Discovering Latent Patterns with Hierarchical Bayesian Mixed-Membership Models

Edoardo M. Airoldi, Stephen E. Fienberg, Cyrille Joutard, Tanzy M. Love

May 9, 2006
CMU-ML-06-101

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

There has been an explosive growth of data-mining models involving latent structure for clustering and classification. While having related objectives these models use different parameterizations and often very different specifications and constraints. Model choice is thus a major methodological issue and a crucial practical one for applications.

In this paper, we work from a general formulation of hierarchical Bayesian mixed-membership models in Erosheva [15] and Erosheva, Fienberg, and Lafferty [19] and present several model specifications and variations, both parametric and nonparametric, in the context of the learning the number of latent groups and associated patterns for clustering units. Model choice is an issue within specifications, and becomes a component of the larger issue of model comparison.

We elucidate strategies for comparing models and specifications by producing novel analyses of two data sets: (1) a corpus of scientific publications from the *Proceedings of the National Academy of Sciences* (PNAS) examined earlier by Erosheva, Fienberg, and Lafferty [19] and Griffiths and Steyvers [22]; (2) data on functionally disabled American seniors from the National Long Term Care Survey (NLTCS) examined earlier by Erosheva [15, 16, 17], Erosheva and Fienberg [18].

Our specifications generalize those used in earlier studies. For example, we make use of both text and references to narrow the choice of the number of latent topics in our publications data, in both parametric and nonparametric settings. We then compare our analyses with the earlier ones, for both data sets, and we use them to illustrate some of the dangers associated with the practice of fixing the hyper-parameters in complex hierarchical Bayesian mixed-membership models to cut down the computational burden. Our findings also bring new insights regarding latent topics for the PNAS text corpus and disability profiles for the NLTCS data.

Keywords: Disability data, Model choice, Model specification, Soft clustering,Text analysis, Survey data

Contents

1 Introduction

A general class of models that has recently gained popularity thanks to its ability to deal with minimal information and noisy labels in a systematic fashion can be viewed as special cases or variants of *Hierarchical Bayesian Mixed-Membership Models* (HBMMMs henceforth). Models in this class are viewed by some as bridging situations in which we have full information about the class labels, i.e., a typical classification setting, and situations in which we have no information about the class labels, i.e., a typical clustering setting. From our perspective, they go beyond these models by allowing each object of study, e.g., words or individuals, to belong to more than one class, group, or cluster [19, 18, 1].

HBMMMs are Bayesian models specified in terms of a hierarchy of probabilistic assumptions (i.e., a directed acyclic graph) that involves:

- observations, x,
- latent variables, θ, and
- parameters for the patterns associated with the groups or clusters, β.

The likelihood of the data is then a function

$$\ell\left(x \mid \beta\right) = \int_\theta \ell\left(x, \theta \mid \beta\right)\, D_\alpha(d\theta). \tag{1}$$

where $D_\alpha(d\theta)$ is a prior distribution over the laten 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 β that describe the patterns associated with the group.

The focus in pattern discovery with HBMMMs is not on the variable amount of information about the labels for the objects, but rather it is on the hierarchy of probabilistic assumptions that we believe provide the structure underlying the data and 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., posterior inference) as it 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. The parameters of this error model inform the average abundance of specific class labels without imposing hard constraints, e.g, must-belong or must-not belong. Rather, the constraints are soft, probabilistic constraints.

Because of their flexibility, instances of HBMMMs have recently gained popularity in a variety of applications, e.g., population genetics [35, 37], scientific publications [11, 19, 22], words and images [8], disability analysis [15, 16, 17], fraud detection [31], biological sequences & networks [2]. Further, we note that the class of HBMMMs is closely related to popular unsupervised data mining methods such as probabilistic principal component analysis [43], parametric independent component analysis, mixtures of Gaussians, factor analysis [21], hidden Markov models [36], and state-space models [3]. Few papers recognize that these methods and diverse applications share with HBMMMs a number of fundamental methodological issues such as those we focus on in the present paper.

1.1 The Issue of Model Choice

As we hinted in the discussion above, in these models classification and clustering tasks correspond to the same mathematical problem of maximizing the likelihood. This, in turn, corresponds to resolving the mixed membership of observations to categories (which are typically observed only for a negligible portion of the data), and to pattern discovery. A fundamental issue of HBMMMs is that of "model choice", that is, 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 model. 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 American seniors.

1.2 Overview of the Paper

In this paper, we present the following ideas and results: (1) we 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) we identify the issue of "model choice" as a fundamental task to be solved in each applied data mining analysis that uses HBMMMs; (3) we survey several of the existing strategies for model choice; (4) we 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) we study what happens as we deviate from statistically sound strategies in order to cut down the computational burden, in a controlled experimental setting.

Although "common wisdom" suggests that different goals of the analysis (e.g., prediction of the topic of new documents or of the disability profile of a new person age 65 or over, versus description of the whole collection of documents in terms of topics or of the elderly in terms of disability profiles) would lead us to choose different models, there are few surprises. In fact, from the case studies we learn that:

1. 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;

2. 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 relevant as input to (1) those managing general scientific journals as they re-examine current indexing schemes or considering the possible alternative of an automated indexing system, and (2) 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 Two Motivating Case Studies

Our study is motivated by two recent analyses about a collection of papers published in the *Proceedings of the National Academy of Sciences* (PNAS) [19, 22], and by two recent analyses of National Long Term Care Survey data about disabled American seniors [15, 16, 17, 18, 42].

2.1 PNAS Biological Sciences Collection (1997–2001)

Erosheva et al. [19] and Griffiths & Steyvers [22] 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 come in the estimates for the numbers of latent topics: Erosheva et al. focus on values like 8 and 10 but admit that a careful study would likely produce somewhat higher values, while Griffiths & Steyvers 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. PNAS, in its information for authors (updated as recently as June 2002), 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 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 [19]. 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, in order to resolve this issue. This case study gives us a basis to discuss and assess the merit of the various strategies.

2.2 Disability Survey Data (1982–2004)

In the second example, we work with an excerpt of data from the National Long-Term Care Survey (NLTCS) 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, 1984, 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). The ADLs are basic activities of hygiene and healthcare: eating, getting in/out of bed, moving inside the house, dressing, bathing and toileting. The IADLs are basic activities necessary to reside in the community: doing heavy housework, doing light housework, doing the laundry, cooking, grocery shopping, moving outside the house, traveling, managing money, taking medicine and telephoning. The subset of data was extracted by [15] from the analytic file of the public use data file of the NLTCS. It consists of combined data from the first four survey waves (1982, 1984, 1989, 1994) with $21,574$ individuals and 16 variables (6 ADL and 10 IADL). For each activity, individuals are either disabled or healthy on that activity (in the data table, this is coded by 1 if the individual is disabled and 0 if he is healthy). We then deal with a 2^{16} contingency table. Of the $2^{16} = 65,536$ possible combinations of response patterns, only $3,152$ are observed in the NLTCS sample.

Here we complement the earlier analyses of Erosheva [15] and Erosheva and Fienberg [18]. In particular, these earlier analyses focused primarily on the feasibility of estimation and model selection under the presumption that K was small, i.e., equal or less than 5. 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.

3 Characterizing HBMMMs

There are a number of earlier instances of mixed-membership models that have appeared in the scientific literature, e.g., see the review in [18]. A general formulation due to [15], and also described in [19], 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 J distinct characteristics. We denote by $f(x_{nj}|\beta_{jk})$ the probability distribution of *j-th* response variable in the *k-th* sub-population for the *n-th* subject, where β_{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 $\theta_n = (\theta_{n[1]}, \ldots, \theta_{n[K]})'$ represent the mixed-membership of the *n-th* subject to the various sub-populations.[1] The distribution of the observed response x_{nj} given the individual membership scores θ_n, is then

$$Pr\ (x_{nj}|\theta_n) = \sum_{k=1}^{K} \theta_{n[k]} f(x_{nj}|\beta_{jk}). \tag{2}$$

Conditional on the mixed-membership scores, the response variables x_{nj} are independent of one another, and independent across subjects.

[1] We denote components of a vector v_n with $v_{n[i]}$, and the entries of a matrix m_n with $m_{n[ij]}$.

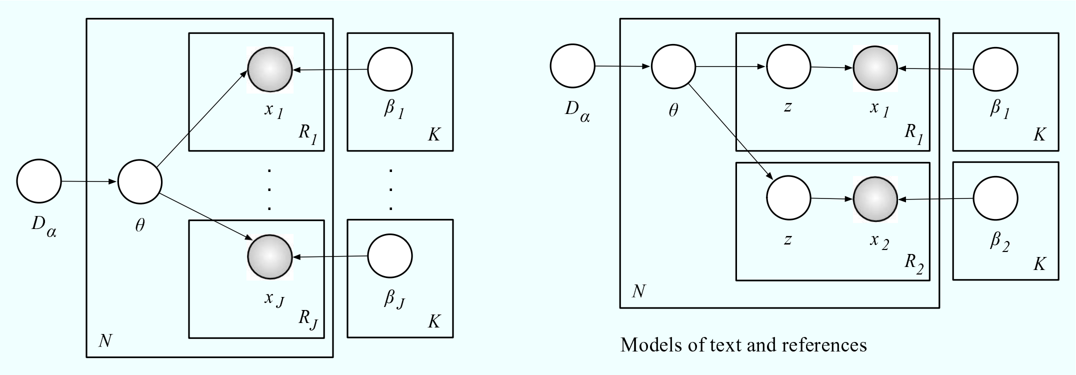

Figure 1: Left: A graphical representation of hierarchical Bayesian models of mixed-membership. Right: Models of text and references used in this paper. Specifically, we pair replicates of variables $\{x_1^r, x_2^r\}$ with latent variables $\{z_1^r, z_2^r\}$ that indicate which latent aspects informs the parameters underlying each individual replicate. The parametric and non-parametric version of the error models for the label discussed in the text refer to the specification of D_α—a Dirichlet distribution versus a Dirichlet process, respectively.

A3–Latent Variable Level. Assume that the vectors θ_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. The probability of observing x_{nj}, given the parameters, is then

$$Pr\ (x_{nj}|\alpha, \beta) = \int \left(\sum_{k=1}^{K} \theta_{n[k]} f(x_{nj}|\beta_{jk}) \right) D_\alpha(d\theta). \tag{3}$$

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

$$Pr\ (\{x_{n1}^r, \ldots, x_{nJ}^r\}_{r=1}^R|\alpha, \beta) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \theta_{n[k]} f(x_{nj}^r|\beta_{jk}) \right) D_\alpha(d\theta). \tag{4}$$

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., $R = R_{nj}$.

3.1 Example 1: Latent Dirichlet Allocation

Our general framework encompasses popular data mining models, such as the one labelled as the "latent Dirichlet allocation" model (LDA) by [29] and [11] for use in the analysis of scientific publications.

For the text component of the PNAS data: sub-populations correspond to latent "topics," indexed by k; subjects correspond to "documents," indexed by n; $J = 1$, i.e., there is only one response variable that encodes which "word" in the vocabulary is chosen to fill a position in a text of known length, so that j is omitted; positions in the text correspond to replicates, and we have a different number of them for each document, i.e. we observe R_n positions filled with words in the *n-th* document. The model assumes that each position in a document is filled with a word that expresses a specific topic, so that each document is the expression of possibly different topics. In order to do so, an explicit indicator variables z_n^r is introduced for each observed position in each document, which indicates the topic that expresses the corresponding word. The function $f(x_n^r|\beta_k)\ Pr\ (x_n^r = 1|z_n^r = k) = Multinomial\ (\beta_k, 1)$, where β_k is a random vector the size of the vocabulary, say V, and $\sum_{v=1}^{V} \beta_{k[v]} = 1$. A mixed-membership vector θ_n is associated to the *n-th* document, which encode the topic proportions that inform the choice of words in that document, and it is distributed according to D_α (i.e., a Dirichlet distribution). We obtain equation 2 integrating out the topic indicator variable z_n^r at the word level—the latent indicators z_n^r are distributed according to a $Multinomial\ (\theta_n, 1)$.

Most of our analyses also incorporate the references and we use the generalization of LDA introduced in [19] for $J = 2$, i.e., words and references which are taken to be independent.

The issue of model choice we introduced in Section 1.1 translates into the choice about the number of non-observable word and reference usage patterns (latent topics) that best describe a collection of scientific publications.

3.2 Example 2: Grade of Membership Model

The "Grade of Membership," or GoM, model is another specific model that can be cast in terms of mixed-membership. This model was first introduced by Woodbury in the 1970s in the context of medical diagnosis [46] and was developed further and elaborated upon in a series of papers and in [27]. Erosheva [15] reformulated the GoM model as a HBMMM.

In the case of the disability survey data, there are no replications, i.e., $R_n = 1$. However we consider several attributes of each american senior, i.e., $J = 16$ daily activities. Further, the scalar parameter β_{jk} is the probability of being disabled on the activity j for a complete member of latent profile k, that is,

$$\beta_{jk} = P(x_j = 1|\theta_k = 1).$$

Since we deal with binary data (individuals are either disabled or healthy), the probability distribution $f(x_j|\beta_{jk})$ is a Bernoulli distribution with parameter β_{jk}. Therefore, a complete member n of latent profile k is disabled on the activity j, i.e., $x_{nj} = 1$, with probability β_{jk}. In other words, introducing a profile indicator variable z_{nj}, we have $P(x_{nj} = 1|z_{nj} = k) = \beta_{jk}$. Each individual n is characterized by a vector of membership scores $\theta_n = (\theta_{n1}, \ldots, \theta_{nK})$. We assume that the membership scores θ_n follow the distribution D_α (for example a Dirichlet distribution with parameter $\alpha = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$. Note that the ratio $\alpha_k / \sum_k \alpha_k$ represents the proportion of the population that "belongs" to the *k-th* latent profile.

In this application, the issue of model choice translates into the choice about the number of non-observable disability propensity profiles (latent profiles) that best describe the population of American seniors.

3.3 Relationship With Other Data Mining Methods

In order to situate HBMMMs in a familiar landscape, we discuss similarities with other *unsupervised* data mining methods. In fact, in many applications including those we present in this paper, HBMMMs are used in an unsupervised fashion, with no information about class labels. Recall that in our problem we want to group observations about N subjects $\{x_n^{1:R_n}\}_{n=1}^N$ into, say, K groups. K-means clustering, for example, searches for K centroids $m_{1:K}$ that minimize

$$MSE = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} \mathbb{I}\left(x_n^{1:R_n} \in k \right) \left\| x_n^{1:R_n} - m_k \right\|^2 ,$$

where the centroids $m_{1:K}$ are centers of respective clusters in the sense of Euclidean norm. Subjects have single group membership in K-means. In the mixture of Gaussians model, a popular HBMMM that extends K-means, the MSE scoring criterion is substituted by the likelihood $\sum_{n,k} \ell(n,k)$. Further, we have unknown mixed-membership vectors θ_n, that relax the single membership of K-means. The connection is given by the fact that the mixed-membership vectors θ_n, i.e., the class abundances, have a specific form in K-means, i.e., for the *n-th* subject we can write

$$\theta_{n[k]} = \begin{cases} 1 & \text{if } k = j_n \\ 0 & \text{otherwise,} \end{cases}$$

where $j_n = \arg\min \left\{ \ell(n,k) : k \in [1,K] \right\}$. In a general specification of HBMMMs we introduce D_α distributed mixed-membership vectors, θ_n, also unknown. Further, in HBMMMs it is possible to have a more complicated likelihood structure, which follows specifications in Section 3.

4 Strategies for Model Choice

Although there are pathological examples, where slightly different model specifications lead to quite different analyses and choices of key parameters, in real situations we expect models with similar probabilistic specifications to suggest roughly similar choices for the number of groups, K.

In our applications to the study of scientific publications and disability survey data we explore the issue of model choice by means of different criteria, of which two popular choices in the data mining community: namely, cross-validation [24], and a Dirichlet process prior [7].

4.1 Choice Informed by the Ability to Predict

Cross-validation is a popular method to estimate the generalization error of a prediction rule [24], and its advantages and flaws have been addressed by many in that context, e.g., [32]. More recently, cross-validation has been adopted to inform the choice about the number groups and associated patterns in HBMMMs [8, 45].

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 with “bad estimates” of such parameters, or with arbitrary 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. To this extent, estimates with “good statistical properties,” e.g., empirical Bayes or maximum likelihood estimates, should be preferred to others [12].

4.2 The Dirichlet Process Prior

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, say, documents or individuals [20, 7]. This is an elegant model selection strategy in that the selection problem become part of the model itself, although in practical situations it is not always possible to justify. A nonparametric alternative to this strategy, recently proposed [28], uses the Dirichlet Process prior is an infinite dimensional prior with a specific parametric form as a way to mix over choices of K. This prior appears reasonable, however, for static analyses of scientific publications that appear in a specific journal. Kuma et al. [26] specify toy models of evolution which justify the scale-free nature of the relation between documents and topics using the Dirichlet process prior for exploratory data analysis purposes.

4.3 Other Criteria for Model Choice

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) [40, 34], the Akaike information criterion (AIC) [5], the deviance information criterion (DIC) [41], minimum description length (MDL) [13]. See [23] for a review of solutions in the data mining community.

AIC has a frequentist motivation and tends to pick models that are too large when then number of parameters its large—it does not pay a high enough penalty. BIC and DIC have Bayesian motivations and thus fit more naturally with the specifications in this paper. Neither is truly Bayesian for HBMMMs; however DIC involves elements that can be computed directly from MCMC calculations, and the variational approximation to the posterior (described in detail below), allows us to integrate out the nuisance parameters in order to compute an approximation to BIC for different values of K. Therefore, we explore the use of both DIC and BIC in connection with the variational approximation for the the NLTCS disability data when we can look at both criteria in action together.

5 Case Study: PNAS Scientific Collection 1997–2001

As we mentioned in Section 2, our analyses are motivated by two recent analyses of extracts of papers published in the *Proceedings of the National Academy of Sciences* (PNAS). Erosheva et al [19, 18] and Griffiths & Steyvers [22] report on wildly different numbers of latent topics, as few as 8 but perhaps a few dozen versus 300. We attempt to provide an explanation for the divergent results here. In the process we explore how to perform the model selection for hierarchical Bayesian models of mixed-membership. 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.

5.1 Modeling Text and References

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. All our models can be subsumed into the general formulation of HBMMMs presented in Section 3. Below, we organize them into finite and infinite mixture models, according to the dimensionality of the prior distribution, D_α, posited at the latent variable level—assumption A3.

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, $D_{1:N}$. The *n-th* document is represented as $D_n = (x_{1n}^{1:R_{1n}}, x_{2n}^{1:R_{2n}})$ where x_{1n}^r is a word in the abstract and x_{2n}^r is a reference in the bibliography, and where R_{1n} is the number of positions in the text of the abstract occupied by a word, and R_{2n} is the number of items in the bibliography occupied by a reference. As in the latent Dirichlet allocation example of Section 3.1, positions (the order of which does not matter), or spots, in the text of the abstracts are modeled as multinomial random variables with V_1 coordinates and unitary size parameter. That is, random variables are associated with spots in the text and their values encode which word in the vocabulary (containing V_1 distinct words) occupies a specific spot. The number of spots is observed, R_{1n}. We model the references in a similar fashion. Each item in the bibliography is modeled as multinomial random variables with V_2 coordinates and unitary size parameter. Values of these variables encode which reference in the set of known citations (V_2 of them) was used as a specific bibliography item. Again, the number of bibliography items is observed, R_{2n}. That is, words and references are vectors of size V_1, respectively V_2, with a single non-zero, unitary component. We denote by $x_{jn[v]}^r$ the *v-th* component of x_{jn}^r, for $j = 1, 2$.

Below, whenever the analysis refers to a single document, the document index n is omitted.

5.1.1 Finite Mixture: The Model

In the finite mixture case, we posit the following generative process for each document.

1. Sample the mixed-membership vector $\theta \sim D_\alpha$.

2. For each of the R_1 spots in the text of the abstract:

2.1. Sample the topic indicator $z_1^r|\theta \sim Multinomial\ (\theta, 1)$.[2]

2.2. Sample $x_1^r|z_1^r \sim Multinomial\ (\beta_1 z_1^r, 1)$.

3. For each of the R_2 items in the bibliography:

3.1. Sample topic indicator $z_2^r|\theta \sim Multinomial\ (\theta, 1)$.

3.2. Sample $x_2^r|z_2^r \sim Multinomial\ (\beta_2 z_2^r, 1)$.

In this model, D_α is a Dirichlet$(\alpha_1, \ldots, \alpha_K)$ distribution with $\alpha_k = \alpha$ for all k, and (β_1, β_2) are two matrices of size $(V_1 \times K)$ and $(V_2 \times K)$ respectively. The topic indicators, (z_1^r, z_1^r), are latent vectors of with K coordinates, only one of which assumes a unitary value.

The hyper-parameters of this model are the symmetric Dirichlet parameter α, and the multinomial parameters for words, $(\beta_{1[\cdot,k]})$, and references, $(\beta_{2[\cdot,k]})$, for each of the latent topics $k = 1, \ldots, K$. That is, through pairs of corresponding columns of the two β matrices we define a parametric representation of the K sub-populations (see assumption A1 in Section 3), 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, k) of β_1 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, i.e., $\beta_{1[v,k]} = Pr\ (x_{1[v]}^r = 1|z_{1[k]}^r = 1)$, with the constraint that $\sum_v \beta_{1[v,k]} = 1$ for each k. Similarly, element (v, k) of β_2 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. Note that, through the latent topic indicators, we associate each spot in the text, i.e., each word instance, with a latent topic. As a consequence, separate instances of the v-th vocabulary word in the same abstract[3] can be generated according to different topics.

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.

5.1.2 Infinite Mixture: The Model

In the infinite mixture case we posit a simpler and more traditional type of clustering model, by assuming that each article D_n 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.

The infinite mixture model is based upon a more compact representation of a document, $D_n = (x_{1n}, x_{2n})$, in terms of a vector of word counts, $x_{1n} = \sum_{r=1}^{R_1} x_{1n}^r$, of size V_1, and a vector of reference counts, $x_{2n} = \sum_{r=1}^{R_2} x_{2n}^r$, of size V_2. In fact, given that word instances and bibliography items in the same document cannot be generated by different topics, we do not need to keep around the corresponding random quantities, $(x_{1n}^{1:R_1}, x_{2n}^{1:R_2})$. Further, given that each article can only be generated by a single topic, the mixed membership vectors, $\theta_{1:N}$, reduce

[2] In this application, we refer to the sub-populations of assumption A1 in Section 3 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, as defined by pairs of corresponding columns of the two β matrices.

[3] On the contrary, any given reference that was picked from the set of known citations typically appears as a unique bibliography item. Thus there are no replicates of any given reference in the same bibliography.

to single membership vectors. This means that each θ_n has a single unitary component, while the remaining components equal zero. However, in the infinite mixture model we do not assume a fixed dimensionality, K, for the membership vectors $\theta_{1:N}$. That is, prior to the analysis, the number of sub-populations (see assumption A1 in Section 3) is unknown and possibly infinite.

It is more convenient to write the infinite mixture model as a generative process for the whole collection of documents altogether, $D_{1:N}$, rather than for a single document as in the previous section. In order to promote clarity in this setting, we change the notation slightly. Instead of working with the single membership vectors, $\theta_{1:N}$, it is more convenient to introduce a latent topic indicator vector, c, whose *n-th* component, $c_{[n]}$, encodes the topic assignment of the corresponding document, D_n. That is, $c_{[n]} = k$ if $\theta_{n[k]} = 1$ for $n = 1, \ldots, N$. Note that, because of single membership, $\theta_{n[k]} = \mathbb{I}(c_{[n]} = k)$ for all k. Further, because of the restriction of one topic per document, $z^r_{1n[k]} = z^r_{2n[k]} = 1$ at the same component k, for all word instances and bibliography items r. This collection of equalities, for a given document D_n, is summarized and simplified by writing $c_{[n]} = k$.

We can now posit the generative process for the whole collection of documents, $D_{1:N}$.

1. $c \sim D_\alpha$.
2. For each of the K distinct values of c:
   2.1. $\beta_{1[\cdot,k]} \sim \mathit{Dirichlet}\ (\eta_{1[1]}, \ldots, \eta_{1[V_1]})$.
   2.2. $\beta_{2[\cdot,k]} \sim \mathit{Dirichlet}\ (\eta_{2[1]}, \ldots, \eta_{2[V_2]})$.
3. For each of the N documents:
   3.1. $x_{1n} | \beta_1, c_{[n]} \sim \mathit{Multinomial}\ (\beta_{1[\cdot,c_{[n]}]}, R_{1n})$.
   3.2. $x_{2n} | \beta_2, c_{[n]} \sim \mathit{Multinomial}\ (\beta_{2[\cdot,c_{[n]}]}, R_{2n})$.

In this model, D_α is be the Dirichlet process prior with parameter α, introduced and discussed in [6, 30]. The distribution D_α models the prior probabilities of topic assignment for the collection of documents. In particular, for the *n-th* document, given the set of assignments for the remaining documents, $c_{[-n]}$, this prior puts on the *k-th* topic (out of K distinct topic assignment observed in $c_{[-n]}$) a mass that is proportional to the number of documents associated with it. It also puts prior mass on a new, $(K+1)$*-th* topic, which is distinct from the topic assignments $(1, \ldots, K)$ observed in $c_{[-n]}$. That is, D_α entails prior probabilities for each component of c as follows,

$$Pr\ \left(\ c_{[n]} = k \ \middle|\ c_{[-n]}\ \right) \ \begin{cases} \frac{m(-n,k)}{N-1+\alpha} & \text{if } m(-n,k) > 0 \\ \frac{\alpha}{N-1+\alpha} & \text{if } k = K(-n)+1 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where $c_{[-n]}$ denotes the latent topic indicator vector without the *n-th* component; $m(-n,k)$ is the number of documents that are associated with the *k-th* topic, other than the *n-th* document, i.e., $m(-n,k) = \sum_{m=1}^{N} \mathbb{I}(c_{[m]} = k, m \neq n)$; and $K(-n)$ is the number of observed, distinct topics that are associated with at least one document, other than the *n-th* document.

The hyper-parameters of this model are the scaling parameter of the Dirichlet process prior, α, and the two vectors of Dirichlet parameters, (η_1, η_2), that control the latent topics of words and references, respectively. Note that the topics, i.e., latent pairs of distributions, are not hyper-parameters of the model in our specification of the infinite mixture model. Rather, we smooth the topics by positing a pair of Dirichlet priors on them, and the corresponding parameter vectors (η_1, η_2) become the hyper-parameters at the top of the model hierarchy. In our implementation we assume symmetric Dirichlet priors for the topics, such that $\eta_{1[k]} = \eta_1$ scalar, and $\eta_{2[k]} = \eta_2$ scalar, for all components $k = 1, \ldots, K$.

In this model, we assume that the number of latent topics, K, is unknown and possibly infinite, through the prior for c, D_α. In order to find the number of topics that best describes the collection of scientific articles, we study the posterior distribution of c.

5.2 Inference

In this section we develop posterior inference for both the finite and the infinite mixture models above. In particular, we use variational methods for the finite mixture model and Monte Carlo Markov chain (MCMC) methods for the infinite mixture model.

5.2.1 Finite Mixture: Inference

In the finite mixture case, we assume the number of topics $(K < \infty)$ is fixed during inference. Unfortunately, the likelihood of a document according to this model,

$$\begin{aligned} & p\left(\, x_1^{1:R_1}, x_2^{1:R_2} \mid \alpha, \beta_1, \beta_2 \, \right) \\ & = \int \left(\prod_{r=1}^{R_1} \sum_{k=1}^{K} \prod_{v=1}^{V_1} (\theta_k \beta_{1[v,k]})^{x^r_{1[v]}} \right) \left(\prod_{r=1}^{R_2} \sum_{k=1}^{K} \prod_{v=1}^{V_2} (\theta_k \beta_{2[v,k]})^{x^r_{2[v]}} \right) D_\alpha(d\theta), \end{aligned} \tag{6}$$

does not have a closed form solution. We need the likelihood to compute the joint posterior distribution of the mixed-membership scores and the topic and reference latent indicator vectors,

$$p\left(\, \theta, z_1^{1:R_1}, z_2^{1:R_2} \mid x_1^{1:R_1}, x_2^{1:R_2}, \alpha, \beta_1, \beta_2 \, \right) = \frac{p\left(\, \theta, z_1^{1:R_1}, z_2^{1:R_2}, x_1^{1:R_1}, x_2^{1:R_2} \mid \alpha, \beta_1, \beta_2 \, \right)}{p\left(\, x_1^{1:R_1}, x_2^{1:R_2} \mid \alpha, \beta_1, \beta_2 \, \right)}, \tag{7}$$

at the denominator of the right hand side of Equation 7. The variational method prescribes the use of a mean-field approximation to the posterior distribution in Equation 7, described below. Such an approximation leads to a lower bound for the likelihood of a document, which depends upon an set of free parameters $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$. These free parameters are introduced in the mean-field approximation, and are set to minimize the Kullback-Leibler (KL henceforth) divergence between true and approximate posteriors.

The "variational EM" algorithm we develop for performing posterior inference, see Figure 2, is then an approximate EM algorithm. During the M step, we maximize the lower bound for the likelihood over the hyper-parameters of the model, $(\alpha, \beta_1, \beta_2)$, to obtain to (pseudo) maximum likelihood estimates. During the E step, we tighten the lower bound for the likelihood by minimizing the KL divergence between the true and the approximate posteriors over the free parameters, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$, given the most recent estimates for the hyper-parameters.

In the M step, we update the hyper-parameters of the model, $(\alpha, \beta_1, \beta_2)$ by maximizing the tight lower bound for the likelihood over such hyper-parameters. Given the most recent updates

```
**Variational EM** ( $\{x_{1n}^{1:R_1}, x_{2n}^{1:R_2}\}_{n=1}^{N}$ )
 1.  initialize $\alpha_{[k]} := 1/K$ for all $k$
 2.  initialize $\beta_{1[kv]} := 1/V_1$ for all $v$ and $k$
 3.  initialize $\beta_{2[kv]} := 1/V_2$ for all $v$ and $k$
 4.  **do**
 5.     **for** $n = 1$ to $N$
 6.        $(\gamma_n, \phi_{1n}^{1:R_1}, \phi_{2n}^{1:R_2}) \longleftarrow$ **Mean-Field Lower-Bound** $(x_{1n}^{1:R_1}, x_{2n}^{1:R_2})$
 7.     $\beta_{1[vk]} \propto \sum_{n=1}^{N} \sum_{r=1}^{R_1} \phi_{1n[k]}^{r} x_{1n[v]}^{r}$ for all $v$ and $k$
 8.     $\beta_{2[vk]} \propto \sum_{n=1}^{N} \sum_{r=1}^{R_2} \phi_{2n[k]}^{r} x_{2n[v]}^{r}$ for all $v$ and $k$
 9.     normalize the columns of $\beta_1$ and $\beta_2$ to sum to 1
10.     find pseudo MLE for $\alpha$ using Newton-Raphson—see main text
11.  **until** convergence
12.  **return** $(\alpha, \beta_1, \beta_2)$
```

Figure 2: The variational EM algorithm to solve the Bayes problem in finite mixture model of text and references, described in Section 5.1.1. Note, the M step updates (steps 7. and 8.) are performed incrementally in our implementation, within step 6. of the algorithm outlined above, thus speeding up the overall computation.

of the free parameters the bound depends on, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$. This leads to the following (pseudo) maximum likelihood estimates for the parameters:

$$\beta_{1[vk]} \propto \sum_{n=1}^{N} \sum_{r=1}^{R_1} \phi_{1n[k]}^{r} x_{1n[v]}^{r},$$

$$\beta_{2[vk]} \propto \sum_{n=1}^{N} \sum_{r=1}^{R_2} \phi_{2n[k]}^{r} x_{2n[v]}^{r},$$

where n is the document index, introduced above. The document index is necessary as we make use of the counts about specific words and references observed in all documents in order to estimate the corresponding conditional probabilities of occurrence, i.e., the latent topics. Unfortunately a closed form solution for the (pseudo) maximum likelihood estimates of α does not exist. We can produce a method that is linear in time by using Newton-Raphson, with the following gradient and Hessian for the log-likelihood

$$\frac{\partial L}{\partial \alpha_{[k]}} = N \left(\Psi \left(\sum_{k=1}^{K} \alpha_{[k]} \right) - \Psi(\alpha_{[k]}) \right) + \sum_{n=1}^{N} \left(\Psi(\gamma_{n[k]}) - \Psi \left(\sum_{k=1}^{K} \gamma_{n[k]} \right) \right), \tag{8}$$

$$\frac{\partial L}{\partial \alpha_{[k_1]} \alpha_{[k_2]}} = N \left(\delta_{k_1 = k_2} \cdot \Psi'(\alpha_{[k_1]}) - \Psi' \left(\sum_{k_2=1}^{K} \alpha_{[k_2]} \right) \right). \tag{9}$$

The variational EM algorithm we is summarized in Figure 2.

In the approximate E step we update the free parameters for the mean-field approximation of the posterior distribution in Equation 7, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$, given the most recent estimates of

```
**Mean-Field Lower-Bound** ( $x_1^{1:R_1}, x_2^{1:R_2}$ )
initialize $\phi^r_{1[k]} := 1/K$ for all $r$ and $k$
initialize $\phi^r_{2[k]} := 1/K$ for all $r$ and $k$
initialize $\gamma_{[k]} := \alpha_{[k]} + R_1/K + R_2/K$ for all $k$
do
  for $r = 1$ to $R_1$
    for $k = 1$ to $K$
      $\phi^r_{1[k]} \propto \beta_{1[vk]} \times \exp\left( \Psi(\gamma_{[k]}) - \Psi\left( \sum_{k=1}^K \gamma_{[k]} \right)\right)$
    normalize $\phi^r_1$ to sum to 1
  for $r = 1$ to $R_2$
    for $k = 1$ to $K$
      $\phi^r_{2[k]} \propto \beta_{2[vk]} \times \exp\left( \Psi(\gamma_{[k]}) - \Psi\left( \sum_{k=1}^K \gamma_{[k]} \right)\right)$
    normalize $\phi^r_2$ to sum to 1
  $\gamma = \alpha + \sum_{r=1}^{R_1} \phi^r_1 + \sum_{r=1}^{R_2} \phi^r_2$
until convergence
return $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$
```

Figure 3: The mean-field approximation to the likelihood for the finite mixture model of text and references, described in Section 5.1.1.

the hyper-parameters of the model, $(\alpha, \beta_1, \beta_2)$, as follows

$$\phi^r_{1[k]} \propto \prod_{v=1}^{V_1} \left[\beta_{1[vk]} \times \exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K} \gamma_{[k]} \right) \right) \right]^{x^r_{1[v]}}, \tag{10}$$

$$\phi^r_{2[k]} \propto \prod_{v=1}^{V_2} \left[\beta_{2[vk]} \times \exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K} \gamma_{[k]} \right) \right) \right]^{x^r_{2[v]}}, \tag{11}$$

$$\gamma_{[k]} = \alpha_k + \sum_{r=1}^{R_1} \phi^r_{1[k]} + \sum_{r=1}^{R_2} \phi^r_{2[k]}. \tag{12}$$

This minimizes the posterior KL divergence between true and approximate posteriors, at the document level, and leads to a new lower bound for the likelihood of the collection of documents. Note that the products over words and references in Equations 10 and 11 serve the purpose of selecting the correct probabilities of occurrence in the respective vocabularies, which correspond to the word and reference observed at a specific position, (r_1, r_2), in the document. That is, the updates of the free parameters $(\phi^{r_1}_{1[k]}, \phi^{r_2}_{2[k]})$ only depend on the probabilities $(\beta_{1[v_1 k]}, \beta_{1[v_2 k]})$, where $v_1 := \{v \in [1, V_1] \text{ s.t. } x^{r_1}_{1[v]} = 1\}$ and $v_2 := \{v \in [1, V_2] \text{ s.t. } x^{r_2}_{2[v]} = 1\}$. Using this notation, the updates simplify to

$$\begin{aligned}
\phi^{r_1}_{1[k]} &\propto \beta_{1[v_1 k]} \times \exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K} \gamma_{[k]} \right)\right), \\
\phi^{r_2}_{2[k]} &\propto \beta_{2[v_2 k]} \times \exp\left(\Psi(\gamma_{[k]}) - \Psi\left(\sum_{k=1}^{K} \gamma_{[k]} \right)\right).
\end{aligned}$$

The mean-field approximation to the likelihood we described above is summarized in Figure 3.

In order to develop the mean-field approximation for the posterior distribution in Equation 7 we used in the E step above, we posit N independent fully-factorized joint distributions over the latent variables, one for each document,

$$q\left(\theta, z_1^{1:R_1}, z_2^{1:R_2} \mid \gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2} \right) = q\left(\theta | \gamma \right) \left(\prod_{r_1=1}^{R_1} q\left(z_1^{(r_1)} \mid \phi_1^{(r_1)} \right) \prod_{r_2=1}^{R_2} q\left(z_2^{(r_2)} \mid \phi_2^{(r_2)} \right) \right),$$

which depends on the set of previously mentioned free parameters, $(\gamma, \phi_1^{1:R_1}, \phi_2^{1:R_2})$. The mean-field approximation consists in finding an approximate posterior distribution,

$$\tilde{p}\left(\theta, z_1^{1:R_1}, z_2^{1:R_2} \mid \tilde{\gamma}, \tilde{z}_1^{1:R_1}, \tilde{z}_2^{1:R_2}, \alpha, \beta_1, \beta_2 \right),$$

where the conditioning on the data is now obtained indirectly, trough the free parameters,

$$\begin{aligned}
\tilde{\gamma} &= \tilde{\gamma}\left(x_1^{1:R_1}, x_2^{1:R_2} \right), \\
\tilde{z}_1^{1:R_1} &= \tilde{z}_1^{1:R_1}\left(x_1^{1:R_1}, x_2^{1:R_2} \right), \\
\tilde{z}_2^{1:R_2} &= \tilde{z}_2^{1:R_1}\left(x_1^{1:R_1}, x_2^{1:R_2} \right)
\end{aligned}$$

The factorized distribution leads to a lower bound for the likelihood; in fact it is possible to find a closed form solution to the integral in Equation 6 by integrating the latent variables out with respect to the factorized distribution. An approximate posterior, $\tilde{p}$, is computed by substituting the lower bound for the likelihood at the denominator of Equation 7. The mean-field approximation in then obtained by minimizing the Kullback-Leibler divergence between the true and the approximate posteriors, over the free parameters.

The mean-field approximation has been used in many applications over the years [38, 39, 33, 9]. Intuitively, the approximation aims at reducing a complex problem into a simpler one by "decoupling the degrees of freedom in the original problem." Such decoupling is typically obtained via an expansion that involves additional, free parameters that are problem dependent, e.g., $\{\gamma_n, \phi_{1n}^{1:R_1}, \phi_{2n}^{1:R_2}\}_{n=1}^N$ in our model above. A thorough treatment of such methods, which focus on applications to statistics and machine learning, is given in [25, 47, 44]. We have adapted these methods for other applications in work we hope to report on in the near future.

5.2.2 Infinite Mixture: Inference

In the infinite mixture case, we assume the total number of topics, K, to be unknown and possibly infinite. The posterior distribution of c, which is the goal of the posterior inference in this model, cannot be derived in closed form. However, the component-specific full conditional distributions, i.e., $Pr\ (c_{[n]}|c_{[-n]})$ for $n = 1, \ldots, N$, are known up to a normalizing constant. Therefore we can explore the desired posterior distribution of the vector c through MCMC sampling methods.

Following Algorithm 3 in [30], we derive the full conditional distribution of the topic assignment vector. The full conditional probability that document D_n belongs in an existing topic k, given all documents, D, and the topic assignment of all other documents, $c_{[-n]}$, is given by

$$\begin{aligned}
&Pr\ (c_{[n]} = k|D, c_{[-n]}) \\
&\qquad \propto \frac{m(-n,k)}{N-1+\alpha} \\
&\qquad\quad \times \binom{R_{1n}}{x_{1n}} \frac{\Gamma(\eta_1 + \sum_v \sum_{i\neq n:c_i=k} x_{1i[v]})}{\prod_v \Gamma(\eta_1/V_1 + \sum_{i\neq n:c_i=k} x_{1i[v]})} \frac{\prod_v \Gamma(x_{1n[v]} + \eta_1/V_1 + \sum_{i\neq n:c_i=k} x_{1i[v]})}{\Gamma(\sum_v x_{1n[v]} + \eta_1 + \sum_v \sum_{i\neq n:c_i=k} x_{1i[v]})} \\
&\qquad\quad \times \binom{R_{2n}}{x_{2n}} \frac{\Gamma(\eta_2 + \sum_v \sum_{i\neq n:c_i=k} x_{2i[v]})}{\prod_v \Gamma(\eta_2/V_2 + \sum_{i\neq n:c_i=k} x_{2i[v]})} \frac{\prod_v \Gamma(x_{2n[v]} + \eta_2/V_2 + \sum_{i\neq n:c_i=k} x_{2i[v]})}{\Gamma(\sum_v x_{2n[v]} + \eta_2 + \sum_v \sum_{i\neq n:c_i=k} x_{2i[v]})}, \quad (13)
\end{aligned}$$

where $c_{[-n]}$ is the topic assignment vector for all documents other than D_n. The full conditional probability that document D_n belongs to a topic which no other D_j belongs to is the following:

$$\begin{aligned}
Pr(c_{[n]} \neq c_{[i]}\ \forall\ i \neq n|D, c_{[-n]}) \quad \propto \quad & \frac{\alpha}{N-1+\alpha} \\
& \times \binom{R_{1n}}{x_{1n}} \frac{\Gamma(\eta_1) \prod_v \Gamma(x_{1n[v]} + \eta_1/V_1)}{\Gamma(\eta_1/V_1)^{V_1} \Gamma(\sum_v x_{1n[v]} + \eta_1)} \\
& \times \binom{R_{2n}}{x_{2n}} \frac{\Gamma(\eta_2) \prod_v \Gamma(x_{2n[v]} + \eta_2/V_2)}{\Gamma(\eta_2/V_2)^{V_2} \Gamma(\sum_v x_{2n[v]} + \eta_2)}. \qquad (14)
\end{aligned}$$

The sparseness of D and symmetry of the Dirichlet prior leads to a forms of Equations 13 and 14 that are more quickly computed.

The parameters of the model estimated in this way are the vector c of topic assignments and the total number of topics, K. The posterior distributions of c and K can be found using a Gibbs sampler with these full conditional distribution as shown in Figure 4.

In order to asses convergence of the Markov chain, we examine the total number of topics (which varies by Gibbs sample) and consider the Markov chain converged when the number of topics has converged. Convergence was diagnosed when several independent chains sampled close values of K. We started chains with 10, 25, 40, and 11988 topics and they converged after approximately 30 iterations. Thus we are reasonably confident of convergence despite the small number of iterations because of the diversity of chain starting values.

In the estimation of the posterior distribution of c and K, there are two hyperparameters which must be chosen. The prior distribution on c depends on the value of α; larger values of α greater than one encourage more groups while values of α smaller than one discourages new groups. We interpret α as the number of documents that we *a priori* believe belong in a new topic started by one document. However, once an document has started a new group, other

```
**MCMC** $( \{x_{1n}, x_{2n}\}_{n=1}^{N} )$
initialize $K$ between 1 and $N$
**for** $k = 1$ to $(K-1)$
    initialize $c_{[n]} := k$ for $n = (k-1) \times \lfloor N/K \rfloor + 1$ to $k \times \lfloor N/K \rfloor$
initialize $c_{[n]} := K$ for $n = (K-1) \times \lfloor N/K \rfloor + 1$ to $N$
**do**
    **for** $n = 1$ to $N$
        sample $c_{[n]}$ from a Multinomial with probabilities from Eq. 13 and Eq. 14
        update $K := \max_n(c_{[n]})$
**until** 50 iterations after convergence (see discussion)
**return** posterior distribution of $(c, K)$
```

Figure 4: The MCMC algorithm to find the posterior distribution of classification in the infinite mixture model of text and references, described in Section 5.1.2.

documents will be less likely to join that group based on its small size. Therefore, $\alpha = 1$ is used here as the standard value.

The posterior distribution of c also depends, through β, on the η parameters. This is the Dirichlet prior on the probability vector over words or references for each topic. A value of η smaller than V, the vocabulary size, implies a prior belief that the word distributions will be highly skewed (a few likely words and many words with almost no probability of use). These values of η cause all documents to appear in one large group, $K = 1$. A value of η larger than V implies a prior belief that all words are equally likely to be used in a given topic. Here, we take $\eta_1 = 1000 \times V_1$ and $\eta_2 = 1000 \times V_2$ as values that encourages a range of values of K.

5.3 Empirical Results

In Figure 5, we give the log-likelihood obtained for the four finite mixture models (at $K = 5, 10, \cdots, 50, 100, 200, 300)$).

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. The plots of the log likelihood in Figure 5 suggest we choose a number of topics between 20 and 40 whether words or words and references are used. The infinite model generates a posterior distribution for the number of topics, K, given the data. Figure 6 shows the posterior distribution ranges from 23 to 33 profiles. We expect that the infinite model will require more topics than the finite mixed-membership because it is a hard clustering.

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. In Table 1, we list 12 high-probability words from these topics after filtering out the stop words. Table 2 shows the 5 references with the highest probability for 6 of the topics.

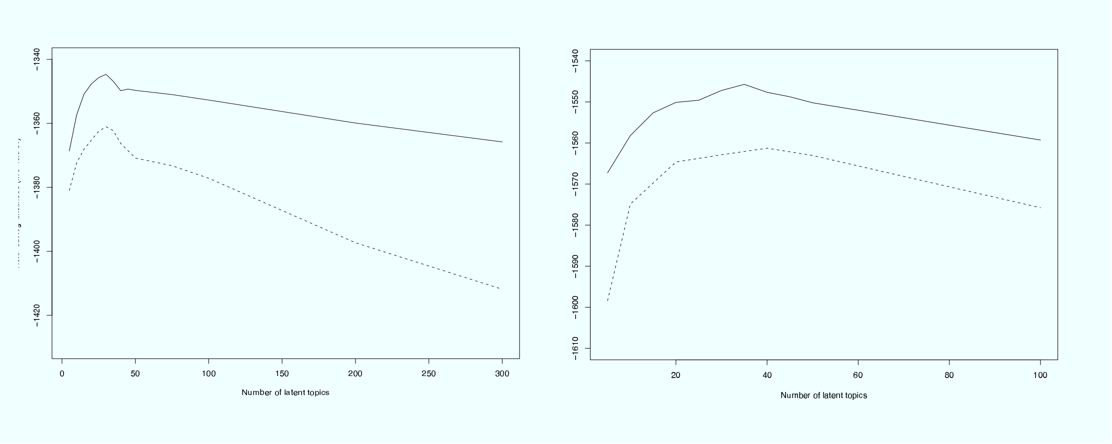

Figure 5: Left Panel: Log-likelihood (5 fold cv) for $K = 5, \ldots, 50, 75, 100, 200, 300$ topics. We plot: text only, α fitted (solid line); text only, α fixed (dashed line). Right Panel: Log-likelihood (5 fold cv) for $K = 5, \ldots, 50, 100$ topics. We plot: text and references, α fitted (solid line); text and references, α fixed (dotted line).

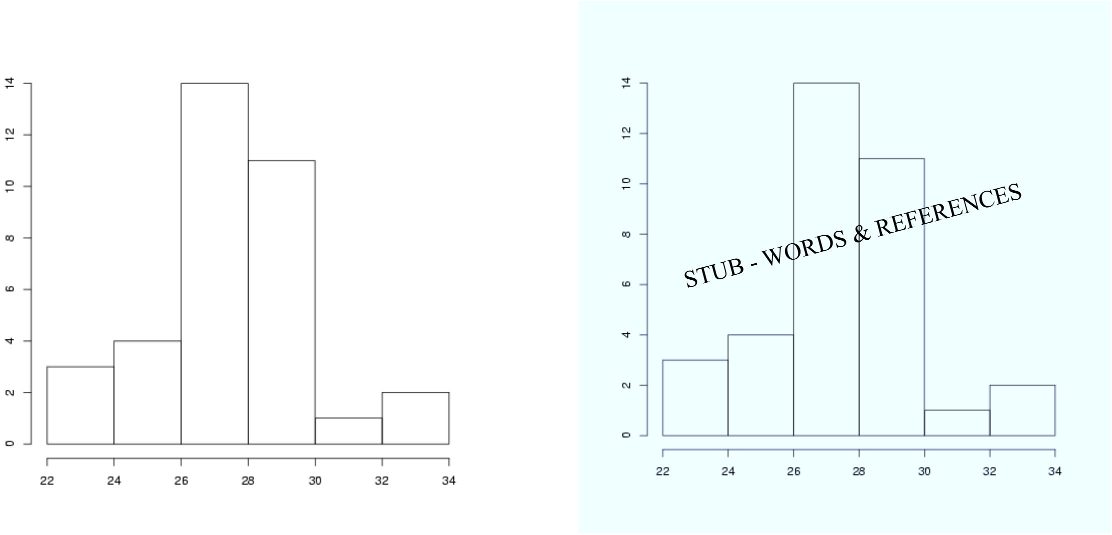

Figure 6: Posterior distribution of K for the PNAS scientific collection corresponding to the infinite mixture models of text (left panel) and of text and references (right panel).

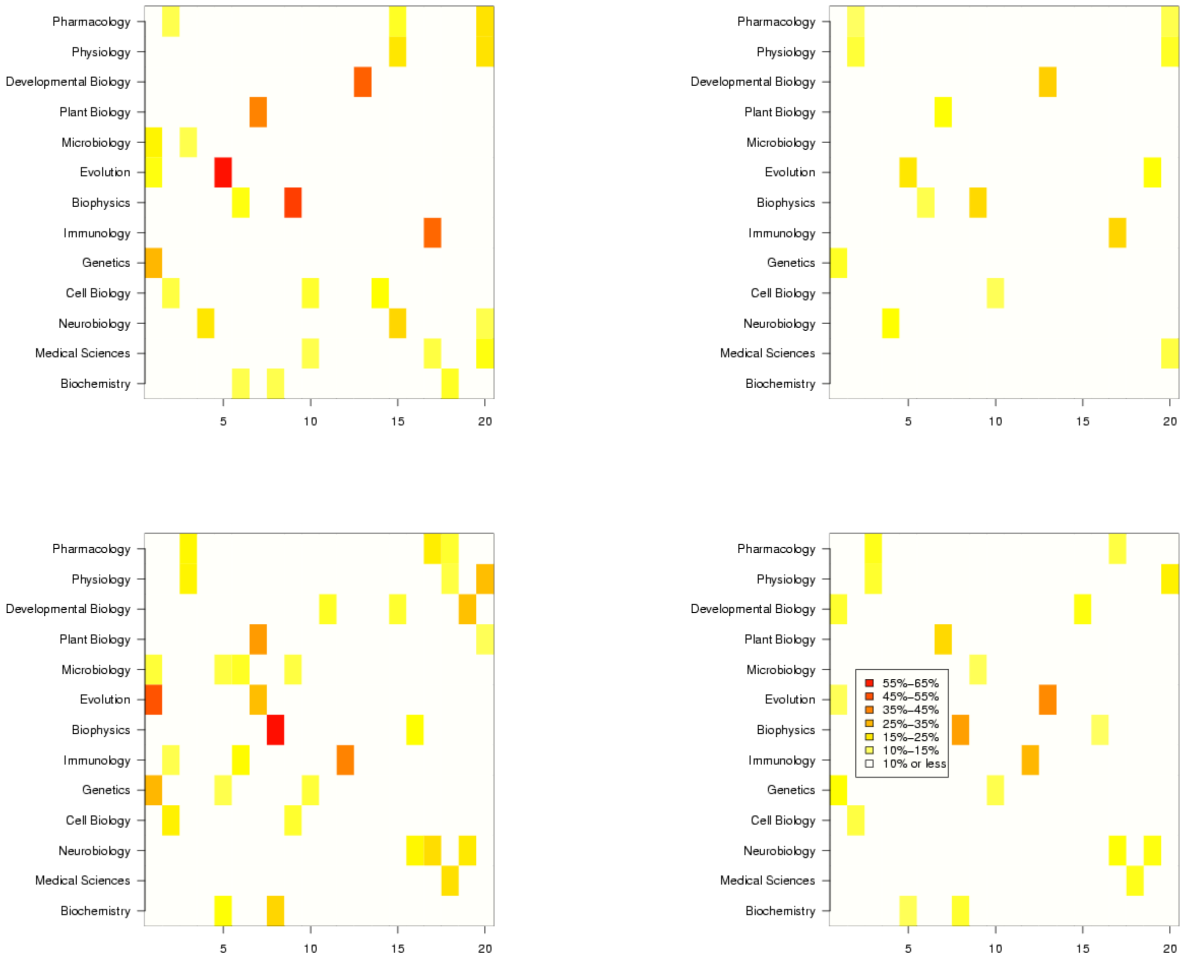

Figure 7: The average membership in the 20 latent topics (columns) for articles in thirteen of the PNAS editorial categories (rows). Darker shading indicates higher membership of articles submitted to a specific PNAS editorial category in the given latent topic and white space indicates average membership of less than 10%. Note that the rows sum to 100% and therefore darker topics show concentration of membership and imply sparser membership in the remaining topics. These 20 latent topics were created using the four finite mixture models with words only (1^{st}, 2^{nd}) or words and references (3^{rd}, 4^{th}) and α estimated (1^{st}, 3^{rd}) or fixed (2^{nd}, 4^{th}).

Table 1: Word usage patterns corresponding to the model of text & references, with $K = 20$ topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
gene	kinase	cells	cortex	species
genes	activation	virus	brain	evolution
sequence	receptor	gene	visual	population
chromosome	protein	expression	neurons	populations
analysis	signaling	human	memory	genetic
genome	alpha	viral	activity	selection
sequences	phosphorylation	infection	cortical	data
expression	beta	cell	learning	different
human	activated	infected	functional	evolutionary
dna	tyrosine	vector	retinal	number
number	activity	protein	response	variation
identified	signal	vectors	results	phylogenetic
Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
enzyeme	plants	protein	protein	cells
reaction	plant	rna	model	cell
ph	acid	proteins	folding	tumor
activity	gene	yeast	state	apoptosis
site	expression	mrna	energy	cancer
transfer	arabidopsis	activity	time	p53
mu	activity	trna	structure	growth
state	levels	translation	single	human
rate	cox	vitro	molecules	tumors
active	mutant	splicing	fluorescence	death
oxygen	light	complex	force	induced
electron	biosynthesis	gene	cdata	expression
Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
transcription	dna	cells	protein	ca2+
gene	rna	cell	membrane	channel
expression	repair	expression	proteins	channels
promoter	strand	development	atp	receptor
binding	base	expressed	complex	alpha
beta	polymerase	gene	binding	cells
transcriptional	recombination	differentiation	cell	neurons
factor	replication	growth	actin	receptors
protein	single	embryonic	beta	synaptic
dna	site	genes	transport	calcium
genes	stranded	drosophila	cells	release
activation	cdata	embryos	nuclear	cell
Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
peptide	cells	domain	mice	beta
binding	cell	protein	type	levels
peptides	il	binding	wild	increased
protein	hiv	terminal	mutant	insulin
amino	antigen	structure	gene	receptor
site	immune	proteins	deficient	expression
acid	specific	domains	alpha	induced
proteins	gamma	residues	normal	mice
affinity	cd4	amino	mutation	rats
specific	class	beta	mutations	treatment
activity	mice	sequence	mouse	brain
active	response	region	transgenic	effects

Table 2: References usage patterns for 6 of the 20 topics corresponding to the model of text & references, with $K = 20$ topics.

Author	**Journal**
Topic 2	
THOMPSON,CB	SCIENCE, 1995
XIA,ZG	SCIENCE, 1995
DARNELL,JE	SCIENCE, 1994
ZOU,H	CELL, 1997
MUZIO,M	CELL, 1996
Topic 5	
SAMBROOK,J	MOL. CLONING. LAB. MANU., 1989
ALTSCHUL,SF	J. MOL. BIOL., 1990
EISEN,MB	P. NATL. ACAD. SCI. USA, 1998
ALTSCHUL,SF	NUCLEIC. ACIDS. RES, 1997
THOMPSON,JD	NUCLEIC. ACIDS. RES, 1994
Topic 7	
SAMBROOK,J	MOL. CLONING. LAB. MANU,1989
THOMPSON,JD	NUCLEIC. ACIDS. RES,1994
ALTSCHUL,SF	J. MOL. BIOL,1990
SAITOU,N	MOL. BIOL. EVOL,1987
ALTSCHUL,SF	NUCLEIC. ACIDS. RES,1997
Topic 8	
SAMBROOK,J	MOL. CLONING. LAB. MANU,1989
KIM,NW	SCIENCE, 1994
BODNAR,AG	SCIENCE, 1998
BRADFORD,MM	ANAL. BIOCHEM., 1976
FISCHER,U	CELL, 1995
Topic 17	
SHERRINGTON,R	NATURE,1995
HO,DD	NATURE,1995
SCHEUNER,D	NAT. MED.,1996
THINAKARAN,G	NEURON,1996
WEI,X	NATURE,1995
Topic 20	
CHOMCZYNSKI,P	ANAL. BIOCHEM., 1987
BRADFORD,MM	ANAL. BIOCHEM., 1976
KUIPER,GGJM	P. NATL. ACAD. SCI. USA, 1996
MONCADA,S	PHARMACOLREV, 1991
KUIPER,GG	ENDOCRINOLOGY, 1998

Using both tables, we offer the following interpretations of topics:

- Topics 1 and 12 focus on nuclear activity (genetic) and (repair/replication).
- Topic 2 concerns protein regulation and signal transduction.
- Two topics are associated with the study of HIV and immune responses: topic 3 is related to virus treatment and topic 17 concerns HIV progression.
- Two topics relate to the study of the brain and neurons: topic 4 (behavioral) and topic 15 (electrical excitability of neuronal membranes).
- Topic 5 is about population genetics and phylogenetics.
- Topic 7 is related to plant biology.
- Two topics deal with human medicine: topic 10 with cancer and topic 20 with diabetes and heart disease.
- Topic 13 relates to developmental biology.
- Topic 14 concerns cell biology.
- Topic 19 focus on experiments on transgenic or inbred mutant mice.
- Several topics are related to protein studies, e.g., topic 9 (protein structure and folding), topic 11 (protein regulation by transcription binding factors), and topic 18 (protein conservation comparisons).
- Topics 6, 8, and 16 relate to biochemistry.

These labels for the topics are primarily convenience, but they do highlight some of the overlap between the PNAS sections (Plant Biology and Developmental Biology) and the latent topics (7 and 13). However, many plant biologists may do molecular biology in their current work. We can also see by examining the topics that small sections such as Anthropology do not emerge as topics and broad sections such as Medical Science and Biochemistry have distinct subtopics within them. This also suggests special treatment for general sections such as Applied Biology and cutting-edge interdisciplinary papers when evaluating the classification effectiveness of a model.

To summarize the distribution of latent aspects over distributions, we provide graphical representations of the distribution of latent topics for each of the PNAS topics in Figure 7. The third figure represents the model used for Tables 1 and 2. The two figures on the right represent models where the α parameter of the Dirichlet prior over topics is fixed. These two models are less sparse than the corresponding models with α fit to the data. For twenty latent topics, we fix $\alpha = 50/20 = 2.5 > 1$ and this means 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 are less than one lead to models that expect articles to have high membership in a small number of topics. See Section 5.4 for further consequences of these assumptions. The PNAS

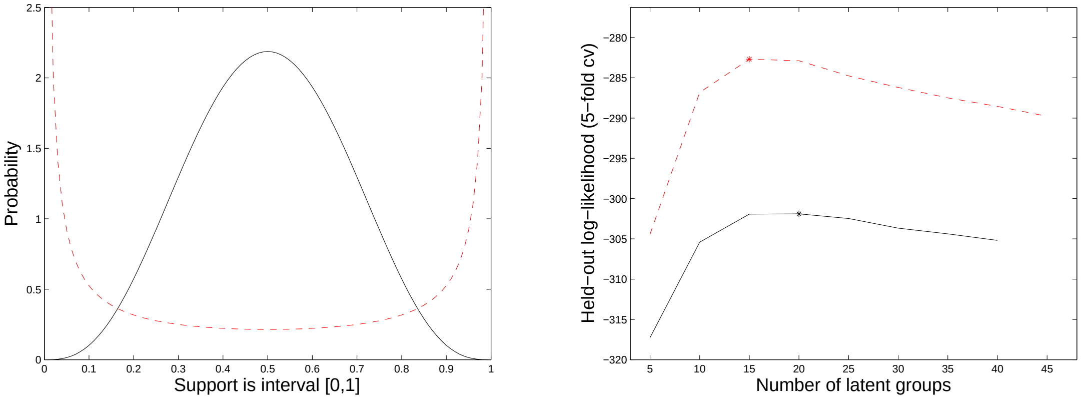

Figure 8: Left: 2D symmetric Dirichlet densities underlying mixed-membership vectors $\theta = (\theta_1, \theta_2)$, with parameter $\alpha = 4 > 1$ (solid, black line) and with parameter $\alpha = 0.25 < 1$ (dashed, red line). Right: held-out log-likelihood for the simulation experiments described in the text. The solid, black line corresponds to the strategy of fixing $\alpha = 50/K$, whereas the dashed, red line corresponds to the strategy of fitting α via empirical Bayes. K^* is denoted with an asterisk.

topics tend to have a few latent topics highly represented when α is fit and low to moderate representation in all topics when α is fixed (as seen by white/light colored rows).

Further examining Figure 7, note that topic 1, identified with genetic activity in the nucleus, was highly represented in articles from Genetics, Evolution, and Microbiology. Also 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 topics and the discovered latent topics. The assigned topics used in PNAS 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 interest of Biological Sciences. Figure 7 also shows that there is a lot of hope for collaboration and interest between separate fields which are researching the same ideas.

As we saw in Figure 5, the held-out log likelihood plot corresponding to five-fold cross validation suggest a number between 20 and 40 topics for the finite mixture model. Othe analyses with finite finite mixture with words and references supports support values towards the lower end of this range, i.e., $K = 20$, more than other choices. This is also true in the posterior distribution of K for the infinite mixture model. We fixed $\alpha = 50/K$ following the choice in [22] and estimated α from the data,. This produced a similar conclusion. While [22] found posterior evidence for nearly 300 topics, a number on the order of 20 or 30 provides a far better fit to the data, assessed robustly by multiple criteria and specifications. Moreover, we find this simpler more interpretable in a meaningful way that is not possible with 300 topics.

5.4 Evidence from a Simulation Study: A Practice to Avoid

To conclude, with the aim of highlighting the dangers of fixing the hyper-parameters according to some ad-hoc strategy that is *not* supported by the data, e.g., fixing $\alpha = 50/K$ in the models of the previous section, we report some anecdotal evidence we gathered from synthetic data. We simulated a set of 3,000 documents according to the finite mixture model of text only described in Section 5.1, with $K^* = 15$ and a vocabulary of size 50. We then fitted the correct finite mixture model on a grid for $K = 5, 10, 45$ that included the true underlying number of groups and associated patterns, using a five-fold cross-validation scheme. In a first batch of experiments we fitted alpha using empirical Bayes [12], whereas in a second batch of experiments we set $\alpha = 50/K$, following the analysis in [22]. The held-out log-likelihood profiles are reported in Figure 8.

In this controlled experiment, the optimal number of non-observable groups is $K^* = 15$. This implies a value of $\alpha = \frac{50}{15} = 3.33 > 1$ for the ad-hoc strategy, whereas $\hat{\alpha} = 0.052 < 1$ according to the empirical Bayes strategy. Intuitively, the fact that $\alpha > 1$ has a disrupting effect on the model fit: each topic is expected to be present in each document, or in other words each document is expected to belong equally to each group/topic, rather than only to only a few of them, as it is the case when $\alpha < 1$. As an immediate consequence, the estimates of the components of mixed-membership vectors, $\{\theta_{nk}\}$, tend to be diffuse, rather than sharply peaked, as we would expect in text mining applications. We can observe this effect, for example, in Figure 7, where the plots in the right column display latent topics that are more "diffuse" than those estimated by fitting the hyper-parameter α with maximum likelihood as well. Further, in our simulation, setting the hyper-paramter α to a value greater than one when the data supports values in a dramatically different range, e.g., $0.01 < \alpha < 0.1$, ultimately bias the estimation of the number of latent groups. This effect can be observed by looking at the entries in Figure 7, where diffuse estimates are found corresponding to the strategy of fixing α. Further, Figure 8 shows that the empirical Bayes strategy correctly recovers $K^* = 15$, whereas the ad-hoc strategy finds $K^* = 20$.

Our experiments in a controlled setting suggest that it is desirable not to fix the hyper-parameters, e.g., the non-observable category abundances α, according to ad-hoc strategies, unless such strategies are supported by previous analyses. Ad-hoc strategies will affect inference about the number of non-observable groups and associated patterns in non-controllable ways, and ultimately bias the analysis of data.

6 Case Study: Disability Profiles of American Seniors

As we mentioned in Section 2, the analysis we present here complements the analyses of the data from the *National Long Term Care Survey* (NLTCS) presented in Erosheva [15] and Erosheva and Fienberg [18]. In particular, [15] considers finite mixture models that feature up to five latent disability profiles and concludes that the model with four profiles is the most appropriate to describe the NLTCS data. In this section, we explore a larger set of finite mixture models that feature up to ten latent disability profiles, and we also present a nonparametric model that does not fix the number of profiles prior to the analysis.[4]

As in the previous case study, the focus on the analysis is on the selection of the number of latent disability profiles, i.e., on the selection of the model, which best describes the data.

[4] Rather, the nonparametric model implicitly encodes a prior on the number of latent profiles such that $K \approx ln(N)$, where N is the number of seniors in the sample. In the NLTCS data, $N = 21,574$ and $ln(N) \approx 10$.

6.1 Modeling Disability

In this section we introduce model specifications to analyze the sample of American seniors included in the NLTC panel survey. For our purposes it will be sufficient to ignore the temporal dimension of the data collection—we refer to [14] for a longitudinal analysis. All our models can be subsumed into the general formulation of HBMMMs presented in Section 3. 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 American senior 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 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 the survey), whereas it is recorded as one if the n-th individual has problems performing the j-th activity (an individual is considered disabled to that extent for the purpose the survey).

6.1.1 Finite Mixture: The Model

To carry out the analysis of the NLTCS data in the finite mixture setting we use the GoM model of Section 3.2, which posits the following generative process for the n-th individual.

1. Sample $\theta_n \sim D_\alpha$.
2. For each of the J responses

   2.1. Sample $z_{jn}|\theta_n \sim Multinomial\ (\theta_n, 1)$.

   2.2. Sample $x_{jn}|z_{jn} \sim Bernoulli\ (\beta z_{jn})$.

Here, we take D_α to be a Dirichlet distribution with hyper-parameter $\alpha = (\alpha_1, \ldots, \alpha_K)$. Note that this is not the symmetric distribution we used in the previous case study, in the finite setting. In this model, β 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, if we denote as before the latent profile indicator vector with z_{jn}, then $\beta_{[jk]} = P(x_{jn} = 1|z_{jn[k]} = 1)$ is the probability of being disabled with respect to the j-th activity for a senior 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, $\sum_{j=1}^{J} \beta_{[jk]}$ is not necessarily one as in the model of Section 5.2.1[5]. The hyper-parameters of this model are α and β. In Section 6.2.1 we develop a variational approximation to perform posterior inference on such hyper-parameters, and on the latent variables θ_n and z_{jn} for all j's and n's.

In our analysis, we also consider a fully Bayesian version of the GoM model, following [15], which posits the following generative process for all N individuals in the survey.

[5] Note another subtle difference from the generative process of Section 5.2.1. In this model we loop over 1 replicate of each of the J responses observed for the n-th American senior, whereas in the previous model we loop over R_1 word instances and R_2 bibliography items observed in the n-th document.

1. Sample $\xi \sim D_\alpha$
2. Sample $\alpha_0 \sim Gamma\ (\tau_1, \tau_2)$
3. Sample $\beta_{[jk]} \sim Beta\ (\sigma_1, \sigma_2)$ for all j and k
4. For each of the N individuals
   - 4.1. Sample $\theta_n \sim Dirichlet\ (\alpha_0 \xi_{[1]}, \ldots, \alpha_0 \xi_{[K]})$.
   - 4.2. For each of the J responses
     - 4.2.1. Sample $z_{jn} | \theta_n \sim Multinomial\ (\theta_n, 1)$
     - 4.2.2. Sample $x_{jn} | z_{jn} \sim Bernoulli\ (\beta z_{jn})$

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, θ_n. The elements of β are sampled from a symmetric Beta distribution with fixed hyper-parameter $\sigma_1 = \sigma_2 = 1$. Note that a symmetric Beta sampling scheme with unitary parameter is equivalent to a Uniform sampling scheme on $[0, 1]$.

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.

6.1.2 Infinite Mixture: The Model

In the infinite setting we do not fix the number of sub-populations K underlying the population of American seniors surveyed prior to the analysis. As in the previous case study, the mixed membership vectors $\theta_{1:N}$ reduce to single membership vectors. We denote membership with c, where $c_{[n]} = k$ indicates that $\theta_{n[k]} = 1$. We posit the following generative process.

1. Sample $c \sim D_\alpha$
2. For each of the K distinct values of c
   - 2.1. Sample $\beta_{[jk]} \sim Beta\ (\tau_1, \tau_2)$ for all j
3. For each of the N seniors
   - 3.1. For each of the J responses
     - 3.1.1. sample $x_{jn} | \beta_{[jc_{[n]}]} \sim Bernoulli\ \left(\beta_{[jc_{[n]}]} \right)$

Here D_α is the Dirichlet process prior described in Section 5.1.2. In our implementation, we specify a symmetric Beta distribution for the disability probabilities, β_{kj}, with $\tau_1 = \tau_2 = \tau$. Further, we fix the hyper-parameter of the Dirichlet process prior D_α at $\alpha = 1$, which encodes “indifference” toward additional groups.

In this model, we assume that the number of latent disability profiles, K, is unknown and possibly infinite, through the prior for c, D_α. In order to find the number of profiles that best describes the population of seniors, we study the posterior distribution of c.

6.2 Inference

In this section we develop posterior inference for both specifications of the finite mixture model, and for the infinite mixture model above. In particular, we use variational methods for the "basic" finite mixture model and Monte Carlo Markov chain (MCMC) methods for "fully Bayesian" finite mixture models and for the infinite mixture model.

6.2.1 Finite Mixture: Inference

A—The Variational approximation for the "basic" model. As in Section 5.1.1, in the finite mixture case, the coupling between the mixed-membership scores, $\theta_{1:N}$, and the conditional disability probabilities given profile, β, results in an intractable likelihood. Likewise, the algorithm that leads to the mean-field solution to the Bayes problem is an variational EM algorithm. This is an approximate EM algorithm, which involves evaluating a lower bound for the likelihood that depends on additional free parameters.

In the M step we maximize such lower bound with respect to the hyper-parameters of the model, (α, β), given the updates of the free parameters, $(\gamma_n, \phi_{1n:Jn})$ for the *n-th* individual. We then obtain the (pseudo) maximum likelihood estimates for the hyper-parameters as follows.

$$\beta_{[jk]} \quad \propto \quad \sum_{n=1}^{N} \phi_{jn[k]} x_{jn},$$

where n is the index that runs over the N individuals in the sample. The (pseudo) maximum likelihood estimates for α are derived using the Newton-Raphson algorithm, with gradient and Hessian given in Equations 8 and 9.

In the approximate E step we update the free parameters corresponding to the *n-th* individual, $(\gamma_n, \phi_{1n:Jn})$, given the update estimates for the parameters of the model, (α, β), as follows.

$$\phi_{j[k]} \quad \propto \quad \beta_{j[k]}^{x_j} (1 - \beta_{[jk]})^{1-x_j} \times \left(\Psi(\gamma_{[k]}) - \Psi(\sum_{k=1}^{K} \gamma_{[k]}) \right), \tag{15}$$

$$\gamma_{[k]} \quad = \quad \alpha_{[k]} + \sum_{j=1}^{J} \phi_{j[k]}. \tag{16}$$

As before, the approximation is introduced because the integral used to evaluate the likelihood for an individual,

$$p\ (\ x_1, \cdots, x_J \mid \alpha, \beta\ ) \int \prod_{j=1}^{J} \left(\left(\sum_{k=1}^{K} \theta_{[k]} \beta_{[jk]} \right)^{x_j} \left(1 - \sum_{k=1}^{K} \theta_{[k]} \beta_{[jk]} \right)^{1-x_j} \right) D_\alpha(d\theta), \tag{17}$$

does not have a closed form solution. As in the model for text and references, we then posit N independent fully factorized joint distributions over the latent variables, one for each individual,

$$q\left(\theta, z_{1:J} \mid \gamma, \phi_{1:J}\right) = q\left(\theta|\gamma\right) \prod_{j=1}^{J} q\left(z_j \mid \phi_j\right)$$

which depend on a set of free parameters, $(\gamma, \phi_{1:J})$. We then develop a mean-field approximation to the true posterior of the latent variables given data and hyper-parameters, which leads to the approximate EM described above.

B—The MCMC for the "fully Bayesian" model. We derived a Metropolis-within-Gibbs MCMC sampler for these model specifications, following [15]. One iteration for this algorithm consists of a Gibbs sampler for drawing z, θ and β and two Metropolis-Hasting steps for drawing α_0 and ξ. The joint distribution for the fully Bayesian version of the GoM model is:

$$p(x, z, \theta, \beta, \alpha_0, \xi) = p(\xi)p(\alpha_0)p(\beta) \prod_{n=1}^{N} \left(p(\theta_n|\alpha) \prod_{j=1}^{J} \prod_{k=1}^{K} \theta_n z_{jn} \beta_{[jz_{jn}]}^{x_{jn}} (1 - \beta_{[jz_{jn}]})^{1-x_{jn}} \right) \tag{18}$$

where $p(\beta) = \prod_{j=1}^{J} \prod_{k=1}^{K} p(\beta_{[jk]})$. The exact specifications for $p(\beta_{[jk]})$, $p(\xi)$ and $p(\alpha_0)$ are given by in Section 6.1.1. From the factorization of the joint distribution in Equation (18), we are able to derive the full conditional distributions of β, z and θ.

The Gibbs sampler algorithm can then be used to obtain the posterior distribution of the model parameters β and θ. To obtain the parameters update for the $(i+1)$-*th* step, we do the following.

- For $n = 1, \ldots, N$, for $j = 1, \ldots, J$, sample

  $$z_{jn}^{(i+1)} \sim Multinomial\ (q, 1),$$

  where $q(q_1, \ldots, q_K)$ and $q_k = \theta_{n[k]}^{(i)} (\beta_{[jk]}^{(i)})^{x_{jn}} (1 - \beta_{[jk]}^{(i)})^{1-x_{jn}}$.
- For $j = 1, \ldots, K$, for $k = 1, \ldots, K$, sample

  $$\beta_{[jk]}^{(i+1)} \sim Beta\ \left(1 + \sum_{\{n: z_{jn}^{(i+1)} = k\}} x_{jn}, 1 + \sum_{\{n: z_{jn}^{(i+1)} = k\}} (1 - x_{jn}) \right).$$

- For $n = 1, \ldots, N$, sample

  $$\theta_n^{(i+1)} \sim Dirichlet\ \left(\alpha_{[1]} + \sum_{j=1}^{J} \delta(z_{jn}^{(i+1)} = 1), \ldots, \alpha_{[K]} + \sum_{j=1}^{J} \delta(z_{jn}^{(i+1)} = K) \right).$$

We use Metropolis-Hasting steps to draw from the posterior distribution of α_0 and ξ, given that $\alpha = \alpha_0 \xi$ is random. For α_0, we consider the proposal distribution $p(\alpha_0^*|\alpha_0) = Gamma\ (\gamma, \gamma/\alpha_0)$ where γ is an adjustable tuning parameter. The Metropolis-Hasting step for α_0 is:

- Sample $\alpha_0^* \sim p(\alpha_0^*|\alpha_0^{(i)})$.
- Compute the proposal ratio

$$r(\alpha_0) = \frac{q(\alpha_0^*|.)p(\alpha_0^{(i)}|\alpha_0^*)}{q(\alpha_0^{(i)}|.)p(\alpha_0^*|\alpha_0^{(i)})}.$$

- Let

$$\alpha_0^{(i+1)} = \begin{cases} \alpha_0^* & \text{with probability } \min(1, r(\alpha_0)) \\ \alpha_0^{(i)} & \text{with probability } 1 - \min(1, r(\alpha_0)). \end{cases}$$

Here, $q(\alpha_0|.)$ is the full conditional distribution of α_0, conditioning on all of the other variables. From (18), it follows that

$$q(\alpha_0|.) \quad \propto \quad p(\alpha_0)\left(\frac{\Gamma(\alpha_0)}{\Gamma(\xi_{[1]}\alpha_0)\ldots\Gamma(\xi_{[K]}\alpha_0)}\right)^N \prod_{n=1}^{N}\prod_{k=1}^{K}(\theta_{n[k]})^{\alpha_0\xi_{[k]}}. \tag{19}$$

Next, for ξ, we consider the proposal distribution $p(\xi^*|\xi) = \mathit{Dirichlet}\ (\delta K\xi_1, \ldots, \delta K\xi_K)$ where δ is a tuning parameter which can be adjusted. The Metropolis-Hasting step for ξ is described below:

- Sample $\xi^* \sim p(\xi^*|\xi^{(i)})$.
- Compute the proposal ratio

$$r(\xi) = \frac{q(\xi^*|.)p(\xi^{(i)}|\xi^*)}{q(\xi^{(i)}|.)p(\xi^*|\xi^{(i)})}$$

- Let

$$\xi^{(i+1)} = \begin{cases} \xi^* & \text{with probability } \min(1, r(\xi)) \\ \xi^{(i)} & \text{with probability } 1 - \min(1, r(\xi)). \end{cases}$$

Here, $q(\xi|.)$ is the full conditional distribution of ξ, conditioning on all of the other variables. From (18), we have

$$q(\xi|.) \quad \propto \quad p(\xi)\left(\frac{\Gamma(\alpha_0)}{\Gamma(\xi_{[1]}\alpha_0)\ldots\Gamma(\xi_{[K]}\alpha_0)}\right)^N \prod_{n=1}^{N}\prod_{k=1}^{K}(\theta_{n[k]})^{\alpha_0\xi_{[k]}}. \tag{20}$$

6.2.2 Infinite Mixture: Inference

In the infinite mixture case, where we assume the total number of disability profiles to be infinite with an unknown number, K of observed profiles in this data, the posterior distribution of c does not have a closed form solution. However, the full conditional distributions of the c_n for $n = 1, \ldots, N$ are known up to a normalizing constant. Using the algorithm in Figure 4, we substitute the following full conditional probabilities into step 7. The full conditional probability that senior x_n belongs in an existing (without x_n) profile k is

$$p\ (\ c_{[n]} = k \mid c_{[-n]}, x\ ) \\ \propto \frac{m(-n,k)\prod_{j=1}^{J}\Gamma\left(\tau + \displaystyle\sum_{\{i:i\neq n, c_i=k\}} x_{ji} + x_{jn}\right)\Gamma\left(\tau + \displaystyle\sum_{\{i:i\neq n, c_i=k\}} (1-x_{ji}) + 1 - x_{jn}\right)}{(N-1+\alpha)(2\tau + m(-n,k))^J\left(\Gamma(\tau)\Gamma(\tau+1)\right)^J},$$

where $c_{[-i]}$ is the profile assignment vector for all seniors other than x_i. The full conditional probability that senior x_i belongs to a profile which no other $x_{i'}$ belongs to is the following:

$$p\ (\ c_{[n]} \neq c_{[i]}\ \forall\ i \neq n \mid c_{[-n]}, x\ ) \propto \frac{\alpha}{2^J(N-1+\alpha)}.$$

The parameters of the model estimated in this way are the vector c of profile assignments and the total number of profiles, K. The posterior distributions of c and K can be found using a Gibbs sampler with these full conditional distribution. In order to asses convergence of the Markov chain, we examine the total number of profiles (which varies by Gibbs sample) and consider the Markov chain converged when the number of profiles has converged.

We diagnosed the algorithm to have converged when several independent chains sampled close values of K. We started chains with 10, 25, 40, and 21,574 profiles and they converged after approximately 25 iterations. We can be reasonably confident of convergence despite the small number of iterations because of the diversity of chain starting values.

Again, the posterior distributions of c and K depend on the values of α (the Dirichlet process parameter) and τ (the parameter of the symmetric Beta priors on the β_{jk}. Using $\alpha = 1$ is a standard value which assumes prior indifference toward groups of one member. Values of τ less than one represent a prior belief that ADL/IADL disability probabilities will tend to be close to 0 or 1 for each profile. Values of τ greater than one represent a prior belief that many disability probabilities will be close to 0.5. We choose a value of $\tau = 10$ to represent a belief that there should be profiles with intermediate probabilities.

6.3 Empirical Results

We fit three models for disability propensity profiles: the finite mixture with random Dirichlet parameter α, the finite mixture with fixed but unknown α, and the infinite mixture model.

We carry out the analysis of the NLTCS data using both MCMC and variational methods, and fitting the data with K-profiles GoM models, for $K = 2, 3, \cdots, 10$. To choose the number of latent profiles that best describes the data, we use a method that focuses on the most frequent response patterns. In the NLTCS data, what we mean by most frequent response patterns are the response patterns with observed counts greater than 100. For example, the "all-zero" response pattern (which concerns individuals with no disabilities on the 16 ADLs/IADLs) has the largest observed count of $3,853$. They are actually 24 response patterns with observed counts greater than 100 and they account for 41% of the total number of observations (which is here $21,574$). Then, using the estimates of the model parameters obtained via an MCMC algorithm

Table 3: Observed and expected cell counts for frequent response patterns under GoM models with $K = 2, 3, \cdots, 10$. The model with $K = 9$ replicate marginal pattern abundance best.

n	response pattern	observed	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$	$K = 9$	$K = 10$
1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	3853	1249	2569	2055	2801	2889	3093	2941	3269	3016
2	0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0	216	212	225	172	177	186	180	180	202	205
3	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0	1107	1176	1135	710	912	993	914	937	1010	944
4	0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0	188	205	116	76	113	200	199	181	190	198
5	0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0	122	259	64	88	58	199	90	89	116	127
6	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0	351	562	344	245	250	274	274	259	331	303
7	0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0	206	69	20	23	116	86	80	137	116	111
8	0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0	303	535	200	126	324	255	236	213	273	264
9	0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0	182	70	44	71	170	169	162	200	172	187
10	0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0	108	99	51	39	162	105	85	117	97	108
11	0 0 1 0 1 0 1 0 0 0 0 1 0 0 0 0	106	16	32	94	94	123	125	133	142	157
12	0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0	195	386	219	101	160	46	25	24	25	31
13	0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0	198	369	127	111	108	341	170	169	189	200
14	0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0	196	86	41	172	90	104	224	214	174	187
15	0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0	123	174	96	86	132	131	120	109	95	108
16	0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0	176	44	136	162	97	67	167	149	152	167
17	0 0 1 0 0 0 1 0 0 0 1 1 1 0 0 0	120	9	144	104	41	57	47	96	75	72
18	0 0 0 0 1 0 1 0 0 0 1 1 1 0 0 0	101	12	127	90	54	41	68	72	70	74
19	0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0	102	57	44	38	22	18	18	85	103	85
20	1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0	107	35	88	104	96	84	87	43	37	31
21	0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0	104	122	269	239	202	52	50	50	63	53
22	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0	164	55	214	246	272	274	276	224	166	143
23	0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	153	80	291	261	266	250	230	235	189	167
24	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	660	36	233	270	362	419	418	582	612	474

Table 4: Sum of Pearson residuals for GoM models with $K = 2, 3, \cdots, 10$.

No. of latent profiles, K	2	3	4	5	6	7	8	9	10
Sum of squares$\times 10^5$	75	20	37	13	11	7.7	9.4	4.4	8.2
Sum of absolute residuals	20684	4889	5032	1840	2202	2458	1908	1582	1602

or a variational EM algorithm, we can compute the expected cell counts for the 24 response patterns and compare with the observed cell counts. Eventually, to choose the model that best fits the data, we can compute the sum of absolute values of the "Pearson chi-square" residuals [10],

$$\frac{\text{Observed Count} - \text{Expected Count}}{\sqrt{\text{Expected Count}}},$$

for each model .

Table 3 provides the expected cell counts for the 24 most frequent response patterns (to be compared with the observed cell counts) using MCMC methods (for $K = 2, \ldots, 10$). We could observe from this results that the model with K=9 has a better fit for the "all-zero" response pattern, the "all-one" response pattern and the pattern number $n = 3$ (pattern with only one 1 on the IADL "doing heavy housework"). The computation of the sum of Pearson residuals confirms that $K = 9$ seems to be a good choice. This is also true when one computes the expected cell counts using the variational methods.

To deal with this issue of model choice, we can also compute a version of DIC directly using the output from MCMC simulations. Indeed, if we focus on parameters θ and β, the computation is done using draws from the posterior distribution of β_{jk} and θ_k. Figure 9 shows the plot of DIC for models with $K = 2, 3, \cdots, 10$ latent profiles. According to the DIC plot, we choose models with $K = 8$ or $K = 9$ latent profiles. Using variational approximation methods, we

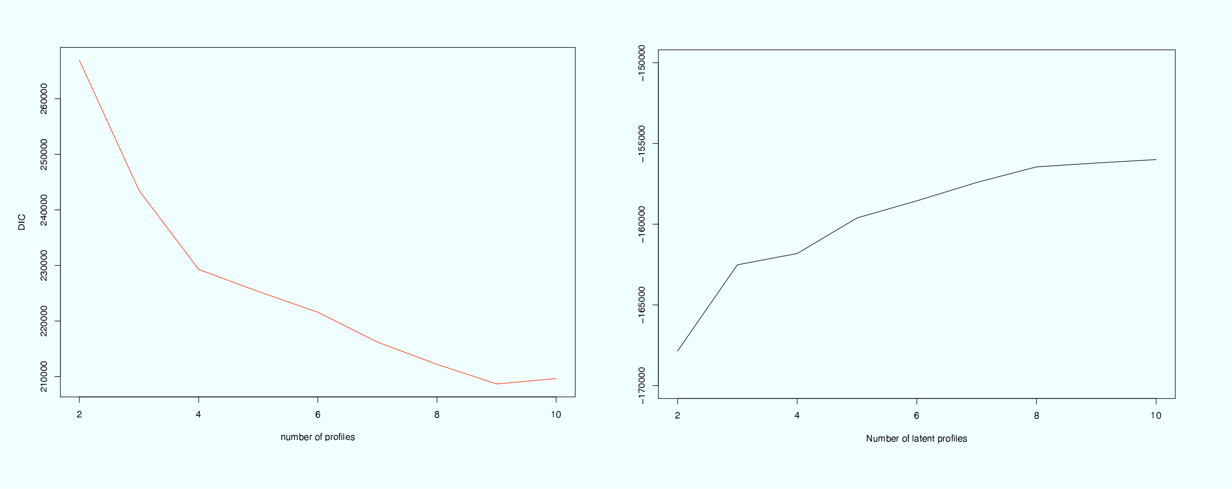

Figure 9: Left Panel: DIC for $K = 2, \cdots, 10$ latent profiles (GoM model). Right Panel: BIC for $K = 2, \cdots, 10$ latent profiles (GoM model).

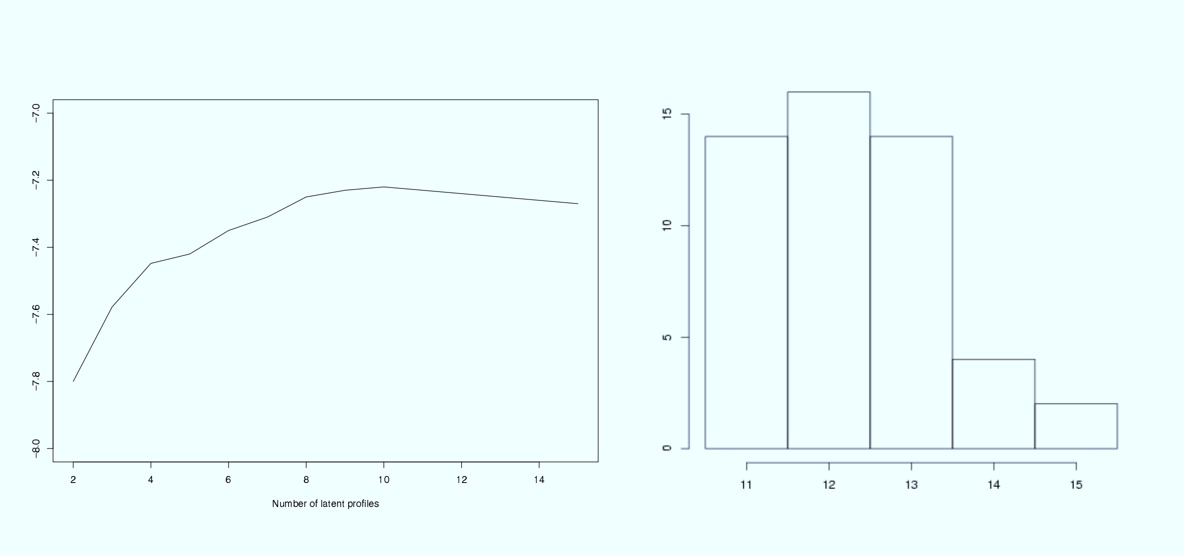

Figure 10: Left Panel: Log-likelihood (5 fold cv) for $K = 2, \ldots, 10, 15$ (GoM model). Right Panel: Posterior distribution of K.

also computed an approximate version of BIC based on the variational approximation. Figure 9 shows the plot of BIC for models with $K = 2, 3, \cdots, 10$ latent profiles. This criterion suggests a number of profiles around 8. The cross validation results shown in Figure 10 (using variational approximation methods) also suggest the choice of 8 or 9 profiles.

The infinite model generates a posterior distribution for the number of profiles, K, given the data. Figure 10 shows the posterior distribution ranges from 11 to 15 profiles. We expect that the infinite model will require more profiles because it is a hard clustering.

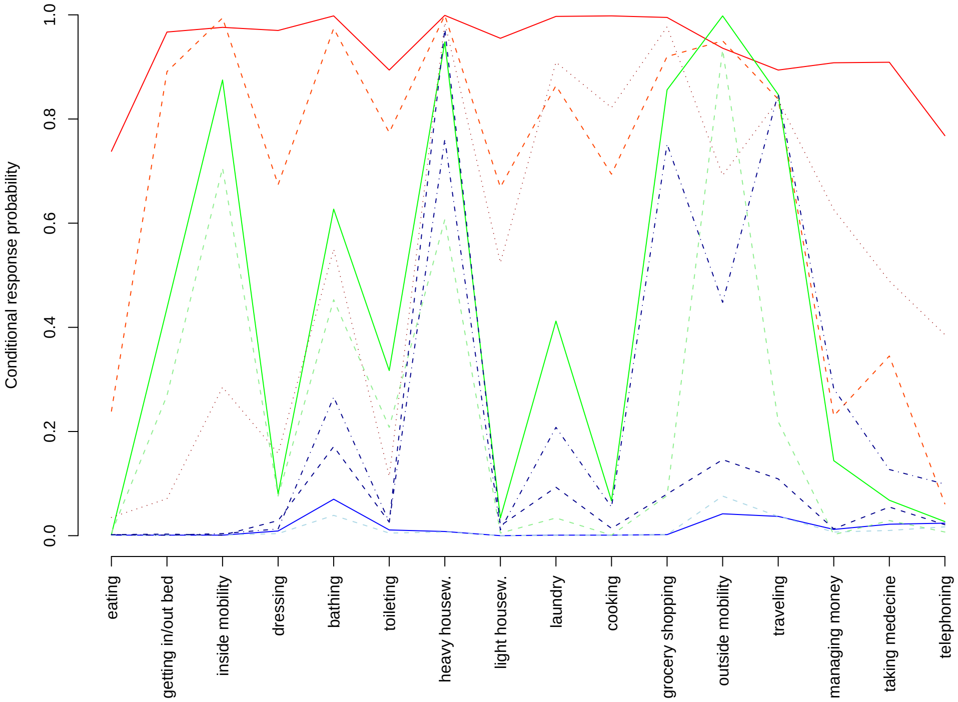

Figure 11: Latent profiles for the GoM model with K=9.

According to the array of different criteria we have considered, $K = 9$ seems to be an appropriate choice for the NLTCS data. Figure 11 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 $\beta_{jk} = P(x_{nj} = 1|\theta_{nk} = 1)$, the probability of being disabled on the activity j for a complete member of latent profile k. The profiles have the following interpretation:

- We can clearly distinguish two profiles for "healthy" individuals; these are the lower curves (the solid, blue curve and the dashed, light blue curve).
- The upper curve (solid, red curve) corresponds to seriously "disabled" individuals since most of the probabilities are greater than 0.8.
- One profile (dotted, brown 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 (dotted, orange curve) characterizes "semi-disabled" individuals.
- The profile with very high probabilities for all the activities involving mobility including the IADL "outside mobility" (solid, green curve) characterizes mobility-impaired individuals.
- Another profile characterizes individuals who are relatively healthy but can't do "doing heavy housework" (dashed, blue curve). Note that in Table 3, the response pattern $n = 3$ has the second largest observed cell count.

- The two remaining profiles (the dot-dashed, blue curve and the dashed, light green curve) corresponds to individuals who are "semi-healthy" since they show limitation s 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 β_{jk}.

Because the NLTCS data is characterized by a large amount of healthy individuals with "all zero" response patterns (there are $3,853$ all zero response patterns and they represent a little less than 18% of the sample population), we would like to take into account this excess of healthy individuals. In a forthcoming paper focusing on the results of analyses with the GoM model, we plan carry out an extended analysis using a modified version of the GoM model which adjusts for this excess.

7 Concluding Remarks

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 have used "latent Dirichlet allocation" which has some breadth of currency in the data mining literature, and shown how extensions to it to analyze a corpus of PNAS biological sciences publications from 1997 to 2001. Among the approaches to select the number of latent topics which we study are k-fold cross-validation and the use of a Dirichlet process prior. Our results focus on six combinations of models and model choice strategies. They lead us to report on and interpret results for $K = 20$ topics, a value that appears to be within the range of possibly optimal numbers of topics. The resulting topics are also easily interpretable and profile the most popular research subjects in biological sciences, in terms of the corresponding words and references usage patterns. Much higher choices for K, lead to faulty and difficult to interpret conclusions. Incidentally, our 20 topics correlate well with the PNAS editorial categories.

For the analysis of the NLTCS data, we have developed parametric and nonparametric variations the GoM model. We performed posterior inference using variational methods and MCMC. We have used different criteria to assess model fit and choose K; in particular a method based on the sum of Pearson residuals for the most frequent response patterns, and information criteria such as DIC and BIC. We have then reached the conclusion that $K = 9$ latent profiles is an appropriate choice for the data set. This choice allows us to identify profiles that did not appear in the analysis performed in [18]; for instance, the profile for individuals who are pretty healthy on all the activities but "doing heavy housework." Further, we were able to interpret all the 9 profiles, whereas with $K = 4$ and $K = 5$, these profiles could not be ordered by severity. Nonetheless, once we reach $K = 5$, the fit seems not to improve markedly.

Acknowledgments

This work was partially supported by National Institutes of Health under Grant No. R01 AG023141-01, by the Office of Naval Research under Contract No. N00014-02-1-0973, and by the National Science Foundation Grant No. DMS-0240019, all to Carnegie Mellon University. We are especially grateful to Elena Erosheva for comments on an earlier draft of this paper.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Latent mixed-membership allocation models of relational and multivariate attribute data. In *Valencia & ISBA Joint World Meeting on Bayesian Statistics*, 2006. Forthcoming.

[2] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the International Biometrics Socity (ENAR) Annual Meetings*, 2006. Forthcoming.

[3] E. M. Airoldi and C. Faloutsos. Recovering latent time-series from their observed sums: network tomography with particle filters. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 10, pages 30–39, 2004.

[4] E. M. Airoldi, S. E. Fienberg, C. Joutard, and T. M. Love. Discovery of latent patterns with hierarchical bayesian mixed-membership models and the issue of model choice. Technical Report CMU-MLD-06-101, School of Computer Science, Canregie Mellon University, April 2006.

[5] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281, 1973.

[6] D. J. Aldous. Exchangeability and related topics. In *Lecture Notes in Mathematics*, pages 1–198. Springer, Berlin, 1985.

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

[8] 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.

[9] K. J. Bathe. *Finite Element Procedures.* Englewood Cliffs, NJ: Prentice Hall, 1996.

[10] Y. Bishop, S. E. Fienberg, and P. Holland. *Discrete Multivariate Analysis: Theory and Practice.* The MIT press, 1975.

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

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

[13] D. Chakrabarti, S. Papadimitriou, D. Modha, and C. Faloutsos. Fully automatic cross-associations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 10, pages 79–88, 2004.

[14] J. Connor. *Multivariate Mixture Models to Describe Longitudinal Patterns of Frailty in American Seniors.*. PhD thesis, Carnegie Mellon University, Department of Statistics, 2006.

[15] 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.

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

[17] E. A. Erosheva. Bayesian estimation of the grade of membership model. In *Bayesian Statistics*, volume 7, pages 501–510, 2003.

[18] E. A. Erosheva and S. E. Fienberg. Bayesian mixed membership models for soft clustering and classification. In C. Weihs and W. Gaul, editors, *Classification—The Ubiquitous Challenge*, pages 11–26. Springer-Verlag, 2005.

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

[20] T. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1:209–230, 1973.

[21] Z. Ghahramani. Unsupervised learning. In O. Bousquet, G. Raetsch, and U. von Luxburg, editors, *Advanced Lectures on Machine Learning*, volume LNAI 3176. Springer-Verlag, 2005.

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

[23] J. Han and M. Kamber. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2000.

[24] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, 2001.

[25] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[26] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Random graph models for the web graph. In *Annual Symposium on Foundations of Computer Science*, pages 57–65, 2000.

[27] K. G. Manton, M. A. Woodbury, and H. D. Tolley. *Statistical Applications Using Fuzzy Sets.* Wiley, 1994.

[28] 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*, 2006. Forthcoming.

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

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

[31] 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*, volume 11, 2005.

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

[33] G. Parisi. *Statistical Field Theory*. Redwood City, CA: Addison-Wesley, 1988.

[34] D. Pelleg and A. W. Moore. X-means: Extending K-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning*, volume 17, pages 727–734, 2000.

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

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

[37] 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.

[38] J. Rustagi. *Variational Methods in Statistics*. New York: Academic Press, 1976.

[39] J. Sakurai. *Modern Quantum Mechanics*. Redwood City, CA: Addison-Wesley, 1985.

[40] G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[41] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–639, 2002.

[42] E. Stallard. Trajectories of disability and mortality among the U.S. elderly population: Evidence from the 1984-1999 NLTCS. In *Living to 100 and Beyond International Symposium*. Society of Actuaries, 2005.

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

[44] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, 2003.

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

[46] M. A. Woodbury, J. Clive, and A. Garson. Mathematical typology: Grade of membership technique for obtaining disease definition. *Computational Biomedical Research*, 11(3):277–298, 1978.

[47] E. P. Xing, M. I. Jordan, and S. Russell. A generalized mean field algorithm for variational inference in exponential families. In *Uncertainty in Artificial Intelligence*, volume 19, 2003.