Joint Covariance Estimation (Preliminary Theorems)

Jiaxin Hu

Feb 21, 2023

Consider n independent p-dimensional multivariate normal variables

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

where $\{V_k\}$ is a non-overlapped partition of $[n]$ with $|V_k| = n_k$, $\Sigma_0 = \boldsymbol{U}_0 \Lambda_0 \boldsymbol{U}_0^T$ is the common low-rank r_0 covariance component, $\Sigma_k = \boldsymbol{U}_k \Lambda_k \boldsymbol{U}_k^T$'s are group-specific low-rank r_k covariance components.

Our goal is to estimate the $\boldsymbol{U}_0, \boldsymbol{U}_k$ with given partition V_k's.

For simplicity, we use $r_k, \boldsymbol{U}_k, \Sigma_k$ for the sets $\{r_k\}_{k\in[K]}, \{\boldsymbol{U}_k\}_{k\in[K]}, \{\Sigma_k\}_{k\in[K]}$ when there is no notational confusion. Let $C(\cdot)$ denote the matrix column space, $\lambda_r(\cdot)$ denote the r-th largest eigenvalue.

1 Identifiability

Consider the parameter space

$$\begin{aligned}\mathcal{P}(r_0, r_k) = \Big\{ (\Sigma_0, \Sigma_k) \in \mathbb{S}_p(r_0) \times \mathbb{S}_p(r_k) : \boldsymbol{U}_0 \in \mathbb{O}_{p,r_0},\ \boldsymbol{U}_k \in \mathbb{O}_{p,r_k}, \boldsymbol{U}_0^T \boldsymbol{U}_k = \mathbf{0},\ k \in [K], \\ \dim(\cap_{k\in[K]} C(\boldsymbol{U}_k)) = 0, \min\{\operatorname{diag}(\Lambda_0), \operatorname{diag}(\Lambda_k)\} > 0 \Big\},\end{aligned} \tag{1}$$

where $\mathbb{S}_p(r)$ refers to the collection of all rank r symmetric matrices with dimension p and $\mathbb{O}_{n,m}$ refers to the collection of all n-by-m matrices with orthonormal columns.

Theorem 1.1 (Identifiability)**.** *For any parameters $(\Sigma_0, \Sigma_k) \in \mathcal{P}(r_0, r_k)$ with $r_0, r_k \geq 1, \max_k 2(r_0 + r_k) < p, K \geq 2$, and $\sigma > 0$, the low-rank factors $\boldsymbol{U}_0, \boldsymbol{U}_k$'s are unique up to orthogonal transformation, and Λ_0, Λ_k, Σ_0, Σ_k are unique.*

Proof of Theorem 1.1. Consider two sets of parameters $(\Sigma_0, \Sigma_k), (\Sigma_0', \Sigma_k') \in \mathcal{P}(r_0, r_k)$, $\sigma, \sigma' > 0$, and the corresponding low-rank factors. Suppose that two sets of parameters lead to the same set of covariance matrices; i.e.,

$$\Sigma_0 + \Sigma_k + \sigma^2 \boldsymbol{I} = \Sigma_0' + \Sigma_k' + (\sigma')^2 \boldsymbol{I}, \quad k \in [K] \tag{2}$$

We firstly show the identifiability of σ^2. Since $\max_k 2(r_0+r_k) < p$, there exists $\boldsymbol{W} \in \mathbb{O}_{p,p-2\max_k(r_0+r_k)}$ such that $\boldsymbol{U}_0^T\boldsymbol{W} = \boldsymbol{U}_k^T\boldsymbol{W} = (\boldsymbol{U}_0')^T\boldsymbol{W} = (\boldsymbol{U}_k')^T\boldsymbol{W} = \boldsymbol{0}$. Right multiply $\boldsymbol{W}$ on both sides of equation (2), we have

$$\sigma^2\boldsymbol{W} = (\sigma')^2\boldsymbol{W},$$

which indicates $\sigma^2 = (\sigma')^2$ and

$$\Sigma_0 + \Sigma_k = \Sigma_0' + \Sigma_k' \quad k \in [K]. \tag{3}$$

Next, we show the identifiability of $\boldsymbol{U}_0$. Right multiply $\boldsymbol{U}_0$ on both sides of equation (3). We have

$$\boldsymbol{U}_0\Lambda_0 = \boldsymbol{U}_0'\Lambda_0'(\boldsymbol{U}_0')^T\boldsymbol{U}_0 + \boldsymbol{U}_k'\Lambda_k'(\boldsymbol{U}_k')^T\boldsymbol{U}_0. \tag{4}$$

For right hand side, let $C_0' = C(\boldsymbol{U}_0'\Lambda_0'(\boldsymbol{U}_0')^T\boldsymbol{U}_0) \subset C(\boldsymbol{U}_0')$ and $C_k' = C(\boldsymbol{U}_k'\Lambda_k'(\boldsymbol{U}_k')^T\boldsymbol{U}_0) \subset C(\boldsymbol{U}_k')$ for all $k \in [K]$. For the left hand side, we have $C(\boldsymbol{U}_0\Lambda_0) = C(\boldsymbol{U}_0)$ due to the boundedness of Λ_0. Noticed that $C_0' \perp C_k'$, we have

$$C_k' = C(\boldsymbol{U}_0)/C_0' \quad \text{for all } k \in [K].$$

Then, by the intersection constraint in parameter space (1), we have $\dim(\cap_k C_k') \leq \dim(\cap_k C(\boldsymbol{U}_k')) = 0$ and thus $C(\boldsymbol{U}_0)/C_0' = 0$. Therefore, we have $C(\boldsymbol{U}_0) = C(\boldsymbol{U}_0')$ and $\boldsymbol{U}_0$ is unique up to orthogonal transformation. Let $\boldsymbol{U}_0' = \boldsymbol{U}_0\boldsymbol{O}$ for some $\boldsymbol{O} \in \mathbb{O}_{r_0}$. Left multiply $\boldsymbol{U}_0$ on both sides of (4). We obtain

$$\Lambda_0 = \boldsymbol{O}\Lambda_0'\boldsymbol{O}^T,$$

which indicates $\Lambda_0 = \Lambda_0', \Sigma_0 = \Sigma_0'$, and $\boldsymbol{O}$ is identity matrix when diagonal elements in Λ_0 are distinct.

Last, given $\Sigma_0 = \Sigma_0'$, we have $\Sigma_k = \Sigma_k' \in \mathbb{S}_{r_k}$. Hence, by eigendecomposition, we have $\Lambda_k = \Lambda_k'$ and $\boldsymbol{U}_k = \boldsymbol{U}_k'\boldsymbol{O}_k$ for some $\boldsymbol{O}_k \in \mathbb{O}_{r_k}$ for all $k \in [K]$. $\square$

2 Guarantee for Algorithm 1

We consider the $\sin\Theta$ metric to evaluate the distance between the rank r estimated factor $\hat{\boldsymbol{U}}$ and true factor $\boldsymbol{U}$; i.e.,

$$\sin\Theta(\hat{\boldsymbol{U}}, \boldsymbol{U}) = \text{diag}(\sqrt{1-\sigma_1^2}, \ldots, \sqrt{1-\sigma_r^2}),$$

where σ_i is the i-th singular value of $\hat{\boldsymbol{U}}^T\boldsymbol{U}$.

Theorem 2.1 (Guarantee for Algorithm 1)**.** *Consider the true parameters $(\Sigma_0, \Sigma_k) \in \mathcal{P}(r_0, r_k)$ with $r_0, r_k \geq 1, 2(r_0 + r_k) < p, K \geq 2$ and noise σ^2. Let $\hat{\boldsymbol{U}}_0$ denote the output of Algorithm 1. Assume $n_k \asymp n/K$ for all $k \in [K]$. As $n \to \infty$, with high probability, we have*

$$\|\sin\Theta(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0)\|_{op} \lesssim \sqrt{\frac{p}{n}} \frac{\sigma^2}{\min_k \min\{\lambda_{r_0}(\Lambda_0), \lambda_{r_k}(\Lambda_k)\}}.$$

Proof sketch of Theorem 2.1. The main idea to prove Theorem 2.1 is to use Davis-Kahan Theorem twice.

Firstly, let $\boldsymbol{M}_k = \frac{1}{n_k-1}\boldsymbol{Y}_{V_k}^T\boldsymbol{Y}_{V_k}$ denote the sample covariance matrices, where $\boldsymbol{Y}_{V_k} \in \mathbb{R}^{n_k \times p}$ is the observation matrix of the k-th group, for all $k \in [K]$. Let $\hat{\boldsymbol{V}}_k = \text{SVD}_{r_0+r_k}(\boldsymbol{M}_k)$ and $\boldsymbol{V}_k = \text{SVD}_{r_0+r_k}(\Sigma_0 + \Sigma_k)$ is the true factor. By Davis-Kahan Theorem, for high probability, we have

$$\|\sin\Theta(\hat{\boldsymbol{V}}_k, \boldsymbol{V}_k)\|_{op} = \|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T - \boldsymbol{V}_k\boldsymbol{V}_k^T\|_{op} \lesssim \frac{\|\boldsymbol{M}_k - (\Sigma_0 + \Sigma_k - \sigma^2\boldsymbol{I})\|_{op}}{\min\{\lambda_{r_0}(\Lambda_0), \lambda_{r_k}(\Lambda_k)\}}.$$

Particularly, by Gaussian covariance estimation, with high probability, we have

$$\|\boldsymbol{M}_k - (\Sigma_0 + \Sigma_k - \sigma^2\boldsymbol{I})\|_{op} \lesssim \sqrt{\frac{p}{n}}\sigma^2. \tag{5}$$

Secondly, let $\boldsymbol{E}_k = \hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T - \boldsymbol{V}_k\boldsymbol{V}_k^T$ denote the symmetric differential matrices. Then, we have $\sum_{k\in[K]}\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T = \sum_{k\in[K]}\boldsymbol{V}_k\boldsymbol{V}_k^T + \sum_{k\in[K]}\boldsymbol{E}_k$, where the estimate $\hat{\boldsymbol{U}}_0 = \text{SVD}_{r_0}(\sum_{k\in[K]}\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T)$ and the true parameter $\boldsymbol{U}_0 = \text{SVD}_{r_0}(\sum_{k\in[K]}\boldsymbol{V}_k\boldsymbol{V}_k^T)$. We apply the Davis-Kahan Theorem for $\sum_{k\in[K]}\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T$. With high probability, we have

$$\|\sin\Theta(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0)\|_{op} \lesssim \frac{\|\sum_{k\in[K]}\boldsymbol{E}_k\|_{op}}{\lambda_{r_0}(\sum_{k\in[K]}\boldsymbol{V}_k\boldsymbol{V}_k^T)} \leq \frac{\sum_{k\in[K]}\|\hat{\boldsymbol{V}}_k\hat{\boldsymbol{V}}_k^T - \boldsymbol{V}_k\boldsymbol{V}_k^T\|_{op}}{K}.$$

Combining the inequality (5), we obtain

$$\|\sin\Theta(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0)\|_{op} \lesssim \sqrt{\frac{p}{n}}\frac{\sigma^2}{\min_k \min\{\lambda_{r_0}(\Lambda_0), \lambda_{r_k}(\Lambda_k)\}}.$$

□

3 Extensions

Heteroskedastic joint estimation

Inspired by Heteroskedastic PCA Zhang et al. (2018), we consider the heteroskedastic version of joint covariance estimation with model

$$Y_i \sim \mathcal{N}_p(\boldsymbol{0}, \Sigma_0 + \Sigma_k + \boldsymbol{D}), \quad \boldsymbol{D} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2).$$

Unlike homogeneous case, the unequal variance in $\boldsymbol{D}$ would affect the estimation of $\boldsymbol{U}_0, \boldsymbol{U}_k$. We would replace the regular SVD step in Algorithm 1 by the the diagonal deletion SVD (or so-called HeteroPCA in Zhang et al. (2018)). The traditional Davis-Kahan Theorem is non-applicable in this case, neither the concentration inequality for the sample covariance matrix. Instead, we may apply the variants in Zhang et al. (2018)). The conjectured guarantee of Hetero-Algorithm 1 is

$$\|\sin\Theta(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_0)\|_{op} \lesssim \sqrt{\frac{\tilde{p}}{n}}\frac{\sigma_{\max}^2}{\min_k \min\{\lambda_{r_0}(\Lambda_0), \lambda_{r_k}(\Lambda_k)\}},$$

where $\tilde{p} = \sigma_{\text{sum}}^2/\sigma_{\max}^2$ is the "effective dimension" for heteroskedastic PCA.

Joint precision matrix estimation

Inspired by the GLasso project, we may assume the low-rank structure on precision matrix, rather than on covariance matrix; i.e.,

$$Y_i \sim \mathcal{N}_p(\mathbf{0}, \Omega_k), \quad \Omega_k^{-1} = \Sigma_0 + \Sigma_k + \sigma^2 \boldsymbol{I},$$

where Σ_0, Σ_k are low-ranked and $\sigma^2 \boldsymbol{I}$ is involved to ensure the positive-definiteness of precision matrix. Then, in Algorithm 1, we apply the decomposition on the estimated precision matrices $\hat{\Omega}_k^{-1} = [\boldsymbol{Y}_{V_k}^T \boldsymbol{Y}_{V_k}/(n_k - 1)]^{-1}$ and consider the estimation error $\|\hat{\Omega}_k^{-1} - \Omega_k^{-1}\|$.

Unlike GLasso project, we assume low-rankness on the factors Σ_0, Σ_k while GLasso model assumes ℓ_0 sparsity on Σ_0, Σ_k. Also, previous GLasso project does not provide any efficient algorithm to estimate the precision matrices but only consider the MLE analysis.

Clustering via covariance/precision matrix

Previous GLasso project assumes the group partition V_k's are unknown while current covariance project is established with given V_k. The EM algorithm is the only way in my mind at this point to estimate V_k practically. However, EM algorithm is computationally expensive, especially for the case with $p > n$.

References

Zhang, A. R., Cai, T. T., and Wu, Y. (2018). Heteroskedastic pca: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*.