CONTRIBUTIONS TO MODELING PARASITE DYNAMICS AND DIMENSION REDUCTION

BY

NA CUI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Yuguo Chen, Chair, Co-Director of Research
Professor Feng Liang, Co-Director of Research
Professor John I. Marden
Professor Annie Qu

Abstract

For my thesis, I have worked on two projects: modeling parasite dynamics (Chapter 2) and complementary dimensionality analysis (Chapter 3).

In the first project, we study a longitudinal data of infection with the parasite *Giardia lamblia* among children in Kenya. Understanding the infection and recovery rate from parasitic infections is valuable for public health planning. Two challenges in modeling these rates are (1) infection status is only observed at discrete times even though infection and recovery take place in continuous time and (2) detectability of infection is imperfect. We address these issues through a Bayesian hierarchical model based on a random effects Weibull distribution. The model incorporates heterogeneity of the infection and recovery rate among individuals and allows for imperfect detectability. We estimate the model by a Markov chain Monte Carlo algorithm with data augmentation. We present simulation studies and an application to an infection study about the parasite Giardia lamblia among children in Kenya.

The second project focuses on supervised dimension reduction. The goal of supervised dimension reduction (SDR) is to find a compact yet informative representation of the original data space via some transformation. Most SDR algorithms are formulated as an optimization problem with the objective being a linear function of the second order statistics of the data. However, such an objective function tends to overemphasize those directions already achieving large between-class distances yet making little improvement over the classification accuracy. To address this issue, we introduce two objective functions, which are directly linked to the classification accuracy, then present an algorithm that sequentially solves the nonlinear objective functions.

To My Family.

Acknowledgments

This dissertation would not have been possible without the support of many people. I must first express my deepest gratitude to my advisers, Professor Yuguo Chen and Professor Feng Liang, for all their guidance, support, and invaluable suggestions. They helped me go through difficult times with constant encouragement and influential discussions. I am also very grateful to my committee members: Professor John I. Marden and Professor Annie Qu, who have generously given their time and insightful suggestions.

I thank all the faculties, staff, graduate students and aluminies in the Department of Statistics. My sincere thanks goes to Professor Adam T. Martinsek, Professor Annie Qu and Dr. Maria Muyot for offering me the opportunity to work in Illinois Statistics Office and leading me working on diverse exciting projects. I have benefited a lot from this wonderful job. A very special thank you to Professor Jeffrey A. Douglas for his generous help all the time. Many thanks must also go to Yang Feng, Peng Wang, Jing Xia, Yunwen Yang, Zhi He, Juan Shen, Xianyang Zhang, Bin Li, Yufei Liu, Lu Gan, and Wei Sun for their support and friendship.

Last but not the least, I would like to thank my family for their understanding and patience and in particular, I must acknowledge my husband, without whose love and encouragement, I would not be the person I am today.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to Modeling Parasite Dynamics

Longitudinal studies play an important role in many areas, such as psychology and sociology. In this thesis, we focus on a longitudinal study of *Giardia lamblia* descried by Chunge (1989). Giardia lamblia, an intestinal parasite, is the most common cause of parasitic gastrointestinal disease and is especially prevalence in young children. The background information on this parasite can be found in Svärd et al. (1998), Hetsko et al. (1998), Warrell et al. (2003), and Huang and White (2006). In the study, eighty-four children were chosen and the stools of children were examined for the presence of *Giardia* every week. The weekly testing result of each child was recorded as 1 if the parasite was found and 0 if the parasite was not found.

Understanding the infection and recovery rate from parasitic infections is valuable for public health planning. There are three main challenges in modeling this type of data. First, the disease is often imperfectly detected due to imperfect diagnostic instruments and procedures. Therefore, the recorded data may not be consistent with the real infected status. Second, the transition rates may change over time. For example, it is possible that an individual may have high immunity shortly after the clearance of an infection but the immunity wanes over time. Third, there is evidence that individuals are heterogeneous in their infection probabilities and dynamics for many parasitic diseases (Woolhouse et al., 1997). Such realistic situations complicate the study of malaria dynamics.

In the literature, many methods have been proposed to deal with this type of data. For example, Bekessy et al. (1976) proposed a first-order Markov model to study the dynam-

ics of malaria in Garki, Nigeria, without considering the above three issues. Nagelkerke et al. (1990) generalized Bekessy's model to the situation with imperfect detectability. Ng and Cook (1997) introduced a mixed continuous-time two-state process that accommodates the heterogeneity among individuals by the bivariate log-normal distribution. Cook (1999) adopted the exponential survival function with random effects to deal with heterogeneity among individuals. Smith and Vounatsou (2003) and Rosychuk et al. (2009) developed a hidden two-state Markov model with imperfect detectability. Crespi et al. (2005) developed Markov and semi-Markov models describing recurrence and time-inhomogeneous transition rates.

However, none of these methods considers all the three issues together. Thus, in this project, we aim to address the non-homogeneous transition rates, imperfect detectability and heterogeneity between individuals simultaneously. We propose a Bayesian hierarchical model based on a random effects Weibull distribution. The key steps are that we introduce a latent process to denote the true parasite dynamics and further model the latent process as a non-Markov continuous stochastic process. To estimate the parameters in such a model, we propose a Markov chain Monte Carlo algorithm with data augmentation for full Bayesian inference.

1.2 Introduction to Complementary Dimension Analysis

Dimension reduction techniques are very crucial in high-dimensional data analysis. One reason is that many statistical methods are designed for low-dimensional data. Take the simple linear regression model for example, as the number of the covariates increases, the standard error of the predicted value will be accumulated. Another reason is that the computational cost for high-dimensional data analysis is very high. In order to handle real-word data adequately, it is of primary interest in many applications to perform a dimension

dimension first.

Due to its importance, many dimension reduction techniques have been proposed in the literature. Depending on whether or not the response information is considered, dimension reduction algorithms can be categorized into supervised and unsupervised ones. In this project, we focus on supervised dimension reduction (SDR) whose goal is to find a compact yet informative representation of the original data space via some transformation. In the setting of K-class classification problem, the objective of SDR is to transform the p-dimensional feature space to a lower m-dimensional space while keeping the most discriminative information. The most well-known method in this category is Fisher Discriminant Analysis (FDA) where a transformation matrix is determined by maximizing the Fisher criterion defined as the between-class over the within-class scatter matrices. The objective function is constructed based on a linear function of the second order statistics (L_2 norm) of the data, which is shared by most SDR algorithms. An advantage of using this kind of objective functions is that the solution is in closed form and can be solved easily by eigen-decomposition.

However, Loog et al. (2001) pointed out that the objective function of FDA is suboptimal when dealing with multi-class problem as it overemphasizes large class distances. Furthermore, as illustrated in our toy example in Section 3.1, such an objective function tends to overemphasize those directions already achieving large between-class distances yet making little improvement over the classification accuracy. In other words, such methods may overlook the directions leading a small margin of between-class distances but a big improvement over the classification accuracy. In summary, the sub-optimality of FDA for multi-class classification is due to the discrepancy between the objective function and the classification accuracy: classification accuracy does not increase *linearly* with respect to the between-class distance.

To address this issue, we introduce two objective functions, which are directly linked to the classification accuracy: one for parametric case and the other for non-parametric case. We further provide a general objective function which not only accommodates these two

particular measures, but also incorporates many existing SDR methods as special cases. Then the reduced subspace is directly guided by maximizing the accuracy of classification which is performed in this subspace. The challenge here is that the objective function may take a nonlinear function of the L_2 norm of the data. Therefore, we cannot apply eigen-decomposition directly in this situation. In this thesis, we present an algorithm that sequentially solves the nonlinear objective functions. The key motivation of this algorithm is that each sequentially added direction should boost the discriminative power of the reduced space. This is why we term our new algorithm as *Complementary* Dimension Analysis. We evaluate the performance of our algorithm on several simulated datasets and real world datasets.

Chapter 2

Modeling Parasite Infection Dynamics When There is Heterogeneity and Imperfect Detectability

2.1 Introduction

Parasitic infections in humans are often characterized by repeated infections and clearance of parasites; examples include malaria and Giardia lamblia. To understand this dynamic behavior of parasitic infections, repeated observations of the infection and recovery status in the same group of individuals are required. This type of longitudinal data is often modeled as a two-state stochastic process. For example, Bekessy et al. (1976) proposed a first-order Markov model to study the dynamics of malaria in Garki, Nigeria. Under the assumptions of constant transition rates and perfect detectability, the model can be easily fitted by the maximum likelihood method where the results are always based solely on raw counts of infected and uninfected observations.

However, these assumptions are often not satisfied in real world situations for three main reasons. First, the disease is often imperfectly detected due to imperfect diagnostic instruments and procedures. For example, false negatives may be common in the detection of Giardia lamblia by stool samples because even if a person harbors Giardia lambdlia, these parasites may not be excreted in every stool sample (Nagelkerke et al., 1990). Second, the pattern of transitions may not correspond to the first-order Markov model. For example, it is possible that an individual may have high immunity shortly after the clearance of an infection but the immunity wanes over time. Third, there is evidence that individuals are heterogeneous in their infection probabilities and dynamics for many parasitic diseases (Woolhouse et al., 1997). The assumption of homogeneous infection rates without identifying

those individuals who are frequently infected from others will introduce biased estimates of the transition rates. Such realistic situations complicate the study of malaria dynamics.

To address the above issues, many Markov and semi-Markov approaches and corresponding solutions have been proposed. Nagelkerke et al. (1990) generalized Bekessy's model to the situation with imperfect detectability, and used numerical maximization of the partial likelihood to estimate the transition rates and rate of detectability. Ng and Cook (1997) introduced a mixed continuous-time two-state process that accommodates the heterogeneity among individuals by the bivariate log-normal distribution. They estimated the model by maximizing the approximated likelihood through numerical methods. Cook (1999) adopted the exponential survival function with random effects to deal with heterogeneity among individuals. Smith and Vounatsou (2003) and Rosychuk et al. (2009) developed a hidden two-state Markov model with imperfect detectability and designed a Markov chain Monte Carlo (MCMC) algorithm to estimate the parameters. Crespi et al. (2005) developed Markov and semi-Markov models describing recurrence and time-inhomogeneous transition rates.

However, none of the above methods considers simultaneously non-homogeneous transition rates, imperfect detectability and heterogeneity between individuals. This is partly due to the numerical challenges in obtaining the marginal likelihood function. Here, we propose a Bayesian hierarchical model based on a random effects Weibull model to explicitly address these difficulties. The model assumes that the true parasite dynamics is a continuous two-state stochastic process which is hidden due to the discrete observations and imperfect detection. Given the latent process, the observed data is assumed to be conditionally independent of each other. We further model the latent process as a non-Markov continuous stochastic process based on a Weibull survival function where the parameters of the Weibull have a random distribution across individuals; the Weibull distribution allows for more flexible transition rates than the constant transition rates of an exponential distribution, and the randomness of the Weibull parameters across individuals allows for heterogeneity among individuals. To estimate the parameters in such a model, we propose a MCMC algorithm

with data augmentation for full Bayesian inference, and construct a series of efficient moves to explore the space of the latent process. In the simulation study, we show the performance of the proposed method for models with different settings. We also apply the proposed method to the infection study of Giardia lamblia of children in Kenya (Chunge, 1989).

The rest of this chapter is organized as follows. In Section 2.2, we show the longitudinal data of *Giardia lamblia* (Chunge, 1989). The statistical model is described in Section 2.3 and the proposed MCMC algorithm is presented in Section 2.4. Next, in Section 2.5, we evaluate the performance of our method under different settings for both simulated data sets and real data set. The conclusion is concluded in Section 2.6.

2.2 Data

In this thesis, we focus on a longitudinal study of *Giardia lamblia* descried by Chunge (1989). Giardia lamblia, an intestinal parasite, is the most common cause of parasitic gastrointestinal disease and is especially prevalence in young children. The background information on this parasite can be found in Svärd et al. (1998), Hetsko et al. (1998), Warrell et al. (2003), and Huang and White (2006). In the study, eighty-four children were chosen and the stools of children were examined for the presence of *Giardia* every week. The weekly testing result of each child was recorded as 1 if the parasite was found and 0 if the parasite was not found. At the end of 44 consecutive weeks, 58 children with 10 to 44 consecutive weekly registration were selected to form the data set. The data are presented in Table 2.1.

2.3 Statistical Model

The observed data is the weekly recorded status of the presence or absence of the parasite for n individuals. Let $X_{ij} = 1$ if individual i is infected by the parasite at the jth week and $X_{ij} = 0$ if the parasite is not detected ($i = 1, \cdots, n;\ j = 1, \cdots, n_i$). Although the data is recorded at discrete times, the actual infection and recovery take place in continuous time.

ID	Infection status (negative 0, positive 1)
1	1 0 1 1 0 0 1 1 1 1 1 0
2	0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 1
3	1 1 0 0 0 0 0 0 0 0
4	0 0 0 1 1 1 1 1 1 1 1
5	1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1
6	0 0 1 0 1 0 0 0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0
7	1 1 1 1 1 0 0 1 1 1 1 1
8	1 0 1 0 1 1 1 0 0 1 0 0
9	0 0 0 0 0 0 0 0 0 0 0
10	1 1 1 1 1 1 1 1 1 0 1 1 0
11	0 1 1 0 0 0 0 0 1 0 0 0 0 0
12	0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
13	1 1 1 1 1 0 0 0 0 0 0 1 1 1 0 1 1
14	1 0 0 0 0 1 1 0 0 1
15	0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1
16	1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1
17	1 1 0 0 1 1 1 1 1 0 0 0 0 1 1 1 1
18	0 0 0 0 0 0 0 0 1 1 1 1
19	0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
20	0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 1 0 0 1 1 1 1 1 0
21	0 0 0 0 0 0 0 0 0 0 0
22	1 1 1 1 1 1 1 1 1 1 1 1 1
23	1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0
24	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
25	1 0 1 1 0 0 1 1 1 1 1
26	1 1 1 1 1 1 1 0 1 1 1 1 1
27	0 0 0 0 0 0 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0
28	0 0 1 1 1 1 1 1 1 1 1 1 0 1
29	1 0 0 0 1 0 0 0 1 1 1
30	0 1 0 0 0 0 0 0 0 0 0 0
31	0 0 0 0 0 1 1 0 1 1
32	1 1 1 1 1 1 1 1 0 1
33	0 1 1 0 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1
34	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
35	0 0 0 0 0 1 1 1 0 0 0 0 0 1
36	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
37	0 0 0 0 1 0 0 1 0 0 0 1
38	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
39	0 1 1 1 1 1 1 0 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 0 1 1 1 0 0 0 1 0 0 0 1 1 1 0 0
40	0 0 1 1 1 1 0 1 1 0
41	1 1 1 1 1 0 1 1 0 0 1 1
42	1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1
43	1 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1 1 1 0 0 0 0 0 1 1 0 0
44	0 0 0 0 0 0 0 0 0 0 0 0
45	0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1
46	1 0 0 0 0 1 1 0 1 1 1 1 0 0 1 1
47	1 1 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1
48	0 0 0 0 1 1 1 1 1 1 1 0 1 1
49	1 1 1 1 1 1 1 1 1 1 1
50	1 1 1 1 1 1 1 1 1 0 1 0 0 0 1 1 0 1 1 0 0 1 1 1 1 0 0 0 0 1
51	0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1
52	1 1 1 1 0 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 0 0
53	0 0 1 1 1 1 1 1 1 1 1
54	0 0 0 0 1 1 1 1 1 1
55	0 0 1 0 0 0 0 0 0 0 0 1 0
56	0 1 0 0 0 0 0 0 0 0 1 0
57	1 1 1 1 1 1 0 0 0 0 1 0 0 0
58	0 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0

Table 2.1: Weekly infection status of *Giardia lamblia* on children from Kenya.

Let $Z_i(t)$, $1 \leq t \leq n_i$, denote the hidden continuous true presence or absence of the parasite for individual i at time t. The underlying true status $Z_i(t)$ also takes value 1 or 0 depending on whether individual i at time t is infected or not. See Figure 2.1 for an illustration. The primary goal is to model the joint distribution of observed data $\mathbf{X}= (X_{ij},\ i = 1, \cdots, n;\ j = 1, \cdots, n_i)$ and latent process $\mathbf{Z}= (Z_i(t),\ i = 1, \cdots, n;\ 1 \leq t \leq n_i)$. We will first address the relation between $\mathbf{X}$ and $\mathbf{Z}$ and then discuss the modeling for $\mathbf{Z}$ in the following sections.

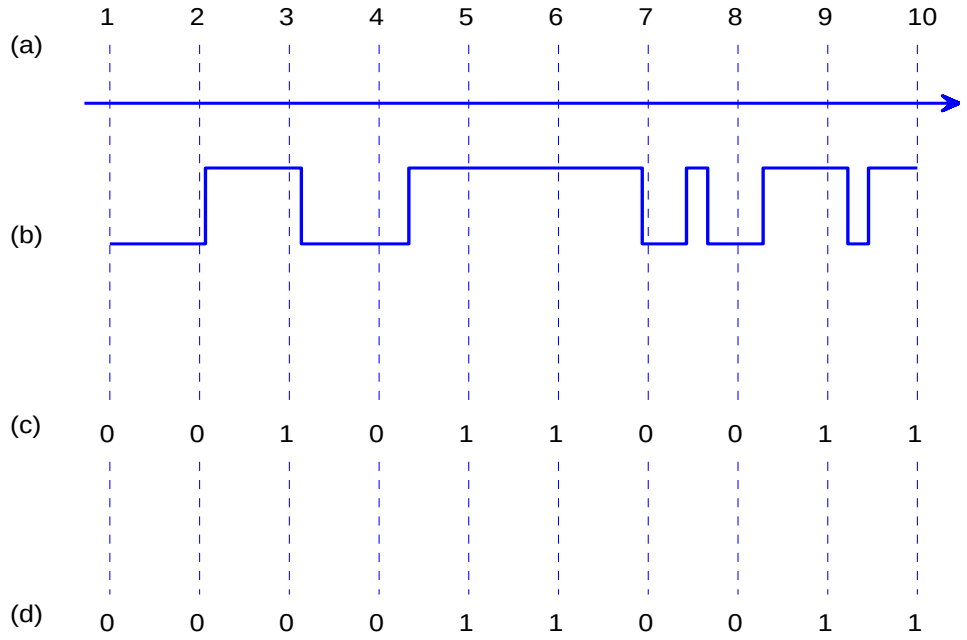


Figure 2.1: Relation between observed values and the underlying latent process. (a) Time index from 1 to 10; (b) The continuous latent process $Z_i(t)$ $(1 \leq t \leq 10)$; (c) The value of $Z_i(t)$ at discrete time points $Z_i(j)$ $(j = 1, \cdots, 10)$; (d) The corresponding observed value at discrete time points, X_{ij} $(j = 1, \cdots, 10)$.

2.3.1 Imperfect detectability

Under the perfect detectability assumption, we have $X_{ij} = Z_i(j)$ for all $i = 1, \cdots, n$ and $j = 1, \cdots, n_i$. If the diagnostic procedure is imperfect, the observation X_{ij} may be different from the true value $Z_i(j)$, which leads to misclassification of X_{ij}. The example shown

in Figure 2.1 has a misclassification at time 3. There are two types of misclassification: imperfect sensitivity ($Z_i(j) = 1,\ X_{ij} = 0$) and imperfect specificity ($Z_i(j) = 0,\ X_{ij} = 1$).

To deal with the two-state data with misclassification, several approaches have been proposed. Nagelkerke et al. (1990) considered the imperfect sensitivity in the continuous-time Markov model. Bureau et al. (2003) and Rosychuk et al. (2009) incorporated both types of misclassification. A more detailed review dealing with misclassification issues can be found in Ji and Fan (2009).

Following Nagelkerke et al. (1990) who also analyzed the *Giardia lamblia* infection in Kenya children, we assume the measurement has perfect specificity and imperfect sensitivity. Moreover, the observed values are assumed to be conditionally independent of each other given the true process. Let $1 - p$ denote the probability of false negatives, then the relation between X_{ij} and $Z_i(j)$ can be parameterized as

$$P\left(X_{ij} = 1 | Z_i(j) = 1\right) = p$$

and

$$P\left(X_{ij} = 0 | Z_i(j) = 0\right) = 1.$$

The probability of observing $\mathbf{X}$ given $\mathbf{Z}$ is

$$p(\mathbf{X}|\mathbf{Z}, p) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} \left[Z_i(j)\left(pX_{ij} + (1-p)(1-X_{ij})\right) + \left(1 - Z_i(j)\right)\left(1 - X_{ij}\right)\right]. \tag{2.1}$$

2.3.2 Latent process

The parasite dynamics are dependent on the modeling of the latent process. Through the modeling, we are trying to learn the transition rate $h_{0i}(t)$ from uninfected state to infected state and $h_{1i}(t)$ from infected state to uninfected state for individual i at time t, where $h_{0i}(t)$

and $h_{1i}(t)$ are defined by

$$h_{0i}(t) = \lim_{\Delta t \to 0} \frac{P(Z_i(t+\Delta t)=1 | Z_i(t)=0)}{\Delta t} \tag{2.2}$$

$$h_{1i}(t) = \lim_{\Delta t \to 0} \frac{P(Z_i(t+\Delta t)=0 | Z_i(t)=1)}{\Delta t}. \tag{2.3}$$

Discrete and continuous time Markov models have been widely used to model the latent process $Z_i(t)$ by assuming constant hazard rates over time for all individuals. However, the Markov assumption with the same transition rate for all individuals may not be appropriate for at least two reasons. First, the transition rate may depend on the duration that the individual has been in a state. For example, an individual may have a high immunity shortly after clearance of an infection. Second, the transition rates may vary considerably between individuals (Woolhouse et al., 1997). Also, individuals have different levels of immunity. Some individuals could have frequent transitions; while other individuals have relatively inactive transitions. Therefore, to specify a more realistic stochastic process, we consider the distribution for the transition rates with the flexibility to be time dependent and to incorporate the heterogeneity.

Here we adopt the random effects Weibull distribution to model the duration time and include the between-individual variation (Butler and Worrall, 1985; Morris and Christiansen, 1995; Sohn et al., 2007, 2006). Let h_{si} denote the hazard rate for individual i at state s ($s = 0, 1$). We have

$$h_{si}(t) = u_{si}\, h_s^0(t), \quad t \geq 0, \tag{2.4}$$

where u_{si} denotes the random effect introduced by between-individual variation and $h_s^0(t)$ is the baseline Weibull hazard function which is defined as

$$h_s^0(t) = \alpha_s^{-\beta_s} \beta_s\, t^{\beta_s - 1}, \quad t \geq 0. \tag{2.5}$$

Here $\alpha_s > 0$ is the scale parameter and $\beta_s > 0$ is the shape parameter of the distribution.

This is a versatile family of distributions that can take on many shapes based on the value of β_s. For example, this distribution is degenerated to exponential distribution when $\beta_s = 1$. The corresponding probability density function of the hazard function in (2.4) is

$$f_{si}(t) = u_{si}\ \alpha_s^{-\beta_s}\beta_s\ t^{\beta_s - 1}\mathrm{e}^{-u_{si}\alpha_s^{-\beta_s}t^{\beta_s}}\ ,\ \ t \geq 0. \tag{2.6}$$

For the random effects u_{0i} and u_{1i}, gamma distribution with mean 1 and an unknown variance is used quite often for its conjugacy property. However, the possible correlation between u_{0i} and u_{1i} is not specified under this situation. We model the random effects u_{0i} and u_{1i} jointly as a bivariate log-normal distribution. Let $U_i = (u_{0i}, u_{1i})^T$, $i = 1, \cdots, n$. We then assume U_i's are independent and identically distributed as:

$$\log\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \ \sim\ \text{Normal}\left(\begin{pmatrix} -0.5\sigma_0^2 \\ -0.5\sigma_1^2 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \tau\sigma_0\sigma_1 \\ \tau\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix}\right)$$

where σ_0 and σ_1 are non-negative unknown parameters and τ is the correlation coefficient between $\log(u_{0i})$ and $\log(u_{1i})$. The mean is chosen to make the expectation of the random effects equal to 1. Larger values of σ_s^2, $s = 0, 1$, correspond to greater heterogeneity of individuals and positive or negative correlation between the logarithm of random effects is determined by the sign of τ.

2.3.3 The likelihood

Let $t_i = (t_{i,1}, \cdots, t_{i,m_i})$ denote the times that individual i changes its state in the time interval from 1 to n_i, where m_i denotes the total number of transitions. Then the latent process $Z_i(t)$, $1 \leq t \leq n_i$, can be represented by the transition time points t_i and its initial state $Z_i(1)$. Assume the latent process Z_i started at $-\infty$ and is in equilibrium at time 1.

Then the density function of the first observed transition given $Z_i(1)$ is

$$p(t_{i,1}|Z_i(1)=s)=\frac{P_{si}(t>t_{i,1}-1)}{\mu_{si}}, \tag{2.7}$$

where $\mu_{si}=\int_0^\infty tf_{si}(t)dt$ denotes the expected duration time in state s for individual i (Cox and Isham, 1980). Let $\pi_i(s)=P(Z_i(1)=s)=\mu_{si}/(\mu_{0i}+\mu_{1i})$. Then the density function of Z_i with $Z_i(1)=s$ is

$$\begin{aligned} p(Z_i|\alpha_0,\alpha_1,\beta_0,\beta_1,u_{si}) &= \pi_i(s)\frac{P_{si}(t>t_{i,1}-1)}{\mu_{si}} \\ &\quad\times\left[\prod_{k=1}^{m_i-1} f_{a(s,k),i}(t_{i,k+1}-t_{i,k})\right] P_{a(s,m_i),i}(t>n_i-t_{i,m_i}), \end{aligned} \tag{2.8}$$

where $a(s,k)$ equals $1-s$ if k is odd and s if k is even. The likelihood for the complete data $(\mathbf{X},\mathbf{Z})$ is

$$p(\mathbf{X},\mathbf{Z}|p,\alpha_0,\alpha_1,\beta_0,\beta_1,\mathbf{U})=\prod_{i=1}^{n}p(Z_i|\alpha_0,\alpha_1,\beta_0,\beta_1,U_i)\prod_{j=1}^{n_i}p(X_{ij}|Z_i(j),p), \tag{2.9}$$

where $\mathbf{U}=(U_i,\ i=1,\cdots,n)$.

2.3.4 Prior and posterior distributions

Following the Bayesian framework, we specify the prior distribution for all the parameters $\Theta=(\alpha_0,\alpha_1,\beta_0,\beta_1,\sigma_1^2,\sigma_2^2,\tau,p)$ as follows:

$$\begin{aligned} p &\sim \text{Beta}(\gamma_1^p,\gamma_2^p) \\ \alpha_s &\sim \text{Gamma}(\gamma_1^\alpha,\gamma_2^\alpha),\ s=0,1 \\ \beta_s &\sim \text{Gamma}(\gamma_1^\beta,\gamma_2^\beta),\ s=0,1 \\ \begin{pmatrix} \sigma_0^2 & \tau\sigma_0\sigma_1 \\ \tau\sigma_0\sigma_1 & \sigma_1^2 \end{pmatrix} &\sim \text{Inverse-Wishart}(W,v), \end{aligned}$$

where all the γ's take positive values, W is a 2×2 positive definite matrix, and $v > 1$ is the degree of freedom. Then, the posterior distribution of interest is

$$\begin{aligned} p(\Theta, \mathbf{Z}, \mathbf{U}|\mathbf{X}) &\propto p(\mathbf{X}|\mathbf{Z}, \Theta, \mathbf{U}) p(\mathbf{Z}|\Theta, \mathbf{U}) p(\mathbf{U}|\Theta) p(\Theta) \\ &\propto p(\mathbf{X}|\mathbf{Z}, p) p(\mathbf{Z}|\alpha_0, \alpha_1, \beta_0, \beta_1, \mathbf{U}) p(\mathbf{U}|\sigma_0^2, \sigma_1^2, \tau) p(\Theta). \end{aligned} \tag{2.10}$$

2.4 Markov Chain Monte Carlo Algorithm

In this section we discuss the MCMC algorithm for estimating the parameters in the hierarchical Bayesian model. As the posterior distribution in (2.10) is too difficult to sample directly, we use Metropolis-within-Gibbs algorithm (Geman and Geman, 1984; Hastings, 1970; Metropolis et al., 1953) to sample from the conditional distribution of each variable. We divide the parameters Θ, the latent process $\mathbf{Z}$, and the random effects $\mathbf{U}$ into five groups: $L_1 = (\alpha_0, \alpha_1, \beta_0, \beta_1)$, $L_2 = p$, $L_3 = (\sigma_0^2, \sigma_1^2, \tau)$, $L_4 = \mathbf{U}$, and $L_5 = \mathbf{Z}$. A sketch of the Metropolis-within-Gibbs algorithm is given in Algorithm 1, where M is the number of Markov chain iterations and $L_{i:j}^{(t)} = (L_i^{(t)}, L_{i+1}^{(t)}, \cdots, L_j^{(t)})$.

Algorithm 1 Metropolis-within-Gibbs algortihm

Initialize all the parameters and variables: $L_1^{(0)}, L_2^{(0)}, L_3^{(0)}, L_4^{(0)}$, and $L_5^{(0)}$.
for $t = 1$ to M **do**
  for $i = 1$ to 5 **do**
    Given $L_{1:i-1}^{(t)}$, $L_{i+1:5}^{(t-1)}$, generate a sample $L_i^\star$ from a proposal distribution $q_i(L_i|L_i^{t-1})$. Let

$$L_i^{(t)} = \begin{cases} L_i^\star, & \text{with probability } r_i \\ L_i^{(t-1)}, & \text{otherwise,} \end{cases} \tag{2.11}$$

    where

$$r_i = \min\left\{ \frac{p(L_i^\star|L_{1:i-1}^{(t)}, L_{i+1:5}^{(t-1)}) q_i(L_i^{t-1}|L_i^\star)}{p(L_i^{t-1}|L_{1:i-1}^{(t)}, L_{i+1:5}^{(t-1)}) q_i(L_i^\star|L_i^{t-1})}, 1 \right\} \tag{2.12}$$

  end for
end for

Here are more details of the algorithm.

Initialization: Assign arbitrary initial values for the parameters $L_1 = (\alpha_0, \alpha_1, \beta_0, \beta_1)$, $L_2 = p$, and $L_3 = (\sigma_0^2, \sigma_1^2, \tau)$ in the corresponding parameter space. Generate initial values of $L_4 = (U_i, i = 1, \cdots, n)$ independently from the log-normal distribution with parameters L_3. For latent process $L_5 = \mathbf{Z}$, we first assign values at time points $Z_i(j),\ i = 1, \cdots, n;\ j = 1, \cdots, n_i$, using the posterior probability $P(Z_i(j) = 1|X_{ij} = 1) = 1$ and $P(Z_i(j) = 0|X_{ij} = 0) = \frac{1}{2-p}$ by assuming $P(Z_i(j) = 0) = P(Z_i(j) = 1) = \frac{1}{2}$; and then generate the transition points $t_i = (t_{i,1}, \cdots, t_{i,m_i})$ for individual i in the following way:

1. Let $T = 1$, $s = Z_i(1)$, and $l = 1$.

2. Find the next nearest data point $d \in \mathbb{N}$, $T < d \leq n_i$, such that $Z_i(d) \neq s$.

3. Generate $\Delta t \sim f_{si}(t)1_{\{0<t\leq d-T\}}$. Let $T = T + \Delta t$. If $T < n_i$, set $s = 1 - s$, $t_{i,l} = T$, $l = l + 1$, and go back to step 2; otherwise stop.

Proposal for $\mathbf{L_1}$: Generate the four parameters of Weibull distribution $(\alpha_0, \alpha_1, \beta_0, \beta_1)$ from a truncated normal distribution with the current values as the means and a small standard deviation $\epsilon_1 > 0$ in one Markov chain iteration, that is,

$$\begin{aligned} q(\alpha_s^\star|\alpha_s^{old}) &\sim \mathrm{N}(\alpha; \alpha_s^{old}, \epsilon_1)1_{\{\alpha_s>0\}},\quad s = 0, 1 \\ q(\beta_s^\star|\beta_s^{old}) &\sim \mathrm{N}(\beta; \beta_s^{old}, \epsilon_1)1_{\{\beta_s>0\}},\quad s = 0, 1. \end{aligned}$$

Proposal for $\mathbf{L_2}$: Generate the parameter p from a truncated normal distribution with the current value as the mean and a small standard deviation $\epsilon_2 > 0$, that is,

$$q(p^\star|p^{old}) \sim \mathrm{N}(p; p^{old}, \epsilon_2)1_{\{0\leq p\leq 1\}}.$$

Proposal for $\mathbf{L_3}$: The parameters $\sigma_0^2, \sigma_1^2, \tau$ denote the variance and correlation coefficient

of the logarithm of the random effects. We perform a random walk on the parameters as:

$$\begin{aligned}(\sigma_0^2)^\star &\sim (\sigma_0^2)^{old}\,\mathrm{Unif}(1-\epsilon_3, 1+\epsilon_3)\\ (\sigma_1^2)^\star &\sim (\sigma_1^2)^{old}\,\mathrm{Unif}(1-\epsilon_3, 1+\epsilon_3)\\ \tau^\star &\sim \mathrm{N}(\tau; \tau^{old}, \epsilon_4)1_{\{-1\le\tau\le 1\}},\end{aligned}$$

where $\epsilon_3 > 0$ and $\epsilon_4 > 0$ control the scale of the perturbation. All the three parameters are updated once in one Markov chain iteration.

Proposal for $\mathbf{L_4}$: Randomly choose an individual i and generate the random effects of this individual from a truncated normal distribution with the current values as the means and a small standard deviation $\epsilon_5 > 0$ in one Markov chain iteration, that is,

$$\begin{aligned}q(u_{0i}^\star|u_{0i}^{old}) &\sim \mathrm{N}(u; u_{0i}^{old}, \epsilon_5)1_{\{u_{0i}>0\}}\\ q(u_{1i}^\star|u_{1i}^{old}) &\sim \mathrm{N}(u; u_{1i}^{old}, \epsilon_5)1_{\{u_{1i}>0\}}.\end{aligned}$$

Proposal for $\mathbf{L_5}$: To find a proposal distribution for the latent process $L_5 = \mathbf{Z}$, we need to construct a continuous path where the values of the path at the predetermined time points match the observed values with imperfect detectability p. We use the following proposal distribution which works well for our model. Given the current latent process $Z_i(t)$ of an individual i, we consider three possible moves: M1 – randomly perturb a transition time point; M2 – split a transition interval into three pieces by adding two transition points in the interval; and M3 – merge three consecutive duration intervals into one long interval. See Figure 2.2 for an illustration of the three moves. The detailed procedure of updating the i-th individual's latent process $Z_i(t)$ with transition time points $t_i = (t_{i,1}, \cdots, t_{i,m_i})$ is explained in the following:

1. Let c equal to 1, 2 and 3 with equal probability;

2. If $c = 1$, the move M1 is chosen. Choose one integer number l from 1 to m_i with

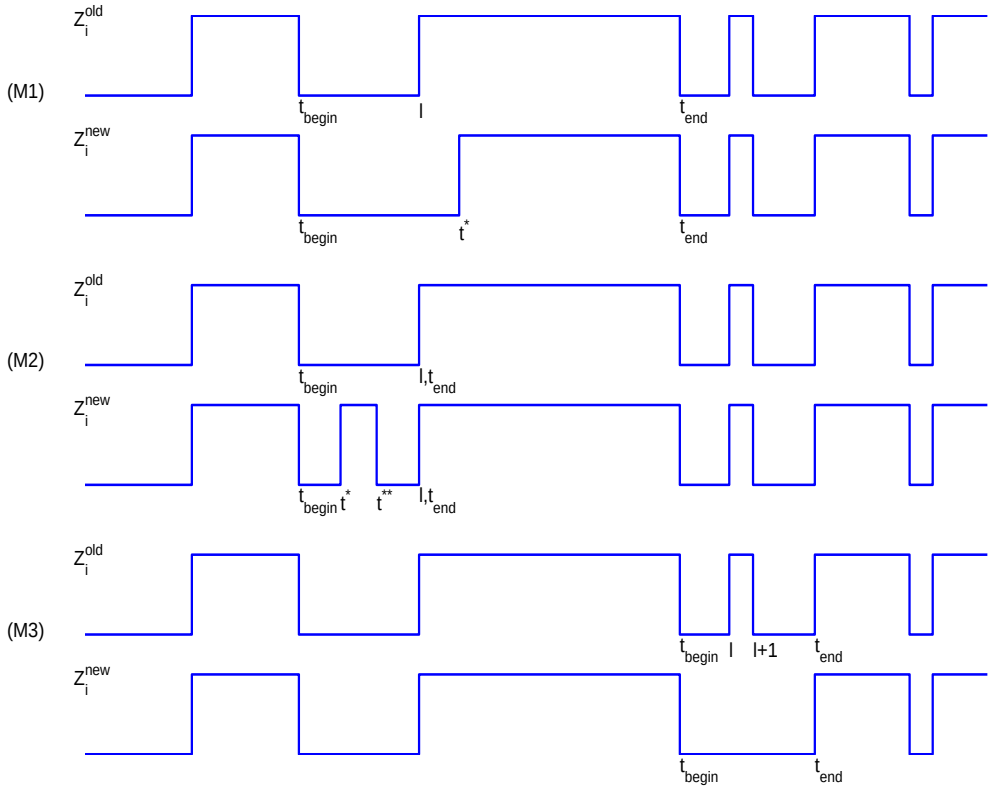


Figure 2.2: Three moves for latent process Z_i. M1 chooses $l = 3$ and changes $t_{i,l}$ to $t^\star$; M2 chooses $l = 3$ and inserts two transition time points $t^\star$ and $t^{\star\star}$ between t_{begin} and t_{end}; and M3 chooses $l = 5$ and deletes two transition time points $t_{i,l}$ and $t_{i,l+1}$ between t_{begin} and t_{end}.

equal probability. Randomly generate a value $t^\star$ from $\text{Unif}(t_{\text{begin}}, t_{\text{end}})$, where $t_{\text{begin}} = t_{i,l-1}1_{\{l>1\}} + 1_{\{l=1\}}$ and $t_{\text{end}} = t_{i,l+1}1_{\{l<m_i\}} + n_i 1_{\{l=m_i\}}$. Let Z_i^{new} be the old m_i transition points with the l-th element replaced by $t^\star$. Calculate the proposal probability:

$$q(Z_i^{new}|Z_i^{old}) = \frac{1}{3m_i(t_{\text{end}} - t_{\text{begin}})},$$

and $q(Z_i^{old}|Z_i^{new})$ is the same.

3. If $c = 2$, the move M2 is chosen. Choose one integer number l from 1 to $m_i + 1$ with equal probability. Randomly generate two values $t^\star < t^{\star\star}$ independently from $\text{Unif}(t_{\text{begin}}, t_{\text{end}})$, where $t_{\text{begin}} = t_{i,l-1}1_{\{l>1\}} + 1_{\{l=1\}}$ and $t_{\text{end}} = t_{i,l}1_{\{l<m_i+1\}} + n_i 1_{\{l=m_i+1\}}$. Let Z_i^{new} be the old m_i transition points with two more transition time points $t^\star$ and

$t^{\star\star}$ inserted after the first l components. Calculate the proposal probabilities

$$q(Z_i^{new}|Z_i^{old}) = \frac{1}{3(m_i+1)}\frac{2}{(t_{\text{end}}-t_{\text{begin}})^2} \text{ and } q(Z_i^{old}|Z_i^{new}) = \frac{1}{3(m_i+1)}.$$

4. If $c = 3$, the move M3 is chosen. Choose one integer number l from 1 to $m_i - 1$ with equal probability. Let Z_i^{new} be the old m_i transition points with two transition time points $t_{i,l}$ and $t_{i,l+1}$ deleted. Calculate the proposal probabilities

$$q(Z_i^{new}|Z_i^{old}) = \frac{1}{3(m_i-1)} \text{ and } q(Z_i^{old}|Z_i^{new}) = \frac{1}{3(m_i-1)}\frac{2}{(t_{\text{end}}-t_{\text{begin}})^2},$$

where $t_{\text{begin}} = t_{i,l-1}1_{\{l>1\}} + 1_{\{l=1\}}$ and $t_{\text{end}} = t_{i,l+2}1_{\{l<m_i-1\}} + n_i 1_{\{l=m_i-1\}}$.

With the three moves, the Markov chain is irreducible as all possible latent processes can be reached no matter what the initial latent process is.

2.5 Numerical Studies

In this section, we first evaluate the performance of our model on simulated data, and then we apply it to the real data introduced in Section 2.2. We start with the simple model with constant hazard rates, imperfect detectability and homogeneity among individuals (Model 1), then we consider the more complicate model with constant hazard rates, imperfect detectability and heterogeneity among individuals (Model 2), and finally we study the general model with nonconstant hazard rates, imperfect detectability and heterogeneity among individuals (Model 3).

Model	n	n_i	p	α_1	β_1	α_0	β_0	σ_1	σ_0	τ
1	50	(30, 50)	0.95	3.33	1.0	5	1.0	–	–	–
2	50	(30, 50)	0.95	3.33	1.0	5	1.0	0.2	0.2	0.25
3	50	(30, 50)	0.90	3.33	1.2	5	0.8	0.2	0.2	0

Table 2.2: Simulated data settings for all three models.

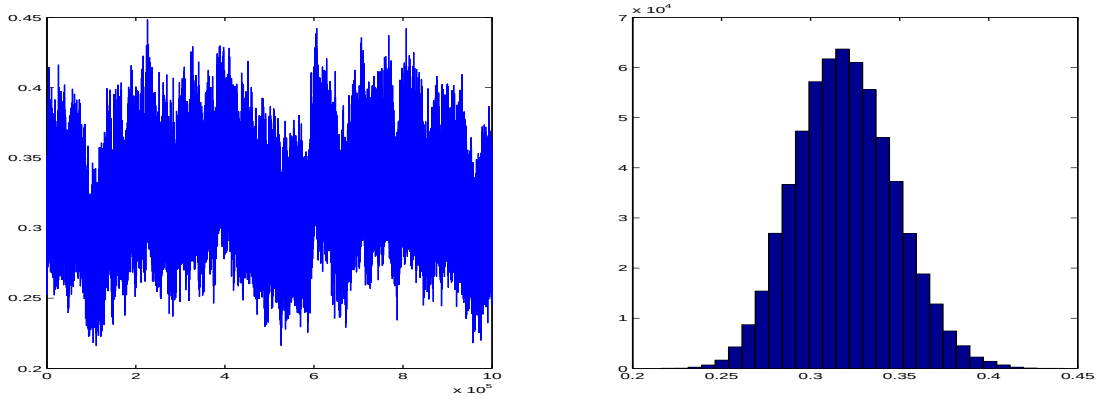
For each model, we simulated 500 data sets with the same parameter values from the model. Each data set contains $n = 50$ individuals with the number of observations for each individual n_i ranging from 30 to 50. The true values of the parameters that have been used to generate the simulated data are given in Table 2.2. In Model 1, there is no heterogeneity among individuals, so there are no values for the three parameters of the covariance matrix.

As discussed before, the probability density function of Weibull distribution when $\beta_s = 1$, $s = 0, 1$, is degenerated to exponential distribution with rate parameter $\lambda_s = 1/\alpha_s$, $s = 0, 1$. For Models 1 and 2, we estimate α_s only by fixing β_s at its true value, which is equivalent to assuming exponential survival distribution and estimating its rate parameter $\lambda_s = 1/\alpha_s$.

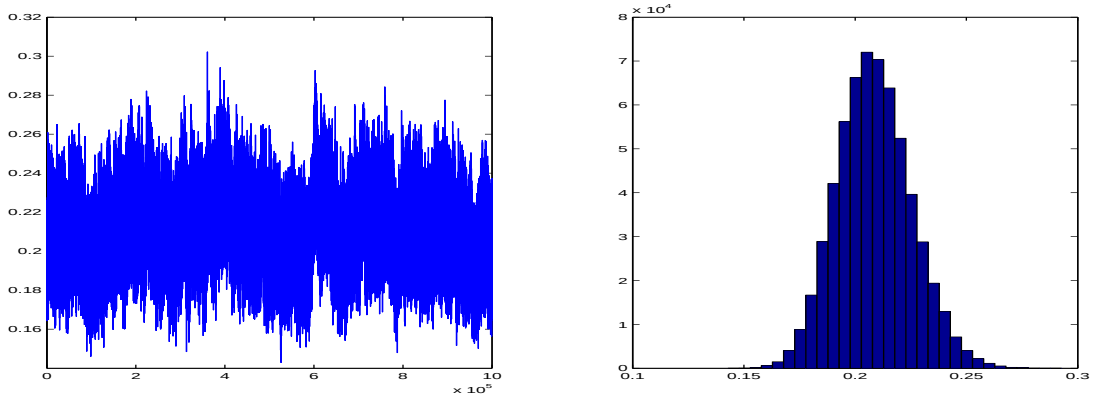
To estimate the parameters in each data set, we follow the initialization procedure and generate samples from the proposal distributions under the MCMC framework described in Section 2.4. The tuning parameters $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5)$ of the proposal distributions are set as $(0.1, 0.03, 0.03, 0.05, 0.1)$ to get a reasonable acceptance rate for each parameter. Then the posterior mean and 95% posterior credible interval of each parameter are derived based on samples from Markov chains after some burn-in period. Here the 95% posterior credible interval of each parameter is derived by using the 2.5-th and the 97.5-th percentiles of the samples. For the overall performance of each model, we also report the bias and the coverage probabilities of 95% posterior credible intervals of each parameter based on 500 simulated data sets.

Model 1: By fixing β_s, $s = 0, 1$, at its true value, we estimate the remaining three parameters $(\lambda_1, \lambda_0, p)$ together. The prior distribution for λ_s $(s = 0, 1)$ is set as Gamma$(0.01, 0.01)$ and the prior distribution for p is chosen as Beta$(0.01, 0.01)$. All the following estimates are based on samples from 1,000,000 MCMC iterations with a 400,000 burn-in period.

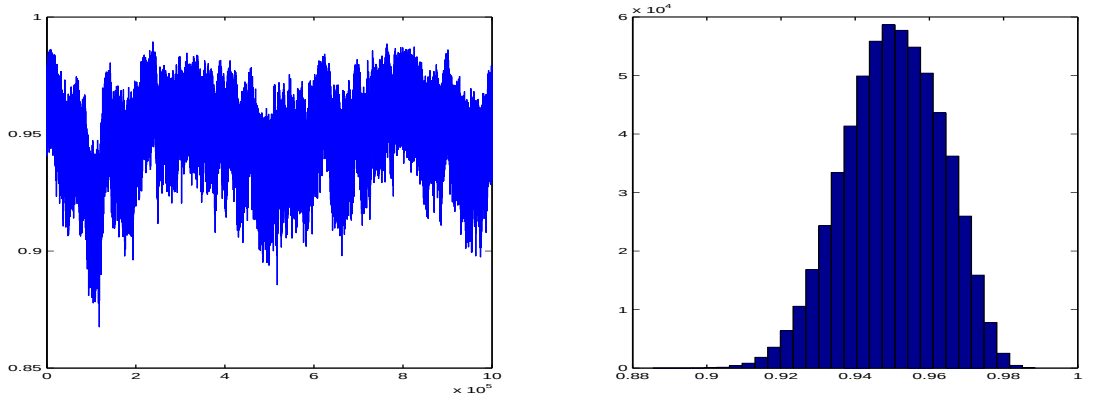
Let's first look at the performance for one simulated data set. The trace plots and histograms of samples from the MCMC procedure are given in Figure 2.3. The trace plots show that the Markov chain converges quickly and mixes well; the histograms are roughly



(a) λ_1



(b) λ_0



(c) p

Figure 2.3: Sample trace plots and histograms of the parameters in Model 1 based on one simulated data set.

Parameter	Posterior Mean	95% Credible Interval	True Value
λ_1	0.3196	[0.2685, 0.3763]	0.30
λ_0	0.2087	[0.1779, 0.2430]	0.20
p	0.9502	[0.9240, 0.9737]	0.95

Table 2.3: Parameter estimation for one simulated data set from Model 1.

bell-shaped with means close to the true values. Table 2.3 gives the posterior means and 95% credible intervals of all parameters. We can see that the bias of each estimate is small and all the 95% credible intervals cover the true values.

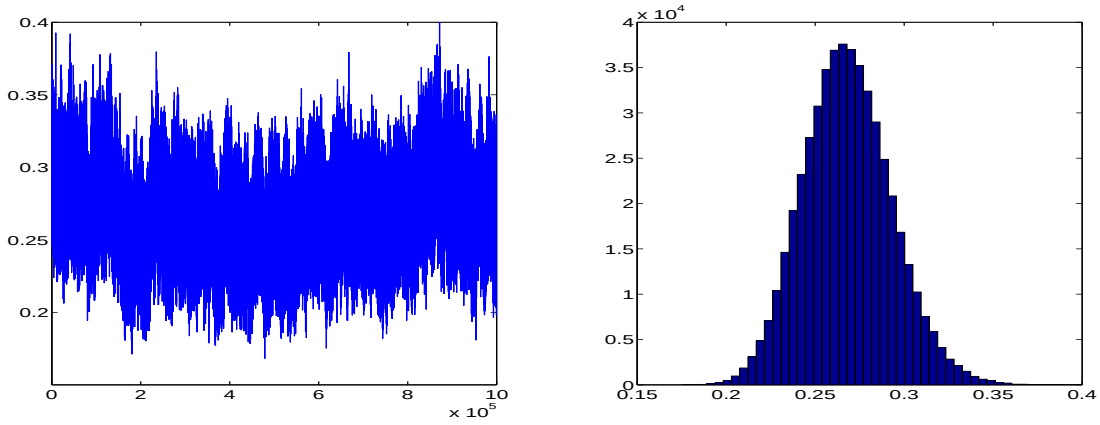
Parameter	Bias	95% CI Coverage
λ_1	0.0218	94.2%
λ_0	0.0133	95.6%
p	0.0121	93.2%

Table 2.4: The bias and coverage probability of 95% credible intervals of each parameter in Model 1 based on 500 simulated data sets.

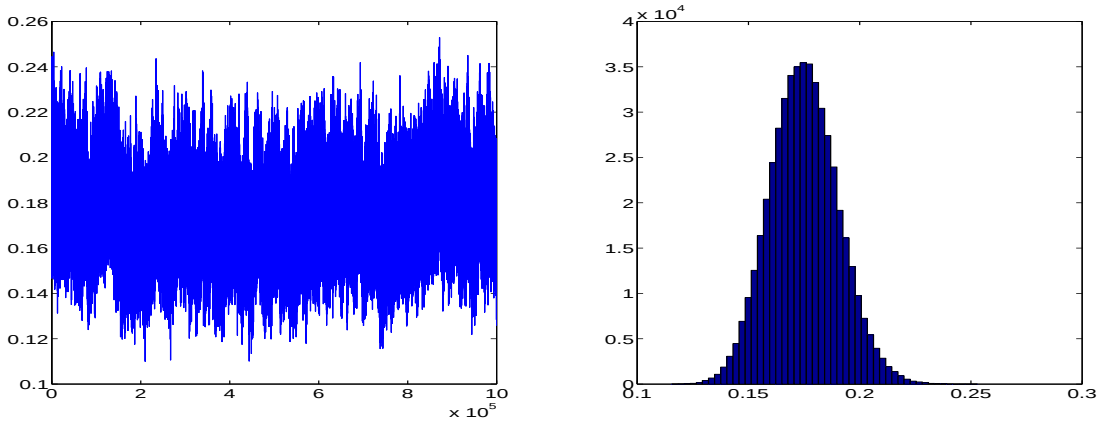
For the overall performance of this model, we provide in Table 2.4 the coverage probability of 95% credible intervals and the bias of each parameter based on 500 simulated data sets. The averaged bias for each parameter is very small with a maximum value around 0.02. The coverage probabilities for all parameters are very close to the nominal level 95%. In summary, we can obtain reasonable estimates of all parameters in this model.

Model 2: Similar to Model 1, we fix β_s at its true value and estimate the remaining parameters $(\lambda_1, \lambda_0, p, \sigma_1, \sigma_0, \tau)$ together. The priors for λ_s, $(s = 0, 1)$ and p are the same as those of Model 1. The prior distribution for the covariance matrix of the logarithm of random effects is set as Inverse-Wishart$(I_2, 3)$ where I_2 is a 2×2 identity matrix. All the following results are based on samples from 1,000,000 MCMC iterations with a 500,000 burn-in period.

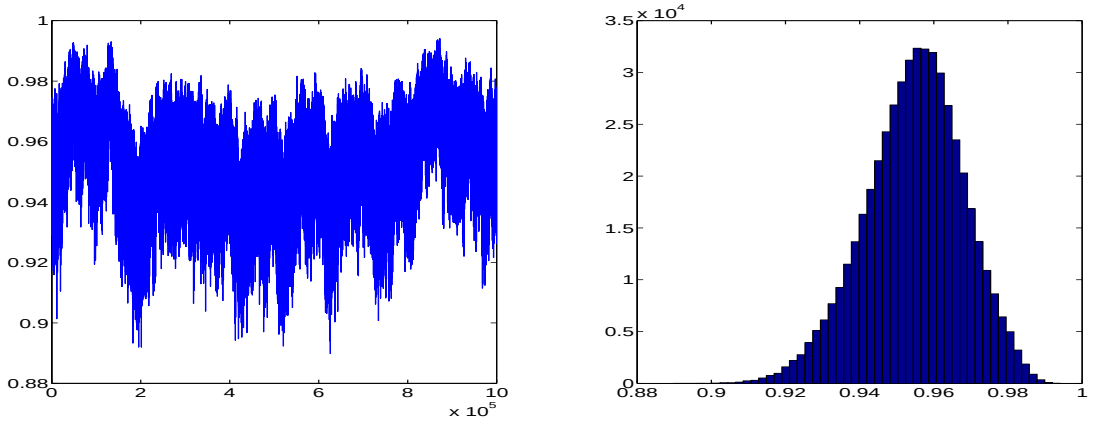
For one simulated data set of Model 2, we show the trace plots and histograms of MCMC samples in Figures 2.4 and 2.5. Similar to Model 1, the Markov chain exhibits fast conver-



(a) λ_1

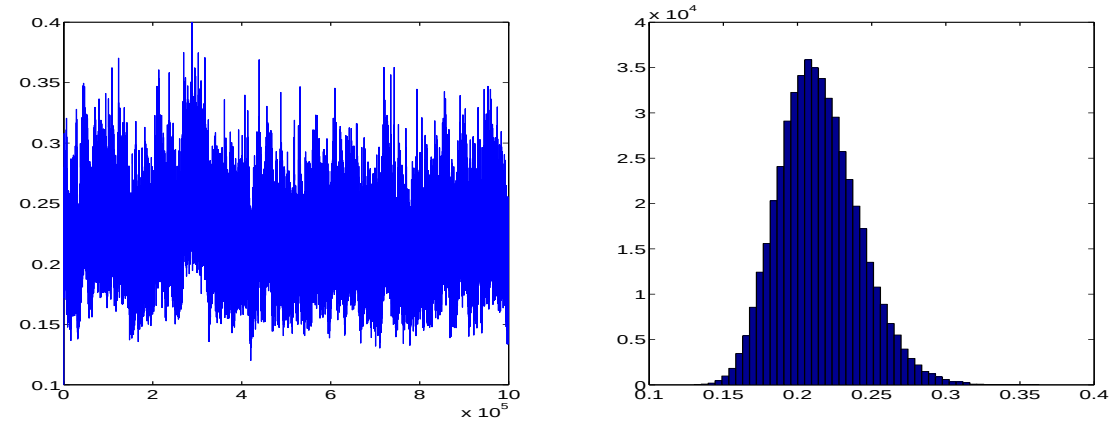


(b) λ_0

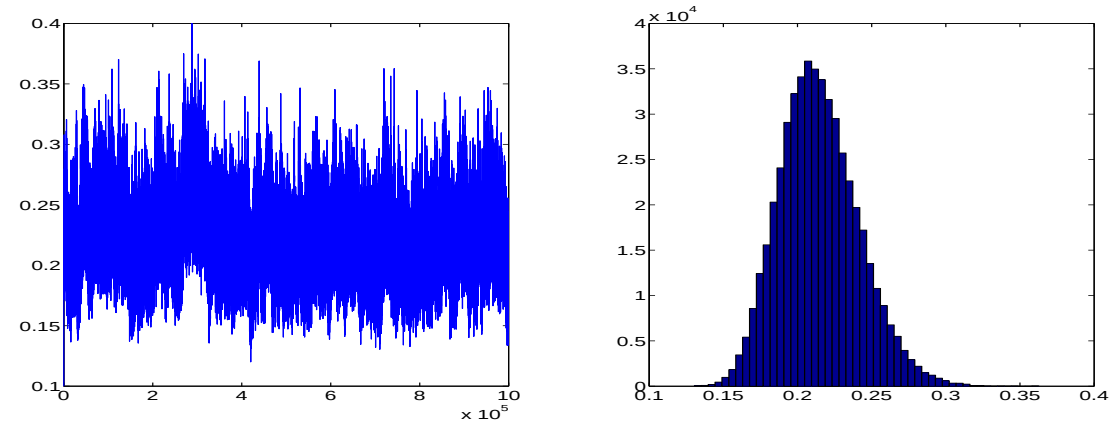


(c) p

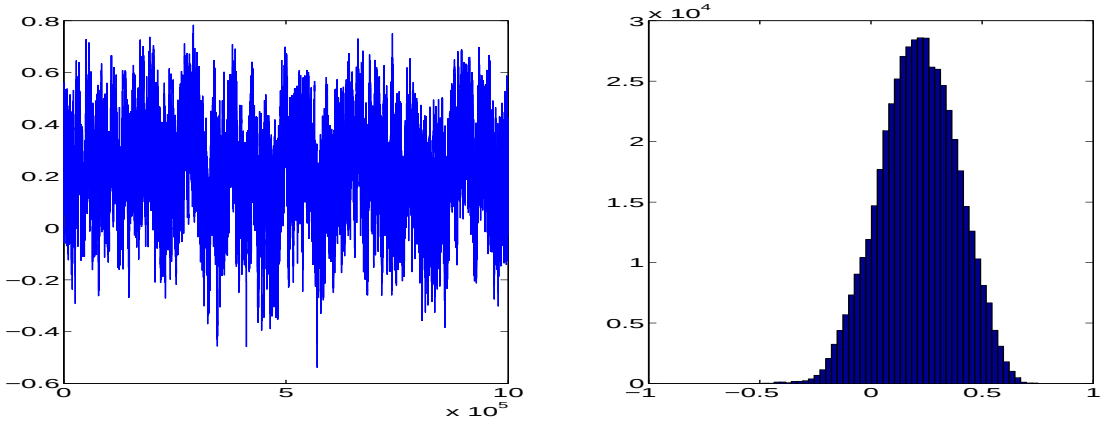
Figure 2.4: Sample trace plots and histograms of the parameters λ_1, λ_0 and p in Model 2 based on one simulated data set.



(a) σ_1



(b) σ_0



(c) τ

Figure 2.5: Sample trace plots and histograms of the parameters σ_1, σ_0 and τ in Model 2 based on one simulated data set.

Parameter	Posterior Mean	95% Credible Interval	True Value
λ_1	0.2682	[0.2222, 0.3196]	0.30
λ_0	0.1751	[0.1461, 0.2062]	0.20
p	0.9547	[0.9269, 0.9791]	0.95
σ_1	0.2147	[0.1677, 0.2721]	0.20
σ_0	0.1914	[0.1458, 0.2471]	0.20
τ	0.2143	[-0.1280, 0.5381]	0.25

Table 2.5: Parameter estimation for one simulated data set from Model 2.

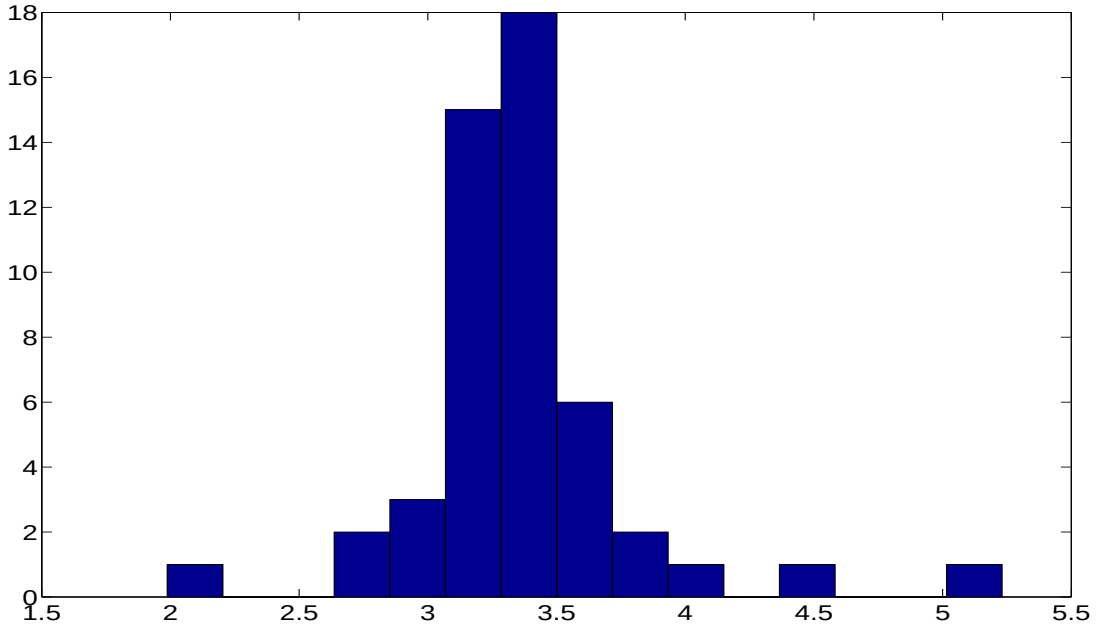
gence and most of the histograms are centered around the true values. The corresponding posterior means and 95% credible intervals are given in Table 2.5. The bias of each estimate is small and all the 95% credible intervals cover the true values.

In this model, we introduced three parameters σ_1, σ_0 and τ to describe the heterogeneity among individuals. To see how the individuals behave differently, we plot in Figure 2.6 the average duration times at infected and uninfected states for all individuals based on one simulated data set from this model. For infected state, most of the individuals have an average duration time from 3 to 4 time units, but one individual's average duration time is 5 time units and another one is 2 time units. Similarly, for uninfected state, all the individuals have an average duration time from 5 to 6 time units except a few with much larger or smaller duration times.

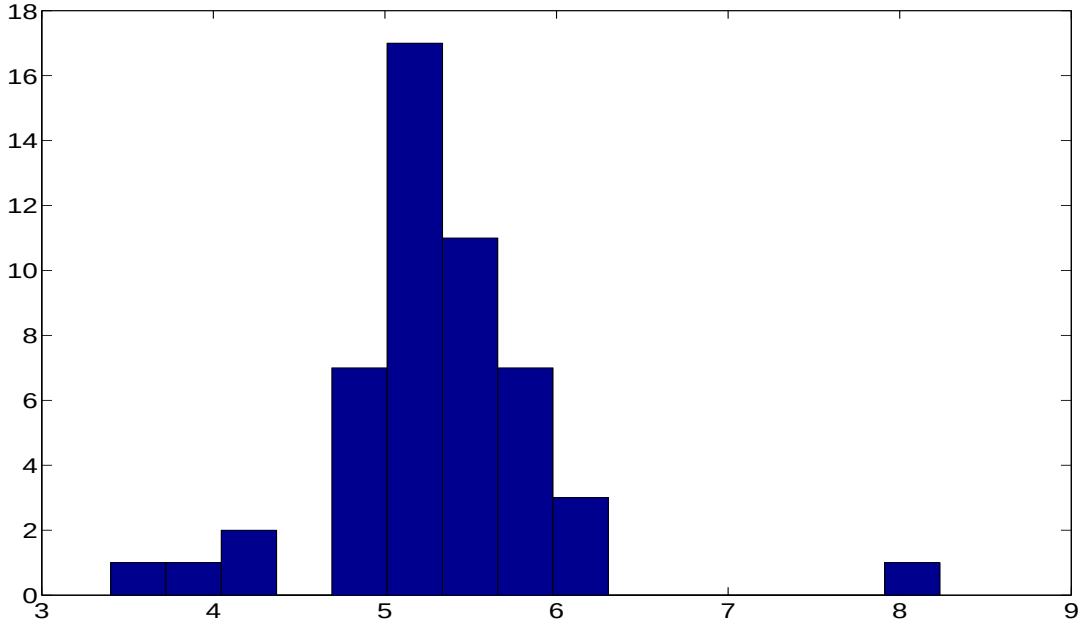
Parameter	Bias	95% CI Coverage
λ_1	0.0227	94.2%
λ_0	0.0142	94.4%
p	0.0116	94.8%
σ_1	0.0230	96.4%
σ_0	0.0239	95.0%
τ	0.1656	95.8%

Table 2.6: The bias and coverage probability of 95% credible intervals of each parameter in Model 2 based on 500 simulated data sets.

To see the overall performance of this model, we list in Table 2.6 the coverage probability of 95% credible intervals and the bias of each parameter based on 500 simulated data sets. The biases of all parameters are small except τ which has a relatively large bias 0.1656. The



(a) Duration times at infected state



(b) Duration times at uninfected state

Figure 2.6: Duration times at infected and uninfected states for all individuals based on one simulated data set from Model 2.

coverage probabilities of the 95% credible intervals for all parameters are close to the nominal level 95%. Therefore, the estimates of all parameters in this model are still satisfactory.

Model 3: Here we work on the general model using Weibull survival function with random effects which allows for nonconstant hazard rates, imperfect detectability and heterogeneity among individuals. The parameters are $\alpha_1, \beta_1, \alpha_0, \beta_0, p, \sigma_1, \sigma_0$ and τ. We set the prior distribution for α_s and β_s, $s = 0, 1$, as Gamma$(0.01, 0.01)$, the prior distribution of p as Beta$(0.01, 0.01)$, and the prior distribution for the covariance matrix of the logarithm of random effects as Inverse-Wishart$(I_2, 3)$. All the following results are based on samples from 1,000,000 MCMC iterations with a 500,000 burn-in period.

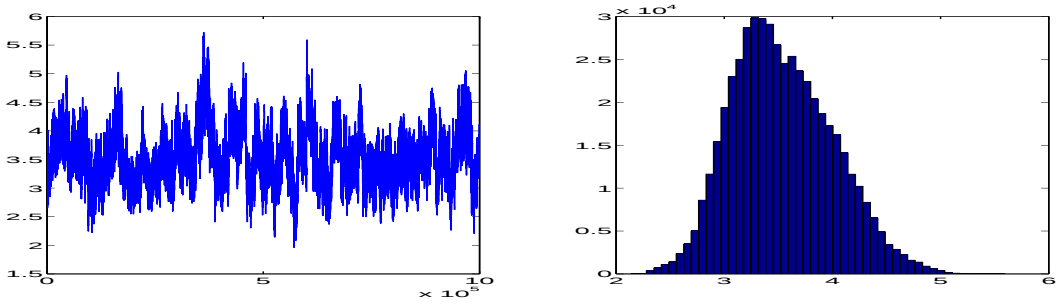
Parameter	Posterior Mean	95% Credible Interval	True Value
α_1	3.5381	[2.7388, 4.5162]	3.33
β_1	1.2395	[0.9519, 1.6884]	1.20
α_0	5.7437	[4.1842, 7.9212]	5.00
β_0	0.8106	[0.6778, 0.9837]	0.80
p	0.8961	[0.8565, 0.9354]	0.90
σ_1	0.2249	[0.1346, 0.3445]	0.20
σ_0	0.2270	[0.1323, 0.3564]	0.20
τ	-0.0303	[-0.6063, 0.5618]	0

Table 2.7: Parameter estimation for one simulated data set from Model 3.

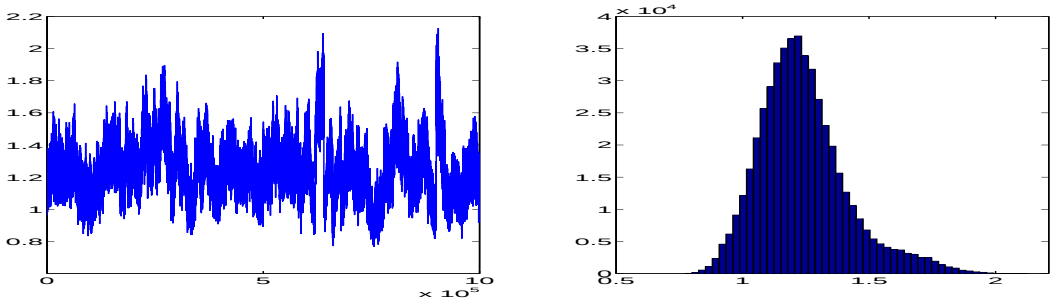
First, we show in Figures 2.7 and 2.8 the trace plots and histograms of MCMC samples based on one simulated data set from Model 3. The corresponding posterior means and 95% credible intervals are given in Table 2.7. We can see that the estimates of all the parameters are close to the true values and the maximum bias is 0.7437 for parameter α_0. For this specific data set, all the 95% credible intervals cover the true values.

Next, we show in Figure 2.9 the average duration times at infected and uninfected states for all individuals based on one simulated data set from this model. The heterogeneity among individuals is quite obvious from the plot.

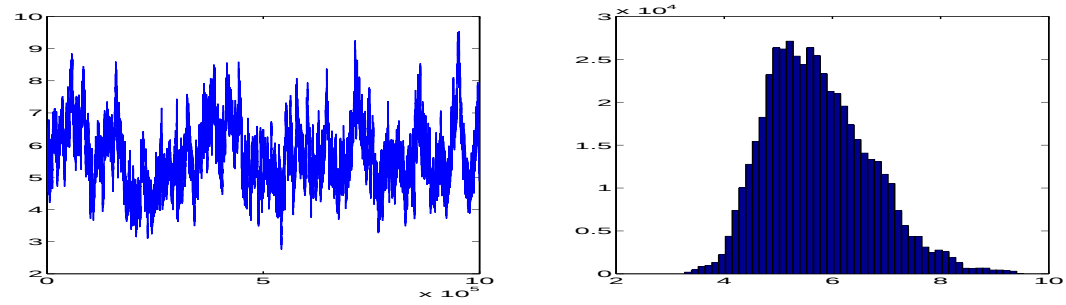
Table 2.8 shows the coverage probability of 95% credible intervals and the bias of each



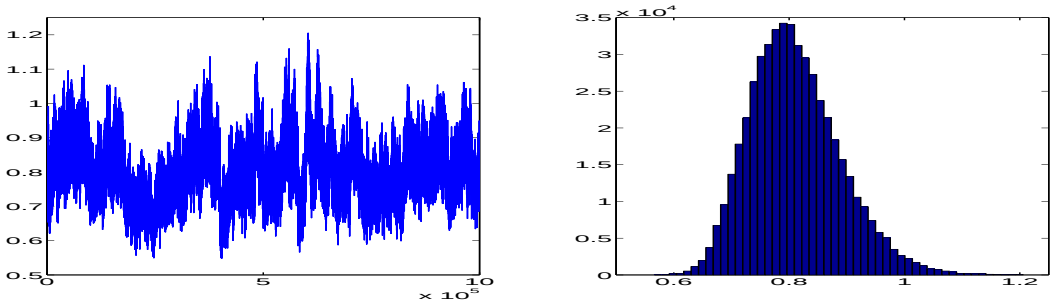
(a) α_1



(b) β_1

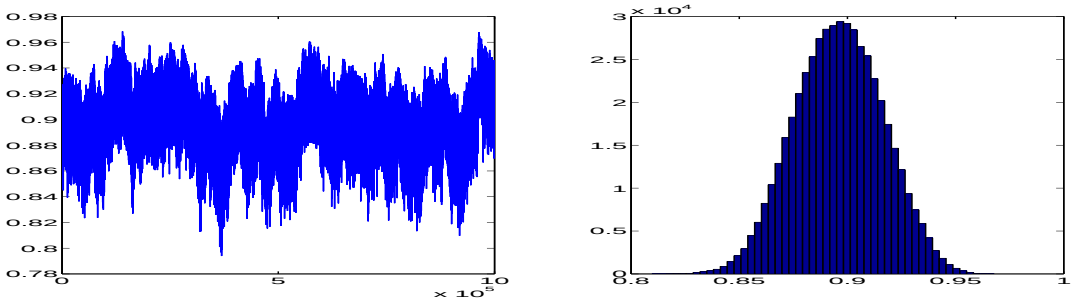


(c) α_0

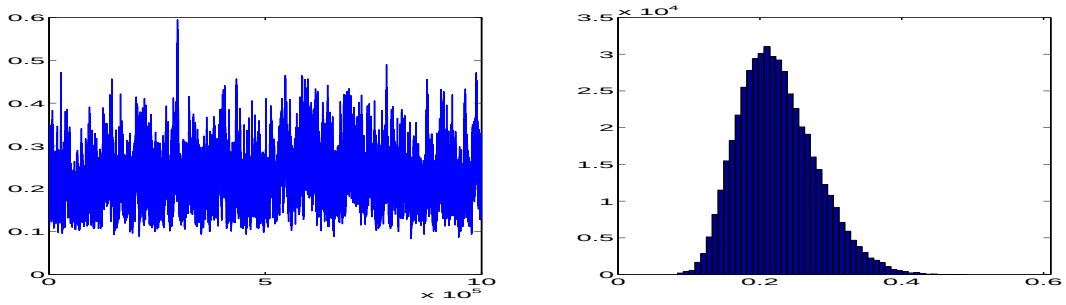


(d) β_0

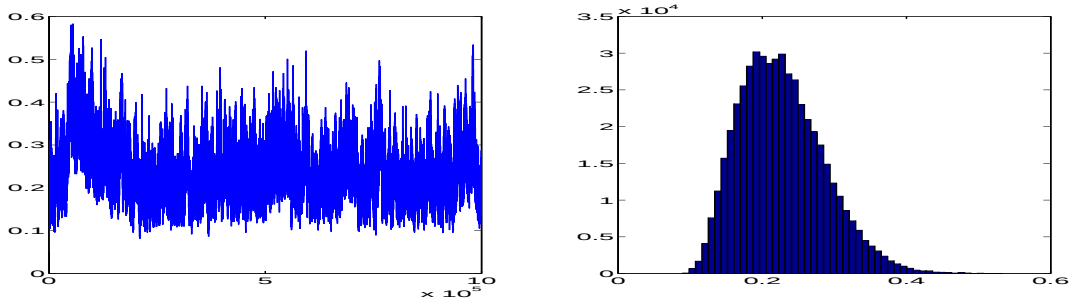
Figure 2.7: Sample trace plots and histograms of the parameters α_s and β_s $(s = 0, 1)$ in Model 3 based on one simulated data set.



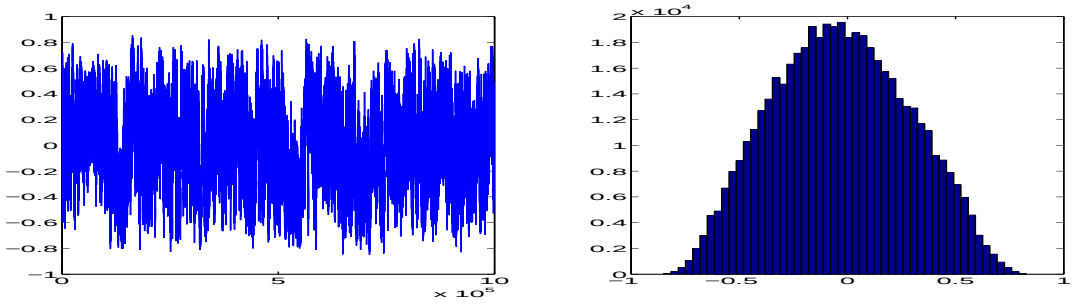
(a) p



(b) σ_1

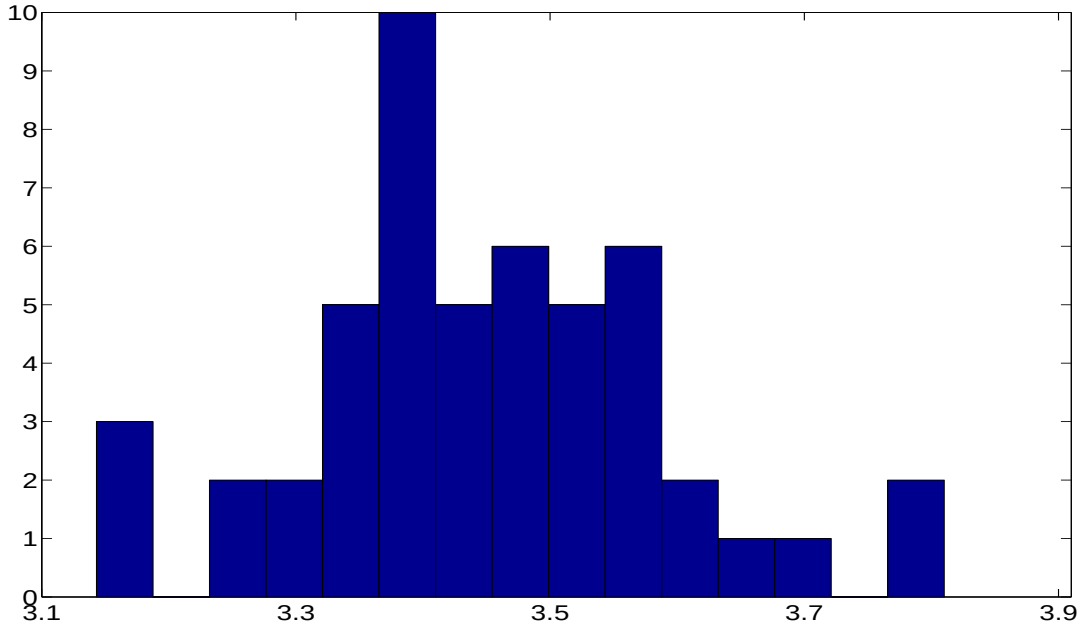


(c) σ_0

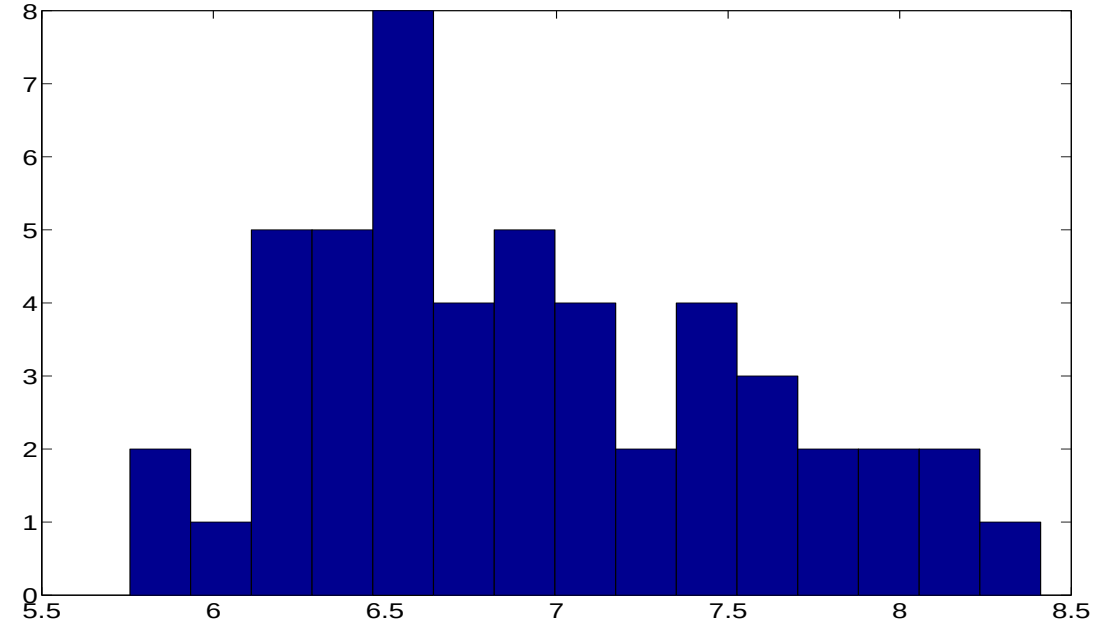


(d) τ

Figure 2.8: Sample trace plots and histograms of the parameters p, σ_1, σ_0 and τ in Model 3 based on one simulated data set.



(a) Duration times at infected state



(b) Duration times at uninfected state

Figure 2.9: Duration times at infected and uninfected states for all individuals based on one simulated data set from Model 3.

Parameter	Bias	95% CI Coverage
α_1	0.4907	88.6%
β_1	0.1615	87.8%
α_0	0.7851	89.0%
β_0	0.1119	84.8%
p	0.0205	90.0%
σ_1	0.0487	98.0%
σ_0	0.0521	98.4%
τ	0.1145	96.2%

Table 2.8: The average bias and coverage probability of 95% credible intervals of each parameter in Model 3 based on 500 simulated data sets.

parameter based on 500 simulated data sets from this general model. Overall, the coverage probabilities of the four parameters of Weibull survival function $\alpha_s, \beta_s, (s = 0, 1)$ as well as p are a little smaller than the nominal level 95%, and those of the parameters σ_1, σ_0 and τ are a little larger than the nominal level 95%. By introducing two more parameters in the Weibull survival function, we can still obtain reasonable estimates of all the parameters.

Real data: In this part, we re-analyze the data about the infection of *Giardia lamblia* introduced in Section 2.2. Following the same procedure as the simulated data, we start with the simple model assuming constant hazard rates and continue with the model allowing for nonconstant hazard rates. The prior distribution of each parameter is chosen to be the same as in the simulated data.

The results of the three models by assuming constant hazard rates: R1, R2, and R3 are shown in Table 2.9. Besides the constant hazard rates assumption, model R1 also assumes perfect detectability and homogeneity among individuals; model R2 assumes homogeneity among individuals; while model R3 has no additional assumptions. For models R1 and R2 with the assumptions of constant hazard rates and homogeneity among individuals, the maximum likelihood estimate (MLE) can be obtained as a result of the Markov property. Bekessy et al. (1976) gave the MLEs with perfect detectability and later Nagelkerke et al. (1990) provided the MLEs based on both the partial likelihood and the full likelihood with

Model	Parameter	Posterior Mean	95% Credible Interval	MLE
R1	λ_1	0.3696	[0.3052, 0.4409]	0.3581
	λ_0	0.3328	[0.2744, 0.3971]	0.3311
R2	λ_1	0.2724	[0.2074, 0.3469]	0.2359
	λ_0	0.3018	[0.2349, 0.3764]	0.3311
	p	0.8953	[0.8494, 0.9374]	0.9090
R3	λ_1	0.2986	[0.2354, 0.3687]	–
	λ_0	0.3139	[0.2517, 0.3843]	–
	p	0.9272	[0.8866, 0.9585]	–
	σ_1	0.2706	[0.1604, 0.3656]	–
	σ_0	0.2059	[0.1447, 0.2829]	–
	τ	0.2150	[-0.2182, 0.5913]	–

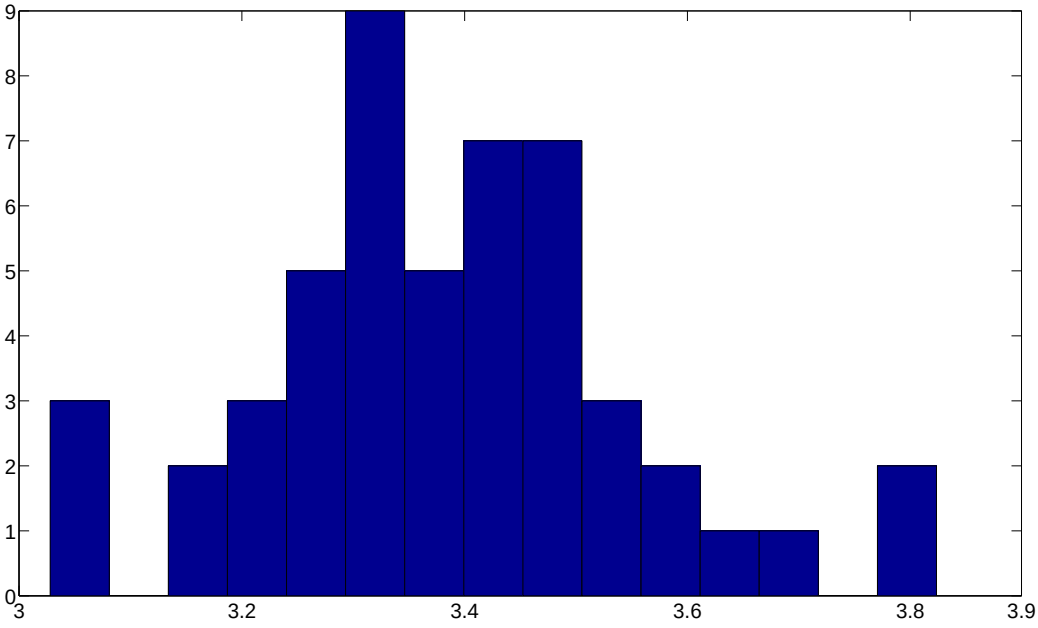
Table 2.9: Parameter estimation for the longitudinal data of infection with the parasite *Giardia lamblia* among children in Kenya by assuming constant hazard rates.

imperfect detectability. For comparison, we also list the MLEs of each parameter in these two models. Note that all the estimates including MLEs in Table 2.9 use week as the unit instead of day.

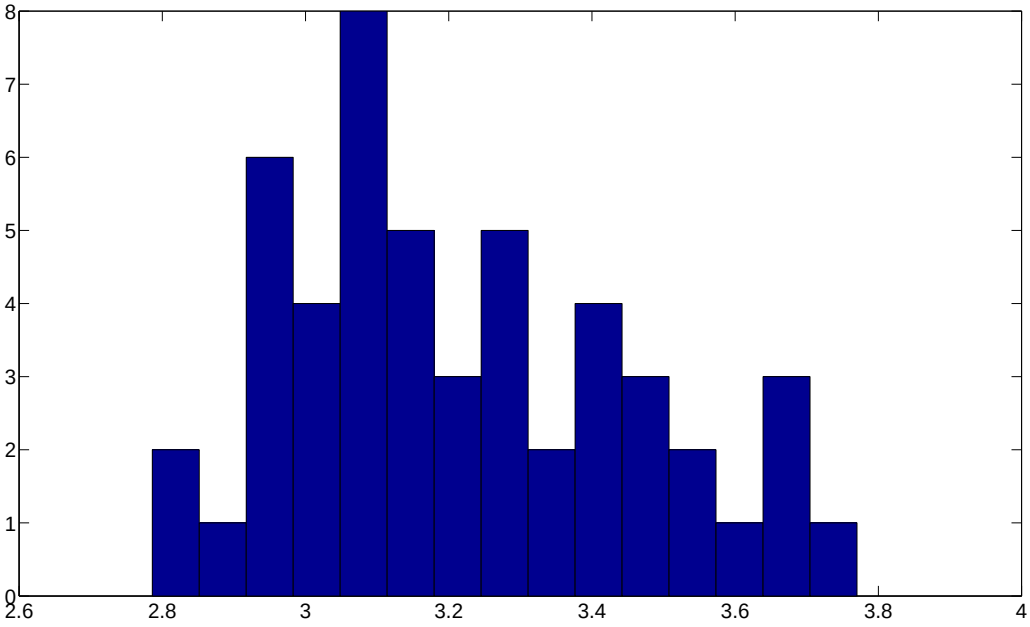
For model R1, we only have two parameters λ_1 and λ_0. Both posterior means are close to the MLEs. From the estimates, we observe that the transition rate from infected to uninfected status (λ_1) is slightly higher than the hazard rate from uninfected to infected status (λ_0), but the 95% credible intervals of both parameters overlap a lot, which implies the difference is not significant.

For model R2, the posterior mean of p is 0.8953, which indicates that the detection procedure fails to detect around 10.5% infection cases. The posterior means and MLEs are still close for all the parameters with the 95% credible intervals covering the MLEs. By allowing for imperfect detectability, the estimates of λ_1 and λ_0 are different from those of model R1, especially for λ_1. Now the transition rate from infected to uninfected status is smaller than the hazard rate from uninfected to infected status, but the difference is not significant. Allowing imperfect detectability makes it possible to discover more about the underlying disease process.

Compared to model R2, model R3 incorporates the heterogeneity among individuals.



(a) Duration times at infected state



(b) Duration times at uninfected state

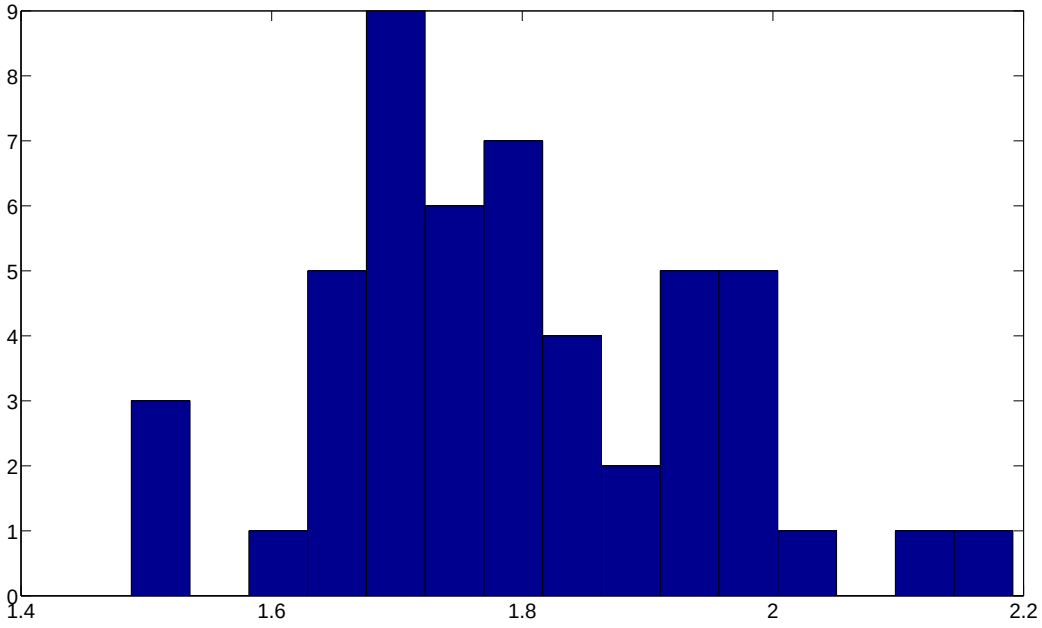
Figure 2.10: Duration times at infected and uninfected states for all individuals of the real data in model R3.

The estimates of λ_1, λ_0 and p are similar to model R2, which indicates that allowing heterogeneity among individuals does not affect the estimates of other parameters. Moreover, the correlation between the logarithm of random effects is not significantly different from 0 as the corresponding 95% credible interval covers 0. To check the heterogeneity among individuals, we show in Figure 2.10 the average duration times at infected and uninfected states for all individuals. The durations times at each state are similar but still show some variation.

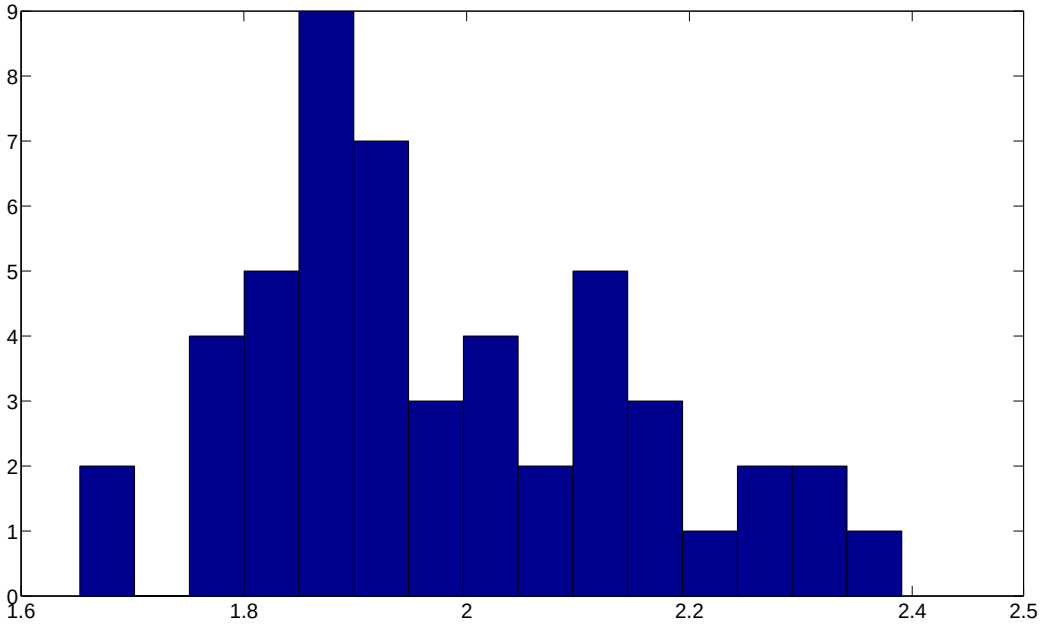
Model	Parameter	Posterior Mean	95% Credible Interval
R4	α_1	1.0937	[0.7007, 1.6612]
	β_1	0.6089	[0.5119, 0.7564]
	α_0	1.4741	[0.9172, 2.1138]
	β_0	0.7236	[0.5574, 0.9022]
R5	α_1	1.0679	[0.5318, 2.2026]
	β_1	0.5220	[0.4201, 0.7166]
	α_0	1.6608	[1.0737, 2.5884]
	β_0	0.7935	[0.6089, 1.0025]
	p	0.9285	[0.8908, 0.9622]
R6	α_1	1.2758	[0.6546, 2.1114]
	β_1	0.6981	[0.5412, 0.9823]
	α_0	1.7384	[0.8608, 2.7434]
	β_0	0.9034	[0.6256, 1.3331]
	p	0.9487	[0.9068, 0.9821]
	σ_1	0.3170	[0.1606, 0.5494]
	σ_0	0.4683	[0.2135, 0.8676]
	τ	-0.4531	[-0.8851, 0.2977]

Table 2.10: Parameter estimation for the longitudinal data of infection with the parasite *Giardia lamblia* among children in Kenya by allowing nonconstant hazard rates.

Next, we investigate the models allowing for nonconstant hazard rates. The corresponding results are also shown in Table 2.9. Model R4 assumes perfect detectability and homogeneity among individuals. We can see that the estimates of β_1 and β_0 are smaller than 1 and the 95% credible intervals do not cover 1. This indicates the actual hazard rates are not constant and they decay as the time goes on. Therefore, adopting the Weibull survival function allows us to detect non-constant hazard rates.



(a) Duration times at infected state



(b) Duration times at uninfected state

Figure 2.11: Duration times at infected and uninfected states for all individuals of the real data in Case R6.

Model R5 allows imperfect detectability besides nonconstant hazard rates. The posterior mean of p is 0.9285, which means that the detection procedure fails to detect 7.15% infected cases. The estimates of the parameters of the Weibull survival function also changed slightly comparing to model R4. Note that the credible interval for the shape parameter β_0 has an upper bound about 1.

Finally, model R6 considers the most general situation. The estimates of the parameters of the survival function are similar to those of model R5 and the estimate of the imperfect detectability is a little higher in model R6. Again, there is no significant correlation between the logarithm of random effects given that the 95% credible interval of τ includes 0. There exists heterogeneity among individuals according to Figure 2.10.

Model R6 fits the data better than the other models. First, the imperfect detectability is non-negligible due to the imperfect diagnostic instruments and procedures. The estimated value of p is below 0.95. Second, the hazard rates appear to be nonconstant especially for the infected state given that the 95% credible interval of β_1 does not cover 1. Third, the data reveals clear heterogeneity among individuals. Therefore, we conclude that it is important to consider nonconstant hazard rates, imperfect detectability and random effects in the model. In this way, we can detect the influence of imperfect detectability and random effects and learn more about the underlying true dynamics of parasites.

2.6 Conclusion

We proposed a Bayesian hierarchical model to study the behavior of Giardia lamblia based on the longitudinal data in Chunge (1989). Our model is flexible by allowing (1) imperfect detectability, (2) non-constant hazard rates, and (3) heterogeneity among individuals. We also proposed an MCMC algorithm with data augmentation to estimate the parameters in such a model. Simulation studies show that we can obtain reasonable estimates of all parameters under different settings.

Chapter 3

Complementary Dimensionality Analysis

3.1 Introduction

Dimension reduction plays an important role in high-dimensional data analysis. The goal is to map the p-dimensional covariate X to a lower dimensional space while keeping the meaningful information. Many dimension reduction algorithms have been proposed in the literature. Depending on whether or not the response information Y is considered, dimension reduction algorithms can be categorized into supervised and unsupervised ones.

Unsupervised dimension reduction algorithms aim to find the most informative representation of the data X. Popular unsupervised algorithms include Principal Component Analysis (PCA), Independent Component Analysis (ICA) (Comon, 1994), Multidimensional Scaling (MDS) (Mardia et al., 1979), Locally-Linear Embedding (LLE) (Roweis and Saul, 2000), Locality Preserving Projections (LPP) (He and Niyogi, 2003), and many others.

In the setting of supervised dimension reduction algorithms, the goal is to extract the the most relevant information in X for the prediction of a response variable Y. For example, Fisher Discriminant Analysis (FDA) (Fisher et al., 1936) is one well-known supervised method to search a linear combination of covariates that maximizes the between-class distance with respect to within-class distance. Later, Sugiyama (2007) proposed an extension of FDA to manifolds learning, named Local FDA, by re-weighting the between-class and within-class distance in FDA. There are also various dimension reduction methods from the inverse regression perspective including Sliced Inverse Regression (SIR) (Li, 1991) and Sliced Average Variance Estimation (SAVE) (Cook and Weisberg, 1991).

On the other hand, dimension reduction algorithms can be categorized into linear or non-linear ones depending on the form of the transformation. The previously mentioned PCA, ICA, FDA, Local FDA, LPP, SIR, and SAVE, are linear algorithms, while MDS and LLE are non-linear algorithms.

In this chapter, we focus on linear supervised dimension reduction. It is equivalent to finding a projection matrix $\mathbf{V}_{p\times m} = (v_1, \cdots, v_m)$, such that the m-dimensional summary $\mathbf{V}^t X$ (where $m << p$) keeps the most discriminative information of Y. Mathematically, the optimal $\mathbf{V}$ is the projection matrix with the smallest m, such that the conditional distribution of $Y|X$ is the same as the one of $Y|\mathbf{V}^t X$ (Li, 1991). However, the assumption that the lower-dimensional representation is a lossless compression of the information in X (relative to the prediction of Y) is too stringent in practice. First, in many real data analysis, every feature is more or less relevant to the prediction. So a lossless representation usually ends up with the original data. Second, in some applications, the dimension of the reduced representation m is not up to the user's choice but subject to exterior constraints such as the capacity of the transmitting channel or the limit of storage space. In light of these practical concerns, we do not aim to retrieve a lossless representation of the data, or discuss the "correct" dimension m. Instead we focus on developing a framework for supervised dimension reduction (SDR), in which directions v_l's are retrieved sequentially by the order of decreasing importance to prediction.

Sequential SDR algorithms are often formulated as an optimization problem: the l-th direction is retrieved by solving

$$v_l = \underset{v\perp M_{l-1}}{\arg\max}\, G(v), \tag{3.1}$$

where M_{l-1} denotes the linear space spanned by the previously solved $(l-1)$ directions: $v_1, \ldots, v_{l-1}$. For example, the aforementioned FDA (Fisher et al., 1936) uses the Rayleigh quotient $v^t\mathbf{B}v/v^t\mathbf{W}v$ as the objective function, where $\mathbf{B}$ and $\mathbf{W}$ are, respectively, the

between-class and within-class scatter matrices; the corresponding solution is given by the eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$. Note that the objective function of FDA is a linear function of the second order statistics (L_2-norm) of the data, a feature shared by many other SDR algorithms as will be shown in Section 3.2. An advantage of the L_2-norm objective functions is that the solution is in closed form and can be solved by eigen-decomposition. The drawback, however, is that the retrieved subspace is suboptimal for multi-class classification or regression problems[1], as illustrated in the following toy example.

Consider a dataset with four classes located in $\mathbb{R}^3$ (see Figure 3.1). The data are generated from a mixture of four Gaussian distributions with a common identity covariance matrix and different mean vectors located at a, b, c, and d, where d is relatively far away from the others. The goal of SDR here is to transform the data into a lower-dimensional subspace while maintaining the maximum discriminative information. When only one dimension is allowed to be kept, the direction chosen by FDA is well aligned with Z-axis, which separates all the classes except classes a and b. In the remaining XY-plane, FDA roughly chooses the X-axis as the second most important direction, due to the large between-class distance of class c and the others. This leaves class a and b being still mixed together. Alternatively, if Y-axis were chosen as the second direction in the reduced space, then class a and b could be separated. Apparently FDA failed to select the direction which best discriminates the response variable.

The sub-optimality of FDA for multi-class classification is due to the discrepancy between the objective function and the classification accuracy: classification accuracy does not increase *linearly* with respect to the between-class distance. In other words, a direction that keeps large between-class distance may not necessarily result in optimal separation of multiple classes. However, the objective function used by FDA, as well as any objective function that is a linear function of the L_2-norm of the data, tends to overemphasize those directions

[1]Although FDA is designed for classification, we can apply it on regression problems too, simply by discretizing the continuous response into multiple categories like what has been done in the sliced inverse regression (Li, 1991). So in the remaining part of the paper, we will just focus on the classification setting.

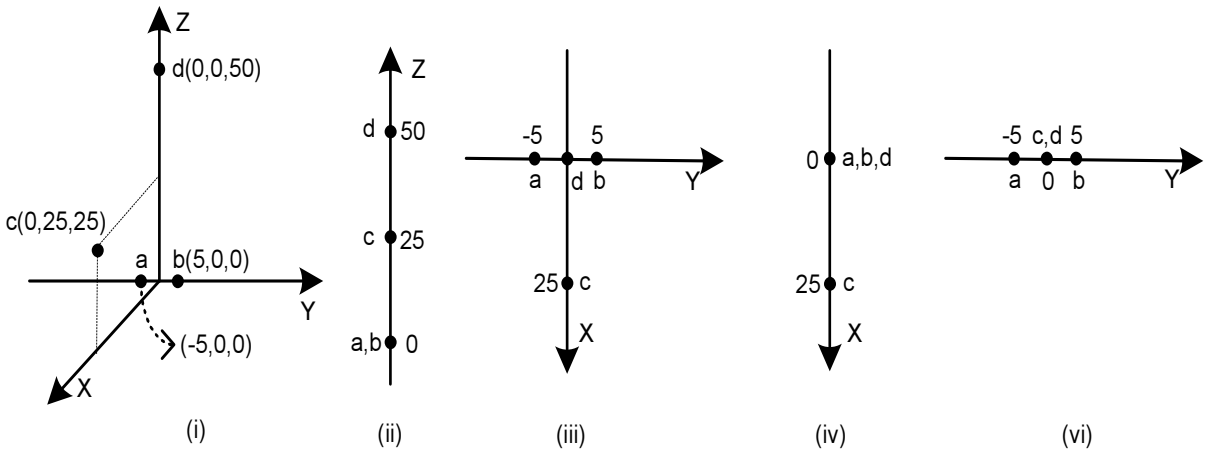


Figure 3.1: The 3-dimensional data consists of 4 classes, each of them containing 10000 data points. We generate each class of data from a Gaussian distribution with an identity covariance but different mean vectors denoted by $a(-5, 0, 0)$, $b(5, 0, 0)$, $c(0, 25, 25)$ and $d(0, 0, 50)$ in (i) respectively. Each dot represents a whole class of data. The projection of the 3-dimensional data onto Z-axis, XY-plane, X-axis and Y-axis are shown in (ii), (iii), (iv) and (vi) respectively.

already achieving large between-class distances yet making little improvement over the classification accuracy. In contrast, such methods may overlook the directions leading a small margin of between-class distances but a big improvement over the classification accuracy, especially if such directions can discriminate classes that have not been well separated yet.

To address this issue, we choose to work with two objective functions that are directly linked to the classification accuracy of the projected data $\mathbf{V}^t X$. We further generalize the two specific objective functions and propose a generalized optimization framework to solve the reduced dimensional subspace. The generalized objective function contains many popular SDR methods as special cases, including FDA, Local FDA (Sugiyama, 2007), and LPP (He and Niyogi, 2003). The challenge here is the objective function, which may be a nonlinear function of the L_2 norm of the data, cannot be solved by a simple eigendecomposition any more. In Section 3.3, we present an algorithm that sequentially solves the nonlinear objective function. The key motivation of this algorithm is that each sequentially added direction should boost the discriminative power of the reduced space. Specifically, when retrieving the l-th direction v_l, we update the objective function $G_l(v)$ so that v_l complements the

previously solved directions $v_1, \cdots, v_{l-1}$ in terms of classification accuracy. This is why we term our new algorithm as *Complementary* Dimension Analysis (CDA). We evaluate CDA on several simulated datasets and real world datasets in Section 3.4, and close with discussion and conclusions in Section 3.5.

Before closing this section, we make some remarks on related work.

- Loog et al. (2001) has pointed out the sub-optimality of FDA, and proposed a new objective function based on re-weighting, called approximate pairwise accuracy criterion (aPAC). Their approach differs from ours: 1) their approach heavily relies on the parametric Gaussian assumption and does not consider the non-parametric case, and 2) they didn't consider updating the weights at each step, therefore the optimality of directions, except the 1st one, returned by aPAC cannot be justified in their framework. Nevertheless, our work is indeed motivated by Loog et al. (2001).

- The idea of modifying our utility from a linear function of the L_2 norm of the data to a nonlinear one also appears in many manifolds learning algorithms. For example, in the aforementioned Local FDA, Sugiyama (2007) proposed to down-weight the contribution to the calculation of $\mathbf{B}$ and $\mathbf{W}$ from data pairs with large L_2 distance. However, the weights, which are nonlinear functions of the L_2 distance, are calculated in the original space and do not depend on the projection matrix $\mathbf{V}$. So the objective function of Local FDA is still a linear function of the L_2 norm, with the data points being weighted differently a priori.

3.2 A Unified SDR Framework for Classification

A major motivation of our work is to find the reduced dimension subspace directly guided by maximizing the accuracy of classification which is performed in this subspace. The classification accuracy, however, may comes at different forms, leading to different solutions to the optimization. Next we will introduce two criteria measuring the accuracy of multi-

class classification: one for parametric case and the other for non-parametric case. We then demonstrate our unified SDR framework and show that it not only accommodates these two particular measures, but also generalizes several existing dimension reduction methods such as FDA, Local FDA and SIR, which are made special cases in our framework with proper specification.

3.2.1 Parametric measure of classification accuracy

As a starting point, assume the data of each class follow a Gaussian distribution with a common covariance shared by all classes. Let m_k, W_k and p_k denote the mean, the within-class covariance and the prior of class k ($k \in \{1, \ldots, K\}$) respectively. The Bayes accuracy between class k and k' ($k \neq k'$) in a reduced space projected by a projection matrix $\mathbf{V}$ is given by

$$A_{kk'}(\mathbf{V}) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{d^{\mathbf{V}}_{kk'}}{2\sqrt{2}}\right), \tag{3.2}$$

where $d^{\mathbf{V}}_{kk'} = \|\mathbf{V}^t\mathbf{W}^{-\frac{1}{2}}(m_k - m_{k'})\|$ is the mean distance of class k and k' in the reduced space, $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the normal error function, and $\mathbf{W} = \sum_{k=1}^K p_k\mathbf{W}_k$ is the pooled within class variance. The detailed deduction of the Bayes accuracy is provided in Appendix A. By rescaling the data by $\mathbf{W}^{-\frac{1}{2}}$, we can assume the pooled within class variance is the identity matrix. Then the between class distance can be simplified as $d^{\mathbf{V}}_{kk'} = \|\mathbf{V}^t(m_k - m_{k'})\|$.

Write projection matrix at the l-th step as $\mathbf{V}_l = [v_1, \ldots, v_{l-1}, v]$. Here v_1 to v_{l-1} are calculated from the previous $l-1$ steps, and v is the new direction to solve subject to $v \perp M_{l-1}$ where M_{l-1} is the subspace spanned by $v_1, \ldots, v_{l-1}$. The averaged pairwise Bayes accuracy

$$G_B(\mathbf{V}_l) = \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} p_k p_{k'} A_{kk'}(\mathbf{V}_l) \tag{3.3}$$

can be used as a criterion to measure the contribution of a direction v as an addition to the existing $(l-1)$-dimensional subspace. Note that the averaged two-class classification accuracy is not the same as the actual accuracy of multi-class classification. Nevertheless, it

provides a reasonable quantification of the discriminative power of a new direction v.

3.2.2 Non-parametric measure of classification accuracy

In some applications, it may not be realistic to make parametric assumptions about the location or shape of the data. In such a scenario, classification procedures based on the neighboring information are more robust. Therefore we introduce a non-parametric measure of classification accuracy. Goldberger et al. (2005) proposed a novel method for learning a Mahalanobis distance measure which optimizes the expected leave-one-out classification error on the training data when used with a stochastic neighbor selection rule. Here we adopt the similar distance measure which aims to optimize the classification error rate for n-nearest neighbors.

First define the following similarity measure between any pair of data points

$$A(x, z) = \exp\left(-\|x - z\|^2/\epsilon\right), \quad A(x, x) = 0, \tag{3.4}$$

where $\epsilon > 0$ is a pre-specified scale parameter. Then consider a modified version of the simple n-nearest neighbors classification rule: given any query data point x^*, classify it as class k with probability proportional to $\sum_{i:y_i=k} A(x^*, x_i)$.

So at the l-th step when we try to solve v_l, the leave-one-out prediction accuracy in the reduce space is measured by

$$\begin{aligned} G_{NN}(\mathbf{V}_l) &= \frac{\sum_{k=1}^{K} \sum_{(i,j),y_i=y_j=k} A(\mathbf{V}_l^t x_i, \mathbf{V}_l^t x_j)}{\sum_{(i,j)} A(\mathbf{V}_l^t x_i, \mathbf{V}_l^t x_j)} \\ &= 1 - \frac{\sum_{(i,j),y_i \neq y_j} A(\mathbf{V}_l^t x_i, \mathbf{V}_l^t x_j)}{\sum_{(i,j)} A(\mathbf{V}_l^t x_i, \mathbf{V}_l^t x_j)}, \end{aligned} \tag{3.5}$$

where $\mathbf{V}_l = [v_1, \ldots, v_{l-1}, v]$ is defined before. The equalities above indicate that maximizing the objective function $J_{NN}(v)$ is equivalent to maximizing the within-class similarity and meanwhile minimizing the between-class similarity.

3.2.3 A unified framework

In the following we propose a unified framework for supervised dimension reduction in classification. The framework involves a general form of objective function parameterized by the unknown reduced dimension subspace. The objective function accommodates a wide variety of forms of classification accuracy, such as the two accuracy measures introduced above. Interestingly, quite a few existing supervised dimension reduction methods, such as FDA, Local FDA, SIR, and the unsupervised techniques such as PCA and LPP (He and Niyogi, 2003) can be seen as special cases within this framework, with properly chosen index sets, weights and functions. The computation of the optimization will be explained in detail in Section 3.3.

Suppose we have a dataset represented as a $p \times n$ matrix $\mathbf{X}$ consisting of n data points $x_i \in \mathbb{R}^p$ $(i \in \{1, \ldots, n\})$ and a set of labels $y_{n\times 1} = (y_1, \ldots, y_n)$ where y_i takes a value from 1 to K. The general objective function of SDR is constructed in the following form:

$$G(\mathbf{V}) = \frac{\sum_{I\in\mathcal{I}} f(\|\mathbf{V}^t X h_I\|^2) w_I}{\sum_{J\in\mathcal{J}} \tilde{f}(\|\mathbf{V}^t X \tilde{h}_J\|^2) \tilde{w}_J}, \tag{3.6}$$

where $\mathcal{I}$ and $\mathcal{J}$ are two summation index sets, h_I and $\tilde{h}_J$ are $n \times 1$ vectors indicating weights on the n data points, and constants w_I and $\tilde{w}_J$ are index-dependent weights on the functions f and $\tilde{f}$ respectively. When $\mathcal{I}$ or $\mathcal{J}$ is empty, we set the corresponding sum to be 1. The objective function is constructed using the projected data $\mathbf{V}^t X$ to establish direct connection with the classification accuracy in the reduced space. Then the optimal reduced dimension subspace can be obtained by maximizing $G(\mathbf{V})$, that is,

$$\mathbf{V} = \underset{\mathbf{V}, v_i^t v_i = 1}{\arg\max}\, G(\mathbf{V}). \tag{3.7}$$

Quite a few dimension reduction methods can be written as a special case of our general framework, such as FDA, Local FDA, SIR and LPP. We demonstrate only with FDA here

and more examples will be shown in Appendix B. Recall the notation in FDA that the between-class covariance $\mathbf{B} = \sum_{k=1}^{K} p_k(m_k - \bar{m})(m_k - \bar{m})^t$ and within-class covariance $\mathbf{W} = \sum_{k=1}^{K} p_k W_k$, where p_k, m_k, and W_k are the frequency, within-class mean, and within-class covariance of class k respectively. By rescaling the data by $\mathbf{W}^{-\frac{1}{2}}$, we can assume the pooled within class variance is the identity matrix. Then the objective function of FDA is $v^t \mathbf{B} v$ which can be written in the form of (3.6) by setting $\mathcal{I} = \{1, \ldots, K\}$, $\quad f(t) = t$, $\quad w_I = p_I$, h_I as a vector with the ℓ-th element equal to $\frac{(1-p_I)}{n_I}\mathbf{1}_{\{y_\ell = I\}} - \frac{p_{y_\ell}}{n_{y_\ell}}\mathbf{1}_{\{y_\ell \neq I\}}$, and $\mathcal{J}$ as an empty set, where n_k denotes the number of observations in class k.

The objective function for FDA can be easily solved via eigen-decomposition. As we will show in the next section, similar results hold true for dimension reduction algorithms where f's and $\tilde{f}$'s are linear functions of the L_2 norm of the data, namely, $\|v^t \mathbf{X} h_I\|^2$. However, there is no analytical solution of the objective function when f' and $\tilde{f}$ are nonlinear. We address this issue in in the following section.

3.3 Algorithm

In this section we present an optimization approach to the proposed unified SDR framework for classification. For notation simplicity, we denote the objective function $G(\mathbf{V})$ by $F(\mathbf{V})/\tilde{F}(\mathbf{V})$. As we show next, when f and $\tilde{f}$ are linear functions, we can pre-normalize the data to get rid of the denominator. Then the objective function can be solved easily. When f and $\tilde{f}$ are nonlinear functions, we propose a numerical method to sequentially solve the directions in the order of decreasing importance. Specifically, at the l-th step, given the previously solved $(l-1)$ directions $(v_1, \ldots, v_{l-1})$, we solve the following optimization problem

$$v_l = \underset{v \perp M_{l-1}}{\arg\max}\, G_l(v), \tag{3.8}$$

$G_l(v) = G(\mathbf{V}_l)$ with $\mathbf{V}_l$ defined as $[v_1, \ldots, v_{l-1}, v]$. It turns out that for the linear case, the solution given by this sequential approach agrees with the global solution.

3.3.1 Linear functions

Assume the functions $f(\|\mathbf{V}^t X h_I\|^2)$ and $\tilde{f}(\|\mathbf{V}^t X \tilde{h}_J\|^2)$ take the following linear forms:

$$f(\|\mathbf{V}^t X h_I\|^2) = a_I \|\mathbf{V}^t X h_I\|^2 \quad \text{and} \quad \tilde{f}(\|\mathbf{V}^t X \tilde{h}_J\|^2) = \tilde{a}_J \|\mathbf{V}^t X h_I\|^2,$$

where a_I and $\tilde{a}_J$ are non-zero constants. Then simple calculations reveal that $F(\mathbf{V})$ and $\tilde{F}(\mathbf{V})$ can be rewritten as $\mathrm{tr}(\mathbf{V}^t S_1 \mathbf{V})$ and $\tilde{F}(\mathbf{V}) = \mathrm{tr}(\mathbf{V}^t S_2 \mathbf{V})$ respectively where $S_1 = \sum_{I \in \mathcal{I}} a_I w_I X h_I {h_I}^t X^t$ and $S_2 = \sum_{J \in \mathcal{J}} \tilde{a}_J X \tilde{h}_J \tilde{h}_J^t X^t \tilde{w}_J$. In this case, we can pre-normalize the data by $S_2^{-\frac{1}{2}}$, then the objective function is

$$G(\mathbf{V}) = \mathrm{tr}\left(\mathbf{V}^t S_2^{-\frac{1}{2}} S_1 S_2^{-\frac{1}{2}} \mathbf{V}\right). \tag{3.9}$$

Note that the orthonormal matrix $\mathbf{V}$ that maximizes the $G(\mathbf{V})$ defined above is different from the solution of (3.6), due to the pre-normalization procedure. However, the subspaces spanned by the two matrices are the same. So it suffices for us to maximize the objective function $G(\mathbf{V})$ defined in (3.9). This becomes a generalized eigen-decomposition problem and the solutions $v_1, \ldots, v_m$ are given by the eigenvectors corresponding to the top m eigenvalues of $S_2^{-\frac{1}{2}} S_1 S_2^{-\frac{1}{2}}$.

One could also solve the directions using the sequential algorithm described previously. It is easy to check that the first direction v_1 is given by the largest eigenvalue of $S_2^{-\frac{1}{2}} S_1 S_2^{-\frac{1}{2}}$. Similarly we can find the next direction v, orthogonal to v_1, such that $v^t S_2^{-\frac{1}{2}} S_1 S_2^{-\frac{1}{2}} v$ is maximized and the solution v_2 is given by the second largest eigenvalue of $S_2^{-\frac{1}{2}} S_1 S_2^{-\frac{1}{2}}$, and so on. Finally, the columns of $\mathbf{V}$ are given by the first m largest eigenvalues of $S_2^{-\frac{1}{2}} S_1 S_2^{-\frac{1}{2}}$, which is indeed equivalent to the global solution.

3.3.2 Nonlinear functions

As mentioned before, there is no closed-form analytical solution to solve the multiple directions simultaneously through the objective function (3.6) when the functions f and $\tilde{f}$ are nonlinear. An intuitive way is to retrieve the multiple directions sequentially. However, the sequentially independent approach is not a good choice here as it ignores the influence of the previously found directions. Instead, we use a sequentially dependent approach where the objective function dependents on the previously found directions and therefore is updated at each step. In this way, each new retrieved direction is complementary to the existing directions. Furthermore, we adopt local linearization to approximate the nonlinear functions by linear functions to keep the simple eigen-decomposition solution as in the case of linear functions. The major procedures can be summarized as follows:

- Write the objection function at the l-th step as in Equation (3.8) which depends on the previously found $(l-1)$ directions;
- Approximate the nonlinear functions f and $\tilde{f}$ through linear formulas through local linearization;
- Solve for v_l by applying eigen-decomposition to the approximated objective function.

Recall that $\mathcal{M}_l$ is the subspace spanned by $v_1, \ldots, v_{l-1}$. Then $\mathrm{M}_l = \sum_{i=1}^{l-1} v_i v_i^t$ is the projection matrix of $\mathcal{M}_l$ and $\mathrm{I}_p - \mathrm{M}_l$ is the projection matrix for $\mathcal{M}_l^{\perp}$, the orthogonal space of $\mathcal{M}_l$. Using the property that $X = \mathrm{M}_l X + (\mathrm{I}_p - \mathrm{M}_l)X$, the objective function of v_l for our sequentially dependent approach is

$$\arg\max_{v} \frac{\sum_{I\in\mathcal{I}} f_I(\|\mathrm{M}_l X h_I\|^2 + \|v^t(\mathrm{I}_p - \mathrm{M}_l)X h_I\|^2)w_I}{\sum_{J\in\mathcal{J}} \tilde{f}_J(\|\mathrm{M}_l X \tilde{h}_J\|^2 + \|v^t(\mathrm{I}_p - \mathrm{M}_l)X \tilde{h}_J\|^2)\tilde{w}_J} \tag{3.10}$$

Note that the constraint $v \in \mathcal{M}_l^{\perp}$ is taken into account by the term $(\mathrm{I}_p - \mathrm{M}_l)$.

When the functions f and $\tilde{f}$ are nonlinear, the eigen-decomposition method cannot be applied directly to the above objective function. However, it is still desirable to keep

the simple form of eigen-decomposition solution. Therefore, we approximate the nonlinear functions through linear functions such that the computation can be carried out by eigen-decomposition and incorporate the previously selected directions to make the searching algorithm efficient. Specifically, to solve v_l, we first transform the sequential dependent Equation (3.10) as linear functions through local linearization under minimum mean squared error criteria and then apply eigen-decomposition method to derive the projection direction.

Here we only show the approximation procedure for the numerator $F(\mathbf{V}_l)$ in (3.10). Similar approximation can be easily applied to the denominator $\tilde{F}(\mathbf{V}_l)$. Define vector $e_{Il} = (\mathrm{I}_p - \mathrm{M}_l)Xh_I$. Then $F(\mathbf{V}_l) = \sum_{I\in\mathcal{I}} F_I(\mathbf{V}_l)$ in (3.10), by local linearization, can be approximated by the following linear form

$$A(\mathbf{V}_l) = \sum_{I\in\mathcal{I}} A_I(\mathbf{V}_l) = \sum_{I\in\mathcal{I}} a_{Il} v^t e_{Il} e_{Il}^t v + b_{Il}, \tag{3.11}$$

where a_{Il} and b_{Il} are unknown constants we need to estimate later.

Similarly, the denominator $\tilde{F}(\mathbf{V}_l)$ can be approximated by

$$\tilde{A}(\mathbf{V}_l) = \sum_{J\in\mathcal{J}} \tilde{A}_G(\mathbf{V}_l) = \sum_{J\in\mathcal{J}} \tilde{a}_{Jl} v^t \tilde{e}_{Jl} \tilde{e}_{Jl} v + \tilde{b}_{Jl}, \tag{3.12}$$

where $\tilde{a}_{Jl}$ and $\tilde{b}_{Jl}$ are unknown constants and $\tilde{e}_{Jl} = (\mathrm{I}_p - \mathrm{M}_l)X\tilde{h}_J$. Then the objective function is approximated by:

$$\arg\max_v \frac{\mathrm{tr}\,(v^t S_1 v)}{\mathrm{tr}\,(v^t S_2 v)}, \tag{3.13}$$

where $S_1 = \sum_{I\in\mathcal{I}} a_{Il} e_{Il} e_{Il}^t + b_{Il} I_p$ and $S_2 = \sum_{J\in\mathcal{J}} \tilde{a}_{Jl} \tilde{e}_{Jl} \tilde{e}_{Jl}^t + \tilde{b}_{Jl} I_p$. Therefore, the solution v_l is give by applying eigen-decomposition of $S_2^{-1} S_1$.

Next we consider choosing parameters a_{Il} and b_{Il} that minimize the distance of $F(\mathbf{V}_l)$ and $A(\mathbf{V}_l)$ for all possible v's. First define $\alpha_{Il} \in (-\pi/2, \pi/2]$ as the angle between v and the vector e_{Il}, so $\cos\alpha_{Il} = v^t e_{Il}/\|e_{Il}\|$. Then we reformulate $F_I(\mathbf{V}_l)$ and $A_I(\mathbf{V}_l)$ as a function of

$\cos^2 \alpha_{Il}$:

$$F_I(\cos^2 \alpha_{Il}) = f\left(\|\mathrm{M}_l X h_I\|^2 + \|e_{Il}\|^2 \cos^2 \alpha_{Il}\right) w_I,$$
$$A_I(\cos^2 \alpha_{Il}) = a_{Il} \cos^2 \alpha_{Il} + b_{Il}.$$

A simple approach to finding a_{Il} and b_{Il} is through the first order Taylor expansion of $F_I(\cos^2 \alpha_{Il})$ at some particular value of $\cos^2 \alpha_{Il}$, which is used in the optimization process for parametric and non-parametric multi-class classification problems in the next section. Rigorously, we solve for the parameter a_{Il} and b_{Il} by the minimum mean squared error of $F_I(\cos^2 \alpha_{Il})$ and $A_I(\cos^2 \alpha_{Il})$, that is

$$\arg \min_{a_{Il},\, b_{Il}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \left(a_{Il} \cos^2 \alpha + b_{Il} - f_I\left(\|\mathrm{M}_l X h_I\|^2 + \|e_{Il}\|^2 \cos^2 \alpha\right) w_I\right)^2 \mathrm{d}\alpha. \tag{3.14}$$

3.3.3 Two examples of optimization

In the following we give the solutions to the optimization using the parametric classification accuracy measure and non-parametric measure introduced in Section 3.2.1.

Example 1: Parametric classification accuracy measure. The objective function is given in (3.3) and the corresponding settings in our general framework are given in Appendix B. With the non-linear error function involved, the solution of the exact minimum mean square error is analytically infeasible. We approximate $G(\mathbf{V}_l) = F_I(\cos^2 \alpha_{Il})$ by its first order Taylor expansion and then solve for the parameters by minimizing the mean square error in Equation (3.14). The Taylor expansion is performed at the point $\cos^2 \alpha_{Il} = 1$ as shown below:

$$\begin{aligned} F_I(\cos^2 \alpha_{Il}) &= w_I \left[\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{1}{2\sqrt{2}}(\|\mathrm{M}_l X h_I\|^2 + \|e_{Il}\|^2 \cos^2 \alpha_{Il})^{1/2}\right)\right] \\ &\approx F_I(1) + F_I'(1)\left(\cos^2 \alpha_{Il} - 1\right) \\ &= w_I \left[\frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\|X h_I\|}{2\sqrt{2}}\right)\right] + \frac{w_I}{4\sqrt{2\pi}} \frac{\|e_{Il}\|^2}{\|X h_I\|} \exp\left(-\frac{\|X h_I\|^2}{8}\right)(\cos^2 \alpha_{Il} - 1), \end{aligned} \tag{3.15}$$

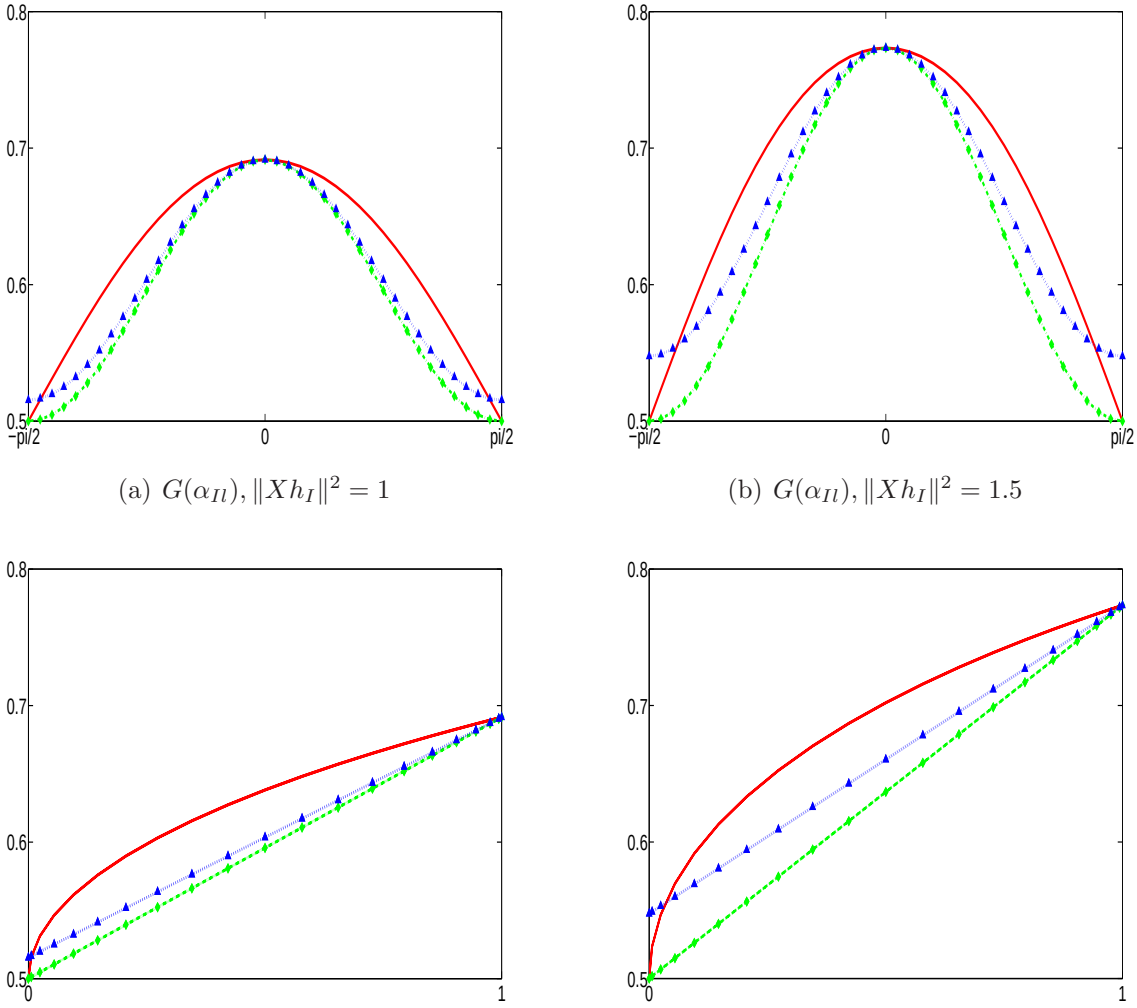
where $F_I'(1)$ denotes the first derivative of $F_I(t)$ at the point $t = 1$. Then the optimal values for the parameters a_{Il} and b_{Il} of $A_I(\cos^2 \alpha_{Il})$ would be:

$$\begin{aligned} a_{Il} &= \frac{w_I}{4\sqrt{2\pi}} \frac{\|e_{Il}\|^2}{\|Xh_I\|} \exp\left(-\frac{\|Xh_I\|^2}{8}\right) \\ b_{Il} &= w_I \left[\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\|Xh_I\|}{2\sqrt{2}}\right)\right] - \frac{w_I}{4\sqrt{2\pi}} \frac{\|e_{Il}\|^2}{\|Xh_I\|} \exp\left(-\frac{\|Xh_I\|^2}{8}\right). \end{aligned}$$

To get more precise approximation of $F_I(\cos^2 \alpha_{Il})$, we can use piecewise Taylor expansion which first divides the value of $\cos^2 \alpha_{Il}$ into several pieces and then takes the Taylor expansion for each piece. As the number of pieces grows, we get better and better approximation. Here we illustrate using the simple one-piece Taylor expansion.

Loog et al. (2001) proposed an approximate pairwise accuracy criterion (aPAC) to approximate the Bayesian accuracy in (3.3) by using the linear approximation function $\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\|Xh_I\|}{2\sqrt{2}}\right)\cos^2(\alpha_{Il})$ where the corresponding parameters are $a_{Il} = \frac{1}{2}\text{erf}\left(\frac{\|Xh_I\|}{2\sqrt{2}}\right)$ and $b_{Il} = \frac{1}{2}$. With such a linear function, the approximate objective function equals the Bayesian criteria at the particular points $\alpha_{Il} = 0, \frac{\pi}{2}, \pi$. We can see there are three major differences between aPAC and the proposed CDA: (i) the chosen linear approximation function of aPAC only matches the true Bayesian accuracy at some particular positions where $\alpha_{Il} = 0, \frac{\pi}{2}, \pi$, which tends to arise large errors in the other positions; whereas our method is trying to minimize the mean square error over all the positions jointly. (ii) aPAC always assume the intercept of the approximation to be 1/2, which restricts the performance of approximation; while CDA allows for arbitrary intercept. (iii) aPAC solves for multiple directions in a sequentially independent manner which ignores the influence of the previously found directions; whereas our method finds the multiple dimensional in a complimentary manner by updating the objective function at each step.

The approximation curves $A_I(\cos^2 \alpha_{I1})$ of our method CDA and aPAC against the true function curve $F_I(\cos^2 \alpha_{I1})$ when solving for the first direction v_1 are shown in Figure 3.2. The figure indicates that our method approximates the Bayesian accuracy better than aPAC



(a) $G(\alpha_{Il}), \|Xh_I\|^2 = 1$

(b) $G(\alpha_{Il}), \|Xh_I\|^2 = 1.5$

(c) $G(\cos^2(\alpha_{Il})), \|Xh_I\|^2 = 1$

(d) $G(\cos^2(\alpha_{Il})), \|Xh_I\|^2 = 1.5$

Figure 3.2: Comparison of the true objective function (solid line) in (3.3) and the approximated linear functions by CDA (dotted line) and aPAC (dashed line) versus angle α (as shown in (a) and (b)) and $\cos^2(\alpha)$ (as shown in (c) and (d)) when solving for the first direction v_1. Figures (a) and (c) are plotted with $\|Xh_I\|^2 = 1$, while figures (b) and (d) are plotted with $\|Xh_I\|^2 = 1.5$.

and are stable with different class distances $\|Xh_I\|^2$. On the other hand, aPAC tends to underestimate the true function when solving for the first projection vector. This is due to the restriction that the approximation function should match the true function at the three particular points and that the intercept is always set to 1/2. As the class distance D_{I1} becomes larger, the performance of aPAC is worse. Moreover, the improvement of CDA becomes more obvious for the second projection direction because we apply the sequentially dependent searching strategy. More illustration will be shown in the experiment section.

In addition, it's worth pointing out that the objective function of our method after linear approximation can be viewed as a weighted FDA. Recall that the objective function of FDA at the l-th step is given by:

$$G_l(v) = \sum_{I \in \mathcal{I}} a_{Il} \ \mathrm{tr}(v e_{Il} e_{Il}^t v^t), \tag{3.16}$$

where $\mathcal{I} = \{(i,j)\,; 1 \le i < j \le K\}$ and $a_{Il} = w_I = p_i p_j$. Assume the prior probabilities of all classes are the same, then $G_l(v)$ is simply the summation of all class-pair distance. We can rewrite our objective function after linear approximation in (3.15) in the same form of (3.16) with the coefficient $a_{Il} = \frac{w_I}{4\sqrt{2\pi}} \frac{\|e_{Il}\|^2}{\|Xh_I\|} \exp\left(-\frac{\|Xh_I\|^2}{8}\right)$. Besides the item w_I and the constant, we have one more item $\frac{1}{\|Xh_I\|} \exp\left(-\frac{\|Xh_I\|^2}{8}\right) \cdot \|e_{Il}\|^2$ where the left part is a decreasing function of $\|Xh_I\|$ and the right part is an increasing function of $\|e_{Il}\|$. Therefore, the large class-pair distance is weighted down by the first part and the importance of the remaining class-pair distance is controlled by the second part. Similarly, we can formulate aPAC into the form of (3.16) by setting $a_{Il} = w_I \frac{1}{2\|Xh_I\|^2} \mathrm{erf}(\frac{\|Xh_I\|}{2\sqrt{2}})$. The extra term $\frac{1}{2\|Xh_I\|^2} \mathrm{erf}(\frac{\|Xh_I\|}{2\sqrt{2}})$ is a decrease function of $\|Xh_I\|$, which shrinkages the contribution of large class distance. From this point of view, we can see that both aPAC and CDA improve the objective function of FDA by re-weighting and CDA further considers the influence of class distance on the remaining space at each step.

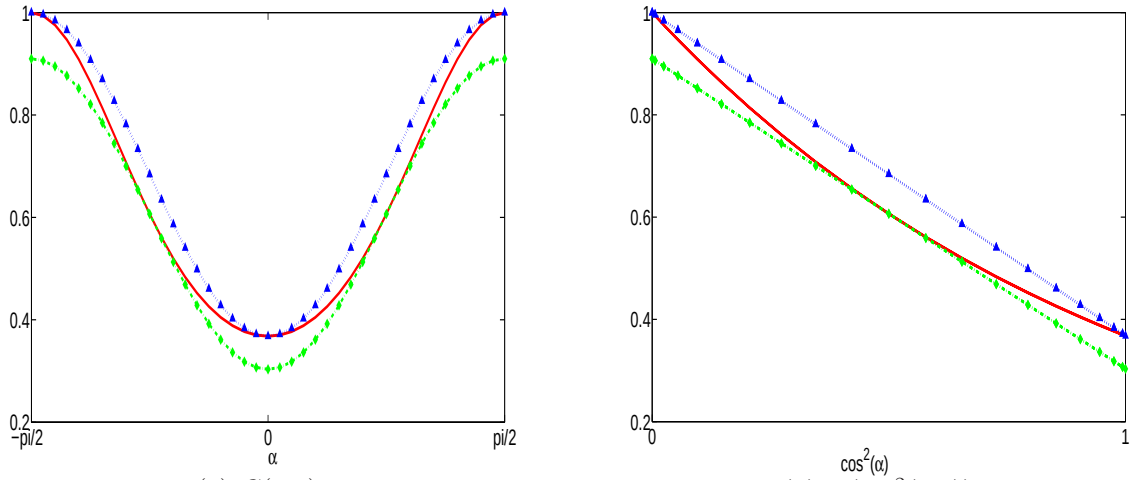
Example 2: Non-parametric classification accuracy measure. Following the same procedures as in the parametric case, we first take Taylor expansion of $F_I(\cos^2 \alpha_{Il})$ at point

$\cos^2(\alpha_{Il}) = \frac{1}{2}$, that is,

$$
\begin{aligned}
F_I(\cos^2 \alpha_{Il}) &= \exp\left(-\frac{\|M_l X h_I\|^2 + \|e_{Il}\|^2 \cos^2 \alpha_{Il}}{\epsilon}\right) \\
&\approx F_I\left(\frac{1}{2}\right) + F_I'\left(\frac{1}{2}\right)\left(\cos^2 \alpha_{Il} - \frac{1}{2}\right) \\
&= \left(1 + \frac{\|e_{Il}\|^2}{2\epsilon}\right) \exp\left(-\frac{\|M_l X h_I\|^2 + \|e_{Il}\|^2/2}{\epsilon}\right) \\
&\quad - \frac{\|e_{Il}\|^2}{\epsilon} \exp\left(-\frac{\|M_l X h_I\|^2 + \|e_{Il}\|^2/2}{\epsilon}\right) \cos^2 \alpha_{Il}. \qquad (3.17)
\end{aligned}
$$

Then we can derive a_{Il} and b_{Il} as:

$$
\begin{aligned}
a_{Il} &= -\frac{\|e_{Il}\|^2}{\epsilon} \exp\left(-\frac{\|M_l X h_I\|^2 + \|e_{Il}\|^2/2}{\epsilon}\right) \\
b_{Il} &= \left(1 + \frac{\|e_{Il}\|^2}{2\epsilon}\right) \exp\left(-\frac{\|M_l X h_I\|^2 + \|e_{Il}\|^2/2}{\epsilon}\right).
\end{aligned}
$$



(a) $G(\alpha_{Il})$ (b) $G(\cos^2(\alpha_{Il}))$

Figure 3.3: Comparison of the true objective function (solid line) in (3.5) and the approximated linear functions by CDA (dotted line) and aPAC (dashed line) versus angle α (as shown in (a)) and $\cos^2(\alpha)$ (as shown in (b)) when solving for the first direction v_1. For both plots, $\|Xh_I\|^2$ is set to be 1.

The approximation curve against the nonlinear objective function is shown in Figure 3.3.

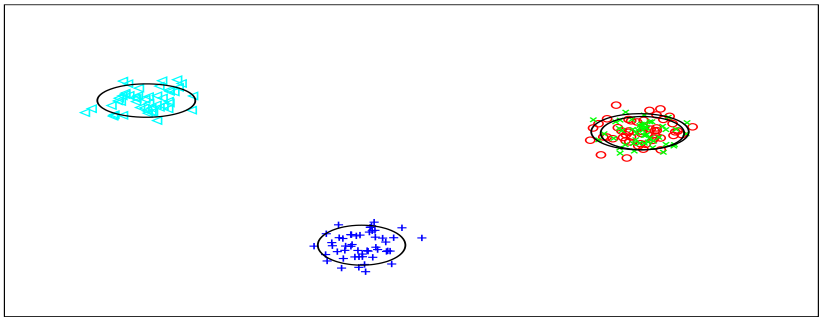
We can see that the approximation curve by one piece first-order Taylor expansion at point $1/2$ is very close to the objective function. The aPAC approximation in the figure has the form $1 + (\exp(-\|Xh_I\|^2/\epsilon) - 1)\cos^2\alpha_{Il}$ which matches the true objective function at points $\alpha_{Il} = 0, \frac{\pi}{2}, \pi$.

3.4 Experiments

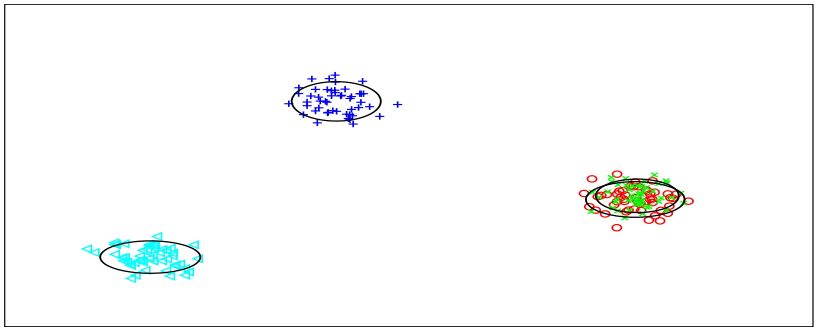
In this section, we evaluate the performance of our proposed optimization approach of the general framework on a synthetic dataset and several real data sets.

3.4.1 Toy example revisited

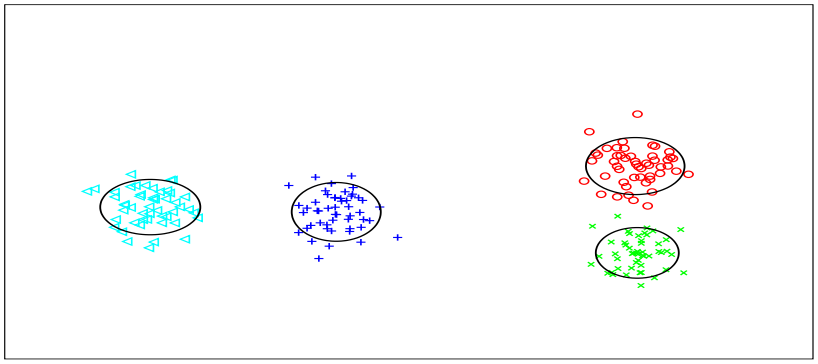
We first revisit the toy example introduced in Section 3.1. As discussed before, the solution given by FDA is not optimal due to the discrepancy between the objective function and the classification accuracy. Loog et al. (2001) proposed a new objective function named approximate pairwise accuracy criterion (aPAC) to deal with the sub-optimality of FDA. However, they did not incorporate the influence of the previously found directions in their algorithm. Our method CDA first reconstructs the objective function directly linked with the classification accuracy and then solves for the complementary directions sequentially. We compare the performance of these methods in Figure 3.4 by showing the data points in the 2-dim reduced space derived by these methods. It is clear that FDA separates two classes away from others and leaves classes a and b mixed together, which is consistent with the explanation in Section 3.1. Not surprisingly, aPAC also fails to separate these two classes even with the revised objective function. Only CDA successfully separates the four classes in the 2-dim reduced space.



(a) FDA



(b) aPAC



(c) CDA

Figure 3.4: Visualization of the data points in the 2-dim reduced subspace derived by FDA, aPAC and CDA for the toy example. We only show the randomly selected 50 data points in each class for a better view here.

3.4.2 Synthetic data

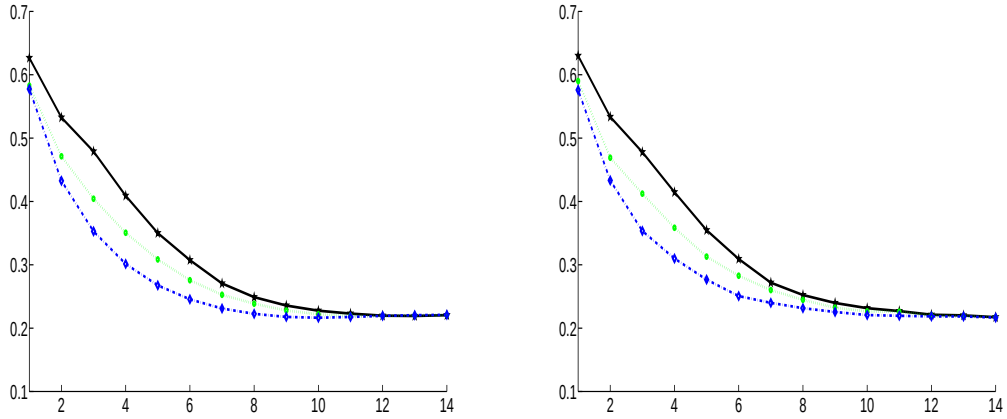
In the simulation study we generate $K = 15$ classes of data in $\mathbb{R}^{15}$. There are $N = 100$ data points in each class. Data points in each class are generated from a multivariate normal distribution with known mean μ_i $(i = 1, \ldots, 15)$ and covariance Σ. The covariance Σ, shared by all classes, is set as the sample variance of $2p$ random number generated from Unif$(-4, 4)$. The means for the 15 classes are generated from the following three different ways:

- C1: Normal distribution with mean 0 and standard deviation 1,
- C2: Log-normal distribution with mean 0 and standard deviation 1,
- C3: t distribution with degree of freedom 10.

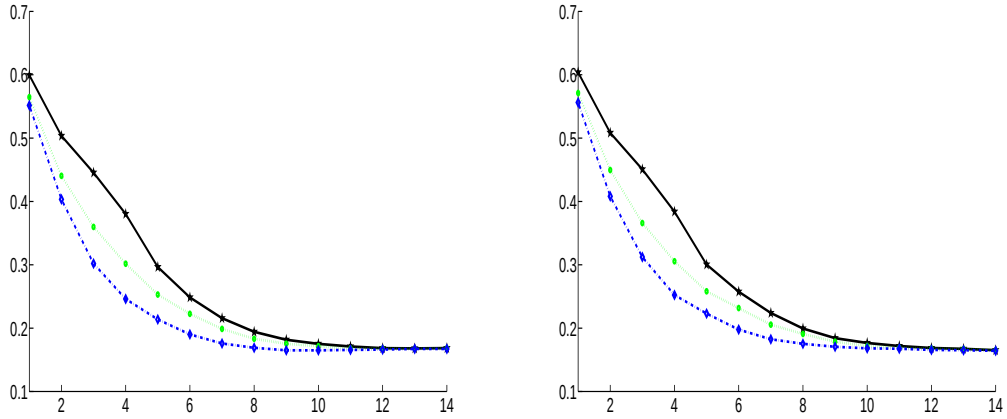
To test how our method handles data with outlier classes, we first randomly selected five classes and then replace one randomly chosen dimension mean with some large number for each of the five classes.

The reduced subspace with dimensionality m ranging from 1 to 14 is computed by learning the projection matrix $V_{m\times p}$ using our parametric measure in (3.3). The performance of the reduced subspace is measured by classification error, where we use the maximum a posterior classification based on the mixture of Gaussians. We use 70% of the whole dataset for training purpose and the rest for testing. Then the classification error rates averaged over 20 experiments with respect to the dimensionality of the reduced space based on FDA, aPAC and CDA on the training and testing data are reported. Figure 3.5 show the averaged classification error rates as a function of the subspace dimensionality ranging from 1 to 14. Table 3.1 lists the averaged classification error rates and the standard deviation of the 2-dim, 4-dim, 6-dim and 8-dim reduced space.

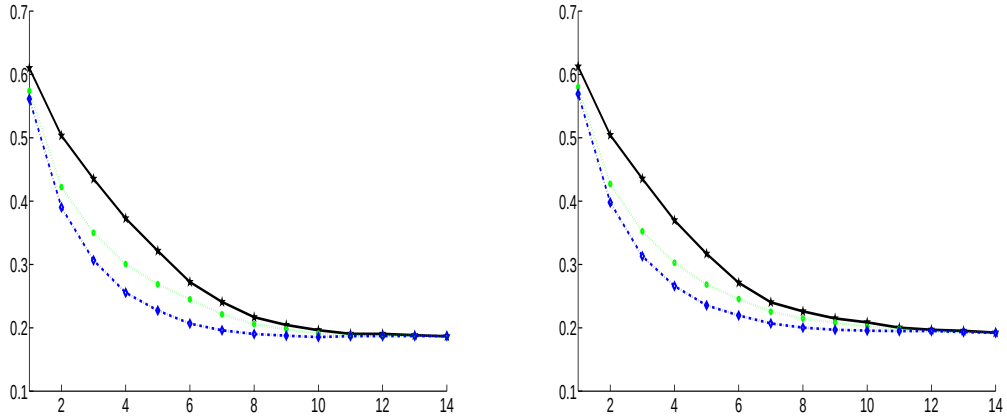
From these figures, we can see that our method CDA achieves the best performance compared to aPAC and FDA with subspace dimensionality ranging from 1 to 10 for both training and testing data sets. Take the 4-dim reduced space of the simulated data from C1 for example: FDA achieves 41.47% classification error rate, which is reduced to 35.84% by



(a) C1 (b) C1



(c) C2 (d) C2



(e) C3 (f) C3

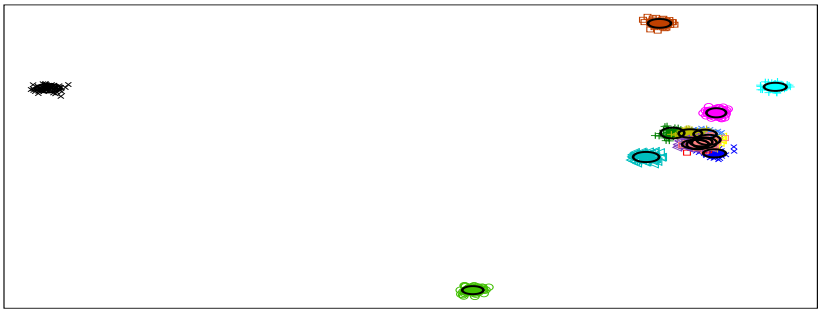
Figure 3.5: Averaged classification error rates as a function of the dimensionality of the reduced space for training data set (a, c, e) and testing data set (b, d, f) with 15 classes in 15 dimensional space generated by three different ways.

aPAC and further to 30.99% by CDA. All the three methods perform similarly with larger subspace dimensionality as there is little information lost in this situation. Moreover, there is no obvious difference of the performance for the three experiment settings.

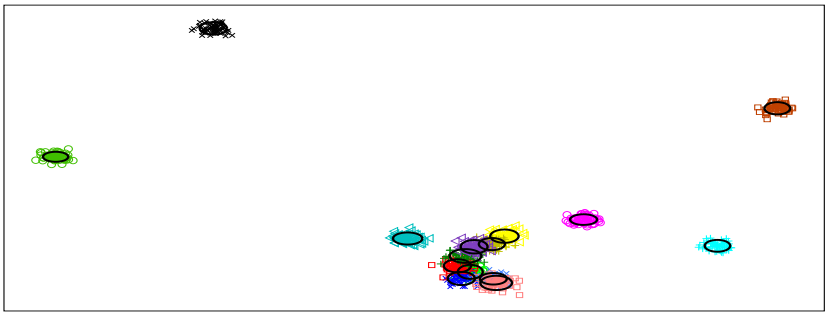
C1			
m	FDA	aPAC	CDA
2	0.5336 (0.0492)	0.4690 (0.0484)	0.4332 (0.0408)
4	0.4147 (0.0469)	0.3584 (0.0393)	0.3099 (0.0428)
6	0.3092 (0.0440)	0.2828 (0.0372)	0.2507 (0.0396)
8	0.2523 (0.0359)	0.2447 (0.0372)	0.2316 (0.0414)
C2			
m	FDA	aPAC	CDA
2	0.5083 (0.0583)	0.4496 (0.0503)	0.4081 (0.0704)
4	0.3839 (0.0442)	0.3056 (0.0645)	0.2523 (0.0607)
6	0.2576 (0.0585)	0.2318 (0.0558)	0.1978 (0.0581)
8	0.1994 (0.0590)	0.1909 (0.0601)	0.1753 (0.0536)
C3			
m	FDA	aPAC	CDA
2	0.5043 (0.0665)	0.4271 (0.0717)	0.3979 (0.0628)
4	0.3697 (0.0662)	0.3028 (0.0627)	0.2660 (0.0586)
6	0.2712 (0.0636)	0.2454 (0.0582)	0.2196 (0.0553)
8	0.2260 (0.0573)	0.2146 (0.0534)	0.2003 (0.0542)

Table 3.1: Averaged classification error rates and standard deviation of the testing sets over 20 experiments for the simulation data.

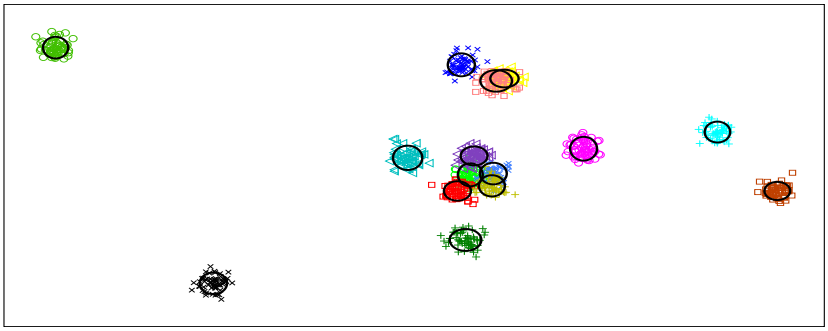
Similarly to the toy example, we also show the data points projected to the 2-dim reduced space given by FDA, aPAC and FDA in Figure 3.6. With the 15-class data, none of the methods successfully separates all of them. A direct way to compare the performance of these methods is to count how many classes are still mixed together. From the figure, we observe that FDA still leaves many class mixed together, aPAC separates more classes than FDA but still leaves quite a few classes overlapping with each other, while our method CDA separates most of the classes.



(a) FDA



(b) aPAC



(c) CDA

Figure 3.6: Visualization of the data points in 2-dim reduced subspace given by FDA, aPAC and CDA for the testing set generated by C1.

3.4.3 PIE database

We further investigate the performance of the proposed method on the CMU PIE database for face recognition. This database contains 41,368 face images from 68 individuals. For each individual, face images of varying pose, illumination, and expression are captured by 13 synchronized cameras under 21 flashes. We choose the five near frontal poses (C05, C07, C09, C27, C29) and use all such images under different illuminations, lighting and expressions, which leaves us 170 near frontal face images for each individual. The image size is 32×32, which gives a 1,024-dim feature. The dimension of each image is reduced to 120 by PCA as a preprocessing step for all the following experiments. A random subset with 50 images per individual is taken to form the training set, while the rest of the database is considered to be the testing set. We report the averaged classification error rates based on 1NN over 20 random splits in Fig. 3.7 and Table 3.2 where results obtained using FDA and aPAC are also reported respectively.

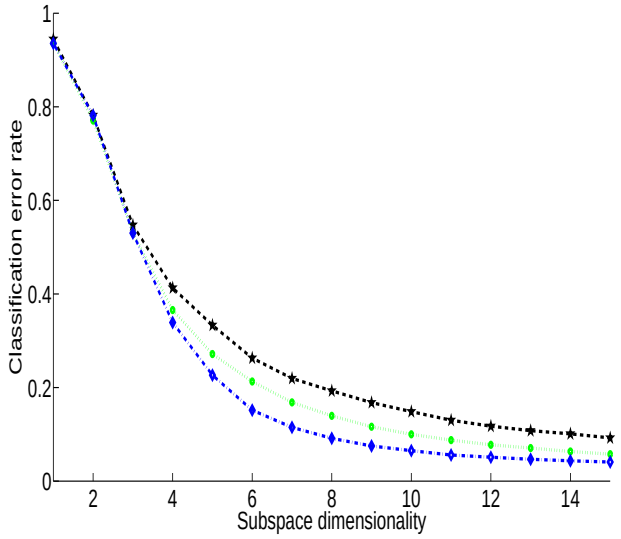


Figure 3.7: Classification error rate based on FDA (dashed line), aPAC (dotted line), and CDA (dash dotted line) versus the reduced dimension for the CMU PIE database.

For this real data set, the performance of CDA is comparable with those of FDA and aPAC when projected to the subspace with dimensionality ranging from 1 to 3. For ex-

d	FDA	aPAC	CDA
2	0.7818 (0.0207)	0.7706 (0.0078)	0.7815 (0.0092)
4	0.4128 (0.0108)	0.3655 (0.0150)	0.3390 (0.0167)
6	0.2631 (0.0158)	0.2129 (0.0273)	0.1517 (0.0125)
8	0.1928 (0.0084)	0.1395 (0.0162)	0.0915 (0.0062)
10	0.1484 (0.0089)	0.1000 (0.0071)	0.0651 (0.0052)
12	0.1175 (0.0065)	0.0775 (0.0055)	0.0511 (0.0035)
14	0.1011 (0.0068)	0.0633 (0.0042)	0.0434 (0.0035)

Table 3.2: Averaged classification error rates and standard deviation as a function of subspace dimensionality over 20 experiments for CMU PIE data set

ample, all the three methods can only achieve roughly 22% classification accuracy in the 3-dim reduced space. With larger subspace dimensionalities, the differences among the three methods become more apparent. When the dimensionality of the subspace is allowed to be 8, our method obtains an averaged classification error rate of 9.15%, followed by 13.95% for aPAC and 19.28% for FDA.

3.4.4 UCI data

In this part, we evaluate the performance of our criteria of non-parametric classification on the following four data sets from the UCI repository: a glass with 214 observations belonging to 6 classes, an E coli dataset with 336 proteins sequences each-labeled as one of the eight classes, a wine dataset including the quantities of 13 constituents for 178 instances found in each of the three types of wines, an ionosphere data with 34 continuous variables for 351 observations in 2 classes. For each data set, we use 70% of the observations for training and the left for testing. We report the averaged classification accuracy rates based on KNN over 20 random splits in Figure 3.8 and Table 3.3 when projected to the 2-dim subspace.

Figure 3.8 shows that our proposed method CDA outperforms PCA and FDA consistently and obtains comparable results with NCA in all the four data sets. Note that NCA used a gradient based optimizer to solve for the projection directions while our method applies a simple approximation and derives the projection directions by eigen-decomposition.

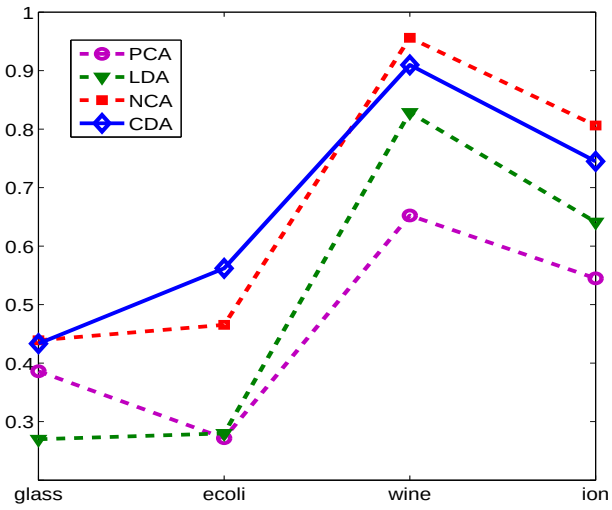


Figure 3.8: Classification accuracy rates in 2-dim subspace given by PCA, FDA, aPAC and CDA for chosen UCI data sets over 20 experiments.

glass	ecoli	wine	ion
0.3862 (0.0382)	0.2700 (0.0224)	0.4391 (0.0506)	0.4334 (0.0511)
0.2718 (0.0285)	0.2800 (0.0300)	0.4655 (0.0789)	0.5620 (0.0600)
0.6524 (0.0397)	0.8284 (0.0448)	0.9560 (0.0323)	0.9098 (0.0427)
0.5449 (0.0243)	0.6411 (0.0291)	0.8062 (0.0472)	0.7448 (0.0382)

Table 3.3: Averaged classification accuracy rates and standard deviation on the four UCI data sets over 20 experiments.

Therefore, we can get very good performance using simple optimization approach in this experiment.

3.5 Discussion and Conclusions

In this paper we proposed a unified dimension reduction framework Complementary Dimension Analysis for classification. Two objective functions are introduced which are directly linked to the classification accuracy of the projected data onto the reduced dimension subspace. We then presented a unified objective function which generalizes the two particular objective functions. A numerical algorithm is proposed to solve the proposed general opti-

mization problem. Wide connections are established between our general framework with existing dimension reduction methods including PCA, FDA, Local FDA, etc. Experiments on simulated data and real data demonstrate superior performance of the proposed method.

Appendix A

Bayes Accuracy for Binary Classification

Recall that we have a dataset represented as a $p \times n$ matrix $\mathbf{X}$ consisting of n data points $x_i \in \mathbb{R}^p$ ($i \in \{1, \ldots, n\}$) and a set of labels $y_{n\times 1} = (y_1, \ldots, y_n)$ where $y_i \in 1, \ldots, K$. The data X is assumed to follow a mixture of normal distributions, the same as the assumption for parametric classification in Section 3.2.1.

Here we only show how to derive Bayes accuracy for two classes ($K = 2$) as K-class ($K > 2$) Bayes accuracy can be decomposed into $\frac{1}{2}K(K-1)$ two-class Bayes accuracy. For simplicity, we set the within-class covariances for both classes as I_p and assume that the two classes have equal prior probabilities, i.e. $p(y_i = 1) = \frac{1}{2}$ and $p(y_i = 2) = \frac{1}{2}$. Then the probability density function of X can be written as:

$$f(X) = \frac{1}{2}\mathsf{N}(X; \mu_1, I_p) + \frac{1}{2}\mathsf{N}(X; \mu_2, I_p)$$

where μ_1 and μ_2 are the means for classes 1 and 2 respectively.

For the simple case with $p = 1$, the optimal classifier is determined by the center of the two means: $\frac{\mu_1+\mu_2}{2}$. Without loss of generality, we assume $\mu_1 \leq \mu_2$. Then the Bayes decision is

$$y_i = \begin{cases} 1, & \text{if } x_i \leq \frac{\mu_1+\mu_2}{2} \\ 2, & \text{otherwise} \end{cases}$$

The corresponding Bayes assuracy can be computed as follows:

$$\begin{aligned} A_1 &= \frac{1}{2}\int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} \phi_1(x-\mu_1)dx + \frac{1}{2}\int_{\frac{\mu_1+\mu_2}{2}}^{\infty} \phi_1(x-\mu_2)dx \\ &= \int_{-\infty}^{\frac{|\mu_1-\mu_2|}{2}} \phi_1(z)dz \\ &= \frac{1}{2} + \frac{1}{2}\mathrm{erf}(\frac{|\mu_1-\mu_2|}{2\sqrt{2}}) \end{aligned} \tag{A.1}$$

where $\phi_1(x) = \mathsf{N}(x; 0, 1)$ and $\mathrm{erf}(\frac{x}{\sqrt{2}}) = \int_{-x}^{x} \phi_1(z)dz$.

When $p > 1$, the optimal Bayesian decision is given by:

$$y_i = \begin{cases} 1, & \text{if } \log\frac{p(y_i=1|x_i)}{p(y_i=2|x_i)} > 0 \\ 2, & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} \log\frac{p(y_i=1|x_i)}{p(y_i=2|x_i)} &= \log\frac{p(x_i|y_i=1)p(y_i=1)}{p(x_i|y_i=2)p(y_i=2)} \\ &= -\frac{1}{2}(x_i-\mu_1)^t(x_i-\mu_1) + \frac{1}{2}(x_i-\mu_2)^t(x_i-\mu_2) \\ &= (\mu_1-\mu_2)^t x_i - \frac{1}{2}(\mu_1^t\mu_1 - \mu_2^t\mu_2)\ . \end{aligned}$$

Therefore the decision boundary is given by $\frac{(\mu_1-\mu_2)^t X}{\|\mu_1-\mu_2\|} = \frac{\mu_1\mu_1^t-\mu_2\mu_2^t}{\|\mu_1-\mu_2\|} := C$ and the Bayes accuracy is:

$$A_p = \frac{1}{2}\int \phi_p(x-\mu_1)\mathbf{1}_{\{\frac{(\mu_1-\mu_2)^t x}{\|\mu_1-\mu_2\|}>C\}}dx + \frac{1}{2}\int \phi_p(x-\mu_2)\mathbf{1}_{\{\frac{(\mu_1-\mu_2)^t x}{\|\mu_1-\mu_2\|}\le C\}}dx, \tag{A.2}$$

where $\phi_p(x-\mu_i) = \mathsf{N}(x; \mu_i, I_p)$.

Let $z_i = W^t x_i = (z_1, \cdots, z_p)^t$ where W is a $p \times p$ orthogonal matrix such that $WW^t = I_p$, and the first column of W is equal to $\frac{(\mu_1-\mu_2)^t}{\|\mu_1-\mu_2\|}$. Then we have $f(z_i|y_i=k) \sim \mathsf{N}(W\mu_i, I_p)$ for $k = 1, 2$. Using the transformed variable z_i, we could rewrite the Bayesian decision boundary

as $z_1 = C$ and the corresponding Bayes accuracy in Equation (A.2) as:

$$\begin{aligned} A_p &= \frac{1}{2}\int_C^{\infty} f(z_1|y_i=1)dz_1 + \frac{1}{2}\int_{-\infty}^{C} f(z_1|y_i=2)dz_1 \\ &= \frac{1}{2}\int_C^{\infty} \phi_1(z_1 - \frac{(\mu_1-\mu_2)^t\mu_1}{\|\mu_1-\mu_2\|})dz_1 + \frac{1}{2}\int_{-\infty}^{C} \phi_1(z_1 - \frac{(\mu_1-\mu_2)^t\mu_2}{\|\mu_1-\mu_2\|})dz_1 \\ &= \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}\int_{-C_1}^{C_1}\phi_1(z)dz\right) + \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}\int_{-C_2}^{C_2}\phi_1(z)dz\right) \\ &= \frac{1}{2} + \frac{1}{2}\int_{-C}^{C}\phi_1(z)dz \\ &= \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\|\mu_1-\mu_2\|}{2\sqrt{2}}) \end{aligned} \tag{A.3}$$

where we have used the following two equations in the derivation.

$$\begin{aligned} C_1 &= -(C - \frac{(\mu_1-\mu_2)^t\mu_1}{\|\mu_1-\mu_2\|}) = \frac{\|\mu_1-\mu_2\|}{2}, \\ C_2 &= C - \frac{(\mu_1-\mu_2)^t\mu_2}{\|\mu_1-\mu_2\|} = \frac{\|\mu_1-\mu_2\|}{2}. \end{aligned}$$

We can see the Bayes accuracy in 1-dim (A.1) and p-dim (A.3) have exactly the same form. Therefore we can write the Baye accuracy for any two classes in the general form:

$$A = \frac{1}{2} + \frac{1}{2}\text{erf}(\frac{\|\mu_1-\mu_2\|}{2\sqrt{2}}).$$

Appendix B

Connection of Our General Framework to Most Popular Dimension Reduction Algorithms

As mentioned in Section 3.2.3, quite a few dimension reduction algorithms can be reformulated into our general framework in (3.6) and FDA was shown as an example. In this part, we will include a few more examples.

1. For the objective function of parametric measure of classification accuracy in (3.3), it can be written in the form of (3.6) by setting $\mathcal{I} = \{(i,j)\,; 1 \le i < j \le K\}$, $f(t) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{\sqrt{t}}{2\sqrt{2}}\right)$, $w_I = p_i p_j$, the ℓ-th ($1 \le \ell \le n$) entry of h_I equal to $\mathbf{1}_{\{y_\ell = y_i\}} n_{y_i}^{-1} - \mathbf{1}_{\{y_\ell = y_j\}} n_{y_i}^{-1}$, and $\mathcal{J}$ as an empty set.

2. Similarly for the objective function of non-parametric measure of classification accuracy in (3.5), we can set $\mathcal{I} = \{(i,j)\,; y_i = y_j = k,\ 1 \le k \le K\}$, $\mathcal{J} = \{(i,j)\,; 1 \le i < j \le n\}$, $f(t) = \tilde{f}(t) = \exp(-t/\epsilon)$, $w_I = \tilde{w}_J = 1$, and the ℓ-th entry of h_I and $\tilde{h}_J$ equal to $\mathbf{1}_{\{\ell=i\}} - \mathbf{1}_{\{\ell=j\}}$ in our general framework to include it as a special case.

3. Principal Component Analysis (PCA) is a commonly used algorithm for unsupervised dimension reduction. The goal is to seek a $p \times m$ projection matrix $\mathbf{V}$ such that the maximum variability is reserved on the reduced space, that is,

$$\arg\max_{\mathbf{V},\, v_i^t v_i = 1} \text{tr}(\mathbf{V}^t S \mathbf{V}) \tag{B.1}$$

   where $S = n^{-1}\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^t$ and $\bar{x} = n^{-1}\sum_{i=1}^n x_i$. We can reformulate the objective function into our framework by setting $\mathcal{I} = \{i; 1 \le i \le n\}$, $f(t) = t$, $w_I = 1$, the ℓ-th entry of h_I equal to $\mathbf{1}_{\{\ell=i\}} - n^{-1}$, and $\mathcal{J}$ as an empty set.

4. Locality Preserving Projections (LPP) attempts to preserve the the local structure of the data in the reduced space. The objective function is given by:

$$\arg\max_{\mathbf{V},\, v_i^t v_i=1} \frac{\sum_{(i,j),1\le i<j\le n} ||\mathbf{V}^t(X_i - X_j)||^2 S_{ij}}{\sum_{i=1}^n ||\mathbf{V}^t X_i||^2 \sum_{j\neq i} S_{ij}} \tag{B.2}$$

where the weight function $S_{ij} = \exp\left(-||X_i - X_j||^2/t\right)$ if $||X_i - X_j||^2 < \epsilon$ and $S_{ij} = 0$ otherwise. Here t is a non-zero real number and $\epsilon > 0$ defines the radius of the local neighborhood. Apparently, this objective function can be reformulated into our general framework by setting $\mathcal{I} = \{(i,j), 1 \le i < j \le n\}$, $f(t) = t$, $w_I = S_I$ and the ℓ-th entry of h_I equal to $\mathbf{1}_{\{\ell=i\}} - \mathbf{1}_{\{\ell=j\}}$ for the numerator, and $\mathcal{J} = \{i, 1 \le i \le n\}$, $\tilde{f}(t) = t$, $\tilde{w}_J = \sum_{j,j\neq i} S_{ij}$ and the ℓ-th element of $\tilde{h}_J$ equal to $\mathbf{1}_{\{\ell=i\}}$ for the denominator.

5. Li (1991) introduced Sliced Inverse Regression (SIR) to find the effective dimension reduction directions to reduce the dimension of input data X without loss of information on the conditional distribution of $Y|X$ where Y is the response variable. For convenience, the input data X is usually standardized as $Z = \Sigma^{-1/2}(X - \mu)$ where μ and Σ are the mean and variance of the data respectively. When the response variable Y is discrete, the SIR kernel M is directly defined as $M = \sum_{k=1}^K \frac{n_k}{n} m_k m_k^t$ where n_k and m_k denote the number of observations and the mean of class k respectively. When the response variable Y is continuous, the algorithm first divides Y into K slices according to its range and then computes the sample mean of each slice. Withe the kernel matrix M, the objective function of SIR is:

$$\arg\max_{\mathbf{V},\, v_i^t v_i=1} \operatorname{tr}(\mathbf{V}^t M \mathbf{V}). \tag{B.3}$$

Again, this objective function can be written as a special case of our general form by setting $\mathcal{I} = \{i, 1 \le i \le K\}$, $f(t) = t$, $w_I = \frac{n_i}{n}$, the ℓth entry of h_I equal to $n_i^{-1}\mathbf{1}_{\{y_\ell=i\}}$, and $\mathcal{J}$ as an empty set.

6. Local FDA was proposed by Sugiyama (2007) to improve FDA by re-weighting the contribution to the calculation of $\mathbf{B}$ and $\mathbf{W}$ from pairs of the L_2 distance. The objective function is in the same form of FDA but with $\mathbf{B}$ and $\mathbf{W}$ defined as:

$$\mathbf{B} = \frac{1}{2}\sum_{i,j=1}^{n} B_{i,j}(x_i - x_j)(x_i - x_j)^t \quad \text{and} \quad \mathbf{W} = \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}(x_i - x_j)(x_i - x_j)^t,$$

where $B_{i,j} = A_{i,j}/n_k$ if $y_i = y_j = k$, 0 otherwise, and $W_{ij} = A_{i,j}(1/n - 1/n_k)$ if $y_i = y_j = k$, $1/n$ otherwise. Here A is a $n \times n$ affinity matrix with $A_{i,j} \in [0, 1]$ describing the similarity between data points x_i and x_j. Similar to FDA, we can include this algorithm as a special case of our general framework by setting $\mathcal{I} = \mathcal{J} = \{(i, j), 1 \leq i < j \leq n\}$, $f(t) = \tilde{f}(t) = t$, $w_I = B_I$, $\tilde{w}_J = W_J$, and the ℓ-th entry of h_I and $\tilde{h}_J$ equal to $\mathbf{1}_{\{\ell=i\}} - \mathbf{1}_{\{\ell=j\}}$.

References

Bekessy, A., L. Molineaux, and J. Storey (1976). Estimation of incidence and recovery rates of Plasmodium falciparum parasitaemia from longitudinal data. *Bulletin of the World Health Organization 54*(6), 685.

Bureau, A., S. Shiboski, and J. Hughes (2003). Applications of continuous time hidden Markov models to the study of misclassified disease outcomes. *Statistics in Medicine 22*(3), 441–462.

Butler, R. and J. Worrall (1985). Work injury compensation and the duration of nonwork spells. *The Economic Journal 95*(379), 714–724.

Chunge, R. (1989). *Intestinal Parasites In A Rural Community In Kiambu District, Kenya With Special Reference To Giardia Lamblia*. Ph. D. thesis, University College Galway.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing 36*(3), 287–314.

Cook, R. (1999). A mixed model for two-state Markov processes under panel observation. *Biometrics 55*(3), 915–920.

Cook, R. and S. Weisberg (1991). Discussion of Li (1991). *Journal of the American Statistical Association 86*, 328–332.

Cox, D. and V. Isham (1980). *Point Processes*. Chapman & Hall/CRC.

Crespi, C., W. Cumberland, and S. Blower (2005). A queueing model for chronic recurrent conditions under panel observation. *Biometrics 61*(1), 193–198.

Fisher, R. et al. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*, 179–188.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (6), 721–741.

Goldberger, J., S. Roweis, G. Hinton, and R. Salakhutdinov (2005). Neighbourhood components analysis. *Advances in Neural Information Processing Systems 17*, 513–520.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

He, X. and P. Niyogi (2003). Locality preserving projections. *Advances in Neural Information Processing Systems 16*, 153–160.

Hetsko, M., J. McCaffery, S. Svärd, T. Meng, X. Que, and F. Gillin (1998). Cellular and transcriptional changes during excystation of Giardia lamblia in vitro. *Experimental Parasitology 88*(3), 17283.

Huang, D. and A. White (2006). An updated review on Cryptosporidium and Giardia. *Gastroenterology Clinics of North America 35*(2), 291.

Ji, Y. and Z. Fan (2009). Analysis of longitudinal binary data with misclassification. *Preprint*.

Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 316–327.

Loog, M., R. Duin, and R. Haeb-Umbach (2001). Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(7), 762–766.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis*. Academic Press.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Chemical Physics 21*(6), 1087–1091.

Morris, C. and C. Christiansen (1995). Fitting Weibull duration models with random effects. *Lifetime Data Analysis 1*(4), 347–359.

Nagelkerke, N., R. Chunge, and S. Kinoti (1990). Estimation of parasitic infection dynamics when detectability is imperfect. *Statistics in Medicine 9*(10), 1211–1219.

Ng, E. and R. Cook (1997). Modeling two-state disease processes with random effects. *Lifetime Data Analysis 3*(4), 315–335.

Rosychuk, R. et al. (2009). Parameter estimation in a model for misclassified Markov data–a Bayesian approach. *Computational Statistics & Data Analysis 53*(11), 3805–3816.

Roweis, S. and L. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science 290*(5500), 2323.

Smith, T. and P. Vounatsou (2003). Estimation of infection and recovery rates for highly polymorphic parasites when detectability is imperfect, using hidden Markov models. *Statistics in medicine 22*(10), 1709–1724.

Sohn, S., I. Chang, and T. Moon (2007). Random effects Weibull regression model for occupational lifetime. *European Journal of Operational Research 179*(1), 124–131.

Sohn, S., K. Yoon, and I. Chang (2006). Random effects model for the reliability management of modules of a fighter aircraft. *Reliability Engineering & System Safety 91*(4), 433–437.

Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *The Journal of Machine Learning Research 8*, 1027–1061.

Svärd, S., T. Meng, M. Hetsko, J. McCaffery, and F. Gillin (1998). Differentiation-associated surface antigen variation in the ancient eukaryote Giardia lamblia. *Molecular Microbiology 30*(5), 979–989.

Warrell, D., T. Cox, and J. Firth (2003). *Oxford Textbook Of Medicine*, Volume 1. Oxford University Press.

Woolhouse, M., C. Dye, J. Etard, T. Smith, J. Charlwood, G. Garnett, P. Hagan, J. Hii, P. Ndhlovu, R. Quinnell, et al. (1997). Heterogeneities in the transmission of infectious agents: implications for the design of control programs. *Proceedings of the National Academy of Sciences of the United States of America 94*(1), 338.