A Sparse Singular Value Decomposition Method for Spiked-mean Model

Qing Sun

Penn State

December 2, 2018

Introduction

Basic Problem: Given a high-dimension observation matrix data X with noise:

$$X = Y + Z, \qquad Z \text{ is the noise}$$

Want to reconstruct or approximate the true structure Y from X.

Spiked Mean Model: Given a large, noisy matrix data X, assume the variables $X_{i,j}$ can be modeled as

$$X = UDV' + Z, \qquad Z_{i,j} \sim \text{ iid }, E(Z_{i,j}) = 0,\ \text{var}(Z_{i,j}) = \sigma^2$$

$U_{n\times r}$, $V_{p\times r}$ are sparse singular vectors, with $r \ll \min(n, p)$.

Example: two-way functional data, $Y_{i,j} = Y(s_i, t_j)$. As smooth function of (s, t), if expand Y in suitable basis, the coefficient should be sparse.

Challenge and a Solution

Difficulties:

1. X is high-dimension, the accumulation of the noise $Z_{i,j}$ results in poor estimate when apply classical SVD on X;
2. The computation involves many structureless cells $Z_{i,j}$, thus computation expensive.

A possible Solution:

Fast Iterative Thresholding-SparseSVD (FIT-SSVD):
Combine classical SVD with thresholding step in each iteration.

Methods

For the **Classical SVD Iteration**, given a right starting frame $V^{(0)}$, a $p \times r$ orthonormal columns, repeat

(1) Right-to-Left: $\tilde{U}^{(k)} = XV^{(k-1)}$
(2) Left QR: $U^{(k)} R_u^{(k)} = \tilde{U}^{(k)}$
(3) Left-to-Right: $\tilde{V}^{(k)} = X' U^{(k)}$
(4) Right QR: $V^{(k)} R_v^{(k)} = \tilde{V}^{(k)}$

FIT-SSVD:

(1) Right-to-Left: $\tilde{U}^{(k)} = XV^{(k-1)}$
(2) Left Thresholding: $\tilde{U}^{(k),thr} = \eta(\tilde{U}^{(k)}, \gamma_u^{(k)})$
(3) Left QR: $U^{(k)} R_u^{(k)} = \tilde{U}^{(k),thr}$
(4) Left-to-Right: $\tilde{V}^{(k)} = X' U^{(k)}$
(5) Right Thresholding: $\tilde{V}^{(k),thr} = \eta(\tilde{V}^{(k)}, \gamma_v^{(k)})$
(6) Right QR: $V^{(k)} R_v^{(k)} = \tilde{V}^{(k),thr}$

FIT-SSVD: threshold level

Choose suitable threshold level γ is tricky:

- If γ too small, only kick out few structureless elements, make little benifit.
- If γ too large, shave off too many elements, give results with high bias.

Recall $X = UDV' + Z$, Given the present estimate $V^{(k-1)}$

$$\tilde{U}^{(k)} = XV^{(k-1)} = UDV'V^{(k-1)} + ZV^{(k-1)}$$

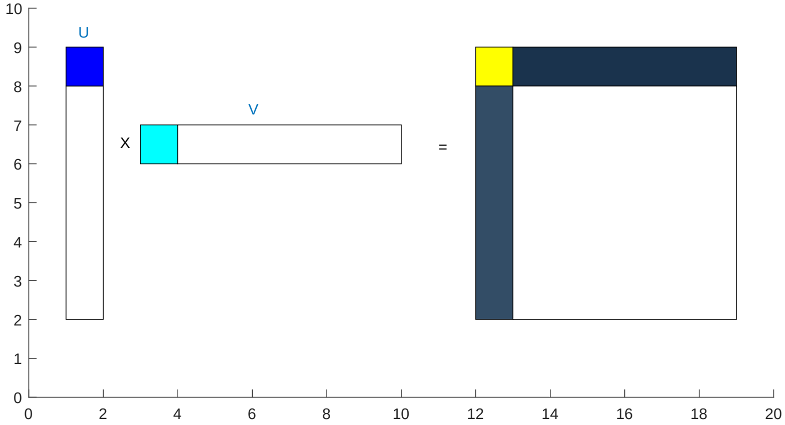
Theoretical threshold level:

$$\gamma = E(\|ZV^{(k-1)}\|_\infty)$$

The element in $\tilde{U}^{(k)}$ with absolute value less than γ can be regarded low signal since it is weaker than the expected noise level.

FIT-SSVD: threshold level

Question: $\gamma = E(\|ZV^{(k-1)}\|_\infty)$ requires the information of Z, we only has the observation data X.
Given the present estimates, $U^{(k-1)}, V^{(k-1)}$



Regard elements in white area as the samples from noise,
nonparametric bootstrap Z^*, let $\gamma = \text{median}\{\|Z_i^* V^{(k-1)}\|_\infty\}$

real data study

Lung Cancer data: $X = [\quad]_{12625 \times 56}$

- gene expression levels of 12,625 genes
- 56 cases (56 patients)
- 4 types of lung cancer

Only a part of the genes regulate the type of the cancer, thus the singular vector should be sparse.

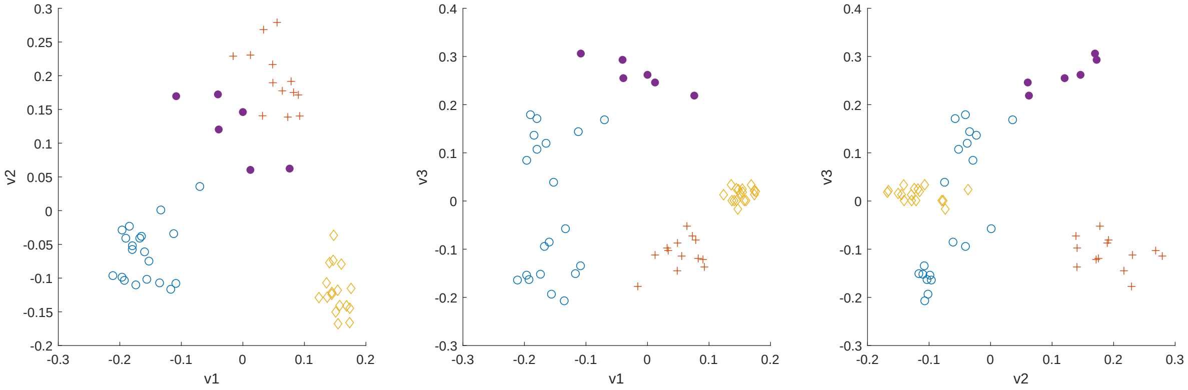


Figure 1: scatter plot of the right eigenvectors

real data study

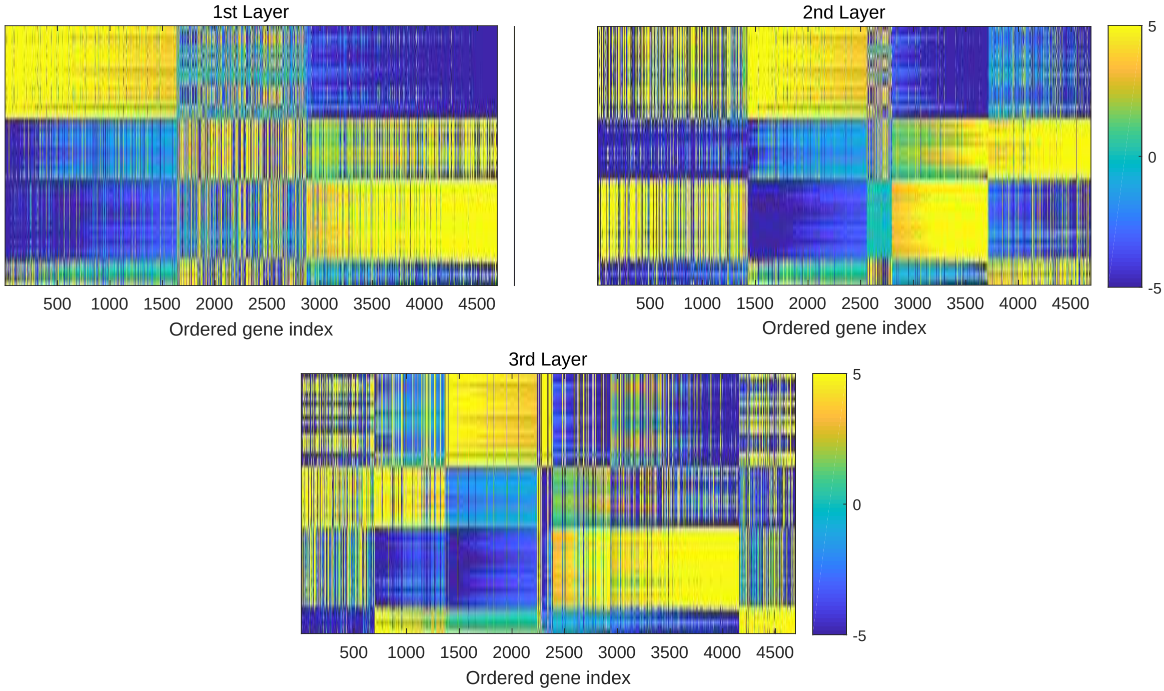


Figure 2: reconstructed rank-three approximation by FIT-SSVD

LSHM: more than 1 hour

FIT-SSVD: 3.7 seconds

Future work: In the real data study I use the hard threshold. In the literature, FIT-SSVD can use many other threshold precedure, e.g. soft threshold. is there an optimal one? produce the good estimate with fastest speed.

Reference

[1] Dan Yang, Zongming Ma, and Andreas Buja, A Spares Singular Value Decomposition Method for High-Dimensional Data, *Journal of Computational and Graphical Statistics* **23**(2014), 923–942.
[2] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron. Biclustering via sparse singular value decomposition, *Biometrics*, **66**(2010),1087-–1095.