11

Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference

John Paisley

Department of Electrical Engineering, Columbia University, New York, NY 10027, USA

David M. Blei

Computer Science Department, Princeton University, Princeton, NJ 08544, USA

Michael I. Jordan

Department of EECS, University of California, Berkeley, CA 94720, USA

CONTENTS

We present stochastic variational inference algorithms for two Bayesian nonnegative matrix factorization (NMF) models. These algorithms allow for fast processing of massive datasets. In particular, we derive stochastic algorithms for a Bayesian extension of the NMF algorithm of Lee and Seung (2001), and a matrix factorization model called correlated NMF, which is motivated by the correlated topic model (Blei and Lafferty, 2007). We apply our algorithms to roughly 1.8 million documents from the *New York Times*, comparing with online LDA (Hoffman et al., 2010b).

11.1 Introduction

In the era of "big data," a significant challenge for machine learning research lies in developing efficient algorithms for processing massive datasets (Jordan, 2011). In several modern data-modeling environments, algorithms for mixed membership and other hierarchical Bayesian models no longer have the luxury of waiting for Markov chain Monte Carlo (MCMC) samplers to perform the tens of thousands of iterations necessary to approximately sample from the posterior, especially when

per-iteration runtime is long in the presence of much data. Instead, stochastic optimization methods provide another non-Bayesian learning framework that is better suited to big data environments (Bottou, 1998).

This may seem unfortunate for Bayesian methods in machine learning, however, recent advances have combined stochastic optimization with hierarchical Bayesian modeling (Sato, 2001; Hoffman et al., 2010b; Wang et al., 2011) allowing for approximate posterior inference for "big data." Called *stochastic variational Bayes*, this method performs stochastic optimization on the objective function used in mean field variational Bayesian (VB) inference (Jordan et al., 1999; Sato, 2001). Like maximum likelihood (ML) and maximum a posteriori (MAP) inference methods, VB inference learns a point estimate that locally maximizes its objective function. But unlike ML and maximum MAP, which learn point estimates of a model's parameters, VB learns a point estimate on a set of probability distributions on these parameters.

Since maximizing the variational objective function minimizes the Kullback-Leibler divergence between the approximate posterior distribution and the true posterior (Jordan et al., 1999), variational Bayes is an approximate Bayesian inference method. Because it is an optimization algorithm, it can leverage stochastic optimization techniques (Sato, 2001). This has recently proven useful in mixed membership topic modeling (Hoffman et al., 2010b; Wang et al., 2011), where the number of documents constituting the data can be in the millions. However, the stochastic variational technique is a general method that can address big data issues for other model families as well.

In this paper, we develop stochastic variational inference algorithms for two nonnegative matrix factorization models, which we apply to text modeling. Integrating out the latent indicators of a probabilistic topic model results in a nonnegative matrix factorization problem, and thus the relationship to mixed membership models is clear. The first model we consider is a Bayesian extension of the well-known NMF algorithm of Lee and Seung (2001) with a KL penalty that has an equivalent maximum likelihood representation. This extension was proposed by Cemgil (2009), who derived a variational inference algorithm. We present a stochastic inference algorithm for this model, which significantly increases the amount of data that can be processed in a given period of time.

The second model we consider is motivated by the correlated topic model (CTM) of Blei and Lafferty (2007). We first present a new representation of the CTM that represents topics and documents as having latent locations in $\mathbb{R}^m$. In this formulation, the probability of any topic is a function of the dot-product between the document and topic locations, which introduces correlations among the topic probabilities. The latent locations of the documents have additional uses, which we show with a document retrieval example. We carry this idea into the nonnegative matrix factorization domain and present a stochastic variational inference algorithm for this model as well.

We apply our algorithms to 1.8 million documents from the *New York Times*. Processing this data in the traditional batch inference approach would be extremely expensive computationally since parameters for each document would need to be optimized before global parameters could be updated; MCMC methods are even less feasible. Using stochastic optimization, we show how stochastic VB can quickly learn the approximate posterior of these nonnegative matrix factorization models. Before deriving these inference algorithms, we give a general review of the stochastic VB approach.

We organize the chapter as follows: In Section 11.2 we review the latent indicator approach probabilistic topic modeling, which forms the jumping-off point for the matrix factorization models we consider. In Section 11.3 we review the Bayesian extension to NMF and present an alternate mixture representation of this model that highlights to relationship to existing models (Blei et al., 2003; Teh et al., 2007). In this section we also present correlated NMF, a matrix factorization model with similar objectives as the CTM. In Section 11.4 we review mean field variational inference in both its batch and stochastic forms. In Section 11.5 we present the stochastic inference algorithm for Bayesian NMF and correlated NMF. In Section 11.6 we apply the algorithm to 1.8 million documents from the *New York Times*.

11.2 Background: Probabilistic Topic Models

Probabilistic topic models assume a probabilistic generative structure for a corpus of text documents. They are an effective method for uncovering the salient themes within a corpus, which can help the user navigate large collections of text. Topic models have also been applied to a wide variety of data modeling problems, including those in image processing Fei Fei and Perona (2005) and political science Grimmer, J. (2010(@), and are not restricted to document modeling applications, though modeling text will be the focus of this chapter.

A probabilistic topic model assumes the existence of an underlying collection of "topics," each topic being a distribution on words in a vocabulary, as well as a distribution on these topics for each document. For a K-topic model, we denote the set of topics as $\beta_k \in \Delta_V$, where β_{kv} is the probability of word index v given that a word comes from topic k. For document d, we denote the distribution on these K topics as $\theta_d \in \Delta_K$, where θ_{dk} is the probability that a word in document d comes from topic k.

For a corpus of D documents generated from a vocabulary of V words, let $w_{dn} \in \{1, \ldots, V\}$ denote the nth word in document d. In its most basic form, a latent-variable probabilistic topic model assumes the following hierarchical structure for generating this word,

$$w_{dn} \overset{ind}{\sim} \text{Discrete}(\beta_{z_{dn}}), \quad z_{dn} \overset{iid}{\sim} \text{Discrete}(\theta_d). \tag{11.1}$$

The discrete distribution indicates that $\Pr(z_{dn} = i|\theta_d) = \theta_{di}$.

Therefore, to populate a document with words, one first selects the topic, or theme of each word, followed by the word-value itself using the distribution indexed by its topic. In this chapter, we work within the "bag-of-words" context, which assumes that the N_d words within document d are exchangeable; that is, the order of words in the document does not matter according to the model. We next review two bag-of-words probabilistic topic models.

Latent Dirichlet Allocation. A Bayesian topic model places prior distributions on β_k and θ_d. The canonical example of a Bayesian topic model is latent Dirichlet allocation (LDA) (Blei et al., 2003), which places Dirichlet distribution priors on these vectors,

$$\beta_k \overset{iid}{\sim} \text{Dirichlet}(c_0 \mathbf{1}_V / V), \quad \theta_d \overset{iid}{\sim} \text{Dirichlet}(a_0 \mathbf{1}_K). \tag{11.2}$$

The vector $\mathbf{1}_a$ is an a-dimensional vector of ones. LDA is an example of a conjugate exponential family model; all conditional posterior distributions are closed-form and in the same distribution family as the prior. This give LDA a significant algorithmic advantage.

Correlated Topic Models. One potential drawback of LDA is that the Dirichlet prior on θ_d does not model correlations between topic probabilities. This runs counter to a priori intuition, which says that some topics are more likely to co-occur than others (e.g., topics on "politics" and "military" versus a topic on "cooking"). A correlated topic model (CTM) was proposed Blei and Lafferty (2007) to address this issue. This model replaces the Dirichlet distribution prior on θ_d with a logistic normal distribution prior (Aitchison, 1982),

$$\theta_{dk} = \exp\{y_{dk}\} / \textstyle\sum_{j=1}^{K} \exp\{y_{dj}\}, \quad y_d \sim \text{Normal}(0, \mathrm{C}). \tag{11.3}$$

The covariance matrix C contains the correlation information for the topic probabilities. To allow for this correlation structure to be determined by the data, the covariance matrix C has a conjugate inverse Wishart prior,

$$\mathrm{C} \sim \text{invWishart}(A, m). \tag{11.4}$$

The correlated topic model can therefore "anticipate" co-occuring themes better than LDA, but since the logistic normal distribution is not conjugate to the multinomial, inference is not as straightforward.

Matrix Factorization Representations. As mentioned, these hierarchical Bayesian priors are presented within the context of latent indicator topic models. The distinguishing characteristic of this framework is the hidden data z_{dn}, which indicates the topic of word n in document d. Marginalizing out these random variables, one enters the domain of nonnegative matrix factorization (Lee and Seung, 2001; Gaussier and Goutte, 2005; Singh and Gordon, 2008). In this modeling framework, the data is restructured into a matrix of nonnegative integers, $X \in \mathbb{N}^{V \times D}$. The entry X_{vd} is a count of the number of times word v appears in document d. Therefore,

$$X_{vd} = \sum_{n=1}^{N_d} \mathbb{I}(w_{dn} = v). \tag{11.5}$$

Typically, most values of X can be expected to equal zero. Several matrix factorization approaches exist for modeling this representation of the data. In the next section, we discuss two NMF models for this data matrix.

11.3 Two Parametric Models for Bayesian Nonnegative Matrix Factorization

As introduced in the previous section, our goal is to factorize a $V \times D$ data matrix X of nonnegative integers. This matrix arises by integrating out the latent topic indicators associated with each word in a probabilistic topic model, thus turning a latent indicator model into a nonnegative matrix factorization model. The matrix to be factorized is not X, but an underlying matrix of nonnegative latent variables $\Lambda \in \mathbb{R}_+^{V \times D}$. Each entry of this latent matrix is associated with a corresponding entry in X, and we assume a Poisson data-generating distribution, with $X_{vd} \sim \text{Poisson}(\Lambda_{vd})$.

A frequently used model for X is simply called NMF, and was presented by Lee and Seung (1999). This model assumes Λ to be low-rank, the rank K being chosen by the modeler, and factorized into the matrix product $\Lambda = B\Theta$, with $B \in \mathbb{R}_+^{V \times K}$ and $\Theta \in \mathbb{R}_+^{K \times D}$. Lee and Seung (2001) presented optimization algorithms for two penalty functions; in this chapter we focus on the Kullback-Leibler (KL) penalty. This KL penalty has a probabilistic interpretation, since it results in an optimization program for NMF that is equivalent to a maximum likelihood approximation of the Poisson generating model,

$$\{B^*, \Theta^*\} = \max_{B,\Theta} \; P(X|B,\Theta) = \max_{B,\Theta} \; \textstyle\prod_{v,d} \text{Poisson}(X_{vd}|(B\Theta)_{vd}).$$

A major attraction of the NMF algorithm is the fast multiplicative update rule for learning B and Θ (Lee and Seung, 2001). We next review the Bayesian extension of NMF (Cemgil, 2009). We then present a correlated NMF model that takes its motivation from the the latent-indicator correlated topic model.

11.3.1 Bayesian NMF

The NMF model with KL penalty was recently extended to the Bayesian setting under the name Bayesian NMF (Cemgil, 2009). This extension places gamma priors on all elements of B and Θ. The generative process of Bayesian NMF under our selected parameterization is

$$X_{vd} \sim \text{Poisson}(\textstyle\sum_{k=1}^{K} \beta_{vk}\theta_{kd}), \tag{11.6}$$

$$\beta_{vk} \overset{iid}{\sim} \text{Gamma}(c_0/V, c_0), \quad \theta_{kd} \overset{iid}{\sim} \text{Gamma}(a_0, b_0). \tag{11.7}$$

Note that $\sum_v \beta_{vk} \neq 1$ with probability one. We also observe that this is not a matrix factorization approach to LDA, though β and θ serve similar functions and have a similar interpretation, as discussed below. Therefore, it is still meaningful to refer to $\beta_{:k}$ as a "topic," and we adopt this convention below.

Just as the latent-variable probabilistic topic models discussed in Section 11.2 have nonnegative matrix factorization representations, the reverse direction holds for Bayesian NMF. The latent-variable representation of Bayesian NMF is insightful since it shows a close relationship with LDA. Using the data-generative structure given in Equation (11.1) (with an additional $\hat{}$ to distinguish from Equation (11.7)), the latent topics and distributions on these topics have the following generative process:

$$\hat{\beta}_k \overset{iid}{\sim} \text{Dirichlet}(c_0 \mathbf{1}_V / V), \quad \hat{\theta}_{dk} := \xi_{dk} \tilde{\theta}_{dk} \tag{11.8}$$

$$\xi_{dk} := \frac{e_k}{\sum_{j=1}^K \theta_{dj} e_j}, \quad \tilde{\theta}_d \overset{iid}{\sim} \text{Dirichlet}(a_0 \mathbf{1}_K), \quad e_k \overset{iid}{\sim} \text{Gamma}(c_0, c_0). \tag{11.9}$$

The vectors $\hat{\beta}_k$ and $\hat{\theta}_d$ correspond to the topics and document distributions on topics, respectively. Note that $\sum_k \hat{\theta}_{dk} = 1$. We see that when $e_k = 1$, LDA is recovered.[1] Thus, when the columns of B are restricted to the probability simplex, that is, when $e_k = 1$ with probability 1 for each k, one obtains the matrix factorization representation of LDA, also called GaP (Canny, 2004). Relaxing this constraint to gamma distributed random variables allows for a computationally simpler variational inference algorithm for the matrix factorization model, which we give in Section 11.5.1.

The representation in Equations (11.8) and (11.9)) shows the motivation for parameterizing the gamma distributions on β as done in Equation (11.7). The desire is for ξ to be close to one, which results in a model close to LDA. This parameterization gives a good approximation; since c_0 is commonly set equal to a fraction of V in LDA, for example $c_0 = 0.1V$, and because V is often on the order of thousands, the distribution of e_k is highly peaked around 1, with $\mathbb{E}[e_k] = 1$ and $\text{Var}(e_k) = 1/c_0$. Though this latent variable representation affords some insight into the relationship between Bayesian NMF and LDA, we derive a cleaner inference algorithm using the hierarchical structure in Equations (11.6) and (11.7).

11.3.2 Correlated NMF

We next propose a correlated NMF model, which we build on an alternate representation of the correlated topic model (CTM) (Blei and Lafferty, 2007). To derive the model, we first present the alternate representation of the CTM. Following a slight alteration to the prior on the covariance matrix C, we show how we can "unpack the information" in the CTM to allow for a greater degree of exploratory data analysis.

Recall that an inverse Wishart prior was placed on C, the covariance of the document-specific lognormal vectors, in Equation (11.4). Instead, we propose a Wishart prior,

$$\text{C} \sim \text{Wishart}(\sigma^2 I_K, m), \tag{11.10}$$

and assume a diagonal matrix parameter. Though this change appears minor, it allows for the prior to be expanded hierarchically in a way that allows the model parameters to contain more information that can aid in understanding the underlying dataset.

There are two steps to unpacking the CTM. For the first step, we observe that one can sample C from its Wishart prior by first generating a matrix $L \in \mathbb{R}^{m \times K}$, where each entry

[1] This additional random variable e_k arises out of the derivation by defining $e_k := \sum_{v=1}^V \beta_{vk}$, with β_{vk} drawn as in Equation (11.7). Nevertheless, e_k can be shown to be independent of all other random variables.

$L_{ij} \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$, and then defining $\text{C} := L^T L$. It follows that C has the desired Wishart distribution.

Intuitively, with this expansion each topic now has a "location" ℓ_k, being the kth column in L. That is, column k of topic matrix B now has an associated latent location $\ell_k \in \mathbb{R}^m$, where $\ell_k \overset{iid}{\sim} \text{Normal}(0, \sigma^2 I_m)$. Note that when $m < K$, the covariance C is not full rank. This provides additional modeling flexibility to the CTM, which in previous manifestations required a full rank estimate of C (Blei and Lafferty, 2007).

The second step in unpacking the CTM is to define an alternate representation of $y_d \sim \text{Normal}(0, L^T L)$. We recall that this is the logistic normal vector that is passed to the softmax function in order to obtain a distribution on topics for document d, as described in Equation (11.3). We can again introduce Gaussian vectors, this time to construct y_d:

$$y_d := L^T u_d, \quad u_d \sim \text{Normal}(0, I_m). \tag{11.11}$$

The marginal distribution of y_d, or $p(y_d|L) = \int_{\mathbb{R}^m} p(y_d|L, u_d)p(u_d)\, du_d$, is a $\text{Normal}(0, L^T L)$ distribution, as desired. To derive this marginal, first let $y_d|L, u_d, \epsilon \sim \text{Normal}(L^T u_d, \epsilon)$, next calculate $p(y_d|L, \epsilon) = \text{Normal}(0, \epsilon I + L^T L)$, and finally let $\epsilon \to 0$. As with topic location ℓ_k, the vector u_d also has an interpretation as a location for document d. These locations are useful for search applications, as we show in Section 11.6.

For the latent variable CTM, this results in a new hierarchical prior for topic distribution θ_d. The previous hierarchical prior of Equation (11.3) becomes the following,

$$\theta_{dk} = \frac{\exp\{\ell_k^T u_d\}}{\sum_j \exp\{\ell_j^T u_d\}}, \quad \ell_k \overset{iid}{\sim} \text{Normal}(0, \sigma^2 I_m), \quad u_d \overset{iid}{\sim} \text{Normal}(0, I_m). \tag{11.12}$$

Transferring this into the domain of nonnegative matrix factorization, we observe that the normalization of the exponential is unnecessary. This is for a similar reason as with the random variables in Bayesian NMF, which made the transition from being Dirichlet distributed to gamma distributed. We also include a bias term α_d for each document. This performs the scaling necessary to account for document length.

The generative process for correlated NMF is similar to Bayesian NMF, with many distributions being the same. The generative process below for correlated NMF is

$$X_{vd} \sim \text{Poisson}(\textstyle\sum_{k=1}^K \beta_{vk} \exp\{\alpha_d + \ell_k^T u_d\}), \tag{11.13}$$

$$\beta_{vk} \overset{iid}{\sim} \text{Gamma}(c_0/V, c_0), \quad \ell_k \overset{iid}{\sim} \text{Normal}(0, \sigma^2 I_m), \quad u_d \overset{iid}{\sim} \text{Normal}(0, I_m).$$

The scaling performed by α_d allows the product $\ell_k^T u_d$ to only model random effects. We learn a point estimate of this parameter.

The latent locations introduced to the CTM and this model requires the setting of the latent space dimension m. Since we are in effect modeling an m-rank covariance matrix C with these vectors, the variety of correlations decreases with m, and the model becomes more restrictive in the distributions on topics it can model. On the other hand, one should set $m \leq K$, since for $m > K$ there are $m - K$ redundant dimensions.

11.4 Stochastic Variational Inference

Text datasets can often be classified as a "big data" problem. For example, Wikipedia currently indexes several million entries, and the *New York Times* has published almost two million articles

in the last 20 years. In other problem domains the amount of data is even larger. For example, a hyperspectral image can contain a hundred million voxels in a *single* data cube. With so much data, fast inference algorithms are essential. Stochastic variational inference (Sato, 2001) is a significant step in this direction for hierarchical Bayesian models.

Stochastic variational inference exploits the difference between *local* variables, or those associated with a single unit of data, and *global* variables, which are shared among an entire dataset. In brief, stochastic VB works by splitting a large dataset into smaller groups. These small groups can be quickly processed, with each iteration processing a new group of data. In the context of probabilistic topic models, the unit of data is a document, and the global variables are the topics (among other possible variables), while the local variables are document-specific and relate to the distribution on these topics.

Recent stochastic inference algorithms developed for LDA (Hoffman et al., 2010b), the HDP (Wang et al., 2011), and other models (e.g., in Paisley et al., 2012) have shown rapid speed-ups in inference for probabilistic topic models. Though mainly applied to latent-indicator topic models thus far, the underlying theory of stochastic VB is more general, and applies to other families of models. One goal of this chapter is to show how this inference method can be applied to nonnegative matrix factorization, placing the resulting algorithms in the family of online matrix factorization methods (Mairal et al., 2010). Specifically, we develop stochastic variational inference algorithms for the Bayesian NMF and correlated NMF models discussed in Section 11.3.

We next review the relevant aspects of variational inference that make deriving stochastic algorithms easy. Our approach is general, which will allow us to immediately derive the update rules for the stochastic VB algorithm for Bayesian NMF and correlated NMF. We focus on conjugate exponential models and present a simple derivation on a toy example—one for which online inference is not necessary, but which allows us to illustrate the idea.[2]

11.4.1 Mean Field Variational Bayes

Mean field variational inference is an approximate Bayesian inference method (Jordan et al., 1999). It approximates the full posterior of a set of model parameters $p(\Phi|X)$ with a factorized distribution $Q(\Phi) = \prod_i q(\phi_i)$ by minimizing their Kullback-Liebler divergence. This is done by maximizing the variational objective $\mathcal{L}$ with respect to the variational parameters Ψ of Q. The objective function is

$$\mathcal{L}(X, \Psi) = \mathbb{E}_Q[\ln p(X, \Phi)] + \mathbb{H}[Q]. \tag{11.14}$$

When the prior and likelihood of all nodes of the model falls within the conjugate exponential family, variational inference has a simple optimization procedure (Winn and Bishop, 2005). We illustrate this with the following example, which we extend to the stochastic setting in Section 11.4.2. This generic example gives the general form of the stochastic variational inference algorithm, which we later applied to Bayesian NMF and correlated NMF.

Consider D independent samples from an exponential family distribution $p(x|\eta)$, where η is the natural parameter vector. The data likelihood under this model has the standard form

$$p(X|\eta) = \left[\prod_{d=1}^{D} h(x_d)\right] \exp\left\{\eta^T \sum_{d=1}^{D} t(x_d) - DA(\eta)\right\}.$$

The sum of vectors $t(x_d)$ forms the sufficient statistic of the likelihood. The conjugate prior on η has a similar form

$$p(\eta|\chi, \nu) = f(\chi, \nu) \exp\left\{\eta^T \chi - \nu A(\eta)\right\}, \tag{11.15}$$

[2]Although Bayesian NMF is not in fact fully conjugate, we will show that a bound introduced for tractable inference modifies the joint likelihood such that the model effectively is conjugate. For correlated NMF, we will also make adjustments for non-conjugacy.

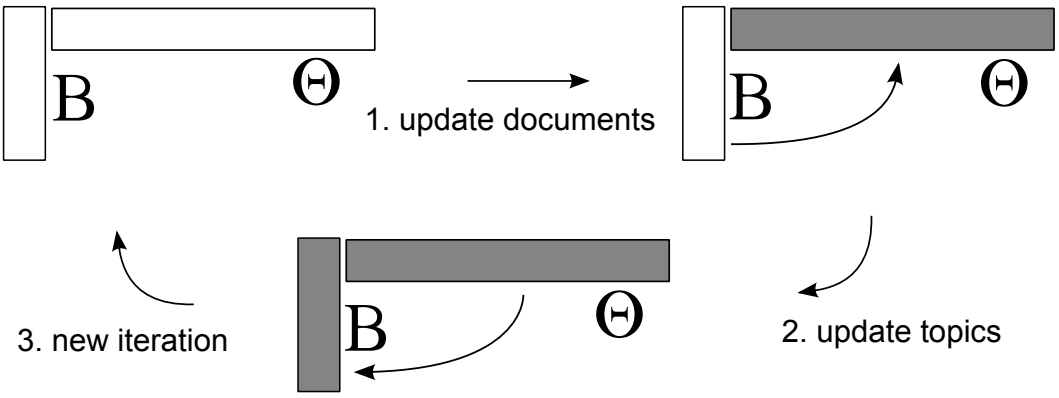

FIGURE 11.1
A graphic describing batch variational inference for Bayesian nonnegative matrix factorization. For each iteration, all variational parameters for document specific (local) variables are updated first. Using these updated values, the variational parameters for the global topics are updated. When there are many documents being modeled, i.e., when the number of columns is very large, Step 1 in the image can have a long runtime.

and conjugacy motivates selecting a q distribution in this same family,

$$q(\eta|\chi',\nu') = f(\chi',\nu')\exp\left\{\eta^T\chi' - \nu' A(\eta)\right\}. \tag{11.16}$$

After computing the variational lower bound given in Equation (11.14), which can be done explicitly for this example, inference proceeds by taking gradients with respect to variational parameters, in this case the vector $\psi := [\chi'^T, \nu']^T$, and then setting to zero to find their updated values. For conjugate exponential family models, this gradient has the general form

$$\nabla_\psi \mathcal{L}(X,\Psi) = -\begin{bmatrix} \frac{\partial^2 \ln f(\chi',\nu')}{\partial\chi'\partial\chi'^T} & \frac{\partial^2 \ln f(\chi',\nu')}{\partial\chi'\partial\nu'} \\ \frac{\partial^2 \ln f(\chi',\nu')}{\partial\nu'\partial\chi'^T} & \frac{\partial^2 \ln f(\chi',\nu')}{\partial\nu'^2} \end{bmatrix} \begin{bmatrix} \chi + \sum_{d=1}^{D} t(x_d) - \chi' \\ \nu + D - \nu' \end{bmatrix}, \tag{11.17}$$

as can be explicitly derived from the lower bound. Setting this to zero, one can immediately read off the variational parameter updates from the right vector, which in this case are $\chi' = \chi + \sum_{d=1}^{D} t(x_d)$ and $\nu' = \nu + D$. Though the matrix in Equation (11.17) is often very complicated, it is superfluous to batch variational inference for conjugate exponential family models. In the stochastic optimization of Equation (11.14), however, this matrix cannot be similarly ignored.

We show a visual representation of batch variational inference for Bayesian matrix factorization in Figure 11.1. The above procedure repeats for each variational Q distribution; first for all distributions of the right matrix, followed by those of the left. We note that, if conjugacy does not hold, gradient ascent can be used to optimize ψ.

11.4.2 Stochastic Optimization of the Variational Objective

Stochastic optimization of the variational lower bound involves forming a noisy gradient of $\mathcal{L}$ using a random subset of the data at each iteration. Let $C_t \subset \{1,\ldots,D\}$ index this subset at iteration t. Also, let ϕ_d be the model variables associated with observation x_d and Φ_X the variables shared among all observations. In Table 11.1, we distinguish the local from the global variables for Bayesian NMF and correlated NMF.

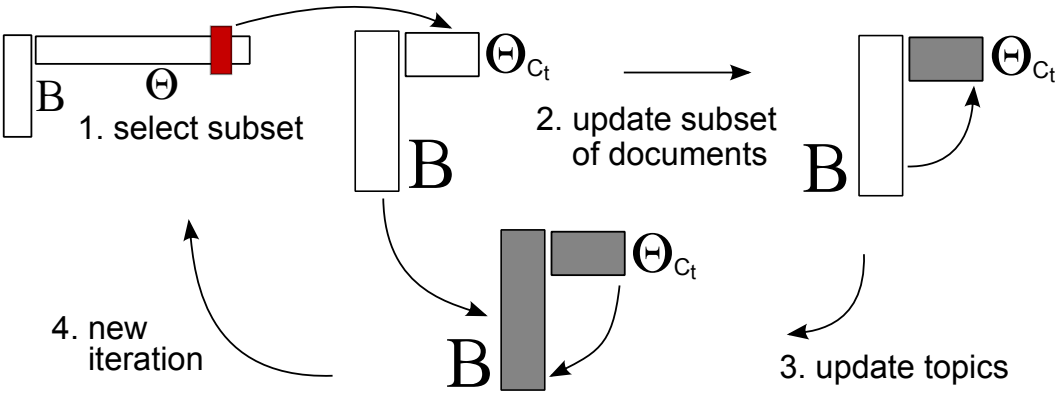

FIGURE 11.2
A graphic describing stochastic variational inference for Bayesian nonnegative matrix factorization. From the larger dataset, first select a subset of data (columns) uniformly at random, indexed by C_t at iteration t; for clarity we represent this subset as a contiguous block. Next, fully optimize the local variational parameters for each document. Because the subset is much smaller than the entire dataset, this step is fast. Finally, update the global topic variational parameters using a combination of information from the local updates and the previously seen documents, as summarized in the current values of these global variational parameters.

The *stochastic variational objective function* $\mathcal{L}_s$ is the noisy version of $\mathcal{L}$ formed by selecting a subset of the data,

$$\mathcal{L}_s(X_{C_t}, \Psi) = \frac{D}{|C_t|} \sum_{d \in C_t} \mathbb{E}_Q[\ln p(x_d, \phi_d | \Phi_X)] + \mathbb{E}_Q[\ln p(\Phi_X)] + \mathbb{H}[Q]. \tag{11.18}$$

This constitutes the objective function at step t. By optimizing $\mathcal{L}_s$, we are optimizing $\mathcal{L}$ in expectation. That is, since each subset C_t is equally probable, with $p(C_t) = \binom{D}{|C_t|}^{-1}$, and since $d \in C_t$ for $\binom{D-1}{|C_t|-1}$ of the $\binom{D}{|C_t|}$ possible subsets, it follows that

$$\mathbb{E}_{p(C_t)}[\mathcal{L}_s(X_{C_t}, \Psi)] = \mathcal{L}(X, \Psi). \tag{11.19}$$

Therefore, by optimizing $\mathcal{L}_s$ we are stochastically optimizing $\mathcal{L}$. Stochastic variational optimization proceeds by optimizing the objective in Equation (11.18) with respect to ψ_d, $d \in C_t$, followed by an update to Ψ_X that blends the new information with the old. For example, in the simple conjugate exponential model of Section 11.4.1, the update of the vector $\psi := [\chi'^T, \nu']^T$ at iteration t follows a gradient step,

$$\psi_t = \psi_{t-1} + \rho_t G \nabla_\psi \mathcal{L}_s(X_{C_t}, \Psi). \tag{11.20}$$

The matrix G is a positive definite preconditioning matrix and ρ_t is a step size satisfying $\sum_{t=1}^{\infty} \rho_t = \infty$ and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$, which ensures convergence (Bottou, 1998).

The key to stochastic variational inference for conjugate exponential models is in selecting G. Since the gradient of $\mathcal{L}_s$ has the same form as Equation 11.17, the difference being a sum over $d \in C_t$ rather than the entire dataset, G can be set to the inverse of the matrix in (11.17) to allow for cancellation. An interesting observation is that this matrix is

$$G = -\left(\frac{\partial^2 \ln q(\eta|\psi)}{\partial \psi \partial \psi^T} \right)^{-1}, \tag{11.21}$$

which is the inverse Fisher information of the variational distribution $q(\eta|\psi)$. This setting of G

Model	Local variables	Global variables
Bayesian NMF	$\{\theta_{kd}\}_{k=1:K,d=1:D}$	$\{\beta_{vk}\}_{v=1:V,k=1:K}$
Correlated NMF	$\{u_d, \alpha_d\}_{d=1:D}$	$\{\beta_{vk}, \ell_k\}_{v=1:V,k=1:K}$

TABLE 11.1
Local and global variables for the two Bayesian nonnegative matrix factorization models considered in this chapter. Stochastic variational inference partitions the local variables into batches, with each iteration of inference processing one batch. Updates to the global variables follow each batch.

gives the natural gradient of the lower bound, and therefore not only simplifies the algorithm, but gives an efficient step direction Amari (1998); Sato (2001). We note that this is the setting of G given in the stochastic variational algorithm of Sato (2001) and was used in Hoffman et al. (2010b) and Wang et al. (2011) for online LDA and HDP, respectively.

In the case where the prior-likelihood pair does not fall within the conjugate exponential family, stochastic variational inference still proceeds as described, instead using an appropriate G for the gradient step in Equation 11.20. The disadvantage of this regime is that the method truly is a gradient method, with the attendant step size issues. Using the Fisher information gives a clean and interpretable update.

This interpretability is seen by returning to the example in Section 11.4.1, where the stochastic variational parameter updates are

$$\begin{aligned} \chi'_t &= (1-\rho_t)\chi'_{t-1} + \rho_t \left\{ \chi + \frac{D}{|C_t|} \sum_{d \in C_t} t(x_d) \right\}, \\ \nu' &= \nu + D. \end{aligned} \tag{11.22}$$

We see that, for conjugate exponential family distributions, each step of stochastic variational inference entails a weighted averaging of sufficient statistics from previous data with the sufficient statistics of new data *scaled up* to the size of the full data set. We show a visual representation of stochastic variational inference for Bayesian matrix factorization in Figure 11.2.

11.5 Variational Inference Algorithms

We present stochastic variational inference algorithms for Bayesian NMF and correlated NMF. Table 11.2 contains a list of the variational q distributions we use for each model. The variational objective functions for both models are given below. Since an expectation with respect to a delta function is simply an evaluation at the point mass, we write this evaluation for correlated NMF:

$$\begin{aligned} \text{B-NMF:} \quad & \mathcal{L} = \mathbb{E}_q \ln p(X|B, \Theta, \boldsymbol{\alpha}) + \mathbb{E}_q \ln p(B) + \mathbb{E}_q \ln p(\Theta) + \mathbb{H}[Q] \\ \text{C-NMF:} \quad & \mathcal{L} = \mathbb{E}_q \ln p(X|B, L, U, \boldsymbol{\alpha}) + \mathbb{E}_q \ln p(B) + \ln p(L) + \ln p(U) + \mathbb{H}[Q]. \end{aligned}$$

As is evident, mean field variational inference requires being able to take expectations of the log joint likelihood with respect to the predefined q distributions. As frequently occurs with VB inference, this is not possible here. We adopt the common solution of introducing a tractable lower bound for the problematic function, which we discuss next.

Model	Variational q distributions
Bayesian NMF	$q(\beta_{vk}) = \text{Gamma}(\beta_{vk}\|g_{vk}, h_{vk})$
	$q(\theta_{kd}) = \text{Gamma}(\theta_{kd}\|a_{kd}, b_{kd})$
Correlated NMF	$q(\beta_{vk}) = \text{Gamma}(\beta_{vk}\|g_{vk}, h_{vk})$
	$q(\ell_k) = \delta_{\ell_k},\ q(u_k) = \delta_{u_k}$

TABLE 11.2
The variational q distributions for Bayesian NMF and correlated NMF.

A Lower Bound of the Variational Objective Function

For both Bayesian NMF and correlated NMF, the variational lower bound contains an intractable expectation in the log of the Poisson likelihood. To speak in general terms about the problem, let ω_{kd} represent the document weights. This corresponds to θ_{kd} in Bayesian NMF and to $\exp\{\alpha_d + \ell_k^T u_d\}$ in correlated NMF.

The problematic expectation is $\mathbb{E}_q \ln \sum_{k=1}^K \beta_{vk}\omega_{kd}$. Given the concavity of the natural logarithm, we introduce a probability vector $p^{(vd)} \in \Delta_K$ for each (v, d) pair in order to lower bound this function,

$$\ln\left(\sum_{k=1}^K \beta_{vk}\omega_{kd}\right) \;\geq\; \sum_{k=1}^K p_k^{(vd)} \ln(\beta_{vk}\omega_{kd}) - \sum_{k=1}^K p_k^{(vd)} \ln p_k^{(vd)}. \tag{11.23}$$

All expectations of this new function are tractable, and the vector $p^{(vd)}$ is an auxiliary parameter that we optimize with the rest of the model. After each iteration, we optimize this auxiliary probability vector to give the tightest lower bound. This optimal value is

$$p_k^{(vd)} \propto \exp\{\mathbb{E}_q[\ln \beta_{vk}] + \mathbb{E}_q[\ln \omega_{kd}]\}. \tag{11.24}$$

Section 11.5.1 contains the functional forms of these expectations.

11.5.1 Batch Algorithms

Given the relationship between batch and stochastic variational inference, we first present the batch algorithm for Bayesian NMF and correlated NMF, followed by the alterations needed to derive their stochastic algorithms. For each iteration of inference, batch variational inference cycles through the following updates to the parameters of each variational distribution.

Parameter Update for $q(\beta_{vk})$

The two gamma distribution parameters for this q distribution (Table 11.2) have the following updates,

$$g_{vk} = \frac{c_0}{V} + \textstyle\sum_{d=1}^D X_{vd} p_k^{(vd)}, \tag{11.25}$$

$$h_{vk} = c_0 + \textstyle\sum_{d=1}^D \mathbb{E}_q[\theta_{kd}], \qquad \textit{(Bayesian NMF)} \tag{11.26}$$

$$h_{vk} = c_0 + \textstyle\sum_{d=1}^D \exp\{\alpha_d + \ell_k^T u_d\}. \quad \textit{(Correlated NMF)} \tag{11.27}$$

Expectations used in other parameter updates are $\mathbb{E}_q[\beta_{vk}] = g_{vk}/h_{vk}$ and $\mathbb{E}_q[\ln \beta_{vk}] = \psi(g_{vk}) - \ln h_{vk}$.

Parameter Update for $q(\theta_{kd})$ (Bayesian NMF)

The two gamma distribution parameters for this q distribution (Table 11.2) have the following updates,

$$a_{kd} = a_0 + \sum_{v=1}^{V} X_{vd} p_k^{(vd)}, \tag{11.28}$$

$$b_{kd} = b_0 + \sum_{v=1}^{V} \mathbb{E}_q[\beta_{vk}]. \tag{11.29}$$

Expectations used in other parameter updates are $\mathbb{E}_q[\theta_{kd}] = a_{kd}/b_{kd}$ and $\mathbb{E}_q[\ln \theta_{kd}] = \psi(a_{kd}) - \ln b_{kd}$.

Parameter Updates for $q(\ell_k)$ and $q(u_d)$ (Correlated NMF)

Since these parameters do not have closed-form updates, we use the steepest ascent gradient method for inference. The gradients of $\mathcal{L}$ with respect to ℓ_k and u_k are

$$\nabla_{\ell_k} \mathcal{L} = \sum_{v=1}^{V} \sum_{d=1}^{D} \left(X_{vd} p_k^{(vd)} - \mathbb{E}_q[\beta_{vk}] \exp\{\alpha_d + \ell_k^T u_d\} \right) u_d - \sigma^{-2} \ell_k, \tag{11.30}$$

$$\nabla_{u_d} \mathcal{L} = \sum_{v=1}^{V} \sum_{k=1}^{K} \left(X_{vd} p_k^{(vd)} - \mathbb{E}_q[\beta_{vk}] \exp\{\alpha_d + \ell_k^T u_d\} \right) \ell_k - u_d. \tag{11.31}$$

For each variable, we take several gradient steps to approximately optimize its value before moving to the next variable.

Paramter Update for α_d (Correlated NMF)

The point estimate for α_d has the following closed-form solution,

$$\alpha_d = \ln \sum_{v=1}^{V} X_{vd} - \ln \sum_{v,k} \mathbb{E}_q[\beta_{vk}] \exp\{\ell_k^T u_d\}. \tag{11.32}$$

We update this parameter after each step of u_d.

11.5.2 Stochastic Algorithms

By inserting the lower bound (11.23) into the log joint likelihood and then exponentiating, one can see that the likelihood β_{vk} is modified to form a conjugate exponential pair with its prior for both models. Hence, the discussion and theory of natural gradient ascent in Section 11.4.2 applies to both models with respect to the topic matrix B. For correlated NMF, this does not apply to the global variable ℓ. For this variable, we use the alternate gradient method discussed in Section 11.4.2.

After selecting a subset of the data using the index set $C_t \in \{1, \ldots, D\}$, stochastic inference starts by optimizing the local variables, which entails iterating between the parameter updates for θ_d and $p^{(vd)}$ for Bayesian NMF, and u_d and $p^{(vd)}$ for correlated NMF. Once these parameters have converged, we take a single step in the direction of the natural gradient to update the distributions on $\beta_{:,k}$ and use Newton's method in the step for ℓ_k. We use a step size of the form $\rho_t = (t_0 + t)^{-\kappa}$ for $t_0 > 0$ and $\kappa \in (.5, 1]$. This step size satisfies the necessary conditions for convergence discussed in Section 11.4.2 (Bottou, 1998). We also recall from Section 11.4.2 that D is the corpus size to which each batch C_t is scaled up.

Stochastic Update of $q(\beta_{vk})$

As with batch inference, this update is similar for Bayesian NMF and correlated NMF. In keeping with the generalization at the beginning of this section, we let ω_{kd} stand for θ_{kd} or $\exp\{\alpha_d + \ell_k^T u_d\}$,

depending on the model under consideration. The update of the variational parameters of $q(\beta_{vk})$ is

$$g_{vk}^{(t)} = (1-\rho_t)g_{vk}^{(t-1)} + \rho_t \left\{ \frac{c_0}{V} + \frac{D}{|C_t|} \sum_{d \in C_t} X_{vd} p_k^{(vd)} \right\}, \tag{11.33}$$

$$h_{vk}^{(t)} = (1-\rho_t)h_{vk}^{(t-1)} + \rho_t \left\{ c_0 + \frac{D}{|C_t|} \sum_{d \in C_t} \mathbb{E}_q[\omega_{kd}] \right\}. \tag{11.34}$$

As expected from the theory, the variational parameters are a weighted average of their previous values and the sufficient statistics calculated using batch C_t.

Stochastic Update of $q(\ell_k)$ (Correlated NMF)

Stochastic inference for correlated NMF has an additional global variable in the location of each topic. The posterior of this variable—a point estimate—is not conjugate with the prior, and therefore we do not use the natural gradient stochastic VB approach discussed in Section 11.4.2. However, as pointed out in that section, we can still perform stochastic inference according to the general update given in Equation (11.20). With reference to this equation, we set the preconditioning matrix G to be the inverse negative Hessian and update ℓ_k at iteration t as follows,

$$\begin{aligned} \ell_k^{(t)} &= \ell_k^{(t-1)} + \rho_t G \nabla_{\ell_k} \mathcal{L}_s(X_{C_t}, U_{C_t}, \boldsymbol{\alpha}_{C_t}, L, B), \\ G^{-1} &= \sigma^{-2} I_m + \frac{D}{|C_t|} \sum_{d \in C_t} \sum_{v=1}^{V} \mathbb{E}_q[\beta_{vk}] \exp\{\alpha_d + \ell_k^T u_d\} u_d u_d^T. \end{aligned} \tag{11.35}$$

In batch inference, we perform gradient (steepest) ascent optimization as well. A key difference there is that we fully optimize each ℓ_k and u_d before moving to the next variable—indeed, for stochastic VB we still fully optimize the local variable u_d with steepest ascent during each iteration. For stochastic learning of ℓ_k, however, we only take *one* step in the direction of the gradient for the stochastic update of ℓ_k before moving on to a new batch of documents. In an attempt to take the best step possible, we use the Hessian matrix to construct a Newton step.

11.6 Experiments

We perform experiments using stochastic variational inference to learn the variational posteriors of Bayesian NMF and correlated NMF. We compare these algorithms with online LDA of Hoffman et al. (2010b). We summarize the dataset, parameter settings, experimental setup, and performance evaluation method below.

Dataset. We work with a data set of 1,819,268 articles from the *New York Times* newspaper. The article dates range from January 1987 to May 2007. We use a dictionary of $V = 8000$ words learned from the data, and randomize the order of the articles for processing.

Parameter Settings. In all experiments, we set the parameter $c_0 = 0.05V$. When learning a K-topic model, for online LDA we set the parameter for the Dirichlet distribution on topic weights to $a_0 = 1/K$, and for Bayesian NMF we set the weight parameters to $a_0 = 1/K$ and $b_0 = 1/K$. For correlated NMF we use a latent space dimensionality of $m = 50$ for all experiments, and set $\sigma^2 = 1/m$.

Experimental setup. We compare stochastic inference for Bayesian NMF and correlated NMF with online LDA. For all models, we perform experiments for $K \in \{50, 100, 150\}$ topics. We also evaluate the inference method for several batch sizes using $|C_t| \in \{500, 1000, 1500, 2000\}$. We use a step size of $\rho_t = (1 + t)^{-0.5}$. Stochastic inference requires initialization of all global variational parameters. For topic-related parameters, we set the variational parameters to be the prior plus a Uniform(0,1) random variable that is scaled to the size of the corpus, similar to the scaling performed on the statistics of each batch. For correlated NMF, we sample $\ell_k \sim \text{Normal}(0, \sigma^2 I_m)$.

Performance Evaluation. To evaluate performance, we hold out every tenth batch for testing. On each testing batch, we perform threefold cross validation by partitioning each document into thirds. Using the current values of the global variational parameters, we then train the local variables on two-thirds of each document, and predict the remaining third. For prediction, we use the mean of each variational q distribution. We average the per-word log likelihoods of all words tested to quantify the performance of the model at the current step of inference. After testing the batch, stochastic inference proceeds as before, with the testing batch processed first—this doesn't compromise the algorithm since we make no updates to the global parameters during testing, and since every testing batch represents a new sample from the corpus.

Experimental Results. Figure 11.3 contains the log likelihood results for the threefold cross validation testing. Each plot corresponds to a setting of K and $|C_t|$. From the plots, we can see how performance trends with these parameter settings. We first see that performance improves as the number of topics increases within our specified range. Also, we see that as the batch size increases, performance improves as well, but appears to reach a saturation point. At this point, increasing the batch size does not appear to significantly improve the direction of the stochastic gradient, meaning that the quality of the learned topics remains consistent over different batch sizes.

Performance is roughly the same for the three models considered. For online LDA and Bayesian NMF, this perhaps is not surprising given the similarity between the two models discussed in Section 11.3.1. Modeling topic correlations with correlated NMF does not appear to improve upon the performance of online LDA and Bayesian NMF. Nevertheless, correlated NMF does provide some additional tools for understanding the data.

In Figure 11.4 and Table 11.3, we show results for correlated NMF with $K = 150$ and $|C_t| = 1000$. In Table 11.3 we show the most probable words from the 40 most probable topics. In Figure 11.4 we show the correlations learned using the latent locations of the topics. The correlation between topic i and j is calculated as

$$\text{Corr}(\text{topic}_i, \text{topic}_j) = \ell_i^T \ell_j / \|\ell_i\|_2 \|\ell_j\|_j. \tag{11.36}$$

The learned correlations are meaningful. For example, two negatively correlated topics, topic 19 and topic 25, concern the legislative branch and football, respectively. On the other hand, topic 12, concerning baseball, correlates positively with topic 25 and negatively with topic 19. The ability to interpret topic meanings does not decrease as their probability decreases, as we show in Table 11.4.

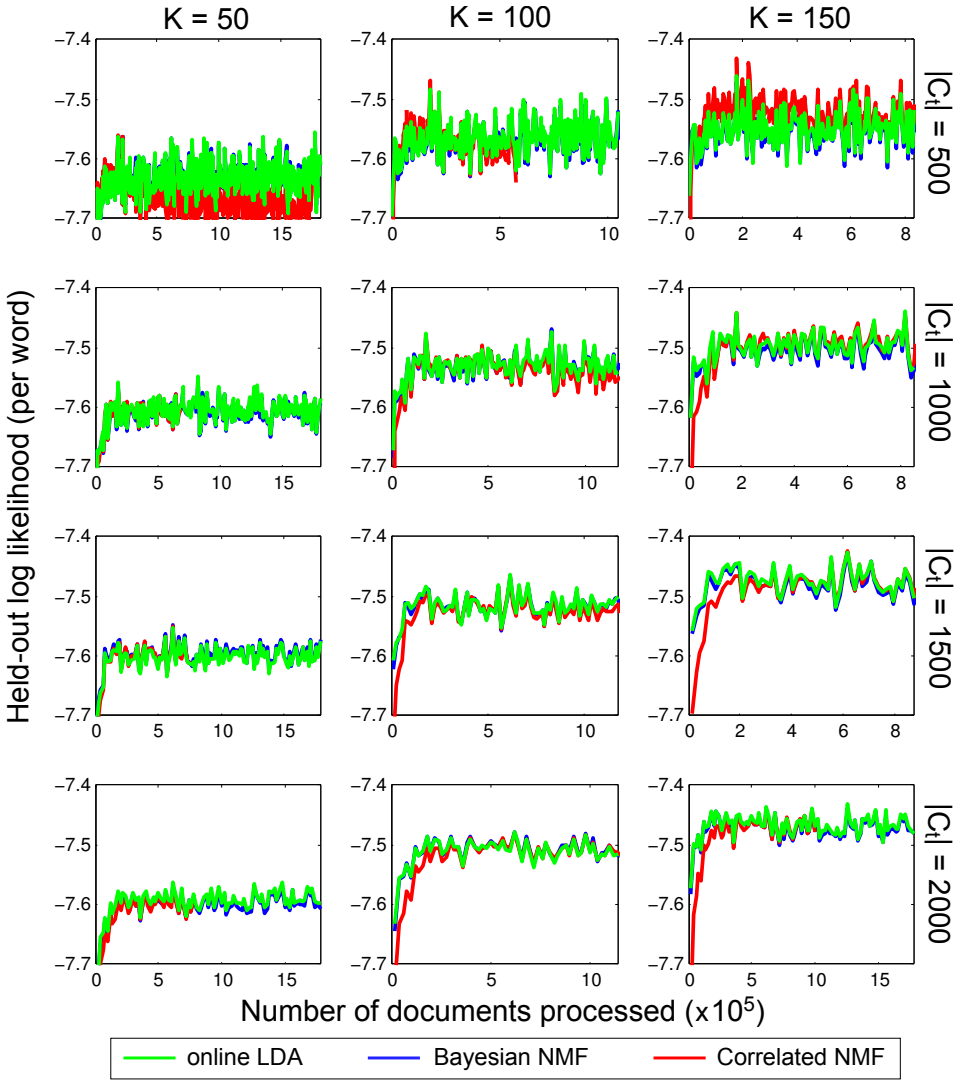

FIGURE 11.3
Performance results for Bayesian NMF, correlated NMF, and online LDA on the *New York Times* corpus. Results are shown for various topic settings and batch sizes. In general, performance is similar for all three models. Performance tends to improve as the number of topics increases. There appears to be a saturation level in batch size, that being the point where the increasing the number of documents does not significantly improve the stochastic gradient. Performance on this data set does not appear to improve significantly as $|C_t|$ increases over 1000 documents.

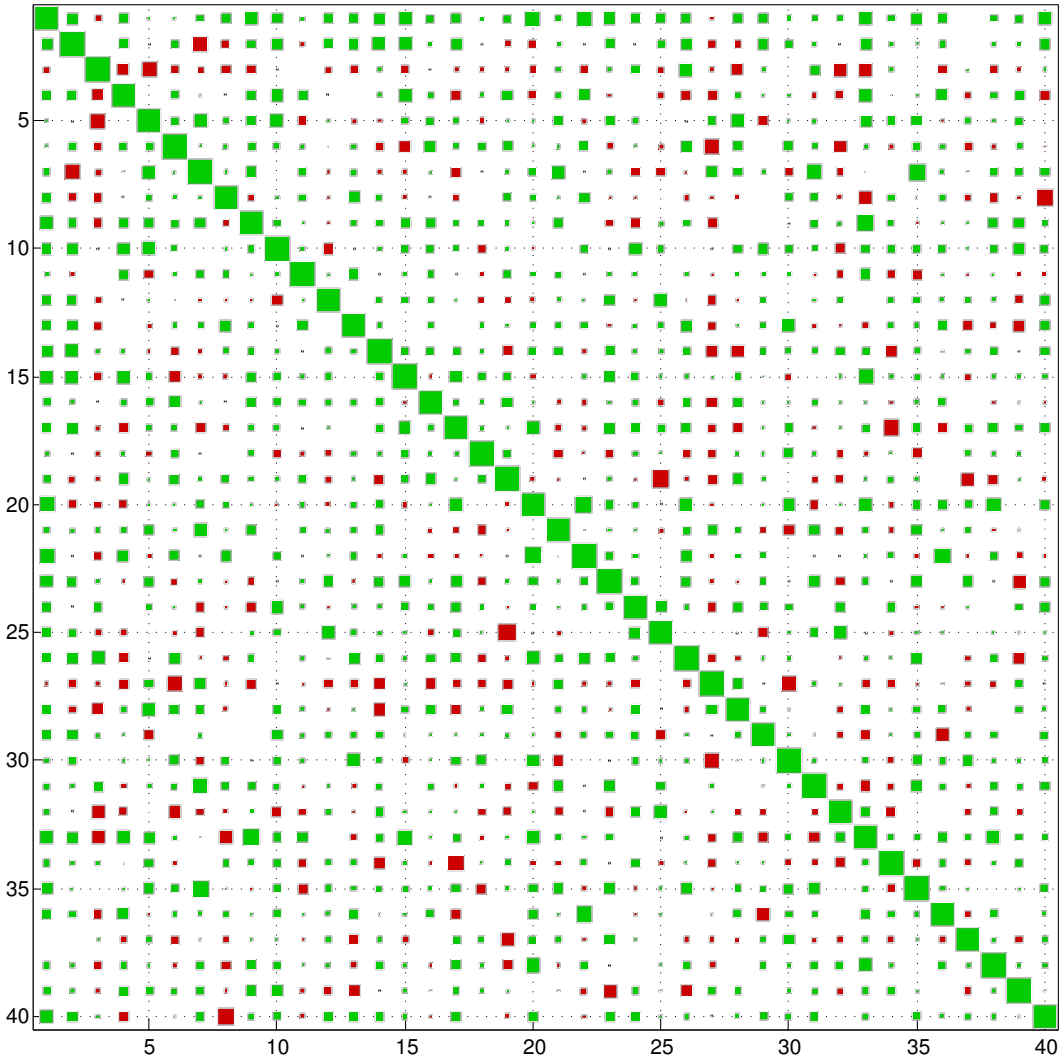

FIGURE 11.4
Correlations learned by correlated NMF for $K = 100$ and $|C_t| = 1000$. The figure contains correlations for the 40 most probable topics sorted by probability. Table 11.3 contains the most probable words associated with each topic in this figure. A green block indicates positive correlation, while a red block indicates negative correlations. The size of the block increases with an increasing correlation. The diagonal corresponds to perfect correlation, and is left in the figure for calibration.

Correlated NMF: Most probable words from the most probable topics
1. think, know, says, going, really, see, things, lot, got, didn
2. life, story, love, man, novel, self, young, stories, character, characters
3. wife, beloved, paid, notice, family, late, deaths, father, mother
4. policy, issue, debate, right, need, support, process, act, important
5. percent, prices, market, rate, economy, economic, dollar, growth, rose
6. government, minister, political, leaders, prime, officials, party, talks, foreign, economic
7. companies, billion, percent, corporation, stock, share, largest, shares, business, quarter
8. report, officials, department, agency, committee, commission, investigation, information, government
9. going, future, trying, likely, recent, ago, hopes, months, strategy, come
10. social, professor, society, culture, ideas, political, study, harvard, self
11. court, law, judge, legal, justice, case, supreme, lawyers, federal, filed
12. yankees, game, mets, baseball, season, run, games, hit, runs, series
13. trial, charges, case, prison, jury, prosecutors, federal, attorney, guilty
14. film, theater, movie, play, broadway, director, production, show, actor
15. best, need, better, course, easy, makes, means, takes, simple, free
16. party, election, campaign, democratic, voters, candidate, republican
17. town, place, small, local, visit, days, room, road, tour, trip
18. military, army, forces, troops, defense, air, soldiers, attacks, general
19. senate, bill, house, congress, committee, republicans, democrats
20. went, came, told, found, morning, away, saw, got, left, door
21. stock, investors, securities, funds, bonds, market, percent, exchange
22. asked, told, interview, wanted, added, felt, spoke, relationship, thought
23. restaurant, food, menu, cook, dinner, chicken, sauce, chef, dishes
24. school, students, education, college, teachers, public, campus
25. team, season, coach, players, football, giants, teams, league, game, bowl
26. family, father, mother, wife, son, husband, daughter, friends, life, friend
27. inc, net, share, reports, qtr, earns, sales, loss, corp, earnings
28. tax, budget, billion, spending, cuts, income, government, percent
29. art, museum, gallery, artists, show, exhibition, artist, works, paintings
30. police, officers, man, arrested, gun, shot, yesterday, charged, shooting
31. executive, chief, advertising, business, agency, marketing, chairman
32. game, points, knicks, basketball, team, nets, season, games, point, play
33. bad, far, little, hard, away, end, better, keep, break, worse
34. study, cancer, research, disease, tests, found, blood, test, cells
35. business, sold, market, buy, price, sell, selling, bought, sale, customers
36. public, questions, saying, response, criticism, attack, news, answer
37. street, park, avenue, west, east, side, village, neighborhood, central
38. feet, right, head, foot, left, side, body, eye, see, eyes
39. system, technology, research, program, industry, development, experts
40. music, band, songs, rock, jazz, song, pop, singer, album, concert

TABLE 11.3
Most probable words from the most probable topics for $K = 150,\ |C_t| = 1000$.

Correlated NMF: Most probable words from less probable topics	
41.	television, films, network, radio, cable, series, show, fox, nbc, cbs
44.	minutes, salt, add, oil, pepper, cup, heat, taste, butter, fresh
48.	computer, internet, technology, software, microsoft, computers, digital, electronic, companies, information
61.	john, thomas, smith, scott, michael, james, lewis, howard, kennedy
83.	iraq, iran, iraqi, hussein, war, saudi, gulf, saddam, baghdad, nations
84.	mayor, governor, giuliani, council, pataki, bloomberg, cuomo, assembly
98.	rangers, game, goal, devils, hockey, games, islanders, team, goals, season
111.	israel, israeli, palestinian, peace, arab, arafat, casino, bank, west
114.	china, chinese, india, korea, immigrants, immigration, asia, beijing
134.	british, london, england, royal, prince, sir, queen, princess, palace
141.	catholic, roman, irish, pope, ireland, bishop, priest, cardinal, john, paul

TABLE 11.4
Some additional topics not given in Table 11.3. Topics with less probability still capture coherent themes. Topics not shown were similarly coherent.

11.7 Conclusion

We have presented stochastic variational inference algorithms for two Bayesian nonnegative matrix factorization models: Bayesian NMF (Cemgil, 2009), a Bayesian extension of NMF (Lee and Seung, 1999); and correlated NMF, a new matrix factorization model that takes its motivation for the correlated topic model (Blei and Lafferty, 2007). Many other nonnegative matrix factorization models are candidates for stochastic inference, for example those based on Bayesian nonparametric priors such as the gamma process (Hoffman et al., 2010a) and the beta process Paisley et al. (2011).

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B* 44: 139–177.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation* 10: 251–276.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *Annals of Applied Statistics* 1: 17–35.

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

Bottou, L. (1998). *Online learning and stochastic approximations*. Cambridge, UK: Cambridge University Press.

Canny, J. (2004). GaP: A factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR '04)*. New York, NY, USA: ACM, 122–129.

Cemgil, A. (2009). Bayesian inference in non-negative matrix factorisation models. *Computational Intelligence and Neuroscience*. Article ID 785152.

Fei-Fei, L. and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the 10th IEEE Computer Vision and Pattern Recognition (CVPR 2005)*. San Diego, CA, USA: IEEE Computer Society, 524–531.

Gaussier, E. and Goutte, C. (2005). Relation between PLSA and NMF and implication. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. New York, NY, USA: ACM, 601–602.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring Expressed Agendas in Senate press releases. *Political Analysis* 18: 1–35.

Hoffman, M. D., Blei, D. M., and Bach, F. (2010a). Online learning for latent Dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A. (eds) *Advances in Neural Information Processing Systems 23*. Red Hook, NY: Curran Associates, Inc., 856–864.

Hoffman, M. D., Blei, D. M., and Cook, P. (2010b). Bayesian nonparametric matrix factorization for recorded music. In Fürnkranz, J., Joachims, T. (eds) *Proceedings of the 27th International Conference on Machine Learning (ICML '10)*. Omnipress, 439–446.

Jordan, M. I. (2011). Message from the president: The era of big data. *ISBA Bulletin* 18: 1–3.

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

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788–791.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Leen, T. K., Dietterich, T. G., and Tresp, V. (eds) *Advances in Neural Information Processing Systems 13*. Cambridge, MA: The MIT Press, 556–562.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* 11: 19–60.

Paisley, J., Carin, L., and Blei, D. (2011). Variational inference for stick-breaking beta process priors. In *Proceedings of the 28th International Conference on Machine Learning (ICML '11)*. Omnipress, 889–896.

Paisley, J., Wang, C., and Blei, D. M. (2012). The discrete logistic normal distribution. *Bayesian Analysis* 7: 235–272.

Sato, M. (2001). Online model selection based on the variational Bayes. *Neural Computation* 13: 1649–1681.

Singh, A. and Gordon, G. (2008). A unified view of matrix factorization models. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases – Part II (ECML/PKDD '08)*. Berlin, Heidelberg: Springer-Verlag, 358–373.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2007). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101: 1566–1581.

Wang, C., Paisley, J., and Blei, D. M. (2011). Online learning for the hierarchical Dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. Palo Alto, CA, USA: AAAI, 752–760.

Winn, J. and Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research* 6: 661–694.