Statistica Sinica **14**(2004), 789-808

THE FUNCTIONAL DATA ANALYSIS VIEW OF LONGITUDINAL DATA

Xin Zhao[1], J. S. Marron[2] and Martin T. Wells[1]

[1]*Cornell University and* [2]*University of North Carolina*

Abstract: Longitudinal data can be viewed as a type of functional data. The functional viewpoint is not typical for most analysts of longitudinal data, but provides a route for powerful new insights. The potential of this approach is demonstrated through an analysis of periodicities in a microarray gene expression data set.

Key words and phrases: Classification, Fourier subspace, functional data analysis, longitudinal data, SiZer analysis, time-course micro-array experiments, yeast cell cycle.

1. Introduction

The statistical area of Functional Data Analysis seems to have evolved as a "parallel statistical culture" to the better known and larger area of Longitudinal Data Analysis. While the data sets are quite similar, the viewpoints and ways of thinking about the data tend to be quite different. A major goal of this paper is to make some directly applicable ideas from FDA more accessible to the LDA community.

This is done through a simple toy example in Section 2, and through a novel real data example in Section 4. The data, described in Section 3, are from a micro-array gene expression study of cell cycles. The goal is to identify genes which have periodic behavior, that are thus strongly associated to the cell cycle process. FDA ideas are seen to provide a clear path towards achieving this goal. This is intended to illustrate the idea that FDA methods can be a very useful addition to the toolbox of longitudinal data analysts.

Direct understanding of the FDA viewpoint comes from consideration of the "atom" of a statistical analysis. In a first course in statistics, atoms are "numbers", and methods are learned for understanding populations of numbers. In a course in multivariate analysis, the atoms are vectors, and methods for understanding populations of vectors are the focus. FDA can be viewed as the generalization of this, where the atoms are more complicated objects, such as curves, images or shapes. In the statistical literature, the case of curves is most commonly treated, see Ramsay and Silverman (1997) for a good overview. Functional data analysis of images was done by Locantore, Marron, Simpson, Tripoli,

Zhang and Cohen (1999). For a range of different approaches to FDA of populations of shapes, see for example, Cootes, Hill, Taylor and Haslam (1993), Kelemen, Szekely and Gerig (1997), Dryden and Mardia (1998) and Yushkevich, Pizer, Joshi and Marron (2001).

While there are many types of FDA, most can be connected with one of two common goals. One goal is "understanding population structure". Relatively simple analyses of this type are the focus of this paper. The second common goal of FDA is discrimination (also called classification). The latter goal is not covered in this paper. However, readers interested in this topic may want to investigate Distance Weighted Discrimination, developed in Marron and Todd (2002), which is an improved version of the Support Vector Machine of Vapnik (1982, 1995).

A useful framework for understanding functional data analyses, and a good starting point for a new analysis, is the concept of parallel "data spaces". The original data objects (curves, images or shapes), are members of the "object space". However, for numerical manipulation, objects are typically represented as vectors. Borrowing terminology from statistical pattern recognition, these vectors are called "feature vectors", and the collection of all possible feature vectors is called the "feature space". A one to one mapping between the object space and the feature space is a very useful device for understanding FDA, as illustrated in Section 2.

Additional insight comes from thinking of the feature space (a set of vectors) in terms of a "point cloud". If the dimension of the feature vectors is 3, then the point cloud is the conventional 3d scatterplot. For higher dimensions, it is useful to think of a "higher dimensional point cloud". This point cloud intuition, used in tandem with the mapping to the object space, provides a useful conceptual framework. This structure provides major insights into a toy example in Section 2, and provides a clear pathway to pursue a wide variety of data analyses, that is followed in one partiuclar direction for some real data in Section 4. Some concluding remarks, and indications of future work are given in Section 5.

2. Functional Toy Example

A toy example data set, that is used to illustrate simple functional data analytic methods, in the context of curves, is shown in Figure 1a. Visually one sees "a set of 50 parabolas plus low noise", with some additional structure. Different colors are used to allow good visual separation of the different curves. FDA methods provide a tool for understanding the variation in this population. Figure 1a is a graphical representation of the object space.

Curves can be easily represented as vectors, i.e., the object space can be mapped to the feature space, by "digitization". Digitization means that the

curves are evaluated at an equally spaced grid, and the values form the entries of a vector.

The object space can be mapped back to the feature space by the "parallel coordinates" methodology. Parallel coordinates were proposed by Inselberg (1985), see also Wegman (1990), as a means of visualizing high dimensional vectors. The idea is to plot the sequence of coordinate values as a time series (where "time" is coordinate number), and then linearly interpolate the plotted values. Figure 1a was actually generated as $n = 50$ vectors of dimension $d = 10$, then parallel coordinates are used in the display. In all such plots in this paper, the horizontal axis should be viewed as "time", and the vertical axis should be thought of as a longitudinal variable of interest.

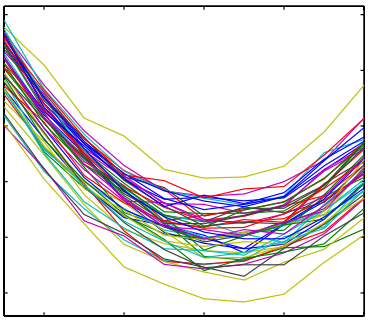

Figure 1a. Toy "raw data set" of curves. For illustration of concept of "understanding population structure". The horizontal axis represents time, and the longitudinal variable of interest is on the vertical axis.

Principal Component Analysis is frequently a very powerful method for understanding population structure in FDA. The essential idea goes back at least to Rao (1958), who proposed this method for studying populations of growth curves. The general idea of PCA is so good, that it has been rediscovered and renamed many times. For example, it is called the Karhunen Loeve method in electrical engineering, Empirical Orthogonal Functions in geophysical areas, Proper Orthogonal Decomposition in applied mathematics, and Factor Analysis in many other fields (particularly unfortunate, since the term has a deeper meaning in psychometrics, where the name was coined).

Figure 1b shows part of a PCA of the data set of curves shown in Figure 1a. The upper left hand panel shows the first important aspect of the population: its "centerpoint". In the particular, this is the sample mean. It is computed as a vector mean in the feature space, and the corresponding curve in the object space is shown here (again using a parallel coordinate view).

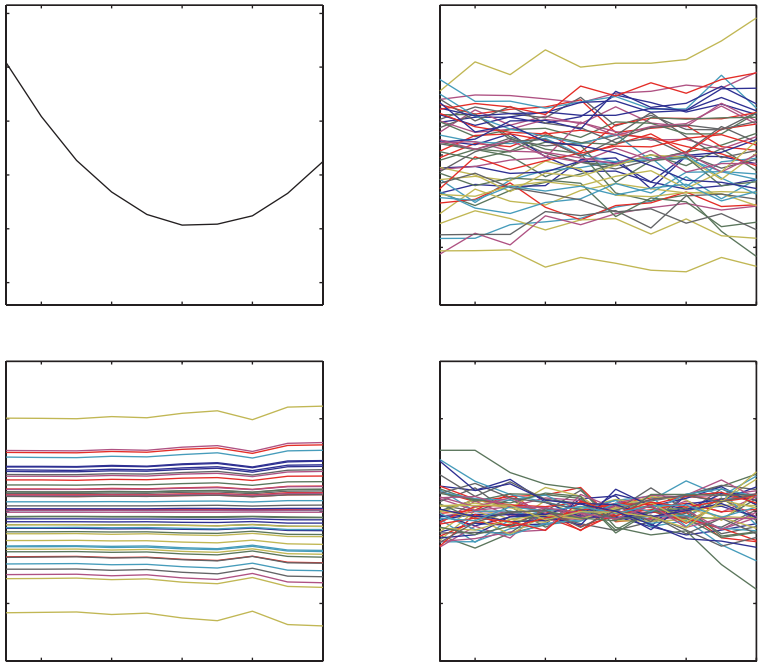

Figure 1b. Decomposition of variation in toy data set from Figure 1a (using the same axes). Upper left shows the mean. Upper right is residuals from mean. Lower left is projections of the mean residuals in the PC1 direction. Lower right is further residuals from PC1 projections.

The next step in the analysis is to study variation in the data about the mean. This is done by subtracting the data (as vectors in the feature space) from the mean, to get mean residuals. Insight into the mean residuals comes from the object space plot, shown in the upper right panel of Figure 1b. Note that the "parabolic structure" so visible in the data is now seen to be reflected only in the mean, and there is no variation having that basic shape. In the point cloud space (where there is the most intuition) these residuals correspond to recentering the point cloud so the mean is at the origin.

PCA gives a further insightful decomposition of the variation in the population. In the point cloud space, PCA seeks to find the direction vector with greatest population variation, in the sense of maximizing the variance of the data projected onto the vector. This direction is found by either an eigenvalue analysis of the covariance matrix, or by a singular value decomposition of the data matrix (these are equivalent). The results of PCA are classically studied by looking at the numerical values of the entries of the eigenvector (the "loadings"). But for FDA applications, a more useful view is usually that shown in the bottom left panel of Figure 1b. In this view, each data point is projected onto the direction

vector (this is computed as multiples of the eigenvector, where the cofficients are the inner products of each data curve with the eigenvector). Each of these is a point in the feature space, and thus has a curve (i.e., object space) representation. These curves are overlaid to provide insight into the "population structure". The bottom left panel shows that the dominant component of variability is "vertical shift". Armed with this insight, note that the same structure can also be seen in the mean residuals on the upper right, and also in the raw data in Figure 1a.

The still unexplained structure in the data are summarized in the residuals shown in the lower right panel of Figure 1b. These can be viewed as the difference of the upper right and lower left panel. They are also the result of projecting the data on the subspace that is orthogonal to the PC1 direction. Note that there is substantially less "variation" apparent in this family of curves than can be seen in the upper right, because the PC1 projections in the lower left "explains most of the variation".

These "amounts of variation" can be usefully quantified by an ANOVA style sum of squares analysis. The sum of squares explained by PC1 (i.e., of the curves on the lower left) is 86% of the sum of squares of the mean residuals in the upper right (thus given on the usual R^2 scale). The sum of squares in the PC1 residuals on the lower right is thus 14%. These numbers fit well with the visual impression.

Further insight into the structure of the population comes from finding the direction of greatest variation for the residuals shown in the bottom right of Figure 1b. This gives the second PCA direction, and projections of the data in this direction are shown in the top panel of Figure 1c. This shows a "random tilt" structure that is hard to see even in the residuals on the lower right of Figure 1b, and is virtually impossible to see in the raw data, or the mean residuals. This "tilt component" represents 10% of the mean residual variation.

The upper right panel shows the residuals from PC2. These are now much smaller in terms of variation, and represent only about 4% of the mean residual sum of squares.

The lower row of Figure 1c shows the analogous projection in the third principal component direction. The PC3 direction, and the corresponding residuals, are both very small, and also uninteresting. The reason is that this toy data set was simulated by starting with a single parabola, adding a random shift, adding a smaller random tilt, and adding some independent and identically distributed Gaussian noise. Once the parabola, the shift and the tilt have been subtracted, the resulting point cloud is essentially spherical Gaussian. Thus there are no strong directions of interest, and PCA decomposition finds random and uninteresting directions. This pattern has been seen in further analysis as expected, but is not shown here to save space.

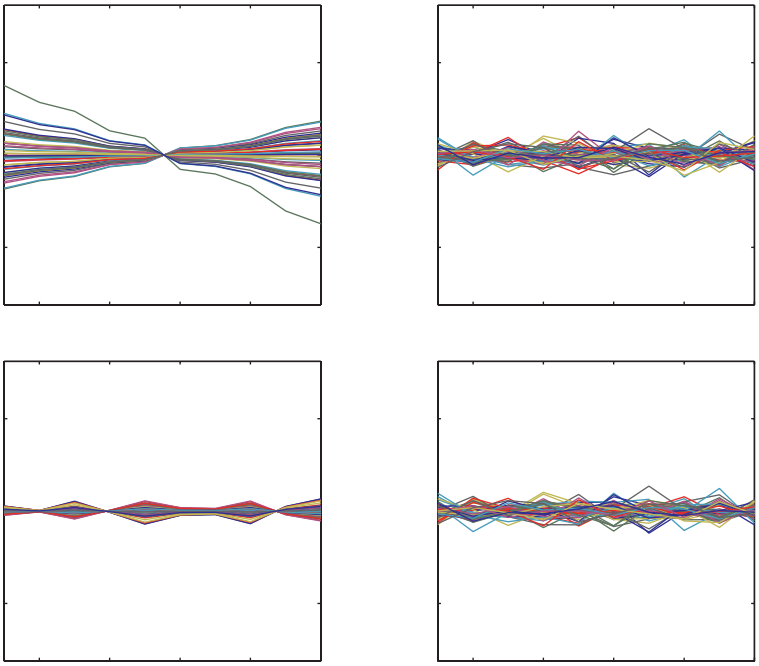

Figure 1c. Further visual decomposition of the variation of the data set from Figure 1a (using the same axes). Left side are projections in PC2 (upper) and PC3 (lower) directions. Right side are respective further residuals (starting from the lower right of Figure 1b).

In summary, this was an example of how PCA, coupled with FDA visualizations, can effectively analyze "population structure", through decomposition of the variation into intuitive components. It found some non-obvious characteristics of the population, and the visualization was made quantitative using the sum of squares analysis. We suggest that this can be a useful addition to the toolbox of Longitudinal Data Analysts.

The performance of this tool, and more important, the use of FDA ideas to generate novel data analyses is illustrated for real data in Section 4. The data are discussed in Section 3.

3. Cell Cycle Gene Expression Data

The data set studied here is from the micro-array measurement of gene expression. Micro-array measurements are complicated, and involve large amounts of pre-processing. That work was done for the data analyzed in this paper by Spellman, Sherlock, Zhang, Iyer, Anders, Eisen, Brown, Botstein and Futcher (1998). The data set is available from the web site: http://genome-www.stanford.edu/cellcycle/. In general, micro-array data represents a world

of new statistical challenges because the data tend to be "High Dimension Low Sample Size" (HDLSS). For example, classical multivariate analysis is useless in HDLSS settings, because the first step of such an analysis is to "sphere the data" by the root inverse of the covariance matrix, which is impossible because the covariance is not of full rank.

The data analyzed here come from Spellman et al. (1998), who ran three experiments to study gene expression during the "yeast cell cycle", the process of yeast cells splitting for reproduction. Here we focus on the α factor-based synchronization experiment. The experiment started with a collection of yeast cells, whose cycles were synchronized by a chemical process. A time series of cDNA micro-arrays was gathered over 18 equally spaced time points, over about two hours, i.e., two cell cycles. Gene expression was measured for the full 6,178 genes in the yeast genome. For simplicity of analysis (recall our main goal is to illustrate FDA ideas) we focus here on only the $n = 4,489$ genes for which there are no missing values.

An important goal of the study was to find which genes have an expression pattern that is related to the cell cycle. Spellman et al. (1998) identified 800 genes as periodic, but for simplicity we will again focus only on the genes with no missing values, leaving us with 612 that have been previously classified. These genes were also classified according to phase. In Section 4, we illustrate FDA ideas by revisiting this gene selection and classification. The data are viewed as a population of $n = 4,489$ time series (curves) of length $d = 18$.

Figure 2a shows a first view of the data, in a format very similar to that of Figure 1a. While there is a large amount of variation in the data, there is no apparent periodic structure (perhaps not surprising, since Spellman et al. (1998) flagged only 612 of the $n = 4,489$ genes as periodic). In all of these plots the horizontal axis is again time, but now the vertical axis is the relative level of gene expression.

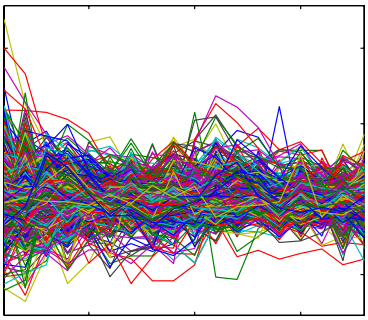

Figure 2a. Raw cell cycle gene expression time series. Horizontal axis represents time (over 2 cell cycles), vertical axis is relative level of gene expression. Each curve shows how the expression of a gene evolves over time.

One approach to finding important underlying structure in the data, such as the expected periodicities, is the PCA method, as illustrated in Figure 1b. This approach is taken in Figure 2b. The mean and the mean residuals (the top row in Figure 1b) are not shown here, because the mean is essentially 0, so the mean residuals are very close to the original data. The top row of Figure 2b shows the projections on the PC1 direction in the upper left, with the corresponding residuals in the upper right. The bottom row shows the projections of the data onto the PC2 direction on the left, and the residuals on the right.

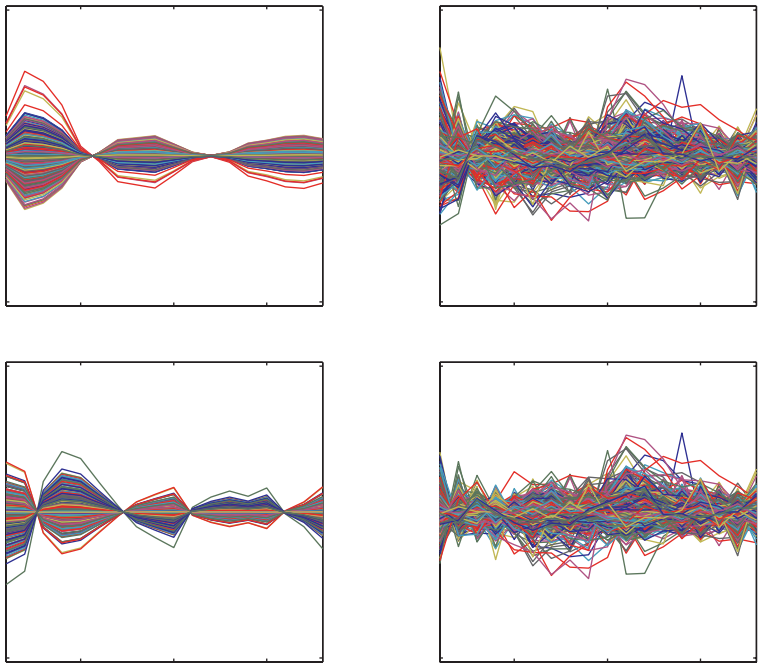

Figure 2b. PCA of gene expression time series from Figure 2b (with the same axes). Left panels are projection onto first (top) and second (bottom) PC directions (analogs of the bottom row of Figure 2b and the top row of Figure 2c). Right panels are corresponding residuals. Show nearly periodic structure, but not quite period 2.

While the PC1 and PC2 directions appear to represent systematic structure (verified by PC1 containing 25% of the variation about the mean, and 16% for PC2), they unfortunately do not reveal the frequency 2 periodic structure expected from the two cell cycle design of the experiment. This seems to be caused by the very large amount of noise present in the micro-array experiment, magnified by the fact that nearly the whole genome is being studied, while only a fraction of the genes are expected to be associated with the cell cycle. Additional principal components were considered, but are not shown here because they were no more insightful than what is shown in Figure 2b.

One approach to the problem of the standard PCA not revealing periodic structure would be to smooth the data curves, to reduce the noise. However, in this case the period is essentially known, so a more powerful approach (avoiding the bias introduced by smoothing) is to focus explicitly on the known periodicity of the data. This is done in the analysis of Section 4, by projection onto appropriate subspaces (i.e., directions in the point cloud space) that better represent periodicities.

Li, Yan and Yuan (2002) use a simple statistical model to describe the expression curves in the Spellman et al. (1998) data. They used a simple three step procedure in their analysis: (1) the use of standard principal component analysis (on the raw data) to suggest basis curves; (2) the use of nested models for organizing gene expression patterns; and (3) the construction of a compass plot using known cycle-regulated genes for phase determination. See Zhang, Yu, Singer and Xiong (2001) and Zhang and Yu (2002) for an interesting recursive partitioning based approach to the identification of important genes.

4. Periodic Functional Data Analysis

The analyses of this section are based on noise reduction, through projection onto the set of frequency two periodic functions. This is implemented by taking the Fourier Transform (representation as projection onto an orthonormal basis of sin and cos functions, see Bloomfield (2000) or Brillinger (1981) for details) of each time series, and keeping only the even frequencies. Intuitively this corresponds to projecting the 18 dimensional point cloud of data onto the 8 dimensional subspace generated by the evenly periodic functions. This subspace includes *all* functions of period 2 (meaning the set of vectors $\{(x_1, \ldots, x_{18})' : x_{i+9} = x_i, i = 1, \ldots, 8\}$), not just the phase shifted period 2 sin wave. This projection can also be viewed as a rotation of the point cloud (the Fourier Transform is orthonormal, and thus can be viewed as "rotation") followed by a reduction of the dimension to only the dimensions of interest. This reduction of the data is shown in Figure 3a.

At first glance the periodicity is not apparent, but a closer look reveals that the left half is an exact copy of the right half. This shows that this 8 dimensional projection includes much more than simple phase shifts of the period 2 sin wave. Strong evidence of the periodic structure in these data is revealed by the fact that the sum of squares of the curves in Figure 3a is 48.4% of the sum of squares of the raw data shown in Figure 2a. This is a larger share of the variation than is explained by the first two principal components, shown in Figure 2b, because the denominator is the sum of squares of projection onto the 8 dimensional subspace, so a large amount of the overall noise is already excluded. Note that again simple frequency 2 structure is not easily seen in the raw data.

The Principal Component Analysis of the "period 2 projected" data in Figure 3a is shown in Figure 3b. As in Figure 2b, PC1 is on the top, PC2 is on the bottom, the projections are shown on the left, and the corresponding residuals are shown on the right.

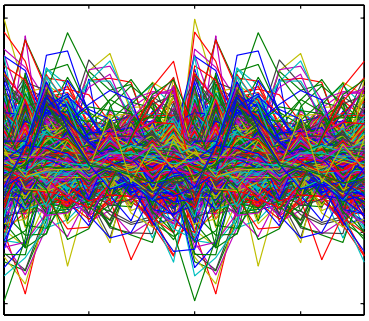

Figure 3a. Cell cycle time series (from Figure 2a, with the same axes), projected onto the subspace of frequency 2 harmonics. This is the "even periodic components" of the data from Figure 2a (note the left half is a replication of the right half).

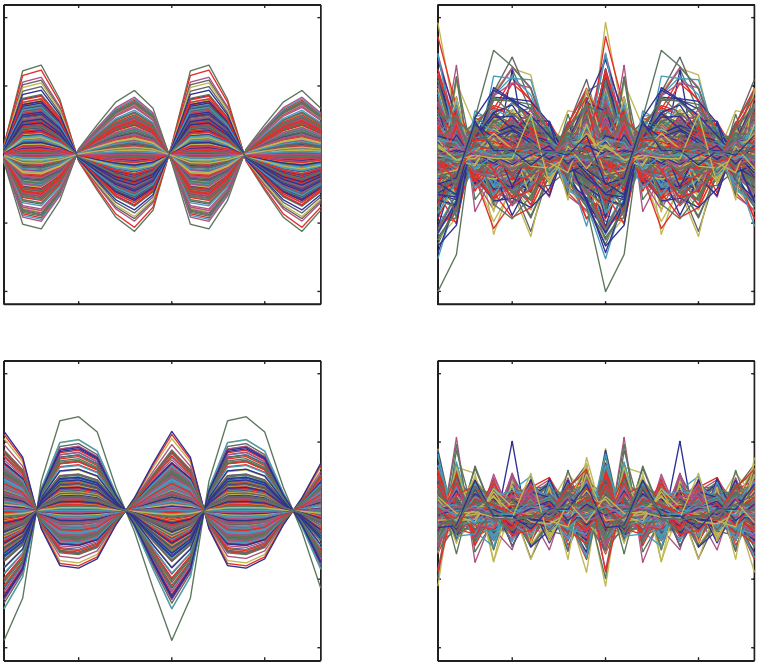

Figure 3b. PCA of frequency 2 projected cell cycle data (from Figure 3a, using the same axes), where the subplots are direct analogs of Figure 2b. PC projections on left, residuals on right. PC1 on top, PC2 on bottom. The two dominant directions are similar to the period 2 sin and cos functions, suggesting strong period 2 components.

An interesting feature of Figure 3b is that the first two Principal Component directions look much like the classical sin and cos functions, respectively. Between the two, these explain 65% of the variation of the data shown in Figure 3a (thus around 32% of the original raw data). Note also that linear combinations of these functions capture all phase shifts of the frequency 2 sinusoids, since the phase shift φ can be written as

$$\cos(x-\varphi) = \cos\varphi\cos x + \sin\varphi\sin x = c_1\cos x + c_2\sin x. \tag{1}$$

This motivates further reduction of the data, projecting onto the two dimensional space of only $\sin x$ and $\cos x$. Projection in this direction results in the representation of the gene as simply a pair of numbers, which is effectively represented as a scatterplot in Figure 4. Each gene appears as a plus sign, where the x-axis (y-axis) shows the projection on the vector corresponding to $\cos x$ ($\sin x$, respectively). It also follows from (1), and the polar coordinate representation of the data, that the phase φ will appear as the angle from the x-axis in the 2d projection plot shown in Figure 4.

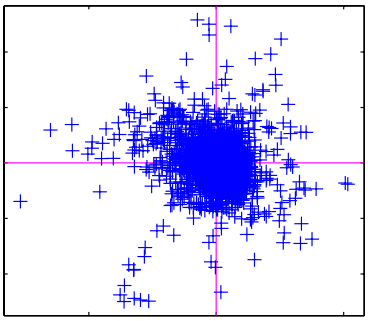

Figure 4. Scatterplot representation of the full data set, projected onto the subspace of the period 2 sin and cos function (cos on the horizontal, and sin on the vertical, axes). Angle from the positive x-axis represents phase. Suggests that several phases are dominant.

The scatterplot of sin-cos projections in Figure 4 shows that most of the $n = 4,489$ genes are very near the center, indicating no periodic behavior. However a reasonable number are farther away, which are the genes of interest because they have strong periodic behavior. These periodic genes appear in several directions from the origin, which indicates that different genes are active with respect to different phases of the cycle. An especially interesting structure is the "ray" appearing on the lower left, of a number of strongly periodic genes with nearly identical phase. Note also that there is no clear, simple division into "periodic"

and "aperiodic" genes. Instead all genes have a "degree of periodicity" (quantified as distance to the origin), which is large for some, and quite small for most, with a full range of values in between.

A referee suggested displaying Figure 4 as a "distance-phase" plot, where the distance to the origin is plotted on the horizontal axis, and the phase (angle from the positive x axis) is displayed on the vertical axis. We can see ways in which that view is useful, but ended up with a personal preference for the sin-cos information, as well as the clock type "angle corresponds to time in the cell cycle" ideas, that are conveyed in this polar coordinate view of the projections.

Spellman et al. (1998) used biological information to group the genes (that they flagged as periodic) into 5 classes according to function during 5 important phases of the cell cycle. Here we revisit this classification, using our two dimensional projection periodic representation. We base this on the phases of the genes, which are just the angles from the positive x-axis in Figure 4. To allow the analysis to be driven by the "most periodic" genes, we focus on the "top 200", in the sense of largest distance from the origin in Figure 4. The choice of 200 was made, after doing the analysis for the thresholds 200, 400, 600, 800 and 1000, on the basis of giving the cleanest division of genes in relation to the previous analysis.

Figure 5 shows a SiZer analysis of the top 200 phases. In all three panels of Figure 5, the x-axis represents the full range of phases $\theta \in [0, 2\pi]$. The top panel shows the phases in two forms. Near the top is a "jitter plot" (first proposed by Tukey and Tukey (1990)), where each gene is represented as a green dot. The x-coordinate shows the phase, and the y-coordinate is a random height that is used simply to spread the data for convenient visualization. The other representations of the data are a family of "smooth histograms" (more precisely "kernel density estimates", see e.g., Wand and Jones (1995)) shown as blue curves. These indicate a number of "bumps", i.e., "clusters in the data". The reason for overlaying a number of curves is that these show a range of different histogram binwidths, representing different amounts of smoothing, important because different clusters show up at different smoothing levels. A minor technical point is that, because of the periodic nature of the data, a "circular design", where copies of the data are placed beyond the opposing ends, is used to avoid boundary effects.

The blue smooth histograms, and also the green jitter plot, suggest both clusters (where the genes are "more dense than usual") and gaps ("less dense", respectively). Most of these clusters correspond well to the five important biological classes found by Spellman et al. (1998), although some of the boundaries are not all that clear. The center and bottom panels of Figure 5 aid in finding these boundaries, shown as the black vertical lines.

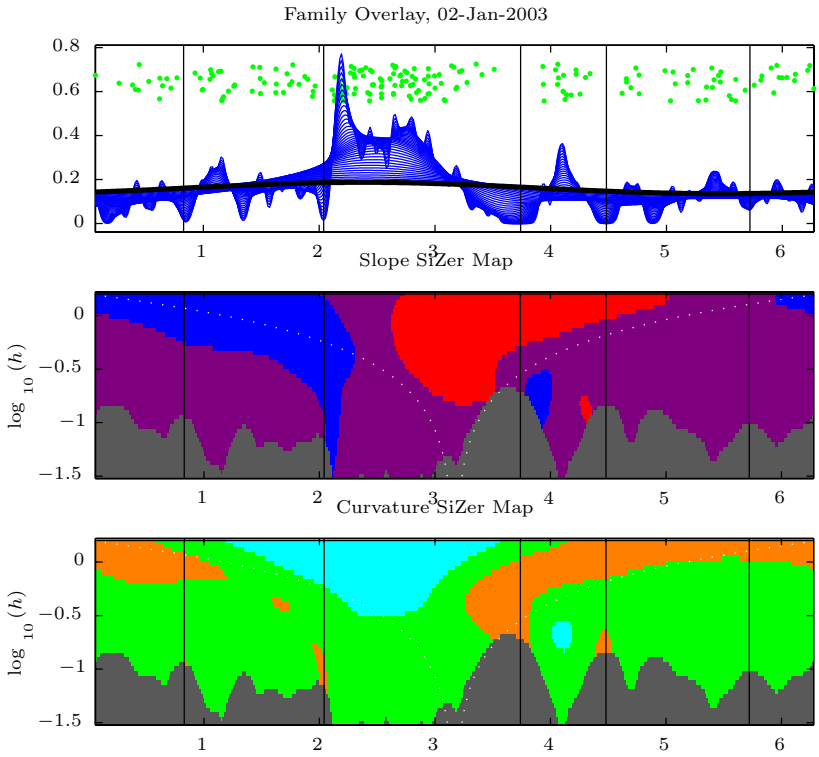

Figure 5. SiZer analysis of phases of top 200 most periodic genes, indicating clusters in the data that are statistically significant. The horizontal axis is phase (angles in radians), and the vertical axes are density (top), and bandwidth, i.e., scale or "level of resolution" (middle and bottom), Vertical lines are the boundaries that we found between the 5 cell phase classes.

The center and bottom panels show SiZer maps, as developed by Chaudhuri and Marron (1999, 2002). These are visualizations that provide useful statistical inference for the blue smoothed histograms. The main goal of the SiZer analysis is to understand which clusters in the data (i.e., bumps in the curves) represent important population structure, and which can be attributed to sampling variability. This is done in the middle panel by studying slopes of the blue curves. The SiZer map uses colors to indicate statistical significance of slopes, with blue for significantly increasing, red for significantly decreasing, and the intermediate color of purple when the slope is not significant. The fourth color of gray is used to indicate locations where the data are too sparse for statistical inference. Rows in the SiZer map correspond to blue curves in the top panel, indexed by the "window width" h. The bottom panel shows a parallel analysis, except the inference is based on curvature, i.e., the second derivative, instead of slope (first derivative) as in the middle panel. The curvature SiZer colors are

cyan (light blue) for significantly concave (i.e., curved downwards), orange for significantly convex (upwards), and the intermediate color of green for situations where there is no statistically significant curvature. The name SiZer comes from the underlying concept of SIgnificance of ZERo crossings.

The slope SiZer analysis in the middle panel shows that the very large cluster between phases $\theta = 2$ and $\theta = 3$ is statistically significant (blue on the left, red on the right). Similarly the smaller cluster near $\theta = 4.2$ is also statistically significant. The curvature SiZer analysis in the bottom panel also flags these clusters with cyan regions.

For finding boundaries between gene clusters, the goal of the SiZer analysis is to identify the valleys, not the peaks. Careful inspection of the SiZer map was used together with the earlier classification of Spellman et al. (1998) for this purpose, resulting in the vertical lines shown in Figure 5. Where possible, the vertical lines were drawn at orange spots, indicating a statistically significant valley, in particular at the phases $\theta = 0.83,\ 2.04,\ 3.74,\ 4.48$. The remaining cell cycle boundary was not highlighted by the SiZer map (probably because the data are quite sparse in this region), but do exist from the Spellman et al. (1999) classification. We chose this boundary to be the local minimizer of the blue curves in the top panel at the phase $\theta = 5.72$.

Figure 6 shows the sin-cos projection scatterplot of Figure 4, with the class boundaries added as purple rays from the origin. Also the data are colored according to membership in the 5 classes, with black used for the genes that are not periodic in our sense.

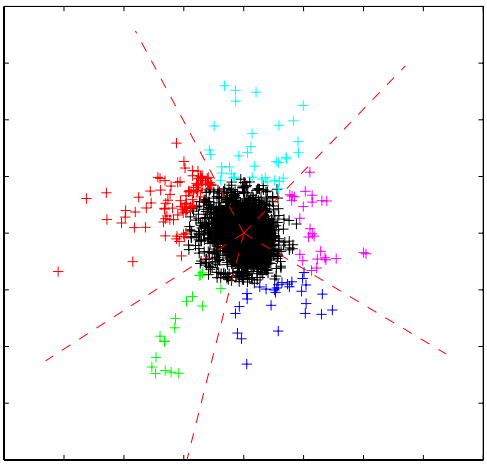

Figure 6. Projected scatterplot view of our gene classification, using the same data and axes as Figure 4. Black genes in the center are unclassified. Other colors represent the 5 classes. Suggests clear groupings of cell cycle phases.

Note that the boundary rays fit nicely into the visually apparent gaps between the colored clusters.

Another view of our top 200 gene classification is given in Figure 7. Again the x-axis is phase $\theta \in [0, 2\pi]$. The solid colored curves are smooth histograms representing the five classes. The dashed black curve is the sum of the colored curves, giving the smooth histogram representation of all 200 periodic genes. The window width was chosen by eye to maximize the visual separation between the classes, at $h = 0.266$.

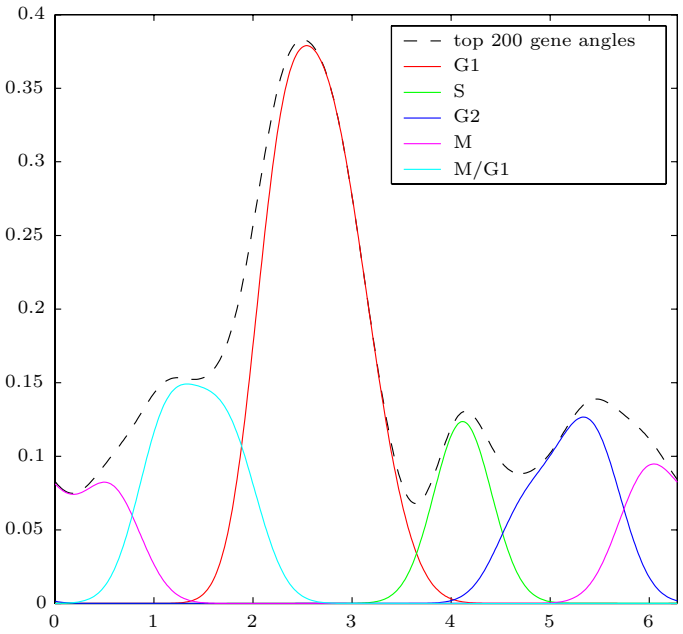

Figure 7. Smoothed histograms of populations of class phases. Phase (angle in radians) is on the horizontal, and density on the vertical, axes. Labels match class colors to named phases of cell cycle. Again suggests clear separation of periodic genes according to part of the cell cycle.

Figure 7 again shows that our classification is quite sensible, with generally good visual separation of the classes. The class boundary at $\theta = 2.04$ is a good example of why both the slope and curvature versions of SiZer are important. Because the red bump is so much larger than the cyan bump, there is no valley between them, so the slope version of SiZer (middle panel of Figure 5) cannot find this boundary. However, there is a change in the convexity, and this is statistically significant as shown in the curvature SiZer map (bottom panel). This same effect of relative sub-population size makes the boundaries of the purple sub-population especially hard to identify, because it has the smallest peak.

While use of the top 200 most periodic genes gave the best division into sub-populations, the number is substantially smaller than the 612 found by Spellman et al. (1998). Because there is no clear boundary between "periodic" and "aperiodic", we chose our top 612 genes as our "final classification". One view of this new classification is very similar to Figure 6, except that the outermost black plus signs now appear in the corresponding groups. This picture is not shown to save space, and because it is too similar to Figure 6.

Another view of the effectiveness of our classification is shown in Figure 8. Here the 612 classified genes are shown in the form of their original raw time series (the full series, not the projected version). These are overlaid as in Figure 2a, but the classification is now illustrated using the same colors as above. The 5 sub-populations are shown individually, and also all 5 are overlaid in the top left panel.

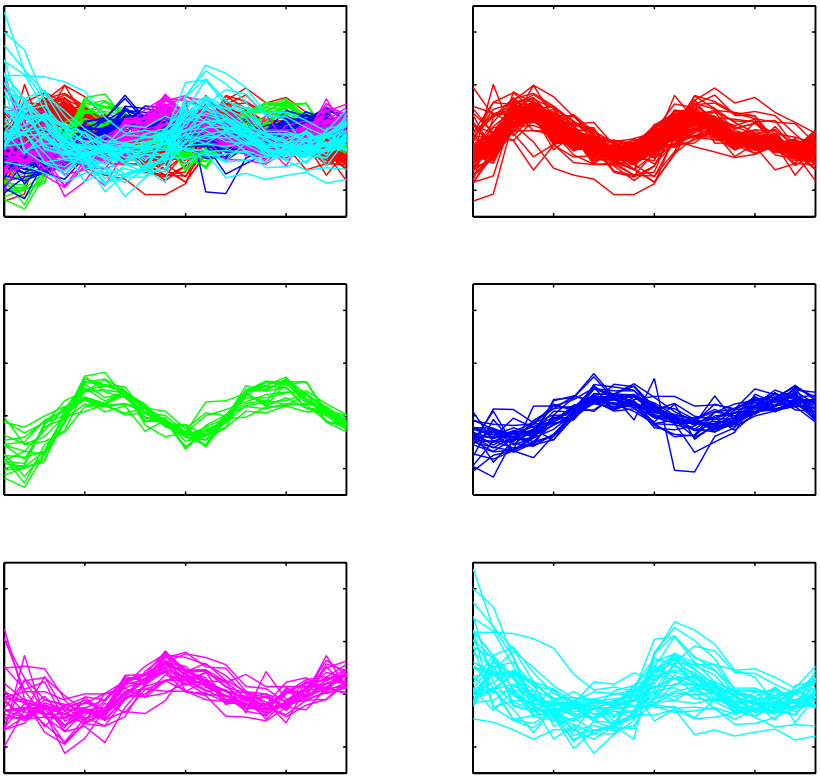

Figure 8. Raw time series view of the results of our classification. For each plot, time is on the horizontal, and relative gene expression level is on the vertical, axes. Show very clear grouping of genes into phase groups.

Figure 8 shows that our analysis did indeed produce 5 classes of genes, where all the members are quite periodic. The different phases of the sub-populations are also clearly apparent.

Heping Zhang suggested that we compare our analysis with that of Spellman et al. (1998), by constructing the same view for their classification. The result of this is shown in Figure 9.

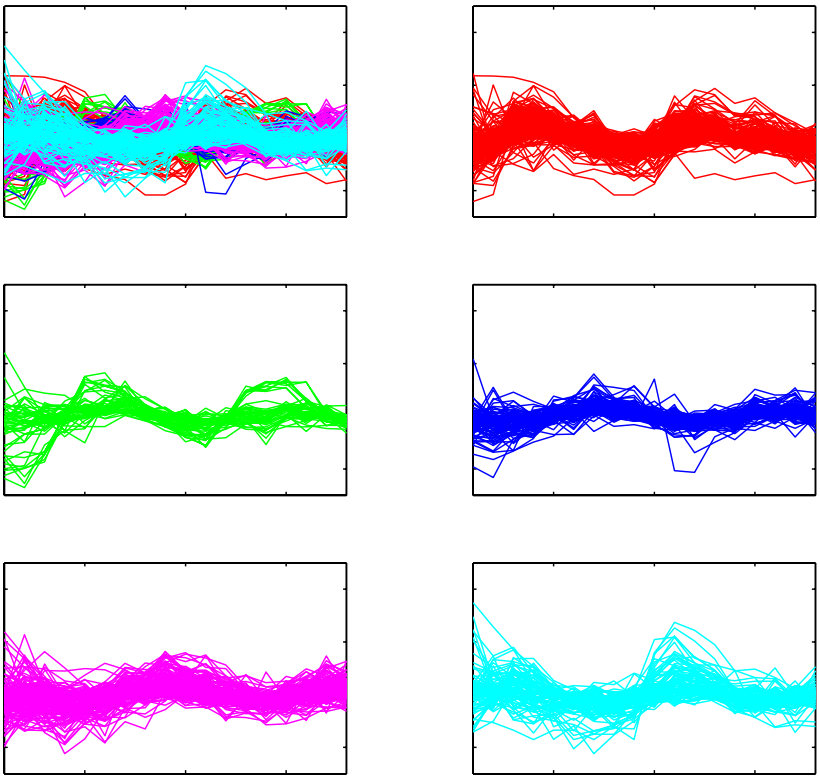

Figure 9. Raw time series view of the results of the Spellman et al. (1998) classification. Format is same as Figure 8. This gives a visual suggestion that our classification gave an improved identification of periodic genes.

The results in Figure 9 are generally similar to those in Figure 8, showing that in general the two analyses found similar sets of genes. However, there are a few differences, most notably in the green sub-populations, shown in the left center panels of Figures 8 and 9. While this is purely personal opinion (often dangerous in the absence of biological information) we suggest that our green

sub-population, in Figure 8, appears to be more periodic than that in Figure 9.

5. Conclusions and Open Problems

The main point of this paper was to introduce longitudinal data analysts to the powerful viewpoint of Functional Data Analysis. The ideas were illustrated in the context of an example from a gene expression study of the yeast cell cycle. In particular, thinking about a population of time series from the FDA viewpoint (in particular, using the framework of object space-feature space-point cloud view) leads to a natural analysis of the interesting periodic structure in the data. This resulted in some improvements over an earlier analysis of the same data. We offer this as motivation for longitudinal analysts to include FDA methods in their toolbox. See Ramsay and Silverman (1997) for more introduction to FDA ideas.

There are several interesting open problems and areas for future research.

Jane Ling Wang pointed out that the period of 2 may not be exactly correct. This is suggested for example by a careful look at the left panels of Figure 3b. Our analysis shows that assuming period 2 is correct gives reasonable answers, however some improvement is indeed possible. A simple way of doing this would be to experiment with periods near 2, and choose one to maximize the sum of squares of the sin-cos projections. More complex methods of fine-tuning the period include complex de-modulation, see e.g., Bloomfield (2000) or Brillinger (1981), or the fitting methods of Thomson (1995).

Spellman et al. (1998) took a different approach to this problem by actually fitting different frequencies to different time series.

Another challenge is choosing the threshold of which genes are periodic. We used a threshold of 200 to identify class boundaries, and 612 in the final analysis. Both choices were rather arbitrary, and a more careful study would probably be useful. Additional biological input seems to be needed for this.

Acknowledgements

The authors are grateful for the chance to present this paper in the stimulating environment of the 2002 AMS-IMS-SIAM Summer Research Conference on Emerging Issues in Longitudinal Data Analysis, and for the interesting comments made at that time, that have been used to improve the presentation. The second author is grateful to the Cornell University School of Operations Research and Industrial Engineering, for providing the stimulating research environment which resulted in this paper, and to Cornell University's College of Engineering Mary Upson Fund for financial support. The first and third authors gratefully acknowledge the support of NSF Grant DMS 99-71586.

References

Bloomfield, P. (2000). *Fourier Analysis of Time Series : An Introduction*. Wiley, New York.

Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. Holden-Day, San Francisco.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structure in curves. *J. Amer. Statist. Assoc.* **94**, 807-823.

Chaudhuri, P. and Marron, J. S. (2002). Curvature vs. slope inference for features in nonparametric curve estimates. Unpublished manuscript.

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

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

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National. Academy of the Sciences of the USA* **95**, 14863-14868.

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

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

Li, K.-C., Yan, M. and Yuan, S. (2002). A simple statistical model for depicting the CDC15-synchronized yeast cell-cycle regulated gene expression data. *Statist. Sinica* **12**, 141-158.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. and Cohen, K. L. (1999). Robust principal component analysis for functional data. *Test* **8**, 1-73.

Marron, J. S. and Todd, M. (2002). Distance weighted discrimination. Tech. Report, internet available at http://www.optimization-online.org/DB_HTML/2002/07/513.html.

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

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics* **14**, 1-17.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

Thomson, D. J. (1995). The seasons, global temperature and precession, *Science* **268**, 59-68.

Tukey, J. and Tukey, P. (1990). Strips displaying empirical distributions: textured dot strips. Bellcore Technical Memorandum.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Verlag, Berlin (Russian version, 1979).

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag, Berlin.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *J. Amer. Statist. Assoc.* **85**, 664-675.

Yushkevich, P., Pizer, S., Joshi, S. and Marron, J. S. (2001). Intuitive, localized analysis of shape variability. In *Information Processing in Medical Imaging*, (Edited by M. F. Insana and R. M. Leahy), 402-408.

Zhang, H., Yu, C. Y., Singer, B. and Xiong, M. (2001). Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of the Sciences* **98**, 6730-6735.

Zhang, H. and Yu, C. Y. (2002). Tree-based analysis of microarray data for classifying breast cancer. *Frontiers in Bioscience* **7**, 62-67.

Department of Statistical Sciences, Cornell University, Ithaca, New York 14853, U.S.A.

E-mail: xz45@comell.edu

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, U.S.A.

E-mail: marron@email.unc.edu

Department of Biological Statistics and Computational Biology, and Department of Social Statistics, Cornell University, Ithaca, New York 14853, U.S.A.

E-mail: mtwl@comell.edu

(Received January 2003; accepted January 2004)