MIXED MEMBERSHIP DISTRIBUTIONS WITH APPLICATIONS TO MODELING MULTIPLE STRATEGY USAGE

april galyardt

Presented in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Statistics
Program for Interdisciplinary Education Research

Carnegie Mellon University
Pittsburgh, PA 15213

July 2012

ADVISORS:

Stephen Fienberg

Brian Junker

COMMITTEE:

David Klahr

John Lafferty

Cosma Shalizi

Dedicated to my husband, Jason.

This has been quite a ride and I never would have made it without your unending love, patience and support.

ABSTRACT

This dissertation examines two related questions. *How do mixed membership models work?* and *Can mixed membership be used to model how students use multiple strategies to solve problems?*

Mixed membership models have been used in thousands of applications from text and image processing to genetic microarray analysis. Yet these models are crafted on a case-by-case basis because we do not yet understand the larger class of mixed membership models.

The work presented here addresses this gap and examines two different aspects of the general class of models. First I establish that categorical data is a special case, and allows for a different interpretation of mixed membership than in the general case. Second, I present a new identifiability result that characterizes equivalence classes of mixed membership models which produce the same distribution of data. These results provide a strong foundation for building a model that captures how students use multiple strategies.

How to assess which strategies students use, is an open question. Most psychometric models either do not model strategies at all, or they assume that each student uses a single strategy on all problems, even if they allow different students to use different strategies. The problem is, that's not what students do. Students switch strategies. Even on the very simplest of arithmetic problems, students use different strategies on different problems, and experts use a different mixture of strategies than novices do.

Assessing which strategies students use is an important part of assessing student knowledge, yet the concept of ‘strategy’ can be ill-defined. I use the Knowledge-Learning-Instruction framework to define a strategy as a particular type of integrative knowledge component. I then look at two different ways to model how students use multiple strategies.

I combine cognitive diagnosis models with mixed membership models to create a multiple strategies model. This new model allows for students to switch strategies from problem to problem, and allows us to estimate both the strategies that students are using and how often each student uses each strategy. I demonstrate this model on a modestly sized assessment of least common multiples.

Lastly, I present an analysis of the different strategies that students use to estimate numerical magnitude. Three smaller results come out of this analysis. First, this illustrates the limits of the general mixed membership model. The properties of mixed membership models developed in this dissertation show that without serious changes to the model, it cannot describe the variation between students that is present in this data set. Second, I develop a exploratory data analysis method for summarizing functional data. Finally, this analysis demonstrates that existing psychological theory for how children estimate numerical magnitude is incomplete. There is more variation between students than is captured by current theoretical models.

ACKNOWLEDGEMENTS

I'd like to thank my advisors. Steve, you never failed to ask exactly the right question to make me work harder or think about a problem differently. Brian, I cannot express my appreciation for all the times you helped me talk through a problem, and helped me figure out how to explain an idea better.

I'd also like to thank Cosma Shalizi, David Klahr, John Lafferty, Ken Koedinger, and Johnathan Templin. Your questions and comments inspired many of the ideas I present here.

This work was supported in part by Graduate Training Grant awarded to Carnegie Mellon University by the Department of Education # R305B040063.

Thanks to Philip Pavlik, Micheal Yudelson, and Ken Koedinger for generously sharing the Least Common Multiples data analyzed in Chapter 4. This research was supported by the U.S. Department of Education (IES-NCSER) #R305B070487.

Thanks to Bob Siegler and John Opfer for generously sharing the Numerical Magnitude Estimation data analyzed in Chapter 5.

CONTENTS

List of Figures xi

List of Tables xv

Acronyms and Notation xvi

1 INTRODUCTION TO MIXED MEMBERSHIP 1

1.1 The General Mixed Membership Model 2

1.2 The Early Development of Mixed Membership 5

1.2.1 Grade of Membership Model 6

1.2.2 Population Admixture Model 6

1.2.3 Latent Dirichlet Allocation 8

1.3 Variations of Mixed Membership Models 8

1.3.1 Distributions of the Membership Parameter 9

1.3.2 Profile Distributions 10

1.4 Contributions of this Dissertation 11

2 PROPERTIES OF MIXED MEMBERSHIP DISTRIBUTIONS 14

2.1 Categorical Data 14

2.2 Conditions for a Between Interpretation 19

2.2.1 The Multinomial Distribution 23

2.3 Relationship between MMM and FMM 27

2.4 Identifiability 34

2.5 Discussion and Implications for Practice 46

3 MODELING MULTIPLE STRATEGIES 51

3.1 Example of Multiple Strategy Use in Addition 52

3.2 Modeling Strategies 54

3.3 The Knowledge-Learning-Instruction Framework 55

3.4 Definition of 'Strategy' 58

3.5 Combining Mixed Membership with Cognitive Diagnosis Models 61

3.6 Response Time and Other Variables 64

3.6.1 Existing models for Responses and Time 66

3.6.2 Time, Accuracy and Strategy 68

3.7 The Multiple Strategies Model 69

3.8 Comments 72

4 MULTIPLE STRATEGIES IN LEAST COMMON MULTIPLES ASSESSMENT DATA 73

4.1 Least Common Multiples Data 74

4.2 Multiple Strategies Model 75

4.2.1 Strategy Profiles 77

4.2.2 Priors for Strategy Profile Parameters 78

4.2.3 Membership in Strategy Profiles 80

4.2.4 Distribution of the Membership Parameter 80

4.2.5 Data Augmentation 82

4.2.6 The Complete Model 83

4.3 MCMC estimation 83

4.3.1 Update for μ and Σ 84

4.3.2 Update for θ 84

4.3.3 Update for Z 85

4.3.4 Update for λ 85

4.3.5 Update for β 86
4.4 Simulations and Results 87
4.4.1 Simulation 1: Average of 30 items per student, flat prior for λ 90
4.4.2 Simulation 2: Average of 15 items per student, flat prior for λ. 97
4.4.3 Simulation 3: Average of 15 items per student, Known λ 100
4.4.4 Simulation 4: Average of 15 items per student, Informative prior for λ_{1j}. 103
4.5 Analysis of Least Common Multiple Assessment Data 106
4.5.1 Model with 2 response times 107
4.5.2 Model without Time 115
4.5.3 Model with Total Time 120
4.6 Comments 125
5 MULTIPLE STRATEGIES IN CHILDREN'S NUMERICAL ESTIMATION 126
5.1 Introduction 126
5.2 Method 128
5.3 Concerns With Common Analysis Techniques 128
5.4 Clustering Method for FDA 132
5.5 Cluster-Based Visualization Results 136
5.5.1 Data from Siegler and Booth, 2004 136
5.5.2 Data from Booth and Siegler, 2006, Experiment 1 140
5.5.3 Data on 0-1000 scale 142
5.6 Model Based Analysis 144
5.6.1 Probability model for the Linear-Logarithmic Representation Theory 144

5.6.2 Probability Model for 2-Piece Linear Theory 146
5.6.3 Probability Model for 3-Piece Linear Model 147
5.7 Results and Discussion 148
6 DISCUSSION AND DIRECTIONS FOR FUTURE WORK 153
6.1 Mixed Membership Distributions 153
6.2 Future Work on Mixed Membership 154
6.2.1 Identifiability 154
6.2.2 An Alternate Mixed Membership Model 156
6.3 Multiple Strategies Model 160
6.4 Future Work on Multiple Strategies 161

Appendix 164

A MSM R-CODE 165
B FDA CLUSTERING R-CODE 172

BIBLIOGRAPHY 174

LIST OF FIGURES

Figure 2.1 Linear map from membership parameter space to individual response probabilities. 17

Figure 2.2 Example of a mixed membership logistic regression model. 27

Figure 2.3 Illustration of how MMM profiles form the basis for FMM components. 33

Figure 2.4 Illustrates how multiple sets of MMM basis profiles generate the same FMM components. 38

Figure 2.5 Example of the complete set of MMM basis profiles which all generate the same FMM mixture components. 39

Figure 2.6 Illustrates different distributions of the membership parameter θ which result in different sizes of the class of equivalent MMM. 43

Figure 3.1 Example of a problem with multiple solution strategies. 59

Figure 3.2 Illustration of the relationship between strategies and items in a multiple-strategies model. 60

Figure 4.1 Relationship between skills and strategies for the Least Common Multiples application of the MSM. 76

Figure 4.2 Distribution of θ in Simulation 1, replication 1 (30 items per student). 92

Figure 4.3 Distribution of θ in Simulation 1, replication 2 (30 items per student). 92

Figure 4.4 Posterior distribution of θ_i for one individual. Simulation 1, replication 1 (30 items per student) 93

Figure 4.5 Posterior distribution of θ_i for one individual. Simulation 1, replication 1 (30 items per student) 94

Figure 4.6 Posterior distribution of θ_i for one individual. Simulation 1, replication 2 (30 items per student) 95

Figure 4.7 Posterior distribution of θ_i for one individual. Simulation 1, replication 2 (30 items per student) 96

Figure 4.8 Distribution of θ in Simulation 2, replication 1 (15 items per student) 98

Figure 4.9 Distribution of θ in Simulation 2, replication 2 (15 items per student) 98

Figure 4.10 Distribution of θ in Simulation 2, replication 3 (15 items per student) 99

Figure 4.11 Distribution of θ in Simulation 2, replication 4 (15 items per student) 99

Figure 4.12 Distribution of θ in Simulation 3 (15 items per student) 100

Figure 4.13 Posterior distribution of θ_i for one individual. Simulation 3 (15 items per student) 101

Figure 4.14 Posterior distribution of θ_i for one individual. Simulation 3 (15 items per student) 102

Figure 4.15 Distribution of θ in Simulation 4, replication 1 (15 items per student) 105

Figure 4.16 Distribution of θ in Simulation 4, replication 2 (15 items per student) 105

Figure 4.17 Posterior distribution of θ in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses a flat prior for λ 108

Figure 4.18 Posterior means for response accuracy parameters in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses a flat prior for λ 110

Figure 4.19 Posterior distribution of θ in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses an informative prior for λ_{1j} 112

Figure 4.20 Posterior means for response accuracy parameters in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses an informative prior for λ 114

Figure 4.21 Posterior distribution of θ in analysis of LCM data. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ 116

Figure 4.22 Posterior means for response accuracy parameters in analysis of LCM data. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ 117

Figure 4.23 Posterior distribution of θ_i for one individual in analysis of LCM data. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ 118

Figure 4.24 Posterior distribution of θ_i for one individual in analysis of LCM data. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ 119

Figure 4.25 Posterior distribution of θ in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij})$, and estimation uses an informative prior for λ_{1j} 122

Figure 4.26 Posterior means of response accuracy parameters in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij})$, and estimation uses an informative prior for λ_{1j} 124

Figure 5.1 Standard visual summary of data used for number line estimation task. 130

Figure 5.2 Figure demonstrates the failings of using a median curve to summarize functional data. 131

Figure 5.3 Number line estimation task data for a first grade participant in Siegler and Booth (2004). 132

Figure 5.4 Clusters of curves from Siegler and Booth, 2004. 137

Figure 5.5 Individual data from a child with an outlier pattern. 138

Figure 5.6 Clusters of curves from Siegler and Booth, 2004, after removing obvious outliers. 139

Figure 5.7 Clusters of individually smoothed curves, from Booth and Siegler, 2006, Experiment 1, after removing obvious outliers. 141

Figure 5.8 Clusters of individually smoothed curves for combined data from Booth and Siegler (2006) Experiment 2 and Opfer and Siegler (2007). 143

Figure 5.9 Posterior mean estimation strategies for participants from Siegler and Booth (2004), shown grouped by grade and strategy cluster. Clusters are shown in columns. 150

Figure 6.1 Illustration of a normal regression version of the alternate mixed membership model. 159

LIST OF TABLES

Table 3.1 Common knowledge component categories (Table 3 from Koedinger et al. (2010)) 58

Table 4.1 Summary of MSM Simulations 89

Table 4.2 Parameter estimates for response time distributions in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses a flat prior for λ 109

Table 4.3 Parameter estimates for response time distributions in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses an informative prior for λ_{1j} 113

Table 4.4 Parameter estimates for response time distributions in analysis of LCM data. $X_{ij} = (C_{ij}, T_{ij})$, and estimation uses an informative prior for λ_{1j} 123

Table 5.1 Details of the four experiments re-analyzed in Chapter 5. 129

Table 5.2 DIC scores for the three probability models on each of the 4 experiments, as well as data simulated from a logarithmic-linear model. 149

ACRONYMS AND NOTATION

cdf	Cumulative Distribution Function
CDM	Cognitive Diagnosis Model (Henson et al., 2009)
FDA	Functional Data Analysis (Ramsay and Silverman, 2005)
FMM	Finite Mixture Model (Titterington et al., 1985)
LDA	Latent Dirichlet Allocation (Blei et al., 2003)
MMM	Mixed Membership Model (Erosheva, 2002)
MSM	Multiple Strategy Model

i indexes individuals observed in the data, $i = 1, \ldots, N$

j indexes observed variables $j = 1, \ldots, J$

k indexes mixed membership profiles $k = 1, \ldots, K$

X_{ij} observed data for feature/item j for individual i

θ_i membership vector for individual i, indicates how much individual i belongs to each profile k.

F_{kj} The distribution of observations for the kth profile on the jth variable

$F_k = \prod_{j=1}^{J} F_{kj}$ The distribution of observations for the kth profile

$\zeta \in \{1, \dots, K\}^J$ indexes the components of the mixture when a MMM is expressed in FMM form.

1

INTRODUCTION TO MIXED MEMBERSHIP

Mixed membership is based on a simple, intuitive idea. Individuals in a population belong to multiple subpopulations, not just a single class. For example, documents may be about multiple topics at the same time (Blei et al., 2003). Patients sometimes get multiple diseases at the same time (Woodbury et al., 1978). Birds may have genetic heritage from multiple subgroups (Pritchard et al., 2000). Children may use multiple strategies to solve mathematics problems (Chapter 4).

The problem of how to turn this intuitive idea into an explicit probability model was originally solved by Woodbury et al. (1978) and later independently by Pritchard et al. (2000), and Blei et al. (2003). Erosheva (2002) and Erosheva et al. (2004) then built a general mixed membership framework to incorporate all three of these models.

This dissertation explores the properties of the general mixed membership model, and whether mixed membership is useful for describing the ways in which students use multiple strategies. This first chapter introduces mixed membership and its historical development, then highlights the contributions of this work.

1.1 THE GENERAL MIXED MEMBERSHIP MODEL

The general mixed membership model (MMM), as defined in Erosheva et al. (2004), specifies an explicit probability model for the idea that individuals belong to multiple classes. The model is defined by four layers of assumptions: population level, subject level, latent variable level, and sampling scheme.

At the population level, the basic assumption is that there are K profiles within the population. If the population is a corpus of documents, then the profiles may represent the topics in the documents. If we are considering the genetic makeup of a population of birds, then the profiles may represent the different genetic heritages present in the populations. In image analysis, the profiles may represent the different categories of objects or components in the images, such as mountain, water, car, etc. When modeling the different strategies that students use to solve problems, then each profile can represent one of the strategies.

Within each profile, each feature or variable that we measure has a different distribution. We index the variables by $j = 1, \ldots, J$, and index the profiles as $k = 1, \ldots, K$. The distribution of observations for the kth profile on the jth variable is then given by the cumulative distribution function (cdf) F_{kj}. In an image processing setting, this indicates that the profile for the water category has a different distribution of features than the mountain profile. In another application, such as an assessment of student learning, different strategies may result in different response times on different problems. The F_{kj} contain all of the information about these differences between the profiles.

The next layer of assumptions is the individual level. Each individual has a membership vector that indicates the degree to which they belong to each profile,

$\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$. The term *individual* here could refer to an image, document, gene, person, etc. The components of θ are non-negative and sum to 1, so that θ can be treated as a probability vector. Thus, if student i used strategies 1 and 2, each about half the time, then this student would have a membership vector of $\theta_i = (0.5, 0.5, 0, \ldots, 0)$. Similarly, if an image was 40% water and 60% mountain then this would be indicated by θ_i.

For a particular variable j, we assume that the response distribution of individual i, conditional on the membership vector θ_i is given by the individual-level mixture;

$$F(x_j|\theta_i) = \sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j). \tag{1.1}$$

Blei et al. (2003) explicitly add a data augmentation indicator vector Z_{ij} to the Latent Dirichlet Allocation model (LDA). $Z_{ijk} = 1$ if individual i acts as a member of profile k for feature j, and $Z_{ijk} = 0$ otherwise. Thus for an image, if segment j of image i contains water, then $Z_{ij,water} = 1$. Or in the assessment setting, $Z_{ijk} = 1$ if student i used strategy k on problem j. The assumption is then that $F(x_j|Z_{ijk} = 1) = F_{kj}$, and that the membership vector θ_i gives the distribution of Z_{ij}, in that $Pr(Z_{ijk} = 1) = \theta_{ik}$. This is equivalent to equation 1.1, since

$$F(x_j|\theta_i) = \sum_{k=1}^{K} F(x_j|Z_{ijk} = 1) Pr(Z_{ijk} = 1|\theta_i) = \sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j). \tag{1.2}$$

The Z_{ij} indicator variables explicitly model partial membership as a *switching* behavior. For example, the student with partial membership in different strategies uses one strategy on some items and switches to another strategy on other items. In LDA, some words come from one topic and other words come from different topics. Equation 1.2 clarifies that even when the Zs are not explicitly included in the model, the individual-level mixture can always be interpreted as modeling a switching behavior.

To continue building the individual level of the model, we assume that observed variables are independent conditional on the membership vector, θ. In psychometrics, this is known as a local independence assumption. This assumption allows us to write the joint distribution of the full response vector $x = (x_1, \ldots, x_J)$, conditional on θ_i

$$F(x|\theta_i) = \prod_{j=1}^{J} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j) \right] \tag{1.3}$$

This conditional independence assumption also contains the assumption that the profile distributions are factorable. If an individual belongs exclusively to profile k (for example, an image contains only water) then $\theta_{ik} = 1$, and all other elements in the vector θ_i are zero. Thus,

$$F(x|\theta_{ik} = 1) = \prod_{j} F_{kj}(x_j) = F_k(x) \tag{1.4}$$

At the sampling scheme level, we may observe replications of each of the J feature variables. For example, in LDA $J = 1$, since only the presence or absence of words is being observed, but there are many replications of this measurement, and a different number of replications for each document in the sample. Let R_{ij} be the number of replications of variable j for individual i. Then the individual response distribution becomes:

$$F(x|\theta_i) = \prod_{j=1}^{J} \prod_{r=1}^{R_{ij}} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_r) \right] \tag{1.5}$$

Note that equations 1.1, 1.3 and 1.5 vary for each individual with the value of θ_i. It is in this sense that MMM is an individual-level mixture model. The distribution of variables for each profile, the F_{kj}, is fixed at the population level, so that the components of the mixture are the same, but the proportions of the mixture change individually with the membership parameter θ_i.

At the latent variable level, we can treat the membership vector θ as either fixed or random. If we wish to treat θ as random, then we can integrate equation 1.5 over the distribution of θ, yielding:

$$F(x) = \int \prod_{j=1}^{J} \prod_{r=1}^{R_{ij}} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_r) \right] dD(\theta) \tag{1.6}$$

The final layer of assumptions about the latent variable θ is crucial for purposes of estimation, but it is unimportant for the theoretical results presented here. The new results in Chapter 2 depend only on the structure of equation 1.3, and not on the distribution of θ or the shape of the F_{kj}.

This general mixed membership model is closely related to finite mixture models. Finite mixture models can be considered a special case of the mixed membership model when the membership parameter θ is restricted to the corners of the simplex, where one component is 1 and all others are 0. However, every mixed membership model can also be expressed as a finite mixture model with a much larger number of classes, and constraints on the class probabilities. Erosheva, et al. (2007) shows that this relationship holds for categorical data, and Theorem 2.3 shows that it holds in general.

1.2 THE EARLY DEVELOPMENT OF MIXED MEMBERSHIP

Woodbury et al. (1978) first introduced the mixed membership idea with the grade-of-membership model to describe individuals with multiple disease profiles. In the early 2000's, two other versions of mixed membership were independently developed. Pritchard et al. (2000) developed an admixture model to describe the presence of individuals in a population with mixed genetic heritage.

Blei et al. (2003) developed Latent Dirichlet Allocation (LDA) to describe the topics present in a corpus of documents where many documents are about multiple topics.

1.2.1 *Grade of Membership Model*

The Grade of Membership model (GoM) is by far the earliest example of mixed membership (Woodbury et al., 1978). The motivation for creating this model came from the problem of designing a system to help doctors diagnose patients. The problems with creating such a system are numerous: Patients may not have all of the classic symptoms of a disease, they may have multiple diseases, relevant information may be missing from a patient's profile, and many diseases have similar symptoms.

In this setting, the mixed membership profiles represent distinct diseases. The observed data X_{ij} are the categorical levels of indicator j for patient i. The profile distributions $F_{kj}(x_j)$ indicate which level of indicator j is likely to be present in disease k. Since X_{ij} is categorical, and there is only one measurement of an indicator for each patient, the profile distributions are multinomial with $n = 1$. In this application, the individual's disease profile is the object of inference, so that θ_i is treated as fixed. Thus the likelihood in Equation 1.3 is used in estimation.

1.2.2 *Population Admixture Model*

Pritchard et al. (2000) models the genotypes of individuals in a heterogeneous population. The profiles represent the distinct populations of origin, with individuals having inherited some genes from the different sub-populations in which they have partial membership.

The variables X_j are the genotypes observed at J locations, and for diploid individuals two replications $R_j = 2$ are observed at each location. Across a population, a finite number of distinct alleles are observed at each location j, so that X_j is categorical, and F_{kj} is multinomial for each sub-population k.

In this application, the distribution of the membership parameters θ_i is of as much interest as the parameters themselves. The parameters θ_i are treated as random realizations from a symmetric Dirichlet distribution. The new results in Chapter 2 indicate that the symmetric distribution is problematic for estimation and interpretation. In a finite mixture model, the model is only identifiable up to permutations of indices, so that there would be $K!$ equivalent model, this is the number which Pritchard et al. (2000) anticipates. However, with a symmetric distribution for θ, there are $K!^{J-1}$ equivalent models, which are created by a very different type of permutation (Theorem 2.7).

One of the more interesting features of the admixture model is that it includes the possibility of both unsupervised and supervised learning. Most mixed-membership models are estimated as unsupervised models. That is, the model is estimated with no information about what the profiles might be, and no information about which individuals might have some membership in the same profiles. Pritchard et al. (2000) considers the unsupervised case, but they also consider the case where we have additional information. In their case, the location where an individual bird was captured provides information about which sub-population it is likely has some membership in, even though it may be the descendant of an immigrant bird. This information is included with a carefully constructed prior on θ, which incorporates rates of migration.

1.2.3 *Latent Dirichlet Allocation*

Latent Dirichlet Allocation Blei et al. (2003) is in some ways the simplest example of mixed membership, as well as the most popular. This is a textual analysis model, where the goal is to identify the topics present in a corpus of documents. Mixed membership is necessary, because many documents are about more than one topic.

LDA uses a bag of words model, where only the presence or absence of words in a document is modeled, and word order is ignored. The individuals i are the documents. The profiles k represent the topics. LDA models only one variable, $J = 1$, the words present in the documents. The number of replications R_{ij} is simply the number of words in document i. The profile distributions are multinomial distributions over the set of words: $F_{kj} = Multinomial(\lambda_k, n = 1)$, where λ_{kw} is the probability that a particular word in topic k will be the word w. LDA uses the integrated likelihood in equation 1.6. The focus here is on estimating the topic profiles, and the distribution of membership parameters, rather than the θ_i themselves. Blei et al. (2003) also uses a Dirichlet distribution for θ, however they do not use a *symmetric* Dirichlet, and so avoid the identifiability issues that are present in Pritchard et al. (2000).

1.3 VARIATIONS OF MIXED MEMBERSHIP MODELS

Though, LDA was not the first mixed membership model, it has quickly become the most popular, currently with over 4000 citations according to Google Scholar. LDA is only one variation of the general mixed membership model developed by Erosheva et al. (2004); however, LDA has become so popular that it continues to inspire new models, based only on LDA. Yet these new models still fit within Erosheva's general model.

1.3.1 *Distributions of the Membership Parameter*

Many mixed membership models have been created as variations of LDA with a different distribution of the membership parameter θ. LDA uses a Dirichlet distribution for the membership parameter θ. The Dirichlet distribution introduces a strong independence condition on the components of θ subject to the constraint $\sum_k \theta_{ik} = 1$ (Aitchison, 1982). In many applications, this strong independence assumption is a problem.

For example, in text analysis an article with partial membership in an evolution topic is more likely to also be about genetics than astronomy. In order to model an interdependence between profiles, Blei and Lafferty (2007) use a logistic-normal distribution for θ. Blei and Lafferty (2006) take this idea a step further and create a dynamic model where the mean of the logistic-normal distribution evolves over time.

Fei-Fei and Perona (2005) analyze images, where the images contain different proportions of the profiles *water, sky, foliage, etc.* However, images taken in different locations will have a different underlying distribution for the mixtures of each of these profiles. For example, rural scenes will have more foliage and fewer buildings than city scenes. Fei-Fei and Perona address this by giving the membership parameters a distribution that is a mixture of Dirichlets.

1.3.2 *Profile Distributions*

Other models have been created by altering the profile distributions, so that they are no longer multinomial. The Latent Process Decomposition model (Rogers et al., 2005) models the different processes that might be responsible for different levels of gene expression observed in microarray data sets. In this application, X_{ij} measures the expression level of the jth gene in sample i, a continuous quantity. This leads to profile distributions $F_{kj} = N(\mu_{kj}, \sigma_{kj})$.

The simplical mixture of Markov chains (Girolami and Kaban, 2005) is a mixed membership model where each profile is characterized by a Markov chain transition matrix. The idea is that over time an individual may engage in different activities, and each activity is characterized by a probable sequence of actions.

Shan and Banerjee (2011) is another interesting extension of LDA which seeks to define a 'generalization' of LDA, which they call a mixed-membership naive Bayes model. This model simply requires the profile distributions F_{kj} to be exponential family distributions. This is a subset of models that fall within the general mixed membership model introduced by Erosheva et al. (2004), and as Theorem 2.2 shows, other exponential family profile distributions will not have the same properties as the multinomial profiles used in LDA. The main contribution of Shan and Banerjee (2011) is a comparison of different variational estimation methods for particular choices of F_{kj}.

1.4 CONTRIBUTIONS OF THIS DISSERTATION

This dissertation makes contributions in several different disciplinary areas. First, are contributions in statistics and machine learning; I establish fundamental properties of mixed membership distributions. Next, are contributions in psychometrics and the learning sciences; I build a statistical model that recognizes strategy choice as an important component of expertise, a model which is capable of estimating both the strategies present in the data and how much each student uses each strategy. Two other contributions include the development of a new exploratory data analysis method for functional data, and new psychological results regarding how children estimate numerical magnitude.

A wide variety of mixed membership models now exist, based on early models, and LDA in particular. The enormous variety in this class of models suggests that we need to better understand the properties of the general mixed membership family of distributions. We need to know which applications represent the special cases, and we need to understand how different choices, such as the choice of a particular distribution for θ affects the identifiability and estimability of the model. The theoretical results in Chapter 2 address these questions. I establish the different model interpretations that are possible in the general case, and in the special case of categorical data. I then develop an identifiability result which characterizes classes of mixed membership models which produce the same data distribution.

Chapter 3 provides a theoretical grounding for using mixed membership as a cognitive model to describe students switching between different strategies. First, I develop a definition of *strategy* within the Knowledge-Learning-Instruction framework Koedinger et al. (2010) that is consistent both with common English usage of the term and is practical for modeling student performance. I then combine cognitive diagnosis models with mixed membership models to build a multiple strategies model that describes students switching strategies between assessment items. The mixed membership framework allows for expanding the multiple strategies model to include additional variables, such as response time or self-reported strategy. The full model is a novel method for jointly modeling all observed student data for a particular item as conditionally dependent on the same cognitive process.

Each strategy is defined by a process-signature, that is each strategy is only identifiable to the extent that it differs from other strategies on the observed variables. If two strategies have the same error rate and have the same distribution of response times, this model will not distinguish between them even if they are arguably different cognitively.

This multiple strategies model represents a substantial contribution in modeling student knowledge. Experts and novices use different strategies to solve problems. Assessing which strategies students use is an important factor in determining what students know.

Simply building the model, however, is not enough. We also need to know that this model is useful for inference in real applications. Chapter 4 demonstrates a simple application of the multiple strategies model. One of the goals of this application is to determine whether we can learn both the strategies and how much students use them from the data, or if we need to know the strategies *apriori* in order to estimate student knowledge. The data set is rather small, with only 15 items per student. Yet, with prior knowledge of only one strategy, we can still estimate how much students are using each of the strategies in the model, and the remaining strategies. The good performance is due, in part, to the incorporation of the response time data. This application functions primarily as a proof-of-concept for the multiple-strategies model.

Chapter 5 illustrates another application where students use multiple strategies. The data are functional data from experiments on how young children estimate numerical magnitude. Due to the properties described in Chapter 2, mixed membership is a poor choice for modeling this data. Instead, I develop a new exploratory data analysis technique to summarize the patterns in this type of data. A model-based analysis then reveals that existing psychological theory is insufficient to describe the strategies that children use to estimate numerical magnitude. This psychological result is relevant to education since the ability estimate numbers accurately is closely related to arithmetic skills (Booth and Siegler, 2008).

One final contribution of this dissertation is that it represents a serious attempt to integrate the cognitive literature on how children use multiple strategies, and how strategy use is an indicator of expertise, with the statistical and psychometric literature to model student performance. The National Research Council report *Knowing What Students Know* Pellegrino et al. (2001) charges that, "Traditional tests do not focus on many aspects of cognition that research indicates are important, and they are not structured to capture critical differences in students' levels of understanding." This dissertation addresses that gap, and builds a model of student performance that is capable of capturing cognitively significant differences in student performance.

2

PROPERTIES OF MIXED MEMBERSHIP DISTRIBUTIONS

This chapter explores the theoretical properties of the general mixed membership model, as defined in Erosheva et al. (2004) and given in Chapter 1. I am not yet concerned with the theoretical properties of different estimators, but rather the more basic question of, “What do the data distributions look like?" The mixed membership model is rather complicated, and the answer to this simple question is not obvious.

In this chapter, I demonstrate that categorical data and other data types behave very differently in mixed membership models. I describe the two possible interpretations of what it means for an individual to have partial membership in multiple profiles, and establish when each interpretation is possible. Last, but certainly not least, I fully characterize equivalence classes of mixed membership models which will produce the same distribution of observed data.

2.1 CATEGORICAL DATA

The intuition for how mixed membership models behave was developed through the early applications including Woodbury et al. (1978); Pritchard et al. (2000);

Erosheva (2002); Blei et al. (2003); Manton et al. (2004) and Erosheva et al. (2004). Each of these papers has one feature in common, the data were categorical.

In the general mixed membership model (MMM), the individual distributions are given by Equation 1.1, which is repeated here for reference:

$$F(x_j|\theta_i) = \sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j). \tag{2.1}$$

For categorical data, this equation simplifies, and that is the key difference between categorical data and any other type of data.

In the general case, MMM models partial membership as a *switching* behavior; sometimes individuals act in accordance with one profile, and sometimes in accordance with another profile (Equation 1.2). The mathematical simplification that occurs for categorical data allows an additional interpretation. In this special case, MMM models individuals residing *between* profiles, and acting with a behavior that is between the profiles in which they have partial membership.

If variable X_j is categorical, then we can represent the possible values for this variable as $\ell_1, \ldots, \ell_{L_j}$. We represent the distribution for each profile as $F_{kj}(x_j) = \mathrm{Multinomial}(\lambda_{kj}, n = 1)$, where λ_{kj} is the probability vector for profile k on feature j, and n is the number of multinomial trials. The probability of observing a particular value l within basis profile k is written as:

$$\Pr(X_j = l|\theta_k = 1) = \lambda_{kj\ell_l} \tag{2.2}$$

The probability of individual i, with membership vector θ_i, having value l for feature j is then

$$\Pr(X_{ij} = l|\theta_i) \;=\; \sum_{k=1}^{K} \theta_{ik} \Pr(X_j = l|\theta_k = 1) \;=\; \sum_{k=1}^{K} \theta_{ik} \lambda_{kj\ell_l} \tag{2.3}$$

Consider as an example, latent Dirichlet allocation (LDA) (Blei et al. (2003), described in Section 1.2.3). Assume that document i belongs to the *sports* and *medicine* topics. The two topics each have a different probability distribution over the lexicon of words, say $\mathtt{Multinomial}(\lambda_s)$, and $\mathtt{Multinomial}(\lambda_m)$. The word *elbow* has a different probability of appearing in each topic, $\lambda_{s,e}$ and $\lambda_{m,e}$ respectively. Then the probability of the word *elbow* appearing in document i is given by $\lambda_i = \theta_{is}\lambda_{s,e} + \theta_{im}\lambda_{m,e}$. Since the vector θ_i sums to one, the individual probability λ_i must be between $\lambda_{s,e}$ and $\lambda_{m,e}$. The individual probability is *between* the probabilities in the two profiles.

We can simplify the mathematics further if we collect the λ_{kj} into a matrix by rows, and call this matrix λ_j. Then $\theta_i^T\lambda_j$ is a vector of length L_j where the lth entry is individual i's probability of value l on feature j, as in equation 2.3.

We can now write individual i's probability vector for feature j as

$$\lambda_{ij} = \theta_i^T\lambda_j. \tag{2.4}$$

The matrix λ_j defines a linear transformation from θ_i to λ_{ij}, as illustrated in Figure 2.1. Since θ_i is a probability vector and sums to one, λ_{ij} is a convex combination of the the profile probability vectors λ_{kj}, and the individual λ_{ij} lie in a simplex where the extreme points are the λ_{kj}. In other words the individual response probabilities lie between the profile probabilities. This leads Erosheva et al. (2004) and others to refer to the profiles as *extreme profiles*; for categorical data the parameters of the profiles form the extremes of the individual parameter space.

Moreover, since the mapping from the individual membership parameters θ_i to the individual feature probabilities λ_{ij} is linear, the distribution of individual response probabilities is effectively the same as the population distribution of membership parameters. (Figure 2.1.)

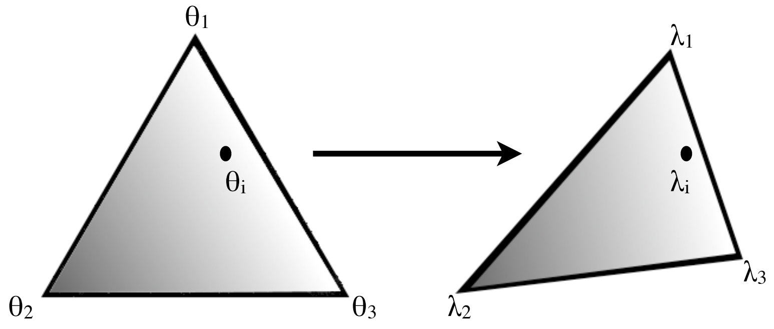

Figure 2.1: For multinomial basis profiles, the distribution of the membership parameter θ is mapped linearly onto a multinomial response probability. This allows us to interpret individual i's position in the θ-simplex as equivalent to their response probability vector.

Thus, when feature x_j is categorical, an individual with membership vector θ_i has a probability distribution of

$$F(x_j|\theta_i) = Multinomial(\theta_i^T\lambda_j, n = 1) \tag{2.5}$$

This is the property that makes categorical data special. When the profile distributions are multinomial with $n = 1$, the individual-level mixture distributions are also multinomial with $n = 1$. Already this is a special property, since we know for example that a mixture of normals is not normal. However, we also have that the parameters of the individual distributions, the $\theta_i^T\lambda_j$ are convex combinations of the profile parameters, the λ_{kj}.

In this sense, when the data are categorical, an individual with mixed membership in multiple profiles is effectively *between* those profiles. Even though the exchangeability structure of MMM, as described in Section 1.1 clearly describes an *switching* behavior, the early applications on categorical data lead to an equivalent *between* interpretation. These two equivalent interpretations have led to certain intuitions about how to interpret mixed membership, intuitions that may not hold in the general case.

We need to reiterate that while categorical data allows a special interpretation of partial membership in the multinomial parameter space, the behavior in data space is the same switching behavior as defined in the general case. Individuals may only give responses that are within the support of at least one of the profiles.

Consider the example of a word appearing in a document. *Camel* may be a high probability word in the *zoo* topic, while *cargo* has high probability in the *transportation* topic. For a document with partial membership in the zoo and transportation topics, the word *camel* will have a probability of appearing that is between the probability of *camel* in the zoo topic and its probability in the transportation topic. Similarly for the word *cargo*. However it doesn't make sense to talk about the word *cantaloupe* being between camel and cargo. With categorical data, there is no 'between' in the data-space. The between interpretation only holds in the parameter space.

Let us consider another example, suppose that we are looking at response times for a student taking an assessment. Suppose that one strategy results in a response time with a distribution $N(10, 1)$, and another less effective strategy has a response time distribution of $N(20, 2)$. In the mixed membership model, an individual with membership vector $\theta_i = (\theta_{i1}, \theta_{i2})$ then has a response time distribution of $\theta_{i1}N(10, 1) + \theta_{i2}N(20, 2)$. This individual may use strategy 1 or strategy 2, but a response time of 15 has a low probability under both strategies, and in the mixture. The individual may switch between using strategy 1 and strategy 2 on subsequent items, but a response time between the two distributions is never likely, no matter the value of θ. Moreover, the individual distribution is no longer normal, but a mixture of normals (Titterington et al., 1985). Thus, for this continuous data, we do not have a between interpretation in the parameter space either.

In Section 2.2, we show that the *between* interpretation does not hold in general. The multinomial distribution with $n = 1$ has a property which is not shared by other common distributions, and it is this property which allows the *between* interpretation of partial membership.

2.2 CONDITIONS FOR A BETWEEN INTERPRETATION

The between interpretation arises out of special property of the multinomial distribution: The individual probability distributions are in the same parametric family as the profile distributions, as both are multinomial with $n = 1$. Further, the individual parameters are between the profile parameters. Thus, for the between interpretation to hold in the general case, we need the individual distributions $F(x|\theta_i)$ to be in the same family of distributions as each profile distribution F_k. If we let ϕ_k represent the parameters of the profile distributions, then the individual parameters must lie between the ϕ_k.

In other words, the property we are looking for is that an individual with membership parameter θ_i would have an individual data distribution of $F(X; \theta_i^T \phi)$, so that for each variable j we have:

$$\sum_k \theta_{ik} F_{kj}(X_j; \phi_{kj}) = F_j(X_j; \theta_i^T \phi_{\cdot j}). \tag{2.6}$$

For simplicity, I will omit the subscript j for the remainder of this section, so that equation 2.6 becomes

$$\sum_k \theta_{ik} F_k(X; \phi_k) \ = \ F\left(X; \sum_k \theta_{ik} \phi_k\right). \tag{2.7}$$

From equation 2.7, we see the property that allows the between interpretation is that the cumulative distribution function F, or equivalently, the density function, is a linear transformation of its parameters ϕ. Theorem 2.1 specifies the form that a cumulative distribution function (cdf) must take in order for the between interpretation to hold. When the distributions of the profiles can be written as finite mixture models, then the between interpretation in the parameter space will hold.

Theorem 2.1. *Suppose that F is a cdf parameterized by the finite vector ϕ. F is a linear transformation of ϕ if and only if F can be written as a finite mixture model, that is:*

$$F(x;\phi) = \sum_{s=1}^{S} \phi_s B_s(X) \tag{2.8}$$

where $\sum \phi_s = 1$, $\phi_s \geqslant 0$ *for each* s, *and each* B_s *is itself a cdf.*

Proof. **I.** Suppose that $F(X;\phi) = \sum_{s=1}^{S} \phi_s B_s(X)$. Then

$$\begin{aligned}
\sum_{k=1}^{K} \theta_k F(X;\phi_k) &= \sum_k \theta_k \left[\sum_s \phi_{ks} B_s(X)\right] \\
&= \sum_k \sum_s \theta_k \phi_{ks} B_s(X) \\
&= \sum_s \sum_k \theta_k \phi_{ks} B_s(X) \\
&= \sum_s \left[\sum_k \theta_k \phi_{ks}\right] B_s(X) \\
&= F\left(X; \sum_k \theta_k \phi_{k\cdot}\right)
\end{aligned}$$

(both sums are finite, so the order is reversible.)

II. Suppose that F is a linear transformation of its parameters ϕ, and ϕ is a vector of length S:

$$\sum_{k=1}^{K} \theta_k F(X; \phi_k) \;=\; F\left(X; \sum_k \theta_k \phi_k\right)$$

A standard result from linear algebra says that any linear map between finite-dimensional vector spaces can be represented as a matrix product, so that F can be written as:

$$F(X; \phi) = B(X)\phi$$

Let e_i be the vector with 1 in the i^{th} entry and 0 otherwise. Note that if the space of possible values of ϕ does not include e_i for $i \in \{1, \dots, S\}$, then we can project ϕ onto ϕ^*, as $\phi^* = V\phi$ so that e_i is in the space of possible values for ϕ^*. Thus we can write:

$$F(X; \phi^*) \;=\; B(X)V\phi^* \;=\; B^*(X)\phi^* \;=\; \sum_{s=1}^{S} B^*_s(X)\phi^*_s$$

We now observe that

$$F(X; e_i) = B^*_i(X)$$

i.e., each $B^*_i(X)$ must be a cdf. In addition, we observe that since $F \rightarrow 1$ as $x \rightarrow \infty$:

$$\lim_{X\rightarrow\infty} F(X; \phi^*) \;=\; \sum_{s=1}^{S} \lim_{X\rightarrow\infty} B^*_s(X)\phi^*_s = \sum_s \phi^*_s = 1$$

Thus, we can parameterize F as a finite mixture model. □

Illustration 1

Let $\Phi(x)$ denote the cdf of the standard normal distribution, and for $s = 1 \dots S$, let c_s be fixed constants. Then the mixture distribution, parameterized by the vector ϕ,

$$F^*(x;\phi) = \sum_{s=1}^{S} \phi_s \Phi(x - c_s) \tag{2.9}$$

satisfies the condition of Theorem 2.1. Thus, if we set the profile distributions for variable j as $F^*(x_j;\phi_k)$, then we will be able to interpret individuals as residing between the basis profiles. In this case, if we have K profiles parameterized by the vectors ϕ_k, then an individual with membership parameter θ_i has the individual distribution:

$$F(x|\theta_i;\phi) = \sum_{k=1}^{K} \theta_{ik} \left[\sum_{s=1}^{S} \phi_{ks} \Phi(x - c_s) \right] = F^*(x;\theta_i^T \phi) \tag{2.10}$$

The parameters for individual i in the family of F^* distributions are $\theta_i^T \phi$ and reside in a simplex defined by the extreme points ϕ_k.

Now suppose the means of the mixture components are also parameters of the mixture, denoted $\phi^\diamond$. Then the basis profiles no longer satisfy the condition of Theorem 2.1:

$$F^\diamond(x;\ \phi, \phi^\diamond) = \sum_{s=1}^{S} \phi_s \Phi(x - \phi_s^\diamond)$$

So if the basis profiles of a mixed membership distribution are $F^\diamond$, then the between interpretation cannot be used, and we must interpret individuals as switching between basis profiles. This illustration clarifies that it is easy to create distributions which satisfy the requirements for a between interpretation, but these requirements are very strict. □

2.2.1 *The Multinomial Distribution*

Interpreting individuals with mixed membership in multiple profiles as being between the profiles requires profile distributions to be linear functions of their parameters. The multinomial distribution with $n = 1$ can be written as a sum of indicator functions, or as a an exponential family distribution, and is unique in this respect. The multinomial distribution with $n = 1$ is the only common distribution that will allow a between interpretation in a mixed membership model.

Let $X \sim \mathtt{Multinomial}(p, n = 1)$. We can write the distribution of X as a mixture model, $\sum_{l=1}^{L} p_l I_{a_l}(X)$, where $I_a(x)$ is an indicator function, and a_l represents the different levels of the multinomial distribution. We can also write the distribution in exponential family form. Recall that the density function of an exponential family distribution has the form $h(x)g(\phi)\exp\left[\eta(\phi)\cdot T(x)\right]$; for the multinomial distribution this is

$$(x_1! \ldots x_L!)^{-1}(n!)\exp\left[\sum_l x_l \log(p_l)\right]. \tag{2.11}$$

Theorem 2.2. *The multinomial distribution with* $n = 1$ *is the only exponential family distribution that can be written as a finite mixture model.*

Proof. If X has a distribution that can be written as a finite mixture model, then its density must be a mixture of densities, which we will write as $f(x) = \phi^T b(x)$. Since f is exponential family, we can then write

$$f(x) \;=\; \phi^T b(x) \;=\; h(x)g(\eta)\exp\{\eta^T t(x)\} \tag{2.12}$$

We can absorb g and h into other terms, so that this simplifies to

$$\phi^T b(x) \;=\; \exp\{\eta^T t(x)\} \tag{2.13}$$

Here we note that ϕ and η are distinct parameterizations, but it must be the case that $\phi = \phi(\eta)$, or likewise $\eta = \eta(\phi)$. Now, if we take the partial derivative with respect to η_i, we have

$$\left(\frac{\partial\phi}{\partial\eta_i}\right)^T b(x) \;=\; t_i(x)\exp\{\eta^T t(x)\} \tag{2.14}$$

Which, substituting, is

$$\left(\frac{\partial\phi}{\partial\eta_i}\right)^T b(x) \;=\; t_i(x)\left(\phi^T b(x)\right) \tag{2.15}$$

Rearranging terms yields

$$b(x)^T\left(\frac{\partial\phi}{\partial\eta_i} - \phi(\eta)t_i(x)\right) = 0 \tag{2.16}$$

Fix any x where $b(x) \neq 0$. Since Equation 2.16 must hold for all η, $\frac{\partial\phi}{\partial\eta_i} - \phi(\eta)t_i(x)$ must be identically zero. This implies that

$$\frac{\partial\phi}{\partial\eta_i} \;=\; t_i(x)\phi(\eta) \tag{2.17}$$

Since $t_i(x)$ is present on only the right hand side, it must be constant on its support; that is, $t_i(x)$ is an indicator function, which we shall write as t_i. The solution to equation 2.17 gives us $\phi_j(\eta) = \exp\{(t+c)^T\eta\}$. This gives us the exact form of a multinomial with $n = 1$ when written in exponential family form. □

Theorem 2.2 indicates that interpretation of mixed membership is restricted to the *switching* interpretation for all exponential family distributions, save one. This result is particularly important when considering the many extensions and variations of LDA. LDA was built for categorical data, the extensions use a variety of profile distributions. Shan and Banerjee (2011) creates a 'naive Bayes mixed membership model' which includes any MMM where the profiles are exponential family. All of these variations are restricted to interpreting mixed membership as individuals switching between profiles, when the LDA model they are extending also allows an interpretation of individuals residing between the profiles. The danger here is that it is easy to believe that extending the model allows an extension of the interpretation as well, when in fact LDA is a very special case of mixed membership.

Illustration from Applications to Data

Let us consider two specific parallel applications of mixed membership. Both applications are mixed membership regression models. For each individual i, and each feature j, we observe a data point X_{ij} and a covariate Z_{ij}.

In Manrique-Vallier (2010), the observed variable X is binary and the covariate Z is continuous. The basis profiles are defined by logistic regression functions.

$$\Pr(X_{ij} = 1|k = 1) = \text{logit}^{-1}(\text{˛}_{ojk} + \text{˛}_{1jk} Z_{ij}) = r_k(Z_{ij}). \tag{2.18}$$

So that

$$F_k(x|z) = \text{Bernoulli}(r_k(z)). \tag{2.19}$$

In Galyardt (2010), both the observed variable X and the covariate Z are continuous. The basis profiles are defined by normal regression functions.

$$X_{ij} = \beta_{0jk} + \beta_{1jk} Z_{ij} + \epsilon_{ij} \qquad \text{with } \text{›}_{ij} \sim N(o, \text{ff}_k^2). \tag{2.20}$$

Thus,

$$F_k(x|z) = N(r_k(z), \sigma_k^2). \tag{2.21}$$

Both of these are mixed-membership regression models, yet the behavior of the models at the individual level are strikingly different. The Bernoulli distribution satisfies the conditions of Theorem 2.1, and so the individual distributions of $X_{ij}|\theta_i$ are also Bernoulli distributions. Moreover, there is an individual regression function $r_i(z)$, where

$$r_i(z) = \sum_{k=1}^{K} \theta_{ik} r_k(z). \tag{2.22}$$

We note that the individual regression function $r_i(z)$ will not be a logistic function, yet $r_i(z)$ is a mean regression function for the individual probability distribution:

$$X_{ij}|\theta_i, Z_{ij} \sim \text{Bernoulli}(r_i(Z_{ij})). \tag{2.23}$$

Figure 2.2 illustrates these individual regression functions in the case where the profiles are logistic regression functions.

Now contrast the behavior of these logistic regression basis profiles with the behavior of the normal regression basis profiles, $F_k = N(r_k(z), \sigma_k^2)$. The individual probability distributions in the normal regression case are

$$X_{ij}|\theta_i, Z_{ij} \sim \sum_{k=1}^{K} \theta_i \, N\left(r_k(Z_{ij}), \sigma_k^2\right). \tag{2.24}$$

The distribution in Equation 2.24 does not simplify. In the Bernoulli case, the individual distribution can be summarized by a single regression function that is a convex combination of logistic regression functions. In the Normal case, the individual distribution is a mixture distribution and cannot be summarized by a single regression function. This result is one of the main reasons why the mixed membership model is an inappropriate analysis for the application in Chapter 5.

The unique dual representation of the multinomial distribution means that it is the only common distribution which allows a between interpretation in a mixed membership model. The early applications of mixed membership were all to categorical data, and thus all of the profile distributions were multinomial with $n = 1$.

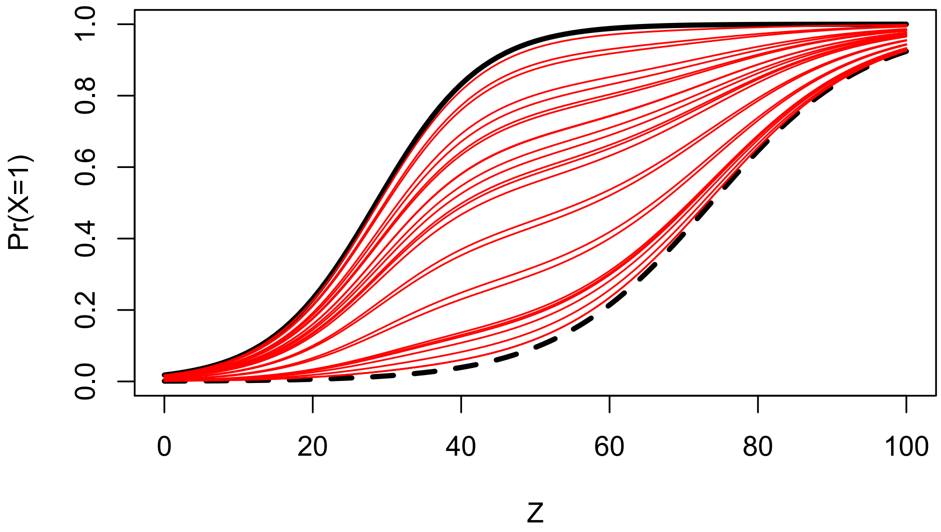

Figure 2.2: Example of a mixed membership regression model, where the profiles are logistic regression functions. This example includes only $K = 2$ profiles for clarity. The black solid and dashed lines show the two profile logistic regression functions, $r_k(z)$. The red lines show individual regression functions $r_i(x) = \sum_k r_k(z)$.

Intuition developed on early applications and versions of mixed membership was developed on categorical data. These intuitions and understandings of how to interpret mixed membership do not hold in general.

2.3 relationship between mixed membership and finite mixture models

The assumption at the subject level of the mixed membership model that features are conditionally independent given the membership vector θ has some surprising consequences for the data distribution. The structural exchangeability assumptions are shared by every mixed membership model whether the data are categorical, discrete or continuous. The more recent models, such as Airoldi et al. (2008) and Manrique-Vallier (2010) have increasingly complicated profile distributions. It is therefore, critical to understand how the basic model assumptions behave on their own and how they interact when combined with other distributions.

In the Section 1.1, we drew a contrast between mixed membership models (MMMs) and finite mixture models (FMMs). Here, we elaborate on this contrast, and define conditions when the two models are equivalent. FMMs are the probability model underlying many machine learning tasks including clustering and classification. FMMs divide a population up into multiple component pieces, each component with its own probability distribution, say $F_w(x)$. Each individual in the population belongs completely to one of these components, and each component makes up a certain proportion of the population, say π_w. Thus for an FMM, we write the population distribution as

$$F^{FMM}(x) = \sum_{w=1}^{W} \pi_w F_w(x) \tag{2.25}$$

In contrast, MMM assumes that each individual belongs to multiple profiles, which results in the individual distribution given in equation 1.3. To get a population distribution, we must integrate equation 1.3 over the population distribution of the individual mixed membership parameter, D, which yields:

$$F(x|\alpha) = \int \prod_{j=1}^{J} \left[\sum_{k=1}^{K} \theta_k F_{kj}(x_j) \right] D(d\theta) \tag{2.26}$$

The FMM in equation 2.25 is a population-level mixture model. MMM is an individual-level mixture model, so that the sum inside equation 2.26 is the same as in the FMM. However, the product is needed to account for measurements on different features, and the integral is needed to account for individuals belonging to different profiles in different proportions.

Despite the additional complications in equation 2.26 compared to 2.25, every mixed membership model can be re-expressed as a finite mixture model with a much larger number of components. Haberman (1995) suggested this relationship in his review of Manton et al. (1994). Erosheva et al. (2007) showed that it holds for categorical data, and recognized that the same result holds in the general case as well. I present the proof of Theorem 2.3 for the general case, because a general version of the proof is not recorded elsewhere. The relationship between mixed membership and finite mixture models forms the foundation for understanding the behavior of mixed membership in the general case.

Theorem 2.3. *Assume a mixed membership model with J features and K profiles. To account for any replications in features, assume that each feature j has R_j replications, and let $R = \sum_{j=1}^{J} R_j$. Write the profile distributions as*

$$F_k(x) = \prod_{r=1}^{R} F_{kr}(x_r).$$

Then the mixed membership model can be represented as a finite mixture model with components indexed by $\zeta \in \{1, \ldots, K\}^R = \mathcal{Z}$, where the classes are

$$F_{\zeta}^{FMM}(x) = \prod_{r=1}^{R} F_{\zeta_r r}(x_r) \tag{2.27}$$

and the probability associated with class ζ is

$$\pi_{\zeta} = \mathbb{E}\left[\prod_{r=1}^{R} \theta_{\zeta_r}\right] \tag{2.28}$$

Proof. Begin with the individual mixed membership distribution, conditional on θ_i.

$$F(x|\theta_i) = \prod_r \sum_k \theta_{ik} F_{kr}(x_r) \tag{2.29}$$

$$= \sum_{\zeta \in \mathcal{Z}} \prod_r \theta_{i\zeta_r} F_{\zeta_r r}(x_r) \tag{2.30}$$

Equation 2.30 reindexes the terms of the finite sum when Equation 2.29 is expanded. Distributing the product over r yields Equation 2.31.

$$F(x|\theta_i) = \sum_{\zeta \in \mathcal{Z}} \left(\left[\prod_r \theta_{i\zeta_r} \right] \left[\prod_r F_{\zeta_r r}(x_r) \right] \right) \tag{2.31}$$

$$= \sum_{\zeta \in \mathcal{Z}} \pi_{i\zeta} F_\zeta(x) \tag{2.32}$$

Integrating Equation 2.32 yields the form of a finite mixture model.

$$F(x) = \mathbb{E}_\theta \left[\sum_{\zeta \in \mathcal{Z}} \pi_{i\zeta} F_\zeta(x) \right] = \sum_{\zeta \in \mathcal{Z}} \pi_\zeta F_\zeta(x) \tag{2.33}$$

□

Corollary 2.4. *Let $\zeta, \zeta' \in \{1, \ldots, K\}^R = \mathcal{Z}$, and F be a set of MMM profiles. If ζ' is a permutation of ζ, then the probability associated with the FMM mixture classes F_ζ and $F_{\zeta'}$ is equal. That is $\pi_\zeta = \pi_{\zeta'}$.*

Proof. From Theorem 2.3, equation 2.31 and equation 2.33,

$$\pi_\zeta = \mathbb{E}\left[\prod_{r=1}^R \theta_{\zeta_r} \right]$$

We simply observe that if ζ' is a permutation of ζ, then

$$\prod_{r=1}^R \theta_{\zeta_r} = \prod_{r=1}^R \theta_{\zeta'_r}$$

□

Theorem 2.3 says that if a mixed membership model needs K profiles to express the diversity in the population, an equivalent finite mixture model will require K^J components. In equation 1.2, we introduced indicator vectors Z_{ij}, let us rewrite them slightly so that $Z_{ij} = k$ if individual i acts as a member of profile k on variable j. Then the FMM class indicator for an individual i is $\zeta_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iJ})$. ζ indicates exactly which profile an individual followed on each and every variable.

The mixed membership model is a much more efficient representation for high dimensional data; however, there is a tradeoff in the constraints on the shape of the data distribution. The equivalent FMM is highly constrained in the shape of the components allowed. Each FMM class F_ζ is a product of components of the profiles across different dimensions F_{kr}. Illustration 2 demonstrates the relationship between the two representations of a mixed membership model, and how the MMM profiles F_{kr} form a “basis" for the FMM components F_ζ.

Corollary 2.4 goes farther, indicating that not only are the shapes of the classes in the FMM constrained, but there are constraints on the class probabilities, so that certain sets of classes will have the same probabilities. From a generative perspective, any data generated by a particular MMM will satisfy the constraints given by Corollary 2.4 when the model is expressed in FMM form. These constraints will play an important part in determining whether an MMM is identifiable.

Illustration 2

Suppose we have a MMM with three profiles, and two variables with no replication ($K = 3; \;\; J = R = 2$), . To make this more concrete, suppose we are looking at student assessment data where the two features represent the amount of time a student takes to reach a solution on two different items. The three profiles then might, for example, represent three possible different solution strategies. Theorem 2.3 says that we could also write this as a FMM with $|\mathcal{Z}| = K^R = 9$ classes.

To highlight how the FMM classes are formed from the MMM profile distributions, I will write each of the two-dimensional profiles as $F_k = F_{k1} \otimes F_{k2}$. Thus $F_{1,2}$ is the distribution for the amount of time required using strategy $k = 1$ on the problem $j = 2$. We suppose that the distribution of time for each strategy on each item is normal, and that strategy 1 takes about 3 minutes per item, strategy 2 takes about 7 minutes per item, and strategy 3 takes about 11 minutes per item. We'll also assume that strategy 3 has a different standard deviation for the two items.

$$\begin{aligned} F_{k=1} &= N(3,1) \otimes N(3,1) \\ F_{k=2} &= N\left(7,0.25\right) \otimes N\left(7,0.25\right) \\ F_{k=3} &= N\left(11,0.5\right) \otimes N(11,1) \end{aligned}$$

If an individual used strategy 3 on the first problem, and strategy 1 on the second problem, then this would be represented by the FMM component $\zeta = \{3,1\}$, and the student's data point would be located in $F_{\zeta=\{3,1\}} \sim N\left(11,0.5\right) \otimes N(3,1)$. Likewise, if a student used strategy 2 on the first item and strategy 3 on the second item, then this is represented by the FMM component $\zeta = \{2,3\}$. The index set $\zeta \in \mathcal{Z}$ represents all the possible combinations of different strategies that could be used on different problems.

The three MMM profiles appear on the left side of Figure 2.3, while the right side shows all 9 FMM classes. The FMM classes are completely determined by the profiles F_k, for $\zeta \in \mathcal{Z}$, we have

$$F_\zeta = F_{\zeta_1,1} \otimes F_{\zeta_2,2} \tag{2.34}$$

In this way the MMM profiles form a “basis" for the data distribution.

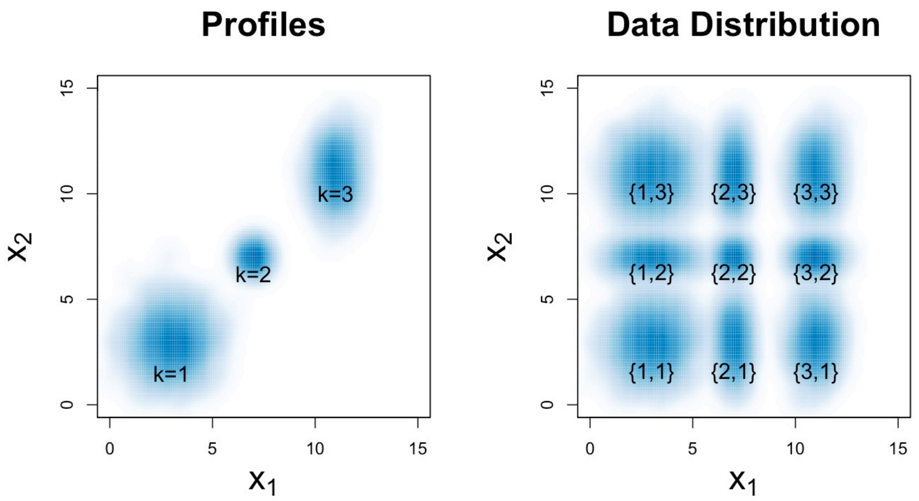

Figure 2.3: Illustration 2. Mixed membership model with three profiles and two features ($K = 3$, $J = R = 2$). The densities of the profiles for the features x_1 and x_2 are shown on the left hand side and labeled with the profile index k. The resulting FMM classes on the right side are labeled with the index $\zeta \in \mathcal{Z}$.

Now consider a particular individual who uses strategies 1 and 3 equally, so that their membership parameter is $\theta_i = (0.5, 0, 0.5)$. This individual’s response time distribution on the first item is then

$$F(x_1|\theta_i) = (0.5)N(3,1) + (0.5)N(11,0.5)$$

It is worth noting that this individual’s mixed membership in each strategy does not make it likely they will have a response time between 3 and 11 minutes, but rather that they will have a response time either near 3 minutes or near 11 minutes. This is, of course, a well-known property of normal mixture models, but it is worth reiterating because of the *between* interpretation possible in other circumstances.

For this individual, since they use strategies 1 and 3 equally, any combination of these two strategies is equally likely. In particular, it is equally likely that they use strategy 1 on the first item and strategy 3 on the second item, $\zeta = \{1,3\}$; as that they use strategy 3 on the first item and strategy 1 on the second item, $\zeta = \{3,1\}$. This is the symmetry under permutation of ζ guaranteed by Corollary 2.4. It is always the case that $\pi_{\{1,3\}} = \pi_{\{3,1\}}$. Thus the corresponding FMM components always have the same probability.

The MMM 'basis' profiles determine the FMM mixture components, as evident in equation 2.27 of Theorem 2.3, and this illustration. A change in the value or distribution of the membership parameter θ cannot change the mixture classes when the model is expressed in FMM form. Changing the distribution of θ only affects the probability associated with each mixture class $\pi_\zeta = \mathbb{E}\left[\prod_r \theta_{\zeta_r}\right]$.

This illustration shows how the small set of mixed membership profiles are recombined by the *switching* behavior to form the larger set of mixture classes. □

2.4 IDENTIFIABILITY

Identifiability is an issue that is often ignored when working with complex hierarchical models. However, if a mixed membership model is not identifiable, we may draw incorrect conclusions about the population based on the basis profiles that are estimated.

Let us consider a highly abstracted version of the applications to the National Long Term Care Survey (Erosheva et al., 2007; Manrique-Vallier, 2010). Suppose that we fit a MMM and find one profile that indicates a high level of mental impairment, but low levels of impairment to mobility. Suppose that another profile indicates the opposite: a low level of mental impairment, but severe mobility restrictions. This might lead us to conclude that mental and physical impairment are relatively independent and severe impairment in both areas rarely occurs.

Alternatively, if we find a profile that is associated with very mild levels of both mental and physical disability, and another profile is associated with high levels of mental and physical disability. This set of profiles might lead us to conclude that physical and mental degeneration are linked.

These are drastically different conclusions, but we cannot know whether either is justified until we understand the conditions under which a mixed membership model is identifiable. Theorem 2.3 shows that a small number of basis profiles can generate a much larger number of classes when written as a finite mixture model. This leads us to wonder whether distinct sets of profiles can generate the same set of finite mixture components.

From Illustration 2, we see that the mixed membership profiles form a sort of 'basis' for the mixture components, and just from looking at Figure 2.3 we might propose an alternate set of basis profiles. Theorem 2.5 shows that we can indeed construct such an alternate set of basis profiles by selecting other mixture components in such a way that they 'span' the set of all mixture components. We note, however, that for the marginal data distribution to be the same, not only must the FMM components be the same, the probabilities for each component must also be equal, so Theorem 2.5 alone is insufficient for identifiability.

Theorem 2.5. *Let* F *and* G *be two sets of* K *MMM profiles over* J *unique variables. Assume that for each feature* $j \in \{1, \dots, J\}$*, the list of profile distributions* $F_{1j}, \dots, F_{Kj}$ *and* $G_{1j}, \dots, G_{Kj}$*, is the same up to permutation. That is, there is a one-to-one and onto mapping from* k *to* k' *such that* $F_{kj} = G_{k'j}$*.*

Then when the distribution is written as a finite mixture model with components indexed by $\zeta \in \mathcal{Z}$, F *and* G *will generate the same set of mixture components.*

Moreover, there are $(K!)^{(J-1)}$ *distinct sets of basis profiles with the same set of mixture model components.*

Proof. Let F_ζ be a component of the mixture model generated by F, then from Theorem 2.3,

$$F_\zeta(x) = \prod_{r=1}^{R} F_{\zeta_r r}(x_r) \tag{2.35}$$

But by assumption for each $\zeta_r \in \{1, ..., K\}$, there exists a $\zeta'_r \in \{1, ..., K\}$ such that $F_{\zeta_r r} = G_{\zeta'_r r}$ so that

$$F_\zeta(x) \;=\; \prod_{r=1}^{R} F_{\zeta_r r}(x_r) \;=\; \prod_{r=1}^{R} G_{\zeta'_r r}(x_r) \;=\; G_{\zeta'}(x) \tag{2.36}$$

Now, we need to count the number of possible distinct sets of basis profiles which generate the same mixture components. First, to avoid double counting, fix the distribution of the first variable for each basis profile. That is, fix $F_{k,j=1}$ for $k = 1, \ldots, K$.

Then, for the first basis profile, there are K possible choices of $F_{k=1,j}$ for $j = 2, \ldots, J$, and thus $K^{(J-1)}$ distinct ways to construct $F_{k=1}$.

For the k^{th} basis profile, there are now $K - (k-1)$ possible choices of $F_{k,j}$ for $j = 2, \ldots, J$, so there are $(K-(k-1))^{(J-1)}$ ways to construct F_k.

Thus the number of possible ways to construct all K components is:

$$\prod_{k=1}^{K} (K-(k-1))^{(J-1)} \;=\; (K!)^{(J-1)} \tag{2.37}$$

□

Illustration 3

Continuing from Illustration 2, we will use the same context of student response times and the same set of basis profiles, F:

$$\begin{aligned} F_{k=1} &= N(3,1) \otimes N(3,1) \\ F_{k=2} &= N(7,0.25) \otimes N(7,0.25) \\ F_{k=3} &= N(11,0.5) \otimes N(11,1) \end{aligned}$$

We can think of constructing an alternate set of basis profiles $\{G_k\}$ as permuting which distributions for the second problem are paired with the distributions for the first problem. For example, we might switch $F_{2,2}$ and $F_{3,2}$ to get the basis profiles:

$$\begin{aligned} G_{k=1} &= N(3,1) \otimes N(3,1) \\ G_{k=2} &= N(7,0.25) \otimes N(11,1) \\ G_{k=3} &= N(11,0.5) \otimes N(7,0.25) \end{aligned}$$

Clearly the interpretation of these two different sets of profiles would be different. Under the F profiles, strategy F_2 is always faster than strategy F_3; however, under the G profiles strategy G_2 is sometimes faster and sometimes slower than strategy G_3. Theorem 2.5 indicates that despite the different interpretation, these two sets of basis profiles will result in the same finite mixture model components. Figure 2.4 shows the profiles F and G, and the FMM components associated with each set of profiles.

In this example, there are $(K!)^{(J-1)} = (3!)^1 = 6$ possible sets of basis profiles with the same mixture components in finite mixture model form. The six possible sets are shown in Figure 2.5. □

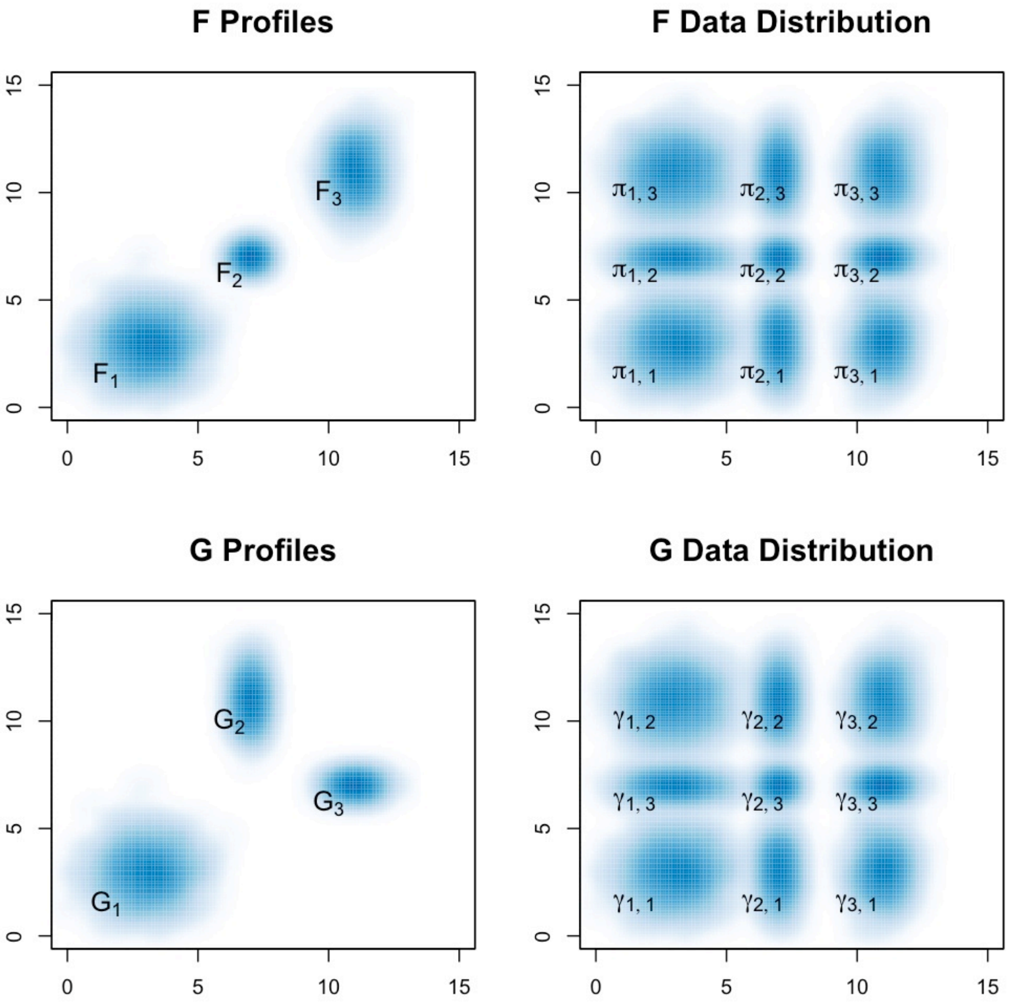

Figure 2.4: Illustrations 3 and 4. The left shows the two sets of basis profiles, F and G. The right shows the finite mixture model with components labeled by their probabilities under each distribution. Note that different components share labels under each set of basis profiles.

The six sets of basis profiles in Illustration 3 and Figure 2.5 generate exactly the same set of mixture components. However, an FMM is defined by both the component distributions F_ζ and the associated probabilities π_ζ. Theorem 2.5 gives the conditions under which different specifications of the basis profiles will result in the same FMM components. The permutation constraints on the probabilities

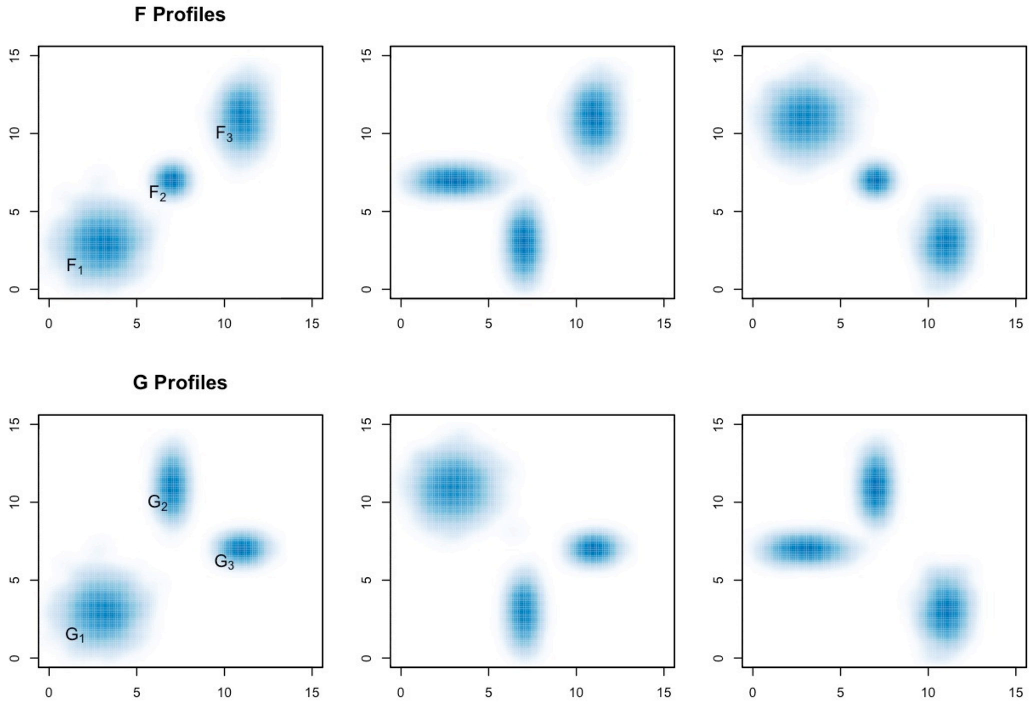

Figure 2.5: Illustration 3. The original basis profiles F are shown in the upper left corner. There are a total of $(K!)^{(J-1)} = 6$ basis profiles which generate the same finite mixture classes. Each box shows one of these sets of basis profiles. The basis profiles G are shown on the lower left.

π_ζ given by Corollary 2.4 lead to Theorem 2.6. Theorem 2.6 says that the probabilities of components will not be equal unless multiple sets of constraints are satisfied simultaneously. Thus, even if the FMM components are the same, the FMM distributions will not be equal unless additional requirements are met.

Theorem 2.7 is the main identifiability result, and characterizes the class of MMM distributions which will generate exactly the same FMM distribution. Essentially, the MMM basis profiles must generate the same FMM components, as in Theorem 2.5, and the distribution of the membership parameter θ must guarantee that the FMM components under each specification of the model have the same probability. The more symmetry there is in the distribution of θ, or equivalently, the more dimensions of θ which are exchangeable, then the larger the class of MMM distributions which generate the same marginal data distribution becomes.

Theorem 2.6. *Let* F *and* G *be two distinct sets of mixed membership profiles which generate the same set of finite mixture model components (as in Theorem 2.5). When the model is expressed in finite mixture model form, the component probabilities,* π_ζ *are subject to equality constraints described in Corollary 2.4. These sets of equality constraints on* π_ζ *under* F *and* G *are distinct.*

Proof. Since F and G are distinct sets of basis profiles, F has at least one basis profile that is not in G, say F_{k^*}. When the data distribution for F is written in finite mixture form, the mixture component F_{ζ^*} where $\zeta^* = \{k^*, k^*, \dots, k^*\}$ is equal to F_{k^*}, and it has probability $\pi_* = \mathbb{E}\left[\prod_r \theta_{k^*}\right] = \mathbb{E}\left[\theta_{k^*}^R\right]$. Note that the set of permutations of ζ^* has only one element, namely ζ^*, so that under F, the probability π_* for mixture component F_{ζ^*} is not constrained.

Since F and G generate the same set of mixture components, there exists a $t \in \mathcal{Z}$ such that $F_{\zeta^*} = G_t$. Under G, the probability of the mixture component G_t is $\gamma_t = \mathbb{E}\left[\prod_r \theta_{t_r r}\right]$, which is constrained so that $\gamma_t = \gamma_{t'}$ for all permuations t' of t.

However, F_{k^*} is not an basis profile in G, so that G_t is not equal to any basis profiles in G, and thus the set of permutations of t has more than one element. Thus the two sets of basis profiles impose distinct sets of constraints on the probabilities of the finite mixture distribution. □

Theorem 2.7. *Fix the distribution of the membership vector,* $\theta = (\theta_1, \dots, \theta_K)$*, and suppose that for some maximal subset* $A \subseteq \{1, \dots, K\}$*, the* θ_a *are exchangeable for* $a \in A$ *. Then two sets of profiles,* F *and* G *will be associated with the same marginal data distribu-*

tion if and only if $G_k = F_k$ *for* $k \notin A$*, and* $G_{kj} = F_{k'j}$ *for* $k, k' \in A$*. (That is* G *permutes the exchangeable dimensions of* F*, otherwise* G *is the same as* F*.)*

There will be $(|A|!)^{J-1}$ *sets of MMM basis profiles which define the same mixed membership distribution.*

Proof. If F and G satisfy $G_k = F_k$ for $k \notin A$, and $G_{kj} = F_{k'j}$ for $k, k' \in A$, then they satisfy the conditions in Theorem 2.5 to generate the same FMM components. Additionally, if F and G have the same marginal data distribution, then they must generate the same FMM components. Thus G must permute the dimensions of F according to Theorem 2.5.

We simply need show that the component probabilities, π_ζ, are equal if and only if G is a permutation of F for only the exchangeable dimensions of θ.

When expressed as a FMM, we have a marginal data distribution of

$$F^{FMM}(x) = \sum_{\zeta \in \mathcal{Z}} \pi_\zeta F_\zeta(x) \tag{2.38}$$

where

$$\pi_\zeta = \mathbb{E}\left[\prod_{r=1}^{R} \theta_{\zeta_r}\right] \tag{2.39}$$

For F and G to have the same marginal data distribution, we must have that

$$\mathbb{E}\left[\prod_{j=1}^{K} \theta_{\zeta_j}\right] = \mathbb{E}\left[\prod_{j=1}^{K} \theta_{\zeta'_j}\right] \tag{2.40}$$

This holds if and only if ζ' is formed from ζ by permuting the elements ζ_a where θ_a is an exchangeable dimension of θ.

The counting is the same as in the proof of Theorem 2.5. □

These identifiability results follow directly from the exchangeability assumption in Equation 1.3. The assumption that features are independent conditional on the latent membership parameter results in both the basis structure identified in Theorem 2.5 and the probability constraints defined in Corollary 2.4. Theorem 7 tells us that if some dimensions of the membership parameter θ are exchangeable, then there are multiple ways to specify the basis profiles which will result in the same data distribution and satisfy the same constraints on class probabilities.

We can now see that the identifiability of a mixed-membership model depends on the exchangeability structure of the membership parameter. If there is no subset of θ_a which are exchangeable, then there is only one set of basis profiles possible which will generate that marginal data distribution. In this case, the MMM is identifiable. On the other hand, if θ is fully exchangeable, as is the case when a symmetric Dirichlet prior is placed on the membership parameter, then there are $(K!)^{J-1}$ equivalent sets of basis profiles. This idea is illustrated in Figure 2.6.

Theorem 2.7 defines an equivalence class of MMM which all have the same marginal data distribution. An MMM is identifiable up to a defined set of permutations that switch distributions for variables from one profile to another profile. This is somewhat similar to how a finite mixture model is identifiable up to permutation (Titterington et al., 1985), yet in the FMM case it is simply a permutation of the indices. For MMM, it is not a simple re-indexing, it is recombining the profiles to form a completely different set of profiles.

Manton et al. (1994) does consider the issue of identifiability for the GoM model. However, these results are flawed because they do not consider the possibility of distinct sets of mixed membership profiles producing the same model for data. The results in Manton et al. (1994) only apply when the profile distributions are multinomial and the equivalence class defined by Theorem 2.7 has a single member.

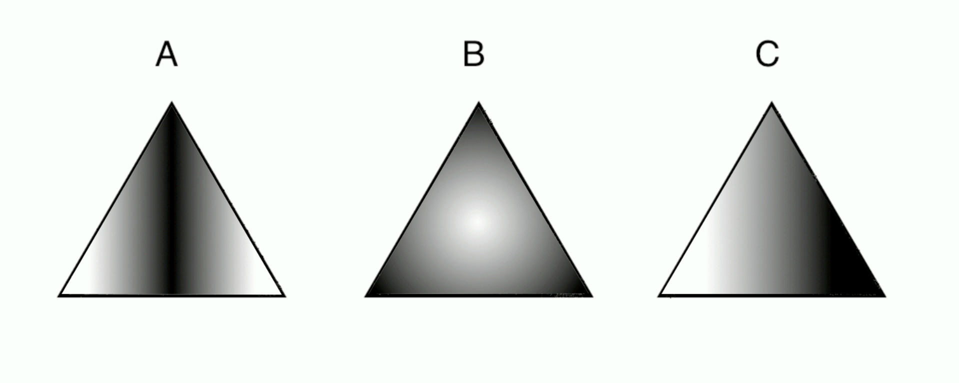

Figure 2.6: Each triangle represents a possible distribution of the membership parameter θ in a MMM with $K = 3$. Distribution A has 2-fold symmetry, and there are $2!^{(J-1)}$ equivalent MMM distributions. Distribution B has complete symmetry and all possible $K!^{(J-1)}$ MMM distributions will have the same distribution of observable data. Distribution C has no symmetry, and the associated MMM will be identifiable.

Illustration 4

Continuing from Illustrations 2 and 3, we have two sets of MMM basis profiles, F and G which generate the same set of FMM mixture components. These two sets of profiles represent two different possible summaries of strategies that students might use in a problem solving context. In order for F and G to represent the same distribution of solution times in data, they must not only generate the same set of FMM mixture components, but the must also generate the same probabilities for each component.

That is, if

$$F^{FMM}(x) = \sum_{\zeta \in Z} \pi_\zeta F_\zeta(x)$$

and

$$G^{FMM}(x) = \sum_{\zeta \in \mathcal{Z}} \gamma_\zeta G_\zeta(x)$$

Then for F^{FMM} to be equal to G^{FMM} we must have that for all $\zeta \in \mathcal{Z}$ there is a ζ' such that $\pi_\zeta = \gamma_{\zeta'}$ and $F_\zeta = G_{\zeta'}$.

Theorem 2.6 says that the constraints on π_ζ and γ_ζ are different. We can think of it this way: If the F strategies are accurate, then a student is equally likely to use strategy F_1 followed by F_2 as to use strategy F_2 followed by F_1. So the student's data point is equally likely to be in the FMM component $N(3,1) \otimes N(7,0.25)$ as in the component $N(7,0.25) \otimes N(3,1)$. This is the equality constraint that $\pi_{\{1,2\}} = \pi_{\{2,1\}}$.

The G basis profiles define strategy 2 differently. If the G strategies are accurate, then a student is still equally likely to use strategy G_1 then G_2 as to use G_2 and then G_1, according to the equality constraint that $\gamma_{\{1,2\}} = \gamma_{\{2,1\}}$. However, since strategy G_2 is different than strategy F_2, now the student's data point is equally likely to be in the FMM component $N(3,1) \otimes N(11,1)$ as in the component $N(7,0.25) \otimes N(3,1)$.

Recall that π and γ are determined completely by the distribution of the membership parameter θ. In order to have $F^{FMM} = G^{FMM}$, the distribution of θ must be one that can simultaneously satisfy the equality constraints on π and those on γ.

In this example, G permutes dimensions of the basis profiles F_2 and F_3. Thus, according to Theorem 2.7, if θ_2 and θ_3 are exchangeable, then the mixed membership model defined by G and the model defined by F will have the same data distribution. If θ_2 and θ_3 are not exchangeable, then F and G define distinct mixed membership distributions. □

There are a few special cases to consider. First, in LDA $J = 1$ with many replications; the only feature being modeled in LDA is the presence or absence of words, but each document has many words, so there are many replications. In this case there is $K!^{J-1} = K!^{0} = 1$ distinct set of basis profiles which will generate the equivalent mixture model components. Thus latent Dirichlet allocation is an identifiable model.

Another special case to consider is the issue of distinct sets of exchangeable dimensions of θ. Suppose that θ_a are exchangeable for $a \in A$ and θ_b are exchangeable for $b \in B$. In this case, there are

$$\left(|A|!^{(J-1)}\right)\left(|B|!^{(J-1)}\right) \tag{2.41}$$

members of the equivalence class with the same marginal data distribution.

A final special case is of particular importance for the myriad extensions of LDA which model more than one feature ($J > 1$). The symmetric Dirichlet distribution is often used as a convenient prior distribution for the membership parameter, θ. For any symmetric distribution of the membership parameter, all $K!^{(J-1)}$ possible basis profiles will have exactly the same marginal data distribution. In this case, the class of equivalent MMM distributions is at its maximal size. This practice may lead to problems with non-identifiability such as different MCMC runs resulting in different, but equivalent estimations of the basis profiles. If we place a strong symmetric prior on θ, then we may estimate the basis profiles to be any one of the $K!^{(J-1)}$ possibilities, because under that prior, they are all equivalent.

2.5 DISCUSSION AND IMPLICATIONS FOR PRACTICE

The results in this chapter address two different facets of mixed membership models. First is the special case of categorical data, addressed by theorems 2.1 and 2.2. Second is the identifiability of the class of general models, described in theorems 2.3-2.7.

Theorems 2.1 and 2.2 show that mixed membership distributions have a different interpretation for categorical data than in the general case. This is a key finding since all of the early papers on mixed membership worked with categorical data exclusively. We demonstrated that the individual-mixture structure in the mixed membership model in the general case may be interpreted as describing a *switching* behavior. Individuals with partial membership sometimes follow one profile and sometimes follow another profile.

For categorical data, and a small set of other cases, an additional interpretation is possible. Partial membership in multiple profiles in this special set of cases may also be interpreted as individuals residing *between* profiles. These two interpretations take place in two different spaces. The *between* interpretation for categorical data is in the parameter space. The *switching* interpretation is in the data space.

This difference in interpretations should be one of the things we consider when deciding if a mixed membership model is appropriate for a given application. For example, suppose that we are again modeling student performance on an assessment. Suppose two profiles represent an immature (bad) strategy and a mature (good) strategy, and mixed membership represents the degree to which a student has moved from the immature strategy to the mature strategy. If the data is categorical, mixed membership is an appropriate and reasonable model whether we believe that students who are learning the mature strategy use some strategy between the two extremes, or if we believe that students switch back and forth between strategies while learning the mature strategy. Either interpretation is possible for categorical data.

If the data are not categorical, but is continuous or discrete, then only the *switching* interpretation is possible. Mixed membership will model students switching back and forth between the immature and mature strategies, but not using some strategy between the two extremes.

Now consider the examples of Rogers et al. (2005) and Blei and Jordan (2003). Rogers et al. (2005) models gene expression levels, and Blei and Jordan (2003) models the content of an image. Both models use continuous data, and both use multivariate Gaussians as basis profile distributions for some features, $F_{kj} = N(\mu_k, \sigma_k)$. Again, the between interpretation is not possible here, only the switching interpretation; but the switching interpretation is fully appropriate. In Rogers et al. (2005), genes switch their expression level based on which processes the tissue sample associated with. In Blei and Jordan (2003), their GM-LDA model describes some image segments being associated with some subjects and other image segments associated with other subjects. This is again a switching behavior, the individual image switches between subjects for different segments of the image. The key here is that in the general case, mixed membership is an appropriate model only if partial membership is reasonably interpreted as switching back and forth between profiles.

Theorems 2.3-2.7 focus on identifiability of mixed membership models. These results are also critical for evaluating the appropriateness of a mixed membership model and interpreting estimation results. First, we showed in Theorem 2.3 that every mixed membership model can be re-written as a constrained finite mixture model. This raises the question of whether mixed membership is worth the extra trouble if we can simply use an equivalent simpler model. The answer is that a finite mixture model is not always simpler. Mixed membership can represent a document being about both healthcare and small-businesses without creating a separate healthcare/small-business topic. The mixed membership model needs only K profiles and the membership parameter θ to describe the diversity in the population. The finite mixture model needs K^J classes to describe the same diversity. The advantage for mixed-membership comes as the number of features increases and data become very high dimensional.

The parsimony in representations comes with some tradeoffs. One well-known tradeoff is that mixed membership models can be difficult to estimate (Blei et al., 2003; Shan and Banerjee, 2011). The probability constraints in Corollary 2.4 are another such tradeoff. We can summarize the data with a small number of profiles at the cost of making assumptions about exchangeability between features. In the image-analysis context we assume that it is equally likely that segment 1 at the top of the image is sky and segment 2 at the bottom of the image is rock, as that the upper segment is rock and the lower segment is sky. In this example, this clearly false assumption probably doesn't make too much difference and the model may perform acceptably regardless. In other applications, we may need to examine this assumption more carefully.

Finally, the main identifiability result in Theorem 2.7 describes classes of mixed membership distributions which all have the same data distribution. If the distribution of the membership vector θ has any symmetry, or equivalently, if any dimensions of θ are exchangeable, then the class of equivalent models has more than one member. Illustration 4 provided a very simple and easy to identify example of this phenomenon. Identifying the equivalence class in practice with high dimensional data will be more tricky.

A large equivalence class might cause different MCMC runs to result in different estimated profiles. When this occurs, it seems prudent to look for equivalent mixed membership representations. Another possibility is to examine the posterior distribution of θ for exchangeable dimensions. Identifying symmetry in the posterior distribution of θ is a good first step toward identifying members of the equivalence class of mixed membership distributions.

A more thorny issue with respect to these equivalence classes is how to interpret them. Any interpretation of the estimated profiles should be applicable to all of the possible profiles in the equivalence class. In our illustration, with the F profiles, strategy F_1 was faster than strategy F_2 which was faster than F_3. This lends itself to a nice interpretation. The alternate G profiles did not have such a nice interpretation. However, if θ_2 and θ_3 were exchangeable, then F and G generated identical data distributions. How we interpret this may depend in large part on the context of the application.

In some situations, we may be able to argue that even though F and G are mathematically equivalent, F is a more reasonable summary because we know that the problems are very similar and the results from each strategy on each problem should be similar. Alternatively, we may decide that G is a more reasonable summary because sometimes strategy G_2 is faster and sometimes strategy G_3 is faster. We must realize, however, that this is an invocation of external knowledge and the two representations are equivalent (for this supposed distribution of θ). It is more honest in this case to report the estimated profiles, the distribution of θ, and note that profiles 2 and 3 are exchangeable.

A final comment on terminology. First, the mixed membership profiles are sometimes referred to as ‘extreme profiles’ due to the properties of mixed membership for categorical data. Since the profiles do not form extremal points of the parameter space in the general case, the term ‘basis profiles’ or simply ‘profiles’ are more appropriate general terms. Additionally, latent Dirichlet allocation is a poor name for the general model for two reasons. First, the distribution of the membership parameter does not need to be Dirichlet. Second, when the distribution of the membership parameter is a symmetric Dirichlet distribution, the class of equivalent mixed membership distributions is at its maximum size. Mixed membership is a more appropriate name for the general class of models.

3

MODELING MULTIPLE STRATEGIES

It seems intuitive that for most problems, there are multiple ways to solve them. Colloquially, these multiple ways to solve a problem are what we refer to as “strategies." There is ample evidence that for all types of problems, individuals use multiple distinct strategies. For example, in mental rotation tasks, we know that different individuals use different strategies, but we also suspect that individuals switch strategies from item to item (Geiser et al., 2006). In another setting, we have solid evidence that children switch strategies on even the simplest arithmetic and spelling problems (Siegler, 1987; Rittle-Johnson and Siegler, 1999).

Not only are different strategies present, the strategies themselves are indicators of expertise. Experts have knowledge that is organized around fundamental principles of a discipline, and they are able to retrieve this knowledge flexibly (Bransford et al., 2000). Experts know more strategies and use different strategies than novices (Lovett, 1998; Pellegrino et al., 2001; Quaiser-Pohl et al., 2010). Since strategies differ by expertise, assessing which strategies individuals are using is an important part of assessing their expertise.

One compounding factor for estimating expertise through strategy use, is that while on the whole, strategies do become more effective with age and experience, growth is not straightforward. Rather it seems that as children learn, the mixture of strategies that each child uses changes; they use one strategy more and another strategy less, but still use both strategies (Pellegrino et al., 2001). In order to estimate expertise from strategies, we have to estimate the proportion of the time each individual uses each strategy. This is the key which leads us to mixed membership for modeling strategies.

This chapter uses foundations in the learning sciences and cognitive psychology to refine the concept of 'strategy.' I then develop a probability model for assessing individual strategy usage by combining mixed membership models with cognitive diagnosis models. I then demonstrate how additional variables such as response time can be incorporated into this model of strategies.

3.1 EXAMPLE OF MULTIPLE STRATEGY USE IN ADDITION

To better understand what modeling multiple strategies entails, let us consider a specific example. Siegler (1987) examines the strategies that young children use in addition, and is one of the first papers to highlight the fact that children switch strategies, even for very simple single-digit addition problems. Five strategies were considered: retrieval, min-counting, count-all, decomposition, and guessing.

The retrieval strategy is the fastest and most accurate if it is available; if $2 + 2$ is memorized, then the student does not need to count. In the min-counting strategy the student begins counting on the bigger number. For example, to solve $3 + 5$, the student counts *five, six, seven, eight.* When a student uses the count-all strategy, they solve $3 + 5$ by counting *one, two, three,. . . , four, five, six, seven, eight*. Decomposition involves transforming the problem into simpler problems.

For example; to solve $12 + 2$; a child may say “12 is 10 and 2; 2 and 2 is 4; 10 and 4 is 14; so 14." Decomposition may be considered a more expert strategy because it relies on a greater conceptual understanding of addition.

The five strategies are separated by both solution time and percentage of errors. Retrieval was the fastest strategy, followed by decomposition and guessing. Min-counting took twice as long on average as retrieval and led to more than twice as many errors. Count-all was used almost exclusively by kindergardeners, was 3 times longer than retrieval and more than twice as long as min-counting. Count-all was also incredibly inaccurate, with an over-all error rate of 54%. Decomposition is worth particular note, because while it was only used about 10% of the time by first and second graders, it is the second fastest strategy and approximately as accurate as retrieval.

The evidence that children switch strategies from problem to problem is overwhelming. 99% of children reported using at least 2 strategies and 62% reported using 3 or more. In kindergarden, retrieval, min-counting, count-all, and guessing were each used at least once by more than 68% of children. In first and second grade, a majority of children used retrieval, min-counting and decomposition at least once. A reasonable conclusion is that by first or second grade, the students knew all 5 strategies but had abandoned the slow and inaccurate strategies.

The key observation here is that even for these very simple addition problems, children switch strategies, and strategy choice is clearly related to expertise. Moreover, the differences in expertise are expressed by differences in the proportion of time that the students use each strategy. Siegler (1987) argues that we should abandon the question “What is *the* strategy that young children use to add?" in favor of the question “Under what conditions do children most often use each of their strategies?" This is the same shift that we now need to make in modeling student knowledge.

3.2 modeling strategies

To consider how strategies are currently modeled in the psychometric literature, it is useful to turn to the categorization created in Junker (1999) and espoused in Pellegrino et al. (2001).

Case 0: No modeling of strategies

Case 1: Model strategy changes between persons

Case 2: Model strategy changes between tasks, within persons

Case 3: Model strategy changes within task, within persons

Most psychometric models are Case 0 models. An item response theory (IRT) model that estimates student ability compared to the difficulty of the questions on the assessment is an example of Case 0 (Junker, 1999). Cognitive diagnosis models (CDM) are also Case 0 models (Henson et al., 2009). CDMs assume that a person can solve a problem if they possess a certain set of skills, but the set of required skills is identical for every individual. That is, the model assumes that everyone solves the problem in the same way, with the same strategy.

Mixture IRT models (Mislevy and Verhelst, 1990) are an example of Case 1. In a mixture IRT model, students are partitioned into latent classes according to which strategy they use, and the responses for all students in a given class is represented by a standard IRT model. Case 1 models assume that different individuals use different strategies, but that each individual uses a single strategy throughout the assessment, thus most Case 1 models are variations of latent class models.

Case 2 and Case 3 models are much more difficult to estimate, because we have to estimate when and how much each individual is using each strategy. As Junker (1999) notes, binary or polytomously scored responses may not contain enough information to identify strategies, and accordingly very little work has been done in modeling Cases 2 and 3. With computer-based assessment becoming more prevalent, it is much easier to collect additional data, and estimate which combination of strategies individuals are using.

Modeling strategy differences between persons, Case 1, is a useful model if individuals predominantly use one strategy. However, if individuals switch strategies between tasks, then using a Case 1 model functions as aggregating the strategies for each individual. Siegler (1987) demonstrates that not only do people switch strategies on even very simple tasks, but also that aggregating strategies can lead to incorrect conclusions about which strategies are present.

In order to capture the differences in expertise that are expressed by different strategy use, we need to be working in Case 2. We need to model individuals switching strategies. In fact, the choice of which strategy to use on which problem is, itself, an indicator of expertise. As established in Chapter 2 mixed membership models are built to describe a switching behavior. This makes mixed membership an ideal foundation for modeling Case 2, where individuals switch strategies between items.

3.3 THE KNOWLEDGE-LEARNING-INSTRUCTION FRAMEWORK

“Strategy" is well defined in common English usage. A common dictionary definition is “a plan of action designed to achieve a specific goal." We need to refine this concept of strategy in a way that is both consistent with common usage and precise enough for the purposes of assessment. For this, we turn to the Knowledge-Learning-Instruction (KLI) framework (Koedinger et al., 2010).

The KLI framework provides the groundwork to connect cognitive learning research to instructional principles at different grain-sizes from neurons to the classroom. It builds a taxonomy of kinds of knowledge focusing on the characteristics that may determine which learning processes are likely to be most effective in producing different kinds of knowledge.

For example, the studying techniques that are effective to memorize facts are not the same techniques that are most effective to learn mathematical problem solving methods. The taxonomy created in the KLI framework focuses on the distinctions that are likely to tell us why the learning process is different in each case. This taxonomy allows us to create a more precise definition of 'strategy.'

The basic unit of knowledge in the KLI framework is dubbed a *knowledge component* (KC) and defined as 'an acquired unit of cognitive function or structure that can be inferred from performance on a set of related tasks.' Defined as such, this term encompasses skills, strategies, concepts, principles, facts, misconceptions, schemas, and production rules. Knowledge components are also distinguished by the time scale on which they occur, taking 1 second to 1 minute.

This framework exists to help researchers study learning. Therefore it is useful to define KCs at the level of students who are in a particular learning environment (such as a course). For example, college students can read regular and irregular verbs with very high accuracy. First or second grade students in comparison, may be able to read regular verbs, but have difficulty with irregular verbs. At the lower level, we need at least two KCs, possibly more since there are different types of irregular verbs, but at the higher level, only one KC is necessary. We unpack knowledge components into smaller components until the smaller components are ones that students can perform with a sufficiently high accuracy and fluency.

The KLI framework decomposes the learning progression into a temporal sequence of events: *instructional events, learning events*, and *assessment events*. Instructional events include any event intended to produce learning: a classroom lesson, a practice session on a computer, or a museum exhibit. Instructional events are observable, as are assessment events. Exams, or particular exam items are the most common assessment events, but formative assessments that occur during instruction are also assessment events. Learning events however, are unobservable. Learning must be inferred from assessment data. The goal of an assessment within this framework is to identify whether a student has acquired a particular set of knowledge components.

Knowledge components represent the conditionalized nature of knowledge by relating the task features and context to the individual's response. An example KC might be: *To get ready to play baseball, get your bat and glove*. However, like learning, KCs are unobservable, and should be inferred from student performance on an appropriate set of tasks. KCs are expressed in a condition-response format to emphasize that a critical part of knowing is knowing when to apply knowledge.

The taxonomy of KCs takes place in four dimensions: Task Features, Response, Relationship, Rationale.Task Features and Responses may be either constant or variable. For example, *To state meaning of the Spanish word rojo, say red*. This simple KC has a constant task feature, and a constant response. The relationship between the features and response may be either implicit or explicit, and the rationale for the relationship may be known or unknown. This yields a possible 16 types of knowledge components, Koedinger et al. leave the question open whether all 16 types exist in practice. Table 3.1 below is a reproduction of Table 3 from Koedinger et al. (2010) showing common knowledge component categories with common labels for each type of KC.

Table 3.1: Common knowledge component categories (Table 3 from Koedinger et al. (2010))

Task Features	Response	Relationship	Rationale	Labels
constant	constant	implicit	no	association
constant	constant	explicit	no	fact
variable	constant	implicit	no	category
variable	constant	explicit	no	concept
variable	variable	implicit	no	production, schema, skill
variable	variable	explicit	no	rule, plan
variable	variable	explicit	yes	principle, rule, model

In addition to these four dimensions, KCs also vary substantially in complexity. Complexity may come from many sources, such as the amount of perceptual encoding required, or the complexity of any motor response. Of interest for us, is that some KCs must be integrated with other KCs to produce behavior, these are referred to as *integrative knowledge components*.

3.4 definition of 'strategy'

We can now define *strategy* as a type of integrative knowledge component that guides how the other KCs should be combined in order to complete a task. For example, the white area in Figure 3.1 can be computed in at least 2 ways. You can compute the area for the large rectangle, and subtract off the area for the black rectangle. Alternatively, you can decompose the white space into 2 smaller rectangles and add the areas together. The strategy chosen dictates which atomic KCs are necessary to solve the problem.

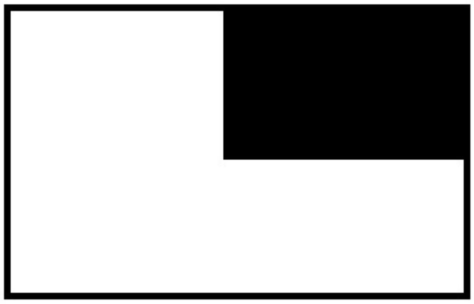

Figure 3.1: To find the area of the white region, students can use numerous different strategies.

Defining strategy as an integrative knowledge component translates nicely to the psychometric literature when compared to conjunctive and disjunctive models. In conjunctive, or noncompensatory models, students need all of the attributes associated with a problem in order to solve the problem. In contrast, disjunctive models assume that if students have at least one of the necessary attributes, they will be able to solve the problem. (See for example Junker and Sijtsma (2001); Henson et al. (2009).)

Junker (1999) introduces the idea that we may think of strategies as disjunctive attributes. As long as a student knows one strategy for solving a problem, then provided the student has the necessary atomic KCs to carry out the strategy, they will be able to solve it. The KCs required to complete the solution are conjunctive, the student must have all of them to carry out the strategy. These conjunctive KCs can be appropriately termed skills. These definitions are consistent with common English usage as well. A student can use only one strategy, one plan of attack, but may require several skills to carry out that plan.

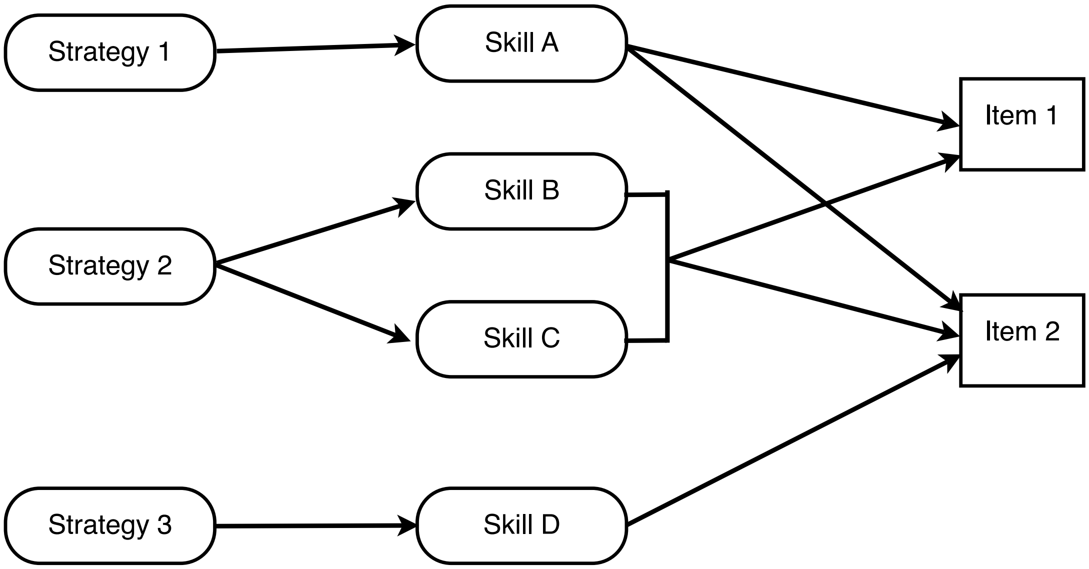

Figure 3.2: Illustration of the relationship between strategies and items in a multiple-strategies model. Strategy 1 may be thought of as an expert strategy, only one skill is required to solve both items. Strategy 2 is less efficient, and requires both Skills B and C, but it is still an effective method for solving both items. Strategy 3, on the other hand, relies on a misconception, and will not generate a correct response to Item 1

Figure 3.2 shows an example of the relationship between strategies, skills and items in a multiple-strategies model. What is missing from Figure 3.2 is a characterization of how students relate to strategies. If students use only one strategy, then we have some flavor of latent class model, but if students switch strategies then we need to construct a new mixed-membership model that allows for this behavior.

3.5 COMBINING MIXED MEMBERSHIP WITH COGNITIVE DIAGNOSIS MODELS

We have now defined what is meant by 'strategy'. We know that students use different strategies, that strategy choice is correlated with expertise, and that students switch strategies from problem to problem. Having defined the problem, we now turn to the question of how to model strategies in assessment data.

In a mixed membership model (MMM), individuals have partial membership in different profiles. If the different profiles can represent different strategies, then partial membership in a profile will correspond to how much an individual uses that strategy. There is a subtle point here. Partial membership is not how well a student knows a particular strategy, but rather how much the student uses the strategy. To represent the individual strategies, we use cognitive diagnosis models (Henson et al., 2009). Later in this chapter, we will demonstrate how to incorporate response time data, and other variables such as self-reported strategy.

Cognitive diagnosis models (CDMs) measure students' expertise by estimating whether they have mastered a pre-defined set of knowledge components. It is more common in the psychometric literature to talk about CDMs measuring 'skills' or 'attributes', but I use the term knowledge component (KC) since it is both more well defined than attribute and more general than skill. A wide variety of CDMs exist in the literature, which are gathered into a single family of models by Henson et al. (2009). I use the general notation of Henson et al. (2009) so that it is clear that the incorporation of CDMs with MMMs is not dependent upon the particular CDM used to represent the individual strategies.

The probability that a student can correctly solve a problem is conditional on the KC's that the item requires and the KCs that the student has acquired. There are many ways that CDMs express this probability. Disjunctive models assume that if a student possesses at least one of the required KCs, then they have a high probability of solving the item. Compensatory models assume that if a student is particularly skilled in one attribute, then that can make up for a lack of another attribute. Conjunctive models require that a student possess all of the required KCs to have a high probability of a correct response.

We have defined strategies as disjunctive KCs that require conjunctive skill KCs to complete a solution. Each strategy is essentially determined by the skills required to carry out the strategy; that is, a strategy is defined by a conjunctive model. We use a separate conjunctive CDM for each profile distribution. Individuals can then switch between the profiles from problem to problem, modeling the strategy switching, but once they have chosen a strategy for a particular item, the CDM associated with that strategy determines the skills required.

In a CDM, the probability that a particular student i can correctly answer a particular item j depends on the KCs that the student possesses and the KCs that the item requires. The KCs that student i possesses are captured by the vector α_i. $\alpha_{is} = 1$ if student i possesses the KC s, and is 0 otherwise. The KCs associated with each item are specified by the matrix Q. The matrix entry $q_{js} = 1$ if skill s is associated with item j, and 0 otherwise. Specifying the Q matrix is equivalent to defining the strategy that this CDM represents.

Henson et al. (2009) gives a common log-linear formulation for the larger family of CDMs:

$$\log\left(\frac{P(X_{ij}=1|\alpha_i)}{1-P(X_{ij}=1|\alpha_i)}\right) = \lambda_j^T h(\alpha_i, q_j) - \eta_j. \tag{3.1}$$

or equivalently:

$$P(X_{ij}=1|\alpha_i) = \frac{\exp\{\lambda_j^T h(\alpha_i, q_j) - \eta_j\}}{1+\exp\{\lambda_j^T h(\alpha_i, q_j) - \eta_j\}}, \tag{3.2}$$

The vector-valued function $h(\alpha_i, q_j)$ determines whether the model is conjunctive, disjunctive or compensatory. Each component of h is a linear combination of α_i and q_j. λ_j is a vector of weights for each of the components of h. The value η_j essentially defines a guessing probability when a student has not mastered any skills required by the item.

We create a multiple strategy model by using one CDM to represent each strategy. We can then use the distinct CDMs as the basis profiles in a mixed membership model. This allows a student i to switch between using different strategies k on different problems j, with the membership parameter θ_i governing how much a student uses each strategy.

Recall that the mixed membership model is defined by

$$F(x|\theta_i) = \prod_j^J \sum_k^K \theta_{ik} F_{kj}(x_j). \tag{3.3}$$

Each MMM basis profile is now a distinct CDM: $F_{kj}(x_j) = P(X_{ij}=1|\alpha_i)$, so that we create a mixed membership cognitive diagnosis model:

$$F(x|\theta_i, \alpha_i) = \prod_j^J \sum_k^K \theta_{ik} \left[\frac{\exp\{\lambda_j^T h(\alpha_i, q_j) - \eta_j\}}{1+\exp\{\lambda_j^T h(\alpha_i, q_j) - \eta_j\}}\right] \tag{3.4}$$

This formulation models a student's mastery of both simple skill knowledge components and integrative strategy knowledge components. θ represents how much the student uses each strategy KC while α represents whether the student has mastered each skill KC. One of the challenges with the full model is estimating both θ and α. Indeed, Junker (1999) suggests that additional information beyond binary or polytomous response data may be necessary to estimate models with multiple strategies. The next section builds a framework for incorporating this additional information.

3.6 response time and other variables

With computer assessment data, response times are easily recorded. In experimental data, researchers often have recorded observed strategies or self-reported strategies. These variables and others are not perfectly predictive of a student's strategy, but they are highly correlated. The multiple strategy model is easily adapted to incorporate additional variables. In this section, we focus on including response times as an illustration of a general way to extend the mixed membership cognitive diagnosis model into a general multiple strategies model.

It is well known that response time and expertise are correlated (van der Linden, 2009). Attempts at including some measure of time into estimates of ability date back to Thurstone (1937), but the question of how to use response data for this purpose is still unresolved, in part due to several issues with how time and expertise are related.

The first issue is an accuracy-time tradeoff. For an individual attempting a single task, they have a choice between going more slowly and doing the task well, or going more quickly and allowing for a higher risk of an error. This within-person tradeoff can only be seen by observing the same individual under different conditions which require more accuracy or faster performance. There is some evidence that during the course of an assessment, students do not alter their speed very much from problem to problem, so that this well-known psychological tradeoff is not observed in assessment data (van der Linden, 2009).

The second issue is that as individuals practice, they get faster (Anderson, 2010; Koedinger et al., 2010). In order to model an accuracy-time tradeoff, you must assume that no learning is taking place. As a student's expertise increases, both accuracy and speed increase, so that any tradeoff between the two changes over time.

This brings up the third issue. Over a population, speed and accuracy are correlated. However, the correlation is sometimes negative and sometimes positive. Van der Linden (2009) argues that this is simply because the higher-ability students have better time management skills, and know when to speed up and slow down. This time-management idea is naive. Time and accuracy depend on strategy.

Siegler (1987) demonstrates that response times vary substantially by strategy, though not in a linear fashion. The most expert strategies were by far the fastest. The most rudimentary strategy of all, guessing, was only a little slower. The slowest strategies were the two novice strategies.

In other task domains, the expert strategies are slower than novice strategies, often because they require more steps. The variable correlation between speed and accuracy depends on whether the expert strategies are slower or faster than the novice strategies. The key though, is that we can model both the responses and response times conditionally on strategy choice.

3.6.1 *Existing models for Responses and Time*

Van der Linden (2009) separates attempts to model speed and accuracy into a couple of categories: distinct models for time and accuracy, incorporating time into a model of ability, incorporating accuracy into a model for speed, and jointly modeling speed and accuracy. Rouder et al. (2003) is an example of the first case. It is a sophisticated model for response times that accounts for differences between items and differences between individuals. An example of the second sort is Thurstone (1937), which incorporates time into a model of ability, and indeed, many more recent models, such as Roskam (1997), are similar to this early model.

Van Breukelen (2005) offers one of the first joint models, where both time and responses are considered random variables and attempts to estimate the correlation between them. Van der Linden (2007) introduces another joint model which has spawned a number of variations, including Loeys et al. (2011) and Entink et al. (2009).

We can think of all of these models as belonging to a larger class of models, inspired by van der Linden (2007). Let C_{ij} indicate whether individual i correctly responded to item j, let T_{ij} be the associated response time, and let $X_{ij} = (C_{ij}, T_{ij})$. The joint distribution is a hierarchical model which accounts for correlation between individual speed and ability. The distribution for C_{ij} depends on item difficulty parameters β_{jc} and individual ability parameters ϕ_{ic}. Similarly, the response time distribution for T_{ij} depends on item intensity parameters β_{jt}, and individual speed parameters ϕ_{it}. The individual parameters are summarized as $\phi_i = (\phi_{ic}, \phi_{it})$, and the item parameters as $\beta_j = (\beta_{jc}, \beta_{jt})$.

In an extension of the usual local independence assumption we treat C_{ij} and T_{ij} as conditionally independent given the item parameters β_j and the individual parameters ϕ_i;

$$F(X_{ij}|\beta_j, \phi_i) = F(C_{ij}|\beta_{jc}, \phi_{ic}) \times F(T_{ij}|\beta_{jt}, \phi_{it}) \tag{3.5}$$

so that,

$$\begin{aligned} F(X|\beta, \theta) &= \prod_j \prod_i \left[F(C_{ij}|\beta_{jc}, \theta_{ic}) \times F(T_{ij}|\beta_{jt}, \theta_{it})\right] && (3.6) \\ &= \left[\prod_j \prod_i F(C_{ij}|\beta_{jc}, \theta_{ic})\right] \left[\prod_j \prod_i F(T_{ij}|\beta_{jt}, \theta_{it})\right] && (3.7) \\ &= F(C|\beta_c, \theta_c) \times F(T|\beta_t, \theta_t) && (3.8) \end{aligned}$$

The second layer of the hierarchical model is what captures the relationship between speed and accuracy and makes this a joint model.

$$\begin{aligned} \theta_i &\sim N(\mu_\theta, \Sigma_\theta) && (3.9) \\ \beta_j &\sim N(\mu_\beta, \Sigma_\beta) && (3.10) \end{aligned}$$

Thus Σ_θ captures the correlation between speed and ability, while Σ_β captures the correlation between item difficulty and item time intensity.

One of the reasons that van der Linden (2007) has inspired several variations is that it is very easy to make changes to the distributions for time and accuracy. The original model proposed used a three-parameter normal-ogive model for $F(C_{ij}|\beta_{jc}, \theta_{ic})$, but the hierarchical structure of the model makes it very simple to substitute another response model, for example, a one-parameter logistic model.

This class of models is useful for detecting test design flaws such as bad test items or ambiguous instructions, analyzing the 'speededness' of the test, and detecting aberrant student behavior. Our goal is to estimate which strategies students are using, and this class of models does not address the issue of multiple strategies (Case 0).

3.6.2 *Time, Accuracy and Strategy*

In Siegler (1987), some strategies were fast, others were up to 3 times slower. Each strategy had a distinct distribution of response times, and a distinct distribution of responses. This is the key observation in using response times to estimate strategies.

As above, let X_{ij} be all the variables collected for student i on item j. When we have only responses and response time, $X_{ij} = (C_{ij}, T_{ij})$. In some applications, we may observe additional variables, such as eye-tracking data, self-reported strategy, or specific intermediate steps. Call these additional variables collectively W_{ij}, so that $X_{ij} = (C_{ij}, T_{ij}, W_{ij})$. Consistent with the rest of this paper, let the mixed membership profiles be indexed $k = 1, \dots, K$, and let each profile represent a different strategy.

A strategy is characterized by a distribution for the vector X_{ij}. This distribution, F_{kj}, is the process-signature for strategy k. It is convenient to assume that the variables are conditionally-independent given the strategy chosen, so that F_k factors. Whether or not a student can carry out strategy k depends on whether they have the required knowledge components, as indicated by α_i. Thus, for a particular item j, the distribution of responses for individuals who used strategy k is assumed to follow:

$$F_{kj}(x_j|\alpha) = F_{kjc}(c_j|\alpha) \times F_{kjt}(t_j|\alpha) \times F_{kjw}(w_j|\alpha) \tag{3.11}$$

Of course, the model is far simpler if we assume that time and other variables do not depend on the KC-skill vector α:

$$F_{kj}(x_j|\alpha) = F_{kjc}(c_j|\alpha) \times F_{kjt}(t_j) \times F_{kjw}(w_j) \tag{3.12}$$

3.7 THE MULTIPLE STRATEGIES MODEL

Section 3.5 developed the idea of combining mixed membership models with cognitive diagnosis models to model students switching strategies during an assessment, using a distinct CDM to represent the response pattern of each strategy. Section 3.6 discussed response time and illustrated how time and other variables can be incorporated into the mixed membership basis profiles that represent each strategy. From here, we can now specify a multiple strategies model that considers how much a student uses each strategy to be an integral part of estimating student expertise.

The data point X_{ij} is the collection of all variables collected for student i on item j. In particular, we consider $X_{ij} = (C_{ij}, T_{ij}, W_{ij})$; where C_{ij} indicates whether student i answered item j correctly, T_{ij} is the response time, and W_{ij} represents any other observed data. Each strategy k is represented by a mixed membership basis profile F_k,

$$F_{kj}(x_j|\alpha) = F_{kjc}(c_j|\alpha) \times F_{kjt}(t_j) \times F_{kjw}(w_j) \tag{3.13}$$

The profile distributions F_k capture the process-signatures for each strategy. We can distinguish two strategies within this model only if they produce different distributions for the observed data X.

At the individual level, the membership parameter θ_i indicates how much individual i uses each strategy, and the KC-skill vector α_i indicates whether the individual has mastered each of the skills necessary to complete each strategy. In the mixed membership model, the distribution of an individual with membership parameter θ_i and KC-skill vector α_i is:

$$\begin{aligned} F(x|\theta_i, \alpha_i) &= \prod_j \left[\sum_k \theta_{ik} F_{kj}(x_j|\alpha_i) \right] && (3.14) \\ &= \prod_j \left[\sum_k \theta_{ik} F_{kjc}(c_j|\alpha_i) F_{kjt}(t_j) F_{kjw}(w_j) \right] && (3.15) \end{aligned}$$

Note that since the sum over k is inside the product over j, the variables C, T, and W are not conditionally independent given θ. Rather, when we write the model in finite mixture model form (Theorem 2.3), then we have an independence relationship conditional on the particular strategies used on each problem.

$$\begin{aligned} F(x|\theta_i, \alpha_i) &= \prod_j \left[\sum_k \theta_{ik} F_{kj}(x_j|\alpha_i) \right] && (3.16) \\ &= \sum_{\zeta \in \mathcal{Z}} \pi_{i\zeta} F_{\zeta}(x|\alpha_i) && (3.17) \\ &= \sum_{\zeta \in \mathcal{Z}} \pi_{i\zeta} F_{\zeta c}(c_j|\alpha_i) F_{\zeta t}(t_j) F_{\zeta w}(w_j) && (3.18) \end{aligned}$$

If we know which strategy individual i used on each problem j, that is the same as knowing which FMM class ζ the individual belongs to. Denote this class with an indicator vector z_i. Equation 3.18 now becomes:

$$F(x|z_i, \alpha_i) = \prod_{\zeta \in \mathcal{Z}} \left[\pi_{i\zeta} F_{\zeta c}(c_j|\alpha_i) F_{\zeta t}(t_j) F_{\zeta w}(w_j) \right]^{z_{i\zeta}} \qquad (3.19)$$

Equation 3.19 demonstrates clearly that C_i and T_i are locally independent given z_i, not θ_i. What's the difference? θ contains information on how much a child uses a certain strategy. On the other hand, z_i contains information about which strategy was used on which problem. $\zeta \in \{1, \ldots, K\}^J$, and $\zeta_{ij} = k$ if individual i used strategy k on item j. z_i is a data augmentation vector that specifically states which strategy a child used on each item. C_{ij} and T_{ij} are independent only if the strategy used by student i on item j is known.

From a cognitive perspective, the response itself and the response latency are outcomes of internal processing (Wenger, 2005). Moreover, the influences on this processing can run all the way from perceptual encoding to motor output, so that the relationship between time and accuracy is not fixed between tasks. By conditioning on the strategy choice for each item, we are conditioning on this cognitive processing event.

The flexible framework provided by the mixed membership model allows us to relate observed variables to each other through the strategies without necessitating specific distributional choices. For example, in one application we may know that response time distributions follow a 3 parameter Weibull distribution (Rouder et al., 2003; Rouder, 2005), in which case we can use that distribution for F_{kjt}. If, on the other hand, the response times follow an exponential or log-normal distribution, then we can adjust F_{kjt} accordingly. In the same way, any member of the CDM family can be used for the the response distribution F_{kjc}; though, since we have defined strategies as disjunctive KCs and skills as conjunctive KCs, we will use conjunctive CDMs. Relating observed variables to each other through solution strategies provides a flexible framework for altering the model to include any other available variables, which we have denoted W.

Both accuracy and latency are random variables that are due to unobserved cognitive processes. The multiple strategies model allows us to condition on the process at the appropriate cognitive grain-size of a knowledge component.

3.8 comments

Assessing expertise by modeling multiple strategy usage is not just building a more complicated model to estimate student “ability." Pellegrino et al. (2001) claims that ″The measurement models in use today include some very sophisticated options, but they have had surprisingly little impact on the everyday practice of educational assessment. The problem lies not so much with the range of measurement models available, but with the outdated conceptions of learning and observation that underlie most widely used assessments."

This model addresses this concern. It represents an effort to capture the dimensions along which cognitive science has shown that experts and novices differ. Almost everyone switches strategies from problem to problem, but experts and novices differ in the mixture of strategies that they use. Experts use efficient strategies more often, but may occasionally fall back on more rudimentary strategies. This mixed membership model is built to capture these individual differences in the mixtures of strategies used.

The concern however is that this model is too complicated to be useful. If we cannot obtain parameter estimates for reasonably sized data sets, then the model is pointless. Chapter 4 tests this model on a simple data set of very modest size to examine whether estimating this model is feasible and useful for inference.

4

MULTIPLE STRATEGIES IN LEAST COMMON MULTIPLES ASSESSMENT DATA

Mixed membership models are undeniably complicated models. The majority of applications where mixed membership models have been used have exceptionally large data sets. For example, Latent Dirichlet Allocation (Blei et al., 2003) is commonly used to analyze corpus of text containing tens of thousands of documents where each document is hundreds or thousands of words long.

Educational data sets exist on a much smaller scale, with hundreds of students and 10-60 items per student. We need to ask whether it is even possible to estimate multiple strategy usage from data sets that can realistically be collected. The purpose of this chapter is to test a simple version of multiple-strategies mixed membership model.

For this application we assume the number of strategy profiles K, is known. In mixed-membership models in general, determining the appropriate number of profiles for a particular data set is difficult (Erosheva et al., 2004). Future work must address the question of determining the appropriate number of strategies.

Ideally, we would like to fit an unsupervised version of this model, so that from the data, we can recover both the strategy process-signatures and how much each child uses each strategy. This may or may not be a reasonable goal with data sets of this size. We may need to use some prior information about the strategies in order to estimate each child's mixture of strategies. The simulations in Section 4.4 explore the question of how much data is necessary to estimate both the strategies and how much children use them.

These simulations indicate that with 300 students and 15 items per student, if we have prior knowledge of one strategy, then we can estimate the process-signatures of the other strategies as well as the student parameters. With 30 items per student, we can estimate the process-signatures of all the strategies without prior knowledge of any strategies.

4.1 LEAST COMMON MULTIPLES DATA

The data come from a computer-based assessment of Least Common Multiples (Pavlik et al., 2011). Two-hundred fifty-five sixth and seventy grade students participated in the experiment ($N = 255$). Each student answered a randomly selected sample from the $J = 24$ items. Most students answered 16 items, but 58 students only received 8 items.

When students answered items incorrectly, they were allowed to review the correct answer for 18 seconds. This provides the opportunity for students to learn during the assessment, and means that if we view the assessment as a whole, then we should certainly observe multiple strategy usage as students switch to a more correct strategy.

Data include accuracy and two kinds of response times. The data point for student i on item j is the ordered triple $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$.

- C_{ij} is binary: 1 indicates the student correctly answered the item, 0 indicates an incorrect response.
- T_{ij1} is the amount of time in milliseconds a student took before beginning to type their response.
- T_{ij2} is the amount of time in milliseconds that the student took to finish typing.

4.2 multiple strategies model

Two distinct strategies are known to be common for computing Least Common Multiples (LMCs). One is essentially a 'correct' strategy, the other is a misconception that produces a correct solution in some cases. We will also include a third 'unknown' strategy in the model, to allow for the possibility that an additional strategy is present in this data, and to account for any additional variation between students. Thus, we use $K = 3$ strategy profiles in each model and in each simulation.

This application is simpler than the general multiple strategies model defined in Chapter 3 because each strategy is associated with a single skill. Thus if a student uses a particular strategy, we can assume that they know the single required skill. This means that we only need to estimate individual strategy parameters θ_i, but not individual skill parameters α_i. The strategy-skill diagram for this application is shown in Figure 4.1.

Now, since the data are collected in a setting where students have an opportunity to learn, it would be nice to estimate when children learned the correct strategy. To do this would require a longitudinal model where the individual membership parameter θ_i changes over time. Creating such a longitudinal model is beyond the scope of this application, but it is certainly a desirable target for future work. The conditional independence assumption in equation 4.10, which is the same as equation 1.3 in the general mixed membership model, means that a student is just as likely to use a particular strategy on the last item as on the first item.

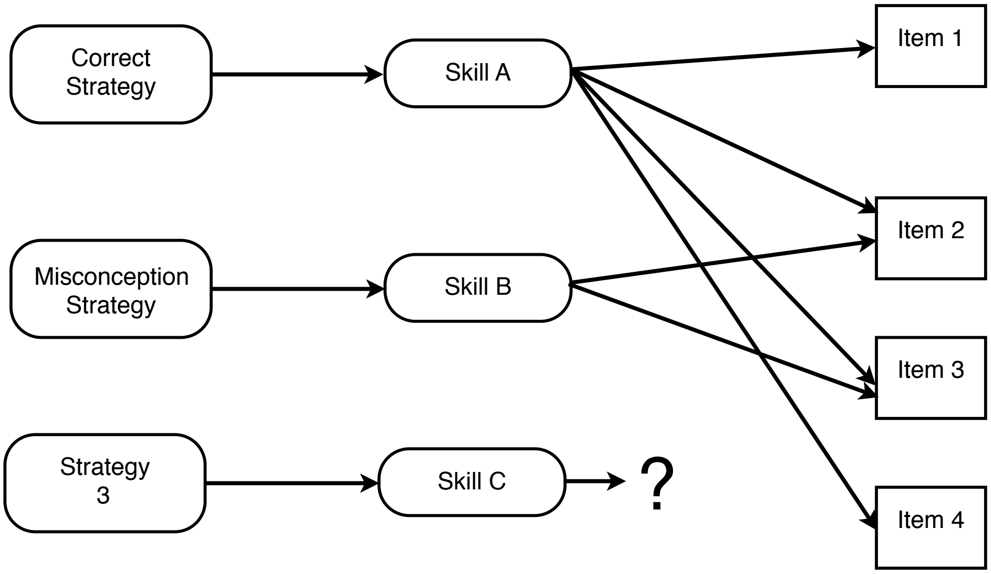

Figure 4.1: The correct strategy is associated with only skill A, and leads to a correct response on all items. The misconception strategy is associated only with skill B, and leads to a correct response on only 2 items out of 4. The third strategy also has only a single skill associated, but it is an unknown strategy, so we do not know when it might lead to a correct response.

4.2.1 *Strategy Profiles*

We assume that there are $K = 3$ different strategies that students might use to to solve the $J = 24$ items.

- Students are indexed $i = 1, \dots, N$.
- Items are indexed $j = 1, \dots, J$.
- Strategy profiles are indexed $k = 1, \dots, K$.

The data point $X_{ij} = (C_{ij},\ T_{ij1},\ T_{ij2})$ includes information on correct responses and two different response times. Thus, the multiple strategies model defined in Chapter 3, gives us strategy profiles of

$$F_{kj}(x_j) = F_{kjc}(c_j) \times F_{kjt_1}(t_{1j}) \times F_{kjt_2}(t_{2j}). \tag{4.1}$$

The distribution of responses, F_{kjc} should be a cognitive diagnosis model, where the probability of a correct response depends on the KCs required by the item and the KCs the student has mastered.

$$F_{kjc}(x_{ij}) = \Pr(x_{ij} = 1) = \frac{\exp\{\lambda_{kj}^T h(\alpha_i, q_{kj}) - \eta_{kj}\}}{1 + \exp\{\lambda_{kj}^T h(\alpha_i, q_{kj}) - \eta_{kj}\}}, \tag{4.2}$$

We assume that each strategy is associated with a single skill. Further, we assume that if a student uses a particular strategy, they possess the the single skill required to execute that strategy, so that $h(\alpha_i, q_{kj})$ is fixed for each profile. Thus, $\Pr(x_{ij} = 1)$ is a constant c_{kj} for each profile k. This is better expressed by writing

$$F_{kjc}(c_j) = \mathtt{Bernoulli}(c_j; \lambda_{kj}). \tag{4.3}$$

Since items have different difficulties, and the difficulty varies with choice of strategy, each strategy k has a distinct probability of a correct response for each item j.

Thus the probability of a correct response is indexed as λ_{kj}. The probability of a correct response depends only on the strategy and the item, there is no additional ability parameter after the student chooses a strategy.

We will use the simplest possible model for the two response times, acknowledging that more sophisticated models, such as the 3 parameter Weibull distribution in Rouder et al. (2003) may refine the results presented here. If a student uses strategy k on a particular item then $T_{ij1} \sim \mathrm{Exp}(\beta_{k1})$ and $T_{ij2} \sim \mathrm{Exp}(\beta_{k2})$. For this data set assessing Least Common Multiples, it is reasonable to assume that the time required to execute a strategy has the same distribution across items. In other words, if a student uses strategy k, the distribution of T_{ij1} is the same for each item j, and similarly for T_{ij2}.

In many other settings, this assumption will not be appropriate, since a particular strategy may be quick on one item and lengthy on another item. In a setting where the profile strategies take different amounts of time for different items, altering the model for $\beta_{kj1}$and β_{kj2} is straightforward, it simply increases the number of parameters in the model. When response time distributions are the same across all items, there are $2K$ parameters for the time distributions. If different items have different response time distributions, there are $2KJ$ parameters.

For a student that uses strategy k on item j, the distribution of X_j is given by

$$F_{kj}(X_j) = \mathrm{Bernoulli}(C_j; \lambda_{kj}) \times \mathrm{Exp}(T_{j1}; \beta_{k1}) \times \mathrm{Exp}(T_{j2}; \beta_{k2}) \tag{4.4}$$

4.2.2 *Priors for Strategy Profile Parameters*

The profile parameters λ and β can be treated as unknown or known. For example, a superficial strategy may work on some items, but not on other items. In this case, we can treat λ_k as known; since if the strategy works, the probability of a correct response is 1, and if the strategy does not work, the probability of a correct response is 0. Note that treating a subset of parameters as known is equivalent to placing a point-mass prior on those parameters. In general, we will not explicitly know λ and β, and will not use a point-mass prior.

Mixed membership models are usually treated as unsupervised models. This provides a strong contrast with CDM models. To estimate a CDM, the association of which skills are required for each item must be known and specified beforehand. In this multiple strategies model, we have the opportunity to 'learn' the strategies from the data.

We do have strong knowledge about two of the strategies that should be present in the data. We expect to find a correct strategy, and a misconception strategy. As I test the multiple strategies model, I will experiment to see how much of this knowledge needs to be incorporated into the priors in order draw reliable inferences from the data.

I use a conjugate Beta prior for λ and a conjugate Gamma prior for β.

$$p(\lambda) = \prod_k \prod_j \mathrm{Beta}(\lambda_{kj}; \gamma) \tag{4.5}$$

$$p(\beta_1) = \prod_k \mathrm{Gamma}(\beta_{k1}; \alpha_1) \tag{4.6}$$

$$p(\beta_2) = \prod_k \mathrm{Gamma}(\beta_{k2}; \alpha_2) \tag{4.7}$$

Since T_{ij2}, the time required to type in an answer, is substantially shorter than the time before a student begins to type, T_{ij1}, the two times require different prior distributions on β.

4.2.3 *Membership in Strategy Profiles*

The amount that each student uses each strategy is parameterized by the non-negative membership vector $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$. Where $\sum_k \theta_{ik} = 1$, so that θ_i lies in the $K-1$ dimensional simplex. We can interpret the component θ_{ik} as the proportion of time which student i uses strategy k. Note that this is different from how much a student *knows* strategy k; θ captures how much a student *uses* each strategy.

For a particular item j, the distribution for student i's response is given by

$$\begin{aligned} X_{ij}|\theta_i &\sim \sum_k \theta_{ik} F_{kj}(X_{ij}) & (4.8) \\ &\sim \sum_k \theta_{ik} \left[\mathrm{Bernoulli}(C_j; \lambda_{kj}) \times \mathrm{Exp}(T_{j1}; \beta_{k1}) \times \mathrm{Exp}(T_{j2}; \beta_{k2})\right] & (4.9) \end{aligned}$$

Items are independent conditional on the student's membership parameter:

$$X_i|\theta_i \sim \prod_j \sum_k \theta_{ik} F_{kj}(X_{ij}) \qquad (4.10)$$

4.2.4 *Distribution of the Membership Parameter*

The membership parameters θ_i resides in the $K-1$ dimensional simplex. There are a couple of common choices for distributions on the simplex, with the most common being the Dirichlet and the Logistic-Normal (Aitchison and Shen, 1980; Aitchison, 1982, 1985).

The most common prior distribution for membership parameters in mixed membership models is the Dirichlet. See, for example: Erosheva (2002); Blei et al. (2003); Erosheva et al. (2004); Airoldi et al. (2008); Manrique-Vallier (2010); Shan and Banerjee (2011). However, the Dirichlet distribution has a strong independence property, where components are independent conditional on summing to 1. This independence is not always appropriate.

In this application, it is reasonable to expect that students who use advanced strategies are less likely to also use immature or inefficient strategies, so that membership in some strategies may be negatively correlated. The Dirichlet distribution is incapable of modeling correlation between membership parameters. Therefore, the Logistic-Normal distribution is a better choice for the prior distribution of θ. This is the same prior used in the Correlated Topic Model (Blei and Lafferty, 2007):

$$\theta_i \sim \mathrm{LogisticNormal}(\mu, \Sigma) \tag{4.11}$$

$$\eta_i = \log\left(\frac{\theta_i}{\theta_{iK}}\right) \tag{4.12}$$

$$\eta_i \sim N(\mu, \Sigma) \tag{4.13}$$

As we discussed in Chapter 2, if θ meets certain partial exchangeability conditions, then a subset of strategy profiles are effectively interchangeable, and there is a class of equivalent models. In other words, if the distribution of θ has any symmetry, then the mixed membership model is not uniquely identifiable. Therefore it is important to estimate μ and Σ, to determine if the distribution of θ meets these conditions.

To complete the distribution for membership parameters, we need a prior for μ and Σ. I chose to use a noninformative Jeffrey's prior:

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2} \tag{4.14}$$

4.2.5 *Data Augmentation*

Let Z_{ij} be a binary vector of length K, where $Z_{ijk} = 1$ if student i uses strategy k on item j, and

$$\Pr(Z_{ijk} = 1 | \theta_i) = \theta_{ik} \tag{4.15}$$

Since Z_{ij} indicates which strategy student i uses on item j, we have that

$$\begin{aligned} X_{ij}|Z_{ij} &\sim \prod_k \left[F_{kj}(X_{ij})\right]^{Z_{ijk}} && (4.16)\\ &\sim \prod_k \left[\mathrm{Bernoulli}(C_{ij};\lambda_{kj}) \times \mathrm{Exp}(T_{ij1};\beta_{k1}) \times \mathrm{Exp}(T_{ij2};\beta_{k2})\right]^{Z_{ijk}} && (4.17)\end{aligned}$$

and items are independent conditional on either θ_i or Z_i so that

$$X_i|Z_i \sim \prod_j \prod_k \left[\lambda_{kj}^{C_{ij}}(1-\lambda_{kj})^{(1-C_{ij})}\right]^{Z_{ijk}} \left[\beta_{k1} e^{-\beta_{k1} T_{ij1}}\right]^{Z_{ijk}} \left[\beta_{k2} e^{-\beta_{k2} T_{ij2}}\right]^{Z_{ijk}} \tag{4.18}$$

This is the same data augmentation technique that was introduced in Erosheva (2002). We can express every mixed membership model as a finite mixture model with K^J classes (Theorem 2.3). The data augmentation variables $Z_i = \{Z_{i1}, \ldots, Z_{iJ}\}$ indicates in which of these K^J classes X_i belongs.

4.2.6 *The Complete Model*

The complete model without data augmentation variables is:

$$p(X, \theta, \lambda, \beta, \mu, \Sigma) = p(X|\theta, \lambda, \beta)p(\theta|\mu, \Sigma)p(\mu, \Sigma)p(\lambda)p(\beta). \tag{4.19}$$

With the addition of the data augmentation variables, we can factor this further.

$$\begin{aligned} p(X, Z, \theta, \lambda, \beta, \mu, \Sigma) &= p(C|Z, \lambda)p(T_1|Z, \beta_1)p(T_2|Z, \beta_2)p(Z|\theta) \times & (4.20)\\ &\qquad p(\theta|\mu, \Sigma)p(\mu, \Sigma)p(\lambda)p(\beta) & (4.21)\\ &= p(\mu, \Sigma)p(\lambda)p(\beta)\prod_{i=1}^{N}[p(C_i|Z_i, \lambda)p(T_{i2}|Z_i, \beta_2) \times & (4.22)\\ &\qquad p(T_{i1}|Z_i, \beta_1)p(Z_i|\theta_i)p(\theta_i|\mu, \Sigma)] & (4.23) \end{aligned}$$

4.3 MCMC ESTIMATION

I used MCMC for estimation. While it is slow, and may not scale, MCMC at least protects from the possible biases of other methods, such as variational approximation. For example, Shan and Banerjee (2011) obtain different results depending on the type of variational approximation that they made. Since the purpose here is to test whether the multiple strategies model can capture strategy switching in real student data, it is desirable to use a more reliable estimation method.

MCMC is also desirable for practical reasons. For comparing several versions of the multiple strategies model, adjusting MCMC is simpler and more straightforward than adjusting variational inference.

4.3.1 *Update for* μ *and* Σ

From equation 4.20, the distribution of μ and Σ given the other parameters is

$$p(\mu, \Sigma|\ldots) \propto p(\theta|\mu, \Sigma)p(\mu, \Sigma). \tag{4.24}$$

For this update we, reparameterize $\theta \sim \mathrm{Logistic-Normal}(\mu, \Sigma)$, as

$$\eta_i = \log\left(\frac{\theta_i}{\theta_{iK}}\right) \tag{4.25}$$

where θ_{iK} is the last component of θ_i. Note that η is effectively of dimension $K-1$, and has the distribution

$$\eta_i \sim N(\mu, \Sigma). \tag{4.26}$$

The prior distribution for μ and Σ is given by

$$p(\mu, \Sigma) \propto |\Sigma|^{-K/2}. \tag{4.27}$$

So that the posterior distribution is

$$\Sigma|\eta \sim \mathrm{Inv-Wishart}_{N-1}(S) \tag{4.28}$$

$$\mu|\Sigma, \eta \sim N\left(\bar{\eta}, \frac{1}{N}\Sigma\right) \tag{4.29}$$

where

$$S = \sum_{i=1}^{N}(\eta_i - \bar{\eta})(\eta_i - \bar{\eta})^T. \tag{4.30}$$

It is worth observing that the posterior distribution for Σ is proper if $N-1 \geqslant K$.

4.3.2 *Update for* θ

The distribution for θ conditional on the other parameters is

$$p(\theta|\ldots) \propto p(Z|\theta)p(\theta|\mu, \Sigma) \tag{4.31}$$

where

$$\begin{aligned} Z_{ij}|\theta_i &\sim \mathrm{Multinomial}(\theta_i, n = 1), && (4.32)\\ \theta_i|\mu, \Sigma &\sim \mathrm{LogisticNormal}(\mu, \Sigma), && (4.33)\\ \eta_i|\mu, \Sigma &= \log\left(\frac{\theta_i}{\theta_{iK}}\right) \sim N(\mu, \Sigma). && (4.34) \end{aligned}$$

The Multinomial and the Logistic-Normal distributions are non-conjugate, so updating θ requires a Metropolis-Hastings step. At iteration b, the proposed point θ_i^* is drawn from the jumping distribution $N(\theta_i^b, \epsilon I)$, where ϵ is a tuning parameter for the MCMC. Setting the tuning parameter to $\epsilon = 0.1$ produces a reasonably efficient algorithm with $K = 3$.

4.3.3 *Update for* Z

The distribution of Z conditional on the other parameters is

$$\begin{aligned} p(Z_i|\ldots) &\propto p(X_i|Z_i, \lambda, \beta)p(Z_i|\theta_i) && (4.35)\\ &\propto \prod_j \prod_k \left[\left(\lambda_{kj}^{C_{ij}}(1-\lambda_{kj})^{(1-C_{ij})}\right)\left(\beta_{k1}e^{-\beta_{k1}T_{ij1}}\right)\left(\beta_{k2}e^{-\beta_{k2}T_{ij2}}\right)\theta_{ik}\right]^{Z_{ijk}} && (4.36) \end{aligned}$$

So that the conditional distribution is $Z_{ij}|\ldots \sim \mathrm{Multinomial}(p_{ij})$, where

$$p_{ijk} = \left(\lambda_{kj}^{C_{ij}}(1-\lambda_{kj})^{(1-C_{ij})}\right)\left(\beta_{k1}e^{-\beta_{k1}T_{ij1}}\right)\left(\beta_{k2}e^{-\beta_{k2}T_{ij2}}\right)\theta_{ik}. \quad (4.37)$$

4.3.4 *Update for* λ

The conditional distribution for λ is

$$p(\lambda|\ldots) \propto p(C|Z, \lambda)p(\lambda). \quad (4.38)$$

where the prior for λ is

$$p(\lambda) = \prod_k \prod_j \mathrm{Beta}(\lambda_{kj}; \gamma_k). \tag{4.39}$$

So that

$$p(\lambda|\ldots) \propto \prod_j \prod_k \left[\lambda_{kj}^{\gamma_{1k}-1}(1-\lambda_{kj})^{\gamma_{2k}-1} \prod_i \left[\lambda_{kj}^{C_{ij}} \left(1-\lambda_{kj}\right)^{1-C_{ij}} \right]^{Z_{ijk}} \right]. \tag{4.40}$$

Thus,

$$p(\lambda_{kj}|\ldots) = \mathrm{Beta}\left(\gamma_{1k} + \sum_i C_{ij} Z_{ijk},\ \gamma_{2k} + \sum_i Z_{ijk}(1-C_{ij}) \right). \tag{4.41}$$

4.3.5 *Update for* β

Recall that the data $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$ include two separate latency times. T_{ij1} is the time before a student begins to type, and is generally a much longer time than T_{ij2}, the time a student takes to finish typing. In this section, I use the index t to represent these two different times.

The profile distributions for T_t are

$$T_{ijt}|\beta_t, Z_{ijk} = 1 \ \sim \ \mathrm{Exp}(\beta_{tk}). \tag{4.42}$$

β_{tk} has a conjugate gamma prior with different parameters for the two times,

$$p(\beta_{tk}) = \mathrm{Gamma}(\alpha_t). \tag{4.43}$$

Thus, the conditional distribution of β_t is

$$\begin{aligned} p(\beta_t|\ldots) &\propto p(\beta_t) p(T_t|Z, \beta_t) && (4.44) \\ &\propto \prod_k \left[{\beta_{tk}}^{\alpha_{t1}-1} e^{-\alpha_{t2}\beta_{tk}} \prod_i \prod_j \left(\beta_{tk} e^{-\beta_{tk} T_{ijt}} \right)^{Z_{ijk}} \right]. && (4.45) \end{aligned}$$

Which leads to the posterior

$$\beta_{tk}|\ldots \sim \mathrm{Gamma}\left(\alpha_{t1} + \sum_i \sum_j Z_{ijk}\,, \quad \alpha_{t2} + \sum_i \sum_j Z_{ijk} T_{ijt}\right). \tag{4.46}$$

4.4 SIMULATIONS AND RESULTS

The least common multiples (LCM) data set has 255 students who each saw 8 or 16 items. We know that two strategies should be present in the data, and want to allow the possibility of a third strategy in the model. The simulations mirror these aspects of the real data.

All simulations use $N = 300$ students and $K = 3$ strategies. The number of items that each student saw in the real data is small, so I considered simulations with both $J = 15$ and $J = 30$ items.

One question I focused on was how much information is necessary for the model to obtain reliable estimates of the strategies, and the distribution of membership parameters. Should we treat the strategies as known, or can we estimate the strategies from the data? The simulations demonstrate that with 30 items per student, we can estimate the strategies, but with only 15 items per student, we need to incorporate some prior information about at least one of the strategies.

Simulations were computationally intensive, taking 5-8 hours for each MCMC chain to run. So I focused on simulations that would provide me with information about parameter-estimation in situations where the model was correct, rather than information about model misfit. Simulations varied in three ways: First, the average number of items each student saw. Second, I varied the types of strategies present in the simulations by varying the generating distribution for the response parameters λ. Finally, I varied the prior information for λ that was used for estimation; from flat priors and informative priors to point mass priors. Table 4.1 summarizes the variations between simulations. One important aspect of model misfit that I did not examine is the case where K is unknown.

Data simulation process:

- number of students $N = 300$
- number of items J, varies by simulation
- number of strategies $K = 3$
- μ and Σ vary with each replication of each simulation.
- for each strategy $k = 1, \ldots, K$, simulate:
  - λ_{kj}, varies by simulation
  - $\beta_{k1} \sim \mathrm{Gamma}(1, 100)$
  - $\beta_{k2} \sim \mathrm{Gamma}(1, 10)$
- for each student $i = 1, \ldots, N$:
  - $\theta_i \sim \mathrm{LogisticNormal}(\mu, \Sigma)$
  - draw the number of items seen, $J_i \sim \mathrm{Poisson}(\gamma)$, where γ varies by simulation.
  - draw a set of items $\mathcal{J}_i$ of size J_i uniformly from $\{1, \ldots, J\}$.
  - for each item $j \in \mathcal{J}_i$
    - $z_{ij} \sim \mathrm{Multinom}(\theta_i, n = 1)$
    - $X_{ij} \sim F_{kj} | z_{ijk} = 1$

Each strategy is defined by the profile distribution

$$F_{kj} = \mathtt{Bernoulli}(C; \lambda_{kj}) \times \mathtt{Exp}(T_1; \beta_{k1}) \times \mathtt{Exp}(T_2; \beta_k 2) \tag{4.47}$$

The different generating distributions for β_{k1} and β_{k2} indicate that the two times T_{ij1} and T_{ij2} are on different scales, but may not differ much between profiles.

Table 4.1: Summary of Simulations

Simulation	Avg. Items per Student	Generative distribution for λ_{kj}	Priors used for estimation
1	30	$\lambda_{kj} \sim \mathtt{Unif}(0,1)$	$\lambda_{kj} \sim \mathtt{Unif}(0,1)$
2	15	$\lambda_{kj} \sim \mathtt{Unif}(0,1)$	$\lambda_{kj} \sim \mathtt{Unif}(0,1)$
3	15	$\lambda_{kj} \sim \mathtt{Unif}(0,1)$	point mass, λ_{kj} known
4.1	15	$\lambda_{1j} \sim \mathtt{Beta}(10,1)$ $\lambda_{2j}, \lambda_{3j} \sim \mathtt{Unif}(0,1)$	$\lambda_{1j} \sim \mathtt{Beta}(10,1)$ $\lambda_{2j}, \lambda_{3j} \sim \mathtt{Unif}(0,1)$
4.2	15	$\lambda_{1j} \sim \mathtt{Beta}(10,1)$ $\lambda_{2j} \sim \mathtt{Bernoulli}(\frac{1}{2})$ $\lambda_{3j} \sim \mathtt{Unif}(0,1)$	$\lambda_{1j} \sim \mathtt{Beta}(10,1)$ $\lambda_{2j} \sim \mathtt{Unif}(0,1)$ $\lambda_{3j} \sim \mathtt{Unif}(0,1)$

The MCMC chains were thinned by saving only every 5th iteration. All of the plots in this section use the thinned chain. In addition, the item parameters β and λ were successfully recovered with appropriately narrow posterior distributions, except in simulation 2 which had very few items per student and used flat priors for λ. The individual parameters θ, and the population-level parameters μ and Σ were much more difficult to estimate. Σ in particular was very difficult to estimate, and took the longest to converge in every case.

4.4.1 *Simulation 1: Average of 30 items per student, flat prior for λ*

This simulation shows that with an average of items per student, we can estimate both the strategies and the distribution of membership parameters with no prior information included. It is especially worth noting that since the strategy response probabilities were simulated as $\lambda_{kj} \sim \mathtt{Unif}(0,1)$, there is no particularly strong distinction between the strategies. Yet with this reasonable number of items per student, we are able to recover both the item and the individual parameters.

This simulation included 60 unique items $J = 60$, but an average of 30 items per student. For each student i, the number of items seen was generated by $\mathtt{Poisson}(30)$, the items were then selected randomly from the 60 items. The profile accuracy parameters were generated by

$$\lambda_{kj} \sim \mathtt{Uniform}(0,1) \tag{4.48}$$

I ran two replications of this simulation, with two different values of μ and Σ. The two resulting distributions of θ are shown in Figures 4.2 and 4.3. Performance was similar in both replications.

Estimation for Simulation 1 I used a flat prior for λ in the MCMC, $\mathtt{Beta}(1,1) = \mathtt{Uniform}(0,1)$. The priors used to estimate the β parameters were the same ones that the data were generated from. These priors provide some information about the scale of the data, but are weak relative to the size of the data set.

From a random start, convergence occurred around 2500 iterations. Accuracy probabilities λ and time distribution parameters β were recovered with high precision and no obvious bias.

Posterior distributions for individual membership parameters θ_i were centered on the simulated value most tightly for individual near the edges of the simplex (Figures 4.4 and 4.6). For individuals with membership parameters closer to the center of the distribution of membership parameters, the posterior of θ_i is more spread out into the population distribution of θ (Figures 4.5 and 4.7).

The parameters μ and Σ which govern the distribution of the membership vector θ were the most unstable and took the longest to converge. They also showed the most auto-correlation. This is expected because μ and Σ are the only two parameters which are updated by a Metropolis-Hastings step. However, the posterior distributions covered the simulated parameters, and more importantly the posterior means $\hat{\mu}$ and $\hat{\Sigma}$ describe a distribution of θ_i very similar to the simulated values (Figures 4.2 and 4.3).

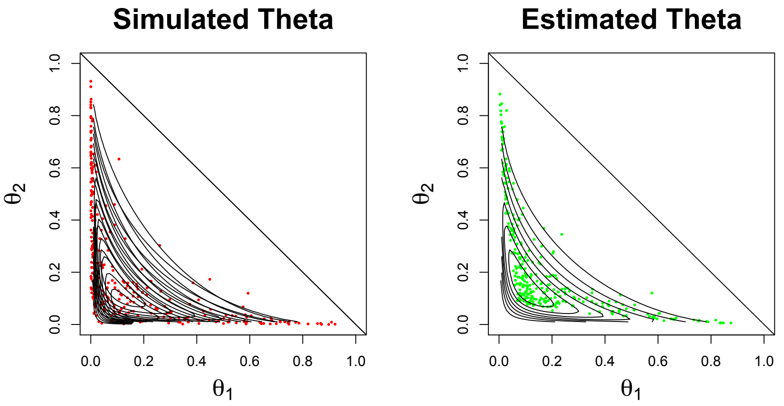

Figure 4.2: Simulation 1, replication 1 (30 items per student). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

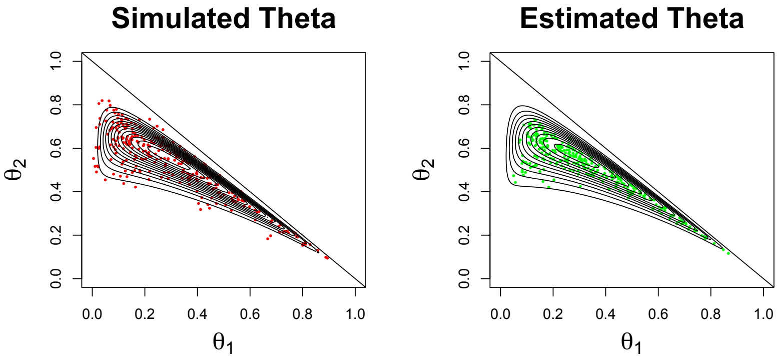

Figure 4.3: Simulation 1, replication 2 (30 items per student). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

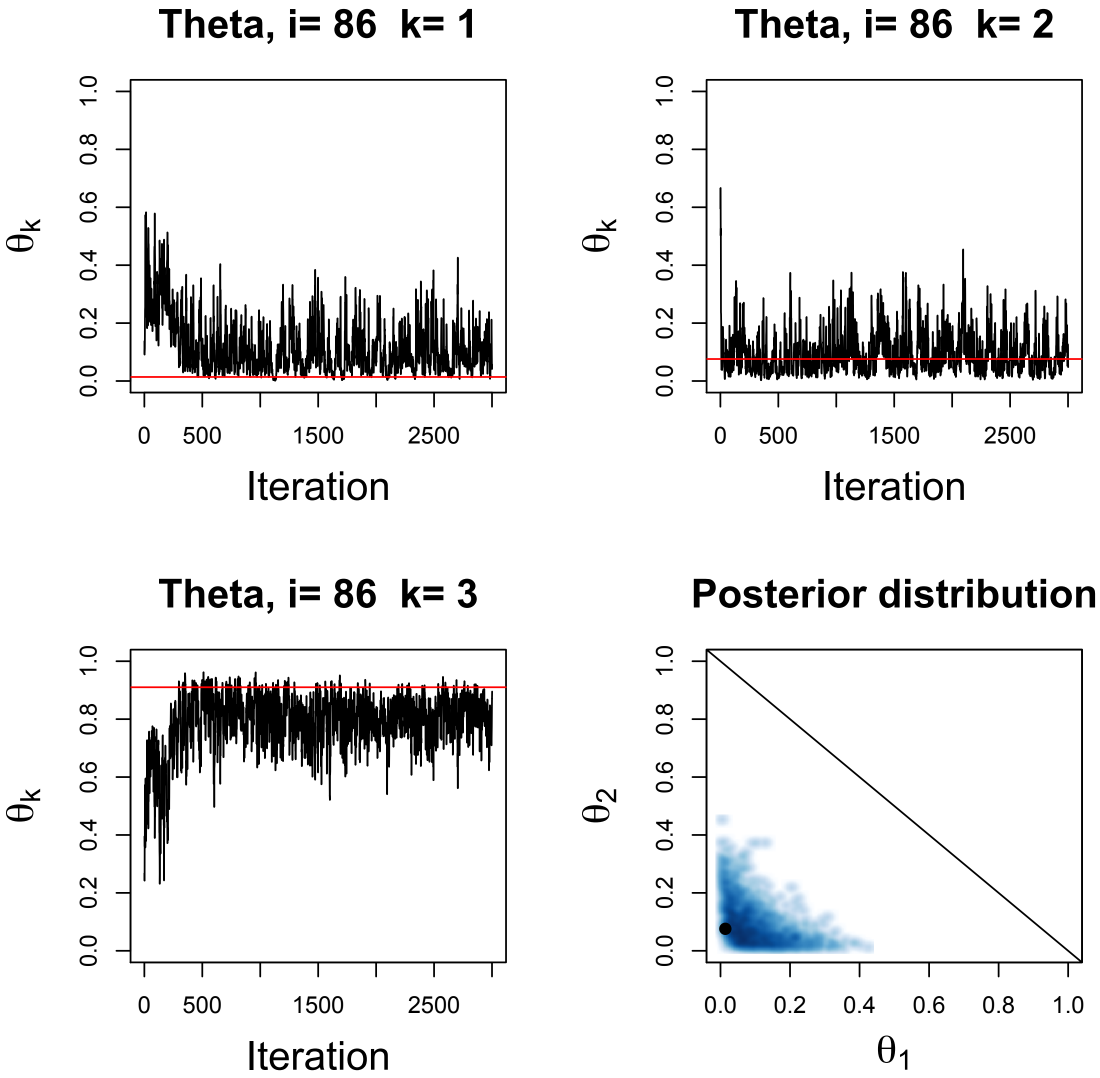

Figure 4.4: Simulation 1, replication 1 (30 items per student). Thinned MCMC chains and posterior distribution for simulated individual $i = 86$. The posterior distribution in the lower right shows the simulated value plotted as a black dot.

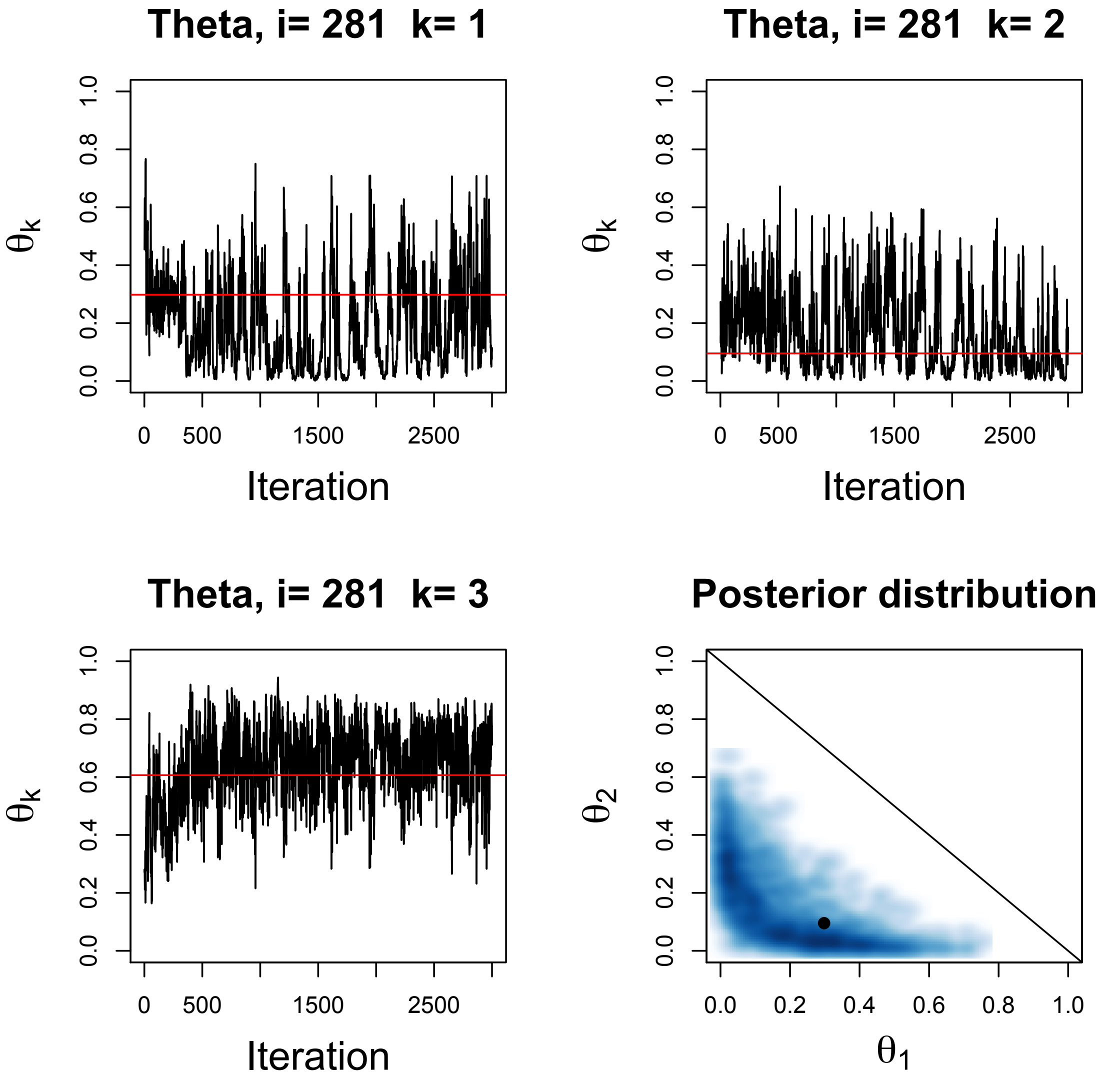

Figure 4.5: Simulation 1, replication 1 (30 items per student). Thinned MCMC chains and posterior distribution for simulated individual $i = 281$. The posterior distribution in the lower right shows the simulated value plotted as a red dot.

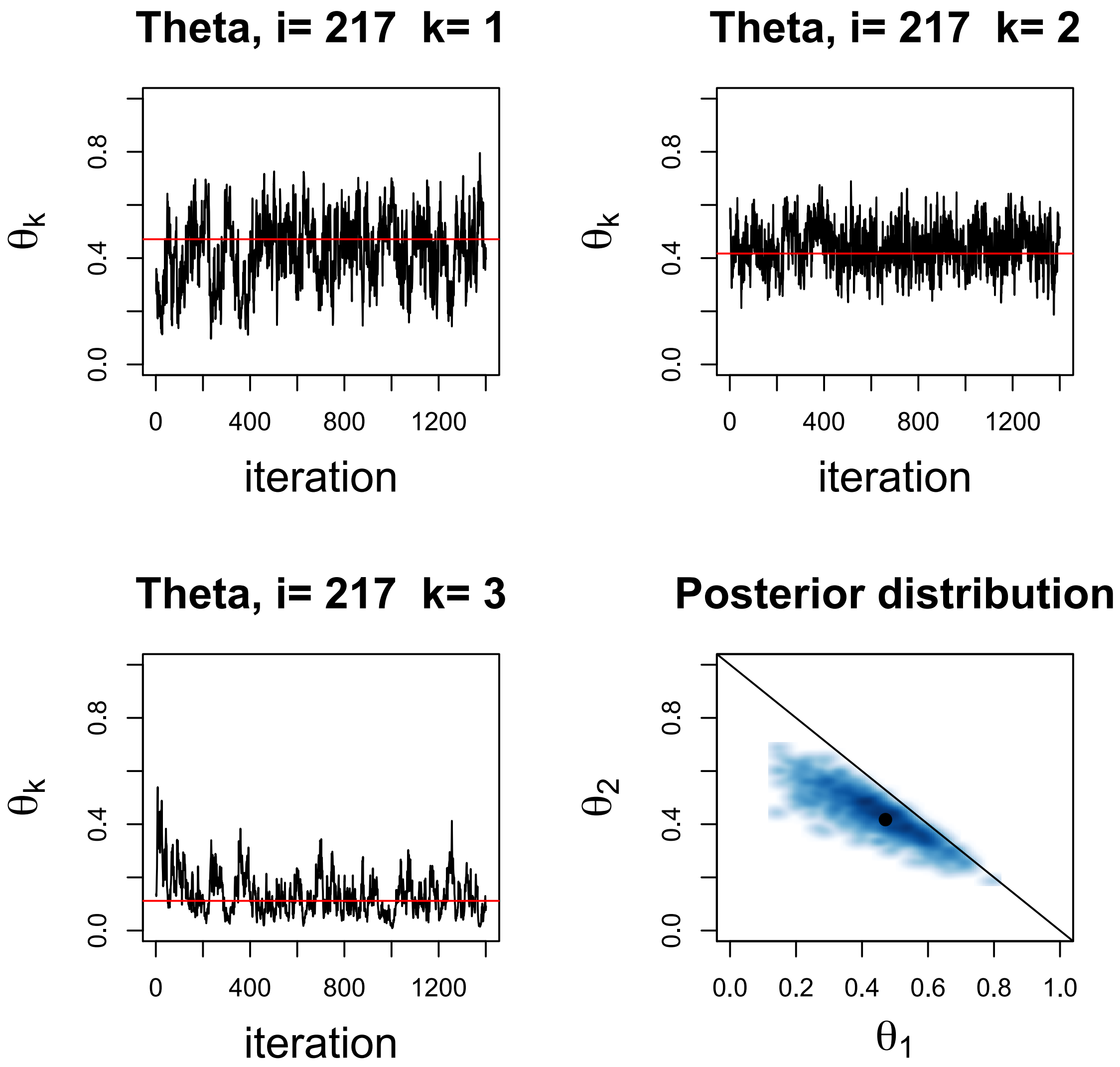

Figure 4.6: Simulation 1, replication 2 (30 items per student). Thinned MCMC chains and posterior distribution for simulated individual $i = 217$. The posterior distribution in the lower right shows the simulated value plotted as a black dot.

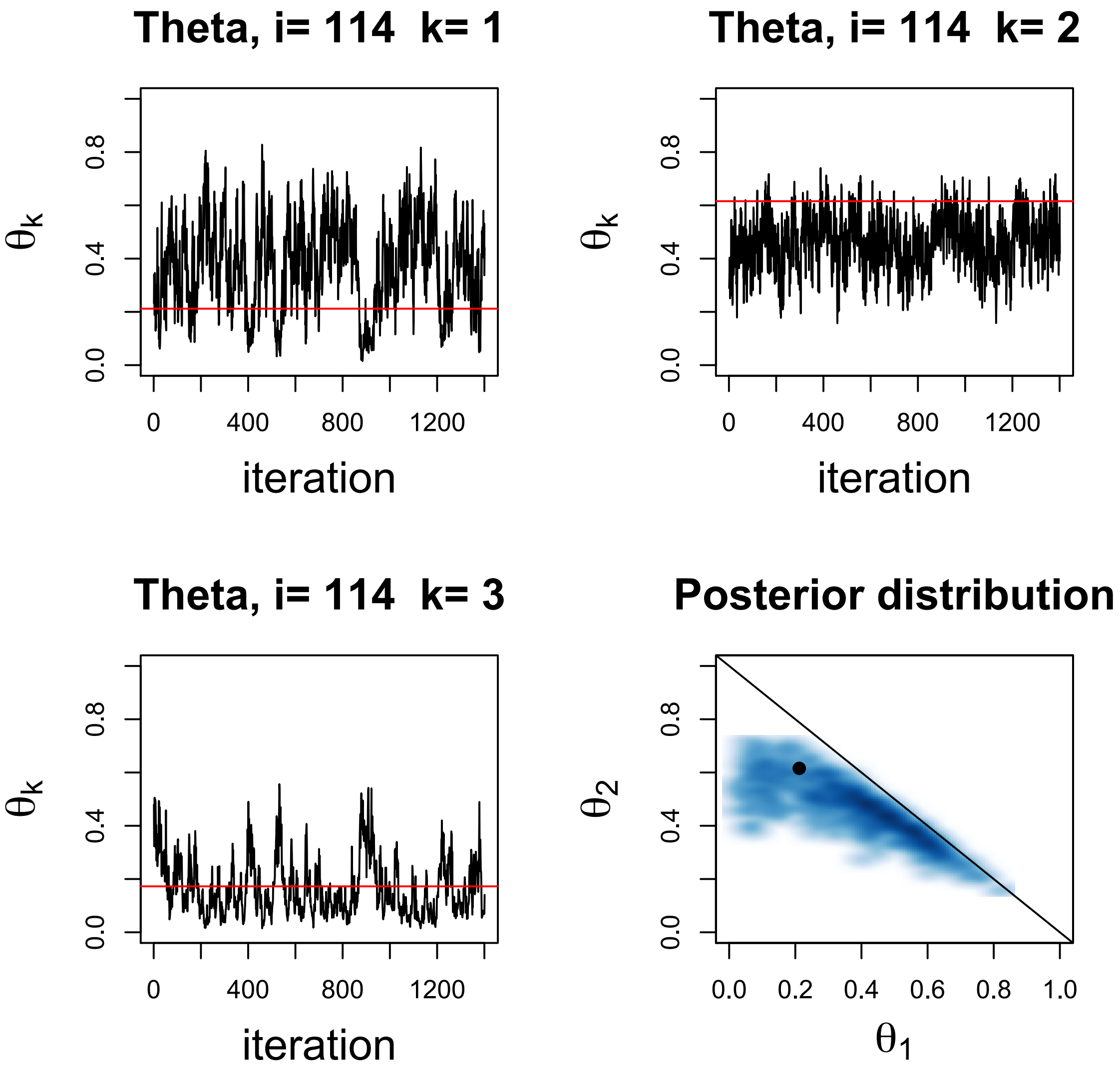

Figure 4.7: Simulation 1, replication 2 (30 items per student). Thinned MCMC chains and posterior distribution for simulated individual, $1 = 114$. The posterior distribution in the lower right shows the simulated value plotted as a black dot.

4.4.2 *Simulation 2: Average of 15 items per student, flat prior for λ.*

This set of simulations was almost identical to those in Section 4.4.1, with one important change. For this set of simulations, there were $J = 30$ unique items, and each student saw $\mathtt{Poisson}(15)$ items. I ran four replications of this simulation, and used the same flat priors for estimation as in Section 4.4.1.

With this smaller number of item per student, estimates for all parameters became much noisier. As might be expected, estimates individual membership parameters had much larger variation, and showed substantial shrinkage toward the estimated population distribution. However the distribution of the membership parameters was itself very poorly estimated. (Figures 4.8, 4.9, 4.10, and 4.11). Reasonable estimates of μ and Σ were found in only one of the four replications. Additionally, several posterior distributions of β_{kt} did not cover the simulated parameters and many posterior distributions of λ_{kj} covered almost the entire unit interval.

These simulations make it clear that even when the model is correct, 15 items per student are simply not enough data to reliably estimate both the strategies and how much students use each one. This is a rather serious concern since the least common multiples data set contains only 8-16 items per student.

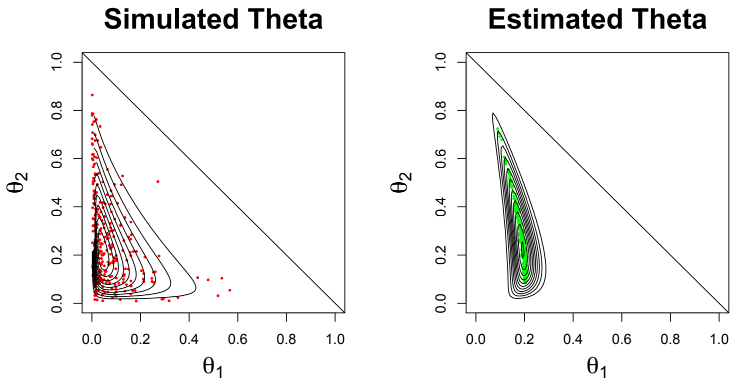

Figure 4.8: Simulation 2, replication 1 (15 items per student). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

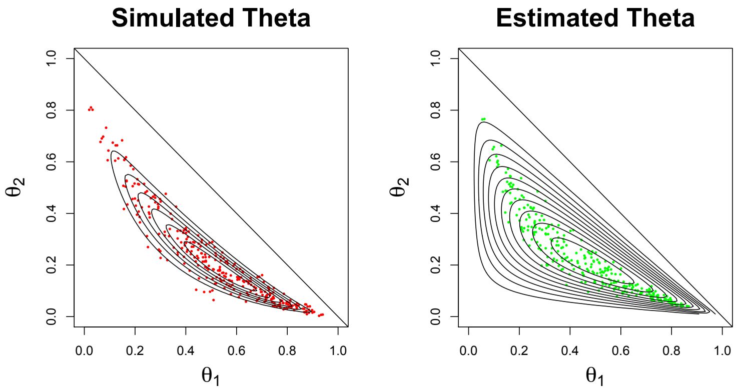

Figure 4.9: Simulation 2, replication 2 (15 items per student). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

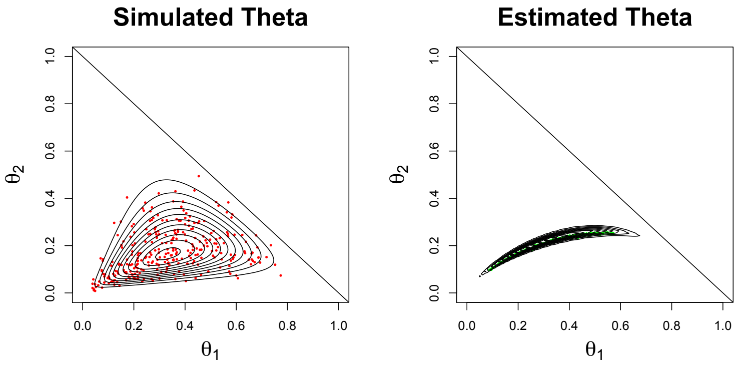

Figure 4.10: Simulation 2, replication 3 (15 items per student). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

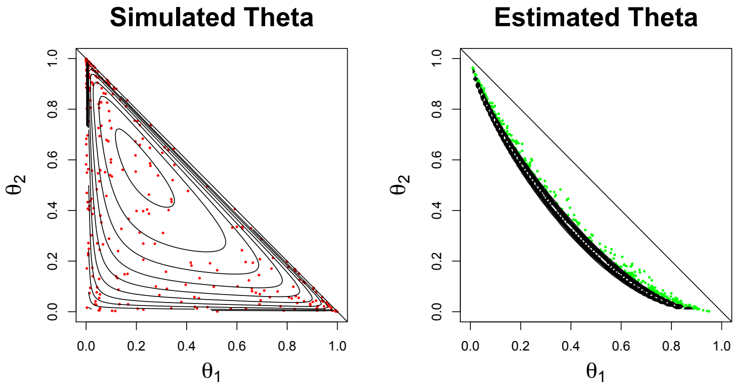

Figure 4.11: Simulation 2, replication 4 (15 items per student). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

4.4.3 *Simulation 3: Average of 15 items per student, Known* λ

This simulation differs from the simulations in sections 4.4.1 and 4.4.2 in how λ is treated. Values were simulated according to $\lambda_{kj} \sim \mathtt{Uniform}(0, 1)$, but during estimation these values were treated as "known". That is, a point-mass prior was placed on the simulated value. The simulation included $J = 30$ unique items, and each student saw $J_i \sim \mathtt{Poisson}(15)$ items.

Convergence in this case happened in less than 500 iterations. The posterior distributions for the remaining "unknown" parameters were appropriately narrow and covered the simulated values (Figure 4.12). It is also worth noting that posterior distributions for θ_i (Figures 4.13 and 4.14) showed markedly less shrinkage towards the population distribution than observed in the simulation with double the number of items per student but unknown item parameters (Simulation 4.4.1).

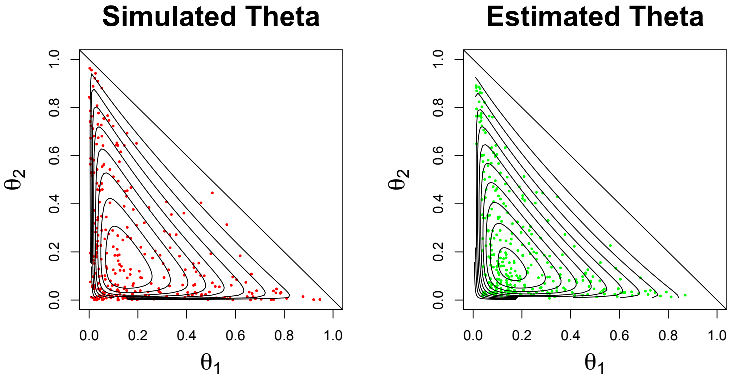

Figure 4.12: Simulation 3 (15 items per student, known λ). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

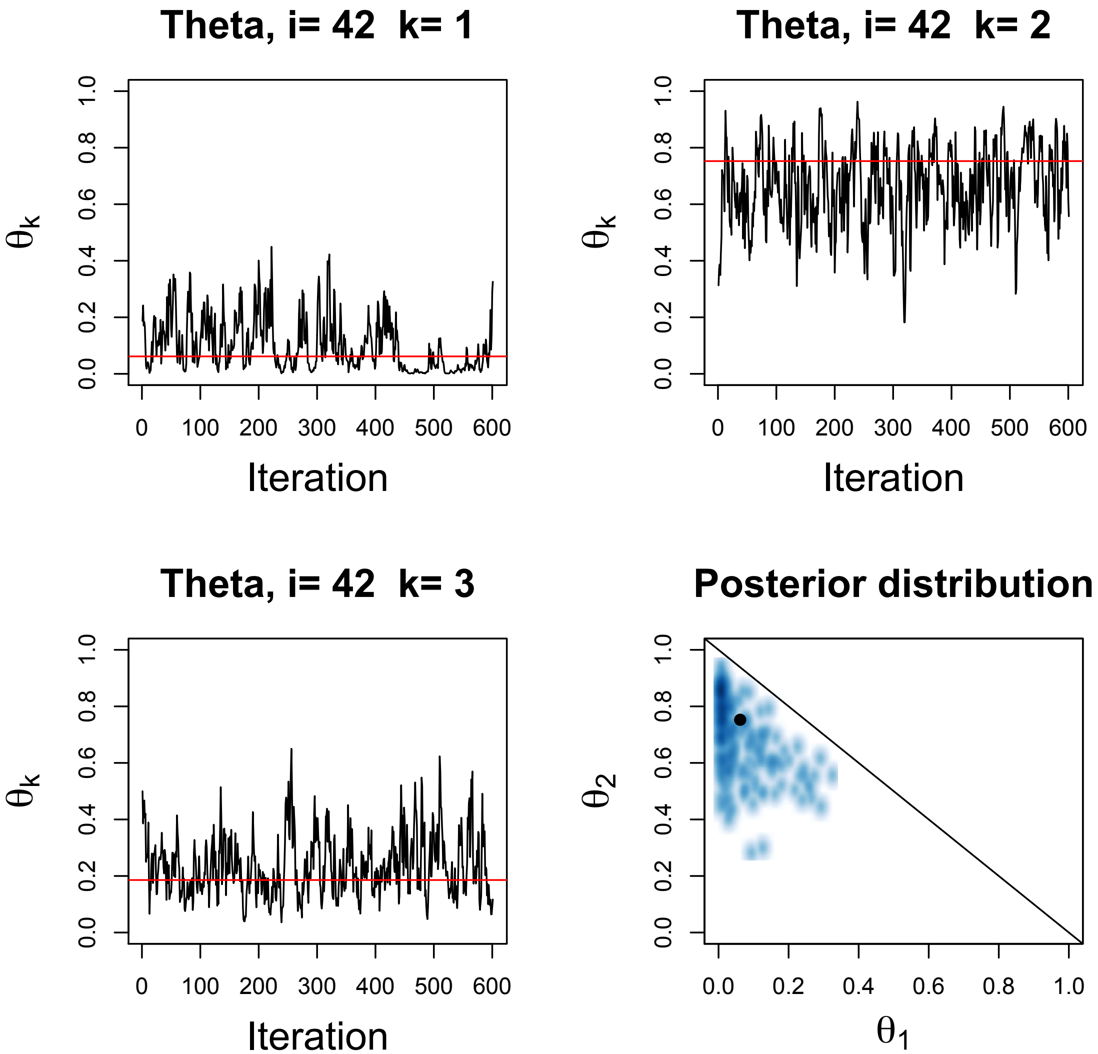

Figure 4.13: Simulation 3 (15 items per student, known λ). MCMC chains and posterior distribution for simulated individual $i = 42$. The posterior distribution in the lower right shows the simulated value plotted as a black dot.

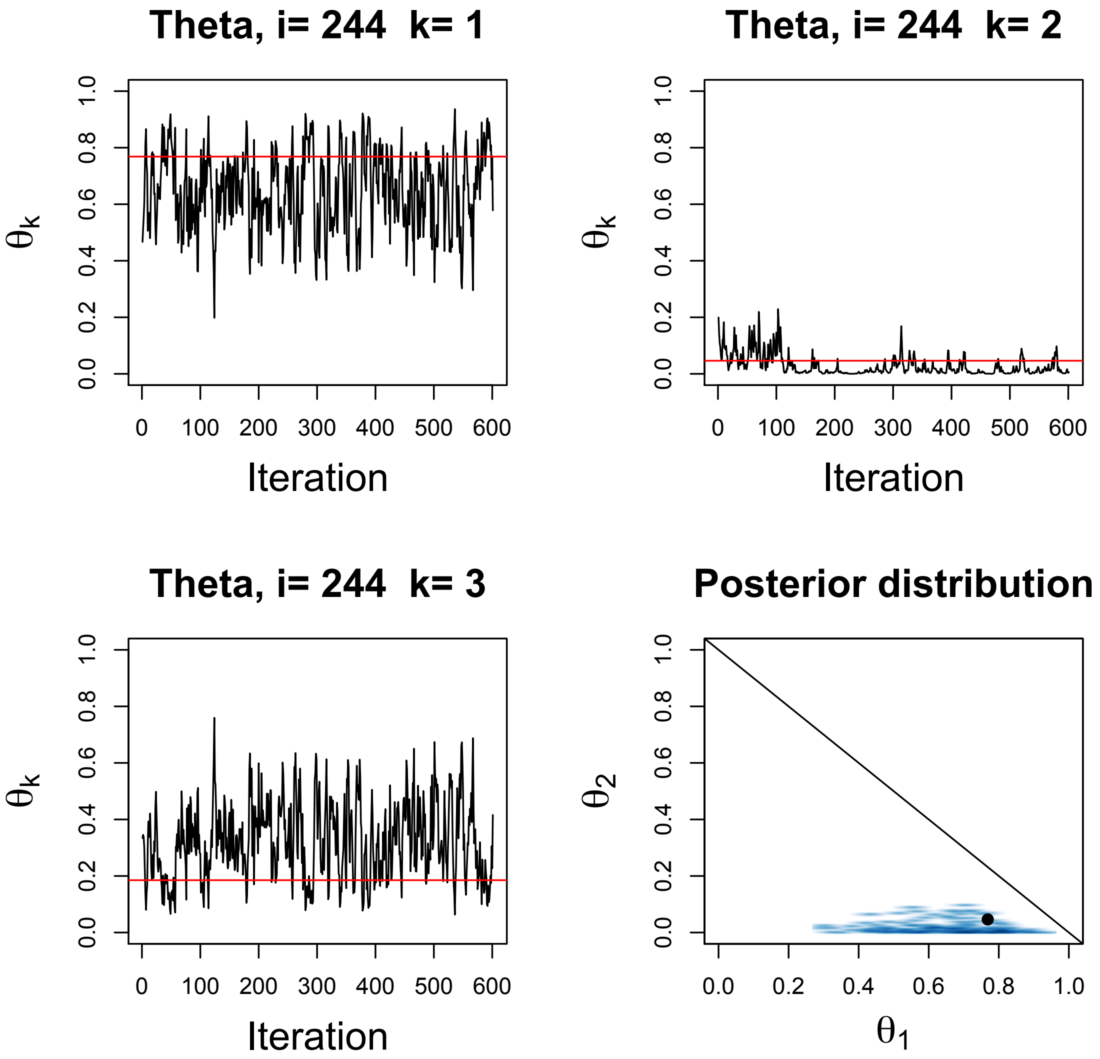

Figure 4.14: Simulation 3 (15 items per student, known λ). MCMC chains and posterior distribution for simulated individual $i = 244$. The posterior distribution in the lower right shows the simulated value plotted as a black dot.

4.4.4 *Simulation 4: Average of 15 items per student, Informative prior for λ_{1j}.*

Simulation 2 demonstrates that with 15 items per student, we cannot reliably estimate both the strategy parameters and individual membership in the strategies. Simulation 3 shows that if the strategies are known, then it is easy to estimate individual membership in each strategy. We now consider whether or not we can estimate the model when we have some prior information for a single strategy.

For example, suppose a particular misconception is common. Understanding this misconception means we have prior information about one of the strategy profiles. As another example, suppose some students know a correct strategy, but we wish to discover if there are any misconceptions present in the data set. We can set a prior distribution for one of the strategy profiles that reflects this prior information. In these situations, we specify an informative prior for λ_{kj} for $k = 1$, but use a flat prior for $k = 2, \dots, K$. This allows us to include prior information about the strategy we understand, but estimate the other strategies for which we have no information about.

For the simulation, I designated $k = 1$ as a 'correct' strategy profile. The probability of correctly answering an item within this profile were simulated as $\lambda_{1j} \sim \mathtt{Beta}(10, 1)$. This distribution has a mean of $10/11 \approx 0.91$, and $\Pr(\lambda_{ij} > 0.7) = 0.97$.

I ran two replications of this simulation with variations in how the remaining 2 profiles were generated. In one replication, the other two profiles had parameters $\lambda_{kj} \sim \mathtt{Beta}(1, 1)$. In the other replication, I considered a profile representing a superficial strategy that worked in some cases and not in others $\lambda_{2j} \sim \mathtt{Bernoulli}(0.5)$. The other profile was generated by $\lambda_{3j} \sim \mathtt{Beta}(1, 1)$. Once again, there were $J = 30$ unique items, and each student saw $J_i \sim \mathtt{Poisson}(15)$ items. In both replications of this simulation for the MCMC estimation, I placed a $\mathtt{Beta}(10, 1)$ prior on profile $k = 1$, and a $\mathtt{Beta}(1, 1)$ prior on profiles $k = 2, 3$.

The estimation results were similar to those for 30 items per student (Section 4.4.1). Anchoring a single profile with an informative prior allows us to estimate both of the other strategies, and students' membership in each strategy as well as if we had twice as many items per student. (Figures 4.15 and 4.16).

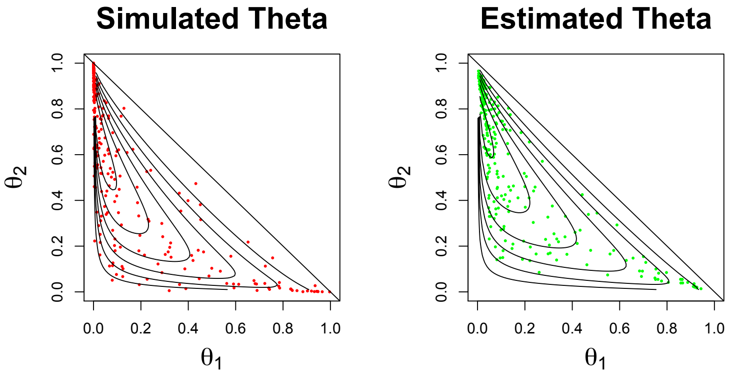

Figure 4.15: Simulation 4, replication 1 (15 items per student, informative prior on λ_{1j}). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

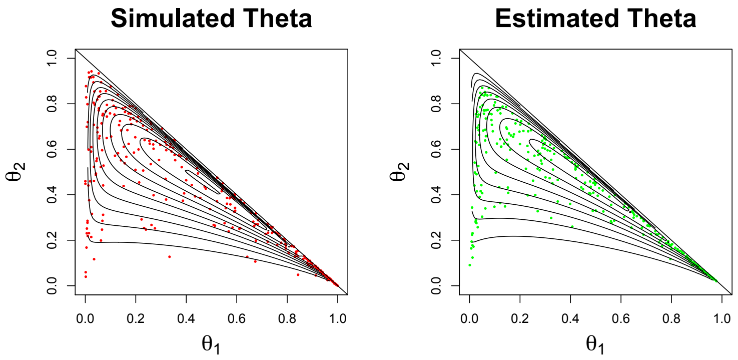

Figure 4.16: Simulation 4, replication 2 (15 items per student, informative prior on λ_{1j}). The contour plot on the left shows the distribution of θ defined by the simulated values of μ and Σ. The red dots show the simulated values of θ_i. On the right, the green dots show the posterior means of θ_i, and the contour plot is based on the posterior means of μ and Σ.

4.5 ANALYSIS OF LEAST COMMON MULTIPLE ASSESSMENT DATA

As described in Section 4.1, we have 255 students with 8 or 16 items per student. The simulations in Section 4.4 indicate that we cannot estimate both the strategy parameters, λ_{kj} and β_k, and the student membership vector θ_i with this amount of data unless we use an informative prior distribution for one strategy.

In this application, there are two known strategies. The first is a correct strategy, where the child computes the least common multiple of two numbers. The second strategy is a misconception, a 'product' strategy where children simply multiply the two numbers together, rather than computing the LCM. The misconception produces a correct answer when two numbers are relatively prime, such as 8 and 9, but the strategy does not always work. For example, the misconception will produce an incorrect response for the numbers 6 and 8. The correct strategy and the misconception represent the theoretical strategies. I include a third strategy profile in the model in order to describe any additional variation in student behavior, so that $K = 3$.

In the following analyses, we experiment with incorporating different amounts of prior information about the strategies, to see how well we can recover the theoretical strategies. I always use a flat prior for λ to estimate the third unknown strategy. This allows me to find additional patterns in the data if they exist, or to represent a guessing strategy in the absence of other meaningful patterns.

For each student, and each item we have the response and two response times, the time before they began to type a solution and the time that it took to enter the solution, $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$. In general T_{ij2} will be much shorter than T_{ij1}, but T_{ij2} may have many outliers. If a student accidentally hits a key before they are ready to enter a solution, or if they begin to enter a solution and change their mind, then T_{ij2} may be very long. Therefore, we also consider a model that simply includes total solution time where $T_{ij} = T_{ij1} + T_{ij2}$, and $X_{ij} = (C_{ij}, T_{ij})$.

4.5.1 *Model with 2 response times*

This is the full model with both times included, $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$. The strategy profiles are of the form:

$$F_{kj}(X_j) = \mathrm{Bernoulli}(C_j; \lambda_{kj}) \times \mathrm{Exp}(T_{j1}; \beta_{k1}) \times \mathrm{Exp}(T_{j2}; \beta_{k2}) \tag{4.49}$$

The simulations indicate that with a data set of this size, estimation will be unstable unless we incorporate some prior information for λ. We can investigate this behavior in the context of the data by comparing results for a model with a flat prior for λ to a model with an informative prior for one of the strategy profiles.

I chose the priors for β to be weak conjugate gamma priors scaled to reflect the distribution of the two distinct time periods. The longer time T_{ij1}, the time for a student to begin entering a response, has the hyper prior $\beta_{k1} \sim \mathrm{Gamma}(1, 40000)$. The shorter time T_{ij2}, the time a student takes to complete the response, has the hyper prior $\beta_{k2} \sim \mathrm{Gamma}(1, 2000)$.

We will compare results for two models with different priors for λ. The first is a flat prior, $\lambda_{kj} \sim \mathrm{Unif}(0, 1)$ for all k and j. The second uses an informative prior for $k = 1$, and a flat prior for $k = 2, 3$. In the model with an informative prior, I chose to let the strategy profile $k = 1$ represent a correct strategy. A student who uses this strategy should have a high probability of correctly responding to each item, and we can consider a student who belongs completely to this profile to have mastered the content. I represented this expert strategy profile with the prior $\lambda_{1j} \sim \mathrm{Beta}(10, 1)$. I ran two MCMC chains for each model.

4.5.1.1 *Model with uniform prior for λ_{kj}.*

Based on the simulations, we expect this model to perform poorly since there are only 16 items per student and a non-informative prior for λ, and indeed this is the case. The two MCMC chains did converge to the same posterior distributions for all parameters; however, the distributions bear remarkable similarity to the posteriors estimates from simulation 2 in Section 4.4.2.

Figure 4.17 shows the estimated posterior distribution for θ, which bears a remarkable similarity to the estimated posterior distributions of θ for simulation 2, (Figures 4.8, 4.10, and 4.11). In each case, the posterior density has collapsed to a narrow curvilinear shape within the simplex.

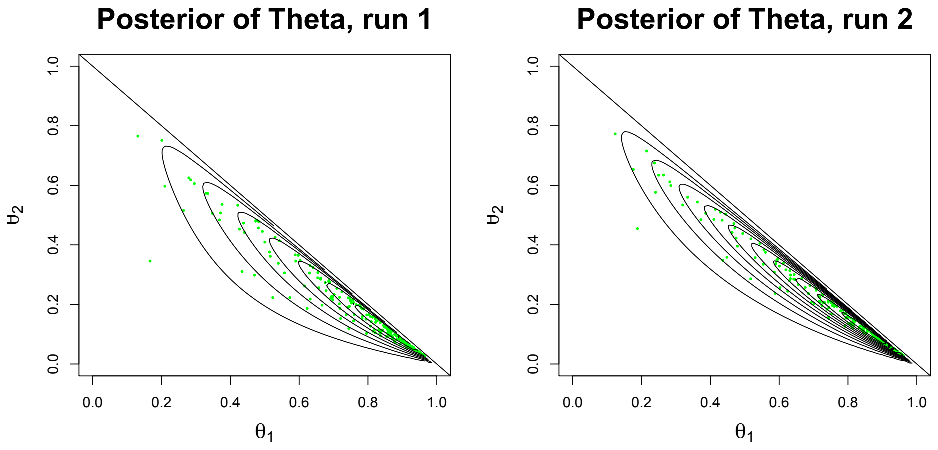

Figure 4.17: Section 4.5.1. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses a flat prior for λ. Contour plots of posterior distribution of θ based on posterior mean of μ and Σ. Green dots are posterior means for each θ_i.

In addition, the posterior estimates of λ_k, the probability of a correct response within each strategy, are not substantially different between the profiles. Each of the three strategy profiles bears some resemblance to the theorized misconception strategy (Figure 4.18).

The distinguishing feature of the profiles in this version of the model are the β_k parameters which govern response time (Table 4.2). Profile 1 has the shortest average time for both T_{ij1} and T_{ij2}. Profile 2 has a very long average T_{ij1}, but the second response time is similar to profile 1. Profile 3 has a longer average T_{ij1} than Profile 1, but an exceptionally long average 2nd response time.

These results agree with the simulations. Sixteen items per student is simply not enough data to estimate a mixed membership multiple strategies model without the use of some prior information for at least one of the strategies.

Table 4.2: Section 4.5.1. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses a flat prior for λ. Posterior means and 95% credible intervals for time parameters β. The mean response time for each profile is $1/\beta_{kt}$.

	β_{k1}	β_{k2}
Profile 1	5.1×10^{-5} $(4.8 \times 10^{-5},\ 5.3 \times 10^{-5})$	8.3×10^{-4} $(7.9 \times 10^{-4},\ 8.7 \times 10^{-4})$
Profile 2	7.6×10^{-6} $(6.7 \times 10^{-6},\ 8.5 \times 10^{-6})$	5.1×10^{-4} $(4.3 \times 10^{-4},\ 6.0 \times 10^{-4})$
Profile 3	1.4×10^{-5} $(1.1 \times 10^{-5},\ 1.6 \times 10^{-5})$	2.9×10^{-5} $(2.3 \times 10^{-5},\ 3.6 \times 10^{-5})$

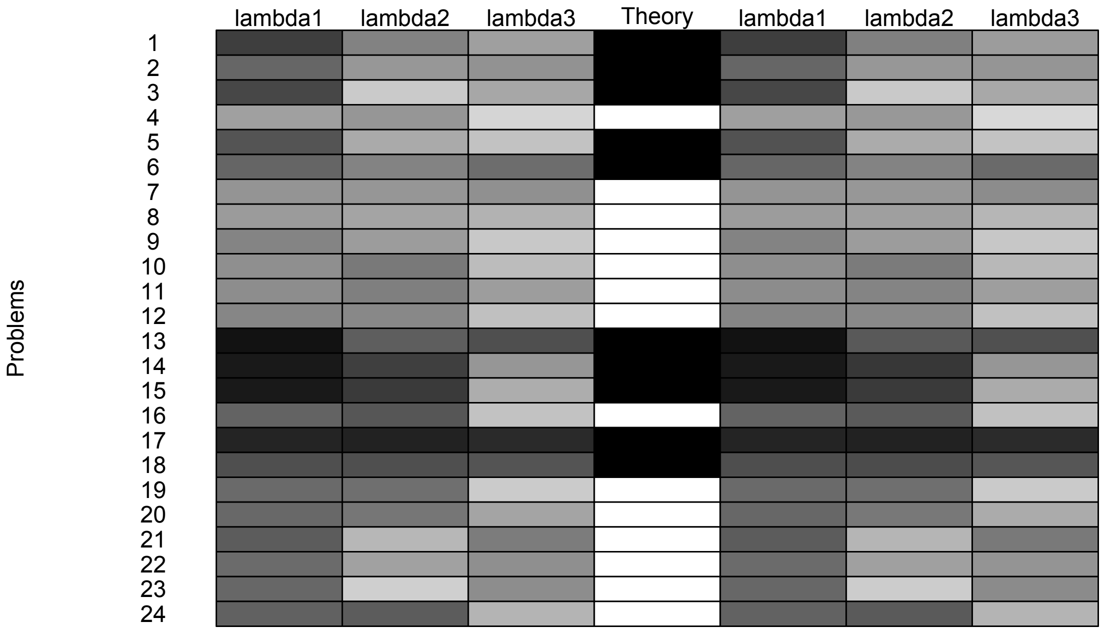

Figure 4.18: Section 4.5.1. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses a flat prior for λ. The first 3 columns are posterior mean estimates of λ_{kj} for the first MCMC run. The last 3 columns are the same for the second MCMC run. The middle column labeled “Theory" indicates the items where the theorized misconception strategy works. Darker cells indicate higher values of λ_{kj} and a higher probability of a correct response.

4.5.1.2 *Model with informative prior for* λ_{1j}.

Based on the simulations, we expect this model to perform better than the model with a flat prior for λ. Figure 4.20 shows the posterior means for λ. These results reflect a better distinction between a correct strategy and the misconception strategy.

The two MCMC runs for this model produced almost identical results, except that the indices of the 2nd and 3rd profiles were permuted between the two runs. Most parameters converged in under 100 iterations, μ and Σ converged after 1000 iterations.

Posterior means of λ for strategy profile $k = 1$ reflect a correct strategy, where the probability of a correct answer is high for each item. We expect this, since the prior distribution $\lambda_{1j} \sim \mathtt{Beta}(10, 1)$ specifies a correct strategy. This strategy is also the fastest strategy (Table 4.3).

Strategy profile $k = 3$ resembles the theoretical misconception strategy most strongly. A student who is using the misconception strategy should have a high probability of a correct response for some items and a very low probability of a correct response for the other items. We see this reflected in the estimated values of λ_{3j} for items $j = 1, 5, 13, 14, 15, 17, 18$. This strategy is slower than the correct strategy. The posterior distributions of β indicate that a student using this misconception strategy takes on average more than twice as long to work before they begin to type their response (T_{ij1}) compared to a student using the correct strategy. The time required to type is not significantly different.

The strategy profile $k = 2$, may be considered a true novice strategy. The probability of a correct response is low for 23 of the 24 items, but the distinguishing feature of this strategy is that it is exceptionally slow. Both of the two response times are on average, more than ten times longer than for the correct strategy.

The posterior distribution of θ (Figure 4.19) is much more distributed across the simplex than in the model which used no prior information (Figure 4.17). The distribution of the membership parameter indicates that most students predominantly use some combination of the correct strategy and the misconception strategy. Membership in the slow profile $k = 2$ is very low.

The strategy profiles in this model are more strongly distinguished by response time than the probability of a correct response. Therefore, we should compare this model which uses $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$ with models which use a single total response time ($X_{ij} = (C_{ij}, T_{ij})$), and a model with no response times included ($X_{ij} = C_{ij}$).

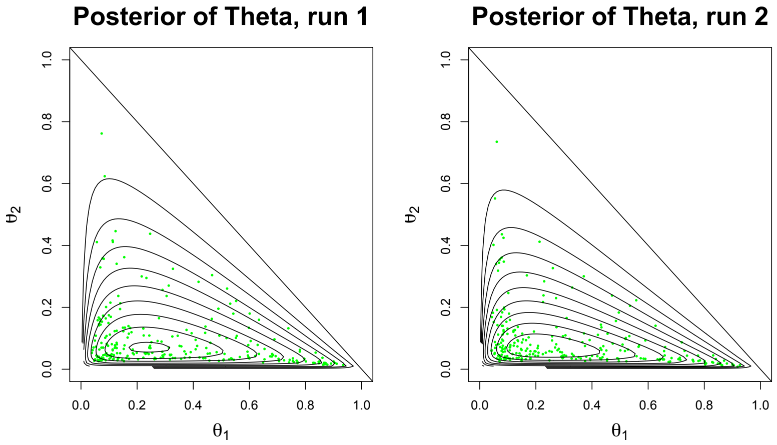

Figure 4.19: Section 4.5.1. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses an informative prior for λ_{1j}. Contour plots of posterior distribution of θ based on posterior mean of μ and Σ. Green dots are posterior means for each θ_i. High values of θ_{i1} correspond to high membership in the ‘correct’ profile. Points near the origin correspond to high values of θ_{i3}, which is the "immature" strategy.

Table 4.3: Section 4.5.1. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses an informative prior for λ_{1j}. Posterior means and 95% credible intervals for time parameters β. The mean response time for each profile is $1/\beta_{kt}$.

	β_{k1}	β_{k2}
Profile 1	6.7×10^{-5} $(5.9 \times 10^{-5},\ 7.6 \times 10^{-5})$	8.6×10^{-4} $(7.9 \times 10^{-4},\ 9.4 \times 10^{-4})$
Profile 2	5.5×10^{-6} $(4.8 \times 10^{-6},\ 6.2 \times 10^{-6})$	5.6×10^{-5} $(4.8 \times 10^{-5},\ 6.5 \times 10^{-5})$
Profile 3	2.9×10^{-5} $(2.7 \times 10^{-5},\ 3.2 \times 10^{-5})$	7.5×10^{-4} $(7.0 \times 10^{-4},\ 8.0 \times 10^{-4})$

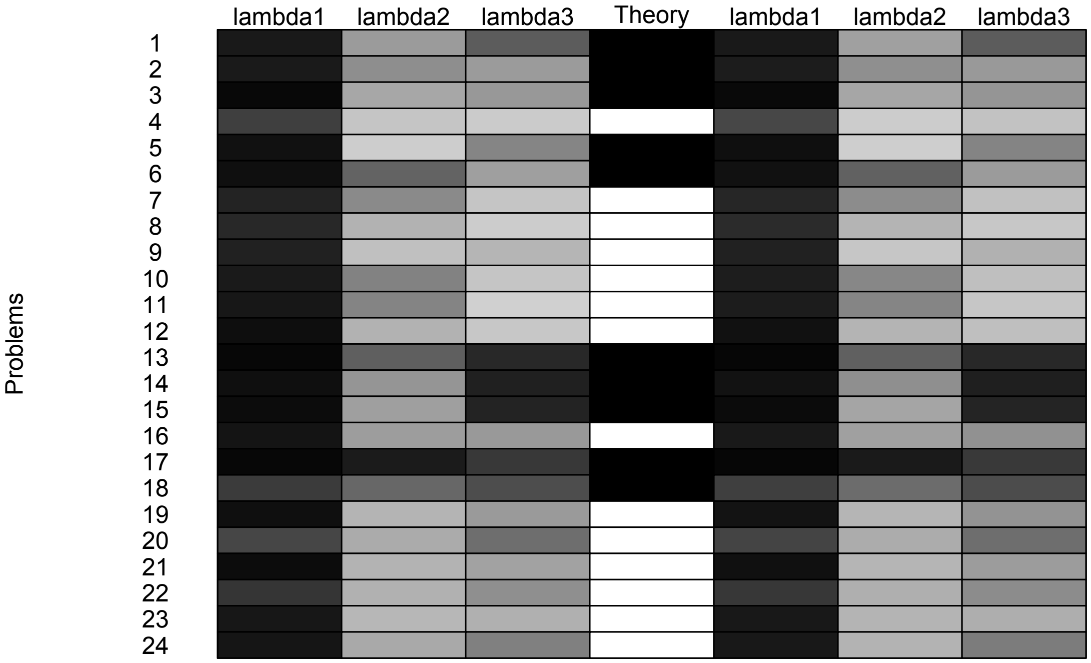

Figure 4.20: Section 4.5.1. $X_{ij} = (C_{ij}, T_{ij1}, T_{ij2})$, and estimation uses an informative prior for λ_{1j}. The first 3 columns are posterior mean estimates of λ_{kj} for the first MCMC run. The last 3 columns are the same for the second MCMC run. The middle column labeled "Theory" indicates the items where the misconception strategy will produce a correct answer. Darker cells indicate higher values of λ_{kj} and a higher probability of a correct response.

4.5.2 *Model without Time*

Let us consider a model based only on the responses, with no time included, $X_{ij} = C_{ij}$. How much information is in the accuracies alone, and how much do we gain from including response time?

In this model, the priors for λ in all three profiles were flat $\mathtt{Uniform}(0, 1)$ distributions. Figure 4.22 shows the posterior means for λ. Profile 1 reflects a correct strategy. Profile 3 reflects the misconception strategy. Profile 2 indicates a higher probability of a correct response for items 13-24 than for 1-12, reflecting that the first 12 items are story problems, the second 12 are not.

The strategy profiles reflect reasonable properties of student behavior, but individual membership in each profile is impossible to estimate well for most students. For the expert students who exclusively use the correct strategy, the posterior distribution of θ_i reflects this (Figure 4.23). For other students, the posterior distribution of θ covers the entire simplex (Figure 4.24). Without using the information in the response times, we cannot estimate how much students use each strategy.

Since it is so difficult to estimate θ_i, it is not surprising that it is also difficult to estimate the distribution of $\theta \sim \mathtt{LogisticNormal}(\mu, \Sigma)$. The posterior means for μ between MCMC run 1 and MCMC run 2 are similar, but the posterior means for Σ are not similar between the two runs (Figure 4.21).

This model without response times confirms the presence of a correct strategy and the presence of a misconception strategy. It also indicates that we may need to pay particular attention to the story problems. Most of all though, we see that the inclusion of response time helps estimate student ability. With this small number of items per student, we need to include response time information in order to estimate individual strategy usage.

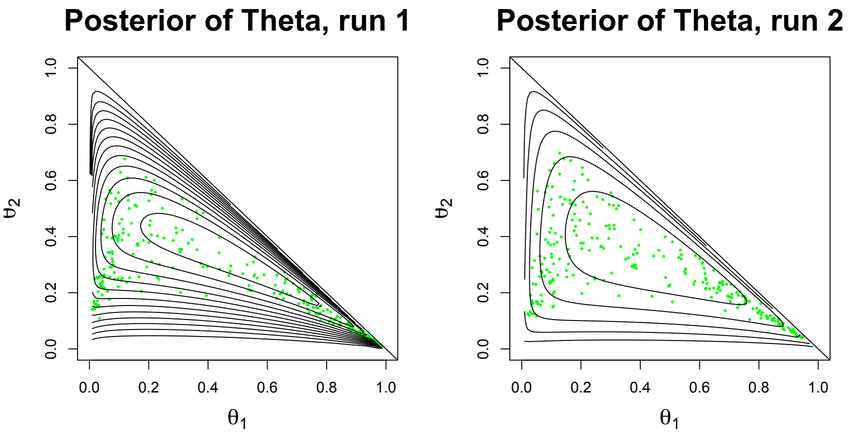

Figure 4.21: Section 4.5.2. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ. Contour plots of posterior distribution of θ based on posterior mean of μ and Σ. Green dots are posterior means for each θ_i. High values of θ_{i1} correspond to high membership in the correct strategy. Points near the origin correspond to high values of θ_{i3}, the misconception strategy.

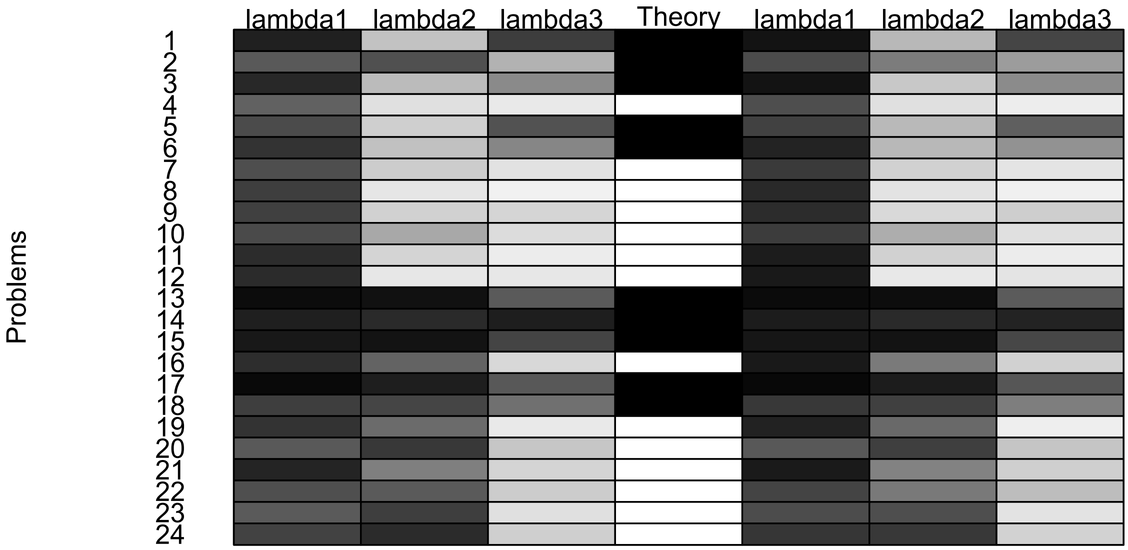

Figure 4.22: Section 4.5.2. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ. First 3 columns are posterior mean estimates of λ_{kj} for the first MCMC run. The last 3 columns are the same for the second MCMC run. The middle column labeled “Theory" indicates the items where the misconception strategy works. Darker cells indicate higher values of λ_{kj} and a higher probability of a correct response.

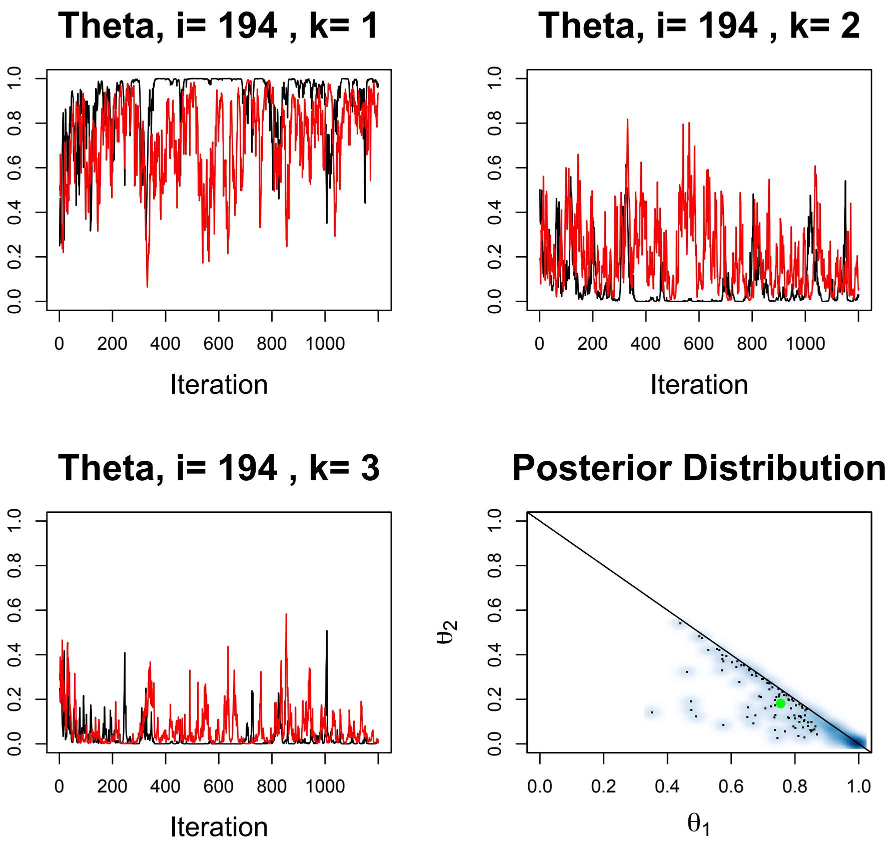

Figure 4.23: Section 4.5.2. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ. MCMC chains and posterior distribution for individual $i = 194$. The posterior distribution in the lower right is based on Run 1, with the posterior mean for Run 2 plotted as the green dot.

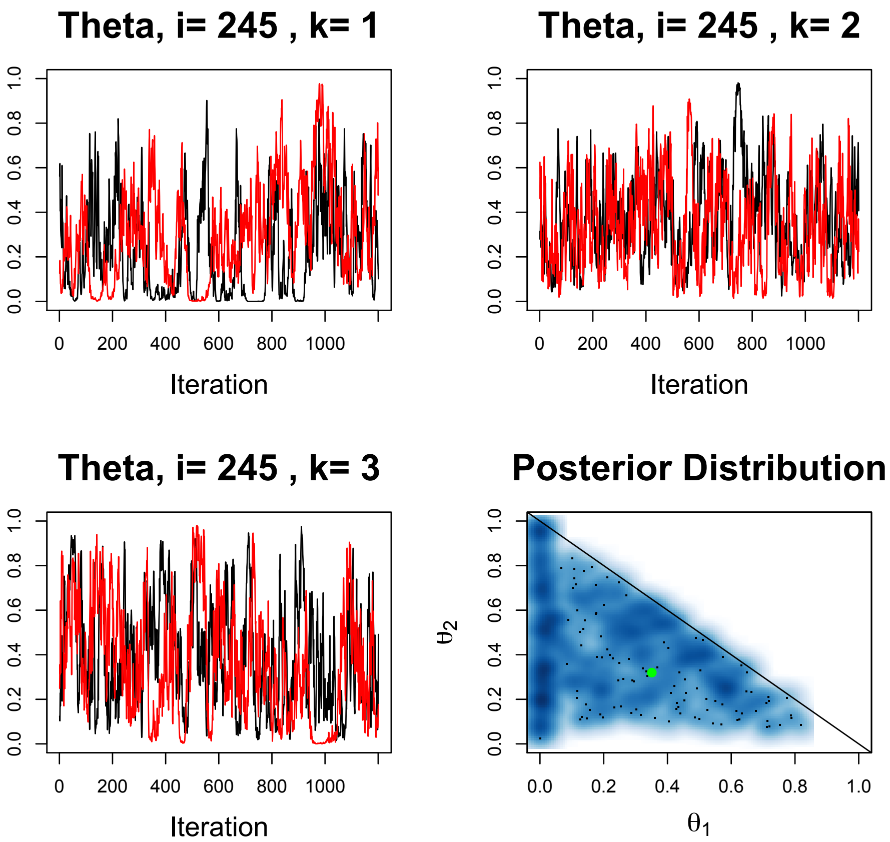

Figure 4.24: Section 4.5.2. $X_{ij} = C_{ij}$, and estimation uses a flat prior for λ. MCMC chains and posterior distribution for individual $i = 245$. The posterior distribution in the lower right is based on Run 1, with the posterior mean for Run 2 plotted as the green dot.

4.5.3 *Model with Total Time*

For this model, the data are $X_{ij} = (C_{ij}, T_{ij})$. The data include two separate solution times, T_{ij1}, the time before a student began to type an answer, and T_{ij2} the time to finish typing. Rather than modeling these as two separate times, it may be more reasonable to combine them into a single “Time to solution", $T_{ij} = T_{ij1} + T_{ij2}$. The strategy profiles are of the form:

$$F_{kj}(X_j) = \mathtt{Bernoulli}(C_j; \lambda_{kj}) \times \mathtt{Exp}(T_j; \beta_k) \tag{4.50}$$

The model with two separate times (Section 4.5.1) and the simulations (Section 4.4) indicate that we need to use an informative prior for λ_{1j} in order to estimate the mixed membership multiple strategies model with 16 items per student. We designate strategy profile $k = 1$ as a correct strategy with the prior $\lambda_{1j} \sim \mathtt{Beta}(10, 1)$. For profiles $k = 2, 3$, I use a flat prior $\lambda_{kj} \sim \mathtt{Uniform}(0, 1)$. For the response times, we use a weak hyper-prior that reflects the scale of the data $\beta_k \sim \mathtt{Gamma}(1,\ 40,000)$.

As with the other models, I ran two separate MCMC chains from random starts to compare convergence. Most parameters converged almost immediately. μ and Σ were again the slowest parameters to converge, taking around 2000 iterations. The indices of profiles 2 and 3 were permuted between the two runs of MCMC. After re-indexing, results from the two runs are identical.

Of all the models considered, this model results in the clearest and most interpretable differences between the strategy profiles. The estimated strategy profiles are similar to those in the model with two response times (Section 4.5.1), but with stronger patterns apparent in the profiles, leading to better interpretation of what strategies the profiles represent.

Strategy profile $k = 1$ has a very high probability of a correct response for all items (Figure 4.25), and the fastest average response time (Table 4.4). This profile represents an expert or correct, strategy.

Strategy profile $k = 3$ is a misconception strategy. The probability of a correct response in this strategy is only relatively high when the known misconception would be successful (Figure 4.25). Though, we note that the story problems, items 1-12, are harder than the non-story problems. Students using the misconception strategy take, on average, about 30% longer than students using the correct strategy (Table 4.4).

Profile $k = 2$ is a slow strategy that may represent fumbling or guessing. The average solution time is more than 7 times longer than for the correct strategy (Table 4.4). Using this strategy, the probability of a correct response on most items is near 0.5. Item 17 is peculiar, $\lambda_{2,17} = 0.9$. There is nothing obvious that makes item 17 special, “What is the least common multiple of 4 and 5?" In comparison, item 13, “What is the least common multiple of 3 and 5?" is not nearly as easy.

The distribution of membership parameters is similar for this model with one time (Figure 4.25), as for the model with two times (Figure 4.19). The density is highest in the corners of the simplex near the expert strategy and the misconception strategy, indicating that many students who do not switch strategies.

Overall, the results from this model with one response time are similar to those from the model with two response times. However, the single response time appears to yield a cleaner, more interpretable model.

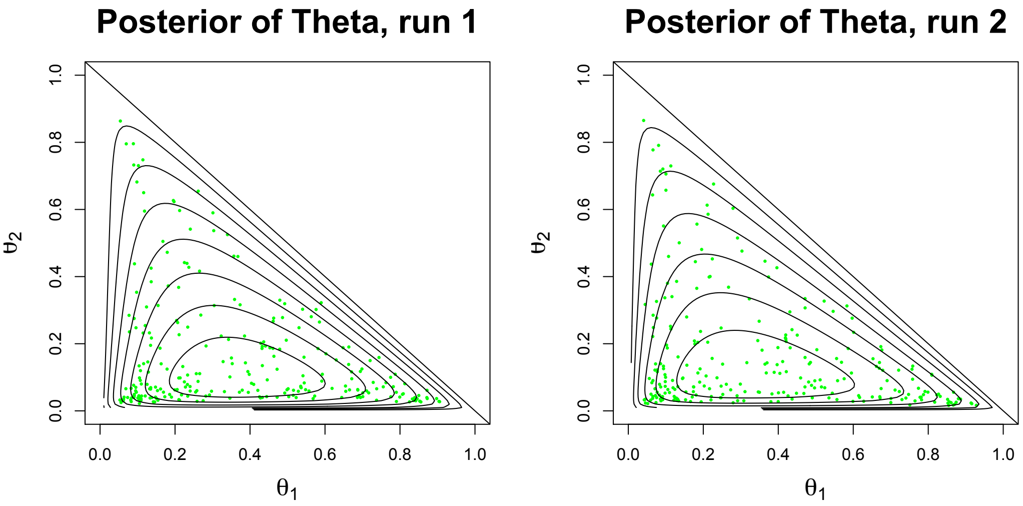

Figure 4.25: Section 4.5.3. $X_{ij} = (C_{ij}, T_{ij})$, and estimation uses an informative prior for λ_{1j}. Contour plots of posterior distribution of θ based on posterior mean of μ and Σ. Green dots are posterior means for each θ_i. High values of θ_{i1} correspond to high membership in the 'correct' profile. Points near the origin correspond to high values of θ_{i3}, which is the "immature" strategy.

Table 4.4: Section 4.5.3. $X_{ij} = (C_{ij}, T_{ij})$, and estimation uses an informative prior for λ_{1j}. Posterior means and 95% credible intervals for time parameters β. The mean response time for each profile is $1/\beta_k$.

	β_k
Profile 1	5.0×10^{-5} $(4.6 \times 10^{-5},\ 5.5 \times 10^{-5})$
Profile 2	6.5×10^{-6} $(5.8 \times 10^{-6},\ 7.4 \times 10^{-6})$
Profile 3	3.9×10^{-5} $(3.6 \times 10^{-5},\ 4.2 \times 10^{-5})$

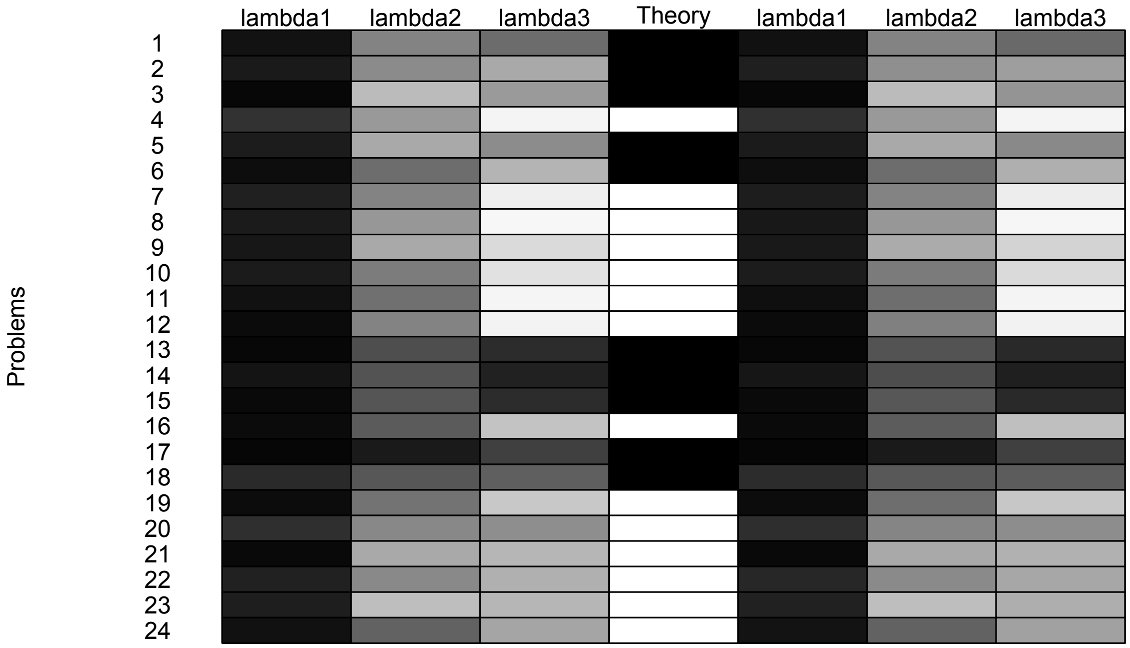

Figure 4.26: Section 4.5.3. $X_{ij} = (C_{ij}, T_{ij})$, and estimation uses an informative prior for λ_{1j}. First 3 columns are posterior mean estimates of λ_{kj} for the first MCMC run. The last 3 columns are the same for the second MCMC run. The middle column labeled "Theory" indicates the items where the superficial strategy works. Darker cells indicate higher values of λ_{kj} and a higher probability of a correct response.

4.6 comments

The multiple strategies model (MSM) successfully discovered the misconception strategy in the Least Common Multiples data. These results compare favorably to Cognitive Diagnosis Models (CDMs) along several dimensions. First, even this simpler form of the MSM allows for students to switch strategies from item to item, whereas CDMs assume that every student uses the same strategy on each item. Second, the MSM is able to learn the strategies, while the association between skills and items must be specified a priori for CDMs.

Perhaps most importantly, MSM performed reasonably well with a moderately sized data set. With 300 students and an average of 15 items per student, we need to have some information about the strategies in the data. However we do not need to know all of the profiles with complete certainty, as with CDMs. We only need some prior information about one of the strategy profiles. If we have more items per student, then we can discover all of the strategies with no prior information.

5

MULTIPLE STRATEGIES IN CHILDREN'S NUMERICAL ESTIMATION

5.1 INTRODUCTION

This chapter explores a second application of multiple strategy modeling. Unlike the Least Common Multiple data in Chapter 4, the numerical magnitude estimation data cannot be modeled with the Multiple Strategies model developed in Chapter 3. The reasons why the mixed membership structure is inappropriate for this data are discussed further in Chapter 6. In this chapter, we focus on the application itself.

Research into how children learn to estimate numerical magnitude has been an active area of research over the last decade. The ability to accurately estimate numerical magnitude is interesting not only from a developmental and cognitive perspective, but has also been shown to be closely related to other areas of mathematical proficiency, such as arithmetic (Booth and Siegler, 2008). Currently, one theory for how this process occurs is quite dominant (Siegler & Opfer, 2003; Siegler, Thompson, & Opfer, 2009), even though other theories have been suggested (Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008; Moeller, Pixner, Kaufmann, & Nuerk, 2008; Barth & Paladino, 2011). This paper re-examines data from early experiments using new analysis methods, and presents evidence that challenges the prevailing theoretical interpretation of these data.

Siegler and Opfer (2003) was a landmark study, and has been widely cited. This study devised a number-line task to compare theories for how children and adults represent numerical magnitude. They observed that young children appeared to use a logarithmic scale for estimation and develop a linear scale as they grew. In addition, second graders appeared to use a linear representation when presented with a 0-100 scale, but revert to a logarithmic representation on the 0-1000 scale. Subsequent research appeared to confirm this finding (Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Booth & Siegler, 2006; Dehaene, Izard, Spelke, & Pica, 2008; Laski & Siegler, 2007; Opfer & Thompson,2008; Siegler & Booth, 2004, 2005). However, the analysis methods for all of these papers are similar, and have many of the same drawbacks.

The current study provides a new tool for summarizing and visualizing data from these numerical magnitude estimation experiments, a method which can be easily utilized for a larger, more general class of data, including longitudinal data. Moreover, by applying this visualization tool to data from Siegler and Booth (2004); Booth and Siegler (2006) and Opfer and Siegler (2007), we build a foundation for more rigorous model-based analysis that challenges the theory of a linear-logarithmic representational shift. We show that while strategies do become more linear as children age, the early strategy is not logarithmic, but may be some combination of linear estimation on small numbers and simple categorization for larger numbers. In addition, we discover a third, intermediate strategy that has not been previously recognized.

5.2 METHOD

We re-examine data from three previously published studies on how young children estimate numerical magnitude (Booth and Siegler, 2006; Opfer and Siegler, 2007; Siegler and Booth, 2004). Booth and Siegler (2006) considers two different experiments, so that in total, data from four experiments were reanalyzed. The central task in each experiment is the number line estimation task. In this task, the stimulus is a number line with only the endpoints marked, and a number to be estimated printed above the line. Participants estimate the magnitude of the number by marking its position on the number line. The number lines are usually around 25 cm long, with a single number line on each page. In two experiments, the number lines were on a 0-100 scale while the other two used a 0-1000 scale. In all four of the experiments, small numbers are over-represented in order to better discriminate between the theorized linear and logarithmic strategies. Table 5.1 provides details on the participants and stimuli for each experiment.

5.3 CONCERNS WITH COMMON ANALYSIS TECHNIQUES

The number line task is the principal task in the vast majority of numerical magnitude estimation experiments, and the basic analysis is similar in each of these papers. The primary visual data summary takes the median estimate of each number for all of the participants in an age group and performs a regression through those median estimates (Figure 5.1). This visualization consistently shows a curve for the younger children that appears logarithmic, and a more linear curve for older children.

Table 5.1: Details of the four experiments being re-analyzed.

Experiment	Scale	Participants	Numbers to be Estimated
Siegler and Booth (2004)	0-100	Kindergartners (21)	{3, 4, 6, 8, 12, 17, 21, 23, 25,
		1st graders (32)	29, 33, 39, 43, 48, 52, 57,
		2nd graders (31)	61, 64, 72, 79, 81, 84, 90, 96}
Booth and Siegler (2006)	0-100	Kindergarteners (20)	{3, 4, 6, 8, 12, 14, 17, 18,
Experiment 1		1st graders (25)	21, 24, 25, 29, 33, 39,
		2nd graders (23)	42, 48, 52, 57, 61, 64,
		3rd graders (22)	72, 79, 81, 84, 90, 96}
Experiment 2	0-1000	2nd graders (30)	{3, 7, 19, 52, 103, 158, 240, 297,
		4th graders (28)	346, 391, 438, 475, 502, 586, 613,
			690, 721, 760, 835, 874, 907, 962}
Opfer and Siegler (2007)	0-1000	2nd graders (93)	{2, 5, 18, 34, 56, 78, 100, 122, 147,
		4th graders (60)	150, 163, 179, 246, 366, 486,
			606, 722, 725, 738, 754, 818, 938}

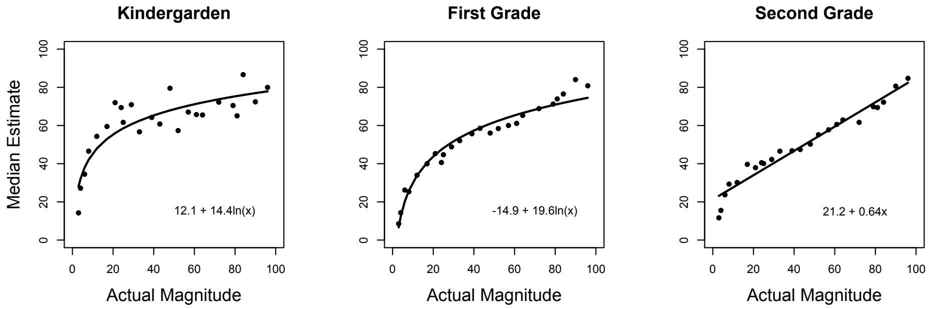

Figure 5.1: Standard visual summary of data used for number line estimation task. This was created with data from Siegler and Booth (2004). Within each grade level, the median estimate of each number is calculated. Regression is performed on the median estimates.

To determine whether individual strategies resemble the median group strategy, standard practice is to fit competing functions to each child's data, and assume that the function with the higher R^2 value is the strategy used by that child. For example, Siegler and Booth (2004) found that on a 0-100 scale, the logarithmic function better described 81% of the kindergarteners, but only 45% of the second graders.

There are three issues with this analysis method. First, visualizations based on median curves may be misleading, as demonstrated in Figure 5.2. In the same way that knowing the mean of a variable without knowing the standard deviation does not give us a complete picture; knowing the shape of a median curve without knowing how individuals vary around this curve is not a complete picture. Any summary of data must include information on variation.

Second, it may be convenient to assume that if a linear model fits better than the alternative logarithmic model, then a linear model is correct. However, it is quite possible that neither model fits very well, even if one is better than the other. A simple visual inspection, such as that in Figure 5.3, can reveal this problem. The

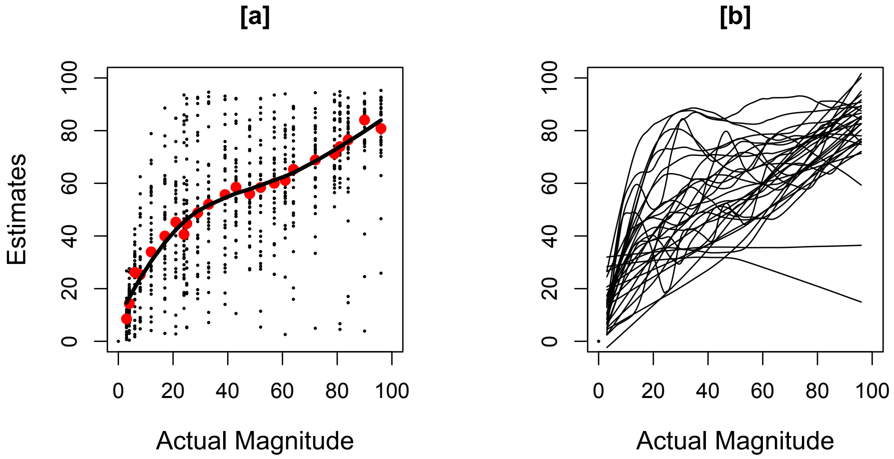

Figure 5.2: [a] Shows the estimates for all the first grade participants in Siegler and Booth (2004) with a smooth curve through the median points. [b] Shows a smooth curve for each individual. Examination of only the median curve hides individual variation.

issue is that if the experiment has more than a few participants it can be difficult and time consuming to perform this inspection for each participant. A good visualization should reveal whether our models are consistent with the observed data, and do this for multiple participants at once.

Finally, each theory for how children estimate numerical magnitude does not simply suggest a set of regression functions that might fit individual data, but rather puts forward a framework for which functions are more likely at different ages and what kinds of variation exist between children (Barth and Paladino, 2011; Ebersbach et al., 2008; Moeller et al., 2008; Siegler et al., 2009). The current practice of analyzing each individual separately with only R^2 as a measure of model fit, makes it impossible to know which theory truly describes the patterns in the data. All information on variation between individuals is lost. It is preferable to

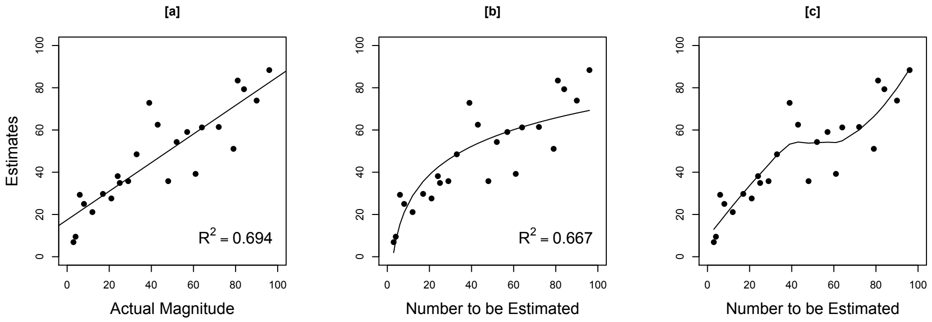

Figure 5.3: Number line estimation task data for a first grade participant in Siegler and Booth (2004). The linear fit in [a] is slightly better than the logarithmic fit in [b]. A smooth nonparametric curve is shown in [c] for comparison.

express each theory as a complete probability model and compare the models on their ability to describe the entire dataset.

5.4 cluster-based method for preliminary analysis of functional data

We are not concerned with children's estimates of any particular number, but rather we wish to draw inferences about the overall pattern of their strategy. We assume that the position where a child estimates a number is a function of the number to be estimated, and the object of inference is the shape of that function.

In introductory statistics courses, we speak of categorical, ordinal, and scalar data. Functional data is another class of data. In general, even though the functional relationship between variables is our concern, it is difficult to directly observe the function. Instead, we usually infer the shape of the function from discrete observations taken at different points. In longitudinal data, we would take measurements of each individual at many time points. In this case, we ask each child to make an estimate of many different numbers.

An initial summary for quantitative data might include a histogram or a scatterplot. These plots provide a quick visual summary that displays overall trends and individual variation, they also provide a simple check for the appropriateness of a model. For example, a scatterplot quickly reveals that the relationship between two quantitative variables may not be linear, and that linear regression may be a poor model for the data.

Our goal is a visual summary of functional data that accomplishes these same objectives. We need to summarize all of the individual curves simultaneously. We need to display the variation between individuals and the overall trends. We would also like to highlight any subgroups within the data. Functional data analysis is presently an active area of research, but methods are computationally intensive and time consuming (Peng & Müller, 2008; Thompson & Rosen, 2003; Crambes, Kneip, & Sarda, 2009; Morris, Baladandayuthapani, Herrick, Sanna & Gutstein, 2011) Our visual summary method is loosely based on Ramsay and Silverman (2005). It is a simple, fast method and appropriate for preliminary analysis and data exploration.

The analysis has 3 flexible steps, and is appropriate for all types of functional data, including longitudinal analysis. We estimate each individual function, and then display the curves grouped by similar shape. Figure 5.2 illustrates that plotting all of the functions simultaneously gives an unclear picture; however, by plotingt the curves in clusters, we can more easily discern patterns.

In order to visually summarize a set of functional data, we create a plot showing clusters of functions with similar shape. This will allow us to view all of the data, display the overall trends and the variation between individuals. It also has the advantages of being very easy to interpret, and being easily created with built-in functions in the statistical software package R. There are only three steps to creating this visualization and summary: (1) Estimate a smooth function based on each individual's data. (2) Cluster discretized versions of the functions. (3) Plot the clusters so that they may be viewed together.

To describe this method fully, we introduce formal notation. Each individual i for $i = 1, \dots, N$ has measurements Y_{ij} taken at the points x_{ij} for $j = 1, \dots, M_i$. In longitudinal data, the x_{ij} represent the different time points at which measurements are taken. For the number line task, the x_{ij} represent the numbers that child i was asked to estimate, and the Y_{ij} represent the positions of their estimates.

We assume that the observations represent discrete measurements along a continuous function f_i, with errors ϵ_{ij}, so that we can write:

$$Y_{ij} = f_i(x_{ij}) + \epsilon_{ij}$$

The notation x_{ij} indicates that each individual may have been measured at different points, but the method described here does require that the x_{ij} span the same interval.

Step 1. Create a nonparametric estimate $\hat{f}_i$ of the function f_i for each individual i. One of the simplest ways to estimate a nonparametric regression function is with some form of linear smoother, such as kernel smoothing, local polynomial smoothing, basis smoothing, or smoothing splines. Tarpey (2007) demonstrates that any linear smoother will produce relatively similar results, so any of these smoothers can be used with subsequent steps.

For this data on numerical magnitude estimation, we need the estimates $\hat{f}_i$ to be accurate near 0, since the small numbers near 0 are where the theorized linear and logarithmic functions will differ the most. Many nonparametric smoothers are greatly influenced by data points near the boundary of the range of x values, this is known as boundary bias. Local linear smoothers have less boundary bias than some of the other methods and are easily implemented with the R function `loess` (Hastie et al., 2009). Therefore, for this data, local linear smoothing is the most appropriate method. For other applications with periodic data, basis smoothing with a Fourier basis would be more appropriate.

Data from different individuals may require different amounts of smoothing. For example, the second graders in this study have better fine motor skills than the kindergarteners, and thus have much smaller errors, ϵ_i, around their curves. Generalized Cross Validation (GCV) is an easily calculated goodness-of-fit measure for linear smoothers (Hastie et al., 2009). Thus, for this application, we chose to create the estimated functions $\hat{f}_i$ with local linear smoothing using GCV to choose an appropriate amount of smoothing.

Step 2. Cluster discretized versions of the $\hat{f}_i$. Let s be a fine grid of points covering the range of x. Evaluate the functions $\hat{f}_i$ on the grid s, to create the discrete vector $\hat{Y}_i$ representing individual i's smooth curve, $\hat{Y}_i = \hat{f}_i(s)$. We can then cluster the $\hat{Y}_i$ with a standard clustering method of our choice, since clustering with Euclidean distance on the $\hat{Y}_i$ is an approximation of clustering the f_i with an L^2 norm in the functional space (Ramsay and Silverman, 2005).

K-means is a very popular clustering method, but when the number of clusters is unknown, it can be difficult to make that choice with k-means. Gaussian mixture models are a more general class of models that includes k-means as a special case. The R implementation of Gaussian mixture model clustering `Mclust` in the package `mclust` automatically calculates Bayesian Information Criterion (BIC) scores, which can then be used to choose the number of clusters.

Step 3. The final step is to plot the clusters of individual curves. If desired, the clusters can be organized to make patterns more apparent. Results are shown in Figures 5.6, 5.7, and 5.8.

Additionally, we may discover outlier and noise patterns through this visualization. Clustering is well known to be sensitive to outliers. If we wish to remove curves that we deem to be noise or outliers, we should re-cluster and re-create the plot. The visible patterns may become more pronounced, or change very little after outliers are removed. We demonstrate this process in our discussion of the visualization results from the Siegler and Booth (2004) data.

5.5 cluster-based visualization results

5.5.1 *Data from Siegler and Booth, 2004*

The best clustering results, as selected by BIC, have 8 clusters (Figure 5.4). The cluster sizes are relatively small, and if the number of clusters and the shape of each cluster were an object of inference, this would be cause for serious concern. For visualizing patterns in the data, though, this is perfectly acceptable. The small clusters allow us to see individual curves more clearly, and make comparisons between the different shapes observed.

Clusters 1-6 all have mean curves that are monotonically increasing. Clusters 7 and 8 do not share this property: all of these curve are flat or decreasing. This previously unreported finding indicates that these students may have not understood the number line estimation task, and were simply plotting numbers at random. These are essentially 'noise' clusters.

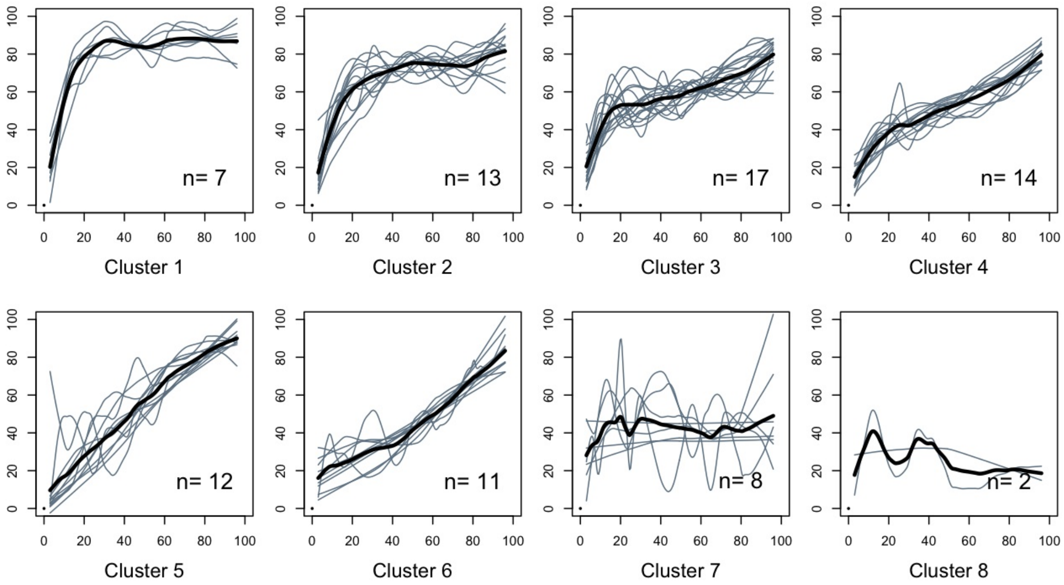

Figure 5.4: Clusters of curves from Siegler and Booth, 2004. Individual curves are shown in gray, the mean curve for each cluster is shown as a thicker black line.

In general, outlier curves will not be grouped together, and outliers can affect the shape of the cluster where they are grouped. Closer examination of cluster 5 reveals just such an outlier. Overall, the trend of cluster 5 is linear, but one child's curve starts near the point (0,80) then dips below the other curves in the cluster before joining them on the larger end of the scale. This is an individual that deserves a closer look. When we plot this child's raw data, in Figure 5.5, we see that this individual seems to recognize numbers greater than 50 as 'large', and near the high end of the 0-100 scale, but responds randomly for numbers less than 50. This is an interesting observation, but since this pattern is an outlier, we will omit this individual along with those in clusters 7 and 8 from subsequent analyses. Omitting these 11 participants leaves us with 73 participants from Siegler and Booth (2004).

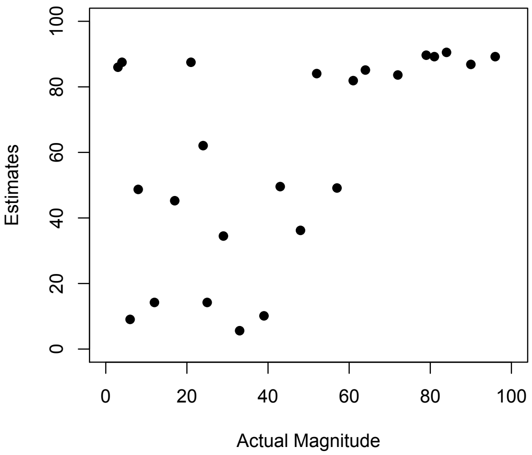

Figure 5.5: Individual data from a child with an outlier pattern. This child was a participant in Siegler and Booth 2004.

Simple visual inspection is usually insufficient to identify outlier curves with any amount of certainty. More robust and computationally intensive methods do exist to definitively identify outliers (Gervini, 2008, 2009). That is beyond the scope of this paper, instead we simply provide an intitial tool to summarize and visualize functional data prior to more rigorous analyses.

Figure 5.6 shows new clusters after outlier patterns are removed. It is coincidence that BIC again chooses 8 clusters. Cluster 1 appears to have a piecewise linear shape with 2 segments. These individual appear to estimate 0-20 on a strong linear scale, but do not differentiate larger quantities. They may simply be classifying any number greater than 20 as 'big' (Laski and Siegler, 2007). Cluster 2 may either be logarithmic or piecewise linear. Cluster 3 appears linear overall, but may have a different slope on 0-10, making it piecewise linear. Clusters 4 and 5 may have three linear segments, rather than the theorized two segments. We may speculate that these children estimate small numbers linearly, recognize numbers

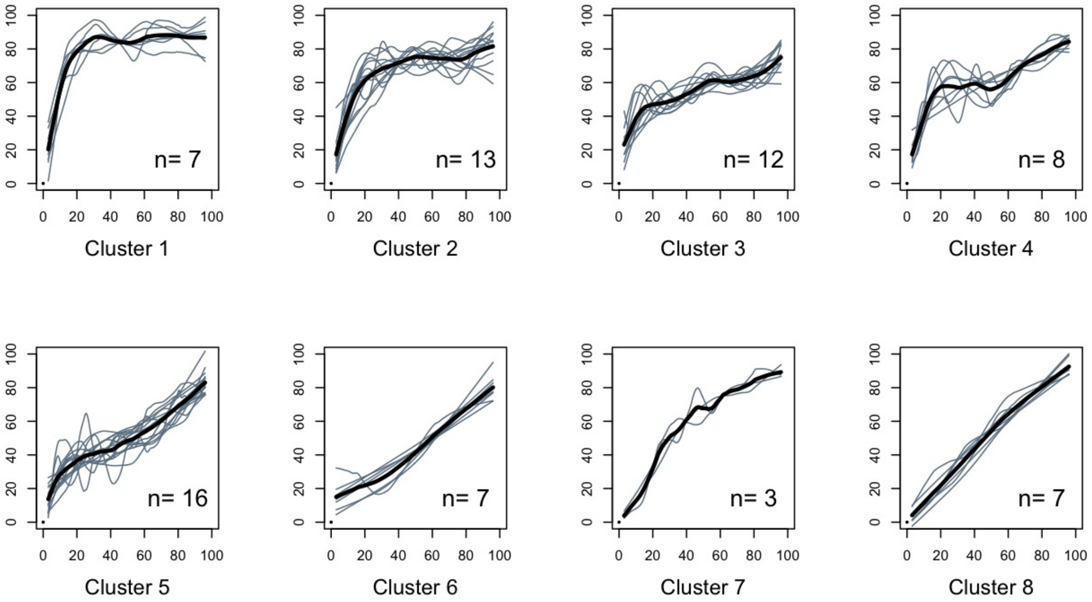

Figure 5.6: Clusters of curves from Siegler and Booth, 2004, after removing obvious outliers. Individual curves are shown in gray, the mean curve for each cluster is shown as a thicker black line.

bigger than 50 as ‘large’, but classify numbers between 20 and 50 as ‘the middle.’ Clusters 6, 7, and 8 are close to the line $y = x$, even though cluster 6 has a lower slope and cluster 7 has some slight curvature.

5.5.2 *Data from Booth and Siegler, 2006, Experiment 1*

For this second dataset on the 0-100 scale, we went through the same process of smoothing and clustering the complete dataset. This identified several flat curves and additional outliers. We removed the 6 children with outlier patterns, leaving 84 in the analysis. The final clusters are shown in Figure 5.7.

In many ways the clusters in Figure 5.7 are very similar to those found in Figure 5.6. Clusters 1, 2 and 4 are piecewise linear with a steep slope for small numbers and flat for larger numbers. Cluster 5 has a similar shape, but is more curved with no obvious corner, the theorized logarithmic estimation function may fit this cluster. We will need to perform more rigorous analysis to be certain whether the logarithmic function is appropriate. Clusters 8, 9 and 10 are very close to accurate linear estimation. Clusters 3, 6, and 7 however, have multiple inflection points. A piecewise linear curve with three segments may be able to describe these clusters.

Cluster 7, in particular, is very different from the clusters found in the Siegler and Booth (2004) data. In the previous experiment, the 4th cluster was flat between 20 and 60, indicating those children did not differentiate between ‘middle’ numbers. This cluster instead has a very steep slope between 40 and 60, which indicates that these students are making a strong differentiation near 50.

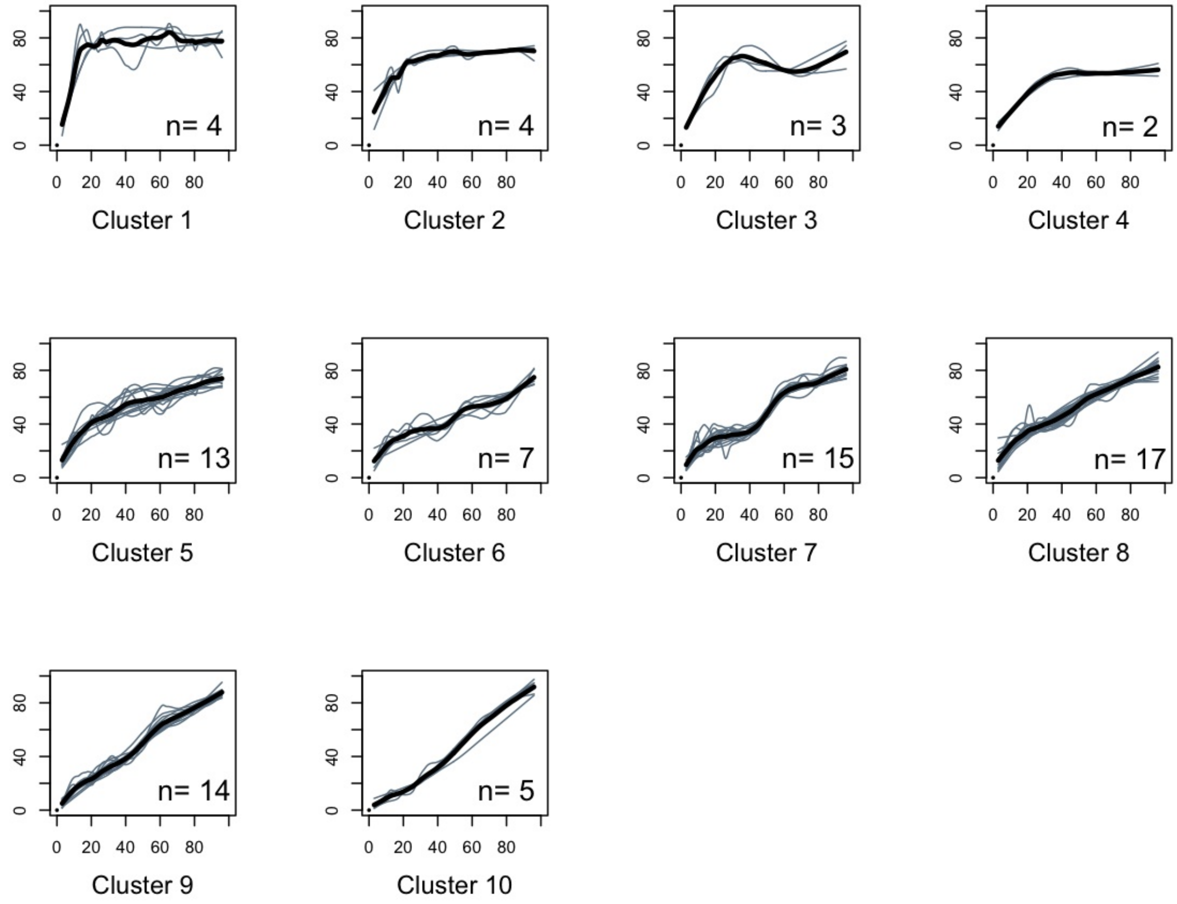

Figure 5.7: Clusters of individually smoothed curves, from Booth and Siegler, 2006, Experiment 1, after removing obvious outliers.

5.5.3 *Data on 0-1000 scale*

One of the benefits of this method is that we can combine multiple datasets easily, even if measurements were taken at different time points or for different inputs. Experiment 2 of Booth and Siegler, 2006; and Opfer and Siegler, 2007 both examined how children estimate numbers on the 0-1000 scale. However, the two different studies include two different sets of numbers to be estimated.

We can combine these datasets by first estimating a smooth curve for each individual in the combined dataset. Then we simply discretize these curves on the same grid of points so that we can cluster the curves as one dataset. Figure 5.8 shows the clusters after the flat curves and obvious outliers have been removed.

The clusters on the 0-1000 scale in Figure 5.8, show a similar pattern to the clusters on the 0-100 scale. Clusters 1, 2, 3, 4, 6, and 8 have a piecewise linear shape with 2 segments. Clusters 5, 7 and possibly 11 could be piecewise linear with 3 segments. Clusters 12 and 13 are very close to accurate linear estimation. Cluster 10 contains individuals who have an overall linear estimation strategy, but make one or two very large estimation errors. Cluster 9 shows a new pattern. It has only 2 children, who might be considered outliers; however, they have a very interesting estimation strategy. These children estimated numbers on (0,100) on a very steep linear scale, while numbers (200,1000) are estimated on a separate and somewhat accurate linear scale. This appears to be a piecewise linear estimation strategy, but unlike other participants, their estimation function is discontinuous.

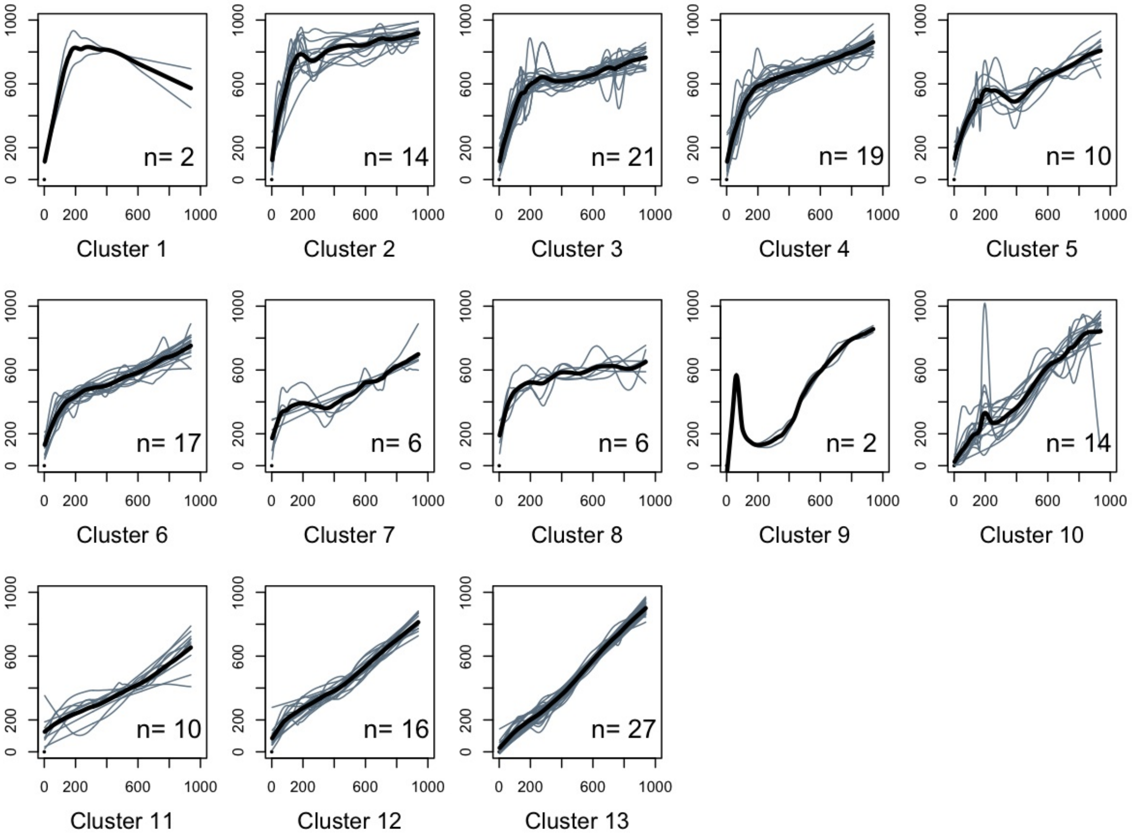

Figure 5.8: Clusters of individually smoothed curves for combined data from Booth and Siegler (2006) Experiment 2 and Opfer and Siegler (2007). Flat curves and obvious outliers have been removed.

5.6 model based analysis

In addition to the dominant theory of a logarithmic to linear representation shift (Dehaene et al., 2008; Siegler and Opfer, 2003; Siegler et al., 2009), another theory has recently emerged. This theory suggests that children use one linear scale to estimate small familiar numbers, and another linear scale to estimate larger numbers, so that the estimation function is piecewise linear with 2 segments (Ebersbach et al., 2008; Moeller et al., 2008).

The clusters in Figures 5.6, 5.7, and 5.8 demonstrate that in all four experiments, a substantial proportion of participants, about 80 children total, did not follow the pattern predicted by either of these existing theories. These graphs themselves suggest one possible model. A piecewise linear function with 3 segments may be appropriate to describe the behavior of the full dataset. So we will compare this third model to the two theory-based models.

Current practice is to compare each model on each individual participant separately. We require instead, a principled way to compare the ability of each theory to simultaneously describe all of the experimental data. Each theory predicts the shape of each child's individual estimation function and how children should vary from each other. We formalize this in hierarchical probability models so that we can compare each theory on the complete dataset.

5.6.1 *Probability model for the Linear-Logarithmic Representation Theory*

The analyses in Berteletti et al. (2010); Booth and Siegler (2006, 2008); Dehaene et al. (2008); Laski and Siegler (2007); Opfer and Siegler (2007); Opfer and Thompson (2008); Opfer and DeVries (2008); Siegler and Booth (2004, 2005); Siegler et al. (2009) all assume that every child uses either a linear strategy or a logarithmic strategy. This assumption corresponds to a two-class hierarchical model where each strategy is represented by one of the classes, and there is some variability between individuals in each class.

We will use the same notation as above, so that individual i provides an estimate Y_{ij} of the quantity X_{ij}. Let Z_i be a latent class indicator variable, so that $Z_i = 0$ if individual i uses the logarithmic strategy, and $Z_i = 1$ if individual i uses the linear strategy. We will place a flat prior on Z, so that $\Pr(Z_i = 1) = \frac{1}{2}$. Within each class, we specify a hierarchical regression model for the children that use that strategy. Thus for students using the logarithmic strategy $(Z = 0)$, we have a mean logarithmic strategy of: $Y = \alpha_0 + \alpha_1 \log(X)$ and each individual strategy has the form:

$$Y_{ij} = \beta_{0,i} + \beta_{1,i} \log(X_j) + \epsilon_{ij}$$

where $\beta \sim N(\alpha, \Sigma_\beta)$, and $\epsilon_{ij} \sim N(0, \sigma_i)$.

Similarly, for the linear strategy $(Z = 1)$, the mean linear strategy is: $Y = \gamma_0 + \gamma_1 X$ and each individual strategy has the form:

$$Y_{ij} = \delta_{0,i} + \delta_{1,i} X_j + \epsilon_{ij}$$

Once again, $\delta \sim N(\gamma, \Sigma_\delta)$, and $\epsilon_{ij} \sim N(0, \sigma_i)$.

In both strategy classes, we place a flat prior on α, γ, Σ_β, and Σ_δ. and model regression errors as $\epsilon_{ij} \sim N(0, \sigma_i)$ with a prior variance distribution of $\sigma_i \sim Inv - \chi^2(\nu, \tau^2)$. On the smaller scale, we set hyper-parameters as $\nu = 10$ and $\tau^2 = 7^2$, and on the larger scale , $\nu = 4$ and $\tau^2 = 80^2$. These weakly informative priors place the expected value of σ_i at a 10% error, but have a wide distribution. This reflects the belief that children will place ninety percent of their marks within 20% of where they intend to place the mark.

5.6.2 *Probability Model for 2-Piece Linear Theory*

Ebersbach et al. (2008) and Moeller et al. (2008) both propose that children estimate small numbers on one linear scale and larger numbers on a separate linear scale. In more formal terms, each child's estimation strategy is a piecewise linear function with 2 segments and there is some variability between individuals.

We use the endpoints of the linear segments to parameterize the hierarchical probability model for this strategy. The first segment has endpoints $(0, \alpha_{y_0})$ and $(\alpha_{x_1}, \alpha_{y_1})$. The second segment begins at $(\alpha_{x_1}, \alpha_{y_1})$, and ends at $(100, \alpha_{y_2})$ or $(1000, \alpha_{y_2})$ as appropriate.

The vector α parameterizes the population mean strategy, with individual strategies similarly parameterized by $\beta_i \sim N(\alpha, \Sigma_\beta)$. The parameters β define regression functions $r(x, \beta)$. Individual data are then modeled as

$$Y_{ij} = r(X_j, \beta_i) + \epsilon_{ij}$$

As in the linear-logarithmic model, regression errors are distributed as $\epsilon_{ij} \sim N(0, \sigma_i)$ with a prior variance distribution of $\sigma_i \sim Inv - \chi^2(\nu, \tau^2)$. On the smaller scale, the hyper-parameters are as $\nu = 10$ and $\tau^2 = 7^2$, and on the larger scale $\nu = 4$ and $\tau^2 = 80^2$.

We set the prior on α as uniform on the scale of the data. So for the 2004 data, we set $\alpha \sim \mathtt{Uniform}(0, 100)$, and so on. Lastly, we restrict Σ_β, the variance of individual strategies around the population mean strategy, to diagonal matrices; and set a wide prior distribution for each diagonal component as $Inv - \chi^2(\nu, \tau^2)$.

On the smaller 0-100 scale, $\nu = 2$ and $\tau^2 = 10^2$. On the 0-1000 scale, $\nu = 3$ and $\tau^2 = 80^2$.

3-piece Linear Model. This is the model suggested by the cluster-based visualization results. Each child's estimation function is represented by a piecewise linear function with 3 segments, and the individual functions are assumed to be normally distributed around a mean strategy.

5.6.3 *Probability Model for 3-Piece Linear Model*

Unlike the linear-logarithmic representation model and the 2-piece linear model, this model is not driven by psychological theory. Rather, this model is a mathematical device developed to describe patterns in the data that we observed in Section 5.5.

These patterns, if they are more than noise, cannot be described by the existing models. We consider this third model in order to determine if these previously undiscovered patterns are indeed real patterns, or if they are in fact just noise.

If we recognize that a piecewise linear function with three segments can also fit both a straight line, and a piecewise linear function with two segments, then it becomes clear that a piecewise linear function with 3 segments can represent all of the observed patterns in the data.

As with the 2-piece model, the endpoints of the linear segments parameterize the 3-piece model: $(0, \alpha_{y_0})$, $(\alpha_{x_1}, \alpha_{y_1})$, $(\alpha_{x_2}, \alpha_{y_2})$, and $(100, \alpha_{y_3})$ or $(1000, \alpha_{y_3})$. Again, α represents the population mean strategy and $\beta_i \sim N(\alpha, \Sigma_\beta)$ represents the individual strategies.

Individual estimates follow the regression model $Y_{ij} = r(X_j, \beta_i) + \epsilon_{ij}$, with errors modeled as $\epsilon_{ij} \sim N(0, \sigma_i)$. We use the same weakly informative Inverse-χ^2 prior for σ_i, with $(\nu, \tau^2) = (10, 7^2)$ on the 0-100 scale, and $(\nu, \tau^2) = (4, 80^2)$ on the larger scale.

The prior for α is again Uniform on the scale of the data; however, since there are two change points, we must require that $\alpha_{x_1} \leqslant \alpha_{x_2}$. Thus the prior on α_{x_2} is $\mathtt{Uniform}(\alpha_{x_1}, 100)$ or $\mathtt{Uniform}(\alpha_{x_1}, 1000)$, as appropriate.

As in the 2-piece model, we assume Σ_β is a diagonal matrix, and each component has an Inverse-χ^2 prior. On the 0-100 scale, $(\nu, \tau^2) = (2, 10^2)$, and $(\nu, \tau^2) = (3, 80^2)$ on 0-1000.

5.7 RESULTS AND DISCUSSION

We fit each of the three models to all 4 datasets via MCMC, and compared model fit using deviance information criterion (DIC) (Gelman, Carlin, Stern & Rubin, 2003). To verify DIC as an appropriate goodness-of-fit measure, we simulated data from the linear-logarithmic model. DIC selected the correct model for the simulation. DIC scores for data are shown in Table 5.2. Both of the piecewise linear models fit all four datasets better than the linear-logarithmic model by a substantial margin.

The best model on the 0-1000 scale is the 2-piece linear model. This corresponds to the theory that children estimate small, familiar numbers on one scale and estimate larger numbers on a different scale. Moeller et al. (2008) argued that this change-point occurred because of difficulty with the place value system. These results support that argument. Even the two outlier curves we observed in Clus-

Table 5.2: DIC scores for the three probability models on each of the 4 experiments, as well as data simulated from a logarithmic-linear model. Lower values indicate a better model fit. * indicates the best fitting model for each dataset.

	2004	**2006, Exp 1**	**2006, Exp 2**	**2007**	**Simulation**
Estimation Scale	0-100	0-100	0-1000	0-1000	0-100
3-piece model	13733*	16826*	14067	27560	11408
2-piece model	13873	16895	14039*	27498*	11331
Log-Linear model	15122	18686	15013	30293	1150*

ter 9 of Figure 5.8 have a non-linear pattern that appears to be due to difficulty integrating place value.

On the 0-100 scale, the 3-piece linear model better describes both datasets, probably due to the presence of intermediate estimation strategies. We observed one group of 8 students in the 2004 data who had a very flat slope in the middle of their curves, from 20-50, indicating a lack of differentiation of these middle numbers. Another group of 22 students in the 2006 data have a very steep slope around the half-way point, indicating an over-differentiation of the middle numbers. Both of these patterns are consistent with both a categorization strategy (Laski and Siegler, 2007) and a proportional reasoning strategy (Barth and Paladino, 2011).

We can further examine the posterior estimates of each child's strategy, as in Figure 5.9. The clusters are created on the change-points of the piecewise linear strategies for participants in Siegler and Booth (2004). This examination shows that as children grow, their strategies do become increasingly linear; however, we observe interesting patterns in the immature strategies. The immature strategy

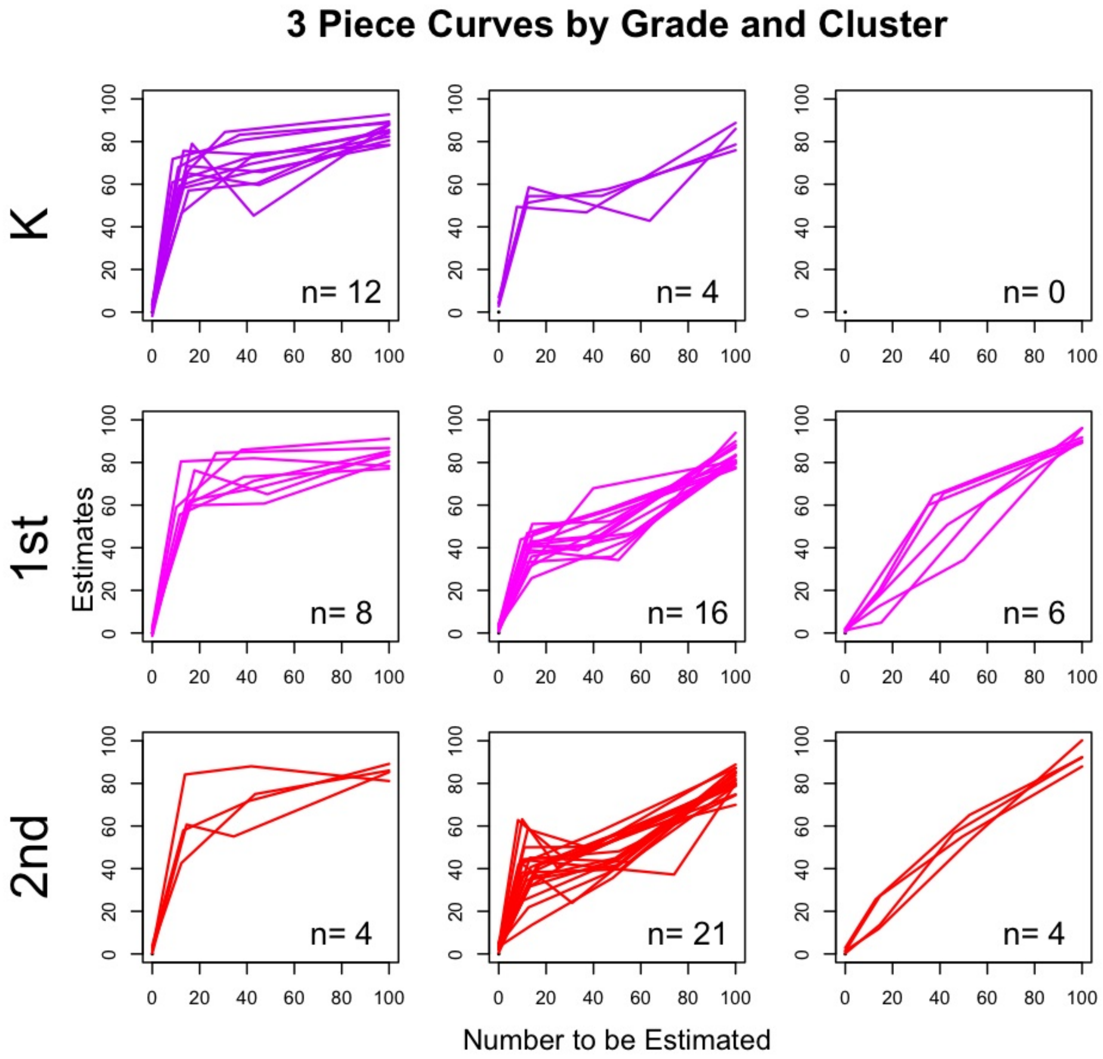

Figure 5.9: Posterior mean estimation strategies for participants from Siegler and Booth (2004), shown grouped by grade and strategy cluster. Clusters are shown in columns. The first row contains kindergarteners, the second row contains first graders and the bottom row contains second graders.

favored by the kindergarteners is linear on 0-10, or 0-20 for some children, larger numbers appear to be simply categorized as 'big' and plotted near the upper end of the number line. The intermediate strategy is the one used by the majority of first and second grade students, and appears to be some combination of linear estimation on 0-20, and categorization or proportional reasoning. The posterior estimates for Experiment 1 of Booth and Siegler (2006) show similar patterns of an immature strategy, a categorization/proportional strategy and a linear strategy.

As a caveat, we note that the experiments on the 0-1000 scale include only second- and fourth-graders. It is possible that if third-graders had been included in the study, the intermediate categorization/proportional reasoning strategy observed on the 0-100 scale would have been present on the larger scale as well.

The intermediate categorization/proportional reasoning strategy answers one of the critiques that Opfer, Siegler, and Young (2011) raised to Barth and Paladino (2011). Opfer et al. (2011) claims that changes in representation are responsible for the abrupt shifts in linearity observed in earlier experiments (Opfer and Thompson, 2008), and the gradual transition model argued for in Barth and Paladino (2011) cannot explain these changes. The patterns observed with thorough visual data summaries and the model-based analysis suggest that children use a mixture of strategies at all stages. The youngest children appear to combine linear estimation of familiar numbers with categorization of less familiar quantities, while older children appear to combine linear estimation, categorization, and/or proportional reasoning. The microgenetic changes that have been observed may be due to feedback activating additional strategies, and thus increasing linearity.

A detailed analysis of the raw data from several widely cited studies does not support the claim that children's understanding of the number line undergoes a transition from a logarithmic to a linear representation with development. Indeed, the analysis presented here reveals that the logarithmic-to-linear shift does not even hold on the data upon which the theory was established. The novel visual data summarization methods described in this paper enabled us to identify previously unrecognized patterns in the data. Children do estimate numerical magnitude with increasing linearity as they grow, but at all ages there is evidence for a combination of linear estimation, categorization and proportional reasoning strategies.

6

DISCUSSION AND DIRECTIONS FOR FUTURE WORK

This dissertation makes substantial contributions in two main areas. The first contribution is to establish the theoretical properties of mixed membership distributions. The second is the development of a model for assessing how much students use different strategies. We discuss each of these in turn.

6.1 MIXED MEMBERSHIP DISTRIBUTIONS

Thousands of applications of mixed membership exist today, the majority based on latent Dirichlet allocation. Each model exists individually with little regard for the general class of models; and while each model appears to stand quite well on its own, there is little understanding of how different choices in building the model will affect how the model functions. The work presented here addresses this gap.

Mixed membership models are constrained finite mixture models (Theorem 2.3). This result is the anchor for all the other results in Chapter 2.

Two interpretations are possible for MMMs, the between and the switching interpretation. We can always interpret an individual with mixed membership in multiple profiles as switching between the profiles, but only in special cases can we interpret an individual having a behavior that is between the profiles. The difference lies in the fact that MMMs are fundamentally FMMs. We know that a mixture of normal distributions is not normal, and we know that a mixture of Bernoulli distributions is also Bernoulli. This is why the interpretation of mixed membership is different for normal profiles than for Bernoulli profiles. It is only when a family of distributions is closed under mixture, and is a linear transformation of its parameters, that we can use the between interpretation in an MMM.

In the same way, the identifiability results in Section 2.4 rely entirely on expressing the MMM in FMM form. We characterize equivalence classes of mixed membership models by first recognizing when profiles generate the same FMM components, and then recognizing when the distribution of the membership parameter guarantees that the FMM components have the same probability.

Mixed membership can summarize data with far more parsimony than a finite mixture, the tradeoff is in the exchangeability assumption that variables are independent conditional on the membership parameter. The results in Chapter 2 are fundamental to understanding how this tradeoff will function in any given application.

6.2 FUTURE WORK ON MIXED MEMBERSHIP

6.2.1 *Identifiability*

Theorem 2.7 characterizes equivalence classes of identifiable MMMs. This result has profound implications for practice. For example, Pritchard et al. (2000) uses a symmetric Dirichlet distribution for the membership vector θ. Theorem 2.7 indicates that this is a poor choice for the distribution of θ. A symmetric distribution here leads to a maximal equivalence class. All possible $K!^{J-1}$ sets of MMM profiles will yield exactly the same data distribution.

The large class of possible equivalent models has two immediate implications for practice: First, when building a mixed membership model, use a non-symmetric distribution for the membership parameter. Second, when estimating the model, it is wise to estimate the distribution of the membership parameter as well.

Beyond this general advice, many open questions remain. Perhaps the first is, after estimating a mixed membership model, how do we recognize the equivalence class? Right now, I can only recommend examining the distribution of θ for exchangeable dimensions, and taking the uncertainty in that distribution into account. Therein lies the problem, it is unclear how uncertainty in the distribution of θ translates to uncertainty in the equivalence class of models.

The second open question lies in how to interpret the class of equivalent models. This is much more complicated than in a finite mixture model where re-indexing is the primary difficulty. Let us suppose that the membership parameter θ has 2 exchangeable dimensions, say θ_1 and θ_2. Then the two profiles

$$F_1 = \prod_{j=1}^{J} F_{1j} \qquad \text{and} \qquad F_2 = \prod_{j=1}^{J} F_{2j} \tag{6.1}$$

are interchangeable: I can swap any F_{1j} for F_{2j}, and the model remains the same. With only two exchangeable dimensions of θ, it is already difficult to characterize F_1 and F_2. How are we to summarize a larger set of exchangeable profiles, and how much will this summary depend on the application at hand?

6.2.2 *An Alternate Mixed Membership Model*

In Chapter 2, we discussed the differences between categorical and continuous data, and the differences in the between interpretation vs. the switching interpretation of mixed membership. The differences between the students in the numerical magnitude estimation application (Chapter 5) are not easily summarized by a mixed membership model.

This is again because the general mixed membership model is at its core a constrained finite mixture model. Each individual in the application has an individual regression function $r(x; \phi_i)$, parameterized by ϕ_i and the parameters come from a joint distribution $\phi_i \sim F_\phi$. This is a quintessential hierarchical regression model.

Yet in this application, a mixed membership interpretation would be undeniably useful. At one extreme of the individual regression functions is a mature linear estimation strategy. At the other extreme is an immature strategy that drastically overestimates small numbers, and treats all other numbers as 'large'. It would be preferable to model individuals who are learning the mature estimation strategy as having mixed membership in the two strategies. The obvious 'solution' is to use the two extreme strategies as the mixed membership basis profiles, yet this solution does not work.

As noted in Section 2.2.1, if the profile distributions are of the form $F_k(x; z) \sim N(r_k(z), \sigma_k^2)$, then the individual mixed membership distribution becomes:

$$X_{ij}|\theta_i, Z_{ij} \sim \sum_{k=1}^{K} \theta_i \, N\left(r_k(Z_{ij}), \sigma_k^2\right). \tag{6.2}$$

This is not a continuous regression function. It is not even a regression model, it's a mixture of regression models. At any given point Z_{ij}, the distribution of X_{ij} is a finite mixture model. A mixture of normals is not normal.

Using regression functions as mixed membership basis profiles is not a solution to this problem. Moreover, no change to the distribution of the membership parameter θ will resolve the issue. Changing θ changes the mixture parameters, the structure of the distribution is still fundamentally a mixture model.

The question we now consider is whether it is possible to propose a version of mixed membership that will allow us to interpret models with continuous data with the *between* interpretation, in the same way that we can with categorical data. To do so, we we must alter the individual-level assumption in the the mixed membership model.

In Erosheva's general mixed membership model (Erosheva, 2002), the primary individual-level assumption is that the distribution for a particular observation is an individual mixture model (Equation 1.1):

$$X_{ij}|\theta_i \sim \sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j) \tag{6.3}$$

We alter the mixed membership distribution in 2 ways. First, we require that all of the population profile distributions be from the same parametric family, so that $F_k(x) = F(x; \phi_k)$. Second, we replace Equation 1.1 with

$$X_{ij}|\theta_i \sim F\left(x_{ij}; \sum_k \theta_{ik}\phi_k\right) \tag{6.4}$$

Now, the individual distribution is in the same parametric family as the profiles, and moreover, the individual parameters lie in a simplex where the extreme points are the profile parameters ϕ_k.

To illustrate how this changes the mixed membership model, let us consider again the mixed membership regression model inspired by the Numerical Magnitude Estimation application. Using this alternate mixed membership model, we can now write the individual level model as:

$$X|\theta_i, Z \sim N\left(\sum_{k=1}^{K} \theta_{ik} r_k(Z), \sum_{k=1}^{K} \theta_{ik} \sigma_k^2\right). \tag{6.5}$$

Now, each child's strategy is described by an individual regression function;

$$r_i(x) = \sum_k \theta_{ik} r_k(z) \tag{6.6}$$

$$Y|\theta_i, X \sim N\left(r_i(X), \sigma_i^2\right). \tag{6.7}$$

Moreover, the individual regression functions are convex combinations of the profile functions, so that the individual function is quite literally between the profiles. Figure 6.1 shows a simple version of this model to illustrate how individual variation is captured. This figure is now analogous to the categorical logistic regression example in Figure 2.2. With this alternative mixed membership model, we can now summarize individual variation in continuous data in the same way that Erosheva's general mixed membership model allows us to summarize individual variation in categorical data.

This alternative mixed membership model has not yet been explored to determine its appropriateness and usefulness in different modeling situations. That is beyond the scope of this dissertation. I am raising the possibility of an alternate specification to highlight how the meaning of “mixed membership" may differ in different applications.

I note that this alternative mixed membership model may have similarities with other existing models and tools. Equation 6.5 is remarkably similar to functional data analysis (Ramsay and Silverman, 2005) and basis expansion of regression

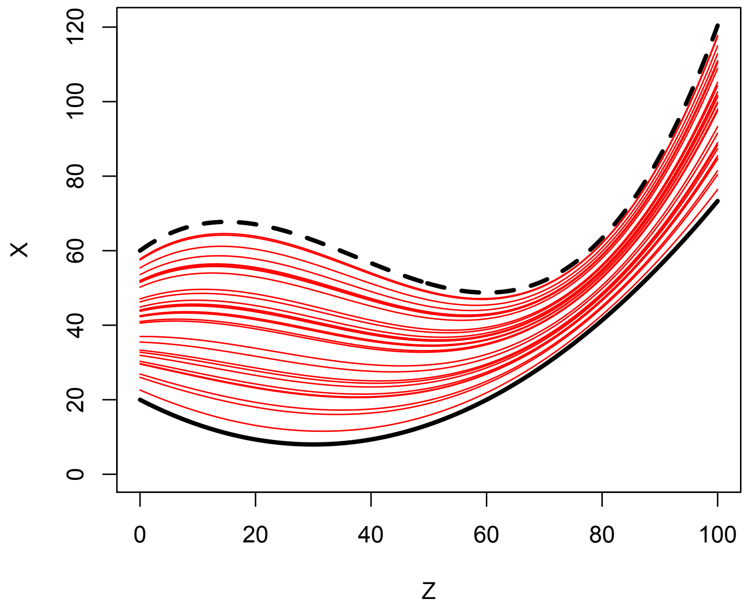

Figure 6.1: Illustration of a regression version of the alternate mixed membership model. The profile distributions are $F_k(x; z) = N(r_k(z), \sigma_k^2)$. The profile functions $r_k(z)$ are the solid and dashed black lines. The red lines in the middle represent the individual regression functions $r_i(x)$ which are a convex combination of the profile functions.

functions (Hastie et al., 2001). Basis expansion represents a function $f(x)$ in terms of a sum of basis functions $b_1(x), b_2(x), \ldots$

$$f(x) = \sum_m \beta_m b_m(x) \tag{6.8}$$

The biggest difference between Equations 6.6 and 6.8 is that $\sum_k \theta_{ik} = 1$, while there is generally no such restriction on $\sum_m \beta_m$. Nonetheless, should this alternate mixed membership model appear useful, the comparison with basis expansion methods may provide a useful tool for estimation and interpretation.

In the case where the profiles F_k are normal, and the parameters ϕ_k are constants rather than functions of covariates; the alternative mixed membership model in Equation 6.4 may reduce to a dimension-reduction technique such as principal components analysis, or a density estimation tool such as kernel density estimation (Hastie et al., 2001). These relationships have not been explored yet, but provide possible directions for future work.

6.3 multiple strategies model

The differences between novices and experts in a domain are the ideal targets for assessing what students know. One of the most crucial differences between novices and experts is how they approach problems, in the strategies they use to solve them. Yet these differences in knowledge are not easy to capture. If there are multiple ways to solve a problem correctly, then in order to distinguish strategies, we need data rich enough to capture the differences in strategies and a probability model that relates the multiple measures of student performance to the strategies.

The multiple strategies model presented in Chapter 3 is a revolutionary solution to this problem. MSM provides a flexible framework for incorporating any variables related to student achievement. This in itself is a remarkable accomplishment, as Wenger (2005) notes in his review of Van Breukelen (2005);

> If you were to browse through recent editions of the most prominent publications in the perceptual and cognitive sciences, you would find that, generally, the papers appearing in these journals concern themselves with patterns of either response choices or latencies. If you were somehow to be able to browse the peer reviews of these published

> works, you would also find that, in the majority of cases, reviewers were requesting information on latencies for those papers that focused on response probabilities, and information on response probabilities (e.g., error rates) for those submissions that focused on latencies. This is because the working consensus in these fields is that, to be interpretable, patterns in one of the variables need to be understood in the context of potentially important convergent or confounding patterns in the other.

Jointly modeling multiple measures of student knowledge is a significant and substantial achievement, but MSM accomplishes more. Very few psychometric models attempt to account for multiple solution strategies, and even fewer account for students who switch between strategies. Yet strategies are one of the dimensions which most strongly distinguishes expert and novice performance. Moreover, the path from novice to expert is not smooth. Individuals switch strategies from task to task, and may revert to immature strategies, even when they know expert strategies (Pellegrino et al., 2001). The multiple strategies model captures the differences in how experts and novices use strategies.

6.4 future work on multiple strategies

The multiple strategies model measures student performance along one of the dimensions which most strongly distinguishes expert and novice performance. Possible applications include everything from psychological studies of mental rotation tasks to large scale assessments of student knowledge. However, the application of the multiple strategies model in Chapter 4, represents a proof of concept. There is much more work to be done.

Some of the necessary work is described in Section 6.2. The multiple strategies model is built upon a mixed membership model. The identifiability and interpretability issues with mixed membership are also concerns for MSM. When two strategy profiles are exchangeable in the probability model, how do we make interpretable inferences about student knowledge?

Other facets of future work are unique to the multiple strategies model. In the application to the least common multiples assessment data, MSM was able to estimate both the strategies and how much students used each strategy for a very modestly sized data set. However, the problem was greatly simplified since each strategy was associated with a single skill. In this application we only needed to estimate which strategies a student uses, θ, and not the skills they possess, α.

In order to take full advantage of the promise of MSM, we need to be able to estimate both the atomic knowledge components, the skills parameterized by α_i, and the integrative knowledge components, the strategies parameterized by θ_i. I have not yet solved the problem of how to estimate both parameters simultaneously. One of the primary difficulties is that skills and strategies are inextricably intertwined. For example, depending on the strategies a student chooses, we may not observe the student use certain skills. The solution to this dilemma may lie in carefully constructing a joint distribution for θ and α.

Another possible direction for future work is to adapt the multiple strategies model into a dynamic or knowledge-tracing model. To accomplish this, we would need to develop a dynamic model of student performance where both θ and α change over time. A dynamic model of this sort could be integrated into a computer testing system, such as the formative assessment program ASSISTment

(Feng et al., 2010). Observing changes in strategy use over time, will provide a powerful tool for research in both psychology and education.

APPENDIX

A

MCMC R-CODE FOR MULTIPLE-STRATEGIES MIXED MEMBERSHIP MODEL IN CHAPTER 4

```
############################################
############################################
###### Full MCMC
############################################
############################################

load(’data_file_name.Rdata’) #load data
## This data file should include:
## K=number of profiles
## J=number of unique problems
## N=number of students
## test.data = matrix where each row is (i, j, C_{ij}, T_{ij1}, T_{ij2})
##
## This code assumes that each student may have seen a
##      different subset of the total items

library(MCMCpack)
library(mvtnorm)
folder.name = 'Save_folder'
        # folder where data will be saved.
        # create folder before you run code.

## MCMC parameters
b = 5 ## Thin the chain by only saving every bth iteration.
max.iter = 5000 # maximum number of iterations

eps.eta = 0.1 # tuning parameter for jumping distribution for eta/theta
        # MH proposal distribution: eta.star ~ N(eta, eps.eta*I)

alpha.1 = c(1,100) # shape & rate prior parameters for beta_{1} -> T_{i1}
alpha.2 = c(1, 10) # shape & rate prior parameters for beta_{2} -> T_{i2}
lambda.prior = rbind(c(10, 1), c(1,1), c(1,1))
        # parameters for beta-priors for lambda -> C_i)
        # row k is the prior for profile k
        # c(a,b)  corresponds to beta(a,b).


######################
## Parameter Initialization
######################
## can initialize from a previous iteration with:
## load('foldername/iter#.Rdata')

# mu, sigma, beta, lambda
# initialized with random starts
mu.0 = rnorm(K-1, 0, 2)
sigma.0 = diag(rchisq(K-1,df=10), nrow=K-1)
lambda.0 = rbeta(J, shape1=lambda.prior[1,1], shape2=lambda.prior[1,2])
for(k in 2:K){lambda.0 = cbind(lambda.0, rbeta(J, shape1=lambda.prior[k,1],
                shape2=lambda.prior[k,2]))}
beta.0 = rbind(rgamma(K, shape=alpha.1[1], rate=alpha.1[2]),
        rgamma(K, shape=alpha.2[1], rate=alpha.2[2]))

# # theta & Z
# # rows of Z.0 correspond to rows of test.data
##################################################################
# # Uncomment to use random starts for theta & Z
# # Note that convergence is slower with random starts.
# Z.0 = t(rmultinom(dim(test.data)[1], 1, rep(1/K,K)))
# eta.0 = rmvnorm(N, mean=rep(0, K-1), sigma = diag(1, K-1))
# theta.0 = exp(eta.0)/(1+apply(exp(eta.0), 1, sum))
# tmp = 1-apply(theta.0[,1:(K-1)], 1, sum)
# theta.0 = cbind(theta.0[, 1:(K-1)], tmp)
##################################################################
# Convergence is faster when initial Z’s are based on initial lambda’s
#       (using accuracy & ignoring response time)
# then initialize theta.0 at the mean of Z.0[i]
# slightly perturb the theta.0’s to avoid taking the log of 0.
Z.0 = matrix(NA, nrow=dim(test.data)[1], ncol=K)
theta.0 = matrix(NA, nrow=N, ncol=K)
eta.0 = matrix(NA, nrow=N, ncol=K-1)
```

```
for(i in 1:N){
        X.i = test.data[test.data[,1]==i,3]
        p.lam = (lambda.0[test.data[test.data[,1]==i,2],]^X.i)*
                (1-lambda.0[test.data[test.data[,1]==i,2],])^(1-X.i)
        Z.tmp = matrix(NA, nrow=length(X.i), ncol=K)
        for(j in 1:length(X.i)){
                Z.tmp[j,] = rmultinom(1,1,p.lam[j,])
                }
        Z.0[test.data[,1]==i,] = Z.tmp
        tmp = abs(apply(Z.tmp, 2, mean) + rnorm(K, 0, 0.001))
        theta.0[i,] = tmp/sum(tmp)
        eta.0[i,] = log(theta.0[i,]/theta.0[i,K])[-K]
        }

# Save starting values
iter = 0
save(mu.0, sigma.0, Z.0, eta.0, theta.0, lambda.0, beta.0,
        alpha.1, alpha.2, lambda.prior,
        file=paste0(folder.name,’/’,’iter’,iter,’.Rdata’))
        #priors are also saved in iter.0




######################
## MCMC
######################
```

```
for(iter in 1:max.iter){
#### Update beta
for(k in 1:K){
        T.1k = test.data[Z.0[,k]==1,4]
        beta.0[1,k] = rgamma(1, shape = alpha.1[1] + length(T.1k),
                rate = alpha.1[2] + sum(T.1k))
        T.2k = test.data[Z.0[,k]==1, 5]
        beta.0[2,k] = rgamma(1, shape = alpha.2[1] + length(T.2k),
                rate = alpha.2[2] + sum(T.2k))
        }

#### Update lambda
for(j in 1:J){
        shape1 = apply(Z.0[test.data[,2]==j, ]*
                test.data[test.data[,2]==j, 3], 2, sum)
        shape2 = apply(Z.0[test.data[,2]==j, ]*
                (1-test.data[test.data[,2]==j, 3]), 2, sum)
        lambda.0[j, ] = rbeta(K, shape1+ lambda.prior[,1],
                shape2+ lambda.prior[,2])
        }

for(i in 1:N){
        #### Update Z
        X.i = test.data[test.data[,1]==i,3]
        T.i1 = test.data[test.data[,1]==i,4]
        T.i2 = test.data[test.data[,1]==i,5]
```

```
p.C.log = log((lambda.0[test.data[test.data[,1]==i,2],]^X.i)*
        (1-lambda.0[test.data[test.data[,1]==i,2],])^(1-X.i))
p.T1.log = matrix(log(beta.0[1,]), nrow=length(T.i1),
          ncol=K, byrow=TRUE) -
        matrix(beta.0[1,], nrow=length(T.i1), ncol=K,
          byrow=TRUE)*T.i1
p.T2.log = matrix(log(beta.0[2,]), nrow=length(T.i1),
                ncol=K, byrow=TRUE) -
        matrix(beta.0[2,], nrow=length(T.i2), ncol=K,
                byrow=TRUE)*T.i2
p.log = p.C.log+p.T1.log+p.T2.log+
        matrix(log(theta.0[i,]), nrow=length(X.i), ncol=K, byrow=TRUE
            )
Z.tmp = matrix(NA, nrow=length(X.i), ncol=K)
for(j in 1:length(X.i)){
        Z.tmp[j,] = rmultinom(1,1,exp(p.log[j,]))
        }
Z.0[test.data[,1]==i,] = Z.tmp

#### Update theta (MH step)
#propose a new theta=f(eta)
eta.star.i = rmvnorm(1, mean=eta.0[i,], sigma=diag(eps.eta, nrow=K-1)
    )
tmp = exp(eta.star.i)/(1+sum(exp(eta.star.i)))
theta.star.i = c(tmp, 1-sum(tmp))
p.log.star.1 = sum(Z.tmp%*%log(theta.star.i))
p.log.0.1 = sum(Z.tmp%*%log(theta.0[i,]))
```

```
        p.log.star.2 = dmvnorm(eta.star.i, mu.0, sigma.0, log=TRUE)
        p.log.0.2 = dmvnorm(eta.0[i,], mu.0, sigma.0, log=TRUE)
        r.log = p.log.star.1 + p.log.star.2 - p.log.0.1 - p.log.0.2
        # accept/reject
        tmp = rbinom(1,1, prob = min(exp(r.log),1))
        if(tmp==1){theta.0[i,] = theta.star.i; eta.0[i,] = eta.star.i}
        }

#### Update mu & sigma
S = (N-1)*cov(eta.0)
sigma.0 = riwish(N-1,S)
mu.0 = rmvnorm(n=1, mean = apply(eta.0, 2, mean), sigma = sigma.0/N)

#### Save every bth iteration
if(round(iter/b) == iter/b) save(mu.0, sigma.0, Z.0, eta.0, theta.0, lambda
    .0,
        beta.0, file=paste0(folder.name,’/’,’iter’,iter,’.Rdata’))

}
```

B

R-CODE FOR CLUSTER-BASED VISUALIZATION IN CHAPTER 5

```
####
# make.smoothed.discrete is a function that returns discretized versions
# of smoothed functions for a functional data set
####
# this function is coded to take advantage of the fact that for the
# numerical magnitude estimation data, each individual estimated the
# same set of numbers x.
####
# The optimal smoothing bandwidth is chosen by generalized cross validation
####

make.smoothed.discrete = function(ests,x, s=seq(0,100, length=50),
        alpha = seq(0.2, 2, length=21), deg=1){
        ####
        # ests is a matrix
        # ests[i,j] is student i’s estimation of the jth quantity.
        # x is a vector, x[j] is the jth quantity to be estimated.
# s = grid of points where the smoothed functions will be evaluated
# alpha = vector of bandwidths over which smoothing is optimized.
# degree of local polynomial smoother (default is local linear smoothing)
####
N = dim(ests)[1] # number of students
n = dim(ests)[2] # number of quantities estimated

Y = matrix(NA, nrow=N, ncol=length(s))
BW = rep(NA, length=N)

for(i in 1:N){
    y = ests[i,]
    GCV = rep(NA, length(alpha))
    # computes GCV for each bandwidth in alpha
    for(a in 1:length(alpha)){
        sm = loess(y~x, span=alpha[a], degree=deg)
        GCV[a] = mean((sm$residuals/(1-sm$trace.hat/n))^2)
        }
    BW[i] = alpha[GCV==min(GCV)]
        #sets BW to the bandwidth with the minimum GCV
    sm = loess(y~x, span = BW[i], degree=deg)
    Y[i,] = predict(sm, s)
        # discretized, smooth function for student i
    }
return(list(Y=Y, BW = BW))
}
```

BIBLIOGRAPHY

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, September 2008. URL http://jmlr.csail.mit.edu/papers/volume9/airoldi08a/airoldi08a.pdf.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):pp. 139–177, 1982. ISSN 00359246. URL http://www.jstor.org/stable/2345821.

J. Aitchison. A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(1):pp. 136–146, 1985. ISSN 00359246. URL http://www.jstor.org/stable/2345555.

J. Aitchison and S. M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–72, 1980.

John R. Anderson. *Cognitive Psychology and Its Implications*. Worth Publishers, New York, NY, 7th edition edition, 2010.

Hilary C. Barth and Annie M. Paladino. The development of numerical estimation: evidence against a representational shift. *Developmental Science*, 14(1):125–135, 2011.

Ilaria Berteletti, Daniela Lucangeli, Manuela Piazza, Stanislas Dehaene, and Marco Zorzi. Numerical estimation in preschoolers. *Developmental Psychology*, 46(2): 545–551, 2010.

David Blei and Michael I. Jordan. Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 2003. ACM Press.

David Blei and John Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

David Blei and John Lafferty. A correlated topic model of science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

David Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1533-7928.

Julie L. Booth and Robert S. Siegler. Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, 41(6):189–201, 2006.

Julie L. Booth and Robert S. Siegler. Numerical magnitude representations influence arithmetic learning. *Child Development*, 79(4):1016–1031, 2008.

John D. Bransford, Ann L. Brown, and Rodney R. Cocking, editors. *How People Learn: Brain, Mind, Experience, and School*. National Academy Press, Washington, DC, 2000.

Christophe Crambes, Alois Kneip, and Pascal Sarda. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1):35–72, 2009.

Stanislas Dehaene, Veronique Izard, Elizabeth Spelke, and Pierre Pica. Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, 320:1217–1220, 2008.

Mirjam Ebersbach, Koen Luwel, Andrea Frick, Patrick Onghena, and Lieven Verschaffel. The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9-year old children: Evidence for a segmented linear model. *Journal of Experimental Child Psychology*, 99:1–17, 2008.

R. H. Klein Entink, J.-P. Fox, and Wim J. van der Linden. A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74(1):21–48, 2009.

Elena Erosheva. *Grade of Membership and Latent Structure Models With Application to Disability Survey Data*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA 15213, August 2002.

Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220–5227, 2004. doi: 10.1073/pnas.0307760101. URL http://www.pnas.org/content/101/suppl.1/5220.abstract.

Elena Erosheva, Stephen E. Fienberg, and Cyrille Joutard. Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, 1(2):502–537, 2007.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Vision and Pattern Recognition*, pages 524–531, 2005.

Mingyu Feng, Neil Heffernan, and Kenneth R. Koedinger. Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI journal)*, 19:243, 2010.

April Galyardt. Mixed membership models for continuous data. In *International Meeting of the Psychometric Society*, Athens, GA, July 2010.

Christian Geiser, Wolfgang Lehmann, and Michael Eid. Separating "rotators" from "nonrotators" in the mental rotations test: A multigroup latent class analysis. *Multivariate Behavioral Research*, 41(3):261–293, 2006.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in statistical science. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition edition, 2003.

Daniel Gervini. Robust functional estimation using the median and spherical principal components. *Biometrika*, 95(3):587–600, 2008.

Daniel Gervini. Detecting and handling outlying trajectories in irregularly sampled functional datasets. *Annals of Applied Statistics*, 3(4):1758–1775, 2009.

Mark Girolami and Ata Kaban. Sequential activity profiling: Latent dirichlet allocation of markov chains. *Data Mining and Knowledge Discovery*, 10:175–196, 2005.

Shelby J. Haberman. Book review of statistical applications using fuzzy sets. *Journal of the American Statistical Association*, 90(431):1131–1133, 1995.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2nd edition edition, 2009.

Trevor J. Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY, 2001.

Robert A. Henson, Jonathan L. Templin, and John T. Willse. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74(2):191–210, June 2009.

Brian W. Junker. Some statistical models and computational methods that may be useful for cognitively-relevant assessment. Technical report, Committee on the Foundations of Assessment, National Research Council, November 1999.

Brian W. Junker and Klaas Sijtsma. Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3):258–272, September 2001.

Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. The knowledge-learning-instruction (KLI) framework: Toward bridging the science-practice chasm to enhance robust student learning. Technical Report CMU-HCII-10-102, Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, 2010.

Elida V. Laski and Robert S. Siegler. Is 27 a big number? Correlation and causal connections among numerical categorization, number line estimation and numerical magnitude comparison. *Child Development*, 78(6):1723–1743, December 2007.

T. Loeys, Y. Rosseel, and K. Baten. A joint model for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, 76(3):487–503, 2011.

Marsha C. Lovett. Cognitive task analysis in service of intelligent tutoring systems design: A case study in statistics. In B. P. Goettl, H. M. Halff, C. L. Redfield, and V. J. Shute, editors, *Intelligent Tutoring Systems, Lecture Notes in Computer Science Volume*, volume 1452, pages 234–243. Springer, 1998.

Daniel Manrique-Vallier. *Longitudinal Mixed Membership Models with Applications to Disability Survey Data*. PhD thesis, Carnegie Mellon University, 2010.

Kenneth G. Manton, Max A. Woodbury, and H. Dennis Tolley. *Statistical Applications Using Fuzzy Sets*. John Wiley, New York, 1994.

Kenneth G. Manton, Xilang Gu, Hai Huang, and Mikhail Kovtun. Fuzzy set analyses of genetic determinants of health and disability status. *Statistical Methods in Medical Research*, 13:395, 2004.

Robert Mislevy and Norman D. Verhelst. Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55(2):195–215, 1990.

Korbinian Moeller, Silvia Pixner, Liane Kaufmann, and Hans-Christoph Nuerk. Children's early mental number line: Logarithmic or decomposed linear? *Journal of Experimental Child Psychology*, 103:503–515, 2008.

Jeffrey S. Morris, Veerabhadran Baladandayuthapani, Richard C. Herrick, Pietro Sanna, and Howard Gutstein. Automated analysis of quantitative image data using isomorphic functional mixed models with application to proteomics data. *Annals of Applied Statistics*, 5(2A):894–923, 2011.

John E. Opfer and Jeffery M. DeVries. Representational change and magnitude estimation: Why young children can make more accurate salary comparisons than adults. *Cognition*, 108(3):834–849, September 2008.

John E. Opfer and Robert S. Siegler. Representational change and children's numerical estimation. *Cognitive Psychology*, 55:169–195, 2007.

John E. Opfer and Clarissa A. Thompson. The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, 79(3):788–804, May/June 2008.

John E. Opfer, Robert S. Siegler, and Christopher J. Young. The powers of noise-fitting: reply to Barth and Paladino. *Developmental Science*, 14(5), 2011.

Philip Pavlik, Micheal Yudelson, and Kenneth R. Koedinger. Using contextual factors anlysis to explain transfer of least common multiple skills. In G. Biswas, S. Bull, J. Kay, and A. Mitrovic, editors, *Artificial intelligence in education*, volume 6738, pages 256–263. Springer Berlin / Heidelberg, 2011.

James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, editors. *Knowing What Students Know: The Science and Design of Educational Assessment*. National Academy Press, Washington, DC, 2001.

Jie Peng and Hans-Georg Müller. Distance-based clustering of sparsely observed stochastic processes with applications to online auctions. *Annals of Applied Statistics*, 2(3):1056–1077, 2008.

Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

Claudia Quaiser-Pohl, Anna M. Rohe, and Tobias Amberger. The solution strategy as an indicator of the developmental stage of preschool children's mental-rotation ability. *Journal of Individual Differences*, 31(2):95–100, 2010.

J.O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd edition edition, 2005.

Bethany Rittle-Johnson and Robert S. Siegler. Learning to spell: Variability, choice, and change in children's strategy use. *Child Development*, 70(2):332–348, 1999.

Simon Rogers, Mark Girolami, Colin Campbell, and Rainer Breitling. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):143–156, 2005.

Edward. E. Roskam. Models for speed and time-limit tests. In W. J. van der Linden and R. K. Hambleton, editors, *Handbook of modern item response theory*, pages 187–208. Springer, New York, 1997.

Jeffery N. Rouder. Are unshifted distributional models appropriate for response time? *Psychometrika*, 70(2):377–381, 2005.

Jeffery N. Rouder, Dongchu Sun, Paul L. Speckman, Jun Lu, and Duo Zhou. A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika*, 68(4):589–606, 2003.

Hanhuai Shan and Arindam Banerjee. Mixed-membership naive Bayes models. *Data Mining and Knowledge Discovery*, 23:1–62, 2011.

Robert S. Siegler. The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psycology: General*, 116(3):250–264, 1987.

Robert S. Siegler and Julie L. Booth. Development of numerical estimation in young children. *Child Development*, 75(2):428–444, March/April 2004.

Robert S. Siegler and Julie L. Booth. Development of numerical estimation: A review. In J. I. D. Campbell, editor, *Handbook of mathematical cognition*, pages 197–212. Psychology Press, Ltd., New York, 2005.

Robert S. Siegler and John E. Opfer. The development of numerical estimation: Evidence for multiple representations of numerical quantitiy. *Psychological Science*, 14(3):237–243, 2003.

Robert S. Siegler, Clarissa A. Thompson, and John E. Opfer. The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain and Education*, 3(3):142–150, 2009.

Thaddeus Tarpey. Linear transformations and the k-means clustering algorithm: Applications to clustering curves. *The American Statistican*, 61(1):34–40, 2007.

Wesley K. Thompson and Ori Rosen. A Bayesian model for sparse functional data. *Biometrics*, 64:54–63, 2003.

L. L. Thurstone. Ability, motivation and speed. *Psychometrika*, 2(4):249–254, 1937.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Chichester, 1985.

Gerard J. P. Van Breukelen. Psychometric modeling of response speed and accuracy with mixed and conditional regression. *Psychometrika*, 70(2):359–376, 2005.

Wim J. van der Linden. A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3):287–308, 2007.

Wim J. van der Linden. Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3):247–272, 2009.

Michael J. Wenger. Models for the statistics and mechanisms of response speed and accuracy. *Psychometrika*, 70(2), 2005.

Max A. Woodbury, Jonathan Clive, and Aruthur Garson, Jr. Mathematical typology: A grade of membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11:277–298, 1978.