Xi'an Jiaotong-Liverpool University

Pattern Recognition

Lecture 17. Dimensionality Reduction : PCA

Dr. Shanshan ZHAO

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University

Academic Year 2023-2024

Table of Contents

1. Introduction
2. PCA
   - Intuition
   - Formulization
   - Examples
3. Exercises
4. Feature Selection*(optional)

Outline

1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

Problems of Dimensionality

Problems of Dimensionality

- In practical multicategory applications, it is not at all unusual to encounter problems involving fifty or a hundred features, particularly if the features are binary valued.
- We might typically believe that each feature is useful for at least some of the discriminations; while we may doubt that each feature provides independent information, intentionally superfluous features have not been included.
- There are two issues that must be confronted.
  - The most important is how classification accuracy depends upon the dimensionality (and amount of training data);
  - the second is the computational complexity of designing the classifier.

Introduction

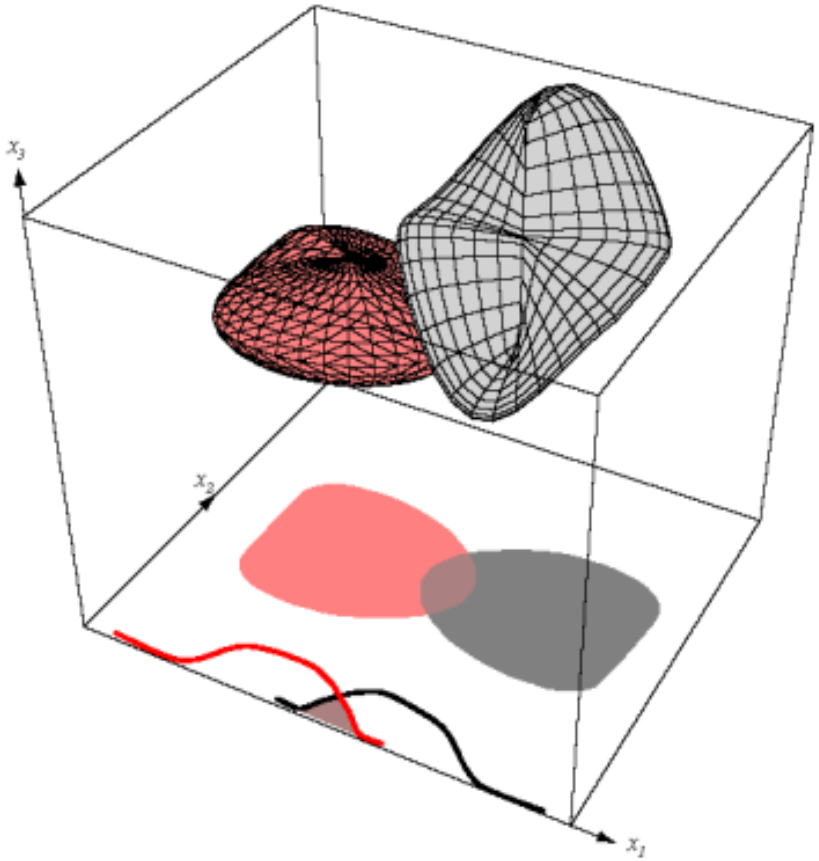


Figure 1: Two three-dimensional distributions have non overlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace here, the two-dimensional x_1, x_2 subspace or a one-dimensional x_1 subspace, there can be greater overlap of the projected distributions, and hence greater Bayes errors.

Curse of Dimensionality

- Unfortunately, it has frequently been observed in practice that, beyond a certain point, **the inclusion of additional features** leads to worse rather than better performance
- This is the so-called **curse of dimensionality**.

Curse of Dimensionality

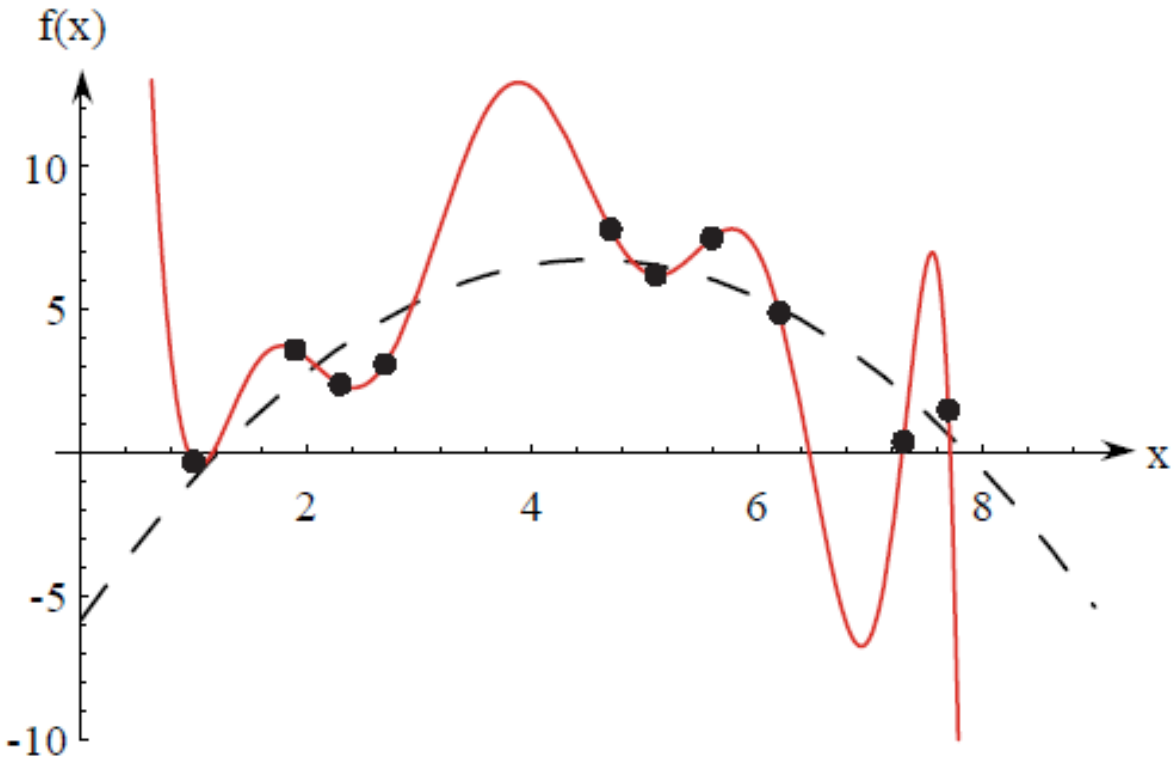


Figure 2: The 'training data' (black dots) were selected from a quadradic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, since it would lead to better predictions for new samples, for better generalization.

Curse of Dimensionality

- Almost **all** commonly used classifiers suffer from the curse of dimensionality.
- The exact relationship between the probability of error, the number of training samples, the number of features, and the number of parameters is **very difficult to establish**.
- Generally, the ratio of the training samples per class to the number of feature ($n/d > 10$) is accepted as good practice.
- Larger **ratio** of sample size to dimensionality should be considered for classifiers with more complexity.

Dimension Reduction

One way of coping with the problem of high dimensionality is to reduce the dimensionality by transforming features.

- considerations in dimension reduction:
  - ***Linear*** vs. *non-linear* transformations.
  - whether to use class *labels* or not (depends on the availability or the application).
- Training objectives:
  - minimizing classification error (discriminative training)
  - minimizing reconstruction error (**PCA**)
  - maximizing class separability (**LDA**)
  - making features as independent as possible (ICA)
  - embedding to lower dimensional manifolds (Isomap, LLE)

Outline

1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

PCA

- Principal components analysis (PCA) is a technique that can be used to simplify a dataset.
- It is a linear transformation that **chooses a new coordinate system** for the data set such that **greatest variance** by any projection of the data set comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis, and so on.
- PCA can be used for reducing dimensionality by eliminating the later principal components.

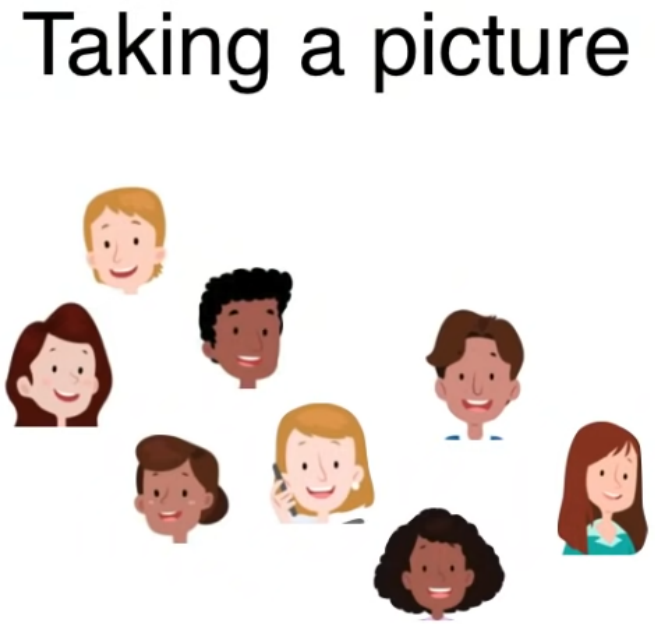
1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

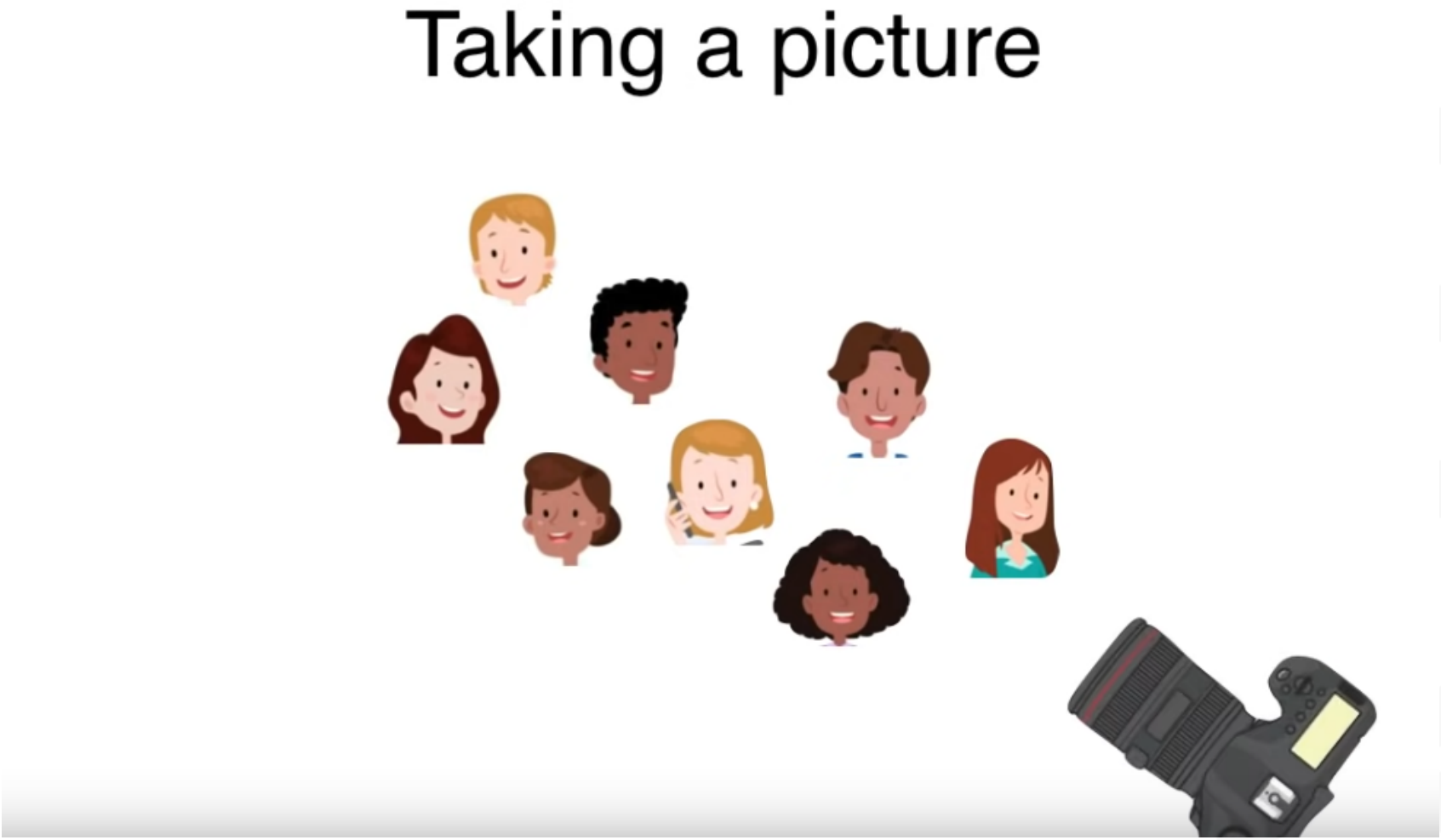
PCA Intuition



PCA Intuition

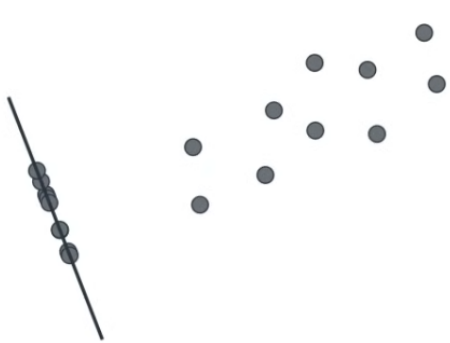
Taking a picture

PCA Intuition

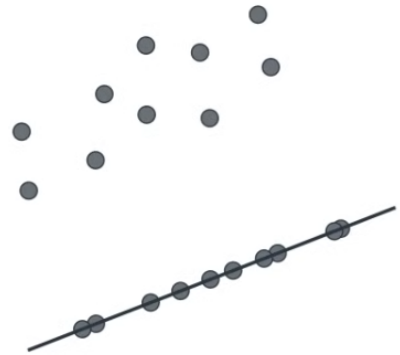


PCA Intuition

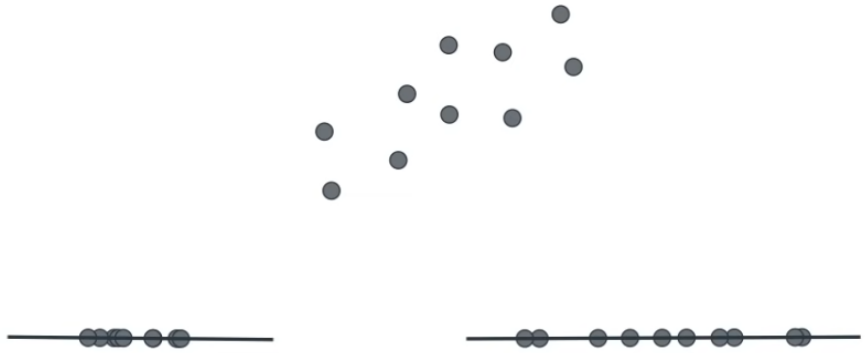
PCA Intuition



(a)

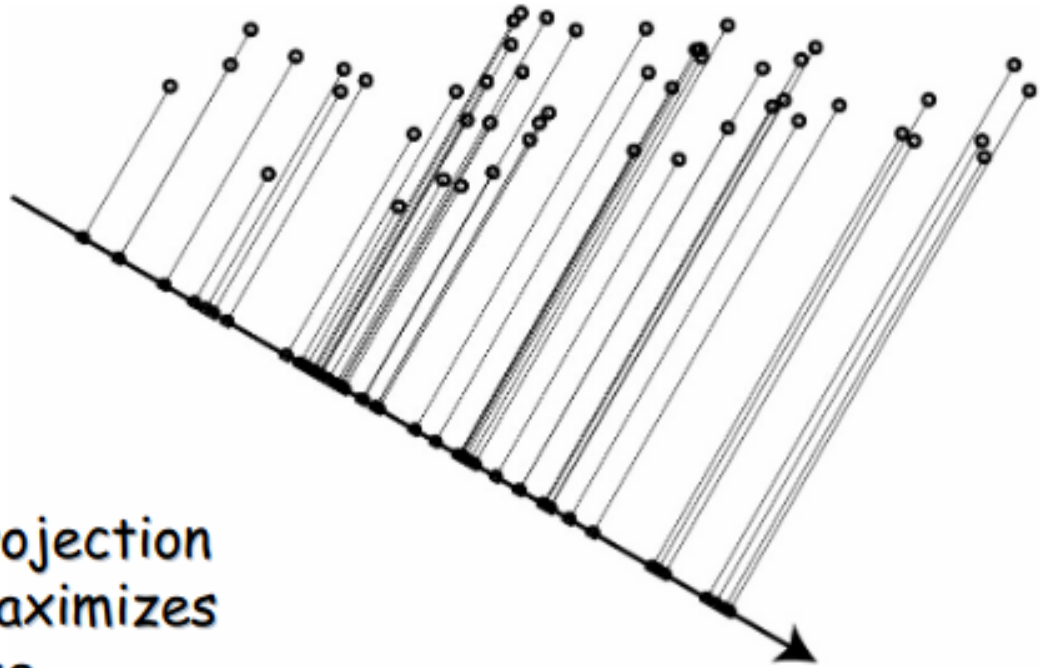


(b)



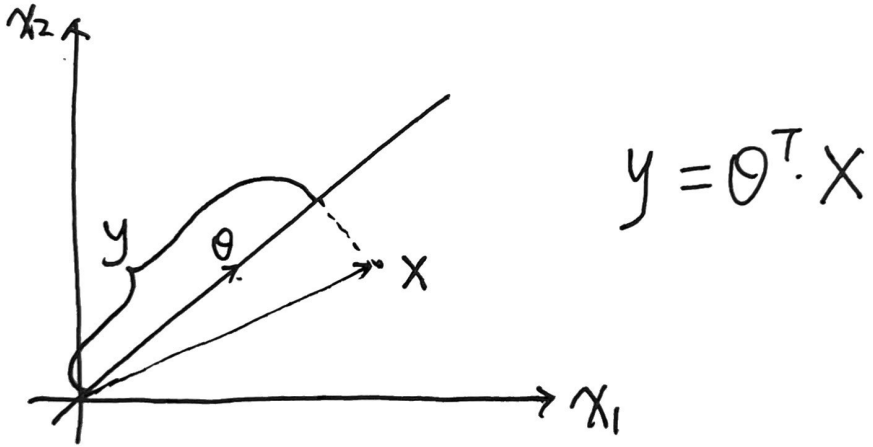
(c)

recall: Variance along Projection

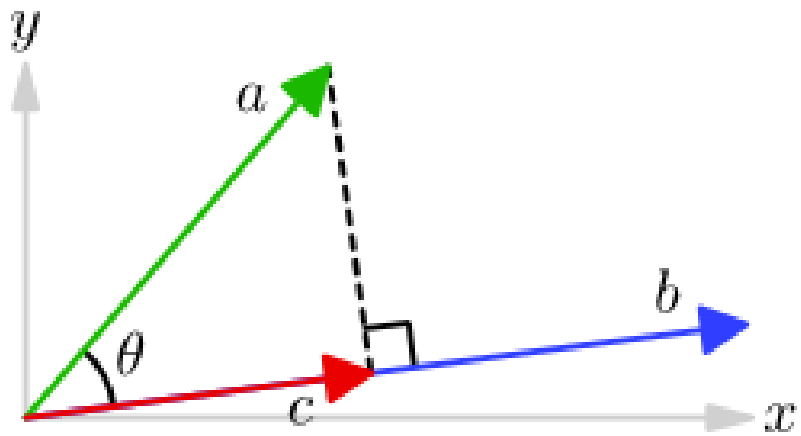


recall: Projection

Projection of a pint x along line with projection weights θ is given by :



recall: Projection



$$a.b = |a||b|cos(\theta)$$

$$a.b = a_x b_x + a_y b_y$$

$$|c| = |a|cos(\theta)$$

Exercise 1

Given two vectors, $x = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, $y = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$, find the projection of y onto x.

Solution Exercise 1

approach 1:

$$\cos(\theta) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|x| \cdot |y|}$$

$$|y| \cdot \cos(\theta) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|x|}$$

$$= \frac{(3,4) \cdot (5,5)}{\sqrt{3^2 + 4^2}}$$

$$= \frac{3 \times 5 + 4 \times 5}{5} = 3 + 4 = 7$$

Solution Exercise 1

approach 2:

The project of vector $\boldsymbol{y}$ onto $\boldsymbol{x}$ is the inner product of the verctor $\boldsymbol{y}$ and the unit vector of $\boldsymbol{x}$.

The unit vector of $\boldsymbol{x}$ is

$$\boldsymbol{u}_x = \frac{\boldsymbol{x}}{|x|} = \frac{(3,4)}{\sqrt{3^2+4^2}} = (\frac{3}{5}, \frac{4}{5})$$

The projection is

$$\boldsymbol{y} \cdot \boldsymbol{u}_x = (5,5) \cdot (\frac{3}{5}, \frac{4}{5}) = 3+4 = 7$$

1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

Optimization Problem

- Consider a dataset of observation $X = [x_1, \cdots, x_i, \cdots, x_N]$ where $i = 1, 2, ..., N$, and x_i is a variable with dimensionality D.
- Our **goal** is to project the data onto a space having dimensionality $M < D$ while maximizing the variance of the projected data.
- Let's consider the projection onto a one-dimensional space ($M = 1$). We can define the direction of the space with a vector a which has D dimensions.

PCA

There are two commonly used definitions of PCA that give rise to the same algorithm *(Bishop, PRML)*.

1.

PCA can be defined as the orthogonal projection of the data onto a lower dimensional linear space, known as the principal subspace, such that **the variance of the projected data is maximized** (Hotelling, 1933).(*This lecture will focus on this definition*)

2.

Equivalently, it can be defined as the linear projection that **minimizes the average projection cost**, defined as the mean squared distance between the data points and their projections (Pearson, 1901)

Optimization Problem

- The **mean** of the projected data is $a^T\overline{x}$, where $\overline{x}$ is the sample set mean given by

$$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1}$$

- and the **variance** of the **projected data** is given by

variance function

$$J_v = \frac{1}{N}\sum_{n=1}^{N}(a^T x_n - a^T\overline{x})^2 \tag{2}$$

Optimization Problem

variance function

$$J_v = \frac{1}{N}\sum_{n=1}^{N}(a^T x_n - a^T \overline{x})^2 = a^T S a \tag{3}$$

- where S is the data covariance matrix defined by

$$S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \overline{x})(x_n - \overline{x})^T = \frac{1}{N} X_c X_c^T \tag{4}$$

where X_c is a centered version of the $N \times D$ design matrix.

Optimization Problem

The variance is a function of both a and S

Maximizing variance along a is not well-defined since we can increase it without limit by increasing the size of the components of a. Impose a normalization constraint on the a vectors such that

$$a^T a = 1 \tag{5}$$

Also, we are only interested in the direction of a instead of its magnitude.

Optimization problem

To solve this optimization problem eq.(2) with constraints eq.(4), we resort to introduce a Lagrange multiplier:

Optimization problem is to maximize

$$J = a^T Sa - \lambda(a^T a - 1)$$

where is λ is a lagrange multiplier.

Optimization problem

Solution: Differentiating w.r.t. a yields

$$\frac{\partial J}{\partial a} = 2Sa - 2\lambda a = 0$$

which reduces to

$$(S - \lambda I)a = 0$$
$$Sa = \lambda a$$

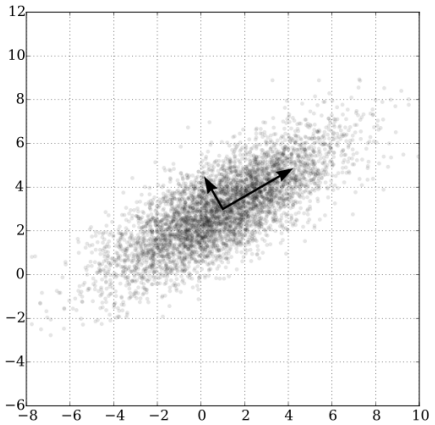
This is an eigen problem

- Hence, it follows that the best one-dimensional estimate (in a least-squares sense) for the data is the eigenvector corresponding to the largest eigenvalue of S.

Principal Components

This is an eigen problem

- Hence, it follows that the best one-dimensional estimate (in a least-squares sense) for the data is the eigenvector corresponding to the largest eigenvalue of S.
- So, we will project the data onto the largest eigenvector of S and translate it to pass through the mean.
- Second Principle component is in direction orthogonal to the first, which corresponds to second largest eigenvalue.



Principal Components

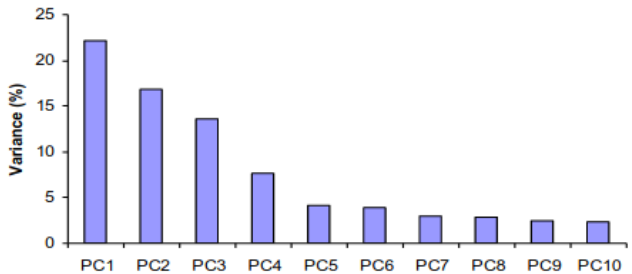
Q: Are the principal components (eigen vectors) always orthogonal? Why?

How Many PCs?

- We can ignore the components of lesser significance. You do lose some information, but if the eigenvalues are small, you don't lose much.
- The percentage of variance for each eigenvector can be represented by

$$\frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{6}$$

where λ_i is the i-th eigenvalue, and i denotes the numbering of the eigenvector/eigenvalue.



1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

Example digits

(from Bishop, PRML)

Figure 12.1 A synthetic data set obtained by taking one of the off-line digit images and creating multiple copies in each of which the digit has undergone a random displacement and rotation within some larger image field. The resulting images each have $100 \times 100 = 10,000$ pixels.

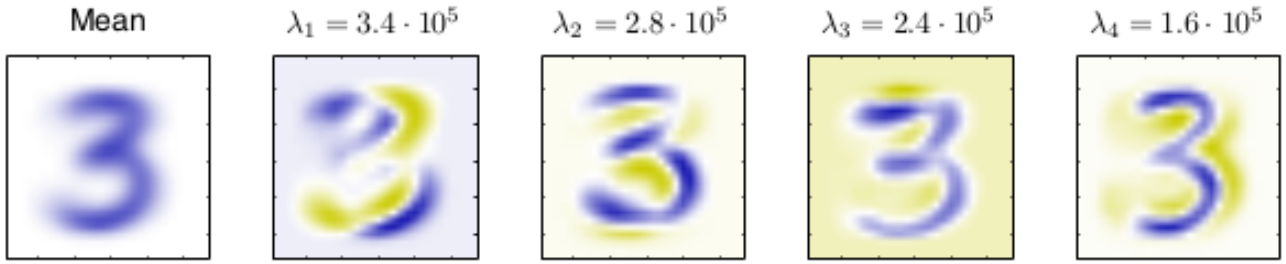


Figure 12.3 The mean vector $\overline{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.

Example digits

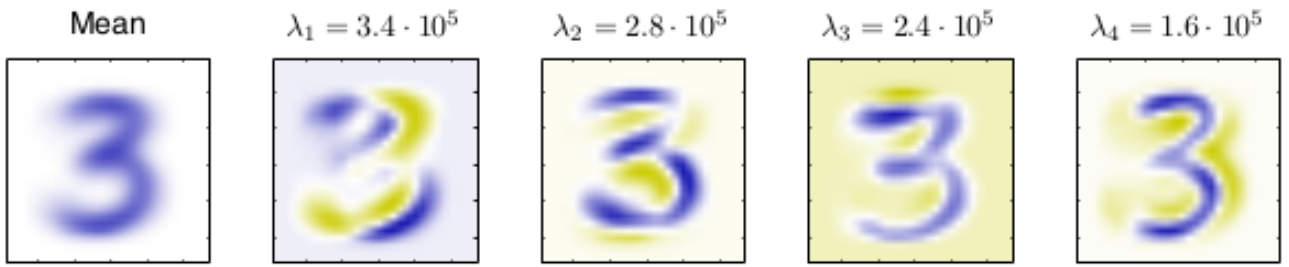


Figure 12.3 The mean vector $\overline{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.

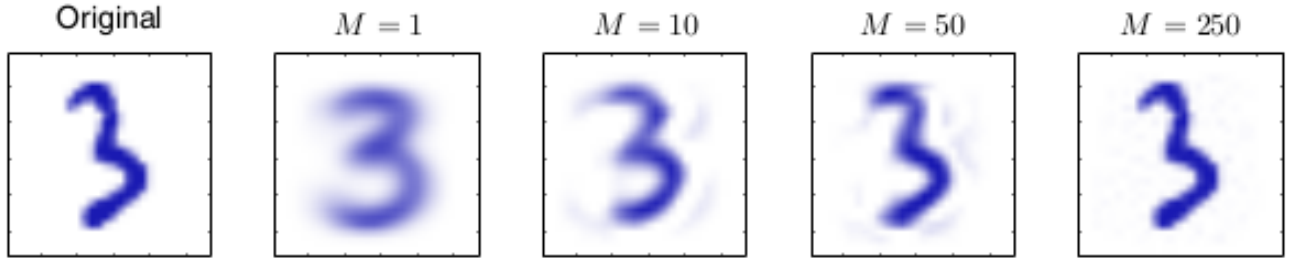


Figure 12.5 An original example from the off-line digits data set together with its PCA reconstructions obtained by retaining M principal components for various values of M. As M increases the reconstruction becomes more accurate and would become perfect when $M = D = 28 \times 28 = 784$.

Example Iris

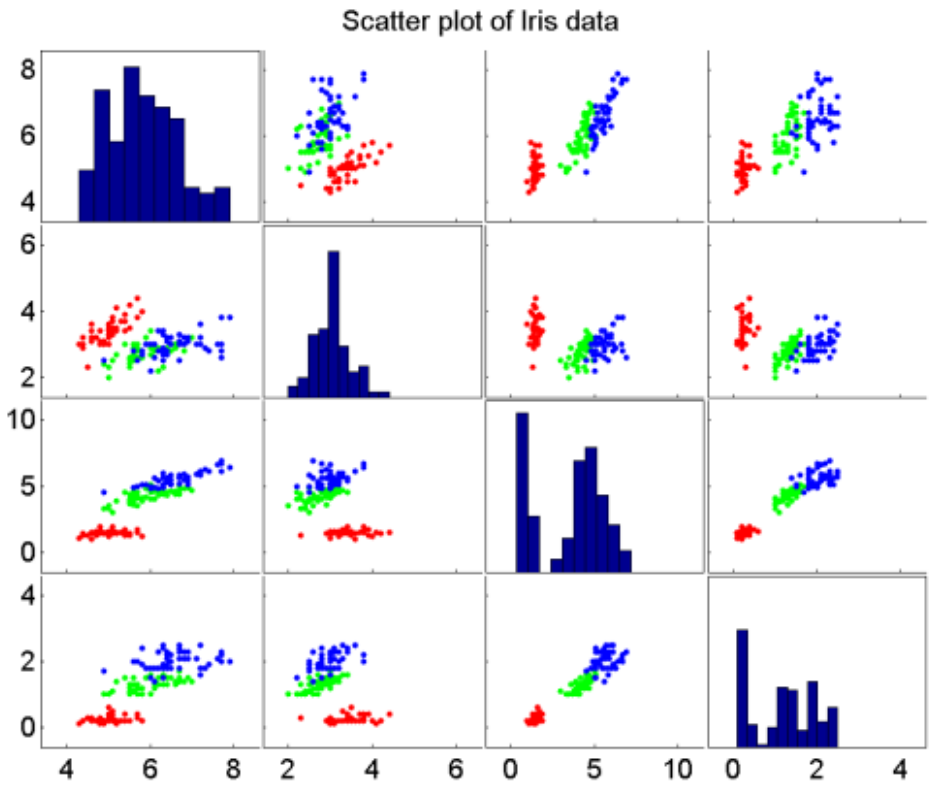


Figure 4: Scatter plot of the iris data. Diagonal cells show the histogram for each feature. Other cells show scatters of pairs of features x_1 , x_2 , x_3 , x_4 in top-down and left-right order. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.

Example Iris

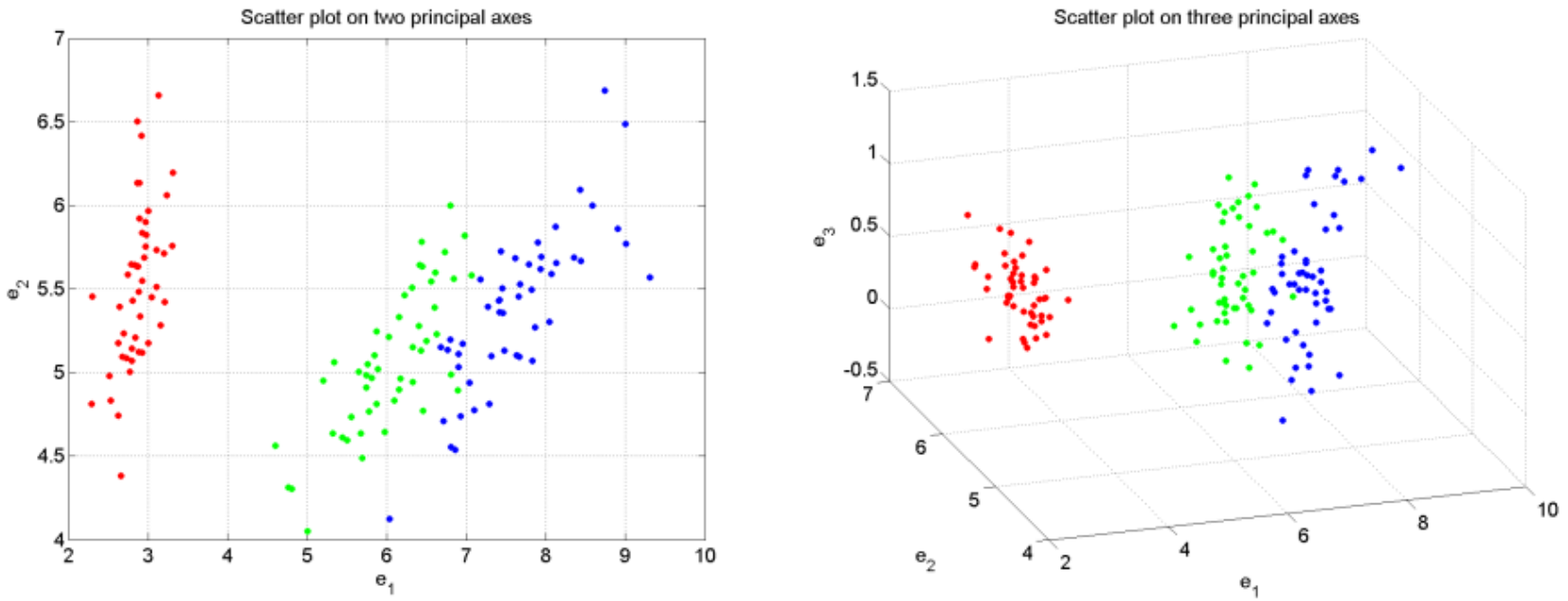
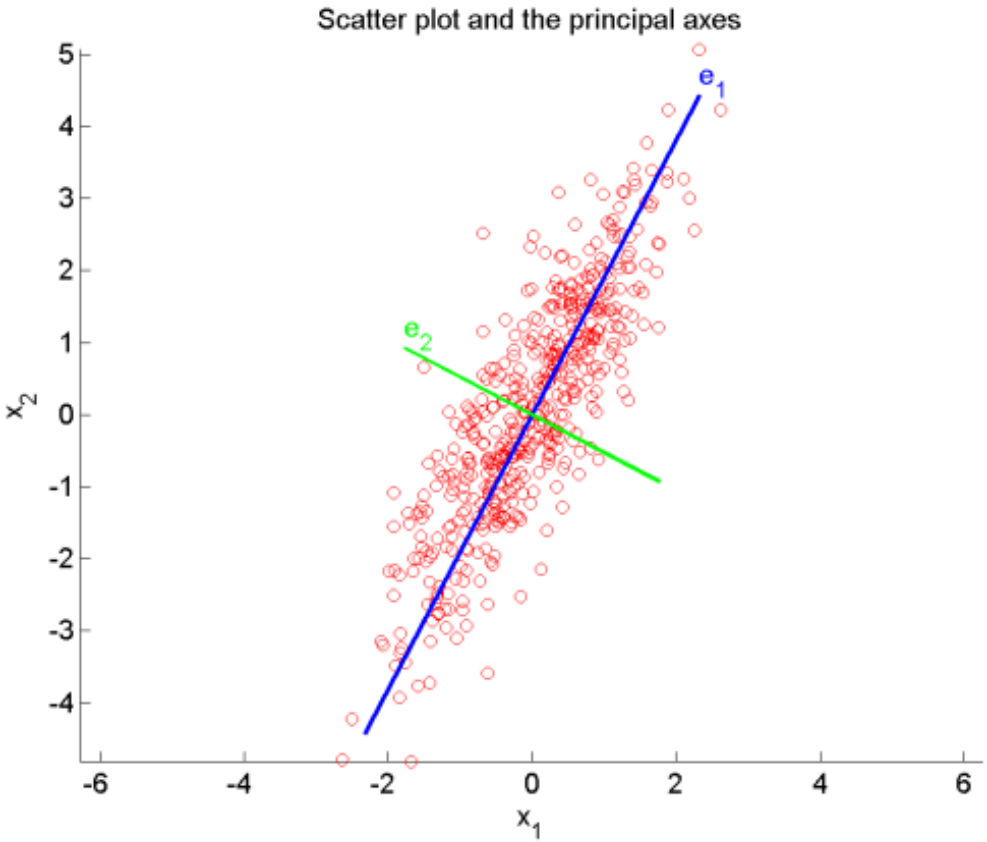
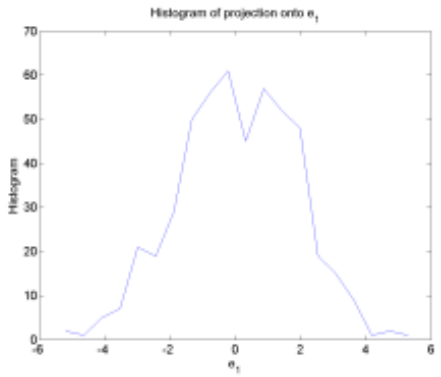


Figure 5: Scatter plot of the projection of the iris data onto the first two and the first three principal axes. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.

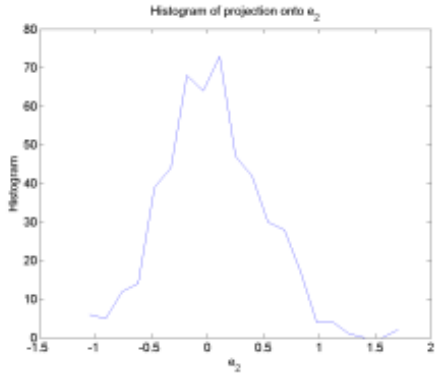
Examples



(a) Scatter plot.



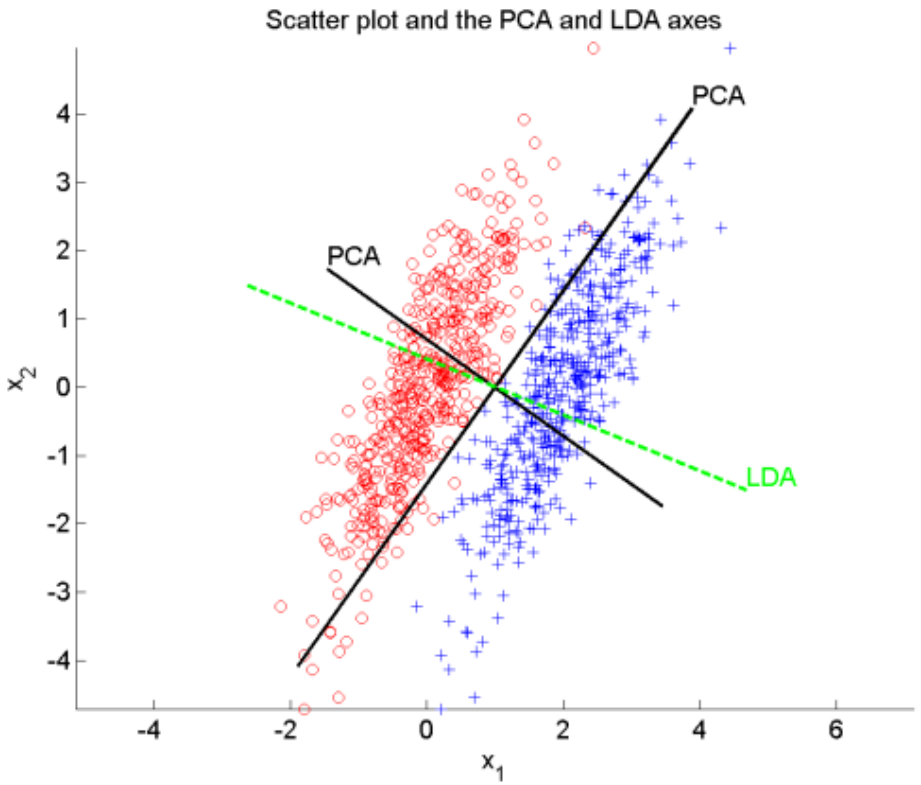
(b) Projection onto e_1.



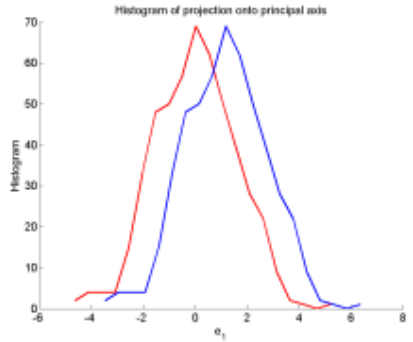
(c) Projection onto e_2.

Figure 6: Scatter plot (red dots) and the principal axes for a bivariate sample. The blue line shows the axis e_1 with the greatest variance and the green line shows the axis e_2 with the smallest variance. Features are now uncorrelated.

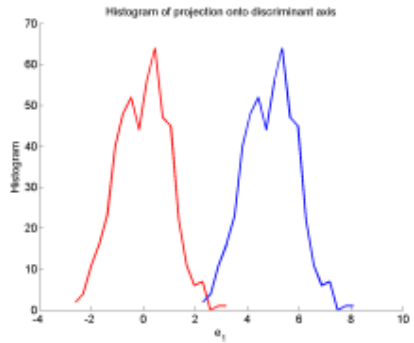
Examples



(a) Scatter plot.



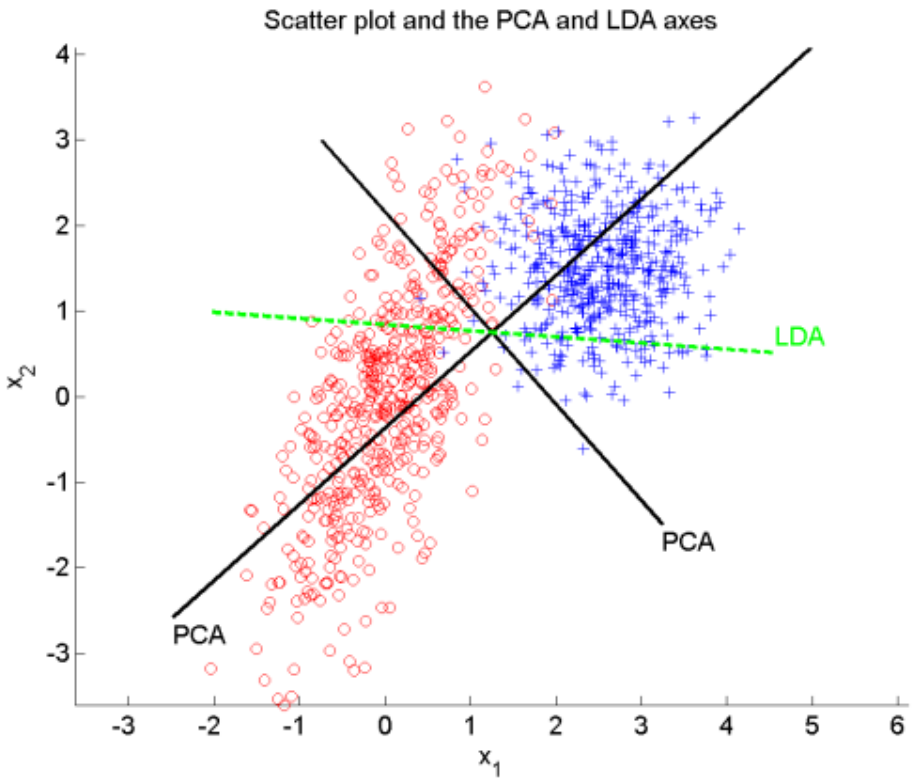
(b) Projection onto the first PCA axis.



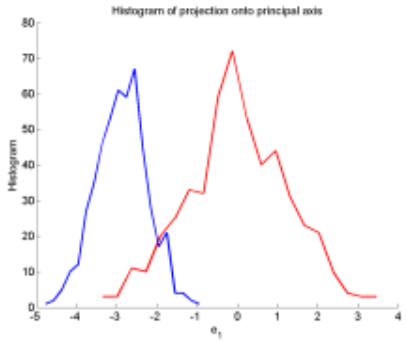
(c) Projection onto the first LDA axis.

Figure 7: Scatter plot and the PCA and LDA axes for a bivariate sample with two classes. Histogram of the projection onto the first LDA axis shows better separation than the projection onto the first PCA axis.

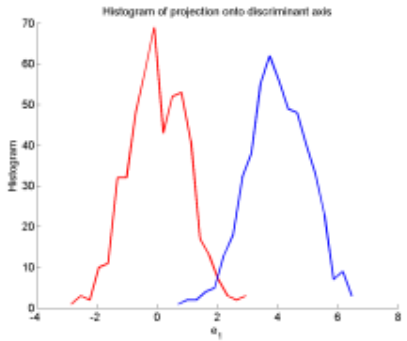
Examples



(a) Scatter plot.



(b) Projection onto the first PCA axis.



(c) Projection onto the first LDA axis.

Figure 8: Scatter plot and the PCA and LDA axes for a bivariate sample with two classes. Histogram of the projection onto the first LDA axis shows better separation than the projection onto the first PCA axis.

Examples satelite image

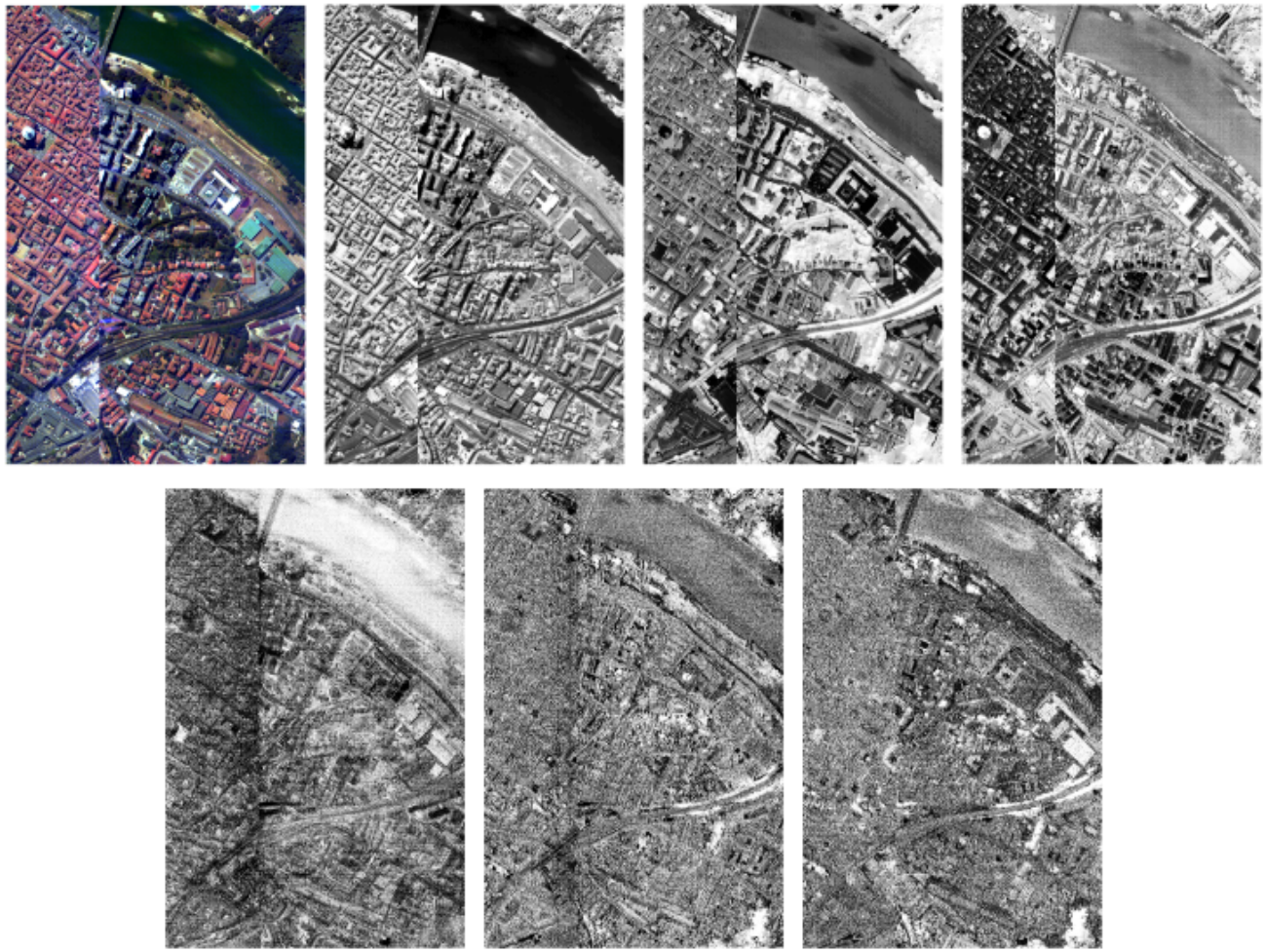


Figure 9: A satellite image and the first six PCA bands (after projection). Histogram equalization was applied to all images for better visualization.

Examples satelite image

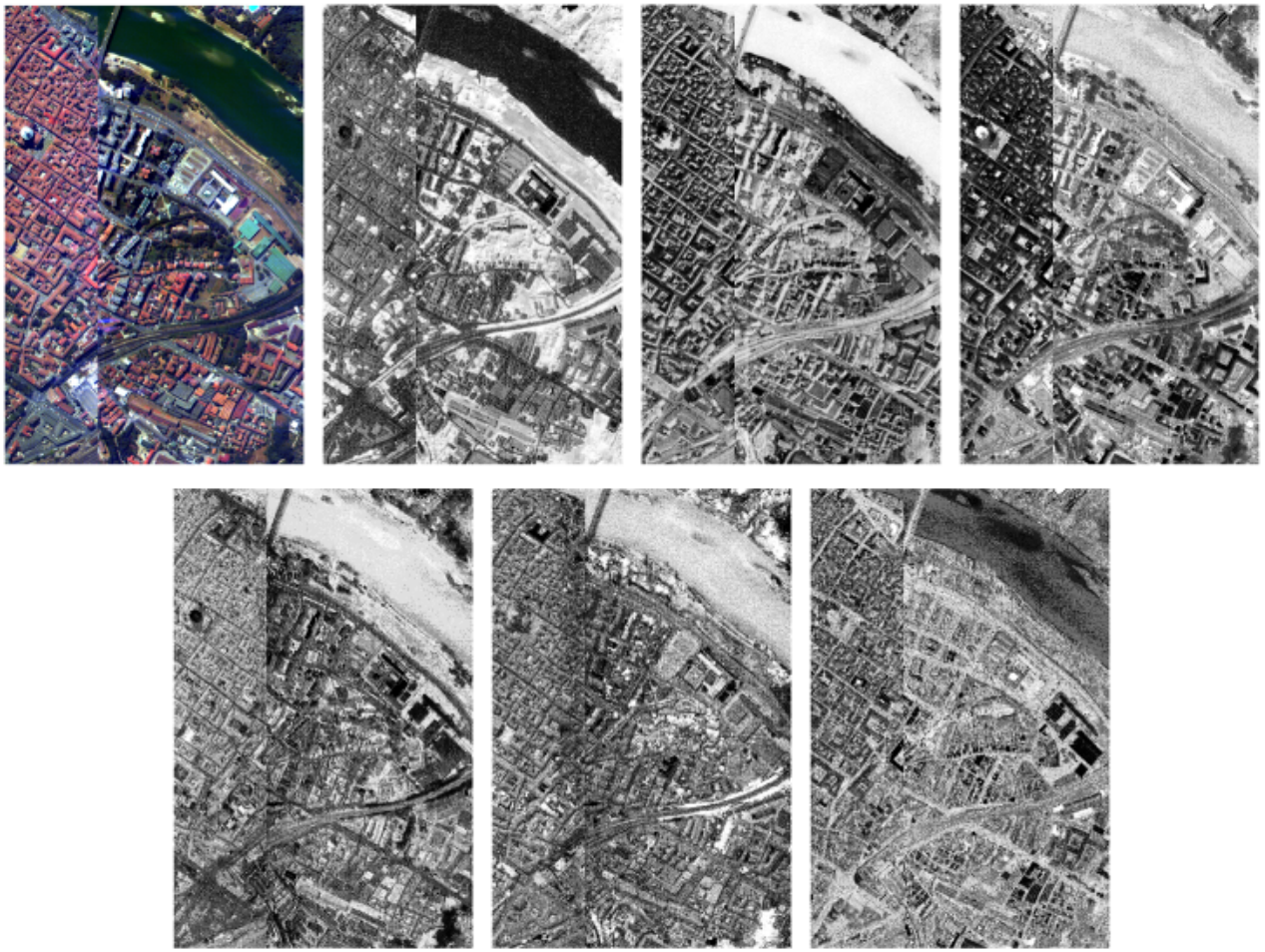


Figure 10: A satellite image and the six LDA bands (after projection). Histogram equalization was applied to all images for better visualization.

Examples eigenfaces

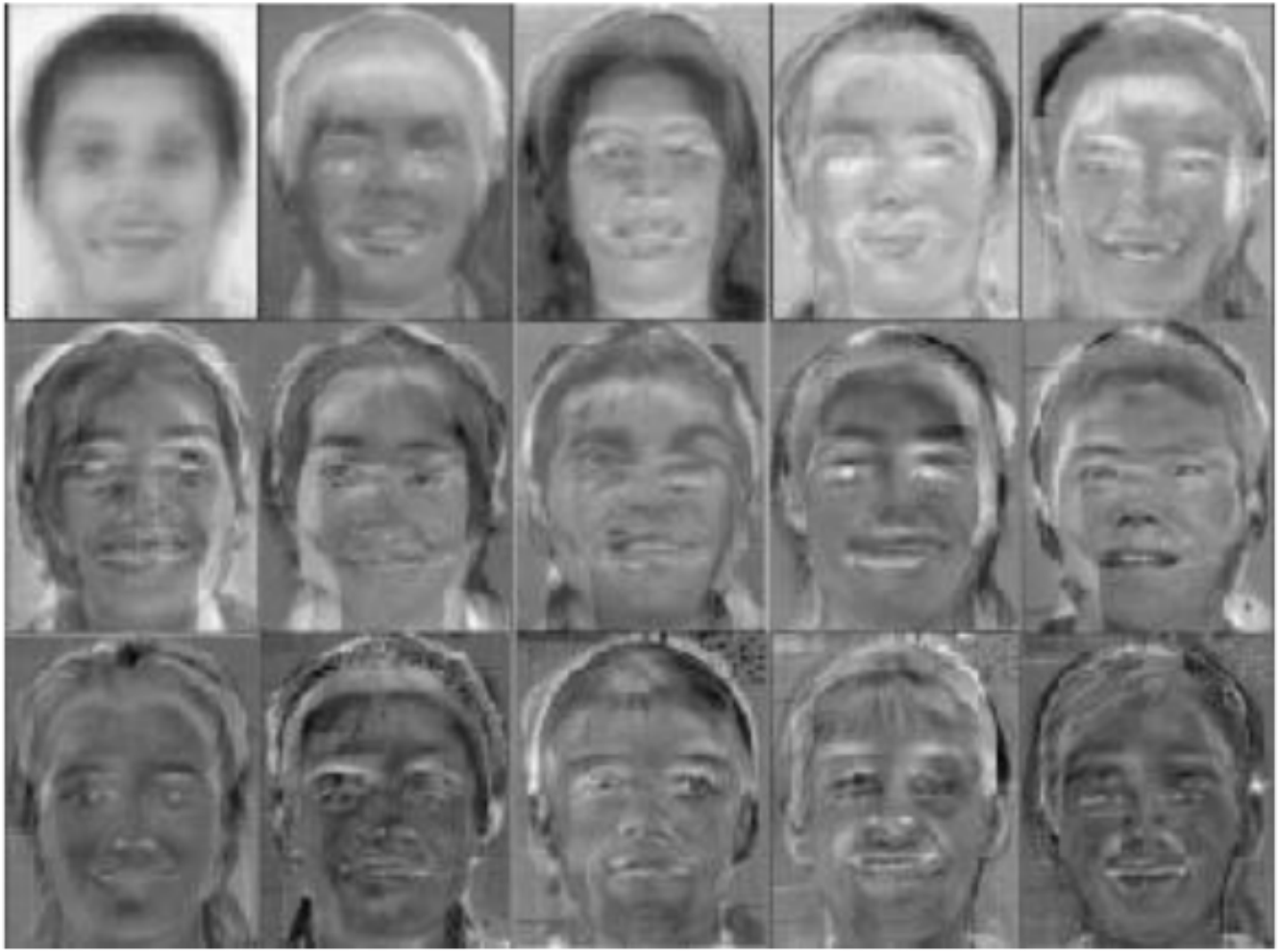


Figure 11: Eigenvectors (principal axes) of the face images (often referred to as eigenfaces).

- **Principal Components Analysis (PCA)**: Seeks a projection that best represents the data in a least-squares sense.
- **Linear Discriminant Analysis (LDA)**: Seeks a projection that best separates the data in a least-squares sense.

Outline

1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

Exercise 2

Find the eigenvalues and eigenvectors of $A = \begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix}$.

Solution to Exercise 2

with $A = \begin{bmatrix} 0 & 3 \\ 1 & 2 \end{bmatrix}$

To find the eigenvalues and eigenvectors $\Rightarrow A\boldsymbol{x} = \lambda\boldsymbol{x}$

$\Rightarrow (A - \lambda I)\boldsymbol{x} = 0$

$\Rightarrow |A - \lambda I| = 0 \Rightarrow \begin{vmatrix} -\lambda & 3 \\ 1 & 2-\lambda \end{vmatrix} = 0$

$\Rightarrow \lambda^2 - 2\lambda - 3 = 0$

$\Rightarrow \lambda = 3, -1$

When $\lambda = 3$, $A - \lambda I = \begin{bmatrix} -3 & 3 \\ 1 & -1 \end{bmatrix}$, then $(A - \lambda I)\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

When $\lambda = -1$, $A - \lambda I = \begin{bmatrix} 1 & 3 \\ 1 & 3 \end{bmatrix}$, then $(A - \lambda I)\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$

Exercise 3

Given a data matrix $X = \begin{bmatrix} 2 & 0 & 3 & -1 \\ 0 & -2 & -3 & 1 \end{bmatrix}$, compute the variance after projecting the dataset onto its first principal component.

Solution Exercise 3

1. $X = \begin{bmatrix} 2 & 0 & 3 & -1 \\ 0 & -2 & -3 & 1 \end{bmatrix}$, the mean of the sample is $\overline{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$
2. The centered version of the data matrix $X_c = \begin{bmatrix} 1 & -1 & 2 & -2 \\ 1 & -1 & -2 & 2 \end{bmatrix}$
3. The data covariance matrix is $S = \frac{1}{N} X_c X_c^T = \begin{bmatrix} \frac{5}{2} & -\frac{3}{2} \\ -\frac{3}{2} & \frac{5}{2} \end{bmatrix}$
4. Then find the eigenvalues for matrix S, we can get $\lambda = 4, 1$
   (detailed steps are like what's in previous exercise)

The the variance after projecting the dataset onto its **first principal component** is 4.

Outline

1 Introduction

2 PCA
- Intuition
- Formulization
- Examples

3 Exercises

4 Feature Selection*(optional)

Importance of Feature Engineering

- Better features means flexibility.
- Better features means simpler models.
- Better features means better results.

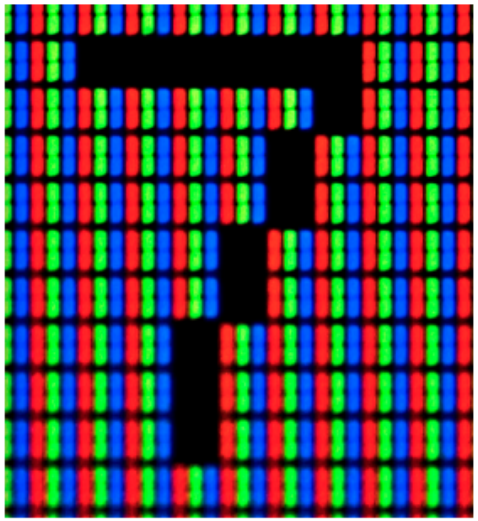
Initial features

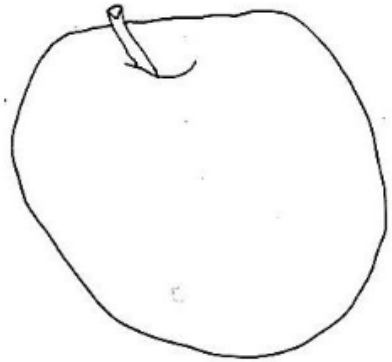
The initial pick of feature is always an expression of prior knowledge.

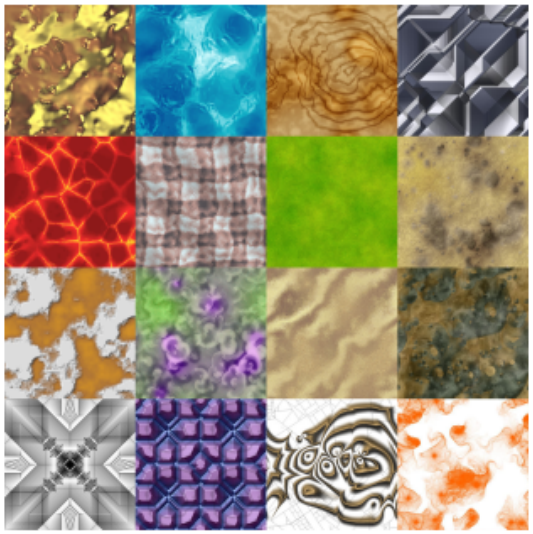
- images
- signal
- time series
- text data

Initial features

- **images** : pixels, contours, textures, etc.

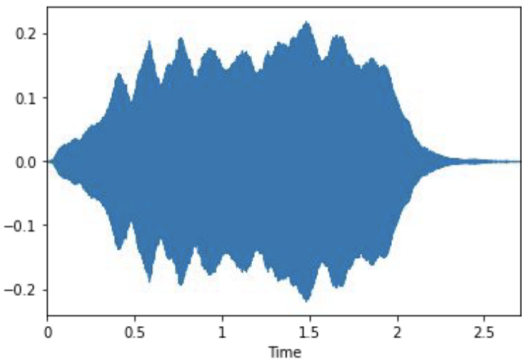


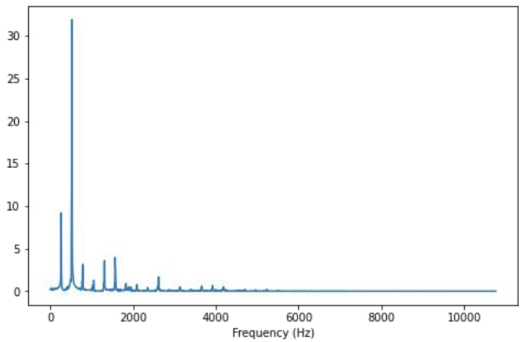


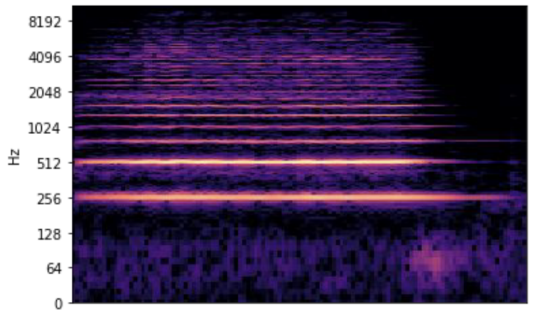


Initial features

- **signal**
  - time domain
  - frequency domain
  - time-frequency representation (e.g., spectrogram, Mel-spectrogram, Constant-Q transform), etc.

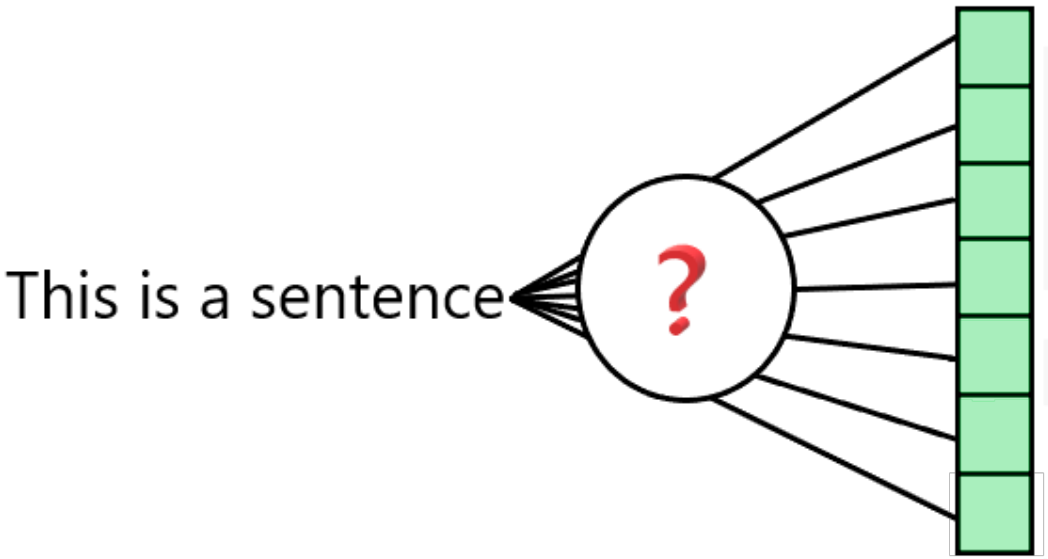






Initial features

- **text data** : Bag-of-Words, Bag-of-n-Grams, etc.



Feature Explosion

Combining features

- Combinations that linear system cannot represent: polynomial combinations, logical conjunctions, decision trees.
- Total number of features then grows very quickly.

Solutions

- Feature selection
- Feature extraction(transformation)
- Feature learning (e.g., CNN)

Selecting features from data

How to select relevant features when p(x, y) is unknown but data is available?

Selecting features from data

- **Training data is limited**
  - Restricting the number of features is a capactity control mechanism.
  - We may want to use only a subset of the relevant features.
- **Notable approaches**
  - Feature selection using regularization.
  - Feature selection using wrappers.
  - Feature selection using greedy algorithms.

Wrapper approaches

Sequential forward selection

- Assume we have chosen a learning system and algorithm.
- Navigate feature subsets by adding/removing features.
- Evaluate on the validation set.

Sequential backward selection

- Start with all features.
- Try removing each feature and measure validation set impact.
- Remove the feature that causes the least harm.
- Repeat.

Notes

- Risk of overfitting the validation set.
- Computationally expensive.
- Quite effective in practice.

Feature Selection

Sequential forward selection:

- First, the best single feature is selected.
- Then, pairs of features are formed using one of the remaining features and this best feature, and the best pair is selected.
- Next, triplets of features are formed using one of the remaining features and these two best features, and the best triplet is selected.
- This procedure continues until all or a predefined number of features are selected.

Feature Selection

Sequential backward selection:

- First, the criterion function is computed for all d features.
- Then, each feature is deleted one at a time, the criterion function is computed for all subsets with $d-1$ features, and the worst feature is discarded.
- Next, each feature among the remaining $d-1$ is deleted one at a time, and the worst feature is discarded to form a subset with $d-2$ features.
- This procedure continues until one feature or a predefined number of features are left.

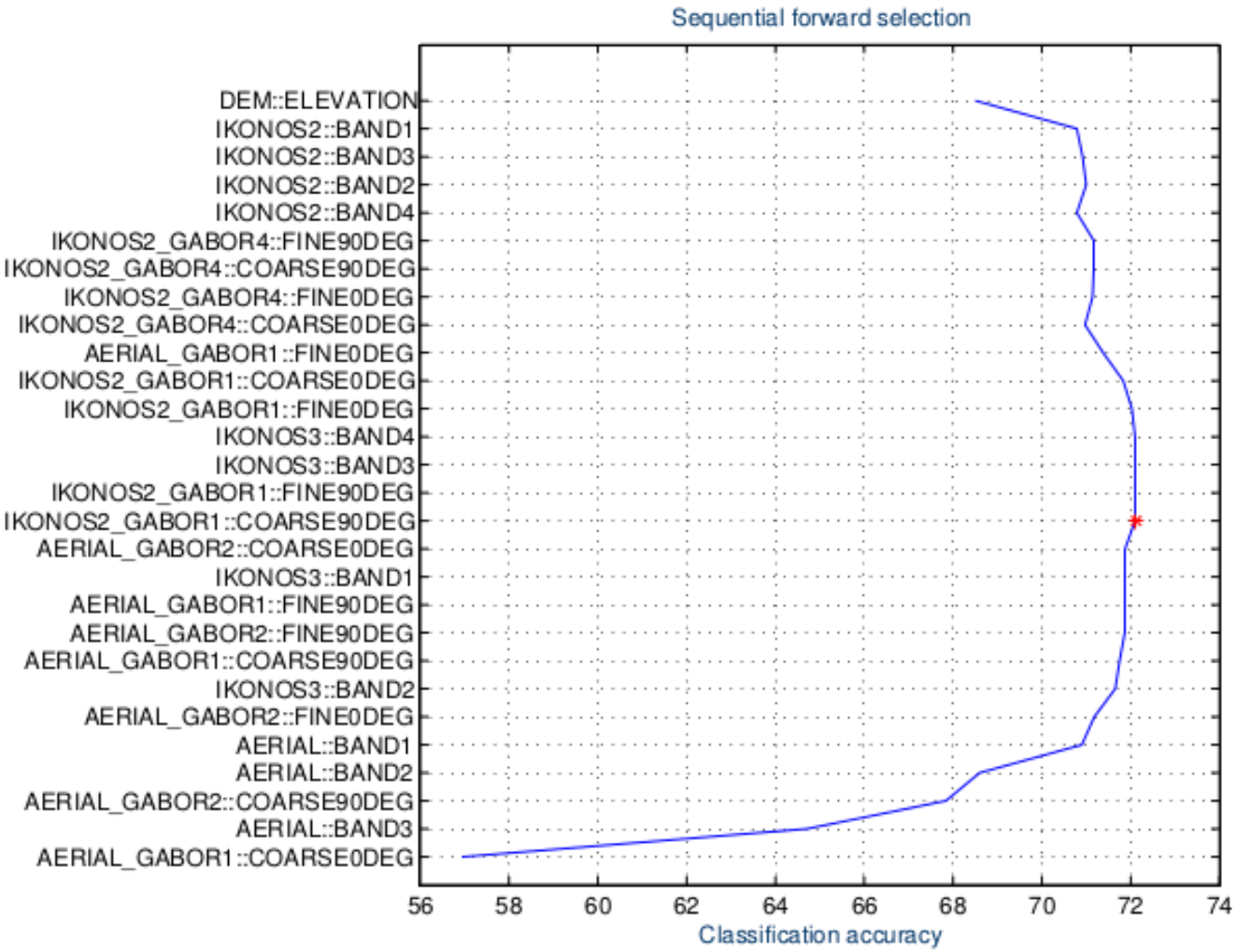


Figure 12: Results of sequential forward feature selection for classification of a satellite image using 28 features. x-axis shows the classification accuracy (%) and y-axis shows the features added at each iteration (the first iteration is at the bottom). The highest accuracy value is shown with a star.

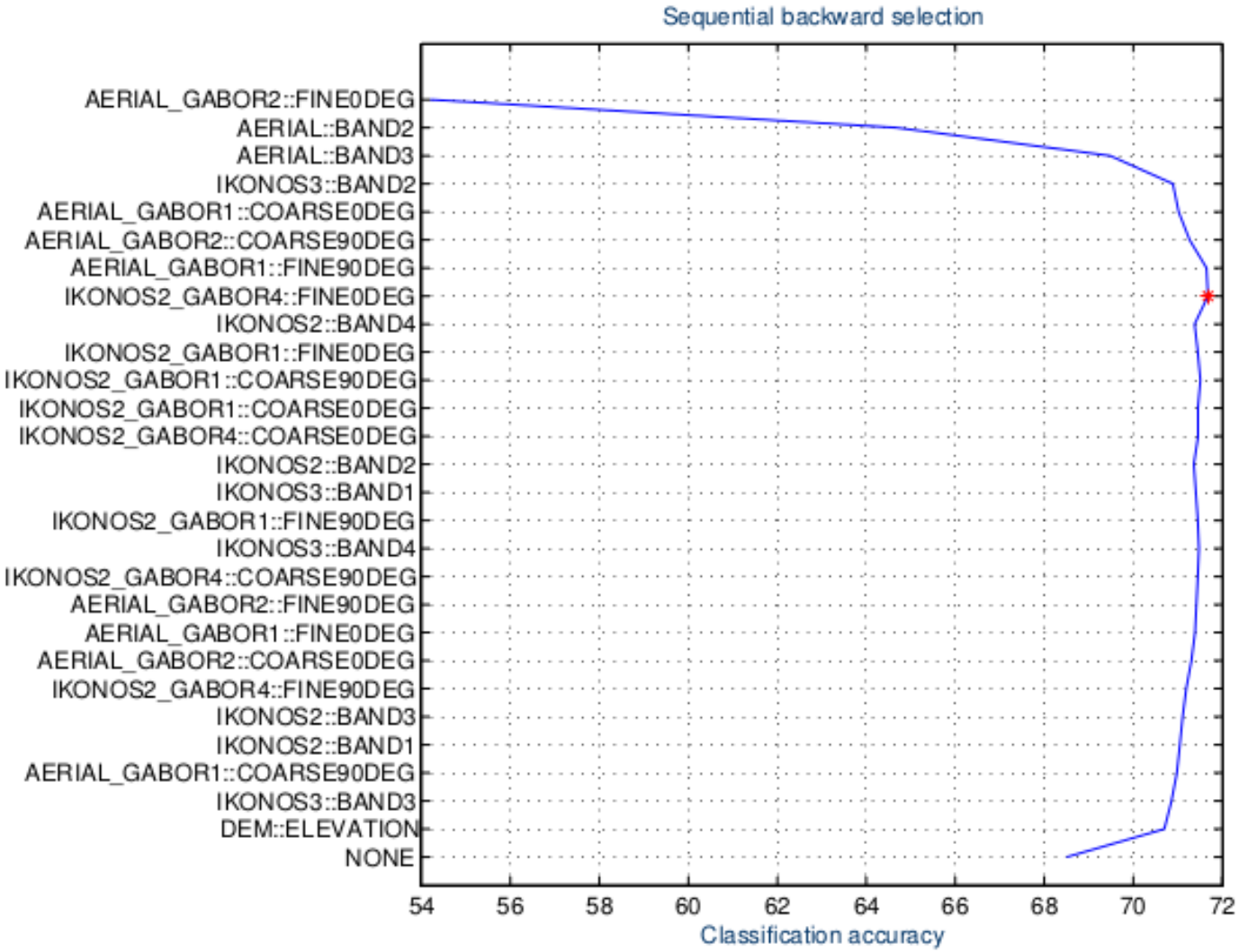


Figure 13: Results of sequential backword feature selection for classification.

Greedy methods

Algorithms that incorporate features one by one.

- Decision trees
  - Each decision can be seen as a feature.
  - Pruning the decision tree prunes the features Ensembles
- Ensembles
  - Ensembles of classifiers involving few features.
  - Random forests.
  - Boosting.

Summary

- The choice between feature reduction and feature selection depends on the application domain and the specific training data.
- Feature selection leads to savings in computational costs and the selected features retain their original physical interpretation.
- Feature reduction with transformations may provide a better discriminative ability but these new features may not have a clear physical meaning.

Thank You !

Q & A