3

Interpreting Mixed Membership Models: Implications of Erosheva's Representation Theorem

April Galyardt

Department of Educational Psychology, University of Georgia, Athens, GA 30602, USA

CONTENTS

The original three mixed membership models all analyze categorical data. In this special case there are two equivalent interpretations of what it means for an observation to have mixed membership. Individuals with mixed membership in multiple profiles may be considered to be 'between' the profiles, or they can be interpreted as 'switching' between the profiles. In other variations of mixed membership, the between interpretation is inappropriate. This chapter clarifies the distinction between the two interpretations and characterizes the conditions for each interpretation. I present a series of examples that illustrate each interpretation and demonstrate the implications for model fit. The most counterintuitive result may be that no change in the distribution of the membership parameter will allow for a between interpretation.

3.1 Introduction

The idea of mixed membership is a simple, intuitive idea. Individuals in a population may belong to multiple subpopulations, not just a single class. A news article may address multiple topics rather than fitting neatly in a single category (Blei et al., 2003). Patients sometimes get multiple diseases at the same time (Woodbury et al., 1978). An individual may have genetic heritage from multiple subgroups (Pritchard et al., 2000; Shringarpure, 2012). Children may use multiple strategies in mathematics problems rather than sticking to a single strategy (Galyardt, 2012).

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.

Erosheva (2002) and Erosheva et al. (2007) also showed that every mixed membership model has an equivalent finite mixture model representation. The proof in Erosheva (2002) shows that the relationship holds for categorical data; Erosheva et al. (2007) indicates that the same result holds in general.

The behavior of mixed membership models is best understood in the context of this representation theorem. The shape of data distributions, the difference between categorical and continuous data, possible interpretations, and identifiability all flow from the finite mixture representation (Galyardt, 2012). This chapter describes the general mixed membership model and then explores the implications of Erosheva's representation theorem.

3.2 The Mixed Membership Model

Due to the history of mixed membership models, and the fact that they were independently developed multiple times, there are now two common and equivalent ways to define mixed membership models. The generative model popularized by Blei et al. (2003) is more intuitive so we will discuss it first, followed by the the general model (Erosheva, 2002; Erosheva et al., 2004).

3.2.1 The Generative Process

The generative version of mixed membership is the more common representation in the machine learning community. This is due largely to the popularity of latent Dirichlet allocation (LDA) (Blei et al., 2003), which currently has almost 5000 citations according to Google Scholar. LDA has inspired a wide variety of mixed membership models, e.g., see Fei-Fei and Perona (2005), Girolami and Kaban (2005), and Shan and Banerjee (2011), though these models still fit within the general mixed membership model of Erosheva (2002) and Erosheva et al. (2004).

The foundation of the mixed membership model is the assumption that the population consists of K profiles, indexed $k = 1, \ldots, K$, and that each individual $i = 1, \ldots, N$ belongs to the profiles in different degrees. 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 original populations that have melded into the current population. 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, each profile can represent a different strategy.

Each individual has a membership vector, $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})$, that indicates the degree to which they belong to each profile. The term *individual* here simply refers to a member of the population and 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. For example, if student i used strategies 1 and 2, each about half the time, then this student would have a membership vector of $\theta_i = (0.5, 0.5, 0, ..., 0)$. Similarly, if an image was 40% water and 60% mountain then this would be indicated by θ_i.

Each observed variable X_j, $j = 1, \ldots, J$ has a different probability distribution within each profile. For example, in an image processing application, the water profile 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. Note that X_j may be univariate or be multidimensional itself, and that we may observe $r = 1, \ldots, R_{ij}$ replications of X_j for each individual i, denoted X_{ijr}. The distribution of X_j within profile k is given by the cumulative distribution function (cdf) F_{kj}.

We introduce the indicator vector Z_{ijr} to signify which profile individual i followed for replication r of the jth variable. For example, in textual analysis, Z_{ijr} would indicate which topic the rth word in document i came from. In genetics, Z_{ijr} indicates which founding population individual i inherited the rth copy of their jth allele from.

The membership vector θ_i indicates how much each individual belongs to each profile so that $Z_{ijr} \sim Multinomial(\theta_i)$. We will write Z_{ijr} in the form that, if individual i followed profile k for replication r of variable j, then $Z_{ijr} = k$. The distribution of X_{ijr} given Z_{ijr} is then

$$X_{ijr}|Z_{ijr} = k \quad \sim \quad F_{kj}. \tag{3.1}$$

The full data generating process for individual i is then given by:

1. Draw $\theta_i \sim D(\theta)$.
2. For each variable $j = 1, \ldots, J$:
   (a) For each replication $r = 1, \ldots, R_{ij}$:
      i. Draw a profile $Z_{ijr} \sim Multinomial(\theta_i)$.
      ii. Draw an observation $X_{ijr} \sim F_{Z_{ijr},j}(x_j)$ from the distribution of X_j associated with the profile Z_{ijr}.

3.2.2 General Mixed Membership Model

The general mixed membership model (MMM) makes explicit the assumptions that are tacit within the general model. These assumptions are collected into four layers of assumptions: population level, subject level, sampling scheme, and latent variable level.

The ***population level*** assumptions are that there are K different profiles within the population, and each has a different probability distribution for the observed variables F_{kj}.

The ***subject level*** assumptions begin with the individual membership parameter θ_i that indicates which profiles individual i belongs to. We then assume that the conditional distribution of X_{ij} given θ_i is:

$$F(x_j|\theta_i) \quad = \quad \sum_{k=1}^{K} Pr(Z_{ijrk} = 1|\theta_i) F(x_j|Z_{ijrk} = 1), \tag{3.2}$$

$$\quad = \quad \sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j). \tag{3.3}$$

Equation (3.3) is the result of combining Steps 2(a)i and 2(a)ii in the generative process. Z_{ijr} is simply a data augmentation vector, and we can easily write the distribution of the observed data without it. Notice that Step 2 of the generative process assumes that the X_{ijr} are independent given θ_i. In psychometrics this is known as a local independence assumption. This exchangeability assumption allows us to write the joint distribution of the response vector $X_i = (X_{i1}, ..., X_{iJ})$, conditional on θ_i as

$$F(x|\theta_i) = \prod_{j=1}^{J} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j) \right]. \tag{3.4}$$

This conditional independence assumption also contains the assumption that the profile distributions are themselves 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,

$$F(x|\theta_{ik} = 1) = \prod_j F_{kj}(x_j) = F_k(x). \tag{3.5}$$

The ***sampling scheme level*** includes the assumptions about the observed replications. Step 2(a) of the generative process assumes that replications are independent given the membership vector θ_i. Thus the individual response distribution becomes:

$$F(x|\theta_i) = \prod_{j=1}^{J} \prod_{r=1}^{R_{ij}} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_{j_r}) \right]. \tag{3.6}$$

Note that Equations (3.3), (3.4), and (3.6) 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.

The ***latent variable level*** corresponds to Step 1 of the generative process. 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 (3.6) over the distribution of θ, yielding:

$$F(x) = \int \prod_{j=1}^{J} \prod_{r=1}^{R_{ij}} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j) \right] d\mathrm{D}(\theta). \tag{3.7}$$

The final layer of assumptions about the latent variable θ is crucial for purposes of estimation, but it is unimportant for the discussion of mixed membership model properties in this chapter. All of the results presented here flow from the exchangeability assumption in Equation (3.4), and hold whether we use Equation (3.6) or (3.7) for estimation.

3.3 The Development of Mixed Membership

Independently, Woodbury et al. (1978), Pritchard et al. (2000), and Blei et al. (2003) developed remarkably similar mixed membership models to solve problems in three very different content areas.

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 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 the likelihood in Equation (3.4) is used for estimation.

Population Admixture Model

Pritchard et al. (2000) models the genotypes of individuals in a heterogeneous population. The profiles represent distinct populations of origin from which individuals in the current population have inherited their genetic makeup.

The variables X_j are the genotypes observed at J locations, and for diploid individuals two replications are observed at each location ($R_j = 2$). 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. It is important to note that a symmetric Dirichlet distribution will result in an identifiability problem that is not present when θ has an asymmetric distribution (Galyardt, 2012).

One interesting feature 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 models are estimated with no information about what the profiles may be and no information about which individuals may have some membership in the same profiles. Pritchard et al. (2000) considers the unsupervised case, but also considers the case where there is additional information. In this application, the location where an individual bird was captured means that it is likely a descendent of a certain population with a lower probability that it descended from an immigrant. This information is included with a carefully constructed prior on θ, which also incorporates rates of migration.

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. LDA 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 present in the corpus. LDA models only one variable, the words present in the documents ($J = 1$). 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 of word w appearing in topic k. LDA uses the integrated likelihood in Equation (3.7). The focus here is on estimating the topic profiles, and the distribution of membership parameters, rather than the θ_i themselves. LDA also uses a Dirichlet distribution for θ, however it does not use a *symmetric* Dirichlet, and so it avoids the identifiability issues that are present in the admixture model (Galyardt, 2012).

3.3.1 Variations of Mixed Membership Models

Variations of mixed membership models fall into two broad groups: The first group alters the distribution of the membership parameter θ, the second group alters the profile distributions F_{kj}.

Membership Parameters

The membership vector θ is non-negative and sums to 1 so that it lies within a $K - 1$ dimensional simplex. The two most popular distributions on the simplex are the Dirichlet and the logistic-normal.

Both LDA and the population admixture model use a Dirichlet distribution as the prior for the membership parameter. This is the obvious choice when the data is categorical, since the Dirichlet distribution is a conjugate prior for the multinomial. However, 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, 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) uses a logistic-normal distribution for θ. Blei and Lafferty (2006) takes this idea a step further and creates a dynamic model where the mean of the logistic-normal distribution evolves over time.

Fei-Fei and Perona (2005) analyzes 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 (2005) addresses this by giving the membership parameters a distribution that is a mixture of Dirichlets.

Profiles

In all three of the original models, the data are categorical and the profile distributions F_{kj} are multinomial. More recently, we have seen a variety of mixed membership models for data that is not categorical, with different parametric families for the F_k distributions.

Latent process decomposition (Rogers et al., 2005) describes the different processes that might be responsible for different levels of gene expression observed in microarray datasets. 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.

The mixed membership naive Bayes model (Shan and Banerjee, 2011) is another extension of LDA which seeks to define a 'generalization' of LDA. This model simply requires the profile distributions F_{kj} to be exponential family distributions. This is a subset of models that falls within Erosheva's general mixed membership model (Erosheva et al., 2004). Moreover, other exponential family profile distributions will not have the same properties as the multinomial profiles used in LDA (Galyardt, 2012). The main contribution of Shan and Banerjee (2011) is a comparison of different variational estimation methods for particular choices of F_{kj}.

3.4 The Finite Mixture Model Representation

Before we discuss the relationship between mixed membership models (MMM) and finite mixture models (FMM), we will briefly review FMM.

3.4.1 Finite Mixture Models

Finite mixture models (FMM) go by many different names, such as "latent class models" or simply "mixture models," and they are used in many different applications from psychometrics to clustering and classification.

The basic assumption is that within the population there are different subgroups, $s = 1, \ldots, S$, which may be called clusters or classes depending on the application. Each subgroup has its own distribution of data, $F_s(x)$, and each subgroup makes up a certain proportion of the population, π_s.

The distribution of data across the population is then given by:

$$F(x) = \sum_{s=1}^{S} \pi_s F_s(x). \tag{3.8}$$

For reference, the distribution of data over the population in a MMM, given by Equation (3.7), is:

$$F(x) = \int \prod_{j=1}^{J} \prod_{r=1}^{R_{ij}} \left[\sum_{k=1}^{K} \theta_{ik} F_{kj}(x_j) \right] d\mathrm{D}(\theta). \tag{3.9}$$

Finite mixture models can be considered a special case of mixed membership models. In a mixed membership model, the membership vector θ_i indicates how much individual i belongs to each of the profiles k, thus θ lies in a $K-1$ dimensional simplex. If the distribution of the membership parameter θ is restricted to the corners of the simplex, then θ_i will be an indicator vector and Equation (3.9) will reduce to the form of Equation (3.8). So a finite mixture model is a special case of mixed membership with a particular distribution of θ.

3.4.2 Erosheva's Representation Theorem

Even though FMM is a special case of MMM, every MMM can be expressed in the form of an FMM with a potentially much larger number of classes. Haberman (1995) suggests this relationship in his review of Manton et al. (1994). Erosheva et al. (2007) shows that it holds for categorical data and indicates that the same result holds in the general case as well. Here the theorem is presented in a general form.

Before we consider the formal version of the theorem, we can build some intuition based on the generative version of MMM. In the generative process, to generate the data point X_{ijr} for individual i's replication r of variable j, we first draw an indicator variable $Z_{ijr} \sim Multinomial(\theta_i)$ that indicates which profile X_{ijr} will be drawn from. Let us write Z_{ijr} in the form: $Z_{ijr} = k$, if X_{ijr} was drawn from profile k. Effectively, Z indicates that individual i 'belongs' to profile k for observation j_r.

The set of all possible combinations of Z defines a set of FMM classes, which we shall write as $\mathcal{Z} = \{1, \ldots, K\}^R$, where R is the total number of replications of all variables. For individual i, let $\zeta_i = (Z_{i11}, \ldots, Z_{iJR_J}) \in \mathcal{Z}$. So ζ_i indicates which profile an individual belongs to for each and every observed variable.

Representation Theorem. *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

$$F_{\zeta}^{FMM}(x) = \prod_{r=1}^{R} F_{\zeta_r, r}(x_r) \tag{3.10}$$

and the probability associated with each class ζ is

$$\pi_{\zeta} = \mathbb{E}\left[\prod_{r=1}^{R} \theta_{\zeta_r} \right]. \tag{3.11}$$

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

$$\begin{aligned} F(x|\theta_i) &= \prod_r \sum_k \theta_{ik} F_{kr}(x_r), && (3.12) \\ &= \sum_{\zeta \in \mathcal{Z}} \prod_r \theta_{i\zeta_r} F_{\zeta_r r}(x_r). && (3.13) \end{aligned}$$

Equation (3.13) reindexes the terms of the finite sum when Equation (3.12) is expanded. Distributing the product over r yields Equation (3.14):

$$\begin{aligned} F(x|\theta_i) &= \sum_{\zeta \in \mathcal{Z}} \left(\left[\prod_r \theta_{i\zeta_r} \right] \left[\prod_r F_{\zeta_r r}(x_r) \right] \right), && (3.14) \\ &= \sum_{\zeta \in \mathcal{Z}} \pi_{i\zeta} F_\zeta(x). && (3.15) \end{aligned}$$

Integrating Equation (3.15) yields the form of a finite mixture model:

$$F(x) = \mathbb{E}_\theta \left[\sum_{\zeta \in \mathcal{Z}} \pi_{i\zeta} F_\zeta(x) \right] = \sum_{\zeta \in \mathcal{Z}} \pi_\zeta F_\zeta(x). \qquad (3.16)$$

□

Erosheva's representation theorem states that if a mixed membership model needs K profiles to express the diversity in the population, an equivalent finite mixture model will require K^R components. In addition, if we compare Equation (3.15) to Equation (3.16), then we see that each individual's distribution is also a finite mixture model, with the same components as the population FMM but with individual mixture proportions.

The mixed membership model is a much more efficient representation for high-dimensional data—we need only K profiles instead of K^R. However, there is a tradeoff in the constraints on the shape of the data distribution (Galyardt, 2012). The rest of this chapter will explore some of these constraints.

3.5 A Simple Example

A finite mixture model is described by the components of the mixture F_ζ and the proportion associated with each component, π_ζ. The representation theorem tells us that when a MMM is expressed in FMM form, the components are completely determined by MMM profiles (Equation 3.10), and that the proportions are completely determined by the distribution of the membership vector θ (Equation 3.11).

We can think of the MMM profiles F_{kj} as forming a *basis* for the FMM components F_ζ. Consider a very simple example with two dimensions ($J = 2$) and two profiles ($K = 2$). Suppose that the first profile has a uniform distribution on the unit square and the second profile has a concentrated normal distribution centered at (0.3, 0.7):

$$\begin{aligned} F_1(x) &= F_{11}(x_1) \times F_{12}(x_2) &= Unif(0,1) \times Unif(0,1), && (3.17) \\ F_2(x) &= F_{21}(x_1) \times F_{22}(x_2) &= N(0.3, 0.1) \times N(0.7, 0.1). && (3.18) \end{aligned}$$

From a generative perspective, an individual with membership vector $\theta_i = (\theta_{i1}, \theta_{i2})$ will have $Z_{i1} = 1$ with probability θ_{i1} and $Z_{i1} = 2$ with probability θ_{i2}, so that $X_{ij} \sim Unif(0,1)$ with probability θ_{i1}, and $X_{ij} \sim N(0.3, 0.1)$ with probability θ_{i2}. Similarly, for variable $j = 2$, with probability θ_{i1}, $Z_{i2} = 1$, and with probability θ_{i2}, $Z_{i2} = 2$. In total, there are $K^J = 4$ possible combinations of $\zeta_i = (Z_{i1}, Z_{i2})$:

$$X_i|\zeta_i = (1,1) \sim Unif(0,1) \times Unif(0,1), \tag{3.19}$$

$$X_i|\zeta_i = (1,2) \sim Unif(0,1) \times N(0.7,0.1), \tag{3.20}$$

$$X_i|\zeta_i = (2,1) \sim N(0.3,0.1) \times Unif(0,1), \tag{3.21}$$

$$X_i|\zeta_i = (2,2) \sim N(0.3,0.1) \times N(0.7,0.1). \tag{3.22}$$

Equations (3.19)–(3.22) are the four FMM components for this MMM model, F_ζ (Figure 3.1), and they are formed from all the possible combinations of the MMM profiles F_{kj}. It is in this sense that the MMM profiles form a basis for the data distribution.

The membership parameter θ_i governs how much individual i 'belongs' to each of the MMM profiles. If $\theta_{i1} > \theta_{i2}$, then $\zeta_i = (1,1)$ is more likely than $\zeta_i = (2,2)$. Notice, however, that since multiplication is commutative, $\theta_{i1}\theta_{i2} = \theta_{i2}\theta_{i1}$, so that $\zeta_i = (1,2)$ always has the same probability as $\zeta_i = (2,1)$.

Figure 3.2 shows the data distribution of this MMM for two different distributions of θ. The change in the distribution of θ affects only the probability associated with each component. Thus the MMM profiles define the modes of the data, and the distribution of θ controls the height of the modes.

Alternate Profiles

Consider an alternate set of MMM profiles, G:

$$G_1(x) = Unif(0,1) \times N(0.7,0.1), \tag{3.23}$$

$$G_2(x) = N(0.3,0.1) \times Unif(0,1). \tag{3.24}$$

The G profiles are essentially a rearrangement of the F profiles, and will generate exactly the same FMM components as the F profiles (Figure 3.3). For any MMM model, there are $K!^{(J-1)}$ sets of basis profiles which will generate the same set of components in the FMM representation (Galyardt, 2012). The observation that multiple sets of MMM basis profiles can generate the same FMM components has implications for the identifiability of MMM, which is explored fully in Galyardt (2012).

Multivariate X_j

The same results hold when X_j is multivariate. Consider an example where each profile F_{kj} is a multivariate Gaussian, as used in the GM-LDA model in Blei and Jordan (2003). Then we can write the profiles as:

$$\begin{aligned} F_1(x) &= F_{11}(x_1) \times F_{12}(x_2) = MvN(\mu_{11}, \Sigma_{11}) \times MvN(\mu_{12}, \Sigma_{12}), \\ F_2(x) &= F_{21}(x_1) \times F_{22}(x_2) = MvN(\mu_{21}, \Sigma_{21}) \times MvN(\mu_{22}, \Sigma_{22}). \end{aligned}$$

The corresponding FMM components are then:

$$X_i|\zeta_i = (1,1) \sim MvN(\mu_{11}, \Sigma_{11}) \times MvN(\mu_{12}, \Sigma_{12}), \tag{3.25}$$

$$X_i|\zeta_i = (1,2) \sim MvN(\mu_{11}, \Sigma_{11}) \times MvN(\mu_{22}, \Sigma_{22}), \tag{3.26}$$

$$X_i|\zeta_i = (2,1) \sim MvN(\mu_{21}, \Sigma_{21}) \times MvN(\mu_{12}, \Sigma_{12}), \tag{3.27}$$

$$X_i|\zeta_i = (2,2) \sim MvN(\mu_{21}, \Sigma_{21}) \times MvN(\mu_{22}, \Sigma_{22}). \tag{3.28}$$

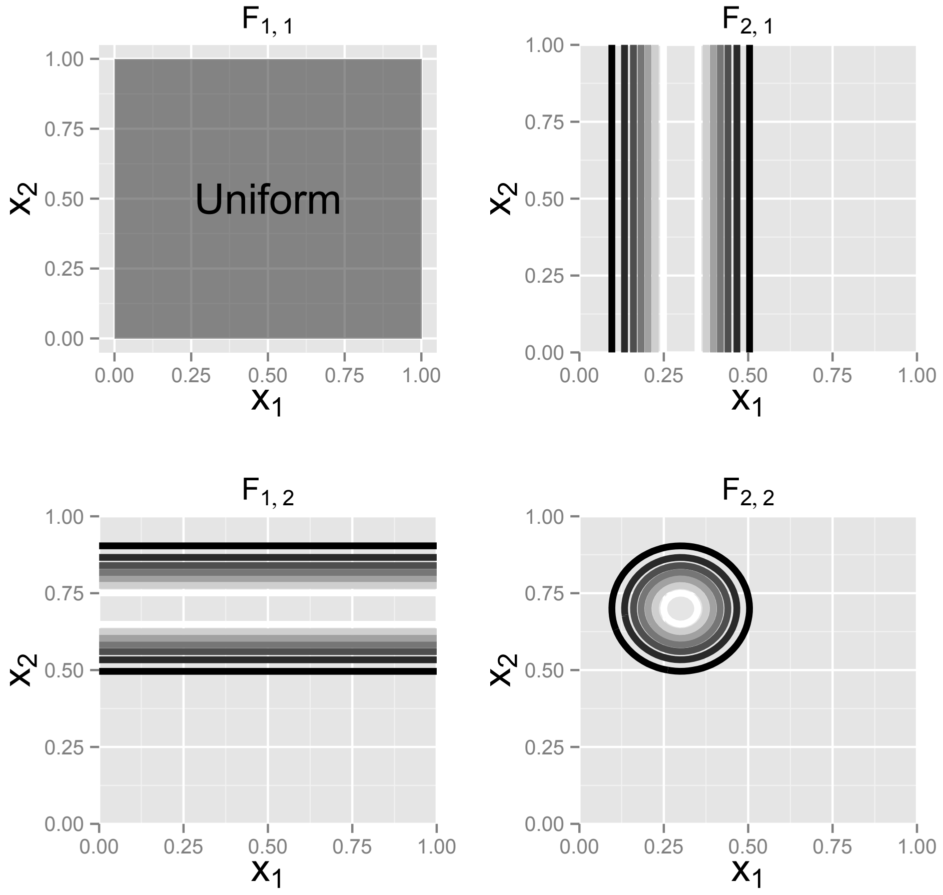

FIGURE 3.1
Each of the four boxes shows the contour plot of an FMM component in Equations (3.19)–(3.22). They correspond to the MMM defined by the F profiles in Equations (3.17)–(3.18). X_1 and X_2 are the two observed variables. Lighter contour lines indicate higher density.

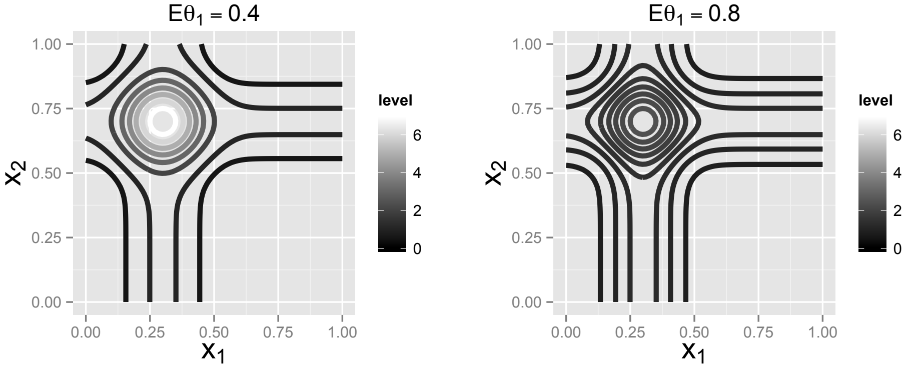

FIGURE 3.2
Contour plot of the MMM defined by the profiles in Equations (3.17)–(3.18) with two different distributions of θ. X_1 and X_2 are the two observed variables. Lighter contour lines indicate higher density; the scale is the same for both figures.

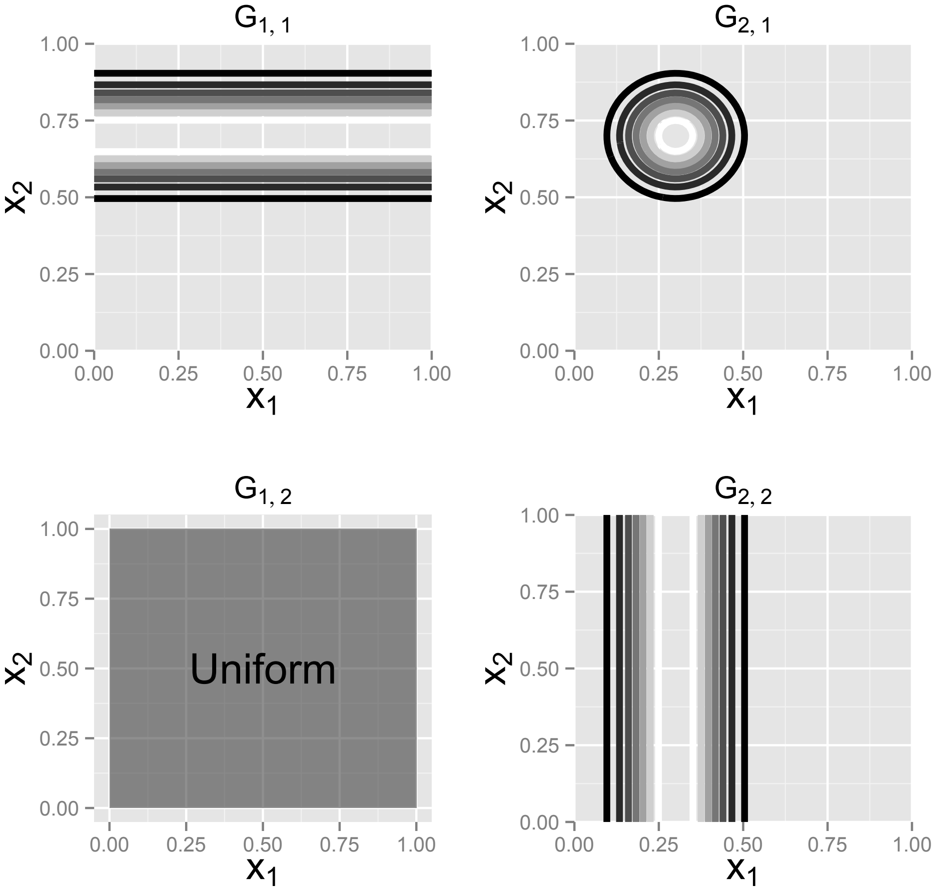

FIGURE 3.3
Each of the four boxes shows the contour plot of an FMM component corresponding to the MMM defined by the G profiles in Equations (3.23)–(3.24). Note that these are the same components as those defined by the F profiles in Figure 3.1 and Equations (3.19)–(3.22), simply re-indexed. X_1 and X_2 are the two observed variables. Lighter contour lines indicate higher density.

There are still K^R FMM components; the only difference is that these clusters are not in an R-dimensional space but a higher-dimensional space, depending on the dimensionality of the X_j.

3.6 Categorical vs. Continuous Data

All three of the original mixed membership models, and a majority of the subsequent variations, were built for categorical data. This focus on categorical data can lead to intuitions about mixed membership models which do not hold in the general case. Since every mixed membership model can be expressed as a finite mixture model, the best way to understand the difference between continuous and categorical data in MMM is to focus on how different data types behave in FMM.

Let us begin by considering the individual distributions conditional on profile membership (Equation 3.3):

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

In general, this equation does not simplify, but in the case of categorical data, it does. This is the key difference between categorical data and any other type of data.

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 $F_{kj}(x_j) = Multinomial(\lambda_{kj}, n = 1)$, 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:

$$Pr(X_j = l|\theta_k = 1) = \lambda_{kjl}. \tag{3.29}$$

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

$$Pr(X_{ij} = l|\theta_i) \ = \ \sum_{k=1}^{K} \theta_{ik} Pr(X_j = l|\theta_k = 1) \ = \ \sum_{k=1}^{K} \theta_{ik}\lambda_{kjl}. \tag{3.30}$$

Consider LDA as an example. 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 $Multinomial(\lambda_s)$ and $Multinomial(\lambda_m)$. 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 1, 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 (3.30).

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

$$\lambda_{ij} = \theta_i^T\lambda_j. \tag{3.31}$$

The matrix λ_j defines a linear transformation from θ_i to λ_{ij}, as illustrated in Figure 3.4. Since θ_i is a probability vector and sums to 1, λ_{ij} is a convex combination of the the profile probability vectors λ_{kj}. Thus the individual λ_{ij} lies within 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 3.4).

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

$$F(x_j|\theta_i) = Multinomial(\theta_i^T \lambda_j, n = 1). \tag{3.32}$$

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 $n = 1$. Moreover, 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.

In general, this between relationship does not hold. The general interpretation is a *switching* interpretation, and is clearly captured by the indicator variable Z_{ijr} in the generative model. Z_{ijr} indicates which profile distribution k generated the observation X_{ijr}. Thus, Z indicates that an individual switched from profile k for the j^{th} variable to profile k' for the $j + 1^{st}$ variable.

The between interpretation for categorical data only holds in the multinomial parameter space: λ_i is between the profile parameters λ_k. 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 LDA as an example. The observation X_{ir} is the rth word appearing in document i; each profile is a multinomial probability distribution over the set of words. "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

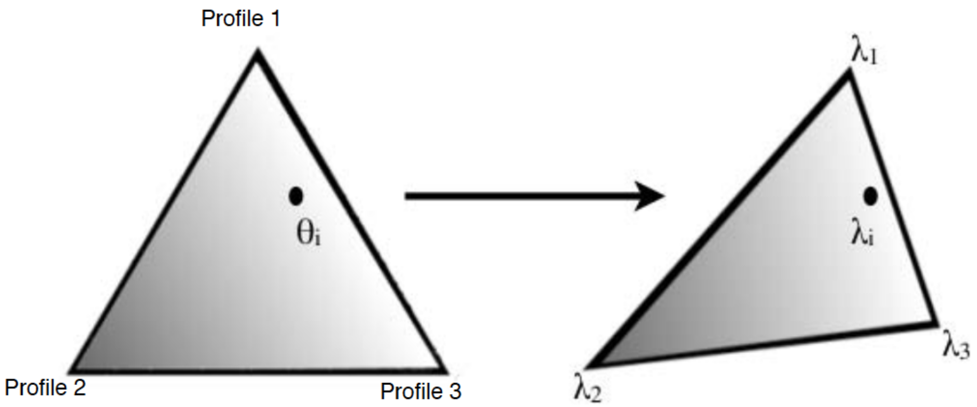

FIGURE 3.4
The membership parameter θ_i lies in a $K - 1$ simplex. When the mixed membership profiles are $F_{kj} = Multinomial(\lambda_{kj}, n = 1)$, the membership parameters are mapped linearly onto response probabilities (Equation 3.31), indicated by the arrow. The density, indicated by the shading, is preserved by the linear mapping. This mapping allows us to interpret individual i's position in the θ-simplex as equivalent to their response probability vector.

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.

Consider another example: suppose that we are looking at response times for a student taking an assessment, where X_{ij} is the response time of student i on item j and each profile represents a particular strategy. 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 can use a switching interpretation, but a between interpretation is unavailable.

3.6.1 Conditions for a ‘Between’ Interpretation

The between interpretation arises out of a special property of the multinomial distribution: the individual probability distributions are in the same parametric family as the profile distributions, multinomial with $n = 1$, and the individual parameters are between the profile parameters (Equation 3.31 and Figure 3.4).

For the between interpretation to be available, this is the property we need to preserve. The individual distributions $F(x|\theta_i)$ must be in the same parametric family as each profile distribution F_k. Additionally, if F is parameterized by ϕ, then the individual parameters ϕ_i must lie between the profile parameters ϕ_k.

Thus, 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 would have:

$$X_{ij}|\theta_i \quad \sim \quad \sum_k \theta_{ik} F_{kj}(X_j; \phi_{kj}) = F_j(X_j; \theta_i^T \phi_{\cdot j}). \tag{3.33}$$

In other words, the between interpretation is only available if the profile cumulative distribution functions (cdfs) are linear transformations of their parameters. The only exponential family distribution with this property is the multinomial distribution with $n = 1$. Thus, it is the only common profile distribution which allows a between interpretation (Galyardt, 2012).

The partial membership models in Gruhl and Erosheva (2013) and Mohamed et al. (2013) use a likelihood that is equivalent to Equation (3.33) in the general case. This fundamentally alters the mixed membership exchangeability assumption for the distribution of $X_{ij}|\theta_i$ and preserves the between interpretation in the general case.

Example

We will focus on a single variable j, omitting the subscript j within this example for simplicity. Let the profile distributions be Gaussian mixture models with proportions $\beta_k = (\beta_{k1}, \ldots, \beta_{kS})$ and fixed means c_s. If we denote the cdf of the standard normal distribution as Φ, then we can write the profiles as

$$F_k(x) = F(x; \beta_k) = \sum_s \beta_{ks} \Phi(x - c_s). \tag{3.34}$$

Define $\beta_{is} = \theta_i^T (\beta_{1s}, \ldots, \beta_{Ks})$. Then the individual distributions, conditional on the membership vector θ_i, are

$$X|\theta_i \quad \sim \quad \sum_k \theta_{ik} \left[\sum_s \beta_{ks} \Phi(x - c_s) \right], \tag{3.35}$$

$$
\begin{aligned}
&= \sum_s \beta_{is}\Phi(x - c_s), && (3.36)\\
&= F(x;\beta_i). && (3.37)
\end{aligned}
$$

Thus, the individual parameter β_i is in between the profile parameters β_k.

Now let us change the profile distributions slightly. Suppose the means are no longer fixed constants but are also variable parameters:

$$F_k^\star(x) = F^\star(x;\beta_k,\mu) = \sum_s \beta_{ks}\Phi(x - \mu_s). \tag{3.38}$$

In this case the individual conditional distributions are given by

$$
\begin{aligned}
X|\theta_i &\sim \sum_k \theta_{ik}\left[\sum_s \beta_{ks}\Phi(x-\mu_s)\right], && (3.39)\\
&= F^\star(x;\beta_i,\mu). && (3.40)
\end{aligned}
$$

Figure 3.5 shows three example profiles of this form and the distribution of $X|\theta_i$ for two individuals. Here, the between interpretation does not hold in the entire parameter space. Individual data distributions are the same form as the profile distributions—both are in the F^* parametric family. However, F^* has two parameters, β and μ. The individual mixing parameter β_i will lie in a simplex defined by the profile parameters β_k, since $\beta_{is} = \theta_i^T(\beta_{1s},\ldots,\beta_{Ks})$.

The fact that the individual mixing parameter β_i is literally 'between' the profile mixing parameters β_k allows us to interpret individuals as a 'blend' of the profiles. The same is not true for the μ parameter. We only have the between interpretation when considering the β parameters.

Now, let's make another small change to the profile distributions. Suppose that the standard deviation of the mixture components is not the same for each profile:

$$F_k^\dagger(x) = F^\dagger(x;\beta_k,\mu,\sigma_k) = \sum_s \beta_{ks}\Phi\left(\frac{x-\mu_s}{\sigma_k}\right). \tag{3.41}$$

Now the conditional individual distributions are

$$
\begin{aligned}
X|\theta_i &\sim \sum_k \theta_{ik}F_k^\dagger(x), && (3.42)\\
&= \sum_k \theta_{ik}\left[\sum_s \beta_{ks}\Phi\left(\frac{x-\mu_s}{\sigma_k}\right)\right], && (3.43)\\
&= \sum_k\sum_s \theta_{ik}\beta_{ks}\Phi\left(\frac{x-\mu_s}{\sigma_k}\right). && (3.44)
\end{aligned}
$$

Equation (3.44) does not simplify in any way. The conditional individual distribution is no longer of the $F^\dagger$ form and as such does not have parameters that are between the profile parameters. Figure 3.6 is an analog of Figure 3.5 and shows three $F^\dagger$ profiles and the distribution of $X|\theta_i$ for two individuals.

This example is analogous to the model of genetic variation, *mStruct* (Shringarpure, 2012). In this model, the population is comprised of K ancestral populations, and each member of the current population has mixed membership in these ancestral populations. *mStruct* also accounts for the fact that the current set of alleles may contain mutations from the ancestral set of alleles.

Each ancestral population has different proportions $\beta_k = (\beta_{k1},\ldots,\beta_{kS})$ of the set of founder

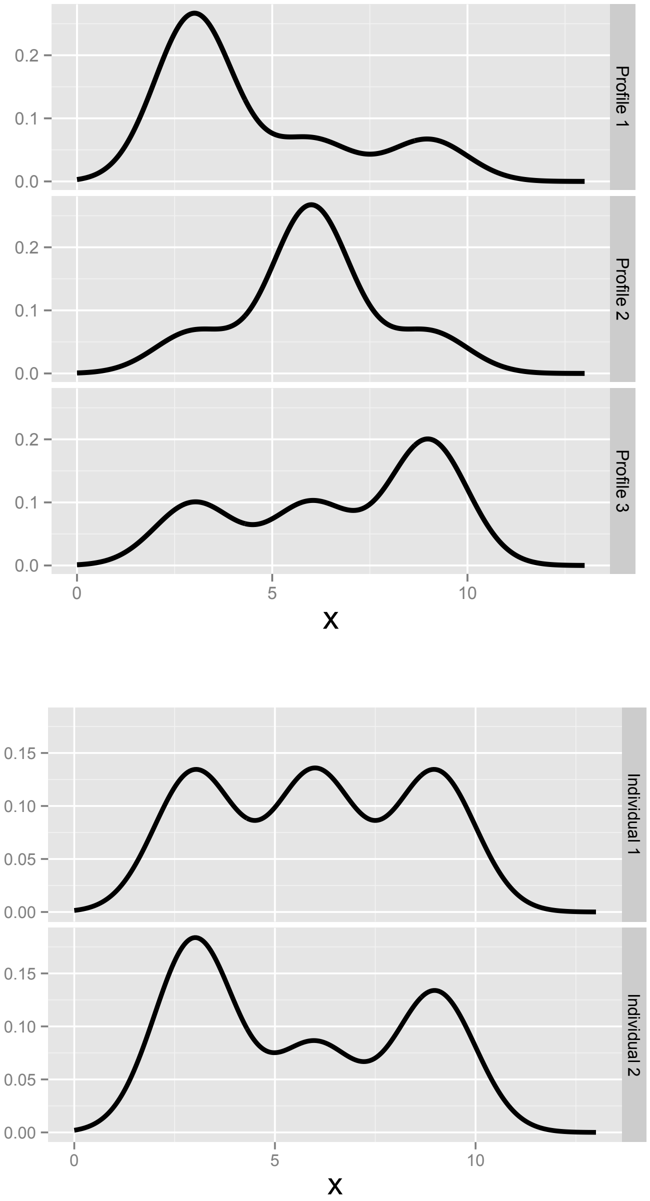

FIGURE 3.5
Non-multinomial profile distributions that preserve the ‘between’ interpretation. The top graph shows three profiles of the form $F_k^* = \sum_s \beta_{ks}\Phi(x - \mu_s)$ (Equation 3.38). The mixture means μ_s and the standard deviations are the same for each profile. The lower graph shows two individual distributions where $X|\theta_i \sim F^\star(x; \beta_i, \mu)$ (Equation 3.40).

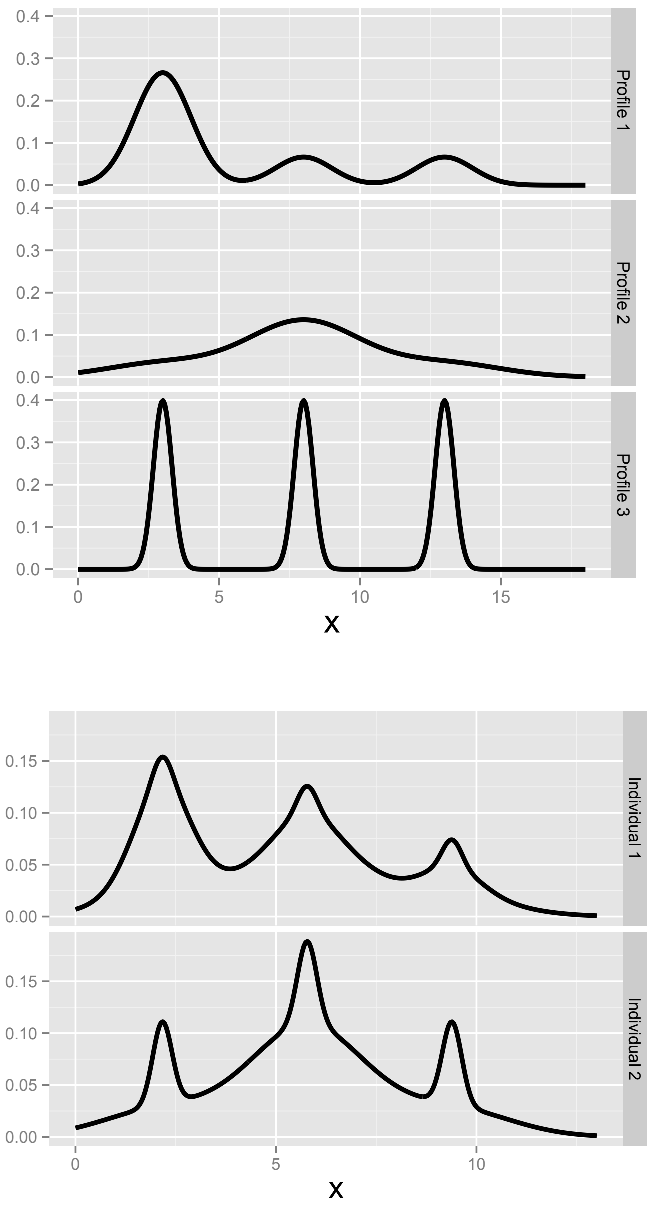

FIGURE 3.6
Profile distributions that do not preserve the 'between' interpretation. The top graph shows three profiles of the form $F_k^\dagger = \sum_s \beta_{ks}\Phi\left(\frac{x-\mu_s}{\sigma_k}\right)$ (Equation 3.41). The mixture means μ_s are the same for each profile, but the standard deviations σ_k are different. The lower graph shows two individual distributions with $X|\theta_i \sim \sum_k \sum_s \theta_{ik}\beta_{ks}\Phi\left(\frac{x-\mu_s}{\sigma_k}\right)$ (Equation 3.44).

alleles at locus j: $\mu_j = (\mu_{j1}, \ldots, \mu_{jS})$. The observed allele for individual i at locus j, X_{ij}, will have mutated from the founder alleles according to some probability distribution $P(\cdot|\mu_{js}, \delta_{kj})$, with the mutation rate δ_{kj} differing depending on the ancestral population. Thus, the profile distributions are

$$F_{kj}(x_j) \quad = \quad F(x_j; \beta_{kj}, \mu_j, \delta_{kj}) \quad = \quad \sum_{s=1}^{S} \beta_{kjs} P(x|\mu_{js}, \delta_{kj}). \tag{3.45}$$

The individual probability distribution of alleles at locus j, conditional on their membership in the ancestral profiles is then given by

$$X_{ij}|\theta_i \;\sim\; \sum_k \theta_{ik} \left[\sum_s \beta_{kjs} P(x|\mu_{js}, \delta_{kj}) \right]. \tag{3.46}$$

In the same way that the conditional individual distributions in the $F^{\dagger}$ model (Equation 3.44) do not simplify, the individual distributions in the *mStruct* do not simplify.

3.7 Contrasting Mixed Membership Regression Models

In this section, we compare and contrast two mixed membership models which are identical in the exchangeability assumptions and the structure of the models. The only difference is that in one case the data is categorical, and in the other case it is continuous. In the categorical case, the between interpretation holds and mixed membership is a viable way to model the structure of the data. In the continuous case, the between interpretation does not hold and mixed membership cannot describe the variation that is present in the data.

Let us suppose that in addition to the variables X_{ij} we also observe a set of covariates T_{ij}. For example, T may be the date a particular document was published or the age of a participant at the time of the observation. In this case, we may want the MMM profiles to depend on these covariates: $F_k(x|t)$. There are many ways to incorporate covariates into F, but perhaps the most obvious is a regression model.

Every regression model, whether linear, logistic, or nonparametric is based on the same fundamental assumption: $\mathbb{E}[X|T = t] = m(t)$. When X is binary, $X|T = t \;\sim$ Bernoulli$(m(t))$. When X is continuous, we most often use $X|T = t \;\sim N(m(t), \sigma^2)$. In general, we tend not to treat these two cases as fundamentally different, they are both just regression. The contrast between these two mixed membership models is inspired by an analysis of the National Long Term Care Survey (Manrique-Vallier, 2010) and an analysis of children's numerical magnitude estimation (Galyardt, 2010; 2012). In Manrique-Vallier (2010), X is binary and T is continuous, so that the MMM profiles are

$$F_k(x|t) = \text{Bernoulli}(m_k(t)). \tag{3.47}$$

In Galyardt (2010), both X and T are continuous, so that the MMM profiles are

$$F_k(x|t) = N(m_k(t), \sigma_k^2). \tag{3.48}$$

Note, however, that for the reasons explained here and detailed in Section 3.7.2, a mixed membership analysis of the numerical magnitude estimation data was wildly unsuccessful (Galyardt, 2010). An analysis utilizing functional data techniques was much more successful (Galyardt, 2012).

The interesting question is why an MMM was successful in one case and unsuccessful in the other. At the most fundamental level, the answer is that a mixture of Bernoullis is still Bernoulli, and a mixture of normals is not normal. This is a straightforward application of Erosheva's representation theorem.

To simplify the comparison, let us suppose that we observe a single variable ($J = 1$), with replications at points $T_r, r = 1, \ldots, R$. For example, X_{ir} may be individual i's response to a single survey item observed at different times T_{ir}. To further simplify, we will use only $K = 2$ MMM profiles with distributions $F(x; m_k(t))$. Thus for an individual with membership parameter θ_i, the conditional data distribution is:

$$X_i|T_i, \theta_i \quad \sim \quad \prod_r \left[\sum_k \theta_{ik} F(X_{ir}; m_k(T_{ir})) \right]. \tag{3.49}$$

3.7.1 Mixed Membership Logistic Regression

When the MMM profiles are logistic regression functions (Equation 3.47), then the conditional data distribution for an individual with membership parameter θ_i becomes

$$X_i|T_i, \theta_i \quad \sim \quad \prod_r \left[\sum_k \theta_{ik} Bernoulli(m_k(T_{ir})) \right], \tag{3.50}$$

with

$$m_k(t) = \text{logit}^{-1}(\beta_{0k} + \beta_{1k} t). \tag{3.51}$$

Equation (3.50) is easily rewritten as

$$X_i|T_i, \theta_i \quad \sim \quad \prod_r \left[Bernoulli \left(\sum_k \theta_{ik} m_k(T_{ir}) \right) \right]. \tag{3.52}$$

In this case, we can write an individual regression function,

$$m_i(t) = \sum_k \theta_{ik} m_k(t). \tag{3.53}$$

This individual regression function m_i does not have the same loglinear form as m_k, so we cannot talk about individual β parameters being between the profile parameters. However, it is a single smooth regression function that summarizes the individual's data, and m_i will literally be between the m_k. Figure 3.7 shows an example with two such logistic regression profile functions and a variety of individual regression functions specified by this mixed membership model.

3.7.2 Mixed Membership Regression with Normal Errors

When the MMM profiles are regression functions with normal errors (Equation 3.48), the conditional distribution for individual i's data is given by

$$X_i|T_i, \theta_i \quad \sim \quad \prod_r \left[\sum_k \theta_{ik} N\left(m_k(T_{ir}), \sigma_k^2\right) \right]. \tag{3.54}$$

Since a mixture of normal distributions is not normal, Equation (3.54) does not simplify. In this case it is impossible to write a smooth regression function m_i. Figure 3.8 demonstrates this by showing two profile regression functions and contour plots of the density for two individuals, $X_i|T_i, \theta_i$.

It can be tempting to suggest that a change in the distribution of the membership parameter

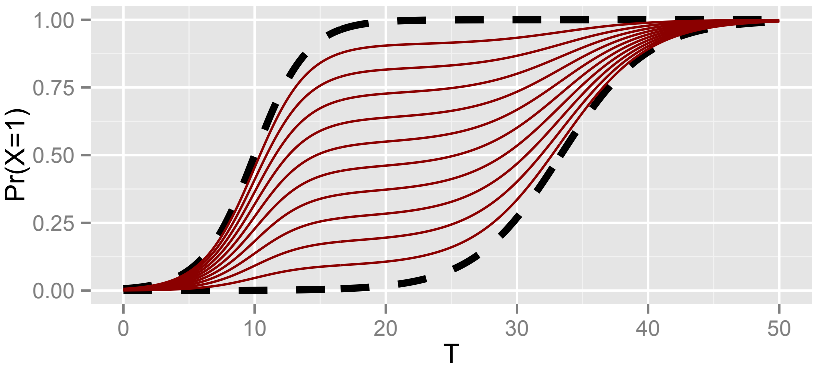

FIGURE 3.7
Profile and individual regression functions in a mixed membership logistic regression model. The thick dashed lines indicate the profile regression functions $m_k(t)$. The thin lines show individual regression functions $m_i(t)$ for a range of values of θ_i.

θ may resolve this issue. However, according to Erosheva's representation theorem, the profile distributions F_k control *where* the data is and θ only controls how much data is in each location (Equations 3.10, 3.11, and Section 3.5). Figure 3.9 illustrates the result of making θ_i a function of t, $\theta_i(t)$.

If the profile distributions F are linear transformations of their parameters (Equation 3.33), then a mixed membership regression model with profiles $F(m(x))$ will have individual regression functions $m_i(x)$. Otherwise a mixed membership model will not produce continuous individual regression functions.

Functional data are a class of data of the form $X_{ij} = f_i(t_{ij}) + \epsilon_{ij}$, where f_i is an individual smooth function, but we only observe a set of noisy measurements X_{ij} and t_{ij} for each individual (Ramsay and Silverman, 2005; Serban and Wasserman, 2005). For example, suppose we observe the height of children at different ages, or temperature at discrete intervals over a period of time. In this type of data analysis, the functions f_i and the similarities and variation between them are the primary objects of inference.

The examples in this section demonstrate that without fundamentally altering the exchangeability assumption of the general mixed membership model (Equation 3.4), a MMM cannot fit functional data. Equation (3.54) will never produce smooth individual regression functions. Galyardt (2012), Gruhl and Erosheva (2013), and Mohamed et al. (2013) suggest a way in which the exchangeability assumption might be altered to model individual regression functions as lying between the profile functions.

3.7.3 Children's Numerical Magnitude Estimation

The mixed membership regression model with normal errors is based on an analysis of the strategies and representations that children use to estimate numerical magnitude. This has been an active area of research in recent years (Ebersbach et al., 2008; Moeller et al., 2008; Siegler and Booth, 2004; Siegler and Opfer, 2003; Siegler et al., 2009). The primary task in experiments studying numerical magnitude estimation is a number line task. The experimenter presents each child with a series of number lines which have only the endpoints marked. The scale of the number lines is most often 0 to 100, or 0 to 1000. The child estimates a number by marking the position where they think the

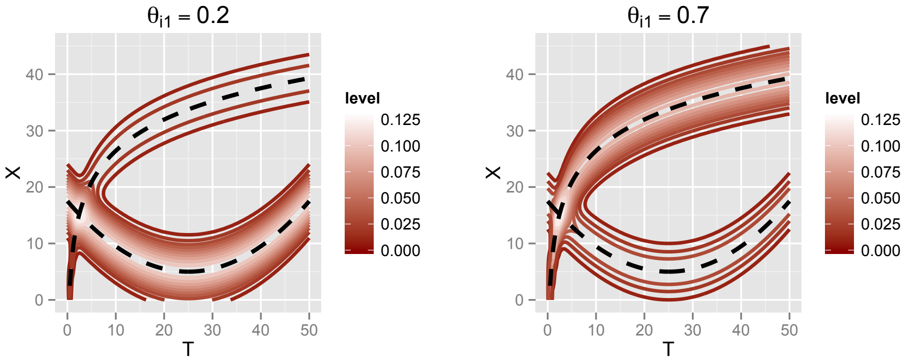

FIGURE 3.8
Mixed membership regression model with normal errors. The two plots show contours of the data distribution for two different values of θ_i. The thick dashed lines indicate the profile regression functions. Lighter contour lines indicate higher density. Note that there is no individual regression function m_i, which can summarize data from this distribution.

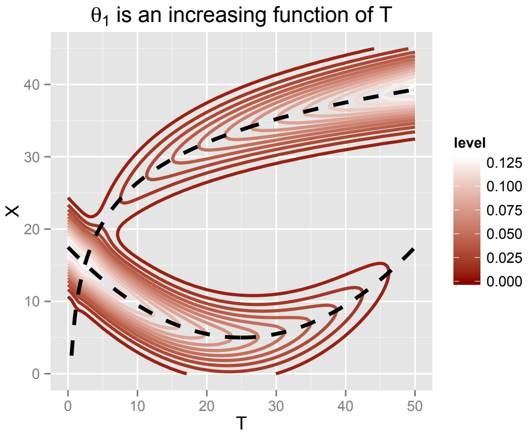

FIGURE 3.9
Mixed membership regression model with normal errors. Contour plot of an individual data distribution where $\theta_{i1}(t)$ is an increasing function of T. The thick dashed lines indicate the profile regression functions. Lighter contour lines indicate higher density. We cannot summarize data from this distribution with any smooth regression function $m_i(t)$.

number 'belongs.' Each child will estimate a series of numbers, with a single number line on each page.

There are competing theories as to how children represent numerical magnitude and the strategies that they use to estimate numbers (Ebersbach et al., 2008; Galyardt, 2012; Moeller et al., 2008; Opfer and Siegler, 2007; Siegler et al., 2009). This argument is not our primary concern. We will focus on the aspect of performance that all of the studies agree upon: there is an immature pattern and a mature pattern. Older children are able to accurately and linearly estimate numerical magnitude. That is, if T_{ir} is the rth number you ask child i to estimate, then their estimates X_{ir} can be modeled as $X_{ir} = T_{ir} + \epsilon_{ir}$.

Young children consistently overestimate small numbers. For example, a kindergardener estimating on the 0–100 scale may place the number 23 three-quarters of the distance from 0 to 100, near a position of 75. These children also appear to not differentiate well between larger quantities, so that they might place both 56 and 84 near a position of 90. The estimate from a child displaying the immature pattern will follow $X_{ir} = m(X_{ir}) + \epsilon_{ir}$. The exact functional form of $m(x)$ is disputed; Opfer and Siegler (2007) and Siegler et al. (2009) suggest that it is logarithmic; Ebersbach et al. (2008) and Moeller et al. (2008) suggest that it is piece-wise linear.

At this point, it seems natural to model children who are learning the mature representation as having mixed membership in both representations (Galyardt, 2010). We can represent each strategy with a MMM profile and use the membership parameter to indicate the degree to which a child has learned the mature strategy. Thus the profiles are mixed membership regression functions with normal errors, as in Equation (3.54). The distribution of individual data predicted by this model would be similar to the distributions shown in Figure 3.8. This mixed membership model would embody a 'switching' interpretation; sometimes the child uses the mature strategy and sometimes the child uses the immature strategy.

This is where the difference between the switching and blending interpretations becomes critical. Children using the immature strategy will estimate the number 30 near the position 80, while those using the mature strategy will estimate the position accurately at 30. If a child is blending the two strategies, then a model should predict an estimate at a position between 30 and 80. On the other hand, if a child is switching between the mature and the immature strategy, then a model should predict estimates near these two points and have lower probability in the middle.

Figure 3.10 shows data from a number line estimation task for six representative individuals. We can see immediately that this is functional data. Each child's strategy can be represented by a single smooth curve, f_i.

Some children clearly display the immature pattern, some children display the mature pattern. The interesting patterns belong to the children between the two extremes. Yet the mixed membership regression model cannot capture this variation, even with the addition of more profiles. The profiles are normal, and since mixtures of normals are not normal, the individual distributions will not be normal. Therefore the exchangeability assumptions in Equation (3.54) will not produce a smooth regression function for each individual.

In this kind of application, we want to model where each individual lies between the two extremes. A mixed membership model cannot capture the patterns of variation that are present in this data. As one measure of model misfit, an attempt to use the mixed membership model with normal errors (Equation 3.54) on this data resulted in estimates of $\sigma > 30$, with data on a scale of 0–100 (Galyardt, 2010). One way to solve this problem is to apply functional data analysis tools, the approach successfully used in Galyardt (2012). Another approach is to alter the exchangeability assumption to allow for a 'between' interpretation (Gruhl and Erosheva, 2013; Mohamed et al., 2013).

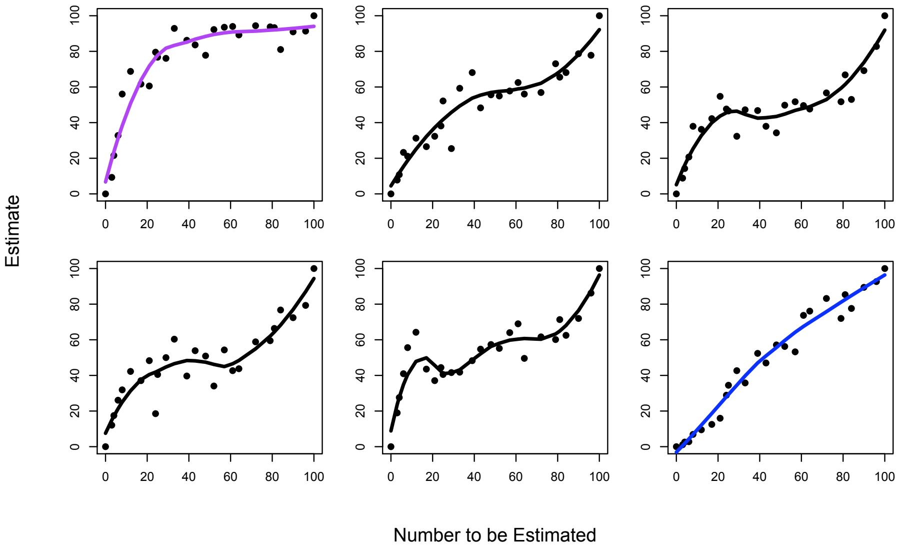

FIGURE 3.10
Each box displays data from a single child participant in Siegler and Booth (2004). Individuals were selected to display the range of strategies observed in the data. The immature and mature patterns are present, but other intermediate patterns are present as well.

3.8 Discussion

Everything presented in this chapter is a straightforward observation based on Erosheva's representation theorem (Erosheva et al., 2007). Every mixed membership model can be expressed as a finite mixture model with a much larger number of classes. Therefore, the best way to understand how mixed membership models behave and how we should interpret them is by focusing on the relationship with finite mixture models.

Categorical data and the multinomial distribution have a unique behavior within the family of finite mixture models. Therefore categorical data have a unique behavior within the family of mixed membership models.

In general, individuals with mixed membership in multiple profiles should be interpreted as switching between the profiles. For example, a student who uses one strategy on one problem and switches to another strategy for the next problem; or one segment of an image from the water profile that then switches its next segment to the tree profile. This switching interpretation is inherent in the exchangeability assumption that observed variables are independent conditional on the individual's membership parameter.

Only in a small set of special cases, including the multinomial distribution, can we interpret mixed membership as individuals being between the profiles. In these cases, the general switching interpretation is also accurate. Think of an individual who has mixed heritage. In the between interpretation, we can consider this individual as blending the two heritages together. Whereas in the switching interpretation, one gene may come from one heritage while the next gene comes from another heritage. In this special case, both interpretations work.

Changing the distribution of the membership parameters has no effect on which interpretations are available. Whether or not the profile distributions are linear transformations of their parameters is the only thing that determines whether the between interpretation is available. The same property is at work in the more complicated regression examples as in the simple examples.

Mixed membership models individuals switching between profiles. Partial membership (Galyardt, 2012; Gruhl and Erosheva, 2013; Mohamed et al., 2013) models individuals blending profiles. Only in very special cases do the two interpretations overlap.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)* 44: 139–177.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*. New York, NY, USA: ACM, 127–134.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. New York, NY, USA: ACM, 113–120.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *Annals of Applied Statistics* 1: 17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Ebersbach, M., Luwel, K., Frick, A., Onghena, P., and Verschaffel, L. (2008). 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.

Erosheva, E. A. (2002). Grade of Membership and Latent Structure Models with Application to Disability Survey Data. Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Erosheva, E. A., Fienberg, S. E., and Lafferty, J. D. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* 101: 5220–5227.

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

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 10th IEEE Computer Vision and Pattern Recognition (CVPR 2005)*. San Diego, CA, USA: IEEE Computer Society, 524–531.

Galyardt, A. (2010). Mixed membership models for continuous data. In *Proceedings of the 75th International Meeting of the Psychometric Society (IMPS 2010)*. Athens, GA, USA.

Galyardt, A. (2012). Mixed Membership Distributions with Applications to Modeling Multiple Strategy Usage. Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Girolami, M. and Kaban, A. (2005). Sequential activity profiling: Latent Dirichlet allocation of Markov chains. *Data Mining and Knowledge Discovery* 10: 175–196.

Gruhl, J. and Erosheva, E. A. (2013). A tale of two (types of) memberships: Comparing mixed and partial membership with a continuous data example. In Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds), *Handbook of Mixed Membership Models and Its Applications*. Chapman & Hall/CRC.

Haberman, S. J. (1995). Book review of statistical applications using fuzzy sets. *Journal of the American Statistical Association* 90: 1131–1133.

Manrique-Vallier, D. (2010). Longitudinal Mixed Membership Models with Applications to Disability Survey Data. Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Manton, K. G., Woodbury, M. A., and Tolley, H. D. (1994). *Statistical Applications Using Fuzzy Sets*. New York, NY: John Wiley & Sons.

Moeller, K., Pixner, S., Kaufmann, L., and Nuerk, H. -C. (2008). Children's early mental number line: Logarithmic or decomposed linear? *Journal of Experimental Child Psychology* 103: 503–515.

Mohamed, S., Heller, K. A., and Ghahramani, Z. (2013). A simple and general exponential family framework for partial membership and factor analysis. In Airoldi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds), *Handbook of Mixed Membership Models and Its Applications*. Chapman and Hall/ CRC Press.

Opfer, J. E. and Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology* 55: 169–195.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. New York, NY: Springer, 2nd edition.

Rogers, S., Girolami, M., Campbell, C., and Breitling, R. (2005). The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2: 143–156.

Serban, N. and Wasserman, L. (2005). CATS: Clustering after transformation and smoothing. *Journal of the American Statistical Association* 100: 990–999.

Shan, H. and Banerjee, A. (2011). Mixed-membership naive Bayes models. *Data Mining and Knowledge Discovery* 23: 1–62.

Shringarpure, S. (2012). Statistical Methods for Studying Genetic Variation in Populations. Ph.D. thesis, Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

Siegler, R. S. and Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development* 75: 428–444.

Siegler, R. S. and Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantitiy. *Psychological Science* 14: 237–243.

Siegler, R. S., Thompson, C. A., and Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain and Education* 3: 142–150.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester, UK: John Wiley & Sons.

Woodbury, M. A., Clive, J., and Garson, A., Jr. (1978). Mathematical typology: A Grade of Membership technique for obtaining disease definition. *Computers and Biomedical Research* 11: 277–298.