A Comparison of Bayesian Inference Techniques for Sparse Factor Analysis

Yong See Foo
School of Mathematics and Statistics
University of Melbourne
Parkville, VIC, 3010, Australia
yongsee.foo@unimelb.edu.au

Heejung Shim
School of Mathematics and Statistics and
Melbourne Integrative Genomics
University of Melbourne
Parkville, VIC, 3010, Australia
heejung.shim@unimelb.edu.au

Abstract

Dimension reduction algorithms aim to discover latent variables which describe underlying structures in high-dimensional data. Methods such as factor analysis and principal component analysis have the downside of not offering much interpretability of its inferred latent variables. Sparse factor analysis addresses this issue by imposing sparsity on its factor loadings, allowing each latent variable to be related to only a subset of features, thus increasing interpretability. Sparse factor analysis has been used in a wide range of areas including genomics, signal processing, and economics. We compare two Bayesian inference techniques for sparse factor analysis, namely Markov chain Monte Carlo (MCMC), and variational inference (VI). VI is computationally faster than MCMC, at the cost of a loss in accuracy. We derive MCMC and VI algorithms and perform a comparison using both simulated and biological data, demonstrating that the higher computational efficiency of VI is desirable over the small gain in accuracy when using MCMC. Our implementation of MCMC and VI algorithms for sparse factor analysis is available at https://github.com/ysfoo/sparsefactor.

1 Introduction

Dimension reduction techniques have been widely used for inferring and explaining an underlying structure in high dimensional data. One of these techniques is *factor analysis*, which linearly maps high dimensional data onto a lower dimensional subspace. This is achieved by finding a set of latent variables, known as *factors*, such that the observed variables may be represented by linear combinations of these factors. The aim of dimension reduction is realised by using a number of factors much smaller than the number of observed variables.

In some applications, it is desirable for each factor to be associated with only a subset of the observed variables. In other words, the factor loadings, which quantify the weighting of each variable on each factor, are expected to be sparse. *Sparse factor analysis* is an extension of factor analysis that allows such sparsity to be captured. The benefit of sparse factor analysis is its increased interpretability of the inferred factors, as each factor is encouraged to have only a few significant loadings.

Sparse factor analysis has been applied to the analysis of gene expression data (West, 2003; Sabatti and James, 2006; Pournara and Wernisch, 2007). One of the aims of such analyses is to infer gene regulatory networks, i.e. to identify sets of genes each regulated by a shared biological pathway. Thus, the use of sparse factor models is appropriate, as it allows the interpretation of factors as biological pathways, which each regulate a small number of the genes. Recent extensions of sparse factor analysis in genomics include Gao et al. (2016); Hore et al. (2016); Buettner et al. (2017); Argelaguet et al. (2018); Wang and Stephens (2021).

The sparsity of factor loadings has been achieved under a Bayesian framework by using sparsity-inducing priors such as a "spike and slab prior" (West, 2003). Markov chain Monte Carlo (MCMC), which relies on sampling from the posterior distribution, has been typically employed for Bayesian inference in sparse factor analysis. On the other hand, the recent extensions of sparse factor models in Hore et al. (2016); Buettner et al. (2017); Argelaguet et al. (2018); Wang and Stephens (2021) have used variational inference (VI). VI reformulates the inference problem to an optimisation problem

of finding an approximate distribution that resembles the posterior distribution. It is known that VI tends to be faster than MCMC, but it does not provide theoretical guarantees of finding the exact posterior distribution, which MCMC provides (Blei et al., 2016).

We aim to investigate the relative strengths and weaknesses of MCMC and VI, when applied to sparse factor models with a spike and slab prior. We derive and implement MCMC and VI algorithms and assess a trade-off between accuracy and computational efficiency using both simulated and biological data. Stegle et al. (2010) performed a similar comparison, but they used a relaxed sparsity prior for their VI algorithm, instead of the exact spike and slab prior. Our work differs from Stegle et al. (2010) as we consider a slightly more flexible sparse factor model, and derive a VI algorithm for the exact spike and slab prior. Our comparison results show that the higher computational efficiency of VI is desirable over the small gain in accuracy when using MCMC, provided that sufficient VI trials are run. We also demonstrate that our model is capable of reproducing the results of a more generalised model (Wang and Stephens, 2021) when applied to a biological dataset. Our implementation of the MCMC and VI algorithms for sparse factor models is available at `https:github.com/ysfoo/sparsefactor`.

2 The sparse factor model

Given N observations $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N]$ each with G features, the sparse factor model describes the data using K factors with a loading matrix $\mathbf{L} \in \mathbb{R}^{G \times K}$ and activation matrix $\mathbf{F} \in \mathbb{R}^{K \times N}$ such that $\mathbf{Y} = \mathbf{LF} + \mathbf{E}$, where $\mathbf{E} \in \mathbb{R}^{G \times N}$ is a matrix of random errors. In the context of gene expression, $\mathbf{Y}$ represents gene expression data across N samples, each measured on G genes. A possible interpretation of the K factors is to view them as biological pathways which regulate gene expression. By assuming independent normal errors with feature-specific variance, the distribution of $\mathbf{Y}$ is given by

$$p(\boldsymbol{y}_{\cdot j} \mid \mathbf{L}, \mathbf{F}, \boldsymbol{\tau}) = \mathcal{N}\Big(\boldsymbol{y}_{\cdot j} \;\Big|\; \mathbf{L}\boldsymbol{f}_{\cdot j}, \operatorname{diag}\Big(\{\tau_i^{-1}\}_{i=1}^{G}\Big)\Big), \tag{1}$$

where $\boldsymbol{y}_{\cdot j}$ and $\boldsymbol{f}_{\cdot j}$ indicate the j-th column of $\mathbf{Y}$ and $\mathbf{F}$ respectively, and τ_i is the precision of normal errors for observations on feature i.

Prior specifications. To induce sparsity in the loading matrix $\mathbf{L}$, we introduce a binary matrix $\mathbf{Z} \in \mathbb{R}^{G \times K}$ whose entries are 1 when the corresponding loading is nonzero. We then specify the following spike-and-slab prior:

$$p(l_{ik} \mid z_{ik}, \alpha_k) = \begin{cases} \delta_0(l_{ik}) & \text{if } z_{ik} = 0 \\ \mathcal{N}\big(l_{ik} \mid 0, \alpha_k^{-1}\big) & \text{if } z_{ik} = 1 \end{cases}, \tag{2}$$

where δ_0 is the Dirac delta distribution, l_{ik} is the loading of factor k on feature i, z_{ik} is a binary variable which indicates whether feature i is related to factor k, and α_k is the factor-specific normal precision of the nonzero values of l_{ik}. Independent Bernoulli priors are placed on the connectivity matrix $\mathbf{Z}$:

$$p(z_{ik}) = \text{Bernoulli}(z_{ik} \mid \pi_k), \tag{3}$$

where $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{K}$ are hyperparameters to be specified. Note that π_k controls the sparsity of column k of $\mathbf{Z}$, which corresponds to factor k. A gamma prior (shape-rate parametrisation) is imposed on the precisions of the loading matrix $\mathbf{L}$:

$$p(\alpha_k) = \Gamma(\alpha_k \mid a_\alpha, b_\alpha), \tag{4}$$

where a_α and b_α are hyperparameters to be specified.

To avoid non-identifiability issues caused by scaling (Pournara and Wernisch, 2007; Stegle et al., 2010), a unit variance normal prior is used for the activation matrix $\mathbf{F}$:

$$p(\boldsymbol{f}_{\cdot j}) = \mathcal{N}(\boldsymbol{f}_{\cdot j} \mid \mathbf{0}, \mathbf{I}), \tag{5}$$

where $\mathbf{I}$ is the identity matrix of size K. Lastly, a gamma prior is placed on the precision parameters of the error model:

$$p(\tau_i) = \Gamma(\tau_i \mid a_\tau, b_\tau), \tag{6}$$

where a_τ and b_τ are hyperparameters to be specified.

Bayesian inference. Bayesian inference aims to find the posterior distribution $p(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha} \mid \mathbf{Y})$. An exact calculation of the posterior distribution is intractable, so we resort to approximate methods to obtain the posterior distribution. The next two sections describe two possible Bayesian inference techniques for the sparse factor model, namely Markov chain Monte Carlo and variational inference.

3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a family of algorithms which simulate the posterior distribution $p(\boldsymbol{\theta}|\mathbf{Y})$, where $\boldsymbol{\theta}$ and $\mathbf{Y}$ denote model parameters and data respectively. In particular, MCMC simulates samples from $p(\boldsymbol{\theta}|\mathbf{Y})$ by constructing a Markov chain $\left\{\boldsymbol{\theta}^{(n)}\right\}_{n=1}$ that converges to $p(\boldsymbol{\theta}|\mathbf{Y})$. *Gibbs sampler* is a MCMC sampler for a multivariate $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ which uses full conditional distributions to construct the Markov chain. Specifically, the transition probability of the chain (assuming a fixed ordering) can be written as

$$p\Big(\boldsymbol{\theta}^{(n)} \;\Big|\; \boldsymbol{\theta}^{(n-1)}\Big) = \prod_{i=1}^{m} p\Big(\theta_i^{(n)} \;\Big|\; \theta_1^{(n)}, \ldots, \theta_{i-1}^{(n)}, \theta_{i+1}^{(n-1)}, \ldots, \theta_m^{(n-1)}, \mathbf{Y}\Big). \tag{7}$$

That is, the Gibbs sampler cycles through sampling each parameter (or parameter block) from its full conditional posterior distribution.

Collapsed Gibbs sampler for the sparse factor model. In the sparse factor model, there is a strong dependence between the parameters l_{ik} and z_{ik}, as they must be either both zero or both nonzero. Hence, applying standard Gibbs sampling to the sparse factor model will lead to slow mixing. To improve the mixing of the chain, a collapsed Gibbs sampler is used, following the approach of Stegle et al. (2010). Specifically, $\mathbf{L}$ is marginalised out from the conditional distribution of $\mathbf{Z}$, so that z_{ik} is sampled from $p(z_{ik} \mid \mathbf{Y}, \mathbf{F}, \mathbf{Z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha})$ instead of the full conditional $p(z_{ik} \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha})$, where $\mathbf{Z}_{-ik}$ denotes the elements of $\mathbf{Z}$ excluding z_{ik}. Algorithm 1 describes this sampler in full, and the derivations of the conditional distributions can be found in Appendix A.

Algorithm 1: Collapsed Gibbs sampler for the sparse factor model

Input: $T, \mathbf{Y}, \boldsymbol{\pi}, a_\tau, b_\tau, a_\alpha, b_\alpha$
Output: T samples approximating the posterior distribution
randomly initialise $\mathbf{L}', \mathbf{F}', \mathbf{Z}', \boldsymbol{\tau}', \boldsymbol{\alpha}'$ (most recent sample);
for $t \leftarrow 1$ **to** T **do**
  for $i \leftarrow 1$ **to** G **do**
    for $k \leftarrow 1$ **to** K **do**
      $z'_{ik} \leftarrow z_{ik}^{(t)} \sim p\big(z_{ik} \mid \mathbf{Y}, \mathbf{F}', \mathbf{Z}'_{-ik}, \boldsymbol{\tau}', \boldsymbol{\alpha}', \boldsymbol{\pi}\big)$;
    end
  end
  $\mathbf{L}' \leftarrow \mathbf{L}^{(t)} \sim p(\mathbf{L} \mid \mathbf{Y}, \mathbf{F}', \mathbf{Z}', \boldsymbol{\tau}', \boldsymbol{\alpha}')$;
  $\mathbf{F}' \leftarrow \mathbf{F}^{(t)} \sim p(\mathbf{F} \mid \mathbf{Y}, \mathbf{L}', \mathbf{Z}', \boldsymbol{\tau}', \boldsymbol{\alpha}')$;
  $\boldsymbol{\tau}' \leftarrow \boldsymbol{\tau}^{(t)} \sim p(\boldsymbol{\tau} \mid \mathbf{Y}, \mathbf{L}', \mathbf{F}', \mathbf{Z}', \boldsymbol{\alpha}', a_\tau, b_\tau)$;
  $\boldsymbol{\alpha}' \leftarrow \boldsymbol{\alpha}^{(t)} \sim p(\boldsymbol{\alpha} \mid \mathbf{Y}, \mathbf{L}', \mathbf{F}', \mathbf{Z}', \boldsymbol{\tau}', a_\alpha, b_\alpha)$;
end
return $\left\{\mathbf{L}^{(t)}, \mathbf{F}^{(t)}, \mathbf{Z}^{(t)}, \boldsymbol{\tau}^{(t)}, \boldsymbol{\alpha}^{(t)}\right\}_{t=1}^{T}$

Handling the symmetry of the sparse factor model.

Given a mode of the posterior distribution, if factors (of equal π_k) are permuted, or if the sign of the entries of $\mathbf{L}$ and $\mathbf{F}$ corresponding to a factor are switched, one obtains another equivalent mode. These symmetries result in up to $2^K K!$ equivalent modes in the posterior distribution, implying that the model is non-identifiable. A MCMC sampler for this model potentially suffers from the label-switching or sign-switching issue. If this occurs, posterior averages will not provide meaningful summaries of the information available; see Stephens (2000) for more discussion.

For our Gibbs sampler, label-switching or sign-switching rarely happens within a chain, and each chain usually explores only one of the equivalent modes. This is because we simulate $\mathbf{L}$ and $\mathbf{F}$ in separate steps. For example, when sampling $\mathbf{L}$, it is unlikely to have one of its column's signs flipped (for large enough G) while $\mathbf{F}$ is held constant. Similar behaviour of MCMC samplers has been previously noted in Pritchard et al. (2000). Exploring a single mode corresponding to a particular labelling of the factors is not a huge problem because the equivalent modes from permuted factors are the same from the point of view of inferring a set of factors. The ambiguity of the sign could be resolved later based on domain-specific knowledge, such as genes known to be up-regulated in a particular pathway.

Nevertheless, model non-identifiability is still an issue as we start MCMC chains from different starting points, so each chain may be exploring a different mode. Thus, we implemented a relabelling algorithm Stephens (2000) so that results

from different chains can be combined, and to deal with any potential label-switching or sign-switching issue during sampling. See Appendix C for details of our relabelling algorithm.

4 Variational inference

Variational inference (VI) is a method from machine learning that approximates probability distributions using optimisation (Blei et al., 2016), serving as an alternative approach to MCMC. We first review VI in Section 4.1, and then describe its application to the sparse factor model in Section 4.2. Further background on VI can be found in Blei et al. (2016).

4.1 Variational inference as a Bayesian inference technique

Let $\boldsymbol{\theta}$ and $\mathbf{Y}$ denote the model parameters and data, respectively. Instead of sampling from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{Y})$, VI approximates the posterior distribution by recasting the inference problem into an optimisation problem. Given a family of probability distributions $\mathcal{D}$, VI aims to find the member of $\mathcal{D}$ (called the variational approximation) which minimises its Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{D}}{\arg\min}\, \mathrm{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{Y})) = \underset{q(\boldsymbol{\theta}) \in \mathcal{D}}{\arg\min}\, \mathbb{E}[\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta} \mid \mathbf{Y})], \tag{8}$$

where the expectation is taken with respect to q. KL divergence penalises choices of q which place significant probability mass on areas where p has little probability mass, thus coercing the density of q to match that of p. It does not however, penalise as much the choices of q which place less probability mass on areas where p has more probability mass. In other words, VI attempts to match the most significant modes of p and q, potentially disregarding other modes of p that are further away. This feature is desirable for us as it implies that q tends to capture only one of the equivalent modes in the sparse factor model. However, it also implies that VI tends to underestimate the variance of the posterior distribution (Blei et al., 2016). Furthermore, a choice of $\mathcal{D}$ that is too restrictive may result in a variational approximation that does not capture the posterior distribution accurately.

Evidence lower bound. In practice, the KL divergence cannot be computed directly, but is related to the *evidence lower bound* $\mathrm{ELBO}(q) = \mathbb{E}[\log p(\mathbf{Y}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})]$ by the equation

$$\mathrm{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{Y})) = \mathbb{E}[\log q(\boldsymbol{\theta}) - \log p(\mathbf{Y}, \boldsymbol{\theta}) + \log p(\mathbf{Y})] = -\mathrm{ELBO}(q) + \log p(\mathbf{Y}). \tag{9}$$

Since $\log p(\mathbf{Y})$ is constant, minimising the KL divergence is equivalent to maximising the ELBO. As the KL divergence is always nonnegative (Kullback and Leibler, 1951), it follows that $\mathrm{ELBO}(q) \leq \log p(\mathbf{Y})$, hence the name evidence lower bound. Provided that the family of distributions $\mathcal{D}$ is simple enough, the ELBO is a tractable quantity to compute.

Mean-field approximation and coordinate ascent variational inference. A common choice of $\mathcal{D}$ is the *mean-field variational family*, where the model parameters $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^m$ are mutually independent in q. In other words, the variational approximation can be written as a product of variational factors,

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{m} q_i(\theta_i). \tag{10}$$

One of the most commonly used algorithms for solving the optimisation problem in equation (8) with the mean-field family is coordinate ascent variational inference (CAVI) (Bishop, 2006). The CAVI algorithm iterates through the variational factors, updating each $q_i(\theta_i)$ while holding the other variational factors fixed:

$$q_i^*(\theta_i) \propto \exp\left\{\mathbb{E}_{-i}[\log p(\theta_i \mid \mathbf{Y}, \boldsymbol{\theta}_{-i})]\right\} \propto \exp\left\{\mathbb{E}_{-i}[\log p(\mathbf{Y}, \boldsymbol{\theta})]\right\}, \tag{11}$$

where the expectation $\mathbb{E}_{-i}$ is taken with respect to the currently fixed variational factors, $\prod_{j \neq i}^m q_j(\theta_j)$. This update maximises the ELBO given the currently fixed variational factors (Blei et al., 2016). This enables the algorithm to monotonically optimise the ELBO, eventually reaching a local optimum.

4.2 Variational inference for the sparse factor model

For the sparse factor model, we choose the following mean-field variational family to approximate the posterior distribution:

$$q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = \prod_{i=1}^{G} \left[q(\tau_i) \prod_{k=1}^{K} q(l_{ik}, z_{ik}) \right] \times \prod_{k=1}^{K} \left[q(\alpha_k) \prod_{j=1}^{N} q(f_{kj}) \right], \tag{12}$$

where

$$q(l_{ik}, z_{ik}) = q(l_{ik} \mid z_{ik}) q(z_{ik}) \tag{13}$$

$$= \mathcal{N}\big(l_{ik} \mid \mu_{l_{ik}}, \sigma^2_{l_{ik}}\big)^{z_{ik}} \times \delta_0(l_{ik})^{1-z_{ik}} \times \text{Bernoulli}(z_{ik} \mid \eta_{ik}) \tag{14}$$

$$q(f_{kj}) = \mathcal{N}\Big(f_{kj} \;\Big|\; \mu_{f_{kj}}, \sigma^2_{f_{kj}}\Big) \tag{15}$$

$$q(\tau_i) = \Gamma\Big(\tau_i \;\Big|\; \hat{a}_{\tau_i}, \hat{b}_{\tau_i}\Big) \tag{16}$$

$$q(\alpha_k) = \Gamma\Big(\alpha_k \;\Big|\; \hat{a}_{\alpha_k}, \hat{b}_{\alpha_k}\Big). \tag{17}$$

Each variational factor we choose is conjugate to the distribution in the likelihood function, so the variational family satisfies the update rule in equation (11), and an analytic computation of the expectation on the right is possible. We use CAVI to optimise the ELBO. Algorithm 2 shows our CAVI for the sparse factor model; see Appendix B for details of the CAVI updates and the derivation of the ELBO.

Note that the variational factor $q(l_{ik}, z_{ik})$ does not factorise into $q(l_{ik})q(z_{ik})$, as it is not possible to remove the dependency of l_{ik} and z_{ik} being either both zero or both nonzero. Moreover, we derived the variational factor for the exact spike and slab prior instead of the relaxed sparsity prior used in Stegle et al. (2010).

Algorithm 2: CAVI for the sparse factor model

Input: $\mathbf{Y}, \boldsymbol{\pi}, a_\tau, b_\tau, a_\alpha, b_\alpha$
Output: variational factors which approximate the posterior distribution
randomly initialise $q(l_{ik}, z_{ik}), q(f_{kj}), q(\tau_i), q(\alpha_k)\ \forall i, j, k$;
while ELBO *has not converged* **do**
  for $i \leftarrow 1$ **to** G **do**
    for $k \leftarrow 1$ **to** K **do**
      $q(l_{ik}, z_{ik}) \propto \exp\big\{\mathbb{E}_{\mathbf{L}_{-ik}, \mathbf{F}, \mathbf{Z}_{-ik}, \tau_i, \boldsymbol{\alpha}}[\log p(l_{ik}, z_{ik} \mid \mathbf{Y}, \mathbf{L}_{-ik}, \mathbf{F}, \mathbf{Z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha})]\big\}$;
    end
  end
  for $k \leftarrow 1$ **to** K **do**
    for $j \leftarrow 1$ **to** N **do**
      $q(f_{kj}) \propto \exp\big\{\mathbb{E}_{\mathbf{L}, \mathbf{F}_{-kj}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}}[\log p(f_{kj} \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}_{-kj}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha})]\big\}$;
    end
  end
  for $i \leftarrow 1$ **to** G **do**
    $q(\tau_i) \propto \exp\big\{\mathbb{E}_{\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha}}[\log p(\tau_i \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha})]\big\}$;
  end
  for $k \leftarrow 1$ **to** K **do**
    $q(\alpha_k) \propto \exp\big\{\mathbb{E}_{\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}}[\log p(\alpha_k \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau})]\big\}$;
  end
end
return $q(l_{ik}, z_{ik}), q(f_{kj}), q(\tau_i), q(\alpha_k)\ \forall i, j, k$

Initialisation. CAVI is a hill-climbing algorithm that may find only a local optimum of the ELBO. In practice, we run multiple VI trials with different initialisations, and select the trial that converges to the largest ELBO for inference. To reduce computation, trials may be stopped early, and only the trial corresponding to the largest ELBO (at early stopping) is run until convergence.

5 Numerical comparisons

We compare the performance of MCMC and VI, focusing on accuracy and computational efficiency. It is expected that VI would converge faster than MCMC, but MCMC will provide more accurate inference in the long run. The comparison is carried out for simulated datasets and a real biological dataset.

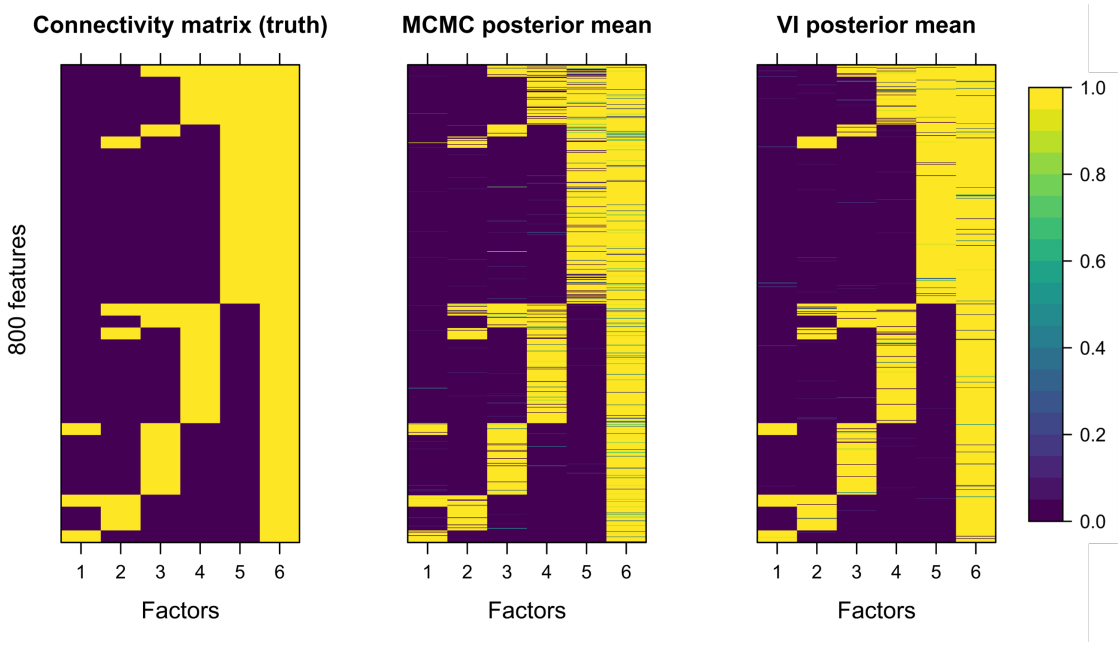

Figure 1: True connectivity structure for $\mathbf{Z}$ (simulated with snr = 5), and inferred structures (posterior mean of $\mathbf{Z}$). Results from a MCMC chain with the best accuracy of $\mathbf{Z}$ and a VI trial with the largest converged ELBO are shown.

5.1 Simulated data

The simulated datasets each consist of G = 800 features over N = 100 samples, explained by K = 6 factors. We simulated three datasets with varying amount of noise to evaluate the robustness of each inference technique. All three datasets share the same underlying connectivity structure $\mathbf{Z}$ (left column of Figure 1), consisting of 5 factors with sparse loadings and 1 factor with full loadings, corresponding to the sparsity hyperparameters $\boldsymbol{\pi}$ = (0.075, 0.15, 0.25, 0.375, 0.5, 1). The entries of $\mathbf{L}$ (that correspond to $z_{ik} = 1$) and $\mathbf{F}$ were simulated from independent standard normal distributions. The random errors present in each dataset was controlled by varying the signal-to-noise ratio (snr = 1, 5, 25). We quantified the signal for feature i using the sample variance V_i of the entries in row i of $\mathbf{LF}$ (the expectation of the data for feature i). The precision of the error is then given by

$$\tau_i = \frac{\text{snr}}{V_i}. \tag{18}$$

We applied MCMC and VI to each of these datasets, assuming *a priori* that we know the correct number of sparse factors and dense factors (5 and 1 respectively). The sparsity hyperparameters $\boldsymbol{\pi}$ were set to be 0.1 and 0.9 for sparse factors and dense factors respectively. The remaining hyperparameters for the gamma priors were set to be $a_\tau = b_\tau = a_\alpha = b_\alpha = 10^{-3}$, corresponding to vague priors. For MCMC, we discarded the first 100 iterations as a burn-in, and then ran 200,000 iterations, keeping one out of every 10 successive samples for inference. We ran MCMC 5 times with different initial values, giving 5 chains of 20,000 samples each. We ran 10 VI trials until the ELBO converged (up to absolute difference of 10^{-10} or relative difference of 10^{-14}).

We first present the results for a dataset with a moderate amount of noise (snr = 5) in the next two subsections, and then assess the robustness of each technique to the noise in data using results from the other two simulated datasets, one with more noise (snr = 1), and one with less noise (snr = 25).

Comparison of accuracy and speed. We quantitatively compare the performance of MCMC and VI when applied to the sparse factor model. Performance is evaluated via the accuracy of the inferred connectivity structure $\mathbf{Z}$, loading matrix $\mathbf{L}$, activation matrix $\mathbf{F}$, and low-dimensional structure $\mathbf{LF}$. The accuracy of $\mathbf{Z}$ is defined as the proportion of correctly inferred entries after rounding the posterior means of $\mathbf{Z}$ to 0 or 1. The accuracy of $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{LF}$ is quantified by the relative root mean squared error (RRMSE). As an example, the RRMSE for $\mathbf{F}$ is

$$\text{RRMSE}(\hat{\mathbf{F}}, \mathbf{F}) = \sqrt{\frac{\sum_{k,j}(\hat{f}_{kj} - f_{kj})^2}{\sum_{k,j} f_{kj}^2}}, \tag{19}$$

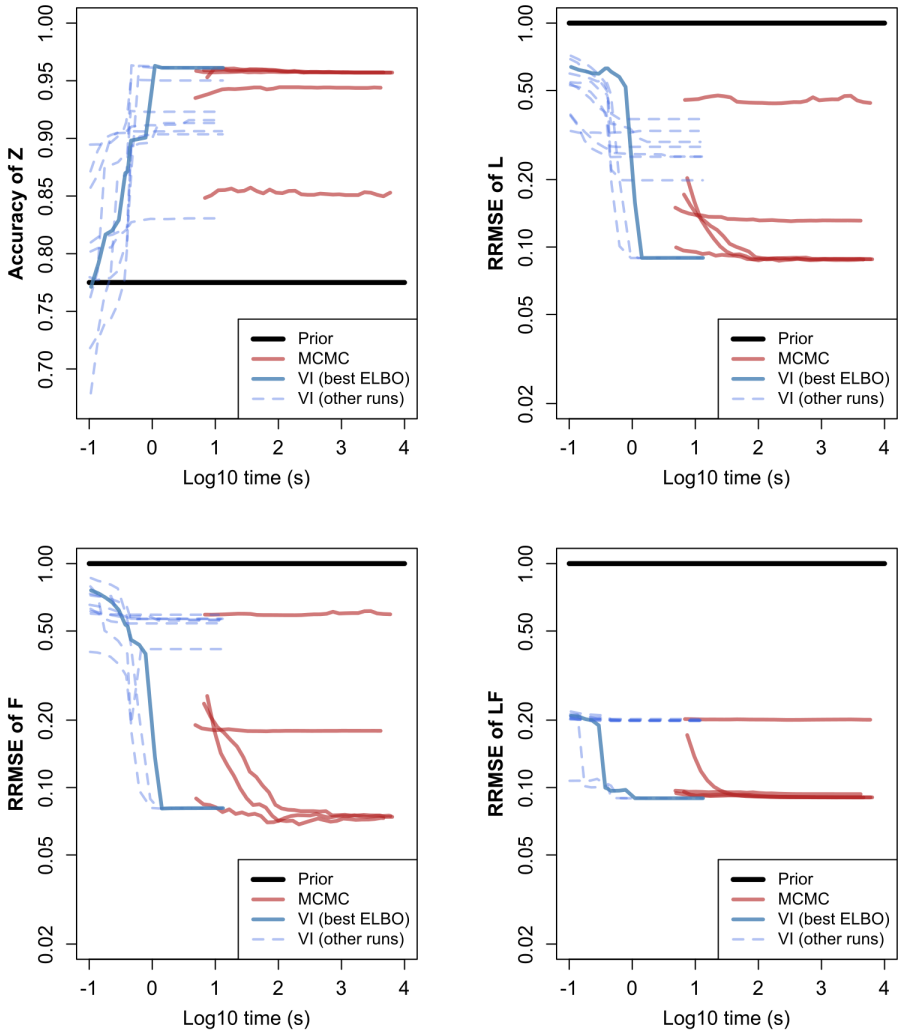

Figure 2: Performance over computation time on a simulated dataset with a moderate amount of noise (snr = 5), based on the posterior mean of the connectivity structure $\mathbf{Z}$, loading matrix $\mathbf{L}$, activation matrix $\mathbf{F}$, and low-dimensional structure $\mathbf{LF}$.

where $\hat{\mathbf{F}}$ is the posterior mean of $\mathbf{F}$. We included the performance measures of the prior mean as a baseline to compare to.

These performance measures were calculated after the inferred model parameters have been permuted and scaled appropriately to match the simulation parameters. Scaling was performed because multiplying column k of $\mathbf{L}$ and dividing row k of $\mathbf{F}$ by the same quantity leaves $\mathbf{LF}$ invariant. We scaled $\mathbf{L}$ and $\mathbf{F}$ such that the Euclidean norm of each row of $\hat{\mathbf{F}}$ matches that of the simulated $\mathbf{F}$.

Figure 2 shows the results from a dataset with a moderate amount of noise (snr = 5). Three out of the five MCMC chains captured the underlying structure well, as evident from a high accuracy of $\mathbf{Z}$ and small RRMSE of $\mathbf{L}$, $\mathbf{F}$ and $\mathbf{LF}$. Two chains failed to converge, even after more than two hours of running 200,000 MCMC iterations for each chain. In contrast, all VI trials converged in about 10 seconds, although the performance varied across trials. This is expected, as each trial climbs the ELBO to a different local optimum. Moreover, the trial which converged to the largest ELBO does not display any significant loss in accuracy when compared to the best accuracy achieved by MCMC.

Robustness against noise. We also compare the performance of MCMC and VI when applied to datasets with different amounts of noise (snr =1, 5, 25). The best VI trial (best in the sense of largest converged ELBO) achieved better performance as the amount of noise decreases (Figure 3). In all cases, its accuracy of $\mathbf{Z}$ roughly matched that of the

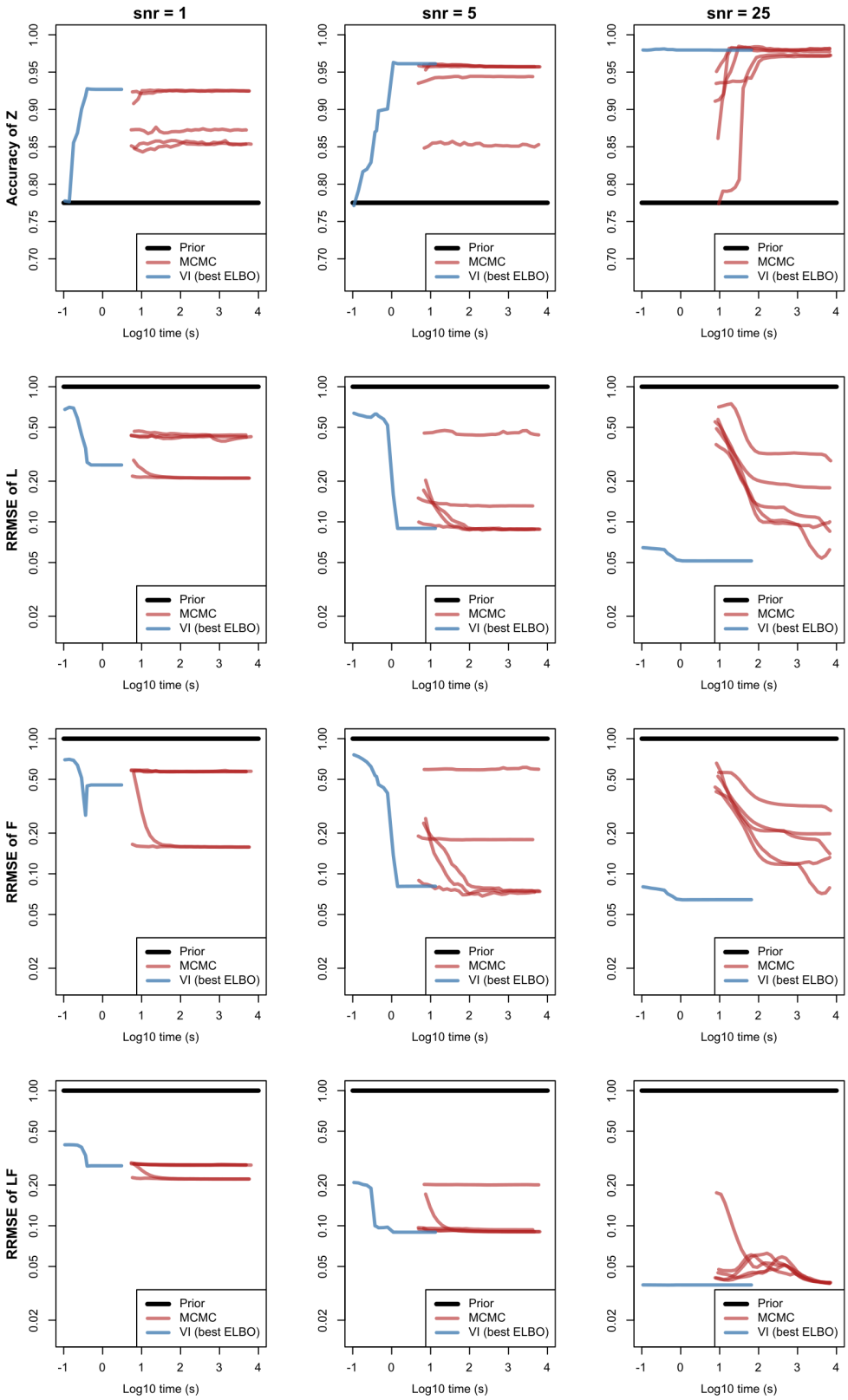

Figure 3: Performance over computation time across different simulated datasets with varying amounts of noise (snr = 1, 5, 25), based on the posterior mean of the connectivity structure $\mathbf{Z}$, loading matrix $\mathbf{L}$, activation matrix $\mathbf{F}$, and low-dimensional structure $\mathbf{LF}$.

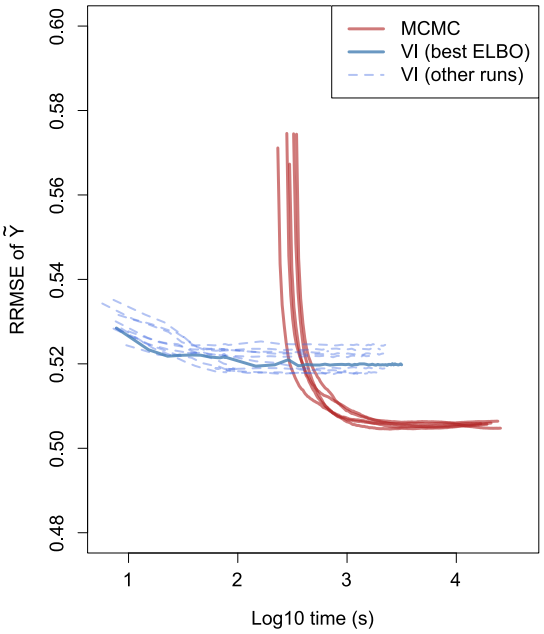

Figure 4: Performance on GTEx data over computation time, based on the posterior mean of predictions on held-out data (10% of full data).

most accurate result from a MCMC chain. The only case where MCMC may have an advantage is the dataset with snr = 1, where 2 out of the 5 MCMC chains achieved a lower error on $\mathbf{L}$, $\mathbf{F}$, and $\mathbf{LF}$ than the best VI trial. In fact, these 2 chains managed to accurately infer factor 1, which is the factor with the most sparsity. The remaining 3 chains and all 10 VI trials did not find this factor. MCMC may be more capable to infer sparse factors from noisy data than VI, but does not do so consistently.

The accuracy measures for MCMC took a longer time to converge for the dataset with the least noise (snr = 25). A possible explanation is that stronger signals make the dependency structure in the posterior distribution stronger, leading to less efficient convergence of the Gibbs sampler. In the noisier datasets, some MCMC chains were clearly stuck in non-optimal modes that do not match the underlying structure.

5.2 Biological data

In this section, we compare the performance of MCMC and VI when applying the sparse factor model to a real dataset. To this end, we used *GTEx eQTL summary data* from Wang and Stephens (2021), which consists of Z-scores measuring the associations of G = 16069 genetic variants with gene expression measured in N = 44 human tissues. In other words, y_{ij} indicates the strength of effect of genetic variant i on gene expression in tissue j. This dataset originates from the Genotype Tissue Expression (GTEx) Project (The GTEx Consortium et al., 2015), which Wang and Stephens (2021) used as part of their evaluation of *flash*, a VI-based method they developed for an Empirical Bayes approach to matrix factorisation. See Wang and Stephens (2021) for a further description of the GTEx eQTL summary data.

The *flash* is capable of automatically selecting the number of factors K, which Wang and Stephens (2021) report to be K = 26 when applied to this dataset. We used the same number of factors as inferred by *flash*, and treated all 26 factors as sparse factors, each with a sparsity hyperparameter of $\pi_k = 0.1$. The remaining hyperparameters for the gamma priors remained at 10^{-3}. The first 2,000 MCMC iterations were discarded as a burn-in. After the burn-in period, 16,000 iterations were run, where one out of every 10 successive samples were kept for inference. We ran MCMC 5 times with different starting points, giving 5 chains of 1,600 samples each. We ran 10 VI trials until the ELBO converged up to a tolerance of 10^{-3}.

Fill-in test. As the ground truth for an underlying structure is not available, we assessed the performance of each method using a *fill-in test*, following Stegle et al. (2010) and Wang and Stephens (2021). We first held-out (masked) 70704 data entries in $\mathbf{Y}$ (denoted as $\tilde{\mathbf{Y}}$), corresponding to 10% of the data entries. Then, we inferred the model parameters using the remaining 90% of the data and predicted (filled-in) these 70704 missing values using the inferred

parameters. Finally, we assessed the performance of each method using the RRMSE of the posterior mean of predictions on the held-out data, against the observed held-out data. The idea is that model parameters which better capture the true underlying structure will predict the held-out entries more accurately (Stegle et al., 2010). As expected, VI is computationally more efficient than MCMC (Figure 4), and its RRMSE is only slightly worse than that of MCMC.

Biological interpretation. To illustrate the utility of sparse factor models, we show how the inferred factors may be interpreted in a biologically meaningful way. We provide in Figure S1a-b (Appendix D) a visualisation for the posterior mean of the activation matrix $\mathbf{F}$ as inferred by the best VI trial.

Although we did not impose sparsity for $\mathbf{F}$, many factors in our activation matrix were still inferred to be sparse. Many factors had a strong effect on only one tissue (e.g. factors 4, 10, 19). Such factors may be interpreted as capturing tissue-specific effects. Moreover, some factors showed strong effects in only a small set of tissues which are known to be biologically related. For example, factors 1 and 17 have a strong effect on the tissues related to the cerebellum and the artery respectively.

In fact, our results are qualitatively similar to that of *flash* (Wang and Stephens, 2021), a sparse factor model that is capable of inducing sparsity in both the loading matrix $\mathbf{L}$ and the activation matrix $\mathbf{F}$. Despite having a more restrictive model, we were able to visually match 17 out of 26 of our factor activations to the factors inferred by *flash* (Table S1).

In both analyses, one of the factors had approximately equal weights for all tissues. In our case, this is factor 2. Wang and Stephens (2021) comment that such a factor aligns well with the fact that many genetic variants show similar effects across all tissues. Indeed, Figure S2 shows that factor 2 is the factor that controls for the largest number of genetic variants.

6 Conclusion

We have compared two Bayesian inference techniques, MCMC and VI, when applied to the sparse factor model. We have derived and implemented MCMC and VI algorithms, and investigated the relative strengths and weaknesses of two methods in terms of accuracy and computational efficiency using both simulated and biological data. Our empirical investigation showed that MCMC gives more slightly accurate inference than VI, however the difference is outweighed by the much faster speed of VI. After taking into account the need of running multiple VI trials to select the trial with the best ELBO, VI achieves similar accuracy as MCMC in significantly less time.

Acknowledgements. Special thanks to Matthew Stephens for sharing the GTEx data used in the numerical comparison. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We thank Yao-ban Chan for helpful comments on a draft manuscript. This research used the Spartan High Performance Computing system at the University of Melbourne. This work was supported by a Vacation Research Scholarship provided by the Australian Mathematical Sciences Institute to Yong See Foo.

Appendix A Conditional distributions of the sparse factor model

Denote $D_{\boldsymbol{v}} = \text{diag}(\boldsymbol{v})$ for any vector $\boldsymbol{v}$. The full conditional distribution of row i of $\mathbf{L}$ and $\mathbf{Z}$ is

$$\begin{aligned}
p(\boldsymbol{l}_{i\cdot}, \boldsymbol{z}_{i\cdot} \mid \mathbf{Y}, \mathbf{F}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \propto & \prod_{k\,:\,z_{ik}=1} \pi_k \sqrt{\frac{\alpha_k}{2\pi}} \times \prod_{k\,:\,z_{ik}=0} (1-\pi_k)\,\delta_0(l_{ik}) \\
& \times \exp\Bigg\{ -\frac{\tau_i}{2} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)^{\mathsf{T}} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right) \\
& \qquad - \frac{1}{2} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}} [D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \Bigg\} \\
\propto & \prod_{k\,:\,z_{ik}=1} \pi_k \sqrt{\frac{\alpha_k}{2\pi}} \times \prod_{k\,:\,z_{ik}=0} (1-\pi_k)\,\delta_0(l_{ik}) \\
& \times \exp\left\{ -\frac{1}{2} \left([\boldsymbol{l}]_{\boldsymbol{z}_{i\cdot}} - \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}\right)^{\mathsf{T}} \Sigma_{\boldsymbol{l}_{i\cdot}}^{-1} \left([\boldsymbol{l}]_{\boldsymbol{z}_{i\cdot}} - \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}\right) + \frac{1}{2} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}^{\mathsf{T}} \Sigma_{\boldsymbol{l}_{i\cdot}}^{-1} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} \right\}
\end{aligned}$$

where

$$\begin{aligned}
[\mathbf{F}]_{\boldsymbol{z}_{i\cdot}} &= \text{matrix consisting of rows of } \mathbf{F} \text{ whose corresponding entries of } \boldsymbol{z}_{i\cdot} \text{ are equal to 1} \\
[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} &= \text{vector consisting of entries of } \boldsymbol{l}_{i\cdot} \text{ whose corresponding entries of } \boldsymbol{z}_{i\cdot} \text{ are equal to 1} \\
[D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}} &= \text{matrix consisting of rows of } D_{\boldsymbol{\alpha}} \text{ whose corresponding entries of } \boldsymbol{z}_{i\cdot} \text{ are equal to 1} \\
\Sigma_{\boldsymbol{l}_{i\cdot}} &= \left(\tau_i [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}} [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}} + [D_{\boldsymbol{\alpha}}]_{\boldsymbol{z}_{i\cdot}}\right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}} &= \tau_i \Sigma_{\boldsymbol{l}_{i\cdot}} [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}} \boldsymbol{y}_{i\cdot}.
\end{aligned}$$

The full conditional distribution of $\boldsymbol{l}_{i\cdot}$ is then

$$p(\boldsymbol{l}_{i\cdot} \mid \mathbf{Y}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) = \mathcal{N}\big([\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}} \mid \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}, \Sigma_{\boldsymbol{l}_{i\cdot}}\big) \times \prod_{k\,:\,z_{ik}=0} \delta_0(l_{ik}).$$

To obtain a collapsed Gibbs sampler, $\boldsymbol{l}_{i\cdot}$ is marginalised out from the full conditional distribution of z_{ik}:

$$p(z_{ik} \mid \mathbf{Y}, \mathbf{F}, \mathbf{Z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \propto \left(\frac{\alpha_k}{2\pi}\right)^{\frac{z_{ik}}{2}} \det|\Sigma_{\boldsymbol{l}_{i\cdot}}|^{\frac{1}{2}} \exp\left\{\frac{1}{2} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}^{\mathsf{T}} \Sigma_{\boldsymbol{l}_{i\cdot}}^{-1} \boldsymbol{\mu}_{\boldsymbol{l}_{i\cdot}}\right\} \pi_k^{z_{ik}} (1-\pi_k)^{1-z_{ik}}.$$

The full conditional distribution of column j of $\mathbf{F}$ is

$$p(\boldsymbol{f}_{\cdot j} \mid \mathbf{Y}, \mathbf{L}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \propto \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j}\right)^{\mathsf{T}} D_{\boldsymbol{\tau}} \left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j}\right) - \frac{1}{2} \boldsymbol{f}_{\cdot j}^{\mathsf{T}} \boldsymbol{f}_{\cdot j} \right\},$$

which is a normal distribution with mean and covariance

$$\begin{aligned}
\Sigma_{\boldsymbol{f}_{\cdot j}} &= \left(\mathbf{L}^{\mathsf{T}} D_{\boldsymbol{\tau}} \mathbf{L} + \mathbf{I}\right)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{f}_{\cdot j}} &= \Sigma_{\boldsymbol{f}_{\cdot j}} \mathbf{L}^{\mathsf{T}} D_{\boldsymbol{\tau}} \boldsymbol{y}_{\cdot j}.
\end{aligned}$$

Lastly, the full conditional distribution of τ_i is

$$p(\tau_i \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha}) = \Gamma\left(\tau_i \;\middle|\; a_\tau + \frac{N}{2}, b_\tau + \frac{1}{2} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)^{\mathsf{T}} \left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]_{\boldsymbol{z}_{i\cdot}}^{\mathsf{T}} [\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)\right),$$

and the full conditional distribution of α_k is

$$p(\alpha_k \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}) = \Gamma\left(\alpha_k \;\middle|\; a_\alpha + \frac{1}{2}\sum_{i=1}^{G} z_{ik}, b_\alpha + \frac{1}{2} \sum_{i\,:\,z_{ik}=1} l_{ik}^2\right).$$

Appendix B Details of VI for the sparse factor model

Throughout this section, all expectations are taken over the distribution $q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha})$.

Coordinate ascent updates. Coordinate ascent on $\boldsymbol{l}_{i\cdot}$ and $\boldsymbol{z}_{i\cdot}$ gives

$$
\begin{aligned}
q^*(l_{ik}, z_{ik}) &\propto \exp\left\{\mathbb{E}_{\mathbf{L}_{-ik},\mathbf{F},\mathbf{Z}_{-ik},\tau_i,\boldsymbol{\alpha}}[\log p(l_{ik}, z_{ik} \mid \mathbf{Y}, \mathbf{L}_{-ik}, \mathbf{F}, \mathbf{Z}_{-ik}, \boldsymbol{\tau}, \boldsymbol{\alpha})]\right\} \\
&\propto \exp\left\{ -\mathbb{E}_{\mathbf{L}_{-ik},\mathbf{F},\mathbf{Z}_{-ik},\tau_i}\left[\frac{\tau_i}{2}\left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^{\mathsf{T}}_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)^{\mathsf{T}}\left(\boldsymbol{y}_{i\cdot} - [\mathbf{F}]^{\mathsf{T}}_{\boldsymbol{z}_{i\cdot}}[\boldsymbol{l}_{i\cdot}]_{\boldsymbol{z}_{i\cdot}}\right)\right]\right. \\
&\qquad \left. + \frac{z_{ik}}{2}\mathbb{E}_{\alpha_k}\left[\log\frac{\alpha_k}{2\pi} - \alpha_k l_{ik}^2\right]\right\} \times \pi_k^{z_{ik}}\left((1-\pi_k)\,\delta_0(l_{ik})\right)^{1-z_{ik}}, \\
&\propto \exp\left\{ -\frac{\hat{a}_{\tau_i}}{2\hat{b}_{\tau_i}}\mathbb{E}_{\mathbf{L}_{-ik},\mathbf{F},\mathbf{Z}_{-ik}}\left[-2\boldsymbol{y}_{i\cdot}^{\mathsf{T}}\boldsymbol{f}_{k\cdot}l_{ik} + 2\sum_{k'\neq k} z_{ik'}\boldsymbol{f}_k^{\mathsf{T}}\boldsymbol{f}_{k'\cdot}l_{ik'}l_{ik} + \boldsymbol{f}_k^{\mathsf{T}}\boldsymbol{f}_k l_{ik}^2\right]\right. \\
&\qquad \left. + \frac{1}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}l_{ik}^2\right)\right\}^{z_{ik}} \times \pi_k^{z_{ik}}\left((1-\pi_k)\,\delta_0(l_{ik})\right)^{1-z_{ik}} \\
&\propto \exp\left\{ \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\left(-\sum_{j=1}^{N}\left(y_{ij}\mu_{f_{kj}} - \sum_{k'\neq k}\eta_{ik'}\mu_{f_{kj}}\mu_{f_{k'j}}\mu_{l_{ik'}}\right)l_{ik} - \sum_{j=1}^{N}\left(\mu_{f_{kj}}^2 + \sigma_{f_{kj}}^2\right)\frac{l_{ik}^2}{2}\right)\right. \\
&\qquad \left. + \frac{1}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}l_{ik}^2\right)\right\}^{z_{ik}} \times \pi_k^{z_{ik}}\left((1-\pi_k)\,\delta_0(l_{ik})\right)^{1-z_{ik}},
\end{aligned}
$$

which corresponds to the updates

$$
\begin{aligned}
\sigma_{l_{ik}}^{2*} &= \left(\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sum_{j=1}^{N}\left(\mu_{f_{kj}}^2 + \sigma_{f_{kj}}^2\right) + \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}\right)^{-1} \\
\mu_{l_{ik}}^* &= \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\sigma_{l_{ik}}^{2*}\sum_{j=1}^{N}\left(y_{ij}\mu_{f_{kj}} - \sum_{k'\neq k}\eta_{ik'}\mu_{f_{kj}}\mu_{f_{k'j}}\mu_{l_{ik'}}\right) \\
q(z_{ik}) &\propto \exp\left\{\frac{z_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} + \frac{\mu_{l_{ik}}^{2*}}{\sigma_{l_{ik}}^{2*}}\right)\right\}\left(\sqrt{2\pi\sigma_{l_{ik}}^{2*}}\pi_k\right)^{z_{ik}}(1-\pi_k)^{1-z_{ik}}.
\end{aligned}
$$

Coordinate ascent on f_{kj} gives

$$
\begin{aligned}
q^*(f_{kj}) &\propto \exp\left\{\mathbb{E}_{\mathbf{L},\mathbf{F}_{-kj},\mathbf{Z},\boldsymbol{\tau},\boldsymbol{\alpha}}[\log p(f_{kj} \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}_{-kj}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha})]\right\} \\
&\propto \exp\left\{\mathbb{E}_{\mathbf{L},\mathbf{F}_{-kj},\mathbf{Z},\boldsymbol{\tau}}\left[-\frac{1}{2}\left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j}\right)^{\mathsf{T}}D_{\boldsymbol{\tau}}\left(\boldsymbol{y}_{\cdot j} - \mathbf{L}\boldsymbol{f}_{\cdot j}\right)\right] - \frac{1}{2}f_{kj}^2\right\} \\
&\propto \exp\left\{\boldsymbol{y}_{\cdot j}^{\mathsf{T}}D_{\boldsymbol{\tau}}\overline{\boldsymbol{l}_{\cdot k}}f_{kj} - \sum_{k'\neq k}\overline{f_{k'j}\boldsymbol{l}_{\cdot k'}^{\mathsf{T}}D_{\boldsymbol{\tau}}\boldsymbol{l}_{\cdot k}}f_{kj} - \frac{1}{2}\left(\overline{\boldsymbol{l}_{\cdot k}^{\mathsf{T}}D_{\boldsymbol{\tau}}\boldsymbol{l}_{\cdot k}} + 1\right)f_{kj}^2\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
D_{\overline{\boldsymbol{\tau}}} &= \operatorname{diag}\left(\left\{\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\right\}_{i=1}^{G}\right) \\
\overline{\boldsymbol{l}_{\cdot k}} &= \{\eta_{ik}\mu_{l_{ik}}\}_{i=1}^{G} \\
\overline{f_{k'j}\boldsymbol{l}_{\cdot k'}^{\mathsf{T}}D_{\boldsymbol{\tau}}\boldsymbol{l}_{\cdot k}} &= \mu_{f_{k'j}}\sum_{i=1}^{G}\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\eta_{ik}\eta_{ik'}\mu_{l_{ik}}\mu_{l_{ik'}} \\
\overline{\boldsymbol{l}_{\cdot k}^{\mathsf{T}}D_{\boldsymbol{\tau}}\boldsymbol{l}_{\cdot k}} &= \sum_{i=1}^{G}\frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}\eta_{ik}\left(\mu_{l_{ik}}^2 + \sigma_{l_{ik}}^2\right),
\end{aligned}
$$

which corresponds to the updates

$$\sigma^{2*}_{f_{kj}} = \left(\overline{\boldsymbol{l}_{\cdot k}^{\mathsf{T}} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} + 1\right)^{-1}$$
$$\mu^{*}_{f_{kj}} = \sigma^{2*}_{f_{kj}} \left(\boldsymbol{y}_{\cdot j}^{\mathsf{T}} D_{\overline{\boldsymbol{\tau}}} \overline{\boldsymbol{l}_{\cdot k}} - \sum_{k' \neq k} \overline{f_{k'j} \boldsymbol{l}_{\cdot k'}^{\mathsf{T}} D_{\boldsymbol{\tau}} \boldsymbol{l}_{\cdot k}} \right).$$

Coordinate ascent on τ_i gives

$$\begin{aligned} q^*(\tau_i) &\propto \exp\left\{\mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z},\boldsymbol{\alpha}}[\log p(\tau_i \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\alpha})]\right\} \\ &\propto \exp\left\{ \left(a_\tau - 1 + \frac{N}{2}\right) \log \tau_i - b_\tau \tau_i - \frac{\tau_i}{2} \mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z}} \left[\left(\boldsymbol{y}_{i\cdot} - \mathbf{F}^{\mathsf{T}} \boldsymbol{l}_{i\cdot}\right)^{\mathsf{T}} \left(\boldsymbol{y}_{i\cdot} - \mathbf{F}^{\mathsf{T}} \boldsymbol{l}_{i\cdot}\right) \right] \right\} \\ &\propto \exp\left\{ \left(a_\tau - 1 + \frac{N}{2}\right) \log \tau_i - \left(b_\tau + \frac{1}{2} \left(\boldsymbol{y}_{i\cdot}^{\mathsf{T}} \boldsymbol{y}_{i\cdot} - 2\overline{\boldsymbol{l}_{i\cdot}}^{\mathsf{T}} \overline{\mathbf{F}} \boldsymbol{y}_{i\cdot} + \overline{\boldsymbol{l}_{i\cdot}^{\mathsf{T}} \mathbf{F}\mathbf{F}^{\mathsf{T}} \boldsymbol{l}_{i\cdot}} \right) \right) \tau_i \right\} \end{aligned}$$

where

$$\begin{aligned} \overline{\boldsymbol{l}_{i\cdot}} &= \{\eta_{ik} \mu_{l_{ik}}\}_{k=1}^{K} \\ \left[\overline{\mathbf{F}}\right]_{kj} &= \mu_{f_{kj}} \\ \overline{\boldsymbol{l}_{i\cdot}^{\mathsf{T}} \mathbf{F}\mathbf{F}^{\mathsf{T}} \boldsymbol{l}_{i\cdot}} &= \sum_{k=1}^{K} \sum_{k'=1}^{K} \left(\eta_{ik} \eta_{ik'}^{1-\delta_{kk'}} \left(\mu_{l_{ik}} \mu_{l_{ik'}} + \delta_{kk'} \sigma^2_{l_{ik}} \right) \sum_{j=1}^{N} \left(\mu_{f_{kj}} \mu_{f_{k'j}} + \delta_{kk'} \sigma^2_{f_{kj}} \right) \right), \end{aligned}$$

which corresponds to the updates

$$\begin{aligned} \hat{a}^*_{\tau_i} &= a_\tau + \frac{N}{2} \\ \hat{b}^*_{\tau_i} &= b_\tau + \frac{1}{2} \left(\boldsymbol{y}_{i\cdot}^{\mathsf{T}} \boldsymbol{y}_{i\cdot} - 2\overline{\boldsymbol{l}_{i\cdot}}^{\mathsf{T}} \overline{\mathbf{F}} \boldsymbol{y}_{i\cdot} + \overline{\boldsymbol{l}_{i\cdot}^{\mathsf{T}} \mathbf{F}\mathbf{F}^{\mathsf{T}} \boldsymbol{l}_{i\cdot}} \right). \end{aligned}$$

Coordinate ascent on α_k gives

$$\begin{aligned} q^*(\alpha_k) &\propto \exp\left\{\mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z},\boldsymbol{\tau}}[\log p(\alpha_k \mid \mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau})]\right\} \\ &\propto \exp\left\{ \left(a_\alpha - 1 + \frac{1}{2} \mathbb{E}_{\mathbf{Z}} \left[\sum_{i=1}^{G} z_{ik} \right] \right) \log \alpha_k - b_\alpha \alpha_k - \frac{\alpha_k}{2} \mathbb{E}_{\mathbf{L},\mathbf{Z}} \left[\sum_{i\,:\,z_{ik}=1} l_{ik}^2 \right] \right\} \end{aligned}$$

which corresponds to the updates

$$\begin{aligned} \hat{a}^*_{\alpha_k} &= a_\alpha + \frac{1}{2} \sum_{i=1}^{G} \eta_{ik} \\ \hat{b}^*_{\alpha_k} &= b_\alpha + \frac{1}{2} \sum_{i=1}^{G} \eta_{ik} \left(\sigma^2_{l_{ik}} + \mu^2_{l_{ik}} \right). \end{aligned}$$

Computing the ELBO. For the sparse factor model, the ELBO is given by

$$\text{ELBO}(q) = \mathbb{E}_{\mathbf{L},\mathbf{F},\mathbf{Z},\boldsymbol{\tau},\boldsymbol{\alpha}}[\log p(\mathbf{Y}, \mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha}) - \log q(\mathbf{L}, \mathbf{F}, \mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\alpha})].$$

Breaking this down into components, the expectation of the first logarithm consists of

$$\begin{aligned} \mathbb{E}_{\mathbf{L},\mathbf{F},\boldsymbol{\tau}}[\log p(y_{ij} \mid \mathbf{L}, \mathbf{F}, \boldsymbol{\tau})] &= \frac{1}{2} \mathbb{E}_{\mathbf{L},\mathbf{F},\boldsymbol{\tau}} \left[\log \frac{\tau_i}{2\pi} - \tau_i \left(y_{ij} - \boldsymbol{l}_{i\cdot}^{\mathsf{T}} \boldsymbol{f}_{\cdot j} \right)^2 \right] \\ &= \frac{1}{2} \left(\psi(\hat{a}_{\tau_i}) - \log 2\pi \hat{b}_{\tau_i} - \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}} \left(-2 y_{ij} \sum_{k=1}^{K} \eta_{ik} \mu_{l_{ik}} \mu_{f_{kj}} \right. \right. \\ &\qquad \left. \left. + y_{ij}^2 + \overline{\left(\boldsymbol{l}_{i\cdot}^{\mathsf{T}} \boldsymbol{f}_{\cdot j} \right)^2} \right) \right) \end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_{\mathbf{L},\mathbf{Z},\boldsymbol{\alpha}}[\log p(l_{ik} \mid \mathbf{Z}, \boldsymbol{\alpha})] &= \mathbb{E}_{\mathbf{L},\mathbf{Z},\boldsymbol{\alpha}}\left[\frac{z_{ik}}{2}\left(\log\frac{\alpha_k}{2\pi} - \alpha_k l_{ik}^2\right) + (1-z_{ik})\log\delta_0(l_{ik})\right] \\
&= \frac{\eta_{ik}}{2}\left(\psi(\hat{a}_{\alpha_k}) - \log 2\pi\hat{b}_{\alpha_k} - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}\left(\mu_{l_{ik}}^2 + \sigma_{l_{ik}}^2\right)\right) \\
&\quad + \mathbb{E}_{\mathbf{L},\mathbf{Z}}[(1-z_{ik})\log\delta_0(l_{ik})] \\
\mathbb{E}_{\mathbf{Z}}[\log p(z_{ik})] &= \eta_{ik}\log\pi_k + (1-\eta_{ik})\log(1-\pi_k) \\
\mathbb{E}_{\mathbf{F}}[\log p(f_{kj})] &= -\frac{1}{2}\left(\mu_{f_{kj}}^2 + \sigma_{f_{kj}}^2 + \log 2\pi\right) \\
\mathbb{E}_{\boldsymbol{\tau}}[\log p(\tau_i)] &= \mathbb{E}_{\boldsymbol{\tau}}[(a_\tau - 1)\log\tau_i - b_\tau\tau_i] + a_\tau\log b_\tau - \log\Gamma(a_\tau) \\
&= (a_\tau - 1)\left(\psi(\hat{a}_{\tau_i}) - \log\hat{b}_{\tau_i}\right) - \frac{\hat{a}_{\tau_i}}{\hat{b}_{\tau_i}}b_\tau + a_\tau\log b_\tau - \log\Gamma(a_\tau) \\
\mathbb{E}_{\boldsymbol{\alpha}}[\log p(\alpha_k)] &= \mathbb{E}_{\boldsymbol{\alpha}}[(a_\alpha - 1)\log\alpha_k - b_\alpha\alpha_k] + a_\alpha\log b_\alpha - \log\Gamma(a_\alpha) \\
&= (a_\alpha - 1)\left(\psi(\hat{a}_{\alpha_k}) - \log\hat{b}_{\alpha_k}\right) - \frac{\hat{a}_{\alpha_k}}{\hat{b}_{\alpha_k}}b_\alpha + a_\alpha\log b_\alpha - \log\Gamma(a_\alpha)
\end{aligned}
$$

where

$$
\overline{\left(\boldsymbol{l}_{i\cdot}^{\mathsf{T}}\boldsymbol{f}_{\cdot j}\right)^2} = \sum_{k=1}^{K}\sum_{k'=1}^{K}\eta_{ik}\eta_{ik'}^{1-\delta_{kk'}}\left(\mu_{l_{ik}}\mu_{l_{ik'}} + \delta_{kk'}\sigma_{l_{ik}}^2\right)\left(\mu_{f_{kj}}\mu_{f_{k'j}} + \delta_{kk'}\sigma_{f_{kj}}^2\right).
$$

Using standard differential entropy results, the expectation of the second logarithm consists of

$$
\begin{aligned}
\mathbb{E}_{\mathbf{L},\mathbf{Z}}[-\log q(l_{ik}, z_{ik})] &= \frac{\eta_{ik}}{2}\left(\log 2\pi\sigma_{l_{ik}}^2 + 1\right) - \eta_{ik}\log\eta_{ik} - (1-\eta_{ik})\log(1-\eta_{ik}) \\
&\quad - \mathbb{E}_{\mathbf{L},\mathbf{Z}}[(1-z_{ik})\log\delta_0(l_{ik})] \\
\mathbb{E}_{\mathbf{F}}[-\log q(f_{kj})] &= \frac{1}{2}\left(\log 2\pi\sigma_{f_{kj}}^2 + 1\right) \\
\mathbb{E}_{\boldsymbol{\tau}}[-\log q(\tau_i)] &= \hat{a}_{\tau_i} - \log\hat{b}_{\tau_i} + \log\Gamma(\hat{a}_{\tau_i}) + (1-\hat{a}_{\tau_i})\,\psi(\hat{a}_{\tau_i}) \\
\mathbb{E}_{\boldsymbol{\alpha}}[-\log q(\alpha_k)] &= \hat{a}_{\alpha_k} - \log\hat{b}_{\alpha_k} + \log\Gamma(\hat{a}_{\alpha_k}) + (1-\hat{a}_{\alpha_k})\,\psi(\hat{a}_{\alpha_k}).
\end{aligned}
$$

The ELBO may be calculated by summing up these expectations appropriately.

Appendix C Relabelling samples

A relabelling algorithm, similar to that of Erosheva and Curtis (2017), is used to deal with these model non-identifiability issues of label symmetry and sign ambiguity. For the sparse factor model, a relabelling consists of permuting the factors, and potentially flipping the signs of all entries of some factors. Following the method of Stephens (2000), a decision-theoretic approach is to define a loss function for a set of actions and relabellings, and select the action and relabelling which minimises the posterior expected loss. This is done with the aim of relabelling samples such that they correspond to being sampled around the same mode. Define an action

$$
\mathbf{a} = \left(\left\{m_{f_{kj}}\right\}_{\substack{j=1:N\\k=1:K}}, \left\{s_{f_{kj}}^2\right\}_{\substack{j=1:N\\k=1:K}}\right)
$$

to be a choice of means and variances of the entries of $\mathbf{F}$. Let $\sigma \in S_K$ and $\boldsymbol{\nu} \in \{-1, 1\}^K$, where S_K is the set of permutations on the set $\{1, 2, \ldots, K\}$. Define a loss function as follows:

$$
\mathcal{L}(\mathbf{a}, \sigma, \boldsymbol{\nu}; \mathbf{F}) = -\sum_{k=1}^{K}\sum_{j=1}^{N}\mathcal{N}(\nu_{\sigma(k)}f_{\sigma(k)j} \mid m_{f_{kj}}, s_{f_{kj}}^2).
$$

Suppose T simulated samples of $\mathbf{F}$, namely $\left\{\mathbf{F}^{(t)}\right\}_{t=1:T}$ are to be relabelled, which may be obtained from multiple chains. The actions and relabellings $\mathbf{a}$ and $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$ are to be chosen such that the Monte Carlo risk

$$
\mathcal{R}_{\text{MC}} = \sum_{t=1}^{T}\mathcal{L}\left(\mathbf{a}, \left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}; \mathbf{F}^{(t)}\right)
$$

is minimised. In the case of multiple chains, it may be more appropriate to first scale each row of $\mathbf{F}$ to unit norm. After initialising $\mathbf{a}$ and $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$, a local optimum may be obtained by alternating between the following steps:

1. Given the current values of $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$, choose $\mathbf{a}$ such that the Monte Carlo risk is minimised.

2. Given the current action $\mathbf{a}$, choose $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$ such that the Monte Carlo risk is minimised.

This procedure is terminated when a fixed point is reached. The final signflips and permutations $\left\{\left(\sigma^{(t)}, \boldsymbol{\nu}^{(t)}\right)\right\}_{t=1:T}$ are then applied to all relevant variables simulated.

Step 1 may be solved analytically, by setting partial derivatives of the Monte Carlo risk with respect to the action parameters to zero. This is equivalent to finding the maximum likelihood estimators, summarised by the following updates:

$$\widehat{m_{f_{kj}}} = \frac{1}{T}\sum_{t=1}^{T} \nu^{(t)}_{\sigma^{(t)}(k)} f^{(t)}_{\sigma^{(t)}(k)j}$$

$$\widehat{s^2_{f_{kj}}} = \frac{1}{T}\sum_{t=1}^{T} \left(\nu^{(t)}_{\sigma^{(t)}(k)} f^{(t)}_{\sigma^{(t)}(k)j} - \widehat{m_{f_{kj}}}\right)^2.$$

Step 2 is equivalent to the linear assignment problem. For each simulated sample, this may be solved by an $\mathcal{O}(K^3)$ algorithm of Jonker and Volgenant (1987) after a cost matrix is constructed. The construction of the cost matrix itself takes $\mathcal{O}\big(K^2(G+N)\big)$ time (for each simulated sample).

Appendix D Supplementary figures

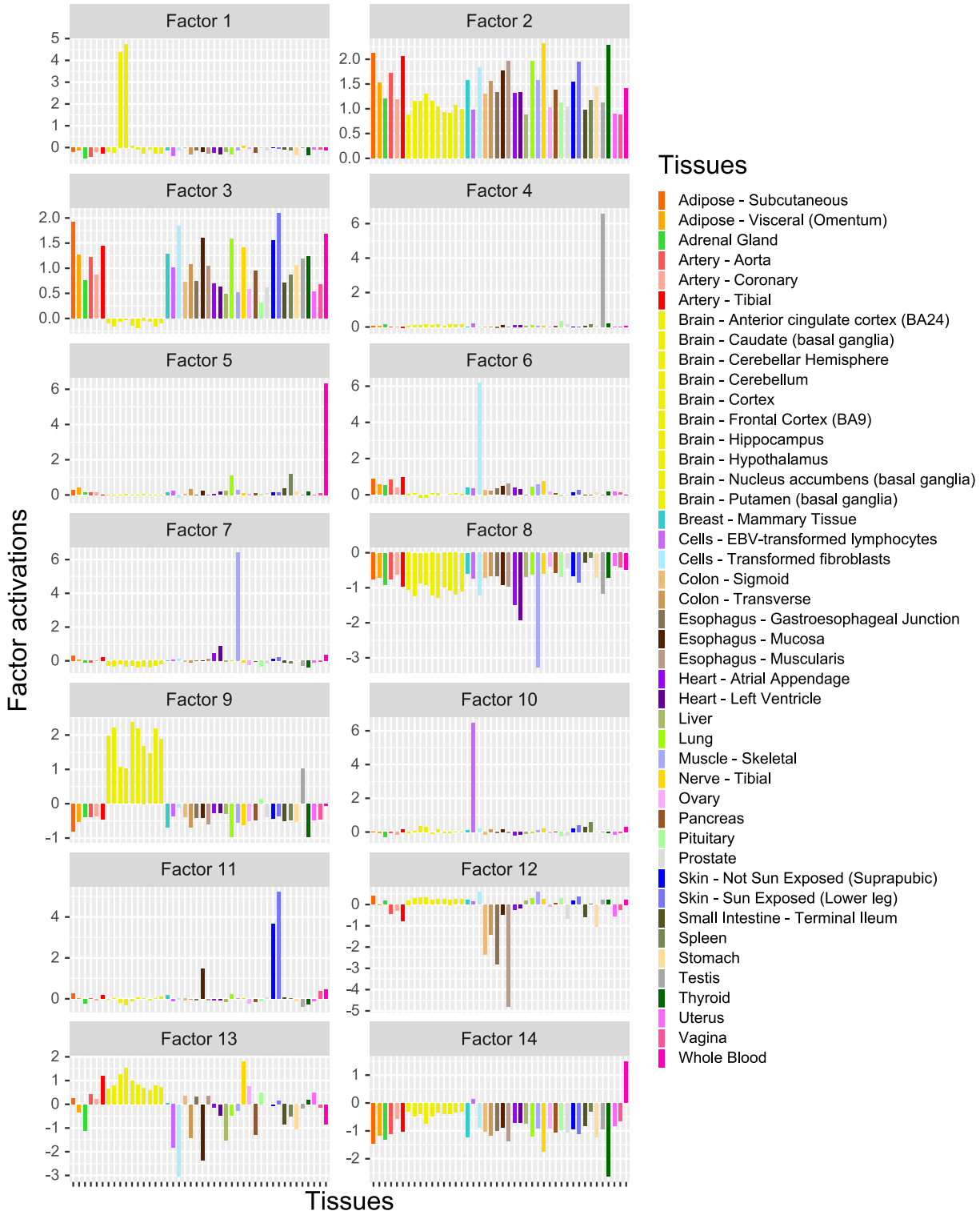

Figure S1a: Posterior means of the activations for factors 1 to 14 from running VI on GTEx data.

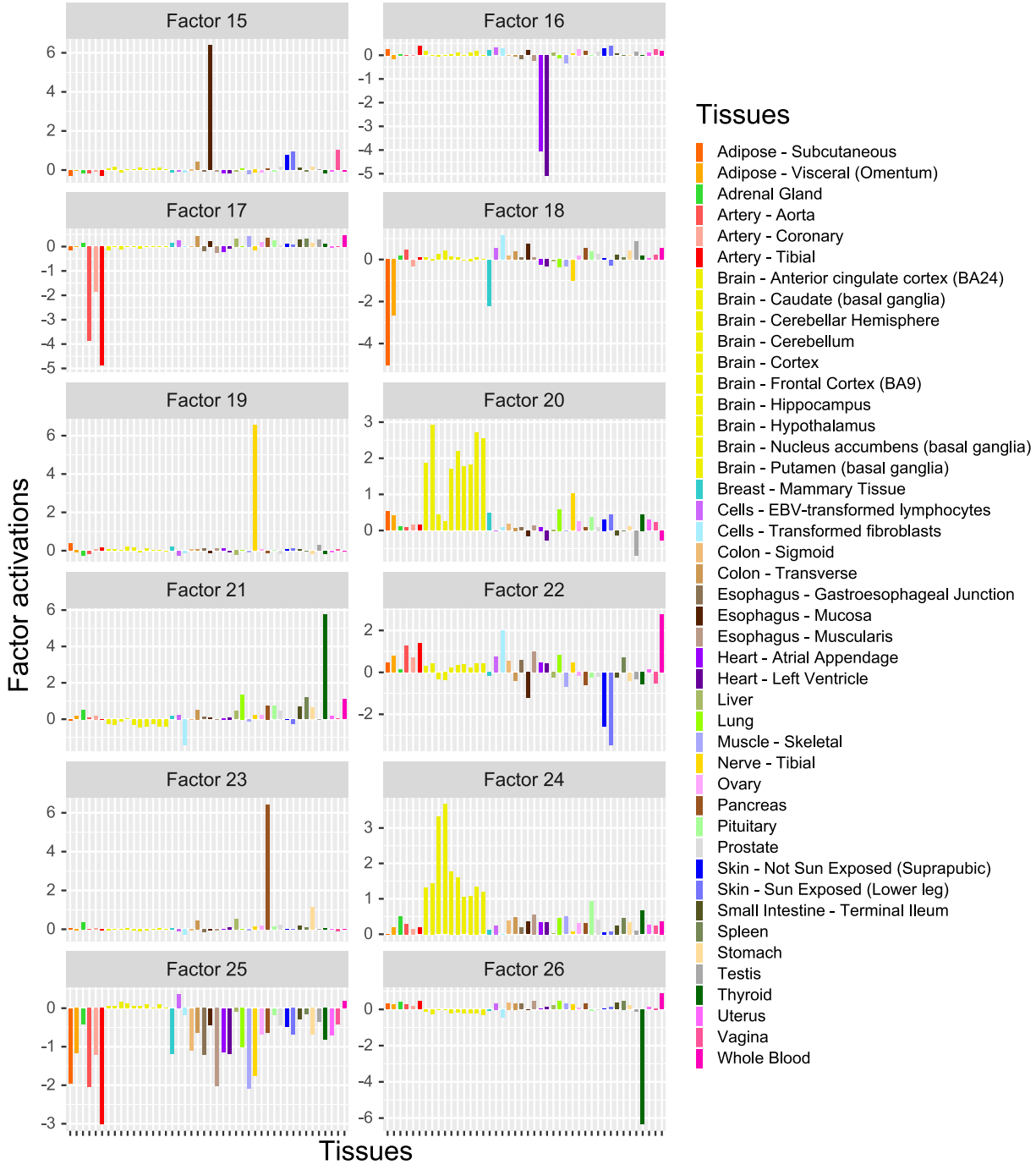

Figure S1b: Posterior means of the activations for factors 15 to 26 from running VI on GTEx data.

Our model	1	2	3	4	5	6	7	8	9	10	11	12	13
flash	9	1	-	7	3	6	5	-	2	13	4	20	-
Our model	14	15	16	17	18	19	20	21	22	23	24	25	26
flash	-	14	12	8	17	16	-	-	-	18	-	-	11

Table S1: Matching factor labels from our VI results to factors inferred by *flash*; see Figures 4 and 5 of Wang and Stephens (2021).

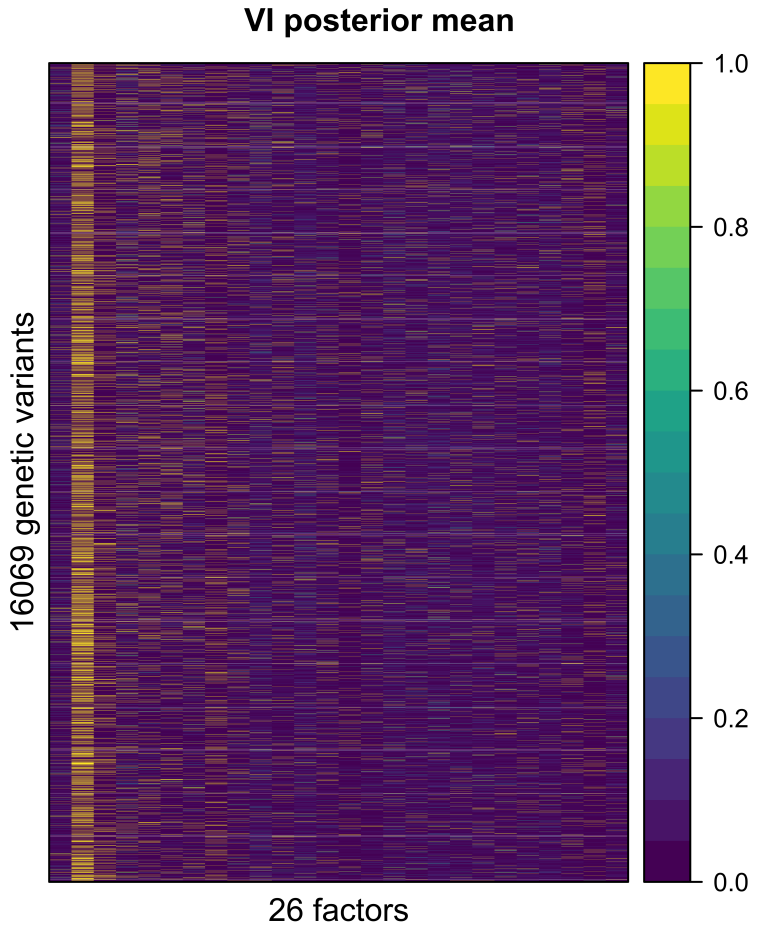

Figure S2: Inferred connectivity structure for the GTEx data obtained from VI.

References

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. Springer, New York.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2016). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877. arXiv: 1601.00670.

Buettner, F., Pratanwanich, N., McCarthy, D. J., Marioni, J. C., and Stegle, O. (2017). f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, 18(1):212.

Erosheva, E. A. and Curtis, S. M. (2017). Dealing with Reflection Invariance in Bayesian Factor Analysis. *Psychometrika*, 82(2):295–307.

Gao, C., McDowell, I. C., Zhao, S., Brown, C. D., and Engelhardt, B. E. (2016). Context specific and differential gene co-expression networks via bayesian biclustering. *PLoS computational biology*, 12(7):e1004791.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., and Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature genetics*, 48(9):1094–1100.

Jonker, R. and Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Pournara, I. and Wernisch, L. (2007). Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8(1):61.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Sabatti, C. and James, G. M. (2006). Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, 22(6):739–746.

Stegle, O., Sharp, K., Winn, J., and Rattray, M. (2010). A comparison of inference in sparse factor analysis. Technical report, Technical report.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

The GTEx Consortium et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Wang, W. and Stephens, M. (2021). Empirical bayes matrix factorization. *Journal of Machine Learning Research*, 22(120):1–40.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742.