PROBABILISTIC MODELS OF TEXT AND IMAGES

by

David Meir Blei

B.S. (Brown University), 1997

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

COMPUTER SCIENCE

with a designated emphasis in

COMMUNICATION, COMPUTATION, AND STATISTICS

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Prof. Michael I. Jordan, Chair
Prof. Stuart J. Russell
Prof. Peter Bickel

Fall 2004

The dissertation of David Meir Blei is approved:

Chair

Date

Date

Date

University of California, Berkeley

Fall 2004

Probabilistic models of text and images

by

David Meir Blei

Abstract

Probabilistic models of text and images

by

David Meir Blei

Doctor of Philosophy in Computer Science

with a designated emphasis in

Communication, Computation, and Statistics

University of California, Berkeley

Prof. Michael I. Jordan, Chair

Managing large and growing collections of information is a central goal of modern computer science. Data repositories of texts, images, sounds, and genetic information have become widely accessible, thus necessitating good methods of retrieval, organization, and exploration. In this thesis, we describe a suite of probabilistic models of information collections for which the above problems can be cast as statistical queries.

We use directed graphical models as a flexible, modular framework for describing appropriate modeling assumptions about the data. Fast approximate posterior inference algorithms based on variational methods free us from having to specify tractable models, and further allow us to take the Bayesian perspective, even in the face of large datasets.

With this framework in hand, we describe latent Dirichlet allocation (LDA), a graphical model particularly suited to analyzing text collections. LDA posits a finite index of hidden topics which describe the underlying documents. New documents are situated into the collection via approximate posterior inference of their associated index terms. Extensions to LDA can index a set of images, or multimedia collections of interrelated text and images.

Finally, we describe nonparametric Bayesian methods for relaxing the assumption of a fixed number of topics, and develop models based on the natural assumption that the size of the index can grow with the collection. This idea is extended to trees, and

to models which represent the hidden structure and content of a topic hierarchy that underlies a collection.

Acknowledgements

This dissertation would not have been possible without the help and support of many friends and colleagues.

Foremost, I thank my advisor Michael I. Jordan. For the past five years, Mike has been an exemplary teacher, mentor, and collaborator. I also thank the dissertation committee, Peter Bickel and Stuart J. Russell, for their insightful comments and suggestions.

I am fortunate to have interacted with a number of superb colleagues, and I would like to recognize their contribution to this work. Francis Bach, Drew Bagnell, Kobus Barnard, Jaety Edwards, Barbara Engelhardt, David Forsyth, Thomas Griffiths, Marti Hearst, Leslie Kaelbling, John Lafferty, Gert Lanckriet, Jon McAuliffe, Andrew McCallum, Brian Milch, Pedro Moreno, Andrew Ng, Mark Paskin, Sam Roweis, Martin Wainwright, Alice Zheng, and Andrew Zimdars have all been influential to my research through collaboration, discussion, and constructive criticism. I particularly thank Andrew Ng, who effectively launched this line of work when he sketched the picture in Figure 3.2 on an envelope one afternoon.

My comfortable life in Berkeley would not have been possible without the generous financial support of a fellowship from the Microsoft corporation. Other financial support came from the Berkeley Microelectronics Fellowship, and travel grants from the NIPS foundation, UAI society, and SIGIR foundation. I also thank Microsoft Research, Whizbang! Labs, and Compaq Research for hosting three excellent and enriching summer internships.

I thank all my friends and family for their support and distraction. I especially thank my parents Ron and Judy Blei and sister Micaela Blei who have given me a lifetime of love and care. Finally, I thank Toni Gantz. Her kindness, patience, humor, and friendship sustain me.

Dedicated to my parents, Ron and Judy Blei.

Contents

Chapter 1

Introduction

The management of large and growing collections of information is a central goal of modern computer science. Data repositories of texts, images, sounds, and genetic information have become widely accessible, thus necessitating good methods of retrieval, organization, and exploration. In 1945, Vannevar Bush predicted the existence of such collections, and anticipated the subsequent challenge. "There may be millions of fine thoughts...all encased within stone walls of acceptable architectural form; but if the scholar can get at only one a week by diligent search, his syntheses are not likely to keep up with the current scene" (Bush, 1945).

Probabilistic models have been paramount to these tasks, used in settings such as speech recognition (Rabiner, 1989), text classification (Pietra et al., 1997; Nigam et al., 1999, 2000), information retrieval (Ponte and Croft, 1998), text segmentation (Beeferman et al., 1999; Blei and Moreno, 2001), information extraction (Lafferty et al., 2001; Blei et al., 2002), collaborative filtering (Popescul et al., 2001; Marlin, 2004), and citation analysis (Taskar et al., 2001; Pasula et al., 2002). These methods entail two stages: (1) estimate or compute the posterior distribution of the parameters of a probabilistic model from a collection of text; (2) for new documents, answer the question at hand (e.g., classification, retrieval, speech recognition) via probabilistic inference.

The goal of such modeling is document *generalization*. Given a new document, how is it similar to the previously seen documents? Where does it fit within them?

What can one predict about it? Efficiently answering such questions is the focus of the statistical analysis of document collections.

Returning to Bush, he dubs a compact storage device for document collections as a *memex*, and speculates on the problem of interacting with it. "Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing." His solution is "*associative indexing*, the basic idea of which is a provision whereby any item may be caused at will to select immediately and automatically another. This is the essential feature of the memex. The process of tying two items together is the important thing."

In this thesis, we develop probabilistic models of collections which represent the underlying hidden patterns of meaning that connect the documents in a network of associations. In statistical terms, we develop *latent variable models*, where the latent variables describe the hidden semantic index to which Bush alludes. Document generalization amounts to probabilistic inference of its constituent indexing terms, and thus a probabilistic relationship to the rest of the collection. Moreover, we allow the term "document" to refer more generally to any data that fits in to a hidden index. Consequently, we develop models appropriate to multimedia collections, such as text and images, with which we can further infer relationships between the media types.

Thus, the contribution of this thesis is a principled development of appropriate statistical models for information collections, and corresponding probabilistic inference algorithms which can accommodate large datasets. The chapters are organized as follows:

- In Chapter 2, we review directed graphical models, exponential family distributions, mixture models, and approximate posterior inference. We derive general forms of Gibbs sampling and variational inference algorithms for approximate posterior inference in the class of graphical models with conditional exponential family distributions.

- In Chapter 3, we develop the *latent Dirichlet allocation* (LDA) model. This

model is central to the rest of the thesis, and reflects a principled approach to document modeling based on exchangeability assumptions and the notion of a hidden semantic index. We present results in language modeling, document classification, and collaborative filtering.

- In Chapter 4, we extend LDA to images, demonstrating that this type of modeling applies to information sources beyond text documents. Moreover, we develop a model of *annotated data*, where two different data-types are associated by the underlying latent factors. This model is applied to image/caption modeling, automatic image annotation, and text-based image retrieval.

- In Chapter 5, we address the problem of having to choose a fixed number of factors (i.e., the size of the index) when using the methods of the previous chapters. We review *Dirichlet process mixture models* as a flexible solution which allows new data to exhibit new factors. We develop mean-field variational inference for such models and provide an empirical comparison to Gibbs sampling.

- In Chapter 6, we develop an extension of the Dirichlet process called the *nested Dirichlet process*, which provides a distribution over hierarchical partitions. We use this distribution as a component in a flexible, structured model of documents, amounting to a culmination of the ideas in the previous chapters. We analyze collections scientific abstracts with this model, automatically discovering the underlying hierarchy of topics which describes a scientific field.

- Finally, in Chapter 7, we summarize the ideas of the thesis and point to directions of future work.

Chapter 2

Graphical models and approximate posterior inference

In this chapter we review latent variable graphical models and exponential families. We discuss variational methods and Gibbs sampling for approximate posterior inference, and derive general forms of these algorithms for a large subclass of models.

2.1 Latent variable graphical models

We use the formalism of *directed graphical models* to describe the independence assumptions of the models developed in the subsequent chapters. A directed graphical model provides a succinct description of the factorization of a joint distribution: *nodes* denote random variables; *edges* denote possible dependence between random variables; and *plates* denote replication of a substructure, with appropriate indexing of the relevant variables.

Graphical models can be used to describe *latent variable models*. Latent variable modeling is a method of developing complicated structured distributions, where the data interact with *latent* or *unobserved* random variables. In the graphical model notation, observed random variables are shaded, and latent random variables are unshaded.

For example, the distribution on the real line in Figure 2.1 (Left) is the *mixture*

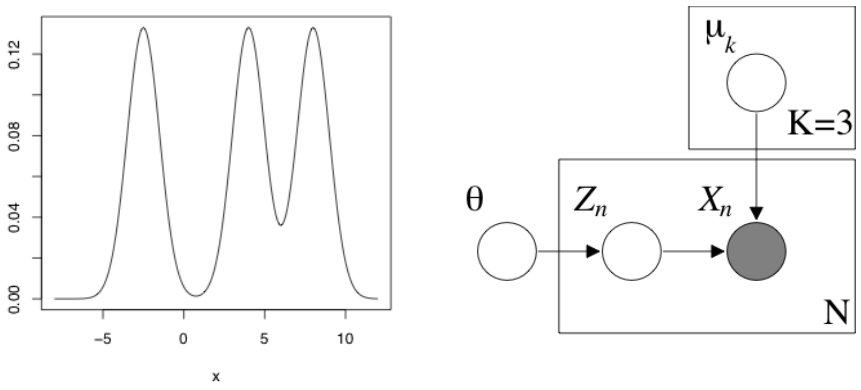


Figure 2.1: (Left) The density of a Gaussian mixture model with three mixture components. (Right) The corresponding graphical model of N data from this density.

distribution formed by combining three unit-variance Gaussian distributions with means $\mu_1 = -2.5$, $\mu_2 = 4$, and $\mu_3 = 8$. A data point is drawn by first choosing a latent variable $Z \in \{1, 2, 3\}$ from a multinomial, and then drawing the data point from $\mathcal{N}(\mu_z, 1)$. This example is illustrated as a graphical model in Figure 2.1 (Right).

The central task of latent variable modeling for data analysis is *posterior inference*, where we determine the distribution of the latent variables conditional on the observations. Loosely, posterior inference can be thought of as a reversal of the generative process which the graphical model illustrates. For example, in the Gaussian mixture with fixed means, we would like to determine the posterior distribution of the indicator Z given a data point x. If $x = 1$, then the posterior $p(Z \mid X = 1, \mu_1, \mu_2, \mu_3)$ is $(0.16, 0.83, 0.01)$.

Traditionally, the structure of the graphical model informs the ease or difficulty of posterior inference. In the models of the subsequent chapters, however, inference is difficult despite a simple graph structure. Thus, we resort to approximate posterior inference, which is the subject of Section 2.2.

Typically, the parameters of the model are not observed (e.g., the means in the Gaussian mixture), and part of the posterior inference problem is to compute their posterior distribution conditional on the data. One option is to adopt the *empirical Bayes* perspective (Morris, 1983; Kass and Steffey, 1989; Maritz and Lwin, 1989),

and find point estimates of the parameters based on maximum likelihood. Such estimates can be found, for example, with the expectation-maximization (EM) algorithm (Dempster et al., 1977), or approximate variant of it (Neal and Hinton, 1999).

Alternatively, we may take a more fully Bayesian approach, placing a prior distribution on the parameters and computing a proper posterior distribution. This is called *hierarchical Bayesian modeling* (Gelman et al., 1995) because it necessitates the specification of a distribution of the parameters, which itself must have parameters called *hyperparameters*.

In a hierarchical Bayesian model, we may still use the empirical Bayes methodology, and find point estimates of the hyperparameters by maximum likelihood. This is often sensible because it affords the advantages of exhibiting uncertainty on the parameters, while avoiding the unpleasant necessity of choosing a fixed hyperparameter or further extending the hierarchy.

2.1.1 Exponential families

All the random variables we will consider are distributed according to *exponential family* distributions. This family of distributions has the form:

$$p(x \,|\, \eta) = h(x) \exp\{\eta^T t(x) - a(\eta)\}, \tag{2.1}$$

where η is the *natural parameter*, $t(x)$ are the *sufficient statistics* for η, and $a(\eta)$ is the *cumulant generating function* or *log partition function*:

$$a(\eta) = \log \int h(x) \exp\{\eta^T t(x)\} dx. \tag{2.2}$$

The derivatives of $a(\eta)$ are the cumulants of $t(x)$. In particular, the first two derivatives are:

$$a'(\eta) = \mathrm{E}_\eta \left[t(X) \right] \tag{2.3}$$

$$a''(\eta) = \mathrm{Var}_\eta \left[t(X) \right]. \tag{2.4}$$

The functions $a(\eta)$ and $h(x)$ are determined by the form and dimension of $t(x)$. For example, if x is real valued and $t(x) = (x, x^2)$, then the corresponding exponential

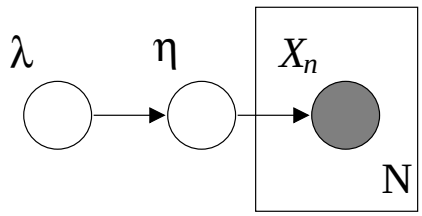


Figure 2.2: Graphical model representation of iid data $X_{1:N}$ from $p(x \mid \eta)$, where η is itself distributed by $p(\eta \mid \lambda)$ for a fixed hyperparameter λ. Computation in this model is facilitated when X_n is in the exponential family, and η is distributed by the conjugate prior.

family is Gaussian. If $t(x)$ is a multidimensional vector with all zeros and a single one, then the corresponding exponential family distribution is multinomial. An exponential family for positive reals is the Gamma distribution, and an exponential family for positive integers is the Poisson distribution.

See Brown (1986) for a thorough analysis of the properties of exponential family distributions.

2.1.2 Conjugate exponential families

In a hierarchical Bayesian model, we must specify a prior distribution of the parameters. In this section, we describe a family of priors which facilitate computations in such a model.

Let X be a random variable distributed according to an exponential family with natural parameter η and log normalizer $a(\eta)$. A *conjugate prior* of η, with natural parameter λ, has the form:

$$p(\eta \mid \lambda) = h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a(\eta)) - a(\lambda)\}.$$

The parameter λ has dimension $\dim(\eta) + 1$ and the sufficient statistic is $t(\eta) = (\eta, -a(\eta))$. We decompose $lambda = (\lambda_1, \lambda_2)$ such that λ_1 contains the first $\dim(\eta)$ components and λ_2 is a scalar. (Note that we overload $a(\cdot)$ to be the log normalizer for the parameter in the argument.)

The conjugate distribution is a convenient choice of prior, because the corresponding posterior will have the same form. Consider the simple model illustrated in Figure 2.2 where $X_{1:N}$ are independent and identically distributed (iid) variables from the exponential family distribution $p(x_n \,|\, \eta)$, and $p(\eta \,|\, \lambda)$ is the conjugate prior. The posterior of η is:

$$\begin{aligned} p(\eta \,|\, \lambda, x_{1:N}) &\propto p(\eta \,|\, \lambda) p(x_{1:N} \,|\, \eta) \\ &\propto h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a(\eta))\} \prod_{n=1}^{N} \exp\{\eta^T t(x_n) - a(\eta)\} \\ &= h(\eta) \exp\{(\lambda_1 + \textstyle\sum_{n=1}^{N} t(x_n))^T \eta + (\lambda_2 + N)(-a(\eta))\}, \end{aligned} \tag{2.5}$$

which is the same type of distribution as $p(\eta \,|\, \lambda)$, with posterior parameters $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)$:

$$\begin{aligned} \hat{\lambda}_1 &= \lambda_1 + \textstyle\sum_{n=1}^{N} t(x_n) \\ \hat{\lambda}_2 &= \lambda_2 + N. \end{aligned} \tag{2.6}$$

The posterior, conditional on any amount of data, can be fully specified by the prior parameters, the sum of the sufficient statistics, and the number of data points.

A second convenience of the conjugate prior is for computing the marginal distribution $p(x \,|\, \lambda) = \int p(x \,|\, \eta) p(\eta \,|\, \lambda) d\eta$. If $p(\eta \,|\, \lambda)$ is conjugate, then:

$$\begin{aligned} p(x \,|\, \lambda) &= h(x) \int \exp\{\eta^T t(x) - a(\eta)\} h(\eta) \exp\{\lambda_1^T \eta + \lambda_2(-a(\eta)) - a(\lambda)\} d\eta \\ &= h(x) \int h(\eta) \exp\{(\lambda_1 + t(x))^T \eta + (\lambda_2 + 1)(-a(\eta))\} d\eta \exp\{-a(\lambda)\} \\ &= h(x) \exp\{a((\lambda_1 + t(x), \lambda_2 + 1)) - a(\lambda)\}. \end{aligned} \tag{2.7}$$

Thus, if the log normalizer is easy to compute then the marginal distribution will also be easy to compute.

Finally, the conjugate prior facilitates computing the *predictive distribution* $p(x \,|\, x_{1:N}, \lambda)$, which is simply a marginal under the posterior parameters in Eq. (2.6).

Example: Gaussian with Gaussian prior on the mean

Suppose the data are real vectors distributed according to a Gaussian distribution with fixed inverse covariance Λ. The exponential family form of the data density is:

$$p(x \,|\, \eta) = \exp\left\{-\frac{1}{2}(d\log 2\pi - \log|\Lambda| + \eta^T\Lambda^{-1}\eta) + \eta^T x - \frac{x^T\Lambda x}{2}\right\},$$

where:

$$\begin{aligned} h(x) &= \exp\left\{-\frac{1}{2}(d\log 2\pi - \log\Lambda)\right\} \\ a(\eta) &= -\eta^T\Lambda^{-1}\eta. \end{aligned}$$

The conjugate prior is thus of the form:

$$p(\eta \,|\, \lambda) \propto \exp\left\{\lambda_1^T\eta - \lambda_2\left(\frac{\eta^T\Lambda^{-1}\eta}{2}\right)\right\},$$

which is a Gaussian with natural parameters λ_1 and $\lambda_2\Lambda^{-1}$. Note that its covariance is the scaled inverse covariance of the data $\frac{\Lambda}{\lambda_2}$. The log normalization is:

$$a(\lambda) = -\frac{1}{2}\log|\lambda_2\Lambda^{-1}| + \frac{\lambda_1^T\Lambda\lambda_1}{\lambda_2},$$

from which we can compute the expected sufficient statistics of η:

$$\begin{aligned} \mathrm{E}[\eta] &= \frac{(\Lambda + \Lambda^T)\lambda_1}{\lambda_2} \\ \mathrm{E}[-a(\eta)] &= \frac{d}{\lambda_2} - \frac{\lambda_1^T\Lambda\lambda_1}{\lambda_2^2}. \end{aligned}$$

2.1.3 Exponential family conditionals

Conditional on all the other variables in a directed graphical model, the distribution of a particular variable depends only on its *Markov blanket*, which is the set containing its parents, children, and other parents of its children. To facilitate approximate posterior inference, we consider models for which the conditional distribution of every node given its Markov blanket is in an exponential family.

One possible substructure which meets this requirement is the conjugate-exponential family model of Figure 2.2. Conditional on η, the distribution of X_n is in an exponential family. Moreover, as we have shown above, the conditional distribution of $\eta \,|\, \{\lambda, x_{1:N}\}$ is also in an exponential family.

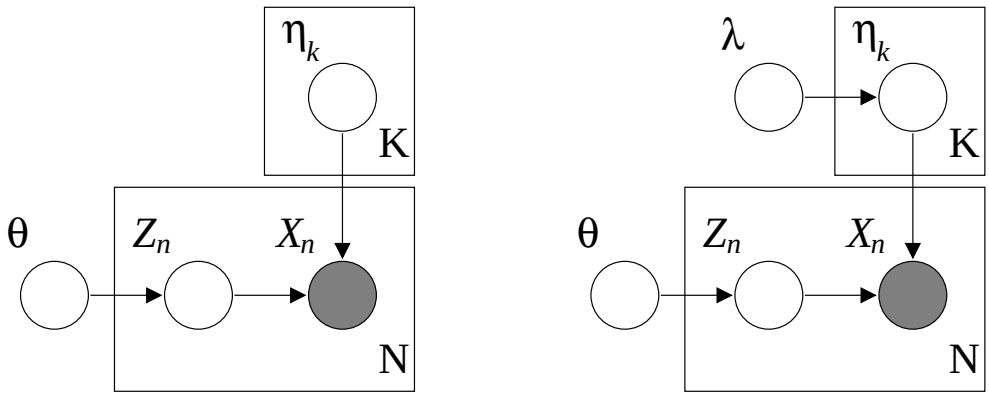


Figure 2.3: (Left) Graphical model representation of a K-mixture model. (Right) A Bayesian K-mixture model.

A second possibility is for the distribution of a variable to be a *mixture* of exponential family distributions. This important substructure is illustrated in Figure 2.3 (Left), where $\eta_{1:K}$ are exponential family parameters and θ is a K-dimensional multinomial parameter. The variables $X_{1:N}$ can be thought of as drawn from a two-stage generative process: first, choose Z_n from Mult(θ); then, choose X_n from the distribution indexed by that value $p(x_n \mid \eta_{z_n})$.

Note that we represent multinomial variables using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, using superscripts to denote components, the kth item is represented by a K-vector z such that $z^k = 1$ and $z^\ell = 0$ for $\ell \neq k$.

We confirm the conditional distributions of nodes X_n and Z_n, given their respective Markov blankets, are in the exponential family. First, by definition, the conditional distribution $p(x_n \mid z_n)$ is a member of the η-indexed exponential family. Second, the conditional distribution $p(z_n \mid x_n)$ is a multinomial:

$$p(z_n \mid x_n, \theta, \eta_{1:K}) \propto p(z_n \mid \theta) p(x_n \mid z_n, \eta_{1:K}),$$

which is also in the exponential family.

In the hierarchical mixture model of Figure 2.3 (Right), we can place the conjugate prior on $\eta_{1:K}$. The distribution of $\eta_k \mid \{z_{1:N}, x_{1:N}\}$ remains in the exponential family as a consequence of the analysis in Eq. (2.5). In particular, we condition only on

those x_n for which $z_n^k = 1$.

By combining the substructures described above, we can build complicated families of distributions which satisfy the requirement that each node, conditional on its Markov blanket, is distributed according to an exponential family distribution. This collection of families contains many probabilistic models, including Markov random fields, Kalman filters, hidden Markov models, mixture models, and hierarchical Bayesian models with conjugate and mixture of conjugate priors.

2.2 Approximate posterior inference

The central computational challenge in latent variable modeling is to compute the posterior distribution of the latent variables conditional on some observations. Except in rudimentary models, such as Figures 2.2 and 2.3, exact posterior inference is intractable and practical data analysis relies on good approximate alternatives. In this section, we describe two general techniques for the class of graphical models which satisfy the conditional exponential family restriction described above.

In the following, we consider a latent variable probabilistic model with parameters η, observed variables $\mathbf{x} = x_{1:N}$ and latent variables $\mathbf{Z} = Z_{1:M}$. The posterior distribution of the latent variables is:

$$p(z_{1:M} \,|\, x_{1:N}, \eta) = \frac{p(x_{1:N}, z_{1:M} \,|\, \eta)}{\int p(x_{1:N}, z_{1:M} \,|\, \eta) dz_{1:M}}.$$

Under the assumptions, the numerator is in the exponential family and should be easy to compute. The denominator, however, is often intractable due to the nature of $z_{1:M}$. For example, if the latent variables are realizations of one of K values, then this integral is a sum over K^M possibilities. (E.g., this is true for the hierarchical mixture model of Figure 2.3 Right.)

2.2.1 Gibbs sampling

Markov chain Monte Carlo (MCMC) sampling is the most widely used method of approximate inference. The idea behind MCMC is to approximate a distribution by

forming an empirical estimate from samples. We construct a Markov chain with the appropriate stationary distribution, and collect the samples from a chain which has converged or "burned in".

The simplest MCMC algorithm is the *Gibbs sampler*, in which the Markov chain is defined by iteratively sampling each variable conditional on the most recently sampled values of the other variables. This is a form of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), and thus yields a chain with the desired stationary distribution (Geman and Geman, 1984; Gelfand and Smith, 1990; Neal, 1993).

In approximate posterior inference, the distribution of interest is the posterior $p(\mathbf{z} \mid \mathbf{x}, \eta)$. Thus, an iteration of the Gibbs sampler draws each latent variable z_i from $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$. After the resulting chain has converged, we collect B samples $\{\mathbf{z}_1, \ldots, \mathbf{z}_B\}$ and approximate the posterior with an empirical distribution:

$$p(\mathbf{z} \mid \mathbf{x}, \eta) = \frac{1}{B} \sum_{b=1}^{B} \delta_{\mathbf{z}_b}(\mathbf{z}).$$

This use of Gibbs sampling has revolutionized hierarchical Bayesian modeling (Gelfand and Smith, 1990).

In the models described in Section 2.1.3, the every variable, conditional on its Markov blanket, is distributed according to an exponential family distribution. Gibbs sampling in this setting is thus straightforward, provided we can easily compute the conditional exponential family parameter for each variable.[1]

2.2.2 Mean-field variational methods

Variational inference provides an alternative, deterministic methodology for approximating likelihoods and posteriors in an intractable probabilistic model (Wainwright and Jordan, 2003). We first review the basic idea in the context of the exponential family of distributions, and then turn to its application to approximating a posterior.

[1] In fact, Gibbs sampling is so straightforward in this case, that one can automatically generate Gibbs sampling code from a graph structure and parameterization using the popular BUGS package (Gilks et al., 1996).

Consider the η-indexed exponential family distribution in Eq. (**??**) and recall the cumulant generating function $a(\eta)$:

$$a(\eta) = \log \int \exp\{\eta^T t(z)\} h(z) dz.$$

As discussed by Wainwright and Jordan (2003), this quantity can be expressed variationally as:

$$a(\eta) = \sup_{\mu \in \mathcal{M}} \{\eta^T \mu - a^*(\mu)\}, \tag{2.8}$$

where $a^*(\mu)$ is the Fenchel-Legendre conjugate of $a(\eta)$ (Rockafellar, 1970), and $\mathcal{M}$ is the set of *realizable expected sufficient statistics*: $\mathcal{M} = \{\mu : \mu = \int t(z) p(z) h(z) dz, \text{for some } p\}$. There is a one-to-one mapping between parameters η and the interior of $\mathcal{M}$ (Brown, 1986). Accordingly, the interior of $\mathcal{M}$ is often referred to as the set of *mean parameters*.

Let $\eta(\mu)$ be a natural parameter corresponding to the mean parameter $\mu \in \mathcal{M}$; thus $\mathrm{E}_\eta [t(Z)] = \mu$. Let $q(z \,|\, \eta(\mu))$ denote the corresponding density. Given $\mu \in \mathcal{M}$, a short calculation shows that $a^*(\mu)$ is the negative entropy of q:

$$a^*(\mu) = \mathrm{E}_{\eta(\mu)} \left[\log q(Z \,|\, \eta(\mu))\right]. \tag{2.9}$$

Given its definition as a Fenchel conjugate, the negative entropy is convex.

In many models of interest, $a(\eta)$ is not feasible to compute because of the complexity of $\mathcal{M}$ or the lack of any explicit form for $a^*(\mu)$. However, we can bound $a(\eta)$ using Eq. (2.8):

$$a(\eta) \geq \mu^T \eta - a^*(\mu), \tag{2.10}$$

for any mean parameter $\mu \in \mathcal{M}$. Moreover, the tightness of the bound is measured by a Kullback-Leibler divergence expressed in terms of a mixed parameterization:

$$\begin{aligned} D(q(z \,|\, \eta(\mu)) \;||\; p(z \,|\, \eta)) &= \mathrm{E}_{\eta(\mu)} \left[\log q(z \,|\, \eta(\mu)) - \log p(z \,|\, \eta)\right] \\ &= \eta(\mu)^T \mu - a(\eta(\mu)) - \eta^T \mu + a(\eta) \\ &= a(\eta) - \eta^T \mu + a^*(\eta(\mu)). \end{aligned} \tag{2.11}$$

Mean-field variational methods are a special class of variational methods that are based on maximizing the bound in Eq. (2.10) with respect to a subset $\mathcal{M}_{\text{tract}}$

of the space $\mathcal{M}$ of realizable mean parameters. In particular, $\mathcal{M}_{\text{tract}}$ is chosen so that $a^*(\eta(\mu))$ can be evaluated tractably and so that the maximization over $\mathcal{M}_{\text{tract}}$ can be performed tractably. Equivalently, given the result in Eq. (2.11), mean-field variational methods minimize the KL divergence $D(q(z \,|\, \eta(\mu)) \;||\; p(z \,|\, \eta))$ with respect to its first argument.

If the distribution of interest is a posterior, then $a(\eta)$ is the log likelihood. Consider in particular a latent variable probabilistic model with hyperparameters η, observed variables $\mathbf{x} = \{x_1, \ldots, x_N\}$, and latent variables $\mathbf{z} = \{z_1, \ldots, z_M\}$. The posterior can be written as:

$$p(\mathbf{z} \,|\, \mathbf{x}, \eta) = \exp\{\log p(\mathbf{z}, \mathbf{x} \,|\, \eta) - \log p(\mathbf{x} \,|\, \eta)\}, \tag{2.12}$$

and the bound in Eq. (2.10) applies directly. We have:

$$\log p(\mathbf{x} \,|\, \eta) \geq \mathrm{E}_q \left[\log p(\mathbf{x}, \mathbf{Z} \,|\, \eta)\right] - \mathrm{E}_q \left[\log q(\mathbf{Z})\right]. \tag{2.13}$$

This equation holds for any q via Jensen's inequality, but, as our analysis has shown, it is useful specifically for q of the form $q(z \,|\, \eta(\mu))$ for $\mu \in \mathcal{M}_{\text{tract}}$.

A straightforward way to construct tractable subfamilies of exponential family distributions is to consider factorized families, in which each factor is an exponential family distribution depending on a so-called *variational parameter*. In particular, let us consider distributions of the form $q(\mathbf{z} \,|\, \boldsymbol{\nu}) = \prod_{i=1}^{M} q(z_i \,|\, \nu_i)$, where $\boldsymbol{\nu} = \{\nu_1, \nu_2, ..., \nu_M\}$ are variational parameters. Using this class of distributions, we simplify the likelihood bound using the chain rule:

$$\log p(\mathbf{x} \,|\, \eta) \geq \log p(\mathbf{x} \,|\, \eta) + \sum_{m=1}^{M} \mathrm{E}_q \left[\log p(Z_m \,|\, \mathbf{x}, Z_1, \ldots, Z_{m-1}, \eta)\right] - \sum_{m=1}^{M} \mathrm{E}_q \left[\log q(Z_m \,|\, \nu_m)\right]. \tag{2.14}$$

To obtain the best approximation available within the factorized subfamily, we now wish to optimize this expression with respect to ν_i.

To optimize with respect to ν_i, reorder $\mathbf{z}$ such that z_i is last in the list. The portion of Eq. (2.14) depending on ν_i is:

$$\ell_i = \mathrm{E}_q \left[\log p(z_i \,|\, \mathbf{z}_{-i}, \mathbf{x}, \eta)\right] - \mathrm{E}_q \left[\log q(z_i \,|\, \nu_i)\right]. \tag{2.15}$$

Given our assumption that the variational distribution $q(z_i \mid \nu_i)$ is in the exponential family, we have:

$$q(z_i \mid \nu_i) = h(z_i) \exp\{\nu_i^T z_i - a(\nu_i)\},$$

and Eq. (2.15) simplifies as follows:

$$\begin{aligned} \ell_i &= \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta) - \log h(Z_i) - \nu_i^T Z_i + a(\nu_i)\right] \\ &= \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] - \mathrm{E}_q \left[\log h(Z_i)\right] - \nu_i^T a'(\nu_i) + a(\nu_i), \end{aligned}$$

because $\mathrm{E}_q [Z_i] = a'(\nu_i)$. The derivative with respect to ν_i is:

$$\frac{\partial}{\partial \nu_i} \ell_i = \frac{\partial}{\partial \nu_i} \left(\mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] - \mathrm{E}_q \left[\log h(Z_i)\right]\right) - \nu_i^T a''(\nu_i). \tag{2.16}$$

Thus the optimal ν_i satisfies:

$$\nu_i = \left[a''(\nu_i)\right]^{-1} \left(\frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] - \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log h(Z_i)\right]\right). \tag{2.17}$$

The result in Eq. (2.17) is general. Under the assumptions of Section 2.1.3, a further simplification is achieved. In particular, when the conditional $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$ is an exponential family distribution:

$$p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta) = h(z_i) \exp\{g_i(\mathbf{z}_{-i}, \mathbf{x}, \eta)^T z_i - a(g_i(\mathbf{z}_{-i}, \mathbf{x}, \eta))\},$$

where $g_i(\mathbf{z}_{-i}, \mathbf{x}, \eta)$ denotes the natural parameter for z_i when conditioning on the remaining latent variables and the observations. This yields simplified expressions for the expected log probability of Z_i and its first derivative:

$$\begin{aligned} \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] &= \mathrm{E} \left[\log h(Z_i)\right] + \mathrm{E}_q \left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta)\right]^T a'(\nu_i) - \mathrm{E}_q \left[a(g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta))\right] \\ \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log p(Z_i \mid \mathbf{Z}_{-i}, \mathbf{x}, \eta)\right] &= \frac{\partial}{\partial \nu_i} \mathrm{E}_q \left[\log h(Z_i)\right] + \mathrm{E}_q \left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta)\right]^T a''(\nu_i). \end{aligned}$$

Using the first derivative in Eq. (2.17), the maximum is attained at:

$$\nu_i = \mathrm{E}_q \left[g_i(\mathbf{Z}_{-i}, \mathbf{x}, \eta)\right]. \tag{2.18}$$

We define a coordinate ascent algorithm based on Eq. (2.18) by iteratively updating ν_i for $i \in \{1, \ldots, N\}$. Such an algorithm finds a local maximum of Eq. (2.13) by

Proposition 2.7.1 of Bertsekas (1999), under the condition that the right-hand side of Eq. (2.15) is strictly convex. Further perspectives on algorithms of this kind can be found in Xing et al. (2003) and Beal (2003).

Notice the interesting relationship of this algorithm to the Gibbs sampler. In Gibbs sampling, we iteratively draw the latent variables z_i from the distribution $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$. In mean-field variational inference, we iteratively update the variational parameters of z_i to be equal to the expected value of the parameter g_i of the conditional distribution $p(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \eta)$, where the expectation is taken under the variational distribution.[2]

2.3 Discussion

In this chapter, we described the directed graphical model formalism, and used it to represent latent variable models for data analysis. For the class of models with conditional exponential family distributions—for which conjugate priors and mixture distributions are sufficient—we derived Gibbs sampling and mean-field variational methods of approximate posterior inference.

Choosing an approximate inference technique is an important part of the data analysis process. In this thesis, we typically prefer mean-field variational methods to Gibbs sampling. Gibbs sampling does have some advantages over variational inference. It gives samples from the exact posterior, while estimates based on variational methods incur an unknown bias. However, obtaining correct samples crucially depends on the Markov chain's convergence to its stationary distribution. This can be a slow process, and assessing whether the chain has converged is difficult. Theoretical bounds on the mixing time are of little practical use, and there is no consensus on how to choose one of the several empirical methods developed for this purpose (**?**).

On the other hand, variational methods are deterministic and have a clear convergence criterion given by the bound in Eq. (2.13). Furthermore, they are typically

[2]This relationship has inspired the software package VIBES (Bishop et al., 2003), which is a variational version of the BUGS package (Gilks et al., 1996).

faster than Gibbs sampling, as we will demonstrate empirically in Section 5.5. This is particularly important in view of the goal of efficient data analysis of large collections of text and images.

Chapter 3

Latent Dirichlet allocation

In this chapter, we begin to consider the problem of modeling text corpora and other collections of discrete data. The goal is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, and similarity and relevance judgments.

Significant progress has been made on this problem by researchers in the field of information retrieval (IR) (Baeza-Yates and Ribeiro-Neto, 1999). The basic methodology proposed by IR researchers for text corpora—a methodology successfully deployed in modern Internet search engines—reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts. In the popular *tf-idf* scheme (Salton and McGill, 1983), a basic vocabulary of "words" or "terms" is chosen, and, for each document in the corpus, a count is formed of the number of occurrences of each word. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a word in the entire corpus (generally on a log scale, and again suitably normalized). The end result is a term-by-document matrix X whose columns contain the *tf-idf* values for each of the documents in the corpus. Thus the *tf-idf* scheme reduces documents of arbitrary length to fixed-length lists of numbers.

While the *tf-idf* reduction has some appealing features—notably in its basic identification of sets of words that are discriminative for documents in the collection—the

approach provides a relatively small amount of reduction in description length and reveals little in the way of inter- or intra-document statistical structure. To address these shortcomings, IR researchers have proposed several other dimensionality reduction techniques, most notably *latent semantic indexing (LSI)* (Deerwester et al., 1990). LSI uses a singular value decomposition of the X matrix to identify a linear subspace in the space of *tf-idf* features that captures most of the variance in the collection. This approach can achieve significant compression in large collections. Furthermore, Deerwester et al. argue that the derived features of LSI, which are linear combinations of the original *tf-idf* features, can capture some aspects of basic linguistic notions such as synonymy and polysemy.

To substantiate the claims regarding LSI, and to study its relative strengths and weaknesses, it is useful to develop a generative probabilistic model of text corpora and to study the ability of LSI to recover aspects of the generative model from data (Papadimitriou et al., 1998). Given a generative model of text, however, it is not clear why one should adopt the LSI methodology—one can attempt to proceed more directly, fitting the model with data using maximum likelihood or Bayesian methods.

A significant step forward in this regard was made by Hofmann (1999b), who presented the *probabilistic LSI (pLSI)* model, also known as the *aspect model*, as an alternative to LSI. The pLSI approach, which is described in detail in Section 3.3.3, models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics." Thus each word is generated from a single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics. This distribution is the "reduced description" associated with the document.

While Hofmann's work is a useful step toward probabilistic modeling of text, it is incomplete in that it provides no probabilistic model at the level of documents. In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers. This leads

to several problems: (1) the number of parameters in the model grows linearly with the size of the corpus, which leads to serious problems with overfitting, and (2) it is not clear how to assign probability to a document outside of the training set.

To see how to proceed beyond pLSI, let us consider the fundamental probabilistic assumptions underlying the class of dimensionality reduction methods that includes LSI and pLSI. All of these methods are based on the "bag-of-words" assumption—that the order of words in a document can be neglected. In the language of probability theory, this is an assumption of *exchangeability* for the words in a document (Aldous, 1985). Moreover, although less often stated formally, these methods also assume that documents are exchangeable; the specific ordering of the documents in a corpus can also be neglected.

A classic representation theorem due to de Finetti (1990) establishes that any collection of exchangeable random variables has a representation as a mixture distribution—in general an infinite mixture. Thus, if we wish to consider exchangeable representations for documents and words, we need to consider mixture models that capture the exchangeability of both words and documents. This line of thinking leads to the *latent Dirichlet allocation (LDA)* model, a hierarchical model of the form found in Section 2.1.

It is important to emphasize that an assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed. Rather, exchangeability essentially can be interpreted as meaning "*conditionally* independent and identically distributed," where the conditioning is with respect to an underlying latent parameter of a probability distribution. Conditionally, the joint distribution of the random variables is simple and factored while marginally over the latent parameter, the joint distribution can be quite complex. Thus, while an assumption of exchangeability is clearly a major simplifying assumption in the domain of text modeling, and its principal justification is that it leads to methods that are computationally efficient, the exchangeability assumptions do not necessarily lead to methods that are restricted to simple frequency counts or linear operations. In this chapter, we aim to demonstrate that, by taking the de Finetti theorem seri-

ously, we can capture significant intra-document statistical structure via the mixing distribution.

3.1 Notation and terminology

We will use the language of text collections, referring to entities such as "words," "documents," and "corpora." This is useful in that it helps to guide intuition, particularly when we introduce latent variables which aim to capture abstract notions such as topics. It is important to note, however, that the LDA model is not necessarily tied to text, and has applications to other problems involving collections of data, including data from domains such as collaborative filtering, content-based image retrieval and bioinformatics. For example, in Section 3.6.3, we present experimental results in the collaborative filtering domain, and Chapter 4 will focus on image/caption data.

Formally, we define the following terms:

- A *word* is the basic unit of discrete data, defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$. Recall that we represent multinomial data using unit-basis vectors, as described in Section 2.1.3.
- A *document* is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the sequence.
- A *corpus* is a collection of M documents denoted by $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

We wish to find a probabilistic model of a corpus that not only assigns high probability to members of the corpus, but also assigns high probability to other similar documents.

3.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics,

where each topic is characterized by a distribution on words.[1]

LDA assumes the following generative process for each document $\mathbf{w}$ in a corpus $\mathcal{D}$.

1. Choose $N \mid \xi \sim \text{Poisson}(\xi)$.
2. Choose proportions $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
3. For $n \in \{1, \ldots, N\}$:
   (a) Choose topic $Z_n \mid \theta \sim \text{Mult}(\theta)$.
   (b) Choose word $W_n \mid \{z_n, \beta_{1:K}\} \sim \text{Mult}(\beta_{z_n})$.

Several simplifying assumptions are made in this basic model, some of which we will remove in the subsequent chapters. First, the dimensionality K of the Dirichlet distribution (and thus the dimensionality of the topic variable Z) is assumed known and fixed. Second, the word probabilities are parameterized by a $k \times V$ matrix $\beta_{1:K}$ where $\beta_{ij} = p(w^j = 1 \mid z^i = 1)$, which for now we treat as a fixed quantity that is to be estimated. Finally, the Poisson assumption is not critical to anything that follows and more realistic document length distributions can be used as needed.[2] Furthermore, note that the number of words N is independent of all the other data generating variables (proportions θ and latent topics $\mathbf{z}$). It is thus an ancillary variable, and we will generally ignore its randomness in the subsequent development.

3.2.1 The Dirichlet distribution

A K-dimensional Dirichlet random variable θ can take values in the $(K-1)$-simplex (a K-vector θ lies in the $(K-1)$-simplex if $\theta_i \geq 0$, $\sum_{i=1}^{K} \theta_i = 1$), and has the following

[1] We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions of sets of words.

[2] However, with the Poisson distribution of the number of words, Buntine and Jakulin (2004) show that LDA can be interpreted as a discrete independent component analysis model.

probability density on this simplex:

$$p(\theta \mid \alpha) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1}, \tag{3.1}$$

where the parameter α is a K-vector with components $\alpha_i > 0$, and where $\Gamma(x)$ is the Gamma function. The Dirichlet is a convenient distribution on the simplex: it is in the exponential family, has finite dimensional sufficient statistics, and is conjugate to the multinomial distribution. In Section 3.4, these properties will facilitate the development of inference and parameter estimation algorithms for LDA.

One way of establishing the conjugacy between the Dirichlet and multinomial distributions is to consider their natural exponential family representations and simply observe that conjugacy holds (see Section 2.1.2). However, we can also proceed directly using Eq. (3.1). First, note that:

$$\mathrm{E}\left[\theta \mid \alpha\right] = \frac{\alpha}{\sum_{j=1}^{K} \alpha_j}. \tag{3.2}$$

From this result, we find that the conjugacy takes a convenient form. If $\theta \mid \alpha \sim \mathrm{Dir}(\alpha)$ and $Z \mid \theta \sim \mathrm{Mult}(\theta)$, then:

$$\begin{aligned} p(\theta \mid z = i, \alpha) &= \frac{p(z = i \mid \theta) p(\theta \mid \alpha)}{p(z = i \mid \alpha)} \\ &= \left(\frac{\sum_{j=1}^{K} \alpha_j}{\alpha_i}\right) \left(\frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \theta_1^{\alpha_1 - 1} \cdots \theta_i^{\alpha_i - 1} \cdots \theta_K^{\alpha_K - 1}\right) \theta_i \\ &= \frac{\Gamma(1 + \sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j + \delta(j = i))} \theta_1^{\alpha_1 - 1} \cdots \theta_i^{\alpha_i} \cdots \theta_K^{\alpha_K - 1} \\ &= \mathrm{Dir}(\alpha_1, \ldots, \alpha_i + 1, \ldots, \alpha_K). \end{aligned} \tag{3.3}$$

From this result, it follows that:

$$\theta \mid \{z_{1:N}, \alpha\} \sim \mathrm{Dir}(\alpha_1 + n_1(z_{1:N}), \ldots, \alpha_K + n_k(z_{1:N})), \tag{3.4}$$

where $n_i(z_{1:N})$ is the number of times that $z_n = i$.

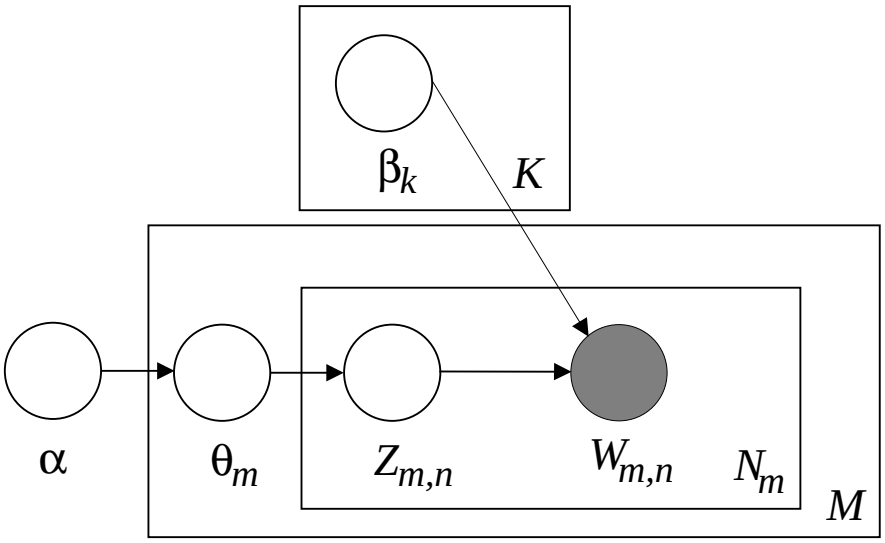


Figure 3.1: Graphical model representation of LDA.

3.2.2 Joint distribution of a corpus

Given the parameters α and $\beta_{1:K}$, the joint distribution of topic proportions θ, a set of N topics $\mathbf{z}$, and a set of N words $\mathbf{w}$ is given by:

$$p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta_{1:K}) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K}), \tag{3.5}$$

where $p(z_n \mid \theta)$ is simply θ_i for the unique i such that $z_n^i = 1$, and $p(w_n \mid z_n, \beta)$ is the analogous component in $\beta_{1:K}$. Integrating over θ and summing over latent topics, we obtain the marginal distribution of a document:

$$p(\mathbf{w} \mid \alpha, \beta_{1:K}) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K}) \right) d\theta. \tag{3.6}$$

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

$$p(\mathcal{D} \mid \alpha, \beta_{1:K}) = \prod_{d=1}^{M} \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(w_{dn} \mid z_{dn}, \beta_{1:K}) \right) d\theta_d.$$

LDA is represented as a probabilistic graphical model in Figure 3.1. As the figure makes clear, LDA is a hierarchical model with three levels. The parameters α and $\beta_{1:K}$

are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled *repeatedly* within the document. Under this model, documents can be associated with multiple topics.

3.2.3 LDA and exchangeability

A finite set of random variables $\{Z_1, \ldots, Z_N\}$ is said to be *exchangeable* if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(N)}).$$

An infinite sequence of random variables is *infinitely exchangeable* if every finite subsequence is exchangeable.

De Finetti's representation theorem (de Finetti, 1990) states that the joint distribution of an infinitely exchangeable sequence of random variables is as if a random parameter were drawn from some distribution and then the random variables in question were *independent* and *identically distributed*, conditioned on that parameter. The elegance of De Finetti's theorem is that from natural assumptions of exchangeability come a principled justification for hierarchical Bayesian modeling; it is thus at the foundation of the Bayesian agenda.

In LDA, we assume that words are generated by topics (by fixed conditional distributions) and that those topics are infinitely exchangeable within a document.

By de Finetti's theorem, the probability of a sequence of words and topics must therefore have the form:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \,|\, \theta) p(w_n \,|\, z_n) \right) d\theta,$$

where θ is the random parameter of a multinomial over topics. We obtain the LDA distribution of documents in Eq. (3.6) by marginalizing out the topic variables and endowing θ with a Dirichlet distribution.

3.2.4 A continuous mixture of unigrams

The LDA model shown in Figure 3.1 is somewhat more elaborate than the two-level models often studied in the classical hierarchical Bayesian literature. By marginalizing over the hidden topic variable z, however, we can understand LDA as a two-level model.

In particular, let us form the word distribution $p(w \,|\, \theta, \beta)$:

$$p(w \,|\, \theta, \beta_{1:K}) = \sum_{z} p(w \,|\, z, \beta_{1:K}) p(z \,|\, \theta),$$

and note that this is a random quantity since it depends on θ.

We now define the following generative process for a document $\mathbf{w}$:

1. Choose $\theta \,|\, \alpha \sim \mathrm{Dir}(\alpha)$.
2. For $n \in \{1, \ldots, N\}$:
   (a) Choose word $W_n \,|\, \theta, \beta_{1:K} \sim \mathrm{Mult}\left(\sum_{i=1}^{K} \theta_i \beta_i\right)$

This process defines the marginal distribution of a document as a continuous mixture distribution:

$$p(\mathbf{w} \,|\, \alpha, \beta_{1:K}) = \int p(\theta \,|\, \alpha) \left(\prod_{n=1}^{N} p(w_n \,|\, \theta, \beta_{1:K}) \right) d\theta,$$

where $p(w_n \,|\, \theta, \beta_{1:K})$ are the mixture components and $p(\theta \,|\, \alpha)$ are the mixture weights.

Figure 3.2 illustrates this interpretation of LDA. It depicts the distribution of $p(w \,|\, \theta, \beta_{1:K})$ which is induced from a particular instance of an LDA model. Note that

this distribution on the $(V-1)$-simplex is attained with only $K+KV$ parameters yet exhibits a very interesting multimodal structure. This perspective also gives an interesting connection to principal component analysis, in which the data arise from a normal distribution whose mean is the inner product of a latent variable (cf., θ) and a collection of means (cf., $\beta_{1:K}$). This connection is highlighted in Buntine and Jakulin (2004), who redub LDA as discrete PCA.

3.3 Other latent variable models for text

In this section we compare LDA to simpler latent variable models for text—the unigram model, a mixture of unigrams, and the pLSI model. Furthermore, we present a unified geometric interpretation of these models which highlights their key differences and similarities.

3.3.1 Unigram model

Under the unigram model, the words of every document are drawn independently from a single multinomial distribution:

$$p(\mathbf{w} \mid \beta) = \prod_{n=1}^{N} p(w_n \mid \beta),$$

where β is a multinomial distribution over words. This is illustrated in the graphical model in Figure 3.3a.

3.3.2 Mixture of unigrams

If we augment the unigram model with a discrete random topic variable (Figure 3.3b), we obtain a *mixture of unigrams* model (Nigam et al., 2000). Under this mixture model, each document is generated by first choosing a topic and then generating words independently from the multinomial associated with that topic. The probability of a document is:

$$p(\mathbf{w} \mid \theta, \beta_{1:K}) = \sum_{z} p(z \mid \theta) \prod_{n=1}^{N} p(w_n \mid z, \beta_{1:K}),$$

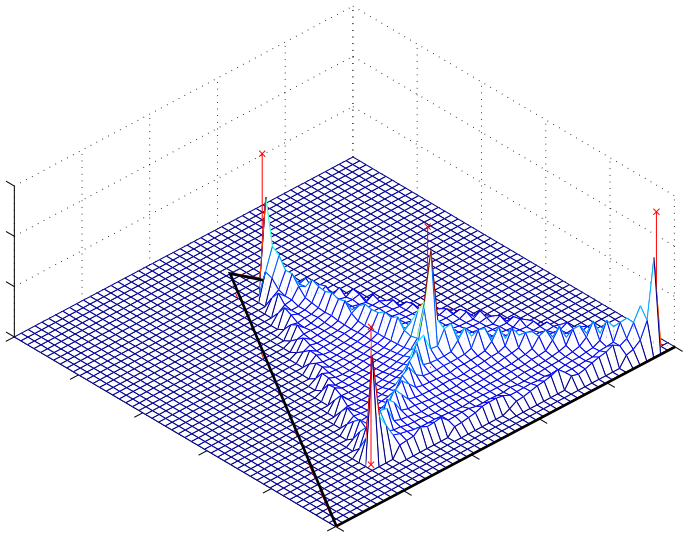
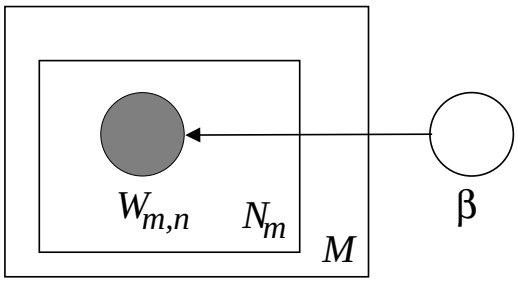
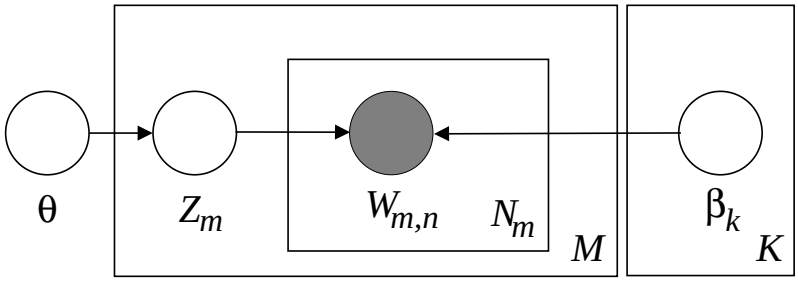


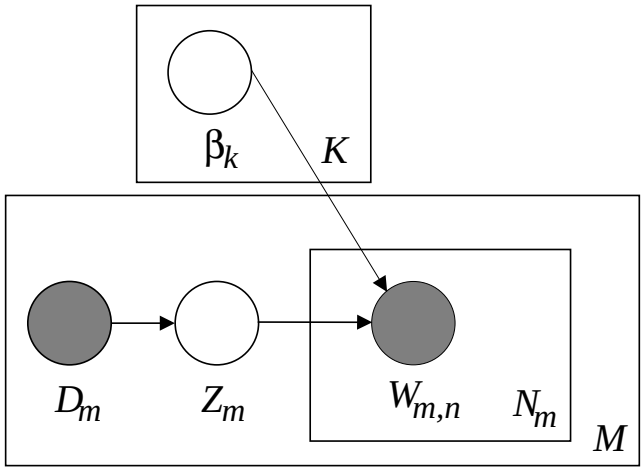
Figure 3.2: An example density on unigram distributions $p(w \mid \theta, \beta_{1:K})$ under LDA for three words and four topics. The triangle embedded in the x-y plane is the 2-D simplex representing all possible multinomial distributions over three words. Each of the vertices of the triangle corresponds to a deterministic distribution that assigns probability one to one of the words; the midpoint of an edge gives probability 0.5 to two of the words; and the centroid of the triangle is the uniform distribution over all three words. The four points marked with an `x` are the locations of the multinomial distributions $p(w \mid z)$ for each of the four topics, and the surface shown on top of the simplex is an example of a density over the $(V-1)$-simplex (multinomial distributions of words) given by LDA.



(a) unigram



(b) mixture of unigrams



(c) pLSI/aspect model

Figure 3.3: Graphical model representation of different models of discrete data.

where θ is a single distribution over the K topics, that is fixed for the entire corpus. When estimated from a corpus, the word distributions $\beta_{1:K}$ can be viewed as representations of topics under the assumption that each document exhibits exactly one topic. As the empirical results in Section 3.6 illustrate, this assumption is often too limiting to effectively model a large collection of documents.

In contrast, the LDA model allows documents to exhibit multiple topics to different degrees. This is achieved at a cost of just one additional parameter: there are $K - 1$ parameters associated with $p(z \mid \theta)$ in the mixture of unigrams, and K parameters associated with $p(\theta \mid \alpha)$ in LDA.

3.3.3 Probabilistic latent semantic indexing

Probabilistic latent semantic indexing (pLSI) is another widely used document model (Hofmann, 1999b). The pLSI model, illustrated in Figure 3.3c, posits that a document label and word are conditionally independent given a latent topic:

$$p(d, w_n \mid \theta_d, \beta_{1:K}) = p(d) \sum_z p(w_n \mid z, \beta_{1:K}) p(z \mid \theta_d),$$

where θ_d are document-specific topic proportions, and $\beta_{1:K}$ is defined as for the mixture of unigrams and LDA.

The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic. In a sense, it does capture the possibility that a document may contain multiple topics since θ_d serves as the mixture weights of the topics for a particular document d. However, θ_d is a parameter and d is a dummy index into the list of documents in the *training set*. Thus, d is a multinomial random variable with as many possible values as there are training documents, and Hofmann estimates the topic proportions θ_d only for those documents on which it is trained. For this reason, pLSI is not a well-defined generative model of documents; there is no natural way to use it to assign probability to a previously unseen document.

A further difficulty with pLSI, which also stems from the use of a distribution indexed by training documents, is that the number of parameters which must be

estimated grows linearly with the number of training documents. The parameters for a K-topic pLSI model are K multinomial distributions of size V and M mixtures over the K hidden topics. This gives $KV + KM$ parameters and therefore linear growth in M. This suggests that the model is prone to overfitting and, empirically, overfitting is indeed a serious problem (see Section 3.6.1). In practice, a tempering heuristic is used to smooth the parameters of the model for acceptable predictive performance. It has been shown, however, that overfitting can occur even when tempering is used (Popescul et al., 2001).

LDA overcomes both of these problems by treating the topic mixture weights as a K-parameter hidden *random variable* rather than a large set of individual parameters which are explicitly linked to the training set. As described in Section 3.2, LDA is a well-defined generative model and generalizes easily to new documents. Furthermore, the $K + KV$ parameters in a K-topic LDA model do not grow with the size of the training corpus. We will see in Section 3.6.1 that LDA does not suffer from the same overfitting issues as pLSI.

3.3.4 A geometric interpretation

A good way of illustrating the differences between LDA and the other latent topic models is by considering the geometry of the latent space, and seeing how a document is represented in that geometry under each model.

All four of the models described above—unigram, mixture of unigrams, pLSI, and LDA—operate in the space of distributions of words. Each such distribution can be viewed as a point on the $(V - 1)$-simplex, which we call the *word simplex*.

The unigram model finds a single point on the word simplex and posits that all words in the corpus come from the corresponding distribution. The latent variable models consider K points on the word simplex and form a sub-simplex based on those points, which we call the *topic simplex*. Note that any point on the topic simplex is also a point on the word simplex. The different latent variable models use the topic simplex in different ways to generate a document.

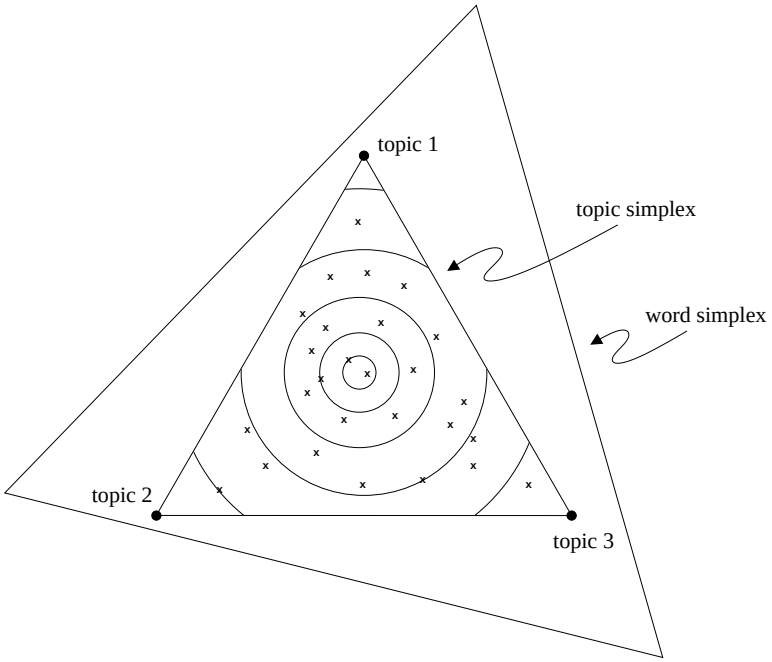


Figure 3.4: The topic simplex for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word (respectively) has probability one. The three points of the topic simplex correspond to three different distributions over words. The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by `x`. LDA places a smooth distribution on the topic simplex denoted by the contour lines.

- The mixture of unigrams model posits that for each document, one of the K points on the word simplex (that is, one of the corners of the topic simplex) is chosen randomly and all the words of the document are drawn from the distribution corresponding to that point.

- The pLSI model posits that each word of a *training* document comes from a randomly chosen topic. The topics are themselves drawn from a document-specific distribution over topics, i.e., a point on the topic simplex. There is one such distribution for each document; the set of training documents thus defines an empirical distribution on the topic simplex.

- LDA posits that each word of both the observed and unseen documents is generated by a randomly chosen topic which is drawn from a distribution with a randomly chosen parameter. This parameter is sampled once per document from a smooth distribution on the topic simplex.

These differences are highlighted in Figure 3.4.

3.4 Posterior inference

We have described the motivation behind LDA and illustrated its conceptual advantages over other latent topic models. In this section, we turn our attention to procedures for posterior inference under LDA.

For the moment, suppose that the topic distributions $\beta_{1:K}$ and Dirichlet parameters α are fixed. The posterior distribution of the document-specific hidden variables given a document is:

$$p(\theta, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta_{1:K}) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \,|\, \alpha, \beta_{1:K})}{p(\mathbf{w} \,|\, \alpha, \beta_{1:K})}.$$

Unfortunately, this distribution is intractable to compute in general. Indeed, to normalize the distribution we marginalize over the hidden variables and write Eq. (3.6) in terms of the model parameters:

$$p(\mathbf{w} \,|\, \alpha, \beta_{1:K}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j}\right) d\theta,$$

a function which is intractable due to the coupling between θ and $\beta_{1:K}$ in the summation over latent topics (Dickey, 1983). Dickey shows that this function is an expectation under a particular extension to the Dirichlet distribution which can be represented with special hypergeometric functions (in fact, this extension is exactly the distribution of words given in Figure 3.2). It has been used in a Bayesian context for censored discrete data to represent the posterior on θ which, in that setting, is a random parameter (Dickey et al., 1987).

Although the posterior distribution is intractable for exact inference, a wide variety of approximate inference algorithms can be considered for LDA, including Laplace approximation, variational approximation, and Markov chain Monte Carlo. In this section, we describe the mean-field variational algorithm of Section 2.2.2 for inference in LDA, and discuss some of the alternatives in Section 3.7.

3.4.1 Mean-field variational inference

The fully factorized distribution of the latent variables is:

$$q(\theta, \mathbf{z} \,|\, \gamma, \phi_{1:N}) = q(\theta \,|\, \gamma) \prod_{n=1}^{N} q(z_n \,|\, \phi_n), \tag{3.7}$$

where the Dirichlet parameter γ and the multinomial parameters $\phi_{1:N}$ are the free variational parameters. We find the setting of these parameters to minimize the Kullback-Leibler divergence to the true posterior:

$$(\gamma^*, \phi^*_{1:N}) = \arg \min_{(\gamma, \phi_{1:N})} \mathrm{D}(q(\theta, \mathbf{z} \,|\, \gamma, \phi_{1:N}) \parallel p(\theta, \mathbf{z} \,|\, \mathbf{w}, \alpha, \beta_{1:K})). \tag{3.8}$$

This minimization can be achieved with the iterative fixed-point method described in Section 2.2.2. In particular, we obtain the following pair of update equations from Eq. (2.18):

$$\phi_{ni} \;\propto\; \beta_{iw_n} \exp\{\mathrm{E}_q[\log(\theta_i) \,|\, \gamma]\} \tag{3.9}$$

$$\gamma_i \;=\; \alpha_i + \textstyle\sum_{n=1}^{N} \phi_{ni}. \tag{3.10}$$

The expectation in the multinomial update can be computed as follows:

$$\mathrm{E}_q[\log(\theta_i) \,|\, \gamma] = \Psi(\gamma_i) - \Psi\left(\textstyle\sum_{j=1}^{K} \gamma_j\right), \tag{3.11}$$

where Ψ is the digamma function, the first derivative of the $\log \Gamma$ function which is computable via Taylor approximations (Abramowitz and Stegun, 1970).

This follows from the natural parameterization of the Dirichlet distribution in Eq. (3.1):

$$p(\theta \mid \alpha) = \exp\left\{\left(\sum_{i=1}^{K}(\alpha_i - 1)\log\theta_i\right) + \log\Gamma\left(\sum_{i=1}^{K}\alpha_i\right) - \sum_{i=1}^{K}\log\Gamma(\alpha_i)\right\}.$$

From this form, we immediately see that the natural parameter for the Dirichlet is $\eta_i = \alpha_i - 1$ and the sufficient statistic is $t(\theta_i) = \log\theta_i$. Using Eq. (2.3), the fact that the derivative of the log normalizer is equal to the expectation of the sufficient statistic, we obtain Eq. (3.11).

Eqs. (3.9) and (3.10) have an appealing intuitive interpretation. The Dirichlet update is a posterior Dirichlet given expected observations taken under the variational distribution, $\mathrm{E}[z_n \mid \phi_n]$. The multinomial update is akin to using Bayes' theorem, $p(z_n \mid w_n) \propto p(w_n \mid z_n)p(z_n)$, where $p(z_n)$ is approximated by the exponential of the expected value of its logarithm under the variational distribution. This matches the intuitions in Section 2.2.2, which link coordinate ascent mean-field variational inference to Gibbs sampling.

It is important to note that the variational distribution is actually a conditional distribution, varying as a function of the document $\mathbf{w}$. This occurs because the optimization problem in Eq. (3.8) is conducted for fixed $\mathbf{w}$, and thus yields optimizing parameters $(\gamma^*, \phi^*_{1:N})$ that are a function of $\mathbf{w}$. We can write the resulting variational distribution as $q(\theta, \mathbf{z} \mid \gamma^*(\mathbf{w}), \phi^*_{1:N}(\mathbf{w}))$, where we have made the dependence on $\mathbf{w}$ explicit. Thus the variational distribution can be viewed as an approximation to the posterior distribution $p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta_{1:K})$.

In the language of text, the optimizing parameters are document-specific. In particular, we view the Dirichlet parameters as providing a representation of a document in the topic simplex.

We summarize the variational inference procedure in Figure 3.5, with appropriate starting points that assume a flat topic distribution for each word in the document. From the pseudocode it is clear that each iteration of variational inference for LDA

(1) initialize $\phi^0_{ni} := 1/k$ for all i and n

(2) initialize $\gamma_i := \alpha_i + N/k$ for all i

(3) **repeat**

(4) **for** $n = 1$ **to** N

(5) **for** $i = 1$ **to** k

(6) $\phi^{t+1}_{ni} := \beta_{iw_n} \exp(\Psi(\gamma^t_i))$

(7) normalize ϕ^{t+1}_n to sum to 1.

(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi^{t+1}_n$

(9) **until** convergence

Figure 3.5: A variational inference algorithm for LDA.

requires $\mathrm{O}((N+1)K)$ operations. Empirically, we find that the number of iterations required for a single document is on the order of the number of words in the document. This yields a total number of operations roughly on the order of N^2K.

3.4.2 Empirical Bayes estimates

In this section we present an empirical Bayes method for parameter estimation in the LDA model (see Section 3.4.3 for a fuller Bayesian approach). In particular, given a corpus of documents $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$, we wish to find parameters that maximize the (marginal) log likelihood of the data:

$$\ell(\alpha, \beta_{1:K}) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \,|\, \alpha, \beta_{1:K}).$$

As we have described above, the quantity $p(\mathbf{w} \,|\, \alpha, \beta_{1:K})$ cannot be computed tractably. However, we can use the variational EM algorithm to maximize the lower bound on the log likelihood given in the mean-field variational framework:

$$\mathcal{L}(\alpha, \beta) = \sum_{d=1}^{M} \mathrm{E}_{q_d}[\log p(\theta_d \,|\, \alpha)] + \mathrm{E}_{q_d}[\log p(\mathbf{z}_d \,|\, \theta_d)] + \mathrm{E}_{q_d}[\log p(\mathbf{w}_d \,|\, \mathbf{z}_d, \beta)]$$
$$- \mathrm{E}_{q_d}[\log q_d(\theta_d)] - \mathrm{E}_{q_d}[\log q_d(\mathbf{z}_d)]$$

We expand this bound in terms of the model parameters and variational parameters. Each of the five lines below expands one of the five terms in the bound:

$$\begin{aligned}
\mathcal{L} = & \sum_{d=1}^{M} \log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} (\alpha_i - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \\
& + \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{dni}\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \\
& + \sum_{n=1}^{N} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} \\
& - \log \Gamma\left(\sum_{j=1}^{k} \gamma_{dj}\right) + \sum_{i=1}^{k} \log \Gamma(\gamma_{di}) - \sum_{i=1}^{k} (\gamma_{di} - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \\
& - \sum_{n=1}^{N} \sum_{i=1}^{k} \phi_{dni} \log \phi_{dni}.
\end{aligned} \tag{3.12}$$

The variational EM algorithm repeats the following two steps until convergence of the bound:

1. (E-step) For each document, find the optimizing values of the variational parameters $\{\gamma_d^*, \phi_{d,1:N}^* : d \in \mathcal{D}\}$ as described in the previous section.

2. (M-step) Maximize the lower bound on the log likelihood with respect to the model parameters α and $\beta_{1:K}$. This corresponds to finding maximum likelihood estimates with expected sufficient statistics under the approximate posterior from the E-step.

To maximize with respect to the multinomial parameters, we isolate terms and add Lagrange multipliers:

$$\mathcal{L}_{[\beta]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} + \sum_{i=1}^{k} \lambda_i \left(\sum_{j=1}^{V} \beta_{ij} - 1\right).$$

We take the derivative with respect to β_{ij}, set it to zero, and find:

$$\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}.$$

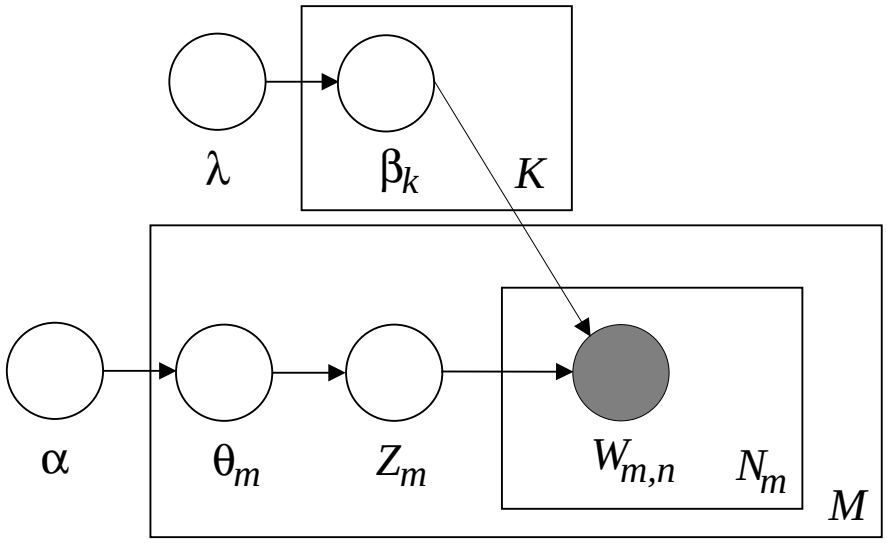


Figure 3.6: Graphical model representation of the smoothed LDA model.

Similarly, to maximize with respect to α, we isolate the appropriate terms:

$$\mathcal{L}_{[\alpha]} = \sum_{d=1}^{M} \left(\log \Gamma \left(\sum_{j=1}^{k} \alpha_j \right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} \left((\alpha_i - 1) \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right) \right) \right) \right).$$

Taking the derivative with respect to α_i gives:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = M \left(\Psi \left(\sum_{j=1}^{k} \alpha_j \right) - \Psi(\alpha_i) \right) + \sum_{d=1}^{M} \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right) \right).$$

This derivative depends on α_j, where $j \neq i$, and we therefore must use an iterative method to find the maximal value. Good methods for finding maximum likelihood estimates of Dirichlet distributions can be found in Minka (2000).

3.4.3 Smoothing

The large vocabulary size that is characteristic of many document corpora creates serious problems of sparsity. A new document is very likely to contain words that did not appear in any of the documents in a training corpus. Maximum likelihood estimates of the multinomial parameters assign zero probability to such words, and thus zero probability to new documents. The standard approach to coping with this problem is to smooth the multinomial parameters, assigning positive probability to

all vocabulary items whether or not they are observed in the training set (Jelinek, 1997). Laplace smoothing is commonly used; this essentially yields the mean of the posterior distribution under a uniform Dirichlet prior on the multinomial parameters.

Unfortunately, in the mixture model setting, simple Laplace smoothing is no longer justified as a maximum a posteriori method (although it is often implemented in practice; cf. Nigam et al., 1999). In fact, by placing a Dirichlet prior on the multinomial parameter we obtain an intractable posterior in the mixture model setting, for much the same reason that one obtains an intractable posterior in the basic LDA model. Our proposed solution to this problem is to simply apply variational inference methods to the extended hierarchical model that includes Dirichlet smoothing on the multinomial parameter. Using variational inference on a parameter with a fixed hyperparameter is known as *variational Bayes* (Attias, 2000; Beal, 2003).

In the LDA setting, we obtain the extended graphical model shown in Figure 3.6. The topic distributions are now random variables, independently drawn from an exchangeable Dirichlet distribution.[3] We now extend our inference procedures, to approximate the posterior of $\beta_{1:K}$ conditional on the data. Thus we move beyond the empirical Bayes procedure of Section 3.4.2 and consider a fuller Bayesian approach to LDA.

Again, we use a mean-field variational approach to Bayesian inference that uses a fully-factorized distribution of the random variables:

$$q(\beta_{1:K}, \mathbf{z}_{1:M}, \theta_{1:M} \mid \rho, \phi, \gamma) = \prod_{i=1}^{k} \mathrm{Dir}(\beta_i \mid \rho_i) \prod_{d=1}^{M} q_d(\theta_d, \mathbf{z}_d \mid \gamma_d, \phi_{d,1:N}),$$

where q_d is the variational distribution defined for LDA in Eq. (3.7). The resulting variational inference procedure yields Eqs. (3.9) and (3.10) as the update equations for the variational parameters ϕ_d and γ_d, with all instances of β replaced by $\exp\{\mathrm{E}[\log \beta \mid \rho]\}$. There is an additional update for the new variational parameter ρ:

$$\rho_{ij} = \lambda + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}. \tag{3.13}$$

[3]An exchangeable Dirichlet is a Dirichlet distribution with a single scalar parameter λ. The density is the same as a Dirichlet (Eq. 3.1) where $\alpha_i = \lambda$ for each component.

Iterating these equations to convergence yields an approximate posterior distribution of $\beta_{1:K}$, $\theta_{1:M}$, and $\mathbf{z}_{1:M}$.

We are now left with the hyperparameter of the exchangeable Dirichlet, as well as the hyperparameter of the Dirichlet for the topic proportions. Our approach is again (approximate) empirical Bayes—we use variational EM to find maximum likelihood estimates of these parameters based on the variational bound of the marginal likelihood.

3.5 Example

In this section, we provide an illustrative example of the use of LDA on real data. Our data are 16,000 documents from a subset of the TREC AP corpus (Harman, 1992). After removing a standard list of stop words, we used the EM algorithm described in Section 3.4.2 to find the Dirichlet and conditional multinomial parameters for a 100-topic LDA model. The top words from some of the resulting topic distributions are illustrated in Figure 3.7 (top). As we have hoped, these distributions seem to capture some of the underlying topics in the corpus (and we have named them according to these topics). Note that the removal of stop words is essential to this analysis; otherwise, they are found to be common in all topics. In Chapter 6, we will define a topic model on trees of topics which can automatically identify such words at different levels of granularity.

As we emphasized in Section 3.3, one of the advantages of LDA over related latent variable models is that it provides well-defined inference procedures for previously unseen documents. Indeed, we can illustrate how LDA works by performing inference on a held-out document and examining the resulting variational posterior parameters.

Figure 3.7 (bottom) is a document from the TREC AP corpus which was not used for parameter estimation. Using the variational inference algorithm in Section **??**, we computed the variational posterior Dirichlet parameters γ for the article and variational posterior multinomial parameters ϕ_n for each word in the article.

Recall that the ith posterior Dirichlet parameter is approximately the ith prior

Dirichlet parameter plus the expected number of words which were generated by the ith topic (see Eq. 3.10). Therefore, the prior Dirichlet parameters subtracted from the posterior Dirichlet parameters indicate the expected number of words which were allocated to each topic for a particular document. For the example article in Figure 3.7 (bottom), most of the γ_i are close to α_i. Four topics, however, are significantly larger (by this, we mean $\gamma_i - \alpha_i \geq 1$). Looking at the corresponding distributions over words identifies the topics which mixed to form this document (Figure 3.7, top).

Further insight comes from examining the ϕ_n parameters. These distributions approximate $p(z_n \mid \mathbf{w})$ and tend to peak toward one of the possible topic values. In the article text in Figure 3.7, the words are color coded according to these values (i.e., the ith color is used if $q(z_n^i = 1 \mid \phi_n) > 0.9$). With this illustration, one can identify how the different topics mixed in the document text.

While demonstrating the power of LDA, the posterior analysis also highlights some of its limitations. In particular, the bag-of-words assumption allows words that should be generated by the same topic (e.g., "William Randolph Hearst Foundation") to be allocated to several different topics. Overcoming this limitation requires an extension of the basic LDA model; in particular, we might relax the bag-of-words assumption by assuming partial exchangeability or Markovianity of word sequences (Girolami and Kaban, 2004).

3.6 Applications and Empirical Results

In this section, we discuss an empirical evaluation of LDA for several problem domains: document modeling, document classification, and collaborative filtering.

In LDA and mixtures of unigrams, the expected complete log likelihood of the data has local maxima at the points where all or some of the mixture components are equal to each other. To avoid these local maxima, it is important to initialize the EM algorithm appropriately. In our experiments, we initialize EM by seeding each conditional multinomial distribution with five documents, reducing their effective total length to two words, and smoothing across the whole vocabulary. This is essentially

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 3.7: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

an approximation to the scheme described in Heckerman and Meila (2001).

3.6.1 Document modeling

We fit a number of latent variable models, including LDA, with two text corpora to compare the generalization performance of these models. The documents in the corpora are treated as unlabeled; thus, our goal is density estimation—we wish to achieve high likelihood on a held-out test set. In particular, we computed the *perplexity* of a held-out test set to evaluate the models. The perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data, and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance.[4] More formally, for a test set of M documents, the perplexity is:

$$\text{perplexity} = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

In our experiments, we used a corpus of scientific abstracts from the C. Elegans community (Avery, 2002) containing 5,225 abstracts with 28,414 unique terms, and a subset of the TREC AP corpus containing 16,333 newswire articles with 23,075 unique terms. In both cases, we held out 10% of the data for test purposes and fit the models with the remaining 90%. In preprocessing the data, we removed a standard list of 50 stop words from each corpus. From the AP data, we further removed words that occurred only once.

We compared LDA to the unigram, mixture of unigrams, and pLSI models described in Section 3.3. We fit all the hidden variable models using EM with exactly the same stopping criteria, that the average change in expected log likelihood is less than 0.001%.

[4]Note that we simply use perplexity as a figure of merit for comparing models. The models that we compare are all unigram ("bag-of-words") models, which—as we have discussed in the beginning of the chapter—are of interest in the information retrieval context. We are *not* attempting to do language modeling—an enterprise that would require us to examine trigram or other higher-order models. We note in passing, however, that extensions of LDA could be considered that involve Dirichlet-multinomial over trigrams instead of unigrams (Girolami and Kaban, 2004).

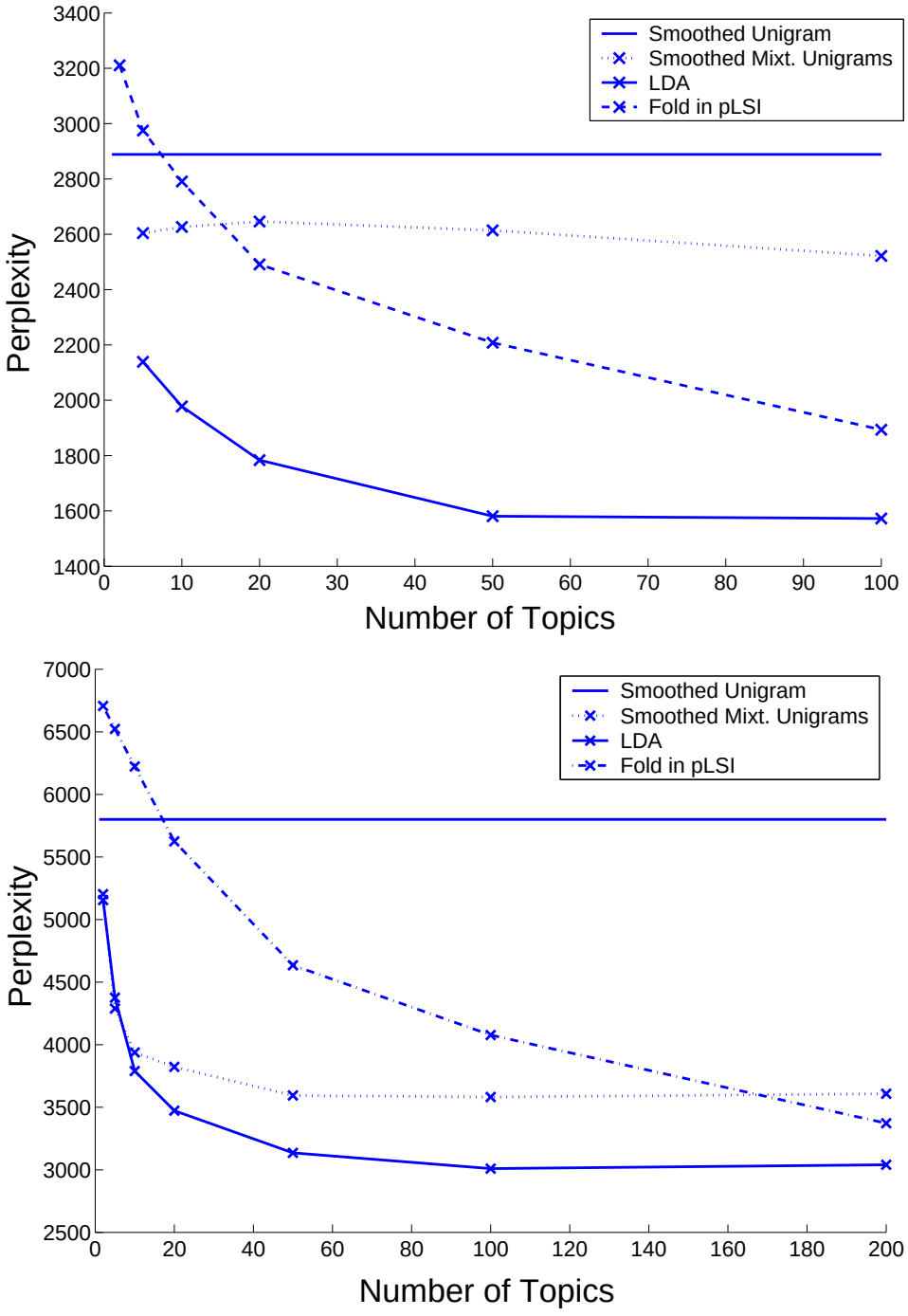


Figure 3.8: Perplexity results on the nematode (Top) and AP (Bottom) corpora for LDA, the unigram model, mixture of unigrams, and pLSI.

Num. topics (K)	Perplexity (Mult. Mixt.)	Perplexity (pLSI)
2	22,266	7,052
5	2.20×10^8	17,588
10	1.93×10^{17}	63,800
20	1.20×10^{22}	2.52×10^5
50	4.19×10^{106}	5.04×10^6
100	2.39×10^{150}	1.72×10^7
200	3.51×10^{264}	1.31×10^7

Table 3.1: Overfitting in the mixture of unigrams and pLSI models for the AP corpus. Similar behavior is observed in the nematode corpus (not reported).

Both pLSI and the mixture of unigrams suffer from serious overfitting issues, though for different reasons. This phenomenon is illustrated in Table 3.1. In the mixture of unigrams, overfitting is a result of peaked posteriors in the training set; a phenomenon familiar in the supervised setting, where this model is known as the naive Bayes model (Rennie, 2001). This leads to a nearly deterministic clustering of the training documents (in the E-step) which is used to determine the word probabilities in each mixture component (in the M-step). A previously unseen document may best fit one of the resulting mixture components, but will probably contain at least one word which did not occur in the training documents that were assigned to that component. Such words will have a very small probability, which causes a large increase in perplexity of the new document. As K increases, the documents of the training corpus are partitioned into finer collections, and thus induce more words with small probabilities.

In the mixture of unigrams, we can alleviate overfitting through the variational Bayesian smoothing scheme presented in Section 3.4.3. This ensures that all words will have some probability under every mixture component.

For pLSI, the problem of deterministic clusterings is alleviated by the fact that each document is allowed to exhibit a different proportion of topics. However, pLSI

only refers to the training documents and a different overfitting problem arises that is due to the dimensionality of the $p(z|d)$ parameter. One reasonable approach to assigning probability to a previously unseen document is by marginalizing over d:

$$p(\mathbf{w}) = \sum_d \prod_{n=1}^{N} \sum_z p(w_n \mid z) p(z \mid d) p(d).$$

Essentially, we are integrating over the empirical distribution on the topic simplex (see Figure 3.4).

This method of inference, though theoretically sound, causes the model to overfit. The document-specific topic distribution has some components which are close to zero for those topics that do not appear in the document. Thus, certain words will have very small probability in the estimates of each mixture component. When determining the probability of a new document through marginalization, only those training documents which exhibit a similar proportion of topics will contribute to the likelihood. For a given training document's topic proportions, any word which has small probability in all the constituent topics will increase the perplexity. As the number of topics gets larger, the chance that a training document will exhibit topics that cover all the words in the new document decreases (and perplexity grows). Note that pLSI does not overfit as quickly, with respect to the number of topics, as the mixture of unigrams.

This overfitting problem essentially stems from the restriction that each future document exhibit the same topic proportions as were seen in one or more of the training documents. Given this constraint, we are not free to choose the most likely proportions of topics for the new document. An alternative approach is the "folding-in" heuristic suggested by Hofmann (1999b), where one ignores the $p(z|d)$ parameters and refits $p(z|d_{\text{new}})$. In a sense, this gives the pLSI model an unfair advantage by allowing it to refit $K-1$ parameters to the test data. However, one can also interpret this procedure as MAP estimation of the posterior in LDA (Girolami and Kaban, 2003).

LDA suffers from neither of these problems. As in pLSI, each document can exhibit a different proportion of underlying topics. However, LDA can easily assign

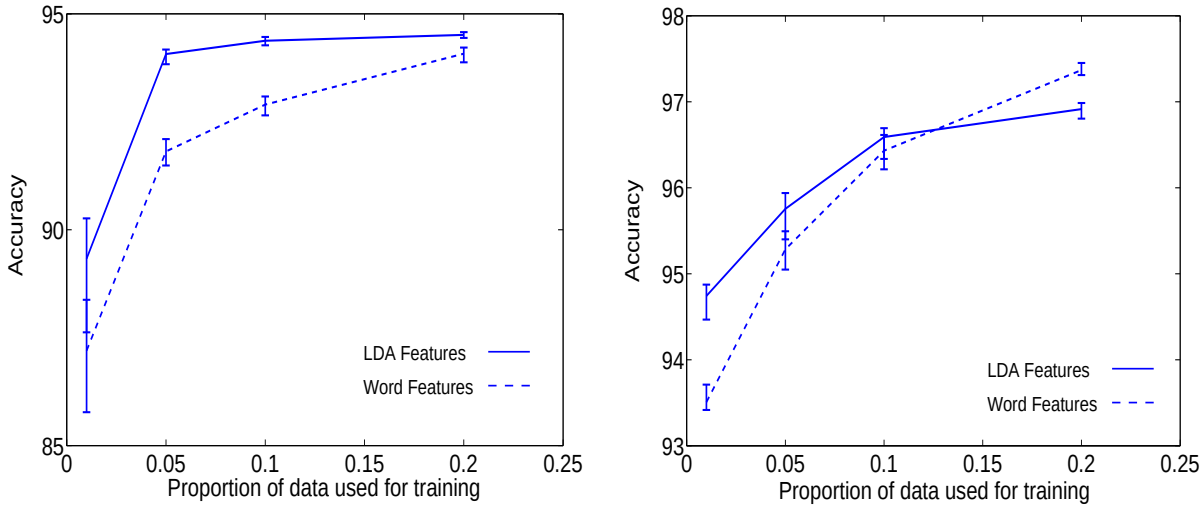


Figure 3.9: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

probability to a new document; no heuristics are needed for a new document to be endowed with a different set of topic proportions than were associated with documents in the training corpus.

Figure 3.8 presents the perplexity for each model on both corpora for different numbers of topics. The pLSI model and mixture of unigrams are suitably corrected for overfitting. The latent variable models perform better than the simple unigram model. LDA consistently performs better than the other models.

3.6.2 Document classification

The text classification problem is to automatically assign a document to a category. As in any classification problem, we may wish to consider generative approaches or discriminative approaches. In particular, by using one LDA module for each class, we obtain a generative model for classification. It is also of interest to use LDA in the discriminative framework, which is the focus of this section.

A challenging aspect of document classification problem is the choice of features. Treating individual words as features yields a rich but very large feature set (Joachims, 1999). One way to reduce this feature set is to use an LDA model for dimensionality reduction. In particular, LDA reduces any document to a fixed set of real-valued features—the posterior Dirichlet parameters $\gamma^*(\mathbf{w})$ associated with the document. It is of interest to see how much discriminatory information we lose in reducing the document description to these parameters.

We conducted two binary classification experiments using the Reuters-21578 dataset. The dataset contains 8000 documents and 15,818 words.

In these experiments, we estimated the parameters of an LDA model on all the documents, without reference to their true class label. We then fit a support vector machine (SVM) with the low-dimensional representations provided by LDA and compared this SVM to an SVM fit with all the word features.

Using the SVMLight software package (Joachims, 1999), we compared an SVM trained on all the word features with those trained on features induced by a 50-topic LDA model. Note that we reduce the feature space by 99.6 percent in this case.

Figure 3.9 shows our results. We see that there is little reduction in classification performance in using the LDA-based features; indeed, in almost all cases the performance is improved with the LDA features. Although these results need further substantiation, they suggest that the topic-based representation provided by LDA may be useful as a fast filtering algorithm for feature selection in text classification.

3.6.3 Collaborative filtering

The final LDA experiment uses the EachMovie collaborative filtering data. In this dataset, a collection of users indicates their preferred movie choices. A user and the movies chosen are analogous to a document and the words in the document (respectively).

The collaborative filtering task is as follows. We fit a model with a fully observed set of users. Then, for each unobserved user, we are shown all but one of the movies preferred by that user and are asked to predict what the held-out movie is. The

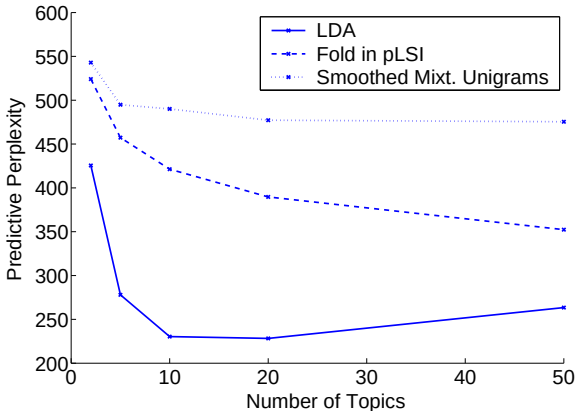


Figure 3.10: Results for collaborative filtering on the EachMovie data.

different algorithms are evaluated according to the likelihood they assign to the held-out movie. More precisely, define the predictive perplexity on M test users to be:

$$\text{predictive-perplexity} = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_{d,N_d} \mid \mathbf{w}_{d,1:N_d-1})}{M}\right\}.$$

We restricted the EachMovie dataset to users that positively rated at least 100 movies (a positive rating is at least four out of five stars). We divided this set of users into 3300 training users and 390 testing users.

Under the mixture of unigrams model, the probability of a movie given a set of observed movies is obtained from the posterior distribution of topics:

$$p(w|\mathbf{w}_{\text{obs}}) = \sum_{z} p(w|z)p(z|\mathbf{w}_{\text{obs}}).$$

In the pLSI model, the probability of a held-out movie is given by the same equation except that $p(z|\mathbf{w}_{\text{obs}})$ is computed by folding in the previously seen movies. Finally, in the LDA model, the probability of a held-out movie is given by integrating over the posterior Dirichlet:

$$p(w|\mathbf{w}_{\text{obs}}) = \int \sum_{z} p(w|z)p(z|\theta)p(\theta|\mathbf{w}_{\text{obs}})d\theta,$$

where $p(\theta \mid \mathbf{w}_{\text{obs}})$ is approximated by $q(\theta \mid \gamma(\mathbf{w}_{\text{obs}}))$, which is computed by variational inference. Note that this integral is tractable to compute, under the variational

distribution. We can interchange the sum and integral sign, and compute a linear combination of K Dirichlet expectations.

With a vocabulary of 1600 movies, we find the predictive perplexities illustrated in Figure 3.10. Again, the mixture of unigrams model and pLSI are corrected for overfitting, but the best predictive perplexities are obtained by the LDA model.

For a thorough treatment of probabilistic models for collaborative filtering problems, including developments with the LDA model, see Marlin (2004).

3.7 Discussion

We have described latent Dirichlet allocation, a flexible generative probabilistic model for collections of discrete data. LDA is based on a simple exchangeability assumption for the words and topics in a document; it is therefore realized by a straightforward application of de Finetti's representation theorem. We can view LDA as a dimensionality reduction technique, in the spirit of LSI, but with proper underlying generative probabilistic semantics that make sense for the type of data that it models. It is interesting to note that LDA-type models have been independently developed in a number of fields, including latent class analysis (Potthoff et al., 2000) and population genetics (Pritchard et al., 2000). For a good survey of the statistical literature, see Erosheva (2002).

It is worth noting that there are a large number of generalizations of the basic notion of exchangeability, including various forms of partial exchangeability, and that representation theorems are available for these cases as well (Diaconis, 1988). Thus, while the work which we discussed in the current chapter focuses on simple "bag-of-words" models, which lead to mixture distributions for single words (unigrams), these methods are trivially extendible to richer models that involve mixtures for larger structural units such as n-grams or paragraphs. Indeed, LDA has been extended to n-grams in Girolami and Kaban (2004).

Exact inference is intractable for LDA, but any of a suite of approximate inference algorithms can be used for inference and parameter estimation. We have

presented the convexity-based variational approach of Section 2.2.2 for approximate inference, showing that it yields a fast algorithm resulting in reasonable comparative performance in terms of test set likelihood. Other approaches that might be considered include Laplace approximation, higher-order variational techniques, and Monte Carlo methods. In particular, Leisink and Kappen (2002) have presented a general methodology for converting low-order variational lower bounds into higher-order variational bounds. It is also possible to achieve higher accuracy by dispensing with the requirement of maintaining a bound, and indeed Minka and Lafferty (2002) have shown that improved inferential accuracy can be obtained for the LDA model via a higher-order variational technique known as expectation propagation, however at greatly increased computational cost (Buntine and Jakulin, 2004). Finally, Griffiths and Steyvers (2002) have presented the Gibbs sampling method of Section 6.2.1 for smoothed LDA. Taking advantage of the conjugacy, they collapse the state-space of the Markov chain to only the latent topic allocation variables.

LDA is a simple model, and although we view it as a competitor to methods such as LSI and pLSI, it is also intended to be illustrative of the way in which probabilistic models can be scaled up to provide useful inferential machinery in domains involving multiple levels of structure. Indeed, the principal advantages of generative models such as LDA include their modularity and their extensibility. As a probabilistic module, LDA can be readily embedded in a more complex model—a property that is not possessed by LSI. As we will see in the subsequent chapters, LDA can readily be extended to model images and their captions, and hierarchies of topics. Furthermore, the assumption of a fixed number of topics can be relaxed in a nonparametric Bayes framework. Applied to LDA, the resulting models allow new documents to express previously unseen topics; this is a desirable property for building statistical models of large and growing document collections.

Chapter 4

Modeling annotated data

Traditional methods of information retrieval and text processing are organized around the representation and processing of a document in word-space. Modern multimedia documents, however, are not merely collections of words, but can be collections of related text, images, audio, and cross-references. When analyzing a collection of such documents, there is much to be gained from representations that explicitly model associations among the different types of data.

In this chapter, we extend the LDA model in two ways: first, we consider more general emission probabilities such as Gaussian image feature data, rather than multinomial word data; second, we extend the model to documents that consist of pairs of data streams. Our focus is on problems in which one data type is an *annotation* of the other data type. Examples of this kind of data include images and their captions, papers and their bibliographies, and genes and their functions. In addition to the traditional goals of retrieval, clustering, and classification, annotated data lends itself to tasks like automatic annotation and retrieval of unannotated data from annotation-type queries.

A number of recent papers have considered generative probabilistic models for such multi-type or *relational* data (Cohn and Hofmann, 2001; Taskar et al., 2001; Barnard et al., 2003; Jeon et al., 2003; Brochu and de Freitas, 2003). These papers have focused on models that jointly cluster the different data types, basing the clustering on latent variable representations that capture low-dimensional probabilistic

relationships among interacting sets of variables.

In many annotation problems, however, the overall goal appears to be that of finding a *conditional* relationship between types, and improved performance may be found in methods with a more discriminative flavor. In particular, the task of annotating an unannotated image can be viewed formally as a classification problem—for each word in the vocabulary we must make a yes/no decision.

Standard discriminative classification methods, however, make little attempt to uncover the probabilistic structure of either the input or output domain. This is ill-advised in the image/word setting because there are relationships among the words labeling an image, and these relationships reflect corresponding relationships among the regions in that image. Moreover, it is likely that capturing these relationships be helpful for annotating new images. Thus, with these issues in mind, we approach the annotation problem within a framework that exploits the best of both the generative and the discriminative traditions.

In this chapter, we develop a set of increasingly sophisticated models for a database of annotated images, culminating in *correspondence latent Dirichlet allocation* (Corr-LDA), a model that finds conditional relationships between latent variable representations of sets of image regions and sets of words. We show that Corr-LDA succeeds in providing both an excellent fit of the joint data and an effective conditional model of the caption given an image. We demonstrate its use in automatic image annotation, automatic region annotation, and text-based image retrieval.

4.1 Hierarchical models of image/caption data

We focus on estimating probabilistic models that can perform three types of inference. First, we would like to model the joint distribution of an image and its caption. This is useful for clustering, classification, and automatically organizing a multimedia database. Second, we would like to model the conditional distribution of words given an image. This is useful for automatic image annotation and text-based image retrieval. Finally, we would like to model the conditional distribution of words given a

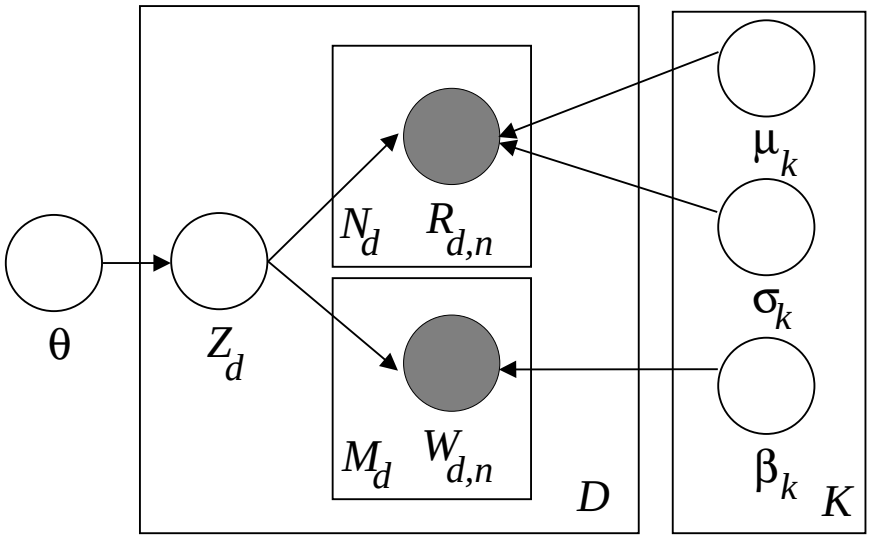


Figure 4.1: The GM-Mixture model of images and captions.

particular region of an image. This is useful for automatically labeling and identifying the objects of an image.

Following Barnard et al. (2003), each image is segmented into regions by the N-cuts algorithm (Shi and Malik, 2000). For each region, we compute a set of 47 real-valued features representing visual properties such as size, position, color, texture, and shape. Analogous to the representation of documents (see Section 3.1), each image and its corresponding caption are a pair $(\mathbf{r}, \mathbf{w})$. The first element $\mathbf{r} = r_{1:N}$ is a collection of N feature vectors associated with the regions of the image. The second element $\mathbf{w} = w_{1:M}$ is the collection of M words of the caption.

We consider hierarchical probabilistic models of image/caption data which involve mixtures of the underlying discrete and continuous variables. Conditional on the values of these latent variables, the region feature vectors are assumed distributed by a multivariate Gaussian with diagonal covariance, and the caption words are assumed distributed by a multinomial distribution over the vocabulary.

4.1.1 Gaussian-multinomial mixture

We begin by considering a simple finite mixture model—the model underlying most previous work on the probabilistic modeling of multi-type data (Taskar et al., 2001; Barnard et al., 2003; Brochu and de Freitas, 2003). In this model—the *Gaussian-multinomial mixture* (GM-Mixture) shown in Figure 4.1—a single discrete latent factor is used to represent a joint clustering of an image and its caption.[1] For a K-factor GM-Mixture model, an image/caption is assumed drawn from the following generative process:

1. Choose factor $Z \mid \theta \sim \text{Mult}(\theta)$.
2. For $n \in \{1, \ldots, N\}$, choose region description $R_n \mid \{z, \mu_{1:K}, \sigma_{1:K}\} \sim \text{Normal}(\mu_z, \sigma_z)$.
3. For $m \in \{1, \ldots, M\}$, choose word $W_m \mid \{z, \beta_{1:K}\} \sim \text{Mult}(\beta_z)$.

The joint distribution of the latent and observed variables is thus:

$$p(z, \mathbf{r}, \mathbf{w} \mid \theta, \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}) = p(z \mid \theta) \prod_{n=1}^{N} p(r_n \mid z, \mu_{1:K}, \sigma_{1:K}) \prod_{m=1}^{M} p(w_m \mid z, \beta_{1:K}). \tag{4.1}$$

Given a collection of images/captions and the choice of a number of factors K, the parameters of a GM-Mixture model can be estimated by the EM algorithm or, in a Bayesian setting, a variational inference procedure. This yields a set of Gaussian distributions on features and multinomial distributions on words which describe a K-clustering of the images/captions. Since each image/caption is assumed generated conditional on the same factor, the resulting multinomial and Gaussian parameters will correspond: an image with high probability under a certain Gaussian distribution will likely contain a caption with high probability under the corresponding multinomial distribution.

Consider the three distributions of interest under this model. First, the joint probability of an image/caption is obtained by marginalizing out the latent factor

[1] In Chapter 3, we descriptively referred to this variable as a *topic* to aid intuitions about the underlying probabilistic assumptions; here, we proceed with *factor*, the more common statistical terminology.

from Eq. (4.1):

$$p(\mathbf{r}, \mathbf{w} \mid \theta, \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}) = \sum_z p(z, \mathbf{r}, \mathbf{w} \mid \theta \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}).$$

Second, the conditional distribution of words given an image is obtained by marginalizing out the latent factor, conditional on the image:

$$p(w \mid \mathbf{r}, \theta, \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}) = \sum_z p(z \mid \mathbf{r}, \theta, \mu_{1:K}, \sigma_{1:K}) p(w \mid z, \beta_{1:K}),$$

where we compute the posterior of the latent topic using Bayes rule:

$$p(z \mid \mathbf{r}, \theta, \mu_{1:K}, \sigma_{1:K}) \propto p(z \mid \theta) p(\mathbf{r} \mid z, \mu_{1:K}, \sigma_{1:K}).$$

Finally, we would like to compute a region-specific distribution of words. This task, however, is beyond the scope of the GM-Mixture model. Conditional on the latent factor, regions and words are generated independently, and the correspondence between specific regions and specific words is necessarily ignored. The mixture model essentially treats collections of regions and collections of words as a single set of average image features and word counts. A model for region correspondence needs to explicitly model the individual words and regions so that it can identify the differences between them.

4.1.2 Gaussian-multinomial LDA

Gaussian-multinomial LDA (GM-LDA) is illustrated in Figure 4.2 and extends the GM-Mixture in the same way that LDA extends the mixture of unigrams. For a K-factor GM-LDA model, an image/caption is assumed drawn from the following generative process:

1. Choose $\theta \mid \alpha \sim \text{Dir}(\alpha)$.
2. For $n \in \{1, \ldots, N\}$:
   (a) Choose factor $Z_n \mid \theta \sim \text{Mult}(\theta)$.
   (b) Choose region description $R_n \mid \{z_n, \mu_{1:K}, \sigma_{1:K}\} \sim \text{Normal}(\mu_{z_n}, \sigma_{z_n})$.

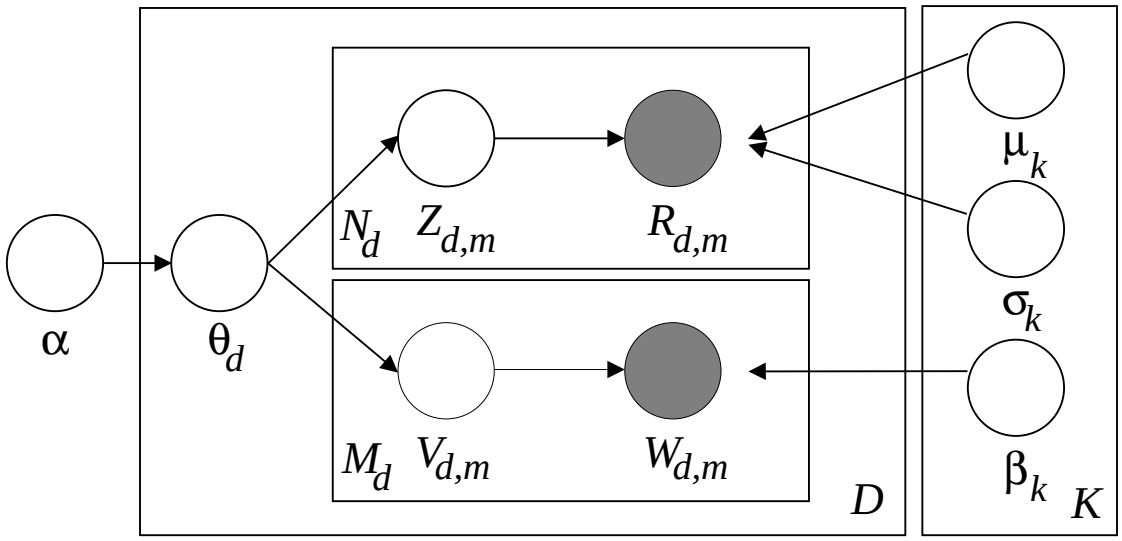


Figure 4.2: The GM-LDA model of images and captions. Unlike GM-Mixture (Figure 4.1), each word and image region is potentially drawn from a different latent factor.

3. For $m \in \{1, \ldots, M\}$:

   (a) Choose factor $V_m \,|\, \theta \sim \text{Mult}(\theta)$.

   (b) Choose word $W_m \,|\, \{v_m, \beta_{1:K}\} \sim \text{Mult}(\beta_{v_m})$.

The resulting joint distribution of image regions, caption words, and latent variables is:

$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{v} \,|\, \alpha, \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}) = p(\theta \,|\, \alpha) \left(\prod_{n=1}^{N} p(z_n \,|\, \theta) p(r_n \,|\, z_n, \mu_{1:K}, \sigma_{1:K}) \right) \left(\prod_{m=1}^{M} p(v_m \,|\, \theta) p(w_m \,|\, v_m, \beta_{1:K}) \right).$$

As in the simpler LDA model, posterior inference is intractable, and we employ the mean-field variational inference method from Section 2.2.2 for approximate inference. Moreover, we can use the resulting variational distributions to find the conditional probabilities needed for image annotation/retrieval and region labeling. See Barnard et al. (2003) for the details of this model.

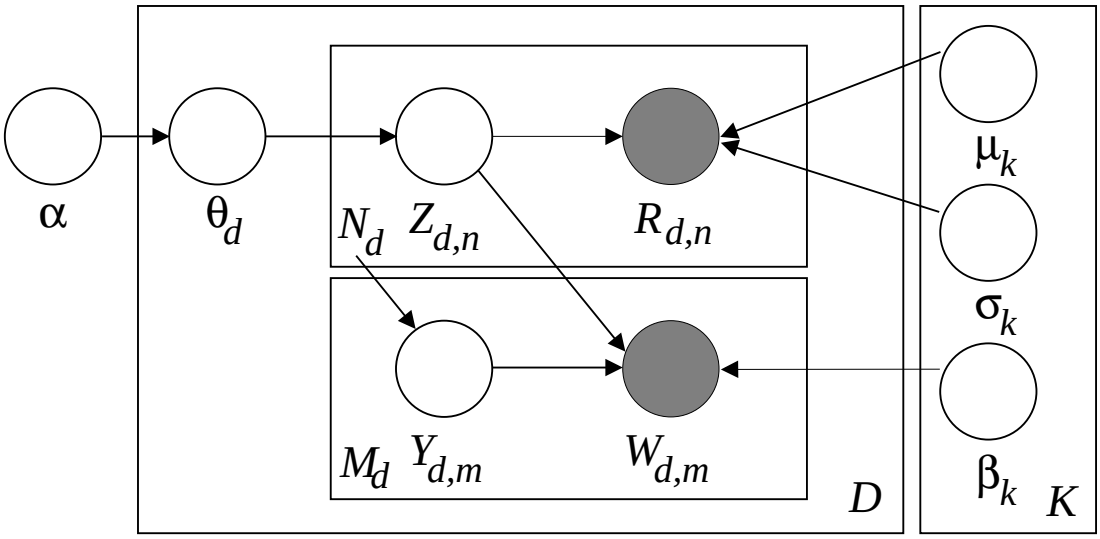


Figure 4.3: The graphical model representation of the CORR-LDA model. Note that the variables y_m are conditioned on N, the number of image regions.

We have demonstrated in Section 3.6 that LDA provides significant improvements in predictive performance over simpler mixture models, and we expect for GM-LDA to provide similar advantages over GM-MIXTURE. Indeed, we will see in Section 4.2.1 that GM-LDA does model the image/caption data better than GM-MIXTURE. We will also see, however, that good models of the joint probability of images/captions do not necessarily yield good models of the *conditional* probabilities that are needed for automatic annotation, text-based image retrieval, and region labeling. This is due to the lack of dependency between the latent factors $Z_{1:N}$ and $V_{1:M}$ which respectively generated the images and their captions. In the next section, we turn to a model that aims to correct this problem.

4.1.3 Correspondence LDA

We describe *correspondence LDA* (CORR-LDA) as a model that combines the flexibility of GM-LDA with the associability of GM-MIXTURE. This model performs dimensionality reduction in the representation of region descriptions and words, while also modeling the conditional correspondence between their respective reduced rep-

resentations.

CORR-LDA is depicted in Figure 4.3. The model can be viewed in terms of a generative process that first generates the region descriptions from a Gaussian LDA model. Then, for each of the caption words, one of the regions is selected from the image and a corresponding caption word is drawn, conditional on the same factor that generated the selected region.

Denote the ensemble of latent factors associated with the image by $\mathbf{Z} = Z_{1:N}$, and let $\mathbf{Y} = Y_{1:M}$ be discrete indexing variables that take values from 1 to N. A K-factor CORR-LDA model assumes the following generative process of an image/caption:

1. Choose $\theta \,|\, \alpha \sim \text{Dir}(\alpha)$.
2. For $n \in \{1, \ldots, N\}$:
   (a) Choose factor $Z_n \,|\, \theta \sim \text{Mult}(\theta)$.
   (b) Choose region description $R_n \,|\, \{z_n, \mu_{1:K}, \sigma_{1:K}\} \sim \text{Normal}(\mu_{z_n}, \sigma_{z_n})$
3. For $m \in \{1, \ldots, M\}$:
   (a) Choose region index $Y_m \,|\, N \sim \text{Unif}(1, \ldots, N)$.
   (b) Choose word $W_m \,|\, \{y_m, \mathbf{z}, \beta_{1:K}\} \sim \text{Mult}(\beta_{z_{y_m}})$

CORR-LDA thus specifies the following joint distribution of image regions, caption words, and latent variables:

$$p(\mathbf{r}, \mathbf{w}, \theta, \mathbf{z}, \mathbf{y} \,|\, \alpha, \beta_{1:K}, \mu_{1:K}, \sigma_{1:K}) = p(\theta \,|\, \alpha) \left(\prod_{n=1}^{N} p(z_n \,|\, \theta) p(r_n \,|\, z_n, \mu_{1:K}, \sigma_{1:K}) \right) \left(\prod_{m=1}^{M} p(y_m \,|\, N) p(w_m \,|\, y_m, \mathbf{z}, \beta_{1:K}) \right).$$

The independence assumptions of CORR-LDA are a compromise between the total correspondence enforced by the GM-MIXTURE , where the entire image and caption are conditional on the same factor, and the lack of correspondence in GM-LDA , where the image regions and caption words can conceivably be associated with two disparate sets of factors. Under the CORR-LDA model, the regions of the image

can be associated with any ensemble of factors, but the words of the caption may only be associated with factors that are present in the image. In effect, this model captures the notion that the image is generated first, and the caption subsequently annotates the image.

Finally, note that the correspondence implemented by CORR-LDA is not a one-to-one correspondence; multiple caption words may be associated with the same region, and some regions may not have any caption words associated with them.

Variational inference

Exact probabilistic inference for CORR-LDA is intractable. As before, we avail ourselves of mean-field variational inference to approximate the posterior distribution of the latent variables given an image/caption.

Define the following factorized variational distribution of the latent variables:

$$q(\theta, \mathbf{z}, \mathbf{y} \,|\, \gamma, \phi_{1:N}, \lambda_{1:M}) = q(\theta \,|\, \gamma) \left(\prod_{n=1}^{N} q(z_n \,|\, \phi_n) \right) \left(\prod_{m=1}^{M} q(y_m \,|\, \lambda_m) \right).$$

The variational parameters are a K-dimensional Dirichlet parameter γ, N K-dimensional multinomial parameters $\phi_{1:N}$, and M N-dimensional multinomial parameters $\lambda_{1:N}$.

Following the mean-field variational inference algorithm of Section 2.2.2, we minimize the KL-divergence between this factorized distribution and the true posterior by the following coordinate ascent algorithm:

$$\begin{aligned} \gamma_i &= \alpha_i + \textstyle\sum_{n=1}^{N} \phi_{ni} \\ \lambda_{mn} &\propto \exp\left\{ \textstyle\sum_{i=1}^{K} \phi_{ni} \log p(w_m \,|\, y_m = n, z_n = i, \beta) \right\} \\ \phi_{ni} &\propto \exp\{S_{ni}\}, \end{aligned}$$

where:

$$S_{ni} = \log p(r_n \,|\, z_n = i, \mu_{1:K}, \sigma_{1:K}) + \mathrm{E}_q\left[\log \theta_i \,|\, \gamma\right] + \sum_{m=1}^{M} \lambda_{mn} \log p(w_m \,|\, y_m = n, z_m = i, \beta_{1:K}).$$

Note that this update takes into account the likelihood, for each caption word, that it was generated by the factor associated with this region.

With the approximate posterior in hand, we can find a lower bound on the joint probability and compute the conditional distributions of interest. For annotation, we fit the variational distribution for the image alone and compute:

$$p(w \mid \mathbf{r}, \alpha, \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}) \approx \sum_{n=1}^{N} \sum_{z_n} q(z_n \mid \phi_n) p(w \mid z_n, \beta).$$

For region labeling, the distribution of words conditional on an image and a region is approximated by:

$$p(w \mid \mathbf{r}, r_n, \alpha, \mu_{1:K}, \sigma_{1:K}, \beta_{1:K}) \approx \sum_{z_n} q(z_n \mid \phi_n) p(w \mid z_n, \beta).$$

Empirical and variational Bayes

Given a collection of image/caption data, we find empirical Bayes point estimates of the model parameters with a variational EM procedure that maximizes the lower bound on the log likelihood of the data induced by the variational approximation described above. Similar to the LDA model (see Section 3.4.2), the E-step computes the variational posterior for each image/caption given the current setting of the parameters. The M-step subsequently finds maximum likelihood estimates of the model parameters using expected sufficient statistics under the variational distribution. The variational EM algorithm alternates between these two steps until the bound on the log likelihood converges.

In the results of Section 4.2.1, we show that overfitting can be a serious problem, particularly when working with the conditional distributions for image annotation. To address this issue, we take a more fully Bayesian approach by placing the conjugate prior distribution on the word multinomial parameters $\beta_{1:K}$. The corresponding variational inference algorithm is practically identical to that for smoothed LDA (see Section 3.4.3). All instances of β_k in the document-specific variational updates are replaced by $\exp\{\mathrm{E}[\log \beta_k \mid \rho]\}$, and we update the corpus-wide variational parameters as for LDA (see Eq. 3.13).

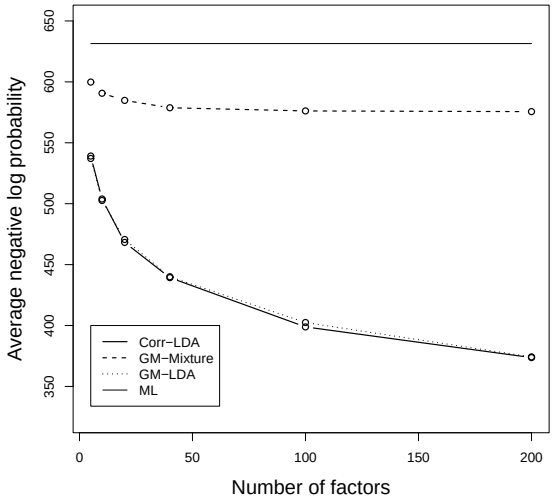


Figure 4.4: The per-image average negative log probability of the held-out test set as a function of the number of latent factors (lower numbers are better). The horizontal line is the model that treats the regions and captions as an independent Gaussian and multinomial, respectively.

4.2 Empirical results

We present an evaluation of the three image/caption models on 7000 images and captions from the Corel database. We held out 25% of the data for testing purposes and used the remaining 75% to estimate parameters (or their posterior). Each image is segmented into 6-10 regions and associated with 2-4 caption words. The vocabulary contains 168 unique terms.

4.2.1 Test set likelihood

To evaluate how well each model fits the observed data, we computed the per-image average negative log likelihood of the test set on all three models for different numbers of factors. A model which better fits the data will assign a higher likelihood to the test set (i.e., lower numbers are better in negative likelihood).

Figure 4.2.1 illustrates the results. Consistent with the results on text data of Sec-

tion 3.6, GM-LDA provides a better fit than GM-MIXTURE. Furthermore, CORR-LDA provides as good a fit as GM-LDA. This is surprising because GM-LDA is a less constrained model. However, both models have the same number of parameters; their similar performance indicates that, on average, the factors needed to model a particular image are adequate to model its caption. Empirically, in a 200-factor model we find that in only two images of the test set does GM-LDA uses more latent factors for the caption than image.

4.2.2 Caption perplexity

Given a segmented image without its caption, we can use the models described above to compute a distribution of words conditioned on the image. This distribution reflects a prediction of the missing caption words for that image.

To measure the annotation quality of the models, we computed the perplexity of the captions under the conditional word distribution for each image in the test set:

$$\text{perplexity} = \exp\left\{ -\sum_{d=1}^{D} \sum_{m=1}^{M_d} \log p(w_m \mid \mathbf{r}_d) / \sum_{d=1}^{D} M_d \right\}.$$

Figure 4.5 (Left) shows the perplexity of the held-out captions under the maximum likelihood estimates of each model for different numbers of factors. We see that overfitting is a serious problem in the GM-MIXTURE model, and its perplexity immediately grows off the graph (e.g., for the 200-factor model, the perplexity is 2922). In related work, Barnard et al. (2003) consider several variants of GM-MIXTURE and rely heavily on ad-hoc smoothing to correct for overfitting.

Figure 4.5 (Right) illustrates the caption perplexity under the smoothed estimates of each model using an empirical Bayes procedure. We place a conjugate prior on the multinomial parameters (see Section 3.4.3), and then find maximum likelihood estimates of the corresponding hyperparameter. The overfitting of GM-MIXTURE has been corrected. Once smoothed, it performs better than GM-LDA despite that model's superior performance in joint likelihood.

We found that GM-LDA does not provide good conditional distributions for two reasons. First, it is "over-smoothed." Computing the conditional word distribution

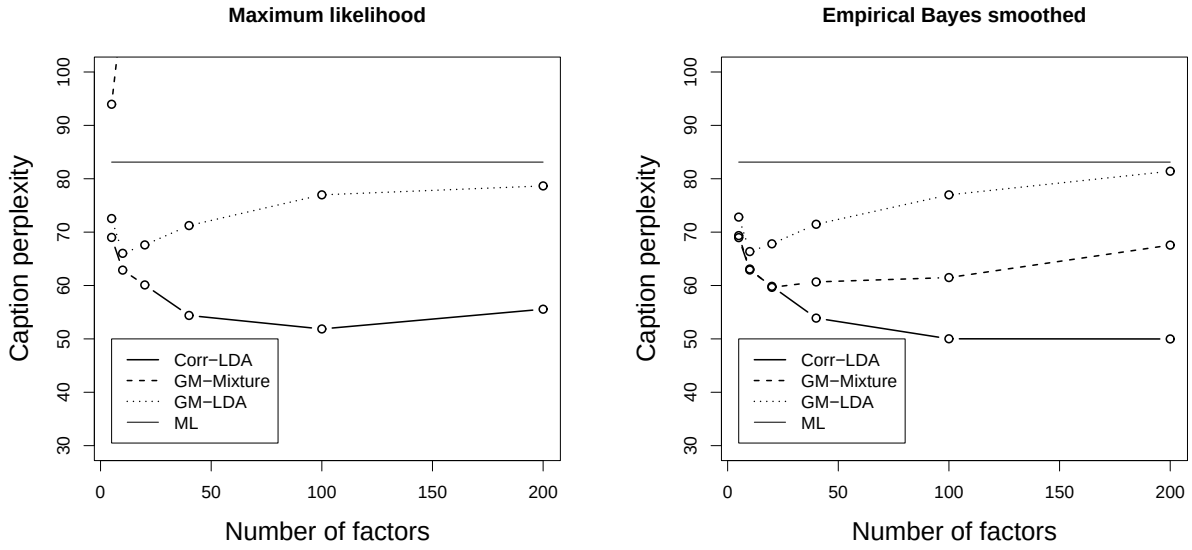


Figure 4.5: (Left) Caption perplexity on the test set for the ML estimates of the models (lower numbers are better). (Right) Caption perplexity for the empirical Bayes smoothed estimates of the models.

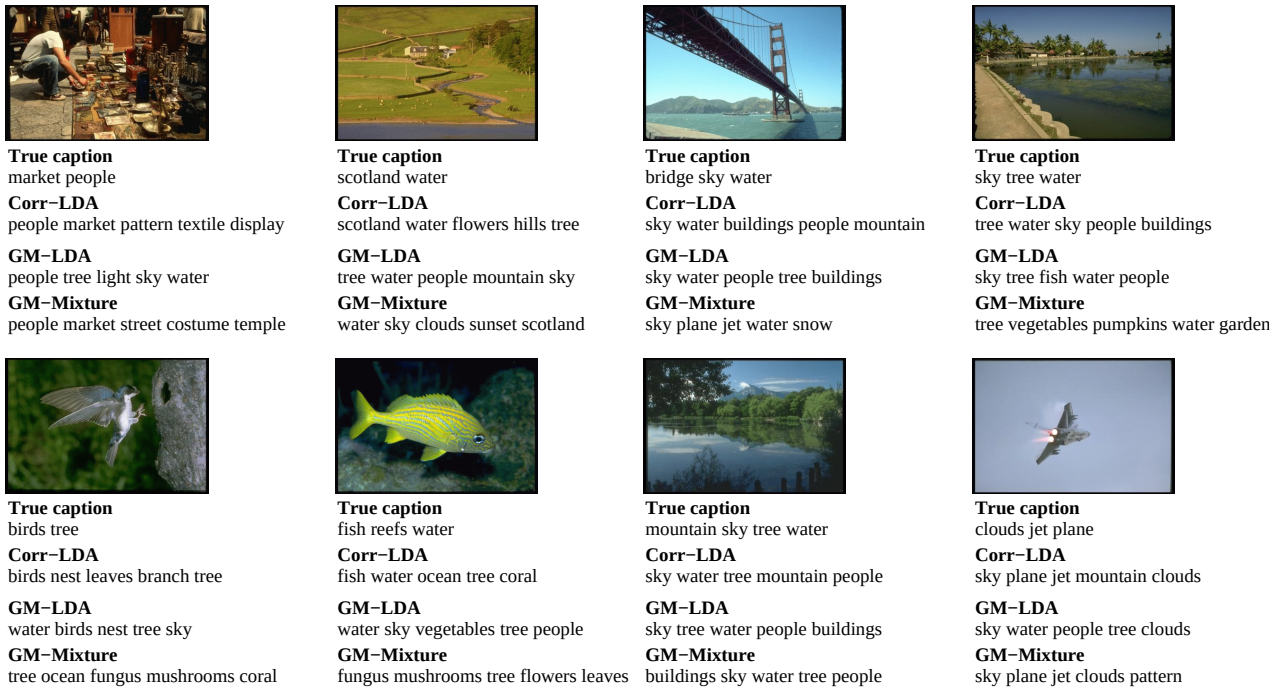


Figure 4.6: Example images from the test set and their automatic annotations under different models.

requires integrating a diffuse posterior (due to the small number of regions) over all the factor dimensions. Thus, the factors to which each region is associated are essentially washed out and, as the number of factors gets large, the model's performance approaches the performance of the simple maximum likelihood estimate of the caption words.

Second, GM-LDA easily allows caption words to be generated by factors that did not contribute to generating the image regions (e.g., in a 200-factor model, 54% of the caption words in the test set are assigned to factors that do not appear in their corresponding images). With this freedom, the estimated conditional Gaussian parameters do not necessarily reflect regions that are correctly annotated by the corresponding conditional multinomial parameters. While it better models the joint distribution of words and regions, it fails to model the relationship between them.

Most notably, Corr-LDA provides better predictive distributions of words than either GM-LDA or GM-Mixture. It provides as flexible a joint distribution as GM-LDA , but guarantees that the conditional Gaussian distributions for image regions correspond with the conditional multinomials for caption words. Furthermore, by allowing caption words to be allocated to different factors, Corr-LDA achieves superior performance to the GM-Mixture which is constrained to associate the entire image/caption with a single factor. Thus, Corr-LDA achieves a competitive fit of the joint distribution, and finds better conditional distributions of words given images.

4.2.3 Annotation examples

We can provide an automatic annotation of an unannotated image, by choosing the top five words from its conditional distribution of words. Figure 4.6 shows ten sample annotations computed by each model with 200 factors. These examples illustrate the power and limitations the models when used for a practical discriminative task.

As shown quantitatively, the GM-LDA model gives the least impressive performance of the three. We see the washing out effect described above by the fact that many of the most common words in the corpus—words like "water" and "sky"—occur

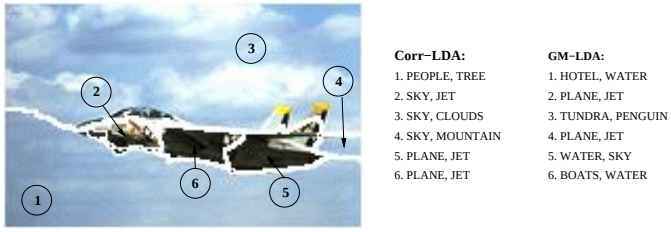


Figure 4.7: An example of automatic region labeling.

in the predicted captions of all of the images. Moreover, the predicted caption rarely contains the correct words for the objects in the image. For example, it misses "jet" in the picture captioned *clouds, jet, plane*, a word that both other models are able to predict.

The GM-Mixture model performs better than GM-LDA, but we can see how this model relies on the average image features and fails to predict words for regions that may not have occurred in other similar images. For example, it omits "tree" from the picture captioned *scotland, water* since the trees are only a small part on the left side of the image. Furthermore, the GM-Mixture predicts incorrect words if the average features do not easily correspond to a common theme. For example, the background of *fish, reefs, water* is not the usual blue, and the mixture model predicts words like "fungus", "tree", and "flowers."

Finally, as reflected by the perplexity results, Corr-LDA gives the best performance and correctly labels most of the example pictures. Unlike the GM-Mixture model, it can assign each region to a different cluster, and the final distribution of words reflects the ensemble of clusters which were assigned to the image regions. Thus, Corr-LDA finds the trees in the picture labeled *scotland, water* and can correctly identify the fish, even without its usual blue background.

As described in Section 4.1, the probabilistic structure of Corr-LDA and GM-LDA allow the computation of a meaningful region-based distribution of words. Figure 4.7 illustrates a sample region labeling on an image from the test set.[2] Though

[2]We cannot quantitatively evaluate this task (i.e., compute the region perplexity) because our data does not provide ground-truth for the region labels.

both models hypothesize the word "plane" and "jet," CORR-LDA places them in the reasonable regions 2, 5, and 6 while GM-LDA places them in regions 2 and 4. Furthermore, CORR-LDA recognizes the top region as "sky, clouds", while GM-LDA provides the enigmatic "tundra, penguin."

4.2.4 Text-based image retrieval

There has been a significant amount of computer science research on *content-based image retrieval*, in which a particular query image , possibly a sketch or primitive graphic, is used to find matching relevant images (Goodrum, 2000; Wang et al., 2001). In another line of research, *multimedia information retrieval*, representations of different data types such as text and images are used to retrieve documents that contain both (Meghini et al., 2001).

Less attention, however, has been focused on *text-based image retrieval*, an arguably more difficult task where a user submits a text query to find matching images for which there is no related text. Previous approaches have essentially treated this task as a classification problem, handling specific queries from a vocabulary of about five words (Naphade and Huang, 2001). In contrast, by using the conditional distribution of words given an image, our approach can handle arbitrary queries from a large vocabulary.

We adapt the language modeling approach of information retrieval (Ponte and Croft, 1998) by deriving the document-specific language models from images rather than words. For each unannotated image, we obtain the conditional distribution of words, and use that distribution to score the query.

Denote an set of query words by $\mathbf{q} = q_{1:N}$. The score of each image, relative to the query, is the conditional probability of the query:

$$p(\mathbf{q} \mid \mathbf{r}_i) = \prod_{n=1}^{N} p(q_n \mid \mathbf{r}_i),$$

where $p(q_n \mid \mathbf{r}_i)$ is the probability of the nth query word under the distribution $p(w \mid \mathbf{r}_i)$. After computing the score for each image, we return a list of images ranked in descending order by conditional likelihood.

Figure 4.8 illustrates three queries performed on the three models with 200 factors and the held-out test set of images. We consider an image to be relevant if its true caption contains the query words (recall that we make no reference to the true caption in the retrieval process). As illustrated by the precision/recall curves, the Corr-LDA model provides superior retrieval performance. It is particularly strong with difficult queries such as "people and fish." In that example, there are only six relevant images in the test set and Corr-LDA places two of them in the top five. This is due to its ability to assign different regions to different clusters. The model can independently estimate the salient features of "fish" and "people", and effectively combine them to perform retrieval.

4.3 Discussion

In this chapter, we have developed Corr-LDA, a powerful model for annotated data that combines the advantages of probabilistic clustering for dimensionality reduction with an explicit model of the conditional distribution from which data annotations are generated. In the setting of image/caption data, we have shown that this model can achieve a competitive joint likelihood and superior conditional distribution of words given an image.

Corr-LDA provides a clean probabilistic model for performing various tasks associated with multi-type data such as images and their captions. We have demonstrated its use in automatic image annotation, automatic image region annotation, and text-based image retrieval. It is important to note that this model is not specially tailored for image/caption data. Given an appropriate parameterization of the data likelihood, Corr-LDA can be applied to any kind of annotated data such as video/closed-captions, music/text, and gene/functions.

Again, our choice of approximate inference technique is motivated by the applicability of this model to large datasets, and the need for computational efficiency in such domains. While mean-field variational inference is convenient in this setting, the usual suite of other methods may also be used.

More importantly, the development of Corr-LDA and GM-LDA illustrates how simple models can be elegantly extended for the analysis of complicated data, such as images and captions. However, the comparative results show that such extensions should be carefully constructed, with the underlying exchangeability assumptions in mind.

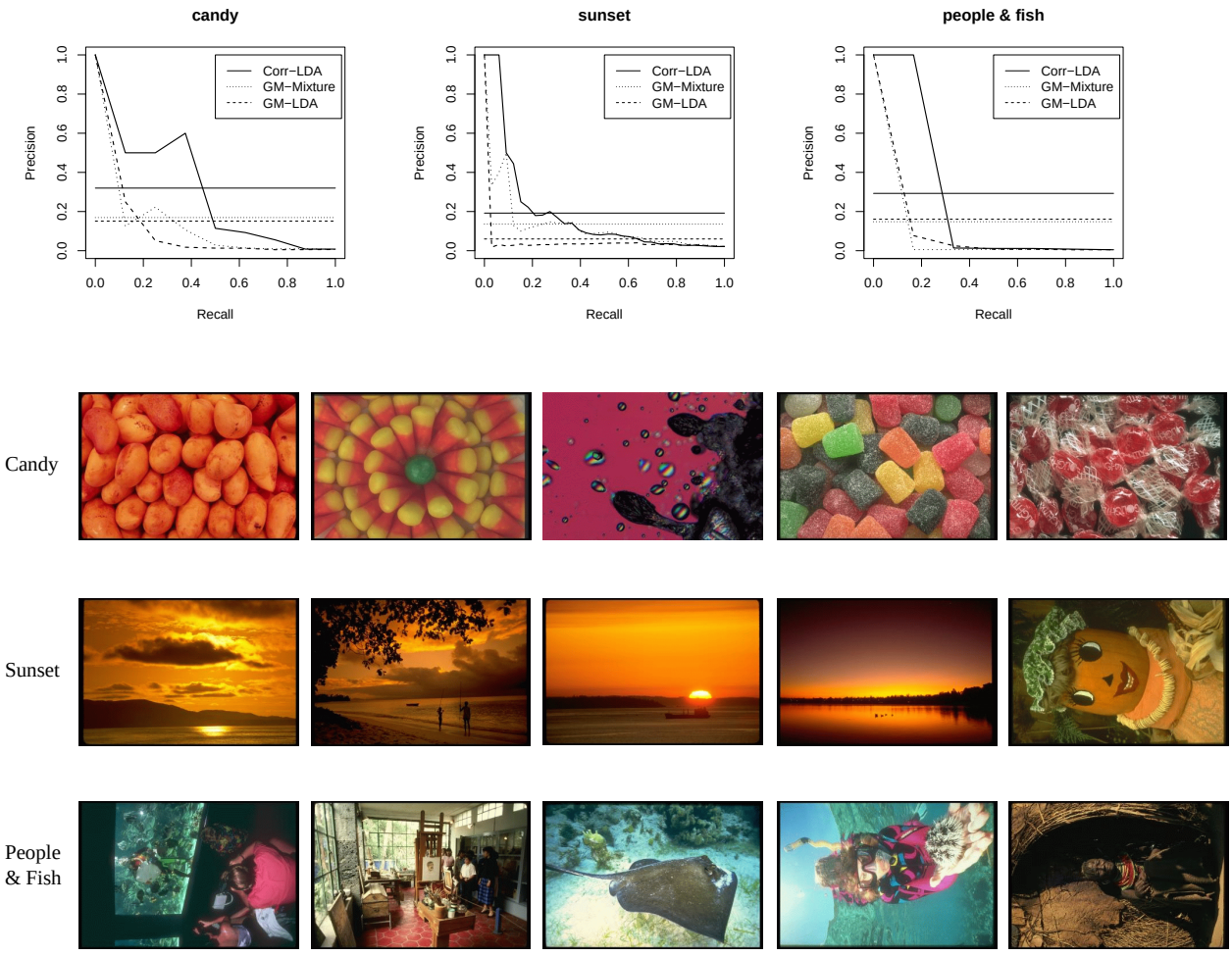


Figure 4.8: Three examples of text-based image retrieval. (Top) Precision/recall curves for three queries on a 200-factor Corr-LDA model. The horizontal lines are the mean precision for each model. (Bottom) The top five returned images for the same three queries.

Chapter 5

Nonparametric Bayesian inference

The models of Chapters 3 and 4 are extensions of mixture models with a fixed number of factors (or topics). In the data analysis, we estimated models with different numbers of factors, and chose one based on held-out likelihood.

Choosing and fixing the number of factors is troublesome for two reasons. First, it is difficult to select an appropriate criterion and, once selected, optimizing with respect to that criterion can be expensive.

Second, many data modeling domains have an open ended nature—data sets often grow over time, and as they grow they bring new entities and new structures to the fore. In this thesis we focus on text and image data which is collected from a continuing stream of information. We aim to use our models to generalize to future data, and it is natural to expect that future text and images might reflect structure that was previously unseen in the original dataset. Thus, we would like to fit flexible models for which the number of factors can grow as new data is observed.

In this chapter, we develop models for which the number of factors is a random variable that depends on how much data has been seen. Implicitly, a dataset induces a posterior distribution of this number, but future data may be representative of previously unseen factors.

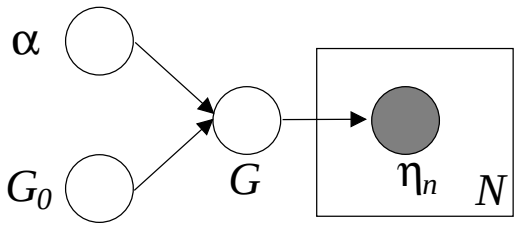


Figure 5.1: Graphical model representation of the Dirichlet process.

5.1 The Dirichlet process

Recall de Finetti's representation theorem, which provides critical justification for Bayesian modeling (de Finetti, 1990). If $X_{1:N}$ are exchangeable random variables, then their joint distribution can be written as:

$$p(X_{1:N}) = \int \left(\prod_{n=1}^{N} G(X_n) \right) dP(G),$$

where G is a measure on X and $p(G)$ is a measure on measures. Typically, we assume a parametric form of the density of x, and thus $p(G)$ is a measure on the underlying parameter. The *Dirichlet process* provides a more general measure on measures, lifting the restriction of G to a particular parametric family.

Let η be a continuous random variable, G_0 be a non-atomic probability distribution for η, and α be a positive, real-valued scalar. A random measure G is distributed according to a Dirichlet process (DP), with scaling parameter α and base measure G_0, if for all natural numbers k and k-partitions $\{B_1, \ldots, B_k\}$:

$$(G(\eta \in B_1), G(\eta \in B_2), \ldots, G(\eta \in B_k)) \sim \mathrm{Dir}(\alpha G_0(B_1), \alpha G_0(B_2), \ldots, \alpha G_0(B_k)). \tag{5.1}$$

Ferguson (1973) proves the existence of such a distribution via these finite dimensional distributions and the Kolmogorov consistency theorem.

Figure 5.1 illustrates $\eta_{1:N}$ drawn iid from a random distribution G. If G is distributed according to a Dirichlet process, then $\eta_{1:N}$ are drawn from the following process:

1. Choose $G \,|\, \{\alpha, G_0\} \sim DP(\alpha, G_0)$

2. For $n \in \{1, \ldots, N\}$, choose $\eta_n \,|\, G \sim G$.

One important property of the DP is that the marginal distribution of η_n is G_0. Consider any k and any k-partition B:

$$\begin{aligned} p(\eta \in B_i) &= \int p(\eta \in B_i \,|\, G) p(G \,|\, G_0, \alpha) dG \\ &= \frac{\alpha G_0(B_i)}{\sum_{j=1}^{k} \alpha G_0(B_j)} \\ &= G_0(B_i), \end{aligned}$$

where we have used the expectation of the Dirichlet distribution in Eq. (3.2).

Another important property of the DP is that its posterior, conditional on a draw from G, is a DP with a point mass added to the base measure. Suppose we draw a sample η_1 from G. Again choose any k and k-partition B. From the posterior Dirichlet distribution of Eq. (3.3) and finite-dimensional distribution of Eq. (5.1), it follows that:

$$(G(\eta \in B_1), \ldots, G(\eta \in B_k)) \,|\, \eta_1 \sim \mathrm{Dir}(\alpha G_0(B_1) + \delta_{\eta_1}(B_1), \ldots, \alpha G_0(B_k) + \delta_{\eta_1}(B_k)).$$

Since this is true for all k and k-partitions, we conclude that:

$$G \,|\, \{\eta_1, \alpha, G_0\} \sim DP(\alpha, G_0 + \delta_{\eta_1}(\cdot)). \tag{5.2}$$

Now consider N samples $\eta_{1:N}$ from G. It follows that the posterior distribution of G is:

$$G \,|\, \{\eta_{1:N}, \alpha, G_0\} \sim DP(\alpha, G_0 + \textstyle\sum_{n=1}^{N} \delta_{\eta_n}(\cdot)), \tag{5.3}$$

and the corresponding marginal distribution of η is:

$$p(\eta \,|\, \eta_{1:N}) \propto \alpha G_0(\eta) + \textstyle\sum_{n=1}^{N} \delta_{\eta_n}(\eta) \tag{5.4}$$

This is known as the *clustering effect*. The variable η will either be equal to one of the previously drawn values, or an independent draw from G_0.

5.1.1 Pólya urns and the Chinese restaurant process

The clustering effect provides a useful representation of the joint marginal distribution of $\eta_{1:N}$:

$$p(\eta_{1:N} \,|\, \alpha, G_0) = \int \left(\prod_{n=1}^{N} p(\eta_n \,|\, G) \right) p(G \,|\, \alpha, G_0). \tag{5.5}$$

Using the chain rule, this distribution can be written as the product:

$$p(\eta_{1:N} \,|\, \alpha, G_0) = \prod_{n=1}^{N} p(\eta_n \,|\, \eta_{1:(n-1)}, \alpha, G_0),$$

which, from Eq. (5.4), follows a generalized Pólya urn scheme (Blackwell and MacQueen, 1973).

Thus, $\eta_{1:N}$ are randomly partitioned into those variables which share the same value. Let $\eta^*_{1:|\mathbf{c}|}$ denote the distinct values of $\eta_{1:N}$, $\mathbf{c} = c_{1:N}$ denote the partition such that $\eta_i = \eta^*_{c_n}$, and $|\mathbf{c}|$ denote the number of groups in that partition. The conditional distribution of $\eta_n \,|\, \eta_{1:n-1}$ is:

$$\eta_n \,|\, \eta_{1:n-1} = \begin{cases} \eta^*_i & \text{with prob } \frac{|\mathbf{c}|_i}{n-1+\alpha} \\ \eta, \eta \sim G_0 & \text{with prob } \frac{\alpha}{n-1+\alpha}, \end{cases} \tag{5.6}$$

where $|\mathbf{c}|_i$ is the size of the ith group in $\mathbf{c}$. Note that $\eta^*_{1:|\mathbf{c}|}$ are iid with distribution G_0.

This specification illuminates the connection between the DP and the Chinese restaurant process (CRP). A CRP is a distribution of partitions obtained by imagining N customers sitting down in a Chinese restaurant with an infinite number of tables.[1] The first customer sits at the first table. The nth subsequent customer sits at a table drawn from the following distribution:

$$\begin{aligned} p(\text{occupied table } i \,|\, \text{previous customers}) &= \frac{n_i}{\alpha+n-1} \\ p(\text{next unoccupied table} \,|\, \text{previous customers}) &= \frac{\alpha}{\alpha+n-1} \end{aligned} \tag{5.7}$$

[1] The terminology was inspired by the Chinese restaurants in San Francisco which seem to have an infinite seating capacity. It was coined by Jim Pitman and Lester Dubins in the early eighties (Aldous, 1985).

where n_i is the number of previous customers at table i and α is a positive parameter, as above. After N customers sit down, the seating plan is a partition of N items into some (random) number of groups.

Comparing Eq. (5.6) and Eq. (5.7), it is clear that the CRP distribution gives the same partition structure as draws from a DP. Furthermore, the CRP allows several variations on the basic rule in Eq. (5.7) without sacrificing exchangeability. These include a data-dependent choice of α, and a more general functional dependence on the current partition (Pitman, 2002).

We recover a DP model using a CRP by associating each table with an independent draw from G_0. Let $CRP_n(\alpha, c_{1:(n-1)})$ denote the table of the nth customer drawn from a CRP with assignments of the first $n-1$ customers given in $c_{1:(n-1)}$. We draw $\eta_{1:N}$ from the following process:

1. For $i \in \{1, 2, \ldots\}$, choose $\eta_i^* \,|\, G_0 \sim G_0$.
2. For $n \in \{1, 2, \ldots, N\}$:
   (a) Choose group assignment $C_n \,|\, \{c_{1:(n-1)}, \alpha\} \sim CRP_n(\alpha, c_{1:(n-1)})$.
   (b) Set $\eta_n = \eta^*_{c_n}$.

Notice that $C_{1:N}$ are *not* independent and identically distributed, but each one is drawn conditional on the previously drawn values. However, they are exchangeable. This can be shown either as a consequence of the form of the joint distribution of $\eta_{1:N}$ in Eq. (5.5) and De Finetti's theorem, or directly from the definition of the CRP in Eq. (5.7).

5.1.2 Sethuraman's stick-breaking construction

Integrating out the random measure G is useful for illuminating the relationship between the DP and exchangeable partition models such as the CRP. However, it is also of interest to explicitly construct G for a better understanding of the type of measures which can be drawn from a DP.

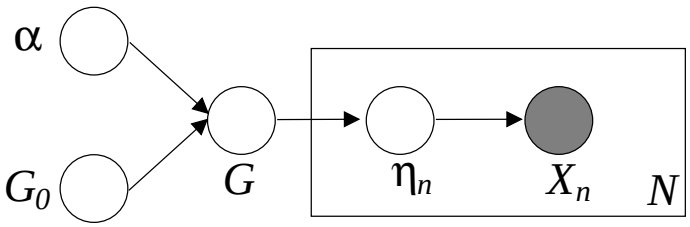


Figure 5.2: Graphical model representation of the Dirichlet process mixture.

Consider two infinite collections of independent random variables, $V_i \sim \text{Beta}(1, \alpha)$ and $\eta_i^* \sim G_0$ for $i = \{1, 2, \ldots\}$. Sethuraman (1994) proves that we can write $G \sim DP(\alpha, G_0)$ as:

$$\theta_i = V_i \prod_{j=1}^{i-1} (1 - V_i) \tag{5.8}$$

$$G(\eta) = \sum_{i=1}^{\infty} \theta_i \delta_{\eta_i^*}(\eta). \tag{5.9}$$

Thus the support of G consists of a countably infinite set of atoms, drawn iid from G_0. The probabilities of the atoms are given by successively breaking a unit length "stick" into an infinite number of pieces. The size of each successive piece, proportional to the rest of the stick, is given by an independent draw from a $\text{Beta}(1, \alpha)$ distribution.

This construction shows that measures drawn from a DP are discrete, even when η lies in a continuous space.

5.2 Dirichlet process mixture models

To overcome the limitation of the discreteness of G, Antoniak (1974) introduced the *Dirichlet process mixture model*, where η_n is the random parameter of the distribution of X_n:

1. Choose $G \,|\, \{\alpha, G_0\} \sim DP(\alpha, G_0)$.

2. For $n \in \{1, 2, \ldots, N\}$:

   (a) Choose $\eta_n \,|\, G \sim G$.

(b) Choose $X_n \,|\, \eta_n \sim p(x_n \,|\, \eta_n)$.

The DP mixture is illustrated in Figure 5.2. Conditional on a dataset, a central goal of nonparametric Bayesian modeling is to compute the predictive density:

$$p(x \,|\, x_1, \ldots, x_N) = \int p(x \,|\, \eta) p(\eta \,|\, x_1, \ldots, x_N) d\eta. \tag{5.10}$$

The practical implications of the DP mixture are readily seen from its CRP interpretation, which we call a *CRP mixture*. Integrating out the random measure G, the observations are partitioned according to those values which were drawn from the same parameter. Thus, the posterior distribution is of partitions of the data and the parameters associated with each group. This is similar to a finite mixture model, except that the number of groups is unknown and determined by the CRP. Furthermore, future data may either be associated with an existing group or drawn from a new, previously unseen parameter value. As described above, this can be a desirable property in data analysis. For example, when modeling text documents with a mixture model, we may naturally assume that future documents exhibit topics which were not yet seen in the given data.

Using the stick-breaking construction of Section 5.1.2, the DP mixture can be interpreted as an mixture model with an infinite number of components:

$$p(x \,|\, v_{1:\infty}, \eta^*_{1:\infty}) = \sum_{i=1}^{\infty} \theta_i p(x \,|\, \eta^*_i), \tag{5.11}$$

where θ is the function of $v_{1:\infty}$ defined in Eq. (5.8). The components of the θ vector are the (infinite) mixing proportions, and $\eta^*_{1:\infty}$ are the infinite number of mixture components. It is useful to consider the variable Z_n which is a multinomial indicator variable of the mixture component associated with X_n. The data thus arise from the following process:

1. For $i \in \{1, 2, \ldots, \}$:

   (a) Choose $V_i \,|\, \alpha \sim \text{Beta}(1, \alpha)$.

   (b) Choose $\eta^*_i \,|\, G_0 \sim G_0$.

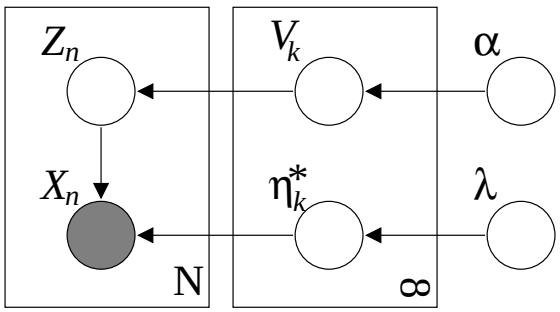


Figure 5.3: Graphical model representation of an exponential family DP mixture.

2. For $n \in \{1, 2, \ldots, N\}$:

   (a) Choose $Z_n \,|\, \mathbf{v} \sim \text{Mult}(\theta)$, where θ is defined in Eq. (5.8).

   (b) Choose $X_n \,|\, z_n, \boldsymbol{\eta}^* \sim p(x_n \,|\, \eta^*_{z_n})$.

5.2.1 The truncated Dirichlet process

Ishwaran and James (2001) have discussed the *truncated Dirichlet process* (TDP), in which V_{K-1} is set equal to one for some fixed value K. This yields $\theta_i = 0$ for $i \geq K$, and thus converts the infinite sum in Eq. (5.8) into a finite sum. Ishwaran and James (2001) show that a TDP closely approximates a true Dirichlet process when the truncation level K is chosen large enough relative to the number of data points. Thus, they can justify substituting a TDP mixture model for a full DP mixture model.

5.2.2 Exponential family mixtures

5.2.3 Exponential family mixtures

In this chapter, we restrict ourselves to DP mixtures for which the observable data are drawn from an exponential family distribution, and where the base measure for the DP is the corresponding conjugate prior.

A DP mixture using the stick-breaking construction is illustrated as a graphical model in Figure 6.1. The distributions of V_k and Z_n are as described above. The

distribution of X_n conditional on Z_n and $\{\eta_1^*, \eta_2^*, \ldots\}$ is:

$$p(x_n \mid z_n, \eta_1^*, \eta_2^*, \ldots) = \prod_{i=1}^{\infty} \left(h(x_n) \exp\{\eta_i^{*T} x_n - a(\eta_i^*)\}\right)^{z_n^i},$$

where $a(\eta_i^*)$ is the appropriate cumulant generating function and we assume for simplicity that x is the sufficient statistic for the natural parameter η.

The vector of sufficient statistics of the corresponding conjugate family is $(\eta^{*T}, -a(\eta^*))^T$. The base measure is thus:

$$p(\eta^* \mid \lambda) = h(\eta^*) \exp\{\lambda_1^T \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)\},$$

where we decompose the hyperparameter λ such that λ_1 contains the first $\dim(\eta^*)$ components and λ_2 is a scalar (see Section 2.1.2).

5.3 MCMC for DP mixtures

As in most of the hierarchical Bayesian models that we consider in this thesis, the posterior distribution of the latent variables is intractable to compute under the DP and TDP mixtures. MCMC methods are the tool of choice for approximating these posteriors (Escobar and West, 1995; Neal, 2000; Ishwaran and James, 2001).

5.3.1 Collapsed Gibbs sampling

In the *collapsed Gibbs sampler* for a DP mixture with conjugate base measure (Neal, 2000), we integrate out the random measure G and distinct parameter values $\{\eta_1^*, \ldots, \eta_{|\mathbf{c}|}^*\}$. The Markov chain is thus defined only on the latent partition of the data $\mathbf{c} = \mathbf{c}_{1:N}$.

Denote the data by $\mathbf{x} = x_{1:N}$. For $n \in \{1, \ldots, N\}$, the algorithm iteratively samples each group assignment C_n, conditional on the partition of the rest of the data $\mathbf{c}_{-n}$. Note that C_n can be assigned to one of $|\mathbf{c}_{-n}| + 1$ values: either the nth data point is in a group with other data points, or in a group by itself.

By exchangeability, C_n is drawn from the following multinomial distribution:

$$p(c_n^k = 1 \mid \mathbf{x}, \mathbf{c}_{-n}, \lambda, \alpha) \propto p(x_n \mid \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n^k = 1, \lambda) p(c_n^k = 1 \mid \mathbf{c}_{-n}, \alpha). \tag{5.12}$$

The first term is a ratio of normalizing constants of the posterior distribution of the kth parameter, one including and one excluding the nth data point:

$$p(x_n \mid \mathbf{x}_{-n}, \mathbf{c}_{-n}, c_n^k = 1, \lambda) = \frac{\exp\left\{a(\lambda_1 + \sum_{m \neq n} c_m^k X_m + X_n, \lambda_2 + \sum_{m \neq n} c_m^k + 1)\right\}}{\exp\left\{a(\lambda_1 + \sum_m c_m^k X_m, \lambda_2 + \sum_{m \neq n} c_m^k)\right\}}. \tag{5.13}$$

The second term is given by the Pólya urn scheme distribution of the partition:

$$p(c_n^k = 1 \mid \mathbf{c}_{-n}) \propto \begin{cases} |\mathbf{c}_{-n}|_k & \text{if } k \text{ is an existing group in the partition} \\ \alpha & \text{if } k \text{ is a new group in the partition,} \end{cases} \tag{5.14}$$

where $|\mathbf{c}_{-n}|_k$ denotes the number of data in the kth group of the partition.

Once this chain has reached its stationary distribution, we collect B samples $\{\mathbf{c}_{1:B}$ to approximate the posterior. The approximate predictive distribution of the next data point is an average of the predictive distributions for each of the collected samples:

$$p(x_{N+1} \mid x_1, \ldots, x_N, \alpha, \lambda) = \frac{1}{B} \sum_{b=1}^{B} p(x_{N+1} \mid \mathbf{c}_b, \mathbf{x}, \alpha, \lambda).$$

For a particular sample, that distribution is:

$$p(x_{N+1} \mid \mathbf{c}, \mathbf{x}, \alpha, \lambda) = \sum_{k=1}^{|\mathbf{c}|+1} p(c_{N+1}^k = 1 \mid \mathbf{c}) p(x \mid \mathbf{c}, \mathbf{x}, c_{N+1}^k = 1).$$

When G_0 is not conjugate, the integral in Eq. (5.13) does not have a simple closed form. Good algorithms for handling this case are given in Neal (2000).

5.3.2 Blocked Gibbs sampling

In the collapsed Gibbs sampler, the distribution of each group assignment variable depends on the most recently sampled values of the other variables. Thus, these variables must be updated one at a time, which can theoretically slow down the algorithm when compared to a blocking strategy. To this end, Ishwaran and James (2001) developed the TDP mixture described in Section 5.2. By explicitly sampling an approximation of G, this model allows for a *blocked* Gibbs sampler, where collections of variables can be simultaneously updated.

The state of the Markov chain consists of the beta variables $\mathbf{V} = \mathbf{V}_{1:K-1}$, the component parameters $\boldsymbol{\eta}^* = \eta^*_{1:K}$, and the component assignment variables $\mathbf{Z} = \mathbf{Z}_{1:N}$. The Gibbs sampler iterates between the following three steps:

1. For $n \in \{1, \ldots, N\}$, independently sample Z_n from:

$$p(z_n^k = 1 \,|\, \mathbf{v}, \boldsymbol{\eta}^*, \mathbf{x}) = \theta_k p(x_n \,|\, \eta_k^*),$$

   where θ_k is the function of $\mathbf{v}$ given in Eq. (5.8).

2. For $k \in \{1, \ldots, K\}$, independently sample V_k from $\text{Beta}(\gamma_{k,1}, \gamma_{k,2})$, where:

$$\begin{aligned} \gamma_{k,1} &= 1 + \textstyle\sum_{n=1}^{N} z_n^k \\ \gamma_{k,2} &= \alpha + \textstyle\sum_{i=k+1}^{K} \sum_{n=1}^{N} z_n^i. \end{aligned}$$

   This follows from the conjugacy between the multinomial data $\mathbf{z}$ and the truncated stick-breaking construction, which is a generalized Dirichlet distribution (Connor and Mosimann, 1969).

3. For $k \in \{1, \ldots, K\}$, independently sample η_k^* from $p(\eta_k^* \,|\, \tau_k)$. This distribution is in the same family as the base measure, with parameters:

$$\begin{aligned} \tau_{k,1} &= \lambda_1 + \textstyle\sum_{i \neq n} z_i^k x_i \\ \tau_{k,2} &= \lambda_2 + \textstyle\sum_{i \neq n} z_i^k. \end{aligned} \tag{5.15}$$

After the chain has reached its stationary distribution, we collect B samples and construct an approximate predictive distribution of the next data point. Again, this distribution is an average of the predictive distributions for each of the collected samples. The predictive distribution for a particular sample is:

$$p(x \,|\, \mathbf{z}, \mathbf{x}, \alpha, \lambda) = \sum_{k=1}^{K} \mathrm{E}\left[\theta_i \,|\, \gamma_1, \ldots, \gamma_k\right] p(x_{N+1} \,|\, \tau_k), \tag{5.16}$$

where $\mathrm{E}\left[\theta_i \,|\, \gamma_1, \ldots, \gamma_k\right]$ is the expectation of the product of independent beta variables given in Eq. (5.8). This distribution only depends on $\mathbf{z}$; the other variables are needed in the Gibbs sampling procedure, but can be integrated out here.

The TDP sampler readily handles non-conjugacy of G_0, provided that there is a method of sampling η_i^* from its posterior.

5.3.3 Placing a prior on the scaling parameter

A common extension to the DP mixture model is to place a prior on the scaling parameter α, which determines how quickly the number of components grows with the data. For the urn-based samplers, Escobar and West (1995) place a gamma(s_1, s_2) prior on α and derive Gibbs updates with auxiliary variable methods.

This gamma distribution has computationally convenient properties in the truncated DP mixture since it is the conjugate prior to the distribution of $V_{1:\infty}$. The V_i are distributed by Beta$(1, \alpha)$:

$$p(v \,|\, \alpha) = \alpha(1-v)^{\alpha-1}.$$

In its exponential family form, this distribution can be written as:

$$p(v \,|\, \alpha) = (1/(1-v)) \exp\{\alpha \log(1-v) + \log \alpha\},$$

where we see that $h(v) = 1/(1-v)$, $t(v) = \log(1-v)$, and $a(\alpha) = -\log \alpha$. Thus, we need a distribution where $t(\alpha) = \langle \alpha, \log \alpha \rangle$.

Consider the gamma distribution for α with shape parameter s_1 and inverse scale parameter s_2:

$$p(\alpha \,|\, s_1, s_2) = \frac{s_2^{s_1}}{\Gamma(s_1)} \alpha^{s_1 - 1} \exp\{-s_2 \alpha\}.$$

In its exponential family form, the distribution of α is:

$$p(\alpha \,|\, s_1, s_2) = (1/\alpha) \exp\{-s_2 \alpha + s_1 \log \alpha - a(s_1, s_2)\},$$

which is conjugate to Beta$(1, \alpha)$. The log normalizer is:

$$a(s_1, s_2) = \log \Gamma(s_1) - s_1 \log s_2,$$

and the posterior parameters conditional on data $\{v_1, \ldots, v_K\}$, are:

$$\begin{aligned} \hat{s}_2 &= s_2 - \sum_{i=1}^{K} \log(1 - v_i) \\ \hat{s}_1 &= s_1 + K. \end{aligned}$$

Thus, in the TDP mixture, auxiliary variable methods are not needed since simple Gibbs updates can be used:

$$\alpha \,|\, \{\mathbf{v}, s_1, s_2\} \sim \text{gamma}\left(\hat{s}_1, \hat{s}_2\right). \tag{5.17}$$

5.4 Variational inference for the DP mixture

We apply the mean-field variational approach to the stick-breaking construction of the DP mixture in Figure 6.1. The bound on the likelihood given in Eq. (2.13) is:

$$\begin{aligned}\log p(\mathbf{x} \mid \alpha, \lambda) \geq & \mathrm{E}_q \left[\log p(\mathbf{V} \mid \alpha)\right] + \mathrm{E}_q \left[\log p(\boldsymbol{\eta}^* \mid \lambda)\right] \\ & + \sum_{n=1}^{N} \left(\mathrm{E}_q \left[\log p(Z_n \mid \mathbf{V})\right] + \mathrm{E}_q \left[\log p(x_n \mid Z_n)\right]\right) \\ & - \mathrm{E}_q \left[\log q(\mathbf{Z}, \mathbf{V}, \boldsymbol{\eta}^*)\right].\end{aligned} \tag{5.18}$$

The subtlety to applying the variational method in this case is in constructing a distribution of the infinite dimensional random measure G, expressed in terms of the infinite set of beta variables $\mathbf{V} = \{V_1, V_2, \ldots\}$ and distinct parameters $\boldsymbol{\eta}^* = \{\eta_1^*, \eta_2^*, \ldots\}$. To do so, we truncate the variational distribution at a value T by setting $q(v_T = 1) = 1$. As in the truncated Dirichlet process, the mixture proportions θ_t will be zero, under the variational distribution, for for $t > T$ and we can subsequently ignore the component parameters η_t^* for $t > T$.

The factorized variational distribution is thus:

$$q(\mathbf{v}, \boldsymbol{\eta}^*, \mathbf{z}, T) = \prod_{t=1}^{T-1} q(v_i \mid \gamma_i) \prod_{t=1}^{T} q(\eta_t^* \mid \tau_t) \prod_{n=1}^{N} q(z_n \mid \phi_n), \tag{5.19}$$

where γ_n are the beta parameters of the distributions of V_i, τ_t are natural parameters of the distributions of η_t^*, and ϕ_n are multinomial parameters of the distributions of Z_n.

We emphasize the difference between this use of truncation and the blocked Gibbs sampler of Ishwaran and James (2001) (see Section 5.3.2). The blocked Gibbs sampler estimates the posterior of a truncated approximation to the DP. In contrast, we use a truncated stick-breaking distribution to approximate the true posterior of a full DP mixture model. The truncation level T is a variational parameter which can be freely set, and is not a part of the prior model specification.

5.4.1 Coordinate ascent algorithm

We now develop the algorithm for optimizing the bound in Eq. (5.18) with respect to the variational parameters. Except for the third term, all the other terms correspond to standard computations in an exponential family distribution. We rewrite the third term with indicator random variables:

$$\begin{aligned} \mathrm{E}_q\left[\log p(Z_n \mid \mathbf{V})\right] &= \mathrm{E}_q\left[\log\left(\textstyle\prod_{i=1}^{T}(1-V_i)^{\mathbf{1}[Z_n>i]}V_i^{Z_n^i}\right)\right] \\ &= \textstyle\sum_{i=1}^{T} q(z_n > i)\mathrm{E}\left[\log(1-V_i)\right] + q(z_n = i)\mathrm{E}\left[\log V_i\right], \end{aligned}$$

where:

$$\begin{aligned} q(z_n = i) &= \phi_{n,i} \\ q(z_n > i) &= \textstyle\sum_{j=i+1}^{K}\phi_{n,j} \\ \mathrm{E}\left[\log V_i\right] &= \Psi(\gamma_{i,1}) - \Psi(\gamma_{i,1}+\gamma_{i,2}) \\ \mathrm{E}\left[\log(1-V_i)\right] &= \Psi(\gamma_{i,2}) - \Psi(\gamma_{i,1}+\gamma_{i,2}). \end{aligned}$$

(Note that Ψ is the digamma function arising from the derivative of the log normalization factor in the beta distribution.)

Even though the exponential family DP mixture is a model with an effectively infinite number of random variables, the stick-breaking construction reveals that the distribution of each variable, conditional on the other variables, is a finite-dimensional exponential family distribution. Thus, we can optimize Eq. (5.18) by employing the coordinate ascent algorithm of Eq. (2.17), with the variational parameters appropriately transformed from their natural parameterization. For $t \in \{1,\ldots,T\}$ and $n \in \{1,\ldots,N\}$:

$$\begin{aligned} \gamma_{i,1} &= 1 + \textstyle\sum_n \phi_{n,i} \\ \gamma_{i,2} &= \alpha + \textstyle\sum_n \sum_{j=i+1}^{K}\phi_{n,j} \\ \tau_{i,1} &= \lambda_1 + \textstyle\sum_n \phi_{n,i}x_n \\ \tau_{i,2} &= \lambda_2 + \textstyle\sum_n \phi_{n,i} \\ \phi_{n,i} &\propto \exp\{S_{n,i}\}, \end{aligned} \tag{5.20}$$

where

$$S_{n,i} = \mathrm{E}\left[\log V_i \mid \gamma_i\right] + \mathrm{E}\left[\eta_i \mid \tau_i\right]^T X_n - \mathrm{E}\left[a(\eta_i) \mid \tau_i\right] - \sum_{j=i+1}^{K} \mathrm{E}\left[\log(1 - V_j) \mid \gamma_j\right].$$

Iterating between these updates optimizes Eq. (5.18) with respect to the variational parameters defined in Eq. (5.19). We thus find the variational distribution that is closest in KL distance, within the confines of its parameters, to the true posterior.

Practical applications of variational methods must address initialization of the variational distribution. As in Gibbs sampling, the algorithm is theoretically valid from any starting values of the variational parameters, but local maxima can be a problem. We initialize the variational distribution by incrementally updating the parameters according to a random permutation of the data points. In a sense, this is a variational version of sequential importance sampling. To avoid local maxima, we repeat the algorithm multiple times and choose the final parameter settings that give the best bound on the marginal likelihood.

Given a (possibly locally) optimal set of variational parameters, the approximate predictive distribution of the next data point is:

$$p(x_{N+1} \mid \mathbf{z}, \mathbf{x}, \alpha, \lambda) = \sum_{t=1}^{T} \mathrm{E}_q\left[\theta_t \mid \boldsymbol{\gamma}\right] \mathrm{E}_q\left[p(x_{N+1} \mid \tau_t)\right]. \tag{5.21}$$

This approximation has a form similar to the approximate predictive distribution under the blocked Gibbs sampler in Eq. (5.16). In the variational case, however, the averaging is done parametrically via the variational distribution, rather than by a Monte Carlo integral.

When G_0 is not conjugate, a simple coordinate ascent update for τ_i may not be applicable if $p(\eta_i^* \mid \mathbf{z}, \mathbf{x}, \lambda)$ is not in the exponential family. However, if G_0 is a mixture of conjugate priors, then there still is a simple coordinate ascent algorithm.

Finally, we extend the variational inference algorithm to posterior updates on the scaling parameter α with a gamma(s_1, s_2) prior. Using the exact posterior of α in Eq. (5.17), the variational posterior gamma(w_1, w_2) distribution is:

$$\begin{aligned} w_1 &= s_1 + T - 1 \\ w_2 &= s_2 - \sum_{i=1}^{T-1} \mathrm{E}_q\left[\log(1 - V_i)\right]), \end{aligned}$$

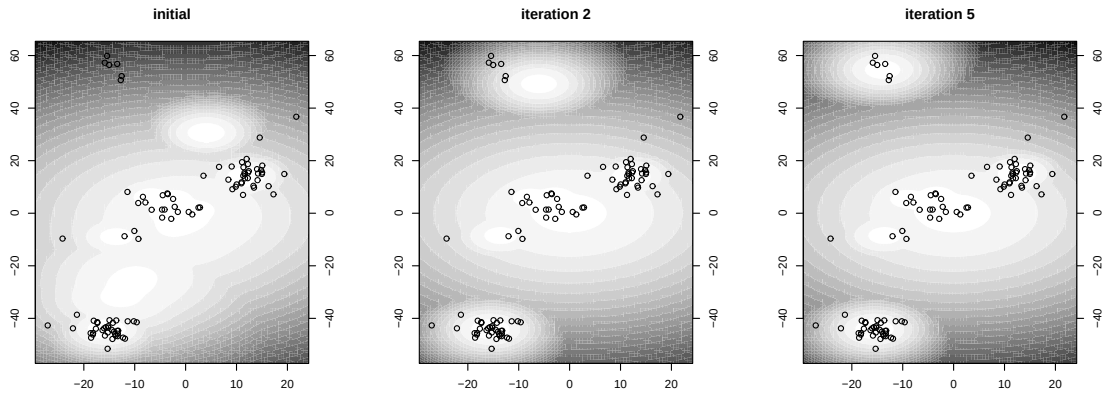


Figure 5.4: The approximate predictive distribution given by variational inference at different stages of the algorithm. The data are 100 points generated by a Gaussian DP mixture model with fixed diagonal covariance.

and we replace α with its expectation $\mathrm{E}_q[\alpha \mid w] = w_1/w_2$ in the updates on $\gamma_{t,2}$ of Eq. (5.20).

5.5 Example and Results

We applied the variational algorithm of Section 5.4 and Gibbs samplers of Section 5.3 to Gaussian DP mixtures. The data are assumed drawn from a DP mixture, where the parameter of the distribution of the nth data point is the mean of a Gaussian with fixed covariance matrix Λ. The base measure is Gaussian, with covariance given by Λ/λ_2 , which is conjugate to the data likelihood (see Section 2.1.2).

In Figure 5.4, we illustrate the variational inference algorithm on a toy problem. We simulated 100 data points from a two-dimensional Gaussian DP mixture with diagonal covariance. Each panel illustrates the data and predictive distribution of the next data point given by the variational inference algorithm, with truncation level 20. In the initial setting, the variational approximation places a largely flat distribution on the data. After one iteration, the algorithm has found the modes of the predictive distribution and, after convergence, it has further refined those modes. Even though 20 mixture components are represented in the variational distribution,

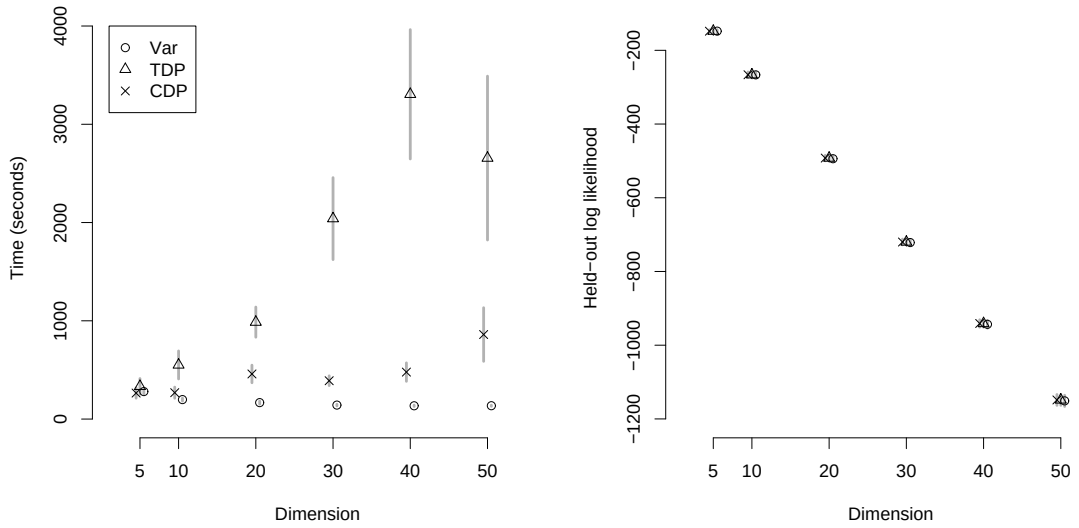


Figure 5.5: (Left) Convergence time across ten datasets per dimension for variational inference, TDP Gibbs sampling, and the collapsed Gibbs sampler (grey bars are standard error). (Right) Average held-out log likelihood for the corresponding predictive distributions.

the fitted approximate posterior only uses five of them.

5.5.1 Simulated mixture models

To compare the approximate inference algorithms described above, we performed the following simulation. We generate 100 data from a Gaussian DP mixture model and 100 additional points as held-out data. In the held-out data, each point is treated as the 101st data point in the collection and only depends on the original data. The fixed covariance is given by a first-order autocorrelation matrix, such that the components are highly dependent ($\rho = 0.9$), and the base measure on the mean is a zero-mean Gaussian with covariance appropriately scaled for comparison across dimensions. The scaling parameter of the DP is fixed at $\alpha = 1$.

We run all algorithms to convergence and measure the computation time. [2] For the Gibbs samplers, we assess convergence to the stationary distribution with the

[2]All timing computations were made on a Pentium III 1GHZ desktop machine.

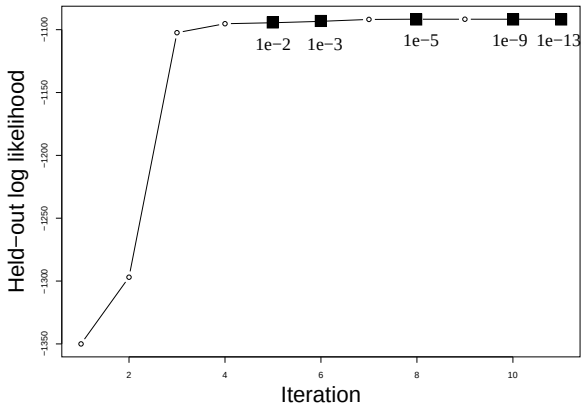


Figure 5.6: Held-out likelihood as a function of iteration of the variational inference algorithm for a 50-dimensional simulated dataset. The proportional change in likelihood bound (not illustrated) is labeled at selected iterations.

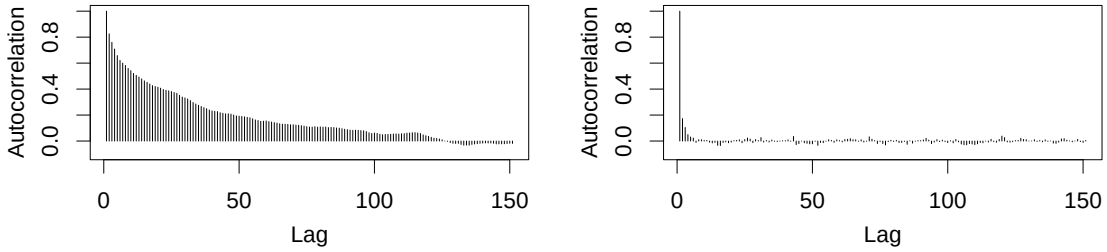


Figure 5.7: Autocorrelation plots on the size of the largest component for the truncated DP Gibbs sampler (left) and collapsed Gibbs sampler (right) in an example dataset of 50-dimensional Gaussian data.

diagnostic given by Raftery and Lewis (1992), and collect 25 additional samples to estimate the predictive distribution (the same diagnostic provides an appropriate lag to collect uncorrelated samples). The TDP approximation and variational posterior approximation are both truncated at 20 components. Finally, we measure convergence for variational inference by the proportional change in the likelihood bound, stopping the algorithm when it is less than $1e^{-10}$. Note that this is a conservative criterion, as illustrated in Figure 5.6.

In this setting, there are clear advantages to the variational algorithm. Figure 5.5 (left) illustrates the average convergence time across ten datasets per dimen-

sion. The variational algorithm was faster and exhibited significantly less variance in its convergence time. Furthermore, note that the collapsed Gibbs sampler converged faster than the TDP Gibbs sampler, giving the truncated approximation no real advantage. Though an iteration of collapsed Gibbs is slower than an iteration of TDP Gibbs, the TDP Gibbs sampler required a longer burn-in and greater lag to obtain uncorrelated samples. This is illustrated in the example autocorrelation plots of Figure 5.7.

Figure 5.5 (right) illustrates the average log likelihood assigned to the held-out data by the approximate predictive distributions. First, notice that the collapsed DP Gibbs sampler assigned the same likelihood as the posterior from the TDP Gibbs sampler—an indication of the quality of a TDP for approximating a DP. More importantly, however, the predictive distribution based on the variational posterior assigned a similar score as those based on samples from the true posterior. Though it is an approximation to the true posterior, the resulting predictive distributions are very accurate.

5.6 Discussion

In this chapter, we have seen how the Dirichlet process allows flexible mixture modeling. The number of mixture components is neither specified nor fixed, which is a natural assumption in many real-world datasets. Approximate posterior inference in such models amounts to simultaneously exploring the space of the number of components, and the specific parameters with which they are associated.

Variational methods for the Dirichlet process provide the general advantages of speed and easy assessment of convergence which were outlined in Section 2.3. Furthermore, in this context, the variational technique gives us an explicit estimate of the infinite-dimensional random parameter G with the truncated stick-breaking construction. The best Gibbs samplers rely on the Pólya urn scheme representation, in which G is marginalized out (Neal, 2000). This representation precludes the simple computation of quantities, such as order statistics, which cannot be expressed as ex-

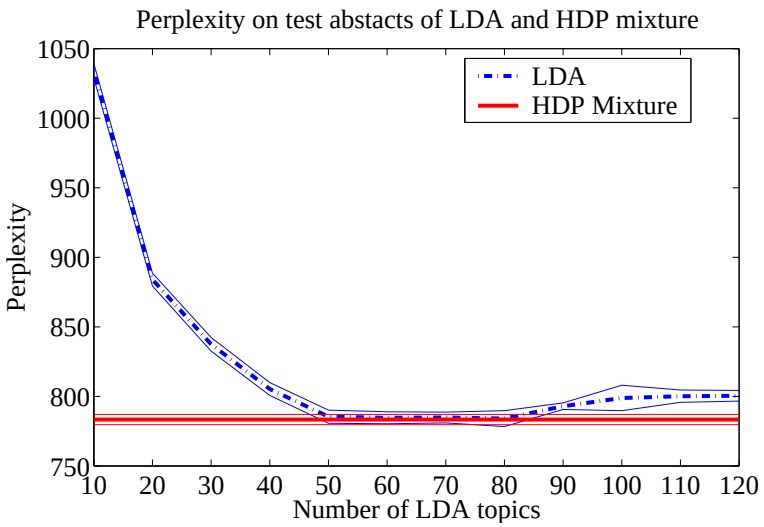


Figure 5.8: Test set perplexity of the nematode abstracts for LDA and HDP. Note that the HDP uses about the same number of topics which is best for LDA. However, no explicit search over K is required.

pectations over G. (See Gelfand and Kottas, 2002, for a method to compute such quantities with the Gibbs sampler.)

Though the DP mixture is a powerful model, one of the main points of Chapter 3 is that single-level mixtures are not always adequate, particularly for text data. The LDA model of that chapter is essentially a *mixture of mixtures*, which is often appropriate for text data and, more generally, for *grouped* data.

Applying the Dirichlet process to a mixture of mixtures requires two levels of DP modeling, which is best explained with the Chinese restaurant process formalism. The mixture components (i.e., the topics) are governed by a corpus CRP, and the words of the documents are governed by a collection of document-specific CRP's. Each document partitions its words according to its CRP, and the topic associated with each partition is drawn from the corpus CRP. Thus, each document uses a different number of topics, while the corpus CRP ensures that the documents will share topics. (Compare this to LDA, where each document uses the same set of topics.)

As a Dirichlet process model, this is referred to as a *hierarchical Dirichlet process* (HDP) because it is equivalent to drawing from a DP for each document, where the

base measure itself is a draw from a corpus DP. Details of this model can be found in Teh et al. (2004). Reanalyzing the nematode abstracts data from Section 3.6.1 with LDA models and an HDP, we find that the HDP uses approximately the same number of topics as best determined, with the held-out likelihood criterion, by LDA (see Figure 5.8). However, the HDP is a more flexible model because different documents can exhibit different number of topics and, as with a simple DP model, future documents can exhibit topics which were previously unseen.

In the next chapter, we will address the problem of finding natural topic hierarchies in collections of text with a model based on the HDP.

Chapter 6

Hierarchical latent Dirichlet allocation

In the previous chapter, we described how Dirichlet process mixtures can extend our suite of models to those for which the number of factors can grow with the data. In this chapter, we extend the Dirichlet process to develop a prior on *factor hierarchies*. Using this prior, we address the important problem of estimating a topic hierarchy from text data, and we develop flexible models that allow it to grow and change as the documents accumulate.

There are several possible approaches to the modeling of topic hierarchies. In our approach, each node in the hierarchy is associated with a topic, i.e., a distribution on words; a document is generated by choosing a path from the root to a leaf, repeatedly sampling topics along that path, and sampling the words from the selected topics.

Thus the organization of topics into a hierarchy aims to capture the breadth of usage of topics across the corpus, reflecting underlying syntactic and semantic notions of generality and specificity. This approach differs from models of topic hierarchies which are built on the premise that the distributions associated with parents and children are similar (Segal et al., 2002). We assume no such constraint—for example, the root node topic may place all of its probability mass on function words, with none of its descendants placing any probability mass on function words. Our model thus more closely resembles the hierarchical topic model considered in Hofmann (1999a), though that work does not address the model selection problem which is our primary focus.

6.1 The nested Chinese restaurant process

The CRP is amenable to mixture modeling because we can establish a one-to-one relationship between tables and mixture components and a one-to-many relationship between mixture components and data. In the models that we will consider, however, each data point is associated with multiple mixture components which lie along a path in a hierarchy. We thus develop an extension of the CRP to specify a prior on trees.

A *nested Chinese restaurant process* is defined by imagining the following scenario. Suppose that there are an infinite number of infinite-table Chinese restaurants in a city. One restaurant is determined to be the root restaurant and on each of its infinite tables is a card with the name of another restaurant. On each of the tables in those restaurants are cards that refer to other restaurants, and this structure repeats infinitely. Each restaurant is referred to exactly once; thus, the restaurants in the city are organized into an infinitely-branched tree. Note that each restaurant is associated with a level in this tree (e.g., the root restaurant is at level 1 and the restaurants it refers to are at level 2).

A tourist arrives in the city for a culinary vacation. On the first evening, he enters the root Chinese restaurant and selects a table using the CRP partition distribution in Eq. (5.7). On the second evening, he goes to the restaurant identified on the first night's table and chooses another table, again from the partition distribution. He repeats this process for L days. At the end of the trip, the tourist has sat at L restaurants which constitute a path from the root to a restaurant at the Lth level in the infinite tree described above. Thus, after M tourists take L-day vacations, the collection of paths describe a particular L-level subtree of the infinite tree (see Figure 6.1 Left for an example of such a tree).

This prior can be used to model random tree structures. Just as a standard CRP is used to express uncertainty about a possible number of components, the nested CRP is used to express uncertainty about possible L-level trees.

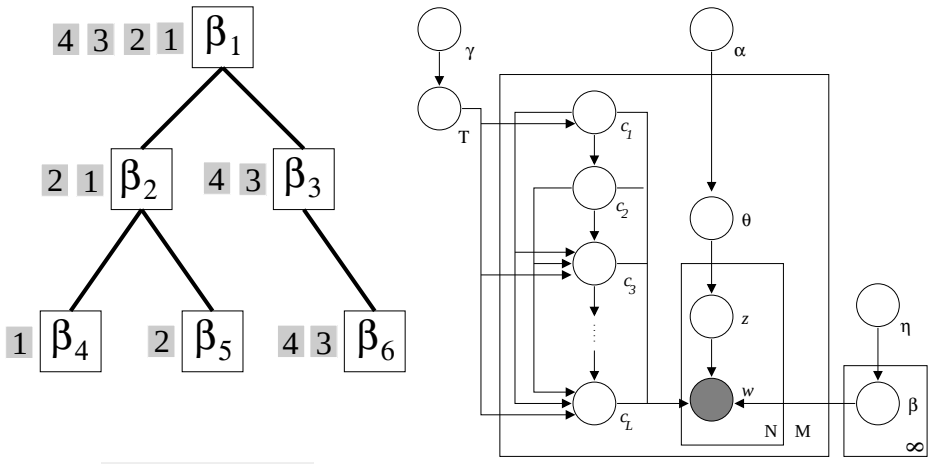


Figure 6.1: (Left) The paths of four tourists through the infinite tree of Chinese restaurants ($L = 3$). The solid lines connect each restaurant to the restaurants referred to by its tables. The collected paths of the four tourists describe a particular subtree of the underlying infinite tree. This illustrates a sample from the state space of the posterior nested CRP of Figure 6.1b for four documents. (Right) The graphical model representation of hierarchical LDA with a nested CRP prior. We have separated the nested Chinese restaurant process from the topics. Each of the infinite β variables is associated with one of the restaurants.

6.2 Hierarchical latent Dirichlet allocation

Recall the LDA model from Chapter 3. Our basic assumption is that the words of a document are generated according to a mixture model, where the mixing proportions are random and document-specific. The documents are assumed to have arisen from the following generative process: (1) choose a K-vector of topic proportions from a Dirichlet distribution on the K-simplex; (2) repeatedly sample words from the mixture distribution resulting from the chosen proportions. LDA is thus a two-level generative process in which documents are associated with topic proportions, and the corpus is modeled as a Dirichlet distribution on these topic proportions.

We now describe an extension of LDA in which the topics lie in a hierarchy. For the moment, suppose we are given an L-level tree and each node is associated with a topic. A document is generated as follows: (1) choose a path from the root of the

tree to a leaf; (2) choose an L-vector of topic proportions from an Dirichlet on the L-simplex; (3) repeatedly sample words from a mixture of the topics along the path chosen in step 1, using the mixture proportions drawn in step 2.

This model can be viewed as a fully generative version of the cluster abstraction model (Hofmann, 1999a). As a text model, it can capture the natural hierarchical topic structure of words in a corpus. For example, the root node topic may contain so-called *function words*—words like "of", "but", or "the"—which are shared across all documents but provide little semantic information. Second level topic distributions may delineate a crude topic structure in the corpus. Lower level topic distributions may further refine that structure.

Finally, we use the nested CRP to relax the assumption of a fixed tree structure. Analogous to using a CRP in a mixture model, we associate each restaurant in the infinite tree with a topic drawn from a symmetric Dirichlet. A document is drawn by first choosing an L-level path through the tree, and then drawing words from the L topics which are associated with the restaurants along that path. Note that all documents share the topic associated with the root restaurant.

Let $c_{1:L}$ denote an L-level path. In *hierarchical LDA* (hLDA), each document is drawn from the following generative process:

1. Set the start of the path C_1 to be the root restaurant.
2. For $\ell \in \{2, \ldots, L\}$, choose $C_\ell \,|\, c_{\ell-1}$ from the CRP indexed by $c_{\ell-1}$.
3. Choose proportions $\theta \,|\, \alpha \sim \text{Dir}(\alpha)$.
4. For $n \in \{1, \ldots, N\}$:
   (a) Choose level $Z_n \,|\, \theta \sim \text{Mult}(\theta)$.
   (b) Choose word $W_n \,|\, \{z_n, c_{1:L}, \beta_{1:\infty}\} \sim \text{Mult}(\beta_{c_{z_n}})$.

By considering the topics of a document, conditional on the previous documents, we illuminate the important difference between hLDA and the CRP mixture from Chapter 5. In the CRP mixture, new topics are created as the documents accumulate.

The nth document is generated either from one of the topics of the previous $n-1$ documents, or an entirely new topic (drawn from the base measure).

In hLDA, the $n-1$ documents trace a *subtree* of the infinite tree. Thus, the nth document is either associated with a path that has been traced before, or a new path. A new path, however, will necessarily share between 1 and $L-1$ components of some of the previously drawn paths, depending on at which level the first new component arises. The topics which it does not share are drawn from the base measure.

The hLDA model is illustrated in Figure 6.1 (Right). The node labeled T refers to a collection of an infinite number of L-level paths drawn from a nested CRP. Given T, the $c_{m,\ell}$ variables are deterministic—simply look up the ℓth level of the mth path in the infinite collection of paths. However, not having observed T, the distribution of $c_{m,\ell}$ will be defined by the nested Chinese restaurant process, conditional on all the $c_{q,\ell}$ for $q < m$.

Conditional on a collection of M documents, the resulting posterior of the path variables is essentially transferred (via the deterministic relationship), to a posterior on the first M paths in T. The posterior path distribution of a new document $\mathbf{w}_{M+1}$ will depend, through the unobserved T, on the posterior paths of all the documents in the original corpus.

Dirichlet process interpretation

We saw in Chapter 5 that a CRP mixture is a DP mixture with the random measure G marginalized out of the model. The hLDA model has an analogous interpretation. The data arise by first drawing G from a Dirichlet process and then, for each document, drawing document-specific parameters from G, and drawing the words from those parameters. In this case, each document is associated with an L-set of multinomial parameters.

The random measure on such sets is drawn from a *nested Dirichlet process*. Suppose for $\ell \in \{1, \ldots, L-1\}$, we recursively define $G_{\ell,0}$ to be the product measure $\text{Dir}(\eta) \times DP(\gamma, G_{\ell+1,0})$, and define $G_{L,0}$ to be $\text{Dir}(\eta)$. The nested Dirichlet process is defined to be $DP(\gamma, G_{0,0})$.

To see how to recover the nested CRP from this definition, consider the discrete structure of G. Each atom on which it places probability is a point mass at a multinomial parameter and another draw from a DP. For that draw, each atom is again a point mass at a multinomial parameter coupled with yet another draw from a DP. This structure continues for L levels, where each atom of the random measure at each level is associated with yet another infinite collection of atoms.

Thus the collection of the atoms of all the draws form an infinite tree as described above. Integrating out the random measures, we recover the Pólya urn scheme given by the nested CRP.

6.2.1 Approximate inference with Gibbs sampling

Given a document collection, we hope to compute the posterior distribution of the hierarchy structure, corresponding topics, document-specific paths, and allocations of each document's words to levels on its path. As in the CRP mixture model, exact posterior inference for hLDA is intractable. We appeal to collapsed Gibbs sampling techniques for posterior approximation.

We can integrate out the level proportions θ_m and topics β_k to construct a Markov chain on the allocation variables $z_{m,n}$ and path assignments $c_{m,\ell}$.

For each document, we divide the Gibbs sampler into two parts. First, given the current path assignment, we sample the level allocation variables of the underlying LDA model (Griffiths and Steyvers, 2002):

$$p(z_{m,n} \mid \mathbf{z}_{-(m,n)}, \mathbf{w}) \propto p(z_{m,n} \mid \mathbf{z}_{m,-n}) p(w_{m,n} \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(m,n)})$$

Let $n_{\mathbf{z}}$ denote the L-vector of topic counts in $\mathbf{z}$, and let $n_{(\mathbf{w},\mathbf{z},\mathbf{c})}$ denote the word counts assigned to the tree defined by the path assignments. The terms in the above equation are:

$$\begin{aligned} p(z_{m,n} \mid \mathbf{z}_{m,-n}) &\propto \alpha + n_{\mathbf{z}_{-m}}^{z_{m,n}} \\ p(w_{m,n} \mid \mathbf{z}, \mathbf{c}, \mathbf{w}_{-(m,n)}) &\propto \eta + n_{(\mathbf{w}_{-(m,n)}, \mathbf{z}_{-(m,n)}, \mathbf{c})}^{c_{z_{m,n}}, w_{m,n}} \end{aligned}$$

Second, given the level allocation variables, we sample the path variables associated with the nested CRP prior:

$$p(\mathbf{c}_m \mid \mathbf{w}, \mathbf{c}_{-m}, \mathbf{z}) \propto p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) p(\mathbf{c}_m \mid \mathbf{c}_{-m}),$$

This expression is an instance of Bayes' rule with $p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z})$ as the likelihood of the data given a particular choice of path, and $p(\mathbf{c}_m \mid \mathbf{c}_{-m})$ as the prior on paths implied by the nested CRP. The likelihood is obtained by integrating over the multinomial parameters, which gives a ratio of normalizing constants for the Dirichlet distribution:

$$p(\mathbf{w}_m \mid \mathbf{c}, \mathbf{w}_{-m}, \mathbf{z}) =$$
$$\prod_{\ell=1}^{L} \frac{\Gamma\left(n^{c_{m,\ell}}_{(\mathbf{w}_{-m},\mathbf{z}_{-m},\mathbf{c}_{-m})} + V\eta\right)}{\prod_w \Gamma\left(n^{c_{m,\ell},w}_{(\mathbf{w}_{-m},\mathbf{z}_{-m},\mathbf{c}_{-m})} + \eta\right)} \frac{\prod_w \Gamma\left(n^{c_{m,\ell},w}_{(\mathbf{w}_{-m},\mathbf{z}_{-m},\mathbf{c}_{-m})} + n^{c_{m,\ell},w}_{(\mathbf{w}_m,\mathbf{z}_m,\mathbf{c}_m)} + \eta\right)}{\Gamma\left(n^{c_{m,\ell}}_{(\mathbf{w}_{-m},\mathbf{z}_{-m},\mathbf{c}_{-m})} + n^{c_{m,\ell}}_{(\mathbf{w}_m,\mathbf{z}_m,\mathbf{c}_m)} + V\eta\right)}$$

Note that the path must be drawn as a block, because its value at each level depends on its value at the previous level. The set of possible paths corresponds to the union of the set of existing paths through the tree, equal to the number of leaves, with the set of possible novel paths, equal to the number of internal nodes.

6.3 Examples and empirical results

We present data analysis of both simulated and real text data to validate this model and its corresponding Gibbs sampler.

Simulated data

In Figure 6.2, we depict the hierarchy structure and allocation count of ten datasets drawn from the hLDA model. For each dataset, we draw 100 documents of 250 words each. The vocabulary size is 100, and the hyperparameters are fixed at $\alpha = (50, 30, 10)$, $\eta = 0.005$, and $\gamma = 1$. The resulting hierarchies illustrate the range of structures on which the prior assigns probability.

In the same figure, we illustrate the approximate posterior mode of the hierarchy structure and allocation counts for the ten datasets. We exactly recover the correct

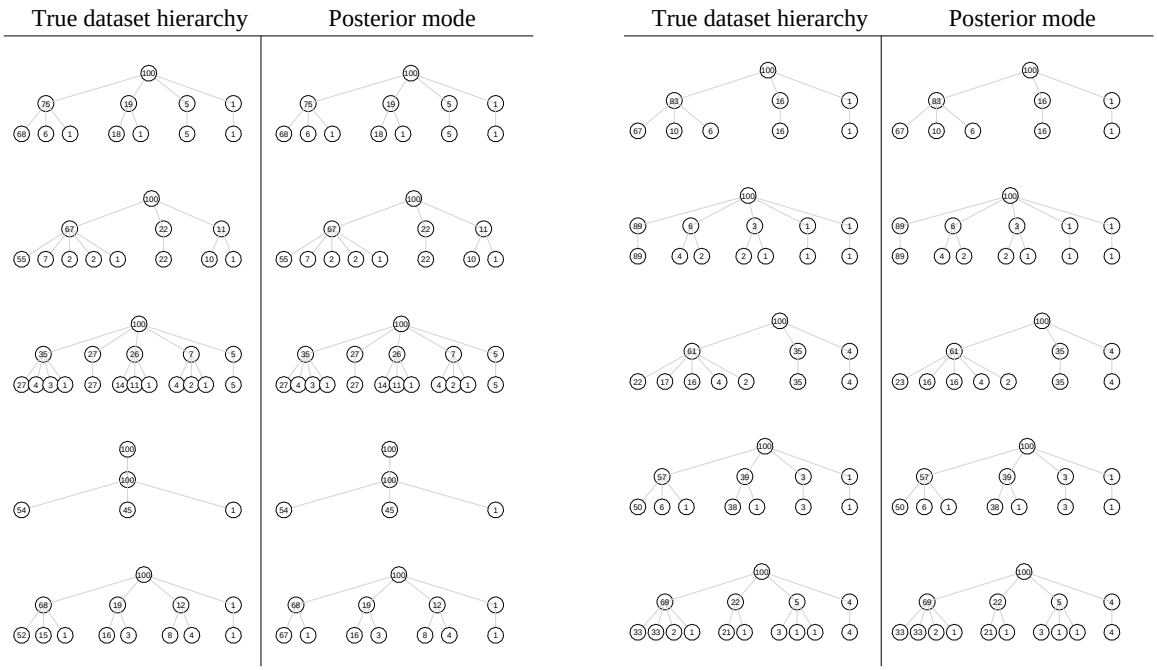


Figure 6.2: Hierarchies of ten datasets drawn from a hLDA model, and the corresponding approximate posterior mode.

hierarchy structure, with only two errors. In one case, the error is a single wrongly allocated path. In the other case, the found mode has higher posterior likelihood than the true tree structure (due to finite data).

In general we cannot expect to always find the exact tree. This is highly dependent on the size of the dataset, and how identifiable the topics are. Our choice of small η yields topics that are sparse and (probably) very different from each other. Datasets which exhibit polysemy and similarity between topics will not have as easily identifiable trees.

Scientific abstracts

The hLDA model is particularly appropriate to analyzing collections of scientific abstracts for recovering the underlying hierarchical structure which embodies many fields.

In Figure 6.3, we illustrate the approximate posterior mode of a hierarchy estimated from a collection of 533 abstracts from the Journal of American Computing

Machinery (JACM). The JACM is a premier computer science journal which seeks to publish the most influential articles in all fields of computer science. The posterior has correctly found the function words of the dataset, assigning words like "the", "of", or, "and" to the root topic. In its second level, the posterior hierarchy captures several subfields of computer science, such as databases, systems, networking, learning, and theory. In the third level, it further refines those fields. For example, it delineates between network routing problems and network balancing problems.

In Figure 6.4, we illustrate an analysis of a collection of psychology abstracts from the Psychological Review (1967-present). Again, we have discovered an underlying hierarchical structure of the field. The top node contains the function words; the second level delineates between large fields, such as cognitive and social psychology; the third level further refines those fields into subfields.

6.4 Discussion

In this chapter, we developed the nested Chinese restaurant process, a distribution on hierarchical partitions. We use this process as a nonparametric prior for a hierarchical extension to the latent Dirichlet allocation model. The result is a flexible, general model for topic hierarchies that naturally accommodates growing data collections.

We used Gibbs sampling for hLDA, rather than mean-field variational inference. However, variational methods are well within reach using the methods developed in Chapter 5. In particular, we appeal to the nested DP representation, and use a nested stick-breaking construction of the posterior.

This model has two natural extensions. First, for simplicity we restrict ourselves to hierarchies of fixed depth, but it is straightforward to consider a model in which the path length can vary from document to document. Each document is still a mixture of topics along the path, but different documents can express paths of different lengths as they represent varying levels of specialization. Second, in our current model a document is associated with a single path. It is also natural to consider models in which documents are allowed to mix over paths. This is a way to take advantage of

syntactic structures such as paragraphs and sentences within a document.

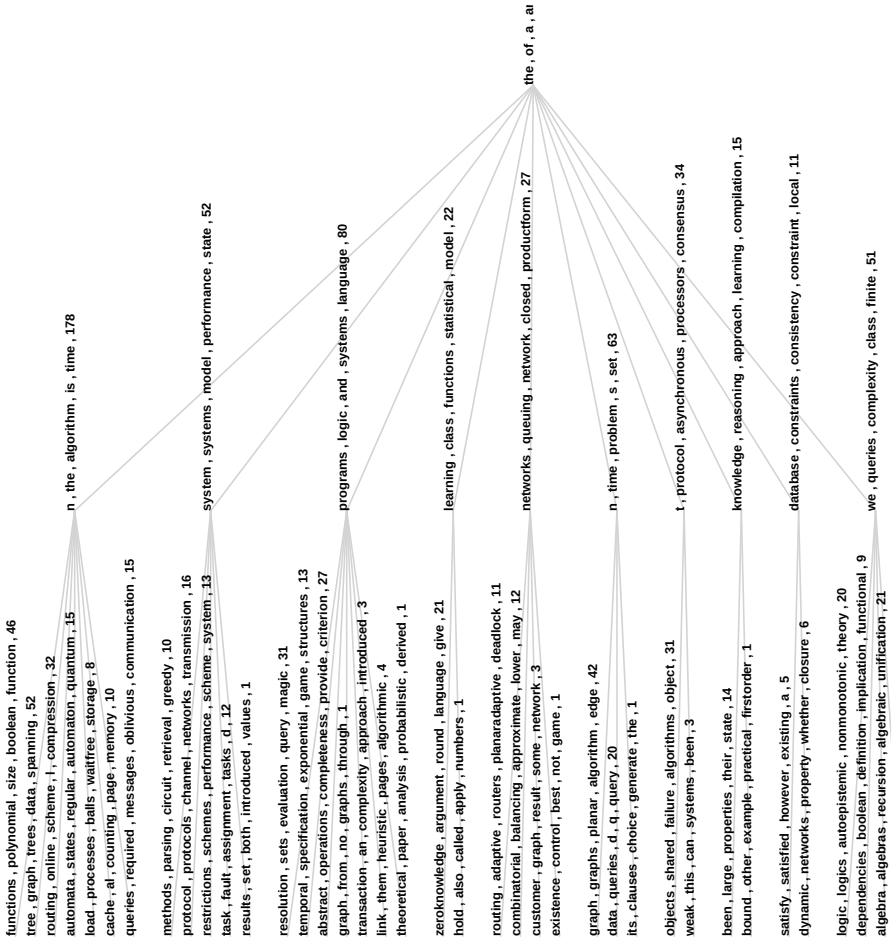


Figure 6.3: Hierarchy learned from the corpus of JACM abstracts.

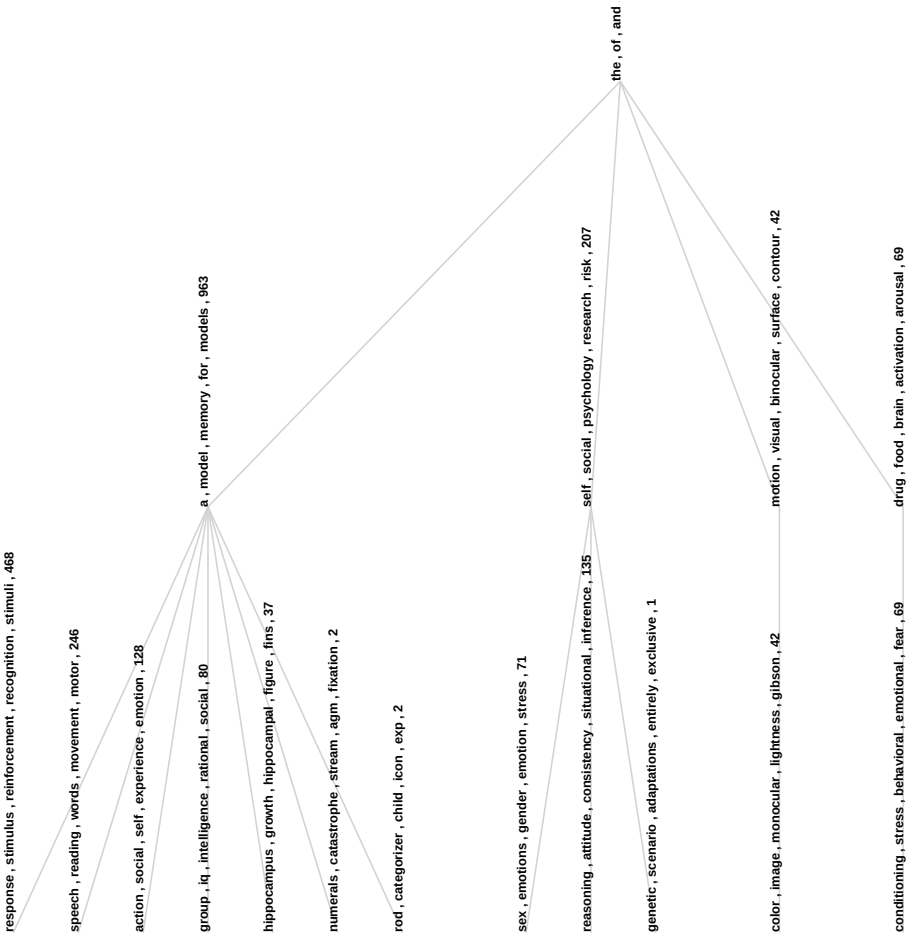


Figure 6.4: Hierarchy learned from a corpus of psychology abstracts.

Chapter 7

Conclusions

In this thesis, we have developed sophisticated statistical techniques for analyzing information collections, such as text or images. We used directed graphical models as a flexible, modular framework to describe modeling assumptions about the data. Furthermore, we derived general posterior inference techniques which free us from having to specify tractable models. These methods allowed us to take the Bayesian perspective, even in the face of large datasets.

With this framework in hand, we developed latent variable models based on principled exchangeability assumptions. In text, these models posit an index of hidden topics that describe the underlying collection. New documents are situated into the collection through posterior inference of those hidden topics. As we have shown, we can use the same types of models to index a set of images, or multimedia collections of interrelated text and images.

Finally, we used nonparametric Bayesian methods (i.e., the Dirichlet process) to relax the assumption of a fixed number of topics, and developed methods for which the size of the latent index grows with the data. We expanded this idea to trees, and can thus discover the structure and content of a topic hierarchy that underlies a collection.

We identify three areas of future work:

- *Partial exchangeability.* LDA assumes full exchangeability of words in a doc-

ument, and documents in a corpus. This is reasonable for capturing semantic content, but we might want to consider more realistic notions of *partial* exchangeability on the words, for which corollaries of de Finetti's representation theorem still hold. For example, we can attach a parse structure to each sentence, or consider the words in a Markovian sequence.

- *Variational methods.* We have consistently appealed to the mean-field variational method, in which the variational distribution is fully-factorized. Structured variational distributions may be appropriate for better approximate inference. We can also explore other kinds of variational methods, such as expectation propagation (Minka and Lafferty, 2002), which consider a different tractable set of mean parameters.

- *Relational data.* In Chapter 4, we developed a model that describes a simple relationship between an image and its caption. In general, multi-type data modeling may require a more complicated relationship structure, and this is the focus of probabilistic entity relation (PER) models (Heckerman et al., 2004). Developing latent variable models for the PER framework, and the corresponding inference techniques, is an important area of further work.

Bibliography

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions*. Dover, New York.

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983*, pages 1–198. Springer, Berlin.

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

Attias, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*.

Avery, L. (2002). Caenorrhabditis genetic center bibliography.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press, New York.

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

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

Beeferman, D., Berger, A., and Lafferty, J. (1999). Statistical models for text segmentation. *Machine Learning*, 34:177–210.

Bertsekas, D. (1999). *Nonlinear Programming*. Athena Scientific.

Bishop, C., Spiegelhalter, D., and Winn, J. (2003). VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. T. and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 777–784. MIT Press, Cambridge, MA.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Blei, D., Bagnell, J., and McCallum, A. (2002). Learning with scope, with application to information extraction and classification. In *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, pages 53–60, San Francisco, CA. Morgan Kaufmann Publishers.

Blei, D. and Moreno, P. (2001). Topic segmentation with an aspect hidden markov model. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 343–348. ACM Press.

Brochu, E. and de Freitas, N. (2003). Name that song: A probabilistic approach to querying on music and text. In *Advances in Neural Information Processing Systems 15*.

Brown, L. (1986). *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA.

Buntine, W. and Jakulin, A. (2004). Applying discrete PCA in data analysis. In *Proceedings of Uncertainty in Artificial Intelligence*.

Bush, V. (1945). As we may think. *The Atlantic Monthly*.

Cohn, D. and Hofmann, T. (2001). The missing link—A probabilistic model of document content and hypertext connectivity. In *Advances in Neural Information Processing Systems 13*.

Connor, R. and Mosimann, J. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206.

de Finetti, B. (1990). *Theory of probability. Vol. 1-2*. John Wiley & Sons Ltd., Chichester. Reprint of the 1975 translation.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39:1–38.

Diaconis, P. (1988). Recent progress on de Finetti's notions of exchangeability. In *Bayesian statistics, 3 (Valencia, 1987)*, pages 111–125. Oxford Univ. Press, New York.

Dickey, J. (1983). Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78:628–637.

Dickey, J., Jiang, J., and Kadane, J. (1987). Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, 82:773–781.

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

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

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

Gelfand, A. and Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11:289–305.

Gelfand, A. and Smith, A. (1990). Sample based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian data analysis*. Chapman & Hall, London.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov chain Monte Carlo Methods in Practice*. Chapman and Hall.

Girolami, M. and Kaban, A. (2003). On an equivalence between PLSI and LDA. In *26th Annual International ACM Conference on Research and Development in Information Retrieval(SIGIR03)*, pages 433–434.

Girolami, M. and Kaban, A. (2004). Simplicial mixtures of Markov chains: Distributed modelling of dynamic user profiles. In *Advances in Neural INformation Procesing Systems 16*, pages 9–16. MIT Press.

Goodrum, A. (2000). Image information retrieval: An overview of current research. *Informing Science*, 3(2):63–67.

Griffiths, T. and Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*.

Harman, D. (1992). Overview of the first text retrieval conference (TREC-1). In *Proceedings of the First Text Retrieval Conference (TREC-1)*, pages 1–20.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Heckerman, D., Meek, C., and Koller, D. (2004). Probabilistic models for relational data. Technical Report MSR-TR-2004-30, Microsoft Research.

Heckerman, D. and Meila, M. (2001). An experimental comparison of several clustering and initialization methods. *Machine Learning*, 42:9–29.

Hofmann, T. (1999a). The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *IJCAI*, pages 682–687.

Hofmann, T. (1999b). Probabilistic latent semantic indexing. *Proceedings of the Twenty-Second Annual International SIGIR Conference*.

Ishwaran, J. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–174.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR 2003*.

Joachims, T. (1999). Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*. M.I.T. Press.

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

Kass, R. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407):717–726.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA.

Leisink, M. and Kappen, H. (2002). General lower bounds based on computer generated higher order expansions. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*.

Maritz, J. and Lwin, T. (1989). *Empirical Bayes methods*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Marlin, B. (2004). Collaborative filtering: A machine learning perspective. Master's thesis, University of Toronto.

Meghini, C., Sebastiani, F., and Straccia, U. (2001). A model of multimedia information retrieval. *Journal of the ACM (JACM)*, 48(5):909–970.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, M., , and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Minka, T. (2000). Estimating a Dirichlet distribution. Technical report, M.I.T.

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

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–65. With discussion.

Naphade, M. and Huang, T. (2001). A probabilistic framework for semantic video indexing, filtering and retrieval. *IEEE Transactions on Multimedia*, 3(1):141–151.

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

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

Neal, R. M. and Hinton, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. pages 355–368.

Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134.

Papadimitriou, C., Tamaki, H., Raghavan, P., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. pages 159–168.

Pasula, H., Marthi, B., Milch, B., Russell, S., and Shpitser, I. (2002). Identity uncertainty and citation matching. In *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press.

Pietra, S. D., Pietra, V. D., and Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:380–393.

Pitman, J. (2002). *Combinatorial Stochastic Processes*. Lecture Notes for St. Flour Summer School.

Ponte, J. and Croft, B. (1998). A language modeling approach to information retrieval. In *ACM SIGIR 1998*, pages 275–281.

Popescul, A., Ungar, L., Pennock, D., and Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*.

Potthoff, R., Manton, K., and Woodbury, M. (2000). Dirichlet generalizations of latent-class models. *Journal of Classification*, 17:315–353.

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

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

Raftery, A. and Lewis, S. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7:493–497.

Rennie, J. (2001). Improving multi-class text classification with naive Bayes. Technical Report AITR-2001-004, M.I.T.

Rockafellar (1970). *Convex Analysis*. Princeton University Press.

Salton, G. and McGill, M., editors (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Segal, E., Koller, D., and Ormoneit, D. (2002). Probabilistic abstraction hierarchies. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):888–905.

Taskar, B., Segal, E., and Koller, D. (2001). Probabilistic clustering in relational data. In *IJCAI-01*, pages 870–876.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2004). Hierarchical Dirichlet processes. Technical Report 653, U.C. Berkeley, Dept. of Statistics.

Wainwright, M. and Jordan, M. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, U.C. Berkeley, Dept. of Statistics.

Wang, J., Li, J., and Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–983.

Xing, E., Jordan, M., and Russell, S. (2003). A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of UAI*.