ATLANTIS PRESS

*Advances in Computer Science Research, volume 91*

International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)

# Statistical Analysis of AIDS Treatment Effectiveness Based on Longitudinal Data Semi-parametric Model

Guoyi Yan[1,*] and Yan Kong[2]

[1]Wuhan Institute of Technology, School of mathematics and physics, Hubei, China 430074

[2]Chongqing University of Posts and Telecommunications, College of computer Science and Technology, Chongqing, China 400065

*Corresponding author

**Keywords:** Longitudinal data, Semi-parametric model, Simulated annealing algorithm, AIDS efficacy.

**Abstract.** Purpose: In this paper, based on the longitudinal data semi-parametric statistical model, the treatment data of more than 1300 AIDS patients are analyzed. Considering the influence of different treatment schemes, age and treatment time on the number of CD4 in patients. Procedures: the co-variance matrix structure of the same individual is considered based on the semi-parametric model, and the parameter estimation value is calculated by using the simulated annealing algorithm. Conclusion: A conclusion about the effect of AIDS treatment is drawn, we get that the scheme 4 is the best treatment scheme under the condition of controlling for the same value of other independent variables, and the trajectory of the average number of CD4 over time is plotted. The new and original in this paper is that Simulated annealing algorithm is used to find the global optimal solution, The proposed procedures can analyze the influencing factors of AIDS treatment more accurately.

## Introduction

In recent years, the incidence of AIDS is increasing year by year, and the treatment of AIDS needs to be solved urgently. WHO reported that in 2010, there were 34 million living HIV carriers and AIDS patients all over the world, with 2.7 million new infections, and 1.8 million deaths throughout the year. UNAIDS released a new report in Paris on July 20, 2017. It is said that more than half of people living with HIV/AIDS (53%) have received treatment and AIDS-related deaths have nearly halved since 2005. But about 30 percent of people infected with HIV worldwide still do not know their infection status, 17.1 million people living with HIV cannot get anti-viral treatment, and more than half of those infected people are not inhibited.

During the treatment of AIDS patients, data from repeated measurements can be obtained for a group of patients followed in chronological order, and the main efficacy indicators are generally measured many times, as well as information such as treatment scheme, gender, age, and so on. Considering the characteristics of time series and cross-sectional data, such data is often referred to as "vertical data." The characteristic of this measurement data is that the observed values of the reaction variables studied change with time, and the relevant covariates may also change with time. Longitudinal data research is mainly concerned with two aspects: one is the analysis of the overall average change trend, and the other is the analysis of specific differences among different individuals. This kind of data structure is complex, the analysis is technical, and the model is diverse. In recent years, the semi-parametric model of longitudinal data has developed rapidly ([6],[7],[8]), which is one of the hot topics in contemporary statistical research ([9],[10],[11],[12]). The longitudinal data semi-parametric model used for analyze the data of AIDS treatment, Describe the internal correlation between the same patient measurement has become a major concern recently.

In China, there are not many literature analyzing AIDS treatment data with the semi-parametric model of longitudinal data. The concept of semi-parametric model and longitudinal data is introduced in detail in the literature [1]. Literature [2] focuses on the evaluation and prediction of AIDS therapy based on the longitudinal data semi-parametric model. There are many differences between literature

[2] and this paper. The actual calculation and analysis in literature [2] is only non-parametric model, and the model does not contain parameters item. The error terms are assumed to be independent of each other, and the correlation between the same individual is not analyzed. The non-parametric estimation method is local polynomial kernel estimation. In this paper, the co-variance matrix structure of the same individual is considered based on the semi-parametric model, and the parameter estimation value is calculated by using the simulated annealing algorithm, this algorithm can avoid local optimization and improve the estimation accuracy.

## Model

The general form of the semi-parametric model of longitudinal data is

$$Y_{ij} = X_{ij}^T \beta + g(t_{ij}) + \varepsilon_{ij} \tag{2.1}$$

Where, we let $\mathrm{Y_{ij}}$, $\mathrm{X_{ij}}$, $t_{ij}$ denote respectively the measured values of dependent variable, independent variables and observation time at subject $i$ on the jth observation, $i = 1, \cdots, n; j = 1, \cdots, n_i, \beta = (\beta_1, \cdots, \beta_p)^T$ is the unknown regression parameter vector of p dimension, $g(\cdot)$ is a smooth, unknown function. Let

$$\mathrm{Y_i} = (\mathrm{Y_{i1}}, \mathrm{Y_{i2}}, \cdots, \mathrm{Y_{in_i}})^{\mathrm{T}}, \mathrm{X_i} = (\mathrm{X_{i1}}, \mathrm{X_{i2}}, \cdots, \mathrm{X_{in_i}})^{\mathrm{T}}, \varepsilon_i = (\varepsilon_{i1}, \cdots, \varepsilon_{in_i})^{\mathrm{T}}, \mathrm{T_i} = (t_{i1}, t_{i2}, \cdots, t_{in_i})^T,$$

the model can be expressed as matrix form is

$$Y_i = X_i^T \beta + g(T_i) + \varepsilon_i,$$

where $g(T_i) = (g(t_{i1}), \cdots g(t_{in_i}))^T, i = 1, \cdots, n$. Assume that the mean value of $\varepsilon_{ij}$ under given $X_i$ and $T_i$ is zero, and the covariance matrix is $\Sigma_i$.

The parameters estimation of our model is described below.

For the semi-parametric model $Y_i = X_i^T \beta + g(T_i) + \varepsilon_i$, firstly fix the value of $\beta$ and use the weight function estimation method to estimate the unknown function $g(\cdot)$ preliminarily. A so-called weight function estimation is defined as follows

**Definition (weight function estimation)** Given model $Y = g(T) + \varepsilon$, let $W_{ni} = W_{ni}(t) = W_{ni}(t; T_1, \cdots T_n)(i = 1, \cdots, n)$ be the selected $n$ functions that depend on $t$ and $(T_1, \cdots, T_n)$, then

$$\hat{g}_n(t) = \sum_{i=1}^{n} W_{ni} Y_i \tag{2.2}$$

is called the weight function estimation of regression function $g(t)$, $\{W_{ni}, i = 1, \cdots, n\}$ is called the weight function. If the weight function satisfies the condition

$$W_{ni}(t; T_1, \cdots, T_n) \geq 0, \sum_{i=1}^{n} W_{ni}(t; T_1, \cdots, T_n) = 1 \tag{2.3}$$

This is called a probability weight function.

Methods of defining weight functions mainly include nuclear function method and neighbor method and so on, here we use the kernel method. Let $K(u)$ be a non-negative kernel function with respect to $u$=0 symmetry.(such as gaussian kernel: $K(u) = e^{-\frac{u^2}{2}}$), select window width $h_n = h_n(t) > 0$ and define

$$w_{ij}^*(t; h_n) = h_n^{-1} K(\frac{t_{ij} - t}{h_n}), w_{ij}(t) = w_{ij}(t; h_n) = \frac{w_{ij}^*(t; h_n)}{\sum_{k=1}^{n} \sum_{l=1}^{n_i} w_{kl}^*(t, h_n)} \tag{2.4}$$

then $w_{ij}(t) \geq 0$ and $\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(t) = 1$ are true for any $t$. Therefore, $\{w_{ij}(t), i = 1,\cdots,n; j = 1,\cdots,n_i\}$ is the probability weight, and the selection of its window width $h_n$ can be obtained by deleting one individual at a time with the cross-validation method. Due to the existence of internal- individual correlation, the cross-validation method of deleting one observation at a time may not be applicable, because when the observation between individuals has positive correlation or the estimation function is very smooth, the estimated smoothness will be poor.

The estimate of $g(t)$ obtained by the kernel estimation method is expressed as

$$\overset{\vee}{g}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(t)(Y_{ij} - X_{ij}^T\beta) \tag{2.5}$$

Let $\bar{Y}_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{n_i} w_{kl}(t_{ij})Y_{kl}$, $\bar{X}_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{n_i} w_{kl}(t_{ij})X_{kl}$, substitute $\overset{\vee}{g}(t)$ for the estimate of $g(t_{ij})$ above into

$$Y_i = X_i^T\beta + g(T_i) + \varepsilon_i,$$

We get a new model

$$Y_{ij} - \bar{Y}_{ij} = (X_{ij} - \bar{X}_{ij})^T\beta + \tilde{\varepsilon}_{ij},$$

where $\tilde{\varepsilon}_{ij} = (g(t_{ij}) - \overset{\vee}{g}(t_{ij})) + \varepsilon_{ij}, i = 1,\cdots,n; j = 1,\cdots,n_i$. Let $\tilde{Y}_{ij} = Y_{ij} - \bar{Y}_{ij}$, $\tilde{X}_{ij} = X_{ij} - \bar{X}_{ij}$, $\tilde{Y}_i = Y_i - \bar{Y}_i$, $\tilde{X}_i = (\tilde{X}_{i1},\cdots,\tilde{X}_{in_i})^T = X_i - \bar{X}_i$, where $\bar{Y}_i = (\bar{Y}_{i1},\cdots\bar{Y}_{in_i})^T$, $\bar{X}_i = (\bar{X}_{i1},\cdots\bar{X}_{in_i})^T$, then the matrix form of the model is

$$\tilde{Y}_i = \tilde{X}_i\beta + \tilde{\varepsilon}_i,$$

where $\tilde{\varepsilon}_i = (\tilde{\varepsilon}_{i1},\cdots\tilde{\varepsilon}_{in_i})^T$. In this way, semi-parametric model is transformed into parametric model in form.

Now we consider to give an estimation method of $\beta$, from the semi-parametric model

$$Y_{ij} = X_{ij}^T\beta + g(T_{ij}) + \varepsilon_{ij},$$

We replace $g(\cdot)$ in (2.1) with $\overset{\vee}{g}$ as the pre-estimate of function $g$

$$\overset{\vee}{g}(t) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} w_{ij}(Y_{ij} - X_{ij}^T\beta).$$

You can optimize the objective function

$$\tilde{Q}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}(Y_{ij} - X_{ij}^T\beta - \overset{\vee}{g}(T_{ij}))^2 \tag{2.6}$$

We can take $\beta$, which minimizes $Q(\beta)$, as its estimate, while get the estimate of $\beta$, then we give a final estimate of $\hat{g}(\hat{\beta})$ for the unknown function $g(\cdot)$. However, based on the optimization of the objective function, the internal correlation of individuals has not been well characterized. In the application, in addition to the regression parameter $\beta$, the parameter or parameter vector (may be denoted as $\alpha$) in the covariance matrix $\Sigma_i$ of $\varepsilon_{ij}$ should also be estimated. The estimate of $\beta$ and $\alpha$ is $\hat{\beta}, \hat{\alpha}$, giving the final estimate of $g$, which is $\hat{g}(\hat{\beta}, \hat{\alpha})$.

In addition to the longitudinal data, there are a large number of related data. Multiple observations, classified data, repeated measurements and spatial data all involve correlation research. There are many kinds of models for describing correlations, we can introduce the working correlation matrix in Marginal models, establish the random-effects models or the Markov transfer model. On the basis of

exploration and analysis, this paper assumes that the reasonable parameter working covariance matrix $V_i$ approximately replaces the real covariance matrix $\Sigma_i$, and transforms the correlation description into the parameter solving problem.

The reasonable assumption of work correlation matrix $V_i$ becomes the key of modelling analysis. The assumption of the working covariance matrix depends on professional theory or working experience, it is also a common method to use scatter plot matrix for exploration and analysis.

Assumed that the working covariance matrix in parameter form is $V_i(\alpha)$, according to the generalized least square method, the objective function can be optimized

$$Q(\beta) = \sum_{i=1}^{n} (Y_i - X_i^T \beta - \overset{\vee}{g}(T_i))^T V_i^{-1}(\alpha)(Y_i - X_i^T \beta - \overset{\vee}{g}(T_i)) \tag{2.7}$$

By comparing (2.6) and (2.7), it can be seen that the introduction of relevant parameters makes the number of variables of the objective function more, the expression more complex, and most of them are non-linear, so the general optimization method of the objective function will be difficult to work. The simulated annealing algorithm adopted in this paper has the advantage of obtaining the global optimal solution (in the sense of probability 1).The value of $\hat{\beta}, \hat{\alpha}$ obtained by calculation, can use such indicators as estimation bias and estimation standard deviation to describe the quality of regression parameter estimation. Under certain regular conditions, the consistency, asymptotic normality and other large sample properties of parameter estimation can be proved.

**Simulated Annealing Algorithm**

Simulated annealing algorithm is a research result of statistical mechanics of materials, which is from principle of solid annealing. The earliest idea of this algorithm was based on the important sampling method proposed by N. Metropolis and others in 1953.In 1983, S. Kirkpatrick successfully introduced the idea of annealing into the field of combinatorial optimization.

Setting the value of the objective function as $f(x)$ and $x$ as the independent variables of the function, the independent variables are analogous to the molecular state and the temperature evolves into the control parameters, and the simulated annealing algorithm for solving the optimization problem can be obtained. The basic steps of the simulated annealing algorithm are as follows

1) randomly set initial state and set reasonable annealing strategy (select parameter values, initial temperature, cooling law, etc.).

2) let $x' = x + \Delta x$ ( $\Delta x$ is a small uniformly distributed random disturbance),calculate $\Delta f = f(x') - f(x)$.

3) If $\Delta f < 0$ ,accept $x'$ as the new state, otherwise accept $x'$ with probability $P = \exp(-\frac{\Delta f}{t})$.The specific approach is to generate a random number $\xi$ between 0 and 1. If $P > \xi$ accepts $x'$, otherwise $x'$ is rejected, and the system remains in state $x$.

4) Repeat steps 2) and 3) until the system reaches equilibrium.

5) Drop the temperature according to the law given in step 1),repeat steps 2)~4) at the new temperature until it approaches 0 or reaches a predetermined low temperature.

Simulated annealing algorithm has many advantages. Firstly, compared with the traditional algorithm, it does not require much analysis property of the target function, and it does not need smooth conditions, such as derivability and high-order derivability. It is good at solving multidimensional problems, suitable for more complex multivariate function optimization problems. Simulated annealing algorithm has asymptotic convergence, the algorithm receives the inferior solution in a random state update way, especially in a certain probability, so as to make the algorithm jump out of the constraint of local extreme value. It tends to achieve better global optimal performance, in theory, it has been proved that the Markov chain associated with the simulated annealing algorithm is strongly ergodic. As long as the initial temperature is high and the annealing

temperature table is appropriate, the simulated annealing algorithm is an optimization algorithm with probability 1 that converges to the global optimal solution.

## AIDS Data Analysis

### Data Description and Model Calculation

We analyze data from the American AIDS medical trial group, the ACTG. More than 1,300 patients were randomly divided into four groups, each taking one of the four treatments below, and the CD4 concentration was tested about every eight weeks. The daily use of the four treatments was 600mg zidovudine or 400mg didanosine, which were rotated monthly. 600 mg zidovudine and 2.25 mg zalcitabine; 600 mg zidovudine and 400 mg didanosine; 600 mg zidovudine and 400 mg didanosine, and 400 mg nevirapine. We compare the four treatments and determine the best treatment time.

The longitudinal data semi-parametric model described above can be used to analyze the above AIDS treatment data. The logarithm of patients' CD4 value is taken as the dependent variable $Y$ of the model, and three dummy variables are introduced into the four treatment schemes, respectively denoted as $x_1, x_2, x_3$, the independent variable of age is denoted as $x_{age}$, and the variable of observation time is denoted as $t$ ,the semi-parametric model of longitudinal data is expressed as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_{age}\beta_4 + f(t) + \varepsilon,$$

where $x_1 = \begin{cases} 1, & treatment1; \\ 0, & otherwise. \end{cases}$, $x_2 = \begin{cases} 1, & treatment2; \\ 0, & otherwise. \end{cases}$, $x_3 = \begin{cases} 1, & treatment3; \\ 0, & otherwise. \end{cases}$

The efficacy of the fourth therapeutic regimen is used as a baseline, $\beta_i (i = 1, 2, 3)$ reflects the average effect of an increase in type $i$ treatment relative to the fourth type of treatment. In the non-parametric part $f(t)$, the kernel estimation method is used, and the difference type matrix of $\varepsilon$ is assumed to be independent, equally correlated and exponential decline respectively to consider the internal correlation of individuals.

Figure 1 shows the change curve of CD4 concentration in the first 7 observed individuals in the above data. From the figure, we can see that the concentration of CD4 changes with time. As a whole, the CD4 of These individuals first goes up, then goes down, and then goes up, showing a nonlinear characteristic.

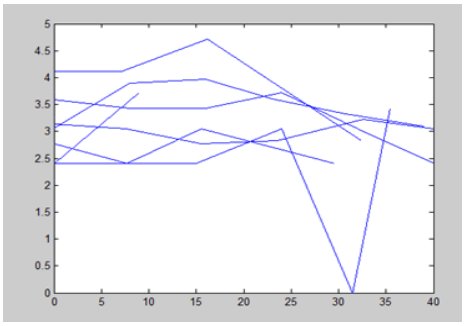


Figure 1. Relational graph of treatment time and treatment effect (number of CD4)

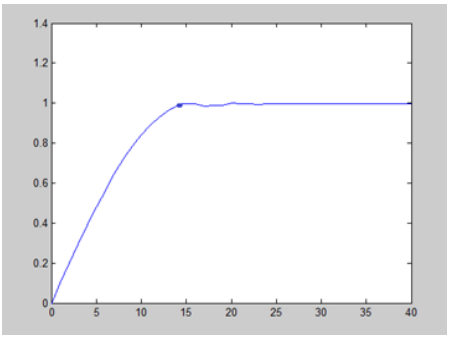


Figure 2. The estimated curve of the function $g(t)$ (the correlation is constant)

Figure 2 is a nonparametric function g($t$) estimation when the correlation is constant, that is, estimation of the function $g(t)$ when the working covariance matrix is $V_i = (V_{kl})_{n_i \times n_i}$ , where $V_{kl} = \begin{cases} \sigma^2, & k = l; \\ \sigma^2\rho, & k \neq l. \end{cases}$ Figure 2 shows that the function image reaches a stable value after 15 weeks when the same individual is assumed to have an equal correlation in the longitudinal data. Table 1 gives estimates of unknown parameters in the model.

Table 1. estimated values of model parameters when individuals are equally correlated

| model parameters | estimated value | estimates of variance |
|---|---|---|
| $\beta_1$ | -0.0273634 | 0.238571 |
| $\beta_2$ | -0.0456932 | 0.345701 |
| $\beta_3$ | -0.037854 | 0.485769 |
| $\beta_{Age}$ | 0.483924 | 0.9246 |
| $\rho$ | 0.527884 | 0.87362 |

Figure 3 shows the function estimation graph of the estimated g(·) in the case of the covariance matrix $V_i = (V_{kl})_{n_i \times n_i}$, where $V_{kl} = \sigma^2 \rho^{|k-l|}, k, l = 1, \cdots, n_i$. (where $n_i$ is the number of observations, that is, the correlation within the individual decreases exponentially).Meanwhile, the estimated values of unknown parameters are given in table 2.

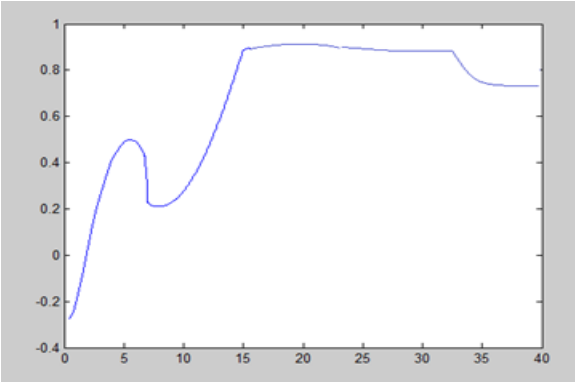


Figure 3. $g(t)$ function estimation graph of individual correlation exponential decline

Table 2. Parameter estimation values of the individual correlation exponential decline model

| parameter | estimated value | estimates of variance |
|---|---|---|
| $\beta_1$ | -0.057383501 | 0.15764 |
| $\beta_2$ | -0.042673912 | 0.57491 |
| $\beta_3$ | -0.02758266 | 0.38642 |
| $\beta_{Age}$ | 0.5736274 | 0.74681 |
| $\rho$ | 0.55897438 | 0.49268 |

## Statistical Result Analysis

The results corresponding to three different covariance matrices can be compared and analyzed as follows:

Table 3. Type of correlation coefficient matrix and corresponding sum of squared residuals

| Correlation coefficient matrix | The value of the Qbeta |
|---|---|
| (1)independent | 97.987 |
| (2)equally correlated | 69.17963 |
| (3)exponential decline | 38.856389 |

The values of Qbeta in table 3 represent the sum of squared residuals. According to the value of Qbeta, by combining the estimated values of each calculated parameter and the function images

obtained by estimating model regression, it can be shown that in the case of exponential decline in the correlation, the variation characteristics of AIDS curative effect are reasonable, that is, when AIDS patients are on medication, the effect of AIDS increases in the first five weeks as the number of treatment weeks increases, but after the fifth week there is a downward trend, and at the same time it reaches a low point in the seventh week, then it continues to rise, reaches a peak in the fifteenth week, then stays at a stable value, and starts to decline after the thirty-second week. It can be seen that the treatment effect began to deteriorate due to physical, mental and other reasons in the later stage of treatment. The therapeutic effect obtained in this covariance matrix is more reasonable compare to the other two situation, and the value of Qbeta is smaller than other cases.

In addition, by analyzing the results of statistical calculation, we come to the following conclusions: since $\hat{\beta}_1 = -0.0573835001$ is the effect of the first scheme based on scheme 4, similar to $\hat{\beta}_2 = -0.042673912$ is the effect of the second scheme, and $\hat{\beta}_3 = -0.02758266$ is the effect of the third scheme, it can be seen that there is a negative mean effect of scheme 1, scheme 2 and scheme 3 on the effect of scheme 4. Therefore, scheme 4 is the best treatment scheme under the condition of controlling for the same value of other independent variables.

**Acknowledgement**

This research was financially Supported by Science Foundation of Wuhan Institute of Technology (17QD53)

**References**

[1] Yan Guoyi. Research on semi-parametric joint model of longitudinal data and survival data.[doctoral thesis], wuhan university, 2013

[2] Yin Xunru. Evaluation and prediction of AIDS therapy based on longitudinal data semi-parametric model [J]. Journal of taishan university.2008,30(6):28-32.

[3] Sun Xiaoqian. You Jinhong. Iterative weighted partial spline least squares estimation in semi-parametric modeling of longitudinal data [J]. Chinese science.2003,33(5):470-480

[4] Gong Guanglu, Qian Minping. Application of stochastic process tutorial and stochastic model in algorithm and intelligent computing [M]. Tsinghua university press, 2003.

[5] Liu Hongwei,Yuan yuan, Wang Xiaoyu and so on. Study on the immunological effects and influencing factors of antiviral therapy for AIDS patients in henan province [J]. Chinese journal of disease control, 2014,18 (9) : 843-846.

[6] Sun Zhihua, Yin junping, Chen feifei. Nonparametric and semiparametric statistics {M}. Tsinghua university press, 2016.

[7] S.L.Zeger and P.J.Diggle. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters[J].Biometrics,1994.50(3):689-699

[8] Wang,N.,Carey,R. and Lin,X. . Efficient semiparametric marginal estimation for longitudinal/clustered data[J]. Journal of American Statistical Association,2005.100:147-157

[9] Sun, L., Song, X., Zhou, J. Regression analysis of longitudinal data with time-dependent covariates in the presence of informative observation and censoring times[J]. Journal of Statistical Planning and Inference,2011, 141: 2902–2919.

[10] Sun, L., Song, X., Zhou, J., Liu, L. Joint analysis of longitudinal data with informative observation times and a dependent terminal event[J]. Journal of the American Statistical Association, 2012, 107: 688–700.

[11] Wu and Briollais. Mixed-effects models for joint modeling of sequence data in longitudinal studies [J]. BMC Proceedings 2014, 8(Suppl 1):S92.

[12] Pei Y B, Du T, Sun L Q. Time-varying latent model for longitudinal data with informative observation and terminal event times [J]. Sci China Math, 2016,59: 2393–2410.