1 Why do we need multivariate statistics?

Multivariate statistical analysis is referring to the analysis of observations which have been observed simultaneously on several variables. This is in contrast to univariate statistical analysis, where we only have observations of a single variable, and we analyze the statistical behavior of these univariate data. In order to analyze multivariate data, we need appropriate tools, called multivariate statistical methods. Multivariate statistics is thus the extension of univariate statistics to more than one dimension

2 What is the Spectral Decomposition Theorem?

Theorem 1.4.3 (Spectral Theorem or eigendecomposition)
Every symmetric matrix $\boldsymbol{\Sigma}$ *of order* $(p \times p)$ *can be decomposed as*

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma A \Gamma}^{\top} = \sum_{i=1}^{p} a_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^{\top} \quad , \tag{1.7}$$

where $\boldsymbol{A} = Diag(a_1, \ldots, a_p)$ *is a diagonal matrix with the eigenvalues of* $\boldsymbol{\Sigma}$*, and* $\boldsymbol{\Gamma}$ *is an orthogonal matrix (i.e.* $\boldsymbol{\Gamma}^{\top} = \boldsymbol{\Gamma}^{-1}$*), where the columns* $\boldsymbol{\gamma}_i$ *of* $\boldsymbol{\Gamma}$*,* $i = 1, \ldots, p$*, are standardized eigenvectors of* $\boldsymbol{\Sigma}$*.*

3 What is the density of the multivariate normal distribution?

Recall the density of the univariate normal distribution,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < \infty \ , \tag{1.14}$$

with expectation μ and variance σ^2.

The density of the multivariate normal distribution is obtained by replacing the distance $(x - \mu)/\sigma$ by a multivariate distance

$$(\boldsymbol{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) := \mathrm{MD}^2(\boldsymbol{x}) \ , \tag{1.15}$$

also called (squared) *Mahalanobis distance*. Moreover, the constant $\frac{1}{\sqrt{2\pi\sigma^2}}$ is replaced by a term such that the volume described by the multivariate density function is standardized to 1.

Thus, let $\boldsymbol{x} = (x_1, \ldots, x_p)^{\top}$ be a p-dimensional random variable with expectation $E(\boldsymbol{x}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\top}$ and covariance $Cov(\boldsymbol{x}) = \boldsymbol{\Sigma} = [(\sigma_{ij})]$. Assume that $\boldsymbol{\Sigma}$ is positive definite (denoted by $\boldsymbol{\Sigma} \geq \boldsymbol{O}$), which is the case if (and only if) all eigenvalues are strictly positive.

4 Name distances for clustering, methods, and their respective objective functions/criteria.

In the following we assume p-variate observations $\boldsymbol{x}_{i\cdot} = (x_{i1}, \ldots, x_{ip})^\top$, for $i = 1, \ldots, n$, sometimes simply called objects.

A widely used *distance or dissimilarity measure* between the i-th and the j-th object is the **Euclidean distance**

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} = \|\boldsymbol{x}_{i\cdot} - \boldsymbol{x}_{j\cdot}\| \ , \tag{2.1}$$

also known under the term L_2 norm distance. An alternative, being less sensitive to outliers, is the **Manhattan distance** (or "city-block" distance), defined as

$$d(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| = \|\boldsymbol{x}_{i\cdot} - \boldsymbol{x}_{j\cdot}\|_1 \ , \tag{2.2}$$

also called L_1 norm distance.

The distance can be computed for every observation pair, which results in an $n \times n$ distance matrix $\boldsymbol{D} = [(d_{ij})]$. Clearly, this matrix is symmetric, and the main diagonal elements are all zero.

Hierarchical Clustering Methods:

The "similarity" of the combined pair can be measured, and a "height" is associated with this newly formed class. At the end of the process there is only one single cluster left.

This calls for a new definition of similarity, expressing the distances between a group of observations in one cluster with indexes in the set C_k, and another group in the second cluster with indexes in C_l. The number of observations in each group is n_k and n_l, respectively. Now we can define different distance measures between clusters, which are also reflecting the name of the corresponding clustering algorithms:

- Complete Linkage:

$$\max_{i \in C_k, j \in C_l} d(i,j)$$

  The similarity between two clusters is thus defined by that pair of objects from the different clusters which has the biggest distance.

- Single Linkage:

$$\min_{i \in C_k, j \in C_l} d(i,j)$$

  Here, the similarity is given by the closest observations from two clusters. Single linkage tends to be unbalanced in the sense that big clusters are quickly combined. This procedure tends to produce many small groups and few large groups. Single linkage is also suitable to detect outliers.

- Average Linkage:

$$\frac{1}{n_k n_l} \sum_{i \in C_k} \sum_{j \in C_l} d(i,j)$$

  The similarity is defined as the average of all pairwise distances.

- Centroid method:

  Here, one first needs to compute the arithmetic means (vectors) of the observations of each cluster, say $\bar{\boldsymbol{x}}(C_k)$ and $\bar{\boldsymbol{x}}(C_l)$. Then the similarity between the two clusters is given by the Euclidean distance of the cluster centers (centroids),

- Ward's method:

  The similarity between two clusters is defined as the increase of the variance when merging the two clusters,

$$\frac{\|\bar{\boldsymbol{x}}(C_k) - \bar{\boldsymbol{x}}(C_l)\|^2}{1/n_k + 1/n_l} \ .$$

Those two clusters will be merged where the increase is the smallest possible.

For all of the above methods, the similarity needs to be computed for every pair of clusters, and then the pair with the smallest values of the measure will be merged. The resulting hierarchy can be presented in a **dendrogram**, where the horizontal axis presents the observations and the vertical axis the "height", which

5 Partitioning Methods

The best known algorithm for creating partitions is the **K-means algorithm**. Consider an index set C_k containing the indexes of the observations of the k-th cluster. One can define the so-called *total point scatter* as

$$T = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} d^2(i,j) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\left(\sum_{j\in C_k} d^2(i,j) + \sum_{j\notin C_k} d^2(i,j)\right) \ .$$

It is then possible to decompose T as $T = W(C) + B(C)$ into a *within-cluster* point scatter

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j\in C_k} d^2(i,j)$$

and a *between-cluster* point scatter

$$B(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in C_k}\sum_{j\notin C_k} d^2(i,j)$$

for a given a cluster partition C. Then, $B(C)$ tends to be large when observations assigned to different clusters are far apart. Thus, it is desirable to find a partition (for fixed K) which maximizes $B(C)$, which is equivalent to minimizing $W(C)$.

For K-means clustering one takes the Euclidean distance $d(i,j) = \|\boldsymbol{x}_{i\cdot} - \boldsymbol{x}_{j\cdot}\|$ as the distance measure. Then one can rewrite $W(C)$ from above as

$$W(C) = \sum_{k=1}^{K} n_k \sum_{i \in C_k} \|\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}}_k\|^2 \,, \tag{2.3}$$

where n_k is the number of observations in the k-th cluster, and $\bar{\boldsymbol{x}}_k$ is the arithmetic mean vector of those observations. Minimizing this criterion means that the n observations are assigned to the K clusters in a way that the (average) distances of the observations from a cluster to their cluster center are minimized. This can be solved by an iterative algorithm:

Step 1. Cluster centers $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$ are initialized. Usually, this is done by randomly selecting K observations as initial cluster centers.

Step 2. Minimize the objective function (2.3) by assigning each observation to the closest cluster center.

Step 3. Minimize (2.3) by replacing the cluster centers from Step 2. by the arithmetic means of the observations per cluster.

Step 4. Repeat Steps 2. and 3. until the assignments do not change.

Each of the Steps 2. and 3. reduce the value of $W(C)$ and thus convergence of this procedure is assured. However, one could end up in a local optimum, and it is thus recommended to re-start the procedure with different random initializations, and take that solution which gives the smallest value $W(C)$.

6 Explain model-based clustering and difficulties that could occur.

Model-based clustering can also be used to obtain a cluster partition, but the result would even give a "probability" for the assignment of an observation to a cluster. As the name indicates, model-based clustering makes use of a statistical model for the shape of the clusters. The standard "model" is multivariate normal distribution, i.e., it is assumed that the cluster has the density of a multivariate normal distribution, with a certain location and covariance. This is a big advantage over K-means clustering, since the cluster shapes can be more flexible. The result of K-means is typically spherically shaped clusters (due to the use of Euclidean distances to the center).

A detailed description of model-based clustering can be found in Fraley and Raftery (2002), and in many other sources of these authors. Assume that the data consist of K clusters, generated by multivariate normal densities with expectation $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$, for $k = 1, \ldots, K$. Further, the class probabilities are given by so-called mixing coefficients $\pi_1, \ldots, \pi_K$, where $\pi_1 + \ldots + \pi_K = 1$. All these parameters are unknown, and they are estimated using the EM (expectation maximization) algorithm. Note that the covariance matrices are $p \times p$ matrices, and for larger p

there are many parameters to estimate from the available data, which can lead to instability. For this reason, the cluster "models" can be simplified, by imposing restrictions on the cluster covariance structures.

The simplest possibility for such restrictions is $\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}$, for $k = 1, \ldots, K$, where $\boldsymbol{I}$ is the identity matrix and σ^2 is a parameter for the variance. This would imply that all clusters are spherical, with the same radius. The estimation of the covariances thus reduces to estimating only one parameter, the variance σ^2. A less restricted covariance structure is $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$, for $k = 1, \ldots, K$. In this case, the clusters are still spherical, but their size can be different according to their variance σ_k^2, which needs to be estimated. Figure 2.3 illustrates different covariance structures.

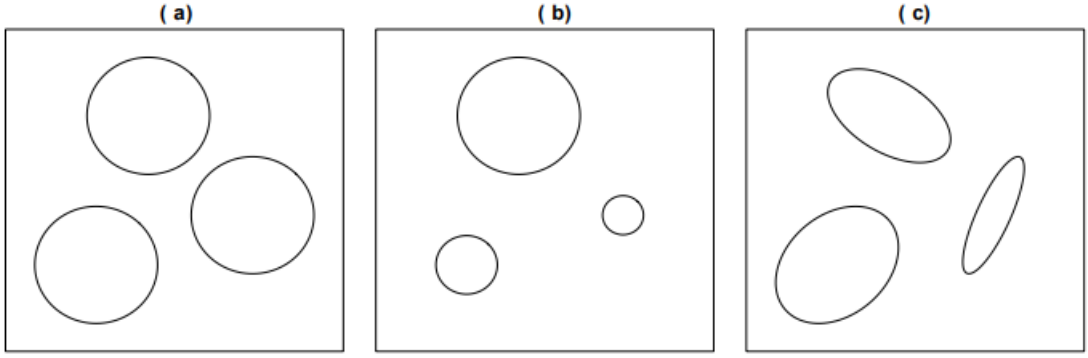


Figure 2.3: Different covariances for three clusters: (a) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \sigma^2 \boldsymbol{I}$; (b) $\boldsymbol{\Sigma}_j = \sigma_j^2 \boldsymbol{I}$, for $j = 1, 2, 3$; (c) all $\boldsymbol{\Sigma}_j$ different and of no special structure.

7 Explain fuzzy clustering, what is the objective function?

Partitioning methods are sometimes called hard clustering methods, since they assign an observation to a cluster or not (1 or 0). In contrast, fuzzy clustering methods.

ods allow for a proportional assignment of an observation to all clusters, where the sum of the proportions is 1. The coefficients for the proportional assignments are called *membership coefficients* $u_{ik} \in [0, 1]$, for $i = 1, \ldots, n$ and $k = 1, \ldots, K$, with $\sum_{k=1}^{K} u_{ik} = 1$ for all i.

The best-known fuzzy clustering algorithm is called **fuzzy K-means** algorithm (Bezdek, 1974; Dunn, 1974), and it works very similar to the K-means procedure. First of all, K has to be given. The objective function (2.3) of K-means clustering is replaced by

$$\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{m}_k\|^2 \,, \tag{2.4}$$

which has to be minimized. Here, $\boldsymbol{m}_k = (m_{k1}, \ldots, m_{kp})^\top$ is the weighted cluster center of cluster k, defined as

$$m_{kj} = \frac{\sum_{i=1}^{n} u_{ik}^2 x_{ij}}{\sum_{i=1}^{n} u_{ik}^2}$$

for $j = 1, \ldots, p$.

One can show that the following equality holds:

$$\sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{m}_k\|^2 = \sum_{k=1}^{K} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{ik}^2 u_{jk}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|^2}{2 \sum_{j=1}^{n} u_{jk}^2} \tag{2.5}$$

This shows that the squared Euclidean distances between the observations enters the objective function, and thus it is also possible to use the distance matrix rather than the data matrix as an input of the procedure.

8 How can we evaluate clustering solutions -- Hetero/Homogeneity, Calinski-Harabasz, Hartigan, silhouette width, Gap statistic (principles)?

The main goal of cluster analysis is to achieve highly homogeneous clusters, i.e. the observations within a cluster should be very similar to each other. On the other hand, different clusters should be dissimilar, because otherwise they should have been merged into one cluster. In other words, heterogeneity between different clusters should be achieved. Heterogeneity can be measured by

$$B_K = \sum_{k=1}^{K} \|\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}\|^2, \tag{2.6}$$

where $\bar{\boldsymbol{x}}_k$ is the k-th cluster center $(k = 1, \ldots, K)$, and

$$\bar{\boldsymbol{x}} = \frac{1}{K} \sum_{k=1}^{K} \bar{\boldsymbol{x}}_k$$

is the overall mean of the cluster centers. This term is also called the *between cluster sum of squares*. Homogeneity within the clusters can be defined by

$$W_K = \sum_{k=1}^{K} \sum_{i \in C_k} \|\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}}_k\|^2. \tag{2.7}$$

This term is called the *within cluster sum of squares*, since it considers squared Euclidean distances from the observations to their own cluster center.

While B_K should be large, W_K should be small. However, both measures depend on the number K of clusters, and thus this needs to be considered in a validity measure. Two prominent measures are the **Calinski-Harabasz index**

$$\mathrm{CH}_K = \frac{B_K/(K-1)}{W_K/(n-K)}$$

and the **Hartigan index**

$$\mathrm{H}_K = \ln \frac{B_K}{W_K}.$$

Another prominent validity measure is the **average silhouette width** (Kaufman and Rousseeuw, 1990). Before computing this value, some definitions have to be provided first. The average dissimilarity of an observation $\boldsymbol{x}_{i.}$ belonging to cluster C_k to all other observations of the same cluster is given by

$$d_{i,C_k} = \frac{1}{n_k - 1} \sum_{i,j \in C_k, i \neq j} d^2(i,j),$$

where n_k is the number of observations in cluster C_k. The average dissimilarity of $\boldsymbol{x}_{i.}$ to observations from another cluster C_l is given by

$$d_{i,C_l} = \frac{1}{n_l} \sum_{j \in C_l} d^2(i,j).$$

The smallest of these values is

$$d_{i,C} = \min_l d_{i,C_l},$$

and it corresponds to the smallest dissimilarity of the i-th observation to its "closest" cluster. The *silhouette value* is defined as

$$s_i = \frac{d_{i,C} - d_{i,C_k}}{\max(d_{i,C_k}, d_{i,C})}.$$

The values of s_i are within the interval $[-1, 1]$. If the value of s_i is close to 1, the observation is well classified, a value of zero means that the observation is in between two clusters, and a value of -1 refers to a poor classification. Observations with negative silhouette values are probably assigned to a wrong cluster. The average silhouette width is

$$\frac{1}{n} \sum_{i=1}^{n} s_i,$$

and the higher this value, the better the classification.

Gap Statistic:

The idea is to consider

$$\tilde{W}_K = \sum_{k=1}^{K} \frac{1}{2n_k} \sum_{i,j \in C_k} d^2(i,j),$$

which is the pooled within-cluster sum of squares around the cluster means. The smaller the value $\tilde{W}_K$, the more compact are the points in the clusters – but this also depends on K: if K is very big, the clusters are naturally compact. The task is thus to find a possibly small value of K which still yields a small value of $\tilde{W}_K$.

The Gap statistic is defined as

$$\text{Gap}_n(K) = E_n^*\{\log(\tilde{W}_K)\} - \log(\tilde{W}_K)\ , \tag{2.8}$$

where E_n^* denotes the expectation under a sample of size n from an appropriate reference distribution. For the reference distribution, a unimodal distribution is simulated, generated from a uniform distribution on the hypercube determined by the ranges of the input data. The expectaton is estimated by an average of B (e.g. $B = 100$) copies $\log(\tilde{W}_K^*)$, each of which is computed from a bootstrap sample $\boldsymbol{x}_{1.}^*, \ldots, \boldsymbol{x}_{n.}^*$ of this reference distribution.

Practically, we search for the smallest K such that the Gap statistic yields a local maximum. Since bootstrapping is used, one cannot just compute the average, but also a standard deviation and thus a standard error. Therefore, it is more advisable to look for the smallest K such that the value of the Gap statistic is not more than one standard error away from the first local maximum.

9 What is the least squares estimator?

Based on the observations y_i and $x_{i1}, \ldots, x_{ip}$, for $i = 1, \ldots, n$, we want to estimate the regression parameters and the error variance σ^2 by focusing on the differences $y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}$, which inform about the deviation between the observed response and the model fit. These differences are called **residuals**.

The *least-squares method* minimizes the sum of the squared residuals,

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \ . \end{aligned} \tag{3.3}$$

The coefficient minimizing this criterion is denoted as $\widehat{\boldsymbol{\beta}}$, and it is the least-squares estimator of the regression parameter $\boldsymbol{\beta}$.

We can multiply the terms in Equation (3.3):

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.4}$$

The resulting terms are scalars (not vectors or matrices), and thus we have

$$\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y})^\top = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.5}$$

Thus, Equation (3.3) is equal to

$$S(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.6}$$

The partial derivative with respect to the vector $\boldsymbol{\beta}$ yields

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 - 2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.7}$$

Setting this equation to zero gives the least-squares estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad . \tag{3.8}$$

Remark: If $\boldsymbol{X}$ does not have full rank $p + 1 \leq n$, we cannot compute $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, and a way out would be to use a generalized inverse (e.g. Moore-Penrose) $(\boldsymbol{X}^\top \boldsymbol{X})^-$.

With $\widehat{\boldsymbol{\beta}}$ one can now compute the *fitted values* $\widehat{\boldsymbol{y}}$ of $\boldsymbol{y}$ by

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \ , \tag{3.9}$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ is called "hat"-matrix. The *estimated residuals* are

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \ . \tag{3.10}$$

We can find that $\boldsymbol{X}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{0}$ and $\widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \widehat{\boldsymbol{\varepsilon}} = 0$. This means that the estimated residuals are orthogonal to the columns of $\boldsymbol{X}$ and to the fitted values $\widehat{\boldsymbol{y}}$.

10 How does multivariate linear regression work, what is the objective function, solution and appropriate inference (estimation of covariance of errors). Basic model selection?

While in the multiple linear regression model we considered one response variable, we have m response variables $y_1, \ldots, y_m$ in the multivariate regression case. Thus, it is possible to construct for each single response variable a regression model based on the explanatory variables $x_1, \ldots, x_p$:

$$\begin{array}{rcl} y_1 & = \beta_{01} + \beta_{11}x_1 + \ldots + \beta_{p1}x_p + \varepsilon_1 \\ \vdots & \vdots \\ y_j & = \beta_{0j} + \beta_{1j}x_1 + \ldots + \beta_{pj}x_p + \varepsilon_j \\ \vdots & \vdots \\ y_m & = \beta_{0m} + \beta_{1m}x_1 + \ldots + \beta_{pm}x_p + \varepsilon_m \end{array} \tag{3.12}$$

For the error term $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ we now have the assumptions $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. This means that the error terms for the different responses can be correlated with each other.

Consider now a sample of size n, then we can write down the regression model separately for every response variable $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})^\top$ $(j = 1 \ldots, m)$ in the same way as in the previous section:

$$\begin{array}{rcl} y_{1j} & = \beta_{0j} + \beta_{1j}x_{11} + \ldots + \beta_{pj}x_{1p} + \varepsilon_{1j} \\ y_{2j} & = \beta_{0j} + \beta_{1j}x_{21} + \ldots + \beta_{pj}x_{2p} + \varepsilon_{2j} \\ \vdots & \vdots \\ y_{nj} & = \beta_{0j} + \beta_{1j}x_{n1} + \ldots + \beta_{pj}x_{np} + \varepsilon_{nj} \end{array}$$

All these equations can be collected in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{3.13}$$

with the matrices

- $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m)$ of dimension $n \times m$,
- $\boldsymbol{X}$, still of dimension $n \times (p+1)$,
- $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ of dimension $(p+1) \times m$,
- $\boldsymbol{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m)$ of dimension $n \times m$.

According to the assumptions in the multiple linear regression model, we have

$$E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$$

$$\mathrm{Cov}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k) = \sigma_{jk}\boldsymbol{I}_n \tag{3.14}$$

for $j, k = 1, \ldots, m$. This means that the m characteristics of the i-th trial ($i = 1, \ldots, n$) have covariance $\boldsymbol{\Sigma} = [(\sigma_{jk})]$, but observations from different trials are uncorrelated.

Also in multiple linear regression we are interested here in deriving the least-squares estimator. Similar as in Equation (3.3), we minimize the sum of squared residuals:

$$S(\boldsymbol{B}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \beta_{0j} - \beta_{1j}x_{i1} - \ldots - \beta_{pj}x_{ip})^2$$

One can show that this is the same as

$$S(\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{Y}^\top\boldsymbol{Y}) - 2\,\mathrm{tr}(\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{B}) + \mathrm{tr}(\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B})\,,$$

where "tr()" denotes the trace of the matrix, i.e. the sum of the diagonal elements. The partial derivative with respect to $\boldsymbol{B}$ yields

$$\frac{\partial S(\boldsymbol{B})}{\partial \boldsymbol{B}} = -2\boldsymbol{X}^\top\boldsymbol{Y} + 2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\,,$$

and this leads to the least-squares estimator

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}\,. \tag{3.15}$$

Interestingly, this is just the same solution as if we would use the least-squares estimator for each single response variable: Denote $\widehat{\boldsymbol{B}} = (\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_m)$, then we have

$$\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}_j\,, \tag{3.16}$$

as the least-squares estimator of $\boldsymbol{\beta}_j$ for the response $\boldsymbol{y}_j$, for $j = 1, \ldots, m$.

The fitted values of the responses are obtained by

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{B}}\,. \tag{3.17}$$

Since the structure of the estimator is still the same as in the multiple linear regression case, still all the results are valid, in particular those of the Gauss-Markov theorem.

11 What are problems with non-robustness? How is it connected to the (empirical) influence function, maxbias curve, breakdown point and efficiency?

The goal of robust statistics is to have less dependency on strict model assumptions. One still works with models and assumptions, but certain deviations and violations are tolerated.

robust methods focus on fitting only the majority of the data, where the requirements need to be fulfilled, but allow for deviations from the minority.

Further, covariance estimation is so crucial in multivariate statistics, so that various methods can be robustified just by plugging in a robust covariance estimate.

Good robust estimators downweight outliers during the estimation, but then provide (robust) diagnostics in order to reveal those observations.

Which properties should a robust estimator have? It should be resistant to a sizeable proportion of outliers or deviation from assumptions. It should also still yield reasonable results if these ideal assumptions are valid. In general,

we are interested in the influence function, the maxbias curve and breakdown point, and in the statistical efficiency.

Influence function:
One of the basic ideas of robustness is that a limited amount of contamination should only have a small effect on the estimator. This can be simply empirically checked by varying single data points and looking at the effect on the estimator. We are interested in the effect of one observation when it is varied along the indicated vertical line. For each position of the data point, we are interested in the change of the slope parameter.
The result is known as the empirical influence function (EIF). Ideally the EIF should be smooth: it should not show local spikes or should not be a step function:

The concept of the EIF can be formalized by the so-called influence function (IF). The IF measures the influence an infinitesimal amount of contamination has on an estimator with respect to its position in space. More precisely, the influence function of an estimator T at a given distribution G is defined as:

$$\mathrm{IF}(\boldsymbol{x}, T, G) = \lim_{\varepsilon \downarrow 0} \frac{T\left[(1-\varepsilon)G + \varepsilon\delta_{\boldsymbol{x}}\right] - T(G)}{\varepsilon}, \tag{4.1}$$

where ε is the fraction of contamination and $\delta_{\boldsymbol{x}}$ is a probability measure which puts all the mass at $\boldsymbol{x}$. For robust estimators, the effect of small contamination will be limited.

Maxbias Curve:
What one expects is that a robust estimator can withstand a certain fraction of contamination. The mathematical tool to examine to which extent an estimator is distorted with respect to the fraction of contamination in the data is the maxbias curve. The maxbias curve measures the bias an estimator has with respect to the percentage of the worst possible type of contamination:

contamination. Let $\boldsymbol{X}$ be the original data set and $\boldsymbol{X}$ be a data set in which m out of n observations have been replaced with arbitrary values, and let $\|\cdot\|$ denote the Euclidean norm. Then the maxbias curve for an estimator T is defined as:

$$\mathrm{maxbias}(m, T, \boldsymbol{X}) = \sup_{\check{\boldsymbol{X}}} \| T(\check{\boldsymbol{X}}) - T(\boldsymbol{X}) \| . \tag{4.2}$$

It is known that for some estimates of regression the worst possible type of outliers is found at points where y, x and the fraction y/x increase to infinity. Non-robust estimators, such as the least-squares regression estimator, turn out to reach a maxbias of infinity already at small amounts of contamination. This brings us to a further important concept.

Breakdown point:

Breakdown point

For every estimators, there exists a point where the maxbias tends to infinity. This point is referred to as the *breakdown point*. Loosely, the breakdown point indicates which percentage of the data may be replaced with outliers before the estimator yields aberrant results. Based on the maxbias curve, the breakdown point of an estimator T at a given sample $\boldsymbol{X}$ is given by:

$$\varepsilon_n^*(T, \boldsymbol{X}) = \min\left\{\frac{m}{n}; \text{maxbias}(m, T, \boldsymbol{X}) = \infty\right\}. \tag{4.3}$$

For $n \to \infty$ one obtains the *asymptotic breakdown point*, denoted ε^*. For least-squares regression it holds that $\varepsilon^* = 0$. The maximal possible value of the asymptotic breakdown point of regression estimators equals 0.5 (if one asks for certain equivariance properties).

One of the goals in designing robust estimators is obtaining a high breakdown point. However, bounded influence and high breakdown should not result in a drastic decrease in efficiency.

Statistical efficiency:

The goal is to design robust estimators which are also efficient under normality, but at the same time achieve high (or at least positive) breakdown point and reasonable behavior of the influence function.

12 What are M-estimators? What are the M-estimating equations? Why is it a weighted least squares estimator?

classical least-squares (LS) estimator is defined as

$$\hat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i(\boldsymbol{\beta})^2 \ .$$

An ancient alternative to LS is the L_1 estimator defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})| \, .$$

4.3.1 M-estimators

One can rewrite Equations (4.5) and (4.6) as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})), \tag{4.7}$$

where $\rho(r) = r^2$ for LS and $\rho(r) = |r|$ for L_1. By taking other ρ functions, different estimators are obtained. In fact, Equation (4.7) is the definition of a whole class of estimators commonly referred to as *M-estimators*.

In order to not depend on the scale of the response, a more suitable definition of M-estimators of regression is

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right), \tag{4.8}$$

where $\hat{\sigma}$ is a robust scale estimator of the residuals, that can be estimated either previously or simultaneously with the regression parameters.

Differentiating (4.8) with respect to $\boldsymbol{\beta}$ we get that the estimate fulfils the system of *M-estimating equations*

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\right) \boldsymbol{x}_i = 0 \tag{4.9}$$

where $\psi = \rho'$.

For LS, $\psi(r) = r$, and (4.9) are the well-known normal equations. We may then in general interpret (4.9) as a robustified version of the normal equations, where the residuals are curbed.

Put $W(r) = \psi(r)/r$ and $w_i = W(r_i(\boldsymbol{\beta})/\hat{\sigma})$. Then (4.9) may be rewritten as

$$\sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \boldsymbol{x}_i = 0\ . \tag{4.10}$$

Then (4.10) is a weighted version of the normal equations, and hence the estimator can be seen as weighted LS, with the weights depending on the data. For LS, W is constant. For an estimator to be robust, observations with large residuals should receive a small weight, which implies that $W(r)$ has to decrease to zero fast enough for large r.

The second row of Figure 4.2 refers to the *Huber* family, with ρ' given by

$$\psi(r) = \begin{cases} r & \text{for} \quad |r| \le b \\ b\ \text{sign}(r) & \text{otherwise} \end{cases} \tag{4.11}$$

The extreme cases $b \to \infty$ and $b \to 0$ correspond to LS and L_1, respectively.

The last row in Figure 4.2 refers to the Tukey *bisquare* family, with

$$\rho(r) = \begin{cases} \left(\frac{r}{k}\right)^2 \left(3 - 3\left(\frac{r}{k}\right)^2 + \left(\frac{r}{k}\right)^4\right) & \text{for} \ \ |r| \le k \\ 1 & \text{else} \end{cases}\ . \tag{4.12}$$

When $k \to \infty$, the corresponding estimate tends to LS and hence becomes more efficient and at the same time less robust. Thus, k is a tuning parameter the choice of which is a compromise between efficiency and robustness. The usual practice is to choose k to attain a given efficiency, such as 0.90.

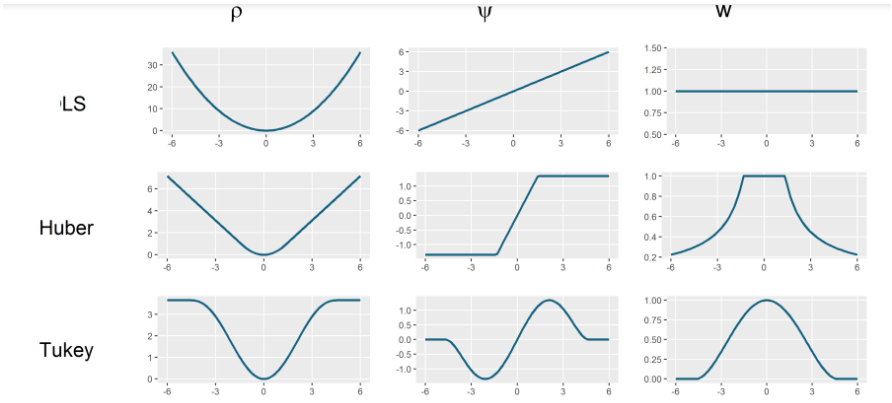


Figure 4.2: Different options for the function ρ, ψ and the weights w.

Note that the Tukey bisquare function has a bounded ρ-function. This is desirable because otherwise if some $\boldsymbol{x}_{i.}$ is large, then the i-th term will dominate the sum in (4.10), which would be unfortunate if $(\boldsymbol{x}_{i.}, y_i)$ is atypical (a so-called bad leverage point). For this reason it is better to use M-estimators given by (4.8) with a *bounded* ρ.

13 What are S-estimators? What is the MM-estimator and the properties inherited from S- and M-estimators?

4.3.2 Regression estimators based on a robust residual scale

Given $\boldsymbol{\beta}$, let $\boldsymbol{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), ..., r_n(\boldsymbol{\beta}))$. We consider an estimator of the form

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \hat{\sigma}(\boldsymbol{r}(\boldsymbol{\beta})) \tag{4.13}$$

where $\hat{\sigma}$ is a robust scale estimator.

Least Median of Squares (LMS) estimator:
The simplest robust scale estimator is the median of the absolute residuals:

$$\hat{\sigma}(\boldsymbol{r}) = \underset{i}{\text{med}} |r_i| \tag{4.14}$$

The corresponding regression estimator achieves highest breakdown point 50%, but a low efficiency. A tradeoff between breakdown and efficiency can be obtained by considering another quantile instead of the median.

Least Trimmed Squares (LTS) estimator:
Call $|r|_{(i)}$ the ordered absolute values of the residuals r_i, i.e. $|r|_{(1)} \leq ... \leq |r|_{(n)}$. A smoother alternative is to consider a scale more similar to the standard deviation, namely the *trimmed squares scale*

$$\hat{\sigma}(\boldsymbol{r}) = \left(\frac{1}{h} \sum_{i=1}^{h} |r|_{(i)}^2 \right)^{1/2}, \tag{4.15}$$

with $h \in [n/2, n]$. Smaller values of h lead to higher breakdown point, but to lower efficiency.

M-estimator of scale
This estimator is defined as the solution σ of an equation of the form

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{r_i}{\hat{\sigma}}\right) = b \tag{4.16}$$

where ρ is a bounded ρ-function and b is a constant. It can be shown that the breakdown point of $\hat{\sigma}$ is $\min(b, 1-b)$. Equation (4.16) is nonlinear, but it is easy to solve iteratively. Put

$$W(z) = \frac{\rho(z)}{z^2}. \tag{4.17}$$

Then (4.16) can be rewritten as

$$\hat{\sigma}^2 = \frac{1}{nb}\sum_{i=1}^{n} w_i r_i^2$$

with $w_i = W(r_i/\hat{\sigma})$, which displays $\hat{\sigma}$ as a weighted RMSE. Given some starting value $\hat{\sigma}_0$, an iterative procedure can be implemented as was done for regression M-estimators.

The choice $\rho(z) = z^2$ and $b = 1$ yields the RMSE. The choice $\rho(z) = \mathrm{I}(|z| > 1)$ and $b = 0.5$ yields $\hat{\sigma} = \mathrm{med}(|r|)$.

Regression estimators with $\hat{\sigma}$ given by (4.16) are called **S-estimators**. Although S-estimators achieve the maximum breakdown point, they have low efficiency. However, they can be used as initial estimator for the M-estimator (4.8). The resulting estimator is called **MM-estimator**; it inherits the breakdown point of the S-estimator, but has controllable efficiency.

14 Define affine equivariance.

4.4.1 Affine equivariance

It is desirable that location and covariance estimates respond in a mathematically convenient form to certain transformations of the data. One can define a transformation that is using a nonsingular $p \times p$ matrix $\boldsymbol{A}$ and a vector $\boldsymbol{b}$ of length p to transform the p-dimensional observations $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}$ by $\boldsymbol{A}\boldsymbol{x}_{j.} + \boldsymbol{b}$. This transformation performs any desired nonsingular linear transformation of the original data. Thus, if $\boldsymbol{t}$ denotes a location estimator, it is requested that

$$\boldsymbol{t}(\boldsymbol{A}\boldsymbol{x}_{1.} + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_{n.} + \boldsymbol{b}) = \boldsymbol{A} \cdot \boldsymbol{t}(\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}) + \boldsymbol{b}, \tag{4.18}$$

and for a covariance estimator $\boldsymbol{C}$ we require

$$\boldsymbol{C}(\boldsymbol{A}\boldsymbol{x}_{1.} + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_{n.} + \boldsymbol{b}) = \boldsymbol{A} \cdot \boldsymbol{C}(\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}) \cdot \boldsymbol{A}^{\top}. \tag{4.19}$$

Location and covariance estimators that fulfil (4.18) and (4.19) are called *affine equivariant* estimators. These estimators transform properly under changes of the origin, the scale, or under rotations.

The property of affine equivariance is only valid for nonsingular transformation matrices. Note that the coordinate-wise median as a robust location estimator is not affine equivariant. Also a robust covariance estimate based on pairwise robust covariances would not be affine equivariant.

15 What is the Minimum Covariance Determinant estimator? How does in work (in concept)? What properties does it have (related to tuning parameter h)?

An estimator of multivariate location and covariance which is affine equivariant and has high breakdown point is the Minimum Covariance

Determinant (MCD) estimator. The idea behind this estimator is in fact related to the LTS estimator. Here, one is searching for those h data points for which the determinant of the empirical covariance matrix is minimal. The location estimator t is the mean of these h observations, and the covariance estimator C is given by the covariance matrix with the smallest determinant, but multiplied by a constant to obtain consistency for normal distribution. The parameter h determines the robustness but also the efficiency of the resulting estimator. The highest possible breakdown point can be achieved if h ≈ n/2 is taken, but this choice leads to a low efficiency. On the other hand, for higher values of h the efficiency increases but the breakdown point decreases. Therefore, a compromise between efficiency and robustness is considered in practice. The data coming from the outlier distribution are inflating the tolerance ellipse based on the classical estimators while that based on the MCD is much more compact and reflects the structure of the majority of data.

Multivariate S-estimators

Similar as in the regression context, see Equation (4.16), it is possible to define S-estimators in the context of robust location and covariance estimation. The idea is to make the Mahalanobis distances small. The squared Mahalanobis distances are defined as

$$\mathrm{MD}^2(\boldsymbol{x}_{i.}, \boldsymbol{t}, \boldsymbol{C}) = (\boldsymbol{x}_{i.} - \boldsymbol{t})^\top \boldsymbol{C}^{-1} (\boldsymbol{x}_{i.} - \boldsymbol{t}) \quad \text{for } i = 1, \ldots, n$$

for a location estimator $\boldsymbol{t}$ and a covariance estimator $\boldsymbol{C}$.

Small Mahalanobis distances can be achieved by using an M-estimator of scale $\hat{\sigma}$, and minimizing

$$\hat{\sigma}(\mathrm{MD}^2(\boldsymbol{x}_{1.}, \boldsymbol{t}, \boldsymbol{C}), \ldots, \mathrm{MD}^2(\boldsymbol{x}_{n.}, \boldsymbol{t}, \boldsymbol{C}))$$

under the restriction that the determinant $|\boldsymbol{C}| = 1$. This restriction avoids a degenerated solution for $\boldsymbol{C}$.

Multivariate MM-estimators

Like in robust regression, a drawback of S-estimators is that their asymptotic efficiency might be rather low. MM estimators for multivariate location and covariance combine both high breakdown point and high efficiency. The resulting estimators are affine equivariant and have bounded influence function. As for the previous estimators, the solution for the estimators can be found by an iterative algorithm.

16 What are problems in classic regression diagnostics (hat matrix)? What are robust regression diagnostics?

LS regression is sensitive to outliers. These could be outliers in the response, so-called "vertical" outliers, or outliers in the explanatory variables, so-called "leverage points". The latter are particularly influential if they have large residual (bad leverage points).

In LS regression, the hat matrix $\boldsymbol{H}$ is used to identify leverage points. The hat matrix is defined as $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, and the LS fit is obtained as $\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{H}\boldsymbol{y}$.

If there is an exact fit, one can assume that this was caused by a leverage point, because an extremely strong influence was caused to the own estimation. In general, one needs to be careful if there are large values of hii. As a rule of thumb one could define the threshold hii > 2 · p+1 n for the identification of leverage points.

What is the reason for the masking effect? In fact, one can show that the following relationship holds in case of regression with intercept:

$$h_{ii} = \frac{1}{n-1}\mathrm{MD}_i^2 + \frac{1}{n} \tag{4.20}$$

Here,

$$\mathrm{MD}_i^2 = (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})$$

Regression diagnostics

The different types of outliers we would like to distinguish are illustrated in the case of simple linear regression schematically in Figure 4.9. These are:

- Regular observations: $\boldsymbol{x}_{i.}$ is in the usual data range, and y_i fits to the model.
- Vertical outliers: $\boldsymbol{x}_{i.}$ is in the usual data range, but y_i does not fit to the model.
- Good leverage points: $\boldsymbol{x}_i$ is an outlier, thus unusual in the x-space, but y_i fits to the model.

- Bad leverage points: $\boldsymbol{x}_i$ is an outlier, thus unusual in the x-space, and y_i does not fit to the model.

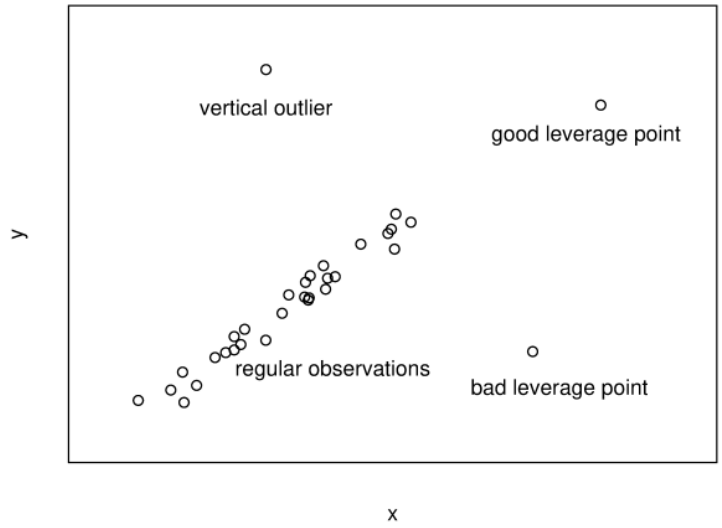


Figure 4.9: Different types of outliers, illustrated in simple linear regression.

In general, good leverage points have the advantage that they are along the regression line (hyperplane) and thus they even allow for a more accurate estimation of the regression parameters. Bad leverage points can have a strong impact on the (LS) estimation, and they can even lead to a leverage of the regression line (hyperplane).

The regression diagnosic plot allows to distinguish these 4 types of observations. Outlyingness in the x-space can be recognized by robust Mahalanobis distances. Large (absolute) residuals can be recognized as scaled deviations in the y-space from the robust regression fit. Thus, the robust residuals need to be scaled by a

robust estimate of the residual scale. Then one can argue according to the normal theory that scaled residuals outside $[-2, 2]$ or $[-2.5, 2.5]$ are extreme and thus very unusual.

Figure 4.10 shows the resulting regression diagnostic plot, where the robust Mahalanobis distances are plotted against the robust scaled residuals. The horizontal lines are at the thresholds ± 2.5 for the scaled residuals, and the vertical line is at the threshold based on the chi-square quantile.

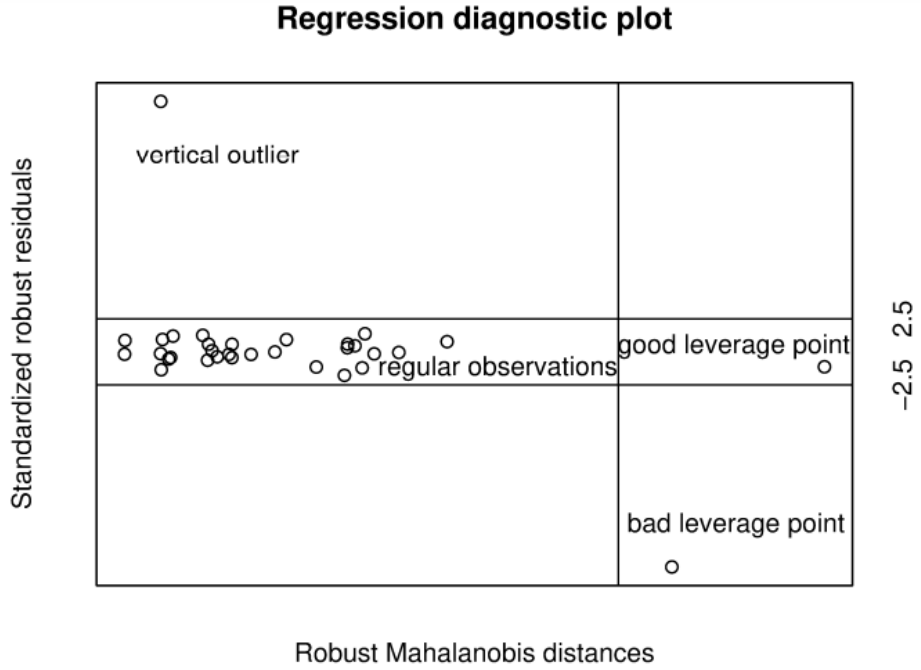


Figure 4.10: Identication of the four different categories of observations in the regression diagnostic plot.

17 How does robust multivariate regression work? (estimate covariance matrix with M-estimator of scale)

4.6 Robust multivariate regression

Consider again the multivariate regression problem with the observations $\boldsymbol{y}_{i.} = (y_{i1}, \ldots, y_{im})^\top$, $\boldsymbol{x}_{i.} = (1, x_{i1}, \ldots, x_{ip})^\top$, and the error terms $\boldsymbol{e}_{i.} = (e_{i1}, \ldots, e_{im})^\top$, and the model

$$\boldsymbol{y}_{i.} = \boldsymbol{B}^\top \boldsymbol{x}_{i.} + \boldsymbol{e}_{i.} \ ,$$

for $i = 1, \ldots, n$, see also Equation (3.13). We are interested in robustly estimating the $(p+1) \times m$ matrix $\boldsymbol{B}$ of regression coefficients, as well as the covariance matrix $\boldsymbol{\Sigma}$ of the error terms. For this purpose, one can use multivariate S-estimators, see Section 4.4.3, and Van Aelst and Willems (2005). Here, the Mahalanobis distances are based on the residuals $\boldsymbol{r}_{i.} = \boldsymbol{y}_{i.} - \boldsymbol{B}^\top \boldsymbol{x}_{i.}$ and on an estimator $\boldsymbol{C}$ of $\boldsymbol{\Sigma}$. Thus, one has to minimize

$$\hat{\sigma}(\boldsymbol{r}_{1.}^\top \boldsymbol{C}^{-1} \boldsymbol{r}_{1.}, \ldots, \boldsymbol{r}_{n.}^\top \boldsymbol{C}^{-1} \boldsymbol{r}_{n.})$$

using an M-estimator of scale $\hat{\sigma}$, under the contraint $|\boldsymbol{C}| = 1$.

18 Principal Component Analysis - how to select the vectors for the transformation, Lagrange problem definition.

Aim is to describe complex relationships in given data in a simpler form. The data are represented by linear combinations of specific components in a way to preserve as much information as possible. Thus, the dimensionality is reduced to the number of those components.

5.2 Definition of principal components

Let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ be a p-dimensional random vector with expectation $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance matrix

$$\boldsymbol{\Sigma} = E\Big[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\Big] \ . \tag{5.1}$$

Further, $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ is a $(p \times p)$ matrix with fixed values (non-random), with the constraint that its column vectors $\boldsymbol{\gamma}_i$ are unitary vectors, i.e. $\boldsymbol{\gamma}_i^\top \boldsymbol{\gamma}_i = 1$ for $i = 1, \ldots, p$. Moreover, different columns of $\boldsymbol{\Gamma}$ are orthogonal, i.e. $\boldsymbol{\gamma}_i^\top \boldsymbol{\gamma}_j = 0$ for $i \neq j$. This implies that $\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^{-1}$.

Consider the linear transformation

$$\boldsymbol{z} = \boldsymbol{\Gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.2}$$

or, expressed in components,

$$z_i = \boldsymbol{\gamma}_i^\top (\boldsymbol{x} - \boldsymbol{\mu}) \qquad \text{for} \qquad i = 1, \ldots, p \ . \tag{5.3}$$

The result of the above transformation is a new random variable $\boldsymbol{z}$ of dimension p. The variance of z_i $(i = 1, \ldots, p)$ is

$$\mathrm{Var}(z_i) = E\Big[\boldsymbol{\gamma}_i^\top (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\gamma}_i\Big] = \boldsymbol{\gamma}_i^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_i \ . \tag{5.4}$$

So far it is not clear which matrix $\boldsymbol{\Gamma}$ we should choose – in spite of the constraints, there are still infinitely many possibilities. Since we are interested in preserving information content, we should look at the variance of the transformed variables. In other words, we would like to obtain such a transformation which maximizes the variances of the components of $\boldsymbol{z}$.

Considering also the contraints on $\boldsymbol{\Gamma}$, we can mathematically formulate a maximization problem in terms of Lagrange optimization.

First principal component:

For $i = 1$ we would like to obtain a component z_1 such that $\mathrm{Var}(z_1)$ is maximized, and $\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 = 1$. The corresponding Lagrangian problem can be stated as:

$$\phi_1 = \boldsymbol{\gamma}_1^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_1 - a_1(\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 - 1) \tag{5.5}$$

The partial derivatives with respect to the unknowns $\boldsymbol{\gamma}_1$ are set equal to zero, and we obtain

$$\frac{\partial \phi_1}{\partial \boldsymbol{\gamma}_1} = 2\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 - 2a_1\boldsymbol{\gamma}_1 = \boldsymbol{0} \tag{5.6}$$

or

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 = a_1\boldsymbol{\gamma}_1 \ . \tag{5.7}$$

This is an eigenvector/eigenvalue problem, and the solution is that the unknown coefficient vector $\boldsymbol{\gamma}_1$ is an eigenvector of the covariance matrix $\boldsymbol{\Sigma}$ to the eigenvalue a_1.

Our problem, however, is that $\boldsymbol{\Sigma}$ has p eigenvectors in total. Which one should we take?

With Equation (5.7) we can see that

$$\mathrm{Var}(z_1) = \boldsymbol{\gamma}_1^\top (\boldsymbol{\Sigma}\boldsymbol{\gamma}_1) = \boldsymbol{\gamma}_1^\top (a_1\boldsymbol{\gamma}_1) = a_1 \boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 = a_1,$$

and since we want to maximize variance, we take that eigenvector $\boldsymbol{\gamma}_1$ corresponding to the largest eigenvalue a_1. Component z_1 now is denoted as *first principal component (PC)*, and $\boldsymbol{\gamma}_1$ is the direction of this component.

Second principal component:

In the next step, for $i = 2$, we again want to maximize variance, more clearly, $\mathrm{Var}(z_2)$ should be maximized under the constraint $\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_2 = 1$. In addition we also want that z_1 and z_2 are uncorrelated (in order to uncover new information in z_2). The latter condition means:

$$Cov(z_1, z_2) = Cov(\boldsymbol{\gamma}_1^\top (\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{\gamma}_2^\top (\boldsymbol{x} - \boldsymbol{\mu})) = Cov(\boldsymbol{\gamma}_1^\top \boldsymbol{x}, \boldsymbol{\gamma}_2^\top \boldsymbol{x}) = \boldsymbol{\gamma}_1^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_2 =$$

$$= \boldsymbol{\gamma}_2^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2^\top a_1 \boldsymbol{\gamma}_1 = a_1 \boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_1 = 0$$

Since $a_1 \neq 0$, the condition of uncorrelated components is equivalent to orthogonality of $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. That's a remarkable fact which you don't easily get for other methods.

Now we can again formulate the Lagrangian problem,

$$\phi_2 = \boldsymbol{\gamma}_2^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_2 - a_2(\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_2 - 1) - b\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_1 \tag{5.8}$$

with the Lagrange coefficients a_2 and b. The partial derivatives with respect to the unknowns $\boldsymbol{\gamma}_2$ are set equal to zero, which yields:

$$\frac{\partial \phi_2}{\partial \boldsymbol{\gamma}_2} = 2\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 - 2a_2\boldsymbol{\gamma}_2 - b\boldsymbol{\gamma}_1 = \mathbf{0} \tag{5.9}$$

Multiplication from the left-hand side with $\boldsymbol{\gamma}_1^\top$ result in

$$2\boldsymbol{\gamma}_1^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_2 - 2a_2\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_2 - b\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 = 0 - 0 - b \cdot 1 = 0 \ ,$$

and thus $b = 0$. Therefore we can reduce (5.9) to

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 = a_2\boldsymbol{\gamma}_2 \ ,$$

and thus $\boldsymbol{\gamma}_2$ is eigenvector to $\boldsymbol{\Sigma}$ to the next largest eigenvalue a_2, and z_2 is called *second PC*.

19 What is the expectation of the Principal Components?

5.4 PCs based on data

Consider an $n \times p$ data matrix $\boldsymbol{X}$ with n observations and p variables. For the PCA transformation we need to estimate the expectation vector and the covariance matrix. This can be done by the empirical estimates, the sample mean and the sample covariance matrix,

$$\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i\cdot} \ , \tag{5.20}$$

and

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}})^\top \ , \tag{5.21}$$

where $\boldsymbol{x}_{i\cdot}$ is the i-th row of $\boldsymbol{X}$.

In analogy to Equation (5.11), the sample principal components are computed by

$$\boldsymbol{Z} = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\Gamma}} \tag{5.22}$$

or

$$\boldsymbol{z}_j = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\gamma}}_j \qquad \text{for} \qquad j = 1, \ldots, p \ . \tag{5.23}$$

Here, $\mathbf{1}$ is a column vector with n entries of 1, and $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ is the matrix of eigenvectors of $\boldsymbol{S}$. Similar to the population definition of PCA, we obtain a decomposition

$$\hat{\boldsymbol{\Gamma}}^\top \boldsymbol{S} \hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{A}} \ , \tag{5.24}$$

where $\hat{\boldsymbol{A}} = \mathrm{Diag}(\hat{a}_1, \ldots, \hat{a}_p)$ is a diagonal matrix with the eigenvalues to the corresponding eigenvectors of $\boldsymbol{S}$, arranged in descending order.

The matrix $\boldsymbol{Z}$ of PCs has the same dimension as the data matrix $\boldsymbol{X}$. In fact, it just represents the data information in an orthgonally rotated coordinate system.

In this representation, the data are centered, and thus the original data center could not be reconstructed. The elements z_{ij} of the matrix $\boldsymbol{Z}$ are called (PC) *scores*.

In analogy to Equation (5.17) we can compute the correlations between the variables and the PCs as

$$\hat{\lambda}_{ij} = \frac{\hat{\gamma}_{ij}\hat{a}_j^{\frac{1}{2}}}{s_{ii}^{\frac{1}{2}}} \qquad \text{for} \qquad i, j = 1, \ldots, p \ , \tag{5.25}$$

where s_{ii} is the i-th diagonal element of $\boldsymbol{S}$. In matrix notation, this correlation is

$$\hat{\boldsymbol{\Lambda}} = \Big(\mathrm{Diag}(\boldsymbol{S})\Big)^{-\frac{1}{2}} \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{A}}^{\frac{1}{2}} \ . \tag{5.26}$$

The contribution of the variables to the PCs are the loadings.

The empirical variances of the PCs are equal to the eigenvalues.

20 **Why is PCA sensitive to scale? What happens if we center-scale the data?**

Remark: We can see that the PC transformation is not scale invariant, which means that the resulting PCs depend on the scale (units) of the variables. If we would first scale (and probably also center) the variables to mean 0 and variance 1,

$$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} \quad \text{for} \quad i = 1, \ldots, p, \tag{5.12}$$

we would in general obtain different PCs, since then we would perform the eigenvector/eigenvalue decomposition on the correlation matrix, and not on the covariance matrix. Therefore, if scale invariance is desired (and in most applications it is), the variables need to be standardized first.

Since our random variable $\boldsymbol{x}$ has been centered, we obtain that the expectation of the PCs is zero:

$$E(\boldsymbol{z}) = \boldsymbol{\Gamma}^\top \Big[E(\boldsymbol{x} - \boldsymbol{\mu}) \Big] = \mathbf{0} \tag{5.13}$$

The covariance matrix of the PCs is

$$\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{\Gamma}^\top \mathrm{Cov}(\boldsymbol{x} - \boldsymbol{\mu}) \boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\top \boldsymbol{\Sigma} \boldsymbol{\Gamma} = \boldsymbol{A} \ . \tag{5.14}$$

21 What are some rules for the number of principal components to select? (for hypothesis tests: only concept, not formulas)

The answer might depend on the purpose, what the user wants to do with the PCs. If it is for visual inspection of the data, it might be sufficient to look at those PCs which are covering the most important data information. One could also argue, that part of the information just consists of noise, and this should be contained in the last few PCs, which are not interesting for the inspection.

PCA tries to explain as much of the total variance as possible with fewer dimensions. What is the total variance? It is:

$$\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

This is the same as,

$$\frac{1}{n-1} \mathrm{trace}\left((\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)^\top (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top) \right) = \mathrm{trace}(\boldsymbol{S}) = \mathrm{trace}(\hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{A}} \hat{\boldsymbol{\Gamma}}^\top) = \mathrm{trace}(\hat{\boldsymbol{A}}) \ ,$$

where "trace" is the sum of the diagonal elements of the matrix, here the sum of the eigenvalues.

There exist several other tests, such as a test of the hypothesis that the explained variance of the first k PCs exceeds a certain threshold, such as 80 % or 90 %.

A further frequently applied criterion is to exclude those PCs which have a variance (eigenvalue) lower than the average. If the data are standardized, the sum of the eigenvalues is p, and thus the average is 1.

A further possibility to select the number of relevant PCs is the *scree graph*, which shows the proportion of explained variance of each PC versus the number of the PC. We exclude those (smallest) PCs where the proportion follows approximately a linear trend.

Example 5.5.3 *The scree graph for the exam data is shown in Figure 5.4. One can see that the contributions of the last 3 PCs follow a linear trend, and thus these PCs can be excluded.*

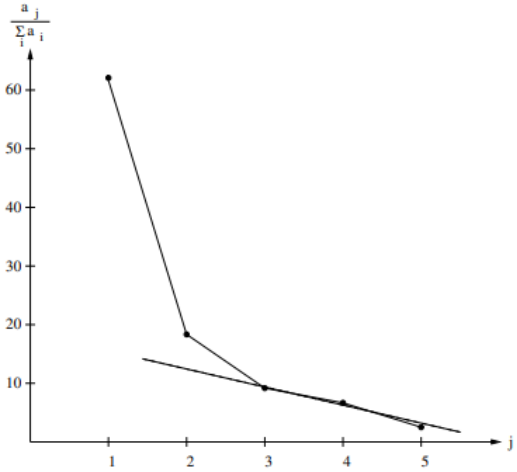


Figure 5.4: Scree graph for the exam data example.

22 What is singular value decomposition, how is it defined, and how is it related to PCA? What are the scores in terms of SVD? When would we prefer SVD to spectral decomposition of the covariance (correlation) matrix?

5.6 Singular value decomposition

Singular value decomposition (SVD) can be viewed as an alternative algorithm to determine the PCs. It is not based on a decomposition of the covariance matrix, but it directly uses the data matrix for the decomposition. This is a particular advantage if $n < p$ ("flat" data matrices), which would always lead to a singular covariance matrix, where the last $n - p$ PCs would have variance 0.

Let us assume in the following that the columns of the real-valued data matrix $\boldsymbol{X}$ are centered to mean 0. Then there exists an orthogonal $n \times n$ matrix $\boldsymbol{U}$ and an orthogonal $p \times p$ matrix $\boldsymbol{V}$, such that we obtain the decomposition

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top , \tag{5.31}$$

where $\boldsymbol{D}$ is an $n \times p$ matrix with "diagonal" elements $d_{ii} \geq 0$ for $i = 1, \ldots, \min(n, p)$, and the remaining elements are 0. The positive values d_{ii} are called *singular values* of $\boldsymbol{X}$. The number of these positive values corresponds to the rank of $\boldsymbol{X}$.

If the rank of $\boldsymbol{X}$ is $k \leq min(n, p)$, then $\boldsymbol{X}$ can be represented as

$$\boldsymbol{X} = \sum_{i=1}^{k} d_{ii}\boldsymbol{u}_i\boldsymbol{v}_i^\top , \tag{5.32}$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the i-th columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Due to the orthogonality of $\boldsymbol{U}$ and $\boldsymbol{V}$ we obtain

$$\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{u}_i = d_{ii}^2\boldsymbol{u}_i \tag{5.33}$$

and

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_i = d_{ii}^2\boldsymbol{v}_i . \tag{5.34}$$

This means that $\boldsymbol{u}_i$ is the i-th eigenvector of $\boldsymbol{X}\boldsymbol{X}^\top$ to the eigenvalue d_{ii}^2, and $\boldsymbol{v}_i$ is the i-th eigenvector of $\boldsymbol{X}^\top\boldsymbol{X}$ to the same eigenvalue d_{ii}^2. The eigenvalues for $i = 1, \ldots, k$ are strictly positive, and the remaining ones are zero.

As a summary we conclude that $\boldsymbol{U}$ has n orthogonal eigenvectors of $\boldsymbol{X}\boldsymbol{X}^\top$ in its columns, and $\boldsymbol{V}$ has p orthogonal eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$ in its columns.

It is now straightforward to see the connection to a covariance-based estimation of the PCs, as it was done in Section 5.4: Consider again mean-centered data $\boldsymbol{X}$.

Then the sample PCs were defined as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$, and thus $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma}^\top$. Further, in this case the sample covariance matrix is

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X} = \hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{A}}\hat{\boldsymbol{\Gamma}}^\top .$$

The matrix $\hat{\boldsymbol{\Gamma}}$ is the matrix with the normed eigenvectors of $\boldsymbol{S}$. In SVD, $\boldsymbol{V}$ is the matrix with the normed eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$, and therefore we can conclude that $\hat{\boldsymbol{\Gamma}} \equiv \boldsymbol{V}$. From Equation (5.34) we see that $d_{ii}^2 = (n-1)\hat{a}_i$. We can thus write

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \tag{5.35}$$

and therefore we have $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$.

23 How can we define the PCA problem in terms of reconstruction error (Frobenius norm)?

Alternative definitions for PCs

With these considerations we can formulate the PCA problem based on a different objective function.

Let us first define the *Frobenius norm* of a matrix: Denote $\boldsymbol{x}_{i\cdot}$ as the rows of $\boldsymbol{X}$, for $i = 1, \ldots, n$. The Frobenius norm of $\boldsymbol{X}$ is defined as:

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{n} \|\boldsymbol{x}_{i\cdot}\|_2^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2}$$

From above we have

$$\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D} = \boldsymbol{Z}.$$

Define $\boldsymbol{V}_m = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$, and m is smaller than the rank of $\boldsymbol{X}$. Then the columns of $\boldsymbol{X}\boldsymbol{V}_m$ are the first m PCs. This is equivalent to a projection of $\boldsymbol{X}$ onto an m-dimensional subspace formed by $\boldsymbol{V}_m$. One can show that

$$\boldsymbol{V}_m = arg\max_{\boldsymbol{B}} \|\boldsymbol{X}\boldsymbol{B}\|_F^2$$

for any $p \times m$ matrix $\boldsymbol{B}$ with $rank(\boldsymbol{B}) \leq m$ and $\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}$.

An equivalent formulation is the following: We have

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{X}\boldsymbol{V}_m\boldsymbol{V}_m^\top + \boldsymbol{E}.$$

Then

$$\boldsymbol{V}_m = arg\min_{\boldsymbol{B}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\|_F^2$$

with the same definition of $\boldsymbol{B}$ as above. Note that $\hat{\boldsymbol{X}} := \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top$ has the same dimension as $\boldsymbol{X}$. Therefore we can view $\hat{\boldsymbol{X}}$ as a rank m approximation of $\boldsymbol{X}$, which is optimal in the above sense. For finding the PC directions we are in fact minimizing residual sum-of-squares, with the residual matrix $\boldsymbol{X} - \hat{\boldsymbol{X}}$.

24 What are Biplots? What is the rank-2 approximation? Define the G/H matrix. What are the properties of the biplot? (inner row product of G and H approximates elements of the X-matrix, etc.)

Biplots have been introduced in Gabriel (1971) to display both variable and object information jointly in one plot – usually in 2 dimensions. The "bi" does not refer to the "2 dimensions" but to the joint presentation of variables and observations. Here we will introduce biplots to represent loadings and scores from a PCA.

We like 2-dimensional plots because they are easy to handle. A projection of the data into the plane should, however, make sure that the data have approximately rank 2. Then we are sure that indeed the projection represents the essential variability.

Let $\boldsymbol{X}$ be a mean-centered $n \times p$ data matrix of rank k (not much larger than 2). The biplot shows a representation of $\boldsymbol{X}$ by means of two groups of vectors with dimension n and p, respectively, which form a rank-2 approximation of $\boldsymbol{X}$. Denote this rank-2 approximation by $\boldsymbol{X}_{(2)}$.

We know that a least-squares based rank-2 approximation is given by the first 2 PCs, and we are aware of SVD to compute these first 2 PCs. In the following we will only need the first 2 columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, but we will still use the notation $\boldsymbol{U}$ and $\boldsymbol{V}$ in order to avoid new symbols. Thus:

$$\boldsymbol{X} \approx \boldsymbol{X}_{(2)} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \end{pmatrix} . \tag{5.36}$$

Since we want to represent the data by n score vectors and p loadings vectors (all 2-dimensional), we need a decomposition of $\boldsymbol{X}_{(2)}$ into an $n \times 2$ and an $p \times 2$ matrix, say

$$\boldsymbol{X}_{(2)} = \boldsymbol{G}\boldsymbol{H}^\top \tag{5.37}$$

with

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c} \tag{5.38}$$

and

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{5.39}$$

for $0 \leq c \leq 1$. Depending on the choice of c, the first 2 singular values are distributed among the matrices $\boldsymbol{G}$ and $\boldsymbol{H}$. The biplot consists of the rows of $\boldsymbol{G}$ and $\boldsymbol{H}$, i.e. of $n + p$ 2-dimensional vectors.

For the choice $c = 0.5$ we use the same scaling for the observation and variable vectors, which seems to be natural. However, we obtain nicer properties with the

choice $c = 1$ (and with re-scaling):

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{g}_{1.}^\top \\ \vdots \\ \boldsymbol{g}_{n.}^\top \end{pmatrix} = \sqrt{n-1}\,(\,\boldsymbol{u}_1\ \boldsymbol{u}_2\,) \tag{5.40}$$

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{h}_{1.}^\top \\ \vdots \\ \boldsymbol{h}_{p.}^\top \end{pmatrix} = \frac{1}{\sqrt{n-1}}\,(\,\boldsymbol{v}_1\ \boldsymbol{v}_2\,) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \tag{5.41}$$

These properties are:

- The inner product between the rows of $\boldsymbol{G}$ and the rows of $\boldsymbol{H}$ approximate the values x_{ij} of the data matrix $\boldsymbol{X}$:

$$\boldsymbol{g}_{i.}^\top \boldsymbol{h}_{j.} = \sqrt{n-1}\ \boldsymbol{u}_{i.}^\top \frac{1}{\sqrt{n-1}} \left(\boldsymbol{v}_{j.}^\top \boldsymbol{D}\right)^\top = \boldsymbol{u}_{i.}^\top \boldsymbol{D}\boldsymbol{v}_{j.} \approx x_{ij}\ . \tag{5.42}$$

- The inner product between the rows of $\boldsymbol{H}$ approximates the covariance:

$$\begin{aligned} \boldsymbol{H}\boldsymbol{H}^\top &= \left(\frac{1}{\sqrt{n-1}}\,\boldsymbol{V}\boldsymbol{D}\right)\left(\frac{1}{\sqrt{n-1}}\,\boldsymbol{D}\boldsymbol{V}^\top\right) = \frac{1}{n-1}\,\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top \\ &= \frac{1}{n-1}\,\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\right) = \frac{1}{n-1}\,\boldsymbol{X}_{(2)}^\top \boldsymbol{X}_{(2)} \\ &\approx \frac{1}{n-1}\,\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{S}\ . \end{aligned} \tag{5.43}$$

  It follows that the squared Euclidean norm of the rows of $\boldsymbol{H}$, i.e. $\|\boldsymbol{h}_{j.}\|^2 = \boldsymbol{h}_{j.}^\top \boldsymbol{h}_{j.}$, approximates the variance. Moreover, the cosine between $\boldsymbol{h}_{i.}$ and $\boldsymbol{h}_{j.}$ $(i, j = 1, \ldots, p)$ approximates the correlation between the variables,

$$\cos(\boldsymbol{h}_{i.}, \boldsymbol{h}_{j.}) = \frac{\boldsymbol{h}_{i.}^\top \boldsymbol{h}_{j.}}{\|\boldsymbol{h}_{i.}\|\|\boldsymbol{h}_{j.}\|} \approx r_{ij}\ . \tag{5.44}$$

- The Euclidean distance between the rows of $\boldsymbol{G}$ approximates the Mahalanobis distance between the observations,

$$\begin{aligned} \|\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\|^2 &= \left(\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\right)^\top \left(\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\right) = (n-1)\left(\boldsymbol{u}_{i.} - \boldsymbol{u}_{j.}\right)^\top \left(\boldsymbol{u}_{i.} - \boldsymbol{u}_{j.}\right) \\ &\approx \left(\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\right)^\top \boldsymbol{S}^{-1} \left(\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\right)\ , \end{aligned} \tag{5.45}$$

  because

$$\boldsymbol{x}_{i.}^\top \boldsymbol{S}^{-1} \boldsymbol{x}_{j.} \approx \left(\boldsymbol{u}_{i.}^\top \boldsymbol{D}\boldsymbol{V}^\top\right)(n-1)\left(\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top\right)\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{u}_{j.}\right) = (n-1)\boldsymbol{u}_{i.}^\top \boldsymbol{u}_{j.} \tag{5.46}$$

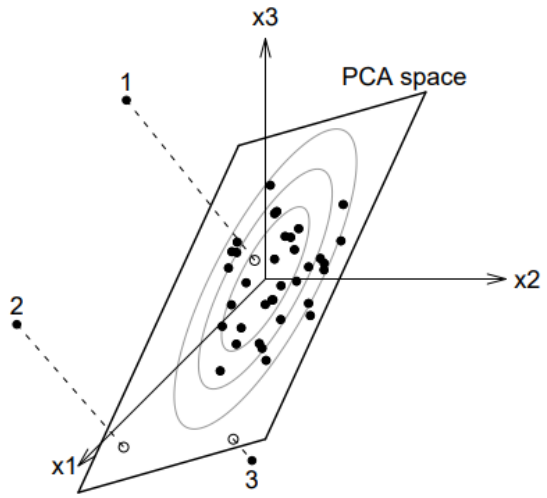
  for $i, j = 1, \ldots, n$.

25 Which diagnostics do we have for PCA (formal definition of orthogonal, score distance)?

Two distance measures have been introduced: the score distance (SD) and the orthogonal distance (OD).

Figure 5.6 tries to explain the ideas behind these distances. We have observations in the 3-dimensional space, and the PCA space is built up with 2 PCs. The SD is a distance measure in the PCA space, and it is equal to the Mahalanobis distances of the observations projected into this space. The ellipses refer to these Mahalanobis distances. The OD is the distance orthogonal to the PCA space.

Here we can see three particular observations: Observation 1 has big OD but small SD; observation 2 has big OD and big SD; observation 3 has small OD but big SD. Similar to the diagnostics in regression one could classify these different types of observations as vertical outliers (1), and good (3) and bad (2) leverage points. In particular, bad leverage points can have a strong influence on the classical estimation of the PCs. Here, "classical" refers to the SVD estimation, or equivalently, the estimation based on the empirical covariance matrix. A robust PCA could easily be obtained through a robust estimation of the covariance matrix. In this case, the diagnostics would also make much more sense.



Formally, these distances are defined as follows. Let $\hat{\mathbf{\Gamma}}_k = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_k)$ be the matrix of the first k estimated PC loadings, $\boldsymbol{z}_{i\cdot} = (z_{i1}, \ldots, z_{ik})^\top$ the i-th score vector, and $\hat{a}_1, \ldots, \hat{a}_k$ the corresponding variances of the PCs. The number k of PCs can be selected according to a criterion defined in Section 5.5, e.g. such that 80% of the variability is explained.

The SD for the i-th observation ($i = 1, \ldots, n$) is defined as

$$\mathrm{SD}_i = \left(\sum_{j=1}^{k} \frac{z_{ij}^2}{\hat{a}_j} \right)^{1/2}.$$

This is equal to the Mahalanobis distance of the score vector to the PCA center (zero) with respect to the covariance matrix ($\boldsymbol{A}$).

The OD for the i-th observation $\boldsymbol{x}_{i\cdot}$ ($i = 1, \ldots, n$) is defined as

$$\mathrm{OD}_i = \|\boldsymbol{x}_{i\cdot} - \hat{\mathbf{\Gamma}}_k \boldsymbol{z}_{i\cdot}\|_2,$$

which is the Euclidean distance of the observation to its projection into the space of the first k PCs.

Similar to multivariate outlier detection, one can define cutoff values for both distance measures, which would refer to unusual values of SD and/or OD. Since SD is a Mahalanobis distance, a suitable cutoff value is $\sqrt{\chi^2_{k;0.975}}$. For the cutoff value for the OD it has been argued that $\mathrm{OD}^{2/3}$ is closer to normality, and thus a suitable cutoff value is

$$\left(\mathrm{median}_i(\mathrm{OD}_i^{2/3}) + \mathrm{MAD}_i(\mathrm{OD}_i^{2/3}) z_{0.975}\right)^{3/2},$$

where $z_{0.975}$ is the 0.975 quantile of the $N(0,1)$. MAD stands for the Median Absolute Deviation, which is defined for univariate values $y_1, \ldots, y_n$ as

$$\mathrm{MAD} = 1.483 \cdot \mathrm{median}_i(\mid y_i - \mathrm{median}_j(y_j) \mid).$$

26 What is the factor analysis model (formal definition, assumptions)? What is the difference to PCA?

In factor analysis we assume that what we observe is basically the result of underlying quantities which are not directly observable. These quantities are called latent variables, and they cannot be measured. The "factors" in factor analysis aim at isolating such latent variables, explaining the relationships in the data.

Note that factor analysis is basically similar to PCA, since we also aim for a dimension reduction. However, we would like to have interpretable factors (the components in PCA are not necessarily interpretable), and we also have a statistical model (PCs were only defined by a linear transformation).

27 Explain the decomposition of the correlation matrix in factor analysis.

6.2 Factor analysis model

6.2.1 Definition

Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ be a p-dimensional vector of random variables, $x_1, x_2, \ldots, x_p$, which will describe our characteristics or variables later on.

Since in PCA we had the problem of the dependency on the scale, we will right away center and scale the random variables to mean zero and variance one. Thus, $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^\top$ is the new random variable, obtained by

$$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}}, \qquad i = 1, \ldots, p \quad .$$

In the model we already assume that the information contained in $\boldsymbol{y}$ can be re-expressed by a smaller number $k < p$ of unknown random variables (factors) $\boldsymbol{f} = (f_1, \ldots, f_k)^\top$, up to an error term $\boldsymbol{e}$. Thus, dimension reduction is already intrinsically stated in the model, called k-factor model:

$$\boldsymbol{y} = \boldsymbol{\Lambda f} + \boldsymbol{e} \quad . \tag{6.1}$$

Here, $\boldsymbol{\Lambda} = [(\lambda_{ij})]$ is a $(p \times k)$ matrix with fixed values. It is called *loadings matrix*, and it describes the relationships between factors and variables (as in PCA). The error term $\boldsymbol{e} = (e_1, \ldots, e_p)^\top$ is often called *unique factor* (or uniqueness).

In this model we have the following assumptions:

$$\begin{aligned} E(\boldsymbol{f}) &= \boldsymbol{0}, \;\; Cov(e_i, e_j) = 0 \qquad (i \neq j), \\ E(\boldsymbol{e}) &= \boldsymbol{0}, \;\; Cov(\boldsymbol{f}, \boldsymbol{e}) = \boldsymbol{O} \;, \\ Var(f_i) &= 1 \;. \end{aligned}$$

With these assumptions, we can see that the covariance matrix of the error term has a diagonal form:

$$Cov(\boldsymbol{e}) = \boldsymbol{\Psi} = Diag(\psi_{11}, \ldots, \psi_{pp}) \;. \tag{6.2}$$

We can thus re-express the correlation matrix $\boldsymbol{\rho} = [(\rho_{ij})]$ of our random variables $\boldsymbol{x}$ by our model:

$$\begin{aligned} \boldsymbol{\rho} &= Cor(\boldsymbol{x}) = Cov(\boldsymbol{y}) = Cov(\boldsymbol{\Lambda f} + \boldsymbol{e}) \\ &= \boldsymbol{\Lambda} \underbrace{Cov(\boldsymbol{f})}_{\boldsymbol{\Phi}} \boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda} \underbrace{Cov(\boldsymbol{f}, \boldsymbol{e})}_{\boldsymbol{O}} + \underbrace{Cov(\boldsymbol{e}, \boldsymbol{f})}_{\boldsymbol{O}} \boldsymbol{\Lambda}^\top + \underbrace{Cov(\boldsymbol{e})}_{\boldsymbol{\Psi}} \\ &= \boldsymbol{\Lambda \Phi \Lambda}^\top + \boldsymbol{\Psi} \quad . \end{aligned}$$

Here, $\boldsymbol{\Phi}$ is the $(k \times k)$ matrix with the correlations between the factors. If we additionally assume that the factors are uncorrelated, i.e.

$$Cov(\boldsymbol{f}) = \boldsymbol{\Phi} = \boldsymbol{I} \quad , \tag{6.3}$$

we obtain

$$\boldsymbol{\rho} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \tag{6.4}$$

or

$$\boldsymbol{\rho}_{red} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \begin{pmatrix} \kappa_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \kappa_2^2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \kappa_p^2 \end{pmatrix} = \boldsymbol{\rho} - \boldsymbol{\Psi} \quad , \tag{6.5}$$

with the *reduced correlation matrix* $\boldsymbol{\rho}_{red}$. The diagonal elements $\kappa_i^2 = 1 - \psi_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 \quad (i = 1, \ldots, p)$ are called *communalities*. They correspond to the row-sums of the squared factor loadings, and describe the proportion of variance of y_i, explained by the factors, because the total variance is trace($\boldsymbol{\rho}$), and thus trace($\boldsymbol{\rho}_{red}$) is the variance explained by the factor model.

28 What are the uniquenesses and communalities, how are they defined, what is their meaning and how are they related?

6.2.2 Non-uniqueness of the factor loadings

Let $\boldsymbol{G}$ be an orthogonal matrix of dimension $(k \times k)$. Because of $\boldsymbol{G}^{-1} = \boldsymbol{G}^\top$ we have

$$\boldsymbol{y} = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{f}) + \boldsymbol{e} \; . \tag{6.6}$$

Since the new factors $\boldsymbol{G}^\top \boldsymbol{f}$ also fulfill our model assumptions,

$$E(\boldsymbol{G}^\top \boldsymbol{f}) = \boldsymbol{0}, \quad Cov(\boldsymbol{G}^\top \boldsymbol{f}) = \boldsymbol{I}$$
$$\text{and} \qquad Cov(\boldsymbol{G}^\top \boldsymbol{f}, \boldsymbol{e}) = \boldsymbol{O} \; ,$$

this k-factor model is also valid with the new factors, i.e.

$$\boldsymbol{\rho} = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{\Lambda}^\top) + \boldsymbol{\Psi} \; . \tag{6.7}$$

This, however, means that the new factor loadings are not uniquely determined.

Basically, this is not a big problem, because later on we want to rotate the factors in any case in order to obtain a better interpretation. For now, if we just want to obtain a unique solution, we can impose further restrictions in order to achieve uniqueness. These are the constraints that either $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ or $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$ is diagonal.

29 What is the maximum number of factors we can include in the factor model, and why?

6.2.3 Number of parameters

Practically, we can estimate the correlation matrix $\widehat{\boldsymbol{\rho}}$ based on real data. Then, $\widehat{\boldsymbol{\rho}}$ is used to estimate loadings $\widehat{\boldsymbol{\Lambda}}$ and error variances $\widehat{\boldsymbol{\Psi}}$, and they need to fulfill either

$$\widehat{\boldsymbol{\Lambda}}^\top \widehat{\boldsymbol{\Psi}}^{-1} \widehat{\boldsymbol{\Lambda}} = Diag \qquad \text{or} \qquad \widehat{\boldsymbol{\Lambda}}^\top \widehat{\boldsymbol{\Lambda}} = Diag \tag{6.8}$$

as well as

$$\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}^\top + \widehat{\boldsymbol{\Psi}} \; . \tag{6.9}$$

Does the factor model lead to a simpler interpretation than the correlation matrix? In other words, is the number of parameters for the factor model lower compared to the correlation matrix? For the correlation matrix we need to estimate $\frac{1}{2}p(p+1)$ parameters. For a k-factor model we need to estimate $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, thus $pk + p$ parameters. For uniqueness of these estimates we ask for diagonal structure of either $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ or $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$, thus $\frac{1}{2}k(k-1)$ restrictions. Therefore, for the k-factor model we have $pk + p - \frac{1}{2}k(k-1)$ parameters to estimate. The difference to an unrestricted model is thus

$$\begin{aligned} s &= \frac{1}{2}p(p+1) - \left(pk + p - \frac{1}{2}k(k-1)\right) \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \quad . \end{aligned} \tag{6.10}$$

Only for the case $s > 0$, the k-factor model leads to a simpler interpretation than the correlation matrix, the other cases would not be valid (or useful) solutions. Thus, the requirement for $s > 0$ leads to an upper bound for the number k of factors.

6.2.4 Parameter estimation by Principal Factor Analysis (PFA)

The main job in factor analysis is to estimate the parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. There are different approaches, such as the *Maximum Likelihood Method (MLM)*, implemented in the R function `factanal()`, or *Principal Factor Analysis (PFA)*, implemented in the R package `StatDA` as function `pfa()`. For MLM we need to assume that the data are generated from a multivariate normal distribution, which is not explicitely required for PFA. In fact, MLM is quite technical, and thus we focus here on PFA only, since this is closely related to PCA.

We start from our model

$$\boldsymbol{\rho} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi} \tag{6.11}$$

and try to first estimate the communalities, which is equivalent to estimating the diagonal elements of $\mathbf{\Psi}$. Afterward we estimate the loadings matrix.

Estimation of the communalities

The communalities

$$\kappa_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = 1 - \psi_{ii} \qquad (i = 1, \ldots, p) \tag{6.12}$$

describe the porportion of variance explained by the k-factor model. They are in the interval $[0, 1]$. There are different options to estimate the communalities:

1. Highest correlation coefficient: $max_{i \neq j} \mid \hat{\rho}_{ij} \mid$
   For the estimation of the communalities we use from each column of $\boldsymbol{\rho}$ the largest (absolute) non-diagonal element.

2. Squared multiple correlation coefficient: $\hat{\rho}^2_{i,12\ldots)i(\ldots p}$
   This measure refers to an R^2 measure from a linear regression of the i-th variable on the remaining variables. Thus, it tells us about the variance porportion of the i-th variable explained from the other variables. We can compute this measure as:

2. Squared multiple correlation coefficient: $\hat{\rho}^2_{i,12\ldots)i(\ldots p}$
   This measure refers to an R^2 measure from a linear regression of the i-th variable on the remaining variables. Thus, it tells us about the variance porportion of the i-th variable explained from the other variables. We can compute this measure as:

$$\hat{\rho}^2_{i,12\ldots)i(\ldots p} = 1 - \frac{1}{\hat{\rho}^{ii}} \quad , \tag{6.13}$$

   where $\hat{\rho}^{ii}$ is the i-th diagonal element of $\hat{\boldsymbol{\rho}}^{-1}$.

3. Iterative estimation:
   We first need to fix the number k of factors. This could be done by using a criterion from PCA on the number of components. Then we start to initialize the communalities by using one of the estimation methods mentioned above. The diagonal elements of $\hat{\boldsymbol{\rho}}$ are replaced by these communalities, and from the resulting reduced correlation matrix we estimate the loadings (see end of this section). Based on that we re-estimate the communalities:

$$\hat{\kappa}_i^2 = \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 \qquad \text{for} \quad i = 1, \ldots, p \,, \tag{6.14}$$

and they are used to form the new reduced correlation matrix. The iteration continues until the communalities stabilize.

If the communalities are over-estimated, part of the uniquenesses are forced into the k-factor model, which might change the pattern and interpretability of the factors. The same may happen if the communalities are under-estimated, because variance of the factors is forced into the uniquenesses. This issue is more severe if the number of variables is small.

Estimation of the loadings

Once the communalities $\hat{\kappa}_i^2 = 1 - \hat{\psi}_{ii}$, for $i = 1, \ldots, p$, have been estimated, we can estimate the loadings based on the reduced correlation matrix

$$\hat{\boldsymbol{\rho}}_{red} = \hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\Psi}} \ , \tag{6.15}$$

where the diagonal consists of these communalities.

Using the Spectral Theorem 1.4.3, we have:

$$\hat{\boldsymbol{\rho}}_{red} = \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{A}} \hat{\boldsymbol{\Gamma}}^\top \ , \tag{6.16}$$

where $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ and $\hat{\boldsymbol{A}} = Diag(\hat{a}_1, \ldots, \hat{a}_p)$, with the eigenvalues $\hat{a}_i$ and the corresponding eigenvectors $\hat{\boldsymbol{\gamma}}_i$ of $\hat{\boldsymbol{\rho}}_{red}$ $(i = 1, \ldots, p)$.

Note that we fixed the number of factors with k, where $1 \leq k < p$, and thus the loadings matrix needs to have dimension $p \times k$. Thus, we shall only use the first k eigenvectors $\hat{\boldsymbol{\Gamma}}_{1:k}$ and eigenvalues $\hat{\boldsymbol{A}}_{1:k}$ in the k-factor model

$$\hat{\boldsymbol{\rho}} - \hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\Lambda}}^\top = \hat{\boldsymbol{\Gamma}}_{1:k} \hat{\boldsymbol{A}}_{1:k} \hat{\boldsymbol{\Gamma}}_{1:k}^\top + \sum_{i=k+1}^{p} \hat{a}_i \hat{\boldsymbol{\gamma}}_i \hat{\boldsymbol{\gamma}}_i^\top \quad . \tag{6.17}$$

The estimated loadings matrix is thus naturally

$$\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Gamma}}_{1:k} \hat{\boldsymbol{A}}_{1:k}^{1/2} \quad , \tag{6.18}$$

where the diagonal elements of $\hat{\boldsymbol{A}}_{1:k}^{1/2}$ are the values $\sqrt{\hat{a}_1}, \ldots, \sqrt{\hat{a}_k}$.

Considering Equation (6.17), it is also natural to update the estimated uniquenesses as

$$\hat{\psi}_{ii} = 1 - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 \qquad \text{for} \quad i = 1, \ldots, p \quad . \tag{6.19}$$

The solution is valid if all $\hat{\psi}_{ii} \geq 0$.

31 How can we interpret factors? Give an overview of factor rotation criteria.

A rotation of the factors will change the loadings, and thus also the interpretation. The goal is to rotate in such a way that the resulting pattern of the loadings matrix is “simple” and thus interpretable. “Simple” basically means that the loadings matrix contains essentially small (absolute) values, and few values close to −1 or 1. If there is a large (absolute) value, then we know that the corresponding variable has a strong contribution on this factor, while others with loadings close to zero do not have a contribution. Many loadings close to zero would simplify the interpretation of a factor.
From the rotated factors we get strong contributions from some variables and weak contributions from others. The rotated factors would thus have a much clearer interpretation.

6.3.1 Orthogonal rotation

The interpretation is "simple" if a point in two dimensions is close to an axis. In that case, the product of both coordinates is small, and by taking the square one gets rid of the sign. Now we can consider the sum of all squared products as a criterion for simplicity. This should be valid for all pairs of factors, resulting in the criterion

$$\sum_{s<j=1}^{k}\sum_{i=1}^{p}\Big(\lambda_{is}\lambda_{ij}\Big)^2 \longrightarrow \min \quad . \tag{6.20}$$

A factor rotation can be achieved by a transformation of the loadings matrix. For orthogonal rotation we thus have to consider an orthogonal $k \times k$ matrix $\boldsymbol{T}$, and the rotated loadings are given by

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{T} \; . \tag{6.21}$$

Orthogonal transformations do not change the communalities, since

$$Diag(\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\Lambda}}^\top) = Diag(\boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}^{-1}\boldsymbol{\Lambda}^\top) = Diag(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top) \; ,$$

and thus

$$\kappa_i^2 = \sum_{j=1}^{k}\lambda_{ij}^2 = \sum_{j=1}^{k}\tilde{\lambda}_{ij}^2 \qquad \text{for} \quad i = 1,\ldots,p \quad . \tag{6.22}$$

Also the squared communalities remain constant under orthogonal transformations,

$$(\kappa_i^2)^2 = \Big(\sum_{j=1}^{k}\tilde{\lambda}_{ij}^2\Big)^2 = \sum_{j=1}^{k}\tilde{\lambda}_{ij}^4 + 2\sum_{s<j=1}^{k}\tilde{\lambda}_{is}^2\tilde{\lambda}_{ij}^2 = const \quad , \tag{6.23}$$

and also the sum over all variables remains constant:

$$\sum_{i=1}^{p}\sum_{j=1}^{k}\tilde{\lambda}_{ij}^4 + 2\sum_{i=1}^{p}\sum_{s<j=1}^{k}\tilde{\lambda}_{is}^2\tilde{\lambda}_{ij}^2 = const \quad . \tag{6.24}$$

Since the sum of both expressions is constant, one term is maximized if the other term is minimized, and vice versa.

The **Quartimax criterion** thus maximizes

$$QMAX = \sum_{i=1}^{p}\sum_{j=1}^{k}\tilde{\lambda}_{ij}^4, \tag{6.25}$$

with the effect that this may lead to one dominant factor.

Another criterion for factor rotation is the **Varimax criterion** (Kaiser, 1958), where the variance of the squared factor loadings for each factor j is considered:

$$s_j^2 = \frac{1}{p}\sum_{i=1}^{p}\left(\widetilde{\lambda}_{ij}^2 - \frac{1}{p}\sum_{l=1}^{p}\widetilde{\lambda}_{lj}^2\right)^2 = \frac{1}{p}\sum_{i=1}^{p}\Big(\widetilde{\lambda}_{ij}^2\Big)^2 - \frac{1}{p^2}\Big[\sum_{i=1}^{p}\widetilde{\lambda}_{ij}^2\Big]^2 \tag{6.26}$$

This variance is summed up over all factors. Maximizing this expression means that we want to have (absolute) large and small loadings. Since variables with larger communality are dominating the criterion, one can normalize with the communalities, which yields (after multiplication with p^2) the varimax criterion

$$VMAX = p\sum_{j=1}^{k}\sum_{i=1}^{p}\Big(\frac{\widetilde{\lambda}_{ij}}{\kappa_i}\Big)^4 - \sum_{j=1}^{k}\Big[\sum_{i=1}^{p}\Big(\frac{\widetilde{\lambda}_{ij}}{\kappa_i}\Big)^2\Big]^2 \quad . \tag{6.27}$$

6.3.2 Oblique rotation

The main difference to orthogonal rotations is in the initial factor model, which is now

$$\boldsymbol{\rho} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Psi} \quad , \tag{6.28}$$

because oblique factors are no longer uncorrelated, and thus the correlation matrix $\boldsymbol{\Phi}$ of the factors needs to be considered.

The **Quartimin criterion** is defined as in Equation (6.20) by

$$QMIN = \sum_{s<j=1}^{k} \sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \widetilde{\lambda}_{ij}^2 \quad . \tag{6.29}$$

The rotated loadings are

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{T} \ , \tag{6.30}$$

where $\boldsymbol{T}$ is no longer orthogonal. Plugging in the rotated loadings in the model yields

$$\begin{aligned} \boldsymbol{\rho} - \boldsymbol{\Psi} &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\top} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}^{-1}\boldsymbol{\Phi}(\boldsymbol{T}^{-1})^{\top}\widetilde{\boldsymbol{\Lambda}}^{\top} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}^{-1}\mathrm{Cov}(\boldsymbol{f})(\boldsymbol{T}^{-1})^{\top}\widetilde{\boldsymbol{\Lambda}}^{\top} \\ &= \widetilde{\boldsymbol{\Lambda}}\mathrm{Cov}(\boldsymbol{T}^{-1}\boldsymbol{f})\widetilde{\boldsymbol{\Lambda}}^{\top} = \widetilde{\boldsymbol{\Lambda}}\mathrm{Cov}(\widetilde{\boldsymbol{f}})\widetilde{\boldsymbol{\Lambda}}^{\top} \quad . \end{aligned} \tag{6.31}$$

Die rotated factors $\widetilde{\boldsymbol{f}}$ are thus

$$\widetilde{\boldsymbol{f}} = \boldsymbol{T}^{-1}\boldsymbol{f} \ . \tag{6.32}$$

The transformation matrix $\boldsymbol{T}$ thus needs to be invertible.

A more flexible criterion is the **Oblimin criterion**

$$OBMIN = \sum_{s<j=1}^{k} \left(\sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \widetilde{\lambda}_{ij}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \sum_{i=1}^{p} \widetilde{\lambda}_{ij}^2 \right) \quad . \tag{6.33}$$

The choice $\gamma = 0$ yields the Quartimin criterion, while $\gamma = 1$ leads to the so-called *Covarimin criterion*.

32 How are factor scores estimated (Bartlett and Regression method). Name the models, formulas and solutions for the factor scores estimates.

6.4.1 Weighted least-squares estimation

In the R function `factanal()` this method is selected by `scores="Bartlett"`.

The factor model is

$$\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{e} \tag{6.34}$$

with the error term $\boldsymbol{e} = (e_1, \ldots, e_p)^{\top}$. This model could also be considered as a regression model, by regressing $\boldsymbol{y}$ on $\boldsymbol{\Lambda}$, with the regression coefficients $\boldsymbol{f}$. However, the model is heteroscedastic, since the error variances $Var(e_i) = \psi_i$ for $i = 1, \ldots, p$ are not necessarily equal. However, one can multiply the equation with weights $\boldsymbol{\Psi}^{-1/2}$, which yields

$$\boldsymbol{\Psi}^{-1/2}\boldsymbol{y} = \boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\Psi}^{-1/2}\boldsymbol{e} \ . \tag{6.35}$$

The covariance of the new error term is the identity matrix, and thus we have a homoscedastic model, and the least-squares estimator can be used, where the loadings and uniquenesses are considered as "true" values. This results in

$$\hat{\boldsymbol{f}} = \left(\boldsymbol{\Lambda}^{\top}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}^{\top}\boldsymbol{\Psi}^{-1}\boldsymbol{y} \tag{6.36}$$

for the estimated factors (random variables).

For a given $n \times p$ data matrix $\boldsymbol{X}$ we obtain the mean-centered and scaled matrix $\boldsymbol{Y}$ as a realization of $\boldsymbol{y}$, which can be substituted into Equation (6.36), to obtain the $n \times k$ matrix of estimated factor scores:

$$\hat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}^{\top}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1} \tag{6.37}$$

6.4.2 Regression method

In the R function `factanal()` this method is selected by `scores="regression"`.

As for the previous method, the loadings $\boldsymbol{\Lambda}$ and uniquenesses $\boldsymbol{\Psi}$ are considered as known. Here we again consider a regression problem, but we regress the (unknown) factors $\boldsymbol{f}$ on $\boldsymbol{y}$ (multivariate regression):

$$\boldsymbol{f} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\delta}\ , \tag{6.38}$$

where $\boldsymbol{B}$ is the $k \times p$ matrix of regression coefficients, and $\boldsymbol{\delta}$ the error matrix. The least-squares estimator is

$$\widehat{\boldsymbol{B}} = \boldsymbol{f}\boldsymbol{y}^{\top}(\boldsymbol{y}\boldsymbol{y}^{\top})^{-1}\ , \tag{6.39}$$

and the estimated factors are

$$\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{B}}\boldsymbol{y} = \boldsymbol{f}\boldsymbol{y}^{\top}(\boldsymbol{y}\boldsymbol{y}^{\top})^{-1}\boldsymbol{y}\ . \tag{6.40}$$

Now we have the factors $\boldsymbol{f}$ again on the right-hand side, but we can plug in our factor model for $\boldsymbol{y}$:

$$\begin{aligned}\widehat{\boldsymbol{f}} &= \boldsymbol{f}(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{e})^{\top}(\boldsymbol{y}\boldsymbol{y}^{\top})^{-1}\boldsymbol{y} \\ &= (\boldsymbol{f}\boldsymbol{f}^{\top}\boldsymbol{\Lambda}^{\top} + \boldsymbol{f}\boldsymbol{e}^{\top})(\boldsymbol{y}\boldsymbol{y}^{\top})^{-1}\boldsymbol{y} \\ &= \boldsymbol{f}\boldsymbol{f}^{\top}\boldsymbol{\Lambda}^{\top}(\boldsymbol{y}\boldsymbol{y}^{\top})^{-1}\boldsymbol{y}\end{aligned}$$

Substituting the sample version yields

$$\begin{aligned}\widehat{\boldsymbol{F}} &= \boldsymbol{Y}(n-1)(\boldsymbol{Y}^{\top}\boldsymbol{Y})^{-1}\boldsymbol{\Lambda}\frac{1}{n-1}\boldsymbol{F}^{\top}\boldsymbol{F} \\ &= \boldsymbol{Y}\boldsymbol{R}^{-1}\boldsymbol{\Lambda}\widehat{\boldsymbol{\Phi}}\ ,\end{aligned}$$

where $\widehat{\boldsymbol{\Phi}}$ is the estimated correlation matrix of the factors. In the orthogonal case, the factors are uncorrelated, and thus

$$\widehat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{R}^{-1}\boldsymbol{\Lambda}\ . \tag{6.41}$$

33 What is the problem setting in multiple correlation analysis? What is the objective function to minimize?

In statistical data analysis, one is often interested in determining relationships and dependencies of features. If one also wants to measure the existence and the strength of dependencies, correlation analysis can be used. The correlation measures the linear relationship between features.

7.1 Multiple correlation analysis

The *multiple correlation* is a measure of the dependency of a feature x on a p - dimensional feature $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$. We assume that both x and $y_1, \ldots, y_p$ are random variables with a joint distribution. The mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ of this distribution (not necessarily normal distribution) are

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \boldsymbol{\mu_y} \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \\ \boldsymbol{\sigma}_{\boldsymbol{y}x} & \boldsymbol{\Sigma_{yy}} \end{pmatrix} ,$$

where

$$E(x) = \mu_x \qquad \mathrm{Var}(x) = \sigma_{xx}$$
$$E(\boldsymbol{y}) = \boldsymbol{\mu_y} \qquad \mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{\Sigma_{yy}}$$

and

$$\mathrm{Cov}(\boldsymbol{y}, x) = \boldsymbol{\sigma}_{\boldsymbol{y}x} \qquad \mathrm{Cov}(x, \boldsymbol{y}) = \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top .$$

If x is now predicted by $\boldsymbol{y}$ (linear), the error is analogous to regression analysis

$$x - a_0 - a_1 y_1 - \ldots - a_p y_p .$$

This error is random, and therefore one would like to choose the coefficients a_0 and $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ such that the *mean squared error* (MSE)

$$\mathrm{MSE} = E(x - a_0 - \boldsymbol{a}^\top \boldsymbol{y})^2$$

is minimal. However, this MSE depends on the joint distribution of x and $\boldsymbol{y}$, and thus on the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

34 What is the linear prediction function in multiple correlation analysis? Describe the structure of the proof.

Theorem 7.1.1 *The linear prediction function* $a_0 + \boldsymbol{a}^\top \boldsymbol{y}$ *with the coefficients*

$$\boldsymbol{a} = \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x} \quad \text{and} \quad a_0 = \mu_x - \boldsymbol{a}^\top \boldsymbol{\mu}_{\boldsymbol{y}}$$

has minimal MSE of all linear prediction functions of x*. It holds that*

$$MSE = E(x - a_0 - \boldsymbol{a}^\top \boldsymbol{y})^2 = E(x - \mu_x - \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}))^2 = \sigma_{xx} - \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x} \ .$$

Furthermore,

$$a_0 + \boldsymbol{a}^\top \boldsymbol{y} = \mu_x + \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}})$$

is the linear prediction function that has maximum correlation with x*, namely*

$$Corr(x, a_0 + \boldsymbol{a}^\top \boldsymbol{y}) = \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}} \boldsymbol{a}}{\sigma_{xx}}} = \sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x}}{\sigma_{xx}}} \ .$$

Proof: See, for example, Johnson and Wichern (1998). For the proof of the minimal MSE one needs to re-express the squared expectation, and then it is visible that with these choices of $\boldsymbol{a}$ and a_0 one obtains a minimum of the MSE. For the maximum correlation one can make use of the

Extended Cauchy-Schwarz Inequality: Let $\mathbf{b}$ and $\boldsymbol{d}$ be two vectors, and $\boldsymbol{B}$ a positive definite matrix. Then

$$(\boldsymbol{b}^\top \boldsymbol{d})^2 \leq (\boldsymbol{b}^\top \boldsymbol{B} \boldsymbol{b})(\boldsymbol{d}^\top \boldsymbol{B}^{-1} \boldsymbol{d}) \ , \tag{7.1}$$

with equality if and only if $\boldsymbol{b} = c\boldsymbol{B}^{-1}\boldsymbol{d}$ (or $\boldsymbol{d} = c\boldsymbol{B}\boldsymbol{b}$), for a constant c.

The correlation between x and the best linear prediction function is called *multiple correlation coefficient* (of the population)

$$\rho_{x,\boldsymbol{y}} = +\sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x}}{\sigma_{xx}}} \ .$$

Its squared form $\rho_{x,\boldsymbol{y}}^2$ is called *multiple coefficient of determination* (of the population), and it indicates how well the feature x is explained by the properties $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$.

If correlations are given instead of the covariances, then the multiple correlation coefficient can also be defined as

$$\rho_{x,\boldsymbol{y}}^2 = \boldsymbol{\rho}_{\boldsymbol{y}x}^\top \boldsymbol{\rho}_{\boldsymbol{yy}}^{-1} \boldsymbol{\rho}_{\boldsymbol{y}x} \ .$$

For a specific sample, one can write the theoretical quantities by the corresponding realizations, and hence obtain for the multiple correlation coefficient

$$r_{x,\boldsymbol{y}}^2 = \boldsymbol{R}_{\boldsymbol{y}x}^\top \boldsymbol{R}_{\boldsymbol{yy}}^{-1} \boldsymbol{R}_{\boldsymbol{y}x} \ .$$

35 Name a hypothesis test for the multiple correlation coefficient.

A test for the hypothesis that the multiple correlation is zero is equivalent to making all bivariate correlations zero. So, we test the hypothesis

$$H_0 : \rho_{x,\boldsymbol{y}} = 0 \ (= \rho_{xy_1} = \ldots = \rho_{xy_p})$$

against the alternative hypothesis

$$H_1 : \exists\, i \in \{1, \ldots, p\} \quad \text{with} \quad \rho_{xy_i} \neq 0$$

at the significance level α. If we can assume multivariate normal distribution, and n is the number of observations in the sample, the test statistic

$$F = \frac{(n-1-p)\; r^2_{x,\boldsymbol{y}}}{p\;(1-r^2_{x,\boldsymbol{y}})}$$

has an $F_{p,n-1-p}$-distribution and the null hypothesis is rejected at the significance level of α, if

$$F > F_{p,n-1-p;1-\alpha}.$$

36 What is the problem setting in canonical correlation, what is the maximization problem?

In canonical correlation analysis we are interested in the linear dependence between two groups of variables. As it turns out, this dependence can no longer be expressed by a single correlation coefficient, but results in a subspace that describes the linear dependence between the groups.

Theorem 7.2.1 *Let $\boldsymbol{x}$ be a p-dimensional and $\boldsymbol{y}$ a q-dimensional random variable $(p \leq q)$ with the expected values*

$$E(\boldsymbol{x}) = \boldsymbol{\mu}_1 \qquad \text{and} \qquad E(\boldsymbol{y}) = \boldsymbol{\mu}_2 \;.$$

The covariance matrices $\boldsymbol{\Sigma}_{ij}$ with $i, j = 1, 2$ are defined by

$$\begin{aligned} \boldsymbol{\Sigma}_{11} &= E[(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top] \quad , \\ \boldsymbol{\Sigma}_{22} &= E[(\boldsymbol{y}-\boldsymbol{\mu}_2)(\boldsymbol{y}-\boldsymbol{\mu}_2)^\top] \quad \text{and} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top &= E[(\boldsymbol{x}-\boldsymbol{\mu}_1)(\boldsymbol{y}-\boldsymbol{\mu}_2)^\top] \end{aligned}$$

and have full rank. We consider the linear combinations $\varphi = \boldsymbol{a}^\top \boldsymbol{x}$ and $\eta = \boldsymbol{b}^\top \boldsymbol{y}$, where $\boldsymbol{a}$ is a p-dimensional and $\boldsymbol{b}$ is a q-dimensional vector.

Then the simple correlation between φ and η is given by

$$\max_{\boldsymbol{a},\boldsymbol{b}} Corr(\varphi, \eta) = \rho_1 \;.$$

37 How do we get the linear combinations for canonical correlation? Why is a matrix product and Eigenvector/Eigenvalue problem involved?

The maximum is achieved by the linear combinations

$$\varphi_1 = \underbrace{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_1^\top} \boldsymbol{x} \quad \text{and} \quad \eta_1 = \underbrace{\boldsymbol{f}_1^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_1^\top} \boldsymbol{y} \ ,$$

which are referred to as **the first pair of canonical variables**. ρ_1 *is called* **the first canonical correlation coefficient**.

The ***k*-th pair of canonical variables** $(k = 2, 3, \ldots, p)$ *is given by*

$$\varphi_k = \underbrace{\boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_k^\top} \boldsymbol{x} \quad \text{and} \quad \eta_k = \underbrace{\boldsymbol{f}_k^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_k^\top} \boldsymbol{y}$$

and maximize

$$Corr(\varphi_k, \eta_k) = \rho_k$$

over all linear combinations that are uncorrelated with the previous $1, 2, \ldots, k-1$ *canonical variables.* ρ_k *is called* ***k*-th canonical correlation coefficient**.

$\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$ *are eigenvalues of the matrix* $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ *and* $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ *are the respective eigenvectors (dimension* p*).*

$\rho_1^2, \rho_2^2, \ldots, \rho_p^2$ *are the* p *largest eigenvalues of the matrix* $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$ *with the respective eigenvectors* $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p$ *(dimension* q*). Every* $\boldsymbol{f}_i$ *is proportional to* $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i$.

The canonical variables have the following properties

$$\begin{aligned} Var(\varphi_k) &= Var(\eta_k) = 1 \\ Cov(\varphi_k, \varphi_l) &= Corr(\varphi_k, \varphi_l) = 0 \qquad k \neq l \\ Cov(\eta_k, \eta_l) &= Corr(\eta_k, \eta_l) \ = 0 \qquad k \neq l \\ Cov(\varphi_k, \eta_l) &= Corr(\varphi_k, \eta_l) \ = 0 \qquad k \neq l \end{aligned}$$

for $k, l = 1, 2, \ldots, p$*; and with the notation* $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)^\top$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$ *and* $\boldsymbol{\rho} = Diag(\rho_1, \ldots, \rho_p)$*, it follows that*

$$Cov\begin{pmatrix} \boldsymbol{\varphi} \\ \boldsymbol{\eta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\rho} \\ \boldsymbol{\rho} & \boldsymbol{I} \end{pmatrix} \ .$$

The canonical correlation coefficients are invariant to linear transformations, which will be shown below.

38 What happens if there is the same variable in X and Y in canonical correlation?

If the same variable is included in both sets XXX and YYY, the first canonical correlation will likely be close to 1, as that variable will have a perfect correlation with itself. This redundancy can lead to overestimation of the strength of the relationship between the two variable sets. It is typically advisable to remove duplicate variables to avoid misleading results in canonical correlation analysis.

Was bedeutet ein kanonischer Korrelationskoeffizient ρ_1 von 1? Es reicht bereits, dass 1 Variable x_i aus $\boldsymbol{x}$ gleich einer Variable y_j aus $\boldsymbol{y}$ ist; ist (oBdA?) z.B. $x_1 = y_1$, (und die übrigen Einträge in $\boldsymbol{x}$ und $\boldsymbol{y}$ beliebig?) kann man mit Vektoren $\boldsymbol{a}_1 = (1, 0, \ldots, 0)^\mathsf{T}$ und $\boldsymbol{b}_1 = (1, 0, \ldots, 0)^\mathsf{T}$ linear kombinieren, so dass sich ein $\rho_1 = 1$ ergibt. Man kann dann nichts über die anderen Variablen aus $\boldsymbol{x}$ und $\boldsymbol{y}$ aussagen. (Kommentar: habe das so in etwa bei der Prüfung erklärt bekommen).

39 What are some hypothesis tests in canonical correlation analysis? (WS20: focus on permutation test)

In canonical correlation analysis, there are a number of **tests**, of which only selected ones are mentioned here, because the distributions of the estimators are very complicated. Assuming that the data are normally distributed, the following applies:

(a) A likelihood ratio test for the hypothesis $H_0 : \boldsymbol{\Sigma}_{12} = \boldsymbol{O}$, i.e. for the hypothesis that $\boldsymbol{x}$ and $\boldsymbol{y}$ are uncorrelated, is given by the test statistic

$$\lambda^{2/n} = |\boldsymbol{I} - \boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}| = \prod_{i=1}^{p}(1 - r_i^2).$$

This test statistic has a $\Lambda(q, n-1-p, p)$ - Wilks distribution. n is the sample size and $r_1, \ldots, r_p$ are the sample canonical correlation coefficients.

(b) Since the Wilks distribution as a product of beta distributions is relatively complex, it can be approximated (with Bartlett's approximation) for large sample sizes n by

$$-\left[n - \frac{1}{2}\Big(p + q + 3\Big)\right] \ln \prod_{i=1}^{p}(1 - r_i^2) \sim \chi_{pq}^2 \quad .$$

(c) A similar test statistic can also be formulated for the hypothesis that only s of the canonical correlation coefficients are non-zero, namely by

$$-\left[n - \frac{1}{2}\Big(p + q + 3\Big)\right] \ln \prod_{i=s+1}^{p}(1 - r_i^2) \sim \chi_{(p-s)(q-s)}^2 \quad .$$

A permutation test is carried out, which does not rely on a distributional assumption. The idea is to permute the observations of one data set, by keeping the other data set unpermuted. Then the canonical correlations are estimated, and this is done many times. Finally, a p-value is determined as the fraction of bootstrap correlation results exceeding the canonical correlation of the unpermuted data.

40 What is the goal of discriminant analysis? What is the expected cost of misclassification, what is involved? How can the ECM be minimized, and how do we arrive at those rules?

It is a multivariate method that deals on the one hand with the classification of different object groups and on the other hand with the assignment of new

objects to previously determined groups. In the former case, the attempt is made to capture the differences of the objects which are known to originate from two or more populations, either graphically or algebraically. One is thus looking for a discriminant function that allows the best possible separation. In the second case, one would like to divide the objects into two or more groups. The goal is then to classify new objects by means of defined rules. The above cases are often directly related, because a function that separates objects can also be used to classify new objects or vice versa.

The objects are classified on the basis of measurements of p underlying random variables $X = (X1, \ldots, Xp)^\top$. Assuming that there are two groups, the objects to be measured are then divided into the classes $\pi1$ and $\pi2$ respectively. Let the sum of the values of the first class be the population of the x-values of $\pi1$, and that of the second class the population of the x-values of $\pi2$. The two populations are then described by the probability distributions $f1(x)$ and $f2(x)$.

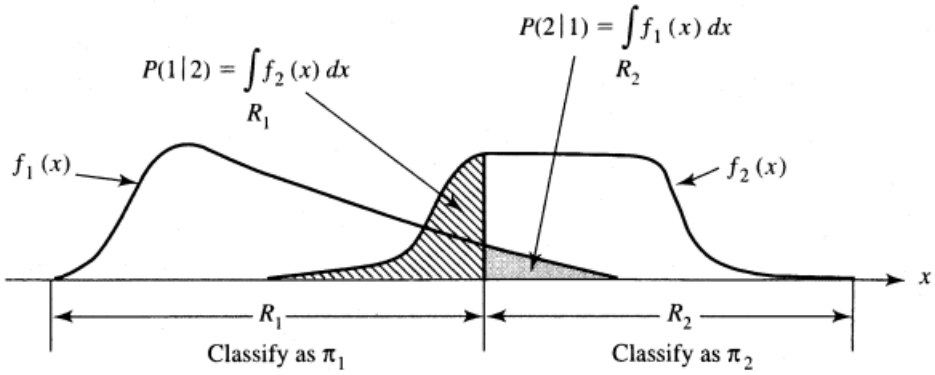
Let Ω be the entire sampling space, i.e. the space containing all observations x. Each object x must come from either population $\pi1$ or $\pi2$. Furthermore, let R1 be the space of observations x to which we assign the objects of $\pi1$, and R2 the space to which the remaining objects of $\pi2$ are assigned. Ω is the union of R1 and R2. It may happen in the case of group assignment that objects which actually belong to population $\pi1$ are falsely classified as $\pi2$. If the probability functions $f1(x)$ and $f2(x)$ are known, this probability of incorrect assignment can be calculated as a conditional probability $P(2|1)$ by:

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.1}$$

Conversely, it may be the case that objects originating from population π_2 are erroneously assigned to π_1. The corresponding probability is

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.2}$$

The integral in (8.1) describes the volume of the density function $f_1(\boldsymbol{x})$ over the region R_2, or analogously for (8.2).



Let p_1 be the probability that the objects come from π_1 (*prior probability*), and p_2 those for π_2, where $p_1 + p_2 = 1$ must hold. Then, the following probabilities can be calculated by applying the formula for conditional probabilities:

$$\begin{aligned}
P(\text{Observation correctly classified as } \pi_1) &= P(\boldsymbol{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \\
P(\text{Observation incorrectly classified as } \pi_1) &= P(\boldsymbol{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2 \\
P(\text{Observation correctly classified as } \pi_2) &= P(\boldsymbol{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2 \\
P(\text{Observation incorrectly classified as } \pi_2) &= P(\boldsymbol{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1
\end{aligned}$$

Missclassification is often directly associated with costs. Of course, the costs are 0 if correctly classified. They are $c(1|2)$ if an observation of π_2 is wrongly classified as π_1. And the cost is $c(2|1)$ if observations of π_1 are erroneously classed as π_2. Together with the probabilities for misclassification, the expected costs for misclassification (ECM) can now be calculated as

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \ . \tag{8.3}$$

The goal of a classification rule is to keep ECM as small as possible.

Theorem 8.2.1 *A classification rule that minimizes ECM is as follows: The set R_1 is defined for observations $\boldsymbol{x}$, for which the following applies:*

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{8.4}$$

The set R_2 is defined for observations $\boldsymbol{x}$, for which the following applies:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{8.5}$$

Theorem 8.2.1 leads to the following special cases:

(a) $p_1 = p_2$:

$$R_1: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2) = c(2|1)$:

$$R_1: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1} \qquad R_2: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{p_2}{p_1} \tag{8.6}$$

(c) $p_1 = p_2$ and $c(1|2) = c(2|1)$:

$$R_1: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1 \qquad R_2: \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < 1$$

However, other criteria can also be used to create a classification rule. For example, one could choose R_1 and R_2 so that the total probability of misclassification (TPM) is minimal, i.e.

$$\begin{aligned} TPM &= P(\text{Misclassification of an observation from } \pi_1 \text{ or } \pi_2) \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \qquad (8.7) \\ &= p_1 P(2|1) + p_2 P(1|2)\ . \qquad (8.8) \end{aligned}$$

Assuming that the cost of misclassification is the same, one immediately realizes that minimizing (8.8) is equivalent to minimizing (8.3).

41 Two-group case: What is linear discriminant analysis, what are the assumptions? How to arrive at the rule for classification? Why is it called linear discriminant analysis, and how do we estimate the involved components?

The rule in case of p-dimensional features reduces the decision to a 1-dimensional variable y, which results from the corresponding linear combinations of the observations of π1 and π2.

42 Two-group case: Explain the downprojection in LDA (graphic would probably help), if the priors and costs are equal.

8.3 The two-group case

We limit ourselves here to multivariate normally distributed populations. Thus, $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are density functions of multivariate normal distributions with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively.

8.3.1 The special case $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

The joint density function of the random variable $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ for the populations π_1 and π_2 are given by

$$f_i(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \quad \text{for } i = 1, 2\,. \tag{8.9}$$

If the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known, then according to (8.4), the region that minimizes ECM is:

$$\begin{aligned} R_1:\ \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \\ \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \\ R_2:\ \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \\ < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \end{aligned} \tag{8.10}$$

The following classification rule can be specified using these regions R_1 and R_2:

Theorem 8.3.1 *Let π_1 and π_2 be normally distributed populations with the same covariance $\boldsymbol{\Sigma}$. Then the classification rule for minimizing ECM is:*
An observation $\boldsymbol{x}_0$ is assigned to π_1, if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right)\,. \tag{8.11}$$

Otherwise $\boldsymbol{x}_0$ is assigned to π_2.

42. **Two-group case: Downprojection in LDA when priors and costs are equal**

- Linear Discriminant Analysis (LDA) projects data onto a one-dimensional axis to maximize class separability.
- If priors and misclassification costs are equal, the decision boundary is determined by the mean vectors and pooled covariance matrix.
- The downprojection means that each sample is projected onto the discriminant axis w, where:

$$w = \Sigma^{-1}(\mu_1 - \mu_2)$$

  with Σ being the within-class covariance matrix and μ_1, μ_2 the class means.
- This projection maximizes the ratio of between-class variance to within-class variance, ensuring optimal classification.

43 Two-group case: Why is QDA called "quadratic"?

8.3.2 The special case $\Sigma_1 \neq \Sigma_2$

The previously formulated classification rules were based on (8.4) and looked at the relation of the density functions $f_1(\boldsymbol{x})/f_2(\boldsymbol{x})$. In the case of the same covariance matrices, this ratio is reduced to a relatively simple term, which is usually expressed by (8.11). In the case of unequal covariance matrices (and unequal means), however, the ratio of density functions becomes a more complicated expression. As with (8.11), the logarithm of the relation can also be considered here:

$$\begin{aligned}\ln\left(\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}\right) &= \ln\left(\frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\right) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ &= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}\end{aligned}$$

$$-\frac{1}{2}\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

This results in the following classification rule:

$$R_1:\ -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}+(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}-k\geq\ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right)$$
$$R_2:\ -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}+(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}-k<\ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right)$$

with

$$k=\frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)+\frac{1}{2}(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1-\frac{1}{2}\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2)\ .$$

The constant k now only depends on the mean and covariance of the two distributions. In the case of $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2$, the above rule is reduced to (8.11). The following statement now follows directly:

Theorem 8.3.3 *Let π_1 and π_2 be normally distributed populations with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively. Then, the classification rule for minimizing ECM is:*
An observation $\boldsymbol{x}_0$ is assigned to π_1, if

$$-\frac{1}{2}\boldsymbol{x}_0^\top(\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0+(\boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}_0-k\geq\ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right)\ . \qquad (8.16)$$

Otherwise $\boldsymbol{x}_0$ is assigned to π_2.

The above discriminant function is *quadratic* in $\boldsymbol{x}$. Since the population parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are not known, they need to be obtained by the respective estimators, e.g. by the empirical sample estimators $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$

43. **Why is QDA called "quadratic"?**

- Quadratic Discriminant Analysis (QDA) extends LDA by allowing each class to have its own covariance matrix Σ_k.
- The decision boundary is no longer linear because the classification rule is based on the log-determinant and inverse of Σ_k, leading to a quadratic form:

$$x^T\Sigma_k^{-1}x+\text{linear terms}+\text{constant}=0$$

- This results in curved decision boundaries, making QDA more flexible than LDA but also more prone to overfitting when data is limited.

44 Two-group case: What is the Fischer criterion to maximize? What is the solution for the projection vector? What is the relation to LDA?

Fisher (1938) developed a linear discriminant function analogously to (8.14), but the idea behind it was different. He tried to transform multivariate observations to univariate, so that the two transformed groups are as strongly separated as possible. The idea was thus to find a direction $\boldsymbol{a} \in I\!R^p$, and to project the observations $\boldsymbol{a}^\top \boldsymbol{x}$ in order to obtain univariate values $y_{11}, y_{12}, \ldots, y_{1n_1}$ for the observations of the first group and values $y_{21}, y_{22}, \ldots, y_{2n_2}$ for the observations of the second group. The separation of the two populations then occurs in such a way that the arithmetic means $\bar{y}_1$ and $\bar{y}_2$ of the univariate y values deviate as much as possible. This difference is expressed in units of the standard deviation and is therefore used as a criterion for the separation

$$\frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \longrightarrow \max$$

with the pooled variance of the y-values

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \; .$$

An attempt is now made to find a linear combination $\boldsymbol{a}$ of $\boldsymbol{x}$, which allows a maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

Theorem 8.3.4 *The linear combination*

$$\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} \tag{8.18}$$

maximizes the ratio

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 - \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled} \hat{\boldsymbol{a}}} \tag{8.19}$$

over all possible vectors $\hat{\boldsymbol{a}}$. $\boldsymbol{S}_{pooled}$ *is defined analogously to the two-group case. The maximum ratio of (8.19) is*

$$D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \; .$$

Theorem 8.3.5 *An observation* $\boldsymbol{x}_0$ *is assigned to* π_1, *if*

$$\hat{y}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}_0 \geq \hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{8.20}$$

Otherwise, $\boldsymbol{x}_0$ *is assigned to* π_2.

This classification rule is shown in Figure 8.4 for $p = 2$. The observations are projected onto a straight line. The direction $\hat{\boldsymbol{a}}$ of the line is varied until the two groups are maximally separated.

In contrast to rule (8.13), Fisher's classification rule does not explicitly require the assumption of normal distribution of both populations. However, in case of violations from this assumption, we know from the previous results that the rule would not be optimal in terms of minimizing TPM (or ECM). Moreover, we also have to assume that the populations have the same covariance matrix, since a pooled estimate of the covariance is used. It can be seen immediately that Fisher's linear discriminant function in (8.20) is a special case of (8.13). If the prior probabilities and the expected costs for misclassification are the same for rule (8.13), which results from minimizing the expected costs of misclassification, we get exactly (8.20).

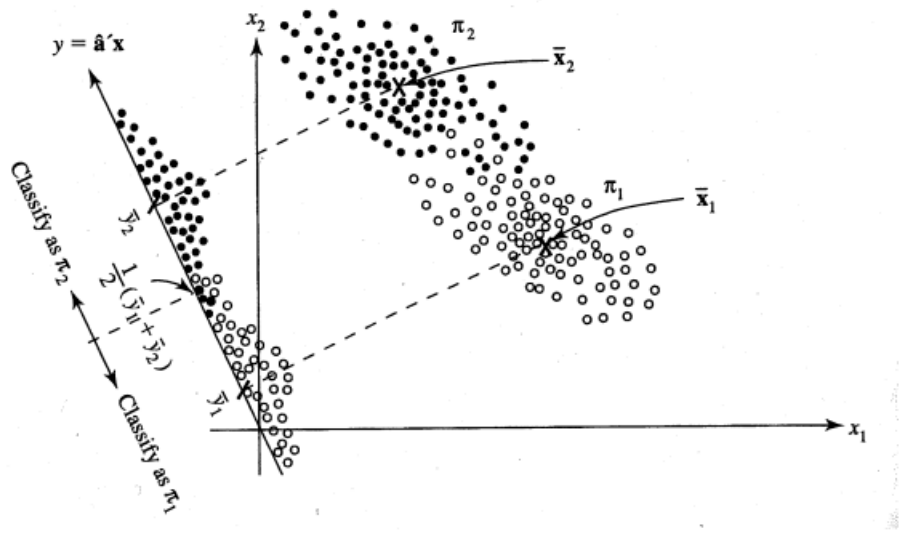


Figure 8.4: Schematic representation of Fisher's classification rule.

45 **Extend the ECM to the multi-group case. What is the resulting decision rule if costs are equal? What are the discriminant functions?**

Deviations from the normal distribution of the groups or different covariances could strongly distort the good theoretical properties.

8.4.1 The method to minimize ECM

Let $f_i(\boldsymbol{x})$ be the density of the observations of population π_i with $i = 1, \ldots, g$. Mostly, it is assumed that $f_i(\boldsymbol{x})$ is the density of a multivariate normal distribution, but this is not a requirement of the following method.

The notation is based on the two-group case. Thus, p_i denotes the prior probability for the population π_i $(i = 1 \ldots, g)$. Further, let R_k be the space of observations

$\boldsymbol{x}$ to which the objects from π_k are assigned. $c(k|i)$ is the cost of misclassification when objects from π_i are mistakenly mapped to π_k $(k = 1, \ldots, g)$, where $k = i$ is of course $c(i|i) = 0$. The probability of this wrong assignment is

$$P(k|i) = P(\boldsymbol{X} \in R_k|\pi_i) = \int_{R_k} f_i(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.21}$$

The conditional expected cost of misclassifying $\boldsymbol{x}$ from π_1 to $\pi_2, \ldots, \pi_g$ are analogous to (8.3)

$$ECM(1) = \sum_{k=2}^{g} P(k|1)c(k|1) \ . \tag{8.22}$$

These expected costs arise with prior probability p_1. $ECM(2), \ldots, ECM(g)$ are analogously defined, and one obtains by multiplying with the prior probabilities and adding up all contributions, the total expected cost of misclassification (ECM) as

$$ECM = \sum_{i=1}^{g} p_i ECM(i) = \sum_{i=1}^{g} p_i \left(\sum_{\substack{k=1 \\ k \neq i}}^{g} P(k|i)c(k|i) \right) \ . \tag{8.23}$$

An optimal classification rule should now yield such ranges $R_1, \ldots, R_g$ (disjoint and complete decomposition of the sample space Ω) so that (8.23) is minimized.

Theorem 8.4.1 *The areas for minimizing ECM (8.23) are given by assigning $\boldsymbol{x}$ to the population π_k $(k = 1, \ldots, g)$ for which the expression*

$$\sum_{\substack{i=1 \\ k \neq i}}^{g} p_i f_i(\boldsymbol{x})c(k|i) \tag{8.24}$$

is minimal.

For simplicity, we assume in the remainder of this chapter that the costs of misclassification are the same for all groups, and thus we can ignore them (or set them equal to 1). Then, minimizing criterion (8.24) corresponds to maximizing the omitted term $p_k f_k(\boldsymbol{x})$, which simplifies the classification rule to:

An observation $\boldsymbol{x}$ is assigned to π_k, if

$$p_k f_k(\boldsymbol{x}) > p_i f_i(\boldsymbol{x}) \quad \text{for all } i \neq k \ . \tag{8.25}$$

Note that the above classification rules can only be applied if the prior probabilities, the misclassification costs, and the density functions are known.

8.4.2 Classification in case of the normal distribution

As an important special case, we consider multivariate normally distributed populations π_i with mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$,

$$f_i(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) \right\} \quad \text{for } i = 1, \ldots, g \ . \tag{8.26}$$

If we now use rule (8.25) or the logarithm of this rule, we get:
An observation $\boldsymbol{x}$ is assigned to π_k, if

$$\ln p_k f_k(\boldsymbol{x}) = \ln p_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) = \max_i \ln p_i f_i(\boldsymbol{x}) \ . \tag{8.27}$$

The constant $(p/2)\ln(2\pi)$ can be omitted in (8.27) since it is the same for all π_i. We therefore define the *quadratic discriminant values* for the i-th population as

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln p_i \quad \text{for } i = 1, \ldots, g \ . \tag{8.28}$$

Thus one obtains the following classification rule:

An observation $\boldsymbol{x}$ is assigned to π_k, if:

$$d_k^Q(\boldsymbol{x}) \quad \text{is the largest of} \quad d_1^Q(\boldsymbol{x}), \ldots, d_g^Q(\boldsymbol{x}) \ . \tag{8.29}$$

Mostly, the $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are unknown and must be estimated. If a "training set" with known class memberships of the observations is available, then the sample means $\bar{\boldsymbol{x}}_i$ and sample covariance matrices $\boldsymbol{S}_i$ can be used to estimate the corresponding parameters of the i-th population ($i = 1, \ldots, g$). This results in the estimated quadratic discriminant values

$$\hat{d}_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i) + \ln p_i \ , \tag{8.30}$$

with the help of which objects are classified analogous to (8.29).

A simplification results if the covariance matrices of the populations are the same, that is $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ for $i = 1, \ldots, g$. In this case, the quadratic discriminant values from (8.28) are

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \ . \tag{8.31}$$

Since the first two terms in (8.31) are the same for all populations, they can be omitted, and we get the *linear discriminant values*

$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad \text{for} \quad i = 1, \ldots, g \ . \tag{8.32}$$

One obtains an estimate of the linear discriminant values from the sample by first estimating a pooled covariance matrix

$$\begin{aligned} \boldsymbol{S}_{pooled} &= \frac{1}{n_1 + \cdots + n_g - g}\Big((n_1 - 1)\boldsymbol{S}_1 + \cdots + (n_g - 1)\boldsymbol{S}_g \Big) \\ &= \frac{1}{\sum_{i=1}^g n_i - g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^\top \ . \end{aligned}$$

This gives the following estimates

$$\hat{d}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} - \frac{1}{2} \bar{\boldsymbol{x}}_i^\top \boldsymbol{S}_{pooled}^{-1} \bar{\boldsymbol{x}}_i + \ln p_i \ . \tag{8.33}$$

The resulting classification rule is:

An observation $\boldsymbol{x}$ is assigned to π_k, if:

$$\hat{d}_k(\boldsymbol{x}) \quad \text{is the largest of} \quad \hat{d}_1(\boldsymbol{x}), \ldots, \hat{d}_g(\boldsymbol{x}) \ . \tag{8.34}$$

Remark: The expression (8.32) is a linear function of $\boldsymbol{x}$. However, one could obtain an analogous rule by ignoring the first term in (8.28), which is the same for all populations in the case of equal covariance matrices. With the corresponding estimated values, one obtains a squared distance

$$D_i^2(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_{pooled}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_i) \ , \tag{8.35}$$

and the classification rule is:

$$\text{Assign } \boldsymbol{x} \text{ to population } \pi_i \text{ for which } -1/2 \ D_i^2(\boldsymbol{x}) + \ln p_i \text{ is the largest.} \tag{8.36}$$

This rule is analogous to rule (8.34), it assigns $\boldsymbol{x}$ to the population that is "closest", with the distance measure penalized with $\ln p_i$. If the a-priori probabilities are not known, one could estimate them by $p_i = 1/g$.

46 What is the idea to extend the two-group Fischer criterion to the multi-group case?

8.4.3 Discriminant analysis by Fisher for the multi-group case

Fisher's discriminant analysis for $g = 2$ can be extended to the multi-group case ($g > 2$), see Rao (1948). For this purpose, we again consider univariate projections of the form $y = \mathbf{a}^T\mathbf{x}$ with $\mathbf{a} \in I\!R^p$ and $\mathbf{a} \neq \mathbf{0}$, but this time a single projection direction $\mathbf{a}$ will not be enough to describe the solution.

Let $\mathbf{a} \neq \mathbf{0}$ be the projection direction we are looking for. The expected value for population π_i (for $i = 1, \ldots, g$) of the random variable $y = \mathbf{a}^T\mathbf{x}$ is then

$$\mu_{i,y} = E(y|\mathbf{x} \in \pi_i) = \mathbf{a}^T E(\mathbf{x}|\mathbf{x} \in \pi_i) = \mathbf{a}^T \boldsymbol{\mu}_i$$

and the variance is

$$\sigma_{i,y}^2 = \mathrm{Var}(y|\mathbf{x} \in \pi_i) = \mathbf{a}^T \mathrm{Cov}(\mathbf{x}|\mathbf{x} \in \pi_i)\mathbf{a} = \mathbf{a}^T \boldsymbol{\Sigma}_i \mathbf{a} \ .$$

The total weighted average of the populations is denoted by $\bar{\boldsymbol{\mu}} = \sum_{i=1}^g p_i \boldsymbol{\mu}_i$, and the corresponding projection into the univariate space by $\bar{\mu}_y = \mathbf{a}^T \bar{\boldsymbol{\mu}}$.

We now make the assumption that the covariances of all groups are the same, that is $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$. Then it also applies to the above variance that

$$\sigma_{1,y}^2 = \ldots = \sigma_{g,y}^2 = \sigma_y^2 = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \ .$$

47 How to arrive at the Fischer multi-group solution (variation within and between groups)?

In Fisher's two-group case, an expression $(\bar{y}_1 - \bar{y}_2)^2/s_y^2$ was maximized. In our notation with random variables, this corresponds to $(\mu_{1,y} - \mu_{2,y})^2/\sigma_y^2$. Additionally, we now consider prior probabilities, and the weighted mean is

$$\bar{\mu}_y = p_1\mu_{1,y} + p_2\mu_{2,y} = \mathbf{a}^T(p_1\boldsymbol{\mu}_1 + p_2\boldsymbol{\mu}_2) \ .$$

Since $p_1 + p_2 = 1$, it is straightforward to see that

$$p_1(\mu_{1,y} - \bar{\mu}_y)^2 + p_2(\mu_{2,y} - \bar{\mu}_y)^2 = p_1p_2(\mu_{1,y} - \mu_{2,y})^2 \ . \tag{8.37}$$

The latter expression (8.37) should therefore be maximized using the Fisher rule (in terms of variance), and this expression describes the weighted sum of the squared distances of the group means to the total mean.

The generalization of (8.37) to the multi-group case is then immediately obvious; now

$$\frac{\sum_{i=1}^{g} p_i(\mu_{i,y} - \bar{\mu}_y)^2}{\sigma_y^2} \tag{8.38}$$

should be maximized. The denominator is according to above $\sigma_y^2 = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$. If the group covariances are not equal, the resulting classification rule will no longer be optimal. $\boldsymbol{\Sigma}$ would then be best replaced by a pooled version, that is

$$\mathbf{W} = \sum_{i=1}^{g} p_i\boldsymbol{\Sigma}_i \ .$$

The matrix $\mathbf{W}$ describes the *variation within groups*.

The numerator of (8.38) can be represented as

$$\sum_{i=1}^{g} p_i(\mu_{i,y} - \bar{\mu}_y)^2 = \sum_{i=1}^{g} p_i(\mathbf{a}^T(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}))^2 = \mathbf{a}^T\mathbf{B}\mathbf{a} \ ,$$

with the matrix

$$\mathbf{B} = \sum_{i=1}^{g} p_i(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^T \ ,$$

which describes the *variation between groups*.

All in all, the maximization problem (8.38) can be expressed as

$$\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \quad \text{for } \mathbf{a} \in I\!R^p,\ \mathbf{a} \neq \mathbf{0} \ . \tag{8.39}$$

Theorem 8.4.2 *The solution of the maximization problem (8.39) is given by the eigenvectors* $\mathbf{a}_1, \ldots, \mathbf{a}_l$ *of the matrix* $\mathbf{W}^{-1}\mathbf{B}$*, which must be scaled, so that* $\mathbf{a}_j^T\mathbf{W}\mathbf{a}_j = 1$ *for* $i = j, \ldots, l$*. The number* l *of the strictly positive eigenvalues of the eigenvalue decomposition of* $\mathbf{W}^{-1}\mathbf{B}$ *is then* $l \leq \min(g-1, p)$*.*

We can now define the *Fisher discriminant functions* as $y_j = \mathbf{a}_j^T \mathbf{x}$, for $j = 1, \ldots, l$, which is the projection of the random variable $\mathbf{x}$ in the direction $\mathbf{a}_j$. If specific data is available, visualizing the first two discriminant functions is particularly interesting because it represents the projection of the data in which the group means appear best separated. Note that in the case of $g = 3$ groups it holds that $l \leq 2$, regardless of whether p is large or not.

Finally, we also want to get a classification rule. For that, consider the *Fisher discriminant values*

$$d_i^F(\mathbf{x}) = \sum_{j=1}^{l} (y_j - \mu_{i,y_j})^2 - 2 \log p_i \tag{8.40}$$

for $i = 1, \ldots, g$. Here $\mu_{i,y_j} = \mathbf{a}_j^T \boldsymbol{\mu}_i$, and one thus has a measure of the deviation from $\mathbf{x}$ to the i-th group mean in the discriminant space, adjusted with the prior probability (analogous to earlier). Note that here in the discriminant space the distance measure is simply the Euclidean distance. A new observation $\mathbf{x}$ is then assigned to population π_k, if $d_k^F(\mathbf{x})$ is the smallest (!) value of all the values of the groups $d_1^F(\mathbf{x}), \ldots, d_g^F(\mathbf{x})$.

By arranging the eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_l$ in the columns of the matrix $\mathbf{A}$, we can also write the *Fisher discriminant values* as

$$d_i^F(\mathbf{x}) = \sum_{j=1}^{l} (\mathbf{a}_j^T (\mathbf{x} - \boldsymbol{\mu}_i))^2 - 2 \log p_i = (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{A}\mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu}_i) - 2 \log p_i \ ,$$

which corresponds to a (squared) Mahalanobis distance in the original space.