Robustness in Risk Classification

Ivan Mendoza

A Thesis

for

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science at

Concordia University

Montréal, Québec, Canada

August 2008

Your file Votre référence
ISBN: 978-0-494-45522-7
Our file Notre référence
ISBN: 978-0-494-45522-7

ABSTRACT

Robustness in Risk Classification

by Ivan Mendoza

Risk classification is a process of grouping together individuals with similar risk levels into categories that insurance companies use in order to decide how premium rates should differ in each category. This process is conditioned on the information available about the insured and the contract, which is stored in many variables.

Because of the large number of variables and the fact that many interactions exist between them, multivariate analysis techniques such as Principal Component Analysys (to reduce the dimensionality of data) and Cluster Analysis (to group individuals with similar characteristics), are applied for this purpose. Here we recommend the application of both methods to obtain better results.

Insurance data usually contains information regarding unexpected extreme losses (catastrophes), modeled with heavy tailed distributions, which may be considered as outliers. Therefore, robust methods for both multivariate techniques are applied by using an algorithm that implies the use of several robust estimators existing nowadays. We compare our results with those obtained from a classical approach.

Acknowledgements

This thesis is the result of a good learning, work, dedication, motivation and effort. First at all, I want to thank someone very special for me who makes me feel alive, energetic and motivated everyday and who I sometimes call God.

I would like to express my gratitude to my supervisor, Dr. José Garrido, whose understanding, expertise and patience, added considerably to my graduate experience. I appreciate his advices in many personal needs. I would also like to thank the Math Department at Concordia University for giving me the financial support and all facilities to conclude my studies. Thanks Dr. P. Gaillardetz and Dr. A. Sen for his time to review my thesis.

A very special thanks goes out to my friends Roudy, Pilar, Edit, Eduardo and Diego, without whose motivation and encouragement I would not have finished this work. Thanks also goes to my friends from Mexico "Los Chiapas" for their invaluable support to start my graduate studies.

Finally, my eternal gratitude to my dear family: Alberto, Teresa, Ivonne and Irving, who perhaps are not able to understand this language and this work, but who have been the real motor in my life. For all your care, advices, support through my entire life and your infinite love, GRACIAS.

Contents

List of Figures

List of Tables

Chapter 1

Risk Classification

1.1 Introduction to Insurance

An insurance contract is a promise of compensation for specific future losses in exchange for periodic *premium* payments. It is designed to protect the financial welfare of an individual, company or entity in the case of unexpected events. Such contracts are called insurance policies. In some cases the policy holder pays part of the loss, called the *deductible*, and the insurer pays the excess of the loss over the deductible. Hence, the purpose of insurance is to transfer the insured's uncertainty of loss to the insurer for the certainty of a smaller premium payment. This uncertainty of loss is called *insurance risk.*

Since the insurer assumes an individual insured's risk, the premium should be based on the expected value of the insured's losses. The expected loss for an insured is the average or probable number of losses (or claims) times the average cost of such claims. The premium should also include the expense of servicing the policy plus a margin for profit and a contingency or reward for taking the risk.

Insurers do not try to predict the actual losses of each insured, but only the expected loss. It is the loss variance that motivates individuals to purchase insurance, while the variation in expected losses, from one individual to another, motivates insurers to price policies differently.

To establish a fair price for insuring an uncertain event, estimates are made from the probabilities associated with the occurrence, frequency and magnitude of such events. These estimates are obtained by using the past experience. In order to derive a price, individuals who are expected to have the same costs are grouped together. The actuary then calculates a price for the group and assumes that this price is applicable to all members of the group.

1.2 Risk Classification Standards

As seen in the previous section, the expected loss of individual policyholders is important in the pricing of insurance. To estimate this vital quantity, insurers use different methods. One of the most important that implies the use of statistical analysis is to observe the experience of a group of similar risks over a short, recent period of time. This grouping of similar risks to estimate costs is called *Risk Classification.*

Furthermore, this group observation process also involves the law of large numbers. If we know the expected losses in advance, then the actual losses will tend to approximate the expected losses at the end of the year for the insurance company as a whole. On the other hand, observing a smaller number of similar risks over a short period of time gives greater confidence that the expected losses are more closely estimated. In Houston (1960), the author notes that if the classes are fairly stable over time, they do not even need to have similar individual expected losses in order to gain a good estimate of the class expected losses. Only the variance of actual losses from the mean for each individual insured in the class should be similar.

Definition 1.1 *We define risk classification as the formulation of different premiums for the same coverage, based on the grouping of risks with similar risk characteristics (rating variables).*

Risk classification is intended simply to group individual risks having reasonably similar expectations of loss. Variables included in a classification system are traditionally chosen so that the following standards or conditions are generally met (for more details, see Walters, 1981):

1. Similar risks should be assigned to the same class with respect to each variable, whereas dissimilar risks should be assigned to different classes, so that there are no clearly identifiable subsets with a significantly different loss potential or expected loss in the same class.

2. The common characteristics used to identify insureds as similar should reasonably relate to the potential for loss.

3. The classes should be exhaustive and mutually exclusive; that is, each insured should belong to at least one, but only one, class with respect to each rating variable.

4. There should be clear and objective phraseology in the definition of classes, with no ambiguity as to what class individual insured belongs.

5. An insured should not be easily able to misrepresent or manipulate his classification.

6. The cost of administering a rating variable should be reasonable in relation to the benefits received.

7. The class rating factors should be susceptible to measurement by actual experience data.

The first standard is what is meant by *homogeneous* classes. Classes that are homogeneous will take fewer risks to obtain reasonable estimates of expected costs. The second standard aids to maintain homogeneous classes by avoiding spurious measures which likely have potentially identifiable subsets. The third, fourth and fifth standards deal with classes being *well–defined* and help to ensure that each risk is actually placed in the right class, avoiding unequal application of the classification system. The two final standards deal with the meaning of *practical*, which is related to "workable, useable and sensible": being cost–effective is important because an inefficient system could increase total costs beyond the value of the information to be obtained.

1.3 Statistical Considerations

Risk classification systems are generally based on statistical analysis, modified by informed judgement. Some considerations of statistical nature, which form part of the design of such a system, are:

- *Homogeneity*: The expected costs for each of the individual risks in a class should be reasonably similar. Significantly dissimilar risks should be assigned to different classes. The occurrence, timing and magnitude of an unexpected event for a specific risk cannot be predicted in advance. Thus, it is inevitable that not all risks in a class will have identical actuarial claim experience. Instead, the individual risk's claim experience will be statistically distributed around the average experience for the class.

  The concept of homogeneity is based upon expected costs as viewed when the risk is originally classified.

- *Credibility*: An important statistical principle is that "the larger the number of observations, the more accurate are the statistical predictions that can be made". Hence, it is desirable that each of the classes in a risk classification system be large enough to allow credible statistical predictions about that class.

- *Predictive Stability*: Elements of a risk classification system must be useful for predicting future costs. This predictive capability must be responsive to changes in the nature of insurance losses. For example, nowadays the impact of automobile bumpers meets certain federal safety standards and at one time very few cars had safe bumpers.

These statistical considerations are sometimes conflicting each other. For example, increasing the number of classes may improve homogeneity, but at the expense of credibility. Consequently, there is no single statistically correct risk classification system. In the final analysis the system adopted will reflect the relative importance of each of these considerations.

The classification of risks is fundamental to any true insurance system in order to group those with similar risk characteristics. This process requires the collection of huge volumes of data from daily operations and it is conditioned on information available about the insured and about the contract, which is called *input profile.* This amount of data is often under-utilized, since insurer's information is stored in hundreds of variables. The analysis of such data involves the application of advanced multivariate statistical analysis and modeling techniques to find useful patterns and relationships, to reduce dimensionality and to create groups in order to provide key facts that can drive decision making. For the purpose of this thesis, multivariate analysis will provide a way to classify the insurer's risk in different classes to apply different premium calculations.

Assuming databases with n vector observations and p variables, one of the most commonly used techniques to reduce dimensionality is *Principal Component Analysis* (PCA). It creates, for each observation, a new set of k transformed variables (*scores*), where $k < p$, that basically reconstructs the original database without loosing much information content and simplifies the graphical representation.

Insurance databases contain information of thousands of insureds and according to the statistical considerations explained above, the purpose of risk classification is to create groups in order to define the premium to be charged to those insureds. This classification can be done by any of the existent multiple clustering methods, detailed in Section 2.3. With these methods we create a partition of the database in subsets, so that individuals in each subset share the same characteristics, based on similarity (or dissimilarity) measures.

In many practical applications, we often find empirical distributions with an asymmetric heavy tail which extend towards positive and negative values skewing results and conclusions. Using multivariate statistical techniques with such distributions is more difficult because extremal points, called *outliers*, have a strong influence on parameter estimation. When the distribution is symmetric (around the mean), the problems caused by outliers can be reduced by using *robust estimation techniques*, which basically try to ignore or downweight the outliers' effect.

In any type of insurance contract we often find individuals who are significantly different to the rest in some attributes in such a way they influence any estimator obtained. Currently, those individuals are treated differently or simply excluded from the database, which sometimes result in errors as insurers do not consider that event's insurer anymore for future predictions. Therefore, the purpose of this thesis is to find a method that allows classifying individuals according to their insurance risk by considering a transformation of their input profile, even when there exist more than one "rare" insureds.

Chapter 2

Multivariate Analysis

2.1 Introduction

Multivariate analysis techniques are useful when observations are obtained from a set of variables of interest, usually called dependent or response variables, and one wants to relate these variables with another set of variables, called independent or predictor variables. The last thirty years have witnessed numerous new results in multivariate analysis, in many different directions.

Some possible practical goals of the analysis are: reduction of the dimensionality (principal component analysis, factor analysis, canonical correlation); identification (discriminant analysis); exploratory models (multivariate linear models). For more details about these techniques, the reader is referred to Johnson (1982).

In the following sections, matrices will be denoted with capital bold letters. We will assume that our data matrix $\mathbf{X}$ has dimension $(n \times p)$, where n denotes the number of observations and p the number of variables. An observation vector will be always indicated in lower case bold letters, e.g., $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ stands for the ith observation and for a random vector we will use capital letters, e.g., the vector

$X = [X_1, ..., X_p]' \in \mathbb{R}^p$ indicates a vector of random vectors $X_1, ..., X_p$, where each random vector contains n observations. Therefore, we can sometimes define the vector $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, where each $\mathbf{x}_i$ is defined as before. Finally, F means a distribution on $\mathbb{R}^p$.

In the classical approach, the *location* parameter of a p-variate random variable X is given by its expectation:

$$\boldsymbol{\mu} = \mathbb{E}(X) = [\mathbb{E}(X_1), ..., \mathbb{E}(X_p)]' = [\mu_1, ..., \mu_p]'$$

and its *dispersion* is described by the covariance matrix:

$$\text{Cov}(X) = \mathbb{E}[(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})'] = [\sigma_{ij}] = \Sigma,$$

where $\sigma_{ij} = \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$ and $\sigma_{ii} = \mathbb{E}[(X_i - \mu_i)^2] = \mathbb{V}(X_i)$ for $i, j = 1, ..., p$ and Σ is a symmetric and positive semidefinite matrix.

As a result we have that for each deterministic vector $\mathbf{a}$ and matrix $\mathbf{A}$, the mean and variance work well under the *affine equivariance* property:

$$\mathbb{E}(\mathbf{A}X + \mathbf{a}) = \mathbf{A}\mathbb{E}(X) + \mathbf{a},$$

$$\text{Cov}(\mathbf{A}X + \mathbf{a}) = \mathbf{A}\Sigma\mathbf{A}'.$$

Hence, the data collected are usually displayed in matrix form, where the rows represent observations and the columns represent the variables. When the multivariate responses are samples from one or more populations one makes the assumption that those samples come from a multivariate probability distribution. In many texts, the multivariate normal distribution (MND) is typically the assumed distribution.

Considering $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, with each $\mathbf{x}_i$ having p-variate normal $N_p(\boldsymbol{\mu}, \Sigma)$ distribution with density:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} \exp\Big(- \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\Big), \tag{2.1}$$

where $|\Sigma|$ stands for the determinant of Σ.

Under this distribution, the MLEs of $\boldsymbol{\mu}$ and Σ for sample X are respectively the sample mean and sample covariance matrix:

$$\overline{\mathbf{x}} = \text{ave}(X) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i,$$

$$\widehat{\Sigma} = \text{ave}\{(X - \overline{\mathbf{x}})(X - \overline{\mathbf{x}})'\},$$

which also work well under the affine transformation of the sample mean and covariances.

As in the univariate case, we define the Mahalanobis distance between vectors $\mathbf{x}$ and $\boldsymbol{\mu}$ with respect to the matrix Σ as:

$$d(\mathbf{x}, \boldsymbol{\mu}, \Sigma) = (\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}). \tag{2.2}$$

It is known (Seber, 1984) that if $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, then $d(\mathbf{x}, \boldsymbol{\mu}, \Sigma) \sim \chi^2_p$.

Then, we can define the multivariate outlyingness measure as $D_i = d(\mathbf{x}_i, \overline{\mathbf{x}}, \mathbf{S})$, with $\mathbf{S} = \widehat{\Sigma}$. Thus, assuming that the estimators $\overline{\mathbf{x}}$ and $\mathbf{S}$ very close to their true values, we may examine the Q-Q plot of D_i vs. the quantiles from a χ^2_p distribution to detect observations for which D_i is "too high". This approach can be useless when n is small.

The zero-order Pearson correlation between two random variables X_i and X_j is given by:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_{ii}\sigma_{jj}} = \frac{cov(X_i, X_j)}{\sqrt{\mathbb{V}(X_i)\mathbb{V}(X_j)}}, \tag{2.3}$$

where $-1 \leq \rho_{ij} \leq 1$. The correlation matrix for the random vector X is:

$$\boldsymbol{\rho} = [\rho_{ij}].$$

As a result, we have:

$$\boldsymbol{\rho} = (\text{diag}\Sigma)^{-1/2}\Sigma(\text{diag}\Sigma)^{-1/2}$$

$$\Sigma = (\text{diag}\Sigma)^{1/2}\boldsymbol{\rho}(\text{diag}\Sigma)^{1/2}$$

where $(\text{diag}\Sigma)^{-1/2}$ represent the diagonal matrix with diagonal elements equal to the square root of the diagonal elements of Σ.

2.2 Principal Component Analysis

The first introduction of Principal Component Analysis (PCA) was made by Karl Pearson in 1890 and a formal treatment was made by Hotelling (1933) and Rao (1964). In PCA we transform a set of p correlated variables into a smaller set of uncorrelated hypothetical constructs called principal components (PCs). None of the variables is assumed as dependent and no grouping of observations is used. The PCs are used to discover the relationship or dependence existing among all variables.

More details about PCA are given in Johnson (1982) and Timm (2002).

2.2.1 Model for PCA

Principal Component Analysis is a dimension reduction technique that deals with a random vector $Y' = [Y_1, ..., Y_p]$, with a vector mean $\boldsymbol{\mu}$ and covariance matrix Σ of full rank p. If the variables are correlated, the swarm of points is not oriented parallel to any of the axes represented by those variables. Therefore, we find natural axis of the ellipsoid generated by those points by translating the origin to $\boldsymbol{\mu}$ and then rotating the axis. After rotation, the new variables (principal components) will be uncorrelated and the goal is that the variance of the jth component is maximal, $j = 1, ..., p$.

The simplest way to reduce the dimension is to take only one element of the observed vector and get rid of all others, but this approach is not reasonable, since we lose the interpretation of the data. An alternative method is to weight all variables equally, but this again is not desirable.

Thus, we take the first principal component of Y as the linear combination $Z_1 = \mathbf{p}_1'Y = p_{11}Y_1 + ... + p_{1p}Y_p$. One aim is to maximize the variance of the projection Z_1, i.e., to choose $\mathbf{p}_1$ such that

$$\mathbb{V}(Z_1) = \mathbb{V}(\mathbf{p}_1'Y) = \mathbf{p}_1'\Sigma\mathbf{p}_1 \tag{2.4}$$

is maximal, subject to the constraint that $\mathbf{p}_1'\mathbf{p}_1 = 1$. This condition is imposed to ensure the uniqueness of the principal component.

Using the method of Lagrangian multipliers, we find that:

$$L = \mathbf{p}_1'\Sigma\mathbf{p}_1 - \lambda_1(\mathbf{p}_1'\mathbf{p}_1 - 1),$$

where $\mathbf{\Sigma}$ is the covariance matrix of Y. Therefore,

$$\frac{\partial L}{\partial \mathbf{p}_1} = 2\Sigma\mathbf{p}_1 - 2\lambda_1\mathbf{p}_1 = \mathbf{0} \Leftrightarrow [\mathbf{\Sigma} - \lambda_1\mathbf{I}]\mathbf{p}_1 = \mathbf{0},$$

where $\mathbf{0}$ is a column vector with each of its p entries equal to zero.

In order to have a nontrivial solution, the determinant $|\mathbf{\Sigma} - \lambda_1\mathbf{I}|$ must be equal to 0. Computing the determinant with λ_1 as an unknown, produces a p degree polynomial in λ_1 that, when we set equal to zero, constitutes the characteristic equation. Any of the nonzero solutions of that equation can be plugged into the matrix equation and the resulting set of equations provide us the coefficients for $\mathbf{p}_1$.

Hence, each root of the characteristic equation λ_1 makes $[\mathbf{\Sigma} - \lambda_1\mathbf{I}]\mathbf{p}_1 = \mathbf{0}$ true and this happens iff $\mathbf{p}_1'\Sigma\mathbf{p}_1 = \lambda_1\mathbf{p}_1'\mathbf{p}_1$. However, we require that $\mathbf{p}_1'\mathbf{p}_1 = 1$, thus

$\lambda_1 = \mathbf{p}_1'\Sigma\mathbf{p}_1$. That means, each root of the characteristic equation is equal to the variance of the combined variable generated by the coefficients of the characteristic vector associated with that root. Since we need this variance to maximize (2.4), the coefficients of the first principal component will be the characteristic vector associated with the *largest* characteristic root.

To determine the second principal component, the linear combination $Z_2 = \mathbf{p}_2'Y$ is constructed such that it is uncorrelated with Z_1 and has maximal variance. To have Z_2 to be uncorrelated with Z_1, the covariance between Z_2 and Z_1 must be zero. However, $\Sigma\mathbf{p}_1' = \mathbf{p}_1\lambda_1$ so that,

$$0 = \mathrm{Cov}(Z_2, Z_1) = \mathbf{p}_2'\Sigma\mathbf{p}_1 = \mathbf{p}_2'\mathbf{p}_1\lambda_1,$$

implies that $\mathbf{p}_2'\mathbf{p}_1 = 0$. Therefore, to maximize $\mathbb{V}(Z_2) = \mathbf{p}_2'\Sigma\mathbf{p}_2$, we find that,

$$L = \mathbf{p}_2'\Sigma\mathbf{p}_2 - \lambda_2(\mathbf{p}_2'\mathbf{p}_2 - 1) - \theta(\mathbf{p}_2'\mathbf{p}_1),$$

then

$$\frac{\partial L}{\partial \mathbf{p}_2} = 2\Sigma\mathbf{p}_2 - 2\lambda_2\mathbf{p}_2 - \theta\mathbf{p}_1 = \mathbf{0} \Leftrightarrow 2[\Sigma - \lambda_2\mathbf{I}]\mathbf{p}_2 = \theta\mathbf{p}_1,$$

which implies (by multiplying both sides by $\mathbf{p}_1'$),

$$2\mathbf{p}_1'[\Sigma - \lambda_2\mathbf{I}]\mathbf{p}_2 = \theta\mathbf{p}_1'\mathbf{p}_1,$$

therefore, after expanding the last equation and using that $\mathbf{p}_2'\mathbf{p}_1 = 0$:

$$\theta = 2\mathbf{p}_1'\Sigma\mathbf{p}_2.$$

However, by definition of $\mathbf{p}_1$, we know that $\Sigma\mathbf{p}_1 = \lambda_1\mathbf{p}_1$, therefore

$$\mathbf{p}_2'\Sigma\mathbf{p}_1 = \mathbf{p}_2'\lambda_1\mathbf{p}_1 = \mathbf{0}.$$

Then, we have that $\theta = 0$, and $\mathbf{p}_2$ must satisfy the equations

$$[\Sigma - \lambda_2\mathbf{I}]\mathbf{p}_2 = 0.$$

Thus $\mathbf{p}_2$ will be one of the characteristic vectors of Σ and, given our goals, it will be the characteristic vector associated with the second largest characteristic root, that means $\lambda_1 \geq \lambda_2$.

More generally, by Theorem $A.1$ (Spectral Decomposition Theorem) there exists an orthogonal matrix $\mathbf{P}(p \times p)$, which is conformed by *eigenvectors* of Σ, such that

$$\mathbf{P}'\mathbf{\Sigma}\mathbf{P} = \mathbf{\Lambda} = diag[\lambda_j],$$

where $\lambda_1 \geq \lambda_2 \geq ...\lambda_p \geq 0$. Therefore, we set

$$\mathbf{Z} = \mathbf{P}'Y, \tag{2.5}$$

where the $\mathbb{E}(\mathbf{Z}) = \mathbf{P}'\boldsymbol{\mu}$, the $\mathrm{Cov}(\mathbf{Z}) = \mathbf{\Lambda}$ and the jth element Z_j of $\mathbf{Z}$ is the jth *principal component* of Y.

As a result, we have:

$$\sum_{j=1}^{p} \mathbb{V}(Y_j) = tr(\Sigma) = tr(\Lambda) = \sum_{j=1}^{p} \mathbb{V}(Z_j).$$

Note that PCA does not require the normality assumption for the estimation of principal components, since any matrix can be decomposed using Theorem $A.1$. However, if $Y \sim N_p(\mathbf{0}, \Sigma)$, we know that Y has associated a constant density ellipsoid of the form $Y'\Sigma^{-1}Y = Q > 0$ centered at the mean $\boldsymbol{\mu} = 0$. Thus, (2.5) represents a rotation of the old axes into the new principal axis so that the new components in the new coordinate system are uncorrelated and have maximum variance. Hence, PCA is a procedure to transform a MND into a set of independent normal distributions, because under normality, no correlation implies independence.

It is well known that principal components are not independent of the scales in which the variables are measured and it is very common to find variables with different scales, specially in Risk Theory. Therefore it is recommended to standardize the variables as a first step, which means to derive the PCs from the *correlation matrix* $\boldsymbol{\rho}$.

Let $Y^* = Y - \boldsymbol{\mu}$, then the principal components with mean zero and variance λ_j are given by:

$$C_j = \mathbf{p}_j' Y^* = \mathbf{p}_j'(Y - \boldsymbol{\mu}). \tag{2.6}$$

To standardize the components with regard to both location and scale, we construct the standardized components

$$Z_j^* = \mathbf{p}_j'(Y - \boldsymbol{\mu})/\sqrt{\lambda_j} = C_j/\sqrt{\lambda_j},$$

so that the $\mathbb{E}(Z_j^*) = 0$ and the $\mathbb{V}(Z_j^*) = 1$ for $j = 1, 2, ..., p$.

In matrix notation,

$$\mathbf{C} = \mathbf{P}'(\mathbf{Y} - \boldsymbol{\mu}\mathbf{1}_n)$$

$$\mathbf{Z}^* = (\Lambda^{1/2})^{-1}\mathbf{P}'(\mathbf{Y} - \boldsymbol{\mu}\mathbf{1}_n) = (\Lambda^{1/2})^{-1}\mathbf{C},$$

where $\mathbf{1}_n$ is the column vector with all n components equal to 1 and Λ is a diagonal matrix with diagonal elements equal to the eigenvalues. Then:

$$\mathrm{Cov}(\mathbf{Z}^*) = (\Lambda^{1/2})^{-1}\Sigma(\Lambda^{1/2})^{-1} = \boldsymbol{\rho}.$$

Hence, the principal components of $\mathbf{Z}$ may be obtained from the eigenvectors of the *correlation* matrix $\boldsymbol{\rho}$ of $\mathbf{Y}$. All previous results hold when using standardize variables, with some simplifications. However, the results derived from $\boldsymbol{\Sigma}$ are, in general, not the same as the ones derived from $\boldsymbol{\rho}$.

The relationship between $\mathbf{Y}$ and the principal components can be expressed as follows:

$$\mathbf{Y} = \mathbf{PZ},$$

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}_n + \mathbf{PC},$$

$$\mathbf{Y} = \boldsymbol{\mu}\mathbf{1}_n + \mathbf{P}\Lambda^{1/2}\mathbf{Z}^* = \boldsymbol{\mu}\mathbf{1}_n + \mathbf{Q}\mathbf{Z}^*,$$

where the matrix $\mathbf{Q} = [q_{hj}]$ is called the covariance loading matrix, that is, the covariance between $\mathbf{Y}$ and $\mathbf{Z}^*$. Therefore, the covariance between Y_h and the j^{th} standardized principal component, $\mathbf{Z}_j^*$, is:

$$\text{Cov}(Y_h, \mathbf{Z}_j^*) = p_{hj}\sqrt{\lambda_{hj}} = q_{hj}, \qquad h = 1, ..., p.$$

By selecting only k components $C_1, C_2, ..., C_p$, where $k \leq p$, the variable Y_h can be estimated as:

$$\widetilde{Y_h} = \boldsymbol{\mu}_h + \sum_{j=1}^{k} q_{hj} Z_j^*, \qquad h = 1, ..., p,$$

so that

$$\widetilde{\mathbf{Y}} = \boldsymbol{\mu} + \mathbf{Q}\mathbf{Z}^*.$$

In practice, we use the sample mean $\overline{Y}$, sample covariance $\mathbf{S}$ and sample correlation matrices $\mathbf{R}$ as estimators for the population parameters $\boldsymbol{\mu}$, Σ and $\boldsymbol{\rho}$, respectively.

2.2.2 Number of Components

In order to effectively summarize the data, we have several criteria to define the number of components:

- Retain sufficient components to have a specified percentage of the total univariate variance (ρ_k^2) accounted by k principal components, say $70\% - 80\%$, that is:

$$\rho_k^2 = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} = \frac{\sum_{j=1}^{k} \lambda_j}{tr(\Sigma)}.$$

- Retain the components whose eigenvalues are greater than the average of the eigenvalues. When using a correlation matrix, this average is 1.

- Use the *scree graph*, a plot of eigenvalues λ_i versus i, and look for a natural break between the large eigenvalues and the small ones.

- Test the significance of the larger components. First, it may be useful to conduct a preliminary test for complete independence (which requires the assumption of normality) among all variables. If independence holds, there is no point in getting principal components, since the variables themselves already provide all the essential information.

To test the significance of the larger components, we test the hypothesis that the last k eigenvalues are small. The implication is that the first sample components capture all the essential information, whereas the last ones reflect noise. The test statistic, which has an approximate χ^2-distribution, is:

$$u = (n - \frac{2p+11}{6})(k \ln \overline{\lambda} - \sum_{i=p-k+1}^{p} \ln \lambda_i),$$

where $\overline{\lambda} = \sum_{i=p-k+1}^{p} \frac{\lambda_i}{k}$. We reject H_0 if $u \geq \chi^2_{\alpha,v}$ and $v = \frac{1}{2}(k-1)(k+2)$.

In practice, when data are highly correlated and data can be represented by a small number of principal components, the first three methods will usually agree on the number of components to retain.

2.2.3 Outlier Detection Using PCA

An outlier is an observation that is numerically distant from the bulk of data. Estimators derived from data sets including outliers may provide wrong results.

Principal Component Analysis is sensitive to the presence of outliers. An extreme outlier may generate a single component. If we cannot remove the outlier from the data matrix, we may use a robust estimator of the covariance or correlation matrix.

Rao (1964) suggests investigate the distances:

$$D_i^2 = (Y - \widetilde{Y})'(Y - \widetilde{Y}) = \sum_{j=k+1}^{p} (q_{ij} Z_j^*)^2 \sim \chi^2_{(p-k)}.$$

An informal plot of $D_1^2, ..., D_n^2$ may be used to detect outliers. Hawkins (1974) suggests a modification of the previous formula by dividing each term D_i^2 by $\widehat{\lambda}_j$ to improve convergence to a chi-square distribution.

To measure the degree of outlyingness, we will use two distances: the *score distance* (SD_i) and the *orthogonal distance* (OD_i). The first one is nothing but the square Mahalanobis distance, defined in (2.2), applied to the principal components and eigenvalue vectors obtained by PCA (that is, information about the covariance of the scores), so that we have:

$$SD_i = \sqrt{C_i' \Lambda^{-1} C_i} = \sqrt{\sum_{j=1}^{k} \frac{c_{ij}^2}{\lambda_j}}, \tag{2.7}$$

for $i = 1, ..., n$. The orthogonal distance, that is, the distance between the observations and their projections in the k-dimensional PCA, is defined as:

$$OD_i = ||\mathbf{y}_i - \widetilde{\mathbf{y}}_i||, \quad i = 1, ..., n. \tag{2.8}$$

To distinguish between regular observations and outliers, we construct a *diagnostic plot*, where on the horizontal axis it graphs the score distance and on the vertical axis, the orthogonal distance of each observation.

To classify the observations we draw two cut–off lines. The *horizontal* cut–off line is given by $\sqrt{\chi^2_{k,1-(\alpha/2)}}$, because the squared Mahalanobis distance of normally distributed scores is approximately χ^2_k distributed and usually α is 5%. The vertical cut-off line is more difficult to determine, because the distribution of the orthogonal distances is not exactly known. However, Box (1954) gives a good approximation for this unknown distribution, given by $g_1 \chi^2_{g_2}$, where g_1 and g_2 are unknown parameters that can be estimated by the method of moments (Nomikos and MacGregor, 1995) or by the approximation of a chi-squared distribution to the normal distribution of Wilson–Hilferty (1931).

Using the latter implies that the orthogonal distance to the power 2/3 is approximately normally distributed with mean and variance:

$$\mu = (g_1 g_2)^{1/3}(1 - \frac{2}{9g_2}),$$

$$\sigma^2 = \frac{2g_1^{2/3}}{9g_2^{1/3}}.$$

We can obtain estimates of $\widehat{\mu}$ and $\widehat{\sigma}$ using the univariate MCD, defined in Section 3.2.4. The *vertical* cut-off value is then equal to $(\widehat{\mu} + \widehat{\sigma} z_{0.975})^{3/2}$, where $z_{0.975}^{3/2} = \Phi(0.975)$ is the 97.5% quantile of the Gaussian distribution.

The orthogonal and the score distances now define four types of observations: *Regular* observations have small orthogonal and score distances. When samples have a large score distance but small orthogonal distance, we call them *Good leverage* observations. In Figure 2.1 we can identify observations 1 and 2 in this latter category, i.e., these observations lie close to the space spanned by the principal components but far from the regular data. This implies that they are very different from the bulk, but there is only a small loss of information, when they are replaced by their fitted values in the PCA–subspace. *Orthogonal* outliers have a large orthogonal distance, but a small score distance, see case 3 in Figure 2.1. They cannot be distinguished from regular observations once they are projected onto the PCA subspace, but they lie far from this subspace. Therefore, it would be dangerous to replace that sample with its projected value, as its outlyingness would not be visible anymore. *Bad leverage* observations, such as observations 4 and 5 in Fig 2.1, have large orthogonal and score distances. They lie far outside the space spanned by the principal components and after projection far from the regular data points. Their degree of outlyingness is high

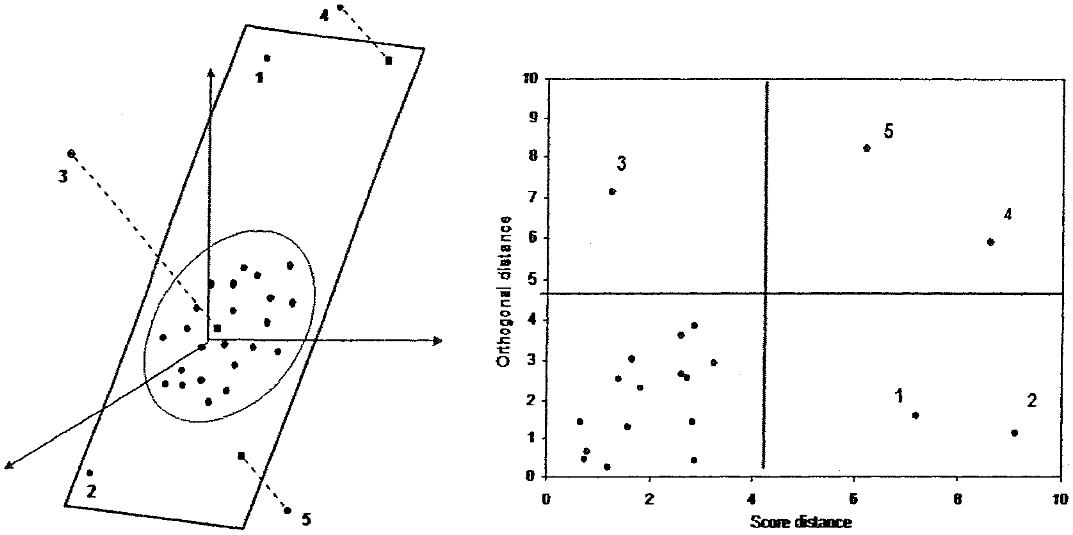


Figure 2.1: Different types of outliers.

in both directions and they have a large influence in the classical PCA.

Thus, the outlier map displays the OD_i vs SD_i and it classifies observations according to the lines given by the cut-offs described before.

2.3 Cluster Analysis

In *cluster analysis*, (CLAN) we look for patterns in data by grouping observations into clusters. The goal is to find an optimal grouping for which the observations within each cluster are similar, while observations in different groups are not similar. For more references and details about CLAN see Timm (2002).

Unlike the previous section data will be represented as a $n \times p$ matrix:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_n' \end{bmatrix},$$

where $\mathbf{y}_i = (y_{i1}, ..., y_{ip}), i = 1, ..., n$ is an observation vector.

Since CLAN attempts to identify the observation vectors that are similar in order to group them into clusters, the technique uses an index of similarity or dissimilarity between each pair of observations. A convenient measure of proximity is the distance between two observations, which is considered a measure of dissimilarity.

One of the most common distance function is the Euclidean distance between two objects. Given the matrix $\mathbf{Y}$, the square of the Euclidean distance between two rows $\mathbf{y}_r$ and $\mathbf{y}_s$ is defined as:

$$d_{rs}^2 = (\mathbf{y}_r - \mathbf{y}_s)'(\mathbf{y}_r - \mathbf{y}_s) = \|\mathbf{y}_r - \mathbf{y}_s\|^2.$$

The $(n \times n)$ data matrix $\mathbf{D} = [d_{rs}]$ is called the Euclidean distance matrix, which is a special case of the Minkowski metric (L_λ-norm), the dissimilarity measure may be represented as:

$$d_{rs} = \left(\sum_{j=1}^{p} |y_{rj} - y_{sj}|^\lambda \right)^{1/\lambda},$$

varying λ to change the weight assigned to larger and smaller distances.

To eliminate the dependence on the units of measurement in variables, we can use the square of the Mahalanobis distance, defined in (2.2), as a proximity measure as follows:

$$d_{rs}^2 = (\mathbf{y}_r - \mathbf{y}_s)'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_r - \mathbf{y}_s) = \|\mathbf{y}_r - \mathbf{y}_s\|^2,$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{Y}$.

To initiate a CLAN, one constructs a proximity matrix to represent the strength of the relationship between pair of rows on $\mathbf{Y}$. There are essentially two types of clustering: *hierarchical* and *partitioning* clustering methods.

The hierarchical algorithm includes agglomerative and splitting procedures. The first type starts from the finest partition possible (each observation forms a cluster) and groups them, whereas the second type starts with a cluster that contains all of the observations. Both of them are used to cluster either observations or variables. The partitioning algorithm starts from a given definition of clusters and proceeds by exchanging elements between groups until a certain score is optimized. The first algorithm is the most used in practice.

When clustering items, hierarchical and partitioning clustering methods may be combined to get a better identification of clusters. As a first step, one may use a hierarchical procedure to identify the seeds and number of clusters, then one can use a partitioning method to refine the cluster solution.

2.3.1 Hierarchical Algorithms

This method uses the elements of a proximity matrix to generate a tree diagram or dendrogram. Steps are as follows:

1. Begin with n clusters, each containing only a single object.

2. Search the dissimilarity matrix $\mathbf{D}$ for the most similar pair. Let the pair chosen be associated with element d_{rs} so that object r and s are selected.

3. Combine objects r and s into a new cluster (rs) using some *criterion* and reduce the number of clusters by 1 by deleting the row and column for objects r and s. Calculate the dissimilarities between the cluster (rs) and all remaining clusters, using the criterion, and add the row and column to the new dissimilarity matrix.

4. Repeat steps 2 and 3, $(n-1)$ times until all objects form a single cluster. At each step, identify the dissimilarity at which the clusters are merged.

Letting $r \in R$ represent any element in cluster R and $s \in S$ be any element in cluster S, one computes the distance between the new group $R + S$ and any other cluster T. For this step, the following distance function is used:

$$d(T, R+S) = \delta_1 d(T,R) + \delta_2 d(T,S) + \delta_3 d(R,S) + \delta_4 |d(T,R) - d(T,S)|.$$

The δ's are weighting factors that change the criterion in Step 3 above, according to Table 2.1. Here n_R, n_S, n_T are the number of observations in cluster R, S and T, respectively.

Name	δ_1	δ_2	δ_3	δ_4
Single linkage (Nearest-Neighbor)	1/2	1/2	0	-1/2
Complete linkage (Farthest-Neighbor)	1/2	1/2	0	1/2
Average linkage (unweighted)	1/2	1/2	0	0
Average linkage (weighted)	$\frac{n_R}{n_R+n_S}$	$\frac{n_S}{n_R+n_S}$	0	0
Centroid	$\frac{n_R}{n_R+n_S}$	$\frac{n_S}{n_R+n_S}$	$-\frac{n_R n_S}{(n_R+n_S)^2}$	0
Median	1/2	1/2	-1/4	0
Ward (Incremental sum of squares)	$\frac{n_R+n_T}{n_R+n_S+n_T}$	$\frac{n_S+n_T}{n_R+n_S+n_T}$	$-\frac{n_T}{n_R+n_S+n_T}$	0

Table 2.1: Different agglomerative hierarchical clustering methods.

2.3.2 Partitioning Algorithms

These methods are only applied to cluster observations and one knows a priori the number of clusters k, which are either centroids or seeds. The process is initiated using the raw data matrix $\mathbf{Y}$ and not a dissimilarity matrix $\mathbf{D}$. It usually follows the following steps:

1. Select k centroids or seeds.

2. Assign each observation to the nearest centroid using some L_λ-norm, usually the Euclidean distance.

3. Reassign each observation to one of the k clusters based upon some criterion.

4. Stop if there is no reallocation of observations or if reassignment meets some convergence criterion; otherwise, return to step 2.

The k seeds might be the first k observations at some defined level of separation, k random seeds and other variations. Once seeds are selected, each observation is evaluated for assignment or reassignment using multivariate statistics that involve the determinant and trace of the within and between-cluster variability.

K-means Algorithm

This method was first studied by MacQueen (1967) and extended by Anderberg (1975) or Hartigan and Wong (1979) provided some modifications. The basics steps are as follows:

1. Select k seeds.

2. Assign each of the $n - k$ observations to the nearest seed and recalculate the cluster centroid (mean, median, or other depending on the L_λ-norm).

3. Repeat step 2 until all observations are assigned or until changes in clusters centroids become small (no reassignments are made in cluster membership).

In step 2, the seed may or may not be updated. Two tests may be made for seed replacement. An observation may replace one of a pair of seeds if the distance between the seeds is less than the distance between an observation and the nearest seed. The former seed becomes an observation in the recalculation of the centroid. If an observation fails this test, we can use another one. The observation replaces the nearest seed if the smallest Euclidean distance from the observation to all seeds, other than the nearest one, is greater than the shortest distance from the nearest seed to all seeds. For one pass of the data, all observations are associated with k clusters.

This process is repeated until all changes in clusters seeds become small based upon a convergence criterion.

2.3.3 The Silhouette Value

In order to measure the power of the algorithm used to classify observations, we can construct the *silhouette plot* (Rosseeuw, 1987) and a corresponding quality index allowing to select the optimal number of clusters. For each object i we denote by A the cluster to which it belongs and compute:

$$a(i) := \frac{1}{|A|-1} \sum_{j \in A, j \neq i} d(i,j),$$

that is, the average dissimilarity of i to all other objects of A.

Now consider any cluster C, different from A, and put

$$d(i,C) := \frac{1}{|C|} \sum_{j \in C} d(i,j),$$

that is, the average dissimilarity of $i \in A$ to all other objects of C. Then, for all clusters $C \neq A$ we take the smallest of the distances:

$$b(i) := \min_{C \neq A \ni i} d(i,C).$$

The cluster B which attains this minimum, that is, $d(i,B) = b(i)$, is called the *neighbor* of object $i \in A$. The *silhouette value* $s(i)$ is then defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{2.9}$$

Clearly, $s(i)$ always lies between -1 and 1. The value $s(i)$ may be interpreted as follows:

- $s(i) \approx 1 \Rightarrow$ object i is well classified (in A).

- $s(i) \approx 0 \Rightarrow$ object i lies intermediate between two clusters (A and B).

- $s(i) \approx$ -1 $\Rightarrow$ object i is badly classified (closer to B than to A).

The silhouette of the cluster A is a plot of all its $s(i)$, ranked in decreasing order. The entire plot shows the silhouettes of all clusters below each other, so that quality of the clusters can be compared: a wide silhouette is better than a narrow one. Finally, the quality index mentioned before is the *overall average silhouette width* of the silhouette plot, defined as the average of the $s(i)$ over all objects i in the data set.

Chapter 3

Robustness

All statistical methods are based on model assumptions about the data analyzed for a problem, where some are only a crude approximations of reality. For instance, the most widely used model assumption is that the observed data follow a normal (*Gaussian*) distribution. In practice this model may describe well the majority of observations, even if some others may seem to follow a different pattern. They are identified as atypical data, in the sense that they are observations far from the bulk of the data. They are called *outliers.*

3.1 Robust Univariate Estimators

The robust approach to statistical modeling and data analysis aims at deriving methods that produce reliable parameter estimates, irrespective of the distribution assumed. That is, if the data does not contain outliers, the robust method gives approximately the same results as the classical approach. While if a small proportion of outliers are presented, the robust method gives approximately the same result as the classical method applied to the "typical" data. The first great steps forward to the idea of statistical robustness occurred in the 1960's with the work of John Tukey (1960, 1962), Peter Huber (1964, 1967) and Frank Hampel (1971, 1974).

Many location estimators, such as the mean, are highly sensitive to outliers. In insurance, outliers can be due to rare events, such as catastrophes, but they can also be due to data entry errors. A robust estimator can limit the influence of a single observation on the parameter estimation.

3.1.1 M-estimators of Location

We assume that the outcome x_i of each observation depends on the "true value" μ and σ, the location and scale parameters, respectively. The location model is given by:

$$x_i = \mu + \sigma u_i, \qquad i = 1, ..., n, \tag{3.1}$$

where $u_1, u_2, ...u_n$ are independent random variables with the same distribution function F and we assume σ is known.

It follows that $x_1, x_2, ..., x_n$ are independent with common distribution function:

$$F_{\mu,\sigma}(x) = F(x; \mu, \sigma) = F\Big(\frac{x-\mu}{\sigma}\Big),$$

that belongs to a family of parametric distributions $\{F_{\underline{\theta}}(x); \underline{\theta} \in \Omega\}$.

Assume that F has a density $f = F'$, given by:

$$f(x; \mu, \sigma) = \frac{1}{\sigma} f\Big(\frac{x-\mu}{\sigma}\Big).$$

Then the joint density of the observations (the *likelihood function*) is:

$$L(x_1, ..., x_n; \mu, \sigma) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f\Big(\frac{x_i-\mu}{\sigma}\Big).$$

Assuming σ known, the *maximum likelihood estimator* (MLE) of μ is the value $\widehat{\mu}$, depending on $x_1, ..., x_n$, that maximizes $L(x_1, ..., x_n; \mu, \sigma)$:

$$\widehat{\mu} = \widehat{\mu}(x_1, ..., x_n) = \arg\max L(x_1, ..., x_n; \mu) \tag{3.2}$$

where arg max stands for "the value maximizing".

The goal will be to find estimators which are: i) "nearly optimal" when F is exactly normal and ii) "nearly optimal" when F is approximately normal, that is, *contaminated normal distribution.*

To formalize this idea, we may imagine that a proportion $(1-\epsilon)$ of the observations is generated by the normal model, while a proportion ϵ is generated by an unknown mechanism. This can be described by supposing:

$$F = (1-\epsilon)G + \epsilon H, \tag{3.3}$$

where $G = N(\mu, \sigma^2)$ and H may be any distribution; for instance, another normal with a larger variance and a possibly different mean. In general, F is called a *mixture* of G and H, and it is called a *normal mixture* when both G and H are normal.

If f is positive everywhere, we can use the logarithm function (which is increasing), to write (3.2) as:

$$\widehat{\mu} = \arg\min \sum_{i=1}^{n} \rho\Big(\frac{x_i - \mu}{\sigma}\Big), \tag{3.4}$$

where $\rho(t) = -\log f(t) + \log f(0)$ is called a ρ-function. If ρ is differentiable, differentiating (3.4) with respect to μ, we have:

$$\sum_{i=1}^{n} \psi\Big(\frac{x_i - \mu}{\sigma}\Big) = 0, \tag{3.5}$$

where $\psi = \rho'$. Therefore, given a function ρ derivable, no decreasing and with $\rho(0) = 0$, an *M-estimator of location* is a solution of (3.4). Note that if f is symmetric, then ρ is even and hence ψ is odd.

For example, if $F = N(0, 1)$, then $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ and up to a constant,

$\rho(x) = x^2/2$ and $\psi(x) = x$. Hence (3.5) becomes:

$$\sum_{i=1}^{n}(x_i - \widehat{\mu}) = 0,$$

which solves for $\widehat{\mu} = \overline{x}$.

For $\rho(x) = |x|$, corresponding to the double exponential distribution, it can be shown that any median of $x_1, ..., x_n$ is a solution of (3.4). In fact, the derivative of $\rho(x)$ exists only for $x \neq 0$ and is given by the sign function:

$$\psi(x) = sgn(x) = \begin{cases} -1 & if \quad x < 0 \\ 0 & if \quad x = 0 \\ 1 & if \quad x > 0 \end{cases},$$

then (3.5) becomes:

$$\sum_{i=1}^{n} sgn(x_i - \widehat{\mu}) = 0,$$

which solves for $\widehat{\mu} = \text{med}(x_1, ..., x_n)$, the sample median.

A very common type of ρ and ψ-functions with important properties is the family of Huber functions, see Figure 3.1:

$$\rho_k(x) = \begin{cases} x^2 & if \quad |x| \leq k \\ 2k|x| - k^2 & if \quad |x| > k \end{cases}, \tag{3.6}$$

with derivative $2\psi_k(x)$, where:

$$\psi_k(x) = \begin{cases} x & if \quad |x| \leq k \\ sgn(x)k & if \quad |x| > k \end{cases}. \tag{3.7}$$

It is seen in Figure 3.1 that ρ_k is quadratic in a central region, but increases only linearly to infinity. The M-estimators corresponding to the limit cases $k \longrightarrow \infty$ and $k \longrightarrow 0$ are the mean and median, and we define $\psi_0(x)$ as $sgn(x)$[1].

[1] The value of k is chosen in order to ensure a given asymptotic variance of the normal distribution.

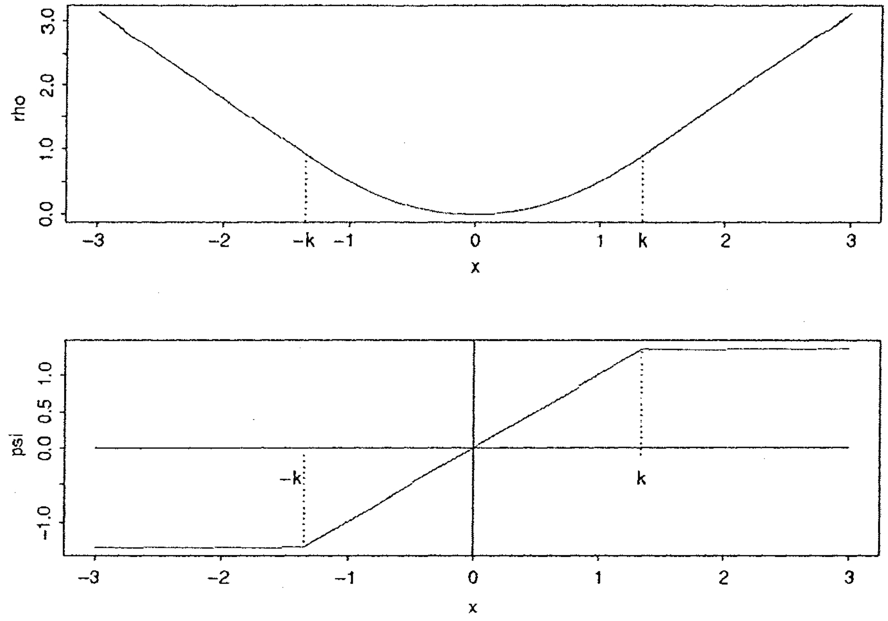


Figure 3.1: Huber ρ– and ψ–functions.

Maronna (2006) showed that when $n \to \infty$, then $\widehat{\mu} \xrightarrow{P} \mu_0$, where $\mu_0 = \mu_0(F)$ is the solution of:

$$\mathbb{E}_F\big[\psi(x-\mu_0)\big] = 0,$$

for a given distribution F and ψ assumed increasing. Therefore, the distribution of $\widehat{\mu}$ is asymptotically:

$$N\Big(\mu_0, \frac{v}{n}\Big), \quad \text{with} \quad v = \frac{\mathbb{E}_F\big[\psi(x-\mu_0)^2\big]}{\Big[\mathbb{E}_F\big[\psi'(x-\mu_0)\big]\Big]^2}. \tag{3.8}$$

In most cases of interest, $\psi(0) = 0$ and $\psi'(0)$ exists, so that ψ is approximately linear at the origin. Let

$$W(x) = \begin{cases} \psi(x)/x & if \quad x \neq 0 \\ \psi'(x) & if \quad x = 0 \end{cases}, \tag{3.9}$$

then (3.5) can be written as:

$$\sum_{i=1}^{n} W(x_i - \widehat{\mu})(x_i - \widehat{\mu}) = 0,$$

or equivalently,

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{3.10}$$

where $w_i = W(x_i - \widehat{\mu})$, which expresses the location M-estimator as a weighted mean.

3.1.2 M-estimators of Scale

To measure the variability of a vector $\mathbf{x} = (x_1, x_2, ..., x_n)$, we use the standard deviation (SD), given by:

$$\mathrm{SD}(\mathbf{x}) = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right], \tag{3.11}$$

where $\overline{x}$ is the sample mean, which is easily influenced by outliers, (see Section 3.1.4) and therefore SD lacks of robustness. One alternative estimator proposed is the *mean absolute deviation* (MD):

$$\mathrm{MD}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|, \tag{3.12}$$

which is still sensitive to outliers, but less that the SD tough (See Tukey, 1977).

A more robust alternative is to subtract the median of the absolute value, which yields the MAD estimate:

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{Med}(|\mathbf{x} - \mathrm{Med}(\mathbf{x})|). \tag{3.13}$$

Now, consider observations x_i satisfying the *multiplicative* model:

$$x_i = \sigma u_i, \tag{3.14}$$

where $u_1, ..., u_n$ are i.i.d. with density f_0 and $\sigma > 0$ is the unknown parameter. The MLE of σ in (3.14) is:

$$\widehat{\sigma} = \arg\max \frac{1}{\sigma^n}\prod_{i=1}^{n} f_0\left(\frac{x_i}{\sigma}\right).$$

Taking logs and differentiating with respect to σ gives:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{x_i}{\widehat{\sigma}}\right) = 1,$$

where $\rho(t) = t\psi(t)$, with $\psi = -f'/f$.

In general, an *M-estimator of scale* is any estimator that satisfies an equation of the form:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\Big(\frac{x_i}{\widehat{\sigma}}\Big) = \delta, \tag{3.15}$$

where ρ is a ρ-function and $0 < \delta < 1$ is a constant needed to have a solution in (3.15).

As in the previous section, the M-scale estimator can be represented as a weighted RMS[2] by defining:

$$W(x) = \begin{cases} \rho(x)/x^2 & if \quad x \neq 0 \\ \rho''(x) & if \quad x = 0 \end{cases}, \tag{3.16}$$

and then (3.15) is equivalent to:

$$\widehat{\sigma}^2 = \frac{1}{n\delta}\sum_{i=1}^{n} W\Big(\frac{x_i}{\widehat{\sigma}}\Big)x_i^2.$$

Therefore, larger x values are assigned smaller weights. In the previous section we defined the M-estimator as a solution of (3.4), when σ is known. However, it does not occur in practice, hence we need to replace it by an estimator $\widehat{\sigma}$, such as those in (3.11), (3.12) or (3.13).

3.1.3 Trimmed Means

Another robust estimation approach of location is to discard a proportion of the largest and smallest values.

[2] The root mean square (RMS), also known as the quadratic mean, of observations $(x_1, ..., x_n)$ is defined by:

$$x_{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Definition 3.1 *Let $x_{(1)}, ..., x_{(n)}$ be the ordered values of $\boldsymbol{x}$. The symmetrically trimmed mean or the δ-trimmed mean is:*

$$\overline{x}_\delta(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} x_{(i)},$$

where $L_n = \lfloor n\delta \rfloor$ and $U_n = n - L_n$.

We can also fix the proportion of cases trimmed and the proportion of cases covered. For the following estimators, we use L_n and U_n as previously defined. Here $\lfloor x \rfloor$ denotes the "greatest integer" less than or equal to x and $\lceil x \rceil$ denotes the "smallest integer" greater than or equal to x.

Definition 3.2 *The Winsorized mean is defined as follows:*

$$W_\delta(L_n, U_n) = \frac{1}{n}\Big[L_n x_{(L_n+1)} + \sum_{i=L_n+1}^{U_n} x_{(i)} + (n - U_n)x_{(U_n)}\Big].$$

Definition 3.3 *A randomly trimmed mean:*

$$R_\delta(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} x_{(i)},$$

where $L_n < U_n$ are integer valued random variables. U_n of the cases are *covered* by the randomly trimmed mean, while $n - U_n + L_n$ of the cases are trimmed.

3.1.4 The Influence Function

Given a random sample $x_1, ..., x_n$ with a common distribution $F_\theta(x) \in \{F_\theta; \theta \in \Omega\}$, we define the empirical cumulative distribution function (c.d.f.), F_n, as follows:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[x_i \leq x]},$$

where $I_{[A]}$ is the indicator function of event A.

By Glivenko Cantelli's theorem, it can be shown (see Devroye, 1997) that:

$$\lim_{n \to \infty} F_n(x) = F(x) \qquad \text{(almost surely)}.$$

Therefore, we consider estimators $\widehat{\theta}$ that are functionals F_n of this empirical c.d.f., $\widehat{\theta}_n = T_n(x_1, ..., x_n) = T(F_n)$, where T is assumed to be continuous, i.e., if $F_n \overset{a.s.}{\to} F$, then $T(F_n) \overset{d}{\to} T(F)$ (see Appendix B). Hence, $\widehat{\theta}_n = T(F_n) \to T(F)$, which is called the *consistency* property.

Now, what happens if we add one more observation with value x to a very large sample? To measure this effect, we give the following definition:

Definition 3.4 *The sensitivity curve (SC) of the estimator T_n for the sample $x_1, ..., x_n$, $n \geq 2$, is:*

$$SC_n(x; x_1, ..., x_{n-1}, T_n) = n\big[T_n(x_1, ..., x_{n-1}, x) - T_{n-1}(x_1, ..., x_{n-1})\big],$$

which is, in general, difficult to compute and it depends on the observed sample values. That is why, we define the asymptotic version of the SC of an estimator, when the sample contains a small fraction ϵ of identical outliers.

Definition 3.5 *The Influence Function (IF) of a functional T at x under a true distribution F is given by:*

$$IF(x; T, F) = \lim_{\epsilon \to 0} \frac{T\big[(1-\epsilon)F + \epsilon I_{[x]}\big] - T(F)}{\epsilon},$$

where $I_[x]$ is the indicator function of x. This function was introduced by Hampel (1971, 1974). Under appropriate regularity conditions, it can be proven that:

$$\int IF(x; T, F) dF(x) = 0, \qquad \text{and by Taylor's expansion:}$$

$$T_n = T(F_n) = T(F) + \frac{1}{n} \sum_{i=1}^{n} IF(x_i; T, F) + o_p\Big(\frac{1}{\sqrt{n}}\Big).$$

Also, $\sqrt{n}\big[T_n - T(F)\big] \to N\big[0, \mathbb{V}(T,F)\big]$ in distribution, where:

$$\mathbb{V}(T,F) = \int IF(x;T,F)^2 dF(x).$$

Hence, the influence function allows us to assess the relative influence of individual observations for the value of an estimate. If it is unbounded, an outlier may cause trouble. For example, if $T_n(x_1, ..., x_n) = \overline{x}_n$, we have:

$$SC_n(x; x_1, ..., x_{n-1}, \overline{x}_n) = x - \overline{x}_{n-1},$$

and

$$IF(x;T,F_\theta) = x - \theta,$$

where $T(F_\theta) = \theta$ is the mean of F_θ. Hence the influence function is unbounded and therefore the sample mean is very influenced by outliers. Now, we define the maximum value of the influence function.

Definition 3.6 *The gross-error sensitivity of T at F_θ is defined by:*

$$\gamma^*(T,F_\theta) = \sup_x |IF(x;T,F_\theta)|,$$

and by re-scaling it, we can compare different estimators:

$$\gamma^{**}(T,F_\theta) = \frac{\sup_x |IF(x;T,F_\theta)|}{\sqrt{\int IF(x;T,F_\theta)^2 dF_\theta(x)}}.$$

The closer $\gamma^{**}(T,F_\theta)$ is to 1, the more robust is the estimator. It is proven (see Maronna, 2006) that the influence function for a location M-estimator of $\mu = T_a(F_\theta)$ given by (3.1), is:

$$IF(x_0;T_a,F_\theta) = \sigma\frac{\psi\big((x_0-\mu)/\sigma\big)}{\mathbb{E}[\psi'(u)]}, \tag{3.17}$$

and for a scale M-estimator of $\sigma = T_b(F_\theta)$ given by (3.14), the influence function is:

$$IF(x_0;T_b,F_\theta) = \sigma\frac{\rho(x_0/\sigma) - \delta}{\mathbb{E}[(x/\sigma)\rho'(x/\sigma)]}. \tag{3.18}$$

3.1.5 The Breakdown Point

The previous section defines the gross-error sensitivity as a measure of robustness of an estimator, when the sample contains a small proportion ϵ of contamination. There is another measure that is used for a bigger (finite or infinite) proportion of contamination. It is called the *Breakdown Point* (BP). Roughly speaking, the BP of an estimator $\widehat{\theta}$ of θ is the largest amount of contamination that the data may contain such that $\widehat{\theta}$ still gives some information about θ, i.e., about the distribution of the "typical" points. In other words the BP is the smallest fraction of data points that we need to replace in order to move the estimator of the contaminated data set arbitrarily far away.

Suppose m arbitrary data points $\mathbf{z} = (z_1, ..., z_m)$ replace m data points from the original data $\mathbf{x} = (x_1, ..., x_n)$, producing a corrupted sample with a proportion $\varepsilon_m = m/n$ of contaminated data. Thus, we observe the difference between the functional on the original sample and the functional on the contaminated sample.

Therefore, we define the maximum bias for an estimator $\widehat{\theta}_n = T(F_n) = T_n(\mathbf{x}) = T_n(x_1, ..., x_n)$ caused by $\varepsilon_m = m/n$ to be:

$$b(\widehat{\theta}_n, \mathbf{x}, m) = \sup_{\mathbf{z}} \big|T_n(\mathbf{z}) - T_n(\mathbf{x})\big|.$$

Definition 3.7 *The breakdown point of $\widehat{\theta}$ in x is defined as:*

$$\epsilon^*(\widehat{\theta}, \boldsymbol{x}) = \frac{m^*}{n},$$

where $m^* = min\big\{m : b(\widehat{\theta}_n, \boldsymbol{x}, m) = \infty\big\}$.

In general, the maximum BP is 1/2 and it is reached by some estimators, for instance, the sample median and M-estimators, based on a bounded ψ.

The finite-sample version of a breakdown point is given by the *replacement finite-sample breakdown point* (FBP) of $\widehat{\theta}_n$ at x. It is the largest proportion $\epsilon^\star(\widehat{\theta}_n, \mathbf{x})$ of data points that can be arbitrarily replaced by outliers such that $\widehat{\theta}_n$ is still bounded and also bounded away from the boundary of Θ, which is nothing else but a range of possible values[3] of θ.

More formally, let χ_m be the set of all data sets $\mathbf{y}$ of size n having $n-m$ elements in common with $\mathbf{x}$, then:

$$\epsilon_n^\star(\widehat{\theta}_n, \mathbf{x}) = \frac{m^\star}{n}, \tag{3.19}$$

where

$$m^\star = \max\left\{m \geq 0 : \widehat{\theta}_n(\mathbf{y}) \text{ is bounded and also bounded away from } \Theta \quad \forall \mathbf{y} \in \chi_m\right\}.$$

For more details, see Donoho and Huber (1983).

3.2 Robust Multivariate Estimators

The estimators of a multivariate location vector $\boldsymbol{\mu}$ and of its scatter matrix $\boldsymbol{\Sigma}$ form the basis of multidimensional data analysis, since they are the input to many classical multivariate techniques. We know that the sample mean and variance are optimal if data comes from multivariate normal distribution, because they correspond to the maximum likelihood estimation (MLE) of the population parameters. However, they are extremely sensitive to the presence of outliers, as reviewed in the univariate case in Section 3.1.4. Therefore it is important to consider robust estimators for multivariate observations.

[3] For example, for a scale or dispersion parameter, we have $\Theta = [0, \infty]$, and the estimator should remain bounded, and also away from 0, in the sense that the distance between $\widehat{\theta}$ and 0 should be larger than some positive value.

Simple robust estimators of a multivariate location parameter can be obtained by applying a robust univariate location estimator to each coordinate, but this lacks of affine equivariance. For dispersion, there exist simple robust estimators of the covariance between two variables (pairwise covariances), which can be used to construct a robust covariance matrix (see Devlin, Gnanadesikan and Kettering, 1981 or Huber, 1981) but apart from not being equivariant, the resulting matrix may not be positive semidefinite. Therefore, it is recommended to estimate them *simultaneously* to get equivariant estimators.

For more details about robustness for multivariate analysis, see Maronna (2006, pp. 175-228).

3.2.1 Multivariate M-estimators

In Section 3.1 we defined M-estimators by generalizing MLE's for univariate location and scale estimators. Recall that a multivariate normal density has the form in (2.1) and also it can be written as follow:

$$f(X,\boldsymbol{\mu},\Sigma)=\frac{1}{\sqrt{|\Sigma|}}h(d(X,\boldsymbol{\mu},\Sigma)), \tag{3.20}$$

where $h(s)=c\exp(-s/2)$, with $c=(2\pi)^{-p/2}$ and $d(X,\boldsymbol{\mu},\Sigma)=(X-\boldsymbol{\mu})'\Sigma^{-1}(X-\boldsymbol{\mu})$.

Any density of this form is called *elliptically symmetric* (or "elliptical" for short) and when $\boldsymbol{\mu}=0$ and $\boldsymbol{\Sigma}=c\mathbf{I}$ it is called *spherically symmetric* (or just "spherical").

Let $\mathbf{x}_1,...\mathbf{x}_n$ be an i.i.d sample from an f of the form in (3.20), where h is assumed to be everywhere positive. Hence, the likelihood function of f is:

$$L(\boldsymbol{\mu},\Sigma)=\frac{1}{|\Sigma|n/2}\prod_{i=1}^{n}h(d(\mathbf{x}_i,\boldsymbol{\mu},\Sigma)),$$

and maximizing $L(\boldsymbol{\mu}, \Sigma)$, yields the following minimization problem:

$$-2 \log L(\boldsymbol{\mu}, \Sigma) = n \log |\widehat{\Sigma}| + \sum_{i=1}^{n} \rho(d_i), \tag{3.21}$$

where $\rho(s) = -2 \log h(s)$ and $d_i = d(\mathbf{x}_i, \boldsymbol{\mu}, \Sigma)$.

Then, we obtain the M-estimators as a solution of (3.21):

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{n} W_1\{d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\Sigma})\}\mathbf{x}_i}{\sum_{i=1}^{n} W_1\{d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\Sigma})\}}, \tag{3.22}$$

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} W_2\{d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\Sigma})\}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})', \tag{3.23}$$

with $W = \rho'$ and where the functions W_1 and W_2 need not to be equal and both depend on the outlyingness measure d_i.

Uniqueness of solutions of (3.22) and (3.23) requires that $d\ W_2(d)$ be a nondecreasing function d, which is the multivariate version of the requirement that the ρ-function of a univariate M-scale be monotone.

It is proved in Huber (1981) that if the $\mathbf{x}_i$ are i.i.d with distribution F, then under general assumptions, when $n \to \infty$, M-estimators converge in probability to the solution $(\boldsymbol{\mu}, \Sigma)$ of:

$$\mathbb{E}[W_1(d)(\mathbf{x}_i - \boldsymbol{\mu})] = \mathbf{0}$$

$$\mathbb{E}[W_2(d)(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'] = \Sigma,$$

where $d = d(\mathbf{x}, \boldsymbol{\mu}, \Sigma)$. Huber also proves that $\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \widehat{\Sigma} - \Sigma)$ tends to a multivariate normal distribution.

From an applied viewpoint, M-estimators can be considered as a simple modification of classical estimators. They give full weight to observations assumed to come from the bulk of the data, but they reduce the weight (influence) of observations from the tails of the contaminating distribution. In many applications, the functions W_1 and W_2 are chosen according to Huber's proposal (see Croux and Haesbroeck, 2000):

$$W_1(y) = \frac{\psi_H(y, q^{1/2})}{y} \quad \text{and} \quad W_2(y) = \frac{\psi_H(y, q)}{\beta y},$$

where $\psi_H(y, k) = \max\{-k, \min(y, k)\}$ is Huber's psi function, β is a constant making the covariance matrix estimator Fisher–consistent at normal models[4] and $q = \chi^2_{p,0.9}$.

3.2.2 The Breakdown Point for Multivariate Robust Estimators

To define the breakdown point (BP) we will consider its asymptotic version. In this case we assume that:

$$W_1(d)\sqrt{d} \quad \text{and} \quad W_2(d)d \quad \text{are bounded for} \quad d \geq 0. \tag{3.24}$$

For multivariate M-estimators, the asymptotic BP depends on the knowledge of $\boldsymbol{\mu}$ and Σ. If Σ is known, Maronna (2006) showed that the BP of $\boldsymbol{\mu}$ is 1/2. On the other hand, if $\boldsymbol{\mu}$ is known, the asymptotic BP of the M-estimator of Σ with W_2 satisfying (3.24) is:

$$\varepsilon^* = \min\left(\frac{1}{K}, 1 - \frac{p}{K}\right), \tag{3.25}$$

where its maximum value is $1/(p+1)$, attained when $K = p + 1$, and hence:

$$\varepsilon^* \leq \frac{1}{p+1}. \tag{3.26}$$

Davies (1987) showed that the maximum FBP of any equivariant estimator for a sample in *general position*[5] is m^*_{max}/n, where.

$$m^*_{max} = \left[\frac{n-p}{2}\right]. \tag{3.27}$$

It is therefore natural to search for estimators whose BP is nearer to this maximum BP than that of monotone M-estimators.

[4] For this purpose, we use the factor defined in (3.38)

[5] To be in general position means that no hyperplane contains more than p points

3.2.3 Using Pairwise Robust Covariances

A robust pairwise covariance proposed initially by Gnanadesikan and Kettenring (1972) and studied by Devlin et al (1981) is based on the identity:

$$\mathrm{Cov}(X,Y) = \frac{1}{4}\Big(\mathrm{SD}(X+Y)^2 - \mathrm{SD}(X-Y)^2\Big). \tag{3.28}$$

They proposed to define a robust correlation by replacing the standard deviation by a robust dispersion σ (they chose a trimmed standard deviation):

$$\mathrm{RCorr}(X,Y) = \frac{1}{4}\left(\sigma\left(\frac{X}{\sigma(X)}+\frac{Y}{\sigma(Y)}\right)^2 - \sigma\left(\frac{X}{\sigma(X)}-\frac{Y}{\sigma(Y)}\right)^2\right), \tag{3.29}$$

and a robust covariance defined by:

$$\mathrm{RCov}(X,Y) = \sigma(X)\sigma(Y)\mathrm{RCorr}(X,Y). \tag{3.30}$$

The above pairwise robust covariances can be used to define a "robust correlation (covariance) matrix" of a random vector $X = (\mathbf{x}_1, ..., \mathbf{x}_p)'$, which is symmetric but not necessarily positive semidefinite and is not affine equivariant. Maronna and Zamar (2002) show that this problem can be overcome by using a robust σ and a set of "principal directions".

3.2.4 Estimators Based on a Robust Scale

For the purpose of this work, it will be very useful to look for multivariate estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ that minimize some measure of size of $d(\mathbf{x}, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$. For a data set X call $\mathbf{d}(X, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ the vector with elements $d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$, $i = 1, ..., n$ and let $\widehat{\sigma}$ be a robust scale estimator. Then we can define the estimators $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ by minimizing:

$$\widehat{\sigma}\Big(\mathbf{d}(X, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})\Big), \tag{3.31}$$

with $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^p$, $\widehat{\boldsymbol{\Sigma}} \in \mathcal{S}_p$ and $|\widehat{\boldsymbol{\Sigma}}| = 1$, where $\mathcal{S}_p$ is the set of symmetric positive definite $p \times p$ matrices. The previous formulation is equivalent to minimizing $|\widehat{\boldsymbol{\Sigma}}|$ subject to a bound $\widehat{\sigma}$.

The Minimum Volume Ellipsoid Estimator

The simplest case is to define $\widehat{\sigma}$ as the sample median and the resulting location and dispersion matrix estimator is called the *minimum volume ellipsoid* (MVE) estimator. This estimator has a consistency rate of only $n^{-1/3}$ and hence is very inefficient[6].

S-estimators

Other estimators with better efficiency are the *S-estimators* (see Davies, 1987), defined by (3.31) and taking for $\widehat{\sigma}$ an M-scale estimator satisfying:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\Big(\frac{d_i}{\widehat{\sigma}}\Big) = \delta, \tag{3.32}$$

where ρ is a smooth bounded ρ-function. If ρ is differentiable, it can be shown that the solution of (3.31) must satisfy equations (3.22) and (3.23), that is:

$$\sum_{i=1}^{n} W_1\Big(\frac{d_i}{\widehat{\sigma}}\Big)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}) = \mathbf{0} \tag{3.33}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_2\Big(\frac{d_i}{\widehat{\sigma}}\Big)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})' = c\widehat{\boldsymbol{\Sigma}}, \tag{3.34}$$

with $W = \rho'$, $\widehat{\sigma} = \widehat{\sigma}(d_1, ..., d_n)$ and c is a scalar such that $|\widehat{\boldsymbol{\Sigma}}| = 1$.

Davies (1987) proved that if ρ is differentiable, then for S-estimators, the distribution of $\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}, \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})$ tends to a multivariate normal distribution.

[6] All estimators considered in this section are considered consistent in the following sense: if $\mathbf{x}_i \sim N_p(\mu, \Sigma)$ then $\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}$ and $\widehat{\boldsymbol{\Sigma}} = c\boldsymbol{\Sigma}$, where c is a constant. Also, we assume that all estimators defined in this section are asymptotically normal:

$$\sqrt{n}(\widehat{\boldsymbol{\mu}}_n - \widehat{\boldsymbol{\mu}}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{V}_\mu) \quad , \quad \sqrt{n}\text{vec}(\widehat{\boldsymbol{\Sigma}}_n - \widehat{\boldsymbol{\Sigma}}) \xrightarrow{d} N_q(\mathbf{0}, \mathbf{V}_\Sigma),$$

where $q = p(p+1)/2$ and for a symmetric matrix $\boldsymbol{\Sigma}$, vec($\boldsymbol{\Sigma}$) is the vector containing the q elements of the upper triangle of $\boldsymbol{\Sigma}$. The matrices $\mathbf{V}_\mu$ and $\mathbf{V}_\Sigma$ are the asymptotic covariance matrices of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$. Hence we say that $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ have a consistency rate of $n^{-1/2}$

We define the bisquare multivariate S-estimator as the one with scale given by (3.32) with:

$$\rho(t) = \min\{1, 1 - (1-t)^3\}, \tag{3.35}$$

and with a weight function

$$W(t) = 3(1-t^2)I_{(t\leq 1)}.$$

In the univariate case, the bisquare scale estimator, call it $\widehat{\eta}$, based on centered data x_i with location μ, is the solution of:

$$\frac{1}{n}\sum_{i=1}^{n} \rho_{bisq}\Big(\frac{x_i - \widehat{\mu}}{\widehat{\eta}}\Big) = \delta, \tag{3.36}$$

where $\rho_{bisq}(t) = \min\left\{1, 1-(1-t^2)^3\right\}$. Note that $\rho_{bisq}(t) = \rho(t^2)$ for ρ defined in (3.35), then (3.36) is equivalent to:

$$\frac{1}{n}\sum_{i=1}^{n} \rho\Big(\frac{(x_i - \widehat{\mu})^2}{\widehat{\sigma}}\Big) = \delta,$$

with $\widehat{\sigma} = \widehat{\eta}^2$. Now $d(\mathbf{x}, \boldsymbol{\mu}, \Sigma)$ is the normalized *squared* distance between $\mathbf{x}$ and $\boldsymbol{\mu}$, which explains the use of ρ.

The Minimum Covariance Determinant Estimator

When the number of variables p is smaller than the data size n, the minimum covariance determinant estimator (MCD) can be used (see Rousseeuw, 1984). This location and covariance estimator is very popular because of its high resistance towards outliers and because a fast algorithm for its computation has been developed by Rousseeuw and Van Driessen (1999).

We consider a vector $X = [\mathbf{x}_1, ..., \mathbf{x}_n]$, then the MCD method searches for the subset of h observations whose covariance matrix $\widehat{\Sigma}$ has the lowest determinant[7].

[7] The number h determines the robustness of the estimator and should be at least $\lfloor (n+p+1)/2 \rfloor$, where $\lfloor y \rfloor$ denotes the "greatest integer" less than or equal to y.

Hence, we obtain the estimators $\widehat{\boldsymbol{\mu}}_{MCD}$ and $\widehat{\Sigma}_{MCD}$ by solving the following objective function:

$$\min|k\widehat{\Sigma}^*|, \tag{3.37}$$

where

$$\widehat{\Sigma}^* = \frac{1}{h}\sum_{i=1}^{h}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}^*)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}^*)'$$

and

$$\widehat{\boldsymbol{\mu}}^* = \frac{1}{h}\sum_{i=1}^{h}\mathbf{x}_i.$$

The value for $h \in \Big[\lfloor(n+p+1)/2\rfloor, n\Big]$ and if we are certain that the dataset contains less than 25% outliers, then we can obtain statistical efficiency by using $h = \lfloor 0.75n \rfloor$[8]. The default value for h in many statistical packages is $\lfloor(n+p+1)/2\rfloor$. The MCD's breakdown is then $(n-h+1)/n$, which means that we need at least $n-h+1$ outliers to make the estimators worthless. Note that, when $h=n$ we have the normal MLE estimators.

The factor k is used to obtain consistency when data come from a multivariate normal distribution and it is defined by:

$$k = \frac{\text{med}\Big(d(\mathbf{x}_1, \widehat{\boldsymbol{\mu}}_{MCD}, \widehat{\Sigma}_{MCD}), ..., d(\mathbf{x}_n, \widehat{\boldsymbol{\mu}}_{MCD}, \widehat{\Sigma}_{MCD})\Big)}{\chi^2_{p,0.5}}, \tag{3.38}$$

where $\chi^2_{p,\alpha}$ denotes the α-quantile of the χ^2_p distribution.

The One–step Re–weighted Estimator

A one-step reweighted estimator is obtained by:

$$\widehat{\boldsymbol{\mu}}_1 = \frac{\sum_{i=1}^n w_i\mathbf{x}_i}{\sum_{i=1}^n w_i},$$

$$\widehat{\Sigma}_1 = \frac{\sum_{i=1}^n w_i(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_1)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_1)'}{\sum_{i=1}^n w_i - 1},$$

[8] Note that in this case $\widehat{\Sigma}^*$ and $\widehat{\boldsymbol{\mu}}^*$ are *trimmed* estimators, but not necessarily symmetric, as the one defined in Section 3.1.3.

where:

$$w_i = \begin{cases} 1 & \text{if} \quad d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}_{MCD}, \widehat{\Sigma}_{MCD}) \leq \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise} \end{cases} .$$

Therefore, this estimator is also called the reweighted MCD estimator and it is equal to the classical mean and covariance matrix of data points with weight one.

The Stahel-Donoho Estimator

Section 2.1 describes a simple approach for detection of outliers using the Mahalanobis distance: for each observation compute an outlyingness measure and identify those points having a large value of this measure. The idea in the multivariate case is that a multivariate outlier should be an outlier in some *univariate* direction.

More precisely, given a direction $\mathbf{a} \in \mathbb{R}^p$ with $||\mathbf{a}|| = 1$, denote by $\mathbf{a}'X = (\mathbf{a}'\mathbf{x}_1, ..., \mathbf{a}'\mathbf{x}_n)$ the projection of the data X along $\mathbf{a}$. Let $\widehat{\mu}$ and $\widehat{\sigma}$ be robust univariate location and dispersion statistics. The outlyingness with respect to X of a point $\mathbf{x} \in \mathbb{R}^p$ along $\mathbf{a}$ is defined by:

$$t(\mathbf{x}, \mathbf{a}) = \frac{\mathbf{x}'\mathbf{a} - \widehat{\mu}(\mathbf{a}'X)}{\widehat{\sigma}(\mathbf{a}'X)}.$$

The outlyingness of $\mathbf{x}$ is then defined by:

$$t(\mathbf{x}) = \max_{\mathbf{a}} t(\mathbf{x}, \mathbf{a}) \tag{3.39}$$

The Stahel-Donoho estimator, proposed by Stahel (1981) and Donoho (1982), is a weighted mean and covariance matrix where the weights of $\mathbf{x}_i$ are nonincreasing functions of $t(\mathbf{x}_i)$. More precisely, let W_1 and W_2 be two weight functions and define:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{\sum_{i=1}^n w_{i1}} \sum_{i=1}^n w_{i1}\mathbf{x}_i, \tag{3.40}$$

$$\widehat{\Sigma} = \frac{1}{\sum_{i=1}^n w_{i2}} \sum_{i=1}^n w_{i2}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})', \tag{3.41}$$

with

$$w_{ij} = W_j(t(\mathbf{x}_i)), \qquad j = 1, 2,$$

where the weight functions must satisfy that $tW_1(t)$ and $t^2W_2(t)$ are bounded for $t \geq 0$, in order that no term dominates.

Maronna and Yohai (1995), showed that under the above condition, the asymptotic BP is 1/2 and for the FBP, Tyler (1990) and Gather and Hilker (1997) showed that the estimator attains the maximum BP given by (3.27) if $\widehat{\boldsymbol{\mu}}$ is the sample median and the scale is:

$$\widehat{\sigma}(z) = \frac{1}{2}(\widetilde{z}_k + \widetilde{z}_{k+1}),$$

where $\widetilde{z}_i$ denotes the ordered values of $|\widetilde{z}_i - \mathrm{Med}(\mathbf{z})|$ and $k = [(n+p)/2]$. The right choice of the weight functions is very important for combining robustness and efficiency.

For more details about this estimator and its influence function, see Gervini (2002) and Zuo, Cui, and He (2004).

Chapter 4

Robust Multivariate Statistical Techniques

4.1 Introduction

Section 2.2.1 describes Principal Component Analysis (PCA) as a classical method to explain the covariance structure using a small number of components. These components are linear combinations of the original variables which allow us to analyze high-dimensional data. PCA is then one of the first steps in the data analysis, followed by discriminant analysis, cluster analysis or other multivariate techniques. Besides, this method can be used for detection of outliers, as we explained in Section 2.2.3.

In this classical approach, the first component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the data points projected on it. Continuing with this procedure, we create all the principal components, which correspond to the eigenvectors of the empirical covariance matrix.

However, as explained in Section 3.1.2, both the classical variance, which is supposed to be maximized, and the classical mean are very sensitive to anomalous observations. As a consequence, the first components are attracted by these outliers and might not catch the variation of regular observations. Therefore, a robust method to reduce the influence of those outlying points is needed.

4.2 Robust Principal Component Analysis

In this section we describe a procedure to robustify PCA to reduce the influence of atypical observations in the obtention of the principal components. For this purpose, a robust estimator $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ of multivariate location and scatter is used to reduce the influence of outliers, instead of the classical sample mean and covariance matrix derived from the Normal distribution.

Consider a sample of p-dimensional observations $\mathbf{x}_1, ..., \mathbf{x}_n$ and denote by $d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}}) = (\mathbf{x}_i - \widehat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$ the statistical distance between $\mathbf{x}_i$ and the $p \times 1$ vector of sample means $\widehat{\boldsymbol{\mu}}$, measured in the metric induced by the positive definite matrix $\widehat{\boldsymbol{\Sigma}}$, the sample covariance or scatter matrix.

An obvious way of modifying the classical PCA is to replace $\widehat{\boldsymbol{\Sigma}}$ by a robust estimator. The method proposed in this thesis is a combination of different robust estimators applied in the method proposed by Maronna (2006) in an effort to reduce the influence of outliers in the estimation of eigenvalues and eigenvectors.

Let $\mathbf{X} = [x_{ij}]$ be an $n \times p$ matrix with rows vectors $\mathbf{x}_i$, $i = 1, ..., n$ and therefore, columns vectors will be represented as $\mathbf{x}^j$, $j = 1, ..., p$. Let $\widehat{\mu}(.)$ and $\widehat{\sigma}(.)$ be univariate M-estimators of location and scale respectively, with ρ-function given by (3.6).

Thus, we obtain a robust dispersion matrix estimator $\widehat{\boldsymbol{\Sigma}}(\mathbf{X})$ and robust location vector estimator $\widehat{\boldsymbol{\mu}}(\mathbf{X})$ by following the next computational steps:

1. Compute a normalized data matrix $\mathbf{Y}$ with columns $\mathbf{y}^j = \frac{\mathbf{x}^j}{\widehat{\sigma}(\mathbf{x}^j)}$, and hence with rows $\mathbf{y}_i = \mathbf{D}^{-1}\mathbf{x}_i$ for $i = 1, ..., n$, where $\mathbf{D} = \text{diag}(\widehat{\sigma}(\mathbf{x}^1), ..., \widehat{\sigma}(\mathbf{x}^p))$, which makes the estimator scale equivariant.

2. Compute a robust correlation matrix $\mathbf{U} = [U_{jk}]$ of $\mathbf{X}$ as the covariance matrix of $\mathbf{Y}$ by applying (3.29) to the columns of $\mathbf{Y}$. That is[1]:

$$U_{jj} = 1, U_{jk} = \frac{1}{4}\left[\widehat{\sigma}(\mathbf{y}^j + \mathbf{y}^k)^2 - \widehat{\sigma}(\mathbf{y}^j - \mathbf{y}^k)^2\right], (j \neq k).$$

3. Compute the eigenvalues λ_j and eigenvector $\mathbf{e}_j$ of $\mathbf{U}$ $(j = 1, ..., p)$ and let $\mathbf{E}$ be the matrix whose columns are the $\mathbf{e}_j$'s. It follows that $\mathbf{U} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ...\lambda_p)$. In this step we use a classical PCA, as described in Section 2.2.

4. Compute the matrix $\mathbf{Z}$ with

$$\mathbf{z}_i = \mathbf{E}'\mathbf{y}_i = \mathbf{E}'\mathbf{D}^{-1}\mathbf{x}_i \quad (\text{for } i = 1, ..., n)$$

so that $(\mathbf{z}^1, ..., \mathbf{z}^p)$ are the principal components of $\mathbf{Y}$.

5. Compute $\widehat{\sigma}(\mathbf{z}^j)$ and $\widehat{\mu}(\mathbf{z}^j)$ for $j = 1, ..., p$ and set

$$\Gamma = \text{diag}\left(\widehat{\sigma}(\mathbf{z}^1)^2, ..., \widehat{\sigma}(\mathbf{z}^p)^2\right)$$

and

$$\boldsymbol{\nu} = \left(\widehat{\mu}(\mathbf{z}^1), ..., \widehat{\mu}(\mathbf{z}^p)\right)'.$$

6. Transform back to $\mathbf{X}$ with

$$\mathbf{x}_i = \mathbf{A}\mathbf{z}_i, \quad \text{where } \mathbf{A} = \mathbf{DE}. \tag{4.1}$$

[1] Recall that this matrix is symmetric but not necessarily positive semidefinite and is not affine equivariant.

7. Set the robust location and covariance estimator of the original matrix $\mathbf{X}$ as

$$\widehat{\boldsymbol{\mu}}(\mathbf{X}) = \mathbf{A}\boldsymbol{\nu}, \quad \text{and } \widehat{\Sigma}(\mathbf{X}) = \mathbf{A}\Gamma\mathbf{A}'. \tag{4.2}$$

   Using this definition of $\widehat{\boldsymbol{\mu}}$ and $\widehat{\Sigma}$ allows us to replace the λ_i's, which may be negative, by the robust variances $\widehat{\sigma}(\mathbf{z}^j)^2$ of the corresponding directions. Maronna and Zamar (2002), show that this simple modification yields a positive definite matrix and approximately equivariant.

   Until this step, the process works well when the data have low correlations. Therefore, in order to have principal components $\mathbf{z}_j$'s less correlated, we iterate the process from Step 1–7 and this is possible by defining:

$$\widehat{\boldsymbol{\mu}}_{(k+1)}(\mathbf{X}) = \mathbf{A}\widehat{\boldsymbol{\mu}}(\mathbf{Z}_{(k)}), \quad \text{and } \widehat{\Sigma}_{(k+1)}(\mathbf{X}) = \mathbf{A}\widehat{\Sigma}(\mathbf{Z}_{(k)})\mathbf{A}'. \tag{4.3}$$

   The resulting estimator it is called the Orthogonalized Gnanadesikan–Kettenring (OGK).

8. In order to increase the estimate's efficiency and to make it more equivariant, we use the robust estimators obtained from iteration in previous step as a starting points for the iteration of either the M– or S–estimator of location and covariance matrix, given in Section 3.2.1 and 3.2.4, respectively, which implies the use of weight functions of the distances $d(\mathbf{x}_i, \widehat{\boldsymbol{\mu}}, \widehat{\Sigma})$

9. Finally, we use this final robust estimators as a parameters in the PCA to obtain the *Robust Principal Components* of $\mathbf{X}$.

Thus, to obtain a robust PCA, we use an estimator $(\widehat{\boldsymbol{\mu}}, \widehat{\Sigma})$ which is: scale equivariant (given by Step 2), with principal components approximately uncorrelated (given by iteration from Steps 1–7), high efficient and more equivariant (given by Step 8).

4.2.1 The Influence Function of Eigenvectors and Eigenvalues

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be an independently and identically distributed sample drawn from a p-variate distribution $F \in \mathfrak{F}$, which will be assumed to be the normal distribution $N(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\Sigma \in \mathcal{S}_p{}^2$. Also, suppose that Σ has distinct eigenvalues $\lambda_1 > ... > \lambda_p$, with corresponding eigenvectors $\mathbf{p}_1, ..., \mathbf{p}_p$.

An influence function, as defined in Section 3.1.4, is essentially the first derivative of the functional version of an estimator. Let T a statistical functional corresponding to an estimator $\widehat{\Sigma}_n$ of Σ that sends an arbitrary distribution $G \in \mathfrak{F}$ to $T(G)$, whenever $T(F_n) = \widehat{\Sigma}_n$, for every empirical distribution function F_n associated with observations $\mathbf{x}_1, ..., \mathbf{x}_n$. If we assume that X is distributed according to G, then the notation $T(X)$ instead of $T(G)$ will be used.

The functional representation of the eigenvectors and eigenvalues from $\widehat{\Sigma}_n$ are denoted by $\mathbf{p}_{T,j}$ and $\lambda_{T,j}$, for $j = 1, ..., p$. At the empirical distribution function, $\mathbf{p}_{T,j}(F_n) = \mathbf{p}_{\widehat{\Sigma}_n,j}$ and $\lambda_{T,j}(F_n) = \lambda_{\widehat{\Sigma}_n,j}$. Also, we assume T to be Fisher consistent for Σ at F, that is $T(F) = \Sigma$, and affine equivariant, meaning that $T(AX + b) = AT(X)A'$ for any $b \in \mathbb{R}^p$ and any $p \times p$ nonsingular matrix A. This implies immediately that Fisher consistency also holds for the eigenvector and eigenvalue functionals, $\mathbf{p}_{T,j}(F) = \mathbf{p}_j$ and $\lambda_{T,j}(F) = \lambda_j$. Moreover, the functionals $\mathbf{p}_{T,j}$ and $\lambda_{T,j}$ are orthogonal equivariant in the sense that:

$$\mathbf{p}_{T,j}(\Gamma X) = \Gamma \mathbf{p}_{T,j}(X) \qquad \text{and} \qquad \lambda_{T,j}(\Gamma X) = \lambda_{T,j}(X),$$

for $j = 1, ..., p$ and for any $p \times p$ orthogonal matrix Γ. All these results are shown in Croux and Haesbroeck (2000).

[2] The set of symmetric positive definite $p \times p$.

By definition, the influence functions of $\mathbf{p}_{T,j}$ and $\lambda_{T,j}$ are given by:

$$IF(\mathbf{x};\mathbf{p}_{T,j},F)=\lim_{\epsilon\to 0}\frac{\mathbf{p}_{T,j}\big[(1-\epsilon)F+\epsilon\delta_{\mathbf{x}}\big]-\mathbf{p}_{T,j}(F)}{\epsilon}, \tag{4.4}$$

$$IF(\mathbf{x};\lambda_{T,j},F)=\lim_{\epsilon\to 0}\frac{\lambda_{T,j}\big[(1-\epsilon)F+\epsilon\delta_{\mathbf{x}}\big]-\lambda_{T,j}(F)}{\epsilon}, \tag{4.5}$$

for $j=1,...,p$ and where $\delta_{\mathbf{x}}$ denotes the point mass 1 at $\mathbf{x}$. For more details on influence functions and statistical functionals, see Hampel et al. (1986).

The following lemma characterizes the general form of the influence function of a covariance matrix and will be used to derive (4.4) and (4.5):

Lemma 4.1 *For any affine equivariant covariance matrix functional T possessing an influence function, there exist two functions α_T, $\beta_T : [0,\infty)\to\mathbb{R}$, such that:*

$$IF(\boldsymbol{x};T,F)=\alpha_T\{d(\boldsymbol{x})\}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})'-\beta_T\{d(\boldsymbol{x})\}\Sigma, \tag{4.6}$$

with $d(\boldsymbol{x})=d(\boldsymbol{x},\boldsymbol{\mu},\Sigma)$ and $F=N_p(\boldsymbol{\mu},\Sigma)$

That influence function has been derived by Huber (1981, pp. 226) for M-estimators and by Lopuha (1989, 1999) for S and for Reweighted estimators. The graph for functions α_T for each estimator is extracted from Croux and Haesbroeck (2000) and showed in Figure 4.1. The corresponding functions α_M, α_S and α_R are nonincreasing, meaning that their contribution to the influence function decreases as the distance between $\mathbf{x}$ and $\boldsymbol{\mu}$ in the metric imposed by $\boldsymbol{\Sigma}$ increases. The function α_{cov}, corresponds to the classic estimator and it is constant, implying that outliers are not given less weight.

Theorem 4.1 *Let F be a multivariate normal distribution with parameters $\boldsymbol{\mu}$ and Σ. Define the scores of X as $Z_k=\boldsymbol{p}_k'(X-\boldsymbol{\mu})$ for $k=1,...,p$ and let $d(\boldsymbol{x})=(\boldsymbol{x}_i-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$. The influence functions of the eigenvectors and eigenvalues of T at F are given by:*

$$IF(\boldsymbol{x};\lambda_{T,j},F)=\alpha_T\{d(\boldsymbol{x})\}Z_j^2-\beta_T\{d(\boldsymbol{x})\}\lambda_j,$$

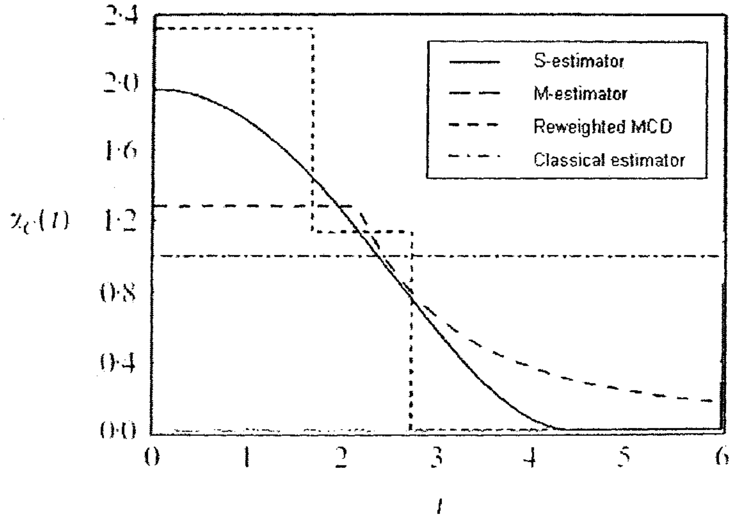


Figure 4.1: Examples of the function α_T for different type of estimators.

$$IF(\boldsymbol{x}; \boldsymbol{p}_{T,j}, F) = \alpha_T\{d(\boldsymbol{x})\} \sum_{k=1,k\neq j}^{p} \frac{Z_k Z_j}{\lambda_j - \lambda_k} \boldsymbol{p}_k,$$

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

It follows now from Critchley (1985) that:

$$IF(\mathrm{x}; \lambda_{T,j}, F) = \alpha_T\{d(\mathrm{x})\} IF(\mathrm{x}; \lambda_{cov,j}, F).$$

Therefore, it is confirmed that the function α_T needs to be interpreted as a down-weighting function. A decreasing α_T function implies a bounded influence function for eigenvectors.

4.3 Robust Clustering Method

Different techniques for identifying clusters of well–separated uncontaminated groups of data have been available for many years. However, this process becomes more difficult when data include outliers since these points can skew the shape estimates of the clusters or distort optimization criterion which mask separation between clusters.

Robust clustering methods often give an accurate representation of data. Still, they do not usually identify particular outlying points which may be of interest or importance. Therefore it is important to have a complementary outlier identification method. For instance, by calculating the Mahalanobis distance for each point to the center of the data, we can then create outlyingness measures, as we defined in previous chapters, to identify possible multivariate outliers, that is, points with a distance larger than some predetermined value. In the clustering context, outlier identification methods can be used individually on each cluster.

4.3.1 Partitioning around medoids

This partitioning around medoids algorithm (PAM) is based on the search for k representative objects, called medoids, among the objects of a data set (Kaufman and Rousseeuw, 1987). These medoids are computed such that the total dissimilarity of all objects to their nearest medoid is minimal, i.e., the goal is to find a subset $\{m_1, ..., m_k\} \subset \{1, ..., n\}$ which minimizes the objective function:

$$\sum_{i=1}^{n} \min_{t=1,...,k} d(i, m_t). \tag{4.7}$$

Each object is then assigned to the cluster corresponding to the nearest medoid. That is, object i is put into cluster R_i when medoid m_{R_i} is nearer to i than any other medoid m_w, i.e.,

$$d(i, m_{R_i}) \leq d(i, m_w) \quad \text{for all } w = 1, ...k.$$

The algorithm of PAM proceeds in two steps:

1. Construct initial medoids:

   - m_1 is the object with the smallest $\sum_{i=1}^{n} d(i, m_1)$
   - m_2,...,m_k decrease the objective (4.7) as much as possible.

2. Repeat until convergence and consider all pairs of objects (i, j) with $i \in \{m_1, ..., m_k\}$ and $j \notin \{m_1, ..., m_k\}$ and make $i \leftrightarrow j$ swap (if any) which decreases the objective most. That is, if the objective can be reduced by interchanging (swapping)

a selected object with an unselected object, then the swap is carried out and this is continued until the objective function can no longer be decreased.

Since the objective function (4.7) only depends on dissimilarities between objects, PAM only needs a dissimilarity matrix. This method can be compared to the k-means method described before. In that method the center of each cluster is defined as the mean of all objects of the cluster and its goal is to minimize a sum of squared euclidean distances, implicitly assuming that each cluster has a spherical normal distribution. The PAM method is, as we will explain further, more robust because it minimizes a sum of unsquared dissimilarities and moreover, PAM does not need an initial guess for cluster centers.

Chapter 5

Application to a Car Classification Database

5.1 Database Description

We use the low-dimensional car data ‘cu.dimensions’ available in S-PLUS and extracted from Consumer Reports (1990), where we have $n = 111$ observations (cars brands) with $p = 11$ variables related to different characteristics, such as:

Variable	Description	Variable	Description
X_1	Length	X_7	Frt.Leg.Room
X_2	Wheel.base	X_8	Rear.Seating
X_3	Width	X_9	Frt.Shld
X_4	Height	X_{10}	RearShld
X_5	Front.Hd	X_{11}	Luggage
X_6	Rear.Hd		

Table 5.1: Variables in Car Database.

5.2 Use of the Robust PCA Method

We obtain, as a preliminary analysis, correlations $\rho(X_i, X_j)$ between some variables, for example, $\rho(X_1, X_2) = 0.83$, $\rho(X_1, X_9) = 0.79$ and $\rho(X_2, X_3) = 0.80$, which indicate a high association among the variables. Therefore a method to reduce dimensions is required, such as Principal Component Analysis. We also check for outliers by computing an outlyingness measure for multivariate observations. Figure 5.1 compares (a) the classical Mahalanobis distance, based on normal estimators versus (b) a robust Mahalanobis distance based on robust estimators. This robust distance is computed using Stahel-Donoho estimators for location and covariance and it is clear that there are atypical observations.

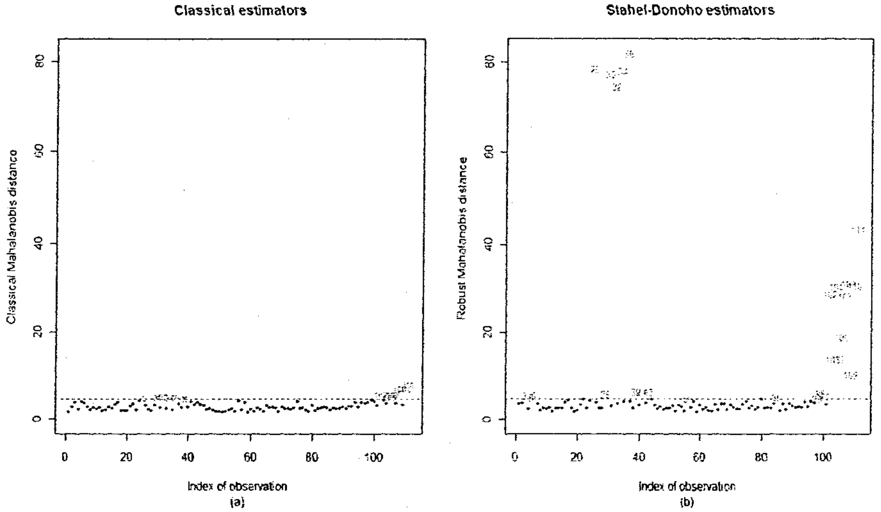
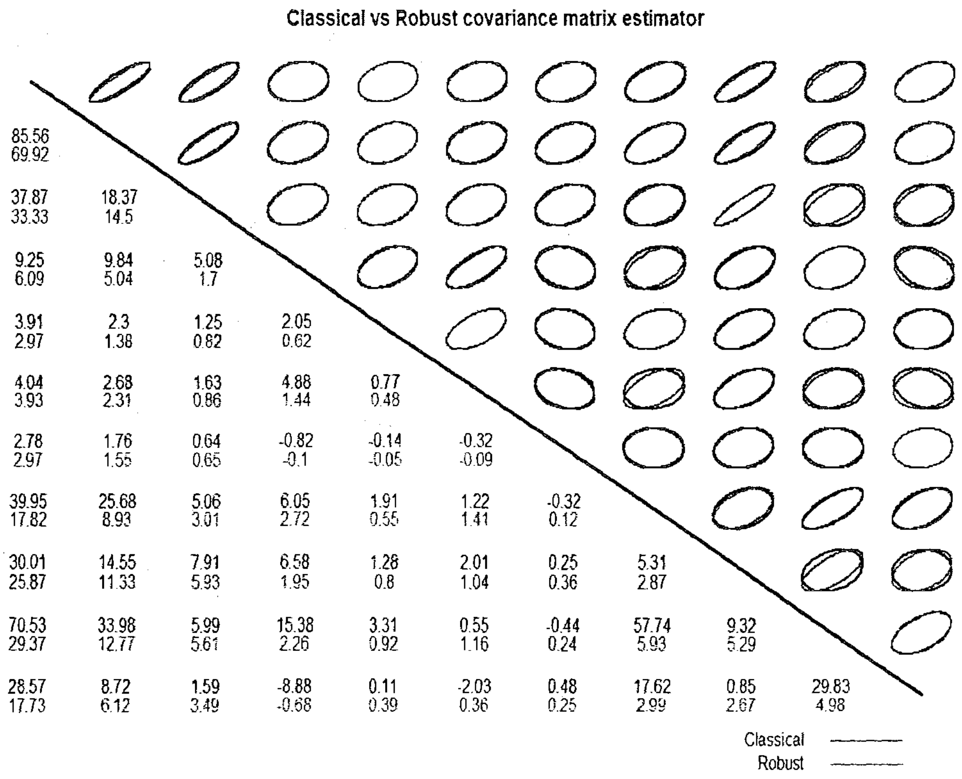


Figure 5.1: (a) Classical Mahalanobis Distance (b) Robust Mahalanobis Distance.

According to Figure 5.1 several outliers are detected by both distances, but there is a significant difference when those are detected using robust estimators (24.32%) from those detected by the classical distance (10.81%) coming from normal estimators. Thus, a method to robustify both location and covariance estimators is needed and the methodology described in Section 4.2 is applied. Under these estimators, a new Mahalanobis distance is computed and approximately 19% of observations are detected as outliers. In Figure 5.2 we compare both, classical and robust covariance estimators. It is clear that there are differences between the $\text{Cov}(X_i, X_j)$, $i \neq j$ and ellipses generated using the corresponding location and covariance estimator in certain variables.


Classical vs Robust covariance matrix estimator

85.56 69.92									
37.87 33.33	18.37 14.5								
9.25 6.09	9.84 5.04	5.08 1.7							
3.91 2.97	2.3 1.38	1.25 0.82	2.05 0.62						
4.04 3.93	2.68 2.31	1.63 0.86	4.88 1.44	0.77 0.48					
2.78 2.97	1.76 1.55	0.64 0.65	-0.82 -0.1	-0.14 -0.05	-0.32 -0.09				
39.95 17.82	25.68 8.93	5.06 3.01	6.05 2.72	1.91 0.55	1.22 1.41	-0.32 0.12			
30.01 25.87	14.55 11.33	7.91 5.93	6.58 1.95	1.28 0.8	2.01 1.04	0.25 0.36	5.31 2.87		
70.53 29.37	33.98 12.77	5.99 5.61	15.38 2.26	3.31 0.92	0.55 1.16	-0.44 0.24	57.74 5.93	9.32 5.29	
28.57 17.73	8.72 6.12	1.59 3.49	-8.88 -0.68	0.11 0.39	-2.03 0.36	0.48 0.25	17.62 2.99	0.85 2.67	29.83 4.98

Figure 5.2: Classical and Robust Covariance Matrix.

In order to know which dimensions are more influenced by outliers, a univariate outlyingness based on the standardization $t = (\mathbf{x}_i - \bar{\mathbf{x}})/\mathbf{s}$ of each observation is computed, where $\bar{\mathbf{x}}$ and $\mathbf{s}$ are the sample vector of means and standard deviations, respectively. Those observations with $|t| > 3$ are considered outliers, a classification based exclusively on that dimension. Under this measure and using both classical and robust estimators, we detect that variables X_4, X_8 and X_{10} contain at least 20% atypical observations. This analysis will be helpful hence forth.

We use Principal Component Analysis to find the most important source of variation among variables. In order to limit the effect of anomalous observations in the reduction of dimensionality in data, results from both classical and robust PCA are showed. Using any of the criteria described in Section 2.2 shows that, with $k = 2$ components, there is an explained variance of 85.42% corresponding to classical PCA (CPCA) and 91.6% for robust PCA (RPCA), which in both cases is acceptable.

In order to compare both analysis, we first display the score distance against the orthogonal distance and it is clear how different outliers are identified by each method. In Figure 5.3 we can distinguish some groups of points in each quadrant given by the cut–offs of 97.5% quantile of the χ^2–distribution. Using (a) CPCA, regular observations represent 76.6% of the total; good leverage represent 10.8% and orthogonal represent 12.6% of the observations. We might expect some bad leverage points, but those outliers are misclassified when using this method. The reason is the presence of outliers, since these observations are influencing the projection of observations onto the corresponding PCA subspace as it depends on the sample mean. Therefore, it would be dangerous to replace those observations with their projected values. However, when using robust estimators for location and covariance matrix, it is possible to handle outliers in a much better way than with classical PCA. The classification given by (b) RPCA is: 77.4% of regular points; 2.7% good leverage; 11.7% orthogonal and 8.0% of bad leverage points, representing a big difference between these two methods. It is important to mention that those bad leverage points given by RPCA are classified as good leverage points in CPCA.

Another difference between CPCA and RPCA is appreciated in the plot of scores, given in Figure. 5.4, where we observe the association between the first two principal components (or *scores*) obtained from both methods. We also compute the Mahalanobis distance based on classical and robust estimators, together with the tolerance ellipses defined by the set of vectors whose squared Mahalanobis distance is equal to

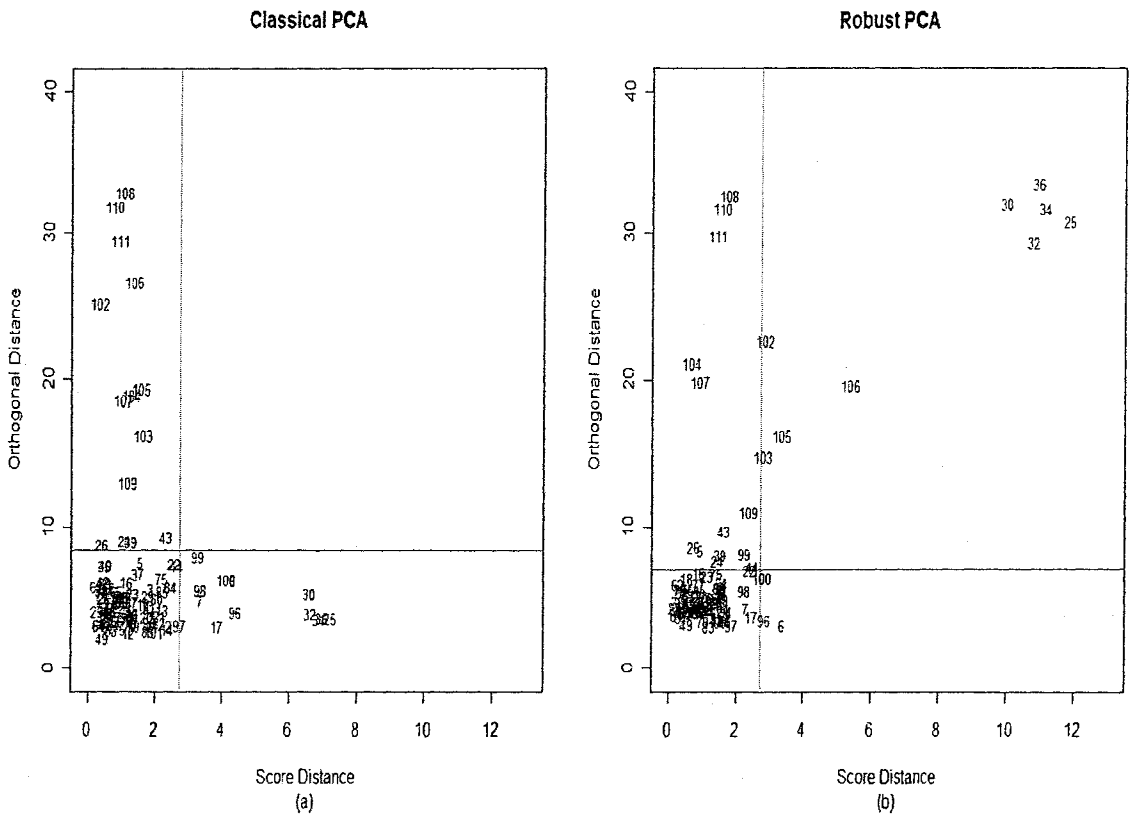


Figure 5.3: Score vs Orthogonal distances with Classical PCA.

the same 97.5% quantile. Observations which fall outside the tolerance ellipse are by definition the good and bad leverages points, but it is clear in Figure 5.4 that (a) the corresponding ellipse is highly inflated towards outliers, which are identified as the bad leverage points. That means that the first two principal components are not lying in the direction of the highest variability of the regular points and they are being magnified by that set of atypical observations. This situation does not occur in Figure 5.4(b), where scores are obtained by RPCA and regular points are enclosed by the ellipse.

On the other hand, assuming we choose two components to reduce dimensionality, we now describe the variables that better explain the first and second components in both the classical and robust PCA and we detect another crucial difference between these two approaches. As we previously mentioned, some variables have a

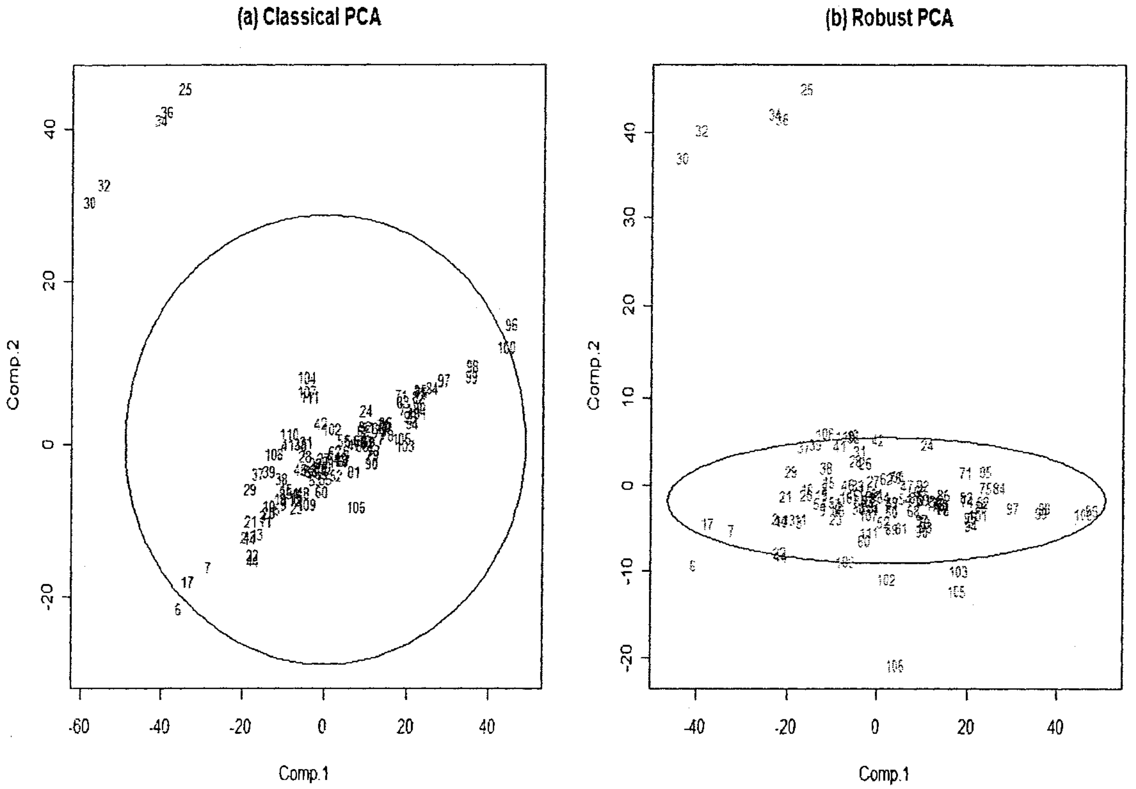


Figure 5.4: Score plot obtained with (a) CPCA and (b) RPCA with their corresponding 97.5% tolerance ellipses.

large amount of outliers in its projection. In Figure 5.5(a), we observe the influence of outliers in the loadings of X_8 and X_{10}, in the first component given by CPCA, which is reduced in (b) when using RPCA. Note that in both cases the directions of projections are attracted by the bad leverage observations. Thus, using the robust approach, the first component is mainly composed by the overall length (X_1), the length of wheelbase (X_2) and the width of car (X_3), while the second component is mostly explained by the rear fore-and-aft seating room (X_8) and rear shoulder room (X_{10}).

So far, we have reduced the original dimensionality $p = 11$ to $k = 2$, which allows us to have a better understanding of the database, since it is possible to plot

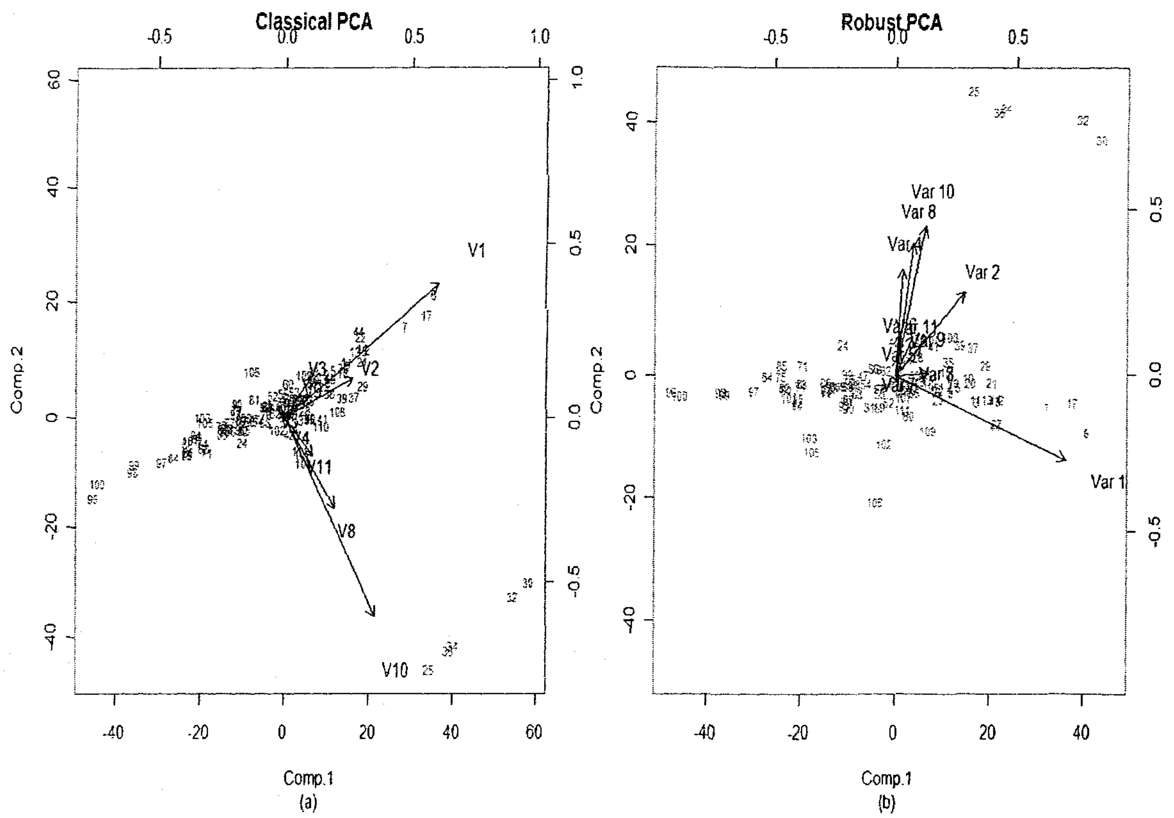
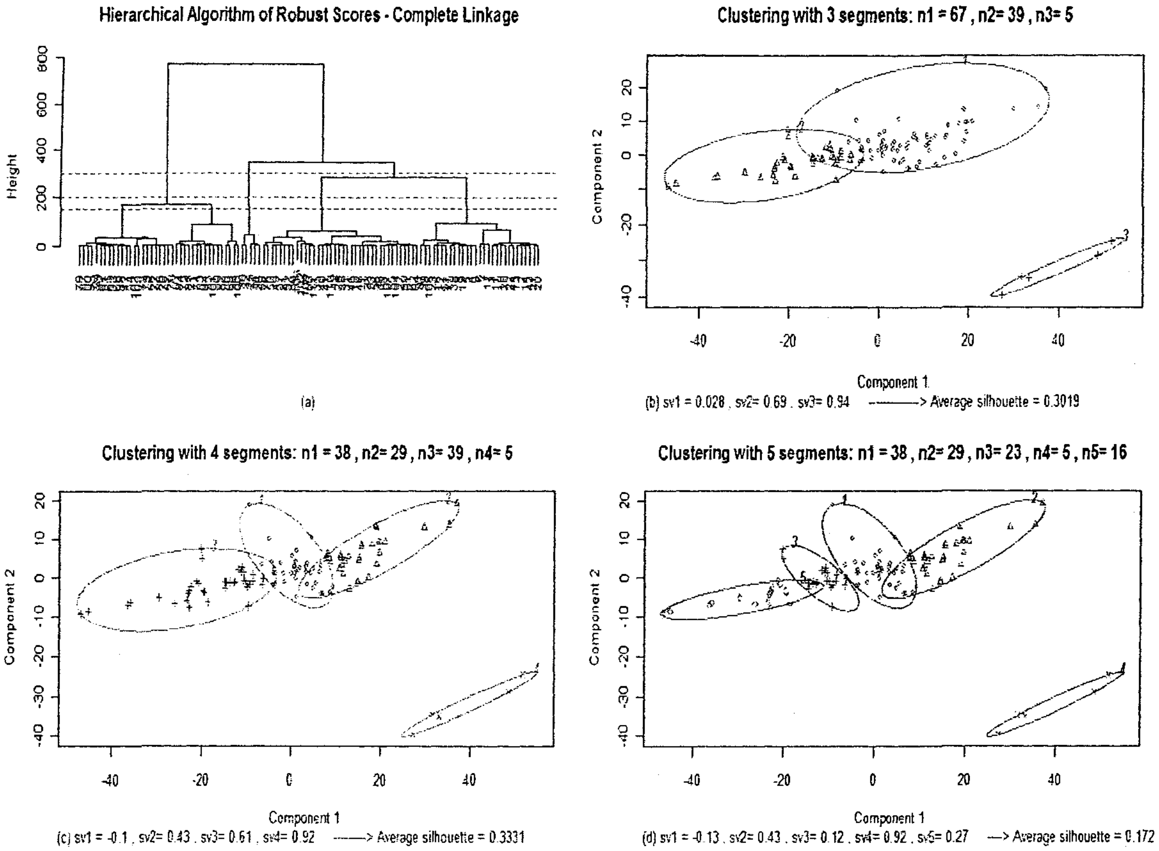


Figure 5.5: Biplots: (a) CPCA, (b) RPCA.

observations without loss of information. The purpose of this thesis is to identify clusters of well-separated groups to create a segmentation of individuals according to their attributes, by using these two scores (Z_1, Z_2), as new variables that summarize the association between original variables.

5.3 Use of the Robust Clustering Method

In order to analyze the influence of outliers in clustering, both the classical method (based on classical dissimilarity measures) and a robust one (based on robust approaches) are analyzed. A dissimilarity matrix is obtained to store all the pairwise distances between observations in a data set, which have been converted into scores according to results from the previous section.


Hierarchical Algorithm of Robust Scores - Complete Linkage

(a)

Clustering with 3 segments: n1 = 67 , n2= 39 , n3= 5

(b) sv1 = 0.028 , sv2= 0.69 . sv3= 0.94 ————> Average silhouette = 0.3019

Clustering with 4 segments: n1 = 38 , n2= 29 , n3= 39 , n4= 5

(c) sv1 = -0.1 , sv2= 0.43 . sv3= 0.61 , sv4= 0.92 ——> Average silhouette = 0.3331

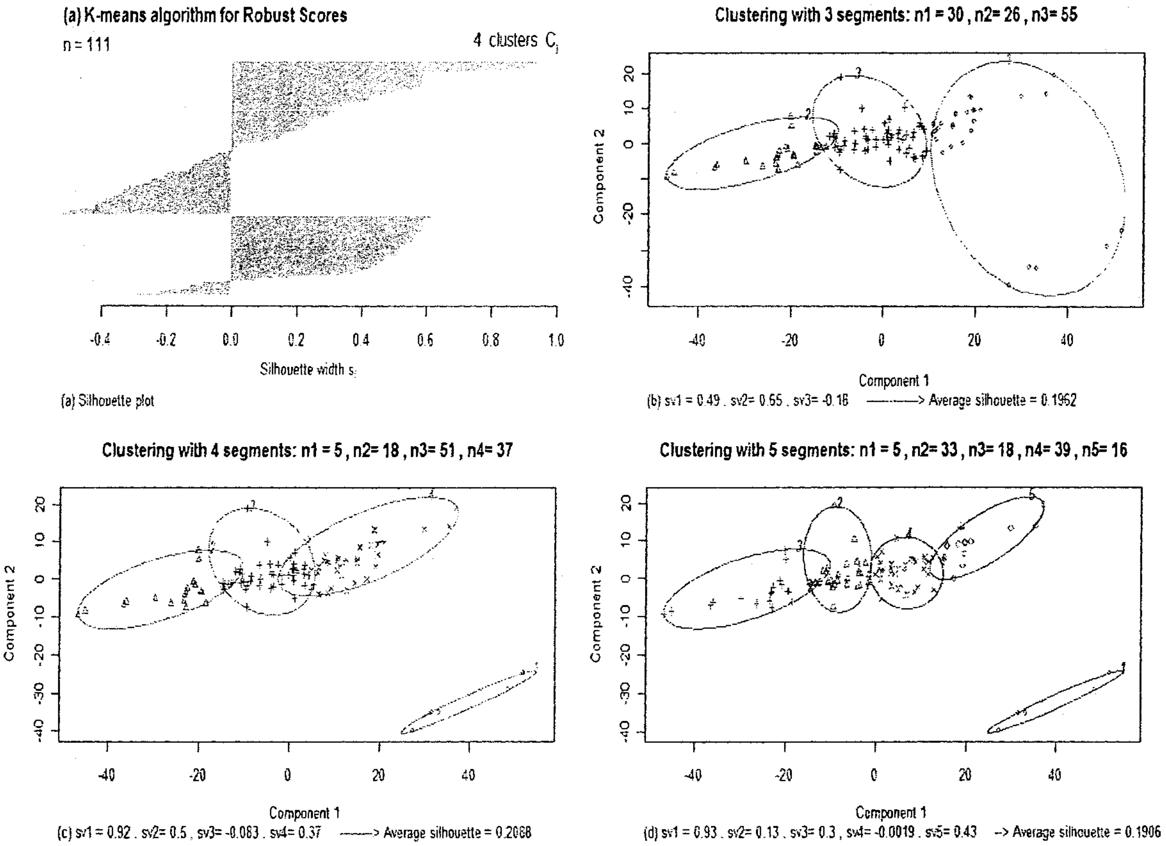
Clustering with 5 segments: n1 = 38 , n2= 29 , n3= 23 , n4= 5 , n5= 16

(d) sv1 = -0.13 . sv2= 0.43 . sv3= 0.12 . sv4= 0.92 . sv5= 0.27 —> Average silhouette = 0.172

Figure 5.6: Hierarchical Complete Linkage algorithm: (a) Dendrogram, (b) 3 clusters, (c) 4 clusters, (d) 5 clusters

With the statistical software **R** (it could be another) a classical clustering procedure is ran, as described in Section 2.3, in order to analyze the impact of outliers in their results. We do this for both hierarchical and non-hierarchical methods. For each segmentation, the silhouette value (*sv*) given in (2.9) is calculated, in order to decide how many clusters will be considered to classify observations.

After running different type of hierarchical algorithms, the dendrogram corresponding to the *Complete linkage* method, as described in Section 2.3.1, is chosen and results are showed in Figure 5.6. According to the *average silhouette value* (*asv*), the best classification is obtained using 4 segments, but this index ($asv = 0.33$) is close to 0, which indicates a high misclassification of individuals in each cluster. Note that Cluster 1 has a very small silhouette value and this is due to the intersection with Clusters 2 and 3.



(a) Silhouette plot

(b) sv1 = 0.49 . sv2= 0.65 . sv3= -0.16 ———> Average silhouette = 0.1962

(c) sv1 = 0.92 . sv2= 0.5 . sv3= -0.083 . sv4= 0.37 ———> Average silhouette = 0.2068

(d) sv1 = 0.93 . sv2= 0.13 . sv3= 0.3 , sv4= -0.0019 . sv5= 0.43 --> Average silhouette = 0.1906

Figure 5.7: Partitioning K-means algorithm: (a) Silhouette plot, (b) 3 clusters, (c) 4 clusters, (d) 5 clusters.

Another important fact is that the average index is highly influenced by the index in Cluster 4 ($sv_4 = 0.92$), which corresponds to a value computed using only 5 observations. The same situation is repeated with all 5 clusters. In general, we conclude that this algorithm does not provide a good clustering of observations. This is mainly explained by the presence of outliers, since this algorithm is based on classical dissimilarity measures. Note that, even when the highest *asv* is reached with 4 segments, there is one cluster with *sv* close to 0, which indicates a poor classification.

Now, we run the K-means algorithm explained in Section 2.3.2. We should expect some improvements in the results with respect to the previous algorithm, since this is an iterative method that looks for the best segmentation by assigning each observation to the nearest centroid in each run. But the problem is that it recalculates the initial clusters and the final result might differ each time. Therefore we generate 5,000 clusters and the "mode" of the composition of the segmentation is chosen as a final result. Results from the final clustering are showed in Figure 5.7 and we do not observe any improvement neither in the *avs* nor in the *sv* for each cluster. Also, we still have important intersections between clusters, which make the average index smaller.

In Section 4.3.1 we discuss the importance of the PAM algorithm as a robust method to classify observations, since it minimizes a sum of dissimilarities, instead of a sum of squared euclidean distances. Results are displayed in Figure 5.8, where there is a clear improvement in the segmentation with respect to previous algorithms. We first observe that the highest *asv* is reached with 4 clusters (asv=0.4355). Within each cluster there is a small intersection among clusters, as the *sv* is still high or at least greater than 0, comparing with previous algorithms. This situation is identified in (a), where the silhouette plot is displayed, showing a good separation between groups. Therefore, the PAM algorithm with 4 clusters will be considered the best one from these three algorithms analyzed.

As a parentheses before going into details, it is important to remember that a first classification of observations was given by RPCA in Figure 5.3, where a group of 86 *regular* observations, 3 *good leverage* points, 13 *orthogonal* observations and 9 *bad*

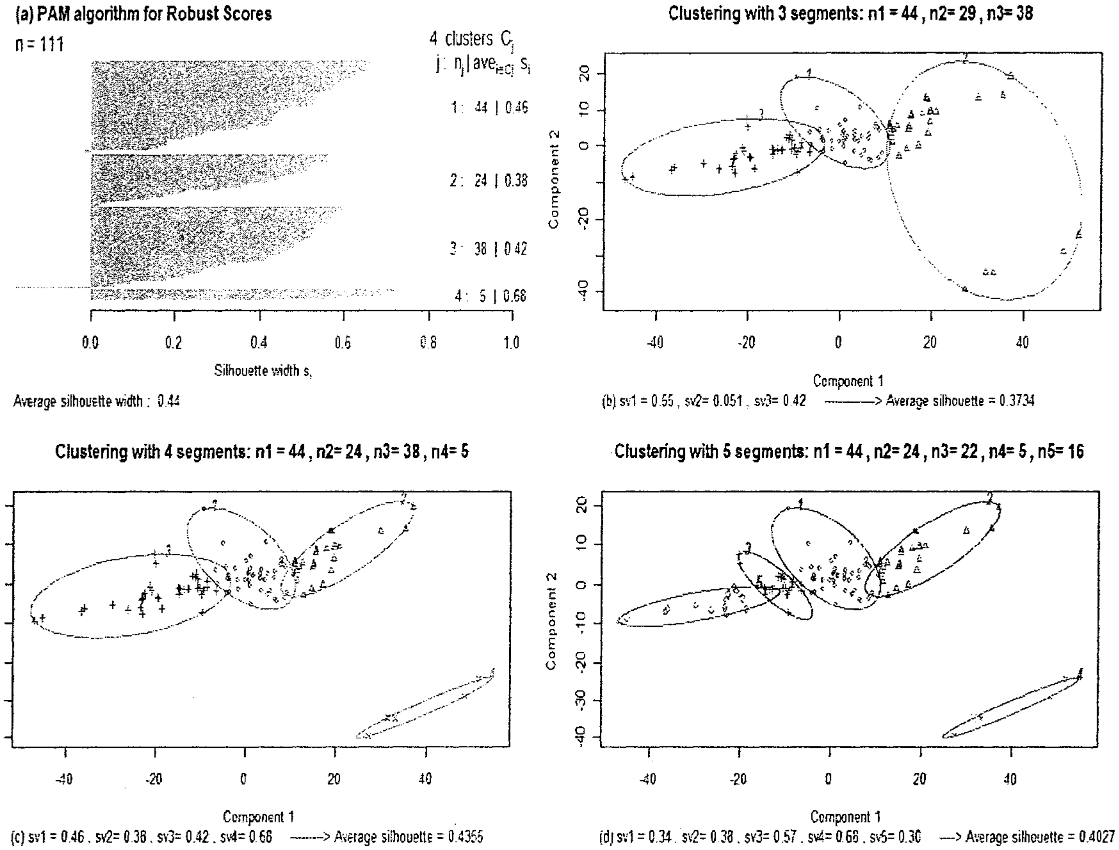


Figure 5.8: PAM algorithm: (a) Silhouette plot, (b) 3 clusters, (c) 4 clusters, (d) 5 clusters.

leverage points were detected. Comparing those groups with segments obtained with cluster algorithms, specially the one obtained by the PAM algorithm with 4 clusters that is illustrated in Figure 5.9, we see that some of the bad leverage observations (25, 30, 32, 34 and 36) form a single cluster. Those points represent the “worst” scores, as they fall far from the bulk of the data and far from the projection. The rest of that group (102, 103, 105 and 106) is classified into the Clusters 1 (102 and 106 that are closer in orthogonal distance) and 2 (103 and 105 that are closer in score distance). This means that PAM assigns some bad leverage points to groups for which they share similar characteristics.

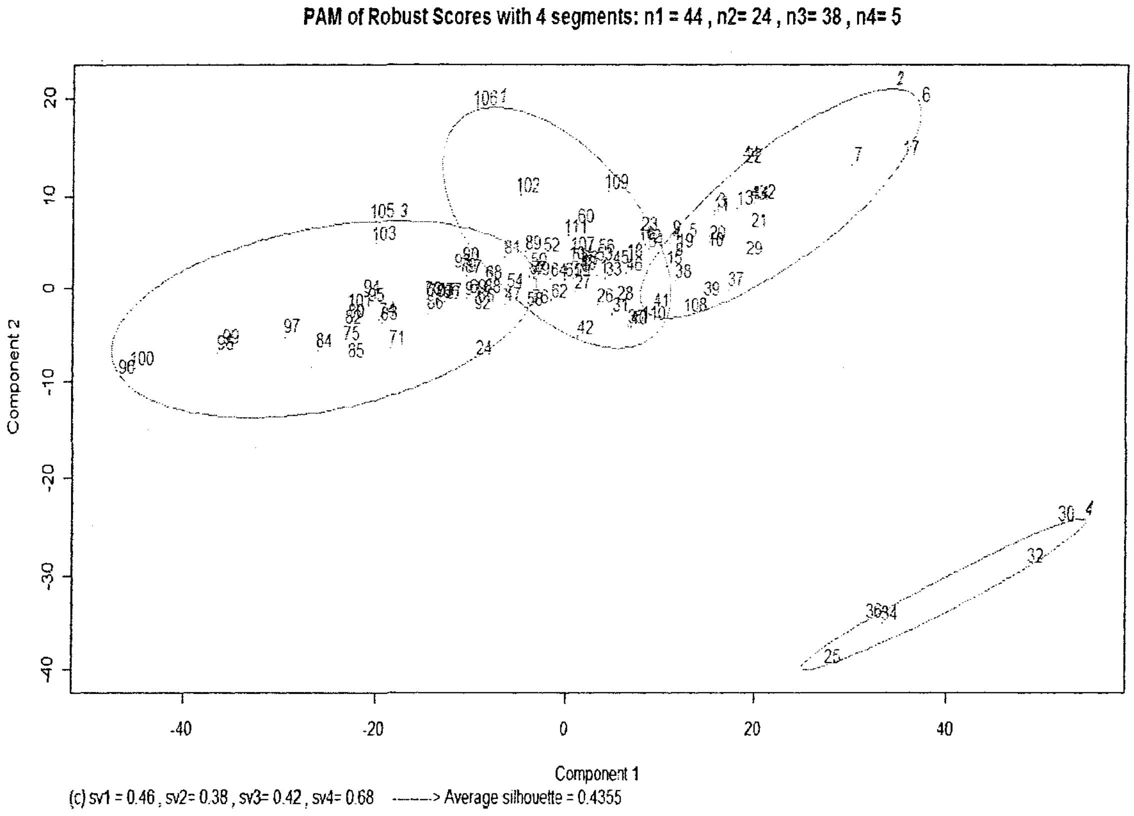


Figure 5.9: PAM algorithm for 4 clusters.

This situation occurs with the orthogonal points (5, 24, 26, 39, 43, 44, 99, 104, 107, 108, 109, 110, 111) where PAM allots observations into different clusters, according to their orthogonal and score distance. For instance, observations 26, 104, 107, 109, 110 and 111 are assigned to Cluster 1, where those points keep a small score distance, which allow them to share characteristics with the rest of this cluster. Observations 5, 39, 43, 44 and 108 are assigned to Cluster 2, where their common property is their closer orthogonal distance, except for observation 108, which looks very far from the rest.

Good leverages points (6, 96 and 100) are allotted into Cluster 2 (6) and Cluster 3 (96, 100) whose ellipses look inflated by the influence of those observations, see Figure 5.9. Therefore the corresponding silhouette values are smallest in this segmentation

($sv_2 = 0.38$ and $sv_3 = 0.42$).

These segments can be used as a predictor in order to estimate the premium in a car insurance database, as it is easier to manage a variable with 4 categories instead of the original one with 111 car brands. Also, there are car characteristics that are not easy to differentiate by only knowing cars dimensions. With this classification it is possible to group cars that share some more detailed attributes.

In conclusion, a reduction of dimensionality of data is first required to better explain the association among observations. We then apply a robust clustering algorithm to create groups with similar characteristics. This way a more precise classification is obtained. In order to describe each cluster and perhaps to name them, some intervals within each variable and then some crosstables may be created.

Bibliography

[1] Anderberg, M.R. (1975) "Cluster analysis for applications", *Society for Industrial and Applied Mathematics*, **17**, pp.580-582.

[2] Billingsley, P. (1999) "Convergence of Probability Measures", *John Wiley and Sons*, 2nd edition, pp.1-28.

[3] Box, G.E.P. (1954) "Some theorems on quadratic forms applied in the study of analysis of variance problems: effect of inequality of variance in one–way classification", *The Annals of Mathematical Statistics*, **25**, pp.33-51.

[4] Campbell, N.A. (1980) "Robust procedures in multivariate analysis I: robust covariance estimation", *Applied Statistics*, **29**, pp.231-237.

[5] "Consumer Reports, April 1990", *Grey Castle Press*, pp.287–288. http://www.consumerreports.org/cro/index.htm

[6] Critchley, F. (1985) "Influence in principal component analysis", *Biometrika*, **72**, pp.627-636.

[7] Croux, C. and Haesbroeck, G. (1999) "Influence function and efficiency of the minimum covariance determinant scatter matrix estimator", *Journal Multivariate Analysis*, **71**, pp.161-190.

[8] Croux, C. and Haesbroeck, G. (2000) "Principal components analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies", *Biometrika*, **87**, pp.603-618.

[9] Davies, P.L. (1987) "Asymptotic behavior of S-estimates of multivariate location parameters and dispersion matrices", *The Annals of Statistics*, **15**, pp.1269-1292.

[10] Devlin, S.J., Gnanadesikan, R. and Kettering, J.R. (1981) "Robust estimation of dispersion matrices and principal components", *Journal of the American Statistical Association*, **76**, pp.354-362.

[11] Devroye, L. (1997) "A Probabilistic Theory of Pattern Recognition", *Springer*, New York, pp.192-196.

[12] Donoho, D.L. (1982) "Breakdown properties of multivariate location estimators", *Ph.D. Qualifying paper*, Harvard University.

[13] Donoho, D.L. and Huber, P.J. (1983) "The notion of breakdown point", *A Festschrift for Erich L. Lehmann, In: P.J. Bickel, K.A. Doksum and J.L. Hodges (Ed.'s)*, Belmont, CA, pp.157-184.

[14] Gather, U. and Hilker, T. (1997). "A note on Tyler's modification of the MAD for the Stahel–Donoho estimator", *The Annals of Statistics*, **25**, pp.2024-2026.

[15] Gervini, D. (2002). "The influence function of the Stahel-Donoho estimator of multivariate location and scatter", *Statistics and Probability Letters*, **60**, pp.425-435.

[16] Gnanadesikan, R. and Kettenring, J.R. (1972). "Robust estimates, residuals, and outlier detection with multiresponse data", *Biometrics*, **28**, pp.81-124.

[17] Hampel, F.R. (1971). "A general definition of qualitative robustness", *The Annals of Mathematical Statistics*, **42**, pp.1887-1896.

[18] Hampel, F.R. (1974). "The influence curve and its role in robust estimation", *The Annals of Statistics*, **69**, pp.383-393.

[19] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W. A. (1986). "Robust Statistics: The Approach Based on Influence Functions", *Wiley*, New York.

[20] Hardin, J. and Rocke, D. (2002). "The distribution of robust distances", *http://www.cipic.ucdavis.edu/ dmrocke/preprints.html*

[21] Hardin, J. and Rocke, D. (2004). "Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator", *Computational and Statistical Analysis*, **44**, pp.625-638.

[22] Hardle, W. and Simar, L. (2003). 2th ed. "Applied Multivariate Statistical Analysis", *Springer*, New York, pp.271-276.

[23] Hartigan, J.A. and Wong, M.A. (1979) "A k-means clustering algorithm", *Applied Statistics*, **28**, pp.100-108.

[24] Hawkins, D.M. (1974) "The detection of errors in multivariate data using principal components", *Journal of the American Statistical Association*, pp.340-344.

[25] Houston, D.B. (1960) "Risk, Uncertainty and Sampling", *Journal of Risk and Insurance*, **27**, pp.77-82.

[26] Huber, Peter J. (1964) "Robust estimation of a location parameter", *The Annals of Mathematical Statistics*, **35**, pp.73-101.

[27] Huber, Peter J. (1967) "The behavior of maximum likelihood estimates under nonstandard conditions", *Proceedings of the Fifth Berkeley Symposium on Mathematics and Statistics Probability, Universidad of California Press*, **1**, pp.221-233.

[28] Huber, Peter J. (1981) "Robust Statistics", *John Wiley & Sons, Inc.*, New York, pp.7-19.

[29] Hubert, Mia and Engelen Sanne. (2004) "Robust PCA and classification in biosciences", *Bioinformatics, Vol.20*, Oxford University Press, pp.1728-1736.

[30] Johnson, Richard A. (1982) "Applied Multivariate Statistical Analysis", *Prentice Hall*, New York, pp.426-465.

[31] Kaufman, L. and Rousseeuw, P.J. (1987) "Clustering by means of medoids", *In: Y. Dodge (Ed.). Statistical Data Analysis based on the L_1 norm*, North-Holland, Amsterdam, pp.405-416.

[32] Lehmann, E. L. (1999) "Elements of Large-Sample Theory", *Springer*, New York, pp.47-93.

[33] Lopuha, H.P. (1989) "On the relation between S-estimators and M-estimators of multivariate location and covariance", *The Annals of Statistics*, **17**, pp.1662-1683.

[34] Lopuha, H.P. (1999) "Asymptotics of reweighted estimators of multivariate location and scatter", *The Annals of Statistics*, **27**, pp.1638-1665.

[35] MacQueen, J. (1967) "Some methods for classification and analysison multivariate observations", *In: L.Le Cam, J. Neyman (Eds.), 5th Berkeley Symp. Math. Stat. Prob.*, **1**, pp.281-297.

[36] Maronna, R.A. (1976) "Robusts M-estimators of multivariate location and scatter", *The Annals of Statistics*, **4** , pp.51-67.

[37] Maronna, R.A. and Yohai, V.J. (1995) "The behavior of the Stahel-Donoho robust multivariate estimator", *Journal of the American Statistical Association*, **90** , pp.330-341.

[38] Maronna, R.A. and Zamar, R.H. (2002) "Robust estimation of location and dispersion for high–dimensional data sets", *Technometrics*, **44** , pp.307-317.

[39] Maronna, R.A. (2006) "Robust Statistics: Theory and Methods", *John Wiley & Sons, Ltd.*

[40] Nomikos P. and MacGregor J.F. (1995) "Multivariate SPC Charts for Monitoring Batch Processers", *Technometrics*, **37**, pp.41-59.

[41] Olive, D.J. (2006) "Applied Robust Statistics", *Department of Math, Southern Illinois University.* Unpublished book.

[42] Rao, C.R. (1964) "The use and interpretation of principal component analysis in applied research", *Sankhya*, **A26**, pp.329-358.

[43] Rao, C.R. (1973) "Linear Statistical Inference and Its Applications", 2nd Ed., *Wiley*, New York, Section 1c.3.

[44] Rousseeuw, P.J. (1984) "Least median of squares regression", *Journal of the American Statistical Association*, **79**, pp.871-880.

[45] Rousseeuw, P.J. and Leroy A.M. (1987) "Robust Regression and Outlier Detection", *Wiley Series in Probability and Statistics*, New York, pp.216-248.

[46] Rousseeuw, P.J. and Van Driessen, K. (1999) "A fast algorithm for the minimum covariance determinant estimator", *Technometrics*, **41**, pp.212–223.

[47] Seber, G.A.F. (1984) "Multivariate Observations", *John Wiley & Sons, Ltd*, New York.

[48] Stahel, W.A. (1981). "Breakdown of covariance estimators, Research Report 31", *Fachgruppe fr Statistik, ETH*, Zurich.

[49] Timm, N.H. (2002). "Applied Multivariate Analysis", *Springer*, 9th Ed., New York, pp.445-472 and pp.515-533.

[50] Tukey, J.W. (1960) "A Survey of Sampling from Contaminated Distributions, Contributions to Probability and Statistics", *Standford University Press*, Standford, C.A.

[51] Tukey, J.W. (1962) "The future of data analysis", *The Annals of Mathematical Statistics*, **33**, pp.1-67.

[52] Tukey, J.W. (1977) "Exploratory Data Analysis", *Addison-Wesley.*

[53] Tyler, D.E. (1990) "Breakdown properties of the M-estimators of multivariate scatter", Technical Report, *Department of Statistics*, Rutgers University.

[54] Walters, M.A. (1981) "Risk classification standards", *Proceedings of the Casualty Actuarial Society*, **68**, pp.1-18.

[55] Wilson E.B. and Hilferty M.M. (1931) "The distribution of chi-square", *Proceeding of the National Academy of Sciences of the United States of America*, **17**, pp.684-688.

[56] Zuo, Y., Cui, H. and He, X. (2004). "On the Stahel-Donoho estimator and depth-weighted means of multivariate data", *The Annals of Statistics*, **32**, pp.167-188.

Appendix A

Matrix Theory

A.1 Random Vector and Matrices

A matrix $\mathrm{A}_{n \times p}$ is a set of observations with n rows and p columns:

$$\mathrm{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{np} \end{bmatrix}$$

We also write (a_{ij}) for A, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$ to indicate the number of rows and columns. Therefore, we call the *transpose* of A (A') the matrix with elements (a_{ji}). A vector is a matrix with one column and is denoted as $x_{p \times 1}$.

Definition A.1 *A set of vectors* $\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_k$ $(k \leq n)$ *are said to be linearly dependent if constants* $c_1, c_2, ..., c_k$*, not all zero can be found such that* $\sum_{i=1}^{k} c_i \boldsymbol{a}_i = \boldsymbol{0}$

Definition A.2 *The rank of a matrix* $\boldsymbol{A}_{n \times p}$*, rank(*$\boldsymbol{A}$*), is defined as the maximum number of linearly independent rows (columns). A set of vectors are linearly independent when they are not linearly dependent. The maximum possible rank of* $\boldsymbol{A}_{n \times p}$ *is the minimum of* n *and* p*, in which case* $\boldsymbol{A}$ *is said to be of full rank.*

Definition A.3 *The trace of a square matrix* $(n=p)$ $\boldsymbol{A}_{p\times p}$, $tr(\boldsymbol{A})$, *is defined as the sum of its diagonal elements:*

$$tr(\boldsymbol{A})=\sum_{i=1}^{p} a_{ii}$$

Definition A.4 *The determinant of a square matrix* $\boldsymbol{A}_{p\times p}$, $det(\boldsymbol{A})$, *is defined as:*

$$det(\boldsymbol{A})=|\boldsymbol{A}|=\sum_{j=1}^{p} a_{ij}C_{ij} \overset{def}{=} \sum_{j=1}^{p} a_{ij}(-1)^{(i+j)}M_{ij},$$

where the M_{ij} *represent the minor matrix, i.e., the determinant of the matrix that results from* $\boldsymbol{A}$ *by removing the* i*th row and the* j*th column. This is called the Laplace's formula and it is very efficient for relatively small matrices.*

When $p=2$, $\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, the determinant of $\mathbf{A}$ is $det(\mathbf{A})=a_{11}a_{22}-a_{12}a_{21}$. A matrix $\mathbf{A}$ is called *singular* if its determinant is 0.

Definition A.5 *If a matrix* $\boldsymbol{A}$ *is square and of full rank, then* $\boldsymbol{A}$ *is said to be nonsingular and* $\boldsymbol{A}$ *has a unique inverse, denoted by* $\boldsymbol{A}^{-1}$, *with the property that* $\boldsymbol{A}\boldsymbol{A}^{-1}$=$\boldsymbol{A}^{-1}\boldsymbol{A}$=$\boldsymbol{I}$, *where* $\boldsymbol{I}$ *is the identity matrix, i.e., a matrix with ones in its diagonal and zeros otherwise, that is,* $\boldsymbol{I}=diag(1,1,...,1)$

Definition A.6 *Two vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$ *of the same size* $(p\times 1)$ *are said to be orthogonal if* $\boldsymbol{a}'\boldsymbol{b}=\sum_{j=1}^{p} a_j b_j = 0$.

Geometrically, orthogonal vectors are perpendicular. If $\mathbf{a}'\mathbf{a}=1$, the vector $\mathbf{a}$ is said to be *normalized.* In fact, a vector $\mathbf{a}$ can always be normalized by dividing it by the scalar $\sqrt{\mathbf{a}'\mathbf{a}}$. Thus $\mathbf{c}=\frac{\mathbf{a}}{\sqrt{\mathbf{a}'\mathbf{a}}}$ is normalized so that $\mathbf{c}'\mathbf{c}=1$.

A matrix $\mathbf{A}_{n\times n}$ whose columns are normalized and mutually orthogonal is called an orthogonal matrix, thus $\mathbf{A}'\mathbf{A}=\mathbf{A}\mathbf{A}'=\mathbf{I}$. Multiplication by an orthogonal matrix has the effect of rotating axes.

If $\mathbf{A}_{n\times m}$, where $m<n$, such that all m columns are orthogonal to each other. In that case we say that $\mathbf{A}$ is *sub-orthogonal* for those m columns.

A.2 Eigenvalues and Eigenvectors

Consider a matrix $\mathbf{A}_{p \times p}$. If there exists a scalar λ and a vector γ such that

$$\mathbf{A}\gamma = \lambda\gamma. \tag{A.1}$$

Then we call λ an *eigenvalue* and γ an *eigenvector.* To find λ and γ, we write (A.1) as

$$(\mathbf{A} - \lambda\mathbf{I}_p)\gamma = 0. \tag{A.2}$$

If $|\mathbf{A} - \lambda\mathbf{I}_p| \neq 0$, then $(\mathbf{A} - \lambda\mathbf{I}_p)$ has an inverse and $\gamma = \mathbf{0}$ is the only solution. Therefore, in order to obtain nontrivial solutions, we set $|\mathbf{A} - \lambda\mathbf{I}_p| = 0$, that is, we need this matrix to be singular to find values of λ that can be substituted in (A.2) to find the corresponding eigenvectors γ. The equation $|\mathbf{A} - \lambda\mathbf{I}_p| = 0$ is called the *characteristic equation.*

Therefore, there are up to p eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ of $\mathbf{A}$. For each eigenvalue λ_j, there exists a corresponding eigenvector γ_j given by equation (A.1).

Suppose that the matrix $\mathbf{A}$ has the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$. Let $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$. The determinant and the trace of $\mathbf{A}$ can be rewritten as:

$$|\mathbf{A}| = |\Lambda| = \prod_{j=1}^{p} \lambda_j, \tag{A.3}$$

$$tr(\mathbf{A}) = tr(\Lambda) = \sum_{j=1}^{p} \lambda_j. \tag{A.4}$$

The following theorem is very useful and important for the purpose of the thesis.

Theorem A.1 *(Spectral Decomposition) For a real symmetric matrix $\boldsymbol{A}_{n \times n}$ there exists an orthogonal matrix $\boldsymbol{P}_{n \times n}$ with columns $\boldsymbol{p}_j$ such that $\boldsymbol{A}$ can be written as:*

$$\boldsymbol{A} = \boldsymbol{P}\Lambda\boldsymbol{P}'$$

where Λ is a diagonal matrix with diagonal elements $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$.

For further details, see Rao (1973).

From Theorem $A.1$ we know that $\mathbf{AP} = \mathbf{P}\Lambda$, $\mathbf{PP}' = \mathbf{I} = \sum_{i=1}^{n} \mathbf{p}_i\mathbf{p}_i'$ and $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}' = \sum_{i=1}^{n} \lambda_i\mathbf{p}_i\mathbf{p}_i'$. If the $r(A) = r \leq n$, then there are r nonzero elements on the diagonal of Λ. A symmetric matrix for which all $\lambda_i > 0$ is said to be positive definite $(p.d.)$ and positive semidefinite $(p.s.d.)$ if some $\lambda_i > 0$ and at least one is equal to zero.

Theorem A.2 *(Singular Value Decomposition) Any matrix $\boldsymbol{B}_{m\times n}$ can be presented as*

$$\boldsymbol{B} = \boldsymbol{UQV}',$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal or sub-orthogonal and $\boldsymbol{Q}$ is a $n \times n$ diagonal matrix.

For further details of this theorem, see Rao (1973).

If m is larger than n, then $\mathbf{U}$ is sub-orthogonal and $\mathbf{V}$ is orthogonal. If m is smaller than n, then after ignoring the last $n - m$ zero columns of $\mathbf{U}$, this reduced matrix, say $\mathbf{U}_*$ and $\mathbf{V}$ are both orthogonal. The diagonal places of matrix $\mathbf{Q}$ contain the singular values of $\mathbf{B}$

Appendix B

Convergence of Random Variables

B.1 Convergence in Probability

Definition B.1 *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Then $\{X_n\}$ is said to converge in probability to X if for every $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0.$$

We write $X_n \xrightarrow{Pr} X$ to indicate convergence in probability.

B.2 Convergence in Distribution

Definition B.2 *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Suppose that X_n has distribution F_n and X has distribution function X. We say that $\{X_n\}$ converges in distribution to the random variable X if:*

$$\lim_{n \to \infty} F_n(t) = F(t), \quad \textit{for any } t \in \mathbb{R}$$

We write $X_n \xrightarrow{d} X$ to indicate convergence in distribution.

B.3 Convergence with Probability 1

Definition B.3 *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that $\{X_n\}$ converges almost surely to the random variable X if and only if:*

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

We write $X_n \overset{a.s.}{\to} X$ to indicate convergence almost surely or almost everywhere or with probability 1.

B.4 Convergence in Mean

Definition B.4 *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that $\{X_n\}$ converges in the rth mean or in the L^r norm to the random variable X if for $r \geq 1$, $\mathbb{E}|X_n|^r < \infty$ for all n, and:*

$$\lim_{n\to\infty} E(|X_n - X|^r) = 0.$$

We write $X_n \overset{L^r}{\to} X$ to indicate convergence in rth mean.

The chain of implications between the various notions of convergence, using the arrow notation, are as follows:

$$\overset{a.s.}{\to} \quad \Rightarrow \quad \overset{Pr}{\to} \quad \Rightarrow \quad \overset{d}{\to}$$

$$\forall r > 0 : \overset{L^r}{\to} \quad \Rightarrow \quad \overset{Pr}{\to}$$

$$\forall r > s \geq 1 : \overset{L^r}{\to} \quad \Rightarrow \quad \overset{L^s}{\to}$$

For more details see Billingsley (1999).

B.5 Consistency

Definition B.5 *A sequence $T_1, T_2, \ldots$ of estimators is consistent for parameter θ if:*

$$T_n \overset{Pr}{\to} \theta, \quad \textit{as } n \to \infty.$$

B.6 The Continuous Mapping Theorem

Theorem B.1 *Let $\{X_n\}$ be a sequence of random variables with values in a metric space such that $X_n \xrightarrow{d} X$ and h is a continuous function on the metric space. Then:*

$$h(X_n) \xrightarrow{d} h(X).$$

For more details see Billingsley (1999).