AMSI VACATIONRESEARCH SCHOLARSHIPS 2020–21

Get a Thirst for Research this Summer

Variational Inference for Bayesian Non-negative Matrix Factorisation

Gyu Hwan Park

Supervised by Heejung Shim

The University of Melbourne

Vacation Research Scholarships are funded jointly by the Department of Education, Skills and Employment and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government
Department of Education, Skills and Employment

Abstract

Non-negative Matrix Factorisation (NMF) models decompose a non-negative data matrix into a product of two low-dimensional non-negative matrices. This is done by discovering factors that represent the intrinsic features of the data, and relating these to original features and samples. NMF models have been widely used in practice in areas including audio signal analysis, computational biology and recommender systems. The Doubly Sparse NMF (DS-NMF) model is an extension of NMF that imposes sparsity in both of the factored matrices to increase the interpretability of the discovered factors. The goal of this research is to derive and implement a structured stochastic variational inference algorithm for the DS-NMF model, and assess its performance on simulated data.

1 Introduction

Non-negative Matrix Factorisation (NMF) aims to decompose an observed non-negative data matrix $\mathbf{X}$ into a product of two low-dimensional non-negative matrices of factor loadings $\mathbf{W}$ and activations $\mathbf{H}$. This is achieved by identifying a set of latent variables called *factors*, which represent the intrinsic features of the data. Each observation (column) in $\mathbf{X}$ can be viewed as a non-negative linear combination of the factors (columns of $\mathbf{W}$), weighted by the columns of $\mathbf{H}$. Variants of NMF have been used in a wide range of tasks, including audio source separation (Liang et al., 2013; Liang and Hoffman, 2014), analysis of gene expression in single cells (Stein-O'Brien et al., 2019), and collaboratively filtered recommender systems (Bobadilla et al., 2018). A particular variant is *Sparse NMF* (S-NMF) that poses sparsity on the activations $\mathbf{H}$. In applications where only a subset of samples is associated with each factor, S-NMF can improve the interpretability of the factors by capturing the sparse structure in $\mathbf{H}$. Liang et al. (2013) and Liang and Hoffman (2014) employ S-NMF in their audio source separation tasks. *Doubly Sparse NMF* (DS-NMF) is an extension of S-NMF that imposes additional sparsity on the factor loadings $\mathbf{W}$ to capture the potential sparse structure in $\mathbf{W}$. Xuan et al. (2018) use this model for co-clustering of word-document data.

The focus of this research is to derive and implement Bayesian inference techniques for DS-NMF. Bayesian inference aims to find the posterior distribution of model parameters from observed data. In this report, I used Variational Inference (VI) (Blei et al., 2016) that casts the inference problem as an optimisation problem of finding a variational distribution that closely approximates the posterior distribution. Mean-field VI (Blei et al., 2016; Hoffman et al., 2013) approximates the posterior distribution using the variational distribution where the parameters are mutually independent. This makes the computation process convenient but is a drawback for models where breaking the dependency structure is critical. To address this issue, Hoffman and Blei (2014) developed Structured Stochastic Variational Inference (SSVI) that restores the dependencies among parameters in the approximating variational distribution using a stochastic optimisation approach. Liang and Hoffman (2014) derived an SSVI algorithm for S-NMF. In this report, I will extend this work by developing an SSVI algorithm for DS-NMF, which, to my knowledge, has not been developed before. I will assess the performance of the SSVI algorithm for DS-NMF

and compare it with the SSVI algorithm for S-NMF where appropriate.

1.1 Related Work

There has been several related works that explore inference techniques for S-NMF and DS-NMF models. Liang et al. (2013) developed a Laplace approximation variational inference algorithm for S-NMF to handle non-conjugate models. But their work suffers from the independent assumption of VI. Liang and Hoffman (2014) improve on this and provide an SSVI framework for S-NMF. Xuan et al. (2018) present a Bayesian inference framework for DS-NMF. However, they use a sampling-based approximate Bayesian inference technique called Markov Chain Monte Carlo (MCMC). Finally, Li et al. (2020) present a VI framework for DS-NMF. In contrast to my approach, they make the independence assumption of parameters, and also make the sparsity controlling parameters shared between $\mathbf{W}$ and $\mathbf{H}$.

Statement of authorship. Under the direction of my supervisor, I conducted the literature review, derived and implemented the methods, performed the simulation study, and wrote this report. My supervisor oversaw the overall direction of this research project, checked my derivations, helped set up the simulation study, and proofread this report.

2 The DS-NMF model

In this report, I work with the DS-NMF model with Poisson likelihood. Poisson likelihood is a natural choice as I aim to analyse single-cell RNA-seq data which is count data, and it corresponds to the widely-used Kullback-Leibler divergence loss function (Liang and Hoffman, 2014). The model is formulated as:

$$\begin{aligned} \mathbb{E}[\mathbf{X}] &= \left(\mathbf{W} \odot \mathbf{S}^{\mathbf{W}}\right)\left(\mathbf{H} \odot \mathbf{S}^{\mathbf{H}}\right), \\ X_{ft} &\sim \text{Poisson}\left(\sum_k W_{fk} S^W_{fk} H_{kt} S^H_{kt}\right) \end{aligned} \tag{1}$$

where $\odot$ denotes the Hadamard (element-wise) product. Here, $\mathbf{X} \in \mathbb{N}_+^{F \times T}$ represents the input data of non-negative counts, with rows and columns corresponding to features and samples respectively. $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ represents the unmasked latent factor loadings where each column represents the relative weights of each feature for the factors. $\mathbf{S}^{\mathbf{W}} \in \{0,1\}^{F \times K}$ represents the binary mask for $\mathbf{W}$. $\mathbf{H} \in \mathbb{R}_+^{K \times T}$ represents the unmasked latent activations where each row represents expression of the factors in each observation. $\mathbf{S}^{\mathbf{H}} \in \{0,1\}^{K \times T}$ represents the binary mask for $\mathbf{H}$.

I adopt the prior specifications for S-NMF from Liang and Hoffman (2014) but with additional priors on the elements of $\mathbf{S}^{\mathbf{W}}$ which are new in DS-NMF. Beta process priors are placed on $\mathbf{S}^{\mathbf{W}}$ and $\mathbf{S}^{\mathbf{H}}$ as they remove the need

to set the precise number of factors K which is typically unknown (Paisley and Carin, 2009). Theoretically, this allows the inference algorithm to find the appropriate number for K, given that a sufficiently large value of K is inputted such that it is larger than the true K. I adopt the finite approximation to the beta process from Liang et al. (2013):

$$S_{fk}^W \sim \text{Bernoulli}(\pi_k^W), \quad \pi_k^W \sim \text{Beta}\left(\frac{a_0^W}{K}, \frac{b_0^W(K-1)}{K}\right)$$
$$S_{kt}^H \sim \text{Bernoulli}(\pi_k^H), \quad \pi_k^H \sim \text{Beta}\left(\frac{a_0^H}{K}, \frac{b_0^H(K-1)}{K}\right)$$

In this formulation, π_k^W and π_k^H are factor-specific parameters that control the sparsity in factor loadings and activations respectively. The rest of the priors are as follows:

$$W_{fk} \sim \text{Gamma}(a, b), \quad H_{kt} \sim \text{Gamma}(c, d)$$

The choice of gamma priors for factor loadings $\mathbf{W}$ and activations $\mathbf{H}$ ensures the non-negativity constraint to be satisfied. Figure 1 depicts the full model.

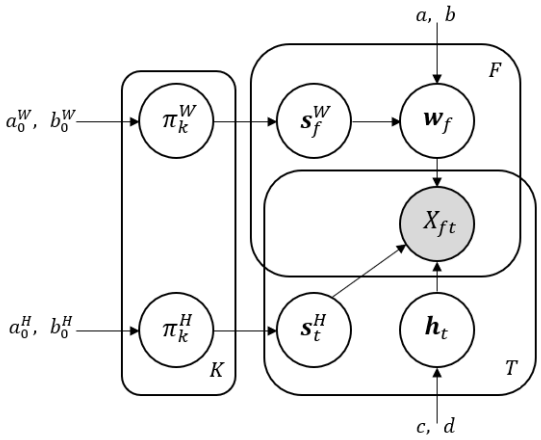

Figure 1: Graphical model representation of DS-NMF. Observed variable is shaded in grey. Hidden variables are unshaded (white). A variable variable a depends on the value of another variable b if there exists an incoming directed edge from b to a. The value in the lower right corner of each plate denotes the replication number.

I introduce auxiliary random variable $\mathbf{Z} \in \mathbb{N}_+^{F \times T \times K}$ similarly to Liang and Hoffman (2014).

$$Z_{ftk} \sim \text{Poisson}\left(W_{fk} S_{fk}^W H_{kt} S_{kt}^H\right)$$
$$X_{ft} = \sum_k Z_{ftk}$$

This will enable the model to enjoy the convenient additive property of Poisson random variables and make the derivation of SSVI inference algorithm more convenient.

3 Bayesian Inference for DS-NMF

In this section, I will first review variational inference (Blei et al., 2016) and structured stochastic variational inference (Hoffman and Blei, 2014), and then present a structured stochastic variational inference algorithm for DS-NMF.

3.1 Variational Inference

Variational Inference (VI) refers to a framework for approximate Bayesian inference algorithms that approximate posterior distributions using an optimisation approach (Blei et al., 2016). Let $\boldsymbol{\theta}$ denote the model parameters and $\mathbf{X}$ denote the data. VI approximates the target posterior distribution with the variational distribution $q^*(\boldsymbol{\theta})$ from a specified family of distributions $\mathcal{Q}$ such that the Kullback-Leibler (KL) divergence between $q^*(\boldsymbol{\theta})$ and target posterior $p(\mathbf{X} \,|\, \boldsymbol{\theta})$ is minimised:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\min}\ \mathrm{KL}\left(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\mathbf{X})\right) = \underset{q(\boldsymbol{\theta})\in\mathcal{Q}}{\arg\min}\ \mathbb{E}_q\left[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{X})\right]$$

The KL divergence cannot be computed directly. But, the problem of minimising the KL divergence is equivalent to maximising the Evidence Lower BOund (ELBO):

$$\mathcal{L} := \mathbb{E}_q\left[\log p(\mathbf{X}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\right] \tag{2}$$

The derivation of equation (2) can be found in the Appendix. The maximisation problem is often solved by the Coordinate Ascent Variational Inference (CAVI) algorithm (Blei et al., 2016). However, CAVI is restrictive as it requires two assumptions: First, the model is conditionally conjugate. Second, the choice of $\mathcal{Q}$ is the *mean-field variational family* where the model parameters $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^m$ are mutually independent in q, that is, the variational distribution is written as a product of independent variational factors:

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{m} q_i(\theta_i).$$

However, the DS-NMF model is not conditionally conjugate. Moreover, it is desirable to capture the dependency between $\mathbf{W}$ ($\mathbf{H}$) and $\mathbf{S}^{\mathbf{W}}$ ($\mathbf{S}^{\mathbf{H}}$) in q. Thus, I develop a structured stochastic variational inference algorithm for DS-NMF.

3.2 Structured Stochastic Variational Inference

Structured Stochastic Variational Inference (SSVI) uses gradient-based stochastic optimisation to solve the problem of maximising the ELBO, while allowing arbitrary dependencies between the *global* and *local* variables (Hoffman and Blei, 2014). Here, global variables are model parameters that possibly govern any of the data, and local hidden

variables are those only associated with specific unit of observation. The word *stochastic* comes from the fact that SSVI only computes a noisy gradient of the ELBO, $\nabla\hat{\mathcal{L}}$, such that:

$$\mathbb{E}_q\left[\nabla\hat{\mathcal{L}}\right] = \nabla\mathcal{L}$$

and uses stochastic optimisation (Robbins and Monro, 1951). The word *structured* comes from the additional structure that is posed on the mean-field variational family, thereby relaxing the independent assumption. Therefore, it allows for an arbitrary dependency structure between local and global variables in the variational distribution.

3.3 Structured Stochastic Variational Inference for DS-NMF

Following the SSVI framework, I divide the model parameters of DS-NMF into local variables $\{\mathbf{Z},\, \mathbf{S^W},\, \mathbf{S^H}\}$ and global variables $\{\mathbf{W},\, \mathbf{H},\, \boldsymbol{\pi^W},\, \boldsymbol{\pi^H}\}$. The posterior distribution is approximated using the following structured variational distribution:

$$\begin{aligned} p(\mathbf{Z},\, \mathbf{W},\, \mathbf{H},\, \mathbf{S^W},\, \mathbf{S^H},\, \boldsymbol{\pi^W},\, \boldsymbol{\pi^H}\,|\mathbf{X}) &\approx q(\mathbf{Z},\, \mathbf{W},\, \mathbf{H},\, \mathbf{S^W},\, \mathbf{S^H},\, \boldsymbol{\pi^W},\, \boldsymbol{\pi^H}) \\ &= \left[\prod_k q(\boldsymbol{w}_k)q(\boldsymbol{h}_k)q(\pi_k^W)q(\pi_k^H)\right] q(\mathbf{Z},\, \mathbf{S^W},\, \mathbf{S^H}|\mathbf{W},\, \mathbf{H},\, \boldsymbol{\pi^H},\, \boldsymbol{\pi^W}) \end{aligned}$$

where the variational distribution on global variables are fully factorised:

$$q(\boldsymbol{w}_k) = \prod_f q(W_{fk}), \qquad q(\boldsymbol{h}_k) = \prod_t q(H_{kt})$$

with

$$\begin{aligned} q(W_{fk}) &= \text{Gamma}(\nu_{fk}^W,\, \rho_{fk}^W), \qquad q(H_{kt}) = \text{Gamma}(\nu_{kt}^H,\, \rho_{kt}^H) \\ q(\pi_k^W) &= \text{Beta}(\alpha_k^{\pi^W},\, \beta_k^{\pi^W}), \qquad q(\pi_k^H) = \text{Beta}(\alpha_k^{\pi^H},\, \beta_k^{\pi^H}). \end{aligned}$$

The ELBO in equation (2) can be re-written as:

$$\mathcal{L} = \mathbb{E}_q\left[\log\frac{p(\mathbf{W},\mathbf{H},\boldsymbol{\pi^W},\boldsymbol{\pi^H})}{q(\mathbf{W},\mathbf{H},\boldsymbol{\pi^W},\boldsymbol{\pi^H})}\right] + \mathbb{E}_q\left[\log\frac{p(\mathbf{X},\mathbf{Z},\mathbf{S^W},\mathbf{S^H}|\mathbf{W},\mathbf{H},\boldsymbol{\pi^W},\boldsymbol{\pi^H})}{q(\mathbf{Z},\mathbf{S^W},\mathbf{S^H}|\mathbf{W},\mathbf{H},\boldsymbol{\pi^W},\boldsymbol{\pi^H})}\right]. \tag{3}$$

The ELBO can be optimised by maximising the *local ELBO*, which is the second term in equation (3), for any particular value of the global variables. The idea is to sample global variables from current variational distributions and use them to maximise the local ELBO with respect to local variables. Then, I update variational parameters for the global variables according to a gradient step. The local ELBO achieves the optimum if the conditional variational distribution equals the exact conditional, i.e. $q(\mathbf{Z},\mathbf{S^W},\mathbf{S^H}|\mathbf{W},\mathbf{H},\boldsymbol{\pi^W},\boldsymbol{\pi^H}) = p(\mathbf{Z},\mathbf{S^W},\mathbf{S^H}|\mathbf{W},\mathbf{H},\boldsymbol{\pi^W},\boldsymbol{\pi^H})$. To compute a noisy gradient for the global variational parameters, SSVI algorithm only requires a sample for local variables from the conditional variational distribution. Thus, I use an MCMC algorithm called collapsed gibbs sampler (Liu, 1994) to sample $\mathbf{S^W}$ and $\mathbf{S^H}$ by marginalising out $\mathbf{Z}$ from the exact conditional distribution.

The full SSVI algorithm for the DS-NMF model is described in Algorithm 1. The derivations of the collapsed gibbs sampler and global variational parameter updates can be found in the Appendix.

Algorithm 1: SSVI for the DS-NMF model

Input: $\mathbf{X}$, a, b, c, d, a_0^W, b_0^W, a_0^H, b_0^H, K

Output: variational factors that approximate the posterior distribution for each global parameter

Randomly initialise global variational parameters: $\{\boldsymbol{\nu^W}, \boldsymbol{\rho^W}, \boldsymbol{\nu^H}, \boldsymbol{\rho^H}, \boldsymbol{\alpha^{\pi^W}}, \boldsymbol{\beta^{\pi^W}}, \boldsymbol{\alpha^{\pi^H}}, \boldsymbol{\beta^{\pi^W}}\}$;

for *i = 1, 2, ...* **do**

Sample $W_{fk}^{(i)} \sim \text{Gamma}(\nu_{fk}^W, \rho_{fk}^W)$;

Sample $H_{kt}^{(i)} \sim \text{Gamma}(\nu_{kt}^H, \rho_{kt}^H)$;

Sample $\pi_k^{W(i)} \sim \text{Beta}(\alpha_k^{\pi^W}, \beta_k^{\pi^W})$;

Sample $\pi_k^{H(i)} \sim \text{Beta}(\alpha_k^{\pi^H}, \beta_k^{\pi^H})$;

Sample $S_{fk}^{W(i)}$ and $S_{kt}^{H(i)}$ using collapsed gibbs sampler and compute $\phi_{ftk}^{(i)} = \frac{W_{fk}^{(i)} S_{fk}^{W(i)} H_{kt}^{(i)} S_{kt}^{H(i)}}{\sum_l W_{fl}^{(i)} S_{fl}^{W(i)} H_{lt}^{(i)} S_{lt}^{H(i)}}$;

Set step-size $\eta^{(i)} = i^{-0.5}$ and perform global variational parameter updates:

$$\nu_{fk}^W \leftarrow (1-\eta^{(i)})\nu_{fk}^W + \eta^{(i)}(a + \sum_t X_{ft}\phi_{ftk}^{(i)});$$

$$\rho_{fk}^W \leftarrow (1-\eta^{(i)})\rho_{fk}^W + \eta^{(i)}(b + S_{fk}^{W(i)} \sum_t H_{kt}^{(i)} S_{kt}^{H(i)});$$

$$\nu_{kt}^H \leftarrow (1-\eta^{(i)})\nu_{kt}^H + \eta^{(i)}(c + \sum_f X_{ft}\phi_{ftk}^{(i)});$$

$$\rho_{kt}^H \leftarrow (1-\eta^{(i)})\rho_{kt}^H + \eta^{(i)}(d + S_{kt}^{H(i)} \sum_f W_{fk}^{(i)} S_{fk}^{W(i)});$$

$$\alpha_k^{\pi^W} \leftarrow (1-\eta^{(i)})\alpha_k^{\pi^W} + \eta^{(i)}(\frac{a_0^W}{K} + \sum_f S_{fk}^{W(i)});$$

$$\beta_k^{\pi^W} \leftarrow (1-\eta^{(i)})\beta_k^{\pi^W} + \eta^{(i)}(\frac{b_0^W(K-1)}{K} + F - \sum_f S_{fk}^{W(i)});$$

$$\alpha_k^{\pi^H} \leftarrow (1-\eta^{(i)})\alpha_k^{\pi^H} + \eta^{(i)}(\frac{a_0^H}{K} + \sum_t S_{kt}^{H(i)});$$

$$\beta_k^{\pi^H} \leftarrow (1-\eta^{(i)})\beta_k^{\pi^H} + \eta^{(i)}(\frac{b_0^H(K-1)}{K} + T - \sum_t S_{kt}^{H(i)});$$

end

4 Simulation Study

I now compare the performance of my SSVI algorithm for the DS-NMF model, denoted *DS-SSVI*, with the existing SSVI algorithm for the S-NMF model by Liang and Hoffman (2014), denoted *S-SSVI*. I focus on the performances of algorithms in capturing the possible sparse structures in masked factor loadings $\mathbf{W} \odot \mathbf{S^W}$ and masked activations

$\mathbf{H} \odot \mathbf{S^H}$. It is expected that DS-SSVI would flexibly capture the sparse structures in both factor loadings and activations, but S-SSVI would only be able to capture the sparse structure in activations. I perform a simulation study where both masked factor loadings and masked activations are sparse.

4.1 A doubly sparse example

I compare the performances of DS-SSVI and S-SSVI on a doubly sparse example, where both masked factor loadings and activations have true sparse structures. My goal is to verify that DS-SSVI is able to capture the sparse structure in the new binary mask $\mathbf{S^W}$, in addition to $\mathbf{S^H}$. I expect DS-SSVI and S-SSVI to have similar performance in inferring the masked activations $\mathbf{H} \odot \mathbf{S^H}$, but DS-SSVI to outperform S-SSVI in inferring the masked factor loadings $\mathbf{W} \odot \mathbf{S^W}$.

The simulated dataset consists of $F = 300$ features over $T = 300$ samples, explained by $K = 4$ factors. The sparsity parameters are set to $\boldsymbol{\pi^W} = (0.3, 0.2, 0.3, 1)$ and $\boldsymbol{\pi^H} = (0.4, 0.5, 0.45, 1)$, leading to the mix of sparse and dense factors. Entries in $\mathbf{W}$ and $\mathbf{H}$ are independently generated from four different Gamma distributions with shape parameters 1.6, 1.4, 1.2, 1 and scale parameter 1, according to their associated factor. This is to enable clusters to be around different values in the mean matrix. The mean matrix, $\mathbb{E}[\mathbf{X}]$, where dataset $\mathbf{X}$ will be simulated from, is computed by $\left(\mathbf{W} \odot \mathbf{S^W}\right)\left(\mathbf{H} \odot \mathbf{S^H}\right)$. Figure 3 in the Appendix shows the true simulated structures. 10 independent datasets are generated from $\mathbb{E}[\mathbf{X}]$, where 3 VI trials with different initialisations are run on each observation. For each dataset, following Liang and Hoffman (2014), I select the VI trial with the highest local likelihood $p(\mathbf{S^W}, \mathbf{S^H}, \mathbf{Z}, \mathbf{X}|\mathbf{W}, \mathbf{H}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H})$ for inference. In VI trials, the sparsity hyperparameters (parameters in the prior) for $\boldsymbol{\pi^W}$ and $\boldsymbol{\pi^H}$ are set to $a_0^W = b_0^W = a_0^H = b_0^W = 1$, corresponding to uniform priors. The gamma entries are set to be dense with hyperparameters $a = b = c = d = 5$, to encourage the sparse structures to be captured by the binary masks $\mathbf{S^W}$ and $\mathbf{S^H}$, not by making entries in $\mathbf{W}$ and $\mathbf{H}$ tiny. Also, the true number of factors $K = 4$ is inputted into the algorithm to verify that the algorithms work appropriately given the correct number of factors. Finally, both algorithms are run for 100 iterations.

To quantitatively assess the performances of DS-SSVI and S-SSVI, I evaluate the accuracy of inferred binary masks $\mathbf{S^W}$ and $\mathbf{S^H}$ using accuracy, defined as the proportion of correctly inferred entries. $\mathbf{S^W}$ and $\mathbf{S^H}$ are inferred by taking the mean of samples generated from the collapsed gibbs sampler after Algorithm 1 has finished running, and rounding the entries to 0 or 1. The accuracy of estimating masked factor loadings $\mathbf{W} \odot \mathbf{S^W}$ and masked activations $\mathbf{H} \odot \mathbf{S^H}$ are evaluated using relative root mean squared error:

$$\text{RRMSE}(\hat{A},\, A) := \sqrt{\frac{\sum(\hat{A}_{ij} - A_{ij})^2}{\sum A_{ij}^2}}$$

Table 1 lists the median accuracy and RRMSE values obtained by DS-SSVI and S-SSVI with standard deviation in the parenthesis. DS-SSVI yields much better (lower) RRMSE for masked activations $\mathbf{W} \odot \mathbf{S^W}$ than S-SSVI, hence it verifies that DS-SSVI does better in estimating the true sparse structure of $\mathbf{W} \odot \mathbf{S^W}$ in a doubly sparse

setting. Also, DS-SSVI does about as good as SSVI in estimating $\mathbf{S^H}$ and $\mathbf{H} \odot \mathbf{S^H}$ in terms of accuracy and RRMSE. Note that for S-SSVI there is no accuracy value obtained for $\mathbf{S^W}$ as S-NMF does not have the binary mask $\mathbf{S^W}$ in its model. The inferred binary mask $\mathbf{S^W}$ from DS-SSVI is shown in Figure 4 in the Appendix.

Table 1: DS-SSVI and S-SSVI results on a doubly sparse example.

	Accuracy: $\mathbf{S^W}$	RRMSE: $\mathbf{W} \odot \mathbf{S^W}$	Accuracy: $\mathbf{S^H}$	RRMSE: $\mathbf{H} \odot \mathbf{S^H}$
DS-SSVI	0.834 (0.041)	0.290 (0.041)	0.735 (0.053)	0.542 (0.272)
S-SSVI	-	0.769 (0.055)	0.745 (0.052)	0.558 (0.075)

Figure 2 visually shows the inferred masked factor loadings by DS-SSVI and S-SSVI. S-SSVI fails to capture the sparse structure in the true masked factor loadings. S-SSVI insufficiently shrinks the zero values compared with DS-SSVI which captures the sparse structure of $\mathbf{W} \odot \mathbf{S^W}$ relatively well. Although with some noise, DS-SSVI can identify the association between factors and features by interpreting the sparse entries as non-related.

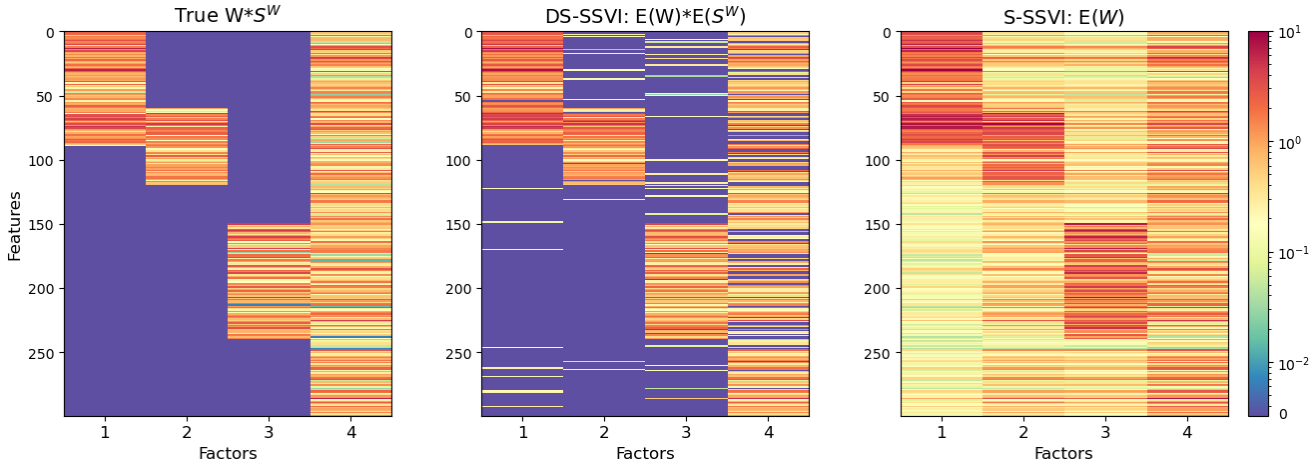

Figure 2: True simulated masked factor loadings $\mathbf{W} \odot \mathbf{S^W}$ (left), estimated masked factor loadings from DS-SSVI (middle), estimated factor loadings from S-SSVI (right).

4.2 Trials with larger values of K

I now compare the performances of DS-SSVI and S-SSVI when the input value of K is larger than the true number of factors $K = 4$. According to their model specifications, the algorithms should attempt to silence insignificant factors where the proportion of non-zero entries are close to 0. 10 trials of DS-SSVI and S-SSVI are run on both the doubly sparse and sparse-activations examples, with input $K = 5$ and $K = 10$.

4.2.1 Doubly sparse example

Table 2 shows the median RRMSE and standard deviation of estimated mean matrix $\mathbb{E}[\hat{\mathbf{X}}]$ for DS-SSVI and S-SSVI using $K = 5$ and $K = 10$ as the input number of factors. I use the RRMSE values of the estimation of mean matrix $\mathbb{E}[\hat{\mathbf{X}}]$ for comparison. I discuss the reason for using this evaluation metric in the Discussion section. For both inputs of $K = 5$ and $K = 10$, DS-SSVI achieves lower RRMSE with lower standard deviation. Thus, DS-SSVI can be considered more accurate and reliable than S-SSVI in the case of doubly sparse example, in the sense that it can more closely estimate the true mean matrix.

Table 2: DS-SSVI and S-SSVI results with input $K = 5$ and $K = 10$ on the doubly sparse example.

Input K	Algorithm	RRMSE $\mathbb{E}[\hat{\mathbf{X}}]$
5	DS-SSVI	0.099 (0.017)
	S-SSVI	0.180 (0.051)
10	DS-SSVI	0.669 (0.001)
	S-SSVI	0.678 (0.003)

5 Discussion

In this section, I address the identified limitations of DS-SSVI, and provide possible directions for future work.

Capturing sparsity in masked activations. A common trend seen in the simulation study was that although DS-SSVI obtained reasonable evaluation metric values for estimating the activations matrix, it did not seem to successfully capture the true sparse structure in visual inspections. One reason could be that instead of the binary mask $\mathbf{S^H}$ capturing sparse entries by masking them out with 0's, actually the unmasked activations $\mathbf{H}$ often tried to capture the sparse pattern by putting small values in $\mathbf{H}$ where corresponding entries in $\mathbf{S^H}$ would still have values of 1. Figure 5 in the Appendix illustrates an example. Here, DS-SSVI failed to learn the sparse pattern of factors 1 and 2 in the binary mask $\mathbf{S^H}$, and instead, it placed very small values (between about 0.1 and 0.4) in the unmasked activations $\mathbf{H}$ corresponding to those entries that should have been masked out. A possible resolution would be to employ the spike and slab prior in modelling the relations between $\mathbf{W}$ with $\mathbf{S^W}$, and $\mathbf{H}$ with $\mathbf{S^H}$. This would impose a more explicit dependency structure such that the entries in $\mathbf{W}$ ($\mathbf{H}$) are non-zero only if corresponding entries in $\mathbf{S^W}$ ($\mathbf{S^H}$) are non-zero. This could improve the interpretibility and performance of the algorithm, but the derivation could become trickier.

Multiple inferred factors representing one true factor. One advantage of the existing S-SSVI algorithm presented by Liang and Hoffman (2014) was its ability to effectively silence insignificant factors and arrive at an

appropriate number for K, even when an arbitrarily large K is inputted into the algorithm. In the simulation study, it was seen that S-SSVI was not able to obtain the correct number of factors $K = 4$ when inputs of 5 and 10 were given. Neither was DS-SSVI able to appropriately reduce down to the correct number of factors. A possible explanation is that two or more significant factors can be produced by the algorithm such that together they estimate one true factor well. Hence, this results in the tendency to keep the larger number of K. Figure 6 in the Appendix shows an example where factors 4 and 5 in the estimated matrix from DS-SSVI seem to together correspond to factor 4 in the true structure. This is why only the RRMSE of estimated mean matrix was used for comparison, as the accuracy and RRMSE of binary masks, masked factor loadings and activations cannot be computed when the values for K do not match between two matrices. Further work would be to investigate a way to interpret the decomposed factors appropriately.

Using mean-field approximation instead of collapsed gibbs sampler. In DS-SSVI, the collapsed gibbs sampler was used to sample $\mathbf{S^W}$ and $\mathbf{S^H}$. This was to obtain an unbiased estimate of the conditional expectation of natural parameters in the local likelihood (see the Global parameter updates section in the Appendix). Instead, the mean-field approximation assumption between the local parameters can be posed, that is, $\mathbf{S^W}, \mathbf{S^H}$, and $\mathbf{Z}$ given the global parameters are independent. This approach would still keep the dependency structure between local and global variables, but possibly make the conditional expectation to be computed in closed-form (Hoffman and Blei, 2014). Although the algorithm's performance would likely be reduced compared to using the collapsed gibbs sampler, there would be a significant reduction in running-time. The trade-off between run-time and performance would could be considered for future work.

Laplace approximation variational inference. In this report, I focused on structured stochastic variational inference for the DS-NMF model, which relaxed the independent assumption of variational factors in the mean-field approximation. Another approach is to use Laplace approximation variational inference (Liang et al., 2013), where non-conjugacy of the model is directly dealt by using the Laplace method to approximate variational factors. This approach does not restore the dependency structure between global and local variables, but algorithm run-time would likely be much faster than DS-SSVI. Again, the trade-off between performance and run-time would need to be investigated.

6 Conclusion

In this report, I reviewed the DS-NMF model, an extension of NMF, where sparsity is imposed on both of the factored matrices. This model is appropriate for use when the factors of the data are expected to be associated with subsets of both the samples and features. Structured stochastic variational inference framework was used in inferring the model parameters. Theoretically, the algorithm should flexibly adapt to different combinations of sparse structures, and infer the matrices accordingly. The simulation study confirmed that DS-SSVI performs

better than S-SSVI in identifying the underlying factors when both the true factor loadings and activations are sparse.

Acknowledgements I would like to sincerely thank my supervisor, Heejung Shim, for her consistent support and guidance throughout the course of this research project. I would also like to thank the Australian Mathematical Scicnes Institute for providing the opportunity to experience research.

Appendix

Derivation of equation (2).

$$\begin{aligned}
\mathrm{KL}\left(q(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta}|\mathbf{X})\right) &= \mathbb{E}_q\left[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{X})\right] \\
&= \mathbb{E}_q\left[\log q(\boldsymbol{\theta}) - \log p(\mathbf{X},\, \boldsymbol{\theta}) + \log p(\mathbf{X})\right] \\
&= -\mathbb{E}_q\left[\log p(\mathbf{X}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\right] + \log p(\mathbf{X}) \\
&= -\mathcal{L} + \log p(\mathbf{X})
\end{aligned}$$

Since $\log p(\mathbf{X})$ is constant with respect to q, minimising the KL divergence is equivalent to maximising the ELBO ($\mathcal{L}$).

Derivation of Algorithm 1.

Collapsed gibbs sampler.

The construction of the auxiliary variables $\mathbf{Z}$ makes it easy to calculate the conditional expectation of Z_{ftk}. Hence, I marginalise Z_{ftk} out from $p(S^W_{fk},\, S^H_{kt}|\mathbf{X},\, \mathbf{W},\, \mathbf{H},\, \boldsymbol{\pi}^{\boldsymbol{W}},\, \boldsymbol{\pi}^{\boldsymbol{H}})$ and only sample S^W_{fk} and S^H_{kt}.

$$p(Z_{ftk}, S^W_{fk},\, S^H_{kt}|\mathbf{X},\, \mathbf{W},\, \mathbf{H},\, \boldsymbol{\pi}^{\boldsymbol{W}},\, \boldsymbol{\pi}^{\boldsymbol{H}}) \propto p(S^W_{fk},\, S^H_{kt}|\mathbf{X},\, \mathbf{W},\, \mathbf{H},\, \boldsymbol{\pi}^{\boldsymbol{W}},\, \boldsymbol{\pi}^{\boldsymbol{H}})$$

First, I derive the active (non-sparse entries) proportion of S^W_{fk} by computing the following quantities, using the marginalised conditional distribution:

$$\begin{aligned}
&\mathbb{P}(S^W_{fk} = 1|S^W_{f,\neg k}, \boldsymbol{x}_f,\, \boldsymbol{w}_f,\, \mathbf{H},\, \mathbf{S}^{\mathbf{H}},\, \boldsymbol{\pi}^{\boldsymbol{W}},\, \boldsymbol{\pi}^{\boldsymbol{H}}) \\
&\qquad \propto \mathbb{P}(S^W_{fk} = 1,\, S^W_{f,\neg k},\, \boldsymbol{x}_f,\, \boldsymbol{w}_f,\, \mathbf{H},\, \mathbf{S}^{\mathbf{H}},\, \boldsymbol{\pi}^{\boldsymbol{W}},\, \boldsymbol{\pi}^{\boldsymbol{H}}) \\
&\qquad \propto \mathbb{P}(S^W_{fk} = 1|\boldsymbol{\pi}^{\boldsymbol{W}}) \cdot \mathbb{P}(\boldsymbol{x}_f|\mathbf{H},\, \boldsymbol{w}_f,\, S^W_{f,\neg k},\, S^W_{fk} = 1,\, \mathbf{S}^{\mathbf{H}}) \\
&\qquad \propto \pi^W_k \cdot \prod_t \left(W_{fk}H_{kt}S^H_{kt} + \sum_{m\neq k} W_{fm}S^W_{fm}H_{mt}S^H_{mt}\right)^{X_{ft}} \exp(-W_{fk}H_{kt}S^H_{kt}) =: P^W_1 \\
&\mathbb{P}(S^W_{fk} = 0|S^W_{f,\neg k}, \boldsymbol{x}_f,\, \boldsymbol{w}_f,\, \mathbf{H},\, \mathbf{S}^{\mathbf{H}},\, \boldsymbol{\pi}^{\boldsymbol{W}},\, \boldsymbol{\pi}^{\boldsymbol{H}}) \\
&\qquad \propto \mathbb{P}(S^W_{fk} = 0|\boldsymbol{\pi}^{\boldsymbol{W}}) \cdot \mathbb{P}(\boldsymbol{x}_f|\mathbf{H},\, \boldsymbol{w}_f,\, S^W_{f,\neg k},\, S^W_{fk} = 0,\, \mathbf{S}^{\mathbf{H}}) \\
&\qquad \propto (1 - pi^W_k) \cdot \prod_t \left(\sum_{m\neq k} W_{fm}S^W_{fm}H_{mt}S^H_{mt}\right)^{X_{ft}} =: P^W_0
\end{aligned}$$

Now, I can sample $S^W_{fk} \sim \text{Bernoulli}(\frac{P^W_1}{P^W_0 + P^W_1})$ after some burn-in period. Similarily to above, I also derive the active

proportion of S^H_{kt}:

$$\begin{aligned}
&\mathbb{P}(S^H_{kt} = 1 | S^H_{\neg k,t}, \boldsymbol{x}_t, \mathbf{W}, \boldsymbol{h}_t, \mathbf{S^W}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H}) \\
&\qquad \propto \mathbb{P}(S^H_{kt} = 1 | \boldsymbol{\pi^H}) \cdot \mathbb{P}(\boldsymbol{x}_t | \mathbf{W}, \boldsymbol{h}_t, S^H_{\neg k.t}, S^H_{kt} = 1, \mathbf{S^W}) \\
&\qquad \propto \pi^H_k \cdot \prod_f \left(W_{fk} S^W_{fk} H_{kt} + \sum_{l \neq k} W_{fl} S^W_{fl} H_{lt} S^H_{lt} \right)^{X_{ft}} \exp(-W_{fk} S^W_{fk} H_{kt}) =: P^H_1 \\
&\mathbb{P}(S^H_{kt} = 0 | S^H_{\neg k,t}, \boldsymbol{x}_t, \mathbf{W}, \boldsymbol{h}_t, \mathbf{S^W}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H}) \\
&\qquad \propto (1 - \pi^H_k) \cdot \prod_f \left(\sum_{l \neq k} W_{fl} S^W_{fl} H_{lt} S^H_{lt} \right)^{X_{ft}} =: P^H_0
\end{aligned}$$

I can then sample $S^H_{kt} \sim \text{Bernoulli}(\frac{P^H_1}{P^H_0 + P^H_1})$ after some burn-in period.

I can calculate the conditional expectation of Z_{ftk} by utilising the property that the conditional of a Poisson distribution is Multinomial. This conditional expectation can be used as a proxy for Z_{ftk}. I have:

$$\boldsymbol{z}_{ft} | X_{ft}, \boldsymbol{w}_f, \boldsymbol{h_t}, \boldsymbol{s}^W_f, \boldsymbol{s}^H_t \sim \text{Multi}(\boldsymbol{z}_{ft}; X_{ft}, \phi_{ft})$$

where $\phi_{ftk} \propto W_{fk} S^W_{fk} H_{kt} S^H_{kt}$. Hence, $\mathbb{E}[Z_{ftk} | X_{ft}, W_{fk}, H_{kt}, S^W_{fk}, S^H_{kt}] = X_{ft} \phi_{ftk}$ and I can use it as a proxy for Z_{ftk} in the global update steps.

Global parameter updates.

In DS-SSVI, an unbiased estimate of the conditional expectation of natural parameters from the local likelihood $p(\mathbf{S^W}, \mathbf{S^H}, \mathbf{Z}, \mathbf{X} | \mathbf{W}, \mathbf{H}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H})$ needs to be computed. This computation becomes convenient as the introduction of auxiliary variables $\mathbf{Z}$ restores conditional conjugacy between parameters in the DS-NMF model in Eq.(1), when conditioning on binary masks $\mathbf{S^W}$ and $\mathbf{S^H}$. This leads to the posterior distribution for each global model parameter being factorised as conjugate pairs with $\mathbf{W}$, $\mathbf{H}$, $\boldsymbol{\pi^W}$, and $\boldsymbol{\pi^H}$ respectively. We have:

$$\begin{aligned}
p(\mathbf{S^W}, \mathbf{S^H}, \mathbf{Z}, \mathbf{X} | \mathbf{W}, \mathbf{H}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H}) &= p(\mathbf{S^W} | \boldsymbol{\pi^W}) p(\mathbf{S^H} | \boldsymbol{\pi^H}) p(\mathbf{Z}, \mathbf{X} | \mathbf{W}, \mathbf{H}, \mathbf{S^W}, \mathbf{S^H}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H}) \\
&= p(\mathbf{S^W} | \boldsymbol{\pi^W}) p(\mathbf{S^H} | \boldsymbol{\pi^H}) p(\mathbf{Z} | \mathbf{W}, \mathbf{H}, \mathbf{S^W}, \mathbf{S^H}, \boldsymbol{\pi^W}, \boldsymbol{\pi^H})
\end{aligned}$$

For the update of variational parameters for each W_{fk}, only the terms containing W_{fk} are considered:

$$\begin{aligned}
p(\boldsymbol{z}_f|\boldsymbol{w}_f, \mathbf{H}, \boldsymbol{s}_f^W, \mathbf{S^H}) &= \prod_t \prod_k p(Z_{ftk}|\boldsymbol{w}_f, \mathbf{H}, \boldsymbol{s}_f^W, \mathbf{S^H}) \\
&\propto \prod_t \Big[(W_{fk} S_{fk}^W H_{kt} S_{kt}^H)^{Z_{ftk}} \exp(-W_{fk} S_{fk}^W H_{kt} S_{kt}^H) \\
&\times \prod_{m \neq k} (W_{fm} S_{fm}^W H_{mt} S_{mt}^H)^{Z_{ftk}} \exp(-W_{fm} S_{fm}^W H_{mt} S_{mt}^H) \Big] \\
&\propto \prod_t \Big[(W_{fk} S_{fk}^W H_{kt} S_{kt}^H)^{Z_{ftk}} \exp(-W_{fk} S_{fk}^W H_{kt} S_{kt}^H) \Big] \\
&= (W_{fk} S_{fk}^W H_{kt} S_{kt}^H)^{\sum_t Z_{ftk}} \exp(-W_{fk} S_{fk}^W \sum_t H_{kt} S_{kt}^H)
\end{aligned}$$

which is Gamma-distributed with natural parameters $\left[\sum_t Z_{ftk},\ S_{fk}^W \sum_t H_{kt} S_{kt}^H\right]$. Here, $\sum_t X_{ft}\phi_{ftk}$ is substituted for $\sum_t Z_{ftk}$ as an unbiased estimator. Now, plugging these, with the natural parameters for the prior on W, into the standard Robbins-Monro stochastic optimisation update step (Hoffman et al., 2013) yields:

$$\begin{aligned}
\nu_{fk}^W &\leftarrow (1-\eta^{(i)})\nu_{fk}^W + \eta^{(i)}(a + \sum_t X_{ft}\phi_{ftk}^{(i)}) \\
\rho_{fk}^W &\leftarrow (1-\eta^{(i)})\rho_{fk}^W + \eta^{(i)}(b + S_{fk}^{W\,(i)} \sum_t H_{kt}^{(i)} S_{kt}^{H\,(i)})
\end{aligned}$$

For the update of variational parameters for each H_{kt}, only the terms containing H_{kt} are considered:

$$\begin{aligned}
p(\boldsymbol{z}_t|\boldsymbol{h}_t, \mathbf{W}, \mathbf{S^W}, \boldsymbol{s}_t^H) &= \prod_f \prod_k p(Z_{ftk}|\boldsymbol{h}_t, \mathbf{W}, \mathbf{S^W}, \boldsymbol{s}_t^H) \\
&\propto \prod_f \Big[(W_{fk} S_{fk}^W H_{kt} S_{kt}^H)^{Z_{ftk}} \exp(-W_{fk} S_{fk}^W H_{kt} S_{kt}^H) \Big] \\
&= (H_{kt} W_{fk} S_{fk}^W S_{kt}^H)^{\sum_f Z_{ftk}} \exp(-H_{kt} S_{kt}^H \sum_f W_{fk} S_{fk}^W)
\end{aligned}$$

which is Gamma-distributed with natural parameters $\left[\sum_f Z_{ftk},\ S_{kt}^H \sum_f W_{fk} S_{fk}^W\right]$. Similarly to above, the update steps become:

$$\begin{aligned}
\nu_{kt}^H &\leftarrow (1-\eta^{(i)})\nu_{kt}^H + \eta^{(i)}(c + \sum_f X_{ft}\phi_{ftk}^{(i)}) \\
\rho_{kt}^H &\leftarrow (1-\eta^{(i)})\rho_{kt}^H + \eta^{(i)}(d + S_{kt}^{H\,(i)} \sum_f W_{fk}^{(i)} S_{fk}^{W\,(i)})
\end{aligned}$$

For the update of variational parameters for each π_k^W, only the terms containing π_k^W are considered:

$$\begin{aligned}
p(S_k^W|\pi_W^H) &= \prod_t p(S_{fk}^W|\pi_k^W) \\
&= \prod_f (\pi_k^W)^{S_{fk}^W} (1-\pi_k^W)^{1-S_{fk}^W} \\
&= (\pi_k^W)^{\sum_t S_{fk}^W} (1-\pi_k^W)^{F-\sum_t S_{fk}^W}
\end{aligned}$$

which is Bernoulli-distributed with natural parameters $\left[\sum_f S_{fk}^W,\ F - \sum_f S_{fk}^W\right]$. The update steps become:

$$\alpha_k^{\pi^W} \leftarrow (1-\eta^{(i)})\alpha_k^{\pi^W} + \eta^{(i)}(\frac{a_0^W}{K} + \sum_f S_{fk}^{W^{(i)}})$$
$$\beta_k^{\pi^W} \leftarrow (1-\eta^{(i)})\beta_k^{\pi^W} + \eta^{(i)}(\frac{b_0^W(K-1)}{K} + F - \sum_f S_{fk}^{W^{(i)}})$$

For the update of variational parameters for each π_k^H, only the terms containing π_k^H are considered:

$$\begin{aligned} p(S_k^H|\pi_k^H) &= \prod_t p(S_{kt}^H|\pi_k^H) \\ &= \prod_t (\pi_k^H)^{S_{kt}^H}(1-\pi_k^H)^{1-S_{kt}^H} \\ &= (\pi_k^H)^{\sum_t S_{kt}^H}(1-\pi_k^H)^{T-\sum_t S_{kt}^H} \end{aligned}$$

which is Bernoulli-distributed with natural parameters $\left[\sum_t S_{kt}^H,\ T - \sum_t S_{kt}^H\right]$. The update steps become:

$$\alpha_k^{\pi^H} \leftarrow (1-\eta^{(i)})\alpha_k^{\pi^H} + \eta^{(i)}(\frac{a_0^H}{K} + \sum_t S_{kt}^{H^{(i)}})$$
$$\beta_k^{\pi^H} \leftarrow (1-\eta^{(i)})\beta_k^{\pi^H} + \eta^{(i)}(\frac{b_0^H(K-1)}{K} + T - \sum_t S_{kt}^{H^{(i)}})$$

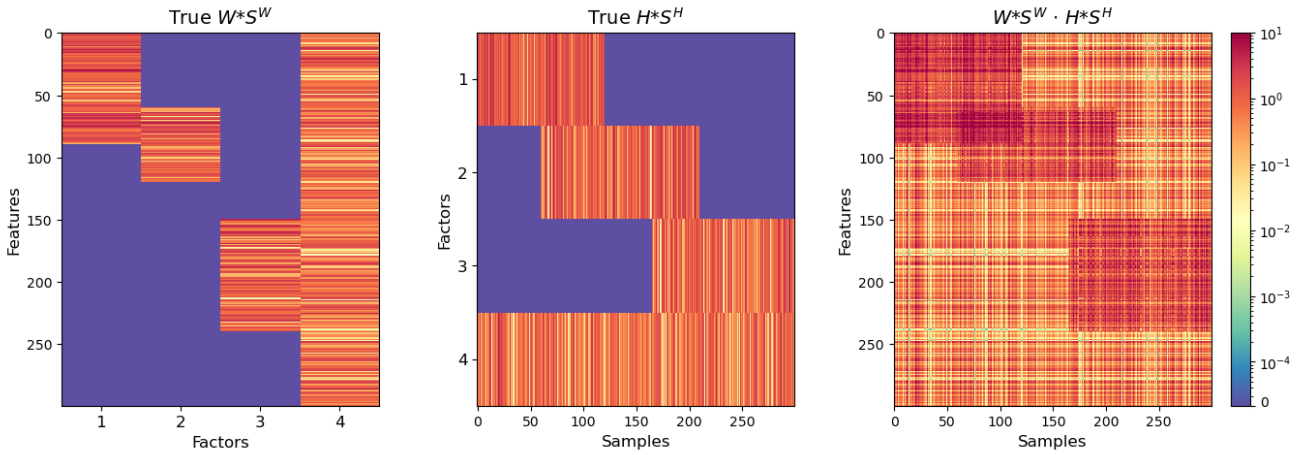

Figure 3: True simulated structures for: masked factor loadings $W \odot S^W$ (left), masked activations $H \odot S^H$ (middle), and mean matrix $\mathbb{E}[\mathbf{X}] = (W \odot S^W)(H \odot S^H)$ (right). Purple entries in the left and middle plots correspond to sparse entries (with values 0).

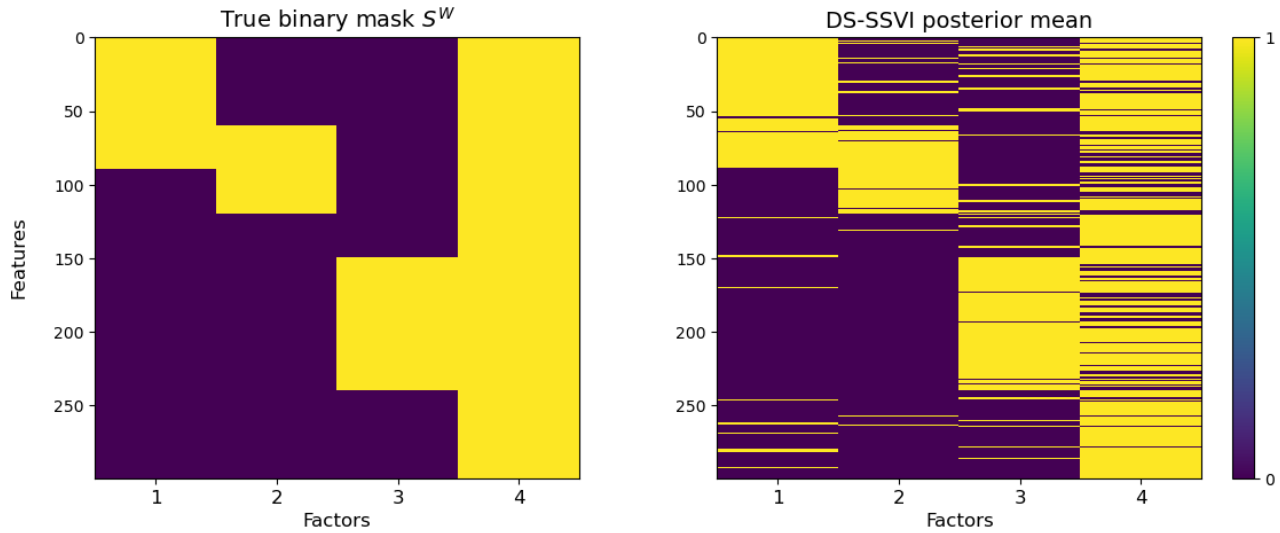

Figure 4: True simulated structure for binary mask $\mathbf{S^W}$ (left), inferred structure of posterior mean of $\mathbf{S^W}$ from DS-SSVI with the highest accuracy of 0.897 (right).

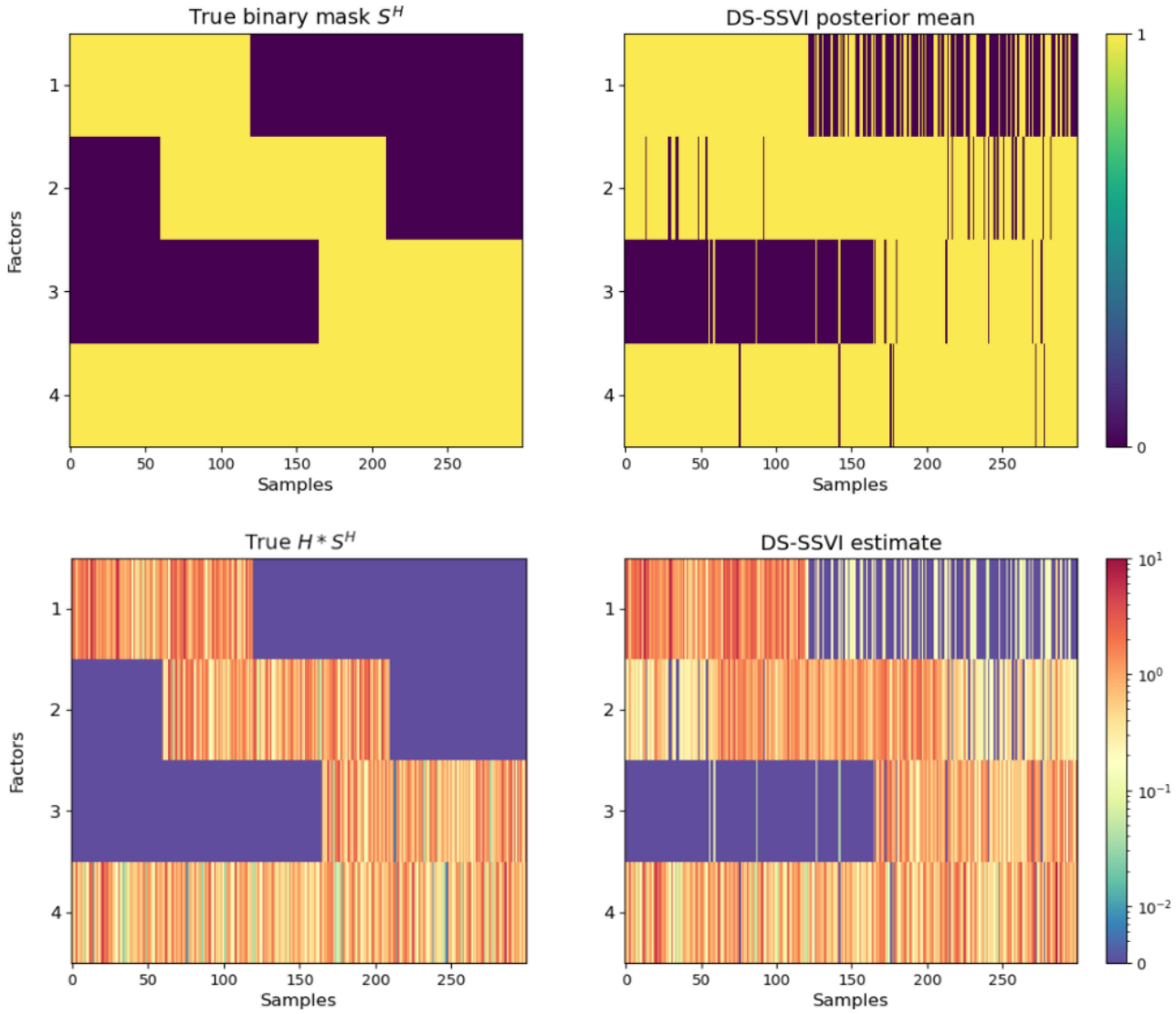

Figure 5: True simulated binary mask $\mathbf{S^H}$ (top-left), inferred binary mask for $\mathbf{S^H}$ (top-right), true masked activations $\mathbf{H} \odot \mathbf{S^H}$ (bottom-left), estimated masked activations obtained by taking an element-wise product between posterior mean of $\mathbf{H}$ and inferred binary mask for $\mathbf{S^H}$ (bottom-right).

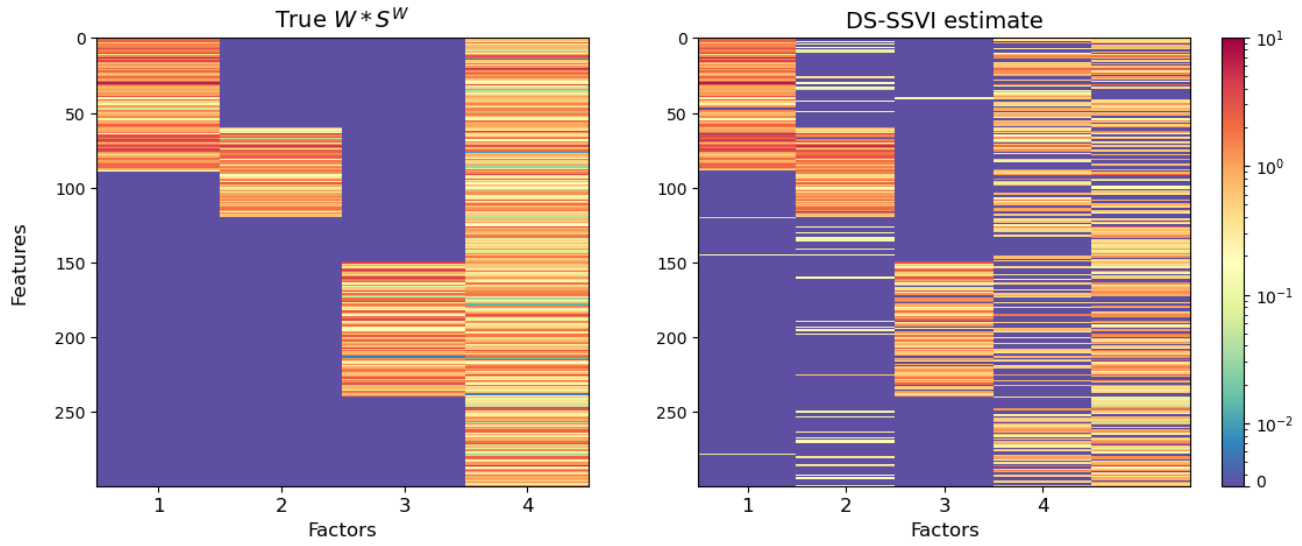

Figure 6: True simulated masked factor loadings $\mathbf{W} \odot \mathbf{S}^{\mathbf{W}}$ (left), estimated masked factor loadings obtained from DS-SSVI with $K = 5$ as input. Factors 4 and 5 on the right are together estimating the true factor 4 in the true structure on the left.

References

D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859 – 877, 2016.

J. Bobadilla, R. Bojorque, A. Hernando Esteban, and R. Hurtado. Recommender systems clustering using bayesian non negative matrix factorization. *IEEE Access*, 6:3549–3564, 2018. doi: 10.1109/ACCESS.2017.2788138.

M. Hoffman and D. Blei. Structured stochastic variational inference. *arXiv: Learning*, 2014.

M. Hoffman, D. Blei, C. Wang, and John W. Paisley. Stochastic variational inference. *ArXiv*, abs/1206.7051, 2013.

C. Li, H. B. Xie, K. Mengersen, X. Fan, R. Y. Da Xu, S. A. Sisson, and S. Van Huffel. Bayesian nonnegative matrix factorization with dirichlet process mixtures. *IEEE Transactions on Signal Processing*, 68:3860–3870, 2020. doi: 10.1109/TSP.2020.3003120.

D. Liang and M. Hoffman. Beta process non-negative matrix factorization with stochastic structured mean-field variational inference. *ArXiv*, abs/1411.1804, 2014.

D. Liang, M. Hoffman, and D. Ellis. Beta process sparse nonnegative matrix factorization for music. In *ISMIR*, 2013.

J. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89:958–966, 1994.

J. W. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *ICML '09*, 2009.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

G. L. Stein-O'Brien, B. S Clark, T. D. Sherman, C. Zibetti, Q. Hu, R. Sealfon, S. Liu, J. Qian, C. Colantuoni, S. Blackshaw, L. Goff, and E. Fertig. Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species. *Cell systems*, 8 5:395–411.e8, 2019.

J. Xuan, J. Lu, G. Zhang, R. Xu, and X. Luo. Doubly nonparametric sparse nonnegative matrix factorization based on dependent indian buffet processes. *IEEE Transactions on Neural Networks and Learning Systems*, 29: 1835–1849, 2018.