Chapter 4

Functional ANOVA with random patient effect: an application to event-related potential modelling

Part of the joint research with Ph. LAMBERT, Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis, to appear in Statistics in Medicine (2006) and published in *Discussion paper number* DP0428 (2004), Institut de Statistique, Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium.

4.1 Introduction

The analysis of longitudinal curve data is a methodological and computational challenge for statisticians. Such data are often generated in biomedical studies. Most of the time, the statistical analysis focuses on simple summary measures, thereby discarding potentially important information. One way to model these curves is to use non parametric regression techniques based on splines. Because of its simplicity

and flexibility, penalized spline regression is a popular smoothing technique. It uses a fixed spline basis and is associated with a penalty for model complexity. This penalty is often based on the second derivative of the fitted function (Schumaker, 1981). As an alternative, Eilers and Marx (1996) proposed to use P-splines: a combination of B-splines and difference penalties on coefficients. We shall present this approach when the curves to model are extracted from electroencephalogram (EEG) data and apply it on P300 curves from the clinical study described in Chapter 1.

This chapter is organized as follows. In Section 4.2, we start by introducing notations for P-splines smoothing and P-splines regression. Next, we apply functional ANOVA with P-splines to the analysis of EEG. A description of the study is given in Chapter 1, where an interesting characteristic of EEG signals named event related potential (ERP) and in particular the P300 peak are introduced. Section 4.3 discusses limitations of standard analyses. In Section 4.4, we introduce a mixed model with random subject effect to analyse the ERP curves. Estimation of the fixed effects, prediction of the random effects and specification of the smoothing parameters are described. Pointwise and simultaneous confidence bands are obtained with model selection. Some computational issues are also discussed. Finally, we apply functional ANOVA with P-splines to the analysis of EEG in Section 4.5.

4.2 Spline smoothing and regression

4.2.1 Literature review

There are several competing approaches to non parametric modelling: serial-based smoothers, including wavelets (Tarter and Lock, 1993); kernel methods (Silverman, 1986; Härdle, 1990), including local regression (Wand and Jones, 1995; Fan and Gijbels, 1996); and splines methods, including smoothing splines (e.g. Eubank, 1988; Wahba, 1990; Green and Silverman, 1994), regression splines (Friedman, 1991; Stone et al., 1997) and B-splines (de Boor, 2001; Dierckx, 1995). The nature of the data

should play a role to select a method.

There is a growing interest in modelling longitudinal data with curves, in an ANOVA setting. Ramsay and Silverman (1997) discuss functional ANOVA in their book. Brumback and Rice (1998) and Rice an Wu (2001) present interesting applications. All these papers use spline regression. The approach to regression used in this paper is based on B-splines.

4.2.2 P-spline smoothing and functional ANOVA with P-splines

B-splines are attractive for non parametric modelling, but the choice of the optimal number of knots and of their positions can be difficult. A solution is to use a large number of equidistant knots and a penalty to control the smoothness of the fitted curve (see O'Sullivan, 1986 and 1988). Eilers and Marx (1996) proposed to use B-splines with a difference penalty on the spline coefficients.

Suppose that we observe a smooth function $y(t)$ at $t_1, t_2, \ldots, t_N$, yielding N pairs of observations $\{(t_n, y_n) : n = 1, \ldots, N\}$. We approximate $y(t)$ by

$$\sum_{s=0}^{S} \beta_s \phi_s(t) \tag{4.1}$$

where $\phi_0, \phi_1, \ldots, \phi_S$ are known functions of t and $\beta = (\beta_0, \ldots, \beta_S)^T$ is an appropriate coefficients vector. The basis functions ϕ_s might be B-splines, tensor products of B-splines, thin plate regression spline basis functions, the truncated power basis for cubic splines, or some radial basis functions (see Ruppert, Wand and Carroll, 2003 for examples) . Here, we use B-spline basis functions.

A B-spline of degree q is a piecewise polynomial function with $q + 1$ pieces of degree q. At the q joining knots, derivatives up to order $q - 1$ are continuous (see de Boor, 2001 and Dierckx, 1995).

The coefficients β in Equation (4.1) can be estimated by minimizing:

$$\sum_{n=1}^{N}\left(y_n - \sum_{s=0}^{S}\beta_s\phi_s(t_n)\right)^2 + PEN(\lambda), \tag{4.2}$$

where $PEN(\lambda)$ is a roughness penalty depending on a smoothing parameter λ. Examples of this penalty are the thin plate spline penalty, the integrated square of second derivative penalty for cubic spline and the various difference penalties used in P-spline smoothing. Eilers and Marx introduced a penalized B-spline regression model with equidistant knots associated to a convenient roughness penalty, and named P-splines. It is based on the difference of adjacent B-spline coefficients yielding the penalized likelihood:

$$\sum_{n=1}^{N}\left(y_n - \sum_{s=0}^{S}\beta_s\phi_s(t_n)\right)^2 + \lambda\sum_{s=d+1}^{S}\left(\Delta^d\beta_s\right)^2, \tag{4.3}$$

where Δ^d is the difference operator of order d.

In the case of longitudinal data, we have several curves $y_i(t)$ $i = 1, \ldots, I$ observed at multiple occasions, say $t = (t_1, t_2, \ldots, t_N)$. We shall now use multiple regression to model these curves and their association to factors. For example, suppose that P binary covariates $(x_{i1}, \ldots, x_{iP})$ are associated with the ith curve (The model can be extended to non-binary variables.). Each curve can be decomposed as the sum of several curves: one reference curve $f(t)$ that we shall name the "mean curve" and P corrections curves $g_p(t)$ $(p = 1, \ldots, P)$ (named "effect curves") summarizing the effect of each factor:

$$E(Y_i(t_n)) = \mu_i^n = f(t_n) + \sum_{p=1}^{P} x_{ip}g_p(t_n). \tag{4.4}$$

These curves can be modelled using B-splines (see Hastie and Tibshirani, 1991).

As with spline smoothing, the evolution of the curvature of a B-spline fit is strongly dependent on the number and the positions of the knots. Instead of trying to optimise the number and the position of the knots (which is a difficult numerical problem), we use a relatively large number of equidistant knots and avoid overfitting by adding a difference penalty on the coefficients of adjacent B-splines. This is the P-spline approach recommended by Eilers and Marx (1996).
The full model for a B-splines basis $\{\phi_s(t) : s = 1, \ldots, S\}$ is, for curve $y_i(t)$,:

$$E(y_i(t)) = f(t) + \sum_{p=1}^{P} x_p g_p(t) = \sum_{s=1}^{S} \beta_{0s}\phi_s(t) + \sum_{p=1}^{P} \left(x_{ip} \sum_{s=1}^{S} \beta_{ps}\phi_s(t) \right). \tag{4.5}$$

If Y is the matrix such that: $Y_{in} = y_i(t_n)$, then Equation (4.5) can be written as:

$$E(Y) = XB\Phi^T, \tag{4.6}$$

where X is the matrix indicating presence or absence of factors of interest ($p = 1, \ldots, P+1$) for each curve ($i = 1, \ldots, I$), Φ is the matrix associated to the B-spline basis (such that $\Phi_{ns} = \phi_s(t_n)$, $s = 1, \ldots, S$; $n = 1, \ldots, N$) and B is the matrix of B-spline coefficients (such that $B_{ks} = \beta_{k-1,s}$, $k = 1, \ldots, P+1$; $s = 1, \ldots, S$).
In the following, we shall transform the notations of the above problem to obtain an index notation which conveys the same information but which is intended for mathematical convenience rather than machine calculation. We define the vectorial form of this model by using the vec(.) operator defined by: $\boldsymbol{x} = \text{vec}(X)$ such that $\boldsymbol{x}_{(j-1)N+i} = X_{ij}$ $(i = 1, \ldots, I;\ j = 1, \ldots, J)$.
For the above spline regression model, we have:

$$\text{vec}(E(Y)^T) = \text{vec}\left((XB\Phi^T)^T\right) = \text{vec}\left(\Phi B^T X^T\right). \tag{4.7}$$

Henderson and Searle (1981) showed that

$$\text{vec}\left(\Phi B^T X^T\right) = (X \otimes \Phi)\,\text{vec}(B^T), \tag{4.8}$$

where $(X \otimes \Phi)$ is the Kroenecker product of the matrices X and Φ. Therefore, Equation (4.5) can be rewritten as

$$E(\boldsymbol{y}) = \widetilde{X}\boldsymbol{b}, \tag{4.9}$$

with $(X \otimes \Phi) = \widetilde{X}$ and $\text{vec}(B^T) = \boldsymbol{b}$.

If D is the difference matrix associated to the difference operator of order d such that $\sum_{s=d+1}^{S} \left(\Delta^d \beta_{ps}\right)^2 = \boldsymbol{\beta}_p^T D^T D \boldsymbol{\beta}_p$, than the estimates of the spline parameters will be obtained by minimizing

$$(\boldsymbol{y} - \widetilde{X}\boldsymbol{b})^T(\boldsymbol{y} - \widetilde{X}\boldsymbol{b}) + \sum_{p=0}^{P} \lambda_p \boldsymbol{\beta}_p^T D^T D \boldsymbol{\beta}_p \tag{4.10}$$

where λ_p is the penalty parameter associated to p^{th} curve. The penalty term in that penalized likelihood can be rewritten as $\boldsymbol{b}^T \left(\Lambda_f \otimes D^T D\right) \boldsymbol{b}$ where $\Lambda_f = \text{diag}\left\{\lambda_0, \lambda_1, \ldots, \lambda_P\right\}$. Conditionally on the penalty parameters, we get the spline parameter estimators:

$$\hat{\boldsymbol{b}} = (\widetilde{X}^T\widetilde{X} + \Lambda_f \otimes D^T D)^{-1}\widetilde{X}^T\boldsymbol{y} \tag{4.11}$$

and the fitted values

$$\hat{\boldsymbol{y}} = \widetilde{X}\hat{\boldsymbol{b}} = \widetilde{X}(\widetilde{X}^T\widetilde{X} + \Lambda_f \otimes D^T D)^{-1}\widetilde{X}^T\boldsymbol{y} = H\boldsymbol{y}. \tag{4.12}$$

The model was presented in a least square setting, to keep the explanation simple. Because of its regression structure, it can be extended straightforwardly to a generalized additive model (see Eilers, 1999).

4.3 Results and limitations of a standard analysis

The classical approach consists in summarizing the P300 data of each subject into one statistic like the amplitude or the time of occurrence of the peak. The corresponding

summary data are then treated using analysis of (co)variance models to compare groups after correction for important covariates. In chapter 1 (Section 1.3.3), we have used an ANOVA model with random effects (Searle, 1971) to account for subject heterogeneity and to analyse the effects of treatments on the amplitudes and latencies of the P300 peaks. The considered fixed effects correspond to treatment, period of recording and their interaction. We found the same qualitative results for electrodes Fz and Cz with a significant decrease of the mean amplitude and a non-significant increase of the associated P300 peak under Lorazepam.

That analysis also pointed a significant evolution of these mean quantities with the time elapsed since treatment administration (= period of recording or shortly period) but no significant interaction between period and treatment.

The analysis of the coefficients shows that the amplitude (the latency) is lowest (largest) when Lorazepam reaches its peak plasmatic concentration and largest (lowest) at the end of the experiment.

The increase (decrease) of the mean amplitude (latency) starts approximately after theoretical Lorazepam concentration starts decreasing in the plasma (i.e. 3 hours after administration).

The problem with this method is that the amplitude and latency of the P300 peak are difficult to compute because the signal is discrete and noisy.

Here, we propose to use P-splines in a functional ANOVA framework to model the P300 curves under placebo and under Lorazepam. The effect of Lorazepam will be assessed from the whole curve and not only from the amplitude and latency of the P300 peak.

4.4 Modelling random patient effects

As shown in Section 4.2, each curve could be modelled as the sum of a mean curve and some other effect curves associated to P different factor levels. We shall also allow the functions to vary with the individuals seen as a random factor. We propose

the following functional ANOVA model, for each curve i:

$$E(Y_i(t)) = \mu_i = f(t) + \sum_{p=1}^{P} x_{ip} g_p(t) + h_{d(i)}(t), \tag{4.13}$$

where i indexes curves, p indicates the factor levels and $d(i)$ indicates the subject associated to curve i ($d(.) = 1, 2, \ldots, \delta$ subjects).
The functions $g_p(.)$ are associated to fixed factor levels; $h_{d(i)}$ correct the reference functional $f(t)$ to give the reference curve for subject $d(i)$. Semi parametric models for subject-specific curves are available in the literature. For example, Durban et al. (2005) suggested individual curves modelled as penalized splines with random intercept. In this paper, the random effect is a curve ($h_{d(i)}$) instead of a random intercept. We adopt linear combinations of B-splines to describe these functionals. The vector of B-splines coefficients associated to $h_{d(i)}$ is assumed to have mean zero as in classical random coefficients models.

Therefore, using the notations introduced in Section 4.2, we propose the linear mixed model formulation:

$$\boldsymbol{y} = \widetilde{X}\boldsymbol{b} + \widetilde{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{4.14}$$

with

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \widetilde{G} & 0 \\ 0 & \widetilde{R} \end{pmatrix} \right] \tag{4.15}$$

where $\widetilde{Z}\boldsymbol{u}$ corresponds to the individual random curves.

4.4.1 Estimation of fixed effects

One can show that the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim N(\widetilde{X}\boldsymbol{b}, V) \tag{4.16}$$

where $V = \widetilde{Z}\widetilde{G}\widetilde{Z}^T + \widetilde{R}$.
A direct minimization of $(\boldsymbol{y}-\widetilde{X}\boldsymbol{b})^T V^{-1}(\boldsymbol{y}-\widetilde{X}\boldsymbol{b})$ to obtain estimates for $\boldsymbol{b}$ is an ill-posed problem because $\widetilde{X}$ is not full rank (see Appendix D for example). A simple solution is to add a penalty $\kappa \boldsymbol{b}^T\boldsymbol{b}$ with a small value for κ (say 10^{-13}). This is the principle of Ridge regression, a biased estimation technique designed to reduce multicollinearity. The main drawback is that the value of the parameter κ could change from one application to another because too large value of κ could impact the results. For the application presented in Section 4.5, a sensitivity analysis didn't show an impact of κ for values in the range $10^{-15} - 10^{-10}$. Cross-validation could be used to choose a good Ridge parameter, but it is time consuming (Golub et al., 1979).

Another alternative is to change the design matrix $\widetilde{X}$ to obtain a well-posed problem. For example, we can reparametrize the matrix X by changing the coding: removing one column and coding the effects with 0, +1 and -1. For the reference level, the dummy variable have a value of -1. Parameter estimates using this coding scheme estimate the difference in the effect of each non reference level compared to the average effect over all levels instead of comparison to a specific level as we defined previously. It is automatically done for classical regression techniques by some software like SAS.

Combined with the roughness penalties for each of the constituting curves, we end up with the penalized log-likelihood

$$(\boldsymbol{y} - \widetilde{X}\boldsymbol{b})^T V^{-1}(\boldsymbol{y} - \widetilde{X}\boldsymbol{b}) + \boldsymbol{b}^T \left(\Lambda_f \otimes D^T D\right) \boldsymbol{b} + \kappa \boldsymbol{b}^T\boldsymbol{b} \tag{4.17}$$

where $\Lambda_f = \text{diag}\,\{\lambda_0, \lambda_1, \ldots, \lambda_P\}$ is the diagonal matrix of the smoothing parameters for the fixed effects curves. The corresponding estimates for $\boldsymbol{b}$ are

$$\hat{\boldsymbol{b}} = (\widetilde{X}^T V^{-1}\widetilde{X} + \Lambda_f \otimes D^T D + \kappa I)^{-1}\widetilde{X}^T V^{-1}\boldsymbol{y}, \tag{4.18}$$

and the associated fitted curves

$$\hat{\boldsymbol{y}} = \widetilde{X}\hat{\boldsymbol{b}} = \widetilde{X}(\widetilde{X}^T V^{-1}\widetilde{X} + \Lambda_f \otimes D^T D + \kappa I)^{-1}\widetilde{X}^T V^{-1}\boldsymbol{y} = H\boldsymbol{y}. \tag{4.19}$$

4.4.2 Prediction of random effects

We know that

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{u} &\sim N(\widetilde{X}\boldsymbol{b} + \widetilde{Z}\boldsymbol{u}, \widetilde{R}) \\ \boldsymbol{u} &\sim N(0, \widetilde{G}). \end{aligned} \tag{4.20}$$

The corresponding conditional log-likelihood function is:

$$-\frac{1}{2}\left[\log(|\widetilde{R}|) + \left(\boldsymbol{y} - \widetilde{X}\boldsymbol{b} - \widetilde{Z}\boldsymbol{u}\right)^T \widetilde{R}^{-1}\left(\boldsymbol{y} - \widetilde{X}\boldsymbol{b} - \widetilde{Z}\boldsymbol{u}\right)\right]. \tag{4.21}$$

Combined with the normal density of the random effects $\boldsymbol{u}$, we obtain the joint likelihood

$$-\frac{1}{2}\left[\log(|\widetilde{R}|) + \left(\boldsymbol{y} - \widetilde{X}\boldsymbol{b} - \widetilde{Z}\boldsymbol{u}\right)^T \widetilde{R}^{-1}\left(\boldsymbol{y} - \widetilde{X}\boldsymbol{b} - \widetilde{Z}\boldsymbol{u}\right) + \log(|\widetilde{G}|) + \boldsymbol{u}^T\widetilde{G}^{-1}\boldsymbol{u}\right] \tag{4.22}$$

from which one can derive the prediction for the random effects:

$$\hat{\boldsymbol{u}} = (\widetilde{Z}^T\widetilde{R}^{-1}\widetilde{Z} + \left(\Lambda_r \otimes D^T D\right) + \kappa I + \widetilde{G}^{-1})^{-1}\widetilde{Z}^T\widetilde{R}^{-1}(\boldsymbol{y} - \widetilde{X}\hat{\boldsymbol{b}}), \tag{4.23}$$

where $\Lambda_r = \lambda^{(r)} I_{(\delta)}$ and $\lambda^{(r)}$ is the common smoothing parameters for the random subjects effects curves.

Of note, $\hat{\boldsymbol{b}}$ and $\hat{\boldsymbol{u}}$ are the solution of the Best Linear Unbiased Prediction (BLUP) (Ruppert et al., 2003).

4.4.3 Selection of the penalty parameters

One common measure of "goodness of fit" is the residual sum of squares (RSS). However, since RSS is minimized at the interpolants, minimization of this criterion to

select the penalty parameters will lead to the spline fit that is closest to interpolation. For P-splines smoothing, this corresponds to zero for the smoothing parameters. Generalized cross-validation[1] gets around this problem. One can select the $P+2$ smoothing parameters $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_P, \underbrace{\lambda^{(r)}, \ldots, \lambda^{(r)}}_{\delta \text{ elements}})$ by minimizing the generalized cross-validation criterion:

$$GCV(\Lambda) = \sum_{i=1}^{I} \sum_{n=1}^{N} \left(\frac{y_i(t_n) - \hat{y}_i(t_n, \Lambda)}{n_y - \text{trace}(H(\Lambda))} \right)^2 \tag{4.24}$$

Of note, in the classical non penalized regression setting, the trace of the hat matrix H corresponds to the number of parameters being fit. The intuitive interpretation of $tr(H)$ is that the degree of freedom is the total influence of all observations: the greater the number of parameters, the greater aggregated influence of the data.

We use the Matlab function `fminsearch` to minimize $GCV(\Lambda)$. It relies on the simplex search method (see Lagarias et al., 1998). This is a direct search method that does not use numerical or analytic gradients.

The parameters involved in matrices $\widetilde{G}$ and $\widetilde{R}$ are estimated using their maximum likelihood estimators, as obtained by a numerical optimisation of the log-likelihood where the fixed and random effects are replaced by their conditional maximum likelihood estimators, see Sections 4.4.1 and 4.4.2. That conditional log-likelihood depends on the smoothing parameters and GCV on the current estimates of the parameters in the $\widetilde{G}$ and $\widetilde{R}$ variance-covariance matrices. So, a good practice is to alternate the estimation of the smoothing and of the variance parameters until convergence is attained. That procedure can be time consuming depending on the initial values.

[1] Leave-one-out cross-validation also gets around this problem. However, with huge quantity of data, the computation of the CV criterion is quite long. To decrease computational time, Ruppert et al. (2003) proposed to use GCV instead of CV.

ALGORITHM:

- *Initial conditions: Estimate the parameters in the* $\widetilde{G}$ *and* $\widetilde{R}$ *variance-covariance matrices for smoothing parameters equal to zero. The parameters involved in matrices* $\widetilde{G}$ *and* $\widetilde{R}$ *are estimated using their maximum likelihood estimators, as obtained by a numerical optimisation of the log-likelihood where the fixed and random effects are replaced by their conditional maximum likelihood estimators, see Sections 4.4.1 and 4.4.2. That means that we define* $\Lambda^{(0)} = 0$ *and compute corresponding* $\widetilde{G}^{(0)}$ *and* $\widetilde{R}^{(0)}$*. Set i=1.*

  At iteration i,

- *Step 1 : Find the optimal smoothing parameters* $\Lambda^{(i)}$ *by minimizing GCV for the current values of the variance-covariance matrices* $\widetilde{G}^{(i-1)}$ *and* $\widetilde{R}^{(i-1)}$*.*

- *Step 2 : Estimate the parameters in the* $\widetilde{G}^{(i)}$ *and* $\widetilde{R}^{(i)}$ *variance-covariance matrices for given smoothing parameters* $\Lambda^{(i)}$*.*

  Iterate until convergence.

In the examples that are considered, less than 10 iterations were required to reach convergence. See details about computation time for our application in Section 5.2.

4.4.4 Confidence bands

The distribution for $\hat{\boldsymbol{y}}$ is

$$\hat{\boldsymbol{y}} \sim N(\widetilde{X}\hat{\boldsymbol{b}} + \widetilde{Z}\hat{\boldsymbol{u}}, \text{var}(\widetilde{X}\hat{\boldsymbol{b}}) + \text{var}(\widetilde{Z}\hat{\boldsymbol{u}})), \tag{4.25}$$

with

$$\begin{bmatrix} \text{var}(\widetilde{X}\hat{\boldsymbol{b}}) \\ \text{var}(\widetilde{Z}\hat{\boldsymbol{u}}) \end{bmatrix} \approx C(C^T\widetilde{R}^{-1}C + B_G + \left(\Lambda \otimes D^T D\right) + \kappa I)^{-1}C^T$$

where $C = [\widetilde{X}, \widetilde{Z}]$, $\Lambda = diag(\lambda_0, \lambda_1, \ldots, \lambda_p, \underbrace{\lambda^{(r)}, \ldots, \lambda^{(r)}}_{\delta \text{ elements}})$ and

$$B_G = \begin{bmatrix} 0_{(S(P+1))} & 0 \\ 0 & \widetilde{G}^{-1} \end{bmatrix} \tag{4.26}$$

(see Appendix B for details).

So, a $100(1-\alpha)\%$ confidence interval for $E(Y_i(t_n))$ is:

$$\hat{\boldsymbol{y}}(N(i-1)+n) \pm z(1-\alpha/2)\sigma, \tag{4.27}$$

where σ is the standard deviation of $\hat{y}_i(t_n)$ estimated by:

$$\hat{\sigma}^2 \quad = c_i^n (C^T \widetilde{R}^{-1} C + B_G + \left(\Lambda \otimes D^T D\right) + \kappa I)^{-1} c_i^{nT}, \tag{4.28}$$

where c_i^n is the $(N(i-1)+n)$th row of C.
One can also compute confidence bands for each effect curve g_p associated to any level p of a factor: simply use Equation (4.27) by substituting to σ the standard deviation of $\hat{g}_p(t_n)$ estimated by:

$$\hat{\sigma} = \sqrt{c_{g_p}^n (C^T \widetilde{R}^{-1} C + B_G + (\Lambda \otimes D^T D) + \kappa I)^{-1} c_{g_p}^{nT}}, \tag{4.29}$$

where $c_{g_p}^{nT}$ is the nth row of the matrix $\left(e_{(p+1)} \otimes \Phi\right)$ with $e_{(p+1)}$ a column vector of length $(P+1+\delta)$ with 1 at the $(p+1)$th position and zeros elsewhere. Then, a $100(1-\alpha)\%$ *simultaneous* confidence band for $y_i(t_n)$ is

$$\hat{y}_i(t_n) \pm m_{1-\alpha}\hat{\sigma}, \tag{4.30}$$

where $m_{1-\alpha}$ is the $(1-\alpha)$ quantile of the random variable:

$$\sup_{t_n,\, n=1,\ldots,N} \left| \frac{\hat{y}_i(t_n) - y_i(t_n)}{\hat{\sigma}} \right| \approx \max_{n=1,\ldots,N} \left| \frac{(\widetilde{x}_i^n \left[\hat{\boldsymbol{b}} - \boldsymbol{b}\right])_n}{\hat{\sigma}} \right|, \tag{4.31}$$

where $\widetilde{x}_i^n$ is the $(N(i-1)+n)$th row of $\widetilde{X}$.

That quantile is obtained by simulating $\hat{\boldsymbol{b}}-\boldsymbol{b}$ using:

$$\hat{\boldsymbol{b}}-\boldsymbol{b} \sim N\left(0, (\widetilde{X}^T V^{-1} \widetilde{X} + \Lambda_f \otimes D^T D + \kappa I)^{-1}\right) \tag{4.32}$$

in combination with Equation (4.31). This process is repeated a large number of times, say 10,000. These simulated values are sorted from the smallest to the largest, and the one with rank $(1-\alpha)N$ is used to approximate $m_{1-\alpha}$.

Confidence intervals presented here are conditional on the choice of the smoothing parameters. A way to overcome this problem is to use bootstrap. However, this solution is not possible for many examples because of computational time. Another way to take into account the uncertainty of the smoothing parameters is to use Bayesian methods (Ruppert et al., 2003).

4.4.5 Model selection

One can select between 2 competing models using an F-test:

$$F_{obs} = \frac{(RSS_1 - RSS_2)/(\gamma_2 - \gamma_1)}{RSS_2/(N-\gamma_2)} \sim F_{\gamma_2-\gamma_1, N-\gamma_2} \tag{4.33}$$

where RSS indicates the residuals sum of squares, N is the number of observations and

$$\gamma_j = 2\ \text{trace}(H) - \text{trace}(HH^T) \tag{4.34}$$

The corresponding p-value is defined by:

$$\text{p-value} = 1 - P(X \leq F_{obs}), \tag{4.35}$$

where $X \sim F_{\gamma_2-\gamma_1,N-\gamma_2}$.

This can be improved by applying a correction to both the numerator and the denominator of F_{obs} (see Hastie and Tibshirani, 1991):

$$F_{obs} = \frac{(RSS_1 - RSS_2)/(\nu_1)}{RSS_2/(\delta_1)} \sim F_{\nu_1^2/\nu_2,\delta_1^2/\delta_2}, \tag{4.36}$$

where

$$\begin{aligned} R_j &= (I - H_j)^T(I - H_j) \ (j = 1, 2), \\ \delta_1 &= \text{trace}(R_2) \\ \delta_2 &= \text{trace}(R_2^2) \\ \nu_1 &= \text{trace}(R_1 - R_2) \\ \nu_2 &= \text{trace}(R_1 - R_2)^2, \text{ and} \end{aligned} \tag{4.37}$$

Notice that if the 2 corrections ν_1/ν_2 and δ_1/δ_2 are equal to one, then we recover the first approximation in Equation (4.33). These corrections are difficult to compute, but can be approximated (see Appendix C).

4.4.6 Computational details

The sizes of the data vector $\boldsymbol{y}$ and of the matrices generated to implement the model (matrices $\widetilde{X}$ and $\widetilde{Z}$) are huge in EEG analysis. Indeed, we have I curves to model ($I = 360$ in our application), each with N time points ($N = 128$ in our application). That means that the vector $\boldsymbol{y}$ is of length $n_y = I \times N$. Matrices $\widetilde{X}$ and $\widetilde{Z}$ have the same number of rows. In addition, we have P factor levels and δ subjects. That means that we obtain a design matrix $\widetilde{X}$ with columns containing P blocks of B-spline bases with S elements for each time point $t_n, n = 1, \ldots, n_y$ and a design matrix $\widetilde{Z}$ with $\delta \times S$ columns. These numbers of rows and columns are huge in EEG analysis.

Consequently, we decided to use Matlab to be able to deal with large matrices. Matlab provides functions to manipulate sparse matrix like $\widetilde{X}$ and $\widetilde{Z}$ using less memory, a decisive element to make the implementation of the above presented models feasible here. Moreover, Eilers (1999) presented formulas to exploit the sparse structure of the equations.

For the generalized cross-validation criterion (see Section 4.4.3) and the degree of freedom in the confidence bands computation (see Section 4.4.4), we have to compute the trace of the matrix:

$$H = C^T(C^T\widetilde{R}^{-1}C + B_G + \left(\Lambda \otimes D^TD\right) + \kappa I)^{-1}C^TR^{-1}, \tag{4.38}$$

such that:

$$\begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = C^T(C^T\widetilde{R}^{-1}C + B_G + \left(\Lambda \otimes D^TD\right) + \kappa I)^{-1}C^TR^{-1}\boldsymbol{y} = H\boldsymbol{y} \tag{4.39}$$

As $\text{trace}(AB) = \text{trace}(BA)$ (for conformable matrices), it is computationally advantageous to use:

$$\begin{aligned} \text{trace}(H) &= \text{trace}(C^T(C^T\widetilde{R}^{-1}C + B_G + \left(\Lambda \otimes D^TD\right) + \kappa I)^{-1}C^TR^{-1}) \\ &= \text{trace}((C^T\widetilde{R}^{-1}C + B_G + \left(\Lambda \otimes D^TD\right) + \kappa I)^{-1}C^TR^{-1}C). \end{aligned} \tag{4.40}$$

The latter expression involves a matrix with smaller dimensions than H since matrix C has more rows (n_y = length of $\boldsymbol{y} = N * I$) than columns ($(P + 1 + \delta)$ blocks $\times$ the number of elements in the B-spline basis ($= (P + 1 + \delta) * S$)). Thus C is a $(n_y \times (P + 1 + \delta)S)$ matrix yielding a H matrix of dimension $(n_y \times n_y)$ whereas $(C^T\widetilde{R}^{-1}C + B_G + \left(\Lambda \otimes D^TD\right) + \kappa I)^{-1}C^TR^{-1}C$ is only $((P+1+\delta)S \times (P+1+\delta)S)$ with $(P+1+\delta)S \ll n_y$.

Of note, Kroenecker products could be transformed into products of multidimensional matrices as proposed in Currie et al. (2006) and Eilers et al. (2006). This should probably improve computational time.

4.5 Application to event-related potentials modelling

4.5.1 Practical issues and model selection

As described in Chapter 1, we have $\delta = 15$ subjects receiving 2 treatments (placebo and Lorazepam) with recordings at 12 periods. After averaging of electrode Fz after stimuli, we obtained $I = 2 * 15 * 12 = 360$ observed ERP curves at $N = 128$ time points. Thus, the length of $\boldsymbol{y}$ is $n_y = 46,080$. We have $S = 34$ elements in the B-spline basis of degree 3 corresponding to 32 knots (= 1/4 of the 128 time points).

In our event-related potentials analysis, we have 2 fixed factors: treatment and period of recording after the treatment administration. With the placebo and the first period as reference, the corresponding model is:

$$E(Y_i(t)) = \mu_i = f(t) + \sum_{p=1}^{12} x_{ip} g_p(t) + h_{d(i)}(t), \tag{4.41}$$

where $i = 1, \ldots, 360$ (observed curves), $d(.) = 1, \ldots, \delta$ (15 subjects). g_1 is associated to the Lorazepam effect and $g_2, g_3, \ldots, g_{12}$ are respectively associated to period $2, 3, \ldots, 12$. That means that the "mean curve" $f(t)$ corresponds to the curve under placebo recorded at period 1 (for a 'typical' subject). The sum of the curves $f(t) + g_1(t)$ corresponds to the curve under Lorazepam recorded at period 1. To obtain the estimated curves at period p, we must still add the effect curve $g_p(t)$. Testing the above model, we saw that period effect is not significant. Moreover, as the drug concentration in the brain is evolving over time (i.e. with period), we also expect a treatment-period interaction. This treatment-period interaction is significant in

spite of the non significant period effect because there is an evolution over time under Lorazepam only (there is no significant period effect under placebo).

The 3 first periods (recorded before treatment administration) correspond to the same level of concentration of the drug. The 9 following periods correspond to 9 other levels of concentration of the drug in the brain. To facilitate the interpretation of the results, we shall use the concentration of Lorazepam as covariate instead of period in the above model. We will define a categorical concentration effect because it is much easier to compute and because we do not have any a priori information on the type of dependence (linear or much more complicated) between concentration and effect on the ERP curve. However, a further quantitative analysis could be interesting.

With the placebo and the first level of concentration of Lorazepam as reference, the model becomes:

$$E(Y_i(t)) = \mu_i = f(t) + \sum_{p=1}^{11} x_{ip} g_{p(i)}(t) + h_{d(i)}(t), \tag{4.42}$$

where g_1 is associated to the Lorazepam effect and $g_2, g_3, \ldots, g_{11}$ are respectively associated to the 10 different levels of concentration of Lorazepam. So, we have $P = 1 + 10 = 11$ factor levels. The "mean curve" $f(t)$ corresponds to the curve under placebo recorded at each period. The curve $g_1(t)$ corresponds to the effect of Lorazepam. To obtain the estimated curves for level p of concentration of Lorazepam, we must still add the effect curve $g_{p+1}(t)$. See Appendix D for construction of the design matrices.

The following block-diagonal forms are taken for $\widetilde{G}$ and $\widetilde{R}$:

$$\widetilde{G} = \left(I_{(\delta)} \otimes G\right), \tag{4.43}$$

where

$$G = \sigma_u^2 I_{(S)} = \sigma_u^2 I_{(34)} \tag{4.44}$$

and

$$\widetilde{R} = (I_I \otimes R) = (I_{360} \otimes R), \tag{4.45}$$

where R has an 'AR(1)' structure:

$$r_{ij} = \sigma_\epsilon^2 \rho^{|j-i|}. \tag{4.46}$$

This structure allows for correlation between measurements coming from the same curve.

Finally, we decided to use 3 smoothing parameters for the fixed effects and 1 smoothing parameter for the random subject effect (see Bugli and Lambert, 2004). For the fixed effects, Λ_f is defined by:

$$\Lambda_f = \text{diag}\{\lambda_0, \lambda_1, \underbrace{\lambda_2, \lambda_2, \ldots, \lambda_2}_{\text{10 elements}}\}, \tag{4.47}$$

with the 3 smoothing parameters λ_0, λ_1 and λ_2 used to add a penalty to the "mean curve", the effect curves corresponding to the effect of Lorazepam and the levels of concentration of Lorazepam respectively. We obtain:

$$\left(\Lambda_f \otimes D^T D\right) = \begin{bmatrix} \lambda_0 D^T D & 0 & 0 \\ 0 & \lambda_1 D^T D & 0 \\ 0 & 0 & \lambda_2 \left(I_{(10)} \otimes D^T D\right) \end{bmatrix} \tag{4.48}$$

For the random subject effect, Λ_r is defined by:

$$\Lambda_r = \text{diag}\{\underbrace{\lambda, \lambda, \ldots, \lambda}_{\delta \text{ elements}}\}, \tag{4.49}$$

with the same smoothing parameter λ for all the $\delta = 15$ subjects. We have:

$$\left(\Lambda_r \otimes D^T D\right) = \lambda \left(I_{(\delta=15)} \otimes D^T D\right) \tag{4.50}$$

4.5.2 Computation time and initial conditions

The problem with our application is the huge quantity of data to manage. Indeed, computation time increases with the size of the involved matrices. We took initial

conditions (that means initial values of smoothing parameters Λ and of parameters in the $\widetilde{G}$ and $\widetilde{R}$ variance-covariance matrices) equal to zero. Less than 10 iterations of the algorithm presented in Section 5.4.3 were required to reach convergence. The determination of the parameters takes less than 2 hours. This leads to quite acceptable computation times for moderately sized data sets. Moreover, the choice of better initial conditions can probably decrease computation time.

To check efficacy of the off-the shelf optimizer provided by Matlab, we tested different initial conditions. We computed GCV (or respectively likelihood) on a grid. We tested 10 values of each parameter that means 10^4 (or respectively 10^3) combinations of parameters. It allows to test different initial values for the optimisation with the Matlab function `fminsearch`. For our example, the determination of the parameters takes approximately 24 hours and provides the same values as before with zero initial values. That proves the accuracy of the optimizer.

4.5.3 Results

We model the P300 curve for the frontal electrode Fz for the 15 subjects under placebo and under Lorazepam during the 12 periods of recordings. Using model in Equation (4.42), we obtained the following estimated values of the parameters:

$$\hat{\sigma}_u^2 = 1.5125, \hat{\rho} = 0.9677, \hat{\sigma}_\epsilon^2 = 4.3307$$

Significant Lorazepam effect

For the above model, we tested the necessity to have a correction curve for treatment effect using the test presented in Section 4.4.5 and we obtained a correction factor ν_1/ν_2 larger than 1. That means that the uncorrected 95% F-test has a higher significance level that the desired one. The sum of square differences between groups equals $16,875$ with degree of freedom equal to 42.55 and the sum of square differences within groups equals $214,500$ with degree of freedom equal to $42,640$. We obtained

a value of F_{obs} equal to 78.83 which is much larger than the 95% quantile 1.33. This corresponds to a p-value smaller than 0.001 suggesting that the Lorazepam effect is significant. The correction factors ν_1/ν_2 and δ_1/δ_2 equal 1.3128 and 0.8774, indicating that the correct significant level is 96%. All the included effects were found to be highly significant.

Simultaneous confidence bands

For example, we obtained $m_{0.95} \approx 2.7001$ for the mean curve of the above model. That means that the simultaneous confidence bands are $\frac{2.7001}{1.96} = 1.38$ times wider than the pointwise intervals. The $m_{0.95}$ quantile can be approximated very accurately. Based on 100 independent simulations of 1000 draws each, we obtained the results summarized for some effect curves in Table 4.1.

effect	mean of $m_{0.95}$	standard deviation of $m_{0.95}$	mean ratio: $\frac{m_{0.95}}{1.96}$
mean	2.7001	0.0487	$\frac{2.7001}{1.96} = 1.38$
treatment	2.7083	0.0388	$\frac{2.7083}{1.96} = 1.38$
period 6	2.7292	0.0429	$\frac{2.7292}{1.96} = 1.39$
period 7	2.7178	0.0504	$\frac{2.7178}{1.96} = 1.39$
subject 1	2.7763	0.0360	$\frac{2.7763}{1.96} = 1.42$

Table 4.1: Mean of the estimations of $m_{0.95}$ for some of the effect curves, standard deviation and approximated ratio between simultaneous and pointwise confidence bands.

Values of $m_{0.95}$ are similar for other time effect curves and other subject effect curves. For each fitted curve $\hat{y}$, $m_{0.95}$ is always approximately 2.82. That means that the simultaneous confidence bands are $\frac{2.82}{1.96} = 1.44$ times wider than the pointwise intervals.

Estimated effect curves

We obtain the results summarized in Figure 4.1. The curve in Figure 4.1(a) is the

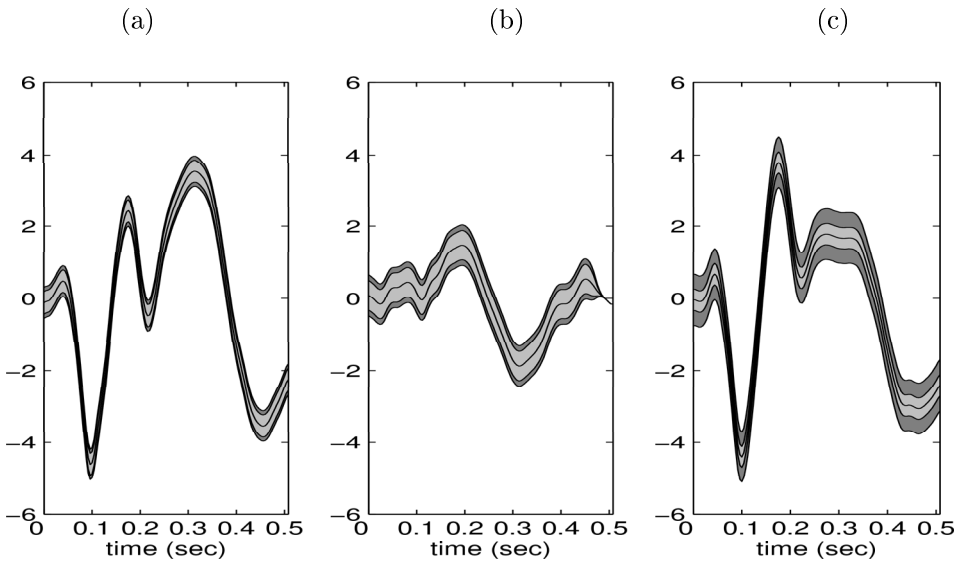

Figure 4.1: (a) Mean curve under placebo (solid line), pointwise confidence bands (light grey) and simultaneous confidence bands (dark grey). (b) Curve representing treatment effect (solid line), pointwise confidence bands (light grey) and simultaneous confidence bands (dark grey). (c) Mean curve under treatment (solid line), pointwise confidence bands (light grey) and simultaneous confidence bands (dark grey). We must still add the concentration effect.

mean curve under placebo as the non active treatment was taken to be the reference. We retrieve the specific pattern of an ERP curve with 3 characteristic positive peaks (P100, P200 and P300). The treatment effect is modelled by the curve in Figure 4.1(b). So, the mean curve under (the active) Lorazepam in Figure 4.1(c) is the sum of the mean curve in Figure 2(a) and the curve representing the treatment effect in Figure 4.1(b). This mean curve under Lorazepam is the average of the curves under Lorazepam for all the concentrations. We observe a significant decrease of the amplitude of the P300 peak and an increase of the P200 peak under Lorazepam. To obtain the mean curve under Lorazepam for a period, we must still add the

concentration effect curve corresponding to the level of concentration for this period (see Figure 4.2 for an illustration): the decrease of the amplitude of the P300 peak under Lorazepam is most important for level 5 (corresponding to the concentration of Lorazepam for period 7 i.e. 3.5 hours after Lorazepam administration).

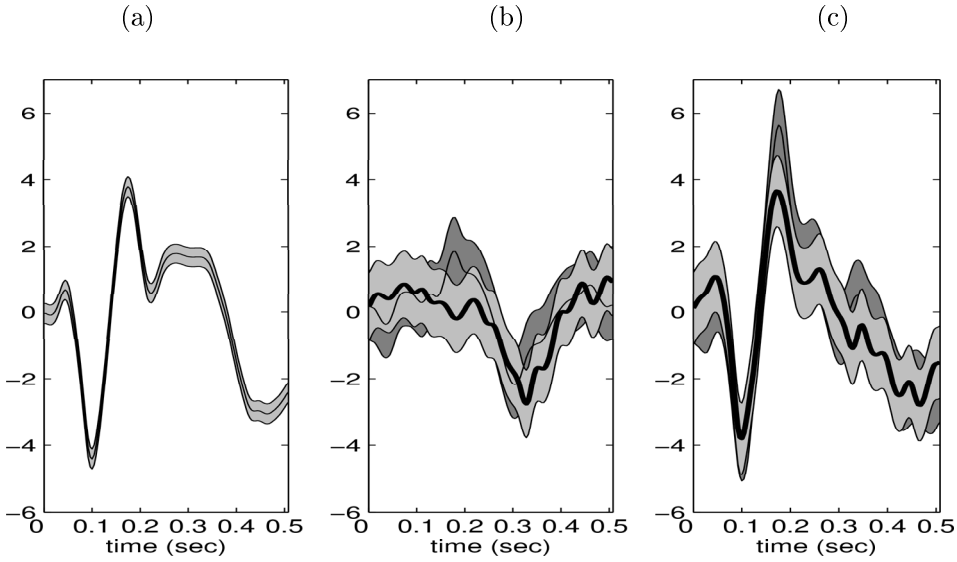

Figure 4.2: (a) Mean curve under treatment. (b) Concentration effect curves for level 4 (corresponding to the concentration of Lorazepam for period 6 i.e. 2.5 hours after Lorazepam administration, thin solid line) and level 5 (corresponding to the concentration of Lorazepam for period 7 i.e. 3.5 hours after Lorazepam administration, bold solid line) and the corresponding pointwise confidence bands. (c) The sum of the mean curve in (a) and the concentration effect curves in (b) gives the mean curve under Lorazepam for concentration level 4 (thin solid line) and 5 (bold solid line) and the pointwise confidence bands (dark grey for level 4 and light grey for level 5): the P300 peak almost disappears.

The last recording (period 12, level of concentration 10) is performed 25.5 hours after the Lorazepam administration. At that time, the concentration of Lorazepam (level 10) in the plasma is almost the same as during periods 1, 2 and 3 (level of concentration 1). This is confirmed in Figure 4.3 that shows very close concentration effects for levels 1 and 10 (corresponding to periods 1, 2, 3 and 12).

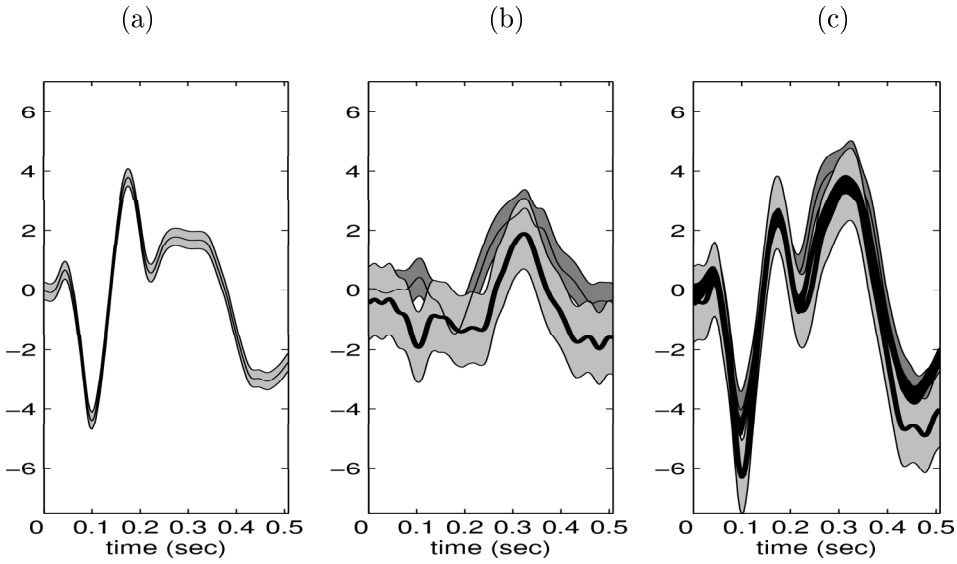

Figure 4.3: (a) Mean curve under treatment. (b) Concentration effect curves for level 1 (corresponding to the concentration of Lorazepam for period 1,2 and 3 i.e. before Lorazepam administration, thin solid line) and level 10 (corresponding to the concentration of Lorazepam for period 12 i.e. 25.5 hours after Lorazepam administration, bold solid line) and the corresponding pointwise confidence bands. (c) The sum of the mean curve in (a) and the concentration effect curves in (b) gives the mean curve under Lorazepam for concentration level 1 (thin solid line) and 10 (bold solid line) and the pointwise confidence bands (dark grey for level 1 and light grey for level 10): the curve is very similar to the mean curve under placebo (pointwise confidence interval in black).

Figure 4.2 suggests that Lorazepam effect decreases the amplitude of the P300 peak but also increases its latency. This effect on latency was not visible on the treatment effect curve because the decrease of amplitude on this curve correspond exactly to the time of the P300 peak on the mean curve. This effect of Lorazepam on latency requires further analysis because this effect is not included in the treatment main effect curve, but is only visible in the period by treatment effect curve.

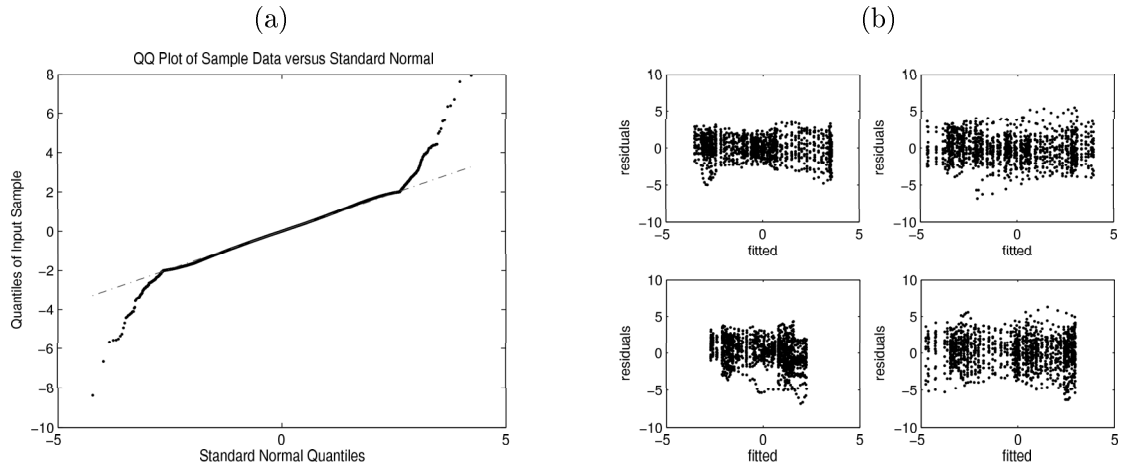

Figure 4.4: (a) Q-Q plot of the residuals: slight deviation from normality. (b) Plots of the residuals against predicted values for subject 1, 8 10 and 12: no evidence of non homogeneity of variance.

4.5.4 Diagnostics

One can check the assumptions underlying our model. The Q-Q plot in Figure 4.4(a) suggests a slight deviation from normality. Plots of the residuals against the predicted values do not show major abnormalities associated with any factor level: there is no evidence of non homogeneity of variance. Figure 4.4(b) shows some examples of these plots. There is no structure associated to any factor level.

Moreover, we wondered if there are outliers curves. Considering a curve containing an outlier point as an outlier curve would probably be too restrictive. The adaptation to functional ANOVA method of outliers detection criteria would require extra developments. However, we found (using leave one out) that no simple data curve had an important effect on the fitted curves.

4.5.5 Interpretation and discussion

The reduction of the amplitude of the P300 peak under Lorazepam is not surprising. Indeed, the amplitude of the P300 peak is related to the process of revision of the

representations kept in working memory. So, the amplitude decreases when there is lesser update of the working memory (see Polish and Kok, 1995). In that sense, Lorazepam would induce cognitive impairment. This result was already confirmed in another analysis of these data where the amplitude and the latency of the peak were directly modelled (Bugli et al., 2004).

We can conclude that functional ANOVA with P-splines is a powerful tool to model curves. Indeed, the model can be easily and clearly stated. In the field of EEG analysis, it allows to bring information about the modification in shape of the peak and not only about some specific characteristics like amplitude or latency of one particular peak that are difficult to identify from noisy EEG signals.

Functional ANOVA using spline is already used to model curves with individual behaviour. For example, Durban et al. (2005) suggested individual curves modelled as penalized splines with random intercept. The originality of our work is that the random effect is a curve instead of a random intercept.

Future work could be the improvement of selection of the smoothing parameters. Bayesian estimation of Λ using linear mixed models (see Wand, 2003) is probably the solution to take into account efficiently the uncertainty of the smoothing parameters in the computation of the confidence intervals.

Another improvement would be to combine registration and functional ANOVA. Problem of identifiability has to be solved. Of note, we tried to apply the same analysis on ICA components (P3a and P3b). In this case, use of ICA component do not give more information: results are the same. The reason is that P300 peak is already easily detected on electrodes Fz. However, registration of ICA component could be useful for other application where the peak of interest is not well identify on electrodes.