USING FUNCTIONAL APPROACHES FOR LONGITUDINAL DATA ANALYSIS

Mozhgan Taavoni – Mahnaz Khalafi – Majid Azimmohseni

Abstract

Longitudinal studies have been highly developed in many scientific research areas such as economic, biomedical researchs and so on. The longitudinal data are mainly resulted from the observations of subjects (human beings, animals, or laboratory samples, etc.), which are measured repeatedly over some period of time.

The classical analysis of longitudinal studies is based on parametric models including marginal models, random effects modals and transition models. Nevertheless, in the recent years, functional data analysis provides a nonparametric approach for the analysis of longitudinal data which observations of the same subject are viewed as a sample from a functional space. In a longitudinal data analysis, one mostly deals with sparsely and irregularly observed data that also corrupted with noise. In contrast, classical functional data analysis requires a large number of regularly spaced measurements per subject. Adjustments of functional data analysis techniques which take these particular features into account are needed to use them for longitudinal data. We review some techniques that have been recently proposed to connect functional data analysis with longitudinal data such as, functional principal components and functional linear regression models. Then performance of these methods are illustrated with real data.

Key words: Longitudinal data, functional data analysis, functional principal component analysis, functional linear models

JEL Code: C23, C33

Introduction

In the last decade, Longitudinal Data Analysis(LDA) has been widely studied in the fields of clinical trials, medicine, social sciences, economy and etc. Longitudinal studies are characterized by data records containing repeated measurements per subject, measured at various points on a suitable time axis. The aim is often to study change over time or time-

dynamics of biological phenomena such as growth. One is also interested in relating these time-dynamics to certain predictors or responses. Therefore LDA represents a connection between regression (cross-sectional) and time series analysis. But longitudinal data have some special properties. Unlike the regression analysis, subjects observed over time and unlike the time series analysis, the data consists of many subjects. Another important property of longitudinal data is that unlike the classical analysis, all of the observations on the same subject tend to be correlated. In such studies, the possible correlations between responses given by the same individual need to be taken into account in order to produce proper analysis. Various models can be used to handle such correlations. One approach is using marginal modeling, which allows for inferences about parameters averaged over the whole population(Liang et al., 1992; Molenberghs and Leasaffre, 1994). Another approach is making use of random effects modeling, which deliberately provide inferences about variability between respondents (Verbeke and Lesaffre, 1996; Diggle et al., 2002). Another appropriate approach is to investigate the reasons for the change of the responses is the use of transition(Markov) models (Reuter et al., 2004; Chung et al., 2005).

The parametric assumptions that are made in these models cause incompatible and complex relationships. Introducing nonparametric components can ameliorate the difficulties of relating various longitudinal models to each other, as it increases the inherent flexibility of the resulting longitudinal models substantially. Taking the idea of modeling with nonparametric components one step further, the Functional Data Analysis (FDA) approach to LDA provides an alternative nonparametric method for the modeling of individual trajectories.

FDA provides an inherently nonparametric approach for the analysis of data which consist of samples of time courses or random trajectories. It is a relatively young field aiming at modeling and data exploration under very flexible model assumptions with no or few parametric components. Basic tools of FDA are smoothing, functional rincipal omponents and functional linear model(Ramsay and Silverman, 2002). While in the usual FDA paradigm the sample functions were considered as continuously observed, in LDA one mostly deals with sparsely and irregularly observed data that also are corrupted with noise. Adjustments of FDA techniques which take these particular features into account are needed to use them for LDA. We review some techniques that have been recently proposed to connect FDA methodology with LDA. The extension of FDA towards LDA is a fairly recent undertaking that presents a promising method for future researchs(Rice, 2004; Zhao et al., 2004 ; and Muller, 2005).

The paper is organized as follows. In section 1 we shall shortly review the specification of three basic models for LDA. Then, we will introduce the connections between FDA with LDA in section 2. Next, in section 3 we also illustrate and compare classical and functional model for a longitudinal study by an analysis of the dynamic relationship between viral load and CD4 cell counts observed in AIDS clinical trials.

1 Review of three basic models for longitudinal data analysis

In this section, we introduce three model strategies that are commonly used to LDA: marginal models, mixed effects models, and transition models. Let us first provide some notation. We assume that *n* subjects are measured repeatedly over time whic $y_{ij}; i = 1,2,\ldots,n; j = 1,2,\ldots,t_i$ is the response variable on j^{th} time order for the i^{th} subject. Each response y_{ij} is associated with a $p \times 1$ vector of covariates, x_{ij}, through the period of study that may change over the time. For example, it can be the age of subject (time varying covariate) or gender of subject (time stationary covariate). Now we introduce marginal, random effects and transition models.

1.1 Marginal models

In marginal model, the relation between response and explanatory variables is modelled separately from within-person correlation. The marginal expectation of the response, $E(y_{ij}) = \mu_{ij}$, depends on the covariates, x_{ij}, through a link function *g* as follows:

$$g[E(y_{ij} \mid x_{ij})] = g(\mu_{ij}) = x_{ij}^{'}\beta . \qquad (1)$$

The marginal variance of the response depends on the marginal mean:

$$\mathrm{Var}(y_{ij}) = \phi v(\mu_{ij}), \qquad (2)$$

where *v* is a known function and the scale parameter ϕ may also depends on some covariates. Also, the correlation between y_{ij} and y_{ik} is a function of the marginal mean:

$$\mathrm{Corr}(y_{ij}, y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha), \qquad (3)$$

where ρ is a known function and the correlation parameters α may depend on covariates. In this model consistensy and effeciency of estimators depends on choice of correlation structure, that is limitation of this model.

1.2 Random-effects models

The basic idea underlying a random-effects model is that there is a natural heterogeneity across individuals in their regression coefficients and that the heterogeneity can be handled by a probability distribution. Correlation among observations for the same individual arises from their sharing unobservable variables, b_i . The model can be given by:

$$g[E(y_{ij} \mid x_{ij}, b_i)] = g(\mu_{ij}) = x_{ij}^{'}\beta + z_{ij}^{'}b_i , \qquad (4)$$

where z_{ij} is a subset of covariates, ie, x_{ij} . Typically, under a random-effects model, the y_{ij} , given b_i , are conditionally independent over time. Correlation among the responses from the same subject arises from their sharing unobservable variables, i.e., random effects. Generally the random effects are assumed to be multivariate Gaussian with mean 0 and variance-covariance matrix D . This assumption is limitation of this model.

1.3 Transition model

Under a transition model, correlation among the responses of same subjects exists because the past response values explicitly influence on the present observation. In this model, the response variable at time t depends on the response at time lags $t-1, t-2, \ldots, t-q$. Correlation between responses is considered by direct effects of *q* past values on present and past values are considered as additional predictive variables(covariates). Let $H_{ij} = (y_{i1}, y_{i2}, \ldots, y_{ij-1})$ denote the history of y_{ij} . The model can be represented as:

$$g[E(y_{ij} \mid x_{ij}, H_{ij})] = g(\mu_{ij}) = x_{ij}^{'}\beta + \sum_{r=1}^{s} f_r(H_{ij}, \alpha) \qquad (5)$$

where f_r. is a known function of history. Limitation of this model is that the estimation of parameters depends on this function f_r .

2 Functional linear regression models for longitudinal data

We propose nonparametric methods for functional linear regression which are designed for sparse longitudinal data, where both the predictor and response are functions of a covariate such as time. Predictor and response processes have smooth random trajectories, and the data consist of a small number of noisy repeated measurements made at irregular times for a sample of subjects. In longitudinal studies, the number of repeated measurements per subject is often small and may be modeled as a discrete random number and, accordingly, only a finite and asymptotically nonincreasing number of measurements are available for each subject or experimental unit. We propose a functional regression approach for this situation,

using Functional Principal Component Analysis(FPCA), where we estimate the Functional Principal Component(FPC) scores through conditional expectations. This allows the prediction of an unobserved response trajectory from sparse measurements of a predictor trajectory. The resulting technique is flexible and allows for different patterns regarding the timing of the measurements obtained for predictor and response trajectories. Asymptotic properties for a sample of n subjects are investigated under mild conditions, as $n \to \infty$, and we obtain consistent estimation for the regression function. In addition to convergence results for the components of functional linear regression, such as the regression parameter function, we construct asymptotic pointwise confidence bands for the predicted trajectories. A functional coefficient of determination as a measure of the variance explained by the functional regression model is introduced, which extend the standard R^2 to the functional case.

2.1 Functional principal components

FPCA has emerged as a major tool for dimension reduction within FDA. One goal is to summarize the infinite-dimensional random trajectories through a finite number of FPC scores. This method does not require distributional assumptions and is merely based on first and second order moments. An important application is a representation of individual trajectories through an empirical Karhunen-Loeve representation. It is always a good idea to check and adjust for smoothing before carrying out an FPCA.

The underlying but unobservable sample consists of pairs of random trajectories $(X_i, Y_i), i = 1, 2, \ldots, n$, with square integrable predictor trajectories X_i and response trajectories Y_i. These are realizations of smooth random processes (X, Y) with unknown smooth mean functions $EY(t) = \mu_Y(t), EX(s) = \mu_X(s)$, and covariance functions $\mathrm{cov}(Y(s), Y(t)) = G_Y(s,t)$, $\mathrm{cov}[X(s), X(t)] = G_X(s,t)$. We usually refer to the arguments of $X(.)$ and $Y(.)$ as time, with finite and closed intervals S and T as domains. We assume that orthogonal expansions of G_X and G_Y (in the L^2 sense) are exist in terms of eigenfunctions ψ_m and φ_k with nonincreasing eigenvalues ρ_m and λ_k, that is, $G_X(s_1, s_2) = \sum \rho_m \psi_m(s_1) \psi_m(s_2); s_1, s_2 \in S$, and $G_Y(t_1, t_2) = \sum \lambda_k \varphi_k(t_1) \varphi_k(t_2); t_1, t_2 \in T$.

We model the actually observed data which consist of sparse and irregular repeated measurements of the predictor and response trajectories X_i and Y_i, contaminated with additional measurement errors. To adequately reflect the irregular and sparse measurements,

we assume that U_{il} (resp. V_{ij}) denote the observation of the random trajectory X_i (resp. Y_i) contaminated with measurement errors ε_{il} (resp. $\in_{ij}$). The errors are assumed to be $i.i.d.$ with $E\,\varepsilon_{il}=0, E[\varepsilon_{il}^2]=\sigma_X^2$ (resp. $E\,\in_{ij}=0, E[\in_{ij}^2]=\sigma_Y^2$), and independent of functional principal component scores ζ_{im} (resp. ξ_{ik}) that satisfy $E\,\zeta_{im}=0, E[\zeta_{im},\zeta_{im'}]=0$ for $m\neq m'$, $E[\zeta_{im}^2]=\rho_m$ (resp. $E\,\xi_{ik}=0, E[\xi_{ik},\xi_{ik'}]=0$ for $k\neq k'$, $E[\xi_{ik}^2]=\lambda_k$). Then we may represent predictor and response measurements as follows:

$$U_{il}=X_i(s_{il})+\varepsilon_{il}=\mu_X(s_{il})+\sum_{m=1}^{\infty}\zeta_{im}\psi_m+\varepsilon_{il},\qquad s_{il}\in\mathcal{S} \tag{6}$$

$$V_{ij}=Y_i(t_{ij})+\in_{ij}=\mu_Y(t_{ij})+\sum_{k=1}^{\infty}\xi_{ik}\varphi_k+\in_{ij},\qquad t_{il}\in\mathcal{T} \tag{7}$$

We note that the response and predictor functions do not need to be sampled simultaneously, that extend the applicability of the proposed functional regression model.

2.2 Functional linear regression model

Consider a functional linear regression model in which both the predictor X and response Y are smooth random functions,

$$E[Y(t)\,|\,X]=\alpha(t)+\int_S\beta(s,t)X(s)ds. \tag{8}$$

Here the bivariate regression function $\beta(s, t)$ is smooth and square integrable, that is, $\int_S\int_T\beta^2(s,t)dsdt<\infty$ Centralizing $X(t)$ by $X^c(s)=X(s)-\mu_X(s)$, and observing $EY(t)=\mu_Y(t)=\alpha(t)+\int_S\beta(s,t)\mu_X(s)ds$, the functional linear regression model becomes

$$E[Y(t)\,|\,X]=\mu_Y(t)+\int_S\beta(s,t)X^c(s)ds. \tag{9}$$

Our aim is to predict an unknown response trajectory based on sparse and noisy observations of a new predictor function. An important step is to estimate the regression function $\beta(s,t)$. We use the following basis representation of $\beta(s,t)$, which is a consequence of the population least squares property of conditional expectation and the fact that the predictors are uncorrelated,

$$\beta(s,t)=\sum_{k=1}^{\infty}\sum_{m=1}^{\infty}\frac{E[\zeta_m\xi_k]}{E[\zeta_m^2]}\psi_m(s)\varphi_k(t). \tag{10}$$

In the first step, smooth estimates of the mean and covariance functions for the predictor and response functions are obtained by scatterplot smoothing; Then a nonparametric FPCA step yields estimates $\hat{\psi}_m, \hat{\varphi}_k$, for the eigenfunctions, and $\hat{\rho}_m, \hat{\lambda}_k$ for the eigenvalues of predictor and response functions. We use two-dimensional scatterplot smoothing to obtain an estimate $\hat{C}(s,t)$ of the cross-covariance surface $C(s,t), s \in S, t \in T$,

$$C(s,t) = \mathrm{cov}(X(s), Y(t)) = \sum_{k=1}^{\infty}\sum_{m=1}^{\infty} E[\zeta_m \xi_k]\psi_m(s)\varphi_k(t). \qquad (11)$$

Finally we obtain estimates for $\sigma_{km} = E[\zeta_m \xi_k]$,

$$\hat{\sigma}_{km} = \int\int \hat{\psi}_m(s)\hat{C}(s,t)\hat{\varphi}_k(t)dsdt \qquad (12)$$

that results an estimate for *β(s, t)*:

$$\hat{\beta}(s,t) = \sum_{k=1}^{K}\sum_{m=1}^{M} \frac{\hat{\sigma}_{km}}{\hat{\rho}_m}\hat{\psi}_m(s)\hat{\varphi}_k(t). \qquad (10)$$

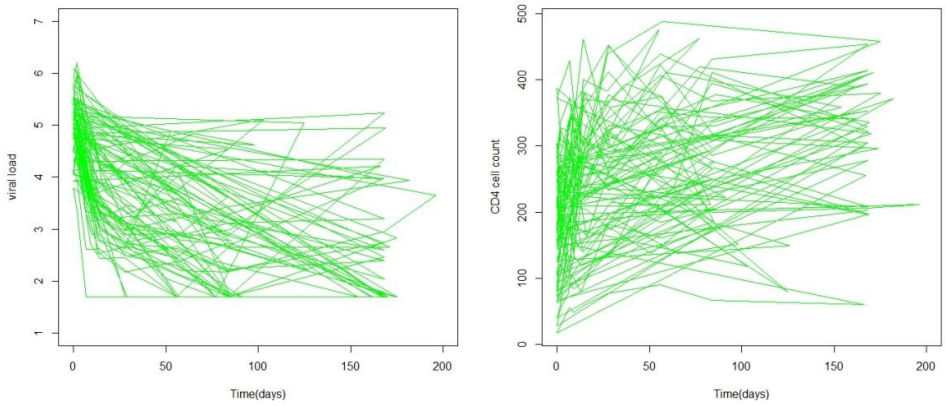
More details on smooting, asymptotic inference, computation of consistent estimation for the regression function, asymptotic pointwise confidence bands for the predicted trajectories and the standard R^2 to the functional case are presented in Staniswslis and Lee(1998), He et al.(2000), Rice and Silverman(1991), Rice(2004), Zhao et al(2004), and Muller(2005).

3 Model fitting results for viral load and CD4 cell counts data

Sinse HIV-1 RNA copies (viral load) and CD4 cell counts are important virologic and immunologic markers for HIV-1 infection, their dynamic relationship during antiviral treatments is of interest. Because the viral load measurement is more difficult to obtain, therefore one has only longitudinal CD4 data available, in which case it is of interest to predict the associated time course of viral load. The data for the following analysis were collected in accordance with AIDS Clinical Trials Group (ACTG). These data are available at [12] and have also been studied in Liang et al. (2003) and Wu and Liang (2004). The data consist of *n* = 46 patients with moderately advanced HIV-1 infection, for whom measurements of viral load and CD4 cell counts were available for the first 24 weeks of treatment. The observed individual trajectories of CD4 cell counts and viral load and descriptive statistics of the data are displayed in Fig.1 and Tab.1 respectively. In order to compare the results of classical and functional approaches, we first analyse this data by marginal models. Then we apply functional linear regression model for analysis of this data.

Fitted viral load trajectories, obtained from both marginal model and functional linear regression model for four randomly selected patients that are shown in figure 2. We compute the mean of observed prediction errors over all subject for every time point, next we use the median criterion of this errors over all times and we use this criterion to compare validity of both models. This criterion are displyed in table 1.

Fig. 1: Observed individual trajectories of the viral load(left panel) and observed individual trajectories of the CD4 cell count(right panel).



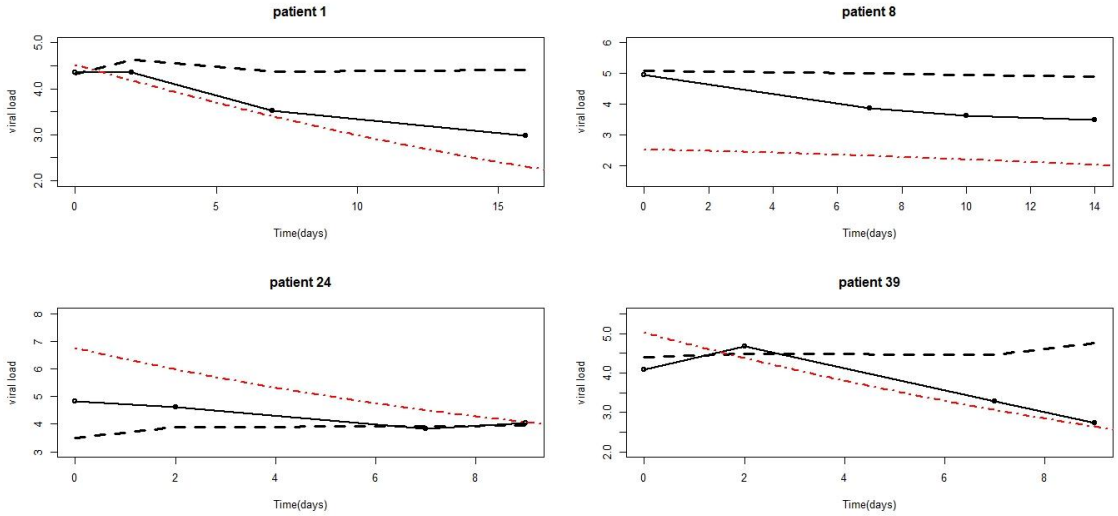
Tab. 1: Descriptive statistics of observed viral load and CD4 cell counts

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
viral load	1.699	3.832	4.398	4.352	4.954	6.204
CD4 cell count	17.28	157.18	216.92	215.63	274.77	461.76

Tab. 2: Median of observed prediction errors over all time points

method	Marginal	Functional
Median	0.5668013	1.084927

Fig. 2: Viral load trajectories(solid), fitted viral load trajectories obtained from both marginal model(dashed) and functional linear regression model(dashed doted red) for four randomly selected patients.



Conclusion

The main problem of inference on longitudinal data is constraint on parametric models with limitative assumption such as choice of correlation structure in marginal models, normality assuption of random effects in random effects models and choice the form of in transition models. These difficulties motivate non-parametric methods for LDA such as FDA method. Classical methods for FDA, which traditionlly have been densely sampled random trajectories observed without errors, are targeted by a new version of FDA. Extention of irregularly measured and noisy trajectories of functional data that is remarkable propertises of longitudinal data have been discussed in this paper. The results of applying this method and marginal models to real longitudinal data are presented in Fig.2 and Tab.2.

In Tab.2 the median criterion of functional regression is greater than marginal model. because in the marginal models we use further assumption and this criterion have been computed based on finit observed time points so information of othere time points are not consedered in this criterion. But in Fig.2 we observed that fitted trajecrories of viral load abtained form the functional model is as well as the marginal model and sometimes perform better rather than the marginal model. Therefor the proposed functional linear regression model is a flexible alternative approach to common classical models, that is applicable to the cases of sparsely sampled longitudinal data. Therefore connection of FDA methodology with longitudinal data is a fairly recent undertaking that presents a promising way for future researchs.

References

[1] Chung, H., Park, Y., & Lanza, S.T. (2005). Latent transition analysis with covariates: pubertal timing and substance use behaviours in adolescent females. Stat. Med. 24, 2895-2910.

[2] Diggle, P.J., Heagerty, P., Liang, K.Y. & Zeger, S.L. (2002). Analysis of Longitudinal Data. Oxford University Press, Oxford.

[3] Liang, K.Y., Zeger, S.L., & Qaqish, B.F. (1992). Multivariate regression analyses for categorical data(with Discussion). J. Roy. Statist. Soc. 54, 3-40.

[4] Molenberghs, G. & Leasaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate plackett distribution. J. Amer. Statist. Assoc. 89, 633-644.

[5] Muller, H.G. (2005). Functional modelling and classification of longitudinal data (with discussion). Scandinavian J. Statistics 32, 223-246.

[7] Ramsay, J., & Silverman, B. (2002). Applied functional data analysis, New York, Springer.

[8] Reuter, M., Hennig, J., Buehner, M., & Hueppe, M. (2004). Using latent mixed Markov models for the choice of the best pharmacological treatment, Stat. Med. 23, 1337-1349.

[9] Rice, J. (2004). Functional and longitudinal data analysis: Perspectives on smoothing. Statistica Sinica 14, 631-647.

[10] Verbeke, G. & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. J. Amer. Statist. Assoc. 91, 217-221.

[12] http://www.urmc.rochester.edu/biostat/people/faculty/WuSite/index.htm

Contact

Mozhgan Taavoni
Department of Statistics, Faculty of Science, Golestan University, Gorgan, Iran
taavonimozhgan@yahoo.com

Mahnaz Khalafi
Department of Statistics, Faculty of Science, Golestan University, Gorgan, Iran
mahnaz_khalafi@yahoo.com

Majid Azimmohseni
Department of Statistics, Faculty of Science, Golestan University, Gorgan, Iran
azim_mohseni@yahoo.com