Differentially Private Subspace Clustering

Yining Wang, Yu-Xiang Wang and Aarti Singh
Machine Learning Department, Carnegie Mellon Universty, Pittsburgh, USA
{yiningwa,yuxiangw,aarti}@cs.cmu.edu

Abstract

Subspace clustering is an unsupervised learning problem that aims at grouping data points into multiple "clusters" so that data points in a single cluster lie approximately on a low-dimensional linear subspace. It is originally motivated by 3D motion segmentation in computer vision, but has recently been generically applied to a wide range of statistical machine learning problems, which often involves sensitive datasets about human subjects. This raises a dire concern for data privacy. In this work, we build on the framework of *differential privacy* and present two provably private subspace clustering algorithms. We demonstrate via both theory and experiments that one of the presented methods enjoys formal privacy and utility guarantees; the other one asymptotically preserves differential privacy while having good performance in practice. Along the course of the proof, we also obtain two new provable guarantees for the agnostic subspace clustering and the graph connectivity problem which might be of independent interests.

1 Introduction

Subspace clustering was originally proposed to solve very specific computer vision problems having a union-of-subspace structure in the data, e.g., motion segmentation under an affine camera model [11] or face clustering under Lambertian illumination models [15]. As it gains increasing attention in the statistics and machine learning community, people start to use it as an agnostic learning tool in social network [5], movie recommendation [33] and biological datasets [19]. The growing applicability of subspace clustering in these new domains inevitably raises the concern of *data privacy*, as many such applications involve dealing with sensitive information. For example, [19] applies subspace clustering to identify diseases from personalized medical data and [33] in fact uses subspace clustering as a effective tool to conduct linkage attacks on individuals in movie rating datasets. Nevertheless, privacy issues in subspace clustering have been less explored in the past literature, with the only exception of a brief analysis and discussion in [29]. However, the algorithms and analysis presented in [29] have several notable deficiencies. For example, data points are assumed to be incoherent and it only protects the differential privacy of any feature of a user rather than the entire user profile in the database. The latter means it is possible for an attacker to infer with high confidence whether a particular user is in the database, given sufficient side information.

It is perhaps reasonable why there is little work focusing on private subspace clustering, which is by all means a challenging task. For example, a negative result in [29] shows that if utility is measured in terms of exact clustering, then no private subspace clustering algorithm exists when neighboring databases are allowed to differ on an entire user profile. In addition, state-of-the-art subspace clustering methods like Sparse Subspace Clustering (SSC, [11]) lack a complete analysis of its clustering output, thanks to the notorious "graph connectivity" problem [21]. Finally, clustering could have high global sensitivity even if only cluster centers are released, as depicted in Figure 1. As a result, general private data releasing schemes like output perturbation [7, 8, 2] do not apply.

In this work, we present a systematic and principled treatment of differentially private subspace clustering. To circumvent the negative result in [29], we use the perturbation of recovered low-

dimensional subspace from the ground truth as the utility measure. Our contributions are two-fold. First, we analyze two efficient algorithms based on the sample-aggregate framework [22] and established formal privacy and utility guarantees when data are generated from some stochastic model or satisfy certain deterministic separation conditions. New results on (non-private) subspace clustering are obtained along our analysis, including a *fully agnostic* subspace clustering on well-separated datasets using stability arguments and *exact clustering* guarantee for thresholding-based subspace clustering (TSC, [14]) in the noisy setting. In addition, we employ the exponential mechanism [18] and propose a novel Gibbs sampler for sampling from this distribution, which involves a novel tweak in sampling from a matrix Bingham distribution. The method works well in practice and we show it is closely related to the well-known mixtures of probabilistic PCA model [27].

Related work Subspace clustering can be thought as a generalization of PCA and k-means clustering. The former aims at finding a *single* low-dimensional subspace and the latter uses zero-dimensional subspaces as cluster centers. There has been extensive research on private PCA [2, 4, 10] and k-means [2, 22, 26]. Perhaps the most similar work to ours is [22, 4]. [22] applies the sample-aggregate framework to k-means clustering and [4] employs the exponential mechanism to recover private principal vectors. In this paper we give non-trivial generalization of both work to the private subspace clustering setting.

2 Preliminaries

2.1 Notations

For a vector $\boldsymbol{x} \in \mathbb{R}^d$, its p-norm is defined as $\|\boldsymbol{x}\|_p = (\sum_i \boldsymbol{x}_i^p)^{1/p}$. If p is not explicitly specified then the 2-norm is used. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, we use $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_n(\mathbf{A}) \geq 0$ to denote its singular values (assuming without loss of generality that $n \leq m$). We use $\|\cdot\|_\xi$ to denote matrix norms, with $\xi = 2$ the matrix spectral norm and $\xi = F$ the Frobenious norm. That is, $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ and $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^n \sigma_i(\mathbf{A})^2}$. For a q-dimensional subspace $\mathcal{S} \subseteq \mathbb{R}^d$, we associate with a basis $\mathbf{U} \in \mathbb{R}^{d\times q}$, where the q columns in $\mathbf{U}$ are orthonormal and $\mathcal{S} = \mathrm{range}(\mathbf{U})$. We use $\mathbb{S}_q^d$ to denote the set of all q-dimensional subspaces in $\mathbb{R}^d$.

Given $\boldsymbol{x} \in \mathbb{R}^d$ and $\mathcal{S} \subseteq \mathbb{R}^d$, the distance $d(\boldsymbol{x}, \mathcal{S})$ is defined as $d(\boldsymbol{x}, \mathcal{S}) = \inf_{\boldsymbol{y}\in\mathcal{S}} \|\boldsymbol{x} - \boldsymbol{y}\|_2$. If $\mathcal{S}$ is a subspace associated with a basis $\mathcal{U}$, then we have $d(\boldsymbol{x}, \mathcal{S}) = \|\boldsymbol{x} - \mathcal{P}_\mathcal{S}(\boldsymbol{x})\|_2 = \|\boldsymbol{x} - \mathbf{U}\mathbf{U}^\top \boldsymbol{x}\|_2$, where $\mathcal{P}_\mathcal{S}(\cdot)$ denotes the projection operator onto subspace $\mathcal{S}$. For two subspaces $\mathcal{S}, \mathcal{S}'$ of dimension q, the distance $d(\mathcal{S}, \mathcal{S}')$ is defined as the Frobenious norm of the $\sin$ matrix of principal angles; i.e.,

$$d(\mathcal{S}, \mathcal{S}') = \|\sin\boldsymbol{\Theta}(\mathcal{S}, \mathcal{S}')\|_F = \|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_F, \tag{1}$$

where $\mathbf{U}, \mathbf{U}'$ are orthonormal basis associated with $\mathcal{S}$ and $\mathcal{S}'$, respectively.

2.2 Subspace clustering

Given n data points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$, the task of subspace clustering is to cluster the data points into k clusters so that data points within a subspace lie approximately on a low-dimensional subspace. Without loss of generality, we assume $\|\boldsymbol{x}_i\|_2 \leq 1$ for all $i = 1, \cdots, n$. We also use $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ to denote the dataset and $\mathbf{X} \in \mathbb{R}^{d\times n}$ to denote the data matrix by stacking all data points in columnwise order. Subspace clustering seeks to find k q-dimensional subspaces $\hat{\mathcal{C}} = \{\hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_k\}$ so as to minimize the Wasserstein's distance or distance squared defined as

$$d_W^2(\hat{\mathcal{C}}, \mathcal{C}^*) = \min_{\pi:[k]\to[k]} \sum_{i=1}^k d^2(\hat{\mathcal{S}}_i, \mathcal{S}^*_{\pi(i)}), \tag{2}$$

where π are taken over all permutations on $[k]$ and $\mathcal{S}^*$ are the optimal/ground-truth subspaces. In a model based approach, $\mathcal{C}^*$ is fixed and data points $\{\boldsymbol{x}_i\}_{i=1}^n$ are generated either deterministically or stochastically from one of the ground-truth subspaces in $\mathcal{C}^*$ with noise corruption; for a completely agnostic setting, $\mathcal{C}^*$ is defined as the minimizer of the k-means subspace clustering objective:

$$\mathcal{C}^* := \mathrm{argmin}_{\mathcal{C}=\{\mathcal{S}_1,\cdots,\mathcal{S}_k\}\subseteq\mathbb{S}_q^d} \mathrm{cost}(\mathcal{C}; \mathcal{X}) = \mathrm{argmin}_{\mathcal{C}=\{\mathcal{S}_1,\cdots,\mathcal{S}_k\}\subseteq\mathbb{S}_q^d} \frac{1}{n}\sum_{i=1}^n \min_j d^2(\boldsymbol{x}_i, \mathcal{S}_j). \tag{3}$$

To simplify notations, we use $\Delta_k(\mathcal{X}) = \mathrm{cost}(\mathcal{C}^*; \mathcal{X})$ to denote cost of the optimal solution.

Algorithm 1 The sample-aggregate framework [22]

1: **Input**: $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, number of subsets m, privacy parameters ε, δ; f, $d_{\mathcal{M}}$.
2: **Initialize**: $s = \sqrt{m}$, $\alpha = \varepsilon/(5\sqrt{2\ln(2/\delta)})$, $\beta = \varepsilon/(4(D + \ln(2/\delta)))$.
3: **Subsampling**: Select m random subsets of size n/m of $\mathcal{X}$ independently and uniformly at random without replacement. Repeat this step until no single data point appears in more than $\sqrt{m}$ of the sets. Mark the subsampled subsets $\mathcal{X}_{S_1}, \cdots, \mathcal{X}_{S_m}$.
4: **Separate queries**: Compute $\mathcal{B} = \{\boldsymbol{s}_i\}_{i=1}^m \subseteq \mathbb{R}^D$, where $\boldsymbol{s}_i = f(\mathcal{X}_{S_i})$.
5: **Aggregation**: Compute $g(\mathcal{B}) = \boldsymbol{s}_{i^*}$ where $i^* = \operatorname{argmin}_{i=1}^m r_i(t_0)$ with $t_0 = (\frac{m+s}{2} + 1)$. Here $r_i(t_0)$ denotes the distance $d_{\mathcal{M}}(\cdot,\cdot)$ between $\boldsymbol{s}_i$ and the t_0-th nearest neighbor to $\boldsymbol{s}_i$ in $\mathcal{B}$.
6: **Noise calibration**: Compute $S(\mathcal{B}) = 2\max_k(\rho(t_0 + (k+1)s) \cdot e^{-\beta k})$, where $\rho(t)$ is the mean of the top $\lfloor s/\beta \rfloor$ values in $\{r_1(t), \cdots, r_m(t)\}$.
7: **Output**: $\mathcal{A}(\mathcal{X}) = g(\mathcal{B}) + \frac{S(\mathcal{B})}{\alpha}\boldsymbol{u}$, where $\boldsymbol{u}$ is a standard Gaussian random vector.

2.3 Differential privacy

Definition 2.1 (Differential privacy, [7, 8]). *A randomized algorithm $\mathcal{A}$ is (ε, δ)-*differentially private *if for all $\mathcal{X}, \mathcal{Y}$ satisfying $d(\mathcal{X}, \mathcal{Y}) = 1$ and all sets S of possible outputs the following holds:*

$$\Pr[\mathcal{A}(\mathcal{X}) \in S] \leq e^{\varepsilon} \Pr[\mathcal{A}(\mathcal{Y}) \in S] + \delta. \tag{4}$$

In addition, if $\delta = 0$ then the algorithm $\mathcal{A}$ is ε-differentially private.

In our setting, the distance $d(\cdot,\cdot)$ between two datasets $\mathcal{X}$ and $\mathcal{Y}$ is defined as the number of different columns in $\mathbf{X}$ and $\mathbf{Y}$. Differential privacy ensures the output distribution is obfuscated to the point that every user has a plausible deniability about being in the dataset, and in addition any inferences about individual user will have nearly the same confidence before and after the private release.

3 Sample-aggregation based private subspace clustering

In this section we first summarize the sample-aggregate framework introduced in [22] and argue why it should be preferred to conventional output perturbation mechanisms [7, 8] for subspace clustering. We then analyze two efficient algorithms based on the sample-aggregate framework and prove formal privacy and utility guarantees. We also prove new results in our analysis regarding the stability of k-means subspace clustering (Lem. 3.3) and graph connectivity (i.e., consistency) of noisy threshold-based subspace clustering (TSC, [14]) under a stochastic model (Lem. 3.5).

3.1 Smooth local sensitivity and the sample-aggregate framework

Most existing privacy frameworks [7, 8] are based on the idea of *global sensitivity*, which is defined as the maximum output perturbation $\|f(\mathcal{X}_1) - f(\mathcal{X}_2)\|_\xi$, where maximum is over all neighboring databases $\mathcal{X}_1, \mathcal{X}_2$ and $\xi = 1$ or 2. Unfortunately, global sensitivity of clustering problems is usually high even if only cluster centers are released. For example, Figure 1 shows that the global sensitivity of k-means subspace clustering could be as high as $O(1)$, which ruins the algorithm utility.

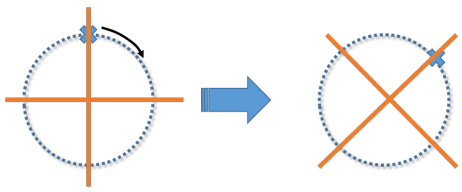

Figure 1: Illustration of instability of k-means subspace clustering solutions ($d = 2, k = 2, q = 1$). Blue dots represent evenly spaced data points on the unit circle; blue crosses indicate an additional data point. Red lines are optimal solutions.

To circumvent the above-mentioned challenges, Nissim et al. [22] introduces the sample-aggregate framework based on the concept of a smooth version of *local sensitivity*. Unlike global sensitivity, local sensitivity measures the maximum perturbation $\|f(\mathcal{X}) - f(\mathcal{X}')\|_\xi$ over all databases $\mathcal{X}'$ neighboring to the *input* database $\mathcal{X}$. The proposed sample-aggregate framework (pseudocode in Alg. 1) enjoys local sensitivity and comes with the following guarantee:

Theorem 3.1 ([22], Theorem 4.2). *Let $f : \mathbb{D} \to \mathbb{R}^D$ be an efficiently computable function where $\mathbb{D}$ is the collection of all databases and D is the output dimension. Let $d_{\mathcal{M}}(\cdot,\cdot)$ be a semimetric on*

the outer space of f. [1] *Set $\varepsilon > 2D/\sqrt{m}$ and $m = \omega(\log^2 n)$. The sample-aggregate algorithm $\mathcal{A}$ in Algorithm 1 is an efficient (ε, δ)-differentially private algorithm. Furthermore, if f and m are chosen such that the ℓ_1 norm of the output of f is bounded by Λ and*

$$\Pr_{\mathcal{X}_S \subseteq \mathcal{X}} [d_{\mathcal{M}}(f(\mathcal{X}_S), \boldsymbol{c}) \leq r] \geq \frac{3}{4} \tag{5}$$

for some $\boldsymbol{c} \in \mathbb{R}^D$ and $r > 0$, then the standard deviation of Gaussian noise added is upper bounded by $O(r/\varepsilon) + \frac{\Lambda}{\varepsilon} e^{-\Omega(\frac{\varepsilon\sqrt{m}}{D})}$. In addition, when m satisfies $m = \omega(D^2 \log^2(r/\Lambda)/\varepsilon^2)$, with high probability each coordinate of $\mathcal{A}(\mathcal{X}) - \bar{\boldsymbol{c}}$ is upper bounded by $O(r/\varepsilon)$, where $\bar{\boldsymbol{c}}$ depending on $\mathcal{A}(\mathcal{X})$ satisfies $d_{\mathcal{M}}(\boldsymbol{c}, \bar{\boldsymbol{c}}) = O(r)$.

Let f be any subspace clustering solver that outputs k estimated low-dimensional subspaces and $d_{\mathcal{M}}$ be the Wasserstein's distance as defined in Eq. (2). Theorem 3.1 provides privacy guarantee for an efficient meta-algorithm with any f. In addition, utility guarantee holds with some more assumptions on input dataset $\mathcal{X}$. In following sections we establish utility guarantees. The main idea is to prove stability results as outlined in Eq. (5) for particular subspace clustering solvers and then apply Theorem 3.1.

3.2 The agnostic setting

We first consider the setting when data points $\{\boldsymbol{x}_i\}_{i=1}^n$ are arbitrarily placed. Under such agnostic setting the optimal solution $\mathcal{C}^*$ is defined as the one that minimizes the k-means cost as in Eq. (3). The solver f is taken to be any $(1+\epsilon)$-approximation[2] of optimal k-means subspace clustering; that is, f always outputs subspaces $\hat{\mathcal{C}}$ satisfying $\text{cost}(\hat{\mathcal{C}}; \mathcal{X}) \leq (1+\epsilon)\text{cost}(\mathcal{C}^*; \mathcal{X})$. Efficient core-set based approximation algorithms exist, for example, in [12]. The key task of this section it to identify assumptions under which the stability condition in Eq. (5) holds with respect to an approximate solver f. The example given in Figure 1 also suggests that identifiability issue arises when the input data $\mathcal{X}$ itself cannot be well clustered. For example, no two straight lines could well approximate data uniformly distributed on a circle. To circumvent the above-mentioned difficulty, we impose the following well-separation condition on the input data $\mathcal{X}$:

Definition 3.2 (Well-separation condition for k-means subspace clustering)**.** *A dataset $\mathcal{X}$ is* (ϕ, η, ψ)-well separated *if there exist constants ϕ, η and ψ, all between 0 and 1, such that*

$$\Delta_k^2(\mathcal{X}) \leq \min\left\{\phi^2 \Delta_{k-1}^2(\mathcal{X}), \Delta_{k,-}^2(\mathcal{X}) - \psi, \Delta_{k,+}^2(\mathcal{X}) + \eta\right\}, \tag{6}$$

where Δ_{k-1}, $\Delta_{k,-}$ and $\Delta_{k,+}$ are defined as $\Delta_{k-1}^2(\mathcal{X}) = \min_{\mathcal{S}_{1:k-1} \in \mathbb{S}_q^d} \text{cost}(\{\mathcal{S}_i\}; \mathcal{X})$; $\Delta_{k,-}^2(\mathcal{X}) = \min_{\mathcal{S}_1 \in \mathbb{S}_{q-1}^d, \mathcal{S}_{2:k} \in \mathbb{S}_q^d} \text{cost}(\{\mathcal{S}_i\}; \mathcal{X})$; and $\Delta_{k,+}^2(\mathcal{X}) = \min_{\mathcal{S}_1 \in \mathbb{S}_{q+1}^d, \mathcal{S}_{2:k} \in \mathbb{S}_q^d} \text{cost}(\{\mathcal{S}_i\}; \mathcal{X})$.

The first condition in Eq. (6), $\Delta_k^2(\mathcal{X}) \leq \phi^2 \Delta_{k-1}^2(\mathcal{X})$, constrains that the input dataset $\mathcal{X}$ cannot be well clustered using $k-1$ instead of k clusters. It was introduced in [23] to analyze stability of k-means solutions. For subspace clustering, we need another two conditions regarding the intrinsic dimension of each subspace. The $\Delta_k^2(\mathcal{X}) \leq \Delta_{k,-}^2(\mathcal{X}) - \psi$ asserts that replacing a q-dimensional subspace with a $(q-1)$-dimensional one is not sufficient, while $\Delta_k^2(\mathcal{X}) \leq \Delta_{k,+}^2(\mathcal{X}) + \eta$ means an additional subspace dimension does not help much with clustering $\mathcal{X}$.

The following lemma is our main stability result for subspace clustering on well-separated datasets. It states that when a candidate clustering $\hat{\mathcal{C}}$ is close to the optimal clustering $\mathcal{C}^*$ in terms of clustering cost, they are also close in terms of the Wasserstein distance defined in Eq. (2).

Lemma 3.3 (Stability of agnostic k-means subspace clustering)**.** *Assume $\mathcal{X}$ is (ϕ, η, ψ)-well separated with $\phi^2 < 1/1602$, $\psi > \eta$. Suppose a candidate clustering $\hat{\mathcal{C}} = \{\hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_k\} \subseteq \mathbb{S}_q^d$ satisfies $\text{cost}(\hat{\mathcal{C}}; \mathcal{X}) \leq a \cdot \text{cost}(\mathcal{C}^*; \mathcal{X})$ for some $a < \frac{1-802\phi^2}{800\phi^2}$. Then the following holds:*

$$d_W(\hat{\mathcal{C}}, \mathcal{C}^*) \leq \frac{600\sqrt{2}\phi^2\sqrt{k}}{(1-150\phi^2)(\psi-\eta)}. \tag{7}$$

The following theorem is then a simple corollary, with a complete proof in Appendix B.

[1] $d_{\mathcal{M}}(\cdot,\cdot)$ satisfies $d_{\mathcal{M}}(x,y) \geq 0$, $d_{\mathcal{M}}(x,x) = 0$ and $d_{\mathcal{M}}(x,y) \leq d_{\mathcal{M}}(x,z) + d_{\mathcal{M}}(y,z)$ for all x, y, z.

[2] Here ϵ is an approximation constant and is not related to the privacy parameter ε.

Algorithm 2 Threshold-based subspace clustering (TSC), a simplified version

1: **Input**: $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathbb{R}^d$, number of clusters k and number of neighbors s.
2: **Thresholding**: construct $G \in \{0,1\}^{n\times n}$ by connecting $\boldsymbol{x}_i$ to the other s data points in $\mathcal{X}$ with the largest absolute inner products $|\langle \boldsymbol{x}_i, \boldsymbol{x}'\rangle|$. Complete G so that it is undirected.
3: **Clustering**: Let $\mathcal{X}^{(1)}, \cdots, \mathcal{X}^{(\ell)}$ be the connected components in G. Construct $\bar{\mathcal{X}}^{(\ell)}$ by sampling q points from $\mathcal{X}^{(\ell)}$ uniformly at random without replacement.
4: **Output**: subspaces $\hat{\mathcal{C}} = \{\hat{\mathcal{S}}_{(\ell)}\}_{\ell=1}^k$; $\hat{\mathcal{S}}_{(\ell)}$ is the subspace spanned by q arbitrary points in $\bar{\mathcal{X}}^{(\ell)}$.

Theorem 3.4. *Fix a (ϕ, η, ψ)-well separated dataset $\mathcal{X}$ with n data points and $\phi^2 < 1/1602$, $\psi > \eta$. Suppose $\mathcal{X}_S \subseteq \mathcal{X}$ is a subset of $\mathcal{X}$ with size m, sampled uniformly at random without replacement. Let $\hat{\mathcal{C}} = \{\hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_2\}$ be an $(1+\epsilon)$-approximation of optimal k-means subspace clustering computed on $\mathcal{X}_S$. If $m = \Omega(\frac{kqd\log(qd/\gamma'\Delta_k^2(\mathcal{X}))}{\gamma'^2\Delta_k^4(\mathcal{X})})$ with $\gamma' < \frac{1-802\phi^2}{800\phi^2} - 2(1+\epsilon)$, then we have:*

$$\Pr_{\mathcal{X}_S}\left[d_W(\hat{\mathcal{C}}, \mathcal{C}^*) \le \frac{600\sqrt{2}\phi^2\sqrt{k}}{(1-150\phi^2)(\psi-\eta)}\right] \ge \frac{3}{4}, \tag{8}$$

where $\mathcal{C}^* = \{\mathcal{S}_1^*, \cdots, \mathcal{S}_k^*\}$ is the optimal clustering on $\mathcal{X}$; that is, $\mathrm{cost}(\mathcal{C}^*; \mathcal{X}) = \Delta_k^2(\mathcal{X})$.

Consequently, applying Theorem 3.4 together with the sample-aggregate framework we obtain a weak polynomial-time ε-differentially private algorithm for agnostic k-means subspace clustering, with additional amount of per-coordinate Gaussian noise upper bounded by $O(\frac{\phi^2\sqrt{k}}{\varepsilon(\psi-\eta)})$. Our bound is comparable to the one obtained in [22] for private k-means clustering, except for the $(\psi-\eta)$ term which characterizes the well-separatedness under the subspace clustering scenario.

3.3 The stochastic setting

We further consider the case when data points are stochastically generated from some underlying "true" subspace set $\mathcal{C}^* = \{\mathcal{S}_1^*, \cdots, \mathcal{S}_k^*\}$. Such settings were extensively investigated in previous development of subspace clustering algorithms [24, 25, 14]. Below we give precise definition of the considered stochastic subspace clustering model:

The stochastic model For every cluster ℓ associated with subspace $\mathcal{S}_\ell^*$, a data point $\boldsymbol{x}_i^{(\ell)} \in \mathbb{R}^d$ belonging to cluster ℓ can be written as $\boldsymbol{x}_i^{(\ell)} = \boldsymbol{y}_i^{(\ell)} + \boldsymbol{\varepsilon}_i^{(\ell)}$, where $\boldsymbol{y}_i^{(\ell)}$ is sampled uniformly at random from $\{\boldsymbol{y} \in \mathcal{S}_\ell^* : \|\boldsymbol{y}\|_2 = 1\}$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma^2/d \cdot \mathbf{I}_d)$ for some noise parameter σ.

Under the stochastic setting we consider the solver f to be the Threshold-based Subspace Clustering (TSC, [14]) algorithm. A simplified version of TSC is presented in Alg. 2. An alternative idea is to apply results in the previous section since the stochastic model implies well-separated dataset when noise level σ is small. However, the running time of TSC is $O(n^2 d)$, which is much more efficient than core-set based methods. TSC is provably correct in that the similarity graph G has no false connections and is connected per cluster, as shown in the following lemma:

Lemma 3.5 (Connectivity of TSC). *Fix $\gamma > 1$ and assume $\max 0.04n_\ell \le s \le \min n_\ell/6$. If for every $\ell \in \{1, \cdots, k\}$, the number of data points n_ℓ and the noise level σ satisfy*

$$\frac{n_\ell}{\log n_\ell} > \frac{\gamma\pi\sqrt{2q}(12\pi)^{q-1}}{0.01(q/2-1)(q-1)};\quad \frac{\sigma(1+\sigma)}{\sqrt{\log n}}\frac{\sqrt{q}}{\sqrt{d}} \le \frac{1}{15\log n} - \sqrt{1 - \min_{\ell\neq\ell'}\frac{d^2(\mathcal{S}_\ell^*, \mathcal{S}_{\ell'}^*)}{q}};$$

$$\bar{\sigma} < \sqrt{\frac{d}{24\log n}}\left[\cos\left(12\pi\left(\frac{\gamma\sqrt{2\pi q}\log n_\ell}{n_\ell}\right)^{\frac{1}{q-1}}\right) - \cos\left(\left(\frac{0.01(q/2-1)(q-1)}{\sqrt{\pi}}\right)^{\frac{1}{q-1}}\right)\right],$$

where $\bar{\sigma} = 2\sqrt{5}\sigma + \sigma^2$. Then with probability at least $1 - n^2e^{-\sqrt{d}} - n\sum_\ell e^{-n_\ell/400} - \sum_\ell n_\ell^{1-\gamma}/(\gamma\log n_\ell) - 12/n - \sum_\ell n_\ell e^{-c(n_\ell-1)}$, the connected components in G correspond exactly to the k subspaces.

Conditions in Lemma 3.5 characterize the interaction between sample complexity n_ℓ, noise level σ and "signal" level $\min_{\ell\neq\ell'} d(\mathcal{S}_\ell^*, \mathcal{S}_{\ell'}^*)$. Theorem 3.6 is then a simple corollary of Lemma 3.5. Complete proofs are deferred to Appendix C.

Theorem 3.6 (Stability of TSC on stochastic data)**.** *Assume conditions in Lemma 3.5 hold with respect to $n' = n/m$ for $\omega(\log^2 n) \leq m \leq o(n)$. Assume in addition that $\lim_{n\to\infty} n_\ell = \infty$ for all $\ell = 1, \cdots, L$ and the failure probability does not exceed $1/8$. Then for every $\epsilon > 0$ we have*

$$\lim_{n\to\infty} \Pr_{\mathcal{X}_S} \left[d_W(\hat{\mathcal{C}}, \mathcal{C}^*) > \epsilon \right] = 0. \tag{9}$$

Compared to Theorem 3.4 for the agnostic model, Theorem 3.6 shows that one can achieve *consistent* estimation of underlying subspaces under a stochastic model. It is an interesting question to derive finite sample bounds for the differentially private TSC algorithm.

3.4 Discussion

It is worth noting that the sample-aggregate framework is an (ε, δ)-differentially private mechanism for any computational subroutine f. However, the utility claim (i.e., the $O(r/\varepsilon)$ bound on each coordinate of $\mathcal{A}(\mathcal{X}) - \boldsymbol{c}$) requires the stability of the particular subroutine f, as outlined in Eq. (5). It is unfortunately hard to theoretically argue for stability of state-of-the-art subspace clustering methods such as sparse subspace cluster (SSC, [11]) due to the "graph connectivity" issue [21][3]. Nevertheless, we observe satisfactory performance of SSC based algorithms in simulations (see Sec. 5). It remains an open question to derive utility guarantee for (user) differentially private SSC.

4 Private subspace clustering via the exponential mechanism

In Section 3 we analyzed two algorithms with provable privacy and utility guarantees for subspace clustering based on the sample-aggregate framework. However, empirical evidence shows that sample-aggregate based private clustering suffers from poor utility in practice [26]. In this section, we propose a practical private subspace clustering algorithm based on the *exponential mechanism* [18]. In particular, given the dataset $\mathcal{X}$ with n data points, we propose to samples parameters $\boldsymbol{\theta} = (\{\mathcal{S}_\ell\}_{\ell=1}^k, \{z_i\}_{i=1}^n)$ where $\mathcal{S}_\ell \in \mathbb{S}_d^q$, $z_j \in \{1, \cdots, k\}$ from the following distribution:

$$p(\boldsymbol{\theta}; \mathcal{X}) \propto \exp\left(-\frac{\varepsilon}{2} \cdot \sum_{i=1}^n d^2(\boldsymbol{x}_i, \mathcal{S}_{z_i}) \right), \tag{10}$$

where $\varepsilon > 0$ is the privacy parameter. The following proposition shows that exact sampling from the distribution in Eq. (10) results in a provable differentially private algorithm. Its proof is trivial and is deferred to Appendix D.1. Note that unlike sample-aggregate based methods, the exponential mechanism can privately release clustering assignment z. This does not violate the lower bound in [29] because the released clustering assignment z is not guaranteed to be exactly correct.

Proposition 4.1. *The random algorithm $\mathcal{A} : \mathcal{X} \mapsto \boldsymbol{\theta}$ that outputs one sample from the distribution defined in Eq. (10) is ε-differential private.*

4.1 A Gibbs sampling implementation

It is hard in general to sample parameters from distributions as complicated as in Eq. (10). We present a Gibbs sampler that iteratively samples subspaces $\{\mathcal{S}_i\}$ and cluster assignments $\{z_j\}$ from their conditional distributions.

Update of z_i: When $\{\mathcal{S}_\ell\}$ and z_{-i} are fixed, the conditional distribution of z_i is

$$p(z_i | \{\mathcal{S}_\ell\}_{\ell=1}^k, z_{-i}; \mathcal{X}) \propto \exp(-\varepsilon/2 \cdot d^2(\boldsymbol{x}_i, \mathcal{S}_{z_i})). \tag{11}$$

Since $d(\boldsymbol{x}_i, \mathcal{S}_{z_i})$ can be efficiently computed (given an orthonormal basis of $\mathcal{S}_{z_i}$), update of z_i can be easily done by sampling z_j from a categorical distribution.

Update of $\mathcal{S}_\ell$: Let $\widetilde{\mathcal{X}}^{(\ell)} = \{\boldsymbol{x}_i \in \mathcal{X} : z_i = \ell\}$ denote data points that are assigned to cluster ℓ and $\tilde{n}_\ell = |\widetilde{\mathcal{X}}^{(\ell)}|$. Denote $\widetilde{\mathbf{X}}^{(\ell)} \in \mathbb{R}^{d \times \tilde{n}_\ell}$ as the matrix with columns corresponding to all data points in $\widetilde{\mathcal{X}}^{(\ell)}$. The distribution over $\mathcal{S}_\ell$ conditioned on z can then be written as

$$p(\mathcal{S}_\ell = \mathrm{range}(\mathbf{U}_\ell) | z; \mathcal{X}) \propto \exp(\varepsilon/2 \cdot \mathrm{tr}(\mathbf{U}_\ell^\top \mathbf{A}_\ell \mathbf{U}_\ell)); \quad \mathbf{U}_\ell \in \mathbb{R}^{d \times q}, \mathbf{U}_\ell^\top \mathbf{U}_\ell = \mathbf{I}_{q \times q}, \tag{12}$$

where $\mathbf{A}_\ell = \widetilde{\mathbf{X}}^{(\ell)} \widetilde{\mathbf{X}}^{(\ell)\top}$ is the unnormalized sample covariance matrix. Distribution of the form in Eq. (12) is a special case of the *matrix Bingham distribution*, which admits a Gibbs sampler [16]. We give implementation details in Appendix D.2 with modifications so that the resulting Gibbs sampler is empirically more efficient for a wide range of parameter settings.

[3] Recently [28] established full clustering guarantee for SSC, however, under strong assumptions.

4.2 Discussion

The proposed Gibbs sampler resembles the k-plane algorithm for subspace clustering [3]. It is in fact a "probabilistic" version of k-plane since sampling is performed at each iteration rather than deterministic updates. Furthermore, the proposed Gibbs sampler could be viewed as posterior sampling for the following generative model: first sample $\mathbf{U}_\ell$ uniformly at random from $\mathbb{S}_q^d$ for each subspace $\mathcal{S}_\ell$; afterwards, cluster assignments $\{z_i\}_{i=1}^n$ are sampled such that $\Pr[z_i = j] = 1/k$ and $\boldsymbol{x}_i$ is set as $\boldsymbol{x}_i = \mathbf{U}_\ell \boldsymbol{y}_i + \mathcal{P}_{\mathbf{U}_\ell^\perp} \boldsymbol{w}_i$, where $\boldsymbol{y}_i$ is sampled uniformly at random from the q-dimensional unit ball and $\boldsymbol{w}_i \sim \mathcal{N}(0, \mathbf{I}_d/\varepsilon)$. Connection between the above-mentioned generative model and Gibbs sampler is formally justified in Appendix D.3. The generative model is strikingly similar to the well-known mixtures of probabilistic PCA (MPPCA, [27]) model by setting variance parameters σ_ℓ in MPPCA to $\sqrt{1/\varepsilon}$. The only difference is that $\boldsymbol{y}_i$ are sampled uniformly at random from a unit ball [4] and noise $\boldsymbol{w}_i$ is constrained to $\mathbf{U}_\ell^\perp$, the complement space of $\mathbf{U}_\ell$. Note that this is closely related to earlier observation that "posterior sampling is private" [20, 6, 31], but different in that we constructed a model from a private procedure rather than the other way round.

As the privacy parameter $\varepsilon \to \infty$ (i.e., no privacy guarantee), we arrive immediately at the exact k-plane algorithm and the posterior distribution concentrates around the optimal k-means solution $(\mathcal{C}^*, \boldsymbol{z}^*)$. This behavior is similar to what a small-variance asymptotic analysis on MPPCA models reveals [30]. On the other hand, the proposed Gibbs sampler is significantly different from previous Bayesian probabilisitic PCA formulation [34, 30] in that the subspaces are sampled from a matrix Bingham distribution. Finally, we remark that the proposed Gibbs sampler is only asymptotically private because Proposition 4.1 requires exact (or nearly exact [31]) sampling from Eq. (10).

5 Numerical results

We provide numerical results of both the sample-aggregate and Gibbs sampling algorithms on synthetic and real-world datasets. We also compare with a baseline method implemented based on the k-plane algorithm [3] with perturbed sample covariance matrix via the SuLQ framework [2] (details presented in Appendix E). Three solvers are considered for the sample-aggregate framework: threshold-based subspace clustering (TSC, [14]), which has provable utility guarantee with sample-aggregation on stochastic models, along with sparse subspace clustering (SSC, [11]) and low-rank representation (LRR, [17]), the two state-of-the-art methods for subspace clustering. For Gibbs sampling, we use non-private SSC and LRR solutions as initialization for the Gibbs sampler. All methods are implemented using Matlab.

For synthetic datasets, we first generate k random q-dimensional linear subspaces. Each subspace is generated by first sampling a $d \times q$ random Gaussian matrix and then recording its column space. n data points are then assigned to one of the k subspaces (clusters) uniformly at random. To generate a data point $\boldsymbol{x}_i$ assigned with subspace $\mathcal{S}_\ell$, we first sample $\boldsymbol{y}_i \in \mathbb{R}^q$ with $\|\boldsymbol{y}_i\|_2 = 1$ uniformly at random from the q-dimensional unit sphere. Afterwards, $\boldsymbol{x}_i$ is set as $\boldsymbol{x}_i = \mathbf{U}_\ell \boldsymbol{y}_i + \boldsymbol{w}_i$, where $\mathbf{U}_\ell \in \mathbb{R}^{d \times q}$ is an orthonormal basis associated with $\mathcal{S}_\ell$ and $\boldsymbol{w}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$ is a noise vector.

Figure 2 compares the utility (measured in terms of k-means objective $\mathrm{cost}(\hat{\mathcal{C}}; \mathcal{X})$ and the Wasserstein's distance $d_W(\hat{\mathcal{C}}, \mathcal{C}^*)$) of sample aggregation, Gibbs sampling and SuLQ subspace clustering. As shown in the plots, sample-aggregation algorithms have poor utility unless the privacy parameter ε is truly large (which means very little privacy protection). On the other hand, both Gibbs sampling and SuLQ subspace clustering give reasonably good performance. Figure 2 also shows that SuLQ scales poorly with the ambient dimension d. This is because SuLQ subspace clustering requires calibrating noise to a $d \times d$ sample covariance matrix, which induces much error when d is large. Gibbs sampling seems to be robust to various d settings.

We also experiment on real-world datasets. The right two plots in Figure 2 report utility on a subset of the extended Yale Face Dataset B [13] for face clustering. 5 random individuals are picked, forming a subset of the original dataset with $n = 320$ data points (images). The dataset is preprocessed by projecting each individual onto a 9D affine subspace via PCA. Such preprocessing step was adopted in [32, 29] and was theoretically justified in [1]. Afterwards, ambient dimension of the entire dataset is reduced to $d = 50$ by random Gaussian projection. The plots show that Gibbs sampling significantly outperforms the other algorithms.

[4] In MPPCA latent variables $\boldsymbol{y}_i$ are sampled from a normal distribution $\mathcal{N}(\mathbf{0}, \rho^2 \mathbf{I}_q)$.

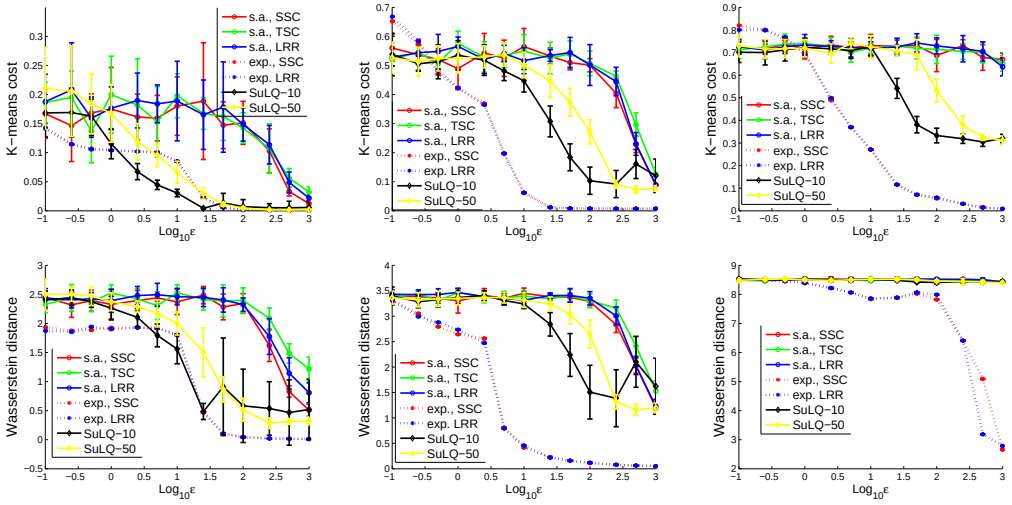

Figure 2: Utility under fixed privacy budget ε. Top row shows k-means cost and bottom row shows the Wasserstein's distance $d_W(\hat{\mathcal{C}}, \mathcal{C}^*)$. From left to right: synthetic dataset, $n = 5000, d = 5, k = 3, q = 3, \sigma = 0.01$; $n = 1000, d = 10, k = 3, q = 3, \sigma = 0.1$; extended Yale Face Dataset B (a subset). $n = 320, d = 50, k = 5, q = 9, \sigma = 0.01$. δ is set to $1/(n \ln n)$ for (ε, δ)-privacy algorithms. "s.a." stands for smooth sensitivity and "exp." stands for exponential mechanism. "SuLQ-10" and "SuLQ-50" stand for the SuLQ framework performing 10 and 50 iterations. Gibbs sampling is run for 10000 iterations and the mean of the last 100 samples is reported.

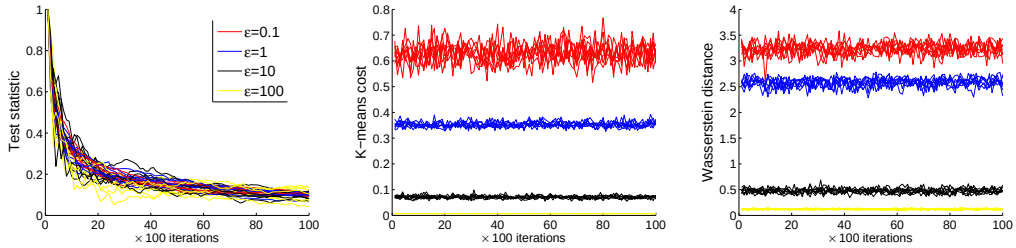

Figure 3: Test statistics, k-means cost and $d_W(\hat{\mathcal{C}}, \mathcal{C}^*)$ of 8 trials of the Gibbs sampler under different privacy settings. Synthetic dataset setting: $n = 1000$, $d = 10$, $k = 3$, $q = 3$, $\sigma = 0.1$.

In Figure 3 we investigate the mixing behavior of proposed Gibbs sampler. We plot for multiple trials of Gibbs sampling the k-means objective, Wasserstein's distance and a test statistic $1/\sqrt{kq} \cdot (\sum_{\ell=1}^{k} \|1/T \cdot \sum_{t=1}^{T} \mathbf{U}_\ell^{(t)}\|_F^2)^{1/2}$, where $\mathbf{U}_\ell^{(t)}$ is a basis sample of $\mathcal{S}_\ell$ at the tth iteration. The test statistic has mean zero under distribution in Eq. (10) and a similar statistic was used in [4] as a diagnostic of the mixing behavior of another Gibbs sampler. Figure 3 shows that under various privacy parameter settings, the proposed Gibbs sampler mixes quite well after 10000 iterations.

6 Conclusion

In this paper we consider subspace clustering subject to formal differential privacy constraints. We analyzed two sample-aggregate based algorithms with provable utility guarantees under agnostic and stochastic data models. We also propose a Gibbs sampling subspace clustering algorithm based on the exponential mechanism that works well in practice. Some interesting future directions include utility bounds for state-of-the-art subspace clustering algorithms like SSC or LRR.

Acknowledgment This research is supported in part by grant NSF CAREER IIS-1252412, NSF Award BCS-0941518, and a grant by Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative administered by the IDM Programme Office.

References

[1] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SULQ framework. In *PODS*, 2015.

[3] P. S. Bradley and O. L. Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16(1), 2000.

[4] K. Chaudhuri, A. Sarwate, and K. Sinha. Near-optimal algorithms for differentially private principal components. In *NIPS*, 2012.

[5] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[6] C. Dimitrakakis, B. Nelson, A. Mitrokotsa, and B. I. Rubinstein. Robust and private bayesian inference. In *Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

[7] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[9] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[10] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *STOC*, 2014.

[11] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[12] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *SODA*, 2013.

[13] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[14] R. Heckel and H. Bölcskei. Robust subspace clustering via thresholding. *arXiv:1307.4891*, 2013.

[15] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *CVPR*, 2003.

[16] P. Hoff. Simulation of the matrix bingham-conmises-fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.

[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Ma, and Y. Yu. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2012.

[18] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

[19] B. McWilliams and G. Montana. Subspace clustering of high-dimensional data: a predictive approach. *Data Mining and Knowledge Discovery*, 28(3):736–772, 2014.

[20] D. J. Mir. *Differential privacy: an exploration of the privacy-utility landscape*. PhD thesis, Rutgers University, 2013.

[21] B. Nasihatkon and R. Hartley. Graph connectivity in sparse subspace clustering. In *CVPR*, 2011.

[22] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

[23] R. Ostrovksy, Y. Rabani, L. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *FOCS*, 2006.

[24] M. Soltanolkotabi, E. J. Candes, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[25] M. Soltanolkotabi, E. Elhamifa, and E. Candes. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

[26] D. Su, J. Cao, N. Li, E. Bertino, and H. Jin. Differentially private k-means clustering. *arXiv*, 2015.

[27] M. Tipping and C. Bishop. Mixtures of probabilistic principle component anlyzers. *Neural computation*, 11(2):443–482, 1999.

[28] Y. Wang, Y.-X. Wang, and A. Singh. Clustering consistent sparse subspace clustering. *arXiv*, 2015.

[29] Y. Wang, Y.-X. Wang, and A. Singh. A deterministic analysis of noisy sparse subspace clustering for dimensionality-reduced data. In *ICML*, 2015.

[30] Y. Wang and J. Zhu. DP-space: Bayesian nonparametric subspace clustering with small-variance asymptotic analysis. In *ICML*, 2015.

[31] Y.-X. Wang, S. Fienberg, and A. Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *ICML*, 2015.

[32] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In *ICML*, pages 89–97, 2013.

[33] A. Zhang, N. Fawaz, S. Ioannidis, and A. Montanari. Guess who rated this movie: Identifying users through subspace clustering. *arXiv*, 2012.

[34] Z. Zhang, K. L. Chan, J. Kwok, and D.-Y. Yeung. Bayesian inference on principal component analysis using reversible jump markov chain monte carlo. In *AAAI*, 2004.

Appendix A Some basic properties regarding the distances

Proposition A.1. *Let $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3 \in \mathbb{S}_q^d$ be three q-dimensional subspaces. Then $d(\mathcal{S}_1, \mathcal{S}_3) \leq d(\mathcal{S}_1, \mathcal{S}_2) + d(\mathcal{S}_2, S_3)$ and $d^2(\mathcal{S}_1, \mathcal{S}_3) \leq 2(d^2(\mathcal{S}_1, \mathcal{S}_2) + d^2(\mathcal{S}_2, \mathcal{S}_3))$.*

Proof. Let $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3 \in \mathbb{R}^{d\times q}$ be orthonormal basis associated with $\mathcal{S}_1, \mathcal{S}_2$ and $\mathcal{S}_3$. We then have $d(\mathcal{S}_1, \mathcal{S}_3) = \|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_3\mathbf{U}_3^\top\|_F \leq \|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F + \|\mathbf{U}_2\mathbf{U}_2^\top - \mathbf{U}_3\mathbf{U}_3^\top\|_F = d(\mathcal{S}_1, \mathcal{S}_2) + d(\mathcal{S}_2, \mathcal{S}_3)$. The other inequality holds due to the fact that $\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_3\mathbf{U}_3^\top\|_F^2 \leq 2(\|\mathbf{U}_1\mathbf{U}_1^\top - \mathbf{U}_2\mathbf{U}_2^\top\|_F^2 + \|\mathbf{U}_2\mathbf{U}_2^\top - \mathbf{U}_3\mathbf{U}_3^\top\|_F^2)$. □

Proposition A.2. *For any $\boldsymbol{x} \in \mathbb{R}^d$ and $\mathcal{S}, \mathcal{S}' \in \mathbb{S}_q^d$, we have $d(\boldsymbol{x}, \mathcal{S}') \leq d(\boldsymbol{x}, \mathcal{S}) + d(\mathcal{S}, \mathcal{S}')$ and $d^2(\boldsymbol{x}, \mathcal{S}') \leq 2(d^2(\boldsymbol{x}, \mathcal{S}) + d^2(\mathcal{S}, \mathcal{S}'))$.*

Proof. By definition, $d(\boldsymbol{x}, \mathcal{S}') = \|\boldsymbol{x} - \mathcal{P}_{\mathcal{S}'}(\boldsymbol{x})\|_2 \leq \|\boldsymbol{x} - \mathcal{P}_{\mathcal{S}'}(\mathcal{P}_{\mathcal{S}}(\boldsymbol{x}))\|_2 \leq \|\boldsymbol{x} - \mathcal{P}_{\mathcal{S}}(\boldsymbol{x})\|_2 + \|\mathcal{P}_{\mathcal{S}}(\boldsymbol{x}) - \mathcal{P}_{\mathcal{S}'}(\mathcal{P}_{\mathcal{S}}(\boldsymbol{x}))\|_2 \leq d(\boldsymbol{x}, \mathcal{S}) + \sup_{\boldsymbol{y}\in\mathcal{S}, \|\boldsymbol{y}\|_2\leq 1} \|\boldsymbol{y} - \mathcal{P}_{\mathcal{S}'}(\boldsymbol{y})\|_2$. Note also that $\sup_{\boldsymbol{y}\in\mathcal{S}, \|\boldsymbol{y}\|_2\leq 1} \|\boldsymbol{y} - \mathcal{P}_{\mathcal{S}'}(\boldsymbol{y})\|_2 \leq \sup_{\boldsymbol{y}\in\mathcal{S}, \|\boldsymbol{y}\|_2\leq 1} \|\mathbf{U}\mathbf{U}^\top\boldsymbol{y} - \mathbf{U}'\mathbf{U}'^\top\boldsymbol{y}\|_2 \leq \|\mathbf{U}\mathbf{U}^\top - \mathbf{U}'\mathbf{U}'^\top\|_2 \leq d(\mathcal{S}, \mathcal{S}')$. Here $\mathcal{U}$ and $\mathcal{U}'$ are orthonormal basis associated with $\mathcal{S}$ and $\mathcal{S}'$. Therefore, $d(\boldsymbol{x}, \mathcal{S}') \leq d(\boldsymbol{x}, \mathcal{S}) + d(\mathcal{S}, \mathcal{S}')$. The other inequality follows by the same argument. □

Proposition A.3. *Fix $\mathcal{S} \in \mathbb{S}_q^d$ and let $\mathbf{U} \in \mathbb{R}^{d\times q}$ be an orthonormal basis associated with $\mathcal{S}$. Suppose $\mathbf{U}' = \mathbf{U} + \mathbf{E}$ and $\mathcal{S}' = \mathrm{range}(\mathbf{U}')$. Then $d(\mathcal{S}, \mathcal{S}') \leq \sqrt{2}\|\mathbf{E}\|_F$.*

Proof. Apply Wedin's Theorem (Theorem F.2 in Appendix F) and note that $\sigma_q(\mathbf{U}) = 1$. □

Appendix B Proofs of sample-aggregate private subspace clustering: the agnostic case

The main objective of this section is to prove Theorem 3.4 for differentially private subspace clustering under the fully agnostic setting. The theorem is a simple consequence of Lemma 3.3 in the main text and the following lemma:

Lemma B.1. *Fix $\gamma > 0$. Suppose $\mathcal{X}_S$ contains $m = \Omega(\frac{kqd\log(qd/\gamma)}{\gamma^2})$ data points subsampled from $\mathcal{X}$ uniformly at random without replacement. Then with probability at least 3/4 over random samples U, the following holds* uniformly *for all candidate subspace sets $\mathcal{C}$:*

$$\mathrm{cost}(\mathcal{C}; \mathcal{X}_S) \leq 2\mathrm{cost}(\mathcal{C}; \mathcal{X}) + \gamma. \tag{13}$$

B.1 Proof of Lemma B.1

Lemma B.2 ([39])**.** *Fix $\mathcal{X}$ and $f : \mathcal{X} \to [0, M]$ for some positive constant $M > 0$. Let $\mathcal{X}_S$ be a subset of $\mathcal{X}$ with t elements, each drawn uniformly at random from $\mathcal{X}$ without replacement. Let $\epsilon, \delta > 0$. Then $\Pr[|\mathbb{E}_{\mathcal{X}}[f(x)] - \mathbb{E}_{\mathcal{X}_S}[f(x)]| \geq \epsilon] \leq \delta$ when $t \geq \frac{M^2\ln(2/\delta)}{2\epsilon^2}$.*

Corollary B.3. *Fix $\mathcal{X}$ and a finite set of functions $\mathcal{F}$, where $0 \leq f(x) \leq M$ for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Let $\mathcal{X}_S$ be a subset of $\mathcal{X}$ with m elements, each drawn uniformly at random from $\mathcal{X}$ without replacement. Let $\epsilon, \delta > 0$. Then $\Pr[\exists f \in \mathcal{F}, |\mathbb{E}_{\mathcal{X}}[f(x)] - \mathbb{E}_{\mathcal{X}_S}[f(x)]| \geq \epsilon] \leq \delta$ when $m \geq \frac{M^2\ln(2|\mathcal{F}|/\delta)}{2\epsilon^2}$.*

Proof. Apply Lemma B.2 and use union bound over all $f \in \mathcal{F}$. □

Lemma B.4. *Fix $\epsilon > 0$. There exists $\mathbb{S} \subseteq \mathbb{S}_q^d$ with $|\mathbb{S}| = O((qd)^{qd/2}/\epsilon^{qd})$ such that for any $\mathcal{S} \in \mathbb{S}_q^d$, $\min_{\mathcal{S}'\in\mathbb{S}} d(\mathcal{S}, \mathcal{S}') \leq \epsilon$.*

Proof. By a standard convering number argument, there exists $\mathbb{L} \subseteq \mathbb{R}^d$ with $|\mathbb{L}| = O((\sqrt{d}/\epsilon)^d)$ such that for any $\boldsymbol{x} \in \mathbb{R}^d$, $\|\boldsymbol{x}\|_2 \leq 1$, we have $\min_{\boldsymbol{x}'\in\mathbb{L}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \leq \epsilon$. Consequently, there exists $\mathbb{L}_q \subseteq \mathbb{R}^{d\times q}$ with $|\mathbb{L}_q| = O((qd)^{qd/2}/\epsilon^{qd})$ such that for any $\mathbf{U} \in \mathbb{R}^{d\times q}$ with unit column norms, $\min_{\mathbf{U}'\in\mathbb{L}_q} \|\mathbf{U} - \mathbf{U}'\|_F \leq \epsilon$. Proposition A.3 then yields the lemma. □

We are now ready to prove Lemma B.1.

Proof of Lemma B.1. Suppose $\mathbb{S}$ is a finite subset of $\mathbb{S}_q^d$ such that for every $\mathcal{S} \in \mathbb{S}_q^d$, $\min_{\mathcal{S}' \in \mathbb{S}} d^2(\mathcal{S}, \mathcal{S}') \leq \gamma/4$. By Lemma B.4, there exists such $\mathbb{S}$ with $|\mathbb{S}| = O((qd/\gamma)^{qd/2})$. Let $\mathcal{X}$ be the set of data points and $\mathcal{F} = \{f(\cdot; \mathcal{C}) | \mathcal{C} = \{\mathcal{S}_1, \cdots, \mathcal{S}_k\} \subseteq \mathbb{S}\}$, where $f(\boldsymbol{x}; \mathcal{C}) = \min_{j=1}^k d^2(\boldsymbol{x}, \mathcal{S}_j)$. By definition, $\mathbb{E}_{\mathcal{X}}[f(\boldsymbol{x}; \mathcal{C})] = \mathrm{cost}(\mathcal{C}; \mathcal{X})$ and $|\mathcal{F}| = O((qd/\gamma)^{kqd/2})$. Subsequently, applying Corollary B.3 we obtain

$$\Pr_{\mathcal{X}_S}\left[\forall \mathcal{C} \subseteq \mathbb{S}, \left|\mathrm{cost}(\mathcal{C}; \mathcal{X}_S) - \mathrm{cost}(\mathcal{C}; \mathcal{X})\right| \leq \frac{\gamma}{2}\right] \geq \frac{3}{4}$$

whenever $|\mathcal{X}_S| = \Omega(\frac{kqd \log(qd/\gamma)}{\gamma^2})$. Consequently, applying Proposition A.2 we have

$$\Pr_{\mathcal{X}_S}\left[\forall \mathcal{C} \subseteq \mathbb{S}_q^d, \mathrm{cost}(\mathcal{C}; \mathcal{X}_S) \leq 2\mathrm{cost}(\mathcal{C}; \mathcal{X}) + \gamma\right] \geq \frac{3}{4}.$$

□

B.2 Proof of Lemma 3.3

We first define some notations that will be used in the proof. Throughout the section we assume the dataset $\mathcal{X}$ is (ϕ, η, ψ)-well separated. Let $\mathcal{X}_i = \{\boldsymbol{x} \in \mathcal{X} : d(\boldsymbol{x}, \mathcal{S}_i^*) \leq d(\boldsymbol{x}, \mathcal{S}_j^*), \forall j\}$ denote the collection of all data points in $\mathcal{X}$ that are clustered to the cluster corresponding to $\mathcal{S}_i^*$. Define $n_i = |\mathcal{X}_i|$. By definition, $\sum_{i=1}^k n_i = n$. Define $r_i^2 = \Delta_1^2(\mathcal{X}_i)$, $D_i = \min_{j \neq i} d(\mathcal{S}_i^*, \mathcal{S}_j^*)$ and $d_i^2 = \phi^2 n \Delta_{k-1}^2(\mathcal{X})/n_i$. Let $\mathcal{X}_i^{\mathrm{cor}} = \{\boldsymbol{x} \in \mathcal{X}_i : d(\boldsymbol{x}, \mathcal{S}_i^*)^2 \leq \frac{r_i^2}{\rho}\}$ for some parameter $\rho \in (0, 1)$.

Proposition B.5. $r_i^2 \leq d_i^2 \leq \frac{2\phi^2}{1-2\phi^2} D_i^2$.

Proof. Since $\mathcal{X}$ is well-separated we have $d_i^2 = \phi^2 \Delta_{k-1}^2(\mathcal{X}) \cdot n/n_i \geq \Delta_k^2(\mathcal{X}) \cdot n/n_i \geq \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{X}_i} d(\boldsymbol{x}, \mathcal{S}_i^*)^2 \geq \Delta_1^2(\mathcal{X}_i) = r_i^2$. Hence the first inequality.

For the second inequality, we only need to prove that $(1 - 2\phi^2) n \Delta_{k-1}^2(\mathcal{X}) \leq 2 n_i D_i^2$. By well-separatedness $(1 - 2\phi^2) n \Delta_{k-1}^2(\mathcal{X}) = n(\Delta_{k-1}^2(\mathcal{X}) - 2\Delta_k^2(\mathcal{X}))$. On the other hand, by diverting all points in $\mathcal{X}_i^*$ into the cluster associated with $\mathcal{S}_j^*$ with $d(\mathcal{S}_i^*, \mathcal{S}_j^*) = D_i$, we have

$$\begin{aligned} n\Delta_{k-1}^2(\mathcal{X}) &\leq \sum_{\ell \neq i} \sum_{\boldsymbol{x} \in \mathcal{X}_\ell} d^2(\boldsymbol{x}, \mathcal{S}_\ell^*) + \sum_{\boldsymbol{x} \in \mathcal{X}_i} d^2(\boldsymbol{x}, \mathcal{S}_j^*) \\ &\leq \sum_{\ell \neq i} \sum_{\boldsymbol{x} \in \mathcal{X}_\ell} d^2(\boldsymbol{x}, \mathcal{S}_\ell^*) + 2 \sum_{\boldsymbol{x} \in \mathcal{X}_i} d^2(\boldsymbol{x}, \mathcal{S}_j^*) + 2 n_i d^2(\mathcal{S}_i^*, \mathcal{S}_j^*) \\ &\leq 2n\Delta_k^2(\mathcal{X}) + 2 n_i D_i^2. \end{aligned}$$

Rearranging the terms we get $n(\Delta_{k-1}^2(\mathcal{X}) - 2\Delta_k^2(\mathcal{X})) \leq 2 n_i D_i^2$. □

Proposition B.6. *For any $\rho \in (0, 1)$, $|\mathcal{X}_i^{\mathrm{cor}}| \geq (1 - \rho)|\mathcal{X}_i| = (1 - \rho) n_i$.*

Proof. By definition $r_i^2 = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \mathcal{X}_i} d^2(\boldsymbol{x}, \mathcal{S}_i^*) = \mathbb{E}[T]$, where T is the random variable of $d^2(\boldsymbol{x}, \mathcal{S}_i^*)$ for a vector chosen from $\mathcal{X}_i$ uniformly at random. By Markov's inequality, $\Pr[T > \frac{r_i^2}{\rho}] \leq \rho$ and hence $|\mathcal{X}_i^{\mathrm{cor}}| = n_i \Pr[T \leq \frac{r_i^2}{\rho}] \geq (1 - \rho) n_i$. □

Lemma B.7. *Suppose $\mathrm{cost}(\hat{\mathcal{C}}; \mathcal{X}) \leq \alpha \Delta_k^2(\mathcal{X})$ for some $\alpha < \frac{1 - 802\phi^2}{800}$. Then there exists a permutation $\pi : [k] \to [k]$ such that $d(\hat{\mathcal{S}}_i, \mathcal{S}_{\pi(i)}^*) \leq D_i/10$ for every $i = 1, \cdots, k$.*

Proof. Pick $\rho = \frac{800\phi^2}{1 - 2\phi^2}$. By conditions on α we have $\alpha \leq (\frac{1}{\rho} - 1)\phi^2$. In the remainder of the proof we show that for every $i \in [k]$, there exists some $j \in [k]$ such that $d(\mathcal{S}_i^*, \hat{\mathcal{S}}_j) \leq \frac{2 d_i}{\sqrt{\rho}} \leq D_i/10$, where the last inequality is due to Proposition B.5 and the choice of ρ. This is sufficient for the conclusion

in Lemma B.7 since no two subspaces $\mathcal{S}_i^*$ and $\mathcal{S}_{i'}^*$ can be within the range of $D_i/10$ to the same subspace $\hat{\mathcal{S}}_j$ due to the definition of D_i and triangle inequality presented in Proposition A.1.

Assume by way of contradiction that there exists $i \in [k]$ such that $d(\mathcal{S}_i^*, \hat{\mathcal{S}}_j) > \frac{2d_i}{\sqrt{\rho}}$. This implies that any point in $\mathcal{X}_i^{\rm cor} = \{\boldsymbol{x} \in \mathcal{X}_i : d(\boldsymbol{x}, \mathcal{S}_i^*) \le \frac{r_i}{\sqrt{\rho}}\}$ is at least $\frac{d_i}{\sqrt{\rho}}$ away from any subspace in $\hat{\mathcal{C}}$, due to $d_i \ge r_i$ and the triangle inequality. Therefore, $\mathrm{cost}(\hat{\mathcal{C}}; \mathcal{X}) \ge \frac{|\mathcal{X}_i^{\rm cor}|}{n}\frac{d_i^2}{\rho} \ge (\frac{1}{\rho} - 1)\frac{n_i}{n} d_i^2$, where the last inequality is due to Proposition B.6. Finally, $(\frac{1}{\rho} - 1)\frac{n_i}{n} d_i^2 = (\frac{1}{\rho} - 1)\frac{n_i}{n} \cdot \phi^2 n \Delta_{k-1}^2(\mathcal{X}) > \alpha n_i \Delta_{k-1}^2(\mathcal{X}) \ge \alpha \Delta_{k-1}^2(\mathcal{X})$, and hence the contradiction. □

Lemma B.8. *Fix a candidate subspace set $\hat{\mathcal{C}} = \{\hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_2\}$. Define $\hat{\mathcal{R}}_i = \{\boldsymbol{x} \in \mathcal{X} : d(\boldsymbol{x}, \hat{\mathcal{S}}_i) \le d(\boldsymbol{x}, \hat{\mathcal{S}}_j) + \hat{D}_i/4, \forall j\}$, where $\hat{D}_i = \min_{j \ne i} d(\hat{\mathcal{S}}_i, \hat{\mathcal{S}}_j)$. Suppose there exists a permutation $\pi : [k] \to [k]$ such that $d(\hat{\mathcal{S}}_i, \mathcal{S}_{\pi(i)}^*) \le D_{\pi(i)}/10$ for every i, where $D_i = \min_{j \ne i} d(\mathcal{S}_i^*, \mathcal{S}_j^*)$. Then we have $\mathcal{X}_{\pi(i)} \subseteq \hat{\mathcal{R}}_i$ and further more $|\mathcal{X}_{\pi(i)}| \ge \beta |\hat{\mathcal{R}}_i|$ for $\beta = \frac{1-2\phi^2}{1+48\phi^2}$.*

Proof. Without loss of generality we assume $\pi(i) = i$; that is, $d(\hat{\mathcal{S}}_i, \mathcal{S}_i^*) \le D_i/10$ for every $i = 1, \cdots, k$. By triangle inequality in Proposition A.1, we have $\frac{4}{5} D_i \le \hat{D}_i \le \frac{6}{5} D_i$. Fix an arbitrary $\boldsymbol{x} \in \mathcal{X}_i$. By definition, $d(\boldsymbol{x}, \mathcal{S}_i^*) \le d(\boldsymbol{x}, \mathcal{S}_j^*)$. Therefore, $d(\boldsymbol{x}, \hat{\mathcal{S}}_i) \le d(\boldsymbol{x}, \mathcal{S}_i^*) + \frac{D_i}{10} \le d(\boldsymbol{x}, \mathcal{S}_j^*) + \frac{D_i}{10} \le d(\boldsymbol{x}, \hat{\mathcal{S}}_j) + \frac{D_i}{5} \le d(\boldsymbol{x}, \hat{\mathcal{S}}_j) + \frac{\hat{D}_i}{4}$. Therefore, $\mathcal{X}_i \subseteq \hat{\mathcal{R}}_i$.

We next prove that $|\mathcal{X}_i| \ge \beta |\hat{\mathcal{R}}_i|$. The approach we take is to assume $|\mathcal{X}_i| = \beta |\hat{\mathcal{R}}_i|$ for some real number β and show that $\beta \ge \frac{1-2\phi^2}{1+48\phi^2}$. Let $a_j = \frac{|\hat{\mathcal{R}}_i \cap \mathcal{X}_j|}{|\hat{\mathcal{R}}_i|}$ and we arbitrarily assign $\frac{a_j n_i}{1-a_i}$ points in $\mathcal{X}_i$ to the cluster associated with subpace $\mathcal{S}_j^*$. This will clear the $\mathcal{S}_i^*$ subspace since $\sum_{j \ne i} \frac{a_j n_i}{1-a_i} = n_i$. As a result, we have

$$\begin{aligned}
n\Delta_{k-1}^2(\mathcal{X}) &\le n\Delta_k^2(\mathcal{X}) - n_i \Delta_1(\mathcal{X}_i) + \sum_{\boldsymbol{x} \in \mathcal{X}_i} d^2(\boldsymbol{x}, \mathcal{S}_j^*) \\
&\le n\Delta_k^2(\mathcal{X}) + n_i \Delta_1(\mathcal{X}_i) + 2\sum_{j \ne i} \frac{a_j n_i}{1-a_i} \cdot d^2(\mathcal{S}_i^*, \mathcal{S}_j^*) \\
&\le 2n\Delta_k^2(\mathcal{X}) + \frac{2\beta}{1-\beta} \sum_{j \ne i} a_j |\hat{\mathcal{R}}_i| d^2(\mathcal{S}_i^*, \mathcal{S}_j^*).
\end{aligned}$$

The last inequality is due to the fact that $\frac{n_i}{a_i} = \frac{|\mathcal{X}_i| \cdot |\hat{\mathcal{R}}_i|}{|\mathcal{X}_i \cap \hat{\mathcal{R}}_i|} = |\hat{\mathcal{R}}_i|$ and $\frac{a_i}{1-a_i} = \frac{|\mathcal{X}_i|}{|\hat{\mathcal{R}}_i| - |\mathcal{X}_i|} \le \frac{\beta}{1-\beta}$.

On the other hand, for any $\boldsymbol{y} \in \mathcal{X}_j \cap \hat{\mathcal{R}}_i$, one has $d(\boldsymbol{y}, \mathcal{S}_i^*) \le d(\boldsymbol{y}, \hat{\mathcal{S}}_i) + \frac{D_i}{10} \le d(\boldsymbol{y}, \hat{\mathcal{S}}_j) + \frac{D_i}{5} + \frac{D_i}{10} \le d(\boldsymbol{y}, \mathcal{S}_j^*) + \frac{3}{10}(D_i + D_j)$. Consequently, $d(\mathcal{S}_i^*, \mathcal{S}_j^*) \le d(\boldsymbol{y}, \mathcal{S}_i^*) + d(\boldsymbol{y}, \mathcal{S}_j^*) \le 2d(\boldsymbol{y}, \mathcal{S}_j^*) + \frac{3}{10}(D_i + D_j) \le 2d(\boldsymbol{y}, \mathcal{S}_j^*) + \frac{3}{5} d(\mathcal{S}_i^*, \mathcal{S}_j^*)$ and hence $d(\mathcal{S}_i^*, \mathcal{S}_j^*) \le 5d(\boldsymbol{y}, \mathcal{S}_j^*)$. Subsequently,

$$\begin{aligned}
n\Delta_{k-1}^2(\mathcal{X}) &\le 2n\Delta_k^2(\mathcal{X}) + \frac{2\beta}{1-\beta} \sum_{j \ne i} a_j |\hat{\mathcal{R}}_i| d^2(\mathcal{S}_i^*, \mathcal{S}_j^*) \\
&= 2n\Delta_k^2(\mathcal{X}) + \frac{2\beta}{1-\beta} \sum_{j \ne i} |\hat{\mathcal{R}}_i \cap \mathcal{X}_j| d^2(\mathcal{S}_i^*, \mathcal{S}_j^*) \\
&\le 2n\Delta_k^2(\mathcal{X}) + \frac{50\beta}{1-\beta} \sum_{j \ne i} \sum_{\boldsymbol{y} \in \hat{\mathcal{R}}_i \cap \mathcal{X}_j} d^2(\boldsymbol{y}, \mathcal{S}_j^*) \\
&\le 2n\Delta_k^2(\mathcal{X}) + \frac{50\beta}{1-\beta} \cdot n\Delta_k^2(\mathcal{X}).
\end{aligned}$$

By the well-separatedness of $\mathcal{X}$, we have

$$\Delta_k^2(\mathcal{X}) \le \phi^2 \Delta_{k-1}^2(\mathcal{X}) \le \left(2 + \frac{50\beta}{1-\beta}\right) \phi^2 \Delta_k^2(\mathcal{X}).$$

Therefore, $2 + \frac{50\beta}{1-\beta} \ge 1/\phi^2$, which implies $\beta \ge \frac{1-2\phi^2}{1+48\phi^2}$. □

Lemma B.9. *Assume $\phi \le 1/2$. Fix a candidate subspace set $\hat{\mathcal{C}} = \{\hat{\mathcal{S}}_1, \cdots, \hat{\mathcal{S}}_n\}$. Let $\hat{\mathcal{X}}_i = \{\boldsymbol{x} \in \mathcal{X} : d(\boldsymbol{x}, \hat{\mathcal{S}}_i) \le d(\boldsymbol{x}, \hat{\mathcal{S}}_j), \forall j\}$ denote the set of all data points that are clustered into $\hat{\mathcal{S}}_i$. Suppose $d(\hat{\mathcal{S}}_i, \mathcal{S}_i^*) \le D_i/10$ for every i. Then $|\hat{\mathcal{X}}_i \triangle \mathcal{X}_i| \le 150\phi^2 |\mathcal{X}_i|$, where $\triangle$ denotes the symmetric difference operator between two sets.*

Proof. We first derive a lower bound on $|\hat{\mathcal{X}}_i \cap \mathcal{X}_i|$. We first claim that for any $\boldsymbol{x} \in \mathcal{X}$, $d(\boldsymbol{x}, \mathcal{S}_i^*) \le \frac{2}{5} D_i$ yields $\boldsymbol{x} \in \hat{\mathcal{X}}_i$. To see this, note that $d(\boldsymbol{x}, \hat{\mathcal{S}}_i) \le d(\boldsymbol{x}, \mathcal{S}_i^*) + d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i) \le \frac{2}{5} D_i + \frac{1}{10} D_i = \frac{1}{2} D_i$ and for every $j \ne i$, $d(\boldsymbol{x}, \hat{\mathcal{S}}_j) \ge d(\mathcal{S}_i^*, \hat{\mathcal{S}}_j) - d(\boldsymbol{x}, \mathcal{S}_i^*) \ge \frac{9}{10} D_i - \frac{2}{5} D_i = \frac{1}{2} D_i$. Therefore, $d(\boldsymbol{x}, \hat{\mathcal{S}}_i) \le d(\boldsymbol{x}, \hat{\mathcal{S}}_j)$ for every $j \ne i$.

On the other hand, by Proposition B.5 $d(\boldsymbol{x}, \mathcal{S}_i^*) \le \frac{r_i}{\sqrt{\rho'}}$ with $\rho' = \frac{25\phi^2}{2(1-2\phi^2)}$ implies $d(\boldsymbol{x}, \mathcal{S}_i^*) \le \frac{2}{5} D_i$. Consequently, by Proposition B.6 we have $|\hat{\mathcal{X}}_i \cap \mathcal{X}_i| \ge |\{\boldsymbol{x} \in \mathcal{X}_i : d(\boldsymbol{x}, \mathcal{S}_i^*) \le \frac{r_i}{\sqrt{\rho'}}| \ge (1-\rho')|\mathcal{X}_i|$. In addition, Lemma B.8 asserts that $|\mathcal{X}_i| \ge \beta |\hat{\mathcal{R}}_i| \ge \beta |\mathcal{X}_i|$. Therefore, $|\hat{\mathcal{X}}_i \triangle \mathcal{X}_i| \le (2\rho' + \frac{1}{\beta} - 1)|\mathcal{X}_i| \le \frac{75\phi^2}{1-2\phi^2} |\mathcal{X}_i| \le 150\phi^2 \mathcal{X}_i$, assuming $\phi \le \frac{1}{2}$. □

Proposition B.10. *Let $\sigma_q(\mathbf{X}_i)$ denote the qth largest singular value of $\mathbf{X}_i$. We then have $\sigma_q^2(\mathbf{X}_i) \ge n\psi$ and $\sigma_{q+1}^2(\mathbf{X}_i) \le n\eta$.*

Proof. By principal component analysis, $n\Delta_k^2(\mathcal{X}) = \sum_{i=1}^k \sum_{p \ge q+1} \sigma_p^2(\mathcal{X}_i)$. Therefore, $n\Delta_{k,-1}^2(\mathcal{X}) \le n\Delta_k^2(\mathcal{X}) + \sigma_q^2(\mathcal{X}_i)$ for any i. Since $\mathcal{X}$ is (ϕ, η, ψ)-well separated, we have $n\Delta_{k,-1}^2 \ge n\Delta_k^2 + n\psi$. Consequently, $\sigma_q^2(\mathcal{X}_i) \ge n\psi$. On the other hand, we have $n\Delta_{k,+1}^2(\mathcal{X}) \le n\Delta_k^2(\mathcal{X}) - \sigma_{q+1}^2(\mathcal{X}_i)$ for any i and $n\Delta_{k,+1}^2(\mathcal{X}) \ge n\Delta_k^2 - n\eta$. Hence $\sigma_{q+1}^2(\mathcal{X}_i) \le n\eta$. □

Lemma B.11. *Following the same notations in Lemma B.11. Suppose $\phi^2 < 1/150$. If $|\hat{\mathcal{X}}_i \triangle \mathcal{X}_i| \le 150\phi^2 |\mathcal{X}_i|$ holds for every i, then $d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i) \le \frac{600\sqrt{2}\phi^2}{(1-150\phi^2)(\psi-\eta)}$.*

Proof. Let $\mathcal{B}_i = \mathcal{X}_i \cap \hat{\mathcal{X}}_i$, $\mathcal{Y}_i = \mathcal{X}_i \backslash \mathcal{B}_i$ and $\mathcal{Z}_i = \hat{\mathcal{X}}_i \backslash \mathcal{B}_i$. Since $|\hat{\mathcal{X}}_i \triangle \mathcal{X}_i| \le 150\phi^2 |\mathcal{X}_i|$, we have $|\mathcal{B}_i| \ge (1 - 150\phi^2)|\mathcal{X}_i|$ and $|\mathcal{Y}_i|, |\mathcal{Z}_i| \le 150\phi^2 |\mathcal{X}_i|$. Let $\mathbf{B}_i, \mathbf{Y}_i, \mathbf{Z}_i$ be the matrices associated with $\mathcal{B}_i, \mathcal{Y}_i$ and $\mathcal{Z}_i$. Define $\mathbf{A}_i = \frac{\mathbf{B}_i \mathbf{B}_i^\top + \mathbf{Y}_i \mathbf{Y}_i^\top}{|\mathcal{B}_i| + |\mathcal{Y}_i|}$ and $\widetilde{\mathbf{A}}_i = \frac{\mathbf{B}_i \mathbf{B}_i^\top + \mathbf{Z}_i \mathbf{Z}_i^\top}{|\mathcal{B}_i| + |\mathcal{Z}_i|}$. By principal component analysis, $\mathcal{S}_i^*$ and $\hat{\mathcal{S}}_i$ are the span of top-q eigenvectors associated with $\mathbf{A}_i$ and $\widetilde{\mathbf{A}}_i$. By Wedin's Theorem (Theorem F.2 in Appendix F), the distance $d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i)$ can be bounded by upper bounding the perturbation between $\mathbf{A}_i$ and $\widetilde{\mathbf{A}}_i$, for example, $\|\mathbf{A}_i - \widetilde{\mathbf{A}}_i\|_F$.

Define $\bar{\mathbf{A}}_i = \frac{\mathbf{B}_i \mathbf{B}_i^\top}{|\mathcal{B}_i|}$ and consider separately $\|\mathbf{A}_i - \bar{\mathbf{A}}_i\|_F$ and $\|\widetilde{\mathbf{A}}_i - \bar{\mathbf{A}}_i\|_F$. By definition, we have

$$
\begin{aligned}
\|\mathbf{A}_i - \bar{\mathbf{A}}_i\|_F &= \left\| \frac{\mathbf{B}_i \mathbf{B}_i^\top}{|\mathcal{B}_i|} - \frac{\mathbf{B}_i \mathbf{B}_i^\top + \mathbf{Y}_i \mathbf{Y}_i^\top}{|\mathcal{B}_i| + |\mathcal{Y}_i|} \right\|_F \\
&\le \left\| \left(\frac{1}{|\mathcal{B}_i|} - \frac{1}{|\mathcal{B}_i| + |\mathcal{Y}_i|} \right) \mathbf{B}_i \mathbf{B}_i^\top \right\|_F + \left\| \frac{\mathbf{Y}_i \mathbf{Y}_i^\top}{|\mathcal{B}_i| + |\mathcal{Y}_i|} \right\|_F \\
&\le \frac{|\mathcal{Y}_i| \cdot \|\mathbf{B}_i\|_F^2}{|\mathcal{B}_i|(|\mathcal{B}_i| + |\mathcal{Y}_i|)} + \frac{\|\mathbf{Y}_i\|_F^2}{|\mathcal{B}_i| + |\mathcal{Y}_i|} \\
&\le \frac{|\mathcal{Y}_i| \cdot |\mathcal{B}_i|}{|\mathcal{B}_i|(|\mathcal{B}_i| + |\mathcal{Y}_i|)} + \frac{|\mathcal{Y}_i|}{|\mathcal{B}_i| + |\mathcal{Y}_i|} \\
&= \frac{2|\mathcal{Y}_i|}{|\mathcal{B}_i| + |\mathcal{Y}_i|} \le \frac{300\phi^2}{1 - 150\phi^2}.
\end{aligned}
$$

Using essentially the same line of argument one can show $\|\widetilde{\mathbf{A}}_i - \bar{\mathbf{A}}_i\|_F \le \frac{300\phi^2}{1-150\phi^2}$ as well. Therefore, $\|\mathbf{A}_i - \widetilde{\mathbf{A}}_i\|_F \le \frac{600\phi^2}{1-150\phi^2}$. Applying Wedin's Theorem and Proposition B.10 we get

$$
d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i) \le \frac{\sqrt{2}\|\mathbf{A}_i - \widetilde{\mathbf{A}}_i\|_F}{\sigma_q(\mathbf{A}_i) - \sigma_{q+1}(\mathbf{A}_i)} \le \frac{\sqrt{2}|\mathcal{X}_i|\|\mathbf{A}_i - \widetilde{\mathbf{A}}_i\|_F}{\sigma_q^2(\mathcal{X}_i) - \sigma_{q+1}^2(\mathcal{X}_i)} \le \frac{\sqrt{2}|\mathcal{X}_i|\|\mathbf{A}_i - \widetilde{\mathbf{A}}_i\|_F}{n(\psi - \eta)} \le \frac{600\sqrt{2}\phi^2}{(1 - 150\phi^2)(\psi - \eta)}.
$$

□

Combining Lemma B.7 to B.11 we arrive at a proof of the key lemma.

Proof of Lemma 3.3. Given $a < \frac{1-802\phi^2}{800\phi^2}$ and $\mathcal{X}$ being (ϕ, η, ψ)-well separated, we have $\mathrm{cost}(\hat{\mathcal{C}}; \mathcal{X}) \le \alpha \Delta_{k-1}^2(\mathcal{X})$ for some $\alpha < \frac{1-802\phi^2}{800}$. By Lemma B.7, $d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i) \le D_i/10$ hold for every $i = 1, \cdots, k$, after possible rearrangement of $\{\hat{\mathcal{S}}_i\}_{i=1}^k$. Applying Lemma B.8, B.9 and B.11 we obtain $d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i) \le \frac{600\sqrt{2}\phi^2}{(1-150\phi^2)(\psi-\eta)}$. Finally, $d_W(\hat{\mathcal{C}}, \mathcal{C}^*) \le \sqrt{k} \max_i d(\mathcal{S}_i^*, \hat{\mathcal{S}}_i) \le \frac{600\sqrt{2}\phi^2\sqrt{k}}{(1-150\phi^2)(\psi-\eta)}$. □

Appendix C Proofs of sample-aggregate private subspace clustering: the stochastic case

In this section we prove Theorem 3.6 that details a stability result for threshold-based subspace clustering under the stochastic datasetting. We first cite the following lemma from [14] which states that (under certain separation conditions) with high probability the similarity graph G recovered by the robust TSC algorithm has no false connections; that is, two data points i and j are connected in G only if they belong to the same cluster (subspace).

Lemma C.1 ([14], Theorem 3; no false connection of TSC)**.** *Suppose* $s \le \min n_\ell/6$ *and*

$$\sqrt{1 - \min_{\ell \neq \ell'} \frac{d^2(\mathcal{S}_\ell^*, \mathcal{S}_{\ell'}^*)}{q}} + \frac{\sigma(1+\sigma)}{\sqrt{\log n}} \frac{\sqrt{q}}{\sqrt{d}} \le \frac{1}{15 \log n} \tag{14}$$

with $d \ge 6 \log n$*; then the similarity graph* G *constructed by Algorithm 2 has* no false connections *with probability at least* $1 - \frac{10}{n} - \sum_\ell n_\ell e^{-c(n_\ell - 1)}$ *for some absolute constant* $c > 0$*.*

Based on Lemma C.1, to prove Lemma 3.5 it remains to show that data points within the same cluster are indeed *connected* in the similarity graph G. The proof is presented in Appendix C.1. With Lemma C.1 and 3.5, Theorem 3.6 can be easily proved as follows:

Proof of Theorem 3.6. By Lemma C.1 and 3.5, we know that under the stated conditions the similarity graph G output by the TSC algorithm has no false connections and is connected per cluster. Fix a cluster ℓ and consider the observed data points $\mathbf{X}^{(\ell)} = \mathbf{Y}^{(\ell)} + \mathbf{E}^{(\ell)}$. Since both the signal $\mathbf{Y}^{(\ell)}$ and the noise $\mathbf{E}^{(\ell)}$ are stochastic, by standard analysis of PCA one can show that the top-q subspace of $\mathbf{X}^{(\ell)}$ converges to the underlying subspace $\mathcal{S}_\ell^*$ in probability as the number of data points n_ℓ goes to infinity [43]. The theorem then holds because $m = o(n)$ and hence $n' = n/m \to \infty$.

□

C.1 Proof of Lemma 3.5

Proposition C.2. *Suppose* $\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{\varepsilon}_i$ *with* $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \frac{\sigma^2}{d}\mathbf{I}_d)$ *and* $\sigma > 0$*. Then with probability at least* $1 - n^2 e^{-\sqrt{d}} - 2/n$ *the following holds:*

$$\left|\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle - \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right| \le (2\sqrt{5}\sigma + 5\sigma^2)\sqrt{\frac{6 \log n}{d}}; \quad \forall i, j \in \{1, \cdots, n\}, i \neq j. \tag{15}$$

Proof. Applying Theorem F.3 (in Appendix F) and set $t = 4$ and $\rho = \sqrt{6 \log n/d}$ in Theorem F.3. Applying also the union bound over all $(i, j) \in \{1, \cdots, n\}$ pairs with $i \neq j$. Then the following

holds uniformly for all $i \neq j$ with probability at least $1 - n^2 e^{-d} - 2/n$:

$$\begin{aligned} \|\boldsymbol{\varepsilon}_i\|_2 &\leq \sqrt{5}\sigma; \\ \left|\langle \boldsymbol{\varepsilon}_i, \boldsymbol{x}_j \rangle\right| &\leq \sqrt{5}\sigma \cdot \sqrt{\frac{6\log n}{d}}, \ \ \forall i, j; \\ \left|\langle \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j \rangle\right| &\leq 5\sigma^2 \cdot \sqrt{\frac{6\log n}{d}}, \ \ \forall i \neq j. \end{aligned}$$

The proof is then completed by noting that

$$\begin{aligned} \left|\langle \boldsymbol{y}_i, \boldsymbol{y}_j \rangle - \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle\right| &\leq \left|\langle \boldsymbol{\varepsilon}_i, \boldsymbol{x}_j \rangle\right| + \left|\langle \boldsymbol{\varepsilon}_j, \boldsymbol{x}_i \rangle\right| + \left|\langle \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j \rangle\right| \\ &\leq (2\sqrt{5}\sigma + 5\sigma^2)\sqrt{\frac{6\log n}{d}}. \end{aligned}$$

□

Lemma C.3 ([14], Lemma 3; extracted from the proof of Lemma 6.2. in [42]). *Let $S^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ denote the d-dimensional unit sphere. For an arbitrary $\boldsymbol{p} \in S^{d-1}$, define $C(\boldsymbol{p}, \theta) = \{\boldsymbol{q} \in S^{d-1} : \vartheta(\boldsymbol{p}, \boldsymbol{q}) \leq \theta\}$ where $\vartheta(\boldsymbol{p}, \boldsymbol{q}) = \arccos(\langle \boldsymbol{p}, \boldsymbol{q} \rangle)$ is the angle between $\boldsymbol{p}$ and $\boldsymbol{q}$. Let $\mathcal{L}(\cdot)$ denote the Lebesgue area of a region and $\Theta(\cdot)$ be the inverse function of $\mathcal{L}(C(\boldsymbol{p}, \theta))$ with respect to θ. Then for each $d \geq 1$ and $M \geq 1$, there exists a partition $R_1, \cdots, R_M$ of the unit sphere $\mathbb{S}^{d-1}$ such that [5] $\sup_{\boldsymbol{x}, \boldsymbol{y} \in R_m} \vartheta(\boldsymbol{x}, \boldsymbol{y}) \leq \theta^*$ for every $m = 1, \cdots, M$. Here $\theta^* = 8\Theta(\mathcal{L}(S^{d-1})/M)$.*

We are now ready to prove Lemma 3.5.

Proof of Lemma 3.5. By Lemma C.1 we already know that the similarity graph G has no false connections. Fix a cluster $\ell \in \{1, \cdots, k\}$. Let $\boldsymbol{x}_i^{(\ell)} = \mathbf{U}^{(\ell)} \boldsymbol{a}_i^{(\ell)}$ where $\boldsymbol{a}_i^{(\ell)} \in S^{q-1}$. Set $M = n_\ell / (\gamma \log n_\ell)$ and let $R_1, \cdots, R_M$ be a partition of the unit sphere S^{q-1} as characterized in Lemma C.3. Here q is the intrinsic rank of an underlying subspace. We need to prove the following two properties hold with high probability: (A) every region R_m contains at least one point in $\mathcal{A}^{(\ell)} = \{\boldsymbol{a}_i^{(\ell)}\}_{i=1}^{n_\ell}$; (B) for every $\boldsymbol{a}_i^{(\ell)}$, all data points $\boldsymbol{a}_j^{(\ell)}$ belonging to the neighboring region of the region containing $\boldsymbol{a}_i^{(\ell)}$ are connected with $\boldsymbol{a}_i^{(\ell)}$.

Property (A) is easy to prove. By union bound, the probability that some region is empty can be upper bounded by

$$M\left(1 - \frac{1}{M}\right)^{n_\ell} \leq M e^{-n_\ell / M} = \frac{n_\ell^{1-\gamma}}{\gamma \log n_\ell}.$$

We next turn to prove Property (B). Unlike the noiseless case, the s-nearest-neighbor graph is computed based on the noise-perturbed data points $\{\boldsymbol{y}_i\}_{i=1}^n$. It is no longer true that data points belonging to neighboring regions have larger inner products compared to data points that do not belong to the same or neighboring regions. Hence, we adopt a different argument from the one presented in [14]. Instead of showing that $|C(\boldsymbol{x}_i^{(\ell)}, 3\theta^*)| \leq \tilde{s} = s/2$, we show that $|C(\boldsymbol{a}_i^{(\ell)}, r\theta^*)| \leq \tilde{s}$ for some $r \gg 3$ and in addition $|\langle \boldsymbol{y}_j^{(\ell)}, \boldsymbol{y}_i^{(\ell)} \rangle| > |\langle \boldsymbol{y}_{j'}^{(\ell)}, \boldsymbol{y}_i^{(\ell)} \rangle|$ for every $\boldsymbol{a}_j^{(\ell)} \notin C(\boldsymbol{a}_i^{(\ell)}, r\theta^*)$ and $\boldsymbol{a}_{j'}^{(\ell)} \in C(\boldsymbol{x}_i^{(\ell)}, 3\theta^*)$. [6] This guarantees that all points in $C(\boldsymbol{a}_i^{(\ell)}, 3\theta^*)$ are connected to $\boldsymbol{y}_i^{(\ell)}$ in the s-nearest-neighbor graph.

Fix $\boldsymbol{a}_i^{(\ell)} \in \mathcal{A}^{(\ell)}$ and set r such that

$$r\theta^* = \left(\frac{0.01(q/2 - 1)(q-1)}{\sqrt{\pi}}\right)^{\frac{1}{q-1}}. \tag{16}$$

[5] By definition $R_1 \cup \cdots \cup R_M = S^{d-1}$ and $R_i \cap R_j = \emptyset$ for $i \neq j$.

[6] Note that $|\langle \boldsymbol{y}, \boldsymbol{y}_i^{(\ell)} \rangle| = |\langle -\boldsymbol{y}, \boldsymbol{y}_i^{(\ell)} \rangle|$ by symmetry. So a point far from $\boldsymbol{y}_i^{(\ell)}$ could have large inner product and be connected with $\boldsymbol{y}_i^{(\ell)}$. We take $\tilde{s} = s/2$ to avoid this issue.

Define $p = \mathcal{L}(C(\boldsymbol{a}_i^{(\ell)}, r\theta^*))/\mathcal{L}(S^{q-1})$, where θ^* is given in Lemma C.3. By definition, $\mathbb{E}[|C(\boldsymbol{a}_i^{(\ell)}, r\theta^*)|] = pn_\ell$. Note that by symmetry p does not depend on $\boldsymbol{a}_i^{(\ell)}$. Set $\bar{s} = (0.01 + p)n_\ell$. We then have

$$\frac{\bar{s}}{n_\ell} = 0.01 + \frac{\mathcal{L}(C(\boldsymbol{a}_i^{(\ell)}, r\theta^*))}{L(S^{q-1})} \leq 0.01 + \frac{\mathcal{L}(S^{q-2})(r\theta^*)^{q-1}}{L(S^{q-1})(q-1)} \leq 0.01 + \frac{\sqrt{\pi}\Gamma(\frac{q-1}{2})(r\theta^*)^{q-1}}{\Gamma(\frac{q}{2})(q-1)} \leq 0.02. \tag{17}$$

The second inequality is an application of Eq. (5.2) in [42] and the last inequality is due to Eq. (16). On the other hand, by tail bounds of binomial distribution (Theorem 1 in [41]) we have

$$\Pr\left[\left|C(\boldsymbol{a}_i^{(\ell)}, r\theta^*)\right| > n_\ell(p + 0.01)\right] \leq e^{-\frac{0.01^2 n_\ell^2}{2(pn_\ell + 0.01n_\ell/3)}} \leq e^{-\frac{n_\ell}{400}}, \tag{18}$$

where in the last inequality we used the fact that $pn_\ell \leq 0.01n_\ell$. Since $\bar{s} \leq 0.02n_\ell \leq \tilde{s}$, we proved that with probability at least $1 - e^{-\frac{n_\ell}{400}}$ there will be no more than $\tilde{s}$ data points contained in $C(\boldsymbol{a}_i^{(\ell)}, r\theta^*)$.

The final step of the proof is to show that $|\langle \boldsymbol{y}_j^{(\ell)}, \boldsymbol{y}_i^{(\ell)}\rangle| > |\langle \boldsymbol{y}_{j'}^{(\ell)}, \boldsymbol{y}_j^{(\ell)}\rangle|$ for every $\boldsymbol{a}_j^{(\ell)} \notin C(\boldsymbol{a}_i^{(\ell)}, r\theta^*)$ and $\boldsymbol{a}_{j'}^{(\ell)} \in C(\boldsymbol{x}_i^{(\ell)}, 3\theta^*)$. By Proposition C.2, we have with probability at least $1 - ne^{-\sqrt{d}}$

$$\left|\langle \boldsymbol{y}_j^{(\ell)}, \boldsymbol{y}_i^{(\ell)}\rangle\right| \leq \left|\langle \boldsymbol{a}_j^{(\ell)}, \boldsymbol{a}_i^{(\ell)}\rangle\right| + (2\sqrt{5}\sigma + 5\sigma^2)\sqrt{\frac{6\log n}{d}} \leq \cos(r\theta^*) + (2\sqrt{5}\sigma + 5\sigma^2)\sqrt{\frac{6\log n}{d}} \tag{19}$$

and

$$\left|\langle \boldsymbol{y}_{j'}^{(\ell)}, \boldsymbol{y}_i^{(\ell)}\rangle\right| \geq \left|\langle \boldsymbol{a}_{j'}^{(\ell)}, \boldsymbol{a}_i^{(\ell)}\rangle\right| - (2\sqrt{5}\sigma + 5\sigma^2)\sqrt{\frac{6\log n}{d}} \geq \cos(3\theta^*) - (2\sqrt{5}\sigma + 5\sigma^2)\sqrt{\frac{6\log n}{d}}. \tag{20}$$

Since $r\theta^*$ is dictated in Eq. (16), we only need to obtain an upper bound on θ^*. Following the same argument on page 25 in [14] we have

$$\theta^* \leq 4\pi\left(\frac{\sqrt{2\pi q}}{M}\right)^{\frac{1}{q-1}} = 4\pi\left(\frac{\gamma\sqrt{2\pi q}\log n_\ell}{n_\ell}\right)^{\frac{1}{q-1}}. \tag{21}$$

Consequently, $|\langle \boldsymbol{y}_j^{(\ell)}, \boldsymbol{y}_i^{(\ell)}\rangle| > |\langle \boldsymbol{y}_{j'}^{(\ell)}, \boldsymbol{y}_j^{(\ell)}\rangle|$ when $\bar{\sigma} = 2\sqrt{5}\sigma + 5\sigma^2$ satisfies

$$\bar{\sigma} < \sqrt{\frac{d}{24\log n}}\left[\cos\left(12\pi\left(\frac{\gamma\sqrt{2\pi q}\log n_\ell}{n_\ell}\right)^{\frac{1}{q-1}}\right) - \cos\left(\left(\frac{0.01(q/2-1)(q-1)}{\sqrt{\pi}}\right)^{\frac{1}{q-1}}\right)\right]. \tag{22}$$

The right-hand side of the above condition is strictly positive if n_ℓ satisfies

$$n_\ell > \frac{\gamma\pi\sqrt{2q}\log n_\ell}{0.01(q/2-1)(q-1)} \cdot (12\pi)^{q-1}.$$

□

Appendix D Supplementary materials for private subspace clustering via the exponential mechanism

D.1 Proof of Proposition 4.1

Proof. Define the score function $h(\cdot; \boldsymbol{\theta})$ as $h(\mathcal{X}; \boldsymbol{\theta}) = \sum_{i=1}^n d^2(\boldsymbol{x}_i, \mathcal{S}_{z_i})$. Since $\|\boldsymbol{x}_i\|_2 \leq 1$, it is straightforward that $h(\cdot; \boldsymbol{\theta})$ has global sensitivity upper bounded by 1; that is, $\sup_{d(\mathcal{X},\mathcal{X}')=1} |h(\mathcal{X}; \boldsymbol{\theta}) - h(\mathcal{X}'; \boldsymbol{\theta})| \leq 1$ for all $\boldsymbol{\theta}$. Eq. (10) is then a direct application of the exponential mechanism. □

Algorithm 3 Gibbs sampling for matrix Bingham distribution (Eq. (23))

Input: symmetric matrix $\mathbf{A}$, diagonal matrix $\mathbf{B}$, current sample $\mathbf{U}$, dimensions d and q.
for each $r \in \{1, \cdots, q\}$ in random order **do**
  Let $\mathbf{U}_{(r)} \in \mathbb{R}^d$ be the rth column of $\mathbf{U}$ and $\mathbf{U}_{(-r)}$ be the matrix excluding $\mathbf{U}_{(r)}$.
  Let $\mathbf{N}$ be an orthonormal basis of the null space of $\mathbf{U}_{(-r)}$.
  Compute $\boldsymbol{z} = \mathbf{N}^\top \mathbf{U}_{(r)}$ and $\widetilde{\mathbf{A}} = \mathbf{B}_{rr}\mathbf{N}^\top \mathbf{A}\mathbf{N}$.
  Update $\boldsymbol{z}$ by Gibbs sampling from the vector Bingham distributrion with parameter $\widetilde{\mathbf{A}}$.
  Set $\mathbf{U}_{(r)} = \mathbf{N}\boldsymbol{z}$.
end for
Output: the updated sample $\mathbf{U}$.

Algorithm 4 Gibbs sampling for vector Bingham distribution (Eq. (24))

Input: symmetric matrix $\mathbf{A}$, current sample $\boldsymbol{x}$, dimension d.
Let $\mathbf{A} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}^\top$, $\boldsymbol{\Lambda} = \mathbf{diag}(\boldsymbol{\lambda})$ be the eigen-decomposition of $\mathbf{A}$. Compute $\boldsymbol{y} = \mathbf{E}^\top \boldsymbol{x}$.
for each $j \in \{1, \cdots, d\}$ in random order **do**
  Compute $q_1, \cdots, q_d$ as $y_1^2/(1-y_i^2), \cdots, y_d^2/(1-y_i^2)$.
  Sample $\theta \in (0,1)$ from the density $p(\theta) \propto e^{(\lambda_i - \boldsymbol{q}_{-i}^\top \boldsymbol{\lambda}_{-i})\theta} \times \theta^{-1/2}(1-\theta)^{(d-3)/2}$.
  Set $s = +1$ or -1 with equal probability.
  Set $y_i = s_i \theta^{1/2}$ and for each $j \neq i$ set $y_j^2 = (1-\theta)q_j$, leaving the sign unchanged.
end for
Output: the updated sample $\boldsymbol{x} = \mathbf{E}\boldsymbol{y}$.

D.2 Gibbs sampling for matrix Bingham distribution

In this section we give details of a Gibbs sampler proposed in [16] for sampling from a matrix Bingham distribution. One component in the Gibbs sampler (the rejection sampling step) is slightly modified to make the sampling more efficient.

The objective is to sample from the following matrix-Bingham distribution:

$$p(\mathbf{U}; \mathbf{A}, \mathbf{B}) \propto \exp(\mathrm{tr}(\mathbf{B}\mathbf{U}^\top \mathbf{A}\mathbf{U})), \tag{23}$$

where $\mathbf{U}$ is a $d \times q$ matrix lying on a Stiefel manifold; that is, $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_{q\times q}$. In our problem $\mathbf{A}$ is an unnormalized sample covariance matrix and $\mathbf{B} = \varepsilon \mathbf{I}_{q\times q}$, with ε the privacy budget. As a simplified case, when $q = 1$ we arrive at a vector version of the Bingham distribution:

$$p(\boldsymbol{x}; \mathbf{A}) \propto \exp(\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x}), \tag{24}$$

with $\boldsymbol{x}$ constrained on the d-dimensional sphere $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$. Gibbs samplers for both Eq. (23) and (24) were proposed in [16] and presented in Algorithm 3 and 4.

In Algorithm 4, step 4 requires sampling from a non-standard 1-dimensional distribution

$$p(x; k, a) \propto x^{-1/2}(1-x)^k e^{ax} \cdot \mathbf{1}_{0<x<1} =: f(x). \tag{25}$$

In [16] a rejection sampling algorithm was proposed to sample x from Eq. (25), with a $\mathrm{Beta}(1/2, 1+\min(k, \max(k-a, -1/2)))$ envelope distribution. However, such a distribution is highly inefficient when $|a| \gg 0$ for which no Beta distribution serves as a good envelope distribution. To address this problem, we propose two separate rejection sampling schemes for Eq. (25) when $|a| \gg 0$.

Case 1: $a \ll 0$ In this case, the mass of the distribution will concentrate on $x \to 0$. We use Gamma distribution $\Gamma(1/2, 1/|a|)$ truncated on $(0,1)$ as an envelope distribution. That is, $x \sim g(\cdot)$ and $g(\cdot)$ is defined as

$$g(x) = \frac{1}{Z} \cdot x^{-1/2} e^{ax} \cdot \mathbf{1}_{0<x<1},$$

with Z a normalization constant. The constant $M = \sup_x f(x)/g(x)$ can be computed as

$$M = Z \cdot \sup_{0<x<1} \frac{x^{-1/2}(1-x)^k e^{ax}}{x^{-1/2}e^{ax}} \leq Z.$$

The step-by-step algorithm is as follows:

1. Sample $x \sim \Gamma(1/2, 1/|a|)$. If $x \geq 1$, throw away x and re-draw the sample.
2. Sample $u \in (0, 1)$ from the uniform distribution over $(0, 1)$.
3. If $u \leq (1-x)^k$, accept the sample; otherwise reject the sample and try again.

The proposed rejection sampling algorithm is efficient because when $a \ll 0$, the envelope distribution g has very high density over the region near zero; consequently, $(1-x)^k$ is close to one and hence the acceptance rate is high.

Case 2: $a \gg 0$ In this case, the mass of the distribution will concentrate on $x \to 1$. However, we have a singularity at $x = 0$ (i.e., $\lim_{x \to 0} f(x) = \infty$). This makes the sampling particularly difficult as a distribution proportional to e^{ax} will be infinitely off at the region near zero. To circumvent the problem, we propose a mixture distribution as the envelope, which have good approximation property at both regions near 0 and 1.

First define density h as

$$h(x) = \frac{1}{Z} \cdot (1-x)^k e^{ax} \cdot \mathbf{1}_{0<x<1},$$

where $Z = \int_0^1 (1-x)^k e^a x \mathrm{d}x$ is the normalization constant. Note that samples from $h(\cdot)$ can be obtained by first sampling z from a Gamma distribution $\Gamma(k+1, 1/a)$ truncated on $(0, 1)$ and then apply transform of variable $x = 1 - z$. The density of the envelope distribution $g(\cdot)$ is then defined as a mixture distribution:

$$g(x) = \frac{1}{Z} \cdot \mathrm{Beta}(x; 1/2, k+1) + \left(1 - \frac{1}{Z}\right) \cdot h(x).$$

The constant $M = \sup f(x)/g(x)$ can be computed by

$$\begin{aligned} M &= \max\left(\sup_{0<x\leq 1/a} \frac{f(x)}{g(x)}, \sup_{1/a<x<1} \frac{f(x)}{g(x)} \right) \\ &\leq \max\left(Z \cdot \sup_{0<x\leq 1/a} \frac{x^{-1/2}(1-x)^k e^{ax} \cdot B(1/2, k+1)}{x^{-1/2}(1-x)^k}, 2 \cdot \sup_{1/a<x<1} \frac{x^{-1/2}(1-x)^k e^{ax}}{(1-x)^k e^{ax}} \right) \\ &\leq Z \cdot \max(2\sqrt{a}, eB(1/2, k+1)). \end{aligned}$$

Here for the second inequality we apply $Z \geq 2$ for reasonably large a and $B(\cdot, \cdot)$ is the Beta function. The normalization constant Z can be approximately computed using numerical integration. However, empirical evidence suggests that Z is huge for large a values (e.g., $Z > 10^{100}$ if $a > k + 500$). Therefore, we could simply take $Z \to \infty$, which simplifies the rejection sampling algorithm as follows:

1. Sample $z \sim \Gamma(k+1, 1/a)$. If $z \geq 1$ then throw away z and re-draw the sample.
2. Compute $x = 1 - z$, $M = \max(2\sqrt{a}, \exp(1) \cdot B(1/2, k+1))$.
3. Sampe u from the uniform distribution over $(0, 1)$.
4. If $u < x^{-1/2}/(M(1 + e^{-ax}\mathrm{Beta}(x; 1/2, k+1)))$, accept the sample; otherwise reject the sample and try again.

The proposed rejection sampling scheme is efficient because when $a \gg 0$ the density $h(\cdot)$ is very skewed to one. Therefore, $x^{-1/2}$ will be close to 1 and e^{-ax} will be very samll, which means the acceptance rate is high.

D.3 Justification of generative model in Section 4.2

Recall the generative model presented in Section 4.2:

1. For each $\ell \in [k]$, sample $\mathbf{U}_\ell$ (orthonormal basis of $\mathcal{S}_\ell$) uniformly at random from $\mathbb{S}_q^d$.
2. For each $i \in [n]$, sample $z_i \in [k]$ such that $\Pr[z_i = j] = 1/k$, $\boldsymbol{y}_i$ uniformly at random from the q-dimensional unit ball, and $\boldsymbol{w}_i \sim \mathcal{N}(0, \mathbf{I}_d/\varepsilon)$. Set $\boldsymbol{x}_i = \mathbf{U}_\ell \boldsymbol{y}_i + \mathcal{P}_{\mathbf{U}_\ell^\perp} \boldsymbol{w}_i$.

Algorithm 5 Differentially private query answering via the SuLQ framework

1: **Input**: query parameters $\mathcal{S}_1, \cdots, \mathcal{S}_k \in \mathbb{R}_d^q$, $\ell \in [k]$; privacy parameters $\varepsilon, \delta > 0$.
2: Let $\mathbf{A}_\ell = \{\boldsymbol{x}_i : \operatorname{argmin}_{\ell'} d(\boldsymbol{x}_i, \mathcal{S}_{\ell'}) = \ell\}$ and form $\mathbf{B} = \mathbf{A}_\ell \mathbf{A}_\ell^\top$.
3: **Noise calibration**: Set $\widetilde{\mathbf{B}} = \mathbf{B} + \sigma \mathbf{W}$, where $\mathbf{W}$ is a standard Normal random matrix and $\sigma = 2\sqrt{2\ln(1.25/\delta)}/\varepsilon$.
4: **Singular value decomposition**: Let $\widetilde{\mathbf{B}} = \mathbf{U}\mathbf{V}\mathbf{D}^\top$ be the top-q singular value decomposition of $\widetilde{\mathbf{B}}$. $\mathbf{U} \in \mathbb{R}^{d \times q}$ denotes the top q left singular vectors of $\widetilde{\mathbf{B}}$.
5: **Output**: new subspace $\mathcal{S}'_\ell$ spanned by columns of $\mathbf{U}$.

In this section we derive a Gibbs sampler for the considered model and show that the derived Gibbs sampler is identical to the one presented in Section 4.1. This result establishes formal connection between our proposed Gibbs sampling algorithm for private subspace clustering and a probabilistic graphical model that resembles the mixtures of probabilistic PCA (MPPCA, [27]) model.

First we note that the prior distribution specified in the generative model is completely non-informative; that is, $p_0(\boldsymbol{\theta}) = p_0(\boldsymbol{\theta}')$ for any $\boldsymbol{\theta} = (\mathcal{C}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ and $\boldsymbol{\theta}' = (\mathcal{C}', \boldsymbol{x}', \boldsymbol{y}', \boldsymbol{z}')$. On the other hand, the likelihood model is as follows:

$$p(\boldsymbol{x}_i | z_i = \ell, \boldsymbol{y}_i, \mathcal{C}) = \begin{cases} \mathcal{N}(\boldsymbol{x}_i; \mathbf{U}_\ell \boldsymbol{y}_i, \mathbf{I}_d/\varepsilon), & \text{if } \mathcal{P}_{\mathcal{S}_\ell} \boldsymbol{x}_i = \mathbf{U}_\ell \boldsymbol{y}_i; \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Here $\mathbf{U}_\ell \in \mathbb{R}^{d \times q}$ is an orthonormal basis associated with $\mathcal{S}_\ell$ and $\mathcal{P}_{\mathcal{S}_\ell}$ stands for the projection operator onto subspace $\mathcal{S}_\ell$. Integrating $\boldsymbol{y}_i$ out we obtain

$$p(\boldsymbol{x}_i | z_i = \ell, \mathcal{C}) \propto \exp\left(-\frac{\varepsilon}{2} \cdot d^2(\boldsymbol{x}_i, \mathcal{S}_\ell)\right). \tag{27}$$

A Gibbs sampler can then be derived as follows:

Update of z_i By Eq. (27), the conditional distribution of z_i is

$$p(z_i = \ell | \boldsymbol{x}_i, \mathcal{C}) \propto p_0(z_i = \ell) p(\boldsymbol{x}_i | z_i = \ell, \mathcal{C}) \propto \exp\left(-\frac{\varepsilon}{2} \cdot d^2(\boldsymbol{x}_i, \mathcal{S}_\ell)\right).$$

Therefore, we can sample z_i from a normalized categorical distribution as specified above.

Update of $\mathcal{S}_\ell$ By Eq. (27), the conditional distribution of $\mathcal{S}_\ell$ is

$$p(\mathcal{S}_\ell | \boldsymbol{x}, \boldsymbol{z}) \propto p_0(\mathcal{S}_\ell) \prod_{z_i = \ell} p(\boldsymbol{x}_i | \boldsymbol{z}_i = \ell, \mathcal{S}_\ell) \propto \exp\left(-\frac{\varepsilon}{2} \cdot \sum_{z_i = \ell} d^2(\boldsymbol{x}_i, \mathcal{S}_\ell)\right).$$

Denote $\mathbf{A}_\ell = \{\boldsymbol{x}_i : z_i = \ell\}$ as all data points in cluster ℓ and let $\mathbf{U}_\ell$ be the orthonormal basis of $\mathcal{S}_\ell$. We then have

$$p(\mathbf{U}_\ell | \boldsymbol{x}, \boldsymbol{z}) \propto \exp\left(\frac{\varepsilon}{2} \cdot \operatorname{tr}(\mathbf{U}_\ell^\top \mathbf{A}_\ell \mathbf{U}_\ell)\right),$$

which corresponds to a matrix Bingham distribution.

The above presented Gibbs sampler is identical to the one proposed in Section 4.1 in the main text, thus justifying our use of the above-mentioned generative model as an equivalent characterization of the proposed private subspace clustering algorithm. This is perhaps not surprising, as the marginal likelihood model Eq. (27) is exactly the same with the sampling distribution dictated by the exponential mechanism, as shown in Eq. (10) in the main text.

Appendix E Private subspace clustering via the SuLQ framework

In this section we introduce a simple iterative subspace clustering algorithm based on the SuLQ framework [2]. Before presenting the algorithm, we first review k-plane [3], a straightforward iterative method for subspace clustering:

1. For each data point $\boldsymbol{x}_i$, compute $z_i = \operatorname{argmin}_{1 \le \ell \le k} d(\boldsymbol{x}_i, \mathcal{S}_k)$.

2. For each cluster ℓ, let $\mathbf{A}_\ell = \{\boldsymbol{x}_i : z_i = \ell\} \in \mathbb{R}^{d \times n_\ell}$ denote all data points assigned to cluster ℓ. Update $\mathcal{S}_\ell$ as the linear subspace spanned by the top-q eigenvectors of $\mathbf{A}_\ell \mathbf{A}_\ell^\top$.
3. Repeat step 1 and 2 until convergence.

Suppose the k-plane algorithm is run for T iterations. From the pseudocode of k-plane, the algorithm needs to query the database $\mathcal{X}$ for kT times, each time asking the following question:

- Given $\mathcal{S}_1, \cdots, \mathcal{S}_k$ and $\ell \in [k]$ as inputs, output the orthonormal basis $\mathbf{U}_\ell \in \mathbb{R}^{d \times q}$ of a q-dimensional subspace $\mathcal{S}'_\ell$ such that $\mathcal{S}'_\ell$ best captures $\mathbf{A}_\ell^\top \mathbf{A}_\ell$; i.e., $\|\mathbf{A}_\ell \mathbf{A}_\ell^\top - (\mathcal{P}_{\mathcal{S}'_\ell} \mathbf{A}_\ell)(\mathcal{P}_{\mathcal{S}'_\ell} \mathbf{A}_\ell)^\top\|_2$ is minimized. Here $\mathbf{A}_\ell$ is defined in terms of $(\mathcal{S}_1, \cdots, \mathcal{S}_k)$.

Algorithm 5 is a simple procedure that approximately answers the above question while preserving (ε, δ)-differential privacy. It is in fact a special case of the SuLQ framework proposed in [2]. The following proposition is immediate.

Proposition E.1. *Algorithm 5 is an (ε, δ)-differentially private algorithm.*

Proof. Define $\boldsymbol{b}(\mathcal{X}) = \mathrm{vec}(\mathbf{A}_\ell^\top \mathbf{A}_\ell) \in \mathbb{R}^{d^2}$. Let $\mathcal{X}'$ be an arbitrary database such that $d(\mathcal{X}, \mathcal{X}') = 1$. That is, exactly one column $\boldsymbol{x}$ in $\mathcal{X}$ is replaced by a new column $\boldsymbol{x}'$ in $\mathcal{X}'$. We then have

$$\|\boldsymbol{b}(\mathcal{X}) - \boldsymbol{b}(\mathcal{X}')\|_2^2 \leq \sum_{i,j=1}^{d} (x'_i x'_j - x_i x_j)^2 \leq 2 \sum_{i,j=1}^{d} (x_i'^2 x_j'^2 + x_i^2 x_j^2) \leq 4,$$

where the last inequality is due to the constraint $\|\boldsymbol{x}\|_2, \|\boldsymbol{x}'\|_2 \leq 1$. Consequently,

$$\Delta_2 \boldsymbol{b} = \sup_{d(\mathcal{X}, \mathcal{X}')=1} \|\boldsymbol{b}(\mathcal{X}) - \boldsymbol{b}(\mathcal{X}')\|_2 \leq 2.$$

The Gaussian mechanism (Theorem A.1, [9]) then suggests that one can release $\boldsymbol{b}$ while preserving (ε, δ)-differential privacy by calibrating i.i.d. Gaussian noise to $\boldsymbol{b}$:

$$\text{Release} \quad \boldsymbol{b}(\mathcal{X}) + \frac{2\sqrt{2\ln(1.25/\delta)}}{\varepsilon} \cdot \boldsymbol{w},$$

where $\boldsymbol{w}$ is a d^2-dimensional standard Normal. The final singular value decomposition step does not affect privacy because differential privacy is close to post-processing. □

The following proposition is then a direct application of advanced composition [9].

Proposition E.2. *Suppose the k-plane algorithm is run for T iterations, each iteration querying Algorithm 5 k times with privacy parameters ε and δ. Then the overall algorithm is (ε', δ')-differentially private with*

$$\begin{aligned} \varepsilon' &= \sqrt{2kT\ln(1/\delta)}\varepsilon + kT\varepsilon(e^\varepsilon - 1), \\ \delta' &= (kT+1)\delta. \end{aligned}$$

Appendix F Concentration theorems

Theorem F.1 ([44], Theorem 1.2)**.** *Let $\mathbf{A}$ be an $n \times n$ matrices with entries i.i.d. sampled from standard Gaussian distribution. Then there exist absolute constants $c_1 > 0, 0 < c_2 < 1$ such that for every $t > 0$,*

$$\Pr\left[\sigma_n(\mathbf{A}) \leq t\sqrt{n}\right] \leq c_1 t + c_2^n,$$

where $\sigma_n(\mathbf{A})$ is the least singular value of $\mathbf{A}$.

Theorem F.2 (Wedin's theorem; Theorem 4.1, pp. 260 in [45])**.** *Let $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{m \times n}$ be given matrices with $m \geq n$. 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*

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

then

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

Theorem F.3 ([32], Lemma 18, Properties of Gaussian random vectors)**.** *Let* $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \frac{\sigma^2}{d}\mathbf{I})$ *be a* d*-dimensional random Gaussian vector with coordinate-wise variance* σ^2*. Then the following holds for some fixed* $\boldsymbol{z} \in \mathbb{R}^d$ *and* $t, \rho > 0$*:*

$$\begin{aligned} \Pr\left[\|\boldsymbol{\varepsilon}_i\|_2^2 > (1+t)\sigma^2\right] &\leq e^{\frac{n}{2}(\log(t+1)-t)}; \\ \Pr\left[\left|\langle \boldsymbol{\varepsilon}_i, \boldsymbol{z}\rangle\right| > \rho\|\boldsymbol{\varepsilon}_i\|_2\|\boldsymbol{z}\|_2\right] &\leq 2e^{-\frac{n\rho^2}{2}}. \end{aligned}$$

References

[35] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SULQ framework. In *PODS*, 2015.

[36] P. S. Bradley and O. L. Mangasarian. k-plane clustering. *Journal of Global Optimization*, 16(1), 2000.

[37] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[38] R. Heckel and H. Bölcskei. Robust subspace clustering via thresholding. *arXiv:1307.4891*, 2013.

[39] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[40] P. Hoff. Simulation of the matrix bingham-conmises-fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009.

[41] S. Janson. On concentration of probability. *Contemporary combinatorics*, 10(3):1–9, 2002.

[42] P. Leopardi. Diameter bounds for equal area partitions of the unit sphere. *Electronic Transactions on Numerical Analysis*, 35:1–16, 2009.

[43] D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17:1617–1642, 2007.

[44] M. Rudelson and R. Vershynin. The littlewood–offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[45] G. W. Stewart, J.-g. Sun, and H. B. Jovanovich. *Matrix perturbation theory*. Academic press New York, 1990.

[46] M. Tipping and C. Bishop. Mixtures of probabilistic principle component anlyzers. *Neural computation*, 11(2):443–482, 1999.

[47] Y.-X. Wang and H. Xu. Noisy sparse subspace clustering. In *ICML*, pages 89–97, 2013.