

# Joint Covariance Estimation

Jiaxin Hu

February 16, 2023

## 1 Model

*Model:*

Consider  $n$  independent  $p$ -dimensional multivariate normal variables

$$Y_i \sim \mathcal{N}_p(\mathbf{0}, \Sigma_0 + \Sigma_k + \sigma^2 \mathbf{I}), \quad i \in V_k, k \in [K] \quad (1)$$

where

- $V_k \subset [n]$  with size  $|V_k| = n_k$  is the set of samples in the group  $k$ ;  $V_a \cap V_b = \emptyset$  for all  $a \neq b$  and  $\sum_k n_k = n$ ;
- $\Sigma_0 = \mathbf{U}_0 \Lambda_0 \mathbf{U}_0^T$  with  $\mathbf{U}_0 \in \mathbb{R}^{p \times r_0}$  and  $r_0$ -dimensional diagonal matrix  $\Lambda_0$ ;  $\Sigma_0$  is the low-rank covariance matrix shared across all clusters.
- $\Sigma_k = \mathbf{U}_k \Lambda_k \mathbf{U}_k^T$  for all  $k \in [K]$  with  $\mathbf{U}_k \in \mathbb{R}^{p \times r_k}$  and  $r_k$ -dimensional diagonal matrix  $\Lambda_k$ ;  $\Sigma_k$  is the low-rank covariance matrix specified for the  $k$ -th group.
- The low-rank singular spaces of shared and specified covariance matrices are orthogonal with each other; i.e.,  $\mathbf{U}_0 \perp \mathbf{U}_k$  or  $\mathbf{U}_0^T \mathbf{U}_k = \mathbf{0}$  for all  $k \in [K]$ .
- The intersection of all spaces  $C(\mathbf{U}_k)$  has dimension 0; i.e.,  $\dim(\cap_{k \in [K]} C(\mathbf{U}_k)) = 0$ .

*Goals:*

1. The primary goal is to estimate the shared low-dimensional space  $\mathbf{U}_0$  and group-specified spaces  $\mathbf{U}_k, k \in [K]$ .
2. The secondary goal is to estimate the whole covariance matrices  $\Sigma_0$  and  $\Sigma_k$ .

*Motivating application:*

1. (Gene covariance network) Suppose we have expression data of  $p$  genes for  $n$  samples. The  $n$  samples are separated into  $K$  groups based on individual features (e.g. sex, age, genotype). The low-rank matrix  $\Sigma_0 + \Sigma_k$  is interpreted as the low-rank gene covariance network for the

$k$ -th group, where  $\Sigma_0$  is the shared low-rank structure across the groups and  $\Sigma_k$ 's are the group-specified structures.

In application, the dimension  $p$  ( $\sim 10k+$ ) is far larger than the sample size  $n$  ( $< 1000$ ). Hence, we encourage the low-rankness,  $r < n$ , in the covariance matrix. Otherwise, it is impossible to estimate the covariance structure, since the number of parameters  $p^2$  is far larger than the sample size  $np$ .

The estimation problem for  $\mathbf{U}_0$  and  $\mathbf{U}_k$  can be considered as an extended version of PCA. Let  $k = 1$ . Then the model (1) reduces to the spike model for regular PCA. Further, with estimated  $\mathbf{U}_0$  and  $\mathbf{U}_k$ , we can find the shared and group-specified community structures among  $p$  genes.

Introducing the common+specified structure in covariance will benefit the gene covariance estimation and different covariance comparison, compared with the separate estimation. The intuition is that we use more information to estimate the shared structure. Thus the estimation error from  $\Sigma_0$  will decrease.

## 2 Preliminary algorithm and simulation

*Algorithm:*

We first consider the joint estimation of the shared space  $\mathbf{U}_0$ . To illustrate the underlying rationale, we explain the algorithm procedures with oracle true matrices  $\mathbf{M}_k = \Sigma_0 + \Sigma_k, k \in [K]$ . The matrices  $\mathbf{M}_k$  can be re-written in a low-rank form

$$\mathbf{M}_k = [\mathbf{U}_0, \mathbf{U}_k] \begin{bmatrix} \Lambda_0 & \mathbf{0} \\ \mathbf{0} & \Lambda_k \end{bmatrix} \begin{bmatrix} \mathbf{U}_0^T \\ \mathbf{U}_k^T \end{bmatrix} = \mathbf{V}_k \mathbf{D}_k \mathbf{V}_k^T.$$

Here  $\mathbf{V}_k \in \mathbb{R}^{p \times (r_0 + r_k)}$  still has orthogonal columns due to the orthogonality between  $\mathbf{U}_0$  and  $\mathbf{U}_k$ . The task is to identify the common space  $\mathbf{U}_0$  from given spaces  $\{\mathbf{V}_k\}$ .

Since the common eigenvector may not correspond to the leading eigenvalues in a specific group, i.e., no assumption (except bounded away from 0) on the magnitudes of the diagonal elements of  $\Lambda_0, \Lambda_k$ , we can not use the gap in eigenvalues to distinguish  $\mathbf{U}_0$  from  $\mathbf{V}_k$ .

Instead, notice that

$$\mathbf{V}_k \mathbf{V}_k^T = \sum_{a=1}^{r_0} \mathbf{u}_a^{(0)} \mathbf{u}_a^{(0),T} + \sum_{b=1}^{r_k} \mathbf{u}_b^{(k)} \mathbf{u}_b^{(k),T},$$

where  $\mathbf{u}_a^{(0)}$ 's are orthogonal basis of the space  $C(\mathbf{U}_0)$ ,  $\mathbf{u}_b^{(k)}$ 's are orthogonal basis of the space  $C(\mathbf{U}_k)$ . Let  $R = \dim(\cup_{k \in [K]} C(\mathbf{U}_k))$  denote the dimension all the union space of all  $\mathbf{U}_k$ , and let  $\{\tilde{\mathbf{u}}_j\}_{j \in [R]} = \{\mathbf{u}_b^{(k)}\}_{b \in [r_k], k \in [K]}$  denote the basis of the union  $\cup_{k \in [K]} C(\mathbf{U}_k)$ . We have

$$\mathbf{W} = \sum_{k \in [K]} \mathbf{V}_k \mathbf{V}_k^T = K \sum_{a=1}^{r_0} \mathbf{u}_a^{(0)} \mathbf{u}_a^{(0),T} + \sum_{b=1}^R c_b \tilde{\mathbf{u}}_b \tilde{\mathbf{u}}_b^T.$$

The key observation is that the constants  $c_b$ 's are smaller than  $K$ , because  $\tilde{\mathbf{u}}_b$  occurs at most in  $K-1$  matrices  $\mathbf{V}_k \mathbf{V}_k^T$ 's under the constraint  $\dim(\cap_{k \in [K]} C(\mathbf{U}_k)) = 0$ . Hence,  $C(\mathbf{U}_0) = \text{span}\{\mathbf{u}_1^{(0)}, \dots, \mathbf{u}_{r_0}^{(0)}\}$  is identifiable by finding  $r_0$  leading eigenvectors of  $\mathbf{W}$ .

In practice, we do not have  $\mathbf{M}_k$ . Alternatively, we estimate the initial space  $\mathbf{V}_k$  by decomposing the sample covariance matrix. Let  $\mathbf{Y}_{V_k} \in \mathbb{R}^{n_k \times p}$  denote the observation matrix of the  $k$ -th group. The sample covariance matrix is  $\hat{\mathbf{M}}_k = \mathbf{Y}_{V_k}^T \mathbf{Y}_{V_k} / (n_k - 1)$ , where

$$\mathbb{E}[\hat{\mathbf{M}}_k] = \Sigma_0 + \Sigma_k + \sigma^2 \mathbf{I}, \quad \text{and} \quad \text{SVD}_{r_0+r_k}(\mathbb{E}[\hat{\mathbf{M}}_k]) = \text{SVD}_{r_0+r_k}(\mathbf{M}_k) = \mathbf{V}_k.$$

Hence, when the smallest eigenvalue in  $\mathbf{M}_k$  is enough far away from 0, we can have a good estimation of  $\mathbf{V}_k$  by decomposing  $\hat{\mathbf{M}}_k$ . Full algorithm procedures are in Algorithm 1.

---

**Algorithm 1** Joint estimation of  $\mathbf{U}_0$

---

**Input:** Observations  $Y_i, i \in [n]$ , ranks  $r_0, r_k$ , group identifications  $V_k$ .

- 1: Let  $\mathbf{Y}_{V_k} \in \mathbb{R}^{n_k \times p}$  denote the observation matrix of the  $k$ -th group. For all  $k \in [K]$ , compute the initial singular spaces

$$\hat{\mathbf{V}}_k = \text{SVD}_{r_0+r_k}(\mathbf{Y}_{V_k}^T \mathbf{Y}_{V_k} / (n_k - 1)).$$

- 2: Find the estimate

$$\hat{\mathbf{U}}_0 = \text{SVD}_{r_0}(\mathbf{W}), \quad \text{where} \quad \mathbf{W} = \sum_{k \in [K]} \hat{\mathbf{V}}_k \hat{\mathbf{V}}_k^T.$$

**Output:** Estimated shared space  $\hat{\mathbf{U}}_0$ .

---

Next, with jointly estimated  $\mathbf{U}_0$ , we find the estimate of  $\mathbf{U}_k$ . The intuition is simple:

$$(\mathbf{I} - \mathbf{U}_0 \mathbf{U}_0^T) \mathbf{V}_k = [\mathbf{0}, \mathbf{U}_k].$$

The matrix  $(\mathbf{I} - \mathbf{U}_0 \mathbf{U}_0^T)$  is the projection matrix onto the complement space  $C^\perp(\mathbf{U}_0)$ , and thus only the matrix  $\mathbf{U}_k \in C^\perp(\mathbf{U}_0)$  is remained after the projection of  $\mathbf{V}_k$ .

In practice, we only have estimates  $\hat{\mathbf{U}}_0$  and  $\hat{\mathbf{V}}_k$ , and thus we may not have pure 0 columns. Instead, we distinguish the columns corresponding to  $\mathbf{U}_k$  by sorting the  $\ell_2$  norms of the columns in  $(\mathbf{I} - \hat{\mathbf{U}}_0 \hat{\mathbf{U}}_0^T) \hat{\mathbf{V}}_k$ . Intuitively, with good estimated  $\hat{\mathbf{U}}_0, \hat{\mathbf{V}}_k$ , the columns corresponding to  $\mathbf{U}_k$  have norms close to 1 while others are close to 0. Full algorithm procedures are in Algorithm 2.

---

**Algorithm 2** Estimation of  $\mathbf{U}_k$

---

**Input:** Estimated  $r_0 + r_k$  singular space  $\hat{\mathbf{V}}_k$ , estimated shared space  $\hat{\mathbf{U}}_0$

- 1: Let  $\mathbf{v}_a, a \in [r_0 + r_k]$  denote the columns of  $\hat{\mathbf{V}}_k$ . Compute the norms

$$m_a = \|(\mathbf{I} - \hat{\mathbf{U}}_0 \hat{\mathbf{U}}_0^T) \mathbf{v}_a\|_2, \quad a \in [r_0 + r_k].$$

- 2: Find the estimate  $\hat{\mathbf{U}}_k = [\mathbf{v}_{b_1}, \dots, \mathbf{v}_{b_{r_0}}]$ , where  $\{b_1, \dots, b_{r_0}\}$  is the set of indices with  $r_0$  largest  $m_a$ 's.

**Output:** Estimated shared space  $\hat{\mathbf{U}}_k$ .

---

*Simulation:*

To evaluate the estimation, we consider the metric

$$\|\hat{\mathbf{U}} \hat{\mathbf{U}}^T - \mathbf{U} \mathbf{U}^T\|_F,$$

where  $\hat{\mathbf{U}}, \mathbf{U}$  are estimated and true orthogonal matrices. This metric is invariant to the orthogonal transformations and the permutation of columns.

The first experiment aims to verify the efficiency of Algorithm 1. We consider the setup with number of groups  $K = 4$ , ranks  $r_0 = r_k = 5$  for all  $k \in [K]$ ,  $\Lambda_0$  and  $\Lambda_k$ 's are identity matrix, and  $\sigma^2 = 0.05$ . We vary the sample size  $n_1 = \dots = n_4 \in [30, 500]$  and dimension  $p \in \{50, 100, 200, 500\}$ . Figure 1 indicates the estimation error decreases with larger sample size and increases with larger dimension, which agrees with our intuition on the hardness of the problem.

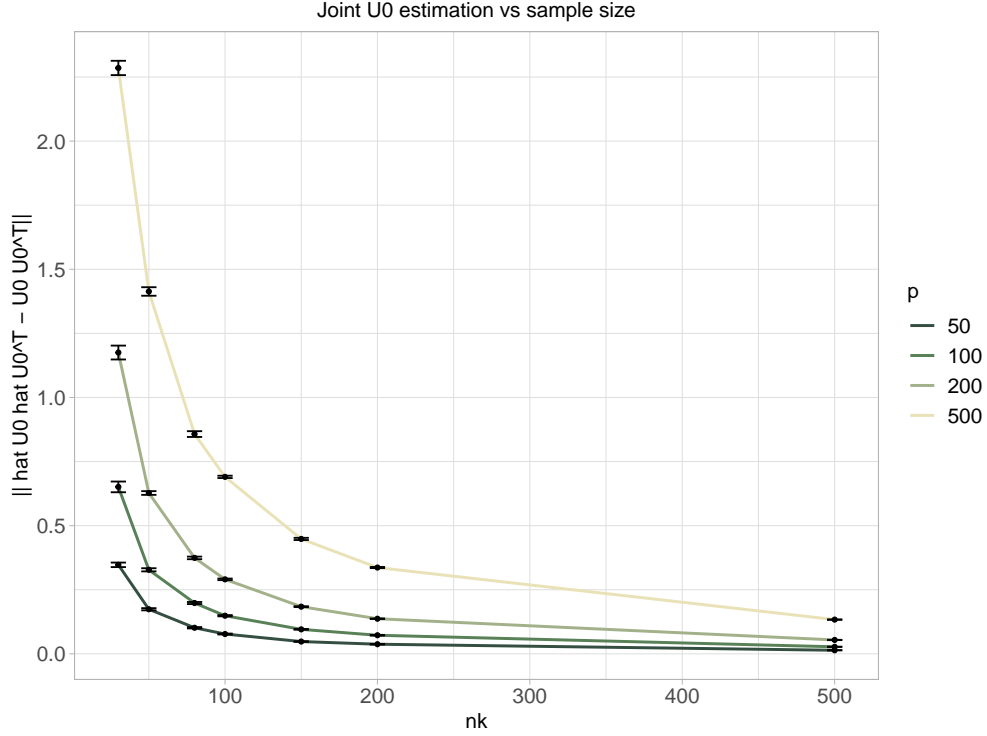


Figure 1: Estimation error  $\|\hat{\mathbf{U}}_0 \hat{\mathbf{U}}_0^T - \mathbf{U}_0 \mathbf{U}_0^T\|_F$  versus sample size  $n_1 = \dots = n_4 \in [30, 500]$  under different dimension  $p \in \{50, 100, 200, 500\}$ . The average error and standard deviation over 20 replications are reported.

### 3 Preliminary theorems

### References