SuSiE PCA: a Scalable Bayesian Variable Selection Technique for Principal Component Analysis

Dong Yuan and Nicholas Mancuso

March 20, 2023

SuSiE PCA: A Scalable Bayesian Variable Selection Technique for Principal Component Analysis

Dong Yuan[1] and Nicholas Mancuso[1,2,3]

[1] Biostatistics Division, Dept of Population and Public Health Sciences,Keck School of Medicine, University of Southern California, Los Angeles, CA
[2] Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA
[3] Dept of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA

Abstract. Traditional latent factor models such as principal component analysis (PCA) provide a statistical framework to infer low-rank latent components across a multitude of biologically relevant settings. However, when this low-rank structure manifests from a sparse subspace, approaches that seek to infer the relevant features either lack the ability to perform feature selection, or fail to quantify the uncertainty in their selected features. In this paper, we present SuSiE PCA, a highly scalable sparse latent factor approach that explicitly models uncertainty in contributing variables through posterior inclusion probabilities (PIPs). We validate our model in extensive simulations and demonstrate that SuSiE PCA outperforms other approaches for detecting relevant signals in observed data, while being robust to model mis-specification. To illustrate its performance in real-data scenarios, we apply SuSiE PCA to multi-tissue eQTL data from GTEx v8 and identify tissue-specific regulatory factors and their contributing eGenes. Next, we investigate its performance to identify gene regulatory modules using large-scale perturbation screen data. We find that SuSiE PCA discovers modules enriched for genes relevant for ribosome function to a greater extent than competing methods (ribosome pathway: FDR $= 9.2 \times 10^{-82}$, 63 genes involved vs. 1.4×10^{-33}, 35 genes involved), while being ∼18x faster. Overall, SuSiE PCA provides an efficient and flexible tool to identify relevant features in high-dimensional structured biological data.

1 Introduction

Principal component analysis (PCA) is a popular dimension reduction technique [1] that has been widely applied for exploratory data analysis in many fields. One notable functionality of PCA is to synthesize crucial information across features into a small number of principal components (PCs). For example, PCA is commonly used to infer population structure from large-scale genetic data [2, 3]. The top PCs explain differences in genetic variation arising from different geographic origins and ancestry of individuals, due to historical migration, admixture, etc. [4]. Moreover, PCA provides a means to rank contributing relevant variables for each latent component, as Tipping and Bishop(1986) proposed the probabilistic reformulation of principal component analysis (PPCA) [5]. Specifically, each PC is independent of other PCs and has its unique weights to represent the "importance" of original features, suggesting different latent components arise from different combinations of variables, or distinct aspects of information from the data.

However, one disadvantage of conventional PCA is that PCs provide limited interpretability, as each results from a linear combination of variables in the data [6]. To improve the interpretability of PCs, while providing an identifiable solution in high-dimensional data, a common approach is to impose sparsity on the PCA loadings. Broadly speaking, there are two types of approaches to achieving sparsity on the loading matrix. The first is the regularization methods such as sparse PCA, which rewrites the PCA as a regression-based optimization problem and then includes a L_1 penalty on the objective function [6] to achieve sparse loadings. The second type of method is the Bayesian treatment of PPCA, which imposes sparsity-induced prior on the factor loading matrix [7, 8, 9, 10, 11, 12]. Despite various methods that focus on inducing sparse solutions for PCA, few provide a statistically rigorous way to select variables relevant to each factor in a post-hoc manner. Although several sparse models are capable of shrinking the loadings of uninformative variables to zero, for those variables with non-zero weights, neither a reasonable threshold nor a formal statistical test is provided to inform feature prioritization for validation or follow-up.

Here, we propose SuSiE PCA, a highly-scalable Bayesian framework for sparse PCA, that quantifies the uncertainty of contributing features for each latent component. Specifically, SuSiE PCA leverages the recent "sum of single effects" (SuSiE) approach [13] to model a loading matrix such that each latent factor contains at most L contributing features. Latent factors and sparse loading weights are learned through an efficient variational algorithm. In addition to providing a sparse loading matrix, SuSiE PCA computes posterior inclusion probabilities (PIPs) for each feature, which enables defining $\rho-$level credible sets for feature selection. We demonstrate through extensive simulations that SuSiE PCA outperforms existing latent factor approaches in identifying relevant features contributing to structured data while being robust to data-generating assumptions. Next, we apply SuSiE PCA to multi-tissue eQTL data from the GTEx v8 [14, 12] study to identify tissue-specific components of regulatory genetic features and contributing eGenes. We also apply SuSiE PCA to high-dimensional perturb-seq data (CRISPR-based screens with single-cell RNA-sequencing readouts) [15] and identify gene sets more enriched in the ribosome, coronavirus disease pathways when compared with sparse PCA (FDR = 9.2×10^{-82}, 63 genes involved vs. 1.4×10^{-33}, 35 genes involved) while requiring 17.8 times less computing time. Overall, we find SuSiE PCA provides an efficient approach to compute interpretable latent factors from high-dimensional biological data. We provide an open-source python implementation that can run seamlessly on CPUs, GPUs, or TPUs available at `http://www.github.com/mancusolab/susiepca`.

2 Materials and Method

2.1 SuSiE PCA Model

Let $\mathbf{X}_{N\times P}$ be the observed data matrix, $\mathbf{Z}_{N\times K}$ be the K dimensional latent vectors, and $\mathbf{W}_{K\times P}$ be the loading matrix. We denote the normal distribution with mean μ and variance σ^2 as $\mathcal{N}(\mu, \sigma^2)$, the multinomial distribution with n choices and probabilities $\boldsymbol{\pi}$ as $\text{Multi}(n, \boldsymbol{\pi})$ and the matrix normal distribution with dimension $N \times K$, mean $\mathbf{M}$, row-covariance $\mathbf{R}$, and column-covariance $\mathbf{C}$ as $\mathcal{MN}_{N,K}(\mathbf{M}, \mathbf{R}, \mathbf{C})$. We denote the basis vector in which k^{th} coordinate is 1 and 0 elsewhere as $\mathbf{e}_k$. The sampling distribution of $\mathbf{X}$ under the SuSiE PCA model is given by,

$$\mathbf{X} \mid \mathbf{Z}, \mathbf{W}, \sigma^2 \sim \mathcal{MN}_{N,P}(\mathbf{ZW}, \mathbf{I}_N, \sigma^2\mathbf{I}_P) \tag{2.1}$$

$$\mathbf{Z} \sim \mathcal{MN}_{N,K}(\mathbf{0}, \mathbf{I}_N, \mathbf{I}_K) \tag{2.2}$$

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{e}_k \mathbf{w}_k^{\intercal} \tag{2.3}$$

$$\mathbf{w}_k = \sum_{l=1}^{L} \mathbf{w}_{kl} \tag{2.4}$$

$$\mathbf{w}_{kl} = w_{kl}\boldsymbol{\gamma}_{kl} \tag{2.5}$$

$$w_{kl} \mid \sigma^2_{0kl} \sim \mathcal{N}(0, \sigma^2_{0kl}) \tag{2.6}$$

$$\boldsymbol{\gamma}_{kl} \mid \boldsymbol{\pi} \sim \text{Multi}(1, \boldsymbol{\pi}), \tag{2.7}$$

where $\mathbf{w}_k$ corresponds to the k^{th} row of $\mathbf{W}$, and contains exactly L non-zero elements determined by the sum of L single-effect vectors $\mathbf{w}_{kl}$. These single-effect vectors are described by a single random effect w_{kl} and indicator vector $\boldsymbol{\gamma}_{kl}$ which assigns the effect to a feature with prior probabilities $\boldsymbol{\pi} = \frac{1}{p}\mathbf{1}$.

2.2 Variational Inference in SuSiE PCA

We seek to perform inference of model variables $\mathbf{Z}, \mathbf{w}_{kl}$, and $\boldsymbol{\gamma}_{kl}$ conditional on observed data $\mathbf{X}$, however the marginal likelihood is intractable to compute and therefore, we cannot evaluate the posterior exactly. While sampling based approaches such as Markov Chain Monte Carlo (MCMC) methods provide a numerical approximation of the exact posterior distribution [16], they often lack computational efficiency in high-dimensional settings. As an alternative, we leverage recent advancements in the variational inference that provides an analytical approximation to the posterior distribution [17] and remains computationally efficient.

Briefly, To approximate the conditional distribution of latent variables $\mathbf{Z}$ given the observed samples $\mathbf{X}$, variational methods first impose a family of densities over the latent variables, $Q(\mathbf{Z})$, usually predefined as known distributions parameterized with a set of variational parameters. Then the goal is to infer those variational parameters such that the variational distribution $Q(\mathbf{Z})$ is as similar as possible to the true posterior distribution $P(\mathbf{Z} \mid \mathbf{X})$. A quantity commonly used to measure dissimilarity between distributions is Kullback-Leibler divergence $D_{KL}(Q \| P)$ [18]. However, since KL divergence contain the unknown true posterior distribution $P(\mathbf{Z} \mid \mathbf{X})$, it cannot be directly computed. Instead, we can show that the log-likelihood of data, $\log P(\mathbf{X})$ can be decomposed as:

$$\log P(\mathbf{X}) = D_{KL}(Q \| P) + \mathcal{L}(Q) \tag{2.8}$$

Where $\mathcal{L}(Q) = \mathbb{E}_Q[\log P(\mathbf{Z}, \mathbf{X}) - \log Q(\mathbf{Z})]$, which is also known as the Evidence Lower Bound (ELBO). Since the $\log P(\mathbf{X})$ is a constant with respect to the variational parameters, minimizing KL divergence is equivalent to maximizing ELBO. As the ELBO does not contain the unknown posterior distribution and therefore is tractable to compute and maximize for variational parameters.

Mean-Field Approximation Mean field approximation [19] is a common solution to find the optimal solution that maximizes ELBO. The basic assumption is that we can factorize the variational distribution into independent components. Then using the calculus of variations, one can show that the distribution $Q_j^*(\mathbf{z}_j)$ minimizing KL divergence for each factor $\mathbf{Z}_j$ can be expressed as:

$$\ln Q_j^*(\mathbf{z}_j \mid \mathbf{X}) = \mathbb{E}_{i \neq j}[\ln P(\mathbf{Z}, \mathbf{X})] + constant \tag{2.9}$$

Applying the Mean-Field approximation to SuSiE PCA the approximate posterior given by,

$$Q(\mathbf{Z}, \mathbf{W}) = Q(\mathbf{Z}) Q(\mathbf{W}) \tag{2.10}$$

$$Q(\mathbf{W}) = \prod_{k=1}^{K} \prod_{l=1}^{L} Q(\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}) Q(\boldsymbol{\gamma}_{kl}) \tag{2.11}$$

Equation (2.10) factorizes the variational densities of the latent variables $\mathbf{Z}$ and the loading matrix $\mathbf{W}$ into independent parts. We further assume that the variational distribution of loadings $\mathbf{w}_{kl}$ from each factor across L single effects are independent as well, leading to equation (2.11). For ease of notation we first define $\tau = \frac{1}{\sigma^2}, \tau_{0kl} = \frac{1}{\sigma_{0kl}^2}$. We obtain the optimal variational distributions of variables $\mathbf{Z}, \mathbf{w}_{kl}$, and $\boldsymbol{\gamma}_{kl}$ is given by,

$$Q(\mathbf{Z}) := \mathcal{MN}_{n,k}(\mathbf{Z} \mid \boldsymbol{\mu}_{\mathbf{Z}}, \mathbf{I}_n, \boldsymbol{\Sigma}_{\mathbf{Z}}) \tag{2.12}$$

$$Q(w_{kl} | \boldsymbol{\gamma}_{kl}) := \mathcal{N}(\mu_{\mathbf{w}_{kl}}, \sigma^2_{\mathbf{w}_{kl}}) \tag{2.13}$$

$$Q(\boldsymbol{\gamma}_{kl}) := \text{Multi}(1, \boldsymbol{\alpha}_{kl}). \tag{2.14}$$

The corresponding update rules for variational parameters from $Q(\cdot)$ can be expressed as,

$$\boldsymbol{\mu}_{\mathbf{Z}} = \tau \mathbf{X} \mathbb{E}[\mathbf{W}^{\intercal}] \boldsymbol{\Sigma}_{\mathbf{Z}} \tag{2.15}$$

$$\boldsymbol{\Sigma}_{\mathbf{Z}} = (\mathbb{E}[\mathbf{W}\mathbf{W}^{\intercal}]\tau + \mathbf{I}_k)^{-1} \tag{2.16}$$

$$\boldsymbol{\mu}_{\mathbf{w}_{kl}} = \tau \sigma^2_{w_{kl}} \mathbb{E}[\mathbf{R}_{kl}^{\intercal} \mathbf{Z}_k] \tag{2.17}$$

$$\boldsymbol{\Sigma}_{\mathbf{w}_{kl}} = \sigma^2_{\mathbf{w}_{kl}} \mathbf{I}_p \tag{2.18}$$

$$\sigma^2_{\mathbf{w}_{kl}} = (\tau \mathbb{E}[\mathbf{Z}_k^{\intercal} \mathbf{Z}_k] + \tau_{0kl})^{-1} \tag{2.19}$$

$$\boldsymbol{\alpha}_{kl} = \text{softmax}(\log \boldsymbol{\pi} - \log \mathcal{N}(0 \mid \mu_{\mathbf{w}_{kl}}, \sigma^2_{\mathbf{w}_{kl}})). \tag{2.20}$$

For details and derivations, please see **Supplemental Note**.

ELBO and Estimating $\boldsymbol{\tau}, \boldsymbol{\tau_{0kl}}$ The ELBO provides a natural criterion for evaluating model performance during model training, and also provides a means to perform hyperparameter optimization for model variance (or equivalently precision) parameters. Given the above definitions for Q, we derive the ELBO for SuSiE PCA as,

$$\begin{aligned}\text{ELBO}(\mathbf{W}, \mathbf{Z}) &= \mathbb{E}_Q\left[\log \Pr(\mathbf{X}, \mathbf{Z}, \mathbf{W}) - \log Q(\mathbf{Z}, \mathbf{W})\right] \\ &= \mathbb{E}_Q[\log \Pr(\mathbf{X}|\mathbf{Z}, \mathbf{W})] + \mathbb{E}_{Q(\mathbf{Z})}[\log \Pr(\mathbf{Z}) - \log Q(\mathbf{Z})] + \mathbb{E}_{Q(\mathbf{W}, \boldsymbol{\Gamma})}[\log \Pr(\mathbf{W}, \boldsymbol{\Gamma}) - \log Q(\mathbf{W}, \boldsymbol{\Gamma})].\end{aligned} \tag{2.21}$$

The maximum likelihood estimates of model precision parameters τ, τ_{0kl}, results in closed-form update equations given by,

$$\hat{\tau}_{0kl} = \frac{\sum_{i=1}^{P} \alpha_{kli}}{\sum_{i=1}^{P} \alpha_{kli}(\mu^2_{\mathbf{w}_{kli}} + \sigma^2_{\mathbf{w}_{kli}})} \tag{2.22}$$

$$\hat{\tau} = \frac{NP}{\sum_{i,j} X_{ij}^2 - 2\text{tr}(\mathbb{E}[\mathbf{W}]\mathbf{X}^\intercal \boldsymbol{\mu}_{\mathbf{Z}})}. \tag{2.23}$$

We provide details on individual components of expectation and a a complete description of our inference procedure in **Supplemental Note**.

2.3 Posterior Inclusion Probability and Credible Set

One of the distinguishing features that the SuSiE model[13] provides is a posterior inclusion probability (PIP). The PIP reflects the posterior probability that a given variable has a non-zero effect given the observed data. Here we extend the PIP definition to include latent factors. Specifically, given variational parameters $\boldsymbol{\alpha}_{kl}$ we can define the PIP that the i^{th} variable has a non-zero effect in the k^{th} latent component as,

$$\text{PIP}_{ki} := \Pr(w_{ki} \neq 0 \mid \mathbf{X}) = 1 - \prod_{l=1}^{L}(1 - \alpha_{kli}) \tag{2.24}$$

Similarly, a level-ρ credible set (CS) refers to a subset of variables that cumulatively explain at least ρ of the posterior density. Here, we define factor-specific level-ρ CSs, which can be computed across each $\boldsymbol{\alpha}_{kl}$ independently, resulting in $K \times L$ total level-ρ credible sets. This lets us reflect on the uncertainty in identified variables to explain a single-effect for each latent factor.

2.4 Simulations

To investigate the performance of SuSiE PCA in variable selection and model fitting, we simulated various data sets that are controlled by 4 parameters: the sample size N, number of features P, number of latent factors K, and number of single effects L in each of the factors. For simplicity, we assume L is the same across different factors. The simulated data $\mathbf{X}$ is generated according to equation (2.1), where $N = 1000, P = 6000$, and $\mathbf{z}_k$ and $\mathbf{w}_k$, for $k = 1, \cdots, 4$ are simulated such that each factor only contain 40 non-zero effects (0.67%) given by,

$$\mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N) \tag{2.25}$$

$$w_{1,i} \sim \mathcal{N}(0, 1) \quad i = 1, \cdots, 40 \tag{2.26}$$

$$w_{2,i} \sim \mathcal{N}(0, 1) \quad i = 41, \cdots, 80 \tag{2.27}$$

$$w_{3,i} \sim \mathcal{N}(0, 2^2) \quad i = 81, \cdots, 120 \tag{2.28}$$

$$w_{4,i} \sim \mathcal{N}(0, 1) \quad i = 121, \cdots, 160, \tag{2.29}$$

with the remaining effects set to zero. Considering the scale of the estimates of loadings may differ from various types of methods, we normalized the loading matrix with respect to Frobenius norm, i.e. $\text{tr}(A^\intercal A) = \text{tr}(B^\intercal B) = 1$.

To evaluate the accuracy of SuSiE PCA, we compare inferred posterior expectations with the true latent variables. However, due to the rotational invariance property in latent factor models, evaluating loading or latent factor accuracy can be challenging. To account for possible rotation, we leverage the Procrustes transformation [20], which finds an orthogonal rotation matrix $\mathbf{P}$ to transform the estimated loading matrix to the true loading matrix space. Specifically, given an estimated loading matrix $\hat{\mathbf{W}} := \mathbb{E}_Q[\mathbf{W}]$ under approximate posterior distribution Q and true effect matrix $\mathbf{W}$, the "Procrustes Norm" can be obtained as following:

$$||\mathbf{W} - \hat{\mathbf{W}}||_P^2 := \min_{\{\mathbf{P}|\mathbf{P}^{-1}=\mathbf{P}^\intercal\}} ||\hat{\mathbf{W}}\mathbf{P} - \mathbf{W}||_F^2. \tag{2.30}$$

Here we perform the Procrustes analysis via Procrustes package [21], from which $\mathbf{P}$ is obtained by performing a singular value decomposition on matrix $\hat{\mathbf{W}}^\intercal\mathbf{W}$ (padding zeros on matrix $\hat{\mathbf{W}}$ would ensure the above operation process correctly).

In addition, we employ the relative root mean squared error (RRMSE) to evaluate the reconstructed data loss as,

$$\text{RRMSE}(\hat{X}, X) = \sqrt{\frac{\sum_{i,j}(\hat{X}_{ij} - X_{ij})^2}{\sum_{i,j} X_{ij}^2}}. \tag{2.31}$$

For model comparison, we also evaluate the performance of sparse PCA[6] and Empirical Bayes Matrix Factorization (EBMF) (a recently described variational approach)[12] on the same data sets with the same K, and compare the model performance with SuSiE PCA via criterion described above.

2.5 Real Data Analysis

To illustrate the application of SuSiE PCA in genetic research, we downloaded the Genotype-Tissue Expression (GTEx) summary statistics data, composed of z-scores computed from the testing association between genetic variants and the gene expression levels across 44 different human tissues. The GTEx project collected genotype data and gene expression data from 49 non-disease tissues across $n = 838$ individuals, providing an ideal resource database to study the relationship between genetic variants and gene expression levels [14]. The genetic variants that are statistically associated with gene expression levels are referred to as expression quantitative trait loci (eQTLs). To identify eQTLs, the GTEx project tested the association between each nearby genetic variant of a certain gene with its expression levels using linear regression to yield a z score. The summary data we explored reflects the most significant eQTL (equivalently, the largest absolute z score in each SNP and gene pair) at each of 16069 genes (row) from 44 tissues (column) curated from GTEx (v8) in ref[12], as those 16069 genes show indication of being expressed in 44 of all 49 human tissues. To identify tissue-specific components of regulatory genetic features and contributing genes, we applied SuSiE PCA across this z-score matrix with a latent dimension of 27 and the number of single effects of 18. The prior information on the number of latent dimensions comes from Wang et al. (2021) [12] who contribute to the z-score dataset and run the EBMF model with 27 factors. To determine the appropriate L that fits the data, we run the SuSiE PCA with L ranged from 10 to 25, and select the model when the increase in the total percentage of variance explained (PVE) is less than 5%. PVE is a measure of the amount of signals in the data captured by the latent component, the PVE of the factor $\mathbf{z}_k$ is calculated based on the following equation:

$$\text{PVE}_k = \frac{s_k}{\sum_k s_k + NP/\tau} \tag{2.32}$$

where $s_k = \sum_{i=1}^N \sum_{j=1}^P (\mathbb{E}[z_{ik}]\mathbb{E}[w_{kj}])^2$.

We next investigated genome-scale Perturb-seq data to discover the co-regulated gene sets affected by some common type of perturbations. To collect the Perturb-seq (CRISPR-based screens with single-cell RNA-sequencing readouts) dataset, the Replogle et al. [15] performed the experiment with 2056 distinct knocked-out genes (perturbations) as well as one non-targeting control group (no perturbations) over an average of 150 different single cells, and then measured the expression levels of the downstream 8563 genes from each cell. The final dataset contains 310385 rows, each representing one perturbation in a specific cell, and the expression levels of 8563 downstream genes as the column. As an exploratory analysis, we omitted the single-cell level information and aggregated the expression levels of downstream genes with the same perturbation over all the cells, which resulted in a "psuedo-bulk" data matrix with 2057 rows and 8563 columns. We then performed the SuSiE PCA and Sparse PCA to investigate the regulatory modules from the common perturbations. To exclude the batch effects and other non-genetic covariates, we regressed out the germ-line group and the mitochondrial percent from the original expression data and then aggregated the expression level of downstream genes with the same perturbation. Finally, the aggregated data is centered and standardized before input into SuSiE PCA and sparse PCA.

3 Results

3.1 Simulation Results

To evaluate the performance of SuSiE PCA, we performed extensive simulations (see **Materials and Methods**). Briefly, we generated 100 simulations by varying model parameters and performed inference using SuSiE PCA with the true number of latent variables (K) and effects (L) known. First, we evaluated the ability of inferred PIPs to discriminate between relevant and non-relevant features for latent factors. Specifically, we compared the sensitivity and specificity of inferred PIPs to normalized posterior mean weights from SuSiE PCA (see Figure 1). When selecting variables based on PIPs > 0.90, SuSiE PCA identifies 88.9% of true positive (non-zero) signals, demonstrating largely calibrated posterior inference. We observed nearly all true negative signals exhibited PIPs < 0.05. As a comparison, the posterior weights performed well on excluding the true negative signals, but fail to capture true positive signals as rapidly as PIP thresholds. Overall, the simulation demonstrates that PIPs provide an intuitive and more efficient indicator for feature selection than normalized posterior weights in SuSiE PCA. In addition, we also examined the sensitivity and specificity using weights estimated from sparse PCA and EBMF (see Figure S1), which have similar trends to the curves in Figure 1 (B) and can only capture a small proportion of the true positive signals as the cutoff threshold increases.

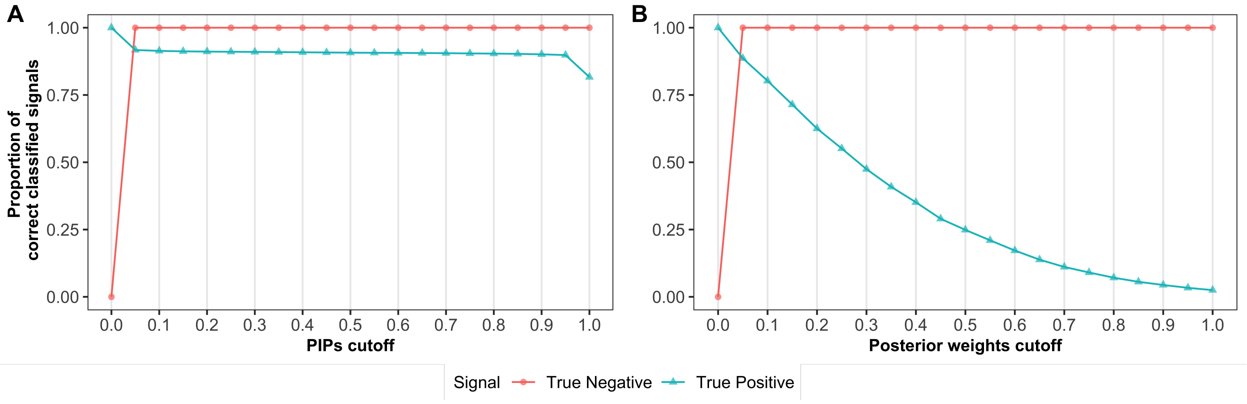

Fig. 1: PIPs exhibit a higher efficiency in selecting the true signals than the posterior weights in SuSiE PCA
The proportion of correct classified signals using PIPs as cutoff (A) or posterior weights as cutoff (B). The green dots represent sensitivity, i.e. Pr(PIPs $\geq$ cutoff | True positive signal), the red dots represent specificity, i.e. Pr(PIPs $<$ cutoff | True false signal). For consistency and to ensure comparability between PIPs and weights, the weights are standardized to be ranged from 0 to 1.

Next, we examined the estimation accuracy of the loading matrix as a function of sample size (N), feature dimension (P), latent dimension (K), and the number of single effects (or sparsity level) (L), via the Procrustes errors defined in equation 2.30 (Figure 2 A-D). We found that SuSiE PCA has the smallest Procrustes errors across all simulation settings compared to sparse PCA and EBMF. And we noticed that the Bayesian methods including SuSiE PCA and EBMF maintain a low error even with a small sample size or high feature dimension. Moreover, we found that SuSiE PCA has the lowest RRMSE across all simulations compared with other methods (Figure S2); And EBMF and SuSiE PCA have a lower level of Procrustes error of factor **Z** than sparse PCA (Figure S3). In summary, SuSiE PCA exhibits the best estimation accuracy, which is consistent with its superior performance in variable selection.

Finally, we investigate the model robustness in model mis-specification. Similar to other latent factor models, SuSiE PCA could be mis-specified as it requires manually inputting the latent dimension K and the number of single effects L. Considering the potential model misspecification setting, the simulation data sets are generated based on (2.25-2.29) and then input into SuSiE PCA, sparse PCA and EBMF with two mis-specified situations: vary L while fixing K, or vary K while fixing L. The model estimation accuracy is then evaluated and compared among three models with Procrustes error (see Figure 2 E-F). We observed that as K and L in the model approaches the true value (i.e. $K = 4$ or $L = 40$), the Procrustes error decreases rapidly to the lower level in SuSiE PCA, and remains the same even when $K > 4$ or $L > 40$. However, the error for sparse PCA has a V-shape and reaches its minimum at the real K. The explanation is that when there are over-specified latent factors in the model, SuSiE PCA and EBMF will not extract any information from the data due to their probabilistic model structure; the sparse PCA, on the other hand, cannot handle the weights as it does not impose any probabilistic assumption on it. Instead, the value of the redundant latent factor in sparse PCA is close to 0, which ensures the latent component does not contribute.

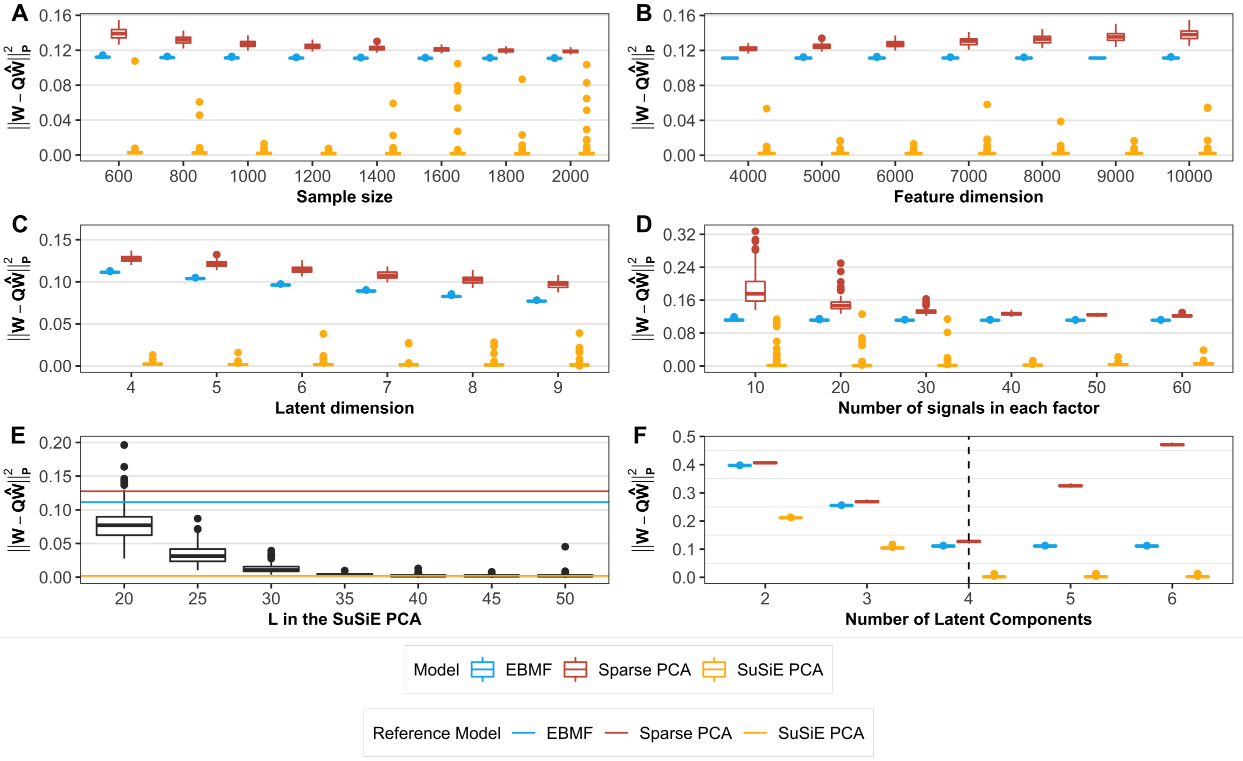

Fig. 2: SuSiE PCA generates the smallest Procrustes error in weight matrix than sparse PCA and EBMF (A-D) and is robust to over-specified K and L (E-F).
For each scenario in (A-D) we vary one of the parameters at a time to generate the simulation data while fixing the other 3 parameters, and then input the true parameters (N, P, K, L) into models. Finally, we compute the Procrustes error and plot them as a function of N, P, K, L. For (E-F), we use the same simulation setting in Figure 1 to generate data but vary the specified L in SuSiE PCA (E) and K in all three models (F). Reference lines refer to the error from the models with correctly specified parameters (i.e. $L = 40, K = 4$).

3.2 GTEx Z-score Data

To illustrate the utility of SuSiE PCA to make inferences in biological data, we analyzed multi-tissue eQTL results from GTEx v8 (see **Material and Methods**). Specifically, we sought to identify latent factors corresponding to tissue-specific and tissue-shared eQTLs similar to ref[12]. Overall, we found that 27 latent factors explained 53.1% of the variance in the data (see Figure S5). Although we set $L = 18$ across all factors, we found the number of tissues with PIP > 0.9 is frequently lower than 18 in different factors (see Figure S4), which is due to inferred τ_{0kl} acting to "shut off" uninformative features. Indeed, we observed $30/486$ τ_{0kl} with estimates greater than e^{10} (see Figure S6) which effectively shrink the effect size of the corresponding single effect toward 0, driving the number of non-zero single effects in some factors smaller than specified L. We found this behavior also reflected in estimated level-0.9 credible sets, where 456/486 contained a single tissue, and the remaining 30 credible sets contained at least two tissues.

To understand what each factor represents, we examined inferred PIPs (Figure S8) and posterior mean weights of each tissue across 27 factors (Figure S7). Here we present the results from factor $\mathbf{z}_1$ and $\mathbf{z}_3$ through the posterior weights (Figure 3; see Figure S7 for the remainder). We observed that the latent factor $\mathbf{z}_1$ with the second largest PVE demonstrates high absolute weights on most tissues except for the brain tissues, while the latent factor $\mathbf{z}_3$ has large weights almost exclusively on brain tissues. Moreover, we observed that brain tissue tends to appear as a group and have similar effects, implying the eQTLs in brain tissue are different from that in other tissue and those strong signals are specifically captured by the factor $\mathbf{z}_1$. For the rest of the factors, we noticed that factors with large PVE such as $\mathbf{z}_2, \mathbf{z}_4, \mathbf{z}_5$ tended to have large weights on multiple tissues; for example, factor $\mathbf{z}_2$ has large weights on esophagus and thyroid, suggesting the eQTLs signals are mostly shared across those tissues; while the factors with small PVE usually have large weights exclusively on one or a few tissues, for example, liver-specific component $\mathbf{z}_{12}$, lung-specific component $\mathbf{z}_{15}$, etc. The only exception is that the factor $\mathbf{z}_0$ with the largest PVE has an exclusively large weight only on the testis, implying the $\mathbf{z}_0$ captures the testis-specific eQTL signals. This is consistent with the investigation of the latent factor values of $\mathbf{z}_0$: the gene with the largest factor value in z_0 is DDT (Figure S9), which is shown to be associated with testis cancer. [22] Overall, we find that SuSiE PCA is able to identify tissue-specific components from multi-tissue eQTL data in an intuitive, interpretable manner.

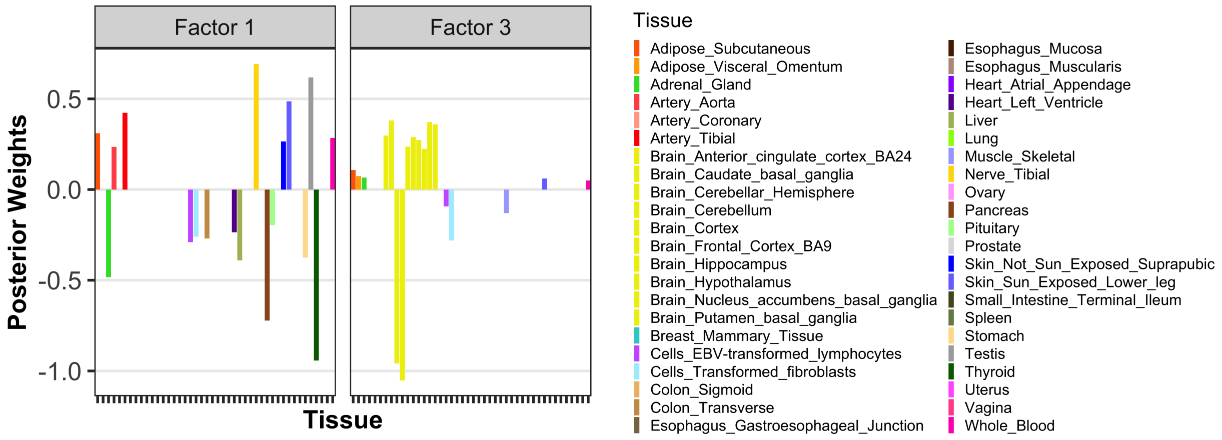

Fig. 3: Factor $\mathbf{z}_1$ and $\mathbf{z}_3$ captures different types of tissues (tissues without brain vs. brain tissues)

3.3 Perturb-seq Data

To identify gene regulatory modules from genome-wide perturbation data, we ran SuSiE PCA on perturb-seq in cell lines [15] (see Methods) with $K = 10$ and $L = 300$. Briefly, we inputted the normalized expression data (2057 × 8563) to SuSiE-PCA to identify gene regulatory modules (i.e. $\mathbf{Z}$) and downstream regulated networks (i.e. $\mathbf{W}$). To ensure our results were robust to K and L, we explored a grid of possible combinations and found that K=10 and L=300 retain the most important information while keeping the relevant gene set much smaller (see Figures S10 for a detailed explanation). As a comparison, we also ran sparse PCA with the default level of sparsity ($\alpha = 1$) and $K = 10$.

Overall, we found the total PVE was 10.71% across all components (Figure S11), with each component exhibiting 299 downstream genes with PIP > 0.9 on average. Focusing on the leading component, we found that perturbations with the top 10 largest absolute factor scores are primarily related to Ribosomal Protein Small (RPS) subunit genes and Ribosomal Protein Large subunit (RPL) family (Figure 4A). To provide a broader characterization of the module function, we extracted downstream genes with PIP greater than 0.9 (298 genes) as input into ShinyGO [23] to perform a gene set enrichment analysis (Figure 4B). We observed the most enriched pathway was related to ribosome function (FDR=9.2×10^{-82},63 genes involved), followed by Coronavirus disease (FDR=2.5×10^{-62},62 genes involved). Inspecting the loadings at these downstream genes, we found nearly all weights were positive, suggesting that the knockout of RPS and RPL genes down-regulate the expression level of those downstream genes. We found multiple elongation factor genes (EEF1G, EEF1A1, EEF1B2, EIF4B, EIF3L) among the leading downstream genes, which are known to be involved in ribosome function. Additionally, recent studies have suggested that the decreased expression of elongation factor genes is associated with less severe conditions among COVID-19 patients [24, 25]. We repeated pathway analysis for each latent factor using corresponding loadings at genes with PIP greater than 0.9 (see Figures S12-S20). To compare with Sparse PCA, we performed the same pathway analysis on factor loadings and assessed enrichments. We observed components identified by Sparse PCA to be less enriched with biological pathways when compared to SuSiE PCA (80 unique enriched pathways in sparse PCA versus 88 pathways in SuSiE PCA), and the top enriched pathways such as ribosome and coronavirus disease are less significant and contain less number of selected genes (FDR $= 1.4 \times 10^{-33}$,35 genes; FDR $= 2.9 \times 10^{-18}$,29 genes). Overall, we find distinct biological functions identified by each component, with groupings consistent with previous works [26, 27, 28].

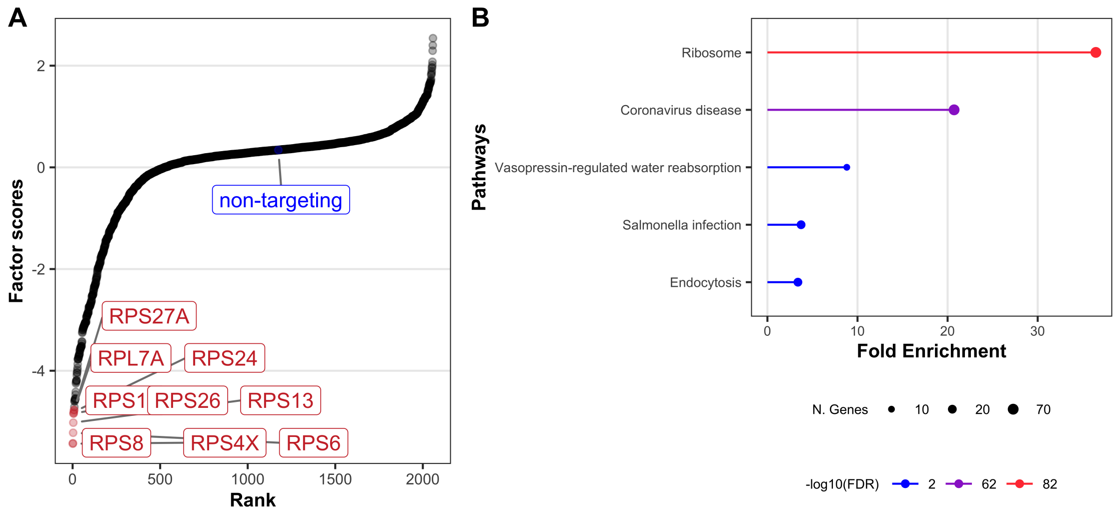

Fig. 4: The perturbations with top factor scores in the first component mostly belong to RPL and RPS family(A), and the enrichment analysis results of downstream genes in the same component are enriched for ribosome and coronavirus disease(B)
Each point in (A) represents the latent factor value of each perturbation. The top 9 points as well as the control group are labeled in the plot and colored red and blue, respectively. In gene set enrichment analysis, we input the downstream genes with PIP > 0.9 and show the top enriched pathways with log(FDR) and the number of genes included in the corresponding pathways.

Table 1: Comparison of mean and standard deviation of running time (seconds) between models

Model [4]	Simulation [5]	GTEx z-score	Perturb-seq
SuSiE PCA	3.14(0.49)	1.20	68.11
Sparse PCA	51.96(33.50)	41.22	1213.21
EBMF	39.83(5.80)	498.60	243.03

4 Discussion

In this paper, we propose SuSiE PCA, an efficient Bayesian variable selection approach to principal components analysis for structured biological data. The sparsity of the loading matrix is achieved by restricting the number of features associated with each factor to be at most L. Through simulations and real-data application, we find that SuSiE PCA outperforms existing approaches to sparse latent structure learning in identifying contributing features, while maintaining a more efficient run time.

There are several advantages of SuSiE PCA as compared to other sparse factor models. First, SuSiE PCA generates the posterior inclusion probabilities (PIPs) for each feature that quantifies the uncertainty of the selected feature, which can not be provided by other sparse models, such as sparse PCA with regularization or the Bayesian treatment of PPCA. And assessing the selected variables based on the probability is more reasonable and convenient than using weights. Second, PIPs are capable of selecting more signals with high confidence. In simulations, we demonstrated that using weights for variable selection from SuSiE PCA, sparse PCA, and EBMF can deliver a high specificity (low false discovery rate) but with low sensitivity as the cutoff value increase, while using PIPs as selection tools can maintain a high sensitivity for any positive cutoff value between 0 and 1. Third, SuSiE PCA provides a more precise estimate of the loadings and higher prediction accuracy, even in the misspecified case, as we impose a probabilistic distribution over the loadings that enables a much more accurate inference on the posterior distribution. Finally, the inference procedure of SuSiE PCA works on the dimension of K and L, which is typically set to be much smaller than feature dimension P; therefore, it is scalable to high-dimensional data and requires less computational demands. Actually, we implement the SuSiE PCA with the JAX library developed by Google to enable fast convergence on CPU, GPU, or TPU.

Although SuSiE PCA only allows one common L specified across all factors, the number of non-zero effects captured across factors can be varied and learned from the data. This is because we treat the inverse of variance τ_{0kl} of the l_{th} single effect in factor $\mathbf{z}_k$ as a random variable. As the Algorithm 1 (**Supplement Note**) demonstrates, the MLE of τ_{0kl} at the step 3 is derived before inference of other parameters. When the L specified in the model, for a certain factor k, is greater than the true number of signals associated with that factor, the MLE of the τ_{0kl} will be extremely large for those excessive single effects, which then shrinks the $\mathbf{w}_{kl}$ and PIP to be 0 or close to 0, and therefore removes the redundant single effects from the model. From this point of view, without prior knowledge of the data, one can specify a relatively larger L during the initial model fitting, and then examine the estimates of τ_{0kl} to explore how many single effects are reasonable for the dataset.

Overall, SuSiE PCA provides a flexible approach to high-dimensional biological data with a low-rank structure.

[4] All run-time data in the table are based on the analyses performed on the same CPU for consistency. The CPU we used is the Apple M2 chip with 16 GB memory.

[5] Run time for simulation are recorded based on simulation setting in Figure 1, i.e. $N = 1000, P = 6000, K = 4, L = 40$, the average run time and corresponding standard deviation are computed for 100 simulations. We presented a more detailed runtime comparison in simulation at Supplement figure S4

References

[1] H. Hotelling. "Analysis of a complex of statistical variables into principal components". In: *Journal of Educational Psychology* 24.6 (1933). Place: US Publisher: Warwick & York, pp. 417–441. ISSN: 1939-2176. DOI: 10.1037/h0071325.

[2] Nick Patterson, Alkes L. Price, and David Reich. "Population Structure and Eigenanalysis". In: *PLOS Genetics* 2.12 (Dec. 22, 2006). Publisher: Public Library of Science, e190. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0020190.

[3] Aman Agrawal et al. "Scalable probabilistic PCA for large-scale genetic variation data". In: *PLoS genetics* 16.5 (May 2020), e1008773. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1008773.

[4] Gil McVean. "A Genealogical Interpretation of Principal Components Analysis". In: *PLoS Genetics* 5.10 (Oct. 16, 2009), e1000686. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1000686. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2757795/ (visited on 05/04/2022).

[5] Jolliffe I.T. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[6] Hui Zou, Trevor Hastie, and Robert Tibshirani. "Sparse Principal Component Analysis". In: *Journal of Computational and Graphical Statistics* 15.2 (June 1, 2006), pp. 265–286. ISSN: 1061-8600. DOI: 10.1198/106186006X113430.

[7] Christopher Bishop. "Bayesian PCA". In: *Advances in Neural Information Processing Systems*. Vol. 11. MIT Press, 1998.

[8] Yue Guan and Jennifer Dy. "Sparse Probabilistic Principal Component Analysis". In: *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. ISSN: 1938-7228. PMLR, Apr. 15, 2009, pp. 185–192. URL: https://proceedings.mlr.press/v5/guan09a.html.

[9] Bo Ning. "Spike and slab Bayesian sparse principal component analysis". In: *arXiv:2102.00305 [stat]* (Jan. 30, 2021). arXiv: 2102.00305. (Visited on 05/04/2022).

[10] Artin Armagan, Merlise Clyde, and David Dunson. "Generalized Beta Mixtures of Gaussians". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc., 2011.

[11] Shiwen Zhao et al. "Bayesian group factor analysis with structured sparsity". In: *Journal of Machine Learning Research* 17.196 (2016), pp. 1–47. ISSN: 1533-7928.

[12] Wei Wang and Matthew Stephens. "Empirical Bayes Matrix Factorization". In: *arXiv:1802.06931 [stat]* (May 2, 2021). arXiv: 1802.06931. URL: http://arxiv.org/abs/1802.06931.

[13] Gao Wang et al. "A simple new approach to variable selection in regression, with application to genetic fine mapping". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.5 (2020). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12388, pp. 1273–1300. ISSN: 1467-9868. DOI: 10.1111/rssb.12388. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12388 (visited on 05/27/2022).

[14] THE GTEX CONSORTIUM et al. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". In: *Science* 348.6235 (May 8, 2015). Publisher: American Association for the Advancement of Science, pp. 648–660. DOI: 10.1126/science.1262110. URL: https://www.science.org/doi/10.1126/science.1262110.

[15] Joseph M. Replogle et al. "Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq". In: *Cell* 185.14 (July 7, 2022). Publisher: Elsevier, 2559–2575.e28. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2022.05.013. URL: https://www.cell.com/cell/abstract/S0092-8674(22)00597-9 (visited on 01/25/2023).

[16] Christophe Andrieu and C Andrieu. "An Introduction to MCMC for Machine Learning". In: (), p. 39.

[17] Michael I. Jordan et al. "An Introduction to Variational Methods for Graphical Models". In: *Machine Learning* 37.2 (Nov. 1, 1999), pp. 183–233. ISSN: 1573-0565. DOI: 10.1023/A:1007665907178.

[18] S. Kullback and R. A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951). Publisher: Institute of Mathematical Statistics, pp. 79–86. ISSN: 0003-4851.

[19] Toshiyuki Tanaka. "A Theory of Mean Field Approximation". In: *Advances in Neural Information Processing Systems*. Vol. 11. MIT Press, 1998.

[20] Ingwer Borg and Patrick Groenen. *Modern Multidimensional Scaling: Theory and Applications (Springer Series in Statistics)*. Aug. 1, 2005. DOI: 10.1007/978-1-4757-2711-1.

[21] Fanwang Meng et al. “Procrustes: A python library to find transformations that maximize the similarity between matrices”. In: *Computer Physics Communications* 276 (July 1, 2022), p. 108334. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2022.108334.

[22] Barbara A Cohn, Piera M Cirillo, and Roberta E Christianson. “Prenatal DDT Exposure and Testicular Cancer: A Nested Case-Control Study”. In: *Archives of environmental & occupational health* 65.3 (2010), pp. 127–134. ISSN: 1933-8244. DOI: 10.1080/19338241003730887. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2936455/.

[23] Steven Xijin Ge, Dongmin Jung, and Runan Yao. “ShinyGO: a graphical gene-set enrichment tool for animals and plants”. In: *Bioinformatics* 36.8 (Apr. 15, 2020). Ed. by Alfonso Valencia, pp. 2628–2629. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz931. URL: https://academic.oup.com/bioinformatics/article/36/8/2628/5688742.

[24] Junedh M. Amrute et al. “Cell specific peripheral immune responses predict survival in critical COVID-19 patients”. In: *Nature Communications* 13.1 (Feb. 15, 2022). Number: 1 Publisher: Nature Publishing Group, p. 882. ISSN: 2041-1723. DOI: 10.1038/s41467-022-28505-3. URL: https://www.nature.com/articles/s41467-022-28505-3 (visited on 02/09/2023).

[25] Manik Garg et al. “Meta-analysis of COVID-19 single-cell studies confirms eight key immune responses”. In: *Scientific Reports* 11 (Oct. 21, 2021), p. 20833. ISSN: 2045-2322. DOI: 10.1038/s41598-021-00121-z. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8531356/ (visited on 02/09/2023).

[26] Anna Signorile et al. “Prohibitins: A Critical Role in Mitochondrial Functions and Implication in Diseases”. In: *Cells* 8.1 (Jan. 18, 2019), p. 71. ISSN: 2073-4409. DOI: 10.3390/cells8010071. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6356732/ (visited on 02/10/2023).

[27] Marta Artal-Sanz et al. “The mitochondrial prohibitin complex is essential for embryonic viability and germline function in Caenorhabditis elegans”. In: *The Journal of Biological Chemistry* 278.34 (Aug. 22, 2003), pp. 32091–32099. ISSN: 0021-9258. DOI: 10.1074/jbc.M304877200.

[28] Marta Artal-Sanz and Nektarios Tavernarakis. “Prohibitin couples diapause signalling to mitochondrial metabolism during ageing in C. elegans”. In: *Nature* 461.7265 (Oct. 8, 2009), pp. 793–797. ISSN: 1476-4687. DOI: 10.1038/nature08466.

Supplement: A Scalable Bayesian Variable Selection Technique for Principal Component Analysis

Dong Yuan[1] and Nicholas Mancuso[1,2,3]

[1] Biostatistics Division, Dept of Population and Public Health Sciences,Keck School of Medicine, University of Southern California, Los Angeles, CA
[2] Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA
[3] Dept of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA

1 Inference on SuSiE PCA

Here we present all mathematical derivations of the inference on SuSiE PCA. The SuSiE PCA model has the following structure:

$$\begin{aligned}
\mathbf{X} \mid \mathbf{Z}, \mathbf{W}, \sigma^2 &\sim \mathcal{MN}_{N,P}(\mathbf{ZW}, \mathbf{I}_N, \sigma^2 \mathbf{I}_P) && (1.1)\\
\mathbf{Z} &\sim \mathcal{MN}_{N,K}(\mathbf{0}, \mathbf{I}_N, \mathbf{I}_K) && (1.2)\\
\mathbf{W} &= \sum_{k=1}^{K} \mathbf{e}_k \mathbf{w}_k^{\intercal} && (1.3)\\
\mathbf{w}_k &= \sum_{l=1}^{L} \mathbf{w}_{kl} && (1.4)\\
\mathbf{w}_{kl} &= w_{kl} \boldsymbol{\gamma}_{kl} && (1.5)\\
w_{kl} \mid \sigma^2_{0kl} &\sim \mathcal{N}(0, \sigma^2_{0kl}) && (1.6)\\
\boldsymbol{\gamma}_{kl} \mid \boldsymbol{\pi} &\sim \text{Multi}(1, \boldsymbol{\pi}). && (1.7)
\end{aligned}$$

First, the complete-data log-likelihood of data and parameters is given by:

$$\begin{aligned}
\ell_c(\sigma^2, \sigma_0^2, \boldsymbol{\pi} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}) &= \log \Pr(\mathbf{X} \mid \mathbf{Z}, \mathbf{W}, \sigma^2) + \log \Pr(\mathbf{Z}) + \log \Pr(\mathbf{W} \mid \sigma_0^2, \boldsymbol{\pi})\\
&= \log \mathcal{MN}_{n,p}(\mathbf{X} \mid \mathbf{ZW}, \mathbf{I}_n, \mathbf{I}_p \sigma^2) + \log \mathcal{MN}_{n,k}(\mathbf{Z} \mid \mathbf{0}, \mathbf{I}_n, \mathbf{I}_k) +\\
&\quad \sum_{l=1}^{L} \sum_{k=1}^{K} \left[\log \text{Multi}(\boldsymbol{\gamma}_{kl} \mid 1, \boldsymbol{\pi}) + \log \mathcal{N}(w_{kl} \mid 0, \sigma_0^2) \right]
\end{aligned}$$

Before proceeding to the full derivation of variational distribution of parameters $\mathbf{Z}, \mathbf{w}_{kl}$, and $\boldsymbol{\gamma}_{kl}$, we first give some helpful definitions, including the expansion of first and second moment of $\mathbf{W}$ and $\mathbf{Z}$, as well as the expansion of the log-likelihood function.

1.1 Helpful definitions

First and Second Moment of $\mathbf{w}_{\boldsymbol{k}}$

$$
\begin{aligned}
\mathbb{E}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] &= \text{p-vector of posterior conditional means} \\
\mathbb{V}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] &= \text{p-vector of posterior conditional variances} \\
\mathbb{E}[\mathbf{w}_k] &= \mathbb{E}[\sum_l \mathbf{w}_{kl}] = \sum_l \mathbb{E}[\mathbf{w}_{kl}] \\
\mathbb{E}[\mathbf{w}_{kl}] &= \sum_l \mathbb{E}[w_{kl} \mid \boldsymbol{\gamma}_{kl}] \circ \mathbb{E}[\boldsymbol{\gamma}_{kl}] \\
\mathbb{V}[\mathbf{w}_k] &= \mathbb{V}[\sum_l \mathbf{w}_{kl}] = \sum_l \mathbb{V}[\mathbf{w}_{kl}] \\
\mathbb{V}[\mathbf{w}_{kl}] &= \mathbb{E}[\mathbf{w}_{kl}\mathbf{w}_{kl}^\intercal] - \mathbb{E}[\mathbf{w}_{kl}]\mathbb{E}[\mathbf{w}_{kl}]^\intercal \\
&= \mathbb{E}[w_{kl}^2 \boldsymbol{\gamma}_{kl}\boldsymbol{\gamma}_{kl}^\intercal] - \mathbb{E}[\mathbf{w}_{kl}]\mathbb{E}[\mathbf{w}_{kl}]^\intercal \\
&= \operatorname{diag}(\mathbb{E}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] \circ \mathbb{E}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] \circ \mathbb{E}[\boldsymbol{\gamma}_{kl}]) - \mathbb{E}[\mathbf{w}_{kl}]\mathbb{E}[\mathbf{w}_{kl}]^\intercal \\
\operatorname{diag}(\mathbb{V}[\mathbf{w}_{kl}]) &= \mathbb{E}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] \circ \mathbb{E}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] \circ \mathbb{E}[\boldsymbol{\gamma}_{kl}] - (\mathbb{E}[\mathbf{w}_{kl} \mid \boldsymbol{\gamma}_{kl}] \circ \mathbb{E}[\boldsymbol{\gamma}_{kl}])^2 \\
\mathbb{E}[\mathbf{w}_k^\intercal \mathbf{w}_k] &= \operatorname{tr}(\mathbb{V}[\mathbf{w}_k]) + \mathbb{E}[\mathbf{w}_k]^\intercal \mathbb{E}[\mathbf{w}_k] \\
\mathbb{E}[w_{kl}^2] &= [\mathbb{E}^2[w_{kl} \mid \boldsymbol{\gamma}_{kl}] + \mathbb{V}[w_{kl} \mid \boldsymbol{\gamma}_{kl}]] \circ \mathbb{E}[\boldsymbol{\gamma}_{kl}]
\end{aligned}
$$

First and Second Moment of $\mathbf{W}$

$$
\begin{aligned}
\mathbb{E}[\mathbf{W}] &= \mathbb{E}[\sum_k \mathbf{e}_k \mathbf{w}_k^\intercal] = \sum_k \mathbf{e}_k \mathbb{E}[\mathbf{w}_k]^\intercal \\
\mathbb{E}[\mathbf{W}\mathbf{W}^\intercal] &= \mathbb{E}[(\sum_k \mathbf{e}_k \mathbf{w}_k^\intercal)(\sum_{k'} \mathbf{e}_{k'} \mathbf{w}_{k'}^\intercal)^\intercal] \\
&= \mathbb{E}[\sum_k \sum_{k'} \mathbf{e}_k \mathbf{w}_k^\intercal \mathbf{w}_{k'} \mathbf{e}_{k'}^\intercal] \\
&= \sum_k \sum_{k'} \mathbf{e}_k \mathbf{e}_{k'}^\intercal \mathbb{E}[\mathbf{w}_k^\intercal \mathbf{w}_{k'}] \\
&= \sum_k \sum_{k'} \mathbf{e}_k \mathbf{e}_{k'}^\intercal \mathbb{E}[\mathbf{w}_k]^\intercal \mathbb{E}[\mathbf{w}_{k'}] + \sum_k \mathbf{e}_k \mathbf{e}_k^\intercal (\mathbb{E}[\mathbf{w}_k^\intercal \mathbf{w}_k] - \mathbb{E}[\mathbf{w}_k]^\intercal \mathbb{E}[\mathbf{w}_k]) \\
&= \mathbb{E}[\mathbf{W}]\mathbb{E}[\mathbf{W}]^\intercal + \sum_k \mathbf{e}_k \mathbf{e}_k^\intercal (\mathbb{E}[\mathbf{w}_k^\intercal \mathbf{w}_k] - \mathbb{E}[\mathbf{w}_k]^\intercal \mathbb{E}[\mathbf{w}_k]) \\
&= \mathbb{E}[\mathbf{W}]\mathbb{E}[\mathbf{W}]^\intercal + \sum_k \mathbf{e}_k \mathbf{e}_k^\intercal \operatorname{tr}(\mathbb{V}[\mathbf{w}_k]) \\
&= \mathbb{E}[\mathbf{W}]\mathbb{E}[\mathbf{W}]^\intercal + \operatorname{diag}(\operatorname{tr}(\mathbb{V}[\mathbf{w}_1]), \ldots, \operatorname{tr}(\mathbb{V}[\mathbf{w}_k]))
\end{aligned}
$$

Second Moment of $\mathbf{Z}$

$$
\begin{aligned}
\mathbb{E}[\mathbf{Z}^\intercal \mathbf{Z}] &= \operatorname{tr}(\mathbf{I}_n)\boldsymbol{\Sigma}_{\mathbf{Z}} + \mathbb{E}[\mathbf{Z}]^\intercal \mathbb{E}[\mathbf{Z}] \\
&= n\boldsymbol{\Sigma}_{\mathbf{Z}} + \mathbb{E}[\mathbf{Z}]^\intercal \mathbb{E}[\mathbf{Z}] \\
\mathbb{E}[\mathbf{Z}_k^\intercal \mathbf{Z}_k] &= \operatorname{tr}(\mathbb{V}[\mathbf{Z}_k]) + \mathbb{E}[\mathbf{Z}_k]^\intercal \mathbb{E}[\mathbf{Z}_k] \\
&= \operatorname{tr}(\mathbf{I}_n(\boldsymbol{\Sigma}_{\mathbf{Z}})_{kk}) + \mathbb{E}[\mathbf{Z}_k]^\intercal \mathbb{E}[\mathbf{Z}_k] \\
&= n(\boldsymbol{\Sigma}_{\mathbf{Z}})_{kk} + \mathbb{E}[\mathbf{Z}_k]^\intercal \mathbb{E}[\mathbf{Z}_k]
\end{aligned}
$$

Other terms in Log Likelihood

$$\begin{aligned}
\log \mathcal{MN}_{n,p}(\mathbf{X} \mid \mathbf{ZW}, \mathbf{I}_n, \mathbf{I}_p\sigma^2) &= -\frac{1}{2\sigma^2}\mathrm{tr}\left[(\mathbf{X}-\mathbf{ZW})^\mathsf{T}(\mathbf{X}-\mathbf{ZW})\right] - \frac{np}{2}\log(2\pi\sigma^2) \\
\log \mathcal{MN}_{n,k}(\mathbf{Z} \mid \mathbf{0}, \mathbf{I}_n, \mathbf{I}_k) &= -\frac{1}{2}\mathrm{tr}\left[\mathbf{Z}^\mathsf{T}\mathbf{Z}\right] - \frac{nk}{2}\log(2\pi) \\
\log \mathrm{Multi}(\boldsymbol{\gamma}_{kl} \mid 1, \boldsymbol{\pi}) &= \sum_{i=1}^{p} \boldsymbol{\gamma}_{kli}\log(\pi_i) \\
\log \mathcal{N}(w_{kl} \mid 0, \sigma_0^2) &= -\frac{1}{2\sigma_0^2}w_{kl}^2 - \frac{1}{2}\log(2\pi\sigma_0^2) \\
\mathrm{tr}\left[\mathbb{E}_{\neg\mathbf{Z}}\left[(\mathbf{X}-\mathbf{ZW})^\mathsf{T}(\mathbf{X}-\mathbf{ZW})\right]\right] &= \mathrm{tr}\left[\mathbb{E}_{\neg\mathbf{Z}}(\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{ZW} - \mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X} + \mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{ZW})\right] \\
&= \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) - 2\mathrm{tr}(\mathbb{E}[\mathbf{W}]\mathbf{X}^\mathsf{T}\mathbf{Z}) + \mathrm{tr}(\mathbf{Z}^\mathsf{T}\mathbf{Z}\mathbb{E}[\mathbf{W}\mathbf{W}^\mathsf{T}]) \\
&= \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) - 2\mathrm{tr}(\mathbb{E}[\mathbf{W}]\mathbf{X}^\mathsf{T}\mathbf{Z}) + \sum_{i=1}^{p}\mathrm{tr}(\mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{W}_i}\mathbf{Z}^\mathsf{T}) + \mathbb{E}[\mathbf{W}^\mathsf{T}]\mathbf{Z}^\mathsf{T}\mathbf{Z}\mathbb{E}[\mathbf{W}] \\
\mathrm{tr}\left[\mathbb{E}_{\neg\mathbf{W}}\left[(\mathbf{X}-\mathbf{ZW})^\mathsf{T}(\mathbf{X}-\mathbf{ZW})\right]\right] &= \mathrm{tr}\left[\mathbb{E}_{\neg\mathbf{W}}(\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{ZW} - \mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X} + \mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{ZW})\right] \\
&= \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) - 2\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbb{E}[\mathbf{Z}]\mathbf{W}) + \mathrm{tr}(\mathbb{E}[\mathbf{Z}^T\mathbf{Z}]\mathbf{W}\mathbf{W}^\mathsf{T})
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[\mathrm{tr}((\mathbf{X}-\mathbf{ZW})^\mathsf{T}(\mathbf{X}-\mathbf{ZW}))] &= \mathbb{E}[\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbf{ZW} - \mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{X} + \mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{ZW})] \\
&= \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X} - \mathbf{X}^\mathsf{T}\mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{W}] - \mathbb{E}[\mathbf{W}^\mathsf{T}]\mathbb{E}[\mathbf{Z}^\mathsf{T}]\mathbf{X} + \mathbb{E}[\mathbf{W}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{ZW}]) \\
&= \mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbf{X}) - 2\mathrm{tr}(\mathbf{X}^\mathsf{T}\mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{W}]) + \mathrm{tr}(\mathbb{E}[\mathbf{Z}^\mathsf{T}\mathbf{Z}]\mathbb{E}[\mathbf{W}\mathbf{W}^\mathsf{T}]) \\
\overline{\mathbf{R}}_{kl} &:= \mathbf{X} - \mathbb{E}[\mathbf{Z}](\sum_{k'\neq k}\mathbf{e}_{k'}\mathbb{E}[\mathbf{w}_{k'}]^\mathsf{T} + \sum_{l'\neq l}\mathbf{e}_{k}\mathbb{E}[\mathbf{w}_{kl'}]^\mathsf{T}) \\
&= \mathbf{X} - \sum_{k'\neq k}\mathbb{E}[\mathbf{Z}_{k'}]\mathbb{E}[\mathbf{w}_{k'}]^\mathsf{T} - \sum_{l'\neq l}\mathbb{E}[\mathbf{Z}_{k}]\mathbb{E}[\mathbf{w}_{kl'}]^\mathsf{T}
\end{aligned}$$

1.2 Derivation of Variational Distributions

In this section, we formally present the detailed derivation of variational distributions of all variables in the model including $Q(\mathbf{Z}), Q(w_{kl} \mid \gamma_{kli} = 1), Q(\gamma_{kl})$. The derivation of the optimal variational distribution is based on the mean-field approximation (see **Methods**) and the corresponding equation (2.10). For the ease of notation, let $\tau = \frac{1}{\sigma^2}$, $\mathbf{p} = \log \boldsymbol{\pi}$ and $\tau_0 = \frac{1}{\sigma_0^2}$.

Derivation of $\log Q(Z)$

$$\begin{aligned}
\log Q(\mathbf{Z}) &= \mathbb{E}_{\neg \mathbf{Z}} \left[\ell_c(\sigma^2, \sigma_0^2, \boldsymbol{\pi} \mid \mathbf{X}, \mathbf{Z}, \mathbf{W})\right] \\
&= \mathbb{E}_{\neg \mathbf{Z}} \left[\log \mathcal{MN}_{n,p}(\mathbf{X} \mid \mathbf{ZW}, \mathbf{I}_n, \mathbf{I}_p \sigma^2)\right] + \log \mathcal{MN}_{n,k}(\mathbf{Z} \mid \mathbf{0}, \mathbf{I}_n, \mathbf{I}_k) \\
&= -\frac{\tau}{2} \left[-\mathrm{tr}(\mathbf{X}^\intercal \mathbf{Z} \mathbb{E}[\mathbf{W}]) - \mathrm{tr}(\mathbb{E}[\mathbf{W}^\intercal] \mathbf{Z}^\intercal \mathbf{X}) + \mathrm{tr}(\mathbf{Z}^\intercal \mathbf{Z} \mathbb{E}(\mathbf{W}\mathbf{W}^\intercal))\right] - \frac{1}{2} \mathrm{tr}(\mathbf{Z}^\intercal \mathbf{Z}) + O(1) \\
&= -\frac{1}{2} \left[\mathrm{tr}(\tau \mathbf{Z}^\intercal \mathbf{Z} \mathbb{E}(\mathbf{W}\mathbf{W}^\intercal)) + \mathrm{tr}(\mathbf{Z}^\intercal \mathbf{Z}) - \mathrm{tr}(\tau \mathbf{X}^\intercal \mathbf{Z} \mathbb{E}[\mathbf{W}]) - \mathrm{tr}(\tau \mathbb{E}[\mathbf{W}^\intercal] \mathbf{Z}^\intercal \mathbf{X})\right] + O(1) \\
&= -\frac{1}{2} \left[\mathrm{tr}(\mathbf{Z}^\intercal \mathbf{Z} (\mathbb{E}(\mathbf{W}\mathbf{W}^\intercal)\tau + \mathbf{I}_k)) - \mathrm{tr}(\tau \mathbf{X}^\intercal \mathbf{Z} \mathbb{E}[\mathbf{W}]) - \mathrm{tr}(\tau \mathbb{E}[\mathbf{W}^\intercal] \mathbf{Z}^\intercal \mathbf{X})\right] + O(1) \\
&= -\frac{1}{2} \left[\mathrm{tr}(\mathbf{Z} \underbrace{(\mathbb{E}(\mathbf{W}\mathbf{W}^\intercal)\tau + \mathbf{I}_k)}_{\boldsymbol{\Sigma}_\mathbf{Z}^{-1}} \mathbf{Z}^\intercal) - \mathrm{tr}(\mathbf{Z} \mathbb{E}[\mathbf{W}] \mathbf{X}^\intercal \tau) - \mathrm{tr}(\tau \mathbf{X} \mathbb{E}[\mathbf{W}^\intercal] \mathbf{Z}^\intercal)\right] + O(1) \\
&= -\frac{1}{2} \left[\mathrm{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\mathbf{Z}^{-1} \mathbf{Z}^\intercal) - \mathrm{tr}(\mathbf{Z} \boldsymbol{\Sigma}_\mathbf{Z}^{-1} \underbrace{\boldsymbol{\Sigma}_\mathbf{Z} \mathbb{E}[\mathbf{W}] \mathbf{X}^\intercal \tau}_{\boldsymbol{\mu}_\mathbf{Z}^\intercal}) - \mathrm{tr}(\underbrace{\tau \mathbf{X} \mathbb{E}[\mathbf{W}^\intercal] \boldsymbol{\Sigma}_\mathbf{Z}}_{\boldsymbol{\mu}_\mathbf{Z}} \boldsymbol{\Sigma}_\mathbf{Z}^{-1} \mathbf{Z}^\intercal)\right] + O(1) \\
&= -\frac{1}{2} \left[\mathrm{tr}(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} \mathbf{Z}^\intercal \mathbf{Z}) - \mathrm{tr}(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} \boldsymbol{\mu}_\mathbf{Z}^\intercal \mathbf{Z}) - \mathrm{tr}(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} \mathbf{Z}^T \boldsymbol{\mu}_\mathbf{Z}) + \mathrm{tr}(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} \boldsymbol{\mu}_\mathbf{Z}^T \boldsymbol{\mu}_\mathbf{Z}) - \mathrm{tr}(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} \boldsymbol{\mu}_\mathbf{Z}^T \boldsymbol{\mu}_\mathbf{Z})\right] + O(1) \\
&= -\frac{1}{2} \left[\mathrm{tr}(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} (\mathbf{Z}^\intercal \mathbf{Z} - \mathbf{Z}^T \boldsymbol{\mu}_\mathbf{Z} - \boldsymbol{\mu}_\mathbf{Z}^T \mathbf{Z} + \boldsymbol{\mu}_\mathbf{Z}^T \boldsymbol{\mu}_\mathbf{Z}))\right] + O(1) \\
&= -\frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}_\mathbf{Z}^{-1} (\mathbf{Z} - \boldsymbol{\mu}_\mathbf{Z})^\intercal (\mathbf{Z} - \boldsymbol{\mu}_\mathbf{Z})\right) + O(1) \Rightarrow \\
Q(\mathbf{Z}) &= \mathcal{MN}_{n,k}(\mathbf{Z} \mid \boldsymbol{\mu}_\mathbf{Z}, \mathbf{I}_n, \boldsymbol{\Sigma}_\mathbf{Z})
\end{aligned}$$

Derivation of $\log Q(w_{kl} \mid \gamma_{kli} = 1)$

$$
\begin{aligned}
\log Q(w_{kl}|\gamma_{kli}=1) &= -\frac{\tau}{2}\mathbb{E}_{\neg w_{kl}}[\mathrm{tr}((\mathbf{R}_{kl}-\mathbf{Z}_k\mathbf{w}_{kl}^\intercal)^\intercal(\mathbf{R}_{kl}-\mathbf{Z}_k\mathbf{w}_{kl}^\intercal))] - \frac{\tau_0}{2}\mathbb{E}_{\neg w_{kl}}[\sum_{l=1}^{L}\sum_{k=1}^{K} w_{kl}^2] + O(1) \\
&= -\frac{\tau}{2}\mathbb{E}_{\neg w_{kl}}[-2\mathrm{tr}(\mathbf{R}_{kl}^\intercal\mathbf{Z}_k\mathbf{w}_{kl}^\intercal) + \mathrm{tr}(\mathbf{Z}_k^\intercal\mathbf{Z}_k\mathbf{w}_{kl}^\intercal\mathbf{w}_{kl})] - \frac{\tau_0}{2}w_{kl}^2 + O(1) \\
&= -\frac{\tau}{2}\mathbb{E}_{\neg w_{kl}}\left[-2\mathrm{tr}((\mathbf{X} - \sum_{k'\neq k}\mathbf{Z}_{k'}\mathbf{w}_{k'}^\intercal - \sum_{l'\neq l}\mathbf{Z}_k\mathbf{w}_{kl'}^\intercal)^\intercal\mathbf{Z}_k\mathbf{w}_{kl}^\intercal) + \mathrm{tr}(\mathbf{Z}_k^\intercal\mathbf{Z}_k w_{kl}^2)\right] \\
&\quad - \frac{\tau_0}{2}w_{kl}^2 + O(1) \\
&= -\frac{\tau}{2}\mathbb{E}_{\neg w_{kl}}\left[-2\mathrm{tr}((\mathbf{X}^\intercal\mathbf{Z}_k - \sum_{k'\neq k}\mathbf{w}_{k'}\mathbf{Z}_{k'}^\intercal\mathbf{Z}_k - \sum_{l'\neq l}\mathbf{w}_{kl'}\mathbf{Z}_k^\intercal\mathbf{Z}_k)\mathbf{w}_{kl}^\intercal) + \mathrm{tr}(\mathbf{Z}_k^\intercal\mathbf{Z}_k w_{kl}^2)\right] \\
&\quad - \frac{\tau_0}{2}w_{kl}^2 + O(1) \\
&= -\frac{\tau}{2}\mathbb{E}_{\neg w_{kl}}\left[-2(\mathbf{X}_i^\intercal\mathbf{Z}_k - \sum_{k'\neq k}\mathbf{w}_{k',i}\mathbf{Z}_{k'}^\intercal\mathbf{Z}_k - \mathbf{Z}_k^\intercal\mathbf{Z}_k\sum_{l'\neq l}\mathbf{w}_{kl'i})w_{kl} + \mathbf{Z}_k^\intercal\mathbf{Z}_k w_{kl}^2\right] \\
&\quad - \frac{\tau_0}{2}w_{kl}^2 + O(1) \\
&= -\frac{1}{2}[-2\tau(\mathbf{X}_i^\intercal\mathbb{E}[\mathbf{Z}_k] - \sum_{k'\neq k}\mathbb{E}[\mathbf{w}_{k',i}]\mathbb{E}[\mathbf{Z}_{k'}^\intercal\mathbf{Z}_k] - \mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k]\sum_{l'\neq l}\mathbb{E}[\mathbf{w}_{kl'i}])w_{kl} + \tau\mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k]w_{kl}^2] \\
&\quad - \frac{\tau_0}{2}w_{kl}^2 + O(1) \\
&= -\frac{1}{2}[-2\tau(\mathbf{X}_i^\intercal\mathbb{E}[\mathbf{Z}_k] - \sum_{k'\neq k}\mathbb{E}[\mathbf{w}_{k',i}]\mathbb{E}[\mathbf{Z}_{k'}^\intercal\mathbf{Z}_k] - \mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k]\sum_{l'\neq l}\mathbb{E}[\mathbf{w}_{kl'i}])w_{kl} + \tau\mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k]w_{kl}^2 \\
&\quad + \tau_0 w_{kl}^2] + O(1) \\
&= -\frac{1}{2}[-2\tau(\mathbf{X}_i^\intercal\mathbb{E}[\mathbf{Z}_k] - \sum_{k'\neq k}\mathbb{E}[\mathbf{w}_{k',i}]\mathbb{E}[\mathbf{Z}_{k'}^\intercal\mathbf{Z}_k] - \mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k]\sum_{l'\neq l}\mathbb{E}[\mathbf{w}_{kl'i}])w_{kl} \\
&\quad + \underbrace{(\tau\mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k] + \tau_0)}_{1/\sigma_{w_{kl}}^2} w_{kl}^2] + O(1) \\
&= -\frac{1}{2\sigma_{w_{kl}}^2}[w_{kl}^2 - 2\underbrace{\tau\sigma_{w_{kl}}^2(\mathbf{X}_i^\intercal\mathbb{E}[\mathbf{Z}_k] - \sum_{k'\neq k}\mathbb{E}[\mathbf{w}_{k',i}]\mathbb{E}[\mathbf{Z}_{k'}^\intercal\mathbf{Z}_k] - \mathbb{E}[\mathbf{Z}_k^\intercal\mathbf{Z}_k]\sum_{l'\neq l}\mathbb{E}[\mathbf{w}_{kl'i}])}_{\mu_{w_{kl}}} w_{kl}] \Rightarrow \\
&= \log\mathcal{N}(\mu_{\mathbf{w}_{kl}}, \sigma_{\mathbf{w}_{kl}}^2)
\end{aligned}
$$

We can update $\mathbf{w}_{kl}$ for all feature at once:

$$\boldsymbol{\mu}_{\mathbf{w}_{kl}} = \tau\sigma_{w_{kl}}^2\mathbb{E}[\mathbf{R}_{kl}^\intercal\mathbf{Z}_k], \boldsymbol{\Sigma}_{\mathbf{w}_{kl}} = \sigma_{w_{kl}}^2\mathbf{I}_p$$

Derivation of $\log Q(\boldsymbol{\gamma}_{kl})$ Note that $\tau\mathbf{R}_{kl}^{\intercal}\mathbb{E}[\mathbf{Z}_k] = \mathbb{E}[w_{kl} \mid \gamma_{kl}]/\sigma^2_{w_{kl}} = \boldsymbol{\mu}_{\mathbf{w}_{kl}}/\sigma^2_{w_{kl}}$.

$$\begin{aligned}
\log Q(\boldsymbol{\gamma}_{kli}=1) &= \mathbb{E}_{\neg\boldsymbol{\gamma}_{kl}}[\ell_c(\sigma^2,\sigma_0^2,\boldsymbol{\pi},\mid \mathbf{X},\mathbf{Z},\mathbf{W})] + \log\mathrm{Multi}(\boldsymbol{\gamma}_{kl}\mid\boldsymbol{\pi}) + O(1)\\
&= -\frac{\tau}{2}\mathbb{E}_{\neg\boldsymbol{\gamma}_{kl}}\mathrm{tr}((\mathbf{R}_{kl}-\mathbf{Z}_k\mathbf{w}_{kl}^{\intercal})^{\intercal}(\mathbf{R}_{kl}-\mathbf{Z}_k\mathbf{w}_{kl}^{\intercal})) + \log\mathrm{Multi}(\boldsymbol{\gamma}_{kl}\mid\boldsymbol{\pi}) + O(1)\\
&= -\frac{\tau}{2}[-2\mathrm{tr}(\mathbf{R}_{kl}^{\intercal}\mathbb{E}[\mathbf{Z}_k]\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]\boldsymbol{\gamma}_{kl}^{\intercal}) + \mathrm{tr}(\mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k]\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1])] + \log\boldsymbol{\pi}_i + O(1)\\
&= -\frac{\tau}{2}\left[-2\mathbf{R}_{kli}^{\intercal}\mathbb{E}[\mathbf{Z}_k]\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1] + \mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k]\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kl}]\right] + \log\boldsymbol{\pi}_i + O(1)\\
&= \tau\mathbf{R}_{kli}^{\intercal}\mathbb{E}[\mathbf{Z}_k]\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1] - \frac{\tau}{2}\mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k]\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1] + \log\boldsymbol{\pi}_i + O(1)\\
&= \frac{1}{\sigma^2_{w_{kl}}}\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2 - \frac{\tau}{2}\mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k]\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1] + \log\boldsymbol{\pi}_i - \frac{\tau_0}{2}\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1] + O(1)\\
&= \frac{1}{\sigma^2_{w_{kl}}}\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2 - \frac{1}{2}\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1](\tau\mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k]+\tau_0) + \log\boldsymbol{\pi}_i + O(1)\\
&= \frac{1}{\sigma^2_{w_{kl}}}\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2 - \frac{1}{2\sigma^2_{w_{kl}}}\mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1] + \log\boldsymbol{\pi}_i + O(1)\\
&= -\frac{1}{2\sigma^2_{w_{kl}}}\left[-2\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2 + \mathbb{E}[w_{kl}^2\mid\boldsymbol{\gamma}_{kli}=1]\right] + \log\boldsymbol{\pi}_i + O(1)\\
&= -\frac{1}{2\sigma^2_{w_{kl}}}\left[-2\mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2 + \sigma^2_{w_{kl}} + \mathbb{E}[w_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2\right] + \log\boldsymbol{\pi}_i + O(1)\\
&= \frac{1}{2\sigma^2_{\mathbf{w}_{kl}}}\mathbb{E}[\mathbf{w}_{kl}\mid\boldsymbol{\gamma}_{kli}=1]^2 + \log\boldsymbol{\pi}_i + O(1) \Rightarrow\\
\log\tilde{\boldsymbol{\alpha}}_{kli} &= \log\boldsymbol{\pi}_i - \log\mathcal{N}(0\mid\mu_{\mathbf{w}_{kl}},\sigma^2_{\mathbf{w}_{kl}})\\
Q(\boldsymbol{\gamma}_{kl}) &= \mathrm{Multi}(1,\boldsymbol{\alpha}_{kl} = \mathrm{softmax}(\log\tilde{\boldsymbol{\alpha}}_{kl}))
\end{aligned}$$

1.3 Derivation of Evidence Lower Bound(ELBO)

To compute the maximum likelihood estimate of the variance terms τ and τ_0 and determine the likelihood of the data under SuSiE PCA, we write out the Evidence Lower Bound (ELBO).

$$\begin{aligned}
\mathrm{ELBO}(\mathbf{W},\mathbf{Z}) &= \mathbb{E}_Q\left[\log\Pr(\mathbf{X},\mathbf{Z},\mathbf{W}) - \log Q(\mathbf{Z},\mathbf{W})\right]\\
&= \mathbb{E}_Q[\log\Pr(\mathbf{X}|\mathbf{Z},\mathbf{W})] + \mathbb{E}_Q[\log\Pr(\mathbf{Z},\mathbf{W}) - \log Q(\mathbf{Z},\mathbf{W})]\\
&= \mathbb{E}_Q[\log\Pr(\mathbf{X}|\mathbf{Z},\mathbf{W})] + \mathbb{E}_{Q(\mathbf{Z})}[\log\Pr(\mathbf{Z}) - \log Q(\mathbf{Z})] +\\
&\quad \sum_{l=1}^{L}\left[\mathbb{E}_{Q(\mathbf{w}_l|\boldsymbol{\Gamma}_l)}[\log\Pr(\mathbf{w}_l|\boldsymbol{\Gamma}_l) - \log Q(\mathbf{w}_l|\boldsymbol{\Gamma}_l)] + \mathbb{E}_{Q(\boldsymbol{\Gamma}_l)}[\log\Pr(\boldsymbol{\Gamma}_l) - \log Q(\boldsymbol{\Gamma}_l)]\right]\\
&= \mathbb{E}_Q[\log\Pr(\mathbf{X}|\mathbf{Z},\mathbf{W})] + \mathbb{E}_{Q(\mathbf{Z})}[\log\Pr(\mathbf{Z}) - \log Q(\mathbf{Z})]\\
&\quad + \mathbb{E}_{Q(\mathbf{W},\boldsymbol{\Gamma})}[\log\Pr(\mathbf{W},\boldsymbol{\Gamma}) - \log Q(\mathbf{W},\boldsymbol{\Gamma})]
\end{aligned}$$

The first is the expectation of the data with respect to all the parameters in the model:

$$\begin{aligned}
\mathbb{E}_Q[\log\Pr(\mathbf{X}|\mathbf{Z},\mathbf{W},\boldsymbol{\Gamma})] &= \mathbb{E}_Q\left[-\frac{1}{2\sigma^2}\mathrm{tr}\left[(\mathbf{X}-\mathbf{Z}\mathbf{W})^{\intercal}(\mathbf{X}-\mathbf{Z}\mathbf{W})\right] - \frac{np}{2}\log(2\pi\sigma^2)\right]\\
&= -\frac{1}{2\sigma^2}\left[\mathrm{tr}(\mathbf{X}^{\intercal}\mathbf{X}) - 2\mathrm{tr}(\mathbf{X}^{\intercal}\mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{W}]) + \mathrm{tr}(\mathbb{E}[\mathbf{Z}^{\intercal}\mathbf{Z}]\mathbb{E}[\mathbf{W}\mathbf{W}^{\intercal}])\right] - \frac{np}{2}\log(2\pi\sigma^2)
\end{aligned}$$

The second term is the negative KL divergence of $\mathbf{Z}$.

$$\begin{aligned}
\mathbb{E}_{Q(\mathbf{Z})}[\log \Pr(\mathbf{Z}) - \log Q(\mathbf{Z})] &= \mathbb{E}[-\frac{1}{2}\mathrm{tr}(\mathbf{Z}^\intercal \mathbf{Z}) - \frac{nk}{2}\log(2\pi) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})^\intercal(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})) \\
&\quad + \frac{nk}{2}\log(2\pi) + \frac{n}{2}\log(|\boldsymbol{\Sigma}_{\mathbf{Z}}|)] \\
&= -\frac{1}{2}\mathrm{tr}(\mathbb{E}[\mathbf{Z}^\intercal \mathbf{Z}]) + \frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}(E[\mathbf{Z}^\intercal \mathbf{Z}] - \boldsymbol{\mu}_{\mathbf{Z}}^\intercal \boldsymbol{\mu}_{\mathbf{Z}})] + \frac{N}{2}\log(|\boldsymbol{\Sigma}_{\mathbf{Z}}|) \\
&= -\frac{1}{2}\mathrm{tr}(\mathbb{E}[\mathbf{Z}^\intercal \mathbf{Z}]) + \frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}_{\mathbf{Z}}^{-1}(N\boldsymbol{\Sigma}_{\mathbf{Z}} + \boldsymbol{\mu}_{\mathbf{Z}}^\intercal \boldsymbol{\mu}_{\mathbf{Z}} - \boldsymbol{\mu}_{\mathbf{Z}}^\intercal \boldsymbol{\mu}_{\mathbf{Z}})] + \frac{N}{2}\log(|\boldsymbol{\Sigma}_{\mathbf{Z}}|) \\
&= -\frac{1}{2}\mathrm{tr}(\mathbb{E}[\mathbf{Z}^\intercal \mathbf{Z}]) + \frac{NK}{2} + \frac{N}{2}\log(|\boldsymbol{\Sigma}_{\mathbf{Z}}|)
\end{aligned}$$

The last term contains joint negative KL divergence of $\mathbf{W}$ and $\boldsymbol{\Gamma}$ can be further decomposed as following:

$$\begin{aligned}
\mathbb{E}_{Q(\mathbf{W},\boldsymbol{\Gamma})}[\log \Pr(\mathbf{W}, \boldsymbol{\Gamma}) - \log Q(\mathbf{W}, \boldsymbol{\Gamma})] &= \mathbb{E}_{Q(\mathbf{W},\boldsymbol{\Gamma})}[\log \Pr(\mathbf{W} \mid \boldsymbol{\Gamma}) \Pr(\boldsymbol{\Gamma}) - \log Q(\mathbf{W} \mid \boldsymbol{\Gamma}) Q(\boldsymbol{\Gamma})] \\
&= \mathbb{E}_{Q(\mathbf{W},\boldsymbol{\Gamma})}[\log \Pr(\mathbf{W} \mid \boldsymbol{\Gamma}) - \log Q(\mathbf{W} \mid \boldsymbol{\Gamma})] \\
&\quad + \mathbb{E}_{Q(\mathbf{W},\boldsymbol{\Gamma})}[\log \Pr(\boldsymbol{\Gamma}) - \log Q(\boldsymbol{\Gamma})] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{E}_{Q(\mathbf{w}_{kl},\boldsymbol{\gamma}_{kl})}[\log \Pr(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl}) - \log Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl})] + \\
&\quad \sum_{k=1}^{K}\sum_{l=1}^{L} \mathbb{E}_{Q(\boldsymbol{\gamma}_{kl})}[\log \Pr(\boldsymbol{\gamma}_{kl}) - \log Q(\boldsymbol{\gamma}_{kl})] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{P} \boldsymbol{\alpha}_{kli} \mathbb{E}_{Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl})}[\log \Pr(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kli} = 1) \\
&\quad - \log Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kli} = 1)] + \\
&\quad \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{P} \mathbb{E}_{\boldsymbol{\gamma}_{kl}}[\log \Pr(\boldsymbol{\gamma}_{kli} = 1) - \log Q(\boldsymbol{\gamma}_{kli} = 1)]
\end{aligned}$$

The first expectation term of the last line of equation $\mathbb{E}_{Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl})}$ can be expanded as following:

$$\begin{aligned}
\mathbb{E}_{Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl})}\left[\log \frac{\Pr(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl})}{Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kl})}\right] &= \sum_{i=1}^{P} \mathbb{E}_{Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kli}=1)}\left[\log \frac{\Pr(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kli} = 1)}{Q(\mathbf{w}_{kl}|\boldsymbol{\gamma}_{kli} = 1)}\right] \\
&= \mathbb{E}[\sum_{i=1}^{P}[-\frac{\tau_0}{2}(w_{kli})^2 + \frac{1}{2\sigma_{w_{kl}}^2}(w_{kli} - \boldsymbol{\mu}_{w_{kli}})^2] \\
&\quad - \frac{p}{2}log(2\pi/\tau_0) + \frac{p}{2}\log(2\pi\sigma_{w_{kl}}^2)] \\
&= \sum_{i=1}^{P}\left[(-\frac{\tau_0}{2} + \frac{1}{2\sigma_{w_{kl}}^2})\mathbb{E}[(w_{kli})^2] - \frac{1}{2\sigma_{w_{kl}}^2}\mu_{w_{kli}}^2\right] - \frac{P}{2}log(2\pi/\tau_0) + \frac{P}{2}\log(2\pi\sigma_{w_{kl}}^2) \\
&= \sum_{i=1}^{P}\left[(-\frac{\tau_0}{2} + \frac{1}{2\sigma_{w_{kl}}^2})[\mu_{w_{kli}}^2 + \sigma_{w_{kl}}^2] - \frac{1}{2\sigma_{w_{kl}}^2}\mu_{w_{kli}}^2\right] + \frac{P}{2}\log(\sigma_{w_{kl}}^2\tau_0) \\
&= \sum_{i=1}^{P}[-\frac{\tau_0}{2}\mu_{w_{kli}}^2 - \frac{\tau_0}{2}\sigma_{w_{kl}}^2 + \frac{1}{2}] + \frac{P}{2}\log(\sigma_{w_{kl}}^2\tau_0)
\end{aligned}$$

And the second expectation term $\mathbb{E}_{\boldsymbol{\gamma}_{kl}}$ can be decomposed as

$$\begin{aligned}
\mathbb{E}_{Q(\boldsymbol{\Gamma})}[\log \Pr(\boldsymbol{\Gamma}) - \log Q(\boldsymbol{\Gamma})] &= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{P} \mathbb{E}_{Q(\boldsymbol{\gamma}_{kli}=1)}\left[(\boldsymbol{\gamma}_{kli}\log \pi_i - \boldsymbol{\gamma}_{kli}\log \alpha_{kli})\right] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{P}[\mathbb{E}(\boldsymbol{\gamma}_{kli})\log(\pi_i) - \mathbb{E}(\boldsymbol{\gamma}_{kli})\log(\alpha_{kli})] \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{P}[\alpha_{kli}(\log(\pi_i) - \log \alpha_{kli})]
\end{aligned}$$

Finally, we provide the algorithm for SuSiE PCA.

Algorithm 1 Algorithm for SuSiE PCA

```
Require: Data $\mathbf{X}_{N\times P}$
Require: Number of Factors $K$; Number of single effects in each factor $L$
Require: Initialize variational parameters $(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}}; \boldsymbol{\mu}_{\mathbf{w}_{kl}}, \boldsymbol{\Sigma}_{\mathbf{w}_{kl}}; \boldsymbol{\alpha}_{kl})$; hyperparameters $\tau, \tau_{0kl}$, for $l = 1, \cdots, L; k = 1, \cdots, K$
Require: update equations on different variables: $F_{\mathbf{Z}}; F_{\mathbf{w}_{kl}}; F_{\boldsymbol{\alpha}_{kl}}; F_{\boldsymbol{\tau}_0}; F_{\tau}$
Require: function to compute ELBO, $F_{\text{ELBO}}$
Ensure: ELBO increase
repeat
    $\mathbf{W} \leftarrow \sum_{l=1}^{L} \boldsymbol{\mu}_{\mathbf{w}} \circ \boldsymbol{\alpha}$.    ▷ Define $\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\alpha}$ as $(L, K, P)$ arrays by arranging $\boldsymbol{\mu}_{\mathbf{w}_{kl}}, \boldsymbol{\alpha}_{kl}$
    $\boldsymbol{\tau}_0 \leftarrow F_{\boldsymbol{\tau}_0}(\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}, \boldsymbol{\alpha})$
    for $k$ in $1, \cdots, K$ do
        $\mathbb{E}[\mathbf{R}_{kl}^{\intercal}\mathbf{Z}_k]^{(1)} = \mathbf{X}^{\intercal}\boldsymbol{\mu}_{\mathbf{z}_k} - \sum_{k' \neq k} \mathbb{E}[\mathbf{w}_{k'}]\mathbb{E}[\mathbf{Z}_{k'}^{\intercal}\mathbf{Z}_k]$    ▷ compute the first two terms in Eq
        for $l$ in $1, \cdots, L$ do
            $\mathbb{E}[\mathbf{w}_{kl'}] = \mathbf{w}_k - \boldsymbol{\mu}_{\mathbf{w}_{kl}} \circ \boldsymbol{\alpha}_{kl}$    ▷ removing the $l_{th}$ effect from $\mathbf{w}_k$
            $\mathbb{E}[\mathbf{R}_{kl}^{\intercal}\mathbf{Z}_k] = \mathbb{E}[\mathbf{R}_{kl}^{\intercal}\mathbf{Z}_k]^{(1)} - \mathbf{w}_k\mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k]$    ▷ complete the calculation of $\mathbb{E}[\mathbf{R}_{kl}^{\intercal}\mathbf{Z}_k]$
            $(\boldsymbol{\mu}_{\mathbf{w}_{kl}}, \boldsymbol{\Sigma}_{\mathbf{w}_{kl}}) \leftarrow F_{\mathbf{w}_{kl}}(\mathbb{E}[\mathbf{R}_{kl}^{\intercal}\mathbf{Z}_k], \mathbb{E}[\mathbf{Z}_k^{\intercal}\mathbf{Z}_k], \tau_{0kl}, \tau)$
            $\boldsymbol{\alpha}_{kl} \leftarrow F_{\boldsymbol{\alpha}_{kl}}(\mathbb{E}[\mathbf{R}_{kl}^{\intercal}\mathbf{Z}_k], \boldsymbol{\mu}_{\mathbf{w}_{kl}}, \boldsymbol{\Sigma}_{\mathbf{w}_{kl}})$
            $\mathbf{w}_k = \mathbb{E}[\mathbf{w}_{kl'}] + \mu_{\mathbf{w}_{kl}} \circ \boldsymbol{\alpha}_{kl}$    ▷ Update the $\mathbf{w}_k$
        end for
    end for
    $(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Sigma}_{\mathbf{Z}}) \leftarrow F_Z(\mathbf{X}, \tau, \mathbb{E}[\mathbf{W}])$
    $\tau = F_{\tau}(\mathbf{X}, \tau, \mathbb{E}[\mathbf{W}], \mathbb{E}[\mathbf{Z}])$
    $ELBO \leftarrow F_{\text{ELBO}}$
until convergence criterion satisfied
```

2 Supplementary Figures

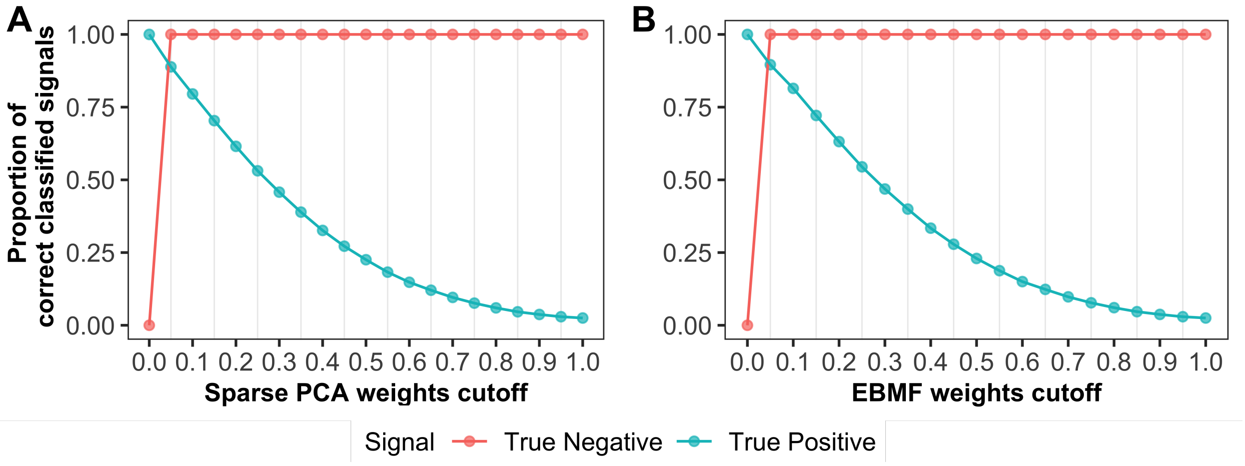

Fig. S1: Sensitivity decreases fast as cutoff value increase when choosing weights from sparse PCA and EBMF for variable selection

The proportion of correct classified signals using posterior weights from sparse PCA (A) and EBMF (B) as the cutoff. The green dots represent sensitivity, i.e. $Pr(\text{weights} \geq \text{cutoff} \,|\, \text{True positive signal})$, the red dots represent specificity, i.e. $Pr(\text{weights} < \text{cutoff} \,|\, \text{True false signal})$. For consistency and comparable between PIPs and weights, the weights are standardized to be ranged from 0 to 1.

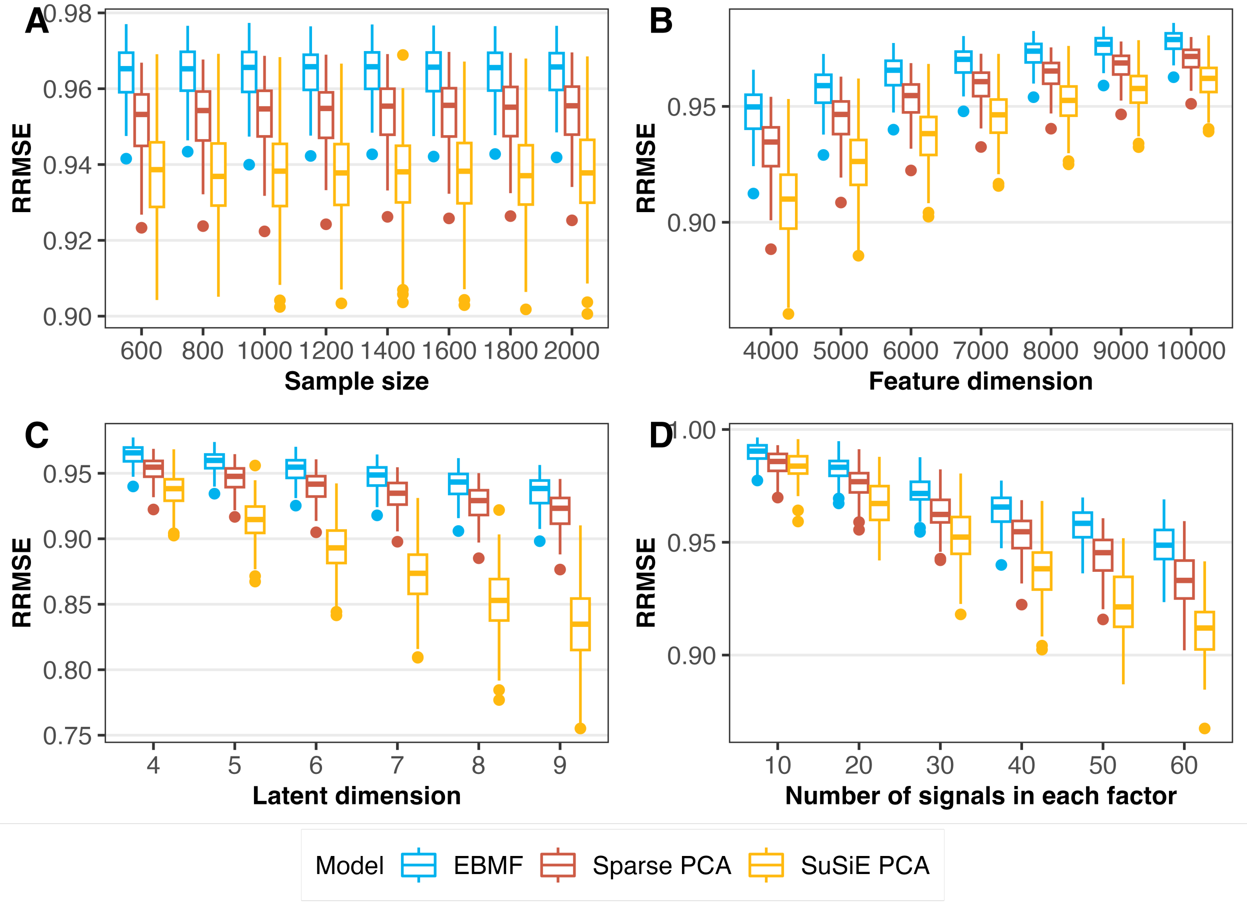

Fig. S2: SuSiE PCA has the lowest RRMSE across all simulations

The RRMSE is defined in **Simulation** equation (2.31), which is an assessment of the model prediction performance. The base simulation data is the same as the simulation setting in Figure 1. For each scenario in (A-D) we only vary one of the parameters at a time to generate the simulation data while fixing the other 3 parameters and then input the true parameters (N,P,K,L) into models. Finally, we compute the RRMSE based on equation (2.31) and plot them as a function of N, P, K, L.

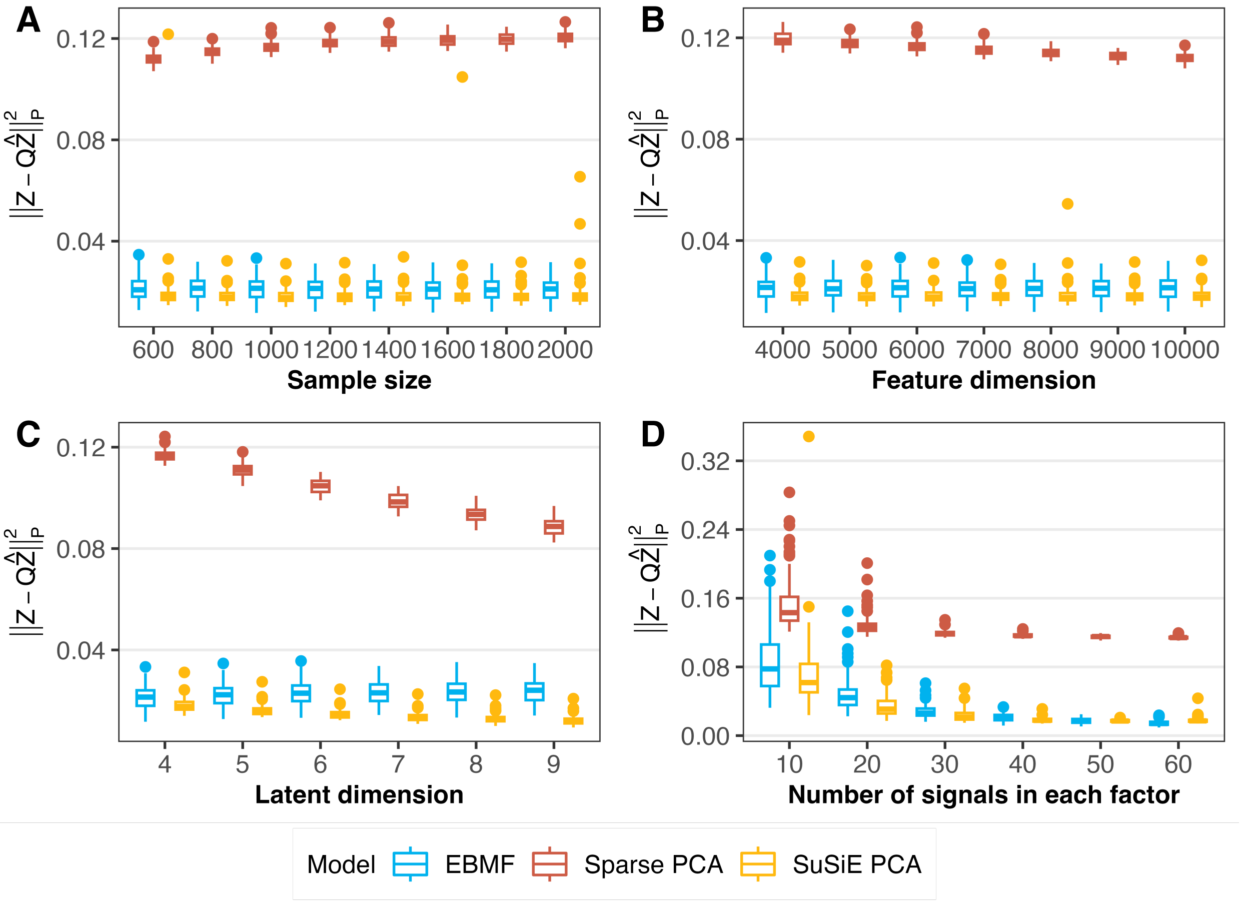

Fig. S3: SuSiE PCA and EBMF has lower Procrustes error of latent factor Z than sparse PCA across all simulations

The Procrustes error for latent factor Z is computed in the same manner of loading (Figure 1) using equation (2.30). For each scenario in (A-D) we only vary one of the parameters at a time to generate the simulation data while fixing the other 3 parameters and then input the true parameters (N,P,K,L) into models. Finally, we compute the Procrustes errors of **Z** based on equation (2.30) and plot them as a function of N, P, K, L.

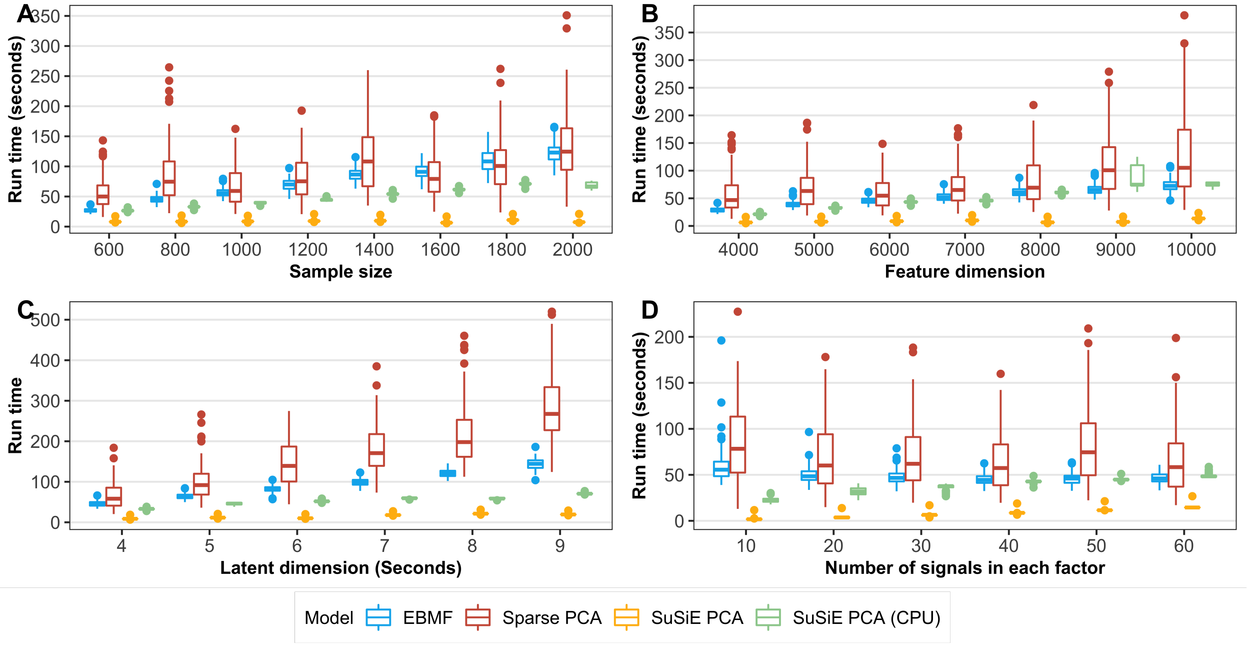

Fig. S4: SuSiE PCA remains the fastest method on either CPU or GPU than sparse PCA and EBMF across all simulations

All analyses are performed on high-performance computing center with the same CPU (AMD EPYC 7302 16-Core Processor) or GPU (Nvidia Tesla A40). Noticed that platform where we collect the runtime data in this figure is different from that in **Table 1** and therefore is not comparable.

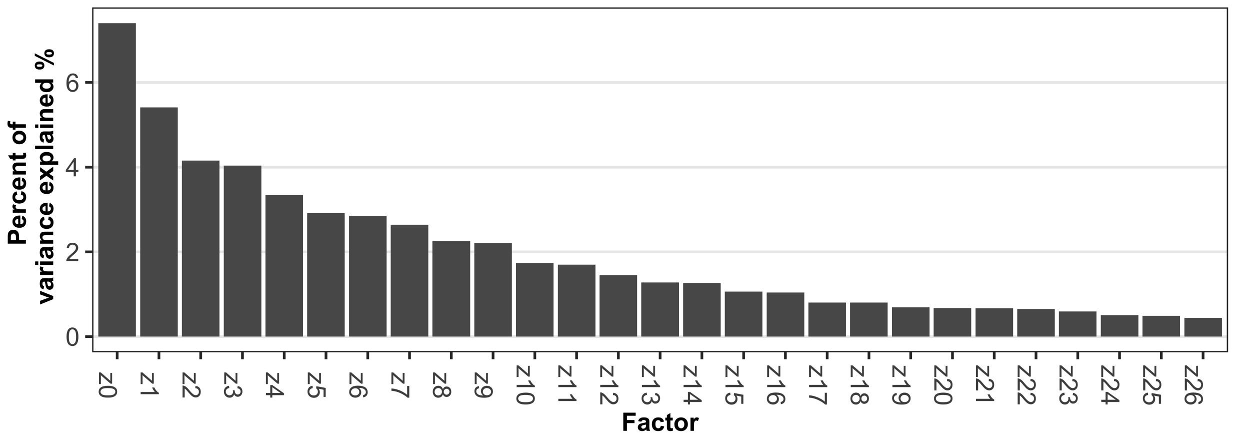

Fig. S5: Percent of variance (PVE) explained by 27 factors from SuSiE PCA in GTEx z score summary data.

PVE is a measurement of variance explained by the model and is computed based on equation (3.1)

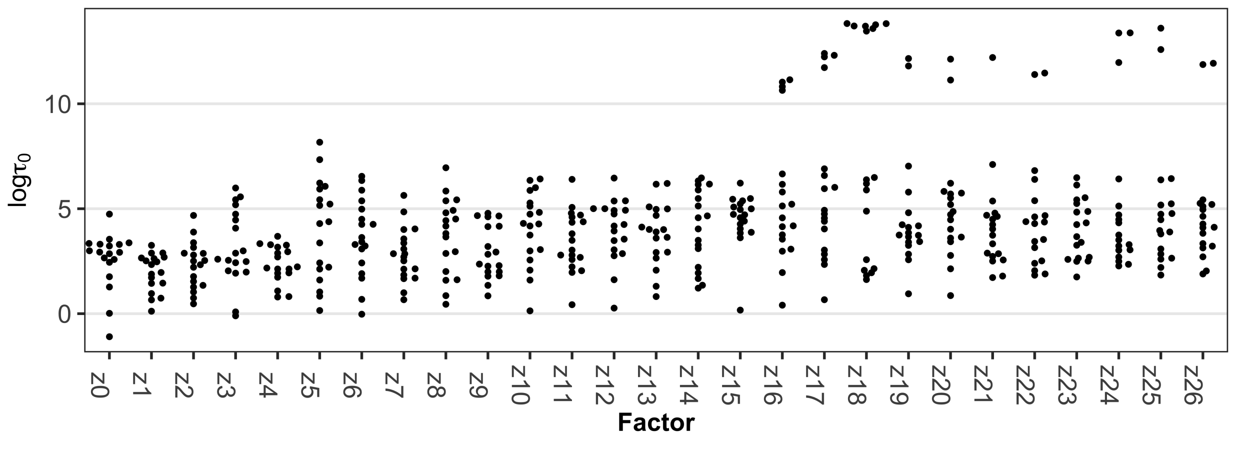

Fig. S6: The MLE of the precision parameter $\log \tau_{0kl}$ in SuSiE PCA will be extremely large for those over-specified single effects in GTEx Z score summary data.

The τ_{0kl} is the inverse variance of the random variable w_{kl}. When there are excessive number of single effects specified in the SuSiE PCA, the MLE of corresponded τ_{0kl} will become extremely large and as a result shrink those redundant single effects to 0.

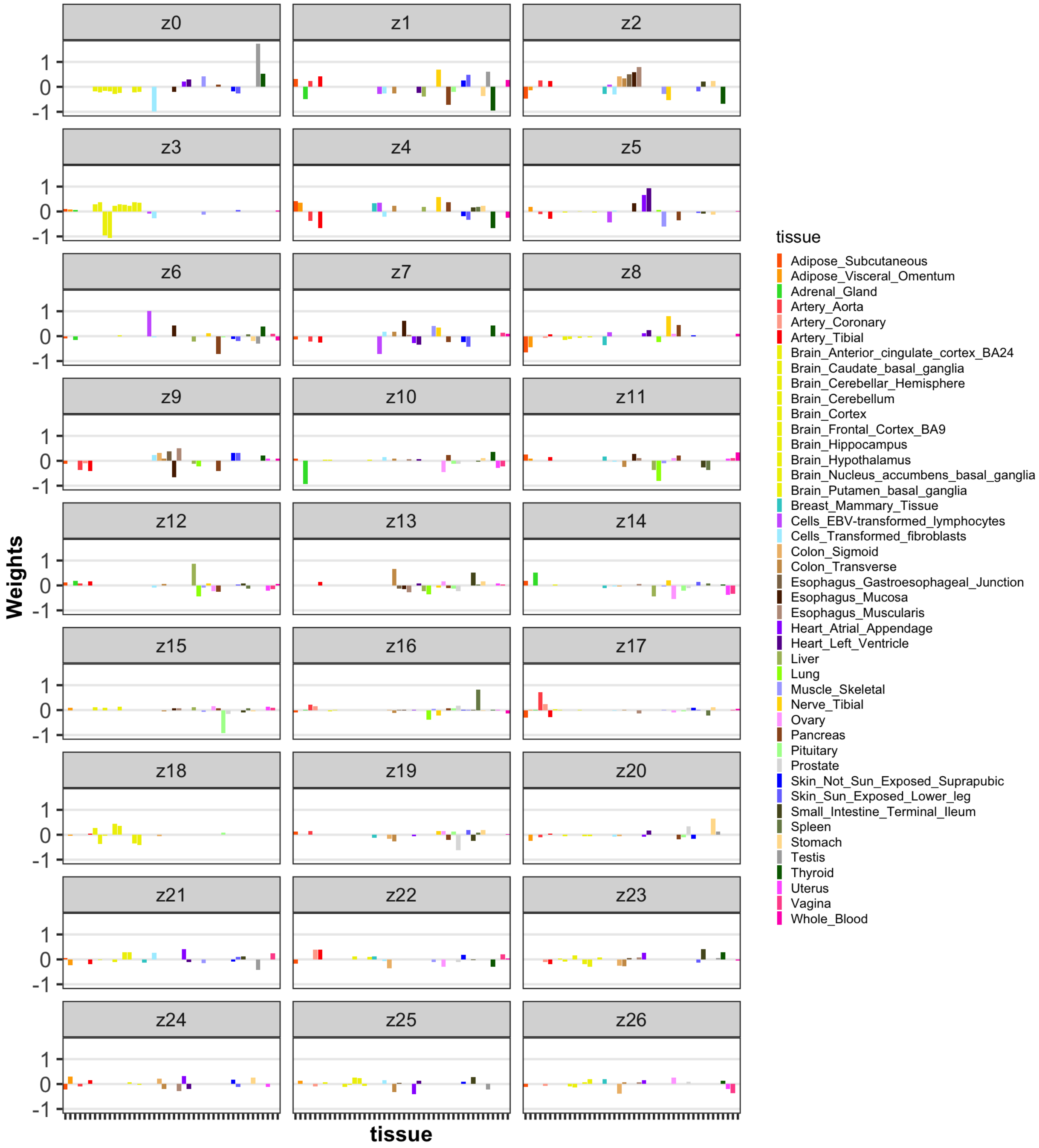

Fig. S7: Posterior weights by factors from SuSiE PCA across different tissues in GTEx Z score summary data.
The posterior weights (or loadings) refers to the strengthen of association of the tissues contribution to the factor. The L is set to be 18 which means each factor has at most 18 tissues with non-zero effects.

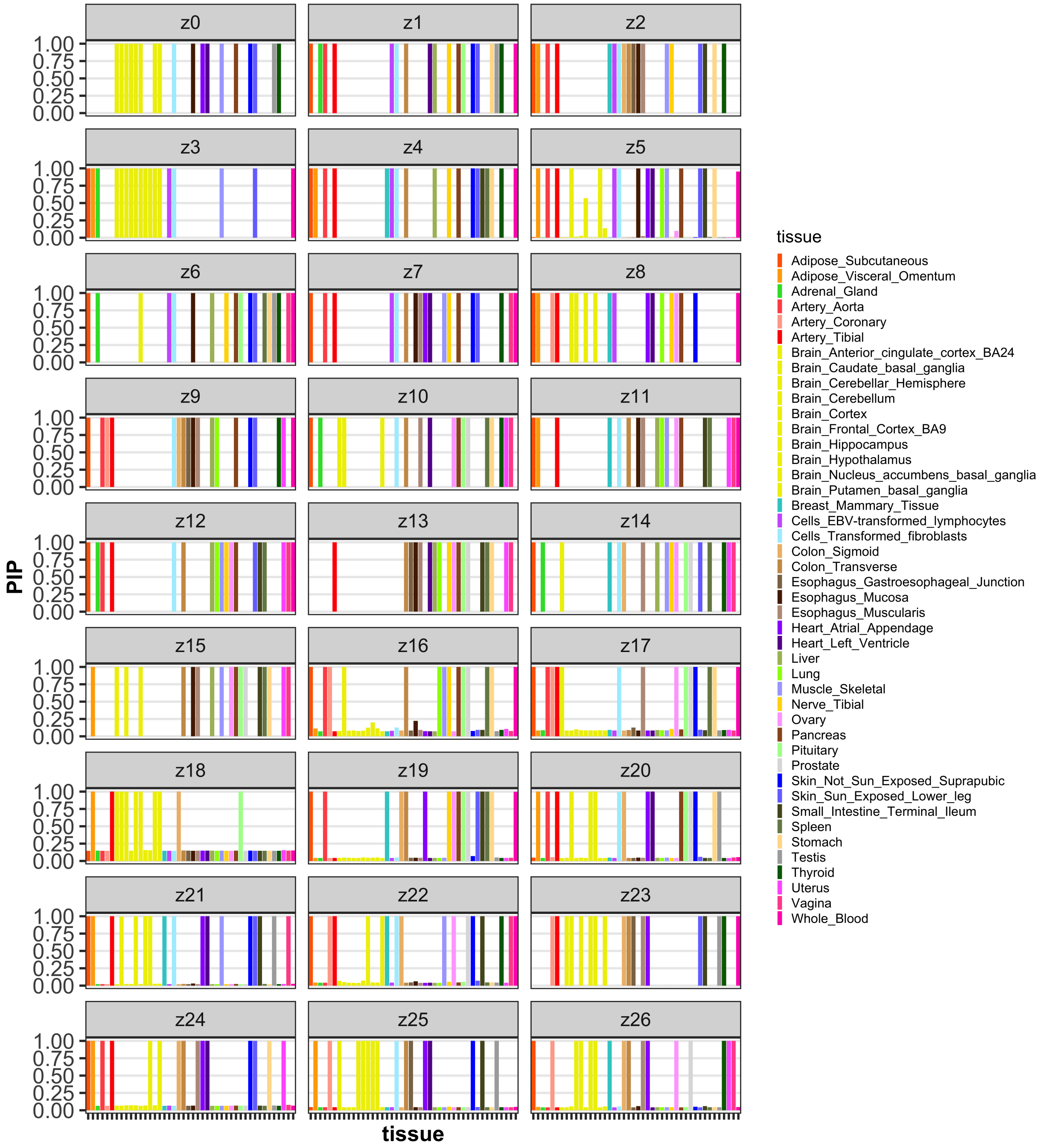

Fig. S8: Posterior inclusion probabilities (PIPs) by factors from SuSiE PCA across different tissues in GTEx Z score summary data.

Most of (PIPs) are exactly 1 across different tissue by factors, implying the model is quite confident in terms of the tissues contributing to each factor

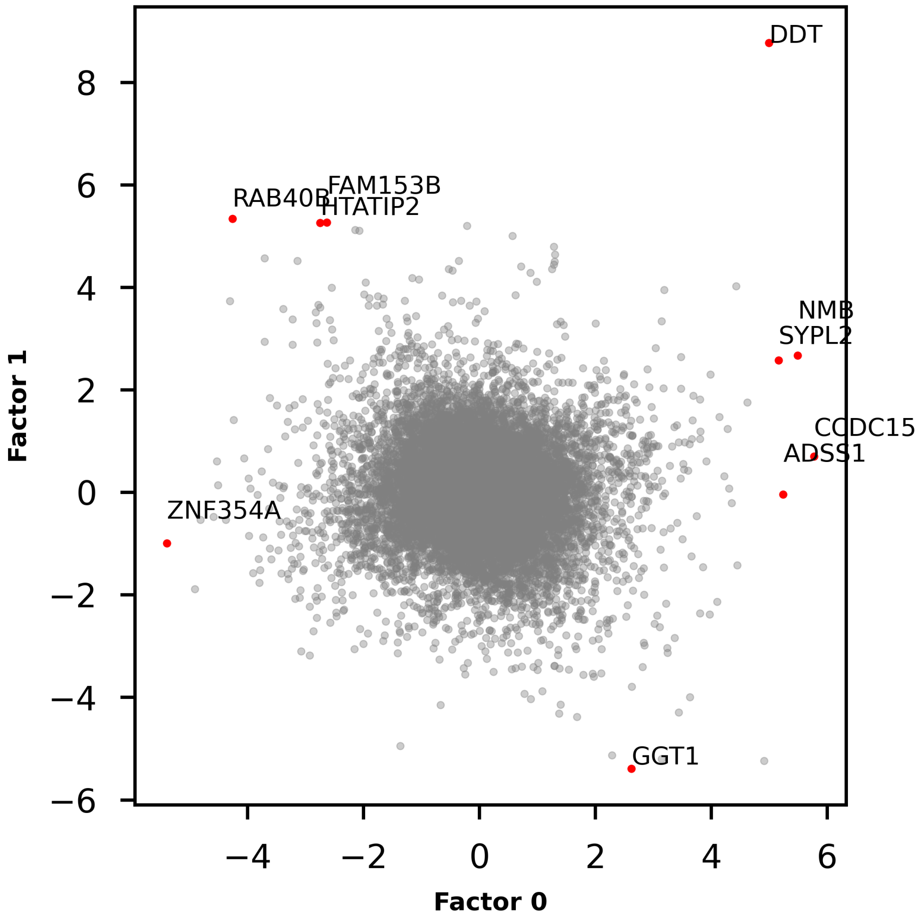

Fig. S9: Scatterplot of latent factor values of $\mathbf{z}_1$ vs. $\mathbf{z}_0$ in GTEx Z score summary data. Latent factor values refer to the posterior means of latent factor **Z**. Each point represents a specific gene, the genes with the top 5 absolute largest latent factor values are the red points with labels. The "outlier" gene DDT is found to be associated with testicular cancer.

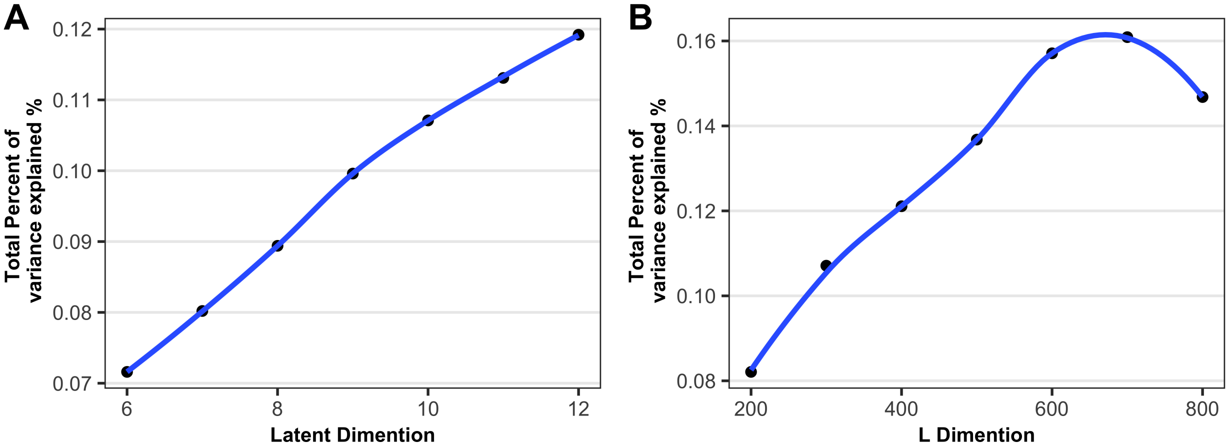

Fig. S10: Total percent of variance explained (PVE) as a function of the number of latent dimension K (A) and the number of single effects L (B) in SuSiE PCA

(A) We first fixed L = 300, and varied K from 6 to 12. The increased amount between two consecutive K in total PVE becomes smaller after K reaches 10. (B) We then fixed K=10 and varied L from 200 to 800. Although the total PVE reaches its maximum at L=700, we noticed that only the first three components have 600 of downstream genes with PIP $>$ 0.9, while the rest of the components only have 200-400 genes with PIP $>$ 0.9. We then compared the results between L=300 and L=700 and realized the smaller L retains the same top significant downstream genes relevant to the component. Considering the parsimony and interpretation of the model, we finally choose K=10 and L=300.

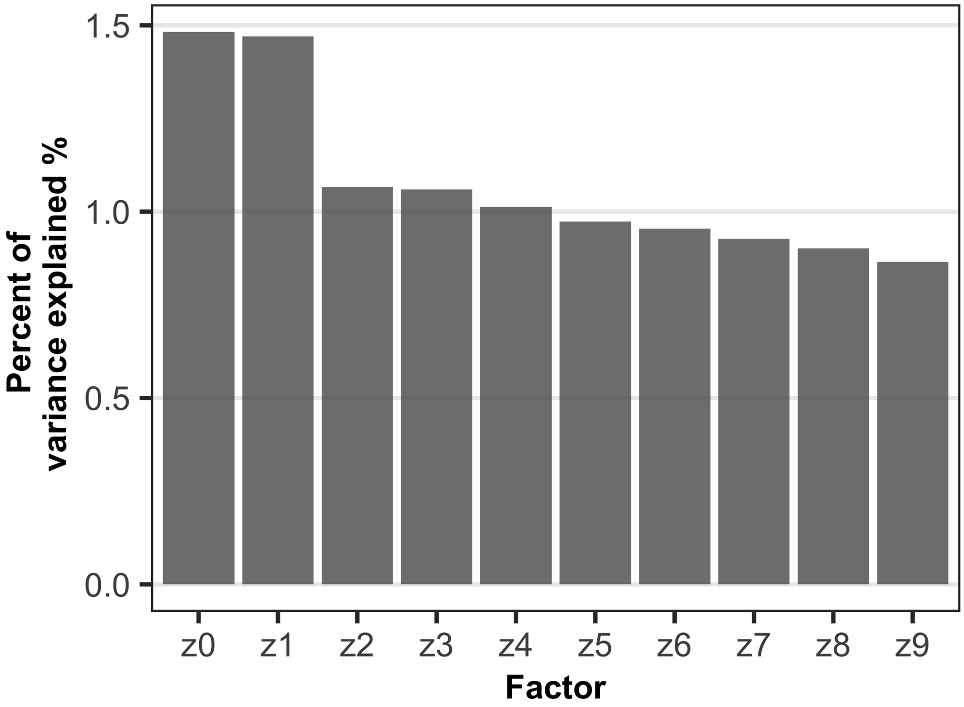

Fig. S11: Percent of variance (PVE) explained by 10 factors from SuSiE PCA in gene expression data from perturb-seq data.
PVE is a measurement of variance explained by the model and is computed based on equation (3.1)

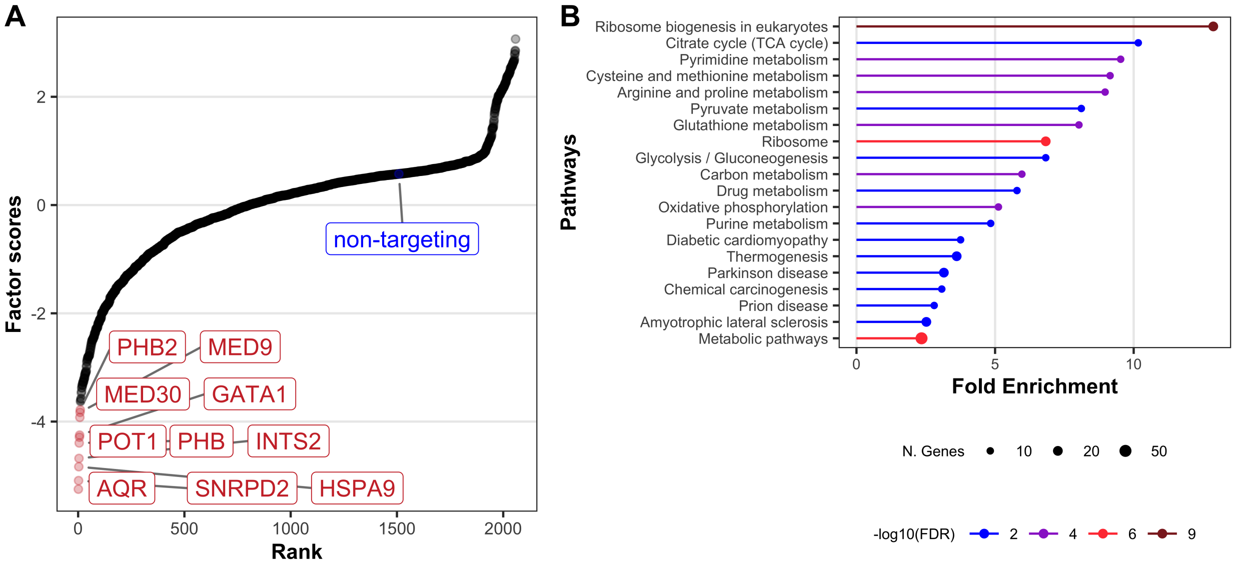

Fig. S12: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 1 in perturb-seq data

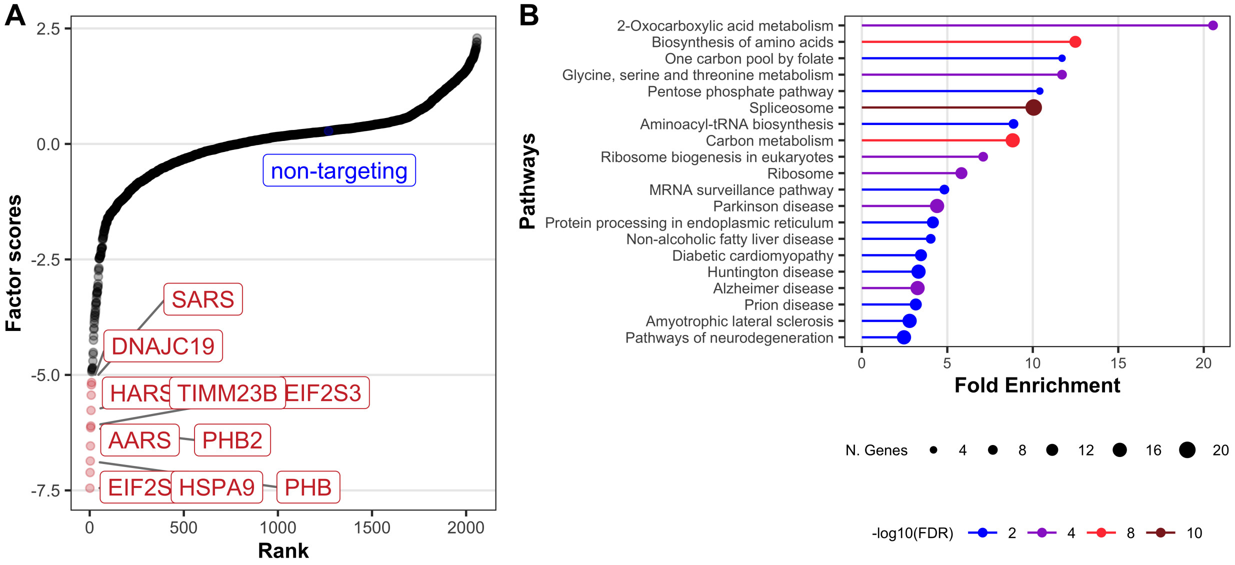

Fig. S13: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 2 in perturb-seq data

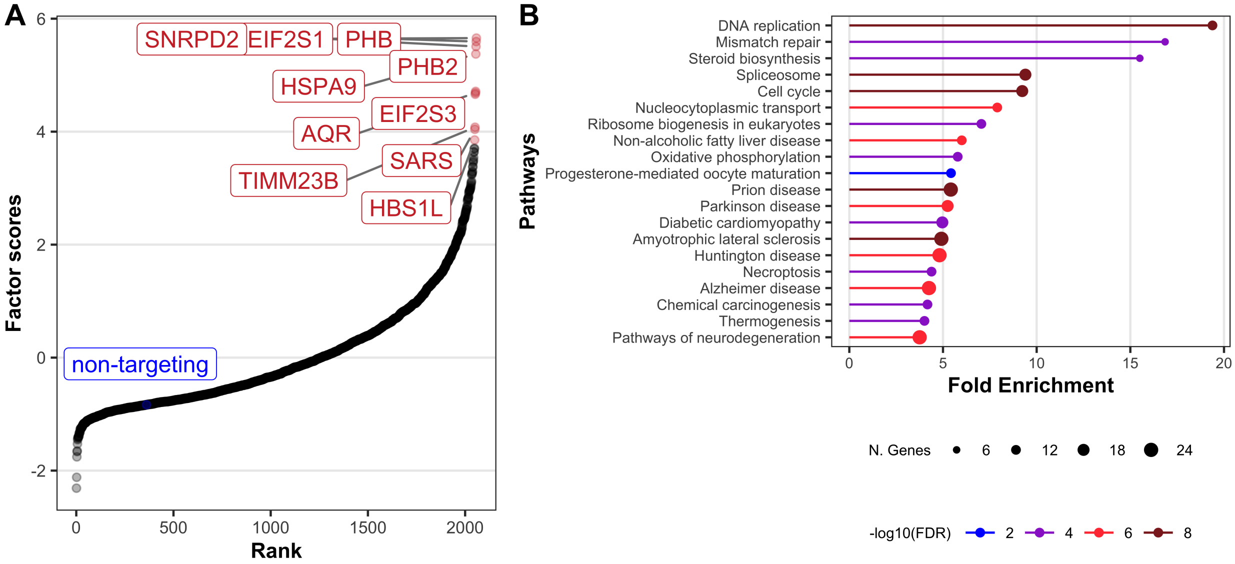

Fig. S14: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 3 in perturb-seq data

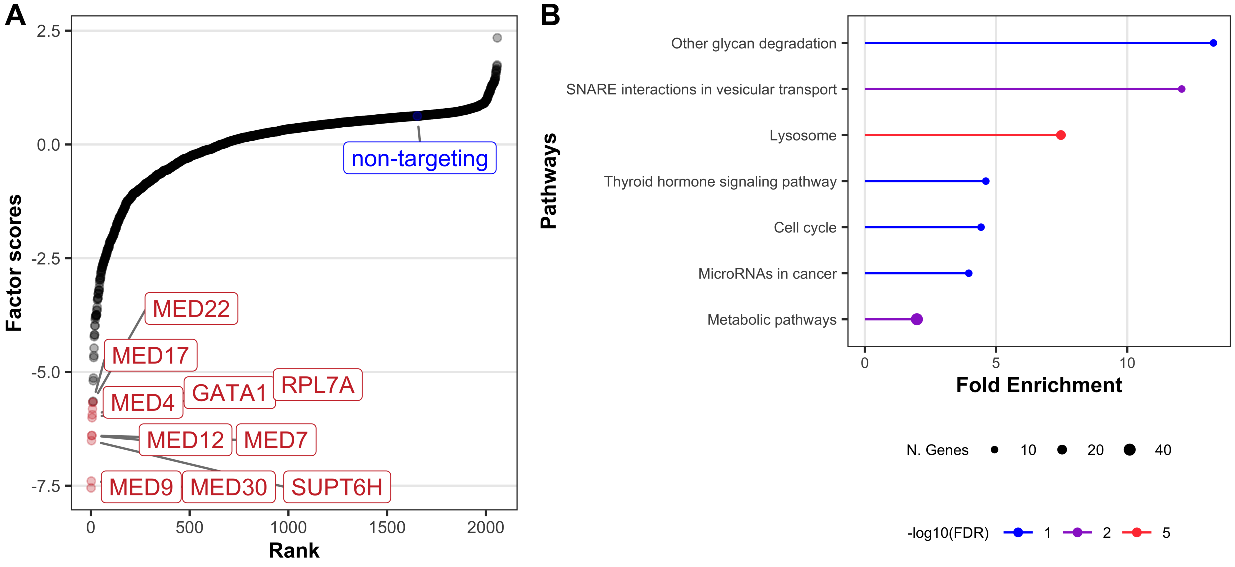

Fig. S15: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 4 in perturb-seq data

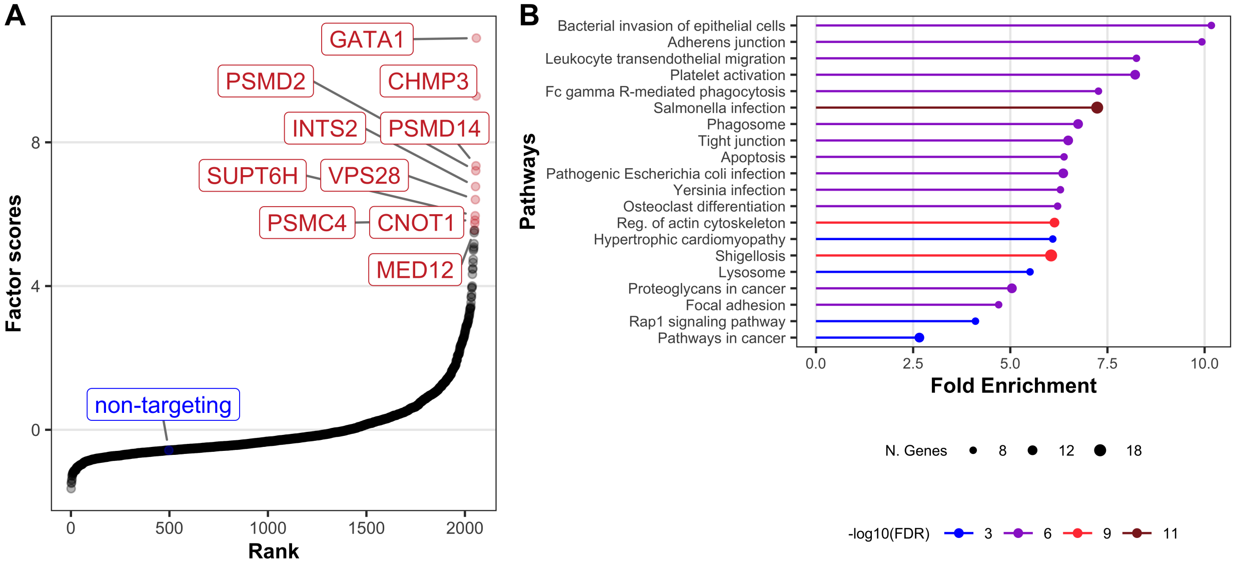

Fig. S16: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 5 in perturb-seq data

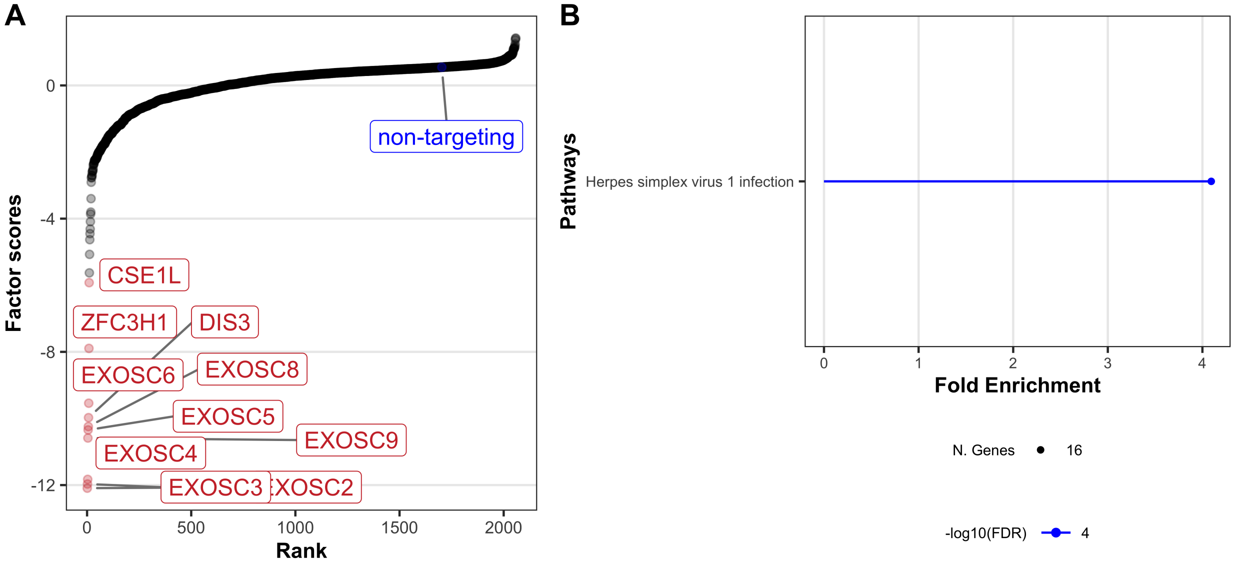

Fig. S17: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 6 in perturb-seq data

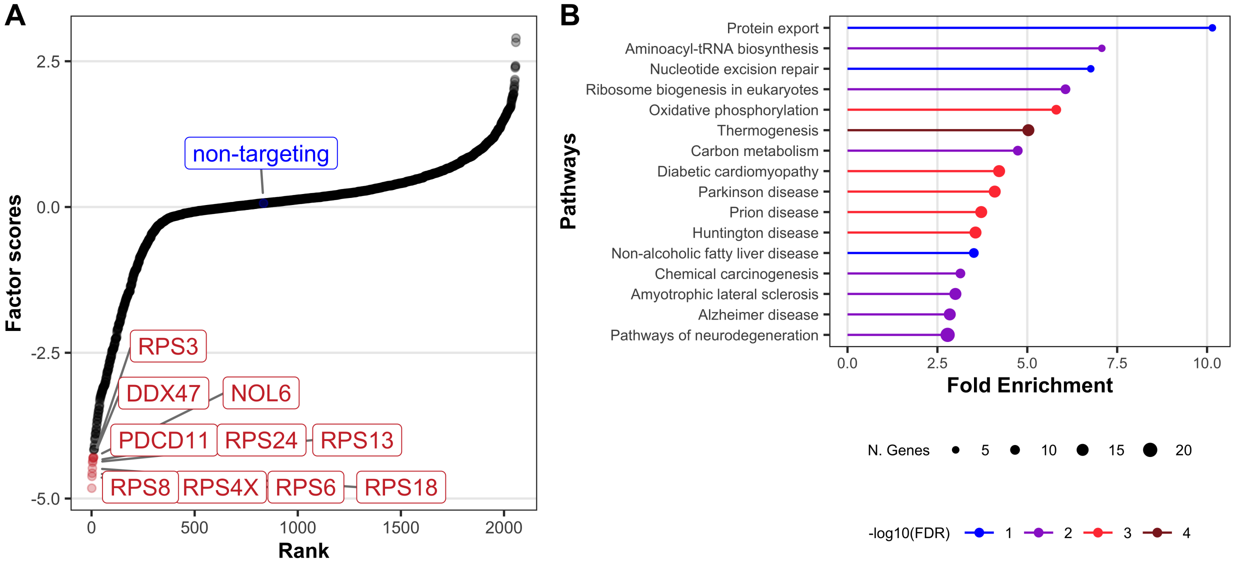

Fig. S18: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 7 in perturb-seq data

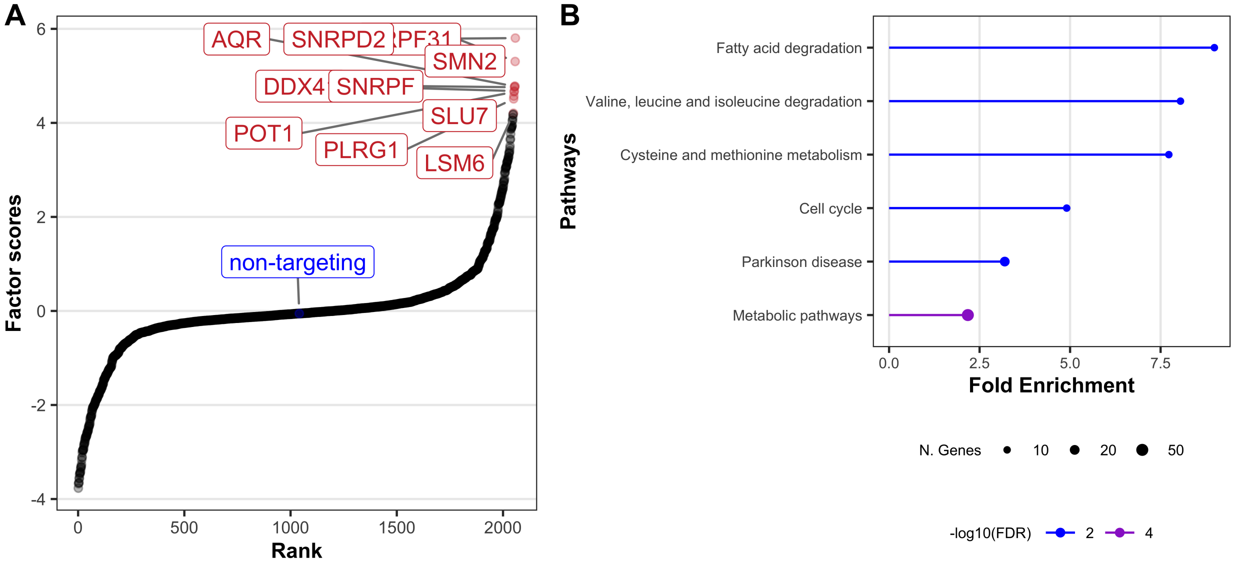

Fig. S19: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 8 in perturb-seq data

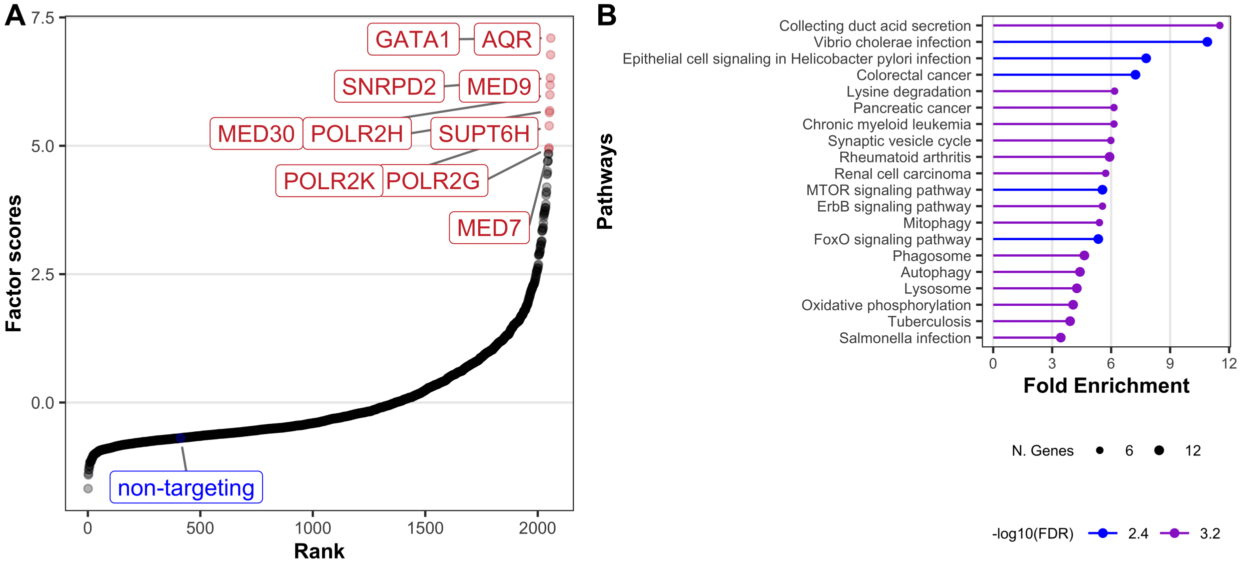

Fig. S20: factor scores (A) and top enriched pathways of gene set enrichment analysis of downstream gene (B) from SuSiE PCA factor 9 in perturb-seq data