Object Oriented Data Analysis

J. S. Marron & Ian L. Dryden

August 18, 2017

Department of Statistics and Operations Research, University of North
Carolina, Chapel Hill, NC 27599-3260, USA

School of Mathematical Sciences, University of Nottingham, Nottingham, NG7
2RD, UK

Contents

1 What is OODA? 4
1.1 Curves as Data Objects 6
1.2 Amplitude and Phase Data Objects 13
1.3 Shapes as Data Objects 17
1.4 Tree Structured Data Objects 24
1.5 Sounds as Data Objects 27
1.6 Images as Data Objects 30

2 Overview of OODA 33
2.1 Data Object Selection 33
2.2 Data Visualization 40
2.2.1 Visualization of Marginals 40
2.2.2 Visualization of Global Structure 44
2.3 Confirmatory Analysis 63
2.4 Further Major Statistical Tasks 67
2.5 OODA Software 69

3 OODA Background and Related Areas 70
3.1 History and Terminology 70
3.2 Compositional Data Analysis 71
3.3 Symbolic Data Analysis 72
3.4 Other Research Areas 73

4 OODA Preprocessing 75
4.1 Marginal Distributions 75
4.2 Standardization 88
4.3 Transformation 90
4.4 Registration 93

5 Distance Based Methods 95
5.1 Fréchet mean 96
5.2 Multi-dimensional scaling 101

6 Directions for Visualization 105

7 Manifold Data Analysis 106
7.1 Shapes as Data Objects 106
7.2 Covariance Matrices as Data Objects 106
7.3 Material from Old Chapter 1 107

8 FDA Curve Registration 111

9 Tree Structured Data Objects 112
9.1 Phylogenetic Trees 112

10 Classification 113
10.1 Classical Methods 113
10.2 Kernel Methods 113
10.3 Support Vector Machines 113
10.4 Distance Weighted Discrimination 113

11 Clustering 115

12 Confirmatory Analysis 116
12.1 DiProPerm 116
12.2 SigClust 117

13 High Dimension Low Sample Size Analysis 118

14 Smoothing 119

15 Robust Methods 120

16 PCA Background and Details 121

17 Multi-Block Methods 122

Chapter 1

What is OODA?

The fields of human endeavor currently known as *statistics*, *data science* and *data analytic* have been radically transformed over the recent past. These transformations have been driven simultaneously by a massive increase in computational capabilities coupled with a rapidly growing scientific appetite for ever deeper understanding and insights. The notion of *data matrix*, with perhaps columns used for cases, and rows for measurements (i.e. *features*) provides a useful paradigm for understanding important aspects of how these fields are evolving. In particular, the currently popular context of *Big Data* is easily seen to have several quite different facets, ranging from *low dimension high sample size* areas (the basis of classical mathematical statistical thought, which is perhaps typified by census data), through both high dimension and sample sizes (common for internet scale data sets of many types), and on to *high dimension low sample size* contexts (frequently encountered in areas such as genetics and other types of extremely rich but relatively expensive measurements). The pressing need to analyze data in this wide array of contexts has generated many exciting new ideas and approaches.

Yet a deeper look into these developments suggests that the organization of data into a matrix may itself be imposing limitations. In particular, there is a growing realization that the challenges presented by Big Data are being eclipsed by the perhaps far greater challenges of *Complex Data*, which are typically not easily represented as an unconstrained matrix of numbers. *Object Oriented Data Analysis* (OODA) provides a useful general framework for the consideration of many types of Complex Data. It is deliberately intended to be particularly useful in the analysis of data in complicated situations, diverse examples of which are given in later sections of this chapter. The phrase OODA in this context was coined by Wang and Marron [223]. An overview of the area was given in Marron and Alonso [145].

The OODA viewpoint is easily understood through taking *data objects* to be the *atoms* of a statistical analysis, where atom is meant in the sense of elementary particle, studied in several contexts of increasing complexity:

- In a first course in statistics atoms are numbers, and the goal is to develop methods for understanding of variation in populations of numbers.
- A more advanced course, termed *multivariate analysis* in the statistical culture, generalizes the atoms, i.e. the data objects from numbers to vectors and involves a host of methods for managing uncertainty in that context.
- A currently very fashionable area in statistics is *functional data analysis* (FDA), where the goal is to analyze the variation in a population of curves. A good introduction to this vibrant research area, where functions are the data objects, can be found in Ramsay and Silver an [177], [178]. An example, illustrating many of the basic concepts of FDA, which are useful for understanding OODA is given in Section 1.1.
- OODA provides the next step in terms of complexity of atoms of a statistical analysis to a wide array of more complicated objects. Several of these are illustrated using real data examples in Sections 1.3 to 1.6.

A good question is: What is the value added to applied statistics and data science from the concept of OODA and its attendant terminology? The terminology is based on very substantial real world experience with a wide variety of complex data sets. A fact that rapidly becomes clear in the course of interdisciplinary research is that there frequently are substantial hurdles in terminology. Especially at the beginning of such endeavors, it can feel like collaborators are even speaking different languages, so often serious effort needs to be devoted to the development of a common set of definitions just to carry on a useful discussion. An added complication is that for complex data contexts, it is frequently not obvious how to even "get a handle on the data". Usually there are many options available, which are most effectively decided upon through careful discussion between domain scientists and statisticians. In such discussions, the issue of *what should be the data objects?* has proven to frequently lead to useful choices, thus resulting in an effective and insightful data analysis.

Real data examples, demonstrating data objects choices in a variety of real data contexts are given in the following sections. In particular, Section 1.3 shows an example where *shapes* are the data objects, which require special treatment as shapes are most naturally viewed as points on a curved manifold. Section 1.4 considers a perhaps even more challenging data set of tree structured data objects, where an overview of various choices that have been made is given. The data objects in Section 1.5 are recordings of sounds, in particular human spoken words, which bring special challenges in the choice of data objects. A deep variation of FDA involves curves with interesting variation in phase in place of, or in addition to the usual amplitude variation, discussed in Section 1.2. It is seen that the notion of data objects provides a particularly useful format for discussing the modes of variation. Finally, in Section 1.6, a fun example with images of faces as data objects in considered.

One more general feature of OODA is that there are frequently three major phases of this type of data analysis:

1. Object definition. This is the phase where the fundamental issue of what should be the data objects is addressed. A number of examples of this provided in the rest of this chapter and also in further examples in other sections.

2. Exploratory Analysis. Here the goal is to find perhaps surprising structure in data, often using some type of visualization method. A wide variety of examples and methods for exploratory analysis are given in the rest of this Chapter and in Chapters 4, 5, 6, 7, 9 and 9. While exploratory analysis frequently only appears sparingly in most classical statistics courses, it is usually more prominent in machine learning. However it has a strong statistical tradition, going back well before the ideas nicely summarize in Tukey [217].

3. Confirmatory Analysis. While many great discoveries have been made using exploratory methods, it is also very easy to make discoveries that are not real, in the sense of being non-reproducible sampling artifacts. For this reason it is very important to validate such discoveries. This critical topic and many variations of approaches to it is discussed in detail in the very large classical statistical literature. Some less well known aspects, that are particularly relevant to OODA are discussed in Chapter 12.

A companion website to this book, containing links to available software, the Matlab programs used to generate the Figures in this book, and additional graphics can be found at Marron [146].

Further discussion on other ideas and nomenclature related to OODA can be found in Chapter 3.

1.1 Curves as Data Objects

An interesting example of functional data analysis (viewed here as an important special case of OODA) is the Spanish Mortality Data, first studied from an FDA viewpoint in Section 2 of Marron and Alonso [145]. Such data sets are available at the Human Mortality Database of Wilmoth and Shkolnikov [231]. For a given population (e.g. citizens of one country) mortality data are generally a matrix with rows and columns indexed by years and ages. The matrix entries are the chance of a person of each age dying in the given year, calculated as the number of deaths during that year - age pair, divided by the number of people. Here we study mortality of males in Spain, mostly because there are interesting features in the data, due to Spain's recent history.

There are several data object choices to be made in the analysis of this data. First, since these probabilities range over several orders of magnitude, logarithms are useful to provide good visual separation across a wide range of scales. Particularly strong interpretability comes from the choice of $\log_{10}$ of the probability. The utility of this data object choice is demonstrated in Figure 1.1, where the raw probabilities are shown in the left panel (with much interesting structure missed since this is very nearly 0 for the important younger age groups) with $\log_{10}$ mortality in the right (highlighting important contrasts among the younger ages). Second there are two different ways to turn the matrix of data into functional data. One is to consider data objects to be curves of mortality as a function of age, with curves indexed by year. The other is (the matrix transpose) where the mortality is viewed as a function of year, as data object curves indexed by age. In this analysis, the former choice is used, because it gives the best illustration of the usefulness of OODA concepts and also gives an interesting narrative. The latter choice is explored in Chapter 17, together with an analysis that also integrates female mortality in an interesting way. This results in $n = 95$ curves corresponding to the years 1908-2002. Ages considered here are 0 through 98, since larger ages are problematic due to occasional small population sizes. The raw data are shown as overlaid curves in Figure 1.1. There the curves are distinguished using the standard graphical technique of a rotating color palate (in this case the default 7 colors in Matlab).

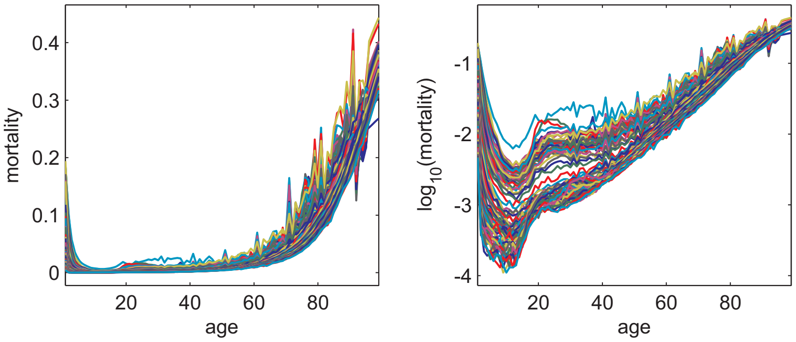

Figure 1.1: Spanish Male mortality curves as a function of age. Raw mortality is in the left panel, with $\log_{10}$ mortality on the right. Years are distinguished using a rotating color palette. Shows age effects and large variation (factors of more than 10 for some age groups) across years.

This view already shows interesting features of the data. For example, being born is a risky activity, with a high mortality rate. However, the chance of dying falls off rapidly, up until the teen years when risky behavior tends to begin. Then through adulthood the death rate slowly increases, becoming quite high in old age. Also note the bundle of curves is quite thick, with the axes indicating approximately a 10 fold change over the years, begging an investigation into how things have changed over time. This is easily provided in Figure 1.2, by applying a different color scheme to the same curves. Here time ordering of the curves is highlighted through coloring with a rainbow scheme to indicate years, starting with magenta for 1908 and ranging through blue, cyan, green, yellow and orange to red for 2002.

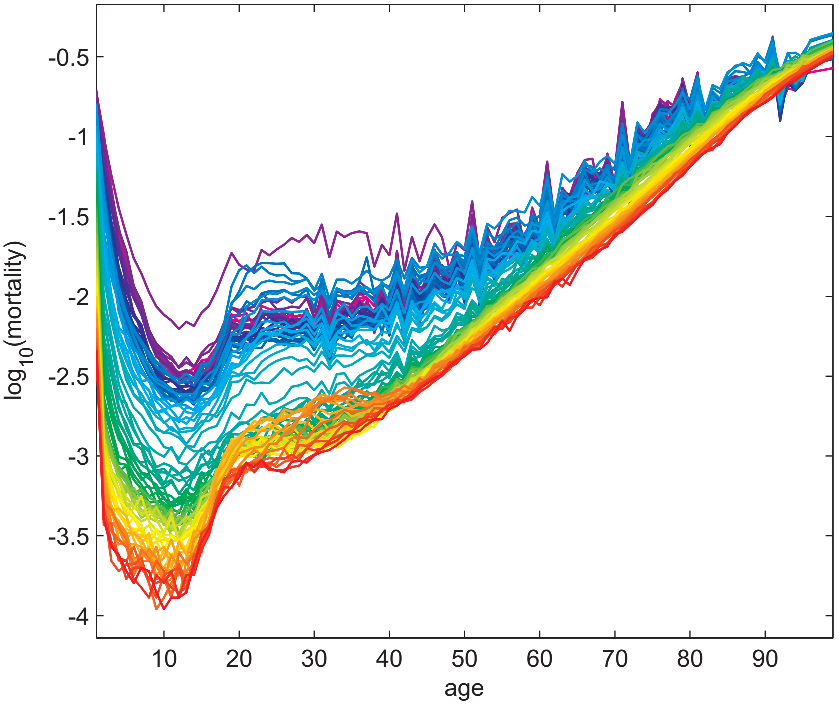

Figure 1.2: Same mortality curves using a rainbow color scheme to indicate progression in time (over years 1908-2002). Shows major improvements in mortality over this time range.

This show a very clear overall improvement over the years in mortality, due mostly to improvements in medicine and public health. Note also that these improvements have benefited younger people more than the old, as there is not yet much treatment available for aging. As happens frequently with OODA data, additional visual insights come from careful decomposition of the variation present in these curves, through a *Principal Component Analysis* (PCA). See Chapters 2 and 16 and Jolliffe [115] for background information concerning the many ways this method is used. One important use of PCA is to gain insight into how data objects relate to each other. Insight comes from considering the data as lying in an abstract *point cloud* in $d = 98$ dimensional space, where low dimensional projections frequently visually illustrate key relationships (e.g. clustering of data objects). An often useful first step of a PCA is *mean centering*, which essentially moves the point cloud so that it is centered at the origin. As seen in Figure 1.3, this centering operation itself can provide an informative decomposition of the data into the mean and residuals about the mean.

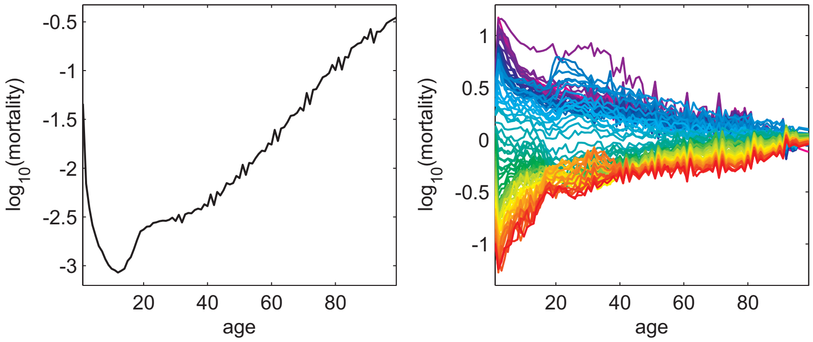

Figure 1.3: Left panel is the mean mortality curve. Right panel contains the mean residuals, where the mean is subtracted from each curve, using the same color scheme. Shows that age effects are essentially common for all, in the sense of appearing in the mean. Improvements over time appear in the residuals, with overall most improvement for the young.

The left panel of Figure 1.3 shows the mean curve, computed as the pointwise mean of the curves in Figure 1.2. The right panel contains the mean residuals, which are computed by subtracting the mean from each of the data curves, while retaining the original year coloring. Note that the mean curve contains many of the important features of the raw data, especially those related to age. In particular, the danger of being born together with low mortality for the young with increasingly higher mortality for the old, are all properties of the mean. These essentially do not appear in the mean residuals, indicating that these are population properties which have not changed much over time. A perhaps surprising feature of the mean is the occasional blips that appear. One might think these are random noise, but note that they are quite periodic and in fact appear at decades. This is a function of historically poor record keeping. The early lack of birth certificates for the full population led to some uncertainty of age at the time of death for some, with subsequent rounding to decades which is clearly visible. The mean residuals also reflect an important aspect of the population structure, being driven by the changes over time. Most important are the dramatic improvements in mortality that have been made over the course of this study. This view also makes it clear that the young have benefited the most with that benefit decreasing as a function of age.

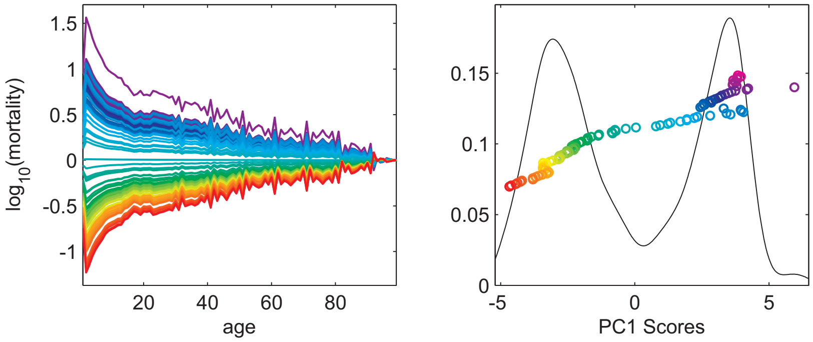

Figure 1.4: PC1 loadings plot (left) and scores distribution plot (right). Loadings plot shows that the dominant mode of variation reflects most of the overall improvement in mortality. Scores plot shows most of the improvements happened relatively rapidly, plus highlights the 1918 Flu Pandemic (magenta outlier on the right) and the Spanish Civil War (light blue sharp trend to the right).

The variation revealed by the first principal component is shown in Figure 1.4. A useful understanding of PCA comes from thinking of the above mentioned point cloud, where each data object (curve in this case) is a point. PCA seeks orthogonal directions of maximal variation within the point cloud. The first PC direction is the unit (i.e. length 1) vector, based at the sample mean, which maximizes the variance of the data projected onto that vector. This direction is easily computed as the first eigenvector of the sample covariance matrix and the entries of that vector (which indicate how it relates to the variables i.e. features, of the data set) are called the *loadings*. Visual insight into these loadings comes from the *loadings plot* in the right panel of Figure 1.4. The horizontal axis indexes the features, which are ages in this case, and the curves are all multiples of the eigenvector. This view highlights the dominant mode of variation, which is seen to be the major overall improvement in mortality. In addition, as life and death record keeping has improved over time, the decline in age rounding effects is reflected in the decade spikes pointing upwards (early) and downwards (later). In particular, the rounding was present earlier, not later, so it shows up partially in the mean in Figure 1.3, and then as this contrast in Figure 1.4 (left panel). The right panel of Figure 1.4 is the PC1 *scores plot*. It shows the one dimensional projection coefficients of the data curves onto the first eigen direction. These are just the coefficients (same color scheme) of the curves in the left panel. Scores are shown as the horizontal coordinates of the circles, with the vertical coordinate (as well as the color) indicating order in the data set, in this case the year. The overall leftward trend again shows the overall improvement in mortality over these years. The black curve shows a *kernel density estimate*, which can be thought of as a smooth histogram. Some discussion of kernel density estimation is in Chapter 14. See Wand and Jones [222] for a more in depth discussion. This shows much higher density of scores in the higher and lower regions, which is another way of seeing that most of the overall transition from higher to lower mortality was relatively rapid. A couple of smaller scale features are also clear in this score plot. The purple year, farthest to the right was the year 1918, when many people around the world died during a flu pandemic, which was the largest ever documented epidemiological event. Also notable is the shift towards higher mortality (i.e. to the right) shown as light blue, which was the time of the Spanish Civil War, just before World War II (in which Spain was not a combatant).

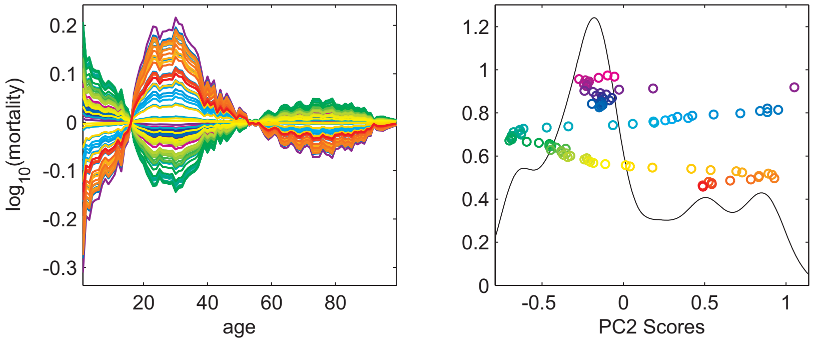

Figure 1.5: PC2 Loadings (left) and Scores (right). The loadings plot shows this provides a contrast between the 20-45 year olds with the rest. The scores plot shows the deep effects of the flu pandemic, the Spanish civil war and automotive death rate.

Figure 1.4 reveals the first mode of variation in the mortality data called PC1. An interesting complementary mode of variation is the second PC, as shown in Figure 1.5. This represents the direction of second strongest variation (in the sense of being orthogonal to the first direction) measured again in terms of variance of projections. It is computed as the second eigen direction of the sample covariance matrix. The PC2 loadings plot (left panel) shows that this direction highlights differences between the 20-45 year old cohort, with the union of the young and the old. The color pattern is harder to interpret in the loadings plot, but is quite direct in the scores plot (right panel). Note that the 20-45 year olds suffered even stronger effects from both the pandemic and also the war, as they died at a substantially higher rate than usual in those times. Another interesting feature is the growing mortality for this cohort in the 1960s to 1980s (green to orange). This period corresponds to growing access to automobiles, and apparently the idea that young males are the group most prone to risky automobile behavior. Note that in the final years, the direction of this trend has fortunately reversed, which has been ascribed to much improved car safety features and also to major improvements in roads.

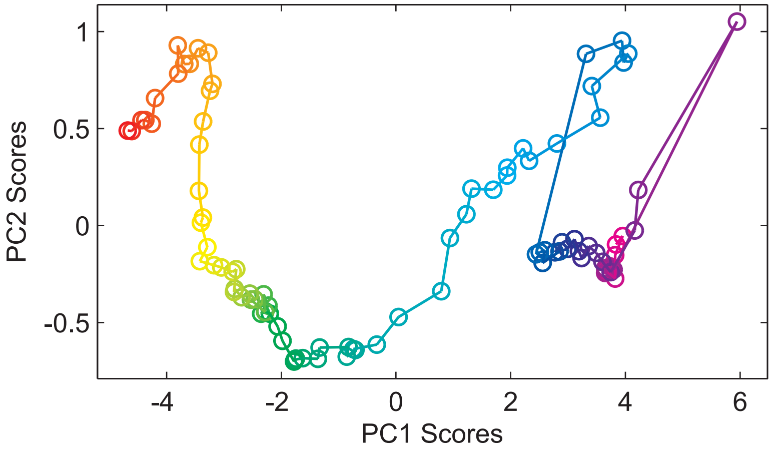

Figure 1.6: Scatterplot of PC1 vs. PC2 scores. This makes many of the above lessons available in a single plot.

Figure 1.6 shows the bivariate distribution of the PC1 and PC2 scores, which provides a useful and concise summary of much of the structure in this data set. The one dimensional PC1 scores distribution in the right panel of Figure 1.4 is on the horizontal axis, while the vertical axis has the corresponding PC2 scores distribution from the right of Figure 1.5. This is the two dimensional projection of the data with maximal variation. Note that the circles representing the data objects (i.e. the mortality curves) are connected with line segments in time order, which facilitates keeping the progression of years in mind when interpreting the plot. The overall improvement in mortality, with the exceptions of flu and war, are clear from the main leftwards progression. Variation over time of the contrast between the 20-45 year old and the rest are also clear on the vertical axis, nicely highlighting the flu, war and automobile effects.

For this data set, the most interesting views are in the first two PC components. For others, more components can also be quite insightful. A standard approach is to create a matrix of such scatterplots, with the axes carefully coordinated over both rows and columns. The diagonal of such a display is most useful when it shows some sort of 1-d distributional summary, e.g. the type used in the left panels of Figures 1.4 and 1.5.

Mortality rates for other countries can be explored in a similar way. For example mortality data from Switzerland (also available in Wilmoth and Shkolnikov [231]) show similar flu pandemic and automobile effects as observed here, but neither the data rounding (due to a longer period of good record keeping) nor the war caused mortality effects are visible as expected.

1.2 Amplitude and Phase Data Objects

The OODA way of thinking has also proven to be especially useful in another area of FDA, as discussed in the survey paper [141]. As noted in Marron et al [140], that part of FDA is sometimes called *curve registration*, because it is very useful in situations where the curve data objects are clearly misaligned. An interesting example of this, from Koch et al [121] and [139], is shown in Figure 1.7. The data objects here are proteomics mass spectrometry profiles from Ho [101], a larger study of bio-markers in Acute Myeloid Leukemia. A detailed description of this data set including a number of pre-processing steps (including median smoothing and interpolation to an equally spaced grid) can be found in Koch et al [121]. Essentially there are 5 patients, represented as colors, with 3 replicate curves for each patient, thus 15 curves in all, shown in the top part of the top panel. Each curve shows Total Ion Counts (TIC). for each mass to charge ratio (horizontal coordinate). The TIC curves have many peaks, which correspond to various peptides. A common goal of mass spectrometry analyses is *curve registration*, i.e. finding deformations, sometimes called *warpings*, of the horizontal axis to properly align the peaks so that they chemically correspond. In most contexts it is hard to quantitatively assess the performance of a given registration, but this data set is special because the locations of several of the actual peptide peaks have been (laboriously) found for each curve using additional information as detailed in Koch et al [121]. These peak locations, for each of the 15 curves, are indicated by peak numbers (1-14), with colors corresponding to the curves. The peak numbers are sorted vertically by height of the corresponding peak and connected with gray line segments to give some visual correspondence. It is hard to see much pattern, showing this to be a challenging curve registration problem.

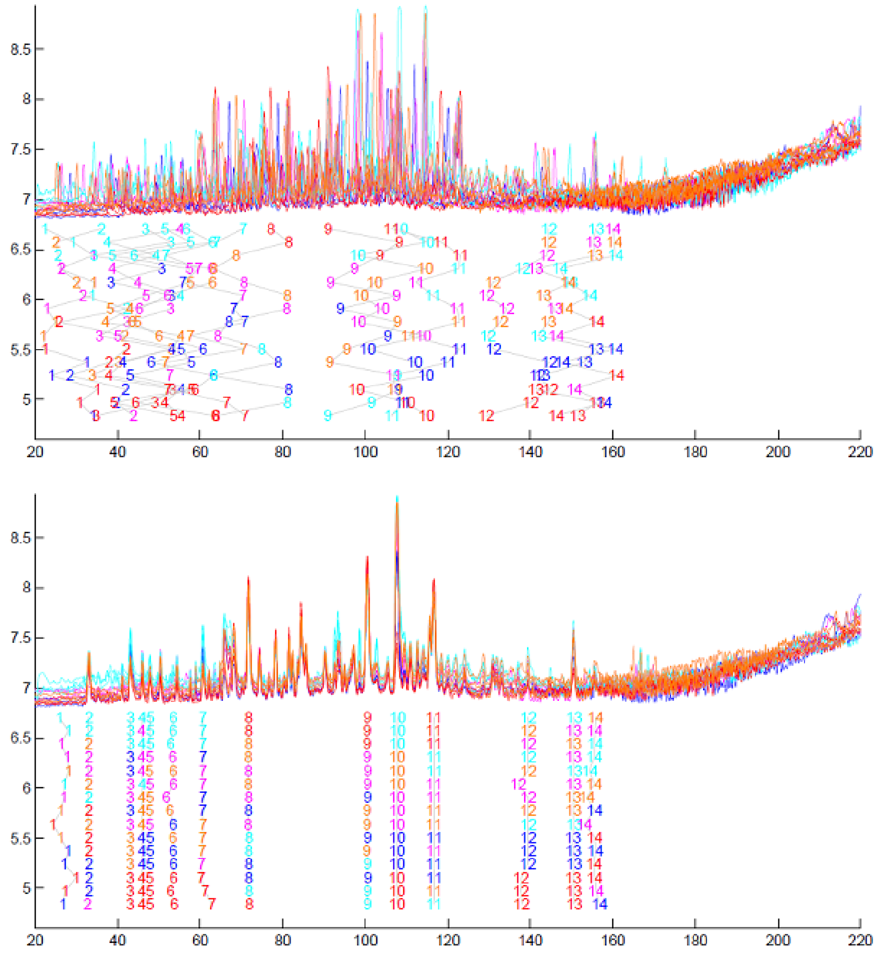

Figure 1.7: Top panel contains raw TIC curves (top), with a labeling of certain important peaks in the lower part of the panel. Bottom panel shows a Fisher-Rao registration of the TIC curves. Numbers under the curves indicate peak locations, showing that the registration has been mostly quite effective.

There are a number of approaches to this type of data challenge, with several such analyses of this data set discussed in Koch et al [139]. The bottom panel of Figure 1.7 shows the results of registration of these same TIC curves using the Fisher Rao method proposed in Srivastava et al. [200] and Kurtek et al [124], using only the curves themselves and not the peak location information. The colored numbers reveal that this is a particularly challenging problem, because the peaks have quite different heights across patients. Peak 10 is particularly challenging as it is quite low for the red patient (especially compared to nearby very tall peaks), yet is the highest peak for other patients. Note the alignment is not perfect for every numbered peak, but it is still of impressively high quality.

Since such time warpings, of the horizontal axis provide an appealing approach to registration as shown in Figure 1.7, many methods have been developed for this. An overview of these have been discussed in [141]. An important point of that paper is that this same mathematical approach is useful more generally than to simply align curves. While in some contexts, such as that of Figure 1.7, the phase component is merely *nuisance variation* to be dealt with but of no intrinsic interest, there are many situations where the warps themselves represent useful modes of variation. In such contexts it is insightful to think of *amplitude data objects*, whose variation is contained in the aligned curves, and *phase data objects* which are the warps used to achieve the alignment. Depending on the context either or both choices of data object can be of primary interest, or either could represent just nuisance variation.

The notions of amplitude and phase data objects are illustrated in the simulated example shown in Figure 1.8. The upper left panel shows a simulated functional data set, where every data object (curve) has two peaks and is a multiple of a beta mixture probability density. A rainbow color scheme is used to distinguish the curves, in order of how separated the peaks are. The peaks have both different heights showing substantial amplitude variation, and also quite different locations reflecting strong phase variation. These modes of variation are decomposed in a useful way by the warping functions shown in the bottom left panel, computed using the Fisher Rao method of Srivastava et al. [200]. The vertical axis is the same as in the upper left panel. Rescaling that axis using the purple warp functions moves the purple peaks inwards, and using the red warp functions moves the red peaks outwards. The top right panel shows the amplitude data objects, i.e. aligned curves. A careful look shows that the random peak heights are linearly related with the left peak being high when the right peak is low. This set of data objects has just a one dimensional mode of variation. The warps in the lower right panel can be thought of as the phase data objects, although they are not easy to interpret. Enhanced interpretation of the variation in the phase data objects comes from the view in the lower left panel. That is an application of each of the warps to the Kärcher mean template from the Fisher-Rao calculation, which nicely reflects the one dimensional phase variation.

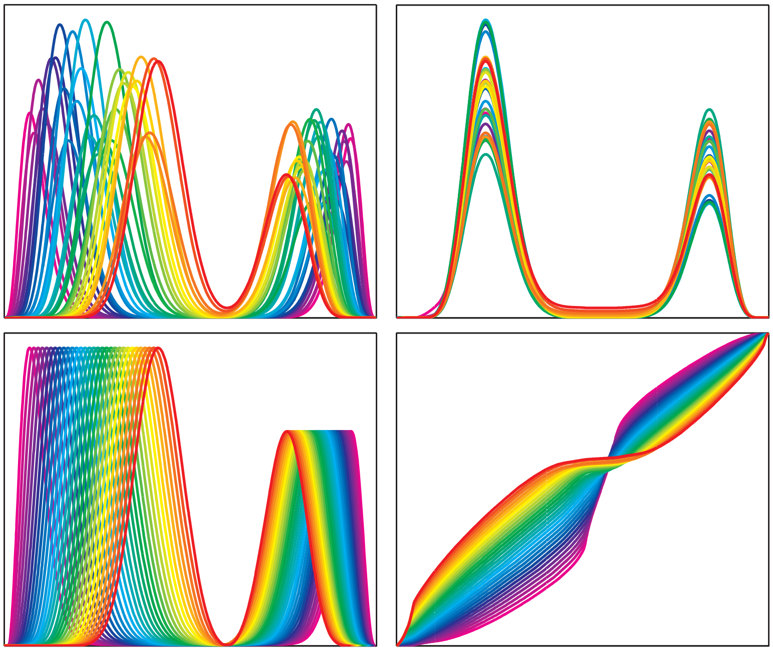

Figure 1.8: Simulated example showing decomposition of an FDA data set (top left panel) into amplitude (bottom left panel) and phase (top right panel) modes of variation. Decomposition is based on the warping functions (bottom right panel). Rainbow color scheme highlights phase variation, with red for closest peaks through magenta for farthest peaks.

The decomposition in Figure 1.8 is much more useful than a standard FDA PCA, which tends to both mix the amplitude and phase components, and also to spread the variation of the phase component over a large number of components, because it is a non-linear mode of variation. As discussed in Marron et al [140] and in Marron et al [141], this type of decomposition is useful in many FDA applications. In some of these, such as the TIC data shown in Figure 1.7, the amplitude data objects are the focus of the analysis, and the phase data objects can be viewed as nuisance parameters. However, in other situations, for example neural spike train data as discussed in Wu at al [232] the phase data objects are of primary interest, and the amplitude data objects are the nuisance component. In still other situations, both amplitude and phase data objects are vital, and in fact their joint variation are important aspects of the variation. These include the variation of theAneuRisk65 vascular shape data in Sangalli et al [184], and in the juggling data discussed in Ramsay et al [179].

Figure 1.9 shows some of the analysis of the juggling data from Lu and Marron [133]. The starting point was positional recordings of location over time of the hand of a juggler, which were reduced to time series of acceleration curves, as discussed in Ramsay et al [179]. These traces were cut into cycles and time registered, to obtain the 113 curves shown in the far left panel of Figure 1.9. Figure 3 of Lu and Marron Ramsay et al [179] shows a variety of PCA type scores plots. The middle left panel shows the version based on the method of Principal Nested Spheres (PNS), from Jung et al [116], which makes special use of the fact that Fisher-Rao data objects naturally lie on a high dimensional sphere. Most PCA variations seem to indicate a homogeneous population. The value added of using this method which takes the curvature of the sphere properly into account, is that it shows two clear clusters, which are highlighted using the graphical technique of *brushing*, i.e. visually separating the cluster through the use of colors and symbols. See Yu et al [237] for more discussion of how and why PNS provides enhanced statistical analysis of Fisher Rao phase data objects. The analysis of Lu and Marron [133] shows that the clusters shown in the center left panel of Figure 1.9 represent important underlying structure in the data. this is also seen in the two right hand panels of Figure 1.9, which show actual vertical and horizontal locations of the paths corresponding to these clusters, using the same colors. These are clearly two quite different types of motions present in the data.

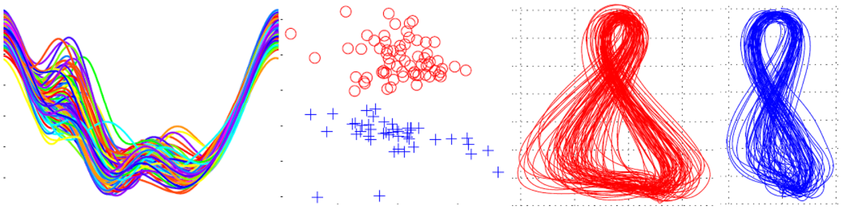

Figure 1.9: Analysis of the Juggling Data. Far left panel shows the input acceleration curves. Center left is the Principal Nested Spheres scatterplot, revealing two distinct clusters, highlighted by brushing. right panels verify these clusters represent two different types of cycles.

1.3 Shapes as Data Objects

A particularly deep and important example of shapes as data objects is the bladder-prostate-rectum data, studied in a series of papers including Chaney et al [37], Broadhurst et al [31], Davis et al [52], Pizer et al [174, 175, 171, 172], Lu et al [132], Stough et al [204], Jeong et al [113], Merck et al [154] and Feng et al [68]. Those analyses were motivated by the challenge of planning radiation treatment of prostate cancer. That treatment is quite effective, but administered over the course of a number of days. The goal is to provide a maximal radiation dose to the prostate while minimizing the impact on nearby sensitive organs such as the attached bladder and the rectum which is adjacent. A major radiation treatment planning challenge is that the locations of all 3 organs vary widely on the time scale of days. Computed Tomography (CT) images are useful for visually locating these organs on a given day, but *segmentation*, i.e. finding the set of voxels inside each organ, was a challenging problem because of poor contrast and noise, as shown in Figure 1.10. That is one slice of a 3-d stack of images, showing a side view of the hip region for one patient. The color scheme of CT is the same as for x-rays, so dense objects such as bones show up as white. Thus the upper right of Figure 1.10 shows the tailbone, and a hipbone passes through this slice in the lower center. Black indicates the least dense regions which are gas bubbles in the rectum, which is the curved lighter region containing the darkest spots starting near the top center and curving down below and to the left of the tail bone. The lighter gray region between the top of the rectum and the small hip bone is the bladder. The prostate, which is the target of the treatment, is a light gray region between the hip bone, the bladder and the lowest visible section of the rectum.

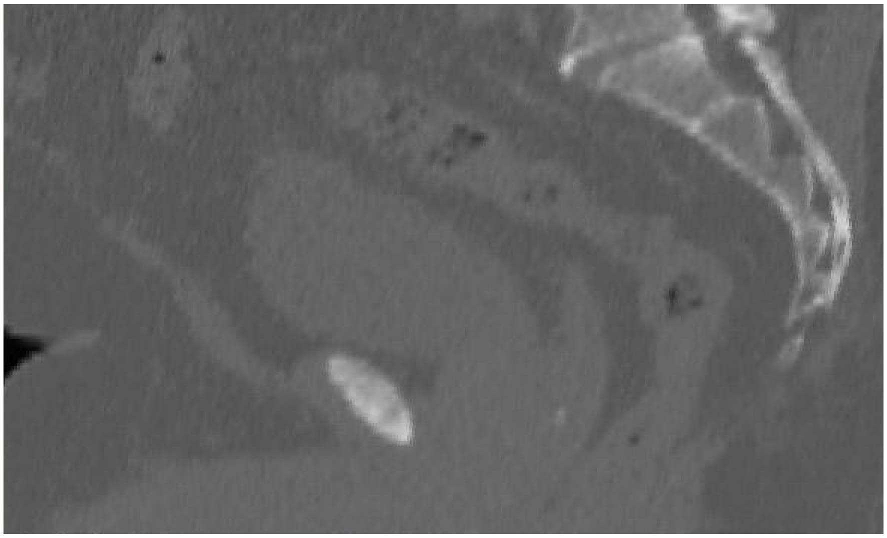

Figure 1.10: One slice of 3-d CT image in bladder-prostate-rectum data. Bones are white, black gas bubbles indicate the rectum. Bladder and prostate are light gray near the center and lower center. Shows that automatic segmentation is very challenging.

Segmentation of the prostate is quite challenging because of very poor contrast with surrounding objects (it is essentially the same shade of gray and has both lighter and darker regions nearby) and because of the relatively high noise level. For these reasons, incorporation of anatomical knowledge is essential to the segmentation process. *Manual segmentation* achieves this through an anatomically trained technician drawing the boundary of an object on each slice of the 3-d image. The union of the interior voxels, aggregated over slices then gives a segmentation of the object. An example of that process is in Figure 1.11, which shows two views of a manual segmentation of the bladder in Figure 1.10. The left panel shows how voxels are aggregated across slices, using a view orthogonal to that where the drawing was done. The right panel is a rotated view of the highlighted collection of blue colored voxels without the CT image, giving a clear impression of the 3-d object.

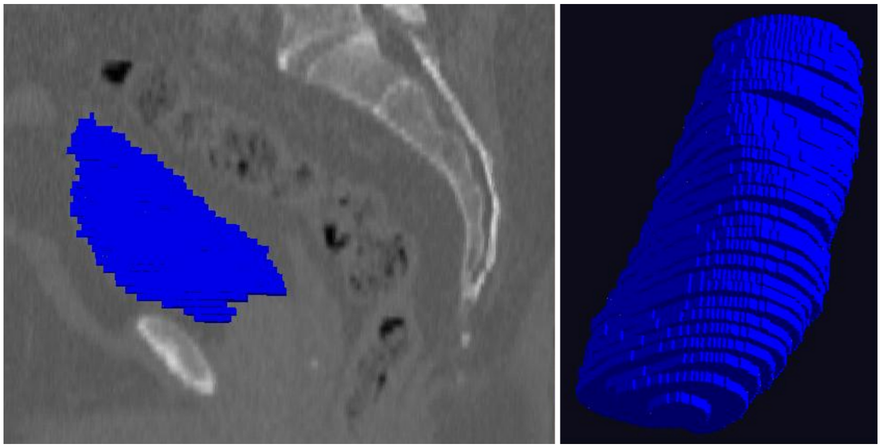

Figure 1.11: Left panel shows the results of a manual segmentation of the bladder, performed on an orthogonal slice. Right panel shows a rotated view of the same bladder, to highlight the 3-d aspect.

While manual segmentation is quite effective at locating these organs for planning radiation treatment, it is time consuming (thus expensive for use in a clinical setting) and hence it is not practical to repeat this operation many times over the course of radiation treatment. This has motivated a lot of research on automatic segmentation of these organs, much of which was developed in the references cited at the beginning of this section. The key idea is to incorporate anatomical information into the training process, using a Bayesian statistical model. The starting point for this is a *shape representation*, i.e. a parametric model for each organ. In some contexts shape is conveniently represented by *landmarks*, i.e. a set of points that correspond across members of the data set, which can be readily found on each. See Dryden and Mardia [61] for introduction to the large literature on statistical analysis of landmark based shape data. Using the coordinates of the landmarks as data objects would not correctly model *shape* because they also include irrelevant aspects such as location, rotation and scaling. Thus shape analysis focuses on data objects where these nuisance aspects have been mathematically removed. There is an interesting parallel here to the idea from Section 1.2 that depending on the context either phase or amplitude data objects could be of primary interest or either could merely represent nuisance variation. In particular, the study of plate tectonics and continental drift is also based on landmark data, as studied in Chang [38] and Royer and Chang [181]. However, an opposite choice of data objects is made, where shape variation is the nuisance, and translations and rotations now become the focus of the analysis.

While landmark approaches are useful for many tasks, they are typically less useful in many medical imaging situations, such as soft tissues, where corresponding (across cases) landmarks can be hard to find, with often very few obvious choices apparent. Hence, there has been much research devoted to *boundary representations*. In the computer graphics world a very common boundary representation is a triangular mesh, see e.g. [161]. A major challenge to the use of mesh representations in shape statistics is *correspondence*, i. e. relating the mesh parameters across instances of shape data objects. Two important approaches to this are Active Shape Models, see Cootes et al [44] for good introduction, and the entropy based ideas of Cates et al [36]. Another major formulation of boundary representations is through Fourier methods, e.g. as in Keleman et al [118]. For sufficiently smooth shapes, Kurtek et al [125] have shown that superior representation comes from enhancing boundary representations by also including surface normal vectors.

As discussed in Siddiqi and Pizer [195], a *medial representation* can provide improvements for a number of imaging tasks. The key idea is to base the representation on the more robust concept of 3-d solids, instead of on 2-d boundary surfaces. For the reasons discussed in Chapter 3 of Siddiqi and Pizer [195], the concept of medial locus has been generalized to give *skeletal representations*. As noted in Pizer et al [176] the enhanced flexibility of skeletal representations allows for superior fits to data. A skeletal representation of one bladder, prostate and rectum instance is illustrated in Figure 1.12.

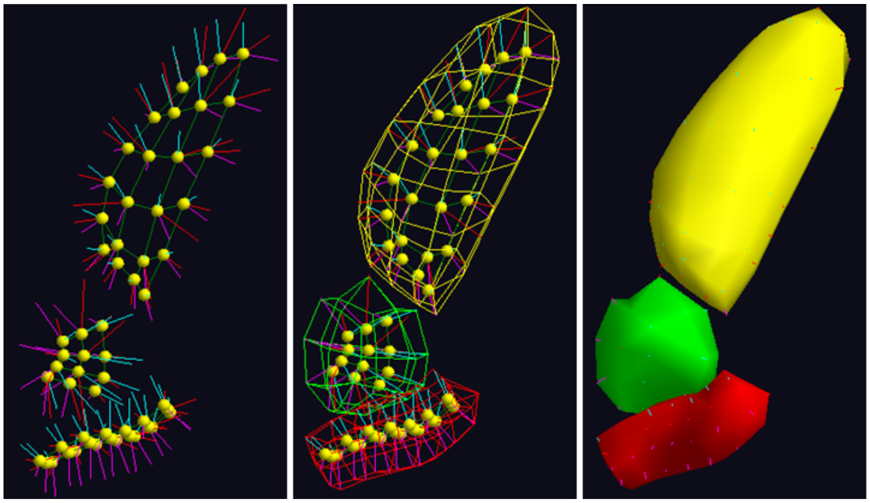

Figure 1.12: Skeletal representation of a single bladder-prostate-rectum. Left panel shows the central skeletal sheets, atoms and spokes for each shape object. Center panel adds the implied boundaries as quad meshes, using yellow for the bladder, green for the prostate, red for the rectum. Right panel represents the implied boundaries using a light source rendering.

The left panel of Figure 1.12 shows the interior components of three skeletal representations, one for each organ. Each has a set of yellow dots, called *skeletal atoms*, connected by green line segments, which are a discretization of the *skeletal sheet*, the 2-d surface which is approximately medial in the sense of being equidistant from both boundaries. Each skeletal atom has *spokes*, shown as cyan and magenta line segments, extending from the skeletal sheet to the boundary of the organ. skeletal atoms at the edge of the sheet each have one additional spoke shown in red, extending to the edge of the organ. The central panel of Figure 1.12 adds three colored meshes (yellow for the bladder, green for the prostate, red for the rectum) which indicate the boundary of each that is implied by the interior components as a quadrilateral mesh that connects the ends of the spokes. The right panel shows the boundary more explicitly by coloring the panels of the quad meshes and using a light source shading in the same colors. The skeletal model is a parametric model of shape, whose parameters are the 3-d locations of the yellow atoms, the lengths of the spokes, and the angles of the spokes, each of which is represented as a point on the sphere S^2.

For CT images where a manual segmentation has been performed, the skeletal shape model can be fit to the binary image shown in blue in Figure 1.11 (i.e. the various parameters estimated), using direct methods such as least squares. However, for clinical applications such as radiation treatment planning, with a need for a technician to perform this operation several tens of times for one course of treatment, manual segmentation is prohibitively expensive. This motivated the work cited at the beginning of this section, on automating fitting of skeletal models, as shown in Figure 1.12 directly to raw CT images as shown in Figure 1.10. As discussed above, this requires incorporation of something akin to anatomical information. That is done using a Bayesian statistical approach. Essentially some manual segmentations are used to train a prior distribution, which is combined with a likelihood based on a new CT image, to generate a posterior distribution which is maximized over the parameters of the skeletal shape representation, to give an automatic segmentation.

The Bayes implementation employed in this type of application differs somewhat from most modern Bayes applications. On one hand, it is relatively simple since it essentially only uses conjugate Gaussian priors, likelihood and hence posteriors. This is a strong contrast with the complicated models involving Monte Carlo Markov Chain methods that are currently very prevalent in applications of Bayes methods. On the other hand, this Bayes application is more complicated than many in two ways. First the number of parameters to fit is typically much higher then the number of training instances. The second complication is the non-Euclidean nature of the parametrization, caused mostly by each spoke naturally lying on the surface of the sphere S^2. The high dimensionality has been handled by a variety of methods related to Principal Component Analysis (PCA). More challenging is that skeletal parametrized data objects are naturally elements of a space of the form $\mathbb{R}^k \times \mathbb{R}^l_+ \times (S^2)^m$. Such spaces are called *manifolds* in differential geometry and are usefully thought of as curved surfaces (e.g. the surface of the sphere). As discussed in Chapter 7, data naturally lying on a manifold present special challenges to statistical analysis. This includes also the statistical area of *directional* data, Mardia [137], where the data objects are angles (e.g. wind or magnetic field directions). Angles are usefully viewed as lying on the unit circle so such data objects are also called *circular data*, as in Fisher [73], and *spherical data*, see Fisher [74]. A good overview of statistical analysis of data on more general manifolds can be found in Patrangenaru and Ellingson [165].

The bladder-prostate-rectum segmentation challenge described above has led to a series of developments in terms of analogs of PCA for data lying on the manifolds of skeletal representations. The Principal Geodesic Analysis (PGA) of Fletcher et al [76] represents an important early landmark in this work. The main idea of PGA is to consider the Euclidean PCA basis as a set of orthogonal lines that (sequentially) best fit the data. In PGA these best fitting lines are replaced by best fitting *geodesics* (e.g. great circles on S^2) which are a natural analog of lines. The results of a PGA, based upon $n = 17$ skeletal representations (collected over a sequence of days) from a single patient are shown in Figure 1.13.

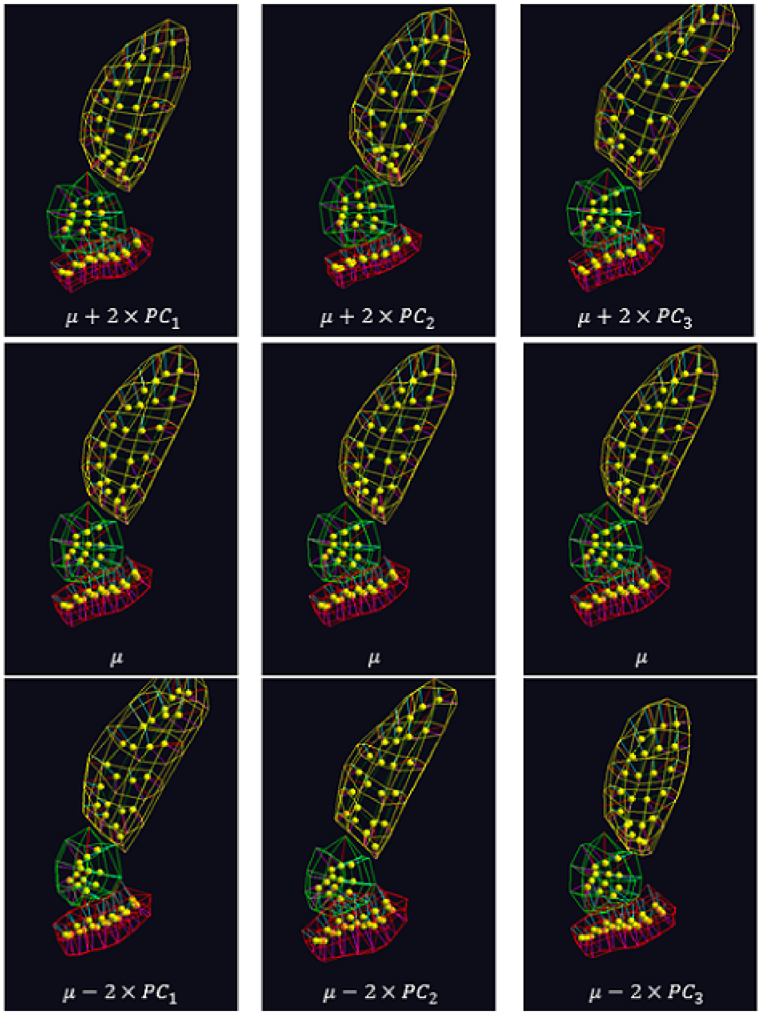

Figure 1.13: Principal geodesic analysis. Modes of variation. Columns give visual impression of first 3 PGA components. All three plots in the second row are the Fréchet mean. Top row shows three +2 standard deviation departures from the mean, and bottom row shows the corresponding -2 standard deviation departures. Shows three interpretable and sensible modes of variation.

Figure 1.13 reveals interesting modes of variation of these organs within this person. The left column (first mode of variation) seems to reflect variation driven by the rectum. The second mode (middle column) shows twisting, while the third (right column) is about emptying and filling of the bladder. This input led to the Bayes segmentation method giving very effective automatic segmentation. This was the basis for the successful start-up company Morphormics, which was subsequently purchased by the radiation treatment equipment manufacturer

Accuray.

More recently there have been a series of improvements to PGA, motivated by a succession of deeper and deeper integrations of statistical ideas with the differential geometry. Detailed discussion of this progression appears in Chapter 7.

While this discussion has focused mostly on segmentation using skeletal shape representations, much important related work has been done on classification as discussed in Chapter 10 and on confirmatory analysis which appears here in Chapter 12. Good overview of the usefulness of skeletal representations, especially in comparison to other types of representations can be found in Pizer [176, 173], Schulz [190] and Hong [103].

1.4 Tree Structured Data Objects

A much different example of OODA is *trees*, in the sense of graph theory, as data objects. An interesting data set, where each data object is essentially the set of arteries in one person's brain, was collected by Bullitt and Aylward [33] and Aylward and Bullitt [13]. While a long term goal is to study pathologies, such as stroke tendency or brain cancer, such cases were deliberately screened out of this data set, to focus on normal variation within the population. Interesting quantities that are useful for various comparisons below are age and gender.

These data objects, for a collection of about 100 people, were collected using a mode of Magnetic Resonance Imaging called Magnetic Resonance Angiography (MRA). This mode flags motion as white, so the flow of blood through the arteries shows up very well. This is seen in Figure 1.14 as the white spots, where the different panels show adjacent horizontal slices of the 3-d image.

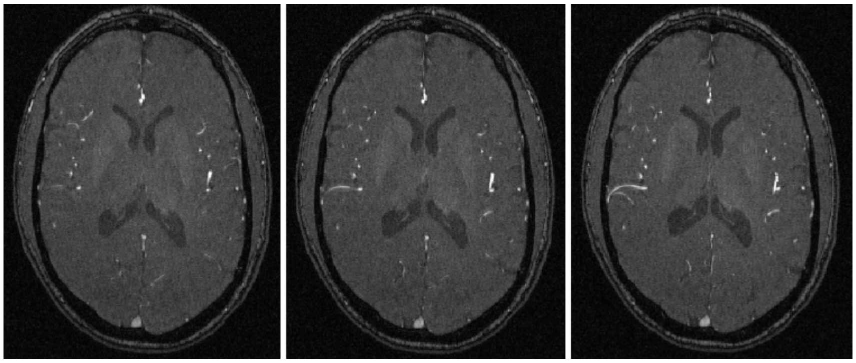

Figure 1.14: Three adjacent slices an MRA image for a single subject. Arteries show up as white dots and curves.

A major contribution of Aylward and Bullitt [13] was the development of a tube tracking algorithm which was used to generate reconstruction of a given artery tree. At this point the data object is the union of many spheres, whose centers follow the central curve of each arterial branch, and whose radii are the branch radius at that point. This tree representation, from the MRA shown in Figure 1.14 can be seen in Figure 1.15. The three panels show different rotations of the same set of arteries. The left and right panels are small rotations, with the closest vessels moved to the left and right respectively.

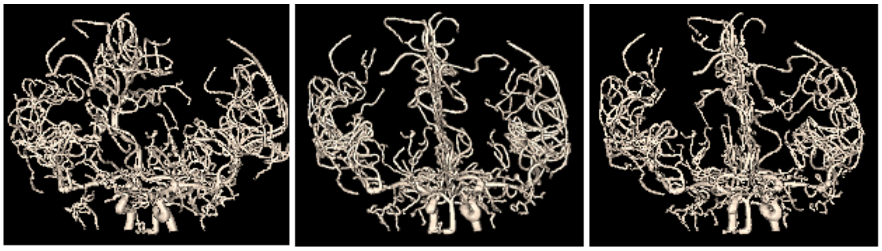

Figure 1.15: Three views of the arterial tree for the subject in Figure 1.14, showing the 3-d structure through somewhat different rotations.

Such data object representations have been computed for approximately 100 people (the original study was a little larger, but some were deleted due to MRA acquisition problems), for example three more of these, for three different subjects are shown in Figure 1.16.

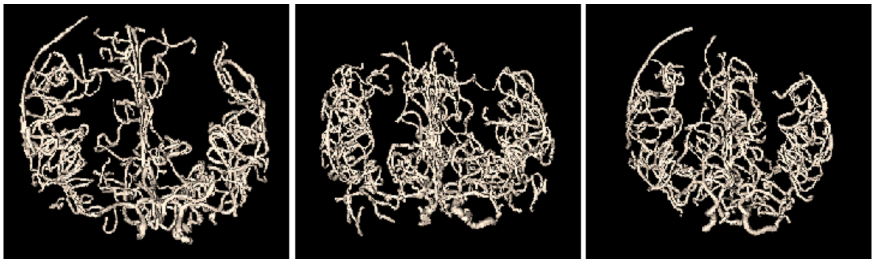

Figure 1.16: Artery tree data objects for three additional subjects.

Data objects of this type present major challenges to doing statistical analysis. For example, it is really not clear how to define even the sample mean of such a set of objects. Understanding variation about the mean, e.g. as done by PCA in Section 1.1, is a further challenge. Some approaches to this that have appeared in the literature are discussed in the rest of this section.

The first versions of PCA like visualizations of these data objects are called *combinatorial*, because they only took into account the branch linkage information, ignoring other aspects such as branch length, thickness, location and physical location. These analyses are in Wang and Marron [223] and Aydin et al [11]. These early analyses involved embedding the trees which naturally lie in a three dimensional ambient space, into a binary two dimensional data object representation, as shown in Figure 1.17. Two arbitrary choices of branch location involving either branch thickness or number of descendant branches, were considered and gave different results.

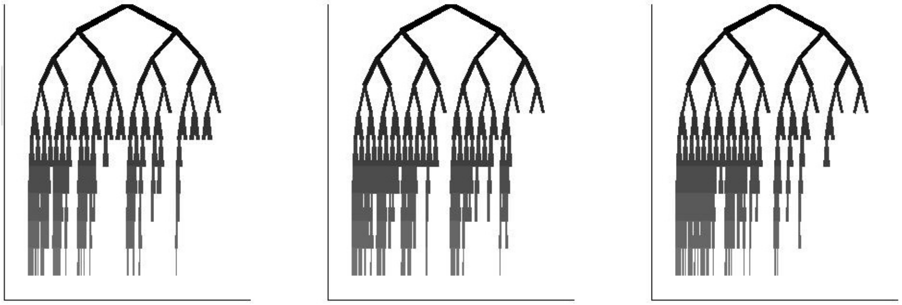

Figure 1.17: Examples of 2-d Embeddings as data objects, for 3 different subjects.

Two challenges with early versions of the brain artery tree data were the linking of tree branches into a tree structure and the starting point of each tree. These issues have been addressed through careful data objects choices. Linking was initially done in Aylward and Bullitt [13] using a thresholding operation combined with manual intervention, and the starting point was arbitrarily chosen by the MR operator. In subsequent analyses, arteries were more accurately linked using a visualization device invented in Aydin et al [12]. Also the starting point issue has been addressed by only including arteries flowing out of the Circle of Willis (a readily identifiable anatomical feature).

Wang et al [224] deeply investigate the relationship between age and artery tree structure and find some unexpected behavior, by inventing an analog of kernel smoothing with a tree structured response variable. More detailed discussion can be found in Chapter 9. See Alfaro et al [4] for another combinatorial approach to PCA of the Brain Artery Data.

A quite different choice of data objects was made in Shen et al [193]. The key idea there was to use the Dyck Path idea of Harris [95] (invented as a tool in the stochastic processes literature for the analysis of branching processes) to represent each data tree as a curve, followed by the use of FDA techniques for the resulting statistical analysis. Several variations were studied. While the above papers were limited to exploratory analyses, Shen et al [193] went on to do confirmatory analyses, which found statistically significant correlations with age, although this is not surprising as this connection was also found in the simple summary based analysis of Bullitt et al [34]. However, a deeper analysis, based on *tree pruning* ideas found the first statistically significant connection of gender with tree structure, see Chapter 9.

Another approach to this data, based on phylogenetic tree representations as data objects, can be found in Skwerer et al [199]. The motivation of that approach was that since phylogenetic trees have been studied for a very long time, in particular going back to Darwin [50] with interesting early graphical representations already in Haeckl [91], much is known about them which should be useful for the study of trees as data objects. The main challenge is that in a typical phylogenetic setting, one works with a common set of species (i.e. leaf set), and the goal is to explore (often to choose between) various ways in which the species could be reasonably organized into an ancestral tree. The main challenge to adapting this idea to the case of the brain artery trees is that the latter do not have a common leaf set. Instead arteries are collected only until they become too thin to show up reliably in the MRA (about 1 mm resolution), so that each person has a different number of arterial endpoints, none of which correspond across individuals in a meaningful way. To create a common set of landmarks and thus create a set of data objects appropriate for a phylogenetic type of analysis, common leaves were artificially generated as a set of corresponding landmarks, based upon the brain cortical surfaces of each person (also collected in the original study), using an elegant algorithm of Oguz et al [160]. See Nye [159] for an early approach to PCA of phylogenetic tree data objects.

A topological data analysis of the brain artery has been done by Bendich et al [18]. That paper uses various *persistent homology* representations as data objects. In confirmatory analysis, these coordinate free representations have given the strongest statistical significance found to date for both age and gender. All of above methods, together with illustrative graphics are discussed in detail in Chapter 9.

Other approaches to data sets of tree structured objects include the *tree kernel* idea discussed in Vert [221] and Yamanishi et al [234]. A mathematically compelling approach to statistical analysis of tree structured data objects, which has not yet been applied to the brain artery data, based on equivalence class ideas can be found in a series of papers including Feragen et al [70], [71] and [5].

1.5 Sounds as Data Objects

Another example of OODA is *sounds* as data objects, which have been studied in a particularly deep way in a series of papers analyzing human speech based on digital recordings. Hadjipantelis et al [90, 89] investigated Mandarin Chinese using a mixed effect model to develop relations between dialects which were consistent with linguistic ideas. Coleman et al [**?**] used these methods to extrapolate back in time to estimate how archaic languages may have sounded. Pigoli et al [170] analyze the relationships between modern romance languages, yielding insights well beyond those available from classical textual linguistic analysis as well as a transformation that provides an estimated reconstruction of how a given speaker would sound speaking a different language. Tavakoli et al [206] combined these analyses with spatial smoothing to produce a dialectic map of the United Kingdom. Shiers et al [194] develop a sound based evolutionary tree for romance languages and dialects.

A typical first step in those analyses is to decompose the raw digital recording of the sound into a spectrogram, which is a moving window version of the Fourier transform, giving a frequency representation in time, as shown in Figure 1.18, from the study of Pigoli et al [170], kindly provided by Davide Pigoli. The top panel is the raw recording of one person saying the word "deux" (two) in French.

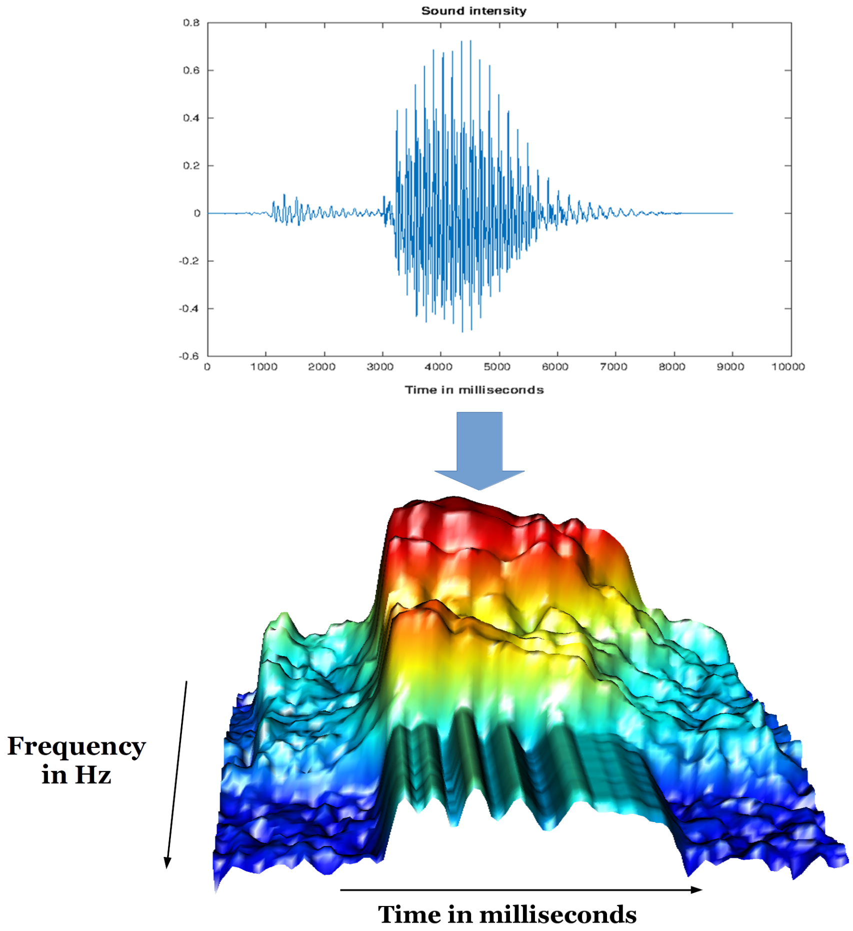

Figure 1.18: Summarization of raw recording of a human speech sound of "deux" in French, top panel, into a corresponding spectrogram which summarizes time and frequency information with color coding height, shown in the bottom panel.

Frequently, the focus is on human speech from the viewpoint that aspects such as pitch and timing are nuisances to be removed from the analysis. For that choice of data objects, those effects are removed by reducing the spectrogram to appropriately defined time and frequency covariance matrices, and also mean vectors sometimes play an important role. Color heatmap representation summaries of five covariance matrices (with entries colored according to the bars on the right, all using the same scale to facilitate comparison) from Pigoli et al [170] are shown in Figure 1.19, also from Davide Pigoli. For each language these summaries are based on aggregating sounds for the spoken digits (1-10). An exploratory visual comparison of these suggest some similarities (e.g. American and Castilian Spanish) and also some stark contrasts such as Portuguese from the others. Confirmatory analysis of these points and a number of others using permutation testing methods can be found in Pigoli et al [170].

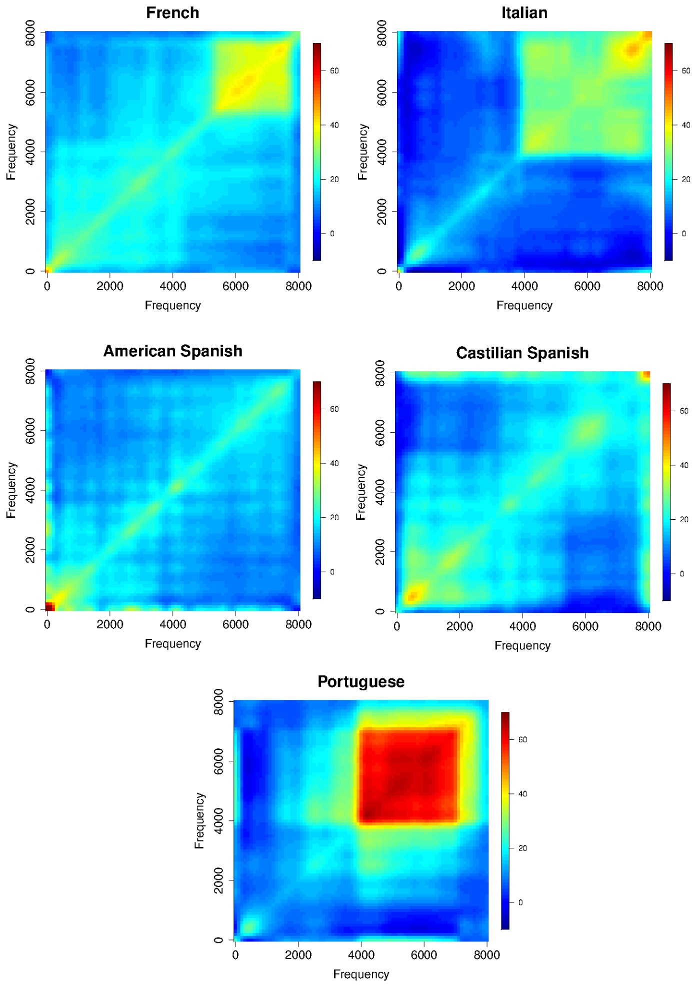

Figure 1.19: Covariance representation summaries of speech sounds from five different languages/dialects. Note strong differences between them, with potentially interesting historical and geographical connections.

In the overall area of sounds as data objects, there is another interesting parallel to the phenomenon noted in Section 1.2, that depending on the context either phase or amplitude data objects could be the major focus of the analysis with the other considered to be nuisance variation. In particular, the above work focuses on a particular type of analysis of sounds as data objects, where the goal is to study human speech, by a variety of speakers. As the human brain does when parsing speech, they deliberately chose data objects which focus on aspects of the sound that are about meaning of the words, which means generally treating issues such as pitch, volume and timing as nuisances, to be mathematically ignored. This a strong contrast with the area of Music Data Analysis, which has been deeply studied in Weihs et al [227] where timing, volume and pitch are actually of keen interest as the data objects.

Statistical analysis of covariance matrices as data objects are particularly challenging. Simple approaches, such as rasterizing the entries of the matrix into a vector and applying conventional Euclidean methods, such as PCA tend to fail, because such analyses typically leave the space of non-negative definite matrices. This issue is generally dealt with by treating the space of covariances as a curved manifold. There are several such manifold representations that are commonly used and may be well distinguished by the metric. The log - Euclidean metric was popularized by Arsigny et al [9]. Fletcher and Joshi [75] point out benefits of a Riemannian metric approach. Dryden et al [59] provide an interesting comparison of different metrics with particular insight coming from studying geodesic paths under each metric and advocate the Procustes metric.

Important work on the statistical analysis of covariance matrices as data objects can be found in Dryden et al [60], Pigoli et al [168, 169] and Aston et al [10]. A large and important application area that uses covariance matrices as data objects in a fundamental way is Diffusion Tensor Imaging, started by Basser et al [15]. Analysis of such data, using local polynomial smoothing methods can be found in Yuan et al [238], and a varying coefficient model approach is given in Yuan et al [239].

The above works demonstrate that it has been very useful to understand covariance matrices as data objects lying on a curved manifold. However, an even more appropriate mathematical context is a *manifold stratified space*. This is a connected set of manifolds of different dimensions. Manifold stratified spaces are appropriate for covariance metrics of varying rank. For each given rank r, the natural data space is a manifold whose dimension is $r\left(r+1\right)/2$. These manifolds are naturally connected across rank through limiting operations where eigenvalues tend to 0.

1.6 Images as Data Objects

The field of image analysis is very large. Statistics has traditionally appeared there in several ways. Early work, with famous papers including Geman and Geman [82] and Besag [21], tended to focus on aspects of mostly a single image, with tasks such as denoising, segmentation and registration being predominant.

However, those fields are now relatively mature, so a currently more important role for statistical ideas comes at the population level which yields a very rich source of potential data objects. For example, the shapes studied in Section 1.3 and the trees featured in Section 1.4 are two types of data objects extracted from images.

But in other situations the images themselves can be treated as data objects. An example of this is shown in Figure 1.20, which shows part of a data set of $n = 108$ images (actually 248×186 gray level photographs) of students from the University of Carlos III in Madrid, kindly provided by Monica Benito and Daniel Peña. Note that there is quite a lot of variation among the faces, yet the human perceptual system clearly indicates that the top row consists of female students, with males on the bottom row.

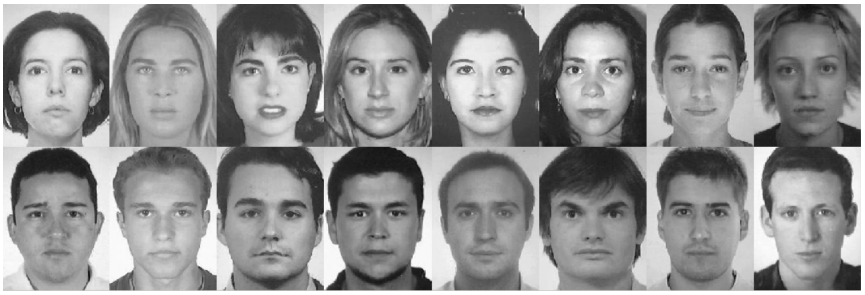

Figure 1.20: Part of the registered student faces image data, females in the top row, males on the bottom.

In an as yet unpublished paper by Benito, García-Portugués, Marron and Peña, male vs. female classification of these data is carefully studied. As discussed in Section 4.4, manual affine registration was used to put each face into a common location in its image. Then the gray level pixels of the images were rasterized into a single long vector, and various classification methods were used to try to understand the difference between males and females. *Classification*, also sometimes called *discrimination*, is an important OODA topic discussed in Section 10. The classification methods used on this face data set were linear methods, as those yield the best interpretation of the results.

Particularly good results came from *Distance Weighted Discrimination* (DWD), proposed by Marron et al [142], as shown in Figure 1.21. DWD is discussed in more detail and compared with other classification methods in Chapter 10. The right panel of Figure 1.21 shows the DWD scores, i.e. the projections of the data onto the DWD separation direction (the normal vector to the DWD separating hyperplane) using a format similar to that of the right panel of Figure 1.4. The red plus signs correspond to the females and the blue circles are the males, which are completely separable using DWD. Also shown are three kernel density estimates, the first for the full population appears in black. Female and Male sub-densities (i. e. rescaled according to sub-sample size) are shown as and red and blue respectively. Top panel gives insight into what DWD is doing with the images, by showing a representative set of 8 reconstructions (i.e. the vectors are converted back into an image) from 8 equally spaced points (locations shown as the 8 equally spaced black bars in the bottom panel) along the DWD separating vector.

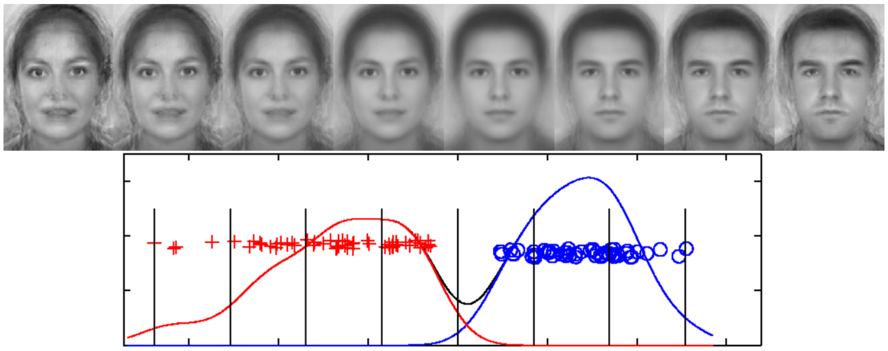

Figure 1.21: Results of DWD discrimination between males and females. Bottom panel shows distribution of DWD scores. Top panel contains 8 reconstructions of faces, corresponding to the 8 points along the DWD separating vector shown as vertical bars in the bottom. Shows clear insight as to how DWD separates males from females.

The array of faces in the top panel is quite compelling. They look clearly very female on the left side, quite androgynous in the middle, and clearly male on the right. Also apparent in perhaps the second and third panel is the idea from Langlois and Roggman [126] that average faces tend to be more beautiful. In addition, note that farther to the right corresponds to stronger masculinity.

Chapter 2

Overview of OODA

This chapter discusses basic aspects of OODA. It also provides an overview of methods discussed in more detail in later chapters.

2.1 Data Object Selection

Any OODA starts with data object selection. This typically has two main components, *determination* of data objects, and their *representation*. Determination involves choice of focus of the analysis, for example in the mortality data of Section 1.1 choosing between age indexed curves over years and year indexed curves over age, and choosing whether to focus on amplitude and / or phase variation in Section 1.2. Representation is more about how data objects should be handled in the analysis, for example studying log probabilities or not in the mortality analysis in Section 1.1, and choosing among the various shape representations discussed in Section 1.3 and tree representations of Section 1.4.

Frequently a *data matrix* is a useful framework for organizing data analytic thoughts. One of the matrix dimensions typically represents the *cases*, i.e. the *elements* of a statistical sample, which are also sometimes called *observations* or *individuals*. Some potentially confusing cross-cultural terminology has arisen in bio-informatics, where a complex biological experiment is used to collect each measurement, i.e. data vector, which itself is sometimes even called a *sample*. The other matrix dimension is used to index *features* or numerical *descriptors* of each data object, with *variables* being a common synonym.

An important issue is that there is a distinct dichotomy in personal preference as to which data matrix dimension is which. From the classical linear

	Number	Synonyms
Cases	n	elements of a statistical sample, observations, individuals, biological samples
Features	d	descriptors, variables

Table 2.1: Commonly used synonyms for cases and features.

algebraic point of view, where vectors are columns, it makes the most sense to treat each data object as a column vector, and then to horizontally concatenate these (i.e. bind the columns), resulting in columns as data objects, with rows representing numerical descriptors. However, from the equally classical statistical tabulation viewpoint, it is perhaps more natural to put variables (i.e. features) in the columns and to hence use row vectors as the data objects.

Keeping this distinction in mind is critical to having meaningful technical conversations. OODA terminology makes this straightforward, by first agreeing whether it will be rows or columns that are the data objects. This choice is often closely connected with software preference. Much mainstream statistical analysis is done using R and SAS, where rows as data objects are the convention. More mathematically oriented work is often done in Matlab where columns as data objects is the more natural choice. Columns as data objects is typical in bio-informatics as well, although this convention appears to be largely driven by the fact that typical data sets tend to have many more features than cases, which were easiest to store in early versions of Excel in that format. The convention here is columns as data objects.

Another point of varying conventions is the letters used to denote the dimensions of the data matrix. Again this is context dependent, with choices like m and n appearing in some areas. Statisticians generally agree that n should be used for sample size, i. e. for the number of data objects. Quite common also is p for the number of variables or features. Less clear is what p might stand for. Some say it stands for *predictors*, but this seems limited to mostly regression contexts. Others suggest *parameters*, which makes sense in contexts where the mean is the focus, but not for consideration of covariance matrices (which typically involve many more than p parameters). The convention here is d standing for *dimension* of the data object vector.

As noted in Marron and Alonso [145], a useful framework for understanding relationships between data objects is through the twin concepts of *object* and *descriptor* spaces. The object space contains the raw curves, images, shapes or trees, while the descriptor space (using terminology coined in Telschow et al [208]) contains some sort of numerical representation, often in vector form.

Example 2.1.1: These spaces are illustrated using the simple FDA example shown in Figure 2.1. The data objects are the $n = 24$ very simple functions shown as black piecewise lines in the left panel of Figure 2.1. This functional form is used here because it is *two dimensional*, in the sense that each data curve is entirely determined by heights of the two x symbols plotted on the vertical lines. Each curve has the constant values of x_1 on [0,1] and x_2 on [2,3], and is piecewise linear between the x symbols (on [1,2]).

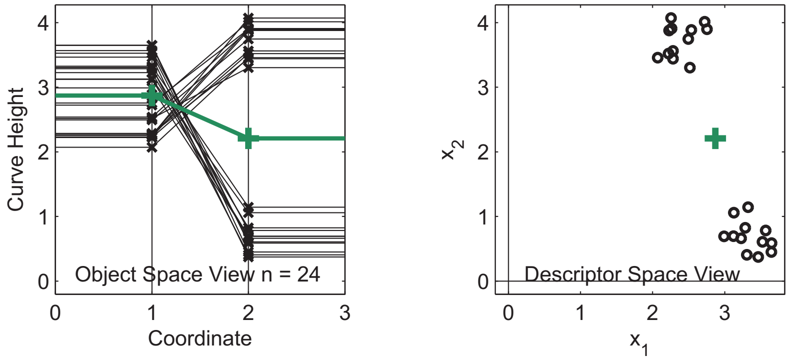

Figure 2.1: Simple 2d toy example illustrating object space (left panel) showing data objects as (simple piecewise linear) curves, and descriptor space representing the same data objects as points (circles in the right panel). Sample mean is shown in green in both panels.

The relationships between these data objects is clearly illustrated in the descriptor space view shown in the right panel, where each black circle represents one piecewise linear function, using a conventional (x_1, x_2) scatterplot. Note that every point (not just the data points) in the descriptor space has a representation as a piecewise curve and vice versa, i.e. there is a one to one correspondence between these spaces. This view clearly shows two very distinct clusters, which are also apparent in the left panel, at least after seeing the right panel. In general higher dimensional cases, the object space is usually of at least somewhat higher dimension, and thus is more challenging to visualize. However, it is often very useful to still think in terms of such a point cloud in descriptor space as a device for considering relationships between data objects (e.g. the clusters apparent here). This concept was used to explain the PCA graphics in Figures 1.4 and 1.5 for the Spanish Mortality data (left hand plots). Graphical devices for visualizing such point clouds are discussed in Section 2.2.

Figure 2.1 also shows the sample mean in green. This is computed as the conventional vector mean in the descriptor space, shown as the green plus sign. The corresponding piecewise curve is shown in green in the object space in the left panel. Note that the green curve is also the point-wise mean of the data curves.

It was seen in Section 1.1 that PCA (recall Principal Component Analysis) can provide an insightful decomposition of variation. This is explored in the context of this same 2d toy example in Figure 2.2. The first principal component is usefully understood in the descriptor space as the direction from the mean (i. e. the red line through the green plus sign), that maximizes projected variation. The projection of each black data point onto the red line is a magenta plus sign, which is the point on the line closest to the data point, connected by a thin cyan line segment. The red line has been carefully chosen to maximize the sample variance of the coefficients of projection, thus giving the direction of maximal variation in the data. By the Pythagorean theorem, it is easily seen that this solution is the same as minimizing the sum of the squared lengths of the cyan line segments.

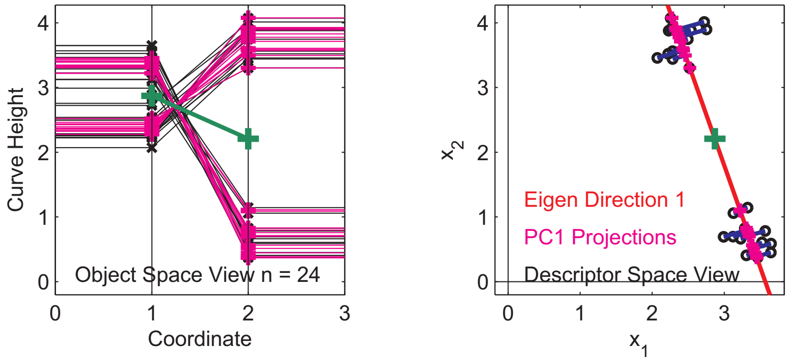

Figure 2.2: First principal component for the 2d toy example. Red line in the right panel is the direction of maximal variation, cyan line segments shows projection to magenta x plus signs. Left panel shows corresponding projected piecewise lines in magenta. Shows how PC1 is the one dimensional approximation which captures most of the variation in the data.

The left panel of Figure 2.2 shows these projections as curves. As the magenta plus signs on the right are close to the black circles the magenta piecewise lines on the left are close to their corresponding black curves. These magenta curves are the best one dimensional approximation of the data, in that they are multiples of the same curve (with respect to the mean). Thus the x_1 heights on the left are negatively correlated with the x_2 heights on the right. This negative correlation is of course also reflected by the circles in the right panel. The magenta curves are also usefully interpreted as a *mode of variation*, with relatively small variation on the left and much larger negatively correlated variation on the right.

Also insightful is the second mode of variation, defined in terms of the second principal component, shown in Figure 2.3. The yellow line in the right panel shows the second principal component direction, which is orthogonal to the red line in Figure 2.2 (generally the orthogonal line with maximal variation, but in this 2d example it is the only choice). Projection of the black data points are shown as cyan plus signs, and the connecting line segments are shown in magenta. Note that lengths of these line segments are the same as the magenta projections in Figure 2.2. Furthermore the coefficients of the cyan projections are the lengths of the cyan line segments in Figure 2.2.

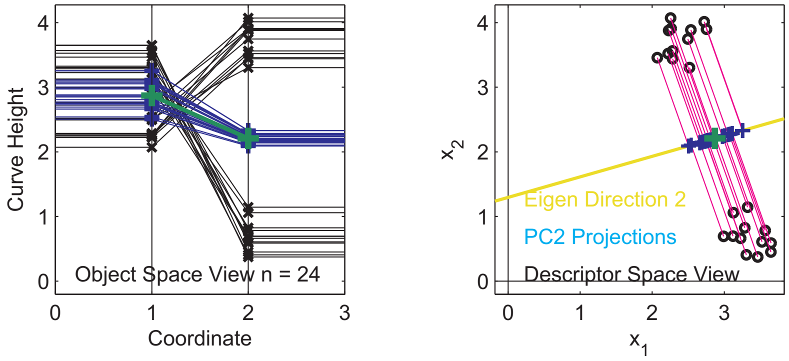

Figure 2.3: Second principal component for the 2d toy example. Yellow line in the right panel is the PC direction, with magenta line segments showing cyan projections. Second mode of variation is shown as cyan piecewise lines in the left panel.

The left panel of Figure 2.3 shows the corresponding mode of variation as the cyan piece-wise curves. Note that there is far less visual variation present which is not surprising because this is the direction of minimal variation. Also note that the x_1 heights on the left are positively correlated with the x_2 heights on the right. This is consistent with the fact that the cyan pluses in the right panel lie on an upward sloping line.

A useful summary of the decomposition of variation shown in the above plots appears in Figure 2.4. The raw data object piece-wise lines shown in black in the upper left are the sum of the components shown in the other panels: the sample mean in green in the upper right, magenta PC1 projections (first mode of variation) in the lower left, cyan PC2 projections (second mode of variation) in the lower right. Note that the magenta first mode focuses on the clustering aspect of the data, in addition to the negative correlation of x_1 and x_2. The cyan second mode of variation is much smaller in magnitude, and contains the smaller scale positive correlation between x_1 and x_2.

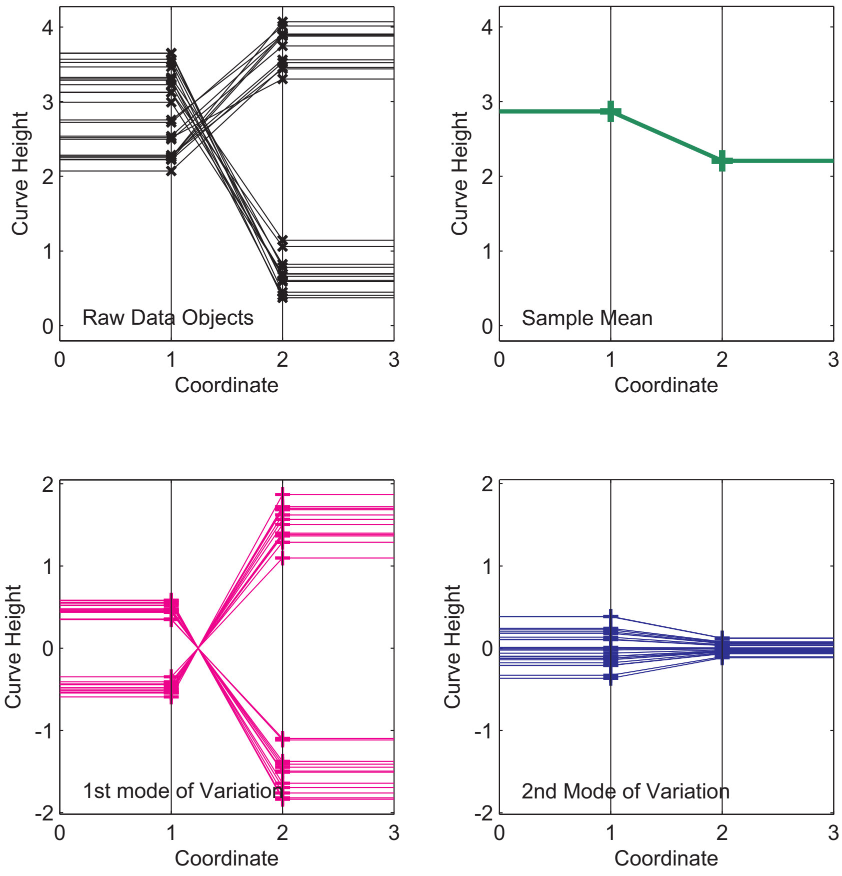

Figure 2.4: PC decomposition of 2d toy example. Raw data objects in the upper left, mean in the upper right (green), 1st mode in the lower left (magenta), 2nd in the lower right (cyan). Raw curves are the sum of others, showing insightful decomposition of variation.

Figures 1.1-1.5, 1.7-1.9 and 2.1-2.4 all include curves as data objects which can be thought of as digital representations of vectors. The relevant plots are all piecewise linear plots, where the heights of the vertices are the entries of the vectors. Such plots have been called *parallel coordinate plots* by Inselberg [110, 111], who advocated them as a general multivariate analysis visualization tool.

In FDA, other representations besides digitization are also commonly used, often based on mathematical *basis* ideas. These include:

- Fourier. This orthogonal basis is very useful for curves which are smooth and periodic. Insightful discussion of Fourier methods in the context of time series analysis can be found in Bloomfield [23] and Brillinger [30].

- Orthogonal Polynomials. There are many such orthogonal bases for curve space. Many useful facts can be found in the classical book Szego [205]. A very useful, and easily accessible summary, of many important aspects can be found in Gradshteyn and Ryzhik [87].

- B-splines. There are many variations of these typically smooth curves, which provide flexible and effective representations of smooth data objects. See Eilers and Marx [64], Stone et al [203] and Ruppert et al [183] for good overviews of statistical aspects of this area. An important classical B-spline reference is de Boor [53].

- Wavelets. This orthonormal basis can give efficient data object representation for curves with varying amounts of smoothness in different locations. See the book Frazier [78] for introduction to this area. Other important references include Mallat [136], Daubechies [51], Donoho and Johnstone [57] and Donoho et al [58]. Different types of useful insights come from exact risk calculation in Marron et al [149], and using spectral ideas in Marron [147].

The Object - Descriptor space concept is also useful for these curve representations, where again the object space consists of curves, but now the descriptor space is the space of basis coefficients. Data analysis methods such as PCA still tend to work quite well performed on the vectors of basis coefficients in that descriptor space, together with insightful visualization of modes of variation seen in the object space, in the spirit illustrated in Figures 2.1 - 2.4. A particularly deep example of this type can be found in Locantore et al [130]. ??? Perhaps add an example later? ???

Another very important aspect of data object representation is *transformation*. The utility of this was illustrated in Figure 1.1, where it was seen that $\log_{10}$ mortality gave much clearer insights than were available from the raw mortality. Data transformation is further studied in Chapter 4.

Sangalli et al [185] gave an interesting discussion of the importance of *sufficiency* in data object choice. A related issue, very important to mathematical statistical analysis of OODA is the data object *environment*. For example, in FDA, there are many ways to measure distance between curves, e.g. there is the whole family of L^p norms. Much of the literature has been dominated by the choice $p = 2$ because of its close relationship with classical least squares, and its tractability. However, when robustness issues are important $p = 1$ can be very useful, and Devroye and Gyorfi [55] offer good reasons why L^1 is more natural in the case of probability densities as data objects. In some cases, such as the occasional need to strongly penalize thin spike departures, the choice $p = \infty$ can be more useful. In other situations performance of derivatives are critical, so Sobolev type norms are the most sensible choice. However, these OODA environment issues run deeper than just the mathematical statistics. In particular, as seen in Chapter 5, even simple data analytis notions such as population center can depend criticially on such choices. Piercesare Secchi nicely summarized this set of ideas as: “Experimental units only become data objects after embedding in an appropriate space”.

There are situations where explicit representation of data objects can be side stepped. An example is when only distances between data objects are measured. There are many methods for handling such situations, discussed in Chapter 5.

2.2 Data Visualization

Data visualization, as illustrated for example in Figures 1.1-1.6 is a very important part of exploratory data analysis. A personal opinion is that it should represent a larger part of statistical training, and of funded research, than it currently does. This seems to be due to statistical models and goals (for example analyzing causality) becoming increasingly complex, which leads to a tendency to co-opt a large share of attention in the field. However visualization is not only important for exploratory analysis and understanding how data objects relate to each other as demonstrated in Figures 1.1-1.6, it is frequently also important to effective choice of data object, and further also provides important reality checks.

Important references on data visualization include Tufte [215], Cleveland [43, 42] and Tukey [218]. These works contain many useful ideas and discussion of what comprises good graphics, although they can sometimes be overly prescriptive. The rest of this section considers two specific types of data visualization that are critical to OODA.

2.2.1 Visualization of Marginals

A perhaps too often ignored, but frequently critical, step in OODA is the study of marginal distributions. Visualizations of marginal distributions, e.g. by histograms or QQ plots, are common when there is time for careful analysis of classical small scale data sets. This often proves very useful in handling variables with strong natural skewness, indicating a potential need for transformation (see Section 4.3 for much more on this), and also in the case of strong outliers, which depending on the context can either be deleted or handled through the use of robust methods (see Chapter 15).

A reason that this step seems challenging in high dimensional contexts is that there are generally just too many variables (i.e. features) to humanly comprehend the structure of all of them. A careful analyst will try to look at some representatives, but it may not be obvious how to choose those. This problem is addressed using the graphical device of *marginal distribution plots* in Section 4.1.

Example 2.2.1.1: An example of a marginal distribution plot is shown in Figure 2.5, for the Spanish male mortality data studied in Section 1.1. This is based on the same data matrix that was used in the left panel of Figure 1.1. Recall that the columns of that matrix (the data objects studied there) were indexed by years. The rows of that matrix are viewed as variables, i.e. features, and correspond to ages. The upper left panel in the marginal distribution plot shows the mean mortality of ages, sorted into increasing order. Note that the first half of these averages all appear to be quite small, with much larger values appearing among the second half. This is consistent with the visual impression from the left half of Figure 1.1 that around half of the ages have mortality orders of magnitude smaller than the rest. This set of sorted means is also the key to finding a *representative* set of variables (ages in this case) to actually visualize. One notion of representative is to look at an equally spaced subset (among the sorted mean ages), as indicated by the vertical dashed lines. The remaining panels show the 8 marginal distributions of ages corresponding to those 8 lines, using the same format as in the right panels of Figures 1.4 and 1.5 and in the bottom panel of Figure 1.21. In particular, the circles correspond to the years (i.e. the data objects, colored using the same year rainbow pattern from Figure 1.2) using mortality as the horizontal coordinate, with the vertical coordinate (and color, red 1908 - magenta 2002) indicating order in the data set (thus the year). The black curve is a smooth histogram, i.e. kernel density estimate, as discussed in Chapter 14.

Note that the first two shown ages, 11 and 19, all have very small mortalities on the order of 10^{-3}. The ages in the middle row, 32, 40 and 59, have medium mortalities on the order of 10^{-2}. On the bottom row, all mortalities are larger. An important issue is that data sets having variables with such diverse scales can be problematic for many forms of statistical analysis. This motivates using one of a number of approaches to data adjustment, discussed in detail in Chapter 4.

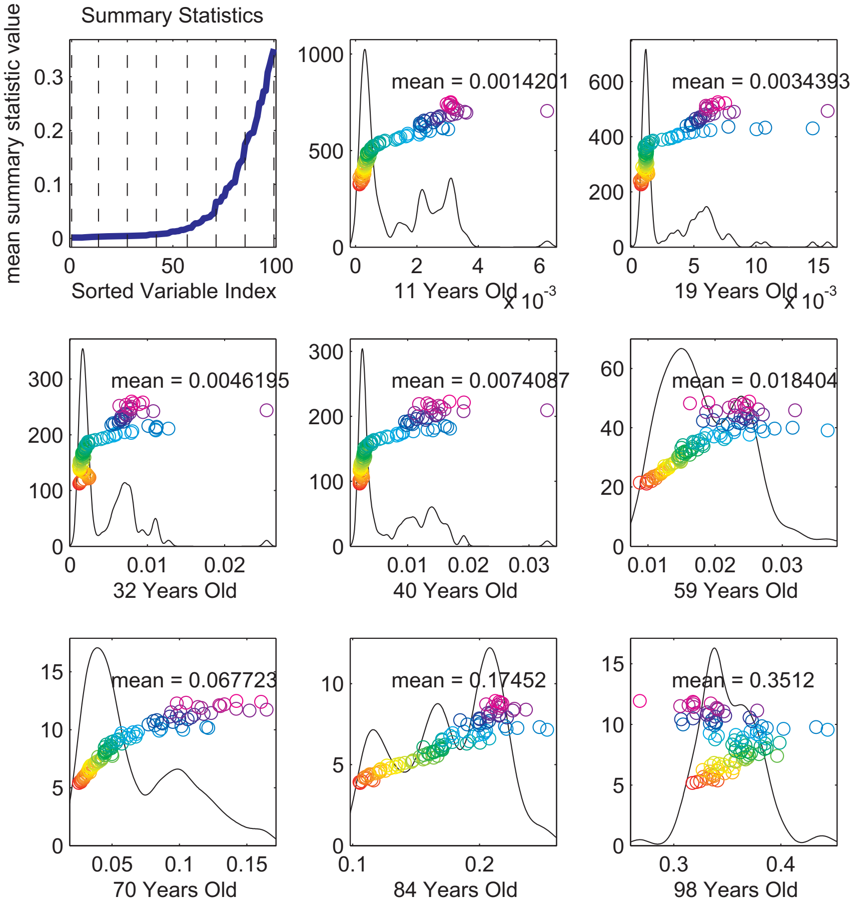

Figure 2.5: Marginal distribution plot of the Spanish Male Mortality data from Section 1.1. Variable means are shown in the upper left panel. Marginal distributions of a representative (equally spaced) set of variables, indicated as vertical dashed lines, appear in the remaining panels. As in Figure 1.2 colors indicate years using a rainbow from 1908 (magenta) to 2002 (red). Shows large variation in variable scaling and many variables have strong skewness and presence of outliers.

These marginal distribution plots show additional challenges to classical analysis methods, such as skewness appearing in most plots, and also the presence of one or more outliers, usually the magenta year 1919 discussed above. These challenges can also be addressed using methods discussed in Chapter 4 on preprocessing.

Given the above described variation across orders of magnitude, log transformation is a natural type of data adjustment to consider for the present data set. Figure 2.6 shows the mean sorted marginal distribution plots for the log transformed mortalities. While there is still natural variation in the means, it no longer spans over several orders of magnitude. This is the reason that the visual impression of variation in the right panel of Figure 1.1 is much more insightful than in the left panel. An added benefit of this transformation is that the skewed distributions above are now transformed into mostly bimodal distribution, which is again very consistent with the fairly rapid overall improvement in mortality, observed in the discussion of Figure 1.4. Note that the impact of the outlying year 1919 is also substantially diminished.

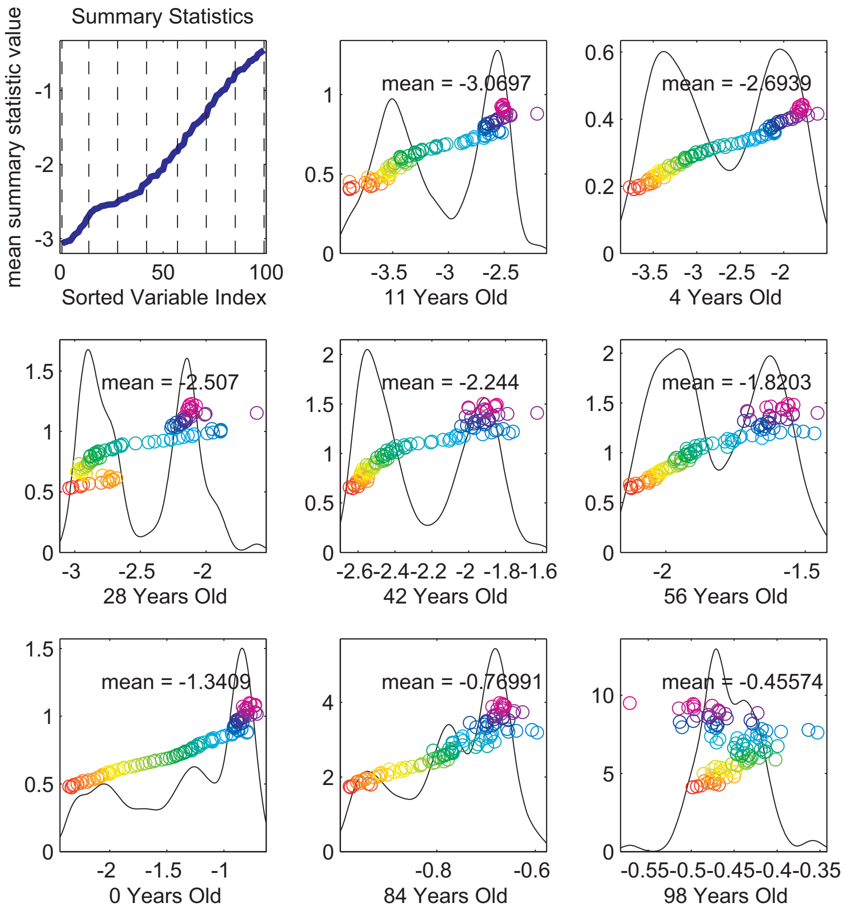

Figure 2.6: Marginal distributions similar to those shown in Figure 2.5, but for $\log_{10}$ Spanish Male mortalities. Shows strong beneficial effects of log transformation.

Finding a representative set of variables (ages) by sorting on the variable means was very effective for understanding critical aspects of this mortality data set, as shown in Figures 2.5 and 2.6. But other summary statistics can highlight different, and very insightful, notions of representative variables as well. For example standard deviation can be a very useful measure of diverse scaling among features. Skewness in distributions can become the focus of such an analysis by sorting on sample skewness. A number of other choices of distributional summaries, and a deep example illustrating their usefulness in a real data context, is discussed in Section 4.1. These include the number of unique numbers in a data set which can be very informative for discrete distributions, and also the number of zeros.

One more issue is the number 8, of representative variables shown in Figures 2.5 and 2.6. This was chosen purely for graphical convenience, in the present format. In other situations $15 = (4 \times 4) - 1$ allows simultaneous viewing of more representatives. A much larger number results in each marginal distribution being too small for easy viewing.

2.2.2 Visualization of Global Structure

This section is about graphical devices that give insights into how data objects relate to each other, for understanding potential clusters and other types of relationships, together with interpretation via *modes of variation*. Several examples are given to demonstrate such relationships.

Principal Components Analysis (PCA) is an effective and commonly used tool for this purpose, as already illustrated in Section 1.1. Ramsay and Silverman [177, 178] made it clear that it is a powerful tool for understanding variation in FDA, i.e. curves as data object contexts. Less well known is that this insightful idea was first published in Rao [180], in the context of analysis of growth curves. Basics of PCA are described in Chapter 16. Main ideas discussed there are the facts that the good idea of PCA has been rediscovered (and generally given different names) a number of times, and that the misconception that PCA is only useful for Gaussian data sets (because one motivation of it is via Gaussian likelihood ideas) is seriously misleading. The latter point is also clear from several of the examples given in this section.

Figure 2.7 shows an FDA toy example to illustrate the concept of decomposition into *modes of variation*, in the spirit of Figure 2.4. The $n = 50$ input raw data curves are shown in the top left panel. These are simulated to have an approximately parabolic shape, but some variation of several types is included as well. Each curve is really just a parallel coordinates plot (as discussed in Section 2.1) of a collection of 10 dimensional vectors, but conceptually it makes sense to think of a bundle of smooth curves.

The object space - descriptor space concept illustrated in Figures 2.1 - 2.3 is useful here, except that explicit visualization of the descriptor space is not done because that space is $\mathbb{R}^{10}$ for this data. None the less, it is still useful to think of statistical analysis as being done in that space, while looking at the corresponding objects space (i.e. curves) view, on the cloud of points that represents the bundle of curves. Colors are based on the Matlab default rotating color palette, with the same colors used in the other panels for visual correspondence.

The top center panel of Figure 2.7 shows a first natural statistical summary:

the sample mean. Again, it is useful to view that curve as the object space representation of the mean of the cloud of points in the descriptor space ($\mathbb{R}^{10}$). The mean curve can also be considered as the point-wise mean of the curves in the top left panel. The top right panel shows the mean residuals, which are a visualization of the curves that correspond to shifting the point cloud so that it is mean centered at the origin. This already highlights an interesting aspect of the data: the parabolic shape of the curves is entirely a feature of the mean, and not the variability about the mean.

The next three rows show the results of a PCA decomposition of variation, of the same type shown in Figures 1.4 - 1.5 and 2.2 - 2.3.

The left plot in the second row, called the *loadings plot* above, shows the first mode of variation. This is based on finding the direction in the descriptor space ($\mathbb{R}^{10}$) that maximizes the projected variation (in the sense illustrated in Figure 2.2), projecting each mean residual curve onto that direction, and then showing the resulting set of curves (as the projection coefficient multiplied by the direction vector). Note that this set of curves has rank one in the sense that they are all multiples of the same curve (which is the curve representation of the direction vector in the descriptor space). This clearly shows that the first mode of variation is essentially a vertical shift. With this knowledge in hand, that mode can clearly be seen also in the mean residuals on the top right, as well as in the raw data on the left. The right panel in the second row, shows the distribution of the projection coefficients, i.e. the *scores*, again with corresponding colors (e.g. the yellow followed by yellow and red on the right correspond to the same colored curves on the bottom of the left hand panel). The format of these scores distribution plots is the same as that used in Figures 1.4, 1.5, 2.5 and 2.6, where each score is represented with a symbol, and the black curve is a smooth histogram. Because there is no special ordering in this data set, the height can be considered to be random, which results in the *jitter plot* idea which was proposed by Tukey and Tukey [216] as a device for visualizing one dimensional data sets. The center panel shows the corresponding PC1 residual curves, each of which is just the centered residual minus its PC1 projection. Note that these are also the projections of the mean residuals onto the hyperplane orthogonal to the PC1 direction.

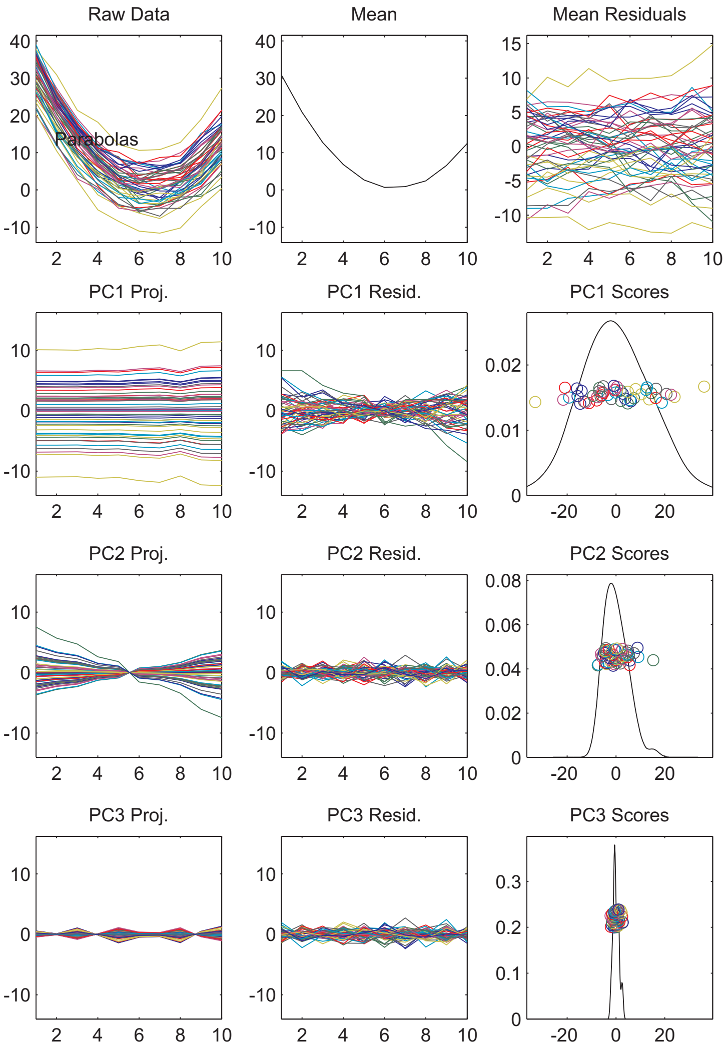

Figure 2.7: A 10-d toy example to illustrate concept of modes of variation in FDA. Top row has the raw data curves on the left, mean center and mean residuals on the right. Remaining rows show PC components (modes of variation), with loadings plots (projections) on the left, residuals center and distribution of scores (projection coefficients) on the right.

The third row shows the second mode of variation, as the object space representation of the projections of the PC1 residuals (2nd row middle panel) onto the 2nd PC direction. Note that this also shows a very interpretable mode of variation, a random tilt. This mode is much harder to see in either the raw data curves, or the mean residuals, demonstrating the ability of PCA to find interesting modes that are not visually apparent. The PC2 scores (i.e. the projection coefficients) in the right panel show much less variation than for PC1, and the PC2 residuals center panel also show relatively less variation.

The fourth row shows the PC3 loadings plot, i.e. third mode of variation. The loadings plot in the left panel looks rather random. This is because the data were simulated as

$$(x-6)^2 + 4Z_{1,j} + 0.5Z_{2,j}\,(x-5) + Z_{3,j},$$

for $j = 1, \cdots, 50$, where x is taken to be equally spaced, and where the $Z_{i,j}$ are independent standard Gaussian. Note the coefficients are deliberately chosen to make these components correctly ordered in PCs 1,2,3. Since the noise terms $Z_{3,j}$ follow an isotropic Gaussian distribution, the PC3 direction is random. The relatively small scale of the noise is also clear from the tightness of the PC3 scores shown in the right panel. Also the PC3 residuals in the center panel show that PC3 explains relatively little of the variation in the PC2 residuals above, again because the noise is isotropic, and thus evenly distributed among the remaining directions in $\mathbb{R}^{10}$. Again these PC3 residuals are the PC2 residuals above minus the PC3 projections to the left.

A useful viewpoint on these issues comes from various sums of squares (in the spirit of Analysis of Variance). The fact that the PC1 projections explain most of the variance is quantified by the sum of squares of the PC1 projections (left, 2nd row) representing 86% of the sum of the mean residual sum of squares. The visual impression that the PC2 projections (left, 3rd row) contain less variation is clear from that some of squares being only 14%. The remaining sum of squares (i.e. summed over all remaining PC components, which is also the sum of the residuals shown center, 4th row) is only 3.6%, confirming that the remaining variation is quite small. The spherical nature of the remaining variation is confirmed by the PC3 variation explained being only 0.7%.

Figure 2.7 also provides an additive decomposition of variation as highlighted in Figure 2.4. In particular, the raw data in the top left panel is the sum of the mean in the top center, the components in the remaining left panels, plus the residuals in the bottom center panel.

As discussed in detail in Chapter 16, the PC direction vectors used in the above data decomposition are easily computed, using either an eigen-analysis of the covariance matrix, or equivalently a singular value decomposition of the mean residual matrix.

A graphical point worth discussion here is the axes used in Figure 2.7. In particular (except for the first row) the vertical axes in the first two columns, as well as the horizontal axes in the third column, are deliberately taken to be the same (even across the rows). Such a view is quite nonstandard for most graphics packages, which generally adhere to the goal of trying to use as much of the graphics space as efficiently as possible, in particular minimizing *white space*. While the minimization of white space is generally a sensible default, in this context it does have an intuitive cost as demonstrated in Figure 2.8, which is a replotting of the bottom 3 rows of Figure 2.7, but this time using axes that minimize white space. The difference between these two figures is perhaps most strong in the bottom row, which is easily understood as small scale noise artifacts in Figure 2.7, but appearing as equal players in Figure 2.8. In particular the important decrease in variation for the higher PC components become much harder to see (only discernible by carefully studying the axis labels). The relative *shapes* (horizontal shift in the first mode, tilt in the second, noise in the third) are now highlighted, at the cost of it being harder to interpret *relative variation*.

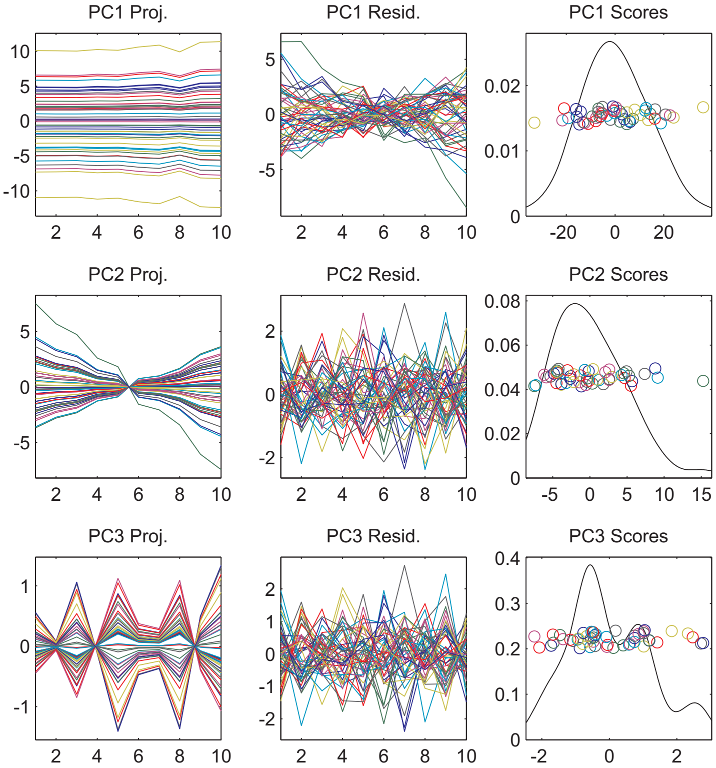

Figure 2.8: Same analysis as in Figure 2.7, showing more typical axis choice to minimize white space. Note less intuitive illustration of decomposition into modes of variation.

One more issue about white space, is that when trying to put a number of plots on a single page, it can make sense to also eliminate the white space *between* plots. The *trellis graphic* ideas of Becker et al [17] provide appropriate ways to do this. ??? May add an example later ???

Another FDA toy example is shown in Figure 2.9, whose format is very similar to Figure 2.8. This time the $n = 50$ data curves shown in the upper left panel, are parallel coordinate plots of vectors in $\mathbb{R}^{50}$. Details of the construction are given below, but at this point consider the data in an exploratory spirit. Apparent is a somewhat higher background noise level, and also some strong structure in the data. The center panel shows the sample mean which this time is essentially constant, so the mean residuals in the left panel are very similar to the original data curves (again using the same axes makes this visually apparent, which would be much harder to see with white space minimizing axes, as in Figure 2.8).

Again the second row shows the PC1 mode of variation. The PC1 loadings plot on the left seems to capture the main twin arch structure, but note that some of these projections are essentially zero. The PC1 scores in the right panel makes it clear that there are actually three strong clusters in the PC1 direction (i.e. this is a very non-Gaussian mode of variation. Perhaps more surprising is the object space representation of the PC1 residuals in the center panel. While some of the residual curves seem to be 0 plus noise, others seem to retain the same arched structure, for reasons discussed below.

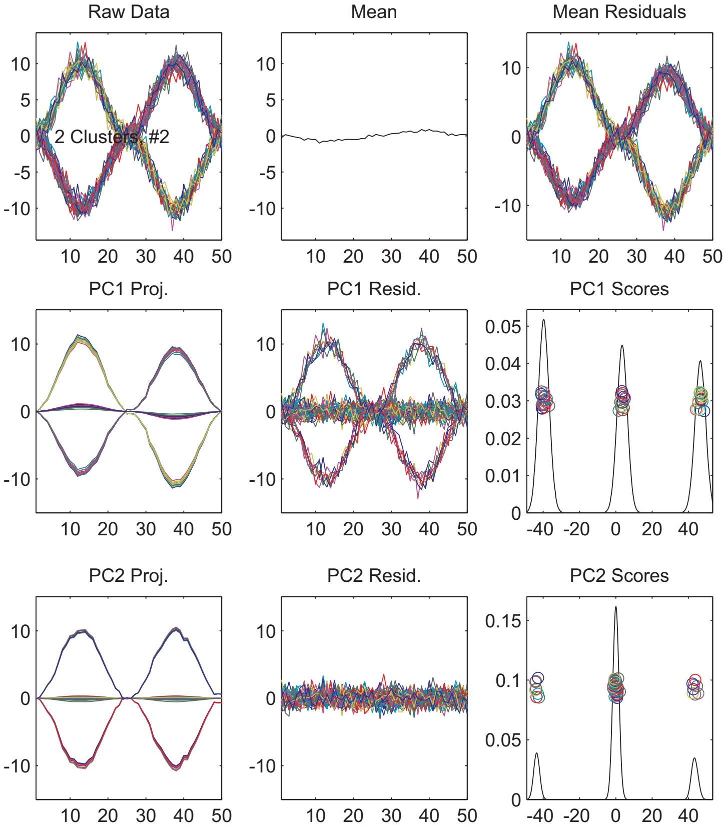

Figure 2.9: Another Toy FDA example in 50-d, illustrating ability of PCA to find insightful modes of variation. This time the mean is negligible and the strong arched structure are artifacts of clustering, i.e. non-Gaussian structure in the data.

The PC2 loadings plot, on the left in the third row, may also at first be surprising. This is because it looks similar to the PC1 loadings just above. But they cannot be similar since these direction vectors (recall each such plot consists of multiples of a single vector) must be orthogonal. A careful look at the colors reveals what is happening. Notice for PC1, the generally yellow color goes up at the first arch and down at the second. Suggesting this function is roughly a sine wave. In PC2 the mostly blue color goes upwards for both arches, while the more red curves go downwards for each, which is a direction orthogonal to the PC1 eigenvector. By the way, these colors have not been deliberately assigned, but are just artifacts of the random generation of the curves, together with over-plotting effects, where the color tends to be dominated by the last plotted curves.

As suggested by the PC2 residual curves, the next components are pure Gaussian noise, with PC directions looking quite random, which thus are not shown here.

As noted in the survey paper by Febrero-Bande and Oviedo de la Fuente [67], a number of FDA software packages aim to integrate PCA with noise reduction in a single step. These include perhaps most notably the FDA package accompanying Ramsay and Silverman [177, 178], and the PACE package started by Yao et al [235]. While this process is critically important in many high noise cases, as well as in the case of uneven and sparse sampling (of the horizontal coordinates of the curves), in perhaps surprisingly many cases such as these two examples, it can be enough to simply do naive PCA on the data. The reason seems to be that often dominant directions of variation, especially those representing interesting modes of variation, tend to lie in smooth directions. In Gaydos et al [80] a variation of PCA, which maximizes smoothness instead of variation, is proposed and integrated with PCA in an interesting way.

An important principal of multivariate analysis is that joint distributions can contain much richer structure than is apparent from the marginals. This concept can be used for a more clear understanding of the structure of the toy data illustrated in Figure 2.9 by studying bivariate projections in addition to the univariate scores distributions shown in the right hand column. Such a view is the *scatterplot matrix* shown in Figure 2.10, which is in the same format as shown in Figure 1.6. This view shows the distributions of the 1-d projections (PC scores) along the diagonal, with in particular the first two being the same as the lower two in the right column of Figure 2.9. The off diagonal plots show corresponding two dimensional plots. For example, the top center panel is the scatterplot of the PC1 versus PC2 scores. Note this is closely linked with the panel below (the horizontal axes are the same, so e.g. the left cluster in the PC2 scores is the same as the left cluster in the scatterplot), and with the plot to the left (where the PC1 score axis becomes the vertical axis so the left cluster in the PC1 scores is the bottom cluster in the scatterplot). This PC1 versus PC2 scatterplot gives a clear view of the underlying structure in this case, there are actually 4 clusters, which project down to 3 clusters in each of the PC1 and PC2 directions. Note that the center left plot is just the transpose of the top center plot. Since this does not convey much new information, the below diagonal plots are sometimes replaced by other graphics.

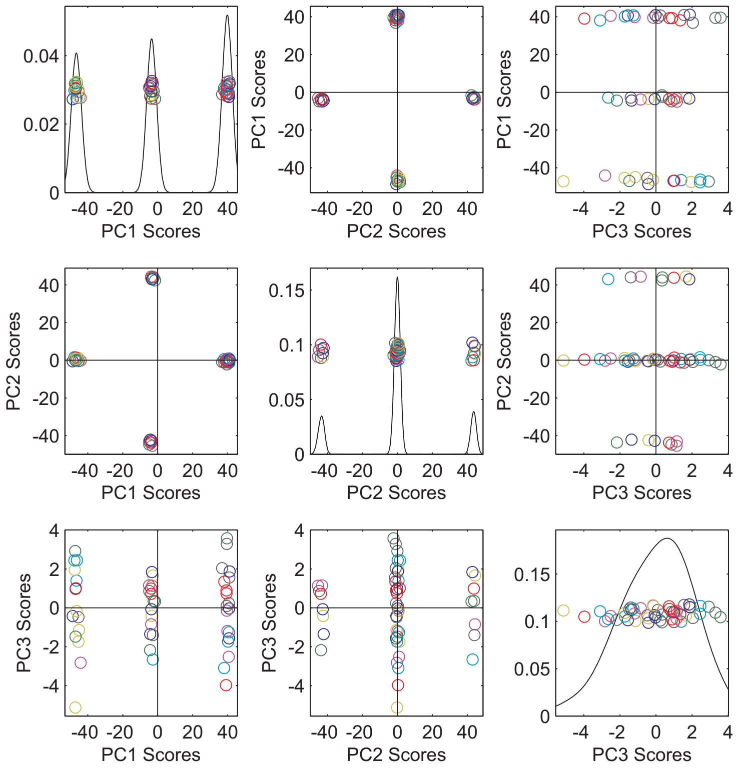

Figure 2.10: Scatterplot matrix view of the data from Figure 2.9. Shows clusters apparent in 1-d marginal scores plot come from marginalizing 4 actual clusters.

The remaining column and row all show the 3rd component. The univariate PC3 scores distributions appear to be Gaussian, which is consistent with how the data were generated. Note that a more conventional white space minimizing choice of axes is used this time, so a careful look at the axis labels is helpful to see that this mode contains much less variation than PC1 and PC2. Again the scatterplots in the right column share the same horizontal axis, so the actually spherical clusters have been strongly stretched by this axis choice. Similarly, for the scatterplots in the bottom row, which are just transposes.

Scatterplot matrix views tend to be very insightful, and are recommended for most situations where relationships between data objects are relevant. A natural question is: if 2-d projections show more than 1-d projections, why not also consider higher dimensional projections? While potential improvements appear to be obvious, and implementation is fairly straightforward for 3-d, it does come with substantial overhead, such as the need for dynamic graphics. These take substantial energy to both implement and to visually explore, which can be a substantial drawback for routine data analysis tasks. For projections of dimension higher than 3, visualization becomes much more challenging, and is thus not frequently done.

Example 2.2.2.3: While toy examples, such as those in Figures 2.7 - 2.10 can give many insights, it is also important to consider real data sets. An interesting example is the set of curves shown in Figure 2.11. These come from an RNAseq study of lung cancer, and in particular was an early data set collected as part of The Cancer Genome Atlas, Weinstein et al [228]. The focus here is on the gene CDKN2A, which has long been known to be involved in many types of cancer. The horizontal axis represents the region on the chromosome that is used to produce the RNA measured here. For each such location, the vertical axis shows the counts (on the $\log_{10}(\cdot + 1)$ scale) of pieces of amplified RNA molecules that match the chromosome at this location. There are $n = 180$ such curves shown here. The $\log_{10}$ scale is useful here, and is the data object representation choice used in the rest of this section, since these counts range over 3 orders of magnitude. An artifact of this $\log_{10}$ scaling for this data set is that very small counts, such as 1 and 2 occupy a large chunk of the bottom of the plot, since $\log_{10}(1 + 1) \approx 0.301$ and $\log_{10}(2 + 1) \approx 0.477$. The same rotating palette of seven colors used in Figures 2.7 - 2.10 is used here as well. The curves seem somewhat chunky in nature, in particular being substantially lower over some intervals, because these coding regions do not appear in a contiguous region on the chromosome, but instead are separated into intervals called *exons*. The union of these exonic regions are used as the horizontal axis in Figure 2.11.

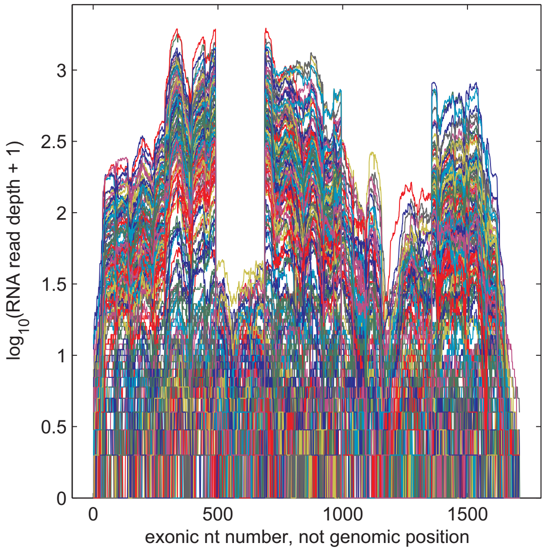

Figure 2.11: Raw log_{10} count curves of Lung Cancer RNA seq data. Colors are standard rotating pallette used for good contrast of curves. It is not easy to discern population structure.

While there is a lot of variation in these curves, it is hard to discern much structure, although they vary over several orders of magnitude. As illustrated above a PCA scatterplot, shown in Figure 2.12, is useful for understanding relationships between data objects. Even the 1-d scores distributions on the diagonal already show multi-modal structure, which is quite apparent in the black smooth histograms. But, as illustrated in Figure 2.10, the 2-d off diagonal scatterplots do a much better job of highlighting the multi-modal structure of the data, where three clusters are immediately apparent. Note that in the spirit of the phenomenon illustrated in Figure 2.10 the top two clusters are combined in the PC1 scores, and the right two clusters combine in the PC2 scores. Only the first 2 components are studied here because the 3rd and 4th components are driven by a few outlying cases, which are not further studied here.

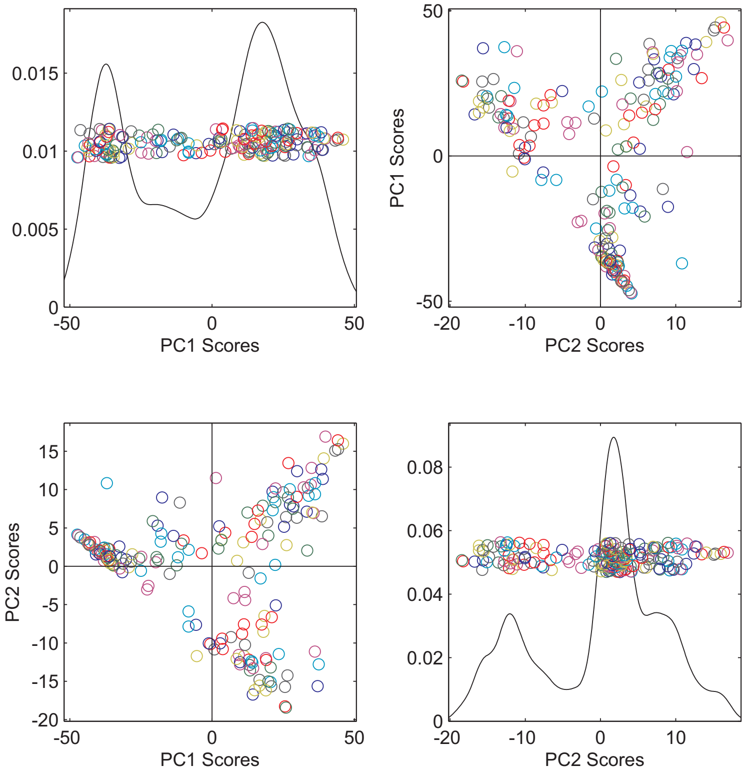

Figure 2.12: Lung Cancer PCA scatterplot matrix. Colored circles represent descriptor space view of corresponding curves in Figure 2.11. Shows three clear clusters in the data.

Insight into the drivers of these clusters comes from a technique called *brushing* in Becker and Cleveland [16]. The idea is to use colors to keep track of subsets of the data in multiple graphics. This is illustrated in Figure 2.13, which shows the same distribution of points as in Figure 2.12, with colors that have now been manually chosen to highlight the three clusters. Automatic versions of clustering can also be done using a variety of methods as discussed in Chapter 11.

Note that the visual impression of these clusters in the 1-d distributions shown on the diagonal is now enhanced with appropriately colored versions of the smooth histogram which focus on each cluster. These are *sub-densities* in the sense that the area under the main black curve is 1, and the areas under each colored curve, which are proportional to the cluster sizes, sum to 1. Note that in regions where one cluster is dominant, the colored sub-density is the same as the black overall density, and thus overplots it (e. g. red on the left side of the PC1 scores and blue on the left side of the PC2 scores). In other regions, the relative curve heights give a clear visual impression of the corresponding cluster proportions.

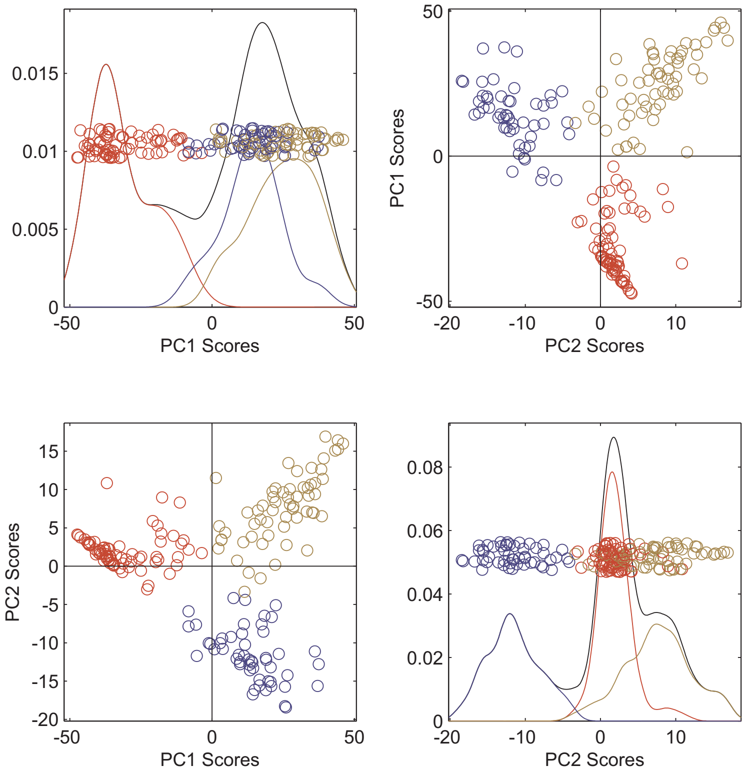

Figure 2.13: Brushed PCA scatterplot matrix, using colors to highlight the three clusters. Also shows sub-density estimates, for visual separation, in the 1-d distributions on the diagonal.

This brushing technique is often especially insightful when used across a

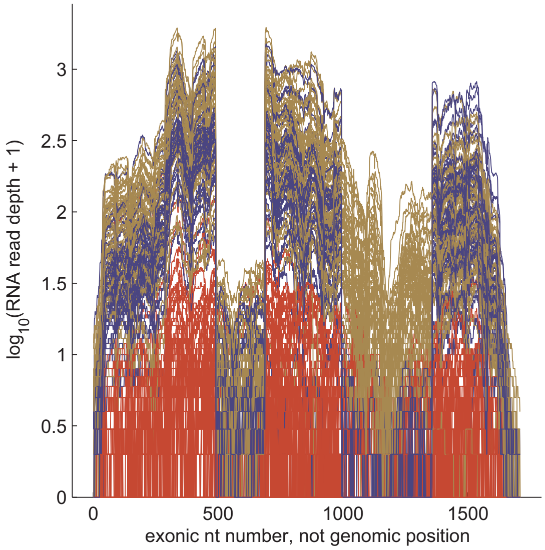

Figure 2.14: Same RNAseq curves as in Figure 2.11, now using brushed colors from Figure 2.13. Shows clusters are very important, in particular representing alternate splicing.

variety of graphical displays. An example of this appears in Figure 2.14. These are the same curves shown in Figure 2.11, but now the color scheme developed in Figure 2.13 is used. Note that the red curves are all substantially lower than the others. In the early days of gene expression these cases would have been labeled as *unexpressed* (which actually means expression at a much lower level, recall the log scale on the vertical axis). Note that for most exons the blue and brown curves both have high expression values. There is an exon to the left of center where all cases seem unexpressed, which is reasonably labeled an annotation error (an on-going issue with such biological data sets). The most interesting issue is the exon right of center, where the brown cases are high, but the blue cases are essentially unexpressed. This is an event called *alternate splicing*, which is very important to the development of new treatments for cancer, because such phenomena can be targeted by appropriate drugs, whereby actually different versions of mRNA are produced from this chromosome region.

While the alternate splicing present in this gene CDKN2A has been well known for some time, the success in finding it with this type of visualization motivated Kimes et al [120] to use this type of idea to scan the whole genome in search of unknown alternate splices. A key challenge was that, as noted in Chapter 11, most automatic clustering methods always find many clusters, whether they represent important biological structure as in Figure 2.14, or not. Hence confirmatory analysis, as discussed in Section 2.3 and Chapter 12 is essential, and was key to the SigFuge method developed in Kimes et al [120].

Another important aspect of data representation is *scale* and *normalization* issues, as illustrated in Figures 2.15 and 2.16. As seen in several examples above PCA can be a powerful visualization device for finding interesting structure in data. But because PCA is driven by finding directions of maximal variation, it can lose effectiveness in situations where differing variables (i.e. features or descriptors) have different scalings. In particular, PCA will tend to be driven by features with the most variation, while ignoring those with smaller scale variation. This challenge can be particularly acute in situations where different descriptors even measure non-commensurate quantities, such as having different units. As noted in many classical texts, such as Mardia et al [138], Muirhead [155], Jolliffe [115] and Anderson [8], this can be handled by *pre-whitening*, i.e. standardizing by subtracting the mean and dividing by the standard deviation. That operation followed by PCA is equivalent to replacing the usual covariaince matrix with the correlation matrix in PCA.

A toy example that underlines this issue is shown in Figure 2.15, whose format is quite similar to Figure 2.7. The $n = 200$ raw data curves in $d = 100$ dimensions appear in the upper left panel, using a rainbow color scheme. Note that the first 20 features (as indexed on the horizontal axis) exhibit a much higher amount of variation than the remaining 80. The mean in the top center panel is essentially 0, so the mean residuals in the top right are nearly the same as the raw data. The first PC mode of variation in the second row is clearly driven by the first 20 features, and is a mode reflecting all 20 features moving up or down together. Similarly the second PC shown in the third row, has a mode of variation which is a contrast between the first and second 10 features, which is orthogonal to the PC1 mode, and of course reflects less total variation. The last row shows some remaining variation of much smaller scale.

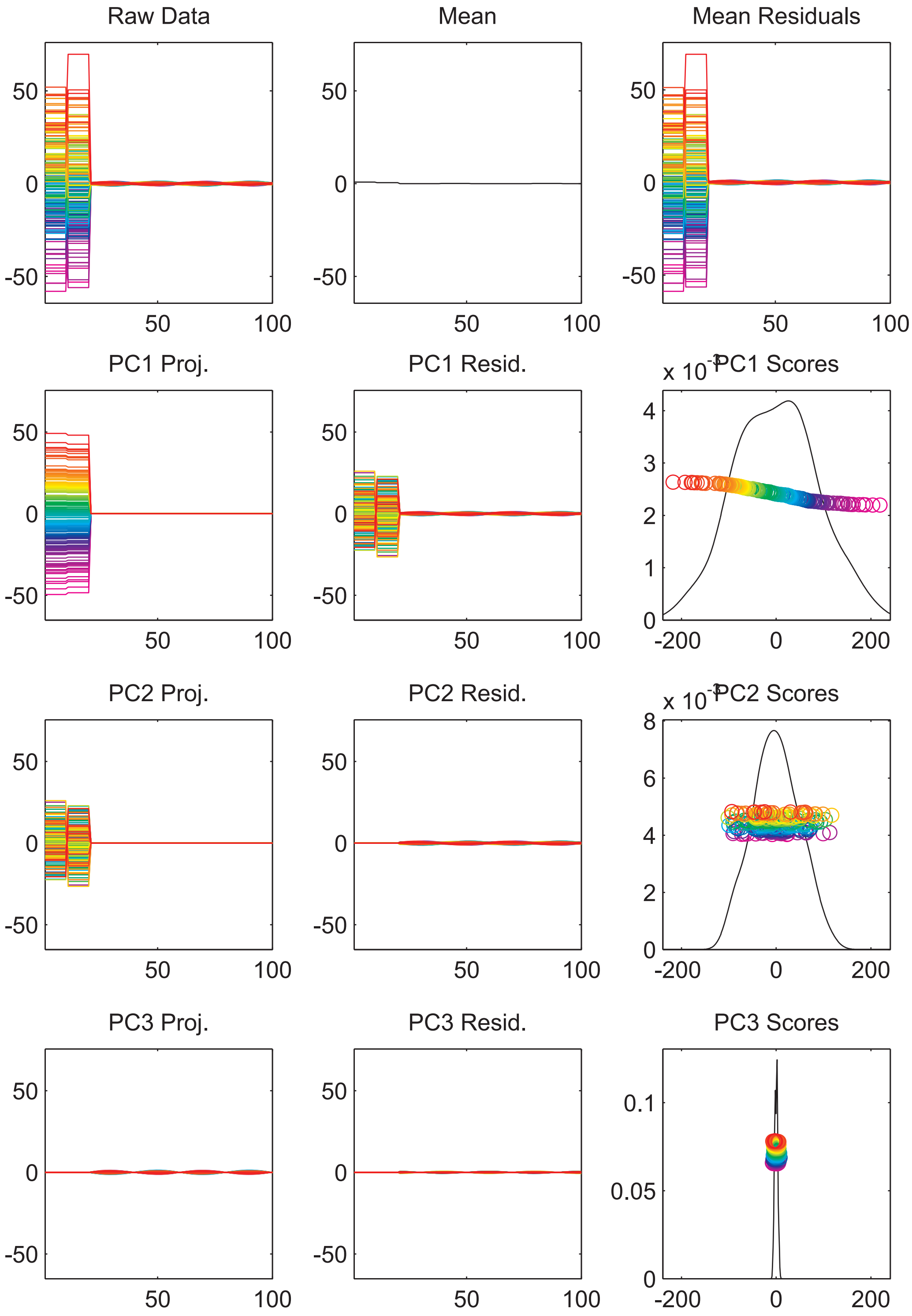

Figure 2.15: Toy FDA example, illustrating challenge of differing variable scaling. The first 20 variables have much more variation and thus drive the first two principal components.

The same data set of curves as in Figure 2.15 is re-analyzed in Figure 2.16, shown in the same format. This time the data are *pre-whitened* by standardizing each variable, in particular each variable has had its sample mean subtracted and been divided by its standard deviation. The similar variation of each feature is immediately clear in the input data plot in the upper left. Unlike Figure 2.15 the last 80 variables are now more prominent. Very different are the discovered modes of variation, as now the first two modes focus mostly on structure in the second 80 features, while the variation that dominated the analysis in Figure 2.15 now shows up only in PC3 and its residuals (which would thus show up in PC4 had that been plotted). The reason that variables 21-100 now drive the analysis is that the magnitude of the signal is now comparable with variables 1-20 and there are simply more of them, giving more overall variation (and they are simulated to be independent, thus the variation goes in essentially orthogonal directions). Quantification of these ideas is given in Table 2.2, which contains the percents of sum sum of squared of each component, with respect to the residuals about the mean. Because the raw data curves in the upper left of Figure 2.15 have almost all their variability on the left side of the range, it is not surprising that part of the range drives two very large PC components explaining almost all the variation in the data, as seen in the top row. The bottom row show a more even spread of variation, which is consistent with the visual impression of the standardized data in the top left of Figure 2.16. In particular, this shows how now it is variation on the right part of the range which has become dominant.

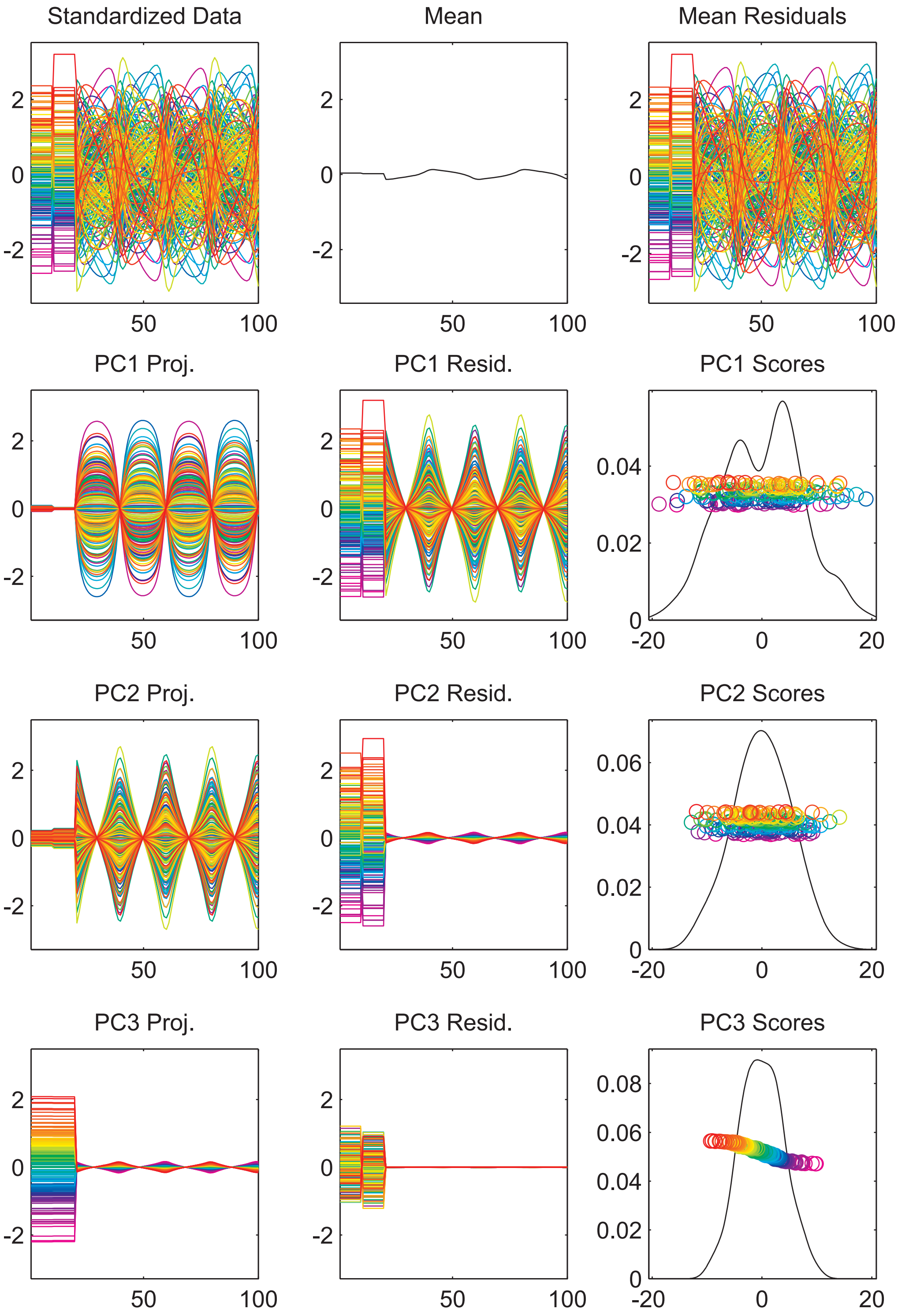

Figure 2.16: Same toy FDA example data set as in Figure 2.15, with prewhitening based on correlations. Now the last 80 variables drive the variation, i.e. appear in the first two components, leading to much different conclusions. This shows data scaling and normalization have a critical impact on this type of analysis.

It is worth considering which of the two very divergent analyses in Figures 2.15 and 2.16 is more appropriate. As noted above, most classical texts on multivariate analysis will recommend doing the analysis based on the correlation matrix (i.e. prewhitening as in Figure 2.16). This is often a sensible default, especially in situations where different variables are measured in different units. However it is important to realize that in other situations the original data scaling may be most appropriate and thus should be preserved. For example, in the Lung Cancer data in Figure 2.14 prewhitening by standardization will result in the small exon starting at exonic nt number 500 playing too large a role in the analysis. Clearly this is an important data object choice, deserving careful consideration in data analysis.

	PC 1	PC 2	PC 3	PC 4
Raw PCA	76%	24%	0.1%	0.03%
Standardized PCA	53%	27%	15%	5%

Table 2.2: Percent sum of squares explained by each PC component for the above two examples. Shows why raw data components focus on structure on the left, while standardization shifts the focus to the right.

More discussion of standardization, together with a real data example, appears in Section 4.2. Other examples studying the tendency of PCA to focus on large scale variation can be found in Chapter 16.

While PCA is a workhorse visualization method, which has frequently found interesting structure in data, one more thing to keep in mind is that because it works through directions that maximize variation, there may be important types of population structure that it might actually obscure. This point is illustrated in the following figures.

Figure 2.17 shows a PCA scatterplot view of another cancer data set, from Hoadley et al [102]. This data set is based on expression of $d = 12478$ genes. Studied here is a subset of $n = 50$ cases (this number gives clear visualization of the main point about the limitations of PCA) from each of the cancer types Bladder Cancer (magenta), Kidney Renal Cancer (blue), Ovarian Cancer (cyan), Head and Neck Squanous Cell Cancer (green), Colon Adenoma Cancer (yellow) and Breast Cancer (red). While each of the six cancer types can be cleary seen, there is substantial overlap of the classes in this view. This is because the PCA directions only maximize variance, and essentially ignore class labels.

Figure 2.18 shows an alternate scatterplot view of the same cancer gene expression data shown in Figure 2.17. The symbols and colors are the same, but instead of using PC directions for the axes, the directions used in the projections are designed to deliberately separate pairs of cancer types. Each direction is based on the DWD method used in Section 1.6 (and discussed in more detail in Chapter 10), this time trained on pairs of cancer types. The projection direction used in the first row and column is DWD trained on only the Kidney (blue) versus the Head and Neck (gren) cancer types. The projections of the full data set onto that direction (although DWD was trained on just those two)

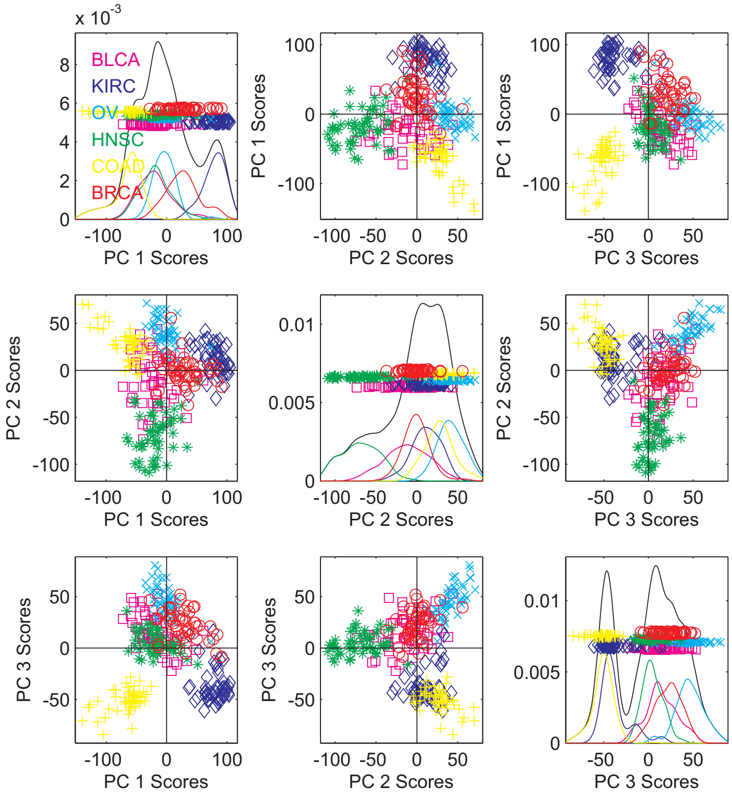

Figure 2.17: NCI 60 microarray data. symbols are tissue samples, colored according to cancer type. Note relatively disappointing separation of cancer types.

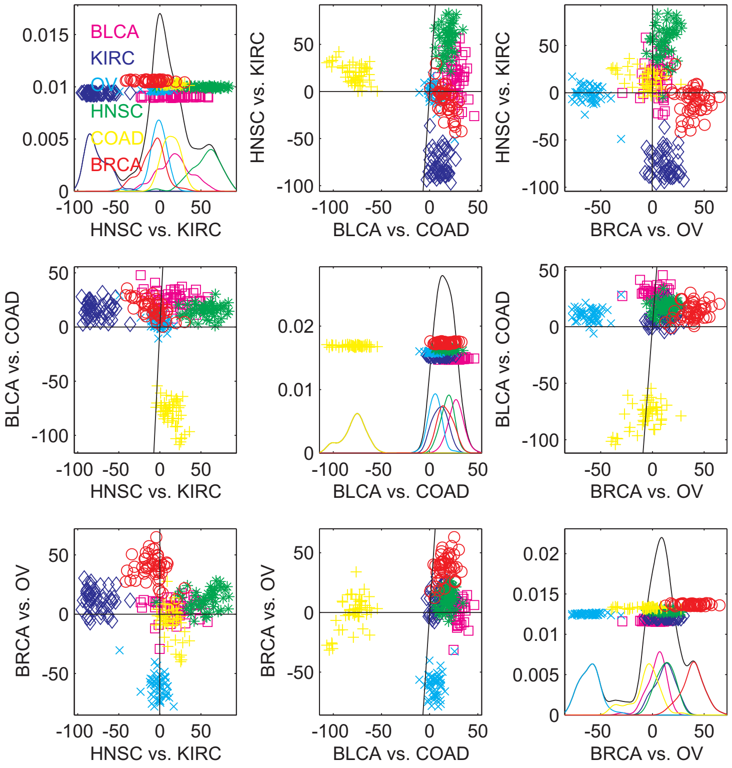

Figure 2.18: Same cancer data from Figure 2.17, with PC directions replaced by directions deliberately aimed at separation of types (highlighted with colors). Shows much better distinction of cancer types, demonstrating that PCA directions may not find all interesting structure in data.

are shown in the upper left and on the same horizontal axis in the other first column plots, as well as the vertical axis of the other plots in the top row. Note that both cancer types stand out as distinct clusters in these views, so DWD has succeeded in separating out the expected biological differences.

Similar excellent separation happens for Colon Cancer (green) in PC2, although the Bladder Cancer (magenta) stands out less well. In the PC3 direction the Ovarian Cancer (cyan) stands out clearly, while Breast Cancer does not. The latter is not surprising because Breast Cancer is well known, see e.g. Perou [167], to have several subtypes which are quite distinct from each other. Presumably each of these would be clearly different from the others, but because of their diversity the union fails to be very distinct in this sense.

One might wonder how these particular pairings of cancer types (on which the DWD directions were trained) were chosen. This was done by considering all pairings and deliberately choosing a set of three on the basis of good visual distinction of types. But the main point of these figures is that there can be visualization directions giving much different visual insights than those available from PCA.

Another striking example of PCA providing not the best separation of cancer classes can be found in Liu et al [129].

The methods and examples studied in this section provide a somewhat non-standard way of thinking about high dimensional data. The currently fashionable notion in much of statistics is that when faced with high dimensional data, one must use approaches such as *sparsity*, i.e. treating most variables as negligible, to reduce the data to a “manageable dimensionality”. While sparsity is a useful approach in some cases and has been tackled effectively using a very large range of methods starting with the LASSO approach of Tibshirani [209], there seem to be many more OODA contexts where the fundamental sparsity assumptions are far from being reasonably well satisfied. These include almost all of the examples discussed in Chapter 1, and also the rich genetic data discussed in Figures 2.11 - 2.14. Yet sparsity ideas seem to be currently both over used and over studied in the statistics literature, perhaps because most statisticians tend to think about high dimensional data in a too *variable centric* way. The OODA viewpoint demonstrated in this section allows taking a more *object centric* approach, where the primary focus is more usefully placed on the data objects and the relationships between them, not the variables. Of course variables are important, but they should be playing the role of descriptors (i.e. representers) of the objects, as opposed to being the focus of the analysis.

2.3 Confirmatory Analysis

The visualization methods discussed in Section 2.2 are very good at providing useful insights and at finding population level structure in data. However an important aspect to keep in mind is that they also have the potential for finding useless artifacts of sampling variation.

This point is illustrated using simulated data in Figure 2.19. Here two classes of data were generated in $d = 1000$ dimensions, with $n_1 = n_2 = 50$ data points in each class. It is hard to see much difference between the red class (shown as circles) and the blue class (shown as plus signs) in the PCA scatterplot view shown in Figure 2.19. However, an important lesson from the Section 2.2 is that for high dimensional data, PCA may not find all interesting aspects of the overall distribution because it focuses only on variation in the data.

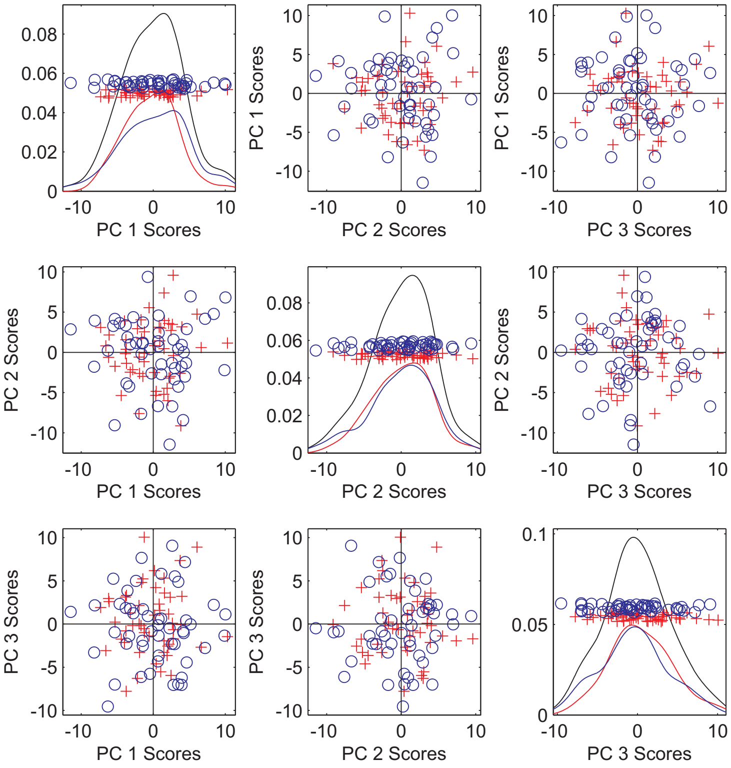

Figure 2.19: PCA scatterplot view of Toy example 2 class data in 1000 dimensions. Shows no apparent difference between classes.

Figure 2.20 more deliberately targets the class difference between the red plus signs and the blue circles, using the DWD direction which was previously used in Figures 1.21 and 2.18. Projections on this DWD direction appear in the upper left panel. Note that this shows a very clear and distinct separation of the two classes which is visually comparable to that seen in Figure 2.18. The other two directions are orthogonal PC directions, which are computed using PCA based on the projection of the data onto the $d = 999$ dimensional subspace orthogonal to the DWD direction. This is useful because it allows the directions used in the scatterplot matrix to be orthogonal, which generally makes the view more interpretable. These issues are discussed in more detail in Chapter 6

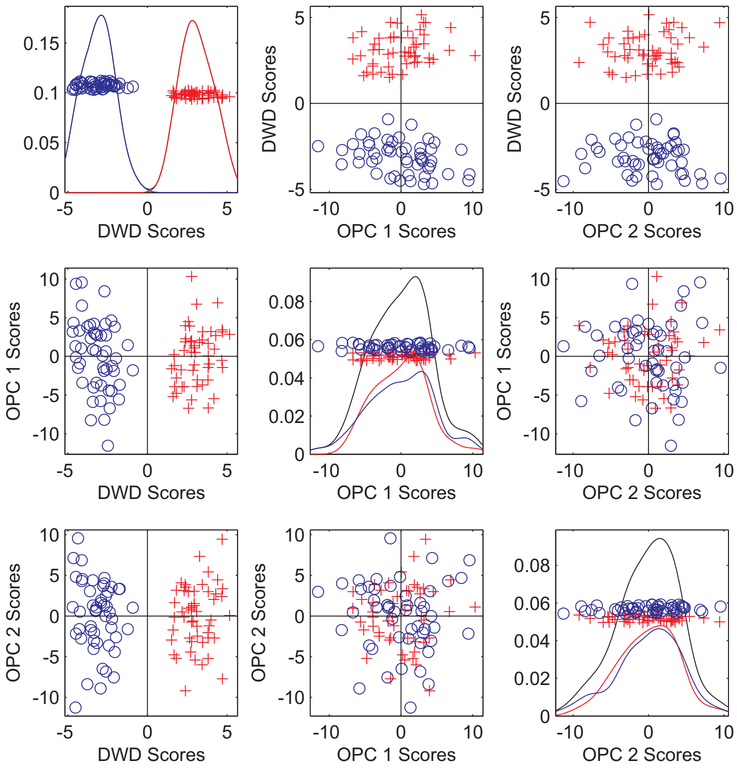

Figure 2.20: DWD and Orthogonal PC view of same data from Figure 2.19. Now shows strong inter-class difference.

While the DWD separation looks very seductive, it is important to keep in mind that DWD is very efficient at finding directions which separate groups of data in high dimensions. But there is another side to this, which is that DWD can be in some sense too good. That is an issue in the example considered, because both the red and blue classes were simulated from the $d = 1000$ dimensional standard normal distribution, $N_d(0_d, I_d)$, where 0_d denotes the d dimensional vector of 0s, and I_d is the $d \times d$ identity matrix. This highlights a fact which is discussed in detail in Chapter 13: high dimensional data can frequently exhibit perhaps surprising behavior. Actually, the specific behavior observed here, and even the amount of visual separation apparent in the DWD direction can be directly explained and even predicted using considerations developed in that Chapter. The fact that this apparent difference is spurious is confirmed in Section 12.1, where it is seen that a hypothesis test for the difference of the means between these two datasets gives a quite non-significant p-value of 0.82.

Figure 2.20 makes a very important point about data visualization in general. While it can be very useful at finding important population structure in data, it is also quite capable of finding things which are just natural artifacts of the sampling variation (which can appear in unexpected ways). For this reason it is critical to combine any exploratory visual analysis with *confirmatory analysis*, as studied in Chapter 12.

In the more complicated areas of OODA, e.g. many of those illustrated in Chapter 1, confirmatory analysis is still in a relative state of infancy (compared to other parts of statistics). One reason for this is that in some of those areas, such as tree-structured or manifold data objects, it can be quite challenging to develop appropriate null probability distributions, which underlie much of classical statistical inference. This has motivated permutation and bootstrap solutions, although careful investigation of their properties remains as a wide open research area in mathematical statistics.

Existing confirmatory analysis methods for OODA are discussed in Chapter 12.

Section 12.1 discusses a generally useful high dimensional permutation type of test, called *DiProPerm*. The key steps are:

- Find a DIRection in the data space, such as the DWD direction used in Figures 1.21, 2.18 and 2.20, although any other systematic direction can used as well.
- PROject the data onto that direction to focus on representative univariate components, e.g. the numbers whose distribution is shown in the top left panel of Figure 2.20. Then summarize the projections with an appropriate summary statistic. A natural choice might be the 2 sample t-statistic. However a surprising result of the careful mathematical analysis of Wei et al [226] is that in the simple difference of sample means provides a more stable hypothesis test in high dimensions.
- PERMute the data to assess statistical significance. In particular randomly reassign the group labels (e.g. red and blue for the data in Figures 2.19 and 2.20), recompute the separating direction, the projections and the summary statistic, to generate one element of a simulated null distribution. Repetition generates a simulated null population and comparison with the original summary statistic provides statistical inference such as p-values.

In Section 12.1 it is seen that while it may not always be the most powerful mean hypothesis test, the DiProPerm test is generally useful because it provides direct confirmation (or not) of visually observed effects, such as the difference between the red and blue groups in Figure 2.20. Further examples exploring these issues, together with real data examples highlighting the importance of this type of confirmatory analysis appear in Section 12.1.

While the DiProPerm test provides a very useful reality check for confirming visualized differences between previously defined groups, care must be taken in the comparison of groups discovered say by *clustering*. The useful operation of clustering can be done in an informal visual way, as for the RNAseq data in Figure 2.13. It can also be done in many more mathematically motivated ways, as discussed in Chapter 11. It is seen in Section 12.2 that application of DiProPerm in clustering contexts can be seriously misleading. Yet the method of clustering has led to many important discoveries in data, so it will continue to be an important tool. In parallel to the challenge of spurious visualization illustrated in Figure 2.20, is the question of "which clusters are really there?" as opposed to being spurious artifacts of the sampling variation. An answer to this question, which becomes particularly challenging in the high dimensional case is the SigClust approach motivated in Section 12.2. Comparison with other approaches is given as well.

There is one more important point aspect of confirmatory analysis in OODA. This is that carefully working from the OODA viewpoint can yield much more powerful and insightful analyses than are available from naive implementation of classical methods. An example of this is the study of osteoarthritis and its impact on knee shape done in An et al [6]. The shape data objects were represented by a set of 60 two-d landmarks, collected from standard x-ray images, using Procrustes methods as discussed in Dryden and Mardia [61]. Earlier work in this area, such as Gregory et al [88] and Nelson et al [158], used PCA to summarize the population structure and then did 2 sample t-tests on the resulting sets of scores. There are 2 ways in which OODA offers improvement in this approach. First is the concept, illustrated in Figures 2.17 and 2.18, that important information in terms of class differences may not show up strongly in any chosen low rank PCA direction. The second is that the multiple testing requires some type of adjustment, such as a Bonferroni correction or False Discovery Rate calculation, which entails additional loss of power. This issue was shown to be serious in the relatively small scale ($n = 65$) study of An et al [6], where a DWD based OODA approach found a statistically significant result, when the PCA and t-test approach did not. See Nelson et al [157] for related results.

2.4 Further Major Statistical Tasks

While data visualization, as illustrated in Section 2.2, and confirmatory analysis as discussed in Section 2.3 are important components of OODA, there are also a number of important analytic methods that are used as well. These include:

- *Distance based analysis.* A number of OODA situations involve data objects which lie in spaces where statistical analysis can be challenging. A straightforward general strategy is to first find a metric on the space and then to compute the matrix of pairwise distances. Chapter 5 discusses various methods for data analysis whose input is only a matrix of distances between data objects. Perhaps most important among these is an analog of PCA called *multidimensional scaling*, Torgerson [211]. A crucial issue in metric based analysis is choice of metric, which is essentially a data object representation issue.

- *Statistics on manifolds.* Chapter 7 discusses data objects lying in *manifolds*, essentially smooth curved surfaces. Relatively simple examples of data objects that are usefully thought of as lying on a manifold include *directional* data, where angles (e.g. wind or magnetic field directions) are the data objects, see Mardia [137] and Fisher [73]. More complicated manifold data objects arise in the study of shape, for example the various type of shape data objects representation discussed in Section 1.3. Statistical analysis of data objects lying on a manifold remains a controversial topic, as there are a number of ways to approach it, with no clear consensus on issues even as to how population centers should be computed.

- *Tree structured data objects.* Even more challenging than manifold data are data objects having a tree structure, in the sense of mathematical *graph theory*. This area is studied in Chapter 9, motivated by a data set where each data object is a representation of the arteries in a human brain. As for manifold data, a number of different analytic methods have been proposed, and it is even less clear which approaches are most natural. A perhaps exotic, but quite successful approach has been *topological data analysis*, done in Bendich et al [18].

- *Classification* (also sometimes called *discrimination* or *pattern recognition*). This is a large field and in fact has become a very important component of the field called *machine learning*. A good overview is available in Duda et al [62] and Hastie et al [97]. This area is reviewed briefly in Chapter 10.

- *Clustering.* Another very large field, with again just some discussion in Chapter 11. The classic reference in this field is Hartigan [96]. In machine learning clustering is often called *unsupervised learning*, to provide useful contrast with classification being called *supervised learning*, since the goals are related, although in the latter class labels are given, while in the former they are derived from the data.

- *Statistical smoothing.* This is one more field with a large literature and many proposed approaches, often with substantial controversy, as reviewed in Chapter 14. It includes *density estimation*, essentially a smoothed version of histograms, and *nonparametric regression* which is essentially scatterplot smoothing. While smoothing methods are commonly used in exploratory data analysis, less well known is the confirmatory method SiZer, proposed by Chaudhuri and Marron [39].

- *Robust Methods.* Once again a very widely studied area of statistics. The main idea is statistical methodologies, which focus on methods with reduced sensitivity to violation of assumptions. Much of that effort has gone towards dealing with outliers, which can be very important in OODA, as discussed in Chapter 15. Major references in this area include Huber [105], Hampel et al [92] and Staudte and Sheather [201].

- *Data Integration.* This relatively new statistical area is driven by the desire in many research areas to make multiple types of measurements and to integrate those in a meaningful way in statistical analyses. In OODA terms, the data objects are typically multiple vectors, which could be merely concatenated into a single vector, but there is often interest in understanding how these relate to each other. This is commonly done using regression methods, which makes sense when the goal is prediction, but not when the goal is a non-directional understanding of the relationship. The latter is accomplished by methods such as canonical correlation analysis, partial least squares and the more general JIVE approach discussed in Chapter 17.

- Thinking about others: Time Series??? (no chapter, add in somewhere? Maybe Chp F, directions for visualization?) Probability distributions & Model based inference??? Design of Experiments???

2.5 OODA Software

Links to available software are provided on the web companion to this book at Marron ???.

- References to available packages

- Marron's Matlab OODA software: http://marron.web.unc.edu/sample-page/marrons-matlab-software/ ??? how to cite? ???

- Show scripts for some of the examples??? Put in an appendix? Just on a referred to website?

- Explain format for later

Chapter 3

OODA Background and Related Areas

This chapter discusses the origins of the OODA terminology in Section 3.1. Other related types of general statistical frameworks are described in Sections 3.2 and 3.3.

3.1 History and Terminology

The terminology Object Oriented Data Analysis (OODA) has a clear connection to the notion of *Object Oriented Programming* from Computer Science. A good definition of that is: *Programming that supports encapsulation, inheritance, polymorphism and abstraction.*

The use of these concepts in a statistical context was pioneered by John M. Chambers and colleagues at the former Bell Laboratories, through the development of the statistical software package S and subsequently S-Plus. See Venables and Ripley [219, 220] for good overview. An important historical point is that S was a major precursor of the currently very popular statistical software package R [207].

OODA itself has its roots in the concept of *Functional Data Analysis* (FDA), which was pioneered by James O. Ramsay and colleagues, see the monographs by Ramsay and Silverman [177, 178] and Ferraty and Vieu [72] for good overview of this area. While this use of "functional" is now quite standard in statistics, it is problematic for researchers with a strong mathematical training, because in that area a *functional* is essentially a function which maps functions into numbers (or more precisely maps a vector space into its underlying field of scalars). Personal discussion with James O. Ramsay led to the realization that the notion of *data objects*, i.e. atoms of the statistical analysis as discussed in Chapter 1, provides the basis of this way of thinking, which led to the coining of the term OODA in Wang and Marron [223].

The perceived value of scientific naming is an interesting cultural issue. Computer scientists seem to enjoy coining many names, trying them out for a while and then frequently abandoning most of them, except for the few that are viewed as having "gained traction". In contrast statisticians have a noticeable tendency to be very careful, in fact are usually quite conservative, about applying new names. Some have observed that at statistical meetings there tends to be too strong a focus on a rather few fashionable areas. At the time of this writing *sparsity* and *FDA* are the over-represented areas, in the past the perhaps overly dominant areas included *kernel smoothing* and *robustness*. A perhaps natural question at this point is whether this apparent narrowness of fashionable research is a consequence of the reluctance to seek new names.

The terminology OODA itself has raised objections on occasion. For example, Lu et al [134] contains an example demonstrating the value of the OODA viewpoint. The example came from the desire to automate the basic biological science practice of growing cells in *wells* on a plate. A challenging part of that automation was making the decision of when to move a subset of the cells to a new well based on digital images, because they have grown to fill the capacity of the current well. The issue of what should be the data objects, between features summarizing aspects of the whole well (e.g. cell counts) and features of individual cells (e.g. shape and size aspects), turned out to be pivotal to the investigation and even led to some interesting theoretical work discussed in that paper. An early submission of that paper was rejected by a well known journal on the grounds that the terminology of "data objects" did not bring added value over the more traditional "experimental units". This point made sense for that particular project, but is limited in the context of the larger data analytic picture. In particular, generally choice of data objects includes not only experimental units, but also data representation issues, for example the choice of original versus log scale illustrated in Figure 1.1 of Section 1.1, the choice to focus on amplitude and/or phase variation in Section 1.2, the choice of shape or tree representation discussed in Sections 1.3 and 1.4, and which aspect of sounds an analysis should be centered on in Section 1.5.

The discussion of the overview paper by Marron and Alonso [145] covers quite a few other interesting aspects of OODA.

In some situations, there have been variations on the name OODA. For example, in 2010-2011 the Statistics and Applied Mathematical Sciences Institute hosted a program on OODA under the name Analysis of Object Data. That version of the name is also prominently featured in the monograph Patrangenaru and Ellingson [165], which provides an important overview of statistical analysis for data lying in manifolds and manifold stratified spaces.

3.2 Compositional Data Analysis

The field of statistical *compositional data analysis* goes back at least to Aitchison [1]. The original motivation was the study of variation in geological composition, in terms of vectors of proportions.

Data objects in that context were typically vectors $v \in \mathbb{R}^d$, each entry of which is the proportion of a given material in the geological sample. Note that each such object $v = (v_1, \cdots, v_d)$ is a point on the unit simplex in $\mathbb{R}^d$, i.e. $v_j \geq 0$ for $j = 1, \cdots, d$ and $\sum_{j=1}^{d} v_j = 1$.

Good insight can come from considering such data objects to be ***probability vectors***, as is common in Markov Chains, see e.g. Hastings [98]. In econometrics terminology, such data objects are sometimes called ***fractional responses***, see Papke and Wooldridge [162, 163] and Murteira et al [156].

Data objects restricted to the unit simplex create some serious statistical challenges. For example, standard Euclidean analysis methods such as PCA (see Sections 1.1 and 2.1 and Chapter 16), or even use of the Gaussian distribution for statistical inference become clumsy at best, because such methodologies tend to leave the unit simplex.

An often advocated choice of data object in this context is the *log-ratio* method, developed by Aitchison and Shen [3] and Aitchison [1, 2]. This approach has worked well in many analyses, and is especially appropriate when the primary focus is on *ratios* of different amounts. However, in other situations, there can be a cost of some distortion, particularly when some entries are 0 or near 0. This has motivated other data objects choices for compositional data analysis, such as the square root transform which moves the data from the unit simplex to the unit sphere, in e.g. Scealy and Welsh [187]. Other power transformations have been proposed and studied by Tsagris et al [214] and Scealy et al [186]. Butler and Glasbey [35] address this issue using a latent Gaussian modeling approach, while Stewart and Field [202] took a mixture modeling approach. Scealy and Welsh [188] provide a fascinating historical discussion of major controversy that has occurred over their data object choices.

See Xiong et al [233] for a quite different example of data objects on the unit simplex, in the context of virus hunting using DNA methods. That paper also considered unit sphere versus simplex data object representations and found the best performance in that case came from working directly on the unit simplex.

3.3 Symbolic Data Analysis

Another statistical area related to OODA is ***Symbolic Data Analysis***, see monographs Bock and Diday [24] and Billard and Diday [22]. The goal of that area is to find intuitive summaries of various aspects of relational databases. These summaries, are called ***symbols***, which are distributional summaries, such as ranges (intervals), frequencies (for categorical variables), histograms or quantiles. There are at least two levels of relationship between Symbolic Data Analysis and OODA. In some situations, some type of symbol (e.g. probability densities) can be the data objects of interest. However, given any set of data objects, the large and well developed set of Symbolic Data Analysis ideas can provide a number of types of useful summarizations via symbols of the data set.

An important historical note is that the terminology Symbolic Data Analysis came first, going back at least to Diday [56].

3.4 Other Research Areas

There are several other areas, not discussed in detail here, where OODA ideas and terminology are potentially very useful, mostly because of the many complicated research questions that are typically addressed there.

One is *Object Oriented Spatial Statistics*, reviewed by Menafoglio and Secchi [153]. In this area a number of the tasks and approaches considered in this book are extended to the important case of spatial data. These include data sets where location plays a key role, and must be properly included in competent analyses.

Another such area is *Natural Language Processing*. This area aims to develop algorithms for the computational extraction of meaning from text. One part of that field is called *latent semantic analysis*, see e.g. Martin and Berry [151], where the key idea is singular value decomposition (essentially PCA without mean centering, as noted in Chapter 16) of some variation of an *occurrence matrix*, which summarize appearance of words in large sets collections of documents. As noted in Berry and Browne [20], there are many data object choices to be made, in terms of both how to summarize word/phrase occurrences and also how to weight various aspects of the decomposition.

Yet another such area, which has had a major impact on both neuroscience and also the study of many aspects of human behavior is *Functional Magnetic Resonance Imaging*, see e.g. Huettel et al [109]. This method involves brain imaging over time, using blood flow as a surrogate for brain activity, measured at a set of *voxels* (the three dimensional version of pixels). Many choices of data objects have been made in this area. In some studies, the focus is on a particular voxel (thus one brain region), so the time series at that point is the data object choice. In other studies the behavior over time is summarized to a single number, so the data objects can be three dimensional objects. Still other studies treat the full 3-d movies over time as data objects. An example, showing joint analysis of how imaged brain function jointly interacts with behavioral scores is discussed in Section 17.

One more research area with close links to OODA is *Deep Learning*, which aims to provide computational methods that work in ways parallel to the human brain. Main methods in this area are based on *neural networks*, which go back at least to McCullough and Pitts [152]. That area was quite popular in the 1990s, but seems to have been over-advertised at the time, with many attempted applications apparently failing to live up to their promise. However, more recently there has been a very strong resurgence, perhaps fueled by much larger typical data sets, together with much more powerful computing capabilities. These ideas have created research revolutions in areas such as computer vision. See Demuth et al [54] for important ideas in this area. Benjio et al [19] suggest that much of the success of deep learning methods comes from the ability of neural networks to provide a type of *automatic data representation*. For example, in classification tasks, the last step is typically a classical method (of the type discussed in Chapter 10), while the preceding neural layers are usefully viewed as providing inputs, via a search over a very large potential feature space. This can be viewed as an interesting way of automating the step of data object representation, as discussed in Section 2.1.

In the context of Natural Language Processing, Baroni et al [14] showed that in many cases neural network based word embedding algorithms gave better performance than traditional matrix factorization based approaches, for a variety of standard measures. However, Levy et al [127] demonstrated that these performance gains are likely due to data object choices that can be easily carried over to make the traditional matrix factorization approaches achieve state of the art performance.

Chapter 4

OODA Preprocessing

An acronym going back at least to the early days of computer programming was GIGO for "Garbage In - Garbage Out". That principal certainly applies to modern data analysis, yet seems to be all too frequently ignored. This chapter describes some useful ways for understanding data problems and some remedies, that scale in a reasonable way to larger data sets. Section 4.1 gives examples demonstrating the importance of a careful study of marginal distributions and how they can be used to guide data object choice. The often useful approach of normalization (usually shifting and scaling of variables, but with some often non-obvious variations) is discussed in Section 4.2. Another data representation point is transformation of variables which is considered in Section 4.3. Finally Section 4.4 studies registration, which is one more data object representation issue that is relevant to image and shape analysis, as well as to phase variation in Functional Data Analysis.

A general term that encompasses all of these issues is *data provenance.*

4.1 Marginal Distributions

Marginal distribution plots were introduced in Section 2.2.1, where Figures 2.5 and 2.6 illustrate how they can provide useful diagnostics. As noted in that section, the challenge of trying to visualize a large number of marginal distributions can be met by selecting a *representative* subset of the variables to actually look at. The idea of sorting on a one dimensional summary statistic (e.g. the mean as in Figures 2.5 and 2.6), is essentially that of Tukey's *scagnostics*, see Wilkinson et al [229, 230] for good overview and discussion. The difference is that scagnostics is about using numerical summaries (e.g. correlation) to find interesting scatterplots from a large collection, while in contrast these marginal distribution plots are about doing this for a large collection of one dimensional marginal distributions.

An important point is that many distributional summaries besides just the mean can be very useful. This point is made here using a chemo-metrics data set which also demonstrates the value of this type of visualization, together with appropriate adaptation, before meaningful analysis can be done. The example studied here comes from the area of drug discovery, or more precisely Quantitative Structure Activity Relationships as discussed in Cherkasov et al [41], with this particular set from Borysov et al [27]. There are $n = 262$ chemical compounds, that are represented by $d = 2489$ chemical descriptors. The primary goal is to distinguish *inactive* compounds shown as blue circles, from *active* ones shown as red plus signs.

A PCA scatterplot matrix, using the same format as Figures 2.10, 2.13 and 2.17, is shown in Figure 4.1. This view of the data is dominated by relatively few of the data objects. Almost all of the $n = 262$ data points are tightly clustered near the origin, which seems to be where any meaningful differences between the actives and inactives may be found. However, as indicated in Figures 2.17 and 2.18, there can be a large amount of interesting structure in data which is not apparent from merely looking at PC scores. There are many potential causes of such behavior. One of these, that is frequently worth checking, is the behavior of the marginal distributions. For example highly skewed marginal distributions (such as a log normal distribution) can frequently generate such data views.

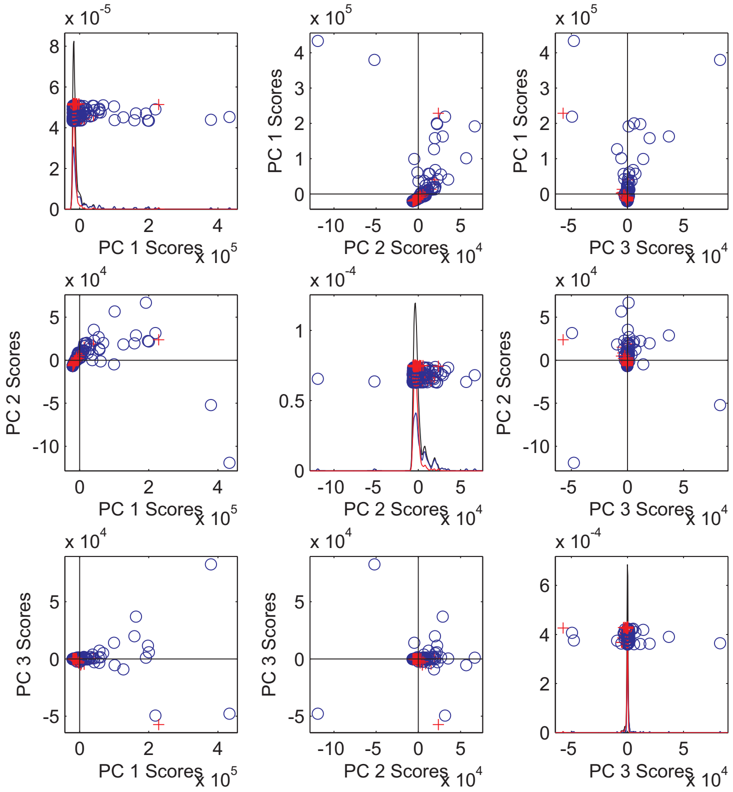

Figure 4.1: PCA scatterplot of raw drug discovery data. This view is driven by a few outliers and gives very poor separation of the active (red) and inactive (blue) compounds.

As noted in Section 2.2.1, while visualization of the marginal distributions is standard in careful small scale data analyses (good analysts will usually consider transformations of variables, etc.), it is less commonly done in Big Data contexts, perhaps because studying all the marginals is intractable. However, the marginal distribution plot approach demonstrated in Figures 2.5 and 2.6, does provide a scalable way to study these. The key idea is to replace visualization of all of the marginals (typically far too many to humanly comprehend) with instead looking only at a *representative subset*. As seen through the following examples, a major challenge for the data analyst is effective choice among many possible ways of determining useful representative variables. But a reasonable start is to consider sorting variables based on standard univariate summary statistics, and then to take either equally spaced representatives, or to focus more on one or both ends.

A number of such marginal distributions for this drug discovery data will now be presented to demonstrate the usefulness of this approach. A reasonable starting point is based on a sort of the sample means, with visualization of an equally spaced set of distributions. This is done for the drug discovery data in Figure 4.2. The upper left panel shows the sorted means as a blue curve. Note that most of the means appear to be around 0 with a few relatively huge values on the far right. Because PCA finds directions of maximal variation, these very few variables are potential drivers of the unfortunate population structure observed in Figure 4.1, and there may yet be useful population structure that will emerge when those variables are properly handled. Note also the rather small downturn in the blue mean curve on the far left.

Eight marginal distributions (that number is chosen merely for convenience of plotting) corresponding to the vertical dashed lines are shown in the remaining panels. These show huge heterogeneity in the variables present in this data set. The first few show no variation at all, i.e. all values are exactly the same. In the first marginal distribution, they are all equal to -999 (this perhaps surprising value is explained below). For the next three variables, all values are 0. The center right variable is all 0's except for a single 1. The variables on the bottom row are also wildly different from each other, with a discrete distribution on the left, and a clearly skewed distribution, with values that are four orders of magnitude larger, on the right.

That value of -999 is sometimes used to code missing values (in fact this is the case here), perhaps with the idea that it is so different from all the others that it would be easily noticed and properly dealt with during an analysis. However, that idea failed in this data set, because there are some variables that are so much bigger in magnitude (which may have added to this combined data set by a different analyst). In particular, because -999 is a number, it would be easy to make the big mistake of treating them as data. This type of effect easily arises in Big Data contexts where there is a lot of merging of diverse data sets in contexts where no individual has a complete understanding of all aspects. This Marginal Distribution type of visualization is often effective in discovering such anomalies.

Figure 4.2 makes it clear that a number variables with no variation (and thus no information about active vs. inactive compounds) can be deleted from the data set with no loss of information. It also indicates that careful attention should be paid to the missing values, coded as -999, and finally that both the relative magnitude and skewness of other variables will need careful consideration.

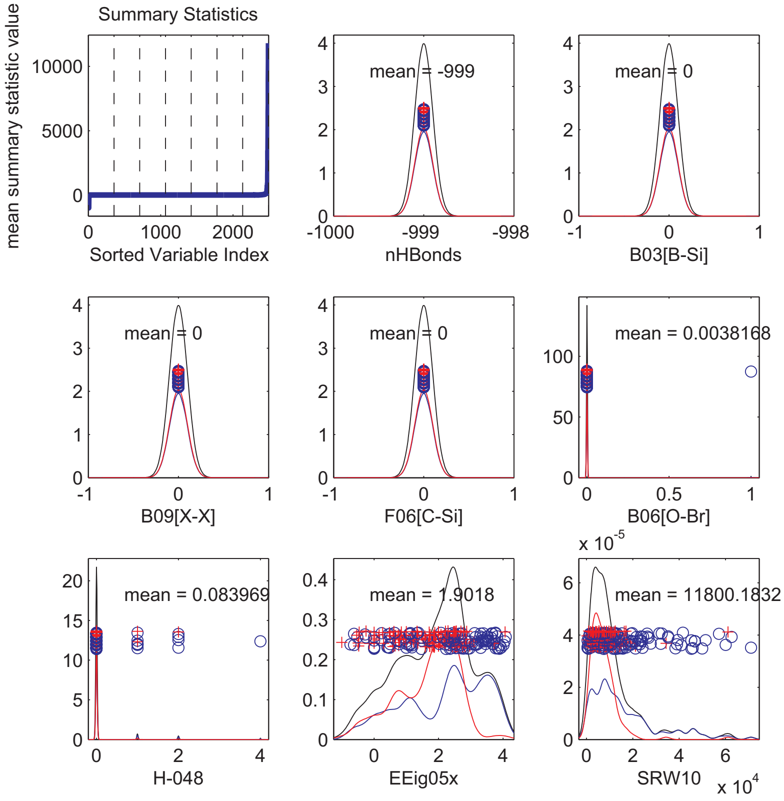

Figure 4.2: Marginal distribution plot view of the drug discovery data, sorted on sample means. Shows great diversity of variation over variables. Many have no variation and a very few are orders of magnitude bigger.

Figure 4.3 shows the marginal distributions, this time sorted on the sample Standard Deviation (SD). This summary statistic gives a more clear focus on those variables with no variation, revealing that they are nearly half of the $d = 2489$ variables. This also sustains several important lessons from Figure 4.2 such as there are a few variables that are several orders of magnitude larger, and the distributional shapes are very heterogeneous.

An aspect not yet discussed is that the text below each distributional subplot is the name of the variable (i.e. feature). This can be very useful for identification of important features in data. Note that the variable with largest SD, labeled ww, is different from the variable SRW10, which has the largest mean (bottom right in Figure 4.2). In addition the variable nHBonds, which was seen to have the smallest mean (with each value being -999) in Figure 4.2, does not appear in Figure 4.3. The reason is that in the latter graphics, there are a large number of variables with SD = 0, and the large number of ties are broken by simply using the original variable ordering.

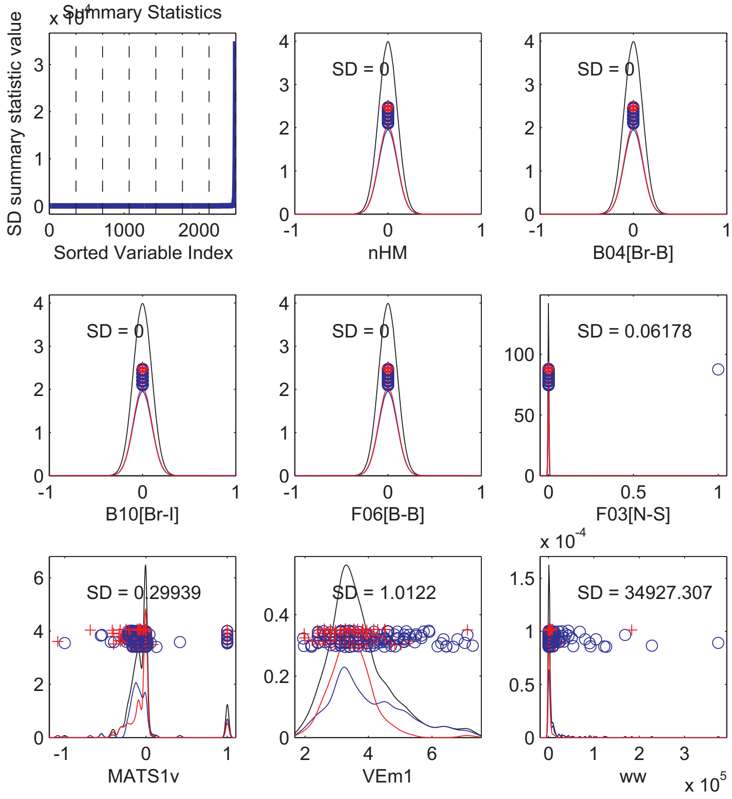

Figure 4.3: Standard deviation sorted marginal distribution view of the drug discovery data. Reveals many variables with no variation.

As seen in Figures 4.2 and 4.3, much can be learned from marginal distribution plots using equally spaced (with respect to the summary statistic) representative distributions. However, in some situations it is very useful to focus in on particular parts of the collection, often those with the smallest and/or the biggest summary measures. For example, with the goal of taking a more careful look into the missing values coded by -999 in the drug discovery data, Figure 4.4 again considers distributions sorted by mean, but now shows the 8 smallest mean values. This is reflected by all eight of the vertical dashed lines appearing on the far right. It reveals that there are six variables that are all missing (i.e. all values are -999) and at least two more with some missings.

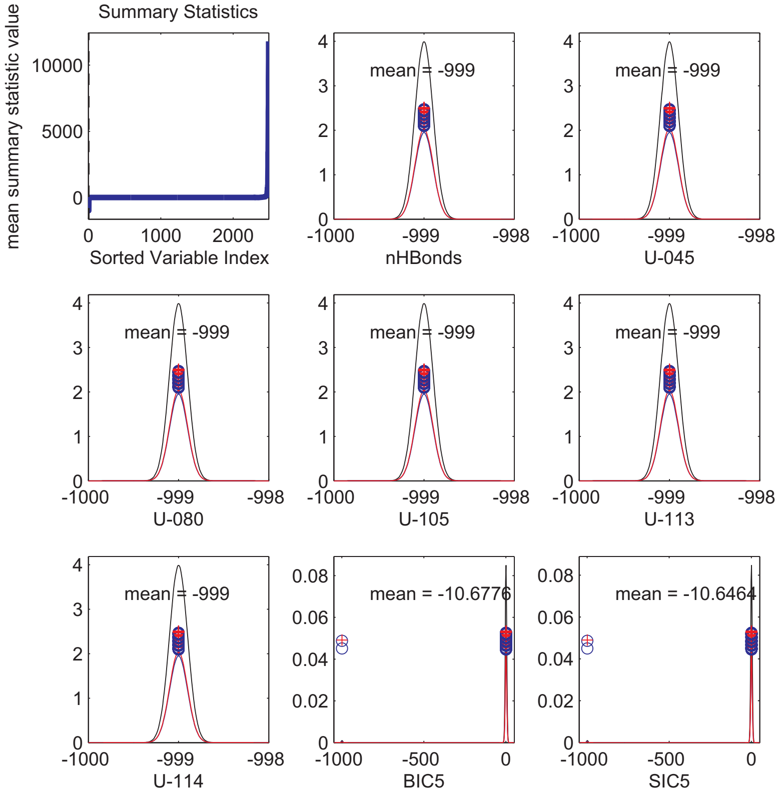

Figure 4.4: Drug discovery data explored using the eight smallest mean marginal distributions. Shows six variables are all -999 and others have a few -999s.

Figure 4.5 highlights a different set of variables, this time by sorting on the minimum value of each marginal distribution. In this situation, this view is less informative than those above, again because of how ties are handled in the sorting. E.g. instead of seeing all values at -999 as in several of the above plots, the shown variable, IC4, only takes on -999 once (and does not appear in Figure 4.4). The next 6 variables all have a minimum of 0, so they appear in an arbitrary order, which is not particularly useful for understanding specifics in this data set, although it does happen to show the wide heterogeneity present here.

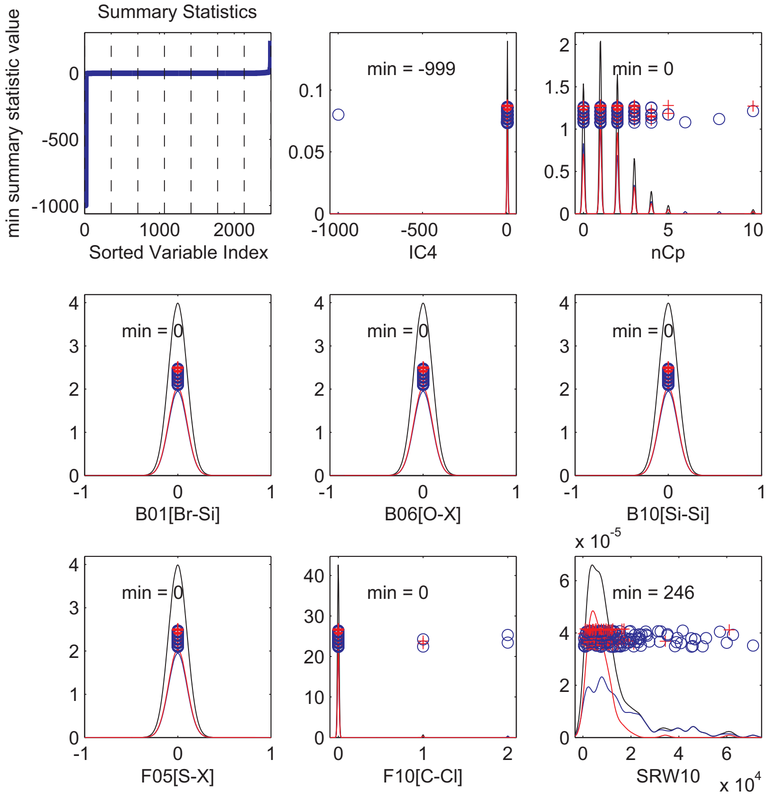

Figure 4.5: Minimum value sorted marginal distributions for the drug discovery data. Again shows wide range of variation, but less informative because of arbitrary handling of many ties.

Based on the above insights, straightforward calculation verified that there were 1315 variables with no variation, and 16 that had at least one value of -999. Removal of these variables resulted in a cleaned data set of $d = 1164$ variables, which is further studied in the following. In other situations, much more care may be needed to deal with missing values. A usually important issue is whether missing values should be deleted (which is sensible here, since there are so few missings), or else imputed in one of various ways. See Gelman and Hill [81] and Enders [65] for good overview of many possible approaches to data imputation. This type of data object choice can be very challenging, nad is often most effectively done in consultation with the data provider.

A PCA view of the cleaned data is not shown here because it looks exactly like Figure 4.1 above. The reason is that both views are driven by the large magnitude variables, so of course eliminating features with no variation changes nothing. Also removing the -999s has no visible impact because the larger variables are so much larger, as revealed by a careful look at the axis labels in

Figure 4.1.

Figure 4.6 studies SD for this cleaned data set. This shows that in addition to a few variables with extremely large variation, there are also some with extremely small variation and seems to indicate the presence of some binary variables. This very wide range of variation suggests that some type of normalization, say rescaling each variable by its standard deviation, will give a much different result, perhaps revealing other types of structure in the data, as discussed in Section 4.2. Note that the lower right (maximal SD) variable, ww, is the same as the lower right variable in Figure 4.3. The others are all different, because they are equally spaced in a smaller set of variables.

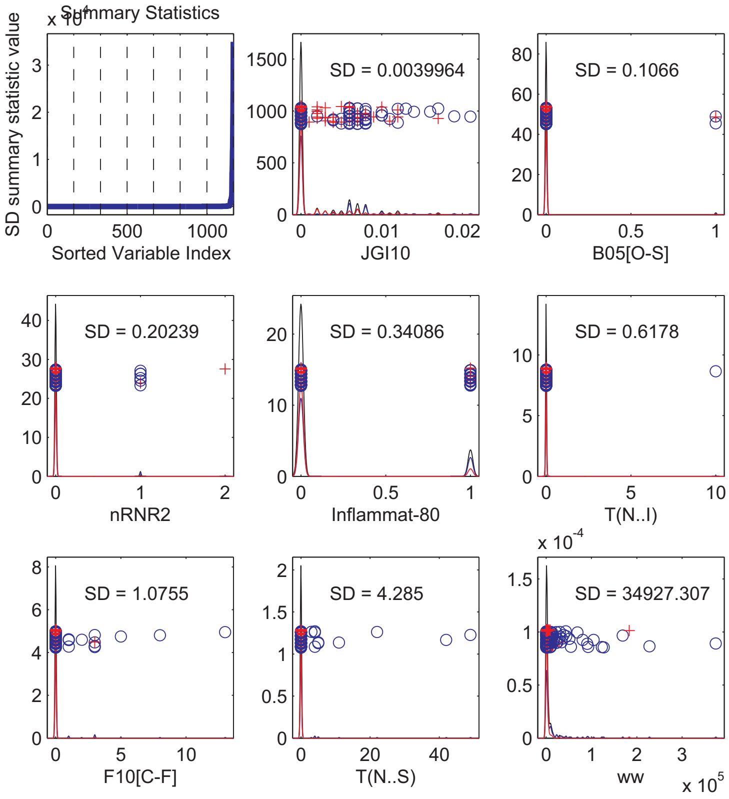

Figure 4.6: Drug discovery data, after cleaning, viewed with marginal distribution plots sorted on standard deviation. Shows widely differing variation.

Several of the above plots, e.g. those in the lower right panels of Figures 4.3, 4.5 and 4.6, suggest that skewness is a serious issue for at least some of these variables. This is explicitly studied in Figure 4.7 by sorting the variables this time on skewness. The upper left blue curve indicates more right skewness than left skewness, although both types are present. It also once again shows a strong presence of binary variables, taking on only the values 0 and 1.

Note in addition that the variables with strongest skewness (typified here by F05|O-Ci| in the lower right panel) take on just a single value of 1, with all other $n - 1 = 261$ values being 0. The large flat spot on the upper right of the blue curve shown in the upper left panel indicates that in fact there are some hundreds of such variables. Conventional wisdom is that such variables contain little useful information and thus should also be eliminated from consideration. However, the large number of them suggested this issue may be worth a second look. This was done in Borysov et al [27] who showed that in this case, there actually is useful information in these variables and developed methods for incorporating them in data analyses.

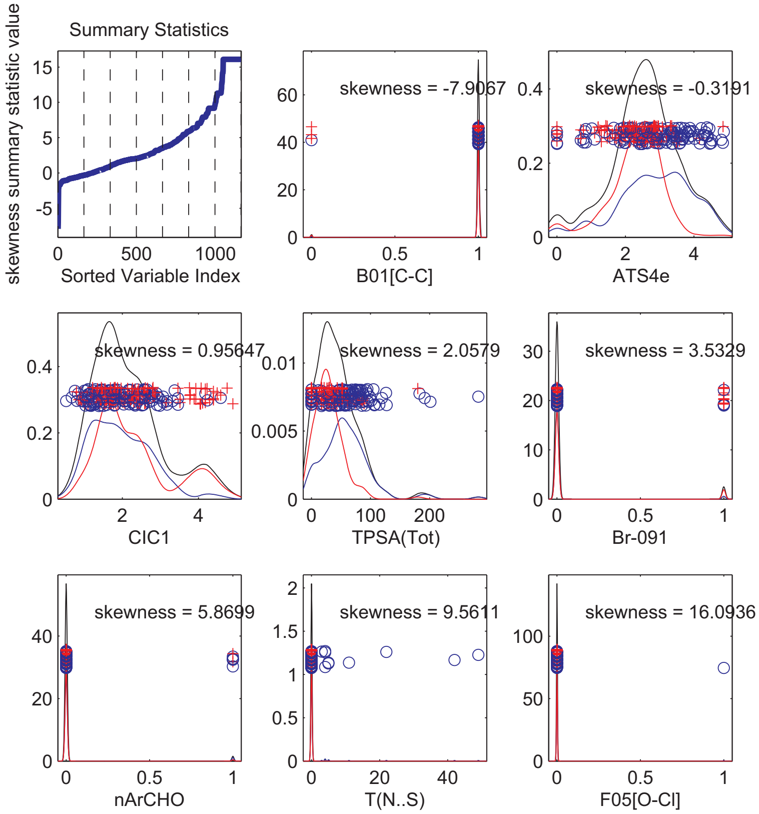

Figure 4.7: Marginal distribution view sorted on skewness, for the clean drug discovery data. Shows many variables with strong skewness, including binary distributions.

When there is a combination of discrete and continuous variables in a data set, as suggested in the above analyses, there are other sortings of marginal distributions that are also quite useful. One of these is the number of unique values for each variable, as shown in Figure 4.8. This clearly highlights binary variables (which have only two unique values) by putting them first, with the blue curve in the upper left panel showing nearly 500 such. In addition to a number of clearly quite discrete variables, there are variables such as ZM1 in the middle left that appear to be continuous and yet contain a large number of exact replicates. The upper left part of the blue curve reveals that there are very few truly continuous variables, for which the number of uniques is $n = 262$, as for AEigp in the lower right panel.

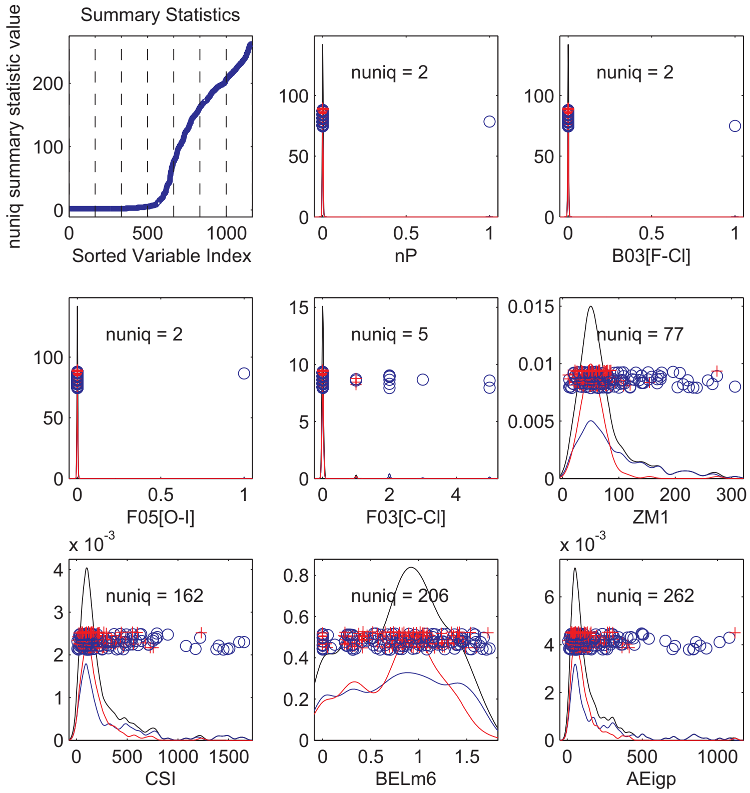

Figure 4.8: Clean marginal distributions of drug discovery data, sorted on number of uniques. Shows wide range, from binary (nearly 500 such) to completely continuous variables.

A different way to study discreteness versus continuity of variables is to sort on *number most frequent* as in Figure 4.9. This measure counts the number of times each value appears in the distribution and reports the largest of those. This number is 1 for continuous variables, because all values are different. Note that the shown continuous variable, TIE, is different from the continuous variable AEgip shown in the lower right panel of Figure 4.8. Hence there are at least two truly continuous variables (namely AEgip and TIE), which appear in different orders in the two sorts. The question of how many continuous variables are presented could be more carefully studied by looking at the smallest variables in this sort, or the largest number of uniques using the ordering in Figure 4.8. In this case there are only 9 variables which are truly continuous in the sense of having $nuniq = 262$. In this view, the binary variables appear on the right, in the order of how many times the majority values appear. This is another way of seeing that hundreds of variables have just a very few non-zero values. Because few statistical methods are designed to handle such a mix of continuous and discrete methods, the need to use methods such as these visualizations is becoming increasingly important in Big Data contexts.

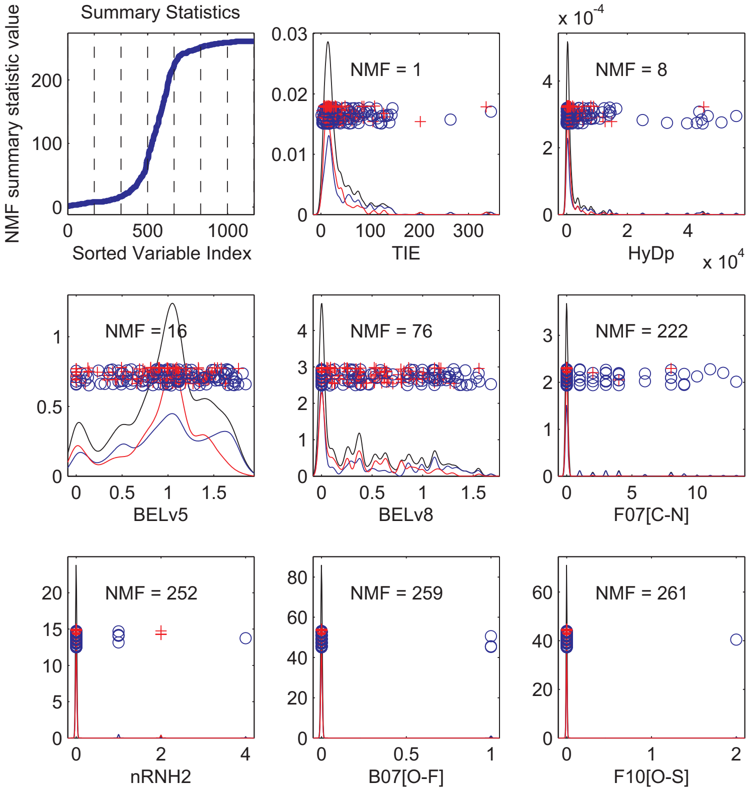

Figure 4.9: Drug discovery data, after cleaning. MargDistPlot sorted on number most frequent. Another way of contrasting discrete and continuous variables, again showing a wide range of both types.

Given the large number of binary variables, a question arises as to how much information for separating active versus inactive compounds is present in those variables only. To study this, the data are reduced to only the $d = 364$ binary variables, with the resulting PCA scatterplot shown in Figure 4.10 This shows some perhaps surprisingly rich structure in this data with most of the active cases focused in just a few regions and larger regions with essentially no active cases. As noted above, Borysov et al [27] gives a more detailed analysis of this binary data set.

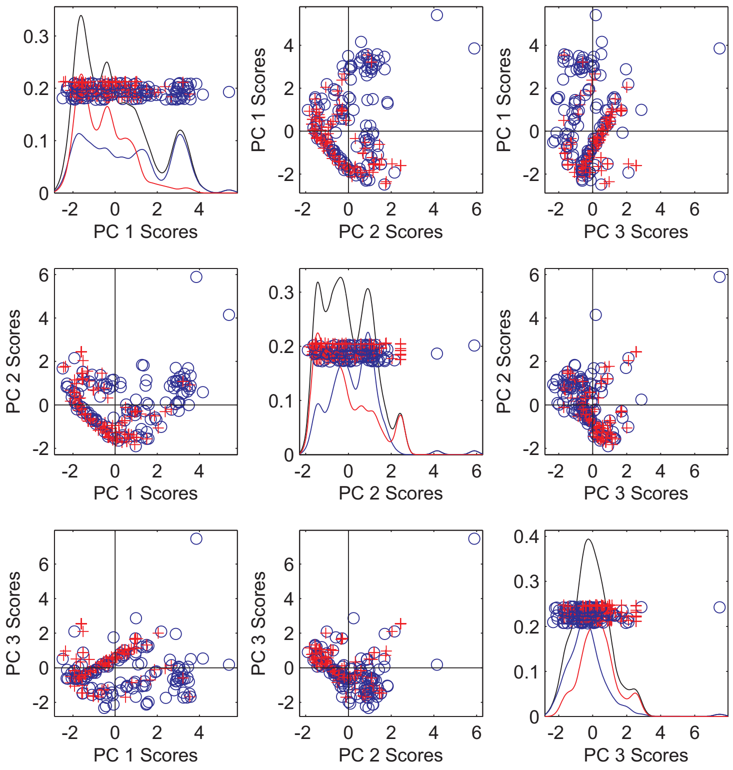

Figure 4.10: Drug discovery data, after cleaning. PCA based on binary variables only. Shows these variables contain rich red-blue separation information. [28]

This section has shown that marginal distribution plots can discover many important aspects of data sets. In many situations, these can be essential in indicating strategies such as variable deletion, scaling (as studied in Section 4.2) and/or transformation (see Section 4.3) that can be very important to arriving at an effective data analysis. A few commonly used summary statistics have been demonstrated here, but many others can be equally revealing in other situations. For example small values of kurtosis can easily find variables with strong bimodal structure. Data sets with endemic outliers can adversely affect conventional moment based summaries, in which case robust summaries, such as the robust skewness-like measure of Bowley [28] and the robust kurtosis-like measure of Ruppert [182], can be very useful. An [7] developed some L-statistics based methods that are quite effective at finding important genes in the context of cancer research.

4.2 Standardization

As illustrated in Figures 2.15 and 2.16, when some variables are orders of magnitude larger than others, the larger ones can completely dominate many types of statistical analysis. As done there, a common strategy is to standardize each variable, subtracting the mean and dividing by the SD. The effect of this on the cleaned drug discovery data from Section 4.1 is demonstrated in Figure 4.11. Unlike the outlier driven PCA shown in Figure 4.1, this view shows much more in the way of interesting relationships between the data objects, i.e. the chemical compounds. In particular, it is clear that there are now very complex (e.g. highly non-linear) relationships between the active (red pluses) and inactive (blue circles) compounds, which is why drug discovery has been a challenging problem over the years. Note for example an indication of small regions where interesting comparisons can be made, which motivates the idea of *activity cliffs*, (regions of abrupt transition between classes) as studied in Maggiorra [135].

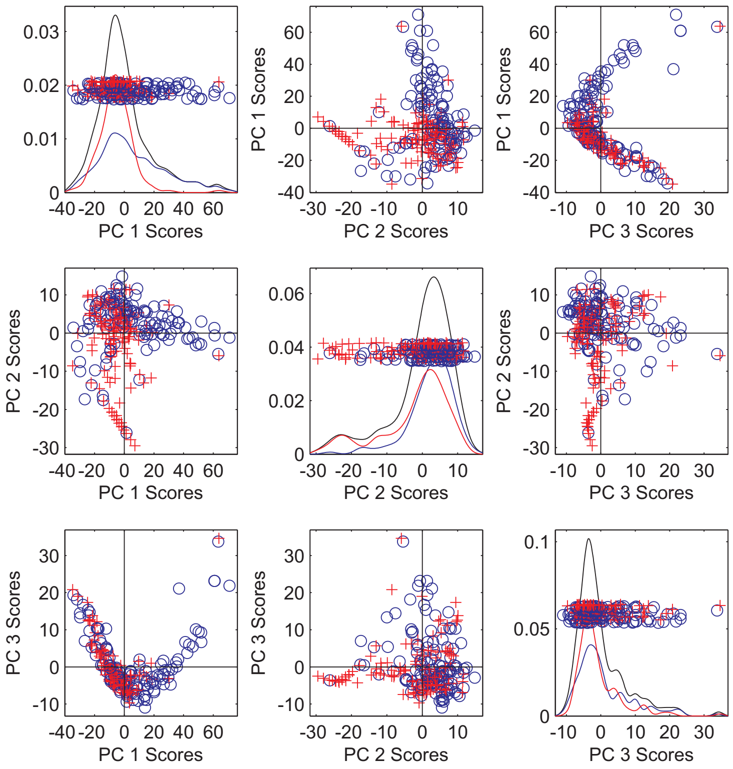

Figure 4.11: Drug discovery data, after cleaning. PCA on non-binary variables only. Shows these also contain red-blue separation information.

There are many other aspects of data normalization that should be kept in mind as needed. For example, while standardization of the variables (i.e. of each row of the data matrix) can be very sensible in some cases as illustrated in Figures 2.15 and 2.16, column standardization can be more useful in others. A canonical example of this is genetic molecular measurements, which are based on amplification of DNA or RNA in a way that is not easy to calibrate, resulting in columns of the data matrix which differ by scale factors. For gene expression studies, for example in Hoadley et al [102], this is commonly handled by scaling each column appropriately. While averages could be used for this, that would not allow for expected differing overall expression across cases, so instead normalization to achieve a common third quartile is used.

In other situations projection of each data (column) vector to the unit sphere, $S^d = \left\{x \in \mathbb{R}^d : \|x\|_2 = 1\right\}$ (where the L^2 norm is $\|x\|_2 = \sqrt{\sum_{j=1}^d x_j^2}$) can be appropriate. This projection is computed as $\frac{x}{\|x\|_2}$. It is useful in situations where vector length contains only nuisance variation, and the interesting variation is in the angles between vectors. In such cases, another option is projection to the unit simplex, as discussed in Section 3.2. For virus detection using DNA data, Xiong et al [233] found that projection to the simplex gave better results than projecting to the sphere.

4.3 Transformation

As seen above relative magnitude of variables is an important consideration. Similarly distributional shape of the marginal distributions can also have a major effect as seen in Figures 2.5 and 2.6. Another example that illustrates this point is shown in Figure 4.12. This data set was a product of the Cancer Genome Atlas, Weinstein et al [228]. In particular the data were preprocessed by Hoadley et al [102], who explored many contrasts between 12 cancer types, based on a variety of measurements. . Here we focus only on gene expression and restrict the cancer types to Ovarian Cancer, which is labeled as OV in TCGA notation and shown as purple circles, and to Uterine Cancer, labeled UCEC and indicated using green plus signs. Furthermore only the 1000 most variable genes, among genes having no missing values, are considered. The raw data are counts indicating gene expression, measured using the RNAseq technology, see Wang et al [225]. The full data set (with a few hundred cases of each type) shows a strongly statistically significant difference between these two cancer types for almost any type of analysis, so for good contrast between statistical methods, randomly chosen subsets of size $n_1 = n_2 = 30$ cancer patients are analyzed here.

The top row of Figure 4.12 shows PCA scatterplot views of the data. Unlike the scatterplot matrices shown above, e.g. in Figures 4.1, 4.10 and 4.11, here each plot shows only PC2 vs. PC1 scores scatterplot (often the left plot in the second row in matrix views). The upper left panel of Figure 4.12 studies the distribution of raw counts. Note that PC1 is dominated by a single very large case (about an order of magnitude bigger than all others). PC2 is driven by a handful of other cases, but still only a relatively few. While one might hope to see a large difference between the OV and UCEC cases, if it can be seen in this scatterplot, it can only be in the lower left part of the plot, but is very hard to perceive due to over-plotting. For a closer view of potential class differences, the top center panel shows a zoomed in (on the lower left corner) version of that plot. This makes it even more clear that this data set suffers from strong skewness (which can also be easily seen using the Marginal Distribution views described in Section 4.1), with essentially no OV-UCEC difference visible. This does not mean that there is no difference, only that it does not appear in the 2 dimensional subspace of the first 2 principal components. For such strongly skewed data, a log transformation of each variable is often very useful, as it tends to strongly reduce the influence of data points that are orders of magnitude larger than the others. The top right panel shows the result of the $\log_2$ transformation applied to each variable. That transformation is usual in this field, where the doubling interpretation of that log base is commonly desired. Note that these two modes of variation highlight a clear and strong difference between the cancer types, appearing as mostly the dominant mode of variation (i.e. the PC 1 Scores).

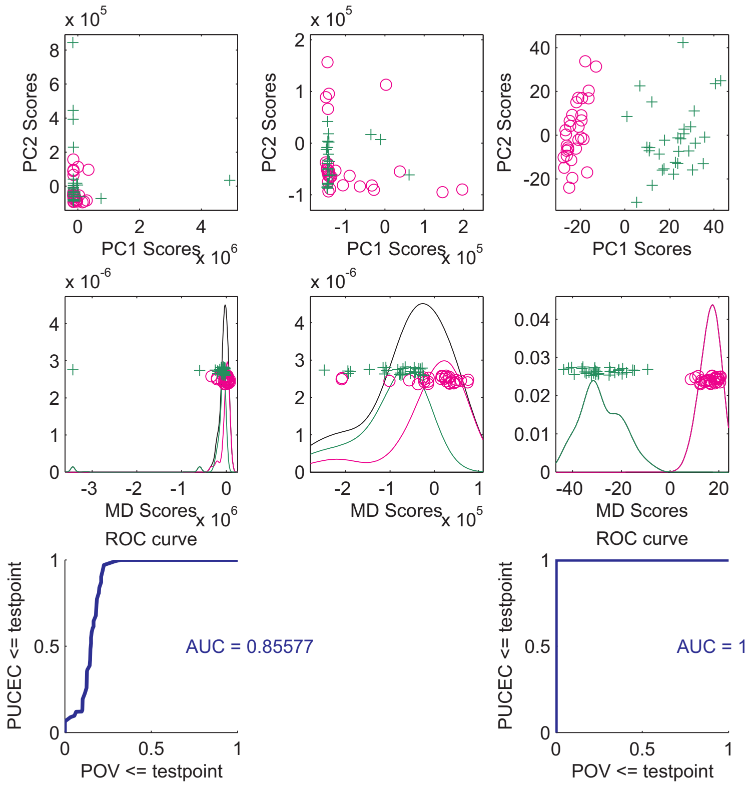

Figure 4.12: Contrast of Ovarian (purple circles) and Uterine (green plus signs) Cancer gene expression. Top row shows PC1 vs. 2 scores, for raw count data (left, with zoomed in version, center) and $\log_2$ transformed data on the right. Middle row shows corresponding MD projected distributions. Bottom row is ROC curve analyses. Shows log transformed analysis gives much better contrast between cancer types in all ways.

Table 4.1 provides another way of seeing that the $\log_2$ transformation provides a much better scale on which to analyze this data set. In particular, the outlier in the upper left panel of Figure 4.12 is seen to dominate the raw data analysis, with the first PC containing 88% of the total variation about the mean. In contrast, on the $\log_2$ scale the first PC explains 25% while the second explains 10%, which are much more reasonable as there are a large number of diverse biological processes whose presence should be reasonably represented in gene expression measurements.

The middle row of Figure 4.12 more directly targets the OV vs. UCEC contrast by showing the univariate distribution of projections onto the *Mean Difference* (MD) direction vector, which is just the normalized (to have length 1) difference between the sample means. This direction vector is also called the *centroid classifier*, see Tibshirani et al [210]. The display format is the same as

	PC1	PC2
Raw Data	88.2	4.7
$\log_2$ Data	24.8	9.9

Table 4.1: Percent of sums of squares about the mean explained by the first two principal components, for the raw gene expression (top row) and the $\log_2$ transofrmed version of the same data. Shows that log transformation gives more sensible distribution of variation in the data.

used several times above, e.g. the Marginal Distribution plots in Section 4.1, using symbols whose x-coordinate reflects the value with y-coordinates simply providing visual separation. The black curve is a kernel density estimate, with the colored curves representing proportional sub-densities for each of the two data types. The middle left panel shows this distribution for the raw counts data. As in the top left panel, the difference is not easy to discern because the view is again dominated by a single large outlier, which obscures how well separated the two groups are. The zoomed view in the center panel shows that actually the MD direction provides decent separation of the classes, with the green plus UCEC cases tending to lie more to the left of the OV purple circles, although there is substantial overlap. This overlap is quantified using the Receiver Operating Characteristic (ROC) curve of Hanley and McNeil [93], in the lower left panel. This curve is generated by sliding a cutoff point along the horizontal axis of the left middle panel, and for each such point displaying the proportion of UCEC (green) points that are smaller on the vertical axis versus the proportion of smaller OV (purple) points on the horizontal axis. The fact that more UCEC points lie to the left is reflected by the curve moving fairly steeply upwards. Once the cutoff point includes all UCEC points the curve remains at height one. The fact that this curve lies mostly towards the upper right of this plot shows that the two populations are relatively well separated. A simple numerical summary of ROC behavior is the *Area Under the Curve* (AUC), which in this case is 0.86.

The middle right panel studies the MD projections for the $\log_2$ transformed version of the data. Given the obvious group separation in the PC 1-2 scatterplot in the upper right panel, it is not surprising that there is a very strong separation between UCEC and OV that is apparent in this projection. The strong visual impression is confirmed in the lower right panel by the ROC plot in the lower right following first the vertical axis, then the top horizontal line, resulting in an AUC of 1.

A related contrast between the raw count data and the $\log_2$ transformed version is provided using the DiProPerm confirmatory method, described in Section 12.1. That hypothesis test for exploring differences between the UCEC and OV cases, using the mean difference direction and mean different summary statistics, for the raw counts gave a non-significant p-value of 0.24, while the log2 counts gave a strongly significant p-value $\ll 10^{-4}$. This is another way of seeing that analyzing this data on the $\log_2$ scale is very well worthwhile.

An important variation of log transformation is the *shifted log transformation*, of the form $\log(\cdot - c)$, where the data are shifted by a constant amount c before application of the logarithm. This is useful both for data which take on 0 or negative values, and in the case of $c < 0$ is also useful as a typically less stringent version of the log transformation, that is useful for data with more mild skewness. This works in the same way as the skewness of the log normal distribution is controlled by the mean parameter of the underlying normal distribution. Good automatic choice of the shifted log transformation has been developed by Feng, Hannig and Marron [69].

Another appealing and widely used family of transformations is the Box-Cox family

$$f(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & for\ \lambda \neq 0 \\ \log x & for\ \lambda = 0, \end{cases}$$

proposed by Box and Cox [29]. A careful calculation of the limit as $\lambda \to 0$ shows that this is a continuous function of the tuning parameter λ which provides a different way of adapting to skewness in data. One more important general family of transformations is described in Johnson [114].

4.4 Registration

Registration, i.e. alignment issues, are often quite important in many types of OODA. This point is illustrated using an FDA (curves as data objects) toy example in Figure 4.13, which is similar to Figure 1.8. The raw data are shown in the left hand panel. Each curve has two peaks, but there is substantial variation in both locations and heights of the peaks. In contrast to Figure 1.8, this time the curves are color coded using the height of the left peak, with a rainbow color scheme ranging from magenta (tallest) through green and yellow to red (shortest) with the goal of highlighting the amplitude variation in this case. The varying locations of the peak creates challenges for standard statistical analysis (which ignores the strong phase variation). For example, the (point-wise) mean curve, shown as a thick black dashed line, is not at all representative of the population. In particular, its peaks are substantially lower than any peak in the data set, and the left peak actually appears as two modes. It will seen in Chapter 8 that PCA (e.g. as in Figures 1.4, 1.5, 2.4, 2.7 and 2.9) of this set of curves also provide very poor low rank representations of the data.

The right panel of Figure 4.13 shows the results of a Fisher Rao registration of these curves as described in Chapter 8. The heights of each curve are the same in both panels, but in the right panel the horizontal axis for each curve has been appropriately warped to make the curves align very well. Note that the mean of this set of curves, again shown as a thick black dashed curve is now a quite sensible notion of center, as it lies clearly in the middle of the data set. As seen in Chapter 8, PCA of this set of curves provides a quite intuitive and much more compact representation (in particular, requiring only a single component) of the data set.

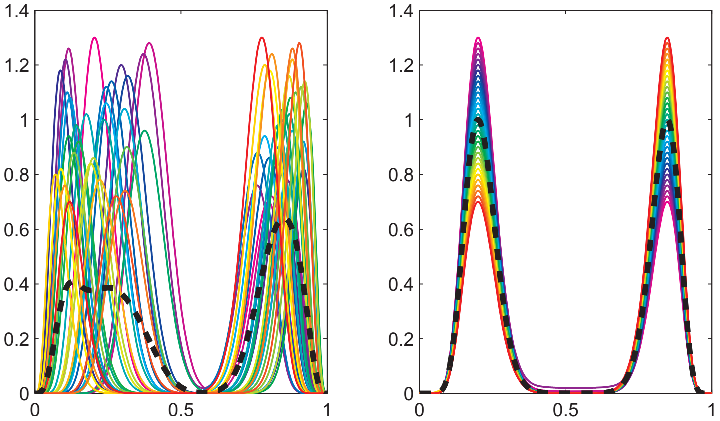

Figure 4.13: Toy example, of functional data, each curve having two peaks, illustrating usefulness of curve registration. Left panel shows original curves, using rainbow color scheme on height of left hand peak to highlight amplitude variation. Right panel shows the result of a careful curve alignment. Dashed curve in each case is the sample mean, which is much more representative of the data curves after alignment.

For quite similar reasons, registration is also critical in image related tasks. For example, the faces data in Figure 1.20, little attempt was made during the initial photographs to ensure that each face was in the same place in each picture, which presented a challenge that is quite similar in spirit to what is seen in the left panel of Figure 4.13. In particular, various facial features (such as eyes, nose, mouth) were seriously misaligned across images, resulting in a far fuzzier analog of Figure 1.21 when the analysis was based on the unaligned data. As noted in Section 1.6, great improvement was made by a simple affine transformation based on landmarks at the center of each eye and the mouth.

Generally in image analysis this type of consideration motivates the study of shapes as data objects. That field is discussed in more detail in Chapter 7.

Chapter 5

Distance Based Methods

This chapter is about OODA methods that are based only on distances between data objects. An advantage of such approaches is that they are quite broadly useful, which can be important in exotic data spaces such as manifolds, or tree/graph spaces, where simply finding analytic methods can be challenging. In particular, the only structure needed on the object space for such methods is presence of a metric. As noted in Sections 2.1 and 2.4, a quick and dirty default approach to OODA is to first find a metric, and then simply do a distance based analysis. While that is useful in some situations, there are others where a more careful use of particular data space structure can result in much improved analyses, as discussed in Chapters 7, 8 and 9.

In this chapter, the symbol δ will be used to denote *metrics*, i.e. distance functions. Given a set of data objects

$$\{X_i : i = 1, \cdots, n\},$$

and a distance δ, the corresponding $n \times n$ symmetric *distance matrix* is

$$D = \begin{bmatrix} 0 & \delta(X_1, X_2) & \cdots & \delta(X_1, X_n) \\ \delta(X_2, X_1) & 0 & & \vdots \\ \vdots & & \ddots & \delta(X_{n-1}, X_n) \\ \delta(X_n, X_1) & \cdots & \delta(X_n, X_{n-1}) & 0 \end{bmatrix}. \tag{5.1}$$

Working with data objects in this type of format is rather common in the machine learning literature, see Cristianini and Shawe-Taylor [46], Schölkopf and Smola [189] and Shawe-Taylor and Cristianini [192] for good overview. Indeed a central idea in that area is called the *kernel trick*, where one works with a matrix of inner products (closely related to the usual distance in Euclidean spaces), with the important goal of large computational benefits.

Distance based notions of *center* are discussed in Section 5.1. Methods for understanding variation about the center, in the spirit of PCA, using only distances are explored in Section 5.2. Clustering is another set of data analytic methods, that are frequently based only on distances between data objects. Cluster analysis is discussed in Chapter 11.

A critical aspect of all of these approaches is that choice of distance has a major impact on the results of the analysis. For a particularly dramatic example consider the *discrete metric*

$$\delta_D(x,y) = \begin{cases} 0 & \text{for } x = y \\ 1 & \text{for } x \neq y \end{cases} .$$

This exists for any space, but is useless for OODA analyses, because it contains no information about how the data objects relate to each other. Section 7.2 contains some perhaps surprising examples on the impact of metric choice in the context of covariance matrices as data objects.

Generally good choice of metric is very situation dependent. For example, in the case of Euclidean data objects, say $x, y \in \mathbb{R}^d$, the standard Euclidean L^2 distance

$$\delta_2(x,y) = \|x-y\|_2 = \left(\sum_{j=1}^{d} (x_j - y_j)^2\right)^{1/2} \tag{5.2}$$

is often very useful. However, when it makes sense to think in terms of polar coordinates, and the important variation happens in the angular direction with mostly distracting noise in the radial direction, a more useful metric can be the cosine distance

$$\delta_C(x,y) = 1 - \frac{2}{\pi}\cos^{-1}\left(\frac{x^t y}{\|x\|_2 \|y\|_2}\right),$$

which is driven only by the angle between x and y projected onto the unit sphere. When outliers are a major concern, the L^1norm

$$\delta_1(x,y) = \|x-y\|_1 = \sum_{j=1}^{d} |x_j - y_j| \tag{5.3}$$

is useful because of its natural tendency to down-weight their influence.

5.1 Fréchet mean

An important notion of *center* of a set of data objects $\{X_1, \cdots, X_n\}$ in an arbitrary metric space S (with distance δ) is the *Fréchet mean*,

$$\arg\min_{x \in S} \sum_{i=1}^{n} \delta(x, X_i)^2, \tag{5.4}$$

from Fréchet [79]. This is a direct generalization of the standard sample mean $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ in Euclidean space $\mathbb{R}^d$, because it is straightforward to show that $\bar{X}$ is the solution of (5.4) in the case of Euclidean distance (5.2). Insight as to how the Fréchet mean works is given in Figure 5.1.

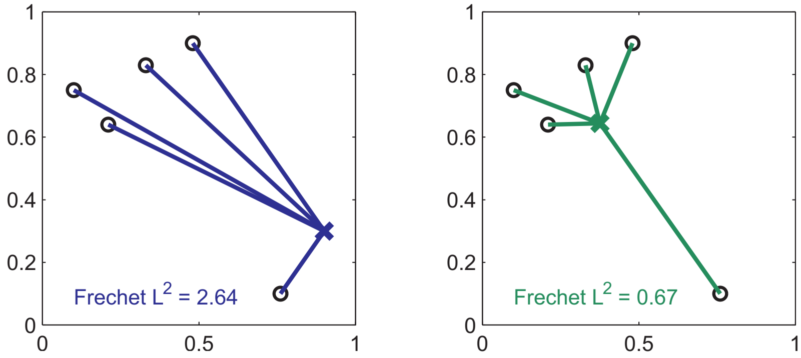

Figure 5.1: Toy two dimensional example illustrating the Fréchet mean. Left panel shows a (far from optimal) candidate point as a blue x, with line segments representing the distance to each data point. Right panel shows the optimal Fréchet mean as a green x, with again line segments showing this point is much closer to the data points than the blue candidate on the left. These issues are quantitated in each panel using the Fréchet sum of squared L^2 distances.

A toy data set of $n = 6$ data points in $\mathbb{R}^2$ are shown as black circles in both panels. The blue x in the left panel is a candidate choice of mean. The Fréchet criterion in (5.4) is the sum of the squared lengths of the blue line segments, whose numerical value for this candidate is seen to be 2.64. The operation of solving the Fréchet optimization problem can be thought of as moving the candidate point to minimize the Fréchet criterion. The solution is shown as the green x in the right panel (actually just the sample mean $\bar{X}$ in this special case), based on an overall shorter collection of line segments and the smaller criterion value of 0.67.

Depending on the choice of metric, many familiar notions of center can be viewed as Fréchet means. For example on the real line $\mathbb{R}^1$, the median can be written in the form (5.4), by taking $\delta = \delta_2^{1/2}$ (note that it is straightforward to show that the square root of any metric is again a metric). Furthermore on $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$ the geometric and harmonic means are Fréchet means, based on $\delta(x, y) = |\log x - \log y|$ and $\delta(x, y) = \left|x^{-1} - y^{-1}\right|$ respectively.

An important variation is the *Fréchet median*,

$$\arg\min_x \sum_{i=1}^{n} \delta(x, X_i),$$

which differs from the Fréchet mean only by replacing the power of 2 by 1. This and other variations of the Fréchet mean have been extensively studied in the field of robust statistics, see e.g. Hampel et al [92], Huber and Ronchetti [105] and Staudte and Sheather [201], because both the median, and also distances δ which down-weight points farther away, have good properties in terms of reduced sensitivity to outliers. See Fletcher et al [77] for interesting applications of this approach in the context of Riemannian manifolds.

Direct comparison between the Fréchet mean and median is provided in Figure 5.2. This example features the same toy data set as in Figure 5.1, but this time the center-point is calculated as the Fréchet median based on δ_2. Again a poor candidate point is shown as the blue x, with the optimal solution shown in green. The key difference is that the sum of lengths of the line segments (not their squares) make up the Fréchet sums that are shown.

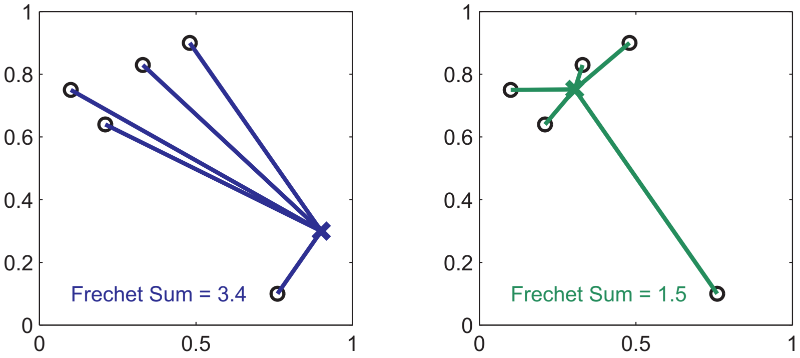

Figure 5.2: Toy example illustrating the Fréchet median, based on the distance δ_2, for the same data set as in Figure 5.1. This estimate is more robust in the sense of reduced sensitivity to the outlying point in the data set.

First note the substantial difference between the Fréchet mean and median in this case. In particular, the improved robustness property of the Fréchet median can be seen in terms of the green x being in the middle of the convex hull of the 4 nearest points in Figure 5.2, while the outlier is far enough away to pull it outside in Figure 5.1, because it has much stronger influence when the distances are squared.

The impact of metric choice is again illustrated in Figure 5.3. This is also the same toy data as in Figures 5.1 and 5.2, but now the distance used in the Fréchet median is the L^1 distance δ_1. Because this distance is the sum of absolute distances in each coordinate direction, each is now represented as a horizontal and a vertical line.

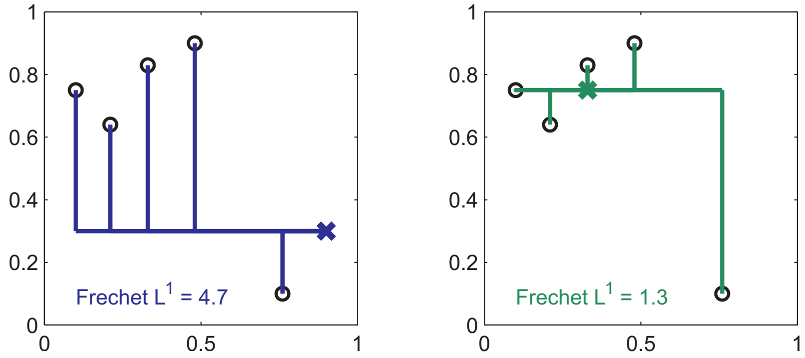

Figure 5.3: Study of the impact of metric choice using the same toy data from Figures 5.1 and 5.2. Center is again estimated using the Fréchet median, but now with δ_1. This version of center also demonstrates reduced influence of outliers.

This variation of the Fréchet median reveals some perhaps surprising differences with the δ_2 version. In particular, the solution here is the point-wise median (i.e. this can be computed by simply taking the median of each coordinate of the data vectors), which is not true for the δ_2 version of the Fréchet median. Thus the δ_1 variation is easier to compute, while an iterative (but rather fast even in high dimensions) computation is need in the case of δ_2. Another basis for the comparison of these notions of center is *rotation invariance*, as illustrated in Figure 5.4. The panels show different rotations of the same toy data set in $\mathbb{R}^2$, together with both the δ_1 and δ_2 Fréchet medians. Because the distance δ_2 is rotation invariant, the blue-green x showing that Fréchet median has the same relative position when the data are rotated. But because the distance δ_1 is not rotation invariant, this is not true for that version, shown as the red + sign. This example is deliberately constructed to show that the δ_1 Fréchet median can be viewed as a poor notion of center, in the sense that it actually lies on the boundary of the convex hull of the data.

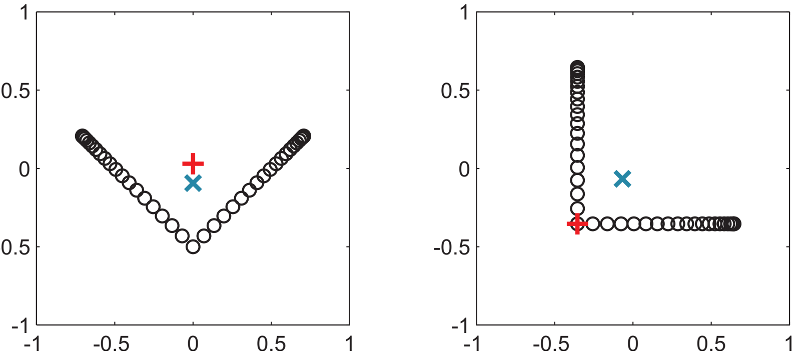

Figure 5.4: Two dimensional toy example studying rotation invariance. Data in the right panel are a 45 degree rotation of the data on the left. In each case the δ_2 based Fréchet median is shown as the red plus sign, with the δ_1 Fréchet median appearing as the blue-green x. Shows the latter is not rotation invariant.

Despite these clear differences, note that both of these medians are direct generalizations of the familiar one dimensional median in $\mathbb{R}^1$ in the sense that they both reduce to that quantity in the case $d = 1$. Actually there are a number of other quite different notions of center in $\mathbb{R}^d$ which also reduce to the median in the case $d = 1$, perhaps the most notable of which are based on *data depth* based ideas, see e.g. Liu et al [128] and Lopéz-Pintado and Romo [131].

Neither the Fréchet mean nor median are guaranteed to be unique. This is typically handled by working with *mean sets* (*median sets*, resp.). For example, on $\mathbb{R}^1$ when n is even, the interval between the central two data points is the median set. In that situation, and in others, sometimes a representative is chosen for the set, e.g. the midpoint of the interval in the $\mathbb{R}^1$ median case. The extreme case of non-uniqueness is the discrete metric δ_D, with respect to which both the Fréchet mean and median sets are the full data set.

From the mathematical statistical perspective an important issue is just what is being estimated by all of these notions of center. When the data objects are thought of as being a random sample from a probability distribution on the object space, then appropriate notions of population center are defined using Fréchet criteria based on expected values. In particular, sample Fréchet means as defined in (5.4) can be usefully treated as estimates of the theoretical (i. e. population) Fréchet means

$$\arg\min_{x \in S} E\delta\left(x, X\right)^2,$$

where X is a random variable with the population probability distribution. Similarly for Fréchet medians. Good overview of many results studying various types of asymptotic convergence of estimates to their population versions, in the challenging context of data objects on manifolds and manifold stratified cases, can be found in Patrangenaru and Ellingson [165]. Some interesting more recent results can be found in Huckemann and Eltzner [108], Huckemann and Hotz [106]. See Hotz et al [104] for a particularly unusual limiting distribution theory for the Fréchet mean in a manifold stratified space.

There are several synonyms for the Fréchet mean. These include *barycenter* (really the center of mass in physics, when using the Euclidean distance δ_2), Kärcher mean (usually used in the context of geometric quotient spaces, see Chapter 7), and geodesic mean (when δ is a geodesic distance, say on a curved manifold, again discussed in Chapter 7). Similarly there are synonyms for the Fréchet median including *geometric median*, *spatial median* and L^1 *M estimate.*

5.2 Multi-dimensional scaling

As with notions of center, there are many ways of quantitating variation about the center, based solely on metrics. A very simple one is the *Fréchet variance* which is just the minimum value attained in (5.4).

But generally more useful for data analytic tasks, such as those discussed in Sections 2.2 - 2.4, are distance based analogs of PCA. One approach to this is *multi-dimensional scaling (MDS).* MDS has been very popular in the psychometrics literature, and at least the nomenclature is usually attributed to Torgersen [211, 212] and Gower [86]. However, the underlying mathematics are substantially older, see Eckart and Young [63] and Young and Householder [236].

In its simplest form, MDS starts with a set of n data objects, with a known set of pairwise distances between them summarized as a distance matrix as in (5.1), and seeks to represent them as a set of points $x_1, \cdots, x_n \in \mathbb{R}^d$ for some d, in such a way that the Euclidean distances $\delta_2(x_i, x_j)$ approximate the elements of D as well as possible, in various senses. When the input distance matrix D is itself composed of pairwise Euclidean distances δ_2, typical basic algorithms return x_i as the vector of the first d PC scores. In that sense MDS extends PCA to cases where only distances are known. A toy example is shown in Figure 5.5.

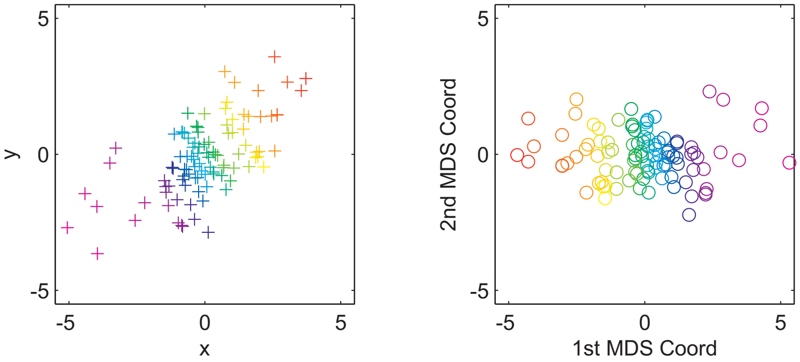

Figure 5.5: Toy 2d example, illustrating MDS. Raw data is shown using + signs in the left panel. Right panel shows the corresponding MDS scores as circles, calculated from the Euclidean distance matrix, which look quite similar to PC scores for the data on the left.

The raw data in the left panel are an elongated Gaussian point cloud, colored along the major axis of elongation. To construct the plot in the right panel, the δ_2 distance matrix D was computed, and then the corresponding MDS coordinates were plotted in the right panel. Note this looks much like what would be expected from PCA scores for this data set, which is consistent with the above discussion. As with PCA scores, these coordinates are only determined up to an arbitrary flip. In particular the colors indicate a reversal along the long axis.

There are many generalizations of the basic MDS idea. An important case in the psychometric literature extends the input of a distance matrix to merely a *dissimilarity matrix*, whose entries are not required to satisfy the metric properties such as the triangle inequality, see Kruskal [123] for good discussion. More overviews of this large area can be found in Cox and Cox [45], Borg and Groenen [25], Buja et al [32] and Chapter 3 of Zhai [240].

For OODA, an important generalization of MDS is to replace the embedding into Euclidean space $\mathbb{R}^d$, with embeddings into curved spaces. This is useful for data spaces that are strongly curved, such as phylogenetic tree space, as discussed in Section 9.1.

A weakness of distance and dissimilarity matrix based methods, is that they only give analyses of the data set at hand, and are challenging to extend to contexts requiring generalization to additional data, such as the classification / discrimination tasks considered in Chapter 10. An interesting approach to this is the *out of sample MDS* ideas of Trosset and Priebe [213]. See Section 3.4 of Zhai [240] for a more detailed overview of out of sample MDS.

Another way of handling data presented only in terms of a distance matrix D, that can be found in the machine learning literature, is to simply take the columns of D as data descriptors (using the object - descriptor terminology from Section 2.1). The impact of this approach on the relationship between data objects is studied in Figure 5.6, where direct understanding comes from using the same data (and colors for keeping track of individual data points) as in Figure 5.5.

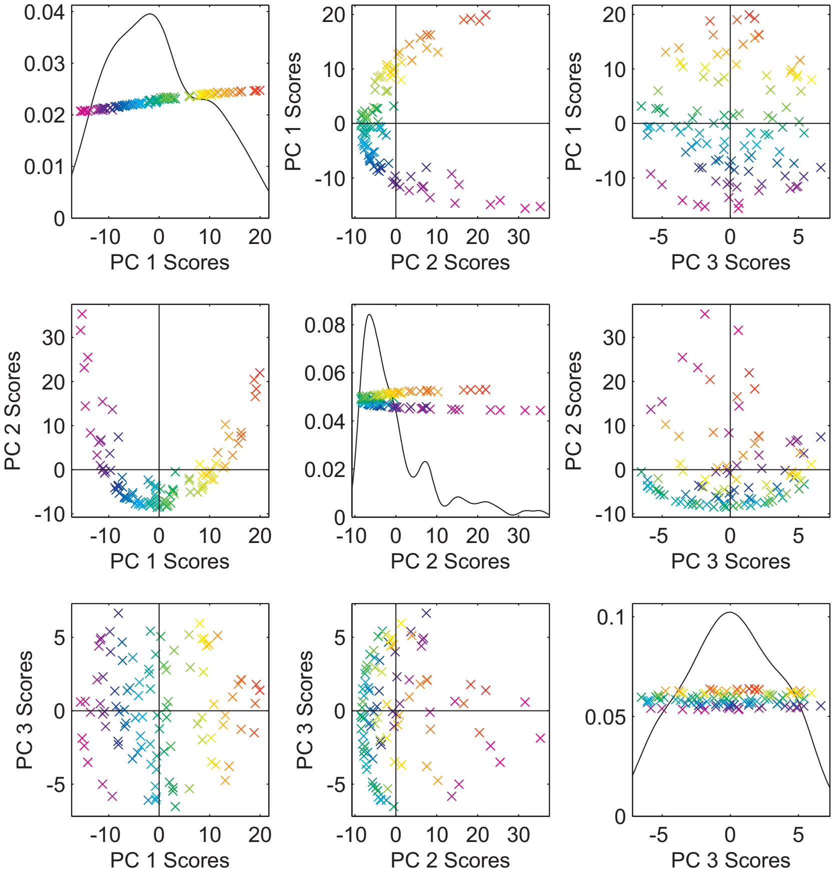

Figure 5.6: PCA of data using columns of the Euclidean distance matrix as descriptors, for the same data and colors as in Figure 5.5. Shows quite strong distortion of the relationships between data objects.

This PCA scatterplot matrix is in the same format as shown several times in Chapter 2. The colors in the upper left panel, indicate that the first PC scores are much as expected. However, the PC1 vs. PC2 scatterplot in the middle panel of the top row shows structure that is harder to interpret. This seems to reflect the fact that points at both the magenta and red ends have some more distant neighbors than the points in the middle, and indicates that this choice of descriptors has introduced a particular very strong distortion in the relationships between the data objects. At first glance, e.g. the PC1 vs. PC2 scatterplot in the top left panel, the 3rd PC component seems to be similar to the 2nd components shown on the vertical axis of the left panel in Figure 5.5. But a careful look at the extremes in the two cases show these are actually quite different, because in Figure 5.6 these scores must follow a direction orthogonal to PC2. For many purposes these issues seem like weaknesses of this representation, although in the spirit of kernel methods described in Section 10.2 this may sometimes provide a useful representation of the data objects.

??? Show some Kernel PCA here or in PCA Chapter???.

Chapter 6

Directions for Visualization

Euclidean space

3-d toy example???

Scatterplot Matrices? Need to refer to previous figures

Visual PCA of toy eg

Coloring of clusters – value of PCA

d=4000 coordinatewise vs. PCA

Limitations of Heat Maps

Curves vs. matrix views

Axis Scaling (e.g. Correlation PCA)

PCA not enough Apple Banana Pear Example

NCI60 data (including DWD directions, maybe MD?)

Fourier Subspace – Yeast Cell Cycle Data – Pointer to Smoothing Chapter

Independent Component Analysis

Classification Directions

Known modes of variation – Kingsolver Caterpillars

Handling of Non-orthogonal directions in scatterplot matrices. E.g. in Figure 2.18.

Chapter 7

Manifold Data Analysis

Toy example on circle to illustrate why need manifold view for directional data.

Connect with Chpater 5, e.g. for Frechet methods.

General theory: extrinsic versus intrinsic, Patrangenaru and Ellingson [165]. Give circular data example and discuss

7.1 Shapes as Data Objects

• Landmark representations o Equivalence Relations

o Equiv. Classes / Orbits as data objects

• Representations (landmark, boundary, medial)

• Manifolds & geodesics

• Circular Data

• Frechet mean (set) (relation to robust statistics)

• Image Analysis

• Bladder – Prostate – Rectum data

• PCA for S-reps (differing levels)

• Variation on landmark based shape – Focus on translation, shape is nuisance

• PNS

• Backwards PCA

o Desirability of Nesting

o NMF

o NNCA

o Principal Curves – Manifold Learning

o Sequence of Constraints

Make sure have lived up statements in 1st paragraph of Chapter E.

7.2 Covariance Matrices as Data Objects

Dryden's multiple metric, and varying geodesics.

Sungkyu and Armin's non-geodesic paths.
Piercesare's spatially indexed covariance matrices

7.3 Material from Old Chapter 1

While it is possible to attempt analysis of angle data in the ambient space, i.e. to treat points on S^2 as lying in $\mathbb{R}^3$, a major problem is that analytic methods such as PCA tend to leave the space where the data lie. One approach to this is called *extrinsic* analysis, see Patrangenaru and Ellingson [165] for good overview of this approach. The idea is to do the statistical analysis (e.g. the mean or PCA) in the ambient space and then project back to the curved manifold. This approach works well when the data lie in a small region of the manifold where there is not much curvature, so there is little distortion caused by the extrinsic approach.

When the data are distributed more broadly across the manifold, curvature matters more, which has motivated *intrinsic* analysis methods. One of these is the Principal Geodesic Analysis (PGA) of Fletcher et al [76]. The main idea of PGA is to consider the Euclidean PCA basis as a set of orthogonal lines that (sequentially) best fit the data. In PGA these best fitting lines are replaced by best fitting *geodesics* (e.g. great circles on S^2) which are a natural analog of lines. A deliberate choice that was made in PGA was to consider only geodesics passing through the Fréchet mean (the minimizer of the sum of squared distances along the manifold to the data). That restriction allowed straightforward computation of PGA, through consideration of the tangent plane, where the log map can be used to represent geodesics as lines in the tangent plane, where PCA can be performed, and then mapped back to the sphere using the exponential map. The results of a PGA, based upon $n = 17$ medial representations from a single patient are shown in Figure Q. ??? This has a little overlap with Chapter 1???

More recent research in medial shape models has gone in two important directions. First, as discussed in Damon [47, 48], Gorczowski et al [85] and Pizer et al [176], medial shape representations are more broadly applicable, and easier to implement and work with, when the medial assumption is weakened to allow models that are only approximately medial, called *skeletal shape representations.*

The second major recent research direction has been major improvements in statistical methodology that have been realized through improved integration with the underlying geometry. These recent methods fall into the *intrinsic* class of methods discussed in Patrangenaru and Ellingson [165]. Intrinsic versions of PCA have seen a series of innovations, that have yielded major statistical successes for skeletal shape representations.

Huckemann et al [107] made the important observation that the effectiveness of the PGA tangent plane analysis of Fletcher et al [76] was strongly tied to the quality of the geodesic mean (also sometimes called the barycenter or the Fréchet or Kärcher mean) as a notion of center-point. To see how this quality can be very poor consider a toy example, where the data objects are distributed along the equator of the ordinary sphere S^2. Because the geodesic mean is the point (or points, as this may not be unique) that minimizes the sum of squared distances (measured as arcs along the surface of the sphere) to the data objects, it will be both the north and south poles, as long as the data are sufficiently distributed around the equator. This is because least squares penalizes most strongly against large distances. Not only is the geodesic mean an unintuitive notion of data center, it is also exceptionally poor as a point of tangency for a PGA analysis. The reason is that the log map projection of the data objects onto the equator form a circle (centered at the geodesic mean). This has the unattractive property that it requires two modes (two PGA components) to quantify the variation. This seems inefficient because the data distributed on the equator of S^2 are one dimensional (in the sense of following a one dimensional curve in the high dimensional space), and thus should be describable by just a single component of variation, resulting in much more efficient statistical analysis (e.g. Bayes model fitting). The solution of Huckemann et al [107] was to consider all geodesics in modeling the data, thus moving beyond the restriction of PGA to only geodesics going through the geodesic mean. An interesting point here is that the first Euclidean PCA component can be viewed as the line that best fits the data, which will necessarily contain the sample mean, as easily seen using an *analysis of variance* decomposition of sums of squares. However, in non-Euclidean situations (e.g. data objects lying on curved manifolds) this is no longer true, so a conscious decision needs to be made. In particular the use of PGA entails the restriction to geodesics to those which go through the geodesic mean. The *Geodesic PCA* proposal of Huckemann et al [107] considers all geodesics, which thus takes the equator in the toy example, resulting in a more appropriate one dimensional representation of the data.

The above toy example of data objects lying on the equator of S^2 may appear to be artificial, but related modes of variation are actually frequently important to medial and skeletal shape representations, as demonstrated in Figure R. The left panel of Figure R shows the distribution of a single spoke over a number of realizations from the bladder-prostate-rectum simulator model of Jeong [112]. Note that the data are quite broadly spread over the sphere, as in the above discussed toy example.

Figure R also demonstrates a limitation of Geodesic PCA. This is that the spoke variation does not quite follow a great circle (i.e. a geodesic such as the equator in the above toy example), but instead follows a small circle, so the Geodesic PCA fit still requires two modes of variation as shown in the center panel. This motivated the *Principal Arc Analysis* proposed by Jung et al [117], which generalizes Geodesic PCA to allowing both small and great circle fits to the data. The benefit of this is shown in the right panel of Figure R, where only one mode of variation is needed to fully model this data set.

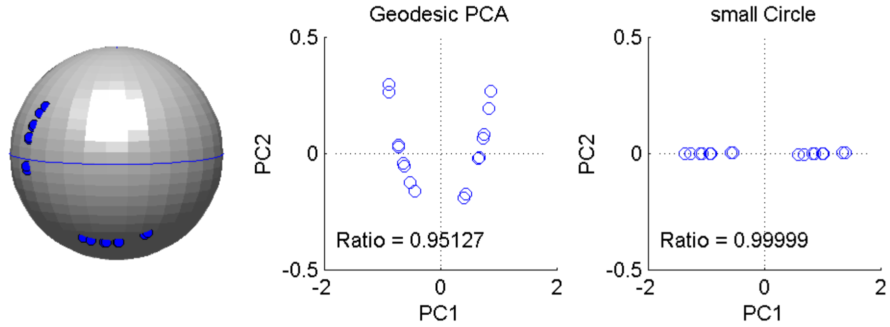

Figure R: Variation of a single spoke, from a bladder-prostate-rectum simulator model. Left panel shows the spherical component. Center panel shows the results of a Geodesic PCA summary, requiring two modes of variation. Right panel shows the corresponding PAA analysis, which is much more efficient since only one mode of variation is needed.

The idea of using small circles to give an analog of PCA was extended to higher dimensional spheres, S^d for $d > 2$, by Jung et al [116], to give a method called *Principal Nested Spheres* (PNS). In the special case of S^2 PNS gives the result show in the right panel of Figure R. Another important contribution of PNS is that it was the first example of *Backwards PCA*. In particular, for $d > 2$ the first step of PNS is to find the sub-sphere of dimension $d-1$ that best fits the data, in the sense of minimum sum of squared residuals (distance measured as arcs along the surface of S^d). Those signed residuals are kept as the highest level PNS scores. This process is then repeated iteratively down through dimension, to generate a full set of PNS scores. The final projections to S^1 play the role of PNS 1 scores. One more feature worth noting is that the geodesic mean of the PNS 1 scores is a compelling notion of center called the *backwards mean*. In particular, for the above toy example of data distributed around the equator of S^2, the backwards mean is the quite reasonable geodesic mean computed along the equator of the data.

A detailed study of the backwards PCA idea can be found in Damon and Marron [49]. The main idea is that ordinary Euclidean PCA can be calculated either forwards (by starting with lower dimensional fits and building up) or backwards (starting with the full data, and successively finding best fitting subspaces), since both just involve components of the eigen-analysis of the sample covariance matrix. However, this equivalence is essentially due to the Pythagorean Theorem, which no longer holds in non-Euclidean contexts. In that case, backwards methods are no longer the same as forwards. Damon and Marron [49] reviews these ideas noting that the backwards approach seems to more frequently give more useful methodologies. An intuitive basis for this observation was developed through viewing PCA in terms of a nested series of constraints. The success of backwards methods is interpretable in terms of it being relatively easy to sequentially find constraints, as done by backwards methods. On the other hand, to compute a forwards method one needs to know the full sequence of constraints in advance, and then sequentially relax them, which takes much more effort. An important contribution to the discussion of backwards versus forwards PCA is the *Barycentric Subspace Analysis* idea of Pennec [166]. The main idea is that both backwards and forwards methods are essentially greedy searches, which one expects can be improved by solving an overall optimization, although that comes at the price of increased complication.

Chapter 8

FDA Curve Registration

Either provide PCA promised in Section 4.4, or change that part to no longer say it is here.

On PCA, consider showing reconstructions at several eigen-levels, to make point that scree plot leading to number of eigencalues can be slippery (using this same toy example, or need another?). Include a scores plot to make point about "actually lower dimensional".

Probably should coordinate all this with Fig. 1.8, which shows pretty similar example. Maybe Chapter D version is best, and should be used everywhere?

- Many approaches
- Fisher-Rao registration
- Data Object Choices
- Chemical Spectra
- PNS on SRVF sphere – blood glucose data

Milan group combine clustering & curve registration

Make sure have lived up statements in 1st paragraph of Chapter E.

Chapter 9

Tree Structured Data Objects

- Brain Arteries
  - Approaches

  o Combinatorics

  2-d embedding – D-L visualization

  o Dyck Path

  o Phylogenetic approach

  o Persistent Homology

  Make sure have lived up statements in 1st paragraph of Chapter E.

9.1 Phylogenetic Trees

Discuss strong curvature, as refered to in Section 5.2. Show pics on this from Zhai [240]. Come back to interesting task for visualizing such data in a curved MDS embedding space.

Chapter 10

Classification

MD was used (and explained in Section 4.3)
DWD was usd in Section 2.2.2

10.1 Classical Methods

Mean Difference
FLD
Maximal Data Piling
Gaussian Likelihood Quadratic

10.2 Kernel Methods

10.3 Support Vector Machines

10.4 Distance Weighted Discrimination

- Stat vs.CS viewpoints & approaches
  - Mean Difference & Naïve Bayes
  - Fisher Linear Discrimination
  - o Nonparametric Derivation
  - o Mahalanobis Interpretation
  - o Likelihood Derivation
  - Gaussian Likelihood Ratio
  - Principal Discriminant Analysis (Generalized eigenanalysis)
  - HDLSS Discrimination – Generalized Inverses
  - Increasing dimension movies
  - Maximal Data Piling
  - o Low dimensional equivalence to FLD
  - o Good performance in autoregressive case

- Kernel Embedding
  - Polynomials
  - Radial Basis functions
  - Explicit vs. Implicit
- SVMs
  - Influence & Robustness
  - Tuning
- DWD
  - Faces
  - Simulations
  - HDLSS analysis
  - Tuning Parameter vs. SVM
- High d asys and kernel methods
- Radial DWD
- Random Forests / Neural Nets
  - Value of Linear Methods
- Mean Difference
- Melanoma Data Classification & ROC curves
- Suman Sen Manifold Classification
- Batch Adjustment
  - Perou Breast Cancer Data – visualizations
  - DWD adjustment
  - NCI 60 data

Include mean & sd
Cross-validation & Various flavors. Get “fold” terminology right.

Chapter 11

Clustering

Pointer to clustering done around Figure 2.13 (RNAseq example)

Milano Clustering curce registration

HDLSS theory for clustering Borysov et al [26].

Integrative Clustering & JIVE, Hellton and Thoresen [99]

- Unsupervised vs. Supervised Learning
- Find by PCA (Mass Flux Data)
- Verify with SiZer (or give pointers???)
- Dependent SiZer – Internet Data
- K-Means Clustering
- SWISS score
- More than 2 classes (relate to Princ. Disc. Anal.)
- Hierarchical Clustering

Chapter 12

Confirmatory Analysis

Confirmatory analysis is a nearly completely dominant part of classical statistics, as typically taught in modern courses. Much of this is based on fitting parametric probability distributions to data, and basing inference such as hypothesis tests and confidence intervals, on such distributions. As noted in Marron [148], this approach lies at the roots of the *scientific method*, which has provided great benefits to science over the years.

So far this approach has not been well developed in OODA contexts, often because in many OODA data settings such as the manifold data objects of Chapter 7 and the tree structured data objects in Chapter 9, suitable probability models are not yet in common use. However as shown in Figures 2.19 and 2.20 from Section 2.3, statistical confirmation is a critically important task, to ensure that non-spurious structure have been discovered. Two main approaches to this are discussed in this chapter. The DiProPerm hypothesis test, for testing the difference between two *previously labelled* subgroups is discussed in Section 12.1. While it is tempting to use DiProPerm to assess the significance of subgroups found by clustering methods, it is seen in Section 12.2 that this can be quite inappropriate. A test for significance of clustering, that is appropriate in the challenging high dimensional context, called DiProPerm, is described in Section .

12.1 DiProPerm

o Toy Examples
  o Caudate & Gene Expression
  o Choice of:
  Direction
  Statistic - Susan Wei analysis
  Compare power with other methods, e.g. Jinting's
  o Revisit Drug Discovery data

Great examples appear in Bioinf/GeneArray/TCGA-PanCan, as output from PanCan2.m. Note some of this appears in Section 4.3.Those used n1-n2=30 reduced data set, to show how log transformed version was better than raw counts. Would be interesting to generate picture along the lines of PanCan2ip6DiProPermDWDraw-ns30.ps, showing how DWD gives big boost over MD.

Include graphic for data shown in Figure 2.20. This is already computed by OODAbookChpBFigT.m, and should be like OODAbookChpBFigT-ShowDiProPerm.ps.

12.2 SigClust

Show example of “why no DiProPerm on cluster labels?”

Stress difference between “previously defined known group labels”, and those discovered by clustering

o Assumptions

o Q-Q plots (here or new earlier section?)

o Diagnostics

o Examples

Other approaches and comparisons

challenge of “what is a cluster?” uniform distributions, outliers, ...

Chapter 13

High Dimension Low Sample Size Analysis

Interesting work on consistency of scores: A particularly interesting characterization of the usefulness of PCA scores can be found in Hellton and Thoresen [100]. Hellton and Thoresen [?] explore the impact of measurement error on HDLSS scores.

- Points on Simplex
- Conditions (mixing, etc. & GWAS)
- Covariance 0 not independence (Gaussian & scale mixture)
- PCA
- DiProPerm insights
- Connection to Concentration of Measure

Remeber to explain, and quantify difference apparent in Figure 2.20.

Chapter 14

Smoothing

Controversies: smoothing methods, smoothing parameter selection. Härdle and Schimeck [94].

L1 view of density estimation, Devroye and Györfi [55].

Density estimation books, Silverman [197], Scott [191], Wand and Jones [222],

- Main focus: Bumps
- Histograms & Limitation

o Stamps (Incomes) Data

- Kernel Density Estimation

o Chondrite Data

o Stamps Data

- Local Linear Regression

o Fossils Data

Confirmatory analysis: Silverman's mode test, Silverman [196], SiZer Chaudhuri and Marron [39, 40], SSS Godtliebsen et al [83, 84].

Computation: Silverman's FFT, Silverman [198], Gasser and Seifert , Fan and Marron [66].

Exact risk analysis, Marron and Wand [144], Marron et al [150].

References to include

Jump SiZer of Kim and Marron [119]

Other smoothers

Wavelet SiZer of Park et al [164].

Bayesian Wavelets of Kohn at al [122].

Error Criteria, MISE, etc., L1...

Visual Error Criterion of Marron and Tsybakov [143]

Chapter 15

Robust Methods

Main Issue: handling violations of classical assumptions. Most often: Gaussianity versus outliers

Main controversy: outliers are "useless, thus completely discount" versus "outliers are too influential, thus downweight, but still should be allowed a vote.

- Cornea Data
- Outliers in PCA
- Zernike basis – Number of basis elements
- Outlier Deletion
- Spherical PCA (Oja's signs & ranks)
- Elliptical PCA
- Toy PCA Outliers
- Multivariate Medians (many, esp. DataDepth)
- Parallel Coordinates motivation of Elliptical PCA
- GWAS & L1 PCA

Chapter 16

PCA Background and Details

PCA as visualization more important than “dimension reduction”.

Point out the renamings of PCA, as mentioned in Section 2.2.

Gaussian Likelihood (go through math & point out seriously misleading as stated in Section 2.2)

Carefully define loadings and scores

• SVD relation (including graphic, and Lingsong’s paper on centering and Ian Carmichael’s sparse PCA computation, even with mean adjustent (sparse matrix plus low rank matrix)

• Differing rank SVD representations

• Graphics & Math of PCA Optimization

• Connect Math & Graphics (probably should refer to example in Figures 2.1-2.4)

• Redistribution of Energy (ANOVA sums of squares)

• Correlation PCA (or already done in Section 2.2?)

• Compare with SVD (= uncentered PCA)

• PCA Data Representation (Full Rotaton? Reduced Rank?)

• PCA Simulation

• PCA visual direction choice (with e.g.)

• Dual PC(same e.v.s) (Include dual of Spasnish Mortality?) Gabriel’s Biplot? Gram matrix

Centering issues, row & column mean

PCA limitations, apple-banana-pear, and pointer to NIC 60 example

Cite:

Generalized PCA Book

Kernel PCA

Integrating PCA with directions of maximal smoothness, in the context of evolutionary biology, Gaydos et al [80].

Point to related areas, such as Bollen’s structural equation modelling.

Chapter 17

Multi-Block Methods

• "Dependence" is critical

o Popular notions and measures, such as Pearson's correlation, and covariance matrices, are inadequate quantification

o (Anscombe's quartet)

o Two Important types of Dependence (say this earlier???):

o Dependence along vectors in descriptor space o Dependence between data objects ("Random sample" vs. "Time series of data objects)

• Partial least squares "Principal Singular Components" (Mueller paper, Byeong Park talk in Berlin, Enno meeting) SAME(?!?) as Partial Least Squares.

• Canonical Correlations

• JIVE:

• Spanish Mortality Males and Females. Pointer back to Section 1.1.

• Lobular Freeze Analysis

Relationship to other methods?

CCA, etc.

Gabriel's biplot?

Integrative Clustering, Hellton and Thoresen [99].

Acknowledgements:
SAMSI, grants, students
Iain Carmichael, Steven M. Pizer, Davide Pigoli, Piercesare Secchi

Appendix ideas:
Basics:

- Linear Algebra
- Multivariate Probability
- Inner and Outer Products
- Principal Angle Analysis???

Colors

- Heat Map visualizations
- Rainbow colors schemes, RGB & HSV
- Topological Colors: blue, lt green, darker green, light brown, darker brown, pink, white

Bibliography

[1] J. Aitchison. The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B*, 44(2):139–177, 1982. With discussion.

[2] J. Aitchison. *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[3] J. Aitchison and S. M. Shen. Logistic-normal distributions: some properties and uses. *Biometrika*, 67(2):261–272, 1980.

[4] Carlos A. Alfaro, Burcu Aydın, Carlos E. Valencia, Elizabeth Bullitt, and Alim Ladha. Dimension reduction in principal component analysis for trees. *Comput. Statist. Data Anal.*, 74:157–179, 2014.

[5] Nina Amenta, Manasi Datar, Asger Dirksen, Marleen de Bruijne, Aasa Feragen, Xiaoyin Ge, Jesper Holst Pedersen, Marylesa Howard, Megan Owen, Jens Petersen, et al. Quantification and visualization of variation in anatomical trees. In *Research in Shape Modeling*, pages 57–79. Springer, 2015.

[6] H An, James Stephen Marron, Todd A Schwartz, Jordan B Renner, F Liu, JA Lynch, NE Lane, Joanne Marie Jordan, and Amanda E Nelson. Novel statistical methodology reveals that hip shape is associated with incident radiographic hip osteoarthritis among african american women. *Osteoarthritis and Cartilage*, 24(4):640–646, 2016.

[7] Hyowon An. *Gaussian Centered L-Moments*. PhD thesis, The University of North Carolina at Chapel Hill, 2017.

[8] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

[9] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic resonance in medicine*, 56(2):411–421, 2006.

[10] John AD Aston, Davide Pigoli, and Shahin Tavakoli. Tests for separability in nonparametric covariance operators of random surfaces. *arXiv preprint arXiv:1505.02023*, 2015.

[11] Burcu Aydin, Gábor Pataki, Haonan Wang, Elizabeth Bullitt, and J. S. Marron. A principal component analysis for trees. *Ann. Appl. Stat.*, 3(4):1597–1615, 2009.

[12] Burcu Aydın, Gábor Pataki, Haonan Wang, Alim Ladha, and Elizabeth Bullitt. Visualizing the structure of large trees. *Electron. J. Stat.*, 5:405–420, 2011.

[13] Stephen R Aylward and Elizabeth Bullitt. Initialization, noise, singularities, and scale in height ridge traversal for tubular object centerline extraction. *IEEE transactions on medical imaging*, 21(2):61–75, 2002.

[14] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247, 2014.

[15] Peter J Basser, James Mattiello, and Denis LeBihan. Mr diffusion tensor spectroscopy and imaging. *Biophysical journal*, 66(1):259–267, 1994.

[16] Richard A. Becker and William S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987. With discussion and a reply by the authors.

[17] Richard A Becker, William S Cleveland, and Ming-Jen Shyu. The visual design and control of trellis display. *Journal of computational and Graphical Statistics*, 5(2):123–155, 1996.

[18] Paul Bendich, J. S. Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.*, 10(1):198–218, 2016.

[19] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[20] Michael W Berry and Murray Browne. *Understanding search engines: mathematical modeling and text retrieval*. SIAM, 2005.

[21] Julian Besag. On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc. Ser. B*, 48(3):259–302, 1986.

[22] Lynne Billard and Edwin Diday. *Symbolic data analysis*. Wiley Series in Computational Statistics. John Wiley & Sons, Ltd., Chichester, 2006. Conceptual statistics and data mining.

[23] Peter Bloomfield. *Fourier analysis of time series*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. An introduction.

[24] Hans-Hermann Bock and Edwin Diday. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer Science & Business Media, 2012.

[25] Ingwer Borg and Patrick J. F. Groenen. *Modern multidimensional scaling*. Springer Series in Statistics. Springer, New York, second edition, 2005. Theory and applications.

[26] Petro Borysov, Jan Hannig, and J. S. Marron. Asymptotics of hierarchical clustering for growing dimension. *J. Multivariate Anal.*, 124:465–479, 2014.

[27] Petro Borysov, Jan Hannig, James Stephen Marron, Eugene Muratov, Denis Fourches, and Alexander Tropsha. Activity prediction and identification of mis-annotated chemical compounds using extreme descriptors. *Journal of Chemometrics*, 2016.

[28] Arthur Lyon Bowley. *Elements of statistics*, volume 2. PS King, 1920.

[29] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

[30] David R. Brillinger. *Time series*. Holden-Day, Inc., Oakland, Calif., second edition, 1981. Data analysis and theory, Holden-Day Series in Time Series Analysis.

[31] Robert E Broadhurst, Joshua Stough, Stephen M Pizer, and Edward L Chaney. Histogram statistics of local model-relative image regions. In *Deep Structure, Singularities, and Computer Vision*, pages 72–83. Springer, 2005.

[32] Andreas Buja, Deborah F. Swayne, Michael L. Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *J. Comput. Graph. Statist.*, 17(2):444–472, 2008.

[33] Elizabeth Bullitt and Stephen R Aylward. Volume rendering of segmented image objects. *IEEE Transactions on Medical Imaging*, 21(8):998–1002, 2002.

[34] Elizabeth Bullitt, Keith E Muller, Inkyung Jung, Weili Lin, and Stephen Aylward. Analyzing attributes of vessel populations. *Medical image analysis*, 9(1):39–49, 2005.

[35] Adam Butler and Chris Glasbey. A latent Gaussian model for compositional data with zeros. *J. Roy. Statist. Soc. Ser. C*, 57(5):505–520, 2008.

[36] Joshua Cates, P Thomas Fletcher, Martin Styner, Martha Shenton, and Ross Whitaker. Shape modeling and analysis with entropy-based particle systems. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 333–345. Springer, 2007.

[37] EL Chaney, S Pizer, S Joshi, R Broadhurst, T Fletcher, G Gash, Q Han, J Jeong, C Lu, D Merck, et al. Automatic male pelvis segmentation from ct images via statistically trained multi-object deformable m-rep models. *International Journal of Radiation Oncology* Biology* Physics*, 60(1):S153–S154, 2004.

[38] Ted Chang. Estimating the relative rotation of two tectonic plates from boundary crossings. *J. Amer. Statist. Assoc.*, 83(404):1178–1183, 1988.

[39] Probal Chaudhuri and J. S. Marron. SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.*, 94(447):807–823, 1999.

[40] Probal Chaudhuri and J. S. Marron. Scale space view of curve estimation. *Ann. Statist.*, 28(2):408–428, 2000.

[41] Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.

[42] William S Cleveland. *Visualizing data*. Hobart Press, 1993.

[43] William S Cleveland et al. *The elements of graphing data*. Wadsworth Advanced Books and Software Monterey, CA, 1985.

[44] Timothy F Cootes, Andrew Hill, Christopher J Taylor, and Jane Haslam. Use of active shape models for locating structures in medical images. *Image and vision computing*, 12(6):355–365, 1994.

[45] Trevor F Cox and Michael AA Cox. *Multidimensional scaling*. CRC press, 2000.

[46] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[47] James Damon. Smoothness and geometry of boundaries associated to skeletal structures. I. Sufficient conditions for smoothness. *Ann. Inst. Fourier (Grenoble)*, 53(6):1941–1985, 2003.

[48] James Damon. Smoothness and geometry of boundaries associated to skeletal structures. II. Geometry in the Blum case. *Compos. Math.*, 140(6):1657–1674, 2004.

[49] James Damon and J. S. Marron. Backwards principal component analysis and principal nested relations. *J. Math. Imaging Vision*, 50(1-2):107–114, 2014.

[50] Charles Darwin. *The origin of species*. Lulu. com, 1872.

[51] Ingrid Daubechies. *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[52] B Davis, Mark Foskey, Julian Rosenman, Lav Goyal, Sha Chang, and Sarang Joshi. Automatic segmentation of intra-treatment ct images for adaptive radiation therapy of the prostate. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2005*, pages 442–450, 2005.

[53] Carl de Boor. *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York, revised edition, 2001.

[54] Howard B Demuth, Mark H Beale, Orlando De Jess, and Martin T Hagan. *Neural network design*. Martin Hagan, 2014.

[55] Luc Devroye and László Györfi. *Nonparametric density estimation*. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York, 1985. The L_1 view.

[56] E Diday. New kinds of graphical representation in clustering. In *Compstat*, pages 169–175. Springer, 1986.

[57] David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[58] David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, 57(2):301–369, 1995. With discussion and a reply by the authors.

[59] Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.*, 3(3):1102–1123, 2009.

[60] Ian L. Dryden, Alfred Kume, Huiling Le, and Andrew T. A. Wood. Statistical inference for functions of the covariance matrix in the stationary Gaussian time-orthogonal principal components model. *Ann. Inst. Statist. Math.*, 62(5):967–994, 2010.

[61] Ian L. Dryden and Kanti V. Mardia. *Statistical shape analysis with applications in R*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, second edition, 2016.

[62] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. Wiley-Interscience, New York, second edition, 2001.

[63] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[64] Paul H. C. Eilers and Brian D. Marx. Flexible smoothing with B-splines and penalties. *Statist. Sci.*, 11(2):89–121, 1996. With comments and a rejoinder by the authors.

[65] Craig K Enders. *Applied missing data analysis.* Guilford Press, 2010.

[66] Jianqing Fan and James S Marron. Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 3(1):35–56, 1994.

[67] Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: the r package fda. usc. *Journal of Statistical Software*, 51(4):1–28, 2012.

[68] Qianjin Feng, Mark Foskey, Wufan Chen, and Dinggang Shen. Segmenting ct prostate images using population and patient-specific statistics for radiotherapy. *Medical Physics*, 37(8):4121–4132, 2010.

[69] Qing Feng, Jan Hannig, and JS Marron. A note on automatic data transformation. *Stat*, 5(1):82–87, 2016.

[70] Aasa Feragen, Francois Lauze, Pechin Lo, Marleen de Bruijne, and Mads Nielsen. Geometries on spaces of treelike shapes. *Computer Vision–ACCV 2010*, pages 160–173, 2011.

[71] Aasa Feragen, Megan Owen, Jens Petersen, Mathilde MW Wille, Laura H Thomsen, Asger Dirksen, and Marleen de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *International Conference on Information Processing in Medical Imaging*, pages 74–85. Springer, 2013.

[72] Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York, 2006. Theory and practice.

[73] N. I. Fisher. *Statistical analysis of circular data.* Cambridge University Press, Cambridge, 1993.

[74] N. I. Fisher, T. Lewis, and B. J. J. Embleton. *Statistical analysis of spherical data.* Cambridge University Press, Cambridge, 1993. Revised reprint of the 1987 original.

[75] P Thomas Fletcher and Sarang Joshi. Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. In *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*, pages 87–98. Springer, 2004.

[76] P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

[77] P Thomas Fletcher, Suresh Venkatasubramanian, and Sarang Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, 2009.

[78] Michael W Frazier. *An introduction to wavelets through linear algebra.* Springer Science & Business Media, 2006.

[79] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, 10(3):215–310, 1948.

[80] Travis L Gaydos, Nancy E Heckman, Mark Kirkpatrick, JR Stinchcombe, Johanna Schmitt, Joel Kingsolver, James Stephen Marron, et al. Visualizing genetic constraints. *The Annals of Applied Statistics*, 7(2):860–882, 2013.

[81] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevelhierarchical models*, volume 1. Cambridge University Press New York, NY, USA, 2007.

[82] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[83] F. Godtliebsen, J. S. Marron, and Probal Chaudhuri. Significance in scale space for bivariate density estimation. *J. Comput. Graph. Statist.*, 11(1):1–21, 2002.

[84] Fred Godtliebsen, James Stephen Marron, and Probal Chaudhuri. Statistical significance of features in digital images. *Image and Vision Computing*, 22(13):1093–1104, 2004.

[85] Kevin Gorczowski, Martin Styner, Ja-Yeon Jeong, JS Marron, Joseph Piven, Heather Cody Hazlett, Stephen M Pizer, and Guido Gerig. Statistical shape analysis of multi-object complexes. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[86] John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

[87] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products.* Elsevier/Academic Press, Amsterdam, eighth edition, 2015. Translated from the Russian, Translation edited and with a preface by Daniel Zwillinger and Victor Moll, Revised from the seventh edition [MR2360010].

[88] Jennifer S Gregory, Jan H Waarsing, Judd Day, Huibert A Pols, Max Reijman, Harrie Weinans, and Richard M Aspden. Early identification of radiographic osteoarthritis of the hip using an active shape model to quantify changes in bone morphometric features: can hip shape tell us anything about the progression of osteoarthritis? *Arthritis & Rheumatism*, 56(11):3634–3643, 2007.

[89] P. Z. Hadjipantelis, J. A. D. Aston, H. G. Müller, and J. P. Evans. Unifying amplitude and phase analysis: a compositional data approach to functional multivariate mixed-effects modeling of Mandarin Chinese. *J. Amer. Statist. Assoc.*, 110(510):545–559, 2015.

[90] Pantelis Z Hadjipantelis, John AD Aston, and Jonathan P Evans. Characterizing fundamental frequency in mandarin: A functional principal component approach utilizing mixed effect models. *The Journal of the Acoustical Society of America*, 131(6):4651–4664, 2012.

[91] Ernst Heinrich Haeckel. *Generelle Morphologie der Organismen allgemeine Grundzuge der organischen Formen-Wissenschaft, mechanisch begrundet durch die von Charles Darwin reformirte Descendenz-Theorie von Ernst Haeckel: Allgemeine Entwickelungsgeschichte der Organismen kritische Grundzuge der mechanischen Wissenschaft von den entstehenden Formen der Organismen, begrundet durch die Descendenz-Theorie*, volume 2. Verlag von Georg Reimer, 1866.

[92] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[93] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[94] W Hardle and MG Schimek. Statistical theory and computational aspects of smoothing: Proceedings of the compstat94 satellite meeting held in semmering, austria, 27–28 august. *Physica-Verlag*, 1996.

[95] T. E. Harris. First passage and recurrence distributions. *Trans. Amer. Math. Soc.*, 73:471–486, 1952.

[96] John A. Hartigan. *Clustering algorithms*. John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.

[97] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

[98] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[99] Kristoffer Hellton and Magne Thoresen. Integrative clustering of high-dimensional data with joint and individual clusters, with an application to the metabric study. *arXiv preprint arXiv:1410.8679*, 2014.

[100] Kristoffer H Hellton and Magne Thoresen. When and why are principal component scores a good tool for visualizing high-dimensional data? *Scandinavian Journal of Statistics*, 2017.

[101] Y. Y. Ho. *Protein profiling of acute myeloid leukemia-specific membrane proteins using label-free liquid chromatography-mass spectrometry.* Honours thesis in Chemistry, The University of Adelaide (available from peter.hoffmann@adelaide.edu.au), 2011.

[102] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[103] Junpyo Hong, Jared Vicory, Jörn Schulz, Martin Styner, JS Marron, and Stephen M Pizer. Non-euclidean classification of medically imaged objects via s-reps. *Medical image analysis*, 31:37–45, 2016.

[104] Thomas Hotz, Stephan Huckemann, Huiling Le, J. S. Marron, Jonathan C. Mattingly, Ezra Miller, James Nolen, Megan Owen, Vic Patrangenaru, and Sean Skwerer. Sticky central limit theorems on open books. *Ann. Appl. Probab.*, 23(6):2238–2258, 2013.

[105] Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2009.

[106] Stephan Huckemann and Thomas Hotz. Nonparametric statistics on manifolds and beyond. In *Rabi N. Bhattacharya*, pages 599–609. Springer, 2016.

[107] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica*, 20(1):1–58, 2010.

[108] Stephan F Huckemann and Benjamin Eltzner. Essentials of backward nested descriptors inference. In *Functional Statistics and Related Fields*, pages 137–144. Springer, 2017.

[109] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

[110] Alfred Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2):69–91, 1985.

[111] Alfred Inselberg. *Parallel coordinates*. Springer, New York, 2009. Visual multidimensional geometry and its applications, With a foreword by Ben Shneiderman, With 1 CD-ROM (Windows).

[112] Ja-Yeon Jeong. *Estimation of probability distributions on multiple anatomical objects and evaluation of statistical shape models.* PhD thesis, Citeseer, 2009.

[113] Ja-Yeon Jeong, Joshua V Stough, J Steve Marron, and Stephen M Pizer. Conditional-mean initialization using neighboring objects in deformable model segmentation. In *Medical Imaging*, pages 69144R–69144R. International Society for Optics and Photonics, 2008.

[114] Norman L Johnson. Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):149–176, 1949.

[115] I. T. Jolliffe. *Principal component analysis.* Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

[116] Sungkyu Jung, Ian L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, 2012.

[117] Sungkyu Jung, Mark Foskey, and J. S. Marron. Principal arc analysis on direct product manifolds. *Ann. Appl. Stat.*, 5(1):578–603, 2011.

[118] András Kelemen, Gábor Székely, and Guido Gerig. Elastic model-based segmentation of 3-d neuroradiological data sets. *IEEE Transactions on medical imaging*, 18(10):828–839, 1999.

[119] C. S. Kim and J. S. Marron. SiZer for jump detection. *J. Nonparametr. Stat.*, 18(1):13–20, 2006.

[120] Patrick K Kimes, Christopher R Cabanski, Matthew D Wilkerson, Ni Zhao, Amy R Johnson, Charles M Perou, Liza Makowski, Christopher A Maher, Yufeng Liu, James Stephen Marron, et al. Sigfuge: single gene clustering of rna-seq reveals differential isoform usage among cancer samples. *Nucleic acids research*, 42(14):e113–e113, 2014.

[121] Inge Koch, Peter Hoffmann, and J. S. Marron. Proteomics profiles from mass spectrometry. *Electron. J. Stat.*, 8(2):1703–1713, 2014.

[122] Robert Kohn, J. S. Marron, and Paul Yau. Wavelet estimation using Bayesian basis selection and basis averaging. *Statist. Sinica*, 10(1):109–128, 2000.

[123] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[124] Sebastian Kurtek, Anuj Srivastava, Eric Klassen, and Zhaohua Ding. Statistical modeling of curves using shapes and related features. *J. Amer. Statist. Assoc.*, 107(499):1152–1165, 2012.

[125] Sebastian Kurtek, Anuj Srivastava, Eric Klassen, and Hamid Laga. Landmark-guided elastic shape analysis of spherically-parameterized surfaces. In *Computer graphics forum*, volume 32, pages 429–438. Wiley Online Library, 2013.

[126] Judith H Langlois and Lori A Roggman. Attractive faces are only average. *Psychological science*, 1(2):115–121, 1990.

[127] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[128] Regina Y. Liu, Jesse M. Parelius, and Kesar Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference. *Ann. Statist.*, 27(3):783–858, 1999. With discussion and a rejoinder by Liu and Singh.

[129] Xuxin Liu, Joel Parker, Cheng Fan, Charles M Perou, and JS Marron. Visualization of cross-platform microarray normalization. *Batch Effects and Noise in Microarray Experiments: Sources and Solutions. Wiley, New York*, pages 167–181, 2009.

[130] N Locantore, JS Marron, DG Simpson, N Tripoli, JT Zhang, KL Cohen, Graciela Boente, Ricardo Fraiman, Babette Brumback, Christophe Croux, et al. Robust principal component analysis for functional data. *Test*, 8(1):1–73, 1999.

[131] Sara López-Pintado and Juan Romo. Depth-based classification for functional data. In *Data depth: robust multivariate analysis, computational geometry and applications*, volume 72 of *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, pages 103–119. Amer. Math. Soc., Providence, RI, 2006.

[132] Conglin Lu, Stephen M Pizer, Sarang Joshi, and Ja-Yeon Jeong. Statistical multi-object shape models. *International Journal of Computer Vision*, 75(3):387–404, 2007.

[133] Xiaosun Lu and J. S. Marron. Analysis of juggling data: object oriented data analysis of clustering in acceleration functions [mr3273599]. *Electron. J. Stat.*, 8(2):1842–1847, 2014.

[134] Xiaosun Lu, J. S. Marron, and Perry Haaland. Object-oriented data analysis of cell images. *J. Amer. Statist. Assoc.*, 109(506):548–559, 2014.

[135] Gerald M Maggiora. On outliers and activity cliffs why qsar often disappoints, 2006.

[136] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[137] Kantilal Varichand Mardia. *Statistics of directional data*. Academic press, 2014.

[138] Kantilal Varichand Mardia, John T. Kent, and John M. Bibby. *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto, Ont., 1979. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

[139] J. S. Marron, Inge Koch, and Peter Hoffmann. Rejoinder: Analysis of proteomics data [mr3273586; mr3273587; mr3273588; mr3273589; mr3273590; mr3273585]. *Electron. J. Stat.*, 8(2):1756–1758, 2014.

[140] J. S. Marron, James O. Ramsay, Laura M. Sangalli, and Anuj Srivastava. Statistics of time warpings and phase variations. *Electron. J. Stat.*, 8(2):1697–1702, 2014.

[141] J. S. Marron, James O. Ramsay, Laura M. Sangalli, and Anuj Srivastava. Functional data analysis of amplitude and phase variation. *Statist. Sci.*, 30(4):468–484, 2015.

[142] J. S. Marron, Michael J. Todd, and Jeongyoun Ahn. Distance-weighted discrimination. *J. Amer. Statist. Assoc.*, 102(480):1267–1271, 2007.

[143] J. S. Marron and A. B. Tsybakov. Visual error criteria for qualitative smoothing. *J. Amer. Statist. Assoc.*, 90(430):499–507, 1995.

[144] J. S. Marron and M. P. Wand. Exact mean integrated squared error. *Ann. Statist.*, 20(2):712–736, 1992.

[145] J. Steve Marron and Andrés M. Alonso. Overview of object oriented data analysis. *Biom. J.*, 56(5):732–753, 2014.

[146] JS Marron. Marron, dryden: Book on object oriented data analysis.

[147] JS Marron. Spectral view of wavelets and nonlinear regression. In *Bayesian Inference in Wavelet-Based Models*, pages 19–32. Springer, 1999.

[148] JS Marron. Big data in context and robustness against heterogeneity. *Econometrics and Statistics*, 2:73–80, 2017.

[149] JS Marron, S Adak, IM Johnstone, MH Neumann, and P Patil. Exact risk analysis of wavelet regression. *Journal of Computational and Graphical Statistics*, 7(3):278–309, 1998.

[150] JS Marron, S Adak, IM Johnstone, MH Neumann, and P Patil. Exact risk analysis of wavelet regression. *Journal of Computational and Graphical Statistics*, 7(3):278–309, 1998.

[151] Dian I Martin and Michael W Berry. Mathematical foundations behind latent semantic analysis. *Handbook of latent semantic analysis*, pages 35–56, 2007.

[152] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[153] Alessandra Menafoglio and Piercesare Secchi. Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European Journal of Operational Research*, 258(2):401–410, 2017.

[154] Derek Merck, Gregg Tracton, Rohit Saboo, Joshua Levy, Edward Chaney, Stephen Pizer, and Sarang Joshi. Training models of anatomic shape variability. *Medical physics*, 35(8):3584–3596, 2008.

[155] Robb J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, Inc., New York, 1982. Wiley Series in Probability and Mathematical Statistics.

[156] José MR Murteira and Joaquim JS Ramalho. Regression analysis of multivariate fractional data. *Econometric Reviews*, 35(4):515–552, 2016.

[157] AE Nelson, Y Shi, R Tiller, TA Schwartz, JB Renner, JM Jordan, R Aspden, JS Gregory, and JS Marron. Baseline knee shape discriminates cases of incident knee radiographic oa from controls: a case-control study using novel methodology from the johnston county osteoarthritis project. *Osteoarthritis and Cartilage*, 25:S70–S71, 2017.

[158] Amanda E Nelson, Felix Liu, John A Lynch, Jordan B Renner, Todd A Schwartz, Nancy E Lane, and Joanne M Jordan. Association of incident symptomatic hip osteoarthritis with differences in hip shape by active shape modeling: the johnston county osteoarthritis project. *Arthritis care & research*, 66(1):74–81, 2014.

[159] Tom M. W. Nye. Principal components analysis in the space of phylogenetic trees. *Ann. Statist.*, 39(5):2716–2739, 2011.

[160] Ipek Oguz, Joshua Cates, Thomas Fletcher, Ross Whitaker, Derek Cool, Stephen Aylward, and Martin Styner. Cortical correspondence using entropy-based particle systems and local features. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1637–1640. IEEE, 2008.

[161] Steven J Owen. A survey of unstructured mesh generation technology. In *IMR*, pages 239–267, 1998.

[162] Leslie E Papke and Jeffrey Wooldridge. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates, 1993.

[163] Leslie E Papke and Jeffrey M Wooldridge. Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, 145(1):121–133, 2008.

[164] Cheolwoo Park, Fred Godtliebsen, Murad Taqqu, Stilian Stoev, and J. S. Marron. Visualization and inference based on wavelet coefficients, SiZer and SiNos. *Comput. Statist. Data Anal.*, 51(12):5994–6012, 2007.

[165] Victor Patrangenaru and Leif Ellingson. *Nonparametric statistics on manifolds and their applications to object data analysis.* CRC Press, Boca Raton, FL, 2016. With a foreword by Victor Pambuccian.

[166] Xavier Pennec. Barycentric subspace analysis on manifolds. *arXiv preprint arXiv:1607.02833*, 2016.

[167] Charles M Perou, Therese Sorlie, Michael B Eisen, Matt Van De Rijn, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747, 2000.

[168] Davide Pigoli, John A. D. Aston, Ian L. Dryden, and Piercesare Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014.

[169] Davide Pigoli, John A. D. Aston, Ian L. Dryden, and Piercesare Secchi. Permutation tests for comparison of covariance operators. In *Contributions in infinite-dimensional statistics and related topics*, pages 215–220. Esculapio, Bologna, 2014.

[170] Davide Pigoli, Pantelis Z Hadjipantelis, John S Coleman, and John AD Aston. The analysis of acoustic phonetic data: exploring differences in the spoken romance languages. *arXiv preprint arXiv:1507.07587*, 2015.

[171] Stephen M Pizer, Robert E Broadhurst, Ja-Yeon Jeong, Qiong Han, Rohit Saboo, JV Stough, Gregg Tracton, and Edward L Chaney. Intrapatient anatomic statistical models for adaptive radiotherapy. In *MICCAI Workshop From Statistical Atlases to Personalized Models: Understanding Complex Diseases in Populations and Individuals*, pages 43–46, 2006.

[172] Stephen M Pizer, Robert E Broadhurst, Joshua Levy, Xioaxiao Liu, Ja-Yeon Jeong, Joshua Stough, Gregg Tracton, and Edward L Chaney. Segmentation by posterior optimization of m-reps: Strategy and results. *Unpublished manuscript*, 2007.

[173] Stephen M Pizer, Junpyo Hong, Sungkyu Jung, JS Marron, Jörn Schulz, and Jared Vicory. Relative statistical performance of s-reps with principal nested spheres vs. pdms. In *Proc. Shape 2014-Symp. of Stat. Shape Models and Appl*, pages 11–13, 2014.

[174] Stephen M Pizer, Ja-Yeon Jeong, Robert E Broadhurst, Sean Ho, and Joshua Stough. Deep structure of images in populations via geometric models in populations. In *Deep Structure, Singularities, and Computer Vision*, pages 49–59. Springer, 2005.

[175] Stephen M Pizer, Ja-Yeon Jeong, Conglin Lu, Keith Muller, and Sarang Joshi. Estimating the statistics of multi-object anatomic geometry using inter-object relationships. In *Deep Structure, Singularities, and Computer Vision*, pages 60–71. Springer, 2005.

[176] Stephen M Pizer, Sungkyu Jung, Dibyendusekhar Goswami, Jared Vicory, Xiaojie Zhao, Ritwik Chaudhuri, James N Damon, Stephan Huckemann, and JS Marron. Nested sphere statistics of skeletal models. In *Innovations for Shape Analysis*, pages 93–115. Springer, 2013.

[177] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. Methods and case studies.

[178] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2002. Methods and case studies.

[179] James O. Ramsay, Paul Gribble, and Sebastian Kurtek. Description and processing of functional data arising from juggling trajectories. *Electron. J. Stat.*, 8(2):1811–1816, 2014.

[180] C Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17, 1958.

[181] Jean-Yves Royer and Ted Chang. Evidence for relative motions between the indian and australian plates during the last 20 my from plate tectonic reconstructions: Implications for the deformation of the indo-australian plate. *Journal of Geophysical Research: Solid Earth*, 96(B7):11779–11802, 1991.

[182] David Ruppert. What is kurtosis? an influence function approach. *The American Statistician*, 41(1):1–5, 1987.

[183] David Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2003.

[184] Laura M. Sangalli, Piercesare Secchi, and Simone Vantini. AneuRisk65: a dataset of three-dimensional cerebral vascular geometries. *Electron. J. Stat.*, 8(2):1879–1890, 2014.

[185] Laura M Sangalli, Piercesare Secchi, and Simone Vantini. Object oriented data analysis: a few methodological challenges. *Biometrical Journal*, 56(5):774–777, 2014.

[186] J. L. Scealy, Patrice de Caritat, Eric C. Grunsky, Michail T. Tsagris, and A. H. Welsh. Robust principal component analysis for power transformed compositional data. *J. Amer. Statist. Assoc.*, 110(509):136–148, 2015.

[187] J. L. Scealy and A. H. Welsh. Regression for compositional data by using distributions defined on the hypersphere. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(3):351–375, 2011.

[188] J. L. Scealy and A. H. Welsh. Colours and cocktails: compositional data analysis 2013 Lancaster lecture. *Aust. N. Z. J. Stat.*, 56(2):145–169, 2014.

[189] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[190] Jörn Schulz, Stephen M Pizer, James Stephen Marron, and Fred Godtliebsen. Non-linear hypothesis testing of geometric object properties of shapes applied to hippocampi. *Journal of Mathematical Imaging and Vision*, 54(1):15–34, 2016.

[191] David W. Scott. *Multivariate density estimation*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2015. Theory, practice, and visualization.

[192] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[193] Dan Shen, Haipeng Shen, Shankar Bhamidi, Yolanda Muñoz Maldonado, Yongdai Kim, and J. S. Marron. Functional data analysis of tree data objects. *J. Comput. Graph. Statist.*, 23(2):418–438, 2014.

[194] Nathaniel Shiers, John A. D. Aston, Jim Q. Smith, and John S. Coleman. Gaussian tree constraints applied to acoustic linguistic functional data. *J. Multivariate Anal.*, 154:199–215, 2017.

[195] Kaleem Siddiqi and Stephen Pizer. *Medial representations: mathematics, algorithms and applications*, volume 37. Springer Science & Business Media, 2008.

[196] B. W. Silverman. Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B*, 43(1):97–99, 1981.

[197] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[198] BW Silverman. Algorithm as 176: Kernel density estimation using the fast fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1):93–99, 1982.

[199] Sean Skwerer, Elizabeth Bullitt, Stephan Huckemann, Ezra Miller, Ipek Oguz, Megan Owen, Vic Patrangenaru, Scott Provan, and J. S. Marron. Tree-oriented analysis of brain artery structure. *J. Math. Imaging Vision*, 50(1-2):126–143, 2014.

[200] Anuj Srivastava, Wei Wu, Sebastian Kurtek, Eric Klassen, and JS Marron. Registration of functional data using fisher-rao metric. *arXiv preprint arXiv:1103.3817*, 2011.

[201] Robert G. Staudte and Simon J. Sheather. *Robust estimation and testing*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, 1990. A Wiley-Interscience Publication.

[202] Connie Stewart and Christopher Field. Managing the essential zeros in quantitative fatty acid signature analysis. *Journal of Agricultural, Biological, and Environmental Statistics*, 16(1):45–69, 2011.

[203] Charles J. Stone, Mark H. Hansen, Charles Kooperberg, and Young K. Truong. Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.*, 25(4):1371–1470, 1997. With discussion and a rejoinder by the authors and Jianhua Z. Huang.

[204] Joshua V Stough, Robert E Broadhurst, Stephen M Pizer, and Edward L Chaney. Regional appearance in deformable model segmentation. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 532–543. Springer, 2007.

[205] Gábor Szeg˝ o. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[206] Shahin Tavakoli, Davide Pigoli, John AD Aston, and John Coleman. Spatial modeling of object data: Analysing dialect sound variations across the uk. *arXiv preprint arXiv:1610.10040*, 2016.

[207] R Core Team. The r project for statistical computing. *Available at www. R-project. org/. Accessed October*, 31:2014, 2014.

[208] Fabian JE Telschow, Stephan F Huckemann, and Michael Pierrynowski. Asymptotics for object descriptors. *Biometrical Journal*, 56(5):781–785, 2014.

[209] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[210] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[211] Warren S. Torgerson. Multidimensional scaling. I. Theory and method. *Psychometrika*, 17:401–419, 1952.

[212] Warren S Torgerson. Theory and methods of scaling. 1958.

[213] Michael W. Trosset and Carey E. Priebe. The out-of-sample problem for classical multidimensional scaling. *Comput. Statist. Data Anal.*, 52(10):4635–4642, 2008.

[214] Michail T Tsagris, Simon Preston, and Andrew TA Wood. A data-based power transformation for compositional data. *arXiv preprint arXiv:1106.1451*, 2011.

[215] E.R. Tufte. *The Visual Display of Quantitative Information.* Encyclopedia of mathematics and its applications. Graphics Press, 1983.

[216] John Tukey and Paul Tukey. Strips displaying empirical distributions: I. textured dot strips. Technical report, Bellcore Technical Memorandum, 1990.

[217] John W Tukey. Exploratory data analysis. 1977.

[218] John W. Tukey. Data-based graphics: visual display in the decades to come. *Statist. Sci.*, 5(3):327–339, 1990.

[219] W. N. Venables and B. D. Ripley. *Modern applied statistics with S-Plus.* Statistics and Computing. Springer-Verlag, New York, 1994. With 1 IBM-PC floppy disk (3.5 inch; HD).

[220] William N Venables and Brian D Ripley. *Modern applied statistics with S-PLUS.* Springer Science & Business Media, 2013.

[221] Jean-Philippe Vert. A tree kernel to analyse phylogenetic profiles. *Bioinformatics*, 18(suppl 1):S276–S284, 2002.

[222] M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, Ltd., London, 1995.

[223] Haonan Wang and J. S. Marron. Object oriented data analysis: sets of trees. *Ann. Statist.*, 35(5):1849–1873, 2007.

[224] Yuan Wang, J. S. Marron, Burcu Aydin, Alim Ladha, Elizabeth Bullitt, and Haonan Wang. A nonparametric regression model with tree-structured response. *J. Amer. Statist. Assoc.*, 107(500):1272–1285, 2012.

[225] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[226] Susan Wei, Chihoon Lee, Lindsay Wichers, and J. S. Marron. Direction-projection-permutation for high-dimensional hypothesis tests. *J. Comput. Graph. Statist.*, 25(2):549–569, 2016.

[227] Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Günter Rudolph. Music data analysis: Foundations and applications, 2016.

[228] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[229] Leland Wilkinson, Anushka Anand, and Robert L Grossman. Graph-theoretic scagnostics. In *INFOVIS*, volume 5, page 21, 2005.

[230] Leland Wilkinson and Graham Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17(2):473–491, 2008.

[231] John R Wilmoth and Vladimir Shkolnikov. Human mortality database. *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*, 2008.

[232] Wei Wu, Nicholas G. Hatsopoulos, and Anuj Srivastava. Introduction to neural spike train data for phase-amplitude analysis. *Electron. J. Stat.*, 8(2):1759–1768, 2014.

[233] Jie Xiong, D. P. Dittmer, and J. S. Marron. “Virus hunting” using radial distance weighted discrimination. *Ann. Appl. Stat.*, 9(4):2090–2109, 2015.

[234] Yoshihiro Yamanishi, Francis Bach, and Jean-Philippe Vert. Glycan classification with tree kernels. *Bioinformatics*, 23(10):1211–1216, 2007.

[235] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, 100(470):577–590, 2005.

[236] Gale Young and Alston S Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

[237] Qunqun Yu, Xiaosun Lu, and J. S. Marron. Principal Nested Spheres for Time-Warped Functional Data Analysis. *J. Comput. Graph. Statist.*, 26(1):144–151, 2017.

[238] Ying Yuan, Hongtu Zhu, Weili Lin, and J. S. Marron. Local polynomial regression for symmetric positive definite matrices. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(4):697–719, 2012.

[239] Ying Yuan, Hongtu Zhu, Martin Styner, John H. Gilmore, and J. S. Marron. Varying coefficient model for modeling diffusion tensors along white matter tracts. *Ann. Appl. Stat.*, 7(1):102–125, 2013.

[240] Haojin Zhai. *Principal component analysis in phylogenetic tree space*. PhD thesis, The University of North Carolina at Chapel Hill, 2016.