Appendix A PROOFS OF THEOREMS FOR NOISELESS SSC

We prove Theorem 3.1, the main theorem for the noiseless clustering consistent SSC algorithm given in Sec. 3.1

Proof of Theorem 3.1. Fix a connected component $G_r = (V_r, E_r) \subseteq G$. By the self-expressiveness property we know that all data points in V_r lie on the same underlying subspace $\mathcal{S}^{(\ell)}$. It can be easily shown that if $\mathbf{X}^{(\ell)}$ is in general position then $|V_r| \geq d_\ell + 1$ because for any $\boldsymbol{x}_i \in \mathcal{S}^{(\ell)}$, at least d_ℓ other data points in the same subspace are required to perfectly reconstruct $\boldsymbol{x}_i$. Consequently, we have $\hat{\mathcal{S}}_{(r)} = \mathcal{S}^{(\ell)}$ because V_r contains at least d_ℓ data points in $\mathcal{S}^{(\ell)}$ that are linear independent. On the other hand, due to the self-expressiveness property, for every $\ell = 1, \cdots, L$ there exists a connected component G_r such that $\hat{\mathcal{S}}_{(r)} = \mathcal{S}^{(\ell)}$ because otherwise nodes in $\mathbf{X}^{(\ell)}$ will have no edges attached, which contradicts Eq. (3.1) and the definition of G. As a result, the above argument shows that Algorithm 1 achieves perfect subspace recovery; that is, there exists a permutation π on $[L]$ such that $\hat{\mathcal{S}}^{(\ell)} = \mathcal{S}^{(\pi(\ell))}$ for all $\ell = 1, \cdots, L$.

We next prove that Algorithm 1 achieves perfect clustering as well, that is, $\pi(\hat{z}_i) = z_i$ for every $i = 1, \cdots, N$. Assume by way of contradiction that there exists i such that $\hat{z}_i = \ell$ and $z_i = \ell' \neq \pi(\ell)$. Let $G_r = (V_r, E_r) \subseteq G$ be the connected component in G that contains the node corresponding to $\boldsymbol{x}_i$. Since $\hat{z}_i = \ell$, by SEP and the above analysis we have $\hat{\mathcal{S}}_{(r)} = \hat{\mathcal{S}}^{(\ell)} = \mathcal{S}^{(\pi(\ell))}$. On the other hand, because $z_i = \ell'$ and data points in V_r are in general position, we have $\hat{\mathcal{S}}_{(r)} = \mathcal{S}^{(\ell')}$. Hence, $\mathcal{S}^{(\pi(\ell))} = \mathcal{S}^{(\ell')}$ with $\ell' \neq \pi(\ell)$, which contradicts the assumption that no two underlying subspaces are identical. ☐

Appendix B DISCUSSION ON IDENTIFIABILITY AND ℓ_0 FORMULATION OF NOISELESS SUBSPACE CLUSTERING

B.1 The identifiability of noiseless subspace clustering

If we use a more relaxed notion of identifiability, even the "general position" assumption could be dropped for consistent clustering. In Theorem B.1 we define such a relaxed notion of identifiability for the union-of-subspace structure.

Theorem B.1. *Any set of N data points in $\mathbb{R}^n$ has a partition that follows a union-of-subspace structure, where points in each subspaces are in general position. We call this partition the* minimal *union-of-subspace structure.*

Proof. Given a finite set $\mathcal{X} \subset \mathbb{R}^n$. We will algorithmically construct a minimal partition. Initialize set $\mathcal{Y} = \mathcal{X}$. Start with $k = 1$, do the following repeatedly until it fails, then increment k, until $\mathcal{Y} = \emptyset$: find the maximum number of points that lie in a hyperplane of dimension $(k+1)$, assign a new partition for these points and remove these points from $\mathcal{Y}$. It is clear that in this way, every partition is a distinct subspace and points in any subspace are in general position. ☐

One consequence of Theorem B.1 is that if SEP holds with respect to any minimal union-of-subspace structure (i.e., a minimal ground truth), then Algorithm 1 will recover the correct ground truth clustering. We remark that SEP does not hold for any finite subset of points in $\mathbb{R}^n$ if ℓ_1 regularization is used, unless the data satisfy certain separation conditions [19]. However, in Section B.2 we propose an ℓ_0 regularization problem which achieves SEP (and hence consistent clustering) for any $\mathcal{X} \subseteq \mathbb{R}^d$.

We note that the minimal union-of-subspace structure may not be unique. An example is that if there is one point in the intersection of two subspaces with equal dimension, then this point can be assigned to either subspaces. Now, suppose the intersection has dimension k, there can be at most k points in the intersection, otherwise these points will form a new k-dimension subspace and the original structure is no longer minimal.

B.2 The merit of ℓ_0-minimization and agnostic subspace clustering

A byproduct of our result is that it also addresses an interesting question of whether it is advantageous to use ℓ_0 over ℓ_1 minimization in subspace clustering, namely

$$\min_{\boldsymbol{c}_i \in \mathbb{R}^N} \ \|\boldsymbol{c}_i\|_0, \quad s.t. \ \ \boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i, \boldsymbol{c}_{ii} = 0. \tag{B.1}$$

If one poses this question to a compressive sensing researcher, the answer will most likely be yes, since ℓ_0 minimization is the original problem of interest and empirical evidence suggests that using iterative re-weighted ℓ_1 scheme to approximate ℓ_0 solutions often improves the quality of signal recovery. On the other hand, a statistician is most likely to answer just the opposite because ℓ_1 shrinkage would often significantly reduce the variance at the cost of a small amount of bias. A formal treatment of the latter intuition suggests that ℓ_1 regularized regression has strictly less "effective-degree-of-freedom" than the "ℓ_0 best-subset selection" [22], therefore generalizes better.

How about subspace clustering? Unlike ℓ_1 solution that is unique almost everywhere, ℓ_0 solutions will not be unique and it is easy to construct a largely disconnected graph based on optimal ℓ_0 solutions. Using the new observation that we do not actually need graph connectivity, we are able to establish that ℓ_0 minimization for SSC is indeed the ultimate answer for noiseless subspace clustering.

Theorem B.2. *Given any N points in $\mathbb{R}^d$, any solutions to the ℓ_0-variant of Algorithm 1 will partition the points into a minimal union-of-subspace structure.*

Proof. Define a *minimal* subspace with respect to point $\boldsymbol{x}_i$ in a set $\{\boldsymbol{x}_i\}_{i=1}^N$ to be the span of any points that minimizes (B.1) for i. Since the ordering of how data points are used does not matter in Algorithm 1, we can sort the points into an ascending order with respect to the dimensionality. Now the merging procedure of these subspaces into a unique set of subspaces is exactly the same as the construction in the proof of Theorem B.1. Therefore, all solutions of the ℓ_0 SSC are going to be the correct partition. □

With slightly more effort, it can be shown that the converse is also true. Therefore, the set of solutions of ℓ_0-SSC completely characterizes the set of minimal union-of-subspace structure for any set of points in $\mathbb{R}^d$. In contrast, ℓ_1-SSC requires additional separation condition to work. That said, it may well be the case in practice that ℓ_1-SSC works better for the noisy subspace clustering in the low signal-to-noise ratio regime. It will be an interesting direction to explore how iterative reweighted ℓ_1 minimizations and local optimization for ℓ_p-norm $(0 < p < 1)$ work in subspace clustering applications.

Appendix C PROOFS OF THEOREMS FOR NOISY SSC

The purpose of this section is to present a complete proof to Theorem 3.2, our main result concerning clustering consistent Lasso SSC on noisy data. We first present and prove two technical propositions that will be used later.

Proposition C.1. *Let $\boldsymbol{u}$ be an arbitrary vector in $\mathcal{S}^{(\ell)}$ with $\|\boldsymbol{u}\|_2 = 1$. Then $\max_{1\leq i\leq N_\ell, i\neq i^*} |\langle \boldsymbol{u}, \boldsymbol{x}_i^{(\ell)}\rangle| \geq \rho_\ell^{-i^*}$ for every $i^* = 1,\cdots,N_\ell$.*

Proof. For notational simplicity let $\mathbf{X}_{-i^*}^{(\ell)} = (\boldsymbol{x}_1^{(\ell)},\cdots,\boldsymbol{x}_{i^*-1}^{(\ell)},\boldsymbol{x}_{i^*+1}^{(\ell)},\cdots,\boldsymbol{x}_{N_\ell}^{(\ell)})$ and $\mathcal{Q}_{-i^*}^{(\ell)} = \mathrm{conv}(\pm\mathbf{X}_{-i^*}^{(\ell)})$. The objective of Proposition C.1 is to lower bound $\|\mathbf{X}_{-i^*}^{(\ell)\top}\boldsymbol{u}\|_\infty$ for any $\boldsymbol{u}\in\mathcal{S}^{(\ell)}$ with $\|\boldsymbol{u}\|_2 = 1$. By definition of the dual norm, $\|\mathbf{X}_{-i^*}^{(\ell)\top}\boldsymbol{u}\|_\infty$ is equal to the objective of the following optimization problem

$$\max_{\boldsymbol{c}\in\mathbb{R}^{N_\ell-1}} \langle \boldsymbol{u}, \mathbf{X}_{-i^*}^{(\ell)}\boldsymbol{c}\rangle \quad s.t.\ \ \|\boldsymbol{c}\|_1 = 1. \tag{C.1}$$

To obtain a lower bound on the objective of Eq. (C.1), note that $\rho_\ell^{-i^*}$ is the radius of the largest ball inscribed in $\mathcal{Q}_{-i^*}^{(\ell)}$ and hence $\rho_\ell^{-i^*}\boldsymbol{u}\in\mathcal{Q}_{-i^*}^{(\ell)}$. Consequently, $\rho_\ell^{-i^*}\boldsymbol{u}$ can be written as a convex combination of (signed) columns in $\mathbf{X}_{-i^*}^{(\ell)}$, that is, there exists $\boldsymbol{c}\in\mathbb{R}^{N_\ell-1}$ with $\|\boldsymbol{c}\|_1 = 1$ such that $\mathbf{X}_{-i^*}^{(\ell)}\boldsymbol{c} = \rho_\ell^{-i^*}\boldsymbol{u}$. Plugging the expression into Eq. (C.1) we obtain

$$\|\mathbf{X}_{-i^*}^{(\ell)\top}\boldsymbol{u}\|_\infty \geq \langle \boldsymbol{u}, \rho_\ell^{-i^*}\boldsymbol{u}\rangle = \rho_\ell^{-i^*}.$$

□

Proposition C.2. *Let $\mathbf{A} = (\boldsymbol{a}_1,\cdots,\boldsymbol{a}_m)$ be an arbitrary matrix with at least m rows. Then $\|\boldsymbol{a}_i - \mathcal{P}_{\mathrm{Range}(\boldsymbol{a}_{-i})}(\boldsymbol{a}_i)\|_2 \geq \sigma_m(\mathbf{A})$, where $\boldsymbol{a}_{-i}$ denotes all columns in $\mathbf{A}$ except $\boldsymbol{a}_i$.*

Proof. Denote $\boldsymbol{a}_i^\perp$ as $\boldsymbol{a}_i^\perp = \boldsymbol{a}_i - \mathcal{P}_{\mathrm{Range}(\boldsymbol{a}_{-i})}(\boldsymbol{a}_i)$. By definition, $\boldsymbol{a}_i^\perp \in \mathrm{Range}(\mathbf{A})$ and $\langle \boldsymbol{a}_i^\perp, \boldsymbol{a}_{i'}\rangle = 0$ for all $i' \neq i$. Consequently,

$$\sigma_m(\mathbf{A}) \leq \inf_{\boldsymbol{u}\in\mathrm{Range}(\mathbf{A})} \frac{\|\mathbf{A}\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \leq \frac{\|\mathbf{A}\boldsymbol{a}_i^\perp\|_2}{\|\boldsymbol{a}_i^\perp\|_2} = \frac{\langle \boldsymbol{a}_i, \boldsymbol{a}_i^\perp\rangle}{\|\boldsymbol{a}_i^\perp\|_2} = \frac{\|\boldsymbol{a}_i^\perp\|_2^2}{\|\boldsymbol{a}_i^\perp\|_2} = \|\boldsymbol{a}_i^\perp\|_2.$$

□

We next present two key lemmas. The first lemma, Lemma C.1, shows that the estimated subspace $\hat{\mathcal{S}}$ from noisy inputs is a good approximation the underlying subspace $\mathcal{S}^{(\ell)}$ as long as the restricted eigenvalue assumption holds and exactly d points from the same subspace are used to construct $\hat{\mathcal{S}}$.

Lemma C.1. *Fix $\ell \in \{1, \cdots, L\}$. Suppose $\hat{\mathcal{S}}$ is the range of a subset of points $\mathbf{Y}_d \subseteq \mathbf{Y}^{(\ell)}$ containing exactly d noisy data points belonging to the ℓth subspace. Let $\mathcal{S}^{(\ell)}$ be the ground-truth subspace; i.e., $\boldsymbol{x}_1^{(\ell)}, \cdots, \boldsymbol{x}_{N_\ell}^{(\ell)} \in \mathcal{S}^{(\ell)}$. Under Assumption 3.1 we have*

$$d(\hat{\mathcal{S}}, \mathcal{S}^{(\ell)}) \leq \frac{2d\xi^2}{\sigma_\ell^2}. \tag{C.2}$$

Proof. Suppose $\mathbf{Y}_d = (\boldsymbol{y}_{i_1}^{(\ell)}, \cdots, \boldsymbol{y}_{i_d}^{(\ell)})$ and $\mathbf{X}_d = (\boldsymbol{x}_{i_1}^{(\ell)}, \cdots, \boldsymbol{x}_{i_d}^{(\ell)})$. By the noise model $\|\mathbf{Y}_d - \mathbf{X}_d\|_F^2 = \sum_{j=1}^d \|\boldsymbol{\varepsilon}_{i_j}\|_2^2 \leq d\xi^2$. On the other hand, by Assumption 3.1 we have $\sigma_d(\mathbf{X}_d) \geq \sigma_\ell$. Wedin's theorem (Lemma D.1 in Appendix D) then yields the lemma. □

In Lemma C.2 we show that if the restricted eigenvalue assumption holds and the regularization parameter λ is in a certain range, the optimal solution to the Lasso problem in Eq. (3.2) has at least d nonzero coefficients, which lead to $|V_r| \geq d+1$ for every connected component V_r in the similarity graph constructed in Algorithm 2. Lemma C.2 is a natural extension to the fact that at least d points should be used to reconstruct a certain data point for noiseless inputs, if the data matrix $\mathbf{X}$ is in general position.

Lemma C.2. *Assume Assumption 3.1 and the self-expressiveness property hold. For each $i \in \{1, \cdots, N\}$, $\|\boldsymbol{c}_i\|_0 \geq d$ if the regularization parameter λ satisfies*

$$2\xi(1+\xi)^2(1+1/\rho_\ell) < \lambda < \frac{\rho_\ell\sigma_\ell}{2}, \quad \ell = 1, \cdots, L. \tag{C.3}$$

Proof. Because the self-expressiveness property holds, we assume without loss of generality that the support set of $\boldsymbol{c}_i$ with $\|\boldsymbol{c}_i\|_0 = t$ is $\{\boldsymbol{y}_1^{(\ell)}, \cdots, \boldsymbol{y}_t^{(\ell)}\}$. Assume by way of contradiction that $\|\boldsymbol{c}_i\|_0 < d$ and define $\boldsymbol{y}^\perp = \boldsymbol{y}_i^{(\ell)} - \sum_{j=1}^{d-1} c_{i,j}\boldsymbol{y}_j^{(\ell)}$, where $c_{i,1}, \cdots, c_{i,d-1}$ contain all nonzero coefficients [4] in $\boldsymbol{c}_i$. Since $\boldsymbol{c}_i$ is optimal, the following must hold for every $\boldsymbol{y}_{i'}^{(\ell)}$ with $i' \neq i$:

$$\mathrm{argmin}_{c\in\mathbb{R}} \left\{ \|\boldsymbol{y}^\perp - c\boldsymbol{y}_{i'}^{(\ell)}\|_2^2 + 2\lambda|c| \right\} = 0. \tag{C.4}$$

To see the necessity of Eq. (C.4), note that the optimal solution to Eq. (C.4) $c^* \neq 0$ implies

$$\|\boldsymbol{y}_i^{(\ell)} - \mathbf{Y}_{-i}^{(\ell)}\tilde{\boldsymbol{c}}_i\|_2^2 + 2\lambda\|\tilde{\boldsymbol{c}}_i\|_1 \leq \|\boldsymbol{y}^\perp - c^*\boldsymbol{y}_{i'}^{(\ell)}\|_2^2 + 2\lambda|c^*| + 2\lambda\|\boldsymbol{c}_i\|_1 < \|\boldsymbol{y}^\perp\|_2^2 + 2\lambda\|\boldsymbol{c}_i\|_1 = \|\boldsymbol{y}_i^{(\ell)} - \mathbf{Y}_{-i}^{(\ell)}\boldsymbol{c}_i\|_2^2 + 2\lambda\|\boldsymbol{c}_i\|_1,$$

where $\tilde{\boldsymbol{c}}_i = \boldsymbol{c}_i + c^* \cdot \boldsymbol{e}_{i'}$. This contradicts the optimality of $\boldsymbol{c}_i$ with respect to Eq. (3.2).

By optimality conditions, Eq. (C.4) implies $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle| \leq \lambda$. In the remainder of the proof we will show that under the assumptions made in Lemma C.2, $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle| > \lambda$, which results in a contradiction.

In order to lower bound $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle|$ we first bound the noiseless version of the inner product $|\langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle|$, where $\boldsymbol{x}^\perp = \boldsymbol{x}_i^{(\ell)} - \sum_{j=1}^{d-1} c_{i,j}\boldsymbol{x}_j^{(\ell)}$. A key observation is that $\boldsymbol{x}^\perp \in \mathcal{S}^{(\ell)}$ and hence by Proposition C.1 and C.2 the following chain of inequality holds for any $\boldsymbol{x}_{i'}^{(\ell)}$ with $i' \neq i$:

$$\left|\langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle\right| \geq \rho_\ell\|\boldsymbol{x}^\perp\|_2 \geq \rho_\ell \left\| \boldsymbol{x}_i^{(\ell)} - \mathcal{P}_{\mathrm{span}(\boldsymbol{x}_{1:d-1}^{(\ell)})}(\boldsymbol{x}_i^{(\ell)}) \right\|_2 \geq \rho_\ell\sigma_\ell. \tag{C.5}$$

[4] Some coefficients in $c_{i_1}, \cdots, c_{i,d-1}$ might be zero because $\|\boldsymbol{c}_i\|_1$ could be smaller than $d-1$.

Our next objective is to upper bound the inner product perturbation $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle - \langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle|$ and subsequently obtain a lower bound on $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle|$. Note that

$$\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle = \langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle + \langle \boldsymbol{y}^\perp - \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle + \langle \boldsymbol{x}^\perp, \boldsymbol{y}_{i'}^{(\ell)} - \boldsymbol{x}_{i'}^{(\ell)}\rangle + \langle \boldsymbol{y}^\perp - \boldsymbol{x}^\perp, \boldsymbol{y}_{i'}^{(\ell)} - \boldsymbol{x}_{i'}^{(\ell)}\rangle;$$

therefore,

$$\left|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle - \langle \boldsymbol{x}^\perp, \boldsymbol{x}_{i'}^{(\ell)}\rangle\right| \le \|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|\|\boldsymbol{x}_{i'}^{(\ell)}\| + \|\boldsymbol{y}^\perp\|\|\boldsymbol{y}_{i'}^{(\ell)} - \boldsymbol{x}_{i'}^{(\ell)}\| \le \|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2 + \xi\|\boldsymbol{y}^\perp\|_2. \tag{C.6}$$

In order to upper bound $\|\boldsymbol{y}^\perp\|_2$ and $\|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2$, note that by definition $\|\boldsymbol{y}^\perp\|_2 = \|\boldsymbol{y}_1^{(\ell)} - \sum_{j=2}^d c_{ij}\boldsymbol{y}_j^{(\ell)}\|_2 \le (1+\|\boldsymbol{c}_i\|_1)(1+\xi)$ and $\|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2 = \|\boldsymbol{\varepsilon}_1^{(\ell)} - \sum_{j=2}^d c_{ij}\boldsymbol{y}_j^{(\ell)}\|_2 \le \xi(1+\|\boldsymbol{c}_i\|_1)$. Hence we only need to upper bound $\|\boldsymbol{c}_i\|_1$, which can be done by the following argument due to the optimality of $\boldsymbol{c}_i$: By arguments on page 21 in [24], the following upper bound on $\|\boldsymbol{c}_i\|_1$ is proven:

$$\|\boldsymbol{c}_i\|_1 \le \frac{1}{\rho_\ell} + \frac{\xi^2}{\lambda}\left(1 + \frac{1}{\rho_\ell}\right)^2. \tag{C.7}$$

The lower bound on λ in Eq. (C.3) implies that $\xi < \lambda(1+1/\rho_\ell)$. Plugging this upper bound into Eq. (C.7) we obtain

$$\|\boldsymbol{c}_i\|_1 \le 1/\rho_\ell + \xi(1+1/\rho_\ell) \le (1+\xi)(1+1/\rho_\ell), \tag{C.8}$$

which eliminates the dependency on λ. We now substitute the simplified upper bound on $\|\boldsymbol{c}_i\|_1$ into the upper bound for $\|\boldsymbol{y}^\perp\|_2$, $\|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2$ and get

$$\|\boldsymbol{y}^\perp\|_2 \le (1+\xi)^2(1+1/\rho_\ell); \quad \|\boldsymbol{y}^\perp - \boldsymbol{x}^\perp\|_2 \le \xi(1+\xi)(1+1/\rho_\ell). \tag{C.9}$$

Combining Eq. (C.5), (C.6) and (C.9) we obtain the following lower bound on $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle|$:

$$\left|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle\right| \ge \rho_\ell\sigma_\ell - 2\xi(1+\xi)^2(1+1/\rho_\ell) \ge \frac{1}{2}\rho_\ell\sigma_\ell, \tag{C.10}$$

where the last inequality is due to the assumption that $2\xi(1+\xi)^2(1+1/\rho_\ell) < \frac{1}{2}\rho_\ell\sigma_\ell$ implied by Eq. (C.3). Finally, since $\frac{1}{2}\rho_\ell\sigma_\ell > \lambda$ as assumed in Eq. (C.3), we have $|\langle \boldsymbol{y}^\perp, \boldsymbol{y}_{i'}^{(\ell)}\rangle| > \lambda$, which results in the desired contradiction. $\square$

Finally, Theorem 3.2 is a simple consequence of Lemma C.1 and C.2 because under the conditions of Lemma C.2, every component V_r will have at least d data points. Define $\mu_\epsilon = \sqrt{2d\xi^2/\min_\ell \sigma_\ell^2}$. Lemma C.1 implies that $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')}) \le \mu_\epsilon$ if V_r and $V_{r'}$ belong to the same cluster. On the other hand, by the separation condition in Eq. (3.6) and Lemma C.1, if V_r and $V_{r'}$ belong to different clusters we would have $d(\hat{\mathcal{S}}_{(r)}, \hat{\mathcal{S}}_{(r')}) > \mu_\epsilon$. Therefore, the single-linkage clustering procedure in Algorithm 2 will eventually merge estimated subspaces correectly.

Appendix D MATRIX PERTURBATION THEOREMS

Lemma D.1 (Wedin's theorem; Theorem 4.1, pp. 260 in [21])**.** *Let $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{m\times n}$ be given matrices with $m \ge n$. Let $\mathbf{A}$ have the following singular value decomposition*

$$\begin{bmatrix} \mathbf{U}_1^\top \\ \mathbf{U}_2^\top \\ \mathbf{U}_3^\top \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_2 \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{V}_1, \mathbf{V}_2$ have orthonormal columns and $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal matrices. Let $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$ be a perturbed version of $\mathbf{A}$ and $(\widetilde{\mathbf{U}}_1, \widetilde{\mathbf{U}}_2, \widetilde{\mathbf{U}}_3, \widetilde{\mathbf{V}}_1, \widetilde{\mathbf{V}}_2, \widetilde{\mathbf{\Sigma}}_1, \widetilde{\mathbf{\Sigma}}_2)$ be analogous singular value decomposition of $\widetilde{\mathbf{A}}$. Let $\mathbf{\Phi}$ be the matrix of canonical angles between $\mathrm{Range}(\mathbf{U}_1)$ *and* $\mathrm{Range}(\widetilde{\mathbf{U}}_1)$ *and $\mathbf{\Theta}$ be the matrix of canonical angles between* $\mathrm{Range}(\mathbf{V}_1)$ *and* $\mathrm{Range}(\widetilde{\mathbf{V}}_1)$*. If there exists $\delta > 0$ such that*

$$\min_{i,j}\left|[\mathbf{\Sigma}_1]_{i,i} - [\mathbf{\Sigma}_2]_{j,j}\right| > \delta \quad \textit{and} \quad \min_i \left|[\mathbf{\Sigma}_1]_{i,i}\right| > \delta,$$

then

$$\|\sin\mathbf{\Phi}\|_F^2 + \|\sin\mathbf{\Theta}\|_F^2 \le \frac{2\|\mathbf{E}\|_F^2}{\delta^2}.$$