Searching for information hidden in multivariate data

Mgr. Radoslav Harman, PhD.

Department of Applied Mathematics and Statistics
Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava

May 2011

The aim and the outline of the talk

- Describe the fundamental ideas behind selected multivariate statistical methods and explain their interrelations.
- We will skip classical multivariate methods such as multivariate regression and multivariate analysis of variance. These two methods focus on hypotheses testing, and are usually dependent on the assumption of multivariate normality.
- We will talk about the modern statistical methods with the following general features:
  - The assumption of normality is often helpful, but not crucial.
  - The fundamental idea is based on an optimization problem.
  - The methods are computationally very intensive.
  - The results enhance our intuition about the structure of the data.
  - Often, people do not care that much why the methods work, but whether they work.

1. Finding a low-dimensional description of data
2. Detecting the structure of similarities in data
3. Discriminating and classifying data

Finding a low-dimensional description: General principles

We are dealing with samples $\mathbf{x}_1, ..., \mathbf{x}_n$ from a p-dimensional distribution $\mathcal{F}$, or we have p-dimensional data vectors $\mathbf{x}_1, ..., \mathbf{x}_n$ without the assumption of an underlying distribution.

$object_1$: $\mathbf{x}_1 = (x_{1,1}, x_{1,2}, ..., x_{1,p})'$
$object_2$: $\mathbf{x}_2 = (x_{2,1}, x_{2,2}, ..., x_{2,p})'$
...
$object_n$: $\mathbf{x}_n = (x_{n,1}, x_{n,2}, ..., x_{n,p})'$

General aim of low-dimensional description:

Try to find a random vector $\mathbf{Z}$ - a "low-dimensional" substitute of the "high-dimensional" random vector $\mathbf{X} \sim \mathcal{F}$, and/or low-dimensional substitutes $\mathbf{z}_1, ..., \mathbf{z}_n$ of the high-dimensional observed vectors $\mathbf{x}_1, ..., \mathbf{x}_n$. The substitute should capture as much of the interesting information as possible.

Different multivariate techniques from this category understand "low-dimensional substitutes" in different ways.

Finding a low-dimensional description: Principal components analysis

Construct a low dimensional "substitute" $\mathbf{Z} = (Z_1, ..., Z_r)'$ of $\mathbf{X} = (X_1, ..., X_p)'$ in the form

$$Z_i = \mathbf{u}_i'\mathbf{X}, \ i = 1, ..., r$$

such that $\mathbf{u}_1, ..., \mathbf{u}_r$ are unit-length vectors, and

$$\begin{aligned} \mathrm{Var}(Z_1) &= \max_{\|\mathbf{u}\|=1} \mathrm{Var}(\mathbf{u}'\mathbf{X}) \\ \mathrm{Var}(Z_i) &= \max_{\|\mathbf{u}\|=1; Z_i \perp Z_1, ..., Z_{i-1}} \mathrm{Var}(\mathbf{u}'\mathbf{X}); \ i = 2, ..., r \end{aligned}$$

Solution: $\mathbf{u}_1, ..., \mathbf{u}_r$ is a system of orthonormal eigenvectors corresponding to the r largest eigenvalues of the covariance matrix Σ of the random vector $\mathbf{X}$.

Finding a low-dimensional description: Principal components analysis

We say that $Z_1, ..., Z_r$ are the r largest principal components of $\mathbf{X}$, which together "explain"

$$\alpha_r = \sum_{i=1}^{r} \mathrm{Var}(Z_i) / \sum_{i=1}^{p} \mathrm{Var}(Z_i)$$

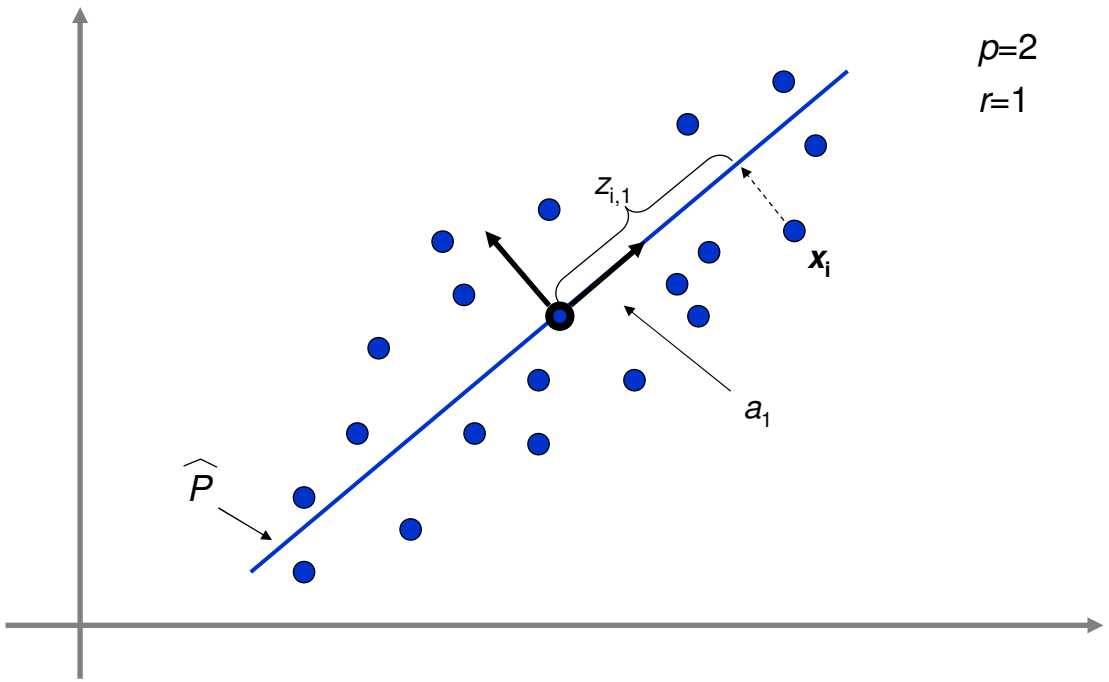
of variance. If α_r is "large", we say that the random vector $\mathbf{Z}$ captures large amount of information about the random vector $\mathbf{X}$.

If we have $\mathbf{x}_1, ..., \mathbf{x}_n$, we construct a sample covariance matrix $\mathbf{S}$, and the eigenvectors $\mathbf{a}_1, ..., \mathbf{a}_r$ corresponding to the r largest eigenvalues of $\mathbf{S}$ are taken as estimates of the eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_r$ of Σ.

Thus, the low-dimensional representation of the vectors $\mathbf{z}_1, ..., \mathbf{z}_n \in \mathbb{R}^r$ of the observations is constructed as $\mathbf{z}_i = \mathbf{a}_i' \mathbf{x}_i$.

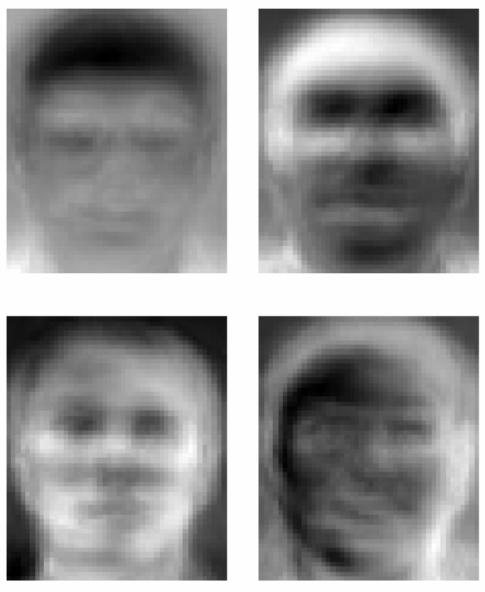
Finding a low-dimensional description: Principal components analysis

Distribution of data is often concentrated "around" an affine subspace $\mathcal{P}$. PCA "identifies" $\mathcal{P}$ as $\mathcal{L}(\mathbf{u}_1, ..., \mathbf{u}_r)$ shifted by the mean. The PCs $Z_1, ..., Z_r$ can be viewed as coordinates of the orthogonal projection of $\mathbf{X}$ onto $\mathcal{P}$. Similarly, $\mathbf{z}_1, ..., \mathbf{z}_n$ are the coordinates of the vectors $\mathbf{x}_1, ..., \mathbf{x}_n$ projected onto $\hat{\mathcal{P}}$, which is an estimate of $\mathcal{P}$ based on data.



Finding a low-dimensional description: Principal components analysis

Eigenfaces: $\mathbf{x}_1, ..., \mathbf{x}_n$ are levels of gray of $p = m \times m$ pixels of n photographs of faces. The eigenvectors $\mathbf{a}_1, ..., \mathbf{a}_r \in \mathbb{R}^p$ of the sample $p \times p$ covariance matrix $\mathbf{S}$ can be visualized as $m \times m$ images:



A characterization of a face with $m \times m$ pixel photograph coded as $\mathbf{x} \in \mathbb{R}^p$ can be represented by a small number r of coordinates $z_1 = \mathbf{a}_1'\mathbf{x}$,...,$z_r = \mathbf{a}_r'\mathbf{x}$.

Finding a low-dimensional description: Other methods

- **Principal Curves and Surfaces**: The vector $\mathbf{Z}$ corresponds to the vector of coordinates of the nearest point to $\mathbf{X}$ on a curve or surface in the p-dimensional space, i.e., instead of the r-dimensional plane $\mathcal{P}$ in the method of principal components we use an r-dimensional nonlinear manifold in $\mathbb{R}^p$.
- **Canonical Correlations**: The low-dimensional characterization of correlations of many-dimensional groups of variables is $\mathbf{Z}$, which is called "a vector of (the first) $r/2$ pairs of canonical correlations".
- **Multidimensional scaling**: For the data $\mathbf{x}_1, ..., \mathbf{x}_n$ we try to find low-dimensional representatives $\mathbf{z}_1, ..., \mathbf{z}_n$ such that for all pairs i, j of indices, the distance between $\mathbf{z}_i$ and $\mathbf{z}_j$ is as close as possible to the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. That is, $\mathbf{z}_1, ..., \mathbf{z}_n$ form a "low-dimensional map" of the high-dimensional data.

Detecting the structure of similarities: General principles

We are dealing with samples $\mathbf{x}_1, ..., \mathbf{x}_n$ from k distinct distributions $\mathcal{F}_1, ..., \mathcal{F}_k$ of high dimension p, or we simply have p-dimensional vectors $\mathbf{x}_1, ..., \mathbf{x}_n$ "belonging" to k clusters without an assumption of underlying probability distributions.

$object_1$: $\mathbf{x}_1 = (x_{1,1}, x_{1,2}, ..., x_{1,p})'$
$object_2$: $\mathbf{x}_2 = (x_{2,1}, x_{2,2}, ..., x_{2,p})'$
...
$object_n$: $\mathbf{x}_n = (x_{n,1}, x_{n,2}, ..., x_{n,p})'$

We do not know which of the data come from which distribution (or a "cluster"), and we usually do not even know the number k.

General aim of cluster analyses:

Divide the set of n multidimensional data into clusters based on their mutual distances, or "proximities".

Different techniques understand the division in different ways.

Detecting the structure of similarities: Model-based clustering

We assume that $\mathcal{F}_j \sim N_p(\mu_j, \Sigma_j)$, $j = 1, ..., k$. We also assume that the vector $\mathbf{x}_i$ originates from $\mathcal{F}_{\gamma_i}$, where $\gamma_i \in \{1, ..., k\}$. As an unknown parameter θ of the model we consider $\mu_1, ..., \mu_k$, $\Sigma_1, ..., \Sigma_k$, as well as the vector $\mathbf{g} = (\gamma_1, ..., \gamma_n)'$. To find an estimate of $\mathbf{g}$ we maximize the likelihood function

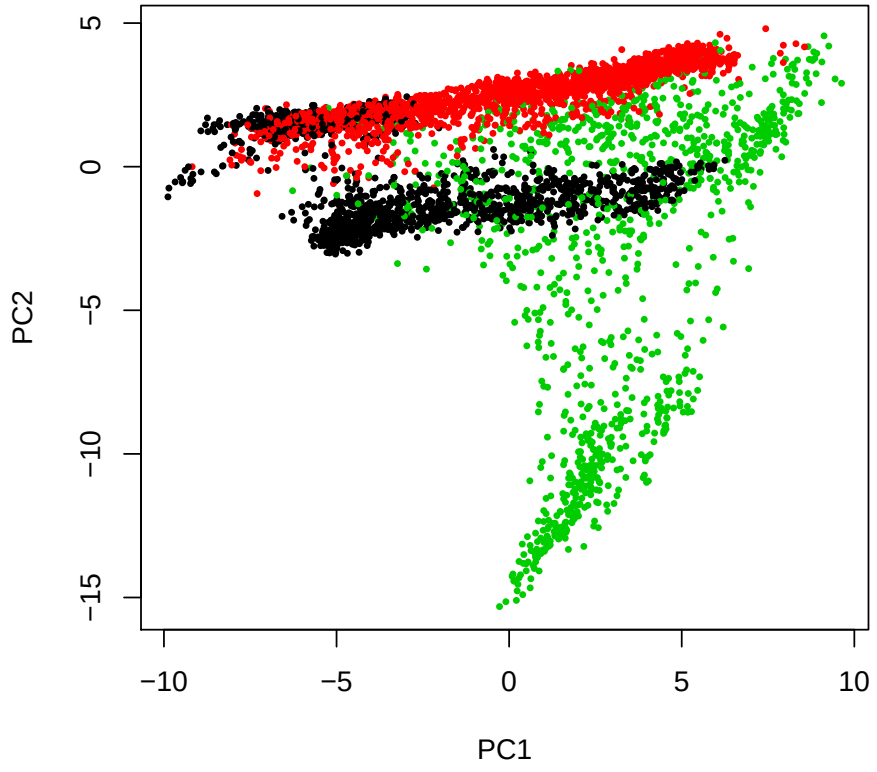
$$L(\theta|\mathbf{x}_1, ..., \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i; \mu_{\gamma_i}, \Sigma_{\gamma_i}),$$

where $f(\cdot; \mu, \Sigma)$ is the density of $N_p(\mu, \Sigma)$.

This is a very hard optimization problem that can be numerically solved using, e.g., the EM-algorithm or a genetic algorithm.

Detecting the structure of similarities: Model-based clustering

Landsat data: $\mathbf{x}_1, ..., \mathbf{x}_n$ are $p = 36$ dimensional measurements of color intensity of $n = 4435$ areas. The result of the normal model clustering with $k = 3$:

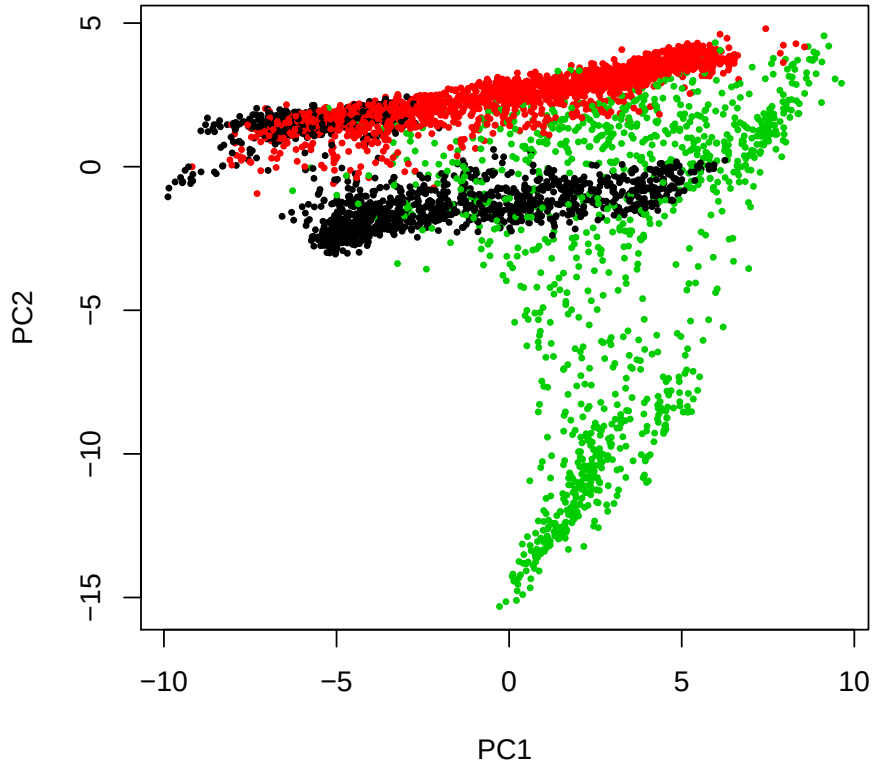


Detecting the structure of similarities: Other methods

- **The k-means method**: We construct a set of k points in the space $\mathbb{R}^p$ called centroids that "best represent" the k clusters. More precisely, the centroids are chosen so that the sum of the squared distances of $\mathbf{x}_1, ..., \mathbf{x}_n$ to the closest centroid is minimal.
- **The k-medoids method**: From the set of n objects we select, in an optimal way, k representatives $\mathbf{x}_{i_1}, ..., \mathbf{x}_{i_k}$ called medoids, and create the groups of data based on proximity to the medoids. The advantage is that we can use not only a metric distance, but general "dissimilarity" between the objects.
- **Hierarchical cluster analysis**: Creates a dendrogram - a "tree model" of similarity of data. Gives us not only an idea of separation of objects into groups, but also of mutual proximity of groups of data, of "groups of groups of data" etc.

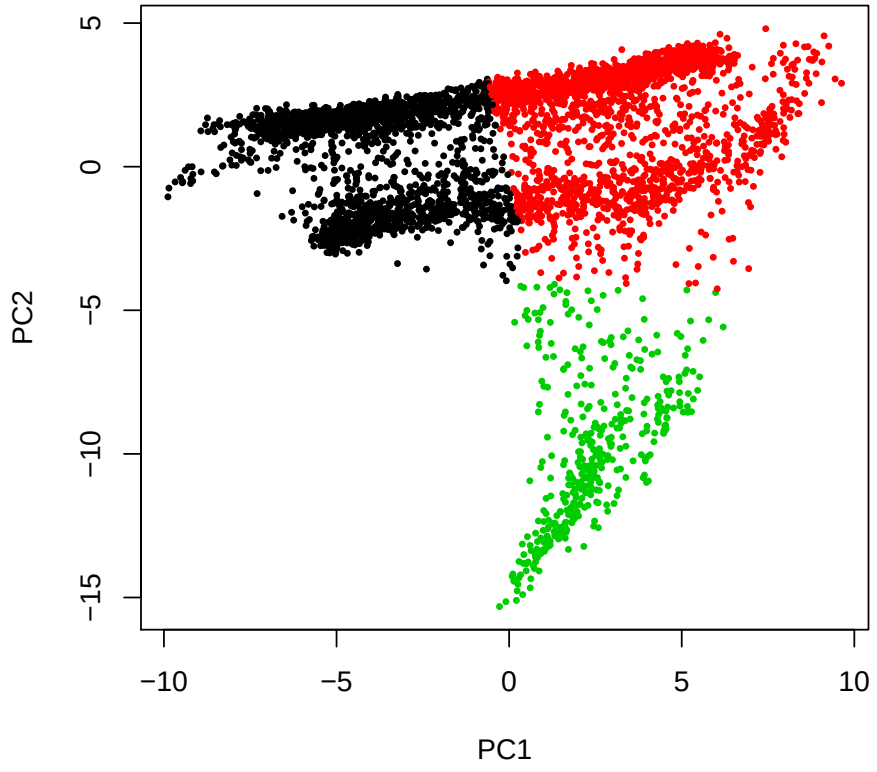
Detecting the structure of similarities: Model-based clustering

Landsat data: $\mathbf{x}_1, ..., \mathbf{x}_n$ are $p = 36$ dimensional measurements of color intensity of $n = 4435$ areas. The result of the normal model clustering with $k = 3$:



Detecting the structure of similarities: The k-means method

Landsat data: $\mathbf{x}_1, ..., \mathbf{x}_n$ are $p = 36$ dimensional measurements of color intensity of $n = 4435$ areas. The result of the k-means algorithm with $k = 3$:



Classifying data: General principles

We are dealing with samples $\mathbf{x}_1, ..., \mathbf{x}_n$ from k distinct distributions $\mathcal{F}_1, ..., \mathcal{F}_k$ of high dimension p with "class labels" $y_1, ..., y_n \in \{1, ..., k\}$.

$object_1$: $\mathbf{x}_1 = (x_{1,1}, x_{1,2}, ..., x_{1,p})' \rightarrow y_1$
$object_2$: $\mathbf{x}_2 = (x_{2,1}, x_{2,2}, ..., x_{2,p})' \rightarrow y_2$
...
$object_n$: $\mathbf{x}_n = (x_{n,1}, x_{n,2}, ..., x_{n,p})' \rightarrow y_n$

In contrast to the previous situation, we know the number k of classes and we also know which data come from which class.

General aim of classification methods:

Based on the data $\mathbf{x}_1, ..., \mathbf{x}_n$ and their classifications $y_1, ..., y_n$ ("training set") find a rule for classifying new data. The rule has the form of a decomposition of $\mathbb{R}^p$ into regions $R_1, ..., R_k$ with the interpretation that a vector $\mathbf{x}$ that belongs to R_i is classified to the i-th class.

Different methods construct the decomposition $R_1, ..., R_k$ in different ways.

Classifying data: Linear discriminant analysis

For simplicity let the number of classes be $k = 2$ and let f_1 and f_2 be densities of the distributions $\mathcal{F}_1$ and $\mathcal{F}_2$. The idea is to decompose $\mathbb{R}^p$ into regions R_1, R_2 such that we minimize the mean value of a loss from misclassification:

$$E(R_1, R_2) = p_2 C(1|2) \int_{R_1} f_2(\mathbf{x}) \mathrm{d}\mathbf{x} + p_1 C(2|1) \int_{R_2} f_1(\mathbf{x}) \mathrm{d}\mathbf{x},$$

where p_1, p_2 are prior expectations of an observation from classes 1 and 2, and $C(1|2), C(2|1)$ are costs of an erroneous classification of an observation from the class 2 to the class 1 and vice versa.

This optimization problem can be solved and the solution is

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2 C(1|2)}{p_1 C(2|1)} \right\}; \; R_2 = \mathbb{R}^p \setminus R_1.$$

Classifying data: Linear discriminant analysis

If f_1 and f_2 are the densities of the p-dimensional normal distributions with the means μ_1 and μ_2 and the same covariance matrix Σ, then R_1 and R_2 are half-spaces:

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p : \mathbf{x}' \Sigma^{-1} (\mu_1 - \mu_2) \geq c \right\}, \; R_2 = \mathbb{R}^p \setminus R_1$$

where

$$c = \frac{1}{2} \left(\mu_1' \Sigma^{-1} \mu_1 - \mu_2' \Sigma^{-1} \mu_2 \right) + \ln \frac{p_2 C(1|2)}{p_1 C(2|1)}.$$

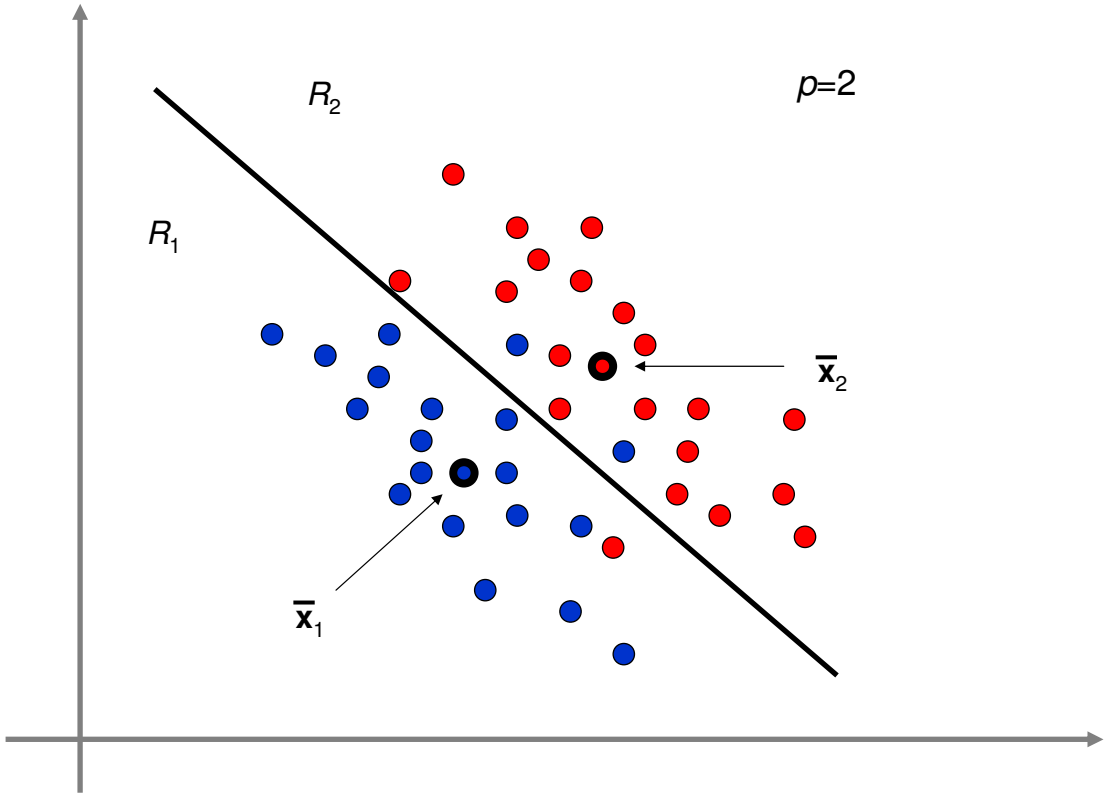
We construct the actual classifier from the training set such that in the previous formula we substitute μ_1 and μ_2 with the means $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ of the classes of data and Σ with a pooled sample covariance matrix $\mathbf{S}$.

This is called a linear discriminant analysis (LDA).

Classifying data: LDA

Geometrical view: Let $p_1 = p_2 = 1/2$ and $C(1|2) = C(2|1)$. Then we classify $\mathbf{x}$ based on the distance to the means $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, but the distance is in the "Mahalanobis sense", reflecting the "shape" of the data:

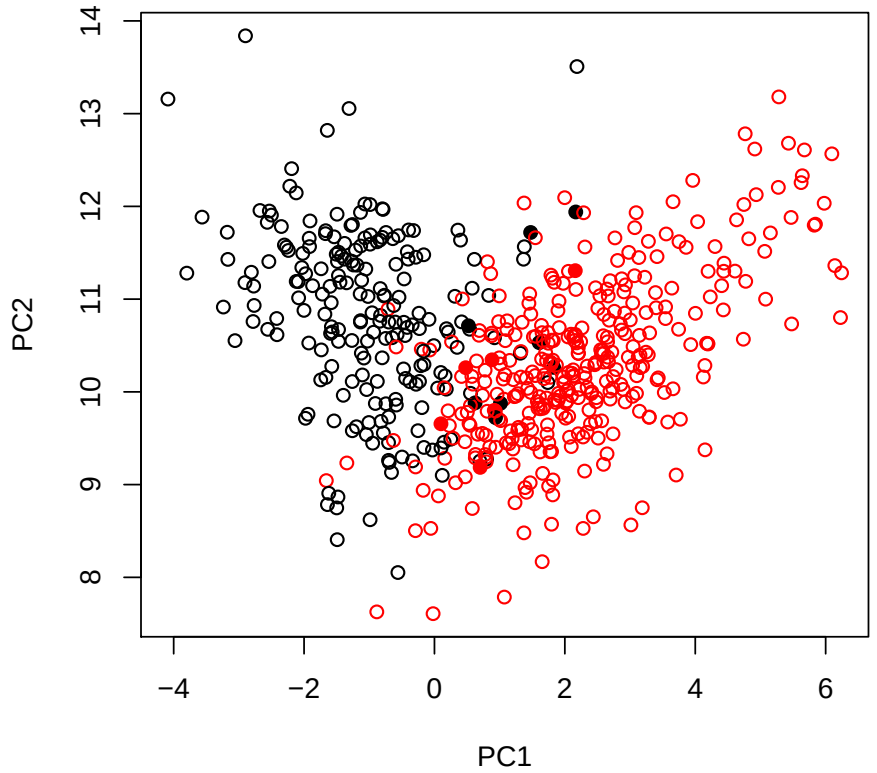
$$\mathbf{x} \in R_1 \Leftrightarrow (\mathbf{x} - \bar{\mathbf{x}}_1)'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_1) \leq (\mathbf{x} - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_2).$$



Classifying data: LDA

Breast cancer data: $\mathbf{x}_1, ..., \mathbf{x}_n$ are measurements of $p = 30$ characteristics of tumor cells of $n = 569$ patients. We will use LDA with $p_1 = 357/569$, $p_2 = 212/569$, $C(1|2) = 2$, $C(2|1) = 1$.

Legend: red = class 1 (benign), black = class 2 (malignant), empty circles = correctly classified, full circles = incorrectly classified.



Classifying data: Other methods

- **Partitioning trees**: Classify the data into categories by asking questions about the values of x_i's in a "tree-like" manner. Decomposes $\mathbb{R}^p$ into regions $R_1, ..., R_k$ that are rectangular. Advantage is that the classification is simple.
- **Support vector machines**: In the simplest version classifies the data into two classes based on values of a linear function of the variables. This method divides $\mathbb{R}^p$ into two half spaces R_1, R_2 by the so-called "maximal margin separating hyperplane". Numerically leads to a convex quadratic programming problem.
- **Neural Networks**: Classifies the data into k classes using a function $f : \mathbb{R}^p \to \{1, ..., k\}$ that can be evaluated by a network of simple computational units (artificial neurons). Decomposes $\mathbb{R}^p$ into highly complex classes $R_1, ..., R_k$ depending on the structure of the network, "synaptic weights" and neuronal "thresholds".

Conclusions

- With the advance of modern electronic devices enormous datasets become the rule rather than the exception.
- For the large multidimensional datasets, the classical assumptions, models and simple mathematical techniques are usually unsatisfactory.
- The methods of multivariate statistics shift from simple hypothesis testing to computationally demanding exploration of the usually complicated structure of the data, with complex results that guide our intuition and decisions.

Thank you for attention.