NIH Public Access
Author Manuscript
IEEE Signal Process Mag. Author manuscript; available in PMC 2014 August 05.

Published in final edited form as:
IEEE Signal Process Mag. 2010 November 1; 27(6): 55–65. doi:10.1109/MSP.2010.938079.

Probabilistic Topic Models:

A focus on graphical model design and applications to document and image analysis

David Blei, **Lawrence Carin**, and **David Dunson**

In this article, we review *probabilistic topic models:* graphical models that can be used to summarize a large collection of documents with a smaller number of distributions over words. Those distributions are called “topics” because, when fit to data, they capture the salient themes that run through the collection. We describe both finite-dimensional parametric topic models and their Bayesian nonparametric counterparts, which are based on the hierarchical Dirichlet process (HDP). We discuss two extensions of topic models to time-series data—one that lets the topics slowly change over time and one that lets the assumed prevalence of the topics change. Finally, we illustrate the application of topic models to nontext data, summarizing some recent research results in image analysis.

INTRODUCTION

Hierarchical mixture modeling has emerged as a powerful methodology for finding patterns and structure in large collections of data. A recent success story for hierarchical mixture modeling is topic modeling where the data under study are large collections of documents and mixture modeling algorithms find the underlying patterns of words that are embedded in the collection (see Figure 1) [15], [35]. Finding these patterns, which are known as topics allows for effective clustering, searching, sorting, exploring, predicting and summarizing a large corpus of documents.

While developed as a way of analyzing documents, topic modeling algorithms— which are fast algorithms for computing with hierarchical mixture models—have been successfully applied in many domains. For example, topic modeling algorithms have been used to find patterns in images, music [34], audio and speech [33], [49], genetic data [46], computer code [4], and even architectural excavations [40]. In particular, applications to computer vision are extensive [6], [10], [16], [19], [28], [37], [57], [59] and researchers have used topic models in a variety of computer vision problems. Examples include sorting multiple images into scene-level classes, annotating images with words, and segmenting and labeling objects within images. Recently, researchers have extended such statistical topic models to analysis of video [60], [61].

Furthermore, new applications of topic modeling have driven new statistical developments in hierarchical mixture modeling. Topic models originally assumed that the data are exchangeable, i.e., that the order of documents in a collection does not matter and that the order of words in a document does not matter. This is too restrictive for many problems, and relaxing this assumption is a central way of building better topic models; topic models have now been applied and extended in spatial settings [3], [24], [53], time series settings [14], [47], [62], and settings that depend on external covariates [39].

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Another notable statistical development in topic modeling is their extension to Bayesian nonparametric methods. The original topic models were finite hierarchical mixtures, i.e., parametric mixture models. Recent advances in Bayesian nonparametric modeling, specifically the HDP [54], has lead to "infinite" topic models. The number of topics need not be specified in advance, and can grow as the collection grows.

In this article, we review recent progress on development of hierarchical mixture models, with a specific focus on topic models. We initially focus on analyzing documents and describe both finite topic models and Bayesian nonparametric topic models (of infinite capacity). We then describe extensions to sequential document collections to show how partial exchangeability may be incorporated into hierarchical mixture models. Finally, we briefly discuss how topic models may be applied to other applications, and show results on analyzing partially annotated images. Throughout, we emphasize the graphical model constructions of the associated models. Graphical models provide a useful schematic of topic modeling assumptions—they are the lever from which researchers have been able to build topic modeling extensions and a myriad of applications.

LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation (LDA) is a hierarchical probabilistic model used to decompose a collection of documents into its salient topics, where a "topic" for LDA is a probability distribution over a vocabulary [15]. LDA and its relatives are called probabilistic topic models.

LDA posits a fixed number of topics in a document collection and assumes that each document reflects a combination of those topics. When a document collection is analyzed under these assumptions, probabilistic inference algorithms reveal an embedded thematic structure (see Figure 1). With this structure, LDA provides a way to quickly summarize, explore, and search massive document collections.

A topic β is a distribution over a fixed vocabulary of V terms, and recall that the Dirichlet distribution is a distribution over multinomial parameter vectors, i.e., vectors of positive values that sum to one. The generative probabilistic assumptions of LDA assume that a document collection is drawn as follows (Figure 2).

1. Draw K topics from a symmetric Dirichlet distribution, $\beta_k \sim \text{Dir}_V(\eta)$, $k \in \{1, \ldots, K\}$.
2. For each document d, draw topic proportions from a symmetric Dirichlet $\theta_d \sim \text{Dir}_K(\alpha)$, $d \in \{1,\ldots,D\}$.
3. For each word n in each document d,
    - Draw a topic assignment from the topic proportions, $z_{d,n} \mid \theta_d \sim \text{Mult}(\theta_d)$.
    - Draw the word from the corresponding topic, $w_{d,n} \mid z_{d,n}, \beta_{1:K} \sim \text{Mult}(\beta_{z_{d,n}})$.

The graphical model in Figure 2 reveals the nested multilevel structure of the LDA assumptions. LDA is composed of a hierarchy of mixture models. Each document is

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

modeled with a finite mixture model, where the mixture proportions (i.e., the topic proportions) are drawn uniquely for each document but the mixture components (i.e., the topics) are shared across the collection. In statistics, this is known as a grade of membership or mixed membership model [25]. LDA builds on seminal work in psychology [23] and machine learning [35]. It has close links to classical principal component analysis [18].

The generative process defines a joint distribution of the latent variables (topics, topic proportions, and topic assignments) and observed variables (words). We analyze a document collection by examining the posterior distribution of the latent variables given the observations

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N} \mid w_{1:D,1:N}). \quad (1)$$

Posterior modes of the topics $\beta_{1:K}$ identify corpus-wide patterns of words; posterior modes of the topic proportions θ_d identify how the *d*th document expresses those patterns; posterior modes of the topic assignments $z_{d,n}$ identify which topic the *n*th word of the *d*th document is associated with.

Posterior inference can be thought of as “reversing” the generative process. Conditioned on a corpus, the goal of posterior inference is to find the posterior distribution over alternative topical structures that generated its documents. This posterior provides structure that is unavailable in the words alone, and this is the structure that can be used to explore and summarize a document collection. For example, Figure 1 illustrates the topics from a 20-topic LDA posterior estimated from articles from *Science*. When fit to language, the posterior distribution of LDA has been shown to match human cognition of themes and associations [32], [52], and to be interpretable for document and corpus understanding [22,] [44].

Exactly computing the posterior of (1) is intractable [15]. Approximating the posterior is the central computational problem for LDA, and devising and improving posterior approximation algorithms is an active area of topic modeling research. Most methods rely either on variational inference [15] [56] or Markov chain sampling [32]. In variational inference, we approximate the posterior by optimizing a distribution to be close to it. In Markov chain sampling, we define a Markov chain whose stationary distribution is the LDA posterior, run the chain for a long time, and collect samples from (what we hope is) the stationary distribution.

These algorithms allow for large scale analysis of document collections, and speeding up LDA inference is an active area of machine learning research. Progress in inference algorithms has included processing on a graphics processing unit [66], parallel processing across machines [43], and combinations of variational and sampling methods [63]. Moreover, several implementations of LDA inference are freely available as open-source software [8], [20], [38]. (The model of Figure 1 was fit with [20].)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

HIERARCHICAL DIRICHLET PROCESSES

Though we can use LDA to successfully learn the topics from a document collection, LDA analysis requires that the number of topics be fixed in advance (e.g., in Figure 1 the number of topics was fixed at 20). This is a serious limitation—one of the tenets of hierarchical modeling with latent variables is to let the data inform the hidden structure. Choosing the number of topics in LDA is usually done by examining the fit to held-out documents [15], or by selecting based on the marginal probability of the whole collection [32], [36].

Bayesian nonparametric methods—methods that place priors on the infinite-dimensional space of probability distributions—provide an elegant solution to this problem. In particular, the discrete HDP can be used in a topic model where the number of topics is "infinite" a priori [54]. Given a collection of documents, the number of topics that it exhibits becomes part of the posterior distribution of the latent structure.

Furthermore, the HDP allows for previously unseen documents to "ignite" previously unseen topics. This is a particularly attractive property for analyzing growing and changing collections. Even with an excellent approach to selecting the number of topics based on a finite sample, it often does not make sense to assume that all future documents will only use those topics. The HDP allows new topics to emerge naturally as a consequence of the probability model.

Before describing the HDP topic model, we review the Dirichlet process (DP) [5], [29]. The DP provides a distribution on distributions over an arbitrary space. It is denoted

$$G \sim DP(\alpha, G_0). \quad (2)$$

In this expression, G is itself a distribution (it is a random distribution) over some space. Its distribution has two parameters: the precision parameter alpha a positive scalar, and the base distribution G_0 is a known distribution over the same space as G. (We will describe the roles of these parameters in detail below.) The DP can be used as a prior over the infinite dimensional space of distributions. In these settings, it is thus referred to as a Bayesian nonparametric method [41].

For developing a Bayesian nonparametric topic model, there are two important properties of the DP. First, draws from the DP are discrete, with positive probability mass placed at values (called "atoms") generated independently from G_0. If G_0 is continuous or contains infinitely many atoms, then each draw from the DP will assign nonzero probability to each of infinitely many values. The allocation of probability across these atoms is controlled by the precision parameter α, with small values of α implying that a few dominant atoms get almost all of the probability. Higher values of α lead to discrete distributions that more closely resemble the base distribution G_0. See Figure 3 for an illustration of this property.

For example, suppose that G_0 is a standard Gaussian distribution with mean zero and unit variance. The random distribution G is a distribution over the real line, with positive mass at a countably infinite set of points drawn from G_0. If α is large then this distribution looks like

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

a discrete version of a standard Gaussian; if α is small, then some individual points will have much higher probability.

The second property of a draw from a DP is the so-called "clustering property." If we draw G from a DP and then draw variables repeatedly from G then our draws will exhibit a partition structure, and we can partition them according to which draws share the same atom. In the example above, this means that drawing X_n repeatedly from G will lead to a collection of real numbers, some of which are exactly the same. Note that repeated draws from the standard Gaussian will not have this property.

This perspective sheds light on the role of the base distribution G_0 and scaling parameter α. The unique values drawn (i.e., the atoms) are independent draws from G_0 and if G_0 is continuous (as in the Gaussian example) then the number of unique values increases with α. Though we focus on the random distribution representation here, the clustering effect links DPs to models of random partitions, specifically the Chinese restaurant process [45] and the Ewens sampling formula [27].

We now return to topic modeling. In LDA, the topic proportions are a distribution over the K topics. To build a Bayesian nonparametric topic model, we replace the topic proportions θ_d, drawn from a finite Dirichlet, with a distribution over topics G_d, drawn from a DP. The atoms of G_d are topics, i.e., multinomial parameters over the vocabulary, and so the base distribution G_0 is a distribution over topics (this replaces the Gaussian base distribution in the example above). To draw a document, we first draw a distribution $G_d \sim DP(\alpha, G_0)$. Then, for each word, we draw a topic β_n from G_d and finally draw the word from β_n. The clustering property is critical—it guarantees that the words of the document will share a smaller subset of topics.

However, this is not enough to fully specify a Bayesian nonparametric topic model. A defining characteristic of LDA is that the topics themselves are shared across the corpus. If the base distribution G_0 is a continuous distribution over topics, e.g., a symmetric Dirichlet distribution, then words within documents will share the same topics but words across documents will not. In the HDP, the base distribution G_0 is itself a draw from a DP, $G_0 \sim DP(\gamma, H)$. The atoms of the per-document distributions over topics G_d are thus shared across documents [54].

Putting this together, the HDP topic model draws a collection of documents from the following process.

1. Draw the base distribution over topics $G_0 \sim \text{DP}(\gamma, H)$, where H is a symmetric Dirichlet on the word simplex.
2. For each document d, draw the per-document distribution over topics $G_d \sim \text{DP}(\alpha, G_0)$.
3. For each word n in each document d,
    - Draw the topic for the word $\beta_{d,n} \sim G_d$.
    - Draw the word $w_{d,n} \sim \text{Mult}(\beta_{d,n})$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

This is illustrated as a graphical model in Figure 4.

In theory, the posterior distribution for the HDP topic model provides per document distributions over topics G_d and the per-corpus distribution over topics H. In practice, however, the random distributions themselves are marginalized out and posterior inference provides topic indices (as for LDA) and posterior distributions of the corresponding distributions over words. The key difference is that in the HDP the number of topics is determined by the data. Moreover, as we mentioned above, when performing prediction a new document can "ignite" a new topic—it is simply considered an atom that had not yet appeared.

As for LDA, exact posterior inference for the HDP is intractable and a number of methods have been developed. The original approximate inference algorithm was based on Markov chain Monte Carlo (MCMC) sampling [54], but faster variational approaches have been recently proposed [55]. When compared to LDA, [54] showed that the HDP topic model finds the "right" number of topics in collection, when using model selection based on marginal likelihood.

MODELING SEQUENTIAL COLLECTIONS

While powerful, LDA assumes that documents are exchangeable. That is, their ordering is irrelevant to determining the corpus-wide topics and the decomposition of each document into those topics. In this section, we describe two topic models that account for sequential document collections. In different ways, these methods model language that changes over time.

DYNAMIC TOPIC MODELING

The dynamic topic model (DTM) accounts for topics changing over time [12]. It assumes that the corpus is organized into epochs, each epoch is associated with its own set of topics, and each topic in each epoch drifts randomly from the same topic in the previous epoch.

Given topics for a particular epoch, the generative process of the documents of that epoch is the same as for LDA. Recall that a topic is a multinomial parameter for a distribution over words. To build the DTM, we need to specify a model how each topic evolves from epoch to epoch.

The building block for this sequential model of multinomial parameter vectors is the logistic normal distribution. A logistic normal distribution is a distribution of multinomial parameters that is an alternative to the more commonly used Dirichlet distribution. It was first developed to allow for complex patterns of correlation between components of the vector [2]. The idea is to draw a real-valued vector from a multivariate Gaussian, and then to transform to a multinomial parameter by first exponentiating it and then renormalizing it to sum to one.

For the DTM, the logistic normal is embedded in a state-space model [64]. The covariance matrix for the topic simplex is diagonal, i.e., the components are independent. However, the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

mean of the distribution is the mean of the previous epoch's real-valued random vector. Thus, to obtain topic k at time t,

$$y_{t,k} \sim N(y_{t-1,k}, \mathrm{I}\sigma^2)$$
$$\beta_{t,k,\nu} = \frac{\exp\{y_{t,k,v}\}}{\Sigma_v \exp\{y_{t,k,v'}\}} \quad (3)$$

with $\beta_{t,k,v}$ representing the probability of word $v \in \{1,\ldots,V\}$, topic k at time t. The variance parameter σ^2 controls the prior drift of the log probability of each word. Notice the difference between the DTM and LDA. In LDA, each word is associated with a probability under each topic. In the DTM, each word is associated with a sequence of probabilities under each topic. These generative assumptions lead to a richer posterior distribution. See Figure 5 for the graphical model representation of the full DTM model.

With an approximation of the posterior distribution (see [12] for a structured variational approach), the richer latent space allows for new ways to investigate long-running sequential corpora. First, we can examine how topics changed over time, by looking at the top words from the topic at each epoch. Second, we can analyze how a particular word's probability changed over time within a topic. Finally, we can examine articles associated with a topic at different times. Note that this associates articles with each other while taking into account how word use has changed. Traditional document clustering methods do not account for such shifts.

As an example, we analyzed the journal *Science* (1880–2002) using the DTM. The articles were scanned by the service JSTOR, and divided by year. Investigations of two topics are illustrated in Figure 6.

The DTM provides a window into the collection that LDA or the HDP do not provide. The DTM has been additionally extended to model continuous time [58] and as a component in a model of document influence, i.e., a language-based approach to finding high impact scholarly documents [31]. Note that developing the HDP analog of the DTM is an open issue in topic modeling. Ideally, a kind of "birth" and "death" of topics over time should be modeled, and perhaps a splitting and merging of topics as well. The beginnings of this kind of model were developed in [48].

DYNAMIC HIERARCHICAL DIRICHLET PROCESSES

The DTM models topics that change over time. In this section, we describe a different kind of time-based topic model, one that models the topic proportions changing over time. This model—the dynamic HDP (dHDP) [49]—is based on extensions of the HDP [54] discussed in the section "Hierarchical Dirichlet Processes."

Rather than positing a document-dependent distribution over topics G_d, as discussed above, the dHDP posits a distribution over topics G_t, which corresponds to all documents at time t. The goal is to impose a generally smooth evolution of G_t as time evolves, with the potential for sharp changes in time, as needed to describe the data. The G_t is are assumed drawn as follows:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ \tilde{w}_t &\sim \mathrm{Beta}(a_t, b_t) \\ H_t &\sim \mathrm{DP}(\alpha_{0t}, G_0) \end{aligned} \quad (4)$$

$$G_t = (1 - \tilde{w}_t) G_{t-1} + \tilde{w}_t H_t \quad (5)$$

with $w_1\tilde{} = 1$.

The graphical model is depicted in Figure 7. The parameters (a_t, b_t) are typically made time independent for simplicity, and set such that with high probability $w_t\tilde{}$ will be small, but infrequently $w_t\tilde{}$ may be near one, indicating a sharp change in the probability of topics at a given time (e.g., because a new subject/ topic becomes important suddenly at a given time). In the dHDP, since G_0 is discrete, each of the G_t is composed of the same discrete set of topics and the probability of each topic being used in a document evolves with time. With this model we again infer the number of topics nonparametrically from the observed data, but the temporal exchangeability of the data (associated with HDP) has been removed.

Using appropriate approximations, the dHDP model has been implemented using variational Bayesian inference [11] and applied to the United States presidential State of the Union addresses from 1790 to 2008, with example results depicted in Figure 8 [47]. The number of topics, the time periods when they are important, and the topic-dependent probability of words are inferred from the data, while human (author) defined labels are associated (imperfectly) to each of the topics.

OTHER APPLICATIONS OF TOPIC MODELS

The examples discussed above have focused on analysis of documents. However, topic models have been extended to many other kinds of data. For example they are widely used in computer vision problems, i.e., for the analysis of images, video, and simultaneous analysis of images and words (e.g., annotations) [3], [6], [10], [19], [24], [28], [37], [51], [53], [57], [59]–[61], [65]. Additionally, researchers have applied topic models to sound features, genetic markers [46], survey responses [26], computer code [4], and social network data [1], [21], [42]. Topic modeling began as a field about finding structure in texts, but has become an area of research that exploits grouped data in many settings.

For these applications, the basic form of the graphical model is typically very similar to that associated with words/documents, with differences manifested in the form of the "words." For example, in image or video processing, if one quantizes the image features, each member of the (discrete and finite) code-book now takes the form of the words [3], [24], [28], [51], [53]. In the time-dependent analysis of documents, the exchangeability assumption in the DP was removed, allowing one to exploit the temporal information (documents that are temporally proximate are likely to be composed of similar topics). In the image processing application, the image words (codes) typically correspond to local regions in the image, and it is again desirable to account for the known inter-relationships between image words, now defined by their spatial location within the image. Specifically, the model parameters associated with proximate regions of an image are likely to be more correlated.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

There have been several recent papers that have addressed this issue [3], [24], [53], explicitly accounting for the spatial location of features within an image.

As an example of recent research on moving beyond the exchangeability ("bag of image words") assumption, consider the logistic stick-breaking process (LSBP) [24]. The LSBP and the related kernel stick-breaking process (KSBP) [3], uses a spatial kernel to define proximity within an image. The radial basis function is a widely employed kernel for such purposes. The probability of whether image features are drawn from the same topic is defined by a probability distribution constituted in terms of spatial kernel functions. Specifically, to map the kernel output to a probability, the kernel is utilized within the logistic link function. Via this construction, image features that are spatially proximate are more probable to be associated with the same topic. It is also possible to associate words with the topics [24], and therefore the topics are manifested in two forms: by spatially dependent image features and via words, the latter constituted in the form of image annotations. Such a model allows one to label localized regions (objects) within an image, based upon a set of annotated images.

The use of a kernel construction to impose spatial structure within a topic model is intuitive and computationally efficient. Alternative means of imposing spatial structure have employed nonparametric constructions, such as spatially dependent Gaussian processes and the Pitman-Yor process [53]. Researchers have also recently developed a distance dependent Chinese restaurant process [9], which removes the exchange-ability assumption inherent to the DP.

We summarize example results from [24]. Figure 9 shows example results that demonstrate the ability to nonparametrically infer the number of topics in a database of images (here the database was designed with ten different classes of images, with this properly inferred by the model), and to associate words with each topic. Once such a model is learned, it may be used to annotate previously nonannotated images, assuming the new images cover the same range of topics as those used for learning. In addition to providing a means of annotating images, such topic models may also be used to label segmented objects within an image. Figure 10 shows example results [24] of jointly segmenting and labeling objects in multiple images, based upon a set of annotated images (the objects are not explicitly labeled in the training set, simply annotated, with the relationship between objects/ segments and words inferred by the model). Figure 10 also shows the evolution in the performance of image-processing-related topic models. Corr-LDA [10] is an early model that seeks to label entities within images, but it does not explicitly account for spatial locations of the image words, and the performance of this model is relatively poor. The DP-based and LSBP-based clustering of spatially dependent image words yields significant improvements in the ability to link objects and words in images, within a Bayesian topic-modeling framework.

DISCUSSION

We have described recent developments in hierarchical mixture models, specifically in topic modeling of large document collections and image collections. We have discussed how the exchangeability assumption can be relaxed in both spatial and sequential settings. We have

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

described flexible Bayesian nonparametric models based on the HDP, to perform model selection of the number of topics as part of posterior inference.

We have focused on graphical model design and on applications to problems of interest to the signal-processing community. Note that we have not addressed computational tools for efficient calculation of the posterior distribution. Modeling and computation are intertwined —many modeling decisions made in developing new topic models keep the ease of approximate posterior inference in mind. For example, when consecutive elements in the graphical model are conjugate-exponential family pairs, then the Gibbs sampling algorithm and variational coordinate ascent algorithm can be derived in closed form [7], [30]. That said, relaxing this restriction allows for more expressive models (the DTM is an example) though also requires more sophisticated approximate methods, e.g., by using the delta method in variational inference [17] or Metropolis-Hastings updates in MCMC [50].

Biographies

David Blei (blei@cs.princeton.edu) is an assistant professor of computer science at Princeton University. He earned his Ph.D. degree in 2004 at the University of California, Berkeley. His main research focuses on developing large scale Bayesian models of high-dimensional data, often employed for better exploration, understanding, and visualization. Much of his work has been on developing models of large document collections, though he has also developed methods for images, music, legislative records, and brain recordings.

Lawrence Carin (lcarin@ece.duke.edu) earned the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1985, 1986, and 1989, respectively. He is now the William H. Younger Professor of Engineering at Duke University. He is the cofounder and director of technology at Signal Innovations Group, Inc. His current research interests include signal processing, sensing, and machine learning. He has published over 200 peer-reviewed papers. He is an IEEE Fellow and a member of the Tau Beta Pi and Eta Kappa Nu honor societies.

David Dunson (dunson@stat.duke.edu) is a professor of statistical science at Duke University. His research focuses on the development and application of novel Bayesian statistical methods motivated by high-dimensional and complex data sets. He is a fellow of the American Statistical Association and of the Institute of Mathematical Statistics. He was won the 2007 Mortimer Spiegelman Award and is the 2010 Myrto Lefkopoulou Distinguished Lecturer at Harvard University.

REFERENCES

1. Airoldi E, Blei D, Fienberg S, Xing E. Mixed membership stochastic blockmodels. J. Mach. Learn. Res. 2008; 9:1981–2014. [PubMed: 21701698]
2. Aitchison J. The statistical analysis of compositional data. J. R. Statist. Soc., Ser. B. 1982; 44(no. 2): 139–177.
3. An Q, Wang C, Shterev I, Wang E, Carin L, Dunson DB. Hierarchical kernel stick-breaking process for multi-task image analysis. Proc. Int. Conf. Machine Learning. 2008
4. Andrzejewski D, Mulhern A, Liblit B, Zhu X. Statistical debugging using latent topic models. Proc. European Conf. Machine Learning. 2007

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

5. Antoniak CE. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Statist. 1974; 2(no. 6):1152–1174.
6. Barnard K, Duygulu P, de Freitas N, Forsyth D, Blei D, Jordan MI. Matching words and pictures. J. Mach. Learn. Res. 2003 Mar.3:1107–1135.
7. Bishop, C.; Spiegelhalter, D.; Winn, J. VIBES: A variational inference engine for Bayesian networks. In: Becker, S.; Thrun, S.; Obermayer, K., editors. Advances in Neural Information Processing Systems. Vol. 15. Cambridge, MA: MIT Press; 2003. p. 777-784.
8. Blei, D. [Online]. Available: http//www.cs.princeton.edu/blei/lda-c/
9. Blei D, Frazier P. Distance dependent Chinese restaurant processes. Proc. Int. Conf. Machine Learning. 2010
10. Blei, D.; Jordan, M. Proc. ACM Special Interest Group on Information Retrieval (SIGIRâ903). 2003. Modeling annotated data.
11. Blei D, Jordan M. Variational inference for Dirichlet process mixtures. Bayesian Anal. 2006; 1(no. 1):121–144.
12. Blei, D.; Lafferty, J. Dynamic topic models; Proc. Int. Conf. Machine Learning (ICML'06); 2006. p. 113-120.
13. Blei, D.; Lafferty, J. Topic models. In: Srivastava, A.; Sahami, M., editors. Text Mining: Theory and Applications. New York: Taylor & Francis; 2009.
14. Blei, D.; Lafferty, JD. Dynamic topic models; Proc. 23rd Int. Conf. Machine Learning; 2006. p. 113-120.
15. Blei D, Ng A, Jordan MI. Latent Dirichlet allocation. J. Mach. Learn. Res. 2003 Mar.3:993–1022.
16. Bosch, A.; Zisserman, A.; Munoz, X. Scene classification via pLSA; Proc. European Conf. Computer Vision; 2006.
17. Braun M, McAuliffe J. Variational inference for large-scale models of discrete choice. J. Amer. Statist. Assoc. 2010:324–335.
18. Buntine, W.; Jakulin, A. Subspace, Latent Structure and Feature Selection. New York: Springer-Verlag; 2006. Discrete component analysis.
19. Cao L, Fei-Fei L. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. Proc. Int. Conf. Computer Vision. 2007
20. Chang, J. [Online]. Available: http://cran.r-project.org/web/packages/lda/
21. Chang J, Blei D. Hierarchical relational models for document networks. Ann. Appl. Statist. 2010; 4(no. 1)
22. Chang, J.; Boyd-Graber, J.; Gerrish, S.; Wang, C.; Blei, D. Reading tea leaves: How humans interpret topic models. In: Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, CKI.; Culotta, A., editors. Advances in Neural Information Processing Systems. Vol. 22. 2009. p. 288-296.
23. Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R. Indexing by latent semantic analysis. J. Amer. Soc. Inform. Sci. 1990; 41(no. 6):391–407.
24. Du, L.; Ren, L.; Dunson, D.; Carin, L. A Bayesian model for simultaneous image clustering, annotation and object segmentation; Proc. Neural and Information Processing Systems (NIPS'09); 2009.
25. Erosheva E. Bayesian estimation of the grade of membership model. Bayesian Statistics. 2003; 7:501–510.
26. Erosheva E, Fienberg S, Joutard C. Describing disability through individual-level mixture models for multivariate binary data. Ann. Appl. Statist. 2007
27. Ewens W. The sampling theory of selective neutral alleles. Theor. Popul. Biol. 1972; 3:87–112. [PubMed: 4667078]
28. Fei-Fei, L.; Perona, P. A Bayesian hierarchical model for learning natural scene categories; Proc. IEEE Computer Vision and Pattern Recognition; 2005. p. 524-531.
29. Ferguson T. A Bayesian analysis of some nonparametric problems. Ann. Statist. 1973; 1(no. 2): 209–230.
30. Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. Bayesian Data Analysis. London, U.K.: Chapman & Hall; 1995.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

31. Gerrish, S.; Blei, D. A language-based approach to measuring scholarly impact; Proc. Int. Conf. Machine Learning; 2010.
32. Griffiths T, Steyvers M. Finding scientific topics. Proc. Nat. Acad. Sci. 2004; 101:5228–5235. [PubMed: 14872004]
33. Hoffman, M.; Blei, D.; Cook, P. Content-based musical similarity computation using the hierarchical Dirichlet process; Proc. Int. Conf. Music Information Retrieval; 2008.
34. Hoffman, M.; Blei, D.; Cook, P. Finding latent sources in recorded music with a shift-invariant HDP; Proc. Int. Conf. Digital Audio Effects; 2009.
35. Hofmann, T. Probabilistic latent semantic analysis; Proc. Uncertainty in Artificial Intelligence; 1999.
36. Kass R, Raftery A. Bayes factors. J. Amer. Statist. Assoc. 1995; 90(no. 430):773–795.
37. Li, J.; Socher, R.; Fei-Fei, L. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework; Proc. IEEE Conf. Computer Vision and Pattern Recognition; 2009.
38. McCallum, A. [Online]. Available: http://mallet.cs.umass.edu
39. Mimno, D.; McCallum, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression; Proc. Uncertainty in Artificial Intelligence; 2008.
40. Mimno, D. Reconstructing Pompeian households; Proc. Applications of Topic Models Workshop; 2009.
41. Muller, P.; Walker, S.; Hjort, N.; Holmes, C. Bayesian Nonparametrics. Cambridge, U.K.: Cambridge Univ. Press; 2010.
42. Nallapati, R.; Cohen, W. Link-pLSA-LDA: A new unsupervised model for topics and influence of blogs; Proc. Int. Conf. on Weblogs and Social Media (ICWSM'08); 2008.
43. Newman, D.; Asuncion, A.; Smyth, P.; Welling, M. Distributed inference for latent Dirichlet allocation; Proc. Neural Information Processing Systems (NIPS'07); 2007.
44. Newman, D.; Lau, JH.; Grieser, K.; Baldwin, T. Automatic evaluation of topic coherence; Proc. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT'10); 2010.
45. Pitman, J. Combinatorial Stochastic Processes. New York, NY: Springer-Verlag; 2002. (Lect. Notes for St. Flour Summer School)
46. Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000 Jun.155:945–959. [PubMed: 10835412]
47. Pruteanu-Malinici, I.; Ren, L.; Paisley, J.; Wang, E.; Carin, L. Hierarchical Bayesian modeling of topics in time-stamped documents; IEEE Trans. Pattern Anal. Machine Intell; 2009.
48. Rao, V.; Teh, YW. Spatial normalized gamma processes. In: Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, CKI.; Culotta, A., editors. Advances in Neural Information Processing Systems. Vol. 22. 2009. p. 1554-1562.
49. Ren L, Dunson D, Lindroth S, Carin L. Dynamic nonparametric Bayesian models for analysis of music. J. Amer. Statist. Assoc. 2009
50. Robert, C.; Casella, G. Monte Carlo Statistical Methods (Springer Texts in Statistics). New York, NY: Springer-Verlag; 2004.
51. Sivic, J.; Russell, B.; Efros, A.; Zisserman, A.; Freeman, W. CSAIL. MIT, Tech. Rep.; 2005. Discovering object categories in image collections.
52. Steyvers, M.; Griffiths, T. Probabilistic topic models. In: Landauer, T.; McNamara, D.; Dennis, S.; Kintsch, W., editors. Latent Semantic Analysis: A Road to Meaning. Lawrence Erlbaum; 2006.
53. Sudderth, EB.; Jordan, MI. Shared segmentation of natural scenes using dependent Pitman-Yor processes; Proc. Neural Information Processing Systems (NIPS'08); 2008.
54. Teh Y, Jordan M, Beal M, Blei D. Hierarchical Dirichlet processes. J. Acoust. Soc. Amer. 2005; 101(no. 476):1566–1582.
55. Teh, Y.; Kurihara, K.; Welling, M. Collapsed variational inference for HDP; Proc. Neural Information Processing Systems (NIPS'07); 2007.
56. Teh Y, Newman D, Welling M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. Proc. Neural Information Processing Systems (NIPS'06). 2006

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

57. Wang, C.; Blei, D.; Fei-Fei, L. Simultaneous image classification and annotation; Proc. IEEE Conf. Computer Vision and Pattern Recognition; 2009.
58. Wang, C.; Blei, D.; Heckerman, D. Continuous time dynamic topic models; Proc. Uncertainty in Artificial Intelligence (UAI); 2008.
59. Wang, X.; Grimson, E. Spatial latent Dirichlet allocation; Proc. Neural Information Processing Systems (NIPS'07); 2007.
60. Wang, X.; Ma, X.; Grimson, E. Unsupervised activity perception by hierarchical Bayesian models; Proc. IEEE Conf. Computer Vision and Pattern Recognition; 2007 Jun. p. 1-8.
61. Wang X, Ma X, Grimson E. Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models. IEEE Trans. Pattern Anal. Machine Intell. 2009
62. Wang, X.; McCallum, A. Topics over time: A non-Markov continuoustime model of topical trends; Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining; 2006. p. 424-433.
63. Welling, M.; Teh, Y.; Kappen, B. Hybrid variational/Gibbs collapsed inference in topic models; Proc. Conf. Uncertainty in Artifical Intelligence (UAI'08); 2008. p. 587-594.Citeseer
64. West, M.; Harrison, J. Proc. Int. Conf. Machine Learning. New York: Springer-Verlag; 1997. Bayesian forecasting and dynamic models.
65. Yakhnenko, O.; Honavar, V. Multi-modal hierarchical Dirichlet process model for predicting image annotation and image-object label correspondence; Proc. SIAM Conf. Data Mining (SDM'09); 2009.
66. Yan, F.; Xu, N.; Qi, Y. Parallel inference for latent Dirichlet allocation on graphics processing units. In: Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, CKI.; Culotta, A., editors. Advances in Neural Information Processing Systems. Vol. 22. 2009. p. 2134-2142.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

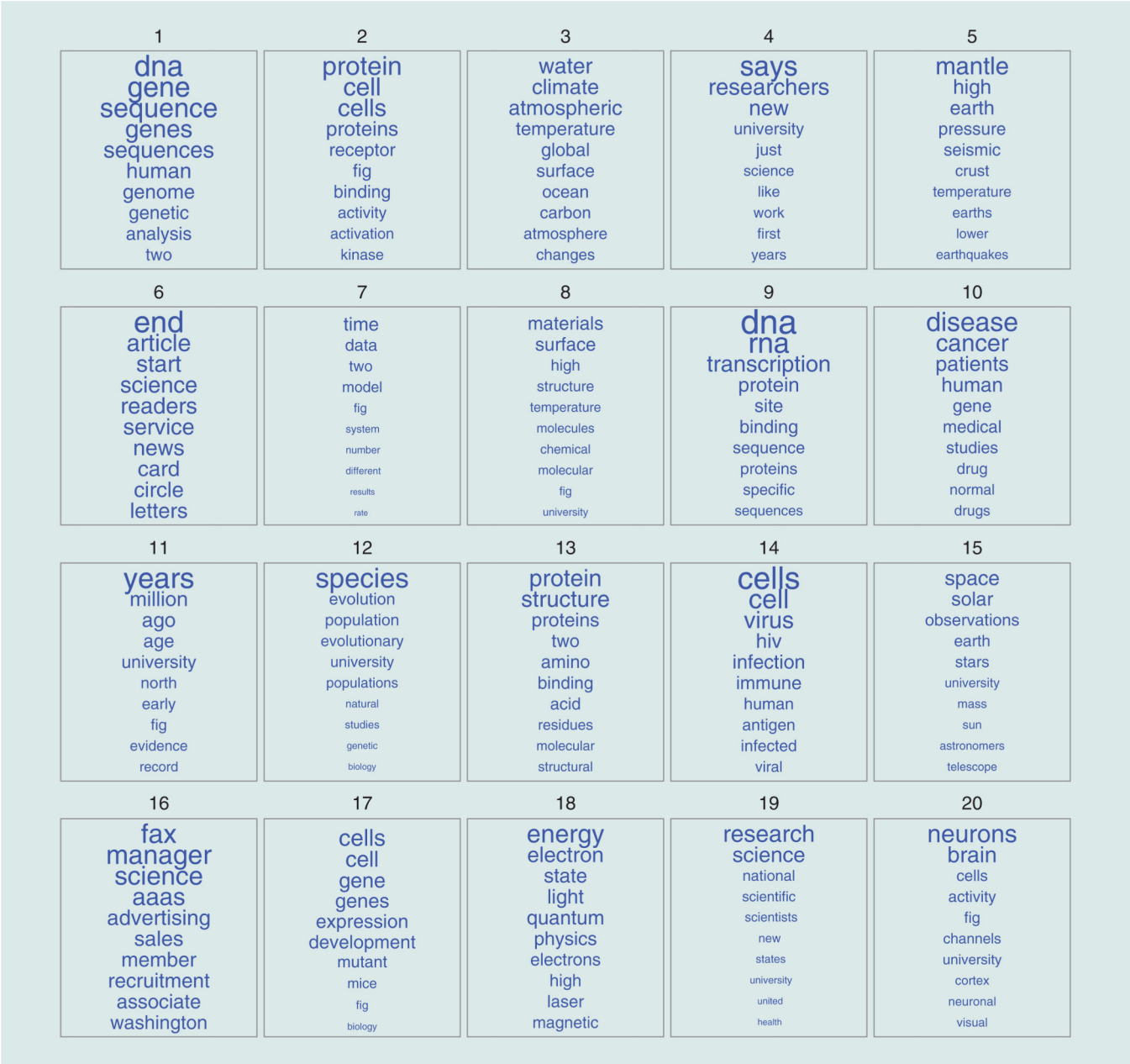

FIG1.

Top words from all 20 topics of a 20-topic LDA model. This posterior was approximated from 17,000 articles from *Science* (all the articles published in the 1990s). The size of each term is proportional to its probability in the topic. This LDA posterior was approximated with Gibbs sampling using open source software [20].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

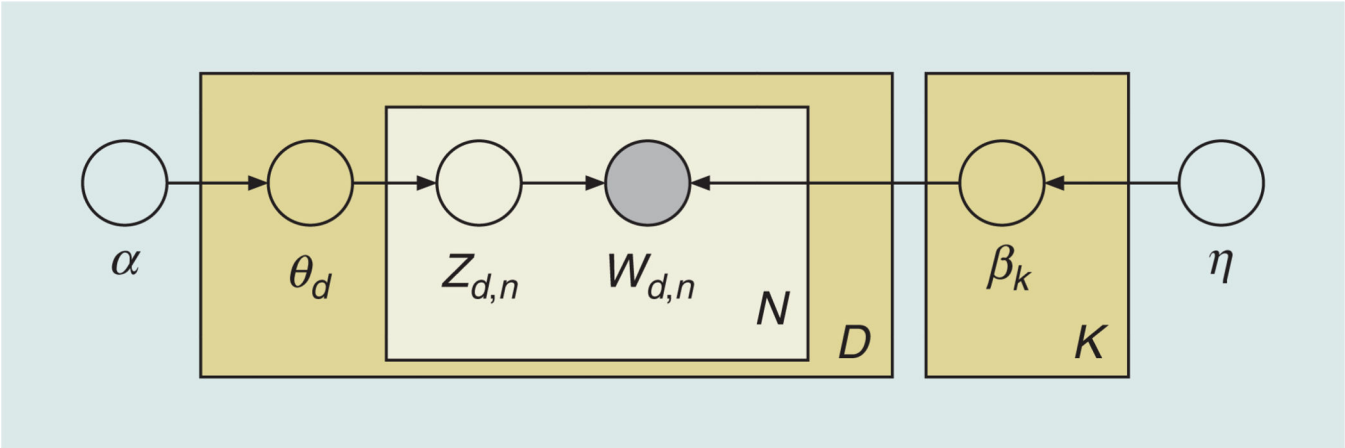

FIG2.
The graphical model representation of LDA [15].

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

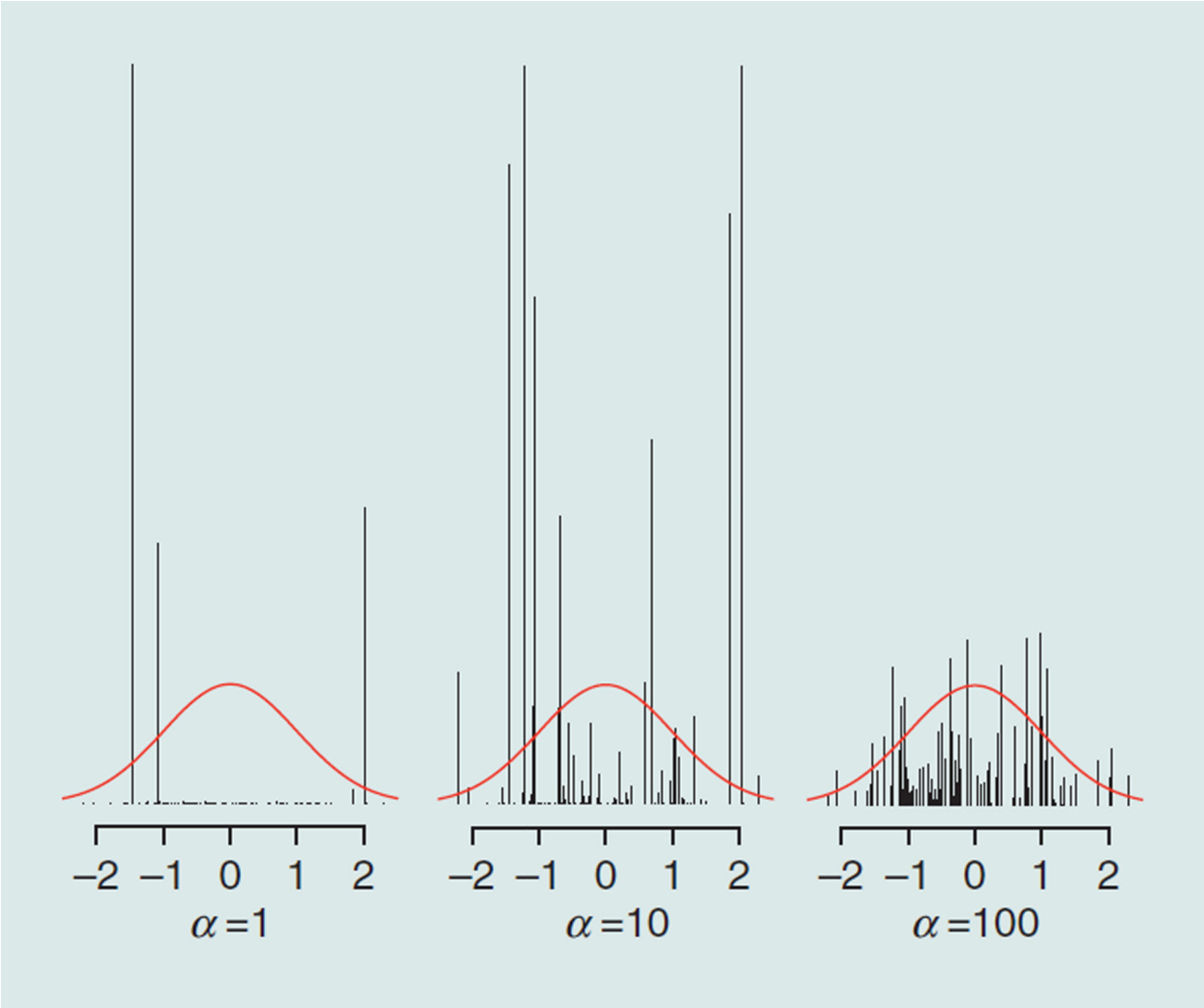

FIG3.

Three draws from a DP with standard normal base distribution. Draws from the DP are discrete; as α increases, the resulting random distribution looks more like the base distribution.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

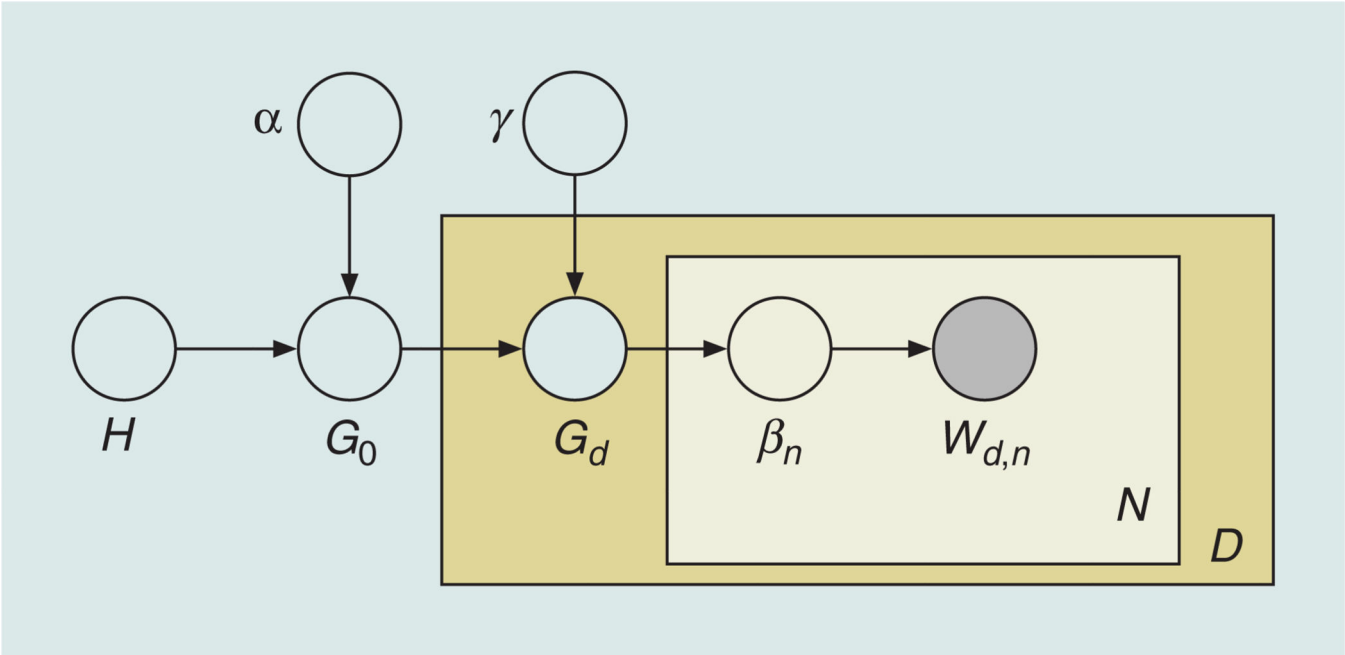

FIG4.
The graphical model representation of the HDP topic model [54].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

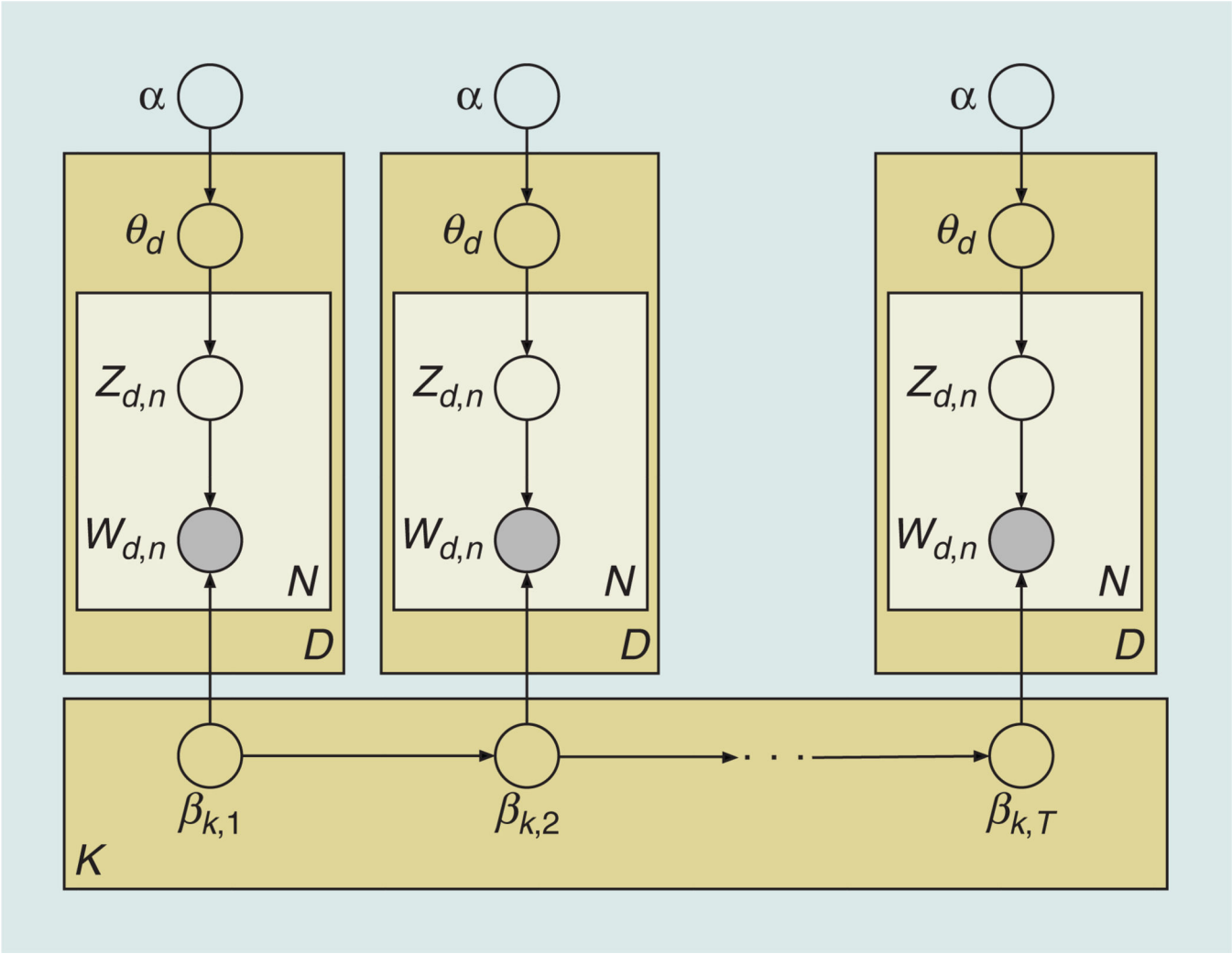

FIG5.
The graphical model representation of the DTM [12].

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

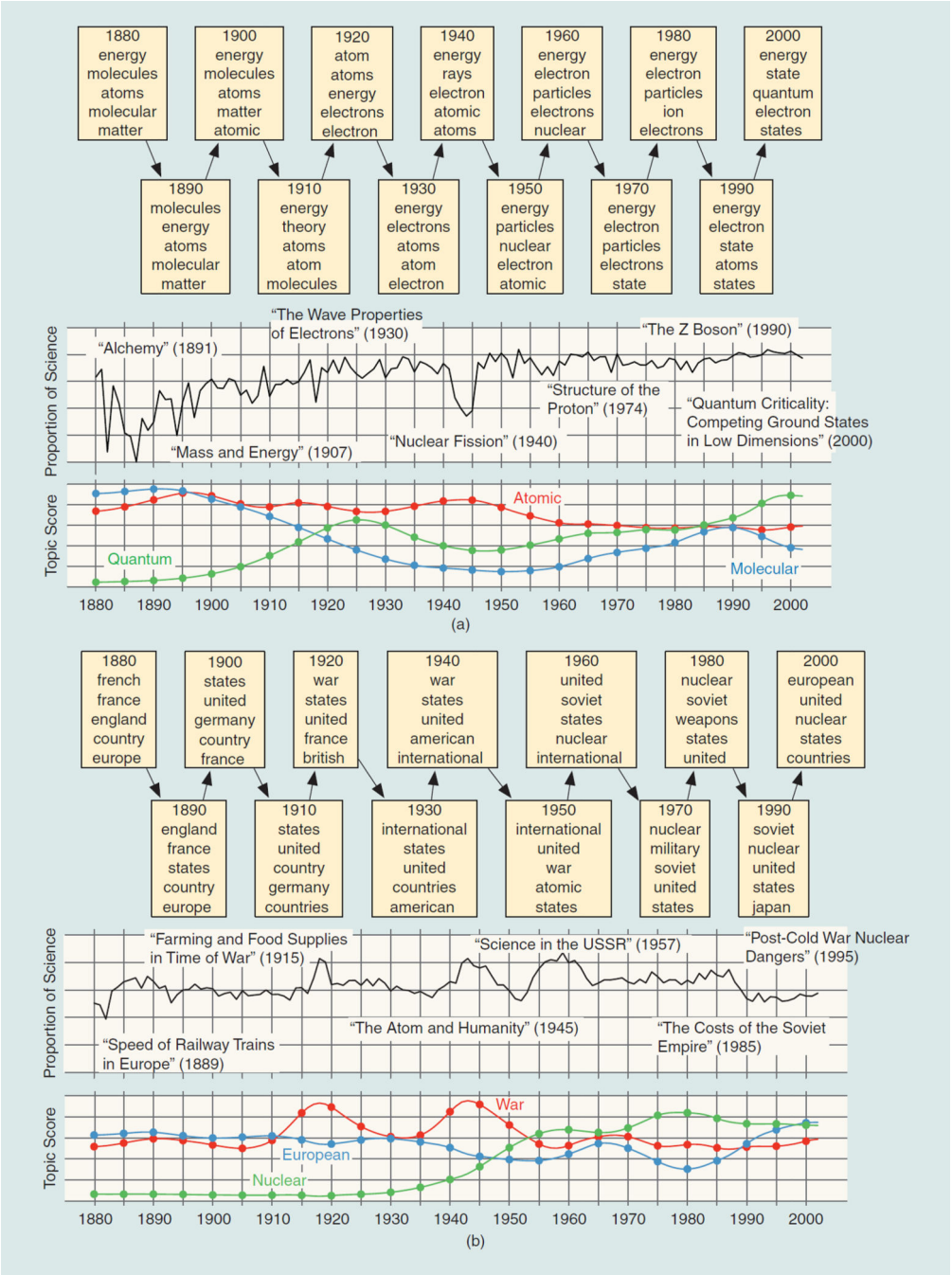

FIG6.

Two example topics from a DTM fit to the *Science* archive (1880–2002). Note that the model posited an epoch at each year of publication. We have illustrated the top words at each decade. This figure is reprinted with permission from [13].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

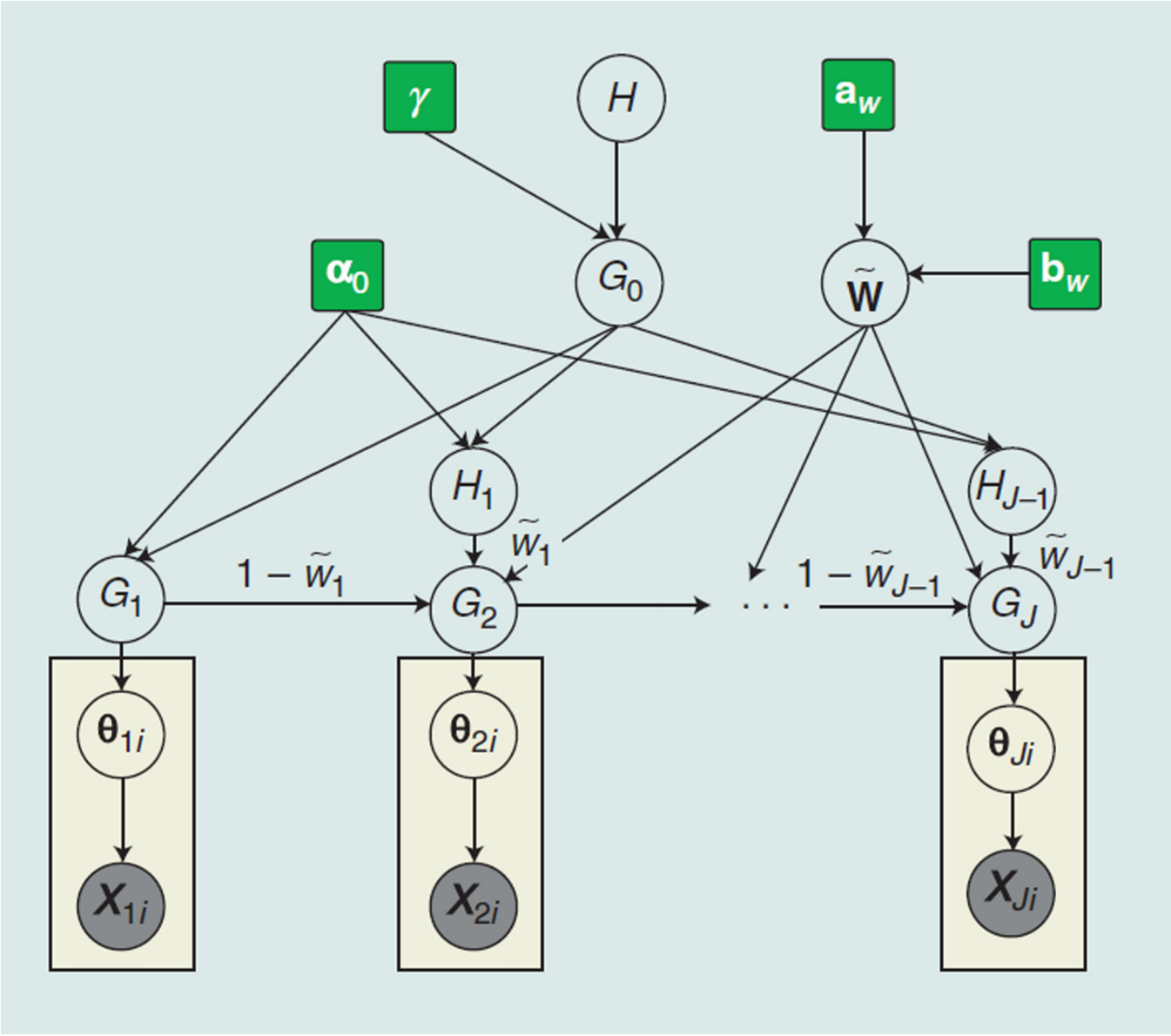

FIG7.
A graphical model of the dHDP model.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

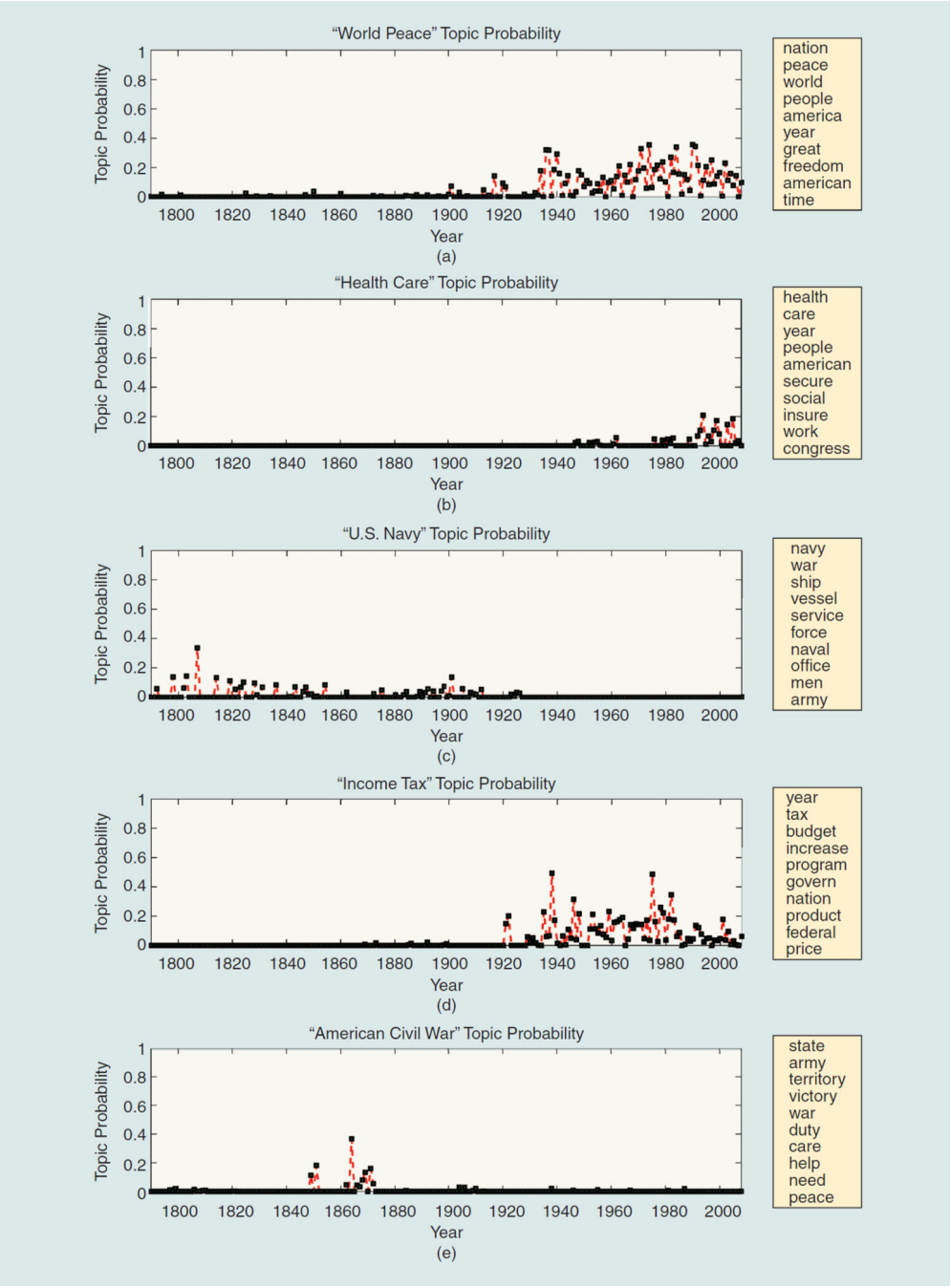

FIG8.

Topics inferred via dHDP analysis of the presidential State of the Union addresses in the United States, from 1790–2008. Each plot depicts topic probability as a function of year, and the most probable words in each topic are shown at right. Notional topics are (a) world peace, (b) health care, (c) U.S. Navy, (d) income taxes, and (e) American Civil War. This figure is reprinted with permission from [47].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

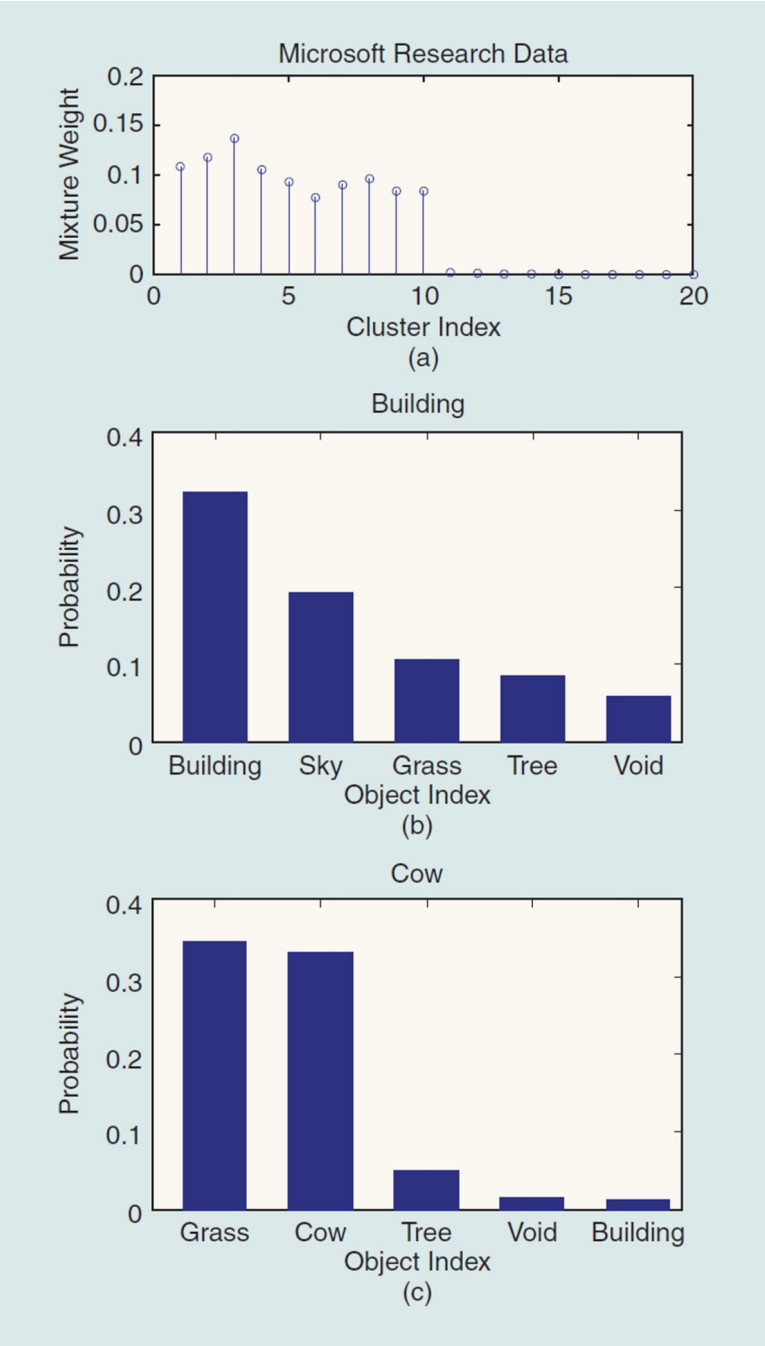

FIG9.

Example inferred latent properties associated with Microsoft data set. (a) Posterior distribution on the mixtureweights, quantifying the probability of scene topics (ten topics are inferred). Parts (b) and (c) show the example probability of objects for a given class of images, or topic (probability of object/ words); here we only give the top five words for each topic. This figure is reprinted with permission from [24].

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

FIG10.

Example segmentation and labeling results. (a) Original images; (b) Corr-LDA [10]; (c) with segmentation of image words performed with DP; (d) segmentation of the visual words accounts for spatial location in the image LSBP [24]). Columns 1–3 from Microsoft data set; Columns 4–6 from UIUC-Sport data set. All of these examples were not annotated originally when performing learning. This figure is reprinted with permission from [24].