Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq

The Cluster Analysis by Using Nonparametric Cubic B-Spline Modeling for Longitudinal Data

Noor Nawzat Ahmed [1]
Department of Statistics,
College of Adminiteration and Economics,
University of Baghdad,
Baghdad, Iraq
Noor.shareef1101@coadec.uobaghdad.edu.iq

Suhail Najm Abdullah [2]
Department of Statistics,
College of Adminiteration and Economics,
University of Baghdad,
Baghdad, Iraq
dr.suhail.najm@coadec.uobaghdad.edu.iq

Received:3/8/2023 **Accepted: 10/9/2023** **Published: 30/12/ 2023**

Abstract

Longitudinal data is becoming increasingly common, especially in the medical and economic fields, and various methods have been analyzed and developed to analyze this type of data.

In this research, the focus was on compiling and analyzing this data, as cluster analysis plays an important role in identifying and grouping co-expressed subfiles over time and employing them on the nonparametric smoothing cubic B-spline model, which is characterized by providing continuous first and second derivatives, resulting in a smoother curve with fewer abrupt changes in slope. It is also more flexible and can pick up on more complex patterns and fluctuations in the data.

The longitudinal balanced data profile was compiled into subgroups by penalizing the pairwise distances between the coefficients of the cubic B-spline model using one of the common penalize functions, the Minimax Concave Penalty function (MCP). This method, in turn, works to determine the number of clusters through one of the model selection criteria, Bayesian information criteria (BIC), and we used optimization methods to solve their equations. Therefore, we applied the alternative direction method of the ADMM multiplier algorithm to reach approximate solutions to find the estimators of the nonparametric model using R statistical software.

Longitudinally balanced data were generated in the simulation study, as the number of subjects was 60 and the number of repeats (time) was 10 for each subject. The simulation was iterated 100 times, and it showed that employing the MCP partial methods on the cubic model can group profiles into clusters, which is the aim of this paper.

Paper type: Research paper.
Keywords: Longitudinal Data, Nonparametric Cubic B-Spline, Cluster Analysis, The Alternating Direction Method for Multiplier Algorithm ADMM.

1. Introduction

Many terms describe longitudinal data. Data of repeated measurements is called (longitudinal data) in clinical and environmental studies, while in economic studies it is called, (panel data), or time series and cross-sectional data. It combines the spatial, the sectional, and the temporal dimensions (Fadaam, 2018; Al-Adieel and Aboodi, 2021).

In longitudinal studies, data is collected from the same individuals or subjects at multiple points in time. This allows researchers to examine changes that occur over time and to study the effects of interventions or treatments (Mohammed and Khaleel, 2012).

Some examples of longitudinal data include tracking the academic performance of students over multiple school years, monitoring the health outcomes of patients over several months or years, or following the career trajectories of workers over some time (Sadik, 2015). When the cross-sectional observations are measured for the same periods, the longitudinal data is called Balanced longitudinal data. However, if the longitudinal data have missing values at some time observations for some of the groups, then it is Unbalanced longitudinal data (Algamal, 2012).

Certain longitudinal data models are only valid for balanced datasets. If the panel datasets are unbalanced, they may need to be condensed to include only the consecutive periods for which there are observations for all individuals in the cross-section (Liu, 2016). As well as the distance between successive observations, there is a case of equal and unequal space.

Cubic B-spline is a widely used mathematical technique in the context of longitudinal data analysis. By using a cubic B-spline, it is possible to efficiently model and analyze smooth trajectories and directions found in longitudinal data. This powerful combination, which is built with a collection of knots and a set of basic functions, provides a valuable way to understand the dynamic behavior and evolution of subjects through time.

Coffey et al. (2014) pointed out that the spline basis functions consist of a set of piecewise polynomials that connect smoothly at specific points in the time interval, which are known as knots, and the number of basis functions used depends on the number of knots selected. The basis functions used in a cubic B-spline are cubic polynomials that are defined over a local interval between adjacent knots. The knots are typically equally spaced over the range of the predictor variable and are used to determine the location and shape of the basis functions. The Cubic B-spline basis functions are created to be continuous and differentiable and to have a smooth second derivative; that is, the curve of the cubic spline regression function will be in the form of curves that make it more accurate in approaching the real regression curve, and this is reflected in reducing the value of the standard of error, which makes it well suited for modeling smooth and flexible trajectories over time.

Subjects' trajectories can be clustered by employing nonparametric smoothing methods like B-spline techniques treated as a convex optimization problem (Chi and Lange, 2015). In this approach, each subject penalizes the pairwise distance between their centers, enabling the estimation of centers of clusters and the simultaneous determination of the number of groups. This method also incorporates the covariates of interest for the univariate model.

In this paper, we are interested in the method of cluster analysis for longitudinal data using a nonparametric cubic B-spline function, but not with the common methods such as the K-mean method. We used the method proposed by Zhu and Qu (2018) and we will investigate whether the method of clustering using penal functions applies to cubic B-spline, that is, we are developing the method by applying it to cubic B-spline functions.

1.1 Literature review

There are many studies concerned with the field of cluster analysis of longitudinal data, such as: Abraham et al. (2003) collected data with a focus on the functional nature of clusters, and the method was based on the two-stage compilation: the B-spline data function and the division of model coefficients using the K-means algorithm . Fitzmaurice and Ravichandran (2008) aimed at studying repeated measurements of heart patients and studying changes in liver function over a 12-month study period. The researchers Genilonini and Falissard (2010) applied the design of kml, which is an application to determine the paths of longitudinal data using k-means, they made a comparison between artificial data and real data (epidemiological data). Rasheed and Abdel-Hafiz (2012) compared the robust M estimates of the cubic smoothing splines technique with the traditional method of estimating time-varying parameter functions for the balanced longitudinal data, by using two criteria differentiation (MADE, WASE) for different sample sizes, the study showed that the method suggested is better than the traditional methods.

Ali and Abd Al-Sattar (2014) studied the mixed linear parametric and non-parametric model (kernel functions) to analyze wind speed data in Iraq that take the form of repeated measurements over a period of years, 8 meteorological stations were chosen randomly among all stations in Iraq, so the researchers considered that each cluster would represent a station for twelve months, and preference was chosen using the mean squared error (MSE). Coffey et al. (2014) proposed an alternative approach aiming to aggregate profiles of gene expression data over a time period using linear mixed effects models and p-spline smoothing. Another study proposed by Schramm and Vial (2015) used an extended baseline. It was method for treatment efficacy clustering in longitudinal data. Zhu and Qu (2018) proposed a grouping method using the pairwise clustering penalty on the coefficients of the nonparametric model to form subgroups on clustering profiles of subgroups of longitudinal data. Yang et al. (2020) studied random effects to capture correlation from multivariate responses and group individuals by penalizing the pairwise distance between the B-spline coefficient vectors. There was a study by Mohamed and Mohammed (2020) that used kernel methods by the k-means method for cluster analysis, which is aimed at clustering observations in the same cluster that data are homogeneous and not homogeneous with the other clusters in nonlinear data, a method algorithm with k-means are misleading. Therefore they used kernel methods. Zhan et al. (2023) proposed a copula kernel mixture model (CKMM) for clustering multivariate longitudinal data in cases where variables exhibit high autocorrelation using Gaussian copula because of its mathematical tractability to estimate marginal distributions.

The problem of this research is to advance the field of clustering in longitudinal data analysis by utilizing the cubic B-spline model through a novel approach- previously employed with the quadratic B-model- by using the method of penalizing pairwise distances between coefficients of the B-spline model, which is the identification of significant features or time points for data collection. This leads to the creation of more interpretable and insightful models for clustering longitudinal data.

The research aims to achieve two main outcomes:

1. The primary objective is: Is it possible to employ the penalty method for clustering on the model nonparametric cubic B-spline with longitudinal data by penalizing pairwise distances of the cubic B-spline coefficient?
2. The researcher seeks to apply the method through simulated longitudinally balanced data. Then comparing it with the k-means method of clustering.

2. Material and Methods

2.1 The **model for longitudinal data**

In general, the subject-wise model for longitudinal data as follows:

$$y_{ij} = f_i(x_{jl}) + \varepsilon_{ij}. \quad (1)$$

Where y_{ij} is the response variable for subject i^{th} , $i=1,2,..,n$, which repeats in j^{th} times, where $j=1, 2, ..., n_i$, $f_i(x_{jl})$ is denoted for a function for each subject, and assumed that x_{jl} , $l=1, 2, ...,p$,is the corresponding covariate of time that can be scaled to compact interval $\chi \in [0,1]$. And ε_{ij} are i.i.d error (noise) with mean 0 and variance σ^2.

Many different types of functions can be used in longitudinal data analysis, but spline based functions are commonly used in many applications. These functions are made of smooth connections between polynomials with many definitions at specific points called nodes. These nodes are denoted by k={ $k_0 < k_1 < \ldots < k_m$ }, and the number of base functions used depends on the number of nodes chosen. (Coffey, 2014)

2.2 Cubic B-Spline

The degree q of a spline basis function refers to the highest power of the polynomial used in the local intervals between adjacent knots. For example, a cubic B-spline uses cubic polynomials (q = 3) in each interval.

The order r of a spline basis function equals to the degree plus one. This is because the number of coefficients needed to represent the basis function equals to the degree plus one. For example, a cubic B-spline has four coefficients (r = 3+1) multiplied by the knots' values and the polynomial terms in each interval (Chaudhuri, 2013). Let r be the r^{th} order B-spline with a set of m knots sequences k={ $0 = k_0 < k_1 < \ldots < k_m = 1$ } , and the values k are monotonically increasing values which may be either equally spaced, integers or positive. The B-spline is defined by (Carl De Boor, 1972) as follow:

$$B_i^q(x) = \frac{x - ki}{k_{i+q-1} - 1} B_i^{q-1}(x) + \frac{k_{i+q} - x}{k_{i+q} + k_{i+1}} B_{i+1}^{q-1}(x), \quad (2)$$

for i = 0 , ±1 , ±2 , ±3 ,…. . The basis functions $B_i^q(x)$, define by (2), are called B-spline of degree q. and there are p=m + r-1 normalized B-spline basis functions of order r for each outcome.

We introduce a special kind of spline function of degree 3, called (cubic B-spline) is given by (Munguia and Bhatta, 2015):

$$B_i^3(x) = \begin{cases} \dfrac{(x-k_i)^3}{(k_{i+3}-k_i)(k_{i+2}-k_i)(k_{i+1}-k_i)} & \text{if } k_i \le x < k_{i+1} \\ \dfrac{(x-k_i)^2(k_{i+2}-x)}{(k_{i+3}-k_i)(k_{i+2}-k_i)(k_{i+2}-k_{i+1})} + \dfrac{(x-k_i)(k_{i+3}-x)(x-k_{i+1})}{(k_{i+3}-k_i)(k_{i+3}-k_{i+1})(k_{i+2}-k_{i+1})} + \dfrac{(k_{i+4}-x)(x-k_{i+1})^2}{(k_{i+4}-k_i)(k_{i+3}-k_{i+1})(k_{i+2}-k_{i+1})} & \text{if } k_{i+1} \le x < k_{i+2} \\ \dfrac{(x-k_i)(k_{i+2}-x)^2}{(k_{i+3}-k_i)(k_{i+3}-k_{i+1})(k_{i+3}-k_{i+2})} + \dfrac{(k_{i+4}-x)(x-k_{i+1})(k_{i+3}-x)}{(k_{i+4}-k_{i+1})(k_{i+3}-k_{i+1})(k_{i+3}-k_{i+2})} + \dfrac{(k_{i+4}-x)(x-k_{i+2})^2}{(k_{i+4}-k_{i+1})(k_{i+4}-k_{i+2})(k_{i+3}-k_{i+2})} & \text{if } k_{i+2} \le x < k_{i+3} \\ \dfrac{(x-k_i)^3}{(k_{i+4}-k_{i+1})(k_{i+4}-k_{i+2})(k_{i+4}-k_{i+3})} & \text{if } k_{i+3} \le x < k_{i+4} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Then, we can write the cubic spline function as an approximation of $f_{in}(x)$

$$f_{in}(x) \approx \Psi_{in}(x) = \sum_i B_i^3(x_{ij})\beta_{in} = B(x)^T\beta_i \ , \quad (4) \ ,$$

where $f_i = (f_i(x_{i1}), \dots, f_i(x_{in_i}))^T$, $\Psi = (\Psi_1^T, \Psi_2^T, \dots, \Psi_n^T)$, $\Psi_i = B_i\beta_i$,

$B = diag(B_1, B_2, \dots, B_n)$, $B_i = (B(x_{i1}), B(x_{i2}), \dots, B(x_{in_i}))^T$ is a matrix $n_i \times p$ for each subject i.

and $\beta = (\beta_1^T, \beta_2^T, \dots., \beta_n^T)$, β_1^T is a p-dimensional coefficient vector with p=m+q.

2.3 Penalized B-spline

In order to estimate the smoothing function, which reduces the sum of the squares of the penalized error, the penalty limit is added. Both indicated this in the following equation (Hmood and Burhan, 2017):

$$\sum_{j=1}^{t_i}[y_{ij} - f_i(x_{ijl})]^2 + \lambda_1 \int_0^1 \left[\beta_l^{(v)}(x)\right]^2 dx. \quad (5)$$

Equation (5) comprises two components: the first component penalizes the lack of fit, which can be considered as modeling bias, while the second component imposes a Roughness Penalty (RP) that addresses the issue of over-parameterization. We introduce the penalty function to address the fact that the least sum of squares in our model adds unnecessary complexity, leading to a large variance in the estimated parameters. In this approach, the residuals $y_{ij} - f_i(x_{ijl})$ are zero, which contradicts our model. For this approach is zero, which contradicts our model, (Fan and Gijbels, 1996).

So the appropriate way to introduce this punishment is through coarseness, which is commonly measured $\lambda_1 \int_0^1 \left[\beta_l^{(v)}(x)\right]^2 dx$, so that differentiable for the time (v=2), λ_1 is the tuning parameter, often called the smoothing parameter, which variates with the change of coefficient functions.

We can rewrite the objective function of penalized regression spline given the r[th]-order difference penalty as a matrix equation:

$$\varphi(\beta) = \frac{1}{2}\|Y - B\beta\|_2^2 + \frac{1}{2}\lambda_1 \beta_i^T G \beta_i \,, \qquad (6)$$

where $\|\,.\,\|_2^2$ is an L_2 norm, $G = diag(G_r\,, G_r\,, \ldots, G_r)$, is penalty matrix with size $(p \times p)$, and $G_r = AC^{-1}A'$, $A = [a_{ls}]$ is a matrix has *$(p\times(p\text{-}r))$ and* G_r can be written as:

$$G_r = \begin{bmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\ h_0 & 2(h_0+h_1) & h_1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & h_1 & 2(h_1+h_2) & h_2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & 0 & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix}_{p\times p} .$$

By minimizing equation (6), we are obtaining the penalized B-spline coefficient estimator as follows:

$$\hat{\beta} = \arg\min_{\beta\in\delta^{\beta}} \varphi(\beta) \; = (B'B + \lambda_1 G_r)^{-1} B'Y. \qquad (7)$$

Where $\delta^{\beta} = \{\beta: \beta \in \mathbb{R}^{np}\}$ is the B-spline coefficients subspace, which corresponds to the group partition.

2.4 Clustering the Subjects

We assumed that each subject has a unique unknown smoothing function and is denoted by $f_i(x) \in C^r(\chi)$, if the subjects share the same smoothing function form if they are the same group, that is $f_i = f_j$ if the subject *i* and *j* are from the same cluster group.

Let $\vartheta = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_w\}$, where $W \leq n$ is the number of distance groups, then we can define the nonparametric function subspace δ_{ϑ}^{f} corresponding to the group partition (Zhu, 2012):

$$\delta_{\vartheta}^{f} = \{f: f_i = f_{(w)}, f_i \in C^q(\chi), \text{for any } i \in \vartheta_w, 1 \leq w \leq W\},$$

and the subspace of the B-spline coefficients corresponding to the group partition as:

$$\delta_{\vartheta}^{\beta} = \{\beta: \beta_i = \beta_{(w)}, \beta_i \in R^q(X), \text{for any } i \in \vartheta_w, 1 \leq w \leq W\}.$$

We use the B-spline approach to estimate B-spline coefficients simultaneously and perform clustering into subgroups (Zhu and Qu, 2018). This involves applying a penalty to the differences between their B-spline coefficients to encourage subjects to be in the same group, which leads to the following objective function as follow:

$$\mathcal{L}(\beta) = \varphi(\beta) + \sum_{i,j\in\nabla}^{\nabla} \rho(\beta_i - \beta_j, \lambda_2) \,. \qquad (8)$$

Where $\rho(.\,, \lambda_2)$ is a penalty function with a tuning parameter λ_2 to determine the number of subgroups. Also, ∇ is the index set containing a total number of possible pairs $|\nabla| = \frac{n(n-1)}{2}$ of $\{d = (i,j): 1 \leq i \leq j \leq n\}$.

We will use a Minimax Concave Penalty (MCP) which is proposed by Zhang (2010) as a penalty concave function for penalizing a cubic smoothing B-spline, the MCP form is as follows:

$$\rho_\lambda(|\beta|) = \begin{cases} \lambda(|\beta| - \dfrac{|\beta^2|}{2\lambda\gamma}, & |\beta| < \lambda\gamma \\ \dfrac{\lambda^2\gamma}{2}, & |\beta| \geq \lambda\gamma \end{cases} . \quad (9)$$

The value of γ is the tuning parameter; it provides the least value of unbiasedness and more concavity. The parameter $\gamma \geq 1$ controls the unbiasedness of the penalty function, ensuring it possesses continuous and scattering properties (Choon, 2012). To achieve nonparametric coefficient estimations and subgroup subjects, we attempt to minimize equation (8). However, we encountered challenges while optimizing the objective function $\mathcal{L}(\beta)$ directly, and thus, we transform it into the following constrained problem:

$$\min \varphi(\beta) + \sum_{i,j\in\nabla}^{\nabla} \rho(\beta_i - \beta_j, \lambda_2).$$

Which is equivalent to:

$$\min \varphi(\beta) + \sum_{d}^{\nabla} \rho_{\lambda_2}(D\beta)_d .$$

Where is $D\beta = (\beta_1-\beta_2, \beta_1-\beta_3, \ldots, \beta_{n-1}-\beta_n)^T$, $D \in \mathbb{R}^{n(n-1)/2\times p}$ is the transformation matrix of pairwise differences (Park and Shin, 2022).

To solve equation (8), we use the Alternative Direction Method of Multipliers (ADMM) algorithm (Boyd et al., 2010), which is a variant of the Augmented Lagrangian Multipliers (ALM) method.

So, we can rewrite the equation as follows:

$$\min \varphi(\beta) + \sum_{d}^{\nabla} \rho_{\lambda_2}([|z_d|]). \quad (10)$$

Subject to $D\beta = z$.

The scaled version of (ALM) of (10) is given by

$$\mathcal{L}(\beta, z, \lambda_2) = \min \varphi(\beta) + \sum_{d}^{\nabla} \rho_{\lambda_2}(|z_d|) + \frac{\theta}{2}\|D\beta - z^s + u\|_2^2 + \frac{\theta}{2}\|u\|_2^2 \quad (11)$$

Where $u = \lambda_2/\theta$

We update the estimation of β, *z*, *λ*, at the (s+1)th iteration step as follows:

$$\beta^{s+1} = \arg\min_\beta \mathcal{L}(\beta, z^s, {\lambda_2}^s)$$

$$\beta^{s+1} = \arg\min_\beta \frac{1}{2}\|Y - B\beta\|_2^2 + \frac{\theta}{2}\|D\beta - z^s + u\|_2^2 \quad (12)$$

$$z^{s+1} = \arg\min_z \mathcal{L}(\beta^{s+1}, z, {\lambda_2}^s)$$

$$z^{s+1} = \arg\min_z \sum_{d}^{\nabla} \rho_{\lambda_2}(|z_d|) + \frac{\theta}{2}\|D\beta - z^s + u\|_2^2 \quad (13)$$

$$\lambda_2{}^{s+1} = \lambda_2{}^{s} + D\beta^{s+1} - z^{s+1}. \quad (14)$$

First, the solution of equation (12) for β has a closed-form solution as follows:

$$\beta^{s+1} = (B^T B + \lambda_1 G_r + \theta D^T D)^{-1}\left(B^T Y + \theta D^T (z^s - u^s)\right). \quad (15)$$

In order to update z -equation (13) - we use the soft threshold operations of the penalty function $S^{MCP}_{\gamma,\lambda_2(z)}$ to approximate the MCP as follows (Pang et al., 2020):

$$S^{MCP}_{\gamma,\lambda_2}(z) = \begin{cases} 0, & |z| \le 2\gamma\lambda_2 \\ sign(z)\dfrac{2\gamma(|z| - \lambda_2)}{2\gamma - 1}, & \lambda_2 < |z| < 2\gamma\lambda_2 \\ z, & |z| \ge 2\gamma\lambda_2 \end{cases}.$$

Then

$$z^{s+1} = S^{MCP}_{\gamma,\lambda_2}(D\beta^{s+1} + \tfrac{\lambda_2^s}{\theta}). \quad (16)$$

Then, we substitute the equations (15) and (16) in (14) to get values $\lambda_2{}^{s+1}$ (the number of clusters).

Now, we can summarize the ADMM algorithm as follows:

ADMM algorithm
Initialize λ^0=0 and z^0=0 , θ and $\gamma > \frac{1}{\theta}$ are fixed. Step1: update $$\beta^{s+1} = (B^T B + \lambda_1 G_r + \theta D^T D)^{-1}\left(B^T Y + \theta D^T (z^s - u^s)\right)$$ Step2: for all *d*=1, 2, 3, …,|∇| , update $$S^{MCP}_{\gamma,\lambda_2(z)} = \begin{cases} 0, & |z| \le \lambda_2 \\ sign(z)\left(\dfrac{2\gamma(|z| - \lambda_2)}{2\gamma - 1}\right), & \lambda_2 < |z| \le 2\gamma\lambda_2. \\ z, & |z| > \lambda_2 \end{cases}$$ Where $z^{s+1} = S^{MCP}_{\gamma,\lambda_2}(D\beta^{s+1} + \frac{\lambda_2^s}{\theta})$ And $\lambda_2{}^{s+1} = \lambda_2{}^{s} + D\beta^{s+1} - z^{s+1}$ Step3: iterate step 1-2 until stopping criteria are met.

2.5 Select the tuning parameter:

There are various methods for choosing the tuning parameters, such as the Generalized Cross-Validation (GCV) method, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC) (Wang and Zhu, 2011). These methods aim to balance the goodness of fit and the complexity of the model.

We used BIC for selecting the tuning parameters λ_1, which controls the smoothness of B-spline approximation, and λ_2 controls the number of clusters. This is done by a two-step procedure that is proposed by (Zhu and Xiaolu, 2018) as follows:

Step 1: We select the optimal λ_1 by minimizing

$$BIC_{\lambda 1} = \sum_{i=1}^{n} log\frac{REE_i}{n_i} + \frac{1}{n_i} log(n_i)\, df_i. \quad (17)$$

Step2: given λ_2 =0, then we select λ_2 given the optimal λ_1 in equation (17) by minimizing:

$$BIC_{\lambda 2} = log\frac{REE}{n} + \frac{1}{n_i} log(n)\ df, where\ df = \frac{\widehat{W}}{n}\sum_{i=1}^{n} df_i. \quad (18)$$

3.Discussion of results

3.1 Data generation

We used the R program to generate data for n= {60, 100} subjects, explanatory variables were generated based on four common models: The first two models were using the Box-Muller method $f_{(1)}(x)=cos(2\pi x)$, $f_{(2)}(x)=sin(2\pi x)$, the third function was $f_{(3)}(x) = 2(1 - 2\ exp(-6x))$, and $f_{(4)}(x) = 1\text{-}2\ exp(-4x)$. For each subject *i* in a subgroup, n_i=10 equally spaced times points in the interval [0, 1]. The longitudinal data have the autocorrelation problem in the subject, but it is independent between subjects. We generated the random error ε_{ij} is independently and identically distributed according to a normal distribution with mean 0 and variance σ^2, where $\sigma \sim (0,0.4)$, which is estimated by generalized least squares GLS. The continuous response y_{ij} for subject *i* at time point *j* is calculated using the corresponding functional pattern $f_{(C)}(x_{ij})$, where *C*=1, 2, 3, 4 represents the subgroup, i.e. $y_{ij} = f_{(c)}(x_{ij}) + \varepsilon_{ij}$. To obtain robust and reliable results, we conducted 100 simulations.

To determine of the number of knots for each subject by choosing the minimum of $n_i/4$, where n_i is the number of observations for subject *i*, i.e. n_i=10, then the number of knots will be k=3 for all subjects. Additionally, we use a B-spline with an order of 4. Figure (1) shows the curve of one subject of data [0,1] vs. the number of coefficient = 7.

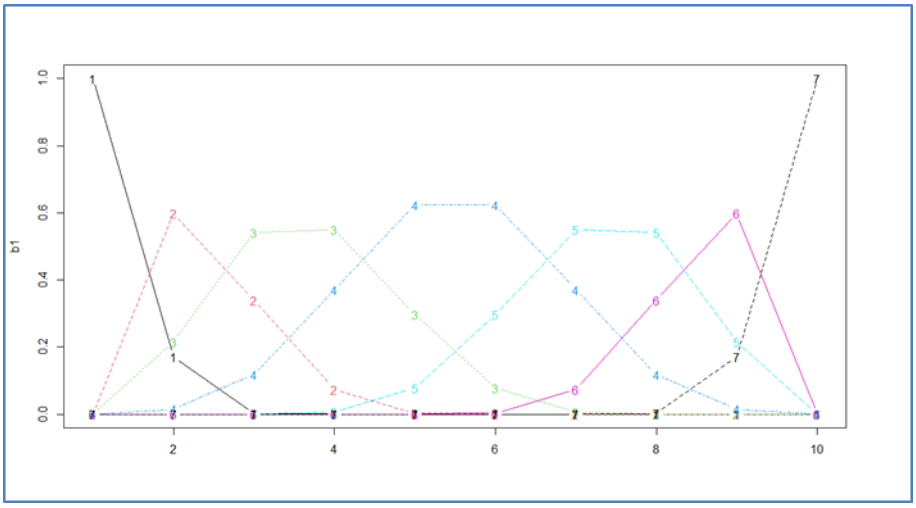

Figure 1: Curve of B-cubic spline for one of the subjects, where x-axis is the time=10, y-axis represents the coefficients =7

By adopting this simulation framework, we can generate data capturing diverse functional patterns that resemble real-world scenarios commonly encountered in scientific studies. In our simulation study, we choose the optimal tuning parameters value by the equations (17) and (18), respectively. λ_1= 0.74 and λ_2=0.08 in case 1 (n=60), and we set the values of θ = 1, 1.25, 1.5 and fixed γ = 1, 2 to ensure the convexity of our objective function.

Case 2: When (n=100), the optimal tuning parameters value, λ_1= 0.8, λ_2=0.05 and we put θ=1.25, 1.5 and $\gamma = 2$.

3.2 The Results

By applying the nonparametric cubic B-spline pairwise grouping, we performed penalty functions MCP, and the results for case 1 are shown in Table (1):

Table (1): The number of clusters and the number of elements in each cluster with γ = 1, 2 and different values of θ

λ_1= 0.74 λ_2=0.08	θ = 1 γ = 1	θ = 1.25 γ = 1	θ = 1.5 γ = 1
Number of clusters	60 clusters	5 clusters	3 clusters
Number in each cluster	1 element in each cluster	17 elements 15 elements 26 elements The other 2 cluster, every one has 1 element	18 elements 16 elements 26 elements
	γ = 2	γ = 2	γ = 2
Number of clusters	3 clusters	3 clusters	3 clusters
Number in each cluster	18 elements 16 elements 26 elements	18 elements 16 elements 26 elements	18 elements 16 elements 26 elements

We calculated the sum of squares within clustering by the cubic B-spline pairwise grouping of 3 clusters, and the result is 8.94772, 12.00920 and 11.91577, and the Mean Squared Error (MSE) is 0.4789956.

And compared it with the MSE for k-means for 3 clusters, which is equal to 0.897 that seems from the Mean Squared Error (MSE), the nonparametric cubic B-spline pairwise grouping is the better than k-means .

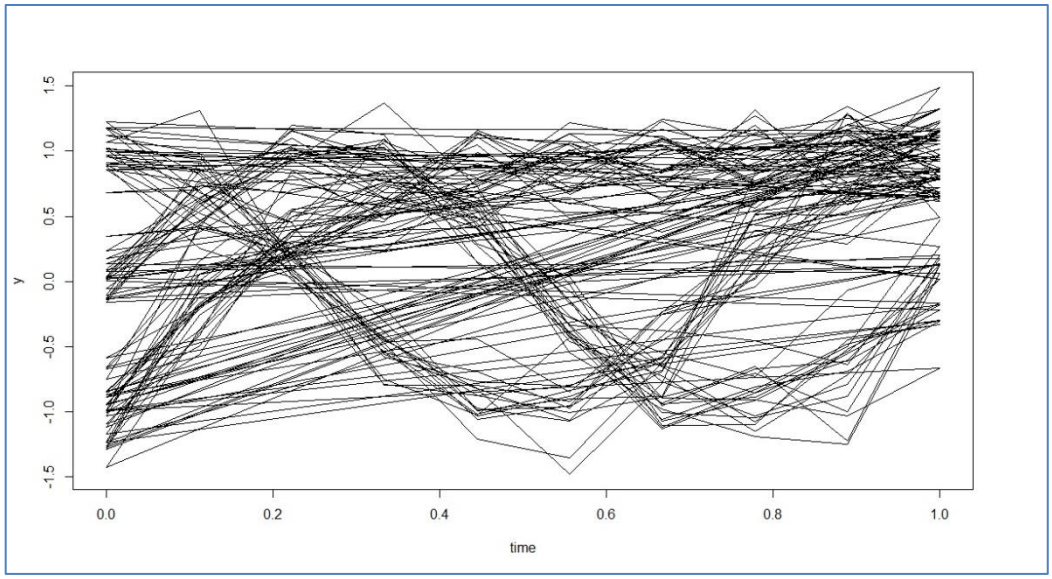

Figure 2: The clusters for 60 subjects, 3 clusters by cubic B-spline, the x-axis representing the repeating(time) and y-axis are the function's curve y, and the curves represented the functions

For the implementation of our study, we utilize the R software. R provides a comprehensive and flexible environment for statistical analysis and algorithm development, making it suitable for our purposes.

In Figure 3, we can see the dendrogram for the clusters of the subjects; we used the (dendextend) R package, which uses to hierarchical cluster analysis.

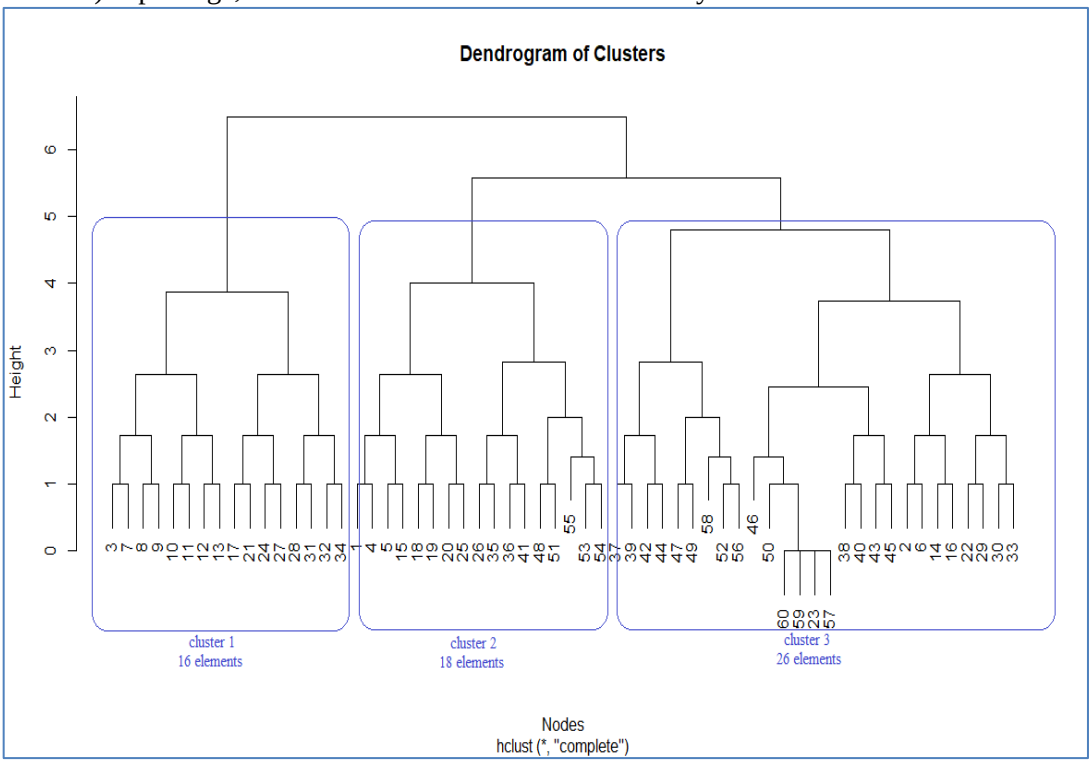

Figure 3: The distribution of 60 subjects using cubic B-spline with MCP penalty function, Having 3 clusters

When we repeated the simulation n=100 subjects (case 2), the results are shown in Table (2):

Table (2): The number of clusters in case 2

λ_1= 0.8 λ_2=0.05	θ = 1.25 γ = 2	θ = 1.5 γ = 2
Number of clusters	4 clusters	3 clusters
Number in each cluster	30 elements 27 elements 32 elements 11 elements	30 elements 27 elements 43 elements

4. Conclusion

By employing the cubic B-spline function to group the longitudinal trajectories over time, we conclude that we can group the subjects into subgroups by penalizing the pairwise distances of the cubic B-spline coefficient vectors, and this method proved successful in grouping by applying it to the generated data of different sizes. For subjects n = 100 and 60, choose the optimal values for the tuning parameters λ_1 and λ_2 and fix the values of both θ and γ suitable for each n, because the characteristic of this sub-grouping method is that it determines the number of clusters by selecting the suitable tuning parameter.

In addition, by comparing this way with k-means through MSE, where we chose the number of clusters as three, we found the nonparametric pairwise grouping was superior to k-means.

References

1.Abraham, C., Cornillon, P.A., Matzner, E. and Molinari, N. (2003). "Unsupervised Curve Clustering Using B-Spline", Board Of Foundation of Scandinavian Journal Of Statistics Vol.30, pp 1-15.
2.Al-adilee, R. and Aboudi, E. (2021). "Comparison of some Methods for Estimating a Semi-Parametric Model for Longitudinal Data", journal of Economics and Administrative sciences, Vol.127, pp.249-261.
3.Algamal, Z. Y. (2012). "Selecting Model in Fixed and Random Penal Data Models", Iraqi Journal of statistical science, Vol.21, pp266-285.
4.Bartels, R. H.; Beatty, J. C.; and Barsky, B. A., (1998), "Hermite and Cubic Spline Interpolation" Ch. 3 in An Introduction to Splines for Use in Computer Graphics and Geometric Modelling. San Francisco, CA: Morgan Kaufmann, pp. 9-17.
5.Boor, Carl.De (1972)." On Calculating With B-Spline" Journal Of Approximation Theory, Vol.6, pp 50-62.
6.Boyd, S. and Parikh, N.,Chu, E., Peleato, B. And Eckstein, J. (2010), "Distributed Optimization and Statistical Learning Via The Alternating Direction Method Of Multiplier", Foundations and Trends in Machine Learning, Vol. 3, No. 1, pp.1-22.
7.Chaudhuri A., (2013). "B-Splines",Samsung R & D Institute Delhi Noida , India.
8.Chi, E. C. and Lange, K. (2015). "Splitting Methods for Convex Clustering", Journal of Computational and Graphical Statistics, Vol.24, No. 4, pp. 994–1013.
9.Choon, C. L.,(2012), "Minimax Concave Bridge Penalty Function for Variable Selection", Department Of Statistics and Applied Probability National University Of Singapore
10. Coffey. N., Hinde, J. and Holian. E. (2014). "Clustering longitudinal profiles using P-spline and mixed effects models applied to time-course gene expression data" , Computational Statistics and Data Analysis, Vol. 71, pp. 14-29.
11. Eilers, P.H. and Marx, B.D.(1996). "Flexible smoothing with B-splines and penalties", Statistical Science, Vol. 11, No.2 , pp. 89 -102.
12. Fadaam, E. (2018). "Compare To Conditional Logistic Regression Models with Fixed and Mixed Effect for Longitudinal Data", Journal of Economics and Administrative Sciences, Vol. 23, No. 98, pp.406-429.
13. Fan,J. and Gijbels, I. (1996). "Local Polynomial Modelling and Its Applications" Chapman and Hall, London.
14. Hmood, M. Y. & Gatie, M. (2014). "A Comparison Of the Semiparametric Estimators Model Using Different Smoothing Methods", Journal of Economics and Administrative Sciences, Vol. 20, No. 75, pp. 376-394.
15. Hmood,M.Y. & Burhan,Y. (2017). "Using Simulation To Compare Between Parametric And Nonparametric Transfer Function Model", Journal Of Economics And Administrative Sciences, vol.24, No.104, pp. 298-313.
16. Liu, Xian (2016). "Methods and Applications of Longitudinal Data Analysis", 1st Edition, Academic Press is an imprint of Elsevier, pp 507 – 530.
17. Manguia, M. and Bhatta, D. (2015). "Use Of Cubic B-Splin in Approximating Solutions of Boundary Value Problem", Applications and Applied Mathematics: An International Journal(AAM) , Vol.10, Issue 2, pp. 750-771.
18. Muhamed, L. A. (2014). " Estimation Mean Wind speed in Iraq by using parametric and nonparametric linear mixed model", Journal of Economics and Administrative sciences, Vol. 20, No. 80, pp. 411-445.
19. Muhamed, L. A. and Mohammed, H.Y. (2020). "On Clustering Scheme for Kernel K-Means", Journal of Al-Rafidain University Collage, Vol. 46, pp. 545-554.
20. Pang. T., Wu. C., Liu. Z. (2020). "A Cubic Spline Penalty for Sparse Approximation Under Tight frame Balanced Model", Springer Science and Business Media, 02 April 2020.

21. Park,S. ,Shin, S.J (2022). "ADMM For Least Square Problem With Pairwise Differences Penalty for Coefficient Grouping" , Communications for Statistical Applications and Methods, Vol. 29, No. 4, pp 441-451.
22. Rasheed, T. and Alhafeth, A. (2012). " Comparison Robust M Estimate with Cubic Smoothing Splines for Time-Varying Coefficient Model for Balance Longitudinal Data", Journal of Economics and Administrative Sciences, Vol. 19, No. 73, pp. 398-413.
23. Rasheed, T. and Alhafeth, A. (2012). "Robust Two-Step Estimation and Approximation Local Polynomial Kernel for Time-Varying Coefficient Model with Balance Longitudinal Data", Journal of Economics and Administrative Sciences, Vol. 19, No. 70, pp. 297-324.
24. Rasheed, T. and Camo, C. (2005). "Comparison of Spline Methods for Estimating Nonparametric Regression Curve", Journal Of Economics and Administrative Sciences, Vol. 8, pp. 40-62.
25. Ruppert, D. (2002), "selecting the number of knots for penalized splines", journal of computational and Graphical Statistics, Vol.11, No.4, pp. 735-757.
26. Ruppert, D. (2002). "Selecting the Number of Knots for Penalized Splines", Journal of Computational and Graphical Statistics Vol. 11, No. 4, pp.735–757.
27. Sadiq, N. and Rasheed, T. (2015). "Estimate The Regression Model of Longitudinal Data With Drop - Outs In Response Variable with Application in Medical Field", University of Baghdad, Collage of Administration and Economics, Department of statistics.
28. Shi, Y., Liu, Y., Jiao, Y. and Cao, Y., (2018), "An Alternating Direction Method of Multipliers for MCP-penalized Regression with High-dimensional Data", Acta Mathematica Sinica Vol.34, No.12, pp. 1-15.
29. Wang, T. , Zhu, L. (2011). "Consistent tuning parameter selection in high dimensional sparse linear regression" , Journal of Multivariate Analysis, Vol.102, pp. 1141–1151 .
30. Yang, L., Zhu, X. , Zhu, Z. and Qu, A., (2020), " Nonparametric Cluster Analysis on Multiple Outcomes of Longitudinal Data", Statistica Sinica, Vol. 30, No. 4, pp. 1829-1856 .
31. Yousif, ali H. & Aboudi, Emad (2017). "Comparison between some of the robust penalized estimators using simulation", Journal Of Economics And Administrative Sciences, vol.23, No.100, pp.490-504.
32. Zhang, C.H. (2010). "Nearly unbiased variable selection under minimax concave penalty", The Annals of Statistics, 38, 894-942.
33. Zhu,X. and Qu, A. (2018). "Cluster analysis of longitudinal profiles with subgroups", Electric Journal Of Statistics, Vol. 12, pp 171-193.
34. Zhu. W. (2012). "Natural Cubic B-Spline Structure At The Boundaries", International Journal of Science & Informatics, Vol. 2, No. 1, pp. 33-39.

التحليل العنقودي باستخدام النموذج اللامعلمي Cubic B-Spline للبيانات الطولية

نور نوزت احمد[1]
جامعة بغداد / كلية الادارة والاقتصاد / قسم الاحصاء
بغداد, العراق
Noor.shareef1101@coadec.uobaghdad.edu.iq

سهيل نجم عبد الله[2]
جامعة بغداد / كلية الادارة والاقتصاد / قسم الاحصاء
بغداد, العراق
dr.suhail.najm@coadec.uobaghdad.edu.iq

Received:3/8/2023 **Accepted: 10/9/2023** **Published: 30/12/ 2023**

مستخلص البحث:
أصبحت البيانات الطولية شائعة بشكل متزايد، خاصة في المجالات الطبية والاقتصادية، وقد تم تحليل وتطوير طرائق مختلفة لتحليل هذا النوع من البيانات.
في الورقة البحثية هذه، تم التركيز على تجميع هذه البيانات وتحليلها، اذ يلعب التحليل العنقودي دورًا مهمًا في تحديد وتجميع الملفات الفرعية والمعبر عنها بمرور الوقت وتوظيفها في نموذج cubic B-spline غير المعلمي، والذي يتميز بكون المشتقات الأولى والثانية لها مستمرة، مما يؤدي إلى منحنى أكثر سلاسة وأكثر مرونة اذ يمكنها التقاط أنماط وتقلبات أكثر تعقيدًا في البيانات.
تم تجميع ملف البيانات الطولية المتوازنة في مجموعات فرعية عن طريق معاقبة المسافات الزوجية بين معاملات نموذج cubic B-spline باستخدام إحدى وظائف العقوبة الشائعة Minimax Concave Penalty function (MCP). وهذه الطريقة بدورها تعمل على تحديد عدد العناقيد من خلال أحد معايير اختيار النموذج وهو معايير المعلومات البايزية (BIC)، واستخدمنا طرائق التحسين لحل معادلاتها. ولذلك قمنا بتطبيق طريقة الاتجاه البديل لخوارزمية مضاعف ADMM للوصول إلى حلول تقريبية لإيجاد مقدرات النموذج غير المعلمي باستخدام برنامج **R** الإحصائي.
وفي دراسة المحاكاة تم توليد بيانات متوازنة طولياً، ذات احجام عينة 60، 100 subjects، وعدد التكرارات (الزمن) 10 لكل subject. تم تكرار المحاكاة 100 مرة، وأظهرت أن استخدام الطرائق الجزائية MCP في النموذج المكعب يمكن أن يعنقد الملفات الشخصية في مجموعات، وهذا هو الهدف من هذه الدراسة.

نوع البحث: ورقة بحثية

المصطلحات الرئيسة للبحث: البيانات الطولية ، نموذج الشرائح B-spline التكعيبية اللامعلمية ، التحليل العنقودي ، طريقة الاتجاه المتناوب لخوارزمية المضاعف ADMM.

* مستل من اطروحة دكتوراه