Pattern Recognition

Lecture 18. Curse of Dimensionality and Principal Component Analysis

Dr. Shanshan ZHAO

shanshan.zhao@xjtlu.edu.cn

School of AI and Advanced Computing

Xi'an Jiaotong-Liverpool University

Academic Year 2021-2022

Table of Contents

Introduction

Problems of Dimensionality[1]

- 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

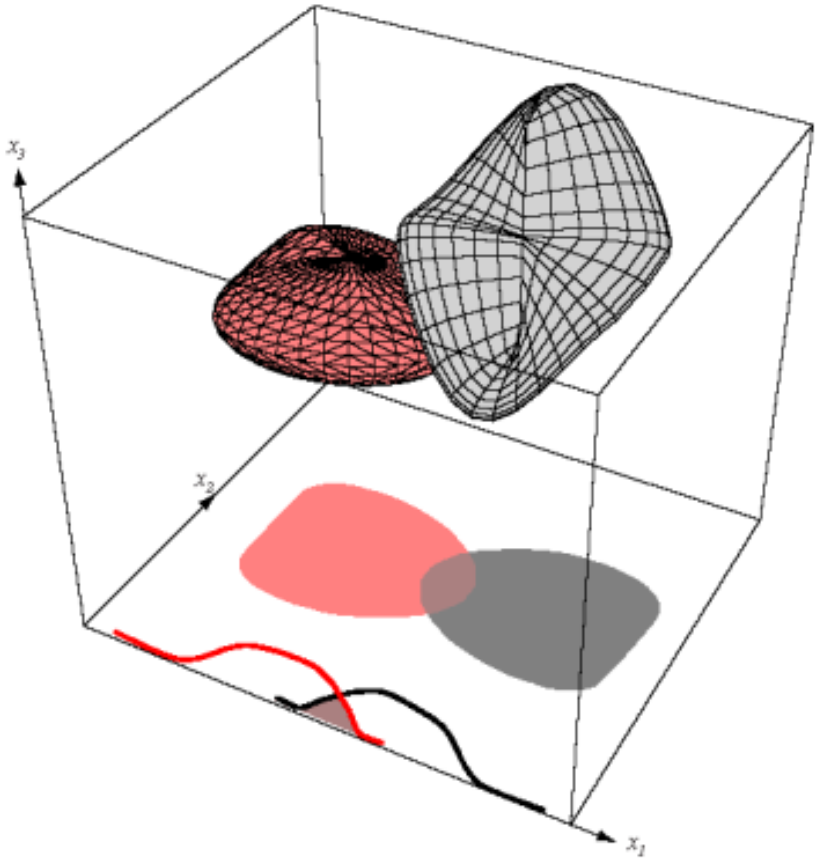


Figure: Two three-dimensional distributions have non overlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace here, the two-dimensional x_1,x_2 subspace or a one-dimensional x_1 subspace, there can be greater overlap of the projected distributions, and 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[1].
- This is the so-called **curse of dimensionality**.

Curse of Dimensionality

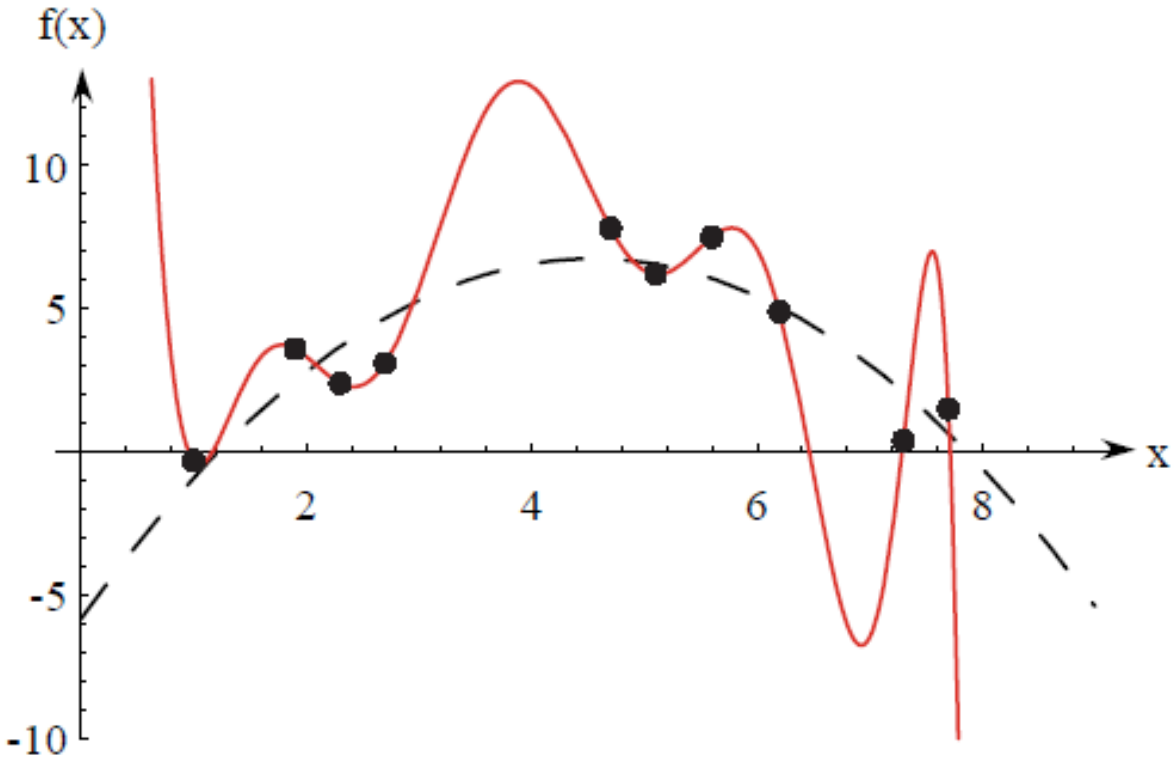


Figure: The 'training data' (black dots) were selected from a quadradic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, since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

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)

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.

PCA

There are two commonly used definitions of PCA that give rise to the same algorithm[3].

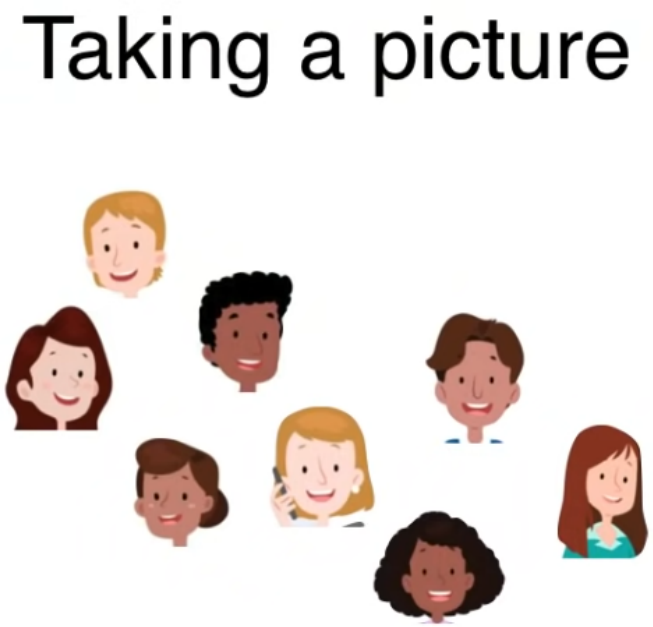
1.

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*)

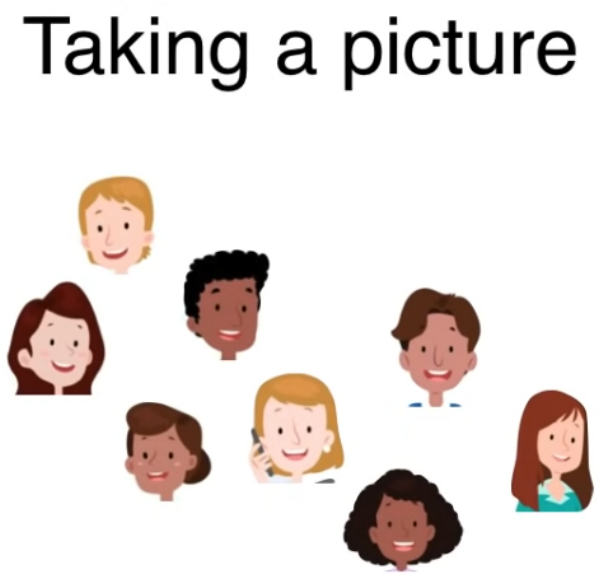
2.

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)

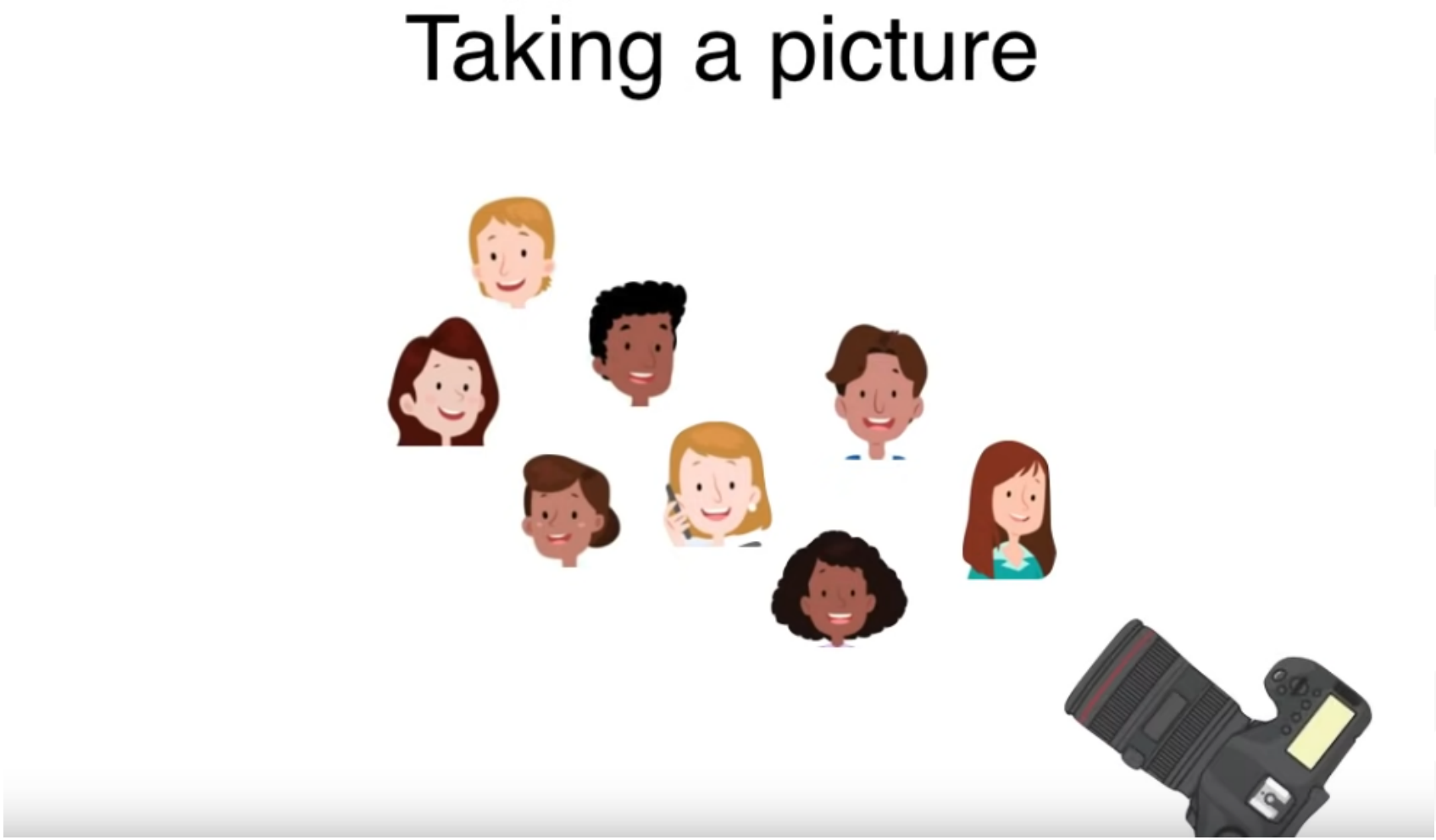
PCA Intuition



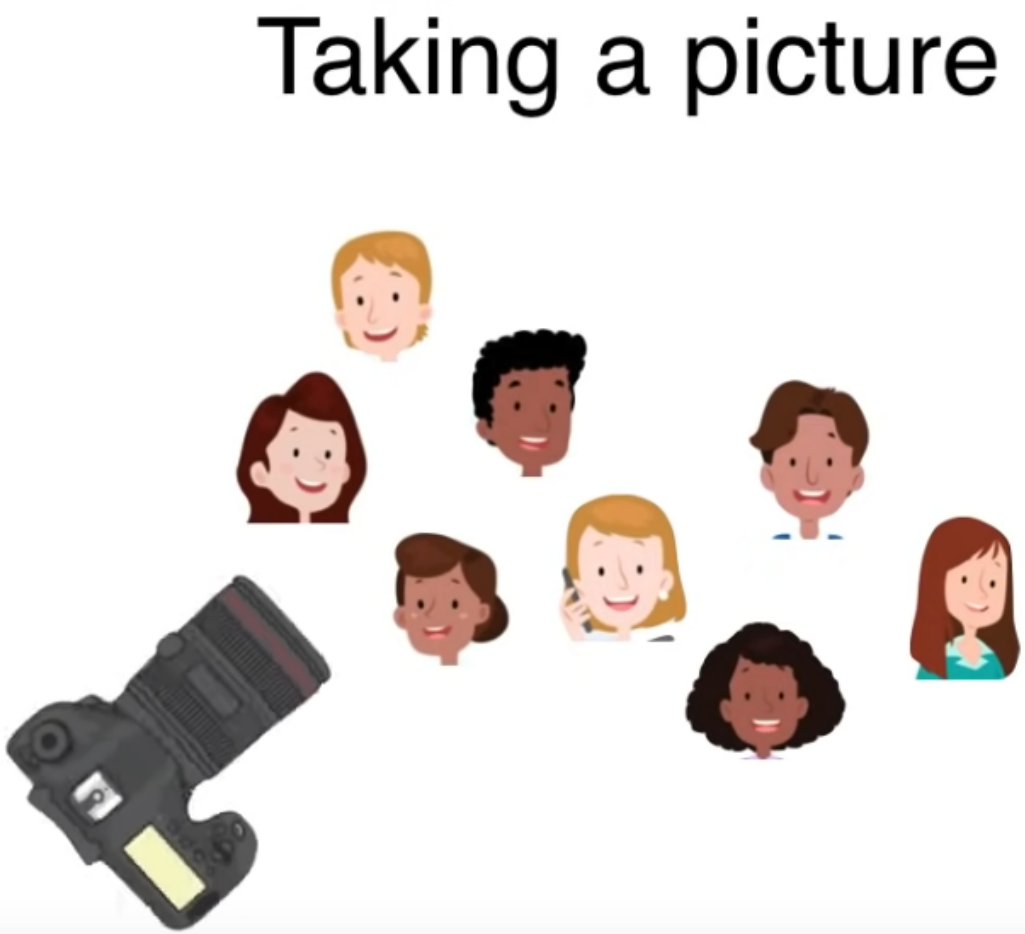
PCA Intuition



PCA Intuition



PCA Intuition



PCA Intuition

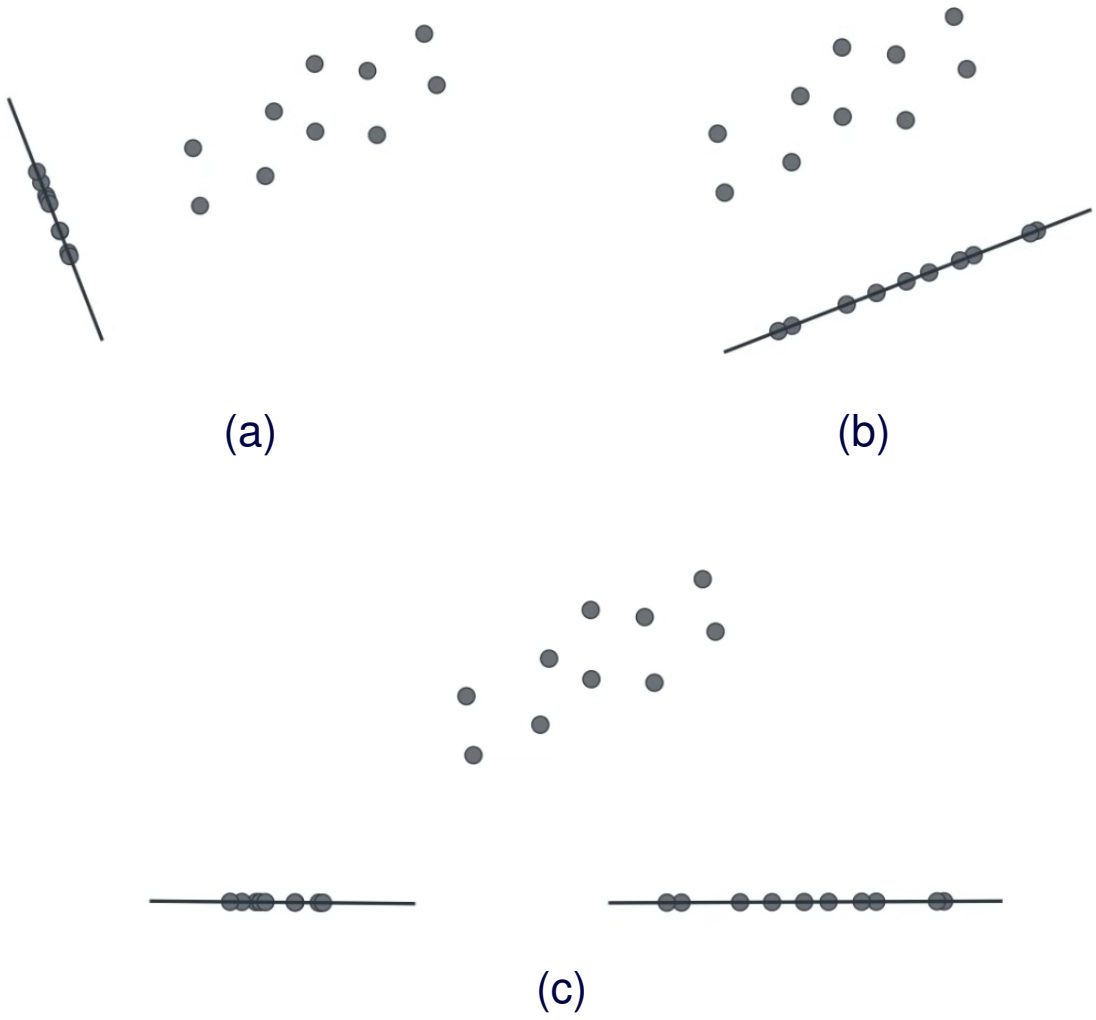
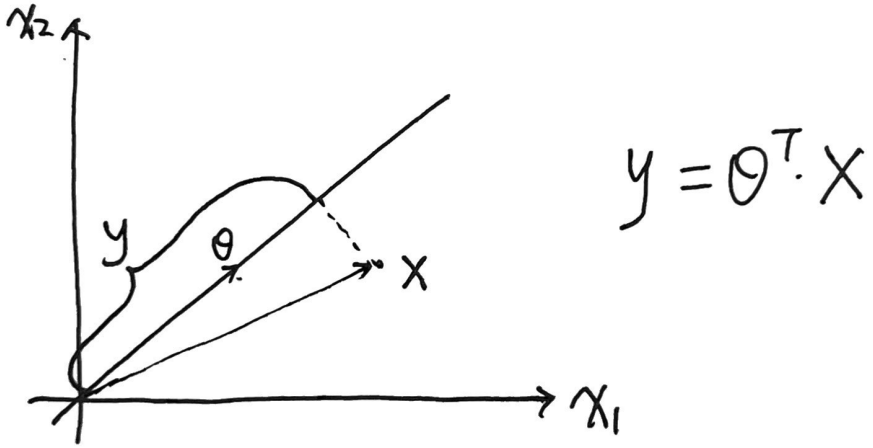


Figure: Projection to lower dimensional space

Preliminaries: Projection

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



Preliminaries: Variance along Projection

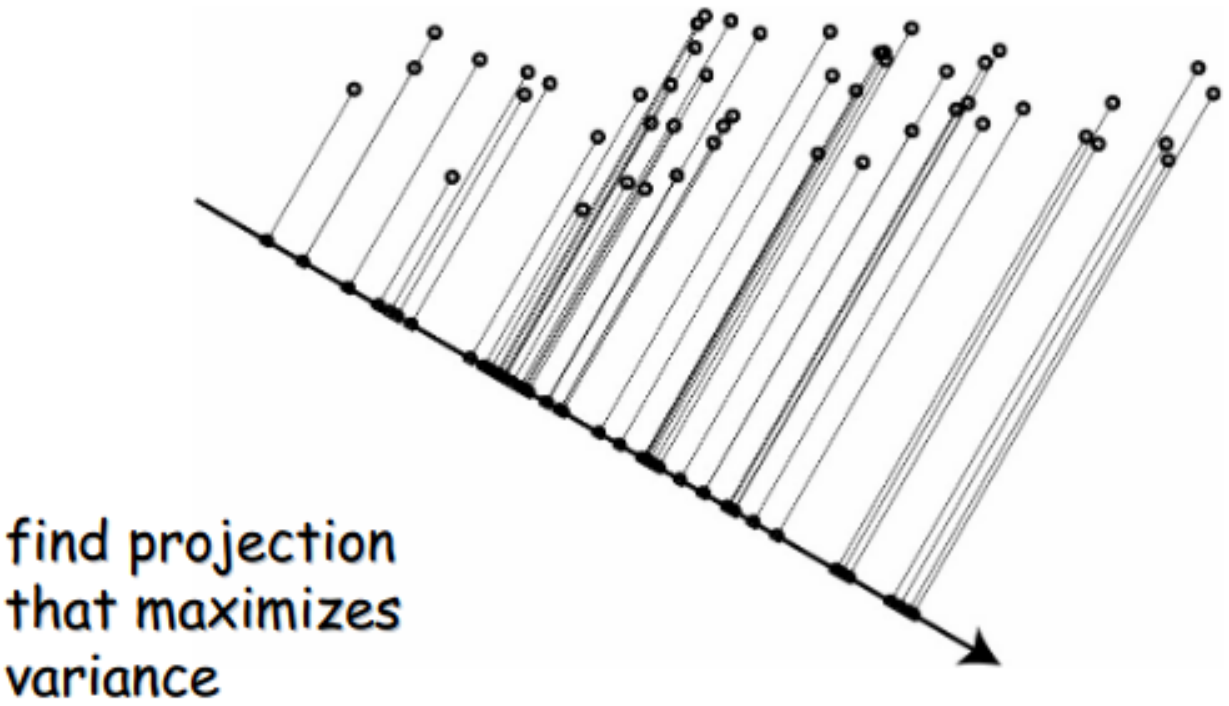


Figure: [4]

Preliminaries: Variance along Projection

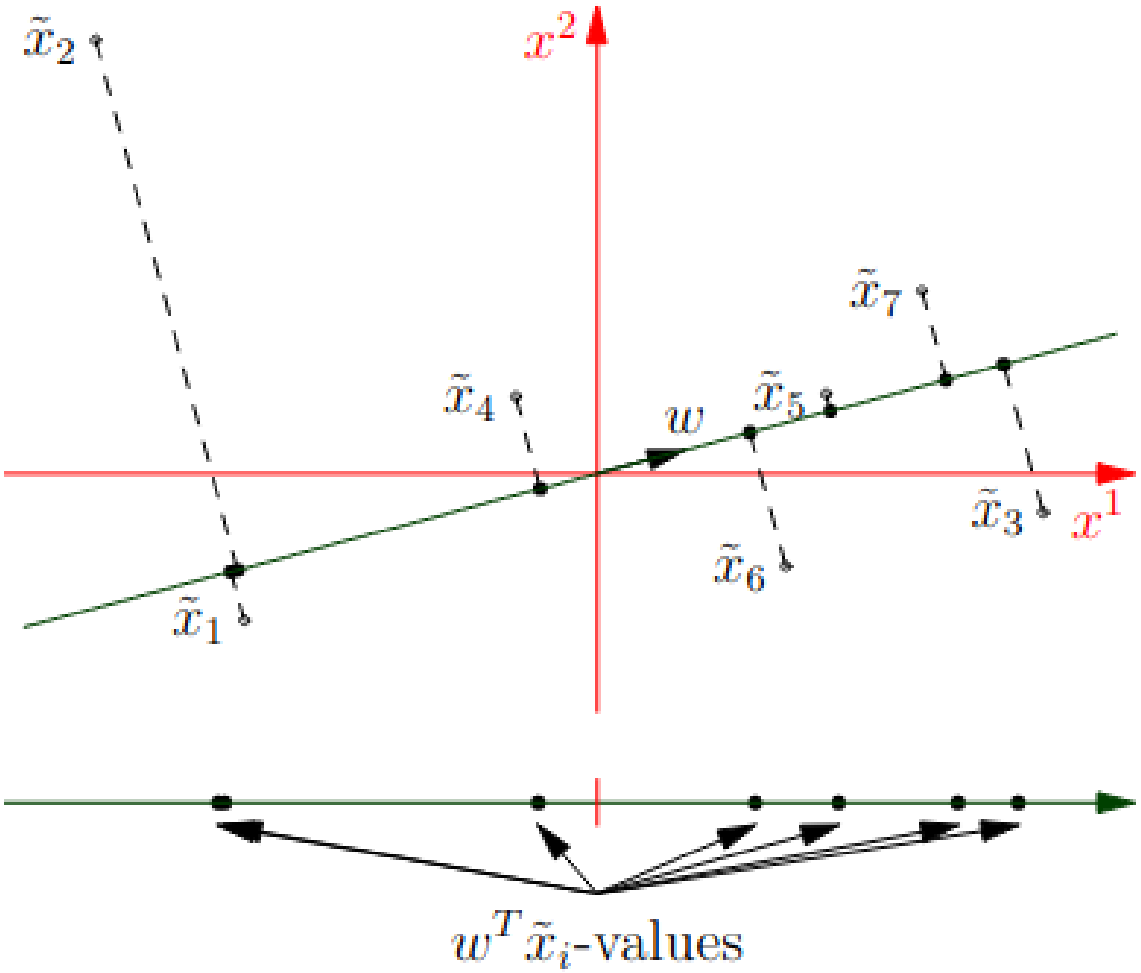


Figure: [5]

Optimization Problem

- Consider a dataset of observation X_n where $n = 1, 2, ..., N$, and x_n is a variable with dimensionality D.
- Our **goal** is to project the data onto a space having dimensionality $M < D$ while maximizing teh 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.

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

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

where S is the data covariance matrix defined by

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

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

$$a^T a = 1 \tag{4}$$

Also, we are only interested in the direction of a instead of its magnitude.
To solve this optimazaition problem eq.(2) with constraints eq.(4), we resort to introduce a Lagrange multiplier:

Optimization problem is to maximize

$$J_L = a^T S a - \lambda(a^T a - 1)$$

where is λ is a lagrange multiplier.

Optimization problem

Solution: Differentiating w.r.t. a yields

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

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 Eigen value.

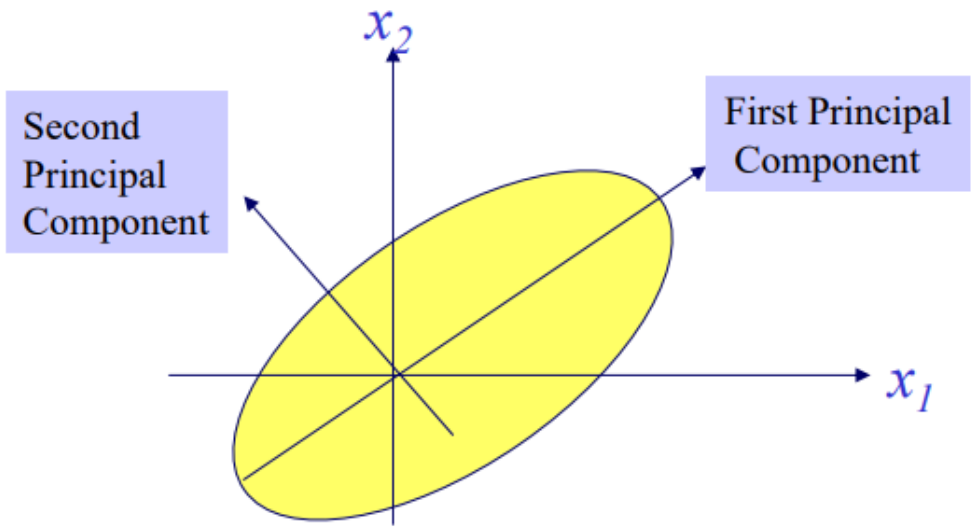


Figure: [6]

Principal Components

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

Recall

- For symmetric matrices, eigenvectors for distinct eigenvalues are orthogonal
- All eigenvalues of a real symmetric matrix are real.
- All eigenvalues of a positive semidefinite matrix are non-negative

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

$$\frac{\lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{5}$$

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

- Usually 5-10 principal components capture 90% of variance in the data.

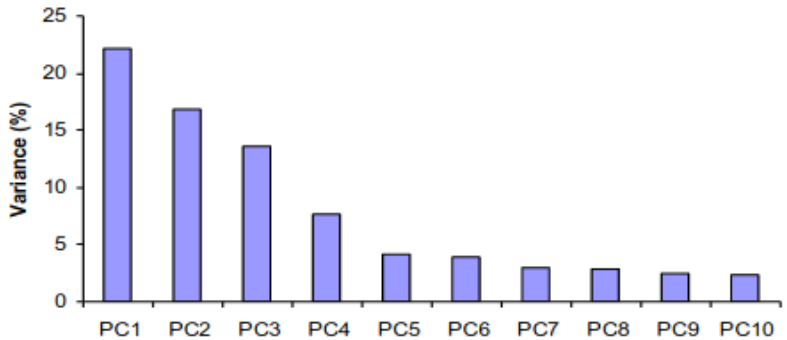


Figure: [6]

Example Iris

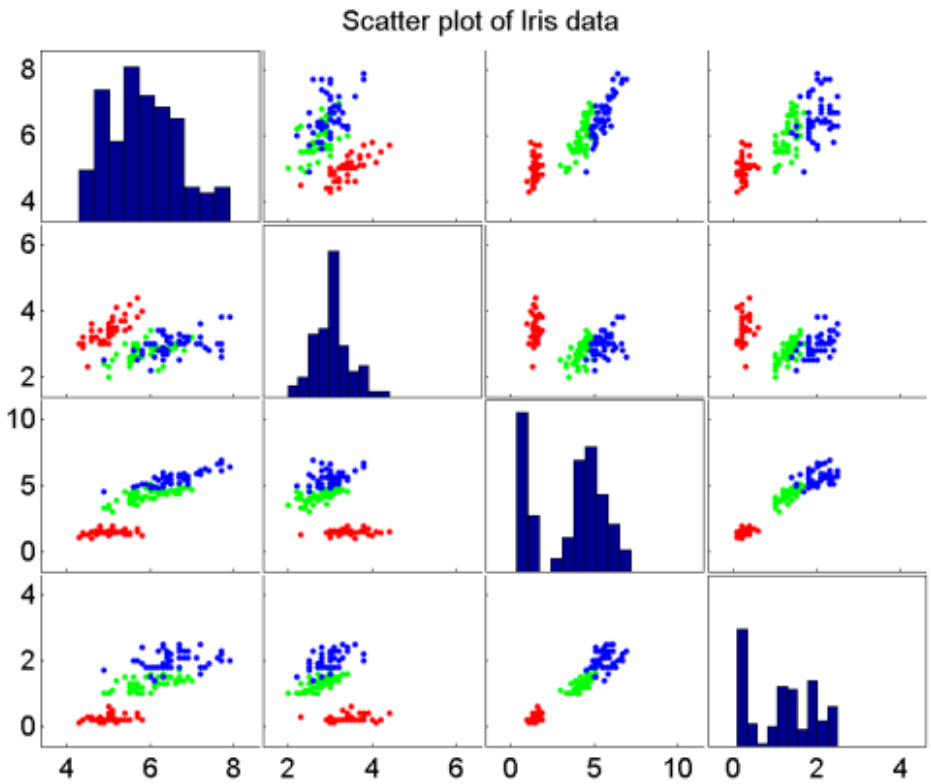


Figure: Scatter plot of the iris data. Diagonal cells show the histogram for each feature. Other cells show scatters of pairs of features x_1 , x_2 , x_3 , x_4 in top-down and left-right order. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.[2]

Example Iris

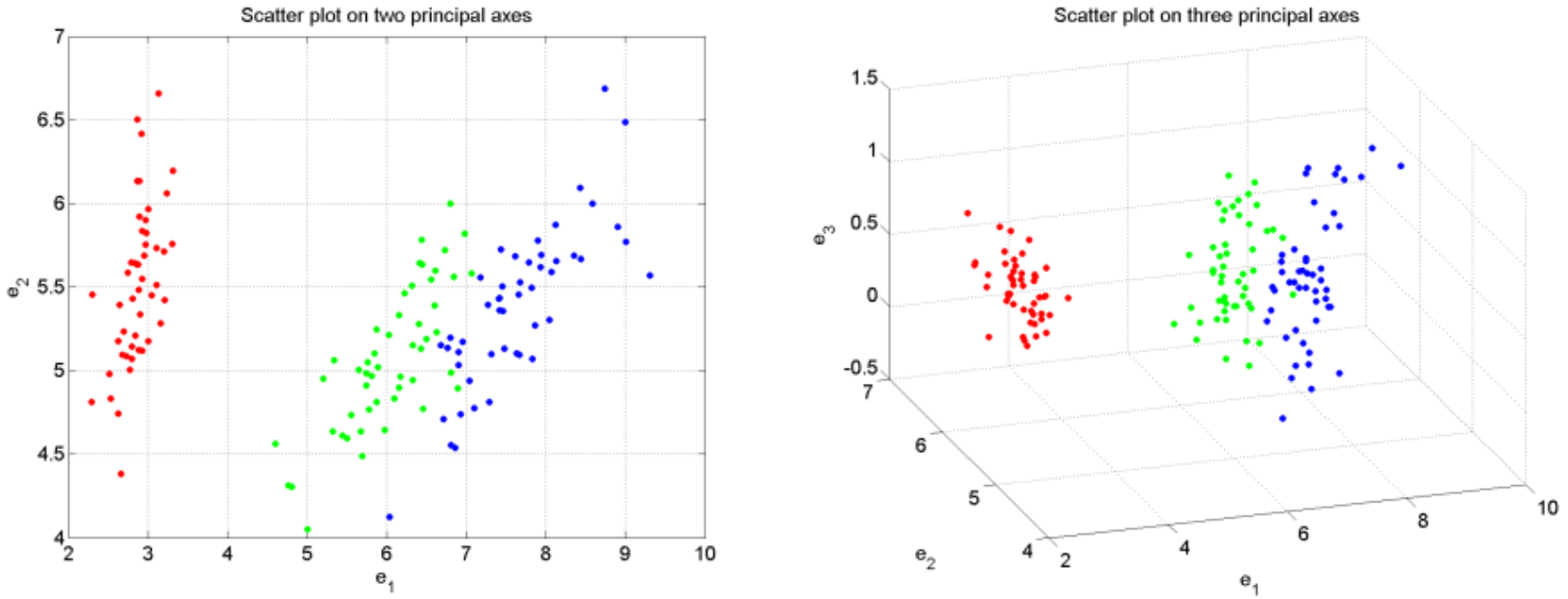
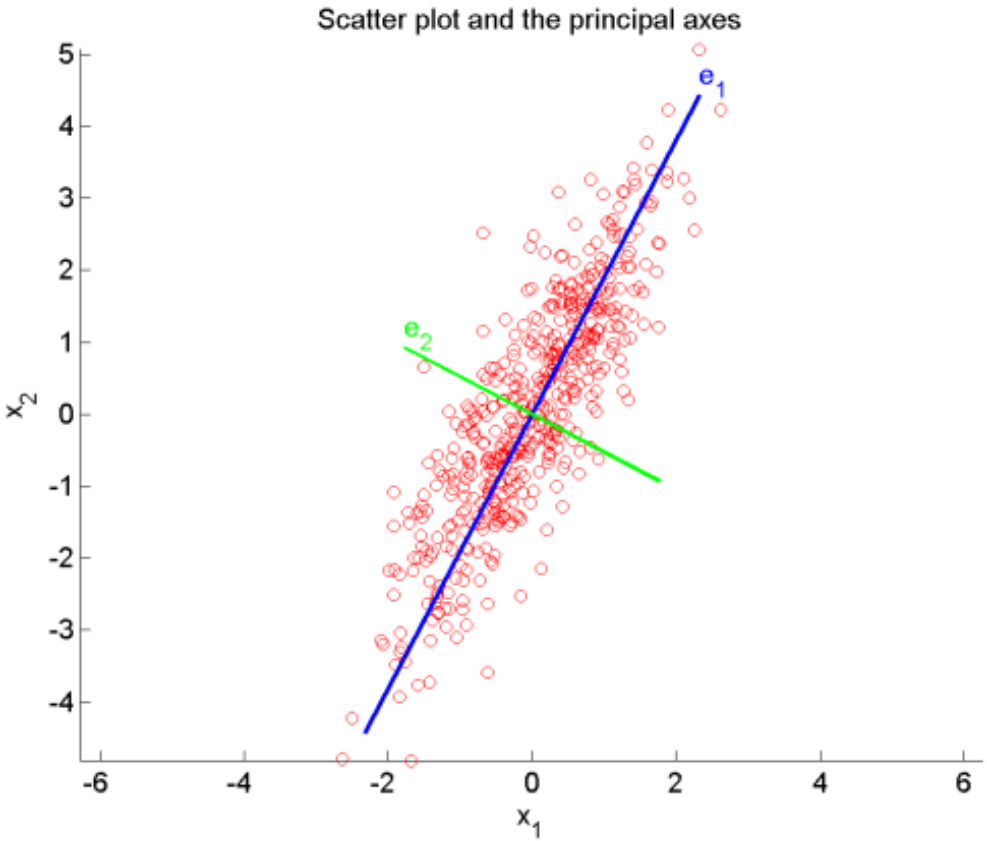
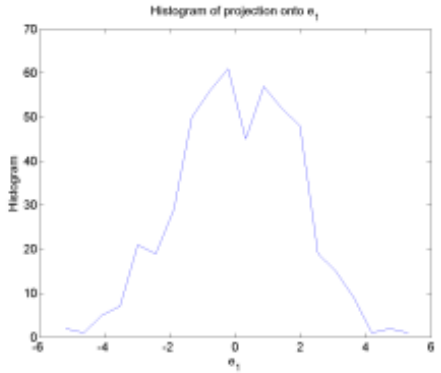


Figure: Scatter plot of the projection of the iris data onto the first two and the first three principal axes. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.[2]

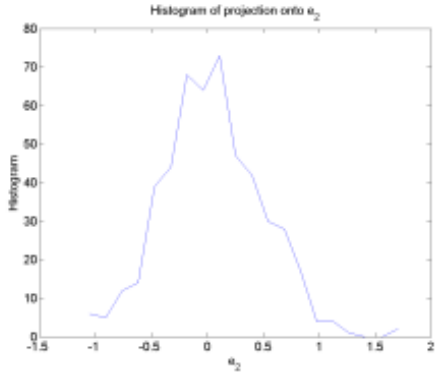
Example 1



(a) Scatter plot.



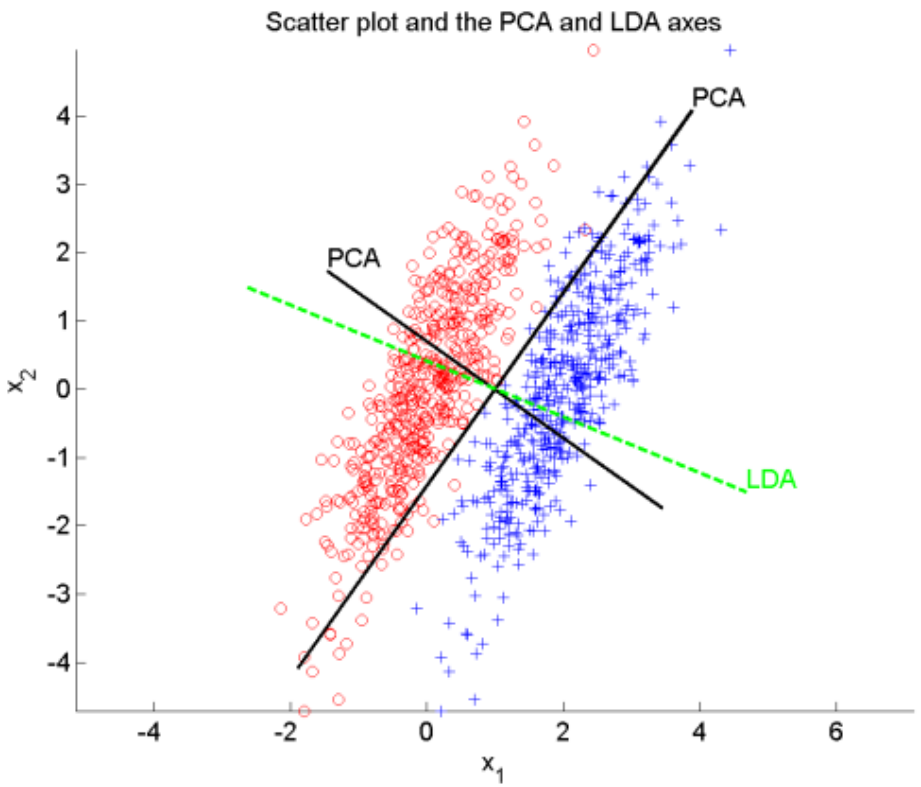
(b) Projection onto e_1.



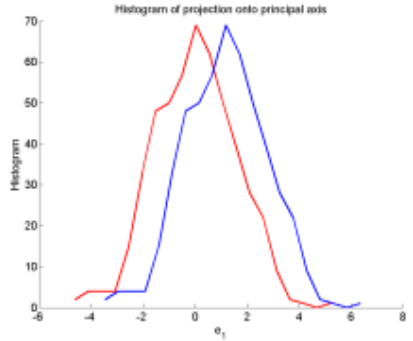
(c) Projection onto e_2.

Figure: 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..[2]

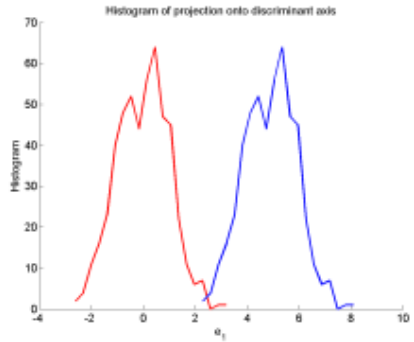
Example 2



(a) Scatter plot.



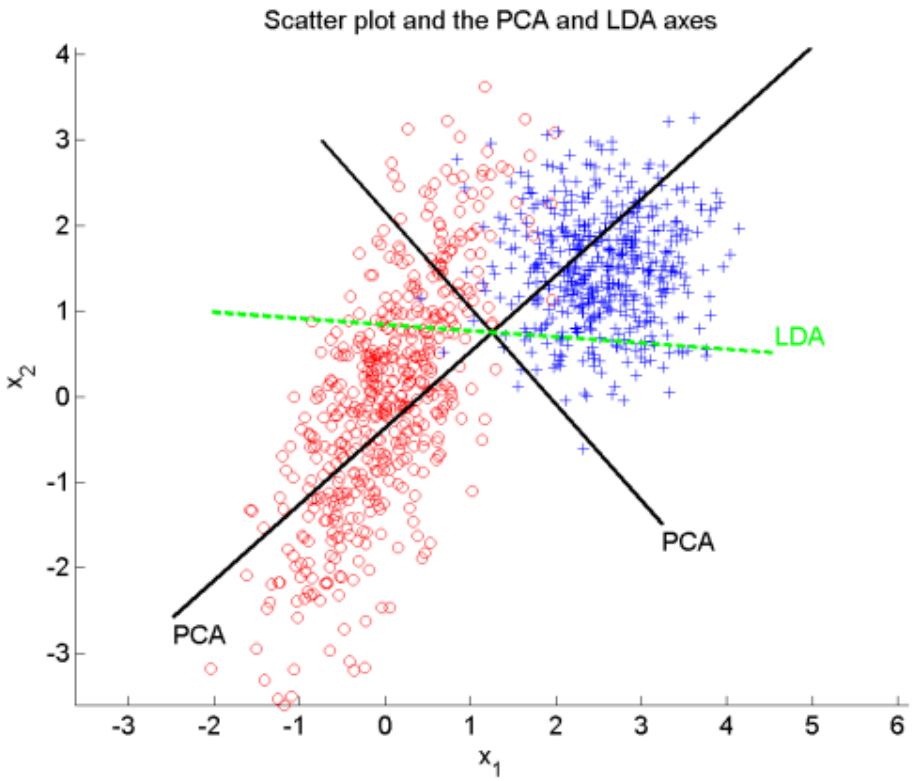
(b) Projection onto the first PCA axis.



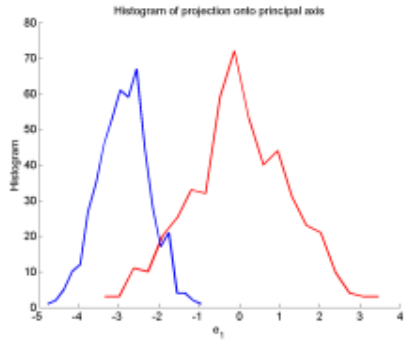
(c) Projection onto the first LDA axis.

Figure: 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.[2]

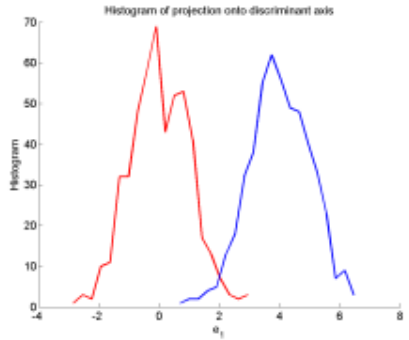
Example 3



(a) Scatter plot.



(b) Projection onto the first PCA axis.



(c) Projection onto the first LDA axis.

Figure: 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.[2]

Example 4

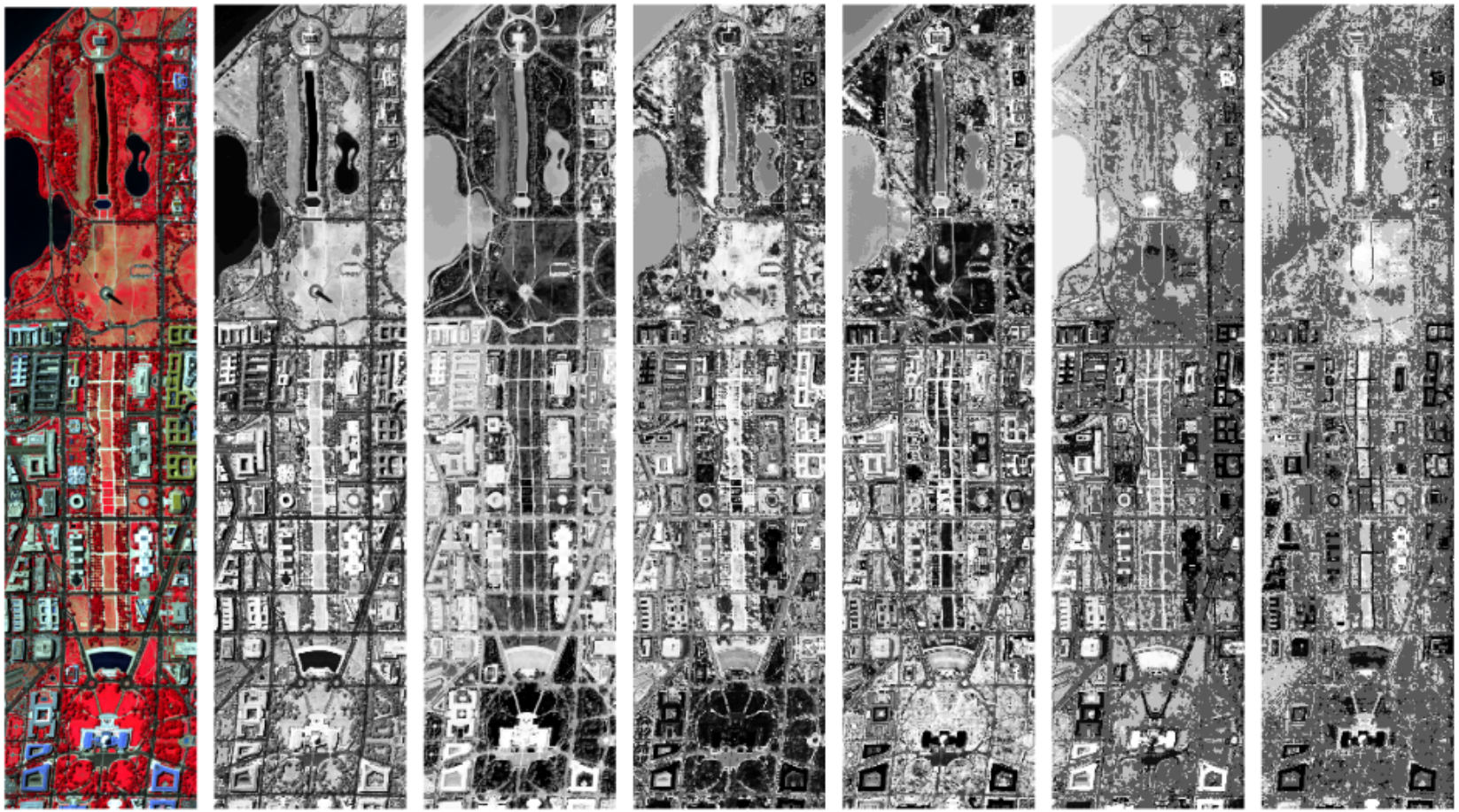


Figure: A satellite image and the first six PCA bands (after projection). Histogram equalization was applied to all images for better visualization.[2]

Example 5

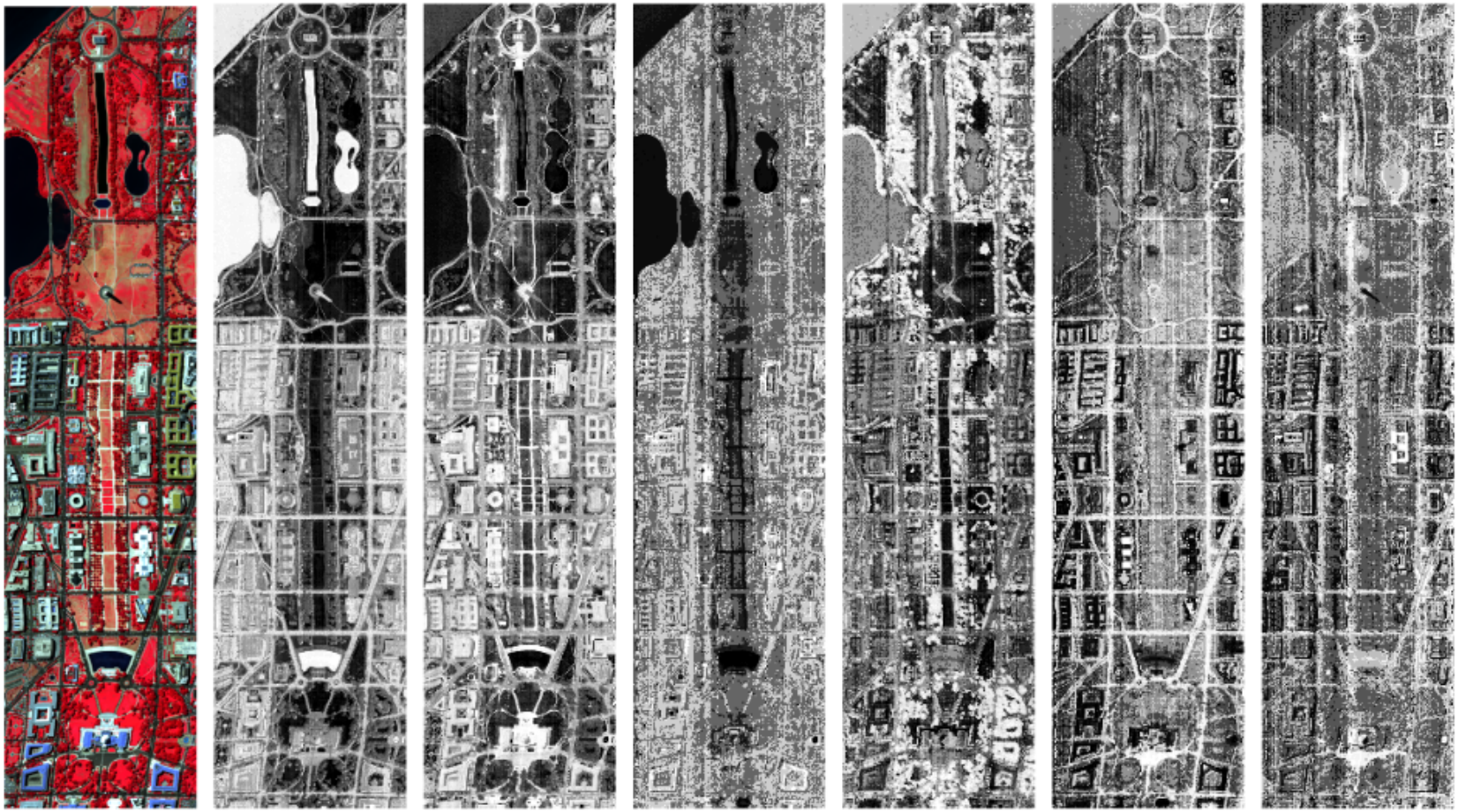


Figure: A satellite image and the six LDA bands (after projection). Histogram equalization was applied to all images for better visualization.[2]

Example 6

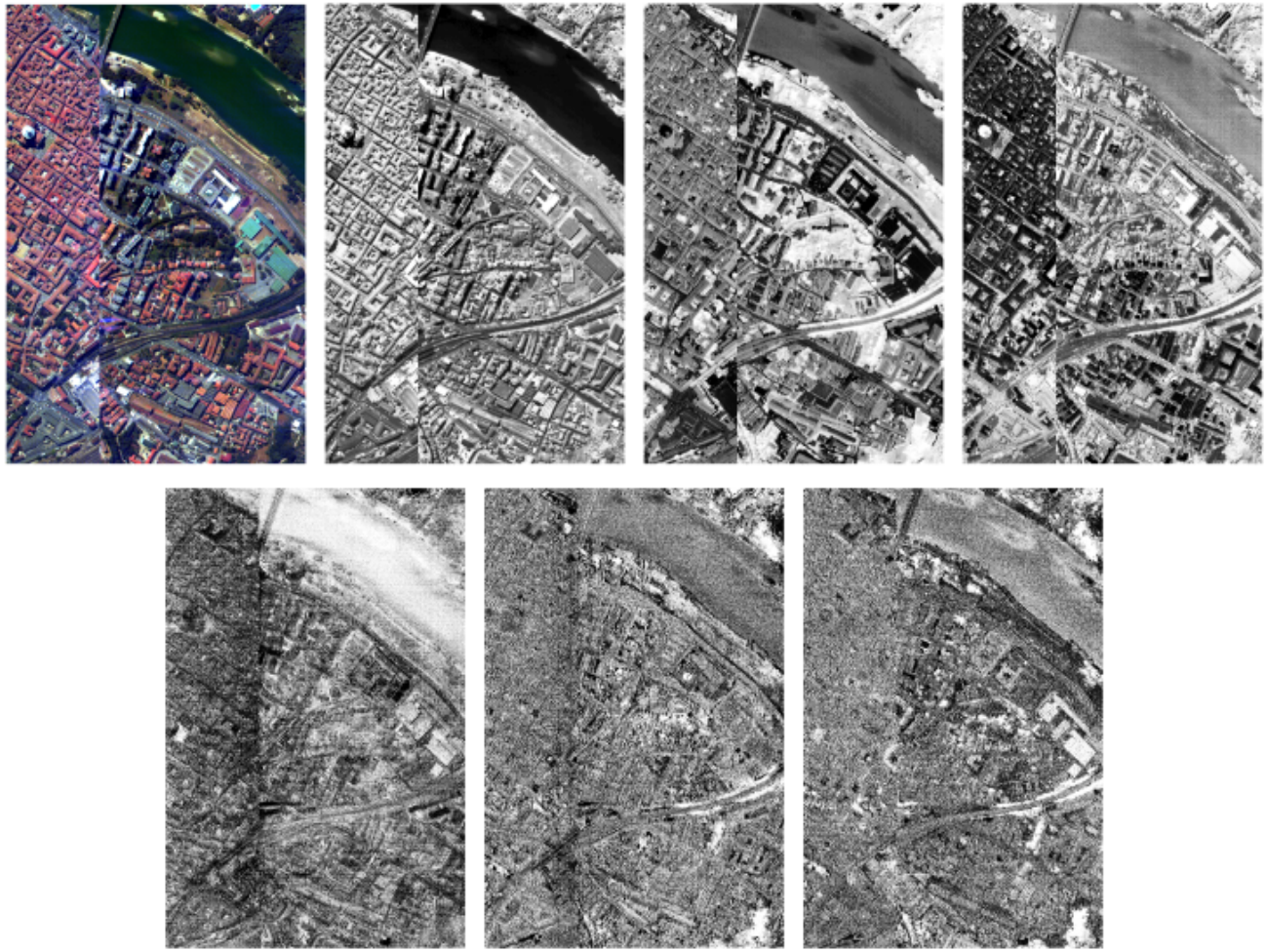


Figure: A satellite image and the first six PCA bands (after projection). Histogram equalization was applied to all images for better visualization.[2]

Example 7

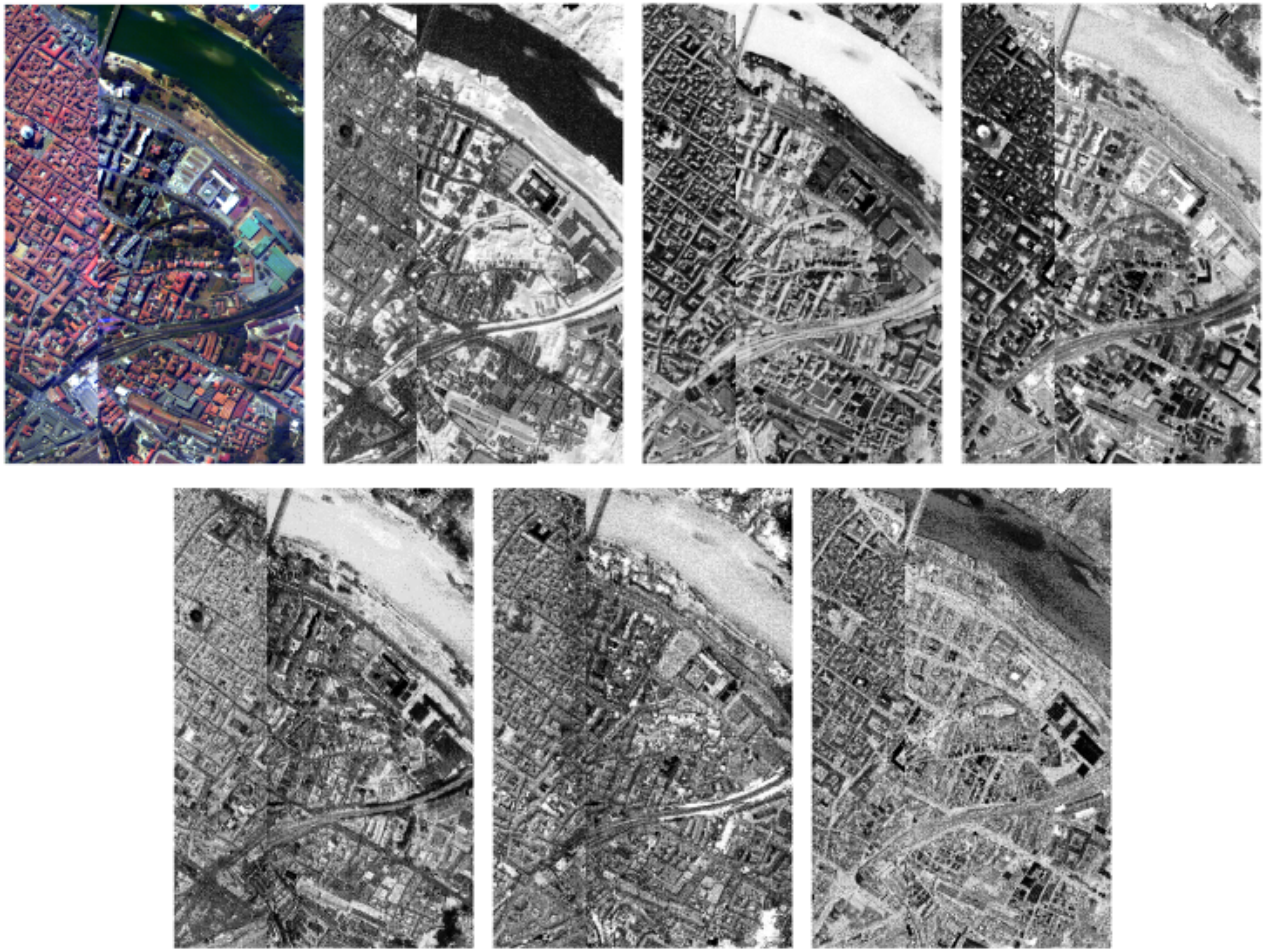


Figure: A satellite image and the six LDA bands (after projection). Histogram equalization was applied to all images for better visualization.[2]

Reference I

[1] Richard O Duda, Peter E Hart, et al. **Pattern Classification**. 2nd ed. Wiley New York, 2000.

[2] **online resources : pca slide**. URL: http://www.cs.bilkent.edu.tr/~saksoy/courses/cs551/slides/cs551_dimensionality.pdf.

[3] Christopher M Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006.

[4] **online resources : pca slide**. URL: http://www.cse.psu.edu/~rtc12/CSE586Spring2010/lectures/pcaLectureShort.pdf.

[5] **online resources : pca slide**. URL: https://davidrosenberg.github.io/mlcourse/Archive/2017/Lectures/13-PCA-Slides.pdf.

[6] **online resources : pca slide**. URL: https://www.cs.cmu.edu/~mgormley/courses/10701-f16/slides/lecture14-pca.pdf.

Thank You !

$\mathcal{Q}$ & $\mathcal{A}$