Robust Subspace Clustering

Mahdi Soltanolkotabi
Stanford University

ICML workshop on Spectral Learning

June 21, 2013

Joint work with
Emmanuel Candes, Stanford
Ehsan Elhamifar, John Hopkins
Fundamental Tool in Data Mining: PCA
Fundamental Tool in Data Mining: PCA
Subspace Clustering: multi-subspace PCA
Subspace Clustering: multi-subspace PCA
Subspace Clustering: multi-subspace PCA
Subspace Clustering: multi-subspace PCA
Subspace Clustering: multi-subspace PCA
Motivation
Many Applications in Computer Vision

Pictures from various papers/websites.
**Problem**

Identify factors contributing to cancer from gene expression data

- different classes correspond to different tumor or tissue types relating to different cancers.
- samples drawn from the same underlying population are often observed to be highly correlated.
Some recent applications of subspace clustering …

- Biostatistics

- Internet network topology inference [Eriksson, Balzano, and Nowak, 2012]

- Security and privacy in recommender systems [Montanari et al., 2012]

- Hyper-spectral imaging [Chen, Nasrabadi and Tran, 2011]
Some more applications...

- Face Clustering. [Yi Ma et. al. 2009, Elhamifar and Vidal 2013]
- Switched linear systems. [Ozay et. al. 2010]
- ...
Key Concepts: What Makes Subspace Clustering Difficult?
How close are the subspaces?
How close are the subspaces?
How close are the subspaces?
How close are the subspaces?
Principal angle and affinity

**Definition (Principal angles)**

$\theta_1, \theta_2, \ldots, \theta_d$ between subspaces $S$ and $S'$ of dim. $d$

\[
\cos(\theta_i) = \max_{u \in S: \|u\|_2 = 1} \max_{v \in S': \|v\|_2 = 1} u^T v = u_i^T v_i
\]

with orthogonality $u \perp u_j$, $v \perp v_j$, $j < i$
Principal angle and affinity

Definition (Principal angles)

\( \theta_1, \theta_2, \ldots, \theta_d \) between subspaces \( S \) and \( S' \) of dim. \( d \)

\[
\cos(\theta_i) = \max_{u \in S : \|u\|_{\ell_2} = 1} \max_{v \in S' : \|v\|_{\ell_2} = 1} u^T v = u_i^T v_i
\]

with orthogonality \( u \perp u_j, v \perp v_j, j < i \)

Definition (Affinity)

\[
\text{aff}(S, S') = \sqrt{\frac{\cos^2 \theta_1 + \ldots + \cos^2 \theta_{d \wedge d'}}{d \wedge d'}}
\]

Note \( 0 \leq \text{aff}(S, S') \leq 1 \)
Principal angle and affinity

Definition (Principal angles)

\[ \theta_1, \theta_2, \ldots, \theta_d \] between subspaces \( S \) and \( S' \) of dim. \( d \)

\[
\cos(\theta_i) = \max_{u \in S: \|u\|_{\ell_2} = 1} \max_{v \in S': \|v\|_{\ell_2} = 1} u^T v = u_i^T v_i
\]

with orthogonality \( u \perp u_j, v \perp v_j, j < i \)

Definition (Affinity)

\[
\text{aff}(S, S') = \sqrt{\frac{\cos^2 \theta_1 + \ldots + \cos^2 \theta_{d \wedge d'}}{d \wedge d'}}
\]

Note \( 0 \leq \text{aff}(S, S') \leq 1 \)

- Related to Pillai-Bartlett trace in multivariate analysis
- If affinity is on the order of \( 1 \) \( \rightarrow \) no algorithm can work.
How well are the points distributed on each subspace?
How well are the points distributed on each subspace?
How well are the points distributed on each subspace?
How well are the points distributed on each subspace?
This lecture: mixture model

- Fixed subspaces (arbitrary)
- Random point distribution near those subspaces

\[ y = x + z \quad x \sim \sum \alpha_S F_S \]

\[ z \text{ normal} \]

\( F_S \) puts mass 1 on subspace \( S \)

Other models are possible
Data

\(N\) points in \(n\) dimensions \(Y = [y_1, y_2, \ldots, y_N]\)

\[Y = X + Z \iff y_j = x_j + z_j\]

- \(X\): clean data points
- Columns of \(X\) belong to union of unknown subspaces \(S_1 \cup S_2 \cup \ldots \cup S_L\) of unknown dimensions \(d_1, \ldots, d_L\)
- \(Z\): noise
- \(Z_{ij}\) iid \(\mathcal{N}(0, \sigma^2/n)\) \(\rightarrow \|z_j\|_2 \approx \sigma\)

\[
Y = \begin{bmatrix}
X^{(1)} & X^{(2)} & X^{(3)} & \cdots & X^{(L)} \\
\end{bmatrix}
\]

\[
X^{(1)} & X^{(2)} & X^{(3)} & \cdots & X^{(L)} \\
\end{bmatrix}
\]

\[
N = \begin{bmatrix}
N_1 & N_2 & N_3 & \cdots & N_L \\
\end{bmatrix}
\]

\[
N_1 & N_2 & N_3 & \cdots & N_L \\
\end{bmatrix}
\]

\[
Z \sim \mathcal{N}(0, \frac{\sigma^2}{n})
\]

\[
Z \sim \mathcal{N}(0, \frac{\sigma^2}{n})
\]
Number of points per subspace

Assume clean points
- are distributed uniformly at random on each subspace
- have unit norm

**Definition (Sampling density)**

\[ N_\ell \text{ points on subspace } S_\ell \text{ of dimension } d_\ell \]

\[ \rho_\ell = \frac{N_\ell}{d_\ell} \]
Number of points per subspace

Assume clean points
- are distributed uniformly at random on each subspace
- have unit norm

**Definition (Sampling density)**

$N_\ell$ points on subspace $S_\ell$ of dimension $d_\ell$

$$\rho_\ell = \frac{N_\ell}{d_\ell}$$

- $\rho < 1$ → not enough samples to span all directions
- Would like efficient methods operating at small values of $\rho$
Methods
1. Construct affinity matrix $W$ between samples $\rightarrow$ weighted graph
2. Construct clusters by applying spectral clustering to $W$
3. Apply PCA to each cluster

Ideal affinity matrix
Sparse subspace clustering:

(Noiseless $X = Y$) Compute affinities via $\ell_1$ minimization

$$W_{i,j} = |\beta_j^{(i)}| + |\beta_i^{(j)}| \quad \beta^{(i)} = \text{arg min} \|\beta\|_{\ell_1} \text{ subject to } x_i = X\beta$$

Motivation
- $\beta_j \neq 0 \rightarrow x_j$ belongs to same cluster
- $x_j$ belongs to other cluster $\rightarrow \beta_j = 0$

Algorithm+some theory
- Elhamifar and Vidal ('09)

Theory
- Soltanolkotabi and Candes ('11)
Robust version?

\[
\begin{align*}
\text{minimize} & \quad \|\beta\|_{\ell_1} \\
\text{subject to} & \quad y_i = Y \beta \\
& \quad \beta_i = 0
\end{align*}
\]

- Apply sparse regression techniques to build affinity graph
  - Lasso
  - Dantzig selector
- Estimate number of clusters and proceed as before
Robust SSC with LASSO

For each $i = 1, \ldots, N$

$$\text{minimize } \frac{1}{2} \| y_i - Y \beta \|_{\ell_2}^2 + \lambda_i \| \beta \|_{\ell_1} \quad \text{subject to } \beta_i = 0$$

- Response is noisy
- Covariates are noisy
- Nonstandard setup and analysis
For each $i = 1, \ldots, N$

$$\text{minimize} \quad \frac{1}{2} \|y_i - Y\beta\|_2^2 + \lambda_i \|\beta\|_1 \quad \text{subject to} \quad \beta_i = 0$$

- Response is noisy
- Covariates are noisy
- Nonstandard setup and analysis

- How should we pick $\lambda_i$?
Robust SSC with LASSO

For each \(i = 1, \ldots, N\)

\[
\text{minimize} \quad \frac{1}{2} \| y_i - Y \beta \|_{\ell_2}^2 + \lambda_i \| \beta \|_{\ell_1} \quad \text{subject to} \quad \beta_i = 0
\]

- Response is noisy
- Covariates are noisy
- Nonstandard setup and analysis

How should we pick \(\lambda_i\)?
Is this provably accurate?
False and true discoveries

\[
\text{minimize} \quad \frac{1}{2} \|y_i - Y\beta\|_2^2 + \lambda \|\beta\|_1 \quad \text{subject to} \quad \beta_i = 0
\]

**Definition (False discovery)**

Call \(\hat{\beta}_j \neq 0\) a false discovery if \(j\)th column not in same cluster

**Definition (True discovery)**

Call \(\hat{\beta}_j \neq 0\) a true discovery if \(j\)th column in same cluster
Performance criteria

minimize $\frac{1}{2} \|y - Y\beta\|_2^2 + \lambda \|\beta\|_1$

Natural trade off:

- Want large $\lambda \longrightarrow$ few false discoveries
- Want small $\lambda \longrightarrow$ many true discoveries
Performance criteria

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \| y - Y\beta \|_{\ell_2}^2 + \lambda \| \beta \|_{\ell_1} \\
\end{align*}
\]

Natural trade off:

- Want large \(\lambda\) \(\longrightarrow\) few false discoveries
- Want small \(\lambda\) \(\longrightarrow\) many true discoveries

In noiseless case (\(\min \| \beta \|_{\ell_1} \text{ s.t. } x = X\beta\))

\(y\) belongs to subspace of dim. \(d\)

\[\begin{array}{c}
\text{no false discovery} \\
\end{array}\]  \(\implies\) \(d\) true discoveries
Performance criteria

\[
\text{minimize } \frac{1}{2} \| \mathbf{y} - \mathbf{Y}\beta \|_2^2 + \lambda \| \beta \|_1
\]

Natural trade off:
- Want large \( \lambda \rightarrow \) few false discoveries
- Want small \( \lambda \rightarrow \) many true discoveries

In noiseless case (min \( \| \beta \|_1 \) s.t. \( \mathbf{x} = \mathbf{X}\beta \))

\( \mathbf{y} \) belongs to subspace of dim. \( d \)
\[ + \quad \Rightarrow \quad d \text{ true discoveries} \]
\[ \text{no false discovery} \]

Proposal

Select \( \lambda \) as large as possible while maintaining \([0.5d, 0.8d]\) true discoveries

Oracle selection: \( \lambda \approx 1/\sqrt{d} \)
Regularization vs. true discoveries

Suppose $y$ and $Y$ noisy random sampled from same $d$-dim subspace.

$$\sigma = 0.25$$

$$\sigma = 0.5$$

$y$ axis: $\frac{\# \text{ of true discoveries}}{d}$

$x$ axis: $\lambda/\lambda_o$  \hspace{1em} $\lambda_o = 1/\sqrt{d}$
False vs. true discovery trade off

<table>
<thead>
<tr>
<th>Total subspaces of dim</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>150</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>500</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>22</strong></td>
</tr>
</tbody>
</table>

- 22 subspaces embedded in $n = 2,000$ dimensions
- Sampling density: $\rho = 5$ (points/dim)
- Full rank (clean) matrix
False vs. true discovery trade off

\[ \text{FPR} = \frac{\text{number of false discoveries}}{n - d} \]

\[ \text{TPR} = \frac{\text{number of true discoveries}}{d} \]
Adaptive regularization?

\[
\text{minimize } \frac{1}{2} \| y_i - Y \beta \|_{\ell_2}^2 + \lambda \| \beta \|_{\ell_1} \quad \text{subject to } \beta_i = 0
\]

Regularization parameter \( \lambda \) should be data dependent

- large for small subspaces (enforce sparsity)
- small for large subspaces (relax sparsity)
Adaptive regularization?

\[
\text{minimize} \quad \frac{1}{2} \left\| y_i - Y \beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 \quad \text{subject to} \quad \beta_i = 0
\]

Regularization parameter \( \lambda \) should be data dependent
- large for small subspaces (enforce sparsity)
- small for large subspaces (relax sparsity)

Oracle estimation:

\[
\lambda \approx \frac{1}{\sqrt{d(i)}}
\]
Adaptive regularization?

\[
\text{minimize} \quad \frac{1}{2} \| y_i - Y \beta \|_{l_2}^2 + \lambda \| \beta \|_{l_1} \quad \text{subject to} \quad \beta_i = 0
\]

Regularization parameter \( \lambda \) should be data dependent
- large for small subspaces (enforce sparsity)
- small for large subspaces (relax sparsity)

Oracle estimation:

\[
\lambda \approx \frac{1}{\sqrt{d(i)}}
\]

Problem: subspace dimension is unknown
Data-driven regularization

**Algorithm 1** Two-step procedure with data-driven regularization

```plaintext
for i = 1, . . . , N do
1. Solve

   \[ \beta^* = \operatorname{arg\,min}_{\beta} \| \beta \|_{\ell_1} \quad \text{subject to} \quad \| y_i - Y \beta \|_{\ell_2} \leq \tau \quad \text{and} \quad \beta_i = 0 \]

2. Set \( \lambda_i = f(\| \beta^* \|_{\ell_1}) \) \((f(t) \propto t^{-1})\) and solve

   \[ \hat{\beta} = \operatorname{arg\,min}_{\beta} \frac{1}{2} \| y_i - Y \beta \|_{\ell_2}^2 + \lambda_i \| \beta \|_{\ell_1} \quad \text{subject to} \quad \beta_i = 0 \]

end for
```

**Insight:** \( \| \beta^* \|_{\ell_1} \propto \sqrt{d} \)
Optimal value for subspaces of dim $d = 200, 150, 100, 50, 20, 10$ and $\tau = 2\sigma$

(a) Value of $\|\beta^*\|_{\ell_1}$

(b) Value of $\|\beta^*\|_{\ell_1} / \sqrt{d}$
Theory
Assumptions

Simplifying (unessential) assumption

\[ d \lesssim \frac{n}{\log^2 N} \]

Key assumption

\[ \sigma \leq \sigma^* \text{ (numerical constant)} \quad y = x + z \quad \mathbb{E} \|z\|_{\ell_2}^2 = \sigma^2 \]

\[ \|x\|_{\ell_2}^2 = 1 \]
No false discoveries

Theorem

Apply two-step procedure with $\tau = 2\sigma$ and $f(t) \geq 0.707\sigma/t$ to $\mathbf{y} \in S$

- Sampling density obeys $\rho(S) \geq \rho^*$
- For all other subspaces $S'$

$$\text{aff}(S, S') = \sqrt{\text{Ave}(\cos^2 \theta)} \lesssim (\log N)^{-1}$$

With high prob., no false discovery
No false discoveries

**Theorem**

Apply two-step procedure with \( \tau = 2\sigma \) and \( f(t) \geq 0.707\sigma/t \) to \( y \in S \)

- **Sampling density obeys** \( \rho(S) \geq \rho^* \)
- **For all other subspaces** \( S' \)

\[
\text{aff}(S, S') = \sqrt{\text{Ave}(\cos^2 \theta)} \lesssim (\log N)^{-1}
\]

*With high prob., no false discovery*

Near information-theoretic limit

- **Constant number of points per dimension:** \( \rho > \rho^* \)
- **Large affinities**
  - significant overlap between subspaces
  - dimension of overlap can grow linearly with \( d \)
Comparison with perfect data: $Y = X \ (Z = 0)$

$$\text{aff}(S, S') \lesssim (\log N)^{-1} \cdot (\log \rho(S))^{1/2}$$

Noiseless case with $\lambda = 0^+$ [S. and Candes ('11)]

- No false discovery
- $d(S)$ true discoveries (neighbors)

$$\|\hat{\beta}\|_{\ell_0} = d(S)$$
Are no-false discoveries enough?

- Easy to get no false discovery no matter noise level ($\lambda \to \infty$ and $\hat{\beta} = 0$)
- Need many true discoveries for correct clustering
Many true discoveries

Theorem

Apply two-step procedure with \( \tau = 2 \sigma \) and \( f(t) \lesssim 1/t \) to \( \mathbf{y} \in S \). Then with high prob., we have

\[
\| \hat{\beta} \|_{\ell_0} \gtrsim \frac{d(S)}{\log \rho(S)}
\]

Roughly proportional to dimension
Asymptotically, when \( d(S) \to \infty \) with \( \rho(S) \) fixed, can take \( f(t) = 0.25t^{-1} \).
Advantages of data-driven approach

- Automatically adapts to approx. correct dimension
- Can have large and small subspaces simultaneously
- Can handle constant SNR since $\sigma \approx \frac{\|z\|_{\ell_2}}{\|x\|_{\ell_2}}$ can be constant
- Accurate subspace estimation
Other Approaches
Other approaches to subspace clustering

- **Algebraic methods**: Generalized PCA
  - Vidal, Ma and Sastry (05)
  - Ma, Yang, Derksen and Fossum (08)
  - Ozay, Sznaier, and Lagoa (10)

- **Iterative methods**:
  - Tseng (00): nearest $q$-flats to $m$ points

- **Statistical methods**: Mixture of singular Gaussians
  - Tipping and Bishop (99): MLE via EM-style algorithm
  - Ma, Derkesen, Hong, and Wright (07): minimize coding length

- **Others**:
  - Lerman and Zhang (11): Fitting via nonconvex minimization over Grassmanian
  - Eriksson, Balzano and Nowak (11): High-rank matrix completion
  - Liu et al. (12): Low-rank representation (LRR)

...
Three important questions

- Is the algorithm computationally tractable?
- Is the algorithm provably robust to noise?
- Can the algorithm function in a challenging regime?
  - high dimensions
  - high noise
  - high affinity
  - low-sampling density
Three important questions

- Is the algorithm computationally tractable?
- Is the algorithm provably robust to noise?
- Can the algorithm function in a challenging regime?
  - high dimensions
  - high noise
  - high affinity
  - low-sampling density

<table>
<thead>
<tr>
<th>Tractability</th>
<th>Robustness</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust Subspace Clustering (RSC)</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>
Three important questions

- Is the algorithm computationally tractable?
- Is the algorithm provably robust to noise?
- Can the algorithm function in a challenging regime?
  - high dimensions
  - high noise
  - high affinity
  - low-sampling density

<table>
<thead>
<tr>
<th>Robust Subspace Clustering (RSC)</th>
<th>Tractability</th>
<th>Robustness</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Other methods may not be tractable, may be sensitive to noise, may require super-polynomial density, may require near orthogonality between subspaces, etc.
Comparison with Sparse regression literature

- Many analysis of LASSO: [Bickel et. al. 08], [Wainwright 09], [Candes and Plan 09, 11], [Bayati and Montanari 12], . . .
Comparison with Sparse regression literature

- Many analysis of LASSO: [Bickel et. al. 08], [Wainwright 09], [Candes and Plan 09, 11], [Bayati and Montanari 12], . . .
  - response and covariates (dictionary) are noisy.
- Regression with corrupted covariates: [Tsybakov et. al. 08, 12], [Loh and Wainwright 12], . . .
Comparison with Sparse regression literature

- Many analysis of LASSO: [Bickel et. al. 08], [Wainwright 09], [Candes and Plan 09, 11], [Bayati and Montanari 12], . . .
  - response and covariates (dictionary) are noisy.
- Regression with corrupted covariates: [Tsybakov et. al. 08, 12], [Loh and Wainwright 12], . . .
  - not LASSO.
  - No target solution/no sparsity.
  - guarantees on support vs. closeness in some norm.
  - Completely different modeling assumptions (standard assumptions are violated).
Numerical Examples
Motion capture Data

Subject 86 from www.mocap.cs.cmu.edu walking, squating, punching, standing, running, jumping, arms-up, and drinking.
Motion capture data
Subspace model

Figure: Eight activities and singular values of the data from three activities
Comparison baseline: KNN

- Build similarity graph based on kernel $B_{ij} = e^{-\|y_i - y_j\|^2_2 / 2\tau^2}$
- Truncate $B$ using $K$ nearest neighbors
- Apply spectral clustering

Minimum clustering error (%) for each $K$
<table>
<thead>
<tr>
<th>Trial</th>
<th>KNN</th>
<th>Robust SSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>17.06%</td>
<td>3.54%</td>
</tr>
<tr>
<td>5</td>
<td>12.47%</td>
<td>4.35%</td>
</tr>
</tbody>
</table>
Robust version of SSC

Parameters: $\tau = 2\sigma$ and $\lambda_o = \frac{1}{4\|\beta^*\|_{\ell_1}} \left( f(t) = \frac{1}{4t} \right)$
Computational issues
run time?

sequence of $\ell_1$ problems $\rightarrow$ highly parallel
run time?

sequence of $\ell_1$ problems $\rightarrow$ highly parallel

what if I want to run this sequentially?
Gene expression level of different groups of patients in *St. Jude Leukemia* data set

Six different cancer types
## No missing data

<table>
<thead>
<tr>
<th></th>
<th># clusters</th>
<th># genes/biological samples</th>
<th># patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>3</td>
<td>999</td>
<td>38</td>
</tr>
<tr>
<td>St. Jude Leukemia</td>
<td>6</td>
<td>985</td>
<td>248</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>4</td>
<td>1000</td>
<td>197</td>
</tr>
<tr>
<td>Novartis multi-tissue</td>
<td>4</td>
<td>1000</td>
<td>103</td>
</tr>
</tbody>
</table>
## No missing data

<table>
<thead>
<tr>
<th></th>
<th># clusters</th>
<th># genes/biological samples</th>
<th># patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>3</td>
<td>999</td>
<td>38</td>
</tr>
<tr>
<td>St. Jude Leukemia</td>
<td>6</td>
<td>985</td>
<td>248</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>4</td>
<td>1000</td>
<td>197</td>
</tr>
<tr>
<td>Novartis multi-tissue</td>
<td>4</td>
<td>1000</td>
<td>103</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>KNN</th>
<th>K-means</th>
<th>Robust SSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>0%</td>
<td>2.63%</td>
<td>5.26%</td>
</tr>
<tr>
<td>St. Jude Leukemia</td>
<td>11.69%</td>
<td>17.34%</td>
<td>9.68%</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>15.74%</td>
<td>30.96%</td>
<td>5.58%</td>
</tr>
<tr>
<td>Novartis multi-tissue</td>
<td>6.80%</td>
<td>35.92%</td>
<td>2.91%</td>
</tr>
</tbody>
</table>
In the last portion of the talk I also discussed how to provably handle missing data in subspace clustering based on joint work with Emmanuel Candes and Lester Mackey, which is not included in these online slides. The methodology and experiments, along with the forthcoming paper will become available on www.stanford.edu/~mahdisol/RSC in the near future.
Summary

Principled approach to subspace clustering
- Computationally tractable
- Works well in practice
- Rigorous theory
- Many applications (some to come very soon)


Last portion of the talk: joint work with Emmanuel Candes and Lester Mackey, In preparation 2013.

Code, latest updates available at: www.stanford.edu/~mahdisol/RSC