LATENT CLASS PROFILE ANALYSIS: INFERENCE, ESTIMATION AND ITS APPLICATIONS

By

Hsiu-Ching Chang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Statistics

2011

ABSTRACT

LATENT CLASS PROFILE ANALYSIS: INFERENCE, ESTIMATION AND ITS APPLICATIONS

By

Hsiu-Ching Chang

Recently, a great deal of attention has been paid to the stage-sequential process for the longitudinal data and a number of methods for analyzing stage-sequential processes have been derived from the family of finite mixture modeling. However, research on the sequential process is rendered difficult by the fact that the number of latent components is not known a priori. To address this problem, we propose two solutions, reversible jump MCMC and the Bayesian non-parametric approach, so as to provide a set of principles for the systematic model selection for the stage-sequential process. The reversible jump MCMC sampler can explore parameter space and automatically learn the model. Nevertheless, we have found that reversible jump Markov chain Monte Carlo requires the efficient design of proposal mechanism as jumping rules. To reduce the technical and computational burdens, we propose a Bayesian non-parametric approach to select the number of latent components. Using a latent class-profile analysis, we test both algorithms on synthesized data sets to evaluate their performances in model selection problems.

Once a model is selected, the model parameters are needed to be estimated. The expectation-maximization algorithm (Dempster et al., 1977) and the data augmentation using MCMC (Hastings, 1970; Tanner and Wong, 1987a) are widely-used techniques to draw statistical inferences of the parameters for the LCPA model. As a number of measurement occasions increases in the LCPA model, however, the computation cost of expectation-maximization or MCMC will become exponentially intensive. On the contrary, if one adapts

recursive scheme in the update steps, calculations will be simplified and become generalized to more time points. In light of this, we formulate each update step with recursive terms which are directly analogous to forward-backward algorithm (Chib, 1996; MacKay, 1997).

The parameter estimation for the LCPA model benefits from recursive formula, but the recursive algorithm still requires careful examination for the existence of multiple local modes of the objective function (i.e., log-likelihood). Applying the recursive formula, we implement deterministic annealing EM (Ueda and Nakano, 1998) and deterministic annealing variant of variational Bayes (Katahiral et al., 2008) in order to find parameter estimates on the global mode of the objective function. Both methods are based on the deterministic annealing framework, in which ω is included as an annealing parameter to control the annealing rate. By adjusting the value of ω, the annealing process tracks multiple local modes and identifies the globalized optimum as a result.

At last, we are interested in analyzing the early onset drinking behaviours among the young generation. We apply latent class-profile analysis to alcohol drinking behaviours as manifest in self-reported items drawn from the National Longitudinal Survey of Youth 1997, which was a survey that explores the transition from school to work and from adolescence to adulthood in the USA. To unveil the stage-sequential bevaviroal progressions, we adopt dynamic Dirichlet learning process to characterize the probable progressions in a discrete manner and then identify patterns in which similar progressions are grouped. For the parameter estimations, we conduct deterministic annealing approaches with predetermined annealing schedule.

Key Words: Dirichlet process; Expectation-Maximization algorithm; Latent class profile analysis; Latent stage-sequential process; Longitudinal data; Multiple local modes; Recursive formula; Reversible jump MCMC; Deterministic annealing; Variatoinal Bayes

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Psychological research is increasingly turning to the idea of stage-sequential process. A common theme of stage-sequential process is that, at any moment, individuals are placed into distinct qualitative stages, and they can change their stage memberships over time. The latent class analysis (LCA) is perhaps the most straightforward mixture model now being used to identify mutually exclusive subgroups of individuals based on their responses to measured variables (Clogg and Goodman, 1984; Goodman, 1974). LCA models explain the relationships among categorical variables in a cross-classified contingency table by assuming the existence of an unobserved or latent classification. The first detailed statistical treatment of the LC model appeared in the textbook of Lazarsfeld and Henry (1968). In their terminology, the LC model is a special case of latent-structure analysis in which the measurement scales and the latent variables are both categorical. General overviews of LC modeling are provided by Goodman (1974), Haberman (1979), McCutcheon (1987), Heinen (1993, Chap. 2), Clogg (1995), Bartholomew and Knott (1999, Chap. 6), Hagenaars and McCutcheon (2002), and others.

Recently, a number of new methods for analyzing stage-sequential processes have been

derived from the family of LCA. For example, latent transition analysis (LTA) (Collins and Wugalter, 1992; Chung et al., 2008) and general growth mixture models (GGMM) (Muthén and Shedden, 1999; Muthén and Muthén, 2004) have been used widely to identify patterns of the progression of adolescent substance use, such as alcohol (Lanza and Collins, 2006) or tobacco (Velicer et al., 2007), or to investigate the initiation and progression of drug-taking behaviors for a number of different substances (Dierker et al., 2007).

Chung et al. (2011) proposed another type of LCA approach, the latent class-profile analysis (LCPA), to identify subtypes of the stage-sequential patterns of early-onset drinking behaviors, where drinking items are treated as fallible indicators of unseen states of drinking behaviors. In LCPA, the identification processes are divided into two steps. In the first step, LCPA identifies discrete subgroups of individuals who have similar responses to items at each measurement occasions. The subgroups identified in the first step are referred as *classes*. In the second step, LCPA examines individuals' class membership over the entire set of time points so as to classify the population into two or more subgroups based on their class sequencing. The subgroups identified in the second step are referred as *class profiles* or simply *profiles*. By applying an LCPA to the longitudinal study of adolescent drinking, for example, all drinkers in a class at a certain time point are expected to be homogeneous in terms of their drinking behavious, and those individuals in a given profile will have similar sequential pattern of class membership over time.

Like any other finite mixtures, the first and most crucial step in LCPA is to choose an appropriate number of classes and profiles since model selction has important ramifications for the analyses performed with the model. This study works on the issues regarding to the selection of the number of classes and profiles in LCPA. Two Bayesian approaches for selecting the number of classes and profiles have been proposed in this study: reversible jump

MCMC (RJMCMC) and Dirichlet process. RJMCMC has been proposed to select the number of latent components (e.g., classes and profiles) in finite mixture models (Green, 1995), where the number of components is considered as an unknown parameter to be estimated. In every step of the RJMCMC algorithm, the current mixture model can be proposed to jump across dimensions, or the model can be simply proposed to update the parameters within the current model. Any proposal is accepted with a probability that preserves reversibility with respect to the target posterior. As a result of the RJMCMC, we can estimate the relative frequencies regarding to the number of classes and profiles given the data. RJMCMC is widely applied on many applications including the finite mixture models (Richardson and Green, 1997) and linear mixed models (Ho and Hu, 2008). For the LCPA model, however, the new RJMCMC procedure should be developed to deal with the stage-sequential process for the latent component membership. In this study, we design a set of *split-and-merge* formula tailored for the multivariate categorical LCPA models.

As a class of non-parametric Bayesian techniques, the Dirichlet process has been utilized in Dirichlet process mixture models (also known as infinite mixture models). The involvement of the non-parametric prior allows us to identify different distributions over observed data. However, there is little literature regarding the Dirichlet process on model selection problems in stage-sequential process with longitudinal data. To perform the Dirichlet process in the longitudinal framework, we develop a dynamic approach that is able to explore the stage-sequential process by elaborating on the stage transition. It can be seen as a special hidden Markov model with no constraints placed. The dynamic approach has the space of all distributions as support and can be easily applied to compute posterior and draw inferences. Since the technique based on the Dirichlet process prior enables a dynamic model learning, we therefore name it dynamic Dirichlet learning process and integrate in the study

to determine the optimal number of latent components.

Once a model is selected, the parameters of the model needed to be estimated. There are several available parameter estimation methods and each is long-held for its own theoretical and practical worth. The expectation-maximization (EM) algorithm (Dempster et al., 1977) and the data augmentation using the Markov chain Monte Carlo (MCMC) (Hastings, 1970; Tanner and Wong, 1987a) are widely-used techniques to draw statistical inferences of the parameters. However, either the E-step of EM algorithm or I-step of MCMC requires the computation of marginal distributions, which appears to be too expensive to be of practical use in more generalized applications. To alleviate this problem, we formulate each update step with recursive formula which are directly analogous to forward-backward algorithm (Chib, 1996; MacKay, 1997). The recursive algorithm will therefore have less computational complexity and storage demands.

Existing algorithms for parameter learning in models for estimating suffer from local maxima problems since the dependency between neighboring starting values is strong. To relax the dependence of the initializations, split-and-merge EM (SMEM) (Ueda et al., 2000), deterministic annealing EM (DAEM) algorithm (Ueda and Nakano, 1998) and deterministic annealing variant of variational Bayes (DAVB) (Katahiral et al., 2008) will be introduced for this purpose. SMEM has some defects that deteriorate its capability of being widely practiced. The major implementational difficulties of SMEM lie in designing suitable split and merge operations and choosing the right component which the proposed mechanisms can be applied to. On the other hand, both the DAEM and the DAVB algorithms are based on the deterministic annealing framework in which ω is included as an annealing parameter to control the annealing rate. By adjusting the value of ω, the annealing process tracks the localized optimum and identifies the globalized optimum as a result.

We organize the rest as follows: Chapter 2 introduces the mathematical structure underlying the LCPA models. Chapter 3 presents the procedures of two Bayesian model selection algorithms: RJMCMC and Dirichlet process. Chapter 4 continues the previous chapter and compares their performances in discovering the true model given different sets of conditions on model parameters. Chapter 5 provides three parameter estimation methods: original EM algorithm with split-and-merge formula; deterministic annealing designs in the application of EM and variational Bayes. Chapter 6 discusses the model identifiability issues. In Chapter 7, we apply model selection and parameter estimation techniques to alcohol drinking items drawn from the National Longitudinal Survey of Youth 1997 (NLSY97) and conclude their performances. In Chapter 8, we summarize the contributions and suggest future researches.

Chapter 2

Latent Class Profile Analysis

Suppose we construct a S-profile LCPA model with C classes from a set of M items over T time periods. Let $\mathbf{C} = (C_1, \ldots, C_T)$ denote the class membership variables from initial time $t = 1$ to time T, where $C_t = 1, \ldots, C$, and let U denote the profile membership variable with S nominal categories. We begin with the profile identification procedure. The basic idea of LCPA in this identification procedure is that associations among class membership across T time points arise from the assumption that the population is composed of S profiles. If the ith individual's class membership $\mathbf{c} = (c_1, \ldots, c_T)$ could be observed, the joint probability that he or she belongs to the sequence $\mathbf{c}$ and the profile s is

$$\begin{aligned} P(U = s, \mathbf{C} = \mathbf{c}) &= P(U = s)P(\mathbf{C} = \mathbf{c} \mid U = s) \\ &= P(U = s) \prod_{t=1}^{T} P(C_t = c_t \mid U = s) \\ &= \gamma_s \prod_{t=1}^{T} \prod_{c=1}^{C} \left[\eta_{c|s}^{(t)}\right]^{I(c_t=c)}, \end{aligned} \tag{2.1}$$

where $\gamma_s = P(U = s)$ denotes the marginal prevalence of the sth profile membership; $\eta_{c|s}^{(t)} = P(C_t = c \mid U = s)$ represents the probability of the cth class membership at time t given the profile membership s; and $I(A)$ is the indicator function such that $I(A) = 1$ if A is satisfied and $I(A) = 0$ otherwise. Here, we assume that the class membership $\mathbf{C} = (C_1, \ldots, C_T)$ are conditionally independent or unrelated within each profile s. This assumption, called *local independence* (Lazarsfeld and Henry, 1968), is the crucial feature of LCPA that allows us to draw inferences about the unseen profile variable. In (2.1), the associations among class membership is explained by latent class theory, which posits that profiles can be identified by class sequencing over time. If individuals' class membership over T time points could be observed, we can simply apply an LCA to identify a number of profiles. However, since individual's class membership is not directly observed, another procedure should be conducted to identify class membership based on their responses to items.

The class identification explains associations among item responses based on the assumption that the population is composed of C classes at each time point. Let $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{Mt})$ represent a vector of discrete M variables measuring latent class membership at time t, and let $\mathbf{y}_{it} = (y_{i1t}, \ldots, y_{iMt})'$ be the observed values of $\mathbf{Y}_t$ for the ith individual, where each response y_{imt} can take values from 1 to r_m for $m = 1, \ldots, M$ and $t = 1, \ldots, T$. The joint probability that the individual belongs to class c_t at time t

and provides responses $\mathbf{y}_t$ would be

$$\begin{aligned} P(C_t = c, \mathbf{Y}_t = \mathbf{y}_{it} \mid U = s) &= P(C_t = c \mid U = s) P(\mathbf{Y}_t = \mathbf{y}_{it} \mid C_t = c) \\ &= P(C_t = c \mid U = s) \prod_{m=1}^{M} P(Y_{mt} = y_{imt} \mid C_t = c) \\ &= \eta_{c|s}^{(t)} \prod_{m=1}^{M} \prod_{k=1}^{r_m} \left[\rho_{mkt|c}\right]^{I(y_{imt}=k)}, \end{aligned} \tag{2.2}$$

where $\rho_{mkt|c_t} = P(Y_{mt} = k \mid C_t = c_t)$ represents the probability of response k to the mth item for a given class c_t at time t. In order to investigate the relationship among items, we assume the following: (a) $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{Mt})$ are conditionally independent given c for $t = 1, \ldots, T$; and (b) the profile membership U is related to the items $\mathbf{Y}_t$ only through the class membership C_t for $t = 1, \ldots, T$. Assumption (b) implies that U depends on the class membership $(C_1, \ldots, C_T)$ but not on the items $(\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$.

The joint probability that the individual belongs to the class sequence $\mathbf{c} = (c_1, \ldots, c_T)$ and the profile s and provides responses $\mathbf{y}_i = (\mathbf{y}_{i1}, \ldots, \mathbf{y}_{iT})$ would be

$$\begin{aligned} L_i^* &= P(U = s, \mathbf{C} = \mathbf{c}, \mathbf{Y} = \mathbf{y}_i) \\ &= P(U = s) \prod_{t=1}^{T} \left\{ P(C_t = c_t \mid U = s) \prod_{m=1}^{M} P(Y_{mt} = y_{imt} \mid C_t = c_t) \right\} \\ &= \gamma_s \prod_{t=1}^{T} \left\{ \eta_{c_t|s}^{(t)} \prod_{m=1}^{M} \prod_{k=1}^{r_m} \left[\rho_{mkt|c_t}\right]^{I(y_{imt}=k)} \right\} \end{aligned} \tag{2.3}$$

Therefore, the ith subject's contribution to the likelihood function of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_T)$, with-

out regard for the latent class and profile, is given by

$$\begin{aligned} L_i &= P(\mathbf{Y} = \mathbf{y}_i) \\ &= \sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} L_i^* \\ &= \sum_{s=1}^{S} \gamma_s \prod_{t=1}^{T} \left\{ \sum_{c_t=1}^{C} \eta_{c_t|s}^{(t)} \prod_{m=1}^{M} \prod_{k=1}^{r_m} \left[\rho_{mkt|c_t}\right]^{I(y_{imt}=k)} \right\} \end{aligned} \quad (2.4)$$

In (2.4), the following three sets of parameters are estimated:

1. $\rho_{mkt|c_t} = P(Y_{mt} = k \mid C_t = c_t)$ represents the probability of the response k to the mth item for a given class c_t at time t;

2. $\eta_{c_t|s}^{(t)} = P(C_t = c_t \mid U = s)$ represents the conditional probability of belonging to class c_t at time t for a given class profile s; and

3. $\gamma_s = P(U = s)$ represents the probability of belonging to the class profile s.

We refer to the ρ-parameter as the *primary measurement parameter* because it describes how individuals in each class tend to respond to the mth item at each occasion for $m = 1, \ldots, M$. The η-parameter, referred to as the *secondary measurement parameter* describes the relation between a class c_t at time t and a class-profile s. The primary measurement parameters are usually constrained to be equal across measurement occasions (i.e., $\rho_{mk1|c} = \cdots = \rho_{mkT|c}$), so that the meaning of classes will not change over time. In practice, this invariance assumption should be carefully checked by comparing the fit of the model with and without constraints.

2.1 Latent Class Profile Model with Covariates

When we consider the stage-sequential patterns, it is reasonable to take external causes into consideration. The nature way to extent the LCPA model in such a manner is to include covariates to examine whether the prevalence of latent profile varies under their influence.

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ denote a vector of time-invariant covariates for individual i that may influence the probability with which he or she falls into the latent class-profile $U_i = 1, \ldots, S$. That is to say, the marginal probability of U_i is affected by the covariates and therefore denoted as $P(U_i = s \mid \mathbf{x}_i)$. Even though the covariates come into play in shaping the profile sizes, the influences of x_i on the data is completely mediated by U_i.

In most cases, the first covariate is fixed as a constant $(x_{i1} = 1)$, so that the model with $p = 1$ reduces to a traditional LCPA model without covariates. The dependence of U_i on $\mathbf{x}_i$ is specified by

$$\begin{aligned} \gamma_s(\mathbf{x}_i) &= P(U_i = s|\mathbf{x}_i) \\ &= \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}{1 + \sum_{j=1}^{S-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_j)} \end{aligned} \tag{2.5}$$

$s = 1, 2, \ldots, S-1$, with $\sum_{s=1}^{S} \gamma_s = 1$. In equation (2.5), $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^T$ is a $p \times 1$ vector of logistic-regression coefficients influencing the log-odds that an individual falls into class j relative profile S, which serves as a baseline,

$$\log \frac{\gamma_s(\mathbf{x}_i)}{\gamma_S(\mathbf{x}_i)} = \mathbf{x}_i^T \boldsymbol{\beta}_s \tag{2.6}$$

Equation (2.6) appears to be a baseline-category logit model for a polytomous response (Agresti,

2002), except that in this case, the response is latent. The likelihood contribution for the ith individual can be written as

$$\begin{aligned} P(\mathbf{Y}=\mathbf{y}_i) &= \sum_{s=1}^{S}\sum_{c_1=1}^{C}\cdots\sum_{c_T=1}^{C} L_i^* \\ &= \sum_{s=1}^{S}\gamma_s(\mathbf{x}_i)\prod_{t=1}^{T}\left\{\sum_{c_t=1}^{C} R_{c_t|s}^{(t)}\right\}, \end{aligned} \tag{2.7}$$

where $R_{c_t|s}^{(t)} = \eta_{c_t|s}^{(t)}\prod_{m=1}^{M}\prod_{k=1}^{rm}\left[\rho_{mkt|c_t}\right]^{I(y_{imt}=k)}$.

The LCPA model with covariates (2.7) has the attractive property that the distribution of $\mathbf{Y}_i$ marginalized over the covariates reduces to that of a traditional LCPA model (Bandeen-Roche et al., 1997). Letting F denote the probability distribution of x_i, the marginal distribution of the item variables for the ith individual is

$$\begin{aligned} Pr(\mathbf{Y}_i=\mathbf{y}_i) &= \int\sum_{s=1}^{S}\gamma_s(\mathbf{x}_i)\prod_{t=1}^{T}R_s^{(t)}dF(\mathbf{x}_i) \\ &= \sum_{s=1}^{S}\int\gamma_s(\mathbf{x}_i)dF(\mathbf{x}_i)\prod_{t=1}^{T}R_s^{(t)} \\ &= \sum_{s=1}^{S}\gamma_s^*\prod_{t=1}^{T}R_s^{(t)}, \end{aligned} \tag{2.8}$$

where $R_s^{(t)} = \left\{\sum_{c_t=1}^{C}\eta_{c_t|s}^{(t)}\prod_{m=1}^{M}\prod_{k=1}^{rm}\left[\rho_{mkt|c_t}\right]^{I(y_{imt}=k)}\right\}$.

That is, the observed log likelihood function can be reduced to an LCPA model with γ_s^* representing the marginal class-profile membership probabilities averaged over the distribution of covariates in the population. This simplified form suggests a researcher can analyze

the data with covariates by two-stage process: first, fit a conventional LCPA model to the data without covariates to determine the nature of the latent variables U_i and $\mathbf{C}_i$; then introduce covariates to assess their influence on the class-profile variable U_i.

2.2 Missing Items

In real applications, it is inevitable to have missing data since some responses to one or more questionnaires are extracted missing. The common attempt for missing data is to delete the incomplete cases. However, the consequence could be far-flung if the removal of the incomplete individuals cause major changes in representing data characteristics.

At fact, case deletion might be unnecessary because the incomplete information can provide substantial true knowledge of the data and estimates of parameters which are meant to represent the full population may be biased if part of the information has been discarded (Little and Rubin, 1987).

A more principled method to deal with missing data is to apply a likelihood function defined by the marginal distribution of the observed items only. Maximizing this likelihood function which eliminates the missing responses is appropriate when the missing items are missing at random (MAR) in the sense defined by Rubin (1976) and Little and Rubin (1987). The contribution of individual i to this function, which we call the observed-data likelihood, is

$$Pr(\mathbf{Y}_{i,obs} = \mathbf{y}_{i,obs}) = \sum_{s=1}^{S} \gamma_s(\mathbf{x}_i) \prod_{t=1}^{T} \left\{ \sum_{c_t=1}^{C} R^{(t)}_{c_t|s} \right\}, \tag{2.9}$$

where $\mathbf{Y}_{i,obs}$ and $\mathbf{y}_{i,obs}$ are the vectors of observed item variables and their realized values,

and obs_i denotes the set of items observed for individual i. We assume that covariates x_i are completely observed. Missing values among the covariates would require us to introduce additional assumptions about the distribution of x_i. The concept for extension to include missing values in the covariates is straightforward but it can somehow wreck a havoc on the implementation because the dimension of the observed item vectors $Y_{i,obs}$ vary from one subject to another. Missing items also introduce complications in assessing goodness of fit and checking for departures from modeling assumptions (e.g., local independence).

Chapter 3

Model Selection

3.1 Introduction

In the finite mixture modeling, the problem to be solved is to estimate the unknown number of components. This question arises in many traditional and novel situations such as variable selections, signal processing and Bayesian non-parametric statistics. There are many model selectors in popular use such as Akaike's information criterion (Akaike, 1974) and Bayesian information criterion (Schwarz, 1978). As an estimated Kullback-Leibler (KL) distance, AIC aims at finding the best approximating model to the unknown true data generating process and penalizes with twice the number of parameters to achieve the parsimony and BIC operates in the similar way but penalizes more with the logarithm of sample size. Besides, the penalized least squares approach with smoothly clipped absolute deviation (SCAD) penalty term has been demonstrated to be an attractive selection approach which not only selects important variables consistently but also produces oracle estimations (Fan and Li, 2001).

Unfortunately, there is no single superior model selection tool which can be applied to all the types of data sets. For example, BIC is consistently better than AIC because if true

model is considered among the candidates, BIC can identify the true model almost surely as sample size grows; however, AIC has been proved to be minimax-rate optimal for both parametric and nonparametric cases for estimating the regression function (Yang, 2003). Even though AIC and BIC are commonly used for comparing different models, computing the relevant AIC and BIC values for each possible model can be very time consuming when the number of competing models is high. In the LCPA models, the structure is intricate and we need to develop methods by which we can fully incorporate data information to better the decision-making process.

3.2 Reversible Jump MCMC

In this section, we explore the algorithm of RJMCMC for the LCPA model to select the appropriate number of classes and profiles. The algorithm aims to compute the joint posterior distribution of the parameters and the number of latent components. The RJMCMC samplers can travel between different dimensions by constructing a reversible Markov chain on the general space with a specified limiting distribution. In other words, the RJMCMC can jump to another LCPA model with different dimensions (*across-model*) or it can simply update the parameters within the LCPA model with same dimension (*within-model*).

3.21 Within-Model Move

In Bayesian analysis, our goal is to compute the posterior distribution, $P(\boldsymbol{\Theta}_{(C,S)} \mid \mathbf{y})$, where $\mathbf{y}$ represents the vectorized observed items from the sample and $\boldsymbol{\Theta}_{(C,S)}$ denotes as the parameter set corresponding to C-class / S-profile LCPA model. The posterior distribution, however, is difficult to portray in the LCPA model. If the class and profile

membership for each individual were known, the augmented posterior $P(\boldsymbol{\Theta}_{(C,S)} \mid \mathbf{y}, \mathbf{z})$ would be easy to simulate.

A within-model MCMC algorithm for the LCPA model is implemented as an iterative two-step procedure which can be regarded as a form of data augmentation (Tanner and Wong, 1987b) or Gibbs sampling (Gelfand and Smith, 1990). In the first step of MCMC procedure—the Imputation or I-step— Let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ indicate the individuals' class and profile memberships where $\mathbf{z}_i$ is a $T+1$ dimensional array for the ith individual such that $z_{i(s,c_1,\ldots,c_T)} \in \{0,1\}$ and $\sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} z_{i(s,c_1,\ldots,c_T)} = 1$. That is, if individual i belongs to the profile s and the class membership $\mathbf{c} = (c_1, \ldots, c_T)$ from initial time $t = 1$ to time T, then $z_{i(s,c_1,\ldots,c_T)}$ equals 1 and 0 otherwise. We assume at $(j+1)$th cycle, the value $\mathbf{z}_i^{(j+1)}$ is drawn from the conditional distribution $P(L = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i, \boldsymbol{\gamma}^{(j)}, \boldsymbol{\eta}^{(j)}, \boldsymbol{\rho}^{(j)})$ given the observed data and previous parameter estimations. We then calculate the marginal indicators of $z_{is}^{(j+1)} = \prod_{t=1}^{T} \sum_{c_t=1}^{C} z_{i(s,c_1,\ldots,c_T)}$, $z_{i(s,c_t)}^{(j+1)} = \prod_{j \neq t} \sum_{c_j=1}^{C} z_{i(s,c_1,\ldots,c_T)}$, and $z_{ic_t}^{(j+1)} = \sum_{s=1}^{S} z_{it(s,c_t)}$ for $i = 1, \ldots, n$. In the second step—the Posterior or P-step—we draw new random values for the parameters from the augmented posterior distribution

$$\boldsymbol{\gamma}^{(j+1)}, \boldsymbol{\eta}^{(j+1)}, \boldsymbol{\rho}^{(j+1)} \sim P(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\rho} \mid \mathbf{y}, \mathbf{z}^{(j+1)}), \tag{3.1}$$

which regards the membership of class and profile as known. Repeating this two-step procedure creates a sequence of iterates,

$$\{(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\eta}^{(1)}, \boldsymbol{\rho}^{(1)}; \mathbf{z}^{(1)}), (\boldsymbol{\gamma}^{(2)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\rho}^{(2)}; \mathbf{z}^{(2)}) \ldots, \} \tag{3.2}$$

which converges to the stationary distribution $P(\boldsymbol{\Theta}_{(C,S)} \mid \mathbf{y})$. This stream of parameter values (after a suitable burn-in period) is summarized in various ways to produce approximate Bayesian estimates, intervals, tests, etc. It is convenient to choose priors that cause $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ to be a posteriori independent given $\mathbf{z}$. One way to achieve this is to make the priors independent from each other. In situations where the priors are independent, the joint posterior distribution for $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ given $\mathbf{z}$ can be expressed as

$$\begin{aligned} P(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\rho} \mid \mathbf{y}, \mathbf{z}) &\propto P(\boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\rho}) P(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{\rho}) \\ &\propto \left[P(\boldsymbol{\gamma}) \prod_{s=1}^{S} \gamma_s^{z_{is}} \right] \times \left[P(\boldsymbol{\eta}) \prod_{s=1}^{S} \prod_{t=1}^{T} \prod_{c_t=1}^{C} \left[\eta_{c|s}^{(t)} \right]^{z_{i(s,c)}^{(t)}} \right] \\ &\times \left[P(\boldsymbol{\rho}) \prod_{t=1}^{T} \prod_{c=1}^{C} \prod_{m=1}^{M} \prod_{k=1}^{r_m} \left\{ \left[\rho_{mkt|c} \right]^{I(yimt=k)} \right\}^{z_{ic}^{(t)}} \right], \end{aligned} \tag{3.3}$$

where $P(\boldsymbol{\gamma})$, $P(\boldsymbol{\eta})$ and $P(\boldsymbol{\rho})$ are the prior distributions for $\boldsymbol{\gamma}$, $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ and the equation is composed of three parts: one pertaining to the $\boldsymbol{\gamma}$ parameters and one pertaining to the $\boldsymbol{\eta}$ and the other pertaining to the $\boldsymbol{\rho}$.

3.21.1 Prior Distribution Specification

In the equation (3.3), the posteriors for the measurement parameters consist of prior distribution and the multinomial distribution. The most straightforward choice for the prior distribution is Dirichlet since it is conjugate to the multinomial distribution. It is said that a random vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S)$ has Dirichlet prior with hyperparameters $\alpha = (\alpha_1, \ldots, \alpha_S)$

if the density of $\boldsymbol{\gamma}$ is

$$P(\boldsymbol{\gamma} \mid \alpha) = \frac{\Gamma(\sum_{l=1}^{S} \alpha_l)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_S)} \prod_{l=1}^{S} \gamma_l^{\alpha-1} \tag{3.4}$$

over the simplex $\gamma_l \geq 0, l = 1, \ldots, S$, and $\sum_{l=1}^{S} \gamma_l = 1$, where $\Gamma(\cdot)$ denotes the gamma function.

We prefer using the noninformative priors to make sure the data-driven inferences will not be drawn under the influence of improper prior specification. The approach to choose a noninformative prior is to adopt Jeffreys' invariance principle (Box and Tiao, 1992). Jeffreys' prior density is proportional to $\prod_{l=1}^{S} \gamma_l^{-1/2}$, which corresponds to the Dirichelt with $\alpha = (1/2, \ldots, 1/2)$. Similarly, we apply the Jeffreys priors to $\boldsymbol{\rho}_{mt|c} = \left(\rho_{m1t|c}, \ldots, \rho_{mr_m t|c}\right)$ and $\boldsymbol{\eta}_s^{(t)} = \left(\eta_{1|s}^{(t)}, \ldots, \eta_{C|s}^{(t)}\right)$.

3.21.2 Posterior Distribution

Adopting a Dirichelt prior to $\boldsymbol{\gamma}, \boldsymbol{\eta}$ and $\boldsymbol{\rho}$, the complete-data posterior distribution is also Dirichlet because of conjugacy. Thus in the P-step, new random values for the parameters are drawn from posterior distributions

$$\begin{aligned}
\boldsymbol{\rho}_{m|c} &\sim \text{Dirichlet}\left(n_{m1|c} + 1/2, \ldots, n_{mr_m|c} + 1/2\right) \\
\boldsymbol{\eta}_s^{(t)} &\sim \text{Dirichlet}\left(n_{1|s}^{(t)} + 1/2, \ldots, n_{C|s}^{(t)} + 1/2\right) \\
\boldsymbol{\gamma} &\sim \text{Dirichlet}\left(n_1 + 1/2, \ldots, n_S + 1/2\right)
\end{aligned} \tag{3.5}$$

for $c_t = 1, \ldots, C$, $t = 1, \ldots, T$, $m = 1, \ldots, M$, and $s = 1, \ldots, S$, where $n^{(t)}_{c|s} = \sum_{i=1}^{n} z^{(t)}_{i(s,c)}$, $n_{mk|c} = \sum_{i=1}^{n} \sum_{t=1}^{T} z^{(t)}_{ic} I(y_{imt} = k)$, and $n_s = \sum_{i=1}^{n} z_{is}$. We repeat this two-step procedure to create a sequence of iterates converging to the stationary posterior distribution. This stream of parameter values (after a suitable burn-in period) is summarized in various ways to produce Bayesian inference and estimation.

3.22 Across-Model Move

For the across-model update, we assume that the sampling starts from the state $\boldsymbol{\Theta}_{(C,S)}$ and then transits to $\boldsymbol{\Theta}^*_{(C^*,S^*)}$. To match up the dimension, we generate a vector of continuous random variables u from some known distributions g and the new proposed state is then constructed by using an invertible deterministic function h such that $(\boldsymbol{\Theta}^*_{(C^*,S^*)}, u^*) = h(\boldsymbol{\Theta}_{(C,S)}, u)$ where u^* are the suitable random variables generated from some known functions g^* that is required for the reversed move using the inverse function h^* of h. The new state updates the current one with probability $\alpha(k, k^*)$, where $k = (C, S)$ and $k^* = (C^*, S^*)$. The probability $\alpha(\cdot, \cdot)$ is usually termed as the acceptance rate.

Assume the probability of choosing move from k to k^* is denoted by $q(\boldsymbol{\Theta}_k, \boldsymbol{\Theta}^*_{k^*})$, a valid choice for the acceptance rate is

$$\alpha(k, k^*) = min \left\{ 1, \frac{P(\boldsymbol{\Theta}^*_{k^*} \mid \mathbf{y}) q(\boldsymbol{\Theta}^*_{k^*}, \boldsymbol{\Theta}_k) g^*(u^*)}{P(\boldsymbol{\Theta}_k \mid \mathbf{y}) q(\boldsymbol{\Theta}_k, \boldsymbol{\Theta}^*_{k^*}) g(u)} \left| \frac{\partial(\boldsymbol{\Theta}^*_{k^*}, u^*)}{\partial(\boldsymbol{\Theta}_k, u)} \right| \right\}, \quad (3.6)$$

note that the last factor in (3.6) is the Jacobian arising from the transformation from (k, u) to (k^*, u^*).

To keep expressions simple, we consider the LCPA model with binary items (i.e., $r_m = 2$

for $m = 1, 2, \ldots, M$) with the invariance constraint on the ρ-parameter (i.e., $\rho_{mk1|c} = \ldots = \rho_{mkT|c} = \rho_{mk|c}$), although an extension to the model defines in (2.4) is straightforward. First, we select the number of classes C; and then explore the number of profiles S with the fixed number of classes. We assume that the random variables C and S are from Poisson distribution with hyperparameter ς with maximum values truncated at S_0 and C_0, respectively. Based on the simulation results, the choices of S_0 and C_0 have little impact on the performance of the RJMCMC and can be chosen as any reasonable values. In this study, we choose $\varsigma = 5$, $S_0 = 10$, and $C_0 = 10$. In RJMCMC, there are two possible ways to accommodate the dimensional changes in latent classes: *split* (e.g., the C-class LCPA moves to the $(C+1)$-class LCPA) and *merge* (e.g., the C-class LCPA moves to the $(C-1)$-class LCPA). Usually, at each stage of dimensional change, split and merge moves are equally preferred (i.e., $b_C = P(\text{ split } | \ C\text{-class LCPA}) = 0.5$).

3.22.1 Jumping Rules for C

To update the number of classes, we first generate U from Uniform$(0, 1)$. If $U < b_C$, split move is executed. Conditional on the current number of profiles, we randomly pick a class c^* from $(1, \ldots, C)$ and split it into classes c_1^* and c_2^*. To accommodate these classes, we draw $u_s^{(t)}$ from Uniform$(0, 1)$ and update the secondary measurement parameters as

$$\begin{aligned} \eta_{c_1^*|s}^{(t)} &= \eta_{c^*|s}^{(t)} u_s^{(t)} \\ \eta_{c_2^*|s}^{(t)} &= \eta_{c^*|s}^{(t)} (1 - u_s^{(t)}) \end{aligned}$$

for $s = 1, \ldots S$ and $t = 1, \ldots, T$. For the primary measurement parameter, we generate u_m from Uniform$(0, 1)$ and update the primary parameters as

$$\rho_{m1|c_1^*} = \rho_{m1|c^*} - u_m \kappa \sqrt{\frac{1-\bar{u}}{\bar{u}}}$$
$$\rho_{m1|c_2^*} = \rho_{m1|c^*} + u_m \kappa \sqrt{\frac{\bar{u}}{1-\bar{u}}}$$

for $m = 1, \ldots M$, where $\bar{u}$ is $\sum_{s=1}^{S} \sum_{t=1}^{T} u_s^{(t)}/ST$ and κ is an adjust term added to enhance the mixing performance. For simulation study, we choose 0.2 for κ since all $\rho_{mk|c}$ parameters are bounded and large κ (e.g. $\kappa \geq 0.4$) would violate the constraint easily and make the proposals invalid, however, small κ (e.g $\kappa \leq 0.1$) will lead to low acceptance rate and slow down the chain.

After the dimension-changing move is made, we need to reallocate the individuals whose class memberships were c^* to the new classes. The class label for each individual is not known a priori and we therefore follow ad hoc rule to assign them to either c_1^* or c_2^* according to the proposed formula.

On the other hand, if merge move is chosen, we randomly select two classes c_1^* and c_2^* from $(1, \ldots, C)$ and merge them into c^*. To preserve the reversibility, we update the parameters as

$$\begin{aligned} \eta_{c^*|s}^{(t)} &= \eta_{c_1^*|s}^{(t)} + \eta_{c_2^*|s}^{(t)} \\ \rho_{m1|c^*} &= \bar{u}\rho_{m1|c_1^*} + (1-\bar{u})\rho_{m1|c_2^*} \end{aligned}$$

for $s = 1, \ldots, S$, $t = 1, \ldots, T$, and $m = 1, \ldots, M$. In this case, the reallocation can be simply done by re-setting the memberships for those who were in the class c_1^* or c_2^* to

c^*.

With the updated parameters, we then calculate the acceptance rate $\alpha(C, C')$ of the proposed C'-class LCPA model ($C' = C - 1$ or $C + 1$) as (3.6) by replacing k with (C, S) and k^* with (C', S).

3.22.2 Jumping Rules for S

Based on the selected number of classes C, we update the number of profiles S using the following procedure. At each stage of dimensional change, split and merge moves are equally preferred (i.e., $b_S = P(\text{ split } \mid S\text{-profile LCPA}) = 0.5$). We generate V from the Uniform(0,1). If split move is chosen (i.e., $V < b_S$), we randomly select one profile s^* from $(1, \ldots, S)$ and split it into s_1^* and s_2^*. We then draw w from the Uniform(0,1) and set

$$\begin{aligned} \gamma_{s_1^*} &= \gamma_{s^*} w \\ \gamma_{s_2^*} &= \gamma_{s^*}(1-w). \end{aligned}$$

For the secondary measurement parameter, we update the parameters by setting them to be proportional to the odds ratio for s^*

$$\begin{aligned} \log\left(\frac{\eta_{c|s_1^*}^{(t)}}{\eta_{C|s_1^*}^{(t)}}\right) &= \log\left(\frac{\eta_{c|s^*}^{(t)}}{\eta_{C|s^*}^{(t)}}\right) + \frac{\beta_{ct}}{w} \\ \log\left(\frac{\eta_{c|s_2^*}^{(t)}}{\eta_{C|s_2^*}^{(t)}}\right) &= \log\left(\frac{\eta_{c|s^*}^{(t)}}{\eta_{C|s^*}^{(t)}}\right) - \frac{\beta_{ct}}{1-w}, \end{aligned}$$

where β_{ct} are generated from $N(0, \sigma^2)$ independently for $c = 1, \ldots, C-1$ and $t = 1, \ldots, T$. In the simulation study, we choose $\sigma = 1$ and for those individuals whose profile memberships were s^*, we relabel each of them to either s_1^* or s_2^* with probabilities based on the proposed formula.

If merge is selected, we randomly choose two profiles s_1^* and s_2^* and merge them into s^* by setting

$$\begin{aligned} \gamma_{s^*} &= \gamma_{s_1^*} + \gamma_{s_2^*} \\ \log\left(\frac{\eta_{1|s^*}^{(t)}}{\eta_{C|s^*}^{(t)}}\right) &= w \log\left(\frac{\eta_{c|s_1^*}^{(t)}}{\eta_{C|s_1^*}^{(t)}}\right) + (1-w)\log\left(\frac{\eta_{c|s_2^*}^{(t)}}{\eta_{C|s_2^*}^{(t)}}\right) \end{aligned}$$

for $c = 1, 2, \ldots, C-1$ and $t = 1, 2, \ldots, T$. And the reallocation can be easily done by labeling those who were in the profile s_1^* or s_2^* to s^*.

With the updated parameters, we then calculate the acceptance rate $\alpha(S, S')$ of the proposed S'-profile LCPA model ($S' = S - 1$ or $S + 1$) as (3.6) by replacing k with (C, S) and k^* with (C, S'). The details of the calculations of acceptance rates for class and profile are summarized in the appendix.

3.3 Dirichlet Process

3.31 Preliminary

Typically, we assume that data is drawn from an unknown distribution which we wish to estimate through the posterior distribution by specifying the prior. In most cases, the prior

distribution is parametrically specified and the Bayesian computing can help to infer the estimate easily. However, assuming prior has parametric form may not be suitable for the data. In the case of Dirichlet process, the prior is a random distribution over probability measures. The Dirichlet process is currently one of the most popular nonparametric Bayesian models since it can be practiced as a modern way to learn dominating components economically.

By the definition from Ferguson (Ferguson, 1973), G is said to be Dirichlet process on a measurable space $(\mathcal{B}, \mathcal{A})$ with concentration parameter λ and base measure G_0 if, for any finite measurable partition $(B_1, B_2, \ldots, B_k)$ of $\mathcal{B}$, the distribution of $(P(B_1), \ldots, P(B_k))$ is Dirichlet with parameter $(\lambda G_0(B_1), \ldots, \lambda G_0(B_k))$ and written as $G \sim DP(G_0, \lambda)$. The parameters G_0 and λ play intuitive roles in the definition of the Dirichlet process. For any measurable set $B \subset \mathcal{B}$, we have

$$\begin{aligned} E(G(B)) &= G_0(B) \\ Var(G(B)) &= G_0(B)(1 - G_0(B))/(\lambda + 1). \end{aligned} \tag{3.7}$$

The value of concentration parameter λ is a positive scalar with larger value of λ leading to small variance of G. In other words, Dirichlet process will concentrate more on the mean and as λ grows, $G(B)$ will converges to $G_0(B)$ weakly for any measurable set B in $\mathcal{B}$. On the contrary, a Dirichlet with small value of λ favors extreme distribution but the belief in prior is easily over-written by the data.

Since G is a prior distribution over the measure spaces, we can draw random samples from G itself. Let $\theta_1, \ldots, \theta_n$ be a sequence of independent draws from G. We can derive the posterior distribution of G given observed values of $\theta_1, \ldots, \theta_n$. Let $A_1, \ldots, A_r$ be a finite measurable partition of $\mathcal{B}$, and let $m_k = \#\{i : \theta_i \in A_k\}$ be the number of

observed values in A_k. Since Dirichlet is the conjugate prior of multinomial, we will have

$$(G(A_1), \ldots, G(A_r)) \mid \theta_1, \ldots, \theta_n \ \sim \ \text{Dirichlet}\left(\alpha G_0(A_1) + m_1, \ldots, \alpha G_0(A_r) + m_r\right)$$

The above expression is true for all the finite measurable partitions $A_1, \ldots, A_r$ of $\mathcal{B}$ and in the light of the definition from Ferguson, we know G must be a Dirichlet process. Generally put, the posterior distribution of G can be written down as:

$$G \mid \theta_1, \ldots, \theta_n \sim DP\left(\lambda + n, \frac{\lambda}{\lambda + n} G_0 + \frac{n}{\lambda + n} \frac{\sum_{i=1}^{n} \theta_i}{n}\right)$$

The posterior mean of G is the weighted average between the prior base distribution G_0 and the sample average $\frac{\sum_{i=1}^{n} \theta_i}{n}$. As $\lambda \to 0$, the prior becomes noninformative since the predictive distribution is controlled solely by the average. That is to say, as the amount of observations outnumbers the concentration parameter $(n >> \lambda)$, the posterior is constructed by the observations themselves which is regarded as a data-driven approximation to the true underlying model. This fully explains that Dirichlet process is consistent in recovering the model by increasing the sample size.

Consider again drawing $G \sim DP(G_0, \lambda)$, and drawing an i.i.d. sequence $\theta_1, \ldots, \theta_n \sim G$. The predictive distribution for θ_{n+1}, conditioned on $\theta_1, \ldots, \theta_n$ with G marginalized out, has the following formulation,

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{1}{\lambda + n}\left(\lambda G_0 + \sum_{i=1}^{n} \delta_{\theta_i}\right) \tag{3.8}$$

The sequence of predictive distributions is called the Blackwell- MacQueen urn scheme (Black-

well and Macqueen, 1973).

Regardless of the question about the existence of the Dirichlet process, we can calculate the joint probability of the sequence of $(\theta_1, \ldots, \theta_n)$ by the conditional rule,

$$P(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} P(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) \tag{3.9}$$

where the conditional probability is defined as (3.8). It is obvious that the order of the sampling will not change the probability and we therefore ensure that the sequence of the drawn samples is infinitely exchangeable. More precisely, we say $(\theta_1, \theta_2, \ldots)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $P(\theta_1, \ldots, \theta_n)$ is invariant to permutation of the indices. That is, for any permutation σ,

$$P(\theta_1, \ldots, \theta_n) = P(\theta_{\sigma(1)}, \ldots, \theta_{\sigma(n)}) \tag{3.10}$$

According to de Finetti's theorem (de Finetti, 1931) and the generalized version of his theorem (Hewitt and Savage, 1955), for any infinitely exchangeable sequence $(\theta_1, \theta_2, \ldots)$, there exists a probability measure μ on the set of probability measures $G(\cdot)$ such that

$$P(\theta_1, \ldots, \theta_n) = \int \prod_{i=1}^{n} G(\theta_i) d\mu(G) \tag{3.11}$$

In our study, the prior over the random distribution is precisely the Dirichlet process $DP(G_0, \lambda)$, thus establishing existence.

The important aspect of the Dirichlet process is its clustering property. Based on the predictive form formulated as (3.8), the first sample is drawn from the distribution G_0

and once $\theta_1, \ldots, \theta_n$ are observed, the next sample θ_{n+1} is either drawn from G_0 with probability $\lambda/(\lambda + n)$ or assigned the same value as θ_i for some $i = 1, \ldots, n$ with probability $1/(\lambda + n)$. After the sequence of draws is long enough, we note that G is a weighted sum of point masses and the same values of draws can be grouped together.

3.31.1 Sethuraman Stick-Breaking Representation

We have already mentioned that Dirichlet process is an almost surely discrete random probability measure which is composed of a weighted sum of point masses. Sethuraman made this precise by providing a constructive definition of the Dirichlet process called the stick-breaking construction (Sethuraman, 1994).

If $\mathcal{P} = DP(G_0, \lambda)$, it can be constructed with the following:

$$\begin{aligned} \beta_k &\sim Beta(1, \lambda) \\ \pi_k &= \beta_k \prod_{l=1}^{k-1} (1 - \beta_k) \\ \mathcal{P}(\cdot) &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}(\cdot), \text{ where } \theta_k^* \sim G_0 \end{aligned} \tag{3.12}$$

The construction of π can be understood through breaking a stick as follows. We imagine there is a stick whose length is 1 and we break it at β_1 and assign weights π_1 to the stick we just broke off. Then repetitively break the other portion to obtain π_2, π_3 and so forth. The stick-breaking distribution over π is sometimes written $\pi \sim GEM(\lambda)$ (GEM stands for Griffiths, Engen and McCloskey). To summarize, the Sethuraman representation indicates that the unknown distribution G can be recovered if the infinite number of θ_i along with their corresponding weights are available. Because of its simple representation, Sethuraman's

construction has lead to a variety of extensions as well as novel inference techniques for the Dirichlet process.

3.31.2 Ferguson Gamma Process Representation

Before the Sethuraman's representation, Ferguson provided a representation for the gamma process based on arrival times from a homogeneous Poisson process (Ferguson and klass, 1972). Let E_k be independent and identically distributed from exp(1) random variables and $\Gamma_k = E_1 + \cdots + E_k$. Let Z_k be i.i.d elements, independent of Γ_k, with a probability distribution G_0 over $(\mathcal{B}, \mathcal{A})$. Then Ferguson showed that the Dirichlet process with parameter (G_0, λ), could be described as the random probability measure

$$G(\cdot) = \sum_{k=1}^{\infty} N^{-1}(\Gamma_k)\delta_{Z_k}(\cdot) / \sum_{l=1}^{\infty} N^{-1}(\Gamma_l) \tag{3.13}$$

where

$$N(x) = \lambda \int_x^{\infty} \frac{e^{-u}}{u} du, \quad \text{for} \quad x > 0 \tag{3.14}$$

is the Lévy measure of a gamma random variable with shape parameter $\lambda > 0$.

In the Sethuraman's stick-breaking construction (3.12), the GEM weights are defined as

$$\pi_1 = \beta_1 \text{ and } \pi_k = (1 - \beta_1) \cdots (1 - \beta_{k-1})\beta_k \text{ where } k \geq 2, \tag{3.15}$$

and they are related to the Poisson process. By ordering them such that $\pi_{(1)} \geq \pi_{(2)} \geq \cdots$, the two sets of weights are also related by the following form (Patil and Taillie, 1977;

Perman and Pitman, 1992; Pitman and Yor, 1997)

$$(\pi_{(1)}, \pi_{(2)}, \cdots,) = \left(\frac{N^{-1}(\Gamma_1)}{\sum_{l=1}^{\infty} N^{-1}(\Gamma_l)}, \frac{N^{-1}(\Gamma_2)}{\sum_{l=1}^{\infty} N^{-1}(\Gamma_l)}, \cdots \right) \tag{3.16}$$

where N is the Lévy measure defined in (3.14).

Since there is no closed form solution for the inversed Lévy measure and since each $\pi_{(i)}$ requires infinite sum calculation, the Ferguson representation is not widely practicable.

3.31.3 Almost Sure Truncation of $DP(G_0, \lambda)$ Measures

We already know that a Dirichlet process can be formulated by the weighted sum of point masses and we are intuited to see if the truncated form of the Dirichlet process can replace the original Dirichlet process without losing accuracy. Let $DP_M(G_0, \lambda)$ be the approximate form of $DP(G_0, \lambda)$ by discarding the $M+1, M+2, \ldots$ terms and denote it as $\mathcal{P}_M$,

$$\mathcal{P}_M(\cdot) = \sum_{k=1}^{M} \pi_k \delta_{\theta_k^*}(\cdot), \text{ where } \theta_k^* \sim G_0, \tag{3.17}$$

and $\pi_M = 1 - \pi_1 - \ldots - \pi_{M-1}$. This corresponds to set $\beta_M = 1$ so that $\sum_{k=1}^{M} \pi_k = 1$. As shown in the paper (Ishwaran and Zarepour, 2000), the $DP_M(\lambda, G_0)$ random measure can be used to approximate integrable funcitonals of the Dirichlet process:

$$DP_M(G_0, \lambda)(g) \to DP(G_0, \lambda)(g) \tag{3.18}$$

for any arbitrary bounded and continuous real valued function g. The key property of (3.18) can be exploited to describe an efficient Gibbs sampler for Bayesian nonparametric problems

in which $\mathcal{P}_M$ is used as an approximating prior to the Dirichlet process. Let's consider the following hierarchical set up where the prior for the measure G follows truncated Dirichlet process:

$$\begin{aligned} \mathbf{Y}_i \mid \boldsymbol{\theta}_i &\sim f(\mathbf{y}_i \mid \boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i \mid G &\sim G \\ G &\sim \mathcal{P}_M \end{aligned} \tag{3.19}$$

The marginal density of the observed data $\mathbf{Y}$ is therefore written as

$$m_M(\mathbf{Y}) = \int \left\{ \prod_{i=1}^{n} \int f(\mathbf{y}_i|\boldsymbol{\theta}_i) dG(\boldsymbol{\theta}_i) \right\} d\mathcal{P}_M(G). \tag{3.20}$$

If m_∞ is the marginal density of Y with Dirichlet process as the prior for G, as Ishwaran (Ishwaran and James, 2001, 2002) showed,

$$\begin{aligned} \int |m_M(\mathbf{Y}) - m_\infty(\mathbf{Y})| d\mathbf{Y} &\leq 4 \left[1 - E \left\{ \left(\sum_{k=1}^{M-1} \pi_k \right)^n \right\} \right] \\ &\approx 4n \exp(-(M-1)/\lambda) \end{aligned} \tag{3.21}$$

The difference between the two marginal densities with each based on exact and approximate sum representation respectively has been proved as by Ishwaran as follows:

$$\begin{aligned}\int |m_M(\mathbf{Y}) - m_\infty(\mathbf{Y})| d\mathbf{Y} &= \int \left| \int \prod_{i=1}^{n} f(\mathbf{y}_i)(\pi^M(d\boldsymbol{\theta}) - \pi_\infty(d\boldsymbol{\theta}_i)) \right| d\mathbf{Y} \\ &\leq \int \int \prod_{i=1}^{n} f(\mathbf{y}_i) d\mathbf{Y} \left| \pi^M(d\boldsymbol{\theta}_i) - \pi_\infty(d\boldsymbol{\theta}_i) \right| \\ &= 2D(\pi^M, \pi_\infty) \end{aligned} \tag{3.22}$$

where D is the total variation distance between two probability measures π^M and π_∞. Let k_i be the indicator of the group membership $\boldsymbol{\theta}_i$ has, that is, $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*_{k_i}$. The sampled values $\boldsymbol{\theta}$ under π^M and π_∞ are identical if k_i is sampled from a value smaller than Mth term. Thus,

$$\begin{aligned} D(\pi^M, \pi_\infty) &\leq 2(1 - \pi^M\{k_i \leq M, \text{ for } i = 1, 2, \ldots, n\}) \\ &= 2\left[1 - E\left\{\left(\sum_{k=1}^{M-1} \pi_k\right)^n\right\}\right] \\ &\approx 2n \exp(-(M-1)/\lambda) \end{aligned} \tag{3.23}$$

where the right most approximation follows by observing that

$$\begin{aligned} \sum_{k=1}^{M-1} \pi_k &=^D 1 - exp(-E_1/\alpha)exp(-E_2/\alpha)\cdots exp(-E_{M-1}/\alpha) \\ &\approx 1 - exp(-(M-1)/\lambda) \end{aligned} \tag{3.24}$$

where $E_1, \cdots E_{M-1}$ are iid exp(1) random variables.

Another method to compare between $\mathcal{P}_M$ and $DP(G_0, \lambda)$ is to compare their clustering behavior under sampling. Let $\phi = \{\phi_1, \ldots, \phi_k\}$ denote the set of distinct $\boldsymbol{\theta}_i$'s, where $k \leq n$ is the number of distinct elements in the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ and s_i is the indicator defined by $s_i = j$ if $\theta_{s_i} = \phi_j$ for $i = 1, \ldots, n$. Let n_j be the number of $s_i = j$, D_M and D_∞ equal the number of distinct values in $\mathbf{Y}$ when sampled under $\mathcal{P}_M$ and $DP(\lambda, G_0)$, Ishwaran proved that

$$\frac{M!}{M^k(M-k)!} \leq \frac{P(D_M = k)}{P(D_\infty = k)} \leq n^{\lambda k/M}, \quad \text{for} \quad k = 1, \ldots, min(n, M) \quad (3.25)$$

Since both sides of inequality (3.25) converge to one for each k as $M \to \infty$, the two distributions agree in the limit by the squeeze theorem. Therefore, the M-truncation prior $\mathcal{P}_M$ has been proved to have similar features in the clustering behavior as the $DP(\lambda, G_0)$. As shown in the corollary 20 of the paper (Pitman, 1996), the posterior distribution $\mathcal{P}_M(\cdot \mid \boldsymbol{\theta})$ is the random probability measure represented as

$$\mathcal{P}_M(\cdot \mid \boldsymbol{\theta}) = \sum_{j=1}^{M} \pi_j^* \delta_{\boldsymbol{\theta}_j^*}(\cdot) + \pi_{m+1}^* \mathcal{P}_M^*(\cdot), \quad (3.26)$$

where $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_M^*$ are distinct values in the full sequence $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ occurring each with frequencies n_j^*, and

$$(\pi_1^*, \cdots, \pi_M^*, \pi_{M+1}^*) \sim Dir(n_1^* + \lambda/n, \cdots, n_M^* + \lambda/n, \lambda(1 - M/n)). \quad (3.27)$$

Let φ be a nonnegative or integrable function, the posterior mean of φ through $\mathcal{P}_M(\cdot|\mathbf{Y})$

is characterized by

$$\int \varphi(G)\mathcal{P}_M(dG \mid \mathbf{Y}) = \int\int \varphi(G)\mathcal{P}_M(dG \mid \theta_1, \ldots, \theta_n)\mu(d\theta_1 \ldots, \theta_n \mid \mathbf{Y}) \quad (3.28)$$

The interior integral on the right head side of (3.28) can be expressed through averaging over the values of $\theta_1, \ldots, \theta_n$ drawn from the Gibbs sampler,

$$\int \varphi(G)\mathcal{P}_M(dG \mid \theta_1, \ldots, \theta_n) = \sum_{j=1}^{M} \frac{n_j^* + \lambda/n}{\lambda + n}\varphi(\delta_{\theta_j^*}) + \frac{\lambda(1 - M/\lambda)}{\lambda + n}\varphi(G_0) \quad (3.29)$$

We can implement the preceding scheme (3.28) to estimate various posterior mean functionals by simplifying (3.28) with (3.29). There are many computational and theoretical advantages by using approximate sum representation since the simplified mathematical representations can be performed a lot more efficiently especially in cases where roundabout working strategy is needed.

3.32 Dirichlet Process Mixture Model

To be consistent with the organization of Chapter 2, we begin with the profile identification defined in (2.1). Let the ith individual's class membership at time t and his or her profile membership are c_{it} and s_i respectively. The class memberships c_{it} over time, $\mathbf{c}_i = (c_{i1}, \ldots, c_{iT})$, $\mathbf{c}_i = (c_{i1}, \ldots, c_{iT})$ is then distributed with the product of multi-

nomial probability densities with the form

$$\begin{aligned} f(\mathbf{c}_i \mid \boldsymbol{\eta}_{s_i}) &= P(\mathbf{C} = \mathbf{c}_i \mid U = s_i) \\ &= \prod_{t=1}^{T} \prod_{c_t=1}^{C} \eta_{c_t|s_i}^{(t)\, I(c_{it}=c_t)}, \end{aligned}$$

where $\boldsymbol{\eta}_{s_i} = (\eta_{1|s_i}^{(1)}, \ldots, \eta_{C|s_i}^{(T)})$ indicates the vector for the secondary measurement parameters associated with profile s_i. Sampling from a Dirichlet process mixture (DPM) can be schemed by forming G with countably infinite number of point masses from G_0 and draw parameters $\boldsymbol{\eta}_{s_i}$ from G,

$$\begin{aligned} \mathbf{c}_i \mid \boldsymbol{\eta}_{s_i} &\sim f(\mathbf{c}_i \mid \boldsymbol{\eta}_{s_i}) \\ \boldsymbol{\eta}_{s_i} &\sim G \\ G &\sim DP(G_0, \lambda). \end{aligned}$$

By marginalizing over the prior for G, the conditional distribution of $\boldsymbol{\eta}_{s_i}$ given the others is shown as

$$\boldsymbol{\eta}_{s_i} \mid \boldsymbol{\eta}_{s_1}, \ldots, \boldsymbol{\eta}_{s_{i-1}} \sim \frac{1}{i-1+\lambda} \sum_{j=1}^{i-1} \delta(\boldsymbol{\eta}_{s_j}) + \frac{\lambda}{i-1+\lambda} G_0, \qquad (3.30)$$

where $\delta(\boldsymbol{\eta}_{s_j})$ is the distribution concentrated at the point $\boldsymbol{\eta}_{s_j}$. Equation (3.30) is associated with the Pólya-urn representation (Blackwell and Macqueen, 1973). In LCPA, this formulation only requires marginal distribution because same values of $\boldsymbol{\eta}_{s_i}$ can be grouped together to form a profile. The evolution of $\boldsymbol{\eta}_{s_1}, \ldots, \boldsymbol{\eta}_{s_n}$ in (3.30) can also be formed by

taking the number of components S in the finite mixture model to infinity:

$$\begin{aligned} \mathbf{c}_i \mid \boldsymbol{\eta}_{s_i} &\sim f(\mathbf{c}_i \mid \boldsymbol{\eta}_{s_i}) \\ \boldsymbol{\eta}_{s_i} &\sim G_0 \\ s_i \mid \boldsymbol{\gamma} &\sim \text{Multinomial}(\gamma_1, \ldots, \gamma_S) \\ \boldsymbol{\gamma} &\sim \text{Dirichlet}(\lambda/S, \ldots, \lambda/S) \end{aligned} \tag{3.31}$$

By integrating over the mixing proportions $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_S)$, the marginal probability of s_i given the profile membership except the ith individual has the following form

$$\begin{aligned} & P(s_i = s \mid s_1, \ldots, s_{i-1}) \\ &= P(s_1, \ldots, s_{i-1}, s_i = s)/P(s_1, \ldots, s_{i-1}) \\ &= \frac{n_{i,s} + \lambda/S}{i - 1 + \lambda}, \end{aligned}$$

where $n_{i,s}$ is the number of s_j for $j < i$ that equals to s. Let us imagine that the ith individual is the last of the n observations, then the conditional probabilities for s_i given $\mathbf{s}_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, $\mathbf{c}_i$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_S)$ can be obtained by multiplying the likelihood, $f(\mathbf{c}_i \mid \boldsymbol{\eta}_s)$, as follows:

$$P(s_i = s \mid \mathbf{s}_{-i}, \mathbf{c}_i, \boldsymbol{\eta}) = b \frac{n_{-i,s} + \lambda/S}{n - 1 + \lambda} f(\mathbf{c}_i \mid \boldsymbol{\eta}_s),$$

where $n_{-i,s}$ is the number of s_j for $j \neq i$ that are equal to s, and b is the appropriate

normalizing constant. If S goes to infinity,

$$P(s_i = s \mid \mathbf{s}_{-i}, \mathbf{c}_i, \boldsymbol{\eta}) \rightarrow \begin{cases} b\frac{n_{-i,s}}{n-1+\lambda} f(\mathbf{c}_i \mid \boldsymbol{\eta}_s) & \text{if } s \text{ is equal to } s_j \text{ for some } j \neq i \\ b\frac{\lambda}{n-1+\lambda} \int f(\mathbf{c}_i \mid \boldsymbol{\eta}_s) dG_0(\boldsymbol{\eta}_s) & \text{if } s \text{ is not equal to } s_j \text{ for all } j \neq i \end{cases} \tag{3.32}$$

Here, $\boldsymbol{\eta}$ is the set of $\boldsymbol{\eta}_S$ currently associated with some observations because we cannot explicitly represent the infinite number of $\boldsymbol{\eta}_S$ as S goes to infinity.

Next, we specify how to perform Gibbs sampling on DPM models to select the number of profiles in the LCPA model. Let the current profile state consist of $\mathbf{s} = (s_1, \ldots, s_n)$. We can repeatedly sample as following three-step procedure:

1. For $i = 1, \ldots, n$, remove the profile s_i from the current state if s_i is not associated with any other observation (i.e., $n_{-i,s_i} = 0$).

2. Generate a new profile membership s_i from the equation defined in (3.32). If the new s_i is not associated with any other observation, generate a set of values for $\boldsymbol{\eta}_{s_i}^{(t)}$ for all t from the posterior distribution defined in (3.5). Repeat this step for $i = 1, \ldots, n$.

3. For all $s \in (s_1, \ldots, s_n)$, generate a set of new values for $\boldsymbol{\eta}_s^{(t)}$ for all t from the posterior distribution.

The procedures described above point out that the ith individual explores a new profile with probability $\lambda/(i - 1 + \lambda)$ and thus the average number of profiles is expected to grow logarithmically in the sample with size n as $O(\lambda \log n)$. It has also been proved that the number of profiles will converge to infinity almost surely as n increases (Korwar and Hollander, 1973). In other words, if λ is fixed to a constant with large sample size, an over-

fitting problem might occur. In Liu's paper (Liu, 1996), λ is estimated by its ML estimate by using sequential imputation but it demands many conditional probability calculations which barely comes with parametric forms. Therefore, we consider a noninformative gamma prior which assigns most equal probabilities to all possibilities to properly update λ with the procedures suggested by Escobar and West (1995). Here we briefly describe how to incorporate the concentration parameter into Gibbs sampling update procedure.

According to Antoniak (1974), given the concentration parameter λ and sample size n, we can represent the uncertainty of the number of components k with the following probability density,

$$P(k|\lambda, n) = C_n(k) n! \lambda^k \frac{\Gamma(\lambda)}{\Gamma(\lambda + n)}, \tag{3.33}$$

where $C_n(k)$ is nothing to do with λ . Assume we already sampled $\boldsymbol{\theta}_i$ for $i = 1, \ldots, n$ by DPM algorithm and are able to classify them into groups. In this way, the value of k is the number of groups we have for configuring the data. Suppose $\lambda \sim \mathcal{G}(a, b)$, a gamma prior with shape a and scale b. For $\lambda \geq 0$, we have

$$\frac{\Gamma(\lambda)}{\Gamma(\lambda + n)} = \frac{(\lambda + n) B(\lambda + 1, n)}{\lambda \Gamma(n)} \tag{3.34}$$

where $B(\cdot, \cdot)$ is the Beta function. Then the posterior of λ given k has the following form:

$$\begin{aligned} P(\lambda \mid k) &\propto P(\lambda) \lambda^{k-1} (\lambda + n) B(\lambda + 1, n) \\ &\propto P(\lambda) \lambda^{k-1} (\lambda + n) \int_0^1 x^{\lambda} (1 - x)^n dx \end{aligned} \tag{3.35}$$

This implies that $P(\lambda \mid k)$ is the marginal distribution from the joint distribution of λ and

a continuous variable ν such that

$$P(\lambda, \nu \mid k) \propto P(\lambda)\lambda^{k-1}(\lambda + n)\nu^{\lambda}(1-\nu)^{n-1}. \tag{3.36}$$

Since we assume λ has $\mathcal{G}(a, b)$ as the prior, we can write down the conditional probabilities for λ and ν as follows:

$$\begin{aligned} P(\nu \mid \lambda, k) &\propto \nu^{\lambda}(1-\nu)^{n-1} \\ P(\lambda \mid \nu, k) &\propto \lambda^{a+k-1}(e^{-b}\nu)^{\lambda} + n\lambda^{a+k-2}(e^{-b}\nu)^{\lambda}. \end{aligned} \tag{3.37}$$

It is obvious that given the value of λ, the value of ν is updated by drawing from $\mathcal{B}(\lambda+1, n)$ and the conditional probability of λ given ν reduces to a mixture of two Gamma densities.

$$P(\lambda \mid \nu, k) \sim \pi_{\nu}\mathcal{G}(a+k, b-\log(\nu)) + (1-\pi_{\nu})\mathcal{G}(a+k-1, b-\log(\nu)) \tag{3.38}$$

with π_{ν} satisfying $\pi_{\nu}/(1-\pi_{\nu}) = (a+k-1)/n(b-\log(\nu)$. At each iteration of Gibbs sampling, we can sample λ from the mixture of Gamma distributions given the currently sampled ν and k.

3.33 Dynamic Dirichlet Process Mixture Model

In the scenarios where data are thought to be produced by the stage-sequential process, hidden Markov model is certainly the most suitable method to describe the data set. However, the task of estimating the number of stages (i.e., classes) is not covered in the hidden Markov model setting. We need to extend the hidden Markov model by assuming each row of the transition matrix follows a Dirichlet process to explore the stage-sequential progression and

infer the number of classes afterwards. To be precise, we consider the data observed at each time point is a group and the observations are exchangeable at this specific time. While each mixture model has mixing proportions specific to the group, we require that the different groups share the same set of mixture components. The idea is that while different groups have different characteristics given by a different combination of mixing proportions, using the same set of mixture components allows statistical strength to be shared across groups, and allows generalization to new groups. It is generally known as hierarchical Dirichlet process (HDP). Overall speaking, The HDP allows the data at a specific time point to have similar structure by providing global layer of hierarchy. The applications of HDP are presented in literatures (Ahmed and Xing, 2008; Xu et al., 2008). In LCPA models, we adopt the concept of HDP to identify the number of classes. Since the number of classes is not necessarily constant over time, we call this process as dynamic Dirichlet process mixture (DDPM) model.

If each individual's class membership over time (i.e., c_{it}) could be observed, we would like a joint model for class membership across T time points, $\prod_{i=1}^{n} P(C_1 = c_{i1}, \ldots, C_T = c_{iT})$, where all possible sequences of class membership can be re-expressed as frequencies in a contingency table with C^T cells for n individuals. This table can produce a reasonable inference about the important aspects of the common progression of class membership over time. In LCPA, we assume there is an extra level of latent variable (i.e., class profile) by which the dependency among classes can be explained. Under this assumption, the classes are conditionally independent given the profile membership. The class memberships, however, are unobserved directly, and it is not possible to learn the class profile before the class memberships over time are known. In addition, it is not possible to know the true structure of dependency among class memberships over time. The LCPA can summarize

the dependency among classes by a small number of class profiles, without regard for the structure of dependency. Therefore, we implement DDPM by generating dependency among classes from the first-order Markov chain and summarizing the dependency through the class profiles.

Let $\mathbf{y}_{i,t}$ be the observed values for M items of the ith individual at time t. If his or her class membership at time t, $c_{i,t}$, could be observed, the distribution of $\mathbf{y}_{i,t}$ is the product of multinomial probability densities with the form

$$\begin{aligned} f(\mathbf{y}_{i,t} \mid \boldsymbol{\rho}_{t|c_{i,t}}) &= P(\mathbf{Y}_t = \mathbf{y}_{i,t} \mid C_t = c_{i,t}) \\ &= \prod_{m=1}^{M} \prod_{k=1}^{r_m} \left[\rho_{mkt|c_{i,t}}\right]^{I(y_{imt}=k)}, \end{aligned}$$

where $\boldsymbol{\rho}_{t|c_{i,t}}$ represents the vector of all ρ-parameters associated with time t and class membership $c_{i,t}$. Let $\tau^{(t)}_{c_t|c_{t-1}} = P(C_t = c_t \mid C_{t-1} = c_{t-1})$ represent the transition probability of class membership from time $t-1$ to t. Suppose we have C classes over T time periods, there is a $C \times C$ transition probability matrix $\boldsymbol{\tau}^{(t)}$ with all elements of each row of $\boldsymbol{\tau}^{(t)}$ summed to one for $t = 2, \ldots, T$. Thus, given the previous class membership c_{t-1}, the row vector $\boldsymbol{\tau}^{(t)}_{c_{t-1}}$, can be used as the mixing proportions for the current class membership.

As described before, sampling from a Dirichlet process mixture can be schemed by taking

the number of classes C to infinity:

$$\begin{aligned}
\mathbf{y}i,t \mid \boldsymbol{\rho}_{t|c_{i,t}} &\sim f(\mathbf{y}_{i,t} \mid \boldsymbol{\rho}_{t|c_{i,t}}) \\
c_{i,t} \mid \boldsymbol{\tau}^{(t)}_{c_{t-1}} &\sim \text{Multinomial}(\tau^{(t)}_{1|c_{t-1}}, \ldots, \tau^{(t)}_{C|c_{t-1}}) \\
\boldsymbol{\rho}_{t|c_{i,t}} &\sim G_0 \\
\boldsymbol{\tau}^{(t)}_{c_{t-1}} &\sim \text{Dirichlet}(n^{(t-1)}_1 + \alpha/C, \ldots, n^{(t-1)}_C + \alpha/C), \qquad (3.39)
\end{aligned}$$

where $n^{(t-1)}_c$ denotes the number of individuals who were assigned class c at time $t-1$, α is the concentration parameter. By integrating over the mixing proportions $\boldsymbol{\tau}^{(t)}_{c_{t-1}}$, the prior for $c_{i,t}$, as conditional probability, has the following form

$$P(c_{i,t} = c_t \mid c_{1,t-1}, \ldots, c_{n,t-1}, c_{1,t}, \ldots, c_{i-1,t}) = \frac{n^{(t-1)}_{c_t} + n^{(t)}_{i,c_t} + \alpha/C}{n + i - 1 + \alpha},$$

$$P(c_{i,t} = c_t \mid c_{1,t-1}, \ldots, c_{n,t-1}, c_{1,t}, \ldots, c_{i-1,t}) = \frac{n^{(t-1)}_{c_t} + n^{(t)}_{i,c_t} + \alpha/C}{n + i - 1 + \alpha},$$

where $n^{(t)}_{i,c_t}$ is the number of $c_{j,t}$ for $j < i$ that equals to c_t at time t. Let us imagine that the ith individual is the last of the n observations, then the conditional probabilities for $c_{i,t}$ given $\mathbf{c}_{t-1} = (c_{1,t-1}, \ldots, c_{n,t-1})$, $\mathbf{c}_{-i,t} = (c_{1,t}, \ldots, c_{i-1,t}, c_{i+1,t}, \ldots, c_{n,t-1})$, and $\boldsymbol{\rho}_t = (\boldsymbol{\rho}_{t|1}, \ldots, \boldsymbol{\rho}_{t|C})$ can be obtained by multiplying the likelihood, $f(\mathbf{y}_{i,t} \mid$

$\boldsymbol{\rho}_{t|c_t}$), as follows:

$$P(c_{i,t} = c_t \mid \mathbf{c}_{t-1}, \mathbf{c}_{-i,t}, \mathbf{y}_{i,t}, \boldsymbol{\rho}_t) = b\frac{n_{c_t}^{(t-1)} + n_{-i,c_t}^{(t)} + \alpha/C}{2n-1+\alpha} f(\mathbf{y}_{i,t} \mid \boldsymbol{\rho}_{t|c_t}),$$

where $n_{-i,c_t}^{(t)}$ is the number of $c_{j,t}$ for $j \neq i$ that are equal to c_t at time t, and b is the appropriate normalizing constant. If C goes to infinity,

$$P(c_{i,t} = c_t \mid \mathbf{c}_{t-1}, \mathbf{c}_{-i,t}, \mathbf{y}_{it}, \boldsymbol{\rho}_t) \rightarrow$$

$$\begin{cases} b\frac{n_{c_t}^{(t-1)} + n_{-i,c_t}^{(t)}}{2n-1+\alpha} f(\mathbf{y}_{it} \mid \boldsymbol{\rho}_{t|c_t}) & \text{if } c_t \text{ is equal to } c_{j,t-1} \text{ or } c_{j,t} \text{ for some } j \neq i \\ b\frac{n_{c_t}^{(t-1)} + n_{-i,c_t}^{(t)}}{2n-1+\alpha} \int f(\mathbf{y}_{it} \mid \boldsymbol{\rho}_{t|c_t}) dG_0(\boldsymbol{\rho}_{t|c_t}) & \text{if } c_t \text{ is not equal to } c_{j,t-1} \text{ and } c_{j,t} \text{ for all } j \neq i \end{cases} \tag{3.40}$$

Here, $\boldsymbol{\rho}_t$ is the set of $\boldsymbol{\rho}_{t|c_t}$ currently associated with some observations because we cannot explicitly represent the infinite number of $\boldsymbol{\rho}_{t|c_t}$ as C goes to infinity. Therefore, we can update the current class membership by sampling $c_{i,t}$ over time from the conditional probability

$$\begin{aligned} & P(c_{i,t} \mid \mathbf{c}_{t-1}, \mathbf{c}_{-i,t}, \mathbf{c}_{t+1}, \mathbf{y}_{i,t}, \boldsymbol{\rho}_t) \\ &= P(c_{it} = c_t \mid \mathbf{c}_{t-1}, \mathbf{c}_{-i,t}, \mathbf{y}_{it}, \boldsymbol{\rho}_t) P(\mathbf{c}_{t+1} \mid \mathbf{c}_t). \end{aligned} \tag{3.41}$$

Note that the first factor in (3.41) was presented in Equation (3.40). The second factor in (3.41) can be computed by integrating over the mixture weights $\boldsymbol{\tau}^{(t+1)}$ which depends

on the number of individuals who were assigned to each of identified classes at time t (i.e., $n_1^{(t)}, \ldots, n_C^{(t)}$). Here, C is the number of existing classes at time t and $t+1$. Then, it is straightforward to show that:

$$P(\mathbf{c}_{t+1} \mid \mathbf{c}_t) = \frac{\Gamma(\sum_{c=1}^{C} n_c^{(t)} + \alpha/C)}{\prod_{c=1}^{C} \Gamma(n_c^{(t)} + \alpha/C)} \times \frac{\prod_{c=1}^{C} \Gamma(n_c^{(t)} + n_c^{(t+1)} + \alpha/C)}{\Gamma(\sum_{c=1}^{C} n_c^{(t)} + n_c^{(t+1)} + \alpha/C)},$$

where Γ is a typical gamma function. We consider Gamma distribution with shape a and rate b, $\mathcal{G}(a, b)$, as the prior for the concentration parameters α and λ used in class and profile learning, respectively. Given the previous constructions, we specify the Gibbs sampling by following four-step procedure.

1. For $t = 1, \ldots, T$, generate a new class membership $c_{i,t}$ from the equation (3.41). If the new $c_{i,t}$ is not associated with any other observation (i.e., $n_{-i,c_{i,t}}^{(t)} = n_{-i,c_{i,t}}^{(t-1)} = 0$), generate a set of values for $\boldsymbol{\rho}_{t|c_{i,t}}$ from the posterior distribution defined in (3.5). Repeat this step for $i = 1, \ldots, n$.

2. To update concentration parameter, α, we sample x and u from Beta($\alpha + 1$) and Uniform(0, 1), respectively. We then calculate $p_\alpha = (a + \mu_\alpha - 1)/(a + \mu_\alpha - 1 + n(b - \log(x)))$, where μ_α is the average of numbers of classes identified at each time. If $u < p_\alpha$, we update α by drawing a sample from $\mathcal{G}(a + \mu_\alpha, b - \log(x))$ or draw a sample from $\mathcal{G}(a + \mu_\alpha - 1, b - \log(x))$, otherwise.

3. With the class membership identified in Step 1, $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,T})$ for $i = 1, \ldots, n$, employ DPM to identify the number of profiles by following the three-step procedure described above.

4. We sample x and u from Beta($\lambda + 1$) and Uniform(0, 1), respectively and calculate $p_\lambda = (a + S - 1)/(a + S - 1 + n(b - \log(x)))$, where S is the number of identified profiles in Step 3. If $u < p_\lambda$, we update λ by drawing a sample from $\mathcal{G}(a + S, b - \log(x))$ or draw a sample from $\mathcal{G}(a + S - 1, b - \log(x))$, otherwise.

Chapter 4

Simulation

In this section, we evaluate the performance of RJMCMC and Dirichlet process over repeated samples. In our simulation, we draw over 200 samples from a pre-specified data set constructed with an LCPA representing two classes (i.e., $C = 2$) and two profiles (i.e., $S = 2$) with four binary items (i.e., $M = 4$) measured over four-time periods (i.e., $T = 4$). To simplify the presentation, we constrained ρ-parameter to be invariant over time (i.e., $\rho_{mk1|c} = \cdots = \rho_{mkT|c} = \rho_{mk|c}$), but imposed no constraint on the η-parameter. Note that the η-parameter describes the class progression over time. If the η-parameters were time invariant (i.e., $\boldsymbol{\eta}_s^{(1)} = \ldots = \boldsymbol{\eta}_s^{(T)}$ where $\boldsymbol{\eta}_s^{(1)} = (\eta_{1|s}^{(1)}, \ldots, \eta_{C|s}^{(1)})$.), given a profile, the probabilities of belonging to a specific class could not change over time, leading to difficulties in interpretation for the identified profiles. We used the balanced probability of profile membership (i.e., $\gamma_1 = \gamma_2 = .5$), but the following three factors were varied for the simulation study: relationship between items and classes (i.e., ρ-parameter), relationship between classes and profiles (i.e., η-parameter), and sample size ($n = 250$ or $n = 500$). The ρ-parameter and the η-parameter have two levels: the strong (i.e, $\rho = .9$ or .1 and $\eta = .9$ or .1) relationship and the mixed (i.e., $.1 < \rho < .4$ or $.6 < \rho < .9$ and $.1 < \eta < .4$ or

$.6 < \eta < .9$) relationship. For each sample, we estimate the number of classes and profiles by the RJMCMC and Dirichlet process. Our purpose is to compare their performances by comparing the relative frequency of selecting the true model (i.e., two-class and two-profile LCPA) over repeated sample. For the prior specifications, we choose $\varsigma = 5$, $\kappa = 0.2$, and $\sigma = 1$ and apply rather diffuse prior, $\mathcal{G}(1.5, 1)$, on both concentration parameters α and λ (i.e., $a = 1.5$ and $b = 1$).

4.1 Simulation with Reversible Jump MCMC

For each data set, we run our sampler for 10,000 iterations with an additional 1,000 iterations for burn-in period. Regrading to the performance in recovering the correct number of classes, RJMCMC never fails to select the true number of classes under all combinations of factors considered. We find that RJMCMC samplers converge to the desired posterior distribution of the number of classes quickly. Once it converged, however, the trans-dimension moves are not easily implemented. In fact, the difficulty in accepting the proposed parameters of a different dimensional space may lead to some biases in the number of classes. To improve low acceptance rates, we modify the across-model scheme by adding birth and death of an empty class. The rate of accepting the birth move of an empty class is controlled by the quantity $\prod_{s=1}^{S}\prod_{t=1}^{T}(1-u_s^{(t)})^{n_s^{(t)}}$, where $u_s^{(t)}$ is a random variable from Uniform$(0,1)$ and $n_s^{(t)}$ is the size of the profile s at time t, for all $s = 1, \ldots, S$ and $t = 1, \ldots, T$. The problem of poor mixing remains, however, since the acceptance rate could diminish exponentially fast even for the moderate size of profile.

Table 4.1 shows the percentage of the most frequently selected profiles over simulated samples. For the number of profiles, RJMCMC is able to identify correct number of profiles

Table 4.1: The percentage of the most frequently selected profiles over simulated samples by reversible jump MCMC

Sample size	ρ	η	Number of profiles 1	2	3	4	5	6	7
250	Strong	Strong	0.0	78.4	19.6	2.0	0.0	0.0	0.0
	Mixed	Strong	0.0	74.8	20.2	4.6	0.4	0.0	0.0
	Strong	Mixed	0.0	45.2	48.6	5.4	0.8	0.0	0.0
	Mixed	Mixed	5.2	34.8	58.8	1.2	0.0	0.0	0.0
500	Strong	Strong	0.0	82.2	17.2	0.6	0.0	0.0	0.0
	Mixed	Strong	0.0	52.2	45.8	1.6	0.3	0.1	0.0
	Strong	Mixed	5.4	19.6	70.8	3.0	0.8	0.3	0.1
	Mixed	Mixed	10.4	14.4	70.2	4.6	0.4	0.0	0.0

when both measurements are strong (78.4% for the small sample and 82.2% for the large sample). However if secondary measurement is mixed, the accurate disparity among competing models is difficult especially when RJMCMC works with large samples. We find out given weak measurement setups, some of the resulting profiles are formalised with small sizes and richly strictured samples would exacerbate the performance even more. We generalize the cases to explain how intrinsic difficulty in RJMCMC impacts the performance in the later section.

4.2 Simulation with Dirichlet Process

Even though Dirichlet process chooses the correct number of classes during the whole simulations, the method is not free from problems. A troubling aspect of the Dirichlet process is that a complex model is falsely preferred because there is probability that some profiles are generated but rarely been visited since then, which would result in sparse decomposition of an LCPA model. In addition to presenting results from Dirichlet process without applying

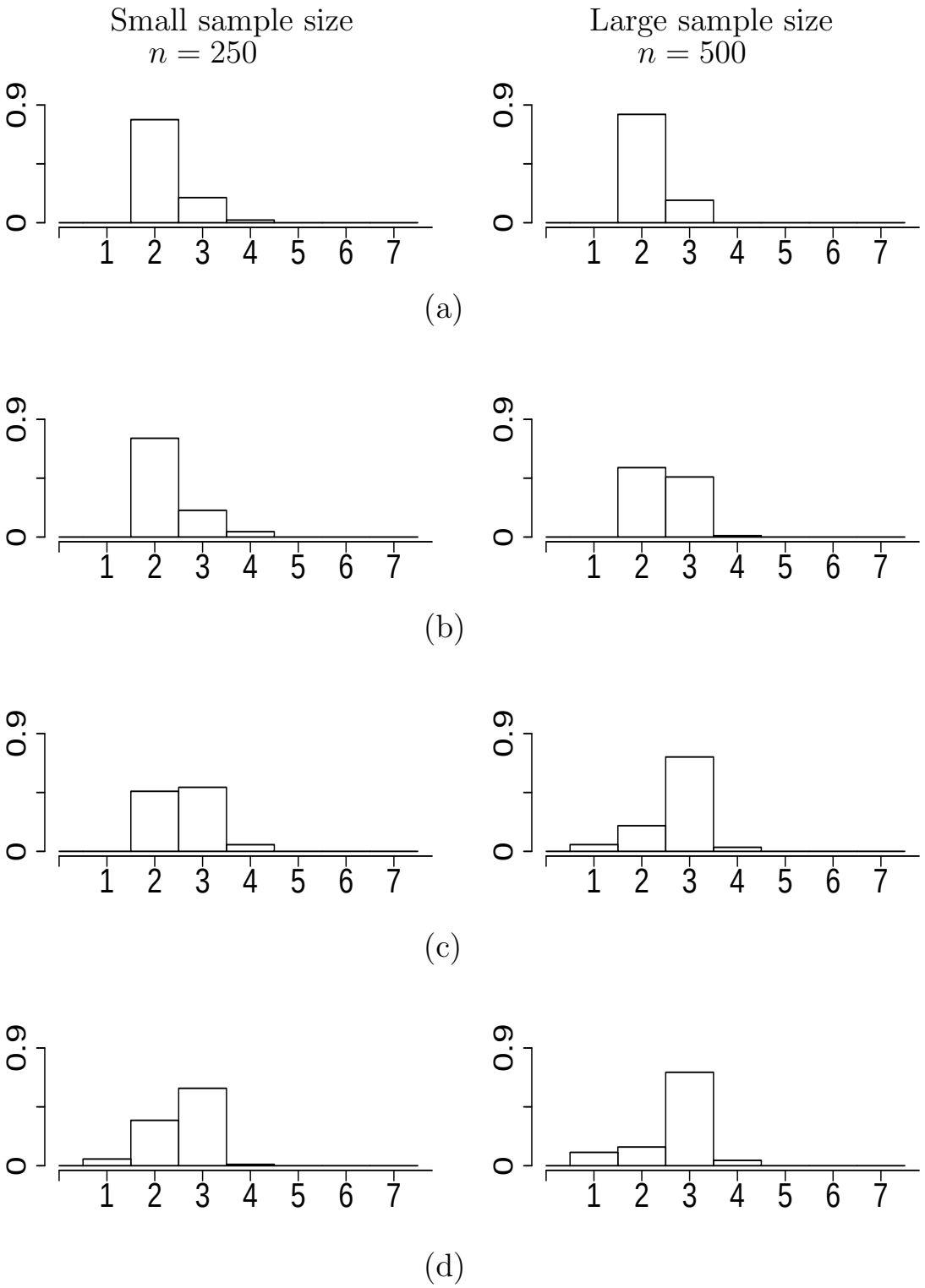

Figure 4.1: The percentage of the most frequently selected profiles over simulated samples by reversible jump MCMC: (a) strong ρ and strong η, (b) mixed ρ and strong η, (c) strong ρ and mixed η, and (d) mixed ρ and mixed η

any inclusion threshold, we will adapt our models to work with two cutoffs (e.g., 1% and 5%) and investigate the changes.

The results of the Dirichlet process for the number of profiles are presented in Figure 4.2. Dirichlet process is known to be consistent in recovering the true model as we increase the sample size and Figure 4.2 illustratively corroborates this attribute since the the correct rates increase when sample size grows.

It is rather interesting to know that Dirichlet process favors models with excessive number of profiles especially when working with mixed primary measurements. We believe the poor performance can be ascribable to the weak relationship between classes and the items since the sequence of the well identified class memberships is the key ingredient for successful classification of the profiles. When the primary measurement is mixed, we can see the poor performance, especially with small samples . Generally, Dirichlet process is poorly suited to identify the correct number of profiles when none of the measurements is strong and sample size is small but the method is vindicated to have a good practical use when sample size is large.

As the matter of fact, in the process of Dirichlet process, there are some profiles developed with sizes not even larger than 1% in many stages of profile learning. To avoid profile redundancy, we need to impose a threshold as an inclusion criterion. We use 1% and 5% as the cut-offs, by which any profile whose size is less than the given cutoff will be considered too sparse to be included and the corresponding distributions for the number of profiles are then given in Figure 4.3 and Figure 4.4.

Clearly, when 5% cutoff is applied, the predictive accuracy increases dramatically especially when strong-strong pairs and large sample sizes are the working scenario. To determine an optimal cutoff value is more of the matter in discovering right dominating profiles.

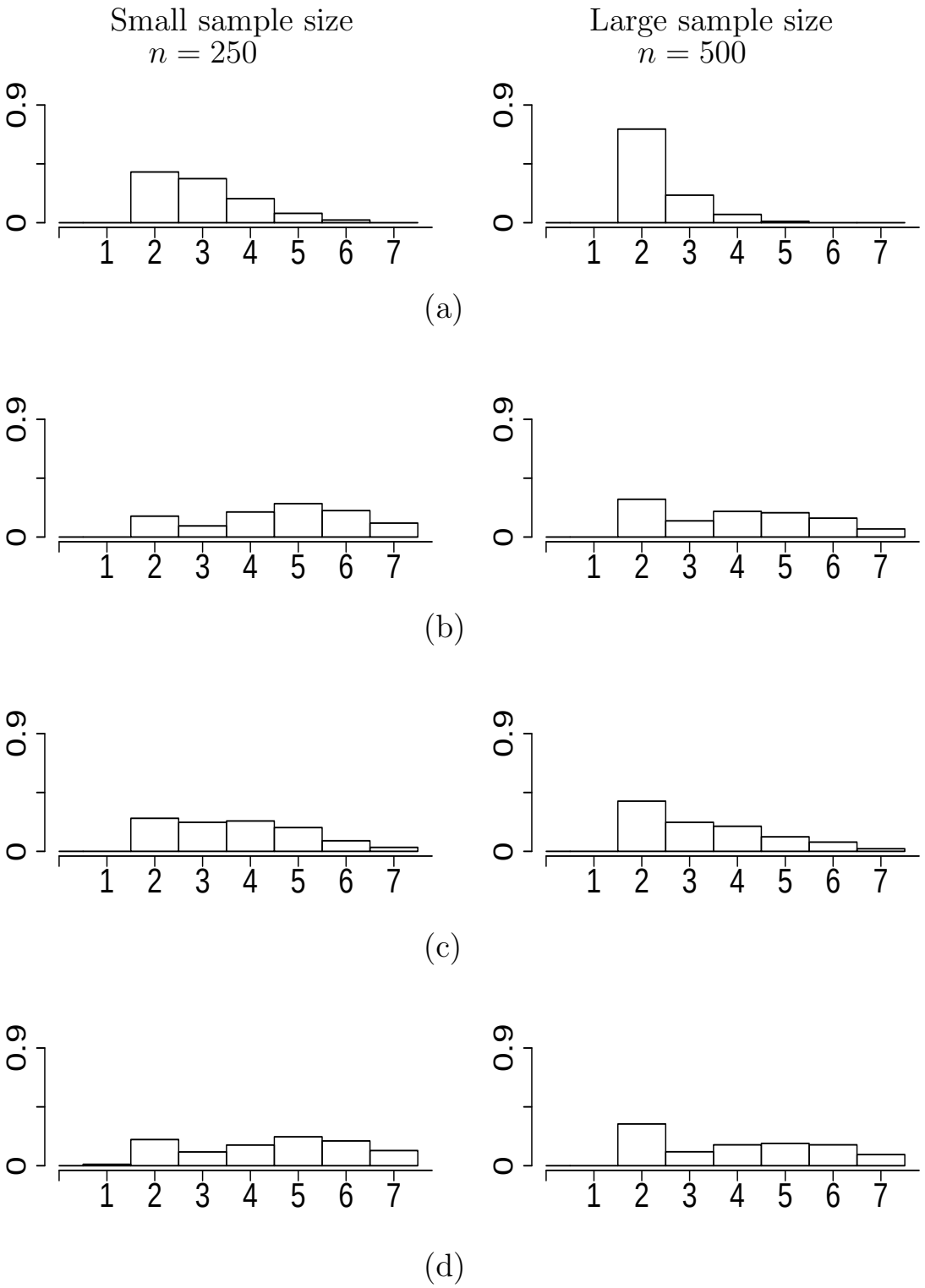

Figure 4.2: The histogram of the most frequently selected profiles over simulated samples from Dirichlet process: (a) strong ρ and strong η, (b) mixed ρ and strong η, (c) strong ρ and mixed η, and (d) mixed ρ and mixed η

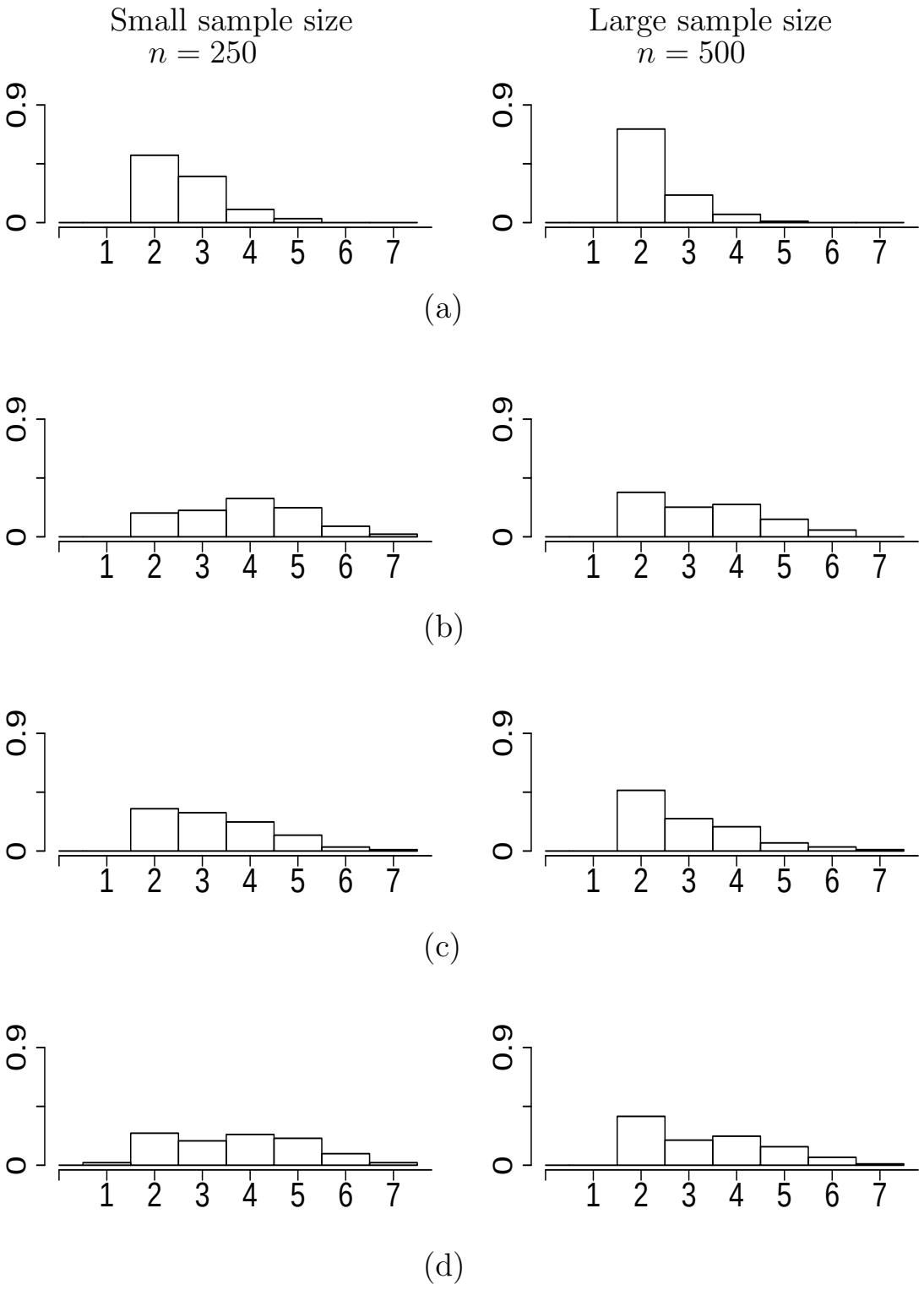

Figure 4.3: The histogram of the most frequently selected profiles over simulated samples from Dirichlet process with cutoff 1%: (a) strong ρ and strong η, (b) mixed ρ and strong η, (c) strong ρ and mixed η, and (d) mixed ρ and mixed η

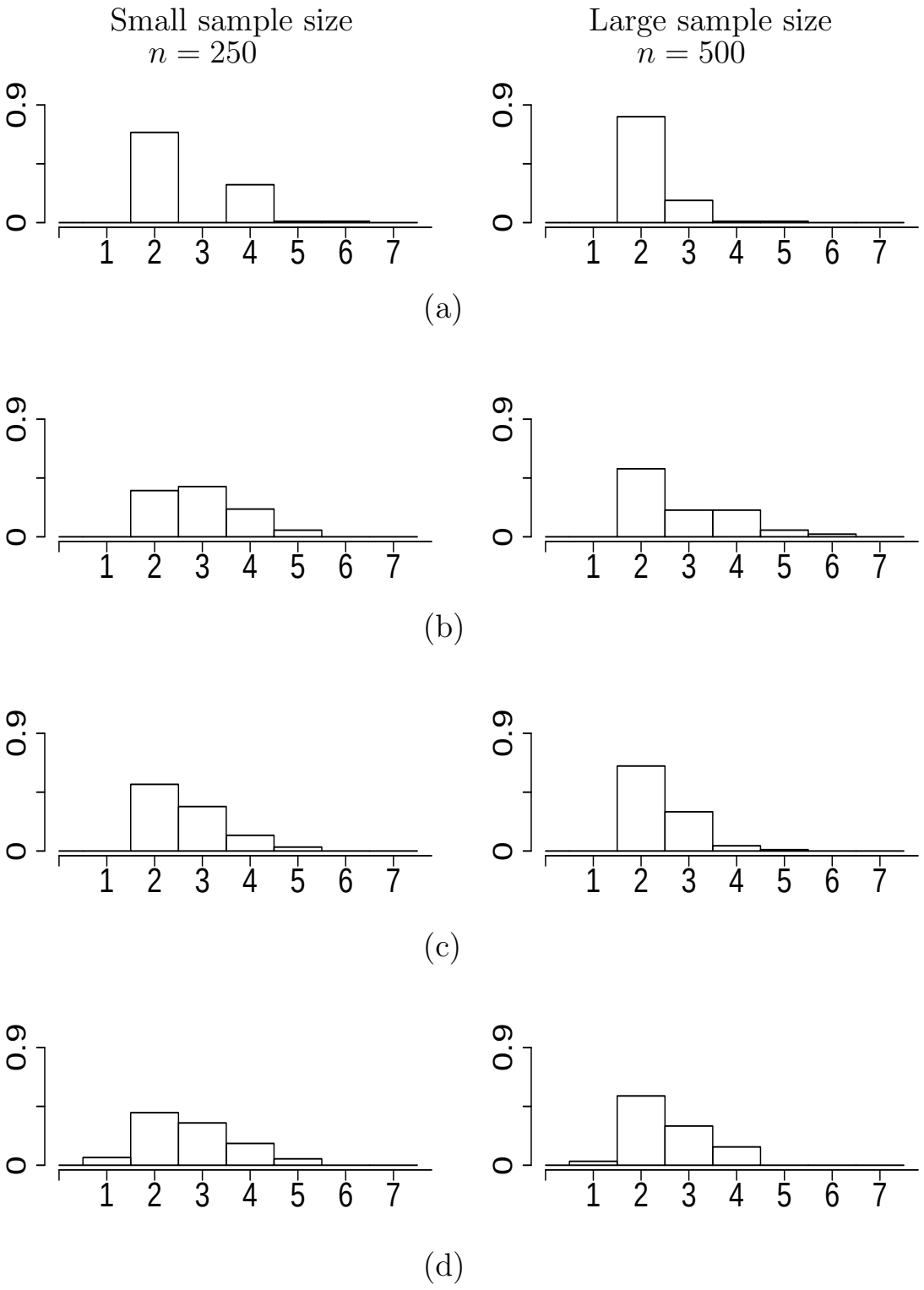

Figure 4.4: The histogram of the most frequently selected profiles over simulated samples from Dirichlet process with cutoff 5%: (a) strong ρ and strong η, (b) mixed ρ and strong η, (c) strong ρ and mixed η, and (d) mixed ρ and mixed η

Table 4.2: The percentage of the most frequently selected profiles over simulated samples from Dirichlet process with cutoff 5%

Sample size	ρ	η	Number of profiles 1	2	3	4	5	6	7
250	Strong	Strong	0.0	69.0	29.0	1.0	1.0	0.0	0.0
	Mixed	Strong	0.0	35.0	38.0	21.0	5.0	1.0	0.0
	Strong	Mixed	0.0	51.0	34.0	12.0	3.0	0.0	0.0
	Mixed	Mixed	6.0	41.0	33.0	17.0	5.0	0.0	0.0
500	Strong	Strong	0.0	81.0	17.0	1.0	1.0	0.0	0.0
	Mixed	Strong	0.0	50.0	20.0	20.0	5.0	2.0	0.0
	Strong	Mixed	0.0	65.0	30.0	4.0	1.0	0.0	0.0
	Mixed	Mixed	3.0	53.0	30.0	14.0	0.0	0.0	0.0

Since 5% cutoff is recognized as more effective in producing satisfying results without being unreasonably liberal in screening profiles, we consider 5% as the appropriate threshold in determining the number of profiles and the percentages of the most frequently selected profiles are distributed in Table 4.2. For cases where more than 7 profiles are generated are not shown in the table since the sizes are insignificantly small.

4.3 Extended Cases

In previous sections, we already provided empirical results showing that both presented techniques are able to learn the class structure correctly. However, to investigate the performance and validate their effectiveness in selecting the right profiles, we consider applications more generally. We draw samples from LCPA models with two classes, two profiles, four items over varying number of time measurements (i.e, $T = 3, 4$ and 5). Table 4.3 indicates that with strong-strong pairs, RJMCMC can optimize the performance when more time measurements are involved. However, the performance degradation is noted when weak measurements are

Table 4.3: The percentage of selecting the correct number of profiles ($S = 2$) over simulated samples from RJMCM and Dirichlet process

Sample size	ρ	η	RJMCMC			DP with 5% cutoff		
			$T = 3$	$T = 4$	$T = 5$	$T = 3$	$T = 4$	$T = 5$
250	Strong	Strong	74.6	78.4	80.5	51.0	69.0	91.0
	Mixed	Strong	76.2	74.8	73.0	34.0	35.0	60.0
	Strong	Mixed	53.6	45.2	58.3	50.0	51.0	65.0
	Mixed	Mixed	54.3	34.8	32.6	38.0	41.0	55.0
500	Strong	Strong	77.9	82.2	86.2	57.0	81.0	92.0
	Mixed	Strong	64.7	52.2	46.9	45.0	51.0	66.0
	Strong	Mixed	27.4	19.6	22.2	51.0	65.0	82.0
	Mixed	Mixed	33.1	14.4	17.7	48.0	53.0	65.0

given, which we believe is because weak measurements vaguely distinguish the competing models and most of time, RJMCMC would prefer a complex model as opposed to a simple one. On the other hand, we found out that Dirichlet process can learn the profiles more accurately when there are more time measurements involved. Besides, the results demonstrate that working with large data is always accompanied by significant improvement but enlarging the sample size might not be a feasible approach in modeling the complexity.

4.4 Discussion

Latent stage-sequential process is an attractive tool for many areas of substantive research. Like other mixture models, however, it can be difficult to estimate the number of latent components such as classes and profiles. We have illustrated two Bayesian approaches, RJMCMC and Dirichlet process, for the latent class profile analysis. Although this model is certainly not representative of all latent stage-sequential process, it nevertheless lends several important insight which we believe have general relevance.

First, both RJMCMC and Dirichlet process can learn the model without assuming the number of latent components in advance. Comparing with the technical challenges that RJMCMC might bring, Dirichlet process may be preferred in terms of flexibility and consistency. Dirichlet process technique relaxes the constraint on the number of classes placed at each measurement occasion and it performs well with large sample size. Dirichlet process is conceptually simple and does not require intensive computation since there is no need for Dirichlet process to design jumping proposals and computing the acceptance rates. Technically, Dirichlet process is easily employed on the intuitive basis and readily extended to study longitudinal data.

In the Bayesian mixture modeling, label switching is one of the most common issues and it will lead to poor parameter estimation. However, It does no harm in finding the number of components for the LCPA model. In the simulation study, we run the algorithms in the presence of label switching, but the conclusions we inferred here are unaffectedly viable. To capture the structural change in classes during experiments, we may allow the RJMCMC to vary the number of classes at different measurement occasions, but it leads to insurmountable technical hurdles. However, allowing various number of classes over time is possible in DDPM and it is an appealing aspect of Dirichlet process approach comparing to the RJMCMC. In this study, we present Bayesian model selection analysis on LCPA without any covariates to predict the prevalence of the profile; however, predicator-dependent kernel stick-breaking process has already increased the interest. It is utilized in choosing the priors for an unknown probability measure (Dunson and Park, 2008) and variable selection problems (Chung and Dunson, 2009). Adding predictors in the prior consideration gives different insights into how the classes are formed under the influence of predictors and it is understood as a dependent Dirichlet Process. Future work should explore the model selection for the LCPA regression

model.

Some specific limitation to this study include the choice of jumping rules in RJMCMC. The simulation study can only serve as a preliminary exploration and it is worth noting that a comparison between RJMCMC and Dirichlet process cannot be generalized unless the ingenuity of the proposal mechanism we proposed in the RJMCMC is ensured. Besides, we note that the results of the Dirichlet process may not be promising when mixed-mixed pairs are considered since it did not gain much accuracy when sample size was increased from 250 to 500. We believe that the appropriate sample size should be investigated for Dirichlet process to produce accurate results under different scenarios on the measurement parameters for future study. We provided a limited demonstration to help elucidate the the two potential methodologies for the model selection in LCPA. Our hope is that substantive practitioners will be able to utilize this demonstration in their research and that it may provide helpful guidance for the implementation of these two methods. In summary, this presentation provided the coverage of the implementation of RJMCMC and the Dirichlet process for the LCPA model and showed the performance of those methods in selecting the number of classes and profiles over repeated samples. Future work should investigate more tailored RJMCMC and the Dirichlet process specialized for LCPA and compare the performance formally via simulation study. We expect that each method will display its own unique strengths and weaknesses under different conditions.

Chapter 5

Parameter Estimation

5.1 Preliminary

Parameters for latent class profile analysis (LCPA) are easily estimated by maximum likelihood via EM algorithm or Bayesian method via Markov chain Monte Carlo. However, the local maximum problem is a long-standing issue in any hill-climbing optimization technique for the LCPA model. In this study, we propose to apply two probabilistic optimization techniques using the deterministic annealing framework in order to deal with multiple local modalities in the LCPA model. The deterministic annealing approaches are implemented with an efficient recursive formula in the step for the parameter update.

5.2 Expectation-Maximization

The EM algorithm is an iterative procedure which is widely employed for finding maximum likelihood and solving missing value problems (Dempster et al., 1977). Estimation of parameters in the LCPA model may be regarded as an estimation problem with missing data:

the realized values of the latent-profile variable U_i and latent-class memberships C_{it} for all t are missing for individuals $i = 1, \ldots, n$. The observed data loglikelihood given the covariates $\mathbf{x}_i$ is

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log \left(Pr(\mathbf{Y}_i = \mathbf{y}_i) \mid \mathbf{x}_i \right) \\ &= \sum_{i=1}^{n} \log \left(\sum_{s=1}^{S} \gamma_s(\mathbf{x}_i) \prod_{t=1}^{T} R_s^{(t)} \right) \end{aligned} \tag{5.1}$$

Direct maximization of the observed-data loglikelihood is complicated because we need to solve optimization problem with sum of logarithm functions; however, ML estimates could be calculated directly in closed form if there is no latent variables. EM is computationally tractable to maximizing the loglikelihood function by repeatedly solving the complete-data problem. EM is an iterative procedure in which each iteration consists of two steps, the E-step and M-step.

In the E-step, we compute the expected values of the unobservable cell counts for the cross-classification by $Y_{i1}, \ldots, Y_{iM}$ and U_i, $C_{i1}, \ldots, C_{iT}$, given the observed data and provious parameter estimates. In the M-step, we maximize the complete-data logliklihood function under the assumption that the missing data are known and the missing data from the E-step are used to substitute the actural missing data.

For the LCPA model, the EM algorithm has some important advantages over other methods for finding ML estimates: it converges to gradually but reliably; it guarantees that the resulting estimates lie within the parameter space; it does not require inversion of the Hessian matrix at each iteration, so it demands less computational time than the Newton-Raphson; and it does not require a carefully chosen set of initial values to start the iterative

process in order to converge to the final solution. For all these reasons, the EM algorithm has been the preferred method for ML-based estimation in the LCPA model. If covariates are included, the expected complete-data loglikelihood can no longer be maximized using closed-form expressions; the M-step requires an iterative procedure equivalent to a routine for fitting a standard baseline-category multinomial logistic-regression model. Van der Heijden et al. (1996) and Bandeen-Roche et al. (1997) utilized an EM algorithm that incorporates a Newton-Raphson steps for the multinomial logit model into the standard latent class estimating procedure devised by Goodman (1974). The E- and M-steps of this procedure are given below.

5.21 E-step

In the E-step, the posterior probabilities of latent class and profile memberships for the ith individual are calculated under the provisional parameter estimates from the previous iteration. By Bayes' Theorem, these posterior probabilities are

$$\begin{aligned}
\theta_{i,(s,c)} &= Pr(U_i = s, \mathbf{C}_i = \mathbf{c}_i \mid \mathbf{Y}_i = \mathbf{y}_i, \mathbf{x}_i) \\
&= \frac{\gamma_s(\mathbf{x}_i) \prod_{t=1}^{T} R^{(t)}_{c_t|s}}{\sum_{s=1}^{S} \gamma_s(\mathbf{x}_i) \prod_{t=1}^{T} \left\{ \sum_{c_t=1}^{C} \eta^{(t)}_{c_t|s} R^{(t)}_{c_t|s} \right\}}
\end{aligned}$$

5.22 M-step

In the M-step, updated parameter estimates are obtained by maximizing the expected complete-data loglikelihod, regarding the latent variables if they were observed. If U_i and c_i were known, let the contribution of the ith individual to the complete-data likelihood

be denoted $L_i^*(\boldsymbol{\theta})$ and the logarithm of the complete-data likelihood function contribution $l_i^*(\boldsymbol{\theta}) = \log(L_i^*(\boldsymbol{\theta}))$ can be written as

$$\begin{aligned} l_i^*(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \log \gamma_s(\mathbf{x}_i)\theta_{is} + \sum_{i=1}^{n}\sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{c_t=1}^{C} \theta_{i(s,c_t)}^{(t)} \log \eta_{c_t|s}^{(t)} \\ &+ \sum_{t=1}^{T}\sum_{c_t=1}^{C} \theta_{ic_t}^{(t)} \sum_{m=1}^{M}\sum_{k=1}^{r_m} I(y_{imt=k}) \log \rho_{mkt|c_t} \end{aligned} \tag{5.2}$$

where $\theta_{is} = \sum_{c_1} \cdots \sum_{c_T} \theta_{i(s,\mathbf{c})}, \theta_{i(s,c_t)}^{(t)} = \prod_{j \neq t} \sum_{c_j} \theta_{i(s,\mathbf{c})}$ and $\theta_{ic_t}^{(t)} = \sum_s \theta_{i(s,c_t)}^{(t)}$. According to (2.6) for each $s = 1, \ldots, S-1$, the maximizing vector $\widehat{\boldsymbol{\beta}}_s$ can be derived by setting

$$\frac{\partial}{\partial \beta_{sr}} \sum_{i=1}^{n} \theta_{is} \log \frac{\exp(x_i' \boldsymbol{\beta}_s)}{\sum_{j=1}^{S} \exp(x_i' \boldsymbol{\beta}_j)} = 0 \text{ for } r = 1, \ldots, p. \tag{5.3}$$

To approach the roots of these non-linear equations, we use NR algorithm to expedite the calculation. In the initial stages of EM, the parameter estimates begin far from the desired values and Newton's method might fail to converge. Burdened with the caveats, we only execute one iteration of NR method within each M-step. In the NR method, an estimate $\boldsymbol{\beta}_s^{new}$ is updated by

$$\boldsymbol{\beta}_s^{new} = \boldsymbol{\beta}_s^{old} - \left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)^{-1} \left(\frac{\partial l}{\partial \boldsymbol{\beta}} \right) |_{\boldsymbol{\beta} = \boldsymbol{\beta}^{old}} \tag{5.4}$$

where $\boldsymbol{\beta}$ is the vectorized parameter containing all parameter elements.

5.3 Recursive Formula

Although direct maximization of the observed-data loglikelihood defined in (5.1) is complicated, ML estimates can be easily calculated if the latent memberships (i.e., $\mathbf{c}$ and s) were known.

In the E-step, the conditional probabilities of class and profile memberships for the ith individual are calculated under the provisional parameter estimates from the previous iteration. By Bayes' theorem, these conditional probabilities are given by

$$\begin{aligned} \theta_{i(s,c_1,\ldots,c_T)} &= P(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i, \mathbf{x}_i) \\ &= \frac{L_i^*}{\sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} L_i^*}, \end{aligned} \tag{5.5}$$

where L_i^* is defined in (2.3).

In the E-step, however, the complexity of computation and the memory demand grow when the number of time periods T increases. Instead of computing the joint posterior probabilities given in (5.5), we apply the recursive formula to the E-step by adopting the forward-backward algorithm (Chib, 1996; MacKay, 1997). Let α and λ represent the forward and backward probabilities, respectively:

$$\begin{aligned} \alpha_{it}(c_t, s) &= P(\mathbf{Y}_1 = \mathbf{y}_{i1}, \ldots, \mathbf{Y}_t = \mathbf{y}_{it}, C_t = c_t \mid U = s, \mathbf{x}_i) \\ \lambda_{it}(c_t, s) &= P(\mathbf{Y}_{t+1} = \mathbf{y}_{i(t+1)}, \ldots, \mathbf{Y}_T = \mathbf{y}_{iT} \mid C_t = c_t, U = s, \mathbf{x}_i). \end{aligned}$$

Both α and λ functions then can be computed by the following recursive representations

using the first-order Markov chain:

$$\begin{aligned}
\alpha_{it}(c_t, s) &= \sum_{c_{t-1}=1}^{C} \alpha_{i(t-1)}(c_{t-1}, s)\eta^{(t)}_{c_t|s} \prod_{m=1}^{M}\prod_{k=1}^{r_m}\left[\rho_{mkt|c_t}\right]^{I(y_{imt}=k)} \\
\lambda_{it}(c_t, s) &= \sum_{c_{t+1}=1}^{C} \lambda_{i(t+1)}(c_{t+1}, s)\eta^{(t+1)}_{c_{t+1}|s} \\
&\times \prod_{m=1}^{M}\prod_{k=1}^{r_m}\left[\rho_{mk(t+1)|c_{t+1}}\right]^{I(y_{im(t+1)}=k)}
\end{aligned}$$

With the forward and backward functions, the marginalized conditional probability of the latent component membership at time t can be computed by

$$\begin{aligned}
\theta^{(t)}_{i(s,c_t)} &= P(U = s, C_t = c_t \mid \mathbf{y}_i, \mathbf{x}_i) \\
&= \frac{\gamma_s(\mathbf{x}_i)\alpha_{it}(c_t, s)\lambda_{it}(c_t, s)}{\sum_{s=1}^{S}\gamma_s(\mathbf{x}_i)\sum_{c_T=1}^{C}\alpha_{iT}(c_T, s)},
\end{aligned} \tag{5.6}$$

for $s = 1, \ldots, S$, $c_t = 1, \ldots, C$, and $t = 1, \ldots, T$.

In the M-step, updated parameter estimates are obtained by maximizing the expected complete-data loglikelihood, regarding the latent variables if they were observed. The contribution of the ith individual to the complete-data loglikelihood $\ell^*_i = \log L^*_i$ can be written as $\ell^*_i(\boldsymbol{\theta})$ as (5.2). The first sum in (5.2), which relates to the regression coefficients (i.e., the $\boldsymbol{\beta}$-parameters), is the loglikelihood function for the multinomial logit model, except that the unobserved counts for s are replaced by the fractional expectations $\sum_{i=1}^{n}\theta_{is}$. Updated estimates for the regression coefficients can be calculated with the standard Newton-Raphson method for multinomial logistic regression, provided that the computational routines allow

fractional responses rather than integer counts. The other model parameters can be obtained by

$$\begin{aligned}
\hat{\eta}^{(t)}_{c_t|s} &= \frac{\sum_{i=1}^{n} \theta^{(t)}_{i(s,c_t)}}{\sum_{i=1}^{n} \theta_{is}} \\
\hat{\rho}_{mkt|c_t} &= \frac{\sum_{i \in obs^{(t)}_m} \theta^{(t)}_{ic_t} I(y_{imt} = k) + \sum_{i \in mis^{(t)}_m} \theta^{(t)}_{ic_t} \rho^{*}_{mkt|c_t}}{\sum_{i=1}^{n} \theta^{(t)}_{ic_t}}
\end{aligned}$$

for $m = 1, \ldots, M$, $k = 1, \ldots, r_m$, $s = 1, \ldots, S$, $c_t = 1, \ldots, C$, and $t = 1, \ldots, T$. Here $obs^{(t)}_m$ denotes the set of individuals who respond to the mth item at time t, $mis^{(t)}_m$ denotes the set of individuals who fail to respond to the mth item at time t, and $\rho^{*}_{mkt|c_t}$ is the provisional parameter estimate.

5.4 Local Modality

We understand that the local maximum problem could be mitigated by starting from multiple initial values and then tracking the optimal solution which achieves the highest likelihood. However, in the case of high-dimensional parameter space, a large number of performing the EM algorithm for each initialization is required and therefore become computationally intractable. In order to relax the initialization dependence for the LCPA model, we introduce two reformulated versions of the standard EM algorithm, namely split-and-merge EM (SMEM) (Ueda et al., 2000), deterministic annealing EM (DAEM) (Ueda and Nakano, 1998) and a variational Bayes learning algorithm, referred as and deterministic annealing variational Bayes (DAVB) (Katahiral et al., 2008).

The general idea of SMEM aims at increasing the log likelihood value gradually by tactically choosing three components as candidates for split and merge. A number of merge and split candidates are selected, usually 5 candidate sets are recommended for each iteration and then using partial expectation step to expedite the expectation computation. The idea of performing split and merge operations has been successfully applied to the EM in the Gaussian mixture models through considering local Kullback divergence as a split criteria and posterior probabilities as a merge index. Here we roughly sketch the steps for partial expectation step on classes and profiles in LCPA models.

Ueda and Nakano (1998) proposed a deterministic annealing version of the EM (DAEM) algorithm to reduce the impact that inappropriate initial values could cause. Even though there is no evidence to prove it is grounded in theory, it has been successfully applied in the realm of local maximum issues and such an annealing framework has proven effective in improving the performance of the standard approaches (Itaya et al., 2004; Park et al., 2005).

5.41 Split-and-Merge EM

Let $J_C = (c_1, c_2, c_3)$ denote the split-and-merge candidates where c_1 and c_2 are chosen by the pre-specified rule to form a new c^* and c_3 is chosen to be divided into two new classes $\tilde{c_1}$ and $\tilde{c_2}$. The partial update step for newly-generated classes $l = c^*, \tilde{c_1}, \tilde{c_2}$ are re-estimated by the following:

$$\theta^{(t)}_{i(s,l)} = \left(\sum_{m \in J_C} \theta^{(t)}_{i(s,m)} \right) \times \frac{\alpha_t(l,s)\lambda_t(l,s)}{\sum_{l=c^*,\tilde{c_1},\tilde{c_2}} \alpha_t(l,s)\lambda_t(l,s)}$$

Similarly, for profiles, we set $J_S = (s_1, s_2, s_3)$ as the split-and-merge candidates where s_1 and s_2 are merged to form s^* and s_3 is split into two new profiles $\tilde{s_1}$ and $\tilde{s_2}$. In the partial update step, the posterior probability of profile components $l = s^*, \tilde{s_1}, \tilde{s_2}$ are re-estimated by

$$\theta^{(t)}_{i(l,c_t)} = \left(\sum_{m \in J_S} \theta^{(t)}_{i(m,c_t)} \right) \times \frac{\gamma_l \alpha_t(c_t, l) \lambda_t(c_t, l)}{\sum_{k=s^*,\tilde{s_1},\tilde{s_2}} \gamma_k \alpha_t(c_t, k) \lambda_t(c_t, k)}$$

SMEM algorithm is expected to be abler to travel low log-likelihood area and solve the problem of initialization dependence only if a clear set of principles for split and merge has been instituted. In the case of LCPA applications, the correspondence between adjacent measurement occasions is inextricably bridged and there is no easy heuristical method to relocate the mixture components in the data space. Generally speaking, SMEM intensifies the extent of the technical difficulties.

5.42 Deterministic Annealing EM Algorithm

Ueda and Nakano (1998) proposed a deterministic annealing version of the EM algorithm (DAEM) to find a set of parameter estimates on the global mode of the loglikelihood function. In the DAEM algorithm for LCPA, the goal to maximize the loglikelihood function is reformulated as the problem of maximizing the function $F(\omega) = \sum_{i=1}^{n} F_i(\omega)$, where

$$F_i(\omega) \;=\; \frac{1}{\omega} \log \sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} (L_i^*)^{\omega}.$$

Note that the function $F(\omega)$ equals to the observed-data loglikelihood ℓ in (5.1) when $\omega = 1$. To maximize the function $F(\omega)$, we introduce a random density function $q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i)$ and find the lower bound of $F_i(\omega)$ by Jensen's inequality (Cover and Thomas, 1991):

$$\begin{aligned} F_i(\omega) &\geq \frac{1}{\omega} \sum_{s,\mathbf{c}} q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i) \log \left[\frac{(L_i^*)^\omega}{q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i)} \right] \\ &= \sum_{s,\mathbf{c}} q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i) \log L_i^* \\ &\quad -\frac{1}{\omega} \sum_{s,\mathbf{c}} q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i) \log q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i). \end{aligned} \tag{5.7}$$

In the first step of DAEM, we find the optimal choice for $q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i)$ by taking functional derivatives with respect to $q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i)$ and setting it as zero under the constraint $\sum_{s,\mathbf{c}} q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i) = 1$. Then the optimal choice for $q(U = s, \mathbf{C} = \mathbf{c} \mid \mathbf{y}_i,)$ would be

$$\underline{\theta}_{i(s,c_1,\ldots,c_T)} = \frac{(L_i^*)^\omega}{\sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} (L_i^*)^\omega}. \tag{5.8}$$

It is worth noting that with values of ω close to zero, the function $\underline{\theta}_{i(s,c_1,\ldots,c_T)}(\omega)$ is uniformly distributed across all the latent components (i.e., classes and profiles). As the value of ω increases toward one, however, $\underline{\theta}_{i(s,c_1,\ldots,c_T)}(\omega)$ agrees with the conditional probabilities $\theta_{i(s,c_1,\ldots,c_T)}$ given in (5.5).

In the second step of DAEM, the model parameters are updated by maximizing the modified expected complete-data loglikelihood $\ell^*(\omega) = \sum_{i=1}^{n} \ell_i^*(\omega)$. The contribution

of the ith individual to the $\ell_i^*(\omega)$ can be written as

$$\begin{aligned}\ell_i^*(\omega) &= \sum_{s=1}^{S} \underline{\theta}_{is}(\omega) \log \gamma_s(\mathbf{x}_i) + \sum_{s=1}^{S}\sum_{t=1}^{T}\sum_{c_t=1}^{C} \underline{\theta}^{(t)}_{i(s,c_t)}(\omega) \log \eta^{(t)}_{c_t|s} \\ &\quad + \sum_{t=1}^{T}\sum_{c_t=1}^{C} \underline{\theta}^{(t)}_{ic_t}(\omega) \sum_{m=1}^{M}\sum_{k=1}^{r_m} I(y_{imt=k}) \log \rho_{mkt|c_t}, \end{aligned} \tag{5.9}$$

where $\underline{\theta}_{is}(\omega) \propto \left(\theta_{is}\right)^{\omega}$, $\underline{\theta}^{(t)}_{i(s,c_t)}(\omega) \propto \left(\theta^{(t)}_{i(s,c_t)}\right)^{\omega}$, and $\underline{\theta}^{(t)}_{ic_t}(\omega) \propto \left(\theta^{(t)}_{ic_t}\right)^{\omega}$. These quantities can be calculated by the recursive formula given in (5.6), and the joint conditional probabilities are not necessary in the second step of DAEM. As the standard EM algorithm, updated estimates for the regression coefficients can be calculated with the Newton-Raphson algorithm for the first sum in (5.9). The other model parameters can be obtained by

$$\begin{aligned}\hat{\eta}^{(t)}_{c_t|s} &= \frac{\sum_{i=1}^{n} \underline{\theta}^{(t)}_{i(s,c_t)}(\omega)}{\sum_{i=1}^{n} \underline{\theta}_{is}(\omega)} \\ \hat{\rho}_{mkt|c_t} &= \frac{\sum^{n}_{i\in obs^{(t)}_m} \underline{\theta}^{(t)}_{ic_t}(\omega) I(y_{imt}=k) + \sum^{n}_{i\in mis^{(t)}_m} \underline{\theta}^{(t)}_{ic_t}(\omega) \rho^*_{mkt|c_t}}{\sum_{i=1}^{n} \underline{\theta}^{(t)}_{ic_t}(\omega)}\end{aligned}$$

for $m = 1, \ldots, M$, $k = 1, \ldots, r_m$, $s = 1, \ldots, S$, $c_t = 1, \ldots, C$, and $t = 1, \ldots, T$.

We adopt ω=(.001, .01, .1, .2, .3, .4, .48, .58, .69, .83, 1) as an annealing schedule in DAEM. The algorithm is initialized with the value of ω close to zero (i.e., .001) and its converged parameters will be used as the starting values for the next one (i.e., .01). By repeating this procedure until ω reaches one, the DAEM algorithm will find the parameter

estimates on the global mode of the loglikelihood function.

5.43 Deterministic Annealing Variational Bayes Algorithm

Let $\mathbf{\Theta}$ denote the vectorized model parameters for the LCPA model. In the deterministic annealing variational Bayes (DAVB) algorithm, we can consider $\mathbf{\Theta}$, a set of unknown parameters, as the random quantities and incorporate the prior information for $\mathbf{\Theta}$. Let $\varphi(\mathbf{\Theta})$ denote a prior distribution of $\mathbf{\Theta}$. Further, let $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ indicate the individuals' class and profile memberships, where $\mathbf{z}_i$ is a $T+1$ dimensional array for the ith individual such that $z_{i(s,c_1,\ldots,c_T)} \in \{0,1\}$ and $\sum_{s=1}^{S}\sum_{c_1=1}^{C}\cdots\sum_{c_T=1}^{C} z_{i(s,c_1,\ldots,c_T)} = 1$. That is, if individual i belongs to the profile s and the class membership $\mathbf{c} = (c_1, \ldots, c_T)$ from initial time $t = 1$ to time T, then $z_{i(s,c_1,\ldots,c_T)}$ equals 1 and 0 otherwise.

Constructed with the similar logic of the DAEM, the goal of DAVB to approximate the distribution over latent variables and model parameters can be rephrased by the problem of maximizing $F(\omega) = \sum_{i=1}^{n} F_i(\omega)$, where

$$F_i(\omega) \;=\; \frac{1}{\omega}\log\sum_{s=1}^{S}\sum_{c_1=1}^{C}\cdots\sum_{c_T=1}^{C}\int P(\mathbf{Y}=\mathbf{y}_i, \mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta})^{\omega} d\mathbf{\Theta}.$$

By introducing a random distribution $q(\mathbf{Z} = \mathbf{z}_i, \mathbf{\Theta})$, $F_i(\omega)$ can be lower bounded by

Jensen's inequality (Cover and Thomas, 1991):

$$F_i(\omega) \geq \sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} \int q(\mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta}) \log P(\mathbf{Y}=\mathbf{y}_i, \mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta}) d\mathbf{\Theta}$$
$$-\frac{1}{\omega} \sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} \int q(\mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta}) \log q(\mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta}) d\mathbf{\Theta}.$$

If $q(\mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta})$ has a factored form (i.e., $q(\mathbf{Z}=\mathbf{z}_i, \mathbf{\Theta}) = Q(\mathbf{Z}=\mathbf{z}_i) r(\mathbf{\Theta})$), the function $F_i(\omega)$ is then lower bounded as

$$\begin{aligned} F_i(\omega) \geq & \sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} \\ & \int Q(\mathbf{Z}=\mathbf{z}_i) r(\mathbf{\Theta}) \log P(\mathbf{Y}=\mathbf{y}_i, \mathbf{Z}=\mathbf{z}_i \mid \mathbf{\Theta}) \varphi(\mathbf{\Theta}) d\mathbf{\Theta} \\ & -\frac{1}{\omega} \left\{ \sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} Q(\mathbf{Z}=\mathbf{z}_i) \log Q(\mathbf{Z}=\mathbf{z}_i) \right. \\ & \left. + \int r(\mathbf{\Theta}) \log r(\mathbf{\Theta}) d\mathbf{\Theta} \right\}. \end{aligned} \tag{5.10}$$

In the first step of DAVB, we iteratively maximizes the lower bound of $F_i(\omega)$ with respective to $Q(\mathbf{Z}=\mathbf{z}_i)$ by taking functional derivative and setting it equal to zero under the constraint $\sum_{s=1}^{S} \sum_{c_1=1}^{C} \cdots \sum_{c_T=1}^{C} Q(\mathbf{Z}=\mathbf{z}_i) = 1$. The optimal choice for $Q(\mathbf{Z}=\mathbf{z}_i)$ would be $\underline{\theta}_{i(s,c_1,\ldots,c_T)}(\omega)$, which is the optimal choice for $q(U=s, \mathbf{C}=\mathbf{c} \mid \mathbf{y}_i)$ given in (5.8).

In the second step, the lower bound of $F_i(\omega)$ is maximized with respective to $r(\mathbf{\Theta})$ and the model parameters are updated based on the variational posteriors. We consider

multivariate normal distribution $N_{p\times(S-1)}(\mathbf{0},\mathbf{I})$ as a prior for the β-parameters. At each cycle of Gibbs sampler, the vectorized coefficients $\boldsymbol{\beta}$ are updated by Metropolis algorithm (Robert and Casella, 2004). A candidate for the next coefficient vector $\boldsymbol{\beta}^{\mathcal{C}}$ is generated from $N_{p\times(S-1)}(\hat{\boldsymbol{\beta}},\delta\boldsymbol{\Sigma})$ at each iteration, where $\hat{\boldsymbol{\beta}}$ is the generated sample from the previous iteration. In this paper, we adjust the value of δ in order to control the acceptance rate within a recommended range from .18 to .30 (Gelman et al., 1997). The variance $\boldsymbol{\Sigma}$ is the negative inverse of the $\boldsymbol{\beta}$-submatrix in the Hessian from the complete-data loglikelihood $\ell^*(\omega)$ evaluated at the DAEM estimates. The candidate vector for $\boldsymbol{\beta}^{\mathcal{C}}$ is accepted with probability $\alpha(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^{\mathcal{C}}) = min\left(1,\exp\left(\omega A\right)\right)$, where

$$\begin{aligned} A &= -\frac{1}{2}\left(|\boldsymbol{\beta}|^2 - |\hat{\boldsymbol{\beta}}|^2\right) + \sum_{i=1}^{n}\sum_{s=1}^{S}\underline{\theta}_{is}(\omega)\mathbf{x}_i'(\boldsymbol{\beta}_s^{\mathcal{C}} - \hat{\boldsymbol{\beta}}_s) \\ &\quad - \sum_{i=1}^{n}\left\{\log\left[\sum_{s=1}^{S}\exp(\mathbf{x}_i'\boldsymbol{\beta}_s^{\mathcal{C}})\right] - \log\left[\sum_{s=1}^{S}\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}_s)\right]\right\}. \end{aligned}$$

Applying the Jeffreys' priors to the measurement parameters $\boldsymbol{\eta}_s^{(t)} = (\eta_{1|s}^{(t)},\ldots,\eta_{C|s}^{(t)})$ and $\boldsymbol{\rho}_{mt|c} = (\rho_{m1t|c},\ldots,\rho_{mr_mt|c})$, new parameters can be drawn from $\boldsymbol{\eta}_s^{(t)} \sim$ Dirichlet$(\tau_{1|s}^{(t)},\ldots,\tau_{C|s}^{(t)})$ and $\boldsymbol{\rho}_{mt|c} \sim$ Dirichlet$(\nu_{m1t|c},\ldots,\nu_{mr_mt|c})$, where

$$\begin{aligned} \tau_{c|s}^{(t)} &= \omega\left[\sum_{i=1}^{n}\underline{\theta}_{i(s,c)}^{(t)}(\omega) - \frac{1}{2}\right] + 1 \\ \nu_{mkt|c} &= \omega\left[\sum_{i\in obs_m^{(t)}} I(y_{imt}=k)\underline{\theta}_{ic}^{(t)}(\omega) + \sum_{i\in mis_m^{(t)}}\underline{\theta}_{ic}^{(t)}(\omega)\rho_{mkt|c}^* - \frac{1}{2}\right] + 1 \end{aligned}$$

for $c = 1, \ldots, C$, $s = 1, \ldots, S$, $t = 1, \ldots, T$, $k = 1, \ldots, r_m$, and $m = 1, \ldots, M$.

In DAVB, we adopt the same annealing schedule and proceed with the same procedure as in DAEM.

Chapter 6

Model Diagnosis

For the inferences drawn from a model being meaningful, we need to establish a correctly specified model and identifiability is fundamental for having valid parameter estimates. According to the definition by Goodman (1974), the parameters of an LC model without covariates are said to be locally identifiable at a particular value $\boldsymbol{\theta}^*$ if for some open neighborhood of it, the loglikelihood function has a unique optimum value at $\boldsymbol{\theta}^*$. The definition is no problem to be further applied to LCPA models. Let us assume the number of a response pattern as a particular combination of responses to the manifest items $\mathbf{Y}_{1t}, \ldots, \mathbf{Y}_{Mt}$ for all $t = 1, \ldots, T$ is $(\prod_{m=1}^{M} r_m)^T$. A saturated model for $\mathbf{Y}_{1t}, \ldots, \mathbf{Y}_{Mt}$ for all $t = 1, \ldots, T$ would have $(\prod_{m=1}^{M} r_m)^T - 1$ nonredundant parameters because the probabilities for the response patterns must sum to 1. A necessary but not sufficient condition to make the LCPA model locally identifiable is that the number of nonredundant parameters in the saturated model must be greater than or equal to the number of free parameters in $\boldsymbol{\gamma}, \boldsymbol{\eta}$ and $\boldsymbol{\rho}$ (Goodman, 1974; Clogg and Goodman, 1984). The saturated model has $(\prod_{m=1}^{M} r_m)^T - 1$ nonredundant parameters and the LCPA model has $S - 1 + ST(C-1) + C(\sum_{m=1}^{C} r_m - M)$ when we assume the number of classes is

fixed as C over time. When

$$\left(\prod_{m=1}^{M} r_m\right)^T - 1 > S - 1 + ST(C-1) + C\left(\sum_{m=1}^{C} r_m - M\right) \tag{6.1}$$

the LCPA model might be identifiable but not necessarily so.

Traditionally, we diagnose the identifiability by carrying out the Hessian matrix evaluated at the ML estimates $\boldsymbol{\theta}^*$. If it has full rank or the inverse exists, we can ascertain the parameter estimates are at least locally identifiable. In practice, however, it is difficult to distinguish between situations where the Hessian is nearly singular and where it is exactly singular because of the imprecision of floating-point computations. As pointed out by Formann (2003), in situations where some estimated parameters lie on the boundary (estimates are close to boundary value), we'd better fix those values in order to make the remaining parameters identifiable. At fact, a covariance matrix with large variance indicates boundary problems. The Hessian matrix for LCPA with covariates may have result in large dimension structure and the derivatives therefore become unappealing. As shown in (2.8), the marginalization implies that an LCPA with logistic regression is locally identifiable if the following three conditions are satisfied (Bandeen-Roche et al., 1997):

1. an LCPA marginalized over covariates is locally identifiable;

2. if the design matrix $(\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ has full column rank;

3. at least one individual has nonzero $\gamma_s(\mathbf{x}_i)$ for all $s = 1, 2, \ldots, S$.

Some new monitoring methods for identifiability have been proposed, for example, Kim and Lindsay (2009) define a new concept, the degree of identifiability of the parameters in the likelihood. This data-dependent identifiability is measured through modal regions that is

the maximum connected subset containing ML estimate $\boldsymbol{\theta}^*$ and determined by a significance level α. The model is said to be identifiable at confidence level $100(1-\alpha)\%$ at $\boldsymbol{\theta}^*$ if the modal region containing $\boldsymbol{\theta}^*$ is disjoint from any other permutated ones. It comes no surprise that as α decreases, the modal regions are hardly well-labeled because overlapping might occur. The level of identifiability can be quantified by finding the infimum of the value α such that the modal region is well separated. Besides, Yao and Lindsay (2009) developed marginal discriminant plots to measure the degree of modal separation.

Chapter 7

Data Analysis on National Longitudinal Survey of Youth 1997 (NLSY97)

7.1 Data

The proposed model selection (i.e., Dirichlet process) and parameter estimation (i.e., deterministic annealing) methods are applied to the drinking items drawn from the National Longitudinal Survey of Youth 1997 (NLSY97), a survey that explores the transition from school to work and from adolescence to adulthood in the United State. Since the detailed information on the data appears in Chung et al. (2011), we briefly describe the data structure here. Complete data for the analysis were available on 1416 adolescents who were identified as the early onset drinkers. The early onset drinkers under this study had started drink by the time when they were 14 years old, at least 7 years before the legal drinking age. There

are three self-report items measuring adolescent drinking behaviors: how many days they had one or more drinks of an alcoholic beverage during the last 30 days (*Recent Drinking*); how many days they had five or more drinks on the same occasion during the last 30 days (*Binge Drinking*); and how many days they had drinks immediately before or during school or work hours in the last 30 days (*Drinking at School*). The responses for *Recent Drinking* were reduced to a three-category indicator, non-drinker (0 days of drinking), occasional drinker (1-5 days of drinking) and regular drinker (6 or more days of drinking). For *Binge Drinking*, respondents who had consumed five or more drinks on the same occasion at least one time were characterized as binge drinkers. The same rule was applied for *Drinking at School*. These three drinking items were tracked over the three survey waves in 1997 (Wave 1), 2000 (Wave 4), and 2003 (Wave 7), corresponding to early adolescence (ages 12–14), middle adolescence (ages 15–17), and late adolescence (ages 18–20), respectively. In addition to these three items, we consider gender and race as covariates in the model.

7.11 Information Criteria versus Dirichlet Process

Under the data set described above, Chung et al. (2011) started by fitting a series of two-class LCPA models where the pathways of the class membership were mapped onto between two and four profiles. This procedure was repeated until it reached a model with six classes and six profiles. Table 7.1 shows a series of LCPA models with evaluations based on the bootstrap p-value for goodness of fit and AIC.

The four-class-three-profile LCPA model and five-class-three-profile LCPA models are favored in terms of AIC, bootstrap p-value and the principle of parsimony. However, due to the unclear interpretation of the classes in the five-class-three-profile LCPA model, four-class-three-profile LCPA model is selected. Even though the traditional model selectors work

Table 7.1: Goodness-of-fit statistics for a series of LCPA models under various numbers of classes and profiles

Number of classes	Number of profiles	LRT	Bootstrapping p-value	AIC
2	2	812.81	0.000	12103
	3	812.81	0.000	12111
	4	812.81	0.000	12119
3	2	492.84	0.000	11803
	3	452.99	0.025	11777
	4	447.16	0.015	11785
	5	445.64	0.020	11798
4	2	449.80	0.005	11780
	3	404.58	0.305	11755
	4	392.61	0.440	11763
	5	387.90	0.315	11778
5	2	429.69	0.025	11780
	3	375.52	0.375	11752
	4	357.65	0.585	11760
	5	347.64	0.595	11776
6	2	421.65	0.015	11792
	3	367.31	0.410	11770
	4	338.81	0.735	11773
	5	325.88	0.810	11792
	6	306.07	0.795	11804

as convenient means for comparison of each competing model, the process would be a lengthy undertaking. Since dynamic Dirichlet learning process has been proved to work well with less computation demands, we are interested in implementing Dirichlet process to see if it comes to the same conclusions without comparing each model one at a time.

Fig. 7.1 (a) - (c) shows scatter diagrams derived by applying DDPM and LOWESS algorithm proposed by Cleveland (1979) with two smoothing curves. LOWESS is a locally weighted scatterplot smoothing used here to summarize the appropriate number of classes over time. The smoothing span gives the proportion of points in the plot that influence the smoothness at each value and larger spans give smoother lines. Each plot is smoothed with two smoother spans (.1 and .5) to locally fit the diagram. It is obvious to see all the wavy lines are around four, which makes the preferences fairly self-explanatory to be four classes over three time measurements. The result is in line with the histogram exploration as shown in Fig. 7.1 (d) where a four-class-over-three-time model has dominant frequency. Chung et al. (2011) selected a four-class LCPA model based on the Table 7.1 and it seems Dirichlet process has equal learning capability of selecting suitable number of classes.

Based on the class transiting routes, we implement DPM to learn the number of profiles. Fig. 7.1 (d) indicates models with three or four profiles are favored by the DPM algorithm because of their higher occurrences, 27% and 29%, respectively. Chung et al. (2011) conclude that three-profile model fit better based on goodness-of-fit index and according to the result DPM presents, we have consistent conclusions. Since DDPM and DPM lead to successful applications without requiring prior knowledge on the unknown system, we believe Dirichlet process can effectively constructs the model by integrating the sequential information.

7.12 Parameter Estimates

Given the picked four-class-three-profile LCPA model, Chung et al. (2011) used 100 different sets of initial parameters to avoid local optimum entrapment. They selected a set of estimates providing the highest loglikelihood value among those from 100 different initializations by using standard EM algorithm. We will focus on the same model structure to compare the performances of two deterministic annealing approaches in finding the global maximum estimates with the pre-determined annealing schedule. Note that we consider the LCPA model where the ρ-parameters are constrained to be equal across time points and both deterministic annealing update the marginalized conditional probability of the latent component membership recursively by using forward and backward functions.

To ensure the comparability of the standard EM algorithm and two deterministic annealing approaches, we implement DAEM with 100 initializations in order to better understand how different starting values induce variations in the loglikelihood values. For the standard EM algorithm, the distribution of the converged loglikelihood values corresponding to 100 sets of initializations are presented in Figure 7.2 (a). Exploring the histogram, we can see that EM solutions are trapped in four local modes based on their initializations. The largest loglikelihood value is -5726.28 presented in the last bar in Figure 7.2 (a), but the most frequent loglikelihood values are observed in the range of -5740 and -5739. On the contrary, Figure 7.2 (b) illustrates that DAEM reaches a single point convergence (loglikelihood value is -5727.14) irrespective of the initial values. Although DAEM does not achieve the optimum value given by the standard EM algorithm, the difference in the loglikelihood values is negligible based on the distributions presented in Figure 7.2. In addition, only 29% of the initial sets are converged to the global maximum with the standard EM algorithm.

Therefore, the DAEM algorithm with recursive formula can be considered as a robust tool to avoid many local entrapments and reduce the computation cost.

For DAVB, we also apply 100 different initializations and calculate the standard deviation for each of the model parameters from the 100 sets of DAVB estimates. Generally speaking, DAVB exhibits the work of satisfaction because the extent to which the variations are incurred due to the numerous initializations can be regarded insignificant. Here we briefly sketch the performance of DAVB in terms of their standard deviations and related concerns. The standard deviation of the ρ-parameter estimates derived from DAVB with 100 sets of initializations are presented in Table 7.2. The values under the *Recent Drinking* column provide the standard deviations of the estimates for probabilities of having reported occasional drinking (one to five days of drinking in the last 30 days) and regular drinking (six or more days of drinking in the last 30 days) for a given class membership. The other two columns show the standard deviations of the estimates for the probabilities of having consumed five or more drinks on the same occasion at least one time in the last 30 days for *Binge Drinking* and consumed alcoholic beverage right before or during school or work hours at least once in the last 30 days for *Drinking at School*. Table 7.2 indicates that Class 3 seems more likely to produce largest deviations among the identified classes. We believe the cause is more related to the small prevalence of Class 3: the prevalence of Class 3 is 7.5%, 7.0%, and 0.2% in 1997, 2000, and 2003, respectively, and the average prevalence over time is 4.9%.

The secondary measurement parameters (i.e., η-parameters) identify common sequential patterns of drinking behaviors. The standard deviation of the η-parameter estimates from the 100 sets of DAVB estimates are presented in Table 7.3. The η-parameter estimates from DAVB are shown to be fairly stable in terms of insignificant degree of spread. We can see

Table 7.2: Standard deviation of the ρ-parameter estimates derived from DAVB with 100 different sets of starting values

	Recent Drinking		*Binge*	*Drinking*
Class	Occasional	Regular	*Drinking*	*at School*
1	0.024	0.000	0.000	0.000
2	0.001	0.009	0.026	0.022
3	0.003	0.062	0.078	0.084
4	0.001	0.022	0.010	0.017

Table 7.3: Standard deviation of the η-parameter estimates derived from DAVB with 100 different sets of starting values

		Year		
Profile	Class	1997	2000	2003
1	1	0.025	0.031	0.027
	2	0.024	0.039	0.030
	3	0.018	0.030	0.020
	4	0.006	0.008	0.026
2	1	0.043	0.044	0.040
	2	0.040	0.056	0.068
	3	0.019	0.026	0.016
	4	0.012	0.035	0.061
3	1	0.030	0.030	0.024
	2	0.039	0.045	0.018
	3	0.040	0.051	0.010
	4	0.020	0.043	0.034

that the estimates corresponding to Profile 2 have more fluctuations comparing to those of other profiles. Although estimates regarding to some classes or profiles might not have ideally small deviations, DAVB can still be considered as a more consistent implementation than the traditional multi-start methods because all the standard deviations are below an acceptable threshold.

The standard deviations of the β-parameter estimates from the 100 sets of DAVB estimates are presented in Table 7.4. As inspected, the standard deviations of Profile 2 relative to Profile 1 are slightly larger than those of Profile 3, but none of them reveal serious insta-

Table 7.4: Standard deviation of β-parameter estimates derived from DAVB with 100 different sets of starting values (Profile 1 is the baseline)

Covariate	Category	Profile 2	Profile 3
Gender *versus* Male	Female	0.122	0.101
Race *versus* White	Black	0.242	0.163
	Hispanic	0.260	0.204
	Other race	0.272	0.283

Table 7.5: Estimated probabilities of responding 'any use' to the drinking items for each class (ρ-parameters)

		Recent Drinking		*Binge*	*Drinking*
Method	Class	Occasional	Regular	*Drinking*	*at School*
EM	1	0.001	0.002	0.000	0.000
	2	0.960	0.040	0.303	0.119
	3	0.675	0.325	0.848	0.606
	4	0.280	0.720	0.960	0.173
DAEM	1	0.102	0.000	0.000	0.000
	2	0.929	0.071	0.378	0.056
	3	0.899	0.101	0.546	0.396
	4	0.260	0.740	0.965	0.205
DAVB	1	0.096	0.000	0.000	0.000
	2	0.922	0.076	0.407	0.044
	3	0.894	0.105	0.541	0.469
	4	0.238	0.761	0.967	0.207

bility among the resulting estimates from 100 sets of starting values with DAVB. Profile 2 tends to produce larger standard deviations as previously shown in Table 7.3 and we believe the small prevalence of Profile 2 affects the results in some degrees.

The estimated primary measurement parameters (i.e., ρ-parameters) using the three estimation algorithms are presented in Table 7.5. The values under the *Recent Drinking* column provide the estimated probabilities of having reported occasional drinking (one to five days of drinking in the last 30 days) and regular drinking (six or more days of drinking in the last 30 days) for a given class membership. The other two columns show the probabilities

of having consumed five or more drinks on the same occasion at least one time in the last 30 days for *Binge Drinking* and consumed alcoholic beverage right before or during school or work hours at least once in the last 30 days for *Drinking at School*. We can see that all three items combined support a meaningful interpretation for each class. An inspection of these values leads to the adoption of the common class names across estimation methods. Adolescents in Class 1 have not been involved in any drinking in the previous 30 days; therefore we label Class 1 as 'non-current drinkers.' Class 2 represents adolescents who drink occasionally but have no history of regular drinking or drinking at work or school. We accordingly identify adolescents classified in Class 2 as 'light drinkers.' For the same grounds, Class 4 labels as 'regular binge drinkers' who both drink regularly and engage in binge drinking. The estimates for Class 3, however, produce largest deviations between EM and two deterministic annealing methods, leading to a difficulty in labeling Class 3 with the common class name across estimation algorithms. The maximal differences in the estimates between EM and two annealing methods are .302 and .307 in *Binge Drinking* for DAEM and DAVB, respectively. We believe the cause is more related to the small prevalence of Class 3: using the EM algorithm, the prevalence of Class 3 is 7.5%, 7.0%, and 0.2% in 1997, 2000, and 2003, respectively, and the average prevalence over time is 4.9%. The other two deterministic alternatives produce similar class prevalences as the EM algorithm. Although three parameter estimation methods produce distinct estimates for Class 3, the difference in estimates of 'small' class may not affect the description of the major classes and the loglikelihood value.

The secondary measurement parameters (i.e., η-parameters) identify common sequential patterns of drinking behaviors. The estimated η-parameters with three estimation methods are presented in Table 7.6. As shown in Table 7.6, in Profile 1 the probabilities of belonging

Table 7.6: Estimated probabilities of belonging to a class sequence for each profile (η-parameters) and estimated profile prevalence (γ-parameters)

			Year		
Method	Profile	Class	1997	2000	2003
EM	1 (45.0)	1	0.694	0.746	0.626
		2	0.250	0.219	0.348
		3	0.055	0.034	0.004
		4	0.001	0.000	0.023
	2 (18.6)	1	0.538	0.058	0.100
		2	0.432	0.766	0.610
		3	0.000	0.000	0.000
		4	0.031	0.175	0.290
	3 (36.4)	1	0.614	0.252	0.133
		2	0.227	0.122	0.015
		3	0.139	0.150	0.000
		4	0.020	0.476	0.853
DAEM	1 (46.1)	1	0.782	0.806	0.688
		2	0.046	0.027	0.203
		3	0.163	0.167	0.078
		4	0.009	0.000	0.031
	2 (19.0)	1	0.610	0.112	0.121
		2	0.239	0.766	0.666
		3	0.130	0.000	0.001
		4	0.021	0.122	0.212
	3 (34.9)	1	0.672	0.253	0.135
		2	0.019	0.034	0.002
		3	0.255	0.158	0.001
		4	0.054	0.555	0.862
DAVB	1 (47.0)	1	0.785	0.780	0.688
		2	0.013	0.052	0.207
		3	0.189	0.165	0.065
		4	0.013	0.003	0.040
	2 (18.4)	1	0.695	0.159	0.094
		2	0.102	0.711	0.749
		3	0.186	0.030	0.014
		4	0.017	0.100	0.143
	3 (34.6)	1	0.665	0.284	0.139
		2	0.033	0.050	0.015
		3	0.258	0.136	0.001
		4	0.044	0.530	0.845

to Class 1, the class of 'not current drinkers,' are consistently higher than the probabilities of belonging to other classes (see the first rows in Table 7.6 for all estimation method), implying that adolescents in this profile are likely to remain as 'not current drinkers' over time. The LCPA identified another two profiles of adolescents who tended to intensify their drinking habits over time. Adolescents in Profile 2 were more likely to belong to Class 1 (not current drinkers) in 1997, advance to Class 2 (light drinkers) in 2000. By 2003, some of them had advanced further to become members of Class 4 (regular binge drinkers), although many others remained in Class 2 (light drinkers). Profile 3 consists of early drinkers who moved toward Class 4 (regular binge drinkers) by the year of 2000 and stayed in that class by 2003. Three parameter estimation methods yield similar estimates for the η-parameters so that we can commonly label the identified profiles across three different estimation methods. The most prevalent profile is Profile 1 for all estimation methods. The largest differences between EM and two annealing methods are .208 for DAEM in Profile 3 and Class 2 in 1997 and .330 for DAVB in Profile 2, Class 2 in 1997.

Point estimates for the logistic regression coefficients (β-parameter) pertaining to each covariate are reported in Table 7.7. Speaking broadly, whites are more likely to be in the profiles involving intensified drinking habits (i.e., Profiles 2 and 3) than in Profile 1 (non-drinking stayers). Interestingly, the odds of belonging to Profile 2 (light drinking advancers) versus Profile 1 are higher for females, but the odds of belonging to Profile 3 (regular binge drinking advancers) are higher for males than their counterparts. Three parameter estimation methods come to the same conclusions in explaining the logistic coefficients even though the yielded values are not indistinguishable.

Table 7.7: Estimated logistic regression coefficients (β-parameters) for the prevalence of profiles (Profile 1, non-drinking stayers, is the baseline)

Method	Covariate	Profile 2	Profile 3
EM	Intercept	−1.144	0.511
	Female *versus* male	1.244	−1.011
	Race *versus* White		
	Black	−1.640	−1.510
	Hispanic	−1.600	−0.198
	Other race	−1.133	−0.481
DAEM	Intercept	−1.161	0.410
	Female *versus* male	1.349	−0.882
	Race *versus* White		
	Black	−1.884	−1.546
	Hispanic	−1.897	−0.240
	Other race	−0.877	−0.459
DAVB	Intercept	−1.146	0.415
	Female *versus* male	1.197	−0.823
	Race *versus* White		
	Black	−1.457	−1.523
	Hispanic	−1.217	−0.261
	Other race	−0.473	−0.301

7.12.1 Discussion

In the standard EM algorithm, the computation of the joint posterior distribution that involves all time measurements is fairly demanding. Instead of computing the joint posterior probabilities given in (5.5), the recursive formula adopts the forward-backward algorithm to calculate the marginal posterior probabilities given in (5.6) directly. The recursive formula enables us to predict latent membership faster than the standard EM algorithm because the forward and backward quantities can easily calculated.

In this study, we apply the two annealing methods, DAEM and DAVB, to estimate the model parameters lying on the global maximum of the objective functions for the LCPA models using the recursive formula at each iteration. Deterministic annealing is an optimization technique aiming to search for a global maximum by gradually increasing the value of ω. However, the performance of the annealing methods is contingent upon the choice of annealing schedule. Geman and Geman (1984) have shown that, if the annealing schedule follows $\omega \propto (\log i)$, where i is the number of current iteration, the global solution is theoretically achievable even though such schedule might not be realistic in practice. Chang and Lin (2002) proposed a novel method to adaptively learn the annealing parameter based on the results of the previous iteration. However, it requires matrix calculation which is limited only for special cases. In this study, we adopt a very tight annealing schedule, but more research is required to select the appropriate annealing schedule for the LCPA models. However, the DAEM and DAVB have their own strengths and therefore are attractive alternatives to the standard EM algorithm in order to discover an optimal solution.

Roughly speaking, when ω is close to zero, for both DAEM and DAVB, the corresponding $\underline{\theta}_{i(s,c_1,\ldots,c_T)}$ generates random moves with uniform weights and pick the membership of

class and profile arbitrarily. As the annealing parameters are gradually adjusted, the localized influences become more discriminating for classes and profiles and enable deterministic annealing approaches to avoid arbitrary local maxima and search in the right direction. The DAEM algorithm is more convenient to operate and typically converges more quickly than DAVB. However, when the annealing rate is too sparse or the prevalence of latent components are not large enough, DAEM may be entrapped into a local mode of the likelihood function. On the contrary, the DAVB algorithm is less influenced by the annealing schedule or the size of latent components, but it may exhibit slow convergence in the search of global maximum. Therefore, more computational burden should be expected as the trade-off. The most troubling aspect of the DAVB is to decide a well-functioning proposal distribution in the MCMC. In the beginning, we applied a Metropolis algorithm in which a candidate for the next logistic coefficient vector was sampled from a multivariate t distribution with degree of freedom ν, $\boldsymbol{\beta}^{\mathcal{C}} \sim t_{\nu}(\hat{\boldsymbol{\beta}}, \delta\boldsymbol{\Sigma})$, where δ was a scalar value used to control the acceptance rate. Using the t distribution, however, the acceptance rate drops when ω approaches one, which causes the slow mixing in the Markov chain. Although decreasing the variation in the proposal distribution by setting a small value to δ may alleviate the problem, it does not make sense to sample the candidate values only from the regions close to the mean $\hat{\boldsymbol{\beta}}$. Therefore, we substitute t distribution with a multivariate normal to expedite the mixing by keeping many undesired extreme values from being selected.

Although many alternatives to the standard estimation algorithm could be pursued in light of the LCPA model, we only mentioned only a few in this presentation. Note that our exploration was intended to demonstrate a possible solution to difficulties in finding a set of estimates on the global maximum of the objective function. We provided a limited demonstration using real data to elucidate the estimation methods. Our hope is that substantive

researchers will be able to identify possible difficulties in estimation for the LCPA model, and consider using the proposed solution in their research.

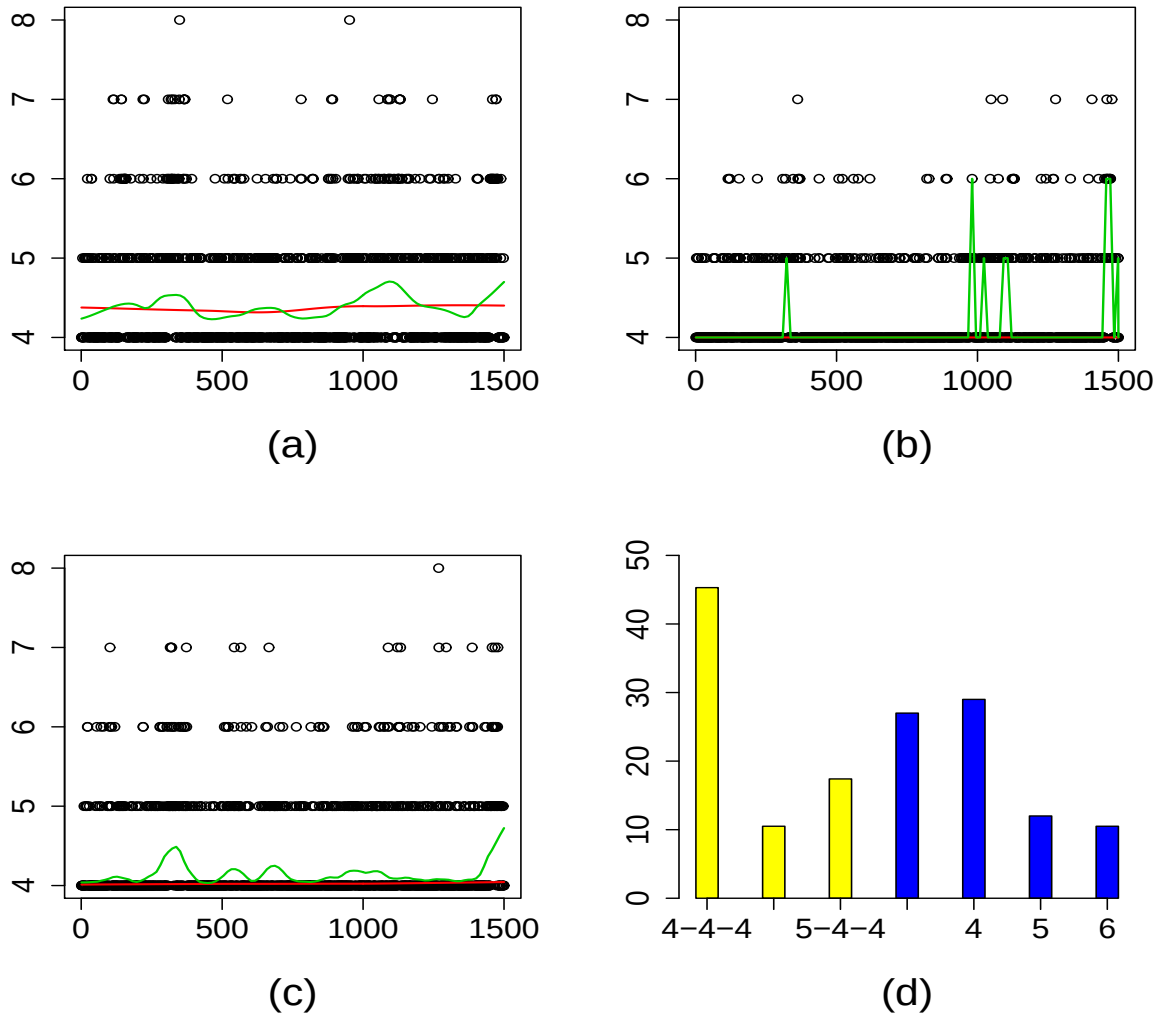

Figure 7.1: (a) - (c) Tracking plots for classes smoothed by LOWESS with two smoother spans (red: .67 and green: .15) from time 1 to time 3; (d) histogram of class progression (yellow bars) and the number of profiles (blue bars).

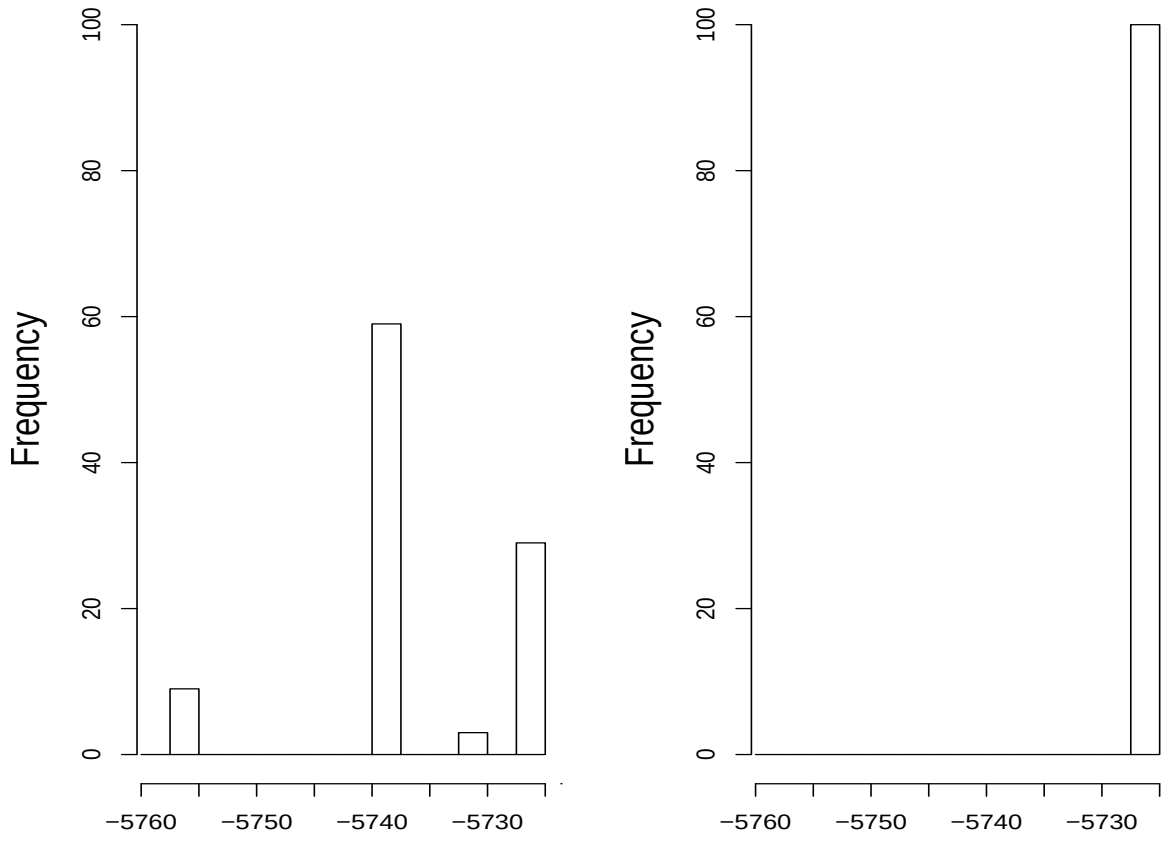

Figure 7.2: Histogram of loglikelihood values derived from (a) the standard EM and (b) the DAEM algorithms with 100 different sets of starting values

Chapter 8

Discussion

8.1 Contributions

Latent stage-sequential process is an attractive tool for many areas of substantive research. We have designed several computational algorithms in order to develop practical understanding and concrete advice for those who may apply the LCPA model to their own data. The unique contributions of this thesis are summarized as follows.

Determining the number of latent components such as classes and profiles can be difficult especially when prior knowledge is not readily available. We adopt two Bayesian approaches to learn the model without presuming the number of latent components in advance. The RJMCMC turns to be less computationally feasible because there is no way to design a best-suited jumping rules. When the proposal construction has trouble landing in regions of low probability density, a rapid mixing to reach the stationary distribution will not be facilitated. However, to design an appropriate jumping principle for dimension changes is not easy. On the contrary, Dirichlet process is easily carried out to predict the underlying sequential structure due to its non-parametric nature. By relaxing the constrains on the

number of classes placed at each measurement occasion, Dirichlet process helps to enhance visibility of future patterns. However, in the profile-learning phase, Dirichlet process has difficulty delivering satisfactory results because weak measurements make competing models indistinguishable.

To estimate the unknown parameters, the widely-known technique is hill climbing optimization such as EM algorithm. This gradient optimization approach can search local neighborhood of the initial values but it may be very poor compared to the global optimal solution. In this study, we firstly compute the forward probability of being in a certain state (i.e. class) at time point t and the backward probability of having a specific type of future observations after t given the current states. Both of them are advantageously represented in recursive forms to expedite each iteration. The forward-backward algorithm proceeds on the basis of parameter updates but the initialization dependence problem remains unsolved.

There is a plethora of literatures aiming at finding a numerically robust parameter estimation tool to deal with local maximum problems (Biernacki et al., 2003; Reddy and Rajaratnam, 2010). For example, Reddy and Rajaratnam (2010) convolute the objective function by kernel functions to flatten the surface and reduce the number of local modes. They demonstrate that the optimal solutions would be reached effectively by adjusting the smoothing factor at each iteration. Instead, we implement deterministic annealing EM and deterministic annealing variant of variational Bayes in order to find parameter estimates on the global mode of the objective function. Two deterministic annealing approaches are presented here to overcome the local maximum issues. It is shown that both DAEM and DAVB are equivalent to a generic gradient method which starts from the unimodal well-shaped function of unknown parameter and by gradually increasing the annealing parameter, the distribution becomes more discriminating and a good approximation to the global optimum

can be located. While we have demonstrated the effectiveness of the deterministic annealing algorithms in combating the local modality problems, a global optimum is not always guaranteed. The most important reason is that the annealing schedule is not unanimously regulated; one must be very discreet when adjusting the increases. Generally, the schedule should be determined so that the annealing rate is either passive nor aggressive.

8.2 Direction for Future Research

In the model selection problem, we adopt reversible jump MCMC because it is the most commonly used MCMC tool by which we can explore variable dimension statistical models. We already discussed about the difficulties in proposing efficient jumping rules especially in the complex LCPA models. Fahimah (2004) suggested using a secondary Markov chain (adding few fixed-dimension MCMC steps) to modify proposed moves before calculating the acceptance rates. They have shown the acceptance probabilities soar even the proposals are poorly matched to the true target distribution but the increased programming costs are expected as a trade-off.

We present dynamic Dirichlet learning process analysis on model selection problems without any covariates to predict the prevalence of the profile; however, predicator-dependent kernel stick-breaking process has already increased the interest. It is utilized in choosing the priors for an unknown probability measure (Dunson and Park, 2008) and variable selection problems (Chung and Dunson, 2009). Adding predictors in the prior consideration gives different insights into how the profiles are formed under the influence of predictors and it is understood as a dependent Dirichlet Process. Future work should explore the model selection for the LCPA regression model.

APPENDICES

Appendix A

Acceptance Rate of $C \to C+1$

For the split move on latent classes (i.e., $C' = C+1$), the acceptance rate $\alpha(C, C')$ is $min\,(1, A)$, where

$$
\begin{aligned}
A &= \frac{P(\boldsymbol{\Theta}_{(C',S)} \mid \mathbf{y})}{P(\boldsymbol{\Theta}_{(C,S)} \mid \mathbf{y})} \times \frac{q(\boldsymbol{\Theta}_{(C',S)}, \boldsymbol{\Theta}_{(C,S)})}{q(\boldsymbol{\Theta}_{(C,S)}, \boldsymbol{\Theta}_{(C',S)})} \times \frac{g^*(u^*)}{g(u)} \times \left| \frac{\partial(\boldsymbol{\Theta}_{(S,C')}, u^*)}{\partial(\boldsymbol{\Theta}_{(S,C)}, u)} \right| \\
&= \text{(likelihood ratio)} \times \frac{\pi_0(\boldsymbol{\Theta}_{(C',S)})}{\pi_0(\boldsymbol{\Theta}_{(C,S)})} \times \frac{d_{C+1}}{b_C} \times \frac{1}{P_{alloc}} \\
&\times \frac{1}{\prod_{s=1}^{S} \prod_{t=1}^{T} g(u_s^{(t)}) \times \prod_{m=1}^{M} g(u_m)} \\
&\times \frac{\prod_{s=1}^{S} \prod_{t=1}^{T} \eta_{c^*|s}^{(t)} \left[(1 + \frac{1-\bar{u}}{\bar{u}})\kappa \right]^M}{[(1-\bar{u})/\bar{u}]^{M/2}},
\end{aligned}
$$

where the likelihood ratio is the ratio of the product of the complete likelihood values for the new parameter sets when new classes are formed to that for the old. The function π_0 is the product of the prior distributions for model parameters, the number of latent classes, and the latent allocation variables. To be specific, writing $D_{\mathbf{V}}(\boldsymbol{\delta})$ to denote a Dirichlet density

evaluated at $\mathbf{v}$ and parametrized by a vector $\boldsymbol{\delta}$,

$$\frac{\pi_0(\boldsymbol{\Theta}_{(C',S)})}{\pi_0(\boldsymbol{\Theta}_{(C,S)})} \propto \prod_{s,t} \left(\frac{\eta_{c_1^*|s}^{(t)\ \delta-1} \eta_{c_2^*|s}^{(t)\ \delta-1}}{\eta_{c^*|s}^{(t)\ \delta-1} B(\delta, C\delta)} \right) \prod_{m=1}^{M} \frac{D_{\boldsymbol{\rho}_{m|c_1^*}}(\boldsymbol{\delta}) D_{\boldsymbol{\rho}_{m|c_2^*}}(\boldsymbol{\delta})}{D_{\boldsymbol{\rho}_{m|c^*}}(\boldsymbol{\delta})}$$

Besides, P_{alloc} is the probability that the specific reallocation for those who were in the class that was chosen to be split is made. The g function is the Uniform$(0, 1)$ density. For the corresponding merge move, the acceptance probability $\alpha(C, C-1)$ is $min(1, A^{-1})$, using the same expression for A but some corrections are required.

Appendix B

Acceptance Rate of $S \rightarrow S+1$

For the split move on latent profiles (i.e., $S' = S+1$), the acceptance rate $\alpha(S, S')$ is $min\,(1, B)$, where

$$
\begin{aligned}
B &= \frac{P(\boldsymbol{\Theta}_{(C,S')} \mid \mathbf{y})}{P(\boldsymbol{\Theta}_{(C,S)} \mid \mathbf{y})} \times \frac{q(\boldsymbol{\Theta}_{(C,S')}, \boldsymbol{\Theta}_{(C,S)})}{q(\boldsymbol{\Theta}_{(C,S)}, \boldsymbol{\Theta}_{(C,S')})} \times \frac{g^*(u^*)}{g(u)} \times \left| \frac{\partial(\boldsymbol{\Theta}_{(S',C)}, u^*)}{\partial(\boldsymbol{\Theta}_{(S,C)}, u)} \right| \\
&= \text{(likelihood ratio)} \times \frac{\pi_0(\boldsymbol{\Theta}_{(C,S')})}{\pi_0(\boldsymbol{\Theta}_{(C,S)})} \times \frac{d_{S+1}}{b_S} \times \frac{1}{P_{alloc}} \\
&\times \frac{1}{g_1(w) \prod_{c=1}^{C-1} \prod_{t=1}^{T} g_2(\beta_{ct})} \\
&\times \left(\frac{\prod_{c=1}^{C} \prod_{t=1}^{T} \eta_{c|s_1}^{(t)} \times \prod_{c=1}^{C} \prod_{t=1}^{T} \eta_{c|s_2}^{(t)}}{\prod_{c=1}^{C} \prod_{t=1}^{T} \eta_{c|s^*}^{(t)}} \times \frac{1}{w(1-w)} \right)^{T(C-1)},
\end{aligned}
$$

where the likelihood ratio is the ratio of the product of the complete likelihood values for the new parameter sets when new profiles are formed to that for the old. The function π_0 is the product of the prior distributions for model parameters, the number of latent profiles,

and the latent allocation variables. To be specific,

$$\frac{\pi_0(\boldsymbol{\Theta}_{(C,S')})}{\pi_0(\boldsymbol{\Theta}_{(C,S)})} \propto \prod_{t=1}^{T} \left(\frac{D_{\boldsymbol{\eta}_{s_1}^{(t)}}(\boldsymbol{\delta}) D_{\boldsymbol{\eta}_{s_2}^{(t)}}(\boldsymbol{\delta})}{D_{\boldsymbol{\eta}_{s*}^{(t)}}(\boldsymbol{\delta})} \right) \times \frac{(\gamma_{s*} w_1)^{\delta-1} (\gamma_{s*} w_2)^{\delta-1}}{\gamma_{s*}^{\delta-1} B(S\delta, \delta)}$$

Besides, $Palloc$ is the probability that the specific reallocation for those who were in the profile that was chosen to be split is made. The g_1 function is the Uniform$(0,1)$ density and g_2 is the $N(0,1)$ density. For the merge move, the acceptance probability $\alpha(S, S-1)$ is $min(1, B^{-1})$, using the same expression for B but some corrections are required.

Appendix C

Hessian Matrix

C.1 Diagonal Entries

Firstly, we introduce several quantities for the future use

1. $\theta_{i(s,c)} = Pr(L_i = s, \mathbf{C}_i = \mathbf{c}_i | \mathbf{Y}_i = \mathbf{y}_i)$
2. $\theta_{is} = \sum_{c_1} \cdots \sum_{c_T} \theta_{i(s,c)}$
3. $\theta^{(t)}_{i(s,c_t)} = \prod_{j \neq t} \sum_{c_j} \theta_{i(s,c)}$
4. $\theta^{(t)}_{ic_t} = \sum_s \theta^{(t)}_{i(s,c_t)}$
5. $\theta^{(t,t^*)}_{i(s,c,c^*)} = P(L_i = s, C_{it} = c, C_{it^*} = c^* \mid \mathbf{Y}_i = \mathbf{y}_i)$
6. $\theta^{(t,t^*)}_{i(c,c^*)} = \sum_s \theta^{(t,t^*)}_{i(s,c,c^*)}$
7. $g(i,m,c,t) = \left(\rho_{m|c_t=c}\right)^{I(y_{imt}=1)} \left(1 - \rho_{m|c_t=c}\right)^{I(y_{imt}=0)}$
8. $\delta(i,c,t) = 1 - \prod_{m=1}^{rm} \frac{g(i,m,t,C)}{g(i,m,t,c)}$

Let $f_i = P(Y_{i1} = y_{i1}, \ldots, Y_{iT} = y_{iT})$ and $l_i = \log f_i$. The s_{ref} and c_{ref} stand for reference group for profile and class respectively. The elements in the diagonal matrix of the Hessian for the profile probabilities $\boldsymbol{\gamma}$ are

$$\frac{\partial^2 l}{\partial \gamma_s \partial \gamma_{s*}} = -\left(\frac{\theta_{is}}{\gamma_s} - \frac{\theta_{is_{ref}}}{\gamma_{s_{ref}}}\right)\left(\frac{\theta_{is*}}{\gamma_{s*}} - \frac{\theta_{is_{ref}}}{\gamma_{s_{ref}}}\right)$$

where s and $s^* = 1, 2, \ldots, S-1$.

To get the diagonal block of the Hessian with respect to $\boldsymbol{\eta}$, we consider the following cases.

1. When $s = s$ and $t = t$, the following equation is true whether c_1 is equal to c_2

$$\frac{\partial^2 l_i}{\partial \eta^{(t)}_{c_1|s} \partial \eta^{(t)}_{c_2|s}} = \left(\frac{\theta^{(t)}_{i(s,c_1)}}{\eta^{(t)}_{c_1|s}} - \frac{\theta^{(t)}_{i(s,c_{ref})}}{\eta^{(t)}_{c_{ref}|s}}\right)\left(\frac{\theta^{(t)}_{i(s,c_2)}}{\eta^{(t)}_{c_2|s}} - \frac{\theta^{(t)}_{i(s,c_{ref})}}{\eta^{(t)}_{c_{ref}|s}}\right)$$

2. When $s = s$ but $t_1 \neq t_2$, the following equation is true whether c_1 is equal to c_2

$$\begin{aligned}\frac{\partial^2 l_i}{\partial \eta^{(t_1)}_{c_1|s} \partial \eta^{(t_2)}_{c_2|s}} &= \frac{\theta^{(t_1,t_2)}_{i(s,c_1,c_2)}}{\eta^{(t_1)}_{c_1|s}\eta^{(t_2)}_{c_2|s}}\delta(i,t_1,c_1)\delta(i,t_2,c_2) \\ &- \left(\frac{\theta^{(t_1)}_{i(s,c_1)}}{\eta^{(t_1)}_{c_1|s}} - \frac{\theta^{(t_1)}_{i(s,c_{ref})}}{\eta^{(t_1)}_{c_{ref}|s}}\right)\left(\frac{\theta^{(t_2)}_{i(s,c_2)}}{\eta^{(t_2)}_{c_2|s}} - \frac{\theta^{(t_2)}_{i(s,c_{ref})}}{\eta^{(t_2)}_{c_{ref}|s}}\right)\end{aligned}$$

3. When $t = t$ but $s_1 \neq s_2$, the following equation is true whether c_1 is equal to c_2

$$\frac{\partial^2 l_i}{\partial \eta^{(t)}_{c_1|s_1} \partial \eta^{(t)}_{c_2|s_2}} = -\left(\frac{\theta^{(t)}_{i(s_1,c_1)}}{\eta^{(t)}_{c_1|s_1}} - \frac{\theta^{(t)}_{i(s_1,c_{ref})}}{\eta^{(t)}_{c_{ref}|s_1}} \right) \times \left(\frac{\theta^{(t)}_{i(s_2,c_2)}}{\eta^{(t)}_{c_2|s_2}} - \frac{\theta^{(t)}_{i(s_2,c_{ref})}}{\eta^{(t)}_{c_{ref}|s_2}} \right)$$

4. When $t_1 \neq t_2$ and $s_1 \neq s_2$, the following equation is true whether c_1 is equal to c_2

$$\frac{\partial^2 l_i}{\partial \eta^{(t_1)}_{c_1|s_1} \partial \eta^{(t_2)}_{c_2|s_2}} = -\left(\frac{\theta^{(t_1)}_{i(s_1,c_1)}}{\eta^{(t_1)}_{c_1|s_1}} - \frac{\theta^{t_1}_{i(s_1,c_{ref})}}{\eta^{(t_1)}_{c_{ref}|s_1}} \right) \times \left(\frac{\theta^{t_2}_{i(s_2,c_2)}}{\eta^{(t_2)}_{c_2|s_2}} - \frac{\theta^{t_2}_{i(s_2,c_{ref})}}{\eta^{(t_2)}_{c_{ref}|s_2}} \right)$$

Next, for the diagonal block of the Hessian with respect to the response measurement parameter $\boldsymbol{\rho}$, there are several cases to be considered.

1. When $t_1 \neq t_2$, the following expression is true irrespective of the relationship between

k and l or c_1 and c_2.

$$\begin{aligned}\frac{\partial^2 l_i}{\partial \rho_{k|c_{t_1}=c_1} \partial \rho_{l|c_{t_2}=c_2}} &= \frac{(2y_{ikt_1}-1)(2y_{ilt_2}-1)}{g(i,k,t_1,c_1)g(i,l,c_2,t_2)} \\ &\times \left(\theta^{(t_1,t_2)}_{i(c_1,c_2)} - \theta^{(t_1)}_{i,c_1}\theta^{(t_2)}_{i,c_2}\right)\end{aligned}$$

2. When $t = t$, $c = c$ but $k \neq l$,

$$\frac{\partial^2 l_i}{\partial \rho_{k|c_t=c} \partial \rho_{k|c_t=c}} = -\frac{(2y_{ikt}-1)(2y_{ilt}-1)}{g(i,k,c,t)g(i,l,c,t)}\theta^{(t)}_{ic}(1-\theta^{(t)}_{ic})$$

3. When $t = t$ but $c_1 \neq c_2$, the following expression is true whether k is equal to l,

$$\frac{\partial^2 l_i}{\partial \rho_{k|c_t=c_1} \partial \rho_{k|c_t=c_2}} = -\frac{\theta^{(t)}_{ic_1}(2y_{ikt}-1)}{g(i,k,c_1,t)}\frac{\theta^{(t)}_{ic_2}(2y_{ilt}-1)}{g(i,l,c_2,t)}$$

4. When all subscripts are the same,

$$\frac{\partial^2 l_i}{\partial \rho_{k|c_t=c} \partial \rho_{k|c_t=c}} = -\frac{{\theta^{(t)}_{i,c}}^2(2y_{ikt}-1)^2}{g(i,k,t,c)^2}$$

C.2 Off-Diagonal Entries

To derive the off-diagonal elements, we start with the Hessian with respect to $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$. Similarly, we need to consider couple of cases.

1. When $s = s$,

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\gamma_s \partial\eta^t_{c|s}} &= \left(\frac{\theta^{(t)}_{i(s,c)}}{\eta^{(t)}_{c|s}\gamma_s} - \frac{\theta^{(t)}_{i(s,c_{ref})}}{\eta^{(t)}_{c_{ref}|s}\gamma_s}\right) \\ &- \left(\frac{\theta_{is}}{\gamma_s} - \frac{\theta_{is_{ref}}}{\gamma_{s_{ref}}}\right)\left(\frac{\theta^{(t)}_{i(s,c)}}{\eta^{(t)}_{c|s}} - \frac{\theta^{(t)}_{i(s,cref)}}{\eta^{(t)}_{C|s}}\right)\end{aligned}$$

2. When $s_1 \neq s_2$,

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\gamma_{s_1} \partial\eta^{(t)}_{c|s_2}} &= -\frac{\theta^{(t)}_{i(s_{ref},c)}}{\gamma_{s_{ref}}\eta^{(t)}_{c|s_{ref}}}\delta(i,c,t) \\ &- \left(\frac{\theta_{is_1}}{\gamma_{s_1}} - \frac{\theta_{is_{ref}}}{\gamma_{s_{ref}}}\right)\left(\frac{\theta^{(t)}_{i(s_2,c)}}{\eta^{(t)}_{c|s_2}} - \frac{\theta^{(t)}_{i(s_2,c_{ref})}}{\eta^{(t)}_{c_{ref}|s_2}}\right)\end{aligned}$$

We then derive the off-diagonal elements of the Hessian with respective to $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$.

$$\begin{aligned}\frac{\partial^2 l_i}{\partial \gamma s_1 \partial \rho_{k|c_t=c}} &= -\frac{(2y_{ikt}-1)}{g(i,k,c,t)} \\ &\times \left(\frac{\theta^{(t)}_{i(s,c)}}{\gamma s} - \frac{\theta^{(t)}_{i(\gamma s_{ref},c)}}{\gamma s_{ref}} - \left(\frac{\theta_{is}}{\gamma s} - \frac{\theta_{i,s_{ref}}}{\gamma s_{ref}} \right) \theta^{(t)}_{ic} \right)\end{aligned}$$

At last, we derive the off-diagonal elements of the Hessian with respective to $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$. There are several cases we need to take into consideration.

1. When $t_1 \neq t_2$, the following expression is applicable whether or not c_1 is equal to c_2,

$$\frac{\partial^2 l_i}{\partial \eta^{(t_1)}_{c_1|s} \partial \rho_{k|c_{t_2}=c_2}} = \frac{2y_{ikt_2} - 1}{g(i,k,c_2,t_2)} R, \tag{C.1}$$

where $R = \left(\frac{\theta^{(t_2,t_1)}_{i(s,c_2,c_1)}}{\eta^{(t_1)}_{c_1|s}} \delta(i,t_1,c_1) - \left(\frac{\theta^{(t_1)}_{i(s,c_1)}}{\eta^{(t_1)}_{c_1|s}} - \frac{\theta^{(t_1)}_{i(s,c_{ref})}}{\eta^{(t_1)}_{c_{ref}|s}} \right) \theta^{(t_2)}_{i,c_2} \right)$

2. When $t = t$ but $c_1 \neq c_2$,

$$\frac{\partial^2 l_i}{\partial \eta^{(t)}_{c_1|s} \partial \rho_{k|c_t=c_2}} = \left(\frac{\theta^{(t)}_{i(s,c_1)}}{\eta^{(t)}_{c_1|s}} - \frac{\theta^{(t)}_{i(s,c_{ref})}}{\eta^{(t_1)}_{c_{ref}|s}} \right) \frac{\theta^{(t)}_{i,c_2}(2y_{ikt-1})}{g(i,k,t,c_2)}$$

3. When all subscripts are the same

$$\frac{\partial^2 l_i}{\partial \eta^{(t)}_{c|s} \partial \rho_{k|c_t=c}} = \frac{2y_{ikt_2} - 1}{g(i,k,c,t)} \left(\frac{\theta^{(t)}_{i(s,c)}}{\eta^{(t)}_{c|s}} - \left(\frac{\theta^{(t)}_{i(s,c)}}{\eta^{(t)}_{c|s}} - \frac{\theta^{(t)}_{i(s,c_{ref})}}{\eta^{(t)}_{c_{ref}|s}} \right) \theta^{(t)}_{i,c} \right)$$

BIBLIOGRAPHY

BIBLIOGRAPHY

A. Agresti. *Categorical Data Analysis*. Wiley, Hoboken, New Jersey, second edition, 2002.

A. Ahmed and E. Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process : with applications to evolutionary clustering. 2008.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 6:716–723, 1974.

C. E. Antoniak. Mixtures of dirichlet processes with applications to non-parametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

K. Bandeen-Roche, D. L. Miglioretti, S. L. Zeger, and P. J. Rathouz. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92: 1375–1386, 1997.

D. J. Bartholomew and M. Knott. *Latent variable models and factor analysis*. Arnold, London, second edition, 1999, Chap. 6.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41:561–575, 2003.

D. Blackwell and J. B. Macqueen. Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1:353–355, 1973.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Wiley & Sons, New York, 1992.

M.-W. Chang and C.-J. Lin. Adaptive determnistic annealing for two applications: competing svr of switching dynamics and travelling salesman problems. *The 9th International Conference on Neural Information Processing, Singapore*, 2002.

S. Chib. Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, 75:79–97, 1996.

H. Chung, J. C. Anthony, and J. L. Schafer. Latent class profile analysis: an application to stage-sequential process of under-age drinking behaviours. *Journal of the Royal Statistical Society, Series A*, 2011.

H. Chung, S. T. Lanza, and E. Loken. Latent transition analysis: inference and estimation. *Statistics in Medicine*, 27:1834–1854, 2008.

Y. Chung and D. B. Dunson. Nonparametric bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, 104:1646–1660, 2009.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.

C. C. Clogg. Latent class models. In G. Arminger, C. C. Clogg, and M. E. Sobel, editors, *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, pages 351–359. Plenum, New York, 1995.

C. C. Clogg and L. A. Goodman. Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, 79:762–771, 1984.

L. M. Collins and S. E. Wugalter. Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, 27:131–157, 1992.

T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6*, 4:251–299, 1931.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39: 1–38, 1977.

L. C. Dierker, F. Vesela, E. M. Sledjeskia, D. Costelloa, and N. Perrine. Testing the dual pathway hypothesis to substance use in adolescence and young adulthood. *Drug and Alcohol Dependence*, 87:83–93, 2007.

D. B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

T. S. Ferguson and M. J. klass. A representation of independentincrement processes without gaussian components. *The Annals of Mathematical Statistics*, 43:1634–1643, 1972.

A. K. Formann. Latent class model diagnosis from a frequentist point of view. *Biometrics*, 59:189–196, 2003.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.

S. Geman and D. Geman. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:401–412, 1984.

L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231, 1974.

P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

S. J. Haberman. *Analysis of Qualitative Data*. Academic Press, New Developments, New York, 1979.

J. A. Hagenaars and A. L. McCutcheon. *Applied latent class analysis*. Cambridge University Press, Cambridge, 2002.

W. Hastings. Monte carlo sampling methods using markov chains and their application. *Biometrika*, 57:97–109, 1970.

T. Heinen. *Discrete latent variable models*. Tilburg University Press, The Netherlands, 1993, Chap. 2.

E. Hewitt and L. J. Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80:470–501, 1955.

K. W. Ho and I. Hu. Flexible modelling of random effects in linear mixed models-a bayesian approach. *Computational Statistics and Data Analysis*, 52:1347–1361, 2008.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

H. Ishwaran and L. F. James. Approximate dirichlet process computing for finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics*, In press, 2002.

H. Ishwaran and M. Zarepour. Markov chain monte carlo in approximate dirichlet and beta twoparameter process hierarchical models. *Biometrika*, 87:371–390, 2000.

Y. Itaya, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura. Deterministic annealing em algorithm in parameter estimation for acoustic model. *Interspeech*, pages 433–436, 2004.

K. Katahiral, K. Watanabe, and M. Okada. Deterministic annealing variant of variational bayes method. *Journal of Physics: Conference Series*, 95, 2008.

D. Kim and B. G. Lindsay. Using condence distribution sampling to visualize confidence sets. *Statistica Sinica*, 2009.

R. M. Korwar and M. Hollander. Contributations to the theory of dirichlet processes. *The Annals of Statistics*, 1:705–711, 1973.

S. T. Lanza and L. M. Collins. A mixture model of discontinuous development in heavy drinking from ages 18 to 30: The role of college enrollment. *Journal of Studies on Alcohol*, 67:552–561, 2006.

P. F. Lazarsfeld and N. W. Henry. *Latent structure analysis*. Houghton Mifflin, Boston, 1968.

R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

J. S. Liu. Nonparametric hierarchical bayes via sequential imputations. *The Annals of Statistics*, 42:911–930, 1996.

D. J. C. MacKay. Emsemble learning for hidden markov models. 1997.

A. L. McCutcheon. Sexual morality, pro-life values, and attitudes toward abortion: A simultaneous latent structure analysis for 1978–1983. *Sociological Methods and Research*, 16: 256–275, 1987.

B. O. Muthén and K. Shedden. Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55:463–469, 1999.

L. K. Muthén and B. O. Muthén. *Mplus user's guide*. Muthén & Muthén, Los Angeles, 3rd edition, 2004.

J. Park, W. Cho, and S. Park. Deterministic annealing em and its application in natural image segmentation. *Computational and Information Science*, 3314:639–644, 2005.

G. P. Patil and C. Taillie. Diversity as a concept and its implications for random communities. *Bulletin of the International Statistical Institute*, 47:497–515, 1977.

M. Perman and J. Pitman. Size-biased sampling of poisson point processes and excursions. probability theory and related fields. *Probability Theory and Related Fields*, 92:21–39, 1992.

J. Pitman. *Some developments of the Blackwell-MacQueen urn scheme. In Statistics, Probability and Game Theory (Edited by T. S. Ferguson, L. S. Shapley and J. B. MacQueen)*, volume 30. 1996.

J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25:855–900, 1997.

C. K. Reddy and B. Rajaratnam. Learning mixture models via component-wise parameter smoothing. *Computational Statistics and Data Analysis*, 54:732–749, 2010.

S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59:731–792, 1997.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

M. Tanner and W. Wong. The calculation of posterior distributions by data augmen- tation. *Journal of the American Statistical Association*, 82:528–550, 1987a.

W. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation,. *Journal of the American Statistical Association*, 82:528–550, 1987b.

N. Ueda and R. Nakano. Deterministic annealing em algorithm. *Neural Networks*, 11:271–282, 1998.

N. Ueda, R. Nakano, Z. Ghahramani, and G. E. Hinton. Split and merge em algorithm for improving gaussian mixture density estimates. *Journal of VLSI Signal Processing Systems*, 26:133–140, 2000.

W. F. Velicer, C. A. Redding, M. D. Anatchkova, J. L. Fava, and J. O. Prochaska. Identifying cluster subtypes for the prevention of adolescent smoking acquisition. *Addictive Behaviors*, 32:228–247, 2007.

T. Xu, Z. Zhang, P. S. Yu, and B. Long. Evolutionary clustering by hierarchical dirichlet process with hidden markov state. 2008.

Y. Yang. Can the strengths of aic and bic be shared? *BIOMETRICA*, 92, 2003.

W. Yao and B. G. Lindsay. Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association*, 104:758–767, 2009.