STATISTICA, anno LXIX, n. 2–3, 2009

FUNCTIONAL MODELLING OF MICROARRAY TIME SERIES WITH COVARIATE CURVES

Maurice Berk
Department of Mathematics, Imperial College London
Giovanni Montana
Department of Mathematics, Imperial College London

1. INTRODUCTION

Biological systems are inherently dynamic and gene expression levels may be temporally regulated for a wide range of reasons including the cell cycle, circadian rythms, developmental processes or in response to stimuli (e.g. drug treatment or environmental stress) (Spellman *et al.*, 1998; Wang and Kim, 2003; Calvano *et al.*, 2005). Microarrays are a high throughput assaying technique for measuring these expression levels of thousands of genes simultaneously. Each microarray hybridisation provides a snapshot of expression levels at a single point in time; by carrying out sequential hybridisations on biological samples arising from the same source (e.g. a human patient), the evolution of these expression levels over time can then be elucidated.

The resulting microarray time series give rise to data that possess certain characteristics which make their analysis particularly challenging. Specifically, due to the large number of genes under study simultaneously, the data is very highly dimensional and there are many more genes than there are time points. Each time series will be replicated typically no more than ten times, and experiments with no replication are not uncommon. The number of genes will often number in the tens of thousands while there are rarely more than ten time points. Even with the falling cost of microarray technology, the limiting factor is often the ability to obtain biological samples which may be restricted due to ethical concerns or other practical, experimental issues. Other challenges include the fact that the data is noisy, with frequent missing observations, and individual heterogeneity.

Our focus is on longitudinal study designs. In this type of microarray experiment, multiple biological units — for example human patients, individual mice or cell lines — are each repeatedly sampled over time to give a collection of observed time series for each gene under study. This type of biological replication is essential for making inference about population parameters but is often overlooked in microarray studies due to experimental issues. A longitudinal microarray experiment is described in Section 2 and provides the data for our case study. The purpose of the study was to follow twelve female and ten male adult human subjects over a period of 6 months, in order to characterise the change in gene expression levels over time in healthy humans. Figure 1 shows some of the raw data for a probe corresponding to the TMEFF1 gene from this example data set where some of the characteristics discussed above can be seen to manifest themselves. A key aspect of human data sets is that the gene expression levels are often collected with covariates - for example, the individual's age, sex and other phenotypic data such as height or weight may be recorded. In the case study, the individuals were stratified by age and gender which allows us to explore not only the evolution of gene expression levels over time but also which genes are differentially expressed between the two gender or age groups.

When modelling experimental data arising from longitudinal microarray experiments there are three distinct challenges: (a) modelling each individual time series, across all genes and individuals, (b) accounting for the correlation between individuals on a gene by gene basis and (c) modelling the correlation between genes. Accounting for each of these sources of correlation — the temporal, the within-gene (between-individual) and the between-gene — is vital for obtaining better parameter estimates and avoiding a loss of power when testing for genes which are differentially or temporally expressed. With less than 10 timepoints, achieving (a) is not possible with standard time series analysis techniques — it is unlikely, for instance, that we would observe any periodicity. Instead, a field which has proven to be quite successful in this area is that of functional data analysis (FDA). In the FDA paradigm, it is assumed that our observations are noisy realisations of an underlying smooth function of time which is to be estimated. These estimated functions, or curves, are then treated as the fundamental unit of data in any subsequent analysis. Formally, the signal-in-noise model assumed is that observation y_i taken at time t_i is given by

$$y_i = f(t_i) + \epsilon_i \tag{1}$$

where $f(\cdot)$ is the function of interest to be estimated and ϵ_i is an error term. Typically the infinite dimensional function $f(\cdot)$ is projected onto some finite dimensional basis using parameterisations such as splines, wavelets or fourier bases. In our discussion we will focus on splines in particular as these regularly occur in the literature in terms of both microarray and functional data analysis. For a thorough treatment of FDA, the monograph Ramsay and Silverman (2005) provides an excellent introduction.

In a longitudinal study, for a particular gene, observations will be collected on not just a single function $f(\cdot)$, but a collection of n functions $f_i(\cdot), i = 1, \cdots, n$, one for each individual biological unit. Often the main quantity of interest is the population mean function $\mu(\cdot)$ characterising the overall population gene expression level over time. In this case we extend the signal-in-noise model (1) so that the jth observation on individual at time t_{ij} is given by

$$y_{ij} = \mu(t_{ij}) + f_i(t_{ij}) + \epsilon_{ij} \tag{2}$$

This is known as the functional mixed-effects model and is an extension of the standard linear mixed-effects model (Harville, 1977) where the fixed- and random-effects are both considered functions. Function $\mu(\cdot)$ is treated as a fixed-effect as it is assumed to be some fixed, but unknown, population function to be estimated. In constrast, the functions $f_i(\cdot), i = 1, \cdots, n$ represent a random sample from the population as a whole and are assumed to be i.i.d realisations of an underlying stochastic process. Model (2) has appeared in a number of different forms depending upon the exact parameterisation of the fixed- and random-effects. For instance, Guo (2002) models both as cubic smoothing splines while Rice and Wu (2001) prefer a B-spline representation.

The task of handling correlations amongst the genes has, to date, generally been overlooked by researchers. It is a challenging, open problem to model both the between- and within-gene correlation simultaneously given the size of the data. Although it is well known that genes are co-regulated, for the sake of tractability the most common approach is to simply model each gene independently. In other words, given the framework outlined thus far, each gene would be modelled as a separate functional mixed-effects model.

In this paper we propose a functional-mixed effects model and a framework for estimation and testing in one-sample problems. The model enables the estimation of a mean response curve with the inclusion of covariates, such as gender and age, also modeled as time-varying smooth functions. We also show how a functional PCA can be applied to the estimated mean curves in order to identify the principal modes of functional variation in the data set, and visually represent the entire set of genes in a low-dimensional plot.

The structure of the paper is as follow. Section 2 provides a description of a data set, previously collected and analysed by Karlovich *et al.* (2009), that we use here as a case study. The proposed model, inferential procedures and functional PCA are provided in Section 3. In Section 4 we present the experimenal results obtained in the context of our case study. In Section 5 we discuss how our methodology compares to related models that have appeared in the literature and compare our experimental results to that of the original study, as well as highlight some of the biological implications. Finally, we conclude in Section 6.

2. Data description

The data set used in our case study is taken from Karlovich *et al.* (2009). The purpose of the study was to characterise the gene expression levels of healthy human individuals over a period of 6 months. 22 subjects were studied, with gene expression levels assayed from blood samples at days 1, 14, 28, 90 and 180. One subject developed lung cancer during the course of the study and died prior to Day 180, thus contributing only a partial time series. All other individuals completed the study and were observed at all 5 time points. Twelve of the individuals were female and ten were male. In the original study, the subjects were divided into two age groups, with the younger group taken to be those subjects less than or equal to 55 years of age, and the older group those subjects over 55.

In the original paper, the observation for a given gene on individual i at time t was

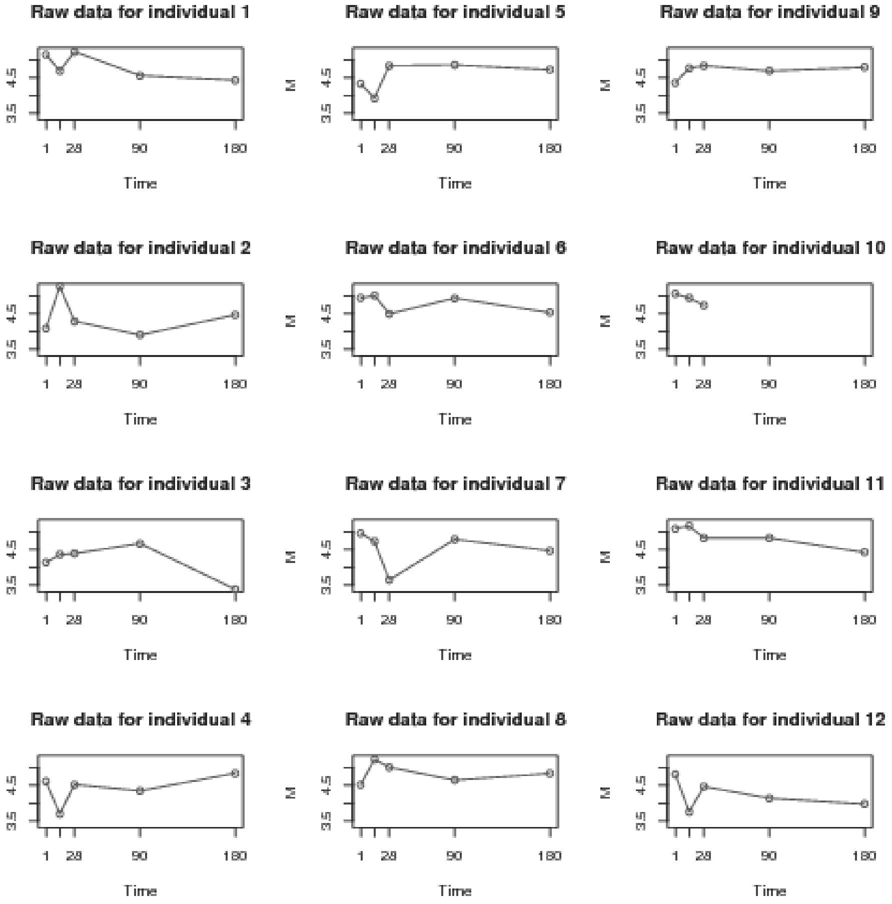

Figure 1 – Raw data for TMEFF1, females. Several key characteristics of the data can be observed: (1) irregularly spaced time points (2) missing data - individual 10 is only observed for the first three time points (3) significant individual heterogeneity (4) noisy observations

modelled as

$$y_{it} = \mu + \alpha_i + \beta_g gender_i + \beta_a age_i + \beta_t time_t + \epsilon_{it}$$

This is a standard linear mixed-effects model. μ is the average gene expression level across all individuals after controlling for gender, age and time effects. α_i is an individual specific term allowing for a deviation in terms of the intercept of the model. The β_g, β_a and β_t parameters separate out the gender, age and time effects respectively while ϵ_{it} is an error term.

The model and study design permitted a wide range of biological issue to be explored. Using t-tests, the significance of the age and gender effects was determined. After correcting for multiple-testing by controlling the false discovery rate (FDR) using the procedure of Benjamini and Hochberg (1995), no genes showed a significant age effect. This was somewhat unexpected given previous studies (Eady *et al.*, 2005; Whitney *et al.*, 2003; Tang *et al.*, 2004) but it was noted that these age effects might be harder to detect in blood than in other tissuses. 78 unique gender genes were identified including XIST, responsible for deactivating one of the X chromosomes in females in order to ensure dosage equivalence, and 23 genes mapped to the Y chromosome. Temporally regulated genes were identified by performing pairwise comparisons between Day 14 and Day 1, Day 28 and Day 14, and Day 180 to Day 90. This was partly due to concerns about a potential batch effect, as Days 1, 14 and 28 were processed in one batch, with Days 90 and 180 being processed in a second batch. No temporally regulated genes were identified in the Day 14 vs Day 1 or the Day 28 vs Day 14 comparisons, but 248 probes were found to be differentially expressed when comparing Day 180 to Day 90, corresponding to 157 unique genes.

Our proposed approach is to replace the original linear mixed-effects model with a functional one. The age and gender effects will be modelled as functions of time, along with the mean and individual curves. To avoid over-parameterisation, all curves will be represented using smoothing splines. The result is a flexible model which permits the interaction of age and gender with time, if the data supports it. During our preprocessing we found little evidence of a batch effect and we will use the entire time course to identify temporally regulated genes, on the basis of the fitted mean function.

3. Methods

We propose the following functional mixed-effects model for the data described in Section 2. Each gene is modelled independently. For a given gene, the observed gene expression level for individual i at time t_{ij} is given by

$$y_i(t_{ij}) = \mu(t_{ij}) + \alpha_k(t_{ij}) + \beta_l(t_{ij}) + \gamma_i(t_{ij}) + \epsilon_{ij} \quad (3)$$

where $\mu(\cdot)$ models the mean expression levels across all individuals after accounting for age and gender effects; $\alpha_k(\cdot)$ is the gender effect for gender k to which individual i belongs with $k = \{\text{Male, Female}\}$; $\beta_l(\cdot)$ is the age group effect for group l to which individual i belongs where $l = \{\text{Young, Old}\}$; $\gamma_i(\cdot)$ is the individual specific effect for individual i and ϵ_{ij} is an error term. The functions $\mu(\cdot)$, $\alpha_k(\cdot)$, $\beta_l(\cdot)$ and $\gamma_i(\cdot)$ are assumed to be smooth functions of time which we wish to estimate based on the noisy observations. We treat $\mu(\cdot)$, $\alpha_k(\cdot)$ and $\beta_l(\cdot)$ as fixed-effects, unknown population functions to be estimated, and the $\gamma_i(\cdot)$ functions which are treated as random-effects as they represent a random sample of functions from the population as a whole. Formally, the $\gamma_i(\cdot)$ are assumed to be i.i.d. realisations of an underlying Gaussian Process with mean 0 and covariance function $\delta(r,s)$.

The functions can be parameterized in a number of ways but we favour smoothing splines as these offer a fine degree of control over the amount to which the data is smoothed. Writing the vector of all observed time points for individual i as $\boldsymbol{t}_i = [t_{i1}, t_{i2}, \cdots, t_{in_i}]^T$ where n_i is the total number of observations on individual i, (3) can be written in matrix form as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\mu} + \boldsymbol{X}_i\boldsymbol{\alpha}_k + \boldsymbol{X}_i\boldsymbol{\beta}_l + \boldsymbol{X}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \tag{4}$$

where $\boldsymbol{y}_i = [y_i(t_{i1}), y_i(t_{i2}), \cdots y_i(t_{in_i})]^T$ and $\boldsymbol{\epsilon}_i = [\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{in_i}]^T$ are vectors of length n_i and $\boldsymbol{\mu} = [\mu(\tau_1), \mu(\tau_2), \cdots, \mu(\tau_M)]^T$ is a vector of length M. The vectors $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}_l$ and $\boldsymbol{\gamma}_i$ are defined similarly to $\boldsymbol{\mu}$. The values $\tau_1, \tau_2, \cdots, \tau_M$ denote the distinct design time points, of which there are M in total, and $\boldsymbol{t}_i$ may differ from these may differ if individual i has missing data or duplicate observations for some time points. The matrix $\boldsymbol{X}_i$ is an incidence matrix of dimension $n_i \times M$ where each row $\boldsymbol{x}_{ij}$ contains all zeroes aside from the column m where $t_{ij} = \tau_m$. Further details on forming the incidence matrices and an example can be found in Appendix A.1. Recall that $\gamma_i(\cdot) \sim GP(0, \delta), i = 1, \cdots, n$, then the vectors $\boldsymbol{\gamma}_i$ are multivariate-normally distributed with mean 0 and covariance matrix $\boldsymbol{D}$ where $\boldsymbol{D}(r,s) = \delta(\tau_r, \tau_s)$. Similarly the noise term $\boldsymbol{\epsilon}_i$ is multivariate-normally distributed with mean 0 and covariance matrix $\boldsymbol{R}_i$, and we assume that the vectors $\boldsymbol{\gamma}_i$ and $\boldsymbol{\epsilon}_i$ are independent. For simplicty we assume that $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}_{n_i \times n_i}$, although a more complicated structure could be modelled at the expense of fitting more parameters. It is further necessary to impose the identifiability constraint that the age and gender fixed-effects for the two groups sum to zero, i.e. $\boldsymbol{\alpha}_{male} + \boldsymbol{\alpha}_{female} = 0$ and $\boldsymbol{\beta}_{young} + \boldsymbol{\beta}_{old} = 0$. For simplicity, therefore, we model a single gender and age effect, $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{female}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_{old}$ respectively. These constraints can equivalently be expressed be rewriting (4) as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\mu} + \boldsymbol{W}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{\beta}_l + \boldsymbol{X}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i \tag{5}$$

where

$$\boldsymbol{W}_i = \begin{cases} -\boldsymbol{X}_i & \text{if } i \text{ is male} \\ \boldsymbol{X}_i & \text{if } i \text{ is female} \end{cases} \quad \boldsymbol{Z}_i = \begin{cases} -\boldsymbol{X}_i & \text{if } i \text{ is young} \\ \boldsymbol{X}_i & \text{if } i \text{ is old} \end{cases}$$

Let $\boldsymbol{\eta} = [\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}]^T$, then (5) can be rewritten more compactly as

$$\boldsymbol{y}_i = \boldsymbol{X}_i^*\boldsymbol{\eta} + \boldsymbol{X}_i\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i$$

where

$$X_i^* = \begin{bmatrix} X_i & W_i & Z_i \end{bmatrix}$$

Finally, the complete data vector for all individuals, $\boldsymbol{y}$, can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\eta} + \widetilde{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{6}$$

where $\boldsymbol{y} = [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \cdots, \boldsymbol{y}_n^T]^T$ is an $N = \sum_i n_i$ length vector, and $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are similarly defined, $\boldsymbol{X}^* = [\boldsymbol{X}_1^{*T}, \boldsymbol{X}_2^{*T}, \cdots \boldsymbol{X}_n^{*T}]^T$ is an $N \times 3M$ matrix and $\widetilde{\boldsymbol{X}} = diag(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)$ is an $N \times nM$ matrix, with the $diag(\cdot)$ operator denoting a block diagonal matrix. The vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are both multivariate-normally distributed with mean $\boldsymbol{0}$ and covariance matrix $\widetilde{\boldsymbol{D}} = diag(\boldsymbol{D}, \cdots, \boldsymbol{D})$ and $\boldsymbol{R} = diag(\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_n)$ respectively.

3.1. Parameter Estimation

Model (6) is in the form of the standard linear mixed-effects model (Laird and Ware, 1982). Standard practice for obtaining estimates of the fixed- and random-effects, $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\gamma}}_i, i = 1, \cdots, n$ would be to maximise the joint likelihood of $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}_i$ (Robinson, 1991). This is equivalent to minimising the following generalized log likelihood (GLL) criterion

$$\begin{aligned} GLL = (\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\eta} - \widetilde{\boldsymbol{X}}\boldsymbol{\gamma})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\eta} - \widetilde{\boldsymbol{X}}\boldsymbol{\gamma}) + \log|\widetilde{\boldsymbol{D}}| \\ + \boldsymbol{\gamma}^T \widetilde{\boldsymbol{D}}^{-1}\boldsymbol{\gamma} + \log|\boldsymbol{R}| \end{aligned} \tag{7}$$

However, in our model the fixed- and random-effects are the fitted values of the smoothing spline estimates of the functions $\mu(\cdot)$, $\alpha(\cdot)$, $\beta(\cdot)$, $\gamma_i(\cdot), i = 1, \cdots, n$, and it is necessary to incorporate a penalty term for the roughness of the smoothing splines into the likelihood. The *penalized* GLL is then given by

$$\begin{aligned} PGLL = GLL + \lambda_\gamma \sum_{i=1}^{n} \left\{ \int_a^b [\gamma_i''(t)]^2 dt \right\} + \lambda \int_a^b [\mu''(t)]^2 dt \\ + \lambda \int_a^b [\alpha''(t)]^2 dt + \lambda \int_a^b [\beta''(t)]^2 dt \end{aligned} \tag{8}$$

where the integrals quantify the roughness of the curves $\mu(\cdot)$, $\alpha(\cdot)$, $\beta(\cdot)$, $\gamma_i(\cdot)$, $i = 1, \cdots, n$ in terms of their squared second derivative, although other penalties could be used. The scalars λ and λ_γ are positive-valued smoothing parameters that control the roughness of the fit. For a given smoothing spline fit, $\lambda = 0$ would correspond to an interpolation of the data points while as λ tends to infinity, the fit tends to a straight line. Note that the same smoothing parameter λ is used for the three fixed-effects functions, $\mu(\cdot)$, $\alpha(\cdot)$, $\beta(\cdot)$, and similarly the same smoothing parameter, λ_γ, is used for all random-effect functions $\gamma_i(\cdot), i = 1, \cdots, n$. This is conceptually justified as each function γ_i is assumed to be a realisation of the same underlying Gaussian Process, but it is possible to envisage selecting a separate smoothing parameter for each fixed- and random-effect function, albeit at the expense of a far greater computational cost.

Minimization of (8) requires calculation of the integral of the squared second derivative of the fixed- and random-effects. In the case of cubic smoothing splines, for a given function $f(t)$ observed at time points $t_1, t_2, \cdots, t_n$ such that $\boldsymbol{f} = [f(t_1), f(t_2), \cdots, f(t_n)]^T$, there is a *roughness matrix* $\boldsymbol{G}$ which can be calculated in a computationally efficient manner that satisfies:

$$\int_a^b [f''(t)]^2 dt = \boldsymbol{f}^T \boldsymbol{G} \boldsymbol{f}$$

this result can be found in Green and Silverman (1994) and we have reproduced the derivation in Appendix A.2 for completeness. Incorporating the roughness matrix into (8) gives

$$\begin{aligned} PGLL &= GLL + \lambda_\gamma \sum_{i=1}^n \gamma_i^T \boldsymbol{G} \gamma_i + \lambda(\mu^T \boldsymbol{G} \mu + \alpha^T \boldsymbol{G} \alpha + \beta^T \boldsymbol{G} \beta) \\ &= GLL + \lambda_\gamma \gamma^T \widetilde{\boldsymbol{G}} \gamma + \lambda \eta^T \boldsymbol{G}^* \eta \end{aligned}$$

where $\widetilde{\boldsymbol{G}}$ is a block diagonal matrix comprised of the matrix $\boldsymbol{G}$ repeated n times. Similarly, $\boldsymbol{G}^*$ is a block diagonal matrix comprised of $\boldsymbol{G}$ repeated three times.

After a rearrangement on the terms featuring in the penalised log-likelihood, the model can be re-written in terms of the *regularised* covariance matrices $\widetilde{\boldsymbol{D}}_\gamma = (\widetilde{\boldsymbol{D}}^{-1} + \lambda_\gamma \widetilde{\boldsymbol{G}})^{-1}$ and $\boldsymbol{V} = \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{D}}_\gamma \widetilde{\boldsymbol{X}}^T + \boldsymbol{R}$, so called because the matrix $\widetilde{\boldsymbol{D}}_\gamma$ is obtained by regularising the covariance matrix $\widetilde{\boldsymbol{D}}$ with the term $\lambda_\gamma \widetilde{\boldsymbol{G}}$. This method of imposing the smoothness constraints by regularisation of the covariance matrix can be credited to Wu and Zhang (2006).

Minimising (8) gives the BLUE and BLUP of the fixed- and random-effects as

$$\hat{\eta} = (\boldsymbol{X}^{*T} \boldsymbol{V}^{-1} \boldsymbol{X}^* + \lambda \boldsymbol{G}^*)^{-1} \boldsymbol{X}^{*T} \boldsymbol{V}^{-1} \boldsymbol{y} \tag{9}$$

$$\hat{\gamma} = \widetilde{\boldsymbol{D}}_\gamma \widetilde{\boldsymbol{X}}^T \boldsymbol{V}^{-1} (y - \boldsymbol{X}^* \eta) \tag{10}$$

The discussion thus far has assumed that the variance components $\boldsymbol{D}$ and σ^2 were known. Of course, in practical applications this will not be the case. Assuming the random-effects γ_i and error terms ϵ are known, the maximum likelihood estimators $\hat{\boldsymbol{D}}$ and $\hat{\sigma}^2$ are given as

$$\hat{\boldsymbol{D}} = \frac{1}{n} \sum_{i=1}^n \gamma_i \gamma_i^T \quad \hat{\sigma}^2 = \frac{1}{N} \epsilon^T \epsilon \tag{11}$$

As the random-effects γ_i and error terms are not, in fact, directly observed, we resort to the Expectation-Maximisation algorithm where they can be treated as missing data. In this procedure the sufficient statistics of $\hat{\boldsymbol{D}}$ and $\hat{\sigma}^2$ — $\gamma_i \gamma_i^T$, $i = 1, \cdots, n$ and $\epsilon^T \epsilon$

respectively — are replaced by their conditional expectations which are calculated at the E-step. In the M-step, the maximum likelihood estimators are then calculated having replaced the sufficient statistics by these conditional expectations, which are given by

$$E[\gamma_i\gamma_i^T|y,\eta=\hat{\eta}] = \hat{\gamma}_i\hat{\gamma}_i^T + \hat{\boldsymbol{D}}_\gamma - \hat{\boldsymbol{D}}_\gamma \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1}\boldsymbol{X}_i\hat{\boldsymbol{D}}_\gamma \quad (12)$$

$$E[\epsilon^T\epsilon|y,\eta=\hat{\eta}] = \hat{\epsilon}^T\hat{\epsilon} + \hat{\sigma}^2 N - \hat{\sigma}^4 tr(V^{-1}) \quad (13)$$

where $tr(\cdot)$ denotes the trace of a matrix and $\boldsymbol{V}_i = \boldsymbol{X}_i\boldsymbol{D}_\gamma\boldsymbol{X}_i^T + \sigma^2\boldsymbol{I}_{n_i\times n_i}$. Derivations of these conditional expectations are given in Appendix A.3.

3.2. Model Selection

Thus far we have treated the smoothing parameters λ and λ_γ as fixed. In reality, optimal values of these parameters must be found using a model selection procedure. Guo (2002) made use of the relationship between a smoothing spline and a linear mixed-effects model in order to treat the smoothing parameters as variances components that could be estimated during the normal course of the EM-algorithm. We prefer, however, to dissociate the model selection from parameter estimation and numerically optimise over the two dimensional space of non-negative reals ($\Lambda \times \Lambda_\gamma$) as this is a much more flexible approach. There are a number of different criteria for scoring the smoothing parameters, all of which essentially trade off between model fit and model complexity.

Ma *et al.* (2006)'s smoothing-spline clustering approach for microarray data, for instance, employed Wahba (1977)'s generalized cross validation (GCV) criterion. It is well known, however, that GCV tends to undersmooth (Lee, 2003). Alternatively, we can employ either the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC):

$$AIC(\lambda,\lambda_\gamma) = -2\text{lik} + 2\text{df}$$

$$BIC(\lambda,\lambda_\gamma) = -2\text{lik} + log(N)\text{df}$$

These two criteria both score the smoothing parameters in terms of the likelihood — measuring the model fit — adjusted for a penalty term for the model complexity, in terms of degrees of freedom. The difference lies in the size of the penalty term, with BIC giving more conservative results when $log(N) > 2$, in other words when there are more than 9 data points.

Both of these criteria, and GCV, have a sound theoretical basis. We suggest, therefore, to choose which one to use on the basis of *a priori* knowledge about the kind of patterns we expect to observe in a given data set. If, as in our example data set, we do not expect there to be many genes with curvy temporal profiles, then we may prefer the more conservative BIC. On the other hand, in a data set with a greater number of time points and with more expected variability — in response to infection for instance — then we may prefer the AIC in order to better capture the more complex patterns expected.

3.2.1. Smoother Matrices

In order to evaluate the criteria, it is necessary to calculate the degrees of freedom of the model. As per Buja *et al.* (1989), the degrees of freedom associated with the fixed- and random-effects, $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$, can be expressed as the trace of some smoother matrix $\boldsymbol{A}$ such that $\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{y}$. Equivalently, it is useful to determine the two smoother matrices $\boldsymbol{A} = \boldsymbol{A}_{\eta} + \boldsymbol{A}_{\gamma}$ so that the degrees of freedom of the fixed- and random-effects can be accounted for separately.

Recall that the fitted values of the fixed-effects at the design time points can be written as $\boldsymbol{X}^{*}\hat{\boldsymbol{\eta}}$. Replacing $\hat{\boldsymbol{\eta}}$ with (9) gives

$$\boldsymbol{X}^{*}\hat{\boldsymbol{\eta}} = \boldsymbol{X}^{*}(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^{*} + \lambda\boldsymbol{G}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{y} = \boldsymbol{A}_{\eta}\boldsymbol{y}$$

and so the smoother matrix $\boldsymbol{A}_{\eta}$ is given by

$$\boldsymbol{A}_{\eta} = \boldsymbol{X}^{*}(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^{*} + \lambda\boldsymbol{G}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}$$

Similarly, the fitted values of the random-effects at the design time points can be written as $\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\gamma}}$, which gives

$$\widetilde{\boldsymbol{X}}\hat{\boldsymbol{\gamma}} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{D}}_{\gamma}\widetilde{\boldsymbol{X}}^{T}\boldsymbol{V}^{-1}(\boldsymbol{I}_{N} - \boldsymbol{A}_{\eta})\boldsymbol{y} = \boldsymbol{A}_{\gamma}\boldsymbol{y}$$

The degrees of freedom of the model can then be calculated as $df = tr(\boldsymbol{A}_{\eta} + \boldsymbol{A}_{\gamma}) + 1$, which is the trace of the smoother matrix plus an additional paramter for fitting the noise variance σ^{2}.

With the scoring function in place any kind of two-dimensional optimisation routine can be used, although in practice a simple grid search or sequential line optimisation is recommended (Wu and Zhang, 2006). We have found that a more sophisticated simplex-search optimiser (Nelder and Mead, 1965) can be employed without incurring a significant computational cost. This allows optimisation over the two smoothing parameters λ and λ_{γ} simultaneously without needing to calculate the derivative of the criterion.

3.3. Confidence Bands

Pointwise confidence bands at the design time points for each of the fixed-effects functions can be determined either theoretically or using a bootstrap resampling procedure. In the case of the former, we have

$$cov(\hat{\boldsymbol{\eta}}) = (\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^{*} + \lambda\boldsymbol{G}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^{*}(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^{*} + \lambda\boldsymbol{G}^{*})^{-1}$$

The diagonal elements of $cov(\hat{\boldsymbol{\eta}})$, therefore, give the variance of the fixed-effects at the design time points with the first M elements corresponding to $\mu(\cdot)$, the next M elements to $\alpha(\cdot)$, and the final M elements to $\beta(\cdot)$. In fact, due to the block diagonal structure of $cov(\hat{\boldsymbol{\eta}})$, these M elements will be the same across all three fixed-effects. Confidence bands for a significance level α at the design time points τ_{i} can then be calculated for $\hat{\mu}$

as $\hat{\mu}(\tau_i) \pm z\sqrt{cov(\hat{\mu}(\tau_i))}$, where z is the critical value under the normality assumption such that $\phi(z) = 1 - \frac{1}{2}\alpha$. These bands can be calculated for the other fixed-effects $\hat{\alpha}$ and $\hat{\beta}$ in an identical fashion.

Alternatively, confidence intervals can be estimated by resampling the between- and within-individual residuals. To construct a bootstrapped sample for a single individual, first one of the individual functions γ_i is randomly selected and evaluated at the design time points - denote this vector as γ^*. Next, M residuals from the noise vector ϵ, are resampled with replacement, writing this vector as ϵ^*. Then, the bootstrapped observation vector $\boldsymbol{y}^*$ is given by

$$\boldsymbol{y}^* = \mu + \boldsymbol{\alpha}^* + \beta^* + \gamma^* + \epsilon^*$$

where $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}$ if the individual is female and $-\boldsymbol{\alpha}$ otherwise, similarly for β. This process is then repeated for n individuals, sampling the individual functions with replacement, to give a complete bootstrapped data set. The model is then fit to this resampled data and new estimates for the fixed-effects obtained. Repeating this process for a large number of iterations gives a large number of fixed-effects estimates from which the confidence bands at a given significance level can be determined empirically.

3.4. Testing for temporal regulation and other effects

Fitting model (6) allows us to separate out the mean, age and gender effects for each gene. It is then possible to determine whether there is a significant group or gender effect by testing the null hypothesis that the corresponding population coefficients are zero. As the effects are modelled as functions, a natural way to quantify their size is the L_2 norm. For instance, the hypothesis of absence of an age effect, for a given gene, versus the alternative hypothesis of an age affect, can be framed as

$$H_0 : ||\alpha(\cdot)||_2 = 0, \quad H_1 : ||\alpha(\cdot)||_2 > 0$$

which is tested using the $L2$ norm of the estimated coefficients.

Assessing the statistical significance in settings similar to ours is complicated by the fact that the sample sizes are generally very small. On the basis of this, and in agreement with previous published studies, we suggest deteriming the null distribution empirically by using data resampling schemes. For example, in Storey *et al.* (2005), the null distribution of their F-type test-statistic was determined using a nonparametric bootstrap procedure by resampling the individual effects and error terms with replacement. In out study, the null distribution of the L_2 norm of the age and gender effects has been estimated empirically using a permutation procedure where the class assignments — male/female or young/old — are randomly permuted. We take a similar approach when testing for temporal regulation. In this case, the null hypothesis of no change over time is formulated as $||\mu'(\cdot)||_2 = 0$ where $\mu'(\cdot)$ is the first derivative of the mean curve. The null distribution in again obtained empirically by randomly permuting the time points.

3.5. Functional Principal Components Analysis

Fitting model (6) to each gene yields a set of mean curves $\mu_i(t), i = 1, \cdots, G$ where G is the total number of genes in the data set. Performing a functional PCA (fPCA) (Ramsay and Silverman, 2005) on this set of curves allows us to identify the main patterns of variation across all genes. We perform this analysis in two stages: (1) the data are smoothed by fitting model (6) to each gene (2) a fPCA is then performed on the smoothed data — in the form of the set of curves $\mu_i(t), i = 1, \cdots, G$. Alternative methods of fPCA such as James *et al.* (2000), which estimate and smooth the PCs directly, cannot be applied in this case where there are two levels of variation — the between and within-gene. Further details of our approach are given below.

Initially, each curve is discretised on a fine grid of n equally spaced points across the range of the time course. If there are N curves in total, this yields a data matrix $\boldsymbol{X}$, of dimension $N \times n$, and a standard PCA can then be performed on $\boldsymbol{X}$. As routinely done, this entails solving the eigenequation

$$\boldsymbol{V}\boldsymbol{u} = \lambda \boldsymbol{u} \tag{14}$$

where $\boldsymbol{V} = N^{-1}\boldsymbol{X}^T\boldsymbol{X}$ is the sample covariance matrix of $\boldsymbol{X}$, λ is one of the eigenvalues of $\boldsymbol{V}$, and $\boldsymbol{u}$ is one of the eigenvectors, or principal components. In the functional setting, we replace $\boldsymbol{V}$ by a covariance function $v(s,t)$, and $\boldsymbol{u}$ by a function of s, $\xi(s)$ such that the eigenequation (14) becomes

$$\int v(s,t)\xi(t)dt = \rho\xi(s) \tag{15}$$

for a given value of s. Noting that after discretisation of the curves the elements of the matrix $\boldsymbol{V} = v(s_j, s_k)$ where j and k are any of the n discretised points on the fine grid, the integral in (15) can be approximated as a summation such that

$$\int v(s,t)\xi(t)dt = w\sum_{k=1}^{n} v(s,s_k)\tilde{\xi}_k$$

where w is the spacing between the points on the fine grid, and $\tilde{\xi}_k$ are the discretised values of the function $\xi(s)$. The approximate discrete form of the functional eigenequation is therefore

$$w\boldsymbol{V}\tilde{\xi} = \rho\tilde{\xi}$$

which corresponds to (14) with $\rho = w\lambda$. Assuming the eigenvectors obtained from the standard PCA have been normalised, the equivalent functional constraint that $\int \xi(s)^2 ds = 1$ is achieved by enforicing $w||\tilde{\xi}||^2 = 1$. The function $\xi(\cdot)$ is then recovered by interpolating the points $\tilde{\xi}$. Assuming the grid is fine enough, the choice of interpolation method is almost irrelevant.

As with a standard PCA, we will wish to retain only a small number of functional PCs. As is standard practice, the eigenvalues ρ can be used to facilitate this choice, by retaining enough PCs to explain most of the variation in the data. Assuming K PCs are retained, for curve i we have

$$y_i(t) = \mu(t) + \sum_k^K x_{ik}\hat{\xi}_k(t) + \epsilon_i(t)$$

where x_{ik} are the PC loadings for curve i. These can be estimated by minimising the residuals $y_i(t) - \sum_k^K x_{ik}\hat{\xi}_k(t)$, which in practice again requires discretisation of the curve i, and the PCs $\hat{\xi}_k(t)$.

4. Results

We fit the functional mixed-effects model described in Section 3 to the example data set described in Section 2, independently for each probe. Convergence of the EM algorithm was confirmed by convergence of the variance components estimates $\hat{\sigma}^2$ and $\hat{\boldsymbol{D}}$ and typically took around 30 iterations. 100 iterations of the simplex optimisation procedure were used to select the smoothing parameters. After obtaining estimates of the mean, age and gender effects, and individual curves, these were assessed for significance. To relieve some of the computational burden, permuted null test statistics were shared across all genes - theoretical results justifying this pooling can be found in Storey *et al.* (2004). Each gene was permuted 32 times, yielding in excess of 1 million null test statistics for each comparison. From these null distributions, empirical p-values were calculated, which were then corrected for multiple testing using the procedure of Benjamini and Hochberg (1995) to control the FDR at 10%.

After applying multiple testing corrections, no significant age genes were identified, as in the original analysis. 21 probes were found to be gender specific. Two of these 21 probes can be found on the Y-chromosome but are not mapped to any known genes. The remaining probes correspond to 7 known genes and 2 open reading frames, given in Table 2. Aside from XIST which, as discussed in Section 2 is only expressed in females and is responsible for X-chromose inactivation to facilitate dosage equivalence between the sexes, all significant genes and the two open reading frames are found on the Y-chromosome.

The highest ranked gender-effect gene on an autosomal chromosome was found to be TUBB2A, located on chromosome 6 and ranked number 23, with an associated FDR of 13%, hence of borderline significance. The gene and fitted mean and gender-effect curves is plotted in Figure 3, where a definite difference between the two groups is apparent, corresponding to between a 3- and 4-fold difference in expression levels.

A total of 299 probes were found to be significantly temporally regulated, corresponding to 183 unique, mapped genes. The highest ranking gene was found to be MBP — myelin basic protein — given as one of the examples in Figure 5. Myelin is an insulating sheath covering nerve cells, essential for the correct functioning of the central nervous system and degredation of myelin can be found in many neurodegenerative diseases such as multiple sclerosis. It is thought that MBP might function to maintain

TABLE 1
19 probes found to be significantly differentially expressed according to gender by Karlovich *et al.* (2009), with a mean log-transformed signal intensity greater than or equal to 7.

Gene Name	Chromosome	Affymetrix ID	Fold Change
-	-	211074_at	0.82
EIF1AX	X	201019_s_at	0.86
TMEFF2	M	224321_at	0.87
FLOT1	6	210142_x_at	0.87
EIF2S3	X	224936_at	0.90
RPS4X	X	213347_x_at	0.91
MGC71993	17	224573_at	0.93
EEF1A1	1	213477_x_at	1.05
EEF1A1	6	206559_x_at	1.07
SPOP	17	204640_s_at	1.07
ERBB2IP	5	217941_s_at	1.09
UHMK1	1	224691_at	1.11
PP784	4	212199_at	1.12
HMGN4	6	209787_s_at	1.13
C10orf45	10	223058_at	1.13
HTATSF1	X	202602_s_at	1.14
GNG2	14	224964_s_at	1.14
HMGN4	6	209786_at	1.17
HMGN4	6	202579_x_at	1.20

the correct structure of myelin, which may explain why we found it to be seasonally regulated, although we could find no existing evidence of this.

We performed a functional PCA of the gene mean curves. Each curve was discretised into 1,000 equally spaced points, then normalised by subtracting the first observation from the rest of the points. Thus, each curve represents the change in expression levels over time, relative to $t = 0$. The first two PC functions are given in Figure 6. The first PC accounts for 99.4% of the variation and corresponds to a linear change in expression levels over time. The second PC accounts for 0.5% of the variation and describes expression levels which rise over the first threee months before falling for the next three months, or vice versa. As these two PCs represent almost all of the variation in the curves, we estimated the loadings for each gene and plotted the results in Figure 4. Four outliers have been highlighted and each of these is plotted in Figure 5. It can be seen that the outliers in the loadings plot correspond to those genes which change most over time, with the distinctive line of points in the center corresponding to genes which change linearly. For these genes with linear dynamics, the size of the first PC loading is relative to the slope. Genes which can be separated on the y-axis are those with a quadratic temporal profile.

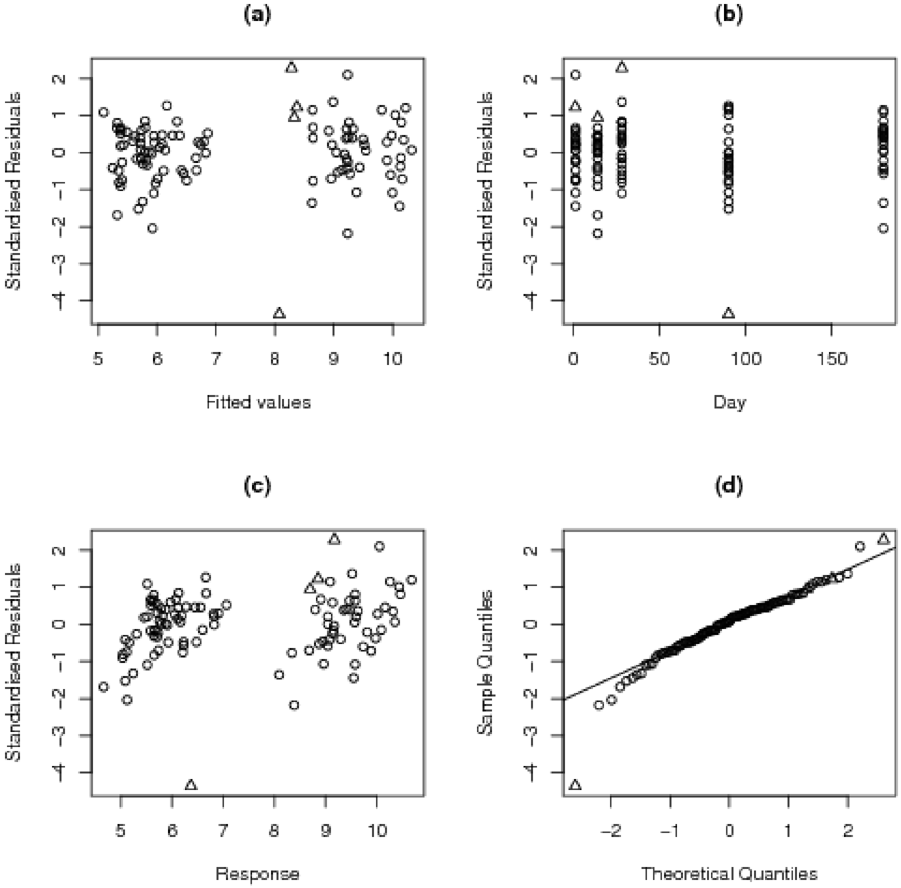

Figure 2 – Residual analysis for the TUBB2A model fit. (a) Standardised residuals against fitted values (b) Standardised residuals against time (c) Standardised residuals against observations (d) QQ-plot of standardised residuals. These plots can be used to detect patterns in the data which the model has failed to capture. Aside from the obvious groupings as a result of the difference in gene expression levels between males and females, there appears to be little structure to the residuals. In all cases, the triangles correspond to observations on subject 174, who developed lung cancer during the course of the study and died prior to the final time point. It can be seen that this subject contributes two obvious outlying residuals, which may have negatively impacted the goodness of fit criteria calculated by Karlovich *et al.* (2009), possibly resulting in its removal from any subsequent analysis.

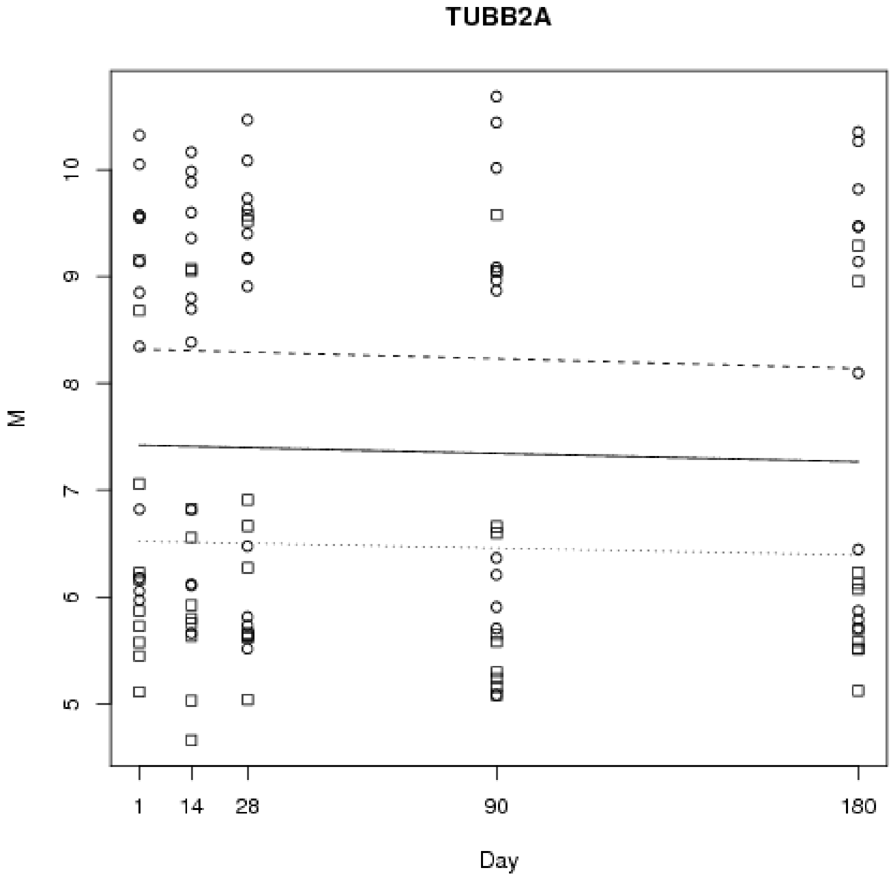

Figure 3 – Plot of TUBB2A's fitted longitudinal profiles. We have identified TUBB2A as a gene with a potentially novel gender effect. Observations on females are shown as squares, and those on males are shown as circles. The solid line is the overall mean expression level over time, after removing age and gender effects. The dotted line is the mean plus gender effect for females, and the dashed line is the mean plus gender effect for males.

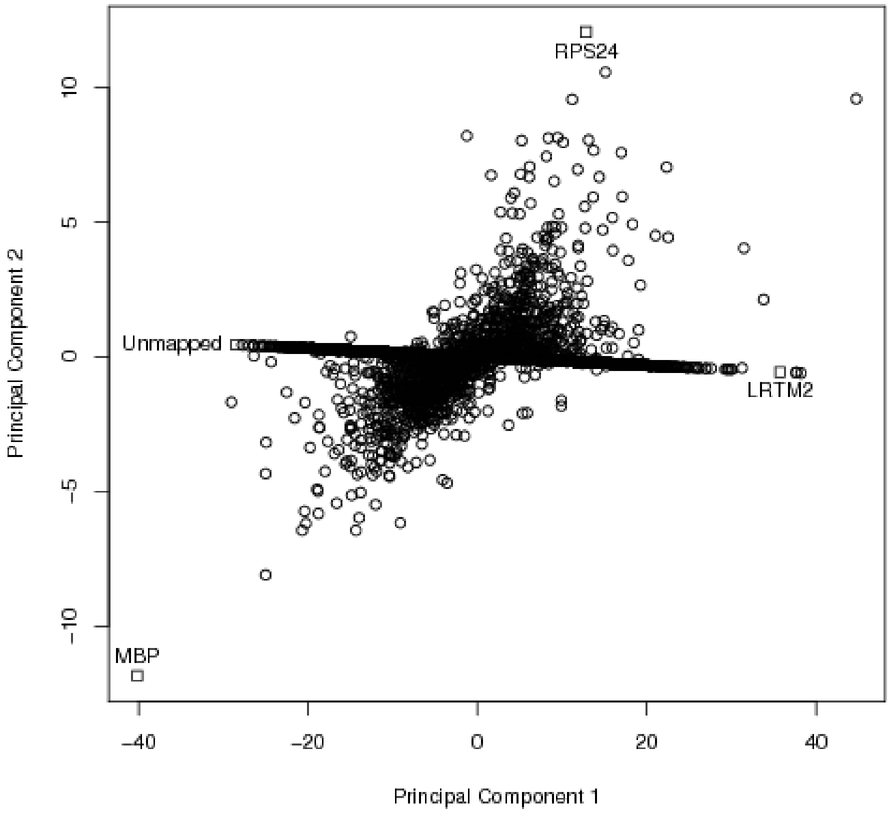

Figure 4 – Functional principal components analysis loadings plot. Two functional principal components capture 99.9% of the observed variation in the fitted mean curves for each gene. The loadings on the first principal component function corresponds to the x-axis, which represents linear variation over time. The second principal component function captures variation which is of a more quadratic nature. These two principal component functions are given in Figure 6. Four outliers representing the spectrum of observed temporal profiles have been highlighted; individual plots for these genes are given in Figure 5.

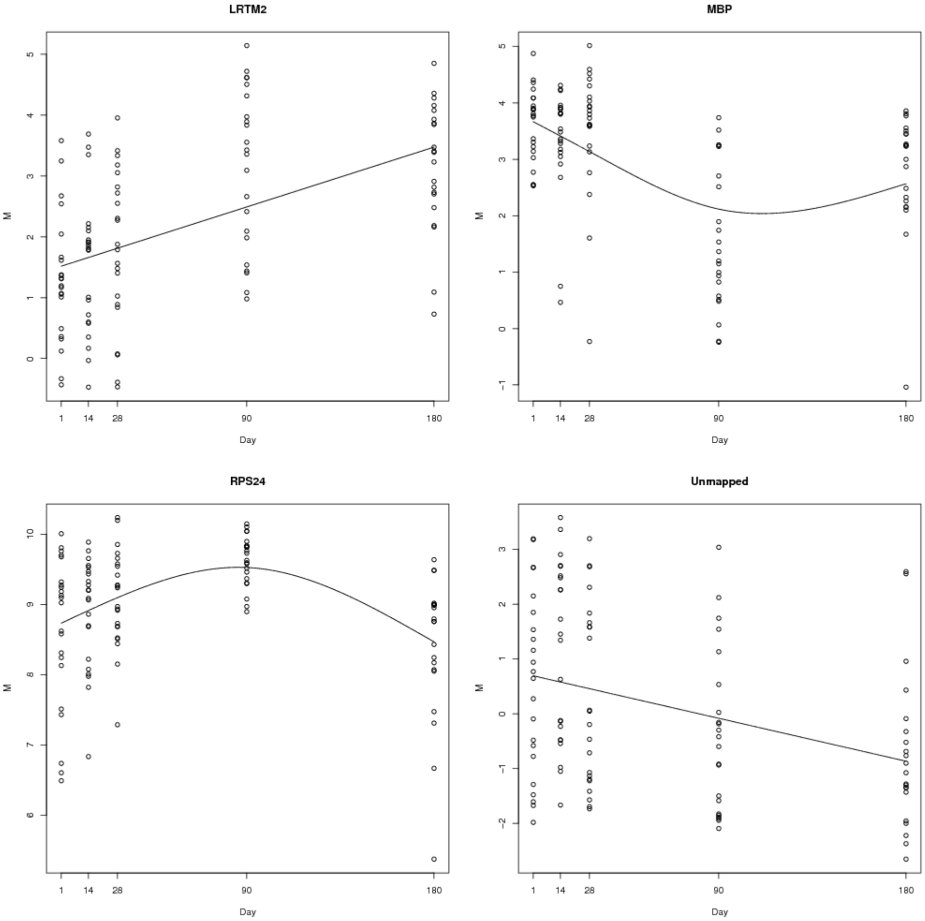

Figure 5 – Outlying genes in the fPCA loadings plot shown in Figure 4. These are some of the genes which show the greatest change in expression levels over time.

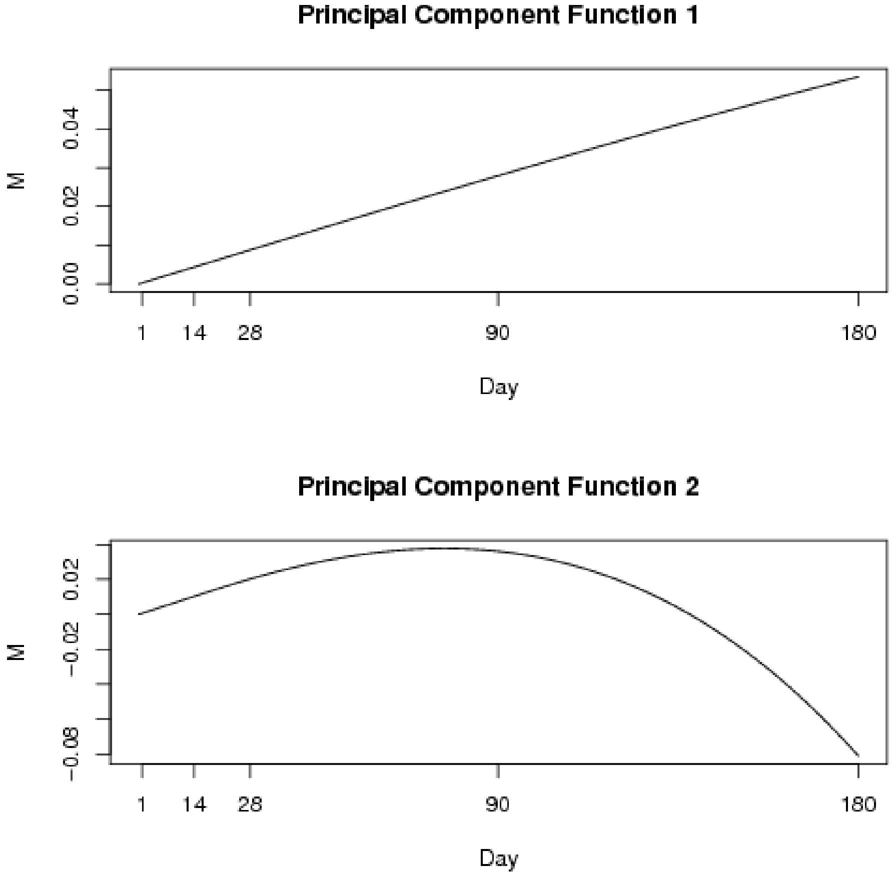

Figure 6 – Two principal component functions which explain 99.9% of the variation observed in the fitted mean curves for each gene. The first principal component describes a linear relationship with time. The second principal component captures a more quadratic fit.

5. Discussion

A number of different models have been proposed in the literature for the analysis of microarray time series data. One of the earliest examples of a FDA approach to the modelling of microarray time series data was Bar-Joseph *et al.* (2003) which dealt with the issue of clustering unreplicated data. In their model, the curves were parameterised using B-splines and functional mixed-effects models were used to estimate the cluster mean curves and model the within-cluster variability. In their approach, the function $\mu(\cdot)$ in (2) represents a given cluster's mean, and the functions $f_i(\cdot), i = 1, \cdots, n$ represent the temporal profiles of each of the genes belonging to this cluster, of which there are n. A specialised EM algorithm was used to handle dynamic cluster assignments. A very similar approach was developed independently by Luan and Li (2003).

A limitation of the models in Bar-Joseph *et al.* (2003) and Luan and Li (2003) is that the B-spline parameterisation of the curves requires selecting both the number and location of the *knots* — breakpoints for the piecewise polynomials — which control the overall smoothness of the fitted curve $\hat{f}(\cdot)$. As the total number of knots is limited by the number of time points, there is limited scope for controlling the smoothness of the fit. Furthermore, each curve was parameterised using the same number of knots which may be unable to fully capture the wide range of temporal profiles we are likely to observe. Ma *et al.* (2006) set out to resolve these issues with their alternative framework for clustering. In their model, the cluster mean curves — $\mu(\cdot)$ in (2) — are represented using smoothing splines, which place a knot at each design time point and use a roughness penalty to avoid fitted curves which are too 'wiggly'. One drawback to their approach, however, is that the individual functions $f_i(\cdot), i = 1, \cdots, n$ are only modelled as scalar shifts rather than smooth curves. This leads to a more parsimonious model which avoids fitting too many parameters but may fail to adequately model the within-cluster variability.

Angelini *et al.* (2009) adopt a fully Bayesian approach to estimation and testing in unreplicated or cross-sectional microarray data sets. Each gene is represented using Legendre polynomials. Three choices for a prior on the noise variance σ^2 allows for errors which are marginally normal, Student t or double exponentially distributed, although σ^2 is assumed the same for all genes. This assumption is unlikely to hold in practice, as a correlation between gene expression intensity and measurement noise is well known (Tusher *et al.*, 2001). Given the fully Bayesian framework, hypothesis testing for differences in expression levels across two biological groups is performed using Bayes Factors.

A handful of models and computer packages have also specifically been suggested to model longitudinal data. For instance, *Timecourse* is an R package based on Tai and Speed (2006), where multivariate analysis techniques are applied directly to the vectors of observations. This treatment of time as an unordered categorical variable — found also in ANOVA approaches as in Wang and Kim (2003) — has some significant drawbacks. In particular, the method cannot handle missing data, the results obtained by an analysis would be invariant to permutation of the time points, and it is assumed that the time points are regularly spaced. Furthermore, this method only ranks the genes with no guidance given as to how to evaluate significance.

TABLE 2
21 probes found to have a significant gender-effect Aside from XIST, all of these probes can be found on the Y-chromosome. Q-value indicates the corresponding false discovery rate (FDR) if a particular gene is taken to be the cut-off between significant and non-significant.

Gene Name	Chromosome	Affymetrix ID	L_2 norm	q-value
XIST	X	224588_at	57.2	0.00248
XIST	X	224590_at	53.9	0.00248
EIF1AY	Y	204409_s_at	48.8	0.00248
RPS4Y1	Y	201909_at	42.9	0.00248
DDX3Y	Y	205000_at	36.7	0.00248
XIST	X	214218_s_at	35.4	0.00248
EIF1AY	Y	204410_at	34.5	0.00248
XIST	X	221728_x_at	33.2	0.00248
CYorf15B	Y	214131_at	30.3	0.00248
CYorf15A	Y	232618_at	29.2	0.00248
USP9Y	Y	228492_at	27.8	0.00248
JARID1D	Y	206700_s_at	25.3	0.00248
XIST	X	224589_at	24.8	0.00248
-	Y	244482_at	22.4	0.00430
XIST	X	227671_at	22.2	0.00430
TSIX	X	231592_at	18.4	0.0247
BCORL2	Y	1562313_at	18.4	0.0247
-	Y	1560800_at	16.3	0.0323
DDX3Y	Y	205001_s_at	16.1	0.0543
CYorf15B	Y	223646_s_at	14.1	0.0597
CYorf15A	Y	236694_at	13.8	0.0845

The EDGE method of Storey *et al.* (2005) is a FDA approach to modelling both longitudinal and cross-sectional microarray data. In their method for longitudinal data analysis, each gene is modelled independently as a separate functional mixed-effects model. The mean curve — $\mu(\cdot)$ in (2) — is modelled as a B-spline while the individual effects are treated as scalar shifts as in Ma *et al.* (2006). A complete framework for detecting genes differentially expressed across two or more biological groupings is presented, with the model estimation performed by an EM algorithm. Differential expression is quantified using an F-type statistic which compares the residuals of a null model where the biological groupings are ignored to an alternative model where the groupings are taken into account. Significance is assessed by using a resampling bootstrap procedure to estimate the null distribution of this F-type statistic, and the multiple testing problem is handled by analysing the empirical p-value histogram (Storey and Tibshirani, 2003) to estimate the positive false discovery rate.

Another way of accounting for the within-gene variance is to perform a functional PCA. As we have pointed out, this is analagous to the standard PCA, except the principal components (PCs) are functions rather than finite dimensional vectors. There have been a number of different methods suggested for estimating the PCs in a functional context including direct estimation in a mixed-effects model framework (James *et al.*, 2000), standard PCA on discretised curves (Ramsay and Silverman, 2005) and 'Principal Components Analysis through Conditional Expectation' (PACE) (Yao *et al.*, 2005). It is this latter approach which Liu and Yang (2009) applied to the analysis of microarray data; however, PACE was originally proposed for data where the observations on each individual are taken at different time points — for example, in the case of growth curve data — in our experience, microarray experiments tend to have much more regular designs, with each individual observed at the same time points, although these may, indeed, be unequally spaced.

Some key shortcomings of these methods should be noted. Firstly, none of the methods can incorporate the gender and age covariates, particularly as functions of time. Secondly, all of these approachs either use B-splines and/or model the individual 'functions' as scalar-shifts, both of which lead to inflexible models. Finally, we are not aware of any existing methods which address the issue of modelling both the within- and between-gene variation. Our proposed methodology has been developed to address some of these limitations.

Our results related to the case study presented in section 2 can be compared to the original findings of Karlovich *et al.* (2009), who used a non-functional mixed-effect model. Those authors listed 19 probes detected as having a significant gender effect and with a log-transformed signal intensity greater than 7, which we have reproduced here in Table 1. No justification for this cut-off of 7 is provided, and this filter gives misleading results. For instance, all of the significant gender genes we have identified fail to meet the cut-off. This is because the mean log-transformed signal intensity is taken across both genders, and all of our genes aside from XIST are found on the Y-chromosome and hence completely unexpressed in females.

We were unable to find any confirmation in the literature that TUBB2A is a sex-related gene, and it does not appear in the 15 probes given by Karlovich *et al.* (2009). With a mean log-transformed signal intesity of 7.4, it meets their cut-off criteria. It is possible that they removed the probe from their analysis if the residuals from their model were found to be non-normally distributed. Indeed, the unadjusted p-value for the Shapiro-Wilk test on the residuals of *our* model for this probe is $2.54e^{-5}$. However, looking at the residual analysis plotted in Figure 2, it is easy to see that there is one very large outlier. This observation corresponds to subject 174 at Day 90. Subject 174 is the individual who developed lung cancer between days 28 and 90, and died prior to day 180. If this observation is removed then the unadjusted Shapiro-Wilk p-value is 0.297, and the null hypothesis that the residuals are normally distributed is no longer rejected. Hence, TUBB2A may indeed be a novel gender regulated gene.

The number of temporally regulated genes we identified are consistent with Karlovich *et al.* (2009), although their method for identifying differentially expressed genes is quite dissimilar to ours (see Section 2). Indeed, although they found 66 significant genes associated with apoptosis, we found only 15, suggesting the actual significant genes found may vary more widely than the numbers suggest.

6. Conclusions

In this paper we have demonstrated a complete framework for the analysis of microarray time series data. The unique characteristics of microarry data lend themselves well to a functional data analysis approach and we have shown how this naturally extends to the inclusion of covariates such as age and sex. Our model presented here is a specialisation of the more general functional mixed-effects model (Rice and Wu, 2001; Guo, 2002) and, to the best of our knowledge, we are the first to show how to derive the maximum-likelihood estimators, EM-algorithm, confidence intervals and smoother matrix with more than one fixed-effects function.

We were motivated by a real data set and we have aimed to improve upon the existing results with a more flexible model. By taking a roughness penalty approach, this is achieved while avoiding overfitting, allowing for a departure from the original linear mixed-effects model when the data permits it. A deeper biological interpretation is required to fully assess our success here, but the results we have highlighted in this paper suggest that we can easily attach meaning to our findings. It may also prove worthwhile performing a comparative analysis with Eady *et al.* (2005), which is another, similar longitudinal study taken over a shorter period of five weeks.

A. Appendix

A.1. Example incidence matrix

In our example data set, there are 5 design time points: Day 1, 14, 28, 90 and 180. Therefore, the incidence matrices for all individuals, $\boldsymbol{X}_i, i = 1, \cdots, n$, all have 5 columns. The first column corresponds to observations at Day 1, the second to observations at Day 14 and so on. The rows correspond to the specific observations on a particular individual. If the individual is observed once at each design time point, then, assuming their vector of observations $\boldsymbol{y}_i$ has been ordered according to the time points, $\boldsymbol{X}_i = \boldsymbol{I}$.

Now consider the case of subject 174 who died prior to Day 180 and hence only contributed 4 observations at each of the remaining design time points. The design matrix for this individual has 4 rows, corresponding to the 4 observations, but still has 5 columns, corresponding to the design time points. Specifically the incidence matrix in this case is:

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Note how there is no 1 in the final column which would correspond to an observation at Day 180.

A.2. Specification of roughness matrix $\boldsymbol{G}$

Green and Silverman (1994) show that there is a straight forward way to calculate the roughness matrix for a smoothing spline given the set of distinct time points $\tau_1, \cdots \tau_M$. The roughness matrix is given as $\boldsymbol{G} = \boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^T$ where the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are defined as follows. First calculate $h_r = \tau_{r+1} - \tau_r, r = 1, \cdots, M-1$, the differences between successive time points. Then matrix $\boldsymbol{A}$ is an $M \times (M-2)$ matrix whose entries $a_{r,s}$ are given by

$$a_{r,r} = h_r^{-1}, \quad a_{r+1,r} = -(h_r^{-1} + h_{r+1}^{-1}), \quad a_{r+2,r} = h_{r+1}^{-1}$$

for $r = 1, \cdots, M-2$ and 0 elsewhere. $\boldsymbol{B}$ is an $(M-2) \times (M-2)$ matrix with the entries given by

$$b_{1,1} = \tfrac{h_1+h_2}{3}, \quad b_{2,1} = \tfrac{h_2}{6}$$
$$b_{r,r+1} = \tfrac{h_{r+1}}{6}, \quad b_{r+1,r+1} = \tfrac{h_{r+1}+h_{r+2}}{3}, \quad b_{r+2,r+1} = \tfrac{h_{r+2}}{6}, \quad r = 1, \cdots, M-4$$
$$b_{M-3,M-2} = \tfrac{h_{M-2}}{6}, \quad b_{M-2,M-2} = \tfrac{h_{M-2}+h_{M-1}}{3}$$

A.3. Derivation of conditional expectations

We begin by first considering the posterior expectation of $\gamma_i \gamma_i^T$ which, using basic properties of expectations, can be rewritten as:

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \gamma_i \gamma_i^T | \boldsymbol{y}, \eta = \hat{\eta}\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\gamma_i \gamma_i^T | \boldsymbol{y}, \eta = \hat{\eta}\right]$$

The definition of covariance allows us to write:

$$\begin{aligned} E\left[\gamma_i \gamma_i^T | \boldsymbol{y}, \eta = \hat{\eta}\right] &= E\left[\gamma_i | \boldsymbol{y}, \eta = \hat{\eta}\right] E\left[\gamma_i^T | \boldsymbol{y}, \eta = \hat{\eta}\right] \\ &+ Cov(\gamma_i | \boldsymbol{y}, \eta = \hat{\eta}, \gamma_i^T | \boldsymbol{y}, \eta = \hat{\eta}) \end{aligned}$$

The problem is now to determine the mean and covariance of $\gamma_i | \boldsymbol{y}$, for which we use a standard result conerning the multivariate normal distribution (See, for example, Anderson, 1958) which says, for any vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ distributed as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}\right)$$

the conditional distribution of $\boldsymbol{x}_1 | \boldsymbol{x}_2$ is given by

$$x_1 | x_2 \sim \mathcal{N}[\mu_1 + V_{12}V_{22}^{-1}(x_2 - \mu_2), V_{11} - V_{12}V_{22}^{-1}V_{21}]$$

If we let $\boldsymbol{x}_1 = \gamma$ and $\boldsymbol{x}_2 = \boldsymbol{y}$, and derive the covariance of γ and $\boldsymbol{y}$ as $Cov(\gamma, \boldsymbol{y}) = \widetilde{\boldsymbol{D}}_\gamma \widetilde{\boldsymbol{X}}^T$ then we have

$$\begin{bmatrix} \gamma \\ y | \eta = \hat{\eta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ X\hat{\eta} \end{bmatrix}, \begin{bmatrix} \widetilde{D}_\gamma & \widetilde{D}_\gamma \widetilde{X}^T \\ \widetilde{X}\widetilde{D}_\gamma & V \end{bmatrix}\right)$$
$$\gamma | y, \eta = \hat{\eta} \sim \mathcal{N}[\widetilde{D}_\gamma \widetilde{X}^T V^{-1}(y - X\hat{\eta}), \widetilde{D}_\gamma - \widetilde{D}_\gamma \widetilde{X}^T V^{-1} \widetilde{X} \widetilde{D}_\gamma]$$

Recognising that, because $\widetilde{D}_\gamma$ and V are block diagonal and $\widetilde{X}\widetilde{D}_\gamma V^{-1}(y - X\hat{\eta}) = \hat{\gamma}$, we have

$$\gamma_i | y, \eta = \hat{\eta} \sim \mathcal{N}[\hat{\gamma}_i, D_\gamma - D_\gamma X_i^T V_i^{-1} X_i D_\gamma]$$

and we can now write

$$E\left[\gamma_i \gamma_i^T | y, \eta = \hat{\eta}\right] \;=\; \hat{\gamma}_i \hat{\gamma}_i^T + [D_\gamma - D_\gamma X_i^T V_i^{-1} X_i D_\gamma]$$

For the posterior expectation of σ^2, we follow exactly the same approach, writing

$$\begin{bmatrix} \epsilon \\ y|\eta = \hat{\eta} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ X\hat{\eta} \end{bmatrix}, \begin{bmatrix} R & R \\ R & V \end{bmatrix}\right)$$
$$\epsilon | y, \eta = \hat{\eta} \sim \mathcal{N}[RV^{-1}(y - X\hat{\eta}), R - RV^{-1}R]$$
$$\epsilon_i | y, \eta = \hat{\eta} \sim \mathcal{N}[R_i V_i^{-1}(y_i - X_i\hat{\eta}), R_i - R_i V_i^{-1} R_i]$$

Note that

$$\begin{aligned} R_i V_i^{-1}(y_i - X_i\hat{\eta}) &= (V_i - X_i D X_i^T) V_i^{-1}(y_i - X_i\hat{\eta}) \\ &= (I - X_i D_\gamma X_i^T V_i^{-1})(y_i - X_i\hat{\eta}) \\ &= (y_i - X_i\hat{\eta}) - X_i D_\gamma X_i^T V_i^{-1}(y_i - X_i\hat{\eta}) \\ &= y_i - X_i\hat{\eta} - X_i\hat{\gamma}_i \\ &= \hat{\epsilon}_i \end{aligned}$$

and

$$\begin{aligned} R_i - R_i V_i^{-1} R_i &= \sigma^2 I_{n_i} - \sigma^4 V_i^{-1} \\ &= \sigma^2 (I_{n_i} - \sigma^2 V_i^{-1}) \end{aligned}$$

and using the identity

$$E[\epsilon_i^T \epsilon_i | y, \eta = \hat{\eta}] \;=\; tr\{E[\epsilon_i \epsilon_i^T | y, \eta = \hat{\eta}]\}$$

allows us to derive

$$\begin{aligned} E[\epsilon_i^T \epsilon_i | y, \eta = \hat{\eta}] &= tr\{E[\epsilon_i \epsilon_i^T | y, \eta = \hat{\eta}]\} \\ &= tr\{E[\epsilon_i | y, \eta = \hat{\eta}] E[\epsilon_i^T | y, \eta = \hat{\eta}] + \sigma^2(I_{n_i} - \sigma^2 V_i^{-1})\} \\ &= tr\{\hat{\epsilon}_i \hat{\epsilon}_i^T + \sigma^2(I_{n_i} - \sigma^2 V_i^{-1})\} \\ &= tr\{\hat{\epsilon}_i \hat{\epsilon}_i^T\} + tr\{\sigma^2(I_{n_i} - \sigma^2 V_i^{-1})\} \\ &= \hat{\epsilon}_i^T \hat{\epsilon}_i + \sigma^2(tr\{I_{n_i}\} - \sigma^2 tr\{V_i^{-1}\}) \\ &= \hat{\epsilon}_i^T \hat{\epsilon}_i + \sigma^2(n_i - \sigma^2 tr\{V_i^{-1}\}) \end{aligned}$$

and so

$$E[\epsilon^T \epsilon | y, \eta = \hat{\eta}] \quad = \quad \sum_{i=1}^{n} [\hat{\epsilon}_i^T \hat{\epsilon}_i + \sigma^2 (n_i - \sigma^2 tr\{V_i^{-1}\})]$$

REFERENCES

T. W. ANDERSON (1958). *Introduction to Multivariate Statistical Analysis.* Wiley.

C. ANGELINI, D. D. CANDITIIS, M. PENSKY (2009). *Bayesian models for two-sample time-course microarray experiments.* Computational Statistics & Data Analysis, 53, no. 5, pp. 1547 – 1565. Statistical Genetics & Statistical Genomics: Where Biology, Epistemology, Statistics, and Computation Collide.

Z. BAR-JOSEPH, G. K. GERBER, D. K. GIFFORD, T. S. JAAKKOLA, I. SIMON (2003). *Continuous representations of time-series gene expression data.* Journal of Computational Biology, 10, no. 3-4, pp. 341–356.

Y. BENJAMINI, Y. HOCHBERG (1995). *Controlling the false discovery rate: A practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society, 57, no. 1, pp. 289–300.

A. BUJA, T. HASTIE, R. TIBSHIRANI (1989). *Linear smoothers and additive models.* The Annals of Statistics, 17, no. 2, pp. 453–510.

S. E. CALVANO, W. XIAO, D. R. RICHARDS, R. M. FELCIANO, H. V. BAKER, R. J. CHO, R. O. CHEN, B. H. BROWNSTEIN, J. P. COBB, S. K. TSCHOEKE, C. MILLER-GRAZIANO, L. L. MOLDAWER, M. N. MINDRINOS, R. W. DAVIS, R. G. TOMPKINS, S. F. LOWRY, L. S. C. R. PROGRAMINFLAMM, H. R. TO INJURY (2005). *A network-based analysis of systemic inflammation in humans.* Nature, 437, no. 7061, pp. 1032–1037.

J. J. EADY, G. M. WORTLEY, Y. M. WORMSTONE, J. C. HUGHES, S. B. ASTLEY, R. J. FOXALL, J. F. DOLEMAN, R. M. ELLIOTT (2005). *Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers.* Physiol. Genomics, 22, no. 3, pp. 402–411.

P. J. GREEN, B. W. SILVERMAN (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall.

W. GUO (2002). *Functional mixed effects models.* Biometrics, 58, no. 1, pp. 121–128.

D. A. HARVILLE (1977). *Maximum likelihood approaches to variance component estimation and to related problems.* Journal of the American Statistical Association, 72, no. 358, pp. 320–388.

G. JAMES, T. HASTIE, C. SUGAR (2000). *Principal component models for sparse functional data.* Biometrika, 87, no. 3, pp. 587–602.

C. KARLOVICH, G. DUCHATEAU-NGUYEN, A. JOHNSON, P. MCLOUGHLIN, M. NAVARRO, C. FLEURBAEY, L. STEINER, M. TESSIER, T. NGUYEN, M. WILHELM-SEILER, J. CAULFIELD (2009). *A longitudinal study of gene expression in healthy individuals.* BMC Medical Genomics, 2, no. 1, p. 33.

N. M. LAIRD, J. H. WARE (1982). *Random-effects models for longitudinal data.* Biometrics, 38, pp. 963–974.

T. C. M. LEE (2003). *Smoothing parameter selection for smoothing splines: a simulation study.* Computational Statistics & Data Analysis, 42, no. 1-2, pp. 139 – 148.

X. LIU, M. C. K. YANG (2009). *Identifying temporally differentially expressed genes through functional principal components analysis.* Biostat, p. kxp022.

Y. LUAN, H. LI (2003). *Clustering of time-course gene expression data using a mixed-effects model with B-splines.* Bioinformatics, 19, no. 4, pp. 474–482.

P. MA, C. I. CASTILLO-DAVIS, W. ZHONG, J. S. LIU (2006). *A data-driven clustering method for time course gene expression data.* Nucleic Acids Res, 34, no. 4, pp. 1261–1269.

J. A. NELDER, R. MEAD (1965). *A Simplex Method for Function Minimization.* The Computer Journal, 7, no. 4, pp. 308–313.

J. RAMSAY, B. W. SILVERMAN (2005). *Functional Data Analysis.* Springer, 2 ed.

J. RICE, C. WU (2001). *Nonparametric mixed effects models for unequally sampled noisy curves.* Biometrics, 57, pp. 253–259(7).

G. K. ROBINSON (1991). *That BLUP is a good thing: the estimation of random effects.* Statistical Science, 6, no. 1, pp. 15–32.

P. T. SPELLMAN, G. SHERLOCK, M. Q. ZHANG, V. R. IYER, K. ANDERS, M. B. EISEN, P. O. BROWN, D. BOTSTEIN, B. FUTCHER (1998). *Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.* Mol. Biol. Cell, 9, no. 12, pp. 3273–3297.

J. D. STOREY, J. E. TAYLOR, D. SIEGMUND (2004). *Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 66, no. 1, pp. 187–205.

J. D. STOREY, R. TIBSHIRANI (2003). *Statistical significance for genomewide studies.* Proceedings of the National Academy of Sciences of the United States of America, 100, no. 16, pp. 9440–9445.

J. D. STOREY, W. XIAO, J. T. LEEK, R. G. TOMPKINS, R. W. DAVIS (2005). *Significance analysis of time course microarray experiments.* Proc Natl Acad Sci U S A, 102, no. 36, pp. 12837–12842.

Y. C. TAI, T. P. SPEED (2006). *A multivariate empirical bayes statistic for replicated microarray time course data.* Annals of Statistics, 34, no. 5, pp. 2387–2412.

Y. TANG, A. LU, R. RAN, B. J. ARONOW, E. K. SCHORRY, R. J. HOPKIN, D. L. GILBERT, T. A. GLAUSER, A. D. HERSHEY, N. W. RICHTAND, M. PRIVITERA, A. DALVI, A. SAHAY, J. P. SZAFLARSKI, D. M. FICKER, N. RATNER, F. R. SHARP (2004). *Human blood genomics: distinct profiles for gender, age and neurofibromatosis type 1.* Molecular Brain Research, 132, no. 2, pp. 155 – 167. Neurogenomics.

V. G. TUSHER, R. TIBSHIRANI, G. CHU (2001). *Significance analysis of microarrays applied to the ionizing radiation response.* Proceedings of the National Academy of Sciences of the United States of America, 98, no. 9, pp. 5116–5121.

G. WAHBA (1977). *Practical approximate solutions to linear operator equations when the data are noisy*. SIAM Journal on Numerical Analysis, 14, no. 4, pp. 651–667.

J. WANG, S. K. KIM (2003). *Global analysis of dauer gene expression in Caenorhabditis elegans*. Development, 130, no. 8, pp. 1621–1634.

A. R. WHITNEY, M. DIEHN, S. J. POPPER, A. A. ALIZADEH, J. C. BOLDRICK, D. A. RELMAN, P. O. BROWN (2003). *Individuality and variation in gene expression patterns in human blood*. Proceedings of the National Academy of Sciences of the United States of America, 100, no. 4, pp. 1896–1901.

H. WU, J.-T. ZHANG (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley.

F. YAO, H.-G. MÜLLER, J.-L. WANG (2005). *Functional data analysis for sparse longitudinal data*. Journal of the American Statistical Association, 100, no. 470, pp. 577–590.