Modeling Multivariate Functional Data using Generalized Canonical Correlation Analysis and Principal Component Analysis

Sohair Fahmy Higazi[1], Dina Hassan Abdel-Hady[2], Hani Ahmed Khedr[3]

[1]Professor of Applied Statistics, Faculty of Commerce - Tanta University, Egypt.
[2]Professor of Statistics, Faculty of Commerce - Tanta University, Egypt.
[3]Statistics Department, Faculty of Commerce, Damanhour University, Egypt.

Article Info
Page Number: 10758-10783
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received:
15 September 2022
Revised: 25 October 2022
Accepted: 14 November 2022
Publication: 21 December 2022

Abstract

The research aimed to study the methods of multivariate analysis Generalized Canonical Correlation Analysis (GCCA), and Principal Components Analysis (PCA) on the functional data, with the aim of finding the most suitable method in analyzing and modeling the functional data. This is done by conducting an applied study of actual data for measuring industry performance for a mobile phone industry company. Where the data included twenty-three variables, divided into six groups (latent variables); As well as the simulation study was applied on fifteen variables for a time series by controlling the form of the relationship between the variables that meets the requirements of the analysis, and these variables are included in three groups. In the actual study data, the search concluded that the GCCA was superior to the principal component method. Whereas the simulation study presented additional results indicating that it is not always possible to assert the superiority of GCCA over the PCA method. Where the simulation study indicated that the results depend on the nature of the correlation matrix of the relationship between the basic variables. If the relationship between the variables within each group is strong, and the interrelationship between the variables in the different groups is weak, it is preferable to perform modeling using the PCA method. The simulation study confirmed that, to model the data using GCCA, there must be activation variables between the groups. The experimental results showed that there are linked variables between the groups that activate the relationship between the latent variables. Whereas, if the relationships between all basic variables within and between groupsare similar, both methods give similar results.

Keywords: Smoothed Data - Multivariate Analysis - Confirmatory Factor Analysis.

1- Introduction

Methods of representing data by functions have received great attention in recent years, as new technologies have made these data increasingly prevalent in science and industries, and these data are known as functional data(Horváth & Kokoszka, 2012). Applications for functional data are numerous, particularly in the sciences of finance, climatology, biology, healthcare, and engineering. and Climate models over time(Pavlidis, Weston, Cai, & Noble, 2002).

Due to significant advancements in data collection technology that have sparked the "Big Data" revolution, Functional Data Analysis (FDA), a field of statistics that deals with the analysis of variables with unlimited dimensions such as curves, groups, and images, has experienced tremendous growth over the past 20 years.According to the study's findings (Aneiros, Cao, Fraiman, Genest, & Vieu, 2019), every methodological issue with multivariate analysis has a functional counterpart. In actuality, the majority of multivariate statistical techniques, like PCA, Canonical Correlation Analysis (CCA), Cluster Analysis, etc., are tailored to functional data(Hanusz, Krzyśko, Nadulski, & Waszak, 2020).

The methods CCA and PCA are statistical tools that are used to analyze and study the interrelationships between data sets. Their primary goal is to minimize the size of a data set made up of numerous interconnected variables(Khan & Farooq, 2012).Furthermore, (Carroll, 1968) proposed Generalized Canonical Correlation Analysis, which is a more comprehensive statistical tool used to analyze more than two sets of variables at the same time. The main goal of GCCA is to generate a series of components aimed at maximizing correlation between groups of multiple variables.

In addition, multivariate statistical techniques such as PCA introduced by (Pearson, 1901) can be relied upon to analyze large sets of data without losing important information (Yishu Wang, Wang, Yang, & Deng, 2014; Yi Wang et al., 2013). PCA can be used to compress datasets from multidimensional vectors into smaller dimensional vectors. Principal components analysis performs linear orthogonal transformation of data that retains maximum variance (Ilin & Raiko, 2010), allowing for reduced dimensionality and better interpretation of results (Helwig, Hong, & Polk, 2012).Hence, this research deals with GCCA and PCA of multivariate functional data. Both are transformational procedures that enable us to minimize the dimensionality of the data and obtain a linear projection of the data.

2- Research Challenges

When performing functional data analysis, the functional data should be analyzed as a complete function defined at a specified continuous time interval, rather than focusing on the observed values at discrete points in the time interval. When conducting the functional analysis of data, researchers may face the problem of multivariate, or that there are certain variables that play the same role as other variables. Accordingly, the research problem is how to integrate PCA and GCCA into functional data analysis and apply it to generate curves that better describe multidimensional variables.

3- Research Objective

This research aims to apply some methods of multivariate analysis (GCCA, and PCA) on functional data, with the aim of finding the best suitable method in analyzing and modeling functional data.

Thus, the research aims to find answers to the following questions:

- How can discrete data be converted into functional data?
- How can the functional data be modeled using the GCCA method and the PCA method?

- Are the results of the statistical data analysis different for the two modeling methods (GCCA, and PCA)?

4- Research Importance

Given that a subfield of statistics that deals with the analysis of unbounded dimensions is functional data analysis, this paper study some multivariate statistical techniques (GCCA, and PCA), as well as the analysis of the main components of functional data. Both GCCA and PCA are important tools for dimensionality reduction (reducing the number of variables), in which the volume of data inputs is reduced by reducing the number of variables included in the model.

5- Research Methodology

The compared methods were implemented using R programming version [4.2.1] to analyze the effect of bothmethods (GCCA, and PCA) on the actual data. The testing data is divided to real data, and simulated data, the real data describe the industry performance of a mobile phone manufacturing company which included twenty-three variables (measured as a percentage), divided into six groups (latent variables). On the other hand, the simulated data was conducted for fifteen variables of a time series by controlling the form of the relationship between the variables that meets the requirements of the analysis. These variables are divided into three groups.

6- Functional Data

Functional data is defined as data obtained from continuous phenomena of space or time and is represented in the form of smoothed functions. Whereas Functional Data Analysis (FDA) deals with data in the form of functions or images and figures in which one or several functions are recorded for each item in the random sample. Functional data, in essence, represents infinite dimensionsof this data pose challenges both in theory and in practice, and these challenges vary according to the mechanism for sampling the functional data. There are several options for data investigation and analysis due to the data's multi-dimensional or infinite structure, which is a rich source of knowledge(J.-L. Wang, Chiou, & Müller, 2016; Yishu Wang et al., 2014).

FDA involves converting data points into continuous functions, which is primarily done using both Fourier, and Spline functions. Functional data analysis relies on functions rather than discrete data points. This has a potential advantage over analyzing discrete data in that it has fewer assumptions over time. The FDA also provides a richer analytics suite than simply comparing means or variances, as functions enable trends and the rate of change to be captured. The FDA provides a very natural method of imposing positional smoothness penalties, which traditional multivariate analysis approaches lack. The functional representations of the curves are highly contradictory after the spline basis is adequate (Ramsay & Silverman, 2005).

Assuming that $X(t)$ is a random process with a continuous parameter $t \in I$, we wish to analyze many multidimensional random processes:

$$X_k(t) = \left(X_{k1}(t), \dots, X_{kp_k}(t)\right)^T \in L_2^{p_k}(I), \qquad t \in I, k = 1, \dots \qquad (1)$$

where $L_2\ (I)$ is the Hilbert space of square-integrable functions in interval I.

A Hilbert space consists of vectors each with an infinite number of coordinates $q_1, q_2, q_3, \dots$. Coordinates are usually considered complex numbers, and each vector has a square length $\Sigma_r\ |q_r|^2$. This squared length must be covered to determine the Hilbert vector by $q's$. A space $L^2\ (a, b)$ is a set of square functions that are integrable in the real or composite interval (a, b), i.e. $\int_a^b |f(t)|^2 dt < \infty$(Dirac, 2012; Ramsay & Silverman, 2005).

Suppose that: l^{th}component of vector $X_k\ (t)$ can be represented by a finite number of orthogonal principal functions $\varphi_b\ (t)$, where $\varphi_b\ (t) \in L_2\ (I)$, $t \in I$. The random process $X(t)$ can be written as:

$$X(t) = \sum_{b=0}^{B} C_b\, \varphi_b(t), t \in I \qquad (2)$$

where φ_b are orthogonal fundamental functions and $c_0, c_1, \dots, c_B$ are the unknown random coefficients. And $E(c) = 0\ , and\ var(c) = \Sigma_c$

This means that the realization of operation $X(t)$ lies in a finite-dimensional subspace of $L_2^p(I)$.

The vector c can be estimated based onn independent outcomes $x_1\ (t), x_2\ (t), \dots, x_n\ (t)$ of the random process $X(t)$ (functional data). The method of least squares can be used as an estimation method. In addition, The majority of financial, meteorological, and other data are often recorded at certain points in time. Thus, we assume that $X_j$represents an observed process value$X(t)$ at time point t_j , where I is a compact set in which$t_j \in I, j\epsilon\ 1, \dots, J$. Thus, the data consists of J of pairs $(t_j\ , x_j\ )$ (Górecki, Krzyśko, Waszak, & Wołyński, 2018; Górecki, Krzyśko, & Wołyński, 2020).

7- Convert discrete data into functional data

The continuous function can be used to initialize discrete data.$x(t)$, where $t \in I$(Ramsay & Silverman, 2005) by supposing:

$$x = \left(x_1, x_2, \dots, x_J\right)'\ , and\ c = (c_0, c_1, \dots, c_B)' \qquad (3)$$

$\Phi(t)$ is a matrix having dimensions $J \times (B + 1)$ with values $\varphi_b\left(t_j\right), b = 0,1, \dots, B, and\ j = 1,2, \dots, J$

It is possible to estimate the coefficient c in equation (2) using the method of least squares, in order to minimize the function:

$$S(c) = (x - \Phi(t)c)'(x - \Phi(t)c) \qquad (4)$$

Differentiating $S(c)$, they found the least squares estimation for vector c:

$$\hat{c} = (\Phi'(t)\Phi(t))^{-1}\Phi'(t)x \tag{5}$$

Then,

$$x(t) = \sum_{b=0}^{B} \hat{c}_b \varphi_b\,(t), \quad t \in I \tag{6}$$

The degree to which $x(t)$ is initialized is determined by the value of B. (the lower the value of B, the more the curves are initialized). The Bayesian Information Criterion(BIC) is used to determine the optimal value of B.

$$BIC = \ln\left(\sum_{j=0}^{J}\left(x_j - \sum_{b=0}^{B} \hat{c}_b \varphi_b\,(t_j)\right)^2\right) + (B+1)\left(\frac{\ln J}{J}\right). \tag{7}$$

This criterion is employed since the BIC and the Akaike Information Criterion (AIC) both evaluate the quality of fit more accurately(Górecki et al., 2018).

And assuming that there are n pairs of independent discrete values are$\left(t_{ij}, x_{ij}\right) j = 1, \ldots, J,$and $i = 1, \ldots, n.$

This discrete data is initialized into continuous functions as follows:

$$x_i(t) = \sum_{b=0}^{B_i} \hat{c}_{ib} \varphi_b\,(t), \quad i = 1, \ldots., n, \qquad t \in I. \tag{8}$$

Among the values $B_1, B_2, \ldots, B_n$, The modal value for one typical value of B is chosen$B_1, B_2, \ldots, B_n$.

The set of functions $x_1(t), x_2(t), \ldots, x_n\,(t){:}\, t \in I$The functional data is what is obtained in this approach (Ramsay & Silverman, 2005). It could be helpful in some circumstances to differentiate Smooth Functions. When analyzing functional data, this is supposed to be the basic aspect of variable selection approaches.

So far, univariate data (p = 1) have been dealt with, and it is possible to generalize to more than one variable $p \geq 2$.

The data is made out ofn independent vector functions:

$$x_i(t) = \left(x_{i1}(t), x_{i2}(t), \ldots, x_{ik}(t)\right)', t \in I, \text{and} i = 1, \ldots, n \tag{9}$$

where$x_1(t), x_2(t), \ldots, x_n\,(t){:}\, t \in I$ with multivariate function data$p = k$. Functional multivariate data can readily be viewed as the output of a multi-dimensional random process with a finite number of dimensions$X(t) = \left(X_1(t), X_2(t), \dots, X_k(t)\right)'$ with a continuous parameter $t \in I$. Also, suppose $X \in L_2^k(I)$, where $L_2\ (I)$is the given inner product present in a Hilbert space of square functions which is integrable in interval I:

$$\langle u, v \rangle = \int_I u'(t)v(t)dt. \tag{10}$$

There are a finite number of orthogonal basis functions $\varphi_b$that can be used to represent the case when the d^{th} component of the process is $X(t)$(Górecki et al., 2018).

$$X_d(t) = \sum_{b=0}^{B_d} c_{db}\varphi_b(t), \quad t \in I, d = 1,2, \dots, p, \tag{11}$$

where c_{db} are random variables. Assuming that:

$$c = \left(c_{10}, \dots, c_{1B_1}, \dots, c_{p0}, \dots, c_{1B_p}\right)',$$

$$\Phi(t) = \begin{bmatrix} \varphi'_{B_1}(t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi'_{B_p}(t) \end{bmatrix},$$

Where

$$\varphi_{B_d}(t) = \left(\varphi_0(t), \dots, \varphi_{B_d}(t)\right)', d = 1, \dots, P.$$

Then

$$X(t) = \Phi(t)c, \qquad t \in I.$$

8- Generalized CanonicalCorrelation Analysis

In this paper, the generalized version of CCA developed by (Carroll, 1968) is studied. This version of data analysis is the most flexible because it enables the solution to be obtained through the self-equation, and it does not require strict distributional assumptions. The main problem of GCCA is the construction of a series of components, or canonical variates, with the aim of maximizing the correlation or homogeneity among groups of multiple variables (Markos & D’Enza, 2016).

GCCA allows several blocks (groups) of variables to be analyzed simultaneously. Assuming that $X_k = \left(X_{k1}, X_{k2}, \dots, X_{kp_k}\right)^T$represents blocks (sets) of random variables, Σ_{kk} where $k = 1, \dots, K$, with covariance matrices and mean vectors with zero values. Also, suppose the total (main) cluster of X variables takes the form $X = (X_1^T, X_2^T, \dots, X_K^T)^T$, and

$$var(X) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1K} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2K} \\ & & & \vdots \\ \Sigma_{K1} & \Sigma_{K2} & \dots & \Sigma_{KK} \end{bmatrix}$$

We seek to make canonical variable vectors $(U_{1i}, \dots, U_{Ki}), i = 1, \dots, s,$and $s = \min_{k \neq j} rank(\Sigma_{kj})$ linear combinations of $(X_1, \dots, X_K)$ on respectively (Górecki et al., 2020). At $i = 1, \dots, s$, the canonical variables maximize the total of their correlations, i.e., they maximize:

$$\sum_{k,j=1,k<j}^{K} corr(U_{ki}, U_{ji})$$

such that U_{ki} has a unit variance, $k = 1, \dots, K$. Furthermore, the vectors $(U_{1i_1}, \dots, U_{Ki_1}$) and $(U_{1i_2}, \dots, U_{Ki_2})$ are uncorrelated at $1 \leq i_1 < i_2 \leq s$.

And assuming $U_{ki} = l_{ki}^T X_k, U_k = (U_{k1}, \dots, U_{ks})^T, L_k = (l_{k1}, \dots, l_{ks}),$

$k = 1,2, \dots, K$,and $i = 1, \dots, s$, then:

$$U_k = L_k^T X_k \ , k = 1,2, \dots, \text{K}.$$

Further, assuming $U = L^T X$, where ($L = (L_1^T, \dots, L_k^T)^T$, we find:

$$Var(U) = L^T \Sigma X = \sum_{k=1}^{K} L_k^T \Sigma_{kk} L_k + 2 \sum_{k,j,k<j}^{K} L_k^T \Sigma_{kj} L_j \qquad (12)$$

The primary issue with GCCA can be described as a maximizing problem, which is similar to the classical scenario:

$$\emptyset(L) = tr(L^T \Sigma L)$$

so that:

$$L^T D L = I_s$$

where D is an agglomeration of diagonal matrices formed with the Σ_{kk} matrices as a diagonal mass.

This leads to the generalized eigenequation:

$$\Sigma L = D L \Delta^2$$

where, Δ^2 is a diagonal matrix made up of the major generalized eigenvalues$sof\Sigma$ with respect to the matrix D, and L is the corresponding generalized eigenvector matrix (Górecki et al., 2020).

And if we have two random vectors:

$$Y = (Y_1, Y_2, \dots, Y_p)' \in \mathbb{R}^P,$$
$$X = (X_1, X_2, \dots, X_q)' \in \mathbb{R}^q$$

One of the biggest issues with CCA is how to obtain the relationship between them. In addition, we search the weight vectors$u \in \mathbb{R}^P$ and $\upsilon \in \mathbb{R}^q$, like the components:

$$U_1 = u_{11}Y_1 + u_{12}Y_2 + \cdots + u_{1p}Y_p = u_1'Y$$
$$V_1 = \upsilon_{11}X_1 + \upsilon_{12}X_2 + \cdots + \upsilon_{1q}X_q = \upsilon_1'X \ ,$$

Which are closely associated and are referred to as the first pair of canonical variables.

When analyzing the canonical correlation of the functional data, it was found that the random processes with limited expansion have simple canonical structures, in a manner like the case of random vectors. This motivates the implementation of regularization by projecting random process onto a limited number of basic functions. The idea of projecting operations based on finite k has been discussed in (He, Müller, & Wang, 2004), and this projection is on a predetermined orthonormal basis.

To explain canonical correlations for multi-variate data, set Y(t)and $X(t)$ are two random processes. In addition,$Y \in L_2^p\ (I_1), X \in L_2^q\ (I_2)$ and each component$Y_g\ (t)$ of operation $Y\ (t)$ and $X_h\ (t)$ of operation $X\ (t)$ can be represented by a determinedset of fundamental orthogonal functions$\varphi_e$and φ_f respectively:

$$Y_{\mathrm{g}}(t) = \sum_{e=0}^{E_{\mathrm{g}}} \alpha_{\mathrm{g}e}\varphi_e(t), \ \ t \in I_1\,, \mathrm{g} = 1,2, \dots, P,$$

$$X_{\mathrm{h}}(t) = \sum_{f=0}^{F_h} \beta_{hf}\varphi_f(t), \ \ t \in I_2\,, h = 1,2, \dots, q.$$

Moreover, assuming that $E\ (Y) = 0, E\ (X) = 0$, Due to the fact that the functional canonical variables are calculated using the functions of process covariance$Y(t)$ and $X(t)$, there is no loss of generalization as a result.

Then:

$$\alpha = \left(\alpha_{10}, \dots, \alpha_{1E_1}, \dots, \alpha_{p0}, \dots, \alpha_{pE_p}\right)',$$

$$\beta = \left(\beta_{10}, \dots, \beta_{1F_1}, \dots, \beta_{q0}, \dots, \beta_{qF_q}\right)',$$

$$\Phi_1(t) = \begin{bmatrix} \varphi'_{E_1}(t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi'_{E_p}(t) \end{bmatrix},$$

$$\Phi_2(t) = \begin{bmatrix} \varphi'_{F_1}(t) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \varphi'_{F_q}(t) \end{bmatrix},$$

where $\varphi_{E_1}, \dots, \varphi_{E_p}$ and $\varphi_{F_1}, \dots, \varphi_{F_q}$are orthogonal fundamental functions in the space $L_2(I_1)$ and $L_2(I_2)$, respectively, and $K_1 = E_1 + E_2 + \cdots + E_p$, $K_2 = F_1 + F_2 + \cdots + F_q$(Górecki et al., 2018).

The processes $Y(t)$ and $X(t)$ can be expressed as follows using the aforementioned matrix:

$$Y(t) = \Phi_1(t)\alpha, \quad X(t) = \Phi_2(t)\beta.$$

The functional canonical variables$U$ and V for random processes $Y(t)$ and $X(t)$ can be defined as:

$$U = \langle u, Y\rangle = \int_{I_1} u'(t)Y(t)dt, \quad V = \langle v, X\rangle = \int_{I_2} v'(t)X(t)dt,$$

where the vector functions $u(t)$ and $v(t)$ are called vector weighting functions. The weighting functions $u(t)$ and $v(t)$ are selected to maximize the coefficients:

$$\rho = \frac{Cov(U,V)}{\sqrt{Var(U)Var(V)}} \in (0,1],$$

To be subject to:

$$Var(U) = Var(V) = 1. \qquad (13)$$

ρ is called the CC coefficient. However, implementing this maximizing does not produce a significant result. The coefficientρdeduced by $u(t)$ and $v(t)$ is 1. The weighted functions of the canonical variables $u(t)$ and $v(t)$ do not give any significant result. Thus, a technique involving data initialization and smoothing is needed. A direct way to initialize the data is to update constraints (13) by adding strong additional conditions (Ramsay & Silverman, 2005) to achieve:

$$Var\left(U^{(N)}\right) = Var\left(\int_{I_1} u'(t)Y(t)dt\right) + \lambda PEN_2(u) = 1, \qquad (14)$$

$$Var\left(V^{(N)}\right) = Var\left(\int_{I_2} v'(t)X(t)dt\right) + \lambda PEN_2(v) = 1, \qquad (15)$$

Where the Penalaized roughness function $PEN2$ is the Integrated Squared Second Derivative

$$PEN_2(u) = \int_{I_1} \left(\frac{\partial^2 u(t)}{\partial t^2}\right)' \frac{\partial^2 u(t)}{\partial t^2} dt.$$

Assuming that the weighted function $u(t)$ and the process $Y(t)$are in the same space, the function $u(t)$ can be expressed as:

$$u(t) = \Phi_1(t)\omega$$

then

$$PEN_2(u) = \int_{I_1} \left(\frac{\partial^2 \Phi_1(t)\omega}{\partial t^2}\right)' \frac{\partial^2 \Phi_1(t)\omega}{\partial t^2} dt$$

$$= \omega' \int_{I_1} \left(\frac{\partial^2 \Phi_1(t)}{\partial t^2}\right)' \frac{\partial^2 \Phi_1(t)}{\partial t^2} dt\ \omega$$

$$= \omega' R_1 \omega,$$

Where

$$R_1 = \int_{I_1} \left(\frac{\partial^2 \Phi_1(t)}{\partial t^2}\right)' \frac{\partial^2 \Phi_1(t)}{\partial t^2} dt. \qquad (16)$$

In a similar way, it can be assumed $\upsilon(t) = \Phi_2(t)v$, So we get $PEN_2(v) = v' R_2 v$, where:

$$R_2 = \int_{I_2} \left(\frac{\partial^2 \Phi_2(t)}{\partial t^2}\right)' \frac{\partial^2 \Phi_2(t)}{\partial t^2} dt. \qquad (17)$$

Now, the first canonical function ρ_1 and the corresponding vector weighting functions $u_1(t)$ and $v_1(t)$ can be defined as follows:

$$\rho_1 = \sup_{u \in L_2^P(I_1),\, v \in L_2^q(I_2)} \frac{Cov(\langle u, Y\rangle, \langle v, X\rangle)}{\sqrt{Var(U^{(N)}) Var(V^{(N)})}},$$

To be subject to:

$$Var\left(U^{(N)}\right) = Var\left(V^{(N)}\right) = 1.$$

Generally, the vector weighting functions $u_k(t)$ and $v_k(t)$, as well as the functional canonical correlation ρ_k, are defined as follows:

In general, the functional canonical correlation $\rho_k$and the corresponding vector weighting functions $u_k(t)$ and $v_k(t)$ are defined as follows:

$$\rho_k = \sup_{u \in L_2^P(I_1), v \in L_2^q(I_2)} Cov(\langle u, Y\rangle, \langle v, X\rangle)$$

$$= Cov(\langle u_k, Y\rangle, \langle v_k, X\rangle)$$

where $u_k(t)$ and $v_k(t)$ are subject to constraints (14) and (15), and the k pair of canonical variables (U_k, V_k) are uncorrelated to the first $(k-1)$ canonical variables, where the canonical variables are:

$$U_k = \langle u_k, Y\rangle, \;\; V_k = \langle v_k, X\rangle$$

This procedure is referred to as symmetric canonical correlation analysis. $(\rho_k, u_k(t), v_k(t))$is called the canonical system k of the process pair $Y(t)$ and $X\ (t)$.

Assuming:

$$Var(\alpha) = E(\alpha\alpha') = \Sigma_{11},$$

$$Var(\beta) = E(\beta\beta') = \Sigma_{22},$$

$$Cov(\alpha, \beta) = E(\alpha\beta') = \Sigma_{12}.$$

Taking into account the canonical variables $U^* = \langle \omega, \alpha\rangle$ and $V^* = \langle v, \beta\rangle$of the random vectors α and β respectively, the canonical correlation k $(\gamma_k)$and the associated vector weights $\omega_k$and ν_k are defined as follows:

$$\gamma_k = \sup_{\omega \in \mathbb{R}^{K_{1+P}}, v \in \mathbb{R}^{K_2+q}} Cov(\langle \omega, \alpha\rangle, \langle v, \beta\rangle) = \omega_k' \Sigma_{12} v_k,$$

To be subject to:

$$\omega_k'(\Sigma_{11} + \lambda R_1)\omega_k = 1,$$

$$v_k'(\Sigma_{22} + \lambda R_2)v_k = 1,$$

Where R_1 and R_2 are determined in equations (16) and (17) respectively, the canonical variables k $(U_k^*, V_k^*)$are not related to the first $k-1$ canonical variables. The expression $(\gamma_k, \omega_k, v_k)$ is named the canonical system k of random vectors α and β(Górecki et al., 2018).

The canonical system k $(\rho_k, u_k(t), v_k(t))$ of the pair of random processes $Y(t)$ and $X(t)$ is related to the canonical system k $(\gamma_k, \omega_k, v_k)$ of the pair of random vectors α and β by equations:

$$\rho_k = \gamma_k,\ u_k(t) = \Phi_1(t)\omega_k, t \in I_1, v_k(t) = \Phi_2(t)v_k, t \in I_2,$$

Where:

$$1 \le k \le \min(K_1 + p, \qquad K_2 + q),$$

$$K_1 = E_1 + \cdots + E_p,$$

$$K_2 = F_1 + \cdots + F_q.$$

CCA of the random vectors α and β is depend on the unknown Σ_{11}, Σ_{22} and Σ_{12} matrices. They are estimated based on n independent results$y_1(t), y_2(t), \dots, y_n(t)$,which have the form $y_i(t) = \Phi_1(t)\hat{\alpha}_i$ of the random process $Y(t)$ and $x_1(t), x_2(t), \dots, x_n(t)$ which has the form$x_i(t) = \Phi_2(t)\hat{\beta}_i$ for the random process $X(t), i = 1,2, \dots, n$, where:

$$\hat{\alpha}_i = \left(\hat{\alpha}_{10}^{(i)}, \dots, \hat{\alpha}_{1E_1}^{(i)}, \dots, \hat{\alpha}_{p0}^{(i)}, \dots, \hat{\alpha}_{pE_p}^{(i)}\right)',$$

$$\hat{\beta}_i = \left(\hat{\beta}_{10}^{(i)}, \dots, \hat{\beta}_{1F_1}^{(i)}, \dots, \hat{\beta}_{q0}^{(i)}, \dots, \hat{\beta}_{qF_q}^{(i)}\right)'.$$

Where:

$$\hat{A} = (\hat{\alpha}_1, \dots, \hat{\alpha}_n)',$$

$$\hat{B} = \left(\hat{\beta}_1, \dots, \hat{\beta}_n\right)',$$

Finally, the estimators of the matrices Σ_{11}, Σ_{22} and Σ_{12} take the form:

$$\hat{\Sigma}_{11} = \frac{1}{n}\hat{A}'\hat{A},$$

$$\hat{\Sigma}_{22} = \frac{1}{n}\hat{B}'\hat{B},$$

$$\hat{\Sigma}_{12} = \frac{1}{n}\hat{A}'\hat{B}.$$

Assuming $\hat{C} = \hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12}$ and $\hat{D} = \hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$where $\hat{\Sigma}_{12}' = \hat{\Sigma}_{21}$, then the matrices $\hat{C}\hat{D}$ and $\hat{D}\hat{C}$ have the same non-zero eigenvalues $\hat{\gamma}_k^2$, and its corresponding eigenvectors $\hat{\omega}_k$and $\hat{v}_k$can be determined by the equations:

$$\left(\hat{C}\hat{D} - \hat{\gamma}_k^2 I_{K_{1+p}}\right)\hat{\omega}_k = 0,$$

$$\left(\hat{D}\hat{C} - \hat{\gamma}_k^2 I_{K_{2+q}}\right)\hat{v}_k = 0.$$

$$1 \le k \le \min(K_1 + p, K_2 + q).$$

Hence, the projection coefficients of the result i$y_i(t)$ of process $Y(t)$ of the canonical function k are:

$$\widehat{U}_{ik} = \langle \hat{u}_k, y_i \rangle = \int_{I_1} \hat{u}_k(t) y_i(t) dt = \hat{\alpha}_i' \widehat{\omega}_k,$$

Similarly, the projection coefficients for the result i$x_i(t)$ of process $X(t)$ for the canonical function k are:

$$\hat{V}_{ik} = \hat{\beta}_i' \hat{v}_k,$$

Where:$i = 1,2, \dots, n, k = 1, \dots, \min(K_1 + p, K_2 + q)$

9- Principal component analysis (PCA)

PCA is widely used for data reduction(Helena et al., 2000; Tanasković, Golobocanin, & Miljević, 2012). This is accomplished by transforming the data into a new set of principal components, which are derived from linear combinations of the original variables, and categorized in such a way that the first principal components are responsible for the most variance in all the original variables(Bošnjak et al., 2012; Charfi, Zouari, Feki, & Mami, 2013).

PCA is interested in explaining the structure of variances and covariances of the original variables using a few linear combinations of these variables. Since obtaining the same total variance requires the use of p of the principal components, the use of only a few k components is usually sufficient to obtain the largest part of the total variance. The use of PCA often leads to the detection of previously unthinkable relationships, allowing us to come up with interpretations that would not usually be obtained without this method.

In addition, functional principal component has a major role to play in the representation of stochastic functions and in supervised or unsupervised learning tasks(Dai & Müller, 2018). The covariance function plays an important role in functional principal component analysis (FPCA)(Ramsay & Silverman, 2005). The main difference between the function of covariance in functional data and the matrix of covariance in multivariate data is that the functional data are measured to the same scale, with large overlaps and possibly irregular sampling. The order of the functional observations is also important but can be easily dealt with by accurate indexing (Xiao et al, 2016).

There are three methods for estimating functional principal component, which are; The first approach is to initialize the functional principal components of the sample covariance function, the second is to initialize the covariance function and then diagonalize it, and the third is to initialize each curve and diagonal the function of covariance of the fitted curves(Xiao, Zipunnikov, Ruppert, & Crainiceanu, 2016).

By reducing the number of correlated variables in a data set while retaining as much variance as possible, PCA aims to reduce the dimensions of large-scale correlated data sets. By using a new set of variables and uncorrelated principal components that are ordered so that the initial principal components preserve the majority of the variation present in all of the original variables, the desired result is achieved.So, if we have a random vector of p dimensions $\boldsymbol{X} = \left(\boldsymbol{X_1}, \boldsymbol{X_2}, \dots, \boldsymbol{X_p}\right)' \in \mathbb{R}^{\boldsymbol{p}}$. In the first stage, a linear combination $\boldsymbol{U_1} = \boldsymbol{u_{11}X_1} + \boldsymbol{u_{12}X_2} + \cdots + \boldsymbol{u_{1p}X_p} = \boldsymbol{u_1'X}$obtained for the elements of vector X that have maximum variance, and the variable U1 is named the first principal component. Next, a linear combination $\boldsymbol{U_2} = \boldsymbol{u_2'X}$ obtained, which has the highest variance, is unrelated to the first principle component U1, and so on, until stage k, where a linear combination result.$\boldsymbol{U_k} = \boldsymbol{u_k'X}$ is found, called the principle component k, which is independent of the first k-1 principal components and has the greatest variance(Jolliffe, 2002).

The observations can be represented graphically as points on a plane (U_1, U_2). The functional case of principal component analysis (FPCA) is a more informative method for structuring the variance-covariance matrix of one-dimensional functional data (Jacques & Preda, 2014). We assume $E(X) = 0$ without losing generality. Thus, in the multivariate functional case, we are concerned with determining the inner product when analysing the principal component:

$$U = \langle u, X \rangle = \int_I u'(t)X(t)dt$$

which has the maximum variance of all $u \in L_2^p(I)$, where $\langle u, u \rangle = 1$. Assuming that:

$$\lambda_1 = \sup_{u \in L_2^p(I)} var(\langle u, X \rangle) = var(\langle u_1, X \rangle),$$

where $\langle u_1, u_1 \rangle = 1$. The first principal component is found using inner multiplication $U_1 = \langle u_1, X \rangle$, and the vector function u_1 is called the weighted function of the first vector. Next, we look for the second principal component$U_2 = \langle u_2, X \rangle$, which maximizes $Var(\langle u, X \rangle)$, such that $\langle u_2, u_2 \rangle = 1$, and is not related to the first function principal componentU_1, i.e. subject to the constraint $\langle u_1, u_2 \rangle = 0$.

In general, the function principal component$U_k = \langle u_k, X \rangle$ satisfies the conditions:

$$\lambda_k = \sup_{u \in L_2^p(I)} var(\langle u, X \rangle) = var(\langle u_k, X \rangle),$$

$$\langle u_{k_1}, u_{k_2} \rangle = \delta_{k_1 k_2}, \quad k_1, k_2 = 1, \dots, k,$$

Where*:*

$$\delta_{k_1 k_2} = \begin{cases} 1 & if \;\; k_1 = k_2 \\ 0 & if \;\; k_1 \neq k_2. \end{cases}$$

The term $(\lambda_k, u_k(t))$ is called the main system k of the process $X(t)$.

In the second section, it turns out that the process $X(t)$ can be represented as $X(t) = \Phi(t)c, t \in I$, and we now consider the principal component of the random vector c. Since $E\ (X) = 0$ then $E(c)\ =\ 0$. If we denote $Var(c)$ as $\sum$, the k principal component$U_k^* = \langle \omega_k, c \rangle$ of this vector satisfies the conditions:

$$\gamma_k = \sup_{\omega \in \mathbb{R}^{k+p}} Var(\langle \omega, c \rangle) = \sup_{\omega \in \mathbb{R}^{k+p}} \omega' Var(c) \omega$$

$$= \sup_{\omega \in \mathbb{R}^{k+p}} \omega' \Sigma \omega = \omega_k' \Sigma \omega_k,$$

$$\omega_{k_1}' \omega_{k_2} = \delta_{k_1 k_2},$$

Where:

$$k_1, k_2 = 1, \dots, k, K = B_1 + \cdots + B_p$$

The term $(\gamma_k, \omega_k)$is known as the k main system of vector c.

Solving the eigenvalue and associated eigenvectors of the covariance matrix $\sum$ of this vector provide the major order k of vector c, normalized such that $\omega_{k_1}' \omega_{k_2} = \delta_{k_1 k_2}$.

The main system $(\lambda_k, u_k(t))$of the random process $X(t)$ is related to the main system (γ_k, ω_k) of the random vector c by the equation:

$$\lambda_k = \gamma_k, u_k(t) = \Phi(t) \omega_k, t \in I,$$

Where:

$$k = 1, \dots, s \ and \ s = rank(\Sigma)$$

The PCA of the random vector c is based on the matrix $\sum$. In practice, this matrix is unknown. This is estimated on the basis of n independent outcomes $x_1(t), x_2(t), \dots, x_n$ which has the form $x_i(t) = \Phi(t)\ \hat{c}_i$ of the random process $X(t)$, where the vectors $\hat{c}_i$ are centered, $i = 1,2, \dots, n$(Górecki et al., 2018).

Assuming that $\widehat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_{\boldsymbol{1}}, \hat{\boldsymbol{c}}_{\boldsymbol{2}}, \dots, \hat{\boldsymbol{c}}_{\boldsymbol{n}})'$, then:

$$\hat{\Sigma} = \frac{1}{n} \hat{C}' \hat{C}.$$

Also, suppose that $\hat{\gamma}_1 \geq \hat{\gamma}_2 \geq \cdots \geq \hat{\gamma}_s$are non-zero eigenvalues of the matrix $\hat{\Sigma}$ and $\widehat{\omega}_1, \widehat{\omega}_2, \dots, \widehat{\omega}_s$ are the corresponding eigenvectors, where $s = rank\big(\hat{\Sigma}\big)$.

Further, the base system k of the random process $X(t)$ selected from the sample has the following form:

$$\left(\hat{\lambda}_k = \hat{\gamma}_k, \hat{u}_k(t) = \Phi(t)\widehat{\omega}_k\right), k = 1, \dots, s.$$

Hence, the projection coefficients for result i of $x_i(t)$of operation $X(t)$ of the functional principal component k are:

$$\widehat{U}_{ik} = \langle \hat{u}_k, x_i \rangle = \int_I \widehat{\omega}'_k \Phi'(t)\Phi(t)\hat{c}_i dt$$

$$= \widehat{\omega}'_k \int_I \Phi'(t)\Phi(t)dt\hat{c}_i = \widehat{\omega}'_k \hat{c}_i,$$

Where: $i = 1,2, \dots, n$, $k = 1,2, \dots, s$.

Finally, the projection coefficients of result i of $x_i(t)$ of process $X(t)$ at the level of the first two functional principal components of the sample are equal to $(\widehat{\omega}'_1\hat{c}_i, \widehat{\omega}'_2\hat{c}_i)$, $i = 1,2, \dots, n$.

10- Applied study

The RGCCA statistical package was relied upon to model the data using the two basic research methods, GCCA and PCA, which are based on three main elements:

- A scheme function (g), that is, a continuous convex function, allows the consideration of different optimization criteria. Typical choices for g are the horst scheme, which leads to maximization of the sum of covariances among group components, the absolute value (a centroid scheme, which leads to maximization of the sum of covariances), and the square function (a factorial scheme, thus maximizing the sum squared covariance).

- The design matrix (C), which is $J \times J$ symmetric matrix of non-negative elements that describes the communication network between the blocks being studied. Normally, $c_{jk} = 1$ for two connected groups and zero otherwise.

- Shrinkage parameters τ_j, ranging in value from 0 to 1 and interpolated by smoothing between covariance maximization and correlation maximization. Setting τ_j to zero forces the group components to unit variance $X_j a_j = 1$, in which case the covariance criterion is correlation. The correlation criterion is better at explaining the correlated structure across data sets, thus ignoring variance within each individual data set. Setting $\tau_j$to 1 causes the block weight vectors to be normally distributed $\left(a_j{}^T a_j = 1\right)$, which applies the covariance criterion. A value between zero and 1 results in a compromise between these two options. When the value of tau is equal to one, this method of analysis is referred to as (Mode A), while if the value of tau is equal to zero, this method of analysis is referred to as (Mode B).

The two methods of GCCA and PCA and other methods for modeling blocks of variables are based on the three axes (design matrix, scheme function (g) and the value of the tau parameter). Then it is possible to compare these methods based on the three axes as follows:

Mode	*Horst*	*Centroid*	*Factorial*
Mode A	*SUMCOV*	*SABSCOV*	*SSQCOV*
Mode B	*SUMCOR*	*SUMCOR*	*SSQCOR*

Table (1): Methods for studying the relationship between groups based on tau and the scheme function.

In addition to the previous methods, the two basic study methods are, GCCA and PCA.

10-1: Data used in the research

Actual data(XLStat, 2022):The research dealt with industry performance data for a mobile phone manufacturing company to analyze the degree of consumer satisfaction. The data included twenty-three variables (measured as a percentage), divided into six groups (latent variables). Whereas, the first group contains five basic variables, the second contains three variables, the third contains seven variables, while the fourth group contains two variables, the fifth contains three variables, and the last group contains three variables. It was dealt with:

- Model actual data using GCCA and PCA.

- Conducting confirmatory factor analysis (CFA) to study the relationship between the basic variables and latent variables and studying the extent of the contribution of confirmatory factor analysis in improving the results of the study methods.

- Smoothing the data using Fourier's rule at different cut-off values from 0.2 to 0.9.

- Modeling smoothed data using GCCA and PCA.

- Conducting CFA of smoothed data to study the relationship between basic and latent variables.

Simulation study:A simulation study of fifteen variables was performed for a time series (500 days). These variables are included in three groups, where the first group contains five variables, the second group contains four variables, and finally the third group includes six variables. The simulation study dealt with:

- Modeling the simulated data based on different correlation matrices, using GCCA and PCA.

- Smoothing the data using Fourier's rule at different cut-off values from 0.2 to 0.9.

- Modeling smoothed data based on different correlation matrices, using GCCA and PCA.

10-2:Results of the actual study

10-2-1: The results of some modeling methods for the actual data that were presented in the RGCCA statistical package:

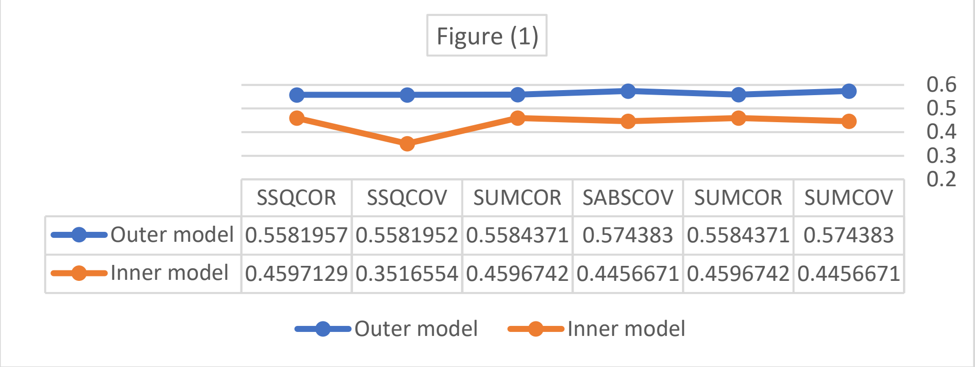


Figure (1)

	SSQCOR	SSQCOV	SUMCOR	SABSCOV	SUMCOR	SUMCOV
Outer model	0.5581957	0.5581952	0.5584371	0.574383	0.5584371	0.574383
Inner model	0.4597129	0.3516554	0.4596742	0.4456671	0.4596742	0.4456671

From the results of Figure (1), we find a very large convergence in the results of the methods SUMCOV, SUMCOR, SABSCOV, and SSQCOR. Then they are followed by the SSQCOV method.

10-2-2:Actual data modeling results before smoothing based on PCA and GCCA methods:

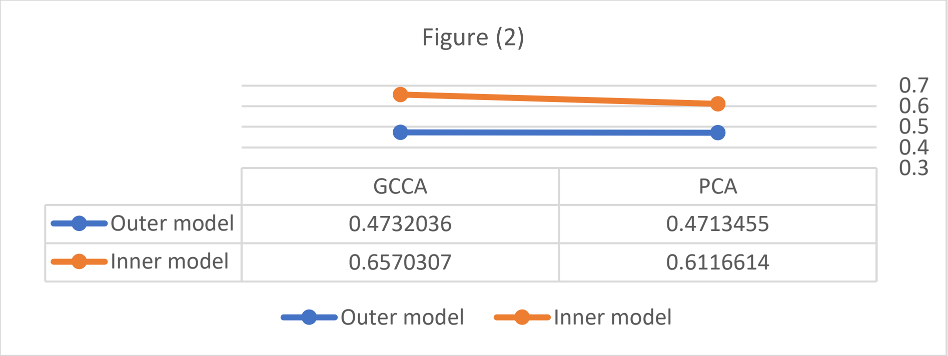


Figure (2)

	GCCA	PCA
Outer model	0.4732036	0.4713455
Inner model	0.6570307	0.6116614

According to Figure (2), we find that the GCCA method is superior to the PCA method. In comparison between these two methods and the previous modeling methods, it was found that the PCA and GCCA methods are superior to the previous modeling methods that were dealt with in the RGCCA statistical package.

10-2-3: Modeling the actual data before smoothing, and after excluding variables with MI greater than 10 based on the PCA and GCCA methods:

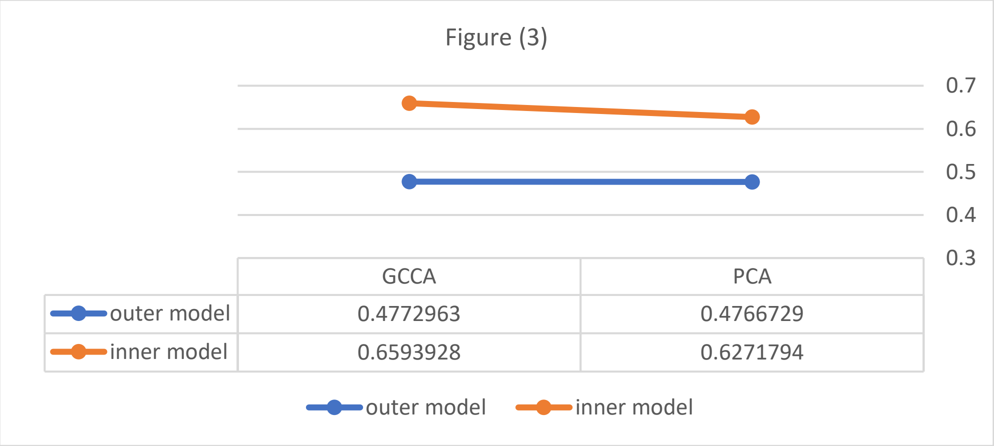


Figure (3)

	GCCA	PCA
outer model	0.4772963	0.4766729
inner model	0.6593928	0.6271794

According to Figure (3), we find that the GCCA method is superior to PCA. By comparing the modeling results of these two methods before and after confirmatory factor analysis (CFA), Figure (4) shows an increase in the average variance explained for both methods after making the modifications approved by CFA, but it still preferred in favor of GCCA method.

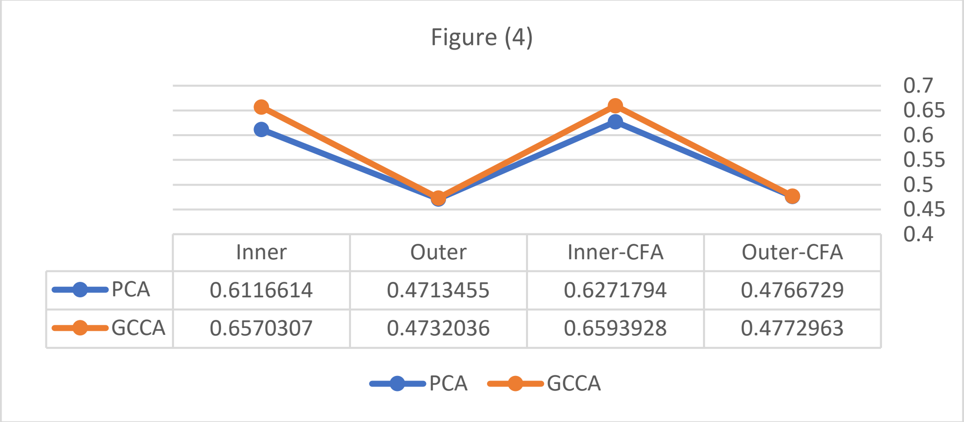


Figure (4)

	Inner	Outer	Inner-CFA	Outer-CFA
PCA	0.6116614	0.4713455	0.6271794	0.4766729
GCCA	0.6570307	0.4732036	0.6593928	0.4772963

10-2-4: Results of the average variance explained for smoothed actual data based on PCA and GCCA methods for different α values:

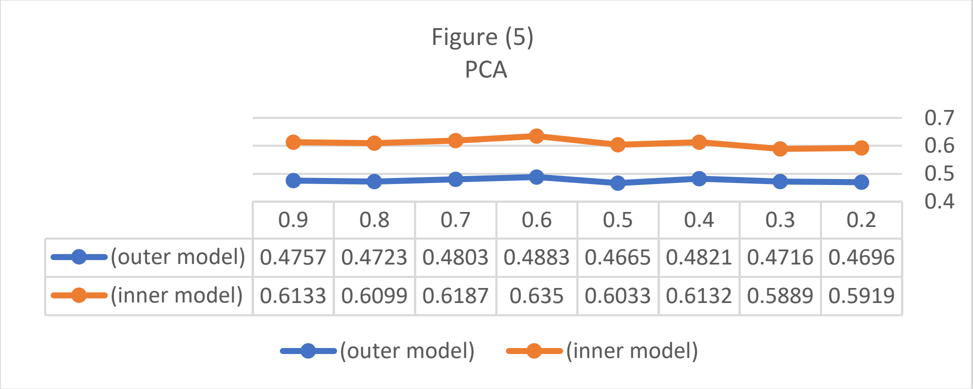


Figure (5)
PCA

	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
(outer model)	0.4757	0.4723	0.4803	0.4883	0.4665	0.4821	0.4716	0.4696
(inner model)	0.6133	0.6099	0.6187	0.635	0.6033	0.6132	0.5889	0.5919

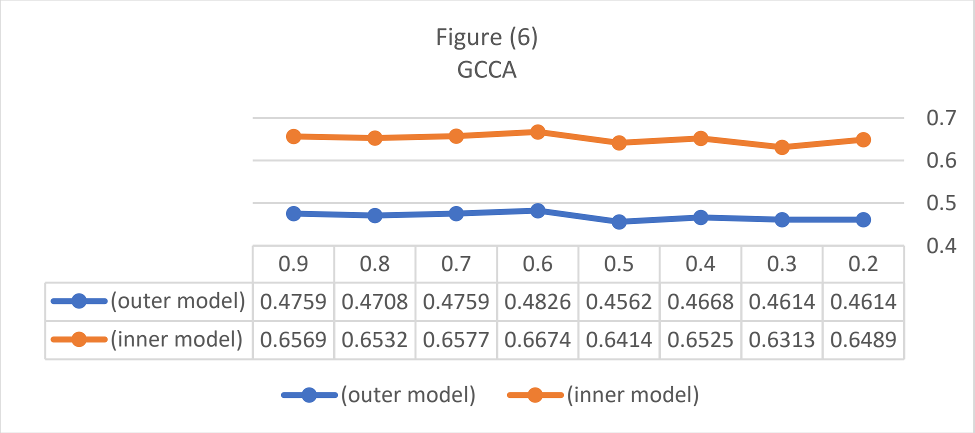


Figure (6)
GCCA

	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
(outer model)	0.4759	0.4708	0.4759	0.4826	0.4562	0.4668	0.4614	0.4614
(inner model)	0.6569	0.6532	0.6577	0.6674	0.6414	0.6525	0.6313	0.6489

By examining the modeling results for the smoothed data at different α values, we find that the explanatory power of the smoothed data is greatly increased according to the PCA and GCCA modeling methods. However, the highest values of the explained average variance were for the smoothed data at the value of α = 0.6, as shown in Figures (5) and (6).

10-2-5: Modeling the smoothed actual data after confirmatory factor analysis:

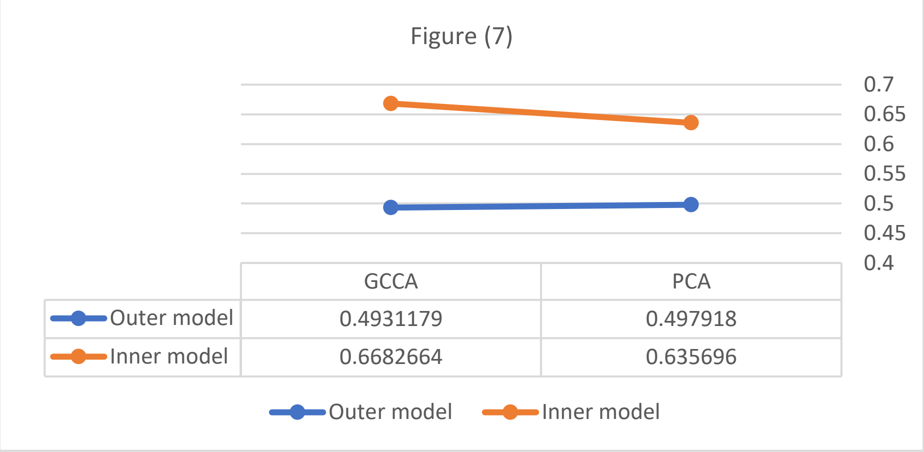


Figure (7)

	GCCA	PCA
Outer model	0.4931179	0.497918
Inner model	0.6682664	0.635696

Also, according to Figure (7), it is clear that the average variance explained by the two study methods increased after making adjustments by excluding some specific variables from CFA modeling.

10-3: Simulation Study Results:

The simulation study data was modeled according to the following:

- Modeling of the simulation studybefore smoothing using methods PCA and GCCA methods, depending on the correlation matrices (a), (b) and (c)[1].
- Modeling of the simulation studyafter smoothing using methods PCA and GCCA methods, depending on the correlation matrices (a), (b) and (c).

10-3-1: Modeling the data of the simulation study before smoothing using PCA and GCCA methods, depending on the correlation matrix (a):

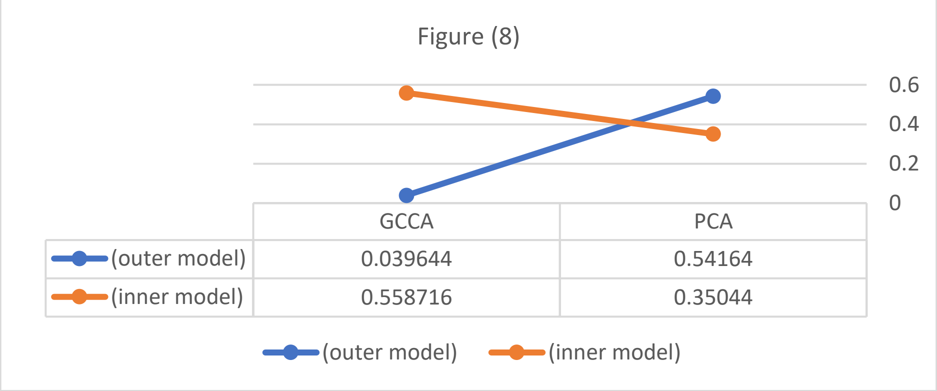


Figure (8)

	GCCA	PCA
(outer model)	0.039644	0.54164
(inner model)	0.558716	0.35044

[1] In the correlation matrix (a), the relationships between variables within the group are strong, while the values of the cross-correlation coefficients between groups variables are weak. The correlation matrix (b) is based on very close correlation coefficients between most groups' variables, both within the group and between groups. That is, the correlations within groups are moderate with some active variables in the relationships between groups. The correlation matrix (c) is based on strong correlation coefficients within groups, with some active variables in the relationships between groups.

According to the data of the simulation study based on the correlation matrix (a), we find that the average variance explained by the PCA method exceeds the average variance explained by the GCCA method, and this is evident in Figure (8).

10-3-2: Modeling the data of the simulation study before smoothing using PCA and GCCA methods, depending on the correlation matrix (b):

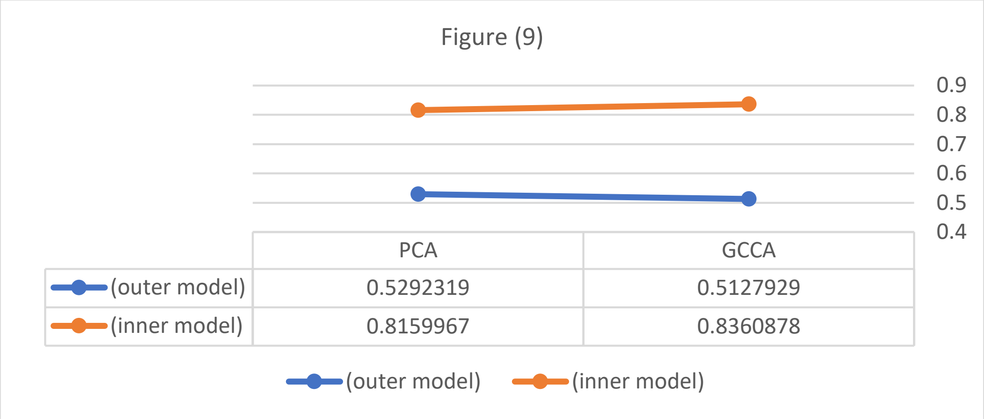


Figure (9)

	PCA	GCCA
(outer model)	0.5292319	0.5127929
(inner model)	0.8159967	0.8360878

According to the data of the simulation study based on the correlation matrix (b), we find a great convergence for the results of the two study methods, PCA and GCCA, although the difference is very small in favor of the GCCA method. This is illustrated in Figure (9).

10-3-3: Modeling the data of the simulation study before smoothing using PCA and GCCA methods, depending on the correlation matrix (c):

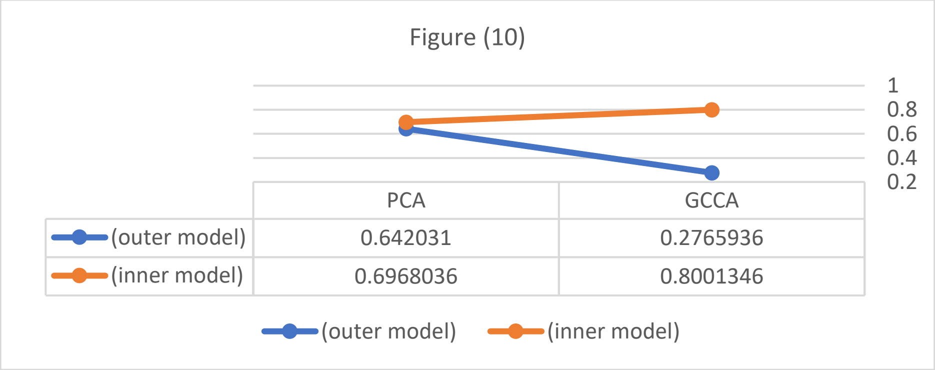


Figure (10)

	PCA	GCCA
(outer model)	0.642031	0.2765936
(inner model)	0.6968036	0.8001346

According to the data of the simulation study based on the correlation matrix (c), we find a great convergence for the results of the two study methods, PCA and GCCA, although the difference is very small in favor of the PCA method. This is illustrated in Figure (10).

10-3-4: Results of the explained average variance of the PCA and GCCA methods for the smoothed simulation study data, depending on the correlation matrix (a) for different α values:

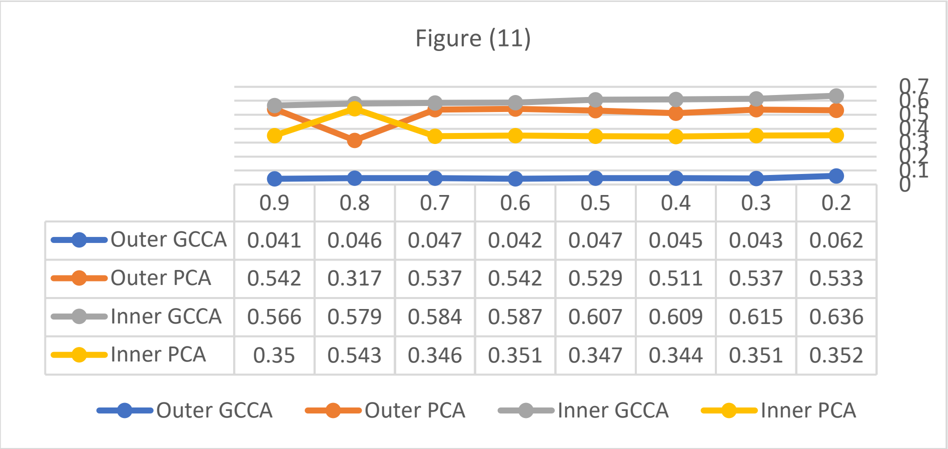


Figure (11)

	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
Outer GCCA	0.041	0.046	0.047	0.042	0.047	0.045	0.043	0.062
Outer PCA	0.542	0.317	0.537	0.542	0.529	0.511	0.537	0.533
Inner GCCA	0.566	0.579	0.584	0.587	0.607	0.609	0.615	0.636
Inner PCA	0.35	0.543	0.346	0.351	0.347	0.344	0.351	0.352

It is clear from Figure (11) regarding the results of modeling the smoothed data for the simulation study based on the correlation matrix (a), a slight improvement in the explained average variance for the smoothed data. It is also evident that the best results are obtained when α = 0.2.

10-3-5: Results of the explained average variance of the PCA and GCCA methods for the smoothed simulation study data, depending on the correlation matrix (b) for different α values:

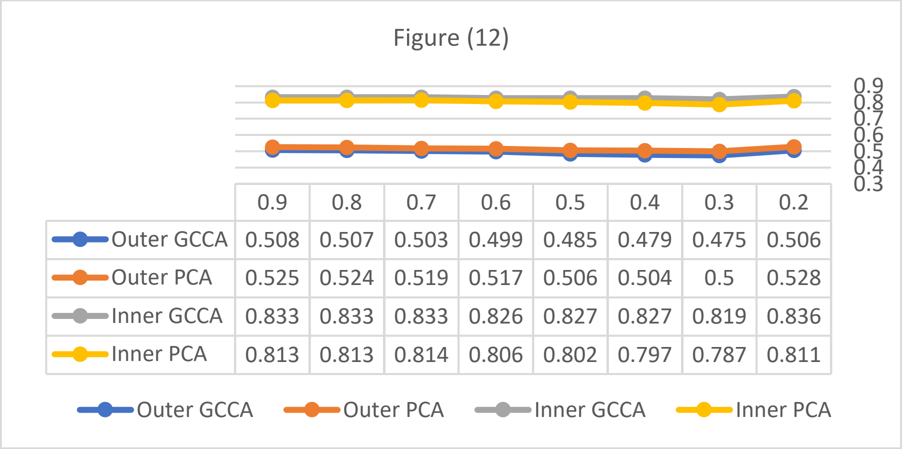


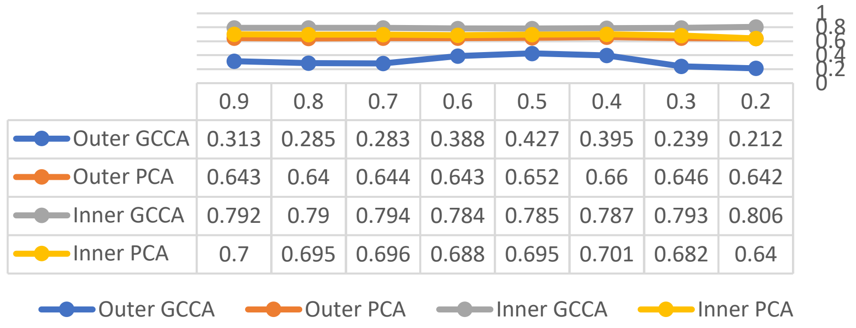
Figure (12)

	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
Outer GCCA	0.508	0.507	0.503	0.499	0.485	0.479	0.475	0.506
Outer PCA	0.525	0.524	0.519	0.517	0.506	0.504	0.5	0.528
Inner GCCA	0.833	0.833	0.833	0.826	0.827	0.827	0.819	0.836
Inner PCA	0.813	0.813	0.814	0.806	0.802	0.797	0.787	0.811

It is clear from Figure (12) about the results of modeling the smoothed data for the simulation study based on the correlation matrix (b), a slight improvement in the explained average variance for the smoothed data. It is also evident that the best results are obtained when α = 0.2.

10-3-6: Results of the explained average variance of the PCA and GCCA methods for the smoothed simulation study data, depending on the correlation matrix (c) for different α values:

Figure (13)



	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
Outer GCCA	0.313	0.285	0.283	0.388	0.427	0.395	0.239	0.212
Outer PCA	0.643	0.64	0.644	0.643	0.652	0.66	0.646	0.642
Inner GCCA	0.792	0.79	0.794	0.784	0.785	0.787	0.793	0.806
Inner PCA	0.7	0.695	0.696	0.688	0.695	0.701	0.682	0.64

It is clear from Figure (12) about the results of modeling the smoothed data for the simulation study based on the correlation matrix (c), a slight improvement in the explained average variance for the smoothed data. It is also evident that the best results are obtained when α = 0.5.

11- A summary of results

- PCA and GCCA methods provided better results than the methods covered in the GCCA Statistical Package.

- In the actual study data, the explanatory power of the model based on the GCCA method was superior to the PCA method. Whereas, the simulation study presented additional results indicating that it is not always possible to assert the superiority of GCCA method over the PCA method. The simulation study indicated that the matter depends on the nature of the correlation matrix of the relationship between the basic variables. If the relationship between the variables within each group is strong, and the interrelationship between the variables in the different groups is weak, it is preferable to perform modeling using the PCA method. The simulation study confirmed that, in order to model the data using generalized canonical analysis, there must be activation variables between the groups, i.e. in the sense that there are linked variables between the groups that activate the relationship between the latent variables. Moreover, if the exploratory analysis of the number of principal components indicates that the number of components corresponds to the number of study groups, and the full saturation of the essential variables with each component is consistent with the actual division of the principal variables within each group, then the preference is given to the PCA method. Whereas, if the relationships between all baseline variables “within and between groups” are similar, both methods give similar results.

- CFA confirmatory factor analysis always adds important results to exploratory analysis methods for multiple data sets, as it increases the power of the model and its explanatory capacity.

- Smoothing the data in the actual study significantly increased the explanatory power, compared to the simulation study data. The smoothing process did not provide a significant addition to the data of the simulation study, and the researcher believes that the reason for the weak effect of smoothing the data of the simulation study is that the data was basically generated according to a certain distribution, and therefore it represents functional data, without the need to perform the smoothing process.

- Do not exaggerate the smoothing process, as the results of the actual study indicated that the best results for the smoothed data were at the cutoff parameter 0.6.

12- Recommendations

- Paying attention to the methods of modeling multiple data sets, as they represent important, more comprehensive statistical tools used to analyze more than two sets of variables at one time. This corresponds to the revolution of information and big data. For example, if there is a large group of images to be clustered according to the similarity between them through a vector for each image that represents its characteristics, then PCA is an alternative measure to well-known data clustering methods such as k-means, DBSCAN, and others.

- Paying attention to more studies towards functional data, as the functional analysis of data is characterized by maintaining the order of data and following up on its development of a function through a continuum. Thus, it provides different insights that are difficult to find with other analyses.

- Always be careful to smooth the time series data, as smoothing the actual data always provides more appropriate results and higher explanatory power.

- Use confirmatory factor analysis (CFA) to improve the results of exploratory analysis methods for multiple data sets.

13- Proposed areas of research

• Studying the effect of the number of groups and the dimensions within each group (dimensions) on the generalized canonical correlation analysis and the principal components analysis.

• Study the effect of outliers on generalized canonical correlation analysis and principal components analysis.

• Studying the missing data processing methods in the generalized cone correlation analysis.

References:

1. Aneiros, G., Cao, R., Fraiman, R., Genest, C., & Vieu, P. (2019). Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis, 170*, 3-9.
2. Bošnjak, M. U., Capak, K., Jazbec, A., Casiot, C., Sipos, L., Poljak, V., & Dadić, Ž. (2012). Hydrochemical characterization of arsenic contaminated alluvial aquifers in Eastern Croatia using multivariate statistical techniques and arsenic risk assessment. *Science of the Total Environment, 420*, 100-110.
3. Carroll, J. D. (1968). *Generalization of Canonical Correlation Analysis to Three of More Sets of Variables*.

4. Charfi, S., Zouari, K., Feki, S., & Mami, E. (2013). Study of variation in groundwater quality in a coastal aquifer in north-eastern Tunisia using multivariate factor analysis. *Quaternary International, 302*, 199-209.
5. Dai, X., & Müller, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *The Annals of Statistics, 46*(6B), 3334-3361.
6. Dirac, P. (2012). *Spinors in Hilbert space*: Springer Science & Business Media.
7. Górecki, T., Krzyśko, M., Waszak, Ł., & Wołyński, W. (2018). Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers, 59*(1), 153-182.
8. Górecki, T., Krzyśko, M., & Wołyński, W. (2020). Generalized canonical correlation analysis for functional data. *Biometrical Letters, 57*(1), 1-12.
9. Hanusz, Z., Krzyśko, M., Nadulski, R., & Waszak, Ł. (2020). Discriminant coordinates analysis for multivariate functional data. *Communications in Statistics-Theory Methods, 49*(18), 4506-4519.
10. He, G., Müller, H.-G., & Wang, J.-L. (2004). Methods of canonical analysis for functional data. *Journal of Statistical Planning Inference, 122*(1-2), 141-159.
11. Helena, B., Pardo, R., Vega, M., Barrado, E., Fernandez, J. M., & Fernandez, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water research, 34*(3), 807-816.
12. Helwig, N. E., Hong, S., & Polk, J. D. (2012). Parallel Factor Analysis of gait waveform data: A multimode extension of Principal Component Analysis. *Human movement science, 31*(3), 630-648.
13. Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications* (Vol. 200): Springer Science & Business Media.
14. Ilin, A., & Raiko, T. (2010). Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research, 11*, 1957-2000.
15. Jacques, J., & Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics Data Analysis, 71*, 92-106.
16. Jolliffe, I. T. (2002). *Principal component analysis for special types of data*: Springer.
17. Khan, A., & Farooq, H. (2012). Principal component analysis-linear discriminant analysis feature extractor for pattern recognition. *arXiv*.
18. Markos, A., & D'Enza, A. I. (2016). Incremental Generalized Canonical Correlation Analysis. In *Analysis of Large and Complex Data* (pp. 185-194): Springer.
19. Pavlidis, P., Weston, J., Cai, J., & Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *Journal of computational biology, 9*(2), 401-411.
20. Pearson, K. (1901). Principal components analysis. *The London, Edinburgh, Dublin Philosophical Magazine Journal of Science, 6*(2), 559.
21. Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis*.
22. Tanasković, I., Golobocanin, D., & Miljević, N. (2012). Multivariate statistical analysis of hydrochemical and radiological data of Serbian spa waters. *Journal of Geochemical Exploration, 112*, 226-234.

23. Wang, J.-L., Chiou, J.-M., & Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics its application, 3*, 257-295.
24. Wang, Y., Wang, L., Yang, D., & Deng, M. (2014). Imputing missing values for genetic interaction data. *Methods, 67*(3), 269-277.
25. Wang, Y., Wang, P., Bai, Y., Tian, Z., Li, J., Shao, X., . . . Li, B.-L. (2013). Assessment of surface water quality via multivariate statistical techniques: a case study of the Songhua River Harbin region, China. *Journal of hydro-environment research, 7*(1), 30-40.
26. Xiao, L., Zipunnikov, V., Ruppert, D., & Crainiceanu, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics computing*, *26*(1), 409-421.
27. XLStat. (2022). https://help.xlstat.com/6479-consumer-satisfaction-analysis-excel-plspm.