Time Series Functional Data Analysis

Blinded
Blinded

June 7, 2004

Abstract

The functional data analysis viewpoint has proved useful in the analysis of a wide variety of modern data sets. Here this approach is applied in the context of a time series of curves, motivated by some chemometric data. A novel visualization is proposed that allows deep insights into the time structure. Principal Component projections reveal additional special structure in the data. Related tools are also used for the comparison of two related time series.

1 Introduction

The set of ideas called "functional data analysis", has provided some powerful data analytic techniques for a wide array of complex modern data sets. This point is well illustrated in Ramsay and Silverman (1997, 2002) in the case of curves as data, by Locantore, et al. (1999) in the case of images as data, and by Cootes, et al. (1993) and e.g. Yushkevich, et al. (2001) where the "data points" are shape representations of body parts.

In this paper, these ideas are extended to the case of a time series of curves. Motivation comes from a time series of chemical spectra, collected to try to understand the evolution of a complex chemical reaction over time.

In Section 2 a single series of spectra is studied. Simple views of the data reveal that care is needed to understand the variation. In Section 2.1, a novel visualization method, that makes heavy use of color, is proposed that allows simple and direct illustration of the time structure. This visualization is combined with Principal Component Analysis (PCA) to reveal interesting insights. Among the lessons learned from time colored PCA is that the projections follow systematic modes of variation over time. This striking behavior is explained by the construction of a simulated experiment in Section 2.2.

In Section 3, the methods are extended to comparison of two spectra, highlighting some important differences. These differences are explained by revisiting the chemical experiment.

Because we believe that these methods will be useful for a wide range of time series of curves, both within and beyond chemometrics, flexible and user-friendly

Matlab software for doing most of the analyses in this paper are internet available at [blinded]. The main function is curvdatSM.m, but it is recommended to download the full zip file, as well as the zip file at [blinded] because they contain a number of needed subroutines.

2 A Single Time Series of Spectra

The motivating data used in this study (courtesy of Darla G. Thompson) are the time-dependent infrared (IR) spectra of various samples of Estane 5703 ®. This polymeric material is part of the binder formulation ("glue") of a high explosive formulation. With heating, the polymer "melts" becoming a viscous fluid and the explosive is formulated and shaped. Upon cooling, this polymer then undergoes a self-organizing, phase separation process that results in its solidification. The reported IR spectra track the phenomena associated with this process for the pure polymer, where the polymer has been artificially aged by various means.

Ideally, the IR spectra would consist of a collection of Gaussian/Lorentzian absorption peaks on flat, zero background. In reality: 1) the peaks are skewed and do not have a precise mathematical form, 2) the peak widths and shapes may be changing during the process in addition to relative intensity changes, and 3) the background can be curved and that can change due to instrumental artifacts, especially during the protracted measurements shown here (collected over a one week timeframe).

The analysis of Schoonover, et al (2003) of this data set indicated that one major process dominated the spectral changes, but that at least one minor process was also occurring. Ideally, to quantitatively interpret these results, the spectra should be expressed as a sum of "real" spectra (positive absorption peaks) whose relative intensities are a function of time. This work describes our efforts to develop methods for this in a manner that requires minimal input (and possible bias) from the analyst.

This process results in a time series (over the week) of curves (each is a single IR spectrum). The series studied in this section represents a sample that was aged in humid air for 27 days. The time points are roughly (but not exactly) logarithmically spaced. This time scale is appropriate to first order, since the chemical reactions exhibit exponential decay.

A data set of this type is shown in Figure 1, with $n = 77$ spectral curves of size (i.e. a vector of dimension) $d = 1556$. The left panel shows an overlay of the raw data curves. The horizontal axes in both panels of Figure 1 are the same, and represent integer "frequency numbers" representing the 1556 channels of the IR spectral measurement.

It is impossible to see that there are $n = 77$ curves present, because they are so similar. This is because the variation in the curves is of a much smaller order than the common structure of the curves. This visual impression is quantified by studying sums of squares (of the entries in the data vectors), where it is seen that the sum of squares for the mean is 99.97% of the total sum of squares.

Thus only 0.03% of the "signal power" in the data is variation about the mean, i.e. represents differences in these curves.

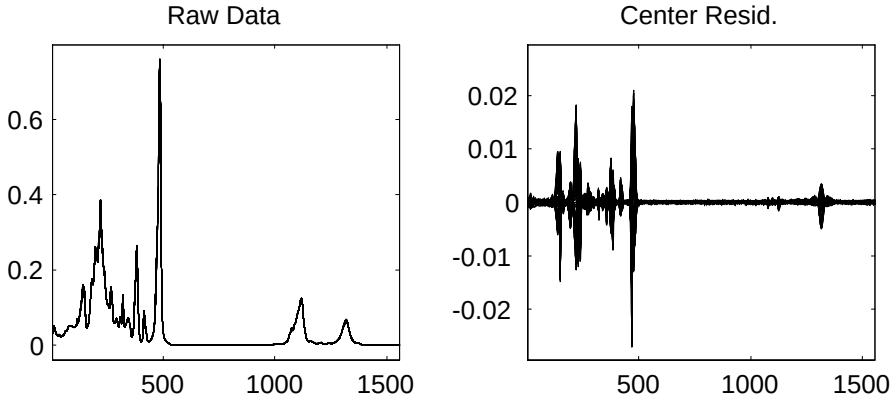

FIGURE 1: *Time series of 77 spectra, raw on left, residuals from the mean on right, allowing better visualization of the variation between the curves.*

A simple visual device that gives better separation between these curves, and thus allows exploration of the chemical reaction of interest, is to study the mean residuals, i.e. take each data vector, and subtract the mean vector. These centered data are shown in the right panel of Figure 1. Now it is apparent that the differences between these curves is in terms of rather few, relatively thin, spikes.

A useful conceptual model for understanding this type of data analysis is to think of each curve as a vector, and each vector as a point in the 1556 dimensional Euclidean space $\mathbb{R}^{1556}$. The collection of curves thus generates a point cloud in $\mathbb{R}^{1556}$, and the analytic ideas can be easily understood as reflecting features of this point cloud. For example the mean residuals in the right panel represents the original point cloud, but shifted so that the mean is now at the origin. This device will also be used to interpret PCA below.

Note that in the left panel of Figure 1 the vertical scale is much smaller, again reflecting the relatively small size of this variation. On this scale, there are some clear differences, although they seem to show up mostly at a fairly narrow range of frequencies. Note also that the exact shape, and also time structure are not easy to see.

A major problem with the mean residuals displayed on the right in Figure 1 is that not much structure can be seen. One reason is that the $d = 1556$ frequencies are perhaps too many for display, especially when a large number of them do not show interesting features of the data. Figure 2 addresses this issue by "zooming in" on an interesting set of frequencies, the range 300-500. This range is small enough that the spectra show up as smooth curves. The raw spectra on the right still appear as nearly one curve, although the "thickenings" in several areas suggest some variation. In the Mean Residual plot on the right, it is now clear that there are in fact a number of curves overlaid. Note that the larger variations seen correspond well with the "thick parts" in the left panel.

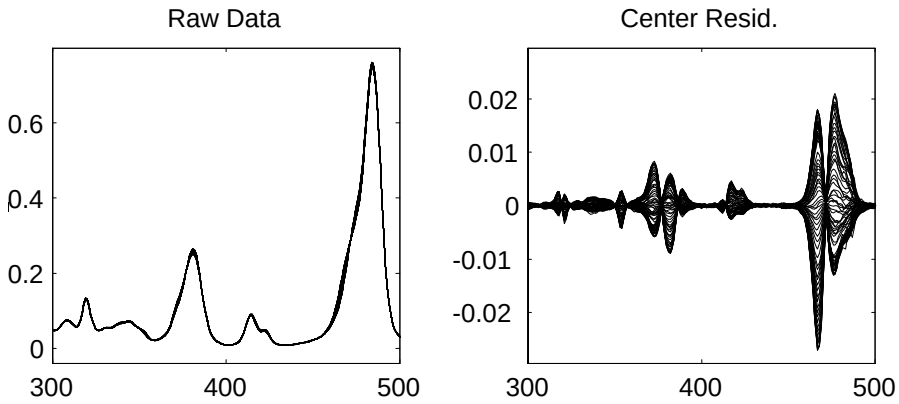

FIGURE 2: *Time series of 77 spectra, shown over reduced range of frequencies, showing smoothi curve structure*

While the display in the right panel of Figure 2 is far more useful than that of Figure 1, it still has some substantial drawbacks. First, it does not show the time ordering. Any variation in the spectra reflected by PC1 that are not ordered by time are likely unimportant, while those that are systematic with respect to time are of keen physical interest. The second major drawback of the display on the right side of Figure 2 is that it is impossible to tell which of the peaks rise and fall together. E.g. how are the peaks around frequencies 370 and 380 associated with those around 470 and 480?

A graphical device which solves these problems is presented in Figure 3.

2.1 Time coloring

The main idea of this section is to apply colors to figures of the type shown above to convey time information. Here a color scheme is used where time is represented as a "spectrum of colors", starting with magenta, moving through blue, green, yellow, and on to red. A standard red-blue-green scheme is used, equally spaced with respect to the logarithm of the time.

The top two panels of Figure 3 are the same as Figure 2, but now colored according to time.

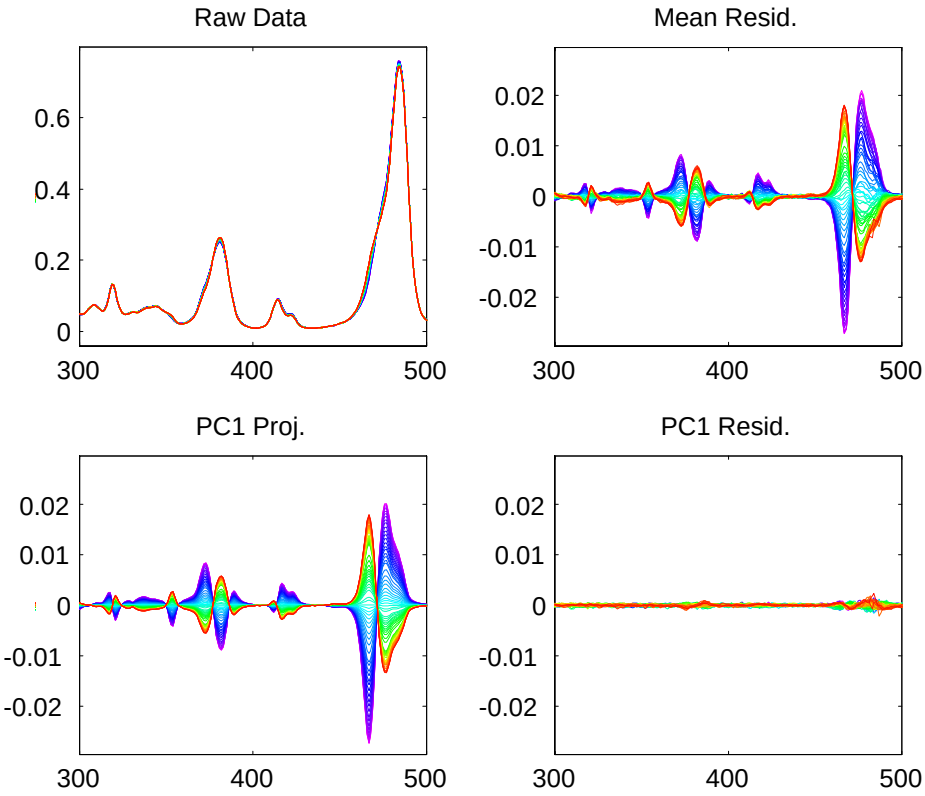

FIGURE 3: *Time series of 77 spectra, colorized to show time ordering (magenta - red). Bottom panel explores variation with PCA.*

In the top left panel of Figure 3, it is now easy to see that there are several curves overlaid. Furthermore, it is apparent which parts of the spectrum are largest at the beginning (magenta), and which parts are larger at the end (red). The coloring gives even more useful information in the Mean Residual plots shown in the upper right panel. Now it is clear that the change is very systematic and smooth with respect to time at all points in the spectrum. Essentially everything represents important chemical change. Also note that peaks at 470 and 480 suggest a combined shifting to the left of a peak, while the smaller peaks at 370 and 380 show a shift in the opposite direction.

Deeper insights into the data come from PCA. This is understood from the viewpoint of the data as a point cloud in $\mathbb{R}^{1556}$, as finding the direction of greatest variation. In particular, when the data are projected onto any one dimensional direction vector, the spread of the projections can be conveniently measured by the variance. PCA seeks the direction with maximal variance. This can be found via an eigenanalysis of the covariance matrix, with the eigenvector having the largest eigenvalue pointing in the direction of maximal variation. The bottom left panel of Figure 3 shows an overlay of the data projections multiplied by the eigenvector. The bottom right panel shows the

residuals from subtracting the curves in the lower left, from those in the upper right. In terms of the point cloud, this is the result of subtracting the projection onto the eigenvector from each point in the cloud, which is equivalent to projecting the data onto the subspace orthogonal to the eigenvector. This looks visually much smaller than the curves in the PC1 direction, suggesting that the first eigendirection explains a large amount of the variation in the population. The Sum of Squares of the PC1 projections of the data, shown in the lower left panel, is 99% of the mean residual Sum of Squares, shown in the upper right panel, leaving only 1% for the residuals in the right panel.

Another useful view, in a variety of functional data analytic contexts, is to study the relationship between the projection coefficients, as done in Figure 4.

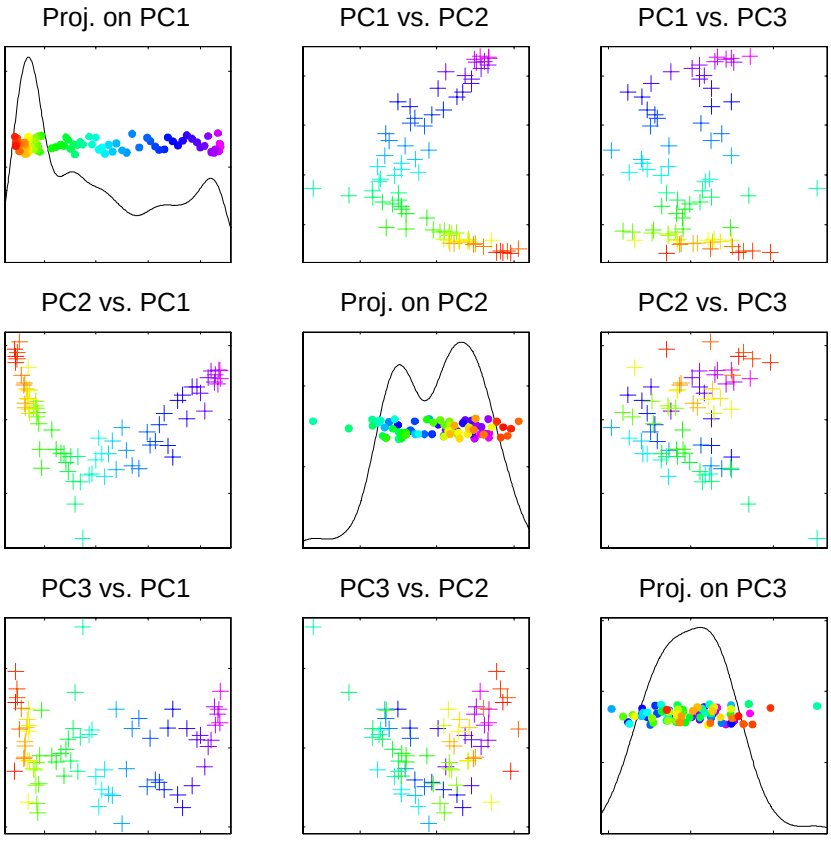

FIGURE 4: *Matrix of scatterplots of PCA projection coefficients. Univariate on the diagonal, bivariate off.*

The top left panel shows the PC1 projection coefficient (i.e. relative location of the projection along the eigenvector) of each curve as a colored dot, using the same coloring scheme as above. Good visual separation comes from using

a random vertical height (the jitter plot idea of Tukey and Tukey 1990). The coloring reveals these PC1 coefficients are nearly exactly ordered according to time, showing that the PC1 direction is clearly driven by the underlying chemical reaction. The black curve is a smooth histogram (more precisely a kernel density estimate, see Wand and Jones 1995), which shows that more of the points cluster at the left end of the eigenvector (as the chemical process stabilizes, recalling that green, yellow and red represent later time points), and they are relatively sparse elsewhere.

The central panel in Figure 4 shows the corresponding projection coefficients in the PC2 direction. Since this eigenvector is orthogonal to the PC2 eigenvector, it is not surprising that a different pattern appears in the colors.

The PC2 pattern is seen to be also systematic via the scatterplot shown in the top center panel, where the PC1 coefficients appear on the vertical axis, with the corresponding PC2 coefficients on the horizontal axis. The transpose of this plot is shown in the center left panel. These scatterplots reveal that as PC1 sweeps from one end to the other, PC2 goes from one extreme, to the other and back (basically PC2 goes through a full cycle compared to the half cycle of PC1). Careful inspection of the center left panel shows that this pattern is discernible in the dot colors. A partial physical explanation of this phenomenon is given in Section 2.2.

The lower left panel of Figure 4 shows the PC3 projection coefficients, and does not reveal any apparent patterns. The off diagonal plots seem to suggest that the PC3 patterns are perhaps driven by noise artifacts.

A different data set, where PC3 captures important structure is studied in Section 2.2.

2.2 Model for Curves in PCA Projections

In this section, the systematic relation between PC projection coefficients is investigated more deeply. The effect shows up even more strongly for a different experiment that is treated here. In this experiment, spectra were collected for some unaged estane, which may be thought of as a control. Figure 5 is the analog of Figure 4 for this experiment, with one dimensional projections shown on the diagonal, and corresponding two dimensional projections shown as scatterplots off of the diagonals.

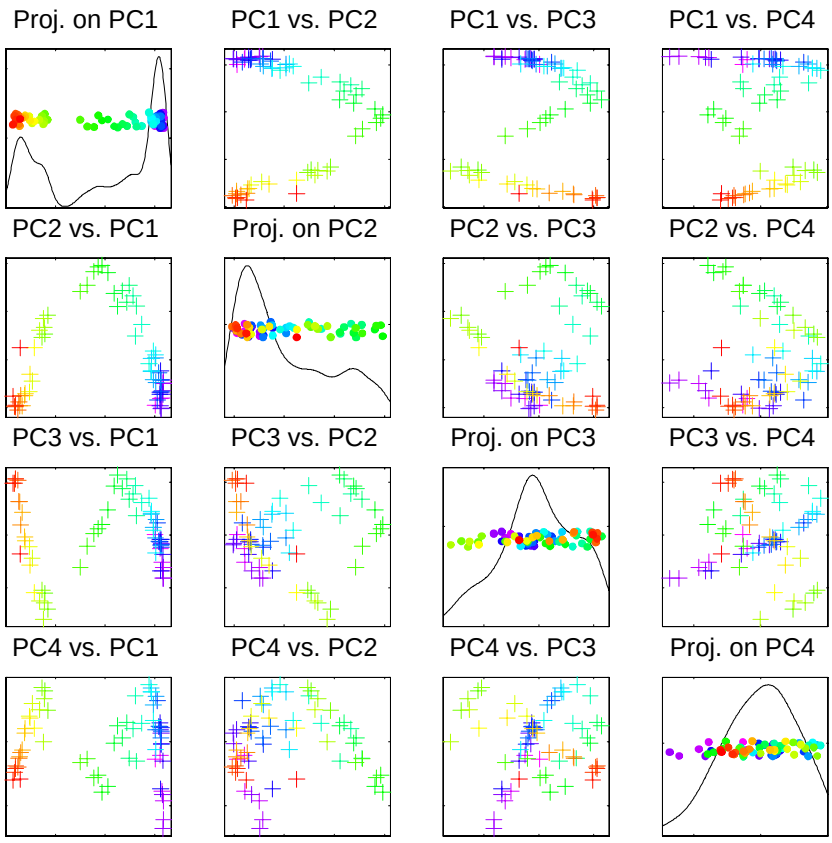

FIGURE 5: *PCA projection coefficient matrix of scatterplots, for a different experiment, showing higher dimensional structure, not obscured by noise.*

The first 2 PCs in Figure 5 show a structure very similar to that seen in Figure 4. The first PC sweeps from one side to the other, although this time there is clustering at both the beginning and end. The second PC starts at one end, moves to other side, then returns. The direction taken by PC2 is the opposite from that in Figure 4, because of the arbitrariness of the sign of the eigenvectors. Again, time ordering is best kept straight using the rainbow color order: magenta, blue, green, yellow, red.

PC3 shows a more interesting structure in Figure 5 than was apparent in Figure 4, where any additional structure appears to be drowned out in the background noise. In the PC1 vs. PC3 scatterplot, it is apparent that the "2 cycle" behavior of PC2 has now become "3 cycle" behavior in PC3. The PC2 vs. PC3 structure is also interesting, starting with magenta at the lower left, sweeping up and to the right, reflecting off the left side, and moving towards the center at the top, and then going back down in a symmetric fashion.

The interesting structure in PC3 motivates exploring more eigendirections, so PC4 is also included in Figure 5. PC4 seems to extend the above ideas in

a very consistent manner, representing "4 cycle" structure. For example, in PC2 vs. PC4, the data follow the shape of a backwards letter C, first starting (magenta) at the lower left corner, arriving at the other tip of the C half way through (green), and then returning back along the C to the beginning point (red at the end).

When the data in Figure 5 are considered from the "point cloud" point of view, there is a strong suggestion that the data are very close to following a curve (i.e. part of a one dimensional manifold) in the Euclidean space $\mathbb{R}^{1556}$. This curve "twists" in a way that is fairly complex. In particular, it is not contained in either a 2 or a 3 dimensional hyperplane.

The special structures exhibited in Figures 4 and 5 bring up questions such as: what type of physical phenomena can generate such behavior?

This question is now explored by construction of a simple simulation experiment that generates such behavior. For simplicity, we consider a toy example of just 5 chemical species (the mass of each of which is viewed as giving the height at a single frequency in the corresponding 5 frequency spectrum). We run a simple discrete time model, starting with equal amounts of species 1, 2, 3 and 4 and none of species 5. At each time step the species change at the rates shown in this table:

Before	After	Rate
1	2	0.0025
2	3	0.0025
3	4	0.0125
4	5	0.0125

Thus, at each step, 0.01 of the amount of Species 1 is turned into Species 2, etc. This process was run for 41 steps, so the data set has $n = 41$ toy spectral curves of dimension $d = 5$.

The analog of Figure 5 for these curves is shown in Figure 6. An animated version of a very similar process, showing the time evolution of the species, and how they relate to the display in Figure 6 is available from the Internet at [blinded].

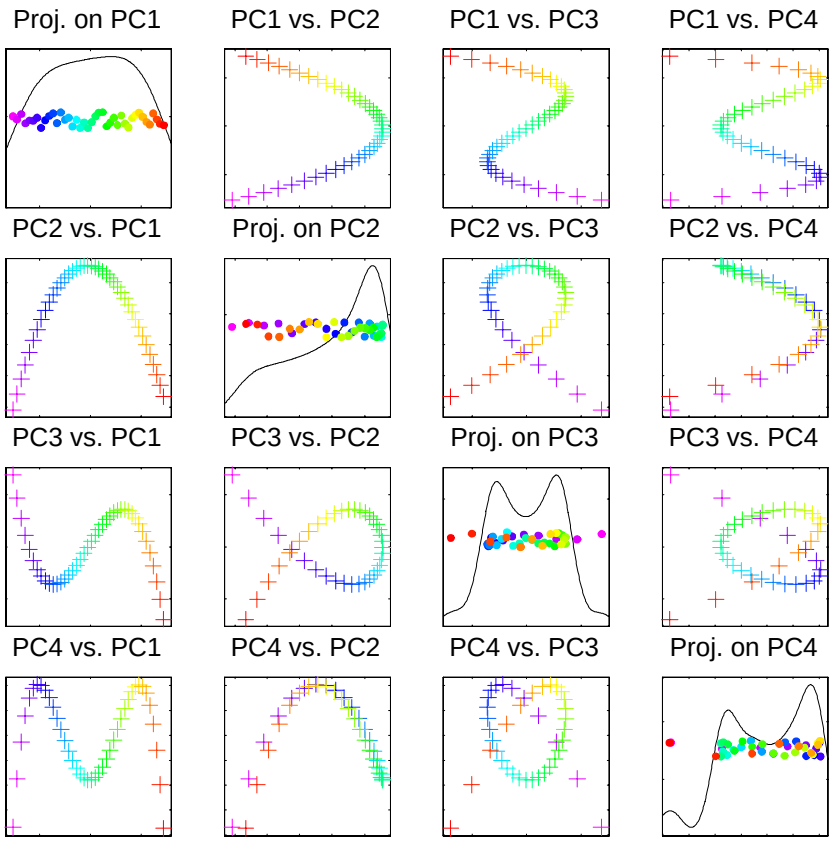

FIGURE 6: *PCA projection coefficient matrix of scatterplots, for a simulated toy example, illustrating a possible mechanism behind the structures seen in Figures 4 and 5.*

The general structure of the data in the scatterplots is similar to Figure 5. In particular, PC1 sweeps through the given range once, PC2 cycles twice, etc. The PC1 vs. others scatterplots are very similar. The backwards C shape noticed in PC2 vs. PC4 is now very clear. This process is one way of generating a family curves that is very reminiscent of those in Figure 5. The traces provide a visual representation of the underlying component behavior.

Some experimentation suggests that at least four chemical reactions are needed to have interesting structure in all four principal components. This makes sense, because with fewer reactions, the resulting spectra will lie in a lower dimensional subspace. An interesting mathematical problem is whether this number is indeed necessary and sufficient for this type of behavior.

A very important open problem is the inverse problem of finding the underlying chemical process from such a trace. It is not clear that these are identifiable, i.e. that there is a unique solution. Work is under way on these

issues, perhaps to be addressed in a future paper.

3 Comparison of Time Series

In this section, we use data analytic methods similar to those developed above to address the different issue of how two time series of curves compare. In addition to the data analyzed in Figures 1 - 4 above, we include an additional related data set. This time the sample was aged in dry air for 59 days. Inspection of the full range of frequencies, as in Figure 1, showed that there is interesting behavior well beyond the frequencies 300-500, which were the focus of Section 2. This motivated the frequency range 300-850, which is used everywhere in this section. Again, the data are analyzed using all frequencies, and only the final graphics are restricted.

The extended range graphic for the first data set is shown in Figure 7.

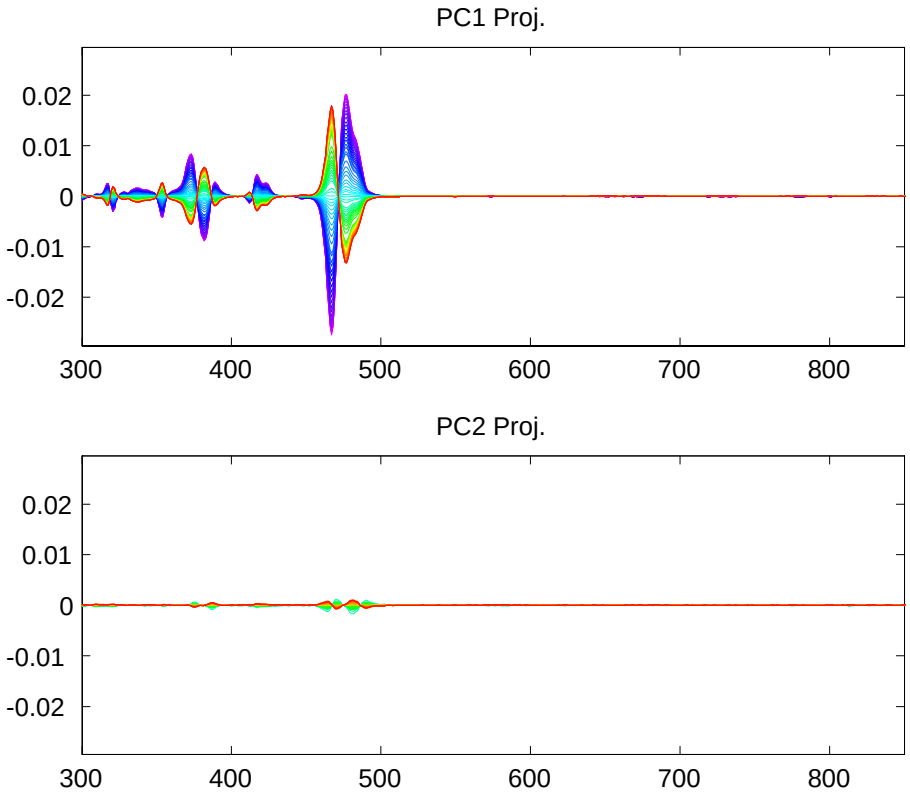

FIGURE 7: *First two principal components, with extended range, for data of Figures 1-5.*

The top panel of Figure 7 is essentially the same as the bottom left panel of Figure 3, except that the horizontal range is much larger. It is clear that for

the analysis of this data set alone, the range of 300-500 shows the main lessons very well. The bottom panel of Figure 7 shows PC2, which as expected from the discussion in Section 2, is very small.

Figure 8 shows the corresponding view of the new data set.

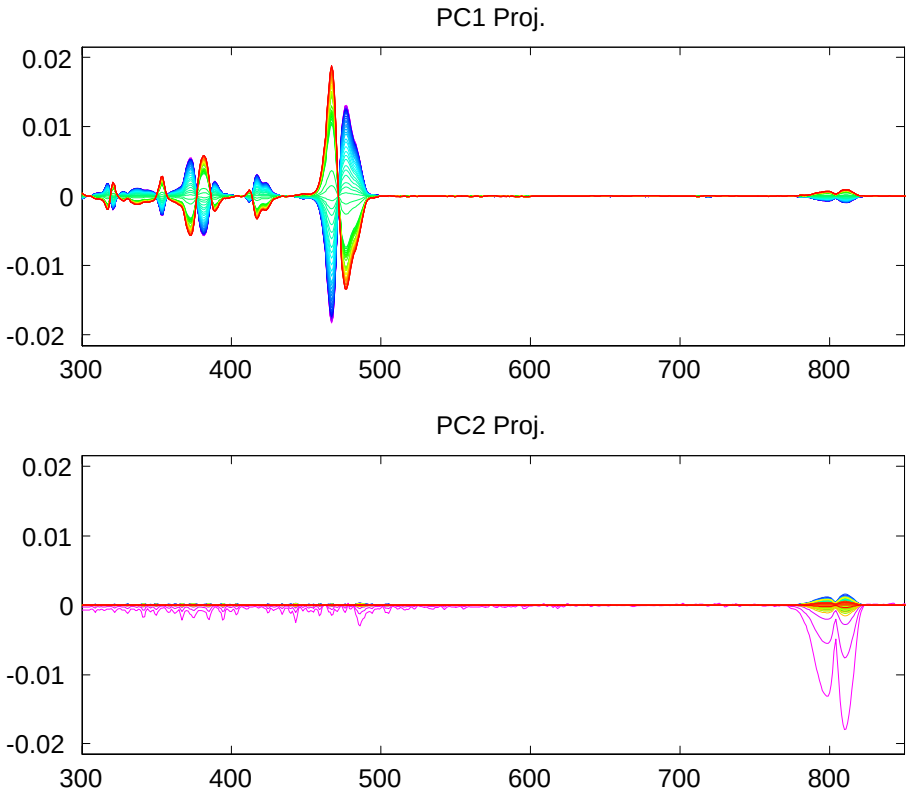

FIGURE 8: *First two principal components, with extended range, for a new data set.*

The PC1 projection in the top panel is quite similar to that for PC1 in Figure 7, not surprising because the chemical process is similar in nature. However, some differences are apparent, including the different relative heights of the peaks at frequencies 470 and 480. In particular, in Figure 7 the magenta peak at 490 is taller, indicating more of this material at the beginning, while in Figure 8, the red peak at 480 is taller, indicating more of this at the end.

The PC2 projection in the bottom panel is very different, showing very few large magenta (indicating this happens early in the process) differences around frequency 800. Investigation revealed that this was due to a rapid purging of CO_2 (which has this far different spectral representation) early in the experiment.

Another view of this is given in Figure 9, using the display device of a scatterplot matrix of projection coefficients, as in Figures 4, 5 and 6.

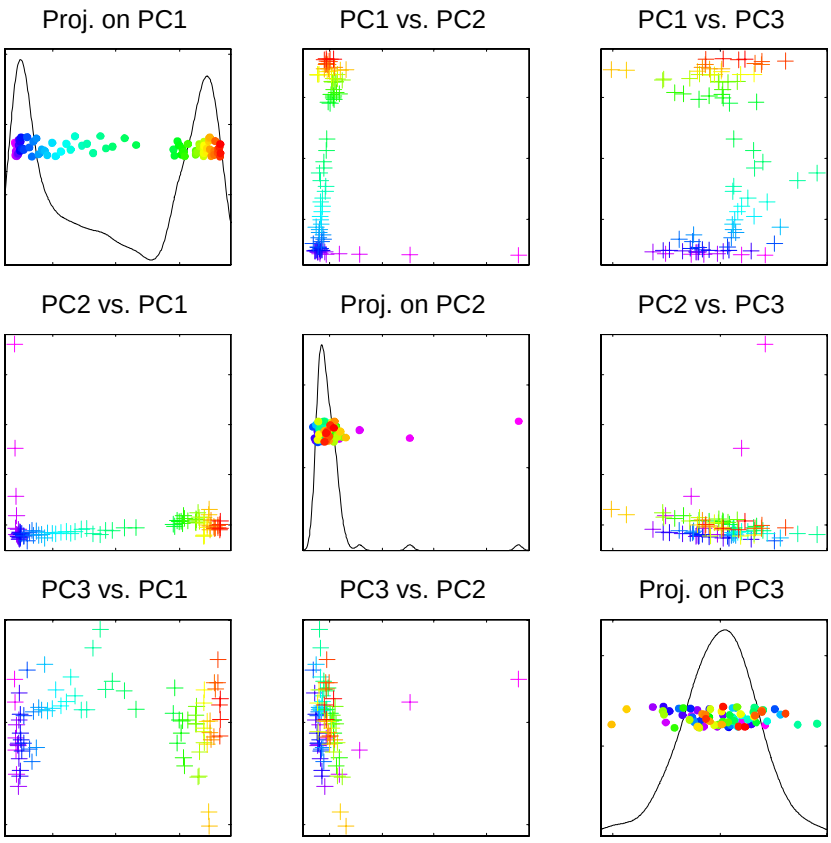

FIGURE 9: *PCA projection coefficient matrix of scatterplots, for the new data in Figure 8. Shows clear effect of two different components of the chemical reaction.*

The univariate PC1 projections in the upper right are quite similar to those observed above. The projections on PC2, shown in the center panel of Figure 9, provide another way of seeing that it is just the first few (magenta) data curves that have this structure (because the CO_2 purging is quite rapid). The PC1 vs. PC2 scatterplot in the top center shows that the directions of these two effects are nearly orthogonal. This is because the CO_2 appears at completely different frequencies from the other components of the chemical process.

The remaining plots, illustrating PC3, show that beyond these two driving components, additional structure is obscured by noise.

Additional insights come from using these same PCA tools on the combined data set, where the two families of curves are combined into a single large family. This is shown in Figures 10 and 11, where the time based color scheme is now replaced by colors that indicate the data set, blue for the first, and green for the second.

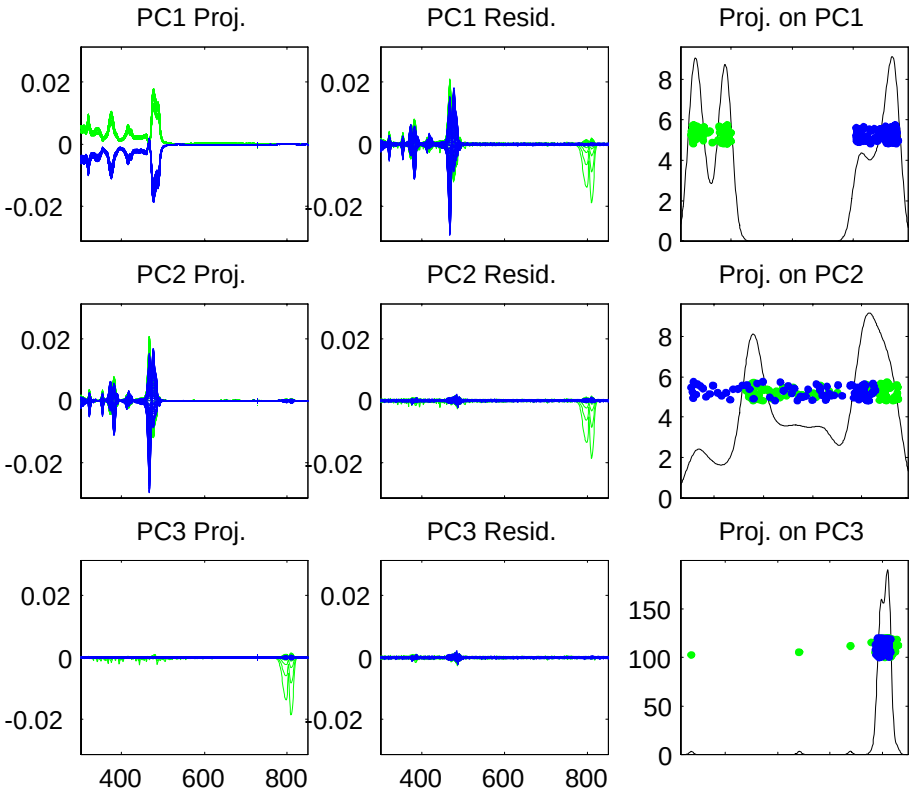

FIGURE 10: *PCA to compare two different time series of spectra.*

Figure 10 shows the decomposition of the combined population in terms of eigenvectors (i.e. PC directions, as in Figures 3, 7 and 8).

The PC1 direction, shown in the upper right of Figure 10, shows a distinct separation (in the frequencies where the spectra are nonzero) between the classes. The residuals, in the top center, show that this effect is effectively removed by subtracting PC1. The upper right is the same type of one dimensional view of the PC1 projection coefficients that is shown on the diagonals in Figures 4, 5, 6 and 9, except that now data set colors are used. This provides another way of seeing that in this direction the two data sets are widely separated. While the difference is dramatic, it turns out not to be of chemical importance, and instead is driven primarily by different scaling and normalization methods used in the two experiments.

The center row of Figure 10 shows the corresponding view in the PC2 direction. The direction, as shown in the center left panel, is quite similar to the PC1 direction shown in the top panels of Figures 7 and 8. The PC2 residuals, shown in the center panel, indicate that almost all of the visible structure is explained by the first two principal components, with the exception of the green structure, caused by an early purging of CO_2 at frequencies around 800.

The corresponding PC3 view is shown in the bottom row of Figure 10. Note

that this direction (bottom left panel) is driven by the CO_2 purging, and the fact that this happens for only a few time points is quite apparent in the plot of the projection coefficients in the bottom right panel.

The corresponding scatterplot matrix of the projection coefficients (same format as Figures 4, 5, 6 and 9), gives additional insight. The diagonal plots are the same as on the left of Figure 10.

The plot of PC1 vs. PC2 in the top center panel suggests that both chemical processes are of a very similar nature, since the lines representing time evolution are essentially parallel. The directions of the population difference (PC1), and the direction of the chemical reaction (the lines of symbols) are not quite orthogonal (since the lines are not horizontal), but this is approximately true.

The remaining scatterplots show that the direction of CO_2 purging (which drives PC3) is essentially orthogonal to both the directions of the data set difference (PC1), and the direction of chemical interest (the lines of points).

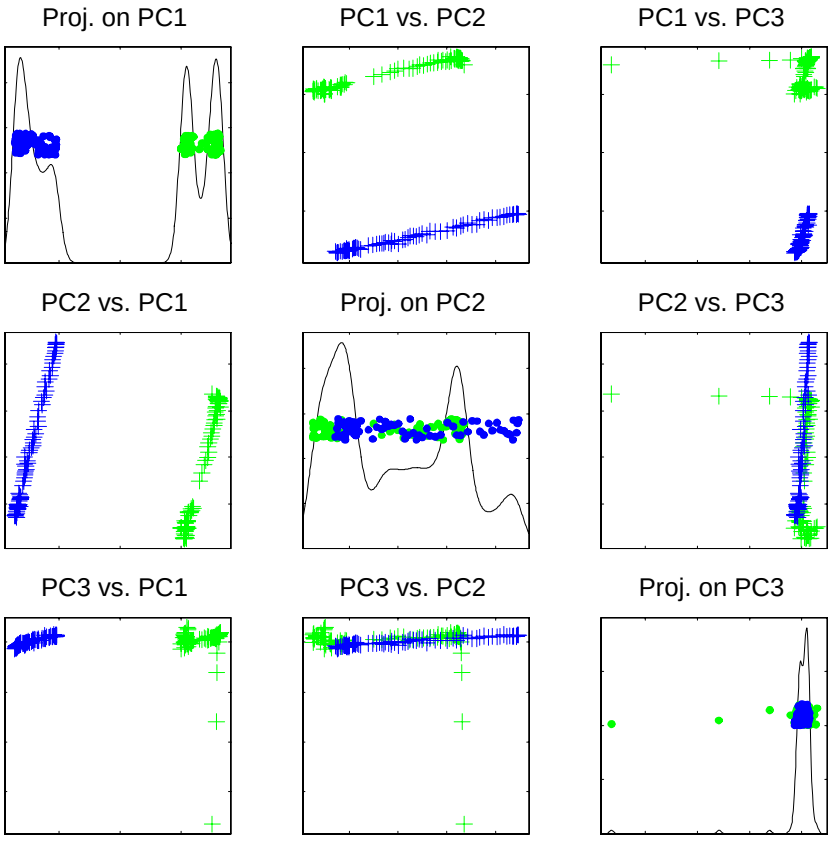

FIGURE 11: *PCA projection coefficient matrix of scatterplots, for comparison of two time series of spectra.*

4 Acknowledgement

The examples used here are based on the work of many people at Los Alamos National Laboratory. We are especially grateful to Jon Schoonover of the Materials Technology / Polymer Coatings Group, and to Darla Graff Thompson of the Materials Dynamics Group, for allowing us to investigate this data set. We also thank David Scott, Rice University, and Bonnie Ray, IBM, for insights gained during previous exploration of this data.

References

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

[2] 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.

[3] [Blinded]

[4] [Blinded]

[5] [Blinded]

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

[7] Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York.

[8] Schoonover, J. R., Marx, R. and Zhang, S. L. (2003) Multivariate Curve Resolution in the Analysis of Vibrational Spectroscopy Data Files, *Applied Spectroscopy*, 57, 483-490.

[9] Tukey, J., and Tukey, P. (1990). Strips Displaying Empirical Distributions: Textured Dot Strips. Bellcore Technical Memorandum.

[10] Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*, Chapman and Hall, New York.

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