

Joint Sparse PCA

(From *Joint sparse principal component analysis* by Shuangyan Yi et. al.)

Ashutosh Maurya (MDS202110)

Athul Prakash (MDS202111)

Avik Das (MDS202112)

Ayush Srivastava (MDS202113)

Chennai Mathematical Institute
Linear Algebra and Its Applications

November 1, 2022

Table of Contents

- 1 Agenda
- 2 Background
 - PCA
 - Sparse PCA
- 3 Motivation for JSPCA
- 4 Methodology
- 5 Algorithm
- 6 Conclusion

Agenda

Background

- A data-processing and dimension-reduction technique.
- Invented in 1901 by Karl Pearson.
- PCA seeks the linear combinations of the input variables in order to derive new uncorrelated variables (or Principal Components) that capture maximal variance. PCA can be computed via the singular value decomposition (SVD) of the data matrix.
- PCA is an unsupervised method.
- PCA's covariance matrix is derived from l_2 -norm

Limitations of PCA

- A particular disadvantage of ordinary PCA is that the PCs are usually linear combinations of all input variables and thus it is often difficult to make sense of what a PC represents.
- PCA is sensitive to outliers.
- Each new feature in a low dimensional subspace is the linear combination of all the original features in high-dimensional space.

Sparse PCA

- An extension of the PCA model
- Aims to reduce the loadings of the PCs by introducing sparsity structures to the input variables.
- Done by writing the problem of PCA as a regression-type optimisation problem with a quadratic penalty, following which the Lasso penalty can be directly integrated into the regression criterion which will lead to the modified PCs with sparse loadings.
- Particular disadvantage of ordinary PCA is difficulty in interpretation of the PCs. Sparse PCA overcomes this disadvantage by finding linear combinations that contain just a few input variables.

- $l_{2,1}$ -norm on both loss term and regularization term; discard the useless features on one hand and reduce the effect of outliers on the other hand.
- JSPCA relaxes the orthogonal constraint and of transformation matrix and introduces another transformation matrix.
- JSPCA uses joint sparse constraints on the objective function.

Motivation for JSPCA

Motivation of JSPCA

- Limitations of other variants of PCA

- l_2 norm cannot zero out coefficients: cannot exclude redundant features.

- Usage of $l_{2,1}$ norm Recall:
$$\|\mathbf{M}\|_{2,1} = \sum_{i=1}^n \sqrt{\sum_{j=1}^m M_{ij}^2}$$

- Combining the pros of l_1 and l_2 norms.
- In loss term, it is able to jointly select features and exclude redundant ones.
- In regularization term, it is more robust to outliers.

Motivation of JSPCA (cont.)

- Limitations of PCA and SPCA

- SPCA does not make zero loadings across all columns, and hence features cannot be ignored. But JSPCA jointly excludes useless features: we get row-sparsity.

Due to adding $l_{2,1}$ norm on loss term.

- Outliers are very common, and PCA doesn't handle it well.

Due to adding $l_{2,1}$ norm on the regularization term.

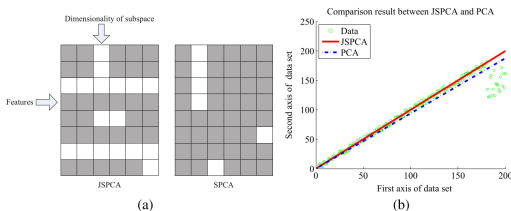


Figure: (a) JSPCA tells us useless features while SPCA cannot and
(b) JSPCA is more robust to outliers

Methodology

Objective Function

(Recall: Given data matrix is $X = [x_1, x_2, \dots, x_n] \in R^{m \times n}$, where m denotes the features and n denotes the number of training samples.)

$$\arg \min_{Q, P} J(Q, P) = \arg \min_{Q, P} \|X - PQ^T X\|_{2,1} + \lambda \|Q\|_{2,1} \quad (1)$$

where transformation matrix $Q \in R^{m \times d}$ is first used to project the data matrix X onto a low-dimensional subspace and another transformation matrix $P \in R^{m \times d}$ is then used to recover the data matrix X . $\lambda \geq 0$ is the regularization parameter.

We transform the objective function in terms of Frobenius norm.

Objective Function (cont.)

$$\arg \min_{Q,P} \|X - PQ^T X\|_{2,1} + \lambda \|Q\|_{2,1}$$

$$= \arg \min_{Q,P} 2\text{tr}((X - PQ^T X)^T D_1 (X - PQ^T X)) + 2\lambda \text{tr}(Q^T D_2 Q)$$

where

$$D_1 = \begin{bmatrix} \frac{1}{\sqrt{\sum_{i=1}^m (X - PQ^T X)_{1i}^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\sum_{i=1}^m (X - PQ^T X)_{mi}^2}} \end{bmatrix}$$

and

$$D_2 = \begin{bmatrix} \frac{1}{\sqrt{\sum_{i=1}^m Q_{1i}^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\sum_{i=1}^m Q_{mi}^2}} \end{bmatrix}$$

Objective Function (cont.)

We arrive at the LHS from the RHS for a general matrix $A = a_{ij}$.

$$\begin{aligned}
 & \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sqrt{\sum_{i=1}^m a_{1i}^2}} & & & \\ & \ddots & & \\ & & \frac{1}{\sqrt{\sum_{i=1}^m a_{mi}^2}} & \\ & & & \ddots \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \\
 = & \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} \frac{a_{11}}{\sqrt{\sum_{i=1}^m a_{1i}^2}} & \cdots & \frac{a_{1n}}{\sqrt{\sum_{i=1}^m a_{1i}^2}} \\ \vdots & \ddots & \vdots \\ \frac{a_{m1}}{\sqrt{\sum_{i=1}^m a_{mi}^2}} & \cdots & \frac{a_{mn}}{\sqrt{\sum_{i=1}^m a_{mi}^2}} \end{bmatrix} \\
 = & \begin{bmatrix} \frac{a_{11}^2}{\sqrt{\sum_{i=1}^m a_{1i}^2}} + \frac{a_{21}^2}{\sqrt{\sum_{i=1}^m a_{2i}^2}} + \cdots + \frac{a_{m1}^2}{\sqrt{\sum_{i=1}^m a_{mi}^2}} & & & \\ & \ddots & & \\ & & \frac{a_{1n}^2}{\sqrt{\sum_{i=1}^m a_{1i}^2}} + \frac{a_{2n}^2}{\sqrt{\sum_{i=1}^m a_{2i}^2}} + \cdots + \frac{a_{mn}^2}{\sqrt{\sum_{i=1}^m a_{mi}^2}} & \end{bmatrix}
 \end{aligned}$$

Objective Function (cont.)

Applying trace, we get

$$\begin{aligned} & \sqrt{\sum_{i=1}^n a_{1i}^2} + \sqrt{\sum_{i=1}^n a_{2i}^2} + \cdots + \sqrt{\sum_{i=1}^n a_{mi}^2} \\ &= \sum_{j=1}^m \sqrt{\sum_{i=1}^n a_{ji}^2} = \|A\|_{2,1} \end{aligned}$$

Objective Function (cont.)

So, the objective function becomes

$$\begin{aligned} & \arg \min_{Q,P} \|X - PQ^T X\|_{2,1} + \lambda \|Q\|_{2,1} \\ &= \arg \min_{Q,P} 2\text{tr}((X - PQ^T)^T D_1 (X - PQ^T X)) + 2\lambda \text{tr}(Q^T D_2 Q) \\ &= \arg \min_{Q,P} 2\text{tr}((X - PQ^T)^T \sqrt{D_1}^T \sqrt{D_1} (X - PQ^T X)) \\ &\quad + 2\lambda \text{tr}(Q^T \sqrt{D_2}^T \sqrt{D_2} Q) \\ &= \arg \min_{Q,P} 2\text{tr}(\sqrt{D_1}^T (X - PQ^T)^T \sqrt{D_1} (X - PQ^T X)) \\ &\quad + 2\lambda \text{tr}((\sqrt{D_2} Q)^T \sqrt{D_2} Q) \\ &= \arg \min_{Q,P} \|\sqrt{D_1}^T (X - PQ^T)\|_F^2 + 2\lambda \|\sqrt{D_2} Q\|_F^2 \end{aligned}$$

Objective Function (cont.)

Hence, eq. (1) becomes

$$\arg \min_{Q,P} J(Q,P) = \arg \min_{Q,P} \|\sqrt{D_1}^T (X - PQ^T X)\|_F^2 + 2\lambda \|\sqrt{D_2} Q\|_F^2 \quad (2)$$

where

$$D_1 = \begin{bmatrix} \frac{1}{\sqrt{\sum_{i=1}^m (X - PQ^T X)_{1i}^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\sum_{i=1}^m (X - PQ^T X)_{mi}^2}} \end{bmatrix} \quad (3)$$

and

$$D_2 = \begin{bmatrix} \frac{1}{\sqrt{\sum_{i=1}^m Q_{1i}^2}} & & \\ & \ddots & \\ & & \frac{1}{\sqrt{\sum_{i=1}^m Q_{mi}^2}} \end{bmatrix} \quad (4)$$

cm

Objective Function (cont.)

- The smaller D_2^{ii} is, the more important the i -th feature is.
- If $\|(X - PQ^T X)^i\|_2$ and $\|Q^i\|_2$ are small, then D_1 and D_2 are large.
So, the minimization of $(X - PQ^T)^T D_1 (X - PQ^T X) + 2\lambda \text{tr}(Q^T D_2 Q)$ forces $\|(X - PQ^T X)^i\|_2$ and $\|Q^i\|_2$ to have very small values.
- Thus, we obtain a joint sparse Q and a small reconstruction matrix P .

Next, let $\sqrt{D_1}P = \bar{P}$ and $\sqrt{D_1}^{-1}Q = \bar{Q}$. Then eq. (2) becomes

$$\arg \min_{Q, P} J(Q, P) = \arg \min_{Q, P} \|\sqrt{D_1}^T X - \bar{P} \bar{Q}^T \sqrt{D_1} X\|_F^2 + 2\lambda \|\sqrt{D_1} \sqrt{D_2} \bar{Q}\|_F^2 \quad (5)$$

Objective Function (cont.)

In order to reduce the feature redundancy, the orthogonality constraint $\overline{P}^T \overline{P} = I^{d \times d}$ is imposed.

Therefore, we have our final objective function

$$\begin{aligned} \arg \min_{Q, P} J(Q, P) = \arg \min_{Q, P} & \|\sqrt{D_1}^T X - \overline{P} \overline{Q}^T \sqrt{D_1} X\|_F^2 \\ & + 2\lambda \|\sqrt{D_1} \sqrt{D_2} \overline{Q}\|_F^2 \\ \text{s.t. } & \overline{P}^T \overline{P} = I^{d \times d} \end{aligned} \quad (6)$$

where $\overline{Q} \in R^{m \times d}$ is first used to project the weighted data matrix $\sqrt{D_1} X$ and $\overline{P} \in R^{m \times d}$ is then used to recover it.

Optimal Solution

Step 1:

Given \overline{P} , there exists an optimal matrix \overline{P}_\perp such that $[\overline{P}, \overline{P}_\perp]$ is $m \times m$ column orthogonal matrix.

Then, optimization problem becomes

$$\arg \min_Q \|\sqrt{D_1}^T X - \overline{P} Q^T \sqrt{D_1} X\|_F^2 + \lambda \|\sqrt{D_1} \sqrt{D_2} Q\|_F^2 \quad (7)$$

Optimal Solution (cont.)

The first part of eq. (7) becomes

$$\begin{aligned} & \|\sqrt{D_1}^T X - \overline{PQ}^T \sqrt{D_1} X\|_F^2 \\ &= \|X^T \sqrt{D_1} - X^T \sqrt{D_1} \overline{QP}^T\|_F^2 \\ &= \|X^T \sqrt{D_1} [\overline{PP}_\perp] - X^T \sqrt{D_1} \overline{QP}^T [\overline{PP}_\perp]\|_F^2 \\ &= \|X^T \sqrt{D_1} \overline{P} - X^T \sqrt{D_1} \overline{QP}^T \overline{P}\|_F^2 + \|X^T \sqrt{D_1} \overline{P}_\perp - X^T \sqrt{D_1} \overline{QP}^T \overline{P}_\perp\|_F^2 \\ &= \|X^T \sqrt{D_1} \overline{P} - X^T \sqrt{D_1} \overline{Q}\|_F^2 + \|X^T \sqrt{D_1} \overline{P}_\perp\|_F^2 \end{aligned}$$

Optimal Solution (cont.)

Since \bar{P} is fixed and $\|X^T \sqrt{D_1} \bar{P}_\perp\|_F^2$ is constant, the optimization problem becomes

$$\arg \min_Q \|X^T \sqrt{D_1} \bar{P} - X^T \sqrt{D_1} \bar{Q}\|_F^2 + \lambda \|\sqrt{D_1} \sqrt{D_2} \bar{Q}\|_F^2 \quad (8)$$

By the derivatives of eq. previous () w.r.t \bar{Q} to be 0, we get,

$$\bar{Q} = (\lambda \sqrt{D_1} D_2 \sqrt{D_1} + \sqrt{D_1} X X^T \sqrt{D_1})^{-1} \sqrt{D_1} X X^T \sqrt{D_1} \bar{P} \quad (9)$$

Therefore,

$$Q = (\lambda D_2 + X X^T)^{-1} X X^T \sqrt{D_1} \bar{P} \quad (10)$$

Optimal Solution (cont.)

Step 2:

Given \overline{Q} to compute \overline{P} , optimization problem becomes

$$\arg \min_{\overline{P}} \|\sqrt{D_1}X - \overline{P}\overline{Q}^T \sqrt{D_1}X\|_F^2 \text{ s.t. } \overline{P}^T \overline{P} = I^{d \times d} \quad (11)$$

The update of \overline{P} of minimizing of eq. (11) with the given constraint means that \overline{P} is orthogonal in the columns. In order to compute \overline{P} , we introduce a lemma.

Optimal Solution: Lemma 1

Lemma 1: Let $Z^{n \times m}$ and $V^{n \times d}$ be two matrices. Consider the constrained minimization problem,

$$\arg \min_{\bar{P}} \|Z - VP^T\|^2 \text{ s.t. } P^T P = I^{d \times d} \quad (12)$$

Suppose the SVD of $Z^T V$ is EDU^T , then the optimal solution is $P = EU^T$.

Proof:

We have to minimize $\|Z - VP^T\|_F^2$ for an optimal P such that $P^T P = I$.

Now,

$$\|Z - VP^T\|_F^2 = \text{tr}((Z - VP^T)^T (Z - VP^T)) \quad (13)$$

But,

$$\begin{aligned} (Z - VP^T)^T (Z - VP^T) &= (Z^T - PV^T)(Z - VP^T) \\ &= Z^T Z - Z^T VP^T - PV^T Z + PV^T VP^T \end{aligned}$$

cm_i

Optimal Solution: Lemma 1 (cont.)

Putting this in eq. (13), we get

$$\begin{aligned}\|Z - VP^T\|_F^2 &= \text{tr}(Z^T Z - Z^T VP^T - PV^T Z + PV^T VP^T) \\ &= \text{tr}(Z^T Z) - \text{tr}(Z^T VP^T) - \text{tr}(PV^T Z) + \text{tr}((PV^T VP^T)) \\ &= \|Z\|_F^2 - \text{tr}(EDU^T P^T) - \text{tr}(PUDE^T) + \text{tr}(V^T V) \\ (\because PV^T VP^T &\sim V^T V \implies \text{tr}(PV^T VP^T) = \text{tr}(V^T V)) \\ &= \|Z\|_F^2 - \text{tr}(DU^T P^T E) - \text{tr}(DE^T PU) + \|V\|_F^2\end{aligned}$$

Here, $\|Z\|_F$ and $\|V\|_F$ are constant. So, we have to minimize the remaining terms.

Optimal Solution: Lemma 1 (cont.)

Now, U , P and E are unitary matrices. We can claim

$$\text{tr}(AB) \leq \text{tr}(A) \quad (14)$$

where A is a diagonal matrix with non-negative entries and B is unitary, since for arbitrary diagonal matrices D and U , we have

$$\begin{bmatrix} d_{11} & & & \\ & d_{22} & & \\ & & \ddots & \\ & & & d_{nn} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1n} \\ u_{21} & u_{22} & \dots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n1} & u_{n2} & \dots & u_{nn} \end{bmatrix} = \begin{bmatrix} d_{11}u_{11} & & & \\ & d_{22}u_{22} & & \\ & & \ddots & \\ & & & d_{nn}u_{nn} \end{bmatrix}$$

Since $|u_{ij}| \leq 1 \quad \forall 1 \leq i, j \leq n$, hence the claim is true.

So, if we choose $P = EU^T$, then $\text{tr}(DU^T P^T E) = \text{tr}(D)$ and $\text{tr}(DE^T P U) = \text{tr}(D)$, and we will get the optimal solution.

Optimal Solution

We have $ZT^V = \sqrt{D_1}XX^T\sqrt{D_1}\bar{Q}$. Let the SVD of $\sqrt{D_1}XX^T\sqrt{D_1}\bar{Q} = EDU^T$, we have,

$$\bar{P} = EU^T \quad (15)$$

Therefore,

$$P = \sqrt{D_1}^{-1}EU^T \quad (16)$$

Convergence Analysis: Lemma 2

For any non-zero vectors $p, q \in R^c$, the following result holds:

$$\|p\|_2 - \frac{\|p\|_2^2}{\|q\|_2} \leq \|q\|_2 - \frac{\|q\|_2^2}{\|q\|_2} \quad (17)$$

Convergence Analysis: Theorem 1

Recall eq. (1)

$$\arg \min_{Q,P} J(Q, P) = \arg \min_{Q,P} \|X - PQ^T X\|_{2,1} + \lambda \|Q\|_{2,1} \quad (18)$$

where transformation matrix $Q \in R^{m \times d}$ is first used to project the data matrix X onto a low-dimensional subspace and another transformation matrix $P \in R^{m \times d}$ is then used to recover the data matrix X . $\lambda \geq 0$ is the regularization parameter.

Theorem 1: Given all the variables in eq. (1) except P , Q , the optimization problem in eq. (1) will monotonically decrease the objective function value in each iteration and converge to the local optimal solution.

Convergence Analysis: Proof of Theorem 1

We denote the objective function as $(J(Q, P) = J(Q, P, D_1, D_2))$. Suppose for the $(t - 1)$ -th iteration, we obtain $P^{(t-1)}$, $Q^{(t-1)}$, $D_1^{(t-1)}$ and $D_2^{(t-1)}$. From eq (10), we can find

$$J(Q^{(t)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) \leq J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) \quad (19)$$

Since the SVD gives the optimal $P^{(t)}$ that further decreases the objective value, we have


$$J(Q^{(t)}, P^{(t)}, D_1^{(t-1)}, D_2^{(t-1)}) \leq J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) \quad (20)$$

Convergence Analysis: Proof of Theorem 1 (cont.)

Once the optimal $P^{(t)}$ and $Q^{(t)}$ are obtained, we have

$$\begin{aligned} & \text{tr}((X - P^{(t)}Q^{(t)}X)^T D_1^{(t-1)}(X - P^{(t)}Q^{(t)}X) \\ & + \lambda(Q^{(t)T}D_2^{(t-1)}Q^{(t)}) \\ & \leq \text{tr}((X - P^{(t-1)}Q^{(t-1)}X)^T D_1^{(t-1)}(X - P^{(t-1)}Q^{(t-1)}X) \\ & + \lambda(Q^{(t-1)T}D_2^{(t-1)}Q^{(t-1)}) \end{aligned} \quad (21)$$

That is,

$$\begin{aligned} & \sum_{i=1}^m \frac{\|X - P_i^{(t)}Q_i^{(t)T}X\|_2^2}{\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\|_2} + \lambda \sum_{i=1}^m \frac{\|Q_i^{(t)}\|_2^2}{\|Q^{(t-1)}\|_2} \\ & \leq \sum_{i=1}^m \frac{\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\|_2^2}{\|X - P_i^{(t-1)}Q_i^{(t-1)T}X\|_2} + \lambda \sum_{i=1}^m \frac{\|Q_i^{(t)}\|_2^2}{\|Q^{(t-1)}\|_2} \end{aligned} \quad (22)$$


Convergence Analysis: Proof of Theorem 1 (cont.)

From Lemma 2, we have,

$$\begin{aligned} & \|X - P_i^{(t)} Q_i^{(t)T} X\|_2 - \frac{\|X - P_i^{(t)} Q_i^{(t)T} X\|_2^2}{\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2} \\ & \leq \|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2 - \frac{\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2^2}{\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2} \end{aligned} \quad (23)$$

Using the matrix calculus on eq. (23), we have,

$$\begin{aligned} & \sum_{i=1}^m \left(\|X - P_i^{(t)} Q_i^{(t)T} X\|_2 - \frac{\|X - P_i^{(t)} Q_i^{(t)T} X\|_2^2}{\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2} \right) \\ & \leq \sum_{i=1}^m \left(\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2 - \frac{\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2^2}{\|X - P_i^{(t-1)} Q_i^{(t-1)T} X\|_2} \right) \end{aligned} \quad (24)$$

cmi

Convergence Analysis: Proof of Theorem 1 (cont.)

Again, from Lemma 2, we have,

$$\|Q_i^{(t)}\|_2 - \frac{\|Q_i^{(t)}\|_2^2}{\|Q_i^{(t-1)}\|_2} \leq \|Q_i^{(t-1)}\|_2 - \frac{\|Q_i^{(t-1)}\|_2^2}{\|Q_i^{(t-1)}\|_2} \quad (25)$$

Again, using matrix calculation, we have,

$$\sum_{i=1}^m \left(\|Q_i^{(t)}\|_2 - \frac{\|Q_i^{(t)}\|_2^2}{\|Q_i^{(t-1)}\|_2} \right) \leq \sum_{i=1}^m \left(\|Q_i^{(t-1)}\|_2 - \frac{\|Q_i^{(t-1)}\|_2^2}{\|Q_i^{(t-1)}\|_2} \right) \quad (26)$$

Convergence Analysis: Proof of Theorem 1 (cont.)

Combining eq. (22), (24) and (26), we have

$$\begin{aligned} J(Q^{(t)}, P^{(t)}, D_1^{(t-1)}, D_2^{(t-1)}) &= \|XP^{(t)}Q^{(t)T}X\|_{2,1} + \lambda\|Q^{(t)}\|_{2,1} \\ &\leq \|XP^{(t-1)}Q^{(t-1)T}X\|_{2,1} + \lambda\|Q^{(t-1)}\|_{2,1} \\ &= J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) \end{aligned} \quad (27)$$

That is,

$$\begin{aligned} J(Q^{(t)}, P^{(t)}) &= J(Q^{(t)}, P^{(t)}, D_1^{(t)}, D_2^{(t)}) \\ &\leq J(Q^{(t-1)}, P^{(t-1)}, D_1^{(t-1)}, D_2^{(t-1)}) = J(Q^{(t-1)}, P^{(t-1)}) \end{aligned} \quad (28)$$

Algorithm

Algorithm

Input: Training sample set X , parameter λ and dimensionality d .

1. Initialize $D_1 = I^{m \times m}$ and random $\bar{P}^{m \times d}$.
2. while not converge do
 - 2.1 Compute \bar{Q} according to eq.
 - 2.2 Compute Q according to eq.
 - 2.3 Compute \bar{P} according to eq.
 - 2.4 Compute P according to eq.
 - 2.5 Compute D_1 according to eq.
 - 2.6 Compute D_2 according to eq.
3. Normalize each column vector of Q to be identity vectors.

Output: Transformation matrix Q

Computational Complexity

- In each iteration, two main steps.
- Computing $Q = (\lambda D_2 + XX^T)^{-1} XX^T \sqrt{D_1} P$ takes $\mathcal{O}(m^3)$ time.
- Computing SVD of $\sqrt{D_1} XX^T \sqrt{D_1} Q = EDU^T$ also takes $\mathcal{O}(m^3)$ time at most.
- So, computational complexity for one iteration will be up to $\mathcal{O}(m^3)$.
- For t iterations, it will be $\mathcal{O}(tm^3)$.

Conclusion

Conclusion

- JSPCA is designed by relaxing the orthogonal constraint of transformation matrix Q , introducing another transformation matrix P
- It is imposing joint 2,1-norms on both loss term and regularization term.
- The proposed method has more freedom to jointly select the useful features for a low-dimensional representation and is robust to outliers.
- A simple yet effective algorithm is designed for the optimization problem.
- The algorithm is quite stable.

References

- F. Nie, H. Huang, X. Cai, C.H. Ding, Efficient and robust feature selection via joint $l_{2,1}$ -norms minimization, in: Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.
- Q. Gu, Z. Li, J. Han, Joint feature selection and subspace learning, in: International Joint Conference on Artificial Intelligence, 2011, pp. 1294–1299.
- H. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, A survey of multilinear subspace learning for tensor data, Pattern Recognit. 44 (7) (2011) 1540–1551.
- W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, Z. Zhu, Robust face recognition via occlusion dictionary learning, Pattern Recognit. 47 (4) (2014) 1559–1572.
- J. Huang, X. You, Y. Yuan, F. Yang, L. Lin, Rotation invariant iris feature extraction using Gaussian Markov random fields with non-separable wavelet, Neurocomputing 73 (4) (2010) 883–894.

Thank you!