Sociedad de Estadística e Investigación Operativa
Test (1999) *Vol.* **8**, *No.* 1, pp. 1–65

Robust principal component analysis for functional data

N. Locantore and J.S. Marron*
Department of Statistics,
University of North Carolina, U.S.A.
D.G. Simpson
Department of Statistics,
University of Illinois, U.S.A.
N. Tripoli
Department of Ophthalmology,
University of North Carolina, U.S.A.
J.T. Zhang
Department of Statistics,
University of North Carolina, U.S.A.
K.L. Cohen
Department of Ophthalmology,
University of North Carolina, U.S.A.

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.

Key Words: Cornea curvature maps, functional data, principal components analysis, robust statistics, spherical PCA, Zernike basis.
AMS subject classification: 62H99

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

*Correspondence to: J.S. Marron, Department of Statistics, University of North Carolina, Chapel Hill, N.C. 27599-3260, U.S.A.

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.

Two such images are shown in Figure 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 2 shows a curvature map from a patient with the disease of keratoconus, in which the cornea grows into a highly curved cone shape

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 the Appendix.

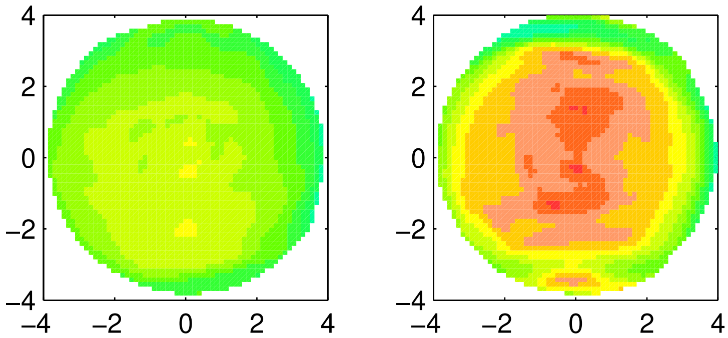

Figure 1: *Two corneal images showing radial curvature. The left shows relatively constant curvature. The right shows more curvature near the center, and a marked vertical astigmatism.*

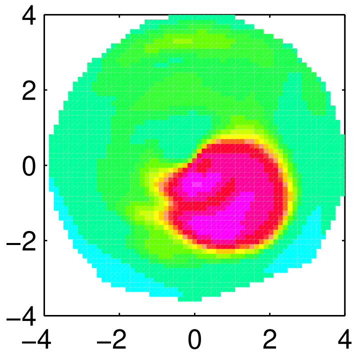

Figure 2: *Radial curvature of a cornea with Kerataconus. The red region is a cone of high curvature.*

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 the Appendix.

The results of Zernike feature vector summarization, for the images of Figure 1, as well as several others, are shown in Figure 3. 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.

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 3 are a subset, chosen to represent the most important features. Note that the raw curvature images from Figure 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

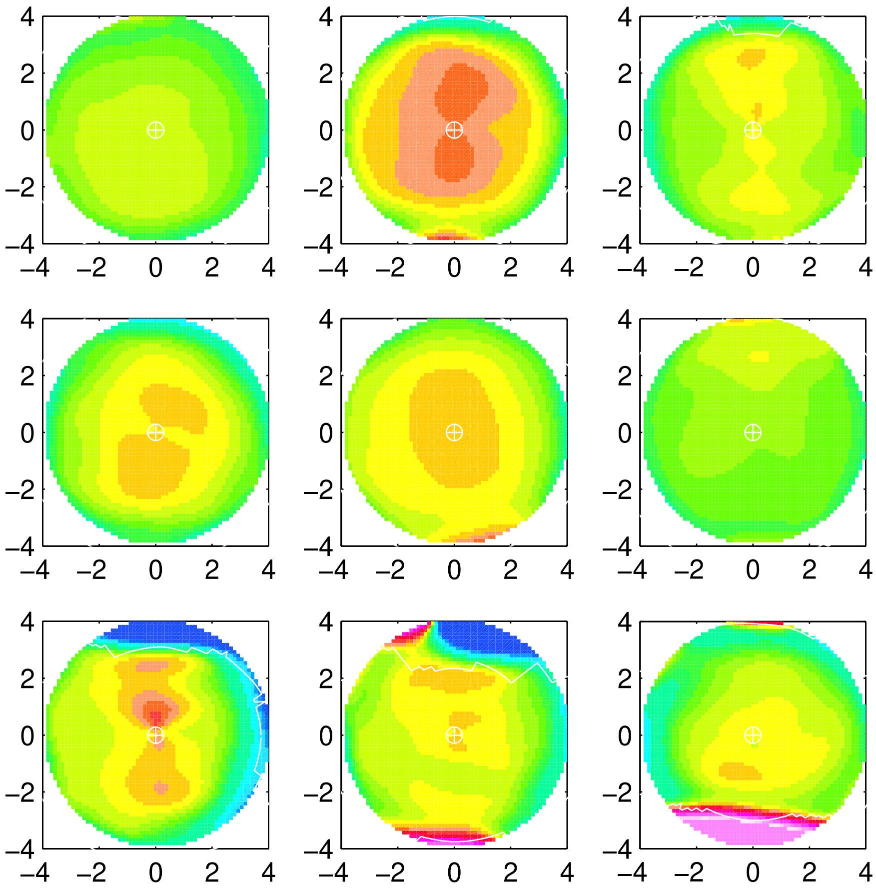

Figure 3: Zernike reconstructions of some normal cornea images.

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/cornea-robust/`, 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 4. 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 4 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

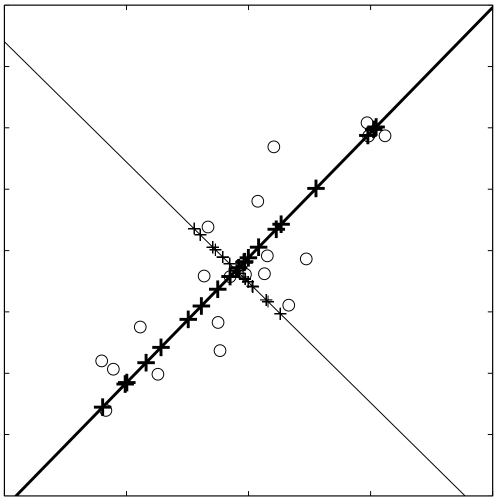

Figure 4: Two dimensional example illustrating PCA. First eigenvector direction (and projections of the data) shown with a thick line (thick plusses). Second eigenvector direction (and projections of the data) shown with a thin line (thin plusses)

of the data.

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 5.

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

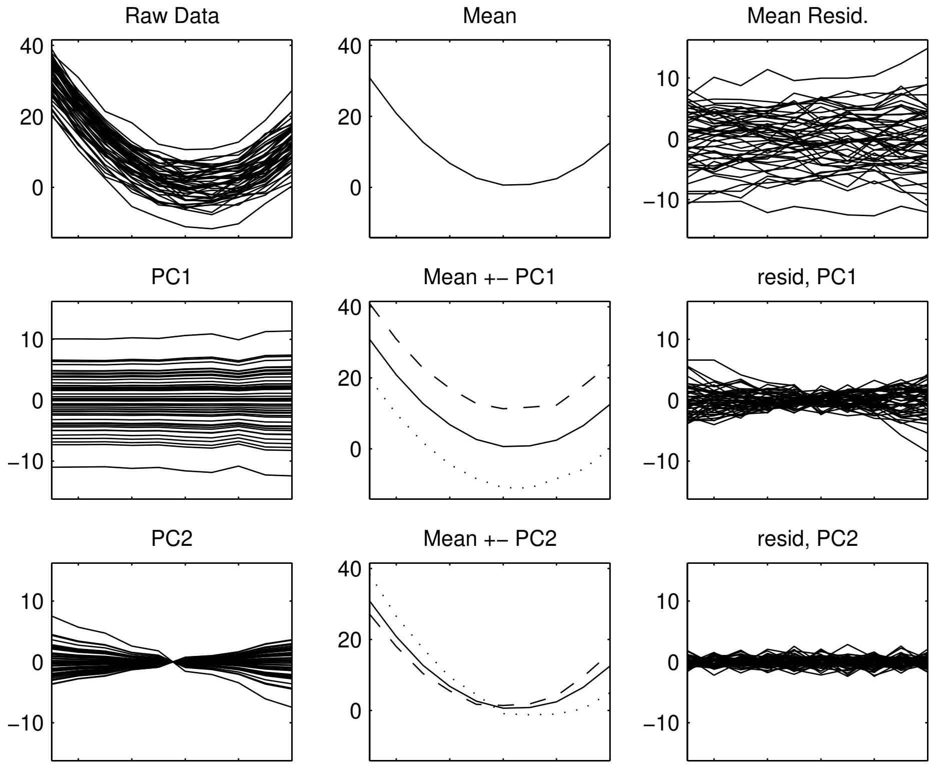

Figure 5: "Curves as data" example illustrating PCA. First row shows results of "recentering". Second row shows strongest component of variability. Third row shows second most important component.

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 4. Two representations of the set of the projections (i.e. the heavy plusses in Figure 4) 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 4). 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 5), it does not give an intuitively useful view when the data are images (as in Figure 3) or more complex structures that are not usefully overlaid on a single plot. For example note that Figure 11, 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 3). 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 5) are no longer useful, representations of the directions in terms of extremes, as shown in the center column of Figure 5, are quite useful. Studying the mean, together with extremes in each direction, gives insight into that "direction of variability". Figure 6 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 3.

The center panel of Figure 6 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 3. The left and right panels of Figure 6 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 7 shows the second most important direction of variability.

The direction in the 66 dimensional feature space, of the second eigenvector, shown in Figure 7, represents a feature of the population that was discussed near Figure 3: 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 8 shows the third direction of 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 6-8 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 9. Note that the single outlier pulls the sample mean actually outside the range of the other observations.

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 Sta-

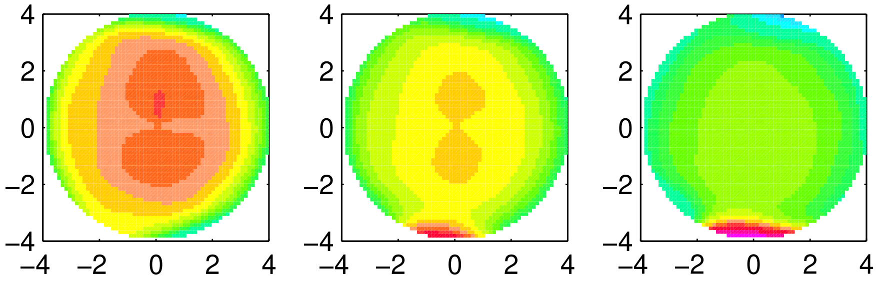

Figure 6: *Mean image of the population of normal corneas in the center. Representatives of the first principal component direction on either side give an impression of the direction of greatest variability.*

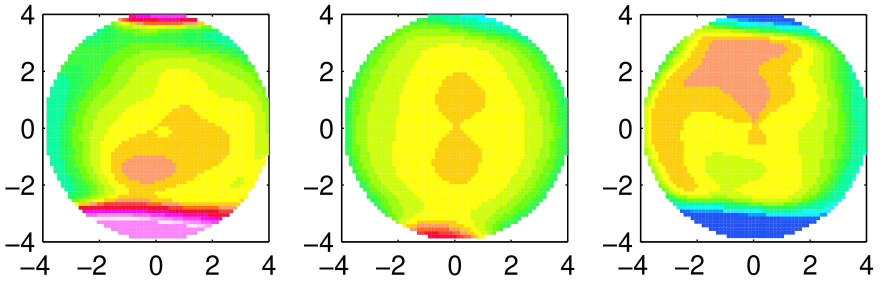

Figure 7: *Mean image of the population of normal corneas in the center. Representatives of the second principal component direction on either side give an impression of the second direction of greatest variability.*

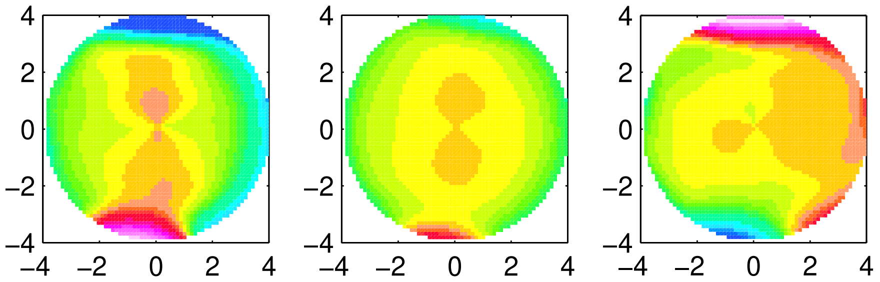

Figure 8: *Mean image of the population of normal corneas in the center. Representatives of the third principal component direction on either side give an impression of the third direction of greatest variability.*

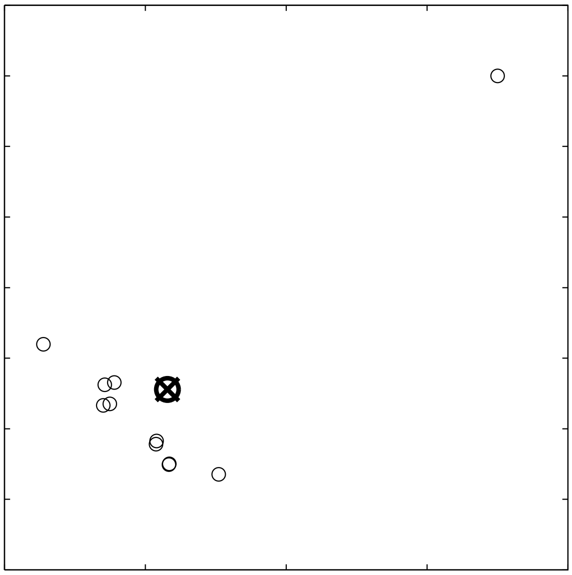

Figure 9: *Two dimensional example to illustrate effect of outliers on sample mean. Data are shown as circles, sample mean as the heavy circle together with the x.*

hel (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:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \|X_i - \theta\|_2^p ,$$

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

$$0 = \frac{\partial}{\partial\theta} \sum_{i=1}^{n} \|X_i - \theta\|_2^p = \sum_{i=1}^{n} \frac{X_i - \theta}{\|X_i - \theta\|_2}. \tag{4.1}$$

Insight as to how this location estimate dampens the effect of outliers

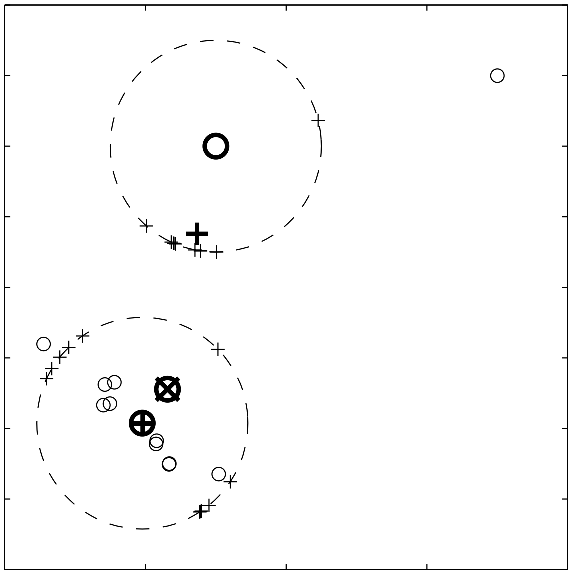

Figure 10: Two dimensional example illustrating the L^1 location estimate. Raw data shown as thin circles, projections onto candidate spheres shown as thin plusses. Averages of projections shown as thick plusses, centers of spheres as thick circles. Sample mean shown as thick circle and x.

comes from recognizing that

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

i.e. the projection of X_i onto the sphere centered at θ, with radius 1. Thus the solution of (4.1) is the solution of

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

Hence $\widehat{\theta}$ may be understood by considering candidate unit spheres centered at θ, 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 10, where the data are the same as in Figure 9, again represented as circles.

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, $\widehat{\theta}$ is any sample median. Hence $\widehat{\theta}$ 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 $\widehat{\theta}$ 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 $\widehat{\theta}$ comes from Gower (1974) or from Section 3.2 of Huber (1981). Given an initial guess, $\widehat{\theta}_0$, iteratively define:

$$\widehat{\theta}_\ell = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$

where

$$w_i = \frac{1}{\left\| X_i - \widehat{\theta}_{\ell-1} \right\|_2}.$$

This iteration can be understood in terms of Figure 10 through the relationship

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

This shows that the next step is in the direction of the vector from the current sphere center $\widehat{\theta}_{\ell-1}$ (shown as the circle in Figure 10) to the mean of the projected data, $\frac{1}{n}\sum_{i=1}^n P_{Sph(\theta_{\ell-1},1)} X_i$ (shown as the plus in Figure 10). 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 $\widehat{\theta}_0$ to be the sample mean, and iterating until either 20 steps had been taken, or the relative difference between $\widehat{\theta}_\ell$ and $\widehat{\theta}_{\ell-1}$ 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 3 is shown in Figure 12. 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 6-8.

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 11.

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 Appendix 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 11 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 13.

This is an improvement, in terms of even less impact by the outliers, over the "centerpoint" shown in Figure 12.

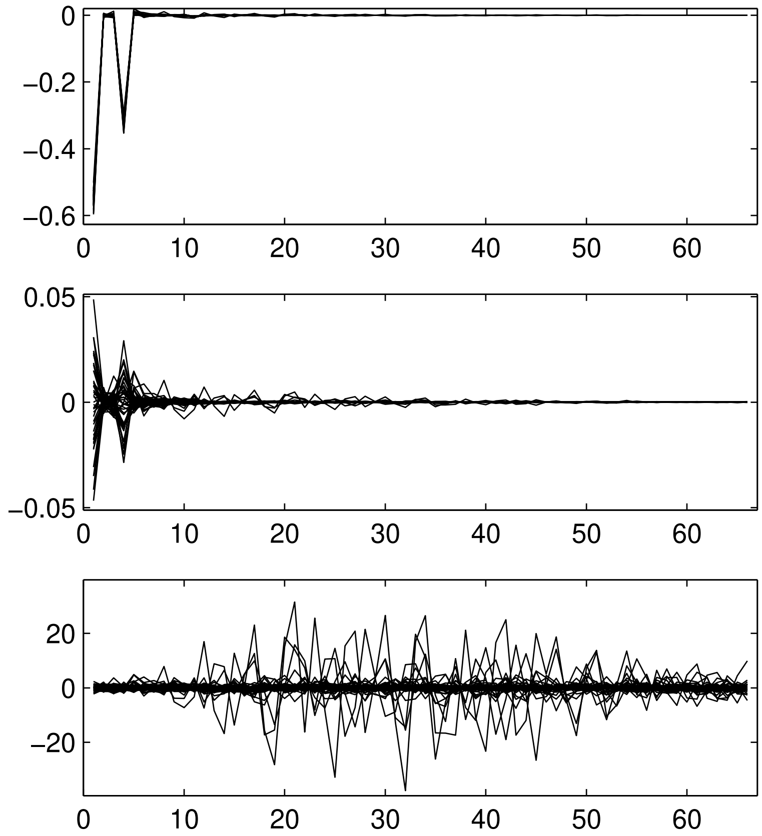

Figure 11: *Parallel Coordinate Plots of Zernike Coefficients, for population of normal corneas. Top uses the original Zernike scale, middle has coordinate-wise median subtracted, bottom is also divided by coordinate-wise MAD.*

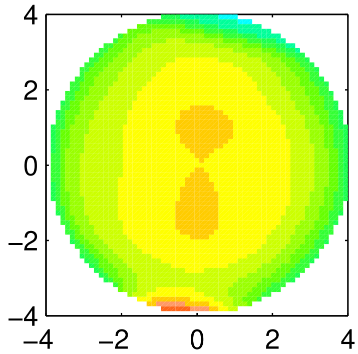

Figure 12: *Spherical L^1 mean. Missing data effects have less influence than on the sample mean (shown in the centers of Figures 6-8.*

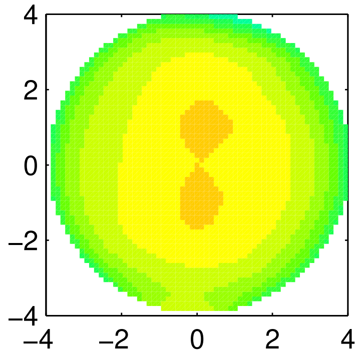

Figure 13: *Elliptical L^1 mean. Here the impact of the missing data is nearly completely eliminated.*

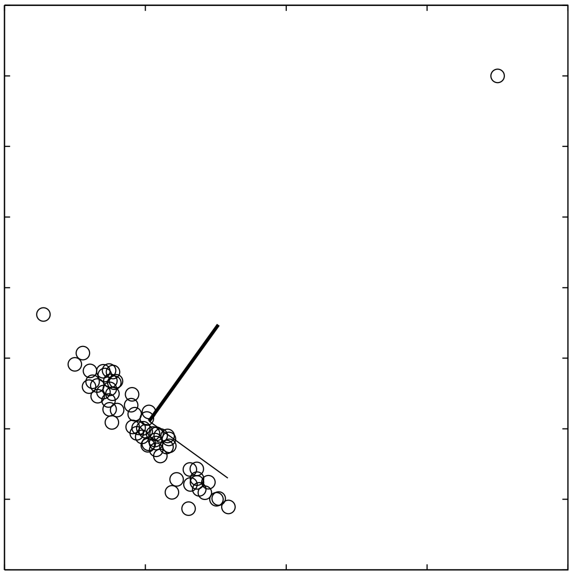

Figure 14: *Two dimensional example showing how outliers affect PCA. Data points are shown as circles. The first eigenvector direction is shown by the thicker line segment, the second by the thinner. The length of each eigenvector is proportional to the eigenvalue.*

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 14. 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 15 shows how a single "outlier" can drastically affect the PCA of the simulated family of curves shown in Figure 5. 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).

The new observation in Figure 15 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 5, 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 16 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 14. 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.

In Figure 16, 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 17, shows the result of a spherical PCA for the data set with the outlier shown in Figure 15.

In Figure 17, 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 15 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

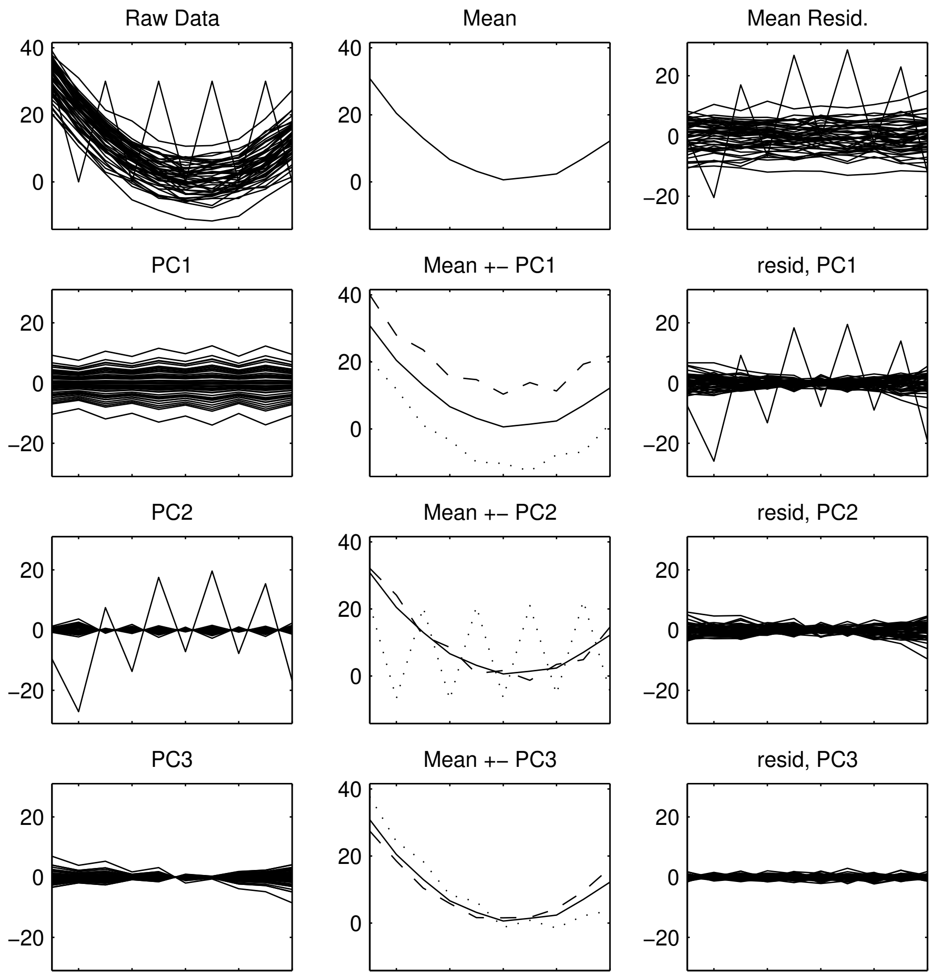

Figure 15: PCA for data of Figure 5 with an outlier added.

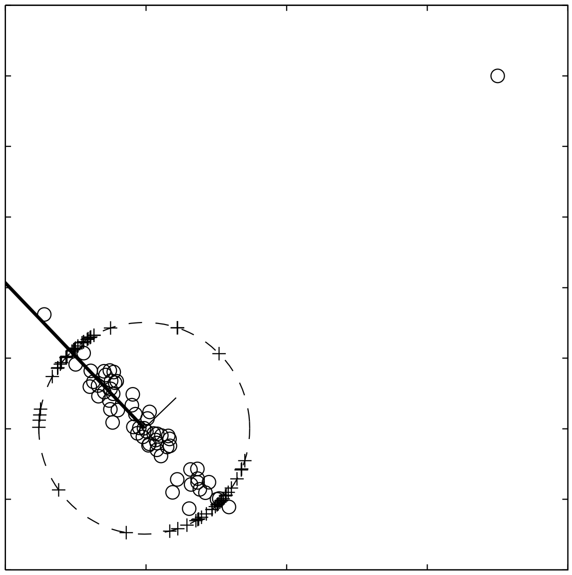

Figure 16: *Two dimensional example showing how spherical PCA downweights the influence of outliers. Data points are shown as circles, projections onto the shown sphere are shown as plusses.. The first eigenvector direction of the projected data is shown by the thicker line segment, the second by the thinner. The length of each eigenvector is proportional to the eigenvalue.*

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 18.

The upper left plot in Figure 18 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 11 suggests that Elliptical PCA is appropriate for the cornea

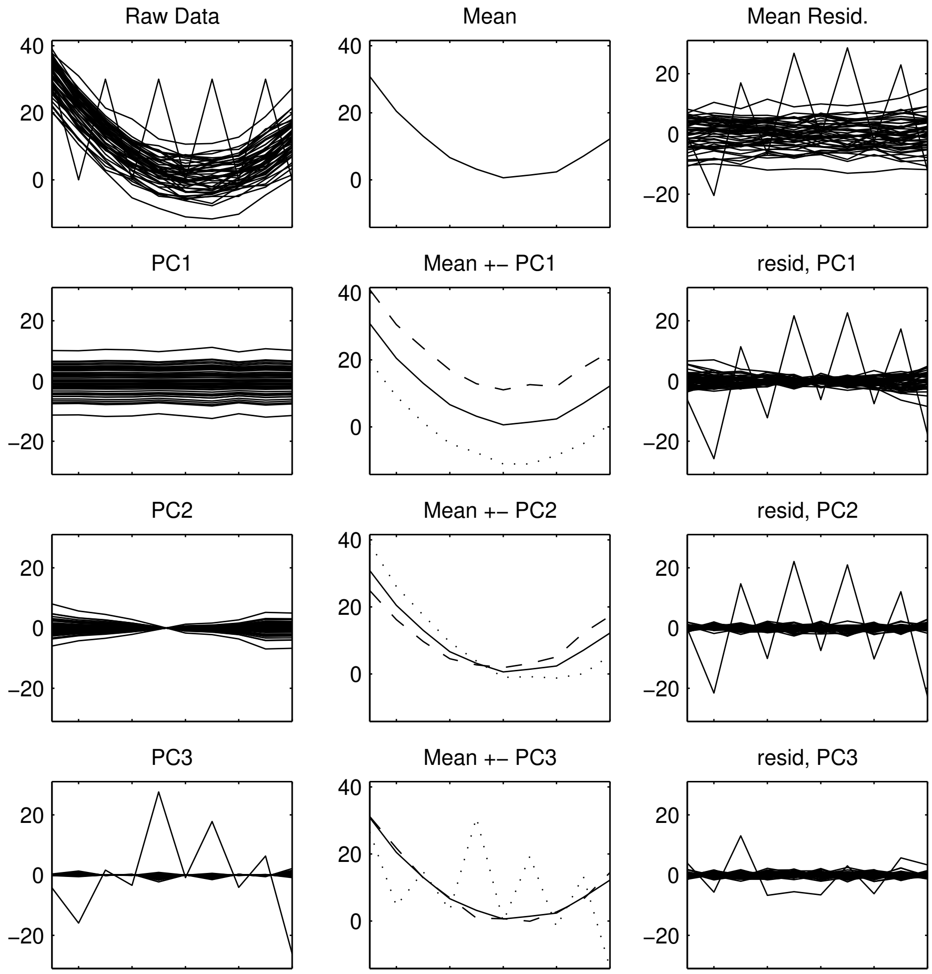

Figure 17: Spherical PCA for data of Figure 15.

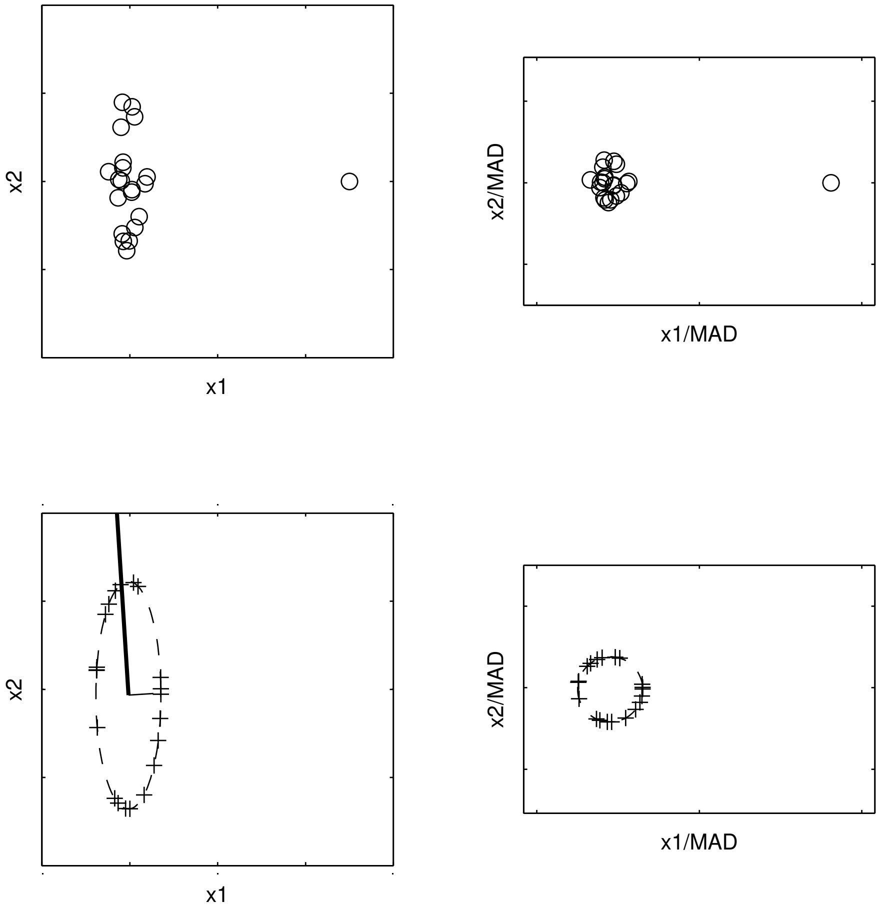

Figure 18: Two dimensional example showing how elliptical PCA correctly accounts for differing axis scaling. Data points are shown as circles (top row), projections onto the shown sphere (or induced ellipse) are shown as plusses (bottom row). Left hand plots are the original scale, right plots are rescaled by the sample Median Absolute Deviations. Elliptical eigenvector directions are shown in the lower left.

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 19 is an improved version of Figure 6, showing the dominant direction.

Figure 19 has the same basic lessons as in Figure 6, except that the stronger vertical astigmatism on the left is now more clear, and the distracting boundary behavior is nearly completely gone.

Figure 20 is an improved version of Figure 7.

Figure 20 has nearly completely eliminated the very strong boundary effects from Figure 7. It also shows the steeper top and bottom regions more clearly (in a way that looks more like these features as seen in Figure 3).

Figure 21 is an improved version of Figure 8.

Figure 21 has also essentially eliminated the very strong missing data artifacts visible in Figure 8. It also makes it more clear that this direction is representing differing axes of the astigmatism.

MPEG movie versions of the Figure 19-21 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.

Appendix: Zernike basics

The Zernike polynomials coefficients are chosen to summarize the cornea data because this basis has natural interpretation in ophthalmology. The

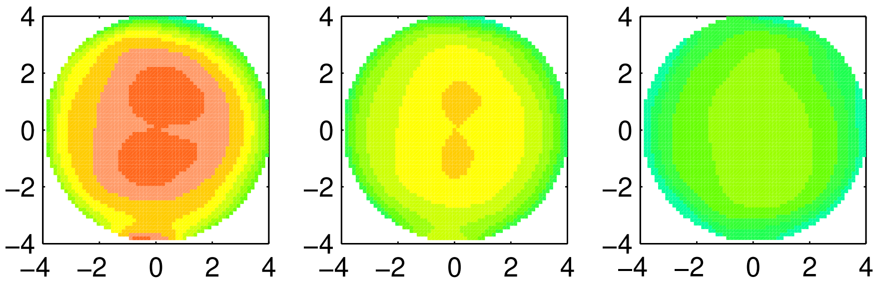

Figure 19: *Center is Elliptical L^1 mean, direction shows first eigenvector of Elliptical PCA.*

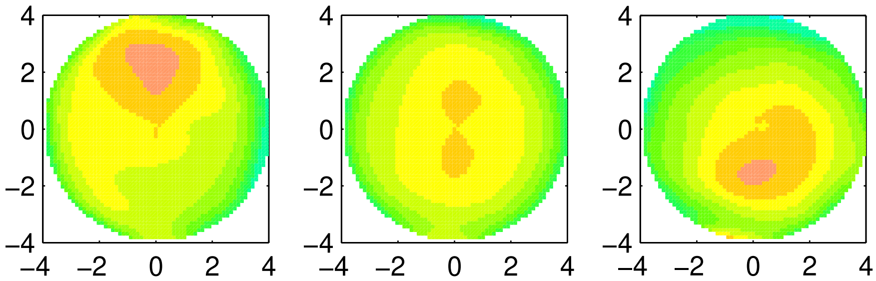

Figure 20: *Center is Elliptical L^1 mean, direction shows second eigenvector of Elliptical PCA.*

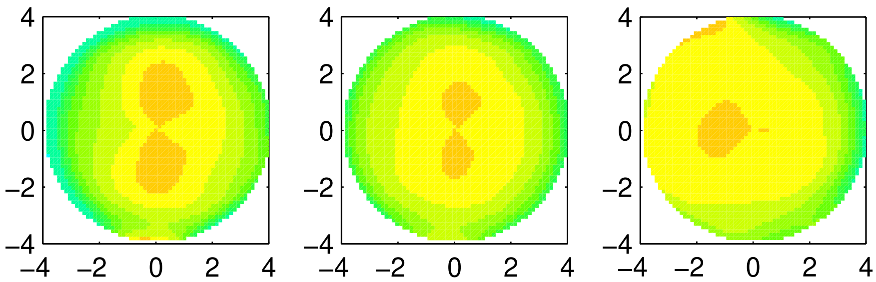

Figure 21: *Center is Elliptical L^1 mean, direction shows third eigenvector of Elliptical PCA.*

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:

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

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

Born, M. and E. Wolf (1980). *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light.* Pergamon Press, New York.

Cootes, T.F., A. Hill, C.J. Taylor and J. Haslam (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, vol. 687. Springer Verlag, Berlin, 33-47.

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

Dryden, I.L. and K.V. Mardia (1998). *Statistical Shape Analysis.* Wiley, New York.

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

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

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

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

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

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

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

Ramsay, J.O. and B.W. Silverman (1997). *Functional Data Analysis.* Springer Verlag, New York.

Rousseeuw, P.J. and A.M. Leroy (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

Staudte, R.G. and S.J. Sheather (1990). *Robust Estimation and Testing.* Wiley, New York.

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

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

DISCUSSION

Graciela Boente
Universidad de Buenos Aires and CONICET, Argentina.

Ricardo Fraiman
Universidad de Buenos Aires and Universidad de San Andrés, Argentina

This article presents several interesting ideas for dimension reduction of complicated data structures. When data are curves, instead of finite dimensional vectors, Ramsay and Silverman (1997) described an extension of principal components analysis, called functional principal component analysis.

Non–smooth principal components of functional data were considered initially by Dauxois, Pousse and Romain (1982). Further analysis of this problem has been developed by Besse and Ramsay (1986), Rice and Silverman (1991), Ramsay and Dalzell (1991), Pezzulli and Silverman (1993), Silverman (1996) and Ramsay and Silverman (1997), where smooth principal components for functional data, based on roughness penalty methods, were considered. Boente and Fraiman (1998) studied a kernel–based approach to this problem. Several examples and applications can be found in these references.

The authors' approach goes beyond that direction and provides many practical insights. They consider more complicated data structures, like images, summarizing them through "feature vectors". A second issue considered in this article is that of robust methods for this high dimensional problem. We expect this article to stimulate more research in the area.

The authors analyze the shape of the outside surface of the cornea measured through the 43 images given by a corneal topography. Their method may be summarized as follows:

i) Smooth dimensional reduction through Zernike decomposition (compression method). A least square fitting of the initial 6912–dimensional vectors to the first 66 coefficients of the Zernike orthogonal basis, gives the 66–dimensional "feature vectors".

ii) Find a robust center for the 43 "feature vectors". The authors consider the spatial median.

iii) Apply the "spherical" principal component analysis proposed by the authors; which is to perform a principal component analysis to the projected data on the unit sphere (centered at the spatial median) in $\mathbb{R}^{66}$. In order to deal with coordinates measured in different scales, they propose an alternative approach which they called "elliptical PCA": it consists in scaling each component of the vector through a robust scale estimate, project the scaled data onto the unit sphere, rescale the projected data and then perform classical PCA.

iv) Visualization of the PCA through a representation of the first principal components directions in the feature space, where intuitive understanding is natural.

As pointed by the authors, the loss due to the compression method is relatively small and not of clinical interest. This first reduction method seems effective and attractive.

With respect to ii) and iii), the authors have to face the extra problem of looking for multivariate robust methods when the number of data n is less than the dimension p of the feature space. When $n > p$, several affine–equivariant robust methods for estimating the location and the covariance matrix are available in practice (see for instance, Tyler (1991) for a review). In addition, the problem of looking for high breakdown point estimates becomes crucial for high-dimensional data. As it is well known, monotone M-estimates have breakdown point less than $1/p$, which makes them resistant only when the number of outliers in the sample is very small (less than n/p) and then inadequate for small data sets in high dimensional spaces. On the other hand, the minimum volume ellipsoid estimators, S-estimators, $\tau-$estimators, CM-estimators and depth-based estimators, such as the Donoho-Stahel estimator, are affine equivariant and have high-breakdown point regardless of the dimension of the data. A shortcoming of these proposals is their computational complexity. Up to our knowledge, the proposed algorithms do not work when the sample size is smaller than the dimension of the space since they are based on resampling methods.

Another drawback of all these estimates of the scatter matrix, except of those based in depth notions, is that they are not well defined when the number of data is less than the dimension of the space.

However, the problems of estimating the location parameter and of finding the main principal components, make sense even in the case where $n < p$.

In fact, a possible approach is that given by the authors. Their proposal is computationally simple, rotationally equivariant but not affine equivariant. Of course that —when possible— consistent, affine equivariant and high breakdown point procedures are desirable, but this is not always possible. An enormous improvement with respect to classical PCA —under the presence of outliers— is obtained with both methods. We think that these proposals point in the right direction: an important improvement over non-robust methods using computationally simple estimates for an "ill-possed" problem.

Going into a more detailed analysis, we found that ellipsoidal principal components have sometimes an asymptotic bias as illustrated in the following example.

We performed 500 replications with samples of size 1000, of a vector $\mathbf{x} \sim N(0, \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}')$, where $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$ with $\boldsymbol{\gamma}_1 = (0.25, \sqrt{15/16})'$, $\boldsymbol{\gamma}_2 = (\sqrt{15/16}, -0.25)'$ and $\mathbf{\Lambda} = \mathrm{diag}(8, 4)$. The mean, median, standard deviation and MAD of the angles (measured in degrees on [-90,90]) between the real and the estimated first principal direction are reported in the following table for the classical and the ellipsoidal principal components.

Principal Components	Mean	Median	SD	MAD
Classical	0.0427	0.0534	1.1473	1.2026
Ellipsoidal	6.4997	6.5188	1.1277	1.1657

Moreover, in 499 of the 500 replications, the angle was greater than 3 degrees while for the classical PCA, 5 of the 500 replications had an angle with absolute value larger than 3 degrees. This is due to the fact that scaling each coordinate is adequate when the principal axes are the canonical basis but not in general.

Indeed, assume that $\mathbf{x}$ has an ellipsoidal distribution, i.e., $\mathbf{x} = \mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z}$ with $\mathbf{z}$ spherically distributed, $\mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{I}$ and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Then, as is well known, if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, the columns $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ of $\mathbf{\Gamma}$ represent the principal components of $\mathbf{x}$ and a q–dimensional reduction is obtained by taking the q eigenvectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ related to the q largest eigenvalues.

In this setting, the proposed ellipsoidal components will be consistent to the eigenvectors of the matrix

$$E\left(\frac{\mathbf{x}\,\mathbf{x}'}{\|\mathbf{D}\mathbf{x}\|^2}\right) = \mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}} E\left(\frac{\mathbf{z}\,\mathbf{z}'}{\|\mathbf{D}\mathbf{\Gamma}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{z}\|^2}\right)\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Gamma}', \tag{1}$$

where $\mathbf{D} = \mathrm{diag}(1/s_{11}, \ldots, 1/s_{pp})$, $s_{ii}^2 = \mathrm{Var}(x_i)$ with x_i the ith component of $\mathbf{x}$ and where we have assumed for simplicity that the vector $\mathbf{x}$ has a finite covariance matrix.

Then if, the principal axes are the canonical basis, i.e., $\mathbf{\Gamma} = \mathbf{I}$, we will have $s_{ii}^2 = \lambda_i$ and so, the matrix on the left of (1) will be proportional to the scatter matrix of $\mathbf{x}$ and thus the elliptical components will be asymptotically unbiased. However, if the principal axis are not the coordinate ones, the matrix

$$E\left(\frac{\mathbf{z}\,\mathbf{z}'}{\|\mathbf{D}\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{z}\|^2}\right)$$

is not necessarily diagonal and then, as in the example, the ellipsoidal components are biased.

On the other hand, the proposed spherical components will always be consistent for any ellipsoidal random vector, since they will be consistent to the eigenvectors of the matrix

$$E\left(\frac{\mathbf{x}\,\mathbf{x}'}{\|\mathbf{x}\|^2}\right) = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}E\left(\frac{\mathbf{z}\,\mathbf{z}'}{\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{z}\|^2}\right)\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Gamma}',$$

which is proportional to the identity in the spherical case, i.e., $\boldsymbol{\Lambda} = \mathbf{I}$.

However, when $\boldsymbol{\Lambda} \neq \mathbf{I}$, $\boldsymbol{\Sigma}$ can be written as

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Gamma},$$

where

$$\boldsymbol{\Psi} = E\left(\frac{\mathbf{z}\,\mathbf{z}'}{\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{z}\|^2}\right).$$

Since,

$$\mathbf{U} = \frac{\mathbf{z}\,\mathbf{z}'}{\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{z}\|^2} = \frac{\mathbf{y}\,\mathbf{y}'}{\|\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{y}\|^2}$$

with $\mathbf{y} = \mathbf{z}/\|\mathbf{z}\|$ being uniform on the sphere, the distribution of $\mathbf{U}$ is the same for any spherically distributed vector $\mathbf{z}$ and so the matrix $\boldsymbol{\Psi}$ can be computed assuming $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$. In this case, it can be shown that Ψ is diagonal and thus, all the basis of eigenvectors is consistently estimated through the spherical principal components.

If $\boldsymbol{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_p) = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Psi}\boldsymbol{\Lambda}^{\frac{1}{2}} = \mathrm{diag}(\lambda_1\psi_1, \ldots, \lambda_p\psi_p)$, it follows easily that $\phi_1 > 1/p$ and $\phi_p < 1/p$, which entails that when $p = 2$, $\phi_1 > \phi_2$ and thus the spherical principal components estimate adequately the principal axis.

Moreover, given $1 \leq j \leq p-1$, since $\lambda_j \geq \lambda_{j+1}$ and $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$, we have that

$$\begin{aligned}
\phi_j &= E\left(\frac{\lambda_j z_j^2}{\lambda_j z_j^2 + \lambda_{j+1} z_{j+1}^2 + \sum_{k \neq j, j+1} \lambda_k z_k^2}\right) \\
&= E\left(\frac{\lambda_j z_{j+1}^2}{\lambda_j z_{j+1}^2 + \lambda_{j+1} z_j^2 + \sum_{k \neq j, j+1} \lambda_k z_k^2}\right) \\
&\geq E\left(\frac{\lambda_{j+1} z_{j+1}^2}{\lambda_{j+1} z_{j+1}^2 + \lambda_j z_j^2 + \sum_{k \neq j, j+1} \lambda_k z_k^2}\right) = \phi_{j+1} \,.
\end{aligned}$$

Also, $\lambda_j > \lambda_{j+1}$ implies $\phi_j > \phi_{j+1}$, which entails that even if the matrix $\mathbf{\Sigma}$ does not estimate consistently the scatter matrix $\mathbf{\Gamma\Lambda\Gamma}'$ of $\mathbf{x}$, it allows to estimate consistently the principal components and the order of the eigenvalues is preserved. The relative importance, i.e., the number of eigenvectors that should be kept in order to obtain a representation which corresponds a high percentage of the trace of the scatter matrix, may be modified, as shown by the following example.

We generated a sample of size 5000 of a vector $\mathbf{x} \sim N(0, \mathbf{\Lambda})$, with $\mathbf{\Lambda}^{\frac{1}{2}} = \mathrm{diag}(9.6, 3, 2, 1.5)$. Thus, the main axis is the first coordinate axis which gives a representation which explains 83% of the total variance. The classical principal components give the following percentages:

$\lambda_1 / \sum_{1<i<4} \lambda_i$	$(\lambda_1 + \lambda_2) / \sum_{1<i<4} \lambda_i$	$\sum_{i \neq 4} \lambda_i / \sum_{1<i<4} \lambda_i$
0.8558304	0.9390217	0.9773178

while the spherical principal components give

$\lambda_1 / \sum_{1<i<4} \lambda_i$	$(\lambda_1 + \lambda_2) / \sum_{1<i<4} \lambda_i$	$\sum_{i \neq 4} \lambda_i / \sum_{1<i<4} \lambda_i$
0.6579615	0.8311586	0.931456

Therefore, the proposed spherical PCA are consistent for any elliptical distribution and thus preferable to ordinary PCA which requires moment conditions.

Obviously, spherical PCA will be resistant for any contamination model which preserves the property of being elliptical. The following small simulations shows, however, that spherical PCA is not resistant with respect to other type of contamination.

As above, we performed 500 replications with samples of size 1000, of a vector $\mathbf{x} \sim N(0, \mathbf{\Gamma\Lambda\Gamma}')$, where $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)'$ with $\boldsymbol{\gamma}_1 = (0.25, \sqrt{15/16})'$, $\boldsymbol{\gamma}_2 = (\sqrt{15/16}, -0.25)'$ and $\mathbf{\Lambda} = \mathrm{diag}(6, 5)$ and we compare the performance of the estimates when we put 10% of contamination at the points $\mathbf{x}_0 = 100\boldsymbol{\gamma}_2 \mathrm{sg}(x_1)$.

In this case, the number of times that the absolute value of the angle between $\boldsymbol{\gamma}_1$ and the estimated first principal direction is greater than 15 (N_{15}), 30 (N_{30}) and 45 (N_{45}) degrees are reported in the following table for the classical and the spherical principal components, for both the normal data sets(C_0) and the contaminated ($C_{0.1}$) ones.

Model	PCA	N_{15}	N_{30}	N_{45}
C_0	Classical	0	0	0
C_0	Spherical	21	1	1
$C_{0.1}$	Classical	500	500	500
$C_{0.1}$	Spherical	463	432	396

In higher dimensions, the behavior is quite similar. We made 100 replications for samples of size 5000 generated as follows. We generated $\mathbf{y} \sim N(\mathbf{0}, \mathbf{\Lambda})$ with $\mathbf{\Lambda} = \mathrm{diag}(6, 5, 4.5, 3)$. With probability 0.9, $\mathbf{x} = \mathbf{y}$ and with probability 0.1, $\mathbf{x} = 100\mathrm{sg}(y_2)\mathbf{e}_2$ where $\mathbf{e}_2 = (0, 1, 0, 0)'$ and $\mathrm{sg}(y_2)$ denotes the sign of the second–coordinate of the vector $\mathbf{y}$. Thus, we have introduced a 10% of contamination at the points $100\mathbf{e}_2$ and $-100\mathbf{e}_2$.

As expected, classical PCA interchanges the first two principal axis but spherical PCA also behaves in the same way. The following table gives the mean of the cosinus of the angles between the theoretical axis and the estimated axis for both the uncontaminated and the contaminated data sets.

Model	First Axis	Second Axis	Third Axis	Fourth Axis
C_0	0.9984	0.9954	0.9961	0.9995
$C_{0.1}$	0.0675	0.0676	0.9992	0.9996

On the contrary, if the contamination is put on the direction of $\mathbf{e}_4$, classical PCA will move all the axis but spherical PCA just moves the third and fourth ones.

The question that naturally arises is if nature is so wild to allow this kind of outliers to appear frequently in practice, in particular, in the problem studied by the authors. As mentioned above, spherical PCA is resistant for any kind of contamination which preserves the property that the underlying distribution is still elliptical. In this sense, spherical PCA are robust, with respect to this kind of neighborhoods, without requiring any moment condition, and thus it represents a great improvement over classical principal components. In particular, it provides a computationally feasible alternative to the classical method for the case where $n < p$.

References

Besse, P. and J.O. Ramsay (1986). Principal component analysis of sampled functions. *Psychometrika*, **51**, 285-311.

Boente, G. and R. Fraiman (1998). Kernel-based functional principal components. Unpublished Manuscript.

Dauxois, J., A. Pousse and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136-154.

Pezzulli, S.D. and B.W. Silverman (1993). Some properties of smoothed principal components analysis for functional data. *Computational Statistics and Data Analysis*, **8**, 1-16.

Ramsay, J.O. and C.J. Dalzell (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistics Society, B*, **53**, 539-572.

Ramsay, J.O. and B.W. Silverman (1997). *Functional Data Analysis.* Springer Series in Statistics, Springer-Verlag, New York.

Rice, J. and B.W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistics Society, B*, **53**, 233-243.

Silverman, B.W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, **24**, 1-24.

Tyler, D. (1991). Some issues in the robust estimation of multivariate location and scatter. In *Directions in Robust Statistics and Diagnostics* (W. Stahel and S. Weisberg eds.) Springer–Verlag, New York, 327-336.

Babette Brumback
Harvard School of Public Health, U.S.A.

Introduction

I would like to thank the authors for this fine piece of work. Their treatment of the corneal image data beautifully illustrates the utility of principal components analysis for characterizing high dimensional data objects. It also firmly motivates the need for special adaptations to standard multivariate techniques in the analysis of functional data. Moreover, it highlights the need to customize these adaptations in practice.

The authors have also contributed a well-organized and insightful presentation of the ideas and methods used in their analysis. In my commentary, I will link their "feature space" analysis to its dual in "data space", raise alternative possibilities for handling the missing data, and relate their methods to others in the literature on functional principal components analysis.

Feature space $\leftrightarrow$ Data space analysis

Let the images be stored sequentially by pixel as J_i-dimensional vectors Y_i, $i = 1, \ldots, n$, where n is the number of images available for analysis (n=43 in the cornea example.) We begin by assuming no data are missing, so that Y_i consists of $J_i = J = 6912$ pixels indexed by polar coordinates (r_{ij}, θ_{ij}). The goal of principal components analysis is to summarize $V = Var(Y_i)$ by its dominant eigenvectors. Modeling the images as i.i.d. realizations from a process such as

$$Y_i = \mu + \delta_i + \epsilon_i, \tag{1}$$

where $\mu + \delta_i$ represents the noiseless image corrupted by additive noise ϵ_i, and $Var(\epsilon_i) = \sigma^2 I$ (with I denoting the identity matrix), the goal is to find the dominant eigenvectors of $Var(Y_i)$, or equivalently, those of $Var(\delta_i)$. For ease of presentation and without much loss of generality, it will be assumed throughout much of the discussion that $\mu = 0$ or, alternatively, that the images Y_i have been pre-centered. The simplest estimate of V is $S_Y = n^{-1} \sum Y_i Y_i^T$, the sample covariance matrix. In the cornea example,

V has dimension 6912×6912, but only 43 realizations of Y_i are observed. Does this necessarily invalidate the naive principal components analysis based upon the dominant eigenvectors of S_Y? If we believe the model in (1), the answer depends upon the relative magnitudes of σ^2 and the eigenvalues of $Var(\delta_i)$. Suppose, for instance, that $\sigma^2 = 0$ and that the largest 5 eigenvalues of $Var(\delta_i)$ dominate the other 6907 by a factor of 100,000. In this case, the first 5 eigenvectors of S should excellently approximate those of V. Unfortunately, however, the sparsity of the cornea dataset renders the relative magnitudes indeterminable; one can only guess whether it is dominant eigenvectors that have been found, or junk.

This curse of dimensionality cannot be overcome by methods for functional data analysis except by making strong untestable assumptions. Traditionally, $\mu + \delta_i$ is assumed smooth. This assumption is a reasonable one in numerous applications, and it certainly appears sensible in Locantore et al., who invoke it by modeling $\mu + \delta_i$ as a Zernike polynomial. That is, they assume $\mu = U\beta_0$ and $\delta_i = U\beta_i$, where the columns of U are the p ($p = 66$ in Locantore et al.) orthonormal Zernike basis functions sampled at locations (r_{ij}, θ_{ij}), $j = 1, \ldots, J$, which index the rows. Again assuming that $\mu = 0$ (or that the mean has already been subtracted), β_i can be estimated via ordinary least squares regression,

$$\hat{\beta}_i = U^T Y_i,$$

where recall that orthonormality of U implies $U^T U = I$. The information stored in Y_i is thereby compressed into the feature space representations $X_i = \hat{\beta}_i = U^T Y_i$. The authors compute the eigenvectors v_{xk}, $k = 1, \ldots, 3$ of $\hat{Var}(X_i) = S_X = n^{-1} \sum X_i X_i^T$ in the feature space; these are then reexpressed in the data space as $v_{yk} = U v_{xk}$. This produces smooth estimated eigenvectors of V.

It can easily be shown that the v_{yk} are, in turn, the eigenvectors of

$$U S_X U^T = U U^T S_Y U U^T = P S_Y P,$$

where $P = UU^T$ is the projection in the data space onto the span of the Zernike basis. Thus, the feature space analysis has a data space dual, in which the eigenvectors of the smoothed sample covariance matrix

$$S_s = P S_Y P \tag{2}$$

are computed directly. The smoothed covariance matrix will have reduced rank equal to p, the dimension of the Zernike basis.

Missing data

When there is missing data, Locantore et al. modify the estimates of v_{yk} as follows. Let U_i be analogous to U but with rows corresponding to missing pixels deleted. The feature space representations are modified to $X_i = \hat{\beta}_i = U_i^T Y_i$; the v_{xk} are computed as before but using the modified version of X_i; the v_{yk} are then obtained as Uv_{xk}. The analogue in data space is to compute the v_{yk} directly as the eigenvectors of $n^{-1} \sum P_i Y_i Y_i^T P_i$, with $P_i = UU_i^T$. As the authors demonstrate, this method can lead to undesirable consequences. Specifically, certain of the projections $X_i = P_i Y_i$ of images with missing data turn up as outliers with excessive influence upon the estimated principal components. I offer a possible explanation by way of a simple example. Consider the toy dataset pictured below, with pixels corresponding to a common angle θ but distinct values of r (i.e. the toy "images" are 1-dimensional):

	subject			
	1	1	1	1
	2	2	2	2
	3	3	3	3
pixel	?	3	3	3
	?	3	3	3
	?	3	3	3

Because the Zernike basis includes the linear functions, the missing pixels from subject 1 will be substituted by 4, 5, and 6, leading to an outlier. Although the data strongly suggest an overall mean of $(1, 2, 3, 3, 3, 3)$ with zero variation thereabout, the first analysis of Locantore et al. would lead to a contaminated estimate of the mean and a nontrivial estimated first principal component. The robust approach adopted by the authors would diminish the problem by downweighting the observation from subject 1 in the analysis.

I would like to propose an alternative method for handling the missing data. The idea is to first estimate the elements of the complete data covariance matrix V with

$$\hat{V}(s,t) = n_{st}^{-1} \sum \Delta_i(s) Y_i(s) \Delta_i(t) Y_i(t),$$

where $V(s,t)$ represents the covariance between pixels s and t, $\Delta_i(s)$ is the missing data indicator ($\Delta_i(s) = 1$ if pixel s is observed for subject i), and

$n_{st} = \sum_i \Delta_i(s)\Delta_i(t)$. The smooth principal components are then computed as the eigenvectors of $P\hat{V}P$. This method presumes that the Y_i have been centered; a smooth group mean could be computed beforehand in an analogous manner by smoothing a weighted mean of the available data. The effect of this approach will be to use other subjects rather than neighboring pixels to fill in the missing data. Alternatively, one could borrow strength simultaneously from other pixels and subjects via an enhanced model for Y_i; i.e., one could let $\delta_i = \gamma + \gamma_i$, where γ is introduced as a shared random effect to induce correlation between observations from different subjects. See Brumback and Rice (1998) for a related discussion in the context of mean estimation.

Related approaches

Still other approaches might be adopted. First, one might choose an alternative basis. The possible effects are illustrated in Figure 1. The top left panel presents the true first three principal components for 20 1-dimensional "images" simulated using `Splus`. The images were each sampled with additive noise at 100 pixels. In all panels, the solid line represents the 1st principal component, the dotted line the second, and the dashed line the third. The true principal components correspond to the 5th, 6th, and 1st functions of the Zernike basis generated for 1-dimensional quintic polynomial images by setting $\theta = 0$ and $(m, n) = (0, 0), (1, 1), (0, 2), (1, 3), (0, 4)$ and $(1, 5)$. The top right panel represents the first three estimated principal components using the unsmoothed covariance matrix. The middle left panel shows the estimates computed as in Locantore et al. by projecting onto the Zernike basis. The middle right panel uses an asymmetric Bspline basis with knots at 60 and 80. The Bspline basis is smoother for the first 50 pixels than for the remainder. Notice the dramatic effect on the estimated principal components, which demonstrates that the wrong choice of basis can lead to serious consequences. This emphasizes the importance of subject matter knowledge in performing functional PCA; when the curse of dimensionality rears its head, the results begin to depend heavily upon untestable assumptions.

Another option is to use a roughness penalty approach as in Rice and Silverman (1991) or Silverman (1996). The first approach estimates the principal components as the eigenvectors of $\hat{V} + \lambda_1\Omega,\ \hat{V} + \lambda_2\Omega, \ldots$, where

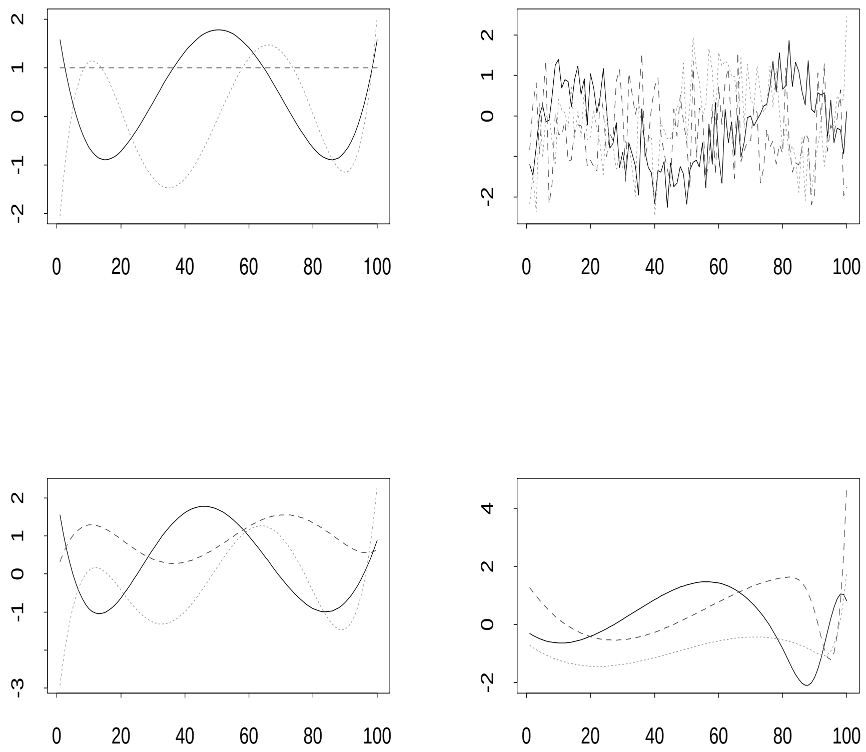

Figure 1: *True components shown at top left. In all panels, the solid line represents the 1st principal component, the dotted line the second, and the dashed line the third. Top right presents the first three estimated principal components using the unsmoothed covariance matrix. Middle left computes the estimates using the method of Locantore et al. to smooth the covariance matrix by first projecting the data onto the Zernike basis. Middle right uses an asymmetric Bspline basis with knots at 60 and 80.*

Ω represents is a covariance matrix with smooth dominant eigenvectors, and λ_k, $k = 1, \ldots$ is a non increasing sequence of tuning parameters mediating between eigenvectors of $\hat{V}$ and Ω; eigenvectors of greater importance bear more resemblance to those of Ω. Silverman (1996) modifies the procedure for efficient computation by constraining the sequence of λ_k. With an empirical choice of Ω, the procedure of Locantore et al. can be embedded within the roughness penalty framework; i.e. letting $\Omega = P\hat{V}P$ leads to the procedure of Locantore et al. for large λ_k. However, the computational cost of the roughness penalty approach may not be within reach. A compromise would be to partially reduce the data using the method of Locantore et al. with a midrange-dimensional basis and then to finish the reduction using roughness penalty methods. But the approach of Locantore et al. has been shown to yield insightful conclusions, and I do not recommend any alterations. I would like to conclude by remarking upon the weighty computational challenges overcome by the authors; data storage and manipulation of high dimensional data objects are often quite difficult, necessitating sophisticated software and clever programming strategies. The video images of the cornea data are particularly impressive.

References

Brumback, B.A. and J.A. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961-994.

Rice, J.A. and B.W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, B*, **53**, 233-243.

Silverman, B.W. (1996). Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, **24**, 1-24.

Christophe Croux
Université Libre de Bruxelles

Introduction

First of all I would like to congratulate the authors with their paper, which presents a very nice analysis of a functional data set from ophthalmology using robust principal components. In this comment, I will focus on the newly proposed method for robust PCA.

Several robustifications for PCA have been proposed in the past. The most simple idea is to compute eigenvalues and eigenvectors of a robust estimator of the covariance or correlation matrix of the data. Many simulation studies, starting with (Devlin et al. 1981), have been carried out to find out which robust estimator should be used, and recently some more theoretical results were obtained by Croux and Haesbroeck (1999). As was pointed out by the authors, these methods require that the number of variables d is higher than the number of observations n, making them less useful for functional data analysis.

Another approach to robustify PCA, based on projection pursuit (PP), has been considered by Li and Chen (1985). As one knows, a classical principal component is determined by the direction for which the projections of the data onto that direction have maximal standard deviation, under the constraint of orthogonality with all previously determined components. Instead of maximizing the standard deviation, one uses now a robust dispersion measure as "PP-index", resulting in a robust PCA. Since the principal components are computed sequentially, this approach can be used even in the high dimensional case $n < d$.

The method proposed in this paper has both a projection aspect and an eigenanalysis aspect. A important virtue of this method is it simplicity and ease of implementation. In contrast with many other highly robust multivariate statistical procedures, the required computation time is extremely limited.

Some Statistical Properties

For a sample $X = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$, the proposed robust PCA is carried out by computing the eigenvectors $v_1(X), \dots, v_k(X)$ of the matrix

$$\tilde{\Sigma}_n(X) = \sum_{i=1}^{n} \frac{(X_i - \hat{\theta}_n)(X_i - \hat{\theta}_n)^t}{\|X_i - \hat{\theta}_n\| \, \|X_i - \hat{\theta}_n\|}, \tag{1}$$

with $k = \text{rank}(\tilde{\Sigma}_n(X))$ and $\hat{\theta}_n$ the L_1 location estimator. The "robust eigenvectors" $v_1(X), \dots, v_k(X)$ are the vectors of interest since the data will be projected on them.

Equivariance Properties: Although $\tilde{\Sigma}_n$ is no *affine equivariant* covariance matrix estimator, it is orthogonal equivariant which suffices in the context of PCA. Indeed, denote $\alpha\Gamma X + b = \{\alpha\Gamma x_1 + b, \dots, \alpha\Gamma x_n + b\}$ where Γ is an orthogonal matrix, b a vector in $\mathbb{R}^d$ and α a scalar, then the usual equivariance property holds

$$v_j(\alpha\Gamma X + b) = \Gamma v_j(X), \tag{2}$$

for $j = 1, \dots, k$.

By first prescaling the data, for example by dividing them by the coordinatewise MAD, an equivalent of a correlation based PCA is obtained. This procedure is called elliptical PCA by the authors, and one has the additional equivariance property $v_j(DX) = v_j(X)$ for any diagonal matrix D.

Influence Function: The authors claim that outliers have bounded influence on their procedure. This can be made formal. To keep things simple, suppose that we are in the bivariate normal case, and due to (2) suppose w.l.o.g.

$$X_1, \dots, X_n \overset{iid}{\sim} F = N\left(0, \begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}\right), \qquad 0 < \gamma < 1.$$

The functional corresponding to the $\tilde{\Sigma}_n(X)$ is given by

$$\tilde{\Sigma}(G) = \int \frac{(y - T(G))(y - T(G))^t}{\|y - T(G)\|^2} dG(y)$$

for an arbitrary distribution G. Denote then $v_1(G), \ldots, v_d(G)$ the eigenvectors of $\tilde{\Sigma}(G)$. It is not difficult to show that

$$\tilde{\Sigma}(F) = N\left(0, \operatorname{diag}(1/(\sqrt{\gamma}+1), 1/(\sqrt{\gamma^{-1}}+1)\right),$$

implying Fisher consistency for the eigenvectors at bivariate normal distributions. Like in Critchley (1985), one can prove quite easily that the influence function for v_1 is given by

$$IF((x_1, x_2), v_1, F) = \frac{1-\gamma}{(1-\sqrt{\gamma})^2} \frac{x_1 x_2}{\|x\|^2} v_2(F) \tag{3}$$

and analogously for the second eigenvector. From (3) boundedness of the influence function follows immediately.

Efficiency: Since emphasis in the paper was on the use of the proposed method as a tool for exploratory data analysis, efficiency considerations are less important but nevertheless interesting. Take once again the simple case where the data come from the bivariate normal distribution F. Assuming that the functional $\tilde{\Sigma}$ is sufficiently regular, the asymptotic variance of v_1 equals

$$\mathrm{ASV}(v_1, F) = \frac{\sqrt{\gamma}}{2(1-\sqrt{\gamma})^2} v_2(F) v_2(F)^t,$$

which needs to be compared with the asymptotic variance of the classical estimator of the first eigenvector $(\gamma/(\gamma-1)^2) v_2(F) v_2(F)^t$. In Figure 1 the associated efficiency (defined as the ratio of the traces of the asymptotic variance matrices) is pictured as a function of γ.

The efficiency of the method depends thus on γ and never exceeds 50%: the more spherical the distribution, the higher the efficiency of the method. This is in contracts with most other methods for robust PCA, where the efficiencies are independent of γ. The same problem will arise for the elliptical version of the method.

Some suggestions

(i) It is not so obvious to interpret the eigenvalues of $\tilde{\Sigma}_n$. As a measure of dispersion of the data in the direction of $v_j(X)$, one could compute

$$\hat{\lambda}_j = S_n(v_j(X)^t X_1, \ldots, v_j(X)^t X_n),$$

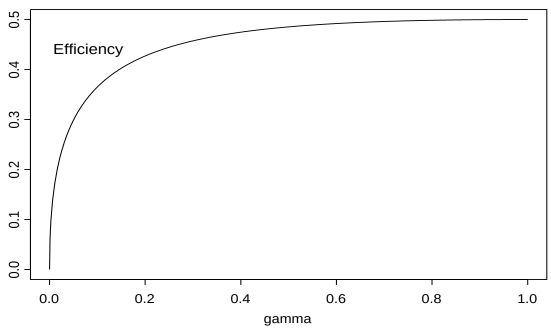

Figure 1: *Efficiency of the proposed estimator for the first eigenvector of a bivariate normal distribution as a function of γ, where γ equals the second divided by the first population eigenvalue.*

for $j = 1, \ldots, k$, with S_n a robust univariate scale estimator like the MAD. Moreover, unlike the eigenvalues of $\tilde{\Sigma}_n$, the $\hat{\lambda}_j$ will be consistent estimators for the eigenvalues of covariance matrices of normal distributions.

(ii) A generalization of (1) is given by

$$\tilde{\Sigma}_n(X) = \sum_{i=1}^{n} w_i \frac{(X_i - \hat{\theta}_n)(X_i - \hat{\theta}_n)^t}{\|X_i - \hat{\theta}_n\| \; \|X_i - \hat{\theta}_n\|},$$

where the assigned weights w_i depends only on the rank of $\|X_i - \hat{\theta}_n\|$ and $\sum_{i=1}^{n} w_i = 1$. The location counterpart of the above estimator has been studied by Hössjer and Croux (1995). By choosing the weights properly, higher efficiencies can be obtained while not loosing too much robustness.

(iii) The choice of the starting value for the algorithm computing the L_1 estimator is not crucial, but the coordinatewise median might yield faster convergence in noisy data sets than the sample mean. Using the Newton steps of Bedall and Zimmerman (1979), the computation time of the L_1 estimator could be even further reduced.

References

Bedall, F.K. and H. Zimmermann (1979). Algorithm AS 143. The mediancentre. *Applied Statistics*, **28**, 325-328.

Critchley, F. (1985). Influence in principal component analysis. *Biometrika*, **72**, 627-636.

Croux, C. and G. Haesbroeck (1999). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Preprint, University of Brussels (ULB), `http://www.sig.egss.ulg.ac.be/Haesbroeck/`.

Devlin, S.J., R. Gnanadesikan and J.R. Kettenring (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, **76**, 354-362.

Hössjer, O. and C. Croux (1995). Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *Nonparametric Statistics*, **4**, 293-308.

Li, G. and Z. Chen (1985). Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. *Journal of the American Statistical Association*, **80**, 759-766.

Jianqing Fan

University of California Los Angeles, U.S.A.

When analyzing complex objects such as images, it is vital to have a simple, effective, robust and computationally feasible approach that can explore salient population features. The authors are to be congratulated for successfully outlining such an elegant method, which extracts and summarizes important clinical features. Deep insights are obtained via graphical presentation and elegant exposition. I am very fortunate to associate with this group for a long time and to wit how this interdisciplinary collaboration yields fruitful statistical methodology innovation and useful clinical results. Such a kind of joint efforts should be strongly encouraged and greatly expanded.

The new method in this paper consists basically of the following steps:

a) extract important features;

b) find the center of the data in the feature space;

c) rescale the centered data in the feature space;

d) carry out the principal component analysis in the feature space by normalizing each data point to have unit length;

e) present results in the original domain of data.

Such a kind of proposal appears ad hoc but effective. I welcome the opportunity to make a few comments on some of these critical steps.

Feature extraction

Feature extraction is extremely vital for analyzing high dimensional data such as noise signals and images. The aim is to significantly reduce dimensionality without losing important information in the original data. This is usually achieved by manually selecting a number of important characteristics that are directly related to the objective of a study or via an orthogonal transform (or local orthogonal transform such as spectrograms in speech recognition). In the current context, the authors reduce the dimensionality from 6912 to 64 via a two-dimensional orthogonal system called the Zernike transform. Other orthogonal bases can also be used, but this system has better known optical properties.

Different orthogonal systems have different ability of information compression, depending on the classes of signals. For example, the Fourier transform is not effective to represent local features such as bumps or short aberrations while wavelet bases are not very efficient in representing sinusoid signals. When choosing an orthogonal basis, the efficiency and the interpretation of the orthogonal basis should be taken into serious considerations. The role of the orthogonal transform here can be intuitively understood as compressing original 6912 highly correlated dimensions (because intensity of neighboring pixels is nearly the same) to 64 nearly independent components.

Statistically, keeping a few coefficients in feature spaces is equivalent to conducting heavy amount of smoothing for the original data. The parameter 64 can be regarded as a smoothing parameter. This results in keeping prominent and stabilized features in the data, since disturbing noises have been reduced. This technique is also very useful for hypothesis testing such as comparing differences of cornea maps between two clinical groups. Due to dimensionality reduction, only prominent features of images are now tested and the power of resulting testing procedures are ameliorated. See for example Fan and Lin (1998).

Orthogonal transforms are linear. They depend sensitively on outliers in the original data. For the current application, outliers are mainly caused by missing data at boundaries of images. A natural question is then if there are some robustness imputation methods to implement or some robust nearly orthogonal transforms to apply at this stage, rather than at a later stage.

Centers in high-dimensional space

To find robust principal directions, the first step is to find robust estimation for the center of high-dimensional data. The authors propose to use L^1 M-estimate of location via minimizing

$$\sum_{i=1}^{n} \|X_i - \theta\|_2.$$

While this method is far more robust than the sample mean, one may naturally ask how robust it is when compared with the componentwise median, which minimizes

$$\sum_{i=1}^{n} \|X_i - \theta\|_1,$$

where $\|X_i - \theta\|_1$ is the sum of componentwise distances. It appears clear that the former is more robust than the componentwise median, but the latter has computational advantages. If outliers do not occur as fearfully as we think, the componentwise median can be an attractive alternative. It is not clear to me why the componentwise median was not chosen as the center of the images in the current application. It is also interesting to relate these methods with the concept of data depths of Liu and Singh (1992).

Functional Principal Component Analysis

After locating the center of the data, the authors propose to carry out the principal component analysis based on the projected data $\{(X - \hat{\theta})/\|X -$

$\hat{\theta}\|_2\}$ (an weighted L_2-norm is used when the data are projected on an ellipse). This is a powerful idea and a useful technique, but there is also some hidden cost. To achieve robustness, we completely change the parameters under estimation. Unless the distribution of $\|X - \theta\|_2$ is nearly constant, the covariance or correlation matrix of the projected data can be quite different from that of the original data. Hence, the population parameters (the principal directions) for the two problems are completely different. Strictly speaking, the principal component analysis based on the projected data is not a robustified version of that based on the original data. Which principal component analysis is more relevant depends critically on the scope of applications.

In a seminal paper, Rice and Silverman (1991) proposed to use smoothed principal component analysis for the original data. An important feature of that method is that the resulting principal component directions are those that capture the greatest variabilities in the original data subject to smoothness constraints. It poses, however, challenges to even modern computers to carry out a 7,000-dimensional principal component analysis in the original data domain, even without imposing smoothness constraints on the principal directions. An important distinction of the current method is that it is carried out in the feature space. This reduces computational cost dramatically.

An intuitive alternative approach to that of Rice and Silverman (1991) is to smooth curves or images first and then apply the ordinary principal component analysis to the smoothed data. This also results in smooth directions in the original data domain that capture the greatest variabilities of smoothed curves or images (instead of original data). This approach is indeed equivalent to that proposed by the authors, when the principal components are obtained based on covariance matrices. To see this, suppose that after an orthonormal transform we decide to keep the first p transformed coefficients. Denote such coefficients by ξ and the corresponding first p orthogonal bases by X, an $n \times p$ orthonormal matrix. Then the resulting smoothed images or curves are just truncated orthonormal reconstruction: $Y = X\xi$. The covariance matrix of the smoothed images is

$$\operatorname{cov}(Y) = X \operatorname{cov}(\xi) X^T = X \Gamma \Lambda_p \Gamma^T X^T,$$

where $\Gamma \Lambda_p \Gamma^T$ is the principal component analysis (eigenvalue decomposition) for the covariance matrix $\operatorname{cov}(\xi)$ in the feature space. It is now very easy to verify that the p directions in the matrix $X\Gamma$ are orthonormal.

Hence, the principal directions for the smoothed images Y are the same as those generated from the feature space. The same conclusion holds for "robustified" principal component analysis with the projection on the sphere since $\|Y\| = \|\xi\|$, but does not hold when the smoothed data are projected on an ellipse.

When the correlation matrix based principal component analysis is used in the feature space, it does not appear to have an intuitive equivalent method in the images domain. Indeed, it is not clear what the resulting orthogonal directions represent in the image domain. Hence, the interpretation and usefulness of such a kind of analysis remain questionable.

Acknowledgements

Fan's research was partially supported by NSF grant DMS-9803200.

References

Liu, R.Y. and K. Singh (1992). Ordering directional data: concepts of data depth on circles and spheres. *Annals of Statistics*, **20**, 1468-1484.

Rice, J.A. and B.W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, B*, **53**, 233-243.

Alois Kneip

Université Catholique de Louvain, Belgium.

The authors have written an interesting article. In the first part of the paper they provide a well-written and very convincing introduction to the use and the benefits of functional principal component analysis. A particularly interesting point is the extension of these concepts developed for the analysis of one-dimensional curves to two-dimensional images.

I only have a minor technical remark in this context. The authors do not directly perform a PCA of the original data, their analysis is based on the coefficients of a Zernike decomposition. This use of "feature vectors" which is common in their field of application seems to be a good idea. However, they write that data are in the form of 6912 measurements at a polar grid of locations, and that classical multivariate analysis of these vectors is numerically intractable. This is not true, and though I think that their alternative approach is very reasonable in their context, I want to show how to deal with such numerical problems in functional PCA, since this might be useful in other applications.

In functional principal component analysis we usually have to treat n different functions f_i (one-dimensional or higher dimensional) which are given at m discretized grid points. In this context we generally have $m \gg n$ (note that m=6912 while n=43 for the application discussed in the paper). In fact, if the number m of data points is very large, then evaluation of the $m \times m$ covariance matrix V of these vectors will indeed be numerically unstable. However, for any grid point x consider the n-vectors $\mathbf{f}(x) = (f_1(x), \ldots, f_n(x))'$, and determine their $n \times n$ covariance matrix M. Calculation of the eigenvalues $\lambda_1 > \lambda_2 > \ldots$ of V as well as of the corresponding eigenvectors $\gamma_1, \gamma_2, \ldots$, defining the principal components, can then be done on the basis of the much smaller and numerically more tractable matrix M. Some simple linear algebra shows that the eigenvalues of M coincide with the eigenvalues $\lambda_1, \lambda_2, \ldots$ of V. Moreover, the principal components $g_1, g_2, \ldots$ are given by

$$g_r(x) = \lambda_r^{-1/2} \sum_i p_{ri} f_i(x),$$

where $p_1 = (p_{11}, \ldots, p_{1n})', p_2 = (p_{21}, \ldots, p_{2n})', \ldots$ are the corresponding eigenvectors of M. If $m \gg n$, this way of calculating principal components in the context of FPCA is numerically much more stable.

In the paper, as in most other work, functional principal component analysis is seen as a tool for analyzing i.i.d. samples of random functions. This certainly is the standard application, but there is another way of representing principal components which is more generally interpretable. Let $f_\mu = \frac{1}{n} \sum_i f_i$ denote the average function. It is well known, the mathematical basis being the famous Kahounen-Loève decomposition, that for any L the first L principal components define the best possible projection of the functions $f_1, \ldots, f_n$ into an L dimensional linear subspace. In other words,

$g_1, \ldots, g_L$ provide a best possible representation

$$f_i(x) - f_\mu(x) \approx \sum_{r=1}^{L} \theta_{ri} g_r(x), \quad i = 1, \ldots, n \tag{1}$$

in the sense that they satisfy

$$\sum_i \|f_i - f_\mu - \sum_{r=1}^{L} \theta_{ri} g_r\|_2^2 = \min_{v_1,\ldots,v_2} \sum_i \min_{\vartheta_{1i},\ldots,\vartheta_{Li}} \|f_i - f_\mu - \sum_{r=1}^{L} \vartheta_{ri} v_r\|_2^2 \tag{2}$$

with respect to all possible $v_1, \ldots, v_n$. An interpretation of principal components in terms of (1) and (2) makes sense for many different families of curves and is quite independent of the underlying mechanism generating the data. For example, it might be used in the context of time series of curves as considered by Bosq (1991). A possible application is analysis of electricity consumption curves for a number of consecutive days.

In the second part of their paper the authors present a robust version of PCA in the context of functional data analysis. To my knowledge this is the first work which explicitly deals with this question. The authors' convincingly argue that the importance of robust methods is even more pronounced if the data are functions than in the multivariate context. The reason is simple: when going over from vectors to functions or surfaces, data structure become more complicated, and there are more and more different ways an observation can be an outlier. Spherical or elliptical PCA as introduced in the paper provides a simple way to robustify functional principal component analysis, and it might thus prove to be an important idea.

In view of (1) and (2), it seems to me that in principle there might exist still more robust versions of PCA. After having replaced f_μ by the spatial median as proposed by the authors, one might try to define more robust principal components by determining the best linear approximation (1) in terms of the L^1 norm, instead of minimizing with respect to the L^2 norm as in (2). Of course, there is no straightforward solution to the resulting complex optimisation problem.

An interesting and important general aspect of the paper is that it demonstrates the possible complexity of functional data. As a final point of my discussion I would like to add some remarks which illustrate this complexity even further. In fact, there are important problems of functional data analysis which do not even possess an analogue in usual multivariate analysis.

I want to consider two of these problems. First of all, in many cases the "true" curves or surfaces are unknown and have to be estimated from discrete data. This is the case, for example, if we want to analyse families of regression curves, or families of noisy images, where the "true" objects of interest are not directly given, but only represented by noisy data. Sometimes data are fairly accurate and noise does not play an important role, but in other applications the estimation error cannot be neglected. In this case we run into two different data analytic and inferential problems: We have to define procedures to analyse similarities and differences between the "true" functions, for example one might rely on PCA, and one has to consider how to estimate principal components or other properties of interest from the given data. In a certain sense one thus has to deal with an unusual type of errors-in-variables problem. There is not much literature on this subject, very often this point is simply ignored (see, however, Kneip, 1994, or Kneip and Utikal, 1999).

When following a term coined by Ramsay and Silverman (1997), a second problem without multivariate analogue is the "registration problem". Many samples of curves like growth curves, brain potentials, etc., do not only differ in amplitudes but also in dynamics. For example assume that there is a collection of one-dimensional functions $f_1, \ldots, f_n$ which only vary in amplitude, location and scale according to the simple model

$$f_i(x) = \theta_i g\left(\frac{x - \alpha_i}{\beta_i}\right) \tag{3}$$

for some basic underlying function g. If the parameters α_i, β_i, quantifying individual dynamics, are very different, then functional principal component analysis is of no use for analyzing the structure of this curve family. Too many principal components will be necessary to explain a large proportion of variability, some of these components being derivatives of others. The point is that, as can be seen from (1), PCA attempts to explain variation between curves by amplitude differences only, and it is not able to incorporate varying dynamics. A possible remedy is first to "register" curves in order to eliminate such differences in dynamics, and then to perform a PCA in a second step. Some registration procedures have been proposed by Kneip and Gasser (1992), Wang and Gasser (1995), Silverman (1995), Ramsay and Li (1996), or Kneip, Li, MacGibbon and Ramsay (1998).

References

Bosq, D. (1991). Modelization, non-parametric estimation and prediction for continuous time processes. *NATO, Asi Series*, Springer Verlag, New York.

Kneip, A. (1994). Nonparametric estimation of common regressors for similar curve data. *Annals of Statistics*, **22**, 1386-1428.

Kneip, A. and T. Gasser (1992). Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, **20**, 1266-1305.

Kneip, A., X. Li, B. MacGibbon and J.O. Ramsay (1998). Curve registration by local regression. *Canadian Journal of Statistics*, to appear.

Kneip, A. and K. Utikal (1999). Inference for density families using functional principal component analysis. Manuscript.

Ramsay, J.O. and X. Li (1996). Curve registration. *Journal of the Royal Statistical Society*, to appear.

Ramsay, J.O. and B.W. Silverman (1997). *Functional Data Analysis*, Springer Verlag, New York.

Silverman, B.W. (1995). Incorporating parametric effects into functional principal component analysis. *Journal of the Royal Statistical Society, B*, **57**, 673-689.

Wang, K. and T. Gasser (1995). Alignment of curves by dynamic time warping. *Biometrics*, **14**, 1-17.

John I. Marden

University of Illinois at Urbana-Champaign, U.S.A.

The movies for visualizing a principal component in the space of images are revealing and delightful. They make a compelling argument for Web-based journals.

Clearly, the authors gave careful consideration to the reduction of the pixel information to the feature vectors. The Zernike basis appears to be an excellent choice given the natural sphericity of eyes. I wonder how poorly the principal components would work if a less appropriate basis was used, such as coordinatewise orthogonal polynomials. Do the authors have any cautionary tales to tell?

The paper demonstrates the utility of the spherically based robust procedures without needing to make restrictive distributional assumptions. If one does make some assumptions, then these procedures are also theoretically reasonable:

1. The spatial median is robust and efficient. In the spherical normal case, Brown (1983) (see also Chaudhuri, 1996) has shown that for estimating the mean, the multivariate median has efficiency that approaches 1 as the dimension approaches infinity. For any dimension, the multivariate median has breakdown of 50%. Thus for high dimensional data as in this paper, the multivariate median (in this special case) is much more robust breakdown-wise, yet practically as efficient as the sample mean.

2. The population spherical principal components are the same as the usual population principal components.

Suppose the m-dimensional random vector X can be written as

$$X = \Omega Z + b, \tag{1}$$

where Ω is a fixed orthogonal matrix, b is a fixed vector, and Z is a random vector that is coordinatewise symmetric about 0, that is, Z and

$$(\pm Z_1, \ldots, \pm Z_m)',$$

have the same distribution. If $\Lambda_Z \equiv Cov(Z)$ exists, then it is diagonal, and

$$Cov(X) = \Omega \Lambda_Z \Omega'.$$

Thus the columns of Ω are the eigenvectors of $Cov(X)$, hence the usual population principal components (in some order).

Marden (1999) shows that

$$Cov\left(\frac{X - b}{\|X - b\|}\right) = \Omega \Lambda \Omega'$$

for some diagonal matrix Λ. Thus the columns of Ω are also the eigenvectors of the covariance of the spherical variables, i.e., they are the population spherical principal components.

There is no guarantee that magnitudes of the diagonal elements in Λ_Z and Λ are in the same order, so it may be, e.g., that the first usual principal component is the second spherical principal component.

Visuri, Koivunen and Oja (1999) note that if one is willing to assume further that X has an elliptically symmetric distribution (a special case of (1)), then the diagonal elements λ_i of Λ are known functions of the diagonal elements λ_{Zi} of Λ_Z:

$$\lambda_i = E\left[\frac{\lambda_{Zi}U_i^2}{\lambda_{Z1}U_1^2 + \cdots + \lambda_{Zm}U_m^2}\right],$$

where the U_i's are independent standard normals. In particular, this relationship shows that the two sets of diagonal elements are in the same order, so that the i^{th} usual principal component is indeed the i^{th} spherical principal component.

The implication of these results is that the sample spherical principal components are estimating the usual population principal components. These papers give evidence that the robust estimates are robust and efficient.

References

Brown, B.M. (1983). Statistical use of the spatial median. *Journal of the Royal Statistical Society, B*, **45**, 25-30.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, **91**, 862-872.

Marden, J.I. (1999). Some robust estimates of principal components. *Statistics and Probability Letters*, **43** (to appear).

Visuri, S., V. Koivunen, H. and Oja (1999). Sign and rank covariance matrices.

Daniel Peña and Javier Prieto
Universidad Carlos III de Madrid, Spain.

This is an interesting paper and the authors should be congratulated for presenting a thought-provoking analysis of a challenging problem. The increasing availability of large data sets in high dimensions has led to a growing need for exploratory tools that can reveal the hidden structure in these data sets. The methods presented in this paper can be very useful in this regard.

The standard multivariate statistical analysis assumes that we have measured a vector variable on each sample point and, therefore, the data is represented by a matrix $\boldsymbol{X}_{n\times p}$, in which usually the columns are the variables, the rows the elements in the sample and $p < n$. A natural generalization of multivariate data is the class of functional data presented in this paper. If instead of assuming a finite vector of p variables we measure a function $x(t_k)$, $k = 1, ..., T$ at a finite set of points t_k and we take the measurements for each realization of $x(t)$ as one element of the sample, the data will be still represented by a matrix $\boldsymbol{X}_{n\times T}$ but now the variables will be highly correlated, typically $T > n$, the covariance matrix will be close to singular and the analysis of the data should take into account the correlations induced by the smoothness of $x(t)$.

The data set used in the paper can be considered as a scalar stochastic process $x(s_l, t_k)$ along two directions, where $l = 1, ..., S$, $k = 1, ..., T$, and the sample data is of the form $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ where the $\boldsymbol{X}_i$ are $S \times T$ matrices that represent the measurements made on the ith sample element of this process. More specifically, the data for each corneal image is a matrix with 6912 elements ($2^7 3^2$) corresponding to values $x(s_l, t_k)$ in a grid of angles and distances. A standard technique to summarize this information is to fit a basis to this bidimensional stochastic process and reduce the analysis of the data to the study of the vector of fitted coefficients. That is, for a specific choice of a finite subset of basis functions $\{b_j(s,t)\}_{j=1}^{f}$ on $[0, 2\pi] \times \Re$ we associate to each matrix $\boldsymbol{X}_i$ a function $g(\boldsymbol{X}_i)$ that returns the f-dimensional vector of coefficients $\boldsymbol{\theta}_i$ providing the best fit of $\sum_i \theta_{ij} b_j(s,t)$ to $\boldsymbol{X}_i$. In this way we can reduced the data space $\boldsymbol{X}_i$ in dimension $S \times T$ to the feature space of dimension f corresponding to the vectors $\boldsymbol{\theta}_i$.

The problem considered in this paper is how to analyze the original data by looking at the structure of these feature vectors in a robust way not affected by outliers. It is clear that any analysis carried out in this manner may depend greatly on the choice of basis functions, and the number of coefficients f to use. Note that the procedure described in the paper is motivated by the need to avoid the high correlation that will appear between the observations, due to the continuous nature of the process generating the data. But the fitting process, and the subsequent analysis in the feature space, will only be of help for this purpose if the number of elements f in the basis is very small. On the other hand, a reduced number of elements may provide a poor fit, implying that the feature data may be an inadequate representation of the original data. As the techniques used for the analysis on the feature space are standard ones, this choice of a representation (a basis and a number of elements) providing a balance between compactness and precision becomes a key issue to justify the advantages of the proposed procedure.

Unfortunately, the paper provides very little information about the advantages of the Zernike basis functions and how well they fit the data. For instance, Fan and Lin (1998) fit Legendre polynomials of order 7 followed by a Fourier transform to a similar set of cornea measures. It would be useful if the authors comment on the pros and cons of different representations for this type of data, and their impact on the subsequent statistical analysis. Also, the crucial issue of the choice of dimension for the feature space is not addressed in the paper. In figure 3 the plot of Zernike coefficients seems to have 66 components but no indication is given on the reasons for this choice, and the fit that can be achieved.

If we understand the paper correctly, the authors first subtract the mean and then compute eigenvectors and eigenvalues of the covariance matrix between the feature vectors $\boldsymbol{M} = \boldsymbol{T}'\boldsymbol{T}/n$, where $\boldsymbol{T}$ has rows corresponding to $\boldsymbol{\theta}_i'$. Let $\boldsymbol{v}_1$ be the $f \times 1$ eigenvector associated to the largest eigenvalue of $\boldsymbol{M}$. The representation of the ith sample point in terms of this first principal component is the function $(\boldsymbol{\theta}_i^T \boldsymbol{v}_1) \sum_i v_{1j} b_j(s,t)$ and this collection of functions represents the best approximation to the data. In the same way, we can compute a second principal component to produce $(\boldsymbol{\theta}_i^T \boldsymbol{v}_2) \sum_i v_{2j} b_j(s,t)$ and so on.

The authors are interested in computing principal components not affected by outliers. In many cases the most interesting problem is the dual one, that is, the identification of outliers, which means detecting structures in the data that deviate from the usual pattern. For instance, in clinical analysis we may be more interested in identifying patterns that may correspond to illness than in describing healthy individuals. This can be carried out in the feature space of the vectors $\boldsymbol{\theta}_i$, because it is sensible to expect that some type of aberrant behavior in the data space $S \times T$ will also be captured in the feature space. The analysis of the relationship between outliers in the data space and in the feature space is an interesting problem that requires a deep study. For instance, a single outlier in the data space due to some measurement error may lead to several outliers in the feature space. But also a group of outliers in the data space due to some differentiated behavior may lead to a single outlier in the feature space. The problem is further complicated as the generation of the information in the feature space (the computation of $\boldsymbol{\theta}_i$) may produce groups of masked outliers, in addition to those that might exist in the original data. Therefore, we should try to use robust estimates with a high breakdown point.

Several authors (see for instance Huber, 1985 and Jones and Sibson, 1987) have suggested that a useful way to detect outliers in multivariate samples is to search for univariate outliers on the projections of the data over a set of directions obtained by maximizing some criterion. This is the projection pursuit method, but the projection criteria to be used in order to have a powerful high-breakdown procedure to identify outliers is not clear. We have shown (Peña and Prieto, 1997) that a useful procedure to identify clusters of multivariate outliers is to look at the directions that maximize either the fourth central moment for the projected data or its kurtosis coefficient. If we apply this idea to the problem considered in this paper, we have to find directions of projection $\boldsymbol{d}_k$ by maximizing

$$\begin{array}{rcc} \boldsymbol{d}_k = & \arg\max & \dfrac{\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{d}^T\boldsymbol{\theta}_i^{(k)} - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{d}^T\boldsymbol{\theta}_j^{(k)})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{d}^T\boldsymbol{\theta}_i^{(k)} - \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{d}^T\boldsymbol{\theta}_j^{(k)})^2\right)^2} \\ & \text{s.t.} & \|\boldsymbol{d}\| = 1 \end{array}$$

where $\boldsymbol{\theta}_i^{(1)} = \boldsymbol{\theta}_i$ and in subsequent iterations

$$\boldsymbol{\theta}_i^{(k+1)} = \left(\boldsymbol{I} - \frac{1}{\boldsymbol{d}^T\boldsymbol{M}\boldsymbol{d}}\boldsymbol{d}\boldsymbol{d}^T\boldsymbol{M}\right)\boldsymbol{\theta}_i^{(k)}.$$

The outliers will be identified by computing univariate measures of distance r_i defined by

$$r_i = \frac{|\boldsymbol{d}^T\boldsymbol{\theta}_i - \text{median}(\boldsymbol{d}^T\boldsymbol{\theta}_j)|}{\text{MAD}(\boldsymbol{d}^T\boldsymbol{\theta}_j)}.$$

These measures can also be used as weights $w(r_i)$ for the computation of the scale estimator as a weighted sample covariance matrix. The approach is related to Stahel (1981) and Donoho (1982), but instead of searching directions at random we use the property proved in Peña and Prieto (1999) that outliers must increase the kurtosis of the projected data.

We would also like to suggest that a possible alternative analysis of this data set can be performed by using spatial time series (Bennet, 1979, Droesbeke, 1987). Then, each corneal image is represented by a realization of a spatial process and the way to identify structural behaviour will be to carry out a factorial time series analysis of this spatial process. We can also try to discriminate between normal corneal images and aberrant ones by performing a cluster analysis of these spatial time series. The ideas of Piccolo (1990) on clustering time series can be generalized to this setting.

Acknowledgements

D. Peña and J. Prieto acknowledge support for this research from DGES-MEC (Spain), grant PB96-0111.

References

Bennett, R.J. (1979). *Spatial Time Series.* Chapman and Hall.

Donoho, D.L. (1982). Breakdown Properties of Multivariate Location Estimators. Ph.D. qualifying paper, Harvard University, Deptartment of Statistics.

Droesbeke, F. (ed.) (1987). *Spatial Processes and Spatial Time Series Analysis.* Facultes Universitaires Saint-Louis.

Huber, P.J. (1985). Projection pursuit. *Annals of Statistics*, **13**, 435-475.

Jones, M.C. and R. Sibson (1987). What is projection pursuit?. *Journal of the Royal Statistical Society, A*, **150**, 1-36.

Peña, D. and F.J. Prieto (1997). Robust covariance matrix estimation and multivariate outlier detection. Working Paper WP97-08, Universidad Carlos III de Madrid.

Piccolo, D. (1990). A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, **11**, 153-164.

Stahel, W.A. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. Thesis, ETH Zurich.

Jim O. Ramsay
McGill University, Canada.

This fascinating paper marries some exciting ideas for the analysis and display of functional models with some important and challenging data. The result is some creative thinking that should have an impact well beyond the context of this application. The new ideas for the display of the data and principal components particularly impressed me. To react to all this is a real pleasure.

The authors wisely reduce the dimensions of the problem by replacing the original nearly 7000 discrete observations by a basis function expansion in terms of 66 Zernicke polynomials. These polynomials are a tensor product of familiar Fourier series functions and the Jacobi polynomials defined to be orthogonal over $[0, 1]$.

The authors refer to these polynomials as "features". What do they mean by this? A basis function can be called a feature if it captures some structure that is known a priori to contribute much of the variation in the data. Examples would be Fourier components in stationary signals, and perhaps B-splines and wavelets for data having local "bumps". Are these variations in corneal curvature that we see really well represented by specific Zernicke polynomial coefficients? If not, these the authors are just using one handy basis among other possibilities that manages to represent the data well in a manageable number of dimensions, and perhaps the term "feature" should be reserved for effects in the data rather than in the basis.

Can we imagine better bases?. The great virtue of B-splines and wavelets is their local character, which neither Jacobi polynomials nor Fourier series possess. Periodic versions of B-splines are available, and perhaps these defined for angular measure could be crossed with the usual B-splines for the radial dimension.

Or, to consider a rather different approach, perhaps these analyses could make use of finite element methods, now used widely to solve what are essentially regularization problems, but defined in terms of partial differential equations rather than explicit roughness penalties. I am finding in my own work, and especially for multidimensional arguments such as here, that the partial differential literature in general and finite element analysis in particular appear to have a great deal to offer problems such as this.

Missing data are indeed a central and difficult problem in functional data analyses. In fact, even in situations usually not thought of as involving missing data, we see similar issues arise. The estimate of the second derivative of a function at a point near the boundary can become dramatically unstable because the data become progressively more one-sided in the information that they convey. While the estimates are unstable, they are not, strictly speaking, outliers. Rather, outliers are usually understood to be actual observations that are wildly inconsistent with sensible model estimates, rather than model estimates that are wild because there are no data to define them.

The authors have elected to handle the missing data problem by ingeniously modifying principal components analysis so as to render eigenfunctions insensitive to this type of instability. In so doing, they have made a real contribution to robust estimation technology for functional data analysis, and this is sure to be useful in the presence of what we usually understand as outlying data.

An alternative approach is to use regularization, involving penalizing the roughness of estimated components, a topic that is a central theme in our book, Functional Data Analysis. When the data are not there, or are sparse, the estimated components simply become smooth, as seems reasonable. Regularization can also be thought of as a Bayesian approach to functional data, since the roughness penalty can correspond to a prior for the estimated function.

The regularization process can be viewed as borrowing information from neighboring data points. We can also borrow information from other entire images. That is, if a piece of an image is missing, and especially if it is on the periphery, it seems reasonable to fill in the image with data from other images that are in other ways similar it. This principle underlies what is variously called in the linear modeling literature empirical Bayes, hierarchical linear models, or multi-level analysis. In that domain, it is postulated that coefficients in a linear expansion are sampled from some population, usually taken to have a Gaussian distribution. One of the main applications of these methods is in fact to compensate for missing time values in longitudinal data. Brumback and Rice (1998), and the commentaries that accompany it, use multi-level analysis for curves with missing data that are represented by a linear combination of B-splines. Since these images are linearly expanded in terms of Zernicke polynomials, the application would seem to be direct.

Finally, an option to explore is the rotation of principal components to provide alternative and perhaps more easily described characterizations of corneal curvature aberrations. PCA has as its main objective the identification of a subspace within which much of the variability of the data can be defined. Of course, the eigenfunctions play a key role in characterizing that subspace, but this role is primarily computational. Once the subspace has been determined, any nontrivial set of basis functions spanning the same subspace, whether orthogonal or not, are potential candidates for describing what is happening within it. This principle, well understood by decades of psychometricians and other specialists in the analysis of behavioral science data, needs to penetrate more deeply other areas of application of PCA and related methods. These alternative coordinate functions should be chosen to more directly evoke the features that ophthamologists actually see through their instruments. Examples of rotating functional principal components using the VARIMAX criterion can be found in our book.

But there is so much to admire in this paper as it stands, and I am sure that analysts of functional data will derive benefit and stimulation from this work for many years to come.

References

Brumback, B. and J. Rice (1998). Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *Journal of the American Statistical Association*, **93**, 961-994.

Mariano J. Valderrama and Ana M. Aguilera
Universidad de Granada, Spain.

This paper presents a nice application of the Functional Principal Component Analysis (FPCA) to model ophtalmological data and an estimation procedure based on a robust approach. In fact, the manuscript can be clearly divided in two separated parts:

First, the functional data space of images is transformed into a space of feature vectors by least-squares fitting on a functional subspace spanned by an orthogonal basis of Zernike polynomials. Then, the stochastic problem is reduced to a finite set of random variables as it is usual in dimensionality reduction techniques.

Second, robust estimators of location and spread of the feature vectors are calculated in order to reduce the outlier influence in the estimation procedure of the FPCA.

The paper is written in a methodological way avoiding to give too much technical developments and involves a very interesting and convincing application about the behaviour of robust estimation and compression procedures with spatial data. Nevertheless, perhaps some specific aspects of the paper could be discussed and clarified by means of the following suggestions:

1. The approach described in this paper is supported on the choice of a Zernike basis so that the spatial FPCA is equivalent to the multivariate PCA of the coefficients (feature vectors) in terms of such a basis. Aguilera et al.(1999) have proved that this issue is valid for any Hilbertian random variable with values in a finite dimensional space.

In order to extend this methodology to more general situations when we deal with real data, different orthogonal basis could be considered such as Bessel functions (Ruiz and Valderrama, 1997) or two-dimensional wavelet functions that can be successful even for non smooth data.

2. Taking into account that the procedure used for obtaining the feature vectors introduces certain noise in the measures, it could be performed an interpolation of the images on the discrete data of the grid by means of two dimensional splines of a suitable order. In fact, Aguilera et al. (1996) have proved for the one-dimensional case that cubic splines provide optimum results with smooth curves. On the other hand, the interpolated images preserve the observed data on the grid by assuming not noisy sample information.

3. The estimation procedure developed in the paper is reduced to robust estimation of the location and spread measures of the feature vectors. Nevertheless, a generalization to the FPCA could be outlined by means of a direct robust estimation of the covariance operator instead of performing it through the Zernike coefficients, although it would give rise to a more

complicated problem.

Finally, an alternative way to find robust functional principal components would be to apply the "projection pursuit" approach developed by Croux and Ruiz-Gazen (1996) by maximizing a robust estimation of the variance.

References

Aguilera, A.M., R. Gutiérrez and M.J. Valderrama (1996). Approximation of estimators in the PCA of a stochastic process using B-splines. *Communications in Statistics: Simulation and Computation*, **25**, 671-690.

Aguilera, A.M., F.A. Ocaña and M.J. Valderrama (1999). Principal component analysis of Hilbertian random variables on finite-dimensional spaces. Technical Report 02/99, Department of Statistics and Operation Research, University of Granada.

Croux, C. and A. Ruiz-Gazen (1996). A fast algorithm for robust principal components based on projection pursuit. *Proceedings in Computational Statistics 1996*, (A. Prat, ed.) Barcelona, Spain, 211-216.

Ruiz, M.D. and M.J. Valderrama (1997). Orthogonal representations of random fields and an application to geophysics data. *Journal of Applied Probability*, **34**, 458-476.

Rejoinder by N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang and K.L. Cohen