Noisy Sparse Subspace Clustering

Yu-Xiang Wang YUXIANGWANG@NUS.EDU.SG
Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Huan Xu MPEXUH@NUS.EDU.SG
Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Abstract

This paper considers the problem of subspace clustering under noise. Specifically, we study the behavior of Sparse Subspace Clustering (SSC) when either adversarial or random noise is added to the unlabelled input data points, which are assumed to lie in a union of low-dimensional subspaces. We show that a modified version of SSC is *provably effective* in correctly identifying the underlying subspaces, even with noisy data. This extends theoretical guarantee of this algorithm to the practical setting and provides justification to the success of SSC in a class of real applications.

1. Introduction

Subspace clustering is a problem motivated by many real applications. It is now widely known that many high dimensional data including 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) can be modelled as samples drawn from the *union* of multiple low-dimensional subspaces (illustrated in Figure 1). Subspace clustering, arguably the most crucial step to understand such data, refers to the task of clustering the data into their original subspaces and uncovers the underlying structure of the data. The partitions correspond to different rigid objects for motion trajectories, different people for face data, subnets for network data, like-minded users in movie database and latent communities for social graph.

Subspace clustering has drawn significant attention in the last decade and a great number of algorithms have been proposed, including K-plane (Bradley & Mangasarian, 2000), GPCA (Vidal et al., 2005), Spectral Curvature Clustering (Chen & Lerman, 2009), Low Rank Representation (LRR) (Liu et al., 2013) and Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2009). Among them, SSC is known to enjoy superb empirical performance, *even for noisy data*. For example, it is the state-of-the-art algorithm for motion segmentation on Hopkins155 benchmark (Tron & Vidal, 2007). For a comprehensive survey and comparisons, we refer the readers to the tutorial(Vidal, 2011).

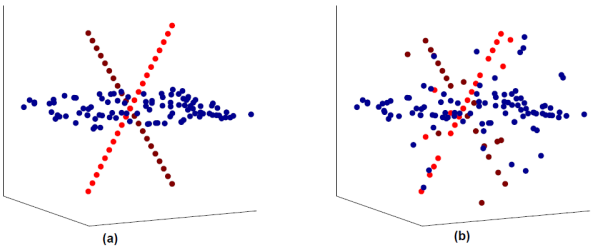

Figure 1. Exact (a) and noisy (b) data in union-of-subspace

Effort has been made to explain the practical success of SSC. Elhamifar & Vidal (2010) show that under certain conditions, *disjoint* subspaces (i.e., they are not overlapping) can be exactly recovered. Similar guarantee, under stronger "independent subspace" condition, was provided for LRR in a much earlier analysis(Kanatani, 2001). The recent geometric analysis of SSC (Soltanolkotabi & Candes, 2012) broadens the scope of the results significantly to the case when subspaces can be overlapping. However, while these analyses advanced our understanding of SSC, one common drawback is that data points are assumed to be lying *exactly* in the subspace. This assumption can hardly be satisfied in practice. For example, motion trajectories data are only *approximately* rank-4 due to perspective distortion of camera.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

In this paper, we address this problem and provide the first theoretical analysis of SSC with noisy or corrupted data. Our main result shows that a modified version of SSC (see (2.2)) when the magnitude of noise does not exceed a threshold determined by a geometric gap between *inradius* and *subspace incoherence* (see below for precise definitions). This complements the result of Soltanolkotabi & Candes (2012) that shows the same geometric gap determines whether SSC succeeds for the noiseless case. Indeed, our results reduce to the noiseless results of Soltanolkotabi & Candes when the noise magnitude diminishes.

While our analysis is based upon the geometric analysis of Soltanolkotabi & Candes (2012), the analysis is much more involved: In SSC, sample points are used as the dictionary for sparse recovery, and therefore noisy SSC requires analyzing noisy dictionary. This is a hard problem and we are not aware of any previous study that proposed guarantee in the case of noisy dictionary except Loh & Wainwright (2012) in the high-dimensional regression problem. We also remark that our results on noisy SSC are *exact*, i.e., as long as the noise magnitude is smaller than the threshold, the obtained subspace recovery is *correct*. This is in sharp contrast to the majority of previous work on structure recovery for noisy data where stability/perturbation bounds are given – i.e., the obtained solution is *approximately* correct, and the approximation gap goes to zero only when the noise diminishes.

2. Problem setup

Notations: We denote the uncorrupted data matrix by $Y \in \mathbb{R}^{n\times N}$, where each column of Y (normalized to unit vector) belongs to a union of L subspaces

$$\mathcal{S}_1 \cup \mathcal{S}_2 \cup ... \cup \mathcal{S}_L.$$

Each subspace $\mathcal{S}_\ell$ is of dimension d_ℓ and contains N_ℓ data samples with $N_1 + N_2 + ... + N_L = N$. We observe the noisy data matrix $X = Y + Z$, where Z is some arbitrary noise matrix. Let $Y^{(\ell)} \in \mathbb{R}^{n\times N_\ell}$ denote the selection of columns in Y that belongs to $\mathcal{S}_\ell$, and let the corresponding columns in X and Z be denoted by $X^{(\ell)}$ and $Z^{(\ell)}$. Without loss of generality, let $X = [X^{(1)}, X^{(2)}, ..., X^{(L)}]$ be ordered. In addition, we use subscript "$-i$" to represent a matrix that excludes column i, e.g., $X^{(\ell)}_{-i} = [x^{(\ell)}_1, ..., x^{(\ell)}_{i-1}, x^{(\ell)}_{i+1}, ..., x^{(\ell)}_{N_\ell}]$. Calligraphic letters such as $\mathcal{X}, \mathcal{Y}_\ell$ represent the set containing all columns of the corresponding matrix (e.g., X and $Y^{(\ell)}$).

For any matrix X, $\mathcal{P}(X)$ represents the symmetrized convex hull of its columns, i.e., $\mathcal{P}(X) = \text{conv}(\pm\mathcal{X})$. Also let $\mathcal{P}^{(\ell)}_{-i} := \mathcal{P}(X^{(\ell)}_{-i})$ and $\mathcal{Q}^{(\ell)}_{-i} := \mathcal{P}(Y^{(\ell)}_{-i})$ for short. $\mathbb{P}_\mathcal{S}$ and $\text{Proj}_\mathcal{S}$ denote respectively the projection matrix and projection operator (acting on a set) to subspace $\mathcal{S}$. Throughout the paper, $\|\cdot\|$ represents 2-norm for vectors and operator norm for matrices; other norms will be explicitly specified (e.g., $\|\cdot\|_1, \|\cdot\|_\infty$).

Method: Original SSC solves the linear program

$$\min_{c_i} \|c_i\|_1 \quad s.t. \quad x_i = X_{-i}c_i \tag{2.1}$$

for each data point x_i. Solutions are arranged into matrix $C = [c_1, ..., c_N]$, then spectral clustering techniques such as Ng et al. (2002) are applied on the affinity matrix $W = |C| + |C|^T$. Note that when $Z \neq 0$, this method breaks down: indeed (2.1) may even be infeasible.

To handle noisy X, a natural extension is to relax the equality constraint in (2.1) and solve the following unconstrained minimization problem instead (Elhamifar & Vidal, 2012):

$$\min_{c_i} \|c_i\|_1 + \frac{\lambda}{2}\|x_i - X_{-i}c_i\|^2. \tag{2.2}$$

We will focus on Formulation (2.2) in this paper. Notice that (2.2) coincide with standard LASSO. Yet, since our task is subspace clustering, the analysis of LASSO (mainly for the task of support recovery) does not extend to SSC. In particular, existing literature for LASSO to succeed requires the dictionary X_{-i} to satisfy RIP (Candès, 2008) or the Null-space property (Donoho et al., 2006), but neither of them is satisfied in the subspace clustering setup.[1]

In the subspace clustering task, there is no single "ground-truth" C to compare the solution against. Instead, the algorithm succeeds if each sample is expressed as a linear combination of samples belonging to the same subspace, as the following definition states.

Definition 1 (LASSO Subspace Detection Property)**.** *We say subspaces $\{\mathcal{S}_\ell\}_{\ell=1}^k$ and noisy sample points X from these subspaces obey LASSO subspace detection property with λ, if and only if it holds that for all i, the optimal solution c_i to (2.2) with parameter λ satisfies: (1) c_i is not a zero vector, (2) Nonzero entries of c_i correspond to only columns of X sampled from the same subspace as x_i.*

This property ensures that output matrix C and (naturally) affinity matrix W are exactly block diagonal with each subspace cluster represented by a disjoint

[1] There may exist two identical columns in X_{-i}, hence violate RIP for 2-sparse signal and has maximum incoherence $\mu(X_{-i}) = 1$.

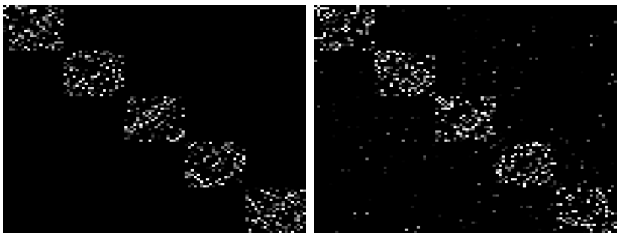

Figure 2. Illustration of LASSO-Subspace Detection Property/Self-Expressiveness Property. **Left:** SEP holds. **Right:** SEP is violated even though spectral clustering is likely to cluster this affinity graph perfectly into 5 blocks.

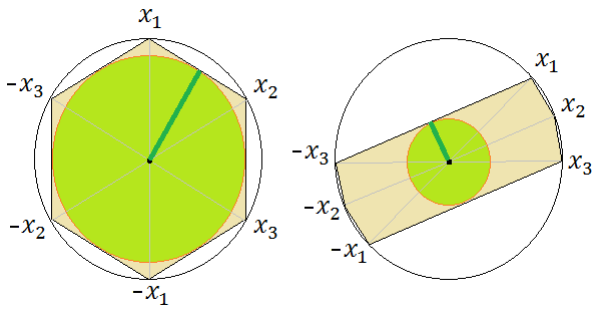

Figure 3. Illustration of inradius and data distribution.

block.[2] The property is illustrated in Figure 2. For convenience, we will refer to the second requirement alone as "*Self-Expressiveness Property*" (SEP), as defined in Elhamifar & Vidal (2012).

Models of analysis: Our objective here is to provide sufficient conditions upon which the LASSO subspace detection properties hold in the following four models. Precise definition of the noise models will be given in Section 3.

- fully deterministic model
- deterministic data+random noise
- semi-random data+random noise
- fully random model.

3. Main results

3.1. Deterministic model

We start by defining two concepts adapted from Soltanolkotabi & Candes's original proposal.

Definition 2 (Projected Dual Direction[3])**.** *Let ν be the optimal solution to*

$$\max_{\nu} \langle x, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu, \quad \textit{subject to:} \quad \|X^T\nu\|_\infty \leq 1;$$

and $\mathcal{S}$ is a low-dimensional subspace. The projected dual direction *v is defined as*

$$v(x, X, \mathcal{S}, \lambda) \triangleq \frac{\mathbb{P}_{\mathcal{S}}\nu}{\|\mathbb{P}_{\mathcal{S}}\nu\|}.$$

Definition 3 (Projected Subspace Incoherence Property)**.** *Compactly denote projected dual direction $v_i^{(\ell)} = v(x_i^{(\ell)}, X_{-i}^{(\ell)}, \mathcal{S}_\ell, \lambda)$ and $V^{(\ell)} = [v_1^{(\ell)}, ..., v_{N_\ell}^{(\ell)}]$. We say that vector set $\mathcal{X}_\ell$ is μ-incoherent to other points if*

$$\mu \geq \mu(\mathcal{X}_\ell) := \max_{y \in \mathcal{Y}\setminus\mathcal{Y}_\ell} \|{V^{(\ell)}}^T y\|_\infty.$$

Here, μ measures the incoherence between corrupted subspace samples $\mathcal{X}_\ell$ and clean data points in other subspaces. As $\|y\| = 1$ by definition, the range of μ is $[0, 1]$. In case of random subspaces in high dimension, μ is close to zero. Moreover, as we will see later, for deterministic subspaces and random data points, μ is proportional to their expected angular distance (measured by cosine of canonical angles).

Definition 2 and 3 are different from their original versions proposed in Soltanolkotabi & Candes (2012) in that we require a projection to a particular subspace to cater to the analysis of the noise case.

Definition 4 (inradius)**.** *The inradius of a convex body $\mathcal{P}$, denoted by $r(\mathcal{P})$, is defined as the radius of the largest Euclidean ball inscribed in $\mathcal{P}$.*

The inradius of a $\mathcal{Q}_{-i}^{(\ell)}$ describes the distribution of the data points. Well-dispersed data lead to larger inradius and skewed/concentrated distribution of data have small inradius. An illustration is given in Figure 3.

Definition 5 (Deterministic noise model)**.** *Consider arbitrary additive noise Z to Y, each column z_i is characterized by the three quantities below:*

$$\delta := \max_i \|z_i\| \quad \delta_1 := \max_{i,\ell} \|\mathbb{P}_{\mathcal{S}_\ell} z_i\| \quad \delta_2 := \max_{i,\ell} \|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_i\|$$

Theorem 1. *Under deterministic noise model, compactly denote*

$$\mu_\ell = \mu(\mathcal{X}_\ell), \quad r_\ell := \min_{\{i: x_i \in \mathcal{X}_\ell\}} r(\mathcal{Q}_{-i}^{(\ell)}), \quad r = \min_{\ell=1,...,L} r_\ell.$$

If $\mu_\ell < r_\ell$ for each $\ell = 1, ..., L$, furthermore

$$\delta \leq \min_{\ell=1,...,L} \frac{r(r_\ell - \mu_\ell)}{3r_\ell^2 + 8r_\ell + 2}$$

[2] Note that this is a very strong condition. In general, spectral clustering does not require the exact block diagonal structure for perfect classifications (check Figure 8 in our simulation section for details).

[3] This definition relate to (4.3), the dual problem of (2.2), which we will define in the proof.

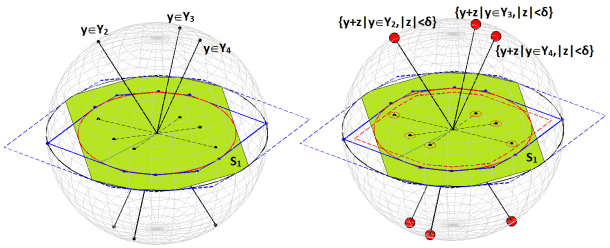

Figure 4. Geometric interpretation and comparison of the noiseless SSC (**Left**) and noisy Lasso-SSC (**Right**).

then LASSO subspace detection property holds for all weighting parameter λ in the range

$$\begin{cases} \lambda > \dfrac{1}{(r-\delta_1)(1-3\delta)-3\delta-2\delta^2}, \\ \lambda < \dfrac{2r}{\delta^2(r+1)} \vee \min\limits_{\ell=1,..,L} \dfrac{r_\ell - \mu_\ell - 2\delta_1}{\delta(1+\delta)(3+2r_\ell-2\delta_1)}, \end{cases}$$

which is guaranteed to be non-empty.

Remark 1 (Noiseless case). *When $\delta = 0$, i.e., there is no noise, the condition reduces to $\mu_\ell < r_\ell$, precisely the form in Soltanolkotabi & Candes (2012). However, the latter only works for the the exact LP formulation (2.1), our result works for the (more robust) unconstrained LASSO formulation (2.2) for any $\lambda > \frac{1}{r}$.*

Remark 2 (Signal-to-Noise Ratio). *Condition $\delta \leq \frac{r(r-\mu)}{3r^2+8r+2}$ can be interpreted as the breaking point under increasing magnitude of attack. This suggests that SSC by (2.2) is provably robust to arbitrary noise having signal-to-noise ratio (SNR) greater than $\Theta\big(\frac{1}{r(r-\mu)}\big)$. (Notice that $0 < r < 1$, we have $3r^2+8r+2 = \Theta(1)$.)*

Remark 3 (Geometric Interpretation). *The geometric interpretation of our results is give in Figure 4. On the left, Theorem 2.5 of Soltanolkotabi & Candes (2012) suggests that the projection of external data points must fall inside the solid blue polygon, which is the intersection of halfspaces defined by dual directions (blue dots) that are tangent planes of the red inscribing sphere. On the right, the guarantee of Theorem 1 means that the whole red sphere (analogous to uncertainty set in Robust Optimization (Ben-Tal & Nemirovski, 1998; Bertsimas & Sim, 2004)) of each external data point must fall inside the dashed red polygon, which is smaller than the blue polygon by a factor related to the noise level.*

Remark 4 (Matrix version of the algorithm). *The theorem suggests there's a single λ that works for all x_i, X_{-i} in (2.2). This makes it possible to extend the results to the compact matrix algorithm below*

$$\begin{aligned} \min_C \ & \|C\|_1 + \frac{\lambda}{2}\|X - XC_i\|_F^2 \\ s.t. \quad & \operatorname{diag}(C) = 0, \end{aligned} \tag{3.1}$$

which can be solved numerically using alternating direction method of multipliers (ADMM) (Boyd et al., 2011). See the supplementary material for the details of the algorithm.

3.2. Randomized models

We analyze three randomized models with increasing level of randomness.

- **Determinitic+Random Noise.** Subspaces and samples in subspace are fixed; noise is random.
- **Semi-random+Random Noise.** Subspace is deterministic, but samples in each subspace are drawn uniformly at random, noise is random.
- **Fully random.** Both subspace and samples are drawn uniformly at random; noise is also random.

Definition 6 (Random noise model). *Our random noise model is defined to be any additive Z that is (1) columnwise iid; (2) spherical symmetric; and (3) $\|z_i\| \leq \delta$ with high probability.*

Example 1 (Gaussian noise). *A good example of our random noise model is iid Gaussian noise. Let each entry $Z_{ij} \sim N(0, \sigma/\sqrt{n})$. It is known that*

$$\delta := \max_i \|z_i\| \leq \sqrt{1 + \frac{6\log N}{n}}\sigma$$

with probability at least $1 - C/N^2$ for some constant C (by Lemma B.2).

Theorem 2 (Deterministic+Random Noise). *Under random noise model, compactly denote r_ℓ, r and μ_ℓ as in Theorem 1, furthermore let*

$$\epsilon := \sqrt{\frac{6\log N + 2\log \max_\ell d_\ell}{n - \max_\ell d_\ell}} \leq \frac{C\log(N)}{\sqrt{n}}.$$

If $r > 3\epsilon/(1-6\epsilon)$ and $\mu_\ell < r_\ell$ for all $\ell = 1, ..., k$, furthermore

$$\delta < \min_{\ell=1,...,L} \frac{r_\ell - \mu_\ell}{3r_\ell + 6},$$

then with probability at least $1 - 7/N$, LASSO subspace detection property holds for all weighting parameter λ in the range

$$\begin{cases} \lambda > \dfrac{1}{(r-\delta\epsilon)(1-3\delta)-3\delta-2\delta^2}, \\ \lambda < \dfrac{2r}{\delta^2(r+1)} \vee \min\limits_{\ell=1,...,L} \dfrac{r_\ell - \mu_\ell - 2\delta\epsilon}{\epsilon\delta(1+\delta)(3+2r_\ell-2\delta\epsilon)}, \end{cases}$$

which is guaranteed to be non-empty.

Remark 5 (Margin of error). *Compared to Theorem 1, Theorem 2 considers a more benign noise which leads to a much stronger result. Observe that in the random noise case, the magnitude of noise that SSC can tolerate is proportional to $r_\ell - \mu_\ell$ – the difference of inradius and incoherence – which is the fundamental geometric gap that appears in the noiseless guarantee of Soltanolkotabi & Candes (2012). We call this gap the **Margin of error**.*

We now analyze this margin of error. We start from the semi-random model, where the distance between two subspaces is measured as follows.

Definition 7. *The* affinity *between two subspaces is defined by:*

$$\text{aff}(\mathcal{S}_k, \mathcal{S}_\ell) = \sqrt{\cos^2 \theta_{k\ell}^{(1)} + ... + \cos^2 \theta_{k\ell}^{(\min(d_k, d_\ell))}},$$

where $\theta_{k\ell}^{(i)}$ is the i^{th} canonical angle between the two subspaces. Let U_k and U_ℓ be a set of orthonormal bases of each subspace, then $\text{aff}(\mathcal{S}_k, \mathcal{S}_\ell) = \|U_k^T U_\ell\|_F$.

When data points are randomly sampled from each subspace, the geometric entity $\mu(\mathcal{X}_\ell)$ can be expressed using this (more intuitive) subspace affinity, which leads to the following theorem.

Theorem 3 (Semi-random+random noise). *Suppose $N_\ell = \kappa_\ell d_\ell + 1$ data points are randomly chosen on each $\mathcal{S}_\ell$, $1 \le \ell \le L$. Use ϵ as in Theorem 2 and let $c(\kappa)$ be a positive constant that takes value $1/\sqrt{8}$ when κ is greater than some numerical constant κ_o. If*

$$\max_{k:k\neq\ell} t \log\left[LN_\ell(N_k+1)\right] \frac{\text{aff}(\mathcal{S}_k, \mathcal{S}_\ell)}{\sqrt{d_k}} > c(\kappa_\ell)\sqrt{\frac{\log \kappa_\ell}{2}} \tag{3.2}$$

and $c(\kappa_\ell)\sqrt{\log \kappa_\ell / 2d_\ell} > 3\epsilon/(1-6\epsilon)$ for each ℓ, furthermore

$$\delta < \frac{1}{9} \min_\ell \left\{ \frac{c(\kappa_\ell)\sqrt{\log \kappa_\ell}}{\sqrt{2d_\ell}} - \max_{k:k\neq\ell} t \log\left[LN_\ell(N_k+1)\right] \frac{\text{aff}(\mathcal{S}_k, \mathcal{S}_\ell)}{\sqrt{d_k d_\ell}} \right\},$$

then LASSO subspace detection property holds for some λ[4] with probability at least $1 - \frac{7}{N} - \sum_{\ell=1}^L N_\ell \exp(-\sqrt{d_\ell(N_\ell - 1)}) - \frac{1}{L^2}\sum_{k\neq\ell} \frac{1}{N_\ell(N_k+1)} \exp(-t/4)$.

Remark 6 (Overlapping subspaces). *Similar to Soltanolkotabi & Candes (2012), SSC can handle overlapping subspaces with noisy samples, as subspace affinity can take small positive value while still keeping the margin of error positive.*

[4]The λ here (and that in Theorem 4) has a fixed non-empty range as in Theorem 1 and 2, which we omit due to space constraints.

Application	Cluster rank
3D motion segmentation (Costeira & Kanade, 1998)	$rank = 4$
Face clustering (with shadow) (Basri & Jacobs, 2003)	$rank = 9$
Diffuse photometric face (Zhou et al., 2007)	$rank = 3$
Network topology discovery (Eriksson et al., 2012)	$rank = 2$
Hand writing digits (Hastie & Simard, 1998)	$rank = 12$
Social graph clustering (Jalali et al., 2011)	$rank = 1$

Table 1. Rank of real subspace clustering problems

Theorem 4 (Fully random model). *Suppose there are L subspaces each with dimension d, chosen independently and uniformly at random. For each subspace, there are $\kappa d + 1$ points chosen independently and uniformly at random. Furthermore, each measurements are corrupted by iid Gaussian noise $\sim N(0, \sigma/\sqrt{n})$. Then for some absolute constant C, the LASSO subspace detection property holds for some λ with probability at least $1 - \frac{C}{N} - Ne^{-\sqrt{\kappa} d}$ if*

$$d < \frac{c^2(\kappa)\log \kappa}{12 \log N} n$$

and

$$\sigma < \frac{1}{18}\left(c(\kappa)\sqrt{\frac{\log \kappa}{2d}} - \sqrt{\frac{6 \log N}{n}} \right).$$

Remark 7 (Trade-off between d and the margin of error). *Theorem 4 extends our results to the paradigm where the subspace dimension grows linearly with the ambient dimension. Interestingly, it shows that the margin of error scales $\tilde{\Theta}(\sqrt{1/d})$, implying a tradeoff between d and robustness to noise. Fortunately, most interesting applications indeed have very low subspace-rank, as summarized in Table 1.*

Remark 8 (Robustness in the many-cluster setting). *Another interesting observation is that the margin of error scales logarithmically with respect to L, the number of clusters (in both $\log \kappa$ and $\log N$ since $N = L(\kappa d + 1)$). This suggests that SSC is robust even if there are many clusters, and $Ld \gg n$.*

Remark 9 (Range of valid λ in the random setting). *Substitute the bound of inradius r and subspace incoherence μ of fully random setting into the λ's range of Theorem 3, we have the the valid range of λ is*

$$\frac{C_1\sqrt{d}}{\sqrt{\log \kappa}} < \lambda < \frac{C_2 n}{\sigma\sqrt{d \log(dL)}}, \tag{3.3}$$

for some constant C_1, C_2. This again illustrates that the robustness is sensitive to d but not L.

4. Roadmap of the Proof

In this section, we lay out the roadmap of the proof for Theorem 1 to 4. Instead of analyzing (2.2) directly, we consider an equivalent constrained version by introducing slack variables:

$$\mathbf{P}_0: \quad \min_{c_i,e_i} \|c_i\|_1 + \frac{\lambda}{2}\|e_i\|^2 \quad s.t. \quad x_i^{(\ell)} = X_{-i}c_i + e_i. \tag{4.1}$$

The constraint can be rewritten as

$$y_i^{(\ell)} + z_i^{(\ell)} = (Y_{-i} + Z_{-i})c_i + e_i. \tag{4.2}$$

The dual program of (4.1) is:

$$\mathbf{D}_0: \quad \max_{\nu} \langle x_i, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu \quad s.t. \quad \|(X_{-i})^T\nu\|_\infty \le 1. \tag{4.3}$$

Recall that we want to establish the conditions on noise magnitude δ, structure of the data (μ and r in deterministic model and affinity in semi-random model) and ranges of valid λ such that by Definition 1, the solution c_i is *non-trivial* and has support indices inside the column set $X_{-i}^{(\ell)}$ (i.e., satisfies SEP).

We focus on the proof of Theorem 1 and 2 and briefly explain the randomized models. Indeed, Theorem 3 and 4 follow directly by plugging to Theorem 2 the bound of r and μ from Soltanolkotabi & Candes (2012) (with some modifications). The proof of Theorem 1 and 2 constitutes three main steps: (1) proving SEP, (2) proving non-trivialness, and (3) showing existence of proper λ.

4.1. Self-Expressiveness Property

We prove SEP by duality. First we establish a set of conditions on the optimal dual variable of D_0 corresponding to all primal solutions satisfying SEP. Then we construct such a dual variable ν as a certificate of proof.

4.1.1. Optimality Condition

Define general convex optimization:

$$\min_{c,e} \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \quad s.t. \quad x = Ac + e. \tag{4.4}$$

We state Lemma 1, which extends Lemma 7.1 in Soltanolkotabi & Candes (2012). The proof is deferred to the supplementary material.

Lemma 1. *Consider a vector $y \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d\times N}$. If there exists a triplet (c, e, ν) obeying $y = Ac + e$ and c has support $S \subseteq T$, furthermore the dual certificate vector ν satisfies*

$$A_s^T\nu = sgn(c_S), \qquad \nu = \lambda e,$$
$$\|A_{T\cap S^c}^T\nu\|_\infty \le 1, \quad \|A_{T^c}^T\nu\|_\infty < 1,$$

then any optimal solution (c^*, e^*) to (4.4) obeys $c^*_{T^c} = 0$.

The next step is to apply Lemma 1 with $x = x_i^{(\ell)}$ and $A = X_{-i}$ and then construct a triplet (c, e, ν) such that dual certificate ν satisfying all conditions and c satisfies SEP. Then we can conclude that all optimal solutions of (4.1) satisfy SEP.

4.1.2. Construction of Dual Certificate

To construct the dual certificate, we consider the following *fictitious* optimization problem that explicitly requires that all feasible solutions satisfy SEP[5] (note that one can not solve such problem in practice without knowing the subspace clusters).

$$\mathbf{P}_1: \quad \min_{c_i^{(\ell)},e_i} \|c_i\|_1 + \frac{\lambda}{2}\|e_i\|^2 \quad s.t. \quad y_i^{(\ell)} + z_i^{(\ell)} = (Y_{-i}^{(\ell)} + Z_{-i}^{(\ell)})c_i^{(\ell)} + e_i. \tag{4.5}$$

This problem is feasible. Moreover, it turns out that the dual solution of this fictitious problem v is a good candidate as our dual certificate. Observe that v automatically satisfies the first three conditions in Lemma 1 and we are left to show that for all data point $x \in \mathcal{X} \setminus \mathcal{X}^\ell$,

$$|\langle x, \nu\rangle| < 1. \tag{4.6}$$

Let ν_1 and ν_2 be the projection of ν to subspace $\mathcal{S}_\ell$ and its complement respectively. The strategy is to provide an upper bound of $|\langle x, \nu\rangle|$ then impose the inequality on the upper bound.

$$\begin{aligned}|\langle x, \nu\rangle| =& |\langle y + z, \nu\rangle| \le |\langle y, \nu_1\rangle| + |\langle y, \nu_2\rangle| + |\langle z, \nu\rangle| \\ \le& \mu(\mathcal{X}_\ell)\|\nu_1\| + \|y\|\|\nu_2\||\cos(\angle(y, \nu_2))| \\ &+ \|z\|\|\nu\||\cos(\angle(z, \nu))|.\end{aligned} \tag{4.7}$$

To complete the proof, we need to bound $\|\nu_1\|$ and $\|\nu_2\|$ and the two cosine terms (for random noise model). The proof makes use of the geometric properties of symmetric convex polytope and optimality of solution. See the supplementary material for the details.

[5] To be precise, it's the corresponding $c_i = [0, ..., 0, (c_i^{(\ell)})^T, 0, ..., 0]^T$ that satisfies SEP.

4.2. Non-trivialness and existence of λ

The idea is that when λ is large enough, trivial solution $c^* = 0$, $e^* = x_i^{(\ell)}$ can never occur. This is formalized by setting

$$\text{OptVal}(\mathbf{D}_0) = \langle x_i^{(\ell)}, \nu \rangle - \frac{1}{2\lambda}\|\nu\|^2 < \frac{\lambda}{2}\|x_i^{(\ell)}\|^2. \quad (4.8)$$

Notice that (4.8) essentially requires that $\lambda > A$ and (4.7) requires $\lambda < B$ for some A and B. Hence, existence of a valid λ requires $A < B$, which leads to the condition on the error magnitude $\delta < C$ and completes the proof. While conceptually straightforward, the details of the proof are involved and left in the supplementary material due to space constraints.

4.3. Randomization

Our randomized results consider two types of randomization: *random noise* and *random data*.

Random noise model improves the deterministic guarantee by exploiting the fact that the directions of the noise are random. By the well-known bound on the area of spherical cap (Lemma B.1), the cosine terms in (4.7) diminishes when the ambient dimension grows. Similar advantage also appears in the bound of $\|\nu_1\|$ and $\|\nu_2\|$ and the existence of λ.

Randomization of data provides probabilistic bounds of inradius r and incoherence μ. The lower bound of inradius r follows from a lemma in the study of isotropy constant of symmetric convex body (Alonso-Gutiérrez, 2008). The upper bound of $\mu(\mathcal{X}_{-i}^{(\ell)})$ requires more effort. It involves showing that projected dual directions $v_i^{(\ell)}$ (see Definition 2) distributes uniformly on the subspace projection of the unit n-sphere, then applying the spherical cap lemma for all pairs of $(v_i^{(\ell)}, y)$. We defer the full proof in the supplementary material.

5. Numerical simulation

To demonstrate the practical implications of our robustness guarantee for LASSO-SSC, we conduct three numerical experiments to test the effects of noise magnitude δ, subspace rank d and number of subspace L. To make it invariant to parameter, we scan through an exponential grid of λ ranging from $\sqrt{n} \times 10^{-2}$ to $\sqrt{n} \times 10^3$. In all experiments, ambient dimension $n = 100$, relative sampling $\kappa = 5$, subspace and data are drawn uniformly at random from unit sphere and then corrupted by Gaussian noise $Z_{ij} \sim N(0, \sigma/\sqrt{n})$. We measure the success of the algorithm by the relative violation of Self-Expressiveness Property defined below.

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

where $\mathcal{M}$ is the ground truth mask containing all (i, j) such that $x_i, x_j \in \mathcal{X}^{(\ell)}$ for some ℓ. Note that $\text{RelViolation}\,(C, \mathcal{M}) = 0$ implies that SEP is satisfied. We also check that there is no all-zero columns in C, and the solution is considered trivial otherwise.

The simulation results confirm our theoretical findings. In particular, Figure 5 shows that LASSO subspace detection property is possible for a very large range of λ and the dependence on noise magnitude is roughly $1/\sigma$ as remarked in (3.3). In addition, the sharp contrast of Figure 6 and 7 demonstrates precisely our observations on the sensitivity of d and L in Remark 7 and 8.

A remark on numerical algorithms: For fast computation, we use ADMM implementation of LASSO solver[6]. It has complexity proportional to problem size and convergence guarantee (Boyd et al., 2011). We also implement a simple solver for the matrix version SSC (3.1) which is consistently faster than the column-by-column LASSO version. Details of the algorithm and its favorable empirical comparisons are given in the supplementary materials.

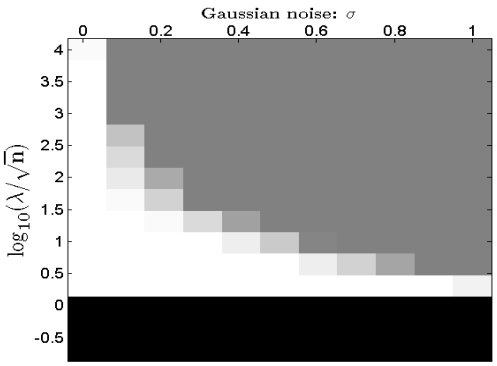

Figure 5. Exact recovery under noise. Simulated with $n = 100, d = 4, L = 3, \kappa = 5$ with increasing Gaussian noise $N(0, \sigma/\sqrt{n})$. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$.

6. Conclusion and future directions

We presented the first theoretical analysis for noisy subspace segmentation problem that is of great practical interests. We showed that the popular SSC algorithm *exactly* (not approximately) succeeds even in

[6]Freely available at:
http://www.stanford.edu/~boyd/papers/admm/

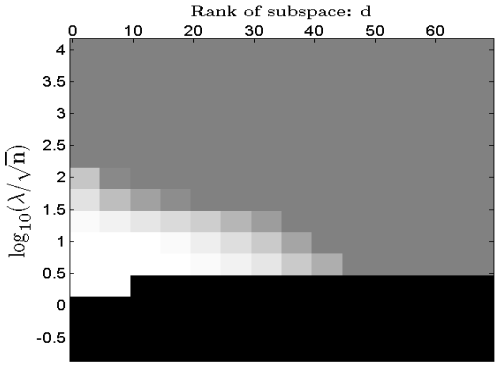

Figure 6. Effects of cluster rank d. Simulated with $n = 100, L = 3, \kappa = 5, \sigma = 0.2$ with increasing d. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$. Observe that beyond a point, subspace detection property is not possible for any λ.

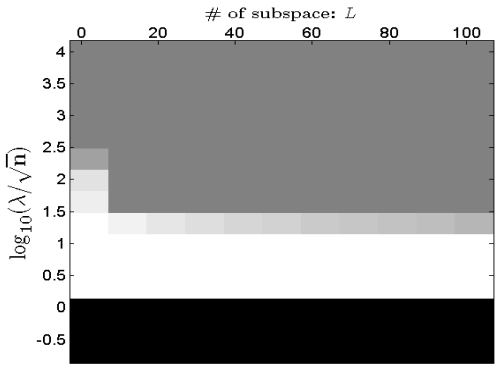

Figure 7. Effects of number of subspace L. Simulated with $n = 100, d = 2, \kappa = 5, \sigma = 0.2$ with increasing L. **Black:** trivial solution ($C = 0$); **Gray:** RelViolation > 0.1; **White:** RelViolation $= 0$. Note that even at the point when $dL = 200$(subspaces are highly dependent), subspace detection property holds for a large range of λ.

the noisy case, which justified its empirical success on real problems. In addition, we discovered a fundamental trade-off between robustness to noise and the subspace dimension, and we found that robustness is insensitivity to the number of subspaces. Our analysis hence reveals fundamental relationships of robustness, number of samples and dimension of the subspace. These results lead to new theoretical understanding of SSC, as well as provides guidelines for practitioners and application-level researchers to judge whether SSC could possibly work well for their respective applications.

Open problems for subspace clustering include the graph connectivity problem raised by Nasihatkon & Hartley (2011), missing data problem (a first attempt

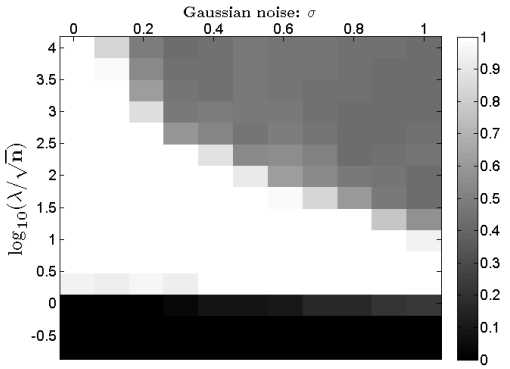

Figure 8. Spectral clustering accuracy for the experiment in Figure 5. The rate of accurate classification is represented in grayscale. White region means perfect classification. It is clear that exact subspace detection property (Definition 1) is not necessary for perfect classification.

by Eriksson et al. (2012), but requires an unrealistic number of data), sparse corruptions on data and others. One direction closely related to this paper is to introduce a more practical metric of success. As we illustrated in the paper, subspace detection property is not necessary for perfect clustering. In fact from a pragmatic point of view, even perfect clustering is not necessary. Typical applications allow for a small number of misclassifications. It would be interesting to see whether stronger robustness results can be obtained for a more practical metric of success.

Acknowledgments

H. Xu was supported by the Ministry of Education of Singapore through National University of Singapore startup Grant R-265-000-384-133.

References

Alonso-Gutiérrez, D. On the isotropy constant of random convex sets. *Proceedings of the American Mathematical Society*, 136(9):3293–3300, 2008.

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

Ben-Tal, A. and Nemirovski, A. Robust convex optimization. *Mathematics of Operations Research*, 23(4):769–805, 1998.

Bertsimas, D. and Sim, M. The price of robustness. *Operations research*, 52(1):35–53, 2004.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Bradley, P.S. and Mangasarian, O.L. k-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

Candès, E.J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Chen, G. and Lerman, G. Spectral curvature clustering (scc). *International Journal of Computer Vision*, 81(3): 317–330, 2009.

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

Donoho, D.L., Elad, M., and Temlyakov, V.N. Stable recovery of sparse overcomplete representations in the presence of noise. *Information Theory, IEEE Transactions on*, 52(1):6–18, 2006.

Elhamifar, E. and Vidal, R. Sparse subspace clustering. In *CVPR'09*, pp. 2790–2797. IEEE, 2009.

Elhamifar, E. and Vidal, R. Clustering disjoint subspaces via sparse representation. In *ICASSP'11*, pp. 1926–1929. IEEE, 2010.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

Eriksson, B., Balzano, L., and Nowak, R. High rank matrix completion. In *AI Stats'12*, 2012.

Hastie, T. and Simard, P.Y. Metrics and models for handwritten character recognition. *Statistical Science*, pp. 54–65, 1998.

Jalali, A., Chen, Y., Sanghavi, S., and Xu, H. Clustering partially observed graphs via convex optimization. In *ICML'11*, pp. 1001–1008. ACM, 2011.

Kanatani, K. Motion segmentation by subspace separation and model selection. In *ICCV'01*, volume 2, pp. 586–591. IEEE, 2001.

Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., and Ma, Y. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

Loh, P.L. and Wainwright, M.J. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

Nasihatkon, B. and Hartley, R. Graph connectivity in sparse subspace clustering. In *CVPR'11*, pp. 2137–2144. IEEE, 2011.

Ng, A.Y., Jordan, M.I., Weiss, Y., et al. On spectral clustering: Analysis and an algorithm. In *NIPS'02*, volume 2, pp. 849–856, 2002.

Soltanolkotabi, M. and Candes, E.J. A geometric analysis of subspace clustering with outliers. *To appear in Annals of Statistics*, 2012.

Tron, R. and Vidal, R. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR'07*, pp. 1–8. IEEE, 2007.

Vidal, R. Subspace clustering. *Signal Processing Magazine, IEEE*, 28(2):52–68, 2011.

Vidal, R., Ma, Y., and Sastry, S. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

Zhang, A., Fawaz, N., Ioannidis, S., and Montanari, A. Guess who rated this movie: Identifying users through subspace clustering. *arXiv preprint arXiv:1208.1544*, 2012.

Zhou, S.K., Aggarwal, G., Chellappa, R., and Jacobs, D.W. Appearance characterization of linear lambertian objects, generalized photometric stereo, and illumination-invariant face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (2):230–245, 2007.