THE UNIVERSITY of ADELAIDE

ISML_5: Dimensionality Reduction

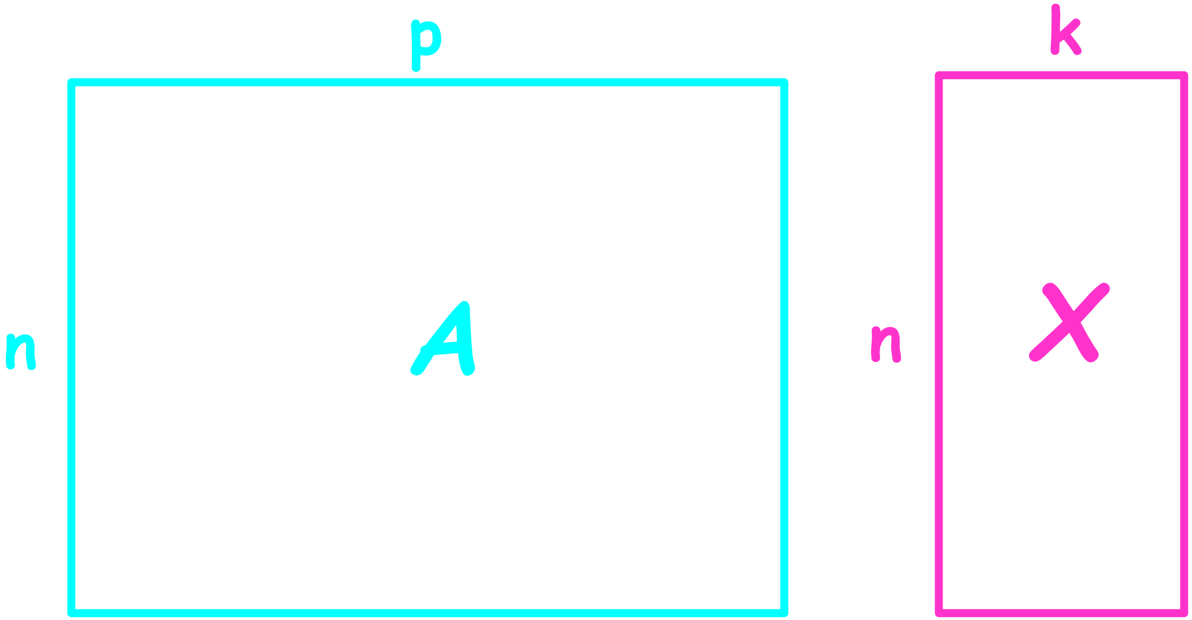
Lingqiao Liu

Outline

- Overview of dimensionality reduction
  - The benefits
  - Why we can perform dimensionality reduction
  - Type of dimensionality reduction methods
- Principal component analysis
  - Mathematical basics
  - Decorrelation and dimension selection
  - Eigenface and high-dimensionality issue
- Linear Discriminative Analysis
- Summary

What is dimensionality reduction

- Reduce dimensions of data



Dimensionality Reduction: why?

- Extract underlying factors

	Disagree		Neutral		Agree
I am the life of the party.					
I feel little concern for others.					
I am always prepared.					
I get stressed out easily.					
I have a rich vocabulary.					
I don't talk a lot.					
I am interested in people.					
I leave my belongings around.					
I am relaxed most of the time.					
I have difficulty understanding abstract ideas.					
I feel comfortable around people.					
I insult people.					
I pay attention to details.					
I worry about things.					
I have a vivid imagination.					

The five factors [edit]

A summary of the factors of the Big Five and their co

- **Openness to experience**: (*inventive/curious* vs intellectual curiosity, creativity and a preference f preference for a variety of activities over a strict r experience.
- **Conscientiousness**: (*efficient/organized* vs. *eas* spontaneous behavior.
- **Extraversion**: (*outgoing/energetic* vs. *solitary/re* talkativeness.
- **Agreeableness**: (*friendly/compassionate* vs. *an* one's trusting and helpful nature, and whether a p
- **Neuroticism**: (*sensitive/nervous* vs. *secure/conf* degree of emotional stability and impulse control

Dimensionality Reduction: why?

- Reduce data noise
  - Face recognition
  - Applied to image de-noising

(a) Noisy image (b) NL means (PSNR=32.90) (c) Local PCA (PSNR=33.70)

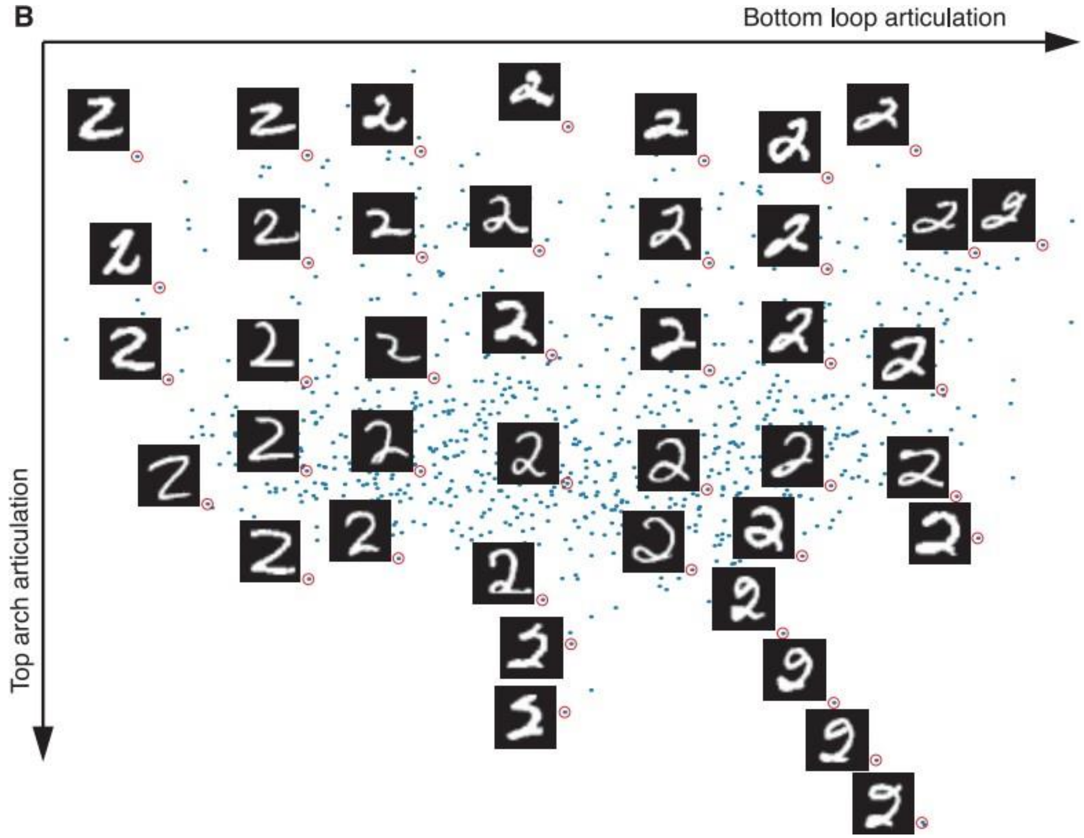
Image courtesy of Charles-Alban Deledalle, Joseph Salmon, Arnak Dalalyan; BMVC 2011
Image denoising with patch-based PCA: local versus global

Dimensionality Reduction: why?

- Reduce the number of model parameters
  - Avoid over-fitting
  - Reduce computational cost

Dimensionality Reduction: why?

- Visualization



Dimensionality Reduction

- General principle:
  - Preserve “useful” information in low dimensional data
- How to define “usefulness”?
  - Many
  - An active research direction in machine learning
- Taxonomy
  - Supervised or Unsupervised
  - Linear or nonlinear
- Commonly used methods:
  - PCA, LDA (linear discriminant analysis), and more.
- Feature Selection vs dimensionality reduction

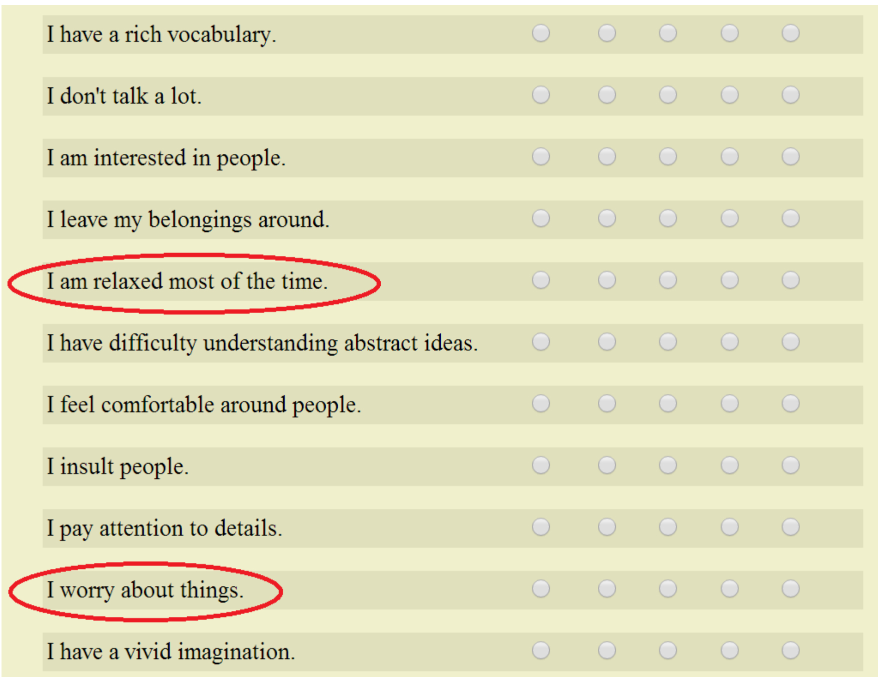
Outline

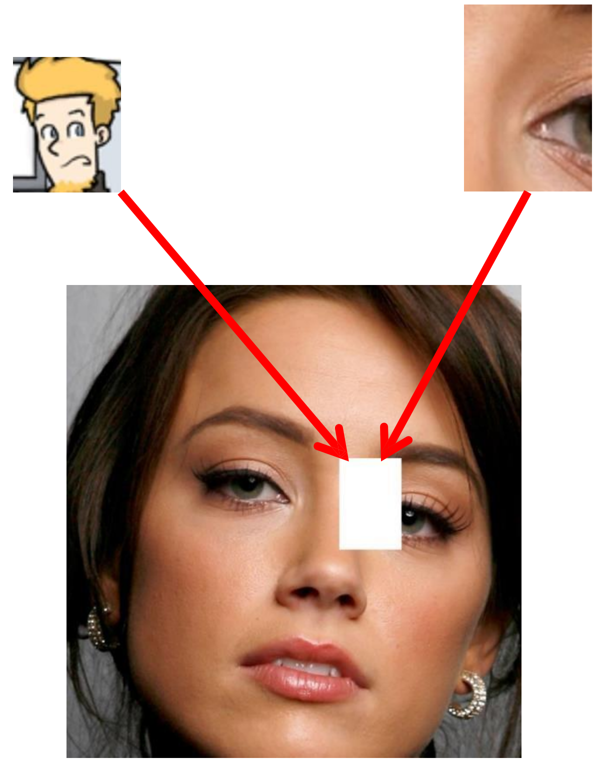
- Overview of dimensionality reduction
  - The benefits
  - Why we can perform dimensionality reduction
  - Type of dimensionality reduction methods
- Principal component analysis
  - Mathematical basics
  - Decorrelation and dimension selection
  - Eigenface and high-dimensionality issue
- Linear Discriminative Analysis
- Summary

Principal Component Analysis: Motivation

- Principal Component Analysis (PCA) is an unsupervised dimensionality reduction method
  - Transform data to remove redundant information
  - Keep the most informative dimensions after the transformation

Data correlation & information redundancy



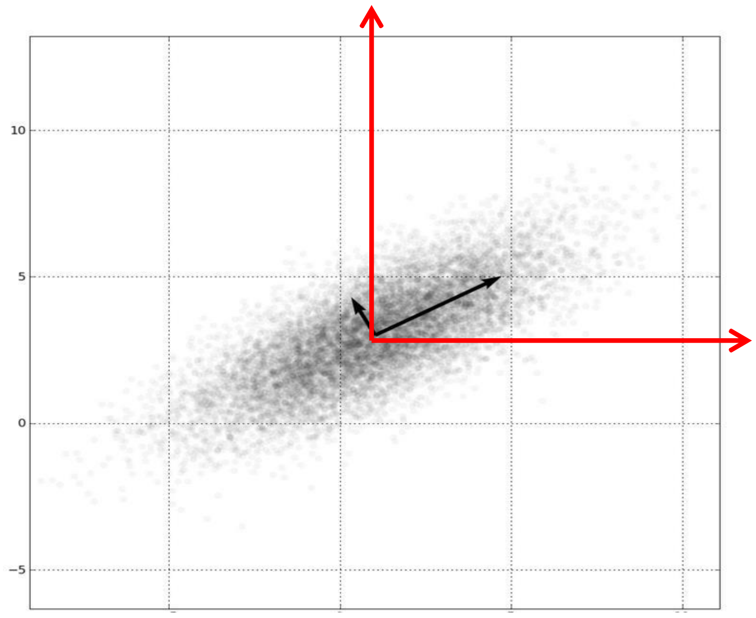


PCA explained: De-correlating data

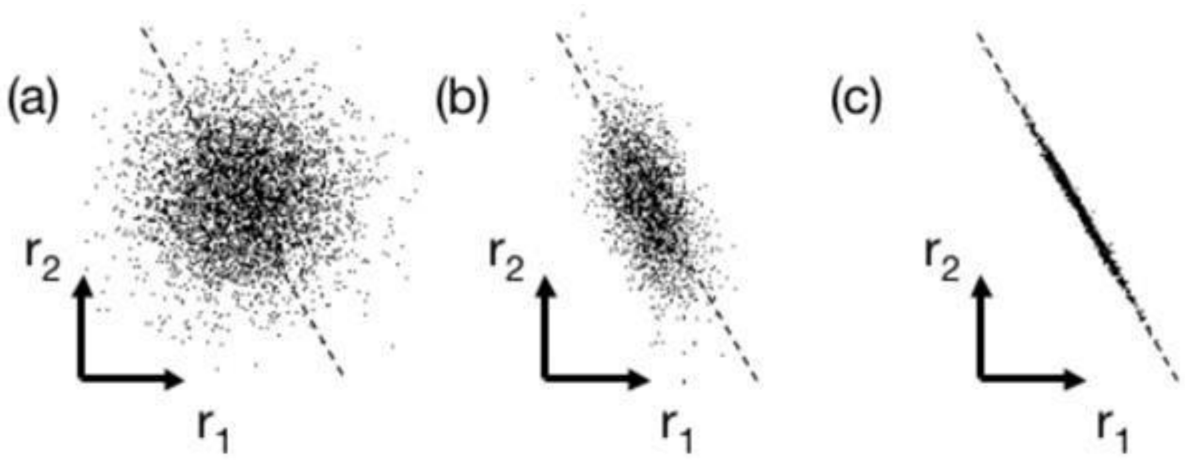
- Dependency vs. Correlation
  - Dependent is a stronger criterion
- Equivalent when data follows Gaussian distribution
- PCA only de-correlates data
  - One limitation of PCA
  - ICA, but it is more complicate

PCA explained: De-correlating data

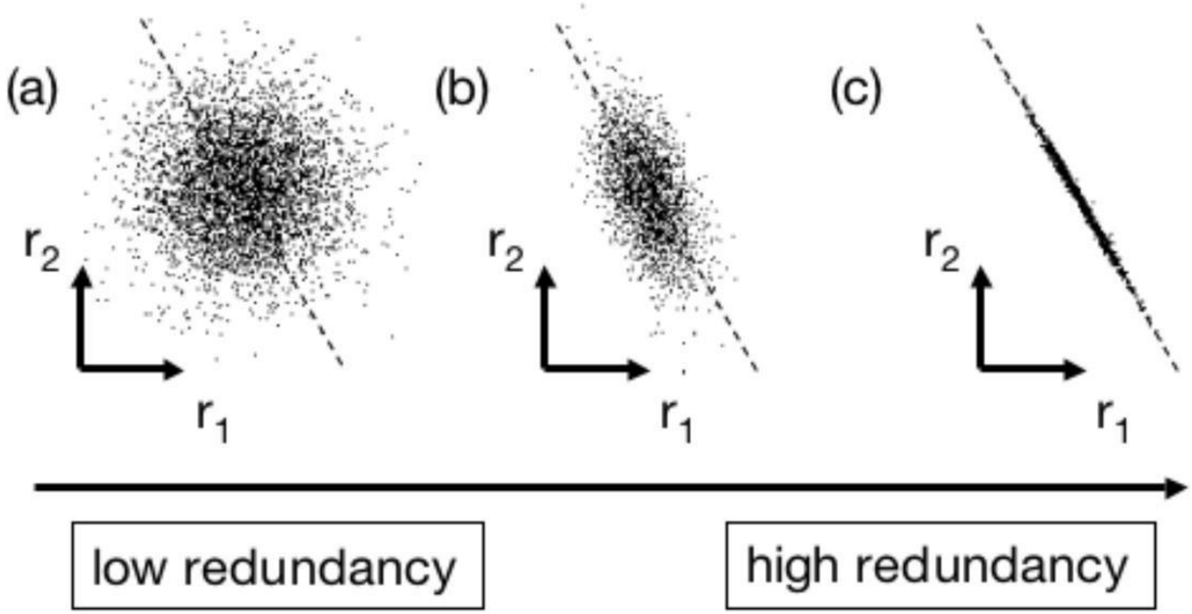
- Geometric interpretation of correlation



Exercise: which figure shows the highest data correlation?

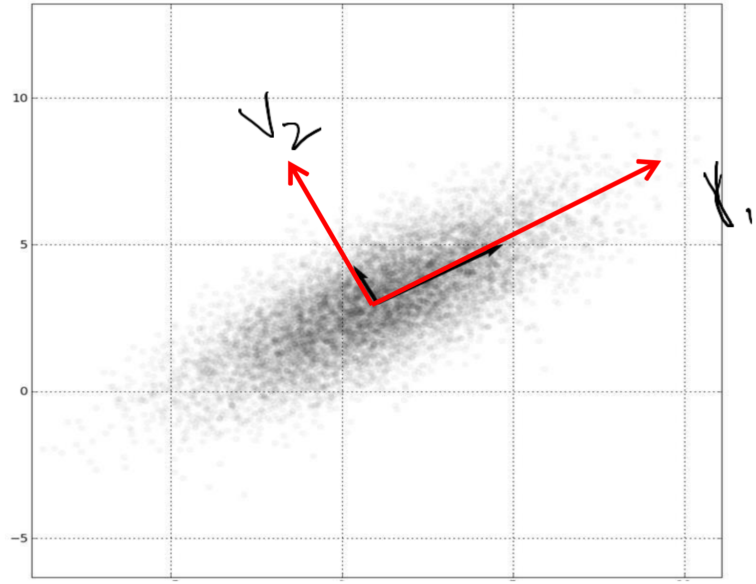


Exercise: which figure shows the highest data correlation?

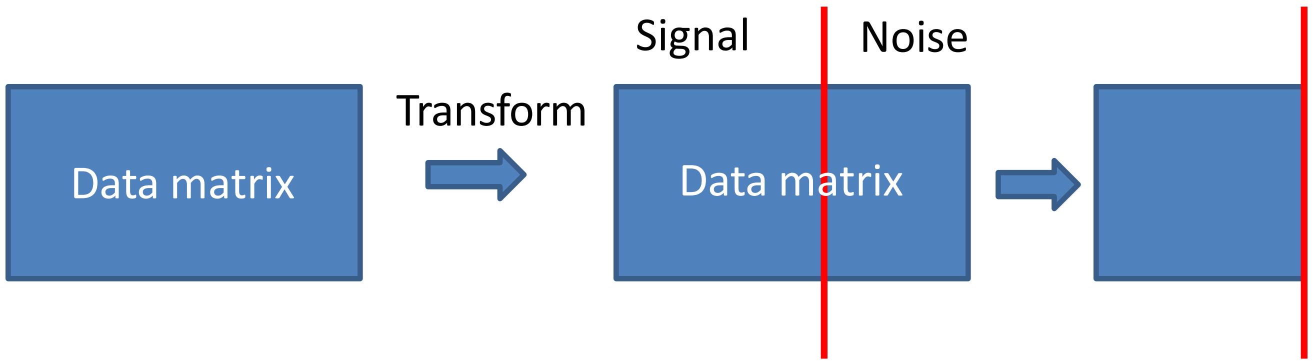


PCA explained: De-correlating data

- Correlation can be removed by rotating the data point or coordinate



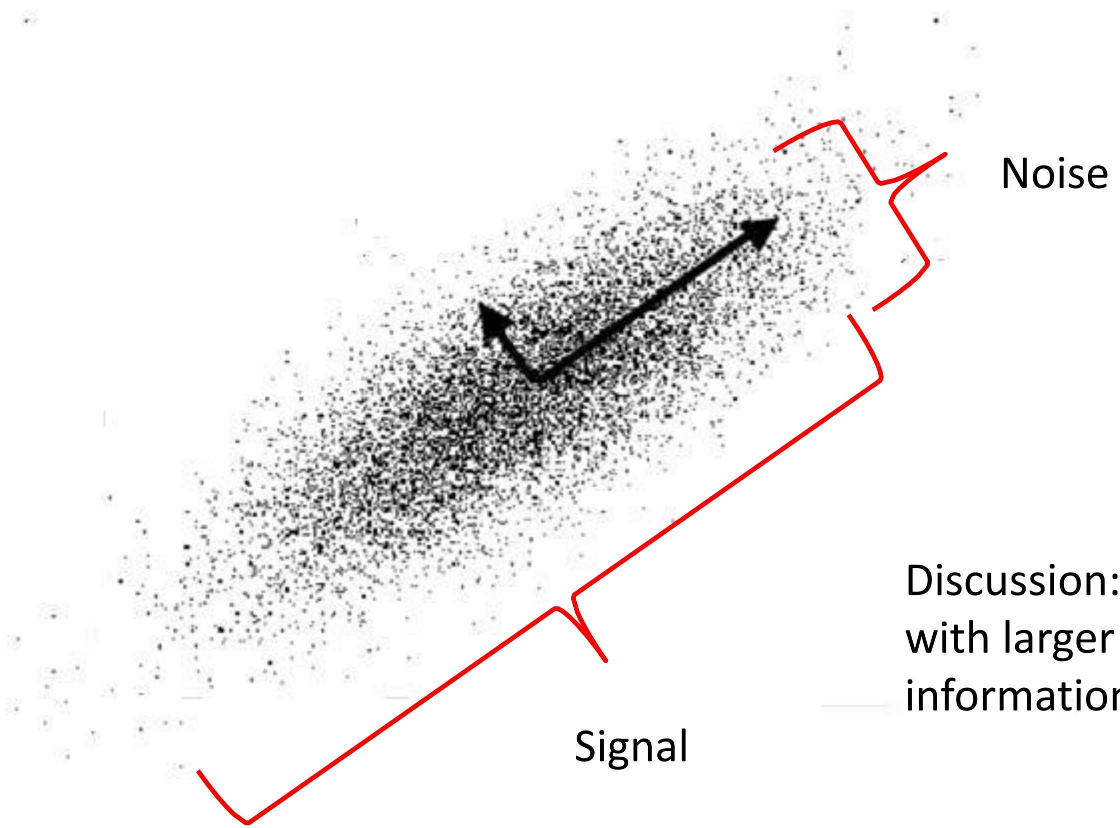
PCA explained: SNR maximization



- Maximize

$$SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}.$$

PCA explained: SNR maximization



Discussion: why we think a dimension with larger covariance contain more information?

PCA explained

- Target
  - 1: Find a new coordinate system which makes different dimensions zero correlated
  - 2: keep the top-k dimensions with largest variance in the new coordinate system
- Method
  - Rotate the data point or coordinate
- Mathematically speaking…
  - How to rotate?
  - How to express our criterion

Mathematic Basics

- Mean, Variance, Covariance
- Matrix norm, trace,
- Orthogonal matrix, basis
- Eigen decomposition

Mathematic Basics

- (Sample) Mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- (Sample) Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

- (Sample) Covariance

$$cov(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Mathematatic Basics

- Covariance Matrix

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top$$

Or compactly

$$\mathbf{C} = \frac{1}{n}\mathbf{X_c}\mathbf{X_c}^\top, \quad \mathbf{X_c} \in \mathrm{R}^{d\times n}$$

$\mathbf{X_c}$ is the centralized data matrix (with mean subtracted)

When C is a diagonal matrix, dimensions of features become zero-correlated

Mathematic Basics

- Orthogonal matrix

$$Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I,$$

- Rotation effect

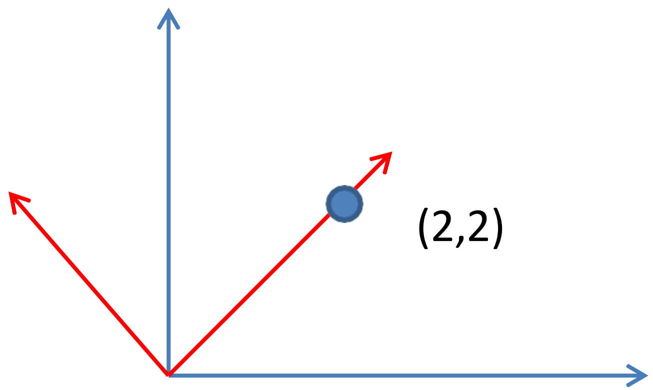
$$\|\mathbf{Qx}\|_2 = \|\mathbf{Qx}\|_F = \sqrt{\mathrm{trace}(\mathbf{x}^T\mathbf{Q}^T\mathbf{Qx})} = \sqrt{\mathrm{trace}(\mathbf{x}^T\mathbf{x})} = \|\mathbf{x}\|_F$$

$$\mathbf{x} = \mathbf{Q}^T\mathbf{Qx}$$

Mathematic Basics

- Relationship to coordinate system
  - A point = linear combination of bases
  - Combination weight = coordinate
- Each row (column) of Q = basis
  - Not unique
  - Relation to coordinate rotation
- New coordinate $\mathbf{Q}\mathbf{x}$

Mathematic Basics



$$\binom{2}{2} = 2\binom{1}{0} + 2\binom{0}{1}$$

$$\binom{2}{2} = 2\sqrt{2}\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} + 0\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$$

New coordinate

$$\binom{2\sqrt{2}}{0} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}\binom{2}{2}$$

Old coordinate

Mathematic Basics

Eigen-decomposition

- If $\mathbf{A}$ is symmetric

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T \quad Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I,$$

PCA: solution

- Target 1: de-correlation

$$\mathbf{C}_X = \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^\top$$ Original covariance matrix of data

$$\mathbf{Y} = \mathbf{P}\mathbf{X_c}$$ We are looking for transform the centralized data with a matrix P

Require $\mathbf{C}_Y = \frac{1}{n}\mathbf{P}\mathbf{X_c}(\mathbf{P}\mathbf{X_c})^\top = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{P}\mathbf{X_c}\mathbf{X_c}^\top\mathbf{P}^\top$ be a diagonal matrix

PCA: solution

- Target 1: de-correlation

$$\mathbf{C}_X = \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^\top$$ Original covariance matrix of data

$$\mathbf{Y} = \mathbf{P}\mathbf{X_c}$$ We are looking for transform the centralized data with a matrix P

Require $\mathbf{C}_Y = \frac{1}{n}\mathbf{P}\mathbf{X_c}(\mathbf{P}\mathbf{X_c})^\top = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{P}\mathbf{X_c}\mathbf{X_c}^\top\mathbf{P}^\top$ be a diagonal matrix

Sounds familiar?

PCA: solution

$$\mathbf{X}_c \in R^{d\times N}$$

$$\mathbf{P} \in R^{d\times d}$$

- Target 1: de-correlation

$\mathbf{C}_X = \frac{1}{n}\mathbf{X}_c\mathbf{X}_c^\top$ Original covariance matrix of data

$\mathbf{Y} = \mathbf{P}\mathbf{X_c}$ We are looking for transform the centralized data with a matrix P

Require $\mathbf{C}_Y = \frac{1}{n}\mathbf{P}\mathbf{X_c}(\mathbf{P}\mathbf{X_c})^\top = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{P}\mathbf{X_c}\mathbf{X_c}^\top\mathbf{P}^\top$ be a diagonal matrix

Sounds familiar? Use eigen decomposition $\mathbf{X_c}\mathbf{X_c}^\top = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ $Q^{\mathrm{T}}Q = QQ^{\mathrm{T}} = I$

$$\begin{aligned}\mathbf{C_Y} &= \frac{1}{n}\mathbf{P}\mathbf{X_c}\mathbf{X_c}^\top\mathbf{P}^\top \\ &= \frac{1}{n}\mathbf{P}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top\mathbf{P}^\top\end{aligned}$$

$\Rightarrow$ set $\mathbf{P} = \mathbf{Q}^\top$

so $\mathbf{C_Y} = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{\Lambda}$

Recap: Matrix decomposition

- Matrix can be decomposed into the combination (usually product) of special matrices
- Eigen decomposition

$$A = Q\Lambda Q^{-1}$$

where Λ is a diagonal matrix, with its i-th diagonal value be the i-th eigenvalue of A. Q is a matrix with its ith column be the eigenvector corresponding to ith eigenvalue.

  - When A is symmetric, i.e. $A = A^\top$

$$A = Q\Lambda Q^\top \qquad Q^\top Q = QQ^\top = I$$

- Related topic: Singular value decomposition

A closer look at the covariance matrix of y

- The covariance matrix of the transformed data is a diagonal matrix, what are the diagonal elements?
- From the perspective of covariance matrix

$$\frac{1}{n}[\mathbf{Y}\mathbf{Y}^\top]_{\mathbf{i},\mathbf{j}} = \mathbf{E}(\mathbf{y_i}\mathbf{y_j})$$

The i, j th element of the covariance matrix is the covariance between the i-th dimension feature and j-th dimension feature. The i-th diagonal element represents the variance of the feature at the i-th dimension

- From the perspective of Eigen-decomposition, the diagonal matrix corresponding to the eigenvalues of $\mathbf{X_c}\mathbf{X_c}^\top$

How to select the most informative dimensions?

- We rank the dimensions by their variance, which is equivalent to rank the dimensions by their corresponding eigen values

$$\mathbf{X_c}\mathbf{X_c}^\top = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$$

$$\overset{Q}{\begin{pmatrix} | & | & & | \\ u_1 & u_2 & \cdots & v_d \\ | & | & & | \end{pmatrix}} \overset{\Lambda}{\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d\end{pmatrix}} \overbrace{\begin{pmatrix} - u_1 - \\ - u_2 - \\ \vdots \\ - v_d - \end{pmatrix}}^{P = Q^T}$$

  - Note that the i-th dimension of y is obtained by using the i-th row of P

$$\mathbf{Y} = \mathbf{P}\mathbf{X_c}, \quad \mathbf{y}_i = \mathbf{p}_i\mathbf{X_c}$$

  - If we only want to have k dimensions, we can simply select the eigen vectors corresponding to the top-k eigenvalues to form the projection matrix P

PCA: algorithm

- 1. Subtract mean
- 2. Calculate the covariance matrix
- 3. Calculate eigenvectors and eigenvalues of the covariance matrix
- 4. Rank eigenvectors by its corresponding eigenvalues
- 4. Obtain P with its row vectors corresponding to the top k eigenvectors

PCA: MATLAB code

```
Mu = mean(fea);
fea = fea - repmat(Mu,[size(fea,1),1]);
Cov = fea'*fea;
[V,D] = eig(Cov);
[value,rank_idx] = sort(diag(D),'descend');
P = V(:,rank_idx(1:10));
```

Note fea is with the size of N x d.

PCA: reconstruction

- From a new feature y, we can reconstruct its original feature x

$$\hat{\mathbf{x}} = \mathbf{P}^{\top}\mathbf{y} = \mathbf{P}^{\top}(\mathbf{P}\mathbf{X})$$

Similar to the rotation back operation, e.g., $\mathbf{x} = \mathbf{Q}^{T}\mathbf{Q}\mathbf{x}$

- Fun fact, we can also derive PCA algorithm from the following optimization problem

$$\min_{\mathbf{P}} \ \|\mathbf{X} - \mathbf{P}^{T}\mathbf{P}\mathbf{X}\|_{F}^{2}$$
$$s.t.\ \mathbf{P}\mathbf{P}^{T} = \mathbf{I}$$

PCA: reconstruction

- Reighley Quotient

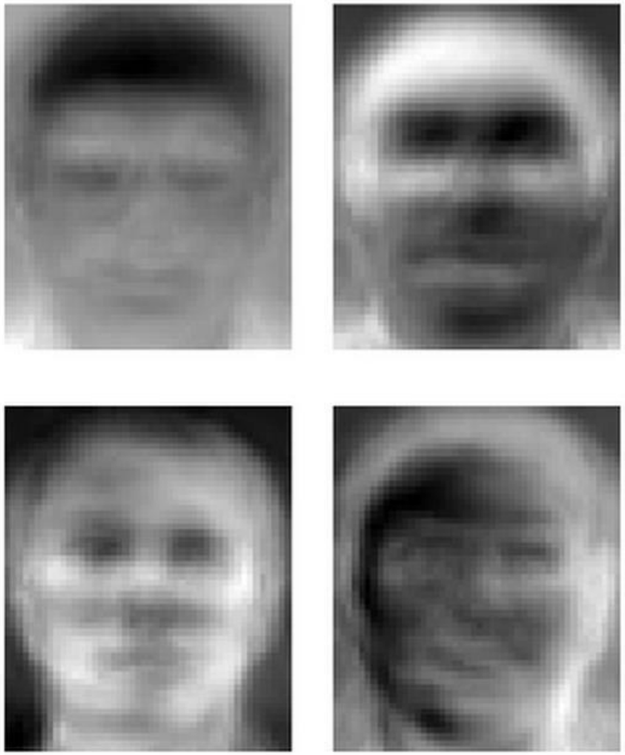
$$\max_{\mathbf{P}} \ \mathrm{trace}(\mathbf{P}\mathbf{X}\mathbf{X}^T\mathbf{P}^T)$$
$$s.t. \ \mathbf{P}\mathbf{P}^T = \mathbf{I}$$

- Solution = PCA

Application: Eigen-face method

- Sirovich and Kirby (1987) showed that PCA could be used on a collection of face images to form a set of basis features.
- Not only limited to face recognition
- Steps
  - Image as high-dimensional feature
  - PCA

Application: Eigen-face method



Some eigenfaces from AT&T Laboratories Cambridge

Application: Reconstruction

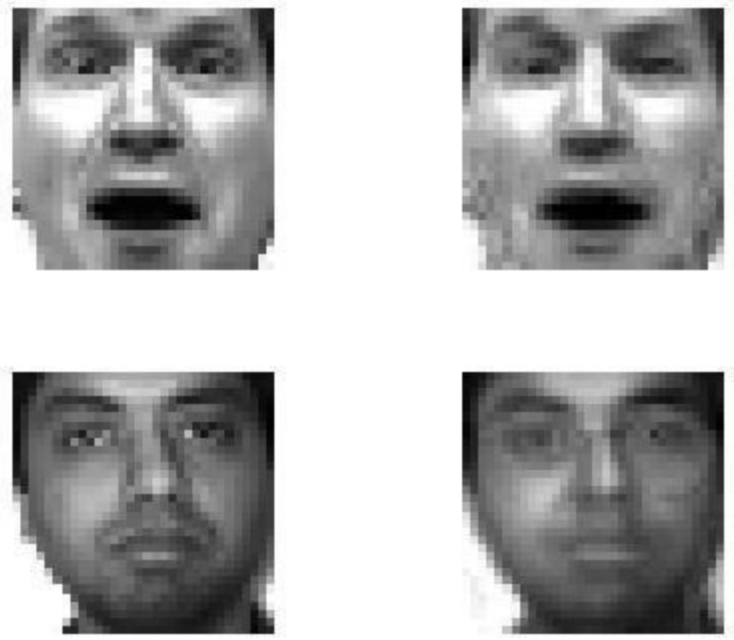
Reconstructed from top-2 eigenvectors

Application: Reconstruction

Reconstructed from top-15 eigenvectors

Application: Reconstruction

Reconstructed from top-40 eigenvectors



Application: Eigen-face method

- From large to small eigenvalues

high dimensionality issue

- For high-dimensional data large
- The number of samples is relatively small, $\mathbf{X_c X_c^\top} \in \mathbf{R^{d \times d}}$ can be too large
- Solution: a useful relation

Suppose $d \gg N$, we first consider the eigen value and vectors of the matrix $\mathbf{X}^T\mathbf{X}$

$$\mathbf{X}^T\mathbf{X}\mathbf{v} = \lambda\mathbf{v}$$

Multiply X on both side of equation

$$\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{v}) = \lambda(\mathbf{X}\mathbf{v})$$

If we define $\mathbf{u} = \mathbf{X}\mathbf{v}$

Then we know $\mathbf{X}\mathbf{X}^T\mathbf{u} = \lambda\mathbf{u}$

high dimensionality issue

- For high-dimensional data large
- The number of samples is relatively small, $\mathbf{X_c X_c^\top} \in \mathbf{R^{d \times d}}$ can be too large
- Solution: a useful relation

Suppose $d \gg N$, we first consider the eigen value and vectors of the matrix $\boxed{\mathbf{X}^T\mathbf{X}}$

$$\mathbf{X}^T\mathbf{X}\mathbf{v} = \lambda \mathbf{v}$$

Kernel matrix

Multiply X on both side of equation

$$\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{v}) = \lambda(\mathbf{X}\mathbf{v})$$

If we define $\mathbf{u} = \mathbf{X}\mathbf{v}$

Then we know $$\boxed{\mathbf{X}\mathbf{X}^T\mathbf{u} = \lambda \mathbf{u}}$$

The definition of the eigen system of covariance matrix

High dimensionality issue

- 1. Centralize data
- 2. Calculate the kernel matrix
- 3. Perform Eigen-decomposition on the kernel matrix and obtain its eigenvector $\mathbf{v}$
- 4. Obtain the Eigenvector of the covariance matrix by $\mathbf{u} = \mathbf{X}\mathbf{v}$
- Question? How many eigenvectors you can obtain in this way?

Mathematic Basics

- Rank of a matrix
  - In linear algebra, the rank of a matrix A is the dimension of the vector space generated (or spanned) by its columns
  - It is identical to the dimension of the vector space spanned by its rows

$$rank(\mathbf{A}) < min\{m, n\}, \quad \mathbf{A} \in R^{m \times n}$$

- Important relationship

$$\text{if} \quad \mathbf{A} = \mathbf{BC}$$
$$\mathrm{rank}(\mathbf{A}) \leq \min\{\mathrm{rank}(\mathbf{B}), \mathrm{rank}(\mathbf{C})\}$$

- Relationship to the eigenvalues

  the rank of a matrix equals to the number of non-zero eigenvalues of the matrix.

Back to the previous question

- So the rank of

$$rank(\mathbf{X}\mathbf{X}^\top) \leq rank(\mathbf{X}) \leq \min\{d, N\}$$

$$rank(\mathbf{X}^\top\mathbf{X}) \leq rank(\mathbf{X}) \leq \min\{d, N\}$$

- So at most $\min\{d, N\}$ meaningful projections from PCA

The eigen vector corresponding to 0 eigenvalue is meaningless in PCA

Outline

- Overview of dimensionality reduction
  - The benefits
  - Why we can perform dimensionality reduction
  - Type of dimensionality reduction methods
- Principal component analysis
  - Mathematical basics
  - Decorrelation and dimension selection
  - Eigenface and high-dimensionality issue
- Linear Discriminative Analysis
- Summary

Discriminative dimensionality reduction

- General principle:
  - Preserve “useful” information in low dimensional data
  - PCA: measure ``usefulness’’ through reconstruction error or covariance structure.
  - Useful for reconstruction ≠ useful for classification
- General principle for discriminative dimensionality reduction
  - Preserve “discriminative” information in low dimensional data

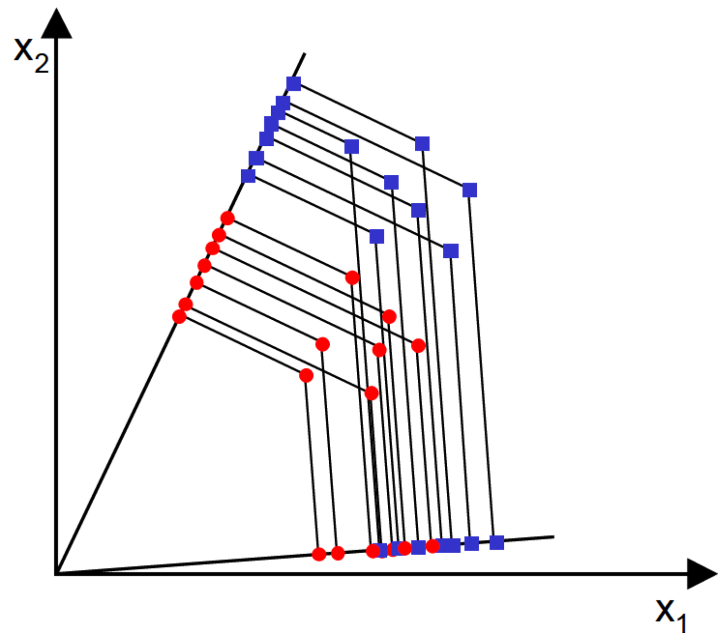
Linear Discriminant Analysis: Basic idea

- Linear Discriminant Analysis (LDA)
  - Discriminative dimensionality reduction
  - Linear dimensionality reduction $\mathbf{PX}$

- Supervised information
  - Class label
  - Data from the same class => Become close
  - Data from different classes => far from each other

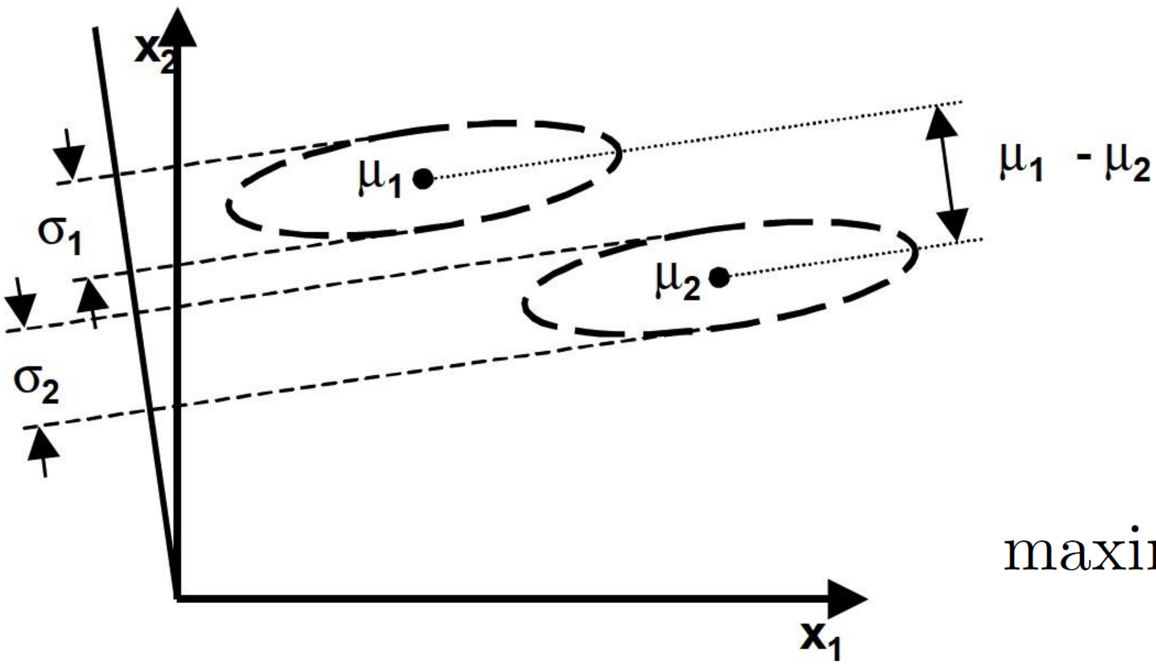
Linear Discriminant Analysis (LDA): Objective

- Two classes:



Linear Discriminant Analysis: Objective

- Two classes:



$$\text{maximize} \quad \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

Linear Discriminant Analysis (LDA): Objective

- Mean after projection:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}^T\mathbf{x}_i = \mathbf{p}^T\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \mathbf{p}^T\mu$$

- Variance after projection:

$$\begin{aligned}
\hat{\sigma}^{\mathbf{2}} &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{p}^T\mathbf{x}_i - \mathbf{p}^T\mu)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbf{p}^T(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T\mathbf{p} \\
&= \mathbf{p}^T\left(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T\right)\mathbf{p}
\end{aligned}$$

Linear Discriminant Analysis (LDA), Objective

- Mean distance

$$\left(\mathbf{p}^T\mu_1 - \mathbf{p}^T\mu_2\right)^2 = \mathbf{p}^T\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^T\mathbf{p}$$

- Between class Scatterness, within class Scatterness

$$\mathbf{S}_b = \left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^T$$

$$\mathbf{S}_w = \sum_{j=1,2}\frac{1}{N_j}\sum_{i=1}^{N_j}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

Linear Discriminant Analysis (LDA): Solution

- Objective

$$\text{maximize} \quad \frac{\mathbf{p}^T \mathbf{S}_b \mathbf{p}}{\mathbf{p}^T \mathbf{S}_w \mathbf{p}}$$

- Solution

$$\text{maximize} \quad \frac{\mathbf{p}^T \mathbf{S}_b \mathbf{p}}{\mathbf{p}^T \mathbf{S}_w \mathbf{p}} \quad \Longrightarrow \quad \begin{aligned} &\text{maximize} \quad \mathbf{p}^T \mathbf{S}_b \mathbf{p} \\ &\qquad s.t. \ \mathbf{p}^T \mathbf{S}_w \mathbf{p} = 1 \end{aligned}$$

$$L = \mathbf{p}^T \mathbf{S}_b \mathbf{p} - \lambda(\mathbf{p}^T \mathbf{S}_w \mathbf{p} - 1)$$

$$\frac{\partial L}{\partial \mathbf{p}} = 0 \quad \Rightarrow \quad \underline{\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{p} = \lambda \mathbf{p}}$$

Linear Discriminant Analysis (LDA), Solution

- Eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$
  - Why?

$$\text{maximize} \quad \mathbf{p}^T \mathbf{S}_b \mathbf{p} \qquad s.t. \; \mathbf{p}^T \mathbf{S}_w \mathbf{p} = 1$$

$$\frac{\partial L}{\partial \mathbf{p}} = 0 \quad \Rightarrow \quad \mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{p} = \lambda \mathbf{p}$$

At optimum, we have $\mathbf{p}^{*\top}\mathbf{S}_b\mathbf{p}^* = \lambda$

- Implementation details
  - What if $\mathbf{S}_w$ is not invertible?

Use $(\mathbf{S}_w + \lambda \mathbf{I})^{-1}$ instead

Linear Discriminant Analysis, Multi-class

- Generalized to multiple classes

$$\mathbf{S}_b = \sum_{i=1,j=1}^{C} (\mu_i - \mu_j)(\mu_i - \mu_j)^T = \sum_i (\mu_i - \mu)(\mu_i - \mu)^T$$

$$\mathbf{S}_w = \sum_{j=1}^{C} \sum_{i \in \mathcal{C}_j} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

$$\text{maximize} \quad \frac{\text{trace}\left(\mathbf{P}^T \mathbf{S}_b \mathbf{P}\right)}{\text{trace}\left(\mathbf{P}^T \mathbf{S}_w \mathbf{P}\right)}$$

- Solution:
  - Top c eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$
  - Discussion: how many projections you can have?

Linear Discriminant Analysis, Multi-class

- Generalized to multiple classes

$$\mathbf{S}_b = \sum_{i=1,j=1}^{C} (\mu_i - \mu_j)(\mu_i - \mu_j)^T = \sum_i (\mu_i - \mu)(\mu_i - \mu)^T$$

$$\mathbf{S}_w = \sum_{j=1}^{C} \sum_{i \in \mathcal{C}_j} (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T$$

$$\text{maximize} \quad \frac{\text{trace}\left(\mathbf{P}^T \mathbf{S}_b \mathbf{P}\right)}{\text{trace}\left(\mathbf{P}^T \mathbf{S}_w \mathbf{P}\right)}$$

- Solution:
  - Top c eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$
  - Discussion: how many projections you can have?

Check the rank of $\mathbf{S}_w^{-1}\mathbf{S}_b$

$$rank(\mathbf{S}_w^{-1}\mathbf{S}_b) \leq rank(\mathbf{S}_b) \leq C$$

At most C projections!

Summary

- The benefit of dimensionality reduction
- Principal component analysis
  - Two basic steps: decorrelation and select the dimensions with top variances
  - How to perform data transformation?
  - Its application, eigen face. The meaning behind eigen vectors
  - High dimensionality issue
- Linear discriminative component analysis
  - Basic objective
  - Two-class case
  - Multi-class case