Rejoinder

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

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

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

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

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

We appreciate the many fine points raised by all of the discussants. They have added much to the paper, and we have learned a lot. We are also very grateful for the many important references that have been added. Because complementary views have been provided by several discussants on a number of topics, we have chosen to organize this rejoinder by topic.

1 Data summarization and choice of basis

Most discussants agreed with our expressed need to summarize the data, and many interesting alternatives were suggested. Ramsay raises an interesting point about the use of the term "features" and "feature vectors" in this context. We borrowed this from the field of statistical pattern recognition, where it has become quite standard terminology. But we agree that it would be better to reserve use of the word "feature" for something found in the data, such as the bright red cone in Figure 2, not just a projection onto a basis element.

We have direct experience with the Zernike basis, as in the paper, and with a tensor product of the Fourier and Legendre bases, as discussed by Fan and by Peña and Prieto. The performance of these bases in this context is not so

different, and we got some similar PCA results using the Fourier-Legendre basis. However, the Zernike basis is more efficient in terms of giving similar representations with fewer terms (66 for the Zernike basis corresponds to about 120 for the Legendre-Fourier basis). A related problem with the Legendre-Fourier basis is that it has some unpleasant singularities at the origin. Most angular tensor product bases will have this problem, and it is a special property of the Zernike basis that its particular Jacobi component can be viewed as carefully chosen to avoid this type of singularity. The singularities were usually not a major problem for this basis, since the signals being fit are smooth, but we believe it was the need for basis functions to adapt properly for these singularities that entailed more coefficients being needed than for the Zernike basis.

As suggested by Marden and by Peña and Prieto, there definitely are some "knobs to turn" in fine tuning our method. The choice of number of terms in the Zernike representation certainly has an impact. As noted by Brumback and by Fan, this is a smoothing parameter, and affects what one sees in the familiar way. Here is a point where the clinical experience of Tripoli and Cohen was essential. We addressed this by looking at a set of raw data images, as in Figures 1 and 2, and comparing with their reconstructions as in Figure 3, for a number of different coefficient numbers. We chose 66 as best highlighting the important clinical features in these images, while at the same time minimizing noise. We have not tried it, but believe that the PCA will still find roughly the same directions for a wide range of coefficient numbers (but with more noise, or else more smoothing). Another knob to turn was the radius of the analysis region (most images extend beyond the 4mm radius shown here). If this radius is taken to be much larger, then even the robust Elliptical PCA can not suppress completely the influence of the missing data (because nearly every image then has edge artifacts). If the radius is taken much smaller, then the edge artifacts are reduced, and thus there is less need for robust PCA.

Summarization by B-splines were suggested by Brumback, and by Ramsay. We considered B-splines at an early point in the research, but only the traditional rectangular tensor product. We rejected it because it did not seem to fit naturally in our circular region of interest. However, Ramsey's idea of an angular tensor product, and the other variations, sound very sensible. The nice Brumback example may leave one skeptical (and it certainly highlights the importance of knot choice), but things are likely not that bad in the 2 dimensional world, because angular information will tend to help with some of the problems shown there. We suspect this approach could give similar performance to what we obtained with the Zernike basis.

Valderrama and Aguilera suggest some additional possibilities, based on wavelets, and on Bessel functions. We expect that as long as these methods can be adapted to our circular analysis region, they will also give similar good performance.

2 Dual problems

Kneip has pointed out a very powerful and promising approach to doing in PCA in high dimensional contexts, that was new to us. I. M. Chakravarti has remarked that this observation can be viewed as a consequence of Lemma (i) of Good (1969). The idea is well worth deeper investigation, and may even prove to be useful in examples such as ours (66 dimensions, but only 43 data points). We are reminded of the "dual problems" found in the simplex method for linear programming.

Brumback has pointed to a different use of the term "dual", which is also an important concept for analyzing populations of complex objects, when they are summarized by feature vectors.

3 Rotation of Principal Components

Ramsay brings up an important point about principal components in general, and functional data in particular, which is that PCA should be viewed as "finding low dimensional subspaces", and these are not always best represented by the eigenvectors found by the original analysis. The original eigenvectors were satisfactory for the set of normal cornea images analyzed here, but we had exactly the problem Ramsay describes with other sets of corneal data.

Even with the normal corneal images, we have contemplated (but have not yet tried) investigating other subspaces. For example, the first optometrical measurement is "spherical curvature", and we could study the component of the data in that direction by doing PCA only on the Fourier order 0 basis elements in the Zernike representation. The second optometrical measurement is astigmatism, which shows up mostly in the Fourier order 2 terms, so something similar could be done in that direction.

4 Performance of spherical and elliptical PCA

The bias effect in the elliptical PCA discovered by Boente and Fraiman is very interesting. We wonder if this small effect could perhaps get worse in higher dimensions. The clinically relevant results we got for the cornea data suggest that this was not a major problem in that particular case. We believe this was because the distributional major axes were roughly parallel to the coordinate axes, as suggested by the middle panel of Figure 11, and thus bias was small. However, it will likely turn out be important to understand this effect for other data sets.

We also enjoyed the simulations of Boente and Fraiman on the "failings" of spherical PCA. With any statistical method, insight comes from "stretching it until it breaks", and a good job of that has been done here. The key to this example is that the eigenvalues are very close, so the "principal direction" becomes a rather fuzzy notion, and then a large amount of contamination is added in a particular direction. When the contamination is large enough, it

easily overwhelms the difference between the eigenvalues and gives systematically wrong answers. This can be viewed as showing that the breakdown point of spherical PCA is quite small when the important eigenvalues are close to each other.

To investigate this, we replicated their experiment, except that we changed the eigenvalue matrix from $\Lambda = (6, 5)$ to $\Lambda = (8, 4)$ and $\Lambda = (\frac{3}{2}, \frac{3}{4})$ (eigenvalues now separated by a factor of 2). The results are shown below.

$\Lambda = (8, 4)$	N_{15}	N_{30}	N_{45}
Classical, C_0	0	0	0
Spherical, C_0	0	0	0
Classical, $C_{0.1}$	500	500	500
Spherical, $C_{0.1}$	77	7	1

$\Lambda = (\frac{3}{2}, \frac{3}{4})$	N_{15}	N_{30}	N_{45}
Classical, C_0	0	0	0
Spherical, C_0	0	0	0
Classical, $C_{0.1}$	500	500	500
Spherical, $C_{0.1}$	58	4	1

It appears that the breakdown in the spherical method, observed by Boente and Fraiman doesn't occur in these examples. In their example, the first theoretical eigenvalue only accounts for 54.5% of the total variation of the data. When we convert the data to the sphere, the eigenvalues become approximately $\Lambda_S = (0.53, 0.47)$. In both of our examples, it is 66.7%, and the converted eigenvalues are approximately $\Lambda_S = (0.58, 0.42)$. Influence function calculations, similar to those in Croux's (3) show why there is breakdown in one case, but not the other. Using notation similar to Croux, the influence function of an eigenvector is bounded by

$$IF((x_1, x_2), v_1, F) \leq \frac{1}{2(\lambda_1 - \lambda_2)}.$$

When ϵ exceeds this value, the estimate will break down, since the amount that the angle moves (in radians) is

$$\sin^{-1}\left(\frac{\epsilon}{2(\lambda_1 - \lambda_2)}\right).$$

Note that this angular measurement will be undefined when $\epsilon > 2(\lambda_1 - \lambda_2)$. The estimate becomes very poor (i.e. off by 45°) when $\epsilon \approx \sqrt{2}(\lambda_1 - \lambda_2)$. In their example the 10% contamination exceeds $\sqrt{2}(0.53 - 0.47) \approx 0.085$, so breakdown of the estimator is to be expected. In our examples, spherical PCA showed some signs of weakness even though there was not total breakdown. We also performed the simulation where the leading eigenvalue accounted for 90% of the total variation, $\Lambda = (9, 1)$, and this admirably withstood one-third contamination. Since the key to breakdown of spherical PCA is the difference between

eigenvalues for the sphered data, we give a table of approximate values for these,

First PC's %	λ_1	λ_2	$\sqrt{2}(\lambda_1 - \lambda_2)$
55%	0.53	0.47	0.085
60%	0.54	0.46	0.11
66.7%	0.58	0.42	0.23
70%	0.59	0.41	0.25
80%	0.67	0.33	0.48
90%	0.75	0.25	0.50

.

We conjecture that other robust approaches will have similar problems with breakdown in this type of situation, which motivates construction of a diagnostic based on eigenvalues. This would not be straightforward because of the problems with interpreting spherical and elliptical eigenvalues pointed out by Brumback, and by Croux. This problem can perhaps be tackled by replacing the eigenvalues with sums of squares of projections of the data (these *are* the eigenvalues for ordinary PCA, but not for spherical PCA) The eigenvalues are quite different for the cornea data, so this does not seem to be a practical problem here, but it seems well worth knowing as it may appear with other data sets.

Croux has elucidated some very interesting and useful properties, including equivariance, the influence function and efficiency. We note that the dependence of the asymptotic relative efficiency (compared with classical PCA) on the eigenvalues is an inherent feature of estimation with $n > d$. An interpretation of his "50% upper bound on the efficiency" may be that when one uses only directional information, in a two dimensional context, half of the information is lost. Given the increasing efficiency of the spatial median as the dimension increases (see Marden's comment), we conjecture that the efficiency bound for spherical PCA will also increase, possibly to 1, as the dimension increases. Croux's suggestion (ii) looks promising to address this inefficiency.

We don't agree with Croux's suggestion that elliptical PCA is an "equivalent of correlation PCA". To understand this point, note that correlation PCA would be nearest to doing PCA in the lower right hand plot of Figure 18. However, elliptical PCA is done in the left hand plot, which can give quite different results.

We were interested to find that Marden (1999) had independently developed the idea of spherical PCA. This seems to be an idea whose "time has come". Marden's remark about the population PCA directions being the same as the spherical PCA directions was very insightful and interesting. A word of caution about the notion of "coordinate-wise symmetry" is that for (non-trivial) empirical distributions, this seems to have the minimal requirement of $n \geq 2^d$, which seems quite far from the $n < d$ situations present for this type of data.

We agree with Kneip's observation that PCA is quite capable of finding "interesting directions", even when the data are not normally distributed. The need for normality is more about classical multivariate hypothesis testing, than it is about finding directions.

5 More approaches to missing data

Ramsay nicely elucidates the difference between "missing data" in individual images and "outliers" in the population sense. However, our missing data problems do cause "outliers" in precisely Ramsay's sense, i.e. data points that are "wildly inconsistent with sensible model estimates". For example, when looking at projections onto eigenvectors that are driven by outliers, e.g. the direction shown in Figure 7, plots of the distributions show that the outliers are easily 8 or more standard deviations from the mean. Such plots were not put in the paper to save space, but are visible in the lower half of the accompanying MPEG movies. See for example `norm200.mpg`. Another way to see that we have "outliers" in this sense is the bottom panel of Figure 11. If these were multivariate Gaussian data vectors, then the bottom curves would all lie in about the range $(-3, 3)$, but there are a number of cases that go far outside that range.

Brumback's analysis of the missing data problem nicely clarifies the problems of the Zernike basis in this context. The wildly discolored regions in Figure 3 are caused by the type of extrapolation illustrated in the simple example, and the effect is heavily magnified by looking at second derivatives. Brumback's suggestion for how to counter this effect provides a nice solution to the problem posed by Fan of "how can we do robust imputation?" A possible downside is that it may be computationally very slow to work with the 6912×6912 covariance matrix $\widehat{V}$.

Also promising is the regularization - Bayesian approach to the outlier problem, suggested by Ramsay. This is especially natural when doing summarization by B-splines.

6 Alternate versions of PCA

The possibility of "depth" based approaches to PCA suggested by Fraiman and Boente sounds promising.

We did consider some projection pursuit approaches, as suggested by Croux, by Peña and Prieto and by Valderrama and Aguilera, but were too intimidated by the very high dimensionality for our data. Examples we have seen tend to be in something like 4 dimensions, with 10 dimensions already causing concern. Our 66 dimensional space is very large, and we were not confident of being able to find an algorithm that would avoid likely problems with multiple optima, etc. Peña and Prieto seem to hint at an approach which could address the multiple optima problem. Another approach may be to use elliptical PCA directions as starting values for this type of approach, and then refine by iteration. The local optimum found by this method would probably be useful, and might be better than the elliptical PCA, especially in view of the bias problem pointed out by Boente and Fraiman.

Robust estimation applied directly to the continuous covariance operator, as suggested by Valderrama and Aguilera sounds well worth further study. One

approach could be to use the spherical or elliptical projection idea in that domain.

7 Robust location estimation

Croux made some very useful suggestions about improving the numerical performance of the L1 location estimate. Although starting with the mean vector caused us no trouble with our data, we anticipate that starting with the componentwise median, as Croux suggests, instead of the median will improve the performance of the algorithm.

Fan asks why we use the spatial median as a final location estimator instead of the componentwise median. Certainly the componentwise median is a possible replacement. We note, however, that the spatial median already has a 50% breakdown point (see Marden's comment), so it has good "global" robustness. He and Simpson (1992) derived optimal "local" robustness of the spatial median for directional data. Brown (1983) preferred the spatial median on the basis of its rotation equivariance and its increasing efficiency as the dimension increases.

Clearly, finding an appropriate tradeoff between robustness, efficiency and smoothness in high dimensions is a challenge. We anticipate many improvements to our initial approach. Indeed the discussants have already introduced many promising ideas.

8 Additional topics

Kneip makes a good point about the importance of "errors in variables" methods. We agree that there is a need to extend these ideas in the direction of functional data analysis. A good starting point may be the monograph Carroll, Ruppert, and Stefanski (1995). We also agree about the importance of registration of functional data, and that was an issue in our analysis, that we solved by using "pupil center" information.

Peña and Prieto have anticipated some upcoming work, by asking how this methodology can be used for more than just describing populations, but in fact to find problems with corneal shape. Work is currently under way on methods for the identification of Kerataconus, as shown in Figure 2.

The spatial time series approach of Peña and Prieto to these data sounds interesting.

Ramsay's suggestion of the use of finite element models also has some appeal. An advantage is that one could perhaps make use of the many known physical properties of the cornea.

References

[1] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*, Chapman & Hall, London.

[2] Good, I. J. (1969) Some applications of the singular value decomposition of a matrix, *Technometrics*, 11, 823-831.

[3] He, X. and Simpson, D.G. (1992) Robust direction estimation, *The Annals of Statistics*, 20, 351-369.