Covariate-Dependent Nonparametric Mixture Models

Contents

Listing of figures

Listing of tables

Acknowledgments

I WISH TO THANK all the people who supported me throughout my education and research, and without whom this work would not have been possible. First and foremost, I wish to express my sincere thanks to my advisor, Dustin Tingley: thank you for taking me on as an inexperienced sophomore, for guiding me through this project, and for helping me develop my scientific confidence. Your sustained support has been indispensable. I also thank Brandon Stewart for being an incredibly patient and effective mentor at the earliest stages of my research experience. A summer spent at the Institute for Quantitative Social Science (IQSS) at Harvard provided me with a friendly environment to conduct research, and allowed me to meet many of the people who shaped my research direction.

My gratitude also goes to many members of the community at Harvard's School of Engineering and Applied Science (SEAS). To Margo Levine, thank you for supporting me through my academic journey at SEAS ever since first encouraging me to join it. I am also grateful for the advice of Finale Doshi-Velez, who first introduced me to the field of Bayesian nonparametrics, and who provided invaluable feedback on this work in its early stages as a project in CS 281. Many other fellow students and faculty members at SEAS were crucial in influencing my academic and personal path, and I extend my thanks and appreciation to them all.

I would like to thank Richard Wrangham, Elizabeth Ross, and all the tutors, staff, and students at Currier House for having provided me with a strong and supportive community and a place to call home. To all my friends both within and outside of Harvard, thank you for your friendship and constant encouragement. Lastly, I am grateful to my parents Alfredo and Lia, to my brother Alessandro, and to Alex, for loving me and always believing in me.

1

Introduction

Unstructured or minimally structured data appears in many domains—from large collections of raw text documents that are of interest in the social sciences, to unannotated genome sequences used in biological and biomedical applications. An important and challenging statistical problem is that of inferring and understanding the hidden underlying structure of such data: this open-ended task is generally referred to as *unsupervised learning*. A common way of imposing some constraints upon unsupervised learning problems is to look for structure in the form of clusters—that is, latent classes to which each datapoint is assigned. By finding groupings in the available observations, we hope to gain greater insight into the data, and in turn improve our predictive power.

In order to address tasks that involve clustering we employ probabilistic *mixture models*. These are generative models that assign likelihood to the observed data by

positing the existence of several underlying subpopulations or classes, each with different characteristics, from which the data is drawn. This generative description corresponds to an additive form of the model likelihood, which is represented as the sum of subpopulation likelihoods. As mixture models can possess a high number of free parameters, a need for regularization often arises in order to address the ill-posedness of the inferential problem and prevent overfitting to the training data. In keeping with the probabilistic nature of the models, regularization is most commonly achieved by imposing priors over the model parameters. This in turn means that inference of the posterior distribution of the model parameters can be naturally achieved through the Bayesian toolkit, which includes sampling and approximation methods.

As in many other areas of statistics and machine learning, an important distinction is the one between *parametric* and *nonparametric* mixture models. The former type of model involves a fixed number of free parameters, while the latter allows the number of parameters to grow with the size of the training data. While parametric models often have the advantage of faster inference, nonparametric models can be more general and flexible, requiring fewer modeling assumptions regarding the structure of the data. In the context of clustering, we are most often interest in allowing the number of classes to grow with the amount of data available, so as to retain model flexibility across multiple training size scales.

In this thesis, I study the problem of performing unsupervised learning using nonparametric mixture models, which allow for great generality. I focus particularly

on scenarios in which the effect of exogenous covariates is relevant to the modeling problem. Such problems may arise in a variety of settings. For instance, researchers in the social sciences might be interested in asking how the contents of documents in a collection vary as a function of author affiliation, date of publication, or other variables of interest. Similarly, biological scientists might be interested in understanding and quantifying how latent genetic patterns are impacted by an individual's population of ancestry, or other characteristics. Since the models needed to adddress these problems all involve incorporating the effect of such external variables into the data likelihood itself, I refer to them as *covariate-dependent nonparametric mixture models*.

1.1 Contributions

Expanding on recent literature on nonparametric mixture models, I present a general modeling framework based on the use of dependent Dirichlet process priors, which provide a natural way to integrate covariate information into the modeling process. Accordingly, I discuss the associated inferential issues. I then demonstrate the practical use of this framework by developing Covariate-Augmented Nonparametric Latent Dirichlet Allocation (C-LDA), a nonparametric mixture model that allows covariates to affect the generative process for data in a very general way. I introduce both Markov Chain Monte Carlo (MCMC) and variational inference procedures for estimating the model from data. After verifying the performance of the model on synthetic data, I test it in a range of practi-

cal applications, which can all be addressed via the use of covariate-dependent nonparametric mixture models.

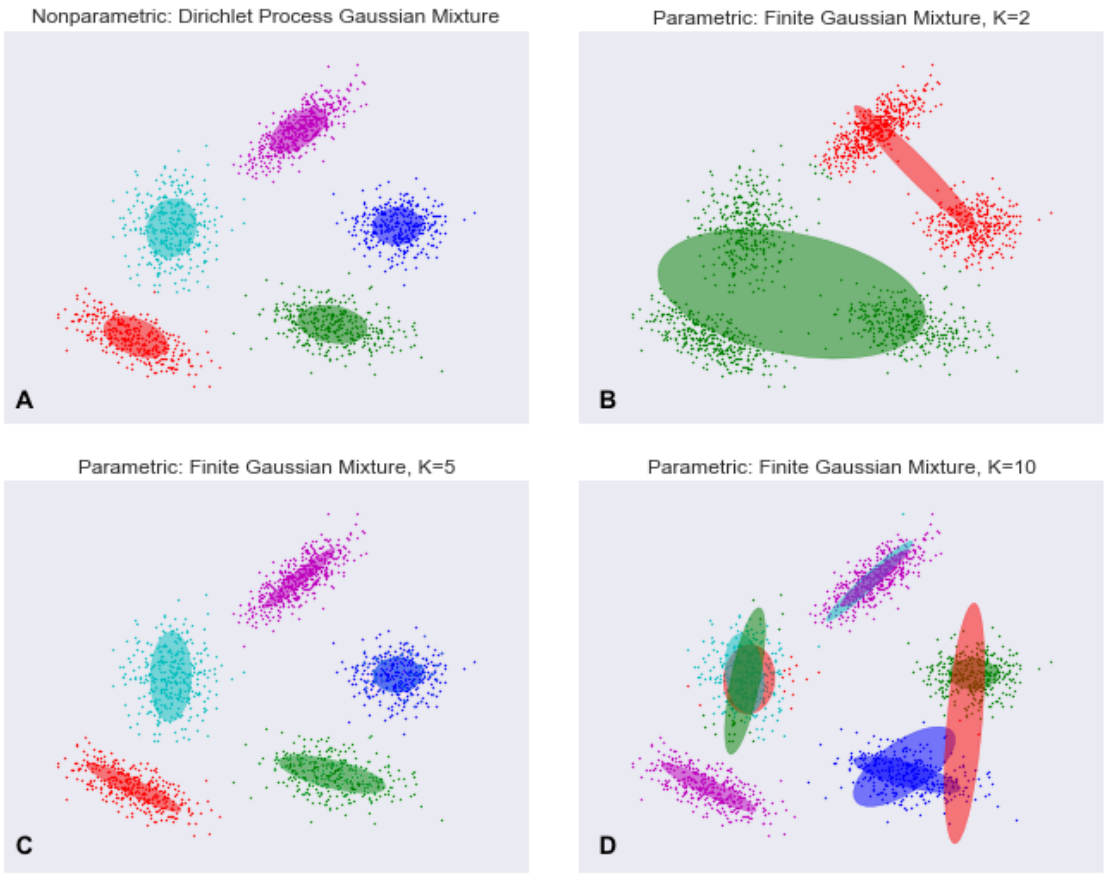

Figure 1.1: **A comparison of parametric and nonparametric statistical models.** This figure demonstrates clustering of a set of points drawn from a mixture of 5 bi-dimensional Gaussian distribution with randomly generated means and covariance matrices. (A) Clustering using a Gaussian mixture model with a nonparametric Dirichlet Proces prior (see chapter 2 for background): with a nonparametric mixture model, there is no need to fix the number of clusters *a priori*, since the complexity of the model can adapt to the size and structure of the data. (B, C, D) Clustering using a finite Gaussian mixture model, in which the number K of Gaussian components is fixed. While the true value $K = 5$ yields a good fit, inaccurate choices such as $K = 2$ and $K = 10$ render the model severely misspecified, since its complexity cannot be adapted dynamically.

The first application concerns haplotype phasing, which consists of the problem of identifying distinct genetic lineages in DNA sequence data. Using C-LDA, I show how the use of covariate information such as an individual's population of ancestry can aid in the estimation of latent haplotypes. The second application consists of modeling a corpus of articles from the opinion section of the New York Times, and studying the effect of date of publication on the topical structure of the corpus.

1.2 Related work

Mixture models have been successfully applied to unsupervised learning problems in a wide range of domains, including genomics and natural language processing. In both fields, much of the literature has focused on learning latent structure in mixed-membership models. While in *single-membership* models we allow each datapoint to belong to a single class only, we talk of *mixed-membership* models (Gross & Manrique-Vallier, 2014) if we admit assignment of a single observation to multiple classes (see figure 1.2 for an illustration of this concept). This additional flexibility is often apt in modeling real-world data.

A particularly popular application of mixed-membership models in the textual domain has been the class of statistical tools known as *topic models* (Blei, 2012). The input for this class of model is a reduced-complexity representation of raw textual documents, in which all word-ordering information is discarded by making the *bag-of-words* assumption: that is, by representing a document simply by

Figure 1.2: **Single-membership vs. mixed-membership clustering models**. (A) Hard K-means is a special case of a Gaussian mixture model, a single-membership mixture model. In such models, each datapoint is assigned to a single class. The red triangles show the imputed cluster means, and the data point coloring indicates the imputed cluster of origin. (B) Soft K-means has a probabilistic interpretation as a mixed-membership mixture model, in which data points have shared cluster responsibilities. In the plot, the shade of grey used to color a particular data point pictorially reflects the relative responsibilities of cluster 1 (black) and cluster 2 (white). For reference on both algorithms, see Bishop (2006). The data in this figure is drawn from a mixture of two bi-dimensional Gaussian distributions.

the counts of the words contained in it. Topic models then postulate that the observed words are sampled from multinomial probability vectors over the vocabulary, which are referred to as *topics*. In this framework, documents have mixed membership in the latent topics, with the respective contribution of each topic encoded in a document-topic distribution. The assignments of words to topics is made via draws from such document-topic distributions. This likelihood model allows for inference of the latent topics, which can then be used for semantic sum-

marization of the corpus. An early and widely adopted model in this class is the Latent Dirichlet Allocation (LDA) model of Blei et al. (2003). In LDA, Dirichlet priors are placed over the topic probability vectors. This choice not only leads to regularization of the model, but is also quite natural in that the Dirichlet distribution is conjugate prior to the multinomial distribution, such that inference in the model becomes particularly tractable.

The LDA model for collections of discrete data of Blei et al. (2003) has been extended numerous times. Some themes of notable interest have been models that introduce correlations among topics, such as Pachinko allocation (Li & McCallum, 2006) and the Correlated Topic Model (Blei & Lafferty, 2006a); nonparametric models that assume an unbounded number of mixture components, such as the Hierarchical Dirichlet Process (HDP) of Teh et al. (2006) as applied to document modeling; and covariate-dependent models such as the Dynamic Topic Model (Blei & Lafferty, 2006b). The latter line of work has resulted in applications ranging from studying the dynamics of a corpus over time, to developing matching methods for causal inference with high-dimensional data (Roberts et al., 2015b).

The notion of parameterizing a model's mixing weights by exogenous covariates is introduced by Roberts et al. (2015a) with the Structural Topic Model (STM). The authors propose a distinction between *content covariates*, which parameterize the topic-word probability vectors, and *prevalence covariates*, which parameterize the document-topic distributions. Allowing both content and prevalence covariates to take part in the generative process for documents has several distinct advantages.

First, if we believe that document-level meta-information does in fact affect the generative process, then it would be reasonable to expect that such generalization would lead to a better fit of the data: Roberts et al. (2015a) show that this is the case. Second, this extension lays down the theoretical groundwork for performing certain kinds of analyses. Suppose that a researcher wants to investigate the effect of age, or treatment assignment, or political affiliation (and so forth, generalizing to any arbitrary covariate) on *how much* authors discuss certain topics and also *how* they use words in discussing them. This type of analysis requires lifting the assumption that the documents are exchangeable with respect to external covariates, which in turn necessitates building the covariates into the model's generative process. An example analysis of this kind is the application to open-ended survey responses by Roberts et al. (2014b).

In a nonparametric setting, the inclusion of covariates in the generative process for a document collection is introduced by Kim & Sudderth (2011) with the Doubly Correlated Nonparametric Topic Model (DCTN). As discussed above, nonparametric models are desirable because they assume an infinite number of topics, allowing the number of *realized* topics to grow with the size of the data. Kim & Sudderth only introduce prevalence covariates in their model. Ideas from the DCTN and the STM will be explored again in chapter 3, where I develop C-LDA as a model that is nonparametric and includes both content and prevalence covariates in its likelihood function.

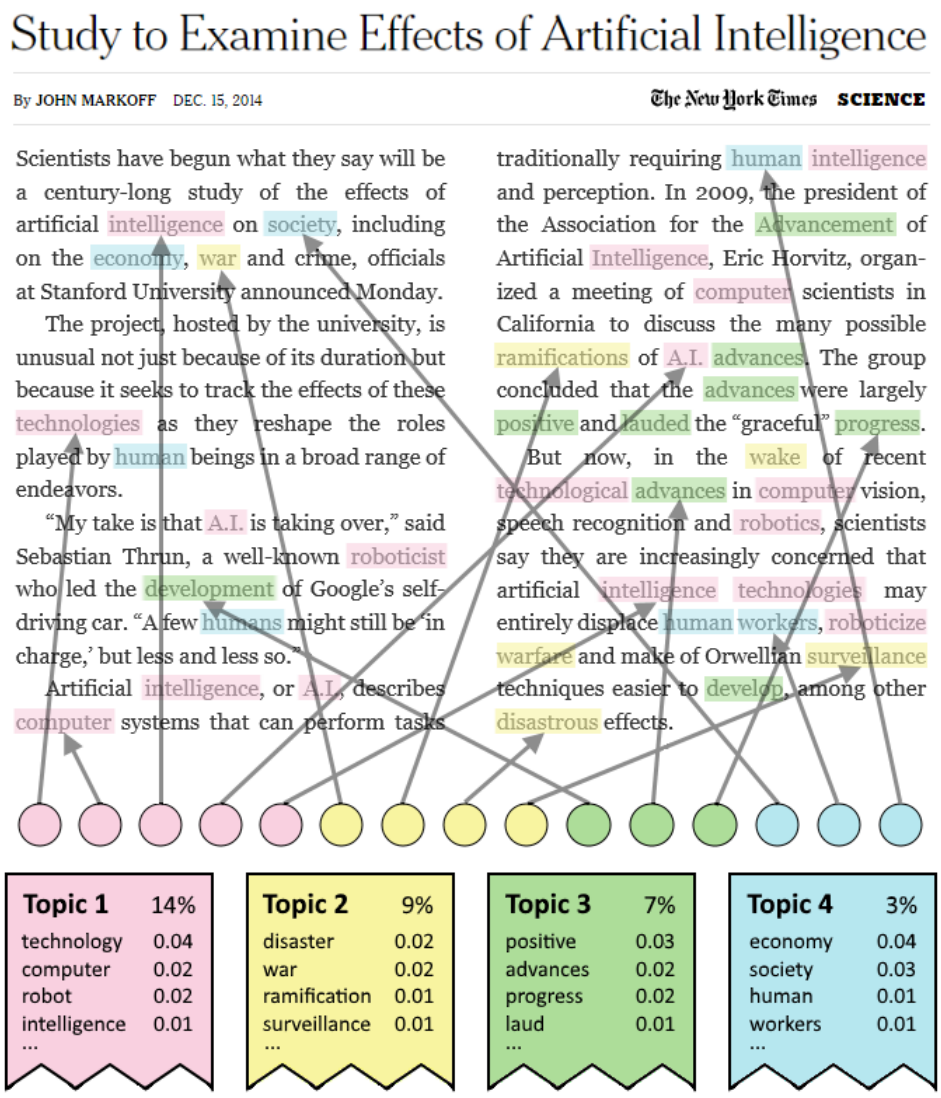

Figure 1.3: **Topic modeling: A textual application of mixed-membership mixture models**. In topic models we posit that each document is drawn from a mixture of latent topics—that is, distributions over the vocabulary. Models such as Latent Dirichlet Allocation (LDA) assume that each observed word in a document is first assigned to a specific topic by performing a multinomial draw from the document-topic distribution, and then sampled from the corresponding topic-word distribution. Figure inspired by Blei (2012).

2

Dependent Dirichlet Processes as Mixture Model Priors

The development of sound theoretical tools and tractable inference procedures for nonparametric Bayesian models has allowed researchers to study and apply models that can grow more complex as more data is observed. In models aimed at clustering data, the nonparametric prior of choice is often the Dirichlet process (DP), a measure on the space of distribution functions. In this chapter I review the theory of infinite mixture models via Dirichlet process priors. I then introduce the dependent Dirichlet process (DDP), an extension of the DP prior that allows for correlation among the process realizations through the intervention of external covariates. The DDP provides a solution to modeling problems where we wish to retain the flexibility of nonparametric Bayesian models, while avoiding restrictive assumptions of independence among observations, and incorporating covariate information into the modeling process.

2.1 Dirichlet processes

In this section I follow the exposition of the theory of Dirichlet processes by Murphy (2012). The Dirichlet process is a stochastic process whose realizations are discrete probability distributions. As we shall see, the Dirichlet process is useful as a prior for the parameters of a data generating process in nonparametric clustering problems. Let Θ be a valid probability space, such as the Borel sets over $\mathbb{R}^n$. Consider an arbitrary valid probability distribution G over this space. Let $T_1, T_2, \cdot, T_k$ be a finite measurable partition of the space Θ, such that $\Big(G(T_1), G(T_2), \ldots, G(T_k)\Big)$ is a random vector. Given a base distribution H over Θ and a scalar concentration parameter α, G is said to follow a Dirichlet process in distribution if this random vector is jointly distributed as

$$\Big(G(T_1), G(T_2), \ldots, G(T_k)\Big) \sim \text{Dir}\Big(\alpha H(T_1), \alpha H(T_2), \ldots, \alpha H(T_k)\Big)$$

One very useful way to construct a probability distribution that follows a Dirichlet process is by the so-called *stick-breaking construction*, which explicitly highlights the properties that make the DP ideally suited to clustering problems. The setup is the following: suppose we wish to make a draw from a stochastic process that is a countably infinite weighted sum of atoms (point masses). In order for this draw to be well-suited to act as a prior in a mixture model we wish for several of the generated mixture weights to be relatively large. In order to accomplish this, we let the infinite set of mixture weights $\{\pi_k\}_{k=1}^{\infty}$ be constructed via the following

generating process:

$$\pi_k = \beta_k \cdot \prod_{j=1}^{k-1}(1-\beta_j) \quad \text{where} \quad \beta_k \overset{iid}{\sim} \text{Beta}(1, \alpha)$$

Informally, we start with a stick of unit length (the full probability mass), and draw a Beta-distributed random variable with support in the unit interval to choose a 'breaking point'. We break the stick at the breaking point, set aside the leftmost part of it, and repeat the breaking process on the remaining part of the stick. In the limit of infinite breaks, this process yields the desired weights.

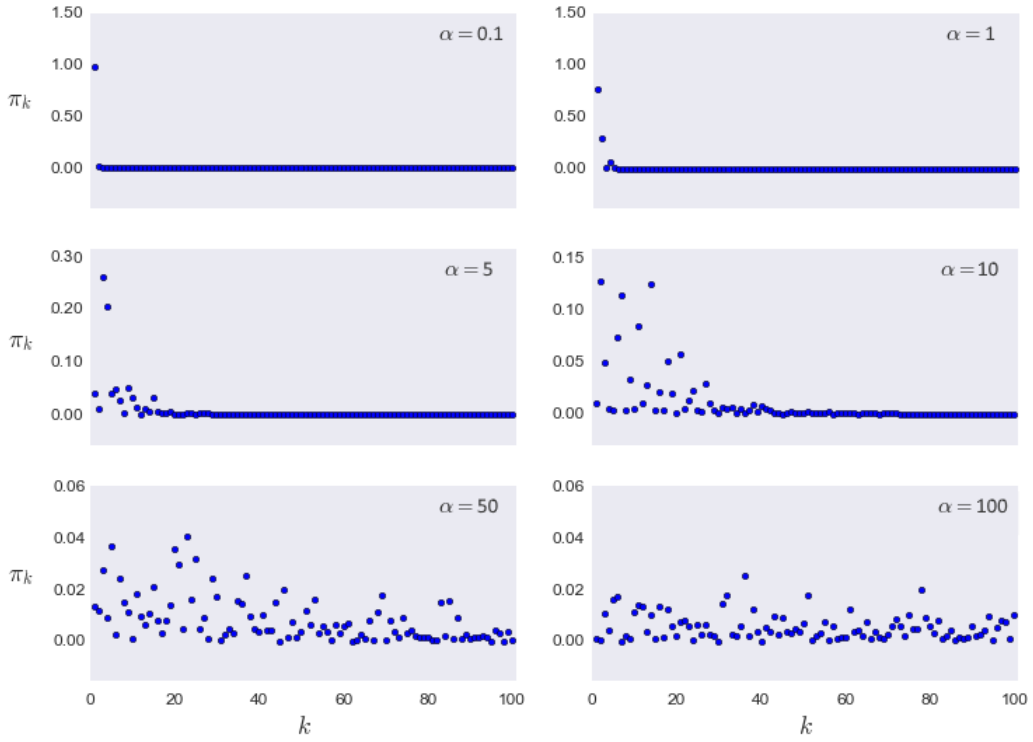

Figure 2.1: **Stick-breaking construction of Dirichlet process weights**. This figure shows sample weights $\{\pi_k\}$ obtained via stick-breaking constructions for different values of the concentration parameter α. Notice that lower values of the concentration parameter induce higher sparsity in the weight distribution by forcing more of the weights to be zero.

Given the base measure H, we can draw atoms $\theta_k \sim H$. These are realizations of H that will serve as the components of a draw from the Dirichlet process. Letting $\delta_{\theta_k}(\theta)$ be a Kronecker delta function centered at θ_k, we can then construct G by taking a weighted average of the atoms according to the stick weights:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \, \delta_{\theta_k}(\theta)$$

Constructed as such, G can be proved to follow a Dirichlet process. The proof is included in appendix B. Having given a formal definition of the Dirichlet process, we can now observe the properties that make it useful in the setting of clustering problems. While the base distribution can be either continuous or discrete, draws from the Dirichlet process are almost surely discrete, which allows for assignment of multiple data points to a single cluster with positive probability. Moreover, the stick-breaking construction guarantees that a few clusters will dominate the solution by enforcing sparsity: this is a desirable property as it leads to more parsimonious and tractable models.

2.2 Dependent Dirichlet processes

Following the exposition by MacEachern (2000) and Müller & Rodriguez (2013), I now turn to discuss the dependent Dirichlet process. The DPP is a generalization of the Dirichlet process that allows for correlation among its realization, mediated by a covariate x. It is particularly elegant in that it builds rather simply on top of the stick-breaking construction described above. To define the DPP, we let

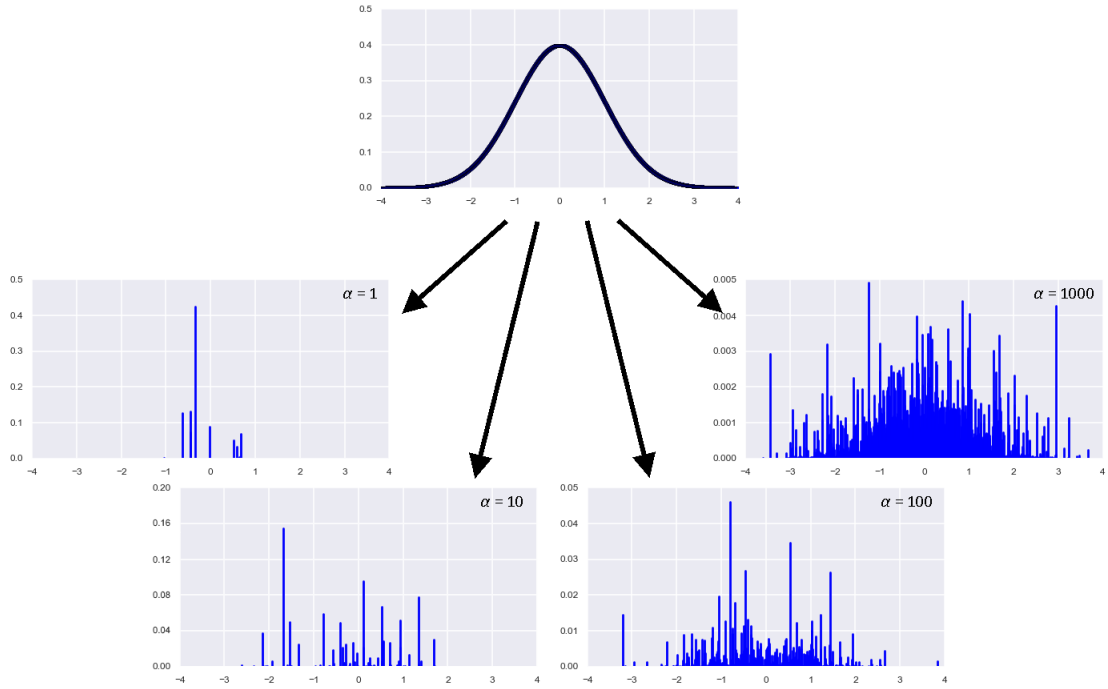

Figure 2.2: **Sample draws from a Dirichlet process**. In this case we have a standard Gaussian distribution as the base measure. The draws are performed using the stick-breaking construction. As noted in figure 2.1, lower values of the concentration parameter α induce more weight sparsity. Also notice that all the draws from the Dirichlet process are discrete probability distributions over the support of the base measure.

$\mathcal{X}$ be the support of the covariate x. The core idea in the theory of the DPP is that we will replace the atoms θ_k with sample paths from a valid stochastic process $\{\theta_{x,k}\}_{k=1}^{\infty}$ on $\mathcal{X}$ (in a simple example, this could be a Gaussian process), which specifies a value for the atom *as a function of the covariate*. This lets the location of the DP point masses be dependent of x. Similarly, we can replace the mixture weights π_k with stochastic processes $\{\pi_{x,k}\}_{k=1}^{\infty}$, which specify weights as functions of the covariate level (thus letting the covariate affect the prevalence of the atoms of G). The only restriction on the process $\{\pi_{x,k}\}_{k=1}^{\infty}$ is that it should be

a map of the type $\mathcal{X} \mapsto C_\infty$, where C_∞ is the infinite-dimensional simplex. Given the processes $\{\theta_{x,k}\}_{k=1}^{\infty}$ and $\{\pi_{x,k}\}_{k=1}^{\infty}$, a draw from the DPP is then constructed analogously to the simple DP case as

$$G_x(\theta) = \sum_{k=1}^{\infty} \pi_{x,k}\, \delta_{\theta_{x,k}}(\theta)$$

An explicit example of the construction of the DDP mixture weights π_k dependent on data ρ_k is given by Ren et al. (2011), who point out that this can be achieved by choosing a link function $g(\cdot)$ whose codomain is the unit interval. Then stick-breaking can be realized by the process

$$\pi_k(\rho_k) = g(\rho_k) \cdot \prod_{j=1}^{k-1}(1 - g(\rho_j))$$

In the limit of $k \to \infty$ the weights will sum to one, and the concentration is now controlled by the variance of the data ρ_k. If we choose $g(\cdot)$ to be sigmoid function, we refer to this as a *logistic stick-breaking process*. In the next chapter, we will see how the logistic stick-breaking construction will be useful as we build practical models for covariate-dependent, nonparametric clustering.

3

Covariate-Dependent Nonparametric LDA

In this chapter I introduce Covariate-Dependent Latent Dirichlet Allocation (C-LDA), a novel model that demonstrates the modeling concepts discussed in chapters 1 and 2. The C-LDA model is concisely described by the graphical model in figure 3.1, or by the generative process detailed in section 3.1. In chapter 4, I introduce inference procedures for the model. In chapter 5, I apply the model to a variety of settings, including textual and genomic data.

C-LDA draws on ideas from the STM model of Roberts et al. (2015a) and the DCTM model of Kim & Sudderth (2011), both of which were discussed in chapter 1. The notation and general setup for C-LDA are based on the DCTM. In contrast with the DCTM, and like in the STM, C-LDA includes the effects of both prevalence and content covariates in its likelihood model. This provides a very

broad and flexible framework with which to study the impact of covariates on the latent structure of the data. Moreover, in contrast with the STM, which introduces covariates into the model by means of a generalized linear model, C-LDA chiefly relies on the Dirichlet-Multinomial conjugacy that is at the heart of the LDA model of Blei et al. (2003).

The following is a detailed description of the model's construction, along with helpful alternate representations of the model that point to its fundamental nature as a derivate of the dependent Dirichlet process.

3.1 Explicit construction of the C-LDA model

We let D be the total number of documents in the corpus and N_d be the number of words w_{id} in document d. We also let $x_d \in \mathbb{R}^F$ be a vector of prevalence covariates associated with each document d. Given also coefficient vectors $\eta_k \in \mathbb{R}^F$ and variance hyperparameter σ_ρ^2, we construct document-topic scores according to a Gaussian distribution:

$$\rho_{dk} | \eta_k, x_d, \sigma_\rho^2 \sim \mathcal{N}\left(\eta_k^T x_d,\, \sigma_\rho^2\right)$$

The scores $\{\rho_{dk}\}$ introduce dependency of the topic proportions on the prevalence covariates, and can be thought of as un-normalized versions of the document-topic frequencies. Normalization according to the logistic stick-breaking process will ensure that the transformed scores result in valid probability vectors. Let $\sigma(\cdot)$

denote the sigmoid function for univariate arguments, and the softmax function for multivariate arguments, so that

$$\sigma(x) = \begin{cases} \frac{1}{1+e^{-x}} & \text{if } x \in \mathbb{R} \\ \left(\frac{e^{x_1}}{\sum_{i=1}^{n} e^{j}}, \ldots, \frac{e^{x_n}}{\sum_{i=1}^{n} e^{j}} \right) & \text{if } x \in \mathbb{R}^n,\ n > 1 \end{cases}$$

As illustrated in chapter 2, nonparametric document-topic distributions can then obtained using the logistic stick-breaking process:

$$\pi_{dk} = \sigma(\rho_{dk}) \prod_{j=1}^{k-1} [1 - \sigma(\rho_{dj})]$$

The values $\{\pi_{dk}\}$ are the normalized document-topic frequencies. Given these frequencies, the word assigments are then drawn multinomially:

$$z_{id} | \pi_d \sim \text{Mult}(\pi_d)$$

Similarly, we have $y_d \in \mathbb{R}^G$ be content covariates, with coefficient vectors α_v. In order to parameterize the topic-word distributions using the content covariates, document-word scores are then constructed by drawing

$$\theta_{dv} | \alpha_v, y_d, \sigma_\theta^2 \sim \mathcal{N}\left(\alpha_v^T y_d,\ \sigma_\theta^2\right)$$

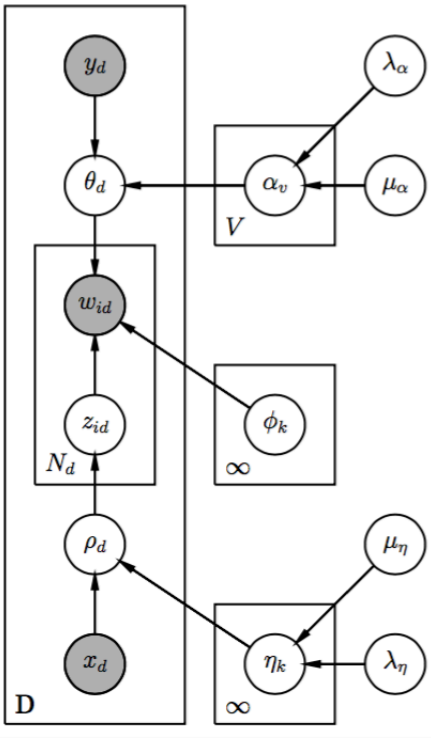

Figure 3.1: **C-LDA as a directed graphical model**. Arrows denote dependencies between variables, and plates denote repetition. The graph specifies a factorization of the joint distribution of the model's variables via a set of conditional independence relations. Prior hyperparameters are not shown.

The base measure for the topic-word distributions is given by $\phi_k \sim \text{Dir}(\beta)$, with β being a vector of hyperparameters that controls the prior concentration of probability mass in the topic-word distributions. Let $\odot$ indicate the elementwise product. The document-specific distributions are then obtained again via a logistic construction, using the softmax function, and words drawn categorically given their topic assignments

$$\psi_{dk} = \sigma\left(\theta_d \odot \phi_k\right) \qquad w_{id}|\{\phi_k\}, \theta_d, z_{id} \sim \text{Mult}(\psi_{d,z_{id}})$$

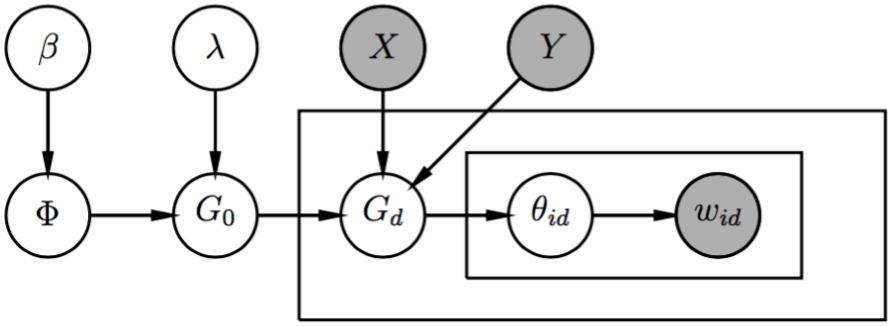

Figure 3.2: **Alternate C-LDA representation as dependent hierarchical Dirichlet process**. The C-LDA model can also isomorphically be represented as a mixture model with a Hierarchical Dirichlet Process prior, where the inner DP is a dependent Dirichlet process.

We assign conjugate priors to the coefficient vectors, letting $\Lambda_\eta \in \mathbb{R}^{F\times F}$ and $\Lambda_\alpha \in \mathbb{R}^{G\times G}$ be diagonal precision matrices with diagonal elements $\lambda_{\eta,f}$ and $\lambda_{\alpha,g}$, respectively:

$$\eta_k \sim \mathcal{N}(\mu_\eta, \Lambda_\eta^{-1}) \qquad \alpha_v \sim \mathcal{N}(\mu_\alpha, \Lambda_\alpha^{-1})$$

The prior parameters are given conjugate hyperpriors. Hence, given hyperparameters a_η, b_η, a_α, b_α, γ_η, γ_α, we have

$$\mu_{\eta,f} \overset{\text{iid}}{\sim} \mathcal{N}(0, \gamma_\eta) \qquad \mu_{\alpha,g} \overset{\text{iid}}{\sim} \mathcal{N}(0, \gamma_\alpha)$$
$$\lambda_{\eta,f} \overset{\text{iid}}{\sim} \Gamma(a_\eta, b_\eta) \qquad \lambda_{\alpha,g} \overset{\text{iid}}{\sim} \Gamma(a_\alpha, b_\alpha)$$

3.2 Alternate representation as a dependent hierarchical Dirichlet process

The generative process described here can also be isomorphically represented as a hierarchical Dirichlet process (Teh et al., 2006) where we allow the inner DP process to be covariate-dependent. With base measure $\Phi \sim \text{Dir}(\beta)$, we define a first DP controlled by concentration parameter λ which acts analogously to the concentration parameters Λ_η, Λ_α in the previous representation:

$$G_0 \sim \text{DP}(\Phi, \lambda)$$

We then sample document-specific distributions G_d from a dependent DP where a random measure is constructed via covariate-dependent components. Note here H is the matrix obtained by stacking of vectors η_k, and A is the stacking of vectors α_v. Letting δ_x be the Kronecker delta function centered at x, we have

$$G_d(x_d, y_d; H, A, G_0) = \sum_{k=1}^{\infty} C_k(x_d; \eta_k)\, \delta_{\phi(G_0, y_d; A)}$$

The function C is parameterized by H and allows the stick weights (topic proportions) to depend on the prevalence covariates. The functional ϕ is parameterized by A and allows the atoms to depend on the content covariates. Given G_d, we sample probability vectors θ_{id} and make categorical draws for the words.

Overall, the structure of the C-LDA model allows for a large amount flexibility in

introducing covariate dependencies within the context of a nonparametric mixture model, while at the same time retaining conjugacies that will make inference tractable. In chapter 5, I will show how the estimates of the latent covariate coefficients $\{\eta_k\}$ and $\{\alpha_v\}$ can be used to construct summary statistics that offer a powerful way to quantify and summarize the relationships between the covariates of interest and the latent structure of the data. This will highlight the effectiveness of C-LDA as a tool of research.

4

Inference in the Model

In this chapter I derive two inference schemes for the C-LDA model introduced in the previous section. A valid inference procedure allows us to estimate the posterior distribution of the free parameters in the model. The first inference scheme presented here, a Gibbs sampler, has the advantage of being exact, but as most sampling-based inference algorithms, it suffers from poor scalability and difficulties in monitoring convergence to the true posterior distribution. The second inference scheme relies on a variational approximation to the true posterior distribution. Although approximate, variational inference has the advantages of better scalability and general performance, especially as it is easily adaptable to online settings. The variational inference scheme also provides more easily monitorable convergence statistics.

4.1 Markov Chain Monte Carlo (MCMC) inference

To start with, for inference in the model we develop a non-collapsed Gibbs sampler based on the explicit stick-breaking representation of the C-LDA generative process. In order to make the problem finite, we approximate the stick-breaking prior via truncation, letting $\tilde{K}$ be a (possibly loose) upper bound on the number of clusters. To achieve maximum generality we can set $\tilde{K} = D$, since we can never observe more clusters than data points, but Ishwaran & James (2001) show that tighter upper bounds of order $O(\log D)$ also result in excellent approximations of the nonparametric prior, which gives significant improvements in computational performance.

4.1.1 Non-conjugate updates via Metropolis-Hastings sampling

The marginal posterior distributions for $\{\rho_d\}$, $\{\theta_d\}$, and $\{z_{in}\}$ cannot be computed in closed form because the respective likelihoods and priors are non-conjugate. In order to derive the Gibbs updates, we follow the algorithm proposed by Neal (2000), which makes use of a Metropolis-Hastings independence sampler. We use the marginal priors as the proposal distributions, so that by the Metropolis-Hastings rule the acceptance probabilities can be computed as a ratio of likelihoods. Hence, in order to sample ρ_d given all other variables, we first propose candidates from the prior distribution

$$q(\rho_d^*|\rho_d) = q(\rho_d^*) = p(\rho_d^*|x_d, \{\eta_k\})$$

We then accept the proposal with probability

$$\begin{aligned} T(\rho_d^*|\rho_d) &= \min\left[1, \frac{q(\rho_d)}{q(\rho_d^*)} \cdot \frac{p(\rho_d^*|\{z_{id}\}, x_d, \{\eta_k\})}{p(\rho_d|\{z_{id}\}, x_d, \{\eta_k\})}\right] \\ &= \min\left[1, \frac{p(\rho_d|x_d, \{\eta_k\})}{p(\rho_d^*|x_d, \{\eta_k\})} \cdot \frac{p(\rho_d^*|x_d, \{\eta_k\})}{p(\rho_d|x_d, \{\eta_k\})} \cdot \frac{p(\{z_{id}\}|\pi_d^*)}{p(\{z_{id}\}|\pi_d)}\right] \\ &= \min\left[1, \prod_{k=1}^{\tilde{K}} \left(\frac{\pi_{dk}^*}{\pi_{dx}}\right)^{\sum_{i=1}^{N_d} \mathbb{I}(z_{id}=k)}\right] \end{aligned}$$

Notice that the priors cancel in the acceptance probability, and we are then left with a tractable ratio of likelihoods. Similarly we sample θ_n by proposing candidates from the distribution $q(\theta_d^*|\theta_d) = q(\theta_d^*) = p(\theta_d^*|y_d, \{\alpha_v\})$ and then accepting with probability

$$\begin{aligned} T(\theta_d^*|\theta_d) &= \min\left[1, \frac{q(\theta_d)}{q(\theta_d^*)} \cdot \frac{p(\theta_d^*|y_d, \{w_{id}\}, \{\alpha_v\}, \{z_{id}\}, \{\phi_k\})}{p(\theta_d|y_d, \{w_{id}\}, \{\alpha_v\}, \{z_{id}\}, \{\phi_k\})}\right] \\ &= \min\left[1, \prod_{i=1}^{N_d} \frac{\psi_{d,z_{id},w_{id}}^*}{\psi_{d,z_{id},w_{id}}}\right] \end{aligned}$$

Lastly, we sample the topic assignments by similarly proposing candidates from $q(z_{id}^*|z_{id}) = q(z_{id}^*) = p(z_{id}^*|\rho_d)$ and accepting with probability

$$\begin{aligned} T(z_{id}^*|z_{id}) &= \min\left[1, \frac{q(z_{id})}{q(z_{id}^*)} \cdot \frac{p(z_{id}^*|\rho_d, w_{id}, \{\phi_k\}, \theta_d)}{p(z_{id}|\rho_d, w_{id}, \{\phi_k\}, \theta_d)}\right] \\ &= \min\left[1, \frac{\psi_{d,z_{id}^*,w_{id}}}{\psi_{d,z_{id},w_{id}}}\right] \end{aligned}$$

4.1.2 Conjugate updates

Using conjugacy results, the marginal posterior distributions of the rest of the variables in the the model can be computed in closed form. Following Kim & Sudderth (2011), we set $\sigma_\rho^2 = \sigma_\theta^2 = 1$. This causes no loss of generality because the values $\{\rho_{dk}\}$ and $\{\theta_{dv}\}$ are subsequently normalized to yield respectively the document-topic distributions $\{\pi_d\}$ and the document-specific topic-word distributions $\{\psi_{dk}\}$. For the precision parameters we have the following Gibbs updates:

$$\begin{aligned} p(\lambda_{\eta f}|\{\eta_k\}, \mu_\eta, \{\rho_d\}, \{x_d\}; a_\eta, b_\eta) &\propto p(\lambda_{\eta f}|a_\eta, b_\eta) \cdot \prod_{k=1}^{\tilde{K}} p(\eta_{fk}|\mu_{\eta f}, \lambda_{\eta f}^{-1}) \\ &\propto \Gamma(\lambda_{\eta f}|a_\eta, b_\eta) \cdot \prod_{k=1}^{\tilde{K}} \mathcal{N}(\eta_{fk}|\mu_{\eta f}, \lambda_{\eta f}^{-1}) \\ &\propto \Gamma\left(\lambda_{\eta f} \middle| \frac{\tilde{K}}{2} + a_\eta,\ \frac{1}{2}\sum_{k=1}^{\tilde{K}}(\eta_{fk} - \mu_{\eta f})^2 + b_\eta\right) \end{aligned}$$

The update for $\lambda_{\alpha,g}$ is analogous. For the coefficient means, which encode the estimated covariate effects, we have

$$\begin{aligned} p(\mu_{\eta f}|\gamma_{\mu_\eta}, \{\eta_k\}, \lambda_\eta) &\propto \mathcal{N}(\mu_{\eta f}|0, \gamma_{\mu_\eta}) \cdot \prod_{k=1}^{\tilde{K}} \mathcal{N}(\eta_{fk}|\mu_{\eta f}, \lambda_{\eta f}^{-1}) \\ &\propto \mathcal{N}\left(\mu_{\eta f} \middle| \frac{\gamma_{\mu_\eta} \cdot \sum_{k=1}^{\tilde{K}} \eta_{fk}}{\tilde{K}\gamma_{\mu_\eta} + \lambda_{\eta f}^{-1}}, (\gamma_{\mu_\eta}^{-1} + \tilde{K}\lambda_{\eta f})^{-1}\right) \end{aligned}$$

The analogous result applies for $\mu_{\alpha,g}$. From Dirichlet-Multinomial conjugacy, the

update for ϕ_k simply relies on word counts for the words assigned to topic k:

$$p(\phi_k|\{w_{id}\},\{z_{id}\},\{\theta_n\},\beta) = \text{Dir}\left(\phi_k \,\middle|\, \left[\beta + \sum_{(i,d)\,:\,z_{id}=k} \mathbb{I}(w_{id}=1)\,,\,\ldots\right]\right)$$

Lastly, the updates for the coefficient vectors η_k are

$$\begin{aligned} p(\eta_k \mid \mu_\eta, \lambda_\eta, \{x_d\}, \{\rho_d\}) &\propto \mathcal{N}\left(\eta_k \mid \mu_\eta, \Lambda_\eta^{-1}\right) \cdot \prod_{d=1}^{D} \mathcal{N}(\rho_{dk} \mid \eta_k^T x_d, 1) \\ &\propto \mathcal{N}\left(\eta_k \mid [\Lambda_\eta + X^T X][X^T \rho_{:k} + \Lambda_\eta \mu_\eta], \Lambda_\eta + X^T X\right) \end{aligned}$$

And similarly for the coefficient vectors α_v we have

$$\begin{aligned} p(\alpha_v \mid \mu_\alpha, \lambda_\alpha, \{y_d\}, \{\theta_d\}) &\propto \mathcal{N}\left(\alpha_v | \mu_\alpha, \Lambda_\alpha^{-1}\right) \cdot \prod_{d=1}^{D} \mathcal{N}(\theta_{dv} | \alpha_v^T y_d, 1) \\ &\propto \mathcal{N}\left(\alpha_v \mid [\Lambda_\alpha + Y^T Y][Y^T \theta_{:v} + \Lambda_\alpha \mu_\alpha], \Lambda_\alpha + Y^T Y\right) \end{aligned}$$

4.2 Variational inference

While solving the inference problem exactly, the sampling-based inference scheme presented above suffers from a number of drawbacks. First and most obviously, there are no easily viable checks to verify convergence of the Gibbs sampler to the stationary posterior distribution. Second, the uncollapsed sampler is rather inefficient memory-wise, as we need to store a simulated path for all the latent variable assignments. Third, and perhaps most importantly, the Metropolis-Hastings updates for the non-conjugate steps require drawing repeated samples from the prior,

which can be very numerous when the sampler becomes trapped in a low-density area of the corresponding acceptance distribution. Practical implementation of the model shows that, unless the model's hyperparameters are very carefully tuned, the prior samples required to achieve a single acceptance can in fact number in the tens of thousands, which slows down the sampling process significantly. A summary comparison of sampling-based methods and variational inference along several relevant dimension is provided in table 4.1.

Performing inference via a variational approximation of the joint posterior distribution provides an alternative that eases these problems, at the expense of the possibility of achieving samples from the *exact* posterior distribution. In contrast with the MCMC procedure, variational inference yields an (approximate) closed-form solution for the posterior distribution rather than samples. It optimizes a lower bound on the model's marginal likelihood, which provides an immediate way to monitor convergence, and it does not require storing an entire sampling path. In addition, variational inference enjoys the advantage of easily being adapted to online data by computing stochastic versions of the bound gradients by either streaming single documents or using minibatches. In either case, this property implies that the entire inference procedure can be completed in the course of a single pass over the dataset, which affords much better performance as compared to a sampling-based scheme.

Given the non-conjugate nature of the C-LDA model, the variational updates cannot be derived via common conjugate methods. Instead, following the recent

work on variational inference in non-conjugate models of Wang & Blei (2013), we use *Laplace variational inference*, which exploits local Laplace approximations after assuming a factorization of the joint posterior distribution over all the factor variables in the model. A Laplace approximation consists of approximating a target density by a Gaussian dentity, whose shape parameters are obtained by performing a Taylor expansion of the original density around its mode. Once again we base the inference procedure on the explicit stick-breaking representation of the model, which is more amenable to this task.

4.2.1 The variational framework

We let $p(\theta, \alpha, \mu, \lambda, \Phi, Z, \pi, \rho, \eta \,|\, X, Y, W)$ be the exact joint posterior distribution of the model's latent variables. We let $q(\theta, \alpha, \mu, \lambda, \Phi, Z, \pi, \rho, \eta)$ be the variational approximation to p, and start by making the mean-field assumption, meaning that we assume q factors over all its component variables, so that

$$q(\theta, \alpha, \mu, \lambda, \Phi, Z, \rho, \eta) = q(\theta)\, q(\alpha)\, q(\mu)\, q(\lambda)\, q(\Phi)\, q(Z)\, q(\rho)\, q(\eta)$$

Notice that for parsimony, we use the notation q to denote several different distributions, each being identified by its argument. From the conditional independence properties of the model, we know that these factors will further decompose, and we can write the following factorizations:

$$q(\theta) = \prod_{d=1}^{D} q(\theta_d), \qquad q(\alpha) = \prod_{v=1}^{V} q(\alpha_v), \qquad q(\mu) = q(\mu_\eta) \cdot q(\mu_\alpha)$$

$$q(\lambda) = q(\lambda_\eta) \cdot q(\lambda_\alpha), \qquad q(\Phi) = \prod_{k=1}^{\tilde{K}} q(\phi_k), \qquad q(Z) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} q(z_{id})$$

$$q(\rho) = \prod_{d=1}^{D} q(\rho_d), \qquad q(\eta) = \prod_{k=1}^{\tilde{K}} q(\eta_k)$$

The mean-field assumption is common in the literature and leads to a tractable, fully-specified model while still allowing great flexibility in the form of the marginal distributions of the model's variables.

The key proposition in variational inference is that we will be turning the inference problem into an optimization problem by minimizing the Kullback-Leibler (KL) divergence between q and p, where the KL divergence is defined as

$$\mathrm{KL}[q(\theta, \ldots, \eta) \,||\, p(\theta, \ldots, \eta \,|\, X, Y, W)] = \mathbb{E}_q \left[\log \frac{q(\theta, \ldots, \eta)}{p(\theta, \ldots, \eta \,|\, X, Y, W)} \right]$$

We cannot directly optimize this quantity given that $p(\theta, \ldots, \eta \,|\, X, Y, W)$ is intractable. In order to get around this limitation, we can work instead with the unnormalized joint distribution of the latent variables and the observed data, which is proportional to $p(\theta, \ldots, \eta \,|\, X, Y, W)$ in terms of the latent variables. Letting $\tilde{p}(\theta, \ldots, \eta\,,\, X, Y, W)$ be the unnormalized joint distribution, we in fact have that

$$\tilde{p}(\theta, \ldots, \eta, X, Y, W) = p(X, Y, W) \cdot p(\theta, \ldots, \eta \,|\, X, Y, W)$$

where $p(X, Y, W)$ is the model evidence, which is independent of the latent variables. To verify correctness, we now consider minimizing $\mathrm{KL}(q||\tilde{p})$, and show that

it is equivalent to minimizing $\mathrm{KL}[q(\theta, \ldots, \eta) \,||\, p(\theta, \ldots, \eta \,|\, X, Y, W)]$:

$$\begin{aligned}
\mathrm{KL}(q||\tilde{p}) &= \int q(\theta, \ldots, \eta) \cdot \log \frac{q(\theta, \ldots, \eta)}{\tilde{p}(\theta, \ldots, \eta, X, Y, W)} \, d\theta \ldots \, d\eta \\
&= \int q(\theta, \ldots, \eta) \cdot \log \frac{q(\theta, \ldots, \eta)}{p(X, Y, W) \cdot p(\theta, \ldots, \eta \,|\, X, Y, W)} \, d\theta \ldots \, d\eta \\
&= \int q(\theta, \ldots, \eta) \cdot \log \frac{q(\theta, \ldots, \eta)}{p(\theta, \ldots, \eta \,|\, X, Y, W)} \, d\theta \ldots \, d\eta - \log p(X, Y, W)
\end{aligned}$$

Note that the last step follows from the fact that $q(\theta, \ldots, \eta)$, being a normalized proability distribution, must integrate to one over its full support. Continuing, this yields

$$\mathrm{KL}(q||\tilde{p}) = \mathrm{KL}(q||p) - \log p(X, Y, W) \tag{4.1}$$

Therefore $\mathrm{KL}(q||\tilde{p})$ corresponds to $\mathrm{KL}(q||p)$ up to an additive constant, which is independent of the variational parameters. This same reasoning proves that $-\mathrm{KL}(q||\tilde{p})$ constitutes a lower bound on the marginal likelihood of the model, since

$$-\mathrm{KL}(q||\tilde{p}) = \log p(X, Y, W) - \mathrm{KL}(q||p) \leq \log p(X, Y, W)$$

We thus define the quantity $\mathcal{L}(q) \equiv -\mathrm{KL}(q||\tilde{p}) = \mathbb{E}_q[\log \tilde{p}(\theta, \ldots, \eta, X, Y, W)] - \mathbb{E}_q[\log q(\theta, \ldots, \eta)]$ to be the objective function for variational inference. Another standard result of mean-field variational theory (Murphy, 2012) is that in order for $\mathcal{L}(q)$ to achieve a maximum, each factor $q_i(\omega_i)$ of the optimal solution q^* must

satisfy the relation

$$q_i(\omega_i) \propto \exp\left\{\mathbb{E}_{-q_i}[\log \tilde{p}(\theta, \ldots, \eta, X, Y, W)]\right\} \tag{4.2}$$

These optimality conditions allow us to derive an iterative optimization procedure, whereby each factor is updated via coordinate ascent until convergence.

4.2.2 Laplace approximations for nonconjugate variables

While equation (4.2) provides a handy rule for constructing the variational updates, these updates remain hardly tractable for the nonconjugate variables, as they do not lead to a known closed-form solution for q_i. An approach to variational inference in nonconjugate models such as C-LDA known as *Laplace variational inference* is discussed in Wang & Blei (2013). The core idea behind Laplace variational inference is to make use of a Laplace approximation in the occurrence on nonconjugacy in the mean-field update equation (4.2).

A Laplace approximation of a twice-differentiable function discards all terms of order higher than two in the Taylor expansion of the logarithm of the function around its mode. For probability densities, this means discarding all information about moments beyond the second one—and thus approximating the unknown distribution by a Gaussian density. The following paragraphs describe the Laplace approximation in greater detail, largely following the treatment given in Wang & Blei (2013).

Consider an intractable posterior $p(\theta|x)$, proportional to a tractable joint distribution $p(\theta, x)$, and let $\hat{\theta}$ be the maximum *a posteriori* (MAP) of $p(\theta|x)$, which can be found by maximizing the joint density. Letting $H(\theta)$ be the Hessian matrix of $\log p(\theta|x)$, a second-order Taylor expansion of $\log p(\theta|x)$ around $\hat{\theta}$ results in

$$\log p(\theta|x) \approx \log p(\hat{\theta}|x) + \frac{1}{2}(\theta - \hat{\theta})^T H(\hat{\theta})(\theta - \hat{\theta}) \tag{4.3}$$

No first-order term appers in this expansion because we assumed that $\hat{\theta}$ is a local optimum of $\log p(\theta|x)$. Exponentiating equation (4.3) yields the desired Gaussian approximation to the posterior distribution, as

$$p(\theta|x) \propto\!\!\!\approx \exp\left\{-\frac{1}{2}(\theta - \hat{\theta})^T[-H(\hat{\theta})](\theta - \hat{\theta})\right\}$$

And therefore

$$p(\theta|x) \approx \mathcal{N}(\hat{\theta}, -H(\hat{\theta})^{-1}) \tag{4.4}$$

In the next subsection, we will use the result in (4.4) to approximate the variational factors q_i in the occurrence of nonconjugacy.

4.2.3 Mean-field variational updates

The conditional independence relation implied by the model's specification allow us to express the updates implied by equation (4.2) in simpler forms. We begin

with $q(\rho_d)$, the optimality condition for which is

$$
\begin{aligned}
q^*(\rho_d) &\propto \exp\{\mathbb{E}_{-q_{\rho_d}}[\log p(\rho_d|\eta, x_d, z_d)]\} \\
&= \exp\{\mathbb{E}_{-q_{\rho_d}}[\log p(\rho_d|\eta, x_d) \cdot p(z_d|\rho_d)]\} \\
&= \exp\{\mathbb{E}_{-q_{\rho_d}}[\log p(\rho_d|\eta, x_d) + \log p(z_d|\rho_d)]\} \\
&= \exp\left\{\mathbb{E}_{-q_{\rho_d}}\left[\log \mathcal{N}(\rho_d|\eta^T x_d, \mathbb{I}) + \sum_{i=1}^{N_d} \mathrm{Mult}(z_{id}|\pi_d)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\rho_d - \mathbb{E}_{q_\eta}(\eta)^T x_d\right)^T \left(\rho_d - \mathbb{E}_{q_\eta}(\eta)^T x_d\right) + \sum_{i=1}^{N_d} \log \mathbb{E}_{q_z}(\pi_{d,z_{id}})\right\} + \text{constant}
\end{aligned}
$$

Because this update is non-conjugate, we must resort to a Laplace approximation. Letting

$$
f(\rho_d) \equiv -\frac{1}{2}\left(\rho_d - \mathbb{E}_{q_\eta}(\eta)^T x_d\right)^T \left(\rho_d - \mathbb{E}_{q_\eta}(\eta)^T x_d\right) + \sum_{i=1}^{N_d} \log \mathbb{E}_{q_z}(\pi_{d,z_{id}})
$$

and also letting $\hat{\rho}_d$ be a mode of f, following (4.4) we approximate $q^*(\rho_d)$ by

$$
q^*(\rho_d) \approx \mathcal{N}(\hat{\rho}_d, -H(\hat{\rho}_d)^{-1})
$$

In practice, we would find $\hat{\rho}_d$ and an approximation $H(\hat{\rho}_d)^{-1}$ by employing a quasi-Newton optimization method such as the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, or limited-memory BFGS (L-BFGS), starting from a random initialization. If f is a multimodal function, the optimization problem is non-convex, and therefore the solution found will depend on the particular initialization—without guarantees of converging to a global optimum. Back-

ground on quasi-Newton optimization algorithms is provided in appendix A.

The update for $q^*(\theta_d)$ is very similar. It is also non-conjugate, and thus requires an approximation. The optimality condition for this factor is

$$
\begin{aligned}
q^*(\theta_d) &\propto \exp\left\{\mathbb{E}[\log p(\theta_d|y_d, w_d, \alpha, z_d, \Phi)]\right\} \\
&= \exp\left\{\mathbb{E}[\log p(\theta_d|y_d, \alpha) + \log p(w_d|\theta_d, z_d, \Phi)]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\theta_d - \mathbb{E}_{q_\alpha}(\alpha)^T y_d\right)^T \left(\theta_d - \mathbb{E}_{q_\alpha}(\alpha)^T y_d\right) + \sum_{i=1}^{N_d} \log \mathbb{E}_{q_{z,\phi}}(\psi_{d,z_{id}})\right\} + \text{constant}
\end{aligned}
$$

As before, we approximate the optimal $q^*(\theta_d)$ by $q^*(\theta_d) \approx \mathcal{N}(\hat{\theta}_d, -H(\hat{\theta}_d)^{-1})$, where H and $\hat{\theta}_d$ are defined with respect to the objective function within the exponential operator.

Table 4.1: **Relative advantages of MCMC and variational inference**.

	MCMC Inference	**Variational Inference**
Exactness	If convergence is reached, inference is exact	Inference is approximate
Ease of Monitoring Convergence	Poor: No easy ways to monitor convergence	Good: At each iteration we estimate a lower bound on the marginal likelihood
Speed	Generally slower, especially if latent variables cannot be marginalized out of the model	Generally faster
Information about the Posterior	We only obtain samples from the posterior distribution	We obtain a full analytical expression for the approximate posterior distribution

The rest of the variational updates in the model are conjugate. So, for instance, for $q(z_{id})$ we have the following optimality condition, which leads to an update to a discrete categorical distribution:

$$\begin{aligned} q^*(z_{id}) &\propto \exp\left\{\mathbb{E}_{-q_{z_{id}}}[\log p(z_{id}|\rho_d, w_d)]\right\} \\ &= \exp\left\{\mathbb{E}_{-q_{z_{id}}}[\log p(z_{id}|\rho_d) + \log p(w_{id}|z_{id})]\right\} \\ &= \exp\left\{\log \mathbb{E}_{q_\rho}(\pi_d) + \log \mathbb{E}_{q_{\theta,\phi}}(\psi_{d,w_{id}})\right\} \end{aligned}$$

Conversely, the variational update for $q(\phi_k)$ exploits the Dirichlet-multinomial conjugacy, resulting in a Dirichlet density.

5

Experiments and Applications

5.1 Inference on synthetic data

In order to verify the validity of the inference schemes outlined in chapter 4, as well as to provide certain baseline measures of performance, I first performed inference on synthetic data generated precisely according to the C-LDA model. Using a textual analogy, $D = 500$ synthetic documents were generated, each consisting of $N = 50$ words drawn from a vocabulary of size $V = 1000$. Each document was associated with content covariates of dimensionality $G = 2$ and prevalence covariates of dimensionality $F = 2$. The content covariates were each drawn independently at random from the distribution $\mathcal{N}(x|\mu = 0, \sigma^2 = 5)$ while the prevalence covariates were drawn independently at random form the distribution $\mathcal{N}(x|\mu = 0, \sigma^2 = 2)$. Hyperparameters in the generative process were set to $\beta = .1$ and $\sigma_\rho^2 = \sigma_\theta^2 = \gamma_\eta = \gamma_\alpha = a_\eta = b_\eta = a_\alpha = b_\alpha = 1$.

The model's latent variables were then re-initialized at random using the same

hyperparameters, and variational inference was performed to recover the posterior distribution of the topic vectors and covariate coefficients. Convergence tolerance for the log marginal likelihood bound was .1.

Figure 5.1 shows the results of reconstructing the mixing proportions $\{\pi_d\}$ for the first 50 documents in this synthetic corpus at a truncation level of $\tilde{K} = 10$, by computing the expectation of $\{\pi_d\}$ under the posterior variational distribution. Since the topic-label combinations are not uniquely identifiable, for the purposes of comparison we perform global alignment of the mixing proportion vectors based on l_1 pairwise similarity scores, as described in Roberts et al. (2014a). After alignment, we observe qualitatively good reconstruction of the true mixing proportions, as well as a tendency of the inferred distributions to understate the posterior variance of the topic assignments. This is a known property of Variational Bayes inference schema, which is due to the characteristics of the KL divergence used as the variational objective. This is discussed in detail in Bishop (2006).

Figure 5.2 shows the true coefficient vectors in $\{\eta_k\}$ for the same dataset, without distinguishing between the two dimensions of the prevalence covariate coefficients, against their inferred mean values from variational inference. There is a notable positive correlation ($r = .53$) between the true and inferred values, although a number of inferred coefficients are directionally incorrect.

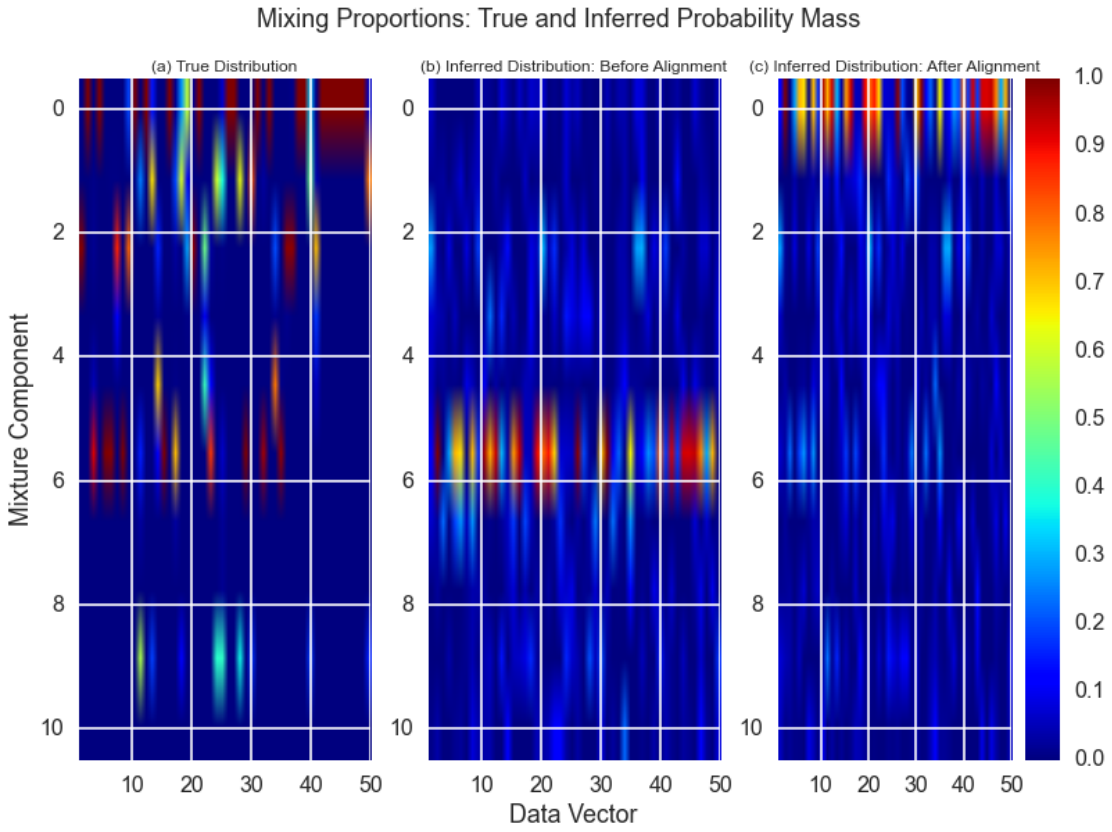

Figure 5.1: **Reconstruction of mixing proportions in synthetic data via variational inference**. (A) Mixing proportions $\{\pi_d\}$ for first fifty data vectors (documents) in the synthetically generated collection at the $\tilde{K} = 10$ truncation level. (B) Expectation of reconstructed distribution from variational inference. Note that the topic-label assignment is unidentifiable, so that the topics are not necessarily aligned with the ones in the leftmost panel. (C) Reconstructed distribution after l_1 global topic alignment (Roberts et al., 2014a): the reconstructed probability mass reflects the qualitative characters of the original distribution, although it understates the posterior variance in the mixing proportions, a common issue in variational inference.

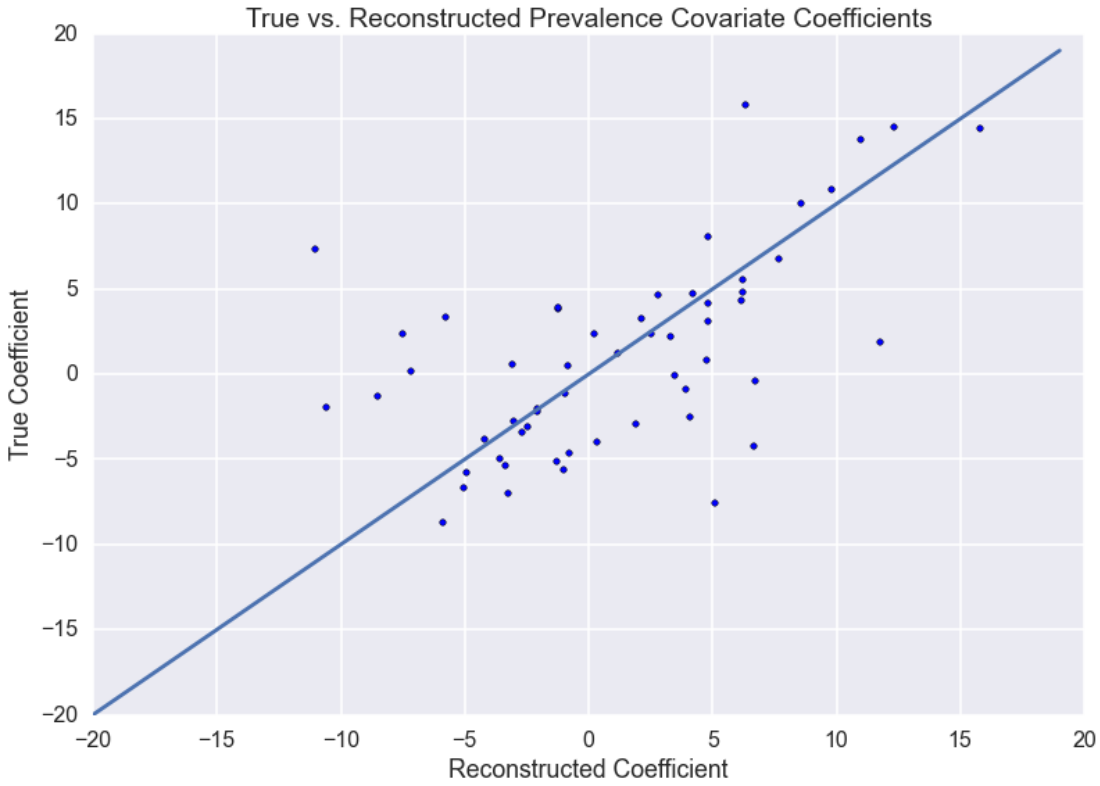

Figure 5.2: **Reconstruction of prevalence covariate coefficients in synthetic data via variational inference**. True and reconstructed prevalence covariate coefficients $\{\eta_k\}$ in the synthetic dataset. The diagonal line shows the identity function. The agreement between the reconstruction and the true coefficients is generally good, as shown by the positive correlation between the true and inferred datapoints, with a few coefficients that are directionally wrong.

Running C-LDA on synthetic data also yields the opportunity to observe the scaling behavior of the inference procedure as a function of the key determinants of the size of the problem—namely, the number of documents D, the size of the vocabulary V, and the truncation level $\tilde{K}$. Figure 5.3 demonstrates the empirically observed scaling behavior of variational inference for C-LDA. Variational inference was performed many times under the same conditions, varying one parameter at

a time. We can observe that the time required for inference scales linearly with D, slightly super-linearly with $\tilde{K}$, and in a more strong super-linear fashion with V. This indicates that the size of the vocabulary V will be the main bottleneck in large-scale implementations of C-LDA.

Figure 5.3: **Empirically observed time scaling behavior of variational inference for C-LDA.** Variational inference was performed on the synthetic data described above several times. The parameters D, V, and $\tilde{K}$ were varied in turn while holding all else constant, and the time for completion of one VI iteration (including gradient evaluation and coordinate ascent) was benchmarked. The charts show the observed scaling behavior in terms of relative time: in each panel, the time value corresponding to the lowest observed setting of the parameter of interest is normalized to 1. Panel A displays the result of a linear regression, while panels B and C display the results of a polynomial regression of degree 2. We observe linear scaling behavior in D, slightly super-linear scaling behavior in $\tilde{K}$, and much more strongly super-linear scaling behavior in V. This indicates that the size of the vocabulary tends to be the bottleneck in large-scale implementations of C-LDA.

5.2 Genomic data: Haplotype phasing

Mixture models have been applied with some success to the problem of haplotype phasing (Xing et al., 2007), which is of great interest in the fields of computational biology and bioinformatics. This section introduces the relevant biological background and presents results from an experiment with data from the international HapMap project, using a variant of the C-LDA model to perform haplotype phasing.

5.2.1 Biological background

Genomic data has become increasingly important in medicine, biology, and the social sciences, as genetic differences often provide valuable insight into disease susceptibility, physiological function, and population heterogeneity. Recent advancements in whole-genome sequencing have enabled the creation of large-scale genomic datasets, and many scientists and statisticians have been focused on the analysis of these datasets. One of the major unresolved challenges in this area is known as the haplotype phasing problem. In short, this problem refers to the inference of haplotypes from genotypes.

A haplotype is an ordered sequence of genetic polymorphisms on a single chromosome, inherited from one parent. A diploid organism, such as a human, has two copies of each chromosome, corresponding to two distinct haplotypes or haplotype mixtures. In any given sample population from a single species, only a small frac-

tion of nucleotides will vary between haplotypes, and these few polymorphisms are known as single-nuleotide polymorphisms, or SNPs. The vast majority of polymorphisms assume only two types within a single-species population, and so they can be represented as a binary indicator. The haplotype is then an ordered vector of these polymorphism indicators. Other polymorphisms include variation in the length of a chromosome, such as the number of repetitions of common DNA sequences, but these are less often implicated in applications.

However, modern methods of genetic sequencing allow for the observation not of an individual's haplotypes, but rather her genotype. The genotype is an ordered sequence of unordered pairs of alleles at each position. For example, a diploid individual with the haplotypes $(0, 0, 0)$ and $(0, 1, 1)$ would have the genotype $(\{0, 0\}, \{0, 1\}, \{0, 1\})$. Although one can easily determine the genotype from the haplotypes, in practice one typically observes the genotype and aims to infer the haplotype. Given the genotype $(\{0, 0\}, \{0, 1\}, \{0, 1\})$, it is not possible to determine whether the individual possesses the haplotype pair $\{(0, 0, 0), (0, 1, 1)\}$ or the pair $\{(0, 0, 1), (0, 1, 0)\}$.

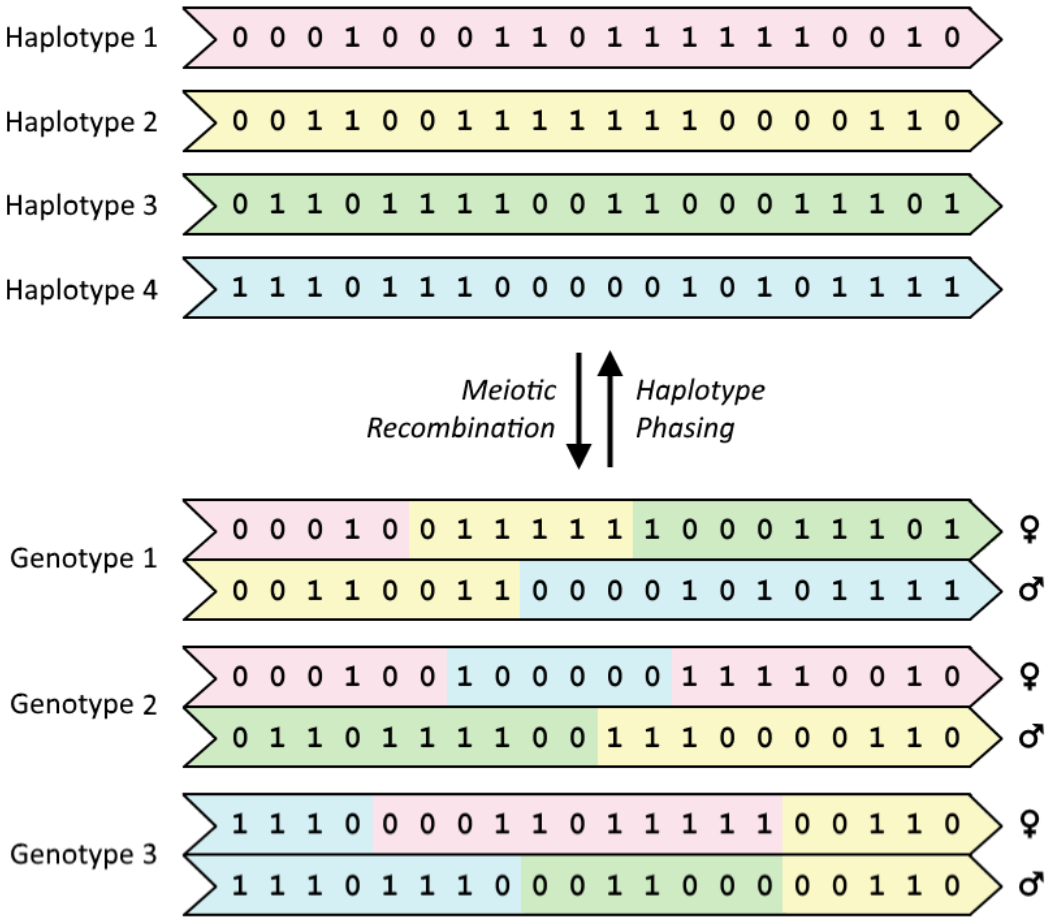

Figure 5.4: **Illustration of the haplotype phasing problem.** Meiotic recombination shuffles single-nucleotide polymorphisms between the parental chromosomes. Genomic sequencing does not allow us to experimentally identify the maternal/paternal lineage of specific alleles, so we perform statistical haplotype phasing to infer the ancestral haplotypes.

5.2.2 A Bayesian approach to haplotype phasing

This problem, fundamental to the analysis of genomic data, has been addressed with varying success using a variety of statistical approaches. These include combinatorial and maximum likelihood approaches. The model parameters can

then be estimated using statistical algorithms including Expectation Maximization (EM), MCMC, and hidden Markov models (HMM).

Here I present a novel approach to the haplotype phasing problem based on the paradigm discussed in this paper, which makes use of a modified variant of the C-LDA model. In a Bayesian fashion, we represent haplotypes as *distributions* over SNP realizations. This is in contrast with the assumption that haplotypes are rather fixed allele sequences—requiring separate modeling of a mutation mechanism that will generate the observed genotypic variance. I argue that it is more natural, from a Bayesian standpoint, to let the haplotypes themselves be probability distributions. This has the added benefit of allowing for a seamless incorporation of basic mutation phenomena into the model.

We consider a population of D individuals with genotypic information available for N chromosomal loci. In order to represent the genotypic data, we construct tokens of the form j_i, indicating the presence of allele i at locus j. The resulting vocabulary $\mathcal{V}$ is therefore of size $2N$, and can be enumerated as follows:

$$\mathcal{V} = \{1_0, 1_1, 2_0, 2_1, \ldots, N_0, N_1\}$$

Each individual genotype will be represented by a vector of $2N$ tokens in V. The tokens for the genotype of individual d are referred to as $\{g_{id}\}_{i=1}^{2N}$. Homozygous loci will contribute two identical tokens to the data vector, while heterozygous loci will contribute two different tokens. We assume that there exist latent ancestral haplotype patterns $\{\phi_k\}_{k=1}^{\infty}$, where a haplotype is a distribution over $\mathcal{V}$. These

haplotype patterns can then be best interpreted by considering the ratios of the probabilities assigned to the two alleles at each locus. We place a dependent Dirichlet process prior on the ancenstral haplotype patterns. This is desirable, since we would like our prior to impose some degree of parsimony on the model (i.e., we wish to limit the use of unnecessary haplotypes, which we can do by tuning the base concentration parameter β), and we would also expect the number of latent haplotypes in a population to be monotonically increasing with the size of the population itself.

Given prevalence covariates x_d (such as ethnic lineage and gender) associated with each individual genotype, the individual-haplotype distributions π_d are constructed precisely as in the standard C-LDA model presented in chapter 3, via a logistic stick-breaking process. On the other hand, we would not expect prevalence covariates to have biological significance, and as such we discard them. Given the individual-haplotype distributions π_d, we have several choices to make in order to complete the description of the genotypic generative process and assign a likelihood to the data. First, we must choose whether to allow the genotype to be drawn from a weighted mixture including all the latent haplotype patterns, or to let the genotype only be drawn from a mixture of two haplotype patterns (maternal and paternal). The first choice corresponds to an assumption that the maternal and paternal haplotypes are *pure* copies of the ancestral patterns, and therefore that no recombination of the ancestral haplotype patterns occurred through generations. The second choice conversely reflects the assumption that recombination of genotypes may have occurred, and that the maternal and paternal copies are

only impure copies of the ancestral patterns. We choose to proceed with the latter scenario, with the understanding that the model could have been altered to reflect the assumption of pure maternal and paternal haplotypes.

We also have to choose whether we wish to explicitly assign each token g_{id} to one of the latent haplotypes according to the distributions ϕ_k, or whether to draw them from a mixture distribution, meaning that each token is drawn as

$$g_{id}|\pi_d, \{\phi_k\} \overset{iid}{\sim} \sum_{i=1}^{k} \pi_{d,k} \cdot \phi_k \tag{5.1}$$

Note that this only amounts to marginalizing out the latent assignments z_{id} of a token g_{id} to a unique haplotype pattern k, and that the model likelihood is identical in these two cases. By marginalizing out these variables, we in fact would expect increased performance of MCMC samplers or variational optimization, at the expense of not being able to sample from the distribution of the latent assignments. In the context of the haplotype phasing problem, we have no interest in the assignments of individual genotype tokens to latent haplotypes, as these carry no information of biological relevance. Correspondingly, we choose to collapse the model and complete the likelihood description as detailed in equation 5.1.

Figure 5.5 shows a representation of this model for haplotype phasing as a directed graphical model. Note that this model corresponds to a special case of C-LDA, where the word-topic assignments z_{id} are marginalized out, and content covariate information is disregarded. As such, inference can be performed using the same results as in chapter 3.

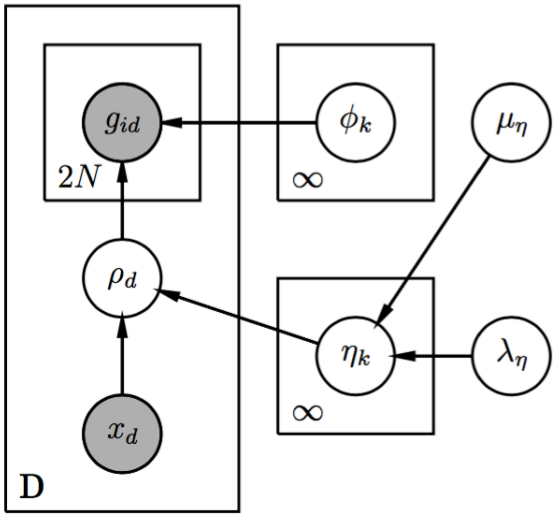

Figure 5.5: **A modified version of C-LDA with applications to haplotype phasing**. Directed graphical model representation of a special case of C-LDA for haplotype phasing. Content covariate information is disregarded and word-topic assignments z_{id} are marginalized out, since we do not expect either of these model components to have biological significance in the context of the problem.

5.2.3 Results

To test the variant of C-LDA presented above, I apply it to the phasing of genotypes from 97 individuals belonging to 11 distinct populations (see table 5.1). All the genotypic data comes from the international HapMap project (Gibbs et al., 2003), whose goal was to develop a full haplotype map of the human genome to study human genetic variation across populations. I focus particularly on genotypic variation in 111 polymorphic loci on chromosome 21. The categorical covariate of interest is population.

Figure 5.5 shows the results of haplotype phasing of the HapMap data using C-LDA alongside the PHASE algorithm of Stephens et al. (2001), one of the most

commonly used Bayesian models for haplotype phasing. I also compare the performance of C-LDA to a later variant of the PHASE algorithm introduced in Stephens & Donnelly (2003), which explicitly models the process of genetic recombination. Since the HapMap data does not contain experimentally verified haplotypes, it is not possible to evaluate the models on the basis of their reconstruction error. Instead, I perform inference in the two variants of the PHASE model using Automatic Differentiation Variational Inference (ADVI) with Stan, a probabilistic programming system (Kucukelbir et al., 2016). This allows us to compare the models by the Evidence Lower Bound (ELBO) on the log marginal likelihood that they achieve on the HapMap data. The results of these experiments are shown in figure 5.6. C-LDA does not attain a marginal likelihood lower bound as high as that attained by the two variants of PHASE, but its performance is overall comparable—especially considering the fact that it is a more generic model than PHASE.

Population code	Description
ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italy
YRI	Yoruba in Ibadan, Nigeria

Table 5.1: **List of populations in HapMap genotype data**. With corresponding population codes (Gibbs et al., 2003).

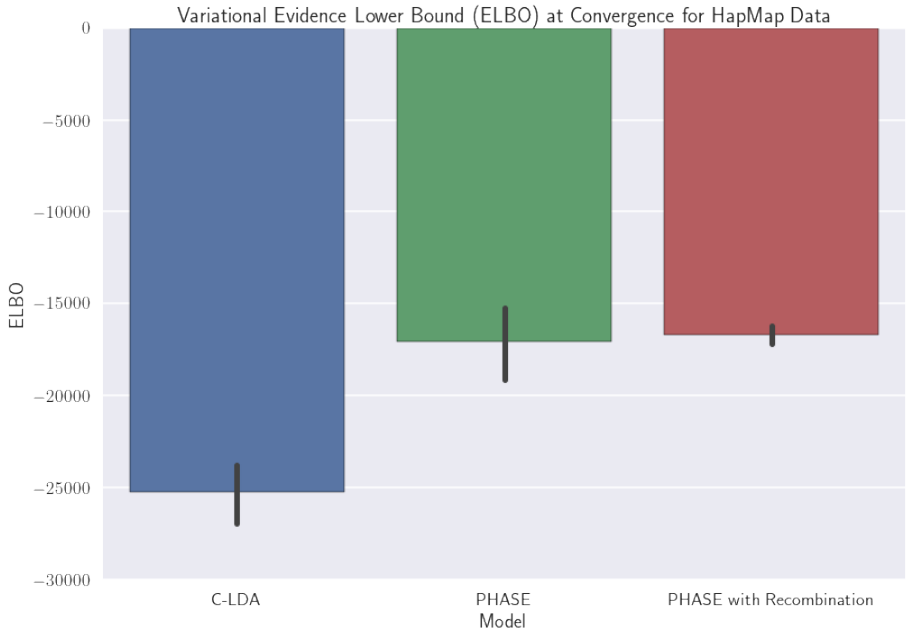

Figure 5.6: **Comparing the performance of C-LDA and other Bayesian models for haplotype phasing**. The figure shows the mean Evidence Lower Bound (ELBO) on the log marginal likelihood of the model achieved in the course of ten runs of variational inference, starting from different random initizations. The models compared are C-LDA, PHASE (Stephens et al., 2001), and PHASE with recombination (Stephens & Donnelly, 2003). The performance of C-LDA is not strictly as good as that of the two PHASE variants, but overall comparable, given its general-purpose rather than ad-hoc nature.

5.3 Textual data: New York Times corpus

As discussed in the introduction, mixture models have been very successfully applied to textual data. Applications in the textual domain often go under the name of topic modeling, as the clustering solutions found via Bayesian inference can facilitate the tasks of document classification and information retrieval, and in the best-case scenario provide insight into the thematic structure of the corpus. Moreover, nonparametric mixture models are particularly well-suited to textual applications, in which the assumption that the realized number of clusters will grow with the size of the corpus (given an infinite numbers of underlying clusters) is especially plausible and desirable. In order to highlight these properties of the class of models discussed in this thesis, here I present an application of C-LDA to the modeling of a large corpus of opinion editorials from New York Times (NYT), a widely circulated American daily newspaper.

5.3.1 Corpus Background

The NYT Annotated Corpus, compiled and distributed by the Linguistic Data Consortium (Sandhaus, 2008), contains approximately 1.8 million articles published by the NYT between the years 1987 and 2008. The articles are accompanied by metadata, including date of publication, originating desk, as well as print page, column, and section. The articles are also manually tagged to highlight people, organizations, and locations. For the purposes of this application, I consider a random sample of approximately 10% of the opinion articles in the

dataset, amounting to 13357 documents.

The corpus is preprocessed by first removing all punctuation and other extraneous elements such as HTML tags, as well as converting all the remaining words to lowercase. Inflected words are then stemmed in order to reduce them to their base roots and simplify the vocabulary (Xu & Croft, 1998). Commonly occurring 'stop words'* in the English language such as articles and prepositions are removed. Words that are either extremely frequent or extremely infrequent in the corpus are also removed, since they are less likely to carry useful information for distinguishing thematic elements that run through the corpus. Letting V be the size of the resulting vocabulary (this is the number of distinct words in the corpus), each document is finally represented as vector in $\mathbb{R}^V$ using the bag-of-words assumption: that is, by simply conserving information about word counts in the document.

5.3.2 Results

To highlight the applicability of models like C-LDA to social science research, I focus on the question of how date of publication affects the prevalence of themes in the NYT opinion corpus. Since our focus will be on exploring the information conveyed by the estimated prevalence coefficients $\{\eta_k\}$, and we would like to learn a model that is more flexible than linear, we will need to introduce nonlinear basis

*In the search and natural language processing literature, the term *stop word* is used to refer to a common word that carries little information about the semantic content of a particular document. These are words such as *the, which, after, a,* and so on.

transformations of the covariate of interest.

The date t_d associated with document d is first mapped to a real-valued variable by means of a linear transformation,[†] and then a nonlinear basis expansion is introduced. While generating polynomial transformations of t_d might be an obvious choice, this option can be problematic because it will tend to induce bias in our estimates by providing systematically larger scaling for dates with higher real-valued images. Instead, we transform t_d by means of Gaussian radial basis functions (RBFs), which obviate this issue. Given a constant c, a Gaussian radial basis function is defined as

$$\phi(x; c) = e^{-(x-c)^2}$$

Note that this function will be maximum at the center point c and decay symmetrically on both sides of it. As such, an RBF expansion will also be particularly apt at capturing peak coverage behavior. To expand t_d using RBFs, we first choose Q equally spaced points $\{c_1, c_2, \ldots, c_Q\}$ that cover the entire date range under consideration, and then we generate the prevalence covariate vector as follows:

$$x_d = [1,\ \phi(t_d; c_1),\ \phi(t_d; c_2),\ \ldots,\ \phi(t_d; c_Q)]$$

The choice of Q degrees of freedom plus an intercept for the RBF expansion is subject to a tradeoff between model complexity and expressivity. A higher number of degrees of freedom will correspond to models with more flexible nonlinearities that

[†]The particular linear transformation used is irrelevant, since the values are always mapped back to date space after model estimation. In this particular application, I mapped dates to the number of seconds following January 1$^{\text{st}}$, 1970, at midnight UCT.

may however overfit the training data. For this application, we set $Q = 10$, which provides good expressivity without introducing unnecessary complexity.

We fit the C-LDA model using $\{x_d\}$ as the prevalence covariate, and a constant matrix as the content covariate. The choice of hyperparameters was as in subsection 5.1. Once the model is estimated, we can construct nonlinear summary prevalence functions $\xi_k(t)$ from the coefficients $\{\eta_k\}$ that will contain information about how the prevalence of each topic k in the corpus varies as a function of the date of publication. Following the C-LDA model specification, the function $\xi_k(t)$ is defined as follows:

$$\begin{aligned}\xi_k(t) &= \eta_k^T \cdot [1,\, \phi(t; c_1),\, \phi(t; c_2),\, \ldots,\, \phi(t; c_Q)] \\ &= \eta_{k,1} + \eta_{k,2} \cdot \phi(t; c_1) + \cdots + \eta_{k,Q+1} \cdot \phi(t; c_Q)\end{aligned}$$

The expected value of $\xi_k(t)$ can be found by simply computing sample averages of the coefficients $\{\eta_k\}$ using either samples obtained via MCMC, or samples from the variational posterior.

In order to allow for better interpretability, the topics are not labeled by choosing the stemmed vocabulary items that are simply most probable in that topic: instead, we use the Frequency-Exclusivity (FREX) method described in Roberts et al. (2014b), which scores each word according to the harmonic mean of its frequency within a topic and a measure of its specificity to the topic. This is done to reflect the intuition that words should be both prevalent and specific to a topic in order for them to convery greater semantic information about the topic.

Figure 5.7 shows the posterior distribution over the number of instantiated topic obtained from the variational approximation, while figure 5.8 shows six selected topics obtained from the corpus by their highest-FREX labels, along with the dynamics of topical prevalence as a function of the date of publication. To illustrate the dynamics of a topic k, we show both the raw topical prevalence π_{dk} against t_d for each document in the corpus, and the expectation of the prevalence summary function $\xi_k(t)$, linearly scaled to be a proportion for easier comparison.

Notice that the prevalence summary functions capture some of the most salient aspect of the topical prevalence dynamics in the corpus. In figure 5.8, starting from the left top corner and going counterclockwise, we observe topics that deal with the USSR and post-Soviet Russia; foreign relations with Iraq; nuclear policy and proliferation; US voting and elections; the political career of Bill and Hillary Clinton; and the Israeli-Palestinian conflict. The prevalence summary functions spike, for instance, at the time of the Lewinsky scandal for the Clinton topic, or at the time of the 2003 invasion of Iraq for the Iraq-related topic. Furthermore, the summary function for the topic related to voting and elections spikes cyclically every four years, in correspondence to American general elections—and shows a marked peak in 2000, at the time of the Florida ballot controversy during the Bush-Gore election. These results highlight the applicability of models such as C-LDA to research in the social sciences, where the quantification of raw textual data can be an effective tool in the hands of researchers.

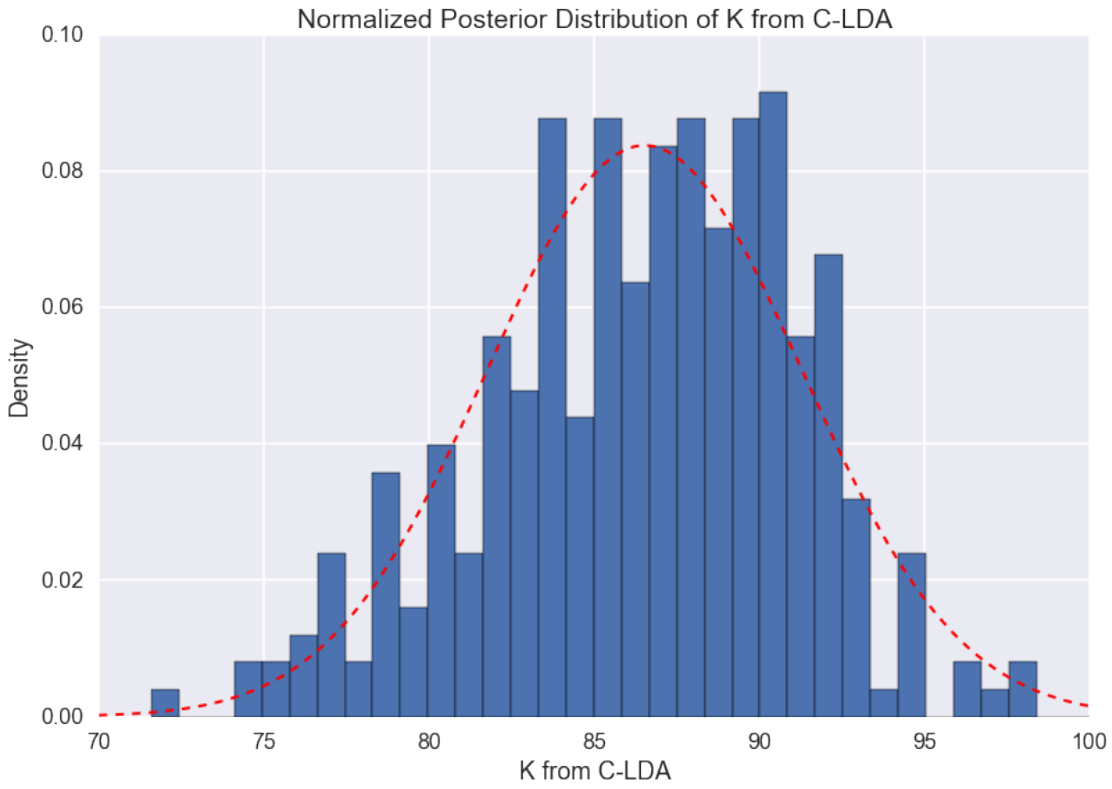

Figure 5.7: **Posterior distribution of K (activated topics) in the New York Times opinion corpus**. Sample mean is 86.4; sample median is 87.0. Gaussian density estimate in red.

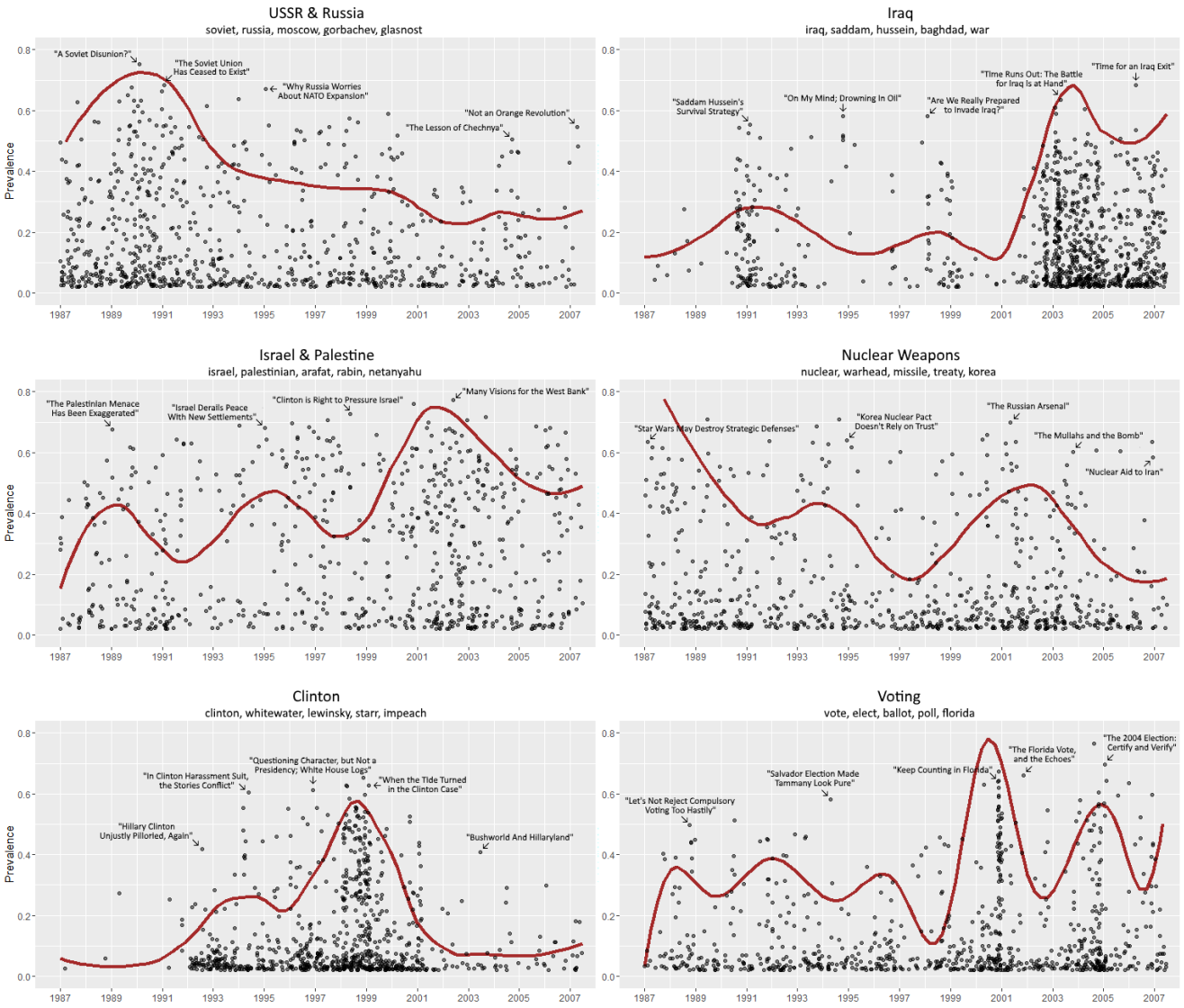

Figure 5.8: **Topical prevalence dynamics in the New York Times opinion corpus**. The dynamics of topical prevalence are plotted as a function of date of publication for six selected topics in the NYT opinion corpus. Model estimation was performed using variational inference. The topic labels are the words with the highest FREX scores in each topic, and are accompanied by a manual title summarizing the semantic content of the topic. For a given topic k and document d, the scatter dots show the prevalence score π_{dk} as a function of the date t_d. The red lines show the linearly scaled means of the summary prevalence functions $\xi_k(t)$. The annotations show selected headlines corresponding to some of the points in the graphs, highlighting some of the salient events that correspond to peaks in the summary prevalence functions.

6

Conclusion

In this thesis, I discussed a general modeling framework for integrating covariate information into the generative process for nonparametric mixture models, which relies on the use of dependent Dirichlet process priors. Covariate-dependent nonparametric mixture models are useful in a range of applicatons, and are apt at capturing structural properties of data that are hypothesized to depend on exogenous variables, while retaining the flexibility to grow the complexity of the model with the amount of training data. I introduced Covariate-Dependent Nonparametric LDA (C-LDA), a model in this class that draws upon the topic modeling literature and provides a very flexible and general way to study the dependencies between data structure and covariates of interest.

I developed both sampling-based Markov-Chain Monte Carlo (MCMC) and posterior variational inference procedures for estimating the model from data. While the MCMC sampler can draw samples from the exact posterior distribution of

the model parameters, variational inference tends to be quicker and more computationally tractable in practical applications, and provides better convergence guarantees. After validating the inferential procedures by recovering parameters from synthetically generated data, I turned to two practical applications intended to demonstrate the wide applicability of the model.

Using genome sequence data with associated ethnic lineage information from the international HapMap project, I showed how nonparametric mixture models can be used to tackle the task of haplotype phasing. This is the problem of reconstructing haplotypes—ordered sequences of genetic polymorphisms on a single chromosome, inherited from one parent—from raw genotypic information. Accuracy in haplotype phasing is important for a range of downstream applications in population genetics, which include disease association studies. I showed that C-LDA achieved comparable performance to other Bayesian models for haplotype phasing in terms of marginal model likelihood.

Turning to a different domain of application, I then used C-LDA to perform topic modeling of a corpus of opinion pieces from the New York Times spanning several decades of publication. I focused particularly on the problem of studying the temporal dynamics of topical prevalence in the corpus, demonstrating the use of covariate-dependent nonparametric mixture models in research that makes use of unstructures textual data. Overall, the statistical approach described in this thesis is a powerful means of describing and quantifying the relationships between the latent structure of data collections and exogenous variables of research interests,

with wide applicability in both the social and the natural sciences.

This thesis lays the framework for future work that will build upon it. A first future direction of work is the development of software for parallel inference in models of this class. Online variational inference is a particularly promising technique for achieving this goal, and some effort in this direction is underway. A second direction of research involves the development of innovative causal inference tools that exploit the newfound understanding of the latent structure of data. An example of such work is that by Roberts et al. (2015b), who show how the STM model can be successfully applied for matching observations in high-dimensional contexts. A continued understanding of principled avenues to perform causal inference using unsupervised learning methods holds much promise for expanding the amount and kinds of data that scientists can explore. Ultimately, the aim of this larger-scale project is to develop robust, scalable tools for the use of complex or unstructured data in research.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.

Blei, D. M. & Lafferty, J. D. (2006a). Correlated topic models. *Advances in Neural Information Processing Systems*, 18, 147.

Blei, D. M. & Lafferty, J. D. (2006b). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning research*, 3, 993–1022.

Coppola, A. & Stewart, B. M. (2014). lbfgs: Efficient L-BFGS and OWL-QN Optimization in R. *CRAN*.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. (2003). The international HapMap project. *Nature*, 426(6968), 789–796.

Gross, J. & Manrique-Vallier, D. (2014). Handbook of mixed membership models and their applications. *Chapman & Hall*.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453).

Kim, D. I. & Sudderth, E. B. (2011). The doubly correlated nonparametric topic model. In *Advances in Neural Information Processing Systems* (pp. 1980–1988).

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2016). Automatic differentiation variational inference. *Journal of Machine Learning Research*.

Li, W. & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine learning* (pp. 577–584).

Liu, D. C. & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1-3), 503–528.

MacEachern, S. N. (2000). Dependent Dirichlet processes. *Unpublished manuscript*.

Müller, P. & Rodriguez, A. (2013). *Nonparametric Bayesian Inference*. Institute of Mathematical Statistics, American Statistical Assocation.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

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

Paisley, J. (2010). A simple proof of the stick-breaking construction of the Dirichlet process. *Technical report*.

Ren, L., Du, L., Carin, L., & Dunson, D. (2011). Logistic stick-breaking process. *The Journal of Machine Learning Research*, 12, 203–239.

Roberts, M., Stewart, B., & Tingley, D. (2014a). Navigating the local modes of big data: The case of topic models. *Unpublished manuscript*.

Roberts, M. E., Stewart, B. M., & Airoldi, E. (2015a). A model of text for experimentation in the social sciences. *Unpublished manuscript*.

Roberts, M. E., Stewart, B. M., & Nielsen, R. (2015b). Matching methods for high-dimensional data with applications to text. *Unpublished manuscript*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014b). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.

Sandhaus, E. (2008). The New York Times annotated corpus. *Linguistic Data Consortium, Philadelphia*.

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

Stephens, M. & Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5), 1162–1169.

Stephens, M., Smith, N. J., & Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68(4), 978–989.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476).

Wang, C. & Blei, D. M. (2013). Variational inference in nonconjugate models. *The Journal of Machine Learning Research*, 14(1), 1005–1031.

Xing, E. P., Jordan, M. I., & Sharan, R. (2007). Bayesian haplotype inference via the Dirichlet process. *Journal of Computational Biology*, 14(3), 267–284.

Xu, J. & Croft, W. B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM Transactions on Information Systems (TOIS)*, 16(1), 61–81.

Appendices

A

Quasi-Newton Optimization

In chapter 4 I introduced the Laplace approximation for variational inference in nonconjugate models, which requires finding the mode and inverse Hessian of a potentially nonlinear function. This appendix discusses the class of quasi-Newton that are used in practice to solve the optimization problem and estimate the inverse Hessian of a generic function. I released the lbfgs R package (Coppola & Stewart, 2014) implementing the methods described in this section, which is available on the Comprehensive R Archive Network (CRAN).

The notation is as follows. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be an objective function to be minimized. We let the $||\cdot||$ operator denote the L_2 norm of a vector, and $||\cdot||_1$ denote the L_1 norm. $H(x_k)$ is the Hessian matrix of f at x_k, and $g(x_k)$ if the gradient of f at the same point. Quasi-Newton optimization methods solve the minimization problem by computing approximations to the Hessian matrix of the objective function. At each iteration, quasi-Newton algorithms locally model f at

the point x_k using a quadratic approximation:

$$Q(x) = f(x_k) + (x - x_k)^T g(x_k) + \frac{1}{2}(x - x_k)^T H(x_k)(x - x_k)$$

A search direction can then be found by computing the vector x^* that minimizes $Q(x)$. Assuming that Hessian is positive-definite, this is

$$x^* = x_k - H(x_k)^{-1} g(x_k)$$

The next search point is then found along the ray defined by $x_k - \alpha H(x_k)^{-1} g(x_k)$. The procedure is iterated until the gradient is zero, with some degree of convergence tolerance.

The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm (Liu & Nocedal, 1989) is a quasi-Newton method that is optimized to reduce memory usage, which is useful in settings where the dimensionality of f is very high, and it is thus expensive to store the gradient and Hessian of f. The L-BFGS algorithm avoids storing sequential approximations of the Hessian matrix. Instead, L-BFGS stores curvature information from the last m iterations of the algorithm, and uses them to find the new search direction. More specifically, the algorithm stores information about the spatial displacement and the change in gradient, and uses them to estimate a search direction without storing or computing the Hessian explicitly.

B

Proof of Validity of Stick-Breaking

In this appendix I prove the validity of the stick-breaking construction in yielding draws from a Dirichlet process, following the proof by Sethuraman (1994) and Paisley (2010). In order to do this, I first establish some notation. As discussed in chapter 2, the Dirichlet distribution of dimension K is a distribution over the simplex in R^K, which we denote by Δ_K. The simplex is defined as

$$\Delta_K = \left\{ (x_1, x_2, \ldots, x_K) \,\middle|\, \forall i : \; 0 \leq x_i \leq 1, \; \sum_{i=1}^{K} x_i = 1 \right\}$$

We can parameterize the Dirichlet distribution by a base vector $g_0 \in \Delta_K$ and a scalar concentration parameter $\alpha > 0$. Then the density function for a vector $\pi \sim \text{Dirichlet}(\alpha\, g_0)$ is

$$p(\pi|\alpha, g_0) = \frac{\Gamma(\alpha)}{\prod_{k=1}^{K} \Gamma(\alpha\, g_{0,k})} \cdot \prod_{k=1}^{K} \pi_k^{\alpha\, g_{0,k}-1}$$

We use the notation δ_k to denote a vector whose all entries are zero, except for the entry at position k, which is instead 1. Similarly, the notation $\delta_{\theta_k}(\theta)$ denotes a distribution whose value is 1 when $\theta = \theta_k$, and otherwise vanishes. Such vectors and distributions are called *Kronecker deltas*, and their dimensionality will be implied from context.

Now that notation is established, I introduce two lemmas relating to the properties of the Dirichlet distribution.

Lemma 1 Consider the random variable $Z \sim \sum_{k=1}^{K} g_{0,k} \cdot \mathrm{Dir}(\alpha\, g_0 + \delta_k)$. The distribution of it is equivalently $Z \sim \mathrm{Dir}(\alpha\, g_0)$.

Proof Notice that we can sample Z according to this distribution by first drawing an intermediate variable $Y \sim \mathrm{Mult}(g_0)$ and then sampling $Z \sim \mathrm{Dir}(\alpha\, g_0 + \delta_k)$. We let $\pi \sim \mathrm{Dir}(\alpha\, g_0)$. Then

$$P(Y = k|\alpha\, g_0) = \int_{\pi \in \Delta_K} P(Y = k|\pi)\, p(\pi|\alpha\, g_0)\, d\pi = E[\pi_k|\alpha\, g_0] = g_{0,k}$$

$$p(\pi|\alpha\, g_0) = \sum_{k=1}^{K} P(Y = k|\alpha\, g_0)\, p(\pi|Y = k, \alpha\, g_0) = \sum_{k=1}^{K} g_{0,k} \cdot \mathrm{Dir}(\alpha\, g_0 + \delta_k) \qquad \blacksquare$$

Lemma 2 Consider the random vectors $W_1 \sim \mathrm{Dir}(w_1, \ldots, w_K)$, $W_2 \sim \mathrm{Dir}(v_1, \ldots, v_K)$, and $V \sim \mathrm{Beta}(\sum_{k=1}^{K} w_k,\, \sum_{k=1}^{K} v_k)$. Define the linear combination

$$Z = VW_1 + (1 - V)W_2$$

Then $Z \sim \text{Dir}(w_1 + v_1, \ldots, w_K + v_K)$.

Proof If $\gamma_k \sim \text{Gamma}(\alpha\, g_{0,k}, \lambda)$ for $k = 1, \ldots, K$ and $\pi = (\gamma_1, \ldots, \gamma_k)/\sum_k \gamma_k$, then $\pi \sim \text{Dir}(\alpha\, g_0)$. Let $\gamma_k \sim \text{Gamma}(w_k, \lambda)$ and $\gamma'_k \sim \text{Gamma}(v_k, \lambda)$, and define

$$
\begin{aligned}
W_1 &= \left(\sum_k \gamma_k \right)^{-1} (\gamma_1, \ldots, \gamma_K) \\
W_2 &= \left(\sum_k \gamma'_k \right)^{-1} (\gamma'_1, \ldots, \gamma'_K) \\
V &= \left(\sum_k \gamma_k + \sum_k \gamma'_k \right)^{-1} \left(\sum_k \gamma_k \right)
\end{aligned}
$$

Then it follows that

$$
\begin{aligned}
W_1 &\sim \text{Dir}(w_1, \ldots, w_k) \\
W_2 &\sim \text{Dir}(v_1, \ldots, v_k) \\
V &\sim \text{Beta}(\textstyle\sum_k w_k, \sum_k v_k)
\end{aligned}
$$

Where the distribution of V results from the fact that $\sum_k \gamma_k \sim \text{Gamma}(\sum_k w_k, \lambda)$, and V is independent of W_1 and W_2. The multiplication $Z = VW_1 + (1 - V)W_2$ yields the representation of $Z \sim \text{Dir}(w_1 + v_1, ..., w_K + v_K)$ as a Gamma-disributed random variable. ∎

Now I use these lemmas to prove our claim of interest—namely, the validity of the stick-breaking construction in valid draws from a Dirichlet Process.

Claim The stick-breaking constructive definition of a Dirichlet process states

that, if G is constructed as follows, then $G \sim \text{Dir}(\alpha\, g_0)$:

$$
\begin{aligned}
G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \\
\pi_k &= \beta_k \prod_{j=1}^{k-1} (1 - \beta_j) \\
\beta_k &\overset{iid}{\sim} \text{Beta}(1, \alpha) \\
\theta &\sim \text{Mult}(g_0)
\end{aligned}
$$

The stick-breaking weights satisfy $\pi_k \in [0, 1]$ for all $k \geq 1$ and $\sum_{k=1}^{\infty} \pi_k = 1$.

Proof Applying lemmas 1 and 2 to $\pi \sim \text{Dir}(\alpha\, g_0 + \delta_\theta)$ allows us to represent the vector by the process

$$
\begin{aligned}
\pi &= VW + (1 - V)\pi' \\
W &\sim \text{Dir}(\delta_\theta) \\
\pi' &\sim \text{Dir}(\alpha\, g_0) \\
V &\sim \text{Beta}(\textstyle\sum_{k=1}^{K} \delta_{\theta_k}, \sum_{k=1}^{K} \alpha\, g_{0,k}) \\
\theta &\sim \text{Mult}(g_0)
\end{aligned}
$$

The resulting random vector π still follows the distribution $\pi \sim \text{Dir}(\alpha\, g_0)$, and we have that $\sum_{k=1}^{K} \delta_{\theta_k} = 1$ and $\sum_{k=1}^{K} \alpha\, g_{0,k} = \alpha$. Yet now we can observe that $P(W = \delta_{\theta_k} | g_0 = \delta_{\theta_k}) = 1$, since only one of the K variables parameterizing the Dirichlet distribution of W is nonzero (in this sense, we say that W is a degenerate random variable). This implies that we can simplify the process by

which we construct π and still achieve an equivalent distribution:

$$\begin{aligned}
\pi &= V\delta_\theta + (1-V)\pi' \\
\pi' &\sim \text{Dir}(\alpha\, g_0) \\
V &\sim \text{Beta}(1, \alpha) \\
\theta &\sim \text{Mult}(g_0)
\end{aligned}$$

Hence we now have that $\pi \overset{d}{=} \pi'$, since both of these random vectors follow the distribution Dirichlet($\alpha\, g_0$). This implies that π' can be decomposed in the exact same way as π. Therefore, for $i = 1, 2$, we have

$$\begin{aligned}
\pi &= V_1\delta_{\theta_1} + V_2(1-V_1)\delta_{\theta_2} + (1-V_1)(1-V_2)\pi'' \\
V_i &\overset{iid}{\sim} \text{Beta}(1, \alpha) \\
\theta &\sim \text{Mult}(g_0) \\
\pi'' &\sim \text{Dir}(\alpha\, g_0)
\end{aligned}$$

Now $\pi \overset{d}{=} \pi' \overset{d}{=} \pi''$. This decomposition process can then proceed following an infinite recursion. For any value i, as well as in the limit $i \to \infty$, the decomposition produces the vector $\pi \sim \text{Dir}(\alpha\, g_0)$. In the limit $i \to \infty$, this process approaches the one described in the original claim, since $\lim_{i\to\infty} \prod_{j=1}^{i}(1 - V_j) = 0$. This concludes the proof. ∎

C

Notation Table

$\mathbb{R}$	The set of real numbers
Θ	A probability space
$\{\cdot\}$	A set
$\stackrel{d}{=}$	Equal in distribution to
$\sim$	Distributed as
$\propto$	Proportional to
$\stackrel{\propto}{\sim}$	Approximately proportional to
$\approx$	Approximately equal to
$\equiv$	Equivalent to
Dir	Dirichlet distribution
Mult	Multinomial distribution
$\mathcal{N}$	Gaussian distribution
Γ	Gamma distribution or Gamma function (depending on context)
Beta	Beta distribution
$\mathbb{E}_q$	Expectation over the distribution q