Project Description

Section 1 discusses results from previous NSF support. Section 2 is for "interdisciplinary research", involving application areas where the development of new statistical ideas and methods are expected to lead to important scientific progress in ophthalmology, computer science, software engineering and geology. These motivate much of Section 3, which is for "infrastructural research", i.e. the creation, development and mathematical analysis of statistical methodologies which are anticipated to be of general purpose use (useful in the above specific fields, but with potential to go far beyond). Three major underlying themes of the proposed research are the analysis of the structure of populations of complex objects, visualization (both of the structure of the populations, and for other purposes) and the significance of features observed in a smooth. Proposed Details about who will be working on these projects is given in Section 4.

1 Results from previous NSF support

1.1 Joint research grant with J. Fan

Work by J. S. Marron, supported by NSF Grant DMS-9504414 resulted in 18 papers published or accepted by the journals *Annals of Statistics* (2), *Biometrika* (2), *Computational Statistics* (2), *Journal of Computational and Graphical Statistics*, *Journal of Nonparametric Statistics* (2), *Journal of the American Statistical Association* (4), *Probability Theory and Related Fields*, *Statistica Neerlandica*, *Statistica Sinica*, *Statistics and Computing*, *Statistics and Probability Letters.*, on topics including bandwidth selection, wavelets, image processing, and the smoothing of sparse data. There are 6 papers published in unrefereed proceedings, or still pending. There were 30 different co-authors on these papers, from 11 different countries.

1.2 National Institute of Statistical Science

J. S. Marron was also supported by NSF Grant SBR-9529926, under the auspices of the National Institute of Statistical Science, for research in organizational theory, management and software engineering. This project involved the statistical study of a very large scale software system. This is a "legacy"

system, meaning it has been continually evolved and adapted over a long period of time, by thousands of programmers. Staudenmayer, Graves, Marron, Mockus, Siy, Votta, Perry (1998) synthesized major management lessons by combining statistical analysis of aspects of the change management history with insights from interviews with the developers. Eick, Graves, Karr, Marron and Mockus (1998) found the first statistical verification that "code decays", i.e. as a legacy system is continually patched, it becomes harder to make future changes. SiZer, developed in Chaudhuri and Marron (1997) was an integral tool for this purpose. Graves, Karr, Marron and Siy (1998) developed a time based model for "fault potential", which has been effective for studying the impact of changes on the system, and has formed the basis for subsequent analysis.

2 Proposed cross disciplinary research

Close scientific contact has already been established, and in some cases substantial work is already completed or under way, with researchers in Ophthalmology, Computer Science, Software Engineering, and Geology. Proposed joint research is organized by discipline.

2.1 Opthalmology

Joint work is proposed with K. L. Cohen, in the UNC Department of Ophthalmology, on cornea curvature maps. These are representations of corneal shape for individual patients, based on radial curvature. We propose studying *populations* of these maps, using methods in the spirit of the rather new statistical area of "curves as data" or "functional data analysis", see Ramsey and Silverman (1997). But this research will carry that field of statistics beyond just curves and into "images as data". A second important contribution of this research, well motivated by this application, but with a great deal of potential for many others, is to make this type of analysis resistant to the influence of outliers.

Ramsey and Silverman elegantly showed that some classical multivariate analysis methods, such as Principal Components Analysis, provide very powerful tools for comprehending the structure of populations of functional data. Direct implementation of their ideas from the corneal data set is impractical, because the dimensionality of the problem is too large. Here (and

we anticipate in many other high dimensional problems) summaries of the individual elements of the population become essential. This need has been well understood for years in the field of statistical pattern recognition, where many methods of summarizing complex data structures by "feature vectors" have been developed. For example, Devijver and Kittler (1982) provide a number of approaches to this problem.

For the analysis of the cornea data, we have summarized the structure of corneal images by feature vectors consisting of the coefficients of least square fits of the Zernike orthogonal basis (see e.g. Schwiegerling, Grievencamp and Miller (1995) for discussion of how this basis relates well to the optical characteristics of lenses). This gives a two phase approach to the analysis of this type of population. The statistical backbone, methods related to PCA, is done in the feature space (where the computations are tractable), but the presentation/interpretation of the results is done in the original curvature map space, which is where the clinician has insight that can make use of the information that is summarized in the analysis.

Part of a preliminary example is shown in Figure 1a. The population here is a set of corneal curvature images from $n = 43$ "normal" eyes. The images in the center of Figure 1a represents the radial curvature in the mean of the feature vectors. A gray level color scheme is used here because of the likely lack of color printing capability of most reviewers, but for serious work color maps highlight important structures quite effectively. The faint hourglass shape at the center of this image is characteristic of the radial curvature of an astigmatic lens. This fits well with clinical knowledge that most eyes have some astigmatism, and that it is typically vertical in direction, called "with the rule".

PCA allows insight into the structure of the population about this mean, by finding "directions of greatest variability". The right and left images in Figure 1a give insight into one such "direction", by representing the radial curvature in the vectors that are at both the positive and negative extremes of the data, in the given direction.

To save space not all of the PCA directions are shown here, but each turned out to highlight a clinically known component of the population. Just one direction, the third, PC3, is shown in Figure 1a.

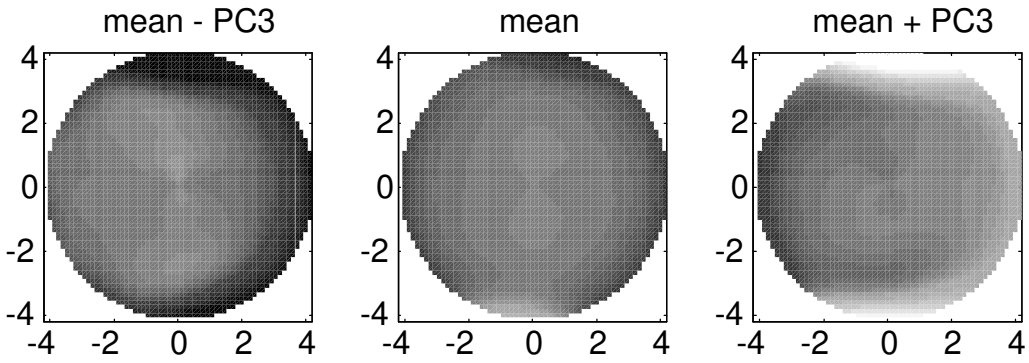

Figure 1a: Third direction of variability, obscured by edge effects.

It is quite difficult for most clinicians to understand which features of the population are presented here because these shapes are not typical in raw data images. Deeper investigation showed that they were due to some "outliers" in the population, where the fit of the Zernike feature vector was poor, due to missing data in a number of the images. At least some missing data is endemic to Keratron images, due to eyelids and other artifacts appearing in the base image. PCA is easily influenced by outliers, and that is what is happening here. Simple attempts at outlier deletion failed, because when some outlying images were deleted, other images then affected the analysis in a similar way. Since this led to deletion of too large a component of the population, and since all of the images contain at least some useful information, a "bounded influence" approach to PCA appears natural.

Discussion of possible approaches to robust PCA, as well as some new ideas, are given in Section 3.2. Figure 1b shows the effect of the method of Elliptical Principal Component Analysis on this same population of normal corneas. Again, the analysis is centered at the sample mean. This practice is very dubious from a robustness point of view, since a robust estimate of the mean is important to avoid many potential pitfalls. But Figure 1b is intended only as preliminary work, and even this crude first approach serves to illustrate that robust methods are very useful in this context.

Large benefits from robustifying the PCA are apparent in the fourth column, where the direction represented by the 3rd Elliptical Principal Component is shown. Unlike the images for PC3, these show a clinically known aspect of the population, which is "with the rule vs. against the rule astigmatism" (note that the hourglass shape changes direction to horizontal in

the top image, which is a less common direction for astigmatism, than the vertical direction shown in the other images).

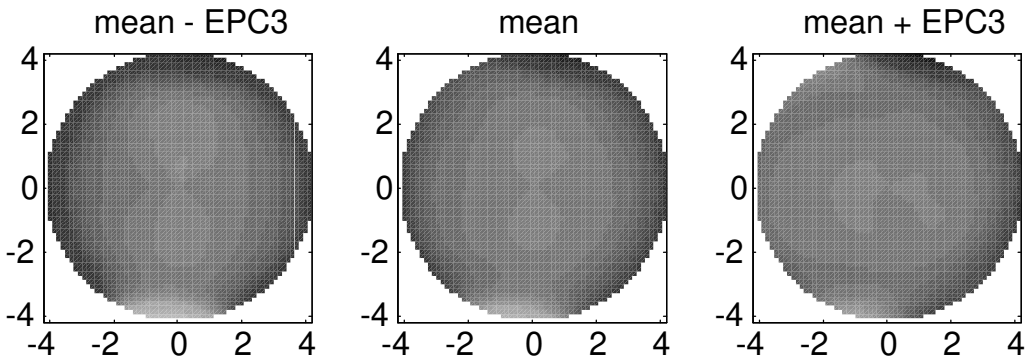

Figure 1b: Clinically clear images from Elliptical PCA, highlighting astigmatic direction.

When a robust modification of the mean is used, further improvements are expected. For example the light region along the bottom left edge of each image here is probably caused by the effect of an outlying observation on the mean. We have found a number of other such examples, where outliers (caused by eyelid and related edge problems) introduce shapes in the PCA that are not of clinical interest, which are removed by this type of robustification.

Using images of this type to understand the structure of populations requires substantial experience and practice. To address this issue, an improved visual interface has been developed based on dynamic graphics, i.e. a movie has been made. Good insight is conveyed by a movie which essentially "morphs" back and forth between the images shown in Figure 1b. Since the human eye is well adapted to incorporate dynamic visual information, we find this makes our approach much easier to explain to others, and also allows for deeper insights. For example, based on his clinical experience, Cohen was able to see the "with the rule vs. against the rule" lesson even in the nonrobust movie version of Figure 1a, despite the substantial distortion caused by the outliers. This same type of dynamic graphics has also been used to visualize the effects of the healing process over time, following PRK Laser surgery.

Proposed new work is to turn this methodology toward solving clinical problems. First will be an aid to the diagnosis of Kerataconus, where a "cone" or region of high curvature appears on the surface of the cornea. Other attempts at "automatic diagnosis" have been proposed, see e.g. ????. These methods have had good success rates, but they use "black box" methods, such as neural nets, which don't give insight as to *how* a given kerataconic cornea is different from the normal population. One improved approach will use the PCA (or a proper robust version) for a large population of normals, and then compare a suspected kerataconic image, in terms of its projections onto the eigenvectors of the PCA. The directions where the projection is outside the normal range gives insights into how the cornea differs from normal,which will be made quantitative by giving tail probabilities which reflect how "non-normal" the diseased image is. An alternate approach will be based on a PCA of a combined population of normal and kerataconic populations, with directions showing how the populations differ, and a likelihood ratio approach (based on the first few PCA directions only) used for discrimination.

Further proposed work in this two population direction is to replace the PCA, which only measures total overall variability, with related analysis which attempt to better separate the two populations. A number of possibilities, including Principal Discriminant Analysis, are described for example in Chapter 6 of Devijver and Kittler (1982).

2.2 Computer Science

The above concept of studying populations of complex objects by summarizing them using feature vectors, doing PCA in the feature space, and then interpreting the results in the original data space has been developed in parallel, by computer scientists, who typically focus on two and three dimensional shapes of body parts. Important references include Coots, Hill, Taylor and Haslam (1993) and Keleman, Székely and Gerig (1998).

Recent close contact with the groups of S. M. Pizer, and G. Gerig, in the UNC Department of Computer Science, has already resulted in cross-fertilization of ideas, and much more is anticipated in this direction. Regular biweekly meetings are currently under way.

2.3 Software Engineering

A problem of current interest in software engineering is measurement of the effect of perfective maintenance on a legacy software system, and assessing how worthwhile it is. Perfective maintenance consists of changes made to the system, that are made for neither the addition of a new feature nor for fixing a fault in the system. For example, organizational changes can be made to a part of the code where increasing complexity has rendered it difficult to understand (and thus to cleanly make future changes). Since such changes are made at substantial cost, the benefits yielded is a critical question.

As a first attempt at addressing this question, for a particular code unit, a time based "rate of perfective maintenance" for a particular code unit is shown as the heavy curve in Figure 2a. This curve is essentially a kernel smooth of the times that perfective changes were made, divided by a kernel smooth of the times that all changes were made, so this curve is on the scale of "proportion of changes". However, the exponential kernel function is used, to model the concept that the effects of a given change in terms of "fault potential" tail off exponentially in time. This model was seen to have good prediction properties in terms of prediction of faults by Graves, Karr, Marron and Siy (1998), and the rate of decay estimated there is used here (essentially the bandwidth of the kernel smooth). Note that the rate of perfective maintenance generally increased over time. This is quite common, because as code grows older, the need for perfective maintenance is more strongly felt.

To understand whether or not perfective maintenance helps in avoiding faults, the times of changes made to the code to fix faults are studied. These times are shown as jittered dots, labelled "event times" in Figure 2a. If perfective maintenance indeed aids in lowering the incidence of faults, then more of these dots should be occurring at times where the perfective maintenance proportion is low, and fewer where it is high. An approach to investigate this is to compare this set of events with a randomly chosen "background set of events", shown as the lower set of dots in Figure 2a. In this case the background events were chosen randomly from all the changes made to this code unit, to properly account for the fact that changes were not made homogeneously in time. The statistical relationship between these two sets of events and the proportion of perfective maintenance curve is then investigated by evaluating the height of the curve at each of the event times, and studying the respective populations. Insight into how these populations

compare is given by ordinary Gaussian kernel density estimates, shown on the left side of Figure 2a, in a "sideways" fashion, which highlights that they are representing these populations of numbers that are "evaluations of the curve at the even times". The density estimates suggest that the actual events occurred more frequently when the curve was lower. The relative shapes suggest that this idea can be quantitated using a hypothesis test for the difference of population means. The p-value for this, shown in Figure 2a is seen to be statistically significant.

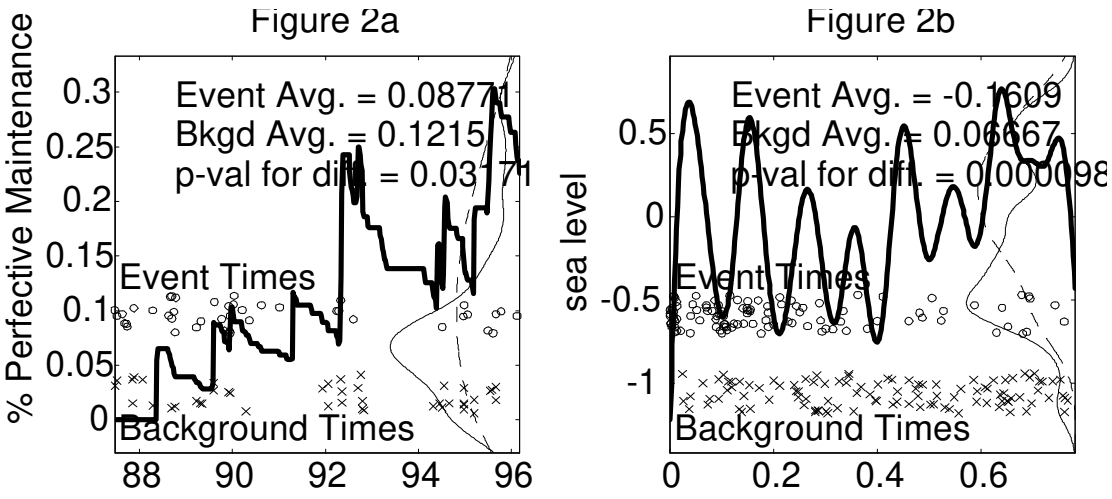

Figure 2: Curve Height at Event Time analysis of data sets from (a) Software Engineering (b) Geology. Both study connections between the heavy curves, and the "event times" shown as jittered points. Both show a statistically significant connection because the structure of the populations evaluated at the curves are different from appropriate "background populations", as respresented by the sample means, and by the thin kernel density estimates, shown vertically, solid for the events and dashed for the background times.

This was the first statistical evidence verifying that perfective maintenance has a significant connection to fault incidence. Future proposed work will be in the direction of trying to establish a causal relationship (here only a connection has been shown), and in quantifying "how much" benefit there is (here only the fact that "there is some" has been established).

This new general type of analysis is called "Curve Height at Event Time Analysis", and has proven to be useful in a quite different application as well: Figure 2b will be discussed in Section 2.4. There are other approaches to studying correlations between point processes that could also have been used for this set of data, for example ???? of Chapter ??? of Cox and Isham could also address this issue. However, these seem less attractive in the present application, since they do not make use of the special knowledge of how changes in the code affect the potential for faults, that was discovered by Graves, Karr, Marron and Siy (1998), and is incorporated in the proportion of perfective maintenance curve.

2.4 Geology

Curve Height at Event Time Analysis has also been useful in an application in geology, in some joint work with Allen Glazner, of the UNC Department of Geology. The question addressed is the possibility of a connection between glaciation and volcanic activity. If such a connection can be established, then it is worth investigating possible mechanisms, such as the weight of the ice impedes volcanic activity.

Level of glaciation is correlated with sea water oxygen isotope ratios (via sea level) which has been precisely measured over time from planktonic foraminifera. As seen in Glazner, Manley, Marron and Rojstaczer (1998), some preliminary smooths of this curve, when compared with similarly smoothed kernel density estimates suggested a tendency for the volcanic activity to be higher during periods of less glaciation. The scale of the smoothing suggests the time scale on which the connection occurs, roughly 100,000 years.

The CHETA plot in Figure 2b gives an assessment of the statistical significance of this. The thick curve in Figure 2b is a smoothed version of the oxygen isotope ratio over time. The smoothing eliminates components which have a higher frequency than the cycles of volcanic activity. This sea level curve is low during periods of major glaciation.

The "event times" shown as jittered dots in Figure 2b are times of volcanic activity. The structure of the population of these times evaluated at the oxygen isotope curve is shown with the (again "sideways") kernel density estimate, shown as the thin solid curve on the left side. The highest peak in this distribution is near the lower part of the oxygen isotope curve, which suggests that these events tend to occur when the curve is low.

Statistical significance of this effect is again quantified by CHETA. This time the background events are simulated uniform random variables. When the oxygen isotope curve is evaluated at these points, the resulting values tend to be substantially larger. Again this is made precise with a two sample mean test, which shows a strongly statistically significant result.

As with the software engineering example above, CHETA only established a connection between glaciation and volcanic activity. This connection motivates interest in possible mechanisms, which will be investigated in future work.

For these data, the point process approach of ??? is unavailable, since the oxygen isotope ratio curve is not point process data, but instead is presented in tabular form as a curve.

Other joint work is proposed with T. J. Bralower, of the UNC Department of Geology. Bralower et. al. (1997) have constructed a curve for Strontium isotope ratios over time, for the interval from 90 to 125 million years ago. These ratios have been determined from fossil shell material, and the fossils themselves also provide age information. The application of their results however is in dating unfossiliferous rocks and in determining the rates of oceanic volcanic crust production and continental weathering in the past. The data are shown in Figure 2 of Chaudhuri and Marron (1997), where SiZer shows which changes in the Strontium isotope are statistically significant, and which can not be distinguished from sampling variability.

Proposed additional work in this direction includes taking an "error in variables" point of view. In particular we will modify the current SiZer analysis to take explicitly into account the fact that both the Strontium isotope dependent variable, and also the dates used for the independent variable have errors. A good summary of why it is important to adopt this point of view (since it can influence results), as well as a good collection of statistical methods to use, may be found in Carroll, Ruppert and Stefanski (1995).

The next proposed project is to address "registration issues", to combine information across several studies, whose differences were noted in Bralower et. al. (1997). An early approach to this problem, based just on linear transformations, was developed in Härdle and Marron (1990). More sophisticated "warping function" approaches, such as described in Section 5.4 of Ramsey and Silverman (1997) and Gasser and Kneip (1995) are expected to be essential.

Another point where advanced statistical methods are anticipated to yield important scientific gains is in the dating of fossils by evolutionary methods.

This is currently done by ad hoc methods based on the shapes of the fossils. But precise mathematical modelling of shapes, and their statistical analysis has recently been rapidly evolving in the statistics community. Dryden and Mardia (1997) provide excellent access to and discussion of this literature. Proposed work in this direction involves adapting "shape space ideas", in particular the "Procustes analysis" methods, to the time series of fossil shapes. Shape analysis analogs of conventional time series methods, such as smoothing, and error modelling, will then yield evolutionary dating methods, which should be more objective and more accurate, and also allowed improved assessment of the accuracy.

3 Proposed infrastructural research

The work proposed in this section addresses problems of a mathematical statistics nature that are motivated by the above real world applications. It is anticipated that this research will both address direct questions raised by the applications, as well as provide classical mathematical statistical theoretical underpinnings for the new methods developed.

3.1 Curve Height at Event Time Analysis

CHETA is seen in Sections 2.3 and 2.4 to be a powerful data analysis technique. Because it is new, none of its mathematical properties have been developed yet. In the simplest case, the curve being evaluated is a "gold standard", which is nearly true in Figure 2b (except some smoothing has been done to remove higher frequency oscillations that are not important to this analysis). Then, when the background events can be chosen appropriately, the validity (i.e. correct size) of the test seems clear, but there are questions to settle about power of the test.

The more mathematically challenging context is where the curve being evaluated is random, e.g. based on a curve estimate, as in Figure 2a. It is conjectured that when the event times are independent of the data used to construct the curve estimate, the test will be asymptotically valid (i.e. the actual size converges to the nominal size). Joint work with P. Hall is proposed on the problem of bandwidth selection for the curve being evaluated. This will different from standard bandwidth selection because the natural criterion will be to maximize the power of the test. Joint experience

with bandwidth selection, see e.g. Hall and Marron (1987a,b,c, 1988, 1990, 1991a,b), Härdle, W., Hall, P. and Marron (1988, 1992), Hall, Jones, Marron and Sheather (1991), Hall, Marron, and Park (1992) , Grund, Hall and Marron (1994) and the survey papers Jones, Marron and Sheather (1996a,b), is expected to be very valuable in this research.

A variation of the test which may be useful in some contexts is to replace the final hypothesis test based on the mean, with a test about the respective population variance, or the two evaluated distributions could be compared by some classical measure of "goodness of fit". See e.g. Stephens (1974) for good discussion of such measures.

3.2 Analysis of Populations of Complex Objects

A robust variation of Principal Component Analysis was seen in Section 2.1 and Section 2.2 to provide a powerful tool for understanding the structure of populations of complex objects. This motivates work in a mathematical statistical direction, which it is anticipated will generate ideas that will feed back into the applications.

Important references on robust statistical methods include the monographs of Huber (1981), Hampel, Ronchetti, Rouseeuw and Stahel (1986), Staudte and Sheather (1990) and Rouseeuw and Leroy (1987). The literature on robust Principal Components can be roughly divided into two groups of methods.

The first approach involves finding a robust estimate of the covariance matrix, and then doing an eigen analysis. An intuitively appealing approach to this is based on "minimum volume ellipsoid" ideas, see Section 7.1 of Rouseeuw and Leroy (1987). See Campbell (1980), Devlin, et. al. (1981) and Galpin and Hawkins (1997) for related approaches. The assumption that the number of data points, n, is smaller than the number of dimensions, d is critical to many of these methods. While it is very convenient, e.g. otherwise it is not even straightforward to produce candidate estimates that are both rotationally invariant and also covariance matrices (i.e. nonnegative definite), the assumption of $d < n$ rules out most data sets of the type discussed in Section 2.1 and Section 2.2. Especially in the psychiatric application of Kelemen, Székely and Gerig (1997), the imaging and rendering of each volume, is an enormously complicated and expensive task.

The second approach to robust PCA is based on projection pursuit type methods. See for example Li and Chen (1985), Wang and Karhunen (1996),

and Chen, Bandoni and Romagnoli (1996). This is based on the characterization of PCA as the "direction of greatest variability". The main idea is to replace the classical variance as the measure of "variability", by some more robust indicator of scale. While this is intuitively attractive, the price for using a robust criterion not based on squared error is that the problem no longer admits an eigen analysis solution, and instead projection pursuit (i.e. numerical iterative search) type methods are needed. This is feasible for small values of d, but becomes very problematic for moderate to high dimensions, such as $d = 66$ as was used in Section 2.1.

These considerations motivate other approaches to robust PCA. S. M. Pizer suggested an approach based on ranks, and the Spearman correlation coefficient (which is the usual Pearson correlation applied to the ranks of the data). The matrix of Spearman correlation coefficients is nonnegative definite, but has the disadvantage of not being rotationally invariant. A more serious drawback is that it only provides good robustness against outliers in the coordinate directions. To understand why these directions are not adequate for high dimensional data, recall that points on the vertices of the unit cube in d dimensional space have distance d from the origin, i.e. they are "gross outliers" with respect to the unit ball! To make the good idea of the Spearman correlation work in this context, one needs a multivariate notion of "ranks" that is rotationally invariant. Since finding important "directions" is the goal of the analysis, a simple approach is to replace each data point, by its corresponding direction vector (where the "origin" is the "center" of the data). The PCA is applied to the set of unit vectors, and the result is called "Spherical PCA".

A weakness of Spherical PCA is it assumes that all coordinates have similar scale. This is not a good assumption for the corneal image data in Section 2.1, since the Zernike basis models these images in the same way the Fourier basis models smooth curves, i.e. some coefficients are orders of magnitude different from others. A simple adjustment is to first scale the variables, e.g. dividing by the Median Absolute Deviation, for each coordinate, mapping to a sphere, and then rescaling. This is equivalent to mapping each data point to the surface of an ellipse. Applying PCA to the data on the ellipse gives "Elliptical PCA", which gave the excellent result shown in Figure 1a.

The mean shown in Figures 1a and 1b also shows some edge effects caused by "outliers", this motivates using also a robust mean estimate. Many such are discussed in the monographs Huber (1981), Hampel, Ronchetti,

Rouseeuw and Stahel (1986), Staudte and Sheather (1990) and Rouseeuw and Leroy (1987). D. Simpson suggested a robust mean that is much in the spirit of Spherical PCA, which can also be modified as for elliptical PCA.

Work of a more mathematical statistical nature will be the analysis of the robustness of these methods, via calculation of influence functions. Of key interest for the corneal data is the effect of outliers on the PCA *directions*. The notion of "bounded influence" will need some modification, since the direction vectors belong to a compact set. See Higuchi and Eguchi (1998) for work on an influence function approach to PCA.

D. Ruppert has suggested a hybrid of the above methods, where one first finds a fairly wide subset of interesting directions by Elliptical PCA, then refines the robust PCA using a projection pursuit method (now more tractable, because of the reduced dimensionality.

3.3 Statistical Significance of Features

SiZer was developed in Chaudhuri and Marron (1997) as a tool for exploratory data analysis. It uses ideas from scale space theory in computer vision. An excellent overview of scale space theory is given in Lindeberg (1994). Scale space is essentially a family of Gaussian kernel smooths indexed by the bandwidth.

SiZer is a visual device which conveys very directly which features in a smooth of the data are significant, and does so simultaneously for a wide range of bandwidths (i.e. over scale space). In addition to providing a visual paradigm which makes analysis much more direct than classical "confidence band" approaches, there are two other important innovations in the SiZer approach.

First, in the spirit of scale space, a wide range of bandwidths is simultaneously considered. Marron and Chung (1997) suggested guidelines for analyzing data by constructing an overlay of a family of smooths. Study of the full scale space (as opposed to classical "bandwidth selection") is important because different types of information can be available at different levels of smoothing. Another advantage is that this avoids the problem of choosing a bandwidth, which has been very challenging (and also contentious, see Chapter 4 of Hall (1992) for a good overview and additional references) for approaches such as confidence bands.

Second the bias problem which has plagued the confidence band approach (see Härdle and Marron (1991) and Chapter 4 of Hall (1992)) is skirted by

using an interpretation which is widely accepted in computer vision, but is new to statistics. The fundamental idea is that "features" exist only as they can be seen in a family of smooths. In particular, their "target" is not the true underlying curve (usual in statistics), but instead the underlying curve after it has been smoothed. This is essentially the expected value of the smooth based on random data, so it is enough to find confidence intervals centered at the expected value, which is relatively simple (in particular there is no bias). Then "significant features" are interpreted as "significant at the given level of resolution (i.e. bandwidth)".

SiZer highlights "features", such as bumps, in terms of the derivative of the smooth (a bump is where the smooth goes significantly up, then significantly down) via a color map. In this map the x-axis represents location, the y-axis represents bandwidth, and pixels are shaded blue (red) where the derivative is significantly increasing (decreasing) in the sense that both ends of a simultaneous confidence interval are above (below) zero. Regions where the confidence intervals contain zero are shaded in the intermediate color of purple. A suitable adjustment deals with the multiple comparison issue of simultaneously studying a number of confidence intervals. Detailed discussion, with real data and simulated examples, can be found in Chaudhuri and Marron (1997) and Marron and Chaudhuri (1998a,b). SiZer provide an important tool in the real data analyses described in Sections 1.2 and 2.4.

The development of several generalizations of SiZer are proposed. Some will be relatively straightforward and direct. Already under development is the straightforward replacement of the "significant slope" aspect of SiZer with "significant curvature". Preliminary examples show that in most cases both methods find essentially the same features. However in a few instances, the curvature version can find features that the slope version cannot, and vice versa. It will also be straightforward to extend the SiZer approach to different estimation settings, such as to hazard estimation, to the censored data case, to local likelihood contexts (such as logistic regression) and to additive models. Also of interest will be the "errors in variables" modification described in Section 2.4

Another variation on the SiZer scheme is to base it on different types of smoothing methods. For example smoothing splines will adapt very readily. B-splines apparently will take additional effort, because of their discrete smoothing parameter. Nearest neighbors will likely be quite interesting, since such a SiZer could be viewed as a"vertical distortion" of the present SiZer, and other such distortions (e.g. multiples of some other local bandwidth

function) may be of interest as well.

A much more challenging extension of SiZer type methods will be to dependent data, i.e. to time series analysis. The need is clear, since often there is the question of significance features (bumps, etc.) in the trend of a time series. At first glance, it may appear that the extension will be straightforward, since one only needs to model the dependence, find suitable estimates, and then appropriately adapt the SiZer significance tests. However deeper thought reveals that there is a fundamental problem in deciding which part of the signal is "trend" and which part is "long range dependence". In some sense this difficulty can not be resolved from a single time series realization (statistician A may prefer to use the label "trend", for something that statistician B prefers to call "long range dependence"). A proposed approach to this problem is somewhat in the spirit of the scale space viewpoint: simultaneously consider a wide range of possible dependencies, from "most the structure in the data is viewed as trend" to "most of the structure in the data is viewed as random". When coupled with an effective display device allowing visual insight into this trade off, the statistician can then see how "significance of features" trades off against different types of assumptions about the underlying dependence.

Another very challenging new context in which to try this type of idea is the two dimensional case. The payoffs for this are potentially very great, since potential applications are very broad in range from image analysis (determining which features in an image are "really there", as opposed to being artifacts of background noise) to epidemiology (learning from maps of locations of disease incidents, which "hot spots" are statistically significant, and which are simply reflecting the natural Poisson variability). Because such problems arise in both density and regression contexts, both will be developed (as was done in 1d with SiZer).

A problem is that scale space is not so easily represented (the direct analog of the SiZer map is now 3 dimensional colored solids, which are much harder to visualize, although some possibilities do exist, see e.g. Foley, et. al. (1990), Kaufman (1991) and Yagel and Hagen (1997)), although this can be overcome by presenting several slices. A harder problem is that "slopes" and "curvatures" are now much more complicated entities than they are in one dimension. Several approaches will be tried and compared. One is to consider local estimates of the gradient, with a small arrow indicating the gradient direction, when it is significantly different from zero. Another is to study directional derivatives, and use a symbol indicating which of these

are significant. Curvature versions will be based on the eigenvalues of the estimated Hessian matrix, with different colored dots used to represent: (i) both eigenvalues significantly positive, (ii) one eigenvalue significantly positive, the other insignificant, (iii) both eigenvalues significant, one positive and the other negative, (iv) one eigenvalue significantly negative, and the other insignifcant, (v) both eigenvalues significantly negative. Finally attempts will be made to combine both slope and curvature information.

Mathematical statistical work on SiZer is also proposed, joint with P. Hall. Scale space motivates work on a not well studied stochastic process, the family of kernel smooths jointly indexed by location and the bandwidth, so Hall's very broad experience with Gaussian processes is viewed as an essential component of this research. The first conjectured result is that "SiZer works", i.e. that the statistical inferences are asymptotically valid, which requires showing that the scale space process converges in law to a particular Gaussian process. A weak version of this was established in Chaudhuri and Marron (1997b), but this result only applied to a range of bandwidths bounded above 0. Deeper results will be attempted first for a range of bandwidth tending to 0 at interesting rates, e.g. $\left[an^{-1/5}, bn^{-1/5}\right]$, and then for broader ranges such as $[n^{-1+\varepsilon}, C]$, and will explore appropriate modes of uniform convergence.

A point where insights gained from this theoretical analysis are expected to directly aid in the development of the method is the multiple comparison problem. The inference underlying the most useful version of SiZer is currently simultaneous only in the direction of location, not bandwidth. Inference in the bandwidth direction is currently implemented with a simulation based approach, but this is quite slow. Appropriate characterization and asymptotic approximation of the full scale space surface is expected to yield a computationally tractable method for obtaining simultaneous inference in both directions.

Research is also proposed in another direction where significance of features is a central issue. A much different approach to image analysis is "medial cores", see Pizer, et. al (1998), Morse, et. al. (1998), Furst, et. al. (1997) and Furst and Pizer (1998). These provide a powerful tool for finding regions of interest called "shape loci" on which deeper analysis is performed. A problem to be addressed in applications of medial core technology is that some cores are "real structure", and others are "noise artifacts". It is proposed to develop SiZer like methods for assessing significance of core structures.

4 Manpower

Funding is requested for two students. The first, N. Locantore, is already working on the cornea project, proposed in Section 2.1, with the theoretical robustness work in Section 3.2 anticipated as the theoretical component of his dissertation. The second student is not yet decided, but the Statistics Department has a large pool of well qualified candidates, and discussion is already under way with R. Z. Li (winner of the 1997 Hoeffding Fellowship, given to our) and Z. J. Zhang (winner of the 1998 Hoeffding Fellowship). The second student will work on computational and theoretical aspects of the generalization of the SiZer idea to complex entities such as cores, as discussed in Section 3.3.

Funding is also requested for a one month visit by P. Hall to UNC, for work on the bandwidth selection aspect of CHETA, and the mathematical analysis of SiZer.