Hierarchical Bayesian Mixed-Membership Models and Latent Pattern Discovery

E. M. AIROLDI,* S. E. FIENBERG,** C. JOUTARD,† and T. M. LOVE‡

* Department of Statistics and FAS Center for Systems Biology, Harvard University
** Department of Statistics and Machine Learning Department, Carnegie Mellon University
† Département de Mathématiques, Université Montpellier 2, France
‡ Department of Biostatistics and Computational Biology, University of Rochester

February 23, 2010

Abstract

Hierarchical Bayesian methods expanded markedly with the introduction of MCMC computation in the 1980s, and this was followed by the explosive growth of machine learning tools involving latent structure for clustering and classification. Nonetheless, model choice remains a major methodological issue, largely because competing models used in machine learning often have different parameterizations and very different specifications and constraints. Here, we utilize hierarchical Bayesian mixed-membership models and present several examples of model specification and variations, both parametric and nonparametric, in the context of learning the number of latent groups and associated patterns for clustering units. We elucidate strategies for comparing models and specifications by producing novel analyses of the following two data sets in both parametric and nonparametric settings: (1) a corpus of scientific publications from the *Proceedings of the National Academy of Sciences* where we use both text and references to narrow the choice of the number of latent topics in our publications data; (2) data on functional disabilities of the elderly from the National Long Term Care Survey.

1 Introduction

Although hierarchical models have dominated the Bayesian literature since the early 1970s, several variations on *Hierarchical Bayesian Mixed-Membership Models* (HBMMMs) have recently gained

popularity thanks to their ability to deal with minimal information and noisy labels in a systematic fashion. These models allow each subject of study, e.g., documents or individuals, to belong to more than one class, group, or cluster (Erosheva, Fienberg, and Lafferty, 2004; Erosheva and Fienberg, 2005; Airoldi *et al.*, 2006; Airoldi, 2007).

We can specify HBMMMs in terms of a hierarchy of probabilistic assumptions (i.e., a directed acyclic graph) that involves: (i) observations, x, (ii) latent variables, λ, and (iii) parameters (constants) for the patterns associated with the groups or clusters, θ. The likelihood of the data is then $\ell(x|\theta) = \int_{\lambda} \ell(x, \lambda|\theta) D_{\alpha}(d\lambda)$, where $D_{\alpha}(d\lambda)$ is a prior distribution over the latent variables. During pattern discovery, i.e., posterior inference, we condition on the values of the observed data and maximize the likelihood with respect to a set of parameters θ that describe the group patterns.

The focus in pattern discovery with HBMMMs is not on the variable amount of information about the labels for the objects; rather, it is on the hierarchy of probabilistic assumptions that the analyst believes provide the structure underlying the data, which ultimately lead to the likelihood function. Whatever the amount of information about the class labels, full, partial, minimal, or none, we simply treat the information as observations about the attributes and we condition upon it. The missing information about the labels or weights on the classes or groups is recovered during pattern discovery (i.e., via posterior inference) as is the information about other non-observable patterns. In this sense, HBMMMs are essentially *soft-clustering* models in that the *mixed-membership* error model for the labels associates each observation with a vector of memberships that sum to one.

Because of their flexibility, instances of HBMMMs have gained popularity in a variety of applications, e.g., population genetics (Pritchard, Stephens, and Donnelly, 2000; Rosenberg *et al.*, 2002), scientific publications (Blei, Ng, and Jordan, 2003; Erosheva, Fienberg, and Lafferty, 2004; Griffiths and Steyvers, 2004), words and images (Barnard *et al.*, 2003), disability analysis (Erosheva, 2002a,b, 2003; Erosheva, Fienberg, and Joutard, 2007), fraud detection (Neville *et al.*, 2005), biological sequences & networks (Airoldi *et al.*, 2006). HBMMMs are closely related to popular unsupervised data mining methods such as probabilistic principal component analysis (Tipping and Bishop, 1999), parametric independent component analysis, mixtures of Gaussians, factor analysis, and hidden Markov models (Rabiner, 1989).

A fundamental issue of HBMMMs is that of *model choice*, involving the choice of the number of latent categories, groups, or clusters. Positing an explicit model for the category labels requires a choice regarding the number of existing categories in the population, i.e., the *choice* of the crucial model dimension. A parametric model for the labels would assume the existence of a predetermined number, K, of categories, whereas a nonparametric error model would let the number of categories grow with the data. We explore the issue of model choice in the context of HBMMMs, both theoretically and computationally, by investigating the nexus between strategies for model choice, estimation strategies, and data integration in the context of data extracted from scientific publications and measures of disability for Americans aged 65+, cf. (Erosheva, Fienberg, and Joutard, 2007; Joutard *et al.*, 2007; Airoldi *et al.*, 2009).

Overview of the Paper In this paper, we (1) describe HBMMMs a class of models that respond to the challenges introduced by modern applications, and we characterize HBMMMs in terms of their essential probabilistic elements; (2) identify the issue of *model choice* as a fundamental task to be solved in each applied data mining analysis that uses HBMMMs; (3) survey several of the existing strategies for model choice; (4) develop new model specifications, as well as use old ones, and we employ different strategies of model choice to find "good" models to describe problems involving text analysis and survey data; (5) study what happens as we deviate from statistically sound strategies in order to cut down the computational burden, in a controlled experimental setting.

PNAS Biological Sciences Collection Our data consists of abstracts and references for a collection of articles from the *Proceedings of the National Academy of Sciences* for 1997–2001. Erosheva, Fienberg, and Lafferty (2004) and Griffiths and Steyvers (2004) report on their estimates about the number of latent topics, and find evidence that supports a small number of topics (e.g., as few as 8 but perhaps a few dozen) *or* as many as 300 latent topics, respectively. There are a number of differences between the two analyses: the collections of papers were only partially overlapping (both in time coverage and in subject matter), the authors structured their dictionary of words differently, one model could be thought of as a special case of the other but the fitting and

inference approaches had some distinct and non-overlapping features. The most remarkable and surprising difference comes in the estimates for the number of latent topics: Erosheva, Fienberg, and Lafferty (2004) focus on values such as 8 and 10, but admit that a careful study would likely produce somewhat higher values, while Griffiths and Steyvers (2004) present analyses they claim support on the order of 300 topics! Should we want or believe that there are only a dozen or so topics capturing the breadth of papers in PNAS or is the number of topics so large that almost every paper can have its own topic. A touchstone comes from the journal itself, which states that it classifies publications in biological sciences according to 19 topics. When submitting manuscripts to PNAS, authors select a major and a minor category from a predefined list of 19 biological science topics, and possibly those from the physical and/or social sciences.

Here, we develop an alternative set of analyses using the version of the PNAS data on biological science papers analyzed in Erosheva, Fienberg, and Lafferty (2004). We employ both parametric and non-parametric strategies for model choice, and we make use of both text and references of the papers in the collection. This case study gives us a basis to discuss and assess the merit of the various model choice strategies.

Disability Survey Data In the second example, we work with data extracted from the National Long-Term Care Survey (NLTCS) by Erosheva (2002a) to illustrate the important points of our analysis. The NLTCS is a longitudinal survey of the U.S. population aged 65 years and older with waves conducted in 1982, 1984 1989, 1994, 1999 and 2004. It is designed to assess chronic disability among the US elderly population especially those who show limitations in performing some activities that are considered normal for everyday living. These activities are divided into *activities of daily living* (ADLs) and *instrumental activities of daily living* (IADLs). ADLs are basic activities of hygiene and healthcare: eating, getting in/out of bed, moving inside the house, dressing, bathing and toileting. IADLs are basic activities necessary to reside in the community: doing light and heavy housework and laundry, cooking, grocery shopping, moving outside the house, traveling, managing money, taking medicine and telephoning. The data extract we work with consists of combined data from the first four survey waves (1982, 1984, 1989, 1994) with $21,574$ individuals and 16 variables (6 ADLs and 10 IADLs). For each activity, individuals are either disabled or

healthy on that activity. We then deal with a 2^{16} contingency table. Of the $2^{16} = 65,536$ possible combinations of response patterns, only $3,152$ occured in the NLTCS sample.

Here we complement the earlier analyses in Erosheva (2002a); Erosheva and Fienberg (2005); Erosheva, Fienberg, and Joutard (2007) and employ both parametric and non-parametric strategies for model choice. We focus on increasing the number of latent profiles to see if larger choices of K result in better descriptions of the data and to find the value of K which best fits the data.

From the case studies we learn that: (i) Independently of the goal of the analysis, e.g., predictive versus descriptive, similar probabilistic specifications of the models often support similar "optimal" choices of K, i.e., the number of latent groups and patterns; (ii) Established practices aimed at reducing the computational burden while searching for the best model lead to biased estimates of the optimal choices for K, i.e., the number of latent groups and patterns.

Arriving at a "good" model is a central goal of empirical analyses. These models are often useful in a predictive sense. Thus our analyses in the present paper are relevant as input to those managing general scientific journals as they re-examine current indexing schemes or consider the possible alternative of an automated indexing system, and to those interested in the implications of disability trends among the US elderly population as the rapid increase in this segment of the population raises issue of medical care and the provision of social security benefits.

2 Characterizing HBMM Models

There are a number of earlier instances of mixed-membership models that have appeared in the scientific literature, e.g., see the review in Erosheva and Fienberg (2005). A general formulation due to Erosheva (2002a), and also described in Erosheva, Fienberg, and Lafferty (2004), characterizes the models of mixed-membership in terms of assumptions at four levels. In the presentation below, we denote subjects with $n \in [1, N]$ and observable response variables with $j \in [1, J]$.

A1–Population Level. Assume that there are K classes or sub-populations in the population of interest and J distinct characteristics measured on each subject. We denote by $f(x_{nj}|\theta_{jk})$ the probability distribution of j-th response variable in the k-th sub-population for the n-th subject, where θ_{jk} is a vector of relevant parameters, $j \in [1, J]$ and $k \in [1, K]$. Within a subpopulation, the

observed responses are assumed to be independent across subjects *and* characteristics.

A2–Subject Level. The components of the membership vector $\lambda_n = (\lambda_{n[1]}, \ldots, \lambda_{n[K]})'$ represent the mixed-membership of the *n-th* subject to the various sub-populations. Conditional on the mixed-membership scores, the response variables x_{nj} are independent of one another, and independent across subjects.

A3–Latent Variable Level. Assume that the vectors λ_n, i.e., the mixed-membership scores of the *n-th* subject, are realizations of a latent variable with distribution D_α, parameterized by vector α.

A4–Sampling Scheme Level. Assume that the R replications of the J distinct response variables corresponding to the *n-th* subject are independent of one another. The probability of observing $\{x^r_{n1}, \ldots, x^r_{nJ}\}_{r=1}^R$, given the parameters, is then

$$Pr\ (\{x^r_{n1}, \ldots, x^r_{nJ}\}_{r=1}^R | \alpha, \theta) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_{n[k]} f(x^r_{nj} | \theta_{jk}) \right) D_\alpha(d\lambda). \tag{1}$$

The number of observed response variables is not necessarily the same across subjects, i.e., $J = J_n$. Likewise, the number of replications is not necessarily the same across subjects and response variables, i.e., $R = R_{nj}$.

3 Strategies for Model Choice

Although pathological cases can be built, where slightly different model specifications lead to quite different analyses, in real situations we expect models with similar probabilistic specifications to suggest an optimal number of groups, K, in the same ballpark.

In our application to scientific publications and survey data we explore the issue of model choice by means of different criteria, two of which are popular in the data mining community; namely, cross-validation and a Dirichlet process prior (Hastie, Tibshirani, and Friedman, 2001; Antoniak, 1974).

Cross-validation is a popular method to estimate the generalization error of a prediction rule (Hastie, Tibshirani, and Friedman, 2001), and its advantages and flaws have been addressed by

many in that context, e.g., see (Ng, 1997). More recently, cross-validation has been adopted to inform the choice about the number of groups and associated patterns in HBMMMs (Barnard *et al.*, 2003; Wang, Mohanty, and McCallum, 2005). Guidelines for the proper use of cross-validation in choosing the optimal number of groups K, however, has not been systematically explored. One of the goals of our case studies is that of assessing to what extent cross-validation can be trusted to estimate the underlying number of topics or disability profiles. In particular, given the non-negligible influence of hyper-parameter estimates in the evaluation of the held-out likelihood, i.e., the likelihood on the testing set, we discover that it is important not to bias the analysis towards unprincipled estimates of such parameters, or with arbitrary ad-hoc choices that are not justifiable using preliminary evidence, i.e., either in the form of prior knowledge, or outcome of the analysis of training documents. Expert prior information was sought, but those consulted expressed views on only a relatively narrow component of the data that did not inform the hyper-parameters. In this situation, estimates obtained following good statistical properties, e.g., empirical Bayes or maximum likelihood estimates, should be preferred to others (Carlin and Louis, 2005).

Positing a Dirichlet process prior on the number of latent topics is equivalent to assuming that the number of latent topics grows with the log of the number of documents or individuals (Ferguson, 1973; Antoniak, 1974). This is an elegant model selection strategy in that the selection problem becomes part of the model itself, although in practical situations it is not always possible to justify. A nonparametric alternative to this strategy uses the Dirichlet Process prior as an infinite dimensional prior with a specific parametric form as a way to mix over choices of K, e.g., see (McAuliffe, Blei, and Jordan, 2006). This prior appears reasonable for static analyses of scientific publications that appear in a specific journal.

The statistical and data mining literatures contain many criteria and approaches to deal with the issue of model choice, e.g., reversible jump MCMC techniques, Bayes factors and other marginal likelihood methods, cross-validation, and penalized likelihood criteria such as the Bayesian Information Criterion (BIC), the Akaike information criterion (AIC), the deviance information criterion (DIC), and minimum description length (MDL). For further details and discussion see Joutard *et al.* (2007).

4 Case Study: PNAS 1997–2001

In this section we introduce model specifications to analyze the collection of papers published in PNAS, which were submitted by the respective authors to the section on biological sciences. Earlier related analyses appear in Erosheva, Fienberg, and Lafferty (2004); Griffiths and Steyvers (2004). After choosing an optimal value for the number of topics, K^*, and its associated words and references usage patterns, we also examine the extent to which they correlate with the actual topic categories specified by the authors.

We organize our models into finite and infinite mixtures, according to the dimensionality of the prior distribution, D_α, posited at the latent variable level. We characterize an article, or document, by the words in its abstract and the references in its bibliography. Introducing some notation, we observe a collection of N documents. The n-th document is represented as (x_{1n}, x_{2n}). We assume that words and references come from finite discrete sets (vocabularies) of sizes V_1 and V_2, respectively. For simplicity, we assume that the vocabulary sets are common to all articles, independent of the publication time, although this assumption can be relaxed (Joutard *et al.*, 2007). We assume that the distribution of words and references in an article is driven by an article's membership in each of K basis categories, $\lambda = (\lambda_1, \ldots, \lambda_K)$, and we denote the probabilities of the V_1 words and the V_2 references in the kth pattern by θ_{k1} and θ_{k2}, for $k = 1, 2, \ldots, K$. These vectors of probabilities define the multinomial distributions over the two vocabularies of words and references for each basis semantic pattern. Below, whenever the analysis refers to a single document, we omit the document index n.

4.1 Finite Mixture Model

For an article with R_1 words in the abstract and R_2 references in the bibliography, the generative sampling process is as follows:

1. Sample $\lambda \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$,
   where $\alpha_k = \alpha$, for all k.

2. Sample $x_1 \sim \text{Multinomial}(p_\lambda, R_1)$,

where $p_\lambda = \sum_{k=1}^{K} \lambda_k \theta_{k1}$.

3. Sample $x_2 \sim \text{Multinomial}(q_\lambda, R_2)$,
   where $q_\lambda = \sum_{k=1}^{K} \lambda_k \theta_{k2}$.

The conditional probability of words and references in a article is then

$$\Pr\left((x_1, x_2)|\theta, \alpha\right) = \int \prod_{j=1}^{2} \prod_{v=1}^{V_j} \left(\sum_{k=1}^{K} \lambda_k \theta_{kj[v]} \right)^{x_{j[v]}} dD_\alpha(\lambda).$$

The hyper-parameters of this model are the symmetric Dirichlet parameter α, and the multinomial parameters for words, θ_{k1}, and references, θ_{k2}, for each of the latent topics[1] $k = 1, \ldots, K$. That is, through corresponding pairs of θ vectors (θ_{k1} and θ_{k2}) we define a parametric representation of each of the K sub-populations (see assumption A1 in Section 2), which we refer to as topics in this application. Technically, they are pairs of latent distributions over the vocabulary and the set of known citations. In other words, element v of θ_{k1} encodes the probability of occurrence of the v-th word in the vocabulary (containing V_1 distinct words) when the k-th topic is active, with the constraint that $\sum_v \theta_{k1[v]} = 1$ for each k. Similarly, element v of θ_{k2} encodes the probability of occurrence of the v-th reference in the set of known citations (V_2 of them) when the k-th topic is active. In this finite mixture model, we assume that the number of latent topics is unknown but fixed at K. Our goal is to find the optimal number of topics, K^*, which gives the best description of the collection of scientific articles.

4.2 Infinite Mixture Model

In the infinite mixture case we posit a simpler and more traditional type of clustering model, by assuming that each article is generated by one single topic. However, in this case we do not need to fix the unknown number of topics, K, prior to the analysis. This full membership model can be thought of as a special case of the mixed membership model where, for each article, all but one of the membership scores are restricted to be zero. As opposed to traditional finite mixture models

[1] In this application, we refer to the sub-populations of assumption A1 in Section 2 as "topics.". Despite the suggestive semantics, topics are pairs of latent distributions over the vocabulary and the set of known citations, from a statistical perspective.

that are formulated conditional on the number of latent categories, however, this model variant allows the joint estimation of the characteristics of the latent categories, θ, and of the number of latent categories, K. That is, prior to the analysis, the number of sub-populations (see assumption A1 in Section 2) is unknown and possibly infinite.

We assume an infinite number of categories and implement this assumption through a Dirichlet process prior for λ, D_α, introduced and discussed in Ferguson (1973); Neal (2000). The distribution D_α models the prior probabilities of latent pattern assignment for the collection of documents. In particular, for the nth article, given the set of assignments for the remaining articles, λ_{-n}, this prior puts a probability mass on the kth pattern (out of K distinct patterns observed in λ_{-n}) which is proportional to the number of documents associated with it. The prior distribution also puts a probability mass on a new, $(K+1)$th latent semantic pattern, that is distinct from the patterns $(1,\ldots,K)$ observed in λ_{-n}. That is, D_α entails prior probabilities for each component of λ as follows:

$$p(\lambda_{n[k]}=1|\lambda_{-n}) = \begin{cases} \frac{m(-n,k)}{N-1+\alpha} & \text{if } m(-n,k)>0 \\ \frac{\alpha}{N-1+\alpha} & \text{if } k=K(-n)+1 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where λ_{-n} denotes the full-membership vectors for all but the nth document; $m(-n,k)$ is the number of documents that are associated with the kth latent pattern, other than the nth document, i.e., $m(-n,k)=\sum_{m=1}^{N}\mathbb{I}(\lambda_{m[k]}=1, m\neq n)$; and $K(-n)$ is the number of distinct latent patterns that are associated with at least one document other than the nth document.

The generative sampling process for the infinite mixture model is as follows:

1. Sample $\lambda \sim \text{DirichletProcess}(\alpha)$

2. For each of the N articles

   2.1. Sample $x_{1n} \sim \text{Multinomial}(\theta_{c1}, R_1)$ where $\lambda_{n[c]}=1$.

   2.2. Sample $x_{2n} \sim \text{Multinomial}(\theta_{c2}, R_2)$ where $\lambda_{n[c]}=1$.

The hyper-parameters of this model are the scaling parameter of the Dirichlet process prior, α, and the multinomial parameters for words, θ_{k1}, and references, θ_{k2}, for each of the latent topics

$k = 1, \ldots, K$.

In this model, we assume that the number of latent topics, K, is unknown and possibly infinite, through the prior for λ, D_α, and we examine the posterior distribution of λ.

4.3 Empirical Results

We fit six models for latent topics in the PNAS dataset: using words alone or with references, finite or infinite mixture models, and (for finite mixture) fitted or fixed Dirichlet parameter α. We used variational methods for the finite mixtures, and MCMC methods for the infinite mixture. For further details see Airoldi (2007); Joutard *et al.* (2007)

In Figure 1, we give the cross-validated log-likelihood obtained for the four finite mixture models (at $K = 5, 10, \cdots, 50, 100, 200, 300$). The plots of the log likelihood in Figure 1 suggest we choose a number of topics between 20 and 40 whether words or words and references are used. Values of K that maximize the held-out loglikelihood are somewhat greater when the database is expanded with references compared to when the database contains only words. Thus, adding references allows for finer refinement of topics.

The infinite model generates a posterior distribution for the number of topics, K, given the data. Figure 2 shows the posterior distribution ranges from 17 to 28 topics. The maximum a posteriori estimate of K is smaller for the model with words and references compared to the model with words only. Further, the posterior range of K is smaller for the model with words and references. Thus adding references to the models reduces the posterior uncertainty about K.

Overall, the values of K in the region of 15-40 are supported by all our models. A number in that range would be a plausible choice for the number of latent basis categories in PNAS biological sciences research reports, 1997-2001. By choosing $K = 20$ topics, we can meaningfully interpret all of the word and reference usage patterns. We then fit the data with a 20 topics model for the finite mixture model using words and references and focused on the interpretation of the 20 topics.

To summarize the distribution of latent aspects over distributions, we provide a graphical representation of the distribution of latent topics for each of the PNAS submission classification in Figure 3. When the references are included, the relationship of estimated latent categories with

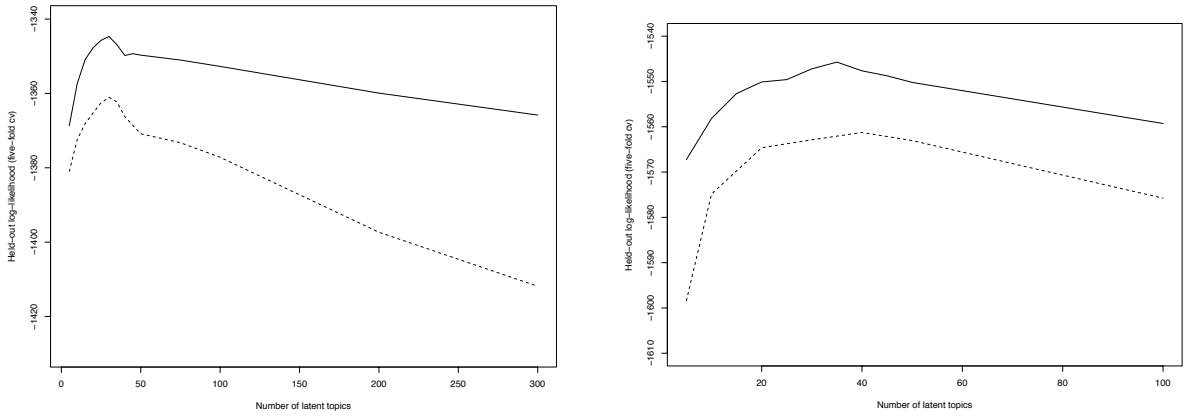

Figure 1: Left Panel: Average held-out log-likelihood corresponding to four mixed-membership models we fit to the PNAS Biological Sciences articles, 1997-2001, using words from article abstracts (left panel) and words and references (right panel). Solid lines correspond to models fitted by estimating the hyperparameter α; dashes lines correspond to models fitted by setting the hyperparameter equal to $\alpha = 50/K$.

designated PNAS classifications becomes more composite for both estimation methods. Models where α is fixed are less sparse than the corresponding models with α fit to the data. For 20 latent topics, we fix $\alpha = 50/20 = 2.5 > 1$—each latent topic is expected to be present in each document and a priori we expect equal membership in each topic. By contrast the fitted values of α less than one lead to models that expect articles to have high membership in a small number of topics. See Joutard *et al.* (2007) for further consequences of these assumptions. The PNAS topics tend to correspond to fewer latent topics when we estimate α and to low to moderate numbers topics when we fix α.

Further, by examining Figure 3, we note that nearly all of the PNAS classifications are represented by several word and reference usage patterns in all of the models. This highlights the distinction between the PNAS submission categories and the discovered latent topics. The assigned PNAS categories follow the structure of the historical development of Biological Sciences and the divisions/departmental structures of many medical schools and universities. These latent topics, however, are structured around the current biological research interests.

We consider the best model of words and references, with $K^* = 20$, and we offer the following interpretation of *all* of the topics to demonstrate what a reasonable model fit should look like:

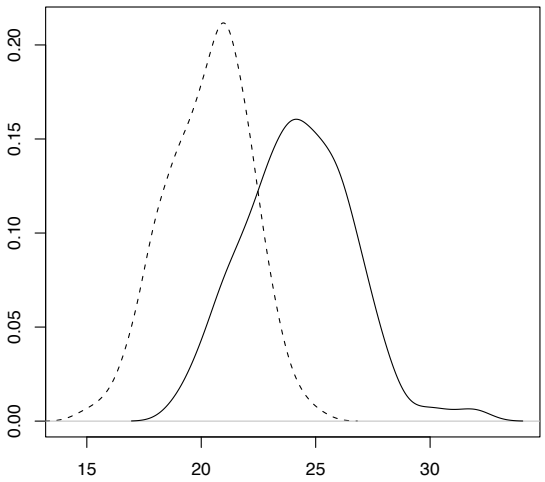

Figure 2: Posterior distribution of the number of mixture components K for infinite models for the PNAS Biological Sciences articles, 1997-2001, using words from article abstracts (solid line) and words and references (dashed line).

5 Case Study: Disability Profiles

All our models are special cases of HBMMMs presented in Section 2. Below, we organize them into finite and infinite mixture models, as before, according to the dimensionality of the prior distribution, D_α, posited at the latent variable level—assumption A3.

We characterize an individual by a set of responses, x_{jn} for $j = 1, \ldots, J$, which were measured through a questionnaire. In our analysis we selected $J = 16$ binary responses that encode answers

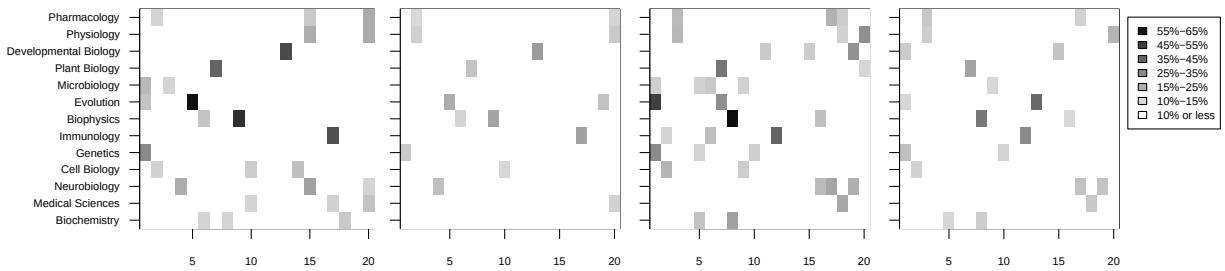

Figure 3: Estimated average mixed membership of articles in 20 estimated topics by PNAS submission classifications. In each panel, we plot average membership values for each submission category (ordered on the Y axis) in the topics (ordered on the X axis). Panels 1 and 2 represent models with only words while panels 3 and 4 use words and references. Panels 1 and 3 represent models with α estimated from the data while panels 2 and 4 use a fixed value of α.

Topic	Interpretation
1	population genetics
2	enzymes by protein kinases
3	problems of hormone levels
4 & 5	nuclear activity production of cdna and mrna & catalysts for dna copying
6 & 12	HIV and immune response & T-cell response to HIV infection
7	plant evolution and phylogenetic relationships
8 & 11	protein structure and folding & protein promotion by transcription binding factors
9	procedural explanations
10	genetic mutation
14	cancer markers
13 & 18	mutant mice and tumor suppression & tumor treatment for mice and humans
15	bone marrow stem cells
16	functional and visual responses to changes in the brain
17	neurons and neurotransmitters
19	nervous system development
20	electrical excitability of cell membranes

Table 1: A post-analysis interpretation for the model with $K^* = 20$ basis catagories.

to questions about the ability to perform six *activities of daily living* (ADL) and ten *instrumental activities of daily living* (IADL). The j-th response, x_{jn}, is recorded as zero if the n-th individual does not have problems performing the j-th activity (he is considered healthy, to that extent, for the purpose of the survey), whereas it is recorded as one if the n-th individual has problems performing the j-th activity (he is considered disabled, to that extent, for the purpose of the survey).

5.1 Finite Mixture Model

To carry out the analysis of the NLTCS data in the finite mixture setting we use the GoM model described in Erosheva, Fienberg, and Joutard (2007); Joutard *et al.* (2007), which posits the following generative process for all N individuals in the survey.

1. Sample $\theta_{jk} \sim$ Beta (σ_1, σ_2) for each j and k.

2. For each of the N seniors

   2.1 Sample $\lambda_n \sim$ Dirichlet $(\alpha_{[1]}, \ldots, \alpha_{[K]})$.

   2.2 Sample $x_{jn} \sim$ Bernoulli $(p_{j\lambda})$ for each j,
   where $p_{j\lambda} = \sum_{k=1}^{K} \lambda_k \theta_{jk}$

We sample the elements of θ from a symmetric Beta distribution with fixed hyper-parameter $\sigma_1 = \sigma_2 = 1$. Note that the distribution on λ is not the symmetric distribution we used in the previous case study, in the finite setting. In this model, θ is a matrix that encodes the probability of being disabled with respect to each one of the 16 activities for seniors who display disability characteristics specific to each of the K latent profiles. That is, θ_{jk} is the probability of being disabled with respect to the j-th activity for a person who "belongs" completely to the k-th latent profile. Note that in this model there are no constraints on the sum of the total probability of having being disabled given any specific profile. For example, $\sum_{j=1}^{J} \theta_{jk}$ is not necessarily one as in the model of Section 4. The hyper-parameters of this model are α and σ. In Joutard *et al.* (2007), we develop a variational approximation to perform posterior inference on such hyper-parameters, and on the latent variables λ_n.

In our analyses, we also consider a fully Bayesian version of the GoM model, following Erosheva (2002a), which posits the following generative process for all N individuals in the survey.

1. Sample $\xi \sim D_\alpha$.
2. Sample $\alpha_0 \sim$ Gamma (τ_1, τ_2).
3. Sample $\theta_{jk} \sim$ Beta (σ_1, σ_2) for each j and k.
4. For each of the N seniors
   - 4.1. Sample $\lambda_n \sim$ Dirichlet $(\alpha_0 \xi_{[1]}, \ldots, \alpha_0 \xi_{[K]})$.
   - 4.2. Sample $x_{jn} \sim$ Bernoulli $(p_{j\lambda})$ for each j,
     where $p_{j\lambda} = \sum_{k=1}^{K} \lambda_k \theta_{jk}$.

In this fully Bayesian setting we fix the hyper-parameter for convenience. According to our model specifications D_α is a symmetric Dirichlet distribution with fixed hyper-parameter $\alpha_1 = \cdots = \alpha_K = 1$. The k-th component of ξ, $\xi_{[k]}$, represents the proportion of the seniors in the survey who express traits of the k-th latent disability profile. Further, we fix a diffuse Gamma distribution, $\tau_1 = 2$ and $\tau_2 = 10$, to control for the tails of the Dirichlet distribution of the mixed membership vectors, λ_n.

In both of the finite mixture models we presented in this section, we assume that the number of latent profiles is unknown but fixed at K. Our goal is to find the number of latent disability profiles, K^*, which gives the best description of the population of seniors.

5.2 Infinite Mixture Model

In the infinite setting we do not fix the number of sub-populations K. As in the previous case study, we restrict subjects (elderly Americans) to complete membership in one group (profile) and the mixed membership vectors $\lambda_{1:N}$ reduce to single membership vectors. The generative sampling process for the infinite mixture model is as follows:

1. Sample $\lambda \sim \text{DirichletProcess}(\alpha)$.
2. Sample $\theta_{jk} \sim \text{Beta}\ (\sigma_1, \sigma_2)$ for each j and k.
3. Sample $x_{jn} \sim \text{Bernoulli}(\theta_{jc})$ where $\lambda_{n[c]} = 1$ for each j and n.

Here D_α is the Dirichlet process prior described in Section 4.2. As in the finite models, we specify a symmetric Beta distribution for the disability probabilities, θ, however, here we fix $\sigma_1 = \sigma_2 = 10$ to make moderate disability probabilities more likely *a priori* than extreme probabilities. Further, we fix the hyper-parameter of the Dirichlet process prior at $\alpha = 1$, which encodes "indifference" toward additional groups.

5.3 Empirical Results

We fit three models for disability propensity profiles to the NLTCS: the finite mixture with random Dirichlet parameter α, the finite mixture with fixed but unknown α using variational and MCMC methods, and the infinite mixture model using MCMC methods. See Joutard *et al.* (2007); Erosheva, Fienberg, and Joutard (2007); Airoldi (2007) for details about inference and a variety of different approaches to the choice of K, including a method based on residuals for the most frequent response patterns, and information criteria such as DIC and BIC. These analyses yield results consistent with those for cross-validation using variational approximation methods, shown in Figure 4, which suggest a choice of 8 or 9 profiles.

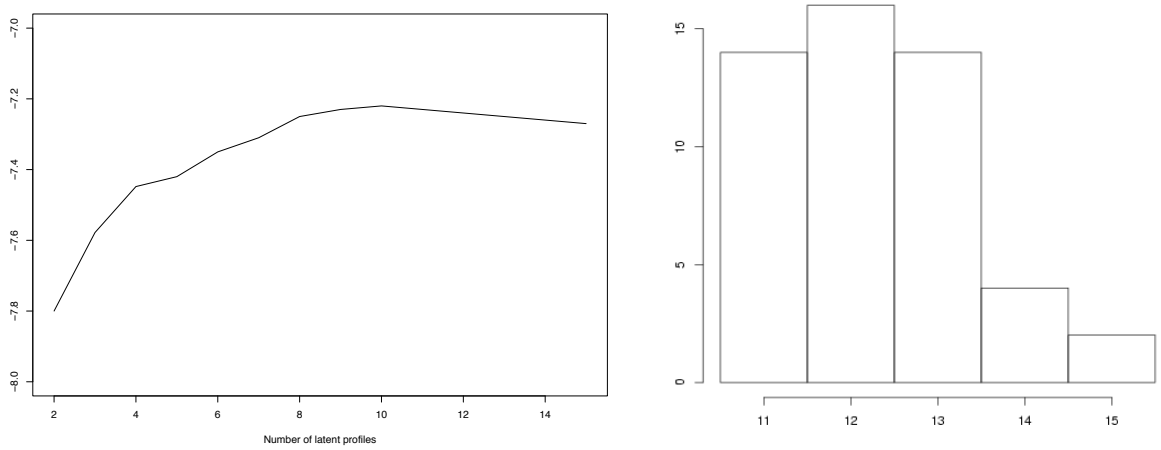

Figure 4: Left Panel: Log-likelihood (5 fold cv) for $K = 2, \ldots, 10, 15$ for the finite model. Right Panel: Posterior distribution of K for the infinite model.

The infinite model generates the posterior distribution for the number of profiles, K, in Figure 4 which in concentrated on from 11 to 15 profiles. We expect that the infinite model requires more profiles because it involves "hard clustering."

Multiple criteria suggest that $K = 9$ is a reasonable choice for the NLTCS data. Figure 5 shows the latent profiles obtained for the 9 profiles GoM model using MCMC methods. The conditional response probabilities represented on the Y-axis are the posterior mean estimates of $\theta_{jk} = P(x_{jn} = 1|\lambda_{n[k]} = 1)$, the probability of being disabled on the activity j for a complete member of latent profile k. We can clearly distinguish two profiles for "healthy" individuals; these are the lower curves (the solid, black curve and the dashed, black curve). The upper curve (solid, grey curve) corresponds to seriously "disabled" individuals since most of the probabilities are greater than 0.8. One profile (long-dashed, grey curve) has the second highest values for the IADLs "managing money", "taking medicine" and "telephoning". This focuses on individuals with some cognitive impairment. The profile with the second highest probabilities for most of the ADLs/IADLs (dashed, grey curve) characterizes "semi-disabled" individuals. The profile with very high probabilities for all the activities involving mobility including the IADL "outside mobility" (dot-dashed, grey curve) characterizes mobility-impaired individuals. Another profile characterizes individuals who are relatively healthy but can't do "heavy housework" (long-dashed, black curve).

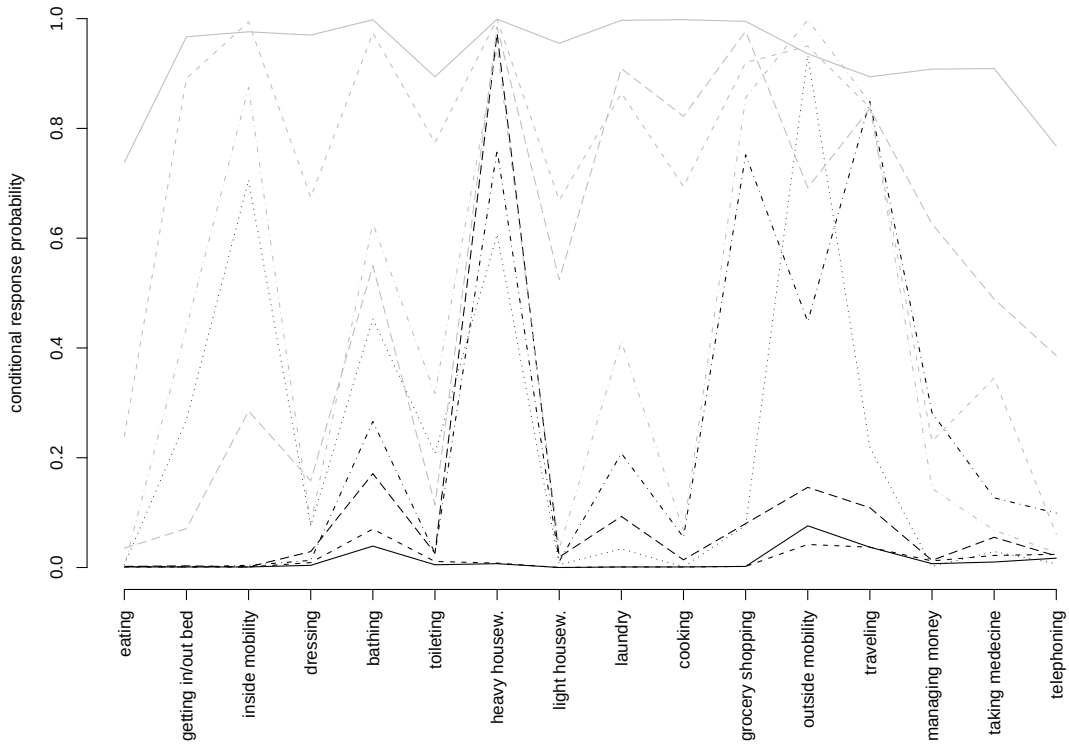

Figure 5: Latent profiles ($\theta_{\cdot k}$'s) for the GoM model with K=9.

The two remaining profiles (the dot-dashed, black curve and the dotted, black curve) corresponds to individuals who are "semi-healthy" since they show limitations in performing some physical activities.

We found similar interpretations with the estimates based on variational methods and MCMC methods despite some differences in the estimated values of the conditional disability propensity probabilities θ_{jk}.

6 Summary

In this paper, we have studied the issue of model choice in the context of mixed-membership models. Often the number of latent classes or groups is of direct interest in applications, but it is always an important element in determining the fit and meaning of the model.

We used extensions of "latent Dirichlet allocation" (LDA) to analyze a corpus of PNAS biological sciences publications from 1997 to 2001. We included k-fold cross-validation and the Dirichlet process prior among our approaches for selecting the number of latent topics, focusing on six combinations of models and model choice strategies. We focused on $K = 20$ topics, a value that appears to be within the range of possibly optimal numbers of topics, and we saw that the resulting topics were easily interpretable and profile popular research subjects in biological sciences, in terms of the corresponding words and references usage patterns. Much higher choices for K lead to far more complex interpretations. For further details see Airoldi *et al.* (2009).

For the NLTCS data, we have developed parametric and nonparametric variations of the GoM model. We performed posterior inference using variational methods and MCMC. We have used different criteria to assess model fit and reached the conclusion that $K = 9$ latent profiles is an appropriate choice for the data set, cf., the related analyses reported in Erosheva, Fienberg, and Joutard (2007). This choice allows us to identify profiles such as the one for individuals who are able to perform all activities except "doing heavy housework." Further, we were able to interpret all 9 of the profiles, although once we reach $K = 5$, the fit seems not to improve markedly.

Acknowledgments This work was partially supported by National Institutes of Health Grant No. R01 AG023141-01, Office of Naval Research Contract No. N00014-02-1-0973, National Science Foundation Grant No. DMS-0240019, and Department of Defense Grant No. IIS0218466, all to Carnegie Mellon University, by National Institutes of Health Grant No. T32 ES007271 to University of Rochester, and by National Science Foundation Grant No. DMS-0907009 to Harvard University. We are especially grateful to Elena Erosheva for comments.

References

E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic block models. *Journal of Machine Learning Research*, 9, 1981–2014, 2008.

E. M. Airoldi. Getting started in probabilistic graphical models. *PLoS Computational Biology*, 3, e252, 2007.

E. M. Airoldi, S. E. Fienberg, C. Joutard, T. M. Love, and S. Shringapure. Re-conceptualizing the Classification of PNAS Articles. Submitted for publication.

D. J. Aldous. Exchangeability and related topics. In *Ecole dété de Probabilités de Saintflour, XIII—1983*, Lecture Notes in Mathematics, Vol. 1117, pp. 1–198. Springer, 1985.

C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, 2(6), 1152–1174, 1974.

K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3, 1107–1135, 2003.

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

B. P. Carlin and T. A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, 2005.

E. A. Erosheva. *Grade of membership and latent structure models with application to disability survey data*. PhD thesis, Department of Statistics, Carnegie Mellon University, 2002.

E. A. Erosheva. Partial membership models with application to disability survey data. In *Proceedings of Conference on the New Frontiers of Statistical Data Mining*, H. Bozdogan, Ed., pp. 117–134. CRC Press, 2002.

E. A. Erosheva. Bayesian estimation of the grade of membership model. In *Bayesian Statistics*, Vol. 7, pp. 501–510. Oxford Univ. Press, 2003.

E. A. Erosheva and S. E. Fienberg. Bayesian mixed membership models for soft clustering and classification. In *Classification—The Ubiquitous Challenge*, C. Weihs and W. Gaul, Eds., pp. 11–26. Springer, 2005.

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

E. A. Erosheva, S. E. Fienberg, and J. Lafferty. Mixed-membership models of scientific publications. *Proceedings of National Academy of Sciences*, 97(22), 11885–11892, 2004.

T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1, 209–230, 1973.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1), 5228–5235, 2004.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed.*. Springer, 2009.

C. J. Joutard, E. M. Airoldi, S. E. Fienberg and T .M. Love. Discovery of latent patterns with hierarchical Bayesian mixed-membership models and the issue of model choice. Chapter 11 in *Data Mining Patterns: New Methods and Applications,* P. Poncelet, F.Masseglia, and M. Teisseire, Eds., pp. 240–275, Information Science Reference, Hershey PA, 2007.

J. McAuliffe, D. Blei, and M. Jordan. Nonparametric empirical Bayes for the Dirichlet process mixture model nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing*, 16, 5–14, 2006. Forthcoming.

T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Uncertainty in Artificial Intelligence*, 2002.

R. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2), 249–265, 2000.

J. Neville, O. Simsek, D. Jensen, J. Komoroske, K. Palmer, and H. Goldberg. Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 11, 2005.

A. Ng. Preventing "overfitting" of cross-validation data. In *Intenational Conference on Machine Learning*, Vol. 14, 1997.

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

L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2), 257–286, 1989.

N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298, 2381–2385, 2002.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61(3), 611–622, 1999.

X. Wang, N. Mohanty, and A. K. McCallum. Group and topic discovery from relations and text. In *Advances in Neural Information Processing Systems*, Vol. 18, 2005.