# Theoretical Analysis of Sparse Subspace Clustering with Missing Entries

**Manolis C. Tsakiris** [*1] **René Vidal** [*2]

## Abstract

Sparse Subspace Clustering (SSC) is a popular unsupervised machine learning method for clustering data lying close to an unknown union of low-dimensional linear subspaces; a problem with numerous applications in pattern recognition and computer vision. Even though the behavior of SSC for complete data is by now well-understood, little is known about its theoretical properties when applied to data with missing entries. In this paper we give theoretical guarantees for SSC with incomplete data, and provide theoretical evidence that projecting the zero-filled data onto the observation pattern of the point being expressed can lead to substantial improvement in performance; a phenomenon already known experimentally. The main insight of our analysis is that even though this projection induces additional missing entries, this is counterbalanced by the fact that the projected and zero-filled data are in effect incomplete points associated with the union of the corresponding projected subspaces, with respect to which the point being expressed is complete. The significance of this phenomenon potentially extends to the entire class of self-expressive methods.

## 1. INTRODUCTION

Clustering data lying close to an unknown union of low-dimensional linear subspaces is a fundamental problem in unsupervised machine learning, known as *Subspace Clustering* or *Generalized Principal Component Analysis* (Vidal et al., 2016). Indeed, this problem is intimately related to the extension of the classical Principal Component Analysis (PCA) to multiple subspaces, and in recent years has found numerous applications in machine learning, computer vision, pattern recognition, bioinformatics and systems theory. Moreover, recent work is beginning to explore connections between subspace clustering and deep learning, with the goal of learning unions of low-dimensional non-linear manifolds (Peng et al., 2016).

Among a variety of subspace clustering methods (Vidal et al., 2016) including algebraic (Vidal et al., 2005; Tsakiris & Vidal, 2017b; 2018a), iterative (Bradley & Mangasarian, 2000), recursive (Fischler & Bolles, 1981; Tsakiris & Vidal, 2017a), and spectral (Aldroubi et al., 2017; Heckel & Bölcskei, 2015; Lu et al., 2012; Chen & Lerman, 2009) techniques, Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2009; 2013) is one of the most popular methods. The reason is that it exhibits a very competitive performance in real-world datasets, it admits efficient algorithmic implementations, and is supported by a rich body of theory (Elhamifar & Vidal, 2013; Soltanolkotabi & Candès, 2012; Wang & Xu, 2016; Soltanolkotabi et al., 2014). In addition, SSC is able to cluster data from incomplete observations reasonably well (Yang et al., 2015), which is an important problem (Ongie et al., 2017; Pimentel-Alarcon & Nowak, 2016; Elhamifar, 2016; Yang et al., 2015; Heckel & Bölcskei, 2015; Pimentel-Alarcon et al., 2015; Eriksson et al., 2012; Recht, 2011; Balzano et al., 2010), since in many applications not all features are available for every data point: Users of recommendation systems only rate a few items, medical patients undergo a few tests and treatments, images are corrupted by occlusions, dynamic processes are observed across short time intervals and so on.

Even though the theoretical foundations of SSC are by now mature, there are many lingering open questions. For example, it is still unclear whether better conditions exist for the performance of SSC even for uncorrupted data; contrast this to the recent study of You & Vidal (2015), who establish a hierarchy of such conditions for sparse subspace recovery. More importantly, even though a satisfactory theory for SSC with general noise does exist (Wang & Xu, 2016), the theoretical properties of SSC for data with missing entries remain elusive. The works of Wang et al. (2016) and Charles et al. (2018) are important recent efforts towards understanding SSC with missing entries. However, the conditions of Wang et al. (2016) are hard to interpret and they refer to the formulation of SSC with exact self-

[1]School of Information Science and Technology, ShanghaiTech University, Shanghai, China. [2]Mathematical Institute for Data Science and Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA. Correspondence to: Manolis C. Tsakiris <mtsakiris@shanghaitech.edu.cn>.

*Proceedings of the $35^{th}$ International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018. 

expressiveness equality constraint, which is not an optimal choice for corrupted data. On the other hand, following Wang & Xu (2016), Charles et al. (2018) provide bounds similar to a subset of the results in the present paper.[1]

In this paper we provide a novel theoretical analysis of SSC for incomplete data. More precisely, we provide theoretical performance guarantees for SSC applied to i) *Zero-Filled* data (ZF-SSC), in which case all unobserved entries are filled with zeros, and ii) *Projected-Zero-Filled* data (PZF-SSC), in which case all unobserved entries are filled with zeros and in addition all data points are projected onto the observation pattern of the point being expressed each time.[2]. A direct comparison of the tolerable bounds of missing entries for ZF-SSC (Theorem 7) and PZF-SSC (Theorem 5) serves as a theoretical indication for the latter being a better method than the former. This is in agreement with experimental evaluation given here and also previously reported by Yang et al. (2015). Since PZF data have in principle many more missing entries than ZF data, this is a remarkable phenomenon, of potentially wider significance to the entire class of self-expressive-based methods, e.g., (Liu et al., 2013; Lu et al., 2012; Elhamifar & Vidal, 2013; Wang et al., 2013; You et al., 2016).

The rest of the paper is organized as follows. In §1.1 we introduce the notation and the main mathematical objects of this paper. In §2 we review SSC for uncorrupted data, and discuss the two known elementary formulations of SSC for incomplete data, i.e., ZF-SSC and PZF-SSC. In §3 we present the main contributions of this paper, which consist of deterministic and probabilistic characterizations of the tolerable percentage of missing entries for ZF-SSC and PZF-SSC, as well as a theoretical and experimental comparison between the two methods (all proofs can be found in our pre-print (Tsakiris & Vidal, 2018b)). We conclude in §4, where we discuss the main insights of this paper as well as existing challenges.

### 1.1. Notation and Main Objects

The nature of the problem studied in this paper calls for a rather heavy notation, which we have strived to simplify and unify as much as possible. To avoid introducing complicated notation amidst other technical developments, we have found it convenient to gather all relevant objects in Definition 1,[3] which the reader is encouraged to refer to when necessary. Other than that, for $\ell$ a positive integer, we define $[\ell] := \{1, \dots, \ell\}$. For a vector $\boldsymbol{w} \in \mathbb{R}^D$ we define $\hat{\boldsymbol{w}} := \boldsymbol{w}/\|\boldsymbol{w}\|_2$, if $\boldsymbol{w} \neq \boldsymbol{0}$ and $\hat{\boldsymbol{w}} := \boldsymbol{0}$, otherwise. For any linear subspace $\mathcal{V}$ of $\mathbb{R}^D$, we denote by $\boldsymbol{P}_{\mathcal{V}}$ the square matrix that represents the orthogonal projection of $\mathbb{R}^D$ onto $\mathcal{V}$. Given a binary relation, RHS stands for *Right-Hand-Side*, and similarly for LHS. Finally, $\langle \cdot, \cdot \rangle$ is the standard inner product of $\mathbb{R}^D$.

**Definition 1.** *We define the following objects:*

1. **The linear subspaces:** *For $i \in [n]$, we let $\mathcal{S}_i$ be a linear subspace of $\mathbb{R}^D$, where $\dim \mathcal{S}_i = d_i < D$.*

2. **The complete data:** *With an abuse of notation we let*

$$\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \dots, \boldsymbol{X}^{(n)}]\boldsymbol{\Gamma} \in \mathbb{R}^{D \times N} \tag{1}$$

*denote a data matrix as well as a set (formed by the columns of this matrix) of unit $\ell_2$-norm points in the union of the linear subspaces $\mathcal{S}_i$, $i \in [n]$, where $\boldsymbol{X}^{(i)} = [\boldsymbol{x}_1^{(i)}, \dots, \boldsymbol{x}_{N_i}^{(i)}] \subset \mathcal{S}_i$, $\mathrm{Span}(\boldsymbol{X}^{(i)}) = \mathcal{S}_i$, and $\boldsymbol{\Gamma}$ is an unknown permutation, indicating that the clustering of the points with respect to the subspaces is unknown. We define $\boldsymbol{X}_{-1}^{(1)} := \boldsymbol{X}^{(1)} \setminus \{\boldsymbol{x}_1^{(1)}\}$, $\boldsymbol{X}_{-1} := \boldsymbol{X} \setminus \{\boldsymbol{x}_1^{(1)}\}$, and $\boldsymbol{X}^{(-1)} := \boldsymbol{X} \setminus \boldsymbol{X}^{(1)}$, where $\setminus$ denotes set-theoretic difference.*

3. **The pattern of missing entries:** *For every point $\boldsymbol{x}_j^{(i)} \in \mathbb{R}^D$ we consider an observation pattern $\boldsymbol{\omega}_j^{(i)} \in \{0,1\}^D$, where a value of $1$ indicates an observed entry, while a value of $0$ indicates an unobserved entry. We assume each $\boldsymbol{\omega}_j^{(i)}$ has precisely $m$ zeros. We let $\tilde{\boldsymbol{\omega}}_j^{(i)} := \boldsymbol{1} - \boldsymbol{\omega}_j^{(i)}$, where $\boldsymbol{1}$ is the vector of all ones.*

4. **The observed/unobserved coordinate subspaces:** *We let $\bar{\mathcal{E}}_j^{(i)} := \mathrm{Span}\{\boldsymbol{e}_k : \boldsymbol{e}_k^\top \boldsymbol{\omega}_j^{(i)} \neq 0\}$, with $\boldsymbol{e}_k$ the canonical vector of $\mathbb{R}^D$ with zeros everywhere and a $1$ at position $k$. The orthogonal projection onto $\bar{\mathcal{E}}_j^{(i)}$ is given by $\bar{\boldsymbol{P}}_j^{(i)} := \mathrm{diag}(\boldsymbol{\omega}_j^{(i)})$, the matrix with $\boldsymbol{\omega}_j^{(i)}$ on its diagonal and zeros everywhere else. $\tilde{\mathcal{E}}_j^{(i)}$ is the orthogonal complement of $\bar{\mathcal{E}}_j^{(i)}$ and $\tilde{\boldsymbol{P}}_j^{(i)} = \mathrm{diag}(\tilde{\boldsymbol{\omega}}_j^{(i)})$ is the orthogonal projection onto $\tilde{\mathcal{E}}_j^{(i)}$.*

5. **The zero-filled data (ZF-data):** *We let $\bar{\boldsymbol{X}} \in \mathbb{R}^{D \times N}$ be the data $\boldsymbol{X}$ with zeros appearing in the unobserved entries, i.e., the column of $\bar{\boldsymbol{X}}$ associated to point $\boldsymbol{x}_j^{(i)}$ is $\bar{\boldsymbol{x}}_j^{(i)} := \bar{\boldsymbol{P}}_j^{(i)} \boldsymbol{x}_j^{(i)}$, $\forall i, j$.*

6. **The projected data:** *We let $\dot{\boldsymbol{X}} := \bar{\boldsymbol{P}}_1^{(1)} \boldsymbol{X}$ be the projection of the data $\boldsymbol{X}$ onto the observed coordinate subspace $\bar{\mathcal{E}}_1^{(1)}$ associated to point $\boldsymbol{x}_1^{(1)}$. The column of $\dot{\boldsymbol{X}}$ associated to $\boldsymbol{x}_j^{(i)}$ is $\dot{\boldsymbol{x}}_j^{(i)} := \bar{\boldsymbol{P}}_1^{(1)} \boldsymbol{x}_j^{(i)}$, $\forall i, j$.*

[1] In the terminology of the present paper Charles et al. (2018) independently study ZF-SSC.

[2] This is called *EWZF-SSC* by Yang et al. (2015); here we have taken the liberty to rename the method according to the more suggestive name PZF-SSC.

[3] For simplicity and clarity, and without loss of generality, we have chosen to present our theoretical results in the context of expressing a single point in terms of the remaining points in the dataset (the precise problem formulation is deferred to §2).

7. **The projected and zero-filled data (PZF-data):** *We let $\dot{\bar{\boldsymbol{X}}}$ be the projection of the zero-filled data onto $\bar{\mathcal{E}}_1^{(1)}$, i.e., $\dot{\bar{\boldsymbol{X}}} := \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}$. The column of $\dot{\bar{\boldsymbol{X}}}$ associated to point $\boldsymbol{x}_j^{(i)}$ is $\dot{\bar{\boldsymbol{x}}}_j^{(i)} := \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{x}}_j^{(i)}$, $\forall i, j$.*

8. **The unobserved data:** *We define $\tilde{\boldsymbol{X}}$ to be the unobserved components of the data, i.e., $\tilde{\boldsymbol{X}} := \boldsymbol{X} - \bar{\boldsymbol{X}}$, and $\tilde{\boldsymbol{x}}_j^{(i)} := \tilde{\boldsymbol{P}}_j^{(i)}\boldsymbol{x}_j^{(i)}$, $\forall i, j$. Similarly, for PZF data we define $\dot{\tilde{\boldsymbol{X}}} := \dot{\boldsymbol{X}} - \dot{\bar{\boldsymbol{X}}}$, and $\dot{\tilde{\boldsymbol{x}}}_j^{(i)} := \bar{\boldsymbol{P}}_1^{(1)}\tilde{\boldsymbol{x}}_j^{(i)}$, $\forall i, j$.*

9. **The projected subspaces:** *For $i \in [n]$, we let $\dot{\mathcal{S}}_i \subset \mathbb{R}^D$ be the orthogonal projection of $\mathcal{S}_i$ onto the subspace $\bar{\mathcal{E}}_1^{(1)}$. In other words, if $\boldsymbol{b}_1^{(i)}, \ldots, \boldsymbol{b}_{d_i}^{(i)}$ is a basis for $\mathcal{S}_i$, then $\dot{\mathcal{S}}_i$ is the subspace of $\mathbb{R}^D$ spanned by the vectors $\bar{\boldsymbol{P}}_1^{(1)}\boldsymbol{b}_k^{(i)}, \forall k \in [d_i]$.*

10. **The inradius:** *We let $r$ be the relative inradius of the symmetrized convex hull $\mathcal{Q}$ of all points $\boldsymbol{X}_{-1}^{(1)}$ lying in subspace $\mathcal{S}_1$, except point $\boldsymbol{x}_1^{(1)}$, i.e., $r$ is the radius of the largest Euclidean ball of $\mathcal{S}_1$ contained in $\mathcal{Q}$.*

11. **The dual directions:** *For $\boldsymbol{W} = \boldsymbol{X}, \bar{\boldsymbol{X}}, \dot{\bar{\boldsymbol{X}}}$ corresponding to complete data $\boldsymbol{X}$, ZF-data $\bar{\boldsymbol{X}}$ and PZF-data $\dot{\bar{\boldsymbol{X}}}$, consider the reduced Lasso-SSC problem*

$$\min_{\boldsymbol{c},\boldsymbol{e}} \|\boldsymbol{c}\|_1 + \frac{\lambda}{2}\|\boldsymbol{e}\|_2^2 \;\; s.t. \;\; \boldsymbol{w}_1^{(1)} = \boldsymbol{W}_{-1}^{(1)}\boldsymbol{c} + \boldsymbol{e}, \quad (2)$$

*corresponding to either complete data $\boldsymbol{X}$, ZF-data $\bar{\boldsymbol{X}}$ or PZF-data $\dot{\bar{\boldsymbol{X}}}$. Consider the dual problem*

$$\max_{\boldsymbol{v}} \langle \boldsymbol{v}, \boldsymbol{w}_1^{(1)} \rangle - \frac{1}{2\lambda}\|\boldsymbol{v}\|_2^2 \;\; s.t. \;\; \|\boldsymbol{v}^\top \boldsymbol{W}_{-1}^{(1)}\|_\infty \leq 1. \quad (3)$$

*Let $\boldsymbol{v}_\lambda^*, \bar{\boldsymbol{v}}_\lambda^*, \dot{\bar{\boldsymbol{v}}}_\lambda^*$ be the optimal solution to problem (3) corresponding to $\boldsymbol{W} = \boldsymbol{X}, \bar{\boldsymbol{X}}, \dot{\bar{\boldsymbol{X}}}$ respectively; these solutions are unique because (3) is strongly convex. Then we define the corresponding dual directions $\hat{\boldsymbol{v}}_{1,\lambda}, \hat{\bar{\boldsymbol{v}}}_{1,\lambda}, \hat{\dot{\bar{\boldsymbol{v}}}}_{1,\lambda}$ to be the normalized projections of $\boldsymbol{v}_\lambda^*, \bar{\boldsymbol{v}}_\lambda^*, \dot{\bar{\boldsymbol{v}}}_\lambda^*$ onto $\mathcal{S}_1, \mathcal{S}_1, \dot{\mathcal{S}}_1$ respectively (if any of these projections is equal to zero, then we define the corresponding dual direction to be the zero vector).*

12. **The inter-subspace coherences:** *We define the inter-subspace coherences for complete data, ZF-data, and PZF-data respectively as*

$$\mu_\lambda := \max_{i>1,\, k\in[N_i]} |\langle \boldsymbol{x}_k^{(i)}, \hat{\boldsymbol{v}}_{1,\lambda} \rangle| \quad (4)$$

$$\bar{\mu}_\lambda := \max_{i>1,\, k\in[N_i]} |\langle \bar{\boldsymbol{x}}_k^{(i)}, \hat{\bar{\boldsymbol{v}}}_{1,\lambda} \rangle| \quad (5)$$

$$\dot{\bar{\mu}}_\lambda := \max_{i>1,\, k\in[N_i]} |\langle \dot{\bar{\boldsymbol{x}}}_k^{(i)}, \hat{\dot{\bar{\boldsymbol{v}}}}_{1,\lambda} \rangle|. \quad (6)$$

13. **The intra-subspace coherences:**

$$\zeta := \|(\boldsymbol{X}_{-1}^{(1)})^\top \boldsymbol{x}_1^{(1)}\|_\infty, \quad (7)$$

$$\bar{\zeta} := \|(\bar{\boldsymbol{X}}_{-1}^{(1)})^\top \bar{\boldsymbol{x}}_1^{(1)}\|_\infty, \quad (8)$$

$$\dot{\bar{\zeta}} := \|(\dot{\bar{\boldsymbol{X}}}_{-1}^{(1)})^\top \dot{\bar{\boldsymbol{x}}}_1^{(1)}\|_\infty, \;\; (\bar{\zeta} = \dot{\bar{\zeta}}) \quad (9)$$

14. **Other quantities:**

$$\bar{\eta} := \|\bar{\boldsymbol{x}}_1^{(1)}\|_2, \quad (10)$$

$$\dot{\bar{\eta}} := \|\dot{\bar{\boldsymbol{x}}}_1^{(1)}\|_2, \;\; (\bar{\eta} = \dot{\bar{\eta}}) \quad (11)$$

$$\bar{\gamma} := \max_{\substack{i>1, k\in[N_i],\\ j\in[N_1]}} |\langle \bar{\boldsymbol{x}}_k^{(i)}, \boldsymbol{P}_{\mathcal{S}_1^\perp}\tilde{\boldsymbol{x}}_j^{(1)} \rangle| \quad (12)$$

$$\dot{\bar{\gamma}} := \max_{\substack{i>1, k\in[N_i],\\ j\in[N_1]}} |\langle \dot{\bar{\boldsymbol{x}}}_k^{(i)}, \boldsymbol{P}_{\dot{\mathcal{S}}_1^\perp}\dot{\tilde{\boldsymbol{x}}}_j^{(1)} \rangle|. \quad (13)$$

## 2. Review of Sparse Subspace Clustering

We begin by reviewing Sparse Subspace Clustering (SSC) for data with no corruptions (§2.1), as well as the two elementary approaches to SSC for incomplete data (§2.2), which this paper is devoted to analyzing.

### 2.1. SSC With Uncorrupted Data

In the absence of data corruptions (noise, missing entries, outliers, etc.) we consider a data matrix $\boldsymbol{X} \in \mathbb{R}^{D\times N}$ as in Definition 1, whose columns are unit-$\ell_2$ points[4] that lie in an unknown union of low-dimensional linear subspaces $\bigcup_{i=1}^n \mathcal{S}_i \subset \mathbb{R}^D$, with $d_i := \dim(\mathcal{S}_i)$. Thus $\boldsymbol{X} = [\boldsymbol{X}^{(1)} \cdots \boldsymbol{X}^{(n)}]\boldsymbol{\Gamma}$, where each $\boldsymbol{X}^{(i)} := [\boldsymbol{x}_1^{(i)} \cdots \boldsymbol{x}_{N_i}^{(i)}] \in \mathbb{R}^{D\times N_i}$ consists of $N_i$ points spanning subspace $\mathcal{S}_i$, and $\boldsymbol{\Gamma}$ is an unknown permutation, indicating that the clustering of the points is unknown.

Among a variety of methods (Vidal et al., 2016) for retrieving the clusters $\{\boldsymbol{X}^{(i)}\}$, one may apply Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2009; 2013), whose main principle is to express each point in $\boldsymbol{X}$ as a sparse linear combination of other points in $\boldsymbol{X}$. Specifically, we seek an expression, say, of point $\boldsymbol{x}_1^{(1)}$ as a sparse linear combination of all other points $\boldsymbol{X}_{-1} := \boldsymbol{X} \setminus \{\boldsymbol{x}_1^{(1)}\}$ by means of the basis pursuit problem (Chen et al., 1998)

$$\min_{\boldsymbol{c}\in\mathbb{R}^{N-1}} \;\; \|\boldsymbol{c}\|_1 \;\; \text{s.t. } \boldsymbol{x}_1^{(1)} = \boldsymbol{X}_{-1}\boldsymbol{c}, \quad (14)$$

and then form an affinity graph in which we connect $\boldsymbol{x}_1^{(1)}$ to those points of $\boldsymbol{X}_{-1}$ that correspond to the support (nonzero coefficients) of the computed optimal solution of (14). Clearly, we want these points to lie in the same subspace as $\boldsymbol{x}_1^{(1)}$, i.e., to be points of $\boldsymbol{X}_{-1}^{(1)} := \boldsymbol{X}^{(1)} \setminus \{\boldsymbol{x}_1^{(1)}\}$, in which

[4] This assumption simplifies the theoretical analysis.

case we say that the solution is *subspace preserving*. When this is true for the expression of each and every point in $\boldsymbol{X}$, then the corresponding affinity graph contains no connections between points in different subspaces, i.e., it is a subspace preserving graph. Assuming that points within each subspace are sufficiently well connected, the affinity graph will have precisely $n$ connected components, and spectral clustering will be guaranteed to furnish the correct clusters.

Often, it is more practical to search for approximate sparse linear combinations rather exact ones as in (14). Thus one may approximately express point $\boldsymbol{x}_1^{(1)}$ by solving the Lasso problem (Tibshirani, 2013)

$$\min_{\boldsymbol{c},\boldsymbol{e}} \ \|\boldsymbol{c}\|_1 + \frac{\lambda}{2}\|\boldsymbol{e}\|_2^2 \ \text{ s.t. } \boldsymbol{x}_1^{(1)} = \boldsymbol{X}_{-1}\boldsymbol{c} + \boldsymbol{e}, \qquad (15)$$

where $\boldsymbol{e}$ represents the self-representation error. We have the following known guarantee:

**Theorem 1** (SSC with uncorrupted data, deterministic (Wang & Xu, 2016))**.** *Recall the notation of Definition 1, and suppose that*

$$\mu_\lambda < r \ \text{ and } \ 1/\zeta < \lambda. \qquad (16)$$

*Then every optimal solution to the Lasso SSC problem* (15) *is non-zero and subspace preserving.*

Theorem 1 can be interpreted as follows: If all data points from $\mathcal{S}_1$ other than $\boldsymbol{x}_1^{(1)}$ are well distributed (large $r$), the data points from other subspaces are sufficiently far from $\mathcal{S}_1$ as measured by their inner product with the dual direction $\hat{\boldsymbol{v}}_{1,\lambda}$ (small $\mu_\lambda$), and the reconstruction error is penalized sufficiently enough (large $\lambda$), then the Lasso problem (15) is guaranteed to furnish non-zero and subspace preserving solutions.

Theorem 2 is an even more interpretable statement and is derived by bounding in probability the terms in Theorem 1 under the following simplified fully random model.

**Definition 2** (Random model)**.** *For each $i \in [n]$, let the $i$th subspace be chosen uniformly at random from the Grassmannian manifold of $d$-dimensional subspaces of $\mathbb{R}^D$. Moreover, let $N/n =: \rho d + 1$ points*[5] *be chosen uniformly at random from the intersection of each subspace and the unit sphere $\mathbb{S}^{D-1}$. Finally, define the quantities*

$$\alpha := \sqrt{\frac{\log(\rho)}{16d}}, \ \ \beta := \sqrt{\frac{6\log(N)}{D}}. \qquad (17)$$

**Theorem 2** (SSC with uncorrupted data, probabilistic (Soltanolkotabi & Candès, 2012; Wang & Xu, 2016))**.** *Consider the random model of Definition 2. If $\rho$ is larger than a universal constant, $\lambda > 1/\alpha$, and*

$$\alpha > \beta, \qquad (18)$$

*then any optimal solution to the Lasso SSC problem* (15) *is non-zero and subspace preserving, with probability at least $1 - 2/N^2 - \exp(-\sqrt{\rho}d)$.*

Condition (18) agrees with intuition, since it effectively says that the subspace preserving property is easier to achieve for small relative subspace dimensions $d/D$, fewer subspaces, and more points per subspace. In §3 we will give analogues of Theorems 1 and 2 for two elementary variants of SSC for incomplete data, described next.

### 2.2. SSC With Missing Entries (ZF-SSC, PZF-SSC)

When the data are incomplete but otherwise uncorrupted, one may consider using a low-rank matrix completion algorithm to first complete the data and then apply SSC to the completed data. However, this procedure is guaranteed to succeed only when the underlying complete matrix $\boldsymbol{X}$ is of low rank and sufficiently *incoherent* (Candès & Recht, 2009; Recht, 2011), an assumption which might become invalid in the presence of data from many distinct subspaces. As a simple alternative, one may fill with zeros the unobserved entries to obtain a zero-filled data matrix $\bar{\boldsymbol{X}}$ exactly as in Definition 1, and subsequently solve the problem

$$\min_{\boldsymbol{c},\boldsymbol{e}} \ \|\boldsymbol{c}\|_1 + \frac{\lambda}{2}\|\boldsymbol{e}\|_2^2 \ \text{ s.t. } \bar{\boldsymbol{x}}_1^{(1)} = \bar{\boldsymbol{X}}_{-1}\boldsymbol{c} + \boldsymbol{e}, \qquad (19)$$

a procedure called *Zero-Filled SSC* (ZF-SSC) (Yang et al., 2015). In spite of its simplicity (after all we are just filling in the missing entries with zeros), as per Figs. 2(a) and 2(c) in Yang et al. (2015), ZF-SSC performs only slightly worse than low-rank matrix completion followed by SSC.

Even so, ZF-SSC has an evident shortcoming: it penalizes the reconstruction error of the zero vector along the unobserved part of the point being expressed, which is clearly an undesirable feature of the method. More precisely, letting $\bar{\mathcal{E}}_1^{(1)}$ and $\tilde{\mathcal{E}}_1^{(1)}$ be, respectively, the observed and unobserved subspaces associated to point $\boldsymbol{x}_1^{(1)}$, and $\bar{\boldsymbol{P}}_1^{(1)}$, $\tilde{\boldsymbol{P}}_1^{(1)}$ the orthogonal projections onto them (see Definition 1), and recalling that $(\bar{\mathcal{E}}_1^{(1)})^\perp = \tilde{\mathcal{E}}_1^{(1)}$, we have that

$$\bar{\boldsymbol{x}}_1^{(1)} = \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{x}}_1^{(1)}, \ \text{ and} \qquad (20)$$

$$\bar{\boldsymbol{X}}_{-1} = \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}_{-1} + \tilde{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}_{-1}, \qquad (21)$$

and so we can rewrite the objective function of ZF-SSC as

$$\|\boldsymbol{c}\|_1 + \frac{\lambda}{2}\|\bar{\boldsymbol{x}}_1^{(1)} - \bar{\boldsymbol{X}}_{-1}\boldsymbol{c}\|_2^2 = \|\boldsymbol{c}\|_1 + \qquad (22)$$

$$\frac{\lambda}{2}\|\bar{\boldsymbol{x}}_1^{(1)} - \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}_{-1}\boldsymbol{c}\|_2^2 + \frac{\lambda}{2}\|\tilde{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}_{-1}\boldsymbol{c}\|_2^2. \qquad (23)$$

We then see that ZF-SSC penalizes the reconstruction error $\|\bar{\boldsymbol{x}}_1^{(1)} - \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}_{-1}\boldsymbol{c}\|_2$ of the observed part of $\boldsymbol{x}_1^{(1)}$, which

[5] For simplicity, we assume that $n$ divides $N$.

is desirable, as well as the norm of the vector $\tilde{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}_{-1}\boldsymbol{c}$. The latter is an artifact of the zero-filling process, and could bias the coefficients $\boldsymbol{c}$ away from a subspace preserving pattern. Thus, it is reasonable to remove this term and obtain self-expressive coefficients for $\bar{\boldsymbol{x}}_1^{(1)}$ by solving instead

$$\min_{\boldsymbol{c},\boldsymbol{e}} \ \|\boldsymbol{c}\|_1 + \frac{\lambda}{2}\|\boldsymbol{e}\|_2^2, \ \text{ s.t. } \boldsymbol{e} = \bar{\boldsymbol{x}}_1^{(1)} - \dot{\bar{\boldsymbol{X}}}_{-1}\boldsymbol{c}, \qquad (24)$$

where $\dot{\bar{\boldsymbol{X}}} := \bar{\boldsymbol{P}}_1^{(1)}\bar{\boldsymbol{X}}$ is the projected and zero-filled data, as in Definition 1. Yang et al. (2015) called this approach *EWZF-SSC*; here we take the liberty to rename it *Projected-Zero-Filled Sparse-Subspace-Clustering (PZF-SSC)*.

PZF-SSC is known to provide accurate clustering while tolerating a higher percentage of missing entries than ZF-SSC or even low-rank matrix completion followed by SSC (e.g., see Fig. 2 in Yang et al. (2015)). This is rather fascinating, since, after all, PZF-SSC works with the projected and zero-filled data $\dot{\bar{\boldsymbol{X}}}$, which have more missing entries than the zero-filled data $\bar{\boldsymbol{X}}$. Because of this reason, direct application of any generic noise bound, such as that of Theorem 6 in Wang & Xu (2016), would naively suggest that ZF-SSC tolerates more missing entries than PZF-SSC, contradicting experimental evidence. This apparent *mystery* is resolved in §3, where we adopt a more sophisticated view of PZF-SSC, which unveils its advantage over ZF-SSC.

## 3. SSC Theory for Incomplete Data

This section contains the main contributions of this paper. In §3.1-3.2 we give deterministic and probabilistic theorems of correctness for PZF-SSC and ZF-SSC, respectively, in analogy with Theorems 1-2 for SSC with uncorrupted data, while in §3.3 we discuss how the conditions for the two methods compare.

### 3.1. PZF-SSC Theory

As already remarked so far, PZF-SSC is experimentally known to be a superior method to ZF-SSC, i.e., it can provide an accurate clustering for a higher percentage of missing entries. This is remarkable, because the projected and zero-filled data $\dot{\bar{\boldsymbol{X}}}$ (see Definition 1 for notation) that PZF-SSC operates on contain more missing entries than the zero-filled data $\bar{\boldsymbol{X}}$ that ZF-SSC operates on. On the other hand, we already saw in §2.2 that the additional zeros in $\dot{\bar{\boldsymbol{X}}}$ are inflicted in such a way, that the objective function minimized by PZF-SSC is, at least on an intuitive level, more accurate than the one minimized by ZF-SSC.

In this paper we give a theoretical justification for the superiority of PZF-SSC over ZF-SSC. Our main insight is the following observation: expressing point $\bar{\boldsymbol{x}}_1^{(1)} = \dot{\bar{\boldsymbol{x}}}_1^{(1)}$ as a sparse linear combination of $\dot{\bar{\boldsymbol{X}}}_{-1}$, can be seen as expressing the *complete* point $\bar{\boldsymbol{x}}_1^{(1)}$ from partial observations $\dot{\bar{\boldsymbol{X}}}_{-1}$ of the *complete* points $\dot{\boldsymbol{X}}_{-1}$, where now the underlying *complete* data $\dot{\boldsymbol{X}}$ lie in the union of subspaces $\bigcup_{i=1}^n \dot{\mathcal{S}}_i$, i.e., the original subspaces projected onto the coordinate subspace defined by the observation pattern of the point being expressed (see Definition 1). With this in mind, inspired by the seminal work of Wang & Xu (2016), and by

1. making more frequent use of strong duality than in the proof of Theorem 6 in Wang & Xu (2016),
2. using a novel bound for the norm of the *dual vector*,
3. and not decoupling the *noise* from the data,[6]

we arrive at the following key result:

**Theorem 3** (PZF-SSC, deterministic)**.** *With the notation of Definition 1, further define the positive quantity*

$$\dot{\lambda}^* := \frac{1}{2}\left\{\frac{1}{2\dot{\zeta}} - \frac{\dot{\mu}_\lambda}{\dot{\gamma}\dot{\eta}} + \sqrt{\frac{9}{4\dot{\zeta}^2} + \frac{\dot{\mu}_\lambda}{\dot{\gamma}\dot{\eta}\dot{\zeta}} + \frac{2}{\dot{\gamma}\dot{\eta}^2} + \frac{\dot{\mu}_\lambda^2}{\dot{\gamma}^2\dot{\eta}^2}}\right\}. \qquad (25)$$

*Then the interval* $\dot{\Lambda} := (1/\dot{\zeta}, \dot{\lambda}^*)$ *is non-empty, if*

$$\dot{\mu}_\lambda\, \dot{\eta} \ < \dot{\zeta}. \qquad (26)$$

*If in addition*[7] $\lambda \in \dot{\Lambda}$*, then every optimal solution to the Lasso SSC problem* (24) *with projected and zero-filled data is non-zero and subspace preserving.*

What is notable about Theorem 3 is the simplicity of the condition $\dot{\mu}_\lambda\, \dot{\eta} < \dot{\zeta}$, as well as its resemblance to the condition $\mu_\lambda < r$ of Theorem 1. In fact, the quantity $\dot{\mu}_\lambda$ is a direct analogue of the inter-subspace coherence $\mu_\lambda$, adjusted for the case of PZF data. Indeed, as seen from its definition in (6), $\dot{\mu}_\lambda$ is the maximum inner product between the dual direction associated to the PZF data of subspace $\mathcal{S}_1$ and the PZF data from the remaining subspaces. The quantity $\dot{\eta} \leq 1$ is the Euclidean norm of the point being expressed, which in the absence of missing entries is equal to 1.

Finally, to understand the quantity $\dot{\zeta}$, we first look at its noiseless counterpart $\zeta$ defined in (7). This measures how well distributed are the points $\boldsymbol{X}_{-1}^{(1)}$ with respect to point $\boldsymbol{x}_1^{(1)}$, or in other words, how coherent they are with that

[6] By that we mean that we allow our conditions to be stated in terms of the corrupted data as opposed to quantities that depend only on clean data and only on noise. This latter approach, e.g. followed by Wang & Xu (2016), usually leads to less tight conditions due to the heavy use of the triangle inequality. Instead, we do this decoupling in the probability analysis.

[7] Since the interval $\dot{\Lambda}$ is a function of $\lambda$, it is misleading to write "for any $\lambda \in \dot{\Lambda}$", as Wang & Xu (2016) do in their Theorem 6: $\dot{\Lambda}$ being non-empty does not alone guarantee that also $\lambda \in \dot{\Lambda}$.

point. Notice here that $\zeta$ is a more relevant quantity than the inradius $r$, since the latter does not involve any information about the point being expressed. In addition, $\zeta$ is directly computable from the data, while the inradius is in principle hard to compute. Furthermore, it is almost always true that $r < \zeta$, so that if we were to replace condition $\mu_\lambda < r$ with condition $\mu_\lambda < \zeta$, we would obtain a better result. This is precisely the condition that Theorem 3 reduces to for complete data, which is a novel result itself:

**Theorem 4** (SSC with uncorrupted data, deterministic)**.** *Consider expressing point $\boldsymbol{x}_1^{(1)}$ in terms of the rest of the points in $\boldsymbol{X}$ via the Lasso SSC formulation* (15)*. If $\mu_\lambda < \zeta$ then the open interval $\Lambda_\lambda := \left(\zeta^{-1}, 0.5\zeta^{-1} + 0.5\mu_\lambda^{-1}\right)$ is non-empty, and if $\lambda \in \Lambda_\lambda$, then any optimal solution is non-zero and subspace preserving.*

Returning to the discussion of Theorem 3, we see that the quantity $\dot{\bar{\zeta}}$ captures how well distributed the PZF data $\dot{\boldsymbol{X}}_{-1}^{(1)}$ are with respect to the point $\bar{\boldsymbol{x}}_1^{(1)}$ that is being expressed, which certainly depends on both how well-distributed the original data $\boldsymbol{X}_{-1}^{(1)}$ are, as well as on how uniform the observation pattern is. We can now interpret condition (26): the PZF data $\dot{\bar{\boldsymbol{X}}}_{-1}^{(1)}$ associated to the same subspace $\mathcal{S}_1$ as the point $\dot{\bar{\boldsymbol{x}}}_1^{(1)}$ being expressed must be well distributed with respect to that point normalized (large $\dot{\bar{\zeta}}/\dot{\bar{\eta}}$), while the PZF points $\dot{\bar{\boldsymbol{X}}}^{(-1)}$ in the remaining subspaces must be sufficiently far away from the projected subspace $\dot{\bar{\mathcal{S}}}_1$, as measured by their inner product with the corresponding dual direction $\dot{\hat{\bar{\boldsymbol{v}}}}_{1,\lambda} \in \dot{\bar{\mathcal{S}}}_1$ (small $\dot{\bar{\mu}}_\lambda$). Note here that as the number $m$ of missing entries increases, the quantity $\dot{\bar{\eta}}$ decreases but so does $\dot{\bar{\zeta}}$; moreover the projection is onto a subspace of even lower dimension $D-m$, which makes $\dot{\bar{\mu}}_\lambda$ increase, thus overall making it harder for (26) to be satisfied.

Next, we derive a probabilistic statement from Theorem 3. This is done by constructing high-probability upper and lower bounds for the LHS and RHS of (26), where we exploit the fact that data corruptions due to missing entries are induced by orthogonal projections, i.e., for every $\boldsymbol{x}_j^{(i)}$,

$$\bar{\boldsymbol{x}}_j^{(i)} = \bar{\boldsymbol{P}}_j^{(i)} \boldsymbol{x}_j^{(i)} = \boldsymbol{x}_j^{(i)} + (-\tilde{\boldsymbol{P}}_j^{(i)} \boldsymbol{x}_j^{(i)}). \tag{27}$$

**Theorem 5** (PZF-SSC, probabilistic)**.** *Consider the random model of Definition 2. Suppose that for each point we do not observe exactly $m < D - d$ entries, with the pattern of missing entries being arbitrary, but otherwise fixed apriori. Suppose that the point density $\rho$ is larger than a universal constant, and let $\epsilon > 0$ be a parameter that controls the probability of success. Then there exists a universal constant $c$, such that if $\omega := m/D$ satisfies*

$$\alpha > \sqrt{2\omega} + \beta\sqrt{1-\omega} + (1+\beta)\sqrt{\epsilon + \beta^2/3}, \tag{28}$$

*then there exists a non-empty interval $\Lambda \subset \mathbb{R}$ such that for any $\lambda \in \Lambda$, any optimal solution to the PZF-SSC problem* (24) *is non-zero and subspace preserving, with probability at least $1 - 2/N^2 - \exp(-\sqrt{\rho}d) - (2/n)\exp(-cD\epsilon)$.*

To get an insight into how the maximal tolerable level of missing entries scales with the subspace dimension $d$, we note that for high-ambient dimensions $D$ the quantity $\beta$ is negligible with respect to the quantity $\alpha$. Similarly, ignoring the small parameter $\epsilon$, (28) becomes approximately $\alpha \geq \sqrt{2\omega}$, which by the definition of $\alpha$ and $\omega$ gives

$$\text{PZF-SSC}: \quad \frac{m}{D} < \frac{1}{2}\frac{\log(\rho)}{16d} = \mathcal{O}\left(\frac{1}{d}\right). \tag{29}$$

Informally, (29) says that the maximal tolerable percentage of missing entries of PZF-SSC as predicted by Theorem 5, scales inversely proportionally to the subspace dimension.

### 3.2. ZF-SSC Theory

Similar techniques that led to Theorems 3 and 5 can be employed to yield deterministic and probabilistic statements about ZF-SSC. In particular, we have:

**Theorem 6** (ZF-SSC, deterministic)**.** *With the notation of Definition 1, further define the positive quantity*

$$\bar{\lambda}^* := \frac{1}{2}\Bigg\{ \frac{1}{2\bar{\zeta}} - \frac{\bar{\mu}_\lambda}{\bar{\gamma}\bar{\eta}} - \frac{1}{2\bar{\eta}^2} + \Bigg( \frac{9}{4\bar{\zeta}^2} + \frac{\bar{\mu}_\lambda}{\bar{\gamma}\bar{\eta}\bar{\zeta}} + \frac{2}{\bar{\gamma}\bar{\eta}^2} + \frac{\bar{\mu}_\lambda^2}{\bar{\gamma}^2\bar{\eta}^2} + \frac{1}{4\bar{\eta}^4} + \frac{1}{\bar{\eta}^2}\Big(\frac{\bar{\mu}_\lambda}{\bar{\gamma}\bar{\eta}} - \frac{1}{2\bar{\zeta}}\Big)\Bigg)^{1/2}\Bigg\}. \tag{30}$$

*Then the interval $\bar{\Lambda} := (1/\bar{\zeta}, \bar{\lambda}^*)$ is non-empty, if*

$$\bar{\mu}_\lambda\, \bar{\eta} + \bar{\gamma} < \bar{\zeta}. \tag{31}$$

*If in addition $\lambda \in \bar{\Lambda}$, then every optimal solution to the Lasso SSC problem* (19) *with zero-filled data is non-zero and subspace preserving.*

The quantities $\bar{\mu}_\lambda$, $\bar{\eta}$, $\bar{\zeta}$ are in direct analogy with the quantities $\dot{\bar{\mu}}_\lambda$, $\dot{\bar{\eta}}$, $\dot{\bar{\zeta}}$ that appeared in Theorem 3, except that now they are defined in terms of ZF data instead of PZF data. In fact, as seen from their definitions in (10) and (7), $\bar{\eta} = \dot{\bar{\eta}}$ and $\bar{\zeta} = \dot{\bar{\zeta}}$, while in principle the inter-subspace coherences $\bar{\mu}_\lambda, \dot{\bar{\mu}}_\lambda$ need not coincide. Instead, the main difference between (31) and (26) is the appearance of the quantity $\bar{\gamma}$, whose PZF counterpart $\dot{\bar{\gamma}}$ appears in Theorem 3 only in the definition of the allowable interval for $\lambda$.

The quantity $\bar{\gamma}$ admits an interesting interpretation: As seen from its definition in (12), $\bar{\gamma}$ captures the coherence between the ZF data $\bar{\boldsymbol{X}}^{(-1)}$ associated to subspaces $\mathcal{S}_i, i > 1$, and a projected version of the unobserved components

$\tilde{\boldsymbol{X}}_{-1}^{(1)}$ of the data from $\mathcal{S}_1$. A large such coherence intuitively means that significant information about $\mathcal{S}_1$, potentially crucial for the reconstruction of $\bar{\boldsymbol{x}}_1^{(1)}$ as a linear combination of points in $\bar{\boldsymbol{X}}_{-1}$, is *leaked away* into $\tilde{\boldsymbol{X}}_{-1}^{(1)}$, with which $\bar{\boldsymbol{X}}^{(-1)}$ highly correlates (assuming large $\bar{\gamma}$). In turn, this may lead the optimization problem to favor points of $\bar{\boldsymbol{X}}^{(-1)}$ in expressing $\bar{\boldsymbol{x}}_1^{(1)}$, thus leading to the loss of the subspace-preserving property by the solutions to (19).

Interestingly, comparison of the proofs of Theorems 3 and 6 reveals that $\dot{\bar{\gamma}}$ did not appear in (26) because $\bar{\boldsymbol{x}}_1^{(1)}$ is complete when the underlying subspace arrangement is taken to be $\bigcup_{i=1}^n \dot{\mathcal{S}}_i$, which is the natural view that we adopted for our analysis of PZF-SSC. On the contrary, such a feature is not available in the analysis of ZF-SSC, as $\bar{\boldsymbol{x}}_1^{(1)}$ is in principle incomplete with respect to $\bigcup_{i=1}^n \mathcal{S}_i$.

As we did for PZF-SSC, we use the deterministic Theorem 6 to derive a probabilistic statement:

**Theorem 7** (ZF-SSC, probabilistic)**.** *Consider the exact setting of Theorem 5. If $\omega := m/D$ satisfies*

$$\alpha > (\sqrt{2}+\sqrt{\epsilon+\beta^2/3})\sqrt{\omega} + (\beta+\sqrt{\epsilon+\beta^2/3})\sqrt{1-\omega} + \sqrt{\omega(1-\omega)} + (1+\beta+\sqrt{\epsilon+\beta^2/3})\sqrt{\epsilon+\beta^2/3}, \quad (32)$$

*then there exists a non-empty interval $\Lambda \subset \mathbb{R}$ such that for any $\lambda \in \Lambda$, any optimal solution to the ZF-SSC problem* (19) *is non-zero and subspace preserving, with probability at least* $1-2/N^2-\exp(-\sqrt{\rho}d)-2(1+1/n)\exp\left(-cD\epsilon\right)$.

Repeating the informal arguments that led to (29), i.e., for high ambient dimension $D$ ignoring $\beta$ and $\epsilon$, (32) becomes $\alpha > \sqrt{2\omega} + \sqrt{\omega(1-\omega)}$. Since $\sqrt{\omega} \geq \sqrt{\omega(1-\omega)}$, we then have that this latter simplified condition is satisfied if the stronger condition $\alpha > (1+\sqrt{2})\sqrt{\omega}$ is true. This gives

$$\text{ZF-SSC}: \quad \frac{m}{D} < \frac{1}{(1+\sqrt{2})^2}\frac{\log(\rho)}{16d} = \mathcal{O}\left(\frac{1}{d}\right), \quad (33)$$

i.e., ZF-SSC can tolerate $1/d$ fraction of missing entries.[8]

### 3.3. A Comparison between PZF-SSC and ZF-SSC

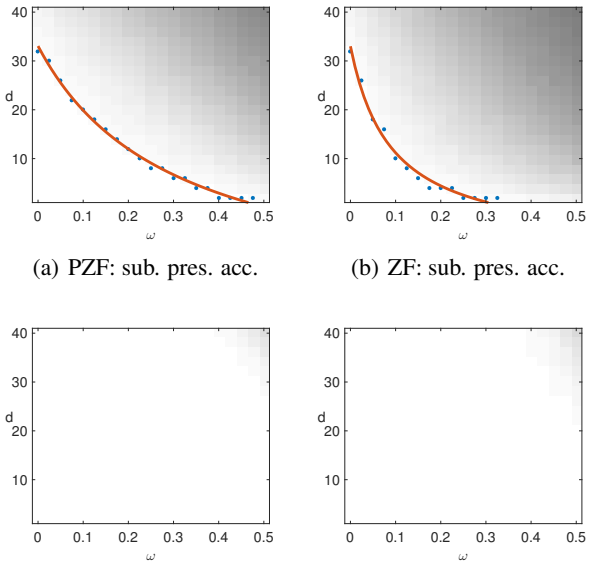
As per (29) and (33), both PZF-SSC and ZF-SSC give subspace preserving solutions as long as the ratio of missing entries scales as $1/d$. On the other hand, the multiplying constant associated to PZF-SSC is about 3 times larger, suggesting a superiority of PZF-SSC. Alternatively, with

$$f_{\text{PZF}}(\omega) := \alpha - \sqrt{2\omega} - \beta\sqrt{1-\omega} - (1+\beta)\sqrt{\epsilon+\beta^2/3}, \quad (34)$$

the PZF Theorem 5 asks that

$$f_{\text{PZF}}(\omega) > 0, \quad (35)$$

[8]This result is in agreement with the result of Charles et al. (2018), who studied only ZF-SSC.




(a) PZF: sub. pres. acc. (b) ZF: sub. pres. acc.

(c) PZF: clustering acc. (d) ZF: clustering acc.

*Figure 1.* Figs. 1(a)-1(b) show the subspace preserving accuracy for both PZF-SSC and ZF-SSC, along with a fitted hyperbola (allowing for vertical and horizontal shift) for the phase transition region, the latter measured with a precision of 0.98. Figs. 1(c)-1(d) show the corresponding clustering accuracies produced by spectral clustering applied on the affinity graphs. Parameters are set as $D = 100$, $\rho = 5$, $n = 3$, $\lambda = 10/\bar{\zeta}$. The complete data are unit norm, drawn uniformly at random from the subspaces, and each point is missing $m = \omega D$ entries also chosen uniformly at random. Results are averaged over 10 trials.

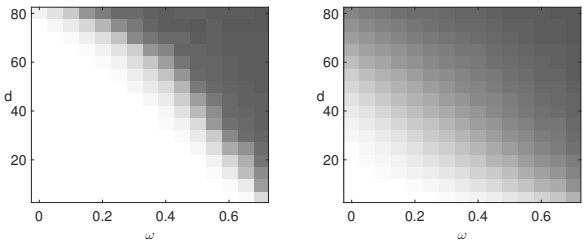
while the ZF Theorem 7 asks that

$$f_{\text{ZF}}(\omega) := -\sqrt{\epsilon+\beta^2/3}(\sqrt{\omega}+\sqrt{1-\omega}+\sqrt{\epsilon+\beta^2/3}) - \sqrt{\omega(1-\omega)} + f_{\text{PZF}}(\omega) > 0, \quad (36)$$

a significantly harder condition to satisfy than (35), due to the dominating negative term $-\sqrt{\omega(1-\omega)}$. Once again, this suggests that PZF-SSC has an advantage over ZF-SSC.

The actual algorithmic behavior is depicted in Fig. 1. In Figs. 1(a)-1(b) we plot the subspace preserving accuracies for PZF-SSC and ZF-SSC, defined as the ratios of the $\ell_1$-norm of the $N \times N$ self-representation matrices $\boldsymbol{C}_{\text{PZF}}$ and $\boldsymbol{C}_{\text{ZF}}$ restricted to intra-subspace connections over the total $\ell_1$ norm of $\boldsymbol{C}_{\text{PZF}}$ and $\boldsymbol{C}_{\text{ZF}}$ respectively. This quantity measures the degree to which points within a subspace use only points from the same subspace for their representation. In Figs. 1(c)-1(d) we show the clustering accuracy that corresponds to spectral clustering applied on the affinity graphs defined by $\boldsymbol{C}_{\text{PZF}}$ and $\boldsymbol{C}_{\text{ZF}}$.

There are at least four notable observations. First, as seen in Figs. 1(a)-1(b), the phase transition between subspace preserving solutions and non-subspace preserving ones is




(a) PZF: clustering acc. (b) PZF: sub. pres. acc.

*Figure 2.* Clustering and subspace preserving accuracies for PZF-SSC plotted with smaller resolution across a wider range of subspace dimensions and missing rates (as in Fig. 1, except $\rho = 3$).

indeed of hyperbolic nature for both methods, as theoretically predicted by (29) and (33). Second, PZF-SSC has indeed higher subspace preserving accuracy than ZF-SCC, as suggested by our two theoretical arguments in the beginning of this section: For example, for 5-dimensional subspaces ($d = 5$) in $\mathbb{R}^{100}$ PZF-SSC can tolerate up to 34 missing entries per point, while ZF-SSC can tolerate up to 19. Third, both methods start breaking down rather quickly as the number of missing entries increases: for $d = 10$ PZF-SSC and ZF-SSC can tolerate, respectively, at most 23 and 10 missing entries per point before their solutions become non-subspace preserving, while for $d = 20$ the maximal tolerable number of missing entries becomes 10 and 4, respectively. Notice how close the values for ZF-SSC are to $D/d$ in each of the above cases. Finally, even though the quality of the connections degrades quickly as $d$ and $\omega$ increase, the clustering accuracy remains very high (close to 1) for both methods, a phenomenon that we attribute to the robustness of spectral clustering (Figs. 1(c)-1(d)).

## 4. Discussion

**Bounding dual vectors and inradius.** A feature of our theory is that the *subspace separation* conditions for complete, ZF and PZF data have the same geometric form, i.e.,

$$\mu_\lambda < \zeta, \;\; \bar{\mu}_\lambda \bar{\eta} + \bar{\gamma} < \bar{\zeta}, \;\; \text{and} \;\; \dot{\bar{\mu}}_\lambda \dot{\bar{\eta}} < \dot{\bar{\zeta}} \qquad (37)$$

respectively. This nice structure comes from a novel bound on the norm of the so-called *dual vector* $\boldsymbol{v}$ that takes into consideration both the objective function as well as the constraint of the reduced dual problem (2). Instead, Soltanolkotabi & Candès (2012) bound $\boldsymbol{v}$ exclusively from the constraint of (2). The two techniques lead to a trade-off between tightness of subspace separation conditions and upper bounds for the Lasso parameter $\lambda$[9] and it is an open problem to optimally bound $\boldsymbol{v}$, which is then expected to lead to jointly better conditions. At any case, the probabilistic lower bound on $r < \zeta$ that we also have used in our analysis is the quantity $\alpha = \sqrt{\log(\rho)/16d}$ (Alonso-Gutierrez, 2008), which even though of fundamental theoretical importance, is too pessimistic: for $\rho = 5$ and $d = 5$ eq. (26) predicts at most 1 tolerable missing entry for PZF-SSC in $\mathbb{R}^{100}$, while as per Fig. 1(a) the method handles 34 missing entries per point. Can we do better than that?

[9] This is more easily seen by comparing the conditions of Theorems 1 and 4 for complete data.

**PZF vs. ZF.** As argued theoretically by comparing Theorems 5 and 7, and corroborated experimentally by Fig. 1 (§3.3), projecting the incomplete dataset onto the observation pattern of the point being expressed increases the robustness of the self-representation of the dataset to missing entries with respect to the subspace preserving property, at least for low-dimensional subspaces. Our study was solely in the context of SSC, yet we believe that working with PZF data instead of ZF data is advantageous regardless of the choice of self-expressive method (Liu et al., 2013; Lu et al., 2012; Elhamifar & Vidal, 2013; Wang et al., 2013; You et al., 2016); a conjecture to be established.

**Beyond PZF-SSC.** Even though the clustering accuracy for PZF-SSC seems rather satisfactory as depicted for higher subspace dimensions and higher missing rates in Fig. 2(a), its rather poor subspace preserving accuracy shown in 2(b), suggests that PZF-SSC is still too simple a method to handle the subspace clustering problem for incomplete data, and that its performance relies to a significant extent on the robustness of spectral clustering. E.g., as per Figs. 2(a)-2(b), for three 60-dimensional subspaces inside $\mathbb{R}^{100}$ and 15 missing entries per point, about $40\%$ of the points a point connects to live in different subspaces; yet the clustering accuracy is $99\%$. On the other hand, the more sophisticated approach of Elhamifar (2016) builds on the SSC formulation and allows for both clustering and completion in a unified framework. Nevertheless, that approach comes with no theoretical guarantees and appears to be computationally burdensome, leaving as an open challenge the proposal of a theoretically sound, efficient and accurate algorithm for clustering incomplete data associated to a union of low-dimensional subspaces.

## Acknowledgements

Work funded by ShanghaiTech University and NSF grants 1447822 and 1618637. The first author thanks Yunzhen Yao for proof-reading the longer version of this manuscript and catching some mistakes, as well as for producing the experiments. He thanks Dr. Chun-Guang Li for insightful comments on an earlier version of this manuscript, Dr. Gregory Ongie for comments on the current manuscript, and Ron Boger for interesting conversations on missing entries and for sharing his code. The authors thank all four anonymous reviewers for their constructive comments.

## References


Aldroubi, A., Sekmen, A., Koku, A. B., and Cakmak, A. F. Similarity matrix framework for data from union of subspaces. *Applied and Computational Harmonic Analysis*, 2017.

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

Balzano, L., Recht, B., and Nowak, R. High-dimensional matched subspace detection when data are missing. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, 2010.

Bradley, P. S. and Mangasarian, O. L. k-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000. ISSN 0925-5001.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.

Charles, Z., Jalali, A., and Willet, R. Subspace clustering with missing and corrupted data. *arXiv:1707.02461v2*, 2018.

Chen, G. and Lerman, G. Spectral curvature clustering (SCC). *International Journal of Computer Vision*, 81 (3):317–330, 2009. ISSN 0920-5691.

Chen, S. S., Donoho, D. L., and Saunders, M. A. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

Elhamifar, E. High-rank matrix completion and clustering under self-expressive models. In *Advances in Neural Information Processing Systems 29*, 2016.

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

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

Eriksson, B., Balzano, L., and Nowak, R. High-rank matrix completion. *Journal of Machine Learning Research*, Proceedings Track 22:373–381, 2012.

Fischler, M. A. and Bolles, R. C. RANSAC random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 26:381–395, 1981.

Heckel, R. and Bölcskei, H. Robust subspace clustering via thresholding. *IEEE Transactions on Information Theory*, 61(11):6320–6342, 2015.

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

Lu, C-Y., Min, H., Zhao, Z-Q., Zhu, L., Huang, D-S., and Yan, S. Robust and efficient subspace segmentation via least squares regression. In *European Conference on Computer Vision*, pp. 347–360, 2012.

Ongie, G., Willett, R., Nowak, R. D., and Balzano, L. Algebraic variety models for high-rank matrix completion. In *34th International Conference on Machine Learning*, volume 70, pp. 2691–2700, 2017.

Peng, X., Xiao, S., Feng, J., Yau, W.-Y., , and Yi, Z. Deep subspace clustering with sparsity prior. In *25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1925–1931, 2016.

Pimentel-Alarcon, D., Boston, N., and Nowak, R. D. Deterministic conditions for subspace identifiability from incomplete sampling. *Information Theory (ISIT), 2015 IEEE International Symposium on*, pp. 2191–2195, 2015.

Pimentel-Alarcon, D.L. and Nowak, R. D. The information-theoretic requirements of subspace clustering with missing data. In *International Conference on Machine Learning (ICML)*, 2016.

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

Soltanolkotabi, M. and Candès, E. J. A geometric analysis of subspace clustering with outliers. *Annals of Statistics*, 40(4):2195–2238, 2012.

Soltanolkotabi, M., Elhamifar, E., and Candès, E. J. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

Tibshirani, R. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Tsakiris, M. C. and Vidal, R. Hyperplane clustering via dual principal component pursuit. In *International Conference on Machine Learning*, 2017a.

Tsakiris, M. C. and Vidal, R. Filtrated algebraic subspace clustering. *SIAM Journal on Imaging Sciences*, 10(1): 372–415, 2017b.

Tsakiris, M. C. and Vidal, R. Algebraic clustering of affine subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(40):482, 2018a.

Tsakiris, M. C. and Vidal, R. Theoretical analysis of sparse subspace clustering with missing entries. *arXiv:1801.00393v3 [cs.LG]*, 2018b.

Vidal, R., Ma, Y., and Sastry, S. Generalized Principal Component Analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1–15, 2005.

Vidal, R., Ma, Y., and Sastry, S. *Generalized Principal Component Analysis*. Springer Verlag, 2016.

Wang, W., Aeron, S., and Aggarwal, V. On deterministic conditions for subspace clustering under missing data. In *International Symposium on Information Theory*, pp. 850–854, 2016.

Wang, Y.-X. and Xu, H. Noisy sparse subspace clustering. *Journal of Machine Learning Research*, 17(12):1–41, 2016.

Wang, Y-X., Xu, H., and Leng, C. Provable subspace clustering: When LRR meets SSC. In *Neural Information Processing Systems*, 2013.

Yang, C., Robinson, D., and Vidal, R. Sparse subspace clustering with missing entries. In *International Conference on Machine Learning*, 2015.

You, C. and Vidal, R. Geometric conditions for subspace-sparse recovery. In *International Conference on Machine learning*, pp. 1585–1593, 2015.

You, C., Li, C.-G., Robinson, D., and Vidal, R. Oracle based active set algorithm for scalable elastic net subspace clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3928–3937, 2016.