

Analysis of proteomics samples from mass spectrography of a longitudinal experiment using Functional Data Analysis (FDA)

Siem H. Heisterkamp

Organon Biosciences, Schering Plough Corporation, and
Groningen Bioinformatics Centre, University of Groningen

Abstract

In a longitudinal experiment cerebral spinal fluid (CSF) was sampled from rats and human volunteers subject to different treatments, at regular time points within a period of 25 hours and analyzed using mass-spectrography. Typically the data from each sample are represented as intensity profiles (spectra) for a large number of mass over charge (m/z) values- each representing a peptide- , with high values (peaks) at some m/z positions (peak positions). The number of peak positions may vary between 20,000 to 80,000. Functional Data Analysis (FDA) for linear models was used to find those peak positions which yield differences between treatments and/or time points. As FDA reduces the resolution of the peak positions - and thus computational efforts - one will find rather ranges of interesting peaks than the individual peak positions themselves. However, the usual linear FDA models ignore correlation within subjects. To overcome this, in a second step a mixed effect linear model for each peak in the selected region(s) is applied. Pre-defined contrast of treatment and time combinations were tested allowing for multiplicity for the numbers of peaks tested, while stabilizing the covariance, similar to the work by Smyth (2004). As an alternative we propose a mixed FDA (mFDA), to find better defined ranges of interesting peaks, and thus reducing the multiplicity in the second step. The connection between FDA with a roughness penalty and linear mixed models as a penalized linear regression is used in an empirical Bayes setting.

1 Introduction

Research and development of drugs for diseases of the Central Nervous System (CNS) is hampered by severe restrictions of the availability of bio-markers in serum or cerebral spinal fluid (CSF). As CSF is in direct contact with brain tissue and is continually refreshed, it possibly contains (fragments of) proteins carrying information of processes in the brain. Using mass-spectrography is a fast method to detect mixtures of molecules in CSF. CSF has been sampled from the spine of either rats or human volunteers, in longitudinal experiments. As observations within subjects are correlated an adequate statistical analysis, using mixed models is required. However, the amount of data is prohibitive to do so directly.

2 Use of linear models in FDA followed by linear mixed models per peak

Consider vectors $y_i(x)$ data of length p , $i = 1 \dots S$, with S the number of samples. The samples are obtained from n subjects each sampled at k points in time, and assume for simplicity that $S = n \times k$. The length of y , p is the number mass over charge positions (m/z) found in the

run of that sample on the mass-spectrograph machine. In the particular experiment, each of the subjects were assigned to one of the available treatments at time t_{treat} , which were active drug(s) or placebo. We assume that the x -values are in one way or the other aligned, such that we effectively deal with one vector x . If a x peak position was not observed in a particular sample, either a zero or the mean of $y_i(x)$ from neighbouring x values are imputed for the corresponding y , see Ramsay and Silverman (2005) chapter 7.

2.1 Basic FDA

Basically FDA starts by choosing a set of basis functions with its domain in the x -values. This basis must suit the given data, e.g. B-splines or a Fourier-basis. We used B-splines, which can be smooth as well as peaked - depending on the choice of internal knots and the order.

Suppose we choose a finite set of B-splines $\phi_k(x)$, $k = 1 \dots K$ of given order m and ordered knots τ . We assume that there exist scalars c_k , $k = 1 \dots K$ such that

$$E[y(x)] = \sum_{k=1}^K c_k \phi_k(x) = \mathbf{c}'\boldsymbol{\phi}$$

The approximation of the $y(x)$ vector is now a matter of linear regression on the basis functions, yielding scalars \hat{c}_k , $k = 1 \dots K$. Using simple weighted least squares the coefficients are given by:

$$\hat{\mathbf{c}} = (\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}'\mathbf{W}'\mathbf{y}$$

where \mathbf{W} is the matrix of weights. The projection operator is given by:

$$\mathbf{S}_\phi = \boldsymbol{\Phi} (\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}'\mathbf{W}'$$

thus $\hat{\mathbf{y}} = \mathbf{S}_\phi\mathbf{y}$. To escape from the crude approximation using the basis functions alone, some 'roughness' penalty function on the coefficients is introduced (Ramsay and Silverman, 2005). Suppose we can write this function as $\hat{\mathbf{c}}'\mathbf{R}\hat{\mathbf{c}}$ where \mathbf{R} is a matrix derived from the basis functions. Ramsay proposes a penalty function being the integral over the squared m th derivative of \mathbf{y} , for some integer m . This derivative can be approximated entirely by the basis functions and the unknown coefficients \mathbf{c} . E.g. $\mathbf{R} = \int D^m\boldsymbol{\phi}(s) D^m\boldsymbol{\phi}'(s) ds$. For a fixed penalty λ , we fit the coefficients, getting the *smoothed* fitted values $\hat{\mathbf{y}} = \mathbf{S}_{\phi,\lambda}\mathbf{y}$ where the modified 'projector' is:

$$\mathbf{S}_{\phi,\lambda} = \boldsymbol{\Phi} (\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \lambda\mathbf{R})^{-1} \boldsymbol{\Phi}'\mathbf{W}'$$

the familiar penalized regression smoother. The degrees of freedom of this smoother is usually defined as $df = \text{trace}(\mathbf{S})$. The choice of the penalty is such that bias and variance is traded off in a reasonable way. One way to do this efficiently is the use of Generalised Cross Validation (GCV).

In the following we use \mathbf{Y} being a $p \times S$ matrix with columns the S observed functions, $\boldsymbol{\Phi}$ the $p \times K_y$ matrix of basis functions $\boldsymbol{\phi}$ and the $S \times K_y$ matrix \mathbf{C} of coefficients and we summarize $\mathbf{Y} = \mathbf{C}\boldsymbol{\Phi}$.

2.2 First step: select ranges of interesting peaks by linear FDA

As a first step in the analysis we apply a functional linear model on the data, in which we consider all samples as independent units, thus without worrying about correlation of the *functions* $y(x)$ within subjects. Here we consider a (non-functional) design matrix \mathbf{X} of size $S \times q$ for which we have to estimate the q functional coefficients $\beta(x)$ for the model:

$$\mathbf{Y}(x) = \mathbf{X}\beta(x) + \varepsilon(x)$$

Although not strictly necessary, we assume that all functions $\beta(x)$ use the same basis Ψ , and in this particular example we even take Φ identically to the Ψ of y . Now, suppose we have a matrix \mathbf{B} of size $q \times K_\beta$ such that $\beta = B\varphi$. In order to get smooth coefficient functions $\hat{\beta}(x)$ we *penalize* these coefficient functions (in addition to the smoothing of the observations). Thus we find $\hat{\beta}(x)$ by solving the columns of matrix \mathbf{B} .

$$\text{vec}(\mathbf{B}) = [\mathbf{J}_{\phi\phi} \otimes \mathbf{X}'\mathbf{X} + \mathbf{R} \otimes \mathbf{\Lambda}]^{-1} \text{vec}(\mathbf{X}'\mathbf{C} \mathbf{J}_{\phi\phi})$$

where $\mathbf{J}_{\phi\phi} = \int \Phi'\Phi$. The diagonal matrix $\mathbf{\Lambda}$ contains the penalties for the different coefficients β . When using a different basis for the latter, a similar equation apply, with $\mathbf{J}_{\phi\phi}$ at the right hand side replaced by $\mathbf{J}_{\phi\psi}$.

For this specific example \mathbf{X} includes time and treatment and its interaction. Variance and covariance of the coefficients are readily obtained, and point wise confidence intervals for *pre-defined* contrasts can be computed and visualized. Multiplicity for each contrast is corrected with a crude Bonferoni correction using as denominator K_y , the number of basis functions used in \mathbf{Y} . All ranges of smoothed contrasts outside the confidence intervals were regarded as interesting. The individual peaks within each ranges were kept for analysis in the second step.

2.3 Second step: mixed models peak within interesting ranges

In this step we used a mixed model analysis for each of the peaks in which we took care of the correlation within subjects. Fixed effects were again time, treatment and interaction between time and treatment. Random subject effects allow individual values for the intercept. For each peak we tested relevant contrasts using (a modified version) of method of Smyth (2004) developed for the analysis of microarray data with *fixed* linear models. To correct for multiplicity we used the adaptive Benjamini-Hochberg method (Yekutieli *et al.*, 2006).

Thus FDA is essentially used as a filter to reduce the amount of computation, and hopefully strengthen the power of the study by reducing multiplicity. However, it is expected that at least theoretically this two-step procedure might be improved by using a mixed model in the FDA itself (Morris *et al.*, 2006).

3 Use of mixed FDA (mFDA)

In principle the FDA can be extended for mixed linear models. Consider a linear mixed functional model (Morris *et al.*, 2006):

$$\mathbf{Y}(x) = \mathbf{X}\beta(x) + \mathbf{Z}u(x) + \varepsilon(x)$$

Now the $S \times q$ matrix \mathbf{X} defines the fixed effects, while the $S \times p$ matrix \mathbf{Z} defines the random effects. The matrix β contains the coefficient functions of the fixed model. The (nuisance) matrices u and ϵ are independent of each other, but there might be a non-trivial covariance structure for each separately. Apparently there is a strong connection between linear mixed models and penalized regression for non-functional data (Pinheiro and Bates, 2000). In a mixed model the coefficients \mathbf{u} are derived by a penalized regression of the *fixed* residuals $\mathbf{r} = \mathbf{Y} - \mathbf{X}\beta$ on \mathbf{Z} with weights \mathbf{W} the representing the co-variance of the coefficients \mathbf{u} apart from a constant σ^2 . In particular

$$\hat{\mathbf{u}} = [\mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{I}]^{-1} \mathbf{Z}'\mathbf{W} \hat{\mathbf{r}}$$

and the ML-estimates of the fixed coefficients are obtained by regression:

$$\hat{\beta} = [\mathbf{X}'\Sigma^{-1}\mathbf{X}]^{-1} \mathbf{X}'\Sigma^{-1}\mathbf{Y}$$

with $\Sigma = \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{I}$. The ML-estimate of the residual variance is $\hat{\sigma}^2 = \mathbf{r}'\Sigma^{-1}\mathbf{r}/S$. Numerical estimation is usually done by a combination of EM and Newton-Raphson (Pinheiro and Bates, 2000).

For functional mixed effects analysis we extend this algorithm to be used for functional data. If \mathbf{D} is the coefficient matrix for the random effects \mathbf{u} we may use the estimation of $\text{vec}(\mathbf{D})$ analogous to the estimation of $\text{vec}(\mathbf{B})$ from section 2.2.

$$\text{vec}(\mathbf{D}) = [\mathbf{J}_{\phi\phi} \otimes \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{R} \otimes \mathbf{I}]^{-1} \text{vec}(\mathbf{Z}'\mathbf{W}(\mathbf{C} - \mathbf{X}\mathbf{B})\mathbf{J}_{\phi\phi})$$

and for $\text{vec}(\mathbf{B})$ a similar equation holds.

In Ramsay and Silverman (2005), multivariate linear analysis is explored without making difference between the nuisance and fixed parameters, by functional principal component analysis. Before using mFDA, one might use this technique for exploring the correlation structure in the functional residuals \mathbf{r} of each subject, which may facilitate in the choice of the weight matrix \mathbf{W} .

References

- Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR (2006). Bayesian Analysis of Mass Spectrometry Proteomics Data using Wavelet Based Functional Mixed Models, University of Texas, MD Anderson Cancer Center, working paper series, paper 22.
- Pinheiro JC, and Bates DM (2000), *Mixed-effects models in S and S-Plus*, Springer Verlag
- JO Ramsay and BW Silverman (2005), *Functional Data Analysis*, Springer Verlag
- Smyth GK (2004), Linear Models and empirical bayes methods for assessing differential expression in Microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, **3**:article 3
- Yekutieli D, Reiner A, Elmer GI, Kafkafi N, Letwin N, Lee NH, and Benjamini Y (2006), Approaches to multiplicity issues in complex research in microarray analysis, *Statistica Neerlandica*, **60**(4), 414-437