

1

Introduction to Mixed Membership Models and Methods

Edoardo M. Airoldi

Department of Statistics, Harvard University, Cambridge, MA 02138, USA

David M. Blei

Departments of Statistics and Computer Science, Columbia University, New York, NY 10027, USA

Elena A. Erosheva

Department of Statistics, University of Washington, Seattle, WA 98195, USA

Stephen E. Fienberg

Department of Statistics, Heinz College, and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

CONTENTS

1.1	Historical Developments	3
1.2	A General Formulation for Mixed Membership Models	4
1.3	Advantages of Mixed Membership Models in Applied Statistics	6
1.4	Theoretical Issues with Mixed Membership Models	8
1.4.1	General Issues Inherent to Mixtures	8
	References	10

Mixed membership models have emerged over the past 20 years as a flexible cluster-like modeling tool for unsupervised analyses of high-dimensional multivariate data where the assumption that an observational unit belongs to a single cluster, or principal component, is violated. Instead, one assumes that every unit partially belongs to all clusters, according to an individual membership vector. Mixed membership models were introduced essentially independently in a number of different statistical application settings: (1) survey data (Berkman et al., 1989; Erosheva, 2002; Erosheva et al., 2007), (2) population genetics (Pritchard et al., 2000b; Rosenberg et al., 2002), (3) text analysis (Blei et al., 2003; Erosheva et al., 2004; Airoldi et al., 2010), and then later on in (4) image processing and annotation (Barnard et al., 2003; Fei-Fei and Perona, 2005), and (5) molecular biology (Segal et al., 2005; Airoldi et al., 2006; 2007; 2013).

1.1 Historical Developments

This volume chronicles recent developments in the area of mixed membership modeling. Mixed membership models are used to characterize complex multivariate data such as those arising in studies of genetic build-up of biological organisms, patterns in disease and disability manifestations,

combinations of topics covered by text documents, political ideology or electorate voting patterns, or heterogeneous relationships in networks. Early applications of mixed membership modeling included the admixture model in genetics (Pritchard et al., 2000a), the Grade of Membership model in medical classification studies (Manton et al., 1994b), and the latent Dirichlet allocation model in machine learning (Blei et al., 2003).

In contrast to the finite mixture or parametric clustering models (McLachlan and Peel, 2000), mixed membership models assume that individuals or observational units may only partly belong to population mixture categories, referred to in various fields as *topics*, *extreme profiles*, *pure* or *ideal types*, *states*, or *subpopulations*. The degree of membership then is a vector of continuous non-negative latent variables that add up to 1 (in mixture models, membership is a binary indicator). The original idea for a mixed membership type of modeling goes back to at least the 1970s when the Grade of Membership (GoM) model was developed by mathematician Max Woodbury to allow for “fuzzy” classifications in medical diagnosis problems (Woodbury et al., 1978). The model had not received a lot of attention from statisticians in the early years, and was later characterized by seemingly controversial statements regarding the nature of the compositional data implied by the GoM model (Haberman, 1995). It was not until the early 2000s, with the widespread use of Bayesian methods and a better explanation of the duality between the discrete and continuous nature of latent structure in the GoM model, that a new Bayesian approach to the GoM model had been developed (Erosheva, 2003). The almost simultaneous and independent development of the admixture model in genetics (Pritchard et al., 2000a) and the latent Dirichlet allocation (LDA) model in computer science (Blei et al., 2003) also relied on the use of Bayesian estimation or approximate Bayesian estimation techniques, as in the case of LDA. This class of mixed membership models (Erosheva, 2002) unifies the LDA, GoM, and admixture models in a common framework and provides ways to construct other individual-level mixture models by varying assumptions on the population, sampling unit and latent variable levels, and the sampling scheme.

The word *mixed* in the name *mixed membership* comes from the alternative latent class specification of the models where each attribute is generated according to its distribution in a certain basis category (Erosheva et al., 2007). For example, each word in an article corresponds to a particular topic, whereas the article’s composition as a whole corresponds to the author’s intention to cover a selection of topics. Thus, the multivariate collection of outcomes for each sampling unit is composed of a mix of attributes that originate from the basis categories, e.g., words within a document that are generated from topics covered by that document. In the case of discrete data, the latent topic indicators for each word do not necessarily have to be the latent variables in the model. An alternative data-generating process that results in the same likelihood can be based on the latent degrees of membership controlling the proportions of attributes originating from each basis category (Erosheva, 2005). For this reason, mixed membership models have been occasionally referred to as partial membership models (e.g., Erosheva, 2004); however, that name has not gained widespread use and the name mixed membership remains the most commonly used descriptor (Erosheva and Fienberg, 2005).

1.2 A General Formulation for Mixed Membership Models

The general mixed membership model relies on four levels of assumptions: population, subject, latent variable, and sampling scheme. Population level assumptions describe the general structure of the population that is common to all subjects. Subject level assumptions specify the distribution of observed responses given individual membership scores. Membership scores are usually unknown and hence can also be viewed as latent variables. The next assumption specifies whether the membership scores are treated as unknown fixed quantities or as random quantities in the model. Finally,

the last level of assumptions specifies the number of distinct observed characteristics (attributes) and the number of replications for each characteristic. We describe each set of assumptions formally in turn.

Population Level

Assume there are K original or basis subpopulations in the populations of interest. For each subpopulation k , denote by $f(x_j|\theta_{kj})$ the probability distribution for response variable j , where θ_{kj} is a vector of parameters. Assume that within a subpopulation, responses to observed variables are independent.

Subject Level

For each subject, membership vector $\lambda = (\lambda_1, \dots, \lambda_K)$ provides the degrees of a subject's membership in each of the subpopulations. The probability distribution of observed responses x_j for each subject is fully defined by the conditional probability $Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj})$, and the assumption that response variables x_j are independent, conditional on membership scores. In addition, given the membership scores, observed responses from different subjects are independent.

Latent Variable Level

With respect to the latent variables, one could either assume that they are fixed unknown constants or that they are random realizations from some underlying distribution.

1. If the membership scores λ are fixed but unknown, the conditional probability of observing x_j , given the parameters θ and membership scores, is

$$Pr(x_j|\lambda; \theta) = \sum_{k=1}^K \lambda_k f(x_j|\theta_{kj}). \quad (1.1)$$

2. If membership scores λ are realizations of latent variables from some distribution D_α , parameterized by vector α , then the probability of observing x_j given the parameters is:

$$Pr(x_j|\alpha, \theta) = \int \left(\sum_{k=1}^K \lambda_k f(x_j|\theta_{kj}) \right) dD_\alpha(\lambda). \quad (1.2)$$

Sampling Scheme

Suppose R independent replications of J distinct characteristics are observed for one subject, $\{x_1^{(r)}, \dots, x_J^{(r)}\}_{r=1}^R$. Then, if the membership scores are treated as realizations from distribution D_α , the conditional probability is

$$Pr\left(\{x_1^{(r)}, \dots, x_J^{(r)}\}_{r=1}^R | \alpha, \theta\right) = \int \left(\prod_{j=1}^J \prod_{r=1}^R \sum_{k=1}^K \lambda_k f(x_j^{(r)}|\theta_{kj}) \right) dD_\alpha(\lambda). \quad (1.3)$$

When the latent variables are treated as unknown constants, the conditional probability for observing R replications of J variables can be derived analogously. In general, the number of observed characteristics J need not be the same across subjects, and the number of replications R need not be the same across observed characteristics.

One can obtain a number of mixed membership models using this general set up by specifying different choices of J and R , and different latent variable assumptions. For instance, the Grade

of Membership model of Manton et al. (1994b) assumes polytomous responses are observed to J survey questions without replications and uses the fixed-effects assumption for the membership scores. Potthoff et al. (2000) employ a variation of the Grade of Membership model by treating the membership scores as Dirichlet random variables; the authors refer to the resulting model as a *Dirichlet generalization of latent class models*. In genetics, Pritchard et al. (2000a) use a *clustering model with admixture*, which they labeled as structure. For diploid individuals the clustering model assumes that $R = 2$ replications (genotypes) are observed at J distinct locations (loci), treating the proportions of a subject's genome that originated from each of the basis subpopulations as random Dirichlet realizations. Variations of mixed membership models for text documents called *probabilistic latent semantic analysis* (Hofmann, 2001) and *latent Dirichlet allocation* (Blei et al., 2003) both assume that a single characteristic (word) is observed a number of times for each document, but the former model considers the membership scores as fixed unknown constants, whereas the latter treats them as random Dirichlet realizations.

The mixed membership model framework presented above unifies several specialized models that have been developed independently in the social sciences, genetics, and text mining applications.

1.3 Advantages of Mixed Membership Models in Applied Statistics

Mixed membership models have had a significant impact on applied statistics. Over the past decade, the data that statisticians analyze have become more diverse and structured, and with this complexity comes the opportunity to model individual data points as belonging to multiple groups. Indeed, for many modern datasets—such as large-scale text documents and complex networks—we believe that there is rarely a case for the simpler models. Statisticians need mixed membership models or alternatives to them, and this is the reason to study them.

The main areas to which mixed membership models have been applied are reflected in the contents of this volume.

Document Collections

Mixed membership models are widely applied to document collections (Blei et al., 2003; Blei, 2012). In document collections, the mixed membership assumptions naturally capture the heterogeneity of language, where documents each exhibit multiple themes and to different degree. When modeling documents as data, each document is a collection of words from a vocabulary. (These are grouped as categorical data.) Mixed membership models allow each document to exhibit multiple components, where each component is a distribution over words. Conditioned on a collection, inspecting the posterior of the components reveals the “topics” inherent in the documents, i.e., the significant patterns of words associated under a single theme. For this reason, mixed membership models of text are often called *topic models*.

Mixed membership models for text have been extended in a myriad of ways and developed for many text-based applications. As examples, they have been developed into time series (Blei and Lafferty, 2006), into further hierarchicalized models of word contagion (Doyle and Elkan, 2009), into Bayesian nonparametric variants (Teh et al., 2006), and into models of interconnected documents (Chang and Blei, 2009). In some ways, mixed membership models of text have become a benchmark for new innovations in mixed membership modeling.

Network Data

Another central application of mixed membership models is for the analysis of network data. A network consists of a population of units and their relationships, represented via a graph with a set of nodes and edges between them. Networks arise naturally in sociological settings, co-author analysis, and a variety of biological problems. A classical latent-variable model of networks is the *stochastic blockmodel* (Wang and Wong, 1987), which assumes that each node belongs to a community, and that its assigned community mediates its connection to other nodes. While these assumptions may have been appropriate for small scale network analysis, modern networks are heterogeneous. Nodes belong to *multiple* communities, and each node's connections reflect its particular signature of community memberships. This is a natural setting for mixed membership models.

Airoldi et al. (2008) developed the mixed membership extension of the stochastic blockmodel. Each node possesses an associated membership vector containing community proportions; each edge (present or absent) is associated with a community assignment drawn from the corresponding nodes' proportions. Note that modeling networks is fundamentally different from modeling documents because the observations are by definition intertwined. (We typically assume that documents, in contrast, are conditionally independent.) Mixed membership network models remain an active area of research. Further innovations include modeling dynamic networks (Ho et al., 2011) and including node attributes in modeling (Kim and Leskovec, 2011; Azari and Airoldi, 2012; Azizi et al., 2014). More broadly, networks are a type of *dyadic data*—data with entries indexed by a row and column—for which we can conceive more general mixed membership models (Mackey et al., 2010).

Social and Health Sciences Applications

The earliest mixed membership model, the Grade of Membership model (GoM) was developed by the statistician Max Woodbury (Woodbury et al., 1978), in the context of a medical classification problem where subsets of symptoms were observed on each patient. The goal was to identify and characterize sub-patterns of illness in a particular disease such as depression (Davidson et al., 1989), schizophrenia (Manton et al., 1994a), and Alzheimer's (Corder and Woodbury, 1993). GoM model analysis has been applied extensively to disability survey data—to analyze patterns in binary indicators of basic and instrumental activities of daily living—in a frequentist (Berkman et al., 1989; Manton et al., 1991) and Bayesian framework (Erosheva et al., 2007). Mixed membership methodologies have been extended to longitudinal settings to capture heterogeneous pathways of disability and cognitive trajectories at the later portion of life (this volume: Manrique-Vallier, 2014 and Lecci, 2014). In political science, researchers have used mixed membership models to analyze politically-oriented beliefs, values, and attitudes from survey data (this volume: Gross and Manrique-Vallier, 2014) and have developed mixed membership models for rank data to analyze votes in Irish elections (Gormley and Murphy, 2009). Other applications of mixed membership models include assessing the risk of privacy violations in databases (Manrique-Vallier and Reiter, 2012), and even reconstructing the contents of a city based on sparse archeological evidence (Mimno, 2011).

Population Genetics

In computational biology, mixed membership models have had a tremendous impact, most notably following the structure model of Pritchard et al. (2000a). In this setting, we observe a collection of human genomes in which each is a collection of alleles (A,G,C,T) measured at different locations. The model assumes that there are ancestral populations, groups of original humans that share a unique genetic signature, which migrated around the world and mixed. The observed genomes—the data we are analyzing—reflect the results of that mixing. Each genome exhibits the populations with different proportions, and each population is characterized by its allele probabilities across

genome locations. Posterior inference of the proportions and populations reveals the latent genetic structure of modern humans.

This kind of analysis has been used in two ways. First, as for networks and text, it is useful for exploring genetic patterns and forming hypotheses about our genetic history. Second, it is important for correcting analyses that seek to find associations between genes and traits. Patterns in ancestral populations, though not observed, are a confounder to making such associations; inferences from mixed membership models are useful in accounting for them. In this volume Shringarpure and Xing (2014) discuss some interesting variants on the original Pritchard et al. (2000a).

1.4 Theoretical Issues with Mixed Membership Models

The early examples of original mixed membership models described above were developed for discrete data, involving multivariate binary data, multinomial data, and ranks, and researchers using them considered responses to survey questions, counts of words in a document, sequences of genotypes, presence or absence of interactions between units, etc. Even though the general formulation of mixed membership models allows for combining outcomes of different types in a single omnibus model (Erosheva, 2002), the theoretical properties of mixed membership models applied to continuous data and data of mixed outcomes and applications of mixed membership for such problems is quite limited, e.g., see the discussion in Heller et al. (2008) and the the analysis of gene expression data by Rogers et al. (2005).

Extending mixed membership models to continuous data and data of mixed types is nontrivial. In this volume Galyardt (2014) demonstrates that the two interpretations—mixed attributes (the ‘switching’ interpretation) and partial memberships (the ‘between’ interpretation)—which are typically assumed as equivalent interpretations of mixed membership models, can not be taken for granted in the presence of continuous data. In fact, the ‘between’ interpretation no longer applies. Gruhl and Erosheva (2014) consider a broader class of individual-level mixture models and compare two members of this class—the mixed membership and the partial membership model (Heller et al., 2008)—for analyzing continuously-valued data. In essence, given individual-specific weights reflecting membership, mixed membership models assume that data are generated from individual-specific distributions that are weighted arithmetic averages of the subpopulation distributions, and partial membership models assume that individual-specific data are generated from a weighted geometric average of the subpopulation distributions. They explain that multivariate data may not provide researchers with a clear signal about the preferred type of individual-level mixture model. However, in this volume, analyzing a player statistics dataset from the National Basketball Association, Gruhl and Erosheva (2014) argue that the use of partial membership in that specific context is more appropriate. Partial membership models also happen to be more computationally convenient. Galyardt (2014) and Gruhl and Erosheva (2014) raise a number of issues for future work with individual-level mixture models for continuous data; some of these issues bear a clear connection to the large body of statistical literature on mixture models in general and on mixtures of normals in particular (McLachlan and Peel, 2000).

1.4.1 General Issues Inherent to Mixtures

While applications for mixed membership models especially in the form of extensions of topic models for text are widespread, these models suffer from a number of theoretical difficulties they inherit from mixture models. A lack of understanding of such issues may impact the validity of empirical analyses based on mixed membership models. Below we list a few key issues, borrowing material from a blog post on the topic by Wasserman (2012).

These issues are best illustrated in the context of a simple mixture model. Consider a finite mixture of Gaussians,

$$p(x; \psi) = \sum_{j=1}^k w_j \phi(x; \mu_j, \Sigma_j),$$

where $\phi(x; \mu_j, \Sigma_j)$ denotes a Gaussian density with mean vector μ_j and covariance matrix Σ_j . The weights w_1, \dots, w_k are non-negative and sum to 1. The entire set of parameters is $\psi = (\mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k, w_1, \dots, w_k)$. One can also consider k , the number of components, to be another parameter.

Now lets consider some of the weird things that can happen.

Infinite Likelihood. The likelihood function (for the Gaussian mixture) is infinite at some points in the parameter space. This is not necessarily bad, since the infinities are at the boundary and one can use the largest (finite) maximum in the interior as an estimator. But the infinities can cause numerical problems.

Multi-modality of the Likelihood. In fact, the likelihood has many modes (Richards and Buot, 2006). Finding the global (but not infinite) mode is a difficult. The EM algorithm only finds local modes. In this sense, the MLE is not really a well-defined estimator because it cannot be found. In the machine learning literature, there have been a number of papers trying to establish estimators for mixture models that can be found in polynomial time. For example, see Kalai et al. (2012).

Multi-modality of the Density. One may naïvely think that a mixture of k Gaussians would have k modes. But, in fact, it can have less than k or more than k . See Carreira-Perpinan and Williams (2003) and Edelsbrunner, Fasy, Rote (2012).

Non-identifiability. Recall that a model $\{p(x; \theta) : \theta \in \Theta\}$ is identifiable if

$$\theta_1 \neq \theta_2 \text{ implies } p(x; \theta_1) \neq p(x; \theta_2).$$

Mixture models are non-identifiable in two different ways. First, there is non-identifiability due to permutation of labels. This is a nuisance, and there are strategies to deal with it (Stephens, 2000). A bigger issue is local non-identifiability. Suppose that

$$p(x; \eta, \mu_1, \mu_2) = (1 - \eta)\phi(x; \mu_1, 1) + \eta\phi(x; \mu_2, 1).$$

When $\mu_1 = \mu_2 = \mu$, we have that $p(x; \eta, \mu_1, \mu_2) = \phi(x; \mu)$. The parameter η has disappeared. Similarly, when $\eta = 1$, the parameter μ_2 disappears. This means that there are subspaces of the parameter space where the family is not identifiable. The result is that all the usual theory about the distribution of the MLE, the distribution of the likelihood ratio statistic, the properties of BIC, and so on, becomes very complicated.

Irregularity. Mixture models do not satisfy the usual regularity conditions that make parametric models easy to deal with. Consider the following example from Chen (1995). Let

$$p(x; \theta) = \frac{2}{3}\phi(x; -\theta, 1) + \frac{1}{3}\phi(x; 2\theta, 1).$$

Then $I(0) = 0$ where $I(\theta)$ is the Fisher information. Moreover, no estimator of θ can converge faster than $n^{-1/4}$. Compare this to a Normal family $\phi(x; \theta, 1)$ where the Fisher information is $I(\theta) = n$ and the maximum likelihood estimator converges at rate $n^{-1/2}$.

Non-intuitive Group Membership. Mixtures are often used for finding clusters. Suppose that

$$p(x) = (1 - \eta)\phi(x; \mu_1, \sigma_1^2) + \eta\phi(x; \mu_2, \sigma_2^2)$$

with $\mu_1 < \mu_2$. Let $Z = 1, 2$ denote the two components. We can compute $P(Z = 1|X = x)$ and $P(Z = 2|X = x)$ explicitly. We can then assign an x to the first component if $P(Z = 1|X = x) > P(Z = 2|X = x)$. It is easy to check that, with certain choices of σ_1, σ_2 , all large values of x get assigned to component 1 (i.e., the leftmost component). Technically this is correct, yet it seems to be an unintended consequence of the model.

Improper Posteriors. Suppose we have a sample from the simple mixture

$$p(x; \mu) = \frac{1}{2}\phi(x; 0, 1) + \frac{1}{2}\phi(x; \mu, 1).$$

Then any improper prior on μ yields an improper posterior for μ regardless of how large the sample size is. Also, Wasserman (2012) shows that the only priors that yield posteriors in close agreement to frequentist methods are data-dependent priors.

These issues are often exacerbated in more complex mixed membership models. They should be taken seriously. In most applications, however, available additional information can be used to mitigate, and sometimes resolve, the problems listed above. The papers we have collected in this volume provide good examples, and explain why we do not share Wasserman's negative assessment: "that mixtures, like tequila, are inherently evil and should be avoided at all costs."

References

- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9: 1981–2014.
- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2006). Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the International Biometrics Society Annual Meeting*.
- Airoldi, E. M., Erosheva, E. A., Fienberg, S. E., Joutard, C., Love, T., and Shringarpure, S. (2010). Reconceptualizing the classification of PNAS articles. *Proceedings of the National Academy of Sciences* 107: 20899–20904.
- Airoldi, E. M., Fienberg, S. E., and Xing, E. P. (2007). Mixed membership analysis of genome-wide expression studies. Unpublished manuscript.
- Airoldi, E. M., Wang, X., and Lin, X. (2013). Multi-way blockmodels for analyzing coordinated high-dimensional responses. *Annals of Applied Statistics* 7: 2431–2457.
- Azari, H. and Airoldi, E. M. (2012). Graphlet decomposition of a weighted network. *Journal of Machine Learning Research* : 54–63.
- Azizi, E., Airoldi, E. M., and Galagan, J. E. (2014). Learning modular structures from network data and node variables. *Journal of Machine Learning Research, W&CP*. In press.
- Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D. M., and Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research* 3: 1107–1135.
- Berkman, L., Singer, B. H., and Manton, K. G. (1989). Black/white differences in health status and mortality among the elderly. *Demography* 26: 661–678.
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM* 55: 77–84.

- 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: 113–120.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.
- Carreira-Perpinan, M. A. and Williams, C. K. I. (2003). On the number of modes of a Gaussian mixture. In *Scale Space Methods in Computer Vision, Proceedings of the 4th International Conference on Scale Space*. Springer, 625–640.
- Chang, J. and Blei, D. M. (2009). Relational topic models for document networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*. *Journal of Machine Learning Research – Proceedings Track 5*, 81–88.
- Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Annals of Statistics* 23: 221–233.
- Corder, E. H. and Woodbury, M. A. (1993). Genetic heterogeneity in Alzheimer’s disease: A Grade of Membership analysis. *Genetic Epidemiology* 10: 495–499.
- Davidson, J. R., Woodbury, M. A., Zisook, S., and Giller, E. L., Jr. (1989). Classification of depression by Grade of Membership: A confirmation study. *Psychological Medicine* 19: 987–998.
- Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML '09)*. New York, NY, USA: ACM, 281–288.
- Edelsbrunner, H., Fasy, B. T., and Rote, G. (2012). Add isotropic Gaussian kernels at own risk: More and more resilient modes in higher dimensions. In *Proceedings of the ACM Symposium on Computational Geometry (SoCG 2012)*. New York, NY, USA: ACM, 91–100.
- Erosheva, E. A. (2002). Grade of Membership and Latent Structure Models with Application to Disability Survey Data. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
- Erosheva, E. A. (2003). Bayesian estimation of the Grade of Membership model. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds), *Bayesian Statistics 7*. New York, NY: Oxford University Press, 501–510.
- Erosheva, E. A. (2004). Partial membership models with application to disability survey data. In *Statistical Data Mining and Knowledge Discovery*. Chapman & Hall/CRC, 117–134.
- Erosheva, E. A. (2005). Comparing latent structures of the Grade of Membership, Rasch, and latent class models. *Psychometrika* 70: 619–628.
- Erosheva, E. A. and Fienberg, S. E. (2005). Bayesian mixed membership models for soft clustering and classification. In Weihs, C. and Gaul, W. (eds), *Classification – The Ubiquitous Challenge*. Springer, 11–26.
- 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.
- Erosheva, E. A., Fienberg, S. E., and Lafferty, J. D. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences* 97: 11885–11892.
- Galyardt, A. (2014). Interpreting mixed membership: Implications of Erosheva’s representation theorem. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds), *Handbook of Mixed Membership Models and Its Applications*. Chapman & Hall/CRC.

- Gormley, I. C. and Murphy, T. B. (2009). A Grade of Membership model for rank data. *Bayesian Analysis* 4: 265–296.
- Gruhl, J. and Erosheva, E. A. (2014). A tale of two (types of) memberships. 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’, by Kenneth G. Manton, Max A. Woodbury, and Larry S. Corder. *Journal of the American Statistical Association* 90(431): 1131–1133.
- Heller, K. A., Williamson, S., and Ghahramani, Z. (2008). Statistical models for partial membership. In *Proceedings of the 25th International Conference on Machine Learning (ICML '08)*. New York, NY, USA: ACM, 392–399.
- Ho, Q., Song, L., and Xing, E. P. (2011). Evolving cluster mixed-membership blockmodel for time-evolving networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. Palo Alto, CA, USA: AAI.
- Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42: 177–196.
- Kalai, A., Moitra, A., and Valiant, G. (2012). Disentangling Gaussians. *Communications of the ACM* 55: 113–120.
- Kim, M. and Leskovec, J. (2011). Modeling social networks with node attributes using the multiplicative attribute graph model. *arXiv preprint*, arXiv:1106.5053.
- 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.
- Gross, J. H. and Manrique-Vallier, D. (2014). A mixed membership approach to political ideology. 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.
- Lecci, F. (2014). An analysis of development of dementia through the Extended Trajectory Grade of Membership model. 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.
- Mackey, L., Weiss, D. and Jordan, M. I. (2010). Mixed membership matrix factorization. In Fürnkranz, J. and Joachims, T. (eds), *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*. Omnipress, 711–718.
- Manrique-Vallier, D. (2014). Mixed membership trajectory models. 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.
- Manrique-Vallier, D. and Reiter, J. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association* 107: 1385–1394.
- Manton, K. G., Stallard, E., and Woodbury, M. A. (1991). A multivariate event history model based upon fuzzy states: Estimation from longitudinal surveys with informative nonresponse. *Journal of Official Statistics* 7: 261–293.

- Manton, K. G., Woodbury, M. A., Anker, M., and Jablensky, A. (1994a). Symptom profiles of psychiatric disorders based on graded disease classes: An illustration using data from the WHO International Pilot Study of Schizophrenia. *Psychological Medicine* 24: 133–144.
- Manton, K. G., Woodbury, M. A., and Tolley, H. D. (1994b). *Statistical Applications Using Fuzzy Sets*. Wiley-Interscience.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1st edition.
- Mimno, D. (2011). Reconstructing Pompeian households. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. Corvallis, OR, USA: AUAI Press, 506–513.
- Potthoff, R. G., Manton, K. G., Woodbury, M. A., and Tolley, H. D. (2000). Dirichlet generalizations of latent-class models. *Journal of Classification* 17: 315–353.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
- Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. *American Journal of Human Genetics* 67: 170–181.
- Richards, D. and Buot, M. -L. G. (2006). Counting and locating the solutions of polynomial systems of maximum likelihood equations. *I. Journal of Symbolic Computation* 41: 234–244.
- 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.
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science* 298: 2381–2385.
- Segal, E., Pe’er, D., Regev, A., Koller, D., and Friedman, N. (2005). Learning module networks. *Journal of Machine Learning Research* 6: 503–556.
- Shringarpure, S. and Xing, E. P. (2014). Population stratification with mixed membership models. In Airolidi, E. M., Blei, D. M., Erosheva, E. A., and Fienberg, S. E. (eds), *Handbook of Mixed Membership Models and Its Applications*. Chapman & Hall/CRC.
- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Annals of Statistics* 2: 40–74.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101: 1566–1581.
- Wang, Y. and Wong, G. (1987). Stochastic block models for directed graphs. *Journal of the American Statistical Association* 82: 8–19.
- Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B* 62(1): 159–180.
- Wasserman, L. (2012). Mixture models: The twilight zone of statistics. Blog post, <http://normaldeviate.wordpress.com/2012/08/04/mixture-models-the-twilight-zone-of-statistics/>. Accessed on December 12, 2013.
- 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.

