International Journal of Mathematics, Statistics, and Computer Science
Vol. 2, 2024
ISSN: 2704-1077 eISSN 2704-1069, DOI: 10.59543/ijmscs.v2i.8337

Cluster Analysis on Longitudinal Data of Patients with Kidney Dialysis Using a Smoothing Cubic B-Spline Model

Noor Nawzat Ahmed, Suhail Najm Abdullah
Department of Statistics, College of Administration and Economics, University of Baghdad, Iraq.
nooalior@uomosul.edu.iq; dr.suhail.najm@coadec.uobaghdad.edu.iq

ABSTRACT: Longitudinal data analysis is gaining prominence, particularly in medicine and economics. This research centers on collecting and analyzing longitudinal data, specifically cluster analysis. The thesis emphasizes the nonparametric cubic B-spline model, known for its smoothness, flexibility, and ability to capture intricate patterns and data fluctuations due to the continuity of its derivatives. The penalization method was employed to accomplish clustering. It categorizes longitudinally balanced data by penalizing the pairwise distances between cubic B-spline model coefficients using a penalization function, such as the recently devised concave penalization function. The cubic spline penalty CSP, part of the pair distance penalty, employs the nonparametric pairwise grouping (NPG) method. Model selection criteria, like Bayesian Information Criteria (BIC), help determine the number of clusters. Optimization methods, including the alternative direction method of the multiplier ADMM algorithm, are applied to approximate solutions within the R statistical program. A simulation study generated balanced longitudinal data for 60 and 100 subjects with ten replicates each. The experiments demonstrated the effectiveness of the CSP penalty function in the clustering process. For practical application, the study involved the analysis of data from kidney failure patients collected from Ibn Sina Teaching Hospital for Dialysis in Mosul over seven consecutive months in 2023. The NPG aggregation method and CSP penalty functions were used, resulting in two groups based on the glomerular filtration rate of the kidneys. According to medical criteria, this rate determines the required dialysis frequency twice a week or thrice.

Keywords: cubic B-spline, cubic spline penalty CSP, ADMM algorithm, kidney failure, nonparametric pairwise grouping.

1. Introduction

Many words are used to describe longitudinal data. In clinical and environmental studies, repeated measurement data is referred to as (longitudinal data). In contrast, it is referred to as (panel data), time series, and cross-sectional data in economic studies. It combines the geographic, sectional, and temporal dimensions [1, 2].

In longitudinal studies, data is collected from the same individuals or points at many points. This enables researchers to study changes over time and the effects of interventions or treatments [3].

Some examples of longitudinal data include tracking students' academic performance 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 [4]. When cross-sectional observations are collected for the same periods, the longitudinal data is referred to as balanced longitudinal data. However, if the longitudinal data has missing values at some point in the observations for some of the groups, it is considered unbalanced longitudinal data [5].

Specific longitudinal data models can only be applied to balanced datasets. If the panel datasets are unbalanced, they may need to be reduced to include only the consecutive periods for which all individuals in the cross-section have observations [6]. There is a situation of equal and unequal space and the distance between successive measurements.

Cubic B-spline is a popular mathematical method of analyzing longitudinal data. Using a cubic B-spline, it is possible to efficiently represent and analyze smooth trajectories and directions seen in longitudinal data. This powerful combination, comprised of a collection of knots and a set of basic functions, provides a valuable way to comprehend subjects' dynamic behavior and evolution through time.

Coffey et al. [7] pointed out that the spline basis functions consist of a set of piecewise polynomials that connect smoothly at specific points in the time interval [8], 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 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 [9]. 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.

2. Literature review

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

Many studies have been conducted in the field of longitudinal data cluster analysis, such as Abraham et al. [10] collected data with a focus on the functional nature of clusters, and the method was based on a two-stage compilation: the B-spline data function and the division of model coefficients using the K-means algorithm. Fitzmaurice and Ravichandran [8] sought to investigate repeated measures of heart patients and changes in liver function over a 12-month. Genolini et al. [11] compared artificial data to real data (epidemiological data) using the kml design, an application for determining paths of longitudinal data using k-means. Ali and Abd al-Sattar [4] studied the mixed linear parametric and nonparametric models (kernel functions) to analyze wind speed data in Iraq that take the form of repeated measurements over the years. Eight meteorological stations were chosen randomly from all stations in Iraq. Hence, the researchers assumed that each cluster would represent a station for twelve months, and preference was chosen using the *mean squared error* (MSE). Coffey et al. [7] provided an alternate method for aggregating gene expression patterns over time using linear mixed effects models and p-spline smoothing [12]. Schramm and Vial [13] offered another study that used an extended baseline for treatment efficacy clustering in longitudinal data. Zhu and Qu [14] suggested a grouping approach that uses the pairwise clustering penalty on the nonparametric model coefficients to construct subgroups on clustering profiles of longitudinal data subgroups. Mohamed and Mohammed [15] conducted a study in which they 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 is misleading, so they used kernel methods. Because of its mathematical tractability in estimating marginal distributions, Zhang et al. [16] presented a copula kernel mixture model (CKMM) for clustering multivariate longitudinal data in cases where variables exhibit high autocorrelation using Gaussian copula.

The problem of this research is how to use one of the partial clustering methods and apply it to the cubic B-spline method. However, most penalty methods used in penalty aggregation are common, so the idea was to use modern penalty functions, which were not dealt with in data collection.

The research aims to achieve two primary outcomes:

1. 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 a new penalty function, cubic spline penalty (CSP), to cluster the profiles of longitudinal data.
3. The researcher aims to improve the nonparametric penalizing clustering method using nonparametric pairwise grouping.
4. In addition, applying these methods to data on patients with kidney failure and clarifying whether there are differences between patients by analyzing kidney function tests. So that the specialized medical staff can treat each classified group in a manner appropriate to it.

3. Material and Methods

The model for longitudinal data

In general, the subject-wise model for longitudinal data is 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 $j=1, 2, \ldots, n_i$, $f_i(x_{jl})$ is denoted for a function for each subject, and assumed that x_{jl} , $l=1, 2, \ldots,d$,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.

In longitudinal data analysis, many different types of functions can be used. However, spline-based functions are commonly employed in many applications. These are constructed through smooth connections between polynomials with several definitions at specified 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 [8] .

Cubic B-Spline

A spline basis function's degree q denotes the highest power of the polynomial used for the local intervals between neighboring knots. A cubic B-spline, for example, employs cubic polynomials (q = 3) at each interval. The order r of a spline basis function is equal to the degree plus one. This is because the number of coefficients needed to represent the basis function is equal 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 [17]. Let r is 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 are defined by Carl De Boor in1972 as follow [18] :

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

For $i = 0 , \pm1 , \pm2 , \pm3 , \ldots .$ the basis function $B_i^q(x)$ as define by (2) are call 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 [7]:

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

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

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

where $f_i = (f_i(x_{i1}), \ldots, f_i(x_{in_i}))^T$, $\Omega = (\Omega_1^T, \Omega_2^T, \ldots, \Omega_n^T)$, $\Omega_i = B_i\beta_i$, $B = \mathrm{diag}(B_1, B_2, \ldots, B_n)$,$B_i = (B(x_{i1}) , B(x_{i2}), \ldots , B(x_{in_i}))^T$ is a matrix ni × p for each subject i, and $\beta = (\beta_1^T, \beta_2^T, \ldots., \beta_n^T)$, β_1^T is a p-dimensional coefficient vector with p=m+q .

Penalized B-spline

In order to estimate the smoothing function, which works to reduce the sum of the squares of the penalized error, the penalty limit is added. Both indicated this in the following equation [19]:

$$\sum_{j=1}^{t_i} [y_{ij} - f_i(x_{ijl})]^2 + \lambda_1 \int_0^1 \left[\beta_l^{(v)}(x)\right]^2 dx \quad (4)$$

Equation (4) has two components: the first penalizes the lack of fit, which can be considered as modeling bias, while the second applies a *Roughness Penalty* (RP), which addresses the issue of over-parameterization. The penalty function is introduced to address the fact that the least sum of squares in our model adds unnecessary complexity, resulting in 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, [20].

So the appropriate way to introduce this punishment is through coarseness, which is commonly measured $\lambda_1 \int_0^1 [\beta_l^{(v)}(x)]^2 dx$, 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 :

$$\varphi(\beta) = \frac{1}{2}\|Y - B\beta\|_2^2 + \frac{1}{2}\lambda_1 \beta_i^T \Lambda \beta_i \quad (5)$$

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

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

By minimizing Equation (5), we are obtaining the penalized B-spline coefficient estimator as follows:

$$\hat{\beta} = \arg\min_{\beta \in \delta^\beta} \varphi(\beta) = (B'B + \lambda_1 \Lambda_r)^{-1} B'Y \quad (6)$$

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

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 $\vartheta = \{\vartheta_1, \vartheta_2, \dots, \vartheta_w\}$,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 [21]:

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

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

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

To estimate B-spline coefficients simultaneously and perform clustering into subgroups, we use the B-spline approach [14]. 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:

$$\mathcal{L}(\beta) = \varphi(\beta) + \sum_{i,j\in\nabla}^{\nabla} \rho(\beta_i - \beta_j, \lambda_2), \quad (7)$$

Where $\rho(., \lambda_2)$ is a penalty function with a tuning parameter λ_2 to determine the number of subgroups. and ∇ is the index set containing a total number of possible pair $|\nabla| = \frac{n(n-1)}{2}$ of $\{d = (i,j): 1 \leq i \leq j \leq n\}$. the value of γ is the tuning parameter, it provides the least value of unbiasedness and more concavity.

We will use a *Cubic Spline Penalty* (CSP) which is proposed by [22] as a penalty concave function for penalizing a cubic smoothing B-spline , the CSP form as follow:

$$\rho_{\gamma,\lambda_2}^{csp} = \rho(\beta_i - \beta_j, \lambda_2) = \gamma\lambda_2 \int_0^{\|\beta_i - \beta_j\|_2} \frac{(x - 2\gamma\lambda_2)^2}{4} dx,$$

where $\in [0, +\infty)$, and $\gamma \geq 1$ is a parameter that controls the unbiasedness of the penalty function which gives a continuous and scattering property. To achieve nonparametric coefficient estimations and subgroup subjects, we attempt to minimize Equation (7). However, we encountered challenges while optimizing the objective function *L(β)* 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 transformation matrix of pairwise differences [23,24].

To solve Equation (9), we use the Alternative Direction Method Of Multipliers (ADMM) algorithm [24], which is a variant of *the Augmented Lagrangian Multipliers* (ALM) method.

So, we can rewrite the equation as follows:

$$\min \varphi(\beta) + \sum_{d}^{\nabla} \rho_{\lambda_2}([|z_d|]) \quad (8)$$

Subject to $D\beta = z$

The scaled version of (ALM) of (8) is given by

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

Where $u = \lambda_2/\theta$

We update the estimation of $\beta, z, \lambda,$ at the (s+1)th iteration step as follow:

$$\beta^{s+1} = \arg\min_{\beta} \mathcal{L}(\beta, z^s, {\lambda_2}^s) \quad (9)$$

$$z^{s+1} = \arg\min_{z} \mathcal{L}(\beta^{s+1}, z, {\lambda_2}^s) \quad (10)$$

$${\lambda_2}^{s+1} = {\lambda_2}^s + D\beta^{s+1} - z^{s+1} \quad (11)$$

First, the solution of equation (10) for β has a closed-form solution as follows:

$$\beta^{s+1} = (B^T B + \lambda_1 G_r + \theta D^T D)^{-1}\left(B^T Y + \theta D^T (z^s - u^s)\right) \quad (12)$$

In order to update z -equation (8)- we use the soft threshold operations of the penalty function $S^{CSP}_{\gamma,\lambda_2(z)}$ to approximate the CSP as follows [22]:

$$S^{CSP}_{\gamma,\lambda_2(z)} = \begin{cases} 0, & |z| \le \lambda_2 \\ sign(z)2\left((\gamma-\gamma^2)\lambda_2 + \gamma\sqrt{\lambda_2(|z| - 2\gamma\lambda_2 + \gamma^2\lambda_2)}\right), & \lambda_2 < |z| \le \gamma\lambda_2 \\ z, & |z| > \lambda_2 \end{cases}$$

and

$$z^{s+1} = S^{CSP}_{\gamma,\lambda_2}\left(D\beta^{s+1} + \frac{\lambda_2^s}{\theta}\right) \quad (13)$$

Then

$$z_d^{s+1} = \begin{cases} \|z_d^{s+1}\|_2 & if \|z_d^{s+1}\|_2 \ge \gamma\lambda_2 \\ (\gamma-\gamma^2)\lambda_2 + \gamma\theta u\sqrt{\left(\frac{1}{\theta u} - \frac{\gamma(2-\gamma)}{\|z_d^{s+1}\|_2}\right)} \cdot \underline{z} & if \|z_d^{s+1}\|_2 < \gamma\lambda_2 \end{cases} \quad (14)$$

Then we substitute the equations (12) and (14) in (11) to get values ${\lambda_2}^{s+1}$ (the number of clusters) .

Now, we can summaries the ADMM algorithm to solve equating (9) as follows:

ADMM algorithm

Intialize 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

$$z_d^{s+1} = \begin{cases} \|z_d^{s+1}\|_2 & if \|z_d^{s+1}\|_2 \geq \gamma\lambda_2 \\ (\gamma - \gamma^2)\lambda_2 + \gamma\theta u \sqrt{\left(\frac{1}{\theta u} - \frac{\gamma(2-\gamma)}{\|z_d^{s+1}\|_2}\right)} \ . \underline{z} & if \|z_d^{s+1}\|_2 < \gamma\lambda_2 \end{cases}$$

Where $z^{s+1} = S^{CSP}_{\gamma,\lambda_2}(D\beta^{s+1} + \frac{\lambda_2^s}{\theta})$

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.

Choose the tuning parameter

There are several tuning parameter selection methods, including the *Generalized Cross-Validation* (GCV) method, the *Akaike Information Criterion* (AIC), and the *Bayesian Information Criterion* (BIC) [25].

These methods seek to find a balance between the goodness of fit and the model's complexity. We will use a two-step technique suggested by to select the tuning parameter λ_1 which controls the smoothness of B-spline approximation and λ_2 which controls the number of clusters selected, and this way has been selected 1 by minimizing [26]:

$$BIC_{\lambda 1} = \sum_{i=1}^{n} \left\{ log\left(\frac{REE_i}{n_i}\right) + \frac{1}{n_i}\log(n_i)\, df_i \right\} \quad (15)$$

given λ_2 =0 , where df_i for each longitudinal profile, and then we select the last one by minimizing :

$$BIC_{\lambda 2} = log\left(\frac{REE}{n}\right) + \frac{1}{n_i}\log(n)\ \ df, where\ df = \frac{\widehat{W}}{n}\sum_{i=1}^{n} df_i \quad (16)$$

Data Generation

We have conducted simulation experiments using the CSP penalty functions of the nonparametric cubic B-spline model to cluster with the penalty method under study. Two numbers of subjects were used n={60, 100}. Each of them was repeated 100 iterations to obtain robust and reliable results.

The explanatory variables were created based on three common models that have been relied upon in many studies and research [12, 26, 27, 28]:

$f_{(1)}(x) = cos(2\pi x)$, $f_{(2)}(x) = sin(2\pi x)$, $f_{(3)}(x) = 2(1 - 2\exp(-6x))$,

each subject *i* in a subgroup has the same reputation, x_{ij} where j=1, 2, …,10 equally spaced times points in the interval [0, 1]. The longitudinal data have the autocorrelation problem in the subject itself, 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 $\sigma \sim (0, 0.4)$. 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 represents the subgroup, i.e. $y_{ij} = f_{(c)}(x_{ij}) + \varepsilon_{ij}$.

To determine of the number of knots for each subject by choosing the minimum of $n_i/4$ [28], 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 (6) shows the curve of one the subject of data [0,1] vs. The number of coefficient =7.

When the clustering by CSP was performed, in case n = 60, we evaluated the optimal tuning parameters, λ_1 = 0.002 and λ_2 = 0.5, by equations (15) and (16), respectively, and we set the values of θ = 1.5 and fixed γ = 1 to ensure the convexity of our objective function. Then the number of clustering are three, which contain {18, 16, and 26} elements. The package "dendextend" used to plot the hierarchical cluster the of the topic groups, which is used for analysis. It is shown in Figure (1):

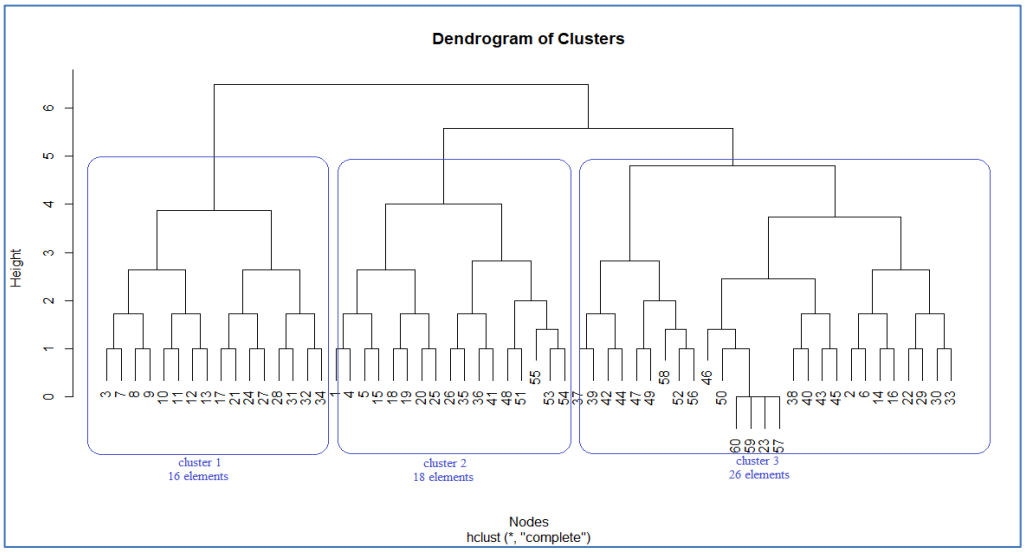

Figure 1: the distribution of 60 subjects using cubic B-spline, having 3 clusters

Figure (2) shows the *Nonparametric Pairwise Grouping* NPG for the cubic B-spline function when the sample size is 60.

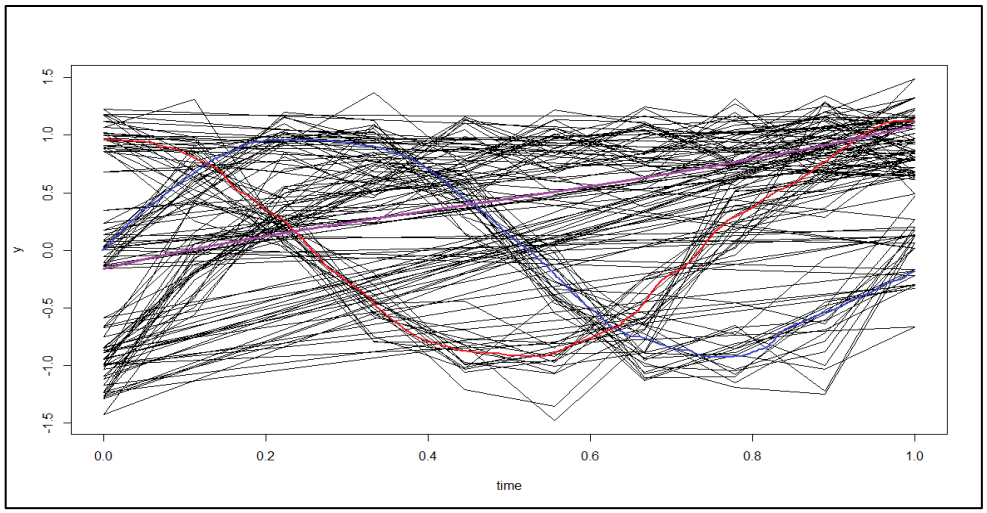

Figure 2: shows The clusters for 60 subjects, 3 clusters by the nonparametric pairwise grouping for the cubic B-spline function, the x-axis representing the repeating (time), and the y-axis is the function's curve y.

When clustering utilizing NPG with a sample size of 100 subjects, the number of clusters after the two tuning parameters $\lambda_1 = 0.747$ and $\lambda_2 = 0.048$ were chosen using equations (15) and (16) after $\theta = 1.25$ was fixed, and $\gamma = 1$, the result was: 3 clusters. These contain {36, 32, 32} elements, respectively. The Figure (3)

shows the 3 clusters as follows:

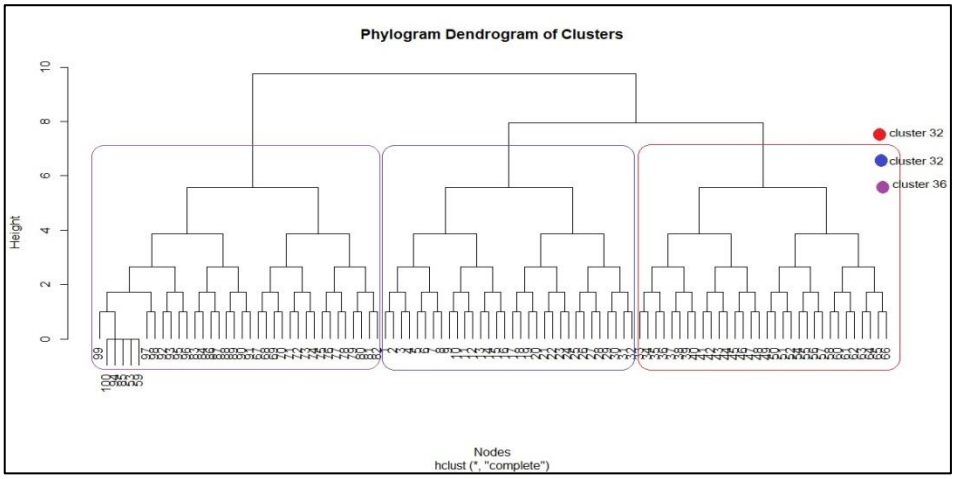

Figure (3): 3 clusters using CSP , when n=100 where cluster1 has 36 elements, cluster2 has 32, and cluster3 has 32 elements

Compare Results

By performing clustering in each of n = 60, 100 subjects using the k-means algorithm for clustering longitudinal data[29,30], the number of clusters k = 3 was chosen and compared with the results of penal clustering using CSP, by calculating the MSE, as shown in Table (3) follows:

Table (3): The comparison between K-means and NPG cubic spline using CSP

Number of subjects	MSE_{CSP}	$MSE_{K-means}$
n=60	0.0162353	0.4789956
n=100	0.6933968	0.8467655

Reviewing the results, we find that the NPG method is generally better than the k-means method and that the use of clustering using CSP has outperformed all of the other methods that were studied. This indicates an improvement in the work of nonparametric clustering in the Cubic B-spline model for longitudinal data.

Data Collection:

This study was applied to the data of kidney failure patients at the Ibn-Sina Teaching Hospital for dialysis, in Nineveh Governorate, for the year 2023, and for 65 patients, where medical examinations were taken on a regular and monthly basis, as the study period was from January until August of this year.

The tests collected are tests for the levels of: blood hemoglobin, white blood cells, creatinine, urea, protein, albumin, and blood glucose, taking into account age and gender. Through a special medical equation, the filtration rate or function of the kidneys' glomeruli is checked for each patient, which determines the condition of the patient's kidneys, whether he suffers from kidney failure and therefore undergoes dialysis or not.

In light of the mentioned above, the following longitudinal data were examined:

n=65 subjects (number of patients), p = 7 replications (months),

$y_{ij} = 65 \times 7$ i.e. 455 observations (Tests related to the functioning of the glomerulus)

Normality Test

Before starting the process of statistical analysis of real data, it is necessary to test whether these data follow a normal distribution or not. The data were tested using the Kolmogrov-Smirnov test, where the hypotheses were tested:

H0: The data follows a normal distribution.

H1: The data doesn't follow a normal distribution.

The following table (4) shows the result of the test:

Table (4): The result of the Kolmogorov-Smirnov normality test

Statistic	N	p-value
5.564	65	0.000

The p-value, as we see, is 0.000, which is less than $\alpha = 0.05$, so H0 will be rejected, which means the data doesn't follow a normal distribution.

Data analysis

In this section, we analyzed kidney failure data, where the model was estimated using a nonparametric cubic B-spline model using equation (7), and penalized clustering was used in nonparametric pairwise clustering using the CSP equation. Through clustered the data using the CSP penalty function, it was grouped into two groups (clusters), where the number of knots = 2, and the two tuning parameters were chosen, λ_1= 0.7 and λ_2 = 0.8, and we fixed θ = 1. 5 and γ=1. The Figure (4) shows the two clusters of the data as follows:

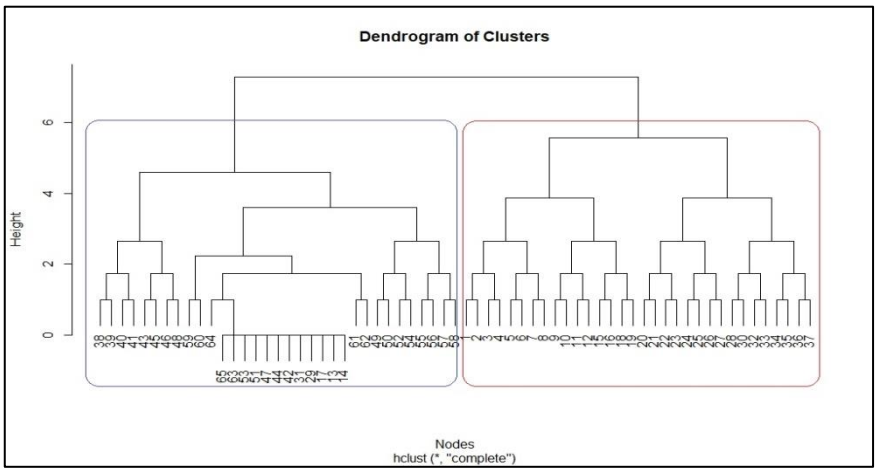

Figure (4): shows 2 clusters using NPG by CSP, cluster 1 has 33 elements, and cluster 2 has 32 elements.

4. Result Interpretation

From what was stated in the above analysis, we find that CSP functions participate in grouping the data into two clusters. The first cluster consists of 33 elements, while the second cluster consists of 32 elements, but by choosing special tuning parameters, The clustering is based on glomerular filtration rates, which in turn affect the functioning of the human kidney. By projecting the elements of the clusters onto the original data, we find that the first cluster, consisting of 33 elements, was caused by diabetes, heredity, kidney atrophy, or an external symptom. As for the second cluster, the cause was high blood pressure, and this is an indication of the seriousness of this cause, which constitutes 49% of kidney failure. These reasons led to a defect in the glomerular filtration function of the kidneys. If the glomerular filtration rate is between (15 - 29) mg/min/ml^3, this indicates the presence of a deficiency in kidney function. The person may suffer from kidney failure and need dialysis twice a week, and the patient may need to Increase the number of washing times to three if the rate is less than that.

5. Conclusions

By implementing simulation experiments and the results presented on the experimental side, as well as implementing a real experiment on a set of balanced longitudinal data for examinations of patients with kidney failure and presenting its results on the applied side, the researcher concluded the following:

1. The results of nonparametric pairwise grouping (NPG) showed its ability to cluster using the cubic spline model using CSP penalty functions in clustering; we find that:

When conducting the simulation experiment n = 60, we found that the number of clusters was 36, 32, and 32 elements, with the appropriate tuning parameters being chosen using the developed BIC standard. In the case of the simulation experiment, when n = 100, we find that the three clusters consisted of 36, 32, and 32.

2. By making a comparison between the NPG method of clustering and the k-means method, we find that the former is highly efficient in clustering when using the cubic B-spline smoothing model, as the NPG method works to group profiles of longitudinal data by penalizing the pairwise distances of cubic B-spline coefficient vectors, it also works on finding estimates of the model parameters at the same time, by choosing $\lambda 1$, which in turn controls the smoothness of the cubic B-spline approximate function, and $\lambda 2$, which controls the number of clusters chosen.
3. In conclusion, the method of penal clustering of longitudinal data subjects using the NPG method of the cubic spline smoothing model and using the penal penalty function CSP is an improvement to the clustering process.
4. After applying the cubic spline smoothing model method to the longitudinal data of examinations of patients with renal failure, targeting the glomerular filtration rate, it was found that patients suffering from renal failure are divided into two groups: the first are patients who suffer from glomerular renal failure resulting from high blood pressure, which in turn leads to a failure in the functioning of the glomeruli, as the glomerular filtration rate ranges between 19-25 mg/min/ml^3, and for this reason, dialysis is performed twice a week. The second group consisted of patients suffering from various diseases such as blood sugar, genetic factors, or the causes combined together, causing a decrease in the glomerular filtration rate, which was less than the mentioned rate. And therefore, it is necessary to resort to dialysis three times a week.

REFERENCES

[1] E.Fadaam, Compare to Conditional Logistic Regression Models with Fixed and Mixed Effect for Longitudinal Data , Journal of Economics and Administrative Sciences, 23(98): 406-429, 2018.

[2] R.Al-adilee, and E.Aboudi, Comparison of some Methods for Estimating a Semi-Parametric Model for Longitudinal Data, journal of Economics and Administrative sciences, 127: 249-261, 2021.

[3] L.A. Muhamed and M.I.Khaleel, The Robust Estimaters in Cluster Analysis with Practical Application in the field of Administrative and Financial Corruption, Journal of Economics and Administrative Sciences, 16(69):278-302, 2012.

[4] L. A. Muhamed, S. Abd al-Sattar , Estimation Mean Wind speed in Iraq by using parametric and nonparametric linear mixed model, Journal of Economics and Administrative sciences, 20(80): 411-445, 2014.

[5] Z. T.Aldabagh, and Z. Y. Algamal, Roubust Penalty Methods To Estimate The Coefficient And Variable Selection of Linear Regression Model, Unpublished ,Phd Thesis, University of Mosul, Iraq, 2020.

[6] Liu, Xian ,Methods and Applications of Longitudinal Data Analysis, 1st Edition , Academic Press is an imprint of Elsevier, 507 – 530, 2016.

[7] M.Manguia, and D.Bhatta, Use of Cubic B-Spline in Approximating Solutions of Boundary Value Problem , Applications and Applied Mathematics: An International Journal(AAM) , 10(2): 750-771, 2015.

[8] G.M. Fitzmaurice, and C.Ravichandran, A Primer in Longitudinal Dat Analysis, Department of Biostatistics, Harvard School of Public Health, Boston, Mass. 2008.

[9] E. C. Chi, and K.Lange, Splitting Methods for Convex Clustering, Journal of Computational and Graphical Statistics , 24(4): 994–1013. 2015.

[10] C.Abraham, P.A. Cornillon, E.Matzner, and N.Molinari ,Unsupervised Curve Clustering Using B-Spline , Board of Foundation of Scandinavian Journal Of Statistics 30: 1-15, 2003.

[11] C. Genolini, R.Ecochard, M. Benghezal, T. Driss, S.Andrieu and F.Subtil, KmlShape:An Efficient Method to Cluster Longitudinal Data (time-Series) According Their Shapes, PLoS ONE, 11(6):1-24, 2016.

[12] N. Coffey , J.Hinde, and E. Holian, Clustering longitudinal profiles using P-spline and mixed effects models applied to time-course gene expression data , Computational Statistics and Data Analysis, 71: 14-29. 2014.

[13] C.Schramm, C.Vial, A.Catherine and S.Katsahian, Clustering of longitudinal data by using an extended daseline: A new method for efficacy clustering in longitudinal data, statistical methods in Medical Research, 27(1): 97-113, 2015.

[14] H. Paul, C.Eilers, and D.Marx, Flexible Smoothing with B-splines and Penalties, statistical Science, 11(2): 89-102, 2012.

[15] L. A. Muhamed, and H.Y. Mohammed, On Clustering Scheme for Kernel K-Means, Journal of Al-Rafidain University Collage, 46: 545-554, 2020.

[16] J.Zhang, Y. Yang, and J.Ding, Information criteria for model selection, School of Statistics, University of Minnesota Twin Cities, Minneapolis, Minnesota, USA, 2023.

[17] A.Chaudhuri, B-spline, Samsung R & D Institute Delhi Noida, India, 2013.

[18] Boor ,Carl.De . On Calculating With B-Spline, Journal Of Approximation Theory , 6: 50-62, 1972.

[19] M.Y. Hmood, and Y.Burhan, Using Simulation To Compare Between Parametric And Nonparametric Transfer Function Model, Journal of Economics and Administrative Sciences, 24(104): 298-313, 2017.

[20] J.Fan, and R. Li, Variable selection via nonconcave penalized likelihood and its oracle propertities, Journal of the American Statistical Association, 96: 1348- 1360, 2001.

[21] Zhu. W. Natural Cubic B-Spline Structure At The Boundaries, University of Maryland Eastern Shore , USA, 2012.

[22] T.Pang, C.Wu, Z.Liu. A Cubic Spline Penalty for Sparse Approximation Under Tight frame Balanced Model, Springer Science and Business Media, 02 April 2020.

[23] Y.Zhu, An augmented ADMM algorithm with application to the generalized lasso problem, Journal of Computational and Graphic statistics, 26: 195-204 ,2015.

[24] S.Park, S.J Shin, ADMM for Least Square Problem With Pairwise Differences Penalty for Coefficient Grouping , Communications for Statistical Applications and Methods, 29(4): 441-451, 2022.
[25] T.Wang and L.Zhu, Consistent tuning parameter selection in high dimensional sparse linear regression, Journal of Multivariate Analysis, Vol(102): 1141–1151, 2011.
[26] Z. Y. Algamal,. Selecting Model in Fixed and Random Penal Data Models, Iraqi Journal of statistical science, 21: 266-285, 2012.
[27] T.Rasheed, and A.Alhafeth, Comparison Robust M Estimate with Cubic Smoothing Splines for Time-Varying Coefficient Model for Balance Longitudinal Data, Journal of Economics and Administrative Sciences, 19(73):398-413, 2012.
[28] D.Ruppert, Selecting the Number of Knots for Penalized Splines, Journal of Computational and Graphical Statistics , 11(4): 735–757, 2002.
[29] J. Wu, H. Xiong, J. Chen and W. Zhou, A Generalization of Proximity Functions for K-means, Research Center for Contemporary Management, Key Research Institute of Humanities and Social Sciences at Universities, Tsinghua University, China, 59: 361-370. 2007.
[30] R. S.king, Cluster Analysis and Data Mining :An introduction, Mercury Learning and Information, Dulles, Virginia, Boston, Massachusetts, 2015.