

---

# Supplementary Materials for “Semi-supervised Vertex Hunting, with Applications in Network and Text Analysis”

---

Yicong Jiang, Zheng Tracy Ke

## A Preliminaries

### A.1 Without-loss-of-generality Normalizations

We begin the appendix with some without-loss-of-generality normalizations. Note that for any constant  $c_\alpha \neq 0$ ,  $M(c_\alpha \alpha) = c_\alpha M(\alpha)$ ,  $\widehat{M}(c_\alpha \alpha) = c_\alpha \widehat{M}(\alpha)$ . Because the scaling of a matrix does not affect its eigenvectors, Algorithm 1 provides the same estimator  $\widehat{V}$  when  $\alpha$  is replaced by  $c_\alpha \alpha$ . Similarly, for different (possibly negative) scalings of  $b$  and  $\hat{b}$ , model (2) and  $\widehat{W}_S, \widehat{V}$  in Algorithm 1 remains the same. Therefore, without the loss of generality, for the convenience of theoretical analysis, we assume in the appendix that

$$\|\alpha\| = 1, \quad \|b\| = \|\hat{b}\| = 1, \quad b'\hat{b} > 0.$$

### A.2 A Lemma on the Singular Values of $V$

To facilitate the evaluation of the singular values of  $V$  in later sections, we first introduce the following lemma.

**Lemma A.1** (Lower Bound of the singular values of  $V$ ). *There exists  $\gamma \in \mathbb{R}^K$ , such that*

$$\max_{\gamma} \lambda_{\min} \left( V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma' \right) \geq \sqrt{0.5} \lambda_{K-1}(V) \quad (\text{A.1})$$

*Proof.* Define  $v_0 = \frac{1}{\sqrt{K}} V' \mathbf{1}_K$ , let  $v_K$  be the unit eigenvector of  $V'V$  corresponding to the smallest eigenvalue. Then, we have that for any vector  $\eta \in \mathbb{R}^K$ ,

$$\begin{aligned} & \eta' \left( V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma' \right)' \left( V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma' \right) \eta \\ &= \eta' V' V \eta + 2(\gamma' \eta) \left( \frac{1}{\sqrt{K}} \mathbf{1}_K' V \eta \right) + (\gamma' \eta)^2 \\ &\geq \lambda_{K-1}(V)^2 \|\eta\|^2 - \lambda_{K-1}(V)^2 (v_K' \eta)^2 + 2(\gamma' \eta)(v_0' \eta) + (\gamma' \eta)^2 \\ &= 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 + 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 - \lambda_{K-1}(V)^2 (v_K' \eta)^2 + 2(\gamma' \eta)(v_0' \eta) + (\gamma' \eta)^2 \end{aligned}$$

Define  $\tilde{v}_0 = (v_0 - (v_K' v_0) v_K) / \|v_0 - (v_K' v_0) v_K\|$  (if  $\|v_0 - (v_K' v_0) v_K\| = 0$ , define  $\tilde{v}_0$  to be any unit vector orthogonal to  $v_K$ ), then  $\|\tilde{v}_0\| = 1$ ,  $v_K' \tilde{v}_0 = 0$ , and there exists some constant  $\phi_1, \phi_2$ , such that

$$v_0 = \phi_1 \tilde{v}_0 + \phi_2 v_K$$

Therefore,

$$\eta' \left( V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma' \right)' \left( V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma' \right) \eta$$

$$\begin{aligned}
&\geq 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 + 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 - \lambda_{K-1}(V)^2 (v'_K \eta)^2 + 2(\gamma' \eta)(v'_0 \eta) + (\gamma' \eta)^2 \\
&\geq 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 + 0.5 \cdot \lambda_{K-1}(V)^2 ((v'_0 \eta)^2 + (v'_K \eta)^2) \\
&\quad - \lambda_{K-1}(V)^2 (v'_K \eta)^2 + 2(\gamma' \eta)(\phi_1 \tilde{v}'_0 \eta + \phi_2 \tilde{v}'_K \eta) + (\gamma' \eta)^2
\end{aligned}$$

Define  $x_1 = \tilde{v}'_0 \eta$ ,  $x_2 = v'_K \eta$ ,  $\gamma_1 = \tilde{v}'_0 \gamma$ ,  $\gamma_2 = v'_K \gamma$ , then we have

$$\begin{aligned}
&\eta'(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma')'(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma') \eta \\
&\geq 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 + 0.5 \cdot \lambda_{K-1}(V)^2 (x_1^2 + x_2^2) - \lambda_{K-1}(V)^2 x_2^2 \\
&\quad + 2(\gamma_1 x_1 + \gamma_2 x_2)(\phi_1 x_1 + \phi_2 x_2) + (\gamma_1 x_1 + \gamma_2 x_2)^2 \\
&= 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2 + (\gamma_1^2 + 2\gamma_1 \phi_1 + 0.5 \cdot \lambda_{K-1}(V)^2) x_1^2 + (\gamma_2^2 + 2\gamma_2 \phi_2 - 0.5 \cdot \lambda_{K-1}(V)^2) x_2^2 \\
&\quad + 2(\gamma_1 \gamma_2 + \gamma_1 \phi_2 + \gamma_2 \phi_1) x_1 x_2
\end{aligned}$$

Hence, if we can find  $\gamma_1, \gamma_2$  such that

$$(\gamma_1 \gamma_2 + \gamma_1 \phi_2 + \gamma_2 \phi_1)^2 \leq (\gamma_1^2 + 2\gamma_1 \phi_1 + 0.5 \cdot \lambda_{K-1}(V)^2)(\gamma_2^2 + 2\gamma_2 \phi_2 - 0.5 \cdot \lambda_{K-1}(V)^2), \quad (\text{A.2})$$

then we have for any  $\eta \in \mathbb{R}^K$

$$\eta'(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma')'(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma') \eta \geq 0.5 \cdot \lambda_{K-1}(V)^2 \|\eta\|^2$$

Hence,

$$\max_{\gamma} \lambda_{\min}(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma') \geq \sqrt{0.5} \cdot \lambda_{K-1}(V)$$

So the desired (A.1) is satisfied.

It suffices to fulfill (A.2). Define  $\tilde{\gamma} = (\gamma_1, \gamma_2, 1)'$ ,  $\tilde{\lambda} = \sqrt{0.5} \lambda_{K-1}(V)$ ,

$$\Phi = \begin{pmatrix} \phi_2^2 + \tilde{\lambda}^2 & -\phi_1 \phi_2 & \phi_1 \tilde{\lambda}^2 \\ -\phi_1 \phi_2 & \phi_1^2 - \tilde{\lambda}^2 & -\phi_2 \tilde{\lambda}^2 \\ \phi_1 \tilde{\lambda}^2 & -\phi_2 \tilde{\lambda}^2 & \tilde{\lambda}^4 \end{pmatrix}$$

Then (A.2) is equivalent to

$$\tilde{\gamma}' \Phi \tilde{\gamma} \leq 0$$

Notice that

$$\det(\Phi) = -\tilde{\lambda}^4(\phi_2^2 - \phi_1^2 + \tilde{\lambda}^2) \leq 0$$

Therefore,  $\Phi$  has at least one non-positive eigenvalue. Consequently, there exists  $\gamma^* = (\gamma_1^*, \gamma_2^*, \gamma_3^*)' \neq \mathbf{0}$  such that  $(\gamma^*)' \Phi \gamma^* \leq 0$ .

If  $\gamma_3^* \neq 0$ , we can choose  $\gamma_1 = \gamma_1^*/\gamma_3^*$ ,  $\gamma_2 = \gamma_2^*/\gamma_3^*$ , and (A.2) is satisfied.

If  $\gamma_3^* = 0$ , then the sub-matrix of  $\Phi$ ,

$$\Phi_{\text{sub}} \stackrel{\text{def}}{=} \begin{pmatrix} \phi_2^2 + \tilde{\lambda}^2 & -\phi_1 \phi_2 \\ -\phi_1 \phi_2 & \phi_1^2 - \tilde{\lambda}^2 \end{pmatrix}$$

has at least one non-positive eigenvalue.

Since  $\text{tr}(\Phi_{\text{sub}}) = \phi_1^2 + \phi_2^2 \geq 0$ , the 2 eigenvalues of  $\Phi_{\text{sub}}$  must be one non-negative and the other non-positive. Hence,  $\det(\Phi_{\text{sub}}) \leq 0$ , so we have  $\phi_1^2 \leq \phi_2^2$ .

Choose  $\gamma_1 = 0$ ,  $\gamma_2 = \tilde{\lambda}^2/\phi^2$ , then (A.2) reduces to

$$(\phi_1^2 - \phi_2^2 - \tilde{\lambda}^2) \gamma_2^2 \leq 0,$$

which is true because  $\phi_1^2 \leq \phi_2^2$ , hence (A.2) is satisfied.

In all, there exists  $\gamma_1, \gamma_2$  such that (A.2) is satisfied, and (A.1) is proved.  $\square$

## B Proof of Theorem 2.2

Recall that

$$M(\alpha) = M(\alpha; S) = \Pi'_S \text{diag}(H\alpha)^{-1} R_S R'_S \text{diag}(H\alpha) \Pi_S.$$

By (2),

$$W_S = \text{diag}(\Pi_S \cdot b)^{-1} \Pi_S \text{diag}(b),$$

hence

$$\Pi_S = \text{diag}(\Pi_S \cdot b) W_S \text{diag}(b)^{-1}$$

Let  $b^{-1}$  to be the entrywise inverse of  $b$ , so that  $b \circ b^{-1} = \mathbf{1}_K$ . Since  $\Pi_S \cdot \mathbf{1}_K = \mathbf{1}_N$ , we have

$$\text{diag}(\Pi_S \cdot b) W_S \text{diag}(b)^{-1} \mathbf{1}_K = \mathbf{1}_N$$

So

$$\text{diag}(\Pi_S \cdot b) W_S b^{-1} = \mathbf{1}_N$$

This indicates that

$$W_S \cdot b^{-1} = (\Pi_S b)^{-1} \quad (\text{B.3})$$

Note that

$$\begin{aligned} \text{rank}(M(\alpha)) &= \text{rank}(\Pi'_S \text{diag}(H\alpha) R_S) \\ &= \text{rank}((\text{diag}(W_S \cdot b^{-1})^{-1} W_S \text{diag}(b^{-1}))' \text{diag}(H\alpha) W_S V) \\ &= \text{rank}(\text{diag}(b^{-1}) W'_S \text{diag}(W_S \cdot b^{-1})^{-1} \text{diag}(H\alpha) W_S V) \\ &= \text{rank}\left(\text{diag}(b^{-1}) W'_S \text{diag}(W_S \cdot b^{-1})^{-1} \text{diag}(H\alpha) W_S \left(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma'\right)\right), \end{aligned}$$

where the last line follows from the fact that  $\text{diag}(H\alpha) W_S \mathbf{1}_K = \text{diag}(H\alpha) \mathbf{1}_N = \mathbf{1}'_N H\alpha = \mathbf{1}'_K \Pi'_S H\alpha = 0$ . By Assumption 3.1,  $\text{diag}(b^{-1})$  is full-ranked. By Lemma A.1, there exists  $\gamma$  such that  $V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma'$  is full ranked. Therefore,

$$\begin{aligned} \text{rank}(M(\alpha)) &= \text{rank}(W'_S \text{diag}(W_S \cdot b^{-1})^{-1} \text{diag}(H\alpha) W_S) \\ &= \text{rank}\left(\sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} w_i w'_i\right) \\ &= \text{rank}\left(\sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} (w_i - \bar{w}_*)(w_i - \bar{w}_*)' - N \bar{w}_* \bar{w}_*' \sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}}\right) \end{aligned}$$

Note that

$$\sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} = \sum_{i \in S} (H\alpha)_i (\Pi_S b)_i = (H\alpha)' \Pi_S b = \alpha' H' \Pi_S b = 0$$

Therefore,

$$\begin{aligned} \text{rank}(M(\alpha)) &= \text{rank}\left(\sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} (w_i - \bar{w}_*)(w_i - \bar{w}_*)'\right) \\ &= \text{rank}\left(\sum_{i \in S} (H\alpha)_i (\pi'_i b) (w_i - \bar{w}_*)(w_i - \bar{w}_*)'\right) \\ &= \text{rank}(N \cdot \Sigma(\alpha)) = \text{rank}(\Sigma(\alpha)) \end{aligned}$$

So when  $\text{rank}(\Sigma(\alpha)) = K - 1$ , we have  $\text{rank}(M(\alpha)) = K - 1$ , which indicates that the null space of  $M(\alpha)$  is one dimensional. By Theorem 2.1 proved in the main paper (its proof is also displayed below for completeness),  $M(\alpha)b = \mathbf{0}_K$ . Therefore, the eigenvector associated with the zero eigenvalue of  $M(\alpha)$  is unique and must be equal to  $b$ .

**Proof of Theorem 2.1:** Let  $J(\alpha) = R'_S \text{diag}(H\alpha)\Pi_S$ . Then,  $M(\alpha) = J(\alpha)'J(\alpha)$ . It suffices to show  $J(\alpha)b = \mathbf{0}_d$ . First, model (1) implies  $R_S = W_S V$ . It follows that  $J(\alpha)b = V'W'_S \cdot \text{diag}(H\alpha)\Pi_S b$ . Second, model (2) implies  $w_i = (\pi_i \circ b)/\|\pi_i \circ b\|_1$ ; in the matrix form, this can be expressed as  $W_S = [\text{diag}(\Pi_S b)]^{-1}\Pi_S \text{diag}(b)$ . We plug  $W_S$  into  $J(\alpha)b$  to obtain:

$$\begin{aligned} J(\alpha)b &= V' \text{diag}(b)\Pi'_S [\text{diag}(\Pi_S b)]^{-1} \text{diag}(H\alpha)\Pi_S b \\ &= V' \text{diag}(b)\Pi'_S \text{diag}(H\alpha) [\text{diag}(\Pi_S b)]^{-1} \Pi_S b \quad (\text{switching diagonal matrices}) \\ &= V' \text{diag}(b)\Pi'_S \text{diag}(H\alpha) \mathbf{1}_N \quad (\text{because } \text{diag}(v)^{-1}v = \mathbf{1} \text{ for a vector } v) \\ &= V' \text{diag}(b)\Pi'_S H\alpha. \quad (\text{because } \text{diag}(v)\mathbf{1} = v \text{ for a vector } v) \end{aligned} \quad (\text{B.4})$$

We recall that  $H$  is the projection matrix to the orthogonal complement of  $\Pi_S$ . Hence,  $\Pi'_S H$  is a zero matrix. It follows that the right hand side of (5) is a zero vector.  $\square$

## C Proof of Lemma 3.1

Denote  $Z = X - R$  and  $Z_S = X_S - R_S$ . Notice that

$$\begin{aligned} \widehat{V} - V &= (\widehat{W}'_S \widehat{W}_S)^{-1} \widehat{W}'_S X_S - V \\ &= (\widehat{W}'_S \widehat{W}_S)^{-1} \widehat{W}'_S (W_S V + Z_S) - V \\ &= (\widehat{W}'_S \widehat{W}_S)^{-1} \widehat{W}'_S (\widehat{W}_S - W_S) V + (\widehat{W}'_S \widehat{W}_S)^{-1} \widehat{W}'_S Z_S \end{aligned}$$

Therefore,

$$\begin{aligned} \|\widehat{V} - V\| &\leq \|(\widehat{W}'_S \widehat{W}_S)^{-1} \widehat{W}'_S (\widehat{W}_S - W_S) V\| + \|(\widehat{W}'_S \widehat{W}_S)^{-1} \widehat{W}'_S Z_S\| \\ &\leq \|(\widehat{W}'_S \widehat{W}_S)^{-1}\| \cdot \|\widehat{W}'_S (\widehat{W}_S - W_S)\| \cdot \|V\| + \|(\widehat{W}'_S \widehat{W}_S)^{-1}\| \cdot \|\widehat{W}'_S Z_S\| \end{aligned} \quad (\text{C.5})$$

We analyze the terms in (C.5) and show that their relation with  $\|\widehat{W}_S - W_S\|_{\max}$  as follows.

### C.1 Error rate of $\|(\widehat{W}'_S \widehat{W}_S)^{-1}\|$

We have

$$\begin{aligned} \|(\widehat{W}'_S \widehat{W}_S)^{-1}\| &= \lambda_{\min}(\widehat{W}_S)^{-2} \\ &\leq (\lambda_{\min}(W_S) - \|\widehat{W}_S - W_S\|)^{-2} \end{aligned} \quad (\text{C.6})$$

Note that

$$\begin{aligned} \lambda_{\min}(W_S) &= \lambda_{\min}(\text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(b)) \\ &\geq \lambda_{\min}(\text{diag}(\Pi_S b)^{-1}) \lambda_{\min}(\Pi_S) \lambda_{\min}(\text{diag}(b)) \\ (\text{Assumption 3.1}) &\geq (\max_{i \in S} \pi'_i b)^{-1} (c_2 \lambda_{\max}(\Pi_S)) (\min_{k \in [K]} b_k) \\ (\text{Assumption 3.1}) &\geq (\max_k b_k)^{-1} \cdot c_2 \cdot \frac{\|\Pi_S \mathbf{1}_K\|}{\|\mathbf{1}_K\|} \cdot c_3 (\max_k b_k) \\ &\geq c_2 c_3 \sqrt{\frac{N}{K}} \end{aligned} \quad (\text{C.7})$$

Hence, when  $\|\widehat{W}_S - W_S\| < \frac{c_2 c_3}{2} \sqrt{\frac{N}{K}}$ , we have

$$\|(\widehat{W}'_S \widehat{W}_S)^{-1}\| \geq \left( \frac{1}{2} c_2 c_3 \sqrt{\frac{N}{K}} \right)^{-2} = \frac{4}{c_2^2 c_3^2} \frac{K}{N}. \quad (\text{C.8})$$

### C.2 Error rate of $\|\widehat{W}_S'(\widehat{W}_S - W_S)\|$

Notice that

$$\begin{aligned}
\|\widehat{W}_S'(\widehat{W}_S - W_S)\| &\leq \|(\widehat{W}_S - W_S)'(\widehat{W}_S - W_S)\| + \|W_S'(\widehat{W}_S - W_S)\| \\
&\leq \|\widehat{W}_S - W_S\|^2 + \|W_S\| \|\widehat{W}_S - W_S\| \\
&\leq \|\widehat{W}_S - W_S\|^2 + \sqrt{N} \|W_S\|_\infty \|\widehat{W}_S - W_S\| \\
&\leq \|\widehat{W}_S - W_S\|^2 + \sqrt{N} \|\widehat{W}_S - W_S\|
\end{aligned}$$

So

$$\|\widehat{W}_S'(\widehat{W}_S - W_S)\| \leq \|\widehat{W}_S - W_S\|^2 + \sqrt{N} \|\widehat{W}_S - W_S\| \quad (\text{C.9})$$

### C.3 Error rate of $\|\widehat{W}_S' Z_S\|$

Notice that

$$\begin{aligned}
\|\widehat{W}_S' Z_S\| &\leq \|W_S' Z_S\| + \|(\widehat{W}_S - W_S)' Z_S\| \\
&\leq \|W_S' Z_S\| + \|(\widehat{W}_S - W_S)\| \cdot \|Z_S\| \\
&\leq \|W_S' Z_S\| + \|(\widehat{W}_S - W_S)\| \cdot \sqrt{mK} \|Z_S\|_{\max} \\
&\leq \text{err}_2 + \sqrt{mK} \|(\widehat{W}_S - W_S)\| \cdot \text{err}_3
\end{aligned}$$

So

$$\|\widehat{W}_S' Z_S\| \leq \text{err}_2 + \sqrt{mK} \|(\widehat{W}_S - W_S)\| \cdot \text{err}_3 \quad (\text{C.10})$$

To sum up, the error rate of the above terms are all connected to  $\|\widehat{W}_S - W_S\|$ . We focus on the analysis of it in the next subsection.

### C.4 Error rate of $\widehat{W}_S$

Recall that

$$\begin{aligned}
W_S &= \text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(b) \\
\widehat{W}_S &= \text{diag}(\Pi_S \hat{b})^{-1} \Pi_S \text{diag}(\hat{b})
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\widehat{W}_S - W_S\| &= \|\text{diag}(\Pi_S \hat{b})^{-1} \Pi_S \text{diag}(\hat{b}) - \text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(b)\| \\
&\leq \|\text{diag}(\Pi_S \hat{b})^{-1} \Pi_S \text{diag}(\hat{b}) - \text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(\hat{b})\| \\
&\quad + \|\text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(\hat{b}) - \text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(b)\| \\
&= \|\text{diag}(\Pi_S (\hat{b} - b)) \text{diag}(\Pi_S b)^{-1} \text{diag}(\Pi_S \hat{b})^{-1} \Pi_S \text{diag}(\hat{b})\| \\
&\quad + \|\text{diag}(\Pi_S b)^{-1} \Pi_S \text{diag}(\hat{b} - b)\| \\
&\leq \left( \max_{i \in S} |\pi_i'(\hat{b} - b)| \right) \cdot \left( \min_{i \in S} \pi_i' b \right)^{-1} \cdot \left( \min_{i \in S} \pi_i' \hat{b} \right)^{-1} \cdot \|\Pi_S\| \cdot \left( \max_k \hat{b}_k \right) \\
&\quad + \left( \min_{i \in S} \pi_i' b \right)^{-1} \cdot \|\Pi_S\| \cdot \left( \max_k |\hat{b}_k - b_k| \right) \quad (\text{C.11})
\end{aligned}$$

We analysis the terms in (C.11) individually as follows.

$$\begin{aligned}
\max_{i \in S} |\pi_i'(\hat{b} - b)| &\leq \max_{i \in S} \sum_{k \in [K]} \pi_i(k) |\hat{b}_k - b_k| \\
&\leq \max_{i \in S} \sum_{k \in [K]} \pi_i(k) \|\hat{b} - b\|
\end{aligned}$$

$$= \|\hat{b} - b\| \quad (\text{C.12})$$

By Assumption 3.1,

$$\begin{aligned} \min_{i \in S} \pi'_i b &= \min_{i \in S} \sum_{k \in [K]} \pi_i(k) b_k \\ &\geq \min_{i \in S} \left( \sum_{k \in [K]} \pi_i(k) \min_l b_l \right) \\ &= \min_l b_l \\ &\geq c_3 \max_l b_l \\ &\geq c_3 \frac{\|b\|}{\sqrt{K}} \end{aligned} \quad (\text{C.13})$$

Similarly,

$$\begin{aligned} \min_{i \in S} \pi'_i \hat{b} &= \min_{i \in S} \sum_{k \in [K]} \pi_i(k) \hat{b}_k \\ &\geq \min_{i \in S} \left( \sum_{k \in [K]} \pi_i(k) \min_l \hat{b}_l \right) \\ &= \min_l \hat{b}_l \\ &\geq \min_l b_l - \max_k |\hat{b}_k - b_k| \\ &\geq c_3 \max_l b_l - \|\hat{b} - b\| \\ &\geq c_3 \frac{\|b\|}{\sqrt{K}} - \|\hat{b} - b\|, \end{aligned}$$

where recall that  $\|b\| = 1$  as in Appendix A.1. Hence when  $\|\hat{b} - b\|/\|b\| < 0.5c_3/\sqrt{K}$ ,

$$\min_{i \in S} \pi'_i \hat{b} \geq 0.5c_3 \frac{\|b\|}{\sqrt{K}} = \frac{0.5c_3}{\sqrt{K}} \quad (\text{C.14})$$

By Assumption 3.1,

$$\begin{aligned} \|\Pi_S\| &\leq \frac{1}{c_2} \lambda_{\min}(\Pi_S) \\ &= \frac{1}{c_2} \min_{x \in \mathbb{R}^K} \frac{\|\Pi_S x\|}{\|x\|} \\ &\leq \frac{1}{c_2} \frac{\|\Pi_S \mathbf{1}_K\|}{\|\mathbf{1}_K\|} \\ &= \frac{1}{c_2} \frac{\|\mathbf{1}_N\|}{\|\mathbf{1}_K\|} \\ &= \frac{1}{c_2} \sqrt{\frac{N}{K}} \end{aligned} \quad (\text{C.15})$$

By Assumption 3.1, when  $\|\hat{b} - b\|/\|b\| < 0.5c_3/\sqrt{K}$ ,

$$\begin{aligned} \max_k \hat{b}_k &\leq \max_k b_k + \|\hat{b} - b\| \\ &\leq \frac{1}{c_3} \min_k b_k + 0.5c_3 \frac{\|b\|}{\sqrt{K}} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{c_3} \frac{\|b\|}{\sqrt{K}} + 0.5c_3 \frac{\|b\|}{\sqrt{K}} \\
&= \left( \frac{1}{c_3} + 0.5c_3 \right) \frac{1}{\sqrt{K}}
\end{aligned} \tag{C.16}$$

Also it is clear that

$$\max_k |\hat{b}_k - b_k| \leq \|\hat{b} - b\| \tag{C.17}$$

Substituting (C.12), (C.13), (C.14), (C.15), (C.16), and (C.17) into (C.11), we have when  $\|\hat{b} - b\|/\|b\| < 0.5c_3/\sqrt{K}$ ,

$$\|\widehat{W}_S - W_S\| \leq \frac{2}{c_2 c_3} \left( 1 + \frac{1}{c_3^2} \right) \sqrt{N} \frac{\|\hat{b} - b\|}{\|b\|} \tag{C.18}$$

Denote  $U_0 = \Pi'_S \text{diag}(H\alpha) R_S$ ,  $U = \Pi_S \text{diag}(H\alpha) X_S$ . Recall that by Appendix A.1.  $\|b\| = \|\hat{b}\| = 1$  and  $b'\hat{b} \geq 0$ . Hence by Davis-Kahan sin  $\Theta$  theorem [14, Theorem 4],

$$\begin{aligned}
\frac{\|\hat{b} - b\|}{\|b\|} &= \|\hat{b} - b\| \leq \sqrt{2} |\sin(\hat{b}, b)| \leq \frac{2^{1.5}(2\|U_0\| + \|U - U_0\|)\|U - U_0\|_{\mathcal{F}}}{\lambda_{K-1}(U_0)^2} \\
&\leq \frac{2^{1.5}(2\|U_0\| + \text{err}_1)\text{err}_1}{\lambda_{K-1}(U_0)^2}
\end{aligned} \tag{C.19}$$

Combining (C.18) and (C.19), we have when  $\text{err}_1 < \min\{\|U_0\|, \frac{c_3}{2^{2.5} \cdot 3 \cdot \sqrt{K}} \frac{\lambda_{K-1}(U_0)^2}{\|U_0\|}\}$

$$\|\widehat{W}_S - W_S\| \leq \frac{2^{3.5}}{c_2 c_3} \left( 1 + \frac{1}{c_3^2} \right) \sqrt{N} \cdot \frac{\|U_0\|}{\lambda_{K-1}(U_0)} \cdot \frac{\text{err}_1}{\lambda_{K-1}(U_0)}. \tag{C.20}$$

It remains to evaluate  $\|U_0\|$  and  $\lambda_{K-1}(U_0)$ , which is addressed in the next subsection.

### C.5 Evaluation of $\|U_0\|$ and $\lambda_{K-1}(U_0)$

With Assumption 3.1, we have  $\min_i |(\Pi_S b)_i| \geq c \max_i |(\Pi_S b)_i|$ , hence the conditional number of  $\text{diag}(\Pi_S b)$  is bounded.

Note that

$$W'_S \text{diag}(H\alpha) W_S \mathbf{1}_K = W'_S \text{diag}(H\alpha) \mathbf{1}_N = W'_S H\alpha = 0$$

Therefore, for any  $\gamma \in \mathbb{R}^K$ ,

$$U_0 = \Pi'_S \text{diag}(H\alpha) W_S (V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma')$$

So,

$$\|U_0\| \leq \|\Pi'_S \text{diag}(H\alpha) W_S\| \cdot \|V\| \tag{C.21}$$

$$\lambda_{K-1}(U_0) \geq \lambda_{K-1}(\Pi'_S \text{diag}(H\alpha) W_S) \cdot \max_{\gamma} \lambda_{\min}(V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma') \tag{C.22}$$

From Lemma A.1 and Assumption 3.1, we have

$$\max_{\gamma} \lambda_{\min} \left( V + \frac{1}{\sqrt{K}} \mathbf{1}_K \gamma' \right) \geq \sqrt{0.5} \lambda_{K-1}(V) \geq \sqrt{0.5} c_1 \|V\| \tag{C.23}$$

Plugging (C.23) into (C.22), we have

$$\lambda_{K-1}(U_0) \geq \sqrt{0.5} c_1 \|V\| \lambda_{K-1}(\Pi'_S \text{diag}(H\alpha) W_S) \tag{C.24}$$

It remains to evaluate  $\|\Pi'_S \text{diag}(H\alpha)W_S\|$  and  $\lambda_{K-1}(\Pi'_S \text{diag}(H\alpha)W_S)$ . We start with  $\|\Pi'_S \text{diag}(H\alpha)W_S\|$ . By (2), we have

$$W_S = \text{diag}(\Pi_S \cdot b)^{-1} \Pi_S \text{diag}(b).$$

Therefore,

$$\begin{aligned} \|\Pi'_S \text{diag}(H\alpha)W_S\| &= \|\Pi'_S \text{diag}(H\alpha) \text{diag}(\Pi_S \cdot b)^{-1} \Pi_S \text{diag}(b)\| \\ &\leq \|\Pi'_S \text{diag}(H\alpha) \text{diag}(\Pi_S \cdot b)^{-1} \Pi_S\| \|\text{diag}(b)\| \\ &= \left\| \sum_{i \in S} (H\alpha)_i / (\pi'_i b) \cdot \pi_i \pi'_i \left( \max_{k \in [K]} b_k \right) \right\| \\ &\leq \left( \max_{\|x\|=1} \sum_{i \in S} (H\alpha)_i / (\pi'_i b) \cdot (\pi'_i x)^2 \right) \left( \frac{\min_{k \in [K]} b_k}{c_3} \right) \\ &\leq \left( \sum_{i \in S} |(H\alpha)_i| \right) / \left( \min_i \pi'_i b \right) \cdot \left( \max_{i \in S, \|x\|=1} (\pi'_i x)^2 \right) \left( \frac{\min_{k \in [K]} b_k}{c_3} \right) \\ &\leq \left( \sqrt{N} \|H\alpha\| \right) / \left( \min_i \pi'_i (1_K \cdot \min_{k \in [K]} b_k) \right) \cdot \left( \max_{i \in S, \|x\|=1} (\pi'_i x)^2 \right) \left( \frac{\min_{k \in [K]} b_k}{c_3} \right) \\ &\leq \sqrt{N} \|\alpha\| / \left( \min_{k \in [K]} b_k \right) \cdot \left( \max_{i \in S} \|\pi\|^2 \right) \left( \frac{\min_{k \in [K]} b_k}{c_3} \right) \\ &\leq \frac{\sqrt{N}}{c_3} \left( \max_{i \in S} \|\pi\|_1^2 \right) \\ &= \frac{\sqrt{N}}{c_3} \end{aligned}$$

where we leverage the fact that  $H$  is projection matrix, so  $\|H\alpha\| \leq \|\alpha\|$ . Plugging into (C.21), we have

$$\|U_0\| \leq \frac{1}{c_3} \sqrt{N} \|V\| \quad (\text{C.25})$$

We then analyze  $\lambda_{K-1}(\Pi'_S \text{diag}(H\alpha)W_S)$ . Recall that by (2),

$$\Pi_S = \text{diag}(W_S \cdot b^{-1})^{-1} W_S \text{diag}(b^{-1}).$$

Hence

$$\begin{aligned} \lambda_{K-1}(\Pi'_S \text{diag}(H\alpha)W_S) &= \lambda_{K-1}((\text{diag}(W_S \cdot b^{-1}) W_S \text{diag}(b^{-1}))' \text{diag}(H\alpha)W_S) \\ &= \lambda_{K-1}(\text{diag}(b^{-1}) W'_S \text{diag}(W_S \cdot b^{-1}) \text{diag}(H\alpha)W_S) \\ &\geq \frac{1}{\max_k b_k} \lambda_{K-1}(W'_S \text{diag}(W_S \cdot b^{-1}) \text{diag}(H\alpha)W_S) \\ (\text{Assumption 3.1}) \quad &\geq \frac{c_3}{\min_k b_k} \lambda_{K-1} \left( \sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} w_i w'_i \right) \\ &\geq \frac{c_3 \sqrt{K}}{\|b\|} \lambda_{K-1} \left( \sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} (w_i - \bar{w}_*)(w_i - \bar{w}_*)' - N \bar{w}_* \bar{w}_*' \sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} \right) \end{aligned}$$

Note that

$$\sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} = \sum_{i \in S} (H\alpha)_i (\Pi_S b)_i = (H\alpha)' \Pi_S b = \alpha' H' \Pi_S b = 0$$

Hence

$$\lambda_{K-1}(\Pi'_S \text{diag}(H\alpha)W_S) \geq \frac{c_3 \sqrt{K}}{\|b\|} \lambda_{K-1} \left( \sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} (w_i - \bar{w}_*)(w_i - \bar{w}_*)' \right)$$



$$\begin{aligned}
(\text{Recall } \|b\| = 1) &= c_3 \sqrt{K} \lambda_{K-1} \left( \sum_{i \in S} (H\alpha)_i (\pi'_i \cdot b) (w_i - \bar{w}_*) (w_i - \bar{w}_*)' \right) \\
&= c_3 \sqrt{K} \lambda_{K-1} (N \Sigma(\alpha)) = c_3 \sqrt{K} N \lambda_{K-1} (\Sigma(\alpha)),
\end{aligned}$$

where we use the equality  $(W'_i \cdot b^{-1})^{-1} = \pi'_i \cdot b$  as shown in (B.3). Note that  $\Sigma(\alpha) \mathbf{1}_K = \sum_{i \in S} \frac{(H\alpha)_i}{W'_i \cdot b^{-1}} (w_i - \bar{w}_*) (w_i - \bar{w}_*)' \mathbf{1}_K = 0$ . Additionally, by definition of  $P$ ,

$$\begin{aligned}
P' \mathbf{1}_K &= (I_{K-1} - \frac{1}{K} \mathbf{1}_{K-1} \mathbf{1}_{K-1}' - \frac{1}{K} \mathbf{1}_{K-1}^\top) \cdot \begin{pmatrix} \mathbf{1}_{K-1} \\ 1 \end{pmatrix} \\
&= \mathbf{1}_{K-1} - \frac{1}{K} (\mathbf{1}_{K-1}' \mathbf{1}_{K-1}) \mathbf{1}_{K-1} - \frac{1}{K} \mathbf{1}_{K-1} = 0.
\end{aligned}$$

Additionally,

$$\begin{aligned}
P'P &= (I_{K-1} - \frac{1}{K} \mathbf{1}_{K-1} \mathbf{1}_{K-1}') (I_{K-1} - \frac{1}{K} \mathbf{1}_{K-1} \mathbf{1}_{K-1}')' + \left( -\frac{1}{K} \mathbf{1}_{K-1}^\top \right) \left( -\frac{1}{K} \mathbf{1}_{K-1}^\top \right)' \\
&= I_{K-1} - \frac{2}{K^2} \mathbf{1}_{K-1} \mathbf{1}_{K-1}'.
\end{aligned}$$

So the singular values of  $P$  are either 1 or  $\sqrt{1 - \frac{2}{K^2}}$ . This, together with  $P' \mathbf{1}_K = 0$ , indicates that  $P$  is full-ranked,  $\|P\| = 1$ , and the column space of  $P$  is exactly the orthogonal space of  $\mathbf{1}_K$ . Therefore, by Assumption 3.3,

$$\begin{aligned}
\lambda_{K-1}(\Sigma(\alpha)) &= \min_{x' \mathbf{1}_K = 0} \frac{\|\Sigma(\alpha)x\|}{\|x\|} \\
&= \min_{x=Py} \frac{\|\Sigma(\alpha)x\|}{\|x\|} \\
&= \min_y \frac{\|\Sigma(\alpha)Py\|}{\|Py\|} \\
&\geq \frac{|\lambda_{\min}(\Sigma(\alpha)P)|}{\|P\|} \\
&\stackrel{(\text{By Assumption 3.3})}{\geq} c_4 \frac{\|\alpha\| \|b\|}{\|P\| \sqrt{Km}}
\end{aligned}$$

Consequently, recall that  $\|P\| = 1$ , and by Appendix A.1,  $\|\alpha\| = \|b\| = 1$ , hence

$$\lambda_{K-1}(\Pi'_S \text{diag}(H\alpha) W_S) \geq c_3 \sqrt{K} N \lambda_{K-1}(\Sigma(\alpha)) \geq c_3 c_4 \sqrt{N} \frac{\|\alpha\|}{\|P\|} = c_3 c_4 \sqrt{N}$$

Plugging the above term into (C.24), by Assumption 3.1, we obtain

$$\lambda_{K-1}(U_0) \geq \sqrt{0.5} c_1 c_3 c_4 \|V\| \sqrt{N} \geq \sqrt{0.5} c_1^2 c_3 c_4 \sqrt{N} \quad (\text{C.26})$$

Combining (C.8), (C.9), (C.10), (C.20), (C.26), and plugging them back into (C.5), we have that for some contact  $C_5 > 0$  depending on  $c_1, \dots, c_4$ , when  $\text{err}_1 \geq \min\{\frac{c_1^3 c_2^2 c_3^5 c_4^2}{2^{5.5}(1+c_3^2)} \sqrt{\frac{N}{K}}, \frac{c_1^3 c_3^4 c_4^2}{2^{3.5 \cdot 3}} \sqrt{\frac{N}{K}}, \sqrt{0.5} c_1^2 c_3 c_4 \sqrt{N}\}$

$$\frac{\|\hat{V} - V\|}{\|V\|} \leq C_5 K \left[ \frac{1}{\sqrt{N}} \text{err}_1 (1 + \text{err}_3) + \frac{1}{N} \text{err}_1^2 + \frac{1}{N} \text{err}_2 \right], \quad (\text{C.27})$$

Choosing  $c_6 = \min\{\frac{c_1^3 c_2^2 c_3^5 c_4^2}{2^{5.5}(1+c_3^2)}, \frac{c_1^3 c_3^4 c_4^2}{2^{3.5 \cdot 3}}, \sqrt{0.5} c_1^2 c_3 c_4\}$ , we conclude the proof.  $\square$

## D Proof of Lemma 3.2

We first focus on  $\text{err}_1$  and  $\text{err}_2$ . Denote  $W_S^{(\alpha)} = \text{diag}(H\alpha) W_S$ ,  $Z_S = X_S - R_S$ . Then  $\text{err}_1 = \|(W_S^{(\alpha)})' Z_S\|_{\mathcal{F}}$ ,  $\text{err}_2 = \|W_S' Z_S\|_{\mathcal{F}}$ . To evaluate  $\text{err}_1$  and  $\text{err}_2$ , we first derive the error bond for a more general expression,  $\|B' Z_S\|_{\mathcal{F}}$ , where  $B$  is an arbitrary  $N$  by  $p$  constant matrix.

Let  $Z_1, \dots, Z_K$  to be the  $N$  dimensional column vectors of  $Z$ . Let  $mK$  dimensional vector  $\text{vec}(Z)$  to be the concatenation of  $Z_1, \dots, Z_K$ , so  $\text{vec}(Z)' = (Z_1', \dots, Z_K')$ . Then we have  $\|B'Z_S\|_{\mathcal{F}}^2 = \text{vec}(Z)'(I_K \otimes (B'B))\text{vec}(Z)$ , where  $\otimes$  denotes the Kronecker product. According to Assumption 3.2, entires of  $\text{vec}(Z)$  are all independent and are sub-Gaussian random variables with scale parameter  $\sigma_n$ . Therefore, by Hanson–Wright inequality [13, Theorem 1.1], there exists an absolute constant  $c_8$  not depending on any constants in our paper, such that for any  $t > 0$ ,

$$\begin{aligned} & \mathbb{P}(|\|B'Z_S\|_{\mathcal{F}}^2 - \mathbb{E}[\|B'Z_S\|_{\mathcal{F}}^2]| > t) \\ &= \mathbb{P}(|\text{vec}(Z)'(I_K \otimes (B'B))\text{vec}(Z) - \mathbb{E}[\text{vec}(Z)'(I_K \otimes (B'B))\text{vec}(Z)]| > t) \\ &\leq 2 \exp\left(-c_8 \min\left\{\frac{t^2}{\sigma_n^4 \|I_K \otimes (B'B)\|_{\mathcal{F}}^2}, \frac{t}{\sigma_n^2 \|I_K \otimes (B'B)\|}\right\}\right) \end{aligned} \quad (\text{D.28})$$

Since

$$\|I_K \otimes (B'B)\| \leq \|I_K \otimes (B'B)\|_{\mathcal{F}} = \sqrt{K} \|B'B\|_{\mathcal{F}} \leq \sqrt{K} \|B\|_{\mathcal{F}}^2,$$

choose  $t = \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 \log(n) \sqrt{K} \|B\|_{\mathcal{F}}^2$ , we have

$$\mathbb{P}\left(|\|B'Z_S\|_{\mathcal{F}}^2 - \mathbb{E}[\|B'Z_S\|_{\mathcal{F}}^2]| > \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 \log(n) \sqrt{K} \|B\|_{\mathcal{F}}^2\right) \leq \frac{2}{n} \quad (\text{D.29})$$

Note that

$$\begin{aligned} \mathbb{E}[\|B'Z_S\|_{\mathcal{F}}^2] &= \mathbb{E}[\text{vec}(Z)'(I_K \otimes (B'B))\text{vec}(Z)] \\ &= \mathbb{E}[\text{tr}(\text{vec}(Z)'(I_K \otimes (B'B))\text{vec}(Z))] \\ &= \mathbb{E}[\text{tr}((I_K \otimes (B'B))\text{vec}(Z)\text{vec}(Z)')] \\ &= \text{tr}((I_K \otimes (B'B))\mathbb{E}[\text{vec}(Z)\text{vec}(Z)']) \\ &= \text{tr}((I_K \otimes (B'B))\text{diag}(\mathbb{E}[\text{vec}(Z) \circ \text{vec}(Z)])) \\ &\stackrel{(\text{Assumption 3.2})}{\leq} \sigma_n^2 \text{tr}(I_K \otimes (B'B)) \\ &= \sigma_n^2 K \|B\|_{\mathcal{F}} \end{aligned} \quad (\text{D.30})$$

Plugging the above (D.30) into (D.29), we have that with probability at least  $1 - 2/n$ ,

$$\|B'Z_S\|_{\mathcal{F}} \leq \sqrt{\sigma_n^2 K \|B\|_{\mathcal{F}} + \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 \log(n) \sqrt{K} \|B\|_{\mathcal{F}}^2} \quad (\text{D.31})$$

With (D.31), it remains to evaluate the corresponding  $\|B\|_{\mathcal{F}}$  for  $\text{err}_1$  and  $\text{err}_2$  respectively.

### D.1 Analysis of $\text{err}_1$

In this section, to evaluate  $\text{err}_1$ , we choose  $B = W_S^{(\alpha)} = \text{diag}(H\alpha)W_S$ , and have

$$\begin{aligned} \|B\|_{\mathcal{F}}^2 &= \|\text{diag}(H\alpha)W_S\|_{\mathcal{F}}^2 \\ &= \sum_{i \in S, k \in [K]} (H\alpha)_i^2 w_i(k)^2 \\ &= \sum_{i \in [S]} (H\alpha)_i^2 \sum_{k \in [K]} w_i(k)^2 \\ &\leq \sum_{i \in [S]} (H\alpha)_i^2 \sum_{k \in [K]} w_i(k) \\ &= \sum_{i \in [S]} (H\alpha)_i^2 \\ &= \|H\alpha\| \leq \|\alpha\| \leq 1, \end{aligned}$$

where the last line is because  $H$  is a projection matrix. Therefore,

$$\|B\|_{\mathcal{F}} = \|W_S^{(\alpha)}\|_{\mathcal{F}} \leq 1 \quad (\text{D.32})$$

Substituting (D.32) into (D.31), we have with probability at least  $1 - 2/n$ ,

$$\text{err}_1 = \|B'Z_S\|_{\mathcal{F}} \leq \sigma_n \sqrt{K + \max\{c_8^{-1}, c_8^{-0.5}\} \log(n) \sqrt{K}} \quad (\text{D.33})$$

Therefore, choosing  $C_{\text{err}} = \max\left\{\sqrt{6}, \sqrt{1 + c_8^{-1}}, \sqrt{1 + c_8^{-0.5}}\right\}$  (here we make  $C_{\text{err}} \geq \sqrt{6}$  for the convenience of evaluating  $\text{err}_3$  as in Appendix D.3), we have that with probability at least  $1 - 2/n = 1 - O(1/n)$ ,

$$\text{err}_1 \leq \sigma_n \sqrt{K + C_{\text{err}}^2 \sqrt{K} \log(n)} \quad (\text{D.34})$$

## D.2 Analysis of $\text{err}_2$

In this section, to evaluate  $\text{err}_2$ , we choose  $B = W_S$ , and have

$$\begin{aligned} \|B\|_{\mathcal{F}}^2 &= \|W_S\|_{\mathcal{F}}^2 \\ &= \sum_{i \in [S]} \sum_{k \in [K]} w_i(k)^2 \\ &\leq \sum_{i \in [S]} \sum_{k \in [K]} w_i(k) \\ &= \sum_{i \in [S]} 1 \\ &= N, \end{aligned}$$

Therefore,

$$\|B\|_{\mathcal{F}} = \|W_S\|_{\mathcal{F}} \leq \sqrt{N} \quad (\text{D.35})$$

Substituting (D.35) into (D.31), we have with probability at least  $1 - 2/n$ ,

$$\text{err}_2 = \|B'Z_S\|_{\mathcal{F}} \leq \sigma_n \sqrt{K\sqrt{N} + \max\{c_8^{-1}, c_8^{-0.5}\} \log(n) \sqrt{K} N} \quad (\text{D.36})$$

Since  $N \geq K$ , we have  $K\sqrt{N} \leq \log(n) \sqrt{K} N$ . Hence,

$$\begin{aligned} \text{err}_2 &\leq \sigma_n \sqrt{(1 + \max\{c_8^{-1}, c_8^{-0.5}\}) \log(n) \sqrt{K} N} \\ &= \sigma_n \sqrt{\max\{1 + c_8^{-1}, 1 + c_8^{-0.5}\} N \sqrt{K} \log(n)} \end{aligned} \quad (\text{D.37})$$

Therefore, recalling that  $C_{\text{err}} = \max\left\{\sqrt{6}, \sqrt{1 + c_8^{-1}}, \sqrt{1 + c_8^{-0.5}}\right\}$  (here we make  $C_{\text{err}} \geq \sqrt{6}$  for the convenience of evaluating  $\text{err}_3$  as in Appendix D.3), we have that with probability at least  $1 - 2/n = 1 - O(1/n)$ ,

$$\text{err}_2 \leq C_{\text{err}} \sigma_n \sqrt{N \sqrt{K} \log(n)} \quad (\text{D.38})$$

## D.3 Analysis of $\text{err}_3$

In this section, we evaluate  $\text{err}_3$ . By Markov inequality, for any  $i \in S, k \in [K], t > 0, C \in \mathbb{R}$

$$\mathbb{P}(e'_i(X_S - R_S)e_k > C) \leq \frac{\mathbb{E}[\exp(t \cdot Z_{ik})]}{\exp(tC)} \leq \frac{\exp(t^2 \sigma_n^2/2)}{\exp(tC)}$$

Choose  $t = C/(\sigma_{max}^2)$ , we have

$$\mathbb{P}(e'_i(X_S - R_S)e_k > C) \leq \exp\left(-\frac{C^2}{2\sigma_n^2}\right)$$

Hence, choose  $C = \sigma_n \sqrt{6 \log(n)}$ , we have

$$\mathbb{P} \left( e'_i(X_S - R_S)e_k > \sigma_n \sqrt{6 \log(n)} \right) \leq \exp\left(-\frac{C^2}{2\sigma_n^2}\right) = \frac{1}{n^3}$$

Similarly, we have

$$\mathbb{P} \left( -e'_i(X_S - R_S)e_k > \sigma_n \sqrt{6 \log(n)} \right) \leq \frac{1}{n^3}$$

Therefore,

$$\begin{aligned} \mathbb{P} \left( \|(X_S - R_S)\|_{\max} > K\sigma_n \sqrt{6 \log(n)} \right) &\leq \sum_{i \in S, k \in [K]} \mathbb{P} \left( e'_i(X_S - R_S)e_k > \sigma_n \sqrt{6 \log(n)} \right) \\ &\quad + \sum_{i \in S, k \in [K]} \mathbb{P} \left( -e'_i(X_S - R_S)e_k > \sigma_n \sqrt{6 \log(n)} \right) \\ &\leq \frac{2K}{n^2} \end{aligned}$$

Since  $C_{\text{err}} \geq \sqrt{6}$ , we have that with probability at least  $1 - 2K/n^2 \geq 1 - 2/n = 1 - O(1/n)$ ,

$$\text{err}_3 = \|(X_S - R_S)\|_{\max} \leq C_{\text{err}} \sigma_n \sqrt{\log(n)}.$$

#### D.4 A Rough Estimate of $C_{\text{err}}$

According to [12], the absolute constant in Hanson–Wright inequality  $c_8$  is approximately lower bounded by 0.1457. Hence,  $C_{\text{err}} = \max \left\{ \sqrt{6}, \sqrt{1 + c_8^{-1}}, \sqrt{1 + c_8^{-0.5}} \right\}$  can be approximately upper bounded by

$$\max \left\{ \sqrt{6}, \sqrt{1 + 0.1457^{-1}}, \sqrt{1 + 0.1457^{-0.5}} \right\} \approx 2.8042.$$

### E Proof of Theorem 3.1 and Lemma 3.3

Theorem 3.1 is a direct result of combining Lemma 3.1 and Lemma 3.2. Because  $\frac{N}{n} \geq \frac{\sigma_n^2(K^2 + C_{\text{err}}^2 K^{1.5} \log(n))}{c_6^2 n}$ , we have

$$\sigma_n \sqrt{K + C_{\text{err}}^2 \sqrt{K} \log(n)} \leq c_6 \frac{N}{k}$$

By Lemma 3.2,

$$\text{err}_1 \leq \sigma_n \sqrt{K + C_{\text{err}}^2 \sqrt{K} \log(n)}$$

So  $\text{err}_1 \leq c_6 \frac{N}{k}$ . Therefore, the condition for Lemma 3.1 is satisfied. Plugging the result of Lemma 3.2 into Lemma 3.1, we have that for some constant  $\tilde{C}_5 > 0$  only depending on  $c_1, \dots, c_4$ , and  $C_{\text{err}}$ , with probability at least  $1 - O(1/n)$ ,

$$\frac{\|\hat{V} - V\|}{\|V\|} \leq \tilde{C}_5 \frac{K^{1.25} \sigma_n \sqrt{\log(n)} \sqrt{\sqrt{K} + \log(n)}}{\sqrt{N}}. \quad (\text{E.39})$$

When  $b$  is ideal, with the same derivation as in Section C.1, (C.5) becomes

$$\|\hat{V}^* - V\| \leq \|(W'_S W_S)^{-1}\| \cdot \|W'_S Z_S\| = |\lambda_{\min}(W_S)| \cdot \text{err}_2 \quad (\text{E.40})$$

From (C.7), we have

$$\lambda_{\min}(W_S) \geq c_2 c_3 \sqrt{\frac{N}{K}}$$

Hence, plugging in the result of Lemma 3.2, we obtain that for constant  $\tilde{C}_5 > 0$  only depending on  $c_1, \dots, c_4$ , and  $C_{\text{err}}$ , with probability at least  $1 - O(1/n)$ ,

$$\frac{\|\hat{V}^* - V\|}{\|V\|} \leq \tilde{C}_5 \frac{K^{1.25} \sigma_n \sqrt{\log(n)}}{\sqrt{N}}. \quad (\text{E.41})$$

## F Proof of Theorem 4.1 and Theorem 4.2

### F.1 Proof of Theorem 4.1

By Definition 4.1, under DCMM model,  $\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \cdot \pi'_i P \pi_j$ . Therefore, if we define  $\Theta = \text{diag}(\theta_1, \dots, \theta_n)$ ,  $\Pi = (\pi_1, \dots, \pi_n)^T$ , then

$$\Omega = \mathbb{E}[A] = \Theta \Pi P \Pi^T \Theta.$$

This is called the matrix form of DCMM [7]. Recall  $r_i = U' \Omega e_i / (\eta' U' \Omega e_i)$ , for  $1 \leq i \leq n$ . We have

$$\begin{aligned} r_i &= U' \Omega e_i / (\eta' U' \Omega e_i) \\ &= U' \Theta \Pi P \Pi^T \Theta e_i / (\eta' U' \Theta \Pi P \Pi^T \Theta e_i) \\ &= U' \Theta \Pi P \Pi^T (\theta_i e_i) / (\eta' U' \Theta \Pi P \Pi^T (\theta_i e_i)) \\ &= U' \Theta \Pi P \pi_i \theta_i / (\eta' U' \Theta \Pi P \Pi^T \pi_i \theta_i) \\ &= U' \Theta \Pi P \pi_i / (\eta' U' \Theta \Pi P \Pi^T \pi_i) \end{aligned}$$

Denote  $B = \Pi' \Theta U \in \mathbb{R}^{K \times K}$ ,  $b = \Pi' \Theta U \eta \in \mathbb{R}^K$ ,  $V = \text{diag}(b)^{-1} B$ , then

$$\begin{aligned} r_i &= B' \pi_i / (b' \pi_i) \\ &= V' \text{diag}(b) \pi_i / (b' \pi_i) \\ &= V' \cdot (b \circ \pi_i) / \|b \circ \pi_i\|_1 \end{aligned}$$

Let  $w_i = (b \circ \pi_i) / \|b \circ \pi_i\|_1$ , and let  $v_1, \dots, v_K \in \mathbb{R}^K$  be the row vectors of  $V$ . Then  $r_i = \sum_{k=1}^K w_i(k) v_k$ , with  $w_i = (b \circ \pi_i) / \|b \circ \pi_i\|_1$ . Furthermore, since  $\eta' U' \Omega e_i > 0$  for all  $i$ , we obtain

$$\begin{aligned} 0 &< \eta' U' \Omega e_i \\ &= \eta' U' \Theta \Pi P \Pi' \Theta e_i \\ &= \eta' U' \Theta \Pi P \Pi' (\theta_i e_i) \\ &= \eta' U' \Theta \Pi P \pi_i \cdot \theta_i \\ &= \theta_i b' \pi_i. \end{aligned}$$

By Definition 4.1,  $\theta_i > 0$ . Therefore,  $b' \theta > 0$  for all  $i$ , which indicates that  $b$  is a positive vector. This concludes the proof for part (a).

Because part (a) is true, if we plug  $(K, \Omega, \Pi_S)$  into Algorithm 2, we have  $x_i = r_i$  for all  $i$  and hence the SSVH algorithm in step 2 provides perfectly estimations  $\hat{V} = V$  and  $\hat{b} = b$ . Hence in step 3,  $\hat{B} = \text{diag}(\hat{b}) \hat{V} = \text{diag}(b) V = B$ . Consequently, before the re-normalization step,

$$\begin{aligned} \tilde{\pi}_i &= e'_i A U \hat{B}' (\hat{B} \hat{B}')^{-1} \\ &= e'_i \Omega U B' (B B')^{-1} \\ &= e'_i (\Theta \Pi P \Pi' \Theta) U B' (B B')^{-1} \\ &= \theta_i \pi'_i (P \Pi' \Theta U) B' (B B')^{-1} \\ &= \theta_i \pi'_i B B' B' (B B')^{-1} \\ &= \theta \pi'_i. \end{aligned}$$

Therefore, after re-normalizing  $\tilde{\pi}_i$  to have a unit  $\ell^1$ -norm ( $\theta \pi'_i$  is a positive vector, so setting the negative entries in  $\tilde{\pi}_i$  to zero has no influence on it), we have  $\hat{\pi}_i = \pi_i$  for all  $i \notin S$ . This concludes the proof of part (b).

In all, we have proved Theorem 4.1. □

## F.2 Proof of Theorem 4.2

The approach to proving Theorem 4.2 is similar to Theorem 4.1. Denote  $\theta_j = \|a_j\|_1 = \|Ae_j\|_1$ . Since  $D_0 = A\Gamma$ , if we plug  $(K, D_0, A_S^*)$  into Algorithm 3, then in step 1,

$$\begin{aligned} x_j &= U'D'e_j/(\eta'U'D'e_j) \\ &= U'D'_0e_j/(\eta'U'D'_0e_j) \\ &= U'\Gamma'A'e_j/(\eta'U'\Gamma'A'e_j) \\ &= U'\Gamma'a_j/(\eta'U'\Gamma'a_j) \\ &= U'\Gamma'a_j^*\theta_j/(\eta'U'\Gamma'a_j^*\theta_j) \\ &= U'\Gamma'a_j^*/(\eta'U'\Gamma'a_j^*) \end{aligned}$$

Denote  $B = \Gamma U \in \mathbb{R}^{K \times K}$ ,  $b = \Gamma U \eta \in \mathbb{R}^K$ ,  $V = \text{diag}(b)^{-1}B$ , then

$$\begin{aligned} x_j &= B'a_j^*/(b'a_j^*) \\ &= V'\text{diag}(b)a_j^*/(b'a_j^*) \\ &= V' \cdot (b \circ a_j^*)/\|b \circ a_j^*\|_1 \end{aligned}$$

Let  $w_j = (b \circ a_j^*)/\|b \circ a_j^*\|_1$ , and let  $v_1, \dots, v_K \in \mathbb{R}^K$  be the row vectors of  $V$ . Then  $x_i = \sum_{k=1}^K w_i(k)v_k$ . Furthermore, since  $\eta'U'D'_0e_j > 0$  for all  $i$ , we obtain

$$\begin{aligned} 0 &< \eta'U'D'_0e_j \\ &= \eta'U'\Gamma'A'e_j \\ &= \eta'U'\Gamma'(\theta_j a_j^*) \\ &= \theta_j b'a_j^*. \end{aligned}$$

Since  $\theta_j = \|a_j\|_1 > 0$ , we have  $b'\theta > 0$  for all  $i$ , which indicates that  $b$  is a positive vector.

Therefore, because  $x_i = \sum_{k=1}^K w_i(k)v_k$  and  $b$  is a positive vector, the SSVH algorithm in step 2 provides perfectly estimations  $\hat{V} = V$  and  $\hat{b} = b$ . Hence in step 3,  $\hat{B} = \text{diag}(\hat{b})\hat{V} = \text{diag}(b)V = B$ , and consequently,

$$\begin{aligned} \hat{A} &= DU\hat{B}'(\hat{B}\hat{B}')^{-1} \\ &= D_0UB'(BB')^{-1} \\ &= A\Gamma UB'(BB')^{-1} \\ &= ABB'(BB')^{-1} \\ &= A. \end{aligned}$$

Therefore, if we plug  $(K, D_0, A_S^*)$  into Algorithm 3, then  $\hat{A} = A$ . This concludes the proof of Theorem 4.2.  $\square$

## G A Review of Existing VH Algorithms

MSCORE allows one to plug in any vertex hunting (VH) algorithm. VH is the task of estimating  $v_1, v_2, \dots, v_K$  from  $\hat{r}_1, \hat{r}_2, \dots, \hat{r}_n$ . The main paper mentions several VH algorithms. Their details are given here.

### G.1 Successive Projection

SP [1] is a VH algorithm that requires the existence of pure nodes and utilizes a geometric property of simplexes: The maximum  $\ell^2$ -norm of vectors in a simplex is always attained at a vertex. Inspired by this observation, SP runs as follows:

- Initialization:  $j_1 = \arg\max_{1 \leq i \leq n} \|\hat{r}_i\|$  and  $\hat{v}_1 = \hat{r}_{j_1}$ .

- For  $k = 2, 3, \dots, K$ , let  $\mathcal{P}_{k-1}$  be the projection matrix to the linear space spanned by  $\hat{v}_1, \hat{v}_2, \dots, \hat{v}_{K-1}$ . Let  $j_k = \arg\max_{1 \leq i \leq n} \|(I_K - \mathcal{P}_{k-1})\hat{r}_i\|$  and  $\hat{v}_k = \hat{r}_{j_k}$ .

The SP algorithm is one of the few VH algorithms that have theoretical guarantees. Under mild conditions, [3] provides the following error bound for SP:

**Lemma G.1.** *Let  $\lambda_k(v)$  denotes the  $k$ th singular value of  $V$  and denote  $\gamma(V) = \max_{1 \leq k \leq K} \{\|v_k\|\}$ . Suppose  $K$ -dimensional vectors  $X_1, \dots, X_n$  satisfies model (1). Assume that for each  $k \in \{1, 2, \dots, K\}$  there exist  $i \in \{1, 2, \dots, n\}$  such that  $\pi_i = e_k$ . Suppose  $\max_{1 \leq i \leq n} \|\epsilon_i\| \leq \frac{\lambda_K(V)}{1+80\gamma(V)^2/\lambda_K^2(V)} \min\{\frac{1}{2\sqrt{K-1}}, \frac{1}{4}\}$ . Let  $\hat{v}_1^{(SP)}, \dots, \hat{v}_K^{(SP)}$  are the output of the SP algorithm. Then, there exists permutation  $\tau : \{1, 2, \dots, K\} \rightarrow \{1, 2, \dots, K\}$ , such that*

$$\max_{1 \leq k \leq K} \{\|\hat{v}_{\tau(k)}^{(SP)} - v_k\|\} \leq \left[1 + 80 \frac{\gamma^2(V)}{\lambda_K^2(V)}\right] \max_{1 \leq i \leq n} \|\epsilon_i\|. \quad (\text{G.42})$$

Comparing SP with our vertex imputation algorithm, we can see that

1. SP requires the existence of “pure point”: for each  $k \in \{1, 2, \dots, K\}$  there exist  $i \in \{1, 2, \dots, n\}$  such that  $\pi_i = e_k$ . In general, this can be hard to be satisfied. For instance, in recommendation system or social network analysis,  $\pi_i$  represents the membership weight of each individual in different communities, and it is very likely that each individual belongs to at least two communities. Furthermore, even if there exists several individuals being in only one community, it can be difficult to make every community to have a pure point, especially when the community numbers  $K$  is large. For example, in the field of statistics, almost all people in the bioinformatics community more or less studies some Bayesian statistics topics, and it is very hard to find a pure person in bioinformatics without any weight on other communities.
2. The theoretical results of SP in [3] requires  $\lambda_K(V) \geq \max_{1 \leq i \leq n} \|\epsilon_i\|$ , which is of order  $O_P(\sigma_n \sqrt{K \log(n)})$  when  $\epsilon_i$  is sub-Gaussian with scale parameter  $\sigma_n$  as in Assumption 3.2 by concentration inequality. See Appendix G.4 for a more detailed derivation. On the other hand, we only need  $\lambda_{K-1}(V)$  to be lower bounded. This assumption of SP is relaxed in [8] where similar to our method, one only needs bounds on  $\lambda_{K-1}(V)$  to guarantee theoretical results for SP.
3. As mentioned previously, by concentration inequality for sub-Gaussian random variable (Appendix G.4),  $\max_{1 \leq i \leq n} \|\epsilon_i\| = O_P(\sigma_n \sqrt{K \log(n)})$ . When  $\epsilon_i$  is sub-Gaussian with scale parameter  $\sigma_n$  as in Assumption 3.2. Define  $\hat{V}_{\tau(k)}^{(SP)} = [\hat{v}_{\tau(1)}^{(SP)}, \dots, \hat{v}_{\tau(K)}^{(SP)}]$ . By the results in Lemma G.1, we have

$$\begin{aligned} \|\hat{V}_{\tau(k)}^{(SP)} - V\| &\leq \|\hat{V}_{\tau(k)}^{(SP)} - V\|_F \\ &\leq \sqrt{K} \max_{1 \leq k \leq K} \{\|\hat{v}_{\tau(k)}^{(SP)} - v_k\|\} \\ &\leq \sqrt{K} \left[1 + 80 \frac{\gamma^2(V)}{\lambda_K^2(V)}\right] \max_{1 \leq i \leq n} \|\epsilon_i\| \\ &\leq \sqrt{K} \left[1 + 80 \frac{\gamma^2(V)}{\lambda_K^2(V)}\right] O_P(\sigma_n \sqrt{K \log(n)}) \\ &\leq O_P(\sigma_n K \sqrt{\log(n)}). \end{aligned} \quad (\text{G.43})$$

Comparing the above result with (13), one can see that the extra information from  $\Pi_S$  render a error rate improvement of  $1/\sqrt{N}$ . When  $N = 30$ , this leads to a 80% decrease of the error. On the other hand, the error rate of SP grows slower than ours as  $K$  increases. However, when  $K$  is large, the pure label assumption of SP can be very hard to satisfied: it is very likely that for some  $k \in \{1, 2, \dots, K\}$ , there do not exists  $i \in \{1, 2, \dots, n\}$  such that  $\pi_i = e_k$ . Without this pure label condition, SP algorithm will lose validity, while our method still remains to work.

We further compare SP with our method in Figure 1 using a simulated network from the DCMM with  $(n, K) = (3000, 3)$ . The model parameters are generated as follows:  $\theta_i$ 's are i.i.d. from

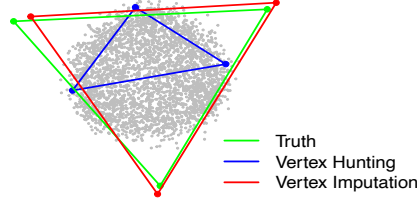


Figure 1: A comparison of VH and VI. The network is simulated from a DCMM with  $(n, K) = (3000, 3)$ , where 2% of nodes are labeled. Each grey point is a node embedding  $\hat{r}_i \in \mathbb{R}^2$ . The VH approach relies on existence of pure nodes (which is not satisfied in this example). The VI approach bypasses the pure node assumption (by leveraging on labeled nodes) and yields much better accuracy.

Uniform( $\sqrt{0.95}, 1$ ),  $\pi_i$ 's are i.i.d. from a disk inside the standard probability simplex, the diagonals of  $P$  are 1, and its off-diagonals are i.i.d. from Uniform(0, 0.2). We randomly pick 2% nodes to label. In Figure 1, the green triangle is the true simplex, and the grey dots are the node embeddings (since  $K = 3$ , each  $\hat{r}_i$  is a 2-dimensional vector). In SP, we ignore the given labels and solve an unsupervised problem. Since there is no pure node in this example, the estimated simplex by VH (in blue) is significantly different from the ground truth. In VI, we leverage on locations of those labeled nodes in the point cloud to “infer” where pure nodes should be located. The estimated simplex by our vertex imputation algorithm (in red) is reasonably close to the ground truth.

In all, compared to SP, our method relaxes the pure points assumption and attains considerable improvement in the estimation error with only a few labeled points.

## G.2 Sketched Vertex Search

It is widely observed that SP is sensitive to outliers. Several robust VH algorithms were introduced in Section 3.4 of [10]. One of them is SVS, which was originally proposed in [7]. SVS has a tuning integer  $L \geq K$  and runs as follows:

- Denoise by k-means: Run the k-means algorithm on  $\{\hat{r}_i\}_{i=1}^n$  assuming  $L$  clusters. Let  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_L$  be the cluster centers output by k-means.
- Exhaustive vertex search: For every  $K$ -tuples  $1 \leq \ell_1 < \dots < \ell_K \leq L$ , consider the simplex formed by  $\hat{y}_{\ell_1}, \dots, \hat{y}_{\ell_K}$ , and let  $d(\ell_1, \dots, \ell_K)$  be the maximum Euclidean distance from any  $\hat{y}_\ell$  outside the simplex to this simplex (such a distance can be computed by a standard quadratic programming). Choose  $(\ell_1^*, \dots, \ell_K^*)$  to minimize  $d(\ell_1, \dots, \ell_K)$ . Let  $\hat{v}_k = \hat{y}_{\ell_k^*}$ ,  $1 \leq k \leq K$ .

The first step in SVS uses k-means to “denoise” the original point cloud, where each cluster center  $\hat{y}_\ell$  is an average of nearby  $\hat{r}_i$ 's, thus less noisy. In the second step of SVS, each output vertex is one of the previous cluster centers. In our empirical study in Section 5, we set  $L$  as  $10 \times K$ . The results suggest that MSCORE-SVS outperforms MSCORE-SP.

However, since SVS exhaustively searches for  $K$  out of  $L$  indices, it is computationally expensive for large  $(K, L)$ . To tackle this issue, [10] proposed several variants of SVS. In one variant, they replaced the exhaustive search in Step 2 by running successive projection on  $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_L$ ; in another variant, they replaced Step 1 by using k-nearest-neighbor to denoise. These two modifications were also combined to create other variants of SVS [10].

## G.3 Minimum Volume Transformation

MVT [2, 4, 11] is another popular VH algorithm. For any  $v_1, v_2, \dots, v_K$ , denote by  $\mathcal{S}(v_1, \dots, v_K)$  the simplex whose vertices are  $v_1, \dots, v_K$ . MVT solves the following optimization:

$$\begin{aligned} \min_{v_1, \dots, v_K} \quad & \text{Volume}(\mathcal{S}(v_1, \dots, v_K)), \\ \text{subject to:} \quad & \hat{r}_i \in \mathcal{S}(v_1, \dots, v_K), \quad 1 \leq i \leq n. \end{aligned}$$

In other words, MVT finds the smallest-volume simplex that contains the whole point cloud.



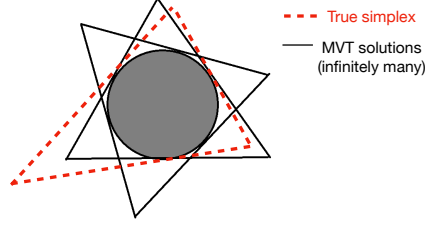


Figure 2: Why MVT does not work for the example in Figure 1. In this example,  $r_i$ 's are contained in a disk inside the true simplex. There are infinitely many MVT solutions (by rotation), but none of them is the true simplex.

For both SP and SVS, the estimated simplex is always in the convex hull of  $\hat{r}_1, \hat{r}_2, \dots, \hat{r}_n$ . In contrast, SVS can output a simplex with vertices far outside the point cloud. This yields a question: If we consider MSCORE-MVT, can we relax the requirement of existence of pure nodes? The answer is NO. We take the example in Figure 1. For simplicity, we consider the noiseless case where  $\hat{r}_i = r_i$ . In this example, all  $r_i$ 's are contained in a 2-dimensional disk as shown in Figure 2. The minimum-volume simplex containing this disk is a circumscribed equilateral triangle. However, such triangles are non unique. In fact, we can rotate the triangle arbitrarily. The optimization in MVT has infinitely many solutions, and there is no guarantee that the true simplex is one of these solutions.

#### G.4 Error Rate of $\max_{1 \leq i \leq n} \|\epsilon_i\|$

For any  $i$  in  $1, \dots, n$ , according to Assumption 3.2, the entries of  $\epsilon_i$  are independent, and sub-Gaussian with scale parameter  $\sigma_n$ . Hence, by Hanson–Wright inequality [13, Theorem 1.1], there exists an absolute constant  $c_8$  not depending on any constants in our paper, such that for any  $t > 0$ ,

$$\begin{aligned} \mathbb{P}(\epsilon'_i \epsilon_i - \mathbb{E}[\epsilon'_i \epsilon_i] > t) &\leq 2 \exp \left( -c_8 \min \left\{ \frac{t^2}{\sigma_n^4 \|I_K\|_F^2}, \frac{t}{\sigma_n^2 \|I_K\|} \right\} \right) \\ &= 2 \exp \left( -c_8 \min \left\{ \frac{t^2}{\sigma_n^4 K^2}, \frac{t}{\sigma_n^2} \right\} \right). \end{aligned}$$

Choose  $t = 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)$ , we have

$$\mathbb{P}(\epsilon'_i \epsilon_i - \mathbb{E}[\epsilon'_i \epsilon_i] > 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)) \leq \frac{2}{n^2}.$$

Since the entries of  $\epsilon_i$  are independent and sub-Gaussian with scale parameter  $\sigma_n$ ,  $\mathbb{E}[\epsilon'_i \epsilon_i] = \sum_{k=1}^K \mathbb{E}[\epsilon_{ik}^2] \leq \sigma_n^2 K$ . Therefore,

$$\begin{aligned} \mathbb{P}(\|\epsilon_i\|^2 > \sigma_n^2 K + 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)) &\leq \mathbb{P}(\|\epsilon_i\|^2 > \mathbb{E}[\epsilon'_i \epsilon_i] + 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)) \\ &\leq \mathbb{P}(\epsilon'_i \epsilon_i - \mathbb{E}[\epsilon'_i \epsilon_i] > 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)) \\ &\leq \frac{2}{n^2}, \end{aligned}$$

Consequently,

$$\begin{aligned} \mathbb{P}(\max_{1 \leq i \leq n} \|\epsilon_i\|^2 > \sigma_n^2 K + 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)) \\ &\leq \sum_{i=1}^n \mathbb{P}(\|\epsilon_i\|^2 > \sigma_n^2 K + 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n)) \\ &\leq \sum_{i=1}^n \frac{2}{n^2} \\ &= \frac{2}{n} \end{aligned}$$

To conclude, with probability at least,  $1 - \frac{2}{n}$ ,

$$\max_{1 \leq i \leq n} \|\epsilon_i\|^2 \leq \sigma_n^2 K + 2 \max\{c_8^{-1}, c_8^{-0.5}\} \sigma_n^2 K \log(n) = O(\sigma_n^2 K \log(n)).$$

Therefore,

$$\max_{1 \leq i \leq n} \|\epsilon_i\| = O_P\left(\sigma_n \sqrt{K \log(n)}\right).$$

□

## H Limitations and Future Extensions

In practice, the prior information on the distorted weight vector  $\pi_i$  may have various forms. Instead of observing its value, one may acquire the knowledge that  $\pi_i$  lies in a certain region. For instance, in the network mixed-membership estimation problem, it may be known that a certain individual has more inclination being in community 1 compared to community 2, so we may have constraints like  $\pi_i(1) \geq \pi_i(2)$ . It is curious how to leverage this sort of inequality information on vertex hunting problems. We think that a properly designed optimization based on our key observation Theorem 2.1 and the vertex imputation algorithm may cast light on a novel solution to this problem, and we believe that this is a thrilling future direction for our work.

## I Statement of Social Impact

In our main paper, we develop a novel semi-supervised vertex hunting algorithm leveraging the structural equation of the distortion matrix  $b$  discussed in Theorem 2.1. As illustrated in Section 4, our method can be applied to both network and text analysis, providing a positive social impact. Our algorithm's usage in mixed-membership estimation can increase the accuracy of recommendation systems built upon social networks, rendering more suitable content for each individuals according to their community label; the application of our method in semi-supervised topic model can help the sub-area classification and archive process of publications, leading to a faster literature search for researchers. One the other hand, we are aware that recommendation system is also a source of idea filtration and polarization, so our algorithm may have potential negative social impact by boosting recommendation system. Still, we believe that this impact is minor, because it is always easy to convert accurate estimations to less accurate one: people can add noise to the recommendation system to prevent self-enhancement of certain communities and information cocoons.

## References

- [1] Mário César Ugulino Araújo, Teresa Cristina Bezerra Saldanha, Roberto Kawakami Harrop Galvao, Takashi Yoneyama, Henrique Caldas Chame, and Valeria Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and intelligent laboratory systems*, 57(2):65–73, 2001.
- [2] Maurice D Craig. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 32(3):542–552, 1994.
- [3] Nicolas Gillis and Stephen A Vavasis. Fast and robust recursive algorithms for separable non-negative matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(4):698–714, 2013.
- [4] Eligius MT Hendrix, Inmaculada García, Javier Plaza, and Antonio Plaza. On the minimum volume simplex enclosure problem for estimating a linear mixing model. *Journal of Global Optimization*, 56(3):957–970, 2013.
- [5] David C Howell. Median absolute deviation. *Encyclopedia of statistics in behavioral science*, 2005.
- [6] Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4), December 2016.

- [7] Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Mixed membership estimation for social networks. *Journal of Econometrics*, 2023.
- [8] Jiashun Jin, Gabriel Moryoussef, Zheng Tracy Ke, Jiajun Tang, and Jingming Wang. Improved algorithm and bounds for successive projection. International Conference on learning and representations, 2024.
- [9] Zheng Tracy Ke, Pengsheng Ji, Jiashun Jin, and Wanshan Li. Recent advances in text analysis. *Annual Review of Statistics and Its Application*, 11(1):347–372, April 2024.
- [10] Zheng Tracy Ke and Jiashun Jin. Special invited paper: The SCORE normalization, especially for heterogeneous network and text data. *Stat*, 12(1):e545, 2023.
- [11] Jun Li, Alexander Agathos, Daniela Zaharie, José M Bioucas-Dias, Antonio Plaza, and Xia Li. Minimum volume simplex analysis: A fast algorithm for linear hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 53(9):5067–5082, 2015.
- [12] Kamyar Moshksar. On the absolute constant in hanson-wright inequality, 2024.
- [13] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *arXiv e-prints*, page arXiv:1306.2872, June 2013.
- [14] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We clearly explained the contributions of the paper in Section 1 using bullets points. We also presented a detailed related work review and emphasize the gap in the current literature.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We summarize our novel semi-supervised vertex hunting algorithm; also, we discuss the limitations and future extensions of our work in Appendix H.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We provide a full set of assumptions, Assumption 3.1-3.3 in Section 3. We provide a complete proof of all the theorems and lemmas in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We disclose all the information needed to reproduce our experiments in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data and code for our experiments are available in the supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the experiment details, such as the parameters/simulation settings we choose, are available in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report the error bars of our result in Table 3 in the form of median absolute deviation (MAD) [5]. We use this robust statistics for error bar to alleviate the effect of extreme cases in the experiments where all of or most of the methods perform very poorly.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Our experiments do not require any computer resources. All the experiments in our paper can be implemented on a personal computer.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We strictly conform the NeurIPS Code of Ethics, without any form of plagiarism or the use of LLM models.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: As illustrated in Section 4, our method can be applied to both network and text analysis, providing a positive social impact. We discuss the possible social impact of our algorithm in Appendix I.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper proposes a vertex hunting problem, which we do not think has any risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cite the sources of the data we use, [6, 9] in Section 5.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.



- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: The details of our method is well explained in Section 2, and the code for our algorithm is available in the supplementary materials.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Our paper does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.