Analyzing time-course microarray data using functional data analysis - a review

1 Introduction

Time-course microarray data have often been analyzed by treating the gene expression profiles as multivariate observations. However treating the data as multivariate does have some limitations. Gene expression data tend to exhibit problems such as high dimensionality, missing values, large amounts of measurement error, correlation between observations over time, etc. Many of the multivariate techniques (e.g. principal components analysis, ANOVA, linear mixed models, etc.) used to analyze such data have difficulties handling missing values, may require uniform sampling for all genes, fail to account for the correlation between measurements made on the same gene and/or do not facilitate the removal of noise from the measured data thus ignoring any smoothness that may be evident in the expression profiles. Functional data analysis (FDA) as outlined in Ramsay and Silverman (2005) is a statistical technique that treats the entire sequence of measurements for an individual expression profile as a single functional entity rather than a set of discrete values. The term functional refers to the belief that the gene expression data are being generated by some underlying smooth *function* and the discrete measurements collected are a snapshot of that function at various points in time (Ramsay and Silverman, 2005). Functional data analysis circumvents many of the difficulties associated with the treating time-course microarray data as multivariate and has proven to be extremely useful in the analysis of such data.

The first step in FDA is creating an estimate of the gene expression curves from the original (and possibly noisy) raw data. This step is called smoothing and involves representing the expression curves as a linear combination of a finite number of basis functions (e.g. spline, Fourier, wavelets, etc.). Representing the expression profiles using basis functions allows for the inclusion of non-uniformly sampled data, enables the experimenter to estimate expression values at times different from those used in the original experiment, allows for the imputation of missing values and facilitates the removal of noise from the measured data. Once the data have been smoothed, many multivariate techniques which have been extended to the functional case, e.g. principal components analysis, discriminant analysis and regression analysis, can be applied. These have been used to satisfy some of the main aims in modeling gene expression data, i.e. dimension reduction and clustering to determine groups of co-expressed genes, tests for differential expression between genes across treatment groups, discrimination and classification of genes, etc. and have been shown to have advantages over multivariate approaches. Sections 2, 3 and 4 of his paper give an outline of some FDA techniques used in the analysis of time-course gene expression data. It should be noted that many technical details regarding methods of computation etc. have been omitted since this paper constitutes a review of FDA procedures in microarray analyses. Interested readers

are referred to Ramsay and Silverman (2002), Ramsay and Silverman (2005) and Ramsay, Hooker, and Graves (2009) for additional information and details regarding computation and performing FDA in the R and MATLAB software suites using the `fda` package. Section 5 demonstrates how FDA has been used in time-course microarray analyses to date and describes how FDA can provide additional information about the behavior of gene expression through time. To date the largest proportion of research papers using FDA techniques have focussed on clustering expression profiles as discussed in Section 5.1. Less work has been carried out in the other main areas of interest though there has been an increase in the use of FDA techniques in other microarray analyses such as tests for differential expression, discriminating between groups of genes and modeling the relationships between expression profiles. These are discussed in more detail in Sections 5.2 and 5.3. Section 6 suggests some possible areas for future research using FDA methods that have not yet been applied to time-course microarray data. Section 7 provides some illustrative examples and gives the results of applying a number of FDA techniques to real time-course gene expression data.

2 Smoothing

Gene expression over time can be thought of as arising from a smooth underlying process or function $g(t)$. However as stated in Section 1, gene expression data are typically measured at a discrete number of time points and often contain large amounts of measurement error, have missing values or measurements taken at different points in time for each gene, etc. Therefore we can write the discrete expression values y_{ij} for the ith gene $i = 1, \ldots, N$ measured at the jth time point $j = 1, \ldots, n_i$ using the model

$$y_{ij} = g_i(t_{ij}) + \varepsilon_{ij}, \tag{1}$$

where n_i denotes the number of measurements for the ith gene, $g_i(t)$ is the smooth expression curve for the ith gene and ε_{ij} is measurement error or noise which can be correlated or uncorrelated. As stated in Ramsay and Silverman (2005), assuming that the error terms are uncorrelated can be unrealistic in a FDA setting since the variance of the errors is likely to change over time or neighboring ε_{ij}'s may be correlated. However, the authors indicate that explicitly modeling variable variance or autocorrelation structure in the errors may not always be necessary if the resulting function estimates are indistinguishable from those obtained from assuming the errors are independent. In any case, it is always pertinent to keep in mind that incorporating more complex error structures may be beneficial and result in better estimates.

A key step in FDA is to determine an estimate of the smooth expression curve $g_i(t)$ which is achieved via smoothing methods. Smoothing methods represent the discrete expression values as a linear combination of K known functions called basis functions $\{\phi_1(t), \ldots, \phi_K(t)\}$ such that

$$g_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) = \Phi \mathbf{c}_i \tag{2}$$

is a smooth expression curve. The basis functions to be used in (2) are chosen to reflect the characteristic behavior of the data, e.g. Fourier basis functions are suitable for periodic data, B-spline basis functions are suitable for non-periodic data, etc. In addition, it is necessary to estimate the vector of basis function coefficients $\mathbf{c}_i$. One way to estimate $\mathbf{c}_i$ is via least squares

$$\|\mathbf{y}_i - \Phi \mathbf{c}_i\|^2, \tag{3}$$

where Φ is an $n_i \times K$ matrix of basis functions evaluated at t_{ij}. In this instance the number of basis functions K affects the smoothness of the results. The choice of an optimal value for K is a complex problem and it is difficult to control the amount of smoothing applied to the data. As a result, Ramsay and Silverman (2005) advocate the use of smoothing splines where $K = n_i$ and over-fitting is controlled by adding a penalty term to the optimization problem. This penalty term penalizes the curvature of the resulting expression curves

$$\|\mathbf{y}_i - \Phi \mathbf{c}_i\|^2 + \lambda \int_{\mathcal{T}} D^2 g_i(t) \mathrm{d}t, \tag{4}$$

where D^2 denotes the 2nd derivative (curvature) of $g_i(t)$ and T denotes the entire time interval. The trade-off between fit to the data and smoothness is controlled by the smoothing parameter λ and a value for λ can be chosen via cross-validation or generalized cross-validation. As stated in Section 1, representing the expression profiles using basis functions has several advantages: facilitating the removal of measurement error, allowing the inclusion of non-uniformly sampled data, enabling the estimation of expression values at times different from those used in the original experiment and allowing for the imputation of missing values. Another key advantage of representing the raw gene expression data as gene expression functions is the availability of derivative information. This is particularly useful for the analysis of time-course microarray data since gene expression is part of a biological system and much information about the behavior of a system is contained in the derivatives. As a result, in some instances smooth estimates of the derivative(s) of the expression curves may be required rather than estimates of the original expression profiles. When the interest is in the estimation of the derivatives, Ramsay

and Silverman (2005) suggest altering the penalty term on the RHS of (4) to ensure smoothness of the *derivative*. For example if velocity curves are required, the curvature of the velocity curves should be penalized. This equates to changing the penalty term to

$$\lambda \int_{\mathcal{T}} D^3 g_i(t) \mathrm{d}t. \tag{5}$$

In general if a derivative of order m is the highest required, derivatives of order $m+2$ should be penalized. Once the smooth expression curves (or derivatives) have been estimated, further analysis can be carried out.

3 Functional Principal Components Analysis

Functional principal components analysis (FPCA) is the functional analogue of multivariate principal components analysis. FPCA is a very common method used to summarize functional data and is used to identify the characteristic features of a set of functions. It also provides a way of looking at the variance structure, which can often be more informative than a direct examination of the variance-covariance function. Intuitively FPCA determines the main modes of variation in a set of curves. The rth functional principal component (FPC) is the weight function $\xi_r(t)$ chosen to maximize the variance of the functional principal components scores

$$f_{ir} = \int_{\mathcal{T}} \xi_r(t)[y_i(t) - \bar{y}(t)]\mathrm{d}t, \tag{6}$$

subject to the constraints

$$\|\xi_r(t)\|^2 = 1 \quad \text{and} \int_{\mathcal{T}} \xi_r(t)\xi_m(t)\mathrm{d}t = 0, \quad r < m. \tag{7}$$

This leads to the eigenequation

$$\int_{\mathcal{T}} v(s,t)\xi(s)\mathrm{d}s = \rho\xi(t), \tag{8}$$

where

$$v(s,t) = N^{-1} \sum_{i=1}^{N} [y_i(s) - \bar{y}(s)][y_i(t) - \bar{y}(t)] \tag{9}$$

is the covariance function, $\bar{y}(t)$ is the cross-sectional mean and ρ is an appropriate eigenvalue. Each functional principal component $\xi_r(t)$ is a function describing a particular pattern of behavior over the entire time interval. A high positive/negative

score on a particular FPC indicates that that gene is exhibiting the behavioral pattern represented by that component. FPCA has been used extensively when analyzing time-course microarray data. It has been applied when clustering expression profiles as discussed in Section 5.1. In addition, since the FPCs form a set of orthonormal basis functions, some authors have used the FPCs to approximate the data (and/or covariate functions) such that

$$\hat{y}_i(t) = \bar{y}(t) + \sum_{r=1}^{R} f_{ir}\xi_r(t), \tag{10}$$

e.g. when using the functional principal components approach to estimate the regression functions as shown in Sections 4 and 5.3.

4 Functional Regression Analysis

Functional linear models attempt to express one dependent variable as a linear combination of other features or measurements. A model can be functional in one of two instances; the dependent variable is a function or one or more independent variables are functions. Coefficient vectors β (as given in standard multivariate regression problems) become coefficient functions $\beta(t)$ and a key theme in functional regression analysis is estimating $\beta(t)$ to ensure the results are interpretable. For example, say we have a scalar response variable y_i and a predictor function $x_i(t)$ then we can write

$$y_i = \alpha + \int_{\mathcal{T}} x_i(t)\beta(t)\mathrm{d}t + \varepsilon_i, \tag{11}$$

where α is the intercept term and $\beta(t)$ is the regression coefficient function. When y_i is binary, this reduces to a functional logistic regression model. Other functional regression models include the varying coefficient model

$$y_i(t) = \alpha(t) + \sum_{m=1}^{M} \beta_m(t)x_{im} + \varepsilon_i(t), \tag{12}$$

where the response is now a function and the predictor is a continuous variable whose relationship with $y_i(t)$ changes over time; the concurrent functional model

$$y_i(t) = \alpha(t) + x_i(t)\beta(t)\mathrm{d}t + \varepsilon_i(t), \tag{13}$$

where both the response and predictor are functions but the value of $y_i(t)$ depends only on the current value of $x_i(t)$ and the non-concurrent model

$$y_i(t) = \alpha(t) + \int_{\mathcal{T}} x_i(s)\beta(s,t)\mathrm{d}s + \varepsilon_i(t), \tag{14}$$

where the value of $y_i(t)$ is influenced by $x_i(t)$ over all t. In each case there is an issue with under-determination, i.e. there are a finite number of observations to determine the infinite-dimensional $\beta(t)$. This results in an infinite number of possible solutions for $\beta(t)$. There are three main ways to overcome this problem. We present the implementation details for the simplest case when the predictor(s) are functions and the response is a scalar. These results can be generalized to the case when both the predictor(s) and the response are functions (see Ramsay and Silverman, 2005, Ramsay et al., 2009 for implementation details).

The first method assumes that both the predictor $x_i(t)$ and the regression function $\beta(t)$ can be represented using a finite number of K_z and K_β basis functions respectively, such that

$$\begin{aligned} x_i(t) &= \sum_{k=1}^{K_z} c_{ik}\psi_k(t) \\ \beta(t) &= \sum_{k=1}^{K_\beta} b_k\phi_k(t) \end{aligned} \tag{15}$$

(we call this the basis function approach). However if K_β is too large the total number of basis functions may still exceed the number of observations available, while if K_β is too small the resulting estimate of $\beta(t)$ may miss important features in the data. As a result a second method can be employed (termed here the roughness penalty approach), which involves estimating $\beta(t)$ using a roughness penalty by minimizing the penalized sum of squares

$$\text{PENSSE}(\alpha,\beta) = \sum_{i=1}^{N}[y_i - \alpha - \int_{\mathcal{T}} x_i(t)\beta(t)\mathrm{d}t]^2 + \lambda\text{PEN}[\beta(t)], \tag{16}$$

where PEN$[\beta(t)]$ is a penalty suitable for the problem under consideration (e.g. penalizing the second derivative as in (4)). This approach allows for more direct control over smoothing which reduces the chance that important features are missed by using too few basis functions. The third method regresses $\mathbf{y}$ on the first R principal component scores for the functional covariate. This involves expressing the functional covariates as

$$x_i(t) = \bar{x}(t) + \sum_{r=1}^{R} f_{ir}\xi_r(t), \tag{17}$$

where $\bar{x}(t)$ denotes the mean curve and

$$f_{ir} = \int_{\mathcal{T}} \xi_r(t)(x_i(t) - \bar{x}(t))\mathrm{d}t \tag{18}$$

denotes the score on the rth component $\xi_r(t)$. y_i can then be represented by the model

$$y_i = \alpha + \sum_{r=1}^{R} f_{ir}\beta_r + \varepsilon_i \tag{19}$$

and the coefficient function $\beta(t)$ can be re-constructed as $\beta(t) = \sum_{r=1}^{R} \beta_r \xi_r(t)$.

Müller and Yao (2008) have extended this approach to functional additive models, where the linear relationship between y_i and the FPC scores shown in (19) is replaced by an additive relationship as given by

$$y_i = \alpha + \sum_{r=1}^{R} g_r(f_{ir}) + \varepsilon_i, \tag{20}$$

where $g_r(\cdot)$ is an arbitrary functional relationship. The functions $g_r(\cdot)$ are estimated using a local linear regression to the data $(f_{ir}, y_i)_{i=1,\ldots,n}$ such that

$$\sum_{i=1}^{n} K_1\left(\frac{f_{ir} - x}{h_r}\right)[y_i - \beta_0 - \beta_1(x - f_{ir})]^2 \tag{21}$$

is minimized with respect to β_0 and β_1, h_r is the bandwidth and $K_1(\cdot)$ is a kernel function. This results in more flexible models and allows for the direct examination of the role of each eigenfunction in predicting the response.

Using functional linear models overcomes problems associated with multivariate regression analysis. These include having observations measured at different time points, the high correlation between observations on the same gene, difficulties encountered due to the high dimensionality of both the response and covariate(s) functions (where the number of observations for each gene may exceed the sample size), the need to use multiple testing procedures and incorporating the smoothness of the underlying expression profiles. Using the FDA approach to analyze data also has the advantage of providing derivative information which greatly extends the power of FDA over multivariate methods. Functional linear models can be used to provide direct examination of relationships between derivatives that could otherwise only be studied indirectly. The use of functional linear models in the study of relationships between derivatives is discussed in Section 6.2. As stated in that section, to date derivative information has not been used extensively when analyzing expression profiles. However, we believe that since gene expression is part of a biological system modeling such relationships may prove insightful for these types of data.

5 Applications to Gene Expression Data

5.1 Clustering

There are many clustering algorithms available to cluster gene expression data, e.g. k-means, hierarchical clustering, self organizing maps, fuzzy clustering, Bayesian clustering, multivariate Gaussian mixture models, etc. However these multivariate methods have their limitations. Some do not account for between time-point correlation or assume the correlation has some specified structure (e.g. autoregressive) that may not be appropriate for microarray data. Others require uniform sampling points for all genes or fail to produce clusters when the number of time points are large (see Wang, Neill, and Miller, 2008 for full discussion). Using FDA techniques has circumvented many of these limitations and much of the work to date using FDA to analyze time-course gene expression data has focussed on cluster analysis. Some early examples include papers by Bar-Joseph, Gerber, Gifford, Jaakkola, and Simon (2002) and Luan and Li (2003) who simultaneously develop a model which represents the mean curve in each cluster, $\mu_c(t)$, using a linear combination of K cubic spline or B-spline basis functions respectively (as in (3)) and cluster using the EM algorithm. As stated in Section 2, it is difficult to choose the optimal value for K and choosing different values can alter the results. To overcome this problem Ma, Castillo-Davis, Zhong, and Liu (2006) model $\mu_c(t)$ using smoothing splines, incorporating a penalty to the optimization criterion as shown in (4) and cluster the expression profiles using the rejection-controlled EM (RCEM) algorithm. Ma and Zhong (2008) extend this to incorporate additional covariate effects into the clustering algorithm. Wang et al. (2008) propose an agglomerative clustering algorithm for functional data based on a new similarity measure and compare the results with many other clustering approaches such as k-means, self-organizing maps, smoothing spline-based clustering using `ssclust` (see Ma et al., 2006), Gaussian finite mixture model-based clustering using `mclust` (see Fraley and Raftery, 2002, Fraley and Raftery, 2006), etc. available in the R statistical programming environment.

Other authors have used functional principal components analysis to cluster time-course expression data. For example Song, Lee, Morris, and Kangd (2007) smooth the data, calculate the FPCs as outlined in Ramsay and Silverman (2005), Chapter 8 and cluster the vector of FPC scores using Gaussian finite mixture models and the `mclust` algorithm. An alternative approach is to cluster expression profiles based on the derivative of the curves. As stated in Section 4, use of the derivative has received little attention in bioinformatics literature to date. The derivative contains information about the shape (change pattern) of the expression profiles and since gene expression can be considered part of a biological system, it could be argued that using the derivative is sensible from a biological perspective. One way

of clustering based on the derivative involves smoothing the raw expression data, calculating the first derivative of the expression profiles, determining the FPCs of the derivative and clustering the resulting scores using an appropriate clustering algorithm. Déjean, Martin, Baccini, and Besse (2007) calculate the multivariate principal components of the matrix resulting from discretizing the first derivative of the smoothed expression profiles. The principal component scores are then clustered using a combination of k-means and hierarchical clustering algorithms. An alternative procedure is suggested by Kim and Kim (2008). These authors cluster change patterns of expression profiles by smoothing the data using a Fourier expansion and calculating the derivatives of the resulting curves using the Fourier coefficients. The Fourier coefficients of the derivative are then clustered using k-means and model-based clustering and the results compared. The authors report that using this method identifies clusters of co-expressed genes not identified by k-means clustering. It could be expected that the derivative may contain different information than the functions and therefore may result in different clustering outcomes. We carried out an initial investigation (results shown in Section 7.1) to examine any differences between clustering results obtained using the FPC scores of the functions versus clustering results obtained using the FPC scores of the derivatives. This analysis involved a subset of data from a well-known experiment examining gene expression during the lifecycle of *Drosophila melanogaster* (Arbeitman, Furlong, Imam, Johnson, Null, Baker, Krasnow, Scott, Davis, and White, 2002). The data consisted of expression levels of 399 cell-cycle regulated genes at 27 time points during the embryonic development phase. These data had no missing values. To form the function estimates the data were smoothed using (4) while derivative estimates were determined using (4) with the modified penalty as shown in (5). The smoothing parameter λ was chosen via generalized cross-validation. The FPCs of the functions and the FPCs of the derivatives were then determined and the corresponding FPC scores were clustered using `mclust`. The number of FPCs to retain was determined by examining a scree plot of the eigenvalues, the proportion of variation accounted for by the first R FPCs and plots of the FPCs. Based on the latter criteria, the first 3 FPCs of the functions were retained while the first 4 FPCs of the derivatives were retained. The clustering results using both approaches are shown in Section 7.1. Preliminary results indicate that clustering using the derivative appears to identify behavior patterns that are not recovered when using the functions. We are currently examining whether these differences are biologically significant and are carrying out further investigations into these results. We believe that use of the derivative may provide additional insight into gene expression dynamics and thus strengthen the argument for using FDA.

One problem encountered when clustering expression profiles is that processes can evolve at different rates and this can affect the clustering results. There-

fore it may be necessary to incorporate some time-warping or allow for gene-specific time-shifts when carrying out the clustering procedure. Removing time variation is called registration in FDA and time-synchronization has been incorporated into the analysis of gene expression data by several authors. Liu and Müller (2003) use an iterative mean-updating technique to synchronize time-scales across genes and thus identify clusters of co-expressed genes. Chudova, Hart, Mjolsness, and Smyth (2004) use a functional mixture model to model the mean expression profile using a specified (parametric) differential equation and allow for possible shifts in expression profiles through time. Clustering is performed using the EM algorithm. However the authors indicate that the method is likely to fail if the true dynamics does not match the restricted class of functions proposed by the differential equation specified. Luan and Li (2004) use the shape-invariant model to model the guide gene profiles (i.e. genes whose pattern of behavior is known) using B-splines. The authors aim to identify other genes following similar patterns as the guide genes and allow for phase and amplitude shifts to model the individual profiles. Leng and Müller (2006b) determine a time-shift characteristic for each gene which facilitates ordering the genes and thus determine groups of co-expressed genes according to time of expression. Smith, Vollrath, Bradfield, and Craven (2009) present a two-step procedure to determine clusters of genes with similar alignment and each cluster is aligned independently of other clusters. The method alternates between assigning genes to clusters and re-computing the alignment for each cluster based on the genes assigned to it. Research has shown (see Gaffney and Smyth, 2004) that performing alignment and clustering as a two-stage procedure can be sub-optimal and thus simultaneous curve alignment and clustering has been examined in Liu and Yang (2009a) and Tang and Müller (2009). The former combine clustering and alignment using a simple time translation model and mixture modeling for clustering and fit the model in a linear mixed effects model framework. The latter allow for more general time-shifts using time-warping functions and combine this with clustering to identify groups of genes with similar shape patterns where genes within a cluster can exhibit time variation. Tang and Müller (2009) use the `PACE` statistical package (Yao, Müller, Clifford, Dueker, Follett, Lin, Buchholz, and Vogel, 2003), freely available online at http://www.stat.ucdavis.edu/ wyang/PACE/index.html.

5.2 Differential Expression

Methods to determine differentially expressed genes have included empirical Bayes, two-way mixed effects ANOVA and hidden Markov models. Each of these techniques is a multivariate technique and thus have drawbacks when analyzing high-dimensional time-course data. For example it may be required that measurements

are made at the same time-points, tests are applied on a gene by gene basis and thus ignore data from all other genes or tests are applied at individual time points and not over the entire time interval. FDA provides a means to examine whether genes are differentially expressed over time (not a a single time point), facilitates non-uniform sampling, etc. Storey, Xiao, Leek, Tompkins, and Davis (2005) test for differential expression of genes over time and differential expression of genes across treatment groups. Determining if a particular gene is differentially expressed over time is achieved by modeling the mean expression profile using cubic splines and testing whether the mean expression profile is flat or not. To determine if genes are differentially expressed across treatment conditions, the mean expression profiles for each gene in each treatment condition are modeled using cubic splines and an F test is performed to test for differences between the mean profiles. The proposed methodology is implemented in the EDGE software package. Hong and Li (2006) use a functional hierarchical model and theory from empirical Bayes analysis techniques to determine differentially expressed genes. Expression profiles are modeled using B-spline basis functions and data from all genes is combined to determine posterior probabilities of differential expression for each gene. Liu and Yang (2009b) use FPCA to determine an estimate of each replicate's trajectory under the null (no differential expression) and alternative hypotheses (differential expression) respectively. The estimates are determined using the `PACE` package. Tests for differential expression are based on a modified F-statistic and a permutation test is performed to determine an appropriate p-value. Ma, Zhong, and Liu (2009) use functional ANOVA mixed-effects models to identify genes that are differentially expressed across several treatment conditions, either with or without a time by condition interaction. The presence of a significant time by condition interaction for a particular gene indicates that the gene is a non-parallel differentially expressed gene and the authors propose a generalized F test to determine if such an interaction exists. If a significant interaction does not exist, the authors then test if the genes are parallel differentially expressed using a modification of the proposed F test. In each case using the FDA approach has been shown to be more powerful than multivariate methods including two-way mixed effects ANOVA, empirical Bayes, etc. and thus has clear advantages when analyzing time-course gene expression data.

5.3 Functional Linear Models

Functional logistic regression models have been used to perform discriminant analysis of two groups of time-course microarray data. This involves solving (11), where $y_i = 1$ if the group is group 1 and $y_i = 0$ if the group is group 2. Araki,

Konishi, and Imoto (2004) use radial basis functions and the roughness penalty approach to determine $\beta(t)$, while Leng and Müller (2006a) and Müller (2005) use the principal components approach outlined in Section 4 as implemented in the `PACE` package. Leng and Müller (2006a) report that using the principal components method results in lower overall classification error rates, while Müller (2005) uses shrinkage estimators of the principal components scores, useful when the data collected for some expression profiles is sparse. Parker and Wen (2009) use FPCA as an exploratory data analysis tool to examine differences in behavior patterns between direct target and indirect target genes. The first derivative of the smooth expression profiles is then determined and used as the predictor variable with a Boolean (i.e. 0/1) response variable for classification. Other examples of using functional regression analysis techniques to examine time-course microarray data can be found in Wang, Chen, and Li (2007) who use model (12) and the basis function approach to determine the transcriptional factors involved in gene regulation during a biological process and model their effect on gene expression levels during that process and Müller, Chiou, and Leng (2008) who use model (14) and the principal components approach (and the `PACE` package) to model the relationship between temporal gene expression profiles of different developmental stages of *Drosophila melanogaster*. Müller and Yao (2008) use (20) to examine the relationship between expression profiles from two developmental stages of *Drosophila melanogaster*. Again, FDA methodologies were found to be extremely useful and Parker and Wen (2009) state that *FDA can provide very good discrimination, substantially better than a standard multivariate analysis*, while Müller et al. (2008) report that *functional regression emerges as a useful tool for relating gene expression patterns from different developmental stages, and avoids the problems with large numbers of parameters and multiple testing that affect alternative approaches*. An application of functional regression analysis to time-course microarray data can be found in Section 7.2.

6 Further Research

The previous sections have outlined the use of several FDA techniques in time-course microarray analyses and have shown the merits of FDA in these contexts. As can be seen above, these applications have primarily focussed on FPCA and functional linear regression. Other FDA procedures may also be of interest in bioinformatics and some of these methods are described in the following section. To our knowledge none of the concepts outlined here have been applied to time-course microarray data to date and are thus interesting areas for future research.

6.1 Functional t-Tests and F-Tests

When modeling functional data we may be interested in questions such as: Is there a statistically significant difference between two groups of curves (e.g. differential expression)? Are there statistically significant relationships among functional random variables? (see Ramsay et al., 2009) Functional t-tests can be used to test for significant differences between groups of curves. The functional version of the t-test statistic has the form

$$T(t) = \frac{|\bar{x}_1(t) - \bar{x}_2(t)|}{\sqrt{\frac{1}{n_1}\text{Var}[x_1(t)] + \frac{1}{n_2}\text{Var}[x_2(t)]}}, \tag{22}$$

where $\bar{x}_1(t)$ and $\bar{x}_2(t)$ denote the mean in group 1 and group 2 respectively, n_1 and n_2 denote the sample sizes in group 1 and group 2 respectively and $\text{Var}[x_1(t)]$ and $\text{Var}[x_2(t)]$ denote the variance in each group. Since functional data are high dimensional, permutation tests are used to determine a null distribution for $T(t)$. This involves randomly permuting the labels on the expression profiles and calculating the maximum of $T(t)$. This is repeated hundreds of times to obtain the null distribution. An example is provided in Section 7.3.

Functional F-tests can be used to determine if significant relationships exist between the functional random variables. The F-test statistic is given by

$$F(t) = \frac{\text{Var}[\hat{\mathbf{y}}(t)]}{\frac{1}{n}\sum_{i=1}^{N}[y_i(t) - \hat{y}_i(t)]^2}, \tag{23}$$

where $\hat{\mathbf{y}}$ are the predicted values from fitting a functional linear model. Again, a null distribution is determined using a permutation test where the response curves (or values) are shuffled and the maximum of $F(t)$ is calculated. We believe that utilizing such tests may provide alternative approaches to testing for differential expression and the significance of relationships between expression profiles and other covariate(s).

6.2 Principal Differential Analysis

There has been limited work focussing on using the derivative in microarray analysis though there has been some work involving derivatives when clustering and discriminating between groups of expression profiles. Functional linear models can be used to model derivatives, i.e. model $Dy_i(t)$, the rate of change of $y_i(t)$ instead

of $y_i(t)$. Such models are called dynamic models and are essentially differential equations. A first-order homogeneous differential equation has the form

$$Dy_i(t) = -y_i(t)\beta(t) + \varepsilon_i(t) \tag{24}$$

while a first-order non-homogeneous equation can be written

$$Dy_i(t) = -y_i(t)\beta_0(t) - \alpha(t)\beta_1(t) + \varepsilon_i(t), \tag{25}$$

where $\alpha(t)$ is called a forcing function and refers to external inputs into the system. It is also possible to model higher order derivatives using higher-order differential equations (see Ramsay and Silverman, 2005, Chapters 17-19). These models are used to model the dynamics of a system and have the advantage of using the function itself as a covariate. Utilizing dynamic models is called principal differential analysis (PDA) in functional data analysis literature and could be highly useful in the analysis of temporal gene expression data since gene expression can be thought of as being part of a biological system. These models could be used to describe the relationship between the rate of change of expression of one gene and the level of expression of other genes in the system. This could provide real insight into how genes are related. See Section 7.4 for a simple example.

6.3 Functional Canonical Correlation Analysis

Functional canonical correlation analysis (fCCA) deals with the case where pairs of functions $(x_i(t), y_i(t))$ are observed for each individual. For example, say we have expression profiles for genes in the embryo phase of development and expression profiles for the same genes in the adult phase of development of *Drosophila melanogaster* (fruit fly). fCCA aims to determine the modes of variability in the two sets of curves that are most associated with one another, i.e. find a pair of functions $(\xi(t), \eta(t))$ such that the canonical variates

$$\rho_{\xi_i} = \int_{\mathcal{T}} \xi(t)[x_i(t) - \bar{x}(t)]\mathrm{d}t \tag{26}$$

and

$$\rho_{\eta_i} = \int_{\mathcal{T}} \eta(t)[y_i(t) - \bar{y}(t)]\mathrm{d}t \tag{27}$$

are highly correlated with each other. $\xi(t)$ and $\eta(t)$ are the modes of variation that account for most of the interaction between expression profiles in the embryo phase and expression profiles in the adult phase (Ramsay and Silverman, 2005). This is related to functional linear regression, though this provides a symmetric view on

the relationship since it does not require one variable to be specified as the response variable and the other as the predictor variable (Ramsay et al., 2009). To determine $\xi(t)$ and $\eta(t)$ it is necessary to optimize the canonical correlation criterion

$$\texttt{ccorsq}(\xi, \eta) = \frac{[\texttt{cov}(\rho_{\xi_i}, \rho_{\eta_i})]^2}{(\texttt{var}\rho_{\xi_i})(\texttt{var}\rho_{\eta_i})}, \tag{28}$$

where ρ_{ξ_i} and ρ_{η_i} are as defined in (26) and (27) respectively. Maximizing ccorsq as given above may not yield interpretable results due to lack of smoothness of the estimates of $\xi(t)$ and $\eta(t)$. Therefore there is a need to incorporate smoothing into the optimization problem. This is achieved by penalizing the curvature of the canonical variate weight functions as shown in Section 2. This yields the smoothed canonical correlation criterion

$$\texttt{ccorsq}_\lambda(\xi, \eta) = \frac{[\texttt{cov}(\rho_{\xi_i}, \rho_{\eta_i})]^2}{[\texttt{var}(\rho_{\xi_i}) + \lambda\|D^2\xi\|^2][\texttt{var}(\rho_{\eta_i}) + \lambda\|D^2\eta\|^2]}, \tag{29}$$

where λ can be chosen via cross-validation or generalized cross-validation. As with FPCA, a series of canonical variate weight functions (ξ_1, η_1), (ξ_2, η_2), $(\xi_3, \eta_3), \ldots$ can be determined by optimizing $\texttt{ccorsq}_\lambda$ subject to the constraint that successive weight functions are orthogonal. Ramsay and Silverman (2005), Chapter 11 and Ramsay et al. (2009), Chapter 7 give full implementation details for fCCA. fCCA may prove to be a useful tool in microarray analyses since it can identify variations in behavior patterns not evident from other techniques. For example, fCCA can examine how expression levels vary from the embryo phase to the adult phase of *Drosophila melanogaster* and establish how changes in one phase influence (or are influenced by) changes in the other phase.

7 Examples

7.1 Clustering Using FPCA

The estimated functions and corresponding derivative estimates for the 399 *Drosophila* genes are displayed on the LHS and RHS of Figure 1 respectively. The LHS of Figure 2 displays the scree plot of the eigenvalues of the FPCs estimated using the original expression curves. This plot indicates that the first 3 FPCs should be retained. These FPCs are shown on RHS of Figure 2 and account for over 94% of the variation in the data. The first 4 FPCs estimated using the first derivative of the expression profiles as displayed in Figure 3 account for 84% of the variation in the data. Though the scree plot indicate that FPC 5 could also be included in the

analysis, the plot of FPC 5 shows that this component contains lots of oscillations and therefore is most likely attributable to noise. The vectors of scores for these 3 (4) FPCs were then supplied to `mclust`. The resulting clusters are shown in Figures 4 and 6. The first derivative of the members of each cluster are given in Figures 5 and 7.

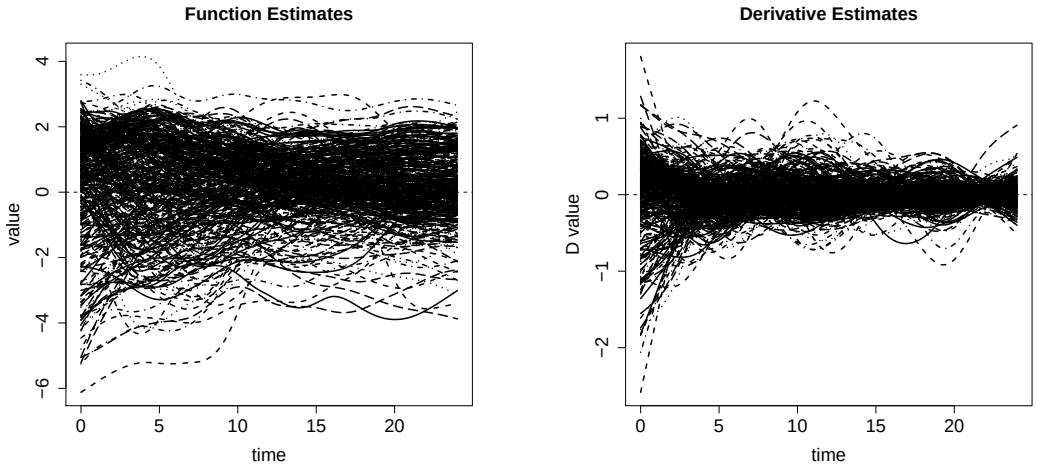

Figure 1: LHS: Estimated gene expression curves. RHS: Estimated first derivative of gene expression curves.

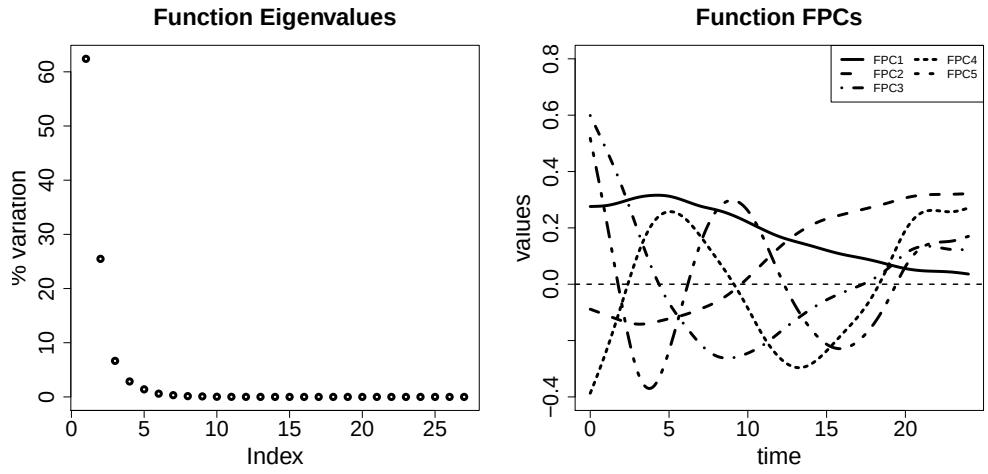

Figure 2: LHS displays the eigenvalues obtained using the original expression profiles. RHS displays the first 3 FPCs.

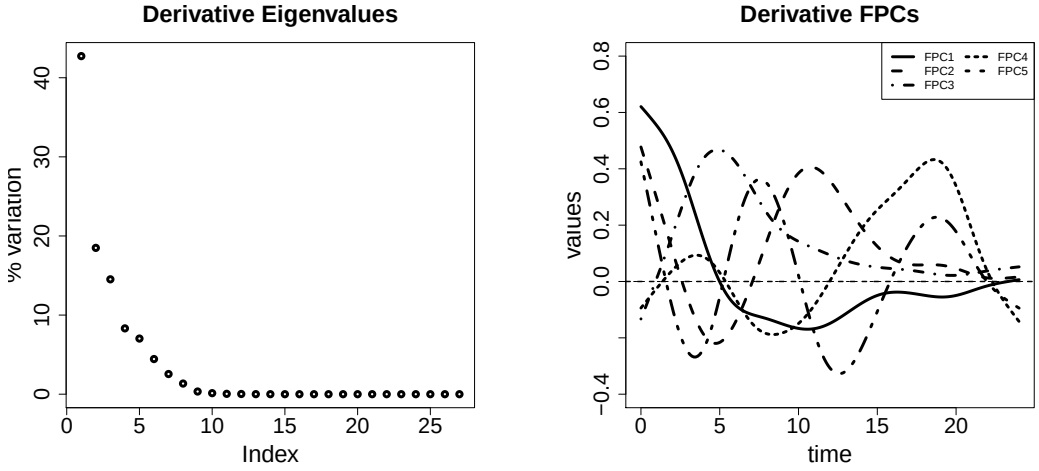

Figure 3: LHS displays the eigenvalues obtained using the first derivative of the expression profiles. RHS displays the first 4 FPCs.

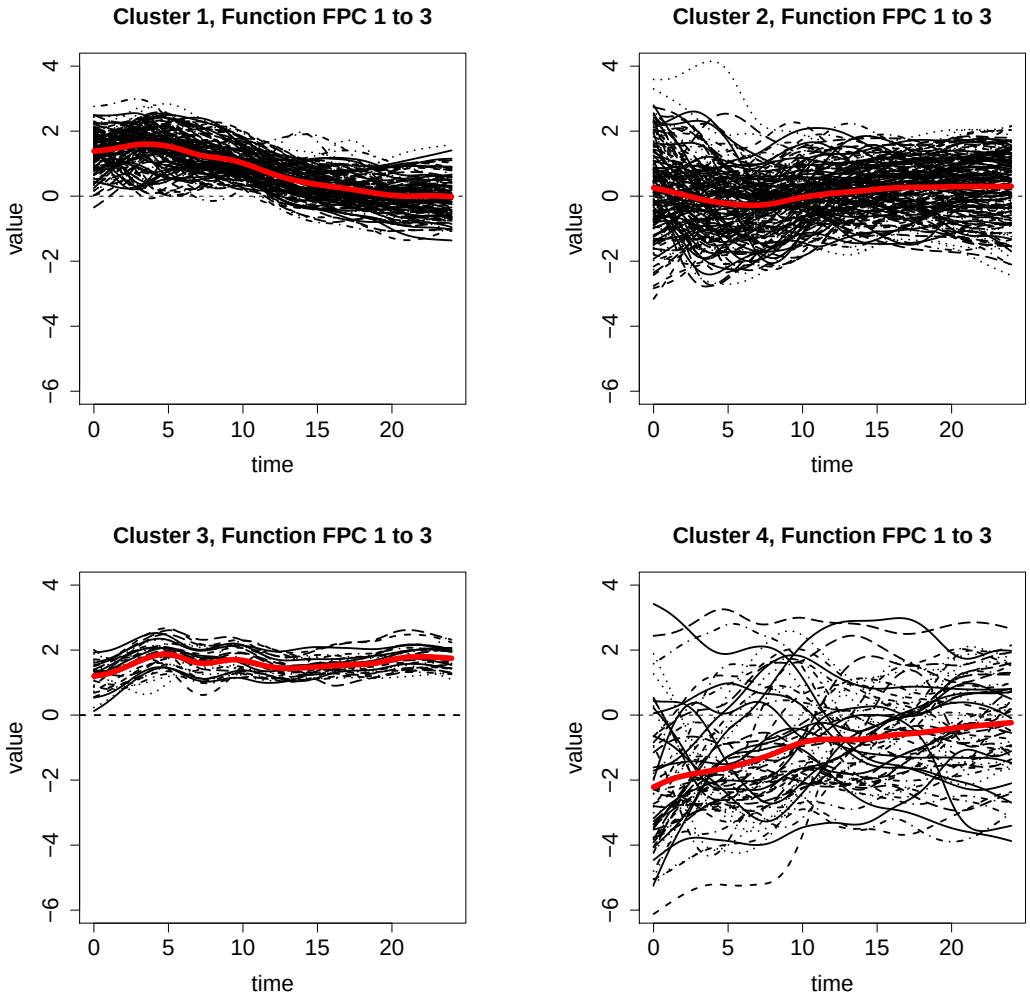

Figure 4: Clusters determined using the scores of the first 3 FPCs extracted using the original expression profiles.

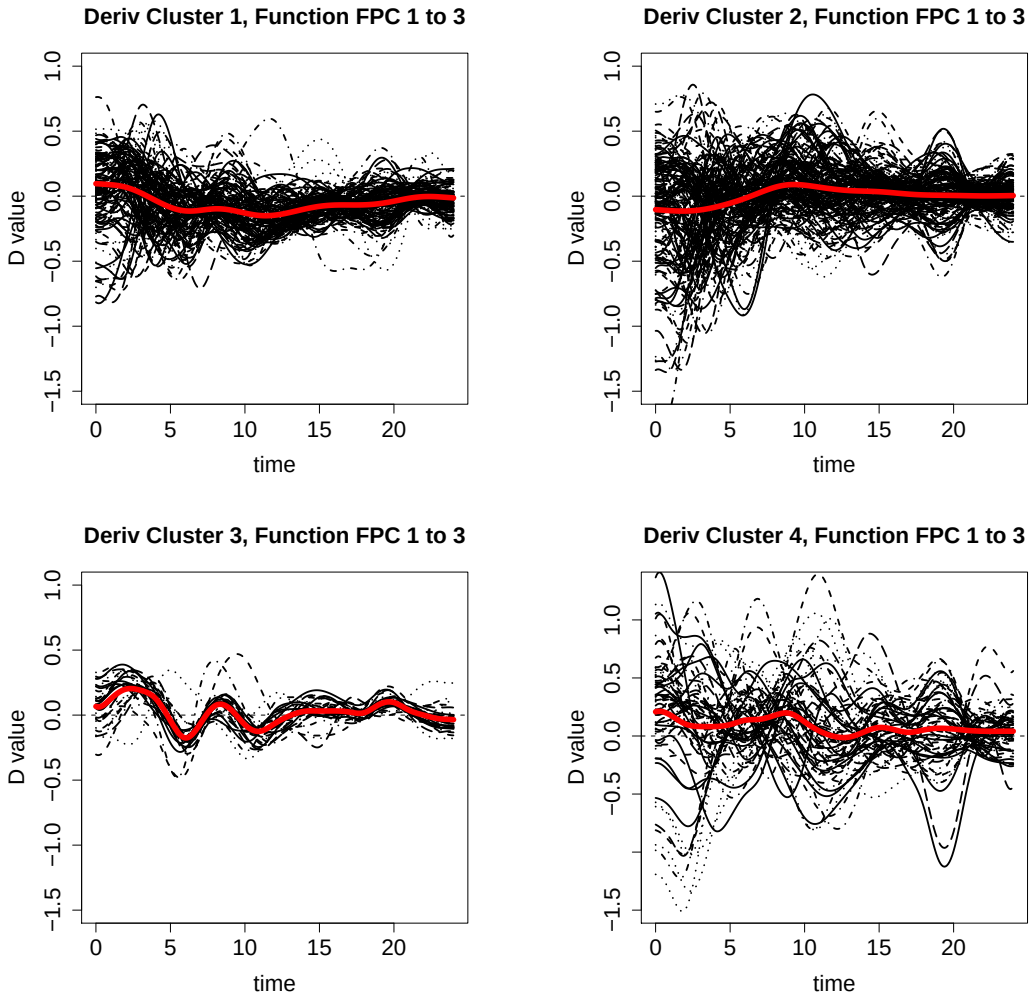

Figure 5: Derivatives of the clusters determined using the scores of the first 3 FPCs extracted using the original expression profiles.

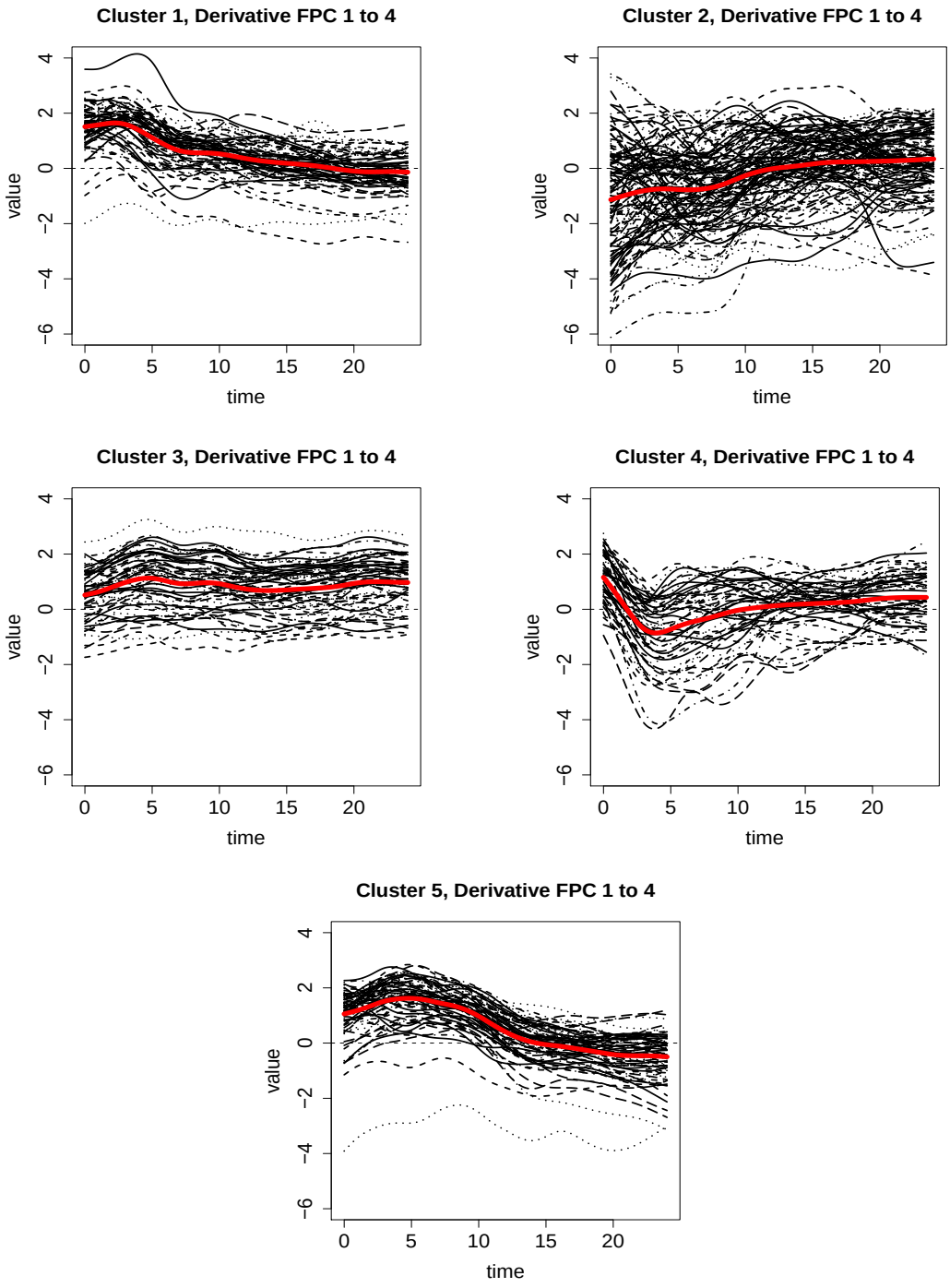

Figure 6: Clusters determined using the scores of the first 4 FPCs extracted using the first derivative of the expression profiles.

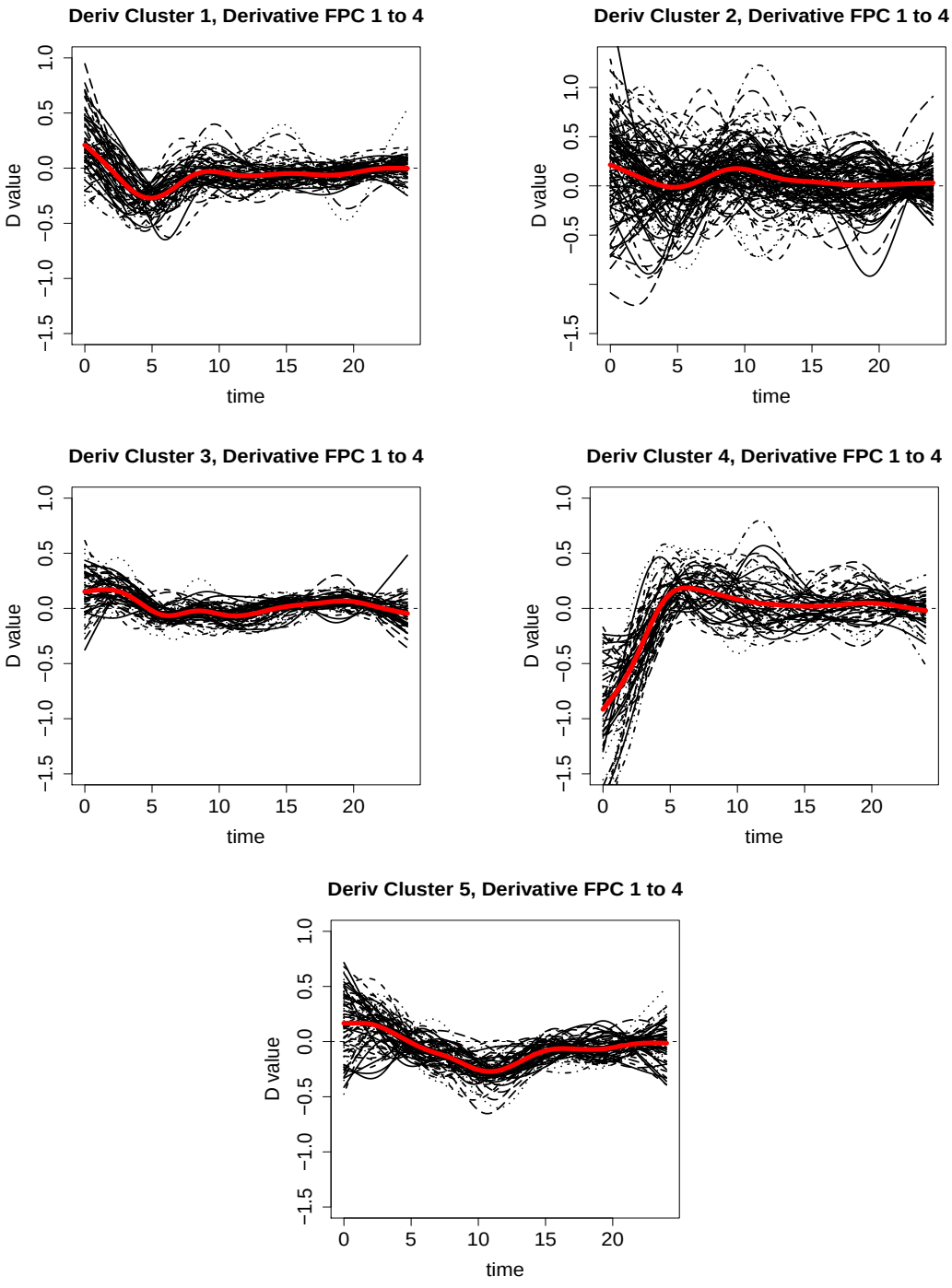

Figure 7: Derivatives of the clusters determined using the scores of the first 4 FPCs extracted using the first derivative of the expression profiles.

	Deriv FPC Class				
Func FPC Class	**1**	**2**	**3**	**4**	**5**
1	47	7	14	14	57
2	22	78	23	36	12
3	1	1	25	1	0
4	2	47	1	6	5

Table 1: Classification table

Table 1 displays a comparison of the classifications for each method. Cluster 1 obtained using the original expression profiles mainly consists of a combination of observations from Clusters 1 and 5 obtained using the first derivative of the expression profiles. The additional separation of these observations into two clusters as achieved using the derivatives appears to differentiate between genes whose expression levels peak almost immediately (typically before time 2) before decreasing for the remainder of the cycle (Cluster 1) and genes with expression levels that exhibit a more gradual increase in expression levels (to a peak value at approximately time 5) before decreasing for the remainder of the cycle (Cluster 5). Cluster 3 for both methods identifies genes whose expression levels change little throughout the cycle though using the derivatives has identified 35 more genes belonging to this cluster. Clusters 2 and 4 determined using the functions contain 232 genes while Clusters 2 and 4 determined using the derivatives contain 190 genes. Cluster 4 is the main difference between both methods. Using the derivative has identified genes exhibiting a rapid decrease in expression levels in the initial stage of the cycle before gradually increasing for the remainder of the cycle. Such a cluster is not evident when using the original expression profiles.

Figures 5 and 7 provide a means of examining the derivatives of the observations in the clusters obtained for each method. The derivatives of the observations in the clusters obtained using the original expression profiles as shown in Figure 5 exhibit lots of oscillations and there is a large amount of variation exhibited in each cluster. Therefore it is difficult to determine any distinct shapes or patterns of behavior in the clusters. In contrast, examining the derivative estimates of the observation in the clusters obtained using the first derivative of the expression profiles as shown in Figure 7 reveals clear patterns of behavior in each cluster. There is reduced variation in the derivative estimates within each cluster suggesting that clustering based on the FPC scores of the derivative determines more homogeneous clusters than clustering based on the FPC scores of the original functions. The homogeneous clustering of the derivatives translates back into homogeneous clusters

in the original data. As a result it is clear that the derivative contains additional information to that given by the functions and use of the derivative may prove a highly useful tool in the analysis of time-course microarray gene expression data.

7.2 Functional Regression Analysis

The following describes an application of functional regression analysis to a subset of the *Drosophila Melanogaster* dataset analyzed by Arbeitman et al. (2002). The authors identified a group of "strictly maternal" genes or genes that were expressed in the embryo phase and then re-expressed in the pupal-adult phase of female flies. Therefore we wish to model the relationship between expression levels in the pupal-adult phase, $y_i(t)$, and expression levels in the embryo phase, $x_i(t)$. In this instance both the response and the predictor variables are functions as shown in Figure 8.

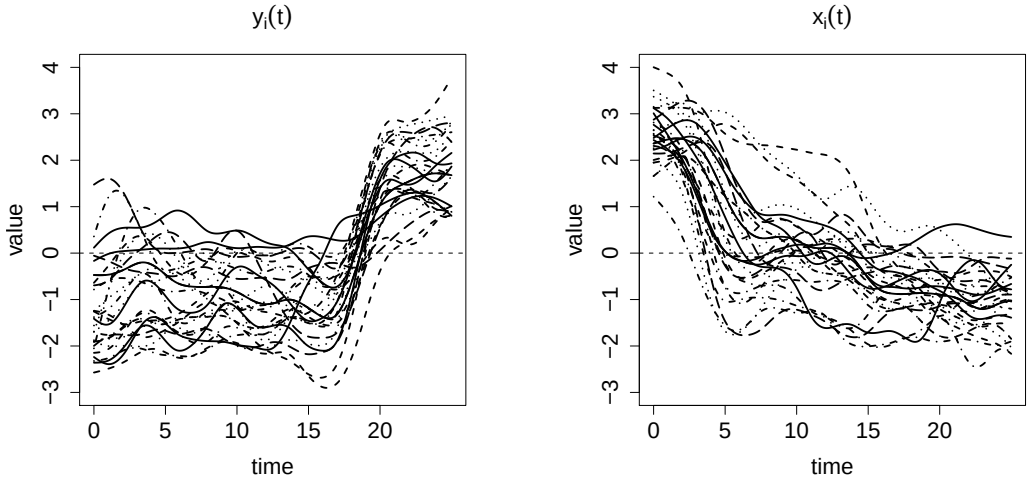

Figure 8: LHS displays the smooth expression profiles in the pupal-adult phase $y_i(t)$. RHS displays the smooth expression profiles in the embryo phase $x_i(t)$.

It is assumed that there is a functional relationship of the form

$$y_i(t) = \alpha(t) + \int_{\mathcal{T}} x_i(s)\beta(s,t)\mathrm{d}s + \varepsilon_i(t), \tag{30}$$

where $\alpha(t)$ is the intercept function and $\beta(s,t)$ is the regression surface to be estimated as in (14). It is assumed that $y_i(t)$, $x_i(t)$ and $\beta(s,t)$ can be represented as a linear combination of basis functions as shown in (15). We chose a large number of B-spline basis functions and accounted for over-fitting using the regularization approach in (16). However since the response is also a function in this example, the criterion to be optimized is now given by

$$\text{PENSSE} = \int_{\mathcal{T}} \sum_{i=1}^{N} [y_i(t) - \alpha(t) - \int_{\mathcal{T}} x_i(s)\beta(s,t)\mathrm{d}s]^2 \mathrm{d}t + \text{PEN}_{\lambda_1,\lambda_2}[\beta(s,t)], \tag{31}$$

where

$$\text{PEN}_{\lambda_1,\lambda_2}[\beta(s,t)] = \lambda_1 \int_{\mathcal{T}} [D^2\beta(s,t)]^2 \mathrm{d}t\mathrm{d}s + \lambda_2 \int_{\mathcal{T}} [D^2\beta(s,t)]^2 \mathrm{d}s\mathrm{d}t, \tag{32}$$

D^2 denotes the second derivative and two penalty terms and associated smoothing parameters (λ_1 and λ_2) are now required, one to account for over-fitting in the s direction and the other to account for over-fitting in the t direction. The above model can be fitted using the `fda` package in `R`.

The estimated regression surface for the strictly maternal genes is shown in Figure 9. Genes with lower expression levels early in the pupal-adult stage have higher embryo expression over time. However, this effect is reversed for the later stages of pupal-adult expression where genes with higher expression levels at this stage have lower embryo expression over time. As stated in Section 5.3, these data are also analyzed in Müller et al. (2008) though these authors use the principal components approach to model the response, predictor and regression functions. It can be seen that both analyses yield similar results, with the same interpretation for $\beta(s,t)$.

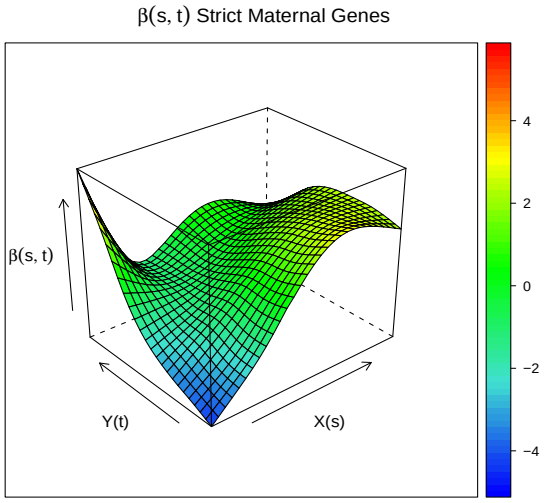

Figure 9: Estimated regression surface $\beta(s,t)$.

7.3 Functional t-Test

The previous section examines the relationship between expression levels in the embryo stage and pupal-adult stage of development of female flies. However, the expression profiles of male flies were also recorded in the pupal-adult stage by Arbeitman et al. (2002). Figure 10 suggests that female flies (black curves) have higher expression levels of strictly maternal genes than male flies (red curves) in the pupal-adult phase.

To test for differential expression between male and female flies, a functional t-test can be performed where the test statistic has the form:

$$T(t) = \frac{|\bar{x}_F(t) - \bar{x}_M(t)|}{\sqrt{\frac{1}{n_F}\text{Var}[x_F(t)] + \frac{1}{n_M}\text{Var}[x_M(t)]}}. \tag{33}$$

Critical values for $T(t)$ are determined using a permutation test by randomly shuffling the male (M) and female (F) labels on the curves and calculating the maximum of $T(t)$ using these new labels. This is repeated many times and a null distribution is constructed. Figure 11 displays the observed test statistic $T(t)$ and corresponding critical values and indicates that there is no significant difference between the expression levels of male and female flies until just after the onset of adulthood where expression levels of these genes in females is significantly higher than males.

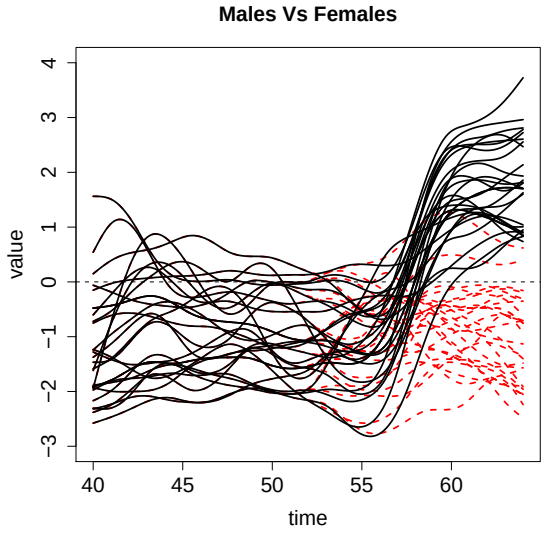

Figure 10: Black lines are the expression profiles of female flies. Red lines are the expression profiles of male flies.

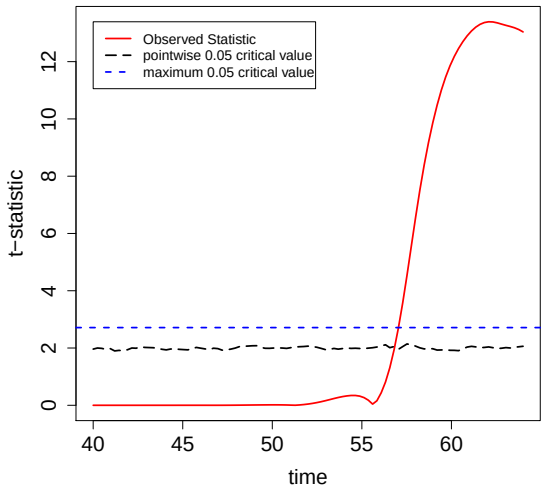

Figure 11: Permutation test for the difference between male and female flies. The blue dashed line gives the permutation 0.05 critical value for the maximum of $T(t)$. The black dotted line gives the permutation critical value for the pointwise statistic.

7.4 Principal Differential Analysis

A major advantage of FDA over multivariate techniques is that it facilitates the use of derivative information in the curves. This may be particularly useful in time-course microarray data analyses since gene expression is part of a biological system and such systems are typically modeled using differential equations. As stated in Ramsay and Silverman (2005), differential equations can describe the features in both the curve and its derivative for a single functional datum but can also describe the variation across observations in a sample of N curves. Principal differential analysis as outlined in Section 6, is a FDA technique which aims to satisfy one of the latter criteria. To demonstrate a simple application of PDA in a time-course microarray setting, the dynamics of expression profiles in the pupal-adult stage of development for the strictly maternal genes of female flies are modeled. It should be noted that the following results are tentative and do not constitute an indepth PDA of these data.

The data used in this example are plotted on the LHS of Figure 8. Let $y_i(t)$ denote the pupal-adult expression profiles and initially assume that a homogeneous first order differential equation is appropriate such that

$$Dy_i(t) = -\beta(t)y_i(t) + \varepsilon_i(t), \tag{34}$$

where $\beta(t)$ is a coefficient function to be estimated. To determine how well this model fits the data, the residual functions are plotted and are displayed on the LHS of Figure 12. It is clear that the first order equation is not fitting the data well between times 15 and 20 as the residual functions show a clear pattern in this interval. Therefore, a second order homogeneous equation is used such that

$$D^2y_i(t) = -\beta_1(t)Dy_i(t) - \beta_0(t)y_i(t) + \varepsilon_i(t), \tag{35}$$

where $D^2y_i(t)$ represents the second derivative of $y_i(t)$ and $\beta_0(t)$ and $\beta_1(t)$ are coefficient functions to be estimated. Each term in the model represents a different force on the system. The first term in this model is proportional to the speed at which the system moves, while the second term position-dependent forces. See Ramsay and Silverman (2002), Ramsay and Silverman (2005) and Ramsay et al. (2009) for additional details. The residuals from this model are displayed on the RHS of Figure 12 and appear to be more satisfactory than those obtained using a first order equation.

It is then necessary to interpret the coefficient functions $\beta_0(t)$ and $\beta_1(t)$. This can be difficult and therefore the discriminant function given by

$$d(t) = \frac{\beta_1(t)^2}{4} - \beta_0(t) \tag{36}$$

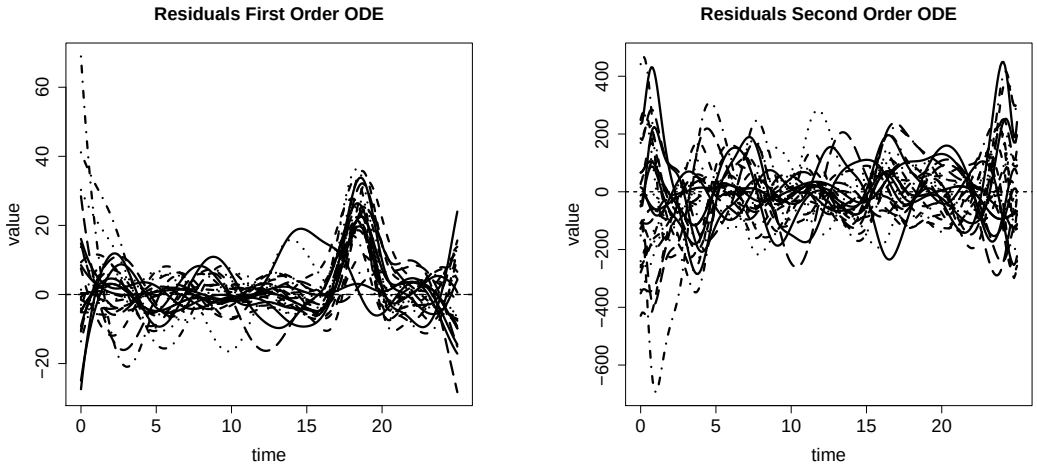

Figure 12: LHS displays residuals of first order equation. RHS displays residuals of second order equation.

is used where the sign of this function is important. When $d(t)$ is negative, $\beta_1(t)$ is small and the system tends to exhibit some oscillation that gradually disappears. When $d(t)$ is positive, $\beta_1(t)$ is relatively large and either becomes stable so quickly that no oscillation is observed ($\beta_1(t) > 0$) or oscillates out of control ($\beta_1(t) < 0$). When $d(t) = 0$ the system is in linear motion. Figure 13 displays $\beta_1(t)$ and $d(t)$ for the pupal-adult data. $d(t)$ is is initially negative which corresponds to an initial increase in energy marking the beginning of gene expression, followed by a period when $d(t) = 0$ and the system is in equilibrium. During this time, expression levels are relatively stable. At approximately time 15 (i.e. the onset of adulthood) $d(t)$ quickly becomes negative. At this point the system is exhibiting some oscillatory behavior corresponding to an increase in energy in the system prior to the large jump in expression levels between times 16 and 21. However after this initial burst the change in expression levels is quite stable. From time 21 onwards $d(t)$ is positive and $\beta_1(t)$ is negative implying that after the sharp increase in expression levels between times 16 and 21, the system contains a vary large amount of energy and is behaving like a rapidly oscillating spring. From this simple example it can be seen that PDA gives real insight into the dynamics of expression profiles in this group of genes. We believe that PDA may be particularly useful when examining differences between the expression levels of genes across two groups, e.g. across two treatment conditions, where information about how the behavior of the derivatives differ across groups may give additional information regarding why expression levels differ. All of the above analysis was carried out using the `fda` package in `R`.

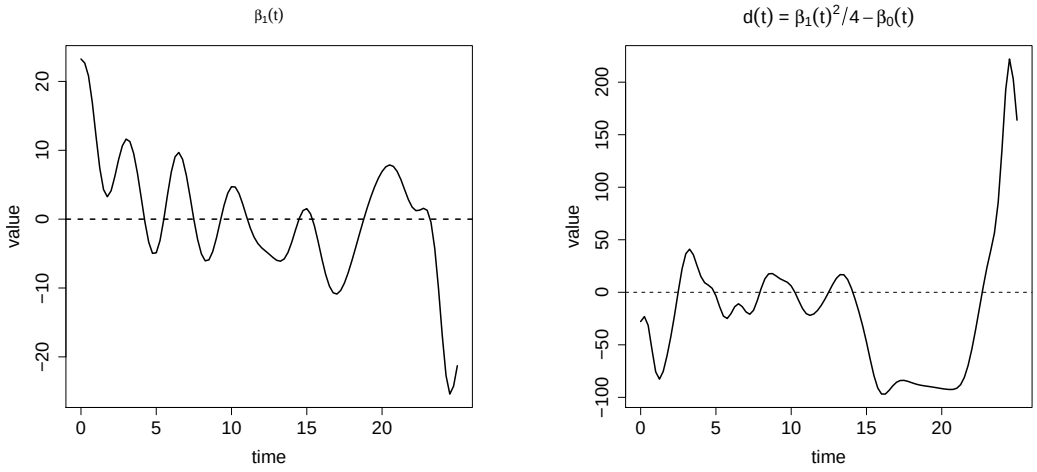

Figure 13: LHS displays the estimate of $\beta_1(t)$. RHS displays the discriminant function $d(t)$.

8 Conclusion

This paper has presented a review of the applications of functional data analysis to time-course microarray experiments. There are a number of issues encountered with such experiments: the high dimensionality of expression profiles and the small sample sizes, correlation between observations on the same gene, unequal sampling of data, missing data, large levels of measurement error, etc. FDA deals with each of these issues and as demonstrated in this article, provides numerous advantages over multivariate methods that have traditionally been popular choices. Though FDA was originally introduced by Ramsay and Dalzell (1991), applications to microarray data have only begun to appear in the last decade. There has been a large increase in the number of papers utilizing FDA in a bioinformatics setting and we believe that as microarray experiments produce larger volumes of data, FDA methods will become increasingly required. There has already been an expansion of FDA into other areas of interest in gene expression analysis other than clustering, discrimination and linear modeling as presented here. For example, Opgen-Rhein and Strimmer (2006) use FDA to identify genetic networks using Gaussian graphical models and again suggest that adopting the FDA approach is advantageous since it allows for the identification of the dependency across the whole time interval rather than at single points in time. We also believe that other FDA techniques

such as functional t-tests and F-tests, principal differential analysis and functional canonical correlation analysis will prove extremely useful in a bioinformatics setting and provide additional insight into gene expression dynamics.

In conclusion, it should be noted that though this paper aimed to provide as thorough a review as possible of FDA procedures in bioinformatics literature, there may be some work that we are not aware of and thus have not included. We therefore apologize to the authors of papers that were not cited in this article.

References

Araki, Y., S. Konishi, and S. Imoto (2004): "Functional discriminant analysis for microarray gene expression data via radial basis function networks," in *Proceedings of COMPSTAT Symposium*, 613–620.

Arbeitman, M. N., E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White (2002): "Gene expression during the life cycle of *drosophila melanogaster*," *Science*, 297, 2270–2275.

Bar-Joseph, Z., G. Gerber, D. K. Gifford, T. S. Jaakkola, and I. Simon (2002): "A new approach to analyzing gene expression time series data," in *Proceedings of RECOMB, Washington DC, USA*.

Chudova, D., C. Hart, E. Mjolsness, and P. Smyth (2004): "Gene expression clustering with functional mixture models," *Advances in Neural Information Processing Systems*, 16, 683–692.

Déjean, S., P. G. P. Martin, A. Baccini, and P. Besse (2007): "Clustering time-series gene expression data using smoothing spline derivatives," *EURASIP Journal on Bioinformatics and Systems Biology*.

Fraley, C. and A. E. Raftery (2002): "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Fraley, C. and A. E. Raftery (2006): "MCLUST version 3 for R: Normal mixture modeling and model-based clustering," Technical Report 504, University of Washington, Department of Statistics, (revised 2009).

Gaffney, S. and P. Smyth (2004): "Joint probabilistic curve clustering and alignment," *Advances in Neural Information Processing Systems*.

Hong, F. and H. Li (2006): "Functional hierarchical models for identifying genes with different time-course expression profiles," *Biometrics*, 62, 534–544.

Kim, J. and H. Kim (2008): "Clustering of change patterns using fourier coefficients," *Bioinformatics*, 24, 184–191.

Leng, X. and H.-G. Müller (2006a): "Classification using functional data analysis for temporal gene expression data," *Bioinformatics*, 22, 68–76.

Leng, X. and H.-G. Müller (2006b): "Time ordering of gene co-expression," *Biostatistics*, 7, 569–584.

Liu, X. and H.-G. Müller (2003): "Modes and clustering for time-warped gene expression profile data," *Bioinformatics*, 19, 1937–1944.

Liu, X. and M. C. K. Yang (2009a): "Simultaneous curve registration and clustering for functional data," *Computational Statistics and Data Analysis*, 53, 1361–1376.

Liu, X. and M. C. K. Yang (2009b): "Identifying temporally differentially expressed genes through functional principal components analysis," *Biostatistics*, 10, 667–679.

Luan, Y. and H. Li (2003): "Clustering of time-course gene expression data using a mixed-effects model with b-splines," *Bioinformatics*, 19, 474–482.

Luan, Y. and H. Li (2004): "Model-based methods for identifying periodically expressed genes based on time-course microarray gene expression data," *Bioinformatics*, 20, 332–339.

Ma, P., C. I. Castillo-Davis, W. Zhong, and J. S. Liu (2006): "A data-driven clustering method for time course gene expression data," *Nucleic Acids Research*, 34, 1261–1269.

Ma, P. and W. Zhong (2008): "Penalized clustering of large-scale functional data with multiple covariates," *Journal of the American Statistical Association*, 103, 625–636.

Ma, P., W. Zhong, and J. S. Liu (2009): "Identifying differentially expressed genes in time-course microarray data," *Statistics in Biosciences*, 1, 144–159.

Müller, H.-G. (2005): "Functional modelling and classification of longitudinal data," *Scandinavian Journal of Statistics*, 32, 223–240.

Müller, H.-G., J.-M. Chiou, and X. Leng (2008): "Inferring gene expression dynamics via functional regression analysis," *BMC Bioinformatics*, 9.

Müller, H.-G. and F. Yao (2008): "Functional additive models," *Journal of the American Statistical Association*, 103, 426–437.

Opgen-Rhein, R. and K. Strimmer (2006): "Inferring gene dependency networks from genomics longitudinal data: a functional data approach," *REVSTAT*, 4, 53–65.

Parker, B. J. and J. Wen (2009): "Predicting microrna targets in time-series microarray experiments via functional data analysis," *BMC Bioinformatics*, 10.

Ramsay, J. O. and C. J. Dalzell (1991): "Some tools for functional data analysis," *Journal of the Royal Statistical Society, Series B*, 53, 539–572.

Ramsay, J. O., G. Hooker, and S. Graves (2009): *Functional Data Analysis with R and MATLAB*, New York: Springer-Verlag.

Ramsay, J. O. and B. W. Silverman (2002): *Applied Functional Data Analysis*, USA: Springer-Verlag.

Ramsay, J. O. and B. W. Silverman (2005): *Functional Data Analysis*, USA: Springer-Verlag.

Smith, A. A., A. Vollrath, C. A. Bradfield, and M. Craven (2009): "Clustered alignments of gene-expression time series data," *Bioinformatics*, 25, 119–127.

Song, J. J., H.-J. Lee, J. S. Morris, and S. Kangd (2007): "Clustering of time-course gene expression data using functional data analysis," *Computational Biology and Chemistry*, 31, 265–274.

Storey, J. D., W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis (2005): "Significance analysis of time course microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 12837–12842.

Tang, R. and H.-G. Müller (2009): "Time-synchronized clustering of gene expression trajectories," *Biostatistics*, 10, 32–45.

Wang, H., J. Neill, and F. Miller (2008): "Nonparametric clustering of functional data," *Statistics and Its Interface*, 1, 47–62.

Wang, L., G. Chen, and H. Li (2007): "Group scad regression analysis for microarray time course gene expression data," *Bioinformatics*, 23, 1486–1494.

Yao, F., H.-G. Müller, A. J. Clifford, S. R. Dueker, J. Follett, Y. Lin, B. A. Buchholz, and J. S. Vogel (2003): "Shrinkage estimation for functional principal component scores, with application to the population kinetics of plasma folate," *Biometrics*, 59, 676–685.