UNIVERSITEIT • STELLENBOSCH • UNIVERSITY

THE VISUALISATION OF MULTIDIMENSIONAL FINANCIAL DATA THROUGH BIPLOTS

Daniel Buitendag & Willem Lodewikus Pretorius

Research assignment presented in partial fulfilment
of the requirements for the degree of
BComHons **(Financial Risk Management**)

&

BComHons **(Mathematical Statistics**)
at the University of Stellenbosch

Department of Statistics and Actuarial Science

Supervisor: Mr. Carel Johannes van der Merwe

Degree of confidentiality: A

October 2018

PLAGIARISM DECLARATION

1. Plagiarism is the use of ideas, material and other intellectual property of another's work and to present it as my own.

2. I agree that plagiarism is a punishable offence because it constitutes theft.

3. I also understand that direct translations are plagiarism.

4. Accordingly, all quotations and contributions from any source whatsoever (including the internet) have been cited fully. I understand that the reproduction of text without quotation marks (even when the source is cited) is plagiarism.

5. I declare that the work contained in this assignment, except otherwise stated, is my original work and that I have not previously (in its entirety or in part) submitted it for grading in this module/assignment or another module/assignment.

18441386 & 17504651	*dbuitendag wlpretorius*
Student number	**Signature**
D Buitendag & W.L. Pretorius	19 October 2018
Initials and surname	**Date**

Acknowledgements

The author acknowledges institutions and people that significantly contributed to the research in this section.

For example:

The Department of Statistics and Actuarial Science (the Department) wishes to acknowledge the University of Stellenbosch Business School (USB) for generously making available their research report template and guidelines. The USB template and guidelines have been adapted for the purposes of the Department.

Abstract

In the modern finance world, the management of companies are responsible for making decisions based on inputs received from teams either within the firm or from external parties. The effectiveness of these inputs is dependent on the managers being able to understand and interpret them. In many cases visual aids such as pie charts and scatterplots are used, but it is believed that there exists a more powerful tool - perhaps one that can venture into the world of high dimensional data. This idea leads to the introduction of the biplot as a medium through which multidimensional data can be displayed. These high dimensional sets could vary from risk measures such as Value at Risk (VaR) or expected tail losses, to residuals of different models for estimating volatility, to financial ratios that are calculated from the management accounts of firms.

The biplot is introduced on a conceptual and theoretical level, followed by an example using a small data set. The two main types used in this research assignment are then revealed, namely Principal Component Analysis (PCA) biplots and Canonical Variate Analysis (CVA) biplots. The fundamentals of the two types are explained followed by two simple plots from the same arbitrary data set to highlight the differences. Methods for dealing with larger sets are introduced, of which some are used later in the research on a chosen data set. The set used, refers to the default status of Polish companies for the period 2000-2012, including 64 financial ratios, where observations of 5910 companies were taken. The set is then cleaned by outlier treatment methods and put through a variable selection process which leads to a large reduction in the number of variables used in the analysis. The end result is a set of 11 financial ratios, with a 12th variable being the defaulted status of the companies. In the default status variable, 410 companies defaulted and 5500 did not default.

Various biplots are plotted from the cleaned set. Due to the imbalance regarding the defaulted and non-defaulted proportion, samples of equal size are taken from the set. Next, various shapes of the different biplots are compared. Different groupings of strongly correlated ratios are made, and the isolated axes are plotted showing the varying relations for variables with different default statuses. The cumulative quality of the plots is then shown for a different number of dimensions, indicating how accurate the prediction would be in the two-dimensional case as used in the research. Lastly, density plots are drawn which are connected to the modelling of default rates in credit risk. The research assignment is ended with a conclusion and recommendations for further studies.

Key words:

Biplot; Canonical Variate Analysis; Principal Component Analysis; Variable selection

Opsomming

In die moderne finansiële wêreld is die bestuur van maatskappye verantwoordelik vir die neem van besluite gebaseer op insette wat van spanne ontvang word, hetsy binne die firma of van eksterne partye. Die effektiwiteit van hierdie insette is nie altyd maklik verstaanbaar en interpreteerbaar vir die bestuurders nie. In baie gevalle word visuele hulpmiddels soos sirkelgrafieke en strooiplate gebruik, maar daar word geglo dat 'n meer kragtige instrument bestaan - dalk een wat in die wêreld van hoë-dimensionele data kan ontstaan. Hierdie idee lei tot die bekendstelling van die biplot as 'n medium waardeur multidimensionele data vertoon kan word. Hierdie hoë-dimensionele stelle kan wissel van uit risiko-maatstawwe soos waarde-op-risiko, verwagte kortval, na finansiële verhoudings wat bereken word uit die bestuursrekeninge van firmas.

Die biplot word op konseptuele en teoretiese vlak aangebied, gevolg deur 'n voorbeeld van 'n klein data stel. Die twee hooftipes wat in hierdie navorsingsopdrag gebruik word, word dan onthul, naamlik die Hoof Komponent Analise biplots en Kanoniese Variasie Analise biplots. Die basiese beginsels van die twee tipes word verduidelik, gevolg deur twee eenvoudige grafieke uit dieselfde arbitrêre data stel om die verskille uit te lig. Metodes vir die hantering van groter stelle word bekend gestel, waarvan sommige later in die navorsing op 'n gekose data stel gebruik is. Die stel wat gebruik word, verwys na die standaardstatus van Poolse maatskappye vir die tydperk 2000-2012, insluitend 64 finansiële verhoudings op waarnemings van 5910 maatskappye geneem is. Die stel word dan skoon gemaak deur uitwisselings behandelings metodes en deur 'n veranderlike seleksie proses geplaas, wat lei tot 'n groot afname in die aantal veranderlikes wat in die analise gebruik word. Die eindresultaat is 'n stel van 11 finansiële verhoudings, met 'n 12de veranderlike die standaardstatus van die maatskappye.

Verskeie biplots word uit die skoongemaakte stel geteken. As gevolg van die wanbalans met betrekking tot die standaardstatus proporsie, is monsters van gelyke grootte uit die stel geneem. Vervolgens word verskillende vorme van die verskillende biplots vergelyk. Verskillende groeperings van sterk gekorreleerde verhoudings word gemaak en die geïsoleerde asse word getoon wat die wisselende verhoudings van veranderlikes met verskillende standaardstatusse toon. Die kumulatiewe kwaliteit van die grafieke word dan vir 'n aantal dimensies getoon, wat aandui hoe akkuraat die voorspelling in die twee-dimensionele geval sal wees soos in die navorsing gebruik word. Laastens word digtheids grafieke geteken wat verband hou met die modellering van wanbetalings koerse in krediet bestuur. Hierdie navorsing word geeindig met ‘n gevolgtrekking en aanbevelings vir toekomstige studies.

Sleutelwoorde:

Biplot; Kanoniese Variasie Analise; Hoof Komponent Analise; Veranderlike seleksie.

Table of contents

List of figures and tables

List of appendices

List of abbreviations and/or acronyms

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CVA	Canonical Variate Analysis
n	Observations
p	Variables
PCA	Principal Component Analysis
RSS	Residual Sum of Squares
SVD	Singular Value Decomposition
VaR	Value at Risk

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

In the modern world, financial data on companies are commonly available and usable by investors, firms, or anyone who finds the information useful. This data forms the basis upon which companies are valued and investigated, whereby in-depth analysis techniques have been designed to extract value from data which is in the form of income statements and balance sheets.

A hypothetical example of such a technique would be a model (built by a statistician working for an investment firm) that provides a valuation of the stock price of a firm by using inputs from a finance data set and then comparing that to the actual value of an investment. This method would involve statistical applications being applied to a multidimensional data set which in turn will output values that can be interpreted by the people who built the valuation model. However, in many situations, the models built cannot be understood by other people that do not have in-depth mathematical statistical knowledge and, therefore, some effectiveness in the process is lost. The moment the manager or decision maker feels that they do not completely comprehend what was done by the analyst, trust and confidence weakens. This leads to a deadweight loss in terms of company resources – and consequently there lies an opportunity for improvement.

It is believed that the visualisation of these statistical methods could be the solution to eliminate the above-mentioned deadweight loss. Visualisation is a powerful tool when it comes to creating links between theoretical knowledge and output-based application (Greenacre, 2010:9). It is a method that is used worldwide in academic scripts and textbooks for students, and it is a technique that should be used in all spaces where value can be added. Therefore, the idea of visualisation will be accompanied by using a biplot to extract different types of information that will relate to the simplification of portraying multidimensional financial data in two-dimensions.

1.2 BACKGROUND

In the finance world there are many inputs to a decision-making process. In many cases there is a panel of people who make the final decisions. Take for example, an investment firm faced with a decision on whether to invest in a certain industry, market, company or country. The decision makers (normally the management) will have teams that provide inputs that are designed to aid their decision-making process. Since it is an investment firm, many mathematical and statistical methods will be applied on different data sets where various models and analyses will be completed. These are then, in some form, presented to the management and a decision is made based on the information available.

Noticeably the method that different firms or institutions use vary, but there is one constant: inputs from parties outside the management are used in the process. This is where the gap is defined: the gap between the decision being made and the analyses being done to aid the decision. It is believed that this is the area where the biplot can act as a ‘bridge’.

The aim of this research assignment is to show how statistical knowledge can be ‘translated’ into common tongue without having a major loss of information using different kinds of biplots. This is a method that can be applied in a number of scenarios and can prove to be highly effective.

1.3 BIPLOTS

In the case where interest lies in the dependencies of different variables and their characteristics, biplots are believed to be an effective tool in the visualisation process. These plots use the data from a table and reveal the main structures in a systematic way – i.e. making the data transparent. They can plot multiple variables and show their dependencies in multidimensional ways, using a two-dimensional plane. The most critical part of biplots is the fact that they are not overly complex to interpret – simply having basic knowledge of the concept and its fundamentals will enable the reader to get information from this plot with relative ease. In the following section the emphasis on how a biplot can be used is elaborated on and in the following Chapter, the fundamentals in the construction process of biplots will be discussed.

1.4 PROBLEM STATEMENT

Financial data consist of thousands of inputs over very large periods of time. When graphical displays are done in general, they consist of a few isolated categories which are compared or plotted over time (e.g. Turnover and gross profit of five candidate firms plotted over a 5-year period). These graphics are useful, however, all the inputs available are not always included, therefore investigation is needed as to whether or not to include as many inputs as possible.

Isolating these values and presenting them graphically can be ineffective and in many cases can mislead people to how the data is being presented to. Also, comparing one or two categories between companies does not add a lot of value to the analysis procedure in general. When most statistical based analyses are done, many more variables are included. This, however, causes a problem in conventional graphical displays. A conventional plot is two-dimensional and thus only has two inputs (two-dimensional data). In the above case more inputs should be displayed, but the plot should be constructed on a two-dimensional field so that it remains interpretable. Two-dimensional plots are well understood in the world of business because of their simple interpretation by people across many fields of study.

This leads to one of the most attractive characteristics of biplots – the capability to plot multidimensional data sets on a two-dimensional space (Greenacre, 2012:399). Therefore, no limit on the number of variables has to be fixed when visualising data – when many more input variables are included it can still, through the use of biplots, provide a two-dimensional result.

Over the past decades there has been some major corporations, showing profitable business models, going bankrupt. This may lead to an uncertainty of the trustworthiness regarding some of these financial organisations and how their financial statements are prepared. Therefore, is it possible that financial ratios can be used for different companies as the input for the biplot whereby graphically and accurately identifying if they are exposed to an extreme risk such as defaulting? Is it possible to create a display that provides warning signs to potential downfall and other noticeable characteristics regarding a company's financial status?

In this research assignment a data set will be used that contains information in determining to what extent certain companies are exposed to default risk. There are many ways in which this form of risk can be analysed but in this case a data set that consists of only financial ratios will be considered.

The focus will extend to two main types of biplots – Principle Component Analysis (PCA) biplots and Canonical Variate Analysis (CVA) biplots. The fundamentals of these two types of plots will be explained in the next chapter. A short description of the techniques are as follows:

PCA biplots demonstrate how certain variables (e.g. financial risk indicators) change over time, as well as displaying the relationships between the changing variables (N J le Roux, S Gardner & P Olivier, 2003:42). This could potentially increase the difficulty of manipulating financial statements since each variable can be plotted to portray interrelationships whereas early warning signs can be detected allowing decision makers to act timely. CVA biplots work with segmentation and is suitable when different classes are being compared (N J le Roux *et al.*, 2003:44). This can be useful when working in the world of credit risk, since a biplot representing financial indicators of each company within a sector can be displayed and connected. CVA biplots also provide means of classification whereby making it possible to build prediction models to identify if a company will go bankrupt or not. Not only can these plots be used as a visualisation technique, but each plot is constructed in such a manner to provide statistical and financial analysis.

1.5 CLARIFICATION OF KEY CONCEPTS

This section highlights the main concepts that are being used in this assignment. For example, in this particular document the concepts of "report", "supervisor" etc. are being discussed.

1.5.1 Report

The report is the product of the research assignment component of the US honours degree.

1.5.2 Supervisor

Mr. CJ van der Merwe

1.6 CHAPTER OUTLINE

Chapter One provides an introduction to this research assignment with a short overview regarding the problem and probable solution to visualising multidimensional data. This Chapter emphasises the use of biplots as a possible visualisation method to ease financial and statistical analysis. Chapter Two provides the underlying theory as to how biplots are constructed. In this Chapter the different types of biplots are explained where the focus turns to understanding these plots in general with the use of simple examples. Chapter Three outlines the data set used in the analysis. In this Chapter, the data set and the data cleaning procedure is explained. Chapter Four provides a discussion of the findings based on the data set explained in Chapter Three. This Chapter outlines different graphical representations of the biplot and how each plot could be used for a specific task. Chapter Five provides a conclusion in conjunction with the problem stated in Chapter One. The Chapter also offers some recommendations for future research.

1.7 SUMMARY AND CONCLUSION

In the first Chapter the problem that the researchers are faced with, was revealed. This problem consists of a gap existing in the work environment between parties with different fields of knowledge. The background to this problem was unpacked and explained, including the introduction of the biplot as a bridge to this gap.

The concept behind a biplot was opened, arguing that visualising data is a common method in the working environment that could be extended to higher dimensions. The most attractive aspect of the biplot being that it continues to use pictures to aid interpretation, but that it removes the constraint that at most three-dimensions may be used.

This argument becomes the fundamental principle of the research assignment – using a complex data set that will later be revealed, as a means to show that the abovementioned is possible through using biplots. In the next Chapter, the underlying theory regarding how these plots are constructed accompanied by simple examples, will be presented.

CHAPTER 2
THE BIPLOT

2.1 INTRODUCTION

The biplot can be seen as a multivariate version of the well-known scatterplot (Greenacre, 2010:9). Biplots and scatterplots have very similar qualities including interpretability and dimension of the display surface. It can be viewed as an expansion of a normal scatterplot, but instead of only having two axes that are orthogonal, the biplot has many axes. The normal scatterplot has an x- and y-axis, while the biplot has numerous axes depending on the number of variables used. The biplot makes it easier to understand how variables in large dimensions interact (or are correlated) with each other. The reason for the name 'BI'-plot is because these plots can display the rows and columns of a matrix at the same time (N J le Roux *et al.*, 2003:43). This means that all of the observations, which are represented by the rows, and all of the variables, which are represented by the columns can be displayed simultaneously.

Biplots have evolved over time – the original biplot as introduced by Gabriel in 1971 contained the idea mentioned above. The covariance biplot was introduced by Barr and Afflek (1987) which was applicable in investment analysis. This method allowed the user to plot deviations for individual financial instruments and covariance's between different financial instruments. This methodology was continued by Barr, Kantor and Underhill (1987) where different weightings were used to refine the previous method. The theoretical basis underlying the biplot was further expanded by Gower and Hand (1996) yielding the modern version of the biplot as followed throughout. This Chapter will try and simplify the mathematical procedures in the construction of these plots with basic explanations and examples.

2.2 UNDERLYING THEORY

2.2.1 The construction

The scatterplot is limited to only two perpendicular axes while biplots have many axes that can take on any orientation. In the case of the scatterplot, the values of the variables can be read off directly from an orthogonal projection to the two axes, however, in the biplot case it is a more complex procedure. To read off values the same method is followed, but the values are now approximations of the true values, since we are displaying higher dimensions in a lower-dimensional plane.

Biplots largely make use of the geometric interpretation of scalar products, which provide a method by which projecting vectors onto each other using a relation. This relation is a function of their distances and the angles between the vectors.

The next fundamental idea is that the axes of a biplot can be calibrated such that the above mentioned calculation becomes easier. Since the scalar products of the points will be proportionate, the user can change the length of the vectors to (e.g. unity) help with the interpretaility. This is the process known as scaling, which is a common pre-processing technique when the input variables have different ranges. The fact that the axes can be calibrated leads to one of the essential ideas behind biplots: the actual values of a variable that can be predicted are not as important as the direction the variable is pointing in, since it can be projected to any value orthogonally onto that line to obtain a value for the specific variable.

Thus, the importance lies in being able to see how variables aline with respect to one another when specific values are not available. From this relation it can be concluded that if two biplots axes point in the same direction then variables will have high inter-variable correlation.

Any two axes that are oblique can represent a two-dimensional space that is familiar to most users. Using a method that is based on Singular Value Decomposition (SVD), the user can create a two-dimensional representation of a multidimensional input where the resulting principal axes provide a two-dimensional scatterplot of the above mentioned input (Gower, Gardner-Lubbe & le Roux, 2011:14). The axes can be translated to make them more visually appealing – e.g. translation to an aspect ratio of unity, or translation such that they pass through zero. These translations are possible since it is allowed to reflect and rotate diagrams that are based on inner products. Therefore, many different types of displays for the same set could be used – reflecting and rotating such that it has a suitable interpretability. It can be projected to any point onto an axis to obtain an approximate value – the biplot axes are calibrated like normal coordinate axes (Gower *et al.,* 2011:20). Thus, *n*-dimensional data can be represented on *n* axes which are not orthogonal but are called biplot axes. These *n* axes can now be used in the same manner as one would use the axes in a conventional scatterplot, as mentioned earlier.

Consider the following arbitrary data for Belgium (Be), Denmark (De), Germany (Ge) and Poland (Pol) in 2008 for three economic variables: X1 = purchasing power per capita (in Euros), X2 = GDP per capita (indexed at 100 for all three countries), X3 = inflation rate (percentage). A simple scatterplot of this set follows in Figure 2.1.

Table 2.1: Arbitrary data for three economic variables

Country	X1	X2	X3
Be	19200	115,2	4,5
De	20400	120,1	3,6
Ge	19500	115,6	2,8
Pol	18400	110,2	5,5

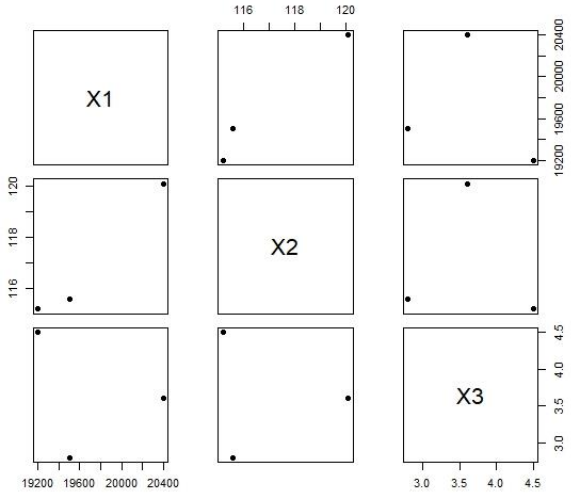

Figure 2.1: Scatterplot of the arbitrary set

Excluding the Poland variable (considering only the other three variables), there are three-dimensions in the set above (each variable represents a dimension in the Euclidean space) which makes it challenging to plot the variables on a two-dimensional space. Initially, scatterplots of each 2 variables could be plotted for all the observations but this would essentially mean that there would be $\binom{p}{2}$ such plots and in this case with $p = 3$, there would be 6 plots of the overall set such as in Figure 2.1. This is not feasible, since not every plot would contain important information and computing cost would increase considerably for a higher value of *p*. Rather than using a scatterplot for analysis of the data, a biplot of the set could be constructed. The biplot below in Figure 2.2, shows that six displays for this set is not required, but in fact that the set can be displayed on a single two-dimensional surface.

From Figure 2.2, various characteristics of the variables in the set are illustrated: Firstly, note that the biplot axes above represent the variables themselves. These axes are constructed by using the sample points and their respective variable values. Note that initially three variables are plotted on a two-dimensional surface, where normally a three-dimensional surface would be needed. When working with higher dimensional data, this will result in a loss of information because of the use of approximation to construct the projections. This however does not yet occur, since the dimension of the input data is still relatively low.

Note also that the observations are displayed in blue. These are the respective countries that are being analysed. The process of reading off values from a biplot is known as predication and is done in a similar way to a conventional scatterplot. As seen by the Denmark sample point, a perpendicular line is drawn from the sample point to the axes to read off the values.

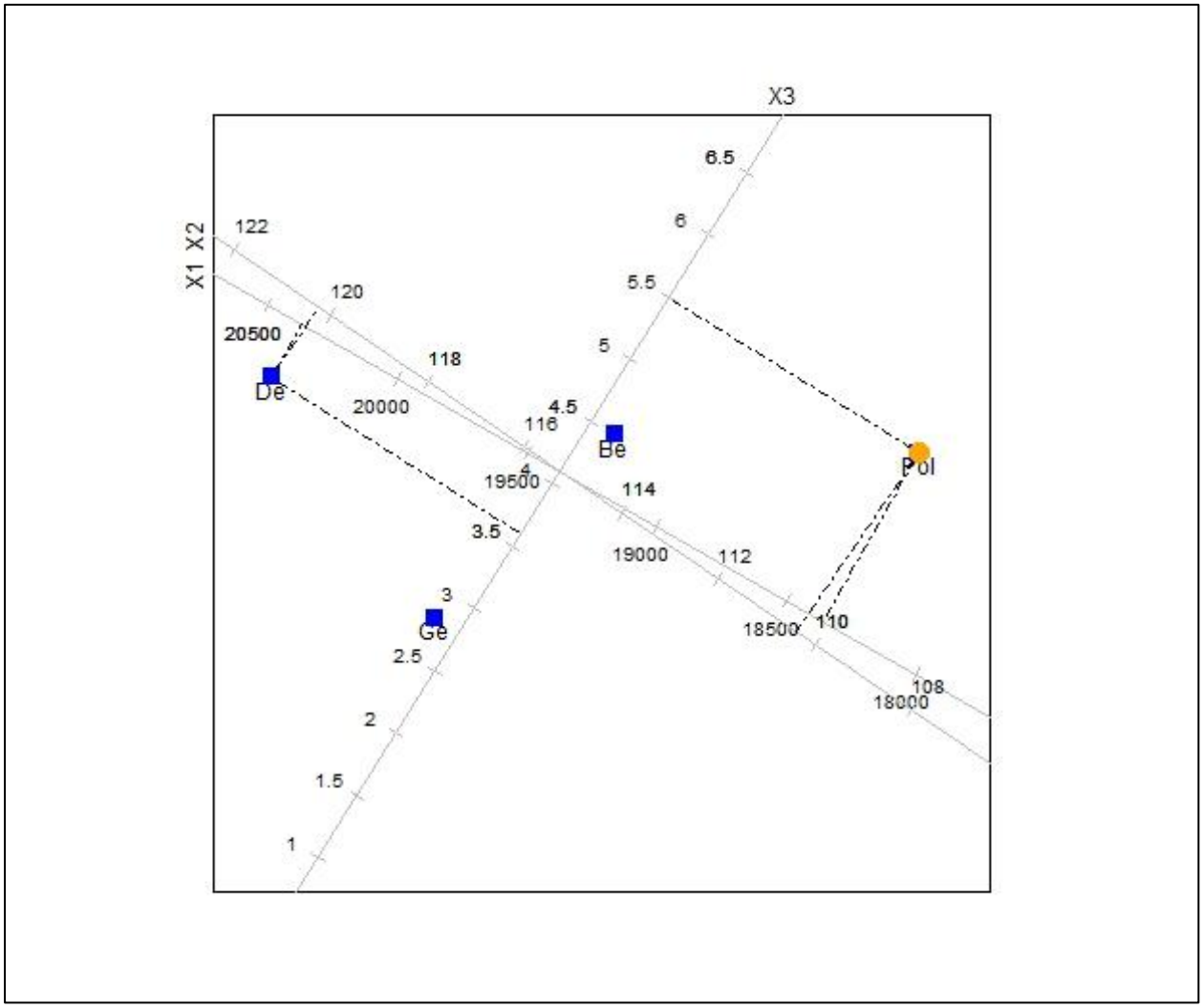

Figure 2.2: PCA biplot of the economic variables from Belgium, Denmark, Germany and Poland

When a biplot is created from a set and new information (samples) is added to the graph, a process called interpolation is used. Figure 2.2 was constructed from the economic variables of Belgium, Denmark and German. A fourth observation was then interpolated relating to Poland and its economic indicators. When a new sample point is interpolated, the point is placed onto the display that was constructed using the sample points in the original set. Now, with interpolation, the Poland sample point (orange) is added to the biplot, from which the approximate values can be read off from the axes.

A line is drawn from each sample point perpendicular (orthogonal) to each axis and a value is read off that axes through the use of interpolation. This is a very important characteristic of biplots when adding new variables (axes) to the plot since it enables the user to obtain approximations of values for variables that are not in the sample originally.

Next, the relations between different variables can be seen in the biplot display. When variables have positive covariance, they “point” in the same direction such as X1 and X2, with the values of the axes increasing in the same direction. However, when variables have negative covariances,

the axes “point” in different directions. When variables are uncorrelated they will be plotted perpendicular on each other, with the values of the axes increasing in any direction. This also illustrates another powerful tool of biplots since one can quickly identify how dependant different variables are by simply looking at the angles between the axes and the direction in which the axes values increase. The above plot is known as a PCA biplot whereby a brief description of the different types of biplots that will be used in the analysis section, follows:

2.2.2 PCA biplots

One of the most popular types of biplots are the Principal Component Analysis (PCA) biplots, which is an asymmetric multivariate plot. This can become rather sophisticated in terms of the mathematical features that underlie this analysis. Therefore, conceptualising the method without an in-depth mathematical explanation is suggested (only a short overview of the underlying mathematical structures regarding this method will be given shortly) seeing that this research assignment intends to simplify the complexity of understanding multidimensional factors, and since existing software does the calculations in fairly easy steps.

PCA is based on a dimension reduction technique applied to a set of correlated variables. It tries to retain as much information as possible when transforming the data set into its principal components, which are uncorrelated and ordered. The first couple of components are made to carry most of the variation that is found in the rest of the variables (Jolliffe, 2011:1). Hence, PCA biplots are primarily for describing variance in multidimensional data.

PCA as a statistical technique can be explained as a method to describe the covariances that occur within a data set. It finds a new set of dimensions, where this new set is orthogonal and ordered by variance (highest to lowest) of the data. This is done by calculating the eigenvectors (and corresponding eigenvalues) and then ranking these vectors from lowest to highest. The first m eigenvalues are chosen to form the new m-dimensional data set (Johnson & Wichern, 2008:432).

This was the intent of PCA as indirectly derived by Pearson (1901) and advanced by Hotelling (1933). In the article that Pearson wrote in 1901, it was shown that there exist expressions for the best fitting lines of a set of points in higher p-dimensional spaces. Hotelling further showed that there may exist a subset of independent variables regarding the original set that contains equivalent information in order to ease dimensionality problems in graphical representations. This is summarised as follows:

- Pearson’s approach to PCA: Pearson’s method makes use of the Pythagorean distance whereby the best fitting hyperplane is defined as: The smallest sum of squared Pythagorean distances between the points in a p-dimensional configuration and their orthogonal projections onto the hyperplane.

- Hotelling's approach to PCA: Hotelling's approach involves checking if there is a more fundamental set of independent variables that may have less values than the original data. This reduction in size of the set decreases the dimensionality of the data.

Research of using this method in the financial setting is plenty of sort. One example is where Stevens (1972) reduced 20 financial ratios to just six that explained 80% of the total variance (leverage, profitability, liquidity, activity, dividend policy, price and earnings). Another is in the case of bankruptcy analysis where Tudor (2009) reduced a set of 16 variables of Romanian listed companies to a new set containing only 3 variables that explain 96.72% of the initial variance. These examples and many more in other settings emphasises the usefulness of using PCA on a high dimensional data set.

PCA biplots are much easier to explain through an example, which will prevent the user from getting trapped in the theory underlying the method. Three-dimensional PCA biplots are available but this research will only consider two-dimensional cases – purely since the objective is that the reader would be able to fully interpret the displays given. A short summary regarding the fundamental mathematical theory needed to construct this type of plot, as explained by Gower *et al*. (2011), is as follows:

Using the SVD as proceeded by Gabriel (1971), there exists an $\mathbf{X} : n \times p$ such that

$$\mathbf{X} : n \times p = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}' \tag{2.1}$$

where, assuming that $n \geq p$, $\mathbf{U}$ is an $n \times n$ orthogonal matrix with columns known as the left singular vectors of $\mathbf{X}$, the matrix $\mathbf{V}$ is a $p \times p$ orthogonal matrix with columns known as the right singular vectors of $\mathbf{X}$, while the matrix $\boldsymbol{\Sigma}$ is of the form

$$\boldsymbol{\Sigma} : n \times p = \begin{matrix} k \\ n-k \end{matrix} \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{2.2}$$
$$\begin{matrix} k & p-k \end{matrix}$$

In the above equation, k denotes the rank of $\mathbf{X}$ while $\boldsymbol{\Sigma}$ is a $k \times k$ diagonal matrix with diagonal elements the nonzero singular values of $\mathbf{X}$, assumed to be presented in nonincreasing order. An r-dimensional approximation of $\mathbf{X}$ is given by $\widehat{\mathbf{X}}_{[r]} = \mathbf{U}\boldsymbol{\Sigma}_{[r]}\mathbf{V}'$ where $\boldsymbol{\Sigma}_{[r]}$ replaces the $p - r$ smallest diagonal values of $\boldsymbol{\Sigma}$ by zero. PCA uses a least-squares criterion as the basis of approximation. To be precise, the sum of squares of the differences between corresponding members of $\mathbf{X}$ and $\widehat{\mathbf{X}}_{[r]}$ is minimised. The r-dimensional Eckart-Young approximation given by $\widehat{\mathbf{X}}_{[r]} = \mathbf{XVJV}'$ with $\mathbf{J} : p \times p = \begin{bmatrix} \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$ minimizes the squared error loss:

$$\left\|\mathbf{X} - \widehat{\mathbf{X}}\right\|^2 = \text{tr}\{(\mathbf{X} - \widehat{\mathbf{X}})\,(\mathbf{X} - \widehat{\mathbf{X}})'\} \tag{2.3}$$

Then the coordinates of the *r*-dimensional approximation of the centred **X** are given by the first *r* columns of **UΣJ** and the directions of the axes by **VJ.** The coordinates given by the *n* rows of **UΣJ** are plotted to obtain the *r*-dimensional plot and the *p* rows of **VJ** to give the directions of the axes. The columns of **V** are often termed the principal components, or the principal component loadings, and **XV** interpreted as new latent variables. Note that these latent variables are uncorrelated.

Now, Consider the following arbitrary data set consisting of 12 observations. The set is split into two types of variables: The variables A, B, D and E are sediment variables which are counts. The variables POL and TEMP are continuous measures for pollution and temperature respectively. The last variable, SED, is a categorical variable classifying the substrate as mainly C (=clay/silt), S (=sand).

Table 2.2: Sample of 12 observations from 7 variables, classified in the last column

Variable	a	b	d	e	Pol	Temp	Sed
S1	0	2	14	2	4,8	3,5	S
S2	26	4	11	0	2,8	2,5	C
S3	0	10	8	0	5,4	2,7	C
S4	0	0	3	0	8,2	2,9	S
S5	13	5	10	7	3,9	3,1	C
S6	31	21	16	5	2,6	3,5	S
S7	9	6	11	2	4,6	2,9	S
S8	2	0	0	1	5,1	3,3	C
S9	17	7	14	6	3,9	3,4	C
S10	0	5	9	0	10,0	3,0	S
S11	0	8	6	7	6,5	3,3	C
S12	14	11	15	0	3,8	3,1	S

A PCA biplot of this data set is constructed, where the sample points are displayed as purple squares for class S and yellow circles for class C. The mean values for the specified classes are also displayed in the middle of the plot, from which the alpha bags are constructed.

The two alpha bags are constructed for the two classes, where the purple bag refers to class S observations, and the yellow to class C. Alpha bags are visual aids that are defined as bags that enclose a certain percentage of the sample points closest to the origin. Different alpha levels can be chosen when constructing the plot. The usefulness of alpha bags increase as the number of observations increases since a more accurate proportion is displayed. In the plot above a 95% alpha bag is illustrated indicating where 95% of the data will lie. Figure 2.3 was constructed using all the mathematical procedures as explained above.

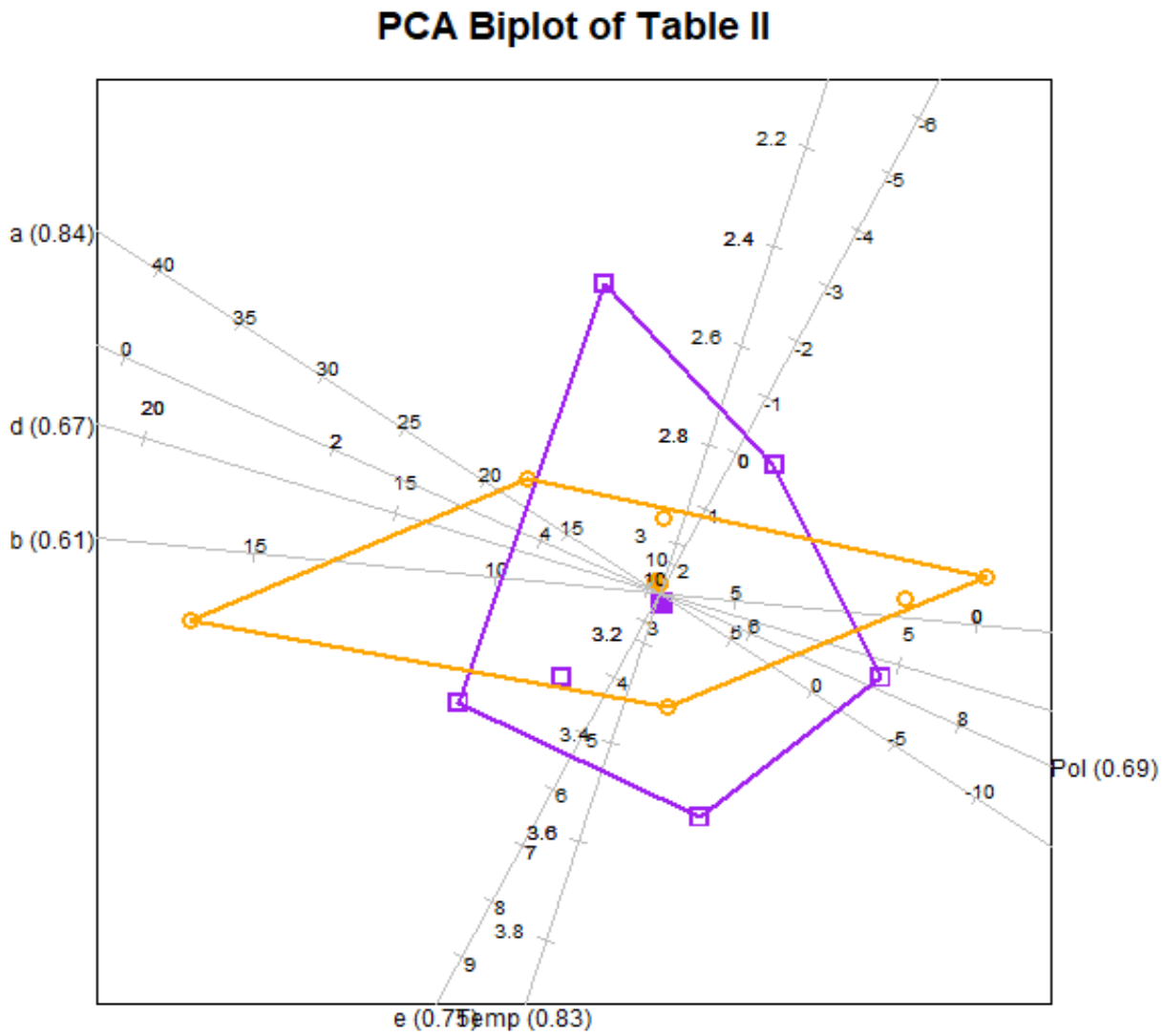

Figure 2.3: PCA biplot of the data in Table 2.2 with 95% alpha bags

2.2.3 CVA biplots

The second type of biplot that will be explained is the Canonical Variate Analysis (CVA) Biplot, which focusses on grouping observations into classes and then studying the between-class and within-class variation. This entails transforming the observed variables into canonical variables.

Canonical variate analysis involves quantifying the relationships between two sets of variables. It focusses on the correlation of a linear combination of the variables in the one set related to the other set. It determines the pair of linear combinations having the largest correlation. The pairs of linear combinations are called the canonical variables and their correlations are called the canonical correlations (Johnson & Wichern, 2008:539).

The differences between PCA and CVA biplots are highlighted as follows:

- The data does not have to be scaled when working with CVA, while scaling is an important tool in the PCA case.
- The CVA components have a value between 0 and 1, while the PCA components are not bounded.
- There is a higher degree of separation between group means in CVA biplots than in PCA biplots.
- PCA biplot scaffolding axes are not determined by group means; CVA group means determine the scaffolding axes.

Considering again the data in Table 2.2, the following CVA biplot is constructed in Figure 2.4. A 95% alpha bag was added to the plot to indicate a clear separation between the classes and their means. Note that the data is classified into two different class of “Sediment” in column seven of Table 2.2. The CVA biplot splits the data according to this column and displays two indexed groups on the plot. This gives the user a demonstration of the form of the two different classes at a 95% confidence interval. Thus, the display in Figure 2.4 has more usefulness than just reading off values for isolated sample points from the axes. At the centre of each of the alpha bags, a group mean is shown, which is simply the average of all the points defined in a class (in this case S vs C). When the mean values are read off the axes, by drawing perpendicular lines towards them, it can be seen which variables contribute most to the separation and which variables are similar, independent of class allocation.

Many other types of biplots exist and can be used for specific cases. Some of these include:

- Log-ratio biplots: When working with data that does not have the same basis scale, we have to scale the data (in most cases to unit variance) to be able to compare the different variables.
- Regression biplots: In this case the axes can be calibrated so that prediction values from regressions can be read off.
- Generalised Linear Model biplots: These plots generalise linear regressions to include different relationships between the conditional mean of the response variable.
- Multidimensional Scaling biplots: Multidimensional scaling is a method that represents a set of objects as a set of points in a map, usually two-dimensional, based on their given inter-point distances. The biplot displays these points.
- Reduced-dimension biplots: These plots rely on approximating a matrix of high dimensionality by a matrix of lower dimensionality.

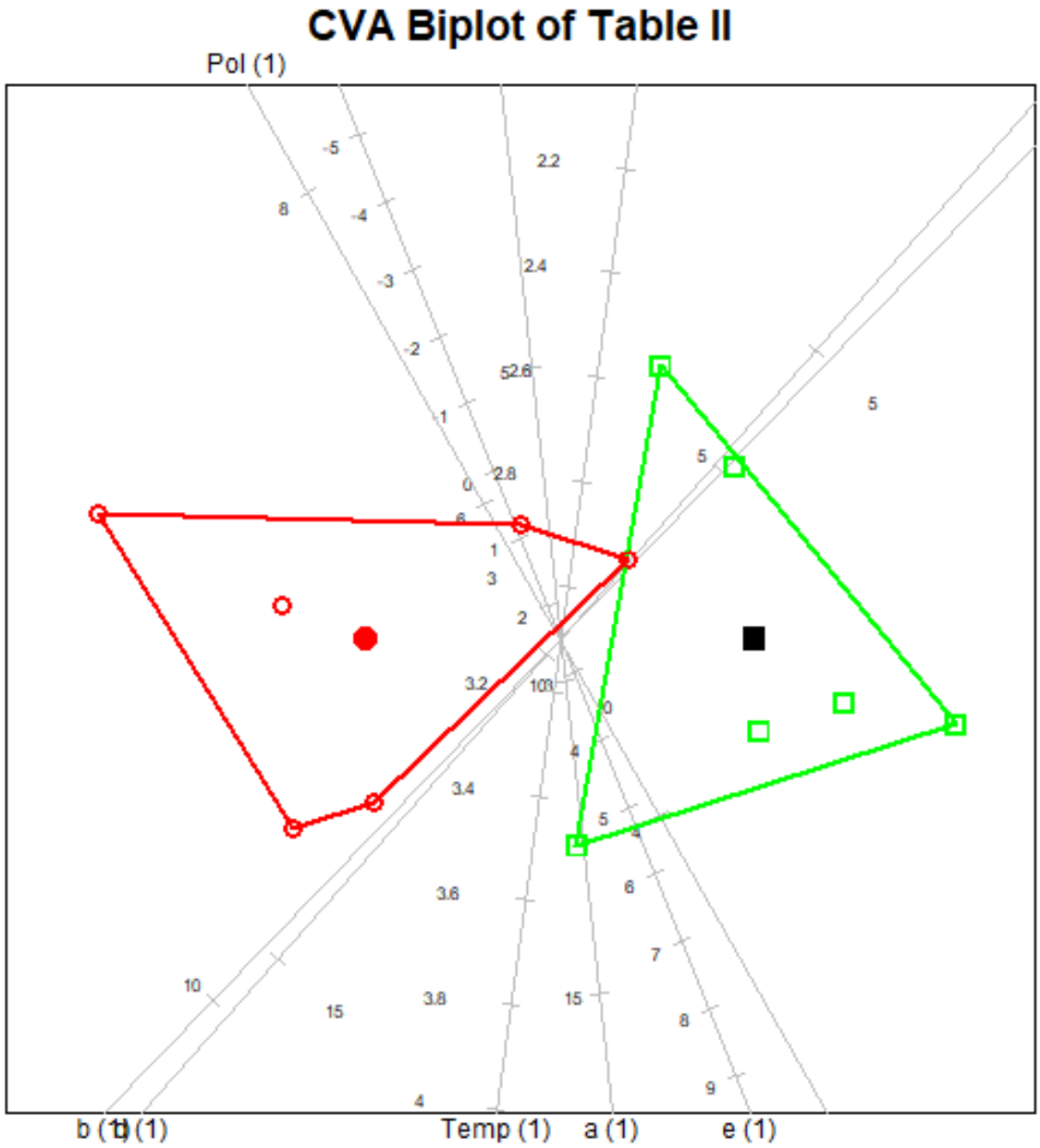

Figure 2.4: CVA biplot of the data in Table 2.2 with 95% alpha bags

2.3 LARGE DATA SETS

When working with larger data sets, biplots can become condensed and more challenging to interpret. When the number of variables (which is the number of biplot axes) becomes larger, it may get to a point where the biplot axes need to be turned off actively to help with interpretation. Alternatively, the sample points can be omitted by replacing condensed zones of sample points with a boundary.

A segment of sample points can be enclosed within a biplot to help with interpretation. The sets of observations can be enclosed to display the shape of a certain type of observation through:

- Spanning ellipse: the smallest ellipse that covers all objects in the display.
- Concentration ellipse: an ellipse that encloses a certain percentage of the objects (based on an interval dependant on the mean and variance of the sample)
- Convex hull: the smallest convex set that encloses all points in the set. Improves on ellipse enclosing large areas with no observations.
- Bag plot: A bag is constructed containing 50% if data points; a fence is constructed by inflating the bag by factor 3 (subjective), where observations outside this fence are labelled as extreme.
- Bivariate Density plots: A colour based display that is based on concentration of observations. This is good for multimodal clusters of points.

A biplot, with the above characteristics, takes the input data and identifies the fundamental structures of the data in a way that it can be easily interpreted. Financial data in many cases can be represented in matrix form – the rows representing dates or different companies, the columns representing the values of different financial parameters. This will become more evident in the next Chapter.

2.4 SUMMARY

In the second Chapter of this research assignment the idea and fundamental theory behind the complex construction of biplots was explained. This was followed by a simple example where a biplot is constructed to aid the idea of showing how to interpret this plot – from reading the relations between the variables to predicting the values of isolated sample points.

The following section then introduced the first kind of biplot, namely the PCA biplot, which uses PCA in the construction of the axes. The underlying theory was explained followed by an example of how a PCA biplot is constructed from a simple data set. A similar process was followed for the second kind of plot, namely the CVA biplot, where the main focus turned to classification. This section also contained an example of a CVA biplot from the same data set as used in the first example.

Chapter 2 was concluded with a section on large data sets and how to handle them when working in the biplot space. Various techniques were revealed in how to manage the sample points so that they remain interpretable. In the following Chapter, the data set used in the analysis Chapter and how it will be cleaned is revealed.

CHAPTER 3
DATA

3.1 INTRODUCTION

The data set chosen for testing the effectiveness of using biplots in the finance space consists of Polish companies that defaulted between the years 2000−2012 with some operating companies being evaluated between 2007−2013 (Zieba, Tomczak, & Tomczak, 2016). This is a secondary data set that is publicly available from the Machine Learning Repository hosted by the Centre for Machine Learning and Intelligent Systems at the University of California, Irvine. Therefore, justification of the project by other researchers are encouraged. In this Chapter, the data cleaning process will be explained in detail.

It is important to note that the global economic crisis was well underway during the period between 2008 and 2009. This led to a substantial decrease in Gross Domestic Product of several European countries. Surprisingly, according to (Poland, 2012:9), Poland did not follow these trends completely since they experienced an increase in averaging growth rates from 2004 throughout the crisis period till 2011. However, this increase did not apply to all companies. This is evident when examining the success rate of company survivability, since capacity utilisation decreased over these periods (Drozdowicz-Biec, 2010). This could potentially be the root of some companies failing and therefore resulting in default status.

The purpose of using this data set was to analyse the default status of a company based on its current financial position using subset selection to identify the important variables that should be used and ultimately visualising these significant variables on specifically selected variations of biplots. This is a complex set and thereby relates to the simplification of understanding multidimensional financial data with the use of these plots. Generally, the term “current financial position” is mostly evaluated in terms of financial indicators or better known as ratios.

The set contains 5910 (=n) observations and 65 (=p) variables of which the first 64 are the predictors or attributes indicating the grade of the company in terms of financial ratios, which can be found in Appendix A, while the 65th observation is the response (a dummy variable) containing the values “Yes” and “No” describing if a company will default or not in the following year. In the set, 410 companies defaulted out of the 5910 observations in the sample space. Taking this imbalance into account (only 7% of the overall sample space consists of companies that went bankrupt), and the fact that there are 4666 missing values in the predictor space, imputation techniques should be performed to clean this set.

In simple terms as explained by Donders *et al*. (2006), “Imputation techniques are based on the idea that any subject in a study sample can be replaced by a new randomly chosen subject from

the same source population. Imputation of missing data on a variable is replacing that missing by a value that is drawn from an estimate of the distribution of this variable".

Now, with cleaning the data and imputing the missing values with the K-nearest neighbors (kNN) technique, the remaining set contained some high correlated variables. Due to multi-collinearity problems, these highly correlated variables should be removed. After this process was performed, some extreme outliers were detected. It is not unreasonable to have these number of outliers since this is perhaps an indication of the complexity of this set which relates to the amount of missing values that were imputed. It could be that some of the information was incorrectly observed, but techniques exist to function around (treat) this problem. It was decided to impute the outliers with a so-called winsorization algorithm in the DescTools R-package. This method shrinks outlying observations to the border of the main part of the data, to reduce the effect of spurious outliers without deleting any observation. Therefore, this method censors the data whereby it replaces extreme values by certain percentiles of the data. Keep in mind that the winsorized mean is not the same as the truncated mean (the mean obtained when outliers are deleted in the data).

Note that in general it is not advisable to change or remove outliers since these observations sometimes provide the most insight of an observed set. The reason for the treatment of outliers in this case is that the different types of biplots examined later would not perform accordingly due to the outliers included, as they explain no significant outcome of the data. The biplot would try to include the outlier points, but the scale would distort since the plot would "zoom-out" as far as possible to include these points. Since these values do not contribute to the analysis in Chapter 4, it would be meaningless to include them.

Finally, also note that these variables are financial ratios as indicated above so their features can vary extremely. For example, some ratios can take on any number and other can take on only positive or negative numbers. This will lead to problems when the set is applied to various methods described in the following Chapter. To overcome this problem, it is advisable to standardise or scale the data, since variables with higher or lower ranges can have different impact in the analysis Chapter.

3.2 VARIABLE SELECTION

Financial data in the real world consists of a massive number of variables, whereby the interpretation of a model regarding high dimensional data can be challenging. This can produce calculational drawbacks and high variance which influences the model building and visualisation procedure. To overcome this problem a popular technique, that has received substantial attention from statisticians over the past century, called variable selection or better known as feature

selection was used on the set (after imputation and standardisation) to exclude irrelevant variables by means of appropriate algorithms.

Many businesses use financial ratios to obtain a better picture of their survivability in competing sectors. They often try to portray themselves being in a better position than their respective competitors by increasing the amount of information showed by their financial statements. This creates a high dimensionality problem since many of the ratios will mostly contain the same inputs or inputs that have a high correlation with one another.

Some machine learning, and visualisation techniques does not always function well in high dimensional space. Variable selection tends to correct this cost. One of the first methods was presented by Efroymson (1960) and is still widely used today in many settings. These methods include forward selection, backward elimination and stepwise regression.

Draper and Smith (1966) explains each method in detail whereby Efroymson (1960) discusses how a large data set can be deduced to contain only a subset of important variables, those that are significant according to some criterion, by making use of variable elimination or addition. Invariably, this idea brought forward new algorithms in statistics and computer science, since there is a high cost of calculating the significance of each feature in a large set.

According to George (2000), a wide-ranging analysis of the methods used before 1990 is explained in Miller (1990). Currently, there are many modernised techniques that can be used such as Best-Subset Selection, Shrinkage Methods like the Least Absolute Shrinkage and Selection Operator (LASSO), the Least Angle Regression, etc. all which can be found in the literature of (Hastie, Tibshirani & Friedman, 2001).

The purpose of a selection feature is to reduce the data to a selected number of variables that will minimise the cost of calculations, i.e. making the calculations easier and more understandable without loss of generality. In doing so, variable selection also clears up some of the noise and degrees of freedom in the model (Hastie, *et al*., 2001:38).

This idea is closely related to Occam's razor or better known as the principle of parsimony stating that the simplest model that fits the data is the best model (Sober, 1981:145), which is the thought process that will be further used in the coming analysis. This does not only free up calculation cost, but it also excludes all unnecessary time spent on fitting different types of models with few modifications. Generally, it may be possible that there are several useful variables and not typically a one-best-model to use (Cox & Snell, 1989).

In this Chapter, stepwise methods are discussed and used in greater detail to obtain the optimal number of variables to use in the analysis without losing too much approximation accuracy.

Stepwise methods include forward and backward procedures that will be used to detect a subset of the predictor space that is mostly related to the response. This method is easily described by the algorithms as set out by James, Witten, Hastie and Tibshirani (2013), after the data has been split into a training set and a test set (Appendix B and C).

The forward stepwise selection assumes no variables in the beginning stage of a model i.e. it starts at the intercept and adds to this model the variables that is the most significant (lowest *p*-value or smallest RSS (residual sum of squares)) until all the significant variables are included in the model. This ensures that the variable describing the fit in the most significant manner is added to the model.

The backward stepwise selection is a reverse of the forward method whereby it starts with a full model and then removes the least significant (highest *p*-value) variable at each step. Note that forward selection can be used in any dimensionality and backward selection can only be used when $n > p$. However, in the Polish set where $n > p$ it is irrelevant which method is used in this regard.

According to the algorithms, both methods make use of the RSS as a measure to find the best model after each selection procedure based on the training data. The problem with this is that the RSS is closely related to the training error. Since the training error is not of interest but rather the test error, a selection criterion for choosing an appropriate model is needed. This error problem is depicted in Figure 3.1:

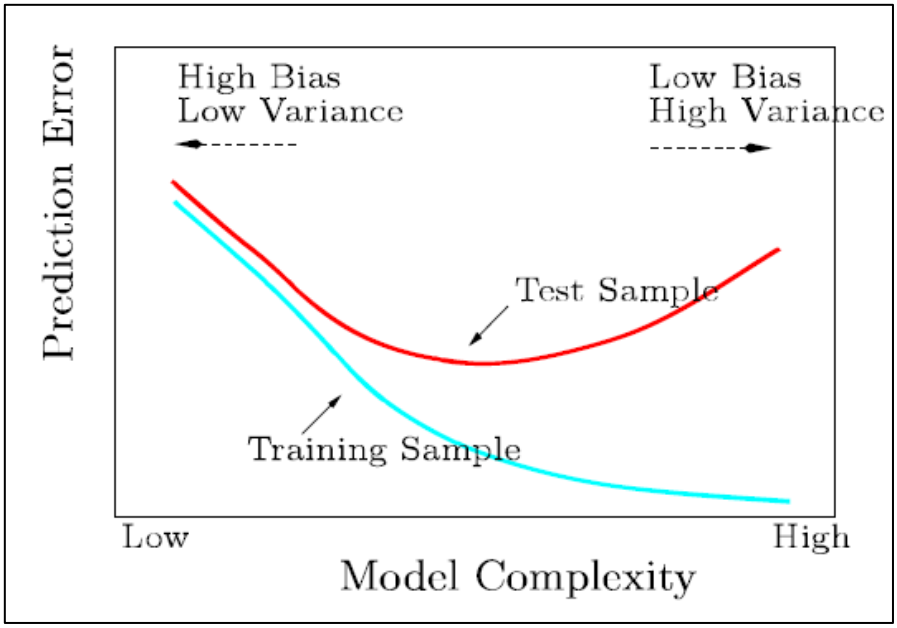

Figure 3.1: Bias-Variance Trade-Off

Source: James, *et al*., 2013

This illustrates a well know principle in statistics called the bias-variance trade-off indicating that as the model complexity increases (as more variables are added to the model) the training error will decrease but not necessarily the test error. Therefore, the training error should not be used to estimate the test error. Fortunately, there are several techniques to overcome this, which is already incorporated in step 3 of the algorithms, such as rather using the adjusted R^2 statistic ($R^2_{adjusted}$), the Akaike information criterion (AIC) and Bayesian information criterion (BIC) as estimation criterions.

Originating from Kullback and Leibler (1951) and improved further, these methods are the most widely used criterion-based procedures. Each method indirectly corrects the training error, but also penalises the variables in different ways. The focus now turns to comparing these methods and finally identifying a single model with the reduced number of variables. A large value for the $R^2_{adjusted}$ and a low value for the AIC and BIC will indicate a model with a low test error.

According to Figure 3.2 and Figure 3.3, it is evident that each method differs regarding the best possible model chosen. Summarising the output as follows to indicate the most important variables as chosen by each method and its corresponding statistic:

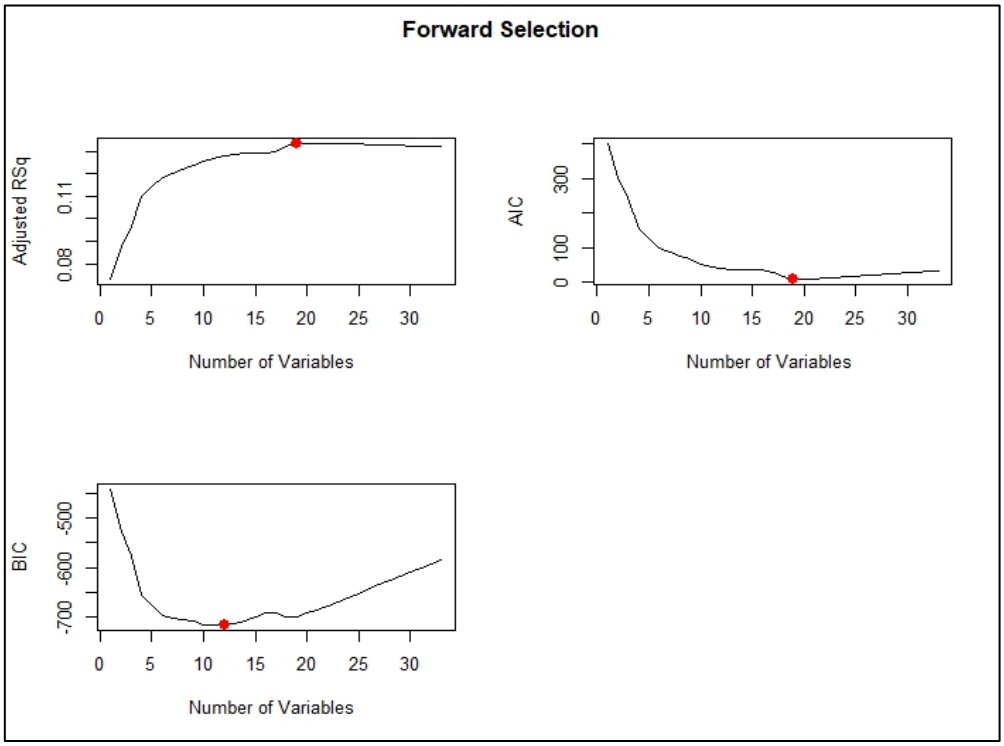

Figure 3.2: Forward Selection

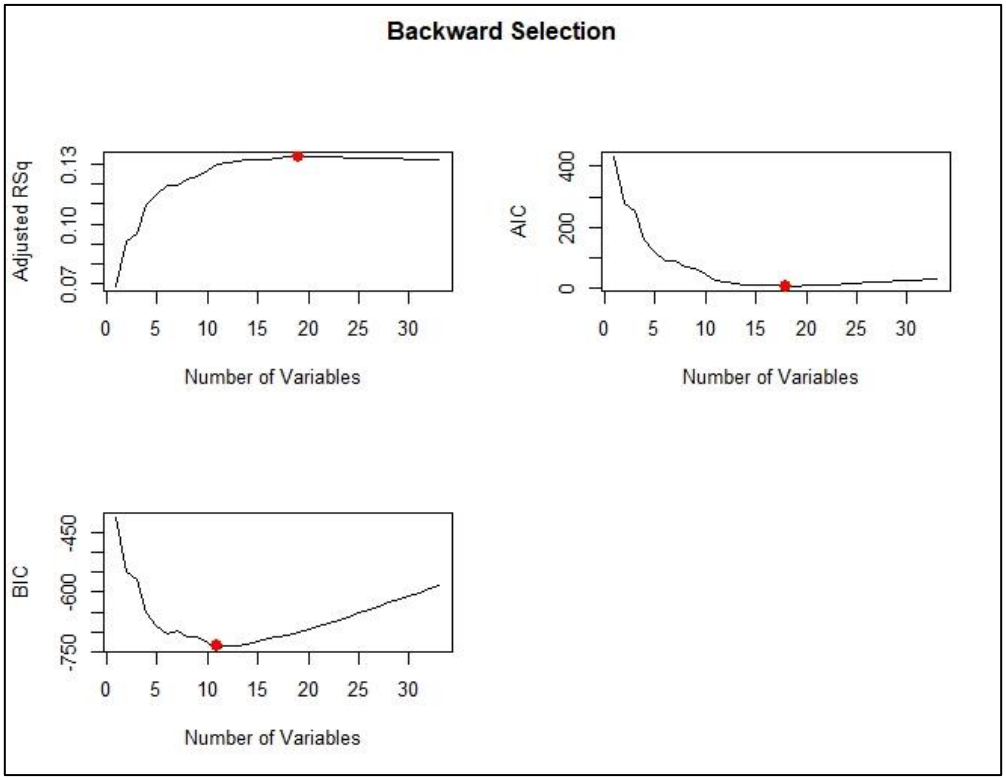

Figure 3.3: Backward Selection

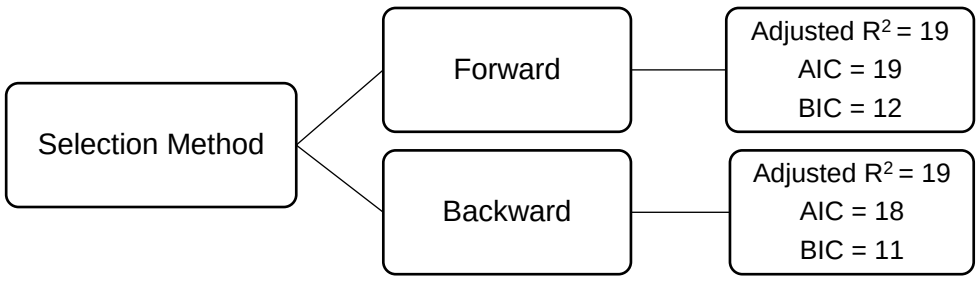

Figure 3.4: The selection process

Now, to choose the best model (i.e. what variables to include) the principle of parsimony was used together with the BIC as a selection criterion. The selection method that included the lowest BIC is the Backward Stepwise approach according to Figure 3.4. The important variables that should be included according to this procedure is as follows:

Table 3.1: Important variables identified by stepwise selection

Important Variables	Ratio
3	Working capital/total assets
16	(gross profit + depreciation)/total liabilities
29	Logarithm of total assets
32	(current liabilities*365)/cost of products sold
34	Operating expenses/total liabilities
36	Total sales/total assets
39	Profit on sales/sales
42	Profit on operating activities/sales
43	Rotation receivables + inventory turnover in days
48	EBITDA/total assets
57	(current assets – inventory – short-term liabilities)/(sales – gross profit – depreciation)

The results above are intuitive and is backed up by the model selection criterion since in practice each of these ratios are regarded as a fundamental indicator in evaluating whether a company has gone bankrupt or not. The variables that were disregarded are not necessarily of less importance but can be omitted in future analysis since it will reduce the difficulty of the model without loss of approximation. It may seem that the reduction is aggressive but the fact that the different ratios in many cases are very alike was based on where they used many similar input values. This leads to many of the ratios behaving in the same way when one input has changed.

A simple example is the use of the input “Total assets” in the financial ratios of the original 64 variable data set. It is directly used in many of the ratios, therefore when the selection is applied, these ratios will not be deemed to add value to the analysis. The end result is that there’s an 83% decrease in the number of variables. These remaining variables and observations will be used in the construction process of various biplots in the following Chapter.

3.3 SUMMARY

In this Chapter the Polish default data set and its characteristics were revealed. The central property of the data set being that an indicator variable existed, with a value of 1 indicating if a company has defaulted and 0 indicating if a company has not defaulted. The data pre-processing procedures were then introduced and justified. These processes included outlier detection, treatment of missing values and the variable selection step.

The methods used in each of these steps were explained alongside the rationale why those specific methods were used. In the end of the Chapter the variables that were selected are shown, amounting to a total of 11 ratios. These 11 ratios will be used as the input variables in the coming analysis.

CHAPTER 4
ANALYSIS

4.1 INTRODUCTION

Defaulting is a major risk that modern firms are exposed to – In many cases, firms believe that they are too big to fail. What has been observed in the financial world, specifically since the turn of the century, is that indeed no company is too big to fail.

In the modern world there are many measures available to attempt and quantify risk in general, and default risk specifically. These measures can vary from Value at Risk (VaR) and its derivatives to Expected Shortfall and Probability of Default. In a more conventional sense, financial ratios have been commonly used as indicators of the health concerning a specific company. Though financial ratios are firmly accounting based, they do carry value when working in the statistical world as it is possible to compare and track the different success levels between companies through time, and between the companies themselves.

The primary objective of this Chapter is to show that by visualising the multidimensional set as described in Chapter 3 may be an effective analytical technique. The emphasis on representing the set in a visual manner is due to the lack of material that exist to display multivariate financial data. People are prone to believe that at most three-dimensions can be displayed but this is not entirely true – through using approximation it can be shown that one can display as many dimensions as desired.

The core principal of biplots is not to accurately be able to read off values from the axes, but rather to interpret the shape and form of the data to obtain different relationships between the variables in the set. The idea being that biplots can be a useful tool through which “unhealthy” companies can be identified by comparing the shapes of various companies that have defaulted and not-defaulted.

For the displays that follow in this Chapter, the R package UBbipl was used to generate the biplots as thoroughly described in Gower *et al*. (2011). The first PCA Biplot in Figure 4.1 contains 5910 observations with the 11 selected variables in Chapter 3.

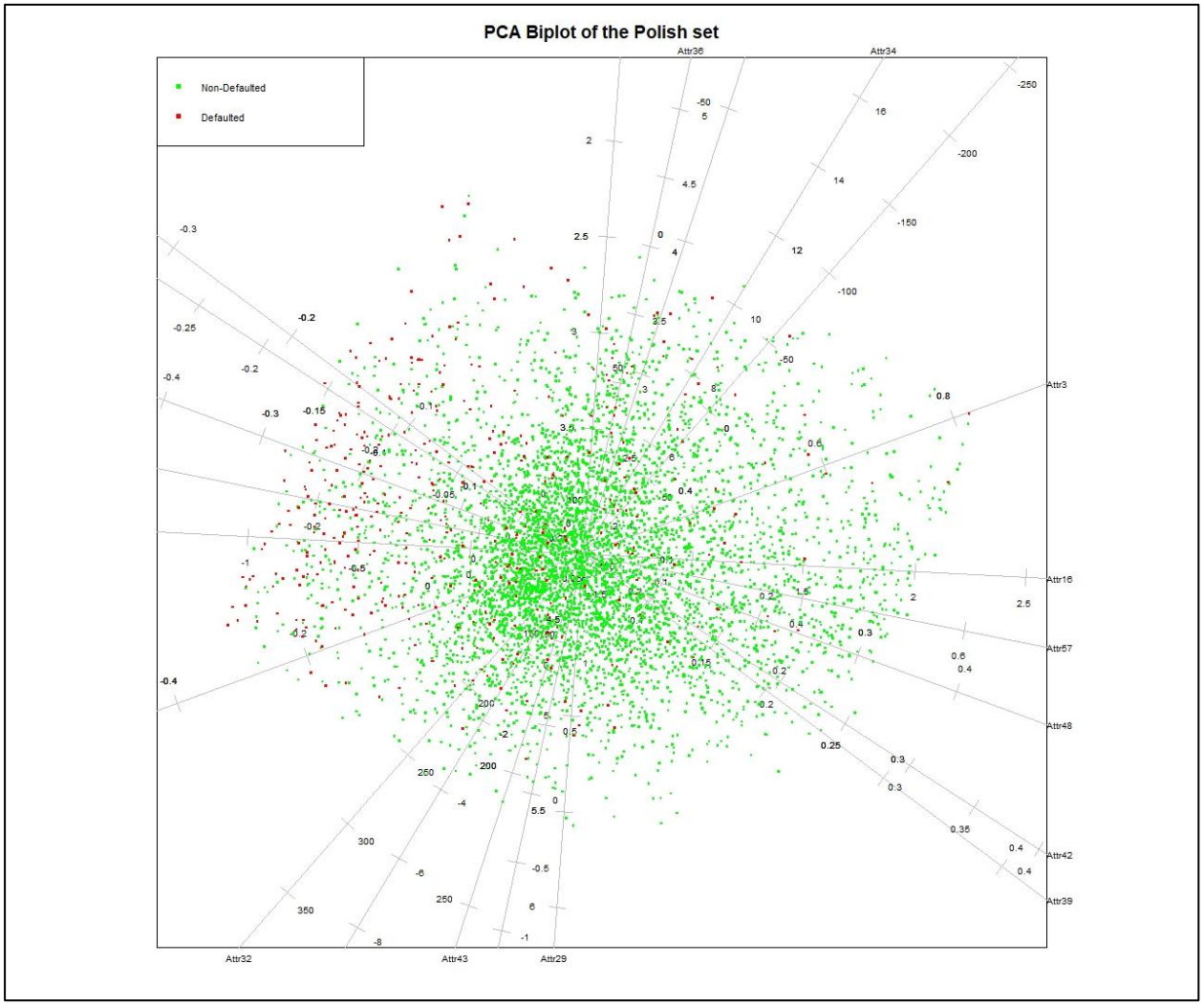

Figure 4.1: PCA biplot of the Polish set

Figure 4.1 is overwhelmed by the non-default sample values that is indicated as a green sample point and correspondingly, the red points are those sample values from the defaulted companies. To obtain a clear picture of how the sample points are positioned, a bagplot containing a 90% percentile were drawn on the biplot obtaining Figure 4.2. These graphs indicate the necessity of working with an equal number of sample points to compare the variables in a more frequent fashion since the non-defaulted points overpowers the graph when compared to the defaulted points. Therefore, it was decided to take a random sample of 410 observations from the 5500 non-defaulted values to obtain a clear comparison between the defaulted and non-defaulted samples in this context. This could possibly ease the analysis when ratios are being compared from defaulted and non-defaulted companies and also, the display of the plot would most probably increase in value.

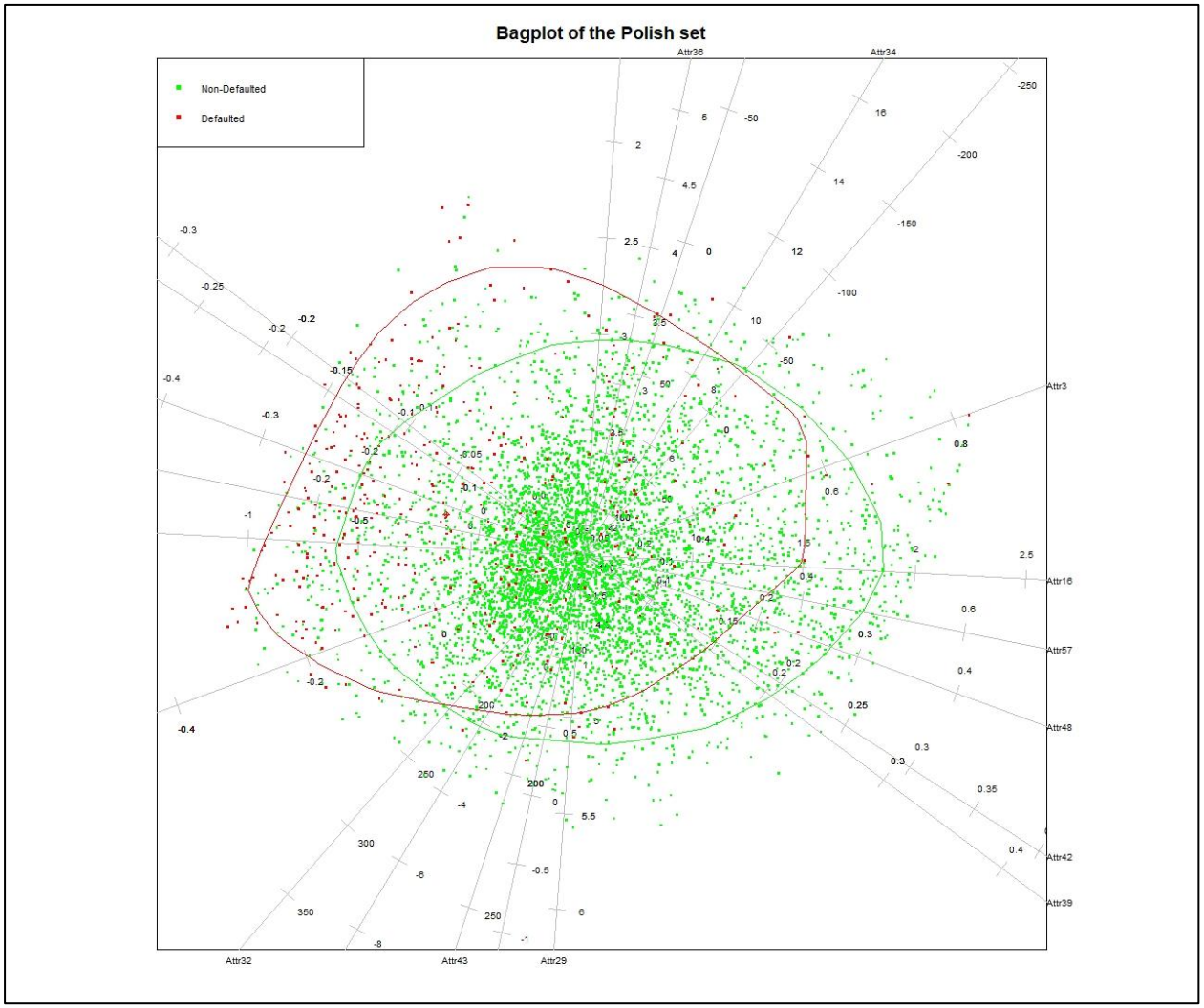

Figure 4.2: Bagplot of the Polish set

4.2 SAMPLING

The plot in Figure 4.3 was created by using the n = 820 by p = 11 input matrix. Of these 820 values, there were 410 observations from companies that defaulted, and 410 observations from companies that have not defaulted. It can be observed that the 11 input variables that are being plotted, are each represented by a biplot axis. In this case the data has been scaled in order to be able to compare the variables on one given plane. Scaling is a critical component of data preparation – in a case where scaling has not been applied to the data, it is very challenging to see the relations between the variables.

One can also observe that no two axes are plotted on top of each other in the display. This is the result of the variable selection process in Chapter 3, leading to selecting variables that are least correlated in order to obtain accurate approximations for the axes. Variables that behaved in the same manner were reduced down to one, in order to minimise the number of variables needed to explain or represent the variation in the data.

Using the biplot above, one can observe the relations that the different variables have with each other. Two axes that are perpendicular indicate no correlation, while axes that are or have similar orientation either have a strong positive or negative correlation. If the axes increment in the same direction then there is a case of positive correlation, while axes incrementing in opposite directions indicate a negative correlation.

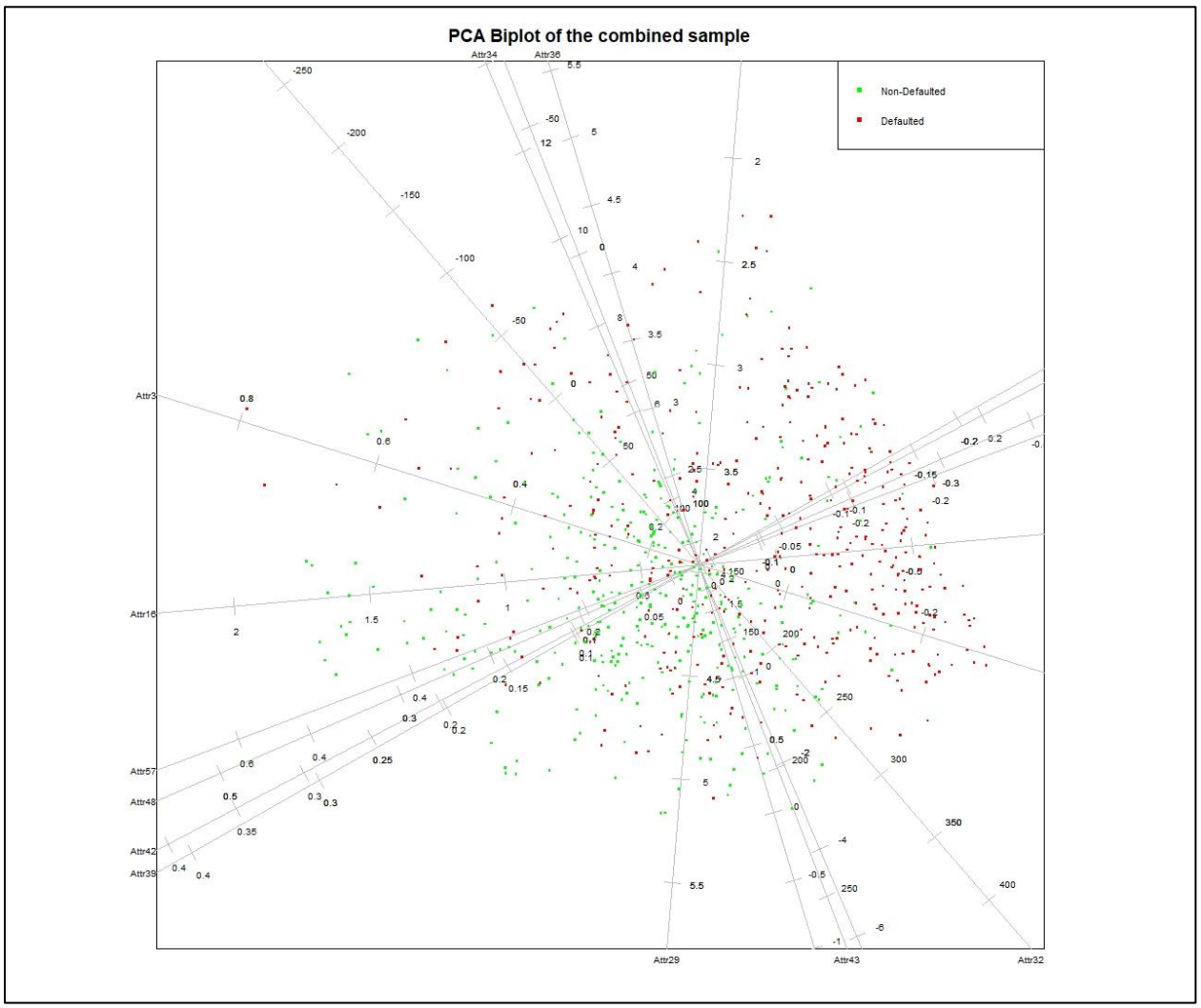

Figure 4.3: PCA biplot of the combined sample consisting of 11 variables and 820 observations; 410 from defaulted and 410 from non-defaulted, respectively

One can represent proportions of the data through various tools within biplots – the two used in Figure 4.4 being bagplots and spanning ellipse. This is again a useful tool in interpreting where different classes of data points lie, and how much variation lies within each subset.

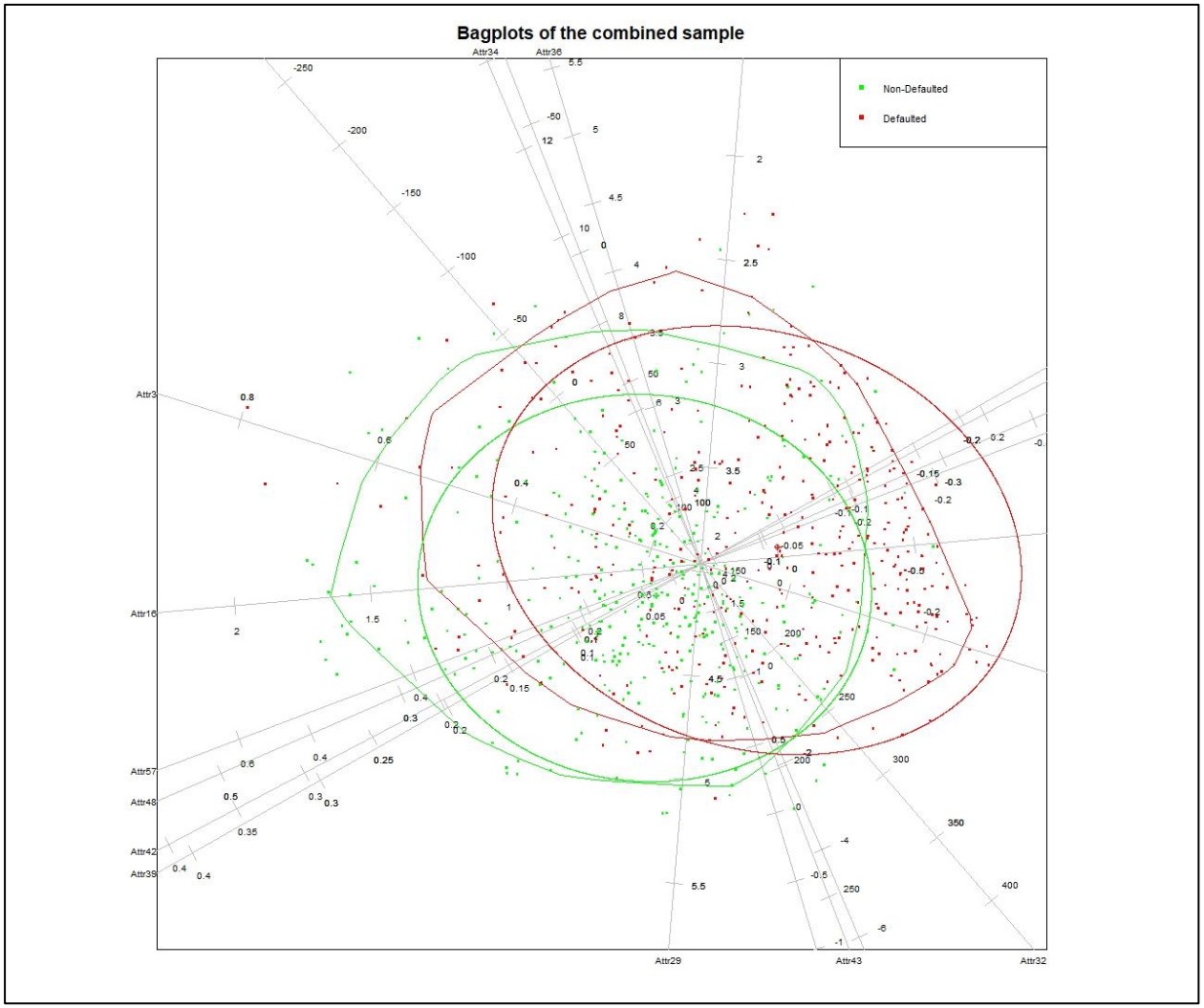

Figure 4.4: PCA biplot of the combined sample consisting of 11 variables and 820 observations, including 90% alpha bags and ellipses for defaulted and non-defaulted companies, respectively

In the case of Figure 4.4, the data has been positioned according to the default or non-default status of the observations. The red ellipse and bagplot indicate lines such that 90% of the default data is enclosed within the shapes, where the green refers to the non-defaulted case. It can be observed that the non-defaulted observations have a smaller variance, since the ellipse and bagplot surrounding is smaller in comparison with the defaulted case. This indicates that many of the companies that have non-defaulted status, have similar financial ratio values.

The event of a company defaulting is extreme by nature; thus, the more variate nature of the defaulted observations intuitively makes sense. For a company to go bankrupt, it may be expected that some of the ratios would be significantly different from the “norm”. This is the case in the above figure where many red observations are detected on the edges of the display, representing extreme ratios in many directions for companies that did not survive.

It should also be noted that there are parts of the ellipses where not many observations are plotted, being a result of the more rigid form of the bagplot. The ellipsoids are plotted around a centroid value, while the bagplots are drawn around the depth median of the data. This results in the difference of location between the two types of grouping mechanisms used, specifically in the defaulted case.

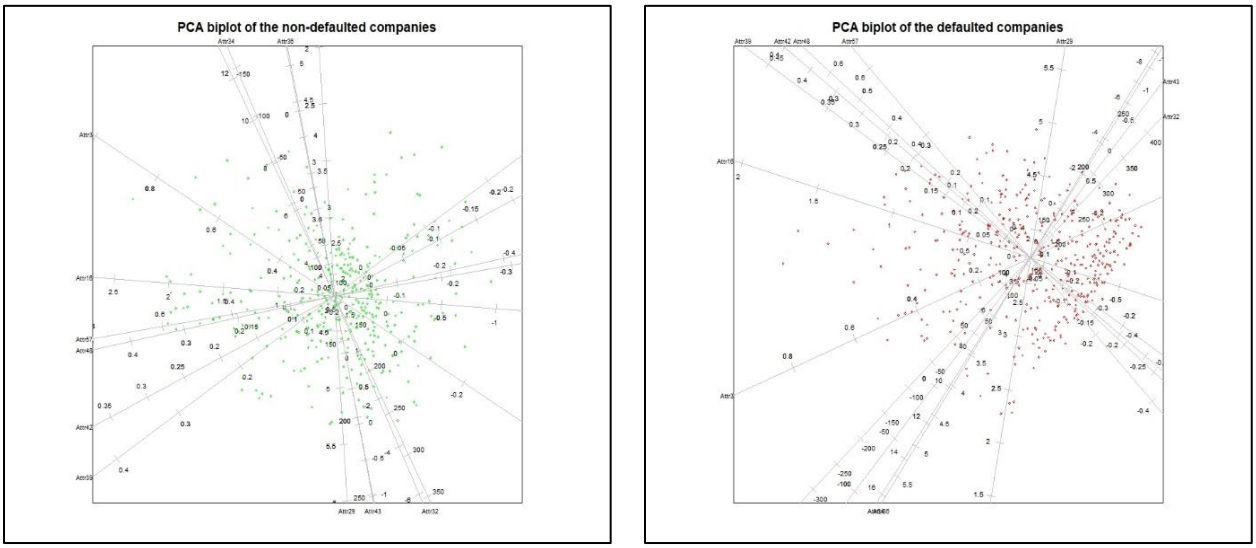

Figure 4.5: PCA biplot of the 410 non-defaulted companies (left) and defaulted (right) companies, independently

In Figure 4.5, the data set is split according to the indicator variable. An important aspect to notice when splitting data in this manner is that the range of the biplot display is dependent on the observations that must be plotted. It can be observed through Figure 4.5 that the non-defaulted observations are grouped closer together, resulting from a lower variance within the class. When the graphs in the figures are compared it can be misleading in the sense that the plot on the right seems to be more condensed, while this is not the case.

In the defaulted plot there is a larger dispersion of sample values and thus the biplot has a larger range, leading to most of the points being close to each other. In the non-defaulted plot, there are fewer extreme values in the observations, leading to a smaller range. When comparing two biplots that have the same variables, it is important to look at the values on the axes in order to compare them relatively. In the left plot it can also be observed that some pairs of variables are plotted very close to each other, indicating that they are either strongly positive or negative correlated. The plot on the right however has a more spread out display of variables.

4.3 MANIPULATING THE AXES

When working with large data sets, the biplot display can become cluttered and more challenging to interpret. There are many ways to treat this problem, including the modification of sample points and axes. The observation sizes can be reduced or removed entirely, leaving a display with only variables. This proves to be useful for scenarios where the variable axes are not of critical importance.

Another way of manipulating the plot is to only display some set of axes. In Figure 4.6 a plot of a grouping of axes that lie approximately diagonal from bottom left to top right are created, and another (the remaining) from bottom right to top left in the non-defaulted set.

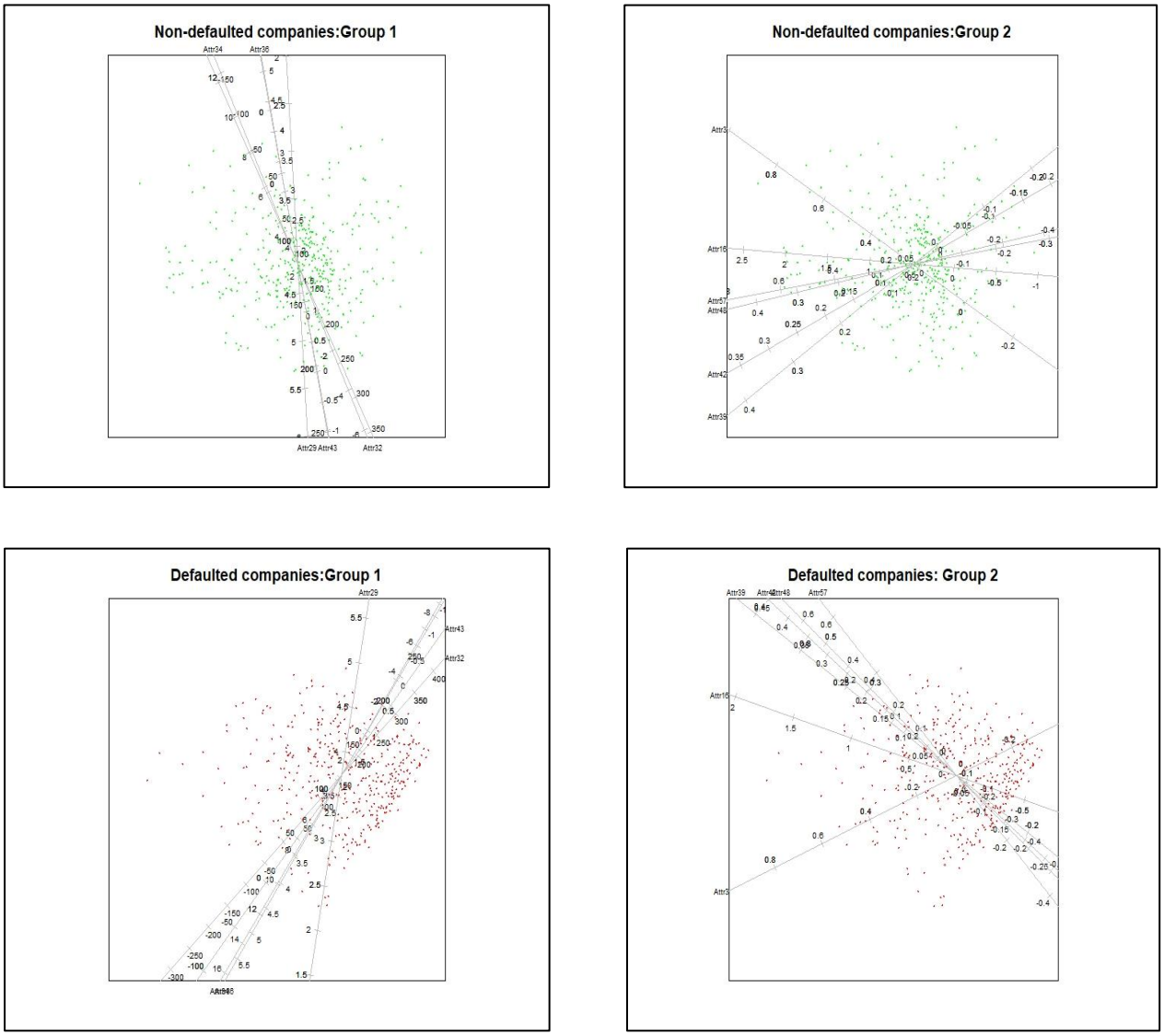

Figure 4.6: PCA biplots of grouping 1 (left) and grouping 2 (right) for non-defaulted (top) and defaulted (bottom) companies

Table 4.1: Groupings

Group 1	Group 2
29: Logarithm of total assets	3: Working capital/total assets
32: (current liabilities×365)/cost of products sold	16: (gross profit + depreciation)/total liabilities
34: Operating expenses/total liabilities	39: Profit on sales/sales
36: Total sales/total assets	42: Profit on operating activities/sales
43: Rotation receivables + inventory turnover	48: EBITDA/total assets
	57: (current assets – inventory – short-term liabilities)/(sales – gross profit – depreciation)

These two groups then consist of two sets of strong positively or negatively correlated variables in the non-defaulted set. When these same two groups of variables are plotted using the defaulted set, a very different result is obtained. While some variables are still positive or negatively correlated, some variables are shown to have no correlation. This is perhaps a result that would not have been so evident if all the axes had simultaneously been plotted. Many of these groupings can be constructed in order to extract value from the biplot displays. In many cases the isolation of either variables or sample points can lead to a better interpretation of covariates within the data.

4.4 QUALITY OF THE PREDICTIVITY

Figures 4.7 and 4.8, indicates the overall quality of the approximated values as predicted by the biplot axes, i.e. how well the PCA biplot approximates the original centred data. The predictivities are plotted for each variable against the number of dimensions fitted. The overall quality is obtained by weighting the predictivities proportionally to each variance attained by each variable, giving it a cumulative increase as one move into higher dimensional subspaces.

Table 4.2: Cumulative overall quality of predictivities vs dimension of the subspaces of non-defaulted companies

1	2	3	4	5	6	7	8	9	10	11
32.54	54.84	68.04	76.59	83.04	88.67	92.83	95.97	97.65	99.25	100.00

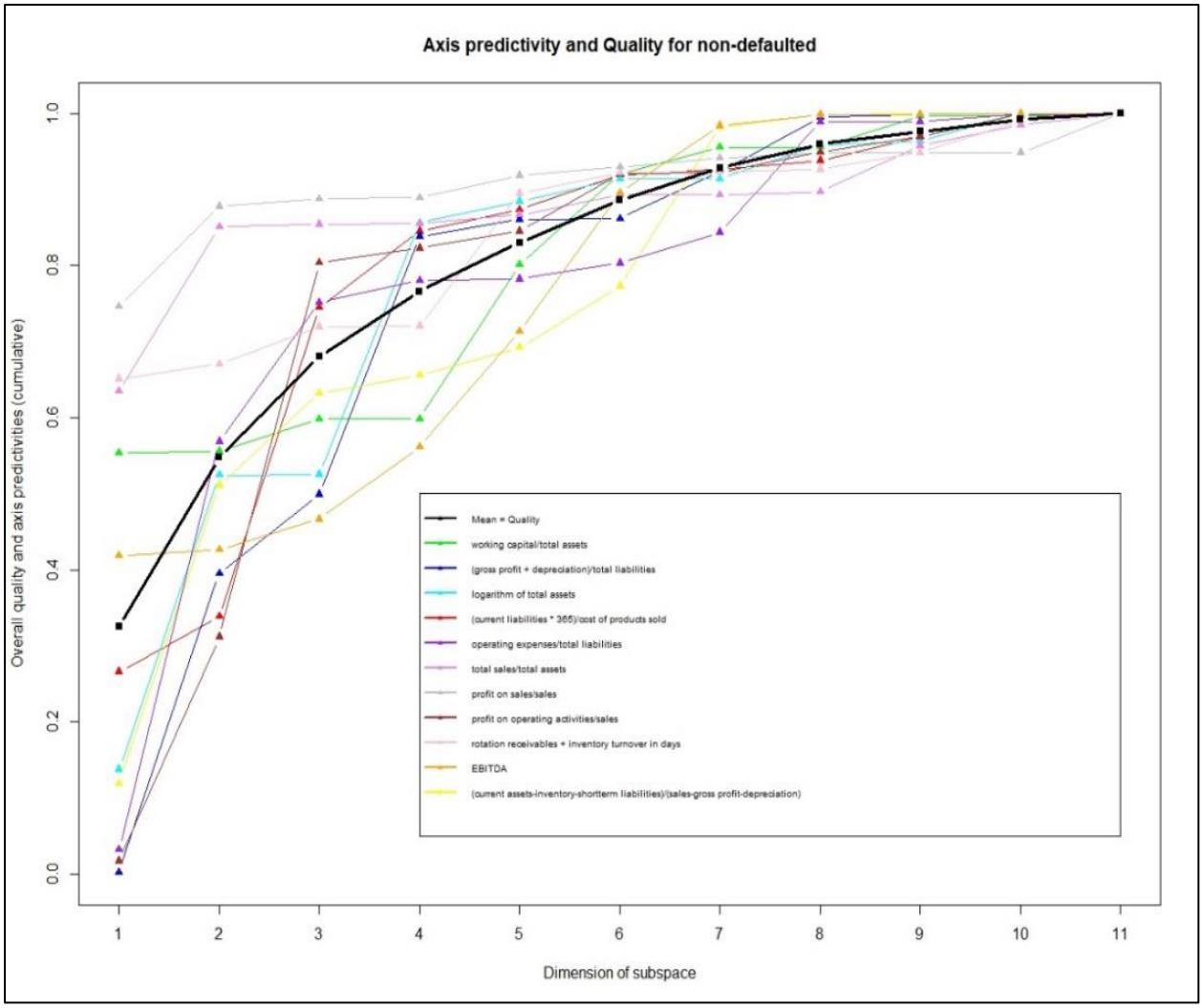

Figure 4.7: Quality of PCA biplot from non-defaulted companies

The plot and table above for non-defaulted companies clearly shows how each possible variable improve in quality, reaching 80% as early in the fifth dimensional subspace. The overall quality for the PCA biplot in Figures 4.7 and 4.8 is approximately 55% in the two dimensional case. This might seem very low, since 45% of the variability is not represented, but keeping in mind that the number of dimensions has been reduced from 11 to 2, in which case this is not too bad.

Another aspect to notice is that the quality of both the non-defaulted and defaulted samples are closely related. This might be an indication of an appropriate sample chosen to compare these indicator variables. The plots regarding the sample of 410 defaulted and non-defaulted companies above are therefore not a bad representation of the overall Polish bankruptcy set.

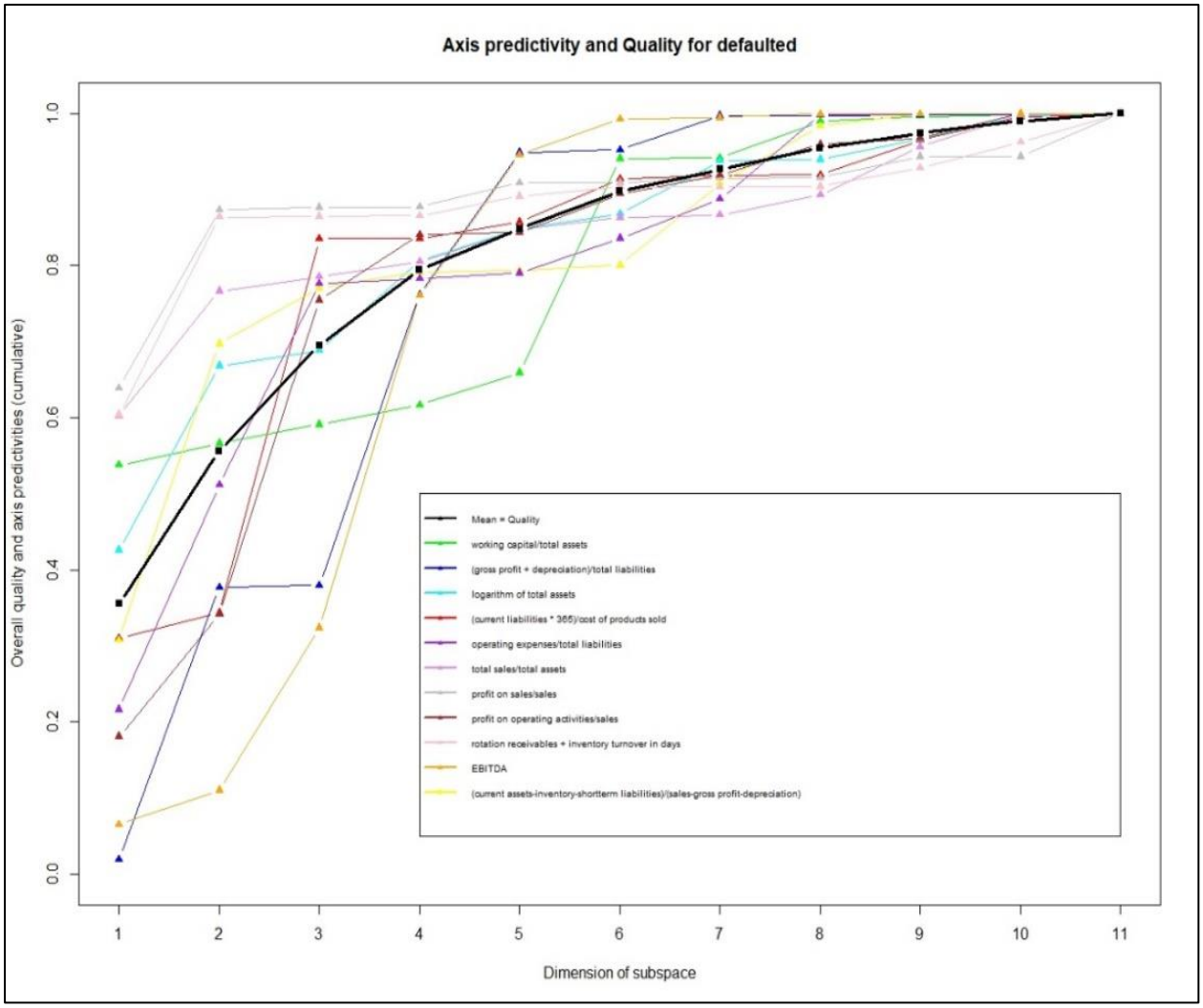

Figure 4.8: Quality of PCA biplot from defaulted companies

Table 4.3: Cumulative overall quality of predictivities vs dimension of subspaces of defaulted companies

1	2	3	4	5	6	7	8	9	10	11
35.53	55.63	69.53	79.49	84.86	89.77	92.64	95.44	97.45	99.03	100.00

4.5 CVA BIPLOTS

From Chapter 2, recall that CVA biplots groups the input data into classes, dependant on some input variable. In Figure 4.9 the data is grouped once again according to the default status of the companies, which is the 12th input variable of the original dataset. In the plot the bagplot is once again utilised – the green corresponding to non-defaulted companies and the red corresponding

to defaulted companies. The different classes in the CVA biplot above are constructed around the mean values which can be determined from the axes. In the data, the class means are represented by the large dots in the plot and are separated by a significant distance. There is, however a large part of the classes that overlap, indicating that there is a portion of data where the classification does not necessarily imply a certain profile of ratios.

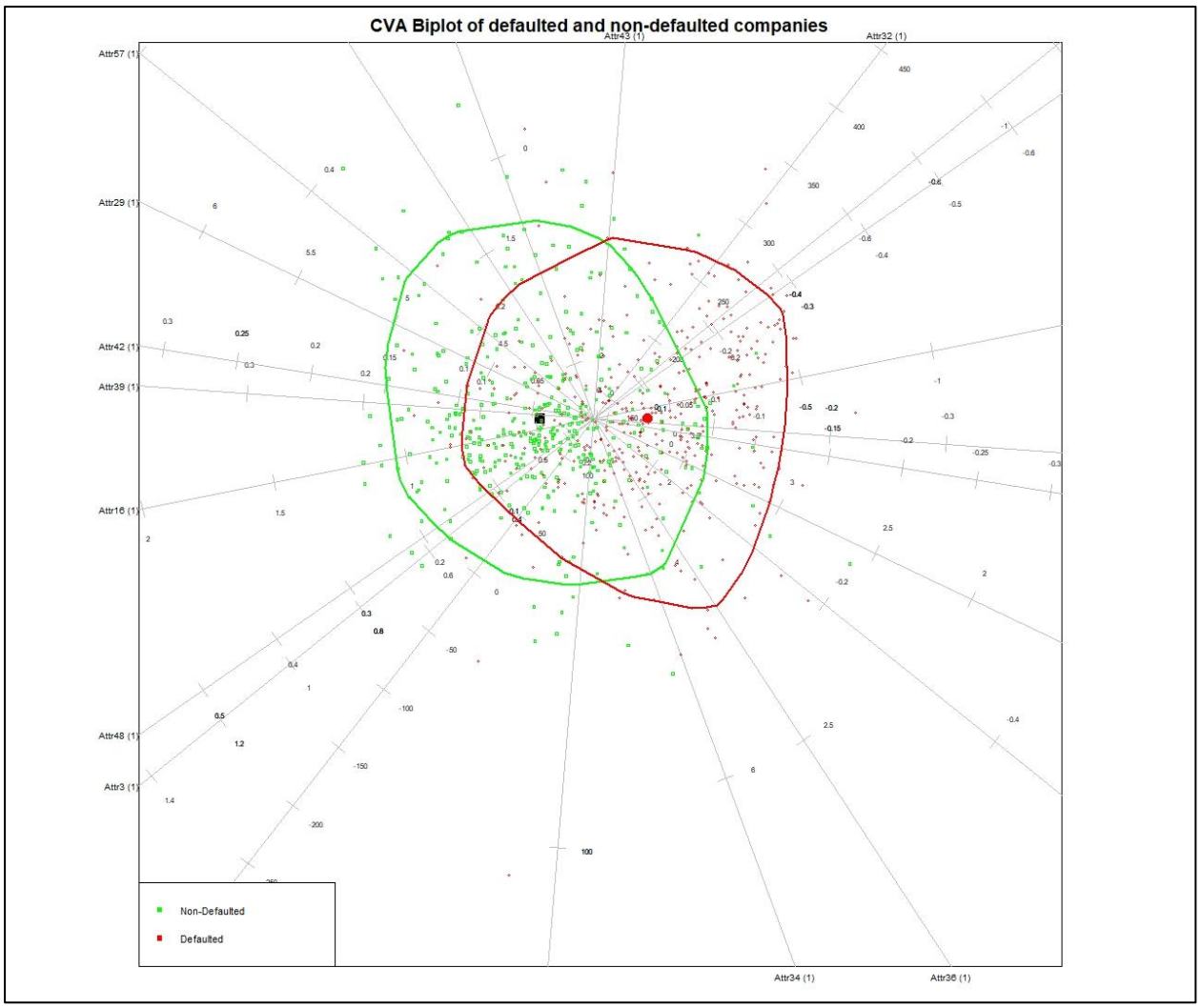

Figure 4.9: CVA biplot of the defaulted and non-defaulted companies

4.6 THE DENSITY PLOTS

Figures 4.10 and 4.11 are two-dimensional PCA biplots on top of a density plot from a two-dimensional PCA approximation of the input set of defaulted, and non-defaulted companies respectively indicating the difference in distribution of the data. These figures were created using the MASS package to perform two-dimensional kernel density estimation. The densities are centred on the origin since the data has been automatically scaled. Initially, the sample points were included in the graph, but these have no meaning since the points overwhelms the plot.

A much clearer picture can be drawn by switching off the sample points, as is done above. Another characteristic of the above graphs is the 0.90-bag enclosing the inner 90% of the sample points.

The bottom bar indicates the height of each density function with the colour red as the highest value and green as the minimum value. The first important feature to notice is that the density of the defaulted companies varies much wider than the non-defaulted companies, i.e. it is more dispersed. The second feature is that the 0.90-bag contains a wider portion of samples for the non-defaulted part compared to the defaulted.

The density of the non-defaulted companies is much more centred but higher than that compared to the defaulted, indicating a significant difference of the two classes. Another important feature is that the two groups formed by the defaulted density, which could indicate the substantial difference in ratio values from defaulted companies whereby relating to the extremeness of going bankrupt. Lastly, the density contours as indicated by the black lines shows fair amounts of variation from the two different classes.

The clear aspect to note in these densities, is that the proportion of defaulted values is not as well spread as in the non-defaulted density. The reader may have difficulty in understanding the mathematics behind the above plots, but various packages exist in many software to ease the manner in how these plots are constructed. Therefore, the readers should not be discouraged when building their own plots even when the underlying mathematics seems rigorous.

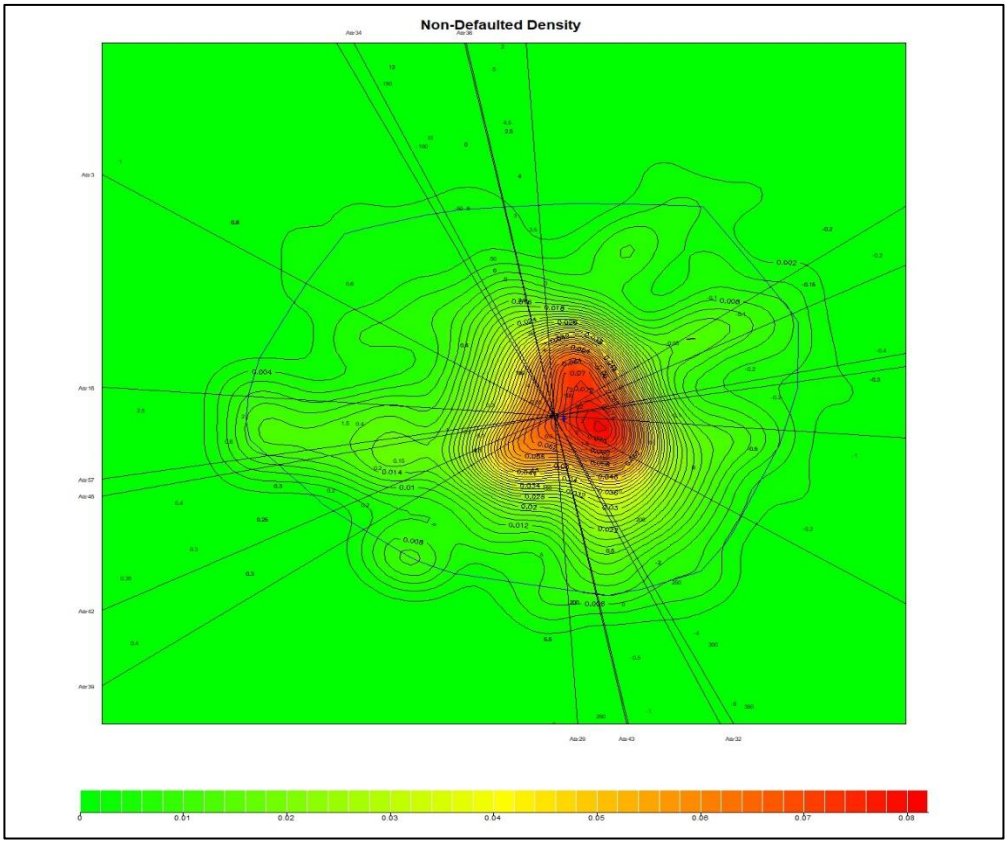

Figure 4.10: PCA density biplot of the non-defaulted companies

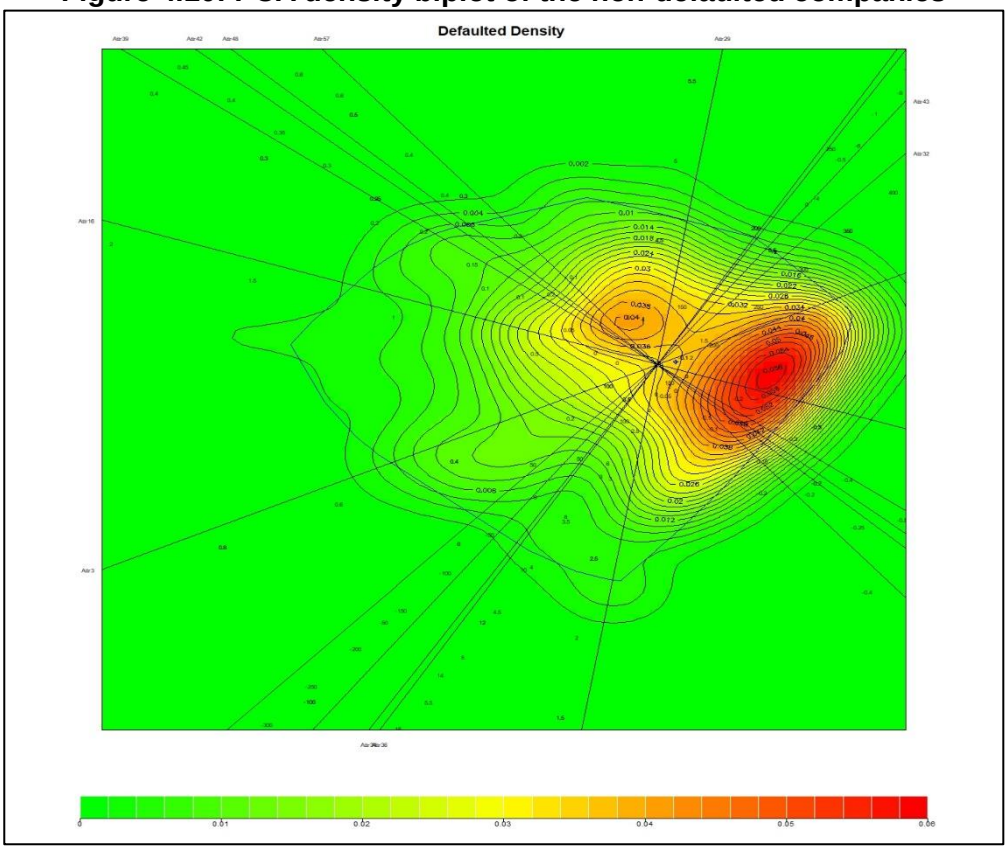

Figure 4.11: PCA density biplot of the defaulted companies

CHAPTER 5
CONCLUSION AND RECOMMENDATIONS

5.1 CONCLUSION

In this research assignment the method of using a visual technique in the form of biplots to display multidimensional financial data was explored. Firstly, the idea behind visualisation as a tool and what utility it has was discussed, including whether or not the method of drawing pictures to aid interpretability can be extended to a multidimensional case. Visualisation is a well-known technique in the two-dimensional space, but once the dimensions increase, the trend is to move away because of the difficulty of interpretation. The human eye can only see in three-dimensions, so it is a natural process to ease away from attempting to plot multidimensional inputs. It was shown through the use of approximation that high dimensional data can in fact be represented in two-dimensions by making use of well-known mathematical procedures.

Second, the fundamental theory underlying biplots was explored, where the focus shifted to the two main types of biplots commonly used: namely PCA and CVA. These different techniques were illustrated using arbitrary data sets in order give a better understanding before investigating a financial set. An important characteristic that was included in the explanation was that it could be understood by individuals from fields of knowledge outside of mathematical statistics.

Thirdly, the data used in this research, namely The Polish defaulting set, was revealed, explaining the unique characteristics of this multidimensional set, including the important default indicator variable. The method of pre-processing the data was shown, including the variable selection techniques as well as the outlier detection and treatment steps. It was evident that the 64 dimensional data could be adequately represented by a significantly smaller dimension (11) because of the similarities amongst many of the financial ratios.

Fourthly, the various biplot techniques were applied to the Polish set whereby value was extracted by looking at the relations between variables. The data was classified according to the default status of the companies, and various plots were made where these two classes were compared. The underlying idea of constructing an analysis on the shape by which the biplot axes take was examined, showing that one can interpret the covariances and correlations of multidimensional data in a two-dimensional plot.

Lastly, the density plots of each sample of the Polish set were examined where substantial differences in the defaulted and non-defaulted sets were observed. This in fact, could give the manager an overview of how well the company, ratio wise, is challenging other companies through the aid of a simple plot.

A fund manager can also compare different ratio values from different companies using this plot by analysing different relationships and groupings between ratios.

It could be a valuable asset to integrate the use of biplots into the analyses done by people outside the realm of statistics, since it is a highly technical method that can be understood by people with limited knowledge of the field. This could create a more effective system where inputs from various parties are better transformed into decision making by the management in the firms.

To conclude this research assignment, the above plots were not time consuming to create. Therefore, it is clear that high dimensional could be adequately represented in two dimensions without a significant loss of information. These plots could potentially contribute in better analysis and decision making processes with regards to a company's financial structure.

5.2 RECOMMENDATIONS

Further research could be done on why the density groups were formed in such a manner by comparing time series data of the companies or by observing the exact nature of the defaults that occurred. The biplot methodology can be linked to modelling default risk through Probability of Default models as used in credit risk.

Secondly, different variable selection techniques can be taken into consideration to possibly yield different outcomes from the Polish dataset.

Lastly, extensions to this analysis can be added using some of the different kinds of biplots mentioned. These varying techniques could add value and highlight relations that were unobserved through this analysis.

REFERENCES

Barr, G.D.I. & Affleck-Graves, J.F. 1987. *The covariance biplot and stock market data: An alternative relative strength chart.* South African Journal of Business Management, 18: 46 – 50.

Barr, G.D.I., Kantor, B.S. & Underhill, L.G. 1987. *The weighted covariance biplot – an application*. South African Statistical Journal, 21: 155 – 171.

Cox, D. R. & Snell, E. J. 1989. *Analysis of Binary Data (second edition)*. Chapman & Hall/CRC.

Donders, A.R.T., Van Der Heijden, G.J., Stijnen, T. & Moons, K.G., 2006. *A gentle introduction to imputation of missing values.* Journal of clinical epidemiology, 59(10), pp.1087-1091.

Draper, N. R. & Smith, H. 1966. *Applied Regression Analysis*. New York: Wiley.

Drozdowicz-Biec, M. 2011. *Reasons Why Poland Avoided the 2007-2009 Recession*. Instytut Rozwoju Gospodarczego (SGH). Prace i Materialy, p.41.

Efroymson. M. A. Multiple regression analysis. 1960. *Mathematical methods for digital computers*. New York: Wiley.

Gabriel, K.R. 1971. *The biplot graphical display of matrices with application to principal component analysis*. Biometrika, 58: 453 – 467.

Gardner, S., le Roux, N.J., & Olivier, P. 2003. *Biplots for displaying multidimensional financial performance data graphically*. South African Journal of Accounting Research, 17:1, 41-64, DOI: 10.1080/10291954.2003.11435105.

George, E. I. 2000. *The Variable Selection Problem*. Journal of the American Statistical Association, 95 (452), 1304-1308.

Gower, J.C. & Hand, D.J. 1996. *Biplots*. London: Chapman & Hall.

Gower, J. Gardner-Lubbe, S. & le Roux, N.J. 2011. *Understanding Biplots*. United Kingdom: John Wiley & Sons, Ltd.

Greenacre, M.J. 2010. *Biplots in Practice*. Spain: S.A. de Litografia.

Greenacre, M.J. 2012. *Biplots: the joy of singular value decomposition*. Wiley Interdisciplinary Reviews: Computational Statistics, 4(4), pp.399-406. doi:10.1002/wics.1200.

Hastie, T.; Tibshirani, R. & Friedman, J. 2001. *The Elements of Statistical Learning*. Springer New York Inc. , New York, NY, USA.

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An introduction to statistical learning (Vol.112).* New York: Springer.

Johnson, A. & Wichern, D. 2008. *Applied Multivariate Statistical Analysis*. USA: Pearson/Prentice Hall, Inc.

Jolliffe, I., 2011. *Principal component analysis*. In International encyclopaedia of statistical science (pp. 1094-1096). Springer, Berlin, Heidelberg.

Kullback, S. & Leibler, R.A., 1951. *On information and sufficiency*. The annals of mathematical statistics, 22(1), pp.79-86.

Miller, A. 1990. *Subset Selection in Regression*. London: Chapman and Hall.

Poland. Report Economy. 2012.

Sober, E., 1981. *The principle of parsimony*. The British Journal for the Philosophy of Science, 32(2), pp.145-156.

Stevens, D. L. 1972. *Financial characteristics of merged firms: a multivariate analysis*. Working Paper of the college of Commerce and Business Administration, University of Illinois at Urbana-Champaign.

Tudor, E. (2009). *Metode de recunoastere a formelor in analiza economico-financiara*. http://www.asecib.ase.ro/simpozion/2009/full_papers/pdf/22_tudor_eugeniu_ro.pdf

Zieba, M., Tomczak, S. K., & Tomczak, J. M. 2016. *Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction*. Expert Systems with Applications [Online]. Available: doi:10.1016/j.eswa.2016.04.001

APPENDIX A:

POLISH BANKRUPTCY DATA SET VARIABLES

The following is a complete list of the 64 quantitative variables from the Polish bankruptcy dataset:

1 net profit / total assets
2 total liabilities / total assets
3 working capital / total assets
4 current assets / short-term liabilities
5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
6 retained earnings / total assets
7 EBIT / total assets
8 book value of equity / total liabilities
9 sales / total assets
10 equity / total assets
11 (gross profit + extraordinary items + financial expenses) / total assets
12 gross profit / short-term liabilities
13 (gross profit + depreciation) / sales
14 (gross profit + interest) / total assets
15 (total liabilities * 365) / (gross profit + depreciation)
16 (gross profit + depreciation) / total liabilities
17 total assets / total liabilities
18 gross profit / total assets
19 gross profit / sales
20 (inventory * 365) / sales
21 sales (n) / sales (n-1)
22 profit on operating activities / total assets
23 net profit / sales
24 gross profit (in 3 years) / total assets
25 (equity - share capital) / total assets
26 (net profit + depreciation) / total liabilities
27 profit on operating activities / financial expenses
28 working capital / fixed assets
29 logarithms of total assets
30 (total liabilities - cash) / sales
31 (gross profit + interest) / sales
32 (current liabilities * 365) / cost of products sold
33 operating expenses / short-term liabilities
34 operating expenses / total liabilities
35 profit on sales / total assets
36 total sales / total assets
37 (current assets - inventories) / long-term liabilities
38 constant capital / total assets
39 profit on sales / sales
40 (current assets - inventory - receivables) / short-term liabilities
41 total liabilities / ((profit on operating activities + depreciation) * (12/365))
42 profit on operating activities / sales
43 rotation receivables + inventory turnover in days
44 (receivables * 365) / sales
45 net profit / inventory
46 (current assets - inventory) / short-term liabilities
47 (inventory * 365) / cost of products sold
48 EBITDA (profit on operating activities - depreciation) / total assets

49	EBITDA (profit on operating activities - depreciation) / sales
50	current assets / total liabilities
51	short-term liabilities / total assets
52	(short-term liabilities * 365) / cost of products sold)
53	equity / fixed assets
54	constant capital / fixed assets
55	working capital
56	(sales - cost of products sold) / sales
57	(current assets - inventory - short-term liabilities) / (sales - gross profit – depreciation)
58	total costs /total sales
59	long-term liabilities / equity
60	sales / inventory
61	sales / receivables
62	(short-term liabilities *365) / sales
63	sales / short-term liabilities
64	sales / fixed assets

APPENDIX B:

FORWARD STEPWISE SELECTION ALGORITHM

Algorithm: Forward Stepwise Selection

1. Let M_0 denote the *null* model, which contains no predictors.

2. For k = 0, 1,…, p-1:

(a) Consider all $p - k$ models that augment the predictors in M_k with one additional predictor.

(b) Choose the *best* among these $p - k$ models and call it M_{k+1}. Here *best* is defined as having smallest RSS or highest R^2.

3. Select a single best model from among $M_0,\ldots,M_p$ using cross-validated

prediction error, C_p (AIC), BIC, or adjusted R^2.

APPENDIX C:

BACKWARD STEPWISE SELECTION ALGORITHM

Algorithm: Backward Stepwise Selection

1. Let M_p denote the *full* model, which contains all p predictors.

2. For k = p, p-1,…,1:

(a) Consider all k models that contain all but one of the predictors in M_k , for a total of k -1 predictors.

(b) Choose the *best* among these k models and call it M_{k-1}. Here *best* is defined as having smallest RSS or highest R^2.

3. Select a single best model from among $M_0,\ldots,M_p$ using cross-validated

prediction error, C_p (AIC), BIC, or adjusted R^2.