Semiparametric Mixed Model for Detection of Rapid Disease Progression

Leo L Duan[1,2], John P Clancy[1], Rhonda D Szczesniak[1*]

Abstract

Semiparametric regression provides a more flexible mean structure to analyze longitudinal data through penalized regression splines and mixed modeling. This representation provides smooth estimates of nonlinear mean functions and derivatives while preventing overfitting. Meanwhile, covariance structures must be included to account for repeated measurements collected on multiple subjects. An additional complication arises in the case of unequally spaced measurements, which are common in observational settings. We propose and implement a semiparametric mixed effects model that fits a complex mean function while incorporating subject-specific random effects and an exponential correlation function to account for irregularly observed measurements from multiple subjects. We illustrate the model using serial lung function measurements from the United States Cystic Fibrosis Foundation Patient Registry. This database has clinical encounter information that has been collected over the lifetimes of more than 30,000 individuals. We present our longitudinal model fit and its derivatives to show how this method can provide further insights into the degree and timing of disease progression.

Keywords: Semiparametric Mixed Model, Longitudinal Data, Intra-subject Correlation, Functional Data Analysis

1 Introduction

Disease progression data are repeated measurements collected over time from multiple subjects. Methods of longitudinal analysis are commonly used in this area. On the other hand, two unique features in disease progression distinguish itself from traditional longitudinal or repeated studies: the data are collected at different time points or within different time intervals. As a result, the observations are unbalanced and unequally spaced; the change of severity over time is usually nonlinear and complex, which reflects the

1: Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center
2: Department of Mathematical Sciences, University of Cincinnati
∗: Corresponding Author, email: rhonda.szczesniak@cchmc.org

different stages of disease progression. Therefore, the modeling of disease progression needs to accommodate these characteristics.

A typical longitudinal model of disease progression for ith subject and time t, denoted as $y_i(t)$, consists of two components: a common mean process as a function of time; an individual stochastic process that measures the deviation from the mean.

$$y_i(t) = f(t) + \epsilon_i(t)$$

Traditional methods model the mean process $f(t)$ with a parametric function. The choice of parametric forms ranges from simple linear or polynomial functions to more advanced exponential, logarithmic functions, etc. The decision of functional form may be arbitrary. A more serious caveat of parametric function is that it only achieve global fit. Important local changes, such as different rates of decline in different time periods will be neglected.

On the second component $\epsilon_i(t)$, it is natural to incorporate correlation within subject. One common practice is the use of a random intercept, specific to subject or cluster. As random intercept models assume uniform effects under the same index, a more realistic model incorporates time series covariance to measure the stochasticity over time [**?**]. Common time series models include autoregressive (AR(p)) or autoregressive-moving-average (ARMA(p,q)), etc. While these models provide a more flexible framework to describe the individual deviation, they have two major limitations: 1) the treatment of time is rigid. Since the computation of correlation are based on lags, it is necessary to force unequally-spaced time points into discrete time intervals. 2) the degree of flexibility is restrained by the number of parameters, such as p and $(p + q)$. Correlation over longer interval will demand an increased number of parameter, which is undesirable.

In this article, we propose a novel longitudinal model that provides a flexible and realistic fit to disease progression data. In our model, the common trend exhibited in the population can be subtly captured through localized fit; the individualized correlation can be flexibly described with covariance functions.

2 Methods

An observation in the disease progression data $y_i(t)$, collected from subject i at time t, can be decomposed into two parts:

$$y_i(t) = f(t) + \epsilon_i(t)$$

where $f(t)$ is the common but unobserved process in the population. We propose the use of semiparametric mixed models. One model choice is the truncated cubic spline:

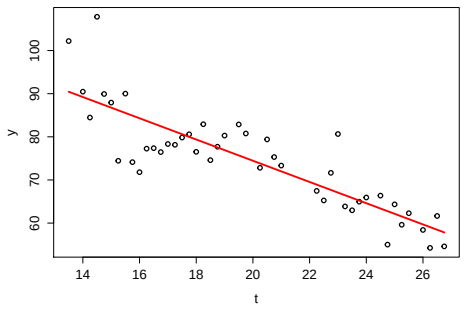
$$f(t) = \beta^T X + \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \gamma_3 t^3 + \sum_{k=1}^{\kappa} \theta_k (t - t_k)_+^3$$

where $\beta^T X$ and the cubic polynomials are the covariates effects and the time effects, respectively. This parametric component is the same as traditional approaches, where global fit is achieved. In the second part, t_k's are the knots selected based on the quantiles of observation time span; $(.)_+$ is a truncation function and only takes value when the inside is positive and is otherwise zero. This nonparametric component not only ensures the continuity and smoothness of $f(t)$, it also enables the local fittig of curves to different periods of time. The derivatives of $f(t)$ can be easily extracted in polynomial differentiation. These derivatives provide important inference to the disease progression rate.

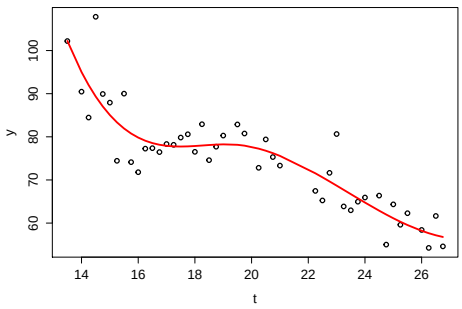
One potential caveat of any nonparametric approach is overfitting. However, this can be avoided by mixed modeling. The treatment of spline coefficients as random effects will penalize overfitting [**?**]. In this case, we assume $\theta_k \overset{iid}{\sim} N(0, \sigma_\theta^2)$.

A comparison between semiparametric and simple cubic polynomials models is shown in Figure 1. The semiparametric method captures the subtle change in mean process, especially the rapid decline during the onset; while cubic function fails to do so since the two thirds of data on the right have dictated an overall linear fit.

The error term $\epsilon_i(t)$ is the individual deviation from the mean process. We use parametric covariance function to model this process. For example, it can be modeled as $\epsilon(t) \sim N(0, \sigma_\epsilon^2 + \sigma_n^2)$ and $Cov(\epsilon(t_1), \epsilon(t_2)) = \sigma_\epsilon^2 exp(-\frac{(t_1 - t_2)^2}{\lambda})$ for $t_1 \neq t_2$. This covariance function is known as double exponential function [**?**]. Compared with traditional time series models, the benefits of using covariance functions include: 1) the observation time can be adopted directly and the function of correlation is continuous, rather than step function; 2) the degree of flexibility is determined by the magnitude of one parameter (λ), instead of number of parameters. The increased flexibility is illustrated in Figure 2.

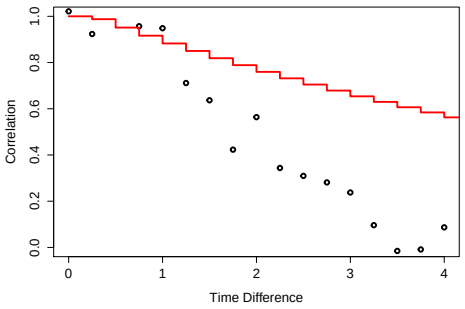


(a) Fitting with cubic function

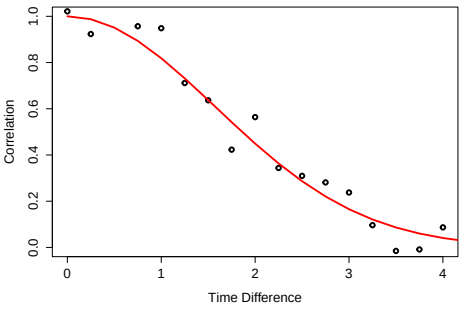


(b) Fitting with semiparametric function

Figure 1: Comparison of data fitting with cubic and semiparametric functions.



(a) Fitting with ARMA(1,1)



(b) Fitting with double exponential function

Figure 2: Comparison of covariance fitting with simple time series structure and covariance function.

3 Data Analysis

We apply the proposed model to the study of a disease named cystic fibrosis (CF), a genetic lung disease that affects over 70,000 people worldwide. Currently no cure exists besides lung transplantation. The data were obtained from the United States Cystic Fibrosis Foundation Patient Registry [**?**]. Forced expiratory volume percentage in 1s (FEV1%) (adjusted with respect to childhood/adulthood, height, gender) is a common measure of CF

lung function, thus we use it for the data analysis. For the regressive nature, the loss of function is non-recoverable; clinical study also demonstrates that the lung condition deteriorates more rapidly during adolescence and early adulthood [**?**].

To study the population level FEV1% loss rate over time, we implement the proposed model in combination with a subject random intercept. This produces an affine transformation of the population curve and increases the robustness of the model. The fitted plots are shown in figure 3. The first derivative shows the magnitudes of disease progression rates. Consistent with clinical findings, the rates are all negative and indicate the monotonically declining nature of lung function. More importantly, the severity between age 10 and 20 is also detected by the model. The signs of the second derivative mark the acceleration or the slow-down of the deterioration.

Comparison of fit statistics shows that the combination of the semiparametric model and covariance function has better fit than each component (Table 1).We also carry out 10-fold cross-validation. The final model has lower error and higher correlation in prediction (Table 2).

	Cubic&RI	**SP&RI**	**Cubic&CFRI**	**SP&CFRI**
AIC	4320076	4313630	4139814	4138807
BIC	4320278	4313827	4140039	4139025

Table 1: Fit statistics of different combinations of mean processes and covariance structures: Cubic polynomial mean (**Cubic**) or Semiparametric mean (**SP**); Random intercept only (**RI**) or Covariance function with random intercept (**CFRI**). Lower is better.

	Cubic	**SP&CFRI**
MAD	10.51	9.81
RMSE	16.80	16.26
RMSE/$\sqrt{\overline{X^2}}$	0.23	0.21
Correlation	0.77	0.79

Table 2: Average of 10-fold cross-validation results: for first three rows, lower is better; for last row, higher is better.

4 Conclusions and Discussions

Semiparametric mixed modeling provides a flexible fit to the nonlinear mean process and a better estimate of the covariance within individual. The derivatives of the mean process enable realistic estimation of the disease progression rate. The model provides improved fit and increased precision in prediction.

Due to the unequal monitoring lengths and irregular observation time in disease progression data, it is reasonable to debate the effects of missingness to the model estimates. In this study, we assume the data are missing at random (MAR). Under this context, the linear mixed model formulation is valid for the analysis, as the missing data is ignorable in the likelihood [**?**]. In the case of missing not at random (MNAR), such as disease-induced death, joint modeling of longitudinal and survival data may be considered [**?**].

One possible extension to this method is to include an additional semiparametric component on the individual level. This will provide predicting and forecasting ability to the model. The population estimates can serve as a guiding trend for each subject. And the individual level estimates and covariance function will regulate and refine the progression trajectory.

References

[1] Diggle, P.J., Heagerty, P., Liang K-Y., Zeger, S.L. (2002). "Analysis of Longitudinal Data". Oxford: OUP

[2] Ruppert D., Wand M.P., Carroll R.J. (2003). "Semiparametric regression". Cambridge University Press.

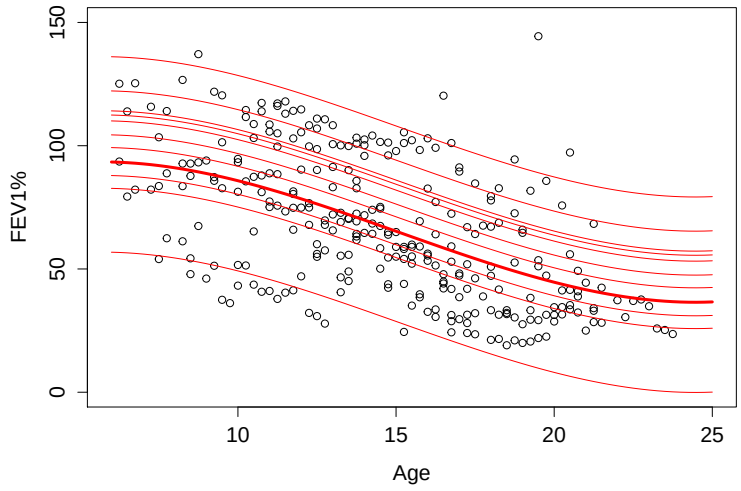
[3] CF Foundation (2012). "Cystic Fibrosis Foundation Patient Registry: 2011 Annual Report".

[4] Vandenbranden S.L., McMullen A., Schechter M.S., Pasta D.J., Michaelis R.L., Konstan M.W., Wagener J.S., Morgan W.J., McColley S.A.; Investigators and Coordinators of the Epidemiologic Study of Cystic Fibrosis (2012). "Lung function decline from adolescence to young adulthood in cystic fibrosis". Pediatr Pulmonol. 47(2): 135-43.

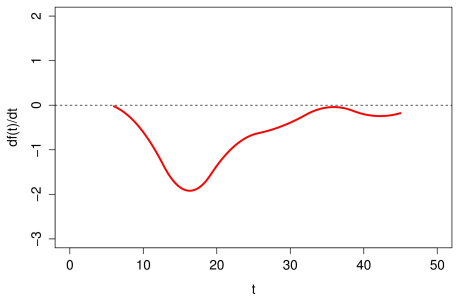
[5] Molenberghs G., Kenward M. (2007), "Missing data in clinical studies", Wiley: Chichester.

[6] Ibrahim J.G., Molenberghs G.(2009),"Missing data methods in longitudinal studies: a review" Test (Madr) **18**(1):1-43.

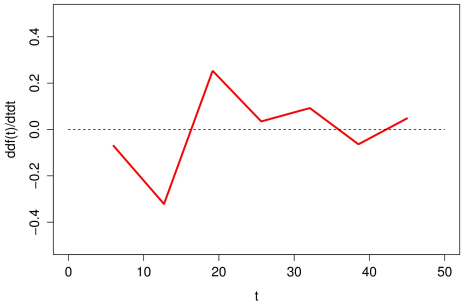
[7] Brockwell PJ, Davis Richard A.(1991) "Time Series: Theory and Methods"



(a) Fitted plot



(b) First derivative



(c) Second derivative

Figure 3: Fitting semiparametric mixed model in cystic fibrosis data. The second derivative is piecewise linear since the function contains cubic polynomials.