An Empirical Study of Stochastic Variational Algorithms for the Beta Bernoulli Process

Amar Shah ⋆ AS793@CAM.AC.UK
David A. Knowles † DAVIDKNOWLES@CS.STANFORD.EDU
Zoubin Ghahramani ⋆ ZOUBIN@ENG.CAM.AC.UK
⋆ University of Cambridge, Department of Engineering, Cambridge, UK
† Stanford University, Department of Computer Science, Stanford, CA, USA

Abstract

Stochastic variational inference (SVI) is emerging as the most promising candidate for scaling inference in Bayesian probabilistic models to large datasets. However, the performance of these methods has been assessed primarily in the context of Bayesian topic models, particularly latent Dirichlet allocation (LDA). Deriving several new algorithms, and using synthetic, image and genomic datasets, we investigate whether the understanding gleaned from LDA applies in the setting of sparse latent factor models, specifically beta process factor analysis (BPFA). We demonstrate that the big picture is consistent: using Gibbs sampling within SVI to maintain certain posterior dependencies is extremely effective. However, we find that different posterior dependencies are important in BPFA relative to LDA. Particularly, approximations able to model intra-local variable dependence perform best.

1. Introduction

The last two decades have seen an explosion in the development of flexible statistical methods able to model diverse data sources. Bayesian nonparametric priors in particular provide a powerful framework to enable models to adapt their complexity to the data at hand (Orbanz & Teh, 2010). In the regression setting this might mean learning the smoothness of the output function (Rasmussen & Williams, 2006), for clustering adapting the number of components (MacEachern & Müller, 1998), and in the case of our interest, latent factor models, finding an appropriate number of latent features (Knowles et al., 2011). While such models are appealing for a range of applied data analysis applications, their scalability is often limited. The posterior distribution over parameters and latent variables is typically analytically intractable and highly multimodal, making MCMC, particularly Gibbs sampling, the norm. Along with concerns over performance and convergence, MCMC methods are often impractical for the applied practitioner: how should the multiple samples be summarized? Variational methods work on the basis that simply finding a good posterior mode, and giving some measure of the associated uncertainty, is typically sufficient. In addition, the predictive performance of variational methods is often comparable to more computationally expensive sampling based approaches (Ghahramani & Beal, 1999).

Recently *stochastic* variational inference has begun to emerge as the most promising avenue for scaling inference in large latent variable models (Hoffman et al., 2013). Marrying variational inference with stochastic gradient descent allows principled updates using only minibatches of observations, greatly improving data scalability. While some MCMC methods have been proposed to work with minibatches (Welling & Teh, 2011; Ahn et al., 2012) they lack theoretical guarantees and apply only to continuous, unbounded latent variables. While SVI has been influential for Bayesian topic modeling, particularly latent Dirichlet allocation (LDA, Mimno et al., 2012; Hoffman & Blei, 2014; Wang & Blei, 2012), the same cannot be said for sparse factor analysis models for continuous data. While the former has been driven by the ready availability of huge text corpora, the scale of continuous data being generated by new genomic technologies is still growing. For example, CyTOF, single cell time of flight mass spectrometry is able to measure the abundance of dozens of proteins in hundreds of thousands of cells in a single run (Bendall et al., 2011). Such complex, large scale, high dimensional datasets require sophisticated statistical models, but are typically analyzed using simple heuristic clustering methods or PCA, which do not capture important structure, such as sparsity. As a result, scaling more advanced factor analysis type models is of great interest.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

When desigining a variational approximation to a posterior distribution, one must trade off the accuracy of the approximation with the complexity of optimizing the evidence lower bound. Mean field approximations are simple to work with, but Hoffman & Blei (2014) demonstrated that in the context of LDA, maintaining posterior dependence between "global" variables (topic vectors) and "local" variables (document vectors) is crucial to finding good solutions. Does this finding hold for sparse factor analysis models? Our results suggest that contrary to the LDA case, maintaining dependencies *amongst local variables* is actually the most important ingredient for obtaining good performance with beta Bernoulli process SVI.

2. Beta Process for Factor Analysis

The *beta process* (Hjort, 1990; Thibaux & Jordan, 2007) is an independent increments process defined as follows:

Definition 1. *Let Ω be a measurable space and $\mathcal{B}$ its σ-algebra. Let H_0 be a continuous probability measure on $(\Omega, \mathcal{B})$ and α a positive scalar. Then for all disjoint, infinitesimal partitions, $\{B_1, ..., B_K\}$, of Ω the beta process is generated as follows,*

$$H(B_k) \overset{\text{iid}}{\sim} \text{Beta}(\alpha H_0(B_k), \alpha(1 - H_0(B_k))) \tag{1}$$

with $K \to \infty$ and $H_0(B_k) \to 0$ for $k = 1, ..., K$. We denote the process $H \sim \text{BP}(\alpha H_0)$.

Hjort considers a generalization of this definition including functions, $\alpha(B_k)$, which we set as constants for the sake of simplicity. Analogous to the Dirichlet process, the beta process may be written in set function form as

$$H(\omega) = \sum_{k=1}^{\infty} \pi_k \delta_{\omega_k}(\omega) \tag{2}$$

with $H(\omega_i) = \pi_i$. Note that the beta process is not a normalized random measure. Hence the $\boldsymbol{\pi}$ of a beta process does not represent a probability mass function on Ω, but instead can be used to parametrize the *Bernoulli process*, a new measure on Ω defined as follows:

Definition 2. *Let $\boldsymbol{z_i}$ be an infinite row vector with k^{th} value, z_{ik}, generated by $z_{ik}|\pi_k \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_k)$. The measure defined by $X_i(\omega) = \sum_k z_{ik}\delta_{\omega_k}(\omega)$ is then a draw from a Bernoulli process, which we denote $X_i \sim \text{BeP}(H)$.*

If we were to stack samples of the infinite-dimensional vector, $\boldsymbol{z_i}$, to form a matrix, $\boldsymbol{Z} = [\boldsymbol{z_1}^\top, ..., \boldsymbol{z_N}^\top]^\top$, we may view the beta-Bernoulli process as a prior over infinite binary matrices (Griffiths & Ghahramani, 2011), where each column in the matrix $\boldsymbol{Z}$ corresponds to a location, δ_ω.

Sampling H directly, as defined in (2), is difficult to do exactly and efficiently. But, just as Aldous (1985) derived the Chinese restaurant process, a marginalized approach used for sampling from the Dirichlet process, there exists an efficient marginalized scheme for sampling from the beta process, called the Indian buffet process (IBP, Griffiths & Ghahramani, 2006; Thibaux & Jordan, 2007).

The IBP sampling procedure introduces strong dependencies between the rows of $\boldsymbol{Z}$. Our goal is to derive a stochastic variational inference scheme where we consider rows in batches. It will hence be crucial to instantiate the global parameters rather than marginalize over them.

For this reason, we shall consider a finite approximation to the beta process which simply set K to a large, finite number. The finite representation is written as

$$H(\omega) = \sum_{k=1}^{K} \pi_k \delta_{\omega_k}(\omega)$$
$$\pi_k \sim \text{Beta}(a/K, b(K-1)/K), \qquad \omega_k \sim H_0 \tag{3}$$

and the K-dimensional vector, $\boldsymbol{z_i}$, is drawn from a finite Bernoulli process parameterized by H.

Consider modelling a data matrix $\boldsymbol{Y} \in \mathbb{R}^{N \times D}$ where rows represent data points. Factor analysis models this data as the product of two matrices $\boldsymbol{L} \in \mathbb{R}^{N \times K}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{K \times D}$, plus an error matrix, $\boldsymbol{E}$.

$$\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{\Phi} + \boldsymbol{E} \tag{4}$$

Prior belief about the structure of the data may be used to induce the desired propeties of $\boldsymbol{L}$ and $\boldsymbol{\Phi}$, e.g. sparsity (West, 2003; Rai & Daumé, 2008; Knowles & Ghahramani, 2007). To encourage sparsity, we model $\boldsymbol{L}$ as the Hadamard (element-wise) product between matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$, $\boldsymbol{L} = \boldsymbol{Z} \circ \boldsymbol{W}$, where $\boldsymbol{Z}$ is binary and $\boldsymbol{W}$ is a Gaussian weight matrix. This idea is described in Section 3 of (Griffiths & Ghahramani, 2011). We model the matrices $\boldsymbol{\Phi}$ and $\boldsymbol{Z}$ as N draws from a beta-Bernoulli process parameterized by a beta process, H.

Using the truncated beta process of (3), we have the following generative process for observation $i = 1, ..., N$ and features $k = 1, ..., K$,

$$\begin{aligned}
\boldsymbol{y_i} &= (\boldsymbol{z_i} \circ \boldsymbol{w_i})\boldsymbol{\Phi} + \boldsymbol{\epsilon_i} && \\
\boldsymbol{w_i} &\sim \mathcal{N}(0, \gamma_w^{-1} I) && \left.\vphantom{\begin{matrix}a\\b\\c\end{matrix}}\right\} \text{Local variables} \\
z_{ik}|\pi_k &\sim \text{Bernoulli}(\pi_k) && \\
\boldsymbol{\epsilon_i} &\sim \mathcal{N}(0, \gamma_{\text{obs}}^{-1} I) && \\
\pi_k &\sim \text{Beta}(a/K, b(K-1)/K) && \left.\vphantom{\begin{matrix}a\\b\end{matrix}}\right\} \text{Global variables} \\
\boldsymbol{\phi_k} &\sim \mathcal{N}(0, D^{-1}\boldsymbol{I}) &&
\end{aligned} \tag{5}$$

where all values are drawn independently. This is the generative model used for beta process factor analysis (Paisley & Carin, 2009). We place independent Gamma(c', d')

and Gamma(e', f') priors on γ_{obs} and γ_w respectively. The separation of local and global variables will be crucial for the stochastic variational inference algorithm which we derive in the next section. For the sake of brevity, we denote the set of global variables $\boldsymbol{\beta} \equiv \{\boldsymbol{\pi}, \boldsymbol{\Phi}, \gamma_w, \gamma_{\text{obs}}\}$ and sets of local variables $\boldsymbol{\psi}_i \equiv \{\boldsymbol{w}_i, \boldsymbol{z}_i\}$ for $i = 1, ..., N$.

3. Variational Inference Schemes

The true posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N} | \boldsymbol{x}_{1:N})$ involves complicated dependencies between latent variables, which makes inference complicated. The goal of variational inference is to approximate the true posterior with a family of distributions $q(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N})$. We choose the best member of the chosen family of distributions by minimizing the KL-divergence between this variational distribution and the true posterior. Equivalently, we maximize the *evidence lower bound* (ELBO),

$$\mathcal{L}(q) = \mathbb{E}_q[\log p(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N}, \boldsymbol{x}_{1:N}) - \log q(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N})]. \quad (6)$$

In this work, we compare the performance of a range of variational approximations. Each of the approximations we consider factorizes as follows

$$q(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N}) = \Big(q(\gamma_{\text{obs}})q(\gamma_w)\prod_k q(\pi_k)q(\boldsymbol{\phi}_k)\Big)q(\boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}),$$

where $q(\gamma_{\text{obs}}) = \text{Gamma}(c, d)$, $q(\gamma_w) = \text{Gamma}(e, f)$, $q(\boldsymbol{\phi}_k) = \mathcal{N}(\tau_k^{-1}\boldsymbol{\mu}_k, \tau_k^{-1}\boldsymbol{I})$ and $q(\pi_k) = \text{Beta}(a_k, b_k)$. The global variational distributions are all of the same exponential family forms as their posterior conditional distributions. The full set of global variational parameters is $\boldsymbol{\lambda} = \{a_k, b_k, c, d, e, f, \tau_k, \boldsymbol{\mu}_k\}$. Due to conjugacy of our model, it is easy to show that the updates for the global variational parameters during the variational M-step are as follows

$$\begin{aligned}
a_k &= a/K + \sum_i \mathbb{E}_q[z_{ik}] \quad (7)\\
b_k &= b(K-1)/K + \sum_i \Big(1 - \mathbb{E}_q[z_{ik}]\Big)\\
c &= c' + \sum_i D/2\\
d &= d' + \sum_i \frac{1}{2}\mathbb{E}_q\Big[\big\|\boldsymbol{y}_i - (\boldsymbol{z_i} \circ \boldsymbol{w_i})\boldsymbol{\Phi}\big\|^2\Big]\\
e &= e' + \sum_i K/2\\
f &= f' + \sum_i \frac{1}{2}\mathbb{E}_q\big[\boldsymbol{w_i}\boldsymbol{w_i}^\top\big]\\
\tau_k &= D + \sum_i \mathbb{E}_q\big[\gamma_{\text{obs}} z_{ik} {w_{ik}}^2\big]\\
\boldsymbol{\mu}_k &= \sum_i \mathbb{E}_q\big[z_{ik} w_{ik} \boldsymbol{y}_i^{-k}\big]
\end{aligned}$$

where $\mathbb{E}_q$ is an expectation over all latent variables with respect to q (except when global parameters are being sampled), and $\boldsymbol{y}_i^{-k} = \boldsymbol{y}_i - \mathbb{E}_q\big[\sum_{j\neq k} z_{ij} w_{ij} \boldsymbol{\phi}_j\big]$. We work with the *natural parameters* of the global variational distributions. Natural gradients give the direction of steepest ascent in Riemannian space, leading to faster convergence for e.g. maximum likelihood estimation (Amari, 1998).

Our aim is to update the global variational parameters stochastically, by considering subsets of the full dataset and making sequential updates. Let $\boldsymbol{\eta}$ denote the vector of global natural parameters of $p(\boldsymbol{\beta})$, and by conditional conjugacy, the vector of global natural parameters of $q(\boldsymbol{\beta})$ is $\boldsymbol{\eta} + \sum_i \boldsymbol{\eta}_i(\boldsymbol{y}_i, \boldsymbol{\psi}_i)$. In fact, $\boldsymbol{\eta} = [a/K, b(K-1)/K, c', d', e', f', D, \boldsymbol{0}]$, and $\boldsymbol{\eta}_i$ is a vector consisting of each of the i^{th} elements of the sums in Equation 7. We have followed the notation of Hoffman & Blei (2014). The general framework of stochastic variational inference we shall follow is summarized in Algorithm 1.

The difficult step is in computing $\hat{\boldsymbol{\eta}}_i$, and it is entirely dependent on the form of the local variable approximation, of which we consider 2 types: 'Unstructured' methods where $q(\boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}) = q(\boldsymbol{\psi}_{1:N})$ and 'structured' methods where this equivalence does not hold. Our notion of 'structure' describes the dependence between local and global variables, as discussed by Hoffman & Blei (2014).

3.1. Unstructured Variational Methods

The simplest, and most commonly used, approximation we can make is the mean field approximation,

$$q_{\text{MF}}(\boldsymbol{\psi}_i) = \prod_{k=1}^{K} q(z_{ik})q(w_{ik}) \quad (8)$$

where $q(z_{ik}) = \text{Bernoulli}(\theta_{ik})$ and $q(w_{ik}) = \mathcal{N}(\kappa_{ik}^{-1}\nu_{ik}, \kappa_{ik}^{-1})$. Given the current set of global parameters, $\boldsymbol{\lambda}^{(t)}$, the local ELBO, $\mathcal{L}_{\text{local}} = \mathbb{E}_{q(\boldsymbol{\beta})q_{\text{MF}}(\boldsymbol{\psi}_{1:N})}[\log p(\boldsymbol{y}_{1:N}, \boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}) - \log q(\boldsymbol{\psi}_{1:N})]$ is optimized as a function of local variational parameters $\{\theta_{ik}, \nu_{ik}, \kappa_{ik}\}$. Up to irrelevant constants,

$$\begin{aligned}
\mathcal{L}_{\text{local}}^{\text{MF-SVI}} = \frac{c}{2d}\sum_{i,k}\Bigg[&2\theta_{ik}\frac{\nu_{ik}}{\kappa_{ik}}\frac{\boldsymbol{\mu}_k}{\tau_k}\boldsymbol{y}_i^\top\\
&- \theta_{ik}\left(\frac{{\nu_{ik}}^2}{{\kappa_{ik}}^2} + \frac{1}{\kappa_{ik}}\right)\left(\frac{\boldsymbol{\mu}_k\boldsymbol{\mu}_k^\top}{{\tau_k}^2} + \frac{1}{\tau_k}\right)\\
&- \sum_{j\neq k}\theta_{ij}\theta_{ik}\frac{\nu_{ij}}{\kappa_{ij}}\frac{\nu_{ik}}{\kappa_{ik}}\frac{\boldsymbol{\mu}_j\boldsymbol{\mu}_k^\top}{\tau_j\tau_k}\Bigg] - \frac{e}{2f}\sum_{i,k}\left(\frac{{\nu_{ik}}^2}{{\kappa_{ik}}^2} + \frac{1}{\kappa_{ik}}\right)\\
&+ \sum_{i,k}\theta_{ik}(\psi(a_k) - \psi(b_k)) - \frac{1}{2}\sum_{i,k}\log(\kappa_{ik})\\
&- \sum_{i,k}\big[\theta_{ik}\log\theta_{ik} + (1-\theta_{ik})\log(1-\theta_{ik})\big] \quad (9)
\end{aligned}$$

where ψ is the digamma function. The mean field approximation breaks dependencies between all local and global variables, and will provide a baseline to compare against. It is possible to compute $\mathbb{E}_{q_{\text{MF}}(\boldsymbol{\psi}_i)}[\boldsymbol{\eta}_i]$ analytically given the optimized local variational parameters. We denote the SVI algorithm which uses a mean field local variable approximation as MF-SVI. It is identical to the original SVI algorithm introduced by Hoffman et al. (2013).

Mimno et al. (2012) suggested an online SVI method which maintains structure between local variables specifically for the LDA. We generalize their idea by suggesting the following variational distribution over local parameters

$$q_{\text{Mimno}}(\boldsymbol{\psi}_i) = \exp(\mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\boldsymbol{\psi}_i|\boldsymbol{y}_{1:N}, \boldsymbol{\beta})]) \tag{10}$$

where $p(\boldsymbol{\psi}_i|\boldsymbol{y}_{1:N}, \boldsymbol{\beta})$ is the true posterior conditional of $\boldsymbol{\psi}_i$. Whilst we are unable to compute $\mathbb{E}_{q_{\text{Mimno}}(\boldsymbol{\psi}_i)}[\boldsymbol{\eta}_i]$ analytically, we are able to estimate it using MCMC. The SVI algorithm using q_{Mimno} as the local variational distribution shall be called Mimno-SVI.

3.2. Structured Variational Methods

Instead of taking an expectation over $q(\boldsymbol{\beta})$ to compute $\hat{\boldsymbol{\eta}}_i$ as in the previous section, we use the current set of global parameters $\boldsymbol{\lambda}^{(t)}$ to draw a sample $\boldsymbol{\beta}^{(t)}$, and compute an estimate of $\mathbb{E}_{q(\boldsymbol{\psi}_i|\boldsymbol{\beta}^{(t)})}[\boldsymbol{\eta}_i]$. Under this framework, Algorithm 1 becomes the SSVI-A algorithm of Hoffman & Blei (2014).

Once again, the simplest approximation that can be made is the conditional mean-field approximation, where z_{ik}, w_{ik} are independent given $\boldsymbol{\beta}$, with $q(z_{ik}) = \text{Bernoulli}(\theta_{ik})$ and $q(w_{ik}) = \mathcal{N}(\kappa_{ik}^{-1}\nu_{ik}, \kappa_{ik}^{-1})$. This time, we optimize the local ELBO, $\mathbb{E}_{q_{\text{MF}}(\boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}^{(t)})}[\log p(\boldsymbol{y}_{1:N}, \boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}^{(t)}) - \log q(\boldsymbol{\psi}_{1:N})]$ as a function of local variational parameters $\{\theta_{ik}, \nu_{ik}, \kappa_{ik}\}$, and compute $\mathbb{E}_{q_{\text{MF}}(\boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}^{(t)})}[\boldsymbol{\eta}_i]$ analytically given the optimized parameters. We shall call this SVI method MF-SSVI.

The Bernoulli-Gaussian products present in the generative process in Equation 5 can be thought of as a spike-and-slab model. Titsias & Lázaro-Gredilla (2011) developed a variational method which maintains dependence between z_{ik} and w_{ik} for eack k, such that

$$q_{\text{Titsias}}(\boldsymbol{\psi}_i|\boldsymbol{\beta}^{(t)}) = \prod_k \text{Bernoulli}(z_{ik}; \theta_{ik}) \tag{11}$$
$$\times \mathcal{N}\big(w_{ik}; z_{ik}\kappa_{ik}^{-1}\nu_{ik}, z_{ik}\kappa_{ik}^{-1} + (1 - z_{ik})\gamma_w^{(t)^{-1}}\big).$$

This approximation has the advantage that it maintains the spike-slab beaviour of the product $z_{ik}w_{ik}$, and matches the exact posterior when $z_{ik} = 0$. However, the dependencies between local variables for which $k \neq k'$ are lost. Analogous to MF-SSVI, we optimize the local ELBO using q_{Titsias} as a function of $\{\theta_{ik}, \nu_{ik}, \kappa_{ik}\}$, and compute

Algorithm 1 General Stochastic Variational Inference

Initialize $t = 1$, $\boldsymbol{\lambda}^{(0)}$.
repeat
  Compute step size $\rho^{(t)} = (t + t_0)^{-\zeta}$.
  Select subset of full data set, $\mathcal{D}$.
  Compute $\hat{\boldsymbol{\eta}}_i$, an (unbiased) estimator of $\mathbb{E}_{q(\boldsymbol{\psi}_i|\boldsymbol{\beta})}[\boldsymbol{\eta}_i]$ for each $i \in \mathcal{D}$
  Set $\boldsymbol{\lambda}^{(t)} = (1 - \rho^{(t)})\boldsymbol{\lambda}^{(t-1)} + \rho^{(t)}\big(\boldsymbol{\eta} + \frac{N}{|\mathcal{D}|}\sum_{i\in\mathcal{D}}\hat{\boldsymbol{\eta}}_i\big)$
until convergence

$\mathbb{E}_{q_{\text{Titsias}}(\boldsymbol{\psi}_{1:N}|\boldsymbol{\beta}^{(t)})}[\boldsymbol{\eta}_i]$ analytically given the optimized parameters. We denote the SVI algorithm which uses the Titsias & Lázaro-Gredilla (2011) local approximation as Titsias-SSVI.

Finally we consider using the exact local conditional distribution given by $q(\boldsymbol{\psi}_i|\boldsymbol{\beta}^{(t)}) = p(\boldsymbol{\psi}_i|\boldsymbol{\beta}^{(t)}, \boldsymbol{y}_i)$. We use MCMC samples to compute $\hat{\boldsymbol{\eta}}_i$ using a Gibbs sampling scheme. We therefore call this method Gibbs-SSVI.

MF-SVI (Hoffman et al., 2013), MF-SSVI, Gibbs-SSVI (Hoffman & Blei, 2014) and Mimno-SVI (Mimno et al., 2012) have been considered in the context of LDA before, but the latter 3 have not been applied to factor analysis to the best of our knowledge. Titsias-SSVI is a new method as Titsias & Lázaro-Gredilla (2011) applied their variational approximation only to regression tasks. More details on the variational approximations over local variables is provided in supplementary variables.

4. Related Work

The idea of applying variational inference to the Indian buffet process was first proposed in Doshi et al. (2008), based on the stick breaking construction of the IBP (Teh et al., 2007). Promising results were shown for the simple but somewhat limited "linear Gaussian" model, which is the model presented here without the weight vector, $\boldsymbol{w}_i$. Paisley & Carin (2009) consider the simpler finite approximation to the beta process described above, and extended the model to include continuous weights w_i. An extension using power-EP, able to handle non-negativity constraints, was developed in (Ding et al., 2010) but has not been widely adopted. Alternative approaches to scale inference in IBP based models have included parallelization (Doshi-Velez et al., 2009) and submodular optimization (Reed & Ghahramani, 2013). The former only performed approximate sampling, and the later is greedy and limited to positive weights. Mean field based stochastic variational inference schemes have been used for large scale dictionary learning, with some success (Li et al., 2012; Polatkan et al., 2014). However, we shall show that preserving dependencies between local variables will greatly improves performance on image interpolation and denoising tasks.

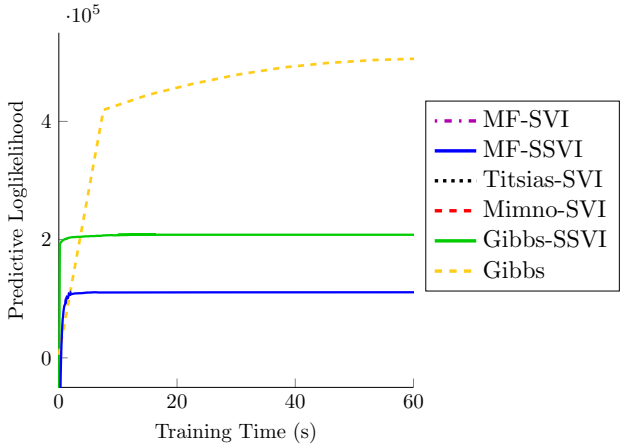


Figure 1. Predicitve loglikelihood versus training time on synthetically generated data, comparing Gibbs-SSVI, MF-SSVI and Gibbs sampling. The same legend is used throughout this paper.

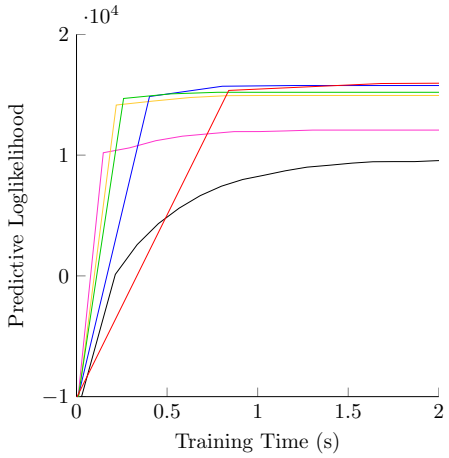


Figure 2. Predicitve loglikelihood versus training time on synthetically generated data using Gibbs-SSVI with burn-in lengths of 0,1,3,5,10 and 25. Converged predictive loglikelihood is monotonically increasing in burn-in length, so no legend is included.

Meanwhile the topic modelling community has taken great strides developing stochastic variational inference methods for latent Dirichlet allocation (Blei et al., 2003), encouraged by the availability of large corpora of text. The idea was initially proposed in Hoffman et al. (2010), and refined in Mimno et al. (2012) where the sparse updates of Gibbs sampling were leveraged to scale inference on just a single machine to 1.2 million books. The latter idea allows non-truncated online learning (Wang & Blei, 2012) of Bayesian non-parametric models, though only the hierachical Dirichlet process (Teh et al., 2004) was demonstrated.

More recently, Hoffman & Blei (2014); Liang & Hoffman (2014) have shown that sampling from the global variational distribution improves predictive performance for the LDA and Bayesian non-negative matrix factorization respectively. In fact, the idea of optimizing an intractable variational inference algorithm by sampling from global variational distributions has been proposed in various contexts to deal with non-conjugacy (Ji et al., 2010; Nott et al., 2012; Gerrish, 2013; Paisley et al., 2012; Ranganath et al., 2014). Kingma & Welling (2014); Titsias & Lázaro-Gredilla (2014); Salimans & Knowles (2013) propose change of variable methods to deal with non-conjugacy or improve convergence speed. In this work we focus more on the quality of the variational approximation and attempt to exploit the conditional conjugacy.

5. Experiments

In this section we discuss our findings from a range of experiments. Results from experiments carried out on synthetically generated data are discussed first. We apply a range of stochastic variational inference algorithms to carry out image inpainting and denoising tasks next. Finally the same algorithms are applied to two large genomic datasets. We choose to compare our models using predictive loglikelihood of held out data, which we compute as

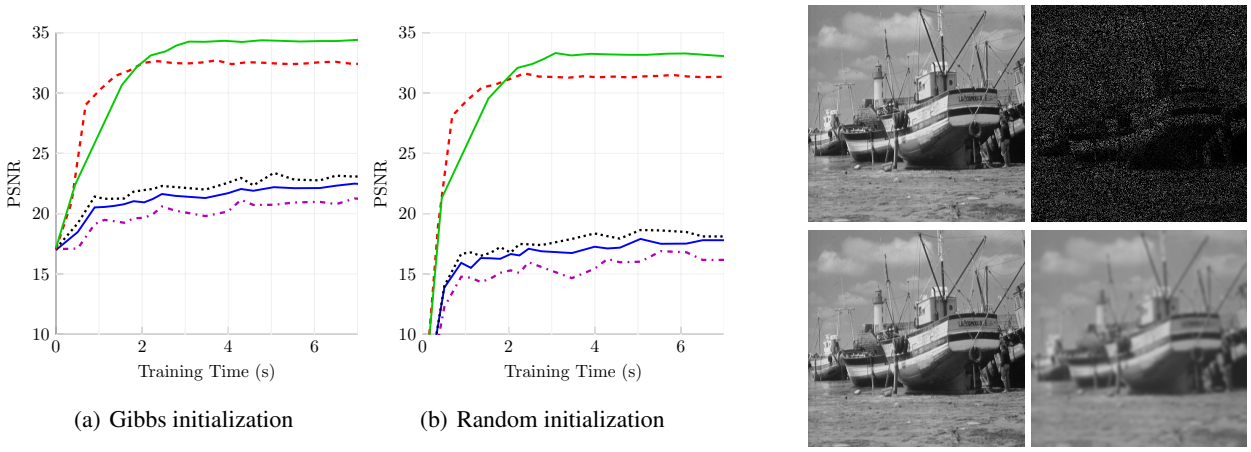
$$p\left(\hat{\boldsymbol{Y}}|\boldsymbol{Y}\right) \approx \int p(\hat{\boldsymbol{Y}}|\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N_{\text{test}}}) q(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N_{\text{test}}}) d(\boldsymbol{\beta}, \boldsymbol{\psi}_{1:N_{\text{test}}})$$
$$\approx \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N_{\text{test}}} \mathcal{N}\left(\hat{\boldsymbol{y}}_n | \boldsymbol{z}_i^{(m)} \circ \boldsymbol{w}_i^{(m)} \boldsymbol{A}^{(m)}, \boldsymbol{I}/\gamma_{\text{obs}}^{(m)}\right) \tag{12}$$

where $\left(\boldsymbol{z}_i^{(m)}, \boldsymbol{w}_i^{(m)}, \boldsymbol{A}^{(m)}, \gamma_{\text{obs}}^{(m)}\right)$ are independent samples from q, for whichever type of variational approximation is being used, and $\hat{\boldsymbol{y}}_i$ is the i^{th} data point in the test set.

In each of our experiments, we transform the data to have empirical mean 0 and variance 1. Hyperparameters are set as follows: $a = b = 10$, $c = 1$, $d = 10$, $e = f = 1$, and a learning rate schedule of $\rho_t = t^{-0.75}$ is employed.

5.1. Synthetic Data

A key question that is important to consider when using variational approximations of a particular form is, 'how close is the approximation to the true posterior?'. We attempt to answer such a question with our first experiment. Data was generated from our prior with parameters $\gamma_w = 1$, $\gamma_{\text{obs}} = 100$, $K = 80$, $N = 1e5$ and $D = 40$, with 7.5% selected uniformly at random held out for testing on 100 independent experiments. We then applied Gibbs-SSVI and MF-SSVI, as well as an uncollapsed Gibbs sampler to the generated data using $K = 150$ potential features and random initialization. The predictive mean squared errors (MSE) of the 3 methods were 0.022 ± 0.002, 0.027 ± 0.004



(a) Gibbs initialization (b) Random initialization

Figure 3. Results from interpolation of the 512×512 pixel 'Boat' image. PSNR vs training time shown using (a) Gibbs initialization and (b) random initialization. The pictures shown are the original image (top left), the image to be reconstructed with 80% of pixels unobserved (top right), the Gibbs-SSVI reconstruction (bottom left) and the MF-SSVI reconstruction (bottom right).

and 0.020 ± 0.002 and the average per iteration training times were 1.5, 0.6 and 7.6 seconds respectively. Figure 1 illustrates our findings. The Gibbs sampler achieves a high predictive likelihood, but the average training time per iteration was very high versus the SVI methods. Warm starting the Gibbs sampler at each iteration helped Gibbs-SSVI to converge in few iterations, whereas stochastically choosing subsets of data points in SSVI methods requires reinitializing local variables at each epoch. Notice that the predictive MSE of Gibbs-SSVI is close to the Gibbs sampler, suggesting that the correct mean is being learnt. The lower likelihood the SSVI methods are able to achieve is therefore due to a poor calibration in posterior variance, a known issue with variational methods (Consonni & Marin, 2007).

Another question of interest to us was, 'what is the empirical trade-off between training time and unbiasedness in the Gibbs-SSVI scheme?'. More specifically, if we allow the Gibbs sampler over the local variables to converge, the subsequent ELBO gradient estimates would be unbiased, whilst using samples from a Gibbs chain which has not converged would lead to biased gradient estimates. Convergence of the Gibbs chain, however, may take a long time. Therefore we experimented with a range of burn-in lengths of the Gibbs chain on synthetically generated data. Various burn-in and sample length combinations were discussed by Mimno et al. (2012). We tried burn-in lengths of 0,1,3,5,10 and 25 whilst fixing the number of samples used after burn-in to 3. The results can be seen in Figure 2. When the burn-in length is below 3 we notice severe loss in predictive power of the Gibbs-SSVI method. We notice diminishing gains in predictive power as we increase the length of burn-in. This experiment suggests that some bias introduced by using samples from an unconverged Gibbs chain may be worth the reduction in training time. For subsequent experiments, we fix the burn-in length to 3.

5.2. Image Interpolation and Denoising

Zhou et al. (2009) first applied the beta process for sparse image representation with good results and much follow up research. The standard metric used for quantifying the quality of a reconstructed image is the peak signal-to-noise ratio (PSNR), defined as $20 \log_{10}(\max_{\text{image}}/\text{rmse})$, where $\max_{\text{image}}$ is the maximum possible pixel value and rmse is the root mean squared error of the reconstruction.

We consider overlapping 8×8 pixel patches as individual 64 dimensional data points. The fact that the patches are overlapping technically breaks the exchangeability assumption of the prior distribution, however the extra model averaging is beneficial to prediction. Five grayscale images originally from Portilla et al. (2003) were used for our study: Boat, Barbara, Lena, House and Peppers. The first 3 are 512×512 in size whilst the last 2 are 256×256. The datasets are therefore of size $N = (512-7)^2 = 255,025$ and $N = (256-7)^2 = 62,001$ for 512×512 and 256×256 images respectively. We use a batchsize of $N_{\text{subset}} = 250$ and $K = 250$ features for our experiments.

For our first experiment, we consider the task of image interpolation, where the task is to reconstruct an image where only 20% of the pixels, chosen uniformly at random, are observed. Li et al. (2012) consider a mean field based variational approximation for such a task, however, the learning rate schedule they used was $\rho_t = (t + 1000)^{-0.5}$. This implies $\rho_t < 0.032$ for all $t \geq 1$, and that their algorithm relied heavily on the initialization of global parameters. They ran an MCMC algorithm over a subset of the data to initialize these global parameters, and we argue that this was integral to the performance of their algorithm. We decided to test how sensitive the variational algorithms were

Figure 4. Results from interpolation and denoising of the 512×512 pixel 'Barbara' image. The pictures shown are the original image (top left), the image to be reconstructed with 50 % of pixels unobserved and remaining pixels corrupted with Gaussian noise (top right), the Gibbs-SSVI reconstruction (bottom left) and the MF-SSVI reconstruction (bottom right).

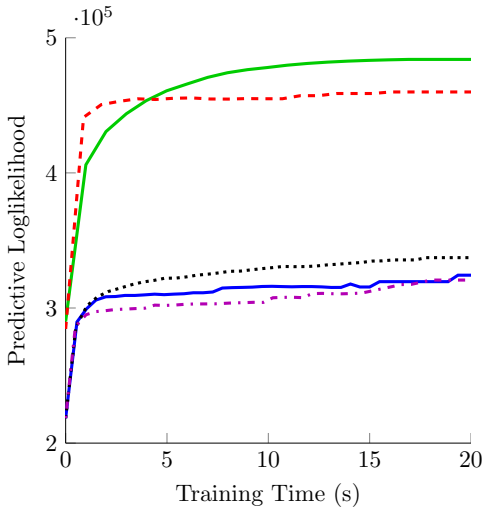


Figure 5. Predicitve loglikelihood versus training time on cell line data comparing five SVI algorithms.

to different initialization methods and an example can be seen in the performance graphs of Figure 3. We found that initializing using MCMC improved the PSNR of MF-SVI, MF-SSVI and Titsias-SSVI by 4.8 on average versus random initialization. However, the analogous improvement for Gibbs-SSVI and Mimno-SVI was 1.0. This suggests that the methods which preserve intra-local variable structure are less sensitive to initialization.

Secondly, we considered the joint task of image interpolation and denoising. Here, we observe 50 % of the pixels chosen uniformly at random, except they are now corrupted with Gaussian noise with standard deviation 15 (the original pixels take integer values in $[0, 255]$). Results for both image interpolation and denoising tasks are summarized in Table 1. Gibbs-SSVI and Mimno-SVI consistently outperform the other methods and an explanation, outlined in Section 5.3, as to why this is the case can be deduced by studying the images in Figures 3 and 4.

For the 512×512 images, the average training time per epoch was $0.12, 0.12, 0.13, 0.46$ and 0.48 secs for the MF-SVI, MF-SSVI, Titsias-SSVI, Mimno-SVI and Gibbs-SSVI methods respectively, on a 2.4GHz dual core machine. Among the multiple experiments, the MF-SSVI reconstructions were similar in appearance to the MF-SVI and Titsias-SSVI methods, whilst the Gibbs-SSVI reconstructions were similar to the Mimno-SVI ones. We believe the blurred appearance of the former 3 methods' reconstructions is a result of the independence between z_{nk} and $z_{nk'}$ for $k \neq k'$ in their variational forms. In contrast, the Gibbs-SSVI and Mimno-SVI methods maintain dependence between z_{nk} and $z_{nk'}$, and are therefore able to select a subset of features which collectively best explain the data. The latter methods are consequently much more capable of capturing structure and detail in the images we tested on, and there is a mild cost to pay in extra training time.

5.3. A Thought Experiment

To illustrate the problem with breaking dependencies between z_{ik} and $z_{ik'}$ for $k \neq k'$, we can consider a simple thought experiment. Suppose $\boldsymbol{y}_i = \boldsymbol{f} + \boldsymbol{\epsilon}_i$, $\boldsymbol{y}_i \in \mathbb{R}^D$, where $\boldsymbol{f} \sim \mathcal{N}(0, \boldsymbol{I})$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, 0.05\boldsymbol{I})$ independently for $i = 1, ..., N$. Now consider applying MF-SSVI and Gibbs-SVI algorithms to this dataset, whilst fixing $w_{ik} = 1$ for each i, k, and using $K = 2$ features. Let's assume that at the current iteration $\pi_1^{(t)} \approx \pi_2^{(t)}$ and $\boldsymbol{\phi}_1^{(t)} \approx \boldsymbol{\phi}_2^{(t)} \approx \boldsymbol{f}$. The local ELBO for the MF-SSVI method will have local optima for $\theta_{i1} = 1 - \theta_{i2}$, since exactly 1 feature is needed to explain the data, so MF-SSVI would find a local optimum of the form $q(z_{i1}) = \text{Bernoulli}(s)$, $q(z_{i2}) = \text{Bernoulli}(1 - s)$ for some $s \in (0, 1)$. (We were able to verify this form of local optimum empircally). The Gibbs-SSVI will generate samples of the form $\boldsymbol{z}_i = (0, 1)$ and $\boldsymbol{z}_i = (1, 0)$.

We would like to predict $\boldsymbol{y}_i$ with each model. Since $q(z_{i1})$ and $q(z_{i2})$ are independent under MF-SSVI, predictions will be $\hat{\boldsymbol{y}}_i = 0$ with probability $s(1 - s)$, $\hat{\boldsymbol{y}}_i \approx \boldsymbol{y}_i$ with probability $s^2 + (1-s)^2$ and finally $\hat{\boldsymbol{y}}_i \approx 2\boldsymbol{y}_i$ with probability $s(1 - s)$. Conversely, the Gibbs samper in Gibbs-SSVI will place little to no probability on both, or neither of the 2 features being used for prediction. In summary, the Gibbs based local variable estimates can handle the strong correlation between the 2 features, whilst the MF based method cannot and suffers dramatically because of it.

One possible solution to this problem would be to encourage all features to have limited correlation apriori, discouraging situations where multiple features are learned to be

Table 1. PSNR performance of image interpolation (left entries) and denoising (right entries) tasks using Gibbs initialization of global parameters on a randomly chosen subset of data.

	BOAT		BARBARA		LENA		HOUSE		PEPPERS	
MF-SVI	21.1	19.5	21.8	20.6	24.1	23.6	25.3	24.2	25.9	24.4
MF-SSVI	22.3	20.8	22.2	21.4	24.7	24.4	26.7	25.4	25.8	24.1
TITSIAS-SSVI	23.2	21.5	22.1	21.7	26.3	25.8	26.7	25.3	27.9	26.8
MIMNO-SVI	32.4	29.7	36.2	35.1	39.4	36.9	**42.8**	**40.1**	43.7	40.4
GIBBS-SSVI	**34.3**	**31.5**	**38.2**	**37.0**	**43.3**	**41.7**	40.5	37.8	**47.4**	**42.3**

similar to each other. Encouraging dissimilarity in such a way is challenging and would complicate the otherwise clean updates that are possible in SVI methods.

5.4. Genomic data

Vast amounts of genomic data are currently being collected as technology advances. It will be crucial to develop machine learning models and more importantly, inference algorithms, which can cope with large data sets, whilst still retaining flexible modelling ability. We consider 2 datasets for which sparse latent feature modelling is appropriate. We use $K = 500$ features for both experiments.

Cancer cell line data. The Cancer Cell Line Encyclopedia is a collection of around 450 cancer samples including gene expression, copy number variation, and drug response information. We focus on modeling the gene expression data, which has measurements for around 15,000 genes. In this setting we are more interested in finding overlapping clusters (sparse features) of *genes* rather than samples, so we effectively have $N = 15000, D = 450$. The latent factors found can then be interpreted as biological pathways, or sets of genes regulated by the same transcription factor. Understanding the structure in this data is valuable as a first step towards associating the cellular characteristics of the cancers to their drug response profiles. We randomly hold out 10 % of the data for testing. Results for this experiment are summarized in Figure 5.

CyTOF data. CyTOF is a novel extremely high throughput technology capable of measuring up to 40 protein abundance levels in thousands of individual cells per second. The cells are controlled using flow cytometry and specific proteins are tagged using heavy metals which can be measured using time-of-flight mass spectrometry. Existing analyses have attempted to group the observed cells into non-overlapping subpopulations, but we here show that the data can be effectively modeled as compromising of a spectrum of cell types expressing different latent factors to differing extents. The sample we analyse consists of human immune cells, so representing the heterogeneity is relevant for understanding disease response. Our dataset has $N = 532,000, D = 40$ and a random 5% is used as test data. The results for the experiments on this data follow a very similar pattern to that of the cell line gene expression data. The converged predictive log-likelihoods after training for 10 minutes are $-1.1e6$, $-9.6e5$, $-9.4e5$, $-3.8e5$ and $-3.2e5$ for the MF-SVI, MF-SSVI, Titsias-SSVI, Mimno-SVI and Gibbs-SSVI methods respectively.

6. Conclusions

In this work, we compare various stochastic variational inference algorithms for beta process factor analysis. Whist many methods in the literature have been proposed, we have chosen to exploit the conditional conjugacy and the exponential family nature of our model to create simple natural parameter updates.

Hoffman & Blei (2014) found that preserving structure between local and global variables significantly boosted performance for the LDA, but based on our experiments, we conclude that preserving intra-local variable dependence is crucial to prediction in the beta-Bernoulli process. This is evident from the fact that both Gibbs-SSVI and Mimno-SVI consistently and significantly outperform MF-SVI, MF-SSVI and Titsias-SSVI on a variety of image interpolation and denoising tasks and on modelling genomic data. The Titsias-SSVI method models dependence between z_{ik} and w_{ik}, but does not appear to significantly outperform MF-SSVI, suggesting that this dependence is not crucial in prediction. Mimno-SVI does not maintain dependence between local and global variables whilst MF-SSVI does, and yet Mimno-SVI leads to better predictions. We discuss why this is the case in a simple thought experiment, showing the benefit of maintaining dependence between local variables where $k \neq k'$. The multi-cluster, sparse nature of the beta-Bernoulli process makes mean field type local variable approximations highly sensitive to correlated features. Gibbs-SSVI does also modestly outperform Mimno-SVI through maintaining dependencies between global and local variables.

In summary, care is needed to ensure that the dependencies encoded by a particular variational approximation are appropriate for the model being considered.

References

Ahn, Sungjin, Korattikara, Anoop, and Welling, Max. Bayesian posterior sampling via stochastic gradient fisher scoring. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 1591–1598, 2012.

Aldous, D. J. Exchangeability and Related Topics. In *École d'Été de Probabilitiés de Saint-Flour XIII - 1983*, pp. 1–198. Springer Berlin Heidelberg, 1985.

Amari, S. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998.

Bendall, S. C., Simonds, E. F., Qiu, P., El-ad, D. A., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030): 687–696, 2011.

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

Consonni, G. and Marin, J. M. Mean-field Variational Approximate Bayesian Inference for Latent Variable Models. *Computational Statistics and Data Analysis*, 52: 790–798, 2007.

Ding, N., Xiang, R., Molloy, I., Li, N., et al. Nonparametric Bayesian matrix factorization by Power-EP. In *International Conference on Artificial Intelligence and Statistics*, pp. 169–176, 2010.

Doshi, F., Miller, K. T., Gael, J. Van, and Teh, Y. W. Variational inference for the Indian buffet process. *Advances in Neural Information Processing Systems*, 2008.

Doshi-Velez, F., Knowles, D. A., Mohamed, S., and Ghahramani, Z. Large Scale Nonparametric Bayesian Inference: Data Parallelisation in the Indian Buffet Process. In *Advances in Neural Information Processing Systems*, volume 22, pp. 2–3, 2009.

Gerrish, S. *Applications of Latent Variable Models in Modeling Influence and Decision Making*. PhD thesis, Princeton University, 2013.

Ghahramani, Z. and Beal, M. J. Variational Inference for Bayesian Mixtures of Factor Analyzers. In *Advances in Neural Information Processing Systems*, 1999.

Griffiths, T. and Ghahramani, Z. Infinite Latent Feature Models and the Indian Buffet Process. *Advances in Neural Information Processing Systems*, 2006.

Griffiths, T. and Ghahramani, Z. The Indian Buffet Process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Hjort, N. L. Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294, 1990.

Hoffman, M. D. and Blei, D. M. Structured Stochastic Variational Inference. *arXiv*, 2014. http://arxiv.org/abs/1404.4114.

Hoffman, M. D., Blei, D. M., and Bach, F. R. Online Learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, volume 2, pp. 5, 2010.

Hoffman, M. D., Blei, D.M., Wang, C., and Paisley, J. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Ji, C., Shen, H., and West, M. Bounded Approximations for Marginal Likelihoods. *Technical Report, Duke University*, 2010. http://ftp.stat.duke.edu/WorkingPapers/10-05.pdf.

Kingma, D. and Welling, M. Auto-Encoding Variational Bayes. *Intl. Conf. on Learning Representations*, 2014.

Knowles, D. and Ghahramani, Z. Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. *7th International Conference on Independent Component Analysis and Signal Separation*, 2007.

Knowles, David, Ghahramani, Zoubin, et al. Nonparametric bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

Li, L., Silva, J., Zhou, M., and Carin, L. Online Bayesian Dictionary Learning for Large Datasets. *Intl. Conf. on Acoustics, Speech and Signal Processing*, 2012.

Liang, D. and Hoffman, M. D. Beta Process Non-negative Matrix Factorization with Stochastic Structured Mean-Field Variational Inference. *arXiv*, 2014. http://arxiv.org/abs/1411.1804.

MacEachern, Steven N and Müller, Peter. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

Mimno, D., Hoffman, M., and Blei, D. Sparse Stochastic Inference for Latent Dirichlet Allocation. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Nott, D., Tan, S., Villani, M., and Kohn, R. Regression Density Estimation with Variational Methods and Stochastic Approximation. *Journal of Computational and Graphical Statistics*, 21(3):797–820, 2012.

Orbanz, Peter and Teh, Yee Whye. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pp. 81–89. Springer, 2010.

Paisley, J. and Carin, L. Nonparametric Factor Analysis with Beta Process Priors. *Proceedings of the 26th International Conference on Machine Learning*, 2009.

Paisley, J., Blei, D., and Jordan, M. Variational Bayesian Inference with Stochastic Search. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Polatkan, G., Zhou, M., Carin, L., Blei, D., and Daubechies, I. A Bayesian Nonparametric Approach to Image Super-resolution. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2014.

Portilla, J., Strela, V., Wainwright, M. J., and Simoncelli, E. P. Image Denoising using Scale Mixtures of Gaussians in the Wavelet Domain. *IEEE Trans on Image Processing*, 2003.

Rai, P. and Daumé, H. The Infinite Hierarchical Factor Regression Model . *Advances in Neural Information Processing Systems*, 2008.

Ranganath, R., Gerrish, S., and Blei, D. Black Box Variational Inference. *Proceedings of the 17th Conference on Artificial Intelligence and Statistics*, 2014.

Rasmussen, Carl and Williams, Chris. Gaussian processes for machine learning. *Gaussian Processes for Machine Learning*, 2006.

Reed, C. and Ghahramani, Z. Scaling the Indian Buffet Process via Submodular Maximization. *Proceedings of the 30th International Conference on Machine Learning*, 2013.

Salimans, T. and Knowles, D. Fixed-Form Variational Posterior Approximation Through Stochastic Linear Regression. *Bayesian Analysis*, 8:837–882, 2013.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, 2004.

Teh, Y. W., Görür, D., and Ghahramani, Z. Stick breaking construction for the Indian buffet process. *Proceedings of the 11th Conference on Artificial Intelligence and Statistics*, 2007.

Thibaux, R. and Jordan, M.I. Hierarchical Beta Processes and the Indian Buffet Process. *Proceedings of the 11th Conference on Artificial Intelligence and Statistics*, 2007.

Titsias, M. K. and Lázaro-Gredilla, M. Spike and Slab Variational Inference for Multi-Task and Multiple Kernel Learning. *Advances in Neural Information Processing Systems*, 2011.

Titsias, M. K. and Lázaro-Gredilla, M. Doubly Stochastic Variational Bayes for Non-Conjugate Inference. *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Wang, C. and Blei, D. Truncation-free stochastic variational inference for bayesian nonparametric models. *Advances in Neural Information Processing Systems*, 25: 422–430, 2012.

Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688, 2011.

West, M. Bayesian Factor Regression Models in the "large p, small n" Paradigm . *Bayesian Statistics*, 7:723–732, 2003.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations. *Advances in Neural Information Processing Systems*, 2009.