# A Deterministic Analysis of Noisy Sparse Subspace Clustering for Dimensionality-reduced Data

**Yining Wang** YININGWA@CS.CMU.EDU
**Yu-Xiang Wang** YUXIANGW@CS.CMU.EDU
**Aarti Singh** AARTI@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

## Abstract

Subspace clustering groups data into several low-rank subspaces. In this paper, we propose a theoretical framework to analyze a popular optimization-based algorithm, Sparse Subspace Clustering (SSC), when the data dimension is compressed via some random projection algorithms. We show SSC provably succeeds if the random projection is a *subspace embedding*, which includes random Gaussian projection, uniform row sampling, FJLT, sketching, etc. Our analysis applies to the most general deterministic setting and is able to handle both adversarial and stochastic noise. It also results in the first algorithm for privacy-preserved subspace clustering.

## 1. Introduction

Subspace clustering groups a collection of data points into $k$ clusters so that data points within a single cluster lie near some low rank subspace. It has found a wide range of applications as many high dimensional data can be approximated by a union of low rank subspaces. Some examples include motion trajectories (Costeira & Kanade, 1998), face images (Basri & Jacobs, 2003), network hop counts (Eriksson et al., 2012), movie ratings (Zhang et al., 2012) and social graphs (Jalali et al., 2011).

A large body of research has been devoted to subspace clustering in the last decade. Recently a class of convex optimization based algorithms, in particular Low Rank Representation (LRR, (Liu et al., 2013)) and Sparse Subspace Clustering (SSC, (Elhamifar & Vidal, 2013)), have drawn much interest from the literature. It is known that SSC enjoys superb performance in practice (Elhamifar & Vidal, 2009) and have theoretical guarantee under fairly general conditions (Soltanolkotabi et al., 2012; Wang & Xu, 2013; Soltanolkotabi et al., 2014).

Let $\mathbf{X} \in \mathbb{R}^{d\times N}$ denote the data matrix, where $d$ is the ambient dimension and $N$ is the number of data points. For noiseless data (i.e., data points lie exactly on low-rank subspaces), the exact SSC algorithm solves the optimization problem in Eq. (1.1) for each data point $\boldsymbol{x}_i$ to obtain self regression solutions $\boldsymbol{c}_i \in \mathbb{R}^N$.

$$\min_{\boldsymbol{c}_i\in\mathbb{R}^N} \|\boldsymbol{c}_i\|_1, \quad \text{s.t. } \boldsymbol{x}_i = \mathbf{X}\boldsymbol{c}_i, \boldsymbol{c}_{ii} = 0. \tag{1.1}$$

For noisy data, the following Lasso version of SSC is often used in practice:

$$\min_{\boldsymbol{c}_i\in\mathbb{R}^N} \|\boldsymbol{x}_i - \mathbf{X}\boldsymbol{c}_i\|_2^2 + 2\lambda\|\boldsymbol{c}_i\|_1, \quad \text{s.t. } \boldsymbol{c}_{ii} = 0. \tag{1.2}$$

Although success conditions for both exact SSC and Lasso SSC have been extensively analyzed in previous literature, in practice it is inefficient or even infeasible to operate on data with high dimension. Some types of dimension reduction is usually required. In this paper, we propose a theoretical framework that analyzes SSC under many popular dimension reduction settings, including

- **Compressive measurement**: For compressive measurement dimensionality-reduced data are obtained by multiplying the original data typically with a random Gaussian matrix. We show that SSC provably succeeds when the projected dimension is at the order of the maximum intrinsic rank of each subspace.
- **Efficient computation**: By using fast Johnson-Lindenstrauss transform (Ailon & Chazelle, 2009) or sketching (Charikar et al., 2004; Clarkson & Woodruff, 2013) one can computationally efficiently reduce the data dimension while still preserving important structures in the underlying data. We prove similar results for both FJLT and sketching.
- **Handling missing data**: In many applications the data matrix may be incomplete due to measurement and sensing limits. It is shown in this paper that when data meet some incoherent criteria uniform feature sampling suffices for SSC.

*Proceedings of the $32^{nd}$ International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

- **Data privacy**: Privacy is an important concern in modern machine learning applications. It was shown that random Johnson-Lindenstrauss transform with added Gaussian noise preserves both information-theoretic (Zhou et al., 2009) and differential privacy (Kenthapadi et al., 2013). We provide a utility analysis which shows that SSC can achieve *exact* subspace detection despite stringent privacy constraints.

A key observation is that all projections for the aforementioned settings are *subspace embeddings*, which means they uniformly preserve the two norm of any vector belonging to a low-rank subspace. Our analysis applies to the *fully deterministic* setting under which both subspaces and data points within each subspace are placed deterministically. It can also handle data corrupted by deterministic or stochastic noise. This generalizes previous work (Heckel et al., 2014) which only applies to semi-random models with noiseless data [1]. The fully deterministic setting poses more challenges because the perturbation of dual directions introduced in (Soltanolkotabi et al., 2012) cannot be easily bounded if exact SSC is used. As a result, even for noiseless data, we employ a Lasso SSC formulation to obtain strong convexity in the dual problem.

## 2. Problem setup

**Notations** The uncorrupted data matrix is denoted as $\mathbf{Y} \in \mathbb{R}^{d\times N}$, where $d$ is the ambient dimension and $N$ is the total number of data points. $\mathbf{Y}$ is normalized so that each column has unit two norm. Each column in $\mathbf{Y}$ belongs to a union of $k$ subspaces $\mathcal{U}^{(1)} \cup \cdots \mathcal{U}^{(k)}$. For each subspace $\mathcal{U}^{(\ell)}$ we write $\mathbf{Y}^{(\ell)} = (\boldsymbol{y}_1^{(\ell)}, \cdots, \boldsymbol{y}_{N_\ell}^{(\ell)})$ for all columns belonging to $\mathcal{U}^{(\ell)}$, where $N_\ell$ is the number of data points in $\mathcal{U}^{(\ell)}$ and $\sum_{\ell=1}^{k} N_\ell = N$. We assume the rank of the $\ell$th subspace $\mathcal{U}^{(\ell)}$ is $r_\ell$ and define $r = \max_\ell r_\ell$. In addition, we use $\mathbf{U}^{(\ell)} \in \mathbb{R}^{d\times r_\ell}$ to represent an orthonormal basis of $\mathcal{U}^{(\ell)}$. The observed matrix is denoted by $\mathbf{X} \in \mathbb{R}^{d\times N}$. Under the noiseless setting we have $\mathbf{X} = \mathbf{Y}$; for the noisy setting we have $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$ where $\mathbf{Z} \in \mathbb{R}^{d\times N}$ is a noise matrix which can be either deterministic or stochastic.

We use "$-i$" to denote all except the $i$th column in a data matrix. For example, $\mathbf{Y}_{-i} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i+1}, \cdots, \boldsymbol{y}_N)$ and $\mathbf{Y}_{-i}^{(\ell)} = (\boldsymbol{y}_1^{(\ell)}, \cdots, \boldsymbol{y}_{i-1}^{(\ell)}, \boldsymbol{y}_{i+1}^{(\ell)}, \cdots, \boldsymbol{y}_{N_\ell}^{(\ell)})$. For any matrix $\mathbf{A}$, let $\mathcal{Q}(\mathbf{A}) = \mathrm{conv}(\pm\boldsymbol{a}_1, \cdots, \pm\boldsymbol{a}_N)$ denote the symmetric convex hull spanned by all columns in $\mathbf{A}$. For any subspace $\mathcal{U}$ and vector $\boldsymbol{v}$, denote $\mathcal{P}_{\mathcal{U}}\boldsymbol{v} = \mathrm{argmin}_{\boldsymbol{u}\in\mathcal{U}}\|\boldsymbol{u} - \boldsymbol{v}\|$ as the projection of $\boldsymbol{v}$ onto $\mathcal{U}$.

[1] In semi/fully random models the underlying subspaces and/or data points are distributed uniformly at random. Detailed definitions can be found in (Soltanolkotabi et al., 2012).

**Methods** The first step is to perform dimensionality reduction on the observation matrix $\mathbf{X}$. More specifically, for a target projection dimension $p < d$, the projected observation matrix $\widetilde{\mathbf{X}}' \in \mathbb{R}^{p\times N}$ is obtained by first computing $\widetilde{\mathbf{X}} = \mathbf{\Psi}\mathbf{X}$ for some random projection matrix $\mathbf{\Psi} \in \mathbb{R}^{p\times d}$ and then normalizing it so that each column in $\widetilde{\mathbf{X}}'$ has unit two norm. Afterwards, Lasso self-regression as formulated in Eq. (1.2) is performed for each column in $\widetilde{\mathbf{X}}'$ to obtain the similarity matrix $\mathbf{C} = \{\boldsymbol{c}_i\}_{i=1}^N \in \mathbb{R}^{N\times N}$. Spectral clustering is then be applied to $\mathbf{C}$ to obtain an explicit clustering of $\mathbf{X}$. In this paper we use the normalized-cut algorithm (Shi & Malik, 2000) for spectral clustering.

**Evaluation measures** To evaluate the quality of obtained similarity matrix $\mathbf{C}$, we consider the *Lasso subspace detection property* defined in (Wang & Xu, 2013). More specifically, $\mathbf{C}$ satisfies Subspace Detection Property (SDP) if for each $i \in \{1, \cdots, N\}$ the following holds: 1) $\boldsymbol{c}_i$ is a non-trivial solution. That is, $\boldsymbol{c}_i$ is not a zero vector; 2) if $c_{ij} \neq 0$ then data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same subspace cluster. The second condition alone is referred to as "Self-Expressiveness Property" (SEP) in (Elhamifar & Vidal, 2013). Note that we do *not* require $c_{ij} \neq 0$ for *every* pair of $\boldsymbol{x}_i, \boldsymbol{x}_j$ belonging to the same cluster. We also remark that in general SEP is not necessary for spectral clustering to succeed, cf. (Wang & Xu, 2013) [2].

## 3. Dimension reduction methods

In this section we review several popular dimensionality reduction methods and show that they are *subspace embeddings*. A linear projection $\mathbf{\Psi} \in \mathbb{R}^{p\times d}$ is said to be a subspace embedding if for some $r$-dimesional subspace $\mathcal{L} \subseteq \mathbb{R}^d$ the following holds:

$$\Pr_{\mathbf{\Psi}}\left[\|\mathbf{\Psi}\boldsymbol{x}\| \in (1\pm\epsilon)\|\boldsymbol{x}\|, \forall \boldsymbol{x} \in \mathcal{L}\right] \geq 1 - \delta. \tag{3.1}$$

The following proposition is a simple property of subspace embeddings, which we prove in Appendix A.1.

**Proposition 1.** *Fix $\epsilon, \delta > 0$. Suppose $\mathbf{\Psi}$ is a subspace embedding with respect to $\mathcal{B} = \{\mathrm{span}(\mathcal{U}^{(\ell)} \cup \mathcal{U}^{(\ell')}); \ell, \ell' \in [k]\} \cup \{\boldsymbol{x}_i, \boldsymbol{z}_i; i \in [N]\}$ with parameters $r' = 2r$, $\epsilon' = \epsilon/3$ and $\delta' = 2\log((k+N)/\delta)$. Then with probability $\geq 1-\delta$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{U}^{(\ell)} \cup \mathcal{U}^{(\ell')}$ we have*

$$\left|\langle \boldsymbol{x}, \boldsymbol{y}\rangle - \langle \mathbf{\Psi}\boldsymbol{x}, \mathbf{\Psi}\boldsymbol{y}\rangle\right| \leq \epsilon\left(\frac{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2}{2}\right); \tag{3.2}$$

*furthermore, for all $\boldsymbol{x} \in \{\boldsymbol{x}_1, \boldsymbol{z}_1, \cdots, \boldsymbol{x}_N, \boldsymbol{z}_N\}$ the following holds:*

$$(1-\epsilon)\|\boldsymbol{x}\|_2^2 \leq \|\mathbf{\Psi}\boldsymbol{x}\|_2^2 \leq (1+\epsilon)\|\boldsymbol{x}\|_2^2. \tag{3.3}$$

[2] It is almost sufficient for perfect clustering both in practice (Elhamifar & Vidal, 2013) and in theory (Wang et al., 2015).

### 3.1. Random Gaussian projection

In a random Gaussian projection matrix $\mathbf{\Psi}$ each entry $\mathbf{\Psi}_{ij}$ is generated from i.i.d. Gaussian distributions $\mathcal{N}(0, 1/\sqrt{p})$, where $p$ is the target dimension after projection. Using standard Gaussian tail bounds and Johnson-Lindenstrauss argument we have the following proposition, which is proved in Appendix A.1.

**Proposition 2.** *Gaussian random matrices $\mathbf{\Psi} \in \mathbb{R}^{p\times d}$ is a subspace embedding with respect to $\mathcal{B}$ if*

$$p \geq 2\epsilon^{-2}(r+\log(2k^2/\delta)+\sqrt{4r\log(2k^2/\delta)}+12\log(4N/\delta)). \tag{3.4}$$

### 3.2. Uniform row sampling

For uniform row sampling each row in the observed data matrix $\mathbf{X}$ is sampled independently at random so that the resulting matrix has $p$ non-zero rows. Formally speaking, each row of the projection matrix $\mathbf{\Omega}$ is sampled i.i.d. from the distribution $\Pr\left[\mathbf{\Omega}_{i\cdot} = \sqrt{\frac{d}{p}}\boldsymbol{e}_j\right] = \frac{1}{d}$, where $i \in [p]$, $j \in [d]$ and $\boldsymbol{e}_j$ is a $d$-dimensional indicator vector with only the $j$th entry not zero.

For uniform row sampling to work, both the observation matrix $\mathbf{X}$ and the column space of the uncorrupted data matrix $\mathbf{Y}$ should satisfy certain incoherence conditions. In this paper, we apply the following two types of incoherence/spikiness definitions, which are widely used in the low rank matrix completion literature (Recht, 2011; Balzano et al., 2010; Krishnamurthy & Singh, 2014).

**Definition 3** (Column space incoherence)**.** *Suppose $\mathcal{U}$ is the column space of some matrix and $\operatorname{rank}(\mathcal{U}) = r$. Let $\mathbf{U} \in \mathbb{R}^{d\times r}$ be an orthonormal basis of $\mathcal{U}$. The incoherence of $\mathcal{U}$ is defined as*

$$\mu(\mathcal{U}) := \frac{d}{r}\max_{i=1,\cdots,d}\|\mathbf{U}_{(i)}\|_2^2, \tag{3.5}$$

*where $\mathbf{U}_{(i)}$ indicates the $i$th row of $\mathbf{U}$.*

**Definition 4** (Column spikiness)**.** *For a vector $\boldsymbol{x} \in \mathbb{R}^d$, the spikiness of $\boldsymbol{x}$ is defined as*

$$\mu(\boldsymbol{x}) := d\|\boldsymbol{x}\|_\infty^2/\|\boldsymbol{x}\|_2^2, \tag{3.6}$$

*where $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$ denotes the vector infinite norm.*

We have the following proposition for the uniform row sampling operator $\mathbf{\Omega}$, which we prove in Appendix A.1.

**Proposition 5.** *Suppose $\max_{\ell=1}^k \mu(\mathcal{U}^{(\ell)}) \leq \mu_0$ and $\max_{i=1}^N \max(\mu(\boldsymbol{x}_i), \mu(\boldsymbol{z}_i)) \leq \mu_0$ for some constant $\mu_0 > 0$. The uniform sampling operator $\mathbf{\Omega}$ is a subspace embedding with respect to $\mathcal{B}$ if*

$$p \geq 8\epsilon^{-2}\mu_0(r\log(4rk^2/\delta) + \log(8N/\delta)). \tag{3.7}$$

### 3.3. FJLT and sketching

The Fast Johnson-Lindenstrauss Transform (FJLT, (Ailon & Chazelle, 2009)) computes a compressed version of a data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$ using $O(d\log d + p)$ operations per column with high probability. The projection matix $\mathbf{\Phi}$ can be written as $\mathbf{\Phi} = \mathbf{PHD}$, where $\mathbf{P} \in \mathbb{R}^{p\times d}$ is a sparse JL matrix, $\mathbf{H} \in \mathbb{R}^{d\times d}$ is a deterministic Walsh-Hadamard matrix and $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a random diagonal matrix. Details of FJLT can be found in (Ailon & Chazelle, 2009).

Sketching (Charikar et al., 2004; Clarkson & Woodruff, 2013) is another powerful tool for dimensionality reduction on sparse inputs. The sketching operator $\mathbf{S} : \mathbb{R}^d \to \mathbb{R}^p$ is constructed as $\mathbf{S} = \mathbf{\Pi\Sigma}$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{\Sigma}$ is a random sign diagonal matrix. The projected vector $\mathbf{S}\boldsymbol{x}$ can be computed in $O(\operatorname{nnz}(\boldsymbol{x}))$ time, where $\operatorname{nnz}(\boldsymbol{x})$ is the number of nonzero entries in $\boldsymbol{x}$.

The following two propositions show that both FJLT and sketching are subspace embeddings. In fact, they are *oblivious* in the sense that they work for any low-dimensional subspace $\mathcal{L}$.

**Proposition 6.** *(Clarkson & Woodruff, 2013) The FJLT operator $\mathbf{\Phi}$ is an oblivious subspace embedding if $p = \Omega(r/\epsilon^2)$, with $\delta$ considered as a constant.*

**Proposition 7.** *(Avron et al., 2014) The sketching operator $\mathbf{S}$ is an oblivious subspace embedding if $p = \Omega(r^2/(\epsilon^2\delta))$.*

## 4. Main results

We present general geometric separation conditions for Lasso sparse subspace clustering (Eq. (1.2)) to succeed for dimensionality-reduced data in the fully deterministic setting; that is, both subspaces and data points within subspaces are deterministically distributed. In addition, our analysis reveals that SSC is able to robustly detect the correct subspaces with substantially compressed data even when the data points are adversarially perturbed, stochastically contaminated, or subject to formal privacy constraints. These contributions significantly expand the previous provable results on the same subject that works only with noiseless data generated from the "semi-random" model (Heckel et al., 2014).

We begin our analysis with two key concepts introduced in the seminal work of Soltanolkotabi and Candes (Soltanolkotabi et al., 2012): *subspace incoherence* and *inradius*. Subspace incoherence characeterizes how well the subspaces associated with different clusters are separated. It is based on the *dual direction* of the optimization problem in Eq. (1.1) and (1.2), which is defined as follows:

**Definition 8** (Dual direction, (Soltanolkotabi et al., 2012; Wang & Xu, 2013))**.** *Fix a column $\boldsymbol{x}$ of $\mathbf{X}$ belonging to subspace $\mathcal{U}^{(\ell)}$. Its dual direction $\boldsymbol{\nu}(\boldsymbol{x})$ is defined as the*

*solution to the following dual optimization problem:* [3]

$$\max_{\boldsymbol{\nu}\in\mathbb{R}^d} \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \frac{\lambda}{2}\boldsymbol{\nu}^\top\boldsymbol{\nu}, \quad \textit{s.t. } \|\mathbf{X}^\top\boldsymbol{\nu}\|_\infty \le 1. \tag{4.1}$$

*Note that Eq. (4.1) has unique solution when $\lambda > 0$.*

The subspace incoherence for $\mathcal{U}^{(\ell)}$, $\mu_\ell$, is defined in Eq. (4.2). Note that it is not related to the column subspace incoherence defined in Eq. (3.5). The smaller $\mu_\ell$ is the further $\mathcal{U}^{(\ell)}$ is separated from the other subspaces.

**Definition 9** (Subspace incoherence, (Soltanolkotabi et al., 2012; Wang & Xu, 2013))**.** *Subspace incoherence $\mu_\ell$ for subspace $\mathcal{U}^{(\ell)}$ is defined as*

$$\mu_\ell := \max_{\boldsymbol{x}\in\mathbf{X}\setminus\mathbf{X}^{(\ell)}} \|\mathbf{V}^{(\ell)\top}\boldsymbol{x}\|_\infty, \tag{4.2}$$

*where $\mathbf{V}^{(\ell)} = (\boldsymbol{v}(\boldsymbol{x}_1^{(\ell)}), \cdots, \boldsymbol{v}(\boldsymbol{x}_{N_\ell}^{(\ell)}))$ and $\boldsymbol{v}(\boldsymbol{x}) = \mathcal{P}_{\mathcal{U}}\boldsymbol{\nu}(\boldsymbol{x})/\|\mathcal{P}_{\mathcal{U}}\boldsymbol{\nu}(\boldsymbol{x})\|_2$. $\boldsymbol{\nu}(\boldsymbol{x})$ is the dual direction of $\boldsymbol{x}$ defined in Eq. (4.1).*

The concept of inradius characterizes how well data points are distributed within a single subspace. More specifically, we have the following definition:

**Definition 10** (Inradius, (Soltanolkotabi et al., 2012; Wang & Xu, 2013))**.** *For subspace $\mathcal{U}^{(\ell)}$, its inradius $\rho_\ell$ is defined as*

$$\rho_\ell := \min_{i=1,\cdots,N_\ell} r(\mathcal{Q}(\mathbf{Y}_{-i}^{(\ell)})), \tag{4.3}$$

*where $r(\cdot)$ denotes the radius of the largest ball inscribed in a convex body.*

The larger $\rho_\ell$ is the more uniformly data points are distributed in the $\ell$th subspace. Note that unlike subspace incoherence, the inradius is defined in terms of the uncorrupted data $\mathbf{Y}$. We also remark that both $\mu_\ell$ and $\rho_\ell$ are between 0 and 1 because of normalization.

Success condition for exact SSC was proved in (Soltanolkotabi et al., 2012) and was generalized to the noisy case in (Wang & Xu, 2013). Below we cite Theorem 6 and Theorem 8 in (Wang & Xu, 2013) for a success condtition of Lasso SSC. In general, Lasso SSC succeeds when there is a sufficiently large gap between subspace incoherence and inradius. Results are restated below, with minor simplification in our notation.

**Theorem 11** ((Wang & Xu, 2013), Theorem 6 and 8)**.** *Suppose $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$ where $\mathbf{Y}$ is the uncorrupted data matrix and $\mathbf{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_N)$ is a determinisitc noise matrix that satisfies $\max_{i=1}^N \|\boldsymbol{z}_i\|_2 \le \eta$. Define $\rho := \min_\ell \rho_\ell$. If*

$$\eta \le \min_{\ell=1,\cdots,k} \frac{\rho(\rho_\ell - \mu_\ell)}{7\rho_\ell + 2}, \tag{4.4}$$

*then subspace detection property holds for the Lasso SSC algorithm in Eq. (1.2) if the regularization coefficient $\lambda$ is in the range*

$$\max_{\ell=1,\cdots,k} \frac{\eta(1+\eta)(2+\rho_\ell)}{\rho_\ell - \mu_\ell - 2\eta} < \lambda < \rho - 2\eta - \eta^2. \tag{4.5}$$

*In addition, if $\mathbf{Z}_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2/d)$ are independent Gaussian noise with variance $\sigma^2 := \max_{i,j}\sigma_{ij}^2$ satisfying*

$$\sqrt{\frac{\log N}{d}}\sigma(1+\sigma) < C \min_{\ell=1,\cdots,k}\left\{\rho, r^{-1/2}, \rho_\ell - \mu_\ell\right\} \tag{4.6}$$

*for sufficiently small constant $C \ge 1/80$, then with probability at least $1 - \frac{10}{N}$ the subspace detection property holds if $\lambda$ is in the range*

$$\frac{C_1\sigma(1+\sigma)}{\rho_\ell - \mu_\ell}\sqrt{\frac{\log N}{d}} < \lambda < \rho - C_2\sigma(1+\sigma)\sqrt{\frac{\log N}{d}}. \tag{4.7}$$

*Here $C_1 \le 80$ and $C_2 \le 20$ are absolute constants.*

In the remainder of this section we prove general success conditions for Lasso SSC on dimensionality-reduced data. We will first describe the result for the noiseless case and then the results are extended to handle a small amount of adversarial perturbation or a much larger amount of stochastic noise. A performance guarantee under differential privacy can then be stated as a simple corollary of the noisy recovery result. The basic idea common in all of the upcoming results is to show that the subspace incoherence and inradius (therefore the geometric gap) are approximately preserved under dimension reduction.

### 4.1. The noiseless case

We first bound the perturbation of dual directions when the data are noiseless.

**Lemma 12** (Perturbation of dual directions, the noiseless case)**.** *Assume $\lambda < 1/4$. Fix a column $\boldsymbol{x}$ in $\mathbf{X}$ with dual direction $\boldsymbol{\nu} = \boldsymbol{\nu}(\boldsymbol{x})$ and $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{x})$ defined in Eq. (4.1) and (4.2). Let $\widetilde{\mathbf{X}}$ denote the projected data matrix $\boldsymbol{\Psi}\mathbf{X}$ and $\widetilde{\mathbf{X}}'$ denote the normalized version of $\widetilde{\mathbf{X}}$. Suppose $\boldsymbol{\nu}^*$ and $\boldsymbol{v}^*$ are computed using the normalized projected data matrix $\widetilde{\mathbf{X}}'$. If $\boldsymbol{\Psi}$ satisfies Eq. (3.2, 3.3) with parameter $\epsilon$ and $\epsilon < 1/\max(1, \|\boldsymbol{\nu}\|)$ then with probability $\ge 1-\delta$ the following holds for all $\boldsymbol{w} \in \mathbf{X}\setminus\mathbf{X}^{(\ell)}$:*

$$\left|\langle \boldsymbol{v}, \boldsymbol{w}\rangle - \langle \boldsymbol{v}^*, \tilde{\boldsymbol{w}}'\rangle\right| \le 32\sqrt{\epsilon/\lambda} + 2\epsilon. \tag{4.8}$$

As a simple corollary, perturbation of subspace incoherence can then be bounded as in Corollary 13.

**Corollary 13** (Perturbation of subsapce incoherence, the noiseless case)**.** *Assume the same notations in Lemma 12.*

[3] For exact SSC simply set $\lambda = 0$.

*Let $\mu_\ell$ and $\tilde{\mu}_\ell$ be the subspace incoherence of the $\ell$th subspace before and after dimension reduction. Then with probability $1 - N\delta$ the following holds:*

$$\tilde{\mu}_\ell \leq \mu_\ell + 32\sqrt{\epsilon/\lambda} + 2\epsilon, \quad \forall \ell = 1, \cdots, k. \tag{4.9}$$

The following lemma bounds the perturbation of inradius for each subspace.

**Lemma 14** (Perturbation of inradius)**.** *Fix $\ell \in \{1, \cdots, k\}$ and $\delta, \epsilon > 0$. Let $\mathbf{Y} = \mathbf{Y}^{(\ell)} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_{N_\ell}) \subseteq \mathcal{U}^{(\ell)}$ be the noiseless $d \times N_\ell$ matrix with all columns belonging to $\mathcal{U}^{(\ell)}$ with unit two norm. Suppose $\widetilde{\mathbf{Y}} = \mathbf{\Psi}\mathbf{Y} \in \mathbb{R}^{p \times N_\ell}$ is the projected matrix and $\widetilde{\mathbf{Y}}'$ scales every column in $\widetilde{\mathbf{Y}}$ so that they have unit norm. Let $\rho_\ell$ and $\tilde{\rho}_\ell$ be the inradius of subspace $\mathcal{U}^{(\ell)}$ before and after dimensionality reduction, defined on $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}'$ respectively. If $\mathbf{\Psi}$ satisfies Eq. (3.2,3.3) with parameter $\epsilon$ then with probability $\geq 1-\delta$ the following holds:*

$$\tilde{\rho}_\ell \geq \rho_\ell/(1+\epsilon). \tag{4.10}$$

With perturbation bounds on both subspace incoherence and inradius we can easily prove the following main theorem, which gives sufficient success condition for Lasso SSC on dimensionality-reduced noiseless data.

**Theorem 15.** *Suppose $\mathbf{X} \in \mathbb{R}^{d \times N}$ is a noiseless input matrix with subspace incoherence $\{\mu_\ell\}_{\ell=1}^k$ and inradii $\{\rho_\ell\}_{\ell=1}^k$. Assume $\mu_\ell < \rho_\ell$ for all $\ell \in \{1, \cdots, k\}$. Let $\widetilde{\mathbf{X}}'$ be the normalized data matrix after compression. Assume $\lambda < 1/4$ and $\lambda < \rho/2$. If $\mathbf{\Psi}$ satisfies Eq. (3.2,3.3) with parameter $\epsilon$ then Lasso SSC satisfies subspace detection property with probability $\geq 1-\delta$, if $\epsilon$ is upper bounded by*

$$\epsilon \leq \min\left\{\frac{1}{2}, \frac{\Delta}{2(2+\rho)}, c_1\lambda\Delta^2\right\}, \tag{4.11}$$

*where $c_1 > 0$ is some absolute constant and $\Delta = \min_\ell(\rho_\ell - \mu_\ell)$ is the minimum gap between subspace incoherence and inradius for each subspace.*

We make several remarks on Theorem 15. First, an upper bound on $\epsilon$ implies a lower bound on projection dimension $p$, and exact $p$ values vary for different data compression schemes. In addition, even for noiseless data the regularization coefficient $\lambda$ cannot be too small if projection error $\epsilon$ is present (recall that $\lambda \to 0$ corresponds to the exact SSC formulation). This is because when $\lambda$ goes to zero the strong convexity of the dual optimization problem decreases. As a result, small perturbation on $\mathbf{X}$ could result in drastic changes of the dual direction and Lemma 12 fails subsequently. On the other hand, as $\lambda$ increases the similarity graph connectivity decreases because the optimal solution to Eq. (1.2) becomes sparser. To guarantee the obtained solution is nontrivial (i.e., at least one nonzero entries in $\boldsymbol{c}_i$), $\lambda$ must not exceed $\rho/2$.

### 4.2. The noisy case

When the input matrix is corrupted with noise, Lemma 14 remains unchanged because the inradius is defined in terms of the noiseless data matrix $\mathbf{Y}$. Therefore, we only need to prove a noisy version of Lemma 12 that bounds the perturbation of dual directions.

**Lemma 16** (Perturbation of dual directions, the noisy case)**.** *Suppose $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$ where $\mathbf{Y}$ is the uncorrupted data matrix and $\mathbf{Z}$ is the noise matrix with $\max_{i=1,\cdots,n} \|\boldsymbol{z}_i\|_2 \leq \eta$. Assume $\lambda < 1/4$. Fix a column $\boldsymbol{x}$ with dual direction $\boldsymbol{\nu}$ and $\boldsymbol{v}$ defined in Eq. (4.1) and (4.2). Suppose $\widetilde{\mathbf{Y}} = \mathbf{\Psi}\mathbf{Y}$ is the projected noiseless data matrix and $\widetilde{\mathbf{Y}}'$ is the normalized version of $\widetilde{\mathbf{Y}}$. Let $\widetilde{\mathbf{X}}' = \widetilde{\mathbf{Y}}' + \widetilde{\mathbf{Z}}$ be the noisy observation after projection, where $\widetilde{\mathbf{Z}} = \mathbf{\Psi}\mathbf{Z}$ is the projected noise. If $\mathbf{\Psi}$ satisfies Eq. (3.2,3.3) with parameter $\epsilon$ and $\epsilon < 1/\max(1, \|\boldsymbol{\nu}\|)$ then with probability $\geq 1 - \delta$ the following holds for all $\boldsymbol{w} \in \mathbf{X} \backslash \mathbf{X}^{(\ell)}$:*

$$\left|\langle \boldsymbol{v}, \boldsymbol{w}\rangle - \langle \boldsymbol{v}^*, \tilde{\boldsymbol{w}}'\rangle\right| \leq 16\sqrt{\frac{5\eta^2}{\rho_\ell} + \frac{8(\epsilon+3\eta)}{\lambda}} + 2\epsilon. \tag{4.12}$$

With Lemma 16 the following corollary on subspace incoherence perturbation immediately follows.

**Corollary 17** (Perturbation of subsapce incoherence, the noisy case)**.** *Assume the conditions as in Lemma 16. Let $\mu_\ell$ and $\tilde{\mu}_\ell$ be the subspace incoherence before and after dimension reduction. Then with probability $\geq 1 - N\delta$,*

$$\tilde{\mu}_\ell \leq \mu_\ell + 16\sqrt{\frac{5\eta^2}{\rho_\ell} + \frac{8(\epsilon+3\eta)}{\lambda}} + 2\epsilon, \quad \forall \ell \in [k]. \tag{4.13}$$

Finally, we have Theorem 18 and Theorem 19 as simple consequences of Corollary 17 and Lemma 14.

**Theorem 18** (Compressed-SSC under Deterministic noise)**.** *Suppose $\mathbf{X} = \mathbf{Y} + \mathbf{Z}$ is a noisy input matrix with subspace incoherence $\{\mu_\ell\}_{\ell=1}^k$ and inradii $\{\rho_\ell\}_{\ell=1}^k$. Assume $\max_i \|\boldsymbol{z}_i\|_2 \leq \eta$ and $\mu_\ell < \rho_\ell$ for all $\ell \in \{1, \cdots, k\}$. Suppose $\widetilde{\mathbf{X}}' = \widetilde{\mathbf{Y}}' + \widetilde{\mathbf{Z}}$ where $\widetilde{\mathbf{Y}}'$ is the normalized uncorrupted data matrix after compression and $\widetilde{\mathbf{Z}} = \mathbf{\Psi}\mathbf{Z}$ is the projected noise matrix. Assume $\eta$ satisfies*

$$\eta \leq \min_{\ell=1,\cdots,k} \frac{\rho(\rho_\ell - \mu_\ell)}{96}. \tag{4.14}$$

*If $\mathbf{\Psi}$ satisfies Eq. (3.2,3.3) with parameter $\epsilon$ and $\lambda = \rho/4$, then Lasso SSC satisfies the subspace detection property with probability $\geq 1 - \delta$. Here $\epsilon$ is upper bounded by*

$$\epsilon \leq \min\left\{\frac{1}{3}, \frac{\Delta}{4(2+\rho)}, \frac{\lambda}{8}\left(c_2\Delta^2 - \frac{5\eta^2}{\rho}\right) - 3\eta\right\}, \tag{4.15}$$

*where $c_2 > 0$ is some absolute constant and $\Delta = \min_\ell(\rho_\ell - \mu_\ell)$ is the minimum gap between subspace incoherence and inradius.*

**Theorem 19** (Compressed-SSC under Gaussian noise)**.** *Define the same positive quantities $\{\mu_\ell\}_{\ell=1}^k, \{\rho_\ell\}_{\ell=1}^k, \rho, \Delta$ and projection matrix $\mathbf{\Psi}$ as in Theorem 18. Assume each column of $\mathbf{Z}$ is sampled from $\mathcal{N}(0, \frac{\sigma^2}{d}\mathbf{I})$. Suppose $\mathbf{\Psi}$ is a linear transform that satisfies Eq. (3.2,3.3) with parameter $\epsilon$, and moreover its spectral norm satisfies $\|\mathbf{\Psi}\| \leq \xi\sqrt{d/p}$ (For Gaussian JL projection $\xi \leq 3$ with high probability). In addition, assume the noise parameter $\sigma$ satisfies*

$$\sqrt{\frac{\log N}{p}}\sigma(1+\sigma) \leq \frac{C}{4\xi^2}\min_{\ell=1,\ldots,k}\left\{\rho, r^{-1/2}, \rho_\ell - \mu_\ell\right\} \tag{4.16}$$

*with the same constant $C$ as in Eq. (4.6). Then Lasso SSC with $\lambda = \rho/4$ satisfies the subspace detection property with probability $\geq 1 - 8/N - \delta$, if $\epsilon$ is upper bounded by*

$$\epsilon \leq \min\left\{\frac{1}{3}, \frac{\Delta}{4(2+\rho)}, \frac{\lambda}{8}\left(c_2\Delta^2 - \frac{45\sigma^2}{\rho}\right) - 9\sigma\right\}. \tag{4.17}$$

*Here $\Delta = \min_\ell(\rho_\ell - \mu_\ell)$ is the minimum gap between subspace incoherence and inradius.*

These results put forward an interesting view of the subspace clustering problem in terms of resource allocation. The critical geometric gap $\Delta$ (called "Margin of Error" in Wang & Xu (2013)) can be viewed as the amount of resource that we have for a problem while preserving the subspace detection property. It can be used to tolerate noise, compress the data matrix, or alleviate the graph connectivity problem of SSC (Wang et al., 2013). For example, if the noise level is high then it will use more of $\Delta$ and as a result we can only compress the data less aggressively, as shown in Eq. (4.15) and (4.17).

### 4.3. Subspace clustering under privacy constraints

Another common motivation to compress the data before data analysis is to protect data privacy. It has been formally shown that random projections (at least with Gaussian random matrices) protect information privacy (Zhou et al., 2009). Stronger privacy protection can be enforced by injecting additional noise to the dimension reduced data (Kenthapadi et al., 2013). Algorithmically, this basically involves adding iid Gaussian noise to the data after we apply a Johnson-Lindenstrauss transform $\mathbf{\Psi}$ of choice to $\mathbf{X}$ and normalize every column. This procedure guarantees differential privacy (Dwork et al., 2006; Dwork, 2006) at the attribute level, which prevents any single entry of the data matrix from being identified "for sure" given the privatized data and arbitrary side information. The amount of noise to add is calibrated according to how "unsure" we need and how "spiky" (Definition 4) each data point can be.

Due to space constraints, we will describe the detailed definition and our technical results on differential privacy preserved subspace clustering in the supplementary document. We show that Lasso-SSC can still achieve exact subspace detection despite differential privacy constraints. To the best of our knowledge, this is the first result of its kind for subspace clustering and it is not possible without dimensionality reduction. So the knife cuts in both sides: dimension-reduction helps in both computational efficiency and privacy protection.

On the other hand, we are not able to generalize the result to an even stronger form of differential privacy that protects each full column in the data matrix. Such privacy requirements make more sense if we consider each column corresponding to an individual. In the supplementary document we present an argument showing that it is impossible to protect differential privacy of this kind if subspace detection property holds with high probability. This calls for a more realistic measure of utility for subspace clustering, for example, percentage of correctly clustered points or closeness of recovered subspaces to the ground truth.

## 5. Proofs

In this section we give proof sketches for the key lemmas. Complete proofs are deferred to Appendix A.

*Proof sketch of Lemma 12.* Let $f: \mathbb{R}^d \to \mathbb{R}, \boldsymbol{\nu} \in \mathbb{R}^d$ and $\tilde{f}: \mathbb{R}^p \to \mathbb{R}, \boldsymbol{\nu}^* \in \mathbb{R}^p$ denote the objective functions and optimal solutions of the dual problem in Eq. (4.1) on the original data and projected data, respectively. Note that for noiseless data, $\boldsymbol{\nu}(\boldsymbol{x})$ lies exactly on the subspace to which $\boldsymbol{x}$ belongs and the same holds after linear projection. Suppose $\tilde{\boldsymbol{\nu}}' \in \mathbb{R}^p$ is a properly shrinked version of $\boldsymbol{\nu}$ after random projection so that $\tilde{\boldsymbol{\nu}}'$ is feasible to the projected dual optimization problem. Since random projection preserves inner products, one can show that with high probability $\tilde{f}(\tilde{\boldsymbol{\nu}}')$ is close to $f(\boldsymbol{\nu})$. On the other hand, $\tilde{f}(\boldsymbol{\nu}^*)$ is close to $f(\bar{\boldsymbol{\nu}}')$ where $\bar{\boldsymbol{\nu}}' \in \mathbb{R}^d$ is some feasible solution to the dual problem on original data, obtained by inversely projecting $\boldsymbol{\nu}^*$ onto the original subspace and properly shrink it so that it is feasible. [4] In general, we have the following:

$$\tilde{f}(\tilde{\boldsymbol{\nu}}') \approx f(\boldsymbol{\nu}) < f(\bar{\boldsymbol{\nu}}') \approx \tilde{f}(\boldsymbol{\nu}^*) < \tilde{f}(\tilde{\boldsymbol{\nu}}'). \tag{5.1}$$

The difference $|\tilde{f}(\boldsymbol{\nu}^*) - \tilde{f}(\tilde{\boldsymbol{\nu}}')|$ can then be upper bounded by applying Eq. (5.1). Consequently, one can bound the dual direction perturbation $\|\boldsymbol{\nu}^* - \tilde{\boldsymbol{\nu}}'\|$ by noting that the dual problem in Eq. (4.1) is strongly convex for both the original data and the projected data. With the upper bound on $\|\boldsymbol{\nu}^* - \tilde{\boldsymbol{\nu}}'\|$ we can easily bound the inner product perturbation $|\langle \boldsymbol{v}, \boldsymbol{w}\rangle - \langle \boldsymbol{v}^*, \tilde{\boldsymbol{w}}'\rangle|$ because $\langle \tilde{\boldsymbol{\nu}}', \tilde{\boldsymbol{w}}'\rangle \approx \langle \boldsymbol{\nu}, \boldsymbol{w}\rangle$ and $\boldsymbol{v}$ is nothing but a normalized version of $\boldsymbol{\nu}$. □

[4] This requires uniform inner product preservation between two low-rank subspaces. Also, there might be multiple $\bar{\boldsymbol{\nu}}$ that correspond to $\boldsymbol{\nu}^*$. Any of them can be taken.

*Proof sketch of Lemma 14.* For notational simpicity redefine $\mathbf{Y} = \mathbf{Y}_{(-i)}$ and $\widetilde{\mathbf{Y}}_{(-i)}$ for some fixed data point $\boldsymbol{x}_i^{(\ell)}$. Let $\mathcal{Q}(\mathbf{Y})$ and $\mathcal{Q}(\widetilde{\mathbf{Y}}')$ denote the convex hull of the original and (normalized) projected data. Suppose $\mathcal{C}, \mathcal{C}'$ are the largest balls inscribed in $\mathcal{Q}(\mathbf{Y})$ and $\mathcal{Q}(\widetilde{\mathbf{Y}}')$. Let $\tilde{\boldsymbol{c}}$ be the point that lies at the intersecion of $\partial\mathcal{C}$ and $\partial\mathcal{Q}(\widetilde{\mathbf{Y}})$. By definition, $\|\tilde{\boldsymbol{c}}\| = r(\mathcal{Q}(\widetilde{\mathbf{Y}}))$. Suppose $\boldsymbol{c}$ lies in the original data space and it corresponds to $\tilde{\boldsymbol{c}}$ after projection (i.e., $\tilde{\boldsymbol{c}} = \boldsymbol{\Psi}\boldsymbol{c}$). It is easy to prove that $\boldsymbol{c}$ does not lie at the interior of $\mathcal{Q}(\mathbf{Y})$ and hence $\|\boldsymbol{c}\|$ is lower bounded by $r(\mathcal{Q}(\mathbf{Y}))$. Subsequently, a lower bound on $\|\boldsymbol{c}\|$ yields a lower bound on $\|\tilde{\boldsymbol{c}}\|$ because a subspace embedding preserves vector norms uniformly on a low-rank subspace. □

*Proof sketch of Lemma 16.* The proof is essentially similar to the one for Lemma 12. The major difference is that under the noisy setting a dual direction $\boldsymbol{\nu}$ no longer falls exactly onto an underlying subspace $\mathcal{U}^{(\ell)}$ and one needs to upper bound the norm of the orthogonal component $\mathcal{P}_{\mathcal{U}^{(\ell)\perp}}\boldsymbol{\nu}$. This can be done using, for example, Eq. (5.16) in (Wang & Xu, 2013), which states that

$$\|\mathcal{P}_{\mathcal{U}^{(\ell)\perp}}\boldsymbol{\nu}\|_2 \leq \lambda\eta\left(1/\rho_\ell + 1\right) \leq 2\lambda\eta/\rho_\ell. \tag{5.2}$$

□

## 6. Related work

Heckel et al. analyzed both SSC and Threshold-based Subspace Clustering (TSC) on projected data (Heckel et al., 2014). The key difference is that the analysis in (Heckel et al., 2014) only applies to noiseless data and is limited to the semi-random model introduced in (Soltanolkotabi et al., 2012), which is arguably less practical. In contrast, our analysis generalizes to fully deterministic settings. It also applies to a broader class of dimensionality reduction methods and can handle data corrupted by noise.

Arpit et al. proposed a novel dimensionality reduction algorithm to preserve independent subspace structures (Arpit et al., 2014). They showed that by using $p = 2k$ one can preserve the independence structure among subspaces. However, their analysis only applies to noiseless and independent subspaces. Furthermore, in our analysis the target dimension $p$ required depends on the intrinsic subspace rank $r$ instead of $k$. Usually $r$ is quite small in practice (Elhamifar & Vidal, 2013; Basri & Jacobs, 2003).

Another relevant line of research is *high-rank matrix completion*. In (Eriksson et al., 2012) the authors proposed a neighborhood selection based algorithm to solve multiple matrix completion problems. Although their method does recover points lying on the same subspace, the completion problem is quite different from subspace clustering as we discuss in Section 8. Furthermore, though their sampling scheme is more practical than ours (does not need sampling entire rows), an exponential number of data points are required. In contrast, in our analysis $N$ only needs to scale polynomially with $r$ if a stochastic model is imposed.

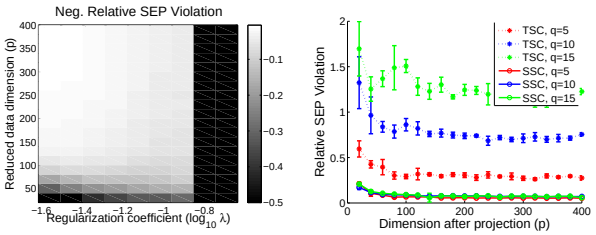


*Figure 1.* Relative SEP violation on extended Yale Face B dataset. Left: Lasso SSC (varying $\lambda, p$); rightmost two columns indicate trivial solutions. White indicates good recovery and black indicates poor recovery. Right: Lasso SSC and TSC (varying $q, p$).

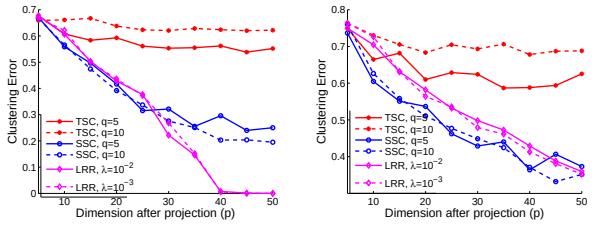


*Figure 2.* Clustering error on the Extended Yale B dataset with five individuals (left) and ten individuals (right).

## 7. Numerical results

In this section we present numerical results that validate our theoretical findings and compare Lasso SSC with TSC (Heckel & Bolcskei, 2013) and LRR (Liu et al., 2013). The Lasso SSC algorithm is implemented using augmented Lagrangian method (ALM) when the regularization coefficient $\lambda$ is fixed and known. We also implement Lasso SSC using a solution path algorithm (Tibshirani & Taylor, 2011) to tune $\lambda$ separately for each data point. The LRR implementation is obtained from (Liu, 2013). Random Gaussian projection is used for all experiments. All algorithms are implemented in Matlab.

We evaluate clustering results by both clustering error and the relative violation of SEP. Clustering error is defined as the percentage of mis-clustered data points up to permutation. The relative violation of SEP characterizes how much the obtained similarity matrix $\mathbf{C}$ violates the self-expressiveness property. It was introduced in (Wang & Xu, 2013) and defined as

$$\text{RelViolation}(\mathbf{C}, \mathcal{M}) = \frac{\sum_{(i,j)\notin\mathcal{M}} |\mathbf{C}|_{ij}}{\sum_{(i,j)\in\mathcal{M}} |\mathbf{C}|_{ij}}, \tag{7.1}$$

where $(i,j) \in \mathcal{M}$ means $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same cluster and vice versa.

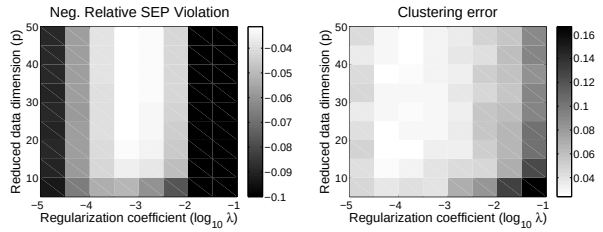


*Figure 3.* Relative SEP violation (left) and clustering error for Lasso SSC on the Hopkins-155 dataset. The rightmost two columns in the left figure indicate trivial solutions. White indicates good similarity graph or clustering and black indicates poor similarity graph or clustering.

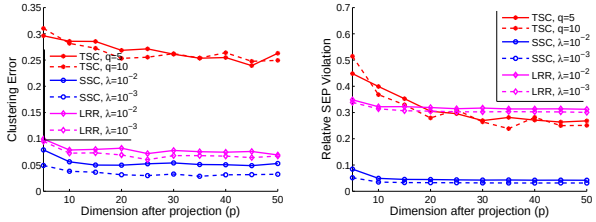


*Figure 4.* Comparison of clustering error (left) and relative SEP violation (right) for Lasso SSC, TSC and LRR on the Hopkins-155 dataset.

### 7.1. Face Clustering

We start by evaluating the performance of Lasso SSC with random Gaussian projection on the extended Yale B face recognition dataset (Lee et al., 2005). We also compare with TSC, which is known to be robust to random projection (Heckel et al., 2014), and LRR. We preprocess the dataset by projecting face images for each individual onto a 9D affine subsapce via PCA. Such preprocess steps were justified in (Basri & Jacobs, 2003) and also adopted in (Wang & Xu, 2013).

In Figure 1 we report the relative SEP violation for both Lasso SSC and TSC. Results are averaged for 10 random projections. We use $q$ to denote the number of solution-path steps taken for each self-regression solution $\boldsymbol{c}_i$. Figure 1 shows that as $\lambda$ decreases the relative SEP violation for Lasso SSC increases, which is predicted by our theoretical analysis. In addition, Figure 1 shows that the relative SEP violation for TSC is rather high compared to Lasso SSC. This is because the analysis for TSC heavily relies upon the semi-random model assumption, which rarely holds true in real-world applications.

Figure 2 shows the clustering accuracy of Lasso SSC, TSC and LRR. For this experiment we randomly selected 5 and 10 individuals from the dataset and report the average clustering error. The total data dimension is $5 \times 9 = 45$ for 5 individuals and $10 \times 9 = 90$ for 10 individuals. We can see that Lasso SSC significantly outperforms TSC under all $p$ and $q$ settings. It outperforms LRR when the projection dimension $p$ is small under which LRR performance guarantee fails because subspaces are no longer independent.

### 7.2. Motion segmentation

We evaluate the performance of Lasso SSC with random projection for motion trajectory segmentation on the Hopkins-155 dataset (Tron & Vidal, 2007). Figure 3 shows the mean relative SEP violation and clustering error for SSC across all 158 video sequences in the dataset. The ambient data dimension ranges from 112 to 240. We can see that the relative SEP violation goes up when $\lambda$ or the projection dimension $p$ decreases. The clustering accuracy acts accordingly, with the exception of very large $\lambda$ values under which we get very sparse self-regression vectors and hence connectivity of the similarity graph is affected.

In Figure 4 we report the clustering error and relative SEP violation for Lasso SSC, TSC and LRR on Hopkins-155. Both clustering error and relative SEP violation are averaged across all 158 sequences. Unlike the face recognition task, we set specific $\lambda$ values instead of solution-path steps ($q$) for Lasso SSC because the former works better on the Hopkins-155 dataset. Figure 4 shows that Lasso SSC outperforms TSC and LRR under various regularization and projection dimension settings, which is consistent with previous experimental results (Elhamifar & Vidal, 2013).

## 8. Discussion

We discuss on the relationship between subspace clustering and high-rank matrix completion. In general, if one can complete a high-rank matrix then exact subspace clustering algorithms can be applied to obtain subspace clusters. On the other hand, once the perfect subspace clustering result is availble we can run separate low-rank matrix completion for each cluster to complete the entire matrix.

However, we remark that under the missing data setting subspace clustering is easier than matrix completion in two ways. First, most matrix completion algorithms require both row and column spaces of a matrix to be incoherent (Recht, 2011), while for subspace clustering we only assume incoherence on the column space. Furthermore, the uniform sampling scheme proposed in Section 3 is a passive sampling scheme because the probability of observing a particular matrix entry is fixed a priori. Although it suffices for the purpose of subspace clustering, it is shown in (Krishnamurthy & Singh, 2014) that any passive sampling scheme fails to complete a column space coherent matrix unless it observes a constant fraction of matrix entries. Adaptive sampling is required to complete a low-rank matrix with coherent column space (Krishnamurthy & Singh, 2014; Chen et al., 2013).

## Acknowledgement

This research is supported in part by grants NSF CAREER IIS-1252412 and AFOSR YIP FA9550-14-1-0285. Yu-Xiang Wang was supported by NSF Award BCS-0941518 to CMU Statistics and Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

## References

Ailon, Nir and Chazelle, Bernard. The fast johnson-lindenstrauss transform and approximate nearest neighbors. *SIAM Journal of Computing*, 39(1):302–322, 2009.

Arpit, Devansh, Nwogu, Ifeoma, and Govindaraju, Venu. Dimensionality reduction with subspace structure preservation. In *NIPS*, 2014.

Avron, Haim, Nguyen, Huy, and Woodruff, David. Subspace embeddings for the polynomial kernel. In *NIPS*, 2014.

Balzano, Laura, Recht, Benjamin, and Nowak, Robert. High-dimensional matched subspace detection when data are missing. In *ISIT*, 2010.

Basri, Ronen and Jacobs, David. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.

Charikar, Moses, Chen, Kevin, and Farach-Colton, Martin. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

Chen, Yudong, Bhojanapalli, Srinadh, Sanghavi, Sujay, and Ward, Rachel. Completing any low-rank matrix, provably. *arXiv:1306.2979*, 2013.

Clarkson, Kenneth and Woodruff, David. Low rank approximation and regression in input sparsity time. In *STOC*, 2013.

Costeira, Joao and Kanade, Takeo. A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, 29(3):159–179, 1998.

Dwork, Cynthia. Differential privacy. In *Automata, languages and programming*, pp. 1–12. Springer, 2006.

Dwork, Cynthia, McSherry, Frank, Nissim, Kobbi, and Smith, Adam. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pp. 265–284. Springer, 2006.

Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2790–2797. IEEE, 2009.

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

Eriksson, Brian, Balzano, Laura, and Nowak, Robert. High rank matrix completion. In *AISTATS*, 2012.

Heckel, Reinhard and Bolcskei, Helmut. Robust subspace clustering via thresholding. *arXiv:1307.4891*, 2013.

Heckel, Reinhard, Tschannen, Michael, and Bolcskei, Helmut. Subspace clustering of dimensionality-reduced data. In *ISIT*, 2014.

Jalali, Ali, Chen, Yudong, Sanghavi, Sujay, and Xu, Huan. Clustering partially observed graphs via convex optimization. In *ICML*, 2011.

Kenthapadi, Krishnaram, Korolova, Aleksandra, Mironov, Ilya, and Mishra, Nina. Privacy via the johnson-lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5(1):39–71, 2013.

Krishnamurthy, Akshay and Singh, Aarti. On the power of adaptivity in matrix completion and approximation. *Arxiv:1407.3619*, 2014.

Lee, Kuang-Chih, Ho, Jeffrey, and Kriegman, David. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684–698, 2005.

Liu, Guangcan. Solving the low-rank representation (LRR) problems. Available online, August 2013. URL https://sites.google.com/site/guangcanliu/.

Liu, Guangcan, Lin, Zhouchen, Shuicheng, Yan, Sun, Ju, Ma, Yi, and Yu, Yong. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (1):171–184, 2013.

Recht, Benjamin. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12: 3413–3430, 2011.

Shi, Jianbo and Malik, Jitendra. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

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

Soltanolkotabi, Mahdi, Elhamifar, Ehsan, Candes, Emmanuel J, et al. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

Tibshirani, Ryan and Taylor, Jonathan. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.

Tron, Roberto and Vidal, Rene. A benchmark for the comparison of 3-D motion segmentation algorithms. In *CVPR*, 2007.

Wang, Yining, Wang, Yu-Xiang, and Singh, Aarti. Clustering consistent sparse subspace clustering. *arXiv:1504.01046*, 2015.

Wang, Yu-Xiang and Xu, Huan. Noisy sparse subspace clustering. *arXiv:1309.1233*, 2013.

Wang, Yu-Xiang, Xu, Huan, and Leng, Chenlei. Provable subspace clustering: When LRR meets SSC. In *NIPS*, 2013.

Zhang, Amy, Fawaz, Nadia, Ioannidis, Stratis, and Montanari, Andrea. Guess who rated this movie: Identifying users through subspace clustering. In *UAI*, 2012.

Zhou, Shuheng, Lafferty, John, and Wasserman, Larry. Compressed and privacy-sensitive sparse regression. *IEEE Transactions on Information Theory*, 55(2):846–866, 2009.

In this document, we provide detailed technical proofs of our main results, as well as the additional results (differentially private subspace clustering), experiments and dicussions that do not fit into the paper due to space constraint.

Appendix A contains proofs for our main results. The proofs are sorted in the order that their corresponding statements appear in the paper. Appendix B formalizes our claims in the paper about attribute privacy and the corresponding utility theorem and includes additional discussions on the difficulty of a stronger user-level privacy claim. Appendix C contains numerical simulations on the performance of compressed SSC under fully random models. Appendix D summarizes a few concentration bounds that we used in the paper.

Lastly, for readers' easy reference, we compile a table of symbols and notations used.

# A. Proofs of the main results

## A.1. Proofs of propositions in Section 3

In this section we prove that a subspace embedding enjoys the property detailed in Proposition 1. We also prove that both random Gaussian projection and uniform row sampling are subspace embeddings with respect to $\mathcal{B} = \{\mathrm{span}(\mathcal{U}^{(\ell)} \cup \mathcal{U}^{(\ell')}); \ell, \ell' \in [k]\} \cup \{\boldsymbol{x}_i, \boldsymbol{z}_i; i \in [N]\}$.

*Proof of Proposition 1.* Fix $\ell, \ell' \in \{1, \cdots, k\}$ and let $\mathcal{U} = \mathrm{span}(\mathcal{U}^{(\ell)} \cup \mathcal{U}^{(\ell')})$ denote the subspace spanned by the union of the two subspaces $\mathcal{U}^{(\ell)}$ and $\mathcal{U}^{(\ell')}$. By assumption, the rank of $\mathcal{U}^{(\ell)} \cup \mathcal{U}^{(\ell')}$, $r'$, satisfies $r' \leq r_\ell + r_{\ell'} \leq 2r$. For any $\boldsymbol{x} \in \mathcal{U}^{(\ell)}$ and $\boldsymbol{y} \in \mathcal{U}^{(\ell')}$ we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{4} \left( \|\boldsymbol{x} + \boldsymbol{y}\|_2^2 - \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right); \tag{A.1}$$

subsequently,

$$\left| \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \langle \boldsymbol{\Psi}\boldsymbol{x}, \boldsymbol{\Psi}\boldsymbol{y} \rangle \right| \leq \frac{1}{4} \left( \left| \|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{\Psi}(\boldsymbol{x} + \boldsymbol{y})\|^2 \right| + \left| \|\boldsymbol{x} - \boldsymbol{y}\|^2 - \|\boldsymbol{\Psi}(\boldsymbol{x} - \boldsymbol{y})\|^2 \right| \right). \tag{A.2}$$

Since $\boldsymbol{\Psi}$ is a subspace embedding, the following holds for all $\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x} + \boldsymbol{y} \in \mathrm{span}(\mathcal{U}^{(\ell)} \cup \mathcal{U}^{(\ell')})$:

$$\begin{aligned} (1-\epsilon)^2 \|\boldsymbol{x} + \boldsymbol{y}\|^2 \leq \|\boldsymbol{\Psi}(\boldsymbol{x} + \boldsymbol{y})\|^2 \leq (1+\epsilon)^2 \|\boldsymbol{x} + \boldsymbol{y}\|^2, \\ (1-\epsilon)^2 \|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq \|\boldsymbol{\Psi}(\boldsymbol{x} - \boldsymbol{y})\|^2 \leq (1+\epsilon)^2 \|\boldsymbol{x} - \boldsymbol{y}\|^2. \end{aligned}$$

The bound for $\left| \langle \boldsymbol{x}, \boldsymbol{y} \rangle - \langle \boldsymbol{\Psi}\boldsymbol{x}, \boldsymbol{\Psi}\boldsymbol{y} \rangle \right|$ then follows by noting that $(1-\epsilon)^2 \geq 1 - 3\epsilon$, $(1+\epsilon)^2 \leq 1 + 3\epsilon$ and $\|\boldsymbol{x} + \boldsymbol{y}\|^2 + \|\boldsymbol{x} - \boldsymbol{y}\|^2 = 2(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2)$. Finally, a union bound over all $k^2$ subspaces and $2N$ data points yields the proposition. □

*Proof of Proposition 2.* Fix $\mathcal{U} \subseteq \mathbb{R}^d$ to be any subspace of dimension at most $r'$ and let $\mathbf{U} \in \mathbb{R}^{d \times r}$ be an orthonormal basis of $\mathcal{U}$. Let $\widetilde{\boldsymbol{\Psi}} = \sqrt{p}\boldsymbol{\Psi}$ denote the unnormalized version of $\boldsymbol{\Psi}$. Since each entry in $\widetilde{\boldsymbol{\Psi}}$ follows i.i.d. standard Gaussian distribution and $\mathbf{U}$ is orthogonal, the projected matrix $\widetilde{\boldsymbol{\Psi}}\mathbf{U} \in \mathbb{R}^{p \times r'}$ follows an entrywise standard Gaussian distribution, too. By Lemma 28 (taking $t = \sqrt{2\delta}$ and scale the matrix by $1/\sqrt{p}$), the singular values of the Gaussian random matrix $\boldsymbol{\Psi}$ obey

$$1 - \sqrt{\frac{r'}{p}} - \sqrt{\frac{2\log(1/\delta)}{p}} \leq \sigma_{r'}(\boldsymbol{\Psi}) \leq \sigma_1(\boldsymbol{\Psi}) \leq 1 + \sqrt{\frac{r'}{p}} + \sqrt{\frac{2\log(1/\delta)}{p}} \tag{A.3}$$

with probability at least $1 - \delta$. Let $\epsilon := \sqrt{\frac{r'}{p}} + \sqrt{\frac{2\log(1/\delta)}{p}}$, then with the same probability, (supposing $\boldsymbol{x} = \mathbf{U}\boldsymbol{\alpha} \in \mathcal{U}$)

$$\begin{aligned} \left| \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{\Psi}\boldsymbol{x}\|_2^2 \right| &= \left| \boldsymbol{\alpha}^\top \mathbf{U}^\top \mathbf{U}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top \mathbf{U}^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \mathbf{U}\boldsymbol{\alpha} \right| \\ &\leq \|\boldsymbol{\alpha}\|_2^2 \|\mathbf{U}^\top \mathbf{U} - \mathbf{U}^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \mathbf{U}\|_2 \\ &= \|\boldsymbol{x}\|_2^2 \|\mathbf{I}_{r' \times r'} - \mathbf{U}^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \mathbf{U}\|_2 \\ &\leq \epsilon \|\boldsymbol{x}\|_2^2. \end{aligned} \tag{A.4}$$

Subsequently,

$$(1-\epsilon)\|\boldsymbol{x}\| \leq \sqrt{1-\epsilon}\|\boldsymbol{x}\| \leq \|\boldsymbol{\Psi}\boldsymbol{x}\| \leq \sqrt{1+\epsilon}\|\boldsymbol{x}\| \leq (1+\epsilon)\|\boldsymbol{x}\|. \tag{A.5}$$

□

*Proof of Proposition 5.* Let $\Omega \subseteq \{1, \cdots, d\}$, $|\Omega| = p$ be the subsampling indices of $\boldsymbol{\Omega}$. By definition, $\Pr[\Omega(j) = i] = 1/d$ for every $i \in \{1, \cdots, d\}$ and $j \in \{1, \cdots, p\}$. Fix any subspace $\mathcal{U} \subseteq \mathbb{R}^d$ of dimension at most $r'$ with incoherence level bounded by $\mu(\mathcal{U}) \leq \mu_0$. Let $\mathbf{U} \in \mathbb{R}^{d \times r'}$ be an orthonormal basis of $\mathcal{U}$. By definition, $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_{r' \times r'}$.

For any $\boldsymbol{x} \in \mathcal{U}$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^{r'}$ such that $\boldsymbol{x} = \mathbf{U}\boldsymbol{\alpha}$. Subsequently, we have

$$\left| \|\boldsymbol{x}\|^2 - \|\boldsymbol{\Omega}\boldsymbol{x}\|^2 \right| = \left| \boldsymbol{\alpha}^\top \boldsymbol{\alpha} - \boldsymbol{\alpha}^\top (\boldsymbol{\Omega}\mathbf{U})^\top (\boldsymbol{\Omega}\mathbf{U}) \boldsymbol{\alpha} \right| \leq \|\boldsymbol{\alpha}\|^2 \cdot \|\mathbf{I} - (\boldsymbol{\Omega}\mathbf{U})^\top (\boldsymbol{\Omega}\mathbf{U})\|. \tag{A.6}$$

Our next objective is to bound the norm $\|\mathbf{I} - (\boldsymbol{\Omega}\mathbf{U})^\top (\boldsymbol{\Omega}\mathbf{U})\|$ with high probability. First let $\mathbf{U}_\Omega := (\boldsymbol{u}_{\Omega(1)}, \cdots, \boldsymbol{u}_{\Omega(p)}) = \sqrt{\frac{p}{d}} (\boldsymbol{\Omega}\mathbf{U})^\top$ be the unnormalized version of subsampled orthogonal operators. By definition we have

$$\|(\boldsymbol{\Omega}\mathbf{U})^\top (\boldsymbol{\Omega}\mathbf{U}) - \mathbf{I}\| = \frac{d}{p} \left\| \mathbf{U}_\Omega \mathbf{U}_\Omega^\top - \frac{p}{d}\mathbf{I} \right\|. \tag{A.7}$$

With Eq. (A.7), we can use noncommutative Matrix Berstein inequality (**?**Recht, 2011) to bound $\|\mathbf{U}_\Omega \mathbf{U}_\Omega^\top - \frac{p}{d}\mathbf{I}\|$ and subsequently obtain an upper bound for the rightmost term in Eq. (A.6). The proof is very similar to the one presented in (Balzano et al., 2010; Krishnamurthy & Singh, 2014), where an upper bound for $\|(\mathbf{U}_\Omega \mathbf{U}_\Omega^\top)^{-1}\|$ is obtained. More specifically, let $\mathbf{B}_1, \cdots, \mathbf{B}_p$ be i.i.d. random matrices such that $\mathbf{B}_j = \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top - \frac{1}{d}\mathbf{I}$. We then have

$$\mathbf{U}_\Omega \mathbf{U}_\Omega^\top - \frac{p}{d}\mathbf{I} = \sum_{j=1}^{p} \mathbf{B}_j \tag{A.8}$$

and furthermore,

$$\mathbb{E}\left[ \mathbf{U}_\Omega \mathbf{U}_\Omega^\top - \frac{p}{d}\mathbf{I} \right] = p \left( \sum_{i=1}^{d} \boldsymbol{u}_i \boldsymbol{u}_i^\top - \mathbf{I} \right) = \mathbf{0}. \tag{A.9}$$

To use Matrix Bernstein, we need to upper bound the range and variance parameters of $\mathbf{B}_j$. Under the matrix incoherence assumption Eq. (3.5) the range of $\mathbf{B}_j$ can be bounded as

$$\|\mathbf{B}_j\| \leq \max_i \left\| \boldsymbol{u}_i \boldsymbol{u}_i^\top - \frac{1}{d}\mathbf{I} \right\| \leq \frac{\sqrt{r'^2}\mu_0}{d} + \frac{1}{d} \leq \frac{r'\mu_0}{d} + \frac{1}{d} \leq \frac{2r'\mu_0}{d} =: R. \tag{A.10}$$

The last inequality is due to the fact that $1 \leq \mu(\mathbf{U}) \leq \frac{d}{r'}$ for any subspace $\mathcal{U}$ of rank $r'$. For the variance, we have

$$\begin{aligned}
\|\mathbb{E}[\mathbf{B}_j^\top \mathbf{B}_j]\| = \|\mathbb{E}[\mathbf{B}_j \mathbf{B}_j^\top]\| &= \left\| \mathbb{E}\left[ \left( \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top - \frac{1}{d}\mathbf{I} \right) \left( \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top - \frac{1}{d}\mathbf{I} \right) \right] \right\| \\
&= \left\| \mathbb{E}\left[ \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top \right] - \frac{1}{d^2}\mathbf{I} \right\| \\
&\leq \left\| \mathbb{E}\left[ \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top \boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top \right] \right\| + \frac{1}{d^2} \\
&\leq \frac{\mu_0 \sqrt{r'^2}}{d^2} \|\mathbb{E}[\boldsymbol{u}_{\Omega(j)} \boldsymbol{u}_{\Omega(j)}^\top]\| + \frac{1}{d^2} \\
&\leq \frac{\mu_0 r'}{d^2} + \frac{1}{d^2} \leq \frac{2\mu_0 r'}{d^2}.
\end{aligned}$$

As a result, we can define $\sigma^2 := 2\mu_0 r'/d^2$ such that $\sigma^2 \geq \max\{\|\mathbb{E}[\mathbf{B}_j \mathbf{B}_j^\top]\|, \|\mathbb{E}[\mathbf{B}_j^\top \mathbf{B}_j]\|\}$ for every $j$. Using Lemma 27, for every $t > 0$ we have

$$\Pr\left[ \|\mathbf{U}_\Omega \mathbf{U}_\Omega^\top - \frac{p}{d}\mathbf{I}\| \geq t \right] \leq 2r \exp\left( -\frac{t^2/2}{\sigma^2 p + Rp/3} \right) = 2r' \exp\left( -\frac{t^2/2}{\frac{2\mu_0 r'}{d^2} p + \frac{2\mu_0 r'}{d} t/3} \right). \tag{A.11}$$

For $\epsilon < 1$ set $t = \frac{p}{d}\epsilon$ and $p = 8\epsilon^{-2}\mu_0 r' \log(2r'/\delta)$. Then with probability $\geq 1 - \delta$ we have

$$\left\| \mathbf{U}_\Omega \mathbf{U}_\Omega^\top - \frac{p}{d}\mathbf{I} \right\| \leq \frac{p}{d}\epsilon. \tag{A.12}$$

The proof is then completed by multiplying both sides in Eq. (A.12) by $\frac{d}{p}$.

□

### A.2. Proof of the main theorems in Section 4

In this section we give rigorous proofs of the three key lemmas in Section 4. We also prove Theorem 15 and 18, which are simple corollaries of Lemma 12, 14 and 16.

*Proof of Lemma 12.* Fix $\ell \in [k]$ and one column $\boldsymbol{x}_i$ in $\mathbf{X}$. Let $\mathcal{U}^{(\ell)}$ and $\widetilde{\mathcal{U}}^{(\ell)}$ denote the low-rank subspaces to which $\boldsymbol{x}_i$ belongs before and after compression. That is, $\widetilde{\mathcal{U}}^{(\ell)} = \{\boldsymbol{\Psi}\boldsymbol{x} : \boldsymbol{x} \in \mathcal{U}^{(\ell)}\}$.

First note that $(1-2\lambda)^2 \leq \|\boldsymbol{\nu}\|^2 \leq 1/(2\lambda)$. $\|\boldsymbol{\nu}\| \geq 1-2\lambda$ because $\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - 2\lambda\|\boldsymbol{\nu}\|^2 \leq \|\boldsymbol{\nu}\|$ and putting $\boldsymbol{\nu} = \boldsymbol{x}$ we obtain a solution with value $1-2\lambda$. On the other hand, $\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - 2\lambda\|\boldsymbol{\nu}\|^2 \leq \|\boldsymbol{\nu}\| - 2\lambda\|\boldsymbol{\nu}\|^2$ and putting $\boldsymbol{\nu} = \mathbf{0}$ we obtain a solution with value 0. Also, under the noiseless setting $\boldsymbol{\nu} \in \mathcal{U}^{(\ell)}$, if $\boldsymbol{x} \in \mathcal{U}^{(\ell)}$.

Define $\tilde{\boldsymbol{\nu}}' = \frac{\sqrt{1-\epsilon}}{1+\epsilon\max(1,\|\boldsymbol{\nu}\|)} \cdot \tilde{\boldsymbol{\nu}}$, where $\tilde{\boldsymbol{\nu}} = \boldsymbol{\Psi}\boldsymbol{\nu}$. Let $f(\boldsymbol{\nu}) = \langle \boldsymbol{\nu}, \boldsymbol{x}\rangle - \frac{\lambda}{2}\|\boldsymbol{\nu}\|_2^2$ and $\tilde{f}(\tilde{\boldsymbol{\nu}}') = \langle \tilde{\boldsymbol{\nu}}', \tilde{\boldsymbol{x}}'\rangle - \frac{\lambda}{2}\|\tilde{\boldsymbol{\nu}}'\|_2^2$ denote the values of the optimization problems. The first step is to prove that $\tilde{\boldsymbol{\nu}}$ is feasible and nearly optimal to the projected optimization problem; that is, $\tilde{f}(\tilde{\boldsymbol{\nu}}')$ is close to $\tilde{f}(\boldsymbol{\nu}^*)$.

We first show that $\tilde{\boldsymbol{\nu}}'$ is a feasible solution with high probability. By Proposition 1, the following bound on $|\tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{\nu}}|$ holds:

$$|\tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{\nu}}| \leq |\boldsymbol{x}_i, \boldsymbol{\nu}| + \epsilon \cdot \frac{\|\boldsymbol{x}_i\| + \|\boldsymbol{\nu}\|}{2} \leq 1 + \epsilon\max(1, \|\boldsymbol{\nu}\|). \quad \forall \boldsymbol{x}_i \in \mathbf{X}. \tag{A.13}$$

Furthermore, with probability $\geq 1-\delta$

$$\|\tilde{\boldsymbol{x}}_i\|_2^2 \geq (1-\epsilon)\|\boldsymbol{x}_i\|_2^2 = 1-\epsilon. \tag{A.14}$$

Consequently, by the definition of $\tilde{\boldsymbol{\nu}}'$ one has

$$\|\widetilde{\mathbf{X}}'^\top \tilde{\boldsymbol{\nu}}'\|_\infty \leq \frac{1}{\sqrt{1-\epsilon}} \cdot \frac{\sqrt{1-\epsilon}}{1+\epsilon\max(1,\|\boldsymbol{\nu}\|)} \|\widetilde{\mathbf{X}}^\top \tilde{\boldsymbol{\nu}}\|_\infty \leq 1. \tag{A.15}$$

Next, we compute a lower bound on $\tilde{f}(\tilde{\boldsymbol{\nu}}')$, which serves as a lower bound for $\tilde{f}(\boldsymbol{\nu}^*)$ because $\boldsymbol{\nu}^*$ is the optimal solution to the dual optimization problem on the projected data.

$$\begin{aligned}
\tilde{f}(\tilde{\boldsymbol{\nu}}') &= \langle \tilde{\boldsymbol{x}}', \tilde{\boldsymbol{\nu}}'\rangle - \frac{\lambda}{2}\|\tilde{\boldsymbol{\nu}}'\|_2^2 \\
&\geq \sqrt{\frac{1-\epsilon}{1+\epsilon}} \frac{\langle \tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\nu}}\rangle}{1+\epsilon\max(1,\|\boldsymbol{\nu}\|)} - \frac{\lambda}{2}(1-\epsilon)\|\tilde{\boldsymbol{\nu}}\|_2^2 \\
&\geq (1-\epsilon)(1-\epsilon\max(1,\|\boldsymbol{\nu}\|))\left(\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \epsilon\max(1,\|\boldsymbol{\nu}\|)\right) - \frac{\lambda}{2}(1-\epsilon)(1+\epsilon)\|\boldsymbol{\nu}\|^2 \\
&\geq \langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \epsilon\max(1,\|\boldsymbol{\nu}\|) - \epsilon\left(\langle \boldsymbol{x}, \boldsymbol{\nu}\rangle - \epsilon\max(1,\|\boldsymbol{\nu}\|)\right) - \frac{\lambda}{2}\|\boldsymbol{\nu}\|^2 \\
&\geq f(\boldsymbol{\nu}) - 2\epsilon\max(1,\|\boldsymbol{\nu}\|).
\end{aligned} \tag{A.16}$$

On the other hand, since $\boldsymbol{\nu}^* \in \widetilde{\mathcal{U}}^{(\ell)}$, there exists $\bar{\boldsymbol{\nu}} \in \mathcal{U}^{(\ell)}$ such that $\boldsymbol{\nu}^* = \boldsymbol{\Psi}\bar{\boldsymbol{\nu}}$. Let $\bar{\boldsymbol{\nu}}'$ be a scaled version of $\bar{\boldsymbol{\nu}}$ so that it is a feasible solution to the optimization problem in Eq. (4.1) before projection. Using essentially similar analysis one can show that $f(\bar{\boldsymbol{\nu}}') \geq \tilde{f}(\boldsymbol{\nu}^*) - 2\epsilon\max(1,\|\boldsymbol{\nu}^*\|)$. Consequently, the following bound on the gap between $\tilde{f}(\tilde{\boldsymbol{\nu}}')$ and $\tilde{f}(\boldsymbol{\nu}^*)$ holds:

$$\left|\tilde{f}(\tilde{\boldsymbol{\nu}}') - \tilde{f}(\boldsymbol{\nu}^*)\right| \leq 4\epsilon\max(1,\|\boldsymbol{\nu}\|,\|\boldsymbol{\nu}^*\|). \tag{A.17}$$

Because the dual problem in Eq. (4.1) is strongly convex with parameter $\lambda$ (this holds for both the projected and the original problem), we can bound the perturbation of dual directions $\|\tilde{\boldsymbol{\nu}}' - \boldsymbol{\nu}^*\|$ by the bounds on their values $|\tilde{f}(\tilde{\boldsymbol{\nu}}') - \tilde{f}(\boldsymbol{\nu}^*)|$ as

$$\|\tilde{\boldsymbol{\nu}}' - \boldsymbol{\nu}^*\|_2 \leq \sqrt{\frac{2|\tilde{f}(\tilde{\boldsymbol{\nu}}') - \tilde{f}(\boldsymbol{\nu}^*)|}{\lambda}} \leq \sqrt{\frac{8\epsilon\max(1,\|\boldsymbol{\nu}\|,\|\boldsymbol{\nu}^*\|)}{\lambda}}. \tag{A.18}$$

Next, note that $\tilde{\boldsymbol{\nu}}', \boldsymbol{\nu}^* \in \widetilde{\mathcal{U}}^{(\ell)}$. Also note that for any two vector $\boldsymbol{a}, \boldsymbol{b}$ the following holds:

$$\left\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right\| = \left\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{a}\|} + \frac{\boldsymbol{b}}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right\|$$

$$\begin{aligned} &\leq \frac{\|\boldsymbol{a}-\boldsymbol{b}\|}{\|\boldsymbol{a}\|}+\frac{\|\boldsymbol{b}\|\cdot\big|\|\boldsymbol{a}\|-\|\boldsymbol{b}\|\big|}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \\ &\leq \frac{\|\boldsymbol{a}-\boldsymbol{b}\|}{\|\boldsymbol{a}\|}+\frac{\|\boldsymbol{a}-\boldsymbol{b}\|}{\|\boldsymbol{a}\|} \\ &= \frac{2\|\boldsymbol{a}-\boldsymbol{b}\|}{\|\boldsymbol{a}\|}. \end{aligned}$$

By symmetry we also have $\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}-\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\|\leq\frac{2\|\boldsymbol{a}-\boldsymbol{b}\|}{\|\boldsymbol{b}\|}$. Therefore,

$$\left\|\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|}-\frac{\boldsymbol{b}}{\|\boldsymbol{b}\|}\right\|\leq\frac{2\|\boldsymbol{a}-\boldsymbol{b}\|}{\max(\|\boldsymbol{a}\|,\|\boldsymbol{b}\|)}. \tag{A.19}$$

Now we can bound $\|\tilde{\boldsymbol{v}}'-\boldsymbol{v}^*\|$ as follows:

$$\begin{aligned} \|\tilde{\boldsymbol{v}}'-\boldsymbol{v}^*\| &= \left\|\frac{\tilde{\boldsymbol{\nu}}'}{\|\tilde{\boldsymbol{\nu}}'\|}-\frac{\boldsymbol{\nu}^*}{\|\boldsymbol{\nu}^*\|}\right\| \\ &\leq \frac{2\|\tilde{\boldsymbol{\nu}}'-\boldsymbol{\nu}^*\|}{\max(\|\tilde{\boldsymbol{\nu}}'\|,\|\boldsymbol{\nu}^*\|)}\leq\frac{2\|\tilde{\boldsymbol{\nu}}'-\boldsymbol{\nu}^*\|}{\max(\|\boldsymbol{\nu}\|/4,\|\boldsymbol{\nu}^*\|)} \\ &\leq \frac{16\sqrt{2\epsilon\max(1,\|\boldsymbol{\nu}\|,\|\boldsymbol{\nu}^*\|)}}{\sqrt{\lambda}\max(\|\boldsymbol{\nu}\|,\|\boldsymbol{\nu}^*\|)}\leq 16\sqrt{\frac{2\epsilon}{\lambda}\max\left(1,\frac{1}{\|\boldsymbol{\nu}\|},\frac{1}{\|\boldsymbol{\nu}^*\|}\right)} \\ &\leq 16\sqrt{\frac{2\epsilon}{\lambda(1-2\lambda)}}\leq 32\sqrt{\frac{\epsilon}{\lambda}}. \end{aligned}$$

Note that after normalization $\tilde{\boldsymbol{v}}'$ is exactly the same with $\tilde{\boldsymbol{v}}$. Subsequently, for any $\boldsymbol{y}\in\mathbf{X}\backslash\mathbf{X}^{(\ell)}$ we have

$$\begin{aligned} \big|\langle\boldsymbol{v},\boldsymbol{y}\rangle-\langle\boldsymbol{v}^*,\tilde{\boldsymbol{y}}'\rangle\big| &\leq \big|\langle\tilde{\boldsymbol{v}}',\tilde{\boldsymbol{y}}'\rangle-\langle\boldsymbol{v}^*,\tilde{\boldsymbol{y}}'\rangle\big|+\big|\langle\tilde{\boldsymbol{v}}',\tilde{\boldsymbol{y}}'\rangle-\langle\boldsymbol{v},\boldsymbol{y}\rangle\big| \\ &\leq \|\tilde{\boldsymbol{v}}'-\boldsymbol{v}^*\|\|\tilde{\boldsymbol{y}}'\|+\big|\langle\tilde{\boldsymbol{v}},\tilde{\boldsymbol{y}}'\rangle-\langle\boldsymbol{v},\boldsymbol{y}\rangle\big| \\ &\leq \|\tilde{\boldsymbol{v}}'-\boldsymbol{v}^*\|+\left|\frac{1}{\|\boldsymbol{\Psi}\boldsymbol{v}\|\|\boldsymbol{\Psi}\boldsymbol{y}\|}\langle\boldsymbol{\Psi}\boldsymbol{v},\boldsymbol{\Psi}\boldsymbol{y}\rangle-\langle\boldsymbol{v},\boldsymbol{y}\rangle\right| \\ &\leq 32\sqrt{\frac{\epsilon}{\lambda}}+\left(1-\frac{1}{\|\boldsymbol{\Psi}\boldsymbol{v}\|\|\boldsymbol{\Psi}\boldsymbol{y}\|}\right)\|\boldsymbol{\Psi}\boldsymbol{v}\|\|\boldsymbol{\Psi}\boldsymbol{y}\|+\big|\langle\boldsymbol{\Psi}\boldsymbol{v},\boldsymbol{\Psi}\boldsymbol{y}\rangle-\langle\boldsymbol{v},\boldsymbol{y}\rangle\big| \\ &\leq 32\sqrt{\frac{\epsilon}{\lambda}}+\left(1-\frac{1}{1+\epsilon}\right)(1+\epsilon)+\epsilon \\ &= 32\sqrt{\frac{\epsilon}{\lambda}}+2\epsilon. \end{aligned}$$

□

*Proof of Lemma 14.* For notational simplicity re-define $\mathbf{Y}=\mathbf{Y}_{(-i)}$ and $\widetilde{\mathbf{Y}}'=\widetilde{\mathbf{Y}}'_{(-i)}$ for some fixed data point $\boldsymbol{x}_i^{(\ell)}$. Let $\mathcal{C}$, $\widetilde{\mathcal{C}}$ be the largest Euclidean balls inscribed in $\mathcal{Q}(\mathbf{Y})$ and $\mathcal{Q}(\widetilde{\mathbf{Y}})$. Since both $\mathcal{Q}(\mathbf{Y})$ and $\mathcal{Q}(\widetilde{\mathbf{Y}})$ are symmetric convex bodies with respect to the origin, the centers of $\mathcal{C}$ and $\widetilde{\mathcal{C}}$ are the origin. Let $\tilde{\boldsymbol{c}}$ be any point in $\widetilde{\mathcal{C}}\cap\partial\mathcal{Q}(\widetilde{\mathbf{Y}})$. By definition, $r(\mathcal{Q}(\widetilde{\mathbf{Y}}))=\|\tilde{\boldsymbol{c}}\|$. Since $\tilde{\boldsymbol{c}}\in\widetilde{\mathcal{U}}^{(\ell)}$, we can find $\boldsymbol{c}\in\mathcal{U}^{(\ell)}$ such that $\tilde{\boldsymbol{c}}=\boldsymbol{\Psi}\boldsymbol{c}$. By Proposition 1, we have (with probability $\geq 1-\delta$)

$$\|\tilde{\boldsymbol{c}}\|\geq\frac{1}{\sqrt{1+\epsilon}}\|\boldsymbol{c}\|. \tag{A.20}$$

On the other hand, $\boldsymbol{c}$ is not contained in the interior of $\mathcal{Q}(\mathbf{Y})$. Otherwise, we can find a scalar $a>1$ such that $a\boldsymbol{c}\in\mathcal{Q}(\mathbf{Y})$ and hence $a\tilde{\boldsymbol{c}}\in\mathcal{Q}(\widetilde{\mathbf{Y}})$, contradicting the fact that $\tilde{\boldsymbol{c}}\in\partial\mathcal{Q}(\widetilde{\mathbf{Y}})$. Consequently, we have $\|\boldsymbol{c}\|\geq r(\mathcal{Q}(\mathbf{Y}))$ by definition. Therefore,

$$r(\mathcal{Q}(\widetilde{\mathbf{Y}}))=\|\tilde{\boldsymbol{c}}\|\geq\frac{1}{\sqrt{1+\epsilon}}\|\boldsymbol{c}\|\geq\frac{r(\mathcal{Q}(\mathbf{Y}))}{\sqrt{1+\epsilon}}. \tag{A.21}$$

Next, we need to lower bound $r(\mathcal{Q}(\widetilde{\mathbf{Y}}'))$ in terms of $r(\mathcal{Q}(\widetilde{\mathbf{Y}}))$. This can be easily done by noting that the maximum column norm in $\widetilde{\mathbf{Y}}$ is upper bounded by $\sqrt{1+\epsilon}$. Consequently, we have

$$r(\mathcal{Q}(\widetilde{\mathbf{Y}}')) \geq r\left(\mathcal{Q}\left(\frac{1}{\sqrt{1+\epsilon}}\widetilde{\mathbf{Y}}\right)\right) \geq \frac{r(\mathcal{Q}(\mathbf{Y}))}{1+\epsilon}. \tag{A.22}$$

□

*Proof of Lemma 16.* Fix $\ell \in \{1, 2, \cdots, k\}$ and a particular column $\boldsymbol{x} = \boldsymbol{x}_i$. Suppose $\boldsymbol{\nu}$ is the optimal solution to the original dual problem in Eq. (4.1). Define $\boldsymbol{\nu}_{\|} = \mathcal{P}_{\mathcal{U}^{(\ell)}}\boldsymbol{\nu}$ and $\boldsymbol{\nu}_{\perp} = \mathcal{P}_{\mathcal{U}^{(\ell)\perp}}\boldsymbol{\nu}$. Let $f(\cdot)$ be the objective value of the dual problem under a specific solution. Then it is easy to observe that

$$f(\boldsymbol{\nu}_{\|}) \geq f(\boldsymbol{\nu}) - \langle \boldsymbol{x}_{\perp}, \boldsymbol{\nu}_{\perp}\rangle \geq f(\boldsymbol{\nu}) - \eta\|\boldsymbol{\nu}_{\perp}\|_2. \tag{A.23}$$

We then cite the following upper bound for $\|\boldsymbol{\nu}_{\perp}\|$, which appears as Eq. (5.16) in (Wang & Xu, 2013).

$$\|\boldsymbol{\nu}_{\perp}\|_2 \leq \lambda\eta\left(\frac{1}{r(\mathcal{Q}(\mathbf{Y}_{-i}^{(\ell)}))} + 1\right) \leq \frac{2\lambda\eta}{\rho_{\ell}}. \tag{A.24}$$

Let $\tilde{\boldsymbol{\nu}} = \boldsymbol{\Psi}\boldsymbol{\nu}_{\|}$ and $\tilde{\boldsymbol{\nu}}' = \frac{\sqrt{1-\epsilon}}{1+(\eta+\epsilon)\max(1,\|\boldsymbol{\nu}\|)} \cdot \tilde{\boldsymbol{\nu}}$. It is easy to verify that $\tilde{\boldsymbol{\nu}}'$ is a feasible solution to the projected dual problem. Define $\eta' := \max_{i=1,\cdots,n}\|\tilde{\boldsymbol{z}}_i\|_2$. Since $\boldsymbol{\Psi}$ is well behaved, $\eta' \leq \sqrt{1+\epsilon}\eta$ with high probability. Applying essentially the same chain of argument as in the proof of Lemma 12 we obtain

$$\begin{aligned}
\tilde{f}(\tilde{\boldsymbol{\nu}}') &= \langle \tilde{\boldsymbol{x}}', \tilde{\boldsymbol{\nu}}'\rangle - \frac{\lambda}{2}\|\tilde{\boldsymbol{\nu}}'\|_2^2 \\
&= \langle \tilde{\boldsymbol{y}}', \tilde{\boldsymbol{\nu}}'\rangle + \langle \tilde{\boldsymbol{z}}, \tilde{\boldsymbol{\nu}}'\rangle - \frac{\lambda}{2}\|\tilde{\boldsymbol{\nu}}'\|_2^2 \\
&\geq \langle \boldsymbol{y}, \boldsymbol{\nu}_{\|}\rangle - \frac{\lambda}{2}\|\boldsymbol{\nu}_{\|}\|^2 - 2(\epsilon+\eta)\max(1,\|\boldsymbol{\nu}\|) - \|\tilde{\boldsymbol{z}}\|_2\|\tilde{\boldsymbol{\nu}}'\|_2 \\
&\geq \langle \boldsymbol{y}, \boldsymbol{\nu}_{\|}\rangle - \frac{\lambda}{2}\|\boldsymbol{\nu}_{\|}\|^2 - 2(\epsilon+\eta)\max(1,\|\boldsymbol{\nu}\|) - \eta' \cdot \frac{\sqrt{(1-\epsilon)(1+\epsilon)}}{1+\epsilon\max(1,\|\boldsymbol{\nu}\|)}\|\boldsymbol{\nu}\| \\
&\geq \langle \boldsymbol{y}, \boldsymbol{\nu}_{\|}\rangle - \frac{\lambda}{2}\|\boldsymbol{\nu}_{\|}\|^2 - 2(\epsilon+\eta)\max(1,\|\boldsymbol{\nu}\|) - \sqrt{1+\epsilon}\eta\|\boldsymbol{\nu}\| \\
&\geq \langle \boldsymbol{x}, \boldsymbol{\nu}_{\|}\rangle - \frac{\lambda}{2}\|\boldsymbol{\nu}_{\|}\|^2 - 2(\epsilon+\eta)\max(1,\|\boldsymbol{\nu}\|) - \sqrt{1+\epsilon}\eta\|\boldsymbol{\nu}\| - \eta\|\boldsymbol{\nu}\| \\
&\geq f(\boldsymbol{\nu}_{\|}) - (2\epsilon+5\eta)\max(1,\|\boldsymbol{\nu}\|) \\
&\geq f(\boldsymbol{\nu}) - \frac{2\lambda\eta^2}{\rho_{\ell}} - (2\epsilon+5\eta)\max(1,\|\boldsymbol{\nu}\|).
\end{aligned}$$

Similarly, one can show that

$$\tilde{f}(\boldsymbol{\nu}^*) \leq f(\boldsymbol{\nu}) + \frac{2\lambda\eta'^2}{\rho_{\ell}} + (2\epsilon+5\eta')\max(1,\|\boldsymbol{\nu}^*\|) \leq f(\boldsymbol{\nu}) + \frac{3\lambda\eta^2}{\rho_{\ell}} + (2\epsilon+6\eta)\max(1,\|\boldsymbol{\nu}^*\|). \tag{A.25}$$

Consequently, noting that $\tilde{f}(\tilde{\boldsymbol{\nu}}') \leq \tilde{f}(\boldsymbol{\nu}^*)$ one has

$$\left|\tilde{f}(\boldsymbol{\nu}^*) - \tilde{f}(\tilde{\boldsymbol{\nu}}')\right| \leq \frac{5\lambda\eta^2}{\rho_{\ell}} + 4(\epsilon+3\eta)\max(1,\|\boldsymbol{\nu}\|,\|\boldsymbol{\nu}^*\|). \tag{A.26}$$

Since both dual problems (before and after projection) are strongly convex with parameter $\lambda$, the following perturbation bound on $\|\boldsymbol{\nu}^* - \tilde{\boldsymbol{\nu}}'\|$ holds:

$$\|\boldsymbol{\nu}^* - \tilde{\boldsymbol{\nu}}'\| \leq \sqrt{\frac{2|\tilde{f}(\boldsymbol{\nu}^*) - \tilde{f}(\tilde{\boldsymbol{\nu}}')|}{\lambda}} \leq \sqrt{\frac{5\eta^2}{\rho_{\ell}} + \frac{8(\epsilon+3\eta)\max(1,\|\boldsymbol{\nu}\|,\|\boldsymbol{\nu}^*\|)}{\lambda}}. \tag{A.27}$$

Subsequently,

$$\begin{aligned}\|\tilde{\boldsymbol{v}}' - \boldsymbol{v}^*\| &\leq \frac{8\|\tilde{\boldsymbol{\nu}}' - \boldsymbol{\nu}^*\|}{\max(\|\boldsymbol{\nu}\|, \|\boldsymbol{\nu}^*\|)} \\ &\leq 8\sqrt{\frac{5\eta^2}{\rho_\ell \max(\|\boldsymbol{\nu}\|^2, \|\boldsymbol{\nu}^*\|^2)} + \frac{8(\epsilon+3\eta)}{\lambda \max(1, \|\boldsymbol{\nu}\|^2, \|\boldsymbol{\nu}^*\|^2)}} \\ &\leq 8\sqrt{\frac{5\eta^2}{\rho_\ell(1-2\lambda)^2} + \frac{8(\epsilon+3\eta)}{\lambda(1-2\lambda)^2}} \\ &\leq 16\sqrt{\frac{5\eta^2}{\rho_\ell} + \frac{8(\epsilon+3\eta)}{\lambda}}.\end{aligned}$$

Finally, the perturbation of the angle between $\boldsymbol{v}$ and $\boldsymbol{y}$ can be bounded by

$$\left|\langle \boldsymbol{v}, \boldsymbol{y}\rangle - \langle \boldsymbol{v}^*, \tilde{\boldsymbol{y}}'\rangle\right| \leq \|\tilde{\boldsymbol{v}}' - \tilde{\boldsymbol{v}}^*\| + 2\epsilon \leq 16\sqrt{\frac{5\eta^2}{\rho_\ell} + \frac{8(\epsilon+3\eta)}{\lambda}} + 2\epsilon. \tag{A.28}$$

□

*Proof of Theorem 15.* Let $\tilde{\mu}_\ell, \tilde{\rho}_\ell$ denote the subspace incoherence and inradius of subspace $\mathcal{U}^{(\ell)}$ after dimensionality reduction. Theorem 11 shows that Lasso SSC satisfies the subspace detection property if $\tilde{\mu}_\ell < \tilde{\rho}_\ell$ for every $\ell$ and $\lambda < \tilde{\rho}$. By Lemma 14, $\tilde{\rho} \geq \rho/2$ with high probability. Note also that $\tilde{\rho}_\ell \geq \frac{\rho_\ell}{1+\epsilon} \geq \rho_\ell(1-\epsilon)$. Subsequently, the following inequality yields $\tilde{\mu}_\ell < \tilde{\rho}_\ell$ for every $\ell$:

$$\mu_\ell + 32\sqrt{\epsilon/\lambda} + (2+\rho_\ell)\epsilon < \rho_\ell, \quad \forall \ell = 1, \cdots, k. \tag{A.29}$$

Taking $32\sqrt{\epsilon/\lambda} < \Delta/2$ and $(2+\rho_\ell)\epsilon < \Delta/2$ where $\Delta = \min_\ell(\rho_\ell - \mu_\ell)$, Eq. (A.29) is subsequently satisfied. This yields

$$\epsilon < \min\left\{\frac{\Delta}{2(2+\rho)}, c_1\lambda\Delta^2\right\} \tag{A.30}$$

for some absolute constant $c_1$. The $\epsilon < 1/2$ term comes from the $\epsilon < 1/\|\boldsymbol{\nu}\|$ condition in Lemma 12. □

*Proof of Theorem 18.* Define $\tilde{\Delta} := \min_\ell(\tilde{\rho}_\ell - \tilde{\mu}_\ell)$ to be the maximum margin of error after dimensionality reduction. First we prove that with $\lambda = \rho/4 < 1/4$ and the upper bound in Eq. (4.15) we have $\tilde{\Delta} \geq \Delta/2$. Essentially, this requires

$$16\sqrt{\frac{5\eta^2}{\rho_\ell} + \frac{8(\epsilon+3\eta)}{\lambda}} < \frac{\Delta}{4}, \tag{A.31}$$

$$2\epsilon + \rho\epsilon < \frac{\Delta}{4}. \tag{A.32}$$

This amounts to

$$\epsilon < \min\left\{\frac{\Delta}{4(2+\rho)}, \frac{\lambda}{8}\left(c_2\Delta^2 - \frac{5\eta^2}{\rho}\right) - 3\eta\right\}, \tag{A.33}$$

where $c_2 > 0$ is an absolute constant.

Next we verify that Eq. (4.5) are satisfied after dimensionality reduction. Let $\tilde{\eta}$ denote the noise level after projection, that is, $\max_i\{\|\tilde{\boldsymbol{z}}_i\|\} \leq \tilde{\eta}$. Because $\epsilon < 1/3$, by Proposition 1 $\tilde{\eta} \leq 2\eta$ with high probability. Consequently, $\eta < \frac{\rho}{96}$ in Eq. (4.14) implies ($\tilde{\rho} = \min_\ell \tilde{\rho}_\ell$ and $\tilde{\mu} = \max_\ell \tilde{\mu}_\ell$)

$$\tilde{\rho} - 2\tilde{\eta} - \tilde{\eta}^2 \geq \rho(1-\epsilon) - 6\eta \geq \frac{2\rho}{3} - \frac{6\rho}{18} = \frac{\rho}{3} \geq \frac{\rho}{4} = \lambda. \tag{A.34}$$

Hence the upper bound on $\lambda$ in Eq. (4.5) is satisfied. For the lower bound, note that $\eta \ll 1$, $\tilde{\rho}_\ell < 1$ and hence

$$\frac{\tilde{\eta}(1+\tilde{\eta})(2+\tilde{\rho}_\ell)}{\tilde{\rho}_\ell - \tilde{\mu}_\ell - 2\tilde{\eta}} \leq \frac{6\tilde{\eta}}{\tilde{\Delta}} \leq \frac{12\eta}{\Delta/2} = \frac{24\eta}{\Delta} < \frac{\rho}{4} = \lambda. \tag{A.35}$$

The last inequality is due to Eq. (4.14). □

*Proof of Theorem 19.* Let the JL transform matrix be $\mathbf{\Psi}$. Since it is a linear transformation, $\boldsymbol{z}_i \sim \mathcal{N}(0, \frac{\sigma^2}{d}\mathbf{I})$ implies that $\mathbf{\Psi}\boldsymbol{z}_i \sim \mathcal{N}(0, \frac{\sigma^2}{d}\mathbf{\Psi}\mathbf{\Psi}^\top)$. Using the fact that this algorithm is invariant to arbitrary unitary transformations, we can apply the rotation that diagonalizes the covariance matrix $\frac{\sigma^2}{d}\mathbf{\Psi}\mathbf{\Psi}^\top$ to every column of the projected (and renormalized) data. This decouples the noise matrix $\mathbf{Z}$ such that every coordinate is independent Gaussian. Moreover, the maximum entrywise variance is upper bounded by

$$\max_{ij} \frac{\sigma_{ij}^2}{p} \le \|\mathbf{\Psi}\|^2 \frac{\sigma^2(1+\epsilon)^2}{d} \le \xi^2 \frac{d}{p}\frac{\sigma^2(1+\epsilon)^2}{d} \le \xi^2 \frac{\sigma^2(1+\epsilon)^2}{p} \le 2\xi^2\frac{\sigma^2}{p},$$

where $\epsilon$ is the JL parameter included to acount for the renormalization of the $y$ part. The last inequality holds because $\epsilon > 1/3$ by our assumption.

Applying the same argument as in the proof of Theorem 18 we get $\tilde{\Delta} = \min_\ell (\tilde{\rho}_\ell - \tilde{\mu}_\ell) \ge \Delta/2$ when Eq. (4.17) is satisfied.

The proof is then completed by invoking the second part of Theorem 11 on the compressed problem with the bounded entrywise independent Gausian noise, we get the condition that

$$\sqrt{\frac{\log N}{p}}\sigma(1+\sigma) \le \frac{C}{4\xi^2}\min_{\ell=1,\dots,k}\left\{\rho, r^{-1/2}, \rho_\ell - \mu_\ell\right\}$$

as claimed in (4.6).

Note that for random Gaussian transforms $\mathbf{\Psi}$, by Lemma 28, $\|\mathbf{\Psi}\| \le 3\sqrt{d/p}$ (hence $\xi^2 \le 9$) with high probability. □

## B. Privacy preserved subspace clustering

In this section, we formalize the claims on attribute-level differential privacy and the corresponding utility guarantee in the paper.

**Privacy Claim** In classic statistical privacy literature, transforming data set $\mathbf{X}$ by taking $\tilde{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{\Delta}$ for some random matrix $\mathbf{A}$ and $\mathbf{\Delta}$ is called *matrix masking*. (Zhou et al., 2009) show that random compression allows the mutual information of the output $\tilde{\mathbf{X}}$ and raw data $\mathbf{X}$ to converge to 0 with rate $O(p/d)$ even when $\mathbf{\Delta} = 0$, their result directly applies to our problem. The guarantee suggests that the amount of information in the compressed output $\tilde{\mathbf{X}}$ about the raw data $\mathbf{X}$ goes to 0 as the ambient dimension $d$ gets large.

On the other hand, if $\mathbf{\Delta} \neq \mathbf{0}$ is an iid Gaussian noise matrix, we can protect the $(\varepsilon, \delta)$-differential privacy of every data entry. Such attribute differential privacy notion is defined below.

**Definition 20** (Attribute Differential Privacy)**.** *Suppose $\mathcal{O}$ is the set for all possible outcomes. We say a randomized algorithm $\mathcal{A} : \mathbb{R}^{d\times N} \to \mathcal{O}$ is $(\varepsilon, \delta)$-differential private at attribute level if*

$$\mathbb{P}(\mathcal{A}(\mathbf{X}) \in \mathcal{S}) \le e^\varepsilon \mathbb{P}(\mathcal{A}(\mathbf{X}') \in \mathcal{S}) + \delta$$

*for any measurable outcome $\mathcal{S} \subset \mathcal{O}$, any $\mathbf{X}$ and $\mathbf{X}'$ that differs in only one entry.*

This is a well-studied setting in (Kenthapadi et al., 2013). It is weaker than protecting the privacy of individual users, which remains an open question, but much stronger than the average protection via mutual information. In fact, it forbids any feature of an individual user from being identified "for sure" by an adversary with arbitrary side information.

**Theorem 21.** *Assume the data (and all other users that we need to protect) satisfy column spikiness conditions with parameter $\mu_0$ as in Definition 4. Let $\mathbf{\Psi}$ be a Johnson-Lindenstrauss transform with parameter $\epsilon$. Releasing compressed data $\widetilde{\mathbf{X}}' = \mathrm{Normalize}(\mathbf{\Psi}\mathbf{X}) + \mathcal{N}(0, \sigma^2\mathbf{I}_{p\times d})$ with $\sigma = \frac{1+\epsilon}{1-\epsilon}\sqrt{\frac{32\mu_0\log(1.25/\delta)}{d\varepsilon^2}}$ preserves attribute-level $(\varepsilon, \delta)$-differential privacy.*

The proof involves working out the $\ell_2$-sensitivity of the operator $\mathrm{Normalize}(\mathbf{\Psi}(\cdot))$ in terms of column incoherence $\mu_0$ and apply "Gaussian Mechanism". By the closeness to post-processing property of differential privacy, the subsequent subspace clustering results protects the same level of privacy. Details are given as follows.

*Proof of Theorem 21.* Let $\mathbf{X}$ and $\mathbf{X}'$ differs by only one entry, w.l.o.g, assume it is the $i$th column and $j$th row,

$$\|\mathbf{\Psi}(\mathbf{X}-\mathbf{X}')\|_F = \|\mathbf{\Psi}(\mathbf{X}_i-\mathbf{X}_i')\|_2 \le \|\mathbf{\Psi}\boldsymbol{e}_j\|\|\mathbf{X}_{ji}-\mathbf{X}'_{ji}| \le 2\sqrt{\frac{\mu}{d}}\|\mathbf{\Psi}\boldsymbol{e}_j\|.$$

Now we derive the $\ell_2$-sensitivity of Normalize$(\mathbf{\Psi}(\cdot))$.

$$\begin{aligned}
&\|\text{Normalize}(\mathbf{\Psi}\mathbf{X}) - \text{Normalize}(\mathbf{\Psi}\mathbf{X}')\|_F \\
=&\left\|\frac{\mathbf{\Psi}\mathbf{X}_i}{\|\mathbf{\Psi}\mathbf{X}_i\|} - \frac{\mathbf{\Psi}\mathbf{X}_i'}{\|\mathbf{\Psi}\mathbf{X}_i'\|}\right\|_2 = \left\|\frac{\mathbf{\Psi}\mathbf{X}_i}{\|\mathbf{\Psi}\mathbf{X}_i\|} - \frac{\mathbf{\Psi}\mathbf{X}_i'}{\|\mathbf{\Psi}\mathbf{X}_i\|} + \frac{\mathbf{\Psi}\mathbf{X}_i'}{\|\mathbf{\Psi}\mathbf{X}_i\|} - \frac{\mathbf{\Psi}\mathbf{X}_i'}{\|\mathbf{\Psi}\mathbf{X}_i'\|}\right\|_2 \\
=&\left\|\frac{\mathbf{\Psi}(\mathbf{X}_i-\mathbf{X}_i')}{\|\mathbf{\Psi}\mathbf{X}_i\|} + \mathbf{\Psi}\mathbf{X}_i'\left(\frac{1}{\|\mathbf{\Psi}\mathbf{X}_i\|} - \frac{1}{\|\mathbf{\Psi}\mathbf{X}_i'\|}\right)\right\| \\
\le&\frac{\|\mathbf{\Psi}(\mathbf{X}_i-\mathbf{X}_i')\|_2}{\|\mathbf{\Psi}\mathbf{X}_i\|} + \|\mathbf{\Psi}\mathbf{X}_i'\|\frac{|\|\mathbf{\Psi}\mathbf{X}_i'\| - \|\mathbf{\Psi}\mathbf{X}_i\||}{\|\mathbf{\Psi}\mathbf{X}_i\|\|\mathbf{\Psi}\mathbf{X}_i'\|} \\
\le&\frac{2\|\mathbf{\Psi}(\mathbf{X}_i-\mathbf{X}_i')\|_2}{\|\mathbf{\Psi}\mathbf{X}_i\|} \le 4\sqrt{\frac{\mu_0}{d}}\frac{\|\mathbf{\Psi}\boldsymbol{e}_j\|}{\|\mathbf{\Psi}\mathbf{X}_i\|} \le 4\sqrt{\frac{\mu_0}{d}}\frac{1+\epsilon}{1-\epsilon}.
\end{aligned}$$

The last step uses the fact that $\mathbf{\Psi}$ is JL with parameter $\epsilon$.

**Lemma 22** (Gaussian Mechanism, (Kenthapadi et al., 2013))**.** *Let $\Delta_2 f$ be the $\ell_2$ sensitivity of $f$, Let $\varepsilon \in (0,1)$ be arbitrary. The procedure that output $f(\mathbf{X}) + \mathcal{N}(0,\sigma^2 I)$ with $\sigma \ge \Delta_2 f\sqrt{2\log(1.25/\delta)}/\varepsilon$ is $(\varepsilon,\delta)$-differentially private.*

Our claim follows by applying Gaussian Mechanism and the closedness to postprocessing property of data privacy. □

**Utility Claim** It turns out that if column spikiness $\mu_0$ is a constant, Lasso-SSC is able to provably detect the correct subspace structures, despite privacy constraints.

**Corollary 23.** *Let the raw data $\mathbf{X}$ be compressed and privatized data $\widetilde{\mathbf{X}}'$ using the above described mechanism. Assume the same set of notations and assumptions in Theorem 15. Suppose $\mathbf{\Psi}$ is a JL transform with parameter $\epsilon$. Let $B := \min_{\ell=1,\dots,k}\{\rho, r^{-1/2}, \rho_\ell - \mu_\ell\}$, and $C$ be the constant in Theorem 15 and 19. If the privacy parameter $\varepsilon$ is set to*

$$\varepsilon > \sqrt{\frac{512\mu_0\log(1.25/\delta)}{d}}\max\left\{\frac{(p\log N)^{1/4}}{(CB)^{1/2}}, \frac{\sqrt{\log N}}{CB}\right\}.$$

*Then the solution to Lasso-SSC using obeys the subspace detection property with probability $1 - 8/N - \delta$.*

The idea is simple. We are now injecting artificial Gaussian noise to a compressed subspace clustering problem with fixed input, and Theorem 19 ( Theorem 8 in (Wang & Xu, 2013) ) directly addresses that. All we have to do is to replace the geometric quantities in $\mu_\ell$ and $\rho_\ell$ by their respective bound after compression in Corollary 13 and Lemma 14.

*Proof of Corollary 23.* The proof involves applying Theorem 19 with $\xi = 1$ and

$$\sigma = \frac{1+\epsilon}{1-\epsilon}\sqrt{\frac{32p\mu_0\log(1.25/\delta)}{d\varepsilon^2}} \le \sqrt{\frac{128p\mu_0\log(1.25/\delta)}{d\varepsilon^2}}$$

according to Theorem 21 and rearranging the expressions in terms of the limit for privacy requirement $\varepsilon$.

Note that the noise here is added after the compression and normalization, but the effect is the same as adding Gaussian noise in the original dimension and scaled orthogonal random projection on a noise. In fact, we can replace $C/4$ with $C$ because there is no renormalization here.

Denote $B := \min_{\ell=1,\dots,k}\{\rho, r^{-1/2}, \rho_\ell - \mu_\ell\}$, and $C$ to be the same as in Theorem 19, the conditions for success is

$$\sigma(1+\sigma) < CB\sqrt{\frac{p}{\log N}}, \tag{B.1}$$

which holds if

$$\sigma < \min\left\{\frac{CB}{2}\sqrt{\frac{p}{\log N}}, \sqrt{\frac{CB}{2}}\frac{p^{1/4}}{(\log N)^{1/4}}\right\}.$$

Substitute the expression of $\sigma$ into (B.1) and rewrite it in terms of $\varepsilon$, we get our claim:

$$\varepsilon > \sqrt{\frac{512\mu_0 \log(1.25/\delta)}{d}} \max\left\{\frac{(p\log N)^{1/4}}{(CB)^{1/2}}, \frac{\sqrt{\log N}}{CB}\right\}.$$

□

### B.1. Discussion of user-level privacy and its impossiblity under perfect subspace detection property

As we described in the main results, attribute-level differential privacy is a much weaker notion of privacy. While it is easy to handle a small group of attributes (in the order of $O(\sqrt{d/p})$ if we consider $B = O(1/r)$) by the composition rule, it does not protect any individual user's complete information. However, this is arguably the best we can do if our measure of utility is in terms of (perfect) subspace detection property.

Let us define formally the user-level differential privacy.

**Definition 24** (User-Level Differential Privacy)**.** *We say a randomized algorithm $\mathcal{A} : \mathbb{R}^{d\times N} \to \mathcal{O}$ is $(\varepsilon, \delta)$-differential private at attribute level if*

$$\mathbb{P}(\mathcal{A}(\mathbf{X}) \in \mathcal{S}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(\mathbf{X}') \in \mathcal{S}) + \delta$$

*for any measurable outcome $\mathcal{S} \subset \mathcal{O}$, any $\mathbf{X}, \mathbf{X}' \in \mathcal{X}^n$ that differs in only one column.*

The only difference to the attribute differential privacy is how $\mathbf{X}$ and $\mathbf{X}'$ may differ. Note that we can arbitrarily replace any single point in $\mathbf{X}$ with any $x \in \mathcal{X}$, to form $\mathbf{X}'$.

**Proposition 25.** *User-level differential privacy is NOT possible for any $0 \leq \varepsilon < \infty$ if we assume perfect subspace detection property, or perfect clustering results. In addition, If an algorithm achieves perfect clustering or subspace detection with probability $1-\delta$, user-level differential privacy is NOT possible for any $\varepsilon < \log\left(\frac{1-\delta}{\delta}\right)$.*

*Proof.* First of all, if a data point can be arbitrarily chosen, then we can change it entirely into a different subspace. Let's first ignore the gap from subspace detection property and perfect clustering. Assume that the output is the clustering result and it is always correct. then if we arbitrarily change the $k$th data point from one Subspace A to Subspace B, the result must reflect the change and cluster this data point correctly to its new subspace and the probability of observing an output that has $k$th data point clustered into Subspace A will change from 1 to 0, which blatantly violates the definition of differential privacy.

The same line of arguments holds if we treat the output as the graph embedding. Note that having subspace detection property for data point $k$ in Subspace A (connected only to a set of points) and having subspace detection for data point $k$ in Subspace B (connected only to another set of points) are two disjoint measurable events. With a perturbation that changes a data point from one subspace to another will blow the likelihood ratio of observing one of these two event to infinity.

The high probability statement holds because

$$\frac{\mathbb{P}(\text{SDP according to } \mathbf{X}|\mathbf{X})}{\mathbb{P}(\text{SDP according to } \mathbf{X}|\mathbf{X}')} \geq \frac{1-\delta}{\delta} \geq e^{\log(\frac{1-\delta}{\delta})}.$$

□

The reason why attribute-level privacy will work is because the promise is much weaker. Also our assumption that the columns are non-spiky ensures that perturbing any attribute of any user will not inject too much error. Intuitively, random projection and the injected dense Gaussian noise makes sure that it is not possible to identify any small changes in one attribute of a single user.

To be fair, the same problem still exists, namely, differential privacy breaks whenever the clustering can be shown to be always correct. What attribute-differential privacy ensures is that it is not possible to tell if a specific attribute of this user used in coming up with the result is actually the same or close to what it truly is.

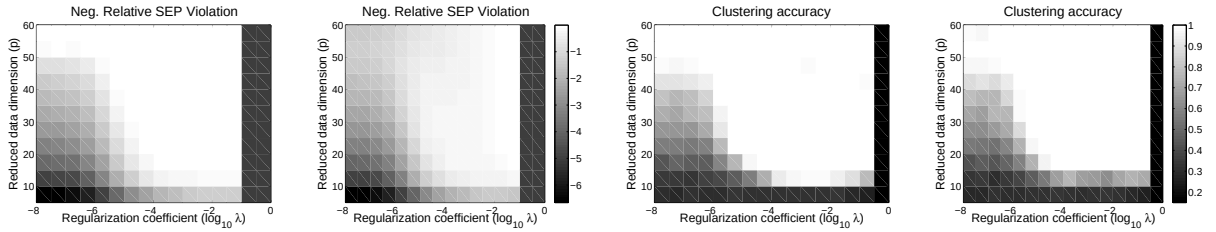


*Figure 5.* Relative Violation (top) and clustering accuracy (bottom) of Lasso-SSC on noiseless and noisy synthetic datasets. Left: noiseless; right: $\sigma/\sqrt{d} = 0.1$. $\lambda$ ranges from $10^{-1}$ to $10^{-8}$ and the projected data dimension ($p$) ranges from 5 to 60. For each figure the rightmost columns indicate trivial solutions.

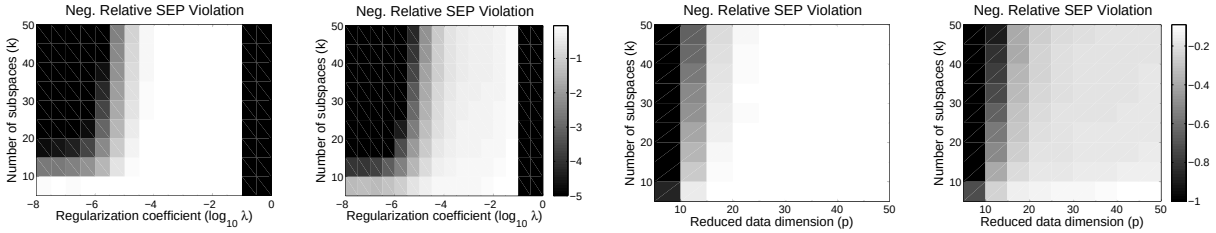


*Figure 6.* Relative violation of Lasso-SSC on noiseless and noisy synthetic datasets with varying number of clusters ($k$). Top row: $\lambda$ ranges from $10^{-8}$ to 1; data dimension after random projection ($p$) is set to 25; rightmost columns indicate trivial solutions. Bottom row: $p$ ranges from 5 to 50; $\lambda$ is set to $10^{-2}$. Left: noiseless; right: $\sigma/\sqrt{d} = 0.1$.

User-level privacy for subspace clustering and for privacy in general remains an important open problem. What we know for sure is that, we need to come up with a different/soft measure of utility other than exact clustering or subspace detection property.

## C. Numerical results on synthetic datasets

We generate synthetic datasets to verify and extend theoretical findings in this paper. All subspaces and data points within each subspace are generated uniformly at random. We fix the ambient dimension ($d$) to be 100 and generate 50 data points per cluster. The intrinsic rank of each subspace is fixed to $r = 5$.

In the first set of experiments we generate $K = 10$ clusters and plot the relative violation of SEP as well as clustering accuracy with respect to different $\lambda$ and $p$ values in Figure 5. It can be shown that when the projected dimension $p$ is smaller than the rank of the union of subspaces (i.e., $p < kr$) the performance of Lasso SSC degrades as $\lambda$ decreases. This holds even for the noiseless case, which nicely justifies our theoretical findings. Note that when $p$ is large (e.g., $p > kr$) both Lasso SSC and exact SSC ($\lambda \to 0$) succeeds when the input data matrix is not corrupted with noise. On the other hand, when $\lambda$ is too large we obtain trivial solutions and clustering fails immediately.

In Figure 6 we report the relative violation of SEP and clustering accuracy with varying number of clusters $k$. It can be seen that even when there are a large number of clusters (e.g., $k = 50$) SEP still holds for a wide range of tuning parameters $\lambda$. In addition, the bottom two plots in Figure 6 show that the choice of projection dimension $p$ is insensitive to the number of clusters ($k$).

## D. Some tail inequalities

**Lemma 26** (Matrix Gaussian and Rademacher Series, the general case (**?**))**.** *Let $\{\mathbf{B}_k\}_k$ be a finite sequence of fixed matrices with dimensions $d_1 \times d_2$. Let $\{\gamma_k\}_k$ be a finite sequence of i.i.d. standard normal variables. Define the summation random matrix* $\mathbf{Z}$ *as*

$$\mathbf{Z} = \sum_k \gamma_k \mathbf{B}_k. \tag{D.1}$$

*Define the variance parameter* $\sigma^2$ *as*

$$\sigma^2 := \max\{\|\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top]\|, \|\mathbb{E}[\mathbf{Z}^\top\mathbf{Z}]\|\}. \tag{D.2}$$

*Then for every* $t > 0$ *the following concentration inequality holds:*

$$\Pr\left[\|\mathbf{Z}\| \geq t\right] \leq (d_1 + d_2)e^{-t^2/2\sigma^2}. \tag{D.3}$$

**Lemma 27** (Noncommutative Matrix Berstein Inequality, (**?**Recht, 2011))**.** *Let* $\mathbf{B}_1, \cdots, \mathbf{B}_p$ *be independent zero-mean square* $r \times r$ *random matrices. Suppose* $\sigma_j^2 = \max\{\|\mathbb{E}[\mathbf{B}_j\mathbf{B}_j^\top]\|, \|\mathbb{E}[\mathbf{B}_j^\top\mathbf{B}_j]\|\}$ *and* $\|\mathbf{B}_j\| \leq R$ *almost surely for every* $j$*. Then for any* $t > 0$ *the following inequality holds:*

$$\Pr\left[\left\|\sum_{j=1}^{p}\mathbf{B}_j\right\|_2 > t\right] \leq 2r\exp\left(-\frac{t^2/2}{\sum_{j=1}^{p}\rho_j^2 + Rt/3}\right). \tag{D.4}$$

**Lemma 28** (Spectrum bound of a Gaussian random matrix,(**?**))**.** *Let* $A$ *be an* $m \times n$ $(m > n)$ *matrix with i.i.d standard Gaussian entries. Then, its largest and smallest singular values* $s_1(A)$ *and* $s_n(A)$ *obeys*

$$\sqrt{m} - \sqrt{n} \leq \mathbb{E}s_n(A) \leq \mathbb{E}s_1(A) \leq \sqrt{m} + \sqrt{n},$$

*moreover,*

$$\sqrt{m} - \sqrt{n} - t \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + \sqrt{n} + t,$$

*with probability at least* $1 - 2\exp(-t^2/2)$ *for all* $t > 0$*.*

The expectation result is due to Gordon's inequality and the concentration follows from the concentration of measure inequality in Gauss space by the fact that $s_1$ and $s_n$ are both 1-Lipchitz functions. Take $t = \sqrt{n}$ in the above inequality we get

$$1 - 2\sqrt{\frac{n}{m}} - \epsilon \leq s_n(A/\sqrt{m}) \leq s_1(A/\sqrt{m}) \leq 1 + 2\sqrt{\frac{n}{m}}$$

with probability $1 - 2\exp(-n^2/2)$.

## Table of symbols and notations

*Table 1.* Summary of symbols and notations

| | |
|---|---|
| $\lvert\cdot\rvert$ | Either absolute value or cardinality |
| $\lVert\cdot\rVert$; $\lVert\cdot\rVert_2$ | 2 norm of a vector/spectral norm of a matrix |
| $\lVert\cdot\rVert_1$ | 1 norm of a vector |
| $\lVert\cdot\rVert_\infty$ | Infinity norm (maximum absolute value) of a vector |
| $\langle\cdot,\cdot\rangle$ | Inner product of two vectors |
| $\lVert\mathbf{A}\rVert_{(i)}$ | The $i$th row of matrix $\mathbf{A}$ |
| $\sigma_1(\cdot), \sigma_r(\cdot)$ | The largest and $r$th largest singular value of a matrix |
| $N$ | Number of data points (number of columns in $\mathbf{X}$) |
| $k$ | Number of subspaces (clusters) |
| $d$ | The ambient dimension (number of rows in $\mathbf{X}$) |
| $N_\ell, r_\ell$ for $\ell = 1, \cdots, k$ | Number of data points and instrinsic dimension for each subspace |
| $r$ | Largest intrinsic dimension across all subspaces |
| $\mathbf{X}$ | Observed data matrix |
| $\mathbf{Y}$ | Uncorrupted (noiseless) data matrix |
| $\mathbf{Z}$ | Noise matrix, can be either deterministic or stochastic |
| $\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}, \widetilde{\mathbf{Z}}$ | Projected matrices of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ |
| $\widetilde{\mathbf{X}}', \widetilde{\mathbf{Y}}', \widetilde{\mathbf{Z}}'$ | Normalized projected matrices of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ |
| $\mathcal{U}^{(\ell)}$, $\mathbf{U}^{(\ell)}$ | Subspace and its orthonormal basis of the $\ell$th cluster |
| $\mathbf{X}_{-i}$, $\mathbf{Y}_{-i}$, $\mathbf{Z}_{-i}$ | All columns in $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ except the $i$th column. |
| $\mathbf{X}^{(\ell)}$,$\mathbf{Y}^{(\ell)}$,$\mathbf{Z}^{(\ell)}$ | All columns in $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ associated with the $\ell$th subspace |
| $\mathbf{X}^{(\ell)}_{-i}$,$\mathbf{Y}^{(\ell)}_{-i}$,$\mathbf{Z}^{(\ell)}_{-i}$ | All columns in $\mathbf{X}^{(\ell)}, \mathbf{Y}^{(\ell)}, \mathbf{Z}^{(\ell)}$ except the $i$th column |
| $\mathcal{Q}(\cdot)$,conv$(\cdot)$ | (Symmetric) convex hull of a set of vectors |
| $r(\cdot)$ | Radius of the largest ball inscribed in a convex body |
| $\mathcal{P}_{\mathcal{U}}(\cdot)$ | Projection onto subspace $\mathcal{U}$ |
| $p$ | Target dimension after random projection |
| $\epsilon$ | Approximation error of random projection methods |
| $\delta$ | Failure probability |
| $\mathbf{\Psi}, \mathbf{\Omega}, \mathbf{\Phi}, \mathbf{S}$ | Projection operators for random Gaussian projection, uniform sampling, FJLT and sketching |
| $\mu_0$ | Column space incoherence or column spikiness |
| $\mu_\ell, \rho_\ell$ for $\ell = 1, \cdots, k$ | Subspace incoherence and inradius for each subspace |
| $\tilde{\mu}_\ell, \tilde{\rho}_\ell$ for $\ell = 1, \cdots, k$ | Subspace incoherence and inradius on the projected data |
| $f(\cdot), \tilde{f}(\cdot)$ | Objective functions of Eq. (4.1) on the original data and projected data |
| $\boldsymbol{\nu}, \boldsymbol{v}$ | Unnormalized and normalized dual direction |
| $\tilde{\boldsymbol{\nu}}$ | Random projection of $\boldsymbol{\nu}$ |
| $\tilde{\boldsymbol{\nu}}'$ | A shrinked version of $\tilde{\boldsymbol{\nu}}$ such that it is feasible for Eq. (4.1) on projected data |
| $\boldsymbol{\nu}^*$ | Optimal solution to Eq. (4.1) on projected data |
| $\bar{\boldsymbol{\nu}}$ | A vector in the original space that corresponds to $\boldsymbol{\nu}^*$ after projection |
| $\bar{\boldsymbol{\nu}}'$ | A shrinked version of $\bar{\boldsymbol{\nu}}$ such that it is feasible for Eq. (4.1) on the original data |
| $\lambda$ | Regularization coefficient for Lasso SSC |
| $\Delta$ | Margin of error (i.e., $\min_\ell \rho_\ell - \mu_\ell$) |
| $\eta, \tilde{\eta}$ | Noise level for deterministic noise, before and after projection |
| $\sigma, \tilde{\sigma}$ | Noise level for random Gaussian noise, before and after projection |
| $\mathbf{C}$ | Similarity matrix |
| $q$ | Number of nonzero entries in regression solutions. Used in solution path algorithms. |