Robust Principal Component Analysis for Functional Data

N. Locantore
Department of Statistics, University of North Carolina,
Chapel Hill, N. C. 27599-3260
J. S. Marron
Department of Statistics, University of North Carolina,
Chapel Hill, N. C. 27599-3260

D. G. Simpson
Department of Statistics, University of Illinois,
Urbana - Champaign, IL 61820

N. Tripoli
Department of Ophthalmology, University of North Carolina,
Chapel Hill, N. C. 27599-7040

J. T. Zhang
Department of Statistics, University of North Carolina,
Chapel Hill, N. C. 27599-3260

K. L. Cohen
Department of Ophthalmology, University of North Carolina,
Chapel Hill, N. C. 27599-7040

February 2, 1999

Abstract

A method for exploring the structure of populations of complex objects, such as images, is considered. The objects are summarized by feature vectors. The statistical backbone is Principal Component Analysis in the space of feature vectors. Visual insights come from representing the results in the original data space. In an ophthalmological example, endemic outliers motivate the development of a bounded influence approach to PCA.

1 Introduction

The "atoms" of traditional statistical analyses are numbers or perhaps vectors. But a number of data sets, from diverse areas of science, provide motivation for generalizing the notion of the atom of the statistical analysis to more general data types. Ramsay and Silverman (1997) have coined the term "functional" for such data. That monograph contains a wide array of examples, and also makes a good start on the development of statistical methods for their analysis.

While this type of new statistical analysis makes use of classical multivariate analysis methods, such as Principal Component Analysis, substantial adaptation and new development is typically needed. For example, when the atoms of the analysis are "curves", e. g. longitudinal data, they can typically be effectively digitized to vectors. However classical methods make little use of the "smoothness" that is present in many data sets. Hence they are poorly suited for analysis in such cases. One reason is that the needed covariance matrices are singular, or nearly so. A second reason is that classical statistical methods tend to be powerful in an "omnibus" way, and thus tend to trade away power in the particular directions that are more important for functional data analysis (e.g. in directions corresponding to "smoothness"). See Fan and Lin (1998) for interesting discussion of this point, and some useful hypothesis testing ideas in functional data analytic contexts.

This paper considers the statistical analysis of data types that go beyond the idea of "curves as data", that was the focus of Ramsay and Silverman (1997), into more complicated data structures. There are two main points. The first is that complicated data types can be effectively handled and analyzed through summarizing them in terms of "feature vectors". The second is that robust methods are very useful, and are perhaps more important in functional situations than in classical ones, since there tend to be more ways for outliers to impact very high dimensional statistical analyses.

The motivating example used in this paper comes from ophthalmology. An important component of the human visual system is the shape of the outside surface of the cornea, the outer surface of the eye. The shape of this surface is responsible for 85% of the refraction that results in an image focused on the retina. Corneal topography measurement instruments such as the Keratron (Optikon 2000, Rome) typically use color-coded maps to display anterior corneal shape information in two dimensions. A useful convention is the mapping of radial curvature that depicts low curvature as blue, then green, yellow, orange, and red as the curvature increases.

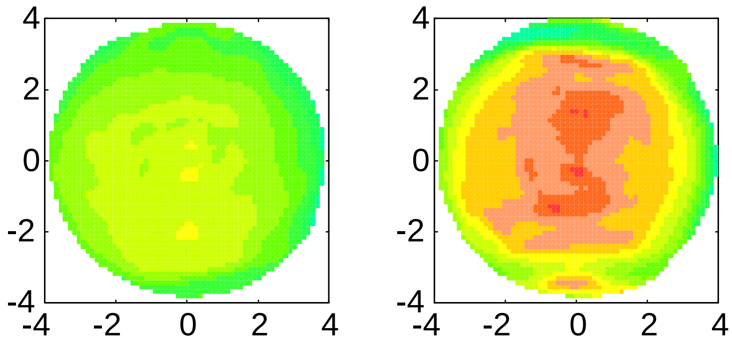

Figure 1.1: Two corneal images showing radial curvature. The left shows relatively constant curvature. The right shows more curvature near the center, and a marked vertical astigmatism.

Two such images are shown in Figure 1.1. These show two features often seen in populations of corneas. The first has fairly constant curvature (shown by nearly constant color), while the second has a vertical orange band, representing astigmatism with a vertical axis.

This type of image provides a useful diagnostic tool. For example, Figure 1.2 shows a curvature map from a patient with the disease of keratoconus, in which the cornea grows into a highly curved cone shape

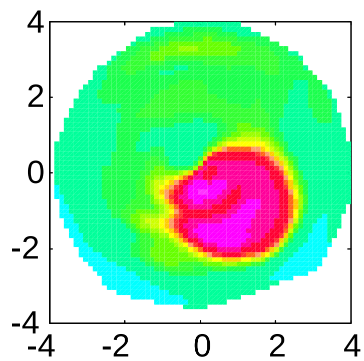

Figure 1.2: Radial curvature of a cornea with Kerataconus. The red region is a cone of high curvature.

In this paper, we study this type of data from a population viewpoint, i.e. the atoms of our analysis are such images. While the example is quite specialized, we believe the methodology developed will be useful for a wide variety of populations of images, and other complex objects.

In Section 2 we discuss effective summarization of each data point into "feature vectors", by fitting the Zernike orthogonal basis to each. In Section 3 Principal Component Analysis is used to understand the structure of a population of normal corneas. The analysis is actually done in the "feature space" of

Zernike vectors, but the results are viewed in the "data space" of curvature images, since this is where visual insights are gained. This idea was independently developed by Cootes, Hill, Taylor and Aslam (1993) and Kelemen, Szekely, and Gerig (1997). In statistics, related methods are often used in "shape analysis", see Dryden and Mardia (1998).

In section 3 it is seen that this PCA reveals several clinically intuitive aspects of the population. But a disturbing feature of the analysis is that it is affected by outliers, caused by some of the images having some missing regions. These outliers motivate a robust bounded influence approach to PCA.

The first step in robust PCA is finding the centerpoint of the population. A suitable robust estimate of "center" is developed in Section 4, which is a modification of the standard L^1 M-estimate. Robust estimates based on a useful surrogate for the covariance matrix are then developed in Section 5. Standard robust estimates of the full covariance matrix are useless here (and we expect this same difficulty to occur in many other very high dimensional contexts) since the number of data points is less than the dimensionality. We overcome this problem using "Spherical Principal Component Analysis", which is a robust version of PCA that is anticipated to be broadly useful. Finally due to the special nature of these data, a simple extension is made to "Elliptical Principal Component Analysis". Details of the Zernike decomposition are given in Section 6.

2 Reduction by Zernike Decomposition

The first challenge in the analysis of the corneal image data is that the raw data are in the form of up to 6912 measurements at a polar grid of locations. Classical multivariate analysis on these vectors is numerically intractable, because of their large size, and because they contain many redundancies and near redundancies.

The problem of reducing data of this type to more manageable "feature vectors" is familiar to the field of statistical pattern recognition, see e.g. Devijver and Kittler (1982). An effective summarization of an image of the type in Figure 1, into a feature vector, is the vector of the coefficients of a least squares fit of the Zernike orthogonal basis.

This two dimensional basis is supported on the disk, and is a tensor product of the Fourier basis in the angular direction, and a special Jacobi basis in the radial direction. The Jacobi basis is very carefully chosen to avoid singularities at the origin. This basis is standard in optics, and is well suited to summarizing optical quantities such as spherical curvature and astigmatism. Mathematical details are discussed in Section 6.

The results of Zernike feature vector summarization, for the images of Figure 1.1, as well as several others, are shown in Figure 2.1. There is some loss in this type of image compression, but it is relatively small, and more important the missing features are not of clinical interest.

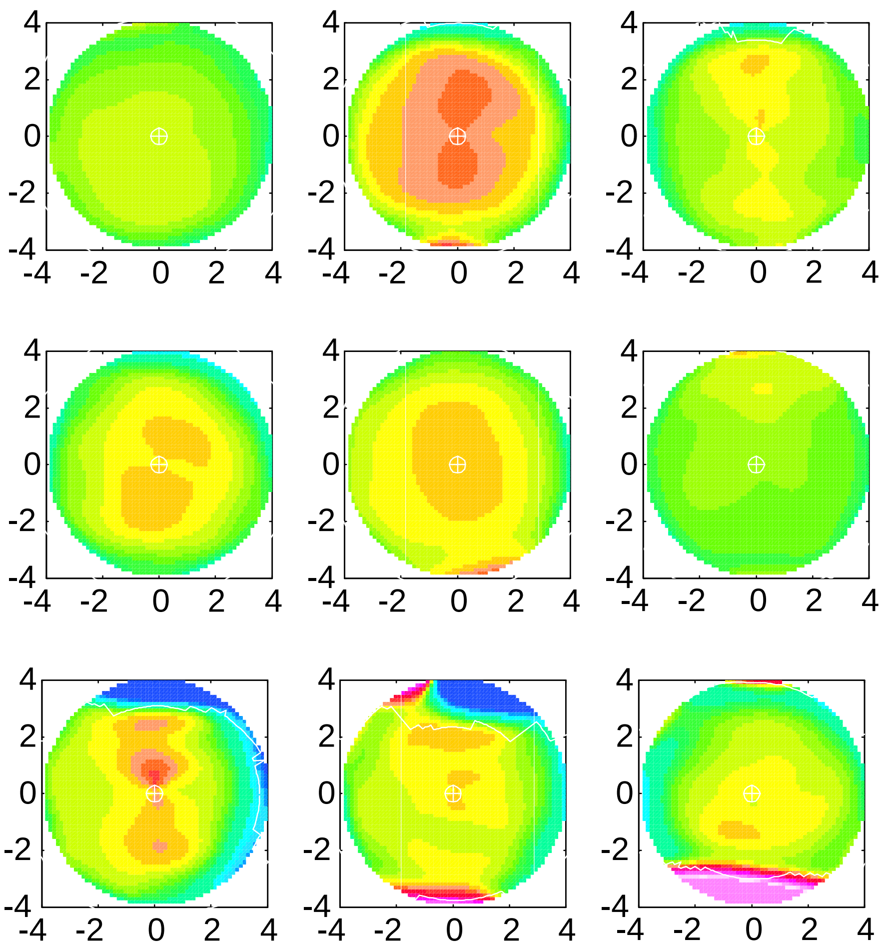

Figure 2.1: Zernike reconstructions of some normal cornea images.

Next we study a population of n = 43 normal corneal images, which were obtained while screening patients for laser surgery. The images shown in Figure 2.1 are a subset, chosen to represent the most important features. Note that the raw curvature images from Figure 1.1 now appear "smoothed". This is the same effect that is observed when a digitized smooth curve is Fourier transformed, and then the transform is inverted using only the low frequency coefficients. The main features are still present, but the rough edges have been smoothed away. Varying degrees of astigmatism are seen as vertical bands of steep curvature in the top center and right, the middle left and center, and the bottom center. Another feature widely observed in normal corneas is the tendency to be steeper either near the top, or near the bottom, shown to varying degrees in the top left and right, middle right and bottom right. Another feature extreme curvature caused by missing data in the images' peripheries, are the red and blue regions of extreme curvature. These are the results of artifacts, such as eyelids blocking the imaging device (the extent of the missing data for each is shown by the thin white lines). The missing data has a serious impact on the Zernike fit, which is reflected by these regions of high curvature. These effects are seen to have an important impact on the analysis of Section 3.

The difficulty of developing an intuitive understanding of the overall structure of the population by viewing a collection of color-coded maps is demonstrated by these nine images. The challenge is overwhelming when all 43 images are included. This can be seen by viewing an MPEG movie of all 43, available from the web page http://www.unc.edu/depts/statistics/postscript/papers/marron/cornearobust/, in the file normlwr.mpg. The reason is simply that there is too much information present, and this information is presented in a visual form that the human perceptual system is not able to effectively comprehend.

3 Ordinary Principal Components Analysis

PCA can provide an effective solution to this quite general problem of understanding the structure of complex populations. Classical PCA seeks one dimensional "directions of greatest variability", by studying projections of the data onto direction vectors starting at the sample mean. The variance of these projections is maximized in the direction of the first eigenvector (i.e. the one with the largest corresponding eigenvalue) of the sample covariance matrix. A simple example is shown in Figure 3.1. Here the data is a simple two dimensional point cloud, where each point is represented by a circle. PCA can be viewed as "decomposing the point cloud" into pieces which reveal the structure of the population. In Figure 3.1 it is centered at the sample mean, where the two lines meet. The heavier line shows the first direction of greatest variability, i.e. the direction of the first eigenvector of the covariance matrix. The thinner line shows the direction of greatest variability in the subspace that is the orthogonal complement (trivial in this example, since that subspace is one dimensional, but otherwise found via the eigenvector with second largest eigenvalue). Each data point is projected onto the thick line to get its "first principal component", shown as a thick +, and is projected onto the thin line to get its "second principal component", shown as a thin +. In each case the principal components give a particular one dimensional view of the data. An important property of PCA is that it allows finding interesting low dimensional representations of the data.

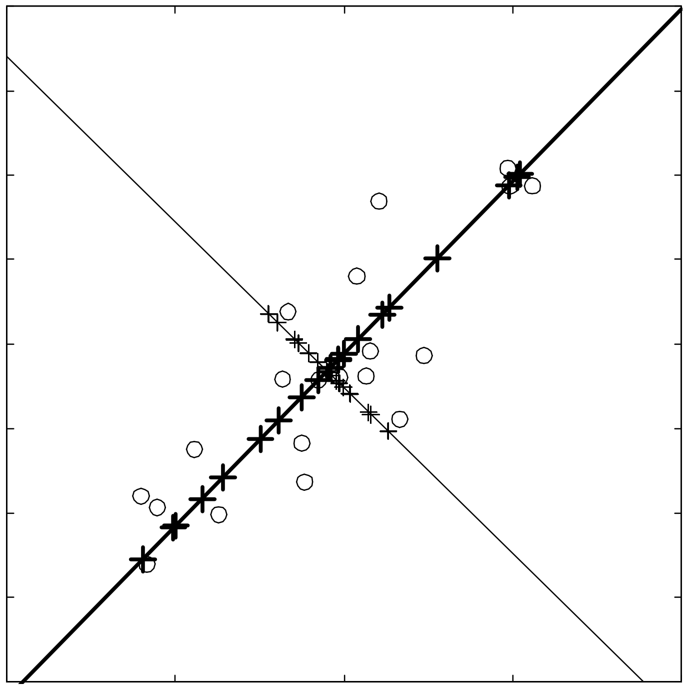

Figure 3.1: Two dimensional example illustrating PCA. First eigenvector direction (and projections of the data) shown with a thick line (thick plusses). Second eigenvector direction (and projections of the data) shown with a thin line (thin plusses).

For application in functional data contexts, the key is to do the PCA "in the feature space" (i.e. on the feature vectors), but then to gain insights "in the data space". For curves as data, Ramsay and Silverman (1997) were successful with overlaying the curves that represent each data point. The PCA directions are effectively displayed by projecting each data point onto the eigenvector, and then representing each projected point again as a curve. The family of curves then clearly displays the intuitive meaning of the component of variability that is represented by that eigendirection. A simulated example of the effectiveness of this type of visual representation is given in Figure 3.2.

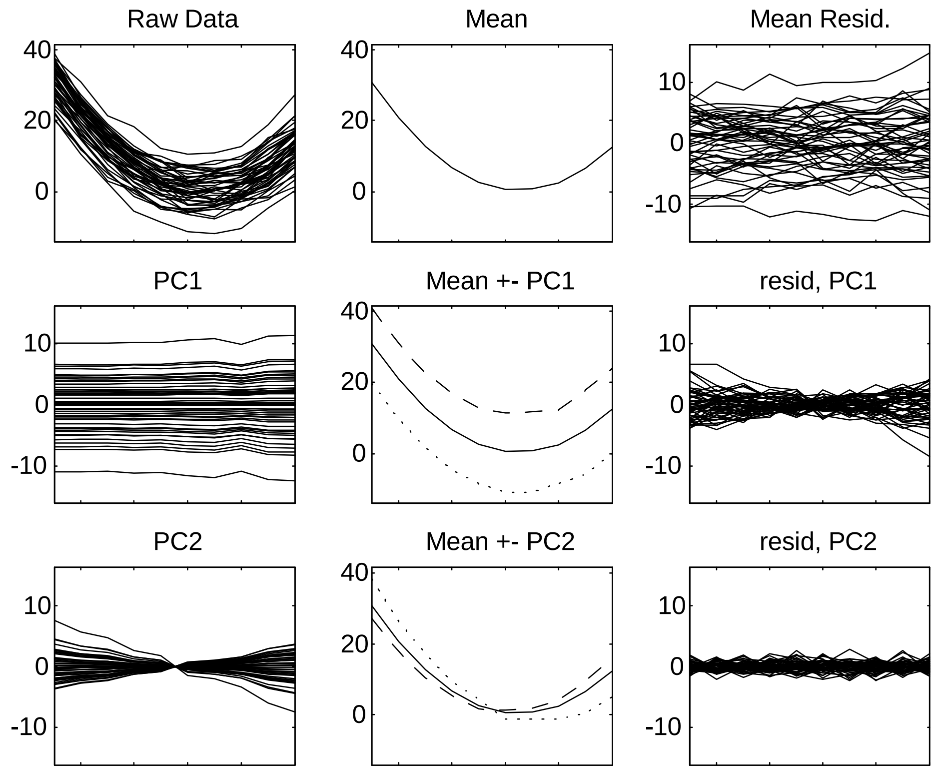

Figure 3.2: "Curves as data" example illustrating PCA. First row shows results of "recentering". Second row shows strongest component of variability. Third row shows second most important component.

The upper left plot shows a simulated family of random curves, that is considered here to be a population whose structure is to be analyzed. This type of visual representation of high dimensional data was termed "parallel coordinates" by Inselberg (1985) and Wegman (1990), who proposed it as a general purpose device for the visualization of high dimensional data (i.e. of point clouds in high dimensional space). The next plot to the right shows the sample mean of this population (i.e. of this point cloud). Since the multivariate mean is calculated coordinate-wise, this is simply the coordinate-wise mean of the curves. The next component shows the residuals from subtracting the mean curve from the raw data. This represents the point cloud which results from shifting the original point cloud so it is now centered at the sample mean.

Next PCA is used to understand the structure of the residual point cloud. The first eigenvector is computed, and the data are projected as in Figure 3.1. Two representations of the set of the projections (i.e. the heavy plusses in Figure 3.1) are shown in the second row. Since these projections are points in the mean residual space (i.e. the data space recentered at the mean), one representation is a parallel coordinate plot overlay, shown in the left plot in the second row. Another representation is shown in the center plot of the second row, in the original data space, which is the mean curve, together with just two extreme projections. Both displays show that the dominant direction of variability is "vertical shift" (which was a feature built into these simulated data). The right hand plot shows the residuals from subtracting the projections from the recentered data (i.e. it is the difference of the plot above, and the plot on the left). This shows the projection onto the complementary subspace (represented by the thin plusses in Figure 3.1). The direction of next greatest variability is analyzed in the same way in the third row. Note that this direction reveals a "tilting component" in the data that is not visually apparent in the raw data plot. This gives a hint about the power of PCA in finding structure in populations of complex objects. Further eigendirections are not shown for this data set, since they do not reveal additional interesting structure.

While the parallel coordinates visual representation is very useful when the data are curves (as shown in the left hand column of Figure 3.2), it does not give an intuitively useful view when the data are images (as in Figure 2.1) or more complex structures that are not usefully overlaid on a single plot. For example note that Figure 4.4, a parallel coordinate plot for the population of 43 normal corneal shapes, does not contain much insight about the population of curvature images (a subset of which can be seen in Figure 2.1). Since intuitive understanding comes in the feature space, that is where the visualization of the PCA must be done. While overlays (as in the left column of Figure 3.2) are no longer useful, representations of the directions in terms of extremes, as shown in the center column of Figure 3.2, are quite useful. Studying the mean, together with extremes in each direction, gives insight into that "direction of variability". Figure 3.3 shows such a representation for the direction of the first eigenvector (i.e. the direction of greatest variability) of the cornea data set shown in Figure 2.1.

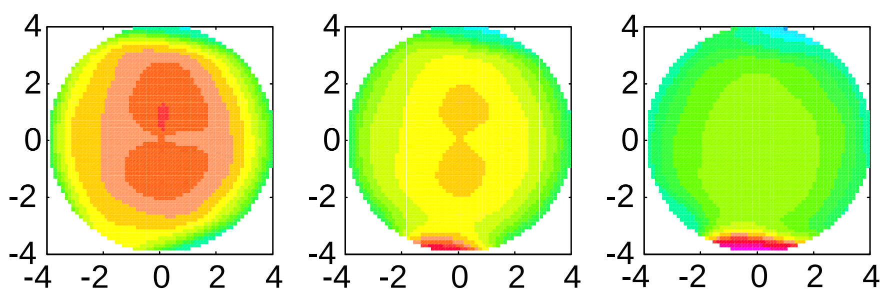

Figure 3.3: Mean image of the population of normal corneas in the center. Representatives of the first principal component direction on either side give an impression of the direction of greatest variability.

The center panel of figure 3.3 shows the population mean. This shows a moderate amount of curvature, and some astigmatism, which are known features of the population of normal corneas. The mean also has been affected somewhat by the edge effects on some of the images, as can be seen in Figure 2.1. The left and right panels of Figure 3.3 give an impression of the direction (in the 66 dimensional feature space) of the first eigenvector. This shows a combination of two known population features. First there is overall higher and lower curvature (shown as overall orange on the left, and green on the right). Second there is stronger (left) and weaker (right) levels of vertical astigmatism. There is some influence from the missing data also on this direction, visible at the bottom.

Figure 3.4 shows the second most important direction of variability.

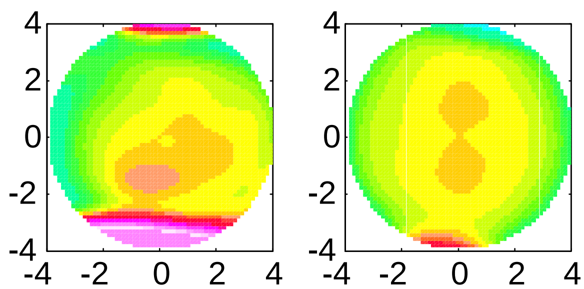

Figure 3.4: Mean image of the population of normal corneas in the center. Representatives of the second principal component direction on either side give an impression of the second direction of greatest variability.

The direction in the 66 dimensional feature space, of the second eigenvector, shown in Figure 3.4, represents a feature of the population that was discussed near Figure 2.1: corneas tend to be steeper either on the top or on the bottom. In this direction, the influence of missing data is quite strong, as indicated by the red and blue regions of extreme curvature at the top and bottom.

Figure 3.5 shows the third direction of variability.

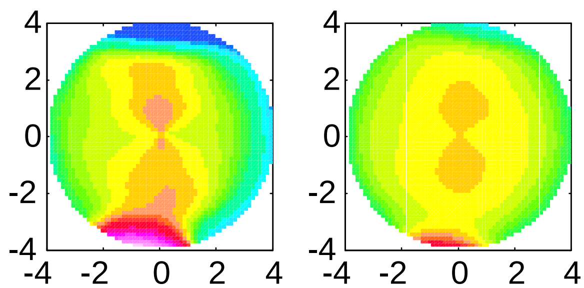

Figure 3.5: Mean image of the population of normal corneas in the center. Representatives of the third principal component direction on either side give an impression of the third direction of greatest variability.

This tertiary variability also seems severely influenced by edge effects, but shows another clinically intuitive aspect of the population: vertical (and stronger than the mean) versus horizontal axes of the astigmatism.

A visually compelling way to study the directions that are suggested by Figures 3.3 - 3.5 is via a movie, which "morphs" between the three images shown. MPEG movies of these can be seen in the files norm100.mpg, norm200.mpg and norm300.mpg, at the same web directory given at the end of section 2.

4 Robust Estimation of Location

A simple example demonstrating the effect of outliers on the mean in two dimensions is shown in Figure 4.1. Note that the single outlier pulls the sample mean actually outside the range of the other observations.

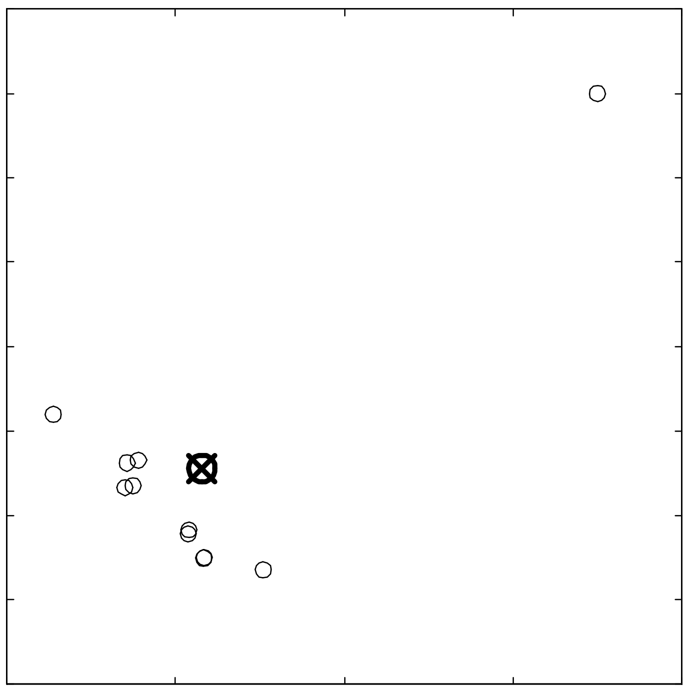

Figure 4.1: Two dimensional example to illustrate effect of outliers on sample mean. Data are shown as circles, sample mean as the heavy circle together with the x.

Simple examples of this type suggest that the impact of outliers may be overcome by simply deleting them. This was not effective for the cornea data set, since as soon as the worst outliers are deleted, other images become the next round of "outliers" (since the missing data problem was endemic to this data set). When these are deleted, then other points appear in this role. Outlier deletion results in loss of too much information, because a very large fraction of the population needs to be deleted.

This motivates a "bounded influence" approach where the goal is to use all of the data, but to allow no single observation to have too much impact.

Much work has been done on the development of such "robust" statistical procedures, see e.g. Hampel, Ronchetti, Rousseeuw and Stahel (1986), Huber (1981), Rousseeuw and Leroy (1987) and Staudte and Sheather (1990).

The robust estimate studied here is the "L^p M-estimate of location", see Section 6.3 of Huber (1981). Given multivariate data $X_1, \ldots, X_n \in \mathbb{R}^d$, this is defined as:

$$\hat{\mu} = \operatorname*{argmin}_{\mu} \sum_{i=1}^{n} \|X_i - \mu\|_2^p ;$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm on $\mathbb{R}^d$. Here we consider only the case $p = 1$, and note that $\hat{\mu}$ may be found as the solution of the equation:

$$0 = \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \|X_i - \mu\|_2^p = \sum_{i=1}^{n} \frac{X_i - \mu}{\|X_i - \mu\|_2} . \tag{1}$$

Insight as to how this location estimate dampens the effect of outliers comes from recognizing that

$$\frac{X_i - \mu}{\|X_i - \mu\|_2} + \mu = P_{Sph(\mu,1)} X_i ;$$

i.e. the projection of X_i onto the sphere centered at μ, with radius 1. Thus the solution of (1) is the solution of

$$0 = \operatorname{avg} \left\{ P_{Sph(\mu,1)} X_i - \mu : i = 1, \ldots, n \right\} .$$

Hence $\hat{\mu}$ may be understood by considering candidate unit spheres centered at μ, projecting the data onto the sphere, then moving the sphere around until the average of the projected values is at the center of the sphere. These ideas are illustrated in Figure 4.2, where the data are the same as in Figure 4.1, again represented as circles.

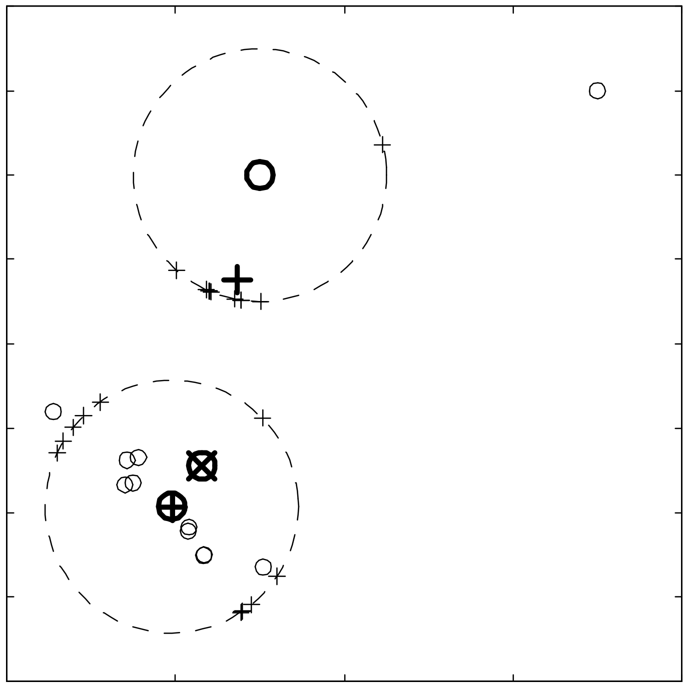

Figure 4.2: Two dimensional example illustrating the L^1 location estimate. Raw data shown as thin circles, projections onto candidate spheres shown as thin plusses. Averages of projections shown as thick plusses, centers of spheres as thick circles. Sample mean shown as thick circle and x.

Note that the upper candidate sphere is not centered near any reasonable "centerpoint of the data". When the data are projected onto the sphere (represented by thin plusses), their centerpoint (shown as the thick plus) is not near the center of the sphere (shown as the thick circle). However, when the sphere is moved until the center of the projected data coincides with the center of the sphere (as for the lower sphere (where the thick plus and the thick circle are the same), that location gives a sensible notion of "center" of the point cloud. In particular, this notion of center gives the outlying point only as much "influence" as the other points receive, it can no longer move the center outside the range of the other points.

This insight makes it clear that in one dimension, $\hat{\mu}$ is any sample median. Hence $\hat{\mu}$ has been called "the spatial median" for higher dimensions. Another consequence is that this location estimate is not unique. However, Milasevic and Ducharme (1987) have shown that in higher dimensions $\hat{\mu}$ is unique, unless all of the data lie in a one dimensional subspace. Other terminology has also been used, e.g. Haldane (1948) called it the "geometric median" and made very early remarks on its robustness properties.

A simple and direct iterative method for calculating $\hat{\mu}$ comes from Gower (1974) or from Section 3.2 of Huber (1981). Given an initial guess, $\hat{\mu}_0$, iteratively define:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}$$

where

$$w_i = \frac{1}{\left\| X_i - \hat{\mu}_{-1} \right\|_2}.$$

This iteration can be understood in terms of Figure 4.2 through the relationship

$$\hat{\mu} = \hat{\mu}_{-1} + \frac{\sum_{i=1}^{n} w_i \left(X_i - \hat{\mu}_{-1} \right)}{\sum_{i=1}^{n} w_i} = \hat{\mu}_{-1} + \frac{\frac{1}{n}\sum_{i=1}^{n} P_{Sph(\hat{\mu}_{-1},1)} X_i - \hat{\mu}_{-1}}{\frac{1}{n}\sum_{i=1}^{n} w_i}.$$

This shows that the next step is in the direction of the vector from the current sphere center $\hat{\mu}_{-1}$ (shown as the circle in Figure 4.2) to the mean of the projected data, $\frac{1}{n}\sum_{i=1}^{n} P_{Sph(\hat{\mu}_{-1},1)} X_i$ (shown as the plus in Figure 4.2). The length of the step is weighted by the harmonic mean distance of the original data to the sphere center (so larger steps are taken when the data are more spread). For the cornea data, and also for a few tests in other high dimensional contexts, we had success taking $\hat{\mu}_0$ to be the sample mean, and iterating until either 20 steps had been taken, or the relative difference between $\hat{\mu}$ and $\hat{\mu}_{-1}$ was less than 10^{-6}. More work needs to be done on verification and fine tuning of these choices, and it may be useful to use a different starting point, such as the coordinate-wise median.

The L^1 estimate of the center of the cornea data from Figure 2.1 is shown in Figure 4.3. Again the calculation is done in the feature space of vectors of Zernike coefficients, but the result is displayed as a curvature image. Note that the impact of the outlying observations, caused by edge effects, is substantially mitigated, when compared to the sample mean, as shown in the center plots of Figures 3.3-3.6.

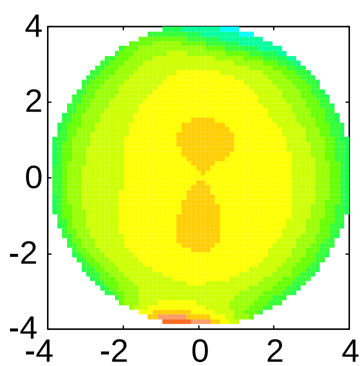

Figure 4.3: Spherical L^1 mean. Missing data effects have less influence than on the sample mean (shown in the centers of Figures 3.2 - 3.6).

The L^1 location estimate is most sensible when the scales of the various dimensions are comparable. However, this is not the case for the cornea data, as shown in Figure 4.4.

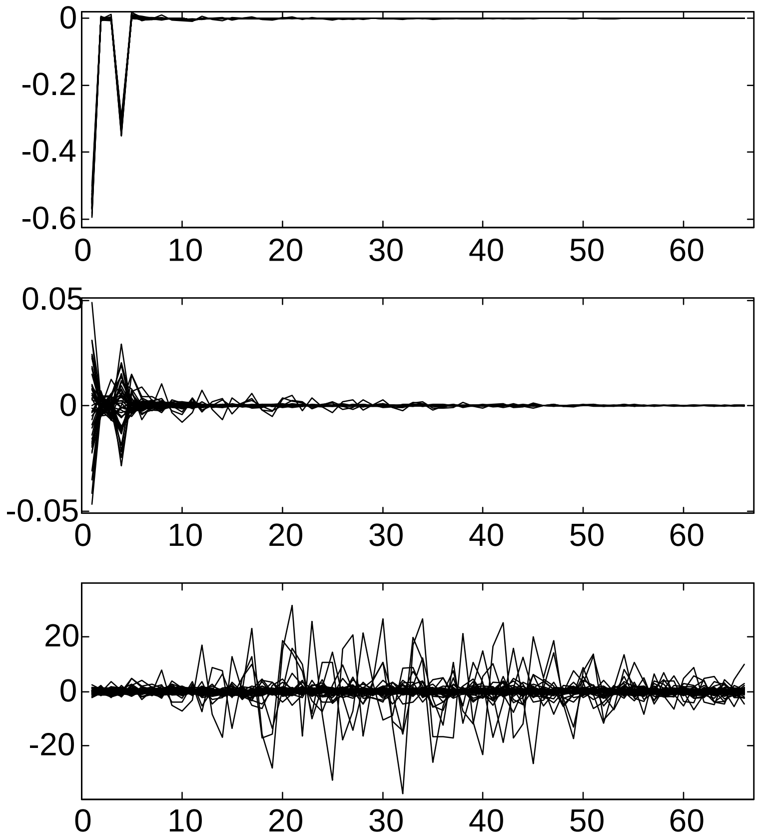

Figure 4.4: Parallel Coordinate Plots of Zernike Coefficients, for population of normal corneas. Top uses the original Zernike scale, middle has coordinate-wise median subtracted, bottom is also divided by coordinate-wise MAD.

The top plot is a parallel coordinate overlay of the raw feature vectors, i.e. the Zernike coefficients, plotted as a function of dimension number (see Section 6 for details). At this scale, it is even impossible to tell how many curves are overlaid, since the dominant features are two very negative coefficients (representing the height and the parabolic curvature components of the eye shapes). The middle plot shows these same feature vectors, with the coordinate-wise median subtracted. Now it is apparent that the data ranges across coordinates differ by orders of magnitude. This effect is similar to the Fourier expansion of a smooth signal having high frequency coefficients that are orders of magnitude smaller than the low frequency coefficients. In this context, it is sensible to modify the L^1 location estimate, by first rescaling each coordinate using some measure of "spread". Here the Median Absolute Deviation from the median is used. The lower plot in figure 4.4 shows the feature vectors when they have been rescaled in this way. The result of modifying the L^1 location estimate, by first dividing by the coordinate-wise MAD, then computing the L^1 location estimate, and finally multiplying by the coordinate-wise MAD, for the cornea data is shown in Figure 4.5.

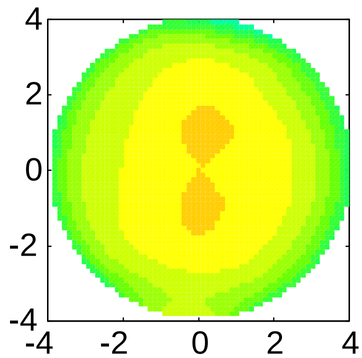

Figure 4.5: Elliptical L^1 mean. Here the impact of the missing data is nearly completely eliminated.

This is an improvement, in terms of even less impact by the outliers, over the "centerpoint" shown in Figure 4.3.

5 Robust Estimation of Spread

While outliers can have a dramatic effect on the mean (the sample first moment), they often have an even stronger impact on traditional measures of scale, such as covariances, since these are based on second moment quantities.

A simple example, showing the potential effect of outliers on PCA is given in Figure 5.1. Note that except for the single outlier, the direction of greatest variability is in the direction of the second and fourth quadrants. But the single outlier completely changes this, so the direction of greatest variability goes towards the first and third quadrants. This is caused by the large effect of the single outlier on the sample covariance matrix.

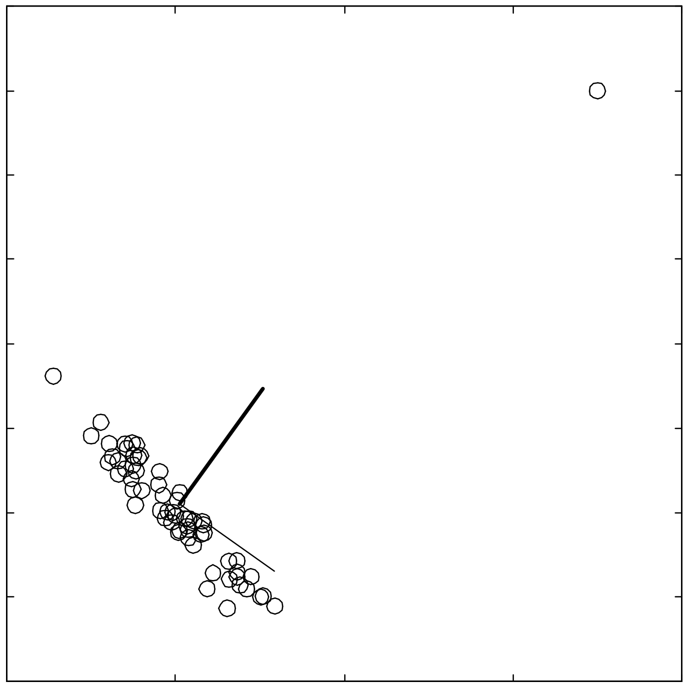

Figure 5.1: Two dimensional example showing how outliers affect PCA. Data points are shown as circles. The first eigenvector direction is shown by the thicker line segment, the second by the thinner. The length of each eigenvector is proportional to the eigenvalue.

Figure 5.2 shows how a single "outlier" can drastically affect the PCA of the simulated family of curves shown in Figure 3.2. A single new data curve is clearly visible in the raw data plot on the upper left. Note that the new data point is not an outlier in any single coordinate direction, but its shape is clearly different from the others (and it is clearly far away in terms of Euclidean distance).

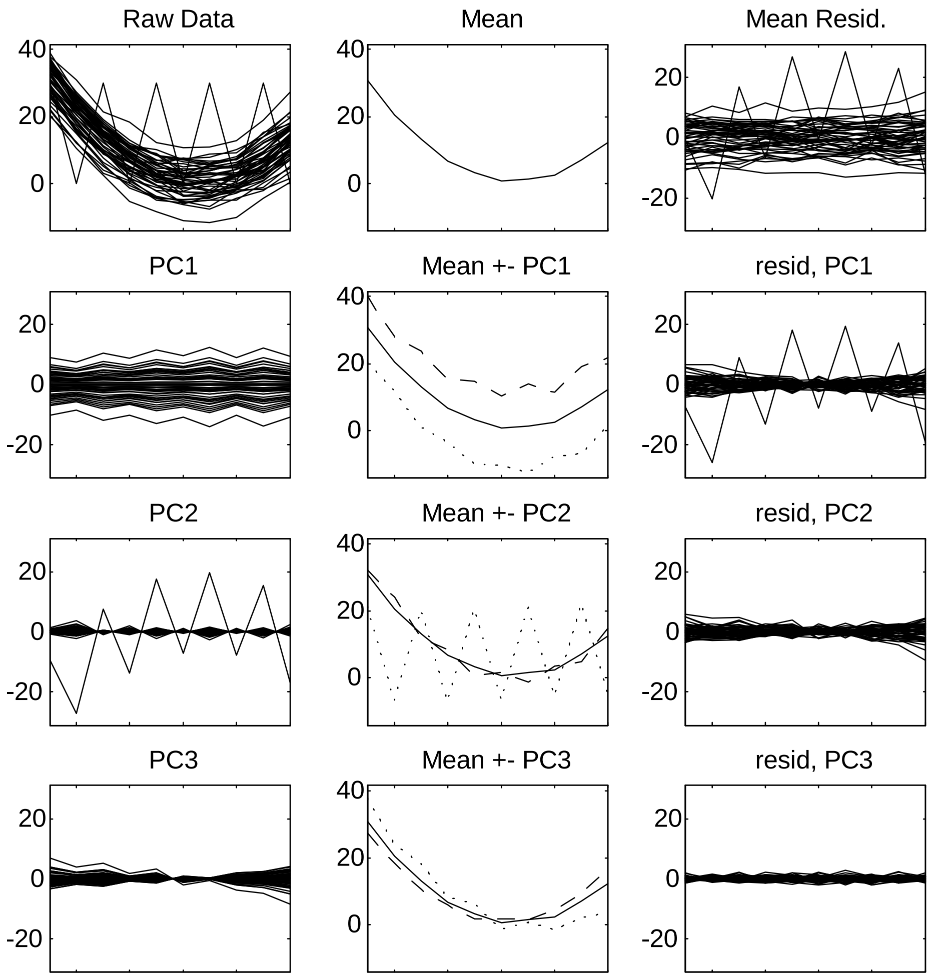

Figure 5.2: PCA for data of Figure 3.2 with an outlier added.

The new observation in Figure 5.2 has negligible impact on the mean, as shown in the center plot on the top row. It has only a small impact on the first principle component direction, as shown in the second row, although it is visible in terms of the "ripples" that can be seen. But this single observation clearly dominates the second PCA direction, as shown in the third row. Because of this major impact, the important second feature of the data, the "tilting" shown in the bottom row of Figure 3.2, now only appears in the third PCA direction. This shows how "outliers" can hide important features of the data. It also shows that a point can be an outlier, even when none of its coordinates is unusually large, which is a perhaps surprising property of high dimensional data (in the spirit of the fact that a point on the vertex of the unit cube in d dimensions is distance $\sqrt{d}$ from the origin).

Figure 5.3 shows how the spherical PCA approach gives a bounded influence version of PCA, for the same simple data set (point cloud oriented towards the second and fourth quadrants, with a single outlier) as in Figure 5.1. The main idea is that of the projection approach to L^1 M-estimation: project the data onto a sphere to reduce the effect of outliers.

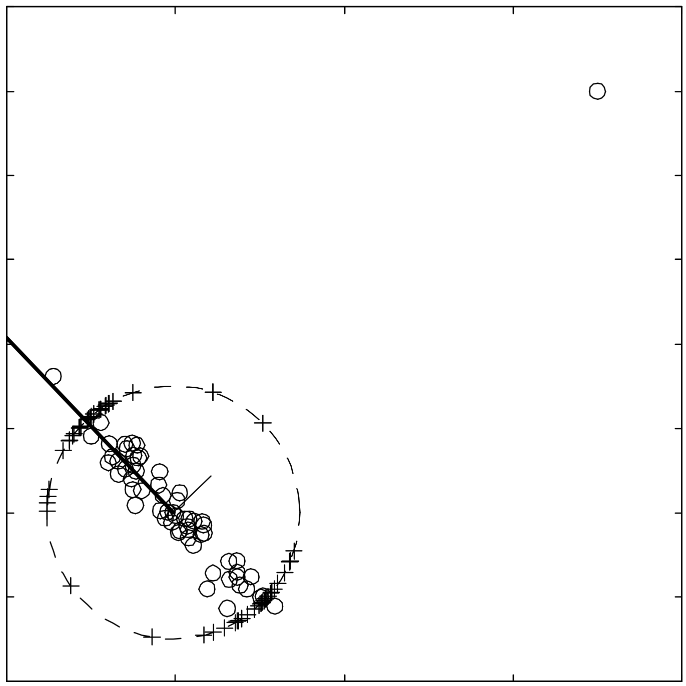

Figure 5.3: Two dimensional example showing how spherical PCA downweights the influence of outliers. Data points are shown as circles, projections onto the shown sphere are shown as plusses.. The first eigenvector direction of the projected data is shown by the thicker line segment, the second by the thinner. The length of each eigenvector is proportional to the eigenvalue.

In Figure 5.3, the circles are the raw data, and the result of projecting them onto a sphere centered at the L^1 M-estimate are shown as the thin plusses. Spherical PCA is based on the eigenanalysis of the covariance matrix of these projected data. As for the location estimate, the influence of the outlying observation is greatly reduced.

Figure 5.4, shows the result of a spherical PCA for the data set with the outlier shown in Figure 5.2.

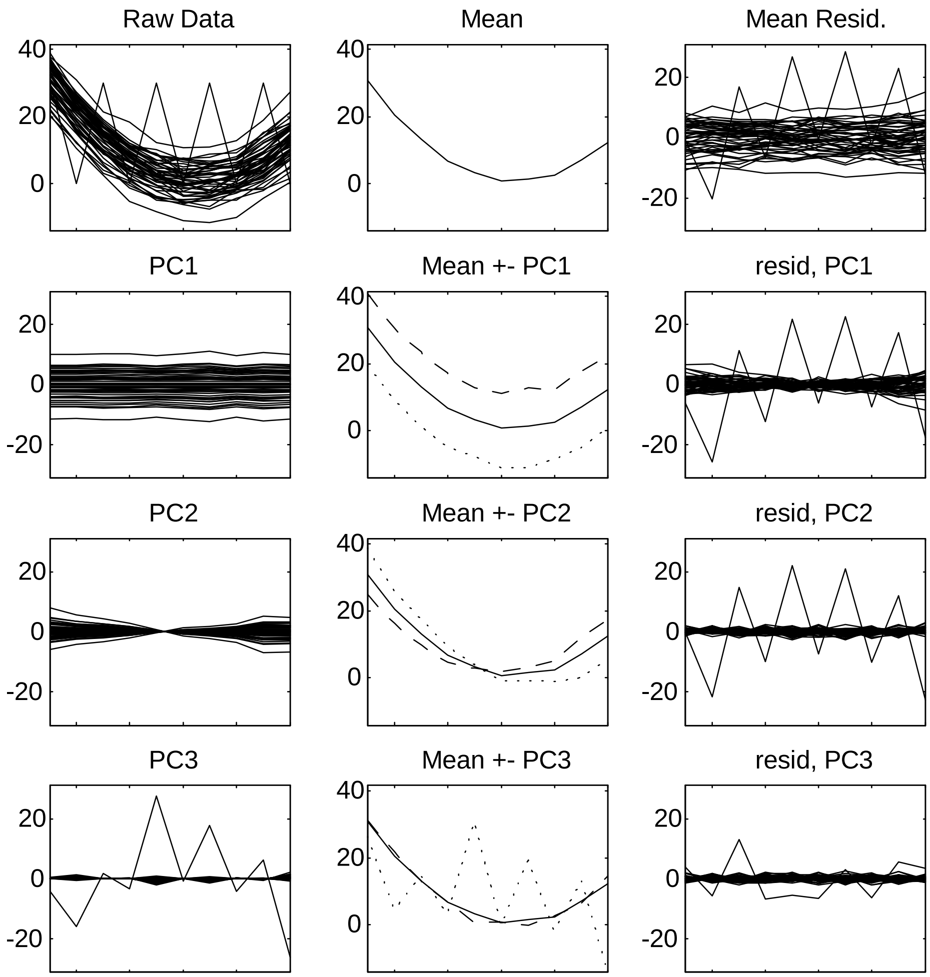

Figure 5.4: Spherical PCA for data of Figure 5.2.

In Figure 5.4, the outlying observation now has almost no effect on the first PCA direction (shown in the second row), i.e. the wiggliness in the second row of Figure 5.2 is gone. But more important, the second PCA direction (shown in the third row) now shows the important tilting feature of the bulk of the data, and the outlier only appears in the third PCA direction. This shows how spherical PCA can limit the effect of outliers on this type of analysis.

As noted near the end of Section 4, projection onto a sphere may not be completely effective if the data are on widely different scales in different coordinate directions. The improvements gained by changing the sphere to a suitable ellipse are present in the present situation also. Visual insight into the corresponding elliptical variation of PCA is given in Figure 5.5.

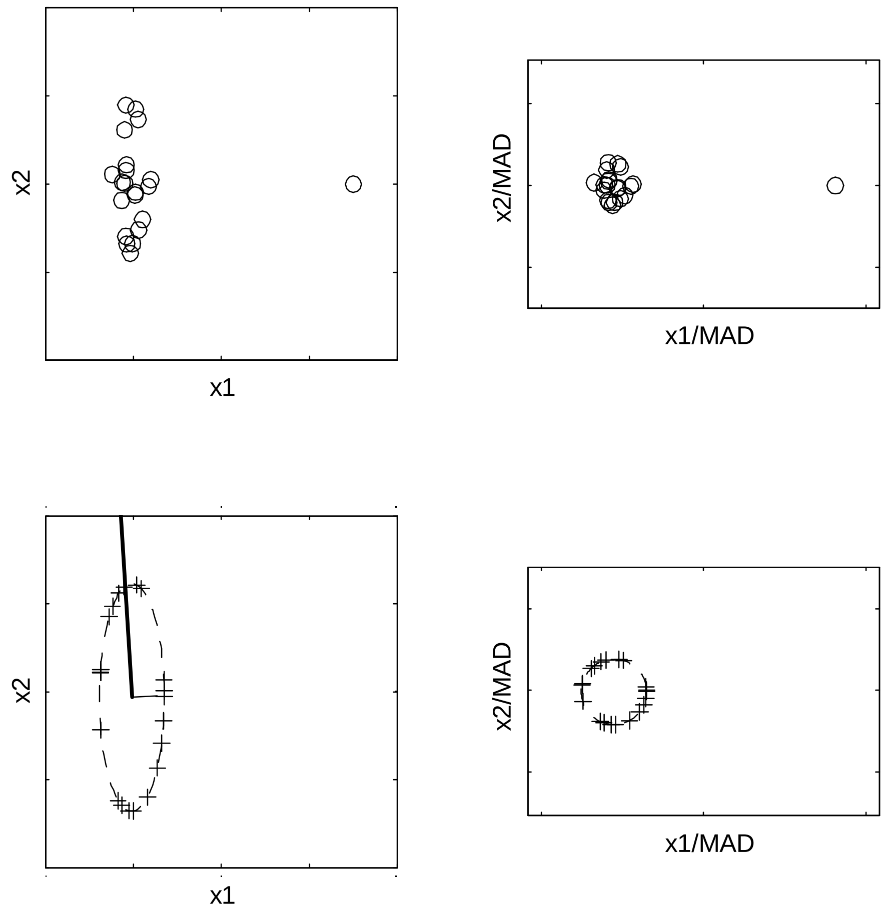

Figure 5.5: Two dimensional example showing how elliptical PCA correctly accounts for differing axis scaling. Data points are shown as circles (top row), projections onto the shown sphere (or induced ellipse) are shown as plusses (bottom row). Left hand plots are the original scale, right plots are rescaled by the sample Median Absolute Deviations. Elliptical eigenvector directions are shown in the lower left.

The upper left plot in Figure 5.5 shows a simple data set where elliptical PCA is a substantial improvement over spherical PCA. The upper right plot shows the results of transforming the data so that the MAD of each coordinate axis is 1. The vertical axis has been substantially compressed, so that the bulk of the data now look spherical. Projection onto the sphere is now done on this scale, as shown in the lower right plot. Finally the data are transformed back to the original scale, as shown in the lower left plot. Note that now the projected data lie on the surface of an ellipse, that appropriately reflects the different scalings of the axes.

Figure 4.4 suggests that Elliptical PCA is appropriate for the cornea data, and we observed the expected improvements over Spherical PCA (not shown here to save space). The results are shown in the following figures. Again the main idea is to do the numerics of the statistical analysis in the 66 dimensional feature space of Zernike coefficient vectors, but to represent the results in the visually intuitive space of curvature maps.

Figure 5.6 is an improved version of Figure 3.3, showing the dominant direction.

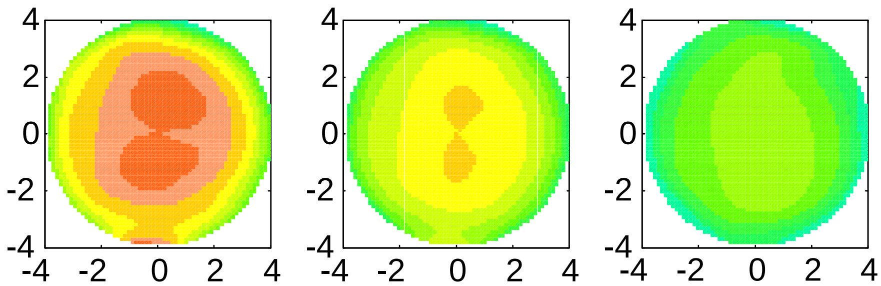

Figure 5.6: Center is Elliptical L^1 mean, direction shows first eigenvector of Elliptical PCA.

Figure 5.6 has the same basic lessons as in Figure 3.3, except that the stronger vertical astigmatism on the left is now more clear, and the distracting boundary behavior is nearly completely gone.

Figure 5.7 is an improved version of Figure 3.4.

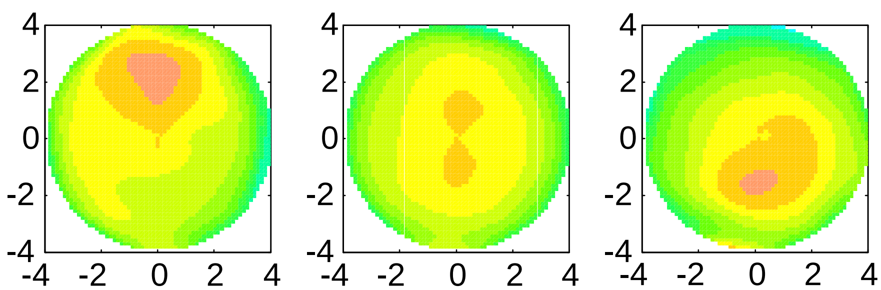

Figure 5.7: Center is Elliptical L^1 mean, direction shows second eigenvector of Elliptical PCA.

Figure 5.7 has nearly completely eliminated the very strong boundary effects from Figure 3.4. It also shows the steeper top and bottom regions more clearly (in a way that looks more like these features as seen in Figure 2.1).

Figure 5.8 is an improved version of Figure 3.5.

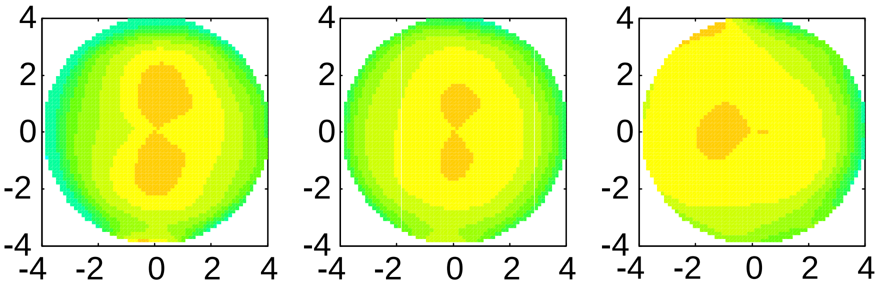

Figure 5.8: Center is Elliptical L^1 mean, direction shows third eigenvector of Elliptical PCA.

Figure 5.8 has also essentially eliminated the very strong missing data artifacts visible in Figure 3.5. It also makes it more clear that this direction is representing differing axes of the astigmatism.

MPEG movie versions of the Figure 5.6-5.8 are available at the web address mentioned at the end of Section 2, in the files norm122.mpg, norm222.mpg, norm322.mpg.

A final comment concerns the relationship between PCA and Gaussian data. Some have offered the opinion that the Gaussian assumption is important to PCA. This reservation is well justified when distribution theory is used, for example in classical multivariate hypothesis testing. However, it is not necessarily a problem when the goal, as here, is simply to find "interesting directions". The problems with outliers shown in Section 3 could be viewed in terms of "non-Gaussianity" of the data, but the solution recommended in Section 5 works effectively in a non-Gaussian way.

6 Appendix: Zernike basics

The Zernike polynomials coefficients are chosen to summarize the cornea data because this basis has natural interpretation in ophthalmology. The Zernike polynomials are orthonormal on the unit sphere, and are radially symmetric. Zernike polynomials are a combination of two components. One component is a Fourier component in the angular direction. The other is a Jacobi polynomial in the radial direction. The general form of the Zernike polynomials (see Schwiegerling, et al. 1995) is defined as:

$$Z_n^{\pm m}(r;\theta) = \begin{cases} \sqrt{2(n+1)}\,R_n^m(r)\cos(m\theta) & \text{for } +m \\ \sqrt{2(n+1)}\,R_n^m(r)\sin(m\theta) & \text{for } -m \\ \sqrt{(n+1)}\,R_n^m(r) & \text{for } m=0 \end{cases}$$

where n is the polynomial order, m is the Fourier order, and $R_n^m(r)$ is the representation for the Jacobi polynomial.

The Jacobi polynomial is given by:

$$R_n^m(r) = \sum_{s=0}^{\frac{1}{2}(n-m)} \frac{(-1)^s (n-s)!}{s!\left(\frac{n+m}{2}-s\right)!\left(\frac{n-m}{2}-s\right)!} r^{n-2s}.$$

An easier computational formula (Born and Wolf, 1980) for $R_n^m(r)$ is:

$$R_n^m(r) = \frac{1}{\left(\frac{n-m}{2}\right)! \, r^m} \left\{ \frac{d}{d(r^2)} \right\}^{\frac{n-m}{2}} \left\{ (r^2)^{\frac{n+m}{2}} (r^2-1)^{\frac{n-m}{2}} \right\}.$$

References

[1] Born, M. and Wolf, E. (1980) Principles of optics: electromagnetic theory of propagation, interference and diffraction of light. Pergamon Press, New York.

[2] Cootes, T. F., Hill, A., Taylor, C. J. and Haslam, J. (1993) The use of active shape models for locating structures in medical images, Information Processing in Medical Imaging, H. H. Barret and A. F. Gmitro, eds., Lecture Notes in Computer Science 687, 33-47, Springer Verlag, Berlin.

[3] Devijver, P. A. and Kittler, J. (1982) Pattern Recognition: A Statistical Approach, Prentice Hall, London.

[4] Dryden, I. L. and Mardia, K. V. (1998) Statistical Shape Analysis, Wiley: New York.

[5] Fan, J. and Lin S. K. (1998) Test of significance when the data are curves, Journal of the American Statistical Association, 93, 1007-1021.

[6] Gower, J.C. (1974). The mediancentre. Applied Statistics, 23, 466-470.

[7] Haldane, J. B. S. (1948) Note on the median of a multivariate distribution. Biometrika, 35, 414-415.

[8] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) Robust statistics: the approach based on influence functions, Wiley: New York.

[9] Huber, P. J. (1981) Robust Statistics, Wiley: New York.

[10] Inselberg, A. (1985) The plane with parallel coordinates, The Visual Computer, 1, 69-91.

[11] Kelemen, A., Szekely, G. and Gerig, G. (1997) Three dimensional model-based segmentation, TR-178 Technical Report Image Science Lab, ETH Zurich.

[12] Milasevic, P. and Ducharme, G. R. (1987) Uniqueness of the spatial median, Annals of Statistics, 15, 1332-1333.

[13] Ramsey, J. O. and Silverman, B. W. (1997) Functional Data Analysis, Springer Verlag, New York.

[14] Rousseeuw, P. J. and Leroy, A. M. (1987) Robust regression and outlier detection, Wiley: New York.

[15] Staudte, R. G. and Sheather, S. J. (1990) Robust estimation and testing, Wiley: New York

[16] Schwiegerling, J., Greivenkamp, J. E., and Miller, J. M. (1995) Representation of videokeratoscopic height data with Zernike polynomials, Journal of the Optical Society of America, Series A, 12, 2105-2113.

[17] Wegman, E. J. (1990) Hyperdimensional data analysis using parallel coordinates, Journal of the American Statistical Association, 85, 664-675.