Sparse Additive Subspace Clustering

Xiao-Tong Yuan[1,2] and Ping Li[2]

[1] S-mart Lab, Nanjing University of Information Science and Technology
Nanjing, 210044, China
[2] Department of Statistics and Biostatistics, Department of Computer Science
Rutgers University, Piscataway, New Jersey, 08854, USA
xtyuan@nuist.edu.cn, pingli@stat.rutgers.edu

Abstract. In this paper, we introduce and investigate a sparse additive model for subspace clustering problems. Our approach, named **SASC** (**S**parse **A**dditive **S**ubspace **C**lustering), is essentially a functional extension of the Sparse Subspace Clustering (SSC) of Elhamifar & Vidal [7] to the additive nonparametric setting. To make our model computationally tractable, we express SASC in terms of a finite set of basis functions, and thus the formulated model can be estimated via solving a sequence of grouped Lasso optimization problems. We provide theoretical guarantees on the subspace recovery performance of our model. Empirical results on synthetic and real data demonstrate the effectiveness of SASC for clustering noisy data points into their original subspaces.

1 Introduction

This paper deals with the problem of subspace clustering which assumes that a collection of data points lie near a union of unknown linear subspaces and aims to fit these data points to their original subspaces. This is an unsupervised learning problem in that we do not know in advance to which subspace these data points belong. It is thus of interest to simultaneously cluster the data points into multiple subspaces and uncover the low-dimensional structure of each subspace. Subspace clustering has been widely applied in numerous scientific and engineering domains, including computer vision [30,35], data mining [20,1], networks analysis [8,10], switched system identification in control [2,16] and computational biology [17,11]. In many such applications, the lower dimensional representations are characterized by multiple low-dimensional manifolds which can be well approximated by subspaces with only slightly higher dimensions than those of the underlying manifolds. For example, in a video sequence, geometric argument shows that trajectories of same rigid-body motion lie on a subspace of dimension 4 [25]. The partitions of observed trajectories correspond to different rigid objects and thus are useful for understanding the scene dynamics. The task of finding these partitions serves as a standard application of subspace clustering which is known as multi-body motion segmentation [30] in computer vision.

1.1 Problem Setup and Motivation

Assume that the underlying K subspaces $\{S_k\}_{k=1}^K$ of $\mathbb{R}^p$ have unknown dimensions $\{d_k\}_{k=1}^K$ respectively. Let $\mathcal{Y} \subset \mathbb{R}^p$ be a given data set of cardinality n which

D. Fleet et al. (Eds.): ECCV 2014, Part III, LNCS 8691, pp. 644–659, 2014.

may be partitioned as $\mathcal{Y} = \bigcup_{k=1}^{K} \mathcal{Y}_k$ with each $\mathcal{Y}_k$ being a collection of n_k vectors that are distributed around subspace S_k. The task of subspace clustering is to approximately segment the point points in $\mathcal{Y}$ into their respective underlying subspaces. Due to the presence of noise, it is usually assumed in real applications that each observation $\boldsymbol{y} = [y^{(1)}, ..., y^{(p)}]^\top \in \mathcal{Y}$ is generated from the following superposition stochastic model [24]:

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e},$$

where $\boldsymbol{x}$ is a sample belonging to one of the subspaces and $\boldsymbol{e}$ is a random perturbation term with a bounded Euclidean norm. For noiseless samples, it is intuitive to assume that these points are distributed uniformly at random on each subspace. The state-of-the-art robust subspace clustering algorithms take advantage of the so-called self-expressiveness property of linear subspaces, i.e., each noiseless data point from one of the subspaces can be reconstructed by a combination of the other noiseless data points from the same subspace. Formally, the self-expressiveness model is defined as:

Definition 1 (Self-Expressiveness (SE) property [7]). *For each data point* $\boldsymbol{x}_i$*, there exist coefficients* $\{\theta_{ij}\}$ *such that* $\boldsymbol{x}_i = \sum_{j \neq i} \theta_{ij} \boldsymbol{x}_j$.

Note that the coefficients $\{\theta_{ij}\}$ are sparse as we assume that the subspace S_k has d_k-dimensionality. Ideally, it is expected that the non-zero coefficients are from those points belonging to the same subspace as $\boldsymbol{x}_i$, and thus the joint parameter matrix $\boldsymbol{\Theta} = (\theta_{ij}) \in \mathbb{R}^{n \times n}$ has block diagonal structure with blocks corresponding to clusters. Since $\boldsymbol{y}_i = \boldsymbol{x}_i + \boldsymbol{e}_i$, the SE property of clean data leads to the following noisy linear representation model for the noisy observations:

$$\boldsymbol{y}_i = \sum_{j \neq i} \theta_{ij} \boldsymbol{y}_j + \boldsymbol{z}_i, \tag{1}$$

where the perturbation term $\boldsymbol{z}_i = \boldsymbol{e}_i - \sum_{j \neq i} \theta_{ij} \boldsymbol{e}_j$. Inspired by this intuitive property, Elhamifar & Vidal [7] introduced the **SSC** (Sparse Subspace Clustering) approach using sparse reconstruction coefficients as similarity measures; the sparse coefficients are obtained by reconstructing each sample $\boldsymbol{y}_i$ using all the rest samples $\{\boldsymbol{y}_j\}_{j \neq i}$, while regularizing the coefficient vector by ℓ_1-norm to promote sparsity. Hence SSC amounts to solving a sequence of ℓ_1-minimization problems (for noiseless data) or Lasso problems (for noisy data) which are computationally tractable and statistically efficient in high dimensional setting [24][23]. The task of clustering is then finalized by applying a spectral clustering method [18] to a symmetric affinity matrix constructed from these representation coefficients.

While SSC has strong theoretical guarantees and impressive practical performances, it essentially fits the data with a high dimensional linear regression model in (1). Unfortunately, due to the presence of distortion beyond random additive perturbation, real-world observations often do not conform exactly to such a linear model assumption.

1.2 Our Contribution

In this paper, we relax the strong linear model assumption made by SSC and investigate a novel class of nonparametric subspace clustering models called SASC (Sparse Additive Subspace Clustering). We assume that for each observed data point $\boldsymbol{y}_i$, there exists a potentially nonlinear transformation f_i such that the transformed data point $\{f_i(\boldsymbol{y}_i)\}$ obeys the noisy SE model in (1). This problem setup is more challenging than SSC in the sense that the transformations $\{f_i\}$, subspaces $\{S_k\}$ and random noises are all unknown. Indeed, SSC is a special case of SASC when $f_i(a) = a$. Our method combines the ideas from SSC and SpAM which is a sparse additive model for nonparametric regression tasks [22]. To make our model computationally tractable, we follow SpAM to project the unknown f_i onto a functional subspace with a known basis (e.g., the truncated Fourier basis). While the two methods share some common principles, there are two fundamental differences between SASC and SpAM: i) obviously the problem setups are different; ii) as the regressors are dependent on noisy data, we need to deal with noisy design matrix which was not addressed by SpAM.

We provide some theoretical guarantees on the subspace recovery capability of our model. Since there are no "true" parameters to compare the solution against, a subspace clustering algorithm succeeds if each data point is represented by those data points belonging to the same subspace. Therefore, it is desirable that the output representation coefficients matrix has block diagonal structure (under proper arrangement of data). We conduct a deterministic sparsity recovery analysis for our model followed by a stochastic extension. Our results show that under mild conditions, the underlying diagonal block structure of the representation matrix can be efficiently recovered with high probability. We test the numerical performance of our model on simulated and real data. The experimental results show that in many cases our model significantly outperforms the state-of-the-art methods in in terms of clustering accuracy.

1.3 Notation and Outline

Notation. We denote scalars by lower case letters (e.g., x and a), and vectors/matrices by bold face letters (e.g. $\boldsymbol{x}$ and $\boldsymbol{A}$). The Euclidean norm of a vector $\boldsymbol{x}$ is denoted by $\|\boldsymbol{x}\|$. Given a disjoint group structure G over a vector $\boldsymbol{x}$, we use the notation $\boldsymbol{x}_g$ as the tuple formed by the components of $\boldsymbol{x}$ belonging to group $g \in G$. We define $\|\boldsymbol{x}\|_{G,2} = \sum_{g \in G} \|\boldsymbol{x}_g\|$ and $\|\boldsymbol{x}\|_{G,\infty} = \max_{g \in G} \|\boldsymbol{x}_g\|$. With this notation, we extend the signum function as $\text{sgn}(\boldsymbol{x}_g) = \boldsymbol{x}_g / \|\boldsymbol{x}_g\|$ in which we adopt the convention that $\text{sgn}(\mathbf{0})$ can be taken to be any vector with norm less than or equal to one. Also, given a disjoint group structure G over the rows of a matrix $\boldsymbol{A}$, we denote $\|\boldsymbol{A}\|_{G,\infty} = \max_i \|\boldsymbol{A}_{i,\cdot}\|_{G,2}$ where $\boldsymbol{A}_{i,\cdot}$ is the i-th row of $\boldsymbol{A}$. For $T \subseteq G$, we define $T^c = G \setminus T$. The element-wise infinity norm of a matrix $\boldsymbol{A}$ is denoted by $\|\boldsymbol{A}\|_{\infty,\infty}$ and the row-wise infinity norm by $\|\boldsymbol{A}\|_{\infty}$.

Outline. The remainder of this paper is organized as follows: Some prior works are briefly reviewed in §2. We introduce in §3 the sparse additive subspace clustering model along with statistical analysis. Monte-Carlo simulations and real data experiments are presented in §4. Finally, we conclude this paper in §5.

2 Prior Work

In the last decade, various algorithms have been proposed for subspace clustering. A vast body of these algorithms are contributed by researchers in computer vision to address problems such as multi-body motion segmentation, face clustering and image compression. Among them, several representative algorithms include factorization based methods [6], algebraic methods such as Generalized Principal Component Analysis (GPCA) [29] and Local Subspace Affinity (LSA) [34], sparsity/low-rank induced methods such as Sparse Subspace Clustering (SSC) [7] and Low Rank Representation (LRR) [12,19], to name a few. Some other well-known methods include K-plane [3] and Spectral Curvature Clustering (SCC) [4]. For a comprehensive review and comparison of subspace clustering algorithms, we refer the interested readers to the tutorial [28] and references there in.

In the current investigation, we are particularly interested in a popular class of subspace clustering methods which are built upon the SE property as defined in Definition 1. Inspired by this intuitive property, Elhamifar & Vidal [7] introduced the SSC approach using sparse reconstruction coefficients as similarity measures. An identical framework was independently considered by Cheng *et al.* [5] to construct ℓ_1-graph for subspace learning and semi-supervised image analysis. In order to further capture the global structures of data, Liu *et al.* [12] proposed LRR to compute the reconstruction collaboratively by penalizing the nuclear norm of the joint representation matrix. They also provided a robust version to resist random perturbation and element-wise sparse outliers. More recently, Lu *et al.* [15] proposed a unified convex optimization framework for SE based subspace clustering of which SSC and LRR can be taken as special cases. Most existing SE based subspace clustering methods are restricted to linear or affine models. There is a recent trend to extend SE models to nonlinear manifolds. For example, Patel *et al.* [21] proposed Non-Linear Latent Space SSC (NLS3C) as a nonlinear extension of SSC via kernel embedding. Our work advances in this line of research.

Theoretical justification of SSC has received significant interests from computer vision researchers as well as statisticians. It is shown in [7] that when subspaces are disjoint, i.e. they are not overlapping, the block structure of affinity matrix can be exactly recovered. Similar block structure guarantees were established for LRR and LSR (Least Square Regression) [15]. When data is noise free, Soltanolkotabi & Candès[23] provided a geometric functional analysis for SSC which broadens the scope of the results significantly to the case where subspaces are allowed to be overlapping. Under the circumstances of corrupted data, Wang & Xu[32] and Soltanolkotabi *et al.* [24] independently showed that high statistical efficiency could still be achieved by SSC when the underlying subspaces are well separated and the noise level does not exceed certain geometric gap.

As nonparametric extensions of linear models, additive models [9] assume that the nonlinear multivariate regression function admits an additive combination of univariate functions, one for each covariate. In high dimensional analysis, progress has been made on additive models by imposing various sparsity-inducing functional penalties [22,36]. More recently, the idea of component-wise nonparametric extensions has been investigated in Gaussian graphical models learning [14,13,33]. Our approach shares some spirits with these methods in the sense that we model for each data point a nonlinear transformation.

3 Sparse Additive Subspace Clustering (SASC)

In this section, we propose the SASC (Sparse Additive Subspace Clustering) method. We first introduce in §3.1 an additive self-expressive model as a nonparametric extension of the SE model (1). Then we investigate the related parameter estimation and clustering issues in §3.2. Finally, we provide some sparse recovery guarantees on SASC in §3.3.

3.1 Additive Self-Expressive Model

Our method follows the same problem setup as discussed in §1.1. We further assume that there exist n univariate functions $\{f_i(\cdot)\}_{i=1}^n$ such that the transformed data points $\{f_i(\boldsymbol{y}_i)\}_{i=1}^n$ obey the following superposition model:

$$f_i(\boldsymbol{y}_i) = \boldsymbol{x}_i + \boldsymbol{e}_i,$$

where $f_i(\boldsymbol{y}_i) = [f_i(y_i^{(1)}), ..., f_i(y_i^{(p)})]^\top$ and $\boldsymbol{e}_i$'s are random noises. From the SE property of the clean data points$\{\boldsymbol{x}_i\}$,

$$f_i(\boldsymbol{y}_i) = \sum_{j \neq i} \theta_{ij} f_j(\boldsymbol{y}_j) + \boldsymbol{z}_i. \tag{2}$$

Clearly, this is a nonparametric extension of the noisy linear model (1). It depends on $\{f_i\}$ as well as the coefficient matrix $\boldsymbol{\Theta}$, all of which are unknown in advance and need to be estimated from data. In this paper, we assume $f_i \not\equiv 0$ to avoid degeneration.

Remark 1. One may compare the additive self-expressive model in (2) with the SpAM model for sparse nonparametric regression [22]. Given a random response y and a fixed regressor $\boldsymbol{x} = (x_j)$, SpAM considers the additive regression model $y = \sum_j f_j(x_j) + \varepsilon$ in which f_j's are unknown. Although the basic ideas are similar, the model in (2) is more general than SpAM as its response $f_i(\boldsymbol{y}_i)$ is also nonparametric. Moreover, the regressors $\{f_j(\boldsymbol{y}_j)\}_{j \neq i}$ in (2) are random. Such differences contrast our method to SpAM.

Remark 2. It is noteworthy that Patel *et al.* [21] recently proposed NLS3C as a kernel extension of SSC to nonlinear manifolds. Without imposing latent space assumption, NLS3C essentially considers the following linear model in a extended feature space $\mathcal{F}$:

$$\phi(\boldsymbol{y}_i) = \sum_{j \neq i} \theta_{ij} \phi(\boldsymbol{y}_j) + \boldsymbol{z}_i,$$

where $\phi(\cdot)$ is a feature map from $\mathbb{R}^p$ to $\mathcal{F}$. By using standard kernel trick, NLS3C can be formulated as an ℓ_1-regularized kernel regression problem. Although sharing some similar elements, our model in (2) is apparently different from the above NLS3C model in two aspects: (i) NLS3C uses an arbitrary feature map ϕ over all the data points for nonlinear embedding while our model allows different nonlinear functions $\{f_i\}$ to be applied to different data points $\{\boldsymbol{y}_i\}$, without embedding; and (ii) the kernel matrix appeared in NLS3C is typically user specified while in our model the univariate functions f_i are unknown and as we will see shortly that they can be learned in a data-driven manner.

Identifiability. In the nonparametric SE model (2), since both the coefficients θ_{ij} and the functions f_i are all unknown, there might be different interpretations of the same data which lead to different segmentations. For instance, as an extreme example, one could pick f_i identically equal to a constant. This would put all the points in a single cluster, independently from their original distribution. Thus, in general, the model (2) is unidentifiable.

To make this model tractable, we need to impose certain restrictions on the space from which the functions f_i are drawn. To this end, we propose to express $\{f_i(\boldsymbol{y}_i)\}$ appeared in (2) in terms of basis functions. For each data point $\boldsymbol{y}_i$, we denote $\{\psi_{i\ell}(\cdot), \ell = 1, 2, ...\}$ as a set of uniformly bounded, orthonormal functional bases with respect to proper Lebesgue measure. We consider

$$f_i(\boldsymbol{y}_i) = \sum_{\ell=1}^{q} \alpha_{i\ell}\psi_{i\ell}(\boldsymbol{y}_i), \tag{3}$$

where q is the truncation order parameter. It is well-known that for sufficiently large q, the above defined f_i can accurately approximate the function $\tilde{f}_i(\boldsymbol{y}_i) = \sum_{\ell=0}^{\infty} \alpha_{i\ell}\psi_{i\ell}(\boldsymbol{y}_i)$ defined in terms of infinity basis. Therefore, in this paper we will only pursue the truncated formulation (3) which is more of practical interests. For the sake of identifiability, we assume that $\alpha_{i\ell} \neq 0$ for all pairs (i, ℓ). By combining (2) and (3), we obtain

$$\psi_{i\ell}(\boldsymbol{y}_i) = -\sum_{t\neq\ell} \alpha_{it}/\alpha_{i\ell}\psi_{it}(\boldsymbol{y}_i) + \sum_{j\neq i} \theta_{ij} \sum_{t=1}^{q} \alpha_{jt}/\alpha_{i\ell}\psi_{jt}(\boldsymbol{y}_j) + \varepsilon_{i\ell}. \tag{4}$$

Obviously, this is a linear model with respect to the data image under the (known) basis functions and thus is generally identifiable. Next, we will show how to use such a linear model for nonlinear subspace clustering.

Re-Parametrization. One issue with the linear model (4) is that its parameters $\alpha_{i\ell}$ and θ_{ij} are coupled which complicates the optimization and analysis. To address this challenge, we introduce the following re-parametrization scheme:

$$\beta_{it}^{i\ell} := \alpha_{it}/\alpha_{i\ell}, \text{ for } t \neq \ell \quad ; \quad \beta_{jt}^{i\ell} := \theta_{ij}\alpha_{jt}/\alpha_{i\ell}, \text{ for } j \neq i.$$

Then, for each pair $(i, \ell) \in \{1, ..., n\} \times \{1, ..., q\}$, we arrive at the following Additive Self-Expressive (ASE) model:

$$\psi_{i\ell}(\boldsymbol{y}_i) = \sum_{t \neq \ell} \beta_{it}^{i\ell} \psi_{it}(\boldsymbol{y}_i) + \sum_{j \neq i} \sum_{t=1}^{q} \beta_{jt}^{i\ell} \psi_{jt}(\boldsymbol{y}_j) + \varepsilon_{i\ell}, \tag{5}$$

where $\{\beta_{it}^{i\ell}, t = 1, ..., \ell - 1, \ell + 1, ..., q\}$ and $\{\beta_{jt}^{i\ell}, j \neq i, t = 1, ..., q\}$ are unknown parameters to be estimated.

Remark 3. Note that the introduced re-parametrization is not invertible, and thus it is hopeless to recover the parameters θ_{ij} and $\alpha_{i\ell}$ (i.e., f_i) of the original model in (4) from those of the ASE model in (5). Fortunately, for the purpose of subspace clustering, what really matters in (4) is the sparse pattern of the parameters θ_{ij} rather than their exact values and the values of $\alpha_{i\ell}$. Since we have assumed $\alpha_{i\ell} \neq 0$, it is immediately known that $\beta_{jt}^{i\ell} = 0$ if and only if $\theta_{ij} = 0$. That is, the sparse pattern of coefficients θ_{ij} is encoded in the group sparse structure of $\beta_{jt}^{i\ell}$. One interesting implication of this observation is that we may hopefully estimate the sparse pattern of the original model in (4) via estimating that of the re-parameterized model in (5). As we will see shortly that an appealing merit of expression (5) is that it suggests a convex solver which eases the consequent optimization and analysis. Therefore, in the following analysis, we choose to use expression (5) for sparse pattern discover, even though its parameters cannot be readily used to estimate the nonlinear functions $\{f_i\}$.

As discussed in Remark 3, it is expected that the parameters $\{\beta_{jt}^{i\ell}, j \neq i, t = 1, ..., q\}$ exhibit group-level sparsity in terms of the groups defined over the q bases. In the next subsection, we will propose to use grouped Lasso programming to estimate these parameters.

3.2 Parameter Estimation and Clustering

Let us abbreviate $\boldsymbol{\psi}_{i\ell} = \psi_{i\ell}(\boldsymbol{y}_i)$, $\boldsymbol{\Psi}_i = [\boldsymbol{\psi}_{i1}, ..., \boldsymbol{\psi}_{iq}]$ and $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, ..., \boldsymbol{\Psi}_n]$. With obvious notations $\boldsymbol{\beta}_i = [\beta_{i1}, ..., \beta_{iq}]^\top$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^\top, ..., \boldsymbol{\beta}_n^\top]^\top$. Based on these notations, we naturally define a group structure as $G = \{1, 2, ..., n\}$, i.e., the elements inside each $\boldsymbol{\beta}_i$ form a group. In order to estimate ASE model (5) at a given pair of (i, ℓ), we consider a grouped lasso estimator which is defined as the solution to the following convex optimization problem:

$$\hat{\boldsymbol{\beta}}^{i\ell} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2p} \|\boldsymbol{\psi}_{i\ell} - \boldsymbol{\Psi}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|_{G,2} \quad \text{subject to } \beta_{i\ell} = 0, \tag{6}$$

where $\lambda > 0$ is the regularization strength parameter and the constraint $\beta_{i\ell} = 0$ is imposed to leave out trivial solutions. The above grouped Lasso estimator is strongly convex, and thus admits a unique global minimizer. In this paper, we use a standard proximal gradient descent algorithm [27] to find the optimal solution $\hat{\boldsymbol{\beta}}^{i\ell}$. After recovering $\hat{\boldsymbol{\beta}}^{i\ell}$ for each index pair (i, ℓ), we define a similarity matrix $\boldsymbol{W} = (w_{ij})$ in which $w_{ij} = \sqrt{\sum_{\ell=1}^{q} \sum_{t=1}^{q} (\beta_{jt}^{i\ell})^2}$ and set the affinity matrix to be

$\boldsymbol{C} = |\boldsymbol{W}| + |\boldsymbol{W}|^\top$. Then we construct clusters by applying spectral clustering algorithms (e.g., [18] as conventionally used in literature) to the affinity matrix $\boldsymbol{C}$. A high level summary of our SASC method is described in Algorithm 1.

Algorithm 1. Sparse Additive Subspace Clustering (SASC)

```
Input  : A collection of data vectors $\mathcal{Y} = \{\boldsymbol{y}_i \in \mathbb{R}^p\}_{i=1}^n$ and a set of pre-fixed
         functional bases $\{\psi_{i\ell}(\cdot), i = 1, ..., n, \ell = 1, ..., q\}$.
1. Compute $\boldsymbol{\psi}_{i\ell} = \psi_{i\ell}(\boldsymbol{y}_i)$ and set $\boldsymbol{\Psi}_i = [\boldsymbol{\psi}_{i1}, ..., \boldsymbol{\psi}_{iq}]$, $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2, ..., \boldsymbol{\Psi}_n]$.
2. for $(i, \ell) \in \{i = 1, 2, ...n\} \times \{1, ..., q\}$ do
    Estimate the minimizer $\hat{\boldsymbol{\beta}}^{i\ell}$ of the grouped Lasso programming (6).
end
3. Construct the $n$-by-$n$ similarity matrix $\boldsymbol{W}$ with entry $(i, j)$ defined as
$w_{ij} = \sqrt{\sum_{\ell=1}^q \sum_{t=1}^q (\beta_{jt}^{i\ell})^2}$. Form the affinity matrix by $\boldsymbol{C} = |\boldsymbol{W}| + |\boldsymbol{W}|^\top$.
4. Let $\gamma_1 \geq \gamma_2 \geq ... \geq \gamma_n$ be the sorted eigenvalues of the normalized Laplacian
matrix of $\boldsymbol{C}$. Estimate the number of clusters as
    $\hat{K} = n - \arg\max_{i=1,\ldots,n-1} (\gamma_i - \gamma_{i+1})$.
5. Apply a spectral clustering method to the affinity matrix $\boldsymbol{C}$ to produce $\hat{K}$
disjoint clusters $\{\mathcal{Y}_k\}_{k=1}^{\hat{K}}$ of the data.
Output: Constructed Clusters $\{\mathcal{Y}_k\}_{k=1}^{\hat{K}}$ of $\mathcal{Y}$.
```

3.3 Theoretical Analysis

This subsection is devoted to analyzing the sparse recovery performance of SASC. We are particularly interested in the conditions under which the grouped Lasso estimator (6) may reliably select out points sharing the same underlying subspace as $\boldsymbol{y}_i$ over those not. In other words, the hope is that whenever $\hat{\boldsymbol{\beta}}_j^{i\ell} \neq \mathbf{0}$, $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ belong to the same subspace. This is formally defined as the following concept of additive subspace detection property:

Definition 2 (Additive Subspace Detection Property). *Let $\boldsymbol{W}$ be the constructed similarity matrix of Step 3 of Algorithm 1. We say the additive subspace detection property holds if (1) for all (i, j) obeying $w_{ij} \neq 0$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same subspace; (2) for all i, the entries $\{w_{ij}\}_{j \neq i}$ are not all zero.*

This property ensures that the weight matrix $\boldsymbol{W}$ has a block diagonal structure with each block representing a subspace cluster, and thus the affinity matrix $\boldsymbol{C}$. In the subsequent subsections, we will provide some sufficient conditions under which the additive subspace detection property holds for Algorithm 1. We start with a deterministic analysis and then extend the results to stochastic settings.

A Deterministic Analysis. Let us consider the ASE model in (5) as a deterministic model. Without loss of generality, we assume that the columns of $\boldsymbol{\Psi}$ are arranged as $\boldsymbol{\Psi} = [\boldsymbol{\Psi}_{T_1}, ..., \boldsymbol{\Psi}_{T_K}]$ in which the sub-matrix $\boldsymbol{\Psi}_{T_k}$ contains those columns associated with $\mathcal{Y}^{(k)}$ from the subspace S_k. Let $\bar{\boldsymbol{\beta}}^{i\ell}$ be the true parameter vector for the ASE model in (5), i.e.,

$$\psi_{i\ell}(\boldsymbol{y}_i) = \sum_{t \neq \ell} \bar{\beta}^{i\ell}_{it} \psi_{it}(\boldsymbol{y}_i) + \sum_{j \neq i} \sum_{t=1}^{q} \bar{\beta}^{i\ell}_{jt} \psi_{jt}(\boldsymbol{y}_j) + \varepsilon_{i\ell}.$$

The following is our deterministic result on the additive subspace detection property of SASC.

Theorem 1. *Assume that there exists a universal constant $\delta \in (0,1)$ such that for any $k \in \{1, ..., K\}$, $\|(\boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k})^{-1} \boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k^c}\|_{T_k^c,\infty} \leq 1 - \delta$. If for any $\boldsymbol{y}_i \in \mathcal{Y}^{(k)}$, the regularization parameter λ satisfies the following two conditions*

(i) $\forall \ell \in \{1, ..., q\}$,

$$\lambda > \frac{\left\| \boldsymbol{\Psi}_{T_k^c}^\top \boldsymbol{\psi}_{i\ell} - \boldsymbol{\Psi}_{T_k^c}^\top \boldsymbol{\Psi}_{T_k} (\boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k})^{-1} \boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\psi}_{i\ell} \right\|_{T_k^c,\infty}}{p\delta},$$

(ii) $\exists \ell \in \{1, ..., q\}, \boldsymbol{y}_j \in \mathcal{Y}^{(k)}, t \in \{1, ..., q\}$ *such that*

$$\lambda < \frac{|\bar{\boldsymbol{\beta}}^{i\ell}_{jt}|}{\left\| \left(\frac{1}{p} \boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k} \right)^{-1} \right\|_{T_k,\infty}} - \left\| \frac{1}{p} \boldsymbol{\Psi}_{T_k}^\top (\boldsymbol{\psi}_{i\ell} - \boldsymbol{\Psi} \bar{\boldsymbol{\beta}}^{i\ell}) \right\|_{T_k,\infty},$$

then the additive subspace detection property holds.

A proof of this result is provided in Appendix A.

Remark 4. The constant $\delta \in (0,1)$ in the theorem is known as *incoherence parameter* in compressive sensing literature [31]. The main message this theorem conveys is that if the K subspaces respectively spanned by the basis $\{\boldsymbol{\Psi}_{T_k}\}_{k=1}^K$ are weakly correlated to each other, and the regularization parameter λ is well bounded from both sides, then the additive subspace detection property holds. Concerning the compatibility between the condition (i) and condition (ii), if the residual term $\boldsymbol{\psi}^i_\ell - \boldsymbol{\Psi} \bar{\boldsymbol{\beta}}^{i\ell}$ is well bounded and $\min_{jt} |\bar{\boldsymbol{\beta}}^{i\ell}_{j,t}|$ is sufficiently large, then these two conditions are compatible. This point will be made more explicit in the following statistical analysis

A Statistical Analysis. We further consider the ASE model (5) as a stochastic model in which the design $\boldsymbol{\Psi}$ and the noise $\boldsymbol{\varepsilon}$ are both random. In this setting, we assume that the $\|\boldsymbol{\Psi}\|_{\infty,\infty} \leq c$ (which is reasonable as the basis functionals $\{\psi_{i\ell}(\cdot)\}$ are assumed to be uniformly bounded) and the noise levels are bounded by σ. The following is our main result on such a stochastic model.

Theorem 2. *Assume that there exist two universal constants $\delta \in (0,1)$ and $l > 0$ such that for any $k \in \{1, ..., K\}$, $\left\|\mathbb{E}[(\boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k})^{-1} \boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k^c}]\right\|_{T_k^c,\infty} \leq 1 - 2\delta$ and $\max_k\{\|\mathbb{E}[(\frac{1}{p}\boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\Psi}_{T_k})^{-1}]\|_{T_k,\infty}\} \leq 0.5l$. If p is sufficiently large and for any $\boldsymbol{y}_i \in \mathcal{Y}^{(k)}$ the regularization parameter λ satisfies the following two conditions* *(i)* $\forall \ell \in \{1, ..., q\}$,

$$\lambda > \frac{c\sigma}{\delta}\sqrt{\frac{n}{p\eta}},$$

(ii) $\exists \ell \in \{1, ..., q\}, \boldsymbol{y}_j \in \mathcal{Y}^{(k)}, t \in \{1, ..., q\}$ *such that*

$$\lambda < \frac{|\bar{\boldsymbol{\beta}}_{jt}^{i\ell}|}{l} - c\sigma\sqrt{\frac{n}{p\eta}},$$

then the additive subspace detection property holds with probability at least $1-n\eta$.

A proof of this theorem is provided in Appendix B.

Remark 5. Clearly, when $|\bar{\boldsymbol{\beta}}_{jt}^{i\ell}| > (1 + 1/\delta)lc\sigma\sqrt{n/(p\eta)}$, the conditions (i) and (ii) are compatible, i.e., the feasible interval of regularization parameter λ is not empty. In this theorem, the dependence of the bounds on p and n is by no means optimal. Indeed, we use the relatively loose Chebyshev's inequality throughout the derivation to bound the concentration behavior. The reason that the much tighter Chernoff's inequality is not directly applicable here is that the entries of each basis vector $\boldsymbol{\psi}_{i\ell}$ are dependent to each other when $p \geq d_k$. Although it is still possible to obtain sharper bounds using Chernoff's inequality with stronger assumptions and more involved analysis, we choose not to pursue in that direction for the sake of presentation clarity. Moreover, in the high dimensional settings where $p \gg n$, the bounds stated in Theorem 2 are still meaningful.

4 Experiments

We evaluate the performance of SASC for robust subspace clustering on synthetic and real data sets. We first investigate subspace detection performance using Monte-Carlo simulation, and then we apply our method to a motion segmentation benchmark data set.

4.1 Monte-Carlo Simulation

This is a proof-of-concept experiment. The purpose of this experiment is to confirm that when the observed data points from each subspace are contaminated by a highly nonlinear transformation, our approach can be significantly superior to existing subspace clustering models for inferring.

Simulated Data. In our simulation study, we generate 5 overlapping subspaces $\{\mathcal{S}_k\}_{k=1}^5 \subset \mathbb{R}^{1000}$ whose bases $\{\boldsymbol{U}_k\}_{k=1}^5$ are generated by $\boldsymbol{U}_{k+1} = \boldsymbol{R}\boldsymbol{U}_k$, $1 \leq k \leq 4$, where $\boldsymbol{R}$ represents a random rotation matrix and $\boldsymbol{U}_1$ a random orthogonal matrix of dimensions 1000×50. Thus each subspace has a dimension of 50. 20 data vectors are sampled from each subspace by $\boldsymbol{X}^{(k)} = \boldsymbol{U}_k\boldsymbol{D}_k$, $1 \leq k \leq 5$ with $\boldsymbol{D}_k$ being a 50×20 matrix whose entries are i.i.d. standard Gaussian variables. The observed samples are generated as $\boldsymbol{Y}^{(k)} = f^{-1}(\boldsymbol{X}^{(k)} + \boldsymbol{\varepsilon}^{(k)})$ where f is a smooth invertible function and $\boldsymbol{\varepsilon}^{(k)}$ is Gaussian noise. We consider two transformations: (i) the polynomial transform: $f(a) = (x - 0.2)^3$; and (ii) the logarithm transform: $f(a) = -\log a$. Note that f is unknown to our algorithm. For the former transformation, we fit the data to ASE model (5) with polynomial basis, and Fourier basis for the latter.

Comparison of Models and Evaluation Criterion. We compare the performance of our estimator to two representative SE based subspace clustering methods: SSC [7] and LRR [12]. Since the subspace information of the data is available, we measure the performance by *Detection Precision* of the top k links on the constructed graph (corresponding to the top $2k$ entries in the affinity matrix $\boldsymbol{C}$). A link is regarded as *true* if and only if it connects two data points belonging to the same subspace. Also, we use the clustering accuracy as a measurement to evaluate the overall clustering performance.

Results. Figure 1 shows the subspace link detection precision curves on the simulated data. From these curves we can see that SASC is significantly better than SSC and LRR. The clustering accuracies of the considered methods are listed in Table 1. It can be seen that SASC succeeds while SSC and LRR perform poorly on these two synthetic data sets. This result makes sense as SASC explicitly models the underlying nonlinear perturbations which are not addressed by SSC and LRR. Concerning the running time, SASC is slightly slower than SSC because it needs to decompose each data point into the combination of multiple basis and then apply grouped Lasso programming on these extracted basis.

4.2 Motion Segmentation Data

We further evaluate SASC on Hopkins 155 motion dataset [26] which is a benchmark for subspace clustering study. This data set consists of 120 sequences of

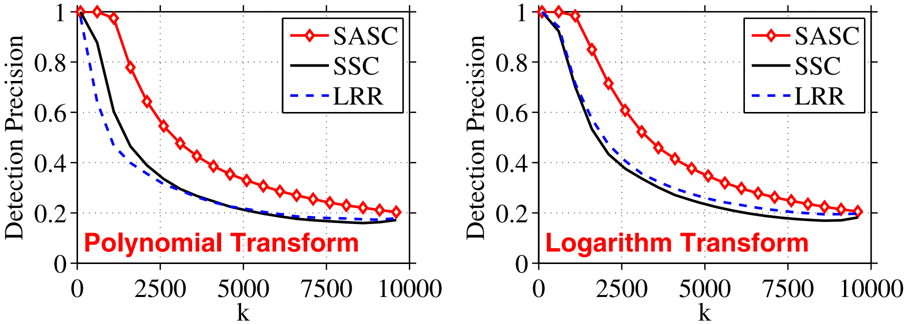

Fig. 1. Precision of the detected top k subspace links on the simulated data

Table 1. Clustering accuracies on the simulated data

Methods	SASC	SSC	LRR
Poly. Trans. $f(a) = (a - 0.2)^3$	**1.00**	0.58	0.36
Log. Trans. $f(a) = -\log a$	**1.00**	0.91	0.72

two motions and 35 sequences of three motions (a motion corresponding to a subspace). Each sequence is a sole segmentation task and so there are 155 subspace segmentation tasks totally. On average, each sequence of two motions has $N = 266$ point trajectories and $F = 30$ frames, while each sequence of three motions has $N = 398$ point trajectories and $F = 29$ frames. We compare the performance of SASC with SSC and LRR which are two representative state-of-the-art subspace clustering algorithms on this data. We follow the experimental protocol in [7] to apply the considered algorithms on the original $2F$-dimensional trajectories and on the $4n$-dimensional subspace (n is the number of subspaces) extracted by PCA. In this experiment, we implement SASC with Fourier basis.

Table 2(a) lists the mean and median clustering errors of the considered methods on the original $2F$-dimensional data points. It can be clearly seen from this table that SASC performs favorably. Table 2(b) lists the clustering errors of the considered methods on the $4n$-dimensional data points obtained by applying PCA. In this setting, SASC achieves the lowest clustering errors on two motion sequences and all sequences, while SCC is the best on three motion sequences. Overall, the observation is that SASC performs the best in most cases. This group of results reveal that the motion trajectories in Hopkins 155 might be contaminated by nonlinear distortions that can be robustly captured by SASC.

Table 2. Hopkins 155: Mean and median clustering errors (%) of the three considered algorithms

(a) $2F$-dimensional data points

Methods	2 Motions		3 Motions		All	
	Mean	Med.	Mean	Med.	Mean	Med.
SASC	**0.90**	0	**3.33**	0.60	**1.45**	0
SSC	1.52	0	4.40	**0.56**	2.18	0
LRR	2.13	0	4.03	1.43	2.56	0

(b) $4n$-dimensional data points by PCA

Methods	2 Motions		3 Motions		All	
	Mean	Med.	Mean	Med.	Mean	Med.
SASC	**0.91**	0	4.46	0.81	**1.71**	0
SSC	1.83	0	**4.40**	**0.56**	2.41	0
LRR	3.41	0	4.86	1.47	3.74	0

5 Conclusions

In this paper, we proposed SASC as a novel nonparametric subspace clustering method. The main idea is to assume that there exists an unknown function for each data point such that the elementwise transformed data point lies near a subspace. This assumption allows us to capture complex perturbations beyond additive random noises in the observed data. In order to make our model computationally tractable, we project the unknown univariate mapping functions onto proper truncated functional spaces. Based on the self-expressiveness property of the clean data, SASC can be formulated as a sequence of nonparametric additive models whose parameters can be estimated via grouped Lasso programming. Statistical analysis shows that under mild conditions, with high probability, SASC is able to successfully recover the underlying subspace structure. Experimental results show that SASC is consistently better than or comparable to the best state-of-the-art methods in clustering accuracy, at a cost of only slightly increased computational time.

Acknowledgements. Xiao-Tong Yuan is supported by NSFJP-BK20141003, NSFJP-BK2012045, NSFC-61272223, ONR-N00014-13-1-0764, and AFOSR-FA9550-13-1-0137. Ping Li is supported by ONR-N00014-13-1-0764, AFOSR-FA9550-13-1-0137, NSF III 1360971, and NSF BIGDATA 1419210.

A Proof of Theorem 1

We need a technical lemma before proving the theorem. Given a response $\boldsymbol{z} \in \mathbb{R}^p$ and a design matrix $\boldsymbol{\Psi} \in \mathbb{R}^{p\times n}$, let us consider the following general grouped Lasso estimator associated with a disjoint group structure G over the parameters:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2p}\|\boldsymbol{z} - \boldsymbol{\Psi}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_{G,2}. \tag{7}$$

Lemma 1. *Let $T \subseteq G$ be a subset of the groups. Assume that there exists a universal constant $\delta \in (0,1)$ such that $\|(\boldsymbol{\Psi}_T^\top\boldsymbol{\Psi}_T)^{-1}\boldsymbol{\Psi}_T^\top\boldsymbol{\Psi}_{T^c}\|_{T^c,\infty} \le 1-\delta$. If the regularization parameter λ satisfies*

$$\lambda > \frac{\|\boldsymbol{\Psi}_{T^c}^\top\boldsymbol{z} - \boldsymbol{\Psi}_{T^c}^\top\boldsymbol{\Psi}_T(\boldsymbol{\Psi}_T^\top\boldsymbol{\Psi}_T)^{-1}\boldsymbol{\Psi}_T^\top\boldsymbol{z}\|_{T^c,\infty}}{p\delta},$$

then

(a) any optimal solution

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2p}\|\boldsymbol{z} - \boldsymbol{\Psi}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_{G,2}$$

must satisfy $\hat{\boldsymbol{\beta}}_{T^c} = \mathbf{0}$.

(b) Moreover, for any $\bar{\boldsymbol{\beta}}$ satisfying $\bar{\boldsymbol{\beta}}_{T^c} = \mathbf{0}$, the element-wise estimation error is bounded by

$$\|\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_\infty \le \left\|\left(\frac{1}{p}\boldsymbol{\Psi}_T^\top\boldsymbol{\Psi}_T\right)^{-1}\right\|_{T,\infty}\left(\left\|\frac{1}{p}\boldsymbol{\Psi}_T^\top(\boldsymbol{z} - \boldsymbol{\Psi}\bar{\boldsymbol{\beta}})\right\|_{T,\infty} + \lambda\right).$$

A proof of this lemma is given in the supplementary material. We are now in the position to prove Theorem 1.

Proof (of Theorem 1). Let us consider a fixed data point $\boldsymbol{y}_i \in \mathcal{Y}^{(k)}$. From the condition (i) and the part (a) of Lemma 1 we know that $\forall \ell \in \{1,...,q\}$, $\hat{\boldsymbol{\beta}}^{i\ell}_{T_k^c} = \mathbf{0}$. From the condition (ii) and the part (b) of Lemma 1 we obtain that $\exists \ell \in \{1,...,q\}, \boldsymbol{y}_j \in \mathcal{Y}^{(k)}, t \in \{1,...,q\}$, such that $\hat{\boldsymbol{\beta}}^{i\ell}_{jt} \neq \mathbf{0}$. Combining these two results and from the construction of $\boldsymbol{W}$ we get that $w_{ij} = 0$ whenever $\boldsymbol{y}_j \notin \mathcal{Y}^{(k)}$, and $\exists \boldsymbol{y}_j \in \mathcal{Y}^{(k)}, j \neq i$ such that $w_{ij} \neq 0$. This verifies the additive subspace detection property.

B Proof of Theorem 2

We start with a technical lemma needed in the proof. Let us consider the following stochastic model

$$\boldsymbol{z} = \boldsymbol{\Psi}\bar{\boldsymbol{\beta}} + \boldsymbol{\varepsilon},$$

where the design $\boldsymbol{\Psi}$ is random and $\boldsymbol{\varepsilon} = [\varepsilon_1, ..., \varepsilon_p]$ are p i.i.d. Gaussian noisewith zero mean and variance σ^2.

The following lemma is a statistical extension of Lemma 1.

Lemma 2. *Let $T \subseteq G$ be a subset of the groups. Assume that there exists a constant $\delta \in (0,1)$ and a constant $l > 0$ such that*

$$\left\|\mathbb{E}[(\boldsymbol{\Psi}_T^\top \boldsymbol{\Psi}_T)^{-1}\boldsymbol{\Psi}_T^\top \boldsymbol{\Psi}_{T^c}]\right\|_{T^c,\infty} \le 1 - 2\delta, \qquad \left\|\mathbb{E}\left[\left(\frac{1}{p}\boldsymbol{\Psi}_T^\top \boldsymbol{\Psi}_T\right)^{-1}\right]\right\|_{T,\infty} \le 0.5l.$$

If p is sufficiently large and the regularization parameter λ satisfies

$$\lambda > \frac{c\sigma}{\delta}\sqrt{\frac{n}{p\eta}},$$

then with probability at least $1 - \eta$

(a) any optimal solution

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{\Psi}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_{G,2}$$

must satisfy $\hat{\boldsymbol{\beta}}_{T^c} = \mathbf{0}$.

(b) Moreover, for any $\bar{\boldsymbol{\beta}}$ satisfying $\bar{\boldsymbol{\beta}}_{T^c} = \mathbf{0}$, then the element-wise estimation error is bounded by

$$\|\hat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_\infty \le \|(\boldsymbol{\Psi}_T^\top \boldsymbol{\Psi}_T)^{-1}\|_{T,\infty}(\|\boldsymbol{\Psi}_T^\top(\boldsymbol{z} - \boldsymbol{\Psi}\bar{\boldsymbol{\beta}})\|_{T,\infty} + \lambda).$$

A proof of this lemma is provided in the supplementary material. Now we prove Theorem 2.

Proof (of Theorem 2). Let us consider a fixed data point $\boldsymbol{y}_i \in \mathcal{Y}^{(k)}$. From the condition (i) and the part (a) of Lemma 2 we know that $\forall \ell \in \{1, ..., q\}$, $\hat{\boldsymbol{\beta}}^{i\ell}_{T_k^c} = \mathbf{0}$ holds with probability at least $1 - \eta$. It is easy to check that with probability at least $1 - \eta$

$$\left\|\frac{1}{p}\boldsymbol{\Psi}_{T_k}^\top(\psi_\ell^i - \boldsymbol{\Psi}\bar{\boldsymbol{\beta}}^{i\ell})\right\|_{T_k,\infty} = \left\|\frac{1}{p}\boldsymbol{\Psi}_{T_k}^\top \boldsymbol{\varepsilon}_\ell^i\right\|_{T_k,\infty} \le c\sigma\sqrt{\frac{n}{p\eta}}.$$

When p is sufficiently large, from the condition (ii) and the part (b) of Lemma 2 we obtain that $\exists \ell \in \{1, ..., q\}, \boldsymbol{y}_j \in \mathcal{Y}^{(k)}, t \in \{1, ..., q\}$, such that $\hat{\boldsymbol{\beta}}^{i\ell}_{jt} \neq \mathbf{0}$ holds with probability $1 - \eta$. Combining these two results and from the construction of $\boldsymbol{W}$ we get that $w_{ij} = 0$ whenever $\boldsymbol{y}_j \notin \mathcal{Y}^{(k)}$, and $\exists \boldsymbol{y}_j \in \mathcal{Y}^{(k)}, j \neq i$ such that $w_{ij} \neq 0$ holds with probability $1 - \eta$. By union of probability, we know that the additive subspace detection property holds with probability at least $1 - n\eta$. This proves the claim.

References

1. Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD 1998), pp. 94–105 (1998)
2. Bako, L.: Identification of switched linear systems via sparse optimization. Automatica 47(4), 668–677 (2011)
3. Bradley, P.S., Mangasarian, O.L.: K-plane clustering. Journal of Global Optimization 16(1), 23–32 (2000)
4. Chen, G., Lerman, G.: Spectral curvature clustering (scc). International Journal of Computer Vision 81(3), 317–330 (2009)
5. Cheng, B., Yang, J., Yan, S., Fu, Y., Huang, T.: Learning with ℓ_1-graph for image analysis. IEEE Transactions on Image Processing 19(4), 858–866 (2010)
6. Costeira, J., Kanade, T.: A multibody factorization method for independently moving objects. International Journal of Computer Vision 29(3), 159–179 (1998)
7. Elhamifar, E., Vidal, R.: Sparse subspace clustering: Algorithm, theory, and applications. IEEE Transactions on Pattern Analysis And Machine Intelligence 35(11), 2765–2781 (2013)
8. Eriksson, B., Balzano, L., Nowak, R.: High-rank matrix completion. In: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS 2012), pp. 373–381 (2012)
9. Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman & Hall/CRC (1990)
10. Jalali, A., Chen, Y., Sanghavi, S., Xu, H.: Clustering partially observed graphs via convex optimization. In: Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML 2011). ACM (2011)
11. Kriegel, H.P., Kröger, P., Zimek, A.: Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data (TKDD) 3, 1–58 (2009)
12. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y., Ma, Y.: Robust recovery of subspace structures by low-rank representation. Transactions on Pattern Analysis and Machine Intelligence 35(1), 171–184 (2013)
13. Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L.: High dimensional semiparametric gaussian copula graphical models. The Annals of Statistics 40(4), 2293–2326 (2012)
14. Liu, H., Lafferty, J., Wasserman, L.: The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. Journal of Machine Learning Research 10, 2295–2328 (2009)
15. Lu, C.-Y., Min, H., Zhao, Z.-Q., Zhu, L., Huang, D.-S., Yan, S.: Robust and efficient subspace segmentation via least squares regression. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 347–360. Springer, Heidelberg (2012)
16. Ma, Y., Vidal, R.: Identification of deterministic switched arx systems via identification of algebraic varieties. In: Morari, M., Thiele, L. (eds.) HSCC 2005. LNCS, vol. 3414, pp. 449–465. Springer, Heidelberg (2005)
17. McWilliams, B., Montana, G.: Subspace clustering of high-dimensional data: A predictive approach. Data Mining and Knowledge Discovery 28(3), 736–772 (2014)
18. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings of the 16th Annual Conference on Neural Information Processing Systems, NIPS 2002 (2002)

19. Ni, Y., Sun, J., Yuan, X.T., Yan, S., Cheong, L.F.: Robust low-rank subspace segmentation with semidefinite guarantees. In: Proceedings of the Workshop on Optimization Based Methods for Emerging Data Mining Problems (OEDM 2010 in conjunction with ICDM 2010) (2010)
20. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: A review. ACM SIGKDD Explorations Newsletter 6(1), 90–105 (2004)
21. Patel, V., Nguyen, H., Vidal, R.: Latent space sparse subspace clustering. In: Proceedings of IEEE International Conference on Computer Vision, ICCV 2013 (2013)
22. Ravikumar, P., Lafferty, J., Liu, H., Wasserman, L.: Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) (JRSSB) 71(5), 1009–1030 (2009)
23. Soltanolkotabi, M., Candès, E.J.: A geometric analysis of subspace clustering with outliers. The Annals of Statistics 40(4), 2195–2238 (2012)
24. Soltanolkotabi, M., Elhamifar, E., Candès, E.J.: Robust subspace clustering. The Annals of Statistics (to appear, 2014)
25. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: A factorization method. International Journal of Computer Vision 9(2), 137–154 (1992)
26. Tron, R., Vidal, R.: A benchmark for the comparison of 3-d motion segmentation algorithms. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, CVPR 2007 (2007)
27. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Submitted to SIAM Journal of Optimization (2008)
28. Vidal, R.: Subspace clustering. IEEE Signal Processing Magazine 28(3), 52–68 (2011)
29. Vidal, R., Ma, Y., Sastry, S.: Generalized principal component analysis (gpca). Transactions on Pattern Analysis and Machine Intelligence 27(12), 1945–1959 (2005)
30. Vidal, R., Tron, R., Hartley, R.: Multiframe motion segmentation with missing data using power factorization and gpca. International Journal of Computer Vision 79, 85–105 (2008)
31. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory 55(5), 2183–2202 (2009)
32. Wang, Y., Xu, H.: Noisy sparse subspace clustering. In: Proceedings of the 30 th International Conference on Machine Learning (ICML 2013), pp. 849–856 (2013)
33. Xue, L., Zou, H.: Regularized rank-based estimation of high-dimensional nonparanormal graphical models. The Annals of Statistics 40(5), 2541–2571 (2012)
34. Yan, J., Pollefeys, M.: A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 94–106. Springer, Heidelberg (2006)
35. Yang, A., Wright, J., Ma, Y., Sastry, S.: Unsupervised segmentation of natural images via lossy data compression. Computer Vision and Image Understanding 110, 212–225 (2008)
36. Yin, Y., Chen, X., Xing, E.: Group sparse additive models. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012 (2012)