LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-499647

Accelerated Gibbs Sampling for Infinite Sparse Factor Analysis

D. M. Andrzejewski

September 19, 2011

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Accelerated Gibbs Sampling for Infinite Sparse Factor Analysis

David Andrzejewski
andrzejewski1@llnl.gov

September 12, 2011

Abstract

The Indian Buffet Process (IBP) gives a probabilistic model of sparse binary matrices with an unbounded number of columns. This construct can be used, for example, to model a fixed numer of observed data points (rows) associated with an unknown number of latent features (columns). Markov Chain Monte Carlo (MCMC) methods are often used for IBP inference, and in this technical note, we provide a detailed review of the derivations of collapsed and accelerated Gibbs samplers for the linear-Gaussian infinite latent feature model. We also discuss and explain update equations for hyperparameter resampling in a "full Bayesian" treatment and present a novel slice sampler capable of extending the accelerated Gibbs sampler to the case of infinite sparse factor analysis by allowing the use of real-valued latent features.

1 Introduction

Say that we wish to model a set of observations by assuming that there are unobserved or latent features associated with each observation. For example, a gene may be associated with latent features describing which biological pathways it participates in. Observed levels of gene activity could then be modeled in terms of these latent features.

In the linear-Gaussian latent feature model, we assume that each observed vector $\mathbf{x} \in \mathbb{R}^D$ is associated with a binary latent feature vector $\mathbf{z} \in \{0,1\}^K$. Given a $K \times D$ matrix of feature weights $\mathbf{A}$, we say that $\mathbf{x}$ is distributed as a multivariate Gaussian random variable with mean $\mathbf{zA}$ and diagonal covariance $I\sigma_X^2$. For convenient reference, these matrices and their dimensionalities are summarized in Table 1.

Putting a zero-mean Gaussian prior on the weights $\mathbf{A}$, we then have the following generative model for a $N \times D$ matrix $\mathbf{X}$ representing N observations drawn i.i.d. from this model (temporarily treating the $N \times K$ latent feature matrix $\mathbf{Z}$ as given)

$$\begin{aligned}\mathbf{A} &\sim \mathcal{N}(0,\ I\sigma_A^2) \\ \mathbf{X} &\sim \mathcal{N}(\mathbf{ZA},\ I\sigma_X^2).\end{aligned}$$

Matrix	Dimensionality	Description
$\mathbf{X}$	$N \times D$	Observed $D-$dimensional features for items $1 \ldots N$
$\mathbf{Z}$	$N \times K$	Binary latent $K-$dimensional features for items $1 \ldots N$
$\mathbf{A}$	$K \times D$	Weight matrix mapping latent features to observations
$\mathbf{V}$	$N \times K$	Real-valued latent features (Infinite FA)

Table 1: Matrices of interest.

For a fixed, finite number of latent features K, it may be convenient to consider each element of a given $\mathbf{z}$ to be a Bernoulli random variable, with each dimension being associated with a Bernoulli parameter drawn from a conjugate Beta prior. However, the Indian Buffet Process (IBP) [3] lifts the restriction to finite K, allowing models with an unbounded number of latent features. The IBP can be expressed in terms of one or two hyperparameters; here we assume the one-parameter model $\mathbf{Z} \sim IBP(\alpha)$ for simplicity.

Bayesian inference under this model requires the use of approximate inference techniques such as Markov Chain Monte Carlo (MCMC) [3] or variational inference [1]. MCMC inference via *collapsed* Gibbs sampling (i.e., integrating out the weight matrix $\mathbf{A}$ in closed form) has been a successful approach, but suffers from scalability problems even after exploiting the Sherman-Morrison-Woodbury formula (also known as the matrix inversion lemma).

This computational bottleneck was eased in a scheme known as "accelerated Gibbs" [2] by maintaining the *posterior* over $\mathbf{A}$ instead of integrating out $\mathbf{A}$ entirely. This approach gets the best of both worlds, achieving computational complexity similar to an uncollapsed sampler and mixing similar to a collapsed sampler.

Infinite Factor Analysis (FA) [4] uses the IBP prior over binary matrices $\mathbf{Z}$, but then takes the entry-wise Hadamard product of $\mathbf{Z}$ with a real-valued weight matrix $\mathbf{V} \in \mathbb{R}^{N \times K}$. Unfortunately this extension "breaks compatibility" with the accelerated Gibbs sampler.

In this technical note, we review the derivations of the standard and accelerated Gibbs samplers for the standard IBP model, derive a slice sampling scheme for accelerated sampling under infinite FA, and present the hyperparameter sampling equations for the "full Bayesian" treatment of these models.

2 Collapsed Gibbs sampling

Given only the observed $\mathbf{X}$ and the model hyperparameters $\Theta = \{\alpha, \sigma_A^2, \sigma_X^2\}$, it is fairly straightforward to derive the Gibbs sampling equations for the binary latent elements of $\mathbf{Z}$ and the real-valued hidden weight matrix $\mathbf{A}$. However, this sampler may suffer from long mixing times [2], especially if dimensionality D is large.

One approach to this issue is to integrate out the hidden weights $\mathbf{A}$ in closed form, which is possible due to conjugacy between the Gaussian likelihood model and the Gaussian prior on the weight matrix $\mathbf{A}$. This reduces the state space of our sampler,

typically resulting in reduced convergence time. In this approach, we would sample $\mathbf{Z}$ from the collapsed posterior $P(\mathbf{Z}|\mathbf{X}, \Theta)$. From Bayes' rule we have

$$P(\mathbf{Z}|\mathbf{X}, \Theta) = \frac{P(\mathbf{Z}, \mathbf{X}|\Theta)}{\sum_{\mathbf{Z}} P(\mathbf{X}|\mathbf{Z}, \sigma_X^2, \sigma_A^2) P(\mathbf{Z}|\alpha)}. \tag{1}$$

When sampling an individual z_{ik}, the denominator will cancel out. We therefore restrict our attention to the numerator

$$P(\mathbf{Z}, \mathbf{X}|\Theta) = \int P(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \sigma_X^2) P(\mathbf{A}|\sigma_A^2) P(\mathbf{Z}|\alpha) d\mathbf{A}. \tag{2}$$

The IBP term $P(\mathbf{Z}|\alpha)$ is insensitive to $\mathbf{A}$ and can therefore be pulled out of the integral. Ignoring normalization, the remaining terms in the integral are given by

$$P(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \sigma_X^2) \propto \exp\left\{ \frac{-1}{2\sigma_X^2} \text{tr}[(\mathbf{X} - \mathbf{ZA})^T (\mathbf{X} - \mathbf{ZA})] \right\} \tag{3}$$

$$P(\mathbf{A}|\sigma_A^2) \propto \exp\left\{ \frac{-1}{2\sigma_A^2} \text{tr}[\mathbf{A}^T \mathbf{A}] \right\} \tag{4}$$

where the diagonal covariance matrices $I\sigma_X^2$ and $I\sigma_A^2$ resolve to simple scalar terms. Multiplying these expressions and simplifying recovers the following terms

$$\sigma_X^{-2} \mathbf{X}^T \mathbf{X} - \sigma_X^{-2} \mathbf{X}^T \mathbf{ZA} - \sigma_X^{-2} \mathbf{A}^T \mathbf{Z}^T \mathbf{X} + \sigma_X^{-2} \mathbf{A}^T \mathbf{Z}^T \mathbf{ZA} + \sigma_A^{-2} \mathbf{A}^T \mathbf{A} \tag{5}$$

wrapped inside a $\exp\left\{\text{tr}[\frac{-1}{2}(\cdot)]\right\}$ function. While this appears quite messy, we can appeal to the linear algebra version of a classic trick.

2.1 Completing the square

While following two expressions are equivalent

$$x^2 + bx \tag{6}$$

$$(x + \frac{b}{2})^2 - \frac{b^2}{4}, \tag{7}$$

$$\tag{8}$$

the lower expression conveniently isolates the variable x within a single quadratic term. The addition of the $\frac{-b^2}{4}$ is therefore known as "completing the square". This trick has an analog in linear algebra which applies when the quadratic coefficient matrix A is nonsingular and symmetric

$$X^T Q X + X^T L + L^T X \tag{9}$$

$$(X + Q^{-1}L)^T Q (X + Q^{-1}L) - (Q^{-1}L)^T L = 0. \tag{10}$$

We are going to integrate out $\mathbf{A}$ by recovering a quadratic form in terms of $\mathbf{A}$, so we apply this technique to $\mathbf{A}$. Note that the quadratic coefficient matrix of $\mathbf{A}$ is $(\sigma_X^{-2}\mathbf{Z}^T\mathbf{Z} + \sigma_A^{-2}I)$, which is indeed symmetric and invertible. Let $\mathbf{M} = (\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}I)^{-1}$. Then our quadratic and linear terms in $\mathbf{A}$ are

$$Q = (\sigma_X^2\mathbf{M})^{-1} \tag{11}$$

$$L = -\sigma_X^{-2}\mathbf{Z}^T\mathbf{X}. \tag{12}$$

$$\tag{13}$$

Applying this rearrangement to Equation 5 and re-organizing terms, we have the following expression

$$(\mathbf{M}\mathbf{Z}^T\mathbf{X} - \mathbf{A})^T(\sigma_X^2\mathbf{M})^{-1}(\mathbf{M}\mathbf{Z}^T\mathbf{X} - \mathbf{A}) + \sigma_X^{-2}(\mathbf{X}^T(I - \mathbf{Z}\mathbf{M}^T\mathbf{Z}^T) \tag{14}$$

which isolates $\mathbf{A}$ within the single quadratic term on the left above. Substituting the quadratic term *only* back into the $\exp$ of a trace function, we have a multivariate Gaussian distribution via the following substitutions

$$\mu = \mathbf{M}\mathbf{Z}^T\mathbf{X} \tag{15}$$

$$\Sigma = \sigma_X^2\mathbf{M}. \tag{16}$$

$$\tag{17}$$

Recall that Equation 14 is a rearranged version of the expression occurring within the integral $\int \exp\{-\frac{1}{2}\text{tr}[(\cdot)]\}$. Since Equation 14 has the form of a multivariate Gaussian with the mean and covariance shown above, this integral must evaluate to the normalization term of that Gaussian distribution, which is given by

$$\det[\sigma_X^2 M]^{D/2}(2\pi)^{KD/2}. \tag{18}$$

We can now express the likelihood of $\mathbf{X}$ (with $\mathbf{A}$ integrated out) as the product of this normalization term and the expression $\exp\{-\frac{1}{2}\text{tr}[(\cdot)]\}$ taken with respect to the $\mathbf{X}$ term not depending on $\mathbf{A}$ from Equation 14. Restoring the normalization terms from the original Equations 3 and 4, the full collapsed likelihood is

$$P(\mathbf{X}|\mathbf{Z}, \Theta) = \left[(2\pi)^{ND/2}\sigma_X^{(N-K)D}\sigma_A^{KD}\det[\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}I]^{D/2}\right]^{-1} \tag{19}$$

$$* \exp\left\{\frac{-1}{2\sigma_X^2}\text{tr}[\mathbf{X}^T(I - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2}I)^{-1}\mathbf{Z}^T)\mathbf{X}]\right\}. \tag{20}$$

We can now finally evaluate the collapsed Gibbs sampling equation for a single z_{ik} entry as

Gibbs Sampling step	Operation	Updated inversion	Easy update
before	remove $\mathbf{z}_i$	$(\mathbf{M}^{-1} - \mathbf{z}_i^T\mathbf{z}_i)^{-1}$	$\mathbf{M} - \frac{\mathbf{M}\mathbf{z}_i^T\mathbf{z}_i\mathbf{M}}{\mathbf{z}_i\mathbf{M}\mathbf{z}_i^T - 1}$
after	add $\mathbf{z}_i$	$(\mathbf{M}^{-1} + \mathbf{z}_i^T\mathbf{z}_i)^{-1}$	$\mathbf{M} - \frac{\mathbf{M}\mathbf{z}_i^T\mathbf{z}_i\mathbf{M}}{\mathbf{z}_i\mathbf{M}\mathbf{z}_i^T + 1}$

Table 2: Easy updates with matrix inversion lemma.

$$P(z_{ik} = 1|\mathbf{Z}_{-(i,k)}, \mathbf{X}, \Theta) = P(\mathbf{X}|z_{ik} = 1, \mathbf{Z}_{-(i,k)}, \sigma_X^2, \sigma_A^2)P(z_{ik} = 1|\mathbf{Z}_{-(i,k)}, \alpha) \tag{21}$$

where the first term can be calculated from Equation 19 and the second is simply the IBP posterior probability [3]

$$P(z_{ik} = 1|\mathbf{Z}_{-(i,k)}, \alpha) = \frac{m_{-i,k}}{N} \tag{22}$$

where m_{ik} is the number of latent feature vectors $\mathbf{z}_i$ where $z_{ik} = 1$, and $m_{-i,k}$ is the same count, omitting the latent vector $\mathbf{z}_i$.

2.2 Rank-one updates

One difficulty with this approach is the need to recompute $M = (\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2} * I)^{-1}$ with both $z_{ik} = 1$ and $z_{ik} = 0$ for every entry in $\mathbf{Z}$ during a single Gibbs "sweep". It is undesirable to have a computationally expensive operation like matrix inversion inside the innermost loop of our sampler.

However, note that the matrix $\mathbf{Z}$ can be expressed as a sum of outer products of each latent feature vector $\sum_i \mathbf{z}_i^T\mathbf{z}_i$. This means that calculating $\mathbf{Z}_{-i,k}$ by removing the influence of a single vector $\mathbf{z}_i$ is a *rank-one* update to $\mathbf{M}^{-1}$. Appealing to the *matrix inversion lemma*, we can easily add or remove the influence of a single $\mathbf{z}_i$ from $\mathbf{M}$ as shown in Table 2. It is advisable to periodically recomputing the full inverse in order to avoid the accumulation of numerical errors.

Unfortunately the $\mathbf{X}^T(\cdot)\mathbf{X}$ computation in $P(\mathbf{X}|\mathbf{Z}, \sigma_X^2, \sigma_A^2)$ is still expensive, and must be computed $2 \times N \times K$ times per full Gibbs sweep.

3 Sampling new features

The collapsed Gibbs sampling scheme discussed in Section 2 allows us to sample z_{ik}, the value of an existing feature k for a given observation i. However, a critical feature of the IBP prior [3] is that the number of latent features K is itself a random variable.

Let $s_i = |\{k|z_{ik} = 1 \text{and} \sum_{i'} z_{i',k} = 1\}|$ be the number of active "singleton" features for z_i. Under the IBP prior, this quantity is distributed according to $\text{POISSON}(\frac{\alpha}{N})$. Given the observed $\mathbf{X}$ and latent $\mathbf{Z}$, we can therefore calculate the posterior over s_i as

$$P(s_i|\mathbf{X}, \mathbf{Z}, \alpha, \sigma_X^2, \sigma_A^2) \propto \text{POISSON}\left(s_i|\frac{\alpha}{N}P(\mathbf{X}|s_i, \mathbf{Z}, \sigma_X^2, \sigma_A^2\right). \tag{23}$$

This observation allows the use of a Metropolis-Hastings [6] sampling scheme where we propose a new number of singleton features $\tilde{s}_i$. If we set our proposal distribution $Q(\tilde{s}_i|s_i) = \text{POISSON}(\frac{\tilde{s}_i}{N})$, the Poisson distributions will cancel out of the acceptance probability

$$\min\left(\frac{P(\tilde{s}_i|\mathbf{X},\mathbf{Z},\Theta)Q(s_i|\tilde{s}_i)}{P(s_i|\mathbf{X},\mathbf{Z},\Theta)Q(\tilde{s}_i|s_i),1}\right) \tag{24}$$

leaving the simple acceptance probability

$$\min\left(\frac{P(\mathbf{X}|\tilde{s}_i,\mathbf{Z},\sigma_X^2,\sigma_A^2)}{P(\mathbf{X}|s_i,\mathbf{Z},\sigma_X^2,\sigma_A^2)},1\right). \tag{25}$$

If we accept a proposal where $\tilde{s}_i < s_i$, we can simply delete any $s_i - \tilde{s}_i$ columns of $\mathbf{Z}$ where $z_{ik} = 1$ and $\sum_{i'} z_{i',k} = 1$. Since these latent features are only active within z_i, it does not matter which of them are deleted. Likewise if we accept a proposal where $\tilde{s}_i > s_i$, we simply add new columns to $\mathbf{Z}$ where only $z_{ik} = 1$.

4 Sampling hyperparameters

Our model has three hyperparameters $\{\alpha, \sigma_A^2, \sigma_X^2\}$, which we collectively refer to as Θ. These values can be set manually in order to yield a model with specific properties (e.g., the level of latent feature sparsity). Another approach is to treat these quantities as random variables themselves, endowing them with hyperpriors. We can once more make use of the mathematical convenience of conjugacy with the following priors

$$\alpha \sim \text{GAMMA}(a_\alpha, b_\alpha) \tag{26}$$
$$\tau_X \sim \text{GAMMA}(a_X, b_X) \tag{27}$$
$$\tau_A \sim \text{GAMMA}(a_A, b_A) \tag{28}$$
$$\tag{29}$$

where τ_X is the *precision*, equal to the inverse variance σ_X^{-2}, and likewise for τ_A and σ_A^{-2}. We now step through the sampling updates for each of these hyperparameters.

4.1 Weight variance σ_A^2

In general form, the posterior update equations for a Gamma prior on Gaussian precision after n i.i.d. observations x_i are given by

$$\tilde{a} = a + \frac{n}{2} \tag{30}$$
$$\tilde{b} = b + \frac{1}{2}\sum_i^n (x_i - \mu)^2. \tag{31}$$

For the weights $\mathbf{A}$, we have $n = KD$ observations with mean $\mu = 0$ and variance σ_A^2. Unfortunately the weights $\mathbf{A}$ are unobserved, but fortunately the favorable properties of the exponential family allow us to substitute the expected sufficient statistics under the posterior $P(\mathbf{A}|\mathbf{X}, \mathbf{Z}, \Theta)$. Letting E_A be the expectation over this distribution, we have

$$\begin{aligned}
E_A\left[\sum_{k,d}(A_{kd}-0)^2\right] &= \sum_{k,d}\left(E_A\left[(A_{kd}-0)^2\right]\right) \\
&= \sum_{k,d}\left(E_A\left[((A_{kd}-\mu_{kd})-(0-\mu_{kd}))^2\right]\right) \\
&= \sum_{k,d}(E_A\left[(A_{kd}-\mu_{kd})^2\right] \\
&\quad - 2E_A\left[(A-\mu_{kd})(0-\mu_{kd})\right] \\
&\quad + E_A\left[(0-\mu_{kd})^2\right]) \\
&= \sum_{k,d}\left(\sigma_{kd}-0+\mu_{kd}^2\right)
\end{aligned}$$

where μ_{kd} and σ_{kd}^2 are the mean and variance of the element A_{kd} under the Gaussian posterior over $\mathbf{A}$.

Sampling a new value of σ_A^2 then consists of taking the inverse of a sample from $\textsc{Gamma}(a_A + \frac{N}{2}, b_A + \frac{1}{2}(\sum_{k,d}(\sigma_{kd} + \mu_{kd}^2)$.

4.2 Noise variance σ_X^2

Each data point $\mathbf{x}_i$ is distributed according to a multivariate Gaussian distribution with mean $A^T\mathbf{z}_i$ and covariance $\sigma_X^2 I$. Again, it is computationally convenient to place a $\textsc{Gamma}(a, b)$ prior on σ_X^2. Letting A_{*d} be all rows of the d^{th} column of A, again we can substitute the quantity $\sum_i \sum_d (\mathbf{x}_{id} - \mathbf{z}_i^T A_{*d})^2$ with its expectation with respect to $\mathbf{A}$. This expectation can be easily computed after a few manipulations.

$$
\begin{aligned}
E_A\left[\sum_{i,d}(\mathbf{x}_{id}-\mathbf{z}_i^T A_{*d})^2\right] &= \sum_{i,d} E_A\left[(\mathbf{x}_{id}-\mathbf{z}_i^T A_{*d})^2\right] \\
&= \sum_{i,d} E_A\left[(\mathbf{x}_{id}-\mathbf{z}_i^T A_{*d})^2\right] \\
&= \sum_{i,d} E_A\left[((\mathbf{x}_{id}-\mathbf{z}_i^T \mu_{A_{*d}})-(\mathbf{z}_i^T A_{*d}-\mathbf{z}_i^T \mu_{A_{*d}}))^2\right] \\
&= \sum_{i,d}(x_{id}-\mathbf{z}_i^T A_{*d})^2 - 2E_A\left[(\mathbf{x}_{id}-\mathbf{z}_i^T \mu_{A_{*d}})(\mathbf{z}_i^T A_{*d}-\mathbf{z}_i^T \mu_{A_{*d}})\right] \\
&+ \mathbf{z}_i^T \Sigma_A \mathbf{z}_i \\
&= \sum_{i,d}\left[(x_{id}-\mathbf{z}_i^T A_{*d})^2 + \mathbf{z}_i^T \Sigma_A \mathbf{z}_i\right]
\end{aligned}
$$

As before, we use this value to update the Gamma parameter $\tilde{b} = b + \frac{1}{2}E_A\left[\sum_{i,d}(\mathbf{x}_{id}-\mathbf{z}_i^T A_{*d})^2\right]$, and we update the other parameter as $\tilde{a} = a + \frac{ND}{2}$.

4.3 IBP parameter α

The Poisson distribution is also conjugate to the Gamma prior. After n i.i.d. Poisson observations x_i the the Gamma posterior update equations are

$$\tilde{a} = a + \sigma_i x_i \tag{32}$$

$$\tilde{b} = b + n. \tag{33}$$

Under the IBP, the quantity $\mathbb{1}^T\mathbf{z}_i$ is distributed according to $\text{POISSON}(\alpha)$ for each latent vector $\mathbf{z}$ [3]. We have $n = N$ observations and we can then drop the sum of all nonzero latent features $\mathbb{1}^T\mathbf{Z}\mathbb{1}$ into the $\tilde{a}$ update.

5 Accelerated Gibbs sampling

Thus far we have focused on the collapsed Gibbs sampler. The primarily computational bottleneck of the fully collapsed approach is that the marginalization of $\mathbf{A}$ induces dependencies among all observations $\mathbf{X}$. Even using the rank-one updates instead of full matrix inversions, there is still an expensive matrix multiplication in the exponent of the collapsed likelihood computation (Equation 20). This multiplication requires $O(DN^2)$ operations and must be performed twice for each of the NK elements of $\mathbf{Z}$ resulting in $O(DKN^3)$ complexity for a single collapsed Gibbs sampling sweep.

At the other extreme, the individual observations of $\mathbf{X}$ can be rendered completely conditionally independent given $\mathbf{A}$, $\mathbf{Z}$, and Θ in the uncollapsed Gibbs sampler. However the resulting Markov chain will have a larger state space and therefore longer convergence time.

Accelerated Gibbs Sampling [2] cuts a middle path between these approaches by maintaining the *posterior* over $\mathbf{A}$, $P(\mathbf{A}|\mathbf{Z}, \mathbf{X}, \Theta)$. Individual observations are then coupled only through this posterior, simplifying our calculations without blowing up the state space. Gibbs sampling of $\mathbf{Z}$ will consist of *removing* the influence of a single observation $\mathbf{x}_i$ from this posterior, sampling $\mathbf{z}_i$, and then reincorporating the observation into $P(\mathbf{A}|\mathbf{Z}, \mathbf{X}, \Theta)$. Gaussian conjugacy allows us to easily update this posterior.

5.1 The posterior over $\mathbf{A}$

In order to compute the posterior over $\mathbf{A}$ we return to the algebraic rearrangements we performed when deriving the collapsed Gibbs sampler in Section 2. In that derivation, our goal was to perform the integration

$$P(\mathbf{Z}, \mathbf{X}|\Theta) = \int P(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \sigma_X^2) P(\mathbf{A}|\sigma_A^2) P(\mathbf{Z}|\alpha) d\mathbf{A}, \tag{34}$$

by rearranging the integrand into a Gaussian form in $\mathbf{A}$, such that the integration itself had a closed-form solution (the normalization of that Gaussian form). Revisiting the Gaussian form with parameters given by Equations 15 and 16, we can easily express the posterior over $\mathbf{A}_{*d}$, the latent feature weights for observed feature d

$$P(\mathbf{A}_{*d}|\mathbf{X}, \mathbf{Z}, \Theta) \propto \exp\{-\frac{1}{2}(\mathbf{M}\mathbf{Z}^T\mathbf{X}_{*d} - \mathbf{A}_{*d})^T(\sigma_X^2\mathbf{M})^{-1}(\mathbf{M}\mathbf{Z}^T\mathbf{X}_{*d} - \mathbf{A}_{*d})\}. \tag{35}$$

Expanding our convenience variable $\mathbf{M}$, we can express the parameters of this multivariate Gaussian as

$$\mu_{\mathbf{A}_{*d}} = (\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2} I)^{-1}\mathbf{Z}^T\mathbf{X}_{*d} \tag{36}$$

$$\Sigma_A = \sigma_X^2(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma_X^2}{\sigma_A^2} I)^{-1}. \tag{37}$$

5.2 Sampling z_{ik} using the posterior over $\mathbf{A}$

We will now see how the use of this posterior can simplify our calculations by revisiting the Gibbs sampling equation

$$P(z_{ik} = 1|\mathbf{Z}_{-(i,k)}, \mathbf{X}, \Theta) \propto P(z_{ik}|\mathbf{Z}_{-(i,k)}, \alpha) \int P(\mathbf{X}|\mathbf{Z}, \mathbf{A}, \sigma_X) P(\mathbf{A}|\sigma_A) d\mathbf{A}. \tag{38}$$

$$\tag{39}$$

We can break the integrand likelihood term into two parts: one for the i^{th} observation, and one for everything else

$$\int P(\mathbf{x}_i|\mathbf{z}_i, \mathbf{A}, \sigma_X)P(\mathbf{X}_{-i}|\mathbf{Z}_{-i}, \mathbf{A}, \sigma_X^2)P(\mathbf{A}|\sigma_A)d\mathbf{A}. \tag{40}$$

$$\tag{41}$$

We then apply Bayes Rule, noting that the right-hand terms are *proportional* to the posterior over $\mathbf{A}$ given all of the data *except* i^{th} point

$$\int P(\mathbf{x}_i|\mathbf{z}_i, \mathbf{A}, \sigma_X)P(\mathbf{A}|\mathbf{X}_{-i}, \mathbf{Z}_{-i}, \mathbf{A}, \sigma_A^2, \sigma_X^2)d\mathbf{A}. \tag{42}$$

$$\tag{43}$$

The integral now represents the expectation of $P(\mathbf{x}_i|\mathbf{z}_i, \mathbf{A}, \sigma_X)$ (a linear-Gaussian distribution) taken over the (Gaussian) posterior of $\mathbf{A}$ given all other data points. This allows us to once again evaluate the integral in closed form, yielding the conclusion that $\mathbf{x}_i$ is multivariate Gaussian with parameters

$$\mu_{x_i} = \mathbf{z}_i^T \mu_A \tag{44}$$

$$\Sigma_{x_i} = \mathbf{z}_i^T \Sigma_A \mathbf{z}_i + \Sigma_X. \tag{45}$$

$$\tag{46}$$

We now have an easy way to compute the $\mathbf{x}_i$ likelihood term for purposes of Gibbs sampling each element of $\mathbf{z}_i$. The only additional complication is adding and removing the influence of the i^{th} data point over the posterior of $\mathbf{A}$. For numerical stability reasons, we compute these on the *information form* representation of the $\mathbf{A}$ posterior

$$P_A = \Sigma_A^{-1} = \sigma_X^- 2\mathbf{Z}^T\mathbf{Z} + \sigma_A^- 2I \tag{47}$$

$$h_A = P_A\mu_A = \sigma_X^- 2\mathbf{Z}^T\mathbf{X}, \tag{48}$$

$$\tag{49}$$

in which case the updates to add or remove the influence of the i^{th} data point consist of

$$P_A \leftarrow P_A \pm \sigma_X^{-2}\mathbf{z}_i^T\mathbf{z}_i \tag{50}$$

$$h_A \leftarrow h_A \pm \mathbf{z}_i^T\mathbf{x}_i\sigma_X^{-2}. \tag{51}$$

We can now bring all the pieces together - the pseudocode for doing T full sweeps of accelerated Gibbs sampling with hyperparameter resampling is given in Algorithm 1.

Algorithm 1: Pseudocode for accelerated Gibbs sampling.

```
for t = 1, ..., T do
    for i = 1, ..., N do
        (P_A, h_A) ← remove influence of x_i
        for k = 1, ..., K do
            Sample z_ik ∝ P(z_ik | Z_{-ik}, α) P(x_i | z_i, μ_{x_i} Σ_{x_i})
        end
        (P_A, h_A) ← restore influence of x_i
        Metropolis sample new features
    end
    Resample hyperparameters {σ_X, σ_A, α}
end
```

6 Real-valued latent features

Thus far we have considered the following generative model

$$\begin{aligned}
\mathbf{ZZ} &\sim IBP(\alpha) \\
\mathbf{A} &\sim \mathcal{N}(0,\ I\sigma_A^2) \\
\mathbf{X} &\sim \mathcal{N}(\mathbf{ZA},\ I\sigma_X^2).
\end{aligned}$$

However, the restriction to *binary* latent features may be restrictive. We can loosen this restriction by allowing real-valued latent features in the following model

$$\begin{aligned}
\mathbf{Z} &\sim IBP(\alpha) \\
\mathbf{V} &\sim \mathcal{N}(0,\ I) \\
\mathbf{A} &\sim \mathcal{N}(0,\ I\sigma_A^2) \\
\mathbf{X} &\sim \mathcal{N}((\mathbf{Z} \circ \mathbf{V})\mathbf{A},\ I\sigma_X^2)
\end{aligned}$$

where $\circ$ denotes the element-wise (or Hadamard) matrix product. Here z_{ik} retains its previous role of determining whether or not a latent feature k is "active" for a given example i, and v_{ik} now specifies the "strength" of the feature within that example. This formulation is known as infinite sparse factor analysis [4], and extends traditional sparse factor analysis by explicitly modeling uncertainty with respect to the number of latent factors.

The inclusion of $\mathbf{V}$ requires modification to previously developed inference schemes. One approach would be to simply return to the uncollapsed Gibbs sampler, i.e., maintain the values of $\mathbf{A}$ in our sampler state. This results in relatively straightforward sampling equations, but has the disadvantage of increasing the state space and slowing convergence, as before.

This leads us to the development of collapsed sampling schemes for real-valued latent features. Given $\mathbf{V}$, we could simply use $\tilde{\mathbf{Z}} = \mathbf{Z} \circ \mathbf{V}$ as a drop-in replacement for

$\mathbf{Z}$ in the equations of the collapsed Gibbs sampler. We would also need to sample the $\mathbf{V}$ values themselves[1]. However, this approach would suffer from the same computational challenges as the collapsed sampler for the binary latent feature model.

6.1 Slice sampling V

We would therefore like to adapt the accelerated scheme to real-valued latent features. As in the collapsed case, we can simply drop $\tilde{\mathbf{Z}}$ into the calculations for the posterior of $\mathbf{A}$ and the likelihood of $\mathbf{X}$. However, it is not immediately clear how to efficiently sample $\mathbf{V}$ using the posterior of $\mathbf{A}$.

Assuming $\mathbf{z}_{ik} = 1$, then we want to sample a value of $\mathbf{v}_{ik}$ from the posterior

$$P(\mathbf{v}_{ik}|\mathbf{v}_{-ik}, \mathbf{Z}, \mathbf{X}, \Theta) \propto \int P(\mathbf{x}_i|\mathbf{z}_i, \mathbf{v}_i, \mathbf{A}, \sigma_X) P(\mathbf{A}|\mathbf{X}_{-i}, \mathbf{Z}_{-i}, \mathbf{V}_{-i}, \sigma_A), P(\mathbf{V}_i) d\mathbf{A} \tag{52}$$

which reduces to the following equation after discarding terms not dependent on $\mathbf{v}_{ik}$

$$P(\mathbf{v}_{ik}|\mathbf{v}_{-ik}, \mathbf{Z}, \mathbf{X}, \Theta) \propto \mathcal{N}(\tilde{\mathbf{z}_i}^T \mu_A, \tilde{\mathbf{z}_i}^T \Sigma_A \tilde{\mathbf{z}_i} + \Sigma_X) P(\mathbf{v}_{ik}) \tag{53}$$

where (μ_A, Σ_A) are the parameters from the posterior over $\mathbf{A}$, as described in Equations 44.

This distribution is an unnormalized product of Gaussians, and therefore not amenable to direct sampling. However, we can use a technique known as *slice sampling* [5] to sample from this distribution. Slice sampling is a form of *rejection sampling* [5] that decomposes the problem of sampling from an arbitrary unnormalized distribution $\tilde{p}(y)$ into two discrete steps, given a current value y and window boundaries (L, R) such that $L < y$ and $R > y$

1. sample u uniformly from $[0, \tilde{p}(y)]$
2. sample $\hat{y}$ uniformly from $[y - L, y + L]$
3. if $\tilde{p}(\hat{y}) \geq u$, accept new $\hat{y}$ value (else reject).

This procedure defines a Markov chain whose stationary distribution consists of points $(y, \tilde{p}(y))$ uniformly sampled from the hypograph of $\tilde{p}(y)$. The y values can then be considered to be samples from the distribution corresponding to the normalization of $\tilde{p}(y)$. We can apply this idea to sample $\mathbf{v}_{ik}$ by treating Equation 53 as the unnormalized distribution.

A further complication is raised because of a common trick in statistical computing, where the idea is to avoid numeric underflow by working in the log-domain. Here this means that we modify Equation 53 to compute $\log(P(\mathbf{v}_{ik}|\ldots)$ instead of $P(\mathbf{v}_{ik}|\ldots)$. Letting $y = P(\mathbf{v}_{ik}|\ldots)$, this means Step 1 in our slice sampler must now sample $\ell = \log(u)$ such that u is drawn uniformly from $[0, \exp(\log \tilde{p}(y))]$.

A bit of manipulation shows how we can do this without leaving the log-domain. First, we define our log-transformation $\ell = f(u) = \log u$ and its inverse $u = f^{-1}(\ell) =$

[1] would this be easy, or not?

$\exp \ell$. Since $u \in [0, \tilde{p}(y)]$ and $\tilde{p}(y) \geq 0$, these functions are well-defined. Furthermore, note that $f^{-1}(\ell)$ is strictly monotonically increasing in ℓ. Given our definition of the log-variable ℓ and the uniform distribution over u, we can derive the inverse cumulative distribution function of ℓ with respect to a uniform random variable a as

$$
\begin{aligned}
pdf_U(u') &= \exp(-\log(\tilde{p}(y))) && (54)\\
cdf_U(u) &= P(U \leq u) && (55)\\
&= \int_0^u pdf_U(u')du' && (56)\\
cdf_L(\ell) &= P(L \leq \ell) && (57)\\
&= P(f^{-1}(L) \leq f^{-1}(\ell)) && (58)\\
&= \int_{f^{-1}(-\infty)}^{f^{-1}(\ell)} pdf_U(u')du' && (59)\\
&= \int_{f^{-1}(-\infty)}^{f^{-1}(\ell)} \exp(-\log(\tilde{p}(y)))du' && (60)\\
&= \exp(-\log(\tilde{p}(y))) \int_{f^{-1}(-\infty)}^{f^{-1}(\ell)} du' && (61)\\
&= \exp(-\log(\tilde{p}(y)))(u'|_{f^{-1}(-\infty)}^{f^{-1}(\ell)} && (62)\\
&= \exp(-\log(\tilde{p}(y)))(f^{-1}(\ell) - 0) && (63)\\
&= \exp(-\log(\tilde{p}(y)))\exp(\ell) && (64)\\
&= \exp(\ell - \log(\tilde{p}(y)))cdf_L^{-1}(a) \qquad = \log(a) + \log(\tilde{p}(y)). && (65)\\
& && (66)
\end{aligned}
$$

We can therefore sample ℓ by drawing a uniformly from $[0, 1]$ and then computing $\ell = cdf_L^{-1}(a) = \log(a) + \log(\tilde{p}(y))$. Given this ℓ value, we can now either accept or reject a newly sampled candidate value of $\mathbf{v}_{ik}$ based on the log of the unnormalized probability in Equation 53.

6.2 Sampling $\mathbf{z}_{ik}$

We notice another complication raised by the use of real-valued latent feature weights $\mathbf{V}$ when we attempt to sample $\mathbf{z}_{ik}$. Previously in accelerated Gibbs sampling, we had to evaluate the integral

$$\int P(\mathbf{x}_i|\mathbf{z}_i, \mathbf{A}, \sigma_X)P(\mathbf{A}|\mathbf{X}_{-i}, \mathbf{Z}_{-i}, \mathbf{A}, \sigma_A^2, \sigma_X^2)d\mathbf{A}. \tag{67}$$

The addition of real-valued latent features forces us to either maintain "phantom" $\mathbf{v}_{ik}$ values connected to currently inactive binary latent features $\mathbf{z}_{ik}$, or to integrate with respect to the currently inactive $\mathbf{v}_{ik}$

$$\int \left(\int P(\mathbf{x}_i|\mathbf{z}_i, \mathbf{v}_i, \mathbf{A}, \sigma_X) P(\mathbf{A}|\mathbf{X}_{-i}, \mathbf{Z}_{-i}, \mathbf{A}, \sigma_A^2, \sigma_X^2) d\mathbf{A} \right) P(\mathbf{v}_{ik}) d\mathbf{v}_{ik}. \quad (68)$$

In the "phantom" approach, we could simply re-sample $\mathbf{v}_{ik}$ from the Gaussian prior $P(\mathbf{v}_{ik})$ each iteration. Alternatively, we could numerically evaluate the Monte Carlo estimate of the integral in Equation 68 by taking the average value of the inner expression over multiple $\mathbf{v}_{ik}$ samples from $P(\mathbf{v}_{ik})$.

6.3 New features via Metropolis-Hastings

Finally, we examine how the sampling of totally new latent features is affected by real-valued latent feature weights. Recall that our Metropolis step for the binary latent feature model proposes a value for k_{new}, the number of "singleton" latent features for the current example. We can extend this proposal distribution by simply jointly proposing real-valued latent values $\mathbf{v}_{ik}$ for each of these singleton features, drawn from the Gaussian prior $P(\mathbf{v}_{ik})$. The only change to the likelihood term is that Σ_{x_i} is increased by $\sum_{k'=k+1}^{k+k_{new}} \mathbf{v}_{ik}^2$. This result is intuitive: we are simply considering a latent model where we are adding a $\mathbf{v}$-weighted sum of i.i.d. Gaussian $\mathbf{A}$ values, increasing the variance Σ_X. Other than this modification, the Metropolis-Hastings procedure is unchanged.

6.4 Practical issues

Initialization of this model can be very important to good performance. A common option is to run a *parametric* version of the model with some fixed K and using that state to initialize the infinite model. Another strategy is to run inference with *fixed* hyperparameters for some number of samples before incorporating hyperparameter re-sampling.

As the MCMC chain runs, it may be useful to hard-code "guards" on hyperparameter values - for example, we may wish to enforce that α should not grow beyond a certain value. Finally, it can be extremely helpful to monitor predictive log-likelihood of both in-sample and out-of-sample observations.

7 Acknowledgments

I would like to thank Finale Doshi-Velez for making MATLAB code available and for answering detailed questions about inference. The idea for the slice-sampling scheme was suggested by David Knowles.

References

[1] F. Doshi, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian Buffet Process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, volume 12, 2009.

[2] Finale Doshi-Velez and Zoubin Ghahramani. Accelerated sampling for the Indian Buffet Process. In *ICML*, page 35, 2009.

[3] Tom Griffiths and Zoubin Ghahramani. Infinite latent feature models and the Indian Buffet Process. In *NIPS*, 2005.

[4] David Knowles and Zoubin Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In Mike Davies, Christopher James, Samer Abdallah, and Mark Plumbley, editors, *Independent Component Analysis and Signal Separation*, volume 4666 of *Lecture Notes in Computer Science*, pages 381–388. Springer Berlin / Heidelberg, 2007.

[5] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[6] Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In *NIPS*, pages 977–984, 2006.