Chapter 1

Graphical Methods for High-dimensional Data

1.1 Objectives

This chapter explains

- interactive and dynamic graphical methods for exploring data, and for diagnosing models.
- the use of linked brushing, dynamic rotations and interactive re-scaling on high-dimensional data
to explore cluster structure, binary response data, and trends over space and time.
- how to use tour methods to examine data beyond 3 dimensions.

1.2 Introduction

We will discuss interactive and dynamic graphics, using the following principles for visualizing multivariate data:

1. focusing: use simple, easy-to-read plots,
2. linking: provide mechanisms for relating information in one plot to the information in another,
3. arranging: provide mechanisms for shifting and reformatting the layout of multiple plots.

The methods introduced for multivariate visualization are relevant for examining different stages of modeling.

In the initial stage we explore the independent, predictor variable space for non-homogeneous patterns such as outliers (potentially influential points), clusters (sub-populations better modelled as stratifications), multicollinearity (dependence amongst the independent variables). These patterns affect the type of modelling that can be conducted and the inference that subsequently can be made.

Also, in the initial stage we inspect the relationship between the predictor and the independent variables for hints at non-linear dependency. Non-linear dependencies affect the type of modelling that should be done.

After modelling the adequacy of the model can be examined visually by representing the model along with the data, supplementing the data set with diagnostic measures such as residuals, influence measures to explore in relations to the model and actual data values.

The software package that will be used are GGobi (http://www.ggobi.org), a public domain dynamic graphics package. GGobi provides the multivariate graphical tools, encompassing the above visualization principles, that can be applied to problems involving cluster analysis, discriminant analysis, principal component analysis, factor analysis, outlier detection and multivariate analysis of variance.
1.2.1 Notation

We consider a multivariate data set to have a matrix form as follows:

$$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \\ X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}_{n \times p}$$

A 1-D projection of the data into a vector $a_{p \times 1}$ takes the form:

$$Xa = \begin{bmatrix} a_1 X_{11} + a_2 X_{21} + \cdots + a_p X_{p1} \\ \vdots \\ a_1 X_{n1} + a_2 X_{n2} + \cdots + a_p X_{np} \end{bmatrix}_{n \times 1}$$

where $||a|| = \sqrt{a_1^2 + \cdots + a_p^2} = 1$. A 2-D projection of the data can be generated by expanding $a$ to $A_{p \times 2} = [a_1 \ a_2]$ where the columns are orthonormal, $a_1^t a_2 = 0$. Similarly this notation can be expanded to represent $d$-D projections. In GGobi we work with 1-D and 2-D projections only. In VRGobi we work with 3-D, or arbitrary dimensional projections.

1.2.2 Plotting Methods

Beyond using simple univariate plots, such as dotplots, histograms, boxplots, to graph multivariate data variable by variable, there are a wealth of ways that have been devised to draw multiple variables in a single graphic: most notable are matrices of pairwise plots, Andrews’ curves and parallel coordinate plots, and icon plots like Chernoff faces and star charts.

Andrews curves and parallel coordinate plots use a line to represent each observation, where the line tracks the observation’s values on each variable or combination of variables. Icon plots represent each observation by a single icon, and the observation’s value on each variable is represented by a different feature in the icon. For example, with Chernoff faces each observation is represented as a face, and the facial features correspond to different variables, such as the length of the nose represents the numerical value of the observation on variable 1. These methods truly attempt to provide graphics for revealing multivariate structure, most notably clustering. In contrast, matrices of pairwise scatterplots simply extend the univariate graphics an additional dimension so that pairwise relationships between variables can be examined.

Complementary to these unique plotting methods, numerical techniques are often used to reduce the dimensionality of a data set in a logical fashion so that a low-dimensional representation comprehensively summarizes the multivariate data. For example, multidimensional scaling is often used to find a lower-dimensional representation that preserves the interpoint distances of the data. A cluster tree can be used to summarize the natural grouping of points in the multivariate space. Principal component analysis reduces dimensionality by determining the low-dimensional subspace that explains the most variation in the data.

1.2.3 Interaction and motion

Take any of these plotting methods and add interaction and motion linking multiple graphics and we make huge gains in what we can understand and absorb about the variable relationships. The key to multivariate visualization are three modes of interacting:

**Focusing:** is placing attention on a particular aspect of the data by selecting subsets (panning and zooming or slicing) or dimension reduction (projection or variable selection), in particular using simple, easy-to-read plots.

**Linking:** is connecting multiple focused views, in parallel (simultaneous) by brushing or as a sequence over time by using animation and motion.
Arranging: is shifting and regrouping multiple pictures to provide a more informative layout of the information.

These principles are discussed in detail in Buja, Cook & Swayne (1996). In GGobi we concentrate on particularly simple graphics, almost all the plots are based on scatterplots, but we add numerous ways to interact with the plots. This tutorial is an introduction to using the interaction tools.

Note that some similar tools are available in other software packages such as XLispStat (Tierney 1991) (http://stat.umn.edu/~luke/xls/xlsinfo/xlsinfo.html), Data Desk (http://www.datadesk.com/datadesk/), Mondrian (www.theusRus.de), XmdvTool (http://cs.wpi.edu/~matt/research/XmdvTool/index.html), Data Explorer (http://www.almaden.ibm.com) and SAS JMP (http://www.sas.com/). Some of these packages are commercial and some are freeware.

1.2.4 Linked Brushing (and Identification)

Linked brushing is the dynamic changing of symbols (glyphs) or colors in one plot which simultaneously changes corresponding points in other plots (Figure 1.1). The most classic example is brushing in a matrix of pairwise scatterplots. Linked identification is where brushed points are identified, for example, by labels rather than colored.

Figure 1.1: Two plots of the same data linked by brushing: (left) is a barchart of a categorical variable, (right) a scatterplot of two real-valued variables.

1.2.5 Rotations and Tours

While linked brushing provides information on conditional distributions, tours provide information on joint distributions. They are particularly useful for detecting clusters, outliers, distributional shape, including covariance, and some non-linear structure.

Grand Tour

Definition 1 A grand tour is a continuous 1-parameter family of d-dimensional projections of p-dimensional data which is dense in the set of all d-dimensional projections in \( \mathbb{R}^p \). The parameter is usually thought of as time.
This means that each projection shown can be indexed by a time parameter. As time is allowed to wander off to $\infty$ the grand tour will show all possible $d$-dimensional projections of the data, which is the meaning of “dense in the set of all projections”.

A grand tour offers a multitude of aspects simultaneously in relationship to one another. If the data is intrinsically 0-, 1-, or 2-dimensional (that is, clusters, curves or surfaces) the human eye can pick up the “gestalt” almost instantly. (We are adept at detecting and recognizing moving objects.)

Three-dimensional rotation can be considered a special case of the tour, where the dimension of the data is $p = 3$.

The grand tour provides the overview.

**Guided Tour**

To find more specific types of structure intelligent search engines can be connected to the tour, which can automatically provide more informative views than the random ones provided by the grand tour.

The guided tour leads the user to rare views.

**Manual Tour**

Prior knowledge can be incorporated with manually controlled tours. The user can increase or decrease the contribution of a particular variable to a view to examine how a particular variable contributes to any structure. In addition manual tools allow us to assess the sensitivity of the structure to a particular variable or sharpen or refine structure exposed with the grand or guided tour.

The manual tour refines the views.
1.3 Applications

1.3.1 Italian Olive Oils (Classification)

About the data:
This data consists of the percentage composition of 8 fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, eicosanoic, linolenic, eicosenoic) found in the lipid fraction of 572 Italian olive oils. (An analysis of this data is given in Forina, Armanino, Lanteri & Tiscornia (1983). There are 9 collection areas, 4 from southern Italy (North and South Apulia, Calabria, Sicily), two from Sardinia (Inland and Coastal) and 3 from northern Italy (Umbria, East and West Liguria).

Purpose: Did you know that the olive oils from different regions of Italy can be distinguished by their fatty acid composition? The aim of the study on this data is to find combinations of the fatty acids which distinguish the oils from different regions.

Data Setup: Several files were constructed to be passed into GGobi:
- **olive.dat** the data matrix, each row contains measurements on one sample
- **olive.col** the variable labels, one label per row, so there are as many rows as there are variables
- **olive.row** the row labels, one label per row, so there as many rows as there are samples
- **olive.colors** colors for each sample in the data set, and the colors code the region of Italy which produced the sample
- **olive.glyphs** glyphs for each sample in the data set, and the glyphs code the area within the region
- **olive.doc** documents where the data came from, not really used in GGobi
- **olive.xml** Metadata file format.

Action:
Start up GGobi by double-clicking the icon and open the olive oil data.

Exercise 1
a. Brush the points so that colors and glyphs code region and area.

b. The region of southern Italy’s oils can be recognized as different from all other regions by the
presence or absence of one fatty acid. Which fatty acid is it? (Using a 1D plot change between variables until one variable shows a separation. A dotplot is a univariate plot, like a histogram to be read sideways. The horizontal axis is used for randomly jittering points that have the same numerical value. An ASH plot is a smoothed histogram.)

c. Because it is so easy to distinguish the southern oils, remove them for now. Erase these points (using the “Color & Glyph groups” control panel from the “Brush” panel or the “Tools” menu).

d. Now try to work on the separating the northern oils from the sardinian oils. Work back through the dotplots to find the variables which do the best job of separating the two groups. Then look at pairwise plots, and find which pair of variables do well at separating the groups. (Eicosenoic is not useful at any more.)

e. I think you will do best with 3 variables, so using 3-D rotation look for three variables which best separate the groups. Pick the three that you see as doing the best job, and find the projection which best separates the groups. Print this out, so that you can construct a classification rule.

At the end of this exercise you should be able to write down a decision rule to classify oil samples into regions of Italy. If you get a new sample, with the same measured variables you can follow the decision rule to classify the oil as coming from the North, South or Sardinia.

Here is my solution, at this point, for separating the Northern Italian oils from those of Sardinia. First, I looked at the three variables Oleic, Linoleic and Eicosenoic Acid, and found a projection of the three variables where we could discriminate between the two regions with a vertical line (Figure ??). I printed out the coefficients using the “I/O” button:

<table>
<thead>
<tr>
<th>Var</th>
<th>Horizontal</th>
<th>Vertical</th>
</tr>
</thead>
<tbody>
<tr>
<td>oleic</td>
<td>-0.000077</td>
<td>-0.000318</td>
</tr>
<tr>
<td>linoleic</td>
<td>-0.000968</td>
<td>-0.000265</td>
</tr>
<tr>
<td>eicosanoic</td>
<td>-0.005010</td>
<td>0.011945</td>
</tr>
</tbody>
</table>

and used these to generate the plot again in S using the following code:

```r
par(pty="s")
x<-d.olive[d.olive[,1]!=1,6:8]
reg<-d.olive[d.olive[,1]!=1,1]
ds1<-x[,1]*(-0.000077)+x[,2]*(-0.000968)+x[,3]*(-0.005010)
ds2<-x[,1]*(-0.000318)+x[,2]*(-0.000265)+x[,3]*0.011945
plot(ds1, ds2,xlab="Discrim 1", ylab="Discrim 2",pch=".")
points(ds1[reg==2],ds2[reg==2],pch=0)
points(ds1[reg==3],ds2[reg==3],pch=1)
abline(v=-1.68)
```

This gives me a discrimination rule as follows:

If \( \text{oleic} \times (-0.000077) + \text{linoleic} \times (-0.000968) + \text{eicosanoic} \times (-0.005010) > 1.68 \) then the oil comes from region 3 (Northern Italy),

Else the oil comes from East Liguria.

To separate the southern Italian oils, looking at the dotplots just of these regions it appears that palmitoleic and linoleic, stearic and oleic might be useful variables for separating the areas North and South Apulia from the other two areas. Eicosenoic, eicosanoic, and linolenic would be needed to separate Calabria from Sicily, although it is clear that it is not easy to cleanly distinguish this group. Figure 1.4 shows some plots to support these statements.
Figure 1.4: Olive oil data: projections showing combinations of variables where separation of oils from different areas of Southern Italy is visible, from GGobi.
### 1.3.2 Broncopulmonary Displasia (Binary Response)

**About the data:**
This data comes from *Biostatistics Casebook*, by (Miller 1980), and this data is a subset of a much larger data set. Broncopulmonary displasia (BPD) is a deterioration of the lung tissue. Evidence of BPD is given by scars in the lung as seen on a chest X-ray or from direct examination of lung tissue at the time of death. BPD is seen in newborn babies with respiratory distress syndrome (RDS) and oxygen therapy. BPD also occurs in newborns who do not have RDS but who have gotten high levels of oxygen for some other reason. Having BPD or not having BPD is a binary response. Since incidence of BPD may be tied to the administration of oxygen to newborns, exposure to oxygen, $O_2$, could be a significant predictor. Oxygen is administered at different levels: **Low** (21–39%), **Medium** (40–79%), and **High** (80–100%). The number of hours of exposure to the different levels of oxygen is recorded for each newborn in the study. These are used to model the incidence of BPD.

**Purpose:** Ideally we would like to know the number of hours of exposure to the different levels of oxygen that lead to much higher incidence of BPD.

![Figure 1.5: Scatterplot Matrix plot of raw Broncopulmonary Displasia data (left) raw, (right) square root](image)

Figure 1.5 shows the scatterplot matrix of the raw data. The variance for the non-BPD babies is much less than the variance for babies with BPD in number of hours of exposure to BPD. This variance difference needs to be corrected before modeling, hence the plots showing the square root of the oxygen exposure variables. On both scales, we can see that clearly at high doses of oxygen less exposure is critical to the presence or absence of BPD. In fact, if the number of hours at high exposure is less than 130, there is no incidence of BPD. The trend is less obvious for low and medium levels of oxygen, although babies with no incidence of BPD had less than 250 hours of exposure.

There is some advantage to maintaining the raw scale in the examination of this data, because it is easier to interpret the results in the scale of the data. If we use the log scale then we need to invert the measurements back to the raw scale for interpretation.

**Data Setup:** Several files were constructed to be passed into GGobi:
- `bpd.dat` the data matrix, each row contains measurements on one sample
- `bpd.col` the variable labels, one label per row, so there are as many rows as there are variables
- `bpd.xml` metadata format
1.3. APPLICATIONS

Action:
Start up GGobi, and open the bpd.

a. Exercise 2  a. Look at a pairwise plot of BPD vs Low. Brush the points difference colors and symbols according to presence or absence of BPD.

b. Look at univariate plots (textured dot plots) of the exposure variables. Starting from High is easier. Notice that at high exposure, the incidence rate increases dramatically after 150 hours. For medium levels the pattern is similar. For low levels the incidence increases dramatically after 200 hours of exposure.

c. Use the transformation panel to transform the data to the log + 1 scale. Examine the univariate patterns again. The patterns are clearer to see on this scale, but the direct interpretation in hours is trickier.

d. Now look at pairwise plots of the exposure variables. In the Low vs Medium levels, notice the general correlation between the two variables. Also using the brush tool, to transiently brush the (0,0) points, notice that 6 babies had no exposure to low or medium levels of oxygen. Also notice that there are 3 babies who only had exposure to medium levels of oxygen and had no exposure to low level oxygen. These 3 babies were diagnosed to have BPD, and they had been exposed for longer than 220 hours. (Restoring to raw scale interactively enables this interpretation.) Examining Low vs High again shows some correlation. There are 15 babies who were exposed to high levels of oxygen but not low levels of oxygen. Of these the 4 with the longer exposure times shows BPD incidence. (Jitter tool can be used to ensure that overplotting hasn’t obscured difference BPD incidence.) There are two babies (obs 1, 2) that have been exposed to fairly long times for low levels of oxygen but a short time at a high level that don’t have BPD. Examining medium vs High again shows some correlation between the two variables. The same two babies (obs 1, 2) identified before having no incidence of BPD, were also exposed to moderate lengths of time of medium levels of oxygen. There are 6 babies that were exposed to high levels of oxygen but not to medium levels. Of these the baby exposed to the highest level was diagnosed with BPD.

e. Now take a look at the exposure variables touring in 3D. You can see that there is a definite split between the two groups. It is possible to draw a line of separation between them (Figure 1.6). It is probably appropriate to use the hide/exclude tool to only include the observations that correspond to babies having exposure to all levels of oxygen. Rotate the data using the manual tour controls, so that a split is present in the horizontal direction. This projection should correspond (up to a scale factor) to the logistic regression equation seen in chapter 3 for this data. It says to get the lowest incidence of BPD we need to keep the combined exposure to less than 1.5. Now interpreting this back into the raw coordinates is left as a challenge!

1.3.3 Precision Farming, Baker Field Data (Regression Analysis)

About the Data:
Since the time that the Global Positioning System (GPS) became available for public use nearly ten years ago, there has been an increased interest in precision farming. The Global Positioning System allows the agronomist or soil scientist to ascertain her location within a field to a high level of accuracy. This precise measure of location in a field, combined with soil characteristics, yield measurements and other data associated with the location can then be used to extract valuable information about the complex process of plant growth. This information collected at multiple locations within a field can be used to formulate yield-maximization strategies such as variable-rate fertilizer, herbicide or pesticide application.

The data to be analyzed in this study was drawn from part of a privately owned farm in southeastern Boone County, Iowa (Colvin, Jaynes, Karlen, Laird & Ambuel 1997). A total of 224 sites within an approximately 350 m x 350 m portion of the field were studied. The sites are located along 8 equally-spaced east-west transects. Along each transect there are 28 sites, spaced approximately 12.2 m apart.
Figure 1.6: Broncopulmonary Displasia data: selected tour views revealing separations between patients with and without the disease.
At harvest time, a combine is driven down each transect, stopping every 12.2 m to measure yield in bushels/acre.

The field of interest has been on a corn-soybean rotation since 1957. The 1997 data set to be discussed contains corn yields and measurements of ten soil nutrients in parts per million, including Boron (B), Calcium (Ca), Copper (Cu), Iron (Fe), Potassium (K), Magnesium (Mg), Manganese (Mn), Sodium (Na), Phosphorus (P), and Zinc (Zn). The corn was harvested on October 6, 1997 and the soil samples were taken on May 22-23, 1997. The soil samples consisted of 6 cores, 1 inch in diameter and collected to a depth of 8 inches. Samples were air dried, ground, extracted, and analyzed for the various nutrients using inductively-coupled plasma (ICP) techniques (Karlen and Colvin, 1998). Because soil nutrient data were not available at 9 of the sites, only 215 of the sites are included in the analyses to follow.

**Purpose:**
Explore the relationship between yield and soil characteristics to make recommendations about improving the yield.

**Action:**
Start ggobi up on the data.

**Graphical Analysis: Simple Plots Relating Yield to Soil**

The first step in our analysis was to look at the relationship or correlation between each of the soil nutrients and corn yield.

GGobi allows the user to identify points in a plot by selecting the “Identify” option and then clicking on the points of interest. This is a valuable tool for identifying outliers. Point number 200 was identified as an outlier for this data set because it lies far away from the bulk of the points in several of the scatterplots that were examined.

Also, GGobi allows the user to make transformations on the fly. In the data we noticed many non-linear patterns, so we explored the use of transformations to linearize the variable relationships. This makes future analysis easier. All correlation coefficients discussed in this section are calculated in the absence of point number 200. When examining each soil nutrient, a transformation of the nutrient was often chosen in order to linearize its relationship with corn yield. A discussion of each nutrient’s relationship with yield follows.

**Boron.** The correlation after removing the outlier (point number 200) is 0.137. There did not appear to be a transformation which could linearize this relationship, but it is interesting to note that higher amounts of Boron lead to consistently higher yields.

**Calcium.** Note how the outliers (point numbers 81 and 82) become part of the main body of the data after applying the inverse transformation. Additionally, the relationship between 1/Calcium and yield is more linear than the relationship between Calcium and yield. This graphical observation is supported by the fact that the correlation improves from 0.329 to -0.457 when the Calcium values are inverted. (Without the points numbered 81 and 82, the correlation between the untransformed Calcium values and yield was 0.387.)

**Copper.** Taking the log of Copper makes its relationship with corn yield more linear, as illustrated by the increase in correlation from 0.625 to 0.654.

**Iron.** There is a highly nonlinear relationship between Iron and corn yield with a linear correlation coefficient of .134. This nonlinearity is accentuated some by taking the log of Iron.

However, there exists an interesting relationship between Iron, Calcium and yield which can be visualized using a technique in GGobi called “brushing.” One can open two separate (but “linked”) GGobi windows and brush or highlight points in one window while viewing the associated points in the second window. Brush high and low values of Calcium (values in excess of 5200 ppm) different colors.

Hence, it appears that the “7”-shape of this plot is due to two separate phenomena. That is, the relationship of iron with yield is different for different values of Calcium. When Calcium is high (with values in excess of 5200 ppm), an increase in Iron results in little practical increase in yield. The correlation between log(Iron) and yield for the high-Calcium group is 0.309. However, when Calcium
values are low (with values in less than 5200 ppm), an increase in Iron results in a substantial increase
in yield. The correlation between log(Iron) and yield for the low-Calcium group is 0.535.

**Potassium.** The plot illustrates a nonlinear relationship between Potassium and yield. Inverting the
Potassium values results in a fairly strong linear relationship between 1/Potassium and yield. The
transformation improves the correlation coefficient from a value of 0.467 to -0.544.

**Magnesium.** The plot of Magnesium versus yield is similar to the relationship between Potassium
and yield. Applying the same transformation (1/x) yields a linear relationship between 1/Magnesium
and yield. The transformation improves the correlation coefficient from a value of 0.491 to -0.528.

**Manganese.** The relationship between the log of Manganese and yield is slightly more linear than
the relationship between Manganese and yield. The transformation increases the correlation coefficient
slightly, from a value of 0.456 to a value of 0.475.

**Sodium.** Note that the discrete Sodium measurements did not have a strong relationship with yield.
Neither the plot nor the low correlation coefficient of .183 could be improved by a transformation, but
it is interesting to note that higher amounts of Sodium lead to consistently higher yields.

**Phosphorus.** The nonlinear shape of the plot of Phosphorus versus yield was changed to a more linear
shape by taking the log of the Phosphorus values. This transform increases the correlation coefficient
from 0.490 to 0.537.

**Zinc.** Because the shape of the plot of Zinc versus yield is similar to the shape of the plots of Copper,
Manganese and Phosphorus, a natural tendency would be to use the same (log) transform. But since
some of the Zinc values are equal to zero, the transform log(Zinc + 1) was used. This increases the value
of the correlation coefficient from 0.615 to 0.645.

**Relating Soil Variables to Each Other**

A convenient way to view the pairwise relationships between variables is using a matrix of scatterplots
of all pairs of variables. The pairwise plots are laid out in “matrix” format. For example, all the
plots across the top row have Boron plotted vertically and all other variables are sequentially plotted
horizontally. Down the first column, Boron is plotted horizontally, with the other variables plotted
vertically. The second row shows Calcium plotted vertically and the other variables plotted horizontally.

In the plots of the raw variables it can be seen that almost all of the pairs of variables have a linear
positive relationship. That is, the more of one nutrient the more of the other nutrient. There are a few
noticeable exceptions: Calcium and Iron, and Copper and Iron. The plot of Copper vs Iron indicates a
non-linear relationship, and the plot of Calcium vs Iron indicates a negative relationship.

The plots of the transformed variables are much clearer to read, but intuitively harder to interpret.
For these plots it is clear that some of the variables are very strongly related to each other: Calcium,
Copper, Magnesium, Manganese and Zinc, and, Potassium and Phosphorus, and Boron and Sodium.
Note also, that the non-linear relationship between Copper and Iron is still visible, and the relationship
between Calcium and Copper now looks somewhat non-linear. There are also a few noticeably unusual
points in several of the plots.

**A Grand Tour of the Soil Variables**

The Grand Tour is a data analysis tool that can be helpful for identifying outliers, clusters and general
structure in a data set. The GGobi software package describes the Grand Tour as a graphical method
which “successively samples planes in p-space, where p is the number of variables presently selected,
and connects the planes by moving along a geodesic interpolation path between them....By allowing the
grand tour to run uninterrupted the viewer can get a global view of the linear combinations among the
variables” (GGobi online help, see also Swayne, Lang, Buja & Cook (2002)).

As the Grand Tour progresses dynamically, one can notice observations that do not “move” with the
rest of the data through the projection space. These observations might be considered to be outliers. In
viewing this Grand Tour, one can readily notice that the points labeled 81 and 82 should be considered outliers when considering all ten untransformed soil variables jointly. Points 81 and 82 have the two highest values for Calcium. There are also a few more less extreme outliers but these also look to be mostly in the measurement of Calcium. Doing analysis on the inverse of Calcium makes the outlying nature of these observations less pronounced.

A cluster in high-dimensional data is a group of points that has “similar” values for all of the variables. We generally think of clustered data as having two or more such clusters, and generally there is a distinct separation between clusters. For detecting clusters we would look for groups of points which “move” together, but differently from points in other clusters in the Grand Tour. With this soil data there are no obvious clusters. Rather the untransformed data takes an interesting non-linear form, for example, there is a distinct “C” shape visible. What this means is that there are strong non-linear dependencies amongst the nutrients in the soil, and there are regions where the presence of some nutrients will be accompanied by the absence of other nutrients. Now a careful inspection of the axes in the lower left of the first plot indicates that almost all of the variables contribute to this view: it is a mess of variable axes, meaning that B, K, Ca, Mg, Zn, Cu, Mn and Fe all are important contributors to this view.

With this many variables it is usually not easy to make a simple interpretation of the non-linear shape, other than there is a strong non-linear dependency between the values of nutrients in the soil. But, there is some similarity to the relationship we saw in the last section between Copper and Iron, and these two variables have large contribution to this view. So it is probably safe to assume that the non-linear relationship that we see here is mostly due to the relationship between Copper and Iron, and also to some extent Calcium with Iron. If we remove all the variables from the view except for these three we can see that the non-linear structure remains.

Now an interesting aside to this structure is that it relates to the yield in the field. At the top end of the “C” the yield is almost exclusively low (red points), and as we move around the “C” to the bottom the yield gets progressively better (green). What does this mean? Clearly, the yield is different according to the combination of nutrients available. To understand this we would need to better characterize the constituents of the soil at both ends. A closer study of this phenomenon in the numerical studies helps to unravel the mystery a little.

**Linking Yield to Spatial Location**

Using linked brushing we can display the locations of the high, medium and low yields in the field. There appears to be very strong spatial contiguity between sites with the different yield classifications. Note that these sites also have fairly well-defined differences in the nutrient make-up: the high yield sites had more Calcium, Copper, Potassium, Magnesium, Manganese, Phosphorus and Zinc on average.

**Conclusions from Graphical Analysis**

From the graphical analysis of the soil nutrient data, it appears that all the nutrients have a positive relationship with yield, with the exception of Iron. Seven of the ten (transformed) nutrients have a strong linear relationship with yield. Ordered from strongest to weakest, the seven transformed variables are log(Copper), log(Zinc+1), 1/Potassium, log(Phosphorus), 1/Magnesium, log(Manganese), and 1/Calcium. That is, higher yield was obtained with higher Copper, higher Zinc, higher Potassium, higher Phosphorus, higher Magnesium, higher Manganese, and higher Calcium.

With Boron and Sodium, the yield is uniformly higher when the content is higher; with lower content yield ranges across the full spectrum. There is one exception to this: one site had very low yield despite the relatively high quantities of Boron and Sodium (and also Phosphorus). There is clearly something unusual happening at this site.

Now there are some interesting relationships between the nutrients when Iron is involved. Yield is good regardless of Iron content if Calcium content is high; however, if Calcium is low Yield improves with higher Iron! Similarly this holds for Copper and Iron also.

**1.4 Exercises**

**1.4.1 Flea beetles**

This data is from Lubischew (1962).
There are 3 species of flea-beetles: Ch. concinna, Ch. heptapotamica, and Ch. heikertingeri, and 6 measurements on each. 1 = width of the first joint of the first tarsus in microns (the sum of measurements for both tarsi). 2 = the same for the second joint. 3 = the maximal width of the head between the external edges of the eyes in 0.01 mm. 4 = the maximal width of the aedeagus in the fore-part in microns. 5 = the front angle of the aedeagus (1 unit = 7.5 degrees). 6 = the aedeagus width from the side in microns.

**Exercise 3**

a. Describe how to set up colors (and glyphs) in order to find classification rules for identifying the 3 species based on the 6 measured variables.

b. Right down a strategy for visually constructing the classification rules.

c. Follow your scheme and write down your final classification rules.

### 1.4.2 Australian Crabs

This data consists of 2 species of crabs, orange and blue, and males and females in each species, and 5 measured variable on each crab.

**Exercise 4**

a. Describe how to set up colors (and glyphs) in order to find classification rules for identifying the 2 species and which sex based on the 5 measured variables.

b. Right down a strategy for visually constructing the classification rules.

c. Follow your scheme and write down your final classification rules.

### 1.4.3 Evaporation data

This data comes from the S blue book, and consists of measurements made on 46 consecutive days, between June 6 and July 21, no year given. Measured variables are:

- avat = average soil temperature (deg F)
- avat = average air temperature (deg F)
- avh = average relative humidity
- wind = total wind (miles per day)
- evap = daily amount of evaporation from the soil

The reason for collecting this data is described as:

It is desired to estimate the daily amount of evaporation from the soil as a function of air temperature, relative humidity and wind.

**Exercise 5**

a. Explore the relationship between evaporation and the 4 independent variables.

b. Do you see a need for transformations?

c. Explore the independent variable space for clusters of similar days, and abnormal weather days.
Bibliography


