Multivariate Analysis
# of Trace Elements in Pyrite

Lyron J. Winderbaum

Supervisor: Associate Professor Andrew Metcalfe

October 2011

Thesis submitted for the degree of Honours in Mathematical Sciences

SCHOOL OF MATHEMATICAL SCIENCES

THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

Declaration

Except where stated this thesis is, to the best of my knowledge, my own work and my supervisor has approved its submission.

Signed by student:

Date:

Signed by supervisor:

Date:

Acknowledgements

I would like to thank Andrew Metcalfe, for constant encouragement, patience, and advice. Thanks to Nigel Cook and Cristiana Ciobanu who provided the data, and classified each grain of pyrite by painstakingly analyzing images of each individual sample, and helping in explaining some basic geology to me. Last but not least I would also like to thank Inge Koch for providing continuous motivation throughout the year, and inspiration through her AMHD (Analysis of Multivariate and High-Dimensional Data) AMSI course [8].

Abstract

I introduce and develop a variety of methods for the exploration of LAICPMS data on trace element concentrations. I demonstrate how to produce and interpret dendrograms of hierarchical agglomerative cluster analysis by correlation-based distance measures to identify interesting clusters of mutually correlated elements. I introduce the use of parallel coordinate plots, principle component analysis, and factor analysis as exploratory techniques in identifying interesting features of data. I introduce and develop a bootstrap-based method for detection of influential points that may be difficult to detect by other, more standard methods, and can be adapted to specifically detect particular types of influence on specific statistics which may be of particular interest. Through these methods, and more, I produce results that are of geological interest, some (such as the presence of a strong positive correlation between cobalt and nickel) which support known scientific theories, and others that may be new, and warrant further research.

Contents

1 Introduction **2**
1.1 The Data 2

2 Descriptive Statistics **5**
2.1 Introduction 5
2.1.1 The LA-ICP-MS Data 5
2.1.2 Zero Values 7
2.1.3 Notation 7
2.2 Raw Data 9
2.2.1 Univariate Statistics 9
2.2.2 Boxplots 10
2.2.3 Depth Plots 12
2.3 Log-transformed Data 14
2.3.1 Univariate Statistics 14
2.3.2 Boxplots 15
2.3.3 Depth Plots 17
2.3.4 Normal QQ (quantile-quantile) Plots 19
2.4 A Zero-inflated Log-Normal Model 21

3 Correlations **24**
3.1 Correlation Matrices 24
3.1.1 Pearson Product-moment Correlation Coefficient 24
3.1.2 Raw Correlations 25
3.1.3 Log-transformed Correlations 25
3.2 Cluster Analysis 26
3.2.1 Distance Measures 27
3.2.2 Linkage 27
3.2.3 Hierarchical Agglomerative Cluster Analysis 28
3.3 Correlation-based Cluster Analysis 29
3.3.1 Log-transformed Data 32
3.3.2 Rank-based Measures of Association 33
3.4 Summary 35

4 Exploratory Analyses **37**
4.1 Introduction 37
4.2 Parallel Coordinate Plots 37
4.3 Principal Component Analysis 40
4.3.1 PCA - Principal Component Analysis 40
4.3.2 Demonstration of Results 41
4.3.3 PCA on the Raw Concentrations 43
4.3.4 PCA on the Standardized Data 46
4.3.5 Log-transformed Data 50
4.4 Factor Analysis 52
4.4.1 Maximum Likelihood Approach 52
4.4.2 Principal Factor Analysis 56

5 Bootstrap based Influence Diagnostics 59
5.1 The Nonparametric Bootstrap . . . 59
5.2 An illustrative example of Bootstrapping . . . 60
5.3 Bootstrap Distributions of the Correlation Coefficients . . . 61
5.4 Distributional Results . . . 62
5.5 An illustrative example of Influence Diagnostics . . . 63
5.6 Notes on Distributional Results for Joint Proportions of Absence . . . 67
5.7 Bootstrap Diagnostics . . . 68
5.8 Future Work . . . 69

6 Discriminant Analysis 70
6.1 k-means Clustering . . . 72
6.1.1 4-class problem . . . 72
6.1.2 2-classproblem . . . 74
6.2 Fisher's Linear Discrimination Rule . . . 75
6.3 Another method, lda() in R . . . 77
6.4 Futher Work . . . 78

7 Regression 79
7.1 Introduction . . . 79
7.2 Use of the log-transformed concentrations . . . 79
7.2.1 Box-Cox and other MLE methods . . . 79
7.3 No Interaction Models . . . 81
7.4 Interaction Models . . . 84

8 Conclusion 86

Chapter 1

Introduction

1.1 The Data

In this thesis i investigate a dataset of measurements made on grains of pyrite [9], from Moonlight Prospect near Pajingo Mine, south of Townsville (Figure 1.1). The primary motivation for the collection of this data was determining if the concentration of gold in the deposit was sufficient to make it economically viable to mine. The concept of ‘hidden gold’ made this a particularly interesting problem. Hidden gold is where some proportion of the gold is chemically locked into the crystal structure of minerals (such as primarily pyrite) and this complicates the process of estimating the total amount of available gold somewhat. The chemistry involved in this process of trace elements (such as gold) being included in the chemical structure of the minerals is not entirely explained, and so geologists and geo-chemists are particularly interested in further investigating it. As such, this investigation is the primary focus of this thesis while the estimation of gold concentration is reduced to being of peripheral interest. Often in these situations it is difficult to elucidate the processes which lead to the inclusion of this hidden gold, and it is only a proportion of this gold that exists in the crystal structure, while the rest exists as in its native state, or possibly even as nano-particles embedded in the crystals. In most cases, it can be very difficult to distinguish between these different forms.

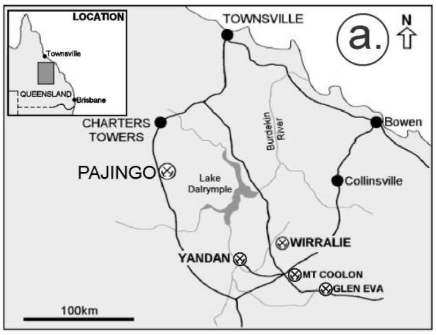


Figure 1.1: Location of Mine

This is one of the reasons this data is particularly interesting, and rare as far as such datasets are concerned. All the measurements on gold concentration lie below what is referred to as the “solid-state solubility of gold” line [in arsenic-rich pyrite]. Which is the empirically determined maximum amount of gold that can be held in the crystal structure of the pyrite [10] (described by Equation 1.1), and so it may be reasonable to assume that most of the measured gold is hidden gold, as measurements above this line tend to indicate the presence of native ‘visible’ gold, or other minerals. This is illustrated in Figure 1.2, on which the line from Equation 1.1 is shown, and it can be seen that all but one of the measurements lie below the line.

$$C_{Au} = 0.02 \times C_{As} + 4 \times 10^{-5} \tag{1.1}$$

(C_{Au} is the concentration of Au,
and similarly C_{As} is the concentration of As)

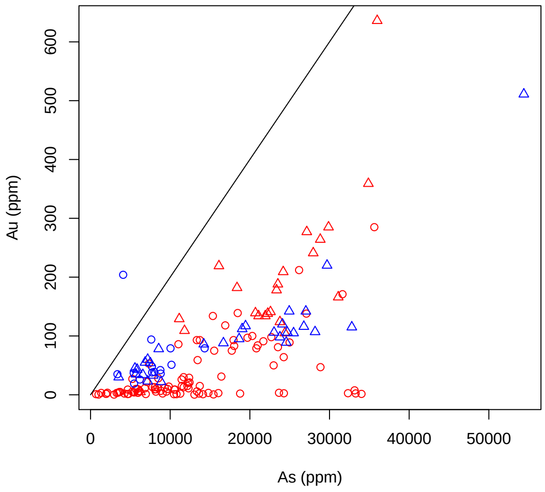
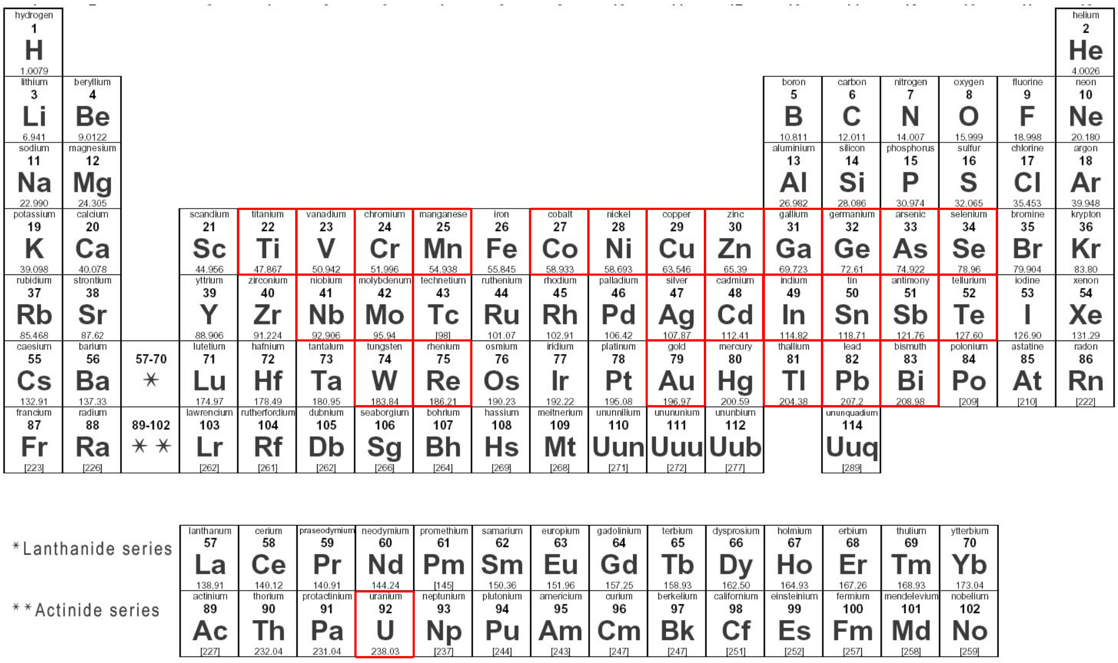


Figure 1.2: Scatterplot of the concentration of gold against concentration of arsenic with the line of maximum “solid-state solubility” from Equation 1.1 drawn

Symbol	Morphology	Location	#
○	Granular	Rock	98
△	Granular	Vein	22
○	Replacement	Rock	14
△	Replacement	Vein	30
		Total:	164

Table 1.1: Classes of Pyrite Legend

Although estimating the total amount of gold in the area was the original motivation for collecting these data, they provide an opportunity to investigate the nature of pyrite, and hopefully gain some insight into the geological processes involved in its formation. I start by looking at the data, first as single-dimensional data, in Chapter 2 then as multi-dimensional data, in Chapter 3 and Chapter 4. I try to visualize it in a way that will provide some insight into the relationships between the trace elements in the pyrite. One theory I investigate is that there is a distinct group of ‘lithophilic’ elements, that have some common correlation to each other across all the pyrite (due to being mutually related by the underlying source of magma that formed it). This theory is to some degree validated by the hierarchical agglomerative cluster analysis with a pseudometric based on the pairwise Pearson’s product-moment correlation coefficient. In Chapter 7 I take a brief look at some models that could improve the prediction of gold concentration. And in Chapter 6 I investigate some classification problems and discriminant analysis in trying to differentiate between subclasses of pyrite (shown in Table 1.1). Finally, as large influential values are a recurring theme throughout most of these analyses, I investigate a potentially very general influence diagnostic method based on the principle component analysis of Chapter 5.



hydrogen 1 H 1.0079																	helium 2 He 4.0026
lithium 3 Li 6.941	beryllium 4 Be 9.0122											boron 5 B 10.811	carbon 6 C 12.011	nitrogen 7 N 14.007	oxygen 8 O 15.999	fluorine 9 F 18.998	neon 10 Ne 20.180
sodium 11 Na 22.990	magnesium 12 Mg 24.305											aluminium 13 Al 26.982	silicon 14 Si 28.086	phosphorus 15 P 30.974	sulfur 16 S 32.065	chlorine 17 Cl 35.453	argon 18 Ar 39.948
potassium 19 K 39.098	calcium 20 Ca 40.078		scandium 21 Sc 44.956	titanium 22 Ti 47.867	vanadium 23 V 50.942	chromium 24 Cr 51.996	manganese 25 Mn 54.938	iron 26 Fe 55.845	cobalt 27 Co 58.933	nickel 28 Ni 58.693	copper 29 Cu 63.546	zinc 30 Zn 65.39	gallium 31 Ga 69.723	germanium 32 Ge 72.61	arsenic 33 As 74.922	selenium 34 Se 78.96	bromine 35 Br 79.904	krypton 36 Kr 83.80
rubidium 37 Rb 85.468	strontium 38 Sr 87.62		yttrium 39 Y 88.906	zirconium 40 Zr 91.224	niobium 41 Nb 92.906	molybdenum 42 Mo 95.94	technetium 43 Tc [98]	ruthenium 44 Ru 101.07	rhodium 45 Rh 102.91	palladium 46 Pd 106.42	silver 47 Ag 107.87	cadmium 48 Cd 112.41	indium 49 In 114.82	tin 50 Sn 118.71	antimony 51 Sb 121.76	tellurium 52 Te 127.60	iodine 53 I 126.90	xenon 54 Xe 131.29
caesium 55 Cs 132.91	barium 56 Ba 137.33	57-70 *	lutetium 71 Lu 174.97	hafnium 72 Hf 178.49	tantalum 73 Ta 180.95	tungsten 74 W 183.84	rhenium 75 Re 186.21	osmium 76 Os 190.23	iridium 77 Ir 192.22	platinum 78 Pt 195.08	gold 79 Au 196.97	mercury 80 Hg 200.59	thallium 81 Tl 204.38	lead 82 Pb 207.2	bismuth 83 Bi 208.98	polonium 84 Po [209]	astatine 85 At [210]	radon 86 Rn [222]
francium 87 Fr [223]	radium 88 Ra [226]	89-102 * *	lawrencium 103 Lr [262]	rutherfordium 104 Rf [261]	dubnium 105 Db [262]	seaborgium 106 Sg [266]	bohrium 107 Bh [264]	hassium 108 Hs [269]	meitnerium 109 Mt [268]	ununnilium 110 Uun [271]	unununium 111 Uuu [272]	ununbium 112 Uub [277]		ununquadium 114 Uuq [289]				

*Lanthanide series	lanthanum 57 La 138.91	cerium 58 Ce 140.12	praseodymium 59 Pr 140.91	neodymium 60 Nd 144.24	promethium 61 Pm [145]	samarium 62 Sm 150.36	europium 63 Eu 151.96	gadolinium 64 Gd 157.25	terbium 65 Tb 158.93	dysprosium 66 Dy 162.50	holmium 67 Ho 164.93	erbium 68 Er 167.26	thulium 69 Tm 168.93	ytterbium 70 Yb 173.04
* *Actinide series	actinium 89 Ac [227]	thorium 90 Th 232.04	protactinium 91 Pa 231.04	uranium 92 U 238.03	neptunium 93 Np [237]	plutonium 94 Pu [244]	americium 95 Am [243]	curium 96 Cm [247]	berkelium 97 Bk [247]	californium 98 Cf [251]	einsteinium 99 Es [252]	fermium 100 Fm [257]	mendelevium 101 Md [258]	nobelium 102 No [259]

Figure 1.3: Periodic Table of the Elements with the trace elements measured in this dataset highlighted in red

Chapter 2

Descriptive Statistics

2.1 Introduction

2.1.1 The LA-ICP-MS Data

Each grain of pyrite measured had its concentration of 27 trace elements (shown in Table 2.2) recorded in ppm (parts per million) recorded by LA-ICP-MS (Laser Ablation Inductively Coupled Plasma Mass Spectrometry). Each grain also had three associated variables recorded:

- Depth: the depth from which that particular pyrite grain was taken, in meters. Grains were collected from 11 different levels of depth.
- Location: the drill hole to which that grain belongs. There are three different drill holes from which data was obtained, referred to as hole no. 3542, 3430, and 3559a, here on in coded as numbers 1, 2, and 3 respectively.
- Class: a classification provided by the Nigel Cook and Cristiana Ciobanu (by looking at images of the grains) into two binary classes,
  - Morphology: Granular or Replacement.
  - Location: Rock or Vein.

It is worth noting that originally the data contained 28 elements, but one of these was Fe, Iron. This ‘variable’ had zero variance, the reason it was there is because as part of the data collection process the machine used was calibrated using a known standard, and these were the calibration measurements (always the same). Fe was used in the calibration (a process involving the transformation of a count variable to concentration, and the total mass ionized) as pyrite is the mineral FeS_2, and the measurements are all on trace elements (in the ppm scale), and so relative to these values the values of the concentration of Fe in the sample should be constant. The reason it is worth noting is that if one of the samples measured was not pyrite, and actually did not contain the appropriately large amount of Fe, then this would result in a calibration error, which would cause a significant error in the measurement. To get a rough idea of the distribution of these peripheral variables, consider we have LA-ICP-MS data from a 3 different drill holes, and several different depths from each:

- Hole: 3542 coded drill no. 1 from here on, (94 of 164 grains), has observations at depths:
  - 412.2m (7 of 94 grains)
  - 431.0m (8 of 94 grains)

- 435.2m (12 of 94 grains)
- 445.0m (37 of 94 grains)
- 448.2m (7 of 94 grains)
- 453.8m (11 of 94 grains)
- 479.7m (12 of 94 grains)

- Hole: 3430 coded drill no. 2 from here on, (58 of 164 grains), has observations at depths:
  - 419.45m (17 of 58 grains)
  - 445.30m (18 of 58 grains)
  - 447.80m (23 of 58 grains)
- Hole: 3559a coded drill no. 3 from here on, (12 of 164 grains), has observations at depth:
  - 664.0m (all 12 grains)

To get a better idea of the distribution of these values, and particularly the distribution of the classes of pyrite within these subsets (as shown in Table 1.1) consider Figure 2.1.

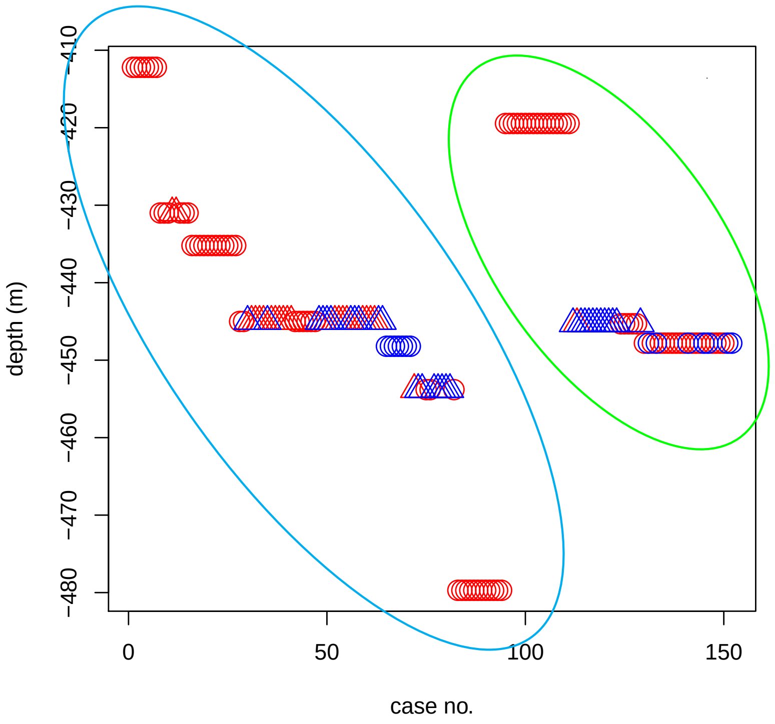


Figure 2.1: cases 1 to 152 vs depth, coloured by class as in Table 1.1

In Figure 2.1 the observations circled in cyan are from drill hole no. 1, and those circled in green are from drill hole no. 2. There are only 12 observations from drill hole no. 3, all of which are from a depth of 664m (far below on the scale shown in Figure 2.1, (this distance is illustrated in scatterplots of depth later in this chapter), and all 12 of these observations belong to the Morphology = Granular, Location = Rock (○) class of pyrite.

2.1.2 Zero Values

For each measurement, an MDL (Minimum Detection Limit) was calculated based on the precision and sensitivity of the machine used to take the measurements. Any values that fell below this threshold were set to zero as there was not sufficient evidence to conclude that they were significantly non-zero. This mechanism causes there to be a number of zero values amongst the data, which can cause a range of issues. For example, we cannot simply take the log-transform of the data (as the log of zero is undefined) and the zero-values have tied ranks, which can cause issues with rank-based measures. The last column in Table 2.2 was included primarily in order to give an indication of the significance this could have on any analyses. Visually it appears a log-transformation might be appropriate but zero values could cause a problem in this case. In the following sections we will discuss some methods for dealing with these issues.

2.1.3 Notation

In some datasets, defining what the variables are and what the observations are is quite easy. For example a dataset of n people with measurements of their heights and weights: clearly there are two variables, height and weight, and n observations (the people). However there are other cases in which the distinction between variable and observation is not as clear. In this case, the LA-ICP-MS data has an intuitive interpretation: that the 27 trace elements are the variables, and the 164 measurements are the observations. I will adhere to this interpretation for most of this thesis, however due to the nature of the dataset this distinction of variable and observation is not immutable. In particular applications (mainly the hierarchical cluster analysis discussed in Chapter 3), it is useful to 'switch perspective' and treat the 27 trace elements as the cases, or observations, and consider them as 164 dimensional objects. This switch of perspective can be very disorienting to readers who are accustomed to having clear cut dependent and independent variables, and certainly is not applicable in general. However in an exploratory analysis such as this, particularly in the example of this data, considering the 'variables' in this way can provide useful insight into relationships between them. In order to minimize this confusion, I introduce some notation here in Table 2.1 that I will consistently use throughout this thesis. I will also introduce most of the objects defined in Table 2.1 in the appropriate sections, but this provides a reference to come back to in the event of any confusion.

	Population		**Sample**
	Data		
d	number of variables	d	number of trace elements ($d = 27$)
		n	number of observations ($n = 164$)
$\boldsymbol{X}$	$d \times 1$ vector of random variables	$\mathbb{X}$	$n \times d$ data matrix s.t.
		x_{ij}	$= [\mathbb{X}]_{ij}$; the $(i, j)^{th}$ element of $\mathbb{X}$
X_j	$[\boldsymbol{X}]_j$; the j^{th} element of $\boldsymbol{X}$	$\boldsymbol{x}_{\bullet j}$	the j^{th} column of $\mathbb{X}$ (j^{th} variable)
		$\boldsymbol{x}_{i\bullet}^T$	the i^{th} row of $\mathbb{X}$ (i^{th} observation)
$\boldsymbol{\mu}$	$= E[\boldsymbol{X}]$; $d \times 1$ mean vector s.t.	$\bar{\boldsymbol{x}}$	$d \times 1$ sample mean vector s.t.
μ_j	$= [\boldsymbol{\mu}]_j = E[X_j]$	$\bar{x}_j$	$= [\bar{\boldsymbol{x}}]_j = mean(\boldsymbol{x}_{\bullet j}) = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$
	Variance Structure		
Σ	$d \times d$ covariance matrix s.t.	S:	$d \times d$ covariance matrix s.t.
σ_{jk}	$= [\Sigma]_{jk} = E[(X_j - \mu_j)(X_k - \mu_k)]$	s_{jk}	$= [S]_{jk}$
			$= \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$
σ_j^2	$= \sigma_{jj} = var(X_j)$	s_j^2	$= s_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$
Σ_{diag}	$d \times d$ diagonal matrix of $\{\sigma_j^2\}$	S_{diag}	$d \times d$ diagonal matrix of $\{s_j^2\}$
C	$= \Sigma_{diag}^{-\frac{1}{2}} \Sigma \Sigma_{diag}^{-\frac{1}{2}}$	C	$= S_{diag}^{-\frac{1}{2}} S S_{diag}^{-\frac{1}{2}}$ s.t.
		r_{jk}	$= [C]_{jk} = \frac{s_{jk}}{\sqrt{s_j^2 s_k^2}}$
	Transformations		
	Raw (Original) Data	$\mathbb{X}^{(raw)}$:	(Superscripts $^{(raw)}$ follow
$\boldsymbol{X}^{(raw)}$		$x_{ij}^{(raw)}$	$= \left[\mathbb{X}^{(raw)}\right]_{ij}$ notation as in *Data*)
	Centered Data	$\mathbb{X}^{(cent)}$:	
$\boldsymbol{X}^{(cent)}$	$= \boldsymbol{X}^{(raw)} - \boldsymbol{\mu}$	$\boldsymbol{x}_{i\bullet}^{(cent)}$	$= \left(\boldsymbol{x}_{i\bullet}^{(raw)} - \bar{\boldsymbol{x}}\right) \; \forall i \in \mathbb{Z} \cap [1, n]$
	Scaled Data	$\mathbb{X}^{(scal)}$:	
$\boldsymbol{X}^{(scal)}$	$= \Sigma_{diag}^{-\frac{1}{2}} \boldsymbol{X}^{(raw)}$	$\boldsymbol{x}_{i\bullet}^{(scal)}$	$= S_{diag}^{-\frac{1}{2}} \boldsymbol{x}_{i\bullet}^{(raw)} \; \forall i \in \mathbb{Z} \cap [1, n]$
	Standardized Data	$\mathbb{X}^{(stan)}$:	
$\boldsymbol{X}^{(stan)}$	$= \Sigma_{diag}^{-\frac{1}{2}} \left(\boldsymbol{X}^{(raw)} - \boldsymbol{\mu}\right)$	$\boldsymbol{x}_{i\bullet}^{(stan)}$	$= S_{diag}^{-\frac{1}{2}} \boldsymbol{x}_{i\bullet}^{(cent)} \; \forall i \in \mathbb{Z} \cap [1, n]$
	Log-Transformed Data	$\mathbb{X}^{(log)}$:	
$\boldsymbol{X}^{(log)}$	$= ln\left(\boldsymbol{X}^{(raw)} + \boldsymbol{1}\right)$	$x_{ij}^{(log)}$	$= ln\left(x_{ij}^{(raw)} + 1\right)$
			$\forall i \in \mathbb{Z} \cap [1, n], \quad j \in \mathbb{Z} \cap [1, d]$

Table 2.1: Some Definitions

Note that the subscript $j \in \mathbb{Z} \cap [1, d]$ refers to a particular trace element in the context of this dataset, and so in certain cases I will replace the subscript with the corresponding elemental symbol (as in the order shown in Table 2.2). For example, $\boldsymbol{x}_{\bullet 3} = \boldsymbol{x}_{\bullet As}$. Also often, $\boldsymbol{X}$ and $\mathbb{X}$ (rather than $\boldsymbol{X}^{(raw)}$ and $\mathbb{X}^{(raw)}$, or $\boldsymbol{X}^{(log)}$ and $\mathbb{X}^{(log)}$) will simply be used, and the specific transformation of the data being referred to will be defined at the beginning of the section, but this notation is here for consistency and for cases where clarification is needed. Furthermore, note how I have set up the notation in Table 2.1 such that each of the vector objects is a column vector, so that I can treat all vectors as column vectors. This can however cause some confusion because in the vector of random variables $\boldsymbol{X}$ the variables are its 'rows' while in the data matrix $\mathbb{X}$ the variables are its columns, and that $\boldsymbol{x}_{i\bullet}$ is actually the *transpose* of the i^{th} row of $\mathbb{X}$. I use this form of the data matrix, rather than its transpose (as [8] does), because I make extensive use of R, and this is the form in which R displays its `data.frame` objects.

2.2 Raw Data

2.2.1 Univariate Statistics

In any exploratory analysis the first thing to do is to look at the data. So here we begin by looking at the data by considering each element as an independent variable, to give us some idea of the basic shape of the data, and what possible underlying distributions could be used to model it.

First we take a look at the raw data, calculating the following basic statistics for each of the 27 elements (shown in Table 2.2), to give some basic understanding of the location, spread, and skewness of the data:

- Mean: $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$ (note here $n = 164$) $\bar{x}_j$ is the mean of the j^{th} Element, in the order listed in Table 2.2.
- Standard Deviation: $s_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ where similarly s_j is the sample standard deviation of the j^{th} Element, in the order listed in Table 2.2.
- Coefficient of Variation: $c_j = \frac{s_j}{\bar{x}_j}$, a standardized measure of spread. It is worth noting that $\bar{x}_j > 0\ \forall j$
- Skewness: $skew_j = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^3}{(\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2)^{\frac{3}{2}}}$, a measure of the skewness of the data.
- Proportion of Zeros: $p_j = \frac{n_j}{n}$ where n_j is the number of zeros in the j^{th} element, and $n = 164$ is the total number of observations.

I use a denominator of $n - 1$ for the standard deviation, but n for skewness, and this may seem like an inconsistency. However in this context is makes very little difference as these values are only being used to get a rough idea of the structure of the data, not for any inference or tests, particularly none depending on asymptotic results.

				Standard	Coefficient of		Proportion
No.	Variable		Mean (ppm)	Deviation	Variation	Skewness	of Zeroes
1	Gold	Au	7.03 x 10^{1}	9.24 x 10^{1}	1.32	2.7	.00
2	Silver	Ag	2.67 x 10^{2}	4.11 x 10^{2}	1.54	2.2	.00
3	Arsenic	As	1.53 x 10^{4}	9.81 x 10^{3}	0.64	0.7	.00
4	Antimony	Sb	4.37 x 10^{2}	5.92 x 10^{2}	1.36	2.6	.00
5	Titanium	Ti	3.02 x 10^{2}	1.89 x 10^{2}	6.28	10.6	.16
6	Vanadium	V	5.73 x 10^{0}	2.07 x 10^{1}	3.61	5.6	.17
7	Chromium	Cr	1.21 x 10^{0}	8.73 x 10^{0}	7.21	11.7	.69
8	Manganese	Mn	2.00 x 10^{1}	6.96 x 10^{1}	3.48	7.0	.15
9	Cobalt	Co	1.70 x 10^{2}	6.38 x 10^{2}	3.76	7.8	.09
10	Nickel	Ni	1.22 x 10^{2}	3.11 x 10^{2}	2.54	5.8	.16
11	Copper	Cu	4.73 x 10^{2}	4.50 x 10^{2}	0.95	1.9	.00
12	Zinc	Zn	2.88 x 10^{1}	1.00 x 10^{2}	3.49	7.4	.06
13	Gallium	Ga	2.02 x 10^{-1}	4.31 x 10^{-1}	2.13	4.0	.31
14	Germanium	Ge	5.55 x 10^{0}	4.42 x 10^{-1}	0.08	0.3	.00
15	Selenium	Se	8.60 x 10^{1}	8.15 x 10^{1}	0.95	2.5	.01
16	Niobium	Nb	3.12 x 10^{-1}	9.77 x 10^{-1}	3.13	4.0	.62
17	Molybedenum	Mo	1.76 x 10^{3}	4.55 x 10^{3}	2.58	3.8	.19
18	Cadmium	Cd	2.17 x 10^{0}	3.95 x 10^{0}	1.82	3.9	.15
19	Indium	In	5.21 x 10^{-2}	1.35 x 10^{-1}	2.59	8.5	.15
20	Tin	Sn	4.74 x 10^{-1}	1.42 x 10^{0}	3.00	8.9	.35
21	Tellurium	Te	1.75 x 10^{0}	6.52 x 10^{0}	3.72	6.5	.55
22	Tungsten	W	2.36 x 10^{0}	1.72 x 10^{1}	7.31	11.8	.41
23	Rhenium	Re	1.20 x 10^{-1}	2.46 x 10^{-1}	2.04	3.0	.54
24	Thallium	Tl	4.04 x 10^{1}	7.11 x 10^{1}	1.76	3.1	.01
25	Lead	Pb	2.96 x 10^{2}	5.72 x 10^{2}	1.93	2.5	.00
26	Bismuth	Bi	2.07 x 10^{0}	7.36 x 10^{0}	3.57	5.1	.40
27	Uranium	U	2.33 x 10^{-1}	1.06 x 10^{0}	4.54	8.7	.43

Table 2.2: Basic statistics on concentration (ppm) of each of the trace elements

There are a number of things to note from Table 2.2:

- In location (means), Arsenic, with $\bar{x}_3 = \bar{x}_{As} = 1.53 \times 10^4$ is a full order of magnitude higher than any other elements, followed by Molybdenum a full order of magnitude lower, with $\bar{x}_{17} = \bar{x}_{Mo} = 1.76 \times 10^3$ which is an order of magnitude higher than the remaining 25 elements.
- I include the coefficient of variation because due to these large changes in mean, there is a clear relationship between the mean and standard deviation, so the coefficient of variation provides a statistic that, as it is standardized by the sample mean, allows us to compare the variability between elements, even though they exist in very different scales.
- A major issue with this data is highlighted in the large values of skewness throughout, which in some cases is influenced by the large proportion of zero values observed, and this is an issue that will have to be addressed.
- The element with the highest proportion of zeroes is Chromium (Cr), with 69% zeros, closely followed by Niobium at 62% zeroes, and Uranium (U), Bismuth (Bi), Rhenium (Re), Tungsten (W) and Tellurium (Te) in the 40% – 60% zeroes range.

Each of these poses interesting issues in analyzing these data, as the difference in scale will overwhelm any measure based on Euclidean distance, and the high skewness and large proportion of zeros will throw off any method which assumes any symmetry or normality. Furthermore the zeros mean we need to be cautious in how we transform our data (as the usual transformation for such data would be a logarithm-transform, which is undefined at zero).

2.2.2 Boxplots

So now we take a look at the data, first we visualize the data in Boxplots (Figures 2.2, 2.3, & 2.4), still treating each element separately, and now separating by location, to see if we can pick up any differences in location visually, to see if it is worth investigating.

Immediately the high values of skewness and large proportion of zeros in some of the elements become very clear. An important note to make is that looking at these figures, any assumption of normality on this data would be quite unrealistic, unless perhaps we consider some type of hidden Gaussian Markov model [12]. However in comparing the three drill holes, there does seem to be a visually apparent trend, but it is likely to be insignificant when we realize that, although there is an absence of high values from Drill No. 3, this could likely be due to the small sample size (12).

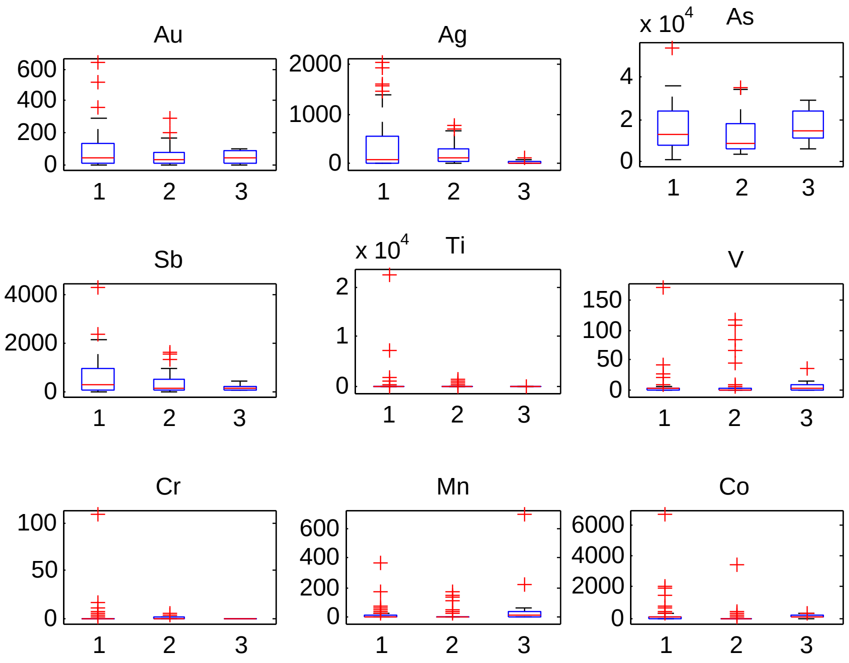


Figure 2.2: Boxplots of the concentrations (ppm) for elements 1 to 9 by drill no. as defined above

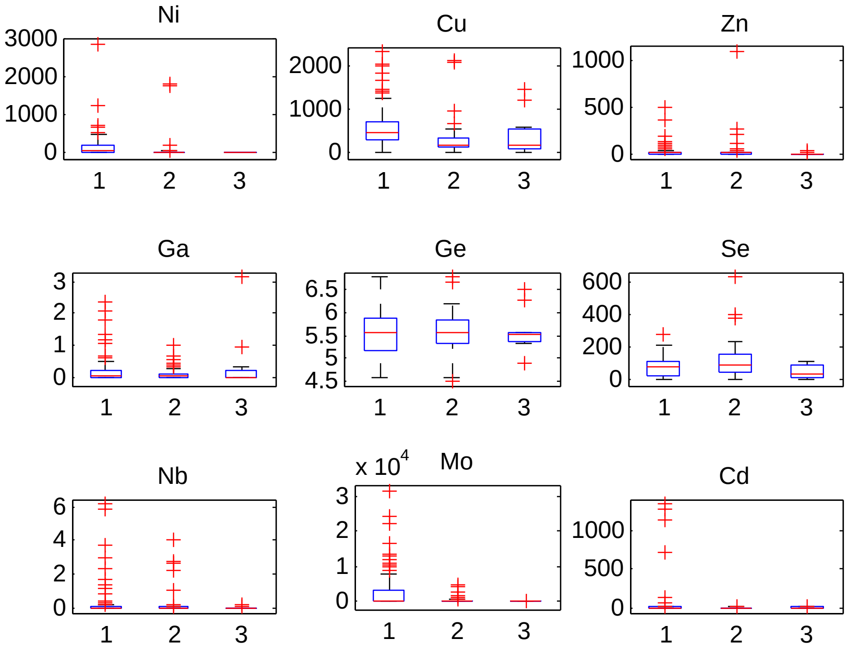


Figure 2.3: Boxplots of the concentrations (ppm) for elements 10 to 18 by drill no. as defined above

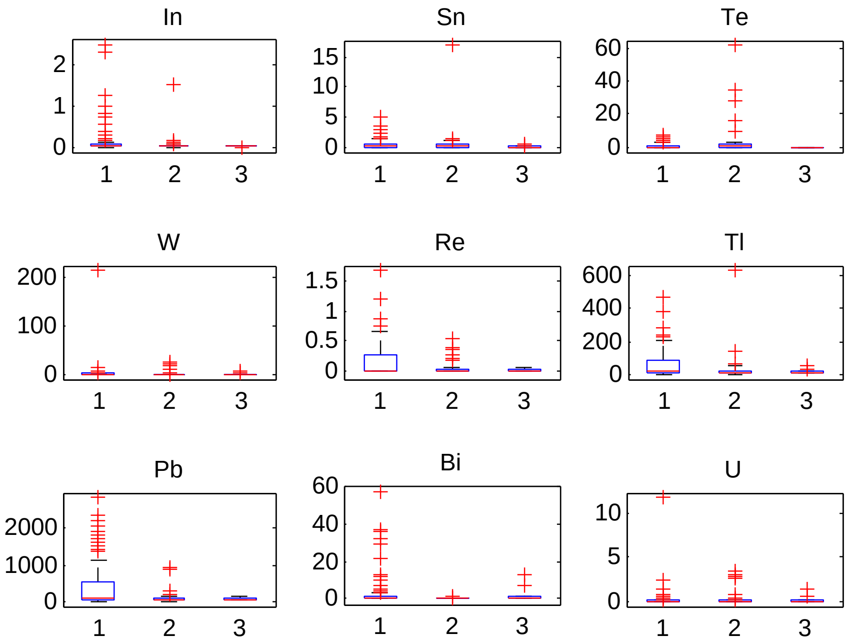


Figure 2.4: Boxplots of the concentrations (ppm) for elements 19 to 27 by drill no. as defined above

2.2.3 Depth Plots

Similarly, we want to separate the data by depth, to get a visual impression of if there could be any relevant trends in depth, as we did in the Boxplots for location, so here we have scatter plots of each element (still being treated independently) vs. Depth. Observe there does not seem to be any visually apparent trends here either, apart from that noted in the boxplots by location, i.e. the absence of high values from drill no. 3 (depth 664m).

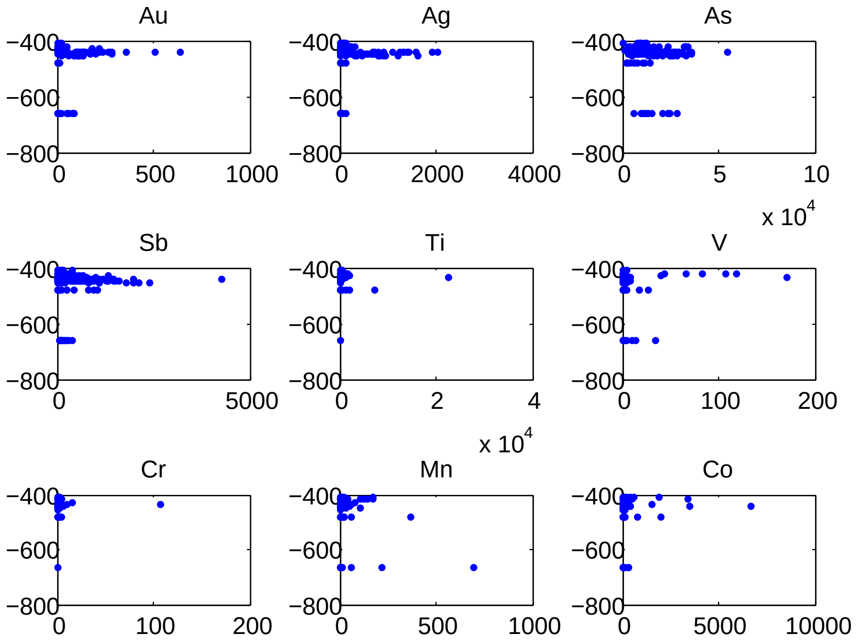


Figure 2.5: Plots of depth (m) vs. concentrations (ppm) of elements 1 to 9

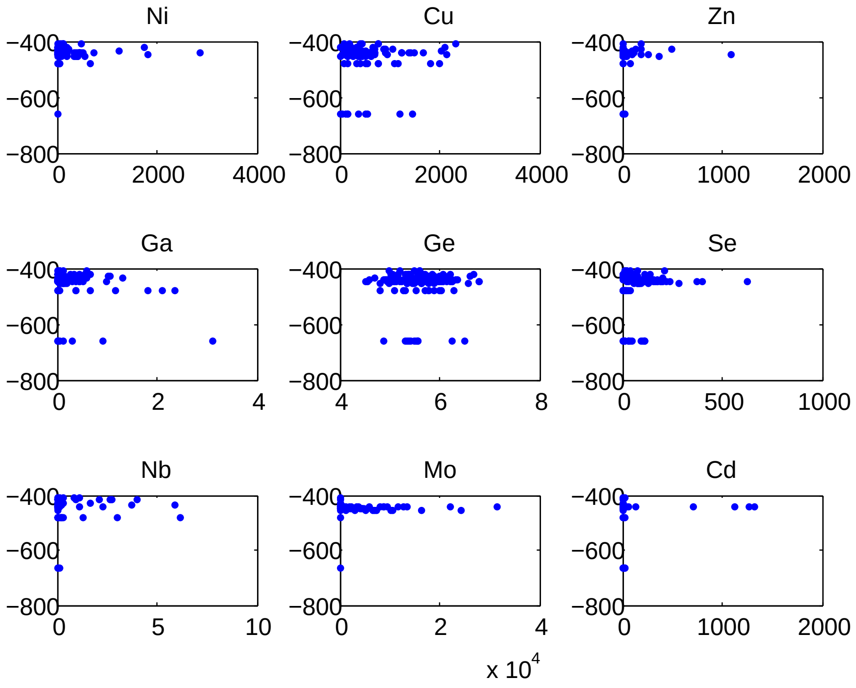


Figure 2.6: Plots of depth (m) vs. concentrations (ppm) of elements 10 to 18

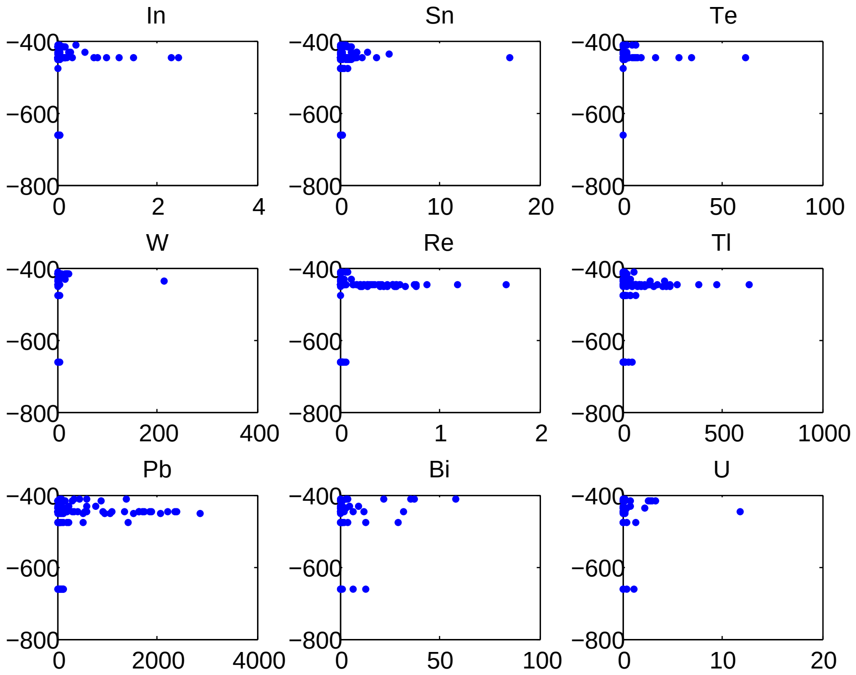


Figure 2.7: Plots of depth (m) vs. concentrations (ppm) of elements 19 to 27

2.3 Log-transformed Data

2.3.1 Univariate Statistics

We will also consider the log-transformed data and so this section will provide similar summary statistics, but for the log-transformed data $X^{(log)}$ as in Table 2.1. This type of trace element data is often considered on a log scale in the context of geochemistry and so it is interesting to consider the log-transformed data, and note how it changes the analyses. For example, we know that using the log-transformed data will in most cases reduce the influence of outliers on statistics that are outlier-sensitive, such as the Pearson's correlation coefficient, which will be used quite extensively in the following analyses.

				Standard	Coefficient of	
No.	Variable		Mean	Deviation	Variation	Skewness
1	Gold	Au	3.37	1.54	0.46	-0.57
2	Silver	Ag	4.63	1.5	0.32	0.11
3	Arsenic	As	9.36	0.76	0.08	-0.51
4	Antimony	Sb	5.2	1.52	0.29	-0.45
5	Titanium	Ti	1.92	2.24	1.16	4.83
6	Vanadium	V	0.77	1.06	1.38	2.29
7	Chromium	Cr	0.32	0.65	2.05	1.6
8	Manganese	Mn	1.43	1.49	1.04	2.16
9	Cobalt	Co	2.79	2.43	0.87	0.65
10	Nickel	Ni	2.81	2.28	0.81	0.5
11	Copper	Cu	5.72	1.04	0.18	-0.68
12	Zinc	Zn	1.95	1.43	0.73	1.72
13	Gallium	Ga	0.15	0.24	1.66	0.32
14	Germanium	Ge	1.88	0.07	0.04	0
15	Selenium	Se	4.04	1.17	0.29	-1.31
16	Niobium	Nb	0.13	0.37	2.78	0.74
17	Molybedenum	Mo	3.56	3.4	0.96	2.89
18	Cadmium	Cd	1.06	1.26	1.19	4.11
19	Indium	In	0.07	0.18	2.47	0.35
20	Tin	Sn	0.25	0.38	1.52	0.73
21	Tellurium	Te	0.42	0.73	1.73	1.52
22	Tungsten	W	0.33	0.74	2.21	2.34
23	Rhenium	Re	0.1	0.18	1.84	0.16
24	Thallium	Tl	2.57	1.66	0.64	0.41
25	Lead	Pb	4.18	1.75	0.42	0.75
26	Bismuth	Bi	0.4	0.83	2.09	2
27	Uranium	U	0.11	0.32	3	0.84

Table 2.3: Basic statistics on the log-transformed concentration for each of the trace elements

We see how in this case, the log-transformation we consider appears to somewhat compensate for the skewness of the data (note the much lower values of skewness), causing the log-transformed data to appear much more symmetric.

2.3.2 Boxplots

Now we see a significant change in the distribution of the data, as on the log scale, the elements with no, or a small proportion of zero values, appear to be much more spread out, however those with a significant proportion of zeros appear to still have a very strong positive skew.

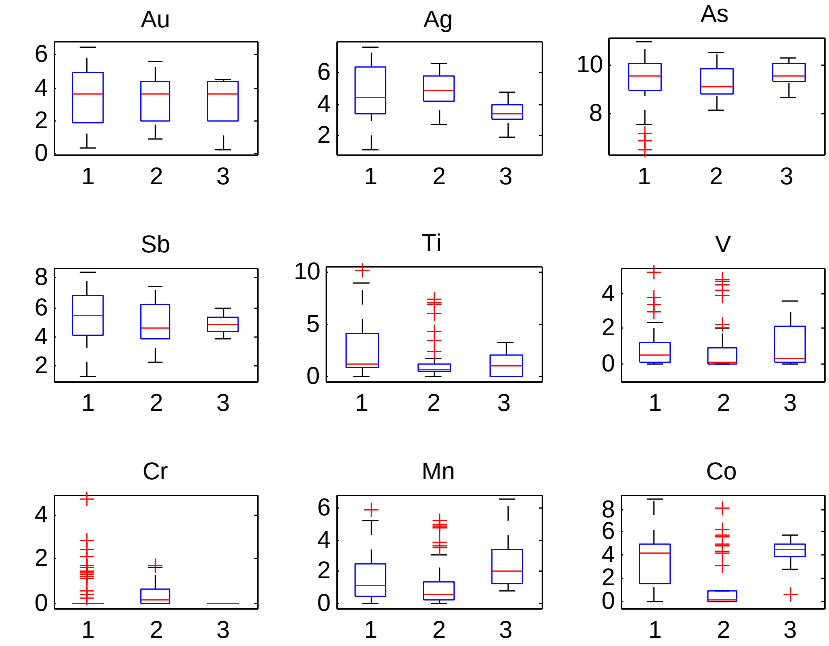


Figure 2.8: Boxplots of the log-transformed concentrations of elements 1 to 9

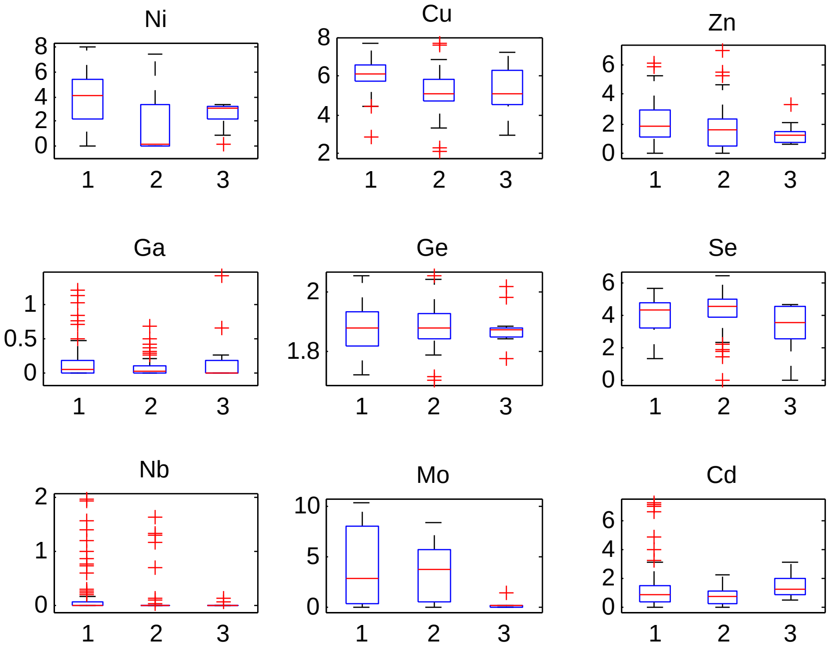


Figure 2.9: Boxplots of the log-transformed concentrations of elements 9 to 18

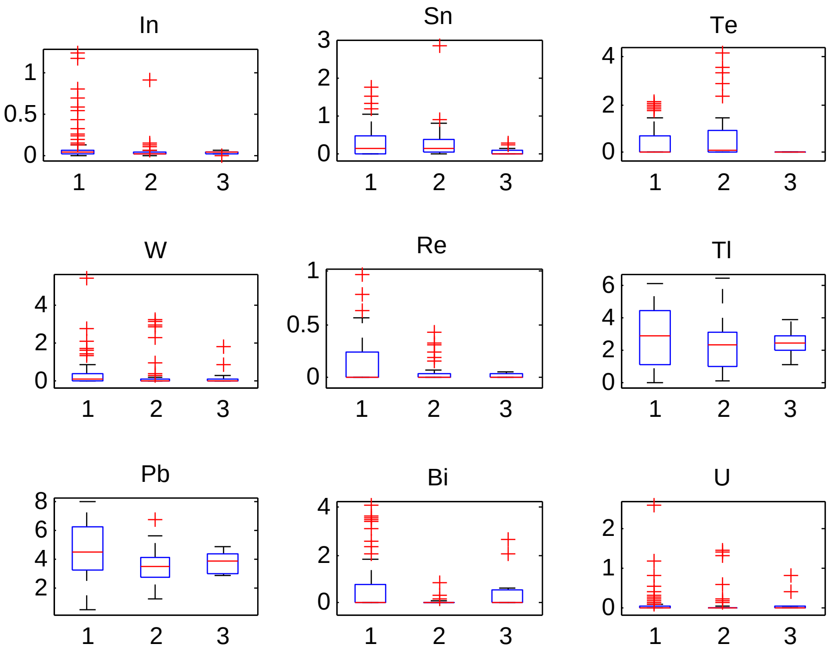


Figure 2.10: Boxplots of the log-transformed concentrations of elements 18 to 27

2.3.3 Depth Plots

And we note how there is still no notable trend with depth on the log scale, however we do see the same difference in skewness as in the boxplots appear here as well.

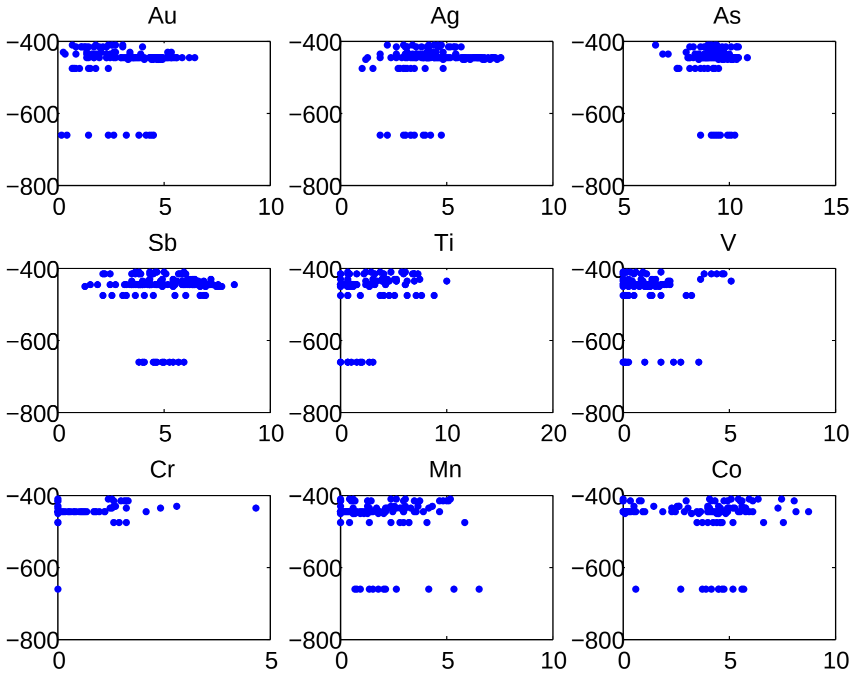


Figure 2.11: Plots of depth (m) vs. log-transformed concentrations (ppm) of elements 1 to 9

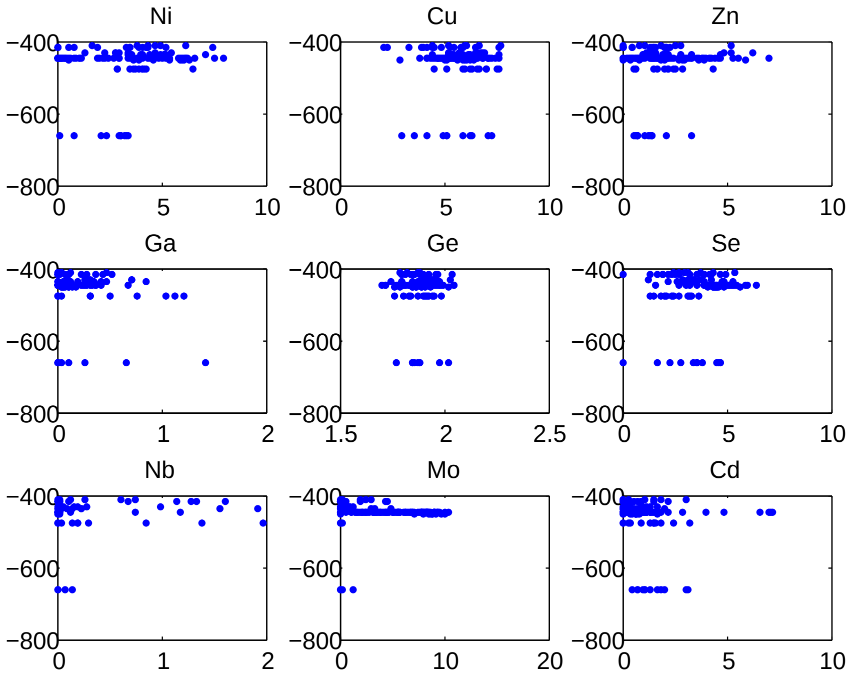


Figure 2.12: Plots of depth (m) vs. log-transformed concentrations (ppm) of elements 10 to 18

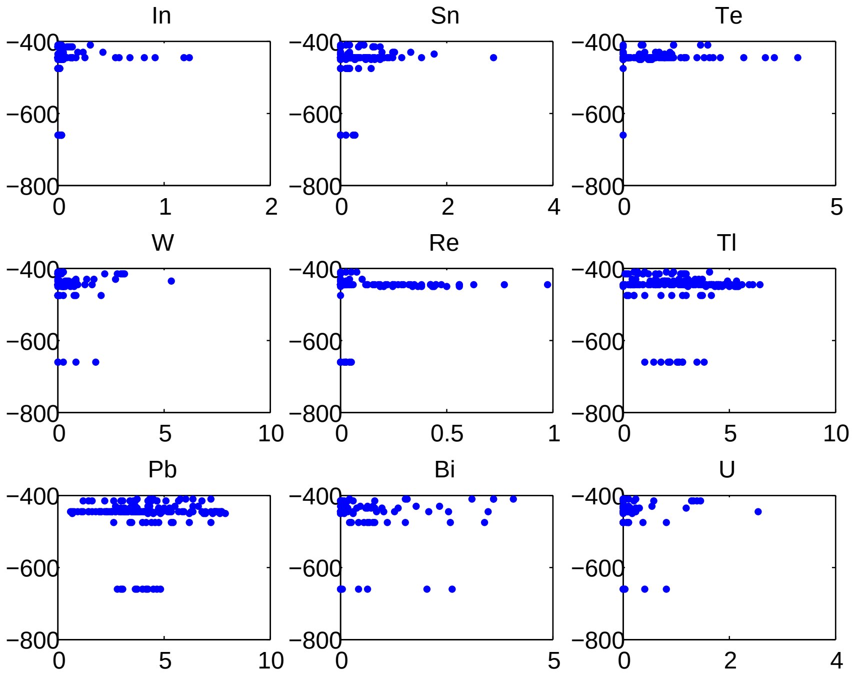


Figure 2.13: Plots of depth (m) vs. log-transformed concentrations (ppm) of elements 19 to 27

2.3.4 Normal QQ (quantile-quantile) Plots

So we see from Figures 2.14, 2.15 and 2.16 that a normal model (on the log-scale) seems appropriate for many of the variables with no, or very few zero values. However it does not seem appropriate for elements with a significant number of zeros. There does also seem to be some deviation from the log-normal model in some of the other trace elements (see Au) but the normality assumption turns out be not very important in most of the following analyses, so this is not cause for concern.

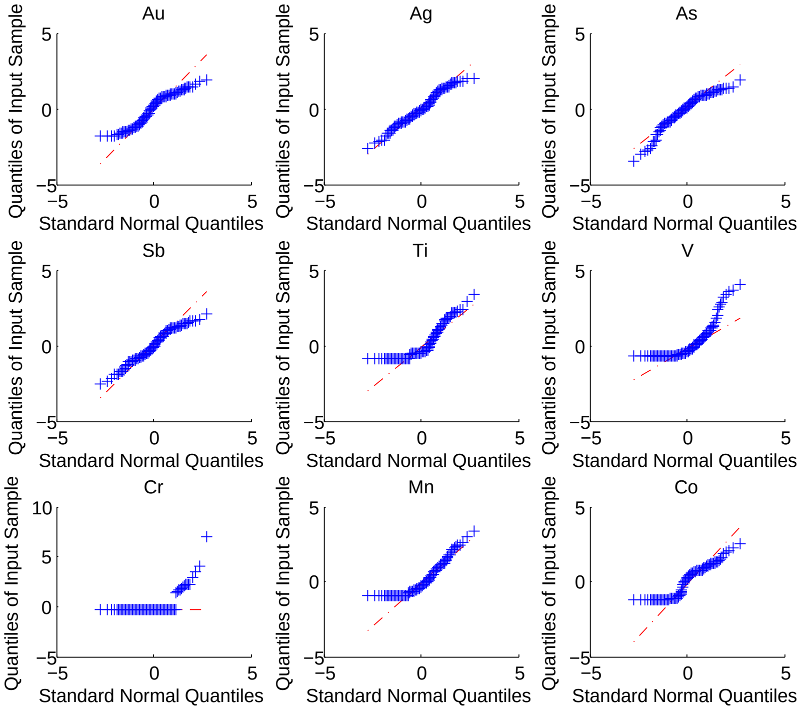


Figure 2.14: QQ Plots of the log Data Elements 1 to 9

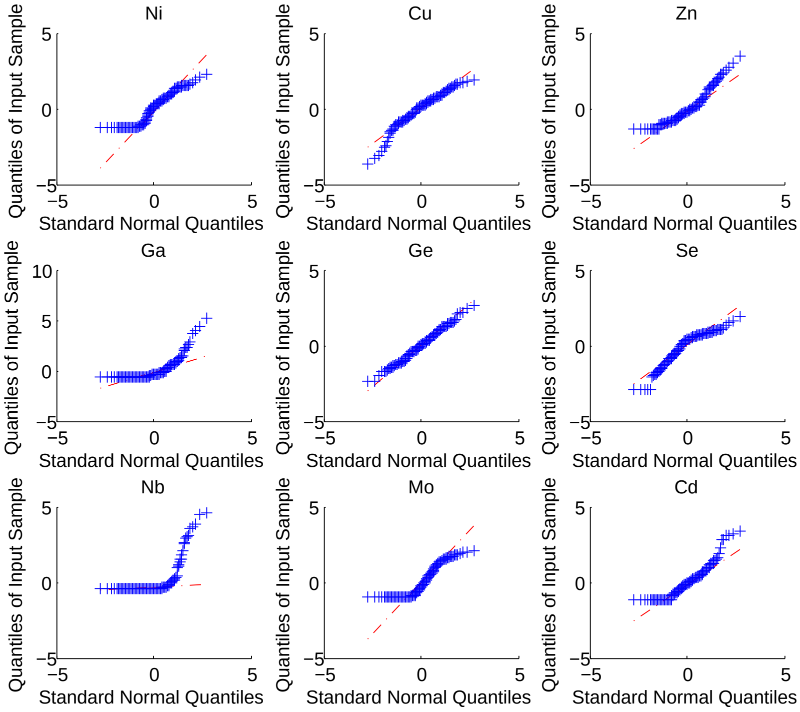


Figure 2.15: QQ Plots of the log Data Elements 9 to 18

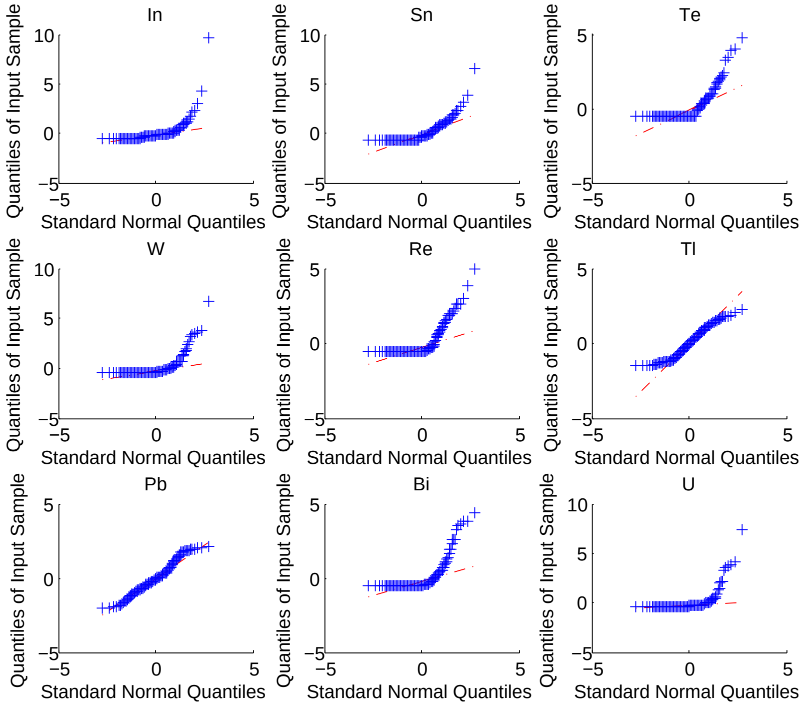


Figure 2.16: QQ Plots of the log Data Elements 19 to 27

Note the piecewise linear behavior apparent in many of these plots. This is caused by the zero values noted above, particularly the zero values contribute the perfectly linear, zero slope section to the left. Then the non-zero values contribute the remaining points. This suggests some kind of model for this data which incorporates this dual nature of the data.

2.4 A Zero-inflated Log-Normal Model

This would seem to indicate a good model to consider for this data might be some kind of latent variable model, similar to the truncated normal power model (2.1) suggested in [12],

$$x = \begin{cases} w^{\beta} & w > 0 \\ 0 & w \leq 0 \end{cases} \tag{2.1}$$

where w follows a normal distribution. The idea for the model I shall consider (2.2) comes from 2.1, as the distribution appears to have a similar heavy-tailed distribution, but adapted to this type of data (which is commonly known to follow an approximate log-normal distribution, in the absence of zero values), i.e. 2.2 has somewhat more physical meaning in this context than 2.1. Here we let x be the observed concentration of a particular trace element; then suppose that

$$x = \begin{cases} e^{w} & z > 0 \\ 0 & z \leq 0 \end{cases} \tag{2.2}$$

Where w similarly follows a normal distribution, and z independently follows a Bernoulli distribution with some probability parameter. This is effectively ignoring the zero values (saying they come from a separate distribution), and fitting a log normal model to the non-zero values. Fitting the model shown in Equation 2.2

gives us the QQ-plots shown below, which appear to be quite a reasonable fit to the data (at least for most of the trace elements).

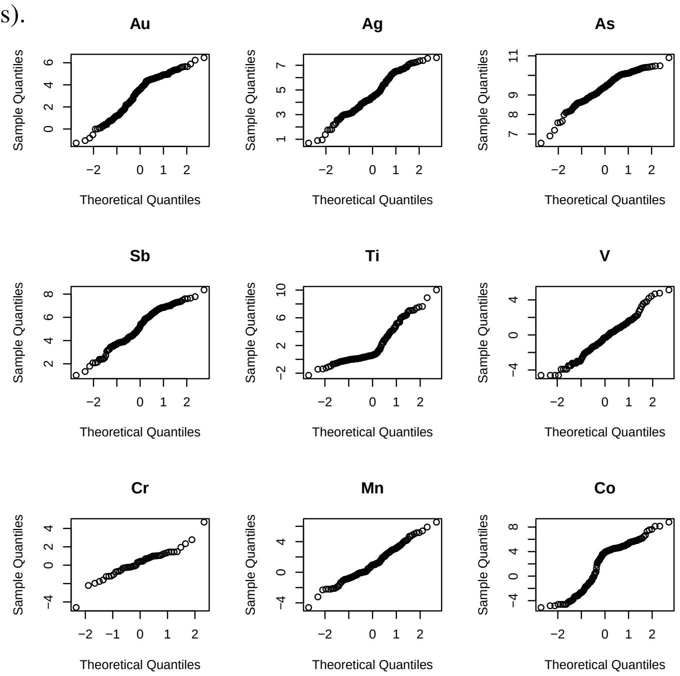


Figure 2.17: Log-Normal QQ Plots of the non-zero cases in Elements 1 to 9

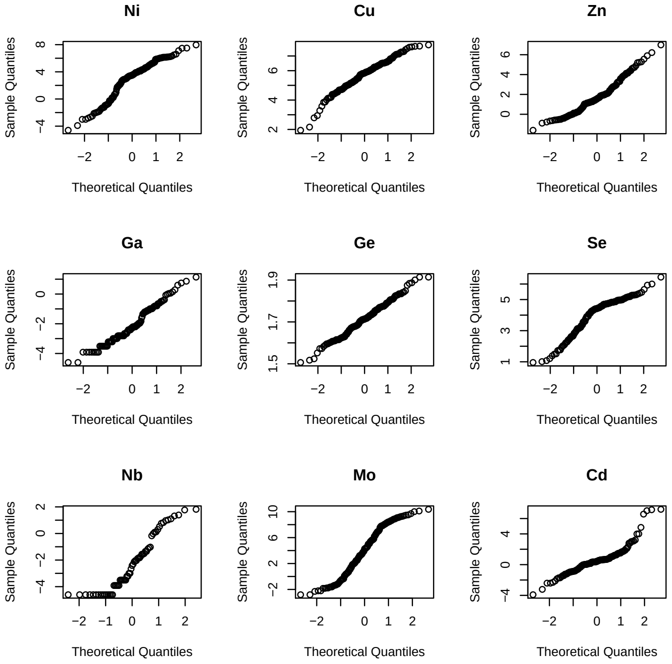


Figure 2.18: Log-Normal QQ Plots of the non-zero cases in Elements 10 to 18

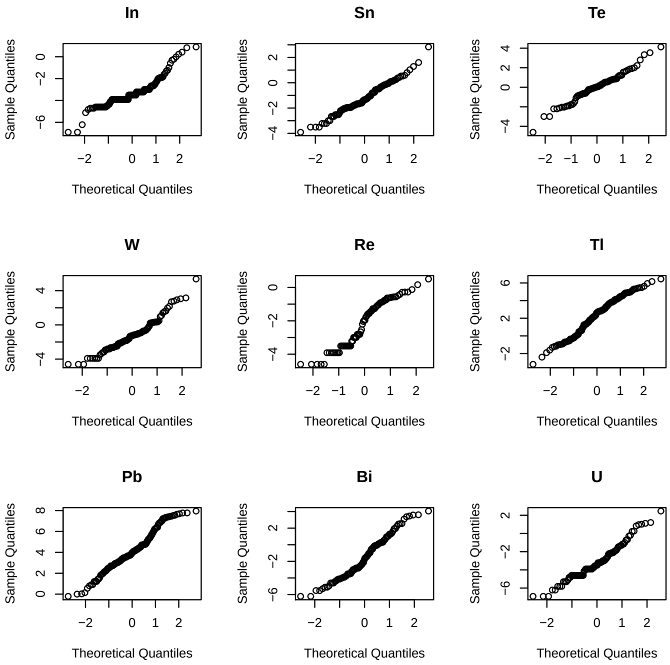


Figure 2.19: Log-Normal QQ Plots of the non-zero cases in Elements 19 to 27

Chapter 3

Correlations

3.1 Correlation Matrices

3.1.1 Pearson Product-moment Correlation Coefficient

Now that we have visualized the data for each element independently, we look at relationships between elements, so we start with the Pearson product-moment correlation coefficient,

$$r_{jk} = \widehat{corr}(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = \frac{\widehat{cov}(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k})}{s_j s_k} = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}} \tag{3.1}$$

as it is a measure of linear dependence between a pair of variables (so in this case we consider each trace element as a variable), this provides us with a basic investigative tool into associations between variables. However caution is required in interpreting these values, as they will only pick up on linear relationships, and not necessarily more complex relationships, as shown in figure 3.1.1, The Pearson's correlation coefficient is not a robust statistic, in that it is heavily influenced by outlying points, and heavy-tailed distributions, and these points will be addressed later.

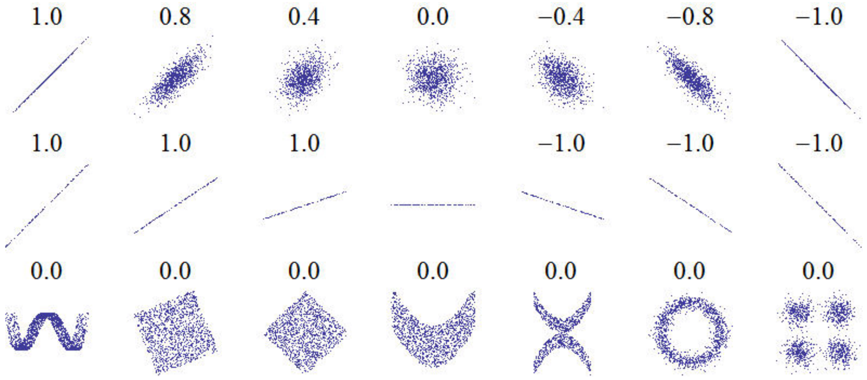


Figure 3.1: Examples of r values by Imagecreator at en.wikipedia [Public domain], from Wikimedia Commons

3.1.2 Raw Correlations

Tables 3.1 depict the pair-wise Pearson's correlation between the concentrations of the 27 trace elements in question (any values not less than 0.6 in bold font).

	Ag	As	Sb	Ti	V	Cr	Mn	Co	Ni	Cu	Zn	Ga	Ge
Au	0.32	**0.67**	0.24	-0.07	-0.14	-0.06	-0.17	-0.14	0.01	0.11	0.08	-0.19	-0.17
Ag	**1.00**	0.30	**0.76**	-0.04	-0.02	-0.03	-0.11	-0.06	0.30	0.05	-0.04	-0.10	0.09
As		**1.00**	0.23	0.03	0.18	0.06	-0.01	-0.13	0.03	-0.06	0.10	-0.18	0.01
Sb			**1.00**	0.05	0.03	0.07	-0.10	-0.01	0.27	0.03	-0.02	-0.01	0.10
Ti				**1.00**	**0.70**	**0.94**	0.08	0.00	0.02	-0.05	0.05	0.37	0.16
V					**1.00**	**0.70**	0.28	-0.01	-0.02	-0.03	-0.00	0.31	0.25
Cr						**1.00**	0.06	0.05	0.07	-0.05	0.05	0.28	0.15
Mn							**1.00**	0.11	-0.01	0.11	-0.03	0.31	0.23
Co								**1.00**	**0.89**	0.27	-0.04	0.14	-0.09
Ni									**1.00**	0.27	-0.04	0.07	-0.07
Cu										**1.00**	0.17	0.22	-0.03
Zn											**1.00**	-0.01	0.12
Ga												**1.00**	0.26
Ge													**1.00**

	Se	Nb	Mo	Cd	In	Sn	Te	W	Re	Tl	Pb	Bi	U
Au	0.32	-0.20	0.20	-0.14	0.17	0.09	0.05	-0.05	0.32	0.26	0.14	-0.19	0.07
Ag	0.20	-0.13	**0.85**	-0.02	-0.03	0.26	0.09	-0.01	**0.77**	**0.85**	**0.77**	-0.14	-0.07
As	0.14	-0.01	0.23	-0.17	0.16	0.08	0.05	0.10	0.30	0.28	0.19	-0.23	0.17
Sb	0.02	-0.05	0.49	0.11	-0.07	0.20	-0.02	0.08	**0.65**	**0.77**	0.52	-0.08	-0.06
Ti	-0.13	**0.68**	-0.06	-0.04	0.01	0.25	-0.04	**0.95**	-0.08	-0.04	-0.05	-0.01	0.22
V	-0.20	**0.67**	-0.06	-0.04	0.06	0.20	-0.06	**0.77**	-0.09	-0.04	-0.06	-0.03	0.43
Cr	-0.10	0.51	-0.05	-0.04	0.01	0.25	-0.03	**0.98**	-0.06	-0.03	-0.03	0.01	0.19
Mn	-0.20	0.20	-0.10	0.00	-0.02	-0.02	-0.06	0.07	-0.12	-0.10	-0.04	0.14	0.12
Co	-0.16	0.06	-0.05	0.18	-0.04	-0.05	-0.03	-0.02	-0.06	-0.01	0.29	0.37	-0.02
Ni	-0.09	0.02	0.28	0.08	-0.03	0.02	-0.00	0.01	0.22	0.31	**0.60**	0.21	-0.05
Cu	-0.13	-0.03	0.05	0.12	0.23	0.05	-0.11	-0.06	-0.01	0.04	0.23	0.26	-0.07
Zn	-0.03	-0.03	-0.03	0.06	0.32	0.11	-0.05	0.05	-0.04	-0.03	-0.03	-0.04	-0.04
Ga	-0.21	0.47	-0.08	0.12	-0.04	0.05	-0.10	0.25	-0.14	-0.08	0.03	0.20	0.15
Ge	-0.02	0.24	0.10	0.03	0.12	0.11	0.04	0.15	0.19	0.08	0.02	-0.09	0.05
Se	**1.00**	-0.28	0.16	-0.19	-0.03	0.08	**0.69**	-0.10	0.29	0.07	0.02	-0.19	-0.09
Nb		**1.00**	-0.12	-0.04	0.00	0.13	-0.08	0.56	-0.15	-0.11	-0.06	0.04	0.35
Mo			**1.00**	-0.04	-0.02	0.17	0.07	-0.04	**0.65**	**0.78**	**0.81**	-0.11	-0.08
Cd				**1.00**	0.02	-0.04	-0.07	-0.05	0.07	0.07	0.05	0.31	-0.05
In					**1.00**	0.10	-0.03	0.02	-0.05	-0.06	-0.04	-0.04	0.02
Sn						**1.00**	-0.03	0.26	0.09	0.19	0.11	-0.06	0.05
Te							**1.00**	-0.03	0.26	0.04	0.08	-0.00	-0.05
W								**1.00**	-0.06	-0.02	-0.04	-0.03	0.25
Re									**1.00**	**0.72**	0.53	-0.13	-0.08
Tl										**1.00**	**0.71**	-0.08	-0.08
Pb											**1.00**	0.19	-0.08
Bi												**1.00**	-0.04

Table 3.1: Correlations between the concentrations of trace elements

3.1.3 Log-transformed Correlations

Similarly, tables 3.1.3 depict the pair-wise Pearson's correlation between the concentrations of each the 27 trace elements on the log-transformed scale (similarly, any values not less than 0.6 highlighted in bold font).

	Ag	As	Sb	Ti	V	Cr	Mn	Co	Ni	Cu	Zn	Ga	Ge
Au	0.51	**0.61**	0.17	-0.54	-0.23	-0.21	-0.44	-0.39	-0.19	0.15	0.23	-0.26	-0.11
Ag	**1.00**	0.28	**0.72**	-0.28	0.16	-0.01	-0.16	-0.09	0.12	0.23	0.19	-0.04	0.09
As		**1.00**	0.00	-0.26	0.03	-0.01	-0.17	-0.23	-0.09	-0.09	0.09	-0.18	0.03
Sb			**1.00**	0.07	0.30	0.14	0.17	0.21	0.33	0.24	0.18	0.13	0.12
Ti				**1.00**	**0.67**	**0.64**	**0.72**	0.53	0.44	0.14	0.04	0.55	0.18
V					**1.00**	**0.61**	**0.66**	0.32	0.33	0.15	0.10	0.59	0.28
Cr						**1.00**	0.50	0.27	0.27	0.07	0.12	0.50	0.23
Mn							**1.00**	**0.61**	0.46	0.23	0.11	0.57	0.20
Co								**1.00**	**0.91**	0.35	-0.01	0.33	0.01
Ni									**1.00**	0.40	0.08	0.26	-0.00
Cu										**1.00**	0.55	0.24	0.06
Zn											**1.00**	0.20	0.11
Ga												**1.00**	0.27
Ge													**1.00**

	Se	Nb	Mo	Cd	In	Sn	Te	W	Re	Tl	Pb	Bi	U
Au	**0.76**	-0.32	**0.66**	-0.09	0.22	0.31	0.30	-0.21	0.39	0.22	-0.01	-0.39	-0.15
Ag	0.48	-0.14	**0.81**	0.15	0.05	0.41	0.29	0.08	**0.64**	**0.71**	0.34	-0.24	-0.03
As	0.30	-0.08	0.36	-0.27	0.24	0.25	0.19	0.07	0.26	0.10	-0.02	-0.36	0.09
Sb	0.09	0.07	0.54	0.39	-0.07	0.32	0.07	0.24	0.48	**0.89**	0.45	0.00	0.04
Ti	-0.52	**0.80**	-0.41	0.01	-0.09	0.09	-0.19	**0.68**	-0.24	0.02	0.23	0.44	0.50
V	-0.28	**0.64**	-0.04	0.05	-0.01	0.28	-0.07	**0.84**	0.03	0.26	0.25	0.12	**0.66**
Cr	-0.26	0.58	-0.18	0.01	0.02	0.22	-0.11	**0.70**	-0.17	0.09	0.09	0.24	0.46
Mn	-0.48	0.56	-0.36	0.16	-0.09	0.03	-0.15	0.52	-0.16	0.13	0.36	0.44	0.40
Co	-0.33	0.27	-0.19	0.22	-0.12	-0.05	-0.05	0.20	0.04	0.22	**0.67**	0.57	0.09
Ni	-0.19	0.23	0.07	0.11	-0.05	0.12	0.06	0.26	0.23	0.36	**0.78**	0.41	0.09
Cu	0.15	0.09	0.09	0.34	0.25	0.30	0.02	0.07	0.11	0.16	0.45	0.32	0.01
Zn	0.08	0.04	0.06	0.26	0.49	0.46	-0.09	0.06	0.01	0.03	0.19	0.03	-0.03
Ga	-0.26	0.53	-0.20	0.14	-0.03	0.16	-0.15	0.43	-0.16	0.05	0.19	0.35	0.33
Ge	-0.12	0.24	0.01	0.09	0.12	0.17	0.07	0.28	0.19	0.10	0.08	-0.09	0.17
Se	**1.00**	-0.42	0.58	-0.10	0.04	0.15	0.42	-0.34	0.35	0.14	-0.01	-0.24	-0.25
Nb		**1.00**	-0.26	-0.00	0.01	0.21	-0.15	**0.75**	-0.19	0.01	0.11	0.20	**0.64**
Mo			**1.00**	-0.06	-0.06	0.30	0.34	-0.11	**0.74**	**0.64**	0.24	-0.35	-0.15
Cd				**1.00**	0.09	0.05	-0.12	-0.03	0.08	0.32	0.15	0.25	-0.03
In					**1.00**	0.33	0.06	0.04	-0.06	-0.15	0.00	-0.10	0.05
Sn						**1.00**	0.08	0.33	0.22	0.26	0.21	-0.17	0.19
Te							**1.00**	-0.11	0.43	0.15	0.34	-0.01	-0.11
W								**1.00**	-0.05	0.17	0.12	0.01	**0.76**
Re									**1.00**	**0.61**	0.36	-0.23	-0.12
Tl										**1.00**	0.45	-0.06	-0.01
Pb											**1.00**	0.37	-0.02
Bi												**1.00**	-0.02
U													**1.00**

Table 3.2: Correlations between the log-transformed concentrations of the trace elements

3.2 Cluster Analysis

This section is largely based on [8, Chap. 5-6].

In this example, due to the relatively small number of variables (27), it is possible to consider the matrix of Pearson's correlation coefficients as shown in table 3.1 or 3.1.3. However it is (even at this relatively low dimensionality) laborious, and difficult to reach any meaningful conclusions by doing so. Furthermore, if the number of variables where much higher it would quickly become completely infeasible to do so. So we would like a better (more efficient, easier to interpret) method for presenting the information contained in the correlation coefficients, and for this purpose I suggest a hierarchical agglomerative cluster analysis dendrogram.

Cluster Analysis is, in broad terms, a method of describing and quantifying the structure of a set of data objects with respect to some distance measure (commonly Euclidean distance), and by what is called linkage, which I will explain below. Traditionally the objects being clustered are the 'observations', i.e. $\{\boldsymbol{x}_{1\bullet}, ..., \boldsymbol{x}_{n\bullet}\}$. But in this case, we are interested in the relationships between the variables, and so will cluster the trace elements, and use the notation $\mathcal{X} = \{\boldsymbol{x}_{\bullet 1}, ..., \boldsymbol{x}_{\bullet d}\}$ so in a sense, I will be treating the variables as 164 dimensional observations for this section. An example of the traditional approach to clustering of these data (clustering the actual observations, and treating them as 27 dimensional objects) will be covered in Chapter 6. It is worth noting that I use the notation $\mathcal{X} = \{\boldsymbol{x}_{\bullet 1}, ..., \boldsymbol{x}_{\bullet d}\}$ for the data objects (and $d = 27$ for the number of them, in this case the 'sample size'), in order to be consistent with these data, but for this section this can apply to any set of vector valued data objects.

3.2.1 Distance Measures

A distance measure provides information on 'how far apart' two data points are from each other, and is any function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the following properties $\forall j, k, m \in \mathbb{Z} \cap [1, d]$

- $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) \geq 0$
- $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet j}) = 0$
- $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = d(\boldsymbol{x}_{\bullet k}, \boldsymbol{x}_{\bullet j})$
- $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) \leq d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet m}) + d(\boldsymbol{x}_{\bullet m}, \boldsymbol{x}_{\bullet k})$

Which makes it a psuedometric. A psuedometric is a metric with the property $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = 0 \iff j = k$ relaxed to $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet j}) = 0$, i.e. $j = k \implies d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = 0$, but $d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = 0$ can hold when $j \neq k$.

By far the most common example is Euclidean distance, or the Euclidean norm:

$$d^{(Euclid)}(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = ||\boldsymbol{x}_{\bullet j} - \boldsymbol{x}_{\bullet k}|| = \sqrt{(\boldsymbol{x}_{\bullet j} - \boldsymbol{x}_{\bullet k})^T \cdot (\boldsymbol{x}_{\bullet j} - \boldsymbol{x}_{\bullet k})}$$

However what many people overlook is that any psuedometric can be used as a distance measure, and some non-Euclidean distances can often (depending on the context) be more meaningful than Euclidean distance. In particular in high-dimensional cases, for example (as suggested in [8]).

$$d^{(cosine)}(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = \arccos\left(\frac{\boldsymbol{x}_{\bullet j} \cdot \boldsymbol{x}_{\bullet k}}{||\boldsymbol{x}_{\bullet j}|| ||\boldsymbol{x}_{\bullet k}||}\right) = \arccos\left(\frac{\boldsymbol{x}_{\bullet j} \cdot \boldsymbol{x}_{\bullet k}}{\sqrt{\boldsymbol{x}_{\bullet j}^T \cdot \boldsymbol{x}_{\bullet j}}\sqrt{\boldsymbol{x}_{\bullet k}^T \cdot \boldsymbol{x}_{\bullet k}}}\right)$$

which measures the internal angle, or angle subtended at the origin between two vectors.

3.2.2 Linkage

Linkage then gives us information on 'how far apart' two *sets* of data points are, i.e. if $\mathfrak{X}$ is the set of all non-empty (non-trivial) subsets of $\mathcal{X}$ (i.e. $\mathfrak{X}$ is the power set of $\mathcal{X}$ without the empty set, $\mathfrak{X} = \mathcal{P}(\mathcal{X}) \setminus \phi$) then the linkage is a function $\mathcal{L} : \mathfrak{X} \times \mathfrak{X} \to \mathbb{R}$, which generally defines a rule involving the (chosen) distance measure. For $C_\alpha, C_\beta \subset \mathfrak{X}$ a common example (that I will use throughout) is single linkage

$$\mathcal{L}^{(single)}(C_\alpha, C_\beta) = min\left\{d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) : \boldsymbol{x}_{\bullet j} \in C_\alpha, \boldsymbol{x}_{\bullet k} \in C_\beta\right\}$$

There are several other commonly used measures of linkage including

- $\mathcal{L}^{(complete)}(C_\alpha, C_\beta) = max\left\{d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) : \boldsymbol{x}_{\bullet j} \in C_\alpha, \boldsymbol{x}_{\bullet k} \in C_\beta\right\}$
- $\mathcal{L}^{(average)}(C_\alpha, C_\beta) = \frac{1}{|C_\alpha||C_\beta|} \sum\limits_{\boldsymbol{x}_{\bullet j} \in C_\alpha} \sum\limits_{\boldsymbol{x}_{\bullet k} \in C_\beta} d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k})$

and several others (Wald, centroid) however in this context they all yield very similar results, and so for simplicity from here on I use single linkage. Single and complete linkage are the easiest to understand, as they are basically the smallest, or largest distance between the clusters respectively. Mean or average linkage is similarly simple, as it is the average of all the pairwise distances between the two clusters. I constructed an example in $\mathbb{R}^2$ with two clusters, one of 4 points (left), and the other of 3 (right) to illustrate these concepts. Figures 3.2 and 3.3 provide diagrams of the difference between single, and complete linkage.

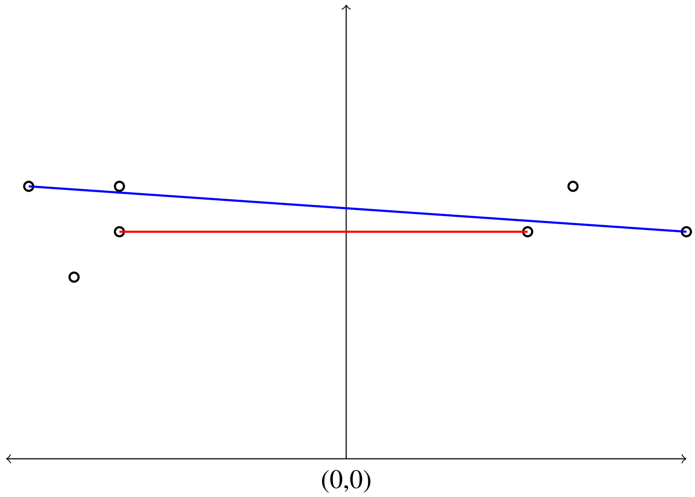


Figure 3.2: Illustrative example with two clusters, showing (using Euclidean distance) single linkage in red and complete linkage in blue.

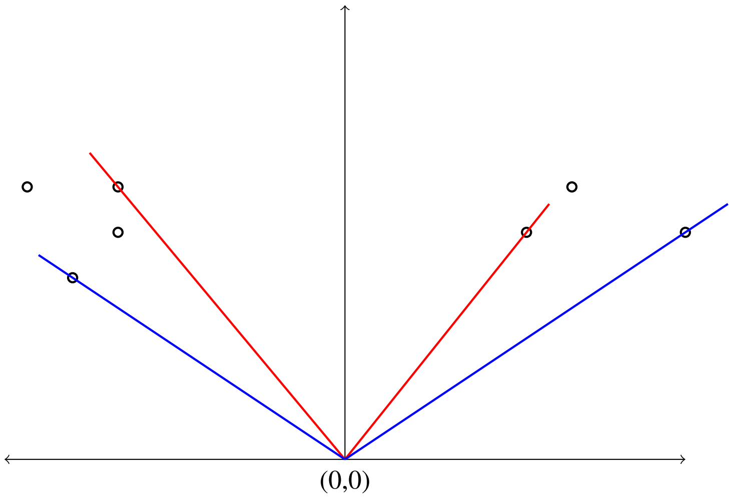


Figure 3.3: Illustrative example with two clusters, showing (using *cosine* distance) single linkage as the angle between the two red lines and complete linkage as the angle between the blue lines.

3.2.3 Hierarchical Agglomerative Cluster Analysis

As I mentioned, the method I recommend for visualizing these correlations is hierarchical agglomerative cluster analysis, this is a particular type of cluster analysis performed by using Algorithm 1.1.

```
Algorithm 1.1:
    Step 1.  Begin with $d$ singleton clusters, which form a partition of $\mathcal{X}$.
    Step 2.  Combine the two clusters with minimum linkage into a single cluster.
    Step 3.  IF(There remains more than one cluster)      Go back to Step 2.
             ELSE()                                       Stop.
```

A dendrogram is a way to visualize this process. It is worth noting that dendrograms can be constructed either vertically or horizontally and that these are equivalent, simply with the axes swapped. I will consistently use vertical dendrograms to avoid any confusion.

In a dendrogram

- Vertical lines represent clusters,
- Horizontal lines (connecting two vertical lines from below into one vertical line above) represent the joining of two clusters into one, and
- The height of the horizontal lines is equal to the value of the linkage of the two clusters that line joins.

3.3 Correlation-based Cluster Analysis

Here an agglomerative hierarchical cluster analysis was performed on the trace elements, using

$$d(\boldsymbol{x}_{\bullet j}, \boldsymbol{x}_{\bullet k}) = 1 - |r_{jk}| \text{ where } r_{jk} \text{ is the correlation as in 3.1} \tag{3.2}$$

for the distance measure between $\boldsymbol{x}_{\bullet j}$ and $\boldsymbol{x}_{\bullet k}$. This is very similar to the correlation distance in [8]), but slightly different (in that it uses the absolute value of the Pearson's correlation coefficient) and thus has a range of $[0, 1]$ rather than $[0, 2]$). I prefer this for the distance measure over the correlation distance as in [8] (which is the same as Equation 3.1 but without the absolute value) because it has a more intuitive, and meaningful interpretation in this particular context. That is, values close to 0 correspond to elements that are highly correlated (i.e. have a strong linear relationship) to each other (either positively or negatively, in this context most of the interesting correlations are positive anyway) while values closer to 1 can be interpreted as being nearly uncorrelated (or having a lack of any linear trend) and as such are described as 'further away' by the distance measure.

Using the distance 3.1, and single linkage, on the concentrations of the elements we obtain the dendrogram shown in Figure 3.4 below,

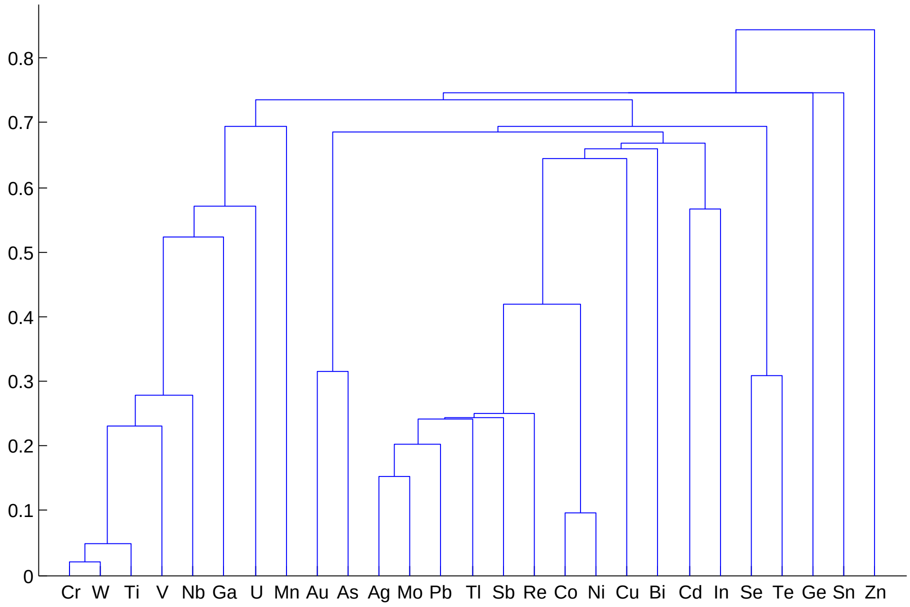


Figure 3.4: Dendrogram of Cluster Analysis on the concentrations of elements by *d* as in Equation 3.2

This provides us with a summary of all the correlation coefficients, combined into one figure, and many of the analyses in the following chapters will be compared back to it. There is a wealth of information about the data here, particularly a number of clusters of correlated trace elements in Figure 3.4 working from left to right

- The 'lithophilic' elements, Cr, W and Ti are very strongly correlated (and to a lesser degree V and Nb), which are then loosely associated with Ga and U, and seeing these elements clustered together is encouraging, as it seems to support the scientific theory.
- The Au, As pair shows up separate from everything else and this is interesting, as we would expect there to be a relation between them [10], but furthermore we show other elements are significant in prediction of Au in Chapter 7. So although this method is useful for visualizing the data, it does not pick up all the interesting relationships in the data, as for example in this context, information contained in partial correlations (which is picked up by the linear regression for Au in Chapter 7).
- The other major group that appears is composed of Ag, Mo, Pb, Tl, Sb and Re, which is loosely related to the Co, Ni pair.
- Se and Te appear as a separate pair with moderate correlation, which is interesting, as they are not related to any of the other groups in any significant way, and this is particularly interesting as they are not related to Ag, and Ag-Se minerals where observed [9], so we know they occur together in the rock, but what this seems to indicate is that they do not necessarily occur together in the pyrite, specifically.

We can get a clearer picture of these (in figure 3.5) by taking out the elements that are relatively uncorrelated to any of the other elements (Zn, Sn, Ge, Mn, Cd, In, Bi, and Cu) and running the analysis again.

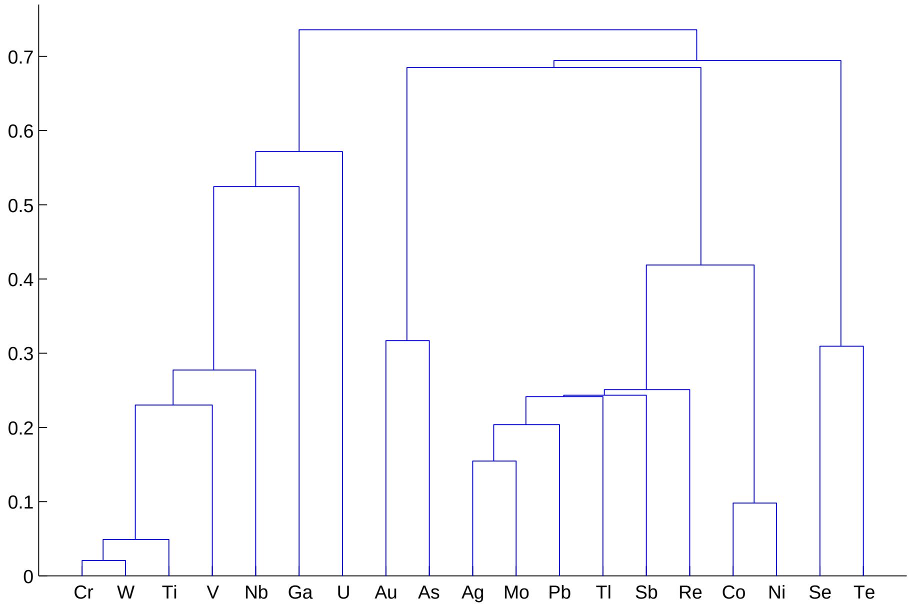


Figure 3.5: Dendrogram of Cluster Analysis on the concentrations of elements by d as in Equation 3.2 with elements that are only slightly correlated to the others removed

However, making conclusions based on these clusters should be done with extreme caution, as we have previously established these data are from distributions with quite a heavy tail, or contain some severe outliers, and it is well known that Pearson's correlation coefficient is highly influenced by outliers and heavy-tailed distributions, and so when making conclusions based on this analysis, this needs to be kept in mind, and in what follows I will investigate a number of methods for dealing with this.

The clearest example of how the correlations are affected by large values, is the Cr, W correlation shown in figure 3.4, referring to table 3.1, we see that $r_{Cr,W} \approx 0.98$ which in and of itself is somewhat suspicious, as we would not expect correlation that close to 1 between any two elements in this type of data, and rightfully so it turns out. If we calculate the contribution of each case to the correlation coefficient,

$$r_{jk}^{(i)} = \frac{(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2}} \text{ so } r_{jk} = \sum_{i=1}^{n} r_{jk}^{(i)} \tag{3.3}$$

Then we quickly discover that $r_{Cr,W}^{(25)} \approx 0.95$ and that $r_i(Cr, W) < 0.008 \;\forall\; i \neq 25$, and that if we recalculate the correlation coefficient without case 25, i.e.

$$r_{jk}^{(-q)} = \frac{\sum\limits_{i=1,i\neq q}^{n} (x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{\sqrt{\sum\limits_{i=1,i\neq q}^{n} (x_{ji} - \bar{x}_j)^2 \sum\limits_{i=1,i\neq q}^{n} (x_{ki} - \bar{x}_k)^2}} \tag{3.4}$$

and find that $r_{Cr,W}^{(-25)} \approx 0.52$ which is hugely different to $r_{Cr,W} \approx 0.98$, so it is clear that there are very influential points in these data (for example case 25, and likely others as well). So we are interested in identifying these interesting points (see Chapter 5), although we cannot justify removing them in any analysis, as they are still valid data. We are also interested in considering robust alternatives to the Pearson's correlation coefficient, and methods that reduce the effect of these 'outliers', to compare and see which correlations are due only to the effect of these large values, and which are not.

3.3.1 Log-transformed Data

Intuitively, using the log-transformed concentrations should dramatically reduce (if not entirely compensate for) the effect of such outliers. Figure 3.6 was produced by the same method as figure 3.4, but applied to the log-transformed concentrations.

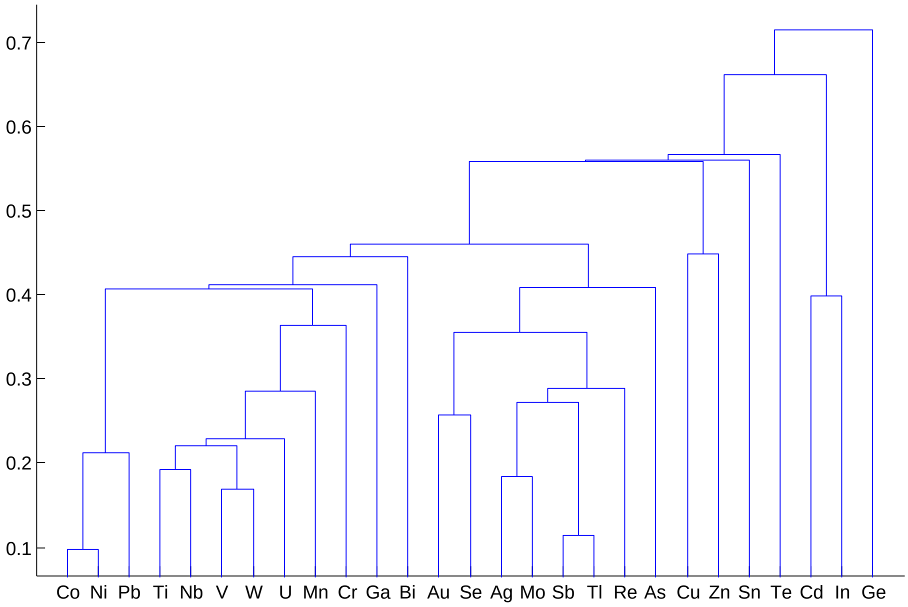


Figure 3.6: Dendrogram of Cluster Analysis on the log-transformed concentrations of the elements by $1 - |r_{jk}|$

Notice the distinctive difference between figures 3.4 and 3.6, how many of the clusters in figure 3.4 noted above are entirely gone in 3.6, and new clusters appear, this is interesting, because for data without extreme values we would generally expect the log-transformed data to have similar correlations as the raw data, but here we have an example of when this is not the case, at all. This might be due to the reduction of the influence of large values, but this has not been demonstrated (and is quite difficult to demonstrate), and also the values of several of the correlations are actually higher after the log-transform, not lower. Regardless, it certainly warrants further investigation. In the log-transformed data, we observe that

- The 'lithophilic' elements, Ti, Nb, V, W (and U) are still in a cluster (although with quite a different internal structure), and are still associated with Mn and Cr.
- The Co, Ni pair is much more strongly correlated in the log-transformed data, and is now associated with Pb, and this cluster of three elements is loosely associated with the 'lithophilic' elements, as well as Ga.
- The other large cluster including Ag, Mo, Sb, Tl, and Re, is also similarly composed (although also with quite different internal structure), but is now loosely associated with the pair Au, Se and, separately, As.

There are also quite a few elements that don't seem to have a significant correlation to any of the other elements, so removing Ge, Te, Sn, Zn, Cu, Bi, Cd and In (elements not strongly correlated to any of the other elements) to get a clearer picture of the structure, yields figure 3.7.

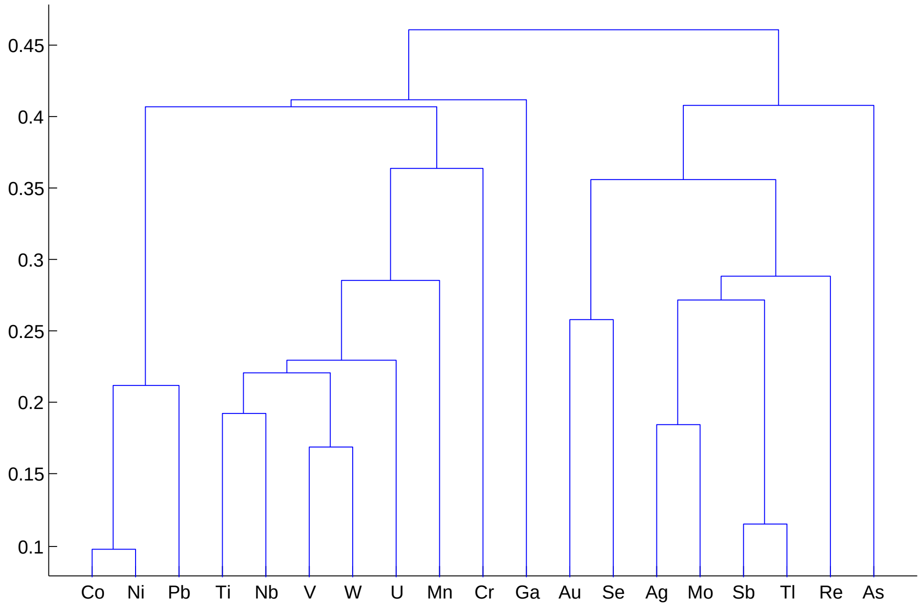


Figure 3.7: Dendrogram of Cluster Analysis on the log-transformed concentrations of the elements by $1 - |r_j k|$ with elements that are only slightly correlated to the others removed

3.3.2 Rank-based Measures of Association

Another straightforward method for compensating for the heavy-tailed distribution or outliers, is to convert the observations x_{ij} to ranks y_{ij} (tied values are assigned their average rank), and to use a rank based measure of association, such as Spearman's rho

$$\rho_{jk} = \frac{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2}} \tag{3.5}$$

Using the distance measure 3.2 but with ρ (3.5) instead of r (3.1), we produce the dendrogram shown in figure 3.8. Kendall's tau statistic could also be used, so it is worth noting that using Kendall's tau produces a dendrogram with no significant differences to that produced using Spearman's Rho, and as such I will use Spearman's Rho as a rank based measure of association over Kendall's Tau from here on, mainly because it is the rank-based equivalent to Pearson's r, which I use extensively.

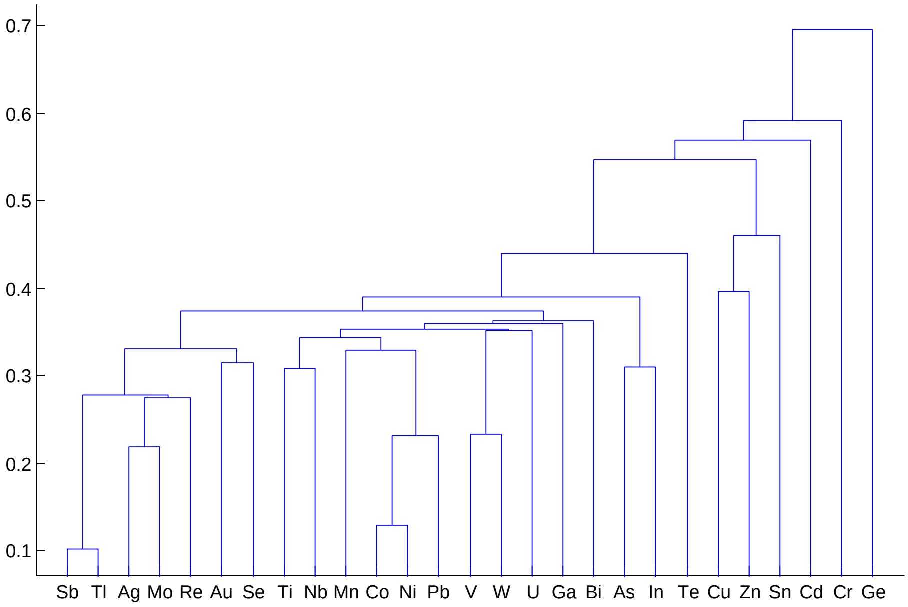


Figure 3.8: Dendrogram of Cluster Analysis on the concentrations of the elements by $1 - |\rho_{jk}|$

Similarly, we can remove the relatively dissociated elements (Ge, Cr, Cd, Sn, Zn, Cu and Te) from Figure 3.8 to get a better picture below in Figure 3.9.

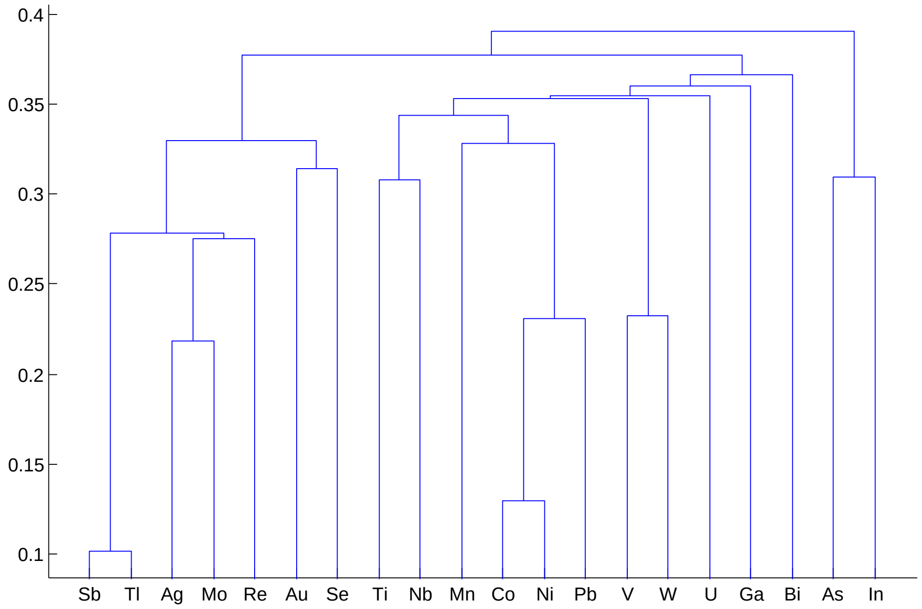


Figure 3.9: Dendrogram of Cluster Analysis on the concentrations of the elements by $1-|\rho_{jk}|$ with elements that are only slightly correlated to the others removed

Interpretation of these results is less straightforward than for Pearson's correlation, but still yields interesting results, particularly

- Co and Ni still appear as a strongly associated pair,
- The lithophilic elements W, U, V, Nb, Ti still appear in the same cluster, although
- Cr is now absent, as it appears to be (relatively) dissociated from the other elements now.
- Sb and Tl appear as a very strongly associated pair as a part of the cluster of elements
- Sb, Tl, Mo, Re, Au and Se which appear together.

Some of these results are interesting such as the Sb, Tl pair, and the absense of other results previously noted is particularly interesting as well, for example the Cr, W and Ti group noted in Figure 3.5. However interpretation of these results should again, be made with care, as the interpretation of Spearman's correlation coefficient differs quite significantly from that of Pearson's, specifically a value of 1 (or -1) is acheived for Pearson's correlation coefficient when one element can be expressed as a *linear* function of the other (with positive, or negative slope respectively), while Spearman's on the other hand acheives a value of 1 (or -1) if one element can be expressed as a *monotonic* (increasing, or decreasing respectively) function of the other.

3.4 Summary

Some of the interesting correlations noted in this chapter are listed below in Table 3.3.

Elements		Correlation Coefficient		
		Raw	Log	Spearmans
Cr	W	0.98	0.70	0.30
Cr	Ti	0.94	0.64	0.30
Ti	W	0.95	0.68	0.55
Au	As	0.67	0.61	0.44
Se	Te	0.69	0.42	0.56
Co	Ni	0.89	0.91	0.87
Sb	Tl	0.77	0.89	0.90
Ag	Mo	0.85	0.81	0.78

Table 3.3: A summary table of some interesting correlations

The effect of case 25 on the Cr, W correlation is quite apparent here in that the correlation is reduced on the log-scale, and even more dramatically reduced in the Spearman's Rho coeficient. Furthermore the Cr, W, Ti group noted in Figure 3.5 is shown to follow the same trend, which indicates it may be an artifact of an extreme value. The Co, Ni and Sb, Tl correlations are particularly interesting however, as they remain very strong throughout. We take a much more detailed look at these relations in Chapter 5, but for now these basic observations will suffice. The primary acheivement of this chapter is the visualization of the large correlations matrix C in the relatively easy to interpret form of a dendrogram.

Comparing Figures 3.5, 3.7 and 3.9, and the respective dot points of notable features of each highlights some interesting trends across the three. Perhaps the clearest of these being the lithophilic group of elements: Cr, W, Ti, Nb, V, and U (featuring the distinctive artifact (caused by case 25) of the trio Cr, W, and Ti being very highly correlated in the raw data) which also includes Mn in the raw data, and Ga in the log-transformed data, but share the other six elements mentioned accross both. This group, and in particular Cr, are missing in the dendrogram constructed using Spearman's rank-based measure of association, which may likely be due to the tied rank zero values in Cr, as well as the difference in interpretation of the Spearman's coefficient causing the Spearman's correlation coefficient being quite small.

Other interesting features include the Co, Ni and Sb, Tl pairs, which appear very highly correlated by all measures of association, which is interesting as Co and Ni have a well known, expected, relationship from geology, but to my (limited) knowledge Sb and Tl may not, and thus may be of geological interest. Thus this is the type of result i would report back to the geologists for interpretation or further investigation. The Sb, Tl pair mentioned above appears as a part of a larger cluster of mutually correlated elements in the raw data including the Co, Ni pair, as well as an Ag, Mo pair, Re, and Pb. However in the log-transformed data, the Co, Ni and Sb, Tl pairs are no longer in the same cluster. The Co, Ni pair appear as a seperate cluster with Pb, while the Sb, Tl and Ag, Mo pairs still appear in a large cluster of elements along with Re, Au, and Se.

Chapter 4

Exploratory Analyses

4.1 Introduction

In previous chapters we have considered the individual trace elements as independent variables, and pairwise correlations and associations between them (which is essentially a method for considering 2 variables at a time). By these methods we have gained some insight into the general shape and structure of the data as well as identified some particular features, such as the large values noted in Chapter 2 and investigated further in Chapter 3. Now we want to start looking at the data as a whole, at the overall structure, without limiting ourselves to the (unrealistic) idea that the relations in the data are going to be simple (such as those entirely describable by pairwise correlations). This chapter will cover a number of different ways of doing this, each of which provides insight into different aspects of the data (and interestingly similar patterns crop up across several different methods: particularly the ones based primarily on the correlations matrix, which I then compare to my visualization of the pairwise correlations in the form of dendrograms from Chapter 3). This turns out to be particularly interesting as *in this case* several of these 'more advanced' methods actually just boil down to the same information as the pairwise correlations, perhaps just providing somewhat different methods for visualizing them, or as in the case of factor analysis, a slightly different interpretation, which as it turns out still lead to very similar conclusions, and this is interesting in and of itself.

4.2 Parallel Coordinate Plots

In order to visualize 2, 3, or even 4 or 5 dimensional data, we can produce scatterplots (extending to multiple scatterplots, and/ or adding colour to define more than 2 dimensions). These can (to varying degrees of effectiveness) display all the information contained in the data. However when the dimension exceeds these small numbers, conventional plots are no longer sufficient. This fact is very nicely demonstrated by the sheer number of plots in Chapter 2, which only succeed in visualizing the relationships between the variables depth and location to the concentrations of the trace elements (and perhaps not even that so well) without showing the relationship between trace elements at all (and this alone takes well over fifty plots). The actual number of 2-dimensional scatter plots needed to fully display a d-dimensional dataset (assuming no use of colour or other methods) is $\binom{d}{2} = \frac{1}{2}d^2 - \frac{1}{2}d = O(d^2)$, i.e. the number of pairwise choices of variables is of quadratic power, which in this case (with $d = 27$) is the same as the number of correlation coefficients, 351, or if we include depth and drill no., and class of pyrite we have $d = 30$ and $\binom{30}{2} = 435$. This provides a nice illustration of just how many scatterplots it would take to visualize this data properly (enough that it would not matter anyway, because it would be nigh impossible to interpret that many plots simultaneously, which we would need to do in order to get a legitimate idea of if there was the presence of any complex patterns involving multiple variables.

A preliminary attempt to solve this, proposed in [8], is to use parallel coordinate plots. It is worth noting that, similar to dendrograms, parallel coordinate plots can be produced as either the vertical or horizontal variety, and that these are entirely analogous, up to a swap of axis. From here on I will only use the horizontal variety, for consistency and to avoid confusion. Parallel coordinate plots are described in more detail in [7], but briefly: In a parallel coordinate plot we plot the dimensions of the data (we treat the 27 trace elements as the dimensions, in this case, as compared to how we treated them as the data objects in the cluster analysis) on the horizontal axis, and the values of the data on the vertical axis (usually on the same scale. This can be generalized to having different scales for different variables (which may be necessary depending on the nature of the variables). The default in MATLAB is to plot all the variables on the same scale, while in R the default is to plot each variable individually on its own scale (from its minimum to its maximum). I use MATLAB to produce these plots, so all the parallel coordinate plots presented are on one scale for all the the variables (and this can be done as all the trace elements are measured in the same units of concentration). Then each case is connected with a line that passes through each dimension (represented in this case by a vertical line) exactly once, at the value that that case takes on that dimension.

So a parallel coordinate plot of the concentrations of the trace elements does provide a method to visualize all the information the data contains in a single plot, somewhat analogously to a scatterplot for 2 dimensional data. But as we shall see, it is not necessarily displayed in a way that is useful, or particularly friendly to easy interpretation. To illustrate this, consider Figure 4.1, which shows us exactly what we would expect from what we have already seen from Table 2.2, i.e. the orders of magnitude difference between the concentrations of As, then Mo, and then Ti, overwhelm all the other information in the data, when plotted on the same scale, as noted in Chapter 2.

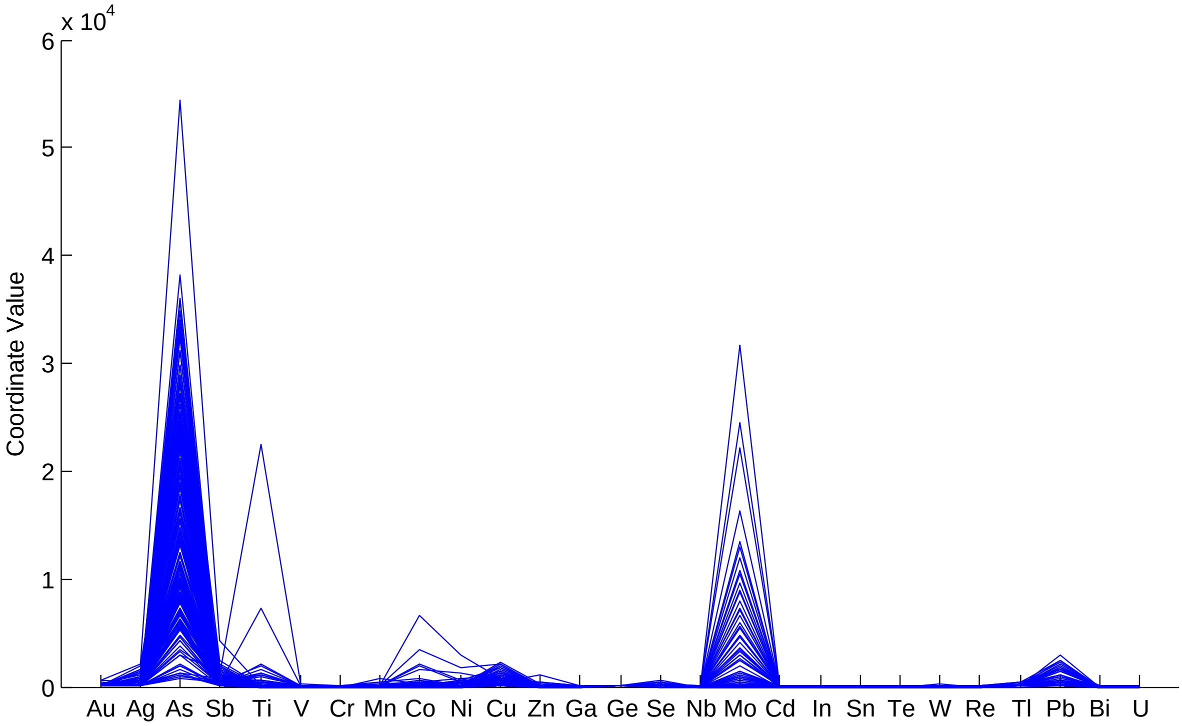


Figure 4.1: Parallel coordinate plot of the concentrations of trace elements

One method for getting a better look at the data, considering this scale difference between variables is, as suggested in [8], to standardize the data first, and then produce the parallel coordinate plot. When I say standardize I mean first center (subtract the mean), then divide each trace element separately by its sample standard deviation. This provides a very similar (but not actually exactly equivalent) plot (see Figure 4.2) to the default method used by R to produce parallel coordinate plots (as R plots the unscaled data, but on a scale that goes from the minimum value, to the maximum value, of each variable separately).

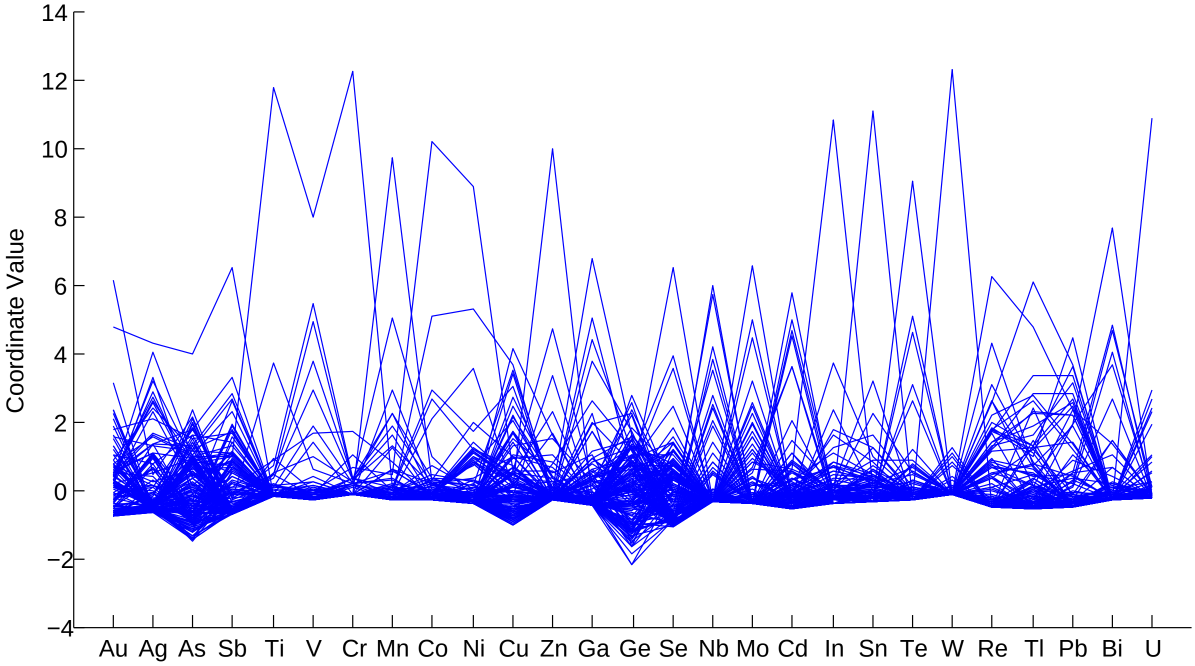


Figure 4.2: Parallel coordinate plot of the standardized concentrations of trace elements

The other transformation worth considering, as mentioned in previous Chapters, is the log-transformed concentrations, of which the parallel coordinate plot is shown in Figure 4.3, and for the standardized log-transformed concentrations shown in Figure 4.4.

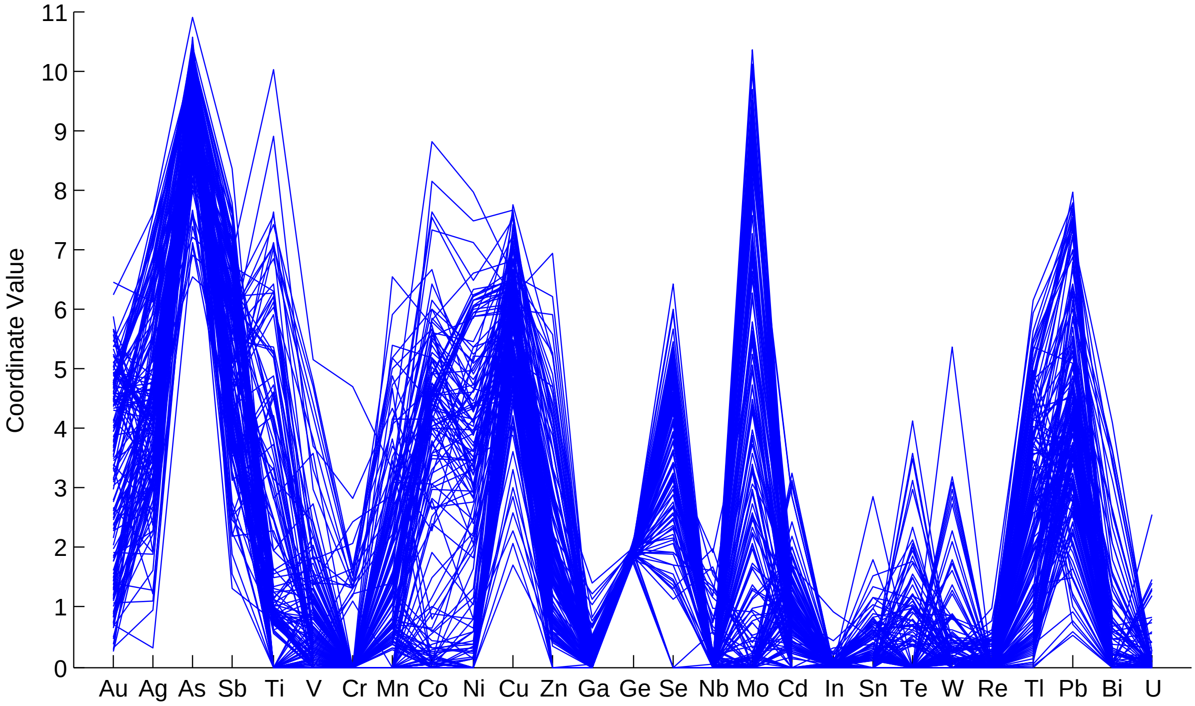


Figure 4.3: Parallel coordinate plot of the log-transformed concentrations of trace elements

From Figure 4.3 we see that the variables are comparable to each other (on the same scale) on the log-transformed scale, which is good, as this was the intention when transforming the data. Figure 4.3 also nicely highlights which variables have zero values and which don't, as due to the nature of the transformation zero gets mapped to zero (this is why I chose the $x \rightarrow ln(x + 1)$ transformation specifically), for example we can see that As has no values anywhere near zero very clearly from this plot.

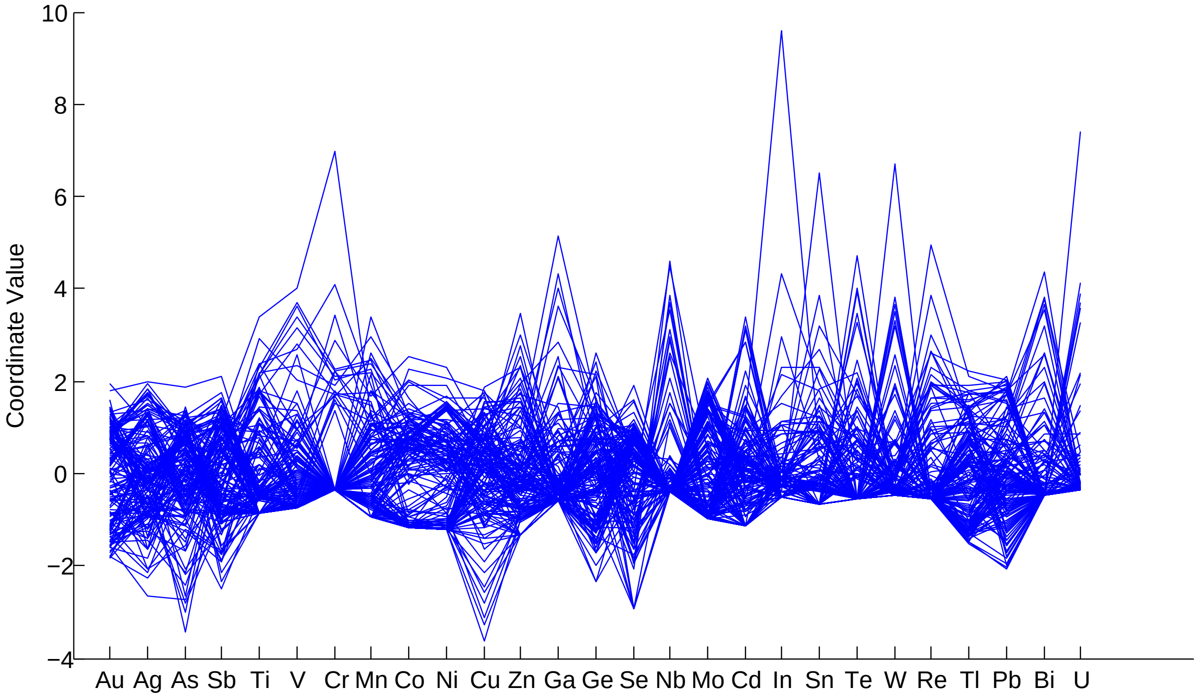


Figure 4.4: Parallel coordinate plot of the standardized log-transformed concentrations of trace elements

However Figure 4.4 demonstrates that although the log-transformed data is on the same scale now, extreme values are still present, as much as eight standard deviations from the mean of individual elements. Figure 4.4 also highlights an interesting property of Cr, in that it not only has a large number of zero values (69% as noted earlier), but the non-zero values are mostly several standard deviations from the mean. As we can see, these parallel coordinate plots provide a great deal of insight into the nature of this data, but it is easy to imagine that there is much more that they fail to show us.

4.3 Principal Component Analysis

Another method for displaying such high dimensional data is to find 'interesting' directions [8] in the data space (in this case 27 dimensional space, so directions in this space correspond to linear combinations of the trace elements) and to project the data into these directions. Most of the rest of this chapter (and thesis, in fact) will be devoted to investigating different such directions, and their usefulness (due to their interpretations).

4.3.1 PCA - Principal Component Analysis

This section draws many of its results from [8, Chap. 2]

In PCA we identify the direction vectors (unit length vectors representing the coefficients of linear combinations of our variables) that yield maximum variance, and are orthogonal. So the first Principal Component,

$\boldsymbol{\eta}_1$ is a unit vector representing the linear combination of the variables that yields the largest variance, then the second Principal Component, $\boldsymbol{\eta}_2$ is the unit vector, orthogonal to $\boldsymbol{\eta}_1$, which yields the largest variance, and so on, each Principal Component orthogonal to the rest. If we consider the data as existing in 27-dimensional space, span$\{\boldsymbol{\eta}_1\}$ is the 1-dimensional subspace which when the data is orthogonally projected into it, gives the largest variance.

It can be shown that these can easily and efficiently be obtained as the eigenvectors of the covariance matrix Σ of the data (demonstrated in the Section below), where the $(j,k)^{th}$ element of Σ is $cov(\boldsymbol{X}_j, \boldsymbol{X}_k)$. As Σ is positive definite in non-trivial cases, we know that its eigenvectors and eigenvalues exist and and are unique up to sign. i.e. $\Sigma = \Gamma^T \Lambda \Gamma$ where $\Gamma = [\boldsymbol{\eta}_1|\boldsymbol{\eta}_2|...|\boldsymbol{\eta}_d]$ and Λ is a diagonal matrix of eigenvalues $\lambda_1, \lambda_2, ..., \lambda_d$. This is then quite naturally extended to the sample case where we estimate Σ with S where the $(j,k)^{th}$ element of S is $\widehat{cov}(\boldsymbol{X}_j, \boldsymbol{X}_k) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$, and we estimate the principal components $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_d$ with the eigenvectors of S, $\hat{\boldsymbol{\eta}}_1, \hat{\boldsymbol{\eta}}_2, ..., \hat{\boldsymbol{\eta}}_d$ corresponding to the eigenvalues of S, $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_d$

4.3.2 Demonstration of Results

This section provides a proof of the property mentioned above (that these direction vectors $\boldsymbol{\eta}_i$ can be obtained as the eigenvectors of the covariance matrix Σ), this result, and more are in [8, Chap. 2]. I will only show the population case here, but all these results generalize very intuitively to the sample case, and this is shown in [8].

From linear combinations of random variables we know that for

$\boldsymbol{X}$ a length $d = 27$ column vector of random variables with $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$, $\quad \mathrm{var}(\boldsymbol{X}) = \Sigma$ (covariance matrix)

that

$$\mathrm{E}[\boldsymbol{a}^T\boldsymbol{X}] = \boldsymbol{a}^T\boldsymbol{\mu} \quad \mathrm{var}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\boldsymbol{a} \quad \mathrm{cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\boldsymbol{b}$$

for $\boldsymbol{a}, \boldsymbol{b}$ length $d = 27$ vectors of constants. Now we know that for non-trivial, $d < n$ (the $d > n$ case is explained in [8]) cases, Σ will have rank d, and so will be diagonalizable

$$\Sigma = \Gamma\Lambda\Gamma^T \tag{4.1}$$

where Λ is a diagonal matrix of the eigenvalues $\lambda_1, ...\lambda_d$ (in decreasing order, without loss of generality) corresponding to the eigenvectors in the columns of Γ ($\boldsymbol{\eta}_1, ..., \boldsymbol{\eta}_d$ as above). So,

$$\begin{aligned}
\mathrm{var}(\Gamma^T\boldsymbol{X}) = \mathrm{var}\begin{pmatrix} \boldsymbol{\eta}_1^T\boldsymbol{X} \\ \vdots \\ \boldsymbol{\eta}_d^T\boldsymbol{X} \end{pmatrix} &= \mathrm{E}[\Gamma^T(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T\Gamma] \\
&= \Gamma^T\,\mathrm{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T]\Gamma \\
&= \Gamma^T\Sigma\Gamma \\
&= \Gamma^T\Gamma\Lambda\Gamma^T\Gamma && (\text{ Equation 4.1}) \\
&= I_{d\times d}\Lambda I_{d\times d} \\
&= \Lambda
\end{aligned}$$

I.e.

$$\mathrm{cov}(\boldsymbol{\eta}_i^T\boldsymbol{X}, \boldsymbol{\eta}_j^T\boldsymbol{X}) = \begin{cases} 0 & \text{for } i \neq j \\ \mathrm{var}(\boldsymbol{\eta}_i^T\boldsymbol{X}) = \lambda_i & \text{for } i = j \end{cases} \tag{4.2}$$

It is worth noting that $\mathrm{var}(\Gamma^T\boldsymbol{X}) = \mathrm{var}(\Gamma^T(\boldsymbol{X} - \boldsymbol{\mu}))$, and $\Gamma^T(\boldsymbol{X} - \boldsymbol{\mu})$ is referred to in [8] as the principal component scores, and are basically the centralized data projected into each of the principal component directions, which if we consider all of them corresponds to a proper rotation (change of axis) of the (centralized) data, but we can also consider subsets of them (such as for example just the first two principal components, then we can produce a scatterplot/ biplot, as shown later in this chapter). As $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ..., \boldsymbol{\eta}_d$ are an orthonormal basis, we can write any unit vector of constants $\boldsymbol{a}$ as

$$\boldsymbol{a} = \sum_{j=1}^{d} c_j \boldsymbol{\eta}_j \text{ such that } \sum_{j=1}^{d} c_j^2 = 1 \tag{4.3}$$

We define the first principal component as the vector of constants, $\boldsymbol{e}_1$ (I will show this is equal to $\boldsymbol{\eta}_1$), such that var($\boldsymbol{e}_1^T \boldsymbol{X}$) maximizes var($\boldsymbol{a}^T \boldsymbol{X}$) over all unit vectors $\boldsymbol{a}$.

$$\begin{aligned}
\text{var}(\boldsymbol{a}^T \boldsymbol{X}) &= \boldsymbol{a}^T \Sigma \boldsymbol{a} \\
&= \sum_{j=1}^{d} \sum_{k=1}^{d} c_j c_k \boldsymbol{\eta}_j^T \Sigma \boldsymbol{\eta}_k && \text{Equation 4.3} \\
&= \sum_{j=1}^{d} c_j^2 \boldsymbol{\eta}_j^T \Sigma \boldsymbol{\eta}_j && \text{Equation 4.2} \\
&= \sum_{j=1}^{d} c_j^2 \lambda_j && (4.4)
\end{aligned}$$

Now by our assumptions that $\lambda_1 > \lambda_2 > ... > \lambda_d$, and $|\boldsymbol{a}|^2 = \sum_{j=1}^{d} c_j^2 = 1$, we get that

$$var(\boldsymbol{a}^T \boldsymbol{X}) = \sum_{j=1}^{d} c_j^2 \lambda_j \leq \sum_{j=1}^{d} c_j^2 \lambda_1 = \lambda_1$$

This reaches equality when $c_j = \begin{cases} 1 & j = 1 \\ 0 & j \neq 1 \end{cases}$, i.e. $\boldsymbol{e}_1 = \underset{\boldsymbol{a}}{\arg\max} \left\{ \text{var}(\boldsymbol{a}^T \boldsymbol{X}) \right\} = \boldsymbol{\eta}_1$. Which is the result we wanted to show for the first principal component, that it is equal to the first eigenvector of the covariance matrix (or the eigenvector corresponding to the largest eigenvalue of the covariance matrix). For the subsequent principal components, the argument is similar. Take $\boldsymbol{e}_2$, the unit vector that maximizes var($\boldsymbol{a}^T \boldsymbol{X}$) in $\boldsymbol{a}$ given the restraint that $\boldsymbol{e}_2$ is orthogonal to $\boldsymbol{e}_1 = \boldsymbol{\eta}_1$. If we write $\boldsymbol{a}$ as in Equation 4.3 we can see that this restraint translates to $c_1 = 0$ as this is equivalent to a vector of this form being orthogonal to $\boldsymbol{\eta}_1$, i.e. having no component in the direction of $\boldsymbol{\eta}_1$. Given this restraint, we now solve the same optimization problem as before, i.e. $\boldsymbol{e}_2 = \underset{\boldsymbol{a}}{\arg\max} \left\{ var(\boldsymbol{a}^T \boldsymbol{X}) \right\}$

$$\begin{aligned}
var(\boldsymbol{a}^T \boldsymbol{X}) &= \sum_{j=1}^{d} c_j^2 \lambda_j && \text{Equation 4.4} \\
&= \sum_{j=2}^{d} c_j^2 \lambda_j && c_1 = 0
\end{aligned}$$

This is then similarly solved as

$$var(\boldsymbol{a}^T \boldsymbol{X}) = \sum_{j=2}^{d} c_j^2 \lambda_j \leq \sum_{j=2}^{d} c_j^2 \lambda_2 = \lambda_2 \text{ as } \sum_{j=1}^{d} c_j^2 = \sum_{j=2}^{d} c_j^2 = 1$$

which similarly reaches equality when $c_j = \begin{cases} 1 & j = 2 \\ 0 & j \neq 2 \end{cases}$, i.e. $\boldsymbol{e}_2 = \underset{\boldsymbol{a}}{\arg\max} \left\{ var(\boldsymbol{a}^T \boldsymbol{X}) \right\} = \boldsymbol{\eta}_2$. The proof for the rest of the principal components follows from this in exactly the same manner, i.e. $\boldsymbol{e}_k = \boldsymbol{\eta}_k \quad \forall k$, as for any k, $\boldsymbol{e}_k = \underset{\boldsymbol{a}}{\arg\max} \left\{ var(\boldsymbol{a}^T \boldsymbol{X}) \right\}$ given the restraint that $\boldsymbol{a}$ is a unit vector, and is orthogonal to the set of vectors $\{\boldsymbol{e}_1, ..., \boldsymbol{e}_{k-1}\}$. So assuming that $\boldsymbol{e}_j = \boldsymbol{\eta}_j$ for $j < k$, this implies $c_1 = ... = c_{k-1} = 0$, and

$$var(\boldsymbol{a}^T \boldsymbol{X}) = \sum_{j=1}^{d} c_j^2 \lambda_j = \sum_{j=k}^{d} c_j^2 \lambda_j \leq \sum_{j=2}^{d} c_j^2 \lambda_k = \lambda_k \text{ as } \sum_{j=1}^{d} c_j^2 = \sum_{j=k}^{d} c_j^2 = 1$$

which reaches equality when $c_j = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$, i.e. $\boldsymbol{e}_k = \underset{\boldsymbol{a}}{\arg\max} \left\{var(\boldsymbol{a}^T\boldsymbol{X})\right\} = \boldsymbol{\eta}_k$ So as we have already shown the $k = 1$ and $k = 2$ cases, by induction we have the result we required (that the principal components, or unit vector solutions to this recursive optimization problem are equal to the eigenvectors of the covariance matrix in order of decreasing eigenvalues). □

4.3.3 PCA on the Raw Concentrations

We would expect, from observing Table 2.2, and Figure 4.1, PCA on the raw data to be completely overwhelmed by the order of magnitude differences in the scale of the variables (As, Mo, then Ti specifically). As intuition would suggest, this is what we observe in the first few Principal Components (i.e. the first principal component is almost entirely influenced by As, and a little by Mo, the second mostly by Mo, and a little by As, and the third largely by Ti):

```
> summary(princomp(X))
Importance of components:
                             Comp.1       Comp.2       Comp.3       Comp.4
Standard deviation     9563.9215477 4360.7052526 1.851834e+03 7.972358e+02
Proportion of Variance    0.7943971    0.1651502 2.978309e-02 5.520002e-03
Cumulative Proportion     0.7943971    0.9595473 9.893304e-01 9.948504e-01

.
.
.

> princomp(X)$loadings

Loadings:
   Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
Au                                                         -0.177
Ag                                0.314                0.218 -0.858
As -0.990  0.143
Sb                                0.926        -0.186 -0.102  0.283
Ti               -0.999
V
Cr
Mn                                                                -0.107
Co                         0.816 -0.106  0.305 -0.274
Ni                         0.388         0.118
Cu                         0.293        -0.940 -0.146
Zn                                                                 0.986
Ga
Ge
Se                                                         -0.233
Nb
Mo -0.141 -0.979
Cd                                                     0.955  0.267
In
Sn
Te
W
Re
Tl
Pb                         0.303  0.117         0.921 -0.117
```

```
Bi
U

.
.
.
```

I use the R output here directly to present these results, as I think it does so quite nicely, very small values are not displayed, and significantly non-zero values are displayed to three decimal places. This makes seeing the important contributions quite easy, in particular we see that η_1 is (as a linear combination of the trace elements) primarily composed of As, with a small contribution from Mo, η_2 is composed primarily of Mo, with a small contribution from As, and η_3 is virtually just make up of Ti. To further illustrate this point, we can construct a scree plot (figure 4.5) which shows the proportion of variation in the data explained by the first principal component, the second, and so forth, (these values correspond to the associated eigenvalues, as shown above) and this quickly drops (as expected) due to the order of magnitude differences in variances.

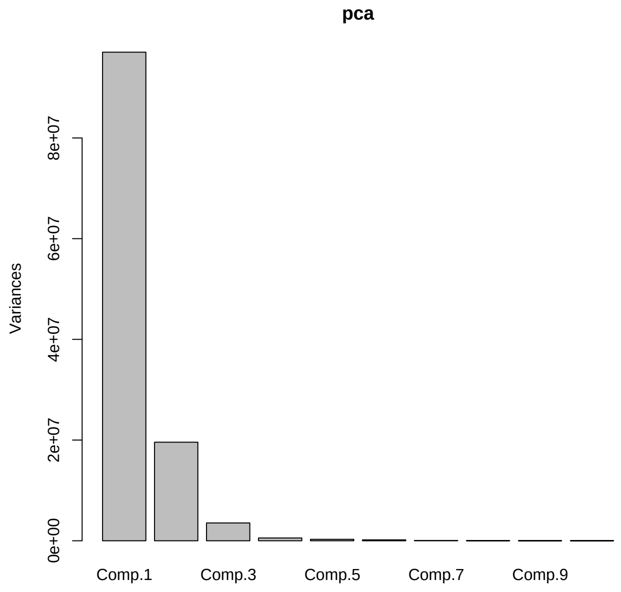


Figure 4.5: Scree plot of the Principal Component Analysis of the raw data.

Another way to display this analysis would be to construct a parallel coordinate plot of what are called the principal component scores, which are what we get if we project the data into the directions of the principal components (as in [8]). Doing this yields Figure 4.6, where the variables (or dimensions) being plotted now are the principal components (i.e. linear combinations of the elements), rather than the trace elements themselves, so effectively it is a change of axes on the data.

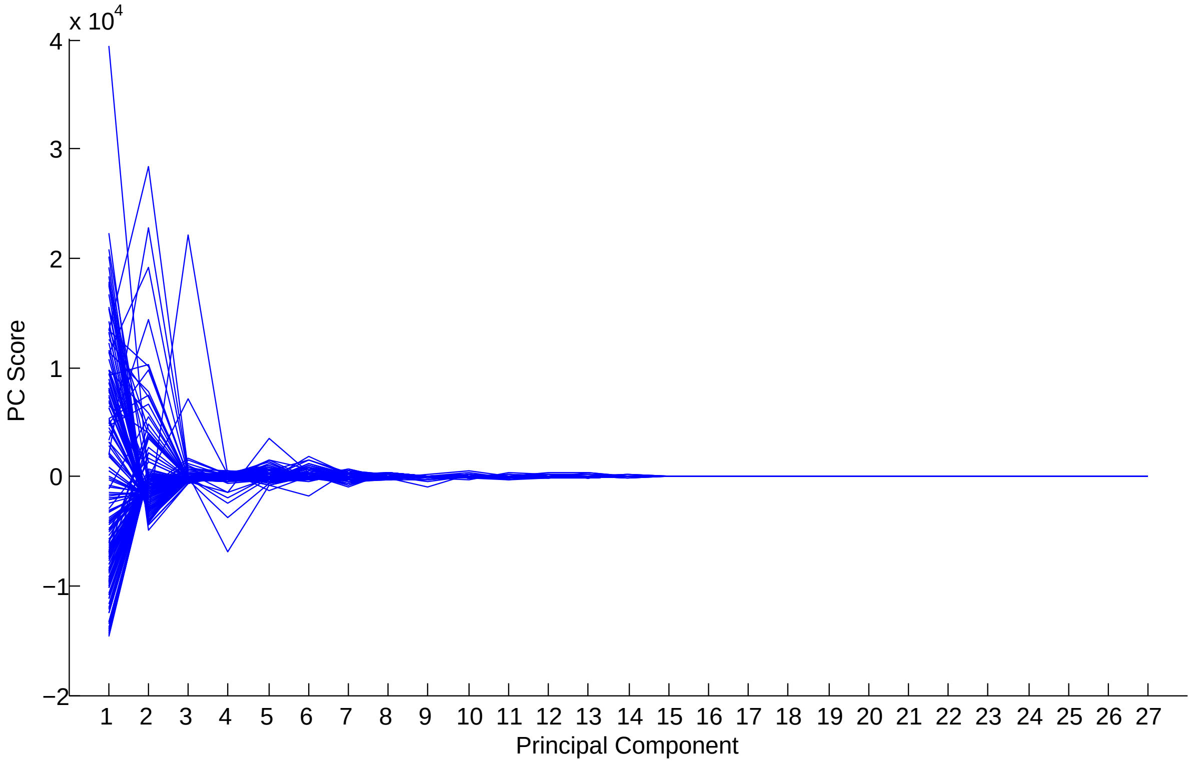


Figure 4.6: Parallel Coordinate Plot of the Principal components of the raw data

It is worth noting of course that as mentioned above, the first few principal components (which explain the vast majority of the total variance in the data) are dominated by the trace elements As, Mo, and Ti, and that what this means is that in this ’rotation of axis’ perspective, that if we consider the loadings shown above, that the first principal component is in a very similar direction to As, and the second is in a very similar direction to Mo, so these axis are actually not rotated very much, and so what we see in Figure 4.6 is in fact very similar to what is shown in Figure 4.1, except with the variables reordered so that As is first, Mo is second, and so on (beyond the first two or three principal components things do start significantly changing, but due to the scale difference, this can not really be visually seen in these plots).

Often, PCA is used as a dimension reduction method, where we get the first k principal component scores (project the data into the subspace spanned by the first k principal component directions), and use this new $k < d$ dimensional dataset for further analysis, and from Figure 4.6 we can see most of the ‘interesting’ or at least visible things occur in the first 7 or so principal components, so we can take a look at these scatterplots (of the principal component scores), in Figure 4.7. However using this (or any other dimension reduction method) should be done with care, as there will always be some loss of information, and just because that information is not a significant component in the variance structure of the data, does not mean it is not important for some other reason (such as, for example, prediction of Au concentration). The exception to this is if some of the later principal component scores have exactly variance zero, i.e. are constant.

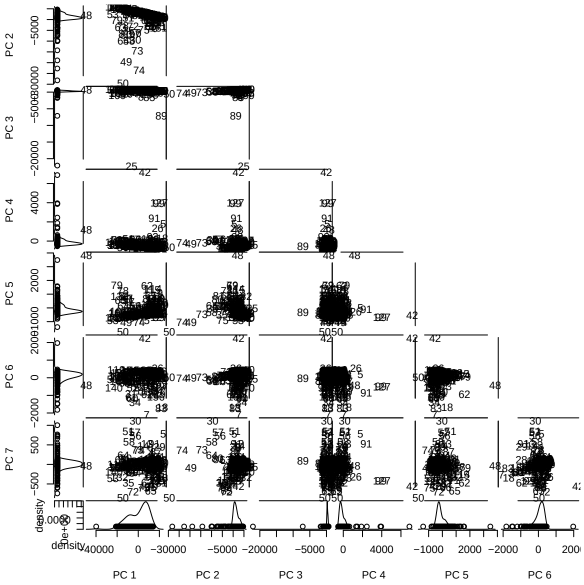


Figure 4.7: Scatterplots of the first 7 principle component scores plotted against each other

From Figure 4.7 we can then see that most of the interesting behavior is really just in the first couple of principal components, so these are shown in Figure 4.8

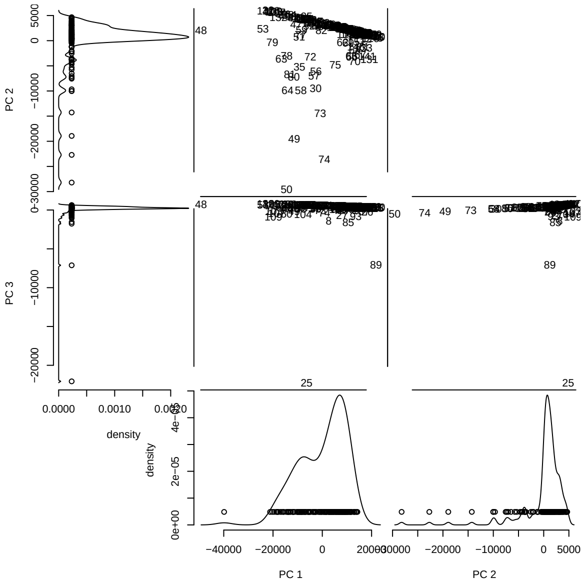


Figure 4.8: Scatterplots of the first 3 principle component scores plotted against each other

Notice how the third principal component has a large contribution from Ti, and looking at Figure 4.8 we can see it is picking up an extreme value in case number 25, which we will look into in more detail later.

4.3.4 PCA on the Standardized Data

Now as could be imagined, variables existing on different scales (or equivalently measured in different units), and thus overwhelming methods such as PCA (and parallel coordinate plots) is not uncommon, and as mentioned above in reference to parallel coordinate plots, often what is done to deal with this and get a look at the data despite it is to standardize. Now with PCA there is an equivalent approach, which can be seen as performing principal component analysis on the scaled (or equivalently standardized) data (i.e. each variable is scaled by its sample standard deviation prior to analysis). This is algebraically equivalent to using the matrix of correlations C instead of the matrix of covariances Σ to obtain the principal components. This is because of how the $(i, j)^{th}$ element of the matrix of correlations, C, is $\frac{cov(\boldsymbol{X}_i,\boldsymbol{X}_j)}{\sqrt{var(\boldsymbol{X}_i)var(\boldsymbol{X}_j)}}$, the Pearson's product moment correlation coefficient as in Equation 3.1 but for the population case. Which we can see from its form, and how the $(i, j)^{th}$ element of Σ is $cov(\boldsymbol{X}_i, \boldsymbol{X}_j)$, means that $C = \Sigma_{diag}^{-\frac{1}{2}}\Sigma\Sigma_{diag}^{-\frac{1}{2}}$ where Σ_{diag} is Σ but with all the off-diagonal terms replaced with zeros. As this is then a diagonal matrix, raising it to a non-natural power is well defined, furthermore it is in fact the diagonal matrix of the values $cov(\boldsymbol{X}_i, \boldsymbol{X}_i) = var(\boldsymbol{X}_i)$ which then makes $\Sigma_{diag}^{-\frac{1}{2}}$ the diagonal matrix of the values $\sqrt{var(\boldsymbol{X}_i)} = s_i$ by the definition of the standard deviation s_i introduced in Chapter 2. Now if we also note how $\Sigma = X_{(raw)}X_{(raw)}^T$ where X is the matrix containing the data (by the definition of covariance) and consider the scaled data, $X_{(scaled)} = \Sigma_{diag}^{-\frac{1}{2}}X_{(raw)}$ which is what I described as scaling above (dividing each trace element by its standard deviation), but written in matrix form. Then if we calculate the covariance matrix for the scaled data, we get that it is equal to $X_{(scaled)}X_{(scaled)}^T = \Sigma_{diag}^{-\frac{1}{2}}X_{(raw)}(\Sigma_{diag}^{-\frac{1}{2}}X_{(raw)})^T = \Sigma_{diag}^{-\frac{1}{2}}X_{(raw)}X_{(raw)}^T\Sigma_{diag}^{\frac{-1}{2}} = \Sigma_{diag}^{-\frac{1}{2}}\Sigma\Sigma_{diag}^{-\frac{1}{2}} = C$. So this shows us that PCA on the scaled data is equivalent to obtaining out principal components from the eigenvectors of the correlations matrix C. This, similarly to ordinary PCA, extends very naturally to the sample case, where we obtain the estimates for the principal components from the eigenvectors of the sample correlations matrix shown in Table 3.1 whose $(i, j)^{th}$ element is then the sample Pearson's correlation as described in Equation 3.1.

I briefly noted above that using the scaled data is equivalent to using the standardized data in the context of PCA, both correspond exactly to the eigenvectors of the correlations matrix C. This is because calculation of the covariance matrix is invariant to location, and centering (subtracting the mean) is the only difference between the scaled and the standardized data (as I have defined them in Table 2.1), so the covariance matrix of the scaled data is exactly the covariance matrix of the standardized data, both of which are simply the correlations matrix C.

Moving along, PCA on the scaled (or standardized) data yields:

```
> pcaCor = princomp(lasData,cor=TRUE)
> summary(pcaCor)
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.3016489 2.1249827 1.69856267 1.38225047 1.31320370
Proportion of Variance 0.1826754 0.1557087 0.09948673 0.06588332 0.05946565
Cumulative Proportion  0.1826754 0.3383841 0.43787082 0.50375415 0.56321980
                          Comp.6    Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.2563606 1.1875512 1.12251663 1.01104351 0.96193657
Proportion of Variance 0.0544290 0.0486303 0.04344978 0.03524859 0.03190765
Cumulative Proportion  0.6176488 0.6662791 0.70972888 0.74497746 0.77688512
                         Comp.11   Comp.12    Comp.13    Comp.14    Comp.15
Standard deviation     0.9242204 0.8579515 0.84405705 0.82232302 0.76320736
Proportion of Variance 0.0294546 0.0253821 0.02456663 0.02331776 0.02008571
Cumulative Proportion  0.8063397 0.8317218 0.85628845 0.87960621 0.89969192

.
.
.

> pcaCor$loadings

Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
DrillNo               -0.130 -0.477         0.299  0.117  0.172
Depth                        -0.558 -0.107  0.161  0.184  0.232
Au      -0.188        -0.231  0.149 -0.331  0.203  0.248
Ag      -0.389  0.140
As      -0.152        -0.253        -0.330  0.196  0.332 -0.133
Sb      -0.305  0.159        -0.107        -0.158         0.122
Ti       0.140  0.406                                     0.195
V        0.140  0.376                                    -0.102
Cr       0.123  0.393                                     0.266
Mn       0.104         0.116 -0.346                      -0.397
Co                     0.444  0.123         0.351  0.166
Ni      -0.164  0.108  0.386  0.128         0.294  0.184
Cu                     0.274        -0.353  0.132 -0.174
Zn                                  -0.424        -0.347  0.160
Ga       0.119  0.200  0.190 -0.198               -0.163 -0.209
Ge              0.146        -0.261               -0.442 -0.273
Se      -0.131        -0.282         0.230  0.451 -0.224
Nb       0.163  0.327                                    -0.215
Mo      -0.355  0.116
Cd                     0.230 -0.252               -0.110  0.393
In                            0.111 -0.493  0.116 -0.271
Sn              0.172               -0.107        -0.166  0.186
Te                    -0.160         0.321  0.504 -0.299 -0.101
W        0.125  0.405 -0.113                              0.234
Re      -0.350
Tl      -0.369  0.136        -0.112        -0.137
Pb      -0.334  0.134  0.215
Bi                     0.365                0.112 -0.101
U               0.158                              0.254 -0.385

.
.
.
```

Similarly we produce the scree plot in Figure 4.9, note the distinctive difference as compared to Figure 4.5. Now we observe a much more reasonable distribution of variability (as the variances are now all equal, in the scaled data, these directions could be more useful, as the direction of maximum variance is no longer affected by which variable has the largest variance, but rather by which variables have the highest mutual correlation).

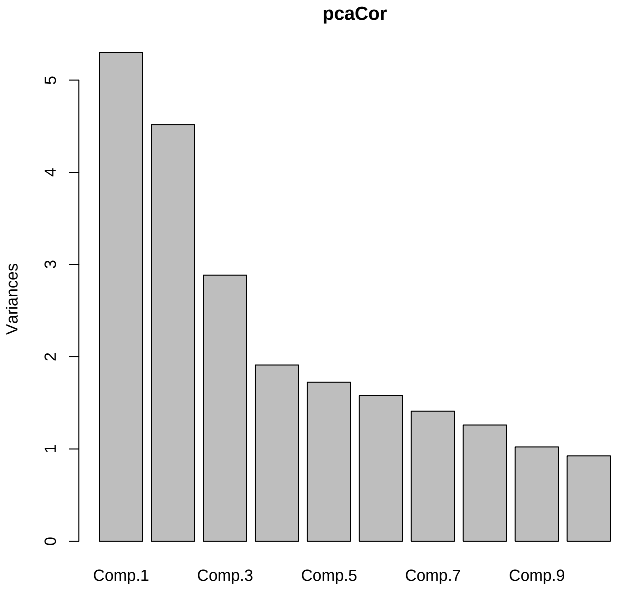


Figure 4.9: Scree plot of the Principal component analysis on the scaled data

So similarly if we project the data into the directions of the principal components obtained from this ‘correlation’ PCA, we can produce the parallel coordinate plot shown in Figure 4.10.

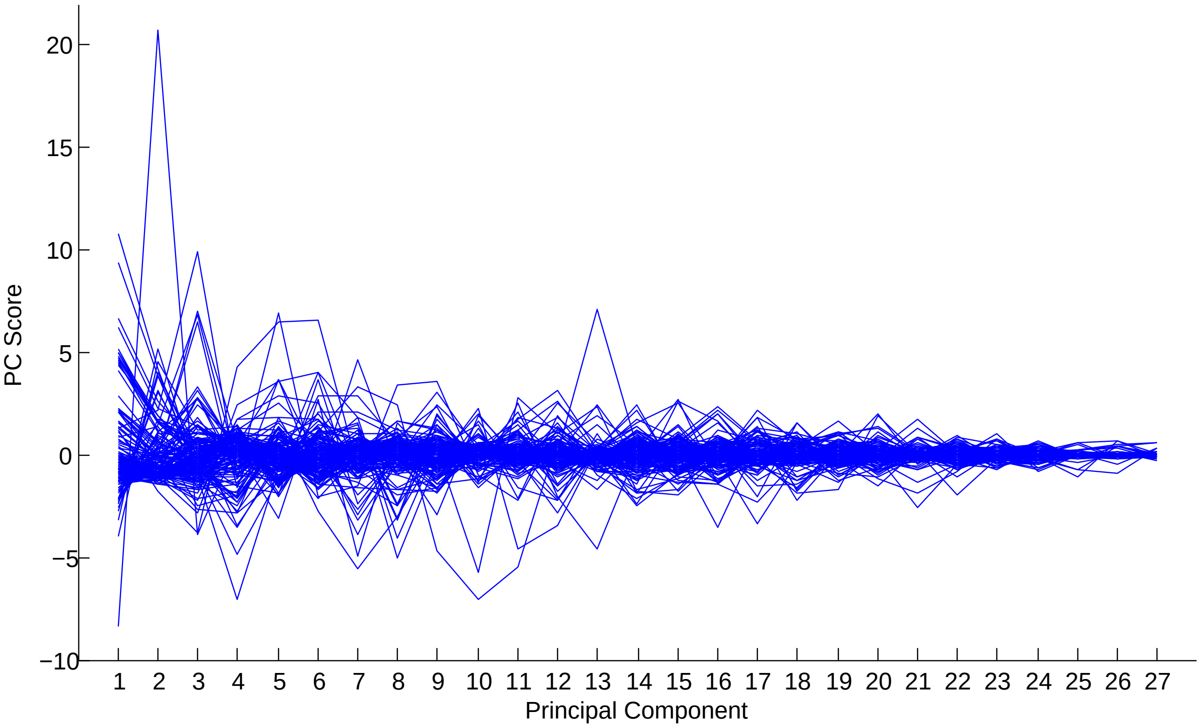


Figure 4.10: Parallel Coordinate Plot of the Principal Components of the standardized data (effectively on the correlation matrix, rather than the covariance matrix)

Similarly to the raw data it can also be interesting to take a look at the first few principal component scores, shown in Figure 4.11

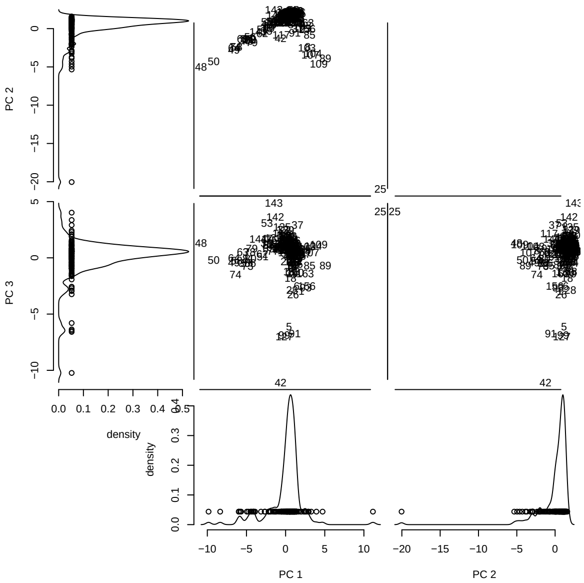


Figure 4.11: Scatterplots of the first 3 principle component scores on the standardized data plotted against each other

Notice how in Figure 4.11 we can see that PC1 and PC2 pick up the extreme value 25, and now the third principle component, PC3, picks up case no 42 as being an extreme value. This is as this analysis is based on the correlations which are standardized by the variances of the elements, and so the first two principle components pick up the direction of the most extreme value, case no 25, rather than picking up the directions of maximum variance (As and Mo) as the PCA on the raw data did.

4.3.5 Log-transformed Data

Similarly, here are parallel coordinate plots of the log-transformed data projected into the directions of the principal components obtained first from the raw, and then the scaled data.

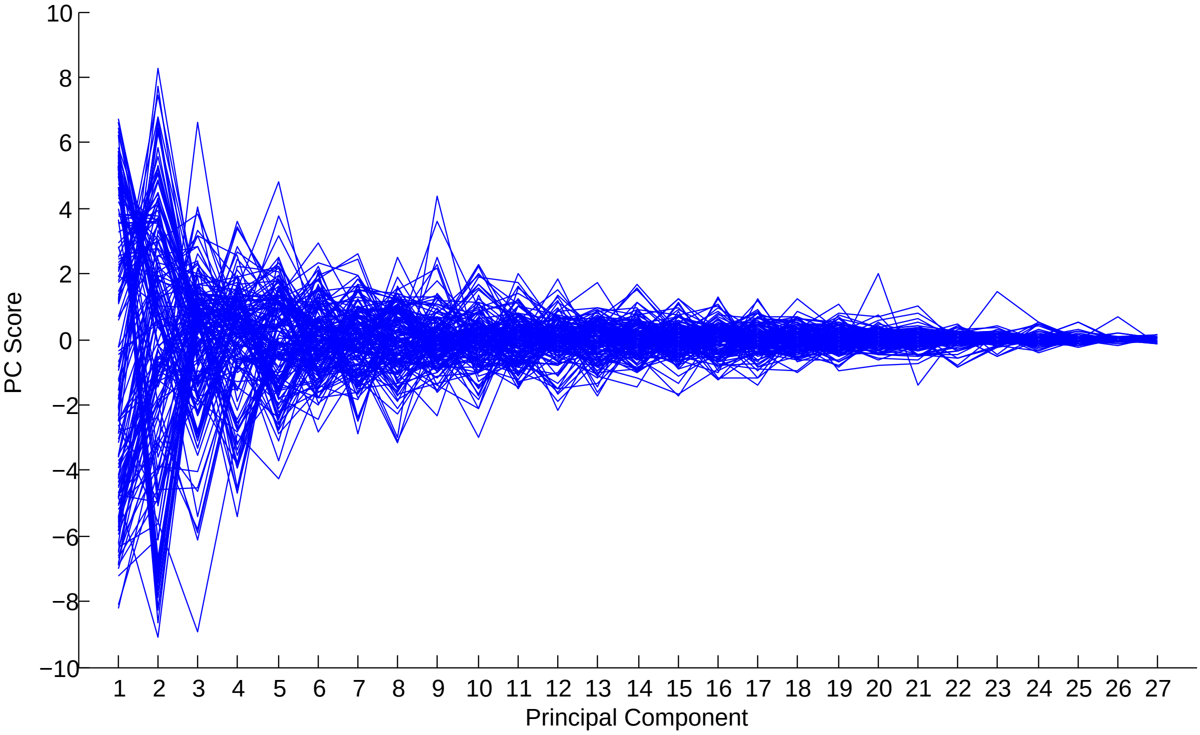


Figure 4.12: Parallel Coordinate Plot of the Principal Components of the log-transformed Data

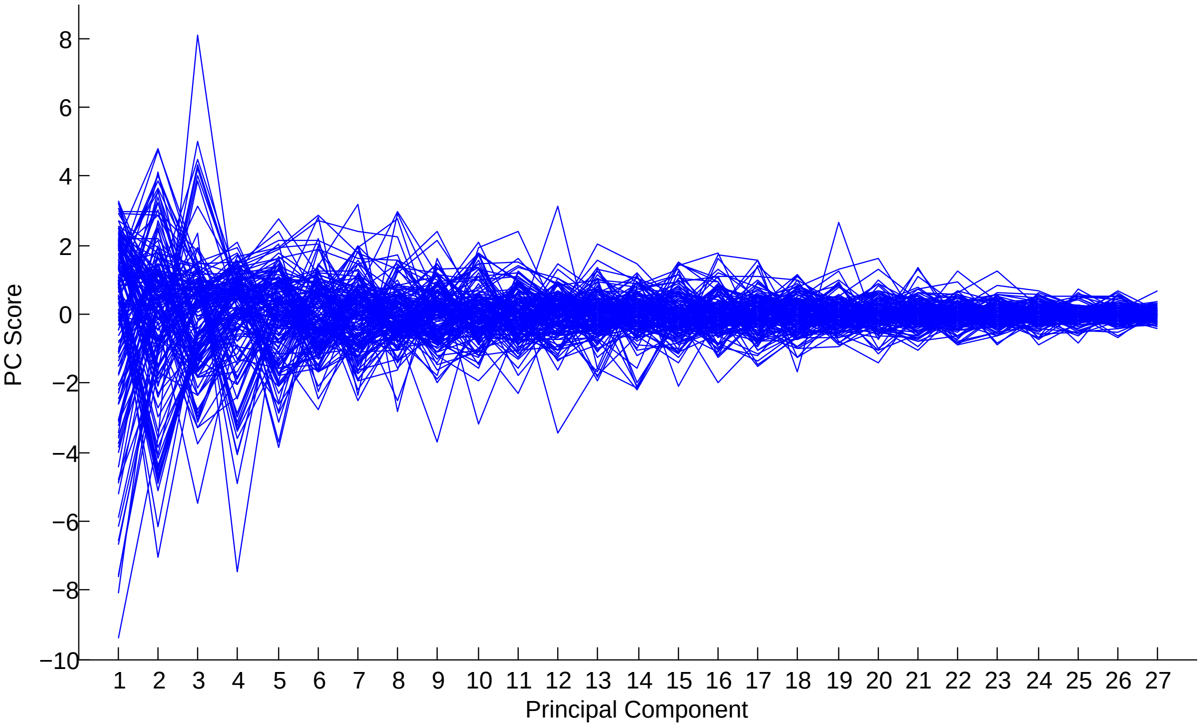


Figure 4.13: Parallel Coordinate Plot of the Principal Components of the standardized log-transformed Data

Notice the lack of a big difference between the two parallel coordinate plots (for PCA on the raw, and scaled log-transformed data) This would be because of how, by the nature of the log-transform, it has brought all the data into the same scale. So this is interesting, as it could be said that PCA on the scaled raw data is similar to PCA on the log-transformed data in this case, as both reduce the effect of certain trace elements having much larger variance than others, but scaling the raw data completely removes this as a factor, while taking the log-transform severely reduces this effect, but does not remove it outright, which could be interesting, as it might be useful to keep the effect in, as it may have some legitimate real-world interpretation.

4.4 Factor Analysis

Factor Analysis is, in its basic idea, somewhat similar to PCA, in how it it interested primarily in the variance structure of the data, and as I will show later, you can actually use your principal components as a non-parametric method for obtaining estimated factor loadings for factor analysis [8, Chap. 7]. The main difference between factor analysis and PCA is that factor analysis has a model (4.5) that asserts there are k ($k < d$) underlying latent factors which can completely (or as completely as possible) 'explain' the variance structure of the data, which changes the interpretation of these factors. What is meant by 'explanation' is put nicely is [3],

> Factor Analysis is concerned with whether the covariances or correlations between a set of observed variables, ... can be 'explained' in terms of a smaller number of unobservable, latent variables $f_1, \ldots, f_k$, ... Explanation in this case means that the correlation between each pair of observed variables results from their mutual association with the latent variables; consequently the *partial* correlation between each pair of observed variables given the values of of $f_1, \ldots, f_k$, should be approximately zero.

I will consider two different approaches to Factor analysis in this section, firstly the MLE approach, which is based on normal-theory assumptions, which are quite inappropriate in this case, but the results that this method yield are interesting, particularly given the lack of normality. The second is, as I mentioned, a non-parametric approach, based on PCA.

4.4.1 Maximum Likelihood Approach

So our model is (in the population case)

$$\boldsymbol{X} = A_{d\times k}\boldsymbol{f} + \boldsymbol{\epsilon}, \text{ where } \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_d \end{pmatrix}, \ \boldsymbol{f} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{pmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_d \end{pmatrix} \tag{4.5}$$

Where A is called the matrix of factor loadings, and contains the linear combinations of the factors that make up each of the observed variables. The different factor analysis methods basically correspond to different methods for obtaining estimates of these factor loadings. As explained in both [8, Chap. 7] and [3], the standard method for estimating these factor loadings is by making the assumptions that the $\boldsymbol{X}_j$ are normally distributed (more specifically that $f_1, f_2, ..., f_k \overset{\text{iid}}{\sim} N(0, 1)$), and maximizing the likelihood function with respect to A and $\Psi = diag(\psi_1, \ldots, \psi_d)$, where Ψ is the covariance matrix of $\boldsymbol{epsilon}$ shown in Equation 4.5 numerically, and so is often referred to as Maximum Likelihood Factor Analysis. It is also worth noting that the formulation of the model in Equation 4.5 assumes that the data (i.e. each $\boldsymbol{X}_j$) is centered, and this is unimportant as the factor analysis is entirely based on the variance structure of the data, not its location, (it is similar in this respect to PCA).

Outline of the Derivation of the Likelihood Function

$$\Sigma = AA^T + \Psi \text{ where } \Psi = diag(\psi_1, \ldots, \psi_d) \text{ and } \psi_i = var(\epsilon_i) \qquad (4.6)$$

First, Equation 4.6 is derived from Equation 4.5 quite simply as

$$\begin{aligned} \text{var}(\boldsymbol{X}) &= \text{var}(A\boldsymbol{f} + \boldsymbol{\epsilon}) && \text{(Equation 4.5)} \\ &= \text{var}(A\boldsymbol{f}) + \text{var}(\boldsymbol{\epsilon}) \\ &= A\,\text{var}(\boldsymbol{f})A^T + \Psi \\ &= AI_{k\times k}A^T + \Psi && (f_j \overset{\text{iid}}{\sim} \text{N}(0,1)) \\ &= AA^T + \Psi \end{aligned}$$

Now consider we have observed values of $\boldsymbol{X}$, particularly $\boldsymbol{x}_{1\bullet}, \boldsymbol{x}_{2\bullet}, ..., \boldsymbol{x}_{n\bullet}$, corresponding to (unknown) $\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_n$. Then log-likelihood function,

$$\begin{aligned} l(A,\Psi) &= ln\left(\prod_{i=1}^{n} p(\boldsymbol{x}_{i\bullet}, \boldsymbol{f}_i|A,\Psi)\right) && (p \text{ denotes the density function}) \\ &= \sum_{i=1}^{n} ln(p(\boldsymbol{x}_{i\bullet}, \boldsymbol{f}_i|A,\Psi)) \\ &= \sum_{i=1}^{n} ln(p(\boldsymbol{x}_{i\bullet}|\boldsymbol{f}_i, A,\Psi)p(\boldsymbol{f}_i|A,\Psi)) \\ &= \sum_{i=1}^{n} ln(p(\boldsymbol{x}_{i\bullet}|\boldsymbol{f}_i, A,\Psi)) + \sum_{i=1}^{n} ln(p(\boldsymbol{f}_i|A,\Psi)) \\ &= \sum_{i=1}^{n} ln(p(\boldsymbol{x}_{i\bullet}|\boldsymbol{f}_i, A,\Psi)) + \sum_{i=1}^{n} log(p(\boldsymbol{f}_i)) && (\text{distribution of } \boldsymbol{f} \text{ independent of } A \text{ and } \Psi) \quad (4.7) \end{aligned}$$

Now as the second term in Equation 4.7 is independent of A and Ψ, and so can be disregarded when finding maximum likelihood estimates for A and Ψ (which is what we are doing in factor analysis). And as $\boldsymbol{X}$ is a linear combination of normally distributed variables, it is normally distributed, with

$$\begin{aligned} \text{E}[\boldsymbol{X}|\boldsymbol{f}] &= \text{E}[A\boldsymbol{f} + \boldsymbol{\epsilon}|\boldsymbol{f}] \\ &= \text{E}[A\boldsymbol{f}|\boldsymbol{f}] + \text{E}[\boldsymbol{\epsilon}|\boldsymbol{f}] \\ &= A\boldsymbol{f} + \boldsymbol{0} \\ &= A\boldsymbol{f} \end{aligned} \qquad \begin{aligned} \text{var}(\boldsymbol{X}|\boldsymbol{f}) &= \text{E}[(\boldsymbol{X} - \text{E}[\boldsymbol{X}])(\boldsymbol{X} - \text{E}[\boldsymbol{X}])^T|\boldsymbol{f}] \\ &= \text{E}[(\boldsymbol{X} - A\boldsymbol{f})(\boldsymbol{X} - A\boldsymbol{f})^T|\boldsymbol{f}] \\ &= \text{E}[((A\boldsymbol{f} + \boldsymbol{\epsilon}) - A\boldsymbol{f})((A\boldsymbol{f} + \boldsymbol{\epsilon}) - A\boldsymbol{f})^T|\boldsymbol{f}] \\ &= \text{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{f}] \\ &= \Psi \end{aligned}$$

So

$$\boldsymbol{X}|\boldsymbol{f} \sim \text{N}_d(A\boldsymbol{f}, \Psi) \text{ I.e. } p(\boldsymbol{X}|\boldsymbol{f}, A, \Psi) = 2\pi^{-\frac{d}{2}}|\Psi|^{-\frac{1}{2}}exp\left(-\frac{1}{2}(\boldsymbol{X} - A\boldsymbol{f})^T\Psi^{-1}(\boldsymbol{X} - A\boldsymbol{f})\right) \qquad (4.8)$$

Using Equation 4.8, the first term in Equation 4.7 becomes

$$
\begin{aligned}
l &\propto \sum_{i=1}^{n} ln(p(\boldsymbol{x}_{i\bullet}|\boldsymbol{f}_i, A, \Psi)) \\
&= \sum_{i=1}^{n} ln\left(2\pi^{-\frac{d}{2}}|\Psi|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{x}_{i\bullet} - A\boldsymbol{f}_i)^T \Psi^{-1}(\boldsymbol{x}_{i\bullet} - A\boldsymbol{f}_i)\right)\right) \\
&\propto -\frac{n}{2} log(|\Psi|) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_{i\bullet} - A\boldsymbol{f}_i)^T \Psi^{-1}(\boldsymbol{x}_{i\bullet} - A\boldsymbol{f}_i) \\
&= -\frac{n}{2} ln(|\Psi|) - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_{i\bullet}^T \Psi^{-1}\boldsymbol{x}_{i\bullet} - 2(A\boldsymbol{f}_i)^T \Psi^{-1}\boldsymbol{x}_{i\bullet} + (A\boldsymbol{f}_i)^T \Psi^{-1} A\boldsymbol{f}_i\right) \\
&= -\frac{n}{2} ln(|\Psi|) - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_{i\bullet}^T \Psi^{-1}\boldsymbol{x}_{i\bullet} - 2(A\boldsymbol{f}_i)^T \Psi^{-1}\boldsymbol{x}_{i\bullet} + tr\left(\boldsymbol{f}_i^T A^T \Psi^{-1} A\boldsymbol{f}_i\right)\right) \\
&= -\frac{n}{2} ln(|\Psi|) - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_{i\bullet}^T \Psi^{-1}\boldsymbol{x}_{i\bullet} - 2(A\boldsymbol{f}_i)^T \Psi^{-1}\boldsymbol{x}_{i\bullet} + tr\left(A^T \Psi^{-1} A\boldsymbol{f}_i\boldsymbol{f}_i^T\right)\right)
\end{aligned}
$$

This can then be evaluated and maximized using the EM-algorithm, details are covered in [11] but I will leave it at that for these purposes.

Results: Biplots

Now a nice way to visualize the results of a factor analysis is a biplot, as shown below in Figure 4.14. A biplot is where for k = 2, or 3-factor analysis, we can plot the factor loadings of each of the observed variables as vectors in the factor-space.

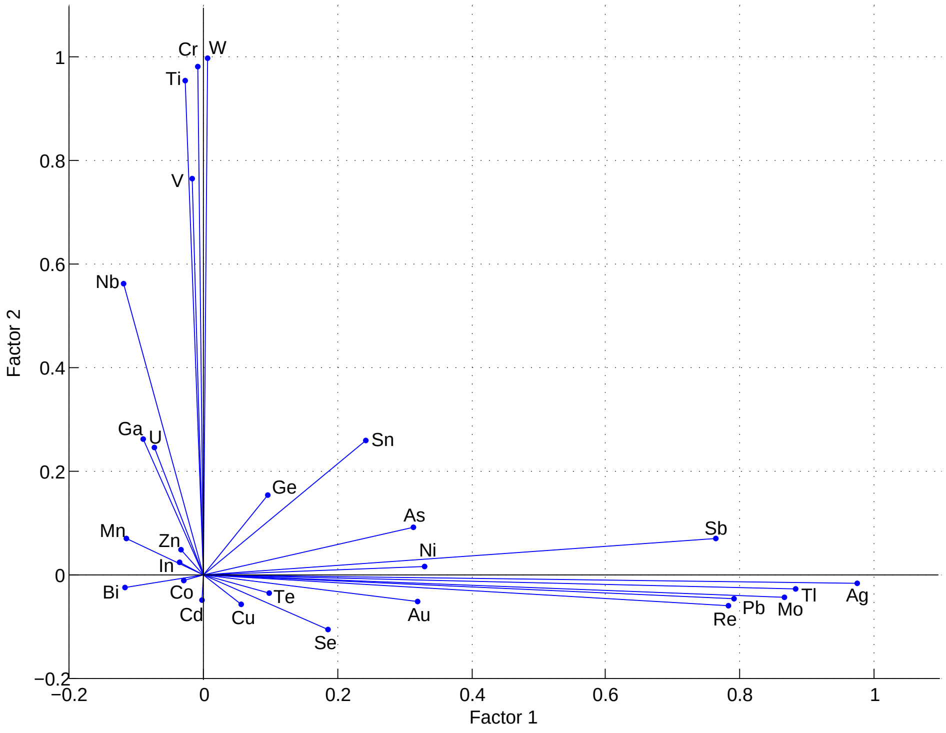


Figure 4.14: Biplot of the 2-factor Analysis on the raw data

Here we notice the same two major clusters noted in Figure 3.4, i.e. The first cluster: Ag, Mo, Tl, Pb, Sb and Re is all oriented in the direction of Factor 1, while the second cluster: Cr, W, Ti, V and Nb, are all oriented in the direction of Factor 2, while all the other elements have relatively small magnitudes. Interestingly, when this analysis is expanded to a 3-factor model, the biplot in the Factor 1 by Factor 2 plane stays very similar, and the only major change is the the pair Co and Ni is raised in the direction of Factor 3, the remaining elements still having quite small magnitudes.

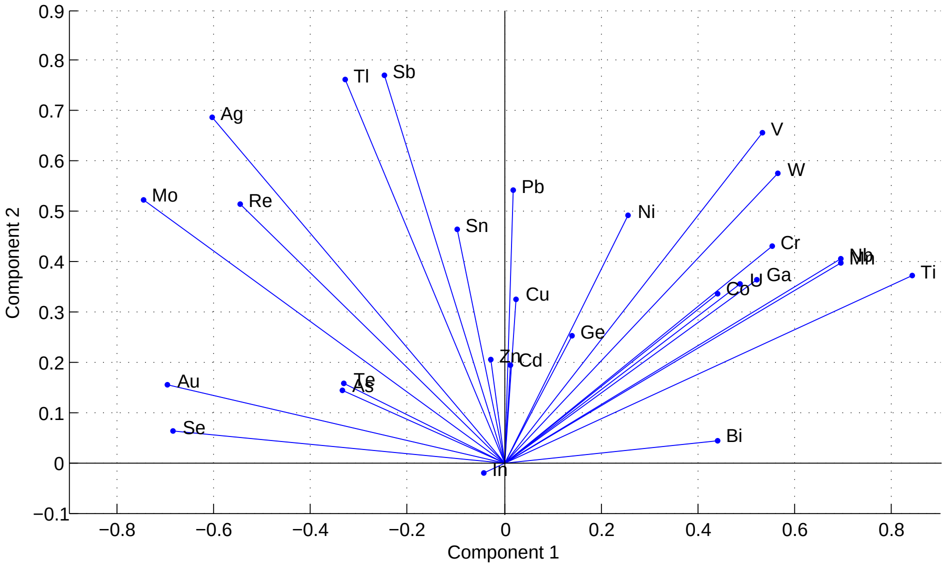


Figure 4.15: Biplot of the 2-factor Analysis on the log data

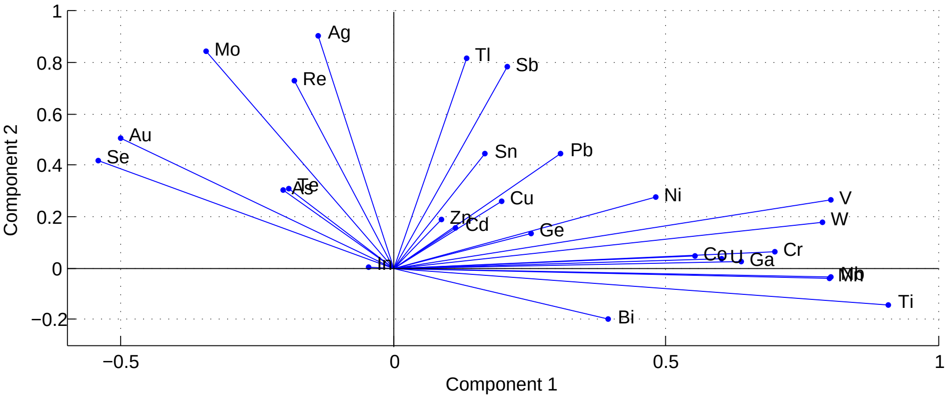


Figure 4.16: Biplot of the 2-factor Analysis on the log data, with Varimax Rotation

Note the lack of interesting trends in the 2-FA of the log-transformed data.

4.4.2 Principal Factor Analysis

If the normality assumptions for the maximum likelihood approach of Maximum Likelihood Factor Analysis do not seem reasonable, but it is still desirable to investigate the possibility of underlying latent variables, a nonparametric approach to finding our factor loadings becomes desirable. One way of doing this is using principal component analysis, i.e. using the eigenvectors of the covariance matrix Σ or the matrix of correlations (as explained above) for the factor loadings.

Principal component based factor analysis on the raw data is somewhat uninformative, see the biplot in Figure 4.17, as the principal component analysis is overwhelmed by the differences in the variances, as we have already discussed. We see this by how the only two elements with magnitude large enough to be visible in Figure 4.17 are As and Mo. On this biplot I also plotted the data projected into the 2-factor plane, which corresponds exactly to the plane spanned by the first two principle components on the raw data, as in Figure 4.8. The strong linear trend apparent in some of the data in Figure 4.17 below is (in vague terms) the line defined by the concentration of Mo being zero, but this is discussed in more detail in Chapter 6 where I make extensive use of this plane. Technically the restraint $x_{iMo} = 0$ defines a hyperplane, but as this plane is a close approximation to a proper rotation of the As, Mo plane, it can (roughly) be considered a line (the intersection of the hyperplane defined by $x_{iMo} = 0$ and the Mo As plane) in this context.

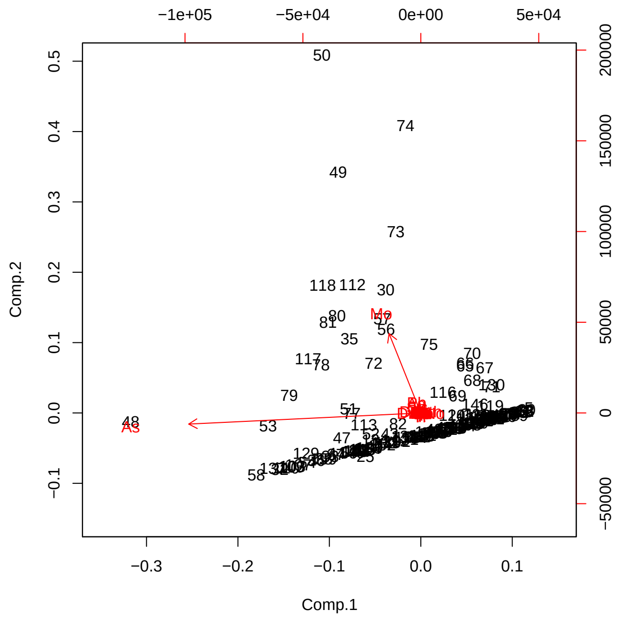


Figure 4.17: Biplot for non-parametric factor analysis based on principal component analysis of the raw data (covariance matrix)

Also, this method provides interesting results when applied to the scaled data (i.e. taking the factor loadings/ directions as the eigenvectors of the correlations matrix), the results of which shown below in Figure 4.18.

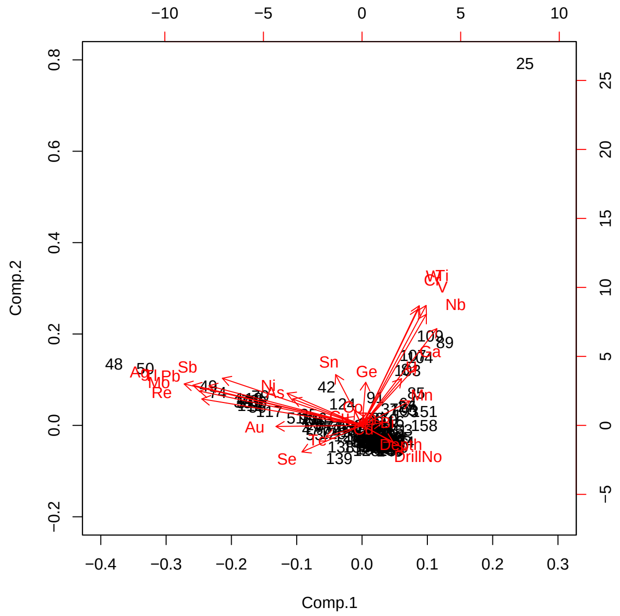


Figure 4.18: Biplot for non-parametric factor analysis based on principal component analysis of the scaled data (matrix of correlations)

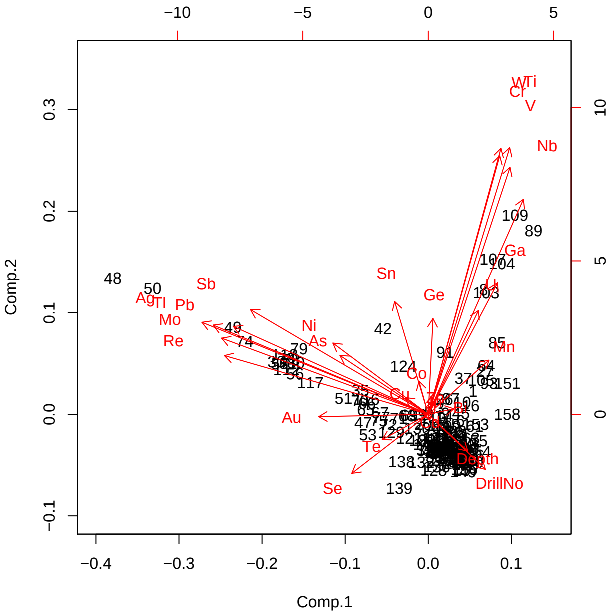


Figure 4.19: Zoom in of Figure 4.18

In fact we see that the biplot in Figure 4.18 shows a separation of the elements nearly identical to that shown in Figure 4.14, which is interesting considering how the normality assumption for the Maximum Likelihood Factor Analysis seems inappropriate (the raw data appears not to fit a normal distribution), and yet it yields nearly identical results to the nonparametric method shown here, so it would seem that the Maximum Likelihood approach is quite robust to deviations from the normality assumption, in this specific example, at least. Although it is worth noting that testing normality in a dataset with such a relatively small $n : d$ ratio is difficult. What is actually happening here is that both these methods are based on the correlations (or scaled covariances) of the data, and this is why they are yielding such similar results, so in fact, these plots are actually displaying information very similar to that displayed in Chapter 3, particularly in Figure 3.4.

Also, note how case no. 25 is extremely far from all the other observations in Figure 4.18, which is encouraging as we identified this point as being highly influential in Chapter 3, and I will investigate the use of correlations (related directly to PCA on the standardized data) in identifying significant points much more thoroughly in Chapter 5.

Chapter 5

Bootstrap based Influence Diagnostics

We noticed in Chapter 3 the influence of large valued points on the Pearson's correlations, as it is a statistic sensitive to the effect of large values. I introduced two methods for more robust estimation of the correlations: using the log-transformed data (which reduces the effect of any large influential points); and Spearman's Rho, a rank-based measure of association (which reduces the effect of any single large values severely). We are however still very interested in the correlations between the valued measures on the trace elements (not just their rank-based association). So in this chapter I will propose a method, primarily based on Efron's Bootstrap [2], that will take advantage of the sensitivity of Pearson's correlation coefficient to identify these influential points.

5.1 The Nonparametric Bootstrap

Suppose we have a data matrix $\mathbb{X}$ (in our case a 164×27 matrix), consisting of a sample of $n = 164$ observations $\{\boldsymbol{x}_{1\bullet}, ..., \boldsymbol{x}_{n\bullet}\}$ and we calculate some statistic from it, $\theta(\mathbb{X}) = \theta(\boldsymbol{x}_{1\bullet}, ..., \boldsymbol{x}_{n\bullet}) = \hat{\theta}$ (we are particularly interested in $\hat{\theta} = r_{jk}$) as an estimate for the population parameter $\theta(\boldsymbol{X}) = \theta$. I use the loose notation $\theta(\mathbb{X}) = \theta(\boldsymbol{x}_{1\bullet}, ..., \boldsymbol{x}_{n\bullet})$ as I consider the data matrix $\mathbb{X}$ to, in this context, be equivalent to the collection of observations $\{\boldsymbol{x}_{1\bullet}, ..., \boldsymbol{x}_{n\bullet}\}$, this is useful later when we construct bootstrap samples. If we want to make some inference on this statistic, such as construct confidence intervals, but we do not have any distributional knowledge, we can use the empirical cumulative distribution function and bootstrap distribution.

We do this by taking B bootstrap samples. Where each bootstrap sample is of size n and is taken with replacement, with equal probability, from our sample. That is, if our sample is (as in Table 2.1) $\{\boldsymbol{x}_{1\bullet}, ..., \boldsymbol{x}_{n\bullet}\}$, then the m^{th} observation in the collection of n that is the b^{th} bootstrap sample is $\boldsymbol{x}^*_{b,m} = \boldsymbol{x}_{i\bullet}$, for $m \in \mathbb{Z} \cap [1, n], \quad b \in \mathbb{Z} \cap [1, B]$ such that independently for each value of b and m, $i = \lfloor n.u + 1 \rfloor$ where u follows the standard uniform distribution, i.e. $u \sim U(0, 1)$. That is to say i follows a discrete uniform distribution on $[1, n]$.

Then for each bootstrap sample, we calculate the statistic of interest, $\theta\left(\boldsymbol{x}^*_{b,1}, ..., \boldsymbol{x}^*_{b,n}\right) = \hat{\theta}^*_b$. However for n even moderately large (even in the case of our relatively small sample size of $n = 164$), it is computationally impossible to enumerate all possible bootstrap samples, as the number of possible different bootstrap samples can be shown to be $\binom{2n-1}{n} = \binom{327}{164} \approx 1.2 \times 10^{97}$. So in this case I use $B = 10^5$, which actually turns out to be unnecessarily large, but with modern computing technology is still very fast, and we do not lose anything from taking B to be larger than necessary (and we gain nice, smooth, histograms). I do briefly consider reducing B and applying these techniques in other applications, in the 'Future Work' section at the end of this chapter, but I wont cover that here.

Then the fundamental core of bootstrapping is that the bootstrap distribution (that the $\hat{\theta}^*_b$'s are drawn from) is to the sample estimate $\hat{\theta}$ as the sampling distribution (from which $\hat{\theta}$ is drawn from), is to the population parameter θ. Then using this relationship we can construct confidence intervals and such for the population parameter θ, based on our sample estimate $\hat{\theta}$. In this case however, we are not so much interested in constructing confidence intervals, as we are not overly concerned with the numerical values of the correlation coefficients, but rather their nature, and in this chapter I will introduce a method for using the bootstrap distribution as an exploratory method for identifying influential points (that influence the correlations between the trace elements).

5.2 An illustrative example of Bootstrapping

Consider a simple example, with $n = 10$, $d = 1$ and $\{\boldsymbol{x}_{1\bullet}, \boldsymbol{x}_{2\bullet}, ..., \boldsymbol{x}_{10\bullet}\} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 100\}$. That is,

$$\boldsymbol{x}_{i\bullet} = \begin{cases} i & \text{for } i \leq 9 \\ i^2 = 100 & \text{for } i = 10 \end{cases}$$

Suppose the statistic we are interesting in estimating, θ above, is the mean, i.e.

$$\theta\left(\boldsymbol{x}_{1\bullet}, \boldsymbol{x}_{2\bullet}, ..., \boldsymbol{x}_{10\bullet}\right) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i\bullet} = \hat{\theta} = 14.5$$

Now consider, that the number of possible distinct bootstrap samples are $\binom{2n-1}{n} = 1910 = \frac{19!}{10!9!} = 92378$ and this is already immense although n is quite small. Now suppose we take $B = 1000$ bootstrap samples. To illustrate what these are, consider some typical examples of bootstrap samples, ordered by their $\hat{\theta}^*_b$ values:

b	$\boldsymbol{x}^*_{b,1}$	$\boldsymbol{x}^*_{b,2}$	$\boldsymbol{x}^*_{b,3}$	$\boldsymbol{x}^*_{b,4}$	$\boldsymbol{x}^*_{b,5}$	$\boldsymbol{x}^*_{b,6}$	$\boldsymbol{x}^*_{b,7}$	$\boldsymbol{x}^*_{b,8}$	$\boldsymbol{x}^*_{b,9}$	$\boldsymbol{x}^*_{b,10}$	$\hat{\theta}^*_b$
245	7	7	1	1	3	1	1	7	1	1	3
151	5	5	5	1	3	5	9	3	2	2	4
73	5	5	2	3	8	3	6	7	1	4	4.4
891	2	9	100	1	5	3	9	7	1	5	14.2
939	6	7	4	5	5	9	1	3	2	100	14.2
980	9	2	8	100	100	8	2	9	6	8	25.2

Table 5.1: Some examples of bootstrap samples from the example

If we then plot a histogram of the bootstrap sample means, this produces Figure 5.1 below.

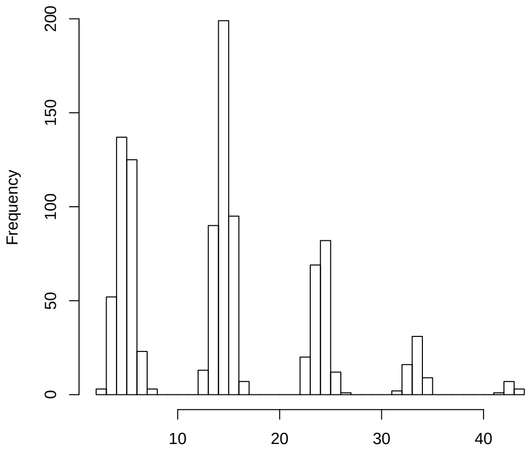


Figure 5.1: Example: Histogram of Bootstrap sample means $\left(\hat{\theta}_b^*\right)$

In Figure 5.1 we see a very clearly multimodal distribution, and this is caused by the fact that any particular bootstrap sample will contain exactly $0, 1, 2, ..., 10$ occurrences of $\boldsymbol{x}_{10\bullet} = 100$. It can be observed that if the b^{th} bootstrap sample contains exactly k occurrences of $\boldsymbol{x}_{10\bullet} = 100$, then $10k < \hat{\theta}_b^* < 10k + 9$ (for example see Table 5.1. These intervals are non-overlapping, and thus result in the bimodality observed in Figure 5.1.

5.3 Bootstrap Distributions of the Correlation Coefficients

So now consider that $\theta(\mathbb{X})$ need not be just a single scalar statistic, and can be several statistics (or a vector of statistics). In particular, here we consider it as the $\binom{27}{2} = 351$ correlation coefficients between the trace elements. The method described above was then carried out, with $B = 10^5$ (I coded the whole thing from scratch in R, because many of the things I want to do are non-standard, code is available at request). Histograms of the bootstrapped values $\hat{\theta}_b^*$ (for $b \in \mathbb{Z} \cap [1, B]$) for each of the 351 correlation coefficients on the raw data, the log-transformed data, and Spearman's correlation coefficients on the raw data are all available at request, but these are not included as they do not add much to my point (other than 60 pages of histograms). The main difference between these is illustrated below in Figure 5.2. So in this case each $\hat{\theta}_b^*$ contains 351 statistics stored conveniently in a 27×27 matrix of correlations, C as in Table 2.1. The large influential point noted in Chapter 3 was case no. 25, which influenced the Cr, W correlation, which in Figure 3.4 was also closely correlated to Ti. So in Figure 5.2 we take a look at the estimated bootstrapped distributions for the correlations between Ti, Cr and W. Notice the very strong bimodality (caused by case no. 25 noted above). This is due to the sensitivity of Pearson's correlation coefficient to large values I mentioned above, that is the bootstrapped correlation coefficient between for example Cr and W is going to be significantly larger if case no. 25 is included in the bootstrap sample as opposed to if it is not, and this is what causes the bimodality in the bootstrap distribution (i.e. the bootstrap samples not including case no. 25 produce the bootstrapped correlations mostly in the left mode in the histogram, and the bootstrap samples including case no. 25 produce the correlation coefficients corresponding to the right mode, but this is further investigated below). Also interesting to note is how in the log-transformed data this bimodality is gone, presumably because the effect of the outlying point has been reduced by the log-transform to the point where it no longer has a strong enough influence to induce such bimodality. Furthermore, notice how similarly the Spearman's rank based correlation coefficient is not bimodal either, but is centered at a slightly lower value, suggesting that although the log-transformed data is less affected by case no. 25, it still has some effect on the location of the bootstrap distribution, while this effect is even further reduced using Spearman's correlation coefficient. This trend of differences between the raw correlation to the log-transformed correlation to the Spearman's correlation is common to all the correlations with bimodality, this being the reason I did not include the histograms for the log-transformed correlations and the Spearman's correlations.

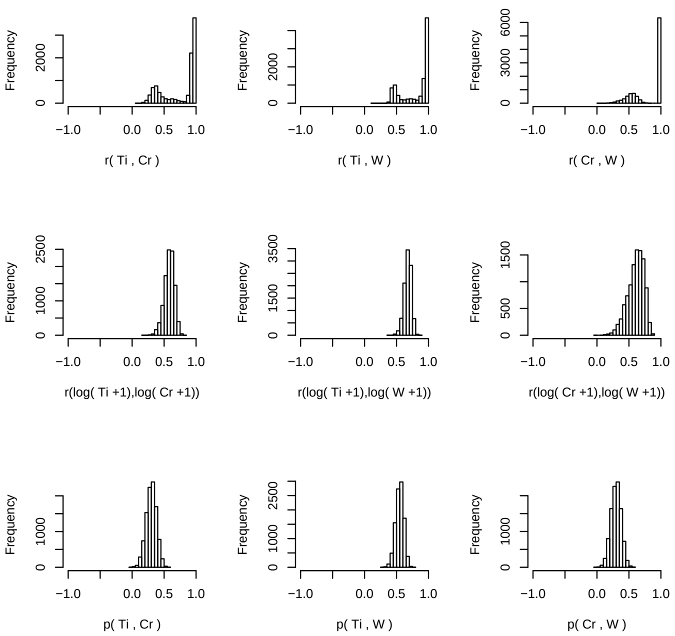


Figure 5.2: Histogram of the estimated bootstrap distribution of the Pearson's Correlation Coefficient on the raw concentrations, on the log concentrations, and the Spearman's Rank Correlation Coefficient on Ti, Cr, and W

5.4 Distributional Results

We have above that *independently* for each $b \in \mathbb{Z} \cap [1, B]$ and $m \in \mathbb{Z} \cap [1, n]$ i is distributed according to a discrete uniform, that is

$$P(i = 1) = P(i = 2) = ... = P(i = n) = \frac{1}{n}$$

So we can consider the number of occurrences of a particular observation i, in a particular bootstrap sample b, as a binomial random variable.

$$X_{b,i} \overset{\text{iid}}{\sim} Bin\left(n, \frac{1}{n}\right) \quad \text{So} \quad P(X_{b,i} = 0) = \binom{n}{0}\left(\frac{1}{n}\right)^0 \left(1 - \frac{1}{n}\right)^n = \left(\frac{n-1}{n}\right)^n$$

For convenience, let us denote

$$q = \left(\frac{n-1}{n}\right)^n \qquad \left(\text{for } n = 164, \quad q = \left(\frac{163}{164}\right)^{164} \approx 0.367\right) \tag{5.1}$$

If we then restrict our attention to the possibilities $X_{b,i} = 0$ and $X_{b,i} > 0$ (in our context, contains the i^{th} observation or not), then we can reduce $X_{b,i}$ to a Bernoulli random variable with probability q, where success is the b^{th} bootstrap sample *not* containing the i^{th} observation. We can then consider the B bootstrap samples as a sample from a binomial distribution

$$X_i \overset{\text{iid}}{\sim} Bin(B, q)$$

The quantity I suggest to consider is the proportion of absence for each observation, in a given subset of the bootstrap samples. That is, the proportion of bootstrap samples that *do not* contain a particular observation i in some subset $\mathcal{B} \subset [1, B] \cap \mathbb{Z}$;

Now varying $|\mathcal{B}|$ can produce interesting results, for example choosing $|\mathcal{B}| = \lfloor qB \rfloor$ or $|\mathcal{B}| = \lfloor (1-q)B \rfloor$ can be interesting when trying to detect particular types of influential points, but will sometimes be less effective in detecting other types of points, so for now, I will use

$$|\mathcal{B}| = \frac{B}{2}$$

Now if $\mathcal{B}$ were chosen randomly, as $\mathcal{B}$ is taken without replacement, the number of bootstrap samples in $\mathcal{B}$ not containing the i^{th} observation is a hypergeometric random variable

$$Y_i \sim Hypergeometric(X_i, B - X_i, |\mathcal{B}|) \text{ i.e. } P(Y_i = k) = \frac{\binom{X_i}{k}\binom{B-X_i}{|\mathcal{B}|-k}}{\binom{B}{|\mathcal{B}|}}$$

5.5 An illustrative example of Influence Diagnostics

What I suggest is to choose $\mathcal{B}$ in a nonrandom manner (related to θ) in order to test if there is a relation between the occurrence of a observation i and θ.

To explain how this works, specifically when θ is a correlation coefficient, consider another illustrative example. This one in $\mathbb{R}^2$, i.e. $d = 2$, and $n = 20$, shown in

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
x	1	2	3	4	5	7	8	9	10	12	13	14	15	17	18	19	20	0	5	20
y	2	1	2	3	3	10	9	10	9	12	12	15	15	18	21	19	19	15	20	0

Consider the scatterplot of the data shown in Figure 5.3. It shows a clear strong linear relationship for observations $1 - 17$ with observations $18 - 20$ clearly not a part of that trend, and as such affecting the least squares line shown in red, while the least squares line in blue (calculated on observations $1 - 17$ only) captures the linear trend mentioned.

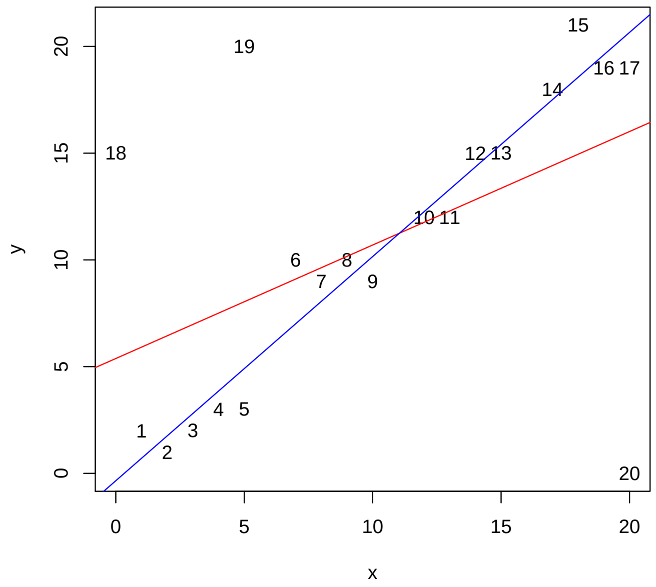


Figure 5.3: Scatterplot of y vs. x, with the linear regression least squares line shown in red, and the least squares line with points 18, 19 and 20 omitted in blue

Similarly to above, we take $B = 1000$ bootstrap samples, and produce the histogram shown in Figure 5.4 of the $\hat{\theta}^*_b$ values. What we notice is that the evident multi-modal behavior apparent in Figure 5.1 is not clearly visible here, although there does seem to be a small peak at 1 separate from the larger, left skewed mode.

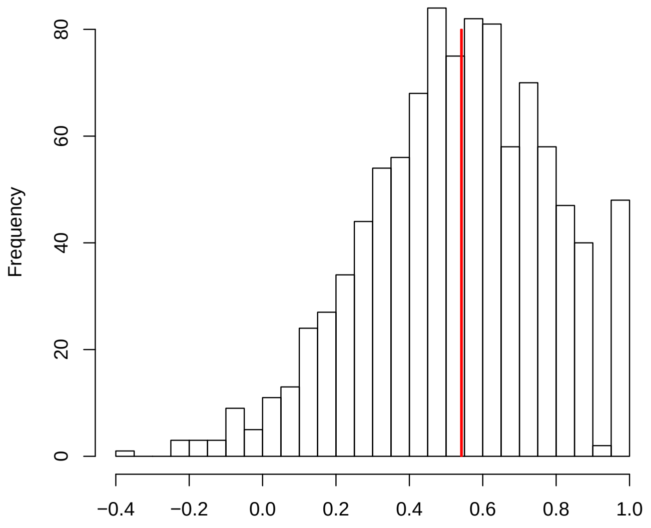


Figure 5.4: Histogram of bootstrapped correlation coefficient values, similar to Figure 5.1 but for the correlation coefficient in this case, with the median shown in red

We choose $|\mathcal{B}| = \frac{B}{2}$, specifically, we take $\mathcal{B}$ to be the set of bootstrap samples that produce values of $\hat{\theta}^*_b$ either above, or below the median value shown in Figure 5.4. Figure 5.5 shows the proportion of absence of each observation given either of these two choices (above or below the median) and the absolute value of their difference.

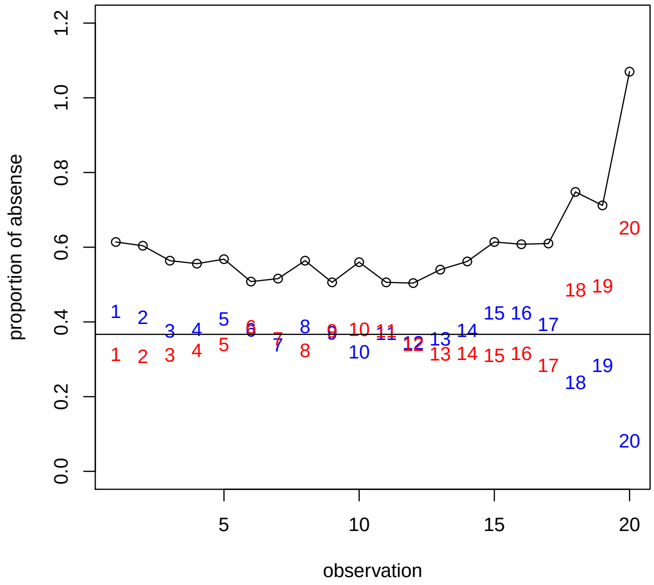


Figure 5.5: For each observation here we have three things plotted, the proportion of absence where $\mathcal{B}$ is the set of bootstrap samples which produce correlation coefficients above the median value shown in Figure 5.4, the proportion of absence where $\mathcal{B}$ is the bootstrap samples producing correlation coefficients below the median, and the absolute value of the difference of the two (○) (shifted up on the scale by 0.5 so it can be seen better)

We immediately see that this picks up points 18 – 20 as being significant, but maybe not much better than conventional methods such as Cook's distance shown in Figure 5.6.

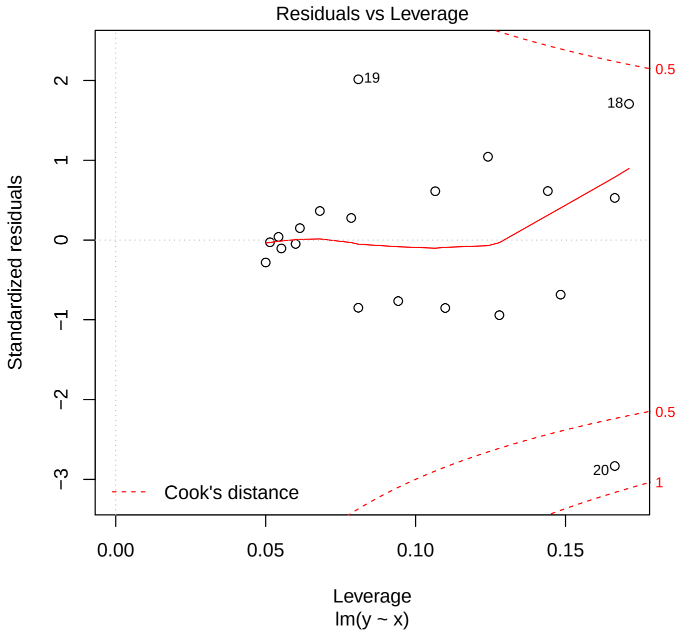


Figure 5.6: Here we have the diagnostic Residual vs. Leverage plot for the regression on y for x, with contour lines of Cook's distance marked in dashed red lines ordinarily a Cook's distance of over 0.5 would indicate a significantly influential point.

What we see here is that my bootstrap method for influence detection performs about as well as Cook's distance in detecting influential points in this context (although cooks distance only works in this context, while the bootstrap method can be applied to any statistic as an influence diagnostic). The main advantage of this method over Cook's distance is its ability to test more general problems. The reason both these methods are having trouble detecting points 18 and 19 is because although (looking at Figure 5.3) both are highly outlying from the obvious linear trend, they are outlying in a similar direction to each other, they are relatively near to each other, relative to the line that is. So when one is removed, the effect of the other is still present. To demonstrate the flexibility of this method, consider that now rather than the proportion of absence, we can consider the proportion of 'joint' absence. That is to say, the proportion of bootstrap that do not contain both observation i and observation j. There are $\binom{20}{2} = 190$ such pairs, not including the $i = j$ case (as this simply reduces to the regular proportion of absence already discussed). So we can produce a similar plot to Figure 5.5 for these joint proportions of absence, in Figure 5.7 below.

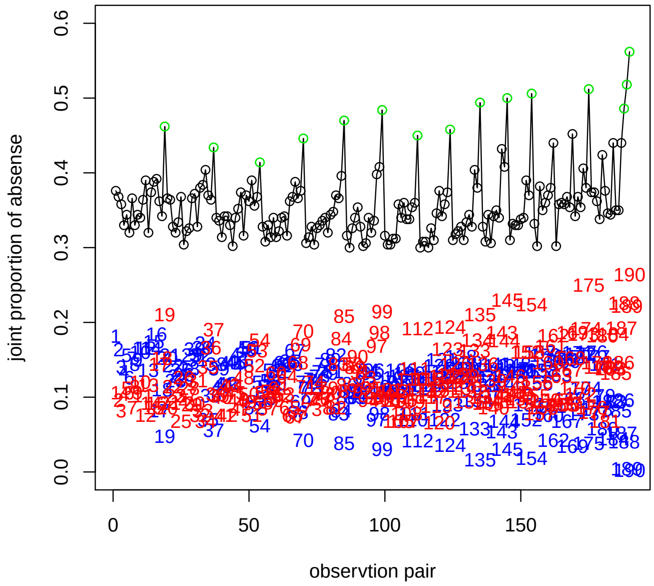


Figure 5.7: Toy2DJointDiag

Similarly we can also easily extend it to three-way joint proportion of absence, shown in Figure 5.8.

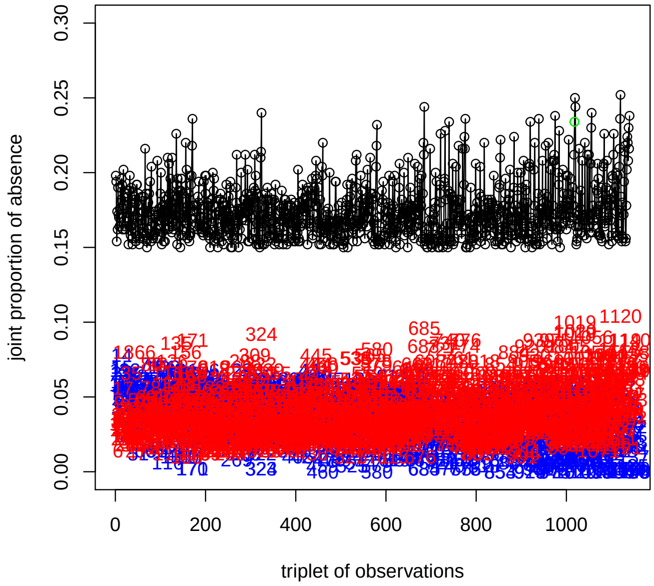


Figure 5.8: Toy2DJoint3Diag

Notice the periodic behavior of the spikes, seemingly increasing in frequency from left to right. Take Figure 5.7, the spikes correspond to pairs including observation 20, so the first spike would be the pair (1,20) while the second would be (2,20) and so on as, left to right, they are ordered

$$(1,2),(1,3),\ldots,(1,20),(2,3),(2,4),\ldots,(2,20),\ldots,(18,19),(18,20),(19,20)$$

and if you look closely at each of the spikes at the left of Figure 5.7, it can be seen that most are immediately preceded by two much smaller, often difficult to see, spikes corresponding to pairs involving observations 18 and 19. The notable feature is how the last spike on the right, where the last three points are all quite high, high enough enough that if we were to set some reasonable threshold (which would detect most of the other peaks) like for example one which detects the points highlighted in green, it would detect these three points as well. This is interesting because these three points correspond to the observation pairs (18,19),(18,20) and (19,20) respectively, so if we were to say that from Figure 5.5 only observation 20 was significantly influential, then Figure 5.7 would be telling us that the observation pair (18,19) is significant, and furthermore if we were to define significant observation pairs as those highlighted in green, it would be the only significant pair that *does not* contain the originally significant observation 20. So what we have detected here is a cluster of significant points (observations 18 and 19) that are, together, significant, but separately not. This is something that Cook's distance (in its standard form) is simply not capable of doing.

5.6 Notes on Distributional Results for Joint Proportions of Absence

It is interesting to realize that as noted above,

$$P(X_{b,j}=0)=\left(\frac{n-1}{n}\right)^n \;(\approx 0.367 \text{ for } n=164)$$

That for $i \neq j$,

$$\begin{aligned}
P\left(X_{b,j}=0 \cap X_{b,i}=0\right) &= P(X_{b,j}=0)P(X_{b,j}=0|X_{b,i}=0)\\
&= \left(\frac{n-1}{n}\right)^n \binom{n}{0}\left(\frac{1}{n-1}\right)^0\left(1-\frac{1}{n-1}\right)^n\\
&\qquad \text{as } X_{b,j}|(X_{b,i}=0)\sim Bin\left(n,\frac{1}{n-1}\right)\\
&= \left(\frac{n-1}{n}\right)^n\left(\frac{n-2}{n-1}\right)^n\\
&= \left(\frac{n-2}{n}\right)^n \;(\approx 0.134 \text{ for } n=164)
\end{aligned}$$

And similarly for $i \neq j$, $j \neq k$, and $i \neq k$,

$$\begin{aligned}
P\left(X_{b,j}=0 \cap X_{b,i}=0 \cap X_{b,k}=0\right) &= P(X_{b,j}=0 \cap X_{b,i}=0)P(X_{b,k}=0|X_{b,i}=0 \cap X_{b,j}=0)\\
&= \left(\frac{n-2}{n}\right)^n \binom{n}{0}\left(\frac{1}{n-2}\right)^0\left(1-\frac{1}{n-2}\right)^n\\
&\qquad \text{as } X_{b,k}|(X_{b,i}=0 \cap X_{b,j}=0)\sim Bin\left(n,\frac{1}{n-2}\right)\\
&= \left(\frac{n-2}{n}\right)^n\left(\frac{n-3}{n-2}\right)^n\\
&= \left(\frac{n-3}{n}\right)^n \;(\approx 0.0484 \text{ for } n=164)
\end{aligned}$$

5.7 Bootstrap Diagnostics

The one-way proportion of absence plots as in Figure 5.5 for each of the $\binom{n}{2} = 351$ correlation coefficients between trace elements are available on request. But in summary the following are some of the more interesting points identified as significant:

Case	Correlations containing this observation as significant are between elements that include:								
25	Au	Ag	Ti	Cr	W	Sb	Cu	Mn	Zn
	Se	Te	Nb	V	Mo	Sn	Tl	Bi	
37	Au	Ti	Cr	W	V	Mn	Ge	Nb	U
	Se	Te							
42	Mn	Ni	Co	Ge					
48	Au	As	Ag	Sb	Co	Re			
117	Au	As	Sn	Cu	Zn	Cd	Tl		
141	Ag	Tl	Sb	Re	Pb				
143	Au	Te	Se						

Table 5.2: Particularly influential points detected by proportion of absence

So we see from Table 5.2 that this picks up the expected case 25, which turns out to be significant in many more correlations than just the Cr, W, Ti group noted in Chapter 3, which along with the Se and Te appear in both correlations affected by cases 25, 37, and 143. Case 42 appears to be related to the Co, Ni pair, while cases 48 and 117 affect correlations involving Au and As. Case 141 appears to be related to the Tl, Sb pair noted in Chapter 3, and case 143 turns up as significant in the Au, Te correlation. The results in Table 5.2 may be somewhat misleading however, as although two elements may be listed next to a case this does not necessarily mean that that case is influential on the correlation coefficient between those two elements. It only means that it is influential on some correlation coefficients between elements including those listed. For example, case 42 does in fact appear as influential on the correlation coefficient between Co and Ni, and case 143 appears influential on the correlation between Se and Te, but although Se and Te are listed next to case no. 25 and 37, neither of these cases seem to be directly influential on the correlation between Se and Te. So although Table 5.2 provides a very basic outline of the most influential points, it does miss some interesting information relating to the structure of these influences.

Furthermore although the points noted in Table 5.2 turn up as significant according to the proportion of absence statistic, some of the correlations in which they are influential such as the correlation between Co and Ni, or Sb and Tl, have already been shown to be robust to extreme values in Chapter 3, i.e. still appear to have very strong (close to 1) Pearson's correlation on the log-transformed scale, and Spearman's Rho coefficient. Intuitively this may seem contradictory, as we have mainly considered influential points as points which significantly increase the correlation coefficient, but this is not the only kind of influence a point can have on a correlation: it could also reduce a correlation coefficient. It can be seen in Table 3.3 that these correlation coefficients are larger on the log-transformed scale than on the raw scale, which supports this idea, i.e. there is a stronger correlation when the effect of extreme values is reduced by considering them on a logarithmic scale. The choice of $|\mathcal{B}| = \frac{B}{2}$ also supports this idea, as this symmetrical choice of $|\mathcal{B}|$ suggests that this statistic should detect negatively influential points just as well as positively influential points (where a negatively influential point is a point that reduces a correlation, and vice versa for a positively influential point).

5.8 Future Work

There is a great deal of research that could still be done on this bootstrapping method, so here I will list some of the things that I think would be particularly interesting to look into:

- Investigate the effect of adjusting $|\mathscr{B}|$, particularly in improving detection of positively influential points at the expense of detection of negatively influential points, or vice versa.
- Investigate the effect of reducing B, as I suspect this method would still be very effective for relatively small values of B, and for larger datasets, smaller B would have the important effect of reducing the required processing time significantly.
- Develop a systematic (and better justified) method for deciding on a cut-off value for what is considered a significant value in proportion of absence (I simply chose an arbitrary value that seemed reasonable).
- Develop a better method for displaying the results of such an analysis, as noted above Table 5.2 is perhaps not the best way to do this.
- Develop a systematic method for detecting clusters of influential points, as in the example in Section 5.5, and further develop the distributional results for the joint proportions of absence. Perhaps even incorporate a method for random selection of joint proportions of absence to test, as for larger datasets the number of such combinations would quickly become very large.
- For datasets with extremely large n, using the standard bootstrapping approach of taking random samples with replacement of the same size as the original dataset would quickly become very slow. So use of other methods could be considered, jackknifing for example (although proportion of absence wouldnt work with jackknifing). Possibly even taking random samples with replacement of a size significantly smaller than n could be considered. Given such a generalization, I think this method could be implemented in an anomaly detection context for extremely large datasets.

Chapter 6

Discriminant Analysis

In this chapter we will try several different methods for clustering and classification in the more ‘classical’ perspective of treating the 164 cases as the observations being clustered (compared with the less conventional clustering performed in Chapter 3). First of all there is a big difference between cluster analysis and discriminant/classification analysis, and I use both in this chapter, I call the chapter Discriminant Analysis because that is the main focus of the chapter. Cluster analysis is the process of separating the data into groups based purely on the data itself, without knowledge of any true class membership. Classification/discriminant analysis is the process of developing a rule to assign any given observation to a class, based on knowledge of actual class membership for some data. We are particularly interested on if it is possible to accurately classify the groups shown in Table 1.1(repeated here as Table 6.1 for convenient reference) for which the true class membership is known (as it was determined by the geologists by careful inspection of the grains of pyrite from which these measurements were taken).

Symbol	**Morphology**	**Location**	#
○	Granular	Rock	98
△	Granular	Vein	22
○	Replacement	Rock	14
△	Replacement	Vein	30
		Total:	164

Table 6.1: Classes of Pyrite Legend

If we consider only the first and second principal components from Figure 4.8, shown below in Figure 6.1 with the known classes shown in colour as in Table 6.1.

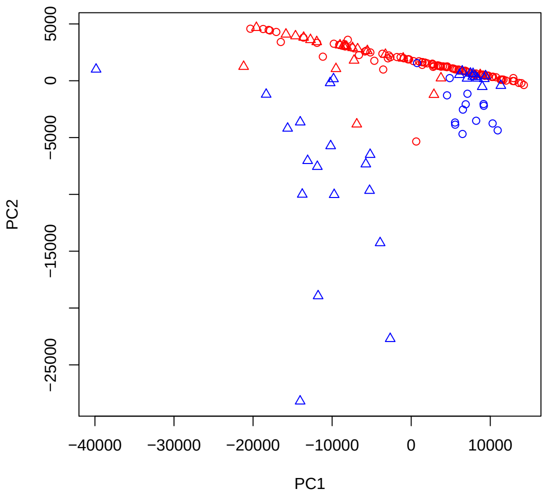


Figure 6.1: known classes displayed on PC1 vs PC2

Figure 6.1 is encouraging in that it contains some interesting information, particularly there are two groups within the RV (△) class, one which is separable from the rest of the data, and the other which might not be, rather being a part of the linear trend followed by much of the data in span$\{\eta_1, \eta_2\}$ i.e. the plane spanned by the first two principal components (which I interchangeably refer to as PC1, PC2 or η_1, η_2). Also, the RR (○) class seems possibly separable, and the GR and GV (○ and △ respectively) seem to be towards the left, and right, ends of the strong linear relation apparent in Figure 6.1 respectively.

I will be using this plane (of the first two principal components, i.e. Figure 6.1) consistently throughout this chapter, so that different classification methods can be directly compared. This is an interesting plane to begin with, as it encompasses 96% of the variability in the data, as shown in the principal component analysis in Chapter 4. Furthermore, as noted in Chapter 4, $PC1 \approx -0.99As - 0.14Mo$ and $PC2 \approx 0.14As - 0.98Mo$ (where PC1 and PC2 are the first and second principal components respectively). So the strongly linear negative line clealry evident in Figure 6.1 actually corresponds to the concentration of Mo being zero, as about 19% of the cases have zero concentration of Mo (as can be seen in Table 2.2). So in Figure 6.2 we see an interesting pattern when we colour the data by the concentration of Mo, specifically:

	colour
$x_{iMo} \geq 10000$	red
$10000 > x_{iMo} \geq 1000$	orange
$1000 > x_{iMo} \geq 100$	green
$100 > x_{iMo} > 0$	blue
$x_{iMo} = 0$	purple

Table 6.2: Colouring scheme for concentration of Mo

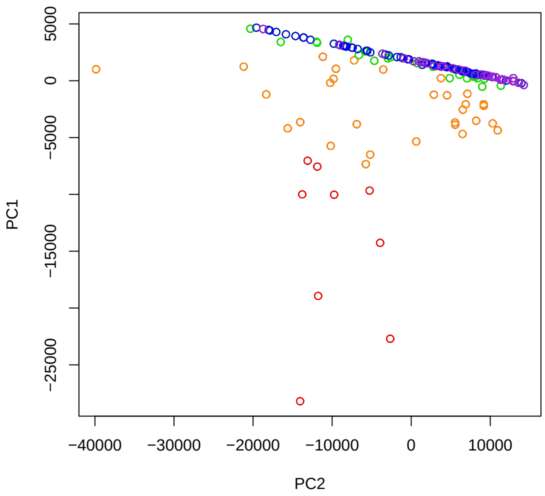


Figure 6.2: PC1 against PC2 plane with data coloured by concentration of Mo as in Table 6.2

6.1 k-means Clustering

First we try a k-means clustering algorithm (from [6], the default method in R, a detailed explanation of which is in [5]), a Euclidean distance based method, to get an idea of how the observations are clustered, then later we will try some Classification/ Discriminant Analysis (otherwise known as supervised or machine learning) and compare the results back to these, to see if they perform better in discriminating between the classes in Table 6.1.

This clustering method splits the observations into a pre-defined constant k number of clusters. Choice of the constant k is a difficult one to justify, but in this case as we know we are later going to try to classify these data into morphologies and locations as described in Table 1.1 we know that we are particularly interested in the $k = 4$ case. Ideally these would separate the data into the classes mentioned, but if it does not, it might pick up some other clusters in the data, which would then be of interest. This method is, as is hierarchical agglomerative cluster analysis from Chapter 3, described more fully in [8, Chap. 6].

6.1.1 4-class problem

This algorithm does not always produce the same clusters, depending on the random seed points it begins with, so here I present the two most dominant clusters, one of which will come up about 75% of the time this algorithm is run.

The first produces clusters with classes as shown in Table 6.3 below:

And we can display this classification on the $PC1$ vs $PC2$ plane shown above in Figure 6.1, marking cluster 1 as ○, cluster 2 as △ cluster 3 as ○ and cluster 4 as △, in Figure 6.3.

	Class				Morphology		Location		
Cluster	GR	GV	RR	RV	Granular	Replacement	Rock	Vein	Total
1	0	1	0	12	1	12	0	13	13
2	39	5	1	0	44	1	40	5	45
3	42	1	13	12	43	25	55	13	68
4	17	15	0	6	32	6	17	21	38
Total	98	22	14	30	120	44	112	52	164

Table 6.3: Assigned cluster membership compared with true class membership for 4-means cluster analysis(1)

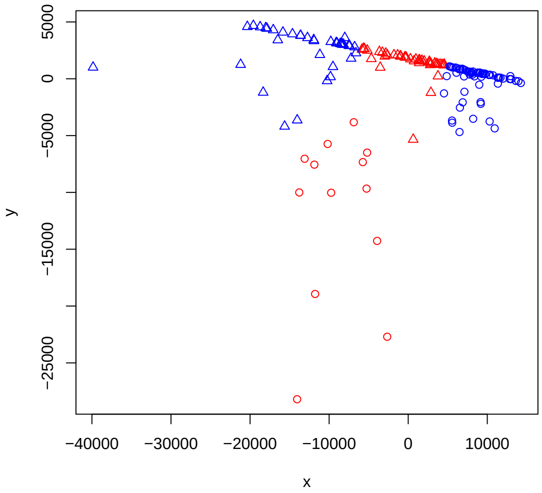


Figure 6.3: clusters displayed on the plane y:PC1 against x:PC2

While the second produces clusters with classes as shown in Table 6.4 below:

	Class				Morphology		Location		
Cluster	GR	GV	RR	RV	granular	replacement	rock	vein	Total
1	34	3	3	0	37	3	37	3	40
2	38	1	11	12	39	23	49	13	62
3	7	6	0	9	13	9	7	15	22
4	19	12	0	9	31	9	19	21	40
Total	98	22	14	30	120	44	112	52	164

Table 6.4: Assigned cluster membership compared with true class membership for 4-means cluster analysis(2)

And we can display this classification on the $PC1$ vs $PC2$ plane shown above in Figure 6.1, marking cluster 1 as ○, cluster 2 as △ cluster 3 as ○ and cluster 4 as △, in Figure 6.3.

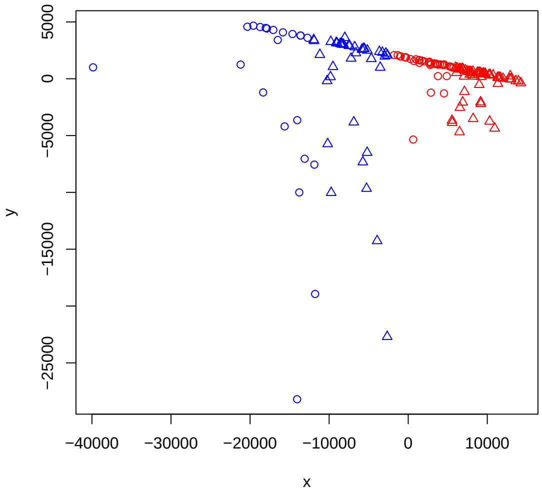


Figure 6.4: clusters displayed on the plane y:PC1 against x:PC2

So we can see that this is picking up the same trend we could see in Figure 6.1: that there are several groups lying more or less on the linear line (that is roughly equivalent to pyrite with very small concentrations of Mo), which are potentially quite difficult to separate by their location on that line, i.e. by their As concentration specifically as the biggest class, GR (○) is spread out across the length of the line (across the range of concentrations of As), while there are two other groups, with significant concentrations of Mo.

6.1.2 2-classproblem

The $k = 2$ kmeans clustering algorithm consistently returns the same clusters consistently, particularly:

```
K-means clustering with 2 clusters of sizes 62, 102

Cluster means:
         Au       Ag       As       Sb       Ti        V
1 136.66452 441.2079 25201.40 558.0635 443.6981 9.790645
2  30.11725 180.9450  8515.38 378.7622 188.9340 2.913235
        Cr       Mn       Co       Ni       Cu       Zn
1 2.116129 11.01258  47.9529 137.4310 468.2544 42.35855
2 0.852549 24.62157 273.8697 126.9736 479.9025 20.86784
         Ga       Ge        Se        Nb        Mo        Cd
1 0.1161290 5.514194 105.16032 0.2824194 3769.7808 65.348226
2 0.2517647 5.593529  80.63667 0.2579412  580.5716  9.526863
          In        Sn       Te         W         Re       Tl
1 0.18596774 0.5708065 2.290161 4.9677419 0.20435484 65.75419
2 0.04965686 0.3816667 1.133039 0.6441471 0.06833333 31.46265
        Pb         Bi          U
1 510.0063 0.02853226 0.42611290
2 158.9097 3.18532353 0.09286275
```

And we can similarly display this classification on the $PC1$ vs $PC2$, marking cluster 1 as ○ and cluster 2 as ○.

	Class				Morphology		Location		
Cluster	GR	GV	RR	RV	granular	replacement	rock	vein	Total
1	26	18	0	18	44	18	26	36	62
2	72	4	14	12	76	26	86	16	102
Total	98	22	14	30	120	44	112	52	164

Table 6.5: Assigned cluster membership compared with true class membership for 2-means cluster analysis

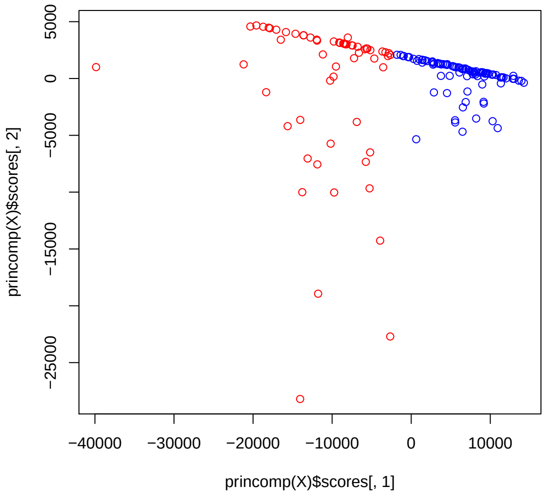


Figure 6.5: clusters displayed on PC1 vs PC2

This is particularly interesting, as if we compare Figure 6.5 to Figure 6.1 we can see what can also be seen in Table 6.5: that cluster one is composed of almost all of the GV (△) class, about a quarter of the GR (○) class, and about half of the RV (△) class. While cluster 2 is composed of the rest. So cluster 1 contains almost all of the GV (△) class, while cluster 2 contains all of the RR (○) class.

6.2 Fisher's Linear Discrimination Rule

This section draws many results from [8, Chap. 4], but uses the method originally proposed by Fisher [4].

K-means cluster analysis is a method for finding groups or clusters of observations within a dataset, without knowledge of existing classes. When we incorporate knowledge about different types of observations, this becomes a different problem, commonly referred to as Discriminant Analysis or Classification. Fisher's Linear Discriminant Rule [4] is possibly the simplest, certainly one of the most intuitive, methods for doing this. The fundamental idea in Fisher's linear discrimination is, somewhat similarly to principal component analysis, to find the direction (or linear combination of the variables) that when the data is projected into that direction, gives us the properties we want. Specifically the properties we are interested in is maximizing the between-class variance, and minimizing the within-class variance. Similarly to principal component analysis, we can obtain this direction as the eigenvector corresponding to the largest eigenvalue of a matrix. Particularly we are interested in the matrix $W^{-1}B$, where loosely W is the within-class covariance matrix, and B is the between-class covariance matrix. More formally, if we have knowledge about the class membership (which observation belongs to which class) for κ classes. Then for each $\nu \in [1, \kappa] \cap \mathbb{Z}$, we can obtain class-specific means and covariance matrices, μ_ν and Σ_ν, respectively (i.e. an observation from class ν, $\boldsymbol{X}^{[\nu]} \sim (\mu_\nu, \Sigma_\nu)$, i.e. $\mathrm{E}[\boldsymbol{X}^{[\nu]}] = \mu_\nu$ and $\mathrm{var}(\boldsymbol{X}^{[\nu]}) = \Sigma_\nu$). Then

$$B = \sum_{\nu=1}^{K} (\mu_\nu - \bar{\mu})(\mu_\nu - \bar{\mu})^T \qquad W = \sum_{\nu=1}^{K} \Sigma_\nu \qquad \left(\bar{\mu} = \frac{1}{K} \sum_{\nu=1}^{K} \mu_\nu\right) \tag{6.1}$$

Taking the first eigenvector, call it $\boldsymbol{\eta}$ (the eigenvector corresponding to the largest eigenvalue) of the matrix $W^{-1}B$, Fisher's linear discriminant rule is to assign an observation $\boldsymbol{x}_{i\bullet}$ to class ν such that ν minimizes the quantity

$$|\boldsymbol{x}_{i\bullet}^T \boldsymbol{\eta} - \boldsymbol{\mu}_\nu^T \boldsymbol{\eta}|$$

This can a nice, intuitive interpretation, basically what we are doing here is projecting the data into the 1-dimensional subspace spanned by η, i.e. the direction which maximizes the between-cluster variability, and minimizes the within-cluster variability, then given we know the true class membership, we then assign each observation to the class, ν, whose class mean $\boldsymbol{\mu}_\nu$ it is closest to. That is $\boldsymbol{\mu}_\nu$, projected into the η direction. Traditionally this is used primarily for 2-class problems, as that is usually the case it is best suited for, but technically there is no reason that restriction is necessary. So applying this technique to the 4-class problem, we get the assignment shown below in Table 6.6

	Assigned	Actual Class				
Symbol	Class	GR	GV	RR	RV	Total
○	GR	25	2	1	0	28
△	GV	17	16	0	13	46
○	RR	44	1	13	12	70
△	RV	12	3	0	5	20
	Total	98	22	14	30	164

Table 6.6: Assignment using Fisher's linear discriminant rule on the 4-class problem

So this method misclassified 105 cases, which corresponds to the sum of the off-diagonal terms in Table 6.6. What is interesting to note here though is that it correctly classifies almost all of the RR (○) class. The other interesting feature of this classification is that although almost all of the RV (△) class are misclassified, they are separated about half half between two different classifications, which likely results from the sub-classes we have noted as seemingly present in this class.

We can also consider the two 2-class problems of classifying either by morphology, or location separately. So applying Fisher's linear discriminant rule to classifying morphology gives the assignment shown in Table 6.7

	Assigned	Actual Class		
Symbol	Class	Granular	Replacement	Total
○ △	Granular	118	18	136
○ △	Replacement	3	25	28
	Total	121	43	164

Table 6.7: Assigned class against true class for Fisher's linear discriminant classification on morphology

Which actually preforms quite well, only misclassifying 21 cases, while applying the same rule to classify location provides the classification shown in Table 6.8.

	Assigned	Actual Class		
Symbol	Class	Rock	Vein	Total
○ ○	Rock	86	16	102
△ △	Vein	26	36	62
	Total	112	52	164

Table 6.8: Assigned class against true class for Fisher's linear discriminant classification on location

Which only misclassified 42 cases. But if we use these two 2-class problems to classify our 4-class problem (i.e. we classify cases that are assigned to granular in Table 6.7 and rock in Table 6.8 to GR, etc.). Using this classification rule provides the assignment shown below in Table 6.9

	Assigned	Actual Class				
Symbol	Class	GR	GV	RR	RV	Total
○	GR	71	4	4	12	91
△	GV	26	17	0	2	45
○	RR	1	0	10	0	11
△	RV	0	1	0	16	17
	Total	98	22	14	30	164

Table 6.9: Classification by 2-way 2-class Fisher's discrimination

This method only misclassified 50 cases, which is significantly better than the 105 when Fisher's rule is directly applied to the 4-class problem as shown in the Table 6.6. This is likely because the Fisher's rule is limited to considering one direction, as it projects the data into a single direction η, and is really optimal for 2-class problems. While this 2-way method has more freedom in that it considers two directions, as it projects the data into two different directions, one to differentiate morphology, and the other to differentiate location. This seems appropriate for this problem, because the four classes results for two bivalued classes, but by a similar logic as that for saying a single direction is appropriate for a two class problem would lead us to state that for a general 4-class problem it would be reasonable to consider up to 3 directions. The default linear discriminant analysis function `lda()` in R automatically does this, and performs extremely well.

6.3 Another method, lda() in R

The default linear discriminant analysis function in R performs surprisingly well, producing the classification shown in Table 6.10.

	Assigned	Actual Class				
Symbol	Class	GR	GV	RR	RV	Total
○	GR	94	6	2	6	108
△	GV	2	15	0	0	17
○	RR	0	0	12	2	14
△	RV	2	1	0	22	25
	Total	98	22	14	30	164

Table 6.10: Assigned classification by `lda()` against true class membership

This method only misclassified 21 cases, which is the lowest misclassification of all the methods I tried. Particularly notable is how it actually identifies the GR (○) class quite well (which none of the other methods succeeded to do), and from Figure 6.1 it was unclear if it would be possible to do so so this result alone is interesting, as it would appear that classification of this class relies on concentrations of elements other than As and Mo.

6.4 Futher Work

These results are very interesting, but the area of discriminant analysis is huge, and there are several things that could still be done to further this chapter.

- Further investigate the mechanics, and thus interpretation of the results produced by the `lda()` method in R
- Implement the method of variable ranking: i.e. rather than using directions (linear combinations of the variables), use the variables themselves. I would expect Mo and As to come up as important, and then hopefully some other elements whose significance would be particularly interesting.

Chapter 7

Regression

7.1 Introduction

In this chapter we demonstrate that it is possible, in the framework of simple linear regression, to construct a model of this data that predicts the amount of gold quite well based on the concentrations of the other trace elements. Particularly, that by such methods we can predict gold significantly better than simply based on As concentration (we use As as a baseline to compare too as the As Au relation is well established [10]). Beyond this demonstration however, most of the methods and processes in this chapter are largely ad hoc, and not systematic as the small size of the data set prevents us from considering larger models (including all interaction terms), and we are restricted to only considering a small subset of all possible models. I try be as rigorous as possible, but there will still be parts that are 'glossed over'. So it is possible we miss important relations (such as high order interactions) between the variables in our approach, which it would be reasonable to assume exist, as higher order interactions are likely to exist in this type of data. Because of this, interpretation of the final model in terms of relationships between the elements, and other variables, should only be considered with caution, as there are likely to be other underlying affects that have not been compensated for. However, this being said, results pertaining to the accuracy of prediction of gold, without interpretation of the source or meaning of such prediction in any more detail, are still perfectly valid and this is the purpose of this chapter.

7.2 Use of the log-transformed concentrations

So far, i have not justified the use of the log-transformation (as in Table 2.1) on the concentrations of the trace elements beyond claiming that it is common practice amongst geologists and geochemists working with trace element data. I do not claim to do so here. What i propose here is a a maximum likelihood method that suggests it is appropriate and although i in no way claim this is a full justification, i do suggest that it is an empirical test of sorts. That is to say that if taking the log-transformed data was appropriate, then we would expect these results, and so if these methods showed something significantly different than what they do, the log-transformation would be inappropriate. So what i provide here is not a justification for using the log-transformation, but rather empirical support for it.

7.2.1 Box-Cox and other MLE methods

In Box-Cox transformations, we consider the transformed variable

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}+1}{\lambda} & \lambda \neq 0 \\ ln(y) & \lambda = 0 \end{cases} \tag{7.1}$$

Thus, under the paradigm of the simple linear model, and the associated normality assumption, this yields the likelihood function (as suggested in [1] for λ as shown below in Equation 7.2

$$L = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{-\frac{(\boldsymbol{y}^{(\lambda)}-\boldsymbol{a\theta})^T(\boldsymbol{y}^{(\lambda)}-\boldsymbol{a\theta})}{2\sigma^2}} J(\lambda;\boldsymbol{y}) \qquad E[\boldsymbol{y}^{(\lambda)}] = \boldsymbol{a\theta} \tag{7.2}$$

So, if we construct the full no-interaction model from the log-transformed data, but before taking the log of the Au (response) variable, and maximize this likelihood function with respect to it, as shown below, we get Figure 7.1 which shows how under those (quite restrictive) conditions, it is optimal to use the ($\lambda = 0$) log transform.

```
> lm1 = lm((X$Au+1)˜.-Au,data=LogX)
```

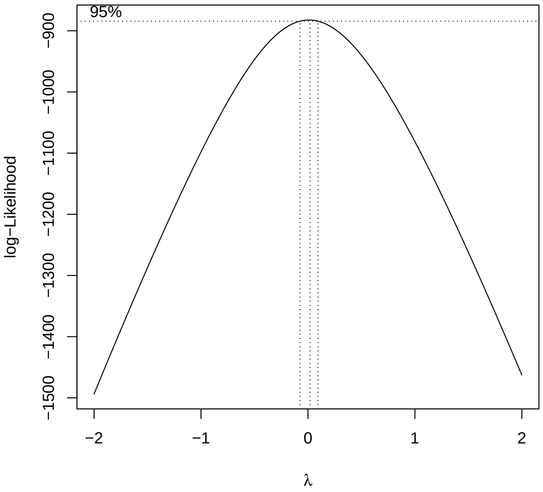


Figure 7.1: log-likelihood plot for lambda on the model lm1 shown above

Furthermore, consider the two full (no interaction) models Au.lm000 and logAu.lm000 as shown below, and the associated diagnostic plots shown in Figure 7.2 and Figure 7.3. The regression assumptions of linearity, homoscedasticity, and even normality of the residuals seem somewhat reasonable for the log-transformed data, but not at all for the raw data.

```
> Au.lm000 = lm(Au˜.,data=X)
> logAu.lm000 = lm(Au˜.,data=LogX)
```

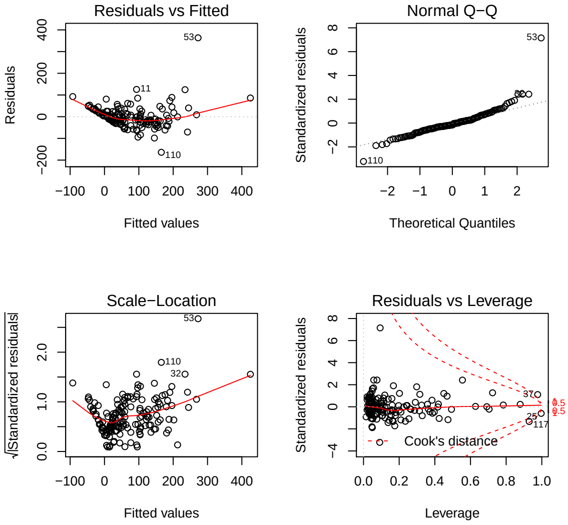


Figure 7.2: Diagnostic Plots for above Au.lm000

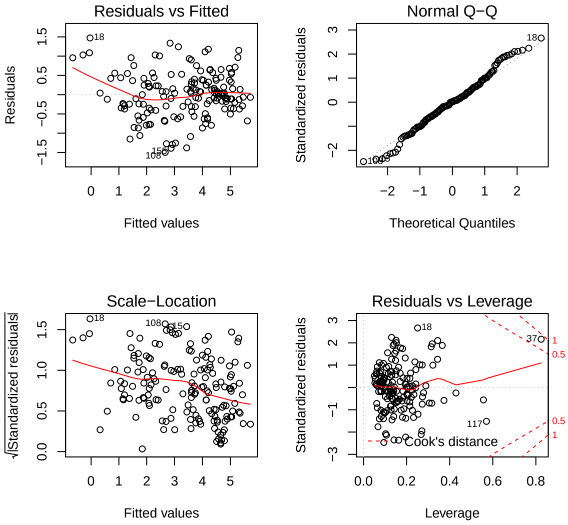


Figure 7.3: Diagnostic Plots for above logAu.lm000

7.3 No Interaction Models

We discard the raw data model, as even in the full no-interaction model, the regression does not satisfy the assumptions, and work with the log-transformed data from here on. Given the known relationship between Au and As [10], we first consider the minimal `logAu.lmAs` model shown below of prediction of Au purely by As concentration.

```
> logAu.lmAs = lm(Au˜As,data=LogX)
> summary(logAu.lmAs)
```

```
Call:
lm(formula = Au ~ As, data = LogX)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7374 -0.6064  0.3811  0.8604  3.2042

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.8400     1.2028  -6.518 8.61e-10 ***
As            1.1973     0.1281   9.350  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.248 on 162 degrees of freedom
Multiple R-squared: 0.3505,     Adjusted R-squared: 0.3465
F-statistic: 87.43 on 1 and 162 DF,  p-value: < 2.2e-16
```

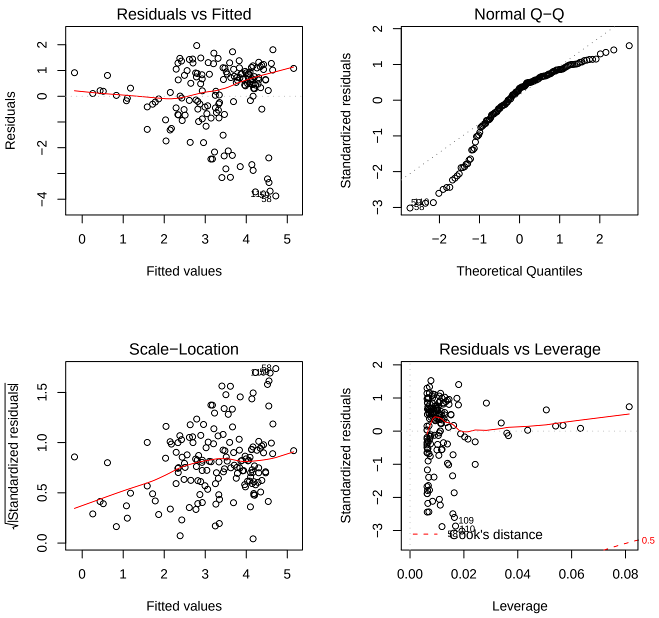


Figure 7.4: Diagnostic Plots for above logAu.lmAs

We can see from Figure 7.4 that the minimal (`logAu.lmAs`) model does not satisfy the regression assumptions of linearity or homoscedasticity, so we want a more parsimonious model than the full no interaction model `logAu.lm000`, but better than the minimal `logAu.lmAs` model. To do this, i use backwards elimination by probability, as shown in Algorithm 7.1:

```
Algorithm 7.1:
    Step 1.   Choose a significance level $\alpha$, where $0 < \alpha < 1$
    Step 2.   Find the term with the highest p-value, call it $p$
    Step 3.   IF($p \geq \alpha$)     Remove the term associated with $p$
                                      Return to Step 2.
              ELSE()                  Stop.
```

If we start with the full no interaction model `logAu.lm000`, and use Algorithm 7.1 above to a significance level of $\alpha = 0.05$, we get the model:

```
> summary(logAu.lm019)

Call:
lm(formula = Au ˜ As + Co + Zn + Ge + Se + Mo + Cd, data = LogX)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48936 -0.35116 -0.04851  0.38320  1.41684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.24831    1.61549  -2.011 0.046078 *
As           0.83528    0.07097  11.770  < 2e-16 ***
Co          -0.08329    0.02382  -3.497 0.000613 ***
Zn           0.15757    0.03637   4.333 2.62e-05 ***
Ge          -2.25251    0.76798  -2.933 0.003863 **
Se           0.63043    0.05782  10.902  < 2e-16 ***
Mo           0.08621    0.01939   4.446 1.66e-05 ***
Cd           0.09829    0.04322   2.274 0.024322 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6541 on 156 degrees of freedom
Multiple R-squared: 0.8282,     Adjusted R-squared: 0.8205
F-statistic: 107.5 on 7 and 156 DF,  p-value: < 2.2e-16
```

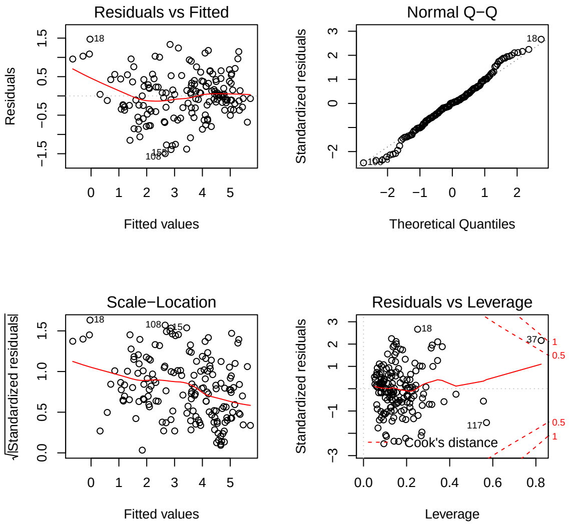


Figure 7.5: Diagnostic Plots for the reduced no-interaction model above

So from Figure 7.5, and the adjusted R-squared value of 0.8205 we can see this reduced no interaction model is quite a good fit, satisfies the regression assumptions of linearity, homoscedasticity, normality of residuals quite well and predicts concentration of Au quite well. We can probably still do better though, so we will also consider some interaction terms.

7.4 Interaction Models

Now although we cannot consider all possible interaction terms due to the small size of the dataset, we can consider all the two-way interaction terms in the reduced `logAu.lm019` model. Then implementing Algorithm 7.1 again with $\alpha = 0.05$ but restricting to only allow removal of interaction terms (to satisfy the marginality principle) we get:

```
Call:
lm(formula = Au ~ (As + Co + Zn + Ge + Se + Mo + Cd) + As:Co +
    As:Zn + As:Se + As:Cd + Co:Cd + Se:Mo + Se:Cd + Mo:Cd, data = logX)

Residuals:
     Min       1Q   Median       3Q      Max
-1.80491 -0.31286  0.00187  0.25964  1.81686

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.36231    3.13892   1.708 0.089672 .
As          -0.26635    0.29267  -0.910 0.364273
Co          -1.00593    0.35145  -2.862 0.004817 **
Zn           1.55781    0.50539   3.082 0.002450 **
Ge          -1.12522    0.69928  -1.609 0.109722
Se          -2.81557    0.57212  -4.921 2.27e-06 ***
Mo           0.43406    0.11648   3.726 0.000276 ***
Cd           0.72295    0.73505   0.984 0.326946
As:Co        0.09000    0.03720   2.419 0.016762 *
As:Zn       -0.15136    0.05292  -2.860 0.004845 **
As:Se        0.35982    0.05929   6.068 1.03e-08 ***
As:Cd       -0.15444    0.07666  -2.015 0.045757 *
Co:Cd        0.05796    0.02368   2.447 0.015557 *
Se:Mo       -0.07099    0.02448  -2.900 0.004303 **
Se:Cd        0.23139    0.06672   3.468 0.000687 ***
Mo:Cd       -0.05860    0.01726  -3.395 0.000881 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5575 on 148 degrees of freedom
Multiple R-squared: 0.8816,     Adjusted R-squared: 0.8696
F-statistic: 73.48 on 15 and 148 DF,  p-value: < 2.2e-16
```

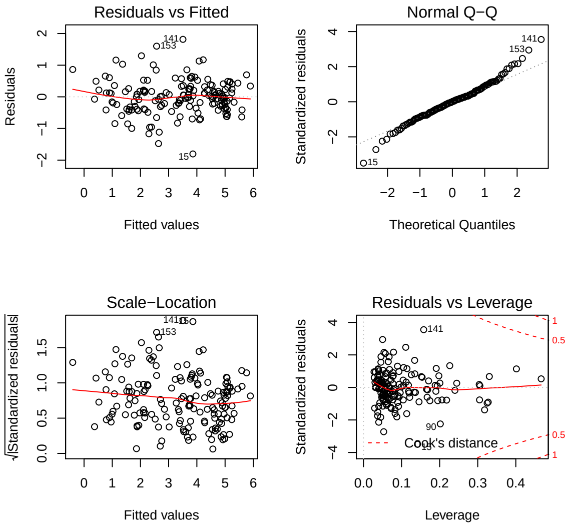


Figure 7.6: Diagnostic Plots for the reduced interaction model shown above

So in short we have demonstrated that accurate prediction of the log-transformed Au concentration from the other trace elements by linear regression is possible, and successfully constructed an accurate model to that effect (Figure 7.6).

Chapter 8

Conclusion

In **Chapter 2** I introduced the associated variables of depth and location, and briefly justified their absence in the remainder of the analyses. I then consider the distribution of each of the trace elements in the dataset, and found that given the skewness and zero-values present in many elements, that the zero-inflated log-normal model described in Equation 2.2 is an appropriate model for this type of data.

In **Chapter 3** I introduced a very nice way of visualizing the pairwise correlations matrix *C* in a dendrogram by hierarchical agglomerative cluster analysis, which introduced the concept of elements being clustered by mutual correlation and produced these plots for the Pearson's correlation coefficient on the raw, and log-transformed data, and for the Spearman's rho rank-based measure of association.

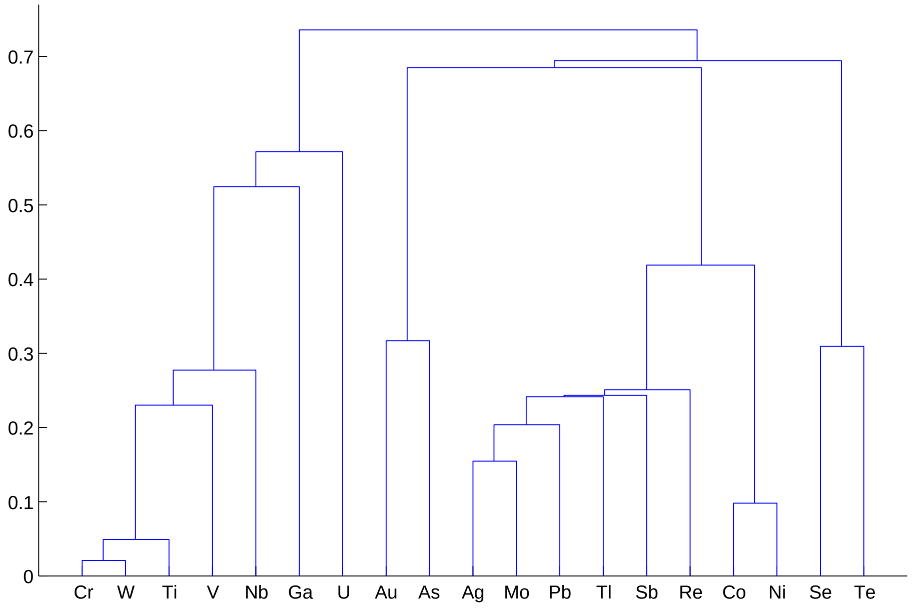


Figure 8.1: Dendrogram of raw Pearson's correlations between selected elements (Figure 3.5 in Chapter 3)

These plots highlighted several interesting clusters of mutually correlated elements, some of which changed from the raw, to the log, to the Spearman's correlation dendrograms, while other particularly interesting ones persisted across the different measures of association. Two such persistent clusters were of particular interest, one corresponding to the lithophilic group of elements: Cr, W, Ti, Nb, V, and U, and the other large group of mutually correlated elements consisting of Sb, Tl, Ag, Mo, and Re. The lithophilic group of elements has a known significance in geology, and so it is encouraging that they turned up through this method, however the other group of elements does not have an obvious interpretation, and as such is particularly interesting, and perhaps warrants further investigation. Several interesting pairs of elements also turned up consistently occurring together, i.e. consistently being highly correlated to each other across the different measures of association, specifically the pairs Co, Ni and Sb, Tl. Co and Ni have a well known relationship in geology, but to my (limited) knowledge Sb and Tl may not, and thus may be geologically significant somehow, and as such this is the type of result to be presented to the geologists to interpret, and if appropriate investigate further.

In **Chapter 4** I introduced a number of methods for visualizing multidimensional data, starting with parallel coordinate plots.

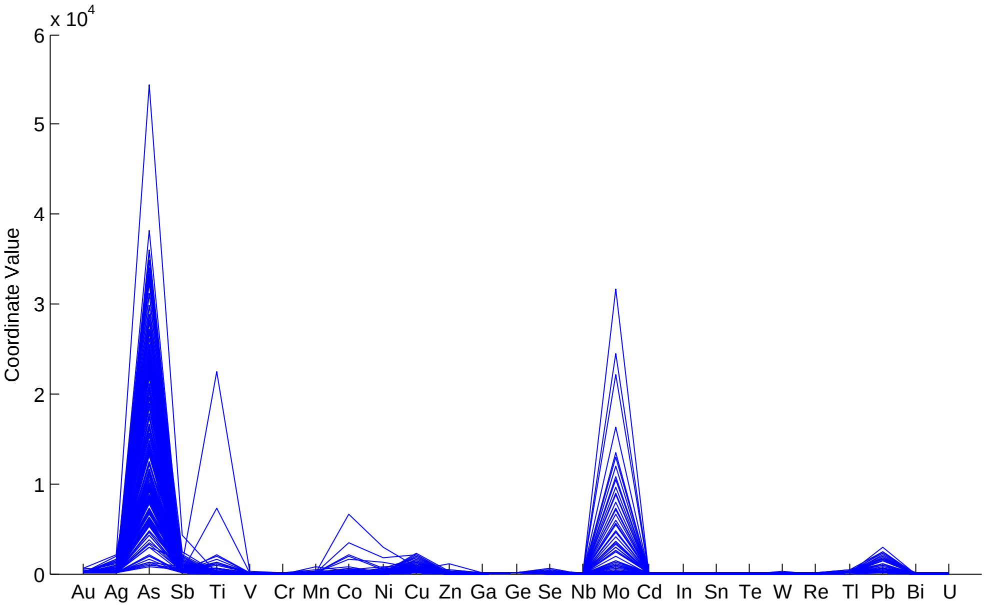


Figure 8.2: Parallel coordinate plot of the raw data(Figure 4.1 in Chapter 4)

These provide us with a great deal of information. The parallel coordinate plot of the raw data (replicated as Figure 8.2 above) provided a very good visual indication of the difference in scale of the different trace elements (that compared with As and Mo, almost all the other elements have insignificantly small values). The parallel coordinate plot of the standardized data then gave a clear indication of the presence of many extreme values in this data as much as much as 10-12 standard deviations from their respective means. Considering parallel coordinate plots of the log-transformed data, and the standardized log-transformed data also yielded interesting conclusions, in that the log-transform successfully brought all the variables onto the same scale as each other, but there are still very extreme values present, although reduced somewhat.

We then investigate the structure of the data further through two more methods, first principle component analysis, and then factor analysis. First we take a closer look at the covariance structure of the data through principle component analysis (finding the directions of maximal variance, and projecting the data into them). PCA on the raw data yielded the expected result; that the vast majority of the variability in the data lies in the As and Mo directions, as the plane spanned by the first two principle components is a close approximation to a rotation of this plane, and explains 96% of the variability in the raw data. We also considered PCA on the standardized data, which is equivalent to using the correlations in place of the covariances. This equivalence explains the results of the principle factor analysis which followed later.

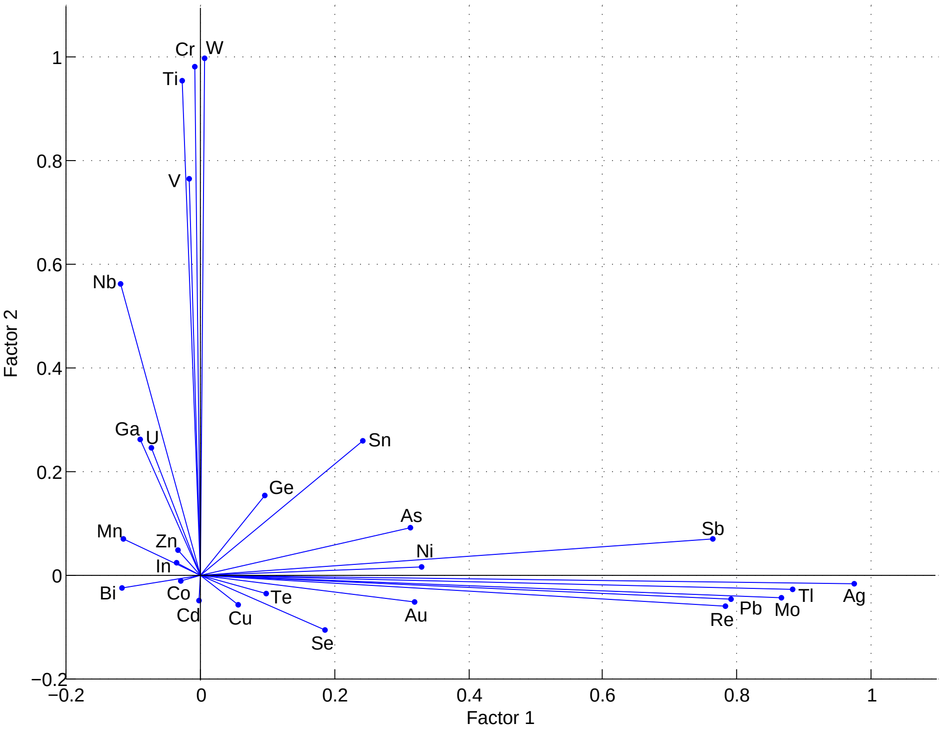


Figure 8.3: Biplot of the 2-factor Analysis on the raw data (Figure 4.14 in Chapter 4)

We considered two types of factor analysis, first using a normal-theory maximum likelihood approach, and then a non-parametric, PCA based approach. Both yielded the interesting result that if we tried to explain the variability in the data with two factors, one would largely explain the variability in a large group of elements (corresponding to the lithophilic elements above), and the other would largely explain the variability in another group of elements (the Sb, Tl, Ag, Mo, Re group mentioned above). Particularly interesting was that when a third factor was added to the model, and the model refitted, the first two factors still explained the same two groups of elements, and the third would largely explain the Co, Ni pair.

In **Chapter 5** I develop a method for identifying influential values that may not be easily identified by other means, and despite there being scope for further research in improving the bootstrap-based method, successfully identified the extreme values which cause most of the discrepancies between Pearson's correlations in the raw data and in the log-transformed data noted in Chapter 3. In particular the most obvious (which was identified in Chapter 3 without this method), case 25, the point responsible for the extremely high correlations between Cr, W, and Ti in the raw data. However this method continued to identify other points of interest as well. For example, how Pb was in a cluster with Sb and Tl in the raw data, but with Co and Ni in the log-transformed data is largely explained by case 141, and the correlation between Se and Te that occurs in the raw data, but is almost completely absent in the log-transformed data is largely explained by case 143.

In **Chapter 6** we consider the third associated variable: class of pyrite, and investigate the relationships these have to the trace element concentrations, which is of particular geological interest. In this chapter I make extensive use of the projection of the data into the span of the first two principle components of the raw data (as mentioned in Chapter 4) to visualize the separation and classification of these data.

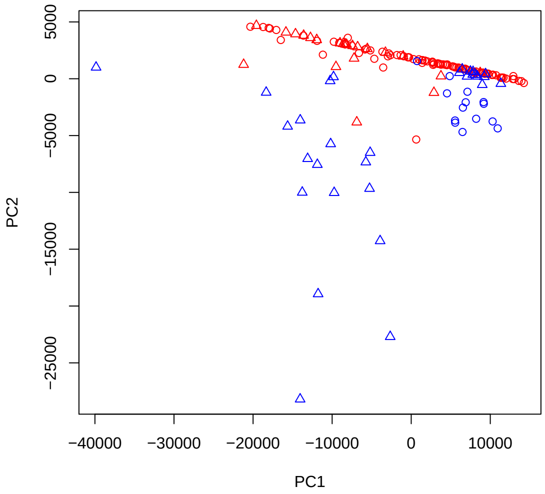


Figure 8.4: Data projected into the plane spanned by the first two principle components of the raw data, with cases coloured according to the true (known) class membership (Figure 6.1 in Chapter 4)

The results of several different clustering and classification methods (k-means clustering, Fisher's linear discriminant rule) confirm what can be seen by careful interpretation of Figure 8.4: that there seems to be a strong relationship between the concentration of Mo and As and the class membership, and that within the replacement, vein (△) class there seems to be two subclasses that can be separated neatly by Mo concentration.

In **Chapter 7** I provide an ad hoc demonstration that accurate prediction of gold concentration is possible from the other trace elements, with surprising accuracy.

So to conclude, we have identified many interesting and potentially meaningful features of this dataset, including

- Relationships between trace elements that could possibly contribute to the understanding of the geochemistry of these pyrite.
- Significant grains of pyrite which can now be singled out and analyzed in more detail, as they are perhaps measurement errors, or special cases which could be of particular interest.
- A particularly interesting separation of one the classes (GV, △) into two sub-classes by their trace element concentrations. So these two sub-groups can now be considered and compared to each other in more detail and perhaps deduce what causes this difference.
- That accurate prediction of gold concentration from the other trace elements is possible in these data.
- Perhaps most notably, as this was the primary purpose of this project, is that we have successfully introduced and developed methods which can now be applied again in the future to visualize and explore this type of data and identify similarly interesting features of other such datasets.

Bibliography

[1] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252, 1964.

[2] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability, 1 edition, May 1994.

[3] B. S. Everitt and G. Dunn. *Applied Multivariate Data Analysis*. Edward Arnold, 1991.

[4] R. A. Fisher. The use of multiple measurements on taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[5] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc. New York, NY, USA, 1975. ISBN: 047135645X.

[6] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):pp. 100–108, 1979.

[7] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 2(2):69–91, 1985.

[8] I. Koch. Analysis of multivariate and high-dimensional data theory and practice. Lecture Notes from 2011 AMSI Summer school at University of Adelaide, to be published as a textbook., December 2010.

[9] M. W. Paul. Geology, geochemistry and mineralogy of epithermal gold ores, moonlight prospect, pajingo, north queensland. Submitted for the degree of Honours in Geology, University of Adelaide, 2010.

[10] M. Reich, S. E. Kesler, S. Utsunomiya, C. S. Palenik, S. L. Chryssoulis, and R. C. Ewing. Solubility of gold in arsenian pyrite. *Geochimica et Cosmochimica Acta*, 69(11):2781–2796, June 2005.

[11] D. B. Rubin and D. T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, March 1982.

[12] B. Sanso and L. Guenni. Venezuelan rainfall data analysed by using a bayesian space-time model. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 48(3):pp. 345–362, 1999.