Supplementary material of "Mixed membership estimation for social networks"

Jiashun Jin*, Zheng Tracy Ke† and Shengming Luo*
Department of Statistics
Carnegie Mellon University* and Harvard University†

Dec. 2, 2022

Contents

A Identifiability and Regularity Conditions

We prove the identifiability of DCMM and discuss Assumption 4 (where we give sufficient conditions for this assumption to hold).

A.1 The Identifiability of DCMM

The following proposition shows that the DCMM model is identifiable if each community has at least one pure node.

Proposition A.1 (Identifiability)**.** *Consider a DCMM model as in* (2.2)*, where P has unit diagonals. When each community has at least one pure node, the model is identifiable: For eligible (Θ, Π, P) and $(\widetilde{\Theta}, \widetilde{\Pi}, \widetilde{P})$, if $\Theta\Pi P\Pi'\Theta = \widetilde{\Theta}\widetilde{\Pi}\widetilde{P}\widetilde{\Pi}'\widetilde{\Theta}$, then $\Theta = \widetilde{\Theta}$, $\Pi = \widetilde{\Pi}$, and $P = \widetilde{P}$.*

Proof of Proposition A.1: Let $G = K\|\theta\|^{-2}\Pi'\Theta^2\Pi$ be the same as in Section 3. We consider two cases: (1) PG is an irreducible matrix. (2) PG is a reducible matrix.

First, we study Case (1). When PG is irreducible, the matrix R is well-defined (see Lemma 2.1). Additionally, by Lemma 2.1, there exists the Ideal Simplex, which is uniquely determined by the eigenvectors $\xi_1, \xi_2, \ldots, \xi_K$ of Ω. For either (Θ, Π, P) or $(\widetilde{\Theta}, \widetilde{\Pi}, \widetilde{P})$, we have an Ideal Simplex. The two Ideal Simplexes can be different only when there are multiple choices of $\xi_1, \xi_2, \ldots, \xi_K$. By Lemma C.1, the first eigenvalue of Ω has a multiplicity 1, so by basic linear algebra, $[\xi_1, \xi_2, \ldots, \xi_K]$ are uniquely defined up to a rotation matrix of the form

$$\left[\begin{array}{cc} a & 0 \\ 0 & S \end{array}\right], \quad \text{where } a \in \{-1, 1\} \text{ and } S \in \mathbb{R}^{K-1,K-1} \text{ is an orthogonal matrix.}$$

Recalling $R = [\mathrm{diag}(\xi_1)]^{-1}[\xi_2, \xi_3, \ldots, \xi_K]$, it is seen that the property of "a row of R falls on one of the vertices of the Ideal Simplex" is invariant to the above rotation. Therefore, a row of Π equals to the corresponding row of $\widetilde{\Pi}$, as long as one of them is pure.

We now proceed to showing $(\Theta, \Pi, P) = (\widetilde{\Theta}, \widetilde{\Pi}, \widetilde{P})$. By the above argument and that each community has at least one pure node, we assume without loss of generality that for $1 \le k \le K$, the k-th node is a pure node in community k. Comparing the first K rows and the first K columns of $\Theta\Pi P\Pi'\Theta$ with those of $\widetilde{\Theta}\widetilde{\Pi}\widetilde{P}\widetilde{\Pi}'\widetilde{\Theta}'$, it follows that

$$\mathrm{diag}(\theta_1, \ldots, \theta_K) \cdot P \cdot \mathrm{diag}(\theta_1, \ldots, \theta_K) = \mathrm{diag}(\tilde{\theta}_1, \ldots, \tilde{\theta}_K) \cdot \widetilde{P} \cdot \mathrm{diag}(\tilde{\theta}_1, \ldots, \tilde{\theta}_K).$$

As both P and $\widetilde{P}$ have unit diagonal entries, $P = \widetilde{P}$ and $\theta_k = \tilde{\theta}_k$, $1 \le k \le K$.

Moreover, note that $P\Pi'\Theta$ has a full row-rank. Since $\Theta\Pi P\Pi'\Theta = \widetilde{\Theta}\widetilde{\Pi}\widetilde{P}\widetilde{\Pi}'\widetilde{\Theta}$, it is seen that $\Theta\Pi = \widetilde{\Theta}\widetilde{\Pi}\Delta$, where $\Delta = \widetilde{P}\widetilde{\Pi}'\widetilde{\Theta}X'(XX')^{-1}$, with $X = P\Pi'\Theta$ for short. We compare the first K rows of $\Theta\Pi$ and $\widetilde{\Theta}\widetilde{\Pi}\Delta$, recalling that the first K rows are pure and that $\theta_k = \tilde{\theta}_k$ for $1 \le k \le K$. It follows that Δ equals to the $K \times K$ identity matrix. Therefore,

$$\Theta\Pi = \widetilde{\Theta}\widetilde{\Pi}.$$

Since each row of Π or $\tilde{\Pi}$ is a PMF, $\Theta = \tilde{\Theta}$, $\Pi = \tilde{\Pi}$, and the claim follows.

Next, we study Case (2). By Lemma C.1,

$$\Xi = \Theta\Pi B, \qquad \text{for a non-singular matrix } B.$$

Row i of Ξ equals to θ_i times a convex combination of rows of B. It follows that *all rows of Ξ are contained in a simplicial cone with K supporting rays, where a pure row falls on one supporting ray, and a mixed row falls in the interior of the simplicial cone.* Note that Ξ is uniquely defined up to a $K \times K$ orthogonal matrix. The effect of this orthogonal matrix is to simultaneously rotate all rows of Ξ. Such a rotation does not change the property that "a pure row falls on one supporting ray". Therefore, a row of Π equals to the corresponding row of $\tilde{\Pi}$, provided that one of them is pure. The remaining of the proof is similar to that of Case (1). □

Remark *(Comparison with the identifiability of other models).* Compared to other models (e.g., MMSB, DCBM), DCMM has many more parameters (for degree heterogeneity and for mixed memberships). These parameters have more degrees of freedom than those in MMSB or DCBM, and so DCMM requires stronger conditions to be identifiable.

- The assumption that P has unit diagonals is not needed for identifiability of MMSB, but it is necessary for identifiability of DCMM. Consider a DCMM with parameters (Θ, Π, P). Given any $K \times K$ diagonal matrix D with positive diagonals, let

  $$\widetilde{P} = DPD, \quad \tilde{\pi}_i = (D^{-1}\pi_i)/\|D^{-1}\pi_i\|_1, \quad \text{and} \quad \tilde{\theta}_i = \|D^{-1}\pi_i\|_1 \cdot \theta_i.$$

  It is seen that $\Theta\Pi P\Pi'\Theta = \widetilde{\Theta}\widetilde{\Pi}\widetilde{P}\widetilde{\Pi}'\widetilde{\Theta}$. This case will be eliminated by requiring P to have unit diagonals.

- The assumption that P has a full rank is not needed for identifiability of DCBM, but it is necessary for identifiability of DCMM. If the rank of P is $< K$, there exists a nonzero vector $\beta \in \mathbb{R}^K$ such that $P\beta = 0$. As long as there is a π_i such that $\pi_i(k) > 0$ for all k, we can change (π_i, θ_i) to $(\tilde{\pi}_i, \tilde{\theta}_i)$ but keep Ω unchanged. To see this, let

$$\tilde{\pi}_i = (\pi_i + \epsilon\beta)/\|\pi_i + \epsilon\beta\|_1, \quad \text{and} \quad \tilde{\theta}_i = \|\pi_i + \epsilon\beta\|_1 \cdot \theta_i,$$

for a sufficiently small $\epsilon > 0$. Since the two vectors, $\theta_i \cdot P\pi_i$ and $\tilde{\theta}_i \cdot P\tilde{\pi}_i$, are equal, Ω remains unchanged.

A.2 Sufficient conditions for Assumption 4 to hold

We give two propositions showing examples where Assumption 4 is satisfied. Below, for a matrix M, let $\lambda_k(M)$ denote the k-th largest eigenvalue in magnitude.

Proposition A.2. *Consider a DCMM model where $\Omega = \Theta\Pi P\Pi'\Theta$ and $\|P\|_{\max} \leq C$. Write $G = K\|\theta\|^{-2}(\Pi'\Theta^2\Pi)$. Let η_1 be the first (unit-norm) right singular vector of PG. As $n \to \infty$, suppose at least one of the following conditions hold, where $c > 0$ is a constant:*

- $\min_{1\leq k,\ell\leq K} P(k,\ell) \geq c$, *and* $\min_k\{\sum_{i=1}^n \theta_i^2\pi_i(k)\} \geq c\max_k\{\sum_{i=1}^n \theta_i^2\pi_i(k)\}$.
- *K is fixed,* $\min_k G(k,k) \geq c$, *and* $|\lambda_1(PG)| \geq c + |\lambda_2(PG)|$. *For a fixed irreducible matrix P_0,* $\|P - P_0\| \to 0$.
- *K is fixed, and* $|\lambda_1(PG)| \geq c + |\lambda_2(PG)|$. *For a fixed irreducible matrix G_0,* $\|G - G_0\| \to 0$.

Then, we can select the sign of η_1 such that all its entries are strictly positive. Furthermore, $[\max_{1\leq k\leq K}\eta_1(k)]/[\min_{1\leq k\leq K}\eta_1(k)] \leq C$.

Proposition A.3. *Consider a DCMM model where $\Omega = \Theta\Pi P\Pi'\Theta$. We assume that* $\max_{1\leq k\leq K}\{\sum_{\ell=1}^K P(k,\ell)\} \leq C\min_k\{\sum_{\ell=1}^K P(k,\ell)\}$. *Suppose π_i's are i.i.d. generated from Dirichlet(α), where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)'$ satisfies $C_1 \leq \alpha_k \leq C_2$ for two constants $C_2 > C_1 > 0$. Write $G = K\|\theta\|^{-2}(\Pi'\Theta^2\Pi)$. Let η_1 be the first (unit-norm) right singular vector of PG. As $n \to \infty$,* $[\max_{1\leq k\leq K}\eta_1(k)]/[\min_{1\leq k\leq K}\eta_1(k)] \leq C$, *with probability* $1 - o(1)$.

Proof of Propositions A.2-A.3: First, we prove Propositions A.2. Consider the first case. Let $x_k = K\|\theta\|^{-2}\sum_{i=1}^n \theta_i^2\pi_i(k)$. It is seen that $\sum_{k=1}^K x_k = K$. The assumption says that

$\min_k x_k \geq c \max_k x_k$. Therefore, $x_k \asymp 1$ for all k. At the same time, $\sum_{\ell=1}^K G(\ell,k) = K\|\theta\|^{-2}\sum_{\ell=1}^K\sum_{i=1}^n \theta_i^2\pi_i(\ell)\pi_i(k) = x_k$. It follows that

$$\max_k\{\sum_\ell G(\ell,k)\} \asymp \min_k\{\sum_\ell G(\ell,k)\} \asymp 1.$$

For any $1 \leq m,k \leq K$, the (m,k)-th entry of PG equals to $\sum_\ell P(m,\ell)G(\ell,k)$, which is between $c\sum_\ell G(\ell,k)$ and $C\sum_\ell G(\ell,k)$ by the assumption on P. It follows that

$$\max_{k,\ell}\{(PG)(k,\ell)\} \asymp \min_{k,\ell}\{(PG)(k,\ell)\} \asymp 1. \tag{A.1}$$

In particular, PG is a positive matrix. By Perron's theorem [7, Theorem 8.2.8], the first right singular value $\lambda_1(PG)$ is positive and has a multiplicity of 1, and the first eigenvector η_1 is a positive vector. Write $\lambda = \lambda_1(PG)$ for short. By definition,

$$\lambda\eta_1 = (PG)\eta_1.$$

It follows that

$$\max_k \eta_1(k) \leq \frac{\|\eta\|_1}{\lambda}\max_{k,\ell}\{(PG)(k,\ell)\}, \quad \min_k \eta_1(k) \geq \frac{\|\eta\|_1}{\lambda}\min_{k,\ell}\{(PG)(k,\ell)\}. \tag{A.2}$$

Combining (A.1)-(A.2) gives $\max_k \eta_1(k) \asymp \min_k \eta_1(k)$. The claim follows.

Consider the second case. We first state and prove a useful result:

> Let A and B be two nonnegative matrices with strictly positive diagonals. If A is irreducible, then AB is irreducible. (A.3)

The proof uses the definition of primitive matrices (a subclass of irreducible matrices; see [7, Section 8.5]). We aim to show AB is a primitive matrix. By [7, Theorem 8.5.2], it suffices to show that there exists $m \geq 1$, such that $(AB)^m$ is a strictly positive matrix. By the assumption, A is an irreducible matrix with positive diagonals; it follows from [7, Theorem 8.5.4] that A is a primitive matrix. By [7, Theorem 8.5.2] again, there exists $m \geq 1$ such that A^m is a strictly positive matrix. Let $\alpha > 0$ be the minimum diagonal entry of B. Since A and B are nonnegative matrices, each entry of $(AB)^m$ is lower bounded by α^m times the corresponding entry of A^m; hence, $(AB)^m$ is also a strictly positive matrix. It follows that AB is a primitive matrix, which is also an irreducible matrix.

We then show the claim. Note that P and G are both nonnegative matrices with positive entries. Since $\|P - P_0\| \to 0$, the support of P has to be a superset of the support of P_0 for

large enough n; as a result, when P_0 is an irreducible matrix, P has to be an irreducible matrix for sufficiently large n. We apply (A.3) to obtain that PG is an irreducible matrix. It follows that $\lambda_1(PG) > 0$ and it has a multiplicity 1; additionally, the first right eigenvector η_1 is a positive vector.

It remains to show $\max_k \eta_1(k) \asymp \min_k \eta_1(k)$. We prove by contradiction. Write $\eta_1 = \eta_1^{(n)}$, $P = P^{(n)}$ and $G = G^{(n)}$ to emphasize the dependence on n. If the claim is not true, then there is a subsequence $\{n_s\}_{s=1}^{\infty}$ such that

$$\lim_{s\to\infty} \left\{ \frac{\min_k \eta_1^{(n_s)}(k)}{\max_k \eta_1^{(n_s)}(k)} \right\} \to 0. \tag{A.4}$$

Since K is fixed, all the entries of $G^{(n_s)}$ are bounded. It follows that there exists a subsequence of $\{n_s\}_{s=1}^{\infty}$, which we still denote by $\{n_s\}_{s=1}^{\infty}$ for notation convenience, such that $G^{(n_s)} \to G^*$ for a fixed matrix G^*. Therefore,

$$\big\|(PG)^{(n_s)} - P_0 G^*\big\| \to 0, \qquad \text{as } s \to \infty. \tag{A.5}$$

Let η_1^* be the first right eigenvector of $P_0 G^*$. Since $|\lambda_1(PG)| \geq c + |\lambda_2(PG)|$, by the sin-theta theorem (e.g., see Lemma D.3), it follows from (A.5) that

$$\|\eta_1^{(n_s)} - \eta_1^*\| \to 0, \qquad \text{as } s \to \infty. \tag{A.6}$$

We now derive a contradiction from (A.4)-(A.6). On the one hand, combining (A.5)-(A.6) and noting that η_1^* is a fixed vector, we conclude that the minimum entry of η_1^* is zero. On the other hand, the assumption of $\min_k G(k,k) \geq c$ ensures that G^* has strictly positive diagonals. We apply (A.3) to conclude that $P_0 G^*$ is a fixed irreducible matrix. By Perron's theorem, η_1^* should be a strictly positive vector. This yields a contradiction.

Consider the third case. The proof is similar to that of the second case, except that we switch the roles of P and G. Note that we do not need additional conditions on the diagonals of P, since P always has unit diagonals.

Next, we prove Propositions A.3. By (A.1) and (A.2), we only need to show that

$$\max_{k,\ell}\big\{(PG)(k,\ell)\big\} \asymp \min_{k,\ell}\big\{(PG)(k,\ell)\big\}.$$

Since the maximum row sum and minimum row sum of P are at the same order, it suffices to show that the maximum and minimum entries of G are at the same order. Let $G_0 = \mathbb{E}_{\pi\sim\text{Dirichlet}(\alpha)}[\pi\pi']$. As $n \to \infty$, it is easy to show that $\|G - G_0\|_F = o(1)$. Therefore, we

only need to show that the maximum and minimum entries of G_0 are at the same order. By direct calculations,

$$\begin{aligned} G_0 &= (\mathbb{E}[\pi])(\mathbb{E}[\pi])' + \mathrm{Cov}(\Pi) \\ &= \frac{1}{\|\alpha\|_1^2}\alpha\alpha' + \frac{1}{1+\|\alpha\|_1}\Big[\frac{1}{\|\alpha\|_1}\mathrm{diag}(\alpha) - \frac{1}{\|\alpha\|_1^2}\alpha\alpha'\Big] \\ &= \frac{1}{\|\alpha\|_1(1+\|\alpha\|_1)}[\mathrm{diag}(\alpha) + \alpha\alpha']. \end{aligned}$$

Since all entries of α are bounded above and below by constants, it is easy to see that the maximum and minimum entries of G_0 are at the same order. This completes the proof. □

B Faster Rates of Mixed-SCORE (Setting 2)

In Section 3.3, we discuss Mixed-SCORE with each specific VH approach in Table 1. For Mixed-SCORE-SVS and Mixed-SCORE-SVS*, we consider two settings where they enjoy faster rates than the generic Mixed-SCORE algorithm. Due to space limit, we only present Setting 1 in Section 3.3. We now present Setting 1.

Setting 2. Let $\mathcal{N}_k$ be the set of pure nodes of community k, $1 \le k \le K$, and let $\mathcal{M}$ be the set of all mixed nodes. Suppose there are constants $c_1, c_2 \in (0,1)$ such that $\min_{1\le k\le K}|\mathcal{N}_k| \ge c_1 n$ and $\min_{1\le k\le K}\sum_{i\in\mathcal{N}_k}\theta^2(i) \ge c_2\|\theta\|^2$. Furthermore, for a fixed integer $L_0 \ge 1$, we assume there is a partition of $\mathcal{M}$, $\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_{L_0}$, a set of PMF's $\gamma_1, \cdots, \gamma_{L_0}$, and constants $c_3, c_4 > 0$ such that (e_k: k-th standard basis vector of $\mathbb{R}^K$) $\big\{\min_{1\le j\ne\ell\le L_0}\|\gamma_j - \gamma_\ell\|,\ \min_{1\le\ell\le L_0, 1\le k\le K}\|\gamma_\ell - e_k\|\big\} \ge c_3$, and for each $1 \le \ell \le L_0$ (note: err_n is the same as that in (3.10)), $|\mathcal{M}_\ell| \ge c_4|\mathcal{M}| \ge n\beta_n^{-2}err_n^2$ and $\max_{i\in\mathcal{M}_\ell}\|\pi_i - \gamma_\ell\| \le 1/\log(n)$.

In this setting, π_i's form several *loose clusters*, where the π_i's in the same cluster are within a distance of $O(\frac{1}{\log(n)})$ from each other. Since $\frac{1}{\log(n)}$ is much larger than the order of noise, $\max_{1\le i\le n}\|H\hat{r}_i - r_i\|$, the assumed clustering structure is indeed "loose". [1]

Theorem B.1. *Consider the DCMM model where Assumptions 1-4 hold and π_i's are from Setting 2. Let H be as in Theorem 3.1. Suppose we apply SVS or SVS* to rows of $\hat{R}$ with*

$$L = \hat{L}_n(A) := \min\big\{L \ge K+1 : \epsilon_L(\hat{R}) < \epsilon_{L-1}(\hat{R})/\log(\log(n))\big\}.$$

[1] In fact, by a slight modification of the proof, we can replace $(1/\log(n))$ in Setting 2 by any $o(1)$ term, or an appropriately small constant $\tilde{c}_3 > 0$ (this constant $\tilde{c}_3$ will depend on the constants in Setting 2 in a quite complicated way). We present the current version for its convenience.

With probability $1 - o(n^{-3})$,

$$\max_{1\le k\le K} \|H\hat{v}_k - v_k\| \le C\sqrt{n^{-1}\sum_{i=1}^{n} \|H\hat{r}_i - r_i\|^2}.$$

Moreover, for Mixed-SCORE-SVS or Mixed-SCORE-SVS,*

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \|\hat{\pi}_i - \pi_i\|^2\Big] \le CK^3\beta_n^{-2}(err_n^*)^2 + o(n^{-2}).$$

C The Oracle Case and Ideal Mixed-SCORE

We consider the oracle case where Ω is observed. In Section C.1, we state a useful lemma, which is the key for analysis of the oracle case. In Section C.2, we prove Lemmas 2.1 in the paper, which inspire Ideal Mixed-SCORE. In Section C.3, we prove Lemma 2.2, which is about recovering (P, θ) from Π. In Section C.4, we study eigenvalues and eigenvectors of Ω and the matrix R; these results are useful for the proofs in Sections D-F.

C.1 A useful lemma and its proof

Let $G = K\|\theta\|^{-2}(\Pi'\Theta^2\Pi)$ is as in Section 3. Let $\lambda_1, \lambda_2, \ldots, \lambda_K$ be the nonzero eigenvalues of Ω, sorted in the descending order of magnitudes. Let $\xi_1, \xi_2, \ldots, \xi_K$ be the corresponding eigenvectors. We have the following lemma:

Lemma C.1. *Consider the DCMM model, where PG is an irreducible matrix and there is at least one pure node for each community. The following statements are true:*

- *There is a non-singular matrix $B \in \mathbb{R}^{K,K}$ such that $\Theta\Pi B = \Xi$, and B is unique once Ξ is chosen.*
- *For $1 \le k \le K$, denote by a_k the kth largest (in magnitude) eigenvalue of PG. Then, a_k's are real, and the nonzero eigenvalues of Ω are $\lambda_k = (K^{-1}\|\theta\|^2)a_k$, $1 \le k \le K$.*
- *For $1 \le k \le K$, denote by b_k the kth column of B. Then, b_k is a (right) eigenvector of PG associated with a_k.*
- *$\lambda_1 > 0$ and it has a multiplicity 1 (so ξ_1 is uniquely determined up to a factor of ± 1).*
- *ξ_1 can be chosen such that all of its entries are positive. For this choice of ξ_1, all the entries of the associated b_1 are also positive.*

Proof of Lemma C.1: Consider the first claim. Denote by $Span(M)$ the column space of any matrix M. It suffices to show that $Span(\Theta\Pi) = Span(\Xi)$. Then, since $\xi_1, \cdots, \xi_K$ form an orthonormal basis of this subspace, there is a unique, non-singular matrix $\tilde{B}$ such that $\Theta\Pi = \Xi\tilde{B}$. We then take $B = \tilde{B}^{-1}$.

We now show $Span(\Theta\Pi) = Span(\Xi)$. By the assumption that there is at least one pure node in each community, we can find K rows of Π such that they form a $K \times K$ identity matrix. So Π has a rank K. Since Θ and P are both non-singular matrices, Ω also has a rank K. By definition, $\Omega\xi_k = \lambda_k\xi_k$, for $1 \leq k \leq K$. It follows that

$$\Theta\Pi(P\Pi'\Theta\xi_k) = \lambda_k\xi_k.$$

Hence, each ξ_k is in the column space of $\Theta\Pi$. This means the column space of Ξ is contained in the column space of $\Theta\Pi$. Since both matrices have a rank K, the two column spaces are the same.

Consider the second claim. Note that P is symmetric and G is positive definite. Let $G^{1/2}$ be the unique square root of G. For any matrices $A \in \mathbb{R}^{m,n}$ and $B \in \mathbb{R}^{n,m}$, if $m \geq n$, then the nonzero eigenvalues of AB are the same as the nonzero eigenvalues of BA [7, Theorem 1.3.22]. As a result, eigenvalues of PG are the same as eigenvalues of the symmetric matrix $G^{1/2}PG^{1/2}$. It implies that $a_1, a_2, \ldots, a_K$ are real.

Furthermore, the nonzero eigenvalues of $\Omega = (\Theta\Pi)(P\Pi'\Theta)$ are the same as the nonzero eigenvalues of $(P\Pi'\Theta)(\Theta\Pi) = (K^{-1}\|\theta\|^2)(PG)$. Hence, the nonzero eigenvalues of Ω are $(K^{-1}\|\theta\|^2)a_1, (K^{-1}\|\theta\|^2)a_2, \ldots, (K^{-1}\|\theta\|^2)a_K$.

Consider the third claim. Write $\tilde{G} \equiv K^{-1}\|\theta\|^2 G = \Pi'\Theta^2\Pi$. Note that $\Omega\xi_k = \lambda_k\xi_k$ and $\xi_k = \Theta\Pi b_k$. Hence, $(\Theta\Pi P\Pi'\Theta)(\Theta\Pi b_k) = \lambda_k(\Theta\Pi b_k)$. Multiplying both sides by $\Pi'\Theta$ from the left, we have

$$\tilde{G}P\tilde{G}b_k = \lambda_k\tilde{G}b_k.$$

Since $\tilde{G}$ is non-singular, $P\tilde{G}b_k = \lambda_k b_k$. Plugging in $\tilde{G} = (K^{-1}\|\theta\|^2)G$ and $\lambda_k = (K^{-1}\|\theta\|^2)a_k$, we obtain $PGb_k = a_k b_k$. This shows that b_k is a (right) eigenvector of PG associated with a_k. Additionally, since η_1 is the first unit-norm right singular vector of PG, it yields that $\eta_1 = b_1/\|b_1\|$.

Consider the fourth claim. Since $\lambda_1 = (K^{-1}\|\theta\|^2)a_1$, it suffices to show that $a_1 > 0$ and that it has a multiplicity 1. This follows immediately from the Perron-Frobenius theorem [7, Theorem 8.4.4] and the assumption that PG is an irreducible matrix.

Consider the last claim. Note that b_1 is the eigenvalue of PG associated with a_1. Since a_1 has a multiplicity 1, $b_1/\|b_1\|$ is unique up to a factor of ± 1 (depending on the choice of ξ_1). By Perron-Frobenius theorem again, $b_1/\|b_1\|$ can be chosen such that all the entries are positive. Recalling that $\Xi = \Theta\Pi B$, we immediately have $\xi_1 = \Theta\Pi b_1$. Since $\Theta\Pi$ is a nonnegative matrix with positive row sums and b_1 has strictly positive entries, all the entries of ξ_1 are also positive. □

C.2 Proofs of Lemma 2.1

Consider the first claim. We have shown in Lemma C.1 that

$$\Xi = \Theta\Pi B, \qquad \text{for a non-singular matrix } B = [b_1, \ldots, b_K] \in \mathbb{R}^{K,K}.$$

Furthermore, by the last two bullet points of Lemma C.1, if we pick the sign of ξ_1 such that $\sum_{i=1}^n \xi_1(i) > 0$, then ξ_1 and b_1 are uniquely determined and have strictly positive entries. This proves the first claim.

Consider the other two claims. We first show there are K affinely independent vectors $v_1, v_2, \ldots, v_K$ such that each r_i is a convex combination of them. For $1 \le k \le K$, define $v_k \in \mathbb{R}^{K-1}$ by

$$v_k(\ell) = b_{\ell+1}(k)/b_1(k), \qquad 1 \le \ell \le K-1. \tag{C.7}$$

The vectors $v_1, v_2, \ldots, v_K$ are affinely independent, if and only if the following matrix

$$Q = \begin{pmatrix} 1 & \cdots & 1 \\ v_1 & \cdots & v_K \end{pmatrix}$$

is non-singular. By (C.7), we observe that $Q' = \mathrm{diag}(b_1)B$. Since B is non-singular and b_1 is a positive vector, Q has to be a non-singular matrix. This proves that $v_1, v_2, \ldots, v_K$ are affinely independent. We then study each r_i. Since $\Xi = \Theta\Pi B$, we have

$$\xi_\ell(i) = \theta(i) \sum_{k=1}^K \pi_i(k) b_\ell(k) = \theta(i) \|b_\ell \circ \pi_i\|_1, \qquad 1 \le \ell \le K.$$

By definition of R, $r_i(\ell) = \xi_{\ell+1}(i)/\xi_1(i)$. It follows that

$$r_i(\ell) = \frac{\theta(i)\sum_{k=1}^K \pi_i(k) b_{\ell+1}(k)}{\theta(i)\|b_1 \circ \pi_i\|_1} = \sum_{k=1}^K \frac{b_1(k)\pi_i(k)}{\|b_1 \circ \pi_i\|_1} \cdot \frac{b_{\ell+1}(k)}{b_1(k)} = \sum_{k=1}^K w_i(k) v_k(\ell).$$

This proves that $r_i = \sum_{k=1}^K w_i(k) v_k$, with $w_i = (b_1 \circ \pi_i)/\|b_1 \circ \pi_i\|_1$. Since b_1 is a positive vector and π_i is a nonnegative vector, we have that w_i is a nonnegative vector and $\|w_i\|_1 = 1$. Therefore, r_i is a convex combination of $v_1, v_2, \ldots, v_K$.

We now show the second claim. Each r_i is in the convex hull of $v_1, v_2, \ldots, v_K$. Since these K vectors are affinely independent, their convex hull is a non-degenerate simplex with K vertices. Recall that $w_i = (b_1 \circ \pi_i)/\|b_1 \circ \pi_i\|_1$, where b_1 is a strictly positive vector. Therefore, for each $1 \le k \le K$, node i is a pure node of community k if and only if $\pi_i = e_k$, which happens if and only if $w_i = e_k$; and $w_i = e_k$ means r_i is located at the vertex v_k.

We then show the last claim, which is the formula for b_1. Write $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_K)$. Then, $\Omega = \Xi\Lambda\Xi'$. First, plugging in $\Xi = \Theta\Pi B$, we find that $\Omega = \Theta\Pi(B\Lambda B')\Pi'\Theta$. Multiplying both sides by $\Pi'\Theta$ from the left and $\Theta\Pi$ from the right, we have $\Pi'\Theta\Omega\Theta\Pi = \tilde{G}(B\Lambda B')\tilde{G}$, where $\tilde{G} = \Pi'\Theta^2\Pi$ is a non-singular matrix. Second, since $\Omega = \Theta\Pi P\Pi'\Theta'$, we have $\Pi'\Theta\Omega\Theta\Pi = \tilde{G}P\tilde{G}$. Combining the above gives

$$\tilde{G}P\tilde{G} = \tilde{G}(B\Lambda B')\tilde{G} \quad \Longrightarrow \quad P = B\Lambda B'. \tag{C.8}$$

It follows that

$$1 = P(k,k) = \sum_{\ell=1}^{K} \lambda_\ell b_\ell^2(k) = b_1^2(k)[\lambda_1 + \sum_{\ell=2}^{K} \lambda_\ell^2 v_k(\ell - 1)].$$

Noting that $b_1(k)$ is positive, the above gives the formula for computing b_1. □

C.3 Proof of Lemma 2.2

Write $V = [v_1, v_2, \ldots, v_K]$. By (C.7), $B = \mathrm{diag}(b_1)[1, V']$. Moreover, by (C.8), $P = B\Lambda B'$. Combining them gives the formula of recovering P. Note that $\Xi = \Theta\Pi B$. It follows that $\xi_1(i) = \theta(i) \cdot \pi_i' b_1$. This gives the formula of recovering θ. □

C.4 Spectral analysis of Ω

First, we study the leading eigenvalues of Ω. Let $\lambda_1, \ldots, \lambda_K$ be the nonzero eigenvalues of Ω, listed in the descending order in magnitude. The following lemma is proved in Section C.4.1:

Lemma C.2. *Under conditions of Theorem 3.1, the following statements are true:*

- $C^{-1}K^{-1}\|\theta\|^2 \le \lambda_1 \le C\|\theta\|^2$. *If* $\beta_n = o(1)$, *then* $\lambda_1 \asymp \|\theta\|^2$.
- $\lambda_1 - |\lambda_2| \asymp \lambda_1$.
- $|\lambda_k| \asymp \beta_n K^{-1}\|\theta\|^2$, *for* $2 \le k \le K$.

Next, we study the leading eigenvectors of Ω. For $1 \le k \le K$, let ξ_k be the eigenvector of Ω associated with λ_k. Write $\Xi_0 = [\xi_2, \xi_3, \cdots, \xi_K] \in \mathbb{R}^{n,K-1}$, and let $\Xi'_{0,i}$ be its i-th row, $1 \le i \le n$. The following lemma is proved in Section C.4.2:

Lemma C.3. *Under conditions of Theorem 3.1, the following statements are true:*

- *If we choose the sign of ξ_1 such that $\sum_{i=1}^n \xi_1(i) > 0$, then the entries of ξ_1 are positive satisfying $C^{-1}\theta(i)/\|\theta\| \le \xi_1(i) \le C\theta(i)/\|\theta\|$, $1 \le i \le n$.*
- *$\|\Xi_{0,i}\| \le C\sqrt{K}\theta(i)/\|\theta\|$, $1 \le i \le n$.*

Last, we study the entry-wise ratio matrix R. Recall that w_i is the barycentric coordinate vector of r_i in the Ideal Simplex. The following lemma is proved in Section C.4.3.

Lemma C.4. *Under conditions of Theorem 3.1, the following statements are true:*

- *The vertices of the Ideal Simplex satisfy that $\max_{1\le k\le K} \|v_k\| \le C\sqrt{K}$ and $\min_{k\neq\ell} \|v_k - v_\ell\| \ge C^{-1}\sqrt{K}$.*
- *$C^{-1}\|\pi_i - \pi_j\|_1 \le \|w_i - w_j\|_1 \le C\|\pi_i - \pi_j\|_1$, for all $1 \le i, j \le n$.*
- *$C^{-1}\sqrt{K}\|w_i - w_j\| \le \|r_i - r_j\| \le C\sqrt{K}\|w_i - w_j\|$, for all $1 \le i, j \le n$.*

Lemmas C.2-C.4 are useful for proofs in Sections D-F. Below, we prove these lemmas.

C.4.1 Proof of Lemma C.2

By Lemma C.1, all nonzero eigenvalues of Ω are $(K^{-1}\|\theta\|^2)a_1, \ldots, (K^{-1}\|\theta\|^2)a_K$, where a_k is the k-th largest eigenvalue (in magnitude) of PG. By Assumption 3,

$$a_1 - |a_2| \ge C^{-1}a_1, \qquad C^{-1}\beta_n \le |a_K| \le |a_2| \le C\beta_n.$$

The second and third claims follow immediately.

It remains to show the first claim, which reduces to studying a_1. For any two matrices A and B, the nonzero eigenvalues of AB are the same as the nonzero eigenvalues of BA. Hence,

$$a_1 = \lambda_1(PG) = \lambda_1(G^{1/2}PG^{1/2}) = \max_{x\neq 0} \frac{x'G^{1/2}PG^{1/2}x}{\|x\|^2}.$$

By Assumption 2, $\|G\| \le C$ and $\|G^{-1}\| \le C$. It is easy to see that $a_1 \le C\lambda_1(P)$. Additionally, $\lambda_1(P) = \max_{y\neq 0} \frac{y'Py}{\|y\|^2} = \max_{x\neq 0} \frac{x'G^{1/2}PG^{1/2}x}{\|G^{1/2}x\|^2}$. Since $\|G^{1/2}x\|^2 = x'Gx \ge C^{-1}\|x\|^2$, it follows that $\lambda_1(P) \le \max_{x\neq 0} \frac{x'G^{1/2}PG^{1/2}x}{C^{-1}\|x\|^2} \le C\lambda_1(PG)$. Together,

$$C^{-1}\lambda_1(P) \le \lambda_1(PG) \le C\lambda_1(P).$$

Note that $\lambda_1(P) \leq K\|P\|_{\max} = O(K)$ and $\lambda_1(P) \geq P(k,k) \geq 1$. We plug them into the above inequality to get

$$C^{-1} \leq a_1 \leq CK. \tag{C.9}$$

This inequality holds in all cases. If, additionally, $\beta_n \to 0$ as $n \to \infty$, we can get a stronger result. Note that P and G are nonnegative matrices, and for each $1 \leq k \leq K$, $P(k,k) = 1$ and $G(k,k) \geq \lambda_{\min}(G) \geq C^{-1}$. It follows that $(PG)(k,k) \geq P(k,k)G(k,k) \geq C^{-1}$. We thus have

$$\mathrm{trace}(PG) \geq C^{-1}K.$$

At the same time, $\mathrm{trace}(PG) = a_1 + \sum_{k=2}^K a_2 = a_1 + O(K\beta_n) = a_1 + o(K)$. It follows that

$$C^{-1}K \leq a_1 \leq CK, \qquad \text{if } \beta_n = o(1). \tag{C.10}$$

The first claim follows from (C.9)-(C.10) and the equality $\lambda_1 = (K^{-1}\|\theta\|^2)a_1$. □

C.4.2 Proof of Lemma C.3

Consider the first claim. From the last item of Lemma C.1, we can choose the sign of ξ_1 such that both (ξ_1, b_1) have strictly positive entries, where this choice of sign corresponds to $\sum_{i=1}^n \xi_1(i) > 0$. Note that $\Xi = \Theta\Pi B$, which implies $\xi_1(i) = \theta(i)\sum_{k=1}^K \pi_i(k)b_1(k)$. Since each π_i is a PMF (a nonnegative vector whose entries sum to 1),

$$\theta(i)\min_{1\leq k\leq K} b_1(k) \leq \xi_1(i) \leq \theta(i)\max_{1\leq k\leq K} b_1(k), \qquad 1 \leq i \leq n.$$

Hence, to show the claim, it suffices to show that

$$C^{-1}\|\theta\|^{-1} \leq b_1(k) \leq C\|\theta\|^{-1}, \qquad \text{for all } 1 \leq k \leq K. \tag{C.11}$$

Write $\tilde{G} = K^{-1}\|\theta\|^2 G = \Pi'\Theta^2\Pi$. Since $\Xi = \Theta\Pi B$ and $X'X = I_K$, we have $B'\Pi'\Theta^2\Pi B = I_K$, or equivalently, $B'\tilde{G}B = I_K$. Multiplying both sides by B from the left and B' from the right, we obtain $BB'\tilde{G}BB' = BB'$. Since BB' is non-singular, it implies

$$BB' = \tilde{G}^{-1} = K\|\theta\|^{-2}G^{-1}. \tag{C.12}$$

We note that $BB' = \sum_{k=1}^K b_k b_k' \succeq b_1 b_1'$. So, $\|b_1\|^2 \leq \|B\|^2 \leq K\|\theta\|^{-2}\|G^{-1}\|$. By our assumption of $\|G^{-1}\| \leq C$. It follows that

$$\|b_1\| \leq C\|\theta\|^{-1}\sqrt{K}.$$

At the same time, $1 = \|\xi_1\|^2 = \|\Theta\Pi b_1\|^2$. By direct calculations, $\|\Theta\Pi b_1\|^2 = \sum_i \theta_i^2(\pi_i' b_1)^2 \leq \sum_i \theta_i^2 \|b_1\|_\infty^2 \leq \|\theta\|^2\|b_1\|_\infty^2$. It follows that

$$\|b_1\|_\infty \geq C^{-1}\|\theta\|^{-1}.$$

In Lemma C.1, we have seen that b_1 is the first right singular vector of PG. Hence, $b_1 \propto \eta_1$, where η_1 is the same as in Assumption 4. By Assumption 4, all the entries of η_1 are at the same order. Hence, all the entries of b_1 are at the same order. It follows that

$$b_1(k) \asymp \|b_1\|_\infty \asymp (1/\sqrt{K})\|b_1\|.$$

This gives (C.11) and completes the proof of the first claim.

Consider the second claim. Since $\Xi = \Theta\Pi B$, for $1 \leq i \leq n$,

$$\|\Xi_{0,i}\| \leq \theta(i)\|B\pi_i\| \leq C\theta(i)\sqrt{\lambda_{\max}(B'B)} \leq C\sqrt{K}\|\theta\|^{-1}\theta(i),$$

where the last inequality is due to (C.12) and and the condition $\|G^{-1}\| \leq C$. □

C.4.3 Proof of Lemma C.4

First, we prove the claim about the connection between $\|w_i - w_j\|_1$ and $\|\pi_i - \pi_j\|_1$. Let $\mathcal{S}_0 \subset \mathbb{R}^K$ be the standard simplex whose vertices are $e_1, e_2, \ldots, e_K$. Define a mapping

$$T_1 : \mathcal{S}_0 \to \mathcal{S}_0, \qquad \text{where} \quad T_1(x) = \frac{x \circ b_1}{\|x \circ b_1\|_1}.$$

Then, $w_i = T_1(\pi_i)$, for $1 \leq i \leq n$. To show the claim, it suffices to show that T_1 and T_1^{-1} are both Lipschitz with respect to the ℓ^1-norm, i.e., for any $x, y \in \mathcal{S}_0$,

$$C^{-1}\|x - y\|_1 \leq \|T_1(x) - T_1(y)\|_1 \leq C\|x - y\|_1. \tag{C.13}$$

We now show (C.13). Fixing any $x, y \in \mathcal{S}_0$, write $x^* = T_1(x)$ and $y^* = T_1(y)$. By definition, $x^*(k) = x(k)b_1(k)/\|x \circ b_1\|_1$ and $y^*(k) = y(k)b_1(k)/\|y \circ b_1\|_1$. We write

$$\begin{aligned} x^*(k) - y^*(k) &= \frac{[x(k) - y(k)]b_1(k)}{\|x \circ b_1\|_1} + y(k)b_1(k)\left[\frac{1}{\|x \circ b_1\|_1} - \frac{1}{\|y \circ b_1\|_1}\right] \\ &= \frac{b_1(k)}{\|x \circ b_1\|_1}[x(k) - y(k)] + \frac{y^*(k)}{\|x \circ b_1\|_1}\left(\|y \circ b_1\|_1 - \|x \circ b_1\|_1\right). \end{aligned}$$

First, by (C.11), $b_1(k) \asymp \|\theta\|^{-1}$ for all $1 \leq k \leq K$. It follows that $|b_1(k)| \leq C\|\theta\|^{-1}$ and $\|x \circ b_1\|_1 \geq \|x\|_1 \cdot C^{-1}\|\theta\|^{-1} \geq C^{-1}\|\theta\|^{-1}$. Hence,

$$\frac{b_1(k)}{\|x \circ b_1\|_1}|x(k) - y(k)| \leq C|x(k) - y(k)|.$$

Second, by the triangle inequality, $|\|y \circ b_1\|_1 - \|x \circ b_1\|_1| \leq \|(y-x) \circ b_1\|_1$. Moreover, since $b_1(k) \asymp \|\theta\|^{-1}$ for all k, we have $\|(y-x) \circ b_1\|_1 \leq C\|\theta\|^{-1}\|x-y\|_1$ and $\|x \circ b_1\|_1 \geq C^{-1}\|\theta\|^{-1}$. It follows that

$$\frac{y^*(k)}{\|x \circ b_1\|_1} |\|y \circ b_1\|_1 - \|x \circ b_1\|_1| \leq Cy^*(k) \cdot \|x-y\|_1.$$

Combining the above gives

$$|x^*(k) - y^*(k)| \leq C|x(k) - y(k)| + Cy^*(k) \cdot \|x-y\|_1.$$

We sum over k on both sides and note that $\sum_k y^*(k) = 1$. It gives

$$\|x^* - y^*\|_1 \leq C\|x-y\|_1.$$

This shows that T_1 is Lipschitz with respect to the ℓ^1-norm. We then consider T_1^{-1}. Define $\tilde{b}_1 \in \mathbb{R}^K$ by $\tilde{b}_1(k) = 1/b_1(k)$, $1 \leq k \leq K$. We can rewrite

$$T_1^{-1}(x) = \frac{x \circ \tilde{b}_1}{\|x \circ \tilde{b}_1\|_1}.$$

T_1^{-1} has a similar form as T_1, where the vector $\tilde{b}_1$ satisfies that $\tilde{b}_1(k) \asymp \|\theta\|$ for all k. Hence, we can similarly prove that T_1^{-1} is Lipschitz with respect to the ℓ^1-norm. This proves (C.13).

Next, we prove the claim about the connection between $\|r_i - r_j\|$ and $\|w_i - w_j\|$. Let $\mathcal{S}_0$ be the same as before, and let $\mathcal{S}^{ideal} = \mathcal{S}^{ideal}(v_1, v_2, \ldots, v_K) \subset \mathbb{R}^{K-1}$ denote the Ideal Simplex. Let $B = [b_1, b_2, \ldots, b_K]$ be as in Lemma C.1. Define a mapping:

$$T_2 : \mathcal{S}_0 \to \mathcal{S}^{ideal}, \qquad \text{where} \quad \begin{pmatrix} 1 \\ T_2(x) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & \cdots & 1 \\ v_1 & \cdots & v_K \end{pmatrix}}_{\equiv Q} x.$$

By Lemma 2.1, $r_i = T_2(w_i)$, for all $1 \leq i \leq n$. To show the claim, it suffices to show that T_2 and T_2^{-1} are both Lipschitz with respect to the ℓ^2-norm, whose Lipschitz constants are $\sqrt{K}$ and $1/\sqrt{K}$, respectively. In other words, we want to prove, for any $x, y \in \mathcal{S}_0$,

$$C^{-1}\sqrt{K}\|x-y\| \leq \|T_2(x) - T_2(y)\| \leq C\sqrt{K}\|x-y\|. \tag{C.14}$$

We now show (C.14). Note that $Qx = (1_K'x, T_2(x))'$. Since $1_K'x = 1_K'y = 1$, we have

$$\|T_2(x) - T_2(y)\|^2 = \|Qx - Qy\|^2 = (x-y)'Q'Q(x-y).$$

It suffices to show that

$$\|Q\| \leq C\sqrt{K}, \qquad \text{and} \qquad \|Q^{-1}\| \leq C/\sqrt{K}. \tag{C.15}$$

By (C.7), we can re-write

$$Q' = [\mathrm{diag}(b_1)]^{-1}B.$$

By (C.11), $b_1(k) \asymp \|\theta\|^{-1}$ for all k. By (C.12), $BB' = K\|\theta\|^{-2}G^{-1}$; we note that by Assumption 2, $\|G\| \leq C$ and $\|G^{-1}\| \leq C$; it follows that $\|B\| \leq C\sqrt{K}\|\theta\|^{-1}$ and $\|B^{-1}\| \leq C\|\theta\|/\sqrt{K}$. Combining them gives (C.15). Then, (C.14) follows.

Last, we prove the claims about the Ideal Simplex (IS). Let $e_1, e_2, \ldots, e_K$ be the standard basis vectors of $\mathbb{R}^K$. It is seen that $v_k = T_2(e_k)$, $1 \leq k \leq K$. By (C.14), for $k \neq \ell$,

$$\|v_k - v_\ell\| \asymp \sqrt{K}\|e_k - e_\ell\| \asymp \sqrt{K}.$$

By definition of Q and (C.15), for all $1 \leq k \leq K$,

$$\|v_k\| \leq \|Q\| = O(\sqrt{K}).$$

The above give the desired claims. □

D Spectral Analysis of A and Large-deviation Bounds for $\hat{R}$

We conduct spectral analysis for A. In Section D.1, we give the large deviation bounds for eigenvalues of A. In Sections D.2, we study the eigenvectors of A and state a key technical lemma. In Section D.3, we prove Theorem 3.1 in the paper, which is about the row-wise large deviation bound for $\hat{R}$. In Section D.4, we give the ℓ^2-norm large deviation bound for $\hat{R}$. In Section D.4, we give a useful property of the rotation matrix H.

D.1 The eigenvalues of A

Let $\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_K$ be the K largest eigenvalues of A (in magnitude), sorted descendingly in magnitude.

Lemma D.1. *Under conditions of Theorem 3.1, with probability* $1-o(n^{-3})$, $\max_{1\leq k\leq K}|\hat{\lambda}_k - \lambda_k| \leq C\sqrt{\theta_{\max}\|\theta\|_1}$.

Proof of Lemma D.1: By Weyl's inequality, $\max_{1\le k\le K}|\hat{\lambda}_k-\lambda_k|\le\|A-\Omega\|$. To show the claim, it suffices to show that with probability $1-o(n^{-3})$,

$$\|A-\Omega\|\le C\sqrt{\theta_{\max}\|\theta\|_1}. \tag{D.16}$$

The following inequality is useful:

$$(\theta_{\max}\|\theta\|_1)/\log(n)\to\infty. \tag{D.17}$$

To see why (D.17) is true, we rewrite $err_n=(\theta_{\max}/\theta_{\min})\|\theta\|^{-2}\sqrt{\theta_{\max}\|\theta\|_1\log(n)}$. Since $\theta_{\max}\ge\theta_{\min}$ and $\theta_{\max}\|\theta\|_1\ge\|\theta\|^2$, we immediately have $err_n\ge\|\theta\|^{-1}\sqrt{\log(n)}$. Therefore, the assumption $err_n\to 0$ implies that $\|\theta\|^2/\log(n)\to\infty$. Then (D.17) is also true because $\theta_{\max}\|\theta\|_1\ge\|\theta\|^2$.

We now prove (D.16). Write

$$A-\Omega=W+\mathrm{diag}(\Omega),\qquad\text{where }\ W\equiv A-E[A].$$

Note that $\pi_i'P\pi_j=\sum_{k,\ell}\pi_i(k)\pi_j(\ell)P_{k\ell}\le\|P\|_{\max}\|\pi_i\|_1\|\pi_j\|_1\le C$. It follows that

$$\Omega(i,j)\le C\theta(i)\theta(j).$$

Note that $\Omega(i,i)=\theta^2(i)(\pi_i'P\pi_i)\le C\theta^2(i)$. As a result,

$$\|\mathrm{diag}(\Omega)\|\le C\theta_{\max}^2\le C\sqrt{\theta_{\max}\|\theta\|_1}, \tag{D.18}$$

where the last inequality follows from (D.17) and $\theta_{\max}^2\le C\ll\sqrt{\log(n)}$. We then apply the non-asymptotic bounds for random matrices in [2] to bound $\|W\|$. By Corollary 3.12 and Remark 3.13 of [2], for the $n\times n$ symmetric matrix W whose upper triangle contains independent entries, for any $\epsilon>0$, there exists a universal constant $\tilde{c}_\epsilon>0$ such that for every $t\ge 0$,

$$\mathbb{P}\big(\|W\|>(1+\epsilon)2\sqrt{2}\tilde{\sigma}+t\big)\le ne^{-t^2/(\tilde{c}\tilde{\sigma}_*^2)}, \tag{D.19}$$

where

$$\tilde{\sigma}=\max_i\sqrt{\sum_j\mathbb{E}[W(i,j)^2]},\qquad\tilde{\sigma}_*=\max_{ij}\|W(i,j)\|_\infty.$$

We fix $\epsilon=1/2$ in (D.19) and write $\tilde{c}=\tilde{c}_\epsilon$ for short. For $t=2\tilde{\sigma}_*\sqrt{\tilde{c}\log(n)}$, it follows from (D.19) that with probability $1-o(n^{-3})$,

$$\|W\|\le 3\sqrt{2}\max_i\sqrt{\sum_j\mathbb{E}[W(i,j)^2]}+C\tilde{\sigma}_*\sqrt{\log(n)}.$$

Note that $\tilde{\sigma}_* \leq 1$ and $\max_i\{\sum_j \mathbb{E}[W(i,j)^2]\} \leq \max_i\{\sum_j \Omega(i,j)\} \leq C\max_i\{\sum_j \theta(i)\theta(j)\} \leq C\theta_{\max}\|\theta\|_1$. We plug them into the above inequality and apply (D.17). It follows that, with probability $1 - o(n^{-3})$,

$$\|W\| \leq C\sqrt{\theta_{\max}\|\theta\|_1} + C\sqrt{\log(n)} \leq C\sqrt{\theta_{\max}\|\theta\|_1}. \tag{D.20}$$

Combining (D.18) and (D.20) gives (D.16). □

D.2 The eigenvectors of A

We state a main technical lemma about the eigenvectors of A. For $1 \leq k \leq K$, let $\hat{\xi}_k$ be the eigenvector associated with $\hat{\lambda}_k$. Write $\hat{\Xi}_0 = [\hat{\xi}_2, \hat{\xi}_3, \ldots, \hat{\xi}_K] \in \mathbb{R}^{n,K-1}$, and let $\hat{\Xi}'_{0,i}$ denote its ith row, $1 \leq i \leq n$.

Lemma D.2. *Suppose the conditions of Theorem 3.1 hold. With probability $1 - o(n^{-3})$, there exist $\omega \in \{\pm 1\}$ and an orthogonal matrix $X \in \mathbb{R}^{K-1,K-1}$ (both ω and X depend on A and are stochastic) such that*

(a) $\|\omega\hat{\xi}_1 - \xi_1\| \leq C\|\theta\|^{-2}K\sqrt{\theta_{\max}\|\theta\|_1}$;

(b) $\|\hat{\Xi}_0 X - \Xi_0\|_F \leq C\beta_n^{-1}\|\theta\|^{-2}K^{3/2}\sqrt{\theta_{\max}\|\theta\|_1}$;

(c) $\|\omega\hat{\xi}_1 - \xi_1\|_\infty \leq C\|\theta\|^{-3}\theta_{\max}^{3/2}K\sqrt{\|\theta\|_1\log(n)}$;

(d) $\max_{1\leq i\leq n}\|X'\hat{\Xi}_{0,i} - \Xi_{0,i}\| \leq C\beta_n^{-1}\|\theta\|^{-3}\theta_{\max}^{3/2}K^{3/2}\sqrt{\|\theta\|_1\log(n)}$.

If $\beta_n = o(1)$, then the factor K in the bounds for $\|\omega\hat{\xi}_1 - \xi_1\|$ and $\|\omega\hat{\xi}_1 - \xi_1\|_\infty$ can be removed.

Proof of Lemma D.2: We first prove claims (a)-(b). The proof is based on the the classical sin-theta theorem [4], where below is a simpler version [3, Theorem 10].

Lemma D.3. *Let M and $\hat{M}$ be two $n \times n$ symmetric matrices. For $1 \leq k \leq n$, let d_k be the k-th largest eigenvalue of M, η_k and $\hat{\eta}_k$ be the eigenvector associated with the k-th largest eigenvalue of M and $\hat{M}$, respectively. Suppose for some $\delta > 0$ and $1 \leq k_1 \leq k_2 \leq n$, we have $d_{k_1-1} > d_{k_1} + \delta$, $d_{k_2+1} < d_{k_2} - \delta$ and $\|\hat{G} - G\| \leq \delta/2$. Write $U = [\eta_{k_1}, \cdots, \eta_{k_2}]$ and $\hat{U} = [\hat{\eta}_{k_1}, \cdots, \hat{\eta}_{k_2}]$. Then, $\|\hat{U}\hat{U}' - UU'\| \leq 2\delta^{-1}\|\hat{G} - G\|$.*

We divide all eigenvalues of Ω into four groups: (i) λ_1, (ii) positive eigenvalues among $\lambda_2, \ldots, \lambda_K$, (iii) zero eigenvalues, and (iv) negative eigenvalues among $\lambda_2, \ldots, \lambda_K$. Define

Ξ_{01} and Ξ_{02} as the submatrices of Ξ_0 by restricting to columns corresponding to eigenvalues in groups (ii) and (iv), respectively. By dividing the empirical eigenvalues and eigenvectors in a similar way, we can define $\hat{\Xi}_{01}$ and $\hat{\Xi}_{02}$. Now, ξ_1, Ξ_{01} and Ξ_{02} contain the eigenvectors associated with eigenvalues in groups (i), (ii) and (iv), respectively. By Lemma C.2, the gap between eigenvalues in group (i) and those in other groups is $\lambda_1 - |\lambda_2| \geq C^{-1}\lambda_1 \geq C^{-1}K^{-1}\|\theta\|^2$, and the eigen-gap between any two remaining groups is $\geq C\beta_n K^{-1}\|\theta\|^2$. It follows from Lemma D.3 that

$$\|\hat{\xi}_1\hat{\xi}_1' - \xi_1\xi_1'\| = O\Big(\frac{K\|A-\Omega\|}{\|\theta\|^2}\Big), \quad \max_{t\in\{1,2\}}\{\|\hat{\Xi}_{0t}\hat{\Xi}_{0t}' - \Xi_{0t}\Xi_{0t}'\|\} = O\Big(\frac{K\|A-\Omega\|}{\beta_n\|\theta\|^2}\Big). \quad \text{(D.21)}$$

By elementary linear algebra, $(\hat{\xi}_1\hat{\xi}_1' - \xi_1\xi_1')$ has two nonzero eigenvalues $\pm[1-(\hat{\xi}_1'\xi_1)^2]^{1/2}$, where $|1-(\hat{\xi}_1'\xi_1)^2| \geq \min_\pm|1\pm\hat{\xi}_1'\xi_1| = (\min_\pm\|\hat{\xi}_1\pm\xi_1\|^2)/2$. It follows that

$$\min_\pm\|\hat{\xi}_1\pm\xi_1\| \leq \sqrt{2}\|\hat{\xi}_1\hat{\xi}_1' - \xi_1\xi_1'\|. \quad \text{(D.22)}$$

Moreover, by [8, Lemma 2.4], there always is an orthogonal matrix X_1 such that $\|\hat{\Xi}_{01} - \Xi_{01}X_1\|_F \leq \|\hat{\Xi}_{01}\hat{\Xi}_{01}' - \Xi_{01}\Xi_{01}'\|_F$. Since the rank of $(\hat{\Xi}_{01}\hat{\Xi}_{01}' - \Xi_{01}\Xi_{01}')$ is at most $2K$, we then have

$$\|\hat{\Xi}_{01} - \Xi_{01}X_1\|_F \leq \sqrt{2K}\|\hat{\Xi}_{01} - \Xi_{01}X_1\|.$$

Similarly, there exists an orthogonal matrix X_2 such that $\|\hat{\Xi}_{02} - \Xi_{02}X_2\|_F \leq \sqrt{2K}\|\hat{\Xi}_{02} - \Xi_{02}X_2\|$. As a result, for the orthogonal matrix $X = \mathrm{diag}(X_1, X_2)$,

$$\|\hat{\Xi}_0X - \Xi_0\|_F \leq 2\sqrt{K}\max_{t\in\{1,2\}}\{\|\hat{\Xi}_{0t}\hat{\Xi}_{0t}' - \Xi_{0t}\Xi_{0t}'\|\}. \quad \text{(D.23)}$$

Plugging (D.22)-(D.23) into (D.21) gives that with probability $1-o(n^{-3})$,

$$\min_\pm\|\hat{\xi}_1\pm\xi_1\| = O\bigg(\frac{K\|A-\Omega\|}{\|\theta\|^2}\bigg) = O\bigg(\frac{K\sqrt{\theta_{\max}\|\theta\|_1}}{\|\theta\|^2}\bigg),$$
$$\|\hat{\Xi}_0X - \Xi_0\|_F = O\bigg(\frac{K\sqrt{K}\|A-\Omega\|}{\beta_n\|\theta\|^2}\bigg) = O\bigg(\frac{\sqrt{K^3\theta_{\max}\|\theta\|_1}}{\beta_n\|\theta\|^2}\bigg),$$

where we have used (D.16). This proves the first two items.

We then prove claims (c)-(d). We borrow the techniques and some results from [1]. The following lemma is adapted from [1, Theorem 2.1] and is proved below. A direct use of [1, Theorem 2.1] will lead to sub-optimal dependence on β_n in the resulting bound, so we have to modify that theorem accordingly.

Lemma D.4. *Let $M \in \mathbb{R}^{n,n}$ be a symmetric random matrix. Write $M^* = \mathbb{E}M$ and $K_0 = \mathrm{rank}(M^*)$. For each $1 \le k \le K_0$, let d_k^* and d_k be the k-th largest nonzero eigenvalue of M^* and M, respectively, and let η_k^* and η_k be the corresponding eigenvector, respectively. Let s and r be two integers such that $1 \le r \le K_0$ and $0 \le s \le K_0 - r$. Write $D = \mathrm{diag}(d_{s+1}, d_{s+2}, \ldots, d_{s+r})$, $D^* = \mathrm{diag}(d_{s+1}^*, d_{s+2}^*, \ldots, d_{s+r}^*)$,*

$$U = [\eta_{s+1}, \eta_{s+2}, \ldots, \eta_{s+r}], \qquad \text{and} \qquad U^* = [\eta_{s+1}^*, \eta_{s+2}^*, \ldots, \eta_{s+r}^*].$$

Define $\Delta^* = \min\{d_s^* - d_{s+1}^*, d_{s+r}^* - d_{s+r-1}^*, \min_{1\le j\le r} |d_{s+j}^*|\}$ and define $\kappa = (\max_{1\le j\le r} |d_{s+j}^*|)/\Delta^*$. Below, the notation $\|\cdot\|_{2\to\infty}$ represents the maximum row-wise ℓ^2-norm of a matrix, and $M_{m,\cdot}^*$ is the m-th row of M^*. Suppose for a number $\gamma > 0$, the following assumptions are satisfied:

- *A1 (Incoherence): $\max_{1\le m\le n} \|M_{m,\cdot}^*\| \le \gamma\Delta^*$.*
- *A2 (Independence): For any $1 \le m \le n$, the entries of the m-th row and column of M are independent with the other entries.*
- *A3 (Spectral norm concentration): For a number $\delta_0 \in (0,1)$, $\mathbb{P}(\|M - M^*\| \le \gamma\Delta^*) \ge 1 - \delta_0$.*
- *A4 (Row concentration): There is a number $\delta_1 \in (0,1)$ and a continuous non-decreasing function $\varphi(\cdot)$ with $\varphi(0) = 0$ and $\varphi(x)/x$ being non-increasing in $\mathbb{R}^+$ such that, for any $1 \le m \le n$ and non-stochastic matrix $Y \in \mathbb{R}^{n,r}$,*

$$\mathbb{P}\left(\|(M - M^*)_{m,\cdot} Y\|_2 \le \Delta^* \|Y\|_{2\to\infty} \varphi\Big(\frac{\|Y\|_F}{\sqrt{n}\|Y\|_{2\to\infty}} \Big) \right) \ge 1 - \delta_1/n.$$

Let $I_0 = (\{1, \ldots, s-1\} \cup \{s+r+1, \ldots, K_0\}) \cap \{j : |d_j^*| > \max_{1\le i\le r} |d_{s+r}^*|\}$ and $\Delta_0^* = \min\{\min_{j\in I_0} |d_j^* - d_s^*|, \min_{j\in I_0} |d_j^* - d_{s+r}^*|\}$. Define $\widetilde{U}^* = [\eta_1, \ldots, \eta_{K_0}]$ and

$$\widetilde{\kappa} = \begin{cases} \max_{j\in I_0}(|d_j^*|/\Delta_0^*), & \text{if } I_0 \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Then, with probability $1 - \delta_0 - 2\delta_1$, for an orthogonal matrix $O \in \mathbb{R}^{r,r}$,

$$\|UO - MU^*(D^*)^{-1}\|_{2\to\infty} \le C\big[\kappa(\kappa + \varphi(1))(\gamma + \varphi(\gamma)) + \widetilde{\kappa}\gamma\big] \cdot \|\widetilde{U}^*\|_{2\to\infty}. \tag{D.24}$$

Proof of Lemma D.4: Fix $1 \le m \le n$. Let $M^{(m)}$ be the matrix by setting the m-th row and the m-th column of M to be zero. Let $\eta_1^{(m)}, \eta_2^{(m)}, \ldots, \eta_n^{(m)}$ be the eigenvectors

of $M^{(m)}$. Write $U^{(m)} = [\eta_{s+1}^{(m)}, \ldots, \eta_{s+r}^{(m)}]$. Let $H = U'U^*$, $H^{(m)} = (U^{(m)})'U^*$ and $V^{(m)} = U^{(m)}H^{(m)} - U^*$. We aim to prove

$$\begin{aligned}\|M_{m\cdot}V^{(m)}\| \leq & 6(\kappa+\widetilde{\kappa})\gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty} \\ & + \Delta^*\varphi(\gamma)\big(4\kappa\|UH\|_{2\to\infty} + 6\|U^*\|_{2\to\infty}\big). \end{aligned} \tag{D.25}$$

Once (D.25) is obtained, the proof is almost identical to the proof of (B.26) in [1], except that we plug in (D.25) instead of (B.32) in [1]. This is straightforward, so we omit it.

What remains is to prove (D.25). Without loss of generality, we only consider the case where $I_0 \neq \emptyset$. In the proof of [1, Lemma 5], it is shown that

$$\begin{aligned}\|M_{m\cdot}V^{(m)}\| &\leq \|M^*_mV^{(m)}\| + \|(M-M^*)_{m\cdot}V^{(m)}\|, \\ \|(M-M^*)_{m\cdot}V^{(m)}\| &\leq \Delta^*\varphi(\gamma)\big(4\kappa\|UH\|_{2\to\infty} + 6\|U^*\|_{2\to\infty}\big).\end{aligned}$$

Combining them gives

$$\|M_{m\cdot}V^{(m)}\| \leq \|M^*_{m\cdot}V^{(m)}\| + \Delta^*\varphi(\gamma)\big(4\kappa\|UH\|_{2\to\infty} + 6\|U^*\|_{2\to\infty}\big). \tag{D.26}$$

We further bound the first term in (D.26). Recall that I_0 is the index set of eigenvalues that are not contained in D^* and have an absolute value larger than $\|D^*\|$. Let $\widetilde{M}^* = \sum_{j\in I_0} d_j^*\eta_j^*(\eta_j^*)'$.

$$\begin{aligned}\|M^*_{m\cdot}V^{(m)}\| &\leq \|\widetilde{M}^*_{m\cdot}V^{(m)}\| + \|(M^*_{m\cdot} - \widetilde{M}^*_{m\cdot})V^{(m)}\| \\ &\leq \|\widetilde{M}^*_{m\cdot}V^{(m)}\| + \|M^* - \widetilde{M}^*\|_{2\to\infty}\|V^{(m)}\| \\ &\leq \|\widetilde{M}^*_{m\cdot}V^{(m)}\| + 6\gamma\|M^* - \widetilde{M}^*\|_{2\to\infty},\end{aligned}$$

where the last line uses $\|V^{(m)}\| \leq 6\gamma$, by (B.12) of [1]. Note that $M^* - \widetilde{M}^* = \sum_{j\notin I_0} d_j^*\eta_j^*(\eta_j^*)'$. By definition of I_0, for any $j \notin I_0$, $|d_j^*| \leq \max_{1\leq i\leq r}|d_{s+r}^*| \leq \kappa\Delta^*$. It follows that

$$\|M^* - \widetilde{M}^*\|_{2\to\infty} \leq \big(\max_{j\notin I_0}|d_j^*|\big)\|\widetilde{U}^*\|_{2\to\infty} \leq \kappa\Delta^*\|\widetilde{U}^*\|_{2\to\infty}.$$

Combining the above gives

$$\|M^*_{m\cdot}V^{(m)}\| \leq \|\widetilde{M}^*_{m\cdot}V^{(m)}\| + 6\kappa\gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty}. \tag{D.27}$$

Write $D_0^* = \mathrm{diag}(d_j^*)_{j\in I_0}$, $U_0^* = [\eta_j^*]_{j\in I_0}$, $U_0 = [\eta_j]_{j\in I_0}$, $U_0^{(m)} = [\eta_j^{(m)}]_{j\in I_0}$, and $H_0^{(m)} = (U_0^{(m)})'U_0^*$. We similarly have $\|U_0^{(m)}H_0^{(m)} - U_0^*\| \leq 6\gamma_0$, where γ_0 is defined in the same

way as γ but is with respect to the eigen-gap Δ_0^*. It is not hard to see that $\gamma_0 = \gamma\Delta^*/\Delta_0^*$. Hence,

$$\|U_0^{(m)}H_0^{(m)} - U_0^*\| \leq 6\gamma\Delta^*/\Delta_0^*.$$

By mutual orthogonality of eigenvectors, $(U_0^{(m)})'U^{(m)} = 0$ and $(U_0^*)'U^* = 0$. It follows that

$$\begin{aligned}
\|\widetilde{M}_{m\cdot}^* V^{(m)}\| &= \|e_m'[U_0^*\Lambda_0^*(U_0^*)'][U^{(m)}H^{(m)} - U^*]\| \\
&= \|e_m'[U_0^*\Lambda_0^*(U_0^*)']U^{(m)}H^{(m)}\| \\
&\leq \|e_m'[U_0^*\Lambda_0^*(U_0^*)']U^{(m)}\| \\
&= \|e_m'U_0^*\Lambda_0^*(U_0^* - U_0^{(m)}H_0^{(m)})'U^{(m)}\| \\
&\leq \|e_m'U_0^*\Lambda_0^*(U_0^* - U_0^{(m)}H_0^{(m)})'\| \\
&\leq \|\widetilde{U}^*\|_{2\to\infty} \cdot \|\Lambda_0^*\| \cdot \|U_0^* - U_0^{(m)}H_0^{(m)}\| \\
&\leq 6(\|\Lambda_0\|^*/\Delta_0^*) \cdot \gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty}.
\end{aligned}$$

We plug it into (D.27) and note that $\widetilde{\kappa} = \|\Lambda_0\|^*/\Delta_0^*$. It gives

$$\|M_{m\cdot}^* V^{(m)}\| \leq 6(\kappa + \widetilde{\kappa})\gamma\Delta^*\|\widetilde{U}^*\|_{2\to\infty}. \tag{D.28}$$

Combining (D.26) and (D.28) gives (D.25). □

We now come back to the proof of Lemma D.2. We have divided nonzero eigenvalues of Ω into four groups: (i) λ_1, (ii) positive eigenvalues in $\lambda_2, \ldots, \lambda_K$, (iii) zero eigenvalues, and (iv) negative eigenvalues in $\lambda_2, \ldots, \lambda_K$. We shall apply Lemma D.4 to each of the four groups. To save space, we only consider applying it to group (ii). The proof for other groups is similar and omitted.

Now, $M = A$ and $M^* = \Omega = \mathrm{diag}(\Omega) + (A - \mathbb{E}A)$. We check conditions A1-A4. By Lemma C.2, $\Delta^* \geq C\beta_n K^{-1}\|\theta\|^2$ and $\kappa \leq C$. For an appropriately large constant $\tilde{C} > 0$, we take

$$\gamma = \tilde{C}\beta_n^{-1}\|\theta\|^{-2}K\sqrt{\theta_{\max}\|\theta\|_1}.$$

Consider A1. Since $\Omega(i,j) \leq C\theta(i)\theta(j)$, we have $\max_{1\leq i\leq n}\|\Omega_{i,\cdot}\| \leq C\theta_{\max}\|\theta\|$. From the universal inequality $\|\theta\| \leq \sqrt{\theta_{\max}\|\theta\|_1}$ and the assumption $\theta_{\max} = O(1)$, this term is $O(\sqrt{\theta_{\max}\|\theta\|_1})$, which is bounded by $\gamma\Delta^*$ when $\tilde{C}$ is appropriately large. Hence, A1 is satisfied. A2 is satisfied because the upper triangle of A contains independent variables. By (D.16), A3 is satisfied for $\delta_0 = o(n^{-3})$. We then verify A4. Since $\|\mathrm{diag}(\Omega)\| \leq C$,

$$\|\mathrm{diag}(\Omega)_{i,\cdot}Y\|_2 \leq C\|Y\|_{2\to\infty}, \qquad 1 \leq i \leq n. \tag{D.29}$$

Fix $1 \le i \le n$ and $1 \le k \le r$. Let $y_k \in \mathbb{R}^n$ be the k-th column of Y. Using the Bernstein's inequality, for any $t \ge 0$,

$$\mathbb{P}\big(|y_k'(A - \mathbb{E}A)_{i,\cdot}| > t\big) \le 2\exp\left(-\frac{t^2/2}{\sum_{j=1}^n \Omega(i,j)y_k^2(j) + t\|y_k\|_\infty/3}\right). \tag{D.30}$$

Note that $\sum_j \Omega(i,j)y_k^2(j) \le C\|y_k\|_\infty^2 \theta_{\max}\|\theta\|_1$. Moreover, $\theta_{\max}\|\theta\|_1 \gg \log(n)$ by (D.17). We take $t = C\|y_k\|_\infty \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ for a large enough constant $C > 0$. It follows that with probability $1 - o(n^{-4})$,

$$|y_k'(A - \mathbb{E}A)_{i,\cdot}| \le \|y_k\|_\infty \cdot C\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}.$$

Combining it with the probability union bound and (D.29), with probability $1 - o(n^{-3})$,

$$\begin{aligned}\|(A - \Omega)_{i,\cdot}Y\|_2 &\le C\sqrt{\theta_{\max}\|\theta\|_1 \log(n)} \cdot \|Y\|_{2\to\infty} \\ &\le \Delta^*\|Y\|_{2\to\infty} \cdot \frac{C\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{K^{-1}\beta_n\|\theta\|^2}.\end{aligned} \tag{D.31}$$

Moreover, in (D.30), if we use an alternative bound $\sum_j \Omega(i,j)y_k^2(j) \le \|y_k\|^2\theta_{\max}^2$, we obtain a different bound as follows: With probability $1 - o(n^{-4})$,

$$|y_k'(A - \mathbb{E}A)_{i,\cdot}| \le C\max\big\{\|y_k\|\theta_{\max}\sqrt{\log(n)},\ \|y_k\|_\infty \log(n)\big\}.$$

Due to the probability union bound and (D.29), with probability $1 - o(n^{-3})$,

$$\begin{aligned}\|(A - \Omega)_{i,\cdot}Y\|_2 &\le C\max\big\{\|Y\|_F\theta_{\max}\sqrt{\log(n)},\ \|Y\|_{2\to\infty}\log(n)\big\} \\ &\le \Delta^*\|Y\|_{2\to\infty}\max\left\{\frac{\theta_{\max}\sqrt{n\log(n)}}{K^{-1}\beta_n\|\theta\|^2}\frac{\|Y\|_F}{\sqrt{n}\|Y\|_{2\to\infty}},\ \frac{\log(n)}{K^{-1}\beta_n\|\theta\|^2}\right\}.\end{aligned} \tag{D.32}$$

Let $t_1 = C(K^{-1}\beta_n\|\theta\|^2)^{-1}\sqrt{\theta_{\max}\|\theta\|_1\log(n)}$, $t_2 = C(K^{-1}\beta_n\|\theta\|^2)^{-1}\theta_{\max}\sqrt{n\log(n)}$, and $t_3 = C(K^{-1}\beta_n\|\theta\|^2)^{-1}\log(n)$. Define the function

$$\widetilde{\varphi}(x) = \min\{t_1,\ \max\{t_2x, t_3\}\}.$$

Then, (D.31)-(D.32) together imply that with probability $1 - o(n^{-3})$,

$$\|(A - \mathbb{E}A)_{i,\cdot}Y\|_2 \le \Delta^*\|Y\|_{2\to\infty}\widetilde{\varphi}\Big(\frac{\|Y\|_F}{\sqrt{n}\|Y\|_{2\to\infty}}\Big). \tag{D.33}$$

We look at the function $\widetilde{\varphi}(x)$. Note that $(\sqrt{n}\|Y\|_{2\to\infty})^{-1}\|Y\|_F$ takes values in the interval $[n^{-1/2}, 1]$. By (D.17), $t_1 \gg t_3$. Moreover, since $\|\theta\|_1 \le n\theta_{\max}$, when $x = 1$, $t_2x \ge Ct_1$. Last, when $x = n^{-1/2}$, $t_2x \ll t_3$. Combining the above, we conclude that in $[n^{-1/2}, \infty)$, the

function $\widetilde{\varphi}(x)$ first stays flat at t_3, then linearly increases to t_1 and then stays flat at t_1. Hence, we construct a function $\varphi(x)$, which linearly increases from 0 to t_3 for $x \in [0, n^{-1/2}]$, then linear increases from t_3 to t_1 for $x \in [n^{-1/2}, t_2/t_1]$, and then stays constant as t_1 for $x \in [t_2/t_1, \infty)$. It is seen that $\varphi(0) = 0$, $\varphi(x)/x$ is non-increasing, and $\widetilde{\varphi}(x) \le \varphi(x) \le t_1$ in the interval $[n^{-1/2}, 1]$. By (D.33) and that $\widetilde{\varphi}(x) \le \varphi(x)$, A4 is satisfied with $\delta_1 = o(n^{-3})$. Furthermore, since $\varphi(x) \le t_1$,

$$\varphi(\gamma) \le \frac{C\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{K^{-1}\beta_n\|\theta\|^2}.$$

So far, we have shown that A1-A4 hold.

We now apply Lemma D.4. As mentioned, we only study the eigenvectors in group (ii), which correspond to positive eigenvalues among $\lambda_2, \ldots, \lambda_K$. Let Λ_1 be the diagonal matrix consisting of these eigenvalues and let Ξ_{01} be the matrix formed by associated eigenvectors. Define their empirical counterparts, $\hat{\Lambda}_1$ and $\hat{\Xi}_{01}$, in the same way. In Lemma D.4, we take $U = \hat{\Xi}_{01}$, $U^* = \Xi_{01}$, and $\widetilde{U}^* = \Xi$. Since $\lambda_2, \ldots, \lambda_K$ are at the same order, $\kappa \le C$. Also, $\widetilde{\kappa} \le \lambda_1/(\lambda_1 - |\lambda_2|) \le C$ by our assumption. It follows from (D.24) that there exists an orthogonal matrix O such that

$$\|\hat{\Xi}_{01}O - A\Xi_{01}\Lambda_1^{-1}\|_{2\to\infty} \le \frac{C\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{K^{-1}\beta_n\|\theta\|^2}\|\Xi\|_{2\to\infty}.$$

By Lemma C.3, $\|\Xi\|_{2\to\infty} = O(\sqrt{K}\|\theta\|^{-1}\theta_{\max})$. Plugging it into the above inequality, we find that

$$\|\hat{\Xi}_{01}O - A\Xi_{01}\Lambda_1^{-1}\|_{2\to\infty} \le \frac{C\theta_{\max}^{3/2}K^{3/2}\sqrt{\|\theta\|_1 \log(n)}}{\beta_n\|\theta\|^3}. \tag{D.34}$$

By definition of eigen-decomposition, $\Omega\Xi_{01} = \Xi_{01}\Lambda_1$. It follows that

$$A\Xi_{01}\Lambda_1^{-1} = \Omega\Xi_{01}\Lambda_1^{-1} + (A-\Omega)\Xi_{01}\Lambda_1^{-1} = \Xi_{01} + (A-\Omega)\Xi_{01}\Lambda_1^{-1}.$$

Plugging it into (D.34) yields

$$\|\hat{\Xi}_{01}O - \Xi_{01}\|_{2\to\infty} \le \frac{C\theta_{\max}^{3/2}K^{3/2}\sqrt{\|\theta\|_1 \log(n)}}{\beta_n\|\theta\|^3} + \|(A-\Omega)\Xi_{01}\Lambda_1^{-1}\|_{2\to\infty}. \tag{D.35}$$

To bound the second term on the right hand side, we apply the first line of (D.31) by letting $Y = \Xi_{01}$. It turns out that with probability $1 - o(n^{-3})$,

$$\begin{aligned}\|(A-\Omega)\Xi_{01}\Lambda_1^{-1}\|_{2\to\infty} &\le (\max_{1\le i\le n}\|(A-\Omega)_{i,\cdot}\Xi_{01}\|_2)\cdot\|\Lambda_1^{-1}\| \\ &\le C\sqrt{\theta_{\max}\|\theta\|_1\log(n)}\cdot\|\Xi_{01}\|_{2\to\infty}\cdot\|\Lambda_1^{-1}\|\end{aligned}$$

$$\leq C\sqrt{\theta_{\max}\|\theta\|_1\log(n)}\cdot\sqrt{K}\|\theta\|^{-1}\theta_{\max}\cdot K\beta_n^{-1}\|\theta\|^{-2}, \quad \text{(D.36)}$$

where in the last inequality, the bound of $\|\Lambda_1^{-1}\|$ is from Lemma C.2 and the bound of $\|\Xi_{01}\|_{2\to\infty}$ is from Lemma C.3. Combining (D.35)-(D.36) gives

$$\|\hat{\Xi}_{01}O-\Xi_{01}\|_{2\to\infty}\leq\frac{C\theta_{\max}^{3/2}K^{3/2}\sqrt{\|\theta\|_1\log(n)}}{\beta_n\|\theta\|^3}.$$

Note that the left hand side only involves eigenvectors in group (ii). We can prove similar results for the other three groups of eigenvectors. For group (i), $\Delta^*\geq CK^{-1}\|\theta\|^{-1}$ and $\|\widetilde{U}^*\|_{2\to\infty}\leq C\|\theta\|^{-1}\theta_{\max}$, and the resulting bound is

$$\|\omega\hat{\xi}_1-\xi_1\|_\infty\leq\frac{C\theta_{\max}^{3/2}K\sqrt{\|\theta\|_1\log(n)}}{\|\theta\|^3}.$$

Furthermore, if $\beta_n=o(1)$, by Lemma C.2, $\lambda_1-|\lambda_2|\geq C^{-1}\lambda_1\geq C^{-1}K\|\theta\|^2$. Compared with the case of $\beta_n\geq c$, the Δ^* of group (i) is larger by a factor of K, so all the bounds concerning $\hat{\xi}_1$ are reduced by a factor of K. □

D.3 Proof of Theorem 3.1

Without loss of generality, we assume $T=\infty$, so that no thresholding is applied in obtaining $\hat{R}$. Note that $\max_i\|r_i\|\leq\max_k\|v_k\|\leq C\sqrt{K}$ by Lemma C.4. For any threshold $\sqrt{K}\ll T<\infty$, the threshold always reduces errors. Therefore, the error bounds for the case of no thresholding immediately imply the error bounds for the case of thresholding.

The second claim is straightforward. We only show the first claim. By Lemma C.3, we can choose the sign of ξ_1 such that it is a strictly positive vector. By definition of err_n, we can re-write

$$err_n=\frac{\|\theta\|}{\theta_{\min}}\cdot\frac{\theta_{\max}^{3/2}\sqrt{\|\theta\|_1\log(n)}}{\|\theta\|^3}.$$

Then, the statements (c)-(d) of Lemma D.2 can be re-expressed as

$$\|\omega\hat{\xi}-\xi\|_\infty=O\Big(\frac{\theta_{\min}}{\|\theta\|}Kerr_n\Big),\quad\max_{1\leq i\leq n}\|X'\hat{\Xi}_{i,0}-\Xi_{i,0}\|=O\Big(\frac{\theta_{\min}}{\|\theta\|}K^{3/2}\beta_n^{-1}err_n\Big). \quad \text{(D.37)}$$

We now show the claim. Let (ω,X) be the same as in Lemma D.2, and define $H=\omega X'\in\mathbb{R}^{K-1,K-1}$. Fix i. By definition of $(r_i,\hat{r}_i)$ and H,

$$r_i=\frac{1}{\xi_1(i)}\Xi_{i,0},\qquad H\hat{r}_i=\omega X'\hat{r}_i=\frac{1}{\omega\hat{\xi}_1(i)}X'\hat{\Xi}_{i,0}.$$

It follows that

$$H\hat{r}_i-r_i=\frac{1}{\omega\hat{\xi}_1(i)}(X'\hat{\Xi}_{i,0}-\Xi_{i,0})+\Big[\frac{1}{\omega\hat{\xi}_1(i)}-\frac{1}{\xi_1(i)}\Big]\Xi_{i,0}$$

$$= \frac{1}{\omega\hat{\xi}_1(i)}(X'\hat{\Xi}_{i,0} - \Xi_{i,0}) - \frac{\omega\hat{\xi}_1(i) - \xi_1(i)}{\omega\hat{\xi}_1(i)} r_i.$$

First, by Lemma C.3, $\xi_1(i) \geq C\theta_{\min}/\|\theta\|$; also, by (D.37), $|\omega\hat{\xi}_1(i) - \xi_1(i)| \ll \theta_{\min}/\|\theta\|$. We thus have $\omega\hat{\xi}_1(i) \geq \xi_1(i)/2 \geq C\theta_{\min}/\|\theta\|$. Second, using the first bullet point of Lemma C.4, we have $\|r_i\| \leq \max_k \|v_k\| \leq C\sqrt{K}$. Plugging these results into the above equation gives

$$\|H\hat{r}_i - r_i\| \leq \frac{C\|\theta\|}{\theta_{\min}}\big(\|X'\hat{\Xi}_{i,0} - \Xi_{i,0}\| + \sqrt{K}|\omega\hat{\xi}_1(i) - \xi_1(i)|\big). \tag{D.38}$$

The claim follows by plugging (D.37) into (D.38). □

D.4 The ℓ^2-norm deviation bound for $\hat{R}$

Theorem 3.1 is about the row-wise large deviation bound for $\hat{R}$. For completeness of theory, we also present the ℓ^2-norm deviation bound for $\hat{R}$. This result will be useful in the proofs of Theorems 3.5-B.1 about faster rates of Mixed-SCORE. Recall the following definition:

$$err_n^* = [(\theta_{\max}^{1/2}\bar{\theta}^{3/2})/(\theta_{\min}\bar{\theta}_*)] \cdot (n\bar{\theta}^2)^{-1/2}.$$

Lemma D.5. *Under conditions of Theorem 3.1, with probability* $1 - o(n^{-3})$,

$$n^{-1}\sum_{i=1}^{n} \|H\hat{r}_i - r_i\|^2 \leq CK^3\beta_n^{-2}(err_n^*)^2.$$

Proof of Lemma D.5: As explained in the proof of Theorem 3.1, we only need to prove the claim for the special case of $T = \infty$ in obtaining $\hat{R}$ (i.e., no thresholding is applied). By definition of err_n^*, we can re-write it as

$$err_n^* = \frac{\|\theta\|}{\theta_{\min}\sqrt{n}} \cdot \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\|\theta\|^2}.$$

Then, the first two bullet points of Lemma D.2 can be re-expressed as

$$\|\omega\hat{\xi} - \xi\| = O\Big(\frac{\theta_{\min}\sqrt{n}}{\|\theta\|}Kerr_n^*\Big), \quad \|\hat{\Xi}_0 X - \Xi_0\|_F = O\Big(\frac{\theta_{\min}\sqrt{n}}{\|\theta\|}K^{3/2}\beta_n^{-1}err_n^*\Big).$$

Combining it with (D.38) gives

$$n^{-1}\sum_{i=1}^{n} \|H\hat{r}_i - r_i\|^2 \leq \frac{C\|\theta\|^2}{n\theta_{\min}^2}\big(\|\hat{\Xi}_0 X - \hat{\Xi}_0\|_F^2 + K\|\omega\hat{\xi}_1 - \xi_1\|^2\big) \leq CK^3\beta_n^{-2}(err_n^*)^2.$$

This proves the claim. □

D.5 A property of the rotation matrix H

Lemma D.6. *Let H be the orthogonal matrix in Theorem 3.1. With probability $1-o(n^{-3})$, $\|H\mathrm{diag}(\hat{\lambda}_2,\ldots,\hat{\lambda}_K) - \mathrm{diag}(\hat{\lambda}_2,\ldots,\hat{\lambda}_K)H\| \leq C\sqrt{\theta_{\max}\|\theta\|_1}$.*

Proof of Lemma D.6: Write for short $\hat{\Lambda}_0 = \mathrm{diag}(\hat{\lambda}_2,\ldots,\hat{\lambda}_K)$. Let $\hat{\Xi}_0$, $\hat{\Xi}_0$, ω and X be the same as in Lemma D.2. In the proof of Theorem 3.1, we have seen that

$$H = \omega X', \qquad \text{where} \quad \omega \in \{\pm 1\}.$$

It follows that

$$\begin{aligned}\|H\hat{\Lambda}_0 - \hat{\Lambda}_0 H\| &= \|(H\hat{\Lambda}_0 - \hat{\Lambda}_0 H)'\| = \|X\hat{\Lambda}_0 - \hat{\Lambda}_0 X\| \\ &= \|(\hat{\Xi}_0'\Xi_0)\hat{\Lambda}_0 - \hat{\Lambda}_0(\hat{\Xi}_0'\Xi_0) + (H - \hat{\Xi}_0'\Xi_0)\hat{\Lambda}_0 - \hat{\Lambda}_0(H - \hat{\Xi}_0'\Xi_0)\| \\ &\leq \|(\hat{\Xi}_0'\Xi_0)\hat{\Lambda}_0 - \hat{\Lambda}_0(\hat{\Xi}_0'\Xi_0)\| + 2\|\hat{\Xi}_0'\Xi_0 - X\| \cdot \|\hat{\Lambda}_0\|. \qquad \text{(D.39)}\end{aligned}$$

We shall apply [1, Lemma 2]: in our setting, their notations H and $\mathrm{sgn}(H)$ correspond to our notations of $\hat{\Xi}_0'\Xi_0$ and X. By their Lemma 2,

$$\|\hat{\Xi}_0'\Xi_0 - X\|^{1/2} \leq C\|A-\Omega\|/\Delta^*, \quad \|(\hat{\Xi}_0'\Xi_0)\hat{\Lambda}_0 - \hat{\Lambda}_0(\hat{\Xi}_0'\Xi_0)\| \leq 2\|A-\Omega\|, \qquad \text{(D.40)}$$

where Δ^* is the eigen-gap quantity defined in the proof of Lemma D.2 and satisfies $\Delta^* \geq C\beta_n K^{-1}\|\theta\|^2$. Additionally, by Lemma C.2 and Lemma D.1, $\|\hat{\Lambda}_0\| \lesssim \|\Lambda_0\| \leq C\beta_n K^{-1}\|\theta\|^2 \leq C\Delta^*$, with probability $1-o(n^{-3})$. Combining these with (D.39)-(D.40), we have: with probability $1-o(n^{-3})$,

$$\begin{aligned}\|H\hat{\Lambda}_0 - \hat{\Lambda}_0 H\| &\leq \|(\hat{\Xi}_0'\Xi_0)\hat{\Lambda}_0 - \hat{\Lambda}_0(\hat{\Xi}_0'\Xi_0)\| + 2\|\hat{\Xi}_0'\Xi_0 - X\| \cdot \|\hat{\Lambda}_0\| \\ &\leq 2\|A-\Omega\| + C(\|A-\Omega\|/\Delta^*)^2 \cdot C\Delta^* \\ &\leq C\|A-\Omega\| \\ &\leq C\sqrt{\theta_{\max}\|\theta\|_1},\end{aligned}$$

where the third line is because $\|A-\Omega\| \ll \Delta^*$ and the last line is from (D.16). □

E Vertex Hunting

Mixed-SCORE as a generic algorithm, where the VH step is a plug-in step. To analyze the errors of Mixed-SCORE, we must first understand the errors of different VH approaches.

Definition E.1 (Efficiency and strong efficiency of Vertex Hunting)**.** *A Vertex Hunting algorithm is said to be efficient if it satisfies* $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\max_{1\le i\le n}\|H\hat{r}_i - r_i\|$*, and it is said to be strongly efficient if* $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\big(n^{-1}\sum_{i=1}^n \|H\hat{r}_i - r_i\|^2\big)^{1/2}$*, where* H *is the same orthogonal matrix as in Theorem 3.1.*

Consider all 4 VH approaches: SVS, SVS*, CVS, and SP in Table 1. We show

- All approaches are efficient under some regularity conditions.
- SVS and SVS* are also strongly efficient in some settings (however, CVS and SP are generally not strongly efficient; this is because SVS and SVS* use a denoise stage while CVS and SP do not).

E.1 Efficiency of SP and CVS

The next lemma gives the efficiency of CVS and SP.

Lemma E.1 (Efficiency of CVS and SP)**.** *Suppose conditions of Theorem 3.2 hold. Suppose we apply either CVS or SP algorithm to the* n *rows of* $\hat{R}$*. With probability* $1 - o(n^{-3})$*, the estimated* $\hat{v}_1, \ldots, \hat{v}_K$ *satisfy that* $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\max_{1\le i\le n}\|H\hat{r}_i - r_i\|$*. Therefore, both the CVS and SP algorithms are efficient.*

Proof of Lemma E.1: Without loss of generality, we only consider the case that H equals to the identity matrix. When H is not the identity matrix, noticing that $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| = \max_{1\le k\le K}\|\hat{v}_k - H'v_k\|$, we only need to plug $H'v_1, \ldots, H'v_K$ into the proof below.

We first prove the efficiency of the CVS algorithm. Write $\hat{h} = \max_{1\le i\le n}\|\hat{r}_i - r_i\|$. We aim to show

$$\min_{1\le \ell\le K}\|v_k - \hat{v}_\ell\| \le C_0\hat{h}, \qquad \text{for all } 1\le k\le K. \tag{E.41}$$

It means for each true vertex v_k, there is at least one of $\{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K\}$ that is within a distance of $C_0\hat{h}$ to v_k. At the same time, since $\hat{h} = o(\sqrt{K})$ and the distance between any two vertices is $\ge C\sqrt{K}$ (see Lemma C.4), each $\hat{v}_\ell$ cannot be simultaneously within a distance $C_0\hat{h}$ to two vertices. The above imply that there is a one-to-one correspondence between true and estimated vertices such that for each true vertex the corresponding estimated vertex is within a distance $C_0\hat{h}$ to it. The claim then follows.

It remains to show (E.41). Fix $1\le k\le K$. Recall that w_i is the unique weight vector such that $r_i = \sum_{s=1}^K w_i(s)v_s$, $1\le i\le n$. For a constant $C_1 > 0$ to be decided, let

$$\mathcal{V}_{0k} = \big\{1\le i\le n : w_i(k) \ge 1 - C_1K^{-1/2}\hat{h}\big\}.$$

Let $\hat{i}_s$ be such that $\hat{v}_s = \hat{r}_{\hat{i}_s}$, $1 \le s \le K$. We shall first prove that

$$\{\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_K\} \cap \mathcal{V}_{0k} \neq \emptyset. \tag{E.42}$$

This means at least one of the estimated vertices has to come from the point set $\{\hat{r}_i : i \in \mathcal{V}_{0k}\}$. We shall next prove that

$$\max_{i \in \mathcal{V}_{0k}} \|\hat{r}_i - v_k\| \le C_0 \hat{h}. \tag{E.43}$$

Then, the estimated vertex which comes from $\{\hat{r}_i : i \in \mathcal{V}_{0k}\}$ is guaranteed to be within a distance $C_0\hat{h}$ to the true v_k, i.e., (E.41) holds.

It remains to show (E.42)-(E.43). First, consider (E.42). In the proof of Lemma C.4, we introduce a one-to-one linear mapping T_2 from the standard simplex $\mathcal{S}_0$ to the Ideal Simplex $\mathcal{S}^{ideal}$ such that $T_2(w_i) = r_i$ for all $1 \le i \le n$. We have shown that both T_2 and T_2^{-1} are Lipschitz with the Lipschitz constants at the order of $\sqrt{K}$ and $1/\sqrt{K}$, respectively. As a result, there is a constant $C_2 > 1$ such that, for any $w, \tilde{w} \in \mathcal{S}_0$,

$$C_2^{-1}\sqrt{K}\|w - \tilde{w}\| \le \|T_2(w) - T_2(\tilde{w})\| \le C_2\sqrt{K}\|w - \tilde{w}\|. \tag{E.44}$$

Below, we first use (E.44) to show the distance from v_k to the convex hull of $\{r_i : i \notin \mathcal{V}_{0k}\}$ is sufficiently large, and then prove (E.42) by contradiction. We take $C_1 = 5C_2$. Take an arbitrary point x^* from the convex hull $\mathcal{H}\{r_i : i \notin \mathcal{V}_{0k}\}$. Since T_2 is a linear mapping, $y^* = T_2^{-1}(x^*)$ is a convex combination of $\{w_i : i \notin \mathcal{V}_{0k}\}$. By definition, for each $i \notin \mathcal{V}_{0k}$, $0 \le w_i(k) \le 1 - C_1 K^{-1/2}\hat{h}$. As a result, $y^*(k)$, as a convex combination of $\{w_i(k) : i \notin \mathcal{V}_{0k}\}$, also satisfies that $0 \le y^*(k) \le 1 - C_1 K^{-1/2}\hat{h}$. This implies

$$\|T_2^{-1}(x^*) - e_k\| = \|y^* - e_k\| \ge C_1 K^{-1/2}\hat{h}, \qquad \text{for any } x^* \in \mathcal{H}\{r_i : i \notin \mathcal{V}_{0k}\}.$$

Combining it with (E.44), we have

$$\|x^* - v_k\| = \|T_2(y^*) - T_2(e_k)\| \ge C_2^{-1}\sqrt{K} \cdot C_1 K^{-1/2}\hat{h} \ge 5\hat{h}.$$

Since x^* is taken arbitrarily from the convex hull $\mathcal{H}\{r_i : i \notin \mathcal{V}_{0k}\}$, we have

$$d\big(v_k, \mathcal{H}\{r_i : i \notin \mathcal{V}_{0k}\}\big) \ge 5\hat{h}. \tag{E.45}$$

Come back to the proof of (E.42). When this claim is not true, the estimated simplex $\widehat{\mathcal{S}}$ is contained in the convex hull of $\{\hat{r}_i : i \notin \mathcal{V}_{0k}\}$. It follows that

$$d(v_k, \widehat{\mathcal{S}}) \ge d\big(v_k, \mathcal{H}\{\hat{r}_i : i \notin \mathcal{V}_{0k}\}\big)$$

$$\geq d\big(v_k, \mathcal{H}\{r_i : i \notin \mathcal{V}_{0k}\}\big) - \hat{h}$$
$$\geq 4\hat{h}.$$

Let j_k be a pure node of community k. Then, $\|\hat{r}_{j_k} - v_k\| = \|\hat{r}_{j_k} - r_{j_k}\| \leq \hat{h}$. It follows that

$$\max_{1\leq i\leq n} d(\hat{r}_i, \widehat{\mathcal{S}}) \geq d(\hat{r}_{j_k}, \widehat{\mathcal{S}}) \geq d(v_k, \widehat{\mathcal{S}}) - \hat{h} \geq 3\hat{h}. \tag{E.46}$$

At the same time, consider the simplex $\widehat{\mathcal{S}}^*$ formed by $\hat{r}_{j_1}, \hat{r}_{j_2}, \ldots, \hat{r}_{j_K}$, where j_s is a pure node of community s, for $1 \leq s \leq K$. Note that $r_{i_1}, r_{i_2}, \ldots, r_{i_K}$ form the Ideal Simplex $\mathcal{S}^*$ and $\max_{1\leq i\leq n} d(r_i, \mathcal{S}^*) = 0$. It follows that

$$\max_{1\leq i\leq n} d(\hat{r}_i, \widehat{\mathcal{S}}^*) \leq \max_{1\leq i\leq n} d(r_i, \mathcal{S}^*) + 2\hat{h} \leq 2\hat{h}. \tag{E.47}$$

Note that $\widehat{\mathcal{S}}$ is the solution of the combinatory search step. It has to satisfy

$$\max_{1\leq i\leq n} d(\hat{r}_i, \widehat{\mathcal{S}}) \leq \max_{1\leq i\leq n} d(\hat{r}_i, \widehat{\mathcal{S}}^*).$$

This yields a contradiction to (E.46)-(E.47). Hence, (E.42) must be true.

Next, consider (E.43). It is easy to see that

$$\begin{aligned}\max_{i\in\mathcal{V}_{0k}} \|\hat{r}_i - v_k\| &\leq \max_{i\in\mathcal{V}_{0k}} \|r_i - v_k\| + \hat{h} \\ &= \max_{i\in\mathcal{V}_{0k}} \|T_2(w_i) - T_2(e_k)\| + \hat{h} \\ &\leq C_2\sqrt{K} \max_{i\in\mathcal{V}_{0k}} \|w_i - e_k\| + \hat{h},\end{aligned}$$

where we have used (E.44) in the last line. For any $i \in \mathcal{V}_{0k}$, $\|w_i - e_k\|^2 = [1 - w_i(k)]^2 + \sum_{\ell\neq k} w_i^2(\ell) \leq [1 - w_i(k)]^2 + [\sum_{\ell\neq k} w_i(\ell)]^2 \leq 2(C_1K^{-1/2}\hat{h})^2 = 50C_2^2K^{-1}\hat{h}^2$. It follows that

$$\max_{i\in\mathcal{V}_{0k}} \|\hat{r}_i - v_k\| \leq (5\sqrt{2}C_2^2 + 1)\hat{h}.$$

Hence, (E.43) is true by choosing $C_0 = 5\sqrt{2}C_2^2 + 1$.

We then prove the efficiency of the SP algorithm. For space limit, the exact description of the SP algorithm is not given in the main paper. We include it here:

- Initialize $Y_i = (1, \hat{r}_i')' \in \mathbb{R}^K$, for $1 \leq i \leq n$.
- At iteration $k = 1, 2, \ldots, K$: Find $i_k = \mathrm{argmax}_{1\leq i\leq n}\|Y_i\|$ and let $u_k = Y_{i_k}/\|Y_{i_k}\|$. Set the k-th estimated vertex as $\hat{v}_k = \hat{r}_{i_k}$. Update Y_i to $(1 - u_ku_k')Y_i$, for $1 \leq i \leq n$.

This algorithm has been analyzed in various literature. We only need to adapt the existing results. The next lemma is from [5, Theorem 3].

Lemma E.2. *Fix $m \geq r$ and $n \geq r$. Consider a matrix $Y = SM + Z$, where $S \in \mathbb{R}^{m\times r}$ has a full column rank, $M \in \mathbb{R}^{r\times n}$ is a nonnegative matrix such that the sum of each column is at most 1, and $Z = [Z_1, \ldots, Z_n] \in \mathbb{R}^{m\times n}$. Suppose M has a submatrix equal to I_r. Write $\epsilon = \max_{1\leq i\leq n} \|Z_i\|$. Suppose $\epsilon = O(\frac{\sigma_{\min}(S)}{\sqrt{r}\kappa^2(S)})$, where $\sigma_{\min}(S)$ and $\kappa(S)$ are the minimum singular value and condition number of S, respectively. If we apply the SP algorithm to columns of Y, then it outputs an index set $\mathcal{K} \subset \{1, 2, \ldots, n\}$ such that $|\mathcal{K}| = r$ and $\max_{1\leq k\leq r} \min_{j\in\mathcal{K}} \|S_k - Y_j\| = O(\epsilon\kappa^2(S))$, where S_k is the k-th column of S.*

Given $\mathcal{K}$, the estimated vertices by SP are $\{Y_j\}_{j\in\mathcal{K}}$. Hence, the above lemma says the maximum ℓ^2-error on estimating vertices is $O(\epsilon\kappa^2(S)) = O\big(\kappa^2(S)\max_{1\leq i\leq n}\|Z_i\|\big)$.

In our setting, we apply SP to $Y_i = (1, \hat{r}_i')'$, $1 \leq i \leq n$. We shall re-write the data in the same form as in Lemma E.2. Recall that H is the orthogonal matrix in Theorem 3.1 and $v_1, \ldots, v_K$ are vertices of the Ideal Simplex. By definition,

$$\begin{pmatrix} 1 & \cdots & 1 \\ H^{-1}v_1 & \cdots & H^{-1}v_K \end{pmatrix} w_i = \begin{pmatrix} 1 \\ H^{-1}r_i \end{pmatrix}.$$

Let $\tilde{v}_k = (1, (H^{-1}v_k)')'$, $\tilde{r}_i = (1, (H^{-1}r_i)')'$, $z_i = (0, (\hat{r}_i - H^{-1}r_i)')'$, $1 \leq k \leq K$, $1 \leq i \leq n$. It is seen that

$$(1, \hat{r}_i')' \equiv Y_i = [\tilde{v}_1, \ldots, \tilde{v}_K]w_i + z_i.$$

Write $Y = [Y_1, \ldots, Y_n] \in \mathbb{R}^{K\times n}$, $\tilde{V} = [\tilde{v}_1, \ldots, \tilde{v}_K] \in \mathbb{R}^{K\times K}$, $W = [w_1, \ldots, w_n] \in \mathbb{R}^{K\times n}$, and $Z = [z_1, \ldots, z_n] \in \mathbb{R}^{K\times n}$. The above can be re-written as

$$Y = \tilde{V}W + Z. \tag{E.48}$$

This reduces to the form in Lemma E.2 with $m = K$. To apply Lemma E.2, we note that $\tilde{V}$ can be re-written as

$$\tilde{V} = \mathrm{diag}(1, H^{-1}) \cdot Q, \qquad \text{where} \quad Q = \begin{pmatrix} 1 & \cdots & 1 \\ & & \\ v_1 & \cdots & v_K \end{pmatrix}.$$

Since $\mathrm{diag}(1, H^{-1})$ is an orthogonal matrix, the singular values of $\tilde{V}$ are the same as the singular values of Q. Moreover, by (C.15), all the singular values of Q are at the order of $\sqrt{K}$. It follows that

$$\sigma_{\min}(\tilde{V}) \asymp \sqrt{K}, \qquad \kappa(\tilde{V}) \asymp 1. \tag{E.49}$$

In particular, $\tilde{V}$ has a full rank, and $\frac{\sigma_{\min}(\tilde{V})}{\sqrt{K}\kappa^2(\tilde{V})} \asymp 1$. By Lemma E.2, the maximum ℓ^2-error on estimating vertices is $O(\max_{1\le i\le n}\|Z_i\|) = O(\max_{1\le i\le n}\|\hat{r}_i - H^{-1}r_i\|) = O(\max_{1\le i\le n}\|H\hat{r}_i - r_i\|)$. The claim follows immediately. □

E.2 Strong efficiency of SVS and SVS*

SVS and SVS* both have a denoise stage, where we use k-means to reduce the n rows of $\hat{R}$ into L "cluster centers", with an L that is (usually a few times) larger than K. We have seen that the denoise stage makes SVS and SVS* more accurate numerically (see Figure 4). We now give a theoretical justification, where we show that SVS and SVS* are strongly efficient (see Definition E.1). Without loss of generality, we focus on SVS. The analysis of SVS* is very similar, which is discussed in the remark in the end.

First, consider Setting 1. Let $\mathcal{S}_0 = \mathcal{S}_0(e_1, e_2, \ldots, e_K)$ be the standard simplex in $\mathbb{R}^K$, where the vertices $e_1, e_2, \ldots, e_K$ are the standard Euclidean basis vectors of $\mathbb{R}^K$. Fix a density g defined over $\mathcal{S}_0$ and let $\mathcal{R} = \{\pi \in \mathcal{S}_0 : g(\pi) > 0\}$ be the support of g. We suppose there is a constant $c_0 > 0$ such that

$$\mathcal{R} \text{ is an open subset of } \mathcal{S}_0, \text{ and distance}(e_k, \mathcal{R}) \ge c_0,\ 1 \le k \le K. \tag{E.50}$$

Let $\delta_v(\pi)$ denote the point mass at $\pi = v$. Let $\epsilon_1, \ldots, \epsilon_K > 0$ be constants such that $\sum_{k=1}^K \epsilon_k < 1$. We invoke a random design model where π_i's are *iid* drawn from a mixture

$$f(\pi) = \sum_{k=1}^K \epsilon_k \cdot \delta_{e_k}(\pi) + \Big(1 - \sum_{k=1}^K \epsilon_k\Big) \cdot g(\pi). \tag{E.51}$$

Lemma E.3 (Efficiency of SVS, Setting 1)**.** *Suppose conditions of Theorem 3.2 hold. Additionally, suppose K is fixed and rows of Π are iid generated from* (E.50)-(E.51)*. We apply the SVS algorithm to rows of $\hat{R}$ with an L that does not change with n. Then, there exists $L_0 = L_0(g, \epsilon_1, \ldots, \epsilon_K)$ such that, as long as $L \ge L_0$, with probability $1 - o(n^{-3})$, the estimated $\hat{v}_1, \ldots, \hat{v}_K$ satisfy $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\max_{1\le i\le n}\|H\hat{r}_i - r_i\|$. As a result, the SVS algorithm is efficient.*

Lemma E.3 is proved in Section E.2.1. Its proof utilizes the Borel-Lebesgue covering theorem to characterize the local centers produced in the denoise stage.

Remark. A noteworthy implication of Lemma E.3 is that the performance of SVS is robust to the choice of L: an overshooting of L only has negligible effects (so as long as computation is not a serious issue, we can choose a larger L in SVS). This is intuitively

explained as follows. As L increases, more local centers emerge, and we have two representative scenarios. In the first scenario, new "local centers" emerge in the interior of the Ideal Simplex, while "local centers" that fall close to one of the vertices of Ideal Simplex remain unaffected. In this case, as "local centers" that fall in the interior of the Ideal Simplex won't be selected in the second stage of SVS, the estimated vertices remain roughly the same as L increases. In the second scenario, near a vertex of the Ideal Simplex, the number of "local centers" increases as L increases. However, all these "local centers" remain close to the vertex, and in its second stage, SVS selects one of these "local centers" as the estimated vertex. In this case, the estimates of vertices also remain roughly the same as L increases. The above heuristic explanation is made rigorous in the proof of Lemma E.3.

Next, consider Setting 2. Let $\mathcal{N}_k = \{1 \le i \le n : \pi_i(k) = 1\}$ be the set of pure nodes of community k, $1 \le k \le K$, and let $\mathcal{M} = \{1 \le i \le n : \max_{1\le k\le K} \pi_i(k) < 1\}$ be the set of all mixed nodes. We assume there are constants $c_1, c_2 \in (0,1)$ such that

$$\min_{1\le k\le K} |\mathcal{N}_k| \ge c_1 n, \qquad \min_{1\le k\le K} \sum_{i\in\mathcal{N}_k} \theta^2(i) \ge c_2\|\theta\|^2. \tag{E.52}$$

Furthermore, for a fixed integer $L_0 \ge 1$, we assume there is a partition of $\mathcal{M}$, $\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_{L_0}$, a set of PMF's $\gamma_1, \cdots, \gamma_{L_0}$, and constants $c_3, c_4 > 0$ such that (e_k: k-th standard basis vector of $\mathbb{R}^K$)

$$\Big\{ \min_{1\le j\ne\ell\le L_0} \|\gamma_j - \gamma_\ell\|, \ \min_{1\le \ell\le L_0, 1\le k\le K} \|\gamma_\ell - e_k\| \Big\} \ge c_3, \tag{E.53}$$

and for each $1 \le \ell \le L_0$ (note: err_n is the same as that in (3.10)),

$$|\mathcal{M}_\ell| \ge c_4|\mathcal{M}| \ge n\beta_n^{-2} err_n^2, \qquad \max_{i\in\mathcal{M}_\ell} \|\pi_i - \gamma_\ell\| \le 1/\log(n). \tag{E.54}$$

In this setting, we assume that the true π_i's form several *loose clusters*, where the π_i's in the same cluster are within a distance of $O(\frac{1}{\log(n)})$ from each other. We note that $\frac{1}{\log(n)}$ is much larger than the order of noise, $\max_{1\le i\le n} \|H\hat{r}_i - r_i\|$ (see Theorem 3.1). Hence, the assumed clustering structure is "loose".

Lemma E.4 (Strong efficiency of SVS, Setting 2)**.** *Suppose conditions of Theorem 3.2 hold. Additionally, suppose K is fixed and (Θ, Π) satisfy* (E.52)-(E.54)*. For any integer $L \ge 1$, denote by $\epsilon_L(\hat{R})$ the sum of squared residuals of applying k-means to rows of $\hat{R}$ to get L clusters. We apply the SVS algorithm to rows of $\hat{R}$, with a data-drive choice of L:*

$$\hat{L}_n(A) = \min\{L \ge K+1 : \epsilon_L(\hat{R}) < \epsilon_{L-1}(\hat{R})/\log(\log(n))\}. \tag{E.55}$$

With probability $1 - o(n^{-3})$*, the estimated* $\hat{v}_1, \ldots, \hat{v}_K$ *satisfy*

$$\max_{1\le k\le K} \|H\hat{v}_k - v_k\| \le C\Big(n^{-1}\sum_{i=1}^{n} \|H\hat{r}_i - r_i\|^2\Big)^{1/2}. \tag{E.56}$$

As a result, the SVS algorithm is strongly efficient.

Lemma E.4 is proved in Section E.2.2. The proof requires unconventional analysis of k-means. The challenge comes from that the clusters of π_i's are loose. Using the conventional analysis of k-means, the VH error is governed by the largest within-cluster variance, which can be as large as $O(\frac{1}{\log(n)})$ for loose clusters (see (E.54)). The key of the proof is to show that the loose clusters in the interior have negligible effects on the estimated vertices.

Remark. Lemmas E.3-E.4 can be easily extended to SVS*. Let $\hat{h} = \max_i \|H\hat{r}_i - r_i\|$. In the proofs of these lemmas, we have shown the following properties of the k-means cluster centers: With high probability, (a) all k-means centers are within a distance of $O(\hat{h})$ to the Ideal Simplex, and (b) for each vertex v_k, there is at least one k-means center that is within a distance of $O(\hat{h})$ to v_k. SVS* applies SP to these k-means centers. Therefore, we can apply Lemma E.1 pretending that the k-means centers are the data points. This gives the desired claims for SVS*.

E.2.1 Proof of Lemma E.3

Lemma E.3 follows directly from the next lemma:

Lemma E.5. *Suppose the conditions of Lemma E.3 hold. We apply the SVS algorithm to* $\{\hat{r}_i\}_{i=1}^n$ *with* L *being a properly large constant. Write* $\hat{h} = \max_{1\le i\le n} \|H\hat{r}_i - r_i\|$*. The following statements are true.*

- *In the local clustering sub-step, all the local centers output by k-means are within a distance of* $C\hat{h}$ *to the Ideal Simplex. Moreover, for each true vertex* v_k*, there is at least one local center that is within a distance of* $C\hat{h}$ *to it,* $1 \le k \le K$*.*
- *The combinatorial search sub-step selects exactly one local center among those within a distance of* $C\hat{h}$ *to a true* v_k*,* $1 \le k \le K$*. As a result, up to a permutation of estimated vertices,* $\max_{1\le k\le K} \|H\hat{v}_k - v_k\| \le C\hat{h}$*.*

Proof of Lemma E.5: As explained in the proof of Lemma E.1, we can assume $H = I_{K-1}$ without loss of generality.

We first argue that, once the first bullet point is proved, the second bullet point follows directly. Let $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_L$ be the local centers by k-means. The combinatorial search step of SVS is an application of CVS on these local centers, and we hope to apply Lemma E.1. Note that when the first bullet point of the claim is true, we have:

- $d(\hat{m}_j, \mathcal{S}^{ideal}) \le C\hat{h}$, $1 \le j \le L$.
- For each $1 \le k \le K$, there exists j_k such that $\|\hat{m}_{j_k} - v_k\| \le C\hat{h}$.

By Lemma C.4, the distance between two different v_k and v_ℓ is lower bounded by a constant times $\sqrt{K}$, while $\hat{h} = o(\sqrt{K})$. As a result, any $\hat{m}_j$ cannot be simultaneously within a distance of $C\hat{h}$ to two vertices, which implies that $j_1, j_2, \ldots, j_K$ are distinct. Define

$$m_j = \begin{cases} \mathrm{argmin}_{x \in \mathcal{S}^{ideal}} \|x - \hat{m}_j\|, & j \notin \{j_1, j_2, \ldots, j_K\}, \\ v_k, & j = j_k, 1 \le k \le K. \end{cases}$$

We then have

- The points $m_1, m_2, \ldots, m_L$ are in the Ideal Simplex $\mathcal{S}^{ideal}$.
- $\|\hat{m}_j - m_j\| \le C\hat{h}$, $1 \le j \le L$.
- For each $1 \le k \le K$, there is at least one m_j located at the vertex v_k.

If we view $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_L$ as the data points and view $m_{j_1}, \ldots, m_{j_K}$ as the "pure nodes", we can apply Lemma E.1 to get $\max_{1 \le k \le K} \|\hat{v}_k - v_k\| \le C \max_{1 \le j \le L} \|\hat{m}_j - m_j\| \le C\hat{h}$.

Therefore, it suffices to prove the first bullet point of the claim. For any $L \ge 1$, let $RSS(L)$ be the objective achieved by applying k-means to mixed r_i's assuming $\le L$ clusters:

$$RSS(L) = \min_{L \text{ cluster centers}} \sum_{\text{mixed nodes } i} \|r_i - (\text{closest-cluster-center})\|^2.$$

In preparation, we study $RSS(L)$ as a function of L.

We provide an upper bound of $RSS(L)$ by constructing a feasible solution to the k-means problem. In the proof of Lemma C.4, we see that there is a one-to-one mapping $T = T_2 \circ T_1$ from the standard simplex $\mathcal{S}_0$ to the Ideal Simplex $\mathcal{S}^{ideal}$ such that $r_i = T(\pi_i)$ and that (note: we have used that K is a constant)

$$C^{-1}\|x - y\| \le \|T(x) - T(y)\| \le C\|x - y\|, \qquad \text{for any } x, y \in \mathcal{S}_0. \tag{E.57}$$

For an integer $s = \lfloor L^{\frac{1}{K-1}} - 1 \rfloor$, we consider the following choice of centers:

$$\Big\{T(x) : x \in \mathcal{S}_0, \text{ entries of } x \text{ take value on } \big\{0, \frac{1}{s}, ..., \frac{s-1}{s}, 1\big\}\Big\}.$$

The total number of centers is bounded by $(s+1)^{K-1} \le L$. We then assign each r_i to the nearest center. The ℓ^∞-distance from each π_i to the nearest x above is at most $1/s$, so the Euclidean distance is at most $\sqrt{K}/s$; combining it with (E.57), the Euclidean distance from $r_i = T(\pi_i)$ to the nearest $T(x)$ above is at most $C\sqrt{K}/s$. It follows that

$$RSS(L) \le n(C\sqrt{K}/s)^2.$$

The choice of s guarantees that $s > L^{\frac{1}{K-1}} - 2$. As a result, for a constant $\tilde{c}$ that does not depend on L,

$$RSS(L) \le n \cdot \tilde{c} L^{-\frac{2}{K-1}}. \tag{E.58}$$

We are now ready to prove the first bullet point. Note that each $\hat{r}_i$ is within a distance $C\hat{h}$ to the corresponding r_i and that all the r_i's are in the Ideal Simplex. Hence, all data points $\{\hat{r}_i\}_{i=1}^n$ are within a distance $C\hat{h}$ to the Ideal Simplex. It is easy to see that all local centers output by k-means must also be within a distance $C\hat{h}$ to the Ideal Simplex. What remains is to show that there is at least one local center within a distance of $C\hat{h}$ to each true vertex v_k. Fix v_k. Our strategy is as follows: for a constant ℓ_0 to be decided,

(a) We first show that there exists at least one local center within a distance ℓ_0 to v_k.

(b) We then show that, for each local center within a distance ℓ_0 to v_k, the associated data cluster consists of only pure $\hat{r}_i$ from community k.

Then, by the nature of k-means, such a local center equals to the average of all the $\hat{r}_i$ assigned to this cluster. Since each $\hat{r}_i$ corresponds to a pure node of community k, it is within a distance $C\hat{h}$ to v_k. As a result, the local center must also be within a distance $C\hat{h}$ to v_k. This gives the first bullet point.

What remains is to prove (a) and (b). Fix v_k. Consider (a). Suppose there are no local centers within a distance ℓ_0 to v_k. Then, each pure r_i from community k has a distance $> \ell_0$ to the nearest local center; hence, the distance from $\hat{r}_i$ to the nearest local center is at least $\ell_0 - C\hat{h} \ge \ell_0/2$. At the same time, by the generating process of π_i's, with probability $1 - o(n^{-3})$, the number of pure nodes of community k is at least $n\epsilon_k/2$. These pure nodes contribute a sum-of-squares of

$$\ge (n\epsilon_k/2) \cdot (\ell_0/2)^2 = n(\ell_0^2\epsilon_k/8).$$

Additionally, the mixed $\hat{r}_i$'s are assigned to at most L clusters. Since $\|\hat{r}_i - x\|^2 \geq \|r_i - x\|^2/2 - O(\hat{h}^2)$ for any point x, we immediately know that the sum-of-squares contributed by mixed $\hat{r}_i$'s is

$$\geq \frac{1}{2} RSS(L) - O(n\hat{h}^2).$$

Combining the above, the objective attained by k-means is

$$\geq \frac{1}{2} RSS(L) + n(\ell_0^2 \epsilon_k/9) \tag{E.59}$$

At the same time, we construct an alternative solution by letting $(L-K)$ of the local centers be those associated with $RSS(L-K)$, letting the remaining K centers be $v_1, v_2, \ldots, v_K$, and assigning each $\hat{r}_i$ to the center closest to the corresponding r_i. Since $\|\hat{r}_i - x\|^2 \leq 2\|r_i - x\|^2 + O(\hat{h}^2)$, the sum of squares attained by this solution is

$$\leq 2RSS(L-K) + O(n\hat{h}^2). \tag{E.60}$$

A contradiction is obtained as long as

$$\begin{aligned} 2RSS(L-K) - \frac{1}{2} RSS(L+K) &< n(\ell_0^2 \epsilon_k/9) - O(n\hat{h}^2) \\ &< n(\ell_0^2/10). \end{aligned}$$

According to (E.58), the above is true if we choose $L > (20\tilde{c}/\ell_0^2)^{\frac{K-1}{2}}$. This proves (a).

Consider (b). Fix k. Let $\hat{m}^*$ be a local center such that $\|\hat{m}^* - v_k\| \leq \ell_0$. By the assumption (E.50), for any $\pi_i \neq e_k$, its distance to e_k (e_k is the k-th standard basis of $\mathbb{R}^K$) is at least c_0. Combining it with (E.57), for any node i that is not a pure node of community k, the distance from r_i to v_k is at least $C^{-1}c_0$. As a result, for any such node,

$$\|\hat{r}_i - \hat{m}^*\| \geq C^{-1}c_0 - \ell_0 - C\hat{h}.$$

By taking $\ell_0 = C^{-1}c_0/4.1$, for any node i not pure of community k,

$$\text{the distance from } \hat{r}_i \text{ to the center } \hat{m}^* \text{ is at least } 3\ell_0. \tag{E.61}$$

We shall also show that, for any node i not pure of community k,

$$\text{the distance from } \hat{r}_i \text{ to the nearest center is at most } 2.5\ell_0. \tag{E.62}$$

By (E.61)-(E.62), these nodes cannot be assigned to $\hat{m}^*$. Therefore, the cluster associated with $\hat{m}^*$ consists of only those $\hat{r}_i$ such that i is a pure node of community k. This proves (b).

What remains is to prove (E.62). If i is a pure node of a different community ℓ, then by (a) above, the distance from $r_i = v_\ell$ to the nearest center is $\ell_0 + C\hat{h} < 2.5\ell_0$. Hence, we only need to consider i that is a mixed node. Since $\max_i \|\hat{r}_i - r_i\| \leq C\hat{h} \ll 0.5\ell_0$, it suffices to show that

$$\text{the distance from a mixed } r_i \text{ to the nearest center is at most } 2\ell_0. \tag{E.63}$$

Let $\mathcal{S}_0 = \mathcal{S}_0(e_1, \ldots, e_K) \in \mathbb{R}^K$ be the standard $(K-1)$-simplex, and denote by $\mathcal{B}(x; c)$ an open ball in $\mathcal{S}_0$ centered at x with a radius c; we notice that here an "open ball" means the intersection of $\mathcal{S}_0$ and an open ball in $\mathbb{R}^K$. Let $\bar{\mathcal{R}}$ be the closure of $\mathcal{R}$, where $\mathcal{R}$ is the support of $f(\cdot)$. We consider the open cover of $\bar{\mathcal{R}}$:

$$\big\{\mathcal{B}(x, C^{-1}\ell_0) : x \in \mathcal{R}\big\}.$$

Since $\bar{\mathcal{R}}$ is closed and bounded, it is a compact set. According to the Borel-Lebesgue covering theorem, the above open cover has a finite sub-cover:

$$\big\{\mathcal{B}(x_1, C^{-1}\ell_0), \mathcal{B}(x_2, C^{-1}\ell_0), \ldots, \mathcal{B}(x_p, C^{-1}\ell_0)\big\}, \qquad \text{where } x_1, \ldots, x_p \in \mathcal{R}.$$

This means each $\pi_i \neq e_k$ is contained in one $\mathcal{B}(x_j, C^{-1}\ell_0)$. Recalling that T is the mapping in (E.57), define

$$\mathcal{B}_j^* = T\big(\mathcal{B}(x_j, C^{-1}\ell_0)\big), \qquad 1 \leq j \leq p.$$

Then, $r_i = T(\pi_i)$ is contained in $\mathcal{B}_j^*$. Moreover, for any $y, \tilde{y} \in \mathcal{B}_j^*$, $\|y - \tilde{y}\| \leq C \max_{x,\tilde{x} \in \mathcal{B}(x_j, C^{-1}\ell_0)} \leq 2\ell_0$. Therefore, if we can show that

$$\text{each } \mathcal{B}_j^* \text{ contains at least one local center}, 1 \leq j \leq p, \tag{E.64}$$

then the distance from r_i to this local center is bounded by $2\ell_0$. This gives (E.63), and in turn gives (E.62).

What remains is to prove (E.64). Note that $\mathcal{R}$ is an open set. By definition of open sets, for each of $x_1, x_2, \ldots, x_p$, there is a $\tau_j > 0$ such that the closed ball $\bar{\mathcal{B}}(x_j, \tau_j)$ is contained in $\mathcal{R}$. We define the closed balls

$$\mathcal{BB}_j \equiv \bar{\mathcal{B}}\big(x_j, \min\{\tau_j, C^{-1}\ell_0/2\}\big), \qquad 1 \leq j \leq p.$$

Let $\omega_j = \int f(\pi) 1\{\pi \in \mathcal{BB}_j\} d\pi = (1 - \sum_{k=1}^K \epsilon_k) \int g(\pi) 1\{\pi \in \mathcal{BB}_j\} d\pi$, $1 \leq j \leq p$. Note that each of these closed balls is contained in the support of g with a nonzero radius and

that g as a probability density is measurable. We immediately know that $\omega_j > 0$. From the assumption (E.51) and elementary large-deviation inequalities (e.g., the Hoeffding's inequality), we know that with probability $1 - o(n^{-3})$, for $1 \le j \le p$,

$$\text{the number of } \pi_i\text{'s contained in } \mathcal{BB}_j \text{ is at least } n\omega_j/2. \tag{E.65}$$

With (E.65), we now prove (E.64) by contradiction. Suppose (E.64) does not hold, i.e., there exists $\mathcal{B}_j^*$ such that

$$\mathcal{B}_j^* \cap \{\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_L\} = \emptyset,$$

where $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_L$ are the local centers output by k-means. By definition of $\mathcal{B}_j^*$ and the fact that T is a one-to-one mapping, we have

$$\mathcal{B}(x_j, C^{-1}\ell_0) \cap \left\{T^{-1}(\hat{m}_1), T^{-1}(\hat{m}_2), \ldots, T^{-1}(\hat{m}_L)\right\} = \emptyset.$$

Note that $\mathcal{BB}_j$ is a ball also centered at x_j but with a radius no larger than half of the radius of $\mathcal{B}(x_j, C^{-1}\ell_0)$. As a result, for any $x \in \mathcal{BB}_j$, its distance to the nearest one of $T^{-1}(\hat{m}_1), \cdots, T^{-1}(\hat{m}_L)$ is at least $C^{-1}\ell_0/2$; combining it with (E.57), the distance from $T(x)$ to the nearest one of $\hat{m}_1, \hat{m}_2, \ldots, \hat{m}_L$ is at least $C^{-2}\ell_0/2$. It follows that

$$\text{for any } \pi_i \in \mathcal{BB}_j, \min_{1 \le s \le L} \|r_i - \hat{m}_s\| \ge C^{-2}\ell_0/2.$$

Note that $\max_i \|\hat{r}_i - r_i\| \le C\hat{h} = o(1)$. We further conclude that

$$\begin{gathered}\text{for any } \pi_i \in \mathcal{BB}_j, \text{ the distance from } \hat{r}_i \\ \text{to the nearest local center is } \ge C^{-2}\ell_0/3.\end{gathered} \tag{E.66}$$

Combining (E.65)-(E.66), the sum-of-squares attained by k-means is

$$\ge (C^{-2}\ell_0/3)^2 \cdot (n\omega_j/2) \ge n(\omega_{\min} C^{-4}\ell_0^2/18),$$

where $\omega_{\min} = \min\{\omega_1, \ldots, \omega_p\}$. At the same time, the objective attained by k-means should be

$$\le RSS(L) + n(C\hat{h}^2).$$

A contradiction is obtained as long as

$$RSS(L) < n(\omega_{\min} C^{-4}\ell_0^2/18) - n(C\hat{h}^2). \tag{E.67}$$

Comparing it with (E.58), as long as $L > (\frac{19C^4\tilde{c}}{\ell_0^2\omega_{\min}})^{\frac{K-1}{2}}$, the inequality (E.67) will be true. We then have a contradiction, which implies that (E.64) must hold. The proof is now complete. □

E.2.2 Proof of Lemma E.4

Lemma E.4 follows directly from the next lemma:

Lemma E.6. *Suppose the conditions of Lemma E.4 hold. We apply the SVS algorithm to* $\{\hat{r}_i\}_{i=1}^n$ *with* $L = \hat{L}_n(A)$*, where* $\hat{L}_n(A)$ *is defined in* (E.55)*. Let* $\hat{h}^* = \sqrt{n^{-1}\sum_{i=1}^n \|H\hat{r}_i - r_i\|^2}$ *and* $\hat{h} = \max_{1\le i\le n}\|H\hat{r}_i - r_i\|$*. With probability* $1 - o(n^{-3})$*, the following statements are true.*

- $\hat{L}_n(A) = L_0 + K$.
- *The local clustering sub-step identifies* $(L_0 + K)$ *local centers, where there is a unique* $(K-1)$*-simplex such that* K *of these centers (denoted by* $\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_K$*) are its vertices, and all other centers are within a distance of* $C\hat{h}$ *to this simplex. These* K *local centers will be identified by the combinatorial search sub-step.*
- *The above* K *local centers satisfy* $\hat{v}_k = |\mathcal{N}_k|^{-1}\sum_{i\in\mathcal{N}_k}\hat{r}_i$*,* $1 \le k \le K$*. As a result, up to a permutation of estimated vertices,* $\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le C\hat{h}^*$.

Proof of Lemma E.6: As explained in the proof of Lemma E.1, we can assume $H = I_{K-1}$ without loss of generality. By Theorem 3.1 and Lemma D.5, with probability $1 - o(n^{-3})$,

$$\hat{h} \equiv \max_{1\le i\le n}\|\hat{r}_i - r_i\| \le \frac{Cerr_n}{\beta_n}, \quad n(h^*)^2 \equiv \sum_{i=1}^n \|\hat{r}_i - r_i\|^2 \le \frac{Cn(err_n^*)^2}{\beta_n^2}, \tag{E.68}$$

where we have absorbed the factors of K into the constants. We also note that $err_n^* \le err_n/\sqrt{\log(n)}$. Below, we restrict to the event of (E.68).

First, we study $\hat{L}_n(A)$. Recall that $\gamma_1, \gamma_2, \ldots, \gamma_{L_0}$ are as in (E.53). Let T be the mapping as in (E.57); note that $T(\pi_i) = r_i$ for $1 \le i \le n$. Introduce

$$m_j = T(\gamma_j), \qquad 1 \le j \le L_0.$$

By (E.57), the assumptions (E.53)-(E.54) imply that the distance between any two of $\{v_1, v_2, \ldots, v_K, m_1, m_2, \ldots, m_{L_0}\}$ is at least c, and $\max_{i\in\mathcal{M}_j}\|r_i - m_j\| \le C_1/\log(n)$, where $c > 0$ and $C_1 > 0$ are constants. In particular,

$$\alpha_n^2 \le \frac{C|\mathcal{M}|}{n\log(n)}, \qquad \text{where} \quad \alpha_n^2 \equiv n^{-1}\sum_{j=1}^{L_0}\sum_{i\in\mathcal{M}_j}\|r_i - m_j\|^2.$$

We now study $\epsilon_L(\hat{R})$. When $L = L_0+K$, by choosing this choice of centers $\{v_1, \ldots, v_K, m_1, \ldots, m_{L_0}\}$, it is easy to see that

$$\epsilon_{L_0+K}(\hat{R}) \leq n\alpha_n^2 + C\sum_{i=1}^{n} \|\hat{r}_i - r_i\|^2 \leq \frac{C|\mathcal{M}|}{\log(n)}, \tag{E.69}$$

where the last inequality is due to (E.68) and the assumption that $|\mathcal{M}| \geq n\beta_n^{-2} err_n^2 \geq n\beta_n^{-2}(err_n^*)^2 \log(n)$. When $K \leq L < L_0 + K$, suppose there are L_1 of $\{v_1, v_2, \ldots, v_K\}$ and L_2 of $\{m_1, m_2, \ldots, m_{L_0}\}$ such that no local centers are within a distance of $c/3$ of them. Since the distance between any two of $\{v_1, v_2, \ldots, v_K, m_1, m_2, \ldots, m_{L_0}\}$ is at least c, we have that $(L_1 + L_2)$ is at least $(L_0 + K) - L$. For any such v_k and $i \in \mathcal{N}_k$ or such m_j and $i \in \mathcal{M}_j$, the distance from $\hat{r}_i$ to the nearest local center is at least $c/3 - \hat{h} \geq c/4$. It follows that

$$\epsilon_L(\hat{R}) \geq (c/4)^2 \cdot (L_1 \min_k |\mathcal{N}_k| + L_2 \min_j |\mathcal{M}_j|) \geq C|\mathcal{M}|, \tag{E.70}$$

where the last inequality is due to $\min_k |\mathcal{N}_k| \geq c_1 n$ and $\min_j |\mathcal{M}_j| \geq c_4|\mathcal{M}|$. At the same time, by choosing the centers to be $\{v_1, v_2, \ldots, v_K\}$ and $(L - K)$ of $\{m_1, m_2, \ldots, m_{L_0}\}$,

$$\epsilon_L(\hat{R}) \leq C(L_0 + K - L)|\mathcal{M}| + C\sum_{i=1}^{n} \|\hat{r}_i - r_i\|^2 \leq C|\mathcal{M}|. \tag{E.71}$$

By (E.69)-(E.71),

$$\epsilon_L(\hat{R})/\epsilon_{L-1}(\hat{R}) \begin{cases} \leq C/\log(n), & L = L_0 + K, \\ \geq C, & K + 1 \leq L \leq L_0 + K. \end{cases}$$

Hence, the definition of $\hat{L}_n(A)$ in (E.55) yields $\hat{L}_n(A) = L_0 + K$. This proves the first bullet point.

Next, we consider the second bullet point. Suppose for L_1 of $\{v_1, v_2, \ldots, v_K\}$ and L_2 of $\{m_1, m_2, \ldots, m_{L_0}\}$, there are no local centers are within a distance of $c/4$ of them. When $L_1 + L_2 \geq 1$, using similar arguments as those for proving (E.70), we can see that the associated sum-of-squares is lower bounded by $C|\mathcal{M}|$. However, in (E.69), we have seen that the sum-of-squares attained by k-means is at most $C|\mathcal{M}|/\log(n)$. Hence, the above situation is impossible, i.e., for each of $\{v_1, v_2, \ldots, v_K, m_1, \ldots, m_{L_0}\}$, there is at least one local center within a distance $c/4$ to it. Since that the distance between any two of $\{v_1, v_2, \ldots, v_K, m_1, \ldots, m_{L_0}\}$ is at least c, these $(L_0 + K)$ local centers must be distinct. Noting that there are at most $\hat{L}_n(A) = L_0 + K$ cluster centers in total, we find that

$$\begin{gathered} \text{there is exactly one local center within a distance } c/4 \\ \text{to each of } \{v_1, v_2, \ldots, v_K, m_1, m_2 \ldots, m_{L_0}\}. \end{gathered} \tag{E.72}$$

Denote by $\hat{m}^*_{(k)}$ the local center nearest to v_k and by $\hat{m}_{(j)}$ the local center nearest to m_j, $1 \le k \le K$, $1 \le j \le L_0$. For any $i \in \mathcal{N}_k$, the distance from $\hat{r}_i$ to $\hat{m}^*_{(k)}$ is at most $c/4 + O(\hat{h}) \le c/3$, but its distance to any other local center is at least $c - c/4 - O(\hat{h}) \ge 2c/3$; hence, $\hat{r}_i$ can only be assigned to the cluster associated with $\hat{m}^*_{(k)}$. Similarly, for any $i \in \mathcal{M}_j$, the distance from $\hat{r}_i$ to $\hat{m}_{(j)}$ is at most $c/4 + O(\frac{1}{\log(n)}) + O(\hat{h}) \le c/3$, but the distance to any other local center is at least $c - c/4 - O(\frac{1}{\log(n)}) - O(\hat{h}) \ge 2c/3$; so $\hat{r}_i$ must be assigned to $\hat{m}_{(j)}$. We have proved that

$$\begin{cases} \text{the cluster associated with } \hat{m}^*_{(k)} \text{ is } \{\hat{r}_i : i \in \mathcal{N}_k\}, 1 \le k \le K, \\ \text{the cluster associated with } \hat{m}_{(j)} \text{ is } \{\hat{r}_i : i \in \mathcal{M}_j\}, 1 \le j \le L_0. \end{cases} \tag{E.73}$$

Then, it is easy to see that

- All the local centers are within a distance $\hat{h}$ to the Ideal Simplex.
- Each $\hat{m}^*_{(k)}$ is within a distance $C\hat{h}$ to v_k, $1 \le k \le K$.
- Each $\hat{m}_{(j)}$ is within a distance $C/\log(n)$ to m_j, $1 \le j \le L_0$.

We now show that $\hat{m}^*_{(1)}, \hat{m}^*_{(2)}, \ldots, \hat{m}^*_{(K)}$ will be selected by the combinatorial search. The proof is similar to that of Lemma E.1 but is simpler. Suppose one $\hat{m}^*_{(k)}$ is not selected by the combinatorial search. By (E.73), the other local centers are contained in the convex hull $\mathcal{H}\{\hat{r}_i : i \notin \mathcal{N}_k\}$. Hence, the estimated simplex $\hat{\mathcal{S}} \subset \mathcal{H}\{\hat{r}_i : i \notin \mathcal{N}_k\}$. We notice that the distance from e_k to the convex hull of all $\pi_i \ne e_k$ is lower bounded by a constant, as a result of the assumptions (E.53)-(E.54). Using (E.57), we know that the distance from v_k to the convex hull $\mathcal{H}\{r_i : i \notin \mathcal{N}_k\}$ is also lower bounded by a constant. Then,

$$\begin{aligned} d(\hat{m}^*_{(k)}, \hat{\mathcal{S}}) &\ge d\big(\hat{m}^*_{(k)},\ \mathcal{H}\{\hat{r}_i : i \notin \mathcal{N}_k\}\big) \\ &\ge d(v_k, \mathcal{H}\{r_i : i \notin \mathcal{N}_k\}) - O(\hat{h}) \\ &\ge C. \end{aligned}$$

At the same time, if we pick the K local centers $\hat{m}^*_{(1)}, \hat{m}^*_{(2)}, \ldots, \hat{m}^*_{(K)}$,

$$\max_{1 \le j \le L_0} d\big(\hat{m}_j, \mathcal{S}(\hat{m}^*_{(1)}, \hat{m}^*_{(2)}, \ldots, \hat{m}^*_{(K)})\big) \le C\hat{h}.$$

This yields a contradiction since $\hat{h} = o(1)$. As a result, all of $\hat{m}^*_{(1)}, \hat{m}^*_{(2)}, \ldots, \hat{m}^*_{(K)}$ will be selected by the combinatorial search.

Last, we prove the third bullet point. So far, we have seen that $\hat{v}_k = \hat{m}^*_{(k)}$ (up to a label permutation). By (E.73) and the nature of k-means solutions,

$$\hat{v}_k = |\mathcal{N}_k|^{-1} \sum_{i \in \mathcal{N}_k} \hat{r}_i, \qquad 1 \le k \le K.$$

We note that $0 \le \sum_{i\in\mathcal{N}_k} \|\hat{r}_i - \hat{v}_k\|^2 = \sum_{i\in\mathcal{N}_k} \{\|\hat{r}_i - v_k\|^2 - 2(\hat{v}_k - v_k)'(\hat{r}_i - v_k) + \|\hat{v}_k - v_k\|^2)\} = \sum_{i\in\mathcal{N}_k} \|\hat{r}_i - v_k\|^2 - |\mathcal{N}_k|\|\hat{v}_k - v_k\|^2$. As a result,

$$\|\hat{v}_k - v_k\|^2 \le \frac{1}{|\mathcal{N}_k|} \sum_{i\in\mathcal{N}_k} \|\hat{r}_i - v_k\|^2 \le \frac{1}{|\mathcal{N}_k|} \sum_{i=1}^n \|\hat{r}_i - r_i\|^2, \quad 1 \le k \le K.$$

Since $|\mathcal{N}_k| \ge c_1 n$, it follows that

$$\max_{1\le k\le K} \|\hat{v}_k - v_k\| \le C \sqrt{n^{-1} \sum_{i=1}^n \|\hat{r}_i - r_i\|^2} \le C\hat{h}^*. \tag{E.74}$$

This proves the third bullet point. □

F Rates of Convergence of Mixed-SCORE

We prove the main results about Mixed-SCORE, including Theorems 3.2-B.1.

F.1 Proofs of Theorem 3.2

Let H be the orthogonal matrix as in Theorem 3.1. We aim to show that, with probability $1 - o(n^{-3})$, for all $1 \le i \le n$,

$$\|\hat{\pi}_i - \pi_i\|_1 \le C\|H\hat{r}_i - r_i\| + C \max_{1\le k\le K} \|H\hat{v}_k - v_k\| + CKerr_n. \tag{F.75}$$

Once (F.75) is true, by efficiency of the VH algorithm (see Definition E.1) and the bound in Theorem 3.1, we immediately have that, with probability $1 - o(n^{-3})$,

$$\max_{1\le i\le n} \|\hat{\pi}_i - \pi_i\|_1 \le CK^{3/2}\beta_n^{-1} err_n. \tag{F.76}$$

Note that $\|\hat{\pi}_i - \pi_i\|^2 \le \|\hat{\pi}_i - \pi_i\|_\infty \|\hat{\pi}_i - \pi_i\|_1 \le \|\hat{\pi}_i - \pi_i\|_1^2$. It follows that $\frac{1}{n}\sum_{i=1}^n \|\hat{\pi}_i - \pi_i\|^2 \le \max_{1\le i\le n} \|\hat{\pi}_i - \pi_i\|^2 \le \max_{1\le i\le n} \|\hat{\pi}_i - \pi_i\|_1^2 \le CK^3\beta_n^{-2} err_n^2$, with probability $1 - o(n^{-3})$. Moreover, $\sum_{i=1}^n \|\hat{\pi}_i - \pi_i\|^2 \le 2$ always holds. Combining these arguments gives

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n \|\hat{\pi}_i - \pi_i^2\|\Big] \le CK^3\beta_n^{-2} err_n^2 + o(n^{-3}).$$

This proves the first claim. The second claim follows directly by noting that $err_n^2 \leq (n\bar{\theta}^2)^{-1}$ if $\theta_{\max} \leq C\theta_{\min}$.

Below, we show (F.75). In the Membership Reconstruction (MR) step, we compute $\hat{w}_i$ and $\hat{b}_1$, then use them to construct

$$\hat{\pi}_i^*(k) = \max\{0, \hat{w}_i(k)/\hat{b}_1(k)\}, \qquad 1 \leq k \leq K, \tag{F.77}$$

and then estimates π_i by $\hat{\pi}_i = \hat{\pi}_i^*/\|\hat{\pi}_i^*\|_1$. We shall study $\hat{w}_i$ and $\hat{b}_1$ separately and then combine their error bounds to get (F.75).

First, we study $\hat{w}_i$. By definition,

$$\underbrace{\begin{pmatrix} 1 & \cdots & 1 \\ v_1 & \cdots & v_K \end{pmatrix}}_{\equiv Q} w_i = \begin{pmatrix} 1 \\ r_i \end{pmatrix}, \qquad \underbrace{\begin{pmatrix} 1 & \cdots & 1 \\ H\hat{v}_1 & \cdots & H\hat{v}_K \end{pmatrix}}_{\equiv \hat{Q}} \hat{w}_i = \begin{pmatrix} 1 \\ H\hat{r}_i \end{pmatrix}. \tag{F.78}$$

We thus write

$$\begin{aligned} \hat{w}_i - w_i &= \hat{Q}^{-1}\begin{pmatrix} 1 \\ H\hat{r}_i \end{pmatrix} - Q^{-1}\begin{pmatrix} 1 \\ r_i \end{pmatrix} \\ &= \hat{Q}^{-1}\Big[\begin{pmatrix} 1 \\ H\hat{r}_i \end{pmatrix} - \begin{pmatrix} 1 \\ r_i \end{pmatrix}\Big] - (Q^{-1} - \hat{Q}^{-1})\begin{pmatrix} 1 \\ r_i \end{pmatrix} \\ &= \hat{Q}^{-1}\begin{pmatrix} 0 \\ H\hat{r}_i - r_i \end{pmatrix} - \hat{Q}^{-1}(\hat{Q} - Q)Q^{-1}\begin{pmatrix} 1 \\ r_i \end{pmatrix} \\ &= \hat{Q}^{-1}\begin{pmatrix} 0 \\ H\hat{r}_i - r_i \end{pmatrix} - \hat{Q}^{-1}(\hat{Q} - Q)w_i. \end{aligned}$$

It follows that

$$\|\hat{w}_i - w_i\| \leq \|\hat{Q}^{-1}\| \cdot \big(\|H\hat{r}_i - r_i\| + \|(\hat{Q} - Q)w_i\|\big). \tag{F.79}$$

This matrix Q is studied in the proof of Lemma C.4, where we prove $\|Q^{-1}\| = O(1/\sqrt{K})$; see (C.15). This means the minimum singular value of Q is lower bounded by $C\sqrt{K}$. Moreover, $\|\hat{Q} - Q\| \leq \|\hat{Q} - Q\|_F \leq \sqrt{K}\max_{1\leq k\leq K}\|H\hat{v}_k - v_k\| = o(\sqrt{K})$. As a result, the minimum singular value of $\hat{Q}$ is also lower bounded by $C\sqrt{K}$. It leads to

$$\|\hat{Q}^{-1}\| \leq C/\sqrt{K}.$$

We note that $(\hat{Q} - Q)w_i \in \mathbb{R}^K$ is a vector whose first entry is 0 and whose remaining entries are equal to $\sum_{k=2}^K w_i(k)(\hat{v}_k - v_k) \in \mathbb{R}^{K-1}$. Since w_i contains the coefficients of writing r_i

as a convex combination of $v_1, \ldots, v_K$, we have $\|w_i\|_1 = 1$. Therefore,

$$\|(\hat{Q} - Q)w_i\| = \Big\|\sum_{k=1}^{K} w_i(k)(H\hat{v}_k - v_k)\Big\| \le \sum_{k=1}^{K} w_i(k)\|H\hat{v}_k - v_k\| \le \max_{1\le k\le K}\|H\hat{v}_k - v_k\|.$$

Plugging in the above results into (F.79) gives

$$\|\hat{w}_i - w_i\| \le CK^{-1/2}\big(\|H\hat{r}_i - r_i\| + \max_{1\le k\le K}\|H\hat{v}_k - v_k\|\big). \tag{F.80}$$

Next, we study $\hat{b}_1$. Recall that

$$\hat{b}_1(k) = [\hat{\lambda}_1 + \hat{v}_k'\mathrm{diag}(\hat{\lambda}_2, \cdots, \hat{\lambda}_K)\hat{v}_k]^{-1/2}.$$

By Lemma 2.1, $b_1(k)$ has the same form except that $(\hat{\lambda}_k, \hat{v}_k)$ are replaced with their population counterparts. Letting $\Lambda_0 = \mathrm{diag}(\lambda_2, \cdots, \lambda_K)$ and $\hat{\Lambda}_0 = \mathrm{diag}(\hat{\lambda}_2, \cdots, \hat{\lambda}_K)$, we write

$$\frac{1}{\hat{b}_1^2(k)} = \hat{\lambda}_1 + \hat{v}_k'\hat{\Lambda}_0\hat{v}_k, \qquad \frac{1}{b_1^2(k)} = \lambda_1 + v_k'\Lambda_0 v_k.$$

By direct calculations,

$$\begin{aligned}
&|\frac{1}{\hat{b}_1^2(k)} - \frac{1}{b_1^2(k)}| \le |\hat{\lambda}_1 - \lambda_1| + |\hat{v}_k'\hat{\Lambda}_0\hat{v}_k - v_k'\Lambda_0 v_k| \\
&= |\hat{\lambda}_1 - \lambda_1| + |\hat{v}_k'H'H\hat{\Lambda}_0\hat{v}_k - v_k'\Lambda_0 v_k| \\
&\le |\hat{\lambda}_1 - \lambda_1| + |\hat{v}_k'H'\hat{\Lambda}_0 H\hat{v}_k - v_k'\hat{\Lambda}_0 v_k| + |\hat{v}_k'H'(H\hat{\Lambda}_0 - \hat{\Lambda}_0 H)\hat{v}_k| + |v_k'(\hat{\Lambda}_0 - \Lambda_0)v_k| \\
&\le |\hat{\lambda}_1 - \lambda_1| + |\hat{v}_k'H'\hat{\Lambda}_0 H\hat{v}_k - v_k'\hat{\Lambda}_0 v_k| + \|\hat{v}_k\|^2\|H\hat{\Lambda}_0 - \hat{\Lambda}_0 H\| + \|v_k\|^2\|\hat{\Lambda}_0 - \Lambda_0\| \\
&\le (1 + \|v_k\|^2)|\max_{\ell}|\hat{\lambda}_\ell - \lambda_\ell| + |\hat{v}_k'H'\hat{\Lambda}_0 H\hat{v}_k - v_k'\hat{\Lambda}_0 v_k| + \|\hat{v}_k\|^2\|H\hat{\Lambda}_0 - \hat{\Lambda}_0 H\|.
\end{aligned}$$

First, by Lemma D.1, $\max_\ell|\hat{\lambda}_\ell - \lambda_\ell| \le C\sqrt{\theta_{\max}\|\theta\|_1}$. Second, by Lemma D.6, $\|H\hat{\Lambda}_0 - \hat{\Lambda}_0 H\| \le C\sqrt{\theta_{\max}\|\theta\|_1}$. Third, by Lemma C.4, $\|v_k\| \le C\sqrt{K}$; since $\max_\ell\|\hat{v}_\ell - v_\ell\| = o(\sqrt{K})$, it follows that $\|\hat{v}_k\| \le C\sqrt{K}$. Combining the above gives

$$|\frac{1}{\hat{b}_1^2(k)} - \frac{1}{b_1^2(k)}| \le |\hat{v}_k'H'\hat{\Lambda}_0 H\hat{v}_k - v_k'\hat{\Lambda}_0 v_k| + CK\sqrt{\theta_{\max}\|\theta\|_1}. \tag{F.81}$$

Since $\hat{v}_k'H'\hat{\Lambda}_0 H\hat{v}_k = v_k'\hat{\Lambda}_0 v_k + 2v_k'\hat{\Lambda}_0(H\hat{v}_k - v_k) + (H\hat{v}_k - v_k)'\hat{\Lambda}_0(H\hat{v}_k - v_k)$, we have

$$|\hat{v}_k'H'\hat{\Lambda}_0 H\hat{v}_k - v_k'\hat{\Lambda}_0 v_k| \le 2\|v_k\|\|\hat{\Lambda}_0\|\|H\hat{v}_k - v_k\| + \|\hat{\Lambda}_0\|\|H\hat{v}_k - v_k\|^2.$$

By Lemma C.2 and Lemma D.1, $\|\Lambda_0\| \le C\beta_n K^{-1}\|\theta\|^2$ and $\|\hat{\Lambda}_0 - \Lambda_0\| \le C\sqrt{\theta_{\max}\|\theta\|_1} = o(K\beta_n^{-1}\|\theta\|^2)$. It follows that $\|\hat{\Lambda}_0\| \le C\beta_n K^{-1}\|\theta\|^2$. Also, as we have argued before,

$\|v_k\| \le C\sqrt{K}$ and $\|H\hat{v}_k - v_k\| = o(\sqrt{K})$. Plugging these results into the above inequality gives

$$|\hat{v}_k' H' \hat{\Lambda}_0 H \hat{v}_k - v_k' \hat{\Lambda}_0 v_k| \le CK^{-1/2}\beta_n \|\theta\|^2 \|H\hat{v}_k - v_k\|.$$

We then plug it into (F.81) to get

$$|\frac{1}{\hat{b}_1^2(k)} - \frac{1}{b_1^2(k)}| \le CK^{-1/2}\beta_n\|\theta\|^2\|H\hat{v}_k - v_k\| + CK\sqrt{\theta_{\max}\|\theta\|_1}. \tag{F.82}$$

In the proof of Lemma C.3, we have shown $b_1(k) \asymp \|\theta\|^{-1}$; see (C.11). Then, $\frac{1}{b_1^2(k)} \asymp \|\theta\|^2$. Combining it with (F.82), we have $\frac{1}{\hat{b}_1^2(k)} = \frac{1}{b_1^2(k)}[1 + o(1)] \asymp \|\theta\|^2$. It follows that

$$\begin{aligned}
|\frac{1}{\hat{b}_1(k)} - \frac{1}{b_1(k)}| &= |\frac{1}{\hat{b}_1(k)} + \frac{1}{b_1(k)}|^{-1} \cdot |\frac{1}{\hat{b}_1^2(k)} - \frac{1}{b_1^2(k)}| \\
&\le C\|\theta\|^{-1} \cdot |\frac{1}{\hat{b}_1^2(k)} - \frac{1}{b_1^2(k)}| \\
&\le CK^{-1/2}\beta_n\|\theta\|\|H\hat{v}_k - v_k\| + C\|\theta\|^{-1}K\sqrt{\theta_{\max}\|\theta\|_1} \\
&\le CK^{-1/2}\beta_n\|\theta\|\|H\hat{v}_k - v_k\| + CK\|\theta\| err_n,
\end{aligned} \tag{F.83}$$

where the last line is because $err_n = (\theta_{\max}/\theta_{\min}) \cdot \|\theta\|^{-2}\sqrt{\theta_{\max}\|\theta\|_1 \log(n)} \gg \|\theta\|^{-2}\sqrt{\theta_{\max}\|\theta\|_1}$.

Last, we combine the results for $(\hat{w}_i, \hat{b}_1)$ to prove (F.75). Recall that $\hat{\pi}_i^*$ is as defined in (F.77). Introduce its non-stochastic counterpart π_i^* by

$$\pi_i^*(k) = w_i(k)/b_1(k), \qquad 1 \le k \le K. \tag{F.84}$$

Since $\pi_i^*(k) \ge 0$, in (F.77), the operation of truncating at zero can only make it closer to $\pi_i^*(k)$. It follows that

$$\begin{aligned}
|\hat{\pi}_i^*(k) - \pi_i^*(k)| &\le |\hat{w}_i(k)/\hat{b}_1(k) - \pi_i^*(k)| \\
&= |\hat{w}_i(k)/\hat{b}_1(k) - w_i(k)/b_1(k)| \\
&\le \frac{1}{\hat{b}_1(k)}|\hat{w}_i(k) - w_i(k)| + w_i(k)|\frac{1}{\hat{b}_1(k)} - \frac{1}{b_1(k)}|.
\end{aligned} \tag{F.85}$$

We sum over k on both sides and note that $\hat{b}_1(k) \asymp \|\theta\|^{-1}$ (see the paragraph above (F.83)) and $\|w_i\|_1 = 1$. It yields

$$\begin{aligned}
\|\hat{\pi}_i^* - \pi_i^*\|_1 &\le C\|\theta\|\|\hat{w}_i - w_i\|_1 + |\frac{1}{b_1(k)} - \frac{1}{\hat{b}_1(k)}| \\
&\le C\|\theta\|\sqrt{K}\|\hat{w}_i - w_i\| + \max_{1\le k\le K}|\frac{1}{b_1(k)} - \frac{1}{\hat{b}_1(k)}| \\
&\le C\|\theta\|\big(\|H\hat{r}_i - r_i\| + \max_{1\le k\le K}\|H\hat{v}_k - v_k\| + Kerr_n\big),
\end{aligned} \tag{F.86}$$

where in the second line we have used Cauchy-Schwarz inequality and in the last line we have plugged in (F.80) and (F.83). By definition, $\hat{\pi}_i = \hat{\pi}_i^*/\|\hat{\pi}_i^*\|_1$. By the triangular inequality,

$$\begin{aligned}|\hat{\pi}_i(k) - \pi_i(k)| &\le \frac{1}{\|\pi_i^*\|_1}|\hat{\pi}_i^*(k) - \pi_i^*(k)| + \hat{\pi}_i^*(k)|\frac{1}{\|\hat{\pi}_i^*\|_1} - \frac{1}{\|\pi_i^*\|_1}| \\ &= \frac{1}{\|\pi_i^*\|_1}|\hat{\pi}_i^*(k) - \pi_i^*(k)| + \frac{\hat{\pi}_i(k)}{\|\pi_i^*\|_1}|\|\hat{\pi}_i^*\|_1 - \|\pi_i^*\|_1| \\ &\le \frac{1}{\|\pi_i^*\|_1}\big(|\hat{\pi}_i^*(k) - \pi_i^*(k)| + \hat{\pi}_i(k)\|\hat{\pi}_i^* - \pi_i^*\|_1\big), \end{aligned} \tag{F.87}$$

where the last inequality is because $|\|\hat{\pi}_i^*\|_1 - \|\pi_i^*\|_1| \le \|\hat{\pi}_i^* - \pi_i^*\|_1$. We sum over k on both sides and note that $\sum_k \hat{\pi}_i(k) = 1$ by definition. It follows that

$$\|\hat{\pi}_i - \pi_i\|_1 \le \frac{1}{\|\pi_i^*\|_1} \cdot 2\|\hat{\pi}_i^* - \pi_i^*\|_1.$$

By (F.84), $\|\pi^*\|_1 \ge \|w_i\|_1 \cdot \min_k \frac{1}{b_1(k)}$. In the paragraph above (F.83), we have seen that $b_1(k) \asymp \|\theta\|^{-1}$. This suggests that $\|\pi_i^*\|_1 \ge C\|\theta\|$. As a result,

$$\begin{aligned}\|\hat{\pi}_i - \pi_i\|_1 &\le C\|\theta\|^{-1} \cdot \|\hat{\pi}_i^* - \pi_i^*\|_1 \\ &\le C\big(\|H\hat{r}_i - r_i\| + \max_{1\le k\le K}\|H\hat{v}_k - v_k\| + Kerr_n\big). \end{aligned} \tag{F.88}$$

This gives (F.75). The proof is now complete. □

F.2 Proof of Theorem 3.3

First, consider $\hat{P} - P$. Let Q and $\hat{Q}$ be the same as in (F.78). Then,

$$P = \mathrm{diag}(b_1)Q'\Lambda Q\mathrm{diag}(b_1), \qquad \hat{P} = \mathrm{diag}(\hat{b}_1)\hat{Q}'\hat{\Lambda}\hat{Q}\mathrm{diag}(\hat{b}_1).$$

It follows that

$$\begin{aligned}\|\hat{P} - P\| \le{}& \|\hat{Q}\mathrm{diag}(\hat{b}_1)\|^2\|\hat{\Lambda} - \Lambda\| + \|\hat{Q}\mathrm{diag}(\hat{b}_1) - Q\mathrm{diag}(b_1)\|\|\Lambda\|\|\hat{Q}\mathrm{diag}(\hat{b}_1)\| \\ &+ \|Q\mathrm{diag}(b_1)\|\|\Lambda\|\|\hat{Q}\mathrm{diag}(\hat{b}_1) - Q\mathrm{diag}(b)_1\|. \end{aligned} \tag{F.89}$$

Recall that we have the following facts (they hold with probability $1 - o(n^{-3})$):

- $\|\Lambda\| \le C\|\theta\|^{-1}$ (by Lemma C.2); $\|\hat{\Lambda} - \Lambda\| \le C\sqrt{\theta_{\max}\|\theta\|} \ll \|\theta\|^2 err_n$ (by Lemma D.1 and the definition of err_n).
- $\|Q\| \le C\sqrt{K}$ (by Lemma C.4); $\|\hat{Q} - \hat{Q}\| \le C\sqrt{K}\max_{1\le k\le K}\|H\hat{v}_k - v_k\| \le CK^2\beta_n^{-1}err_n$ (by Theorem 3.1 and the definitions of Q and $\hat{Q}$).

- $C^{-1}\|\theta\|^{-1} \le b_1(k) \le C\|\theta\|^{-1}$, for $1 \le k \le K$ (by (C.11) in the proof of Lemma C.3); $|\frac{1}{\hat{b}_1(k)} - \frac{1}{b_1(k)}| \le CK^{-1/2}\beta_n\|\theta\|\|H\hat{v}_k - v_k\| + CK\|\theta\|err_n \le CK\|\theta\|err_n$ (by (F.83) in the proof of Theorem 3.2).

From the third bullet point, $|\hat{b}_1(k) - b_1(k)| \le C\|\theta\|^{-2}|\frac{1}{\hat{b}_1(k)} - \frac{1}{b_1(k)}| \le CK\|\theta\|^{-1}err_n$. From the second bullet point, $\|\hat{Q} - Q\| \le CK^2\beta_n^{-1}err_n$, and $\|\hat{Q}\| \le 2\|Q\| \le C\sqrt{K}$. As a result,

$$\begin{aligned}\|\hat{Q}\mathrm{diag}(\hat{b}_1) - Q\mathrm{diag}(b)_1\| &\le \|\hat{Q}\|\|\mathrm{diag}(\hat{b}_1) - \mathrm{diag}(b_1)\| + \|\hat{Q} - Q\|\|\mathrm{diag}(b_1)\| \\ &\le C\sqrt{K}\cdot K\|\theta\|^{-1}err_n + CK^2\beta_n^{-1}err_n\cdot\|\theta\|^{-1} \\ &\le C(K^{3/2} + K^2\beta_n^{-1})\|\theta\|^{-1}err_n. \end{aligned} \tag{F.90}$$

It further implies $\|\hat{Q}\mathrm{diag}(\hat{b}_1)\| \le 2\|Q\mathrm{diag}(b)_1\| \le C\sqrt{K}\|\theta\|^{-1}$. We then plug these results into (F.89) and use the first bullet point above. It gives

$$\begin{aligned}\|\hat{P} - P\| &\le \|\hat{Q}\mathrm{diag}(\hat{b}_1)\|^2\|\hat{\Lambda} - \Lambda\| + 3\|\hat{Q}\mathrm{diag}(\hat{b}_1) - Q\mathrm{diag}(b_1)\|\|\Lambda\|\|\hat{Q}\mathrm{diag}(\hat{b}_1)\| \\ &\le C(\sqrt{K}\|\theta\|^{-1})^2\cdot\|\theta\|^2err_n + C(K^{3/2} + K^2\beta_n^{-1})\|\theta\|^{-1}err_n\cdot\|\theta\|^2\cdot\sqrt{K}\|\theta\|^{-1} \\ &\le C(K^2 + K^{3/2}\beta_n^{-1})err_n. \end{aligned} \tag{F.91}$$

This proves the first claim.

Second, consider $\|\hat{\Theta} - \Theta\|_F^2$, which by definition is equal to $\sum_{i=1}^n |\hat{\theta}(i) - \theta(i)|^2$. Recall that $\theta(i) = \xi_1(i)/(\pi_i'b_1)$ and $\hat{\theta}(i) = \hat{\xi}_1(i)/(\hat{\pi}_i'\hat{b}_1)$. It follows that

$$\begin{aligned}|\hat{\theta}(i) - \theta(i)| &\le \frac{1}{|\pi_i'b_1|}|\hat{\xi}_1(i) - \xi_1(i)| + |\hat{\xi}_1(i)|\Big|\frac{1}{\hat{\pi}_i'\hat{b}_1} - \frac{1}{\pi_i'b_1}\Big| \\ &\le \frac{1}{|\pi_i'b_1|}|\hat{\xi}_1(i) - \xi_1(i)| + |\hat{\xi}_1(i)|\cdot\frac{|\hat{\pi}_i'\hat{b}_1 - \pi_i'b_1|}{|\hat{\pi}_i'\hat{b}_1||\pi_i'b_1|} \\ &\le \frac{|\hat{\xi}_1(i) - \xi_1(i)|}{|\pi_i'b_1|} + \frac{|\hat{\xi}_1(i)|}{|\hat{\pi}_i'\hat{b}_1||\pi_i'b_1|}\big(\|\hat{\pi}_i - \pi_i\|_1\|b_1\|_\infty + \|\hat{\pi}_i\|_1\|\hat{b}_1 - b_1\|_\infty\big). \end{aligned}$$

Note that $\|\hat{\pi}_i\|_1 = 1$, $b_1(k) \asymp \|\theta\|^{-1}$, and $\|\hat{b}_1 - b_1\|_\infty \le CK\|\theta\|^{-1}err_n = o(\|\theta\|^{-1})$. It further implies $\pi_i'b_1 \asymp \hat{\pi}_i'\hat{b}_1 \asymp \|\theta\|^{-1}$. We plug these results into the above inequality to get

$$|\hat{\theta}(i) - \theta(i)| \le C\|\theta\||\hat{\xi}_1(i) - \xi_1(i)| + C\|\theta\||\hat{\xi}_1(i)|\|\hat{\pi}_i - \pi_i\|_1 + CK\|\theta\|err_n|\hat{\xi}_1(i)|.$$

We take the sum of squares of $i = 1, 2, \ldots, n$ on both sides and note that $\|\hat{\xi}\| = 1$. Moreover, by Lemma D.2, $\|\hat{\xi}_1 - \xi_1\| \le C\|\theta\|^{-2}K\sqrt{\theta_{\max}\|\theta\|_1} \ll Kerr_n$. It follows that

$$\|\hat{\Theta} - \Theta\|_F^2 \le C\|\theta\|^2\|\hat{\xi}_1 - \xi_1\|^2 + C\|\theta\|^2\Big(\max_{1\le i\le n}\|\hat{\pi}_i - \pi_i\|_1^2\Big) + CK^2\|\theta\|^2err_n^2$$

$$\le C\|\theta\|^2\big(K^2 err_n^2 + K^3\beta_n^{-2} err_n^2 + CK^2 err_n^2\big)$$
$$\le \|\theta\|^2 \cdot CK^3\beta_n^{-2} err_n^2. \tag{F.92}$$

This proves the second claim. □

F.3 Proofs of Theorems 3.4, 3.5 and B.1

Theorem 3.4 is a direct consequence of Theorem 3.2 and Lemma E.1. For Theorem 3.5 and Theorem B.1, their first claims about the VH step follow from Lemma E.3 and Lemma E.4, respectively. We now show their second claims, where we aim to obtain a faster rate for $\frac{1}{n}\sum_{i=1}^n \|\hat\pi_i - \pi_i\|^2$ when the VH step is strongly efficient.

In (F.87), we have shown that for every $1 \le k \le K$,

$$|\hat\pi_i(k) - \pi_i(k)| \le \frac{1}{\|\pi_i^*\|_1}\big(|\hat\pi_i^*(k) - \pi_i^*(k)| + \hat\pi_i(k)\|\hat\pi_i^* - \pi_i^*\|_1\big).$$

Taking the sum of squares over k on both sides and using the universal inequality $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\|\hat\pi_i - \pi_i\|^2 \le \frac{2}{\|\pi_i^*\|_1^2}\big(\|\hat\pi_i^* - \pi_i^*\|^2 + \|\hat\pi_i\|^2 \cdot \|\hat\pi_i^* - \pi_i^*\|_1^2\big).$$

In the paragraph above (F.88), we have shown that $\|\pi_i^*\|_1 \ge C\|\theta\|$. Additionally, $\|\hat\pi_i\|^2 \le \|\hat\pi_i\|_1\|\hat\pi_i\|_\infty \le 1$. It follows that

$$\|\hat\pi_i - \pi_i\|^2 \le \frac{C}{\|\theta\|^2}\big(\|\hat\pi_i^* - \pi_i^*\|^2 + \|\hat\pi_i^* - \pi_i^*\|_1^2\big). \tag{F.93}$$

In light of (F.93), we first derive upper bounds for $\|\hat\pi_i^* - \pi_i^*\|$ and $\|\hat\pi_i^* - \pi_i^*\|_1$, respectively. By (F.85) and (F.83),

$$|\hat\pi_i^*(k) - \pi_i^*(k)| \le \frac{1}{\hat b_1(k)}|\hat w_i(k) - w_i(k)| + w_i(k)\Big|\frac{1}{\hat b_1(k)} - \frac{1}{b_1(k)}\Big|,$$
$$\Big|\frac{1}{\hat b_1(k)} - \frac{1}{b_1(k)}\Big| \le CK^{-1/2}\beta_n\|\theta\|\|H\hat v_k - v_k\| + C\|\theta\|^{-1}K\sqrt{\theta_{\max}\|\theta\|_1}.$$

Also, $\hat b_1(k) \asymp b_1(k) \asymp \|\theta\|^{-1}$ (see the paragraph above (F.83)). It follows that

$$|\hat\pi_i^*(k) - \pi_i^*(k)| \le C\|\theta\|\,|\hat w_i(k) - w_i(k)| + Cw_i(k)\bigg(\frac{\beta_n\|\theta\|\|H\hat v_k - v_k\|}{\sqrt{K}} + \frac{K\sqrt{\theta_{\max}\|\theta\|_1}}{\|\theta\|}\bigg).$$

Note that

$$err_n^* = [\|\theta\|/(\theta_{\min}\sqrt{n})] \cdot \|\theta\|^{-2}\sqrt{\theta_{\max}\|\theta\|_1} \ge \|\theta\|^{-2}\sqrt{\theta_{\max}\|\theta\|_1}.$$

We further have

$$|\hat{\pi}_i^*(k)-\pi_i^*(k)| \le C\|\theta\||\hat{w}_i(k)-w_i(k)|+Cw_i(k)\|\theta\|\Big(K^{-1/2}\beta_n\|H\hat{v}_k-v_k\|+Kerr_n^*\Big). \quad \text{(F.94)}$$

It follows that

$$\|\hat{\pi}_i^*-\pi_i^*\|^2 \le C\|\theta\|^2\Big[\|\hat{w}_i-w_i\|^2+\|w_i\|^2\Big(K^{-1}\beta_n^2\max_{1\le k\le K}\|H\hat{v}_k-v_k\|^2+K^2(err_n^*)^2\Big)\Big],$$
$$\|\hat{\pi}_i^*-\pi_i^*\|_1 \le C\|\theta\|\Big[\|\hat{w}_i-w_i\|_1+\|w_i\|_1\Big(K^{-1/2}\beta_n\max_{1\le k\le K}\|H\hat{v}_k-v_k\|+Kerr_n^*\Big)\Big].$$

Note that $\|w_i\|_1=1$, $\|w_i\|^2 \le \|w_i\|_1\|w_i\|_\infty \le 1$, and $\|\hat{w}_i-w_i\|_1 \le \sqrt{K}\|\hat{w}_i-w_i\|$. Additionally, by (F.80),

$$\|\hat{w}_i-w_i\| \le CK^{-1/2}\big(\|H\hat{r}_i-r_i\|+\max_{1\le k\le K}\|H\hat{v}_k-v_k\|\big).$$

Combining the above gives

$$\|\hat{\pi}_i^*-\pi_i^*\|^2 \le C\|\theta\|^2\Big(K^{-1}\|H\hat{r}_i-r_i\|^2+K^{-1}\max_{1\le k\le K}\|H\hat{v}_k-v_k\|^2+K^2(err_n^*)^2\Big),$$
$$\|\hat{\pi}_i^*-\pi_i^*\|_1 \le C\|\theta\|\Big(\|H\hat{r}_i-r_i\|+\max_{1\le k\le K}\|H\hat{v}_k-v_k\|+Kerr_n^*\Big). \quad \text{(F.95)}$$

Next, we plug (F.95) into (F.93) to get

$$\|\hat{\pi}_i-\pi_i\|^2 \le C\|H\hat{r}_i-r_i\|^2+C\Big(\max_{1\le k\le K}\|H\hat{v}_k-v_k\|\Big)^2+CK^2(err_n^*)^2.$$

Summing over i on both sides gives

$$n^{-1}\sum_{i=1}^n\|\hat{\pi}_i-\pi_i\|^2 \le Cn^{-1}\sum_{i=1}^n\|H\hat{r}_i-r_i\|^2+C\Big(\max_{1\le k\le K}\|H\hat{v}_k-v_k\|\Big)^2+CK^2(err_n^*)^2.$$

By strong efficiency of the VH step, $\max_{1\le k\le K}\|H\hat{v}_k-v_k\| \le \sqrt{n^{-1}\sum_{i=1}^n\|H\hat{r}_i-r_i\|^2}$ (see Definition E.1). It follows that

$$n^{-1}\sum_{i=1}^n\|\hat{\pi}_i-\pi_i\|^2 \le Cn^{-1}\sum_{i=1}^n\|H\hat{r}_i-r_i\|^2+CK^2(err_n^*)^2.$$

Using Lemma D.5, $n^{-1}\sum_{i=1}^n\|H\hat{r}_i-r_i\|^2 \le CK^3\beta_n^{-2}(err_n^*)^2$. Therefore,

$$n^{-1}\sum_{i=1}^n\|\hat{\pi}_i-\pi_i\|^2 \le CK^3\beta_n^{-2}(err_n^*)^2+CK^2(err_n^*)^2 \le CK^3\beta_n^{-2}(err_n^*)^2.$$

Additionally, $err_n^*=[\|\theta\|/(\sqrt{n}\theta_{\max})]\cdot err_n/\sqrt{\log(n)} \le err_n/\sqrt{\log(n)}$. We thus have

$$n^{-1}\sum_{i=1}^n\|\hat{\pi}_i-\pi_i\|_1^2 \le CK^3\beta_n^{-1}(err_n^*)^2 \le \frac{CK^3\beta_n^{-1}err_n^2}{\log(n)}. \quad \text{(F.96)}$$

This proves the claim.

G More Simulation Results

We present additional simulation results. They are not included in the main article due to space limit. For most experiments below, we set $n = 500$ and $K = 3$. For $0 \leq n_0 \leq 160$, let each community have n_0 number of pure nodes. Fixing $x \in (0, 1/2)$, let the mixed nodes have four different memberships $(x, x, 1-2x)$, $(x, 1-2x, x)$, $(1-2x, x, x)$ and $(1/3, 1/3, 1/3)$, each with $(500-3n_0)/4$ number of nodes. Fixing $\rho \in (0, 1)$, the matrix P has diagonals 1 and off-diagonals ρ. Fixing $z \geq 1$, we generate the degree parameters such that $1/\theta(i) \overset{iid}{\sim} U(1, z)$, where $U(1, z)$ denotes the uniform distribution on $[1, z]$. The tuning parameter L is selected as in (2.8). For each setting, we report $n^{-1} \sum_{i=1}^{n} \|\hat{\pi}_i - \pi_i\|^2$ averaged over 100 repetitions.

Experiment 5: Connectivity across communities. Fix $(x, n_0, z) = (0.4, 80, 5)$ and let ρ range in $\{0.05, 0.1, 0.15, \cdots, 0.5\}$. The larger ρ, the more edges across different communities. The results are presented in Figure 1. We see that the performance of Mixed-SCORE improves as ρ decreases. One possible reason is that, for ρ large, it is relatively more difficult to identify the vertices of the Ideal Simplex. Furthermore, Mixed-SCORE is better than OCCAM in all settings.

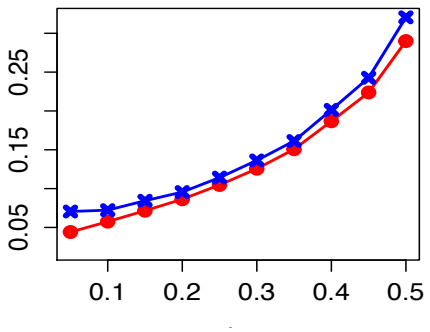

Figure 1: Estimation errors of Mixed-SCORE and OCCAM (y-axis: $n^{-1} \sum_{i=1}^{n} \|\hat{\pi}_i - \pi_i\|^2$).

Experiment 6: Mixed memberships taking continuous values. In this experiment, we generate the mixed memberships from a continuous distribution. Set $(n, K) = (500, 3)$ and let P have diagonals 1 and off-diagonals 0.3. Each community has $n_0 = 25$ pure nodes. The π_i of remaining nodes are *iid* drawn as follows: We generate $\pi_i(1)$ and $\pi_i(2)$ independently from $U(1/6, 1/2)$ and set $\pi_i(3) = 1 - \pi_i(1) - \pi_i(2)$. The degree parameters $\theta(i)$ are *iid* drawn from $\alpha_n \cdot U(1, 2)$, where $\alpha_n > 0$ controls the sparsity of the network. Let α_n range in $\{0.02, 0.04, 0.06, \cdots, 0.20\}$. The results are presented in Table 1. This setting does not satisfy the regularity conditions (E.53)-(E.54) on π_i's, however, Mixed-SCORE

still has a good performance and outperforms OCCAM. It suggests that the regularity conditions on π_i's are only for theoretical convenience, and our method indeed works for broader settings.

Table 1: Estimation errors in Experiment 6, where π_i's take continuous values.

α_n	0.02	0.04	0.06	0.08	0.10	0.12	0.14	0.16	0.18	0.20
Mixed-SCORE	.38	.35	.36	.32	.30	.28	.23	.18	.15	.12
OCCAM	.44	.42	.41	.41	.38	.36	.32	.28	.26	.23

Experiment 7: Tuning parameter selection. We first study the choice of the tuning parameter L in Mixed-SCORE. We aim to see (i) how the estimation errors change for a range of L, and (ii) how the adaptive choice $\hat{L}_n^*(A)$ in (2.8) performs. Fix $(x, \rho, z) = (0.4, 0.2, 5)$ and let n_0 range in $\{60, 80, 100\}$. For each setting, we run Mixed-SCORE with $L \in \{4, 5, \cdots, 9\}$ and $\hat{L}_n^*(A)$. The results are displayed in Figure 2. First, when there are relatively few mixed nodes (e.g., $n_0 = 100$), small values of L yield good performance; but as the number of mixed nodes going up, we favor larger values of L; these match our theoretical results (Lemmas E.3-E.4). Second, under the circumstances of a moderate number of mixed nodes (e.g., $n_0 = 60, 80$), for a range of L (e.g., $L \in \{7, 8, 9\}$), the statistical errors of Mixed-SCORE are similar, and $\hat{L}_n^*(A)$ falls in this range with high probability. Figure 3 shows the estimated 2-simplex in one repetition ($n_0 = 80$), and the simplex changes very little when L falls in a range.

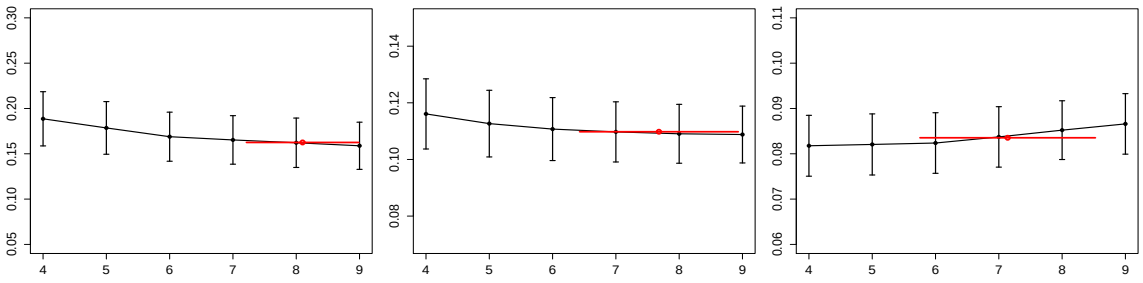

Figure 2: Performance of Mixed-SCORE as the tuning parameter L varies (y-axis: estimation errors; $\hat{L}_n^*(A)$ is plotted in red; both mean and standard deviation are displayed). From left to right, there are $60, 80, 100$ pure nodes in each community, respectively.

Experiment 8: Comparison with latent space approach. We compare Mixed-SCORE with the Bayesian method based on LPC [6] (we use the R package *latentnet*).

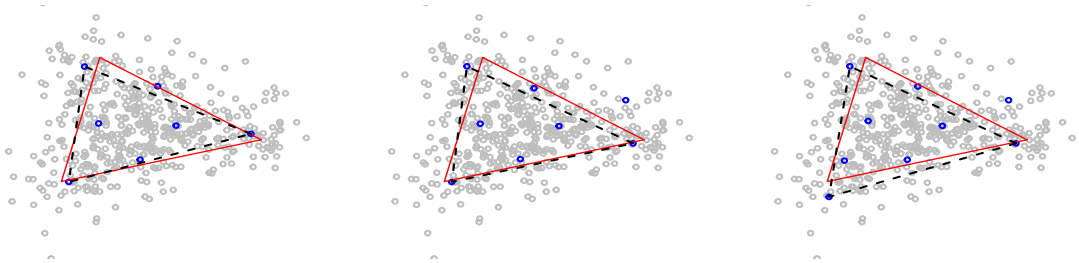

Figure 3: Illustration of the Vertex Hunting step. From left to right, $L = 7, 8, 9$. Although the local cluster centers (blue points) are different, the estimated 2-simplex (dashed black) changes very little, and it approximates the IS (solid red) well.

In this experiment, we fix $n = 120$, $K = 3$, $(x, \rho, z) = (0.4, 0.3, 5)$, and let n_0 range in $\{12, 16, 20, \cdots, 32, 36\}$ (so the number of mixed nodes in each group decreases from 21 to 3). The results are displayed in Figure 4. We find that, when the fraction of mixed nodes is comparably small, LPC has a perfect performance; however, as the fraction of mixed nodes increases to more than 40%, the performance of LPC deteriorates rapidly; one reason is that, when n_0 is not very large, LPC often estimates the PMF of all the nodes as the same. In contrast, the performance of Mixed-SCORE is quite stable. In terms of computing time, Mixed-SCORE takes only seconds for one repetition while LPC takes > 20 minutes (both measured in R).

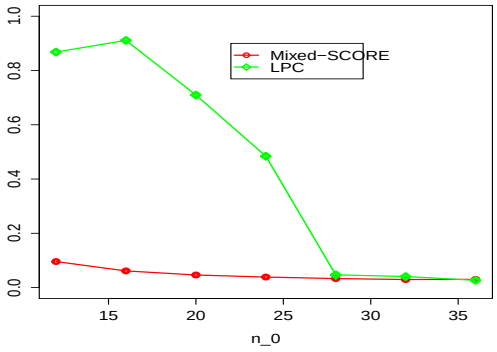

Figure 4: Estimation errors of Mixed-SCORE and LPC (y-axis: $n^{-1}\sum_{i=1}^{n}\|\hat{\pi}_i - \pi_i\|^2$).

H More Real Data Results

We present additional results for the trade networks. First, we plot the rows of $\hat{R}$ for the GOS network (see Figure 6a for a comparison). Recall that edges in the GOS network indi-

cate significant over-estimation of trade flows in the initial gravity model. This embedding is not as informative as the embedding we obtained for the GUS network. One interesting observation is that countries with high GDPs tend to cluster together and countries with low GDPs tend to cluster together.

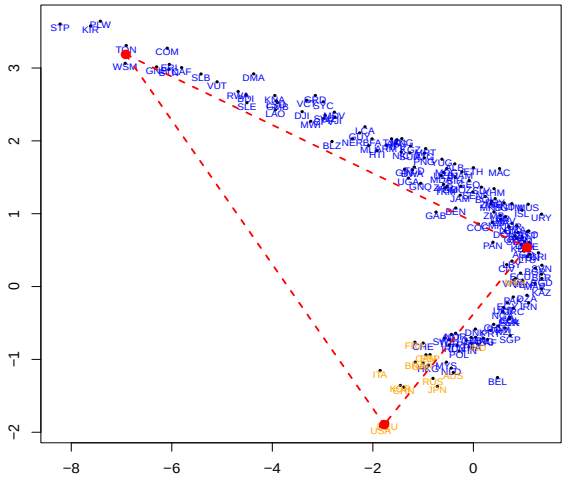

Figure 5: Rows of $\hat{R}$ for the GOS network after fitting a gravity model. We set $K = 3$ in Mixed-SCORE, so the Ideal Simplex is a triangle. Each $\hat{r}_i$ corresponds to a country, whose ISO3 code is shown (orange color: top 15 countries with highest GDPs). In each plot, the dashed triangle is the estimated simplex from SVS with $L = 40$. We note that although each r_i is in the Ideal Simplex, some $\hat{r}_i$'s can be outside the estimated simplex due to noise corruption.

Next, we present the estimated mixed memberthips of representative countries in the trade in service (TIS) network.

Table 2: The estimated $\hat{\pi}_i$ for the 10 countries with largest total service exports. By Figure 6b, the three communities are interpreted as 'North Africa', 'Southeast Asia' and 'South/Central Europe'.

Economy	Service export	degree	$\hat{\pi}_i(1)$	$\hat{\pi}_i(2)$	$\hat{\pi}_i(3)$
USA	3,998,419	45	0.128	0.424	0.448
UK	1,914,255	34	0.202	0.319	0.479
Germany	1,534,393	29	0.348	0.215	0.436
France	1,354,407	26	0.243	0.193	0.564
China	1,146,845	14	0.130	0.606	0.264
Netherlands	1,064,165	19	0.218	0.215	0.567
Japan	882,650	17	0.124	0.611	0.265
India	865,543	6	0.033	0.598	0.369
Singapore	830,975	20	0.313	0.554	0.134
Ireland	811,105	12	0.144	0.269	0.586

We also present additional results for the citee network. The following table shows those "high-degree and relatively pure" nodes in each of the three communities.

Table 3: Estimated PMF of the 100 nodes with the highest degrees in the Citee network, among which only the 12 purist nodes in each community are reported.

Name	Deg.	MulTest	SpatNon	VarSelect	Name	Deg.	MulTest	SpatNon	VarSelect	Name	Deg.	MulTest	SpatNon	VarSelect
Felix Abramovich	366	0.943	0	0.057	Peter Muller	429	0.326	0.613	0.061	Lixing Zhu	432	0.121	0	0.879
Joseph Romano	377	0.868	0	0.132	Jeffrey Morris	452	0.146	0.519	0.335	Zhiliang Ying	382	0.107	0.027	0.866
Sara van de Geer	372	0.834	0	0.166	Michael Jordan	383	0.321	0.495	0.184	Zhezhen Jin	361	0.134	0	0.866
Yoav Benjamini	478	0.821	0	0.179	Mahlet Tadesse	383	0.373	0.493	0.134	Dennis Cook	424	0.253	0	0.747
David Donoho	484	0.819	0	0.181	Naijun Sha	383	0.373	0.493	0.134	Wenbin Lu	405	0.255	0	0.745
Christopher Genovese	521	0.810	0	0.190	Michael Stein	379	0.093	0.449	0.458	Dan Yu Lin	527	0.257	0	0.743
Larry Wasserman	535	0.800	0	0.200	Adrian Raftery	413	0.175	0.446	0.379	Donglin Zeng	489	0.270	0	0.730
Jon Wellner	387	0.798	0.05	0.152	Robert Kohn	429	0.310	0.428	0.262	Gerda Claeskens	404	0.247	0.033	0.720
Alexandre Tsybakov	521	0.784	0	0.216	George Casella	430	0.303	0.425	0.271	Yingcun Xia	358	0.302	0	0.698
Jiashun Jin	441	0.780	0	0.220	Marina Vannucci	571	0.304	0.418	0.278	Naisyin Wang	586	0.283	0.043	0.674
Yingying Fan	410	0.741	0	0.259	Bernard Silverman	577	0.514	0.395	0.091	Hua Liang	509	0.334	0	0.666
John Storey	544	0.737	0	0.263	Catherine Sugar	501	0.450	0.360	0.190	Wolfgang Karl Hardle	456	0.343	0	0.657

I Using Mixed-SCORE for the Estimation of Ω

In Remark 9 of Section 5.1, we mentioned that Mixed-SCORE can be used to estimate Ω, where we let $\hat{\Omega} = \hat{\Theta}\hat{\Pi}\hat{P}\hat{\Pi}'\hat{\Theta}$ by using $\hat{\Pi}$ from Mixed-SCORE and $(\hat{\Theta}, \hat{P})$ in Section 2.4. Alternatively, we may also estimate Ω by the standard PCA, where $\hat{\Omega} = \sum_{k=1}^{K} \hat{\lambda}_k \hat{\xi}_k \hat{\xi}_k$. The following simulation results suggest that the $\hat{\Omega}$ by Mixed-SCORE is much better than the $\hat{\Omega}$ by standard PCA.

Parameters	$\hat{\Omega}=\sum_{k=1}^{K} \hat{\lambda}_k \hat{\xi}_k \hat{\xi}_k$	Mixed-SCORE
$\theta_i^{-1} \sim \text{Unif}(5, 10)$, α_1=(.6, .2, .2), α_2=(.3, .4, .3)	78.84	46.63
$\theta_i^{-1} \sim \text{Unif}(5, 10)$, α_1=(.4, .2, .4), α_2=(.2, .6, .2)	78.78	44.43
$\theta_i^{-1} \sim \text{Unif}(5, 10)$, α_1=(.4, .2, .4), α_2=(.1, .8, .1)	80.65	44.84
$\theta_i \sim \text{Unif}(0.05, 0.2)$, α_1=(.4, .2, .4), α_2=(.2, .6, .2)	71.83	44.31
$\theta_i \sim \text{Unif}(0.05, 0.2)$, α_1=(.6, .2, .2), α_2=(.3, .4, .3)	71.73	38.86

Table 4: Comparison of the Frobenius errors of estimating Ω based on 100 repetitions. Settings: $K = 3$, $n = 540$; There are $n/6$ pure nodes for each community, and the π_i's of the remaining nodes are i.i.d. drawn from a mixture distribution $0.5\,\text{Dirichlet}(\alpha_1) + 0.5\,\text{Dirichlet}(\alpha_2)$. The diagonals of P are 1 and off-diagonals are 0.3.

References

[1] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.*, 48(3):1452,

2020.

[2] Afonso S Bandeira and Ramon Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016.

[3] Tony Cai, Zongming Ma, and Yihong Wu. Sparse pca: Optimal rates and adaptive estimation. *Ann. Statist.*, 41(6):3074–3110, 2013.

[4] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM J. Numer. Anal.*, 7(1):1–46, 1970.

[5] Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2013.

[6] Mark Handcock, Adrian Raftery, and Jeremy Tantrum. Model-based clustering for social networks. *J. Roy. Statist. Soc. A*, 170(2):301–354, 2007.

[7] Roger Horn and Charles Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[8] Jiashun Jin and Wanjie Wang. Influential features pca for high dimensional clustering (with discussion). *Ann. Statist.*, 44(6):2323–2359, 2016.