Multi-Subspace Representation and Discovery

Dijun Luo Feiping Nie **Chris Ding** Heng Huang
Dept of Computer Science & Engineering
University of Texas at Arlington

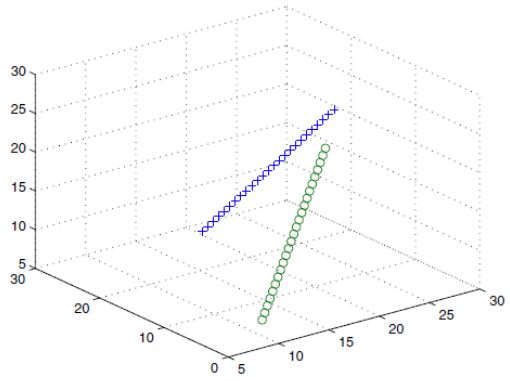
Outline

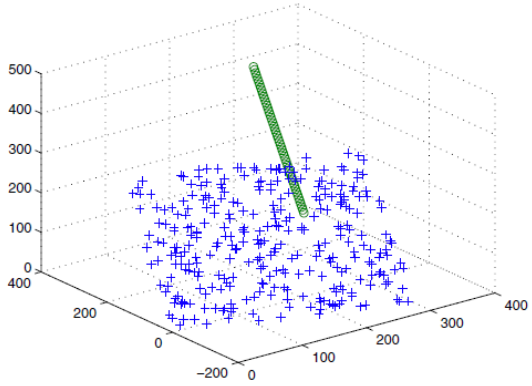
- Introduction
- Background and related work
- Problem formulation
- Our solution
- Theoretical analysis
- Empirical studies
- Conclusions

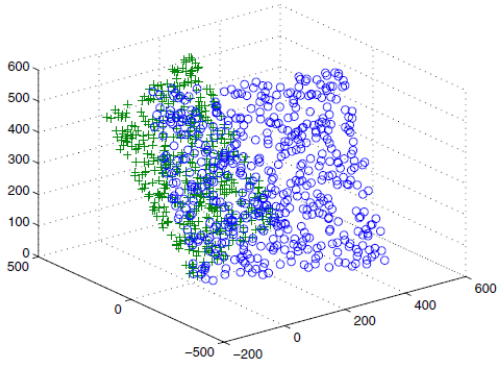
Multi-subspace

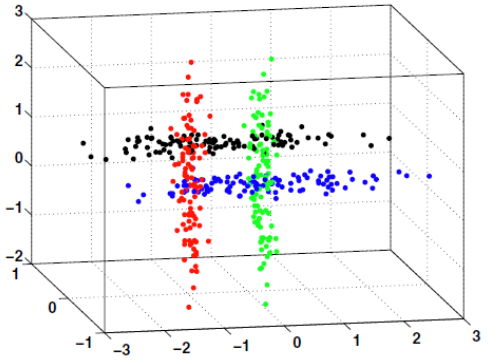
- Data distribution has multiple linear subspaces (extended clusters live in low-dimensions)

  Example: Data points live on a 1D line in 10-dimensional space



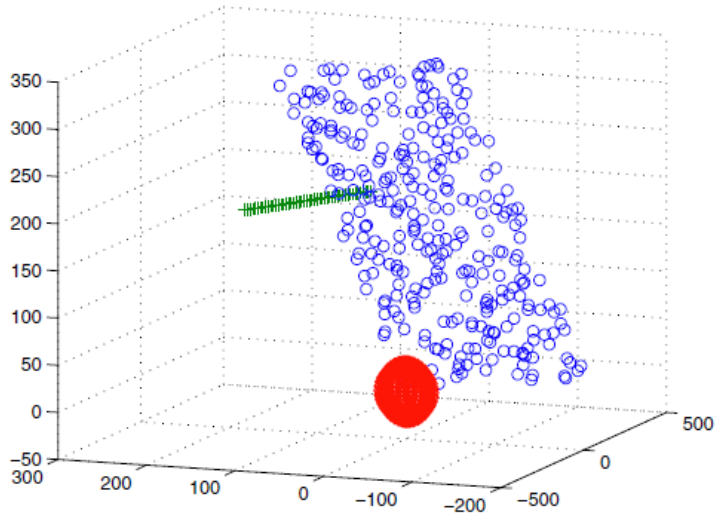






More challenging data distribution: Multi-subspace + Sold Clusters

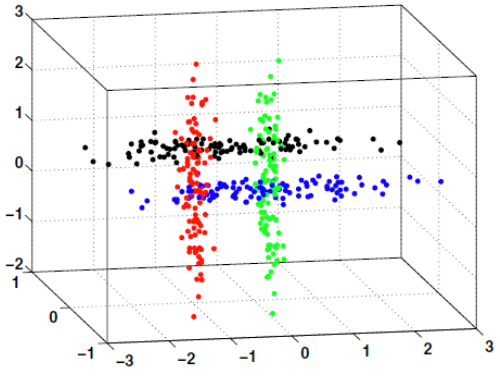
- **Linear subspaces** (extended clusters live in low-dimensions)
- **Solid clusters** (limited linear extension, but live in higher dimensions)

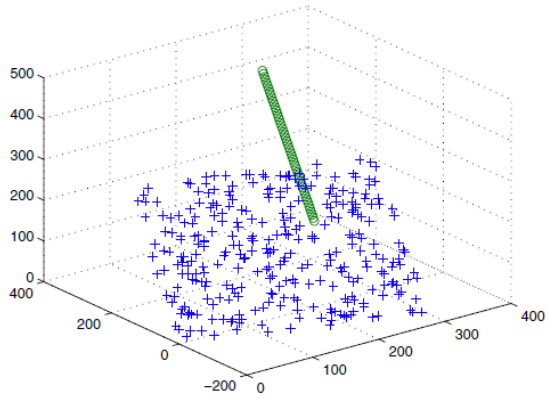


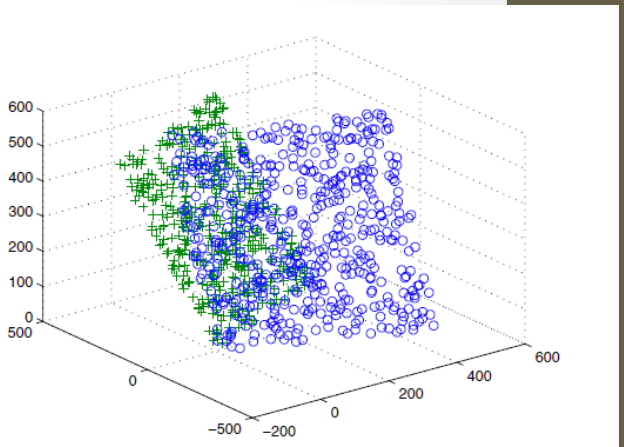
- Use PCA to approximate subspaces
- Detect solid clusters

(Wang, Ding, Li, ECML PKDD 2009)

Data as multi-subspaces







- Earlier research: subspace clustering
- Explicit search in different subspaces
- CLIQUE, MAFIA, CBF, CLtree, Proclus, FINDIT (Survey by Parsons et al)

- New approach: Using sparse coding

Sparse Representation

- The assumption is that data points are represented by the linear (convex or affine) combinations of their neighbors.
  - Perhaps the simplest assumption in representation
  - Intuitive, used in many earlier works (LLE)
  - New emphasis is sparse (not necessarily near neighbors)
- Sparse representation models have been widely studied
  - Simple model
  - Robust performance
  - Sound theoretical foundations [Jenatton 2009, Candes 2008]
  - Works well in many machine learning and data mining applications [Wright 2009, Lin 2010]

Sparse Representation

Generic sparse representation $x_i \approx Xz_i$

given a set of training data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ ($p \times n$ matrix, where p is the dimension of the data) and a testing data point $\mathbf{x}_t$, they solve the following optimization problem

$$\min_{\alpha_t} \|\mathbf{x}_t - \mathbf{X}\alpha_t\|^2 + \lambda\|\alpha_t\|_1, \qquad (16)$$

where α_t ($n \times 1$ vector) has the reconstruction coefficients of x_t

Let t=1,2, …, n, we solve for all representation simultaneously:

$$\min_{Z} \| X - XZ \|^2 + \lambda \| Z \|_1$$

Multi-Subspace Representation

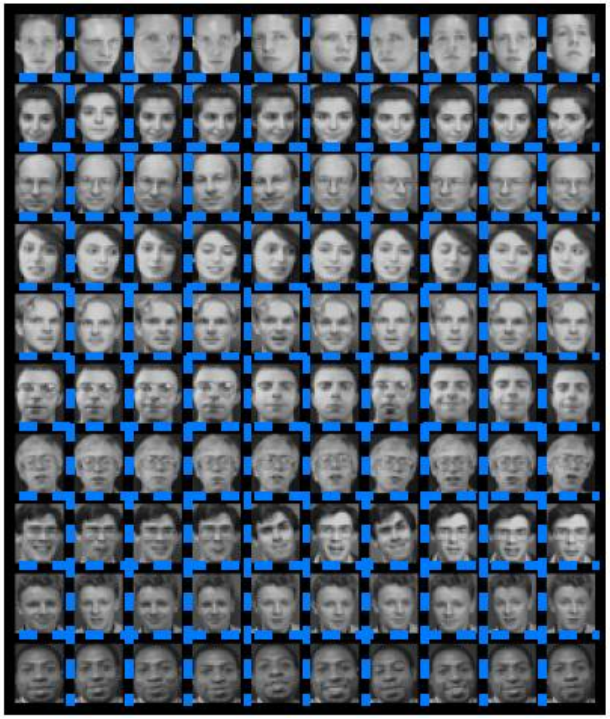
Generic sparse representation $x_i \approx X z_i$

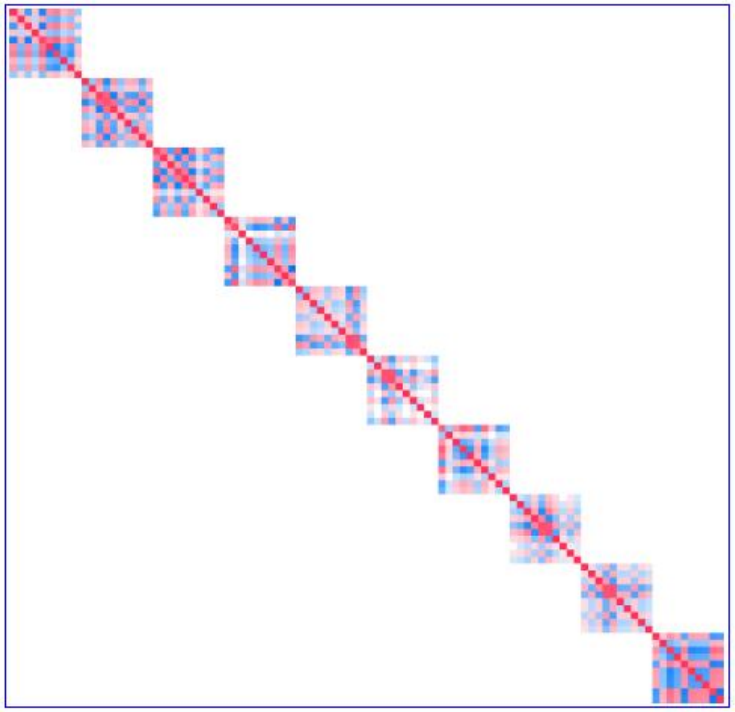
Multi- subspace representation: $X \approx XZ$

where Z has block diagonal structure:

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{Z}_K \end{pmatrix}$$

The Challenges



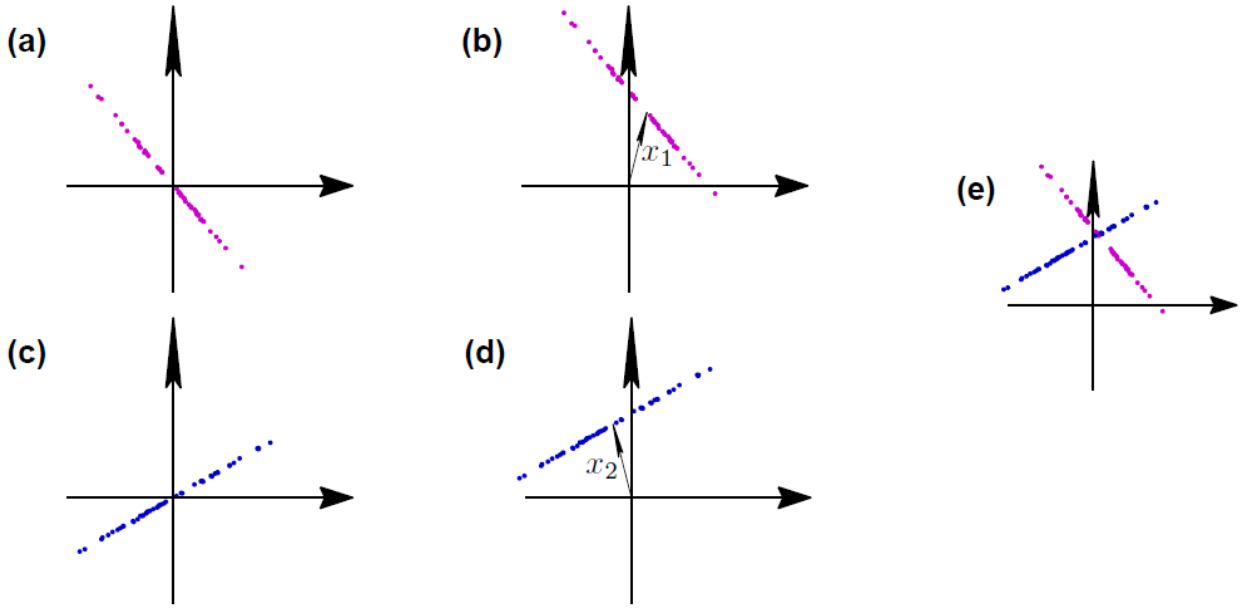


1. The number of subspaces is unknown
2. The dimensions of the subspaces are unknown
3. The memberships of the data points are also unknown

Our Contributions

- Theory
  - Explicit construction of multi-subspace representation
  - Affine construction such that subspace no longer required to pass feature space origin.
  - Reduce strong block structure assumption to weaker assumption.
  - Better understanding and interpretation
- Algorithm
  - An efficient algorithm to compute the solution
  - Guaranteed to converge to global solution
- A new sparse representation based classification and semi-supervised classification method

Affine construction of Multi-subspace



- Affine combination such that contribution to each data point is equal-weighted

$$\sum_{i=1}^{n} Z_{ij} = 1$$

- Padding extra dimension such that subspaces may locate away from feature space origin

$$x_i \Rightarrow \begin{pmatrix} x_i \\ 1 \end{pmatrix}$$

Problem Formulation

- Consider K groups of data points (totally n data points)
  - $X = [X_1, X_2, \dots, X_K]$ with $n_1, n_2, \dots, n_K$ data points
  - $n_1 + n_2 + \dots + n_K = n$
- The dimensions of the subspaces are
  - $d_1, d_2, \dots, d_K$
- For each subspace X_k, there exists $d_k + 1$ bases
  - $U_k = [u_1, u_2, \dots u_{dk+1}]$, such that for each data point $\mathrm{x} \in X_k$, there exist β: $\mathrm{x} = X_k\beta, \beta^T 1 = 1$
  -

Explicit Subspace Construction for K=1

- A constructive solution for K=1

$$\mathbf{x}_i = \mathbf{U}_1 \boldsymbol{\alpha}_i,\ \boldsymbol{\alpha}_i \in \mathbb{R}^{d_1+1},\ \boldsymbol{\alpha}_i^T \mathbf{1} = 1,\ 1 \le i \le n_1$$

- Or using matrix form

$$\mathbf{X}_1 = \mathbf{U}_1 \mathbf{A},\ \mathbf{A}^T \mathbf{1} = \mathbf{1} \qquad A = (\alpha_1 \cdots \alpha_{n_1})$$

- Let $\tilde{\mathbf{X}}_1 = \begin{pmatrix} \mathbf{U}_1 \mathbf{A} \\ \mathbf{1}^T \end{pmatrix}$

- Then $\mathbf{X}_1 = \mathbf{X}_1 \mathbf{Z}_1,\ \mathbf{1}^T \mathbf{Z}_1 = \mathbf{1}^T$
- Where $\mathbf{Z}_1 = \tilde{\mathbf{X}}_1^+ \tilde{\mathbf{X}}_1$ $\qquad X_1^+ = (X_1^T X_1)^{-1} X_1^T$ Pseudo inverse

Explicit Subspace Construction for K ≥ 1

- For arbitrary K
  - If $X = [X_1, X_2, \dots, X_K]$ belong to K subspaces, then there exists a Z such that
    - $X = XZ, Z^T 1 = 1$ and

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{Z}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{Z}_K \end{pmatrix}$$

$$rank(\mathbf{Z}_k) = d_k + 1, 1 \leq k \leq K$$

$$\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_K] = [\mathbf{X}_1\mathbf{Z}_1, \cdots, \mathbf{X}_K\mathbf{Z}_K] = \mathbf{XZ}$$

Reformulation of Our Construction

When data consists of exactly multi-subspaces

- For the following optimization

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_*,$$
$$\text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{Z},\ 1^T\mathbf{Z} = 1^T$$

- We have one of the optimal solution as following

$$\mathbf{Z}^* = \tilde{\mathbf{X}}^+\tilde{\mathbf{X}}$$

where

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X} \\ 1^T \end{pmatrix}$$

Multi-subspace Discovery

When data is approximately multi-subspaces:

- Typically Z is not block diagonal. Inspired by Lemma 1, Theorems 1 and 2, we solve the following problem

$$\min_{\mathbf{Z}} J_1(\mathbf{Z}) = \|\mathbf{Z}\|_* + \delta\|\mathbf{Z}\|_1$$
$$\text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{Z},\ \mathbf{1}^T\mathbf{Z}_1 = \mathbf{1}^T,$$

Our Model

- Typically Z is not block diagonal. Inspired by Lemma 1, Theorems 1 and 2, we solve the following problem

Low rank Sparse

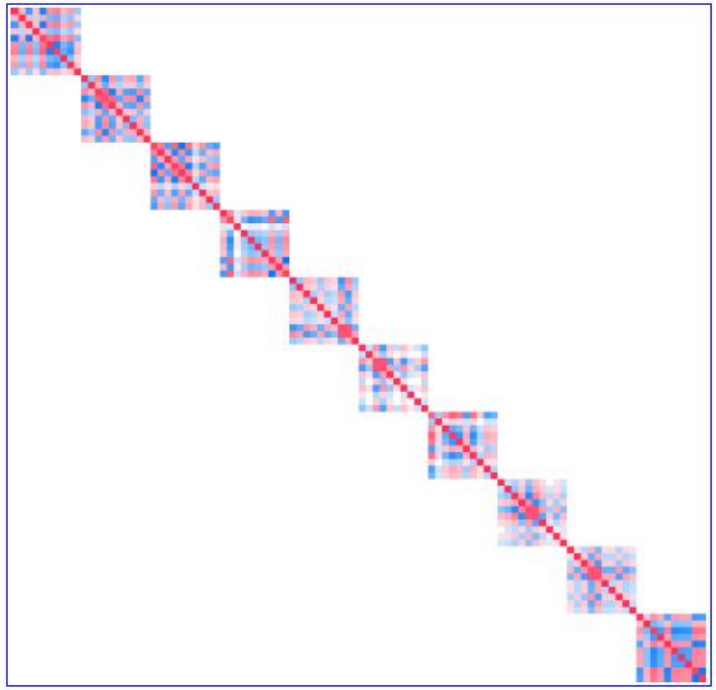
$$\min_{\mathbf{Z}} J_1(\mathbf{Z}) = \|\mathbf{Z}\|_* + \delta\|\mathbf{Z}\|_1$$
$$\text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}, \; \mathbf{1}^T\mathbf{Z}_1 = \mathbf{1}^T,$$

Self representation Affine subspace

The solution of the above problem is guaranteed to have the block diagonal structure. Proposition 1

Example: Large feature size

(a)



(b)

A. Compute SVD(X). Subtract the smallest singular value term.
B. Find the solution Z according to Theorem 1.

The Algorithm

Algorithm 1 $(\mathbf{X}, \lambda, \delta)$

Input: Data $\mathbf{X}$, model parameters λ, δ
Output: $\mathbf{Z}$ which optimizes Eq.(15).
Initialization: Compute $\tilde{\mathbf{X}}$ using Eq. (12), $\mathbf{Z} = \mathbf{0}$.
while not converged **do**
  $\mathbf{B} = \left(\mathbf{Z}\mathbf{Z}^T + \epsilon\mathbf{I}\right)^{-1/2}$
  for $i = 1 : n$ **do**
    $\mathbf{d}_i = \|\tilde{\mathbf{x}}_i - \tilde{\mathbf{X}}\mathbf{z}_i\|,$
    $\mathbf{D}_i = \mathbf{diag}\left(Z_{1i}^{-1}, Z_{2i}^{-1}, \cdots, Z_{ni}^{-1}\right),$
    $\mathbf{z}_i = \left[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} + \lambda\mathbf{d}_i\left(\mathbf{B} + \delta\mathbf{D}\right)\right]^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{x}}_i,$
  end for
end while
Output: $\mathbf{Z}$

Three key theoretical results

Lemma 3.

$$\|\mathbf{Z}\|_* = \lim_{\epsilon \to 0} \mathbf{tr}\left(\mathbf{Z}\mathbf{Z}^T + \epsilon \mathbf{I}\right)^{1/2},$$

and

$$\lim_{\epsilon \to 0} \left(\mathbf{Z}\mathbf{Z}^T + \epsilon \mathbf{I}\right)^{-1/2} \mathbf{Z} \in \partial \|\mathbf{Z}\|_*,$$

where $\partial \|\mathbf{Z}\|_*$ *is the subgradient of trace norm.*

Lemma 4. *Assume matrices* $\mathbf{Z}$ *and* $\mathbf{Y}$ *have the same size. Let* $\mathbf{A} = \left(\mathbf{Y}\mathbf{Y}^T + \epsilon \mathbf{I}\right)^{1/2}$ *and* $\mathbf{B} = \left(\mathbf{Z}\mathbf{Z}^T + \epsilon \mathbf{I}\right)^{1/2}$. *Then the following holds*

$$\mathbf{tr}\mathbf{A} - \mathbf{tr}\mathbf{B} + \frac{1}{2}\mathbf{tr}\mathbf{Z}^T\mathbf{B}^{-1}\mathbf{Z} - \frac{1}{2}\mathbf{tr}\mathbf{Y}^T\mathbf{B}^{-1}\mathbf{Y} \leq 0. \qquad (23)$$

Theorem 3. *Algorithm 1 monotonically decreases the following objective,*

$$\min_{\mathbf{Z}} J(\mathbf{Z}) = \|\tilde{\mathbf{X}} - \tilde{\mathbf{X}}\mathbf{Z}\|_{\ell_2/\ell_1} + \lambda \mathbf{tr}\left(\mathbf{Z}\mathbf{Z}^T + \epsilon \mathbf{I}\right)^{\frac{1}{2}} + \delta \|\mathbf{Z}\|_1,$$

i.e. $J(\mathbf{Z}_{t+1}) \leq J(\mathbf{Z}_t)$, *where* $\mathbf{Z}_t$ *is the solution of* $\mathbf{Z}$ *in the* t*-th iteration.*

The Algorithm

- Guaranteed to be converged
  - Theoretically guaranteed by Theorem 3
- Guaranteed to converge to global solution
  - Later work for this paper, not published yet
- Converge fast, using very small number of iterations, (10~50)
- Our optimization techniques are applied to general trace norm and l_1 norm optimization
  - Lemmas 3 and Lemma 4

By-product: classification

- Once of the representation is solved, as a by-product, we can do sparse low rank representation classifier

$$\arg\min_k r_k = \|\mathbf{X}\mathbf{z}_t - \hat{\mathbf{x}}_t^k\|, \quad \hat{\mathbf{x}}_t^k = \sum_{i \in C_k} \mathbf{x}_i Z_{it}$$

Representation Error

Choose the class with lowest representation error.

Empirical Studies

- As preprocessing
  - To use XZ instead of X
- Can be used in
  - Clustering
  - Semi-supervised learning
  - Clustering
- Representation based classification

Experiments

- From input data X, compute Z which contains sunspaces
- Use XZ as the corrected/denoised data, do
  - classification
  - clustering
  - semi-supervised learning
- Multi-Subspace Representation (MSR) based classification

Experiments

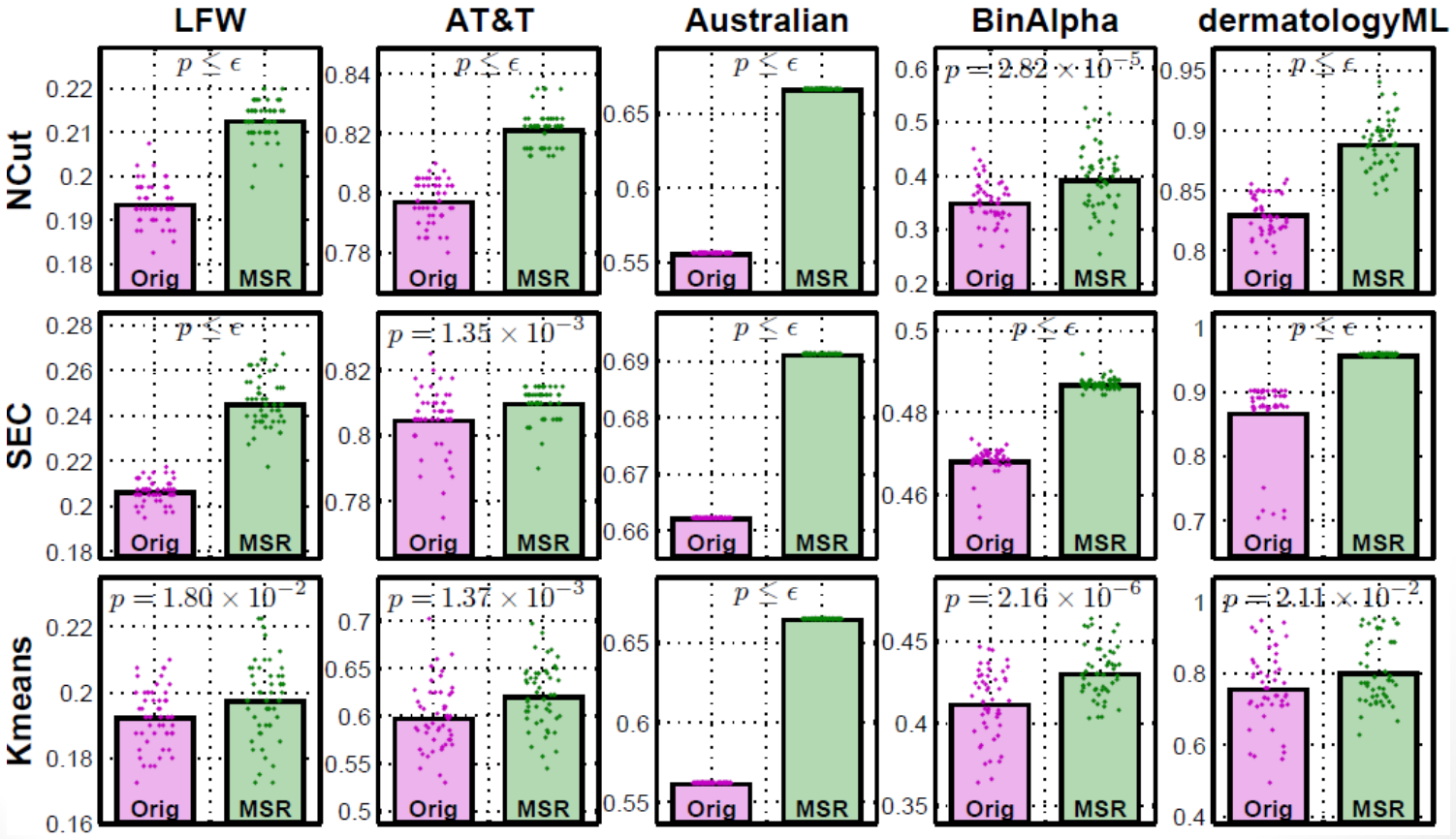
- Data sets
- LFW (labeled Faces in the Wild)
- AT&T face data
- UCI: Australia Sign Language
- UCI: Dermatology
- BinAlpha: hand-written letters

Experiments

- Compared methods
  - **Clustering:**
    - Normalized Cut,
    - Embedded Spectral Clustering,
    - K-means
  - **Classification:**
    - Support Vector Machine
    - KNN
  - **Semi-supervised learning:**
    - local-global consistency
    - harmonic function

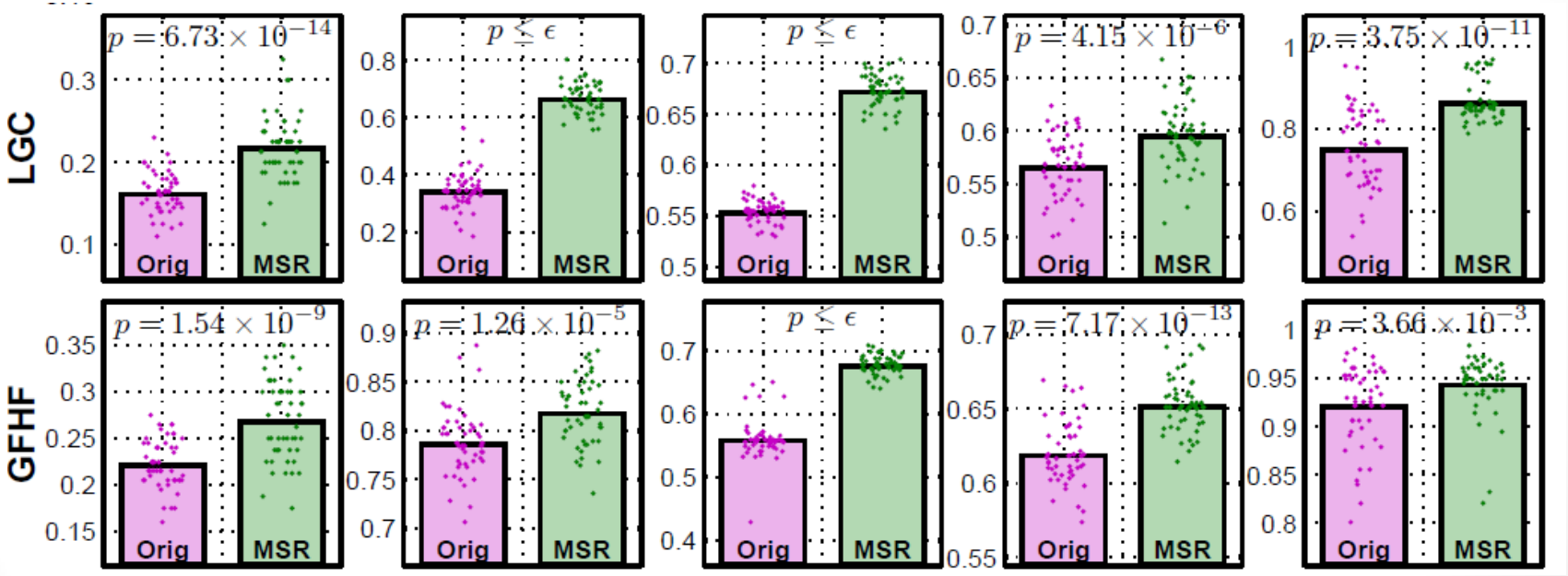
Experiment Results

- Used preprocessing in clustering (Orig: before preprocessing, MSR: preprocessing using our method)



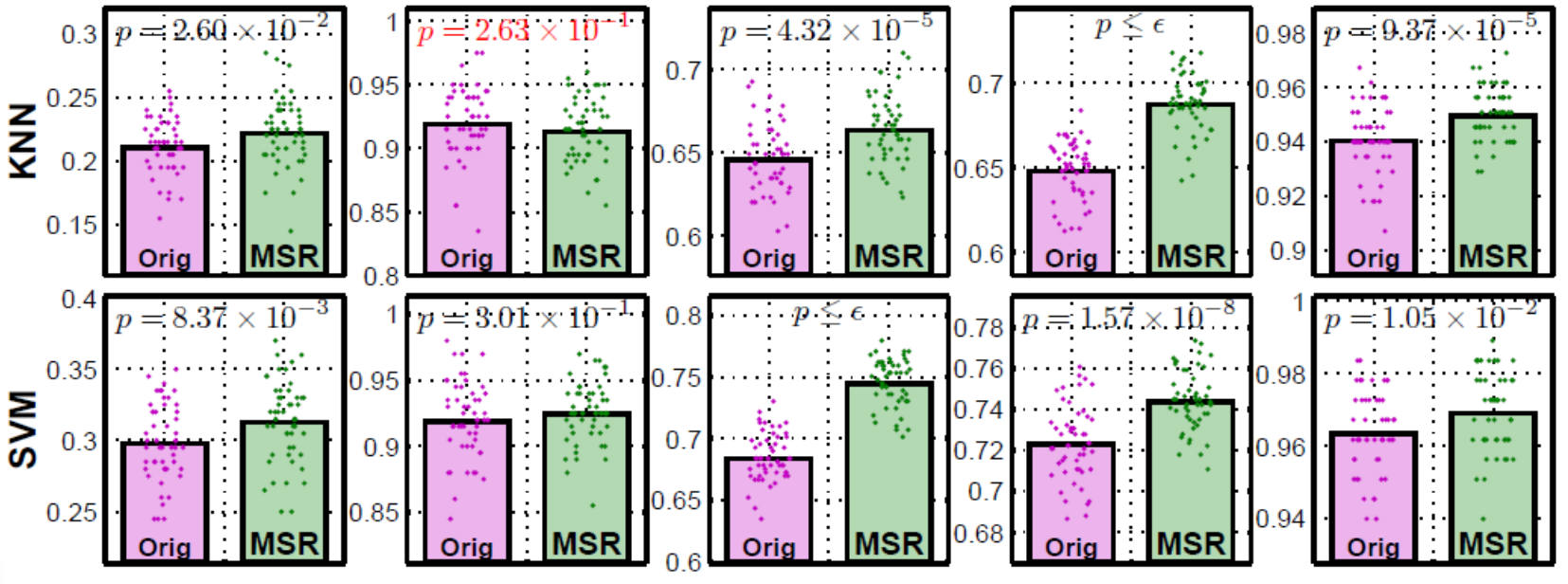
Experiment Results

- Used preprocessing in Semi-supervised Learning (Orig: before preprocessing, MSR: preprocessing using our method)



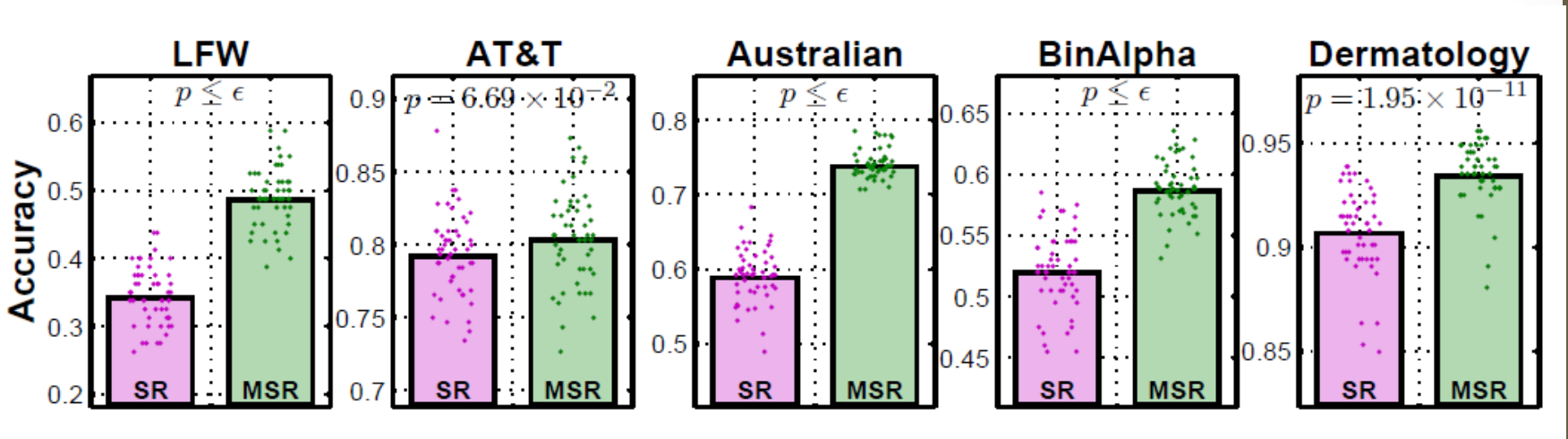
Experiment Results

- Used preprocessing in classification (Orig: before preprocessing, MSR: preprocessing using our method)



Experiment Results

- As representation-based classification (SR: Sparse Representation based classification, Wright 2009, MSR: our method)



Conclusions

- We present multi-subspace representation and discovery model
  - solve the multi-subspace discovery problem by providing block diagonal representation matrix
  - extend our approach to handle noisy real world data
- Efficient optimization algorithm is presented
  - Global optimal solution is guaranteed
- Our optimization technique is general for other trace norm and L1 norm optimization
- Our method can be used in classification, clustering, and semi-supervised learning.

Thank you!

- Questions are welcome!

Introduction

- Sparse representation models have been widely studied
  - Simple model
  - Robust performance
  - Sound theoretical foundations [Jenatton 2009, Candes 2008]
- The assumption is that data points are represented by the linear combinations of their neighbors.
  - Perhaps the simplest assumption in representation
  - Works well in many machine learning and data mining applications [Wright 2009, Lin 2010]
- The linear assumption in previous studies is yet too strong
  - We extend the model with weaker assumption
  - We develop more fundamental properties are represented

Background and Related Work

- Sparse representation
  - To representation data points using a linear but sparse combination of a set of bases.
- Multi-Subspace discovery
  - Given a set of data points, to discover the number of linear subspaces, the dimensions of the subspaces and the membership of the data points to the subspaces
- In previous study
  - Lin et al. presented the fundamental connection between the two in 2010 [Lin et al. ICML 2010]
  - The multi-subspace discovery problem is formulated as sparse representation
  - The assumption is two strong
  - No theoretical guarantee is given for the optimal results

Our Contributions

- Theoretical extension Lin et al. 2010's model
  - Weaker assumption
  - Theoretical guarantees of the multi-subspace discovery problem
- A new classification and semi-supervised learning framework is represented
  - Explicitly leverage the sparse and low rank representation
  - Robust in performance
- A new optimization approached is develop for efficient l_1 and trace norm optimization problems
  - Efficient – requires very small number of iterations to converge
  - Guarantee to converge to global solution
  - Can be easily applied in other related machine learning problems for optimization

Multi-Subspace Representation

Generic sparse representation $x_i \approx Xz_i$

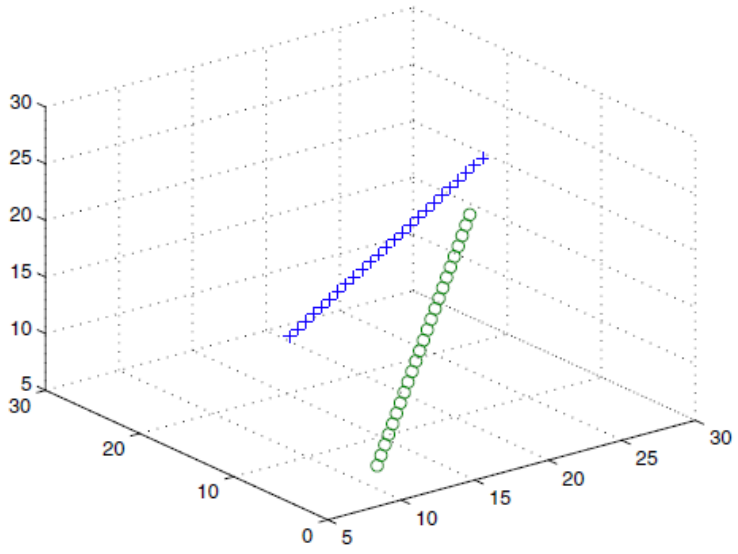
$$\min_Z \| X - XZ \|^2$$

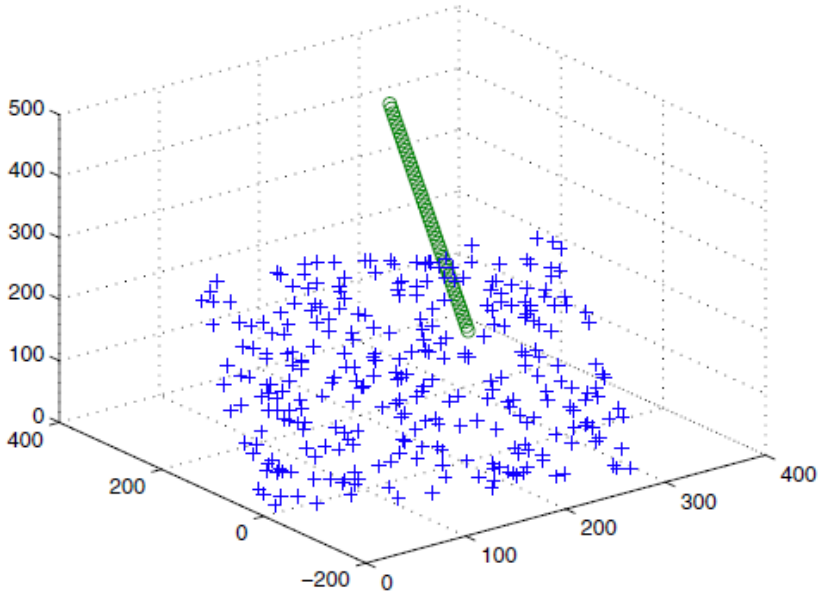
s.t. Z has block diagonal structure

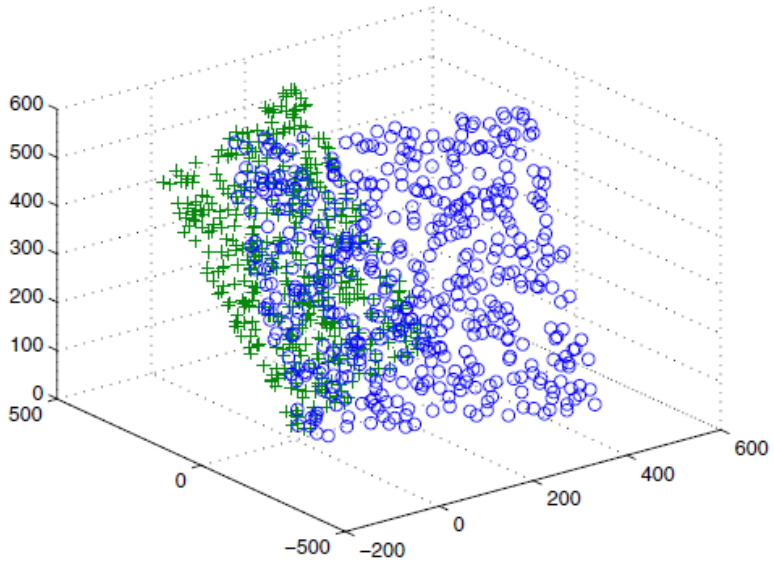
Let t=1,2, …, n

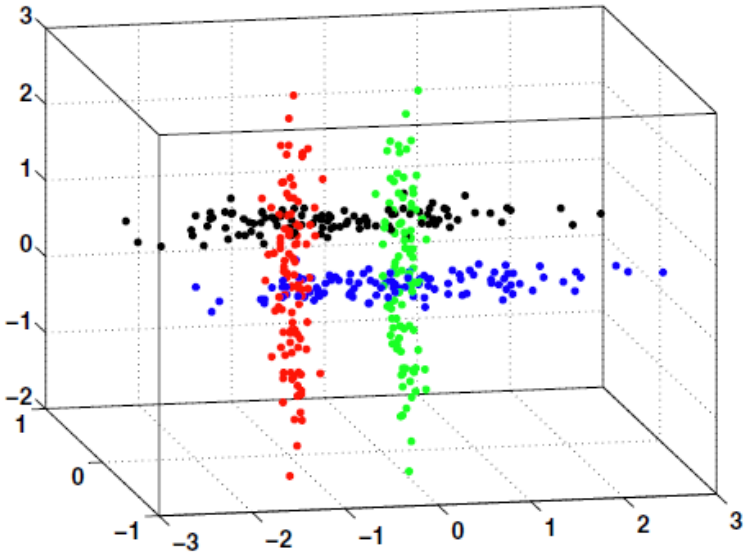
Generic sparse representation

Data as multi-subspaces









The Challenges

- The input is the data points
  - The number of subspaces is unknown
  - The dimensions of the subspaces are unknown
  - The memberships of the data points are also unknown