Multidimensional Data Visualization

Linear Projection Methods

Dimensionality Reduction

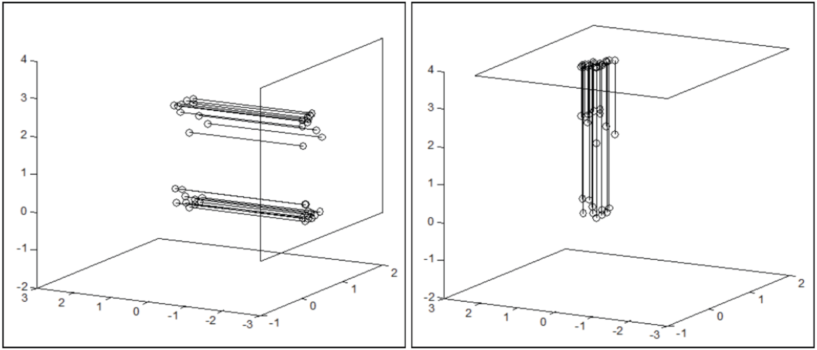
- It is difficult to perceive the data structure using the direct visualization methods, particularly when we deal with large data sets or data of high dimensionality.
- Projection methods are based on reduction of the dimensionality of data.
- Their advantage is that each n-dimensional object is represented as a point in the space of low-dimensionality d, $d < n$, usually $d = 2$.
- There exists a lot of methods that can be used for reducing the dimensionality.
- The aim of these methods is to represent the multidimensional data in a low-dimensional space so that certain properties (such as distances, topology or other proximities) of the data set were preserved as faithfully as possible.
- These methods can be used to visualize the multidimensional data, if a small enough resulting dimensionality is chosen.

Projection Methods

- Methods that allow us to represent multidimensional data from $\mathbb{R}^n$ in a low-dimensional space $\mathbb{R}^d$, $d < n$, are called projection (dimensionality reduction) methods.
- If the dimensionality of the *projection space* is small enough ($d = 2$ or $d = 3$), these methods may be used to visualize the multidimensional data.
- In such a case, the projection space can be called a *display*, *embedding* or *image space*.
- These methods usually invoke formal mathematical criteria by which the projection distortion is minimized.
  1. Linear projection methods:
     a) principal component analysis,
     b) linear discriminant analysis,
     c) projection pursuit.
  2. Nonlinear projection methods:
     a) multidimensional scaling,
     b) locally linear embedding,
     c) isometric feature mapping,
     d) principal curves.

Example of Projection of Three-Dimensional Points

- Figure shows two possible ways of projections of the three-dimensional points ($n = 3$) onto a plane ($d = 2$). We can see two clusters of points on the projection plane on the left and only one cluster on the projection plane on the right.



Example of Projections

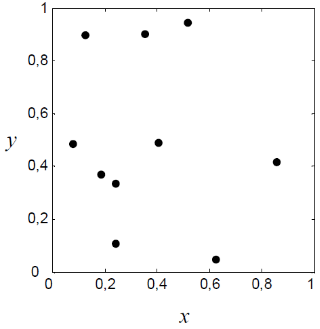
- The example demonstrates that different projections of the same data can reveal different aspects of the data structure (clusters, outliers, etc.).
- Indeed, some projections can fail to reveal any structure.
- Therefore, the proper choose of projection is an important problem.
- When visualizing multidimensional data, we confront with two often contradictory aims.
- On one hand, we want to reduce the dimensionality of data in the simplest way.
- On the other hand, we want to preserve the original information as much as possible.

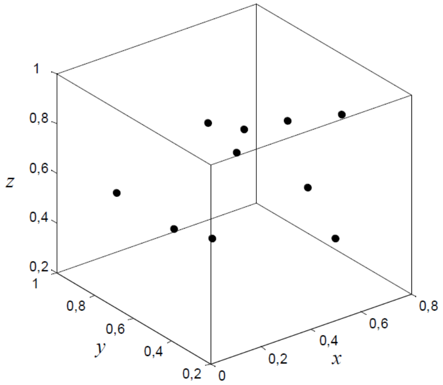
Projection Methods

- The projection methods are used for *transformation* of multidimensional data to a low-dimensional space.
- The aim of these methods is to represent the multidimensional data in a low-dimensional space so that certain properties of the data set were preserved as faithfully as possible.
- These methods can be used to visualize the multidimensional data, if a sufficiently small dimensionality of the projection space $\mathbb{R}^d$ is chosen ($d = 2$ or $d = 3$).
- We call the space $\mathbb{R}^d$ as a display or image space, since its points can be observed visually.

Scatter Plots

- *Scatter plots* are one of the most commonly used techniques for data representation on a plane $\mathbb{R}^2$ or space $\mathbb{R}^3$. Points are displayed in the classic (x, y) or (x, y, z) format.





Projection Methods

- Suppose that the multidimensional data set is defined by a matrix

  $$X = \{X_1, X_2, \ldots, X_m\} = \left\{x_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\right\}.$$

  Here m is the number of objects (n-dimensional points $X_i \in \mathbb{R}^n$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i \in \{1, \ldots, m\}$). x_{ij} is the jth coordinate, corresponding to the jth feature.
- One needs to find a transformation of the points $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, $i = 1, \ldots, m$, into points $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $i = 1, \ldots, m$, that are on a low-dimensional space $\mathbb{R}^d$, $d < n$. One-dimensional space ($d = 1$) can also be used, however more information can be preserved when observing points on a plane ($d = 2$) or a 3D space ($d = 3$).

Criteria of the Projection Quality

- There are some formal mathematical criteria of the projection quality. These criteria are optimized in order to get the optimal projection of multidimensional data onto a low-dimensional space.
- The main goal is to preserve the proportions of distances or estimations of other proximities between the multidimensional points in the image space as well as to preserve, or even to highlight other characteristics of the multidimensional data (for example, clusters).
- Let us remind that the proximity is the general term of *similarity* and *dissimilarity*. A high value of similarity indicates that the objects X_i and X_j are very similar. For dissimilarities, a small value indicates that the objects are very similar, e.g. dissimilarity may be measured using the Euclidean (or other) distances.

Proximity of Data

- A (dis)similarity is a proximity that indicates how two objects X_i and X_j are (dis)similar. The (dis)similarity is denoted by δ_{ij}.
- If δ_{ij} is a similarity, a high δ_{ij} value indicates that the objects X_i and X_j are very similar.
- For dissimilarities, a small δ_{ij} value indicates that the objects are very similar.
- When the proximities are known, the visualization of objects $X_1, X_2, \ldots, X_m$ may be carried out using the matrix of their proximities $\Delta = \{\delta_{ij}, i, j = 1, \ldots, m\}$. The advantage is that the dimensionality n can be unknown. This often happens, for example, in psychological tests.

Proximity Measures

- A proximity matrix can be obtained from matrix X applying some proximity measure, too.
- Often the proximity is measured using the Euclidean distance, which belongs to the group of Minkowski distances. The Minkowski distance between two objects $X_k = (x_{k1}, x_{k2}, \ldots, x_{kn})$ and $X_l = (x_{l1}, x_{l2}, \ldots, x_{ln})$ is defined by the formula:

$$d_q(X_k, X_l) = \left\{ \sum_{j=1}^{n} |x_{kj} - x_{lj}|^q \right\}^{\frac{1}{q}} .$$

- Some other proximity measures are also possible: Canberra distance, Bray-Curtis dissimilarity, correlation, etc.

Minkowski Distances

- The following Minkowski distances may be derived for different q:
- City-block or Manhattan distance, $q = 1$:

$$d_1(X_k, X_l) = \sum_{j=1}^{n} |x_{kj} - x_{lj}|.$$

- Euclidean distance, $q = 2$:

$$d_2(X_k, X_l) = \sqrt{\sum_{j=1}^{n} |x_{kj} - x_{lj}|^2}.$$

- Chebyshev distance, $q = \infty$:

$$d_\infty(X_k, X_l) = \max_j \left| x_{kl} - x_{lj} \right|.$$

Linear Transformation

- There are linear and nonlinear projection methods.
- Linear projection methods pursue a linear transformation of data. There are various linear transformations: rotation, shearing, reflection, scaling, etc.
- A *linear transformation* may be described by linear equations
$$Y_i = X_i A + B.$$
- If $d = n$, i.e. $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ and $X_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, then A is a square matrix, consisting of n rows and n columns. The matrix A is called a *transformation matrix*.
- If a linear transformation is used for dimensionality reduction, then $d < n$, $Y_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $i = 1, \ldots, m$, and A is a matrix, consisting of n rows and d columns.

Example of Linear Transformation

- Let us analyze a simple case of a linear transformation, when $n = d = 2$. Let us have a point $X_i = (x_{i1}, x_{i2})$. Transform it linearly to a point $Y_i = (y_{i1}, y_{i2})$ using a matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

- In the case of rotation, the elements of the matrix A can be expressed using trigonometric functions:

$$A = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix},$$

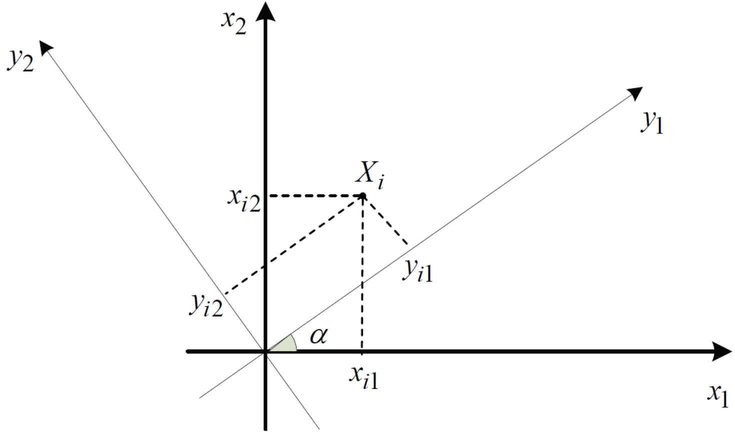
  where α is the rotation angle between the axes x_1 and y_1, as well as between the axes x_2 and y_2.
- The coordinate system (x_1, x_2) is rotated around the origin $(0, 0)$ counterclockwise by α to get a coordinate system (y_1, y_2).
- Such a matrix A is called a *rotation matrix*.

Example of Linear Transformation

- The coordinates of the point Y_i may be expressed as:

$$\begin{aligned} y_{i1} &= x_{i1}\cos(\alpha) + x_{i2}\sin(\alpha), \\ y_{i2} &= x_{i2}\cos(\alpha) - x_{i1}\sin(\alpha). \end{aligned}$$

- Actually (y_{i1}, y_{i2}) is a linear transformation of the point X_i to the coordinate system (y_1, y_2).



Nonlinear Transformation

- A nonlinear transformation is such that cannot be expressed in the linear form.
- The nonlinear transformation may be described as follows:

  $$Y = f(X),$$

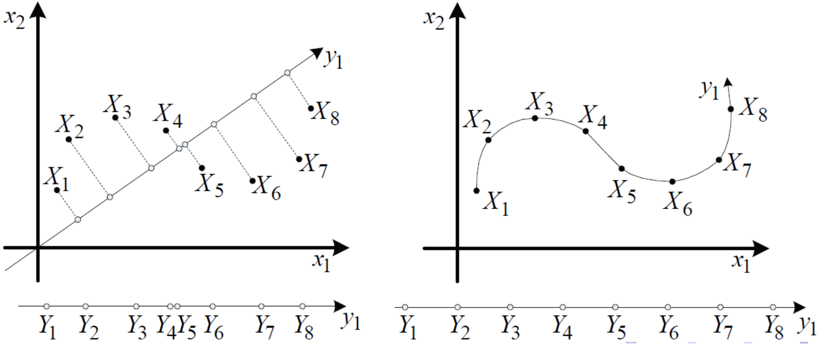
  where f is a nonlinear function and

  $$Y = \{Y_1, Y_2, \ldots, Y_m\} = \{y_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\}.$$

- The nonlinear transformation is more complicated than the linear one and requires more time-consuming computations. However, such a transformation allows us to preserve the characteristics of multidimensional data better as compared with the linear transformation if $d < n$, i.e. the data are projected to a lower-dimensional space.

Linear and Nonlinear Projection

- Let the two-dimensional points $X_1, X_2, \ldots, X_8$ be spread so that the distances between the nearest points are equal.
- If we project to the one-dimensional space using the linear projection (to the line y_1), equal distances between the nearest points are not preserved. However, in the case of the nonlinear projection, when the proper transformation is found, the distances between the nearest points remain equal.



Principal Component Analysis

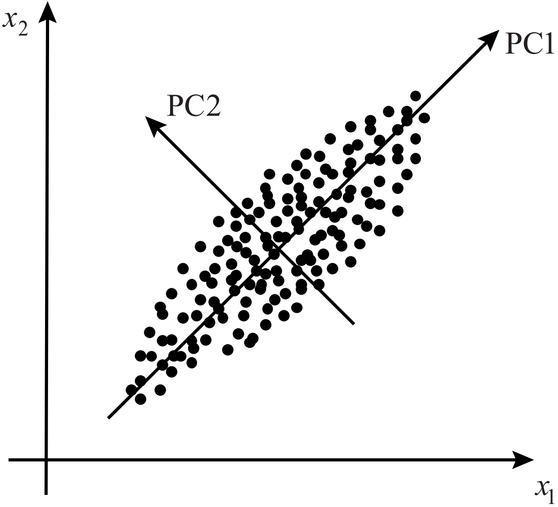
- The principal component analysis (PCA) is a well-known data analysis technique invented in 1901 by Pearson.
- It is a way of linear transforming a set X of n-dimensional points $X_1, X_2, \ldots, X_m$ into another set Y of n-dimensional points $Y_1, Y_2, \ldots, Y_m$.
- The property of the set is that the largest part of its information content is stored in the first few coordinates (components) of points Y_i, $i = 1, \ldots, m$.
- The principal component analysis is often used to reduce the dimensionality of multidimensional points X_i, $i = 1, \ldots, m$, by discarding some of the components of the points Y_i and by leaving only the first (principal) d ones.
- The principal component analysis projects the data linearly into a low-dimensional space preserving the variance of the data best.

Principal Component Analysis

- The main idea of PCA is to reduce the dimensionality of data by performing a linear transformation and rejecting a part of the components, variances of which are the smallest ones.
- When analyzing the data set X, a direction in $\mathbb{R}^n$ with the maximal variance is found.
- This direction defines the first principal component.
- Other principal components maximize the variance of a data set in the directions orthogonal to the previous principal components.
- So, the principal components are uncorrelated and ordered by decreasing variances.

Illustration of Principal Component Analysis

- Figure illustrates a two-dimensional case with two principal components PC1 and PC2.



Principal Component Analysis

- The principal component analysis needs a correlation or covariance matrix of features.
- Suppose we have a data matrix X:

$$X = \{X_1, X_2, \ldots, X_m\} = \left\{x_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n\right\}.$$

- The rows of this matrix correspond to the objects $X = \{X_1, X_2, \ldots, X_m\}$ and the columns correspond to the features $x_1, x_2, \ldots, x_n$ characterizing the objects.

Correlation

- A *correlation* is a number that describes the degree of relationship between two features.
- The *correlation coefficient* r_{kl} between the features x_k and x_l is computed by the formula:

$$r_{kl} = \frac{\sum_{i=1}^{m}(x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{m}(x_{ik} - \bar{x}_k)^2 \sum_{i=1}^{m}(x_{il} - \bar{x}_l)^2}},$$

  where

$$\bar{x}_k = \frac{1}{m}\sum_{i=1}^{m} x_{ik} \text{ and } \bar{x}_l = \frac{1}{m}\sum_{i=1}^{m} x_{il}.$$

- The *correlation matrix* $R = \{r_{kl},\ k, l = 1, \ldots, n\}$ consists of the correlation coefficients. The diagonal elements $r_{kk},\ k = 1, \ldots, n$, are equal to 1. This matrix is symmetric.

Covariance

- The *covariance coefficient* c_{kl} between the features x_k and x_l is computed by the formula:

$$c_{kl} = \frac{1}{m-1} \sum_{i=1}^{m} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l).$$

- If $k = l$, the expression is a variance formula, i.e. c_{kk} is the variance of feature x_k. The *covariance matrix* C consists of the covariance coefficients:

$$C = \{c_{kl},\ k, l = 1, \ldots, n\}.$$

- It follows that the correlation coefficient is equal to:

$$r_{kl} = \frac{c_{kl}}{\sqrt{c_{kk} c_{ll}}}.$$

- If the features x_k and x_l are not correlated, their covariance coefficient is equal to zero: $c_{kl} = c_{lk} = 0, k \neq l$.

Principal Component Matrix

- Let us describe the eigenvector and the eigenvalue of the covariance matrix.
- The *eigenvalue* λ_k and the *eigenvector* E_k corresponding to λ_k are solutions of the equation $CE_k = \lambda_k E_k$. Here E_k is a vector-column. The value of λ_k is found from the characteristic equation $|C - \lambda_k I| = 0$, where I is an identity matrix of the same order as the matrix C and $|.|$ denotes a determinant of the matrix.
- The number of eigenvectors is equal to n.
- Let us sort the eigenvectors E_k, $k = 1, \dots, n$, in descending order of the corresponding eigenvalues $(\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n)$.
- The matrix $A = (E_1, E_2, \dots, E_n)$ is called a principal component matrix. The columns of this matrix are the eigenvectors E_k, $k = 1, \dots, n$, corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n$.
- Each column of the matrix A is orthogonal to any other column. Usually E_k, $k = 1, \dots, n$ of unit length are used.

Principal Component Transformation

- Let us transform the points X_i, $i = 1, \dots, m$, to points Y_i, $i = 1, \dots, m$, by the formula:

  $$Y_i = (X_i - \bar{X})A, \; i = 1, \dots, m,$$

  where $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$, $A = (E_1, E_2, \dots, E_n)$, A is a transformation matrix.
- $Y_i = (y_{i1}, y_{i2}, \dots, y_{in})$, obtained by the formula, are points in the new coordinate system $(y_1, y_2, \dots, y_n)$.
- The eigenvectors E_k, $k = 1, \dots, n$ represent the basis set of this system.
- The covariance matrix of components $y_1, y_2, \dots, y_n$ of the points Y_i, $i = 1, \dots, m$ is equal to:

  $$\begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Principal Component Transformation with Dimensionality Reduction

- We may use only a few first eigenvectors for transforming multidimensional data instead of all the eigenvectors of the covariance matrix.
- Suppose that the matrix A_d consists of the first d eigenvectors. Then it is possible to define a transformation

$$Y_i = (X_i - \bar{X})A_d,\ i = 1, \ldots, m.$$

- In this way, a projection of the point X_i to the d-dimensional space is derived.

Properties of Eigenvalues

- Some properties of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are as follows:
  1. $\sum_{k=1}^{n} \lambda_k = \sum_{k=1}^{n} c_{kk}$.
  2. $\lambda_1 \geq \max_k c_{kk}$.
  3. $\lambda_n \leq \min_k c_{kk}$.
- It follows from the second property that the first eigenvector E_1 describes the first principal component y_1, the variance of which is the highest one among $\lambda_1, \lambda_2, \ldots, \lambda_n$. The second eigenvector E_2 describes the second principal component y_2, the variance of which is the second one according to the value.

Principal Components in Data Visualization

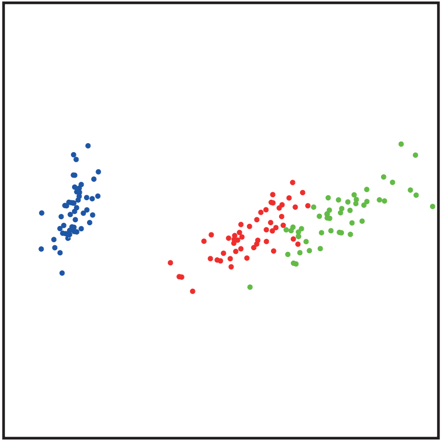
- It follows from the third property that the last eigenvector E_n describes the principal component y_n, the variance of which is smallest. Therefore, if only the first d eigenvectors are used, the components with the smallest variances will be rejected.
- In order to derive the principal components, it suffices to find the highest d eigenvalues and the corresponding eigenvectors of matrix C. This matrix has specific properties: it is symmetric and non-negative definite, therefore, special fast algorithms are used.
- The advantage of the principal component analysis is the simplicity of its idea. This fact influences its popularity and wide application.

Application of Principal Component Analysis in Visualization of Multidimensional Data

- The examples of application of principal component analysis in visualization of multidimensional data are presented.
- The Iris and the Breast Cancer data sets are visualized by two principal components.
- We do not present labels and units for both axes in the figure, because we are interested in observing the interlocation of points on a plane only.
- In figures we can observe clusters of points, corresponding to particular classes of *n*-dimensional objects.

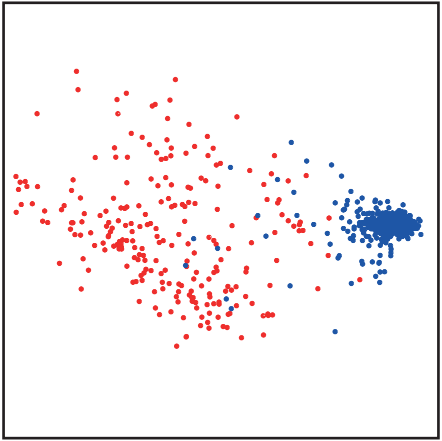
Iris Data Set Visualized Using PCA

- Setosa irises (marked in blue) are faraway from Versicolor (red) and Virginica (green) irises. There is no exactly expressed boundary between these two species.



Breast Cancer Data Set Visualized Using PCA

- A large amount of the points, corresponding to the benign tumor data (blue points), are concentrated in one area, and the other points, corresponding to the malignant tumor data (red), are spread widely.



Principal Component Analysis

- In the literature, some authors prefer to define the principal components using the correlation matrix instead of the covariance one. The correlation between a pair of features is equivalent to the covariance divided by the product of the standard deviations of two features.
- Although PCA is widely used for multidimensional data visualization, it has some shortcomings. It is not good for data of nonlinear structures, consisting of arbitrarily shaped clusters or curved manifolds.
- During the past 50 years many works have appeared proposing extensions of the principal components to data with a nonlinear structure. *Principal curves* are a nonlinear generalization of principal components. The principal curve provides a nonlinear summary of the data. The idea of the principal curves can be extended to principal surfaces.

Linear Discriminant Analysis

- In contrast to most other dimensionality reduction methods, a *linear discriminant analysis* (LDA) is a supervised method.
- The method is often called Fisher's discriminant analysis.
- In a supervised strategy, some known properties of data (for example, belonging of the objects to one of classes) are applied.
- LDA transforms multidimensional data to a low-dimensional space, maximizing the linear separability between objects belonging to different classes.

Linear Discriminant Analysis

- Suppose that the data matrix X

  $$X = \{X_1, X_2, \dots, X_m\} = \left\{x_{ij},\ i = 1, \dots, m,\ j = 1, \dots, n\right\}.$$

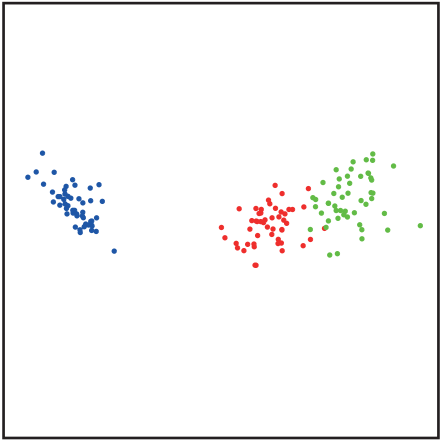
  consists of k submatrices $X^{(1)}, X^{(2)}, \dots, X^{(k)}$, where k is the number of classes.
- The rows of $X_i^{(j)}$, $i = 1, \dots, m_j$, of $X^{(j)}$ correspond to objects that belong to the jth classes. Here m_j is the number of objects in the jth class. The number of all objects $m = \sum_{j=1}^{k} m_j$.

Scheme of Linear Discriminant Analysis

1. Covariance matrix C of the whole data set X and covariance matrices $C^{(j)}$, $j = 1, \ldots, k$, of each class are computed. The within-class scatter is defined: $S_w = \sum_{j=1}^{k} p_j C^{(j)}$, where $p_j = \frac{m_j}{m}$. The between-class scatter is defined: $S_b = C - S_w$.
2. The eigenvectors and eigenvalues of the matrix $S = S_w^{-1} S_b$ are computed. The eigenvectors are sorted in descending order of the corresponding eigenvalues. d eigenvectors, corresponding to the highest eigenvalues, are selected (under the requirement that $d < k$).
3. The transformation $Y_i = (X_i - \bar{X})A_d$, $i = 1, \ldots, m$, is performed, where A_d is a d-column matrix consisting of the eigenvectors, corresponding to the highest d eigenvalues of matrix S. Here $\bar{X} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ is the vector of averages of the features.

Iris Data Set Visualized Using LDA

- The difference between LDA and PCA is that, in addition, the known classes of objects are applied.



Projection Pursuit

- A projection pursuit is a type of statistical technique which involves finding the most “interesting” projection in multidimensional data.
- The aim of the projection pursuit, as many other projection methods, is to find such linear combinations of components (commonly two or three dimensional) that the transformed data preserve a structure of the initial data.
- Suppose, we have the data matrix X, rows of which are the n-dimensional vectors. A projection of points of the space $\mathbb{R}^n$ to the space $\mathbb{R}^d$ can be expressed as follows: $Y = XA$, where $X = \{X_1, X_2, \dots, X_m\} = \{x_{ij}, i = 1, \dots, m, j = 1, \dots, n\}$, A is a matrix consisted of n rows and d columns, $Y = \{Y_1, Y_2, \dots, Y_m\} = \{y_{ij}, i = 1, \dots, m, j = 1, \dots, d\}$ is a matrix of the data obtained after projection. A question arises how to choose the matrix A.

Projection Pursuit

- At first, we have to decide what kind of feature (property) we wish to detect (highlight) in a visualization process. After that, one needs to define a measure that reflects this property. Call this measure $I(Y)$. It is sometimes called the index function.
- Suppose we wish to highlight data clusters. In such a case, a statistical clustering measure – mean nearest neighbor distance could be used. Lower values of this measure indicate greater clustering.
- Let's $I(Y)$ is the mean nearest neighbor distance. The expression $I(Y)$ can be written in the form $I(XA)$. The problem of projection choice can be formulated as an optimization problem: it is necessary to choose such a matrix A that the value of the function $I(XA)$ was minimal.
- If we are interested in a presence of outliers, then the problem is changed so that the maximal mean nearest neighbor distance were derived. More complex index functions could be used too.