Robust Subspace Segmentation by Low-Rank Representation ICML 2010

Latent Low-Rank Representation for Subspace Segmentation and Feature Extraction

ICCV 2011

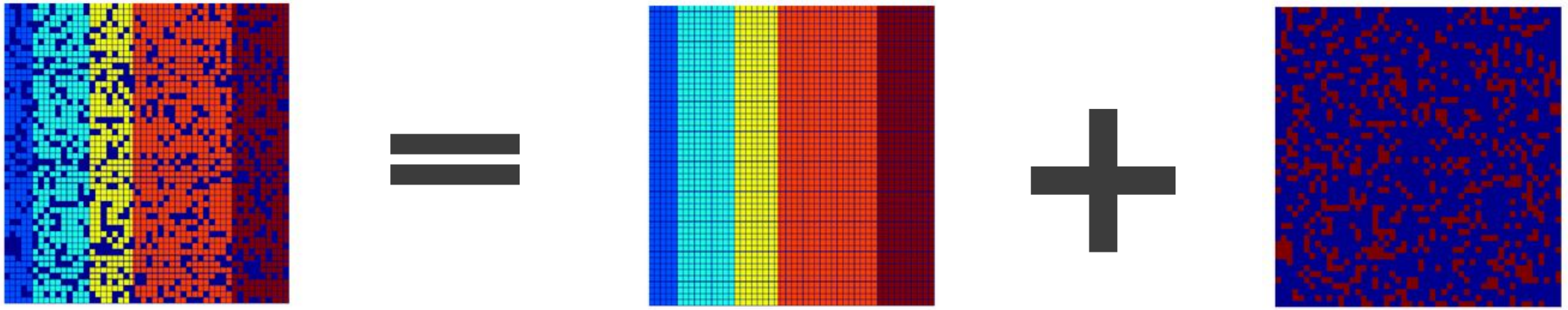
Integrated Low-Rank-Based Discriminative Feature Learning for Recognition

TNNLS 2016

2018.5.17

Problem Description

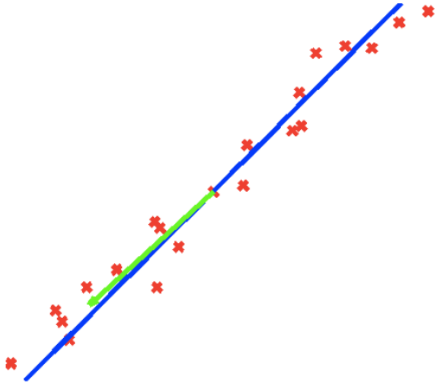
Rank-r matrix A of size m x n, where r << min(m,n). Model the observed matrix *D* to **be a set of linear measurements** on the matrix *A*, subject to **noise and gross corruptions** i.e., $D = L(A) + \eta$, where *L* is a **linear operator**, and η represents **the matrix of corruptions**. We seek to recover the true matrix *A* from *D*.

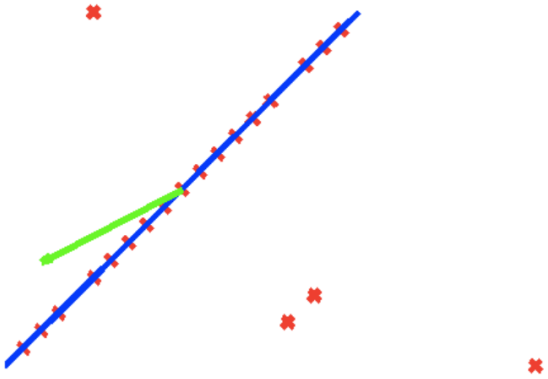


Matrix of corrupted observations = *Underlying low-rank matrix* + *Sparse error matrix*

PCA

L is the **identity operator** and the entries of η are **independent and identically distributed according to a isotropic Gaussian distribution**, then classical PCA provides the optimal estimate to *A*.





Matrix Completion

η is **zero**, L is the **matrix subsampling operator**, the problem is to use information from some entries of A to infer its missing entries.

$$\min_X \operatorname{rank}(X) \quad s.t. \;\; L(X) = D$$

nuclear norm: the sum of the singular values of the matrix

$$\min_X \|X\|_* \quad s.t. \;\; L(X) = D$$

$$\operatorname{rank}(A^H A) = \operatorname{rank}(AA^H) = \operatorname{rank}(A) \qquad A^H A = \sigma^2 u$$

Robust PCA

L is the identity operator and η is a sparse matrix, the problem is to **find the matrix of lowest rank that could have generated *D* when added to an unknown sparse matrix η**.

convex relaxation

$$\min_{X,E} \operatorname{rank}(X) + \gamma\|E\|_0 \quad s.t. \;\; D = X + E \longrightarrow \min_{X,E} \|X\|_* + \gamma\|E\|_1 \quad s.t. \;\; D = X + E$$

The **A**ugmented **L**agrange **M**ultiplier method for exact recovery of corrupted low-rank matrices

The general method of augmented Lagrange multipliers is introduced for solving constrained optimization problems of the kind:

$$\min f(X) \quad s.t. \; h(X) = 0$$

$$L(X, Y, \mu) = f(X) + \langle Y, h(X)\rangle + \frac{\mu}{2}\|h(X)\|_F^2$$

$$\langle Y, h(X)\rangle = tr(Y^T h(X))$$

Algorithm 3 (General Method of Augmented Lagrange Multiplier)

1: $\rho \geq 1$.
2: **while** not converged **do**
3: Solve $X_{k+1} = \arg\min_X L(X, Y_k, \mu_k)$.
4: $Y_{k+1} = Y_k + \mu_k h(X_{k+1})$;
5: Update μ_k to μ_{k+1}.
6: **end while**
Output: X_k.

$$\min f(X) \quad s.t.\ h(X) = 0$$

$$L(X, Y, \mu) = f(X) + \langle Y, h(X) \rangle + \frac{\mu}{2} \|h(X)\|_F^2$$

$$\min_{X,E} \|X\|_* + \gamma \|E\|_1 \quad s.t.\ \ D = X + E$$

$$L(X, E, Y, \mu) = \|X\|_* + \gamma \|E\|_1 + \langle Y, D - X - E \rangle - \frac{\mu}{2} \|D - X - E\|_F^2$$

Problem

Given a set of sufficiently dense data vectors $X = [x_1, x_2, \dots, x_n]$ (each column is a sample) drawn from a union of k subspaces $\{S_i\}_{i=1}^{k}$ of unknown dimensions, in a D-dimensional Euclidean space, segment all data vectors into their respective subspaces.

Assumption

The subspaces are low-rank and independent, and the data is noiseless. $\sum_{i=1}^{k} S_i = \bigoplus_{i=1}^{k} S_i$

A fraction of the data vectors are corrupted by noise or contaminated by outliers, or to be more precise, the data **contains sparse and properly bounded errors**.

Low-Rank Representation

Consider data vectors $X = [x_1, x_2, \dots, x_n]$, $x_i \in R^D$, each of which can be represented by the linear combination of the basis in a dictionary $\mathrm{A} = [a_1, a_2, \dots, a_m]$

$$X = AZ \qquad Z = [z_1, z_2, \dots, z_n]$$

[2009]Sparse representations using an appropriate dictionaries A may reveal the clustering of the points x_i. However, sparse representation **may not capture the global structures** of the data X. **Low rankness** may be a more appropriate criterion.

$$\min_Z \operatorname{rank}(Z) \quad \text{s.t. } X = AZ$$

a good surrogate

$$\min_Z ||Z||_* \quad \text{s.t. } X = AZ$$

The Basic Messages

$X = [x_1, x_2, \dots, x_n]$ $\{S_i\}_{i=1}^k$ $\{d_i\}_{i=1}^k$ $\{n_i\}_{i=1}^k$ $X = [X_1, X_2, \dots, X_k]$

In order to segment the data into their respective subspaces, we **need to compute an affinity matrix that encodes the pairwise affinities between data vectors**. So we use the data X itself as the dictionary.

$$\min_Z ||Z||_* \quad \text{s.t. } X = XZ$$

There always exist feasible solutions even when the data sampling is insufficient.

Theorem 3.1

Assume that the data sampling is sufficient such that $n_i > rank(X_i) = d_i$. If the subspaces are independent then there exists an optimal solution Z^*.

$$Z^* = \begin{bmatrix} Z_1^* & 0 & 0 & 0 \\ 0 & Z_2^* & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & Z_k^* \end{bmatrix}_{n \times n}$$

Theorem 3.1 **does not guarantee that an arbitrary optimal solution to the problem is block-diagonal**. The difficulty is essentially that the minimizer **is non-unique**. However, in our simulations we have observed that the solution obtained is always block-diagonal, and so we do not pursue this here.

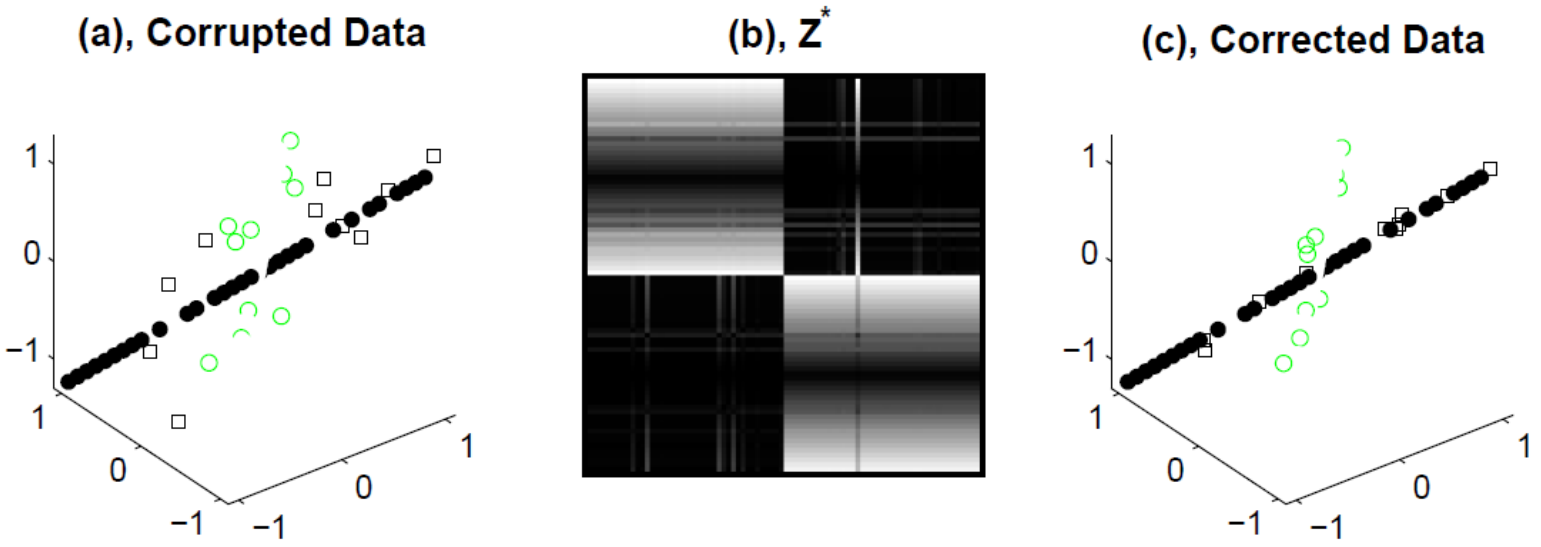
Robustness to Noise and Outliers

$$\min_{Z,E} ||Z||_* + \lambda \|E\|_{2,1} \quad \text{s.t. } X = XZ + E \qquad \|E\|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} (E_{ij})^2}$$

$$(Z^*, E^*) \xrightarrow{\text{recover data}} X - E^*/XZ^*$$

$$X = X_l + X_c$$

In the case that the remainder clean data is still sufficient to represent the subspaces, and the corruptions are properly bounded, it shall **automatically correct the corruptions** so as to obtain the lowest-rank representation.



There are about 80 data vectors sampled from two one-dimensional subspaces embedded in R^3, and about 25% data vectors are corrupted by large Gaussian errors.

sufficient

bounded

it shall automatically correct the corruptions so as to obtain the lowest-rank representation.

$$\min_{Z,E} ||Z||_* + \lambda\|E\|_{2,1} \quad \text{s.t. } X = XZ + E \iff \min_{Z,E,J} ||J||_* + \lambda\|E\|_{2,1} \quad \text{s.t. } X = XZ + E, Z = J$$

ALM

$$\min f(X) \quad s.t.\ h(X) = 0 \qquad L(X, Y, \mu) = f(X) + \langle Y, h(X)\rangle + \frac{\mu}{2}\|h(X)\|_F^2$$

$$\min_{Z,E,J,Y_1,Y_2} \big||J|\big|_* + \lambda\|E\|_{2,1} + tr[Y_1^T(X - XZ + E)] + tr[Y_2^T(Z - J)] + \frac{\mu}{2}(\|X - XZ + E\|_F^2 + \|Z - J\|_F^2)$$

1. fix the others and update J
2. fix the others and update Z
3. fix the others and update E
4. update the multipliers Y_1, Y_2
5. update the parameter μ
6. check the convergence conditions

$$\|X - XZ - E\|_\infty < \varepsilon \ \ and \ \|Z - J\|_\infty < \varepsilon$$

$$\min_{Z,E,Y} \big||Z|\big|_* + \lambda\|E\|_{2,1} + tr[Y^T(X - XZ + E)] + \frac{\mu}{2}(\|X - XZ + E\|_F^2)$$

$$\|X - XZ - E\|_\infty < \varepsilon \ \ and \ \|Z^{t+1} - Z^t\|_\infty < \varepsilon$$

稀疏表示

SR_1: $\min_{Z,E} ||Z||_1 + \lambda\|E\|_1 \quad \text{s.t. } X = XZ + E, Z_{ii} = 0$

$SR_{2,1}$: $\min_{Z,E} ||Z||_1 + \lambda\|E\|_{2,1} \quad \text{s.t. } X = XZ + E, Z_{ii} = 0$

$$\|x\|_1 = \sum_{i=1}^{N} |x_i| \qquad \|Z\|_1 = \max_j \sum_{i=1}^{m} |Z_{i,j}|$$

鲁棒PCA

$$\min_{X,E} \|X\|_* + \gamma\|E\|_1 \quad s.t. \ D = X + E$$

低秩稀疏表示LRR

$$\min_{Z,E} ||Z||_* + \lambda\|E\|_{2,1} \quad \text{s.t. } X = XZ + E$$

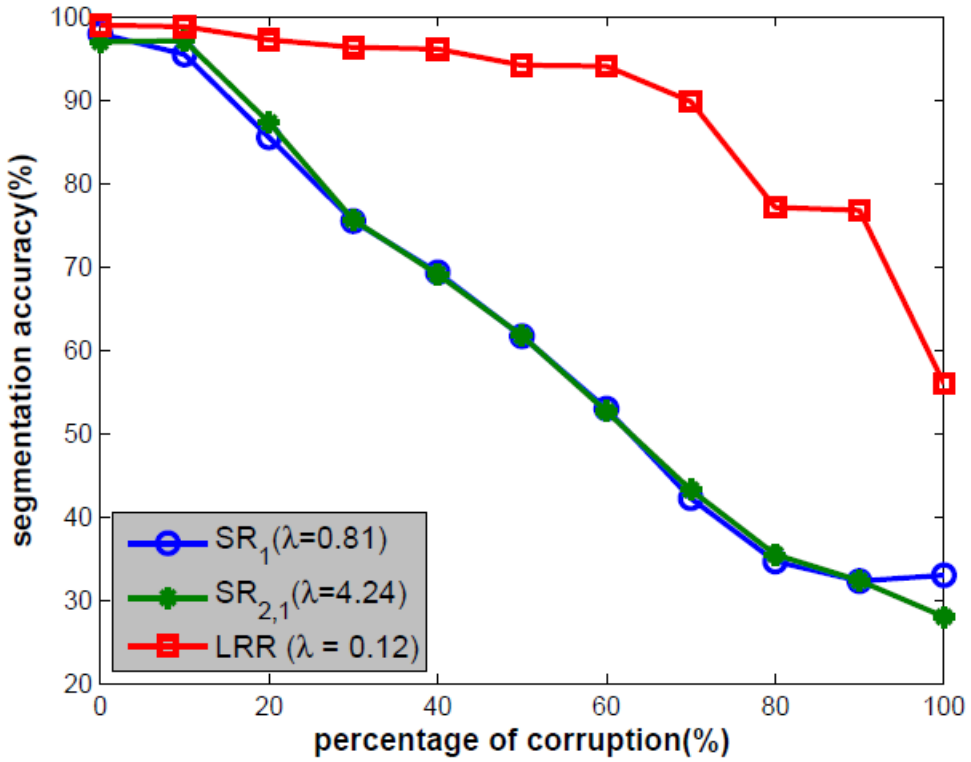
Algorithm 2 Subspace Segmentation by LRR

Input: data matrix X, number of subspaces k

1. obtain the lowest-rank representation by solving problem (5)

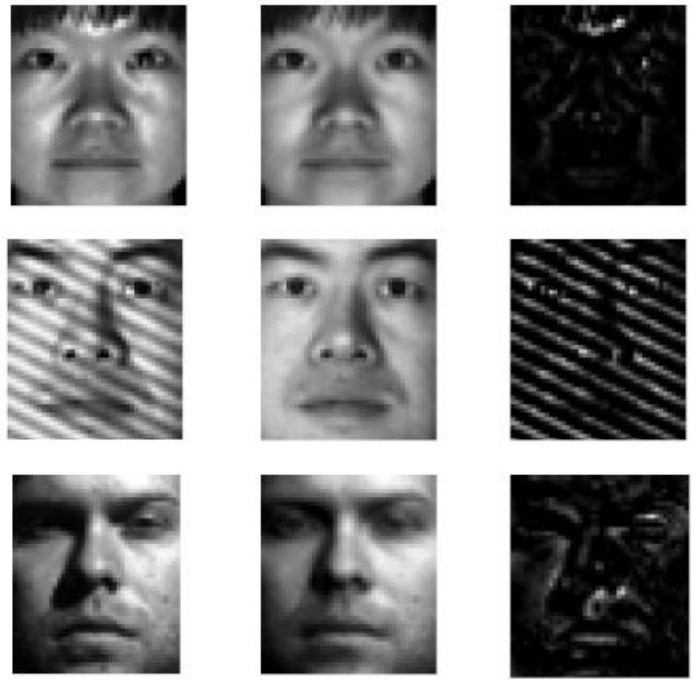
2. construct an undirected graph by using the lowest-rank representation to define the affinity matrix of the graph

3. use NCut to segment the vertices of the graph into k clusters



Some examples of using LRR to correct the corruptions in faces.

$$X = XZ^* + E^*$$



Latent Low-Rank Representation for Subspace Segmentation and Feature Extraction

ICCV 2011

$$\min_Z \|Z\|_* \quad s.t.\ X_O = X_O Z$$

$$Z_O^* = I$$

- cannot use X_o as the dictionary to represent the subspaces if the data sampling is insufficient.

- LRR requires that sufficient noiseless data is available in the dictionary A, i.e., only a part of A is corrupted. this assumption may be **invalid** and the robustness of LRR may be depressed in reality.

$$\min_Z \|Z\|_* \quad s.t.\ X_O = [X_O, X_H]Z$$

X_o is the observed data matrix and X_H represents the unobserved, hidden data

$$Z_{O,H}^* = \left[Z_{O|H}^*; Z_{H|O}^*\right]$$

Problem 1 (Noiseless Data)

$$\min_{Z} \|Z\|_* \quad s.t.\ X_O = [X_O, X_H]Z$$

Problem 2(Corrupted Data)

$$\min_{Z,E} \|Z\|_* + \lambda\|E\|_1 \quad s.t.\ X_O = [X_O, X_H]Z + E$$

Suppose $Z^*_{O,H} = \left[Z^*_{O|H}; Z^*_{H|O}\right]$ is the optimal solution (with respect to the variable Z) and $Z^*_{O|H}$ is the submatrix corresponding to X_o, then our goal is to **recover by using only the observed data** X_o.

$$[X_O, X_H] = U\Sigma V^T = U\Sigma[V_O; V_H]^T \qquad X_O = U\Sigma V_O^T \qquad X_H = U\Sigma V_H^T$$

$$U\Sigma V_O^T = U\Sigma V^T Z \qquad V_O^T = V^T Z$$

$$\min_{Z} \|Z\|_* \quad s.t.\ X_O = [X_O, X_H]Z \xrightarrow{\text{a unique minimizer [PAMI2013]}} Z^*_{O,H} = VV_O^T = [V_O V_O^T; V_H V_O^T]$$

$$X_O = [X_O, X_H]Z^*_{O,H} = X_O Z^*_{O|H} + X_H Z^*_{H|O} = X_O Z^*_{O|H} + X_H V_H V_O^T$$

$$= X_O Z^*_{O|H} + U\Sigma V_H^T V_H V_O^T = X_O Z^*_{O|H} + U\Sigma V_H^T V_H \Sigma^{-1} U^T X_O$$

$$L^*_{H|O} = U\Sigma V_H^T V_H \Sigma^{-1} U^T$$

$$X_O = X_O Z^*_{O|H} + L^*_{H|O} X_O$$

X_o and X_H are sampled from the same collection of low-rank subspaces

$$rank\left(Z^*_{O|H}\right) \le r \quad \text{and} \quad rank\left(L^*_{H|O}\right) \le r$$

$$\min_Z \|Z\|_* \quad s.t.\ X_o = [X_o, X_H]Z$$

$$\min_{Z,L} \|Z\|_* + \|L\|_* \quad s.t. X = XZ + LX$$

$$\min_{Z_{O|H}, L_{H|O}} rank\left(Z_{O|H}\right) + rank(L_{H|O}) \quad s.t.\ X_O = X_O Z_{O|H} + L_{H|O} X_O$$

$\min_Z \|Z\|_* \quad s.t. X = XZ$ Z_Z^* $\min_L \|L\|_* \quad s.t. X = LX$ L_L^*

[PAMI2013] $\|Z_Z^*\|_* = rank(X) = rank(X^T) = \|L_L^*\|_*$ So the strengths of L and Z are balanced naturally.

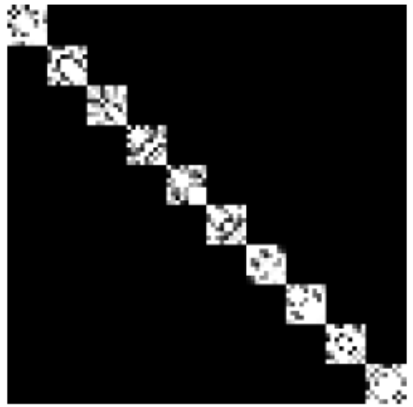
Setting

$\{S_i\}_{i=1}^{10}$ $\{U_i\}_{i=1}^{10}(U_{i+1} = TU_i)$, T is a random rotation and $U_1 \in R^{200\times 10}$

$X_O = [X_1, X_2, \dots, X_{10}]$ $X_i = U_i C_i, 1 \le i \le 10$ hidden matrix $X_H(200 \times 50)$

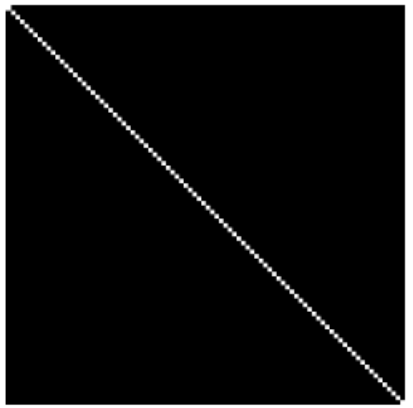
Example

$\min_Z \|Z\|_* \quad s.t.\ X_O = [X_O, X_H]Z$



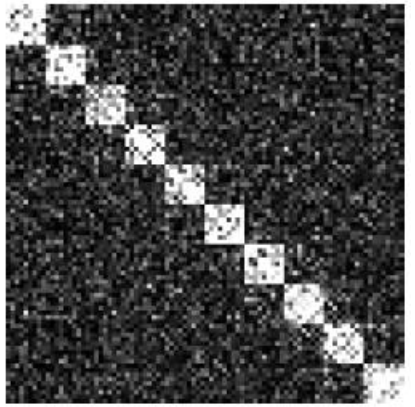
(a) $Z^*_{O|H}$

$\min_Z \|Z\|_* \quad s.t.\ X_O = X_O Z$



(b) LRR

$\min_{Z,L} \|Z\|_* + \|L\|_* \quad s.t. X = XZ + LX$

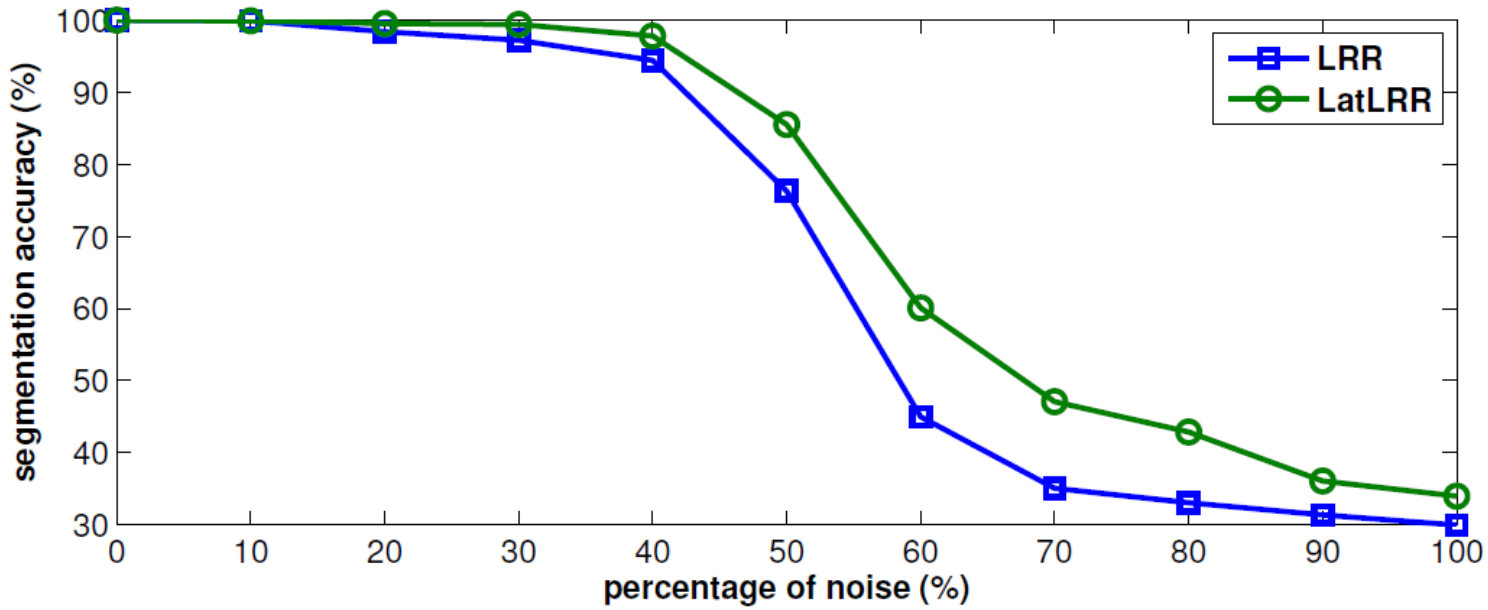


(c) LatLRR

Corrupted Data

$$\min_{Z,L,E} \|Z\|_* + \|L\|_* + \lambda\|E\|_1 \quad s.t. X = XZ + LX + E$$

$$\min_{Z,L,J,S,E} \|J\|_* + \|S\|_* + \lambda\|E\|_1 \quad s.t. \quad X = XZ + LX + E, Z = J, L = S$$



In summary, let (Z^*, L^*, E^*) be the minimizer, then Z^*, L^* are useful for **subspace segmentation** and **feature extraction**, respectively.

Subspace Segmentation

Utilize the **affinity matrix** identified by to define edge weights of an undirected graph, and then use Normalized Cuts (NCut) [23] to produce the final segmentation results.

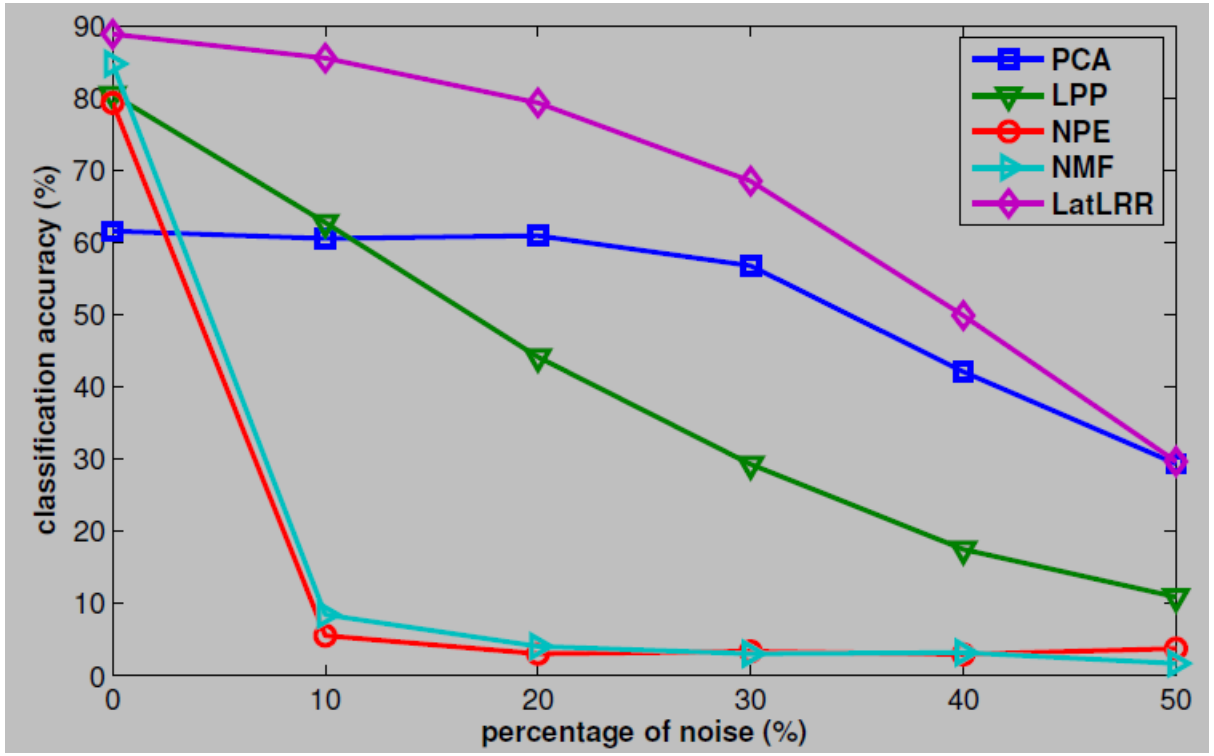
Comparison under the same setting					
	LSA	RANSAC	SR	LRR	LatLRR
Mean	8.99	8.22	3.89	3.16	**2.95**
Std.	9.80	10.26	7.70	5.99	**5.86**
Max	37.74	47.83	**32.57**	37.43	37.97

Comparison to state-of-the-art methods					
	LSA	ALC	SSC	SC	LatLRR
Mean	4.94	3.37	1.24	1.20	**0.85**

Feature Extraction

L^* may be useful for feature extraction, and experimentally find that L^* is able to extract "salient features" (i.e., notable features such as the eyes of faces) from data.

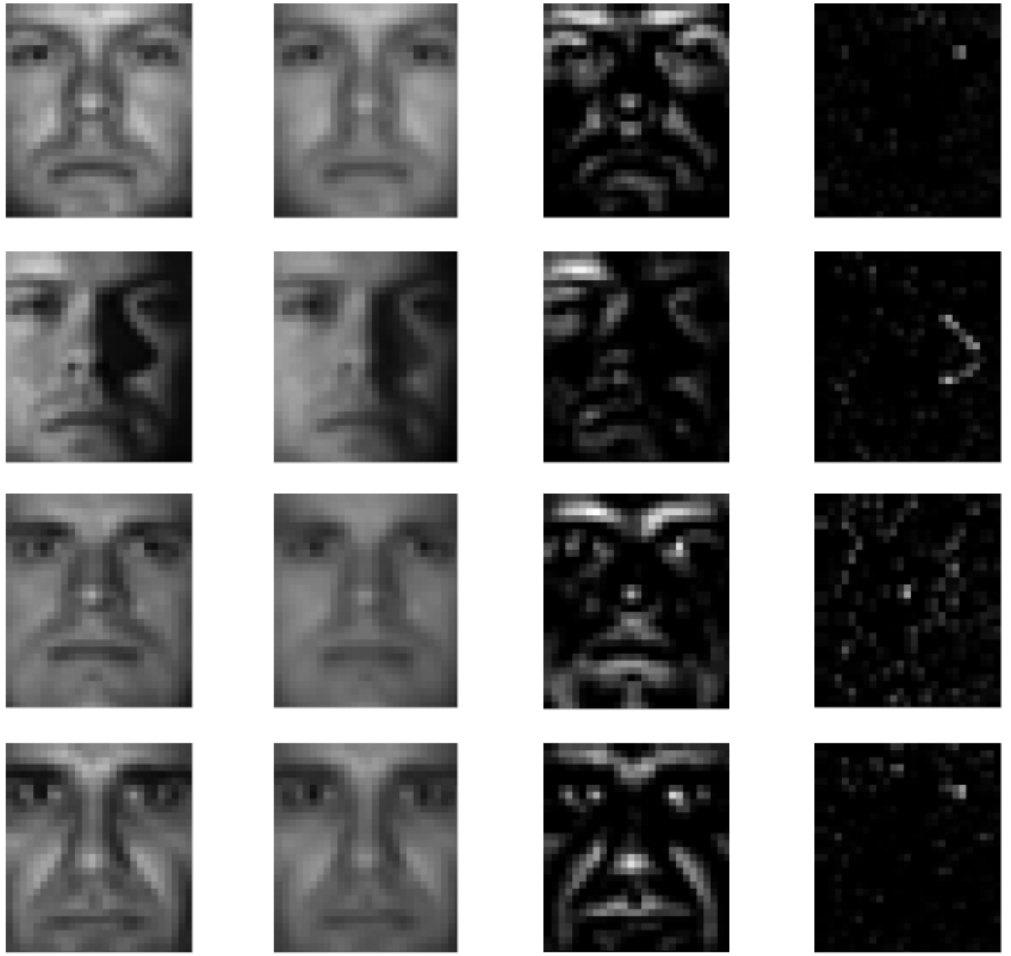
$$y = L^* x$$



The salient features correspond to the key object parts (e.g., the eyes), which are usually discriminative for recognition

$$X = XZ^* + L^*X + E^*$$

data = principal features + salient features + sparse noise



Suppose P is a low-dimensional projection learnt by using L^*x as inputs for some dimension reduction methods, the reduced feature vector y of a testing data vector x can be computed by $y = P^T L^* x$.

	Raw Data	PCA (317D)	LPP (83D)	NPE (325D)	NMF (195D)	LatLRR	LatLRR + PCA(400D)	LatLRR + LPP(52D)	LatLRR + NPE(400D)
1-NN	61.07	61.54	80.46	79.28	84.69	**88.76**	87.28	87.60	82.18
3-NN	59.81	60.03	78.73	79.28	84.07	**87.76**	85.95	87.13	81.71
5-NN	58.16	58.54	76.69	76.69	82.58	**86.03**	85.87	85.56	80.85

Conclusions

LatLRR extends LRR to **handle the hidden effects**

As a subspace **segmentation algorithm**, LatLRR outperforms the state-of-the-art algorithms for motion segmentation.

Be empirically able to automatically **extract salient features from corrupted data**.

Compared to dimension reduction based methods, LatLRR is **more robust to noise**.

Integrated Low-Rank-Based Discriminative Feature Learning for Recognition

TNNLS 2016

The basic idea is to **utilize the supervised information**, e.g., the labels of training samples, to learn discriminative features LX resulting from the LatLRR model.

Theorem

$$\min_{Z,L} \|Z\|_* + \|L\|_* \quad s.t. X = XZ + LX \xrightarrow{\text{complete solutions}} Z^* = V_X(I - S)V_X^T \text{ and } L^* = U_X S U_X^T$$

$SVD(X) = U_X \Sigma_X U_X^T$ S is any block-diagonal matrix that satisfies two constraints:

1) its blocks are compatible with Σ_X, i.e., if $(\Sigma_X)_{ii} \neq (\Sigma_X)_{jj}$, then $S_{ij} = 0$
2) both S and $I - S$ are positive semidefinite.

Note that S can usually be chosen as diagonal with diagonal entries being any number between 0 and 1.

$$\min_{L,W} \sum_{i=1}^{m} \varphi\left(h_i, f(Lx_i, W)\right) + \alpha||W||_F^2$$
$$s.t.\ \ L = U_X S U_X^T$$

$$U_X \in R^{d\times r} \quad S \in R^{r\times r}$$

$$f(x, W) = Wx, \qquad W \in R^{c\times d}$$

$$\min_{W,L} \|H - WLX\|_F^2 + \alpha||W||_F^2$$
$$s.t.\ \ L = U_X S U_X^T$$

$$H = [h_1, h_2, \dots, h_n] \in R^{c\times m}$$
$$h_i = [0,0,\dots,1,\dots,0,0]^T \in R^c$$

solve

1) the singular values of the data matrix X are usually distinct from each other, i.e. $(\Sigma_X)_{ii} \neq (\Sigma_X)_{jj}$, when $S_{ij} = 0$.

2) since only focus on learning the discriminative features, the constraint that $I - S$ is positive semidefinite is not necessary, so, only need to bound $S_{ii} > 0$.

$$\text{diag}(\Lambda) = (S_{11}, S_{22}, \dots, S_{rr}, 0,0, \dots, 0) \in R^d \qquad \Lambda \in R^{d\times d}$$

$$L = U_X S U_X^T = U\Lambda U^T \qquad\qquad UU^T = U^T U = VV^T = I$$

$$\|H - WLX\|_F^2 + \alpha||W||_F^2 = \|H - WU\Lambda U^T U\Sigma V^T\|_F^2 + \alpha||W||_F^2$$

$$= \|HV - WU\Lambda\Sigma\|_F^2 + \alpha||W||_F^2 = \|HV - WU\Lambda\Sigma\|_F^2 + \alpha||WU||_F^2$$

$$\widetilde{H} = HV \quad \widetilde{W} = WU$$

$$\|H - WLX\|_F^2 + \alpha||W||_F^2 = \left\|\widetilde{H} - \widetilde{W}\Lambda\Sigma\right\|_F^2 + \alpha||\widetilde{W}||_F^2$$

$$= \sum_{i=1}^{r}\left\|\widetilde{H}_i - S_{ii}\sigma_i\widetilde{W}_i\right\|_2^2 + \sum_{i=r+1}^{m}\left\|\widetilde{H}_i\right\|_2^2 + \alpha\sum_{i=1}^{r}||\widetilde{W}_i||_2^2 + \alpha\sum_{i=r+1}^{m}||\widetilde{W}_i||_2^2$$

$$\widetilde{W}_i = 0, i = r+1, \dots, m$$

$$\min_{\substack{S_{11},\dots,S_{rr}\\ \widetilde{W}_1,\dots,\widetilde{W}_r}} \sum_{i=1}^{r} \left(\left\| \widetilde{H}_i - S_{ii}\sigma_i \widetilde{W}_i \right\|_2^2 + \alpha ||\widetilde{W}_i||_2^2 \right) \qquad \text{s.t.}\, S_{ii} \geq 0, i = 1,2,\dots,r$$

$$\sum_{i=1}^{r} S_{ii} = t \qquad \sum_{i=1}^{r} S_{ii}\sigma_i = t \qquad g = [S_{11}\sigma_1, \dots, S_{rr}\sigma_r]^T \quad Q = \left[S_{11}\sigma_1 \widetilde{W}_1, \dots, S_{rr}\sigma_r \widetilde{W}_r\right]$$

$$\min_{g,Q} \sum_{i=1}^{r} \left(\left\| \widetilde{H}_i - Q_i \right\|_2^2 + \frac{\alpha}{g_i^2} ||Q_i||_2^2 \right) \qquad \text{s.t.} \sum_{i=1}^{r} g_i = t,\ g_i \geq 0, i = 1,2,\dots,r$$

fix g update of Q $$Q_i = \min_{Q_i} \left\| \widetilde{H}_i - Q_i \right\|_2^2 + \frac{\alpha}{g_i^2} ||Q_i||_2^2 = \frac{g_i^2}{g_i^2 + \alpha} \widetilde{H}_i \qquad i = 1,2,\dots,r$$

fix Q update of g $$\operatorname{argmin}_{g_i} \sum_{i=1}^{r} \frac{\alpha}{g_i^2} ||Q_i||_2^2 \qquad \text{s.t.} \sum_{i=1}^{r} g_i = t, g_i \geq 0, i = 1,2,\dots,r$$

$$L(g,\tau)=\sum_{i=1}^{r}\frac{\alpha}{g_i^2}||Q_i||_2^2+\tau\left(\sum_{i=1}^{r}g_i-t\right)$$

$$\frac{\partial L}{g_i}=-\frac{2\alpha\left|\left|Q_i\right|\right|_2^2}{g_i^3}+\tau=0 \qquad \sum_{i=1}^{r}g_i=\tau \qquad g_i=\frac{t||Q_i||_2^{2/3}}{\sum_{i=1}^{r}||Q_i||_2^{2/3}}$$

$$L=U\Lambda U^T \qquad \Lambda_{ii}=\frac{g_{ii}}{\sigma_i}\quad(i=1,\dots,r) \qquad \widetilde{W}_i=\frac{Q_i}{g_i} \qquad Z=LX$$

[*Neural Comput.* 2015] proposed denoise *X* first, apply the noiseless LRR or the LatLRR to the denoised data.

$$\min_{Z,L,E}\|Z\|_*+\|L\|_*+\lambda\|E\|_1 \quad s.t.\,X-E=(X-E)Z+L(X-E)$$

[*Neural Comput.* 2015] proved solving the above problem is equivalent to denoising X with the robust PCA first to obtain (A, E), and then solving noiseless LatLRR with X replaced by A.

Let the pair (A^*, E^*) be any optimal solution to the robust PCA problem. Then, the new noisy LatLRR model has minimizers (A^*, L^*, E^*) where $Z^* = V_A(I - S)V_A^T \quad L^* = U_A S U_A^T$

Robust PCA

$$\min_{A,E} \|X\|_* + \gamma\|E\|_1 \quad s.t. \;\; D = A + E$$

1) $X_{tr} \rightarrow A_{tr} \rightarrow U_{tr}\Sigma_{tr}V_{tr}^T \qquad A_{ts} \rightarrow U_{tr}U_{tr}^T X_{ts}$

2) $Z_{tr} \quad Z_{ts} \quad W$

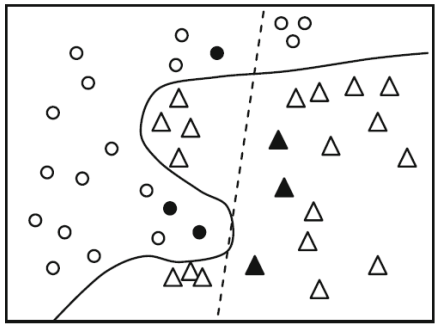
$$\min_{Z,L} \|Z\|_* + \|L\|_* \quad s.t. X = XZ + LX$$

?

$$\min_{L,W} \sum_{i=1}^{m} \varphi\big(h_i, f(Lx_i, W)\big) + \alpha||W||_F^2$$
$$s.t. \;\; L = U_X S U_X^T$$

Active Learning

Cold start ?



Limited budget

Active Learning via

Transductive Experimental Design

The goal of experimental design is to **find a set of experiments x_i that together are maximally informative**.

$$X: [x_1, \dots, x_m]^T \in \mathbb{R}^{m \times d}, |X| = m \quad \mathrm{V}: [v_1, \dots, v_n]^T \in \mathbb{R}^{n \times d}, \quad |V| = n$$

$$\min_{\mathbf{w}} \left\{ J(\mathbf{w}) = \sum_{i=1}^{m} \left(\mathbf{w}^\top \mathbf{x}_i - y_i \right)^2 + \mu \|\mathbf{w}\|^2 \right\} \qquad \mathbf{C}_{\mathbf{w}} = \left(\frac{\partial J(\mathbf{w})}{\partial \mathbf{w} \partial \mathbf{w}^\top} \right)^{-1} = (\mathbf{X}^\top \mathbf{X} + \mu \mathbf{I})^{-1}$$

$\mathbf{f} = \left[f\left(v_1, \dots, f(v_n)\right) \right]$ be the function values on all the available data **V,** the predictive error $\mathbf{f} - \hat{\mathbf{f}}$ has the covariance matrix $\sigma^2 C_{\mathrm{f}}$

$$\mathbf{C_f} = \mathbf{V}\mathbf{C_w}\mathbf{V}^\top = \mathbf{V}(\mathbf{X}^\top\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{V}^\top$$
$$= \frac{1}{\mu}\left[\mathbf{V}\mathbf{V}^\top - \mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right]$$

$$\max_{\mathbf{X}} \quad Tr\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right]$$
$$subject\ to \quad \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m$$

$$\max_{X_o} Tr[V_o X_o^T (X_o X_o^T + \mu I)^{-1}] X_o V_o^T \longrightarrow X_o = (LX)^T \quad V_o = (LV)^T \quad L = USU^T$$

$$V \in R^{d\times n} \quad X \in R^{d\times m}$$

$$\max_{X,L} Tr[V^T L^T LX(X^T L^T LX + \mu I)^{-1}] X^T L^T LV \quad s.t.\ L = USU^T$$

$$= \max_{X,S} Tr[V^T US^2 U^T X(X^T US^2 U^T X + \mu I)^{-1}] X^T US^2 U^T V$$

$$X = VP \quad s.t. X \in R^{m\times n} \quad \sum_i P_{ij} = 1 \quad \sum_j P_{ij} \le 1 \quad \sum_{i,j} P_{ij} = m$$

$$\max_{P,S} Tr[V^T US^2 U^T VP(P^T V^T US^2 U^T VP + \mu I)^{-1}] P^T V^T US^2 U^T V \qquad s.t.\ P, S$$

$$\min_{\mathbf{X},\mathbf{A}} \quad \sum_{i=1}^{n} \|\mathbf{v}_i - \mathbf{X}^\top \mathbf{a}_i\|^2 + \mu\|\mathbf{a}_i\|^2 \qquad \textit{subject to} \quad \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m \qquad \mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times m}$$

$$L(X, A) = \|V - AX\|_F^2 + \mu \mathrm{Tr}(AA^T) = \mathrm{Tr}[(V - AX)(V - AX)^T] + \mu \mathrm{Tr}(AA^T)$$

$$= \mathrm{Tr}[VV^T - AXV^T - VX^TA^T + AXX^TA^T + \mu AA^T]$$

$$= \mathrm{Tr}[VV^T] - Tr[AXV^T + VX^TA^T - A(XX^T + \mu I)A^T]$$

$$\frac{\partial L(X, A)}{A} = 0 \longrightarrow A^* = VX^T(XX^T + \mu I)^{-1}$$

$$\min \|\mathbf{V} - \mathbf{AX}\|_F^2 + \mu \mathrm{Tr}(\mathbf{AA}^\top) = \mathrm{Tr}(\mathbf{VV}^\top) - \mathrm{Tr}\left[\mathbf{VX}^\top(\mathbf{XX}^\top + \mu\mathbf{I})^{-1}\mathbf{XV}^\top\right]$$

$$\updownarrow$$

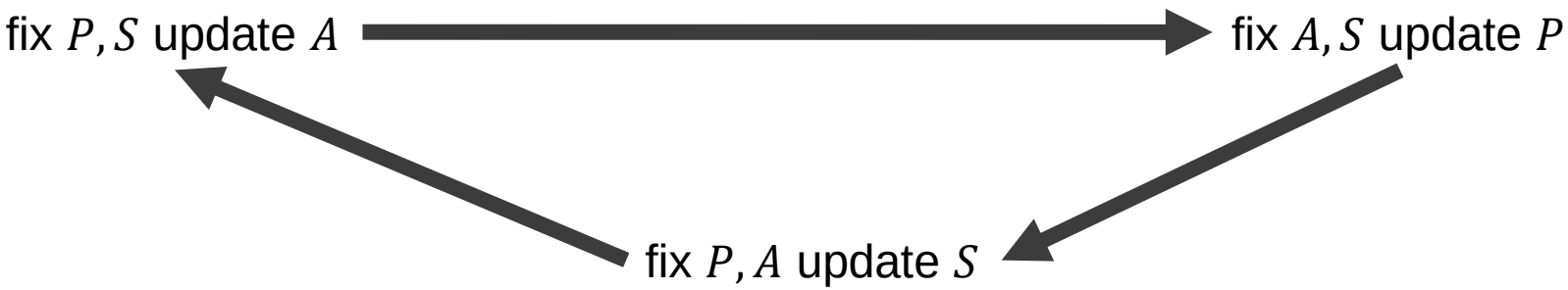
$$\max_{\mathbf{X}} \quad Tr\left[\mathbf{VX}^\top(\mathbf{XX}^\top + \mu\mathbf{I})^{-1}\mathbf{XV}^\top\right]$$
$$\textit{subject to} \quad \mathbf{X} \subset \mathbf{V}, |\mathbf{X}| = m$$

$$\min\|\mathbf{V}-\mathbf{A}\mathbf{X}\|_F^2+\mu\mathrm{Tr}(\mathbf{A}\mathbf{A}^\top)=\mathrm{Tr}(\mathbf{V}\mathbf{V}^\top)-\mathrm{Tr}\left[\mathbf{V}\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top+\mu\mathbf{I})^{-1}\mathbf{X}\mathbf{V}^\top\right]$$

$$\min_{A,P}\|V_o-AV_oP\|_F^2+\mu\|A\|_F^2=\min_{A,P}\|V^TL^T-AV^TL^TP\|_F^2+\mu\|A\|_F^2\,,s.t.\,L=USU^T\;\;V=U\Sigma V_v^T$$

$$=\min_{A,P,S}\|V_v\Sigma SU^T-AV\Sigma SU^TP\|_F^2+\mu\|A\|_F^2$$

$$L=USU^T=U\Lambda U^T\;\;\mathrm{diag}(\Lambda)=(S_{11},S_{22},\dots,S_{rr},0,0,\dots,0)\in R^d\qquad\Lambda\in R^{d\times d}$$



Early Active Learning via Robust Representation and Structured Sparsity

IJCAI 2013

The key idea of TED is to select the samples that can best represent the whole data using a linear representation.

$$\min_{A,B} \sum_{i=1}^{n} (\|x_i - Ba_i\|_2^2 + \gamma \|a_i\|_2^2) \quad s.t. A = [a_1, \dots, a_n] \in R^{m \times n}, B \subset X, |B| = m$$

The TED objective minimizes the **least square error**, hence it is **sensitive to the data outliers**. To solve these two deficiencies, we formulate the early active learning problem **using the structured sparsity-inducing norms** and propose a new robust formulation with efficient optimization algorithm.

$$\min_{A,B} \sum_{i=1}^{n} (\|x_i - Xa_i\|_2^2 + \gamma \|a_i\|_2^2) \quad s.t. A = [a_1, \dots, a_n] \in R^{n \times n} \ \|A\|_{2.0}$$

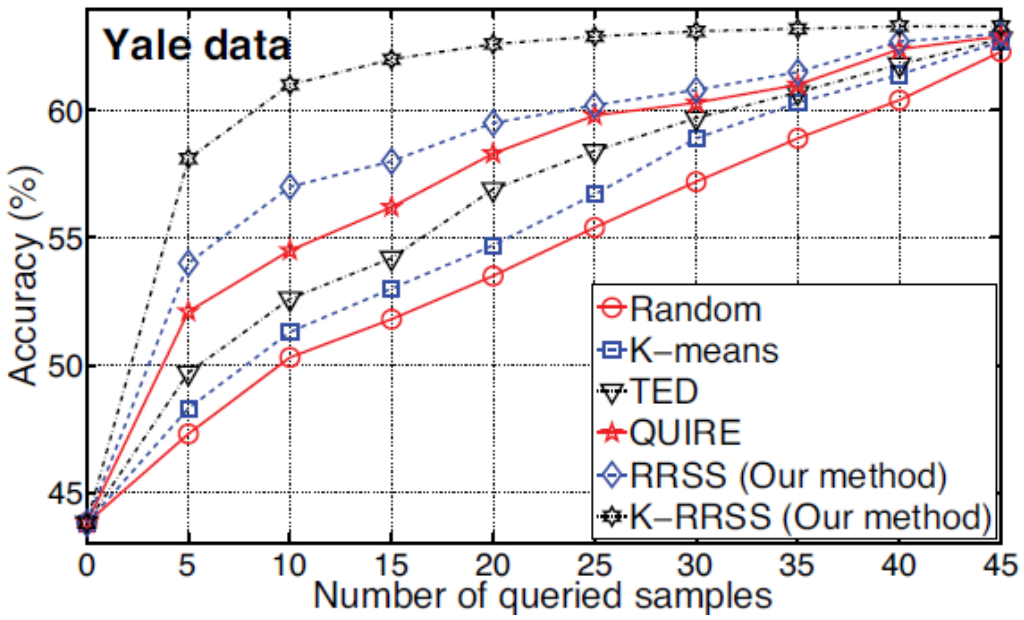
$$\min_{A} \sum_{i=1}^{n} \left(\|x_i - Xa_i\|_2^2 + \gamma \|A\|_{2,0}\right)$$

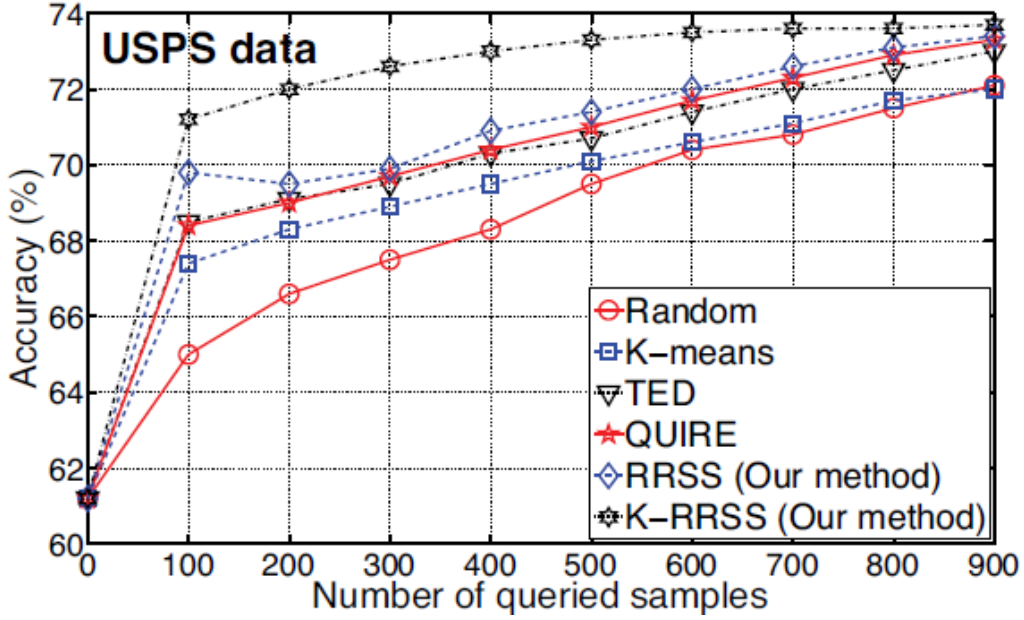
$\|A\|_{2,1}$ is the minimum convex hull of $\|A\|_{2,0}$, and when A is row-sparse enough, one can always minimize $\|A\|_{2,1}$ to obtain the same result of minimizing $\|A\|_{2,0}$.

$$\min_A \sum_{i=1}^{n} \left(\|x_i - Xa_i\|_2^2 + \gamma \|A\|_{2,1} \right) = \min_A \|(X - XA)^T\|_2^2 + \gamma \|A\|_{2,1}$$

$$J = \min_A \ \|(X - XA)^T\|_{2,1} + \gamma \|A\|_{2,1}$$

L1-norm is imposed among data points and the L2-norm is used for features





IJCAI 2013

$$\min_A \sum_{i=1}^{n} \left(\|x_i - Xa_i\|_2^2 + \gamma \|A\|_{2,1} \right) = \min_A \|(X - XA)^T\|_2^2 + \gamma \|A\|_{2,1}$$

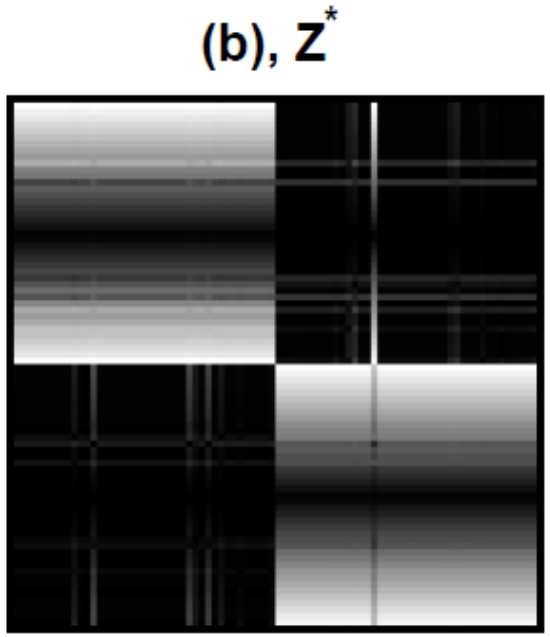
$$J = \min_A \ \|(X - XA)^T\|_{2,1} + \gamma \|A\|_{2,1}$$

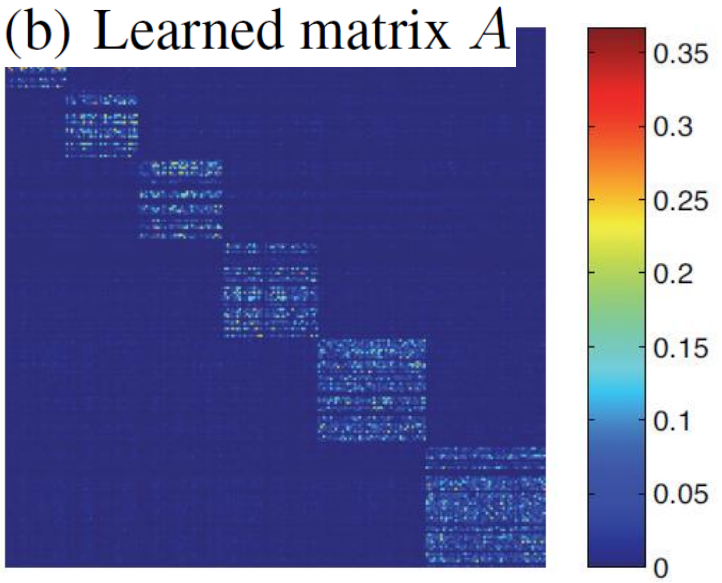
ICML 2010

$$\min_Z ||Z||_* \quad \text{s.t. } X = XZ$$

$$\min_{Z,E} ||Z||_* + \lambda \|E\|_{2,1} \quad \text{s.t. } X = XZ + E$$

$$\min_Z \lambda \|X - XZ\|_{2,1} + ||Z||_*$$





SIGIR 08

$$\min_{\boldsymbol{\beta}, \boldsymbol{\alpha}_i \in \mathbb{R}^N} \ \sum_{i=1}^{M} \|\mathbf{x}_i - \mathbf{X}_C^\top \boldsymbol{\alpha}_i\|^2 + \sum_{j=1}^{N} \frac{\alpha_{i,j}^2}{\beta_j} + \gamma \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad \mathbf{x}_i \in \mathbf{X}_{\mathcal{P}}, \ \ \beta_j \geq 0, \ \ j = 1, \cdots, N$$

$$J = \min_{A} \ \|(X_o - X_o A)^T\|_{2,1} + \gamma\|A\|_{2,1} \qquad X_o \in R^{d\times n} \quad A \in R^{n\times n}$$

$$X_o = LX \qquad L = USU^T = U\Lambda U^T$$

$$\min_{A} \ \|(LX - LXA)^T\|_{2,1} + \gamma\|A\|_{2,1} = \min_{A,S} \ \|(US\Sigma V^T - US\Sigma V^T A)^T\|_{2,1} + \gamma\|A\|_{2,1}$$

$$\min_{A,P,S} \|V_v \Sigma S U^T - AV\Sigma S U^T P\|_F^2 + \mu\|A\|_F^2$$

1) $(\Sigma_X)_{ii} \neq (\Sigma_X)_{jj}$, when $S_{ij} = 0$. 2) S and $I - S$ are positive semidefinite

since only focus on learning the discriminative features, the constraint that $I - S$ is positive semidefinite is not necessary, so, only need to bound $S_{ii} > 0$.

Kernel Extension

$$J = \min_{A} \ \|(\phi(X_o) - \phi(X_o)A)^T\|_{2,1} + \gamma\|A\|_{2,1}$$

$$= \min_{A} \ \|(\phi(US\Sigma V^T) - \phi(US\Sigma V^T)A)^T\|_{2,1} + \gamma\|A\|_{2,1}$$