



## 6.6 PCA + NMF

2.5 Why need feature extraction?

A. Many pattern recognition techniques may not be effective for high-dimensional data. (Aka curse of dimensionality) Meanwhile, pattern itself may just have small variance.

Q2. What is the benefit of feature extraction?

A: Visualisation, data compression, (EER), efficient storage and transmission.

In PCA, we want to capture the big variability and ignore the small variability, and after ordering, we choose the first k' direction to project the raw data on. Each  $P_i$  is orthogonal to each other. Given a data set  $\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$  (using PCA, we project  $x_i = a_i^T x_i$ , where  $a_i$  is chosen such that  $|v[a_i]| = a_i^T a_i = 1$ , where  $v$  is the covariance matrix. The solution:  $a_1$  is the eigenvector of  $S$  corresponding to the largest eigenvalue. And in general,  $a_1, a_2, \dots, a_d$  are the  $k'$  greatest variation.

STEPS:

1. covariance matrix  
 $S = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$
2. eigenvectors:  $\{a_i\}_{i=1}^d$
3. first p eigenvectors:  $\{a_i\}_{i=1}^p$
4. reconstruction:  $G = [a_1 | a_2 | \dots | a_p]_{d \times p}$
5.  $y = G^T x$

Using SVD:

- centering:  $X_{nd} = [(x_1 - \bar{x}), \dots, (x_n - \bar{x})]^T$
- SVD:  $X = U_d \Lambda_d V_{dn}^T$  ( $S = X^T X = U_d U_d^T$ )
- reconstruction:  $\hat{x}_i = U_d \Lambda_d V_{dn}^T x_i$
- eigenvalues:  $G = [a_1 | a_2 | \dots | a_p]_{d \times p}$
- $y = G^T x$

A criterion for  $p$ :  $\frac{\lambda_1 + \lambda_2 + \dots + \lambda_p}{\lambda_1} > \text{Threshold}(\text{e.g. } 0.95)$

In NMF, we express an image vector as the linear combination of a set of basis images.

PCA has some disadvantages due to only (PCA involves subtraction due to some negative values after de-mean). Factorization:

$V_{nm} = W_n H_m$ .  $V_{nm}$  contains m images, which one has n non-negative pixel values.  $W$  has r columns of basis images; each column in  $H$  is called coding.

STEPS:

1. objective:  
 $J_{NMF}(W, H) = \|V - WH\|^2 \quad \text{s.t.} \quad W \geq 0, H \geq 0$
2. a prior constraint:  $H_{ij} = 0 \Rightarrow \sum_{i=1}^m W_{ij} H_{ij} = 0$
3. gradient descent:  $H_{ij} \leftarrow H_{ij} + \eta W_{ij} (V_{ij} - W_{ij} H_{ij})$
4. multiplicative update rule:  
 $H_{ij} \leftarrow H_{ij} \cdot \frac{(V_{ij}^T W)}{(W^T W H_{ij})}$
5.  $(W_{ij})_{ij} = \eta (V_{ij}^T W) / (W^T W H_{ij})$

In summary: PCA has holistic representation, while NMF has part-based; PCA uses eigenvectors as basis image, while NMF uses localized features; each face in PCA is a linear combination of all faces, while in NMF only additive combinations.

Data points are represented in a rotated orthogonal coordinate system: the origin is the mean of the data points and the axes are provided by the eigenvectors.

PCA provides a more fine-grained structure. Drawbacks of PCA:

1. PCA involves adding up some basis vectors the subspaces are not orthogonal.
2. Basis vectors aren't physically intuitive (but based on the assumption that they are).
3. Subtracting doesn't make sense in context of some applications.

• What are differences between NMF and PCA?

NMF	PCA
Representation	Part-based
Basic image	Localized features
Constraints on W and H	Allow multiple basis images to represent a face but only additve combinations

## LT: LDA + GE

LDA finds the most discriminative projection by maximizing between-class distance and minimizing within-class distance. We define:

Class-specific means:  $\mu_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}, C_i$

Class-specific covariance:  $S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T$

Total mean:  $\mu = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}$

Within-class scatter:  $S_w = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \mu_i)(x_{ij} - \mu_i)^T$

Between-class scatter:  $S_b = \sum_{i \neq j} P_{ij} (\mu_i - \mu_j)(\mu_i - \mu_j)^T$

Total covariance:  $S_t = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} (\mu_i - \mu)(\mu_i - \mu)^T = S_w + S_b$

In two-class case, given data:

$\{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$ , we seek a projection w (onto a scalar):  $w^T x$ . Thus within-class scatter:  $S_w = \sum_{i=1}^2 \sum_{j=1}^{n_i} (w^T x_{ij} - w^T \mu_i)^2$ ; between-class scatter:  $S_b = \sum_{i \neq j} P_{ij} (\mu_i - \mu_j)(\mu_i - \mu_j)^T$ . And Fisher's linear discriminant is to maximize  $J(w) = \frac{w^T S_b w}{w^T S_w w}$ . (NOTE that  $S_b$  is the outer product of two vectors, its rank is at most 1). Taking derivative and setting to zero yields:  $(w^T S_w w)S_b w - 0 \Rightarrow S_w^T S_b w = J(w)w$  (Eigenvalue decomposition - since  $S_b$  is at most rank 1, only take the first eigenvalue).

In multi-class case, we can project on  $\{y_1, y_2, \dots, y_k\} \subset \mathbb{R}^d$ , where  $y_i$  is a matrix  $W = [y_1 | y_2 | \dots | y_k]$ . We want to maximize  $J(W) = \frac{W^T S_b W}{W^T S_w W}$  the optimal projections. The eigenvalues/eigenvectors are the eigenvectors corresponding to the largest eigenvalues of the following generalized eigenvalue problem:  $W^* = \arg \max_{W^T S_w W} \frac{W^T S_b W}{W^T S_w W} = \arg \max_{W^T S_w W} \frac{\langle S_b, W \rangle}{\langle S_w, W \rangle}$ .

$\lambda_1, \lambda_2, \dots, \lambda_k$  are singular. Limitations of LDA are 1) at most  $C-1$  feature projections; 2) parametric (assumes multivariate Gaussian likelihoods).

For  $S_b$  that each  $\mu_j - \mu$  is the first. For this, initially协方差矩阵的大小等于子集的阶数的乘积的秩的大小。但是由于我们只处理k-1个类，最后一个类的协方差矩阵的秩为k-1，因此该秩的秩为k-1。

Q3. Is LDA always better than PCA?

A: There has been a tendency in the computer vision community to prefer LDA over PCA.

This is mainly because LDA deals directly with discrimination between classes while PCA does not pay attention to the underlying class structure.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.

This paper shows that when the training set is small, PCA can outperform LDA.

The number of samples is large and representative for each class. LDA outperforms PCA.