Available online at www.sciencedirect.com

SCIENCE DIRECT®

Gynecologic Oncology 99 (2005) S76 – S83

Gynecologic Oncology

www.elsevier.com/locate/ygyno

Original Research Report

Interactive dynamic graphical techniques for the exploration of functional data

E. Neely Atkinson, Michele Follen*

Department of Biostatistics and Applied Math, the University of Texas M.D. Anderson Cancer Center, Center for Biomedical Engineering, Unit 193, 1515 Holcombe Blvd., Houston, TX 77030, USA

Available online 7 October 2005

Abstract

Introduction. The use of interactive dynamic graphics has become a common practice for the exploration of multidimensional data sets. The availability of powerful and inexpensive hardware and software for graphical computing makes the use of such techniques feasible for the examination of complex forms of data. This paper describes some simple techniques, which were implemented in the LISP-STAT environment, for the visualization of functional data arising from studies of optical technologies used for the detection of cervical intraepithelial neoplasia or squamous intraepithelial lesions.

Materials and methods. The methods demonstrated have been implemented in software coded in LISP-STAT, a free statistical computing package available for most computer systems. The data used in this paper are drawn from a previous study in which fluorescence spectroscopy was measured from cervical sites at 337 nm, 380 nm, and 460 nm excitation in cervical screening patients. The goal of the project is to explore biographical variables to better understand the biology of fluorescence.

Results. 199 measurements were taken in 55 women with normal Pap smears. The data are recorded as spectra showing the intensity of emission excitation versus emission in nanometers. Covariate variables available for analysis are current smoker vs. nonsmoker premenopausal vs. postmenopausal, tissue type (columnar, squamous, and 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. Patient age affects the fluorescent spectrum showing increasing intensity with increasing age. Menopausal status affects the fluorescent spectra coincidentally with age. Smoking and race do not appear to affect the spectra in this sample of patients.

Conclusions. The use of interactive graphical techniques permits the data analyst to examine multidimensional data in intuitive ways. These explorations allow non-statisticians to explore the data in a perceptive manner that may lead to new approaches in algorithm development for optical technologies.

Keywords: Dynamic graphical analysis; Fluorescence spectroscopy; Reflectance spectroscopy; CIN; SIL; Cervical screening; Cervical diagnosis; Functional data; Excitation–emission matrices

Introduction

Interactive dynamic graphics, graphical displays that change in real time in response to user actions, have become a standard method for examining multivariate data and are included in many statistical packages. Typically, the user is presented with linked displays of data; when points are selected on one graph, corresponding points are highlighted in the remaining displays. As the user changes the point selection, the corresponding changes appear in the linked graphs. By watching the way the linked plots change in response to varying selections, the analyst can develop insight into the higher dimensional structure of the data. An excellent introduction to the basic methods is provided by Cleveland and McGill [1].

As computer hardware becomes capable of faster processing, it becomes possible to extend these methods to

* Corresponding author. Fax: +1 713 792 4856.
E-mail address: mfollen@mdanderson.org (M. Follen).

0090-8258/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygyno.2005.07.048

more complex types of data. In previous work, we have described some dynamic graphical techniques for censored survival data [2]; in this paper we extend dynamic graphics to functional data.

Functional data are data that 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 by Ramsay and Silverman [3,4]. This paper presents several techniques for the graphical explorations of functional data. These explorations help locate features of the data that 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 experts in the subject matter under study by providing non-statisticians with ways to examine data directly.

The program

The methods demonstrated in this paper (as well as several additional methods) have been implemented in software coded in LISP-STAT, a free statistical computing package available for most computer systems. For more information, see Tierney [5]. The methods described in this paper are much more effective when viewed dynamically than can be demonstrated in a printed article. We encourage readers to download the software and data and to explore these methods for themselves. The complete set of commands available in the program is described in the Appendix.

Data

The data used in this paper to illustrate the techniques are drawn from an ongoing study of optical methods used to diagnose cervical abnormalities. During a gynecological examination, 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 response to this excitation. The intensity of the fluorescence is measured at approximately 60 wavelengths (the emission wavelengths). The device that produced the data reported in this paper used three excitation wavelengths (337 nm, 380 nm, and 460 nm). For a given excitation wavelength, a plot of emission wavelength versus intensity produced a smooth curve or spectrum. The goal of the project was to develop algorithms that would be 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/clinical-trials/cervix.

The particular data examined here are drawn from 199 measurements taken in 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 hoped that adjustments for covariates will remove a large part of this variability.

Since these data are still being collected and edited, the data used here were formed by taking random linear combinations of cases with similar covariate values. This produces pseudo-data sets that resemble the original data sufficiently enough to illustrate the techniques being presented. The pseudo-data sets will be available for download.

Exploring a single covariate

We begin by studying the effects of a single covariate on the emission spectra for a single, fixed excitation wavelength. Since we know that age affects the pathological assessment of normal versus abnormal tissue, we examined the relationship between patient age and the spectra.

All of the emission spectra from the first excitation wavelength (337 nm) are shown in Fig. 1. There is a great deal of variability in intensity from spectrum to spectrum. We wish to determine whether this variability is affected by some covariate value, e.g., patient age. One way to study this visually is by linking the plot of the spectra to a histogram of age. As various values for age are selected in the histogram, we can study the corresponding changes in the displayed spectra. Fig. 2 shows a histogram for age with the lower values of age selected; Fig. 3 shows the corresponding spectra. Figs. 4 and 5 show the corresponding plots for the older patients.

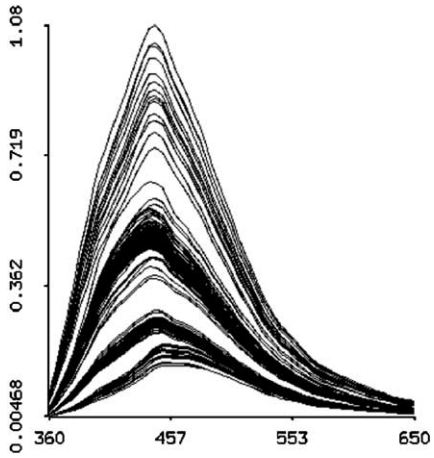

Fig. 1. Plots of the spectral curves (emission frequency versus intensity) for each sample taken at the first excitation wavelength.

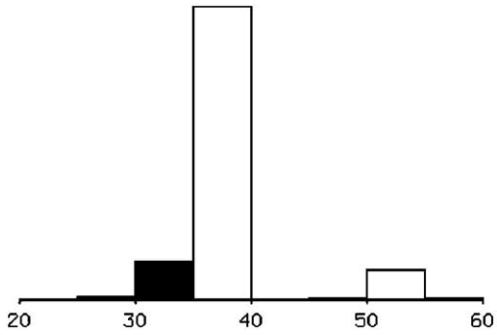

Fig. 2. A histogram of patient age with the lower values selected.

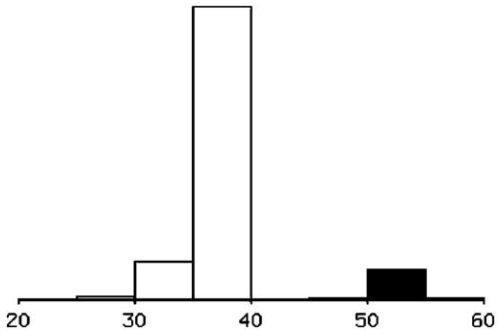

Fig. 4. A histogram of patient age with the higher values selected.

There is clearly a relation between age and the intensity of the spectra; patients with higher age values tend to have higher emission intensities. The differences in intensity from spectrum to spectrum can be immediately seen. We would also like to determine whether there is a relationship between patient age and the shape of the spectra, aside from the intensity. This can be accomplished by registering all the spectra to have a maximum intensity of 1.0 and then selecting low and high values for age from the histogram. The registered spectra corresponding to low and high values for age are shown in Figs. 6 and 7. There does indeed seem to be a shift to the right associated with higher values for age; the effect is more pronounced when viewed by dynamically brushing the 3 histogram for age. This form of inspection could be continued, for example, by further registering the curves by rescaling the emission frequencies so that the peak intensity for each spectrum occurs at the same point in all curves.

We could then look for different degrees of curvature in various subgroups of the data. Examining various aspects of the spectra, such as peak intensity or the emission wavelength at which the peak occurs, is a vital step in determining parametric forms for more classical statistical analyses. We will not pursue this approach here, however. When we rescaled the curves to have a peak intensity of 1.0, we saved the peak intensity for each spectrum and treated it as a separate covariate. For example, when we plotted age against peak intensity, as shown in Fig. 8, we again saw a positive association between patient age and intensity. Of course, it is possible that the effects of age vary for different emission wavelengths, i.e. for different locations in the spectra. Information on the variability of the effects of age will be vital for classifying patients using the spectra. If we find that certain portions of the spectra are relatively unaffected by age, but still differ between normal and abnormal tissue, our classification problem becomes much simpler.

To examine this possibility that tissue type interacts with age, we noted that, for a fixed emission wavelength, the relation between age and emission intensity is totally captured by a simple scatterplot; as we varied the emission wavelength, we could examine how the scatterplots changed. Figs. 9 and 10 show the plots of age versus intensity for wavelengths 375 nm and 485 nm. The figures make clear that the relation between age and intensity did change with emission wavelength. In practice, the emission wavelength is selected using a slider and can be changed dynamically. To explore this relation more formally, we performed a regression of intensity on age for each emission wavelength and plotted the regression coefficient of age against emission wavelength. The results are shown in

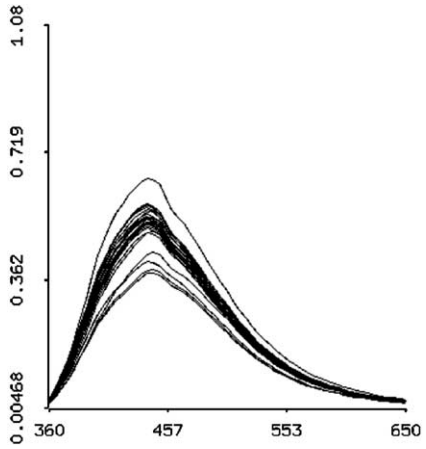

Fig. 3. The spectra of patients with lower ages selected in the histogram of age.

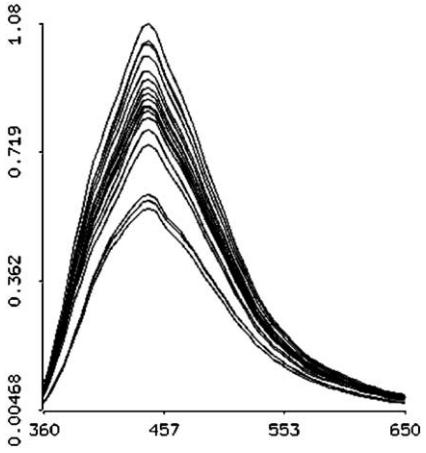

Fig. 5. The spectra of patients with higher ages selected in the histogram of age.

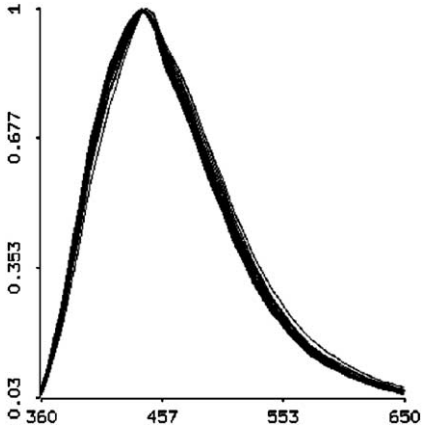

Fig. 6. The registered spectra for low values of patient age.

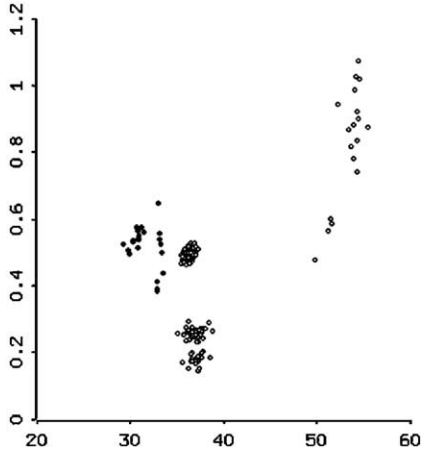

Fig. 8. Patient age versus peak intensity.

Fig. 11. The effect of age seems greater at the lower wavelengths.

Examining multiple covariates

The methods described so far permit us to explore the effects of a single covariate. In general, however, we would have a number of covariates of potential interest. A simple way to examine the relation between multiple covariates and the spectra is to link a scatterplot matrix of the covariates to a display of the spectral data. Fig. 12 shows a scatterplot matrix of the covariates patient age, smoker, tissue type, and menstrual status. 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 there are other relationships as well. In this data set, the women who smoke were all premenopausal. Furthermore, no transition zone samples were available from postmenopausal women as this tissue becomes less accessible following menopausal. 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 of age and other covariates. For example, we can use the computer's mouse to brush the scatterplot matrix. Suppose we wish to determine whether age has an effect in addition to that of menopausal status. First, we focus on the scatterplot for age versus menopausal status. In this plot, we then brush points from youngest to oldest for premenopausal women only, watching at the same time the changes in the spectra displayed. The results of this exploration suggest that, even in premenopausal women, age has an effect on the spectra. Finally, we see if we can predict the expected spectrum for a given patient based on that patient's covariate values. If this can be done, then when the device is used clinically, we can examine the way in which the patient's observed spectra differ from those expected for a normal healthy patient; the adjustment for covariate values may remove much of the patient-to-patient variability. We began by attempting to reduce the spectral data to a manageable number of dimensions using principal components analysis (PCA); we attempted to determine a small number of ortho-normal basis functions such that the observed spectra could be written as weighted sums of

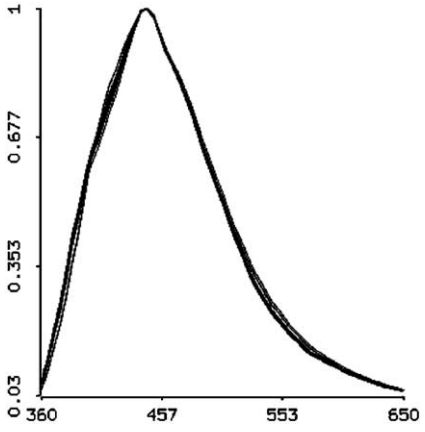

Fig. 7. The registered spectra for high values of patient age.

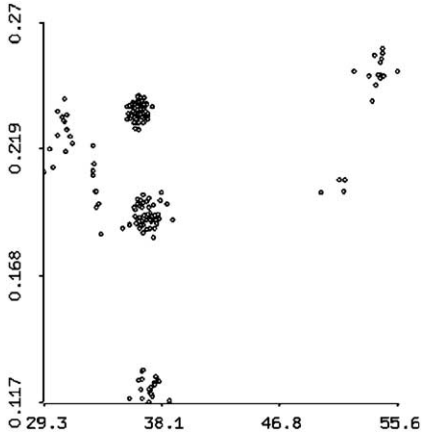

Fig. 9. Patient age versus emission intensity for emission wavelength 375 nm.

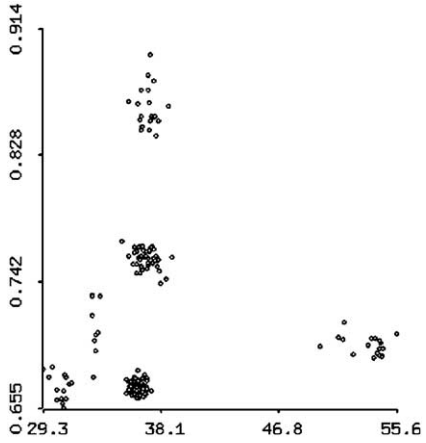

Fig. 10. Patient age versus emission intensity for emission wavelength 485 nm.

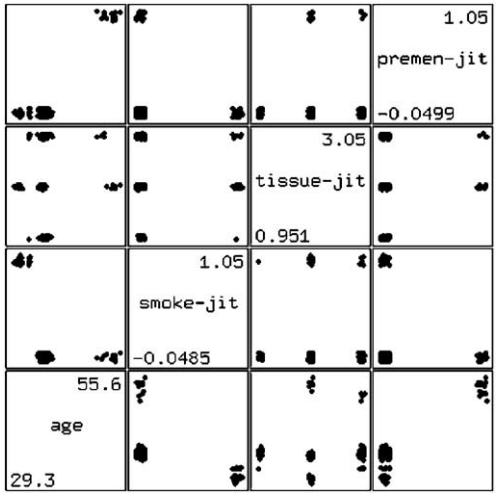

Fig. 12. A scatterplot matrix of the covariates. Categorical variables have been jittered.

these basis functions. To illustrate this, we considered the data scaled to have a peak intensity of 1.0, and we computed the first three principal components for the spectral data. For each observed spectrum, we computed the weights associated with each of the three principal components to give the best least-squares fit when the observed data were approximated by weighted sums of the principal component curves. To get an estimate of the goodness of fit, we plotted the residuals formed when the optimal weighted sum of principal component curves is subtracted from the observed curves. These residuals are shown in Fig. 13. 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 shown in Fig. 14. Thus, the spectra for the first excitation wavelength for each patient can be summarized in three values: the three principal component weights. If we can use the covariates to predict these weights, we can predict the spectra. When we regressed the weights on age and the peak intensity (as calculated earlier), we found that both covariates were significant predictors ($P < 0.01$) of all three weights. Fig. 15 shows a scatterplot of peak intensity versus the first weight. Fig. 16 shows the values observed and predicted for the first weight. These figures suggest that a nonlinear model would be necessary to describe the data. By using the predicted weights, we can get predicted spectra based on the covariates; Fig. 17 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.

Future developments

The analyses presented in this paper all display a single emission spectrum. Versions of the device under development will measure emission spectra at a number of excitation wavelengths and produce an excitation–emission matrix. This matrix can be viewed as a two-dimensional

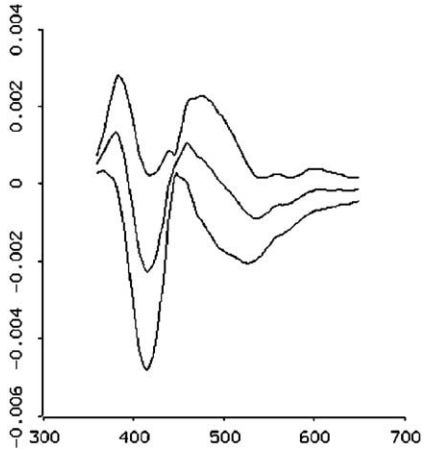

Fig. 11. The coefficient of the regression of intensity on patient age for each emission wavelength; the outer bands are 95% confidence intervals.

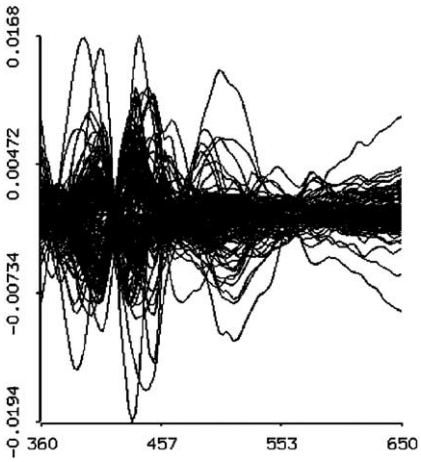

Fig. 13. The residuals formed by subtracting the optimally weighted sum of the PC curves from the observed spectra.

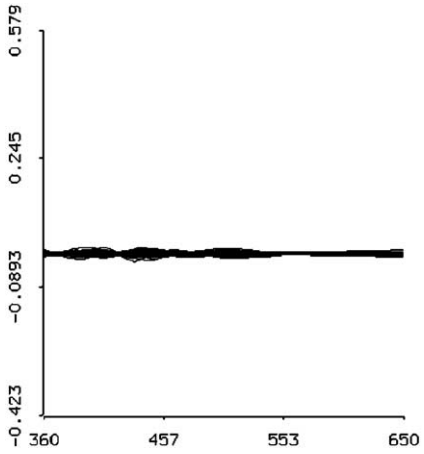

Fig. 14. The residuals formed by subtracting the optimally weighted sum of the PC curves from the observed spectra; the scale if the plot was the same as that of the original data.

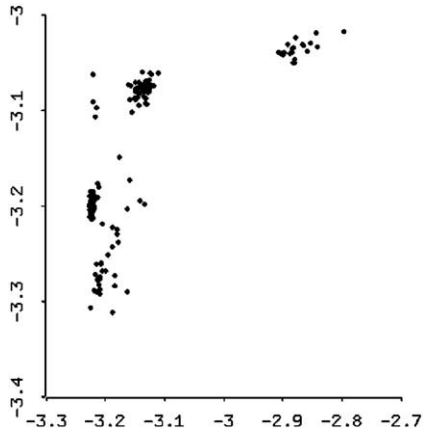

Fig. 16. Observed versus predicted value for the first principal component weight.

surface. We are investigating ways in which the displays presented here can be extended to surfaces. In particular, we are looking at two-dimensional analogs of principal components analysis based on the multilinear, singular value decomposition as described by de Lathauwer et al. [6], a multidimensional equivalent of the standard singular value decomposition.

We have predicted the spectra based on covariates using principal components analysis; Faraway [7] takes an approach to predicting functional outcomes based on regression analysis; we are examining graphical displays based on this approach. It is also possible to consider the functional data as a predictor; for example, we might wish to classify tissue as normal, abnormal, or cancer based on the emission spectra. Marx and Eilers [8] have developed methods based on penalized splines for such problems. We are currently examining these methods. Nason [9] has described extensions of projection pursuit applicable to functional data. Although these techniques are computer intensive, it should be possible to establish dynamic displays based on them.

Finally, an alternative to principal components analysis is provided by independent components analysis, which yields components that are not only orthogonal but also statistically independent [10]. These techniques have been used in signal processing applications to separate competing signals. In situations in which the observed functional data arise from a mixture of several input sources, independent components analysis may provide more interpretable results than principal component analysis. We are implementing graphics similar to those demonstrated here for principal component analysis using independent components instead.

Conclusions

The use of interactive graphical techniques permits the data analyst to examine multidimensional data in intuitive, nonparametric ways. These explorations can guide the

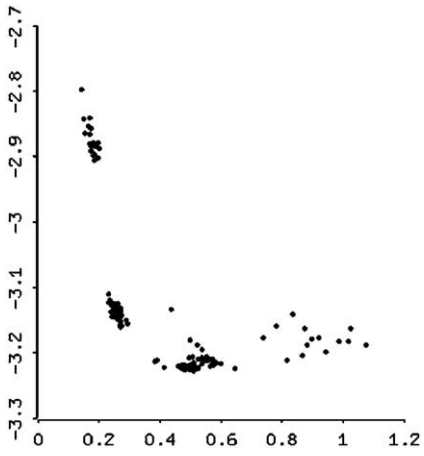

Fig. 15. Peak intensity versus the weight of the first principal component.

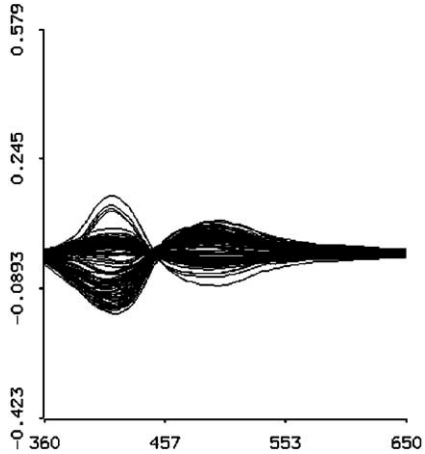

Fig. 17. Residuals formed by subtracting the spectra predicted by the covariates from the observed data.

construction of more conventional statistical models that can be used for hypothesis testing and estimation. The continuing increase in the availability of inexpensive, powerful desktop computers permits analysts to explore more and more complicated forms of data. Since the explorations are based on graphical displays rather than statistical models, researchers who are expert in the subject matter under investigation, but not in statistics, can understand the results and assist in the exploration of the data.

The techniques described in this paper are best understood when they are viewed in real time. Interested readers are encouraged to download the programs and data described in this paper and to examine the techniques for themselves. The author would greatly appreciate any suggestions or recommendations for modifications, extensions, or additions to these techniques.

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 Biomedical Engineering program 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 Cancer Agency.

This research was supported by the National Cancer Institute under Program Project Grant 3PO1-CA82710-04.

Appendix A. Complete program commands

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

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 are 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 affect 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 affect the disk file from which the data was read.
5. Export Covariate—stores a variable from the program into LISP-STAT.
6. Import Covariate—brings a variable from LISP-STAT into the data known to the program. Since the program runs within the LISP-STAT environment, all of the features of LISP-STAT are available. The Import and Export commands facilitate moving data between the program and LISP-STAT. This makes it simple, for example, to perform various transformations of the covariates by exporting them, processing them as desired, and importing them back into the program.

The Analysis menu

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 $f1i(w)$; let the spectrum measured at wavelength 2 be $f2i(w)$. Then, we plot the parametric curve given by $(f1i(w), f2i(w)$ for w 2 (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. 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.

(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* PCs.
   (c) PCA Residuals—displays the residuals when the weighted sum of the PCs is subtracted from the original data.
   (d) PCA Obs vs. Fitted—displays parametric plots of the observed spectra and the weighted sums of PCs.
   (e) Regression Residuals—displays the residuals formed when the PCs summed using the weights predicted by the regression are subtracted from the data. Curves predicted by the regression versus the observed spectra.
   (f) 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.

References

[1] Cleveland WS, McGill ME, editors. Dynamic graphics for statistics. Belmont, CA: Wadsworth and Brooks/Cole; 1988.
[2] Atkinson EN. Interactive dynamic graphics for exploratory survival analysis. Am Stat 1995;49:77–84.
[3] Ramsay JO, Silverman BW. Functional data analysis. New York: Springer; 1997.
[4] Ramsay JO, Silverman BW. Applied functional data analysis. New York: Springer; 2002.
[5] Tierney L. LISP-STAT: An object-oriented environment for statistical computing and dynamic graphics. New York: John Wiley and Sons; 1990.
[6] de Lathauwer L, de Moor B, Vandewalle J. A multilinear singular value decomposition. SIAM J Matrix Anal Appl 2000;4:1253–78.
[7] Faraway JJ. Regression analysis for a functional response. Technometrics 1997;39:254–61.
[8] Marx BD, Eilers PHC. Generalized linear regression on sampled signals and curves: a P-spline approach. Technometrics 1999; 41:1–13.
[9] Nason GP. Functional projection pursuit. Comput Sci Stat 1997; 29:330–6.
[10] Roberts S, Everson R, editors. Independent component analysis: principles and practice. Cambridge: Cambridge Univ. Press; 2001.