

Bayesian Mixed Membership Models for Soft Clustering and Classification

Elena A. Erosheva¹ and Stephen E. Fienberg²

¹ Department of Statistics,
School of Social Work,
Center for Statistics and the Social Sciences,
University of Washington, Seattle, WA 98195, U.S.A.

² Department of Statistics,
Center for Automated Learning and Discovery,
Center for Computer and Communications Security
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Abstract. The paper describes and applies a fully Bayesian approach to soft clustering and classification using mixed membership models. Our model structure has assumptions on four levels: 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 observable responses given individual membership scores. Membership scores are usually unknown and hence we can also view them as latent variables, treating them as either fixed or random in the model. Finally, the last level of assumptions specifies the number of distinct observed characteristics and the number of replications for each characteristic. We illustrate the flexibility and utility of the general model through two applications using data from: (i) the National Long Term Care Survey where we explore types of disability; (ii) semantic decompositions of abstracts and bibliographies from articles published in *The Proceedings of the National Academy of Sciences*. In the first application we use a Monte Carlo Markov chain implementation for sampling from the posterior distribution. In the second application, because of the size and complexity of the data base, we use a variational approximation to the posterior. We also include a guide to other applications of mixed membership modeling.

1 Introduction

The canonical clustering problem has traditionally had the following form: for N units or objects measured on J variables, organize the units into G groups, where the nature, size, and often the number of the groups is unspecified in advance. The classification problem has a similar form except that the nature and the number of groups are either known theoretically or inferred from units in a training data set with known group assignments. In machine learning, methods for clustering and classification are referred to as involving “unsupervised” and “supervised learning” respectively. Most of these methods assume that every unit belongs to exactly one group. In this

paper, we will primarily focus on clustering, although methods described can be used for both clustering and classification problems.

Some of the most commonly used clustering methods are based on hierarchical or agglomerative algorithms and do not employ distributional assumptions. Model-based clustering lets $\mathbf{x} = (x_1, x_2, \dots, x_J)$ be a sample of J characteristics from some underlying joint distribution, $Pr(\mathbf{x}|\theta)$. Assuming each sample is coming from one of G groups, we estimate $Pr(\mathbf{x}|\theta)$ indicating presence of groups or lack thereof. We represent the distribution of the g th group by $Pr_g(\mathbf{x}|\theta)$ and then model the observed data using the mixture distribution:

$$Pr(\mathbf{x}|\theta) = \sum_{g=1}^G \pi_g Pr_g(\mathbf{x}|\theta), \quad (1)$$

with parameters $\{\theta, \pi_g\}$, and G .

The assumption that each object belongs exclusively to one of the G groups or latent classes may not hold, e.g., when characteristics sampled are individual genotypes, individual responses in an attitude survey, or words in a scientific article. In such cases, we say that objects or individuals have mixed membership and the problem involves *soft clustering* when the nature of groups is unknown or *soft classification* when the nature of groups is known through distributions $Pr_g(\mathbf{x}|\theta)$, $g = 1, \dots, G$, specified in advance.

Mixed membership models have been proposed for applications in several diverse areas. We describe six of these here:

1. *NLTCS Disability Data*. The National Long Term Care Survey assesses disability in U.S. elderly population. We have been working with a 2^{16} contingency table on functional disability drawing on combined data from the 1982, 1984, 1989, and 1994 waves of the survey. The dimensions of the table correspond to 6 Activities of Daily Living (ADLs)—e.g., getting in/out of bed and using a toilet—and 10 Instrumental Activities of Daily Living (IADLs)—e.g., managing money and taking medicine. In Section 3, we describe some of our results.
2. *DSM-III-R Psychiatric Classifications*. One of the earliest proposals for mixed membership models was by Woodbury et. al. (1978), in the context of disease classification. Their model became known as the *Grade of Membership* or GoM model, and was later used by Nurnberg et al. (1999) to study the DSM-III-R typology for psychiatric patients. Their analysis involved $N = 110$ outpatients and used the $J = 112$ DSM-III-R diagnostic criteria for clustering in order to reassess the appropriateness of the “official” 12 personality disorders. One could also approach this problem as a classical classification problem but with $J > N$.
3. *Peanut Butter Market Segmentation*. Seetharaman et al. (2001) describe data on peanut butter purchases drawn from A.C. Nielsen’s scanner database. They work with data from 488 households over 4715 purchase

occasions (chosen such that there are at least 5 per household) for 8 top brands of peanut butter. For each choice occasion we have: (a) shelf price, (b) information on display/feature promotion, and a set of household characteristics used to define “market segments” or groupings of households. Market segmentation has traditionally been thought of as a standard clustering problem but Varki et al. (2000) proposed a mixed-membership model for this purpose which is a variant on the GOM model.

4. *Matching Words and Pictures*. Blei and Jordan (2003) and Barnard et al. (2003) have been doing mixed-membership modeling in machine learning combining different sources of information in text documents, i.e., main text, photographic images, and image annotations. They estimate the joint distribution of these characteristics via employing hierarchical versions of a model known as the Latent Dirichlet Allocation in machine learning. This allows them to perform such tasks as automatic image annotations (recognizing image regions that portray, for example, clouds, water, and flowers) and text-based image retrieval (finding unannotated images that correspond to a given text query) with remarkably good performance.
5. *Race and Population Genetic Structure*. In a study of human population structure Rosenberg et al. (2002) used genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 populations and part of their analysis focuses on the soft clustering of individuals in groups. One of the remarkable results of their study which uses the mixed membership methods of Pritchard et al. (2002), is a typology structure that is very close to the “traditional” 5 main racial groups, a notion much maligned in the recent social science and biological literatures.
6. *Classifying Scientific Publications*. Erosheva, Fienberg, and Lafferty (2004) and Griffiths and Styvers (2004) have used mixed membership models to analyse related data bases involving abstracts, text, and references of articles drawn from the Proceedings of the National Academy of Sciences U S A (PNAS). Their mutual goal was to understand the organization of scientific publications in PNAS and we explore the similarities and differences between their approaches and results later in Section 4.

What these examples have in common is the mixed membership structure. In the following sections, we first introduce our general framework for mixed membership models and then we illustrate its application in two of the examples, using the PNAS and NLTCs data sets.

2 Mixed membership models

The general mixed membership model relies on four levels of assumptions: population, subject, latent variable, and sampling scheme. At the population level, we describe the general structure of the population that is common

to all subjects, while at the subject level we specify the distribution of observable responses given individual membership scores. At the latent variable level, we declare whether the membership scores are considered fixed or random with some distribution. Finally, at the last level, we specify the number of distinct observed characteristics and the number of replications for each characteristic. Following the exposition in Erosheva (2002) and Erosheva et al. (2004), we describe the assumptions at the four levels in turn.

Population level. We assume that there are K basis subpopulations (extreme or pure types) in the population of interest. For each subpopulation k , we denote by $f(x_j|\theta_{kj})$ the probability distribution for response variable j , where θ_{kj} is a vector of parameters. Moreover, we assume that, within a subpopulation, responses for the observed variables are independent.

Subject level. For each subject, membership vector $\lambda = (\lambda_1, \dots, \lambda_K)$ represents the degrees of a subject's membership in each of the subpopulations or the consonance of the subject with each of the pure types. The form of the conditional probability, $Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj})$, combined with the assumption that the response variables x_j are independent conditional on membership scores, fully defines the distribution of observed responses x_j for each subject. In addition, given the membership scores, we take the observed responses from different subjects to be independent.

Latent variable level. We can either assume that the latent variables are fixed unknown constants or that they are random realizations from some underlying distribution.

1. If the membership scores λ are fixed but unknown, then

$$Pr(x_j|\lambda; \boldsymbol{\theta}) = \sum_{k=1}^K \lambda_k f(x_j|\theta_{kj}) \quad (2)$$

is the conditional probability of observing x_j , given the membership scores λ and parameters $\boldsymbol{\theta}$.

2. If the membership scores λ are realizations of latent variables from some distribution D_α , parameterized by α , then

$$Pr(x_j|\alpha, \boldsymbol{\theta}) = \int \left(\sum_{k=1}^K \lambda_k f(x_j|\theta_{kj}) \right) dD_\alpha(\lambda) \quad (3)$$

is the marginal probability of observing x_j , given the parameters.

Sampling scheme. Suppose we observe R independent replications of J distinct characteristics for one subject, $\{x_1^{(r)}, \dots, x_J^{(r)}\}_{r=1}^R$. If the membership

scores are realizations from the distribution D_α , the conditional probability is

$$Pr\left(\{x_1^{(r)}, \dots, x_J^{(r)}\}_{r=1}^R | \alpha, \theta\right) = \int \left(\prod_{j=1}^J \prod_{r=1}^R \sum_{k=1}^K \lambda_k f(x_j^{(r)} | \theta_{kj}) \right) dD_\alpha(\lambda). \quad (4)$$

If we treat the latent variables as unknown constants, we get an analogous representation for the conditional probability of observing R replications of J variables. 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.

This mixed membership model framework unifies several specialized models that have been developed independently in the social sciences, in genetics, and in machine learning. Each corresponds to different choices of J and R , and different latent variable assumptions. For example, the standard GoM model of Woodbury and Clive (1974) and Manton et al. (1994) assumes that we observe responses to J survey questions without replications, i.e., $R = 1$, and treats the membership scores as fixed unknown constants (fixed-effects). Examples of the “fixed-effect” GoM analyses include but are not limited to: an analysis mentioned earlier of DSM-III psychiatric classifications in Nurnberg et al. (1999), a study of data on remote sensing (Talbot (1996)), an analysis of business opportunities (Talbot et al. (2002)), and a classification of individual tree crowns into species groups from aerial photographs (Brandtberg (2002)).

Another class of mixed membership models is based directly on the standard GoM model but places a distribution on the membership scores. Thus, Potthoff et al. (2000) treat the membership scores as realizations of Dirichlet random variables and are able to use marginal maximum likelihood estimation in a series of classification examples when the number of items J is small. Eroshova (2002) provides a Markov chain Monte Carlo estimation scheme for the GoM model also assuming the Dirichlet distribution on the membership scores. Varki et al. (2000) employ a mixture of point and Dirichlet distributions as the generating distribution for the membership scores in their work.

Independently from the GoM developments, in genetics Pritchard et al. (2000) use a *clustering model with admixture*. For diploid individuals the clustering model assumes that $R = 2$ replications (genotypes) are observed at J distinct locations (loci) and that the membership scores are random Dirichlet realizations. Again, J and N vary in this and related applications. In the Introduction, we briefly described an example of findings obtained via this model in the study on race and population genetic structure by Rosenberg et al. (2002).

A standard assumption in machine learning of text and other objects is that a single characteristic is observed multiple times. For example, for a text document of length L only one distinct characteristic, a word, is observed with $R = L$ realizations. In this set-up, the work of Hofmann (2001)

on *probabilistic latent semantic analysis* treated membership scores as fixed unknown constants and that of Blei et al. (2003) adopted a Dirichlet generating distribution for the membership scores. More recently, this line of modeling has moved from considering a single characteristic (e.g., words in a document) to working with a combination of distinct characteristics. An example that we discussed in this area is by Barnard et al. (2003) who modeled a combination of words and segmented images via a mixed membership structure.

Given this multiplicity of unrelated mixed membership model developments, we should not be surprised by the variety of estimation methods adopted. Broadly speaking, estimation methods are of two types: those that treat membership scores as fixed and those that treat them as random. The first group includes the numerical methods introduced by Hofmann (2003) and by Kovtun et al. (2004b), and joint maximum likelihood type methods described in Manton et al. (1994) and Varki and Cooil (2003) where fixed effects for the membership scores are estimated in addition to the population parameter estimates. The statistical properties of the estimates in these approaches, such as consistency, identifiability, and uniqueness of solutions, are suspect. The second group includes variational estimation methods used by Blei et al. (2003), expectation-propagation methods developed by Minka and Lafferty (2002), joint maximum likelihood approaches of Potthoff et al. (2000) and Varki et al. (2000), and Bayesian MCMC simulations (Pritchard et al. (2002), Erosheva (2002, 2003a)). These methods solve some of the statistical and computational problems, but many other challenges and open questions still remain as we illustrate below.

3 Disability types among older adults

3.1 National Long Term Care Survey

The National Long-Term Care Survey (NLTCs), conducted in 1982, 1984, 1989, 1994, and 1999, was designed to assess chronic disability in the U.S. elderly Medicare-enrolled population (65 years of age or older). Beginning with a screening sample in 1982, individuals were followed in later waves and additional samples were subsequently added maintaining the sample at about 20,000 Medicare enrollees in each wave. The survey aims to provide data on the extent and patterns of functional limitations (as measured by activities of daily living (ADL) and instrumental activities of daily living (IADL), availability and details of informal caregiving, use of institutional care facilities, and death. NLTCs public use data can be obtained from the Center for Demographic Studies, Duke University.

Erosheva (2002) considered the mixed membership model with up to $K = 5$ subpopulations or extreme profiles for the 16 ADL/IADL measures, pooled across four survey waves of NLTCs, 1982, 1984, 1989, and 1994. For each ADL/IADL measure, individuals can be either disabled or healthy. Thus the

data form a 2^{16} contingency table. The table has 65,536 cells, only 3,152 of which are non-zero and there are a total of $N = 21,574$ observations. This is a large sparse contingency table that is not easily analyzed using classical statistical methods such as those associated with log-linear models.

3.2 Applying the mixed membership model

Following the GoM structure for dichotomous variables, we have $J = 16$ dichotomous characteristics observed for each individual and the number of replications R is 1. For each extreme profile k , the probability distribution for characteristic j , $f(x_j|\theta_{kj})$ is binomial parameterized by the probability of the positive response μ_{kj} .

We assume that the membership scores follow a Dirichlet distribution D_α and employ Monte Carlo Markov chain estimation for the latent class representation of the GoM model (Erosheva (2003a)). We obtain posterior means for the response probabilities of the extreme profiles and posterior means of the membership scores conditional on observed response patterns. Estimated response probabilities of the extreme profiles provide a qualitative description of the extreme categories of disability as tapped by the 16 ADL/IAD measures while the estimated parameters α of the Dirichlet distribution describe the distribution of the mixed membership scores in the population.

Although the Deviance Information Criteria (Spiegelhalter et al., (2002)) indicates an improvement in fit for K increasing from 2 to 5 with the largest improvement for K going from 2 to 3, other considerations point out that a $K = 4$ solution might be appropriate for this data set (Erosheva (2002)). In Table 1, we provide posterior means and standard deviation estimates for the parameters of the GoM model with four extreme profiles. The estimates of ξ_i and α_0 reported in Table 1 and their product gives the vector of Dirichlet distribution parameters. The estimated distribution of the membership scores is bathtub shaped.

One of the most significant findings in this analysis is based on examining interpretations of the extreme profiles for the mixed membership models for $K = 4, 5$ which rejects the hypothesis of a unidimensional disability structure, i.e., the extreme profiles are qualitatively different and can not be ordered by severity. In particular, individuals at two of the estimated extreme profiles can be described as mostly cognitive and mostly mobility impaired individuals. For more details on the analysis and substantive findings see Erosheva (2002).

4 Classifying publications by topic

4.1 Proceedings of the National Academy of Sciences

The *Proceedings of the National Academy of Sciences* (PNAS) is the world's most cited multidisciplinary scientific journal. Historically, when submitting

k	1	2	3	4
$\mu_{k,1}$	0.000 (3e-04)	0.002 (2e-03)	0.001 (6e-04)	0.517 (1e-02)
$\mu_{k,2}$	0.000 (3e-04)	0.413 (1e-02)	0.001 (5e-04)	0.909 (7e-03)
$\mu_{k,3}$	0.001 (5e-04)	0.884 (1e-02)	0.018 (8e-03)	0.969 (5e-03)
$\mu_{k,4}$	0.007 (2e-03)	0.101 (6e-03)	0.016 (4e-03)	0.866 (8e-03)
$\mu_{k,5}$	0.064 (4e-03)	0.605 (9e-03)	0.304 (9e-03)	0.998 (2e-03)
$\mu_{k,6}$	0.005 (2e-03)	0.316 (9e-03)	0.018 (4e-03)	0.828 (8e-03)
$\mu_{k,7}$	0.230 (7e-03)	0.846 (7e-03)	0.871 (7e-03)	1.000 (3e-04)
$\mu_{k,8}$	0.000 (2e-04)	0.024 (4e-03)	0.099 (7e-03)	0.924 (7e-03)
$\mu_{k,9}$	0.000 (3e-04)	0.253 (9e-03)	0.388 (1e-02)	0.999 (1e-03)
$\mu_{k,10}$	0.000 (2e-04)	0.029 (5e-03)	0.208 (1e-02)	0.987 (4e-03)
$\mu_{k,11}$	0.000 (3e-04)	0.523 (1e-02)	0.726 (1e-02)	0.998 (2e-03)
$\mu_{k,12}$	0.085 (5e-03)	0.997 (2e-03)	0.458 (1e-02)	0.950 (4e-03)
$\mu_{k,13}$	0.021 (4e-03)	0.585 (1e-02)	0.748 (1e-02)	0.902 (5e-03)
$\mu_{k,14}$	0.001 (7e-04)	0.050 (5e-03)	0.308 (1e-02)	0.713 (8e-03)
$\mu_{k,15}$	0.013 (2e-03)	0.039 (4e-03)	0.185 (8e-03)	0.750 (8e-03)
$\mu_{k,16}$	0.014 (2e-03)	0.005 (2e-03)	0.134 (7e-03)	0.530 (9e-03)
ξ_k	0.216 (2e-02)	0.247 (2e-02)	0.265 (2e-02)	0.272 (2e-02)
α_0	0.197 (5e-03)			

Table 1. Posterior mean (standard deviation) estimates for $K = 4$ extreme profiles. The ADL items are: (1) eating, (2) getting in/out of bed, (3) getting around inside, (4) dressing, (5) bathing, (6) using toilet. The IADL items are: (7) doing heavy house work, (8) doing light house work, (9) doing laundry, (10) cooking, (11) grocery shopping, (12) getting about outside, (13) traveling, (14) managing money, (15) taking medicine, (16) telephoning.

a research paper to the Proceedings, authors have to select a major category from Physical, Biological, or Social Sciences, and a minor category from the list of topics. PNAS permits dual classifications between major categories and, in exceptional cases, within a major category. The lists of topics change over time in part to reflect changes in the National Academy sections. Since in the nineties the vast majority of the PNAS research papers was in the Biological Sciences, our analysis focuses on this subset of publications. Another reason for limiting ourselves to one major category is that we expect papers from different major categories to have a limited overlap.

In the Biological Sciences there are 19 topics. Table 2 gives the percentages of published papers for 1997-2001 (Volumes 94-98) by topic and numbers of dual classification papers in each topic.

4.2 Applying the mixed membership model

The topic labels provide an author-designated classification structure for published materials. Notice that the vast majority of the articles are members of only a single topic. We represent each article by collections of words in

Topic	Number	% Dual	% Dual	More Dual?	Dual?
1 Biochemistry	2578	21.517	33	18.436	338
2 Medical Sciences	1547	12.912	13	.263	84
3 Neurobiology	1343	11.209	9	5.028	128
4 Cell Biology	1231	10.275	10	5.587	111
5 Genetics	980	8.180	14	7.821	131
6 Immunology	865	7.220	9	5.028	39
7 Biophysics	636	5.308	40	22.346	62
8 Evolution	510	4.257	12	6.704	167
9 Microbiology	498	4.157	11	6.145	42
10 Plant Biology	488	4.073	4	2.235	54
11 Developmental Biology	366	3.055	2	1.117	43
12 Physiology	340	2.838	1	0.559	34
13 Pharmacology	188	1.569	2	1.117	34
14 Ecology	133	1.110	5	2.793	16
15 Applied Biological Sciences	94	0.785	6	3.352	7
16 Psychology	88	0.734	1	0.559	22
17 Agricultural Sciences	43	0.359	2	1.117	8
18 Population Biology	43	0.359	5	2.793	4
19 Anthropology	10	0.083	0	0	2
Total	11981	100	179	100	1319

Table 2. Biological Sciences publications in PNAS volumes 94–98, by subtopic, and numbers of papers with dual classifications. The numbers in the final column represent projections based on our model.

the abstract and references in the bibliography. For our mixed membership model, we assume that there is a fixed number of extreme categories or aspects, each of which is characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). A distribution of words and references in each article is given by the convex combination of the aspects’ multinomials weighted by proportions of the article’s content coming from each category. These proportions, or membership scores, determine soft clustering of articles with respect to the internal categories.

Choosing a suitable value for the number of internal categories or aspects, K , in this type of setting is difficult. We have focused largely on two versions of the model, one with eight aspects and the other with ten. The set of parameters in our model is given by multinomial word and reference probabilities for each aspect, and by the parameters of Dirichlet distribution, which is a generating distribution for membership scores. There are 39,616 unique words and 77,115 unique references in our data, hence adding an aspect corresponds to having $39,615 + 77,114 + 1 = 116,730$ additional parameters. Because of the large numbers of parameters involved, it is difficult to assess the extent to which the added pair of aspects actually improve the fit of the model to the data. In a set of preliminary comparisons we found little to choose between

them in terms of fit and greater ease of interpretation for the eight aspect model. In Erosheva et al. (2004) we report on the details of the analysis of the $K = 8$ aspect model and its interpretation and we retain that focus here.

From our analysis of high probability words and references, the 8 aspects of our model have the following interpretation:

1. Intracellular signal transaction, neurobiology.
2. Evolution, molecular evolution.
3. Plant molecular biology.
4. Developmental biology; brain development.
5. Biochemistry, molecular biology; protein structural biology.
6. Genetics, molecular biology; DNA repair, mutagenesis, cell cycle.
7. Tumor immunology; HIV infection.
8. Endocrinology, reporting of experimental results; molecular mechanisms of obesity.

Based on the interpretations, it is difficult to see whether the majority of aspects correspond to a single topic from the official PNAS classifications. To investigate a correspondence between the estimated aspects and the given topics further, we examine aspect “loadings” for each paper. Given estimated parameters of the model, the distribution of each article’s “loadings” can be obtained via Bayes’ theorem. The variational and expectation-propagation procedures give Dirichlet approximations to the posterior distribution $p(\boldsymbol{\lambda}(d), \boldsymbol{\theta})$ for each document d . We employ the mean of this Dirichlet as an estimate of the “weight” of the document on each aspect.

We can gauge the sparsity of the loadings by the parameters of the Dirichlet distribution, which for the $K = 8$ model we estimate as $\alpha_1 = 0.0195$, $\alpha_2 = 0.0203$, $\alpha_3 = 0.0569$, $\alpha_4 = 0.0346$, $\alpha_5 = 0.0317$, $\alpha_6 = 0.0363$, $\alpha_7 = 0.0411$, $\alpha_8 = 0.0255$. This estimated Dirichlet, which is the generative distribution of membership scores, is “bathtub shaped” on the simplex; as a result, articles will tend to have relatively high membership scores in only a few aspects.

To summarize the aspect distributions for each topic, we provide a graphical representation of these values for $K = 8$ and $K = 10$ in Figure 1 and Figure 2, respectively. Examining the rows of Figure 1, we see that, with the exception of Evolution and Immunology, the subtopics in Biological Sciences are concentrated on more than one internal category. The column decomposition, in turn, can assist us in interpreting the aspects. Aspect 8, for example, which from the high probability words seems to be associated with the reporting of experimental results, is the aspect of origin for a combined 37% of Physiology, 30% of Pharmacology, and 25% of Medical Sciences papers, according to the mixed membership model.

Finally, we compare the loadings (posterior means of the membership scores) of dual-classified articles to those that are singly classified. We consider two articles as having similar membership vectors if their loadings are equal for the first significant digit for all aspects. One might consider singly

classified articles that have membership vectors similar to those of dual-classified articles as interdisciplinary, i.e., the articles that should have had dual classification but did not. We find that, for 11 percent of the singly classified articles, there is at least one dual-classified article that has similar membership scores. For example, three biophysics dual-classified articles with loadings 0.9 for the second and 0.1 for the third aspect turned out to have similar loading to 86 singly classified articles from biophysics, biochemistry, cell biology, developmental biology, evolution, genetics, immunology, medical sciences, and microbiology. In the last column of Table 2, we give the numbers of projected additional dual classification papers by PNAS topic.

4.3 An alternative approach with related data

Griffiths and Steyvers (2004) use a related version of the mixed membership model on the words in PNAS abstracts for the years 1991-2001, involving 28,154 abstracts. Their corpus involves 20,551 words that occur in at least five abstracts, and are not on the “stop list”. Their version of the model does not involve the full hierarchical probability structure. In particular, they employ Dirichlet(α) distribution for membership scores λ , but they fix α at $50/K$, and a Dirichlet(β) distribution for aspect word probabilities, but they fix β at 0.1. These choices lead to considerable computational simplification that allows using a Gibbs sampler for the Monte Carlo computation of marginal components of the posterior distribution.

In Griffiths and Steyvers (2004) they report on estimates of $Pr(data|K)$ for $K=50, 100, 200, 300, 400, 500, 600, 1000$, integrating out the latent variable values. They then pick K to maximize this probability. This is referred to in the literature as a maximum *a posteriori* (MAP) estimate (e.g., see Denison et al. (2002)), and it produces a value of K approximately equal to 300, more than an order of magnitude greater than our value of $K=8$.

There are many obvious and some more subtle differences between our data and those analyzed by Griffiths and Steyvers as well as between our approaches. Their approach differs from ours because of the use of a words-only model, as well as through the simplification involving the fixing of the Dirichlet parameters and through a more formal selection of dimensionality. While we can not claim that a rigorous model selection procedure would estimate the number of internal categories close to 8, we believe that a high number such as $K=300$ is at least in part an artifact of the data and analytic choices made by Griffiths and Steyvers. For example, we expect that using the class of Dirichlet distributions with parameters $50/K$ when $K > 50$ for membership scores biases the results towards favoring many more categories than there are in the data due to increasingly strong preferences towards extreme membership scores with increasing K . Moreover, the use of the MAP estimate of K has buried within it an untenable assumption, namely that $Pr(K)$ constant *a priori*, and pays no penalty for an excessively large number of aspects.

4.4 Choosing K to describe PNAS topics

Although the analyses in the two preceding subsections share the same general goal, i.e., detecting the underlying structure of PNAS research publications, they emphasize two different levels of focus. For the analysis of words and references in Erosheva et al. (2004), we aimed to provide a succinct high-level summary of the population of research articles. This led us to narrow our focus to research reports in biology and to keep the numbers of topics within the range of the current classification scheme. We found the results for $K = 8$ aspects were more easily interpretable than those for $K = 10$ but because of time and computational expense we did not explore more fully the choice of K .

For their word-only model, Griffiths and Steyvers (2004) selected the model based on $K = 300$ which seems to be aimed more at the level of prediction, e.g., obtaining the most detailed description for each paper as possible. They worked with a database of all PNAS publications for given years and considered no penalty for using a large number of aspects such as that which would be associated with the Bayesian Information Criterion applied to the marginal distributions integrating out the latent variables.

Organizing aspects hierarchically, with sub-aspects having mixed membership in aspects, might allow us to reconcile our higher level topic choices with their more fine-grained approach.

5 Summary and concluding remarks

In this paper we have described a Bayesian approach to a general mixed membership model that allows for:

- Identification of internal clustering categories (unsupervised learning).
- Soft or mixed clustering and classifications.
- Combination of types of characteristics, e.g., numerical values and categories, words and references for documents, features from images.

The ideas behind the general model are simple but they allow us to view seemingly disparate developments in soft clustering or classification problems in diverse fields of application within the same broad framework. This unification has at least two salutary implications:

- Developments and computational methods from one domain can be imported to or shared with another.
- New applications can build on the diverse developments and utilize the general framework instead of beginning from scratch.

When the GoM model was first developed, there were a variety of impediments to its implementation with large datasets, but the most notable were technical issues of model identifiability and consistency of estimation,

since the number of parameters in the model for even a modest number of groups (facets) is typically greater than the number of observations, as well as possible multi-modal likelihood functions even when the model was properly identified. These technical issues led to practical computational problems and concerns about the convergence of algorithms. The Bayesian hierarchical formulation described here allows for solutions to a number of these difficulties, even in high dimensions, as long as we are willing to make some simplifying assumptions and approximations. Many challenges remain, both statistical and computational. These include computational approaches to full posterior calculations; model selection (i.e., choosing K), and the development of extensions of the model to allow for both hierarchically structured latent categories and dependencies associated with longitudinal structure.

Keywords

Aging and disability; Grade of membership model; Hierarchical models; Latent Dirichlet allocation; Latent variables; Monte Carlo Markov chain methods; Scientific publication topics; Text classification; Variational approximation.

Acknowledgments

We are indebted to John Lafferty for his collaboration on the analysis of the PNAS data which we report here, and to Adrian Raftery and Christian Robert for helpful discussions on selecting K . Erosheva's work was supported by NIH grants 1 RO1 AG023141-01 and R01 CA94212-01, Fienberg's work was supported by NIH grant 1 RO1 AG023141-01 and by the Centre de Recherche en Economie et Statistique of the Institut National de la Statistique et des Études Économiques, Paris, France.

References

- BARNARD, K., DUYGULU, P., FORSYTH, D., de FREITAS, N., BLEI, D. M., and JORDAN, M. I. (2003): Matching words and pictures. *Journal of Machine Learning Research*, 3, 1107–1135.
- BLEI, D. M., JORDAN, M. I. (2003a): Modeling annotated data. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 127–134.
- BLEI, D. M., JORDAN, M. I., and NG, A. Y. (2003b): Latent Dirichlet models for application in information retrieval. In J. Bernardo, et al. eds., *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*, Oxford University Press, Oxford, 25–44.
- BLEI, D. M., NG, A. Y., and JORDAN, M. I. (2003c): Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1002.

- BRANDTBERG, T. (2002): Individual tree-based species classification in high spatial resolution aerial images of forests using fuzzy sets. *Fuzzy Sets and Systems*, 132, 371–387.
- COHN, D., and HOFMANN, T. (2001): The missing link: A probabilistic model of document content and hypertext connectivity. *Neural Information Processing Systems (NIPS*13)*, MIT Press .
- COOIL, B. and VARKI, S. (2003): Using the conditional Grade-of-Membership model to assess judgment accuracy. *Psychometrika*, 68, 453–471.
- DENISON, D.G.T., HOLMES, C.C., MALLICK, B.K., AND SMITH, A.F.M. (2002): *Bayesian Methods for Nonlinear Classification and Regression*. Wiley, New York.
- EROSHEVA, E. A. (2002): *Grade of Membership and Latent Structure Models With Application to Disability Survey Data*. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University. PhD thesis, Carnegie Mellon University.
- EROSHEVA, E. A. (2003a): Bayesian estimation of the Grade of Membership Model. In J. Bernardo, et al. eds., *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*, Oxford University Press, Oxford, 501–510.
- EROSHEVA, E. A. (2003b): Partial Membership Models With Application to Disability Survey Data In H. Bozdogan, ed. *New Frontiers of Statistical Data Mining, Knowledge Discovery, and E-Business*, CRC Press, Boca Raton, 117–134.
- EROSHEVA, E.A., FIENBERG, S.E., and LAFFERTY, J. (2004): Mixed Membership Models of Scientific Publications. *Proceedings of the National Academy of Sciences*, in press.
- GRIFFITHS, T. L., and STEYVERS, M. (2004): Finding scientific topics. *Proceedings of the National Academy of Sciences*, in press.
- HOFMANN, T. (2001): Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42, 177–196.
- KOVTUN, M., AKUSHEVICH, I., MANTON, K.G., and TOLLEY, H.D. (2004a): Grade of membership analysis: Newest development with application to National Long Term Care Survey. Unpublished paper presented at Annual Meeting of Population Association of America (dated March 18, 2004).
- KOVTUN, M., AKUSHEVICH, I., MANTON, K.G., and TOLLEY, H.D. (2004b): Grade of membership analysis: One possible approach to foundations. Unpublished manuscript.
- MANTON, K. G., WOODBURY, M. A., and TOLLEY, H. D. (1994): *Statistical Applications Using Fuzzy Sets*. Wiley, New York.
- MINKA, T. P., and LAFFERTY, J., (2002): Expectation-propagation for the generative aspect model. *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, Morgan Kaufmann, San Francisco, pages 352–359.
- NURMBERG, H.G., WOODBURY, M.A., and BOGENSCHUTZ, M.P. (1999): A mathematical typology analysis of DSM-III-R personality disorder classification: grade of membership technique. *Compr Psychiatry*, 40, 61–71.
- POTTHOFF, R. F., MANTON, K. G., and WOODBURY, M. A., (2000): Dirichlet generalizations of latent-class models. *Journal of Classification*, 17, 315–353.
- PRITCHARD, J. K., STEPHENS, M., and DONNELLY, P., (2000): Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959.

- 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.
- SEETHARAMAN, P.B., FEINBERG, F.M., and CHINTGUNTA, P.K. (2001): Product line management as dynamic, attribute-level competition. Unpublished manuscript.
- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., and VAN DER LINDE, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Methodological*, 64, 1-34.
- TALBOT, B.G., WHITEHEAD, B.B., and TALBOT, L.M. (2002): Metric Estimation via a Fuzzy Grade-of-Membership Model Applied to Analysis of Business Opportunities. *14th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2002*, 431–437.
- TALBOT, L.M. (1996): A Statistical Fuzzy Grade-of-Membership Approach to Unsupervised Data Clustering with Application to Remote Sensing. Unpublished Ph.D. dissertation, Department of Electrical and Computer Engineering, Brigham Young University.
- VARKI and CHINTGUNTA (2003): The augmented latent class model: Incorporating additional heterogeneity in the latent class model for panel data. *Journal of Marketing Research*, forthcoming.
- VARKI, S., COOIL, B., and RUST, R.T. (2000): Modeling Fuzzy Data in Qualitative Marketing Research. *Journal of Marketing Research*, XXXVII, 480-489.
- WOODBURY, M. A., CLIVE, J. (1974): Clinical pure types as a fuzzy partition. *Journal of Cybernetics*, 4, 111–121.
- WOODBURY, M. A., CLIVE, J., and GARSON, A. (1978): Mathematical typology: A Grade of Membership technique for obtaining disease definition. *Computers and Biomedical Research*, 11, 277–298.

Fig. 1. Graphical representation of mean decompositions of aspect membership scores for $K = 8$. Source: Erosheva et al.(2004).

Fig. 2. Graphical representation of mean decompositions of aspect membership scores for $K = 10$.