MIXED MEMBERSHIP NETWORK WITH AUTOREGRESSIVE STRUCTURE

Tianyi Sun[1], Bo Zhang[1*], Baisuo Jin[2,1], and Yuehua Wu[3]

[1] *University of Science and Technology of China,* [2] *Xinjiang Normal University and* [3] *York University*

Supplementary Material

The supplementary material provides additional information on the proposed model, including the convergence rate of the connection probability estimator, an example using normalized Laplace matrix, a discussion on constant values in EETE, and an analysis of node heterogeneity. Furthermore, the supplementary material includes additional real data analysis and detailed proofs of all propositions and theorems.

A.1 Remark of the mixed membership autoregressive network model

A.1.1 Convergence rate of the connection probability estimator

In Subsection 2.1 of our paper, we give the estimator of the connection probability $\alpha_{i,j}$ and $\beta_{i,j}$ in Jiang et al. (2023), which is:

$$\hat{\alpha}_{i,j} = \frac{\sum_{t=1}^{n} X_{i,j}^{t}(1 - X_{i,j}^{t-1})}{\sum_{t=1}^{n}(1 - X_{i,j}^{t-1})}, \quad \hat{\beta}_{i,j} = \frac{\sum_{t=1}^{n}(1 - X_{i,j}^{t})X_{i,j}^{t-1}}{\sum_{t=1}^{n} X_{i,j}^{t-1}}. \tag{A.1}$$

We now present the uniform convergence rate of this estimator. The following proposition is the same as Proposition 4 in Jiang et al. (2023).

Proposition A.1.1. *If Assumptions 1 - 2 hold, then for any constant $c > 2$, there exists a sufficiently large constant $C > 0$ such that*

$$P\left(\max_{1\le i<j\le p} |\widehat{\alpha}_{i,j} - \alpha_{i,j}| \ge l^{-1} C\sqrt{\log p/n}\right) \le 2p^2 \exp\{-c \log p\},$$
$$P\left(\max_{1\le i<j\le p} |\widehat{\beta}_{i,j} - \beta_{i,j}| \ge l^{-1} C\sqrt{\log p/n}\right) \le 2p^2 \exp\{-c \log p\}.$$

Consequently, as $n, p \to \infty$, we have

$$\max_{1\le i<j\le p} |\widehat{\alpha}_{i,j} - \alpha_{i,j}| = O_p\left(\sqrt{\log p/n}\right) \quad \textit{and} \quad \max_{1\le i<j\le p} |\widehat{\beta}_{i,j} - \beta_{i,j}| = O_p\left(\sqrt{\log p/n}\right).$$

The proof of Proposition A.1.1 can be found in Jiang et al. (2023), and hence omitted.

A.1.2 Example of using the normalized Laplacian matrix

In Subsection 2.2, we find that it is not appropriate to use the normalized Laplacian matrices $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ to obtain the community structure. Here, we set up two networks as examples to support this argument.

For both networks, we set $p = 9$ and $K = 3$. The first network contains only pure nodes, where communities 1, 2, and 3 contain 1, 1, and 7

pure nodes, respectively. The second network contains 6 pure nodes and 3 mixed nodes, where communities 1, 2, and 3 contain 1, 1, and 4 pure nodes, respectively, and the membership vectors of the mixed nodes are $(0.1, 0.2, 0.7)$, $(0.3, 0.5, 0.2)$, and $(0.8, 0.1, 0.1)$, respectively. At the same time, we set the diagonal elements of $\boldsymbol{B}_1$ to 0.8 and all other elements to 0.1.

Then, we consider the latent community structure of $\boldsymbol{L}_1$. Assuming that the eigen-decomposition of $\boldsymbol{L}_1$ is $\boldsymbol{U}\Lambda\boldsymbol{U}^\top$, it is obvious that if $\boldsymbol{L}_1$ satisfies $\boldsymbol{L}_1 = \boldsymbol{\Theta}\boldsymbol{\Psi}\boldsymbol{\Theta}^\top$ for the matrix $\boldsymbol{\Psi}$, then $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{\Theta}^\top(\boldsymbol{\Theta}^\top\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}$. We calculate the maximum singular value of $\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{\Theta}^\top(\boldsymbol{\Theta}^\top\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}$ for both networks.

For the second network, the maximum singular value is about 0.0659. Compared with 5×10^{-16} of the first network, this result cannot be ignored because the first network, though does not contain mixed nodes, is also community-unbalanced. Such results indicate that the normalized Laplacian matrix is not applicable to the mixed membership case.

A.1.3 Discussion on constant values in EETE

As mentioned in Remark 4, we adopt $\xi = 0.02$ and $r = 0.5$ in our analysis of EETE. Here, we provide some explanation for this setting. Note that

our estimator is applied to the eigenvalues of $\boldsymbol{P}^*$, where

$$\boldsymbol{P}^* = \boldsymbol{P}_1^* \boldsymbol{P}_1^* + \boldsymbol{P}_2^* \boldsymbol{P}_2^*. \tag{A.2}$$

In a similar estimator from Zhang et al. (2021), the constant coefficient (analogous to ξ in our estimator) is set to 0.1. However, their estimator operates on the eigenvalues of the adjacency matrix, which corresponds to P_1^* and P_2^* in our work. Given the form of $\boldsymbol{P}^*$, $\xi = (0.1)^2 + (0.1)^2 = 0.02$ may be a more appropriate setting. Therefore, we set $\xi = 0.02$ and find that it performs well in both simulations and real data analysis.

For the constant r, $0 < r < 1$ needs to be satisfied to ensure the theoretical consistency of the estimator. We set $r = 0.5$ and observe good performance in both simulations and real data analysis.

Finally, we present the eigenvalues of $\boldsymbol{P}^*$ and $\widehat{\boldsymbol{P}}^*$ in a simulated network to provide more details and show the rationality of the estimator. The number of communities K is set to 3, and the number of nodes p to 200. For $i = 1, \ldots, K$, we set $\zeta_{i,i} = \eta_{i,i} = 0.4$, and for $1 \leq i < j \leq K$, we let $\zeta_{i,j}$ and $\eta_{i,j}$ be independently drawn from $U[0.05, 0.25]$.We set 50 pure nodes for each community and the remaining 50 nodes as mixed nodes, with the membership of each mixed node randomly set to $(x_i, x_i, 1 - 2x_i), (x_i, 1 - 2x_i, x_i)$, or $(1 - 2x_i, x_i, x_i)$, with a probability of 1/3. Here, for node i, x_i is a random value uniformly drawn from the interval [0,0.5]. The eigenvalues

Table A.1: The six largest eigenvalues of $\boldsymbol{P}^*$ and $\widehat{\boldsymbol{P}}^*$in a simulated network.

Eigenvalue	1	2	3	4	5	6
$\boldsymbol{P}^*$	1.5335	0.2959	0.1706	0	0	0
$\widehat{\boldsymbol{P}}^*$	1.1403	0.1477	0.0799	0.0098	0.0098	0.0093

are shown in Table A.1. It can be seen that, for both matrices, there is a significant gap between the third and fourth largest eigenvalues.

A.1.4 Discussion on node heterogeneity

Node heterogeneity is not addressed in our study, so we discuss it here.

In standard stochastic block models (without autoregressive structure), node heterogeneity often manifests as degree heterogeneity, which can be addressed by degree-corrected models. These models introduce node-specific coefficients to scale the connection probability p_{ij}. However, in our framework, since the static probabilities are replaced by the dynamic probabilities α and β, introducing the degree heterogeneity requires model refinement and parameter tuning. Jiang et al. (2023) proposed a two-way heterogeneity model for this purpose, providing a comprehensive analysis.

At the same time, another type of node heterogeneity is also worth

exploring, that is, the heterogeneity of node change trends. In a mixed-membership autoregressive network model, the dynamic transition probabilities of nodes (i.e., α and β) may contain individual characteristics of each node, meaning that for each node i, there may be a heterogeneity parameter d_i, and the connection probabilities of nodes satisfy

$$\mathbf{P}_1 = \mathbf{D}\mathbf{\Theta}\mathbf{B}_1\mathbf{\Theta}^\top\mathbf{D} \quad \text{and} \quad \mathbf{P}_2 = \mathbf{D}\mathbf{\Theta}\mathbf{B}_2\mathbf{\Theta}^\top\mathbf{D}, \tag{A.3}$$

where $\mathbf{D}$ is a $p \times p$ diagonal matrices with elements d_i. Such networks also exist in reality and have certain research value.

To investigate the heterogeneity of node trends, we apply the AMSC algorithm in simulations. We construct a network consisting of 200 nodes and 3 communities, each containing 50 pure nodes. The memberships of the remaining 50 mixed nodes are randomly assigned from $(x_i, x_i, 1-2x_i)$, $(x_i, 1-2x_i, x_i)$, or $(1-2x_i, x_i, x_i)$, where x_i is a random value uniformly drawn from the interval [0, 0.5]. For the community transition probabilities, $\zeta_{i,i} = \eta_{i,i} = 0.4$, $\zeta_{i,j}$ and $\eta_{i,j}$ are independently drawn from $U[0.05, 0.25]$. For each node i, d_i is a random number uniformly drawn from the interval [0.5,1].

The simulation results are shown in Table A.2. As can be seen, the AMSC algorithm demonstrates excellent performance in estimating the membership matrix for models with heterogeneous node trends. Further-

Table A.2: Mixed-Hamming error rates of the estimated and true membership matrices with 500 replications on the AR-1 mixed membership model with heterogeneity in node change trends.

K	p	n	Mixed-Hamming error rate
3	200	5	0.2211
		20	0.1626
		50	0.0944
		100	0.0575

more, the estimated membership matrix and connection probability matrix can be used to estimate the node heterogeneity parameter d_i. However, within this model framework, the theoretical foundation for the algorithm's application remains to be established. Furthermore, the methods used in static degree correction networks can be extended to dynamic autoregressive networks to improve estimation accuracy. These issues warrant further investigation.

A.1.5 Discussion on BIC

From the definition of BIC_j, we can analyze the difference between consecutive BIC values:

$$\begin{aligned}\mathrm{BIC}_K - \mathrm{BIC}_{K+1} &= (p-1-K)\log\frac{\bar{\lambda}_K}{\bar{\lambda}_{K+1}} + \log\bar{\lambda}_{K+1} - \log\hat{\lambda}_{K+1} - \frac{p-K-1}{p}\log p \\ &= (p-1-K)\log[1 - \frac{(1-\frac{\hat{\lambda}_{K+1}}{\bar{\lambda}_{K+1}})}{p-1-K}] - \log\frac{\hat{\lambda}_{K+1}}{\bar{\lambda}_{K+1}} - \frac{p-K-1}{p}\log p \\ &= (p-1-K)\{\log[1 + \frac{\frac{\hat{\lambda}_{K+1}}{\bar{\lambda}_{K+1}} - 1}{p-1-K}] - \frac{\log\frac{\hat{\lambda}_{K+1}}{\bar{\lambda}_{K+1}}}{p-1-K} - \frac{\log p}{p}\}\end{aligned}$$

For the true number of communities K to be selected, we require $K = \arg\min_{j\le p-2}\mathrm{BIC}_j$, which implies $\mathrm{BIC}_K \le \mathrm{BIC}_{K+1}$. As p tends to infinity, if the ratio $\hat{\lambda}_{K+1}/\bar{\lambda}_{K+1}$ is much larger than $\log p$, i.e., $\hat{\lambda}_{K+1}/\bar{\lambda}_{K+1} >> \log p$, then we would have $\mathrm{BIC}_K > \mathrm{BIC}_{K+1}$, leading to an incorrect estimator. Thus, the condition $\hat{\lambda}_{K+1}/\bar{\lambda}_{K+1} = O(\log p)$ is necessary for K to be the minimum argument. This implies that the performance of BIC is sensitive to the ratio $\hat{\lambda}_{K+1}/\bar{\lambda}_{K+1}$, a ratio that incorporates all of the remaining small eigenvalues $\hat{\lambda}_{K+2}, \ldots, \hat{\lambda}_{p-1}$ through $\bar{\lambda}_{K+1}$, making the criterion unstable.

A.2 Supplement to the real data analysis

A.2.1 Global trade data

In Subsection 6.1 of the paper, we used the proposed algorithm to analyze the global trade data. Here, we present some further details of the results.

First, we provide some analysis results when K is set to 2, Figure A.1 shows the average adjacency matrix of 195 countries from 1991 to 2014, with countries sorted by estimated membership. In Figure A.1, red indicates large values and blue indicates small values. It can be seen that during the estimation period, trade intensity is concentrated in most developed industrial countries in Community 1, while countries in Community 2 have sparse trade, indicating that trade globalization has not yet spread to most underdeveloped countries.

Next, when K is set to 6, we display the specific classification results of the global trade data from 1991 to 2014 in Table A.3. The analytical conclusions we presented in the main text are consistent with this classification results.

Furthermore, To test the robustness of the classification results on global trade data, we estimate the membership using data from 1991 to 2011 and data from 1991 to 2006, and then compare their results.

For data up to 2011, only three of the 195 countries had different classifications than those up to 2014: Malta, Kuwait, and Singapore. For data up to 2006, thirteen countries had different classifications than those up to 2014: Saint Kitts and Nevis, Argentina, Malta, Croatia, Slovenia, Liberia,

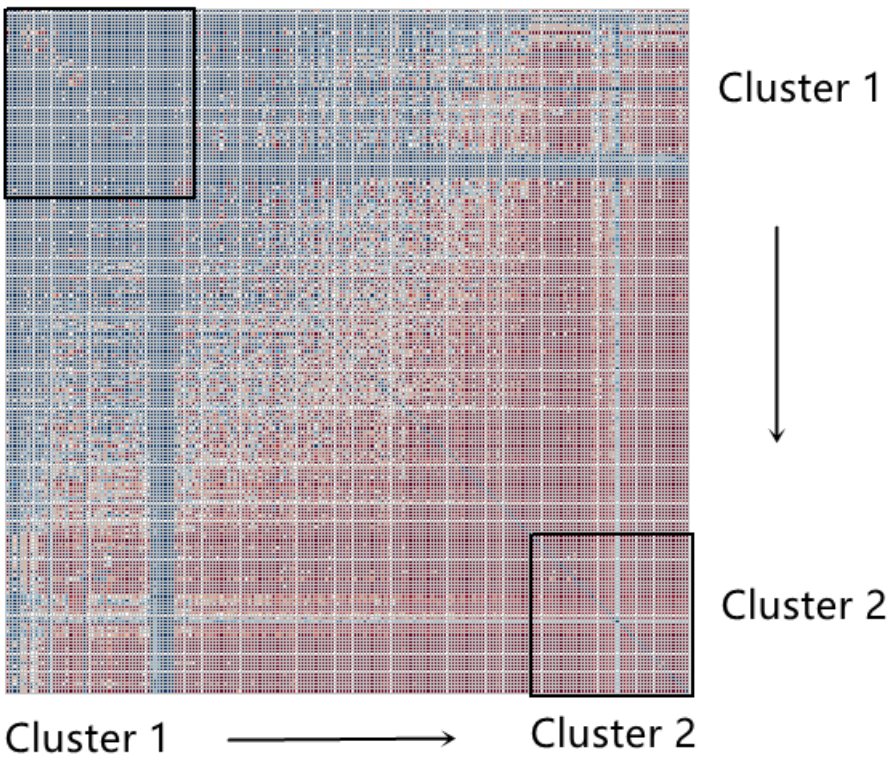

Figure A.1: Average adjacency matrix for the 195 countries in the Global trade data.

Egypt, Yemen, Kuwait, Bahrain, Qatar, Tajikistan, and Singapore. Furthermore, for these thirteen countries whose classifications changed, we find that their maximum memberships in the estimates up to 2006 were all below 0.6, and nine of these countries had maximum memberships below 0.5, meaning that they were considered nodes between multiple communities in the original results, but their distances from any one community were not significantly close. We believe these results indicate that our estimates are

Table A.3: The community detection result of the global trade data (1991-2014) for six communities.

	Country
Cluster1	United States of America, Canada, Colombia, Peru, Brazil, Argentina, United Kingdom, Ireland, Netherlands, Belgium, Luxembourg, France, Switzerland, Spain, Portugal, Germany, Poland, Austria, Slovakia, Italy, Greece, Cyprus, Finland, Sweden, Norway, Denmark, Turkey, Egypt, China, Taiwan, South Korea, Japan, India, Pakistan, Thailand, Malaysia, Singapore, Philippines, Indonesia, Australia, New Zealand .
Cluster2	Bahamas, Cuba, Haiti, Dominican Republic, Jamaica, Trinidad and Tobago, Barbados, Dominica, Grenada, St. Lucia, St. Vincent and the Grenadines, Antigua & Barbuda, St. Kitts and Nevis, Mexico, Belize, Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, Panama, Venezuela, Guyana, Suriname, Ecuador, Bolivia, Paraguay, Chile, Uruguay, Israel.
Cluster3	Monaco, Liechtenstein, Andorra, San Marino, Yugoslavia, Kosovo, Cape Verde, South Sudan, East Timor, Marshall Islands, Palau, Federated States of Micronesia.
Cluster4	Hungary, Czech Republic, Malta, Albania, Montenegro, Macedonia, Croatia, Slovenia, Bulgaria, Moldova, Romania, Russia, Estonia, Latvia, Lithuania, Ukraine, Belarus, Armenia, Georgia, Azerbaijan, Iran, Iraq, Syria, Jordan, Kuwait, Bahrain, Qatar, United Arab Emirates, Afghanistan, Tajikistan, Kyrgyzstan, Kazakhstan, Mongolia, Bangladesh, Sri Lanka, Cambodia, Vietnam.
Cluster5	Bosnia and Herzegovina, Sao Tome and Principe, Guinea-Bissau, Equatorial Guinea, Sierra Leone, Chad, Burundi, Rwanda, Somalia, Djibouti, Eritrea, Malawi, Namibia, Lesotho, Botswana, Swaziland, Comoros, Seychelles, Turkmenistan, Uzbekistan, Bhutan, Myanmar, Maldives, Nepal, Laos, Brunei, Papua New Guinea, Vanuatu, Solomon Islands, Kiribati, Tuvalu, Fiji, Tonga, Nauru, Samoa.
Cluster6	Iceland, Gambia, Mali, Senegal, Benin, Mauritania, Niger, Ivory Coast, Guinea, Burkina Faso, Liberia, Ghana, Togo, Cameroon, Nigeria, Gabon, Central African Republic, Congo, Democratic Republic of the Congo, Uganda, Kenya, Tanzania, Ethiopia, Angola, Mozambique, Zambia, Zimbabwe, South Africa, Madagascar, Mauritius, Morocco, Algeria, Tunisia, Libya, Sudan, Lebanon, Saudi Arabia, Yemen, Oman, North Korea.

sufficiently robust.

In addition to the above analysis, since the original data covers the period from 1870 to 2014, we also analyze trade data before 1991. Using Jiang et al. (2023), we find a change point in the trade data in 1991, likely due to changes in the global situation caused by the end of the Cold War.

Therefore, we analyze the network data before and after 1991 separately to see if there were any significant changes in country classifications.

We initially attempt to use data from 1870 to 1991 for our analysis, but the lack of early trade data made the classification results less interpretable. Therefore, we follow Jiang et al. (2023) and use data from 1950 to 1991. The classification results are shown in Table A.4. Clearly, the classification results before 1991 differ significantly from those after 1991. In addition to the geographic characteristics of other communities, Community 1 and Community 3 include major countries, led by the United States and Russia, respectively—the two major opposing sides of the Cold War—which is consistent with the historical context of the time. These results are consistent with the conclusions of Jiang et al. (2023) and confirm that international trade relations underwent significant changes with 1991 as the dividing point.

A.2.2 French high school contact data

We present details of the analysis of French high school contact data. The dataset is taken from Mastrandrea et al. (2015). Students in the school have different majors: the "MP" classes focus more on mathematics and physics, the "PC" classes focus on physics and chemistry, the "PSI" classes

Table A.4: The community detection results of the global trade data (1950-1991) for six communities

	Country
Cluster1	United States of America, Canada, United Kingdom, Netherlands, Belgium, Luxembourg, France, Switzerland, Spain, Italy, Finland, Sweden, Norway, Denmark, Japan, India.
Cluster2	Jamaica, Trinidad and Tobago, Barbados, Dominica, Guyana, Austria, Malta, Cyprus, Uganda, Kenya, Tanzania, Burundi, Rwanda, Somalia, Djibouti, Mozambique, Zambia, Zimbabwe, Swaziland, Madagascar, Mauritius, Seychelles, Libya, Sudan, Syria, Jordan, Yemen Arab Republic, Yemen, Yemen People's Republic, Kuwait, Bahrain, Qatar, United Arab Emirates, Oman, Afghanistan, North Korea, Bangladesh, Nepal, Malaysia, Singapore, Brunei, Papua New Guinea, Fiji.
Cluster3	Cuba, Brazil, Ireland, Portugal, Poland, Hungary, Czechoslovakia, Yugoslavia, Greece, Bulgaria, Romania, Russia, Iceland, Ethiopia, South Africa, Iran, Turkey, Iraq, Egypt, Lebanon, Israel, Saudi Arabia, China, South Korea, Pakistan, Myanmar, Sri Lanka, Thailand, Philippines, Indonesia, Australia, New Zealand.
Cluster4	Bahamas, Haiti, Dominican Republic, Mexico, Belize, Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, Panama, Colombia, Venezuela, Suriname, Ecuador, Peru, Bolivia, Paraguay, Chile, Argentina, Uruguay.
Cluster5	Grenada, St. Lucia, St. Vincent and the Grenadines, Antigua & Barbuda, St. Kitts and Nevis, Liechtenstein, German Federal Republic, Albania, Cape Verde, Sao Tome and Principe, Guinea-Bissau, Equatorial Guinea, Zanzibar, Namibia, Lesotho, Botswana, Comoros, Mongolia, Bhutan, Maldives, Cambodia, Laos, Republic of Vietnam, Vanuatu, Solomon Islands, Samoa.
Cluster6	Germany, German Democratic Republic, Gambia, Mali, Senegal, Benin, Mauritania, Niger, Ivory Coast, Guinea, Burkina Faso, Liberia, Sierra Leone, Ghana, Togo, Cameroon, Nigeria, Gabon, Central African Republic, Chad, Congo, Democratic Republic of the Congo, Angola, Malawi, Morocco, Algeria, Tunisia, Vietnam.

focus on engineering studies, and the "BIO" classes focus on biology. The data comes from students in nine classes: 3 classes of the "MP" type (MP1, MP2, MP3), two classes of the "PC" type (PC1 and PC2), one class of the "PSI" type (PSI), and 3 classes of the "BIO" type (BIO1, BIO2, BIO3). The dataset contains information on a total of 329 students. Since two of the students did not communicate with anyone else, we excluded them from

the data and used the data of the remaining 327 students for the analysis.

We apply the empirical eigenvalue-threshold estimator to the data set. The result shows that $K = 9$ is the most appropriate setting, which also coincides with the number of classes.

Assume that the community number $K = 9$. We apply the AMSC algorithm to the data and classify students to different communities based on their respective largest membership, we present the result in Table A.5. It can be seen that students of the 9 original classes are identified perfectly into 9 communities, which is better than the result given by Jiang et al. (2023) for the same dataset. In addition, since the algorithm gives a membership vector instead of a simple community label, we can not only classify students by the community that each node belongs to with the largest membership but also explore each node's tendency to belong to other communities, which can help us study the unique communication tendencies of different students in the same class.

Further, to demonstrate the robustness of our estimation of the French high school contact data, we try to randomly delete the contact data of some students and then use the AMSC algorithm to estimate the communities of the remaining students. We try to keep 200 students and 100 students respectively, which means that 127 and 227 students are removed, and

Table A.5: Estimated community distribution for 327 students by the proposed community detection method. The number of communities $K = 9$.

Class	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
BIO1	36	0	0	0	0	0	0	0	0
BIO2	0	34	0	0	0	0	0	0	0
BIO3	0	0	40	0	0	0	0	0	0
MP1	0	0	0	33	0	0	0	0	0
MP2	0	0	0	0	29	0	0	0	0
MP3	0	0	0	0	0	38	0	0	0
PC1	0	0	0	0	0	0	44	0	0
PC2	0	0	0	0	0	0	0	39	0
PSI	0	0	0	0	0	0	0	0	34

we repeat the experiment several times with different random seeds. The results show that the deviation of the estimate gradually becomes larger as the number of students included in the estimation decreases. However, the estimation of our method is always better than the original AR-1 algorithm in Jiang et al. (2023). Table A.6 and Table A.7 show the comparison of our estimation with the estimation using the AR-1 algorithm when the random seed is set to 0. It can be seen that for 200 students, the AMSC algorithm identified all the 9 original classes as the 9 clusters with only 2 students being placed outside their classes, while for the original AR-1 algorithm, 9 students are being placed in the wrong classes. In the situation of 100 students, the AMSC algorithm identified 9 classes as the 9 clusters with 2

Table A.6: Estimated community distribution of the AMSC algorithm after removing 127 students (random seed set to 0), and estimated communities of the AR-1 algorithm for the same dataset. In each cell, the black number on the left is the estimation by the AMSC algorithm, and the red number in parentheses on the right is the estimation by the AR-1 algorithm.

Class	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
BIO1	25(25)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
BIO2	0(0)	18(18)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
BIO3	0(0)	0(0)	22(21)	0(0)	0(0)	0(0)	0(0)	0(0)	0(1)
MP1	0(0)	0(0)	0(0)	14(12)	0(2)	1(1)	0(0)	0(0)	1(1)
MP2	0(0)	0(0)	0(0)	0(0)	19(19)	0(0)	0(0)	0(0)	0(0)
MP3	0(0)	0(0)	0(0)	0(0)	0(0)	23(23)	0(0)	0(0)	0(0)
PC1	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	29(27)	0(1)	0(1)
PC2	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	23(22)	0(1)
PSI	0(0)	0(0)	0(0)	0(0)	0(1)	0(0)	0(0)	0(0)	25(24)

students being placed outside their classes. The original AR-1 algorithm, on the other hand, performed poorly and even divided the majority of students from two different classes "PC1" and "PSI" into the same cluster, Cluster 7.

Table A.7: Estimated community distribution of the AMSC algorithm after removing 227 students (random seed set to 0), and estimated communities of the AR-1 algorithm for the same dataset. In each cell, the black number on the left is the estimation by the AMSC algorithm, and the red number in parentheses on the right is the estimation by the AR-1 algorithm.

Class	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9
BIO1	14(14)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
BIO2	0(0)	13(5)	0(6)	0(0)	0(0)	0(0)	0(0)	0(0)	0(2)
BIO3	0(0)	0(4)	10(4)	0(0)	0(0)	0(0)	0(0)	0(0)	0(2)
MP1	0(0)	0(0)	0(0)	8(6)	0(1)	0(1)	0(0)	0(0)	0(0)
MP2	0(0)	0(0)	0(0)	1(0)	11(10)	0(0)	0(0)	0(0)	0(2)
MP3	0(0)	0(0)	0(0)	0(0)	1(0)	8(8)	0(0)	0(0)	0(1)
PC1	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	12(10)	0(1)	0(1)
PC2	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	11(10)	0(1)
PSI	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(8)	0(0)	11(3)

A.3 Technical proofs

A.3.1 Proof of Theorem 1

proof: It is obvious that for any pure node i belonging to the community k, the corresponding row $\boldsymbol{\theta}_i$ in $\boldsymbol{\Theta}$ satisfies $\boldsymbol{\theta}_i(k) = 1$ and all remaining elements are 0. Since $\mathcal{I}$ is the indices of rows corresponding to K pure nodes from each community, without loss of generality, we can reorder the nodes to ensure that $\boldsymbol{\Theta}(\mathcal{I}, :) = \boldsymbol{I}$. Here, $\boldsymbol{I}$ is the identity matrix.

Note that

$$\begin{aligned}\boldsymbol{P}^* &= \frac{\boldsymbol{P}_1\boldsymbol{P}_1}{\|\boldsymbol{P}_1\|_F^2} + \frac{\boldsymbol{P}_2\boldsymbol{P}_2}{\|\boldsymbol{P}_2\|_F^2} \\ &= \boldsymbol{\Theta}\left(\frac{\boldsymbol{B}_1\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{B}_1}{\|\boldsymbol{P}_1\|_F^2} + \frac{\boldsymbol{B}_2\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{B}_2}{\|\boldsymbol{P}_2\|_F^2}\right)\boldsymbol{\Theta}^\top.\end{aligned}$$

From Assumption 3 and $\boldsymbol{\Theta}(\mathcal{I},:) = \boldsymbol{I}$, it can be seen that $\text{rank}(\boldsymbol{P}^*) = K$.

Then, we have

$$\boldsymbol{\Gamma}(\mathcal{I},:)\boldsymbol{\Lambda}\boldsymbol{\Gamma}(\mathcal{I},:)^\top = \boldsymbol{P}^*(\mathcal{I},\mathcal{I}) = \boldsymbol{\Theta}(\mathcal{I},:)(\frac{\boldsymbol{B}_1\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{B}_1}{\|\boldsymbol{P}_1\|_F^2}+\frac{\boldsymbol{B}_2\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{B}_2}{\|\boldsymbol{P}_2\|_F^2})\boldsymbol{\Theta}(\mathcal{I},:)^\top = \boldsymbol{B}^*,$$

which shows that $\boldsymbol{\Gamma}(\mathcal{I},:) \in \mathbb{R}^{K\times K}$ is full rank. Then, we can obtain that

$$\boldsymbol{\Gamma}(\mathcal{I},:)\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top = \boldsymbol{P}^*(\mathcal{I},:) = \boldsymbol{\Theta}(\mathcal{I},:)(\frac{\boldsymbol{B}_1\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{B}_1}{\|\boldsymbol{P}_1\|_F^2} + \frac{\boldsymbol{B}_2\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\boldsymbol{B}_2}{\|\boldsymbol{P}_2\|_F^2})\boldsymbol{\Theta}^\top = \boldsymbol{B}^*\boldsymbol{\Theta}^\top,$$

therefore, we have

$$\boldsymbol{\Gamma} = \boldsymbol{P}^*\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Theta}\boldsymbol{B}^*\boldsymbol{\Theta}^\top\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Theta}\boldsymbol{\Gamma}(\mathcal{I},:)\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{-1} = \boldsymbol{\Theta}\boldsymbol{\Gamma}(\mathcal{I},:).$$

This completes the proof.

A.3.2 Proof of Proposition 1

Proof: Note that $\lambda_1 \geq \cdots \geq \lambda_K > \lambda_{K+1} = \cdots = \lambda_p = 0$ are the eigenvalues of $\boldsymbol{P}^*$, and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ are the eigenvalues of $\widehat{\boldsymbol{P}}^*$. Since both $\boldsymbol{P}^*$ and $\widehat{\boldsymbol{P}}^*$ are symmetric matrices, by Theorem A.37 in Bai and Silverstein (2010), it follows that

$$\begin{aligned}\max_{i=1,\ldots,p}\left\{\lambda_i - \hat{\lambda}_i\right\}^2 &\leq \sum_{i=1}^{p}\left\{\lambda_i - \hat{\lambda}_i\right\}^2 \leq \text{tr}\left\{\left(\boldsymbol{P}^* - \widehat{\boldsymbol{P}}^*\right)\left(\boldsymbol{P}^* - \widehat{\boldsymbol{P}}^*\right)^{\text{T}}\right\} \\ &\leq \|\boldsymbol{P}^* - \widehat{\boldsymbol{P}}^*\|_F^2.\end{aligned}$$

Therefore, using the conclusion in (A.12), we have that with probability $1-4\exp\{-c\log p\}$, there exists a constant C such that

$$\max_{i=1,\ldots,p}|\lambda_i-\hat{\lambda}_i|\leq\|\boldsymbol{P}^*-\widehat{\boldsymbol{P}}^*\|_F\leq\frac{C(\sqrt{p\log p}+\sqrt{n})}{\sqrt{np}}, \tag{A.4}$$

which, jointly with (A.4), implies that

$$\max_{i=1,\ldots,K}|\lambda_i-\hat{\lambda}_i|\leq\frac{C(\sqrt{p\log p}+\sqrt{n})}{\sqrt{np}}, \tag{A.5}$$

$$\max_{i=(K+1),\ldots,p}\hat{\lambda}_i\leq\frac{C(\sqrt{p\log p}+\sqrt{n})}{\sqrt{np}}. \tag{A.6}$$

If $k\leq K$, by the assumption that $\lambda_K>c_0$, we have $\lambda_k>c_0$, which, jointly with (A.5), yields that $\hat{\lambda}_k>c_0-\frac{C(\sqrt{p\log p}+\sqrt{n})}{\sqrt{np}}$ with probability $1-4\exp\{-c\log p\}$. Since $\frac{\sqrt{p\log p}+\sqrt{n}}{\sqrt{np}}\to 0$ as $n,p\to\infty$, if we set $\xi\leq c_0/2$, it follows that $\hat{\lambda}_k>\xi*\min\left(1,(\frac{\sqrt{p\log p}+\sqrt{n}}{\sqrt{np}})^r\right)$, *a.s.* as $n,p\to\infty$.

Similarly, if $k>K$, since r satisfy $0<r<1$, by (A.6), for any constant c_r, $\hat{\lambda}_k\leq c_r(\frac{(\sqrt{p\log p}+\sqrt{n})}{\sqrt{np}})^r$, *a.s.* as $n,p\to\infty$. Therefore, we have $\hat{\lambda}_k<\xi*\min\left(1,(\frac{\sqrt{p\log p}+\sqrt{n}}{\sqrt{np}})^r\right)$, *a.s.* as $n,p\to\infty$.

A.3.3 Proof of Theorem 2

Proof: We first find a bound of $\|\widehat{\boldsymbol{P}}^*-\boldsymbol{P}^*\|_F$. Note that

$$\begin{aligned}\|\widehat{\boldsymbol{P}}^*-\boldsymbol{P}^*\|_F=&\ \|(\widehat{\boldsymbol{P}}_1^*\widehat{\boldsymbol{P}}_1^*+\widehat{\boldsymbol{P}}_2^*\widehat{\boldsymbol{P}}_2^*)-(\boldsymbol{P}_1^*\boldsymbol{P}_1^*+\boldsymbol{P}_2^*\boldsymbol{P}_2^*)\|_F\\=&\ \|(\widehat{\boldsymbol{P}}_1^*\widehat{\boldsymbol{P}}_1^*-\boldsymbol{P}_1^*\boldsymbol{P}_1^*)+(\widehat{\boldsymbol{P}}_2^*\widehat{\boldsymbol{P}}_2^*-\boldsymbol{P}_2^*\boldsymbol{P}_2^*)\|_F\\\leq&\ \|\widehat{\boldsymbol{P}}_1^*\widehat{\boldsymbol{P}}_1^*-\boldsymbol{P}_1^*\boldsymbol{P}_1^*\|_F+\|\widehat{\boldsymbol{P}}_2^*\widehat{\boldsymbol{P}}_2^*-\boldsymbol{P}_2^*\boldsymbol{P}_2^*\|_F.\end{aligned} \tag{A.7}$$

To bound $\|\widehat{\boldsymbol{P}}^* - \boldsymbol{P}^*\|_F$, we only need to bound $\|\widehat{\boldsymbol{P}}_1^*\widehat{\boldsymbol{P}}_1^* - \boldsymbol{P}_1^*\boldsymbol{P}_1^*\|_F$ since a bound of $\|\widehat{\boldsymbol{P}}_2^*\widehat{\boldsymbol{P}}_2^* - \boldsymbol{P}_2^*\boldsymbol{P}_2^*\|_F$ can be obtained in a similar way. In view that

$$\begin{aligned}
\|\widehat{\boldsymbol{P}}_1^*\widehat{\boldsymbol{P}}_1^* - \boldsymbol{P}_1^*\boldsymbol{P}_1^*\|_F \leq & \ \|\tfrac{\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1}{\|\widehat{\boldsymbol{P}}_1\|_F^2} - \tfrac{\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1}{\|\boldsymbol{P}_1\|_F^2}\|_F + \|\tfrac{\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1}{\|\boldsymbol{P}_1\|_F^2} - \tfrac{\boldsymbol{P}_1\boldsymbol{P}_1}{\|\boldsymbol{P}_1\|_F^2}\|_F \\
= & \ \tfrac{|\|\widehat{\boldsymbol{P}}_1\|_F^2 - \|\boldsymbol{P}_1\|_F^2|\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1\|_F}{\|\widehat{\boldsymbol{P}}_1\|_F^2\|\boldsymbol{P}_1\|_F^2} + \tfrac{\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F^2} \\
\leq & \ \tfrac{(\|\widehat{\boldsymbol{P}}_1\|_F + \|\boldsymbol{P}_1\|_F|\|\widehat{\boldsymbol{P}}_1\|_F - \|\boldsymbol{P}_1\|_F|)\|\widehat{\boldsymbol{P}}_1\|_F^2}{\|\widehat{\boldsymbol{P}}_1\|_F^2\|\boldsymbol{P}_1\|_F^2} + \tfrac{\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F^2} \\
\leq & \ \tfrac{2\|\boldsymbol{P}_1\|_F|\|\widehat{\boldsymbol{P}}_1\|_F - \|\boldsymbol{P}_1\|_F|}{\|\boldsymbol{P}_1\|_F^2} + \tfrac{(\|\widehat{\boldsymbol{P}}_1\|_F - \|\boldsymbol{P}_1\|_F)^2}{\|\boldsymbol{P}_1\|_F^2} + \tfrac{\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F^2} \\
\leq & \ \tfrac{2\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F} + (\tfrac{\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F})^2 + \tfrac{\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F^2},
\end{aligned} \tag{A.8}$$

we first bound $\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F$. Denote $\widetilde{\boldsymbol{P}}_1 = \boldsymbol{P}_1 - \mathrm{diag}(\boldsymbol{P}_1)$. Since $\mathrm{diag}(\widehat{\boldsymbol{P}}_1) = 0$, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F \leq & \ \|\widehat{\boldsymbol{P}}_1 - \widetilde{\boldsymbol{P}}_1\|_F + \|\widetilde{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F \\
= & \ \sqrt{\sum_{\substack{1\leq i,j\leq p \\ i\neq j}} (\hat{\alpha}_{i,j} - \alpha_{i,j})^2} + \|\mathrm{diag}(\boldsymbol{P}_1)\|_F \\
= & \ \sqrt{\sum_{\substack{1\leq i,j\leq p \\ i\neq j}} (\hat{\alpha}_{i,j} - \alpha_{i,j})^2} + \sqrt{\sum_{1\leq k\leq p} (\alpha_{k,k})^2}.
\end{aligned}$$

From Assumption 1, we have for all $1 \leq i, j \leq p$, $l \leq \alpha_{i,j} \leq 1$, which, jointly with Proposition A.1.1, implies that for any constant $c > 0$, there exists a large constant C such that with probability $1 - 2\exp\{-c\log p\}$,

$$\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F \leq C\left(p\sqrt{\frac{\log p}{n}} + \sqrt{p}\right). \tag{A.9}$$

We next bound $\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F$. Since

$$\begin{aligned}
\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F \leq \;& \|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \widetilde{\boldsymbol{P}}_1\widetilde{\boldsymbol{P}}_1\|_F + \|\widetilde{\boldsymbol{P}}_1\widetilde{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F \\
= \;& \sqrt{\sum_{\substack{1\leq i,j\leq p \\ i\neq j}} [\sum_{k=1}^{p}(\hat{\alpha}_{i,k}\hat{\alpha}_{j,k} - \alpha_{i,k}\alpha_{j,k})]^2} + \sqrt{\sum_{1\leq i,j\leq p}[(\alpha_{i,i}+\alpha_{j,j})\alpha_{i,j}]^2 - \sum_{1\leq i\leq p} 3(\alpha_{i,i})^4} \\
= \;& \sqrt{\sum_{\substack{1\leq i,j\leq p \\ i\neq j}} [\sum_{k=1}^{p}(\hat{\alpha}_{i,k}\hat{\alpha}_{j,k} - \alpha_{i,k}\hat{\alpha}_{j,k} + \alpha_{i,k}\hat{\alpha}_{j,k} - \alpha_{i,k}\alpha_{j,k})]^2} \\
+ \;& \sqrt{\sum_{1\leq i,j\leq p}[(\alpha_{i,i}+\alpha_{j,j})\alpha_{i,j}]^2 - \sum_{1\leq i\leq p} 3(\alpha_{i,i})^4} \\
= \;& \sqrt{\sum_{\substack{1\leq i,j\leq p \\ i\neq j}} \{\sum_{k=1}^{p}[\hat{\alpha}_{j,k}(\hat{\alpha}_{i,k} - \alpha_{i,k}) + \alpha_{i,k}(\hat{\alpha}_{j,k} - \alpha_{j,k})]\}^2} \\
+ \;& \sqrt{\sum_{1\leq i,j\leq p}[(\alpha_{i,i}+\alpha_{j,j})\alpha_{i,j}]^2 - \sum_{1\leq i\leq p} 3(\alpha_{i,i})^4}.
\end{aligned} \tag{A.10}$$

Note that by (2.5), it follows that for all $1 \leq i, j \leq p$, $\hat{\alpha}_{i,j} \leq 1$. Therefore, by Assumption 1, Proposition A.1.1 and (A.10), we have

$$\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F \leq C(\sqrt{p^2\left(p\sqrt{\frac{\log p}{n}}\right)^2} + \sqrt{p^2 - p}) \leq C\left(p^2\sqrt{\frac{\log p}{n}} + p\right). \tag{A.11}$$

Note that

$$\|\boldsymbol{P}_1\|_F = \sqrt{\sum_{1\leq i,j\leq p}(\alpha_{i,j})^2} \geq pl.$$

By (A.8), (A.9) and (A.11), it follows that

$$\frac{\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F} \leq \frac{C}{l}\left(\sqrt{\frac{\log p}{n}} + \frac{\sqrt{p}}{p}\right),$$

$$\frac{\|\|\widehat{\boldsymbol{P}}_1\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F^2} \leq \frac{C}{l^2}\left(\sqrt{\frac{\log p}{n}} + \frac{1}{p}\right).$$

When $n, p \to \infty$, it is easy to show that

$$\left(\frac{\|\widehat{\boldsymbol{P}}_1 - \boldsymbol{P}_1\|_F}{\|\boldsymbol{P}_1\|_F}\right)^2 = o\left(\sqrt{\frac{\log p}{n}} + \frac{\sqrt{p}}{p}\right).$$

Therefore, for any constant $c > 0$, there exists a large constant C_{p1} such that with probability $1 - 2\exp\{-c\log p\}$,

$$\|\widehat{\boldsymbol{P}}_1^*\widehat{\boldsymbol{P}}_1^* - \boldsymbol{P}_1^*\boldsymbol{P}_1^*\|_F \leq C_{p1}\left(\sqrt{\frac{\log p}{n}} + \frac{\sqrt{p}}{p}\right).$$

Similarly, we can show that there exists a large constant C_{p2} such that with probability $1 - 2\exp\{-c\log p\}$,

$$\|\widehat{\boldsymbol{P}}_2^*\widehat{\boldsymbol{P}}_2^* - \boldsymbol{P}_2^*\boldsymbol{P}_2^*\|_F \leq C_{p2}\left(\sqrt{\frac{\log p}{n}} + \frac{\sqrt{p}}{p}\right).$$

Therefore, by (A.7), we have that for any constant $c > 0$, with probability $1 - 4\exp\{-c\log p\}$, there exists a constant $C_p = C_{p1} + C_{p2}$ satisfy that

$$\|\widehat{\boldsymbol{P}}^* - \boldsymbol{P}^*\|_F \leq C_p\left(\sqrt{\frac{\log p}{n}} + \frac{\sqrt{p}}{p}\right) = C_p\left(\frac{\sqrt{p\log p} + \sqrt{n}}{\sqrt{np}}\right). \quad \text{(A.12)}$$

To show (4.2), by Davis-Kahan Theorem in Yu et al. (2014), we have that for constant $C_\gamma = 2^{3/2}C_p$, there exist a $K \times K$ orthogonal matrix $\boldsymbol{O}$

such that

$$\|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O} - \boldsymbol{\Gamma}\|_F \leq \frac{C_\gamma(\sqrt{p\log p} + \sqrt{n})}{\sqrt{np}\lambda_K}.$$

This completes the proof.

A.3.4 Proof of Theorem 3

Proof: As mentioned in Section 4 of our paper. The set $\mathcal{S}$ is defined as $\mathcal{S} = \{\boldsymbol{U} \in R^{K\times K}\}$. We show that for any $\boldsymbol{U} \in \mathcal{S}$, $\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \boldsymbol{U})$ converges to $\mathcal{L}(\boldsymbol{\Gamma}; \boldsymbol{U})$. For each i, let $\boldsymbol{u}_{1i}$ and $\boldsymbol{u}_{2i}$ be the rows in $\boldsymbol{U}$ that are closest to $(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:)$ and $\boldsymbol{\Gamma}(i,:)$, respectively, note that

$$\|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{2i}\|_2 \leq \|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{1i}\|_2,$$

$$\|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{1i}\|_2 \leq \|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{2i}\|_2.$$

so we have

$$\begin{aligned}
\|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{2i}\|_2 - \|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{1i}\|_2 &\leq \|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{1i}\|_2 - \|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{1i}\|_2 \\
&\leq \|\boldsymbol{\Gamma}(i,:) - (\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:)\|_2, \\
\|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{1i}\|_2 - \|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{2i}\|_2 &\leq \|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{2i}\|_2 - \|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{2i}\|_2 \\
&\leq \|\boldsymbol{\Gamma}(i,:) - (\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:)\|_2.
\end{aligned}$$

Therefore, we have

$$|\|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{1i}\|_2 - \|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{2i}\|_2| \leq \|\boldsymbol{\Gamma}(i,:) - (\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:)\|_2,$$

Note that the loss function $\mathcal{L}(\boldsymbol{Q}, \boldsymbol{U})$ is defined as

$$\mathcal{L}(\boldsymbol{Q}, \boldsymbol{U}) = \frac{1}{p} \sum_{i=1}^{p} \min_{1 \le k \le K} \|\boldsymbol{Q}(i,:) - \boldsymbol{U}(k,:)\|_2.$$

We get

$$\begin{aligned} |\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \boldsymbol{U}) - \mathcal{L}(\boldsymbol{\Gamma}; \boldsymbol{U})| &= |\frac{1}{p} \sum_{i=1}^{p} (\|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{u}_{1i}\|_2 - \|\boldsymbol{\Gamma}(i,:) - \boldsymbol{u}_{2i}\|_2)| \\ \le \sqrt{\frac{1}{p} \sum_{i=1}^{p} \|(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(i,:) - \boldsymbol{\Gamma}(i,:)\|_2^2} &= \frac{1}{\sqrt{p}} \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O} - \boldsymbol{\Gamma}\|_F. \end{aligned} \tag{A.13}$$

Since (A.13) holds for any matrix $\boldsymbol{U}$, we obtain that

$$\sup_{\boldsymbol{U} \in \mathcal{S}} |\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \boldsymbol{U}) - \mathcal{L}(\boldsymbol{\Gamma}; \boldsymbol{U})| \le \frac{\|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O} - \boldsymbol{\Gamma}\|_F}{\sqrt{p}}.$$

Then, we bound $\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F$. Note that

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Gamma}; \widehat{\boldsymbol{V}}\boldsymbol{O}) - \mathcal{L}(\boldsymbol{\Gamma}; \boldsymbol{V}) &\le |\mathcal{L}(\boldsymbol{\Gamma}; \widehat{\boldsymbol{V}}\boldsymbol{O}) - \mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \widehat{\boldsymbol{V}}\boldsymbol{O})| \\ &\quad + (\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \widehat{\boldsymbol{V}}\boldsymbol{O}) - \mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \boldsymbol{V})) \\ &\quad + |\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \boldsymbol{V}) - \mathcal{L}(\boldsymbol{\Gamma}; \boldsymbol{V})| \\ &\le 2 \sup_{\boldsymbol{U} \in \mathcal{S}} |\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}; \boldsymbol{U}) - \mathcal{L}(\boldsymbol{\Gamma}; \boldsymbol{U})|. \end{aligned}$$

By Assumption 4, we have that with probability at least $1 - 4\exp\{-c \log p\}$, the following holds:

$$\begin{aligned}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F &\leq (\kappa K^{-1})^{-1}(\mathcal{L}(\boldsymbol{\Gamma};\widehat{\boldsymbol{V}}\boldsymbol{O})-\mathcal{L}(\boldsymbol{\Gamma};\boldsymbol{V}))\\ &\leq \frac{2K}{\kappa}\sup_{\boldsymbol{U}\in\mathcal{S}}|\mathcal{L}(\widehat{\boldsymbol{\Gamma}}\boldsymbol{O};\boldsymbol{U})-\mathcal{L}(\boldsymbol{\Gamma};\boldsymbol{U})|\\ &\leq \frac{2K}{\kappa}\left(\frac{\|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}-\boldsymbol{\Gamma}\|_F}{\sqrt{p}}\right)\\ &\leq \frac{2K}{\kappa}\left(\frac{C_\gamma(\sqrt{p\log p}+\sqrt{n})}{p\sqrt{n}\lambda_K}\right)\\ &= \frac{C_v K(\sqrt{p\log p}+\sqrt{n})}{p\sqrt{n}\lambda_K},\end{aligned}$$

where $C_v=\frac{2C_\gamma}{\kappa}$. This completes the proof.

A.3.5 Proof of Theorem 4

Proof: First, let $m=\frac{min_i\{\|\boldsymbol{\Gamma}(i,:)\|_2^2\}}{max_i\{\|\boldsymbol{\Gamma}(i,:)\|_2^2\}}$. Note that $\|\boldsymbol{\Gamma}\|_F=\sqrt{K}$, since $\boldsymbol{V}=\boldsymbol{\Gamma}(\mathcal{I},:)$ is a submatrix of $\boldsymbol{\Gamma}$, we have

$$\|\boldsymbol{V}\|_F=\sqrt{\frac{\|\boldsymbol{V}\|_F^2}{\|\boldsymbol{\Gamma}\|_F^2}\|\boldsymbol{\Gamma}\|_F^2}=\sqrt{\frac{\|\boldsymbol{\Gamma}(\mathcal{I},:)\|_F^2}{(\sum_{i=1}^p\|\boldsymbol{\Gamma}(i,:)\|_2^2)}K}.$$

Therefore, we have

$$\frac{\sqrt{m}K}{\sqrt{p}}=\sqrt{\frac{mK}{p}K}\leq\|\boldsymbol{V}\|_F\leq\sqrt{\frac{K}{mp}K}=\frac{K}{\sqrt{mp}}.$$

Let each row $\boldsymbol{y}_i$ of matrix $\boldsymbol{Y}$ satisfy

$$\boldsymbol{y}_i=\boldsymbol{\theta}_i.$$

It is obvious that

$$\boldsymbol{\theta}_i = \frac{\boldsymbol{y}_i}{\|\boldsymbol{y}_i\|_1},$$

and

$$\boldsymbol{\Gamma}(i,:) = \boldsymbol{y}_i \boldsymbol{V}.$$

Hence,

$$\boldsymbol{Y} = \boldsymbol{\Gamma}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1},$$

Since

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\|_2 &= \|\frac{\widetilde{\boldsymbol{y}}_i}{\|\widetilde{\boldsymbol{y}}_i\|_1} - \frac{\boldsymbol{y}_i}{\|\boldsymbol{y}_i\|_1}\|_2 = \|\frac{\widetilde{\boldsymbol{y}}_i\|\boldsymbol{y}_i\|_1 - \boldsymbol{y}_i\|\widetilde{\boldsymbol{y}}_i\|_1}{\|\widetilde{\boldsymbol{y}}_i\|_1\|\boldsymbol{y}_i\|_1}\|_2 \\
&= \|\frac{\widetilde{\boldsymbol{y}}_i\|\boldsymbol{y}_i\|_1 - \widetilde{\boldsymbol{y}}_i\|\widetilde{\boldsymbol{y}}_i\|_1 + \widetilde{\boldsymbol{y}}_i\|\widetilde{\boldsymbol{y}}_i\|_1 - \boldsymbol{y}_i\|\widetilde{\boldsymbol{y}}_i\|_1}{\|\widetilde{\boldsymbol{y}}_i\|_1\|\boldsymbol{y}_i\|_1}\|_2 \\
&\leq \frac{\|\widetilde{\boldsymbol{y}}_i\|\boldsymbol{y}_i\|_1 - \widetilde{\boldsymbol{y}}_i\|\widetilde{\boldsymbol{y}}_i\|_1\|_2 + \|\widetilde{\boldsymbol{y}}_i\|\widetilde{\boldsymbol{y}}_i\|_1 - \boldsymbol{y}_i\|\widetilde{\boldsymbol{y}}_i\|_1\|_2}{\|\widetilde{\boldsymbol{y}}_i\|_1\|\boldsymbol{y}_i\|_1} \\
&= \frac{\|\widetilde{\boldsymbol{y}}_i\|_2|\|\boldsymbol{y}_i\|_1 - \|\widetilde{\boldsymbol{y}}_i\|_1| + \|\widetilde{\boldsymbol{y}}_i\|_1\|\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2}{\|\widetilde{\boldsymbol{y}}_i\|_1\|\boldsymbol{y}_i\|_1} \\
&\leq \frac{\|\widetilde{\boldsymbol{y}}_i\|_1|\|\boldsymbol{y}_i\|_1 - \|\widetilde{\boldsymbol{y}}_i\|_1| + \|\widetilde{\boldsymbol{y}}_i\|_1\|\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2}{\|\widetilde{\boldsymbol{y}}_i\|_1\|\boldsymbol{y}_i\|_1} \\
&= \frac{|\|\boldsymbol{y}_i\|_1 - \|\widetilde{\boldsymbol{y}}_i\|_1| + \|\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2}{\|\boldsymbol{y}_i\|_1} \leq \frac{\|\boldsymbol{y}_i - \widetilde{\boldsymbol{y}}_i\|_1 + \|\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2}{\|\boldsymbol{y}_i\|_1} \\
&\leq \frac{(\sqrt{K}+1)\|\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2}{\|\boldsymbol{y}_i\|_1} = (\sqrt{K}+1)\|\widetilde{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2,
\end{aligned}$$

it follows that

$$\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_F \leq (\sqrt{K}+1)\|\widetilde{\boldsymbol{Y}} - \boldsymbol{Y}\|_F.$$

Since all elements of $\boldsymbol{\Theta}$ are non-negative, by the definition of $\boldsymbol{Y}$, all elements of $\boldsymbol{Y}$ are also non-negative, which means that for any $1 \leq i \leq p, 1 \leq j \leq K$,

$\boldsymbol{Y}_{ij} = \max\{\boldsymbol{Y}_{ij}, 0\}$), so we have

$$|\widetilde{\boldsymbol{Y}}_{ij} - \boldsymbol{Y}_{ij}| \leq |\widehat{\boldsymbol{Y}}_{ij} - \boldsymbol{Y}_{ij}|,$$

which implies that

$$\|\widetilde{\boldsymbol{Y}} - \boldsymbol{Y}\|_F \leq \|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F.$$

Therefore, we only need to bound $\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F$. By Theorem 2, we have

$$\begin{aligned}
\|\widehat{\boldsymbol{Y}} - \boldsymbol{Y}\|_F &= \|\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{V}}^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{\Gamma}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&= \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{\Gamma}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&= \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{O}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1} \\
&\quad + \widehat{\boldsymbol{\Gamma}}\boldsymbol{O}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1} - \boldsymbol{\Gamma}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \widehat{\boldsymbol{\Gamma}}\boldsymbol{O}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\quad + \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1} - \boldsymbol{\Gamma}\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O}\|_F\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\quad + \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O} - \boldsymbol{\Gamma}\|_F\|\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \|\widehat{\boldsymbol{\Gamma}}\|_F\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\quad + \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O} - \boldsymbol{\Gamma}\|_F\|\boldsymbol{V}\|_F\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \sqrt{K}\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F + \|\widehat{\boldsymbol{\Gamma}}\boldsymbol{O} - \boldsymbol{\Gamma}\|_F\|\boldsymbol{V}\|_F\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \sqrt{K}\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F + \frac{C_\gamma K(\sqrt{p\log p} + \sqrt{n})}{p\sqrt{mn}\lambda_K}\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F.
\end{aligned} \tag{A.14}$$

Here, the second line holds because

$$\widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{V}}^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} = (\widehat{\boldsymbol{\Gamma}}\boldsymbol{O})(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top((\widehat{\boldsymbol{V}}\boldsymbol{O})(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1},$$

while the second line from the bottom hold because $\|\widehat{\boldsymbol{\Gamma}}\|_F = \sqrt{K}$ and $\|\boldsymbol{V}\|_F \leq \frac{K}{\sqrt{mp}}$.

Next, we need to bound $\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1} - \boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F$ and $\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F$. Since $\boldsymbol{V}\boldsymbol{V}^\top$ is a positive definite symmetric matrix, we have

$$\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \leq \frac{\sqrt{K}}{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)}. \tag{A.15}$$

From the fact that $\|\boldsymbol{V}\|_F \leq \frac{K}{\sqrt{mp}}$, $|\lambda_{\min}(\boldsymbol{V})| > 0$, it follows that

$$\begin{aligned}
&\|\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_F \leq \|\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V})^\top + (\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V})\boldsymbol{V}^\top\|_F \\
&\leq (\|\widehat{\boldsymbol{V}}\boldsymbol{O}\|_F + \|\boldsymbol{V}\|_F)\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F \leq (\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F + 2\|\boldsymbol{V}\|_F)\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F \\
&\leq (\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F + 2\frac{K}{\sqrt{mp}})\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F
\end{aligned}$$

Then we have

$$\begin{aligned}
&\|(\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top)^{-1} - (\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \leq \|(\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top)^{-1}\|_F\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F\|\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_F \\
\leq&\sqrt{K}(\lambda_{\min}(\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top))^{-1}\sqrt{K}(\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top))^{-1}\|\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_F \\
\leq&K\left(\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top) - \|\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_F\right)^{-1}(\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top))^{-1}\|\widehat{\boldsymbol{V}}\boldsymbol{O}\widehat{\boldsymbol{V}}\boldsymbol{O}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_F \\
\leq&K(\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top) - (\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F + 2\frac{K}{\sqrt{mp}})\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F)^{-1}(\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top))^{-1} \\
&(\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F + 2\frac{K}{\sqrt{mp}})\|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_F
\end{aligned} \tag{A.16}$$

Since $\left(\sqrt{p\log p}+\sqrt{n}\right)/(\sqrt{np}\lambda_K) \to 0$, we have $\sqrt{p}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F \to 0$. Therefore, $\frac{\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{2\frac{K}{\sqrt{mp}}} \to 0$ and $\frac{2\frac{K}{\sqrt{mp}}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)} \leq \frac{2\sqrt{p}K\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{\sqrt{m}m_V} \to 0$, Thus

$$\|(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \leq C_\lambda\frac{K^2}{\sqrt{mp}}(\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top))^{-2}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F, \tag{A.17}$$

where $C_\lambda = \frac{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)(2+\frac{\sqrt{mp}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{K})}{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)-(2\frac{K}{\sqrt{mp}}+\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F)\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}$.

Next, by (A.15) and (A.17), we have

$$\begin{aligned}
&\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&= \|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}+(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}-\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \|\widehat{\boldsymbol{V}}\|_F\|(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F+\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \frac{K}{\sqrt{mp}}\|(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F+\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F \\
&\leq \frac{K}{\sqrt{mp}}\|(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F+\frac{\sqrt{K}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)} \\
&\leq \frac{C_\lambda K^3\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{mp\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)^2}+\frac{\sqrt{K}\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)}.
\end{aligned} \tag{A.18}$$

By (A.14), (A.15), and (A.18), it follows that

$$\begin{aligned}
\|\widehat{\boldsymbol{Y}}-\boldsymbol{Y}\|_F &\le \sqrt{K}\|(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top(\widehat{\boldsymbol{V}}\boldsymbol{O}(\widehat{\boldsymbol{V}}\boldsymbol{O})^\top)^{-1}-\boldsymbol{V}^\top(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F+\frac{C_\gamma K(\sqrt{p\log p}+\sqrt{n})}{p\sqrt{mn}\lambda_K}\|(\boldsymbol{V}\boldsymbol{V}^\top)^{-1}\|_F\\
&\le \sqrt{K}\left(\frac{C_\lambda K^3}{mp\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)}+\sqrt{K}\right)\frac{\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)}+\frac{C_\gamma K^{3/2}(\sqrt{p\log p}+\sqrt{n})}{p\sqrt{mn}\lambda_K\lambda_{\min}(\boldsymbol{V}\boldsymbol{V}^\top)}\\
&\le \sqrt{K}\left(\frac{C_\lambda K^3}{mm_V}+\sqrt{K}\right)\frac{p\|\widehat{\boldsymbol{V}}\boldsymbol{O}-\boldsymbol{V}\|_F}{m_V}+\frac{C_\gamma K^{3/2}(\sqrt{p\log p}+\sqrt{n})}{\sqrt{mn}\lambda_K m_V}\\
&\le \left(\frac{C_\lambda K^{5/2}}{mm_V}+1\right)C_v\frac{K^2(\sqrt{p\log p}+\sqrt{n})}{\sqrt{n}\lambda_K m_V}+\frac{C_\gamma K^{3/2}(\sqrt{p\log p}+\sqrt{n})}{\sqrt{mn}\lambda_K m_V}\\
&\le \frac{C_Y K^{7/2}(\sqrt{p\log p}+\sqrt{n})}{m\sqrt{n}\lambda_K m_V^2},
\end{aligned}$$

where $C_Y \ge C_\lambda C_v+\frac{C_v m m_V}{K^{3/2}}+\frac{C_\gamma\sqrt{m}m_V}{K^2}$.

Let $C_z \ge C_Y(1+\frac{1}{\sqrt{K}})$. By combining all the results above, we obtain that

$$\frac{\|\widehat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_F}{\sqrt{p}} \le \frac{(\sqrt{K}+1)\|\widetilde{\boldsymbol{Y}}-\boldsymbol{Y}\|_F}{\sqrt{p}} \le \frac{(\sqrt{K}+1)\|\widehat{\boldsymbol{Y}}-\boldsymbol{Y}\|_F}{\sqrt{p}} \le \frac{C_z K^4(\sqrt{p\log p}+\sqrt{n})}{m\sqrt{np}\lambda_K {m_V}^2}$$

holds with probability at least $1-4\exp\{-c\log p\}$, which concludes the proof.

Bibliography

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices*, Volume 20. Springer.

Jiang, B., C. Leng, T. Yan, Q. Yao, and X. Yu (2023). A two-way heterogeneity model for dynamic networks. arXiv:2305.12643v2 [stat.ME].

Jiang, B., J. Li, and Q. Yao (2023). Autoregressive networks. *Journal of Machine Learning Research 24*(227), 1–69.

Mastrandrea, R., J. Fournet, and A. Barrat (2015). Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PloS one 10*(9), e0136497.

Yu, Y., T. Wang, and R. J. Samworth (2014, 04). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika 102*(2), 315–323.

Zhang, W., B. Jin, and Z. Bai (2021). Learning block structures in u-statistic-based matrices. *Biometrika 108*(4), 933–946.