

Graphical Techniques for the Exploration of Functional Data

E. Neely Atkinson
Department of Biomathematics
University of Texas M. D. Anderson Cancer Center
1515 Holcombe Blvd.
Houston, TX 77096

`eatkinso@mdanderson.org`

Abstract

This paper describes several graphical techniques for the exploration of functional data. The functional data may be displayed or transformed in several ways and may be linked to a number of univariate and multivariate plots of covariates of interest. The methods are implemented in LISP-STAT.

1 Introduction

Functional data are data which may be considered to have been sampled from some underlying smooth process. In principal, the data could be sampled at as fine an interval as desired, yielding smooth curves. Techniques for analyzing such data are described in [1]. This paper presents several techniques for the graphical explorations of functional data. These explorations help locate features of the data which may be worthy of further study and suggest models for formal estimation and hypothesis testing. Interactive graphical explorations can be of great value in promoting collaborations between data analysts and colleagues who are expert in the subject matter under study by providing non-statisticians ways to examine data directly.

2 Data

The data used in this paper to illustrate the techniques are drawn from an ongoing study of the use of optical methods to diagnose cervical abnormalities. During a gynecological exam, an optical probe is placed on the surface of the cervix. A brief pulse of light at a fixed wavelength (the excitation wavelength) is produced. The tissue of the cervix fluoresces in responses to this excitation. The intensity of the fluorescence is measured

at approximately 60 wavelengths (the emission wavelengths). The device which produced the data used in this paper produced data from three excitation wavelengths. For a given excitation wavelength, a plot of emission wavelength versus intensity produces a smooth curve or spectrum. The goal of the project is to develop algorithms which are able to distinguish between normal and abnormal tissue based on the characteristics of these curves. More information on this ongoing study can be found at <http://www.mdanderson.org/depts/bio/>.

The particular data examined here are drawn from 199 measurements of women with normal Pap smears; since multiple measurements were taken on some women, this sample represents a total of 55 individual women. Covariates available for analysis are current smoker (0=N, 1=Y), premenstrual (0=N, 1=Y), tissue type (1=columnar, 2=squamous, 3=transition zone), and age in years. Although the optical measurements show consistent changes between normal and abnormal tissue in individual patients, there is wide variation in the intensity of the measurements between patients, even for normal tissue. We hope that adjustments for covariates will remove a large part of this variability.

Since these data are still under collection and editing, the data used here were formed by taking random linear combinations of cases with similar covariate values. This produces a pseudo-dataset which sufficiently resembles the original data to illustrate the techniques being presented.

3 The Program

The methods demonstrated in this paper are coded in the LISP-STAT language, a free statistical computing package available for most computer systems. For more information, see [2]. This paper assumes that the user has LISP-STAT installed and has some familiar-

ity with its operation. LISP-STAT can be obtained at <http://www.stat.umn.edu/ARCHIVES/archives.html>.

The demonstrations given in this paper are necessarily static; to fully exploit the dynamic features of the methods, the program must be run. The code and data are included in the Proceedings CD-ROM. To build and execute the program, load the file `run.lsp`.

When the program has loaded, it will have added two menus to the system: **Data** and **Analysis**.

The **Data** menu contains the following commands.

1. **Read Spectra** - reads in the functional data. The program refers to frequency and intensity, but these can be any sets of x and y values. The intensities and frequencies are in separate files. Each line of the input files gives the values for one curve. The spectra are named as they are read in so they can be referenced in analyses. Note that the frequencies may vary for each spectra, although they do not in the sample data.
2. **Read Covariate** - reads in covariate values. There is an options to jitter the input values; this is useful when plotting categorical data.
3. **Delete Spectra** - removes the selected spectra from the data sets known to the program. This does not effect the disk files from which the data were read.
4. **Delete Covariates** - removes the selected variables from the data known to the program. This does not effect the disk file from which the data was read.
5. **Fetch Covariate** - brings a variable from LISP-STAT into the data known to the program.
6. **Store Covariate** - stores a variable from the program into LISP-STAT.

Since the program runs within the LISP-STAT environment, all of the features of LISP-STAT are available. The **Fetch** and **Store** commands facilitate moving data between the program and LISP-STAT.

The **Analysis** menu contains the following commands.

1. These commands display the spectra.
 - (a) **Data** - displays the raw spectral data; optionally, the mean curve is displayed in red.
 - (b) **Derivative** - displays the derivatives of the spectral data

- (c) **Parametric Curve** - displays a parametric curve representing the relation between the spectra measured at two different excitation wavelengths. Let the value of the spectrum measured for patient i at excitation wavelength 1 and emission wavelength w be $f_{1i}(w)$; let the spectrum measured measured at wavelength 2 be $f_{2i}(w)$. Then we plot the parametric curve given by $(f_{1i}(w), f_{2i}(w))$ for $w \in (0, 1)$. For this plot, the emission wavelengths for each excitation wavelength are arbitrarily scaled to run from 0 to 1.

2. These commands display the covariates.
 - (a) **Histogram** - plots a histogram of the selected covariate.
 - (b) **Scatterplot** - plots a scatterplot of the selected variables.
 - (c) **Scatterplot Matrix** - plots a scatterplot matrix of the selected variables.
 - (d) **Spinning Plot** - produces a spinning 3-D plot of the selected variables.
3. These commands modify the spectra. By aligning the spectra by intensity and location, we are able to focus on aspects of the shapes which would not otherwise be apparent. We are also to extract certain features of the curves and save them as scalars, so that standard statistical techniques can be applied.
 - (a) **Register Inten by Area** - divides the intensity of each curve by the area of that curve. The area of each curve is saved as a covariate.
 - (b) **Register Inten by Max** - divides the intensity of each curve by the maximum intensity of that curve. The maximum intensity is saved as a covariate.
 - (c) **Register Freq to Max** - uses a piecewise linear transformation on the frequency axis to align the peaks of all the spectra.
4. These commands examine how the relation between the spectra and a covariate changes with the emission frequency.
 - (a) **Inten-Covar Scatterplot** - produces a scatterplot of intensity versus the value of a selected covariate for a given emission wavelength. The emission wavelength can be varied using a slider, permitting exploration of which portions of the spectra are most strongly affected by the covariate.

- (b) **Build Inten-Covar Regression** - performs a regression of intensity on selected covariates for a range of emission frequencies.
 - (c) **Plot Inten-Covar Regression** - plots the results of the above regression. This command produces plots of the regression coefficients for each emission frequency with 95% confidence intervals; it will also produce plots of the P-value of each regression and the r^2 . The P-values are, of course, mostly gibberish because of the repeated hypothesis tests which have occurred.
5. These commands perform a principal components analysis of the spectra and see if the dimensionality of the data can be reduced.
- (a) **Perform PCA** - computes the first n principal components of the spectra, where n is specified by the user. Each curve can be written as a weighted sum of the principal components. This command also computes the weight associated with each PC. Finally, the program uses covariates selected by the user to predict the weights for each PC for each case.
 - (b) **Plot PCA Curves** - displays the n PC's.
 - (c) **PCA Residuals** - displays the residuals when the weighted sum of the PC's is subtracted from the original data.
 - (d) **PCA Obs vs Fitted** - displays parametric plots of the observed spectra and the weighted sums of PC's.
 - (e) **Regression Residuals** - displays the residuals formed when the PC's summed using the weights predicted by the regression are subtracted from the data.
 - (f) **Regression Obs vs Fitted** - displays parametric plots of the curves predicted by the regression versus the observed spectra.
 - (g) **Display Regression Results** - displays the results of the regressions predicting the weights for each PC using the selected covariates.

The most important feature of these plots is that they are all dynamically linked. As the points selected in one display change, the changes are reflected in all displays.

4 Sample Plots

Figure 1 shows plots of the 199 spectra measured at the first excitation wavelength. Figure 2 shows derivatives of

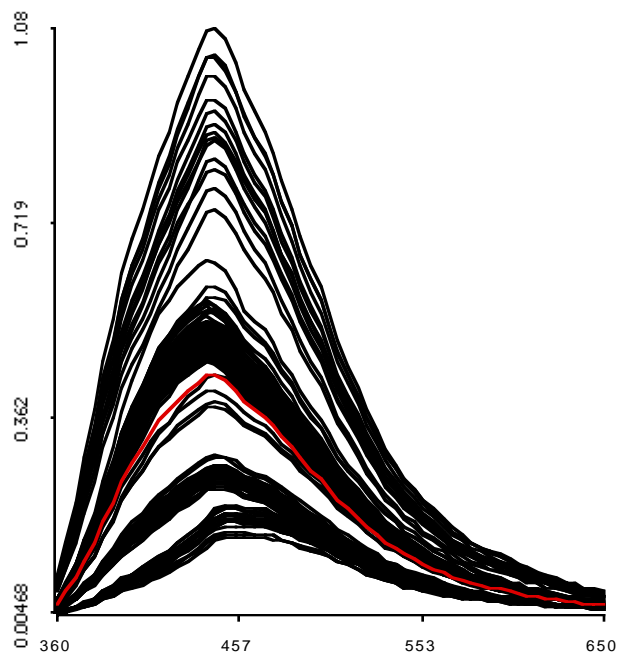


Figure 1: Plots of the spectral curves (emission frequency versus intensity) for each sample taken at the first excitation wavelength. The thick line is the mean of all curves; in a color output it is red.

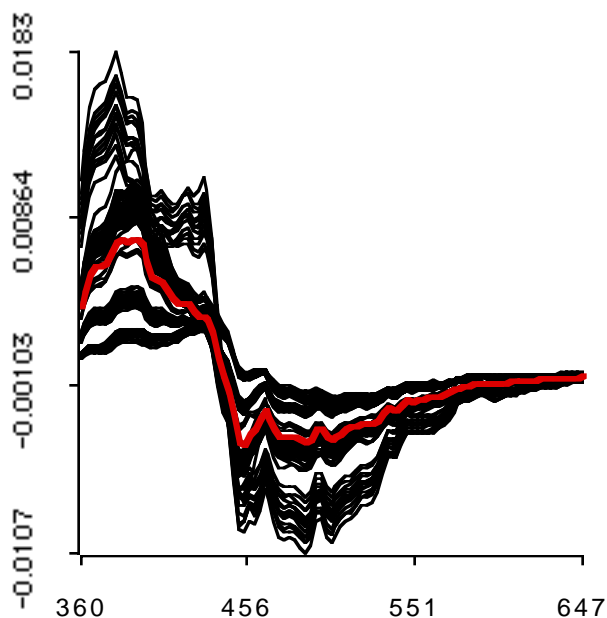


Figure 2: Plots of the derivatives of the spectral curves.

the spectra. The derivatives are computed fitting splines to the data and using finite differences on the splined data. Although the shapes of the curves are similar, there is a great deal of variation from sample to sample. We wish to see if this variation is related to the covariates.

We begin by examining the effects of age on the spectra, since age is well known to effect pathologists assessment of normal versus abnormal cells. Figure 3 shows a histogram for age with the lower values of age selected; Figure 4 shows the corresponding spectra. Figures 5 and 6 show the corresponding plots for the older patients. There is clearly a relation between age and the intensity of the spectra.

To explore this relation further, we need to examine the other covariates. Figure 7 shows a scatterplot matrix of the covariates `age`, `smoke`, `tissue`, and `premen`. The categorical variables have been jittered to improve their visibility on the plot. We note that age is related to the other covariates. Clearly, premenopausal women are younger in general than postmenopausal women, but in this data set the women who smoke are all premenopausal. Further, no transition zone samples are available from postmenopausal women. Thus, the effect of age is difficult to separate from the effects of the other covariates. We can use dynamic techniques to explore this issue. For example, we can use the mouse to brush the scatterplot matrix. First we focus on the scatterplot for age versus menopausal status; then we brush points from youngest to oldest for premenopausal women only. This results of this exploration suggest that even in premenopausal women, age has an effect on the spectra.

It is clear that age has an effect on the intensity of the curves. We can extract the intensity as a separate covariate and see if age has any effects on the shape of the curve in addition to its intensity. We begin by normalizing the curves to all have peak intensity 1.0; this is done with the `Register Inten` by `Max` menu item. The resulting curves are shown in Figure 8. Now we can examine the registered curves for low and high values if age; the resulting plots are shown in Figures 9 and 10. There seems to be a shift in the location of the peak intensity with advancing age; this effect is more striking when seen dynamically using brushing. We can also see the effect of age on the peak intensity of the curves by plotting age versus the peak intensity, which we captured when we registered the curves. This plot is shown in Figure 11. We can see if age has more effect on the spectra at certain emission frequencies by using the `Inten-Covar Scatterplot` menu item. This lets us look at a plot of age versus emission intensity for each emission wavelength. Figures 12 and 13 show the

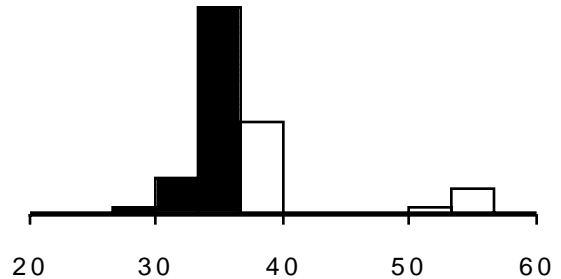


Figure 3: A histogram of age with the lower values selected.

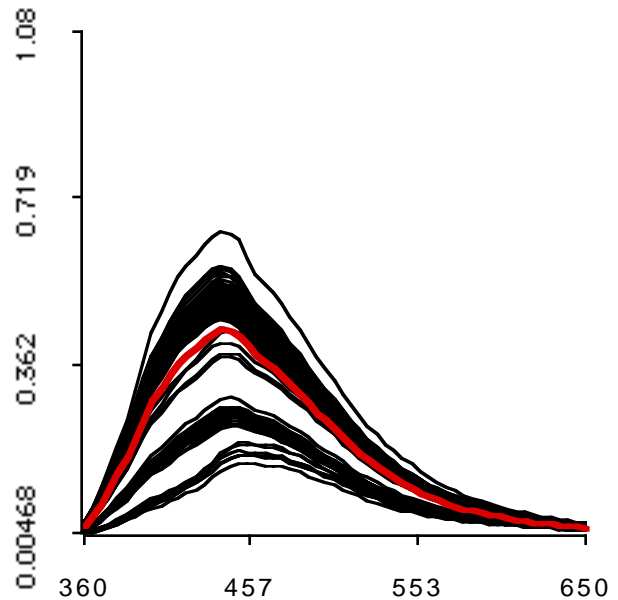


Figure 4: The spectra of patients with lower ages selected in the histogram of age.

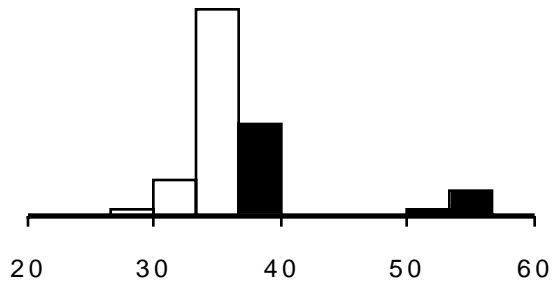


Figure 5: A histogram of age with the higher values selected.

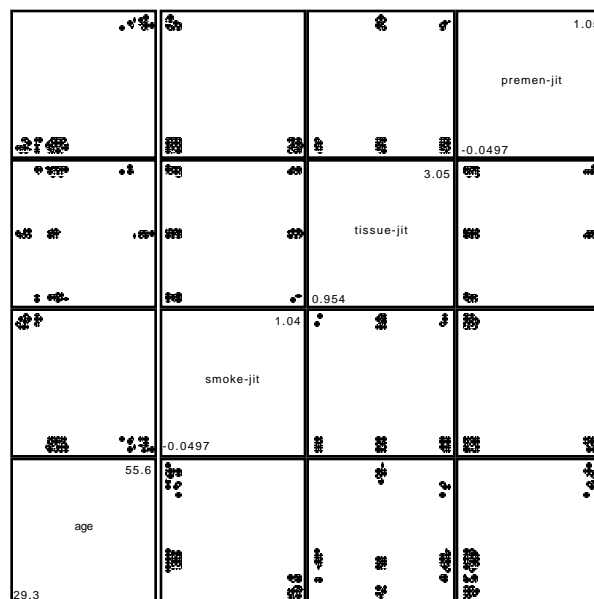


Figure 7: A scatterplot matrix of the covariates. Categorical variables have been jittered.

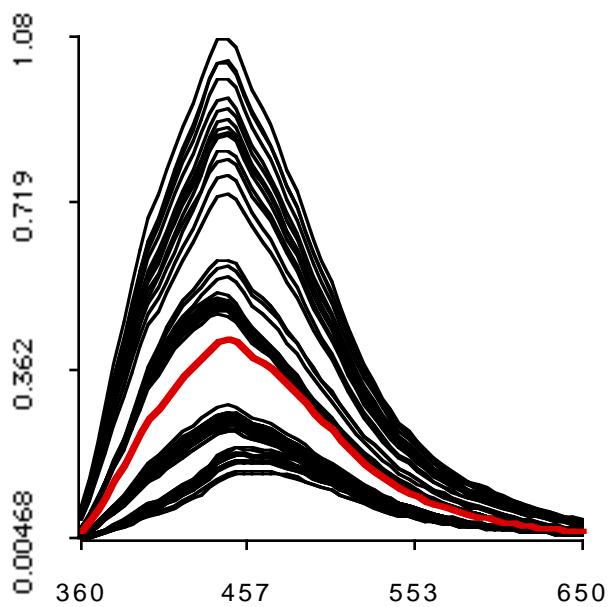


Figure 6: The spectra of patients with higher ages selected in the histogram of age.

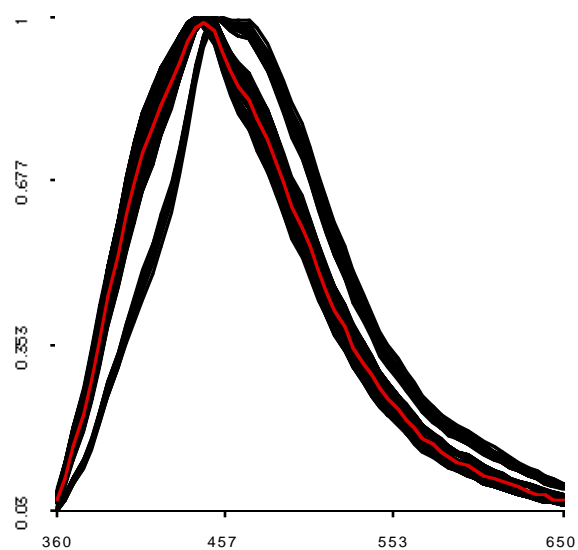


Figure 8: The spectra with all curves registered to have a peak intensity of 1.0.

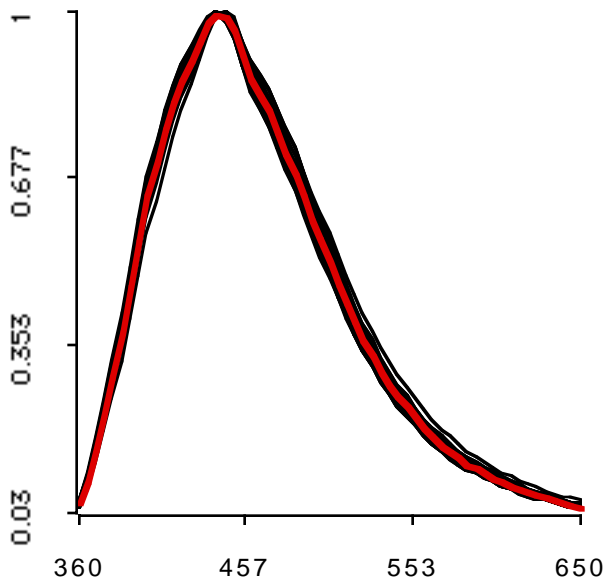


Figure 9: The registered spectra for low values of age.

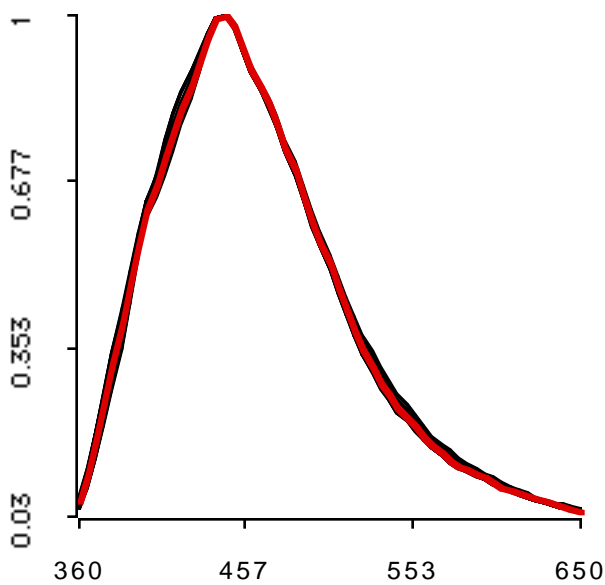


Figure 10: The registered spectra for high values of age.

plots for wavelengths 375 nm and 485 nm. These plots are taken from the spectra registered to peak intensity 1.0. We can also plot the coefficient of the regression of intensity on age for each emission wavelength; this plot is shown in Figure 14. The effect of age seems greater at the lower wavelengths.

Finally, we see if we can use the covariates to predict the spectra. If this can be done, then when the device is used clinically, we can remove much of the patient to patient variability in the spectra. We begin by attempting to reduce the spectral data to a manageable number of dimensions using principal components analysis (PCA); we will attempt to determine a small number of orthonormal basis functions such that the observed spectra can be written as weighted sums of these basis functions. The following analyses are performed on the data scaled to have peak intensity 1.0. We compute the PCA using `Perform PCA`; we set the number of PC's to be 3. The routine computes the PC curves and the weights associated with each observed spectra. These weights give the best least-squares fit when the observed data are approximated by weighted sums of the PC curves. To get an estimate of the goodness of fit, we plot the residuals formed when the optimal weighted sum of PC curves is subtracted from the observed curves. These residuals are shown in Figure 15. By observing the scale of the magnitude of the residuals, it is clear that the fit is quite good; this can be emphasized by plotting the residuals on the same scale as the original data, as is done in Figure 16. Thus, the spectra for the first excitation wavelength for each patient can be summarized in three values - the three PC weights. If we can use the covariates to predict these weights, we can predict the spectra. The `Perform PCA` command performs regressions of the weights on selected variables. For our example data, we regress the weights on age and the peak intensity (calculated earlier). The regression reports that both age and peak intensity are significant predictors ($P < 0.01$) of all three weights. Figure 17 shows a scatterplot of peak intensity versus the first weight; Figure 18 shows a scatterplot of the values observed and predicted for the first weight. These figures suggest a nonlinear model will be necessary to describe the data. By using the predicted weights, we can get predicted spectra based on the covariates; Figure 19 shows the residuals formed when these predicted spectra are subtracted from the original (registered) spectra. A more complete analysis would include all covariates and possible nonlinear effects.

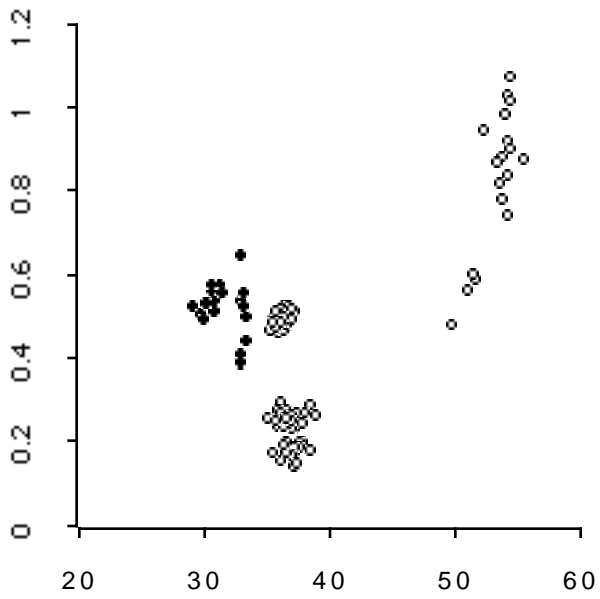


Figure 11: Peak intensity versus age.

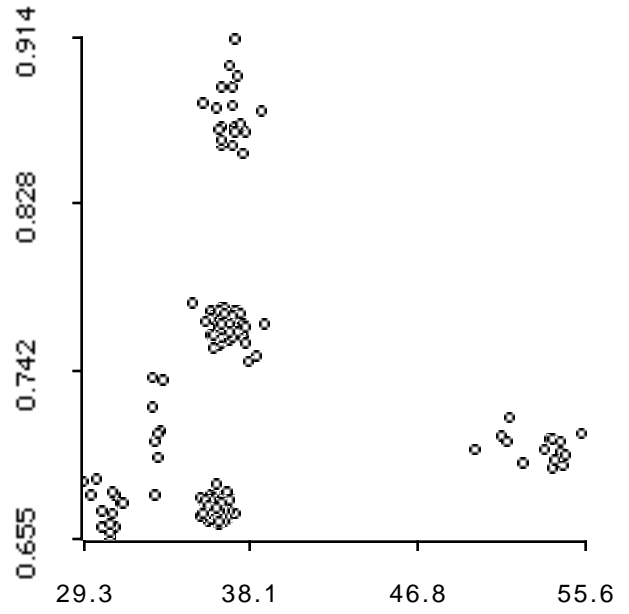


Figure 13: Age versus emission intensity for emission wavelength 485 nm.

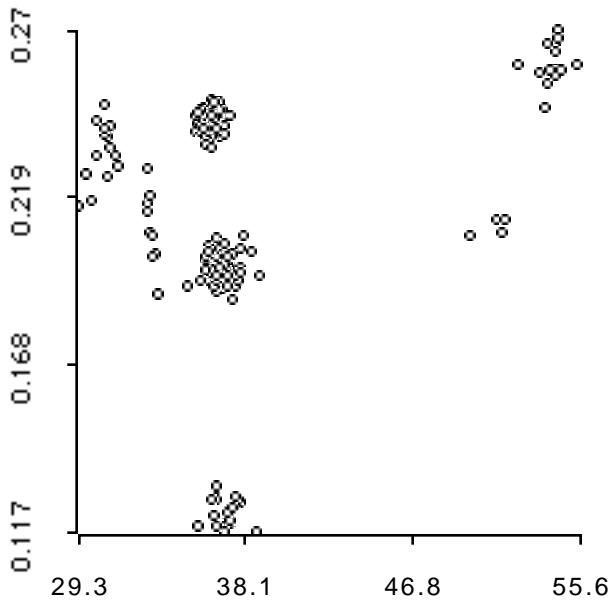


Figure 12: Age versus emission intensity for emission wavelength 375 nm.

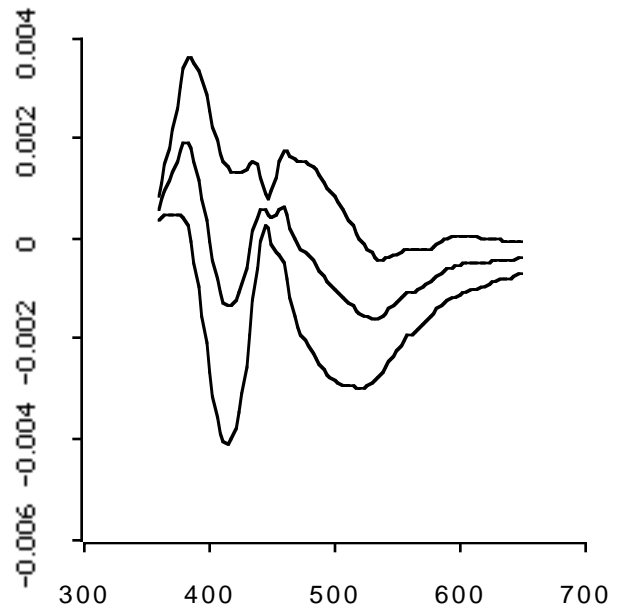


Figure 14: The coefficient of the regression of intensity on age for each emission wavelength; the outer bands are 95% confidence intervals.

Cancer Agency.

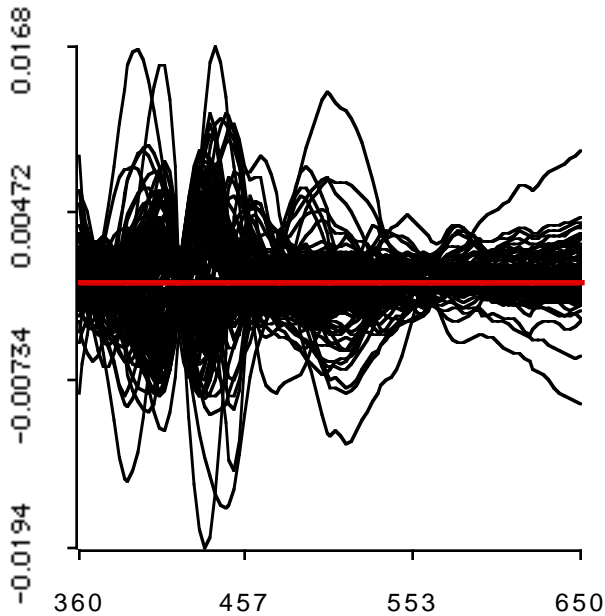


Figure 15: The residuals formed by subtracting the optimally weighted sum of the PC curves from the observed spectra.

5 Summary and Conclusions

This paper has given a brief illustration of some techniques that can be used for the graphical exploration of functional data. Most of the power of these techniques can be seen only when they are used in a dynamic, interactive fashion. The reader is encouraged to explore these techniques using the included code. The code does not represent a finished package; rather, it is meant to demonstrate some approaches and to encourage further development. Please contact the author with all suggestions, criticisms, and comments.

Acknowledgments

The work is supported by grant CA82710 from the National Cancer Institute. The authors wish to acknowledge the extensive contributions of Michele Follen of the Department of Gynecologic Oncology of M. D. Anderson Cancer (UTMDACC), Nan Earle of the Department of Biomathematics UTMDACC, Rebecca Richards-Kortum and Urs Utzinger of the University of Texas at Austin, Dennis Cox of the Department of Statistics, Rice University, and Calum MacAulay of the Cancer Imaging Department of the British Columbia

References

- [1] Ramsay, J.O. and Silverman, B.W. **Functional Data Analysis**, Springer, New York. 1997.
- [2] Tierney, Luke. **LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics**. John Wiley and Sons, New York. 1990.

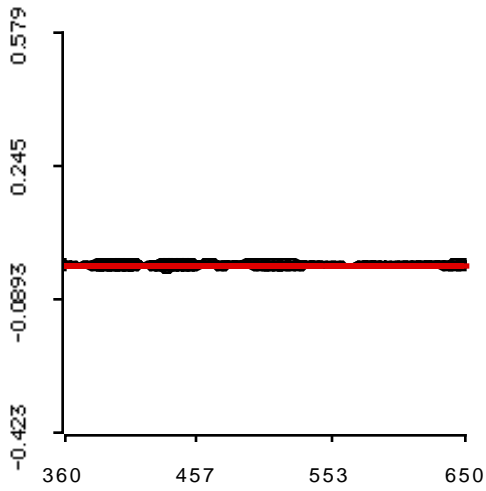


Figure 16: The residuals formed by subtracting the optimally weighted sum of the PC curves from the observed spectra; the scale of the plot is the same as that of the original data.

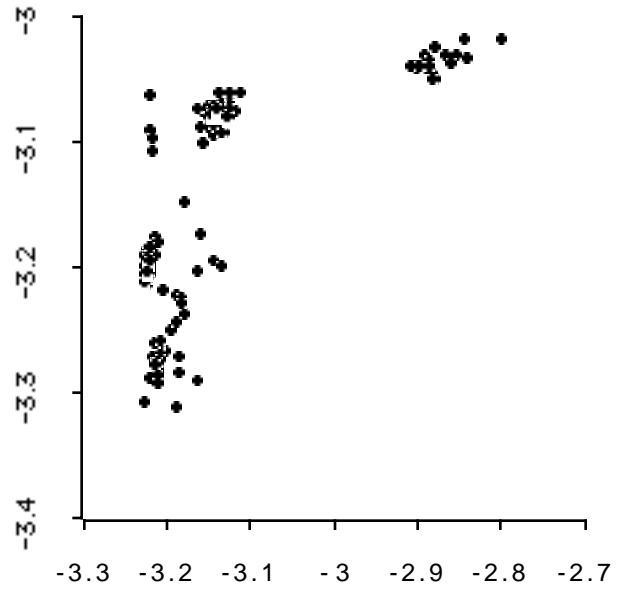


Figure 18: Observed versus predicted value for the first principal component weight.

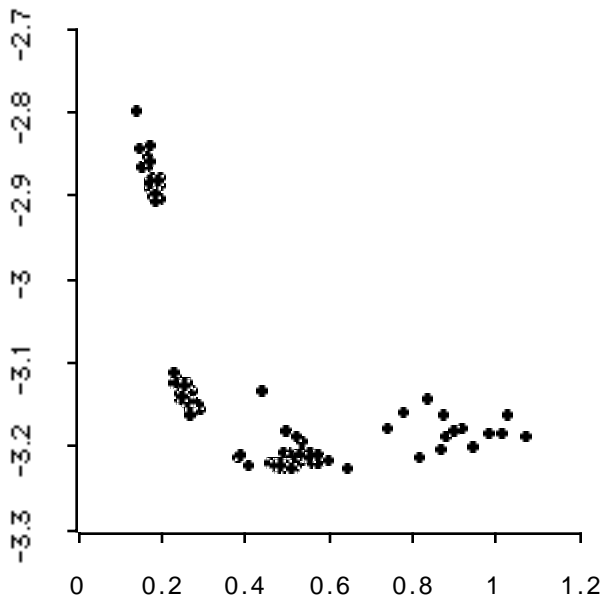


Figure 17: Peak intensity versus the weight of the first principal component.

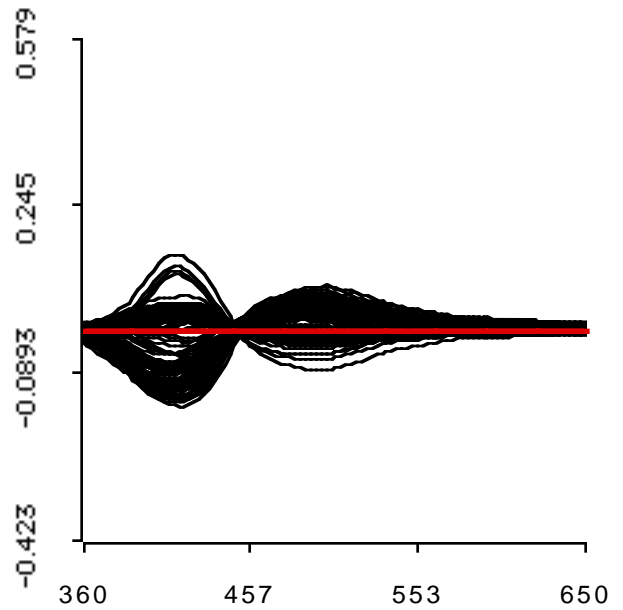


Figure 19: Residuals formed by subtracting the spectra predicted by the covariates from the observed data.