for an individual, utilizing all the data recorded for that individual. Figure 2.3 presents the resulting growth trajectories for the 3rd, 12th, 17th and 25th individuals studied. In Figure 2.3, the red dots are the posterior median estimates of each individual ability and the red dash lines correspond to the 2.5% and 97.5% quantiles of the posterior distributions of the abilities, while the green dots correspond to the estimates of individuals' abilities obtained by solving the equation that the expectation of expected score for the

person's ability is equivalent to the observed score; these can roughly be thought of as the raw test scores put on the same scale as the $\theta_{i,t}$. The most interesting feature of these growth trajectories is that, while indeed there typically does appear to be an overall growth in an ability, this growth needs not be monotone. In particular, when there is a large time gap between subsequent tests, the ability appears to drop for some individuals. One natural explanation is that, during vacations, a student may not read and could actually lose the ability. Another possible explanation is that the student has become less adept at implementation of CAIT after a long break.

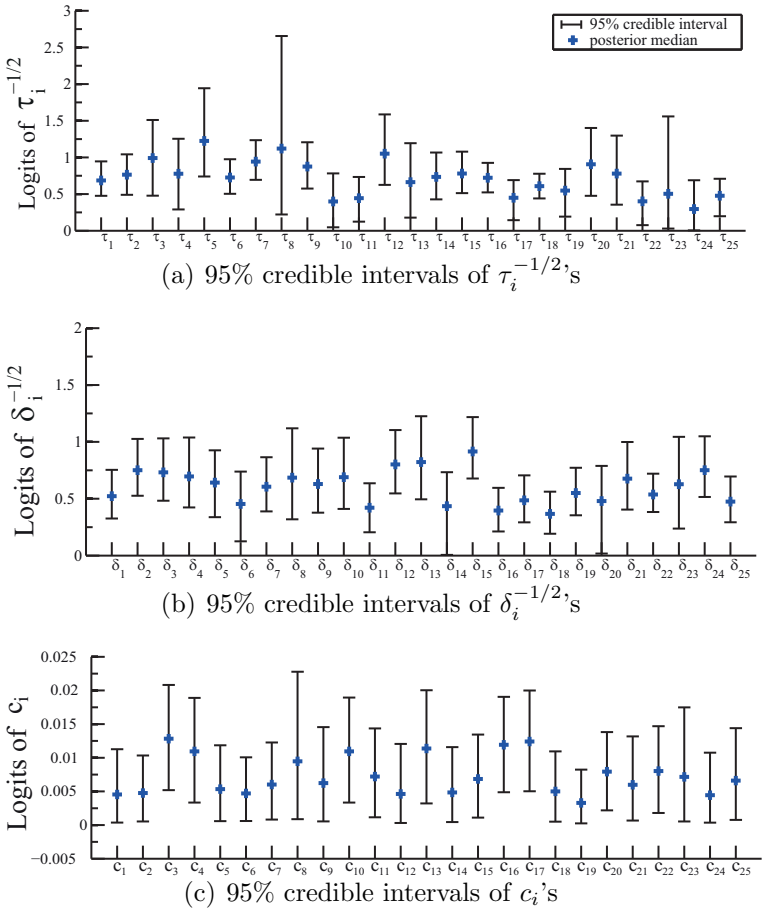


(a) 3rd person's ability growth (b) 12th person's ability growth

(c) 17th person's ability growth (d) 25th person's ability growth

FIGURE 2.3: Estimated ability trajectories of 4 individuals from the dataset collected by the MetaMetrics.

Figure 2.4 gives the summaries of the posterior distributions of the standard deviations of test random effects, $\tau_i^{-1/2}$'s, the standard deviations of daily random effects, $\delta_i^{-1/2}$'s and the average growth rates c_i's, for $i = 1, \cdots, 25$. Moreover, the estimated



(a) 95% credible intervals of $\tau_i^{-1/2}$'s

(b) 95% credible intervals of $\delta_i^{-1/2}$'s

(c) 95% credible intervals of c_i's

FIGURE 2.4: 95% credible intervals of the $\tau_i^{-1/2}$'s, $\delta_i^{-1/2}$'s and c_i's with the Meta-Metrics dataset.

posterior median of $\phi^{-1/2}$ is 0.0612 and its 95% credible interval is $[0.0477, 0.0757]$.

Figures 2.4 (a) and (b) show that the standard deviations of two random effects are almost all quite large with 95% credible intervals well separated from zero. Recall that these were included in the model to account for a possible lack of the local independence; the evidence is thus strong that the local independence is, indeed, not tenable for this data and that both types of random effects are present. The consistency of the standard deviations of the random effects across individuals is somewhat surprising, but lends credence to the notion that random effect modeling of the local dependence is fruitful.

2.5.2 On-line Estimation of an Ability Growth

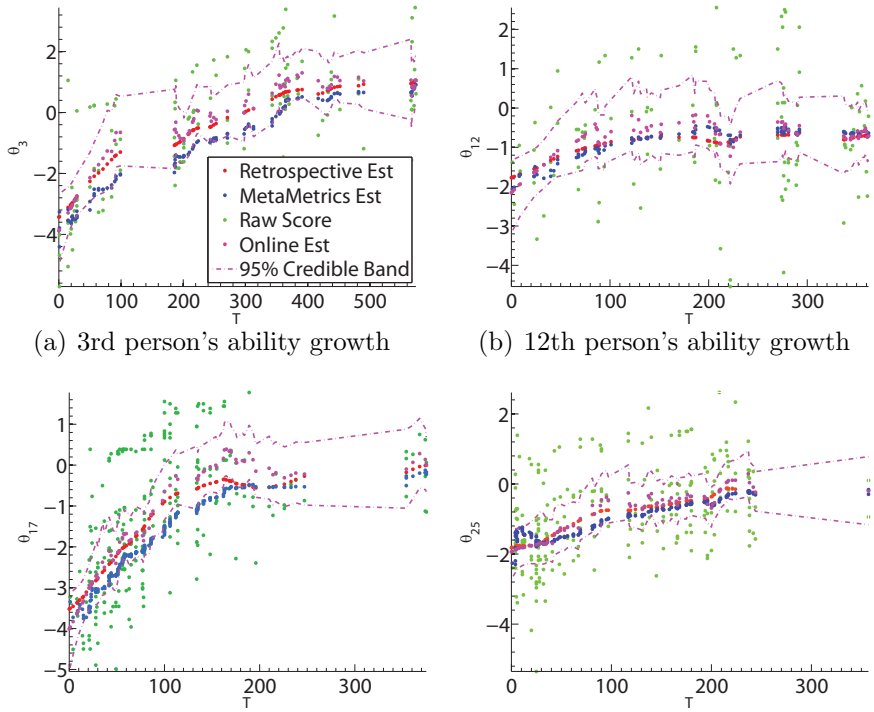
In on-line estimation of a reading ability, essentially the same model is used but, at each time point, only the data up to that time is utilized. Instead of having $\phi^{-1/2}$ unknown, however, we utilize $\phi^{-1/2} = 0.0612$, the estimate arising from the retrospective analysis; $\phi^{-1/2}$ cannot be effectively estimated in the on-line mode.

Applying the Bayesian methodology yields the on-line posterior median ability estimates, as well as the 2.5% and 97.5% quantiles of the posterior distributions of abilities for the 25 individuals being studied; these are the purple dots and dashed purple lines in Figure 2.5, shown for the 3rd, 12th, 17th and 25th individuals. Again the green dots show the raw score estimates of each individual ability at each time point, and the red dots are the retrospective estimates discussed earlier. In these figures we also include, as blue dots, the ability estimates obtained from the current methodology of MetaMetrics, which is a partial Bayesian procedure.

As expected, the on-line ability estimates are much more variable than the retrospective estimates. Sometimes, the on-line estimates seem to be somewhat more variable than the current MetaMetrics estimates (the blue dots). This is because at each online estimation point, the current methodology of MetaMetrics uses a very tight prior (arising from the previous data) for the student's ability.

While we do not know the truth here, it is plausible that the retrospective red dots are our best guesses as to the true abilities, and we can then judge how well the various on-line procedures are doing relative to these best guesses. Our on-line estimates are generally closer to these retrospective estimates than the current MetaMetrics estimates (the 12th individual being the interesting exception). In fact, the average mean squared error of the on-line estimates relative to the retrospective estimates is 0.0851, while the average mean squared error of the current MetaMetrics estimates is 0.1311.

If we do view the retrospective estimates (red dots) as surrogates for the truth, it is interesting to see how often these fall outside the on-line uncertainty bands (purple lines). This happened very rarely; individual 17 in Figure 2.5 was one case in which this sometimes happened. One final observation from Figure 2.5 is that the current MetaMetric estimates usually are lower than our on-line estimates of the person's reading ability.



(a) 3rd person's ability growth (b) 12th person's ability growth

(c) 17th person's ability growth (d) 25th person's ability growth

FIGURE 2.5: On-line estimates of ability trajectories of 4 individuals from the MetaMetrics data.

2.6 Conclusions and Generalizations

The evidence of the violation of the local dependence assumption in CAIT situations is generally strong, and use of test and daily random effects to model the local dependence seems to be necessary and successful. Embedding a dynamic linear model

framework for an individual's ability trajectory within the logistic IRT structure provides a powerful and individually adaptive method for dealing with longitudinal testing data.

The retrospective DIR model analysis seems excellent for assessing actual ability trajectories and, hence, is of considerable use in understanding population behaviors, such as the frequently observed drops in some abilities after a long pause in testing. The on-line DIR analysis provides real-time ability estimates for assignments of materials at the right difficulty level and other possible educational goals.

A key advantage of the Bayesian framework adopted is that uncertainty in all unknowns can be built into the model (e.g., uncertainty in the difficulty of the random test items), and uncertainty of the estimates is available for all inferences. Also, prior information (e.g. knowledge about ability distributions over the population and knowledge that general growth in ability is expected) can be built into the analysis, in a non-dogmatic fashion that allows the data to overrule the prior.

Many extensions are possible, such as the already mentioned extension to two-parameter and three-parameter IRT models. If one also had data for individuals over a period of many years – including years near the maturation point in one's reading ability – it would be possible to include individual-specific ρ_i in the model.

3

Estimating Shape Constrained Functions Using Gaussian Processes

This chapter discusses how to introduce shape constraints such as the class of monotonic functions or convex functions in the Gaussian process priors over the distributions of the nonparametric functions and the paper Wang and Berger (2011) is based on some of materials presented here. Although Gaussian processes are a popular tool for nonparametric function estimations, there are not amenable to assume shape-constrained structures for the unknown functions. However, for Gaussian processes, which have mean square differentiability (see Section 2.2 of Adler (1981)), for example, a Gaussian process with squared exponential correlation function, shape constraints can be incorporated through the use of the derivative process, which are a joint Gaussian process with the original process. The extent to which this is feasible is discussed. The application in this chapter focuses on emulating computer models, where utilizing Gaussian processes to approximate the response surface is routine and shape-constrained prior knowledge is available for the physical processes approximated. But our proposed methods are easily extended into other contexts of which Gaussian process models are extensively used, for instance, supervised-learning for

both regression and classification in machine learning.

3.1 Literature Review and Motivations

In function estimation, prior knowledge about the function shape, such as monotonicity, convexity or concavity, is often available. Examples include a dose response analysis in medicine, option pricing in finance, growth curves in biology and psychology and many others. Often this is the only knowledge about the function, in which case nonparametric function estimation, subject to the shape constraint, is of interest.

Gaussian processes are a popular tool for the nonparametric function estimations, due to their flexibility, their capability of operating as interpolators, and the fact that they produce an automatic estimate of accuracy of the function estimate. Moreover, much of the computation with Gaussian processes is simply parametric Gaussian computation. Gaussian processes are also mathematically equivalent or closely related to many extensively used models, such as Bayesian linear models, spline models, neural networks and support vector machines (see Rasmussen and Williams (2006)). The application of Gaussian processes is widespread in spatial models of meteorology and geology, in the analysis of computer experiments and time series, in machine learning, and elsewhere.

We consider Gaussian processes with squared exponential correlation functions, because they have the useful property that their derivative processes are also Gaussian processes and are jointly Gaussian with the original processes. This fact allows incorporation of shape constraints in modeling nonparametric functions by imposing restrictions on the derivative processes. For instance, if a function is assumed to monotonically increase, one can impose positivity constraints on the first derivative process of the Gaussian process. Convexity can be induced by imposing positivity constraints on the second derivative of the Gaussian process.

The use of derivative information with Gaussian processes is not new. Solak et al. (2003) considered the situation in which observations are available of both the function and its derivative when modeling nonlinear dynamical systems; utilizing the derivative information dramatically improved estimation performance. Banerjee et al. (2003) examined the directional rates of change for selling prices of individual homes utilizing a dataset comprising residential properties in Baton Rouge in Louisiana in 1992. They formalized and extended the notions of directional finite difference processes and directional derivative processes beyond the concept of mean square differentiability. Stephenson (2010) addressed the gain of efficiency in using derivative information with emulators in the statistical analysis of computer models.

This work is more directly related with that of Riihimäi and Vehtari (2010), who considered imposition of monotonicity at selected points through the derivative process. Their work differs from that herein through the way the positivity constraints are induced (through a probit link, rather than a step function), and the computational procedure utilized in the analysis; they utilized an approximate expectation propagation algorithm, whereas we consider a Gibbs sampling method. In addition, constraints other than monotonicity are considered herein.

There have been numerous other Bayesian approaches to nonparametric function estimation under shape constraints. Lavine and Mockus (1995) utilized a Dirichlet process prior to estimate a monotone function; Perron and Mengersen (2001) used mixtures of triangular distribution functions to obtain a monotone function; Neelon and Dunson (2004) used an autoregressive prior distribution for parameters of basis functions in monotone regression; and Shively et al. (2009) assumed a general discrete random probability measure prior proposed by Ongaro and Cattaneo (2004) on a mixing distribution to model monotone functions. Recently, more general shape constraints, such as convexity or concavity, have been considered in Chang et al. (2007), who used the Bernstein polynomial basis for shape constrained regression,

and Bornjamp (2009) which extended Shively et al. (2009)'s method to convexity and monotone convexity.

One of the advantages of the Gaussian process approach to function estimation under shape constraints is that a single technology can be used for any shape constraints that can be expressed in terms of function derivatives. The purpose of this chapter, however, is not to explicitly compare the approach with other Bayesian methods. The perspective is, instead, that the Gaussian process approach has become standard (for a variety of compelling reasons) in certain important areas, such as emulation of computer models, and we seek to understand the benefits and complications of adding shape constraints to the Gaussian process machinery for such areas. Indeed, we illustrate the benefits with an example of emulation of a computer model of vehicle crashworthiness, from Bayarri et al. (2009).

Section 3.2 presents the basic expressions for the chosen Gaussian processes and their derivatives. Section 3.3 formally describes the models we entertain and introduces two alternative ways to express shape constraint information. Section 3.4 examines the performance of the approach with simulated data. Moreover, the example of emulation of a computer model for vehicle crashworthiness is considered, as this is an example where monotonicity of the response is a natural constraint to add. Section 3.5 presents the conclusions and discussions.

3.2 Gaussian Processes and Their Derivative Processes

As indicated in Chapter 1, a Gaussian process (GP) is a stochastic process $Z(t)$ with inputs $t \in \mathcal{T}$, an open subset of the real line, such that any finite number of realizations of the process have a joint Gaussian distribution. From that chapter, we know that a GP is completely specified by its mean function $m(t)$ and its covariance

function $K(t,t')$, i.e.,

$$\begin{aligned} m(t) &= \mathrm{E}[Z(t)] \\ K(t,t') &= \mathrm{cov}(Z(t),Z(t')) = \mathrm{E}[(Z(t)-m(t))(Z(t')-m(t'))]. \end{aligned}$$

The GP will be written as $Z(t) \sim \mathcal{GP}(m(t), K(t,t'))$.

We focus herein on the *squared exponential (SE)* covariance function

$$K(t,t') = \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right), \tag{3.1}$$

where the characteristic length-scale β and the signal variance σ_z^2 are the parameters of the GP model. To simplify the presentation, we will also initially assume that $m(t) = \mu$ with $\mu \in \mathbb{R}$.

The reason for considering the squared exponential covariance function is that it is well known that the resulting process has derivatives of all orders (refer to Theorem 2.2.2 in Adler (1981)). Furthermore, since differentiation is a linear operator, derivatives of the GP remain a GP. For the first order derivative process, the corresponding variance, covariance and mean function (joint with the original process) are as follows:

$$\begin{aligned} \mathrm{E}\left[\frac{\partial Z(t)}{\partial t}\right] &= \frac{\partial m(t)}{\partial t} = 0, \\ \mathrm{cov}\left[\frac{\partial Z(t)}{\partial t}, \frac{\partial Z(t')}{\partial t'}\right] &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right)\frac{1}{\beta^2}\left(1-\frac{1}{\beta^2}(t-t')^2\right), \\ \mathrm{cov}\left[\frac{\partial Z(t)}{\partial t}, Z(t')\right] &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right)\left(-\frac{1}{\beta^2}(t-t')\right). \end{aligned} \tag{3.2}$$

Similarly, the second order derivative process of the GP is still a GP and its corresponding mean and covariance function (joint with the original process and the first

order derivative process) are as follows:

$$
\begin{aligned}
\mathrm{E}\left[\frac{\partial^2 Z(t)}{\partial t^2}\right] &= \frac{\partial^2 m(t)}{\partial t^2} = 0, \\
\mathrm{cov}\left[\frac{\partial^2 Z(t)}{\partial t^2}, \frac{\partial^2 Z(t')}{\partial t'^2}\right] &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right) \frac{1}{\beta^4}\left(\frac{1}{\beta^4}(t-t')^4 - \frac{1}{\beta^2}6(t-t')^2 + 3\right), \\
\mathrm{cov}\left[\frac{\partial^2 Z(t)}{\partial t^2}, Z(t')\right] &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right)\left(\frac{1}{\beta^4}(t-t')^2 - \frac{1}{\beta^2}\right), \\
\mathrm{cov}\left[\frac{\partial^2 Z(t)}{\partial t^2}, \frac{\partial Z(t')}{\partial t'}\right] &= -\sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right)(t-t')\left(\frac{3}{\beta^4} - \frac{1}{\beta^6}(t-t')^2\right).
\end{aligned}
$$

Similarly one can define any kth order derivative process and its joint distribution with the lower order processes, but we will only need the first two derivative processes herein.

When utilizing the first order derivative process to introduce shape constraints, we utilize the following more concise notation. Suppose $t, t', s, s' \in \mathcal{T}$ and denote

$$
\begin{aligned}
K(t,t') &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-t')^2\right), \\
K^{11}(s,s') &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(s-s')^2\right)\frac{1}{\beta^2}\left(1 - \frac{1}{\beta^2}(s-s')^2\right), \\
K^{01}(t,s) &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-s)^2\right)\left(-\frac{1}{\beta^2}(t-s)\right).
\end{aligned}
$$

Figure 3.1 shows these covariance functions of the first order derivative process in one dimension, with hyperparameters $\beta = 1$ and $\sigma_z = 1$.

Denote the input vectors for the GP and for its derivative by $\mathbf{t}$ and $\mathbf{s}$, of lengths n and m, respectively; and let $\mathbf{Z}(\mathbf{t}) = (Z(t_1), \cdots, Z(t_n))^T$ and $\mathbf{Z}'(\mathbf{s}) = (Z'(s_1), \cdots, Z'(s_m))^T$ denote the vectors of corresponding GP and derivative values; the input vectors could contain some (or all) common inputs. It follows from (3.2) that

$$
\begin{bmatrix} \mathbf{Z}(\mathbf{t}) \\ \mathbf{Z}'(\mathbf{s}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \mathbf{1}_n \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{t},\mathbf{t}) & \mathbf{K}^{\mathbf{01}}(\mathbf{t},\mathbf{s}) \\ \mathbf{K}^{\mathbf{10}}(\mathbf{s},\mathbf{t}) & \mathbf{K}^{\mathbf{11}}(\mathbf{s},\mathbf{s}) \end{bmatrix}\right).
$$

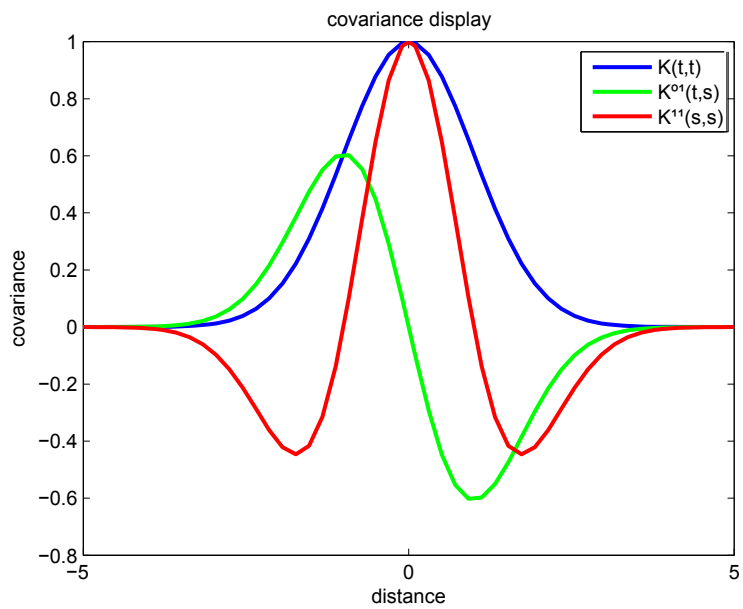


FIGURE 3.1: The GP covariance relationships for $\beta = 1$ and $\sigma_z = 1$.

where $\mathbf{K}(\mathbf{t},\mathbf{t}) = (K(t,t'))$ is the matrix with elements $K(t,t')$, $\mathbf{K^{01}}(\mathbf{t},\mathbf{s}) = (K^{01}(t,s))$, $\mathbf{K^{10}}(\mathbf{s},\mathbf{t}) = \mathbf{K^{01}}(\mathbf{t},\mathbf{s})^T$ and $\mathbf{K^{11}}(\mathbf{s},\mathbf{s}) = (K^{11}(s,s'))$; also $\mathbf{1}_n = (1,\cdots,1)^T_{1\times n}$ and $\mathbf{0_s} = (0,\cdots,0)^T_{1\times m}$.

When utilizing shape constraints on the second derivatives of the GP, we utilize the notation

$$\begin{aligned} K^{22}(s,s') &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(s-s')^2\right)\frac{1}{\beta^4}\left(\frac{1}{\beta^4}(s-s')^4 - \frac{1}{\beta^2}6(s-s')^2 + 3\right), \\ K^{02}(t,s) &= \sigma_z^2 \exp\left(-\frac{1}{2\beta^2}(t-s)^2\right)\left(\frac{1}{\beta^4}(t-s)^2 - \frac{1}{\beta^2}\right). \end{aligned}$$

Then

$$\begin{bmatrix} \mathbf{Z}(\mathbf{t}) \\ \mathbf{Z}''(\mathbf{s}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu\mathbf{1}_n \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{t},\mathbf{t}) & \mathbf{K^{02}}(\mathbf{t},\mathbf{s}) \\ \mathbf{K^{20}}(\mathbf{s},\mathbf{t}) & \mathbf{K^{22}}(\mathbf{s},\mathbf{s}) \end{bmatrix}\right),$$

where $\mathbf{K^{02}}(\mathbf{t},\mathbf{s}) = (K^{02}(t,s))$, $\mathbf{K^{20}}(\mathbf{s},\mathbf{t}) = \mathbf{K^{02}}(\mathbf{t},\mathbf{s})^T$ and $\mathbf{K^{22}}(\mathbf{s},\mathbf{s}) = (K^{22}(s,s'))$.

3.3 Shape Constraints Through the Derivative Processes

A GP can be influenced towards desired shape constraints by constraining its derivatives appropriately. For instance, if an increasing function were desired, one could constrain the GP derivative at a set of inputs to be positive. That is the most direct approach to implementing constraints, and will be considered in Section 3.3.1. A different approach – the development of a new conditional Gaussian process – will be considered in Section 3.3.2.

Before proceeding, it is important to note that one can only constrain the derivative at a discrete set of inputs (i.e., a set of inputs that does not have a limit point). For instance, no GP can have positive derivative at all inputs having a limit point, so constraining the derivative at all points of a nondiscrete set would eliminate all sample paths of the GP.

Since constraints cannot be applied at all inputs, the constrained GP realizations will not strictly follow the constraint. However, through the choice of a dense enough set of constrained inputs, the resulting posterior will be following the constraints for all practical purposes.

3.3.1 Imposing Constraints via Indicator Functions

The Model and Posterior Distribution

Suppose we observe n scalar observations

$$x_i = Z(t_i) + \epsilon_i, \quad i = 1, \ldots n, \tag{3.3}$$

at inputs t_i, where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, and define $\mathbf{X} = (x_1, \cdots, x_n)^T$. (Note that, in emulation of computer models, it is often the case that $\sigma^2 = 0$, i.e., one exactly observes the function being estimated at certain inputs.) Assign $Z(\cdot)$ the GP prior defined in Section 3.2 and suppose that we impose the constraint that $Z(t)$ be nondecreasing at the m points $\mathbf{s} = (s_1, \cdots, s_m)$. Then the marginal constrained prior distribution

of $\mathbf{Z}'(\mathbf{s})$ is simply

$$[\mathbf{Z}'(\mathbf{s})] \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K^{11}}(\mathbf{s},\mathbf{s})\right) \mathbf{1}_{\{Z'(s_i)\geq 0,\ i=1,\cdots,m\}}.$$

Lemma 3.1. *Suppose the goal is to predict $Z(\cdot)$ at a new set of n^* inputs $\mathbf{t}^*$, given the current inputs $\mathbf{t}$ and resulting observations $\mathbf{X}$. The joint conditional posterior distribution of $(\mathbf{Z}(\mathbf{t}^*), \mathbf{Z}'(\mathbf{s}))$ is then*

$$\pi(\mathbf{Z}(\mathbf{t}^*), \mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta) = \pi(\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s}), \mathbf{X}, \theta)\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta),$$

where

$$[\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s}), \mathbf{X}, \theta] \sim \mathcal{N}\left(\mu\mathbf{1}_{n^*} + A^{-1}\left(A_1^T B_1^{-1}(\mathbf{X} - \mu\mathbf{1}_n) + B_2^{-1}A_2\mathbf{Z}'(\mathbf{s})\right), A^{-1}\right), \tag{3.4}$$

$$\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta) \propto \mathcal{N}(\mathbf{m}(\mathbf{s}), \mathbf{S}(\mathbf{s},\mathbf{s}))\, \mathbf{1}_{\{Z'(s_i)\geq 0,\ i=1,\cdots,m\}}, \tag{3.5}$$

with

$$\begin{aligned}
\mathbf{m}(\mathbf{s}) &= \mathbf{K^{10}}(\mathbf{s},\mathbf{t})(\sigma^{\mathbf{2}}\mathbf{I} + \mathbf{K}(\mathbf{t},\mathbf{t}))^{-\mathbf{1}}(\mathbf{X} - \mu\mathbf{1}_n), \\
\mathbf{S}(\mathbf{s},\mathbf{s}) &= \mathbf{K^{11}}(\mathbf{s},\mathbf{s}) - \mathbf{K^{10}}(\mathbf{s},\mathbf{t})(\sigma^{\mathbf{2}}\mathbf{I} + \mathbf{K}(\mathbf{t},\mathbf{t}))^{-\mathbf{1}}\mathbf{K^{01}}(\mathbf{t},\mathbf{s}), \\
A_1 &= \mathbf{K}(\mathbf{t},\mathbf{t}^*)\mathbf{K}(\mathbf{t}^*,\mathbf{t}^*)^{-1}, \\
A_2 &= \mathbf{K^{01}}(\mathbf{t}^*,\mathbf{s})\mathbf{K^{11}}(\mathbf{s},\mathbf{s})^{-1}, \\
B_1 &= \sigma^{\mathbf{2}}\mathbf{I} + \mathbf{K}(\mathbf{t},\mathbf{t}) - \mathbf{K}(\mathbf{t},\mathbf{t}^*)\mathbf{K}(\mathbf{t}^*,\mathbf{t}^*)^{-1}\mathbf{K}(\mathbf{t}^*,\mathbf{t}), \\
B_2 &= \mathbf{K}(\mathbf{t}^*,\mathbf{t}^*) - \mathbf{K^{01}}(\mathbf{t}^*,\mathbf{s})\mathbf{K^{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{K^{10}}(\mathbf{s},\mathbf{t}^*), \\
A &= A_1^T B_1^{-1} A_1 + B_2^{-1},
\end{aligned}$$

and $\theta = (\mu, \sigma^2, \beta, \sigma_z^2)$ are the given parameters of the model and prior.

Proof. For simplicity, we omit reference in this proof to conditioning on θ. It is straightforward to show that the unconstrained joint distribution of $(\mathbf{X}, \mathbf{Z}'(\mathbf{s}))$ is

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Z}'(\mathbf{s}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu\mathbf{1}_n \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} \sigma^2\mathbf{I} + \mathbf{K}(\mathbf{t},\mathbf{t}) & \mathbf{K^{01}}(\mathbf{t},\mathbf{s}) \\ \mathbf{K^{10}}(\mathbf{s},\mathbf{t}) & \mathbf{K^{11}}(\mathbf{s},\mathbf{s}) \end{bmatrix}\right).$$

It is immediate that the unconstrained marginal posterior distribution of $\mathbf{Z}'(\mathbf{s})$ given $\mathbf{X}$ is $[\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}] \sim \mathcal{N}(\mathbf{m}(\mathbf{s}), \mathbf{S}(\mathbf{s},\mathbf{s}))$. Since a constrained posterior is simply proportional to the unconstrained posterior together with the constraint, (3.5) is immediate.

Next, note that

$$\pi(\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s}), \mathbf{X}) \quad \propto \quad f(\mathbf{X} \mid \mathbf{Z}(\mathbf{t}^*))\pi(\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s})).$$

Since $(\mathbf{Z}(\mathbf{t}^*), \mathbf{Z}'(\mathbf{s}))$ has the multivariate normal prior

$$\begin{bmatrix} \mathbf{Z}(\mathbf{t}^*) \\ \mathbf{Z}'(\mathbf{s}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \mathbf{1}_{n^*} \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{t}^*, \mathbf{t}^*) & \mathbf{K^{01}}(\mathbf{t}^*, \mathbf{s}) \\ \mathbf{K^{10}}(\mathbf{s}, \mathbf{t}^*) & \mathbf{K^{11}}(\mathbf{s}, \mathbf{s}) \end{bmatrix} \right),$$

the conditional distribution $\pi(\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s}))$ is

$$[\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s})] \sim \mathcal{N}\left(\mu \mathbf{1}_{n^*} + A_2 \mathbf{Z}'(\mathbf{s}), B_2\right).$$

Similarly, the joint distribution of the observations and the prediction values according to the GP prior is

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Z}(\mathbf{t}^*) \end{bmatrix} \sim \mathcal{N}\left(\mu \begin{bmatrix} \mathbf{1}_n \\ \mathbf{1}_{n^*} \end{bmatrix}, \begin{bmatrix} \sigma^2 \mathbf{I} + \mathbf{K}(\mathbf{t}, \mathbf{t}) & \mathbf{K}(\mathbf{t}, \mathbf{t}^*) \\ \mathbf{K}(\mathbf{t}^*, \mathbf{t}) & \mathbf{K}(\mathbf{t}^*, \mathbf{t}^*) \end{bmatrix} \right).$$

Thus, $f(\mathbf{X} \mid \mathbf{Z}(\mathbf{t}^*))$ is a multivariate normal distribution as well, i.e.

$$[\mathbf{X} \mid \mathbf{Z}(\mathbf{t}^*)] \sim \mathcal{N}\left(\mu \mathbf{1}_n + A_1(\mathbf{Z}(\mathbf{t}^*) - \mu \mathbf{1}_{n^*}), B_1\right).$$

The conditional distribution of $\mathbf{Z}(\mathbf{t}^*)$ given $(\mathbf{Z}'(\mathbf{s}), \mathbf{X})$ is thus as indicated in the lemma, completing the proof.

□

Sampling From the Posterior

To sample from the posterior in (3.4), it is first necessary to sample from the constrained normal distribution $\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta)$ in (3.5). Based on an idea from Alan

et al. (1992), a Gibbs sampling procedure can be developed to sample from this distribution.

Let $j = 1, \cdots, M$ be the iterations of the Gibbs sampler. At step j and for $i = 1, \cdots, m$, draw

$$Z'^{(j+1)}(s_i) \mid \mathbf{Z}'^{(j)}(\mathbf{s}_{-i}), \mathbf{X}, \theta \sim \mathcal{N}_+(\mu_i^{(j)}, \nu_i),$$

where $\mathcal{N}_+$ denotes the normal distribution truncated at the left by 0, $\mathbf{s}_{-i}$ is the set of inputs in $\mathbf{s}$ other than s_i,

$$\begin{aligned} \mu_i^{(j)} &= m(s_i) + \mathbf{S}(s_i, \mathbf{s}_{-i})\mathbf{S}^{-1}(\mathbf{s}_{-i}, \mathbf{s}_{-i})(\mathbf{Z}'^{(j)}(\mathbf{s}_{-i}) - \mathbf{m}(\mathbf{s}_{-i})), \\ \nu_i &= \mathbf{S}(s_i, s_i) - \mathbf{S}(s_i, \mathbf{s}_{-i})\mathbf{S}^{-1}(\mathbf{s}_{-i}, \mathbf{s}_{-i})\mathbf{S}(\mathbf{s}_{-i}, s_i), \end{aligned}$$

$\mathbf{S}(s_i, s_i)$ is the i-th diagonal entry of the covariance matrix $\mathbf{S}(\mathbf{s}, \mathbf{s})$, $\mathbf{S}(s_i, \mathbf{s}_{-i})$ is the i-th row of the covariance matrix $\mathbf{S}(\mathbf{s}, \mathbf{s})$ without the entry from the i-th column, $\mathbf{S}(\mathbf{s}_{-i}, \mathbf{s}_{-i})$ is $\mathbf{S}(\mathbf{s}, \mathbf{s})$ without the i-th row and i-th column, $\mathbf{Z}'^{(j)}(\mathbf{s}_{-i}) = (Z'^{(j+1)}(s_1), \cdots, Z'^{(j+1)}(s_{i-1}), Z'^{(j)}(s_{i+1}), \cdots, Z'^{(j)}(s_m))'$, $\mathbf{m}(\mathbf{s}_{-i})$ is the mean vector $\mathbf{m}(\mathbf{s}_i)$ without the i-th element, and $\mathbf{m}(s_i)$ is the i-th element of $\mathbf{m}(\mathbf{s}_i)$.

After completing these Gibbs updates, one simply block updates the normal distribution $\pi(\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'(\mathbf{s}), \mathbf{X}, \theta)$ in (3.4). Using the Woodbury, Sherman & Morrison formula (see Lemma B.4), $A^{-1} = B_2 - B_2 A_1^T (B_1 + A_1 B_1 A_1^T)^{-1} A_1 B_2$ makes the computation faster and more stable.

Extending to Higher Order Derivative Processes

Suppose that, instead of monotonicity, one wishes to impose convexity constraints on the GP at the m points $\mathbf{s} = (s_1, \cdots, s_m)$. Then, using the notation from Section 3.2, the marginal constrained prior distribution of $\mathbf{Z}''(\mathbf{s})$ is simply

$$[\mathbf{Z}''(\mathbf{s})] \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K^{22}}(\mathbf{s}, \mathbf{s})\right) \mathbf{1}_{\left\{Z''(s_i) \geq 0,\ i=1,\cdots,m\right\}}.$$

The posterior distribution is then described as in Lemma 3.1, with the minor changes of replacing $\mathbf{K^{11}}(\mathbf{s},\mathbf{s})$ by $\mathbf{K^{22}}(\mathbf{s},\mathbf{s})$, $\mathbf{K^{10}}(\mathbf{s},\mathbf{t})$ by $\mathbf{K^{20}}(\mathbf{s},\mathbf{t})$ and $\mathbf{K^{01}}(\mathbf{t},\mathbf{s})$ by $\mathbf{K^{02}}(\mathbf{t},\mathbf{s})$. Sampling from this posterior distribution is also identical to the sampling method above for monotonically constrained functions, with the indicated matrix replacements.

The extension to constraints on higher order derivatives is clear, so that expressions will not be given here. More complicated would be inclusion of constraints on derivatives of differing orders, e.g. simultaneous constraints of monotonicity and convexity, but such can be done in a similar fashion.

3.3.2 Imposing Constraints via a Conditional Gaussian Process

Using the constrained prior on $Z(\cdot)$ from the previous section forces the prior realizations to be strictly increasing (with probability one) at the constraint points $\mathbf{s}$. It is often natural, however, to also allow for the realizations to have derivative zero at the constraint points, catering to the possibility that the function being estimated is essentially constant over a region. (This is of particular importance in the emulation of computer models.)

To allow for this, we define a new process. First, let

$$Z'^{+}(t) = Z'(t) \vee 0,$$

where $Z'(t) \sim \mathcal{GP}\left(0, K^{11}(t,t')\right)$ is the GP derivative process. Note that $Z'^{+}(t)$ now has positive probability of being 0. Then conditioned on $Z'^{+}(\cdot)$ evaluated at m virtual points, $\mathbf{s}$, define the conditional Gaussian process

$$Z^{\triangle}(t) \triangleq Z(t) \mid \{Z'^{+}(s)\}_{s=1}^{m},$$

where

$$Z(t) \mid \{Z'^{+}(s)\}_{s=1}^{m} \sim \mathcal{N}(\mu + K^{01}(t,\mathbf{s})K^{11}(\mathbf{s},\mathbf{s})^{-1}Z'^{+}(\mathbf{s}), K^{\triangle}(t,t)), \tag{3.6}$$

with $K^{\triangle}(t,t) \triangleq K(t,t) - K^{01}(t,\mathbf{s})K^{11}(\mathbf{s},\mathbf{s})^{-1}K^{10}(\mathbf{s},t)$; this conditional distribution is the usual GP conditional distribution given derivative values, but $Z'^{+}(\cdot)$ is not a GP derivative process. Also, this is only a stochastic process conditioned on $\{Z'^{+}(s)\}_{s=1}^{m}$.

Posterior Distribution

Assuming observations from the model (3.3) and use of the conditional Gaussian process prior in (3.6), the posterior distribution for $Z(\cdot)$ at new predictive inputs is given in the following lemma.

Lemma 3.2. *Suppose the goal is to predict $Z(\cdot)$ at a new set of n^* inputs $\mathbf{t}^*$, given the current inputs $\mathbf{t}$ and resulting observations $\mathbf{X}$. The joint posterior distribution of $(\mathbf{Z}(\mathbf{t}^*), \mathbf{Z}'^{+}(\mathbf{s}))$ is then*

$$\pi(\mathbf{Z}(\mathbf{t}^*), \mathbf{Z}'^{+}(\mathbf{s}) \mid \mathbf{X}, \theta) = \pi(\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'^{+}(\mathbf{s}), \mathbf{X}, \theta)\pi(\mathbf{Z}'^{+}(\mathbf{s}) \mid \mathbf{X}, \theta)\,, \tag{3.7}$$

where

$$[\mathbf{Z}(\mathbf{t}^*) \mid \mathbf{Z}'^{+}(\mathbf{s}), \mathbf{X}, \theta] \;\sim\; \mathcal{N}\left(\mu_2 + \Lambda_{21}\Lambda_{11}^{-1}(\mathbf{X} - \mu_1), \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\right),$$

and the posterior distribution of $\mathbf{Z}'^{+}(\mathbf{s})$ given $\mathbf{X}$ and θ can be found by projecting from $\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta)$, i.e.

$$Z'^{+}(\mathbf{s}) = Z'(\mathbf{s}) \vee \mathbf{0}_m,$$

where

$$\begin{aligned}\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta) \;&\propto\; \exp\{-\frac{\mathbf{Z}'(\mathbf{s})^T\mathbf{K}^{11}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'(\mathbf{s})}{2}\} \\ &\cdot\; \exp\{-\frac{(\mathbf{X} - \mu\mathbf{1}_n - A^*\mathbf{Z}'^{+}(\mathbf{s}))^T B^{*-1}(\mathbf{X} - \mu\mathbf{1}_n - A^*\mathbf{Z}'^{+}(\mathbf{s}))}{2}\},\end{aligned}$$

with

$$
\begin{aligned}
A^* &= \mathbf{K}^{01}(\mathbf{t},\mathbf{s})\mathbf{K}^{11}(\mathbf{s},\mathbf{s})^{-1}, \\
B^* &= \sigma^2\mathbf{I} + \mathbf{K}^{\triangle}(\mathbf{t},\mathbf{t}) = \sigma^2\mathbf{I} + \mathbf{K}(\mathbf{t},\mathbf{t}) - \mathbf{K}^{01}(\mathbf{t},\mathbf{s})\mathbf{K}^{11}(\mathbf{s},\mathbf{s})^{-1}\mathbf{K}^{10}(\mathbf{s},\mathbf{t}), \\
\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} &= \begin{bmatrix} \mu\mathbf{1}_n + \mathbf{K^{01}}(\mathbf{t},\mathbf{s})\mathbf{K^{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+}(\mathbf{s}) \\ \mu\mathbf{1}_{n^*} + \mathbf{K^{01}}(\mathbf{t}^*,\mathbf{s})\mathbf{K^{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+}(\mathbf{s}) \end{bmatrix}, \\
\begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} &= \begin{bmatrix} \mathbf{K}^{\triangle}(\mathbf{t},\mathbf{t}) + \sigma^2\mathbf{I} & \mathbf{K}^{\triangle}(\mathbf{t},\mathbf{t}^*) \\ \mathbf{K}^{\triangle}(\mathbf{t}^*,\mathbf{t}) & \mathbf{K}^{\triangle}(\mathbf{t}^*,\mathbf{t}^*) \end{bmatrix},
\end{aligned}
$$

and $\theta = (\mu, \sigma^2, \beta, \sigma_z^2)$ *being the given parameters of the model and prior.*

Proof. The only significant difference from Lemma 3.1 is the derivation of the posterior distribution $\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}, \theta)$. Again to simplify notation, we omit conditioning on θ in the proof.

Given the data $\mathbf{X}$, the posterior distribution of $\mathbf{Z}'(\mathbf{s})$ is proportional to

$$\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}) \propto f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}))\pi(\mathbf{Z}'(\mathbf{s})).$$

From the construction of the conditional Gaussian process (3.6), the marginal conditional distribution of the observed data $\mathbf{X}$ is

$$[\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{t})] \sim \mathcal{N}(\mu\mathbf{1}_n + \mathbf{K}^{01}(\mathbf{t},\mathbf{s})\mathbf{K}^{11}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+}(\mathbf{s}), \mathbf{K}^{\triangle}(\mathbf{t},\mathbf{t}) + \sigma^2\mathbf{I})\,.$$

Thus

$$
\begin{aligned}
\pi(\mathbf{Z}'(\mathbf{s}) \mid \mathbf{X}) &\propto f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}))\pi(\mathbf{Z}'(\mathbf{s})) \\
&\propto \exp\{-\frac{\mathbf{Z}'(\mathbf{s})^T\mathbf{K}^{11}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'(\mathbf{s})}{2}\} \\
&\cdot \exp\{-\frac{(\mathbf{X} - \mu\mathbf{1}_n - A^*\mathbf{Z}'^{+}(\mathbf{s}))^T B^{*-1}(\mathbf{X} - \mu\mathbf{1}_n - A^*\mathbf{Z}'^{+}(\mathbf{s}))}{2}\}\,.
\end{aligned}
$$

The posterior distribution of $Z'^{+}(\mathbf{s}) = Z'(\mathbf{s}) \vee \mathbf{0}_m$ is then clearly obtained by projection, completing the proof. □

Sampling From the Posterior

The difficulty in sampling from this posterior lies in finding a suitable MCMC scheme for sampling $\mathbf{Z}'^{+}(\mathbf{s})$ from $\pi(\mathbf{Z}'^{+}(\mathbf{s}) \mid \mathbf{X}, \theta)$. According to Lemma 3.2, this can be done by sampling $\mathbf{Z}'(\mathbf{s})$ from its posterior distribution and projecting onto the positive quadrant. A Gibbs sampling procedure is considered, wherein we draw $Z'(s_i)$ one at a time from its full conditional distribution $\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X}, \theta)$ for $i = 1, \cdots, m$; this distribution is derived in the following lemma.

Lemma 3.3. *For any $i = 1, \cdots, m$, the full conditional distribution of $Z'(s_i)$ is*

$$\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X}, \theta) = p_{s_i}\mathcal{N}_{-}(Z'(s_i) \mid \kappa_i, \nu_i) + (1 - p_{s_i})\mathcal{N}_{+}(Z'(s_i) \mid \rho_i, \delta_i)\,, \quad (3.8)$$

where

$$p_{s_i} = \frac{C_1}{C_1 + C_2} = \frac{\Phi(\frac{\kappa_i}{\sqrt{\nu_i}})}{\Phi(\frac{\kappa}{\sqrt{\nu_i}}) + \frac{\sqrt{\delta_i}}{\sqrt{\nu_i}}\exp(-\frac{\kappa_i^2}{2\nu_i} + \frac{\rho_i^2}{2\delta_i})(1 - \Phi(\frac{\rho_i}{\sqrt{\delta_i}}))}$$

with $\Phi(\cdot)$ is a normal cdf,

$$\begin{aligned} C_1 &= \int_{-\infty}^{0} f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), 0, \theta)\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \theta)dZ'(s_i), \\ C_2 &= \int_{0}^{\infty} f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i), \theta)\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \theta)dZ'(s_i), \end{aligned}$$

and

$$\begin{aligned} \kappa_i &= \mathbf{K}^{11}(s_i, \mathbf{s}_{-i})\mathbf{K}^{11}(\mathbf{s}_{-i}, \mathbf{s}_{-i})^{-1}\mathbf{Z}'(\mathbf{s}_{-i}), \\ \nu_i &= K^{11}(s_i, s_i) - \mathbf{K}^{11}(s_i, \mathbf{s}_{-i})\mathbf{K}^{11}(\mathbf{s}_{-i}, \mathbf{s}_{-i})^{-1}\mathbf{K}^{11}(\mathbf{s}_{-i}, s_i), \\ \rho_i &= \delta_i(A_i^{*T}B^{*-1}(\mathbf{X} - \sum_{j \neq i}^{m} A_j^{*}Z'^{+}(s_j)) + \nu_i^{-1}\kappa_i), \\ \delta_i &= (\nu_i^{-1} + A_i^{*T}B^{*-1}A_i^{*})^{-1}, \end{aligned}$$

with $A_i^{*} = (a_{1i}, \cdots, a_{ni})^T$, for $i = 1, \cdots, m$, being the m column vectors of the $n \times m$ matrix $A^{*} = [A_1^{*}, \cdots, A_m^{*}]$.

Proof. Again, we omit the conditioning on θ in the proof. Clearly

$$\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X}) \propto \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s})),$$

so that

$$\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X}) = \frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s})) \left(\mathbf{1}_{\{Z'(s)<0\}} + \mathbf{1}_{\{Z'(s)\geq 0\}}\right)}{C_1 + C_2}.$$

Thus

$$\begin{aligned}
\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X}) &= \frac{C_1}{C_1 + C_2} \frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), 0) \mathbf{1}_{\{Z'(s)<0\}}}{\int_{-\infty}^{0} f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), 0) \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) dZ'(s_i)} \\
&+ \frac{C_2}{C_1 + C_2} \frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i)) \mathbf{1}_{\{Z'(s)\geq 0\}}}{\int_{0}^{\infty} \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i)) dZ'(s_i)} \\
&= p_{s_i} \frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) \mathbf{1}_{\{Z'(s)<0\}}}{\int_{-\infty}^{0} \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) dZ'(s_i)} \\
&+ (1 - p_{s_i}) \frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i)) \mathbf{1}_{\{Z'(s)\geq 0\}}}{\int_{0}^{\infty} \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i)) dZ'(s_i)}.
\end{aligned}$$

Since $\mathbf{Z}'(\mathbf{s}) \sim \mathcal{N}(0, \mathbf{K}^{11}(\mathbf{s}, \mathbf{s}))$, the conditional distribution of $Z'(s_i)$ provided that $\mathbf{Z}'(\mathbf{s}_{-i})$ is $[Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})] \sim \mathcal{N}(\kappa_i, \nu_i)$. Thus,

$$\frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) \mathbf{1}_{\{Z'(s)<0\}}}{\int_{-\infty}^{0} \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i})) dZ'(s_i)} = \mathcal{N}_{-}(Z'(s_i) \mid \kappa_i, \nu_i),$$

where $\mathcal{N}_{-}$ denotes the normal distribution truncated from the right by zero.

Next, define

$$\pi^{UC}(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X}) \propto \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-\mathbf{i}})) f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i)),$$

where

$$f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i)) \propto \exp\{-\frac{(\mathbf{X} - \mu \mathbf{1}_n - A^{*}\tilde{\mathbf{Z}}'(\mathbf{s}))^T B^{*-1} (\mathbf{X} - \mu \mathbf{1}_n - A^{*}\tilde{\mathbf{Z}}'(\mathbf{s}))}{2}\}$$

with $\tilde{\mathbf{Z}}'(\mathbf{s}) = (Z'^{+}(s_1), \cdots, Z'^{+}(s_{i-1}), Z'(s_i), Z'^{+}(s_{i+1}), \cdots, Z'^{+}(s_m))^T$. Then, the distribution of $\pi^{UC}(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X})$ is

$$Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}), \mathbf{X} \sim \mathcal{N}(\rho_i, \delta_i),$$

and this means

$$\frac{\pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}))f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i))\mathbf{1}_{\{Z'(s)\geq 0\}}}{\int_0^\infty \pi(Z'(s_i) \mid \mathbf{Z}'(\mathbf{s}_{-i}))f(\mathbf{X} \mid \mathbf{Z}'^{+}(\mathbf{s}_{-i}), Z'(s_i))dZ'(s_i)} = \mathcal{N}_{+}(Z'(s_i) \mid \rho_i, \delta_i)\,.$$

Moreover, from the deviation we know that

$$p_{s_i} = \frac{C_1}{C_1 + C_2},$$

then do some calculus, p_{s_i} is equal to

$$p_{s_i} = \frac{\Phi(\frac{\kappa_i}{\sqrt{\nu_i}})}{\Phi(\frac{\kappa}{\sqrt{\nu_i}}) + \frac{\sqrt{\delta_i}}{\sqrt{\nu_i}}\exp(-\frac{\kappa_i^2}{2\nu_i} + \frac{\rho_i^2}{2\delta_i})(1 - \Phi(\frac{\rho_i}{\sqrt{\delta_i}}))}.$$

This completes the proof. □

The Gibbs sampling procedure is thus as follows for any $j \in \{1, \cdots, M\}$, where M is the iteration numbers.

- Step 1: Generate $r \sim \text{Bernoulli}(p_{s_i}^{(j)})$. Then, $\forall i \in \{1, \cdots, m\}$, $Z'^{(j+1)}(s_i)$ is drawn from

$$Z'^{(j+1)}(s_i) \sim \begin{cases} \mathcal{N}_{-}(Z'(s_i) \mid \kappa_i^{(j)}, \nu_i), & \text{if} \quad r = 1; \\ \mathcal{N}_{+}(Z'(s_i) \mid \rho_i^{(j)}, \delta_i), & \text{if} \quad r = 0. \end{cases}$$

  where

$$\kappa_i^{(j)} = \mathbf{K}^{11}(s_i, \mathbf{s}_{-i})\mathbf{K}^{11}(\mathbf{s}_{-i}, \mathbf{s}_{-i})^{-1}\mathbf{Z}'^{(j)}(\mathbf{s}_{-i})$$

  with $\mathbf{Z}'^{(j)}(\mathbf{s}_{-i}) = (Z'(s_1)^{(j+1)}, \cdots, Z'(s_{i-1})^{(j+1)}, Z'(s_{i+1})^{(j)}, \cdots, Z'(s_m)^{(j)})$ and the analogous expressions for $\rho_i^{(j)}$ and $p_{s_i}^{(j)}$. Set $Z'^{+(j+1)}(s_i) = Z'^{(j+1)}(s_i) \vee 0$.

- Step 2: Block update $\mathbf{Z}^{(j+1)}(\mathbf{t}^*)$ according to

$$\mathbf{Z}^{(j+1)}(\mathbf{t}^*) \mid \mathbf{Z}'^{+(j+1)}(\mathbf{s}), \mathbf{X}, \theta \sim \mathcal{N}\left(\mu_2^{(j+1)} + \Lambda_{21}\Lambda_{11}^{-1}(\mathbf{X} - \mu_1^{(j+1)}), \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}\right)$$

where

$$\begin{aligned}
\mu_1^{(j+1)} &= \mu\mathbf{1}_n + \mathbf{K^{10}}(\mathbf{s},\mathbf{t})\mathbf{K^{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+(j+1)}(\mathbf{s}), \\
\mu_2^{(j+1)} &= \mu\mathbf{1}_{n^*} + \mathbf{K^{10}}(\mathbf{s},\mathbf{t}^*)\mathbf{K^{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+(j+1)}(\mathbf{s}).
\end{aligned}$$

Extending to Higher Order Derivative Processes

To define a conditional GP based on convexity constraints, first define

$$Z''^{+}(t) = Z''(t) \vee 0,$$

where $Z''(t) \sim \mathcal{GP}\left(0, K^{22}(t,t')\right)$. Then conditioned on $Z''^{+}(\cdot)$ at m virtual points, $\mathbf{s}$, the conditional Gaussian process is defined exactly as in the beginning of Section 3.3.2, with the changes of replacing $Z'^{+}(\mathbf{s})$ by $Z''^{+}(\mathbf{s})$, $\mathbf{K^{11}}(\mathbf{s},\mathbf{s})$ by $\mathbf{K^{22}}(\mathbf{s},\mathbf{s})$, $\mathbf{K^{10}}(\mathbf{s},\mathbf{t})$ by $\mathbf{K^{20}}(\mathbf{s},\mathbf{t})$ and $\mathbf{K^{01}}(\mathbf{t},\mathbf{s})$ by $\mathbf{K^{02}}(\mathbf{t},\mathbf{s})$. The posterior distribution is then given as in Lemma 3.2, with the corresponding changes mentioned above, and sampling from this posterior is identical to the Gibbs sampling method above of conditional GP for monotone functions with the indicated changes.

3.3.3 Estimating the Parameters of the Gaussian Process Models

We have assumed that the parameters μ and σ^2 from the model and σ_z^2 and β^2 from the Gaussian process prior are given, but in most situations, they are unknown. A fully Bayesian approach to dealing with them is possible, inserting Gibbs and/or Gibbs-Metropolis updates into the MCMC's discussed earlier. There is considerable confounding between σ_z^2 and β^2, however (and between σ_z^2 and σ^2 if no replicate observations are present), so that the fully Bayesian approach can be problematical.

Also, this would presume that the inputs $\mathbf{s}$ have been pre-determined, whereas we will see that their choice is actually a design question.

We thus turn to marginal maximum likelihood to estimate the parameters. Even that, however, is problematical, in that the constraints on the derivatives of the GP result in marginal likelihoods whose evaluation requires high-dimensional integration. We then take a partial likelihood approach, ignoring the derivative information and basing the maximum likelihood estimates only on the marginal likelihood resulting from consideration of the marginal prior on $\mathbf{Z}(\mathbf{t})$, which yields

$$\mathbf{X} \mid \theta \sim \mathcal{N}(\mu \mathbf{1}_n, \mathbf{K}(\mathbf{t}, \mathbf{t}) + \sigma^2 \mathbf{I}_n). \tag{3.9}$$

Another perspective on this is to imagine that, ordinarily, the analysis would just proceed with the GP on $\mathbf{Z}(\mathbf{t})$; we are considering what gain can be achieved by adding in the derivative information, and can view this as adding in the derivative information after the parameter estimation has been done.

Maximizing (3.9) over the four parameters is straightforward; the maximum likelihood estimate (MLE) for μ given the other parameters can be given in closed form, as can the MLE for σ_z^2 in the situation (common in emulation of computer models) where $\sigma^2 = 0$. Denote the MLE of θ as $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2, \hat{\beta}, \hat{\sigma}_z^2)$.

3.3.4 Determining Locations of the Virtual Derivative Points

Because the computational complexity increases with the number of derivative constraints included, it is of interest to develop a method for selecting locations for the constraints that are likely to be most effective. For clarity, we only discuss the situation in which the function is presumed to be non-decreasing, but the ideas extend directly to other constraints.

Intuitively, constraining the derivative to be non-negative will have the most influence at a given location if the probability that the process has negative derivative

there is comparatively large. Hence, in an iterative fashion, we will seek locations with the largest probabilities of having a negative derivative.

To begin, we define $\mathcal{S}_{\mathbf{t}^*}$ to be the collection of prediction locations or other locations (such as observation points) that are going to be considered as candidates to be the virtual points at which to impose on derivative conditions. The procedure is then defined as follows.

- *Step 1*: Initially and for $\forall\ t^* \in \mathcal{S}_{\mathbf{t}^*}$, compute the probability $\Pr(Z'(t^*) \leq 0 \mid \mathbf{X}, \hat{\theta})$; let t_1^* be a point where this probability is maximized, and include it in the selection set $\mathcal{T}_{\mathcal{S}}$ if the probability exceeds a threshold $p*$.

- *Step 2*: Suppose $\mathcal{T}_{\mathcal{S}} = \{t_1^*, \ldots, t_k^*\}$. $\forall\ t^* \in \mathcal{S}_{\mathbf{t}^*} \setminus \mathcal{T}_{\mathcal{S}}$, compute $\Pr(Z'(t^*) \leq 0 \mid \mathbf{X}, \hat{\theta}, \mathcal{T}_{\mathcal{S}})$, using the marginal distribution of $Z'(t^*)$ based on the current selection set, which is given by

$$\pi(Z'(t^*) \mid \mathbf{X}, \hat{\theta}, \mathcal{T}_{\mathcal{S}}) = \int \pi(Z'(t^*) \mid \mathbf{X}, Z'^*(\mathcal{T}_{\mathcal{S}}), \hat{\theta}) \pi(Z'^*(\mathcal{T}_{\mathcal{S}}) \mid \mathbf{X}, \hat{\theta}) dZ'^*(\mathcal{T}_{\mathcal{S}}). \tag{3.10}$$

  Let t_{k+1}^* be a point where this probability is maximized, and include it in the selection set if the probability exceeds the threshold $p*$.

- *Step 3*: Repeat *Step 2* until all probabilities for the remaining candidate points are less than $p*$, or a pre-determined selection set size K_M has been reached.

Clearly,

$$\begin{aligned} \Pr(Z'(t^*) \leq 0 \mid \mathbf{X}, \hat{\theta}, \mathcal{T}_{\mathcal{S}}) &= \int \Pr(Z'(t^*) \leq 0 \mid \mathbf{X}, Z'^*(\mathcal{T}_{\mathcal{S}}), \theta) \pi(Z'^*(\mathcal{T}_{\mathcal{S}}) \mid \mathbf{X}, \hat{\theta}) dZ'^*(\mathcal{T}_{\mathcal{S}}) \\ &\approx \frac{1}{M} \sum_{l=1}^{M} \Pr(Z'(t^*) \leq 0 \mid \mathbf{X}, Z'^{*(l)}(\mathcal{T}_{\mathcal{S}}), \hat{\theta}), \end{aligned} \tag{3.11}$$

where $\{Z'^{*(1)}(\mathcal{T}_\mathcal{S}), \ldots, Z'^{*(M)}(\mathcal{T}_\mathcal{S})\}$ is a sample from $\pi(Z'^{*}(\mathcal{T}_\mathcal{S}) \mid \mathbf{X}, \hat{\theta})$, which can be obtained by the MCMC schemes from Section 3.3.1 or Section 3.3.2, depending on the derivative constraint method being used. Note that the same sample can be used to compute (3.11) for all remaining candidate t^*.

Finally, normal computations yield

$$\Pr(Z'(t^*) \leq 0 \mid \mathbf{X}, Z'^{*}(\mathcal{T}_\mathcal{S}), \hat{\theta}) = \Phi\left(\frac{\lambda_{t^*}}{\sqrt{\omega_{t^*t^*}}}\right),$$

where λ_{t^*} and $\omega_{t^*t^*}$ are defined through the following expressions for the indicator function method (Here, $\mathbf{t}^*$ indicates the vector of candidates in $\mathcal{S}_{\mathbf{t}^*} \backslash \mathcal{T}_\mathcal{S}$ and $\hat{\mathbf{K}}$ is the MLE estimate of $\mathbf{K}$.)

$$\begin{aligned}
A'_1 &= \hat{\mathbf{K}}^{\mathbf{01}}(\mathbf{t}, \mathbf{t}^*)\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*, \mathbf{t}^*)^{-1}, \\
A'_2 &= \hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*, \mathbf{s})\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s}, \mathbf{s})^{-1}, \\
B'_1 &= \hat{\sigma}^{\mathbf{2}}\mathbf{I} + \hat{\mathbf{K}}(\mathbf{t}, \mathbf{t}) - \hat{\mathbf{K}}^{\mathbf{01}}(\mathbf{t}, \mathbf{t}^*)\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*, \mathbf{t}^*)^{-1}\hat{\mathbf{K}}^{\mathbf{10}}(\mathbf{t}^*, \mathbf{t}), \\
B'_2 &= \hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*, \mathbf{t}^*) - \hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*, \mathbf{s})\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s}, \mathbf{s})^{-1}\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s}, \mathbf{t}^*), \\
A' &= A_1'^T B_1'^{-1} A'_1 + B_2'^{-1}, \\
\lambda &= \hat{\mu}\mathbf{1}_{n^*} + A'^{-1}\left(A_1'^T B_1'^{-1}(\mathbf{X} - \hat{\mu}\mathbf{1}_n) + B_2'^{-1}A'_2 \frac{1}{D-d}\sum_{\ell=d+1}^{D} \mathbf{Z}'(\mathcal{T}_\mathcal{S})^{(\ell)}\right) \\
&= (\lambda_1, \cdots, \lambda_{\mathcal{S}_{\mathbf{t}^*}\backslash\mathcal{T}_\mathcal{S}})^T, \\
\Omega &= \frac{A'^{-1}}{D-d} \\
&= (\omega_{ij}), \quad \omega_{ij} \text{ indicates the entry of } \Omega.
\end{aligned}$$

and the analogous expressions for the conditional GP method:

$$
\begin{aligned}
\mu_1' &= \hat{\mu}\mathbf{1}_n + \hat{\mathbf{K}}^{\mathbf{01}}(\mathbf{t},\mathbf{s})\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+}(\mathbf{s}),\\
\mu_2' &= \hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*,\mathbf{s})\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{Z}'^{+}(\mathbf{s}),\\
\Lambda_{11}' &= \hat{\mathbf{K}}^{\triangle}(\mathbf{t},\mathbf{t}) + \hat{\sigma}^2\mathbf{I},\\
\Lambda_{12}' &= \hat{\mathbf{K}}^{\mathbf{01}}(\mathbf{t},\mathbf{t}^*) - \hat{\mathbf{K}}^{\mathbf{01}}(\mathbf{t},\mathbf{s})\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s},\mathbf{s})^{-1}\mathbf{K}^{\mathbf{11}}(\mathbf{s},\mathbf{t}^*) = \Lambda_{21}'^{T},\\
\Lambda_{22}' &= \hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*,\mathbf{t}^*) - \hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{t}^*,\mathbf{s})\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s},\mathbf{s})^{-1}\hat{\mathbf{K}}^{\mathbf{11}}(\mathbf{s},\mathbf{t}^*),\\
\lambda &= \mu_2' + \Lambda_{21}'\Lambda_{11}'^{-1}\left(\mathbf{X} - \mu_1'\right)\\
&= \left(\lambda_1, \cdots, \lambda_{\mathcal{S}_{\mathbf{t}^*}\setminus\mathcal{T}_{\mathcal{S}}}\right)^T,\\
\Omega &= \frac{\Lambda_{22}' - \Lambda_{21}'\Lambda_{11}'^{-1}\Lambda_{12}'}{D-d}\\
&= (\omega_{ij}), \quad \omega_{ij} \text{ indicates the entry of } \Omega.
\end{aligned}
$$

3.4 Illustrations

3.4.1 Simulated Examples

In this section, some simulated examples are considered in order to investigate the behavior of the proposed methods and indicate the gain from incorporating shape constraints in Gaussian processes.

The model

$$x_i = Z(t_i) + \epsilon_i, \quad \text{for } \ i = 1, \cdots, n$$

is used to simulate the data, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$. The $n = 50$ input points t_i are chosen equally spaced in the interval $[-10, 10]$. Inferences will be performed at $n^* = 100$ predictive points equally spaced over the same interval. The following functions $Z(t)$ are considered; the first three are non-decreasing functions while the last three are convex functions:

Table 3.1: Summary of parameter estimates of GP models

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\beta}$	$\hat{\sigma}_z$
$Z(t) = 4/(1 + \exp(-t/2 + 4))$	0.9593	0.4713	3.6454	1.0655
$Z(t) = \sin(t) + t$	-0.2330	0.5321	2.9964	5.4201
Monotone Stepwise function	0.1113	0.4268	2.5046	2.0887
$Z(t) = 5(t/25)^2 - 2$	-1.6501	0.4816	1.4235	0.3690
$Z(t) = (1/1050)\cosh(t) - 0.55$	2.0273	0.4593	2.0382	3.3787
Convex Stepwise function	2.2765	0.4857	2.6914	2.0648

1. $Z(t) = 4/(1 + \exp(-t/2 + 4))$;

2. $Z(t) = \sin(t) + t$;

3. $Z(t) = \begin{cases} t+5, & \text{if} \quad t < -5, \\ 0, & \text{if} \quad -5 \leq t \leq 5, \\ t-5, & \text{if} \quad t > 5; \end{cases}$

4. $Z(t) = 5(t/25)^2 - 2$;

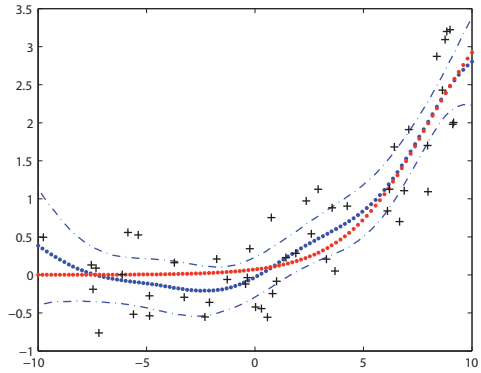
5. $Z(t) = (1/1050)\cosh(t) - 0.55$;

6. $Z(t) = \begin{cases} -t-5, & \text{if} \quad t < -5, \\ 0, & \text{if} \quad -5 \leq t \leq 5, \\ t-5, & \text{if} \quad t > 5. \end{cases}$

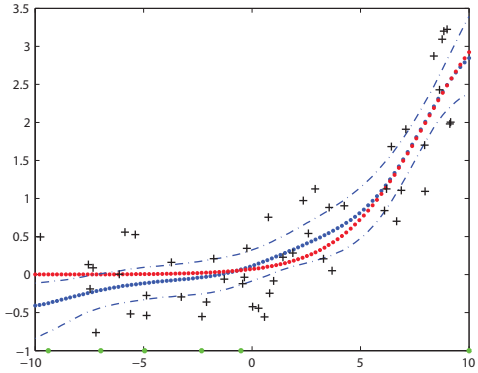
The three monotone functions represent three kinds of monotonic functions. The first one is strictly increasing function, slowly increasing at the beginning and sharply increasing at the end; the second is increasing but has a nearly flat region; the third has a flat region where the derivative is exactly zero. Similarly, the first of the last three convex functions is strictly convex, while the second has a nearly flat region and the third has a flat region with exactly zero second derivative. The partial likelihood estimates of the model and prior parameters for the various functions are given in Table 3.1 and were used in the subsequent analysis.

We compare the two proposed methods from Section 3.3 for imposing shape constraints. Each MCMC was run for 10,000 iterations with a 5,000 burn-in period.

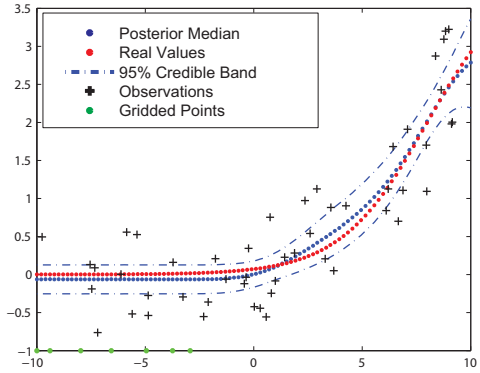
In Figures 3.2 through 3.7, the red dots are the function's real value at the predictive points and the black plus points are the observations. The blue dots in each figure are the value of the posterior median and the the blue dashed lines give the corresponding pointwise 95% credible interval band. The (b) and (c) figures also show the virtual derivative points used, the green dots on the x-axis. These were determined by the method of Section 3.3.4, using the predictive points as the candidate set and setting the probability threshold at $p^* \leq 1^{-20}$.



(a) GP without constraints



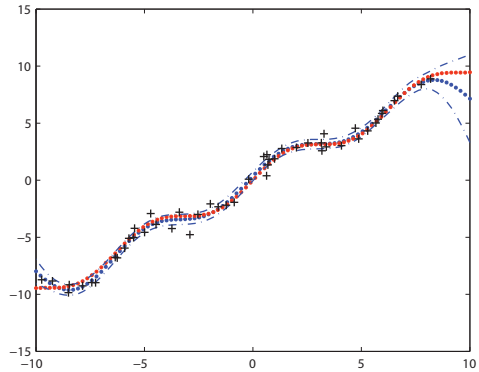
(b) GP with constraints via Indicator Function



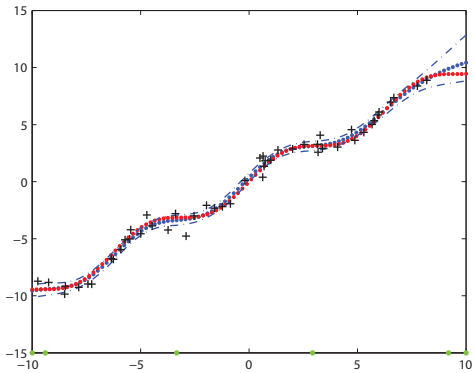
(c) GP with constraints via Conditional GP

FIGURE 3.2: Comparing GP without vs with constraints for $Z(t) = 4/(1 + \exp(-t/2 + 4)$.

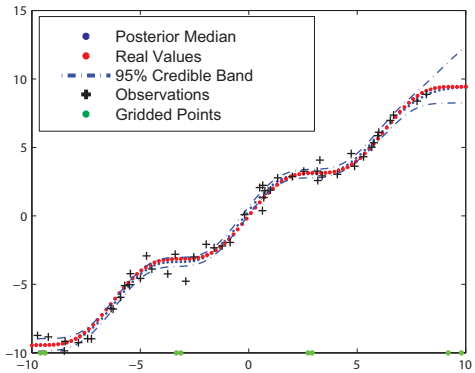
All the figures suggest that placing constraints on the derivative process of the original GP considerably improves estimation of the target function if, in fact, it



(a) GP without constraints



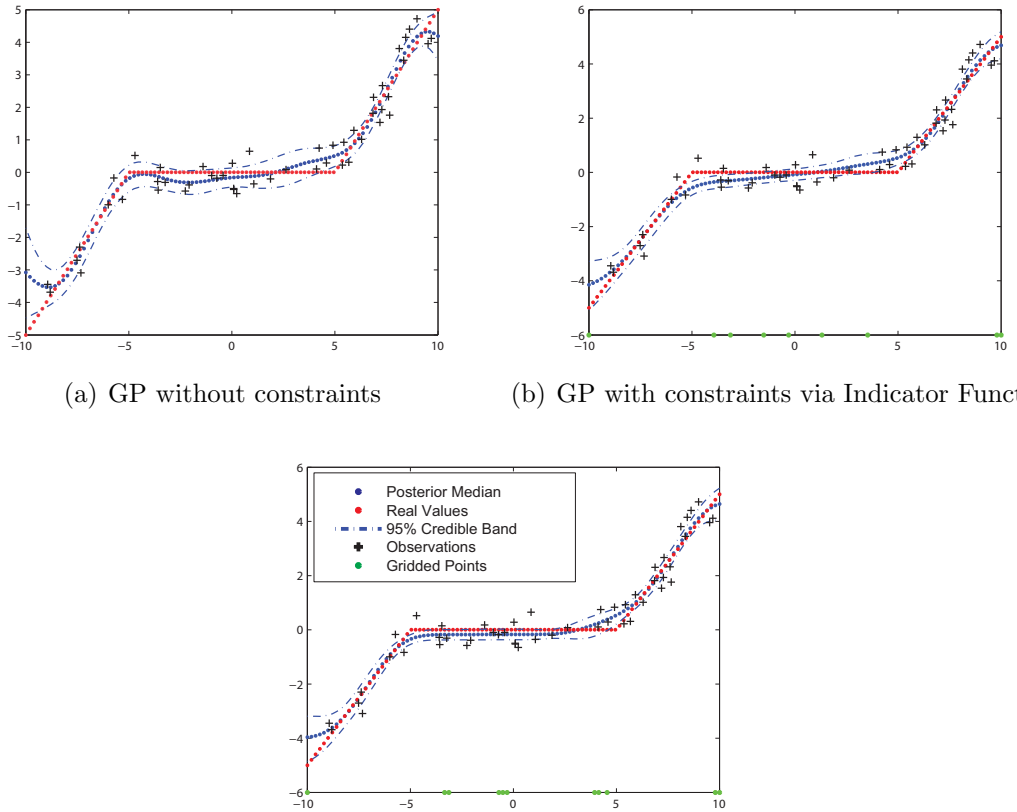
(b) GP with constraints via Indicator Function



(c) GP with constraints via Conditional GP

FIGURE 3.3: Comparing GP without vs with constraints for $Z(t) = \sin(t) + t$.

is monotone or convex. Imposing constraints via the conditional Gaussian process is clearly superior to use of the indicator functions in Figure 3.2, Figure 3.4 and Figure 3.7 situations in which the function has a flat or nearly flat segment. The performance of the conditional Gaussian process method is similar to the use of the indicator function otherwise, and so would seem to be the preferred method. It is more intensive computationally, however, and so use of the indicator function method is sensible if no flat regions in the function are expected. Note that the use of constraints can be particularly useful when data is not available in some region – for instance the right end of Figure 3.3 without the constraints the GP would try to



(a) GP without constraints

(b) GP with constraints via Indicator Function

(c) GP with constraints via Conditional GP

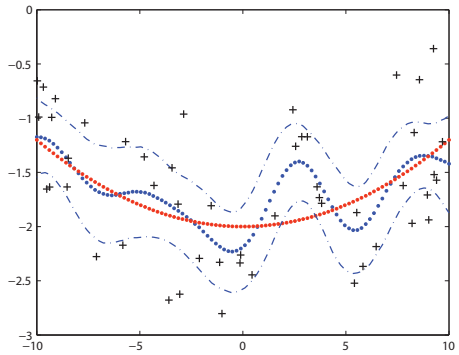
FIGURE 3.4: Comparing GP without vs with constraints for the monotone stepwise function.

revert to its mean in such a situation.

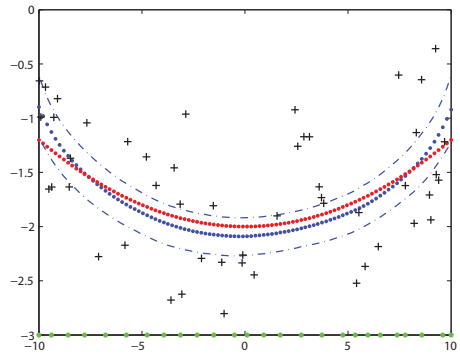
3.4.2 Emulating the CRASH Computer Model

The CRASH computer model was considered in Bayarri et al. (2009). This computer model simulates the effect of a collision of a vehicle, that is analyzing the velocity changes after impacts at key positions on the vehicle. Runs of the computer model itself are very time consuming, so that a fast approximation –called an emulator – is required for the statistical analysis. Bayarri et al. (2009) used a GP to produce the emulator.

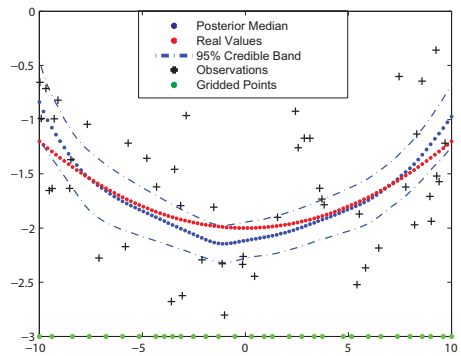
The primary output of interest from the computer model is the relative velocity



(a) GP without constraints



(b) GP with constraints via Indicator Function

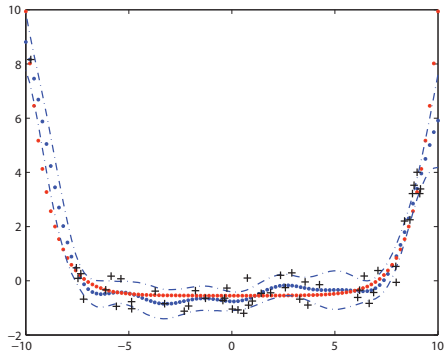


(c) GP with constraints via Conditional GP

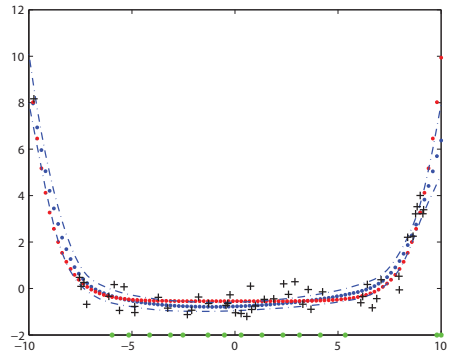
FIGURE 3.5: Comparing GP without vs with constraints for $Z(t) = 5(t/25)^2 - 2$.

over the time period of a crash (the time from initial impact to the vehicle reaching stationarity) of the head of a dummy sitting in the driver's seat. The relative velocity is the difference between the actual velocity and the vehicle impact velocity. Intuitively, this relative velocity should be monotonically decreasing over the crash period, so imposing this shape constraint on the GP emulator is natural.

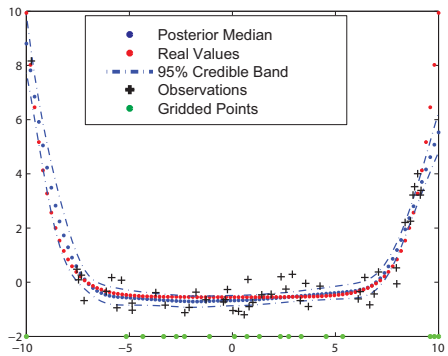
As in Bayarri et al. (2009), we denote the relative velocity output of the CRASH model at input time t by $y(t)$. We assume that $y(t)$ has a prior distribution given by a Gaussian process with mean $m(t) = \mu$ and the squared exponential covariance function. We illustrate the methodology here by considering one of the computer model runs, the resulting data being given in Table 3.2.



(a) GP without constraints



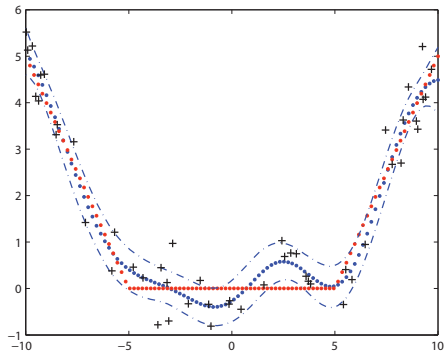
(b) GP with constraints via Indicator Function



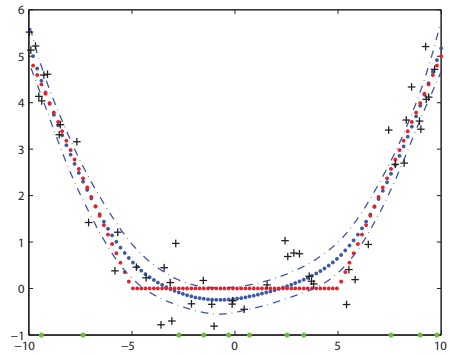
(c) GP with constraints via Conditional GP

FIGURE 3.6: Comparing GP without vs with constraints for $Z(t) = (1/1050)\cosh(t) - 0.55$.

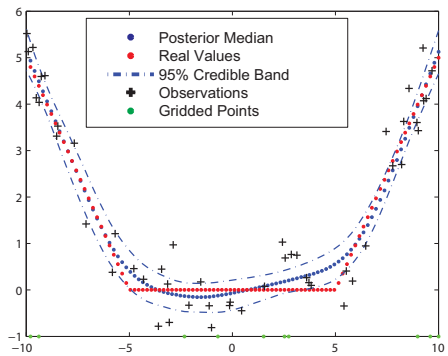
This data is dense enough that the unconstrained Gaussian process leads to a monotone posterior mean. To illustrate the potential of our methodology, therefore, we consider three thinned data sets: in Figure 3.8, it is presumed that only data points 1, 7, 10, 12, and 14 in Table 3.2 are available; in Figure 3.9, that 1, 3, 6, 12, and 14 are available; and, in Figure 3.10, that 4, 5, 9, 10, and 12 are available. The prediction set consists of 50 points equally spaced over the time domain from 0 to 81 milliseconds, the same domain considered in Bayarri et al. (2009), as well as the real data points that were removed in the thinning. Table 3.3 gives the partial likelihood estimates of the GP parameters arising in each situation.



(a) GP without constraints



(b) GP with constraints via Indicator Function



(c) GP with constraints via Conditional GP

FIGURE 3.7: Comparing GP without vs with constraints for the convex stepwise function.

The blue dots in Figure 3.8, Figure 3.9 and Figure 3.10 are the posterior median of the prediction points and the blue dash lines in each figure are the 95% credible bands of the posterior distribution of the prediction sets. The red dots in each figure correspond to the data in Table 3.2 and the black plus points are the selected data used in the analysis. The green dots on the x-axis in Figure 3.8 (b), Figure 3.9 (b), and Figure 3.10 (b) are the locations at which the derivative constraints were placed. It is clear, in each situation, that addition of the monotonicity information greatly improves the function estimate.

Table 3.2: CRASH model dataset

	time (milliseconds)	relative velocity (km/h)
1	1	0.00
2	3	-0.10
3	5	-0.40
4	7	-0.80
5	9	-1.40
6	11	-2.40
7	13	-3.20
8	15	-4.10
9	17	-5.20
10	20	-6.20
11	25	-8.40
12	30	-10.30
13	35	-12.90
14	40	-15.20
15	45	-17.80
16	50	-20.00
17	55	-21.90
18	60	-24.34
19	65	-26.54

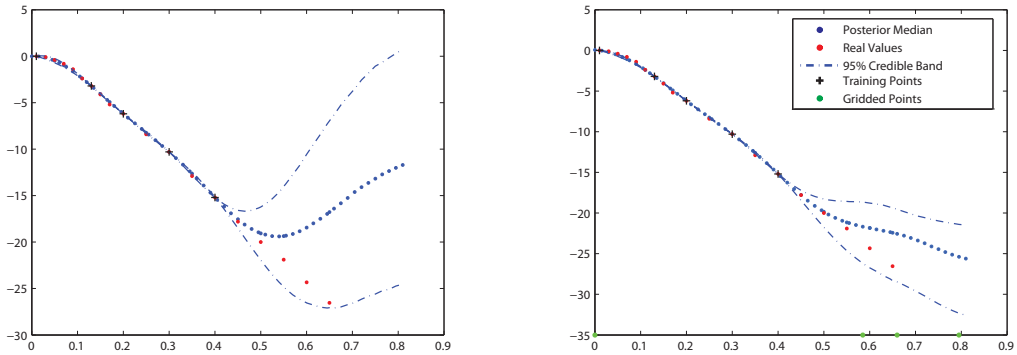
Table 3.3: Estimates of GP parameters

	Figure 3.8	Figure 3.9	Figure 3.10
$\hat{\beta}$	0.1713	0.2094	0.1206
$\hat{\sigma}_z$	6.5004	2.7318	3.0088
$\hat{\mu}$	-9.8728	-10.0011	-10.0002

3.5 Summary and Discussion

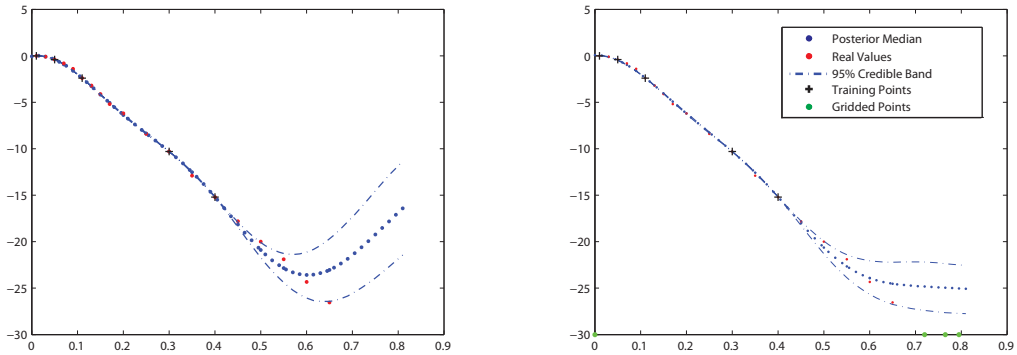
We have put forward two alternative methods to apply shape constraints to a non-parametric Gaussian process model, focusing on monotonicity constraints and convexity/concavity constraints. For each method, we develop a Monte Carlo Markov Chain sampling scheme to carry out the posterior analysis. In addition, a method is proposed to determine good locations at which to specify the derivative constraints.

In the illustrations considered, use of either method for imposing shape con-



(a) GP without constraints (b) GP with constraints via Indicator Function

FIGURE 3.8: Comparing GP without vs with constraints for the training set $1, 7, 10, 12, 14$.

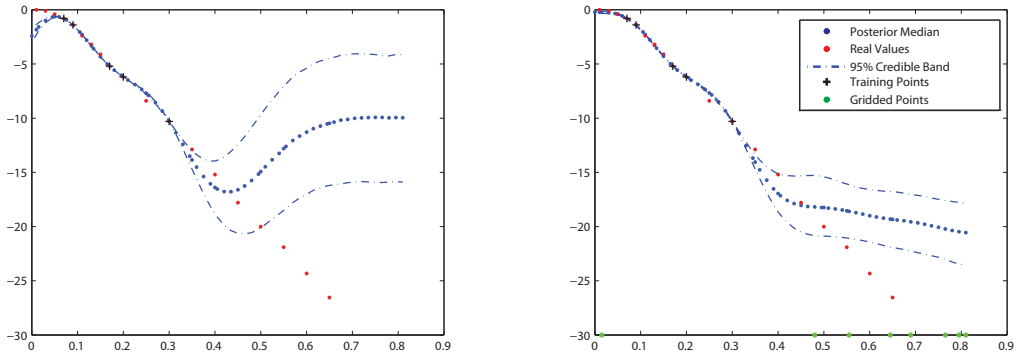


(a) GP without constraints (b) GP with constraints via Indicator Function

FIGURE 3.9: Comparing GP without vs with constraints for the training set $1, 3, 6, 12, 14$.

straints results in a considerable improvement over use of the unconstrained Gaussian process. However, when the target function has flat areas, the procedure for eliciting constraints that utilizes the conditional GP appears to be superior. This is because the conditional GP method allows the derivative of the function to be zero with positive probability.

Although, in this chapter, we focus on Gaussian processes with the squared expo-



(a) GP without constraints (b) GP with constraints via Indicator Function

FIGURE 3.10: Comparing GP without vs with constraints for the training set $4, 5, 9, 10, 12$.

nential covariance function, the two alternative proposed methods are easily extended to other covariance functions which satisfy the conditions of Theorem 2.2.2 in Adler (1981), for instance, the Matérn class of covariance functions. Also, the methods can be directly extended to higher dimensional inputs.

4

Subgroup Analyses Using Tree-based Models

This chapter discusses subgroup analyses, the goal of which is to determine the heterogeneity of the treatment effect across subpopulations. Searching for subgroup effects is difficult because of the issue of multiple testing together with the complication that test statistics for subgroups are typically dependent, making simple multiplicity corrections such as the Bonferroni correction far too conservative.

In this chapter, a Bayesian approach to subgroup analyses is proposed, utilizing tree-based models to assign prior probabilities to subgroup effects that account for multiplicity, and completing the analysis using Bayesian model selection techniques. The main focus of the chapter is on utilization of the tree-based models to incorporate prior information into the model probabilities.

An illustration of the analysis involving the subgroup analysis of biomarker effects on treatments in clinical trials will be considered.

4.1 Literature Review and Motivation

Studies typically just compare the effect of a treatment (or compare treatments) on an entire population. But often it is also of interest to determine if there are differential

treatment effects on subpopulations. For instance, it may be of interest to determine if a drug has a different effect on young patients than on old patients; or it might be of interest to understand whether patients who carry a genetic polymorphism in cytochrome enzyme would respond more actively to a new therapy than those who lack the polymorphism in the enzyme.

Temple and Ellenberg (2000) mentioned that approved treatments for many conditions did not subsequently appear effective in many follow-up clinical trials. They attributed these findings in part to the heterogeneity of effectiveness among subgroups of patients, who often have distinctive genetical biomarker, clinical markers (e.g. medical history, disease severity), demographics (e.g. sex, age) and social or environmental factors (e.g. smoking habits). Thus, diagnosing the subgroups for which there are substantial benefits might salvage some new drugs and improve overall success (Woodcock (2007)). Moreover, even if the treatment does exhibit overall significance, there may be considerable differences of effectiveness in subpopulations (Wang et al. (2007)) that would be important to understand for clinical practice. Indeed, there is an increasing trend (or at least wish) to consider tailored treatments in the pharmaceutical industry. The key to the tailored therapeutics is to identify subgroups that are prone to be active for the drug (Lipkovich et al. (2011)).

While the appeal of subgroup analyses is undeniable, there are a number of concerns with the approach. The main concerns are (e.g. Berry (1990), Pocock et al. (2002), Cui et al. (2002), Lagakos (2006) and Wang et al. (2007)):

1. *Multiplicity*: When multiple subgroup analyses are performed, the issue of multiple testing arises. Standard adjustment for multiplicity such as the Bonferroni correction can be utilized to correct for multiple testing, but it will give up too much power when the test statistics are highly dependent, which is virtually always the situation in subgroup analyses.

2. *Post hoc analyses (unplanned analyses or data dredging)*: Given a plethora of baseline covariates and the tendency not to have prespecified subgroups, there are a large number of possible subgroup analyses that can be performed in a post hoc manner. Proper adjustment for post hoc multiple subgroup testing is extremely difficult due to uncertainty for the number of tests that were actually performed.

3. *Lack of power*: Most studies recruit just enough participants to guarantee that there is sufficient statistical power to detect an overall main effect, if one is present. The power for detecting effects for subgroups is then inherently lower, and this power is further reduced by the needed multiplicity corrections.

In addition to these primarily statistical concerns, there are also issues concerning the interpretation of findings from a subgroup analysis. Pocock et al. (2002) advocated assessing biological plausibility along with consideration of the statistical strength of evidences.

The Bayesian approach to subgroup analyses in this chapter is illustrated by, and partly motivated from, a simulated clinical trial dataset from Eli Lilly. A two-group randomized clinical trial with a continuous outcome variable Y is considered. The treatment group is indicated as T ($T = 0$ gives placebo; $T = 1$ implies receiving treatment) while a $32-$dimensional covariate vector $\mathbf{X}$ of dichotomous (0 or 1) biomarkers is available for each patient in the study. There is interest in determining if these covariates themselves have any predictive value related to the disease (their *prognostic efficacy*), but the primary interest centers on understanding if there are differing treatment effects for subgroups defined by values of these covariates.

Before turning to our approach, two relevant exploratory approaches to identifying subgroup effects are worth mentioning. One common approach is based on fitting a single model which includes both main effects and interactions for all co-

variates simultaneously. For a continuous response variable, a general form of this regression-based model could be written as

$$\mathrm{E}(Y) = \mu_0 + \eta T + \boldsymbol{\beta} h(\mathbf{X}) + \boldsymbol{\theta} T \omega(\mathbf{X}) \tag{4.1}$$

where the main interest would be in the response increment for the treatment-by-covariate interaction, i.e. the term $\omega(\mathbf{X})$ in treatment group. Although an interaction test based on the model above partially overcomes the multiplicity concerns, such tests are likely to be underpowered (Pocock et al. (2002)). Dixon and Simon (1991) and Simon (2002) overcame some of those difficulties by providing a Bayesian approach to analyze an analog of (4.1) but use the first order interactions as the last term. They defined a suitable prior distribution for the parameters of the interaction terms and summarized the point and interval estimates for the subgroup-specific treatment effects as well as posterior probabilities about the effect size of the treatment in each subgroup. Jones et al. (2011) extended Dixon and Simon (1991) and Simon (2002)'s method to be more flexible by including second- and higher-order interaction terms. Hodges et al. (2007) used different variances for different interaction terms and permitted each interaction to shrink to zero. One deficiency of this approach is that only covariates or combination of covariates included in the function $\omega(\mathbf{X})$ are considered as interactions in the model, which might rule out some important situations.

Another popular approach relies on tree-based methods to identify subgroups in which the outcome significantly differs. As mentioned in Chapter 1, the tree-based method is an excellent tool for exploring heterogeneous structure, and can be used to find complex and nonlinear relationships among predictors. Ruberg et al. (2010) also argued that clinicians can better understand and more easily apply the results of such methods in clinical practice. Negassa et al. (2005), Su et al. (2008), and Su et al. (2009), among others, utilized the CART algorithm or random forests to recursively

partition the data into two subgroups that show the greatest heterogeneity in the treatment. Unlike these approaches aiming at partitioning of the entire covariate space, Ruberg et al. (2010), Foster et al. (2011) and Lipkovich et al. (2011) developed a search tool to identify 'interesting' regions in the covariate space (e.g. a region of the covariate space where the treatment effect on the response is substantially better than the average treatment), but ignored the rest of covariate space as 'uninteresting' in their recursive partitioning algorithm to define subgroups.

Enlightened by these two approaches, we develop a Bayesian model selection approach to subgroup analyses which utilizes a tree-based method to define subgroup models and also to assign prior probabilities to the models (the key to Bayesian multiplicity control). At each terminal node of the tree (defining a subgroup), the response is assumed to have a linear regression structure similar to (4.1), replacing $\omega(\mathbf{X})$ in the interaction term by the subgroup mean effects.

Chipman et al. (1998) and Denison et al. (1998) first utilized Bayesian model selection techniques to implement CART. In contrast with the conventional greedy search algorithms for CART, the Bayesian model selection approach induces a posterior distribution over trees that can be used to guide a stochastic search towards 'more promising' treed models, broadening the search of the tree model space and allowing for posterior summaries. Allowing such overall posterior summaries is particularly crucial for the subgroup analysis problem, since particular subgroups can occur in many different trees, so that one must average over trees to determine the effective of a treatment in a subgroup.

The stochastic structure by which the trees are generated defines the prior probabilities of the resulting models, and is the device by which multiplicity is controlled (cf. Scott and Berger (2010)). Sivaganesan et al. (2011) also applied a Bayesian model selection approach to subgroup analyses, but they take a more decision-theoretic approach in which the problem is viewed as a series of decisions concerning

whether or not to delve more deeply into subgroups. In our approach, we simultaneously consider all possible subgroup analyses.

This chapter is organized as below. Section 4.2 introduces the notation for subgroups and the definition of allowable subgroups. Section 4.3 gives the approach for constructing the tree and describes how to model the outcome at the terminal nodes of the trees. Section 4.4 specifies the priors for trees in the model space and the priors for the parameters in the outcome model. Section 4.5 illustrates how to approximate the marginal likelihood in each model. Section 4.6 discusses how to do posterior inferences and summarizes them.

4.2 Notation and Allowable Subgroups

Suppose the data consist of $(Y_m, T_m, X_{1m}, \cdots, X_{pm})'$, for $m = 1, \cdots, n$, where Y_m is the outcome variable, assumed herein to be continuous, observed on an independent sample of subjects, $T_m = 1$ or $T_m = 0$ indicates whether the subject was in the treatment group or the control group, and $(X_{1m}, \cdots, X_{pm})'$, are the covariates for that individual, herein assumed to be dichotomous (0 or 1). A subgroup is then defined as those individuals with a specfied value of one or more covariates.

In practice, there is usually some maximum number of covariates k that will be considered in the subgroup analysis. Clearly $k = 0$ means only the entire cohort is considered (i.e., only an overall treatment effect is being considered); $k = 1$ defines subgroups based on a single covariate in addition to the entire cohort; $k = 2$ additionally allows subgroups based on two covariates, etc. When covariates correspond to demographics, social/enviornmental factors or clinical variables such as gender, age, smoking status and disease severity, $k = 1$ or $k = 2$ will be popular choices. But if covariates correspond to genomic factors such as biomarkers, larger k might be needed. In this chapter, we focus on the situation when $k \leq 3$.

A key principle of our approach is to restrict consideration only to subgroups

that make scientific sense, e.g., obey some biological rationale. To elaborate on this idea and introduce the basic notation for subgroups, we first consider a pedagogical example.

Example 4.1. *Suppose age, gender and smoking status are the three dichotomous covariates under consideration, to be denoted A, B and C, respectively:*

- *Denote discrete subgroup designations as numbered subscripts of the appropriate letter. Young is A_1 and old is A_2. Male is B_1 and female is B_2. Smoking is C_1 and non-smoking is C_2. The notation $A_\bullet$ is used, for simplicity, to denote any variant value of the covariate (rather than the more notationally consistent A_{12}).*

- *Denote an actual subgroup as a concatenation of letters with numbered subscripts. All letters must appear in the subgroup. Hence young males is $A_1B_1C_\bullet$, reflecting the fact that no split has been made according to smoking status. Young male smokers are, of course, $A_1B_1C_1$.*

- *A partition of the population is written as a list of subgroups, with every individual in the population belonging to one and only one subgroup in the list. An example partition is $\{A_1B_\bullet C_\bullet, A_2B_\bullet C_\bullet\}$. In this example, the entire cohort is $\{A_\bullet B_\bullet C_\bullet\}$.*

- *Associated with each subgroup, there is a mean treatment effect which we will write by suppressing the letters, as in $\mu_{1\bullet\bullet}$, representing the mean treatment effect for subgroup $A_1B_\bullet C_\bullet$.*

- *Likewise, in each subgroup, let $\beta_{1\bullet\bullet}$ define the underlying subgroup baseline effect for subgroup $A_1B_\bullet C_\bullet$.*

A key observation is that the above system of defining subgroups limits the subgroups that can be considered. For instance, the subgroup consisting of the union of male smokers and female nonsmokers is not representable by this notation, and is hence excluded from our analysis. But, arguably, it should be excluded, as there is no natural biological reason to think that these two groups of individuals would have the same treatment effect, an effect different from the male nonsmokers or the female smokers.

This restriction to plausible subgroups can incredibly reduce the number of subgroups that must be considered. When $k = 4$, for instance, there are only 81 such allowable subgroups, whereas the total number of possible subgroups is 65,535. The Bayesian approach to multiplicity correction operates by dividing up the total available prior probability of 1 amongst the various models, so limiting the number of allowable subgroups can enormously increase the Bayesian power for detecting subgroup effects. (Of course, the implicit assumption here is that the many implausible subgroups should be assigned a prior probability 0, and this should be checked in applications.)

4.3 Tree-Based Models for Subgroup Analyses

A partition of the population into allowable subgroups is the first step in defining the statistical model. The second step is in specifying which subgroups in the partition exhibit either or both of a zero treatment effect or a zero baseline effect (since, even without treatment, different subgroups may differ in terms of the outcome Y). Both steps of this model construction can be carried out via a tree construction process.

4.3.1 Tree Formulation Rules

Trees (and hence subgroup models) will be generated according to the following rules. When convenient, we illustrate the rules with the 3 covariates pedagogical example.

Covariate choice and ordering: According to probabilities specified later, k out of the p possible covariates are selected, and an ordering of the covariates (e.g., $B_\bullet A_\bullet C_\bullet$) is chosen. This will define the covariates and ordering in which the tree splitting occurs. Thus $B_\bullet A_\bullet C_\bullet$ implies that the first splitting is on Covariate B, which might lead to two possible nodes $B_1 A_\bullet C_\bullet$ and $B_2 A_\bullet C_\bullet$ (i.e., men and women). While the initial node $B_\bullet A_\bullet C_\bullet$ and $A_\bullet B_\bullet C_\bullet$ are logically the same subgroup, we distinguish the splitting order.

Assignment of effect sizes: Let s denote the level of the tree, with the initial node defined as Level $s = 0$.

- When at Level s of the tree, decide with probability γ_s that some node at that level will have zero treatment effect.
  - If this event happens, randomly choose one of the nodes at the s-th level and set its treatment effect to zero. This will be indicated by placing a 0 in front of the node label.
  - If a node is set to be a zero effect node, it is a terminal node and will not be further split.
  - For example, γ_0 would be the probability of setting the initial node to zero effect, and doing so would result in the final node $0A_\bullet B_\bullet C_\bullet$, which here would be the null model of no treatment effect in the entire population.
- Any terminal node that is not a zero treatment effect node is presumed to have an treatment effect size different from any other terminal node.

Splitting of nonzero nodes: All nonzero effect nodes at a given level are subjected to splitting. Splitting will occur probabilistically, according to the following rules:

- All nonzero effect nodes at a given level are associated with the same covariate (the covariate in the corresponding position in the initial covariate ordering) and splits at those nodes can only occur with this covariate.
- If there is a failure to probabilistically split at a node, that node becomes a terminal node.

Assignment of zero baseline effects: Suppose tree growth has stopped (either because the level of the tree has reached its maximum level $s = k$, or the current end nodes of the tree did not split). For the terminal node resulting in subgroup i, set the baseline effect to zero with probability ϱ_i. If this happens, it is so indicated by placing a 0 at the end of the node label.

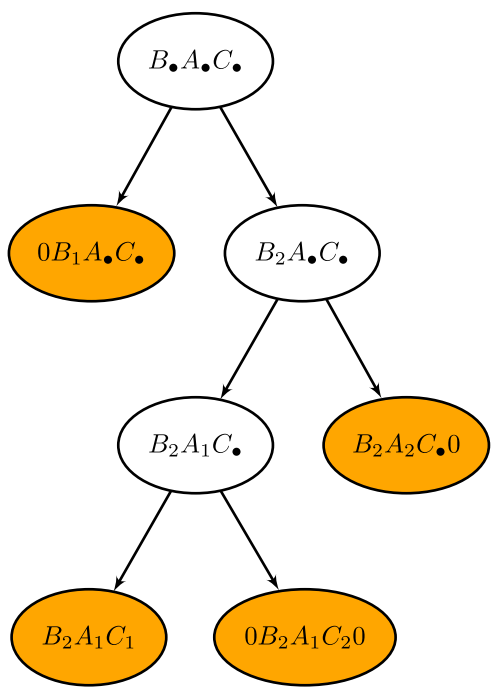


FIGURE 4.1: A three-level tree constructed by the tree formulation rules.

An example is shown in Figure 4.1, using the 3 covariates in Example 4.1. The construction proceeded as follows:

- As here $k = p$, there was no initial selection of covariates. The covariate ordering $B_\bullet A_\bullet C_\bullet$ was chosen.

- At Level $s = 0$, the event of zero mean treatment effect did not occur, and the node was split according to Covariate B, resulting in $B_1 A_\bullet C_\bullet$ and $B_2 A_\bullet C_\bullet$.

- At Level $s = 1$, the event of assigning a zero mean treatment effect did occur; randomly, the left node was chosen to have zero mean and it became the terminal node $0B_1 A_\bullet C_\bullet$.

- The right node at Level 1 did split according to Covariate A, resulting in $B_2 A_1 C_\bullet$ and $B_2 A_2 C_\bullet$.

- At Level $s = 2$, the event of assigning a zero mean treatment effect did not occur, so $B_2 A_1 C_\bullet$ and $B_2 A_2 C_\bullet$ were subjected to possible splitting by Covariate C.

  - The left node at Level 2 split, yielding nodes $B_2 A_1 C_1$ and $B_2 A_1 C_2$.
  - The right node $B_2 A_2 C_\bullet$ did not split and thus became a terminal node.

- At Level $s = 3$, the event of assigning a zero mean treatment effect did occur; randomly, the right node was chosen to have zero mean and it became the terminal node $0B_2 A_1 C_2$.

- Since Level 3 was the maximum tree level, growth stopped, and all terminal nodes (those in orange in Figure 4.1) were probabilistically assigned to have zero or nonzero baseline effects. The assigned zeroes were the rightmost nodes at Level 2 and Level 3, becoming $B_2 A_2 C_\bullet 0$ and $0B_2 A_1 C_2 0$.

It should be emphasized that the statistical model is defined only by the set of terminal nodes, the orange nodes in Figure 4.1. This provides a partition of the entire

population into four subgroups, {males, old females, young smoking females, young non-smoking females}. Furthermore, the model in Figure 4.1 specifies that there is no treatment effect for males or young non-smoking females, while old females and young smoking females have non-zero and differing treatment effects. Finally, the model says that old females and young non-smoking females have zero baseline effects, with the other two subgroups have non-zero baseline effects. It is worth formalizing this notion in the following definition.

Definition 4.2. *A* model *is a partition of the population into subgroups, together with the specification of zero treatment effects and zero baseline effects, arising from the collection of terminal nodes of the tree constructed using the tree formulation rules. The* model space *is the collection of all possible such models.*

4.3.2 Motivation for the Tree Formulation Rules

Clearly the tree formulation rules will only construct trees that yield allowable subgroups as terminal nodes. Additionally, the rules are designed to preclude models (partitions) that are scientifically unreasonable and reduce duplication of models, as explained below.

The motivation for making a node terminal if a split does not occur is to eliminate consideration of scientifically implausible models. For instance, in our pedagogical example, suppose one first split on A and then only split on B for one branch of the A split while only split on C for the other branch, declaring the resulting nodes to be terminal. The resulting partition becomes {young male, young female, old smoker, old non-smoker}, which does not seem a reasonable model (partition) since it would be hard to come with a biological explanation as to why these subgroups – but no finer subgroups – have distinct effects. Note each subgroup itself is a plausible subgroup; it is just that this way of partitioning the populations is questionable.

The motivation for allowing only one zero treatment effect at each level of the

tree is, likewise, based on scientific plausibility of the resulting model (partition). For example, consider the model $\{B_1C_1A_\bullet, 0B_1C_2A_\bullet, 0B_2C_1A_\bullet, B_2C_2A_\bullet 0\}$, which has two zero treatment effect subgroups at Level 2 of the tree. This model says that both non-smoking men and smoking women have zero treatment effect, while the others have non-zero treatment effect, which does not seem scientifically plausible.

There are some scientifically plausible models with more than one zero effect node at a given level, such as $\{0B_1C_1A_\bullet, 0B_1C_2A_\bullet, B_2C_1A_\bullet, B_2C_2A_\bullet 0\}$. But this is equivalent to the model $\{0B_1C_\bullet A_\bullet, B_2C_1A_\bullet, B_2C_2A_\bullet 0\}$, which is already a part of the model space; thus the limitation to one zero treatment effect node at each level also helps to eliminate duplication in the model space.

The convention that each of the nonzero treatment effect subgroups in the model has a different effect size is also a tool to reduce duplication of models. Indeed, taken together, it is easy to see that the tree construction rules do not allow for creation of any duplicate models for a given covariate ordering.

Duplicate models can arise when all the covariate orderings are considered. For instance, with the ordering of $A_\bullet B_\bullet C_\bullet$, the tree which is unable to split on Covariate A would result in the same model as using the ordering of $B_\bullet A_\bullet C_\bullet$ and not splitting on Covariate B, since both are then the entire population. Also, the partition of fully splitting on Covariate A, C, B successively with 8 nonzero treatment effect terminal nodes is equivalent to the partition of 8 terminal nodes from fully splitting on B, C, A successively, assuming no zero treatment effect nodes are introduced. (In a sense, these are the only two ways in which duplicate models can be introduced and it can be seen that the chance of duplicate models arising is actually small.)

In any case, the possibility that duplicate models can arise if different covariate orderings are allowed makes no difference in the posterior analysis; one obtains the same Bayesian result from the duplicate models with separate prior probabilities, as would be obtained by replacing all the duplicates by a single model with its prior

probability equal to the sum of the individual prior probabilities. Having duplicates reduces computational efficiency but, as mentioned earlier, the chance of duplicates is small so that this is not a serious concern.

4.3.3 Outcome Modeling for a Given Model

A Bayesian analysis of the subgroup problem will proceed by considering all models arising from the tree process, computing their marginal likelihoods, and determining their posterior probabilities. Each marginal likelihood computation is separate, internal to the specified model, so it is convenient to switch notation and provide the statistical outcome model specific to each partition.

Suppose the model (partition) h has I_h subgroups, with $n_{i,h}$ individuals in the i-th subgroup. Let $n = \sum_{i=1}^{I_h} n_{i,h}$ be the total number of individuals in the sample. Finally, let $Y_{ij,h}$ denote the outcome for the j-th observation in the i-th subgroup for the model (partition) h.

Usually, in clinical trials, there may be some prognostic covariates $\mathbf{X}_0$ that have a direct effect on the outcome variable regardless of the treatment or control group. Such covariates will always be included in modeling and assumed to be separate from the covariates defining the subgroups. The specific linear regression model assumed for individual j in the i-th subgroup for a given model (partition) h is

$$\begin{aligned} Y_{ij,h} &= \mu_{0,h} + \boldsymbol{X}_{0ij,h}\boldsymbol{\beta}_{0,h} + \mathbf{1}_{\{\beta_{i,h}\neq 0\}}\beta_{i,h} + T_{ij,h}\mu_{i,h}\mathbf{1}_{\{\mu_{i,h}\neq 0\}} + \epsilon_{ij,h}, \\ \epsilon_{ij} &\sim \mathcal{N}(0, \sigma_h^2), \forall i = 1, \cdots, I_h, j = 1, \cdots, n_{i,h} \end{aligned} \tag{4.2}$$

where $\mu_{0,h}$ is an overall baseline, $\boldsymbol{X}_{0ij,h}$ are the common prognostic covariates, $\boldsymbol{\beta}_{0,h}$ is the unknown vector of regression coefficients for these covariates, $T_{ij,h}$ is the treatment indicator (0 or 1), $\mu_{i,h}$ is the treatment effect for subgroup i, $\beta_{i,h}$ is the baseline effect for the i-th subgroup, both $\mathbf{1}_{\{\mu_{i,h}\neq 0\}}$ and $\mathbf{1}_{\{\beta_{i,h}\neq 0\}}$ indicate whether or not they have been set to zero in the tree partition. Note that one reason to allow $\beta_{i,h}$ to be

zero is to minimize the possible confounding of the baseline and treatment effects.

4.4 Specification of Priors

The outcome variable Y can be viewed as arising from a three stage hierarchical mixture model:

- The model (partition) M_h is generated from $\mathrm{P}(M_1), \cdots, \mathrm{P}(M_H)$.
- The model parameter vector $\boldsymbol{\alpha}_h = (\mu_{0,h}, \boldsymbol{\beta}_{0,h}, \beta_{i,h}, \mu_{i,h}, \sigma_h^2)'$ is generated from $\mathrm{P}(\boldsymbol{\alpha}_h \mid M_h)$.
- The data Y is generated from $\mathrm{P}(Y \mid \boldsymbol{\alpha}_h, M_h)$ given in (4.2).

We discuss specification of the prior on each stage in turn.

4.4.1 Specifying Priors on Model Space

According to tree formulation rules in Section 4.3.1, the prior on model space requires specification of

1. the probabilities on the covariate orderings;
2. the probabilities γ_s of assigning zero mean treatment effect to some node at Level s;
3. the tree splitting probabilities;
4. the probabilities ϱ_i of assigning zero baseline effect to subgroup i.

We consider these in order.

Recall that each tree begins with a random selection of k covariates from the $p-$dimensional covariates considered, so that there are $\binom{p}{k}$ possible choices among the covariates to define subgroups. In addition, there are $k!$ possible orderings of

the selected covariates. Here are three possibilities for *selection of the covariates and orderings*:

1. Let each subset and ordering be equally likely, so that the probability of a particular subset and ordering is $(p-k)!/p!$.

2. Assign covariate j a weight w_j and choose the covariates and ordering by selection without replacement proportionally to these weights.

3. Choose the covariates as in Method 2, but then randomly reorder the selected covariates.

Method 1 is natural in problems where there is nothing known about the covariates, and will be used in illustrations herein. Method 2 is attractive when certain covariates are expected to be much more influential, but is much harder to evaluate theoretically. Method 3 is a compromise between the two which seems to preserve the computational simplifications discussed later.

The most natural choice of *tree splitting probabilities* is to associate each covariate with a splitting probability – e.g., A with p_A, B with p_B, and C with p_C. Thus, if one is going to split at Covariate A, with probability p_A one would split to two children and with probability $1-p_A$ one would not split, the latter resulting in a terminal node.

As an example, using Method 1 above and this choice of tree splitting probabilities, the model probabilities in the $k=3$ (with $p=3$) pedagogical example are

$$\begin{aligned} \mathrm{P}(M) &= \frac{(p-k)!}{p!}[p_A^a p_B^b p_C^c][(1-p_A)^{a'}(1-p_B)^{b'}(1-p_C)^{c'}] \\ &\times \prod_{s=0}^{k^*}\left(\frac{\gamma_s}{\ell_s}\right)^{E(s)}(1-\gamma_s)^{1-E(s)} \times \prod_{i=1}^{I}\varrho_i^{\mathbf{1}_{\{\beta_i=0\}}}(1-\varrho_i)^{(1-\mathbf{1}_{\{\beta_i=0\}})}, \end{aligned} \tag{4.3}$$

where a is the number of nonzero effect nodes at which an A split occurs, a' is the number of nonzero effect nodes at which an A split does not occur, etc.; $k^* \leq k$ is the last level of the tree, $E(s)$ is the indicator to show whether or not the tree has a zero node at Level s, ℓ_s is the number of nodes at Level s of the tree, and I is the number of terminal nodes.

An obvious possibility for the *zero treatment effect probabilities* γ_s is to choose them to be equal (and will be done herein), but increasing γ_s might be needed, especially when the splitting probabilities are large and one wants to preclude the trees from becoming too large.

Similarly, an obvious possibility for the *zero baseline effect probabilities* ϱ_i is to choose them to be equal (and will be done herein). However, it might be more reasonable to choose the probability ϱ_i higher when the i-th group has been assigned zero mean treatment effect.

4.4.2 Specifying Priors for Parameters in Outcome Models

For simplicity of notation, we will drop use of the model index h in this subsection. Also, we only consider the case in which there are no general prognostic covariates $\mathbf{X}_0$, although regression coefficients for that situation can be handled in the same fashion as μ_0 below.

The marginal likelihood model for M is

$$\mathrm{P}(\mathbf{Y} \mid M) = \int \mathrm{P}(\mathbf{Y} \mid \boldsymbol{\alpha}, M)\mathrm{P}(\boldsymbol{\alpha} \mid M)d\boldsymbol{\alpha}\,. \tag{4.4}$$

Since there can be many models considered, it is desirable to choose $\mathrm{P}(\boldsymbol{\alpha} \mid M)$ so that the computation does not involve high-dimensional integration. A reasonable

choice of this prior is

$$\begin{aligned}
\mu_i &\sim \mathcal{N}(\mu_i \mid \mu_m, \nu^2), \\
\mu_m &\sim \mathcal{N}(0, \omega^2), \\
\beta_i &\sim \mathcal{N}(\beta_i \mid 0, \tau^2), \\
\pi(\mu_0, \sigma^2) &\propto \frac{1}{\sigma^2},
\end{aligned}$$

where ω^2, ν^2 and τ^2 are hyperparameters that, for the moment, we consider specified. Note that μ_0 and σ^2 are common to all models and, hence, can be assigned the usual objective prior. However, the other parameters are not common to all models (and, in particular, none occur in the null model of no treatment effect with zero baseline effect whatsoever), and thus they must be assigned proper priors.

4.5 Approximating the Marginal Likelihood of a Model

Suspend the model index h for simplicity. Define $\theta = (\mu_0, \beta, \mu)'$ with $\beta = (\beta_1 \cdots, \beta_I)'$ and $\mu = (\mu_1, \cdots, \mu_I)'$, where I is the number of subgroups in the model. Also define

$$\begin{aligned}
\mathbf{Y} &= (Y_{1,1}, \cdots, Y_{1,n_1}, Y_{2,1}, \cdots, Y_{2,n_2}, \cdots, Y_{I,1}, \cdots, Y_{I,n_I})', \\
\boldsymbol{\epsilon} &= (\epsilon_{1,1}, \cdots, \epsilon_{1,n_1}, \epsilon_{2,1}, \cdots, \epsilon_{2,n_2}, \cdots, \epsilon_{I,1}, \cdots, \epsilon_{I,n_I})',
\end{aligned}$$

and the design matrix

$$\mathbf{X} = \begin{pmatrix}
\mathbf{1}_{n_1} & \mathbf{1}_{n_1}\mathbf{1}_{\{\beta_1 \neq 0\}} & & & \mathbf{T}_1\mathbf{1}_{\{\mu_1 \neq 0\}} & & \\
\vdots & & \ddots & & & \ddots & \\
\mathbf{1}_{n_I} & & & \mathbf{1}_{n_I}\mathbf{1}_{\{\beta_I \neq 0\}} & & & \mathbf{T}_I\mathbf{1}_{\{\mu_I \neq 0\}}
\end{pmatrix}_{n \times (2I+1)}$$

where $\mathbf{T}_i = (T_{i,1}, \cdots, T_{i,n_i})'$, $\forall i \in \{1, \cdots, I\}$, and white space in the matrix $\mathbf{X}$ implies that the corresponding entries are zero. We can then rewrite the outcome model as

$$\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon} \tag{4.5}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n\times n})$ and $\mathbf{I}_{n\times n}$ is the $n \times n$ identity matrix.

Finally, note that the two sufficient statistics for this linear model are

$$\begin{aligned} \hat{\theta} &= (\hat{\mu}_0, \hat{\beta}, \hat{\mu})' = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}), \\ s^2 &= (\mathbf{Y} - \mathbf{X}\hat{\theta})'(\mathbf{Y} - \mathbf{X}\hat{\theta}), \end{aligned}$$

and that we can write the outcome model density as

$$\begin{aligned} f(\mathbf{Y}|\theta, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{(2I+1)/2}} \exp\left(-\frac{(\theta - \hat{\theta})'(\mathbf{X}'\mathbf{X})(\theta - \hat{\theta})}{2\sigma^2}\right) \\ &\cdot \frac{1}{(2\pi\sigma^2)^{(n-2I-1)/2}} \exp\left(-\frac{(\mathbf{Y} - \mathbf{X}\hat{\theta})'(\mathbf{Y} - \mathbf{X}\hat{\theta})}{2\sigma^2}\right). \end{aligned} \tag{4.6}$$

We now derive an approximation to the marginal likelihood for the specified model and prior in three steps.

Step 1. Elimination of σ^2. We first approximately integrate out σ^2 in (4.6), by replacing σ^2 in the first factor by its maximum likelihood estimate $\hat{\sigma}^2 = s^2/(n-2I-1)$, and then integrating out over the second factor with respect to the prior $1/\sigma^2$. The result is

$$\begin{aligned} &\int_0^\infty f(\mathbf{Y}|\theta, \sigma^2)\pi(\sigma^2)d\sigma^2 \\ \approx\ & \frac{1}{(2\pi)^{(2I+1)/2} \mid \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \mid^{1/2}} \exp\left(-\frac{(\theta - \hat{\theta})'(\mathbf{X}'\mathbf{X})(\theta - \hat{\theta})}{2\hat{\sigma}^2}\right) \\ \cdot\ & \frac{\mid \mathbf{X}'\mathbf{X} \mid^{-1/2}}{(\pi)^{(n-2I-1)/2}} \frac{\Gamma(\frac{n-2I-1}{2})}{(s^2)^{(n-2I-1)/2}} \\ =\ & \mathcal{N}(\hat{\theta} \mid \theta, \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}) \cdot \frac{\mid \mathbf{X}'\mathbf{X} \mid^{-1/2}}{(\pi)^{(n-2I-1)/2}} \frac{\Gamma(\frac{n-2I-1}{2})}{(s^2)^{(n-2I-1)/2}}. \end{aligned} \tag{4.7}$$

This approximation is typically very good when degrees of freedom $n - 2I - 1$ is at least moderate. Note that some of the β_i and μ_i could be zero, and so one could

increase the degrees of freedom available by including those terms in the second factor of the likelihood rather than the first factor. Again this makes little difference when $n-2I-1$ is at least moderate, and complicates the ensuing expressions considerably.

Step 2. Integrating out θ. With the approximation in Step 1, θ now occurs in a normal likelihood and has a normal prior $\pi(\theta \mid \omega^2, \nu^2, \tau^2)$, allowing closed form integration. The following presents a computationally convenient version of the resulting marginal likelihood.

Lemma 4.3. *Define $\theta_{2I} = (\beta, \mu)'$ and its corresponding estimate $\hat{\theta}_{2I} = (\hat{\beta}, \hat{\mu})'$, which are the vectors θ and $\hat{\theta}$ without their first element. Then*

$$\int \mathcal{N}(\hat{\theta} \mid \theta, \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})\pi(\theta \mid \omega^2, \nu^2, \tau^2)d\theta = \frac{1}{(2\pi)^I \mid \mathbf{\Sigma}_{22} \mid^{1/2}} \exp\left(-\frac{\hat{\theta}_{2I}'\mathbf{\Sigma}_{22}^{-1}\hat{\theta}_{2I}}{2}\right), \tag{4.8}$$

where

$$\mathbf{\Sigma}_{22} = \mathbf{\Sigma}_{22}(\omega^2, \nu^2, \tau^2) = \begin{pmatrix} \nu^2\mathbf{I}_{I\times I} + \omega^2(\mathbf{1}_I\mathbf{1}_I') + \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \tau^2\mathbf{I}_{I\times I} + \mathbf{L}_{22} \end{pmatrix},$$

with

$$\mathbf{L} = \mathbf{D}^{-1}(\mathbf{I}_{2I\times 2I} + m\mathbf{Q}) := \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} a_1 & & & c_1 & & \\ & \ddots & & & \ddots & \\ & & a_I & & & c_I \\ c_1 & & & b_1 & & \\ & \ddots & & & \ddots & \\ & & c_I & & & b_I \end{pmatrix}_{2I\times 2I},$$

$$
\begin{aligned}
m &= \frac{1}{n - \sum_{i=1}^{I} \frac{a_i b_i (a_i + b_i - 2c_i)}{a_i b_i - c_i^2}}, \\
\mathbf{Q} &= \begin{pmatrix}
\frac{a_1 b_1 (a_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{a_1 b_I (a_I - c_I)}{a_I b_I - c_I^2} & \frac{a_1^2 (b_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{a_1 a_I (b_I - c_I)}{a_I b_I - c_I^2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{a_I b_1 (a_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{a_I b_I (a_I - c_I)}{a_I b_I - c_I^2} & \frac{a_1 a_I (b_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{a_I^2 (b_I - c_I)}{a_I b_I - c_I^2} \\
\frac{b_1^2 (a_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{b_1 b_I (a_I - c_I)}{a_I b_I - c_I^2} & \frac{a_1 b_1 (b_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{a_I b_1 (b_I - c_I)}{a_I b_I - c_I^2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\frac{b_1 b_I (a_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{b_I^2 (a_I - c_I)}{a_I b_I - c_I^2} & \frac{a_1 b_I (b_1 - c_1)}{a_1 b_1 - c_1^2} & \cdots & \frac{a_I b_I (b_I - c_I)}{a_I b_I - c_I^2}
\end{pmatrix}_{2I \times 2I}.
\end{aligned}
$$

and

$$
\begin{aligned}
a_i &= n_i \mathbf{1}_{\{\beta_i \neq 0\}}, \\
b_i &= \sum_{j=1}^{n_i} T_{i,j} \mathbf{1}_{\{\mu_i \neq 0\}} = \sum_{j=1}^{n_i} T_{i,j}^2 \mathbf{1}_{\{\mu_i \neq 0\}}, \\
c_i &= \sum_{j=1}^{n_i} T_{i,j} \mathbf{1}_{\{\beta_i \neq 0, \mu_i \neq 0\}}.
\end{aligned}
$$

Proof. Write

$$
\begin{aligned}
\mathbf{X}'\mathbf{X} &= \begin{pmatrix} n & \mathbf{s}' \\ \mathbf{s} & \mathbf{D} \end{pmatrix}, \\
\mathbf{s} &= (a_1, \cdots, a_I, b_1, \cdots, b_I)'.
\end{aligned}
$$

Then, the multivariate normal distribution of $\hat{\theta}$ can be rewritten as

$$
\hat{\theta} \sim \mathcal{N}\left(\begin{pmatrix} \mu_0 \\ \theta_{2I} \end{pmatrix}, \hat{\sigma}^2 \begin{pmatrix} n & \mathbf{s}' \\ \mathbf{s} & \mathbf{D} \end{pmatrix}^{-1} \right).
$$

Since the priors for θ_{2I} and μ_0 are

$$
\begin{aligned}
\pi(\mu_0) &\propto 1, \\
\theta_{2I} &\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \nu^2 \mathbf{I}_{I \times I} + \omega^2 (\mathbf{1}_I \mathbf{1}_I') & \\ & \tau^2 \mathbf{1}_{I \times I} \end{pmatrix} \right),
\end{aligned}
$$

marginalization yields

$$\hat{\theta}_{2I} \sim \mathcal{N}\left(\mathbf{0}, \hat{\sigma}^2 \left(\mathbf{D} - \frac{\mathbf{s}\mathbf{s}'}{n}\right)^{-1} + \begin{pmatrix} \nu^2 \mathbf{I}_{I\times I} + \omega^2(\mathbf{1}_I \mathbf{1}_I') & \\ & \tau^2 \mathbf{1}_{I\times I} \end{pmatrix}\right)$$

Using Lemma B.3 in Appendix B, algebra yields

$$\left(\mathbf{D} - \frac{\mathbf{s}\mathbf{s}'}{n}\right)^{-1} = \mathbf{D}^{-1} + \frac{\mathbf{D}^{-1}\mathbf{s}\mathbf{s}'\mathbf{D}^{-1}}{n - \mathbf{s}'\mathbf{D}^{-1}\mathbf{s}} = \mathbf{L}.$$

Thus, the density of $\hat{\theta}_{2I}$ is

$$f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22}) = \frac{1}{(2\pi)^I \mid \mathbf{\Sigma}_{22} \mid^{1/2}} \exp\left(-\frac{\hat{\theta}_{2I}' \mathbf{\Sigma}_{22}^{-1} \hat{\theta}_{2I}}{2}\right),$$

completing the proof.

□

Step 3. Estimating the hyperparameters. Using (4.7) and (4.8), the marginal likelihood for a given model, conditional on the hyperparmeters, is

$$\mathrm{P}(\mathbf{Y} \mid M, \omega^2, \nu^2, \tau^2) \approx \frac{\mid \mathbf{X}'\mathbf{X} \mid^{-1/2}}{(\pi)^{(n-2I-1)/2}} \frac{\Gamma(\frac{n-2I-1}{2})}{(s^2)^{(n-2I-1)/2}} \cdot \frac{1}{(2\pi)^I \mid \mathbf{\Sigma}_{22} \mid^{1/2}} \exp\left(-\frac{\hat{\theta}_{2I}' \mathbf{\Sigma}_{22}^{-1} \hat{\theta}_{2I}}{2}\right), \tag{4.9}$$

where $\mathbf{\Sigma}_{22} = \mathbf{\Sigma}_{22}(\omega^2, \nu^2, \tau^2)$ depends on the hyperparameters.

A fully Bayesian approach to dealing with the hyperparameters would be to assign these parameters a proper prior and integrate the marginal likelihood over this prior. We instead adopt the empirical Bayesian approach, which is to replace the hyperparameters by their marginal maximum likelihood estimates. The marginal likelihood for these hyperparameters is the model-averaged marginal likelihood

$$\mathcal{L}(\mathbf{Y} \mid \omega^2, \nu^2, \tau^2) = \sum_{h=1}^{H} \mathrm{P}(M_h)\mathrm{P}(\mathbf{Y} \mid M_h, \omega^2, \nu^2, \tau^2),$$

where the sum is over all H models. When the models cannot be enumerated, this will need to be done iteratively, with identification of high probability models to include in the sum. After, obtaining the maximum likelihood estimates of ν^2, τ^2 and ω^2, they will simply be inserted into (4.9) to obtain the final approximation to the model marginal maximum likelihood.

For completeness, we present the equations utilized to compute the model-averaged marginal maximum likelihood estimates for ω^2, ν^2 and τ^2. From Lemma B.2 in Appendix B,

$$\mathbf{\Sigma}_{22}^{-1} = \begin{pmatrix} \mathbf{B}_1^{-1} & -\mathbf{B}_1^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-1} \\ -\mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-1} & \mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-1} + \mathbf{B}_2^{-1} \end{pmatrix},$$

where $\mathbf{B}_1 = \nu^2\mathbf{I}_{I\times I} + \omega^2(\mathbf{1}_I\mathbf{1}_I') + \mathbf{L}_{11} - \mathbf{L}_{12}\mathbf{B}_2^{-1}\mathbf{L}_{21}$ and $\mathbf{B}_2 = \tau^2\mathbf{I}_{I\times I} + \mathbf{L}_{22}$.

In a given model, the logarithm of $f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22})$ is

$$\log f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22}) = -I\log(2\pi) - \frac{1}{2}\log \mid \mathbf{\Sigma}_{22} \mid - \frac{1}{2}\hat{\theta}_{2I}\mathbf{\Sigma}_{22}^{-1}\hat{\theta}_{2I}.$$

The derivatives of $\log(f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22}))$ are

$$\begin{aligned}
\frac{\partial \log f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22})}{\partial \nu^2} &= \frac{1}{2}\hat{\theta}_{2K}\mathbf{\Sigma}_{22}^{-1}\frac{\partial \mathbf{\Sigma}_{22}}{\partial \nu^2}\mathbf{\Sigma}_{22}^{-1}\hat{\theta}_{2I} - \frac{1}{2}\mathrm{Trace}(\mathbf{\Sigma}_{22}^{-1}\frac{\partial \mathbf{\Sigma}_{22}}{\partial \nu^2}) \\
&= \frac{1}{2}\hat{\theta}_{2I}\begin{pmatrix} \mathbf{B}_1^{-2} & -\mathbf{B}_1^{-2}\mathbf{L}_{12}\mathbf{B}_2^{-1} \\ -\mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-2} & \mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-2}\mathbf{L}_{12}\mathbf{B}_2^{-1} \end{pmatrix}\hat{\theta}_{2I} - \frac{1}{2}\sum_{i=1}^{I} b_{ii}^1, \\
\frac{\partial \log f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22})}{\partial \omega^2} &= \frac{1}{2}\hat{\theta}_{2I}\mathbf{\Sigma}_{22}^{-1}\frac{\partial \mathbf{\Sigma}_{22}}{\partial \omega^2}\mathbf{\Sigma}_{22}^{-1}\hat{\theta}_{2I} - \frac{1}{2}\mathrm{Trace}(\mathbf{\Sigma}_{22}^{-1}\frac{\partial \mathbf{\Sigma}_{22}}{\partial \omega^2}) \\
&= \frac{1}{2}\hat{\theta}_{2I}\begin{pmatrix} \mathbf{B}_1^{-1}\mathbf{E}\mathbf{B}_1^{-1} & -\mathbf{B}_1^{-1}\mathbf{E}\mathbf{B}^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-1} \\ -\mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-1}\mathbf{E}\mathbf{B}_1^{-1} & \mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-1}\mathbf{E}\mathbf{B}_1^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-1} \end{pmatrix}\hat{\theta}_{2I} - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{I} b_{ij}^1, \\
\frac{\partial \log f(\hat{\theta}_{2I} \mid \mathbf{\Sigma}_{22})}{\partial \tau^2} &= \frac{1}{2}\hat{\theta}_{2I}\mathbf{\Sigma}_{22}^{-1}\frac{\partial \mathbf{\Sigma}_{22}}{\partial \tau^2}\mathbf{\Sigma}_{22}^{-1}\hat{\theta}_{2I} - \frac{1}{2}\mathrm{Trace}(\mathbf{\Sigma}_{22}^{-1}\frac{\partial \mathbf{\Sigma}_{22}}{\partial \nu^2}) \\
&= \frac{1}{2}\hat{\theta}_{2I}\begin{pmatrix} \mathbf{B}_1^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-2}\mathbf{L}_{21}\mathbf{B}_1^{-1} & \mathbf{B}_1^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-1}\mathbf{B}_L \\ -\mathbf{B}_L\mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-1} & \mathbf{B}_L\mathbf{B}_L \end{pmatrix}\hat{\theta}_{2I} - \frac{1}{2}\sum_{i=1}^{I} b_{ii}^{\ell},
\end{aligned}$$

where $\mathbf{B}_1^{-1} = (b_{ij}^1)_{I\times I}$, $\mathbf{B}_L = \mathbf{B}_2^{-1}\mathbf{L}_{21}\mathbf{B}_1^{-1}\mathbf{L}_{12}\mathbf{B}_2^{-1} + \mathbf{B}_2^{-1} = (b_{ij}^{\ell})_{I\times I}$ and

$$\mathbf{E} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}_{I\times I} .$$

Therefore, the overall marginal maximum likelihood estimates of ν^2, ω^2 and τ^2 are the solution to the system of equations

$$\begin{aligned} \sum_{h=1}^{H} \mathrm{P}(M_h) \frac{\log f(\hat{\theta}_{2I,h} \mid \mathbf{\Sigma}_{22,h})}{\partial \nu^2} &= 0, \\ \sum_{h=1}^{H} \mathrm{P}(M_h) \frac{\log f(\hat{\theta}_{2I,h} \mid \mathbf{\Sigma}_{22,h})}{\partial \omega^2} &= 0, \\ \sum_{h=1}^{H} \mathrm{P}(M_h) \frac{\log f(\hat{\theta}_{2I,h} \mid \mathbf{\Sigma}_{22,h})}{\partial \tau^2} &= 0\,. \end{aligned}$$

4.6 Posterior Inferences for Subgroup Analyses

Information about unknowns is encoded in the posterior distribution, which consists of the posterior model probabilities

$$\mathrm{P}(M_h \mid \mathbf{Y}) = \frac{\mathrm{P}(\mathbf{Y} \mid M_h)\mathrm{P}(M_h)}{\sum_{\kappa=1}^{H} Pr(\mathbf{Y} \mid M_\kappa)\mathrm{P}(M_\kappa)}\,, \tag{4.10}$$

as well as the posterior distributions of model parameters given the models. Of particular interest will be the posterior distributions of the $\mu_{i,h}$ which are the mean treatment effects for the i-th subgroup in model M_h.

Note that computations can be challenging when the number of models is large. For instance, if the dimension of the subgroup covariates is $p = 32$ and one only allows subgroup models of at most $k = 2$ covariates there are still $148,800$ potential models. For $k = 3$, enumeration of the models would not be feasible. In such cases,

intelligent stochastic search schemes in model space would be necessary, and sums over model space become sums over the models found in the stochastic search.

There are also issues in summarizing the information available from the posterior distribution on model space. Individual models are typically not of interest here, since particular subgroups will appear in many models. One marginal posterior quantity of very clear interest is the posterior distribution of the mean effect size of the treatment for a given individual (or, equivalently, for a given specification of the covariates). A useful summary of this individual posterior would be

- P_{0j} = the probability of no treatment effect for individual j,
- the posterior distribution of
  Δ = the posterior treatment effect for individual j given there is an effect,
  which can be summarized by the mean and variance of Δ.

Letting $\mu_{i(j),h}$ denote the mean treatment effect in the subgroup of Model M_h to which individual j belongs, it is clear that these quantities are given through model averaging (see Clyde et al. (1996) and Hoeting et al. (1999)) as

$$\begin{aligned}
P_{0j} &= \sum_h \mathrm{P}(M_h \mid Y)\mathbf{1}_{\{\mu_{i(j),h}=0\}}, \\
\mathrm{E}(\Delta \mid \mathbf{Y}) &= \frac{\sum_h \mathrm{E}(\mu_{i(j),h} \mid M_h, \mathbf{Y})\mathrm{P}(M_h \mid \mathbf{Y})\mathbf{1}_{\{\mu_{i(j),h}\neq 0\}}}{\sum_h \mathrm{P}(M_h \mid \mathbf{Y})\mathbf{1}_{\{\mu_{i(j),h}\neq 0\}}}, \\
\mathrm{Var}(\Delta \mid \mathbf{Y}) &= \frac{\sum_h [\mathrm{Var}(\mu_{i(j),h} \mid M_h, \mathbf{Y}) + (\mathrm{E}(\mu_{i(j),h} \mid M_h, \mathbf{Y}))^2]\mathrm{P}(M_h \mid \mathbf{Y})\mathbf{1}_{\{\mu_{i(j),h}\neq 0\}}}{\sum_h \mathrm{P}(M_h \mid \mathbf{Y})\mathbf{1}_{\{\mu_{i(j),h}\neq 0\}}} \\
&\quad - (\mathrm{E}(\Delta \mid \mathbf{Y}))^2.
\end{aligned}$$

To complete this computation, we need to obtain the posterior mean and variance of each $\mu_{i(j),h}$. The following lemma provides these for the case of known σ_h^2 and known hyperparameters. In practice, we will replace these by their maximum likelihood and empirical Bayes estimates.

Lemma 4.4. *For simplicity, we suspend the model index h. Define $\beta = (\beta_1, \cdots, \beta_I)'$ and $\mu = (\mu_1, \cdots, \mu_I)'$. The posterior mean and covariance matrix of (β, μ) are*

$$\mathrm{E}(\beta, \mu \mid \mathbf{Y}) = \mathbf{\Delta}^{-1}\mathbf{X}'_{-1}(\mathbf{I}_{n\times n} - \frac{\mathbf{1}_n\mathbf{1}'_n}{n})\mathbf{Y}$$

and

$$\mathrm{Cov}(\beta, \mu \mid \mathbf{Y}) = \sigma^2\mathbf{\Delta}^{-1},$$

where

$$\begin{aligned}
\mathbf{\Delta} &= \mathbf{X}'_{-1}(\mathbf{I}_{n\times n} - \frac{\mathbf{1}_n\mathbf{1}'_n}{n})\mathbf{X}_{-1} + \sigma^2\mathbf{\Omega}, \\
\mathbf{\Omega} &= \begin{pmatrix} \nu^{-2}(\mathbf{I}_{I\times I} - \frac{\omega^2}{\nu^2+I\omega^2}\mathbf{1}_I\mathbf{1}'_I) & \\ & \tau^{-2}\mathbf{I}_{I\times I} \end{pmatrix},
\end{aligned}$$

and $\mathbf{X}_{-1}$ is the matrix $\mathbf{X}$ without the first row and the first column.

Proof. Given $\sigma^2, \nu^2, \omega^2, \tau^2$, from model (4.5) and the prior distributions of (β, μ) and μ_0, we have

$$\begin{aligned}
\mathbf{Y} \mid \theta, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\theta, \sigma^2\mathbf{I}_{n\times n}), \\
\beta, \mu \mid \nu^2, \omega^2, \tau^2 &\sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \nu^2\mathbf{I}_{I\times I} + \omega^2(\mathbf{1}_I\mathbf{1}'_I) & \\ & \tau^2\mathbf{1}_{I\times I} \end{pmatrix}\right), \\
\pi(\mu_0) &\propto 1.
\end{aligned}$$

Therefore, the conditional posterior distribution of β, μ is

$$\beta, \mu \mid \mathbf{Y}, \sigma^2, \nu^2, \omega^2, \tau^2 \sim \mathcal{N}\left(\mathbf{\Delta}^{-1}\mathbf{X}'_{-1}(\mathbf{I}_{n\times n} - \frac{\mathbf{1}_n\mathbf{1}'_n}{n})\mathbf{Y}, \sigma^2\mathbf{\Delta}^{-1}\right),$$

completing the proof. □

It may well be of interest to present the posterior summary for a higher level subgroup, such as the entire population. It is not immediately clear how to do

this, in part because it is unclear how to define subgroup effects in general. One natural definition would be to define the treatment effect size for a subgroup to be the average of the treatment effect sizes for all individuals in the subgroup. Likewise the probability there is an treatment effect in the subgroup could be defined as the average of all treatment effect probabilities of the individuals in the subgroup.

5

Concluding Remarks and Future Work

This thesis has investigated a variety of interesting practical problems and developed several novel Bayesian modeling approaches to solve them by using latent structures in the prior specification. These problems cover item response theory, computer model emulation and subgroup analyses, with complex data arising in dynamic, constrained and dependent situations, respectively. The contribution lies in the development of three new models from a Bayesian perspective, i.e. Dynamic Item Response Models, Gaussian Process Models with Shape Constraints, and Tree-based Models for Subgroup Analyses, to cope with modeling complex data, and the development of efficient computational techniques for these models.

Chapter 2 considers a problem motivated from a time series of dichotomous response data collected by adaptive measurement testing in education. The proposed Dynamic Item Response Models generalizes the classical item response models in three important ways, allowing 1) longitudinal observations at variable and irregular time points; 2) potential dependence of items in a test; and 3) uncertainties associated with item difficulties. The developed Bayesian approach to analyzing DIR models not only allows for borrowing strength across individuals, but also enables

the retrospective analysis of an individual's changing ability, as well as online prediction of their ability change. Posterior propriety of the objective Bayesian analysis is rigorously proved and the techniques of the proof should have wider applicability to investigations of posterior propriety in other general state space models with logistic links and objective priors.

Chapter 3 introduces a GP prior for shape constrained functions, with the constraints imposed on the derivative process of the original GP. Two alternative ways to introduce constraints in the GP prior are proposed. Although the second method, i.e. the conditional GP method, is more computationally intensive, it has the additional flexibility of allowing for functions that have flat areas. The methodology is illustrated on emulation of a CRASH computer model, relevant since GP models are extensively used in emulation of computer models, and shape constraints are also common therein.

Chapter 4 develops a tree-based model to analyze subgroup effects from a Bayesian perspective. The subgroups arise from terminal nodes of trees based on covariate splits. Following this approach and utilizing certain rules for tree construction lead to a dramatically reduced number of biologically reasonable subgroups that need be considered. The stochastic mechanism to generate the tree allows subgroups to have zero mean treatment effect and/or zero underlying subgroup baseline effect, which is a key to controlling for multiplicity. The terminal nodes of the tree become the model (partition of the population into subgroups) and Bayesian model selection techniques can be utilized to yield a multiplicity-controlled posterior subgroup analysis.

5.1 Future Work on Dynamic Item Response Models

Typically, as indicated in Figure 5.1, an individual's ability grows quickly initially, but slows with maturity. Indeed, it is typical that one's ability plateaus at some point, e.g. τ in Figure 5.1. Therefore, a valuable extension of DIR models will

incorporate a plateau effect.

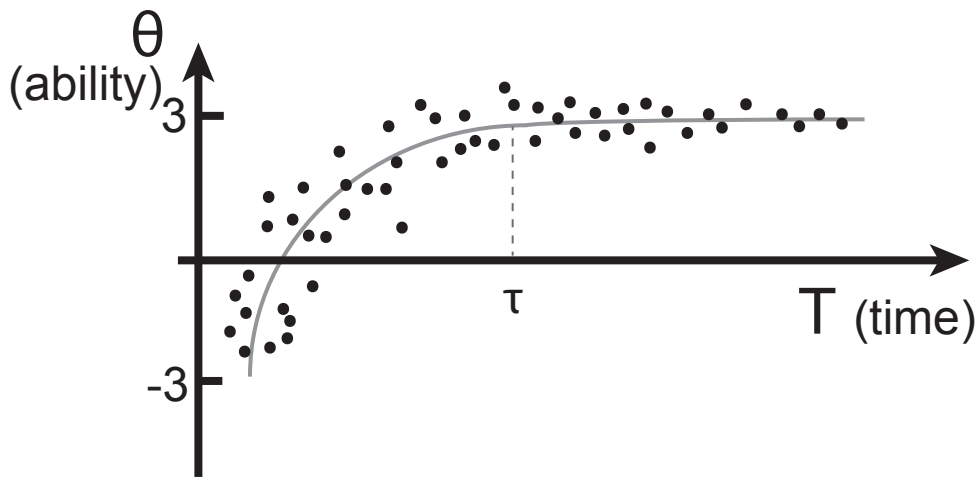


FIGURE 5.1: Typical measurements of an individual's ability over time.

This can naturally be done by introducing a change point in DIR models. Recalling that the current DIR models described in Chapter 2 have the form

$$\theta_{i,t} = \theta_{i,t-1} + c_i(1 - \rho\theta_{i,t-1})\Delta^+_{i,t} + w_{i,t}\,,$$

a reasonable system equation involving a change point to indicate when the plateau has occurred is

$$\theta_{i,t} = \theta_{i,t-1} + c_i(1 - \rho_i\theta_{i,t-1})Z^{(\tau_i)}_{i,t}\Delta^+_{i,t} + w_{i,t}, \tag{5.1}$$

where $Z^{(\tau_i)}_{i,t}$ is the indicator function, i.e.

$$Z^{(\tau_i)}_{i,t} = \begin{cases} 0 & \text{if} \quad t \geq \tau_i, \\ 1 & \text{otherwise} \end{cases}$$

with $\tau_i \in \{1, \cdots, T_i\}$. Note that this also requires using individual ρ_i in (5.1), since plateaus will typically occur at different times.

5.2 Future Work on Gaussian Process Models with Shape Constraints

As mentioned in Chapter 3, one interesting topic is to extend the methodology to higher dimensional inputs in GP models. Thus, suppose $\mathbf{t}$ now denotes a D-dimensional input, and denote $\mathbf{T}$ as an $n \times D$ matrix for all observed inputs. Then the model considered for the unknown function becomes

$$x_i = Z(\mathbf{t}_i) + \epsilon_i, \quad i = 1, \ldots n,$$

at inputs $\mathbf{t}_i$. Note that the other specifications of the model are unchanged. Assume a Gaussian process prior for the unknown function $Z(\mathbf{t}_i)$ as usual, i.e.

$$Z(\mathbf{t}_i) \sim \mathcal{GP}(\mu, K(\mathbf{t}_i, \mathbf{t}_i'))$$

with squared exponential covariance function $K(\cdot, \cdot)$ as

$$K(\mathbf{t}_i, \mathbf{t}_i') = \sigma_z^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \beta_d^{-2}(t_d^i - t_d^{i'})^2\right),$$

where $\mathbf{t}_i = (t_1^i, \cdots, t_d^i)^T$, σ_z^2 and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_D)^T$ are the hyperparameters for the GP prior. By applying Theorem 2.2.2 in Adler (1981), we have

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial Z(\mathbf{t})}{\partial t_d}\right] &= \frac{\partial \mu}{\partial t_d} = 0, \\
\mathrm{cov}\left[\frac{\partial Z(\mathbf{t})}{\partial t_g}, Z(\mathbf{t}')\right] &= \sigma_z^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \beta_d^{-2}(t_d - t_d')^2\right)\left(-\frac{1}{\beta_g^2}(t_g - t_g')\right), \\
\mathrm{cov}\left[\frac{\partial Z(\mathbf{t})}{\partial t_g}, \frac{\partial Z(\mathbf{t}')}{\partial t_h'}\right] &= \sigma_z^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \beta_d^{-2}(t_d - t_d')^2\right)\frac{1}{\beta_g^2}\left(\delta_{gh} - \frac{1}{\beta_h^2}(t_h - t_h')(t_g - t_g')\right),
\end{aligned}$$

where $\delta_{gh} = 1$ if $g = h$, and 0 otherwise.

Obtaining analogs of Lemma 3.1 and Lemma 3.2 to estimate the constrained function is thus quite straightforward. However, the methodology for determining

the locations of derivative points does not easily generalize, and will need to be studied.

Another interesting topic is to develop a fully Bayesian method to estimate the unknown functions. The MCMC algorithm combined with some adaptive procedure may achieve this. Some pseudo code might look like:

Input: Initialize derivative points $\mathbf{Z}(\mathbf{s})^{(0)}$, predictive points $\mathbf{Z}(\mathbf{t}^*)^{(0)}$ and $\theta^{(0)}$.

Repeat for $j = 1$ **to** M **do**

- Sample the ith derivative point for $i = 1, \cdots, m$: $\mathbf{Z}'(s_i)^{(j+1)} \sim \pi(\mathbf{Z}(s_i) \mid \mathbf{Z}(s_{-i})^{(j)}, \theta^{(j)}, \mathbf{X})$,
- Sample $\mathbf{Z}(\mathbf{t}^*)^{(j+1)}$: $\mathbf{Z}(\mathbf{t}^*)^{(j+1)} \sim \pi(\mathbf{Z}(\mathbf{t}) \mid \mathbf{Z}(\mathbf{s})^{(j)}, \theta^{(j)}, \mathbf{X})$,
- Generate $\theta^{(j+1)}$ from some suitable proposal and accept or reject $\theta^{(j+1)}$ with the Metropolis-Hastings algorithm.

end for

Until The convergence of the Markov chain is achieved.

More delicate work might be asked for designing some adaptive procedures for the changes of the dimensionality of $\mathbf{Z}(\mathbf{s})$ if the value of θ differs substantially between each iteration.

5.3 Future Work on Tree-based Models for Subgroup Analyses

While Chapter 4 presents a strategy for Bayesian subgroup analyses, more experience is needed to understand the needed prior inputs to the algorithm. In particular, there is great flexibility in the choice of probabilities defining the tree growth, and understanding how these relate to intuitive prior beliefs of investigators will be a key step in building a tree prior to reflect features of relative importance of covariates in determining its covariate ordering probability as well as specifying the spiltting

probability to split one covariate. This would implicitly help us to reduce the complexity (i.e. size and the shape) of the tree and increase the probability of searching more interested subgroups for researchers.

It is also of interest to further investigate how to summarize the posterior effect for subgroups. For instance, suppose one is interested in an overall effect. One could only consider those models (partitions) for which all the subgroup effects are nonzero. The overall effect probability and the overall effect distribution would be the appropriate averages an mixtures over these models. This idea is simply illustrated in the example below, where we follow the convention of letter notation from the pedagogical example in Chapter 4:

Example 5.1. *Suppose the following are the possible models (ignoring the fixed effects for the moment). Along with each model is given the model posterior probability and the posterior effect distribution (given nonzero). Suppose there are 20 individuals in A_1 and 30 in A_2.*

- $A_\bullet$*: 0.2,* $N(\mu \mid 2, 1)$
- $0A_\bullet$*: 0.3*
- $\{A_1, A_2\}$*: 0.2,* $N\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} \middle| \begin{pmatrix}1\\ 3\end{pmatrix}, \begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}\right)$
- $\{A_1, 0A_2\}$*: 0.1,* $N(\mu_1 \mid 1.5, 3)$
- $\{0A_1, A_2\}$*: 0.2,* $N(\mu_2 \mid 2.5, 3)$

The posterior probability of an overall effect would be 0.2+ 0.2 = 0.4. The posterior distribution of an overall effect, given an effect exists, requires defining what is the overall effect in the situation $\{A_1, A_2\}$. Since we are within a fixed model, it is natural to define the overall effect as the relevant proportional sum of effects μ_1 and

μ_2 in the two subgroups, here $[\frac{20}{50}\mu_1 + \frac{30}{50}\mu_2] = N(2.2, 1.04)$. Thus the final overall effects distribution, given that there is an effect, would be

$$\frac{0.2}{0.4}\, N(2,1) + \frac{0.2}{0.4} N(2.2, 1.04)\,.$$

This has a mean of 2.1.

However, it is not obvious which way is best to measure the spread for the overall effects in the example above. It is also of interest to investigate the use of correct Bayesian marginalization over the hyperparameters to obtain the model marginal likelihoods, instead of the use of the empirical Bayesian method.

Finally, the number of subgroup partitions (models) becomes very large, a stochastic search of the model space of partitions becomes necessary. In this situation, the Markov chain Monte Carlo (MCMC) method are often used, which would facilitate to identify high probability models for selection or model averaging. For example, Chipman et al. (1998) and Denison et al. (1998) simulated a Markov chain sequence of models by using the Metropolis-Hastings algorithm to stochastically gravitate toward the higher posterior regions. Hoeting et al. (1999) adopted the MCMC model composition methodology of Madigan et al. (1995) to generate a stochastic process that moves through model space. Green (1995) used the reversible jump MCMC for this search. Other MCMC methods to explore the model space are used proposals such as Evolutionary Monte Carlo (Liang and Wong (2000) and Wilson et al. (2010)), Swendsen-Wang (Nott and Kohn (2005)) and adaptive MCMC (Nott and Kohn (2005)). Moreover, a more recent development of Bayesian adaptive sampling for model averaging is the paper of Clyde et al. (2011) and their algorithm of sampling without replacement is based on binary trees. This paper is particularly interesting to us since it enlightens us there might be some possibility to connect some of those ideas on searching the model space in the way we build our tree-based models. Thus,

one of major goals in our next step would be focusing on an efficient and suitable stochastic searching algorithm in our context of the tree-based model for analyzing subgroups.

Appendix A

Posterior Propriety of DIR Models

We first give some needed lemmas that may be of independent interest for proving posterior propriety in other logistic modeling scenarios.

Lemma A.1. *For any three real numbers x, ϵ_1 and ϵ_2,*

$$\frac{e^{x+\epsilon_1}}{1+e^{x+\epsilon_1}} \times \frac{1}{1+e^{x+\epsilon_2}} \leq e^{-|x|+|\epsilon_1|+|\epsilon_2|}\,.$$

Proof. It is easy to see that

$$\frac{e^{x+\epsilon_1}}{1+e^{x+\epsilon_1}} \times \frac{1}{1+e^{x+\epsilon_2}} = \frac{1}{1+e^{-x-\epsilon_1}} \times \frac{1}{1+e^{x+\epsilon_2}} = \frac{1}{1+e^{x+\epsilon_2}+e^{-x-\epsilon_1}+e^{\epsilon_2-\epsilon_1}}$$

Moreover, we have

$$\begin{aligned} & \left(1+e^{x+\epsilon_2}+e^{-x-\epsilon_1}+e^{\epsilon_2-\epsilon_1}\right)\left(e^{-|x|+|\epsilon_1|+|\epsilon_2|}\right) \\ = \; & e^{-|x|+|\epsilon_1|+|\epsilon_2|}+e^{x-|x|+|\epsilon_1|+\epsilon_2+|\epsilon_2|}+e^{-x-|x|-\epsilon_1+|\epsilon_1|+|\epsilon_2|}+e^{-|x|+|\epsilon_1|-\epsilon_1+|\epsilon_2|+\epsilon_2} \end{aligned}$$

Let us discuss two situations:

1. if $x \geq 0$, then $x - |x| = 0$. If $\epsilon_2 \leq 0$, we have $\epsilon_2 + |\epsilon_2| = 0$, otherwise, we have $\epsilon_2 + |\epsilon_2| \geq 0$. Thus, $e^{x-|x|+|\epsilon_1|+\epsilon_2+|\epsilon_2|} \geq e^0 = 1$. Then

$$\left(1 + e^{x+\epsilon_2} + e^{-x-\epsilon_1} + e^{\epsilon_2-\epsilon_1}\right)\left(e^{-|x|+|\epsilon_1|+|\epsilon_2|}\right) \geq e^{x-|x|+|\epsilon_1|+\epsilon_2+|\epsilon_2|} \geq 1.$$

2. if $x < 0$, then $-x - |x| = 0$. If $\epsilon_1 \geq 0$, we have $-\epsilon_1 + |\epsilon_1| = 0$, otherwise, we have $-\epsilon_1 + |\epsilon_1| \geq 0$. Thus, $e^{-x-|x|-\epsilon_1+|\epsilon_1|+|\epsilon_2|} \geq e^0 = 1$. Then

$$\left(1 + e^{x+\epsilon_2} + e^{-x-\epsilon_1} + e^{\epsilon_2-\epsilon_1}\right)\left(e^{-|x|+|\epsilon_1|+|\epsilon_2|}\right) \geq e^{-x-|x|-\epsilon_1+|\epsilon_1|+|\epsilon_2|} \geq 1.$$

Therefore, in either situations, we have

$$\left(1 + e^{x+\epsilon_2} + e^{-x-\epsilon_1} + e^{\epsilon_2-\epsilon_1}\right)\left(e^{-|x|+|\epsilon_1|+|\epsilon_2|}\right) \geq 1,$$

i.e.

$$\frac{1}{1 + e^{x+\epsilon_2} + e^{-x-\epsilon_1} + e^{\epsilon_2-\epsilon_1}} \leq e^{-|x|+|\epsilon_1|+|\epsilon_2|}.$$

Therefore,

$$\frac{e^{x+\epsilon_1}}{1 + e^{x+\epsilon_1}} \times \frac{1}{1 + e^{x+\epsilon_2}} \leq e^{-|x|+|\epsilon_1|+|\epsilon_2|}$$

would hold for any situations. □

Lemma A.2. *For $\theta_i \in (-\infty, \infty)$, $i = 1, 2$,*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{-1/2} e^{-\tau(\eta_1^2+\eta_2^2)} e^{-(|\theta_1+\eta_1|+|\theta_1-\eta_1|+|\theta_2+\eta_2|+|\theta_2-\eta_2|)} d\tau d\eta_1 d\eta_2 \leq (4 + 2\pi)\sqrt{\pi} e^{-(|\theta_1|+|\theta_2|)}.$$

Proof. Firstly, let us see the integration below

$$\int_0^{\infty} \frac{1}{\sqrt{\tau}} e^{-\tau(\eta_1^2+\eta_2^2)} d\tau = \frac{\sqrt{\pi}}{\sqrt{\eta_1^2 + \eta_2^2}}.$$

Moreover, it is easy to prove that

$$\begin{aligned} |\theta_1 + \eta_1| + |\theta_1 - \eta_1| &= 2\max\{|\theta_1|, |\eta_1|\}, \\ |\theta_2 + \eta_2| + |\theta_2 - \eta_2| &= 2\max\{|\theta_2|, |\eta_2|\}, \end{aligned}$$

and the relationship of $2\max\{|\theta_1|,|\eta_1|\} \geq |\theta_1|+|\eta_1|$ and $2\max\{|\theta_2|,|\eta_2|\} \geq |\theta_2|+|\eta_2|$.

Thus,

$$\begin{aligned}
& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{-1/2} e^{-\tau(\eta_1^2+\eta_2^2)} e^{-(|\theta_1+\eta_1|+|\theta_1-\eta_1|+|\theta_2+\eta_2|+|\theta_2-\eta_2|)} d\tau d\eta_1 d\eta_2 \\
\leq \; & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\sqrt{\pi}}{\sqrt{\eta_1^2+\eta_2^2}} e^{-(|\theta_1|+|\eta_1|+|\theta_2|+|\eta_2|)} d\eta_1 d\eta_2 \\
= \; & \sqrt{\pi} e^{-(|\theta_1|+|\theta_2|)} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\sqrt{\eta_1^2+\eta_2^2}} e^{-(|\eta_1|+|\eta_2|)} d\eta_1 d\eta_2.
\end{aligned}$$

Let us focus on the integration part, we could decompose it into two parts, i.e.

$$\int_{\sqrt{\eta_1^2+\eta_2^2}\leq 1} \frac{1}{\sqrt{\eta_1^2+\eta_2^2}} e^{-(|\eta_1|+|\eta_2|)} d\eta_1 d\eta_2 \leq \int_{\sqrt{\eta_1^2+\eta_2^2}\leq 1} \frac{1}{\sqrt{\eta_1^2+\eta_2^2}} d\eta_1 d\eta_2 = \int_0^{2\pi}\int_0^1 dr d\theta = 2\pi,$$

and

$$\begin{aligned}
& \int_{\sqrt{\eta_1^2+\eta_2^2}>1} \frac{1}{\sqrt{\eta_1^2+\eta_2^2}} e^{-(|\eta_1|+|\eta_2|)} d\eta_1 d\eta_2 \\
\leq \; & \int_{\sqrt{\eta_1^2+\eta_2^2}>1} e^{-(|\eta_1|+|\eta_2|)} d\eta_1 d\eta_2 \\
\leq \; & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(|\eta_1|+|\eta_2|)} d\eta_1 d\eta_2 \\
= \; & 4.
\end{aligned}$$

Therefore,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} \tau^{-1/2} e^{-\tau(\eta_1^2+\eta_2^2)} e^{-(|\theta_1+\eta_1|+|\theta_1-\eta_1|+|\theta_2+\eta_2|+|\theta_2-\eta_2|)} d\tau d\eta_1 d\eta_2 \leq K e^{-(|\theta_1|+|\theta_2|)},$$

where the constant $K = (4+2\pi)\sqrt{\pi}$, which completes the proof.

□

Lemma A.3. *For $\theta_i \in (-\infty, \infty)$, $i = 1, 2$,*

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} \delta^{-1/2} e^{-\frac{\delta}{2}(\varphi_1^2+\varphi_2^2)} e^{-(|\theta_1+\varphi_1|+|\theta_2+\varphi_2|)} d\delta d\varphi_1 d\varphi_2 \leq \frac{K}{1+|\theta_1|},$$

with some constant K.

Proof. Analogous to Lemma A.2, we have

$$\int_0^{\infty} \frac{1}{\sqrt{\delta}} e^{-\frac{\delta}{2}(\varphi_1^2+\varphi_2^2)} d\delta = \frac{\sqrt{2\pi}}{\sqrt{\varphi_1^2+\varphi_2^2}}.$$

Then, some derivation would yield

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{0}^{\infty} \delta^{-1/2} e^{-\frac{\delta}{2}(\varphi_1^2+\varphi_2^2)} e^{-(|\theta_1+\varphi_1|+|\theta_2+\varphi_2|)} d\delta d\varphi_1 d\varphi_2 \\
= &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\sqrt{2\pi}}{\sqrt{\varphi_1^2+\varphi_2^2}} e^{-(|\theta_1+\varphi_1|+|\theta_2+\varphi_2|)} d\delta d\varphi_1 d\varphi_2 \\
\leq &\frac{K}{1+\sqrt{\theta_1^2+\theta_2^2}},
\end{aligned}$$

where K is a constant and the last inequality holds by using similar idea as the last part of derivations in Lemma A.2. Moreover,

$$\frac{K}{1+\sqrt{\theta_1^2+\theta_2^2}} \leq \frac{K}{1+|\theta_1|},$$

which completes the proof. □

Lemma A.4. *For $T \geq 2$,*

$$\begin{aligned}
\int_0^{\infty}\int_0^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{\phi^{3/2}} \quad &\cdot \quad \frac{1}{1+|\sqrt{\frac{B(c)}{\phi}}z + A(c)|} e^{-\frac{z^2}{2}} \\
&\cdot \quad \frac{1}{1+|\sqrt{\frac{B'(c')}{\phi}}z' + A'(c')|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi < \infty\,,
\end{aligned} \tag{A.1}$$

where

$$
\begin{aligned}
A(c) &= \mu_{G_j}\prod_{t=1}^{T}(1-c\rho\Delta_t^+)+\sum_{t=1}^{T}c\Delta_t^+\prod_{i=t+1}^{T}(1-c\rho\Delta_i^+),\\
B(c) &= \sum_{t=1}^{T}\Delta_t\prod_{i=t+1}^{T}(1-c\rho\Delta_i^+)^2+\phi V_{G_j}\prod_{t=1}^{T}(1-c\rho\Delta_t^+)^2,\\
A'(c') &= \mu_{G_j}\prod_{t=1}^{T}(1-c'\rho\Delta_t^+)+\sum_{t=1}^{T}c'\Delta_t^+\prod_{i=t+1}^{T}(1-c'\rho\Delta_i^+),\\
B'(c') &= \sum_{t=1}^{T}\Delta_t\prod_{i=t+1}^{T}(1-c'\rho\Delta_i^+)^2+\phi V_{G_j}\prod_{t=1}^{T}(1-c'\rho\Delta_t^+)^2\,,
\end{aligned}
$$

and we have dropped the label i in the subscripts for $\Delta_{i,t}$, $\Delta_{i,t}^+$, c_i, $\mu_{G_{j_i}}$ and $V_{G_{j_i}}$.

Proof. Let K' be sufficient large, then when $c' \geq K'$, we have

$$\sum_{t=1}^{T}\Delta_t\prod_{i=t+1}^{T}(1-c'\rho\Delta_i^+)^2 \sim O(c'^{2(T-1)}),$$

and

$$V_G\prod_{t=1}^{T}(1-c'\rho\Delta_t^+)^2 \sim O(c'^{2T}).$$

Similarly, let $c \geq K$, where K be sufficient large, then

$$\sum_{t=1}^{T}\Delta_t\prod_{i=t+1}^{T}(1-c\rho\Delta_i^+)^2 \sim O(c^{2(T-1)}),$$

and

$$V_G\prod_{t=1}^{T}(1-c\rho\Delta_t^+)^2 \sim O(c^{2T}).$$

Let us define $M = \max\{K', K, \frac{1}{\rho\Delta_1}+1, \cdots, \frac{1}{\rho\Delta_T}+1\}$. Moreover, for simplicity, we use Label A instead of $A(c)$ and Label B instead of $B(c)$ and similarly for A' and B'. We are going to decompose the integral of (A.1) into six different regions:

1. Let us consider the integral (A.1) in the region of $c, c' < M$ and $\phi \geq 1$, i.e.

$$\int_1^\infty \int_0^M \int_0^M \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi$$

$$\begin{aligned}
< \quad & M^2 \int_1^\infty \frac{1}{\phi^{3/2}} d\phi \int_{-\infty}^\infty e^{-\frac{z^2}{2}} dz \int_{-\infty}^\infty e^{-\frac{z'^2}{2}} dz' \\
= \quad & -2M^2 \frac{1}{\phi^{1/2}}\Big|_1^\infty \\
= \quad & 2M^2.
\end{aligned}$$

2. Let us look at the region of $c, c' \geq M$ and $\phi \geq 1$. Firstly, define

$$\begin{aligned}
|h| \quad &= \quad \frac{|A|}{|\sqrt{\frac{B}{\phi}}|} \\
&= \quad \frac{|\mu_G \prod_{t=1}^T (1 - c\rho\Delta_t^+) + \sum_{t=1}^T c\Delta_t^+ \prod_{i=t+1}^T (1 - c\rho\Delta_i^+)|}{\sqrt{\frac{\sum_{t=1}^T \Delta_t \prod_{i=t+1}^T (1-c\rho\Delta_i^+)^2}{\phi} + V_G \prod_{t=1}^T (1 - c\rho\Delta_t^+)^2}} \\
&< \quad \frac{|\mu_G \prod_{t=1}^T (1 - c\rho\Delta_t^+) + \sum_{t=1}^T c\Delta_t^+ \prod_{i=t+1}^T (1 - c\rho\Delta_i^+)|}{\sqrt{V_G \prod_{t=1}^T (1 - c\rho\Delta_t^+)^2}} \\
&\sim \quad O(\frac{c^T \rho^{T-1} \prod_{t=1}^T \Delta_t^+}{\sqrt{V_G} c^T \rho^T \prod_{t=1}^T \Delta_t^+}) \quad \text{when } c \geq M \\
&= \quad O(\frac{1}{\sqrt{V_G}\rho})
\end{aligned}$$

Thus, when $c \geq M$, we have $|h| \leq O(\frac{1}{\sqrt{V_G}\rho})$. Similarly, we have $|h'| = |A'|/|\sqrt{\frac{B'}{\phi}}| \leq O(\frac{1}{\sqrt{V_G}\rho})$. Moreover, let us denote $B_1 = V_G \prod_{t=1}^T (1 - c\rho\Delta_t^+)^2$ and $B_1' = V_G \prod_{t=1}^T (1 - c'\rho\Delta_t^+)^2$. Then, after some algebra, the integral (A.1)

in the region of $c, c' \geq M$ and $\phi \geq 1$ would be bounded as

$$
\begin{aligned}
& \int_1^\infty \int_M^\infty \int_M^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi \\
< \; & \int_1^\infty \int_M^\infty \int_M^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} e^{-\frac{z^2}{2}} \frac{1}{1+\sqrt{\frac{B'}{\phi}}|z'+h'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi \\
\leq \; & M_5^2 \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{\log(1+\sqrt{B_1})}{\sqrt{B_1}} \frac{\log(1+\sqrt{B_1'})}{\sqrt{B_1'}} dc dc' d\phi \\
+ \; & M_5 M_6 \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{\log(1+\sqrt{B_1})}{\sqrt{B_1}} \frac{1}{1+\sqrt{B_1'}} dc dc' d\phi \\
+ \; & M_5 M_6 \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+\sqrt{B_1}} \frac{\log(1+\sqrt{B_1'})}{\sqrt{B_1'}} dc dc' d\phi \\
+ \; & M_6^2 \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+\sqrt{B_1}} \frac{1}{1+\sqrt{B_1'}} dc dc' d\phi \\
= \; & O\Bigg(\int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{\log(1+M_7^T c^T)}{M_7^T c^T} \frac{\log(1+M_7^T c'^T)}{M_7^T c'^T} dc dc' d\phi \\
+ \; & \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{\log(1+M_7^T c^T)}{M_7^T c^T} \frac{1}{1+M_7^T c'^T} dc dc' d\phi \\
+ \; & \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+M_7^T c^T} \frac{\log(1+M_7^T c'^T)}{M_7^T c^T} dc dc' d\phi \\
+ \; & \int_1^\infty \int_M^\infty \int_M^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+M_7^T c^T} \frac{1}{1+M_7^T c'^T} dc dc' d\phi \Bigg) \\
\leq \; & O\left(\frac{2M_8^2}{M_7^2} + \frac{2M_8}{M_7^2} \frac{(M_7 M)^{1-T}}{T-1} + \frac{2M_8}{M_7^2} \frac{(M_7 M)^{1-T}}{T-1} + \frac{2}{M_7^2} \frac{(M_7 M)^{2(1-T)}}{(T-1)^2} \right) \\
< \; & \infty,
\end{aligned}
$$

where

$$
\begin{aligned}
M_5 &= \max\{e^{-\frac{z^2}{2}}, z \in (-1-h, 1-h)\}, \\
M_6 &= \int_{\infty}^{\infty} e^{-\frac{z^2}{2}} dz \geq \int_{|z+h|\geq 1} e^{-\frac{z^2}{2}} dz, \\
M_7 &= \rho \times (V_G \prod_{t=1}^{T} \Delta_t^2)^{\frac{1}{T}}, \\
M_8 &= \frac{(M_7 M)^{1-T} T[1+(T-1)log(M_7 M)]}{(T-1)^2}.
\end{aligned}
$$

3. Let us compute the integral (A.1) in the region $c, c' \geq M$ and $\phi < 1$. Define

$$B_2 = \sum_{t=1}^{T} \Delta_t \prod_{i=t+1}^{T} (1 - c\rho\Delta_i^+)^2,$$

and

$$B_2' = \sum_{t=1}^{T} \Delta_t \prod_{i=t+1}^{T} (1 - c'\rho\Delta_i^+)^2.$$

Similarly as derived in $c, c' \geq M$ and $\phi \geq 1$, we would obtain

$$
\begin{aligned}
& \int_0^1 \int_M^{\infty} \int_M^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi \\
< & \int_0^1 \int_M^{\infty} \int_M^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\phi^{3/2}} \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} e^{-\frac{z^2}{2}} \frac{1}{1+\sqrt{\frac{B'}{\phi}}|z'+h'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi
\end{aligned}
$$

Firstly, let us derive the bound of the integral below

$$
\begin{aligned}
& \int_M^\infty \int_{-\infty}^\infty \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} e^{-\frac{z^2}{2}} dz dc \\
= & \int_M^{\frac{M}{\sqrt{\phi}}} \int_{-\infty}^\infty \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} e^{-\frac{z^2}{2}} dz dc + \int_{\frac{M}{\sqrt{\phi}}}^\infty \int_{-\infty}^\infty \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} e^{-\frac{z^2}{2}} dz dc \\
\leq & \int_M^{\frac{M}{\sqrt{\phi}}} \int_{-\infty}^\infty \frac{1}{1+\sqrt{\frac{B_2}{\phi}}|z+h|} e^{-\frac{z^2}{2}} dz dc + \int_{\frac{M}{\sqrt{\phi}}}^\infty \int_{-\infty}^\infty \frac{1}{1+\sqrt{B_1}|z+h|} e^{-\frac{z^2}{2}} dz dc \\
\leq & M_5 \int_M^{\frac{M}{\sqrt{\phi}}} \frac{\log(1+\sqrt{\frac{B_2}{\phi}})}{\sqrt{\frac{B_2}{\phi}}} dc + M_6 \int_M^{\frac{M}{\sqrt{\phi}}} \frac{1}{\sqrt{\frac{B_2}{\phi}}} dc \\
+ & M_5 \int_{\frac{M}{\sqrt{\phi}}}^\infty \frac{\log(1+\sqrt{B_1})}{\sqrt{B_1}} dc + M_6 \int_{\frac{M}{\sqrt{\phi}}}^\infty \frac{1}{\sqrt{B_1}} dc \\
= & O\left(\int_M^{\frac{M}{\sqrt{\phi}}} \frac{\log(1+\frac{M_9^{T-1}c^{T-1}}{\sqrt{\phi}})}{1+\frac{M_9^{T-1}c^{T-1}}{\sqrt{\phi}}} dc + \int_{\frac{M}{\sqrt{\phi}}}^\infty \frac{\log(1+M_7^T c^T)}{M_7^T c^T} dc \right) \\
\leq & O\left(\frac{\sqrt{\phi}\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log(1+\frac{M_9 M}{\phi})}{M_9} + \frac{\phi^{\frac{T-1}{2}} T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]}{M_7^T M^{T-1}(T-1)^2} \right),
\end{aligned}
$$

where $M_9 = \rho \times (\Delta_1 \prod_{t=2}^T \Delta_t^2)^{1/(T-1)}$. Similarly, we have

$$
\begin{aligned}
& \int_M^\infty \int_{-\infty}^\infty \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} e^{-\frac{z^2}{2}} dz dc \qquad \text{(A.2)} \\
\leq & O\left(\frac{\sqrt{\phi}\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log(1+\frac{M_9 M}{\phi})}{M_9} + \frac{\phi^{\frac{T-1}{2}} T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]}{M_7^T M^{T-1}(T-1)^2} \right)
\end{aligned}
$$

Therefore, when $c, c' \geq M$ and $\phi < 1$, and when $T \geq 2$, the integral would be

bounded by

$$\int_0^1 \int_M^\infty \int_M^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi$$

$$\begin{aligned}
\leq \quad & O\left(\int_0^1 \frac{\log^2(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log^2(1+\frac{M_9 M}{\phi})}{\sqrt{\phi} M_9^2} d\phi \right. \\
+ \quad & 2\int_0^1 \frac{\phi^{\frac{T-2}{2}} T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})] \log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log(1+\frac{M_9 M}{\phi})}{M_9 M_7^T M^{T-1}(T-1)^2\sqrt{\phi}} d\phi \\
+ \quad & \left. \int_0^1 \frac{\phi^{T-2} T^2 [1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]^2}{M_7^{2T} M^{2(T-1)}(T-1)^4 \sqrt{\phi}} d\phi \right) \\
\leq \quad & O\left(\int_0^1 \frac{\log^2(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log^2(1+\frac{M_9 M}{\phi})}{\sqrt{\phi} M_9^2} d\phi \right. \\
+ \quad & 2\int_0^1 \frac{T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})] \log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log(1+\frac{M_9 M}{\phi})}{M_9 M_7^T M^{T-1}(T-1)^2\sqrt{\phi}} d\phi \\
+ \quad & \left. \int_0^1 \frac{T^2 [1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]^2}{M_7^{2T} M^{2(T-1)}(T-1)^4 \sqrt{\phi}} d\phi \right).
\end{aligned}$$

Since $\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}$ and $\frac{M_9 M}{\phi}$ are sufficient large and set $d = M_9 M$, then

$$\begin{aligned}
& \int_0^1 \frac{\log^2(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log^2(1+\frac{M_9 M}{\phi})}{\sqrt{\phi} M_9^2} d\phi \qquad \text{(A.3)} \\
= \quad & O\left(\int_0^1 \frac{\log^2(\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log^2(\frac{M_9 M}{\phi})}{\sqrt{\phi}} d\phi \right) \\
= \quad & O\left(\int_0^1 \frac{\left[(T-1)\log d - \frac{T}{2}\log(\phi)\right]^2 \left[\log d - \log(\phi)\right]^2}{\sqrt{\phi}} d\phi \right)
\end{aligned}$$

$$
\begin{aligned}
&= O\Bigg(\int_0^1 \frac{(T-1)^2\log^4 d}{\sqrt{\phi}}d\phi - \int_0^1 \frac{2(T-1)\log^2 d\log\phi}{\sqrt{\phi}}d\phi \\
&+ \int_0^1 \frac{(T-1)\log^2 d\log^2\phi}{\sqrt{\phi}}d\phi + \int_0^1 \frac{T(T-1)\log^3 d\log\phi}{\sqrt{\phi}}d\phi \\
&- \int_0^1 \frac{2T(T-1)\log^2 d\log^2\phi}{\sqrt{\phi}}d\phi + \int_0^1 \frac{T(T-1)\log d\log^3\phi}{\sqrt{\phi}}d\phi \\
&+ \int_0^1 \frac{\frac{T}{4}\log^2 d\log^2\phi}{\sqrt{\phi}}d\phi - \int_0^1 \frac{\frac{T}{2}\log d\log^3\phi}{\sqrt{\phi}}d\phi + \int_0^1 \frac{\frac{T}{4}\log^4\phi}{\sqrt{\phi}}d\phi\Bigg)
\end{aligned}
$$

It is easy to see that

$$
\begin{aligned}
\int_0^1 \frac{1}{\sqrt{\phi}}d\phi &= 2, \\
\int_0^1 \frac{\log\phi}{\sqrt{\phi}}d\phi &= -4, \\
\int_0^1 \frac{\log^2\phi}{\sqrt{\phi}}d\phi &= 16, \\
\int_0^1 \frac{\log^3\phi}{\sqrt{\phi}}d\phi &= -96, \\
\int_0^1 \frac{\log^4\phi}{\sqrt{\phi}}d\phi &= 768.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\int_0^1 \frac{\log^2(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log^2(1+\frac{M_9M}{\phi})}{\sqrt{\phi}M_9^2}d\phi \\
&= O\left(2(T-1)^2\log^4 d + 8(T-1)\log^2 d + 16(T-1)\log^2 d - 4T(T-1)\log^3 d\right. \\
&- \left. 32T(T-1)\log^2 d - 96T(T-1) + 4T\log^2 d + 48T\log d + 192T\right)
\end{aligned}
$$

Then,

$$
\int_0^1 \frac{\log^2(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log^2(1+\frac{M_9M}{\phi})}{\sqrt{\phi}M_9^2}d\phi < \infty.
$$

Similar derivation would lead to

$$\int_0^1 \frac{T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log(1+\frac{M_9 M}{\phi})}{M_9 M_7^T M^{T-1}(T-1)^2\sqrt{\phi}} d\phi < \infty$$

and

$$\int_0^1 \frac{T^2[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]^2}{M_7^{2T} M^{2(T-1)}(T-1)^4\sqrt{\phi}} d\phi < \infty.$$

Therefore,

$$\int_0^1 \int_M^\infty \int_M^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z+A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi < \infty.$$

4. Next, we look at the integral in the region $c, c' < M$ and $\phi < 1$, i.e

$$\int_0^1 \int_0^M \int_0^M \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z+A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi.$$

Since $c, c' < M$, it is easy to prove that $|h|, |h'| \leq M_{10}$, where M_{10} is a constant.

Similarly, as the derivation before, we could obtain that

$$\int_0^1 \int_0^M \int_0^M \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z+A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi$$

$$\leq \quad M_5^2 \int_0^1 \int_0^M \int_0^M \frac{1}{\phi^{3/2}} \frac{\log(1+\sqrt{\frac{B}{\phi}})}{\sqrt{\frac{B}{\phi}}} \frac{\log(1+\sqrt{\frac{B'}{\phi}})}{\sqrt{\frac{B'}{\phi}}} dc dc' d\phi$$

$$+ \quad M_5 M_6 \int_0^1 \int_0^M \int_0^M \frac{1}{\phi^{3/2}} \frac{\log(1+\sqrt{\frac{B}{\phi}})}{\sqrt{\frac{B}{\phi}}} \frac{1}{1+\sqrt{\frac{B'}{\phi}}} dc dc' d\phi$$

$$
\begin{aligned}
&+ \quad M_5 M_6 \int_0^1 \int_0^M \int_0^M \frac{1}{\phi^{3/2}} \frac{1}{1+\sqrt{\frac{B}{\phi}}} \frac{\log(1+\sqrt{\frac{B'}{\phi}})}{\sqrt{\frac{B'}{\phi}}} dcdc'd\phi \\
&+ \quad M_6^2 \int_0^1 \int_0^M \int_0^M \frac{1}{\phi^{3/2}} \frac{1}{1+\sqrt{\frac{B}{\phi}}} \frac{1}{1+\sqrt{\frac{B'}{\phi}}} dcdc'd\phi \\
&\leq \quad M_5^2 M_{11} \int_0^1 \frac{\log^2(1+\frac{M_{12}}{\sqrt{\phi}})}{\sqrt{\phi}} d\phi \\
&+ \quad 2M_5 M_6 M_{11} \int_0^1 \frac{\log(1+\frac{M_{12}}{\sqrt{\phi}})}{\sqrt{\phi}} d\phi \\
&+ \quad M_6^2 M_{11}^2 \int_0^1 \frac{1}{\sqrt{\phi}} d\phi
\end{aligned}
$$

where

$$
\begin{aligned}
M_{11} &= \frac{M}{\Delta_T}, \\
M_{12} &= \max\{\sqrt{B_1}, \sqrt{B_2}, \sqrt{B_1'}, \sqrt{B_2'}, c, c' \in (0, M)\}.
\end{aligned}
$$

It is easy to obtain that

$$
\begin{aligned}
\int_0^1 \frac{\log(1+\frac{M_{12}}{\sqrt{\phi}})}{\sqrt{\phi}} d\phi &\leq 2(\log(M_{12}+1)+1), \\
\int_0^1 \frac{\log^2(1+\frac{M_{12}}{\sqrt{\phi}})}{\sqrt{\phi}} d\phi &\leq 2(\log(M_{12}+1)+1)^2+2, \\
\int_0^1 \frac{1}{\sqrt{\phi}} d\phi &= 2.
\end{aligned}
$$

Therefore, we have

$$
\int_0^1 \int_0^M \int_0^M \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z+A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|} e^{-\frac{z'^2}{2}} dzdz'dcdc'd\phi < \infty.
$$

5. Now, we consider the integral in the region $c \geq M$, $c' < M$ and $\phi < 1$. Since c and c' are symmetric in the integral, if the the integral is finite in the region $c \geq M$, $c' < M$ and $\phi < 1$, then by the symmetric property of c and c', we would easily obtain that the integral is finite in the region $c < M$, $c' \geq M$ and $\phi < 1$. Thus, we only need to prove the integral below is finite,

$$\int_0^1 \int_0^M \int_M^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi.$$

From (A.2), we know that

$$\begin{aligned}
& \int_M^\infty \int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \\
= & \int_M^\infty \int_{-\infty}^\infty \frac{1}{1+\sqrt{\frac{B}{\phi}}|z + h|} e^{-\frac{z^2}{2}} dz dc \\
\leq & \; O\left(\frac{\sqrt{\phi}\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log(1+\frac{M_9 M}{\phi})}{M_9} + \frac{\phi^{\frac{T-1}{2}} T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]}{M_7^T M^{T-1}(T-1)^2} \right)
\end{aligned}$$

and also similarly as the derivation in Situation 4, we know that

$$\begin{aligned}
& \int_0^M \int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz' dc' \\
\leq & \; M_5 M_{11}\sqrt{\phi}\log(1+\frac{M_{12}}{\sqrt{\phi}}) + M_6 M_{11}\sqrt{\phi}.
\end{aligned}$$

Then, when $T \geq 2$, the integral below is bounded by

$$\int_0^1 \int_0^M \int_M^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \frac{1}{\phi^{3/2}} \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} e^{-\frac{z^2}{2}} \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z' + A'|} e^{-\frac{z'^2}{2}} dz dz' dc dc' d\phi$$

$$\begin{aligned}
\leq \; & O\Bigg(\int_0^1 \frac{\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log(1+\frac{M_9 M}{\phi}) \log(1+\frac{M_{12}}{\sqrt{\phi}})}{\sqrt{\phi}} d\phi \\
+ \; & \int_0^1 \frac{T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})] \log(1+\frac{M_{12}}{\sqrt{\phi}})}{M_7^T M^{T-1}(T-1)^2\sqrt{\phi}} d\phi \\
+ \; & \int_0^1 \frac{\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) \log(1+\frac{M_9 M}{\phi})}{\sqrt{\phi}} d\phi + \int_0^1 \frac{T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]}{M_7^T M^{T-1}(T-1)^2\sqrt{\phi}} d\phi \Bigg)
\end{aligned}$$

Since $\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}$ and $\frac{M_9 M}{\phi}$ are sufficient large,

$$\begin{aligned}
\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}) &\cong \log(\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}}), \\
\log(1+\frac{M_9 M}{\phi}) &\cong \log(\frac{M_9 M}{\phi}).
\end{aligned}$$

Moreover, if $\frac{M_{12}}{\sqrt{\phi}}$ is sufficient large, then

$$\log(1+\frac{M_{12}}{\sqrt{\phi}}) \cong \log(\frac{M_{12}}{\sqrt{\phi}}).$$

However, when $\frac{M_{12}}{\sqrt{\phi}}$ is not sufficient large for ϕ closer to 1, then we just depart the integral of ϕ into two parts, one is from 0 to ϵ, where any $\phi \leq \epsilon$ would guarantee that $\frac{M_{12}}{\sqrt{\phi}}$ is sufficient large, and the other one is from ϵ to 1. Since the integral of ϕ from ϵ to 1 is finite, the discussion of finiteness for the integral of ϕ from 0 to ϵ would be the same as what we discuss here for the integral of ϕ

from 0 to 1. Thus, without loss of generality, we assume that $\frac{M_{12}}{\sqrt{\phi}}$ is sufficient large for $\phi \in (0,1)$. Then, by similar derivation as (A.3), we could obtain that

$$\begin{aligned}
\int_0^1 \frac{\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log(1+\frac{M_9 M}{\phi})\log(1+\frac{M_{12}}{\sqrt{\phi}})}{\sqrt{\phi}}d\phi &< \infty, \\
\int_0^1 \frac{T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]\log(1+\frac{M_{12}}{\sqrt{\phi}})}{M_7^T M^{T-1}(T-1)^2\sqrt{\phi}}d\phi &< \infty, \\
\int_0^1 \frac{\log(1+\frac{M_9^{T-1}M^{T-1}}{\phi^{\frac{T}{2}}})\log(1+\frac{M_9 M}{\phi})}{\sqrt{\phi}}d\phi &< \infty, \\
\int_0^1 \frac{T[1+(T-1)\log(\frac{M_7 M}{\sqrt{\phi}})]}{M_7^T M^{T-1}(T-1)^2\sqrt{\phi}}d\phi &< \infty.
\end{aligned}$$

Therefore, the integral

$$\int_0^1\int_0^M\int_M^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \frac{1}{\phi^{3/2}}\frac{1}{1+|\sqrt{\frac{B}{\phi}}z+A|}e^{-\frac{z^2}{2}}\frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|}e^{-\frac{z'^2}{2}}dzdz'dcdc'd\phi < \infty.$$

6. Finally, we consider the integral in the region $c \geq M$, $c' < M$ and $\phi \geq 1$. Since c and c' are symmetric in the integral, if the the integral is finite in the region $c \geq M$, $c' < M$ and $\phi \geq 1$, then by the symmetric the integral is finite in the region $c < M$, $c' \geq M$ and $\phi \geq 1$ as well. Thus, in the following, we are going to prove that the integral in the region $c \geq M$, $c' < M$ and $\phi \geq 1$ is finite, i.e.

$$\begin{aligned}
&\int_1^\infty\int_0^M\int_M^\infty\int_{-\infty}^\infty\int_{-\infty}^\infty \frac{1}{\phi^{3/2}}\frac{1}{1+|\sqrt{\frac{B}{\phi}}z+A|}e^{-\frac{z^2}{2}}\frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|}e^{-\frac{z'^2}{2}}dzdz'dcdc'd\phi \\
\leq\ & 2M\int_M^\infty\int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B'}{\phi}}z'+A'|}e^{-\frac{z'^2}{2}}dz'dc' \\
\leq\ & O\left(\frac{M_8}{M_7}+\frac{1}{M_7^T M^{T-1}(T-1)}\right) \\
<\ & \infty.
\end{aligned}$$

Therefore, the lemma A.4 holds by summing up 8 different regions in the integral, where we dividing them up to 6 situations in the discussion.

□

Lemma A.5. *For $T \geq 2$,*

$$\int_0^\infty \int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B(c)}{\phi}}z + A(c)|} \exp\{-\frac{z^2}{2}\} dzdc < \infty.$$

Proof. As shown in Lemma A.4, let us define A and B in the similar way as during the proof of Lemma A.4, and we have

$$|h| = \frac{|A|}{\sqrt{\frac{B}{\phi}}} \leq M_K$$

where M_K is a constant. Then, following the definition of M_5, M_6, M_7 and M_8 in Lemma A.4, we have

$$\begin{aligned}
& \int_0^\infty \int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} \exp\{-\frac{z^2}{2}\} dzdc \\
= & \int_0^M \int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} \exp\{-\frac{z^2}{2}\} dzdc + \int_M^\infty \int_{-\infty}^\infty \frac{1}{1+|\sqrt{\frac{B}{\phi}}z + A|} \exp\{-\frac{z^2}{2}\} dzdc \\
\leq & M + \int_M^\infty \int_{|z+h|<1} \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} \exp\{-\frac{z^2}{2}\} dzdc \\
+ & \int_M^\infty \int_{|z+h|\geq 1} \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} \exp\{-\frac{z^2}{2}\} dzdc \\
\leq & M + M_5 \int_M^\infty \int_{|z+h|<1} \frac{1}{1+\sqrt{\frac{B}{\phi}}|z+h|} dzdc + M_6 \int_M^\infty \frac{1}{1+\sqrt{\frac{B}{\phi}}} \exp\{-\frac{z^2}{2}\} dzdc
\end{aligned}$$

$$
\begin{aligned}
&\leq O\left(1+\frac{M_8}{M_7}+\frac{1}{M_7}\int_{M_7M}^{\infty}\frac{1}{1+b^T}db\right)\\
&\leq O\left(\frac{M_8}{M_7}+\frac{1}{M_7^T M^{T-1}(T-1)}\right)\\
&< \infty,
\end{aligned}
$$

which completes the proof. □

Theorem A.6. *Suppose $n \geq 2$ and for $i = 1, \cdots, n$, $T_i \geq 2$ and $S_{i,t} \geq 2$ for at least two days $t \in \{1, \cdots, T_i\}$ with at least two of the tests on each of the two days having at least one 0 and one 1 observation. Then the posterior density of the DIR model is proper.*

Proof. In proving posterior propriety, it is easiest to work with the posterior density without the data augmentation, namely

$$
\begin{aligned}
&\pi(\theta, c, \tau, \eta, \epsilon, \phi \mid X)\\
\propto\ & \left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}V_{G_{j_i}}}\exp\left(-\frac{(\theta_{i,0}-\mu_{G_{j_i}})^2}{2V_{G_{j_i}}}\right)\mathbf{I}_{\{c_i\geq 0\}}\frac{1}{\tau_i^{3/2}}\frac{1}{\delta_i^{3/2}}\right\}\frac{1}{\phi^{3/2}}\\
\cdot\ & \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{l=1}^{K_{i,t,s}}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\epsilon_{i,t,s,l}^2}{2\sigma^2}\right)\right\}\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\delta_i}{2\pi}}\exp\left(-\frac{\delta_i\varphi_{i,t}^2}{2}\right)\right\}\\
\cdot\ & \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}(\frac{\tau_i}{2\pi})^{\frac{S_{i,t}-1}{2}}\exp\left(-\frac{\tau_i\eta_{i,t}^{*\prime}\Sigma_{i,t}^{-1}\eta_{i,t}^{*}}{2}\right)\right\}\\
\cdot\ & \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{l=1}^{K_{i,t,s}}\frac{\exp\left[X_{i,t,s,l}(\theta_{i,t}-a_{i,t,s}+\varphi_{i,t}+\eta_{i,t,s}+\epsilon_{i,t,s,l})\right]}{1+\exp(\theta_{i,t}-a_{i,t,s}+\varphi_{i,t}+\eta_{i,t,s}+\epsilon_{i,t,s,l})}I\{\eta_{i,t,S_{i,t}}=-\sum_{s=1}^{S_{i,t}-1}\eta_{i,t,s}\}\right\}\\
\cdot\ & \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi\{\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^{+}\}^2}{2\Delta_{i,t}}\right)\right\},
\end{aligned}
\tag{A.4}
$$

Noting that

$$\frac{\exp\left[X_{i,t,s,l}(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})\right]}{1 + \exp(\theta_{i,t} - a_{i,t,s} + \varphi_{i,t} + \eta_{i,t,s} + \epsilon_{i,t,s,l})} \leq 1\,,$$

an upper bound on the posterior density can be found by dropping all terms except the 0 and 1 test observations in the assumed tests for each individual. Utilizing Lemma A.1 for each pair of observations 0 and 1 then results in the following upper bound on the posterior density (A.4):

$$\begin{aligned}
&\frac{1}{\phi^{3/2}}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}V_{G_{j_i}}}\exp\left(-\frac{(\theta_{i,0}-\mu_{G_{j_i}})^2}{2V_{G_{j_i}}}\right)\mathbf{I}_{\{c_i\geq 0\}}\frac{1}{\tau_i^{3/2}}\frac{1}{\delta_i^{3/2}}\right\}\\
\cdot\quad &\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}\prod_{\ell=1}^{K_{i,t,s}}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\epsilon_{i,t,s,l}^2}{2\sigma^2}\right)\right\}\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\delta_i}{2\pi}}\exp\left(-\frac{\delta_i\varphi_{i,t}^2}{2}\right)\right\}\\
\cdot\quad &\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}(\frac{\tau_i}{2\pi})^{\frac{S_{i,t}-1}{2}}\exp\left(-\frac{\tau_i\eta_{i,t}^{*}{}'\Sigma_{i,t}^{-1}\eta_{i,t}^{*}}{2}\right)\right\}\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}I\{\eta_{i,t,S_{i,t}} = -\sum_{s=1}^{S_{i,t}-1}\eta_{i,t,s}\}\right\}\\
\cdot\quad &\left\{\prod_{i=1}^{n}\exp\left(-|\theta_{i,t_i}+\varphi_{i,t_i}+\eta_{i,t_i,m}| + |a_{i,t_i,m}| + |\epsilon_{i,t_i,m,k}| + |\epsilon_{i,t_i,m,k'}|\right)\right.\\
\cdot\quad &\exp\left(-|\theta_{i,t_i}+\varphi_{i,t_i}+\eta_{i,t_i,m'}| + |a_{i,t_i,m'}| + |\epsilon_{i,t_i,m',h}| + |\epsilon_{i,t_i,m',h'}|\right)\\
\cdot\quad &\left.\exp\left(-|\theta_{i,t_i'}+\varphi_{i,t_i'}+\eta_{i,t_i',r}| + |a_{i,t_i',r}| + |\epsilon_{i,t_i',r,q}| + |\epsilon_{i,t_i',r,q'}|\right)\right\}\\
\cdot\quad &\left.\exp\left(-|\theta_{i,t_i'}+\varphi_{i,t_i'}+\eta_{i,t_i',r'}| + |a_{i,t_i',r}| + |\epsilon_{i,t_i',r',g}| + |\epsilon_{i,t_i',r',g'}|\right)\right\}\\
\cdot\quad &\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi\{\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^{+}\}^2}{2\Delta_{i,t}}\right)\right\}\,.
\end{aligned} \tag{A.5}$$

Ignoring multiplicative constants, and integrating out all the $\epsilon_{i,t,s,l}$, (A.5) has an

upper bound of

$$
\begin{aligned}
&\frac{1}{\phi^{3/2}}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}V_{G_{j_i}}}\exp\left(-\frac{(\theta_{i,0}-\mu_{G_{j_i}})^2}{2V_{G_{j_i}}}\right)\mathbf{I}_{\{c_i\geq 0\}}\frac{1}{\tau_i^{3/2}}\frac{1}{\delta_i^{3/2}}\right\}\\
\cdot\quad&\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\delta_i}{2\pi}}\exp\left(-\frac{\delta_i\varphi_{i,t}^2}{2}\right)\right\}\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}(\frac{\tau_i}{2\pi})^{\frac{S_{i,t}-1}{2}}\exp\left(-\frac{\tau_i\eta_{i,t}^{*}{}'\Sigma_{i,t}^{-1}\eta_{i,t}^{*}}{2}\right)\right\}\\
\cdot\quad&\left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\prod_{s=1}^{S_{i,t}}I\{\eta_{i,t,S_{i,t}}=-\sum_{s=1}^{S_{i,t}-1}\eta_{i,t,s}\}\right\}\\
\cdot\quad&\left\{\prod_{i=1}^{n}\exp\{-|\theta_{i,t_i}+\varphi_{i,t_i}+\eta_{i,t_i,m}|\}\exp\{-|\theta_{i,t_i}+\varphi_{i,t_i}+\eta_{i,t_i,m'}|\}\right.\\
\cdot\quad&\left.\exp\{-|\theta_{i,t_i'}+\varphi_{i,t_i'}+\eta_{i,t_i',r}|\}\exp\{-|\theta_{i,t_{i'}}+\varphi_{i,t_{i'}}+\eta_{i,t_i',r'}|\}\right\}\\
\cdot\quad&\left\{\prod_{i=1}^{n}\prod_{t=1}^{T}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi\{\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^{+}\}^2}{2\Delta_{i,t}}\right)\right\}.
\end{aligned}
\tag{A.6}
$$

We only consider here the 'least information' case in which $S_{i,t_i}=S_{i,t_{i'}}=2$; the more general case can be done similarly. Then $\eta_{i,t_i,m}=-\eta_{i,t_i,m'}$, $\eta_{i,t_{i'},r}=-\eta_{i,t_{i'},r'}$, $\exp\left(-\tau_i\eta_{i,t_i}^{*}{}'\Sigma_{i,t_i}^{-1}\eta_{i,t_i}^{*}/2\right)=\exp\left(-\tau_i\eta_{i,t_i,m}^2\right)$, and $\exp\left(-\tau_i\eta_{i,t_{i'}}^{*}{}'\Sigma_{i,t_{i'}}^{-1}\eta_{i,t_{i'}}^{*}/2\right)=\exp\left(-\tau_i\eta_{i,t_{i'},r}^2\right)$. Using this in (A.6) and integrating out all other η except for $\eta_{i,t_i,m}$

and $\eta_{i,t_{i'},r}$ and all φ except for φ_{i,t_i} and φ_{i,t'_i}, results in the expression

$$\frac{1}{\phi^{3/2}}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}V_{G_{j_i}}}\exp\left(-\frac{(\theta_{i,0}-\mu_{G_{j_i}})^2}{2V_{G_{j_i}}}\right)\mathbf{I}_{\{c_i\geq 0\}}\right\}$$
$$\cdot\ \left\{\prod_{i=1}^{n}\frac{1}{\delta_i^{3/2}}\frac{\delta_i}{2\pi}\exp\left(-\frac{\delta_i\varphi_{i,t_i}^2}{2}\right)\exp\left(-\frac{\delta_i\varphi_{i,t'_i}^2}{2}\right)\cdot\frac{1}{\tau_i^{3/2}}\cdot\frac{\tau_i}{2\pi}\exp\left(-\tau_i(\eta_{i,t_i,m}^2+\eta_{i,t'_i,r}^2)\right)\right.$$
$$\cdot\ \exp\{-(|\theta_{i,t_i}+\varphi_{i,t_i}+\eta_{i,t_i,m}|+|\theta_{i,t_i}+\varphi_{i,t_i}-\eta_{i,t_i,m}|)\}$$
$$\cdot\ \left.\exp\{-(|\theta_{i,t'_i}+\varphi_{i,t'_i}+\eta_{i,t'_i,r}|+|\theta_{i,t'_i}+\varphi_{i,t'_i}-\eta_{i,t'_i,r}|)\}\right\}$$
$$\cdot\ \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi\{\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}\}^2}{2\Delta_{i,t}}\right)\right\}. \tag{A.7}$$

Next integrate out over τ_i, $\eta_{i,t_i,m}$ and $\eta_{i,t'_i,r}$ using Lemma A.2, resulting in the upper bound (again ignoring multiplicative constants)

$$\frac{1}{\phi^{3/2}}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}V_{G_{j_i}}}\exp\left(-\frac{(\theta_{i,0}-\mu_{G_{j_i}})^2}{2V_{G_{j_i}}}\right)\mathbf{I}_{\{c_i\geq 0\}}\right\}$$
$$\cdot\ \left\{\prod_{i=1}^{n}\frac{1}{\delta_i^{3/2}}\frac{\delta_i}{2\pi}\exp\left(-\frac{\delta_i\varphi_{i,t_i}^2}{2}\right)\exp\left(-\frac{\delta_i\varphi_{i,t'_i}^2}{2}\right)\exp\{-(|\theta_{i,t_i}+\varphi_{i,t_i}|+|\theta_{i,t'_i}+\varphi_{i,t'_i}|)\}\right\}$$
$$\cdot\ \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi\{\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}\}^2}{2\Delta_{i,t}}\right)\right\}. \tag{A.8}$$

After that integrate out δ_i, φ_{i,t_i} and φ_{i,t'_i} using Lemma A.3. The resulting upper bound on (A.8) is

$$\frac{1}{\phi^{3/2}}\left\{\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}V_{G_{j_i}}}\exp\left(-\frac{(\theta_{i,0}-\mu_{G_{j_i}})^2}{2V_{G_{j_i}}}\right)\mathbf{I}_{\{c_i\geq 0\}}\cdot\frac{1}{1+|\theta_{i,t'_i}|}\right\}$$
$$\cdot\ \left\{\prod_{i=1}^{n}\prod_{t=1}^{T_i}\sqrt{\frac{\phi}{2\pi\Delta_{i,t}}}\exp\left(-\frac{\phi\{\theta_{i,t}-\theta_{i,t-1}-c_i(1-\rho\theta_{i,t-1})\Delta_{i,t}^+\}^2}{2\Delta_{i,t}}\right)\right\}.$$

Integrating out all the $\theta_{i,t}$ except the θ_{i,t'_i} results in the expression

$$\frac{1}{\phi^{3/2}}\left\{\prod_{i=1}^{n}\mathbf{I}_{\{c_i\geq 0\}}\cdot\frac{1}{1+|\theta_{i,t'_i}|}\right\}$$

$$\cdot\quad\left\{\prod_{i=1}^{n}\sqrt{\frac{\phi}{2\pi V_{G_{j_i}}}}\left(\frac{1}{\sum_{t=1}^{t'_i}\Delta_{i,t}\prod_{i=t+1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})^2+\phi V_{G_{j_i}}\prod_{t=1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})^2}\right)^{1/2}\right.$$

$$\left.\cdot\quad\exp\left(-\frac{\phi(\theta_{i,t'_i}-\mu_{G_{j_i}}\prod_{t=1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})-\sum_{t=1}^{t'_i}c_i\Delta_{i,t}^{+}\prod_{i=t+1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+}))^2}{2(\sum_{t=1}^{t'_i}\Delta_{i,t}\prod_{i=t+1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})^2+\phi V_{G_{j_i}}\prod_{t=1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})^2)}\right)\right\}. \tag{A.9}$$

Finally, defining

$$\begin{aligned}
A_i(c_i) &= \mu_{G_{j_i}}\prod_{t=1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})+\sum_{t=1}^{t'_i}c_i\Delta_{i,t}^{+}\prod_{i=t+1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+}))^2,\\
B_i(c_i) &= \sum_{t=1}^{t'_i}\Delta_{i,t}\prod_{i=t+1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})^2+\phi V_{G_{j_i}}\prod_{t=1}^{t'_i}(1-c_i\rho\Delta_{i,t}^{+})^2,\\
z_i &= \frac{\sqrt{\phi}(\theta_{i,t'_i}-A_i(c_i))}{\sqrt{B_i(c_i)}},
\end{aligned}$$

using Lemma A.5 to integrate out all θ_{i,t'_i} and c_i, except for two individuals, and then using Lemma A.4 for the remaining variables, it follows that the integral is finite, completing the proof. □

Appendix B

Matrix Properties

Lemma B.1. *Suppose a matrix* $\mathbf{G}$ *can be partitioned into blocks as below*

$$\mathbf{G} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

then if $\mathbf{A}$ *is invertible,*

$$\mid \mathbf{G} \mid = \mid \mathbf{A} \mid \mid \mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B} \mid .$$

Lemma B.2. *Suppose a matrix* $\mathbf{G}_1$ *can be partitioned into blocks as below*

$$\mathbf{G}_1 = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix},$$

and if $\mathbf{G}_1$ *and* $\mathbf{D}$ *are invertible, then* $(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})$ *is invertible as well and we have*

$$\mathbf{G}_1^{-1} = \begin{pmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} + \mathbf{D}^{-1} \end{pmatrix}.$$

Lemma B.3. *Suppose* $\mathbf{A}$ *is an invertible square matrix and* $\mathbf{u}$, $\mathbf{v}$ *are vectors. Suppose furthermore that* $1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u} \neq 0$. *Then the Sherman-Morrison formula states that*

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}.$$

The general form shown here is the one published by Bartlett (1951).

Lemma B.4. *The Woodbury, Sherman & Morrison formula states that*

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B})^{-1}\mathbf{DA}^{-1},$$

assuming the relevant inverses all exist.

Lemma B.5. *Derivatives of the elements of an inverse matrix:*

$$\frac{\partial \mathbf{A}^{-1}}{\partial \theta} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \theta}\mathbf{A}^{-1},$$

where $\partial \mathbf{A}/\partial \theta$ is a matrix of elementwise derivatives.

Lemma B.6. *Derivatives of the logarithm determinant of a positive definite symmetric matrix are*

$$\frac{\partial}{\partial \theta}\log|\mathbf{A}| = \mathrm{Trace}(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \theta}).$$

Bibliography

Abramowitz, M. and Stegun, I. A. (1965), *Handbook of Mathematical Functions*, New York: Dover.

Adler, R. J. (1981), *The geometry of random fields*, Chichester:Wiley.

Akaike, H. (1974), "Markovian Representation of Stochastic Processes and Its Application to the Analysis of Autoregressive Moving Average Processes," *Annals of the Institute of Statistical Mathematics*, 26, 363–387.

Andersen, E. B. (1970), "Asymptotic Properties of Conditional Maximum-likelihood Estimates," 32, 282–301.

Andrews, D. F. and Mallows, C. L. (1974), "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 99–102.

Andrich, D. and Kreiner, S. (2010), "Quantifying Response Dependence Between Two Dichotomous Items Using the Rasch Model," *Applied Psychological Measurement*, 34, 181–192.

Bahl, L. R., Brown, P. F., De Souza, P. V., and Mercer, R. L. (1989), "A Tree-Based Statistical Language Model for Natural Language Speech Recognition," *IEEE Transactions on Acoustics Speech and Signal Processing*, 37, 1001–1008.

Banerjee, S., Gelfand, A., and Sirmans, C. (2003), "Directional Rates of Change Under Spatial Process Models," *Journal of the American Statistical Association*, 98, 946–954.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton: Chapman and Hall/CRC, 1 edn.

Bartlett, M. S. (1951), "An Inverse Matrix Adjustment Arising in Discriminant Analysis," *The Annals of Mathematical Statistics*, 22, 107–111.

Bayarri, M. J., Berger, J. O., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Lin, C.-H., and Tu, J. (2009), "Predicting Vehicle Crashworthiness: Validation of Computer Models for Functional and Hierarchical Data," *Journal of the American Statistical Association*, 104, 929–943.

Belson, W. A. (1959), "Matching and Prediction on the Principle of Biological Classification," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8, 65–75.

Berger, J. O. (2006), "The Case for Objective Bayesian Analysis," *Bayesian Analysis*, 1, 385–402.

Berry, D. A. (1990), "Subgroup Analyses," *Biometrics*, 46, 1227–1230.

Blight, B. and Ott, L. (1975), "A Bayesian Approach to Model Inadequacy for Polynomial Regression." *Biometrika*, 62, 79–88.

Bornjamp, B. (2009), "On Nonparametric Bayesian Analysis under Shape Constraints with Applications in Biostatistics," Ph.D. thesis, technische universität, Dortmund.

Bradlow, E., Wainer, H., and Wang, X. (1999), "A Bayesian Random Effects Model for Testlets," *Psychometrika*, 64, 153–168.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984), *Classification and Regression Trees*, Baco Raton: Chapman and Hall/CRC, 1 edn.

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992), "A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling," *Journal of the American Statistical Association*, 87, 493–500.

Chang, I.-S., Chien, L.-C., Hsiung, C. A., Wen, C.-C., and Wu, Y.-J. (2007), "Shape Restricted Regression with Random Bernstein Polynomials," *IMS Lecture Notes-Monograph Series*, 54, 187–202.

Chipman, H., George, E. I., and Mcculloch, R. E. (1998), "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935–960.

Clyde, M., Desimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011), "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101.

Cowell, R. G., Dawid, P. A., Lauritzen, S. L., and Spiegelhalter, D. J. (2003), *Probabilistic Networks and Expert Systems (Information Science and Statistics)*, New York: Springer.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Chichester: Wiley-Interscience, revised edition edn.

Cui, L., James Hung, H. M., Wang, S. J., and Tsong, Y. (2002), "Issues Related to Subgroup Analysis in Clincial Trials," *Journal of Biopharmaceutical Statistics*, 12, 347–358.

Darrell Bock, R. and Lieberman, M. (1970), "Fitting a Response Model for n Dichotomously Scored Items," *Psychometrika*, 35, 179–197.

De Boeck, P. (2008), "Random Item IRT Models," *Psychometrika*, 73, 533–559.

De'ath, G. and Fabricius, K. E. (2000), "Classification and Regression Trees: a Powerful Yet Simple Technique for Ecological Data Analysis," *Ecology*, 81, 3178–3192.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "A Bayesian CART Algorithm," *Biometrika*, 85, 363–377.

Dixon, D. O. and Simon, R. (1991), "Bayesian Subset Analysis," *Biometrics*, 47, 871–881.

Duncan, D. B. and Horn, S. D. (1972), "Linear Dynamic Recursive Estimation from the Viewpoint of Regresion Analysis," *Journal of the American Statistical Association*, 67, 815–821.

Embretson, S. (1991), "A Multidimensional Latent Trait Model for Measuring Learning and Change," *Psychometrika*, 56, 495–515.

Fahrmeir, L. (1992), "Posterior Mode Estimation by Extended Kalman Filtering for Multivariate Dynamic Generalized Linear Models," *Journal of the American Statistical Association*, 87, 501–509.

Fang, K.-T., Li, R., and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments (Computer Science and Data Analysis)*, Boca Raton: Chapman and Hall/CRC.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011), "Subgroup Identification from Randomized Clinical Trial Data," *Statistics in Medicine*, 30, 2867–2880.

Geman, D. and Jedynak, B. (1996), "An Active Testing Model for Tracking Roads in Satellite Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18, 1–14.

Gibbons, R. and Hedeker, D. (1992), "Full-information Item Bi-factor Analysis," *Psychometrika*, 57, 423–436.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Handcock, M. S. and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410.

Hanlon, S. T., Swartz, C. W., Stenner, A., Burdick, H., and Burdick, D. (2010), *Oasis Literacy Research Platform*, MetaMetrics, Inc., Software [V1].

Harrison, P. J. and Stevens, C. F. (1976), "Bayesian Forecasting," *Journal of the Royal Statistical Society: Series B (Methodological)*, 38, 205–247.

Hodges, J. S., Cui, Y., Sargent, D. J., and Carlin, B. P. (2007), "Smoothing Balanced Single-Error-Term Analysis of Variance," *Technometrics*, 49, 12–25.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401.

Huerta, G. and West, M. (1998), "Bayesian Inference for Unequally-spaced Time Series," in *JSM Proceedings, Section on Bayesian Statistical Science*, pp. 17–21, American Statistical Association, Alexandria, VA.

Jannarone, R. (1986), "Conjunctive Item Response Theory Kernels," *Psychometrika*, 51, 357–373.

Johnson, C. and Raudenbush, S. W. (2006), *A Repeated Measures, Multilevel Rasch Model With Application to Self-Reported Criminal Behavior*, pp. 131–164, New York: Routledge.

Jones, H. E., Ohlssen, D. I., Neuenschwander, B., Racine, A., and Branson, M. (2011), "Bayesian Models for Subgroup Analysis in Clinical Trials," *Clinical Trials*, 8, 129–143.

Jones, R. H. (1993), *Longitudinal Data with Serial Correlation: A State-Space Approach*, Boca Raton: Chapman and Hall/CRC, 1 edn.

Jørgensen, B., Christensen, S., Song, X., and Sun, L. (1999), "A State Space Model for Multivariate Longitudinal Count Data," *Biometrika*, 86, 169–181.

Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," *Trans. of the AMSE - Journal of Basic Engineering (Series D)*, 82, 35–45.

Kitagawa, G. (1987), "Non-Gaussian State-Space Modeling of Nonstationary Time Series," *Journal of the American Statistical Association*, 82, 1032–1041.

Kuss, M. (2006), "Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning," Ph.D. thesis, Technische Universität Darmstadt.

Lagakos, S. (2006), "The Challenge of Subgroup Analyses–Reporting without Distorting," *New England Journal of Medicine*, 354, 1667–1669.

Lavine, M. and Mockus, A. (1995), "A Nonparametric Bayes Method for Isotonic Regression," *Journal of Statistical Planning and Inference*, 46, 235–248.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., and Lu, C.-J. (2006), "Mining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines," *Computational Statistics and Data Analysis*, 50, 1113–1130.

Lemon, S., Roy, J., Clark, M., Friedman, P., and Rakowski, W. (2003), "Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression," *Annals of Behavioral Medicine*, 26, 172–181.

Liang, F. and Wong, W. H. (2000), "Evolutionary Monte Carlo: Applications to C_p Model Sampling and Change Point Problem," *Statistica Sinica*, 10, 317–342.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011), "Subgroup Identification Based on Differential Effect Search – a Recursive Partitioning Method for Establishing Response to Treatment in Patient Subpopulations," *Statistics in Medicine*, 30, 2601–2621.

Lord, F. (1953), "The Relation of Test Score to the Trait Underlying the Test," *Educational Psychology Measurement*, 13, 517–548.

Madigan, D., York, J., and Allard, D. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Marshall, R. J. (2001), "The Use of Classification and Regression Trees in Clinical Epidemiology," *Journal of Clinical Epidemiology*, 54, 603–609.

Martin, A. D. and Quinn, K. M. (2002), "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999," *Political Analysis*, 10, 134–153.

Marvelde, J. M. t., Glas, C. A. W., Landeghem, G. V., and Damme, J. V. (2006), "Application of Multidimensional Item Response Theory Models to Longitudinal Data," *Educational and Psychological Measurement*, 66, 5–34.

Matérn, B. (1986), *Spatial Variation*, Berlin: Springer-Verlag, 2 edn.

Morgan, J. N. and Sonquist, J. A. (1963), "Problems in the Analysis of Survey Data, and a Proposal," *Journal of the American Statistical Association*, 58, 415–434.

Neelon, B. and Dunson, D. (2004), "Bayesian Isotonic Regression and Trend Analysis," *Biometrics*, 60, 398–406.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J.-F. (2005), "Tree-structured Subgroup Analysis for Censored Survival Data: Validation of Computationally Inexpensive Model Selection Criteria," *Statistics and Computing*, 15, 231–239.

Nott, D. J. and Kohn, R. J. (2005), "Adaptive Sampling for Bayesian Variable Selection," *Biometrika*, 92, 747–763.

O'Hagan, A. and Kingman, J. F. C. (1978), "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society: Series B (Methodological)*, 40, 1–42.

Oliveira, V. D., Kedem, B., and Short, D. A. (1997), "Bayesian Prediction of Transformed Gaussian Random Fields," *Journal of the American Statistical Association*, 92, 1422–1433.

Ongaro, A. and Cattaneo, C. (2004), "Discrete Random Probability Measures: a General Framework for Nonparametric Bayesian Inference," *Statistics and Probability Letters*, 67, 33–45.

Park, J. H. (2011), *Modeling Preference Changes via a Hidden Markov Item Response Theory Model.*, pp. 479–491, Boca Raton: CRC Press.

Perron, F. and Mengersen, K. (2001), "Bayesian Nonparametric Modeling Using Mixtures of Triangular Distributions," *Biometrics*, 57, 518–528.

Petris, G., Petrone, S., and Campagnoli, P. (2009), *Dynamic Linear Models with R*, New York: Springer.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2002), "Subgroup Analysis, Covariate Adjustment and Baseline Comparisons in Clinical Trial Reporting: Current Practice and Problems," *Statistics in Medicine*, 21, 2917–2930.

Prado, R. and West, M. (2010), *Time Series: Modelling, Computation and Inference*, Boca Raton: Chapman and Hall/CRC Press.

Quiñonero Candela, J. and Rasmussen, C. E. (2005), "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, 6, 1939–1959.

Quinlan, J. R. (1986), "Induction of Decision Trees," *Machine Learning*, 1, 81–106.

Rasch, G. (1961), *On General Laws and the Meaning of Measurement in Psychology*, Danmarks pædagogiske Institut.

Rasmussen, C. E. and Williams, C. (2006), *Gaussian Processes for Machine Learning*, Cambridge: MIT Press.

Riihimäi, J. and Vehtari, A. (2010), "Gaussian Processes with Monotonicity Information," *Journal of Machine Learning Research - Proceedings Track*, 9, 645–652.

Roko, I. and Gilli, M. (2008), "Using Economic and Financial Information for Stock Selection," *Computational Management Science*, 5, 317–335.

Ruberg, S. J., Chen, L., and Wang, Y. (2010), "The Mean Does Not Mean as Much Anymore: Finding Sub-groups for Tailored Therapeutics," *Clinical Trials*, 7, 574–583.

Santner, T. J., B., W., and W., N. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer-Verlag.

Scott, J. G. and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *Annals of Statistics*, 38, 2587–2619.

Shi, J. Q. and Choi, T. (2011), *Gaussian Process Regression Analysis for Functional Data*, Boca Raton:Chapman and Hall/CRC Press.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009), "A Bayesian Approach to Non-parametric Monotone Function Estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, 71, 159–175.

Simon, R. (2002), "Bayesian Subset Analysis: Application to Studying Treatment-by-gender Interactions," *Statistics in Medicine*, 21, 2909–2916.

Sinharay, S., Johnson, M. S., and Williamson, D. M. (2003), "Calibrating Item Families and Summarizing the Results Using Family Expected Response Functions," *Journal of Educational and Behavioral Statistics*, 28, 295–313.

Sivaganesan, S., Laud, P. W., and Mller, P. (2011), "A Bayesian Subgroup Analysis with a Zero-enriched Polya Urn Scheme," *Statistics in Medicine*, 30, 312–323.

Solak, E., Murray-Smith, R., Leithead, W., Leith, D., and Rasmussen, C. (2003), "Derivative Observations in Gaussian Process Models of Dynamic Systems," in *Conference on Neural Information Processing Systems*, eds. S. Becker, S. Thrun, and K. Obermayer, Advances in neural information processing systems 15, pp. 1033–1040, MIT Press.

Song, P. X.-K. (2007), *Correlated Data Analysis: Modeling, Analytics, and Applications*, New York: Springer.

Stenner, A. J. (2010), "Using Technology to Merge Assessment and Instruction," Presented at 2nd International Conference for Teaching and Learning with Technology.

Stephenson, G. (2010), "Using Derivative Information in the Statistical Analysis of Computer Models," Ph.D. thesis, University of Southampton.

Stout, W. (1987), "A Nonparametric Approach for Assessing Latent Trait Unidimensionality," *Psychometrika*, 52, 589–617.

Strout, W. (1990), "A New Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation," *Psychometrika*, 55, 293–325.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008), "Interaction Trees with Censored Survival Data," *The International Journal of Biostatistics*, 4, 1–26.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., and Li, B. (2009), "Subgroup Analysis via Recursive Partitioning," *Journal of Machine Learning Research*, 10, 141–158.

Tan, E. S., Ambergen, A. W., Does, R. J. M. M., and Imbos, T. (1999), "Approximations of Normal IRT Models for Change," *Journal of Educational and Behavioral Statistics*, 24, 208–223.

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Temple, R. and Ellenberg, S. S. (2000), "Placebo-controlled Trials and Active-control Trials in the Evaluation of New Treatments. Part 1: Ethical and Scientific Issues," *Annals of Internal Medicine*, 133, 464–470.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), "Statistics in Medicine–Reporting of Subgroup Analyses in Clinical Trials," *New England Journal of Medicine*, 357, 2189–2194.

Wang, X. and Berger, J. O. (2011), "Estimating Shape Constrained Functions Using Gaussian Processes," in *JSM Proceedings, Section on Nonparametric Statistics*, pp. 5162–5171, American Statistical Association, Alexandria, VA.

Wang, X., Berger, J. O., and Burdick, D. S. (2012), "Bayesian Analysis of Dynamic Item Response Models in Educational Testing," Revision submitted to the Annals of Applied Statistics.

West, M. and Harrison, P. J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer Verlag, 2 edn.

Wilson, A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., Joellen, and Schildkraut, M. (2010), "Bayesian Model Search and Multilevel Inference for SNP Association Studies," *Annals of Applied Statistics*, 4, 1342–1364.

Woodcock, J. (2007), "The Prospects for "Personalized Medicine" in Drug Development and Drug Therapy," *Clinical Pharmacology and Therapeutics*, 81, 164–169.

Wu, Y., Tjelmeland, H., and West, M. (2007), "Bayesian CART: Prior Specification and Posterior Simulation," *Journal of Computational and Graphical Statistics*, 16, 44–66.

Xu, S., Jones, R. H., and Grunwald, G. K. (2007), "Analysis of Longitudinal Count Data with Serial Correlation." *Biometrical journal*, 49, 416–428.

Biography

Xiaojing Wang was born in July 18, 1983, in Yiyang, Hunan, P. R. China. In July 2012, she will be awarded Ph.D. in Statistics and M.A in Economics as scheduled. Before these, she was conferred M.S. in Probability and Mathematical Statistics in July 2008 from Institute of Applied Mathematics, Academy of Mathematics and Systems Science at Chinese Academy of Sciences and B.S. in Information and Computing Science in Hunan University in June 2005. She gained several awards and fellowships, such as Graduate Fellowship from SAMSI and Duke University, Excellent Master's Thesis in Chinese Academy of Sciences in 2008 and Outstanding University Student in Hunan Province in 2004.

Her publication and submitted papers cover "Estimating Shape Constrained Functions Using Gaussian Processes" with James O. Berger (JSM proceedings, 2011); "Bayesian Analysis of Dynamic Item Response Models in Educational Testing" with James O. Berger and Donald S. Burdick (revision submitted to Annals of Applied Statistics, 2012); "Wavelet Analysis of Change-points in a Non-parametric Regression with Heteroscedastic Variance" with Yong Zhou and others (Journal of Econometrics, 2010); " Statistical Analysis of Cellular Aggregates in Immunofluorescence Histology" with Ioanna Manolopoulou, Mike West and etc. (Discussion Paper, the Department of Statistical Science, Duke University, 2009); "Estimating Equations Inference with Missing Data" with Zhou Yong and Alan Wan (Journal of the American Statistical Association, 2008).