Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Studying the Hidden Relationships in Mixed Data Via Principal Component Analysis with Application in Traumatic Brain Injury Data

Zakiah I. Kalantan*, **Rafal Z. Alqarni, Hanan Baaqeel**

Department of Statistics, Faculty of Sciences, King Abdulaziz University, Jeddah, 21589, Saudi Arabia
E-mail: zkalanten@kau.edu.sa

Received: 4 November 2024; **Revised:** 17 December 2024; **Accepted:** 24 December 2024

Abstract: The analysis of complex biomedical datasets often involves a mix of numerical and categorical variables, posing challenges for traditional statistical techniques. To address this limitation, this study proposes the using of Principal Component Analysis for Mixed Data (PCAmix). PCAmix is a powerful technique that can effectively reduce the dimensionality of complex datasets while preserving the most important information. By combining the strengths of Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA), PCAmix can handle both numerical and categorical variables simultaneously. This flexibility allows for a more comprehensive analysis of complex datasets, particularly in biomedical research. In this study, we applied PCAmix to a real-world biomedical dataset to investigate the intricate relationship between brain injury, functional outcomes, and genetic factors. The results we obtained illustrate not only the efficacy of PCAmix but also its practical uses in recognizing underlying frameworks, streamlining analysis by minimizing the number of variables while retaining essential information, creating predictive models to anticipate patient results, including functional recovery and cognitive deficits, and categorizing patients according to shared traits to facilitate tailored treatment approaches. Through the applying PCAmix, we gained a deeper understanding of the complex interplay between these factors and identified potential biomarkers for predicting patient outcomes. These findings have significant implications for the development of more effective diagnostic tools, prognostic models, and therapeutic interventions for traumatic brain injury. Ultimately, researchers can contribute to advancements in healthcare and medicine by unlocking valuable insights from complex biomedical data via leveraging the potential of PCAmix.

***Keywords*:** principal component analysis, PCA mix, MCA, traumatic brain injury (TBI), hidden variables

MSC: 62H25, 62P10, 65S05

1. Introduction

Multivariate analysis is an essential methodology approach in the field of data science and statistics, particularly for the objective of comprehending and interpreting elaborate datasets. As data grows in mass and complexity, the capacity to examine several factors concurrently becomes more crucial. This analytical framework enables researchers and practitioners to reveal complex interactions among variables, identify patterns, and extract significant insights frequently hidden when analyzing individual variables in isolation. Central to multivariate analysis are machine-

DOI: https://doi.org/10.37256/cm.6120256017

learning algorithms, which may be generally categorized into two main types: supervised and unsupervised approaches. Supervised learning approaches rely on labeled data, meaning the model is trained using a dataset with predetermined outcomes or target variables. This configuration allows the algorithm to recognize the fundamental patterns linking input data to known outcomes, making it proficient for classification, regression, and forecasting jobs. In a supervised learning context, a model may be trained on past data to forecast client behavior using multiple attributes such as age, income, and prior purchases [1-4].

Despite supervised learning, unsupervised approaches function without pre-labeled data, concentrating on identifying latent patterns and connections within the dataset. Such approaches seek to recognize patterns, clusters, and correlations among observations, enabling researchers to investigate the data without prior assumptions about the results. This attribute renders unsupervised learning especially beneficial in exploratory data analysis, where the objective often involves formulating hypotheses or detecting variations within the data [5-7]. Principal Component Analysis (PCA) is one of the most often used approaches for dimensionality reduction in unsupervised learning. PCA facilitates the transformation of an extensive array of correlated variables into a reduced number of uncorrelated variables named principal components. This modification simplifies the information and improves its interpretability, enabling researchers to see intricate connections and determine significant contributing components. PCA seeks to preserve maximal variance from the original dataset, reducing information loss during dimensionality reduction [8, 9]. Although PCA is extensively used and beneficial for most multivariate issues, it is fundamentally intended for numerical data, creating difficulties for researchers dealing with datasets that include categorical variables. The incapacity of PCA to adequately manage categorical data has been a subject of critique and discourse in academic literature. As datasets increasingly integrate numerical and categorical characteristics, there is a critical demand for approaches capable of addressing this complexity [10, 11].

To alleviate the limitations of PCA, Multiple Correspondence Analysis (MCA) has arisen as a significant expansion of Correspondence Analysis (CA). MCA has been developed to analyze correlations among many categorical dependent variables, proving it especially valuable in disciplines such as social sciences, marketing research, health studies, and climate [12-17]. By offering a framework similar to PCA but tailored for categorical data, MCA enables researchers to investigate intricate interdependencies and illustrate the relationships among categorical variables [14, 18, 19]. Although several conventional data analysis techniques have primarily concentrated on quantitative or qualitative data, most data applications handle datasets combining both measurements. This mixed-data situation poses distinct problems for analysts since conventional methodologies may inadequately address the intricacies inherent in such datasets. Researchers have developed modifications of PCA for mixed data to provide a more thorough analysis that incorporates both numerical and categorical factors.

The present study intends to examine the implementation of PCA for mixed and general datasets, focusing on adapting the approach to equally address numerical and categorical variables. This technique aims to augment the analytical capacities of researchers working with complex datasets, facilitating deeper insights and comprehension of the underlying patterns within the data. This study emphasizes the versatility of PCA and the need to establish resilient approaches to adeptly address the intricacies of modern data difficulties [20, 21]. It is of tremendous significance for researchers in the field of multivariate analysis to focus on implementing the currently used techniques and actively pursue opportunities for theoretical development and methodological enhancements in light of the future. Pursuing innovation is vital for advancing the profession and ensuring that analytical techniques remain relevant and effective in addressing the complexities of modern datasets. By fostering a culture of inquiry and experimentation, we may enhance the effectiveness and significance of multivariate methods, thus improving our understanding of the intricate nature of data.

2. Materials and methods

Principal Component Analysis (PCA) is a method for dimensionality reduction that seeks to maintain as much variation as possible in a dataset while reducing the number of variables by significant amounts. This is achieved by converting the original variables into a collection of ordered and uncorrelated variables called principal components (PCs). The first principal component account for the bulk of the variance in the original dataset [22]. PCA determines

linear combinations of the original variables that optimize variance by orthogonal transformations, converting correlated variables into a diminished set of independent linear combinations (the PCs). PCA is used in mobile robots, handprint recognition, and face recognition [23-25]. Johnson and Wichern provided a comprehensive Principal Component Analysis (PCA) overview. In population PCA, the covariance matrix is decomposed into eigenvalues and eigenvectors, representing the principal components. These linear combinations of original variables capture the maximum variance in the data. Geometrically, PCA can be visualized as a rotation of the coordinate axes to align with the directions of maximum variance. In sample PCA, the population parameters are estimated from the sample data. The key steps involve estimating eigenvalues and eigenvectors and interpreting the principal components in the context of the sample data. Dimensionality reduction is achieved by selecting a subset of principal components that capture a significant portion of the total variance [4, 6].

To execute PCA on a data matrix including p variables and n samples, we start the process by centering the data according to the means of each variable. Subsequently, we compute the covariance matrix and decompose it into eigenvalues and eigenvectors. The eigenvectors represent the principal components, while the eigenvalues signify the amount of variance explained by each component. To determine the optimal number of principal components, we typically employ techniques like the elbow method or set a threshold for the cumulative explained variance. By selecting a subset of principal components that capture a significant portion of the total variance, we can effectively reduce the dimensionality of the data while preserving its essential information.

Let A be $n*n$ matrix and X is a non-zero p vector for which

$$AX = \lambda X, \tag{1}$$

$$(A - \lambda I)X = 0, \tag{2}$$

where λ is known as the eigenvalue of matrix A and X is the eigenvector of matrix A for the corresponding eigenvalue. Eigenvalues represent the amount of variance explained by each principal component. The first PC (PC_1) is given by the linear combination of the variables $X_1, X_2, ..., X_P$,

$$PC_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p, \tag{3}$$

The first PC (PC_1) is calculated to account for the greatest possible variance in the data set [26]. Weights are computed using constraints as follows:

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1 \tag{4}$$

Similarly, the second PC is calculated under the condition of being uncorrelated with (i.e., perpendicular to) the first PC and accounting for the highest variance below.

$$PC_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p, \tag{5}$$

This process continues until P PCs equal to the number of original variables are computed [27].

Multiple correspondence analysis (MCA) investigates the relationships between two or more categorical variables. MCA is similar to PCA but is tailored for categorical data, functioning as an extension of Correspondence Analysis (CA) for multiple variables [18]. CA is used in many domains, such as archeology, ecology, medicine, and health sciences [14, 19, 21]. Given n observations and K categorical variables, with J_k levels for each categorical variable, let X be the $n \times J$ indicator matrix, where $J = \sum_j J_k$. MCA is performed by applying CA to the indicator matrix X, which provides row and column factor scores. These factor scores are standardized such that their variances are equal to their corresponding eigenvalues. We start computing the probability matrix

$$Z = N^{-1}X, \tag{6}$$

where N is the grand total of matrix X. Let $D_c = diag\{c\}$ and $D_r = diag\{r\}$ be matrices, where c and r denote the vectors of the column and row totals of Z. We compute the factor scores by applying the singular value decomposition (SVD), as follows:

$$D_r^{-\frac{1}{2}}\left(Zrc^T\right)D_c^{-\frac{1}{2}} = P\Delta Q^T, \tag{7}$$

where Δ is the diagonal matrix of the singular values and $\Lambda = \Delta^2$ is the eigenvalues matrix. Then, we obtain the columns factor scores, which are denoted by G, and the rows factor scores, which are denoted by F as follows [18, 14, 19]:

$$G = D_c^{-\frac{1}{2}}Q\Delta, \tag{8}$$

$$\text{and } F = D_r^{-\frac{1}{2}}P\Delta \tag{9}$$

The PCAmix method is dedicated to analyzing mixed data, in which numerical and categorical variables describe the attributes. This was proposed by de Leeuw and Van Rijckevorsel [28] and extended by Kiers [29]. PCAmix is a combination of PCA and MCA, where PCA handles the numerical variables and MCA handles the categorical variables. We implement the PCA on mixed data following the approach proposed by Chavent [14, 19, 30]. The dataset contains n observations which are described by p_1 numerical variables and p_2 categorical variables. The dataset is represented by an $n \times p_1$ numerical data matrix X_1 and the $n \times p_2$ categorical data matrix X_2, with d denoting the total number of levels of the p_2 categorical variables.

Let G be an indicator matrix with $n \times d$ dimensions containing binary coding from each level of the categorical variables. $Y = \left(Y_1|Y_2\right)$ is a numerical matrix with dimensions $n \times (p_1 + d)$ where Y_1 is the standardized matrix constructed by the centered and normalized columns X_1, and Y_2 denotes the centered indicator matrix X_2. Now, we build a diagonal matrix N of the weights of the rows of Y; the n rows are weighted $\frac{1}{n}$, such that $N = \frac{1}{n}I_n$. Suppose $D = diag\left(1, \cdots, 1, \frac{n}{n_1}, \cdots, \frac{n}{n_s}\right)$ is the diagonal matrix of the weights of the columns of Y, and $s = 1, ..., n$ represents the number of observations that belong to the sth level. Then, the eigenvalue of Y is obtained using the generalized singular value decomposition (GSVD) as follows:

$$Y = U\Lambda V^T, \tag{10}$$

where $\Lambda = diag\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \cdots, \sqrt{\lambda_r}\right)$ is the $r \times r$ diagonal matrix, such that $\lambda_1, \lambda_2, \cdots, \lambda_r$ are the eigenvalues of Y, and r denotes the rank of Y. U is a matrix with $n \times r$ dimensions, where the first r eigenvectors of ZDZ^tN, such that $U^tNU = I_r$. V is the $p \times r$ matrix of the first r eigenvectors of Z^tNZD and $V^tDV = I_r$. Therefore, the PC of the PCA mix can be computed as:

$$Y^{mix} = YDV, \tag{11}$$

with the dimensions of $n \times r$. $R = U\Lambda$ indicates scores for rows that represent the PC scores. The scores of columns $C = DV\Lambda$ and the standard PCA are $C = V\Lambda$.

3. Enhancing the stability and efficacy of MCA and PCA

Regarding the innovations of the PCA, MCA, and PCAmix techniques, some approaches can be employed, such as:

Customize PCAmix for specialized fields, for instance healthcare or environmental sciences, where tailored pre-processing and post-processing may uncover new insights.

Enhance approaches for addressing absent categorical and numerical data inside PCA or MCA frameworks, increasing their usefulness. This is done by dividing the data into two and studying the two that reflect the absence of some categorical variables.

In this paper, we try to adapt PCAmix to specific domains, such as environmental sciences or healthcare, where customized pre- and post-processing might provide novel insights. We employ a dataset of traumatic brain injury (TBI), which is a complex disorder that is traditionally stratified based on clinical signs and symptoms. The Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) Pilot multicenter study enrolled 586 acute TBI patients and collected diverse common data elements (TBI-CDEs), including imaging, genetics, and clinical outcomes, across the study population. The dataset consisted of 25 variables, including seven quantitative and 18 qualitative variables, as shown in Table 1 and Table 2.

Table 1. Numerical variables descriptions

Name	Description	Values
BIF1	Marshall CT score	1-6
BIF2	Rotterdam CT score	1-6
FCO1	GOSE score (3 months)	1-8
FCO2	GOSE score (6 months)	18
FCO4	WAIS PSO Composite Score (6 monthes)	50-150
FCO5	CVLT Short Delay Cued Reall Standard Score (6 monthes)	-4.0-2.5
FCO6	CVLT Long Delay Cued Reall Standard Score (6 monthes)	-3.5-2.5

Table 2. Categorical variables descriptions

Name	Description	Values
BIF3	CT brain pathology	no, yes
BIF4	CT skull fracture	no, yes
BIF5	CT Skull base fracture	no, yes
BIF6	CT facial fracture	no, yes
BIF7	CT epidural hematoma	no, yes
BIF8	CT subdural hematoma	no, yes
BIF9	CT subarachnoid hemorrhage	no, yes
BIF10	CT contusion	no, yes
BIF11	CT midline shift	no, yes
BIF12	Cisternal compression	no, yes
BIF13	MRI brain pathology	no, yes, indeterminate
FCO3	PTSD DSM-IV diagnosis (6months)	no, yes
GM1	COMT SNP genotype	Met/Met, Met/Val, Val/Val
GM2	DRD2 SNP genotype	C/C, C/T, T/T
GM3	PARP1 SNP genotype	A/A, A/T, T/T
GM4	ANKK1 SNP Gly318Arg	A/A, A/G, G/G
GM5	ANKK1 SNP Gly442Arg	C/C, C/G, G/G
GM6	ANKK1 SNP Glu713Lys	C/C, C/T, T/T

The following is a brief description of the variables:

Brain imaging findings (BIF):

Marshall CT score: Predictive scoring system for traumatic brain injury outcomes based on computed tomography scans (CT scan) abnormalities.

Rotterdam CT score: Alternative scoring system for traumatic brain injury outcomes derived from CT scans.

CT Brain Pathology: Abnormal findings in brain tissue on CT scans.

CT Skull Fracture: finding of a fracture in the skull bone on a CT scan.

CT Skull Base Fracture: finding of fracture in the bones of the base of the skull on a CT scan.

CT Facial Fracture: finding of a fracture in facial bone on a CT scan.

CT Epidural Hematoma: finding of bleeding above the epidural layer of the meninges on a CT scan.

CT Subdural Hematoma: finding of bleeding below the epidural layer of meninges in computed tomography scan.

CT Subarachnoid Hemorrhage: finding of bleeding below the subarachnoid layer of the meninges on a CT scan.

CT Contusion: finding of scattered bleeding over the brain surface on a CT scan.

CT Midline Shift: finding of the shift of one brain hemisphere to the other side across the midline on a CT scan.

Cisternal Compression: finding of compression over the cisternal part of the brain.

MRI Brain Pathology: abnormal findings in brain tissue on an magnetic resonance imaging (MRI).

Functional and cognitive outcomes (FCO):

The Extended Glasgow Outcome Scale (GOSE): Eight-category functional outcome measure at 3 and 6 months post-injury. PTSD DSM-IV Diagnosis: Diagnosis of post-traumatic stress disorder (PTSD), which occurs when a person experiences, witnesses, or is confronted with an event that involves actual or threatened death or serious injury, or a threat to the physical integrity of the self or others. The diagnosis is based on the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders.

WAIS PSO Composite Score: obtained using the The Wechsler Adult Intelligence Scale (WAIS), which is an IQ test designed to measure intelligence and cognitive ability.

CVLT Short Delay Cued Recall Standard Score and CVLT Long Delay Cued Recall Standard Score: obtained using the California Verbal Learning Test, which is a common assessment instruments used by clinicians to measure a verbal learning and memory. Short refers to short-term memory, and long refers to long-term memory. Free or cued recall is measured. In the latter, the patient is provided material to remember and asked to complete the test.

Genetic markers (GM):

COMT SNP Genotype: type of COMT gene important for cognitive function.

DRD2 SNP Genotype: type of DRD2 gene which is important for verbal learning.

PARP1 SNP Genotype: type of PARP1 gene important for response to stress.

ANKK1 SNP Gly318Arg: type of COMT gene important for cognitive function.

ANKK1 SNP Gly442Arg: type of COMT gene important for cognitive function.

ANKK1 SNP Glu713Lys: type of COMT gene important for cognitive function.

4. Applying PCA to TBI data: A step-by-step guide

4.1 *Understanding the data*

Before applying PCA to TBI data, explore these characteristics further to see how they might be used in PCA for a better understanding of TBI. Standard variables in TBI research include:

Categorical Variables.

Continuous Variables.

4.2 *Steps involved in PCA*

PCA can be applied to TBI data variables to:

Determine the latent factors: Minimize the dimensionality of the data by discerning fundamental elements that elucidate the variability within the data. Uncover hidden patterns and associations among variables that may not be

evident from basic correlation analysis.

Anticipate results: Employ PCA to ascertain the main determinants of long-term outcomes, including cognitive impairment and functional disability. Create predictive algorithms to identify patients at elevated risk of adverse outcomes.

Comprehend the neurobiological mechanisms behind traumatic brain injury (TBI): Examine the correlation between brain morphology and functionality, as assessed by neuroimaging and cognitive and behavioral results.

To understand the PCA results, it is important to consider the clinical and neurological importance of the main components. By correctly applying PCA to TBI data, one can learn a lot about how injuries happen and how people heal, leading to more effective medications and interventions.

5. Results and discussion

The PCAmix tool is utilized to assess the efficacy of PCA for mixed data types, adhering to the methodology outlined by Chavent [30]. This analysis is conducted using traumatic brain injury data, which comprises both numerical and categorical variables. The data was divided into two matrices: the first matrix, designated as dataA, contains seven columns for numerical data, while the second matrix, referred to as dataB, comprises 18 columns for categorical data. The implementation is conducted across three distinct scenarios, as illustrated in the following.

5.1 *Case I: Global PCAmix (TBI data)*

The global PCAmix is implemented on TBI data by combining all variables, Brain Imaging Findings (BIF), Functional and Cognitive Outcomes (FCO), and Genetic Markers (GM), aiming to capture the overall variance structure across these domains. The eigenvalues and proportions of each PC are shown in Table 3, which provides an overview of the global PCAmix implementation outcomes. The eigenvalue indicates the quantity of variation captured by a principal component, with higher values indicating more explained variability. The proportion reflects the percentage of total variation elucidated by each principal component, while the cumulative proportion illustrates the overall variance accounted for when more principal components are included. The first 11 principal components collectively explain over 75% of the total variance, the contributions of all variables to each one of those PCs are presented in Table 4. Figure 1(a) illustrates the primary variances in each fundamental component. The first PC represents 21.48% of the variation. The second principal component accounts for 8.76%, whereas the first two components together explain 30.24% of the variation. The 11th principal component accounts for 3.15%, whereas the first eleven components together explain 77.17% of the total variation. Figure 1(b) shows the factor coordinates, absolute contributions, and squared cosines of the qualitative variables, along with the correlations between the first two principal components and the levels of the qualitative variables. The correlations with the first principal component are represented on the horizontal axis, while those with the second principal component are displayed on the vertical axis. In the positive aspect of , the “yes” levels of “BIF” variables are predominant, indicating the significance of particular brain imaging results in influencing , whereas the “no” levels and certain SNP genotypes (“GM3 = A/A”, “GM5 = C/G”) are inversely associated with . is significantly affected by genetic marker levels, with positive influences from “GM4 = G/G”, “GM5 = G/G”, and “GM2 = C/C”, and negative influences from “GM4 = A/G”, “GM2 = C/T”, and “GM6 = C/T”.

Figure 1(c) presents the results for the quantitative variables. Brain imaging findings, such as “BIF1” and “BIF2,” exhibit strong positive coordinates on , indicating their dominance. In contrast, the functional and cognitive outcome variables, namely “FCO5”, “FCO6”, “FCO2”, and “FCO1”, exhibit negative correlations, indicating an inverse relationship with the “BIF” variables. In the case of , “FCO4” exhibits the highest positive loading, indicating its distinct variance relative to the other variables. Figure 1(d) shows the results for all the variables, the squared contribution correlation for the quantitative variable and the contribution correlation ratio of qualitative variables. It is apparent that is strongly associated with brain imaging findings, particularly “BIF8”, “BIF9”, “BIF10”, “BIF3”, and “BIF1”, while functional and cognitive outcomes (“FCO5”, “FCO6”) and genetic markers (“GM5”) have moderate contributions. On the other hand, is primarily linked to genetic markers, notably “GM4”, “GM2” , and “GM5”, with minimal contributions from functional and cognitive outcomes (“FCO3”, “FCO4”) and some “BIF” variables.

In summary, the first principal component (PC_1) primarily reflects brain injury severity, as indicated by

neuroimaging findings such as CT scans and MRIs. Higher PC_1 scores correspond to more severe brain damage, which is associated with poorer cognitive and functional outcomes. In contrast, the second principal component (PC_2) appears to be influenced by genetic factors, as evidenced by the strong contribution of genetic markers. This suggests that genetic variations may play a significant role in modulating the impact of brain injury on cognitive and functional outcomes. Moderate contributions from other variables, such as demographic and clinical factors, indicate that a combination of genetic and environmental factors may influence the overall outcome.

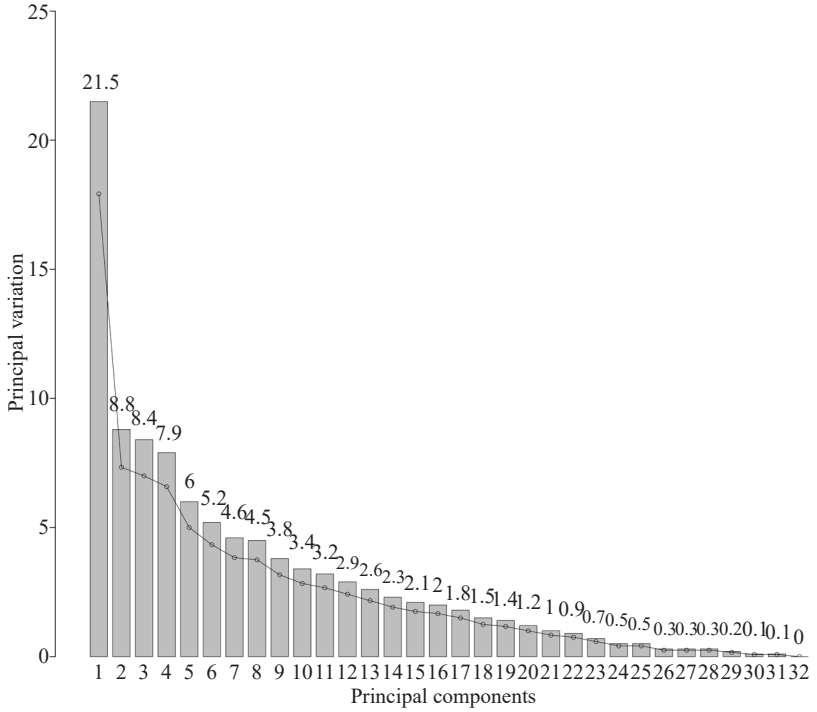
Table 3. Case I: the contribution of each principle component (PC)

	Eigenvalue	Proportion	Cumulative
comp 1	6.873	21.479	21.479
comp 2	2.804	8.7621	30.241
comp 3	2.696	8.4235	38.665
comp 4	2.518	7.8699	46.535
comp 5	1.928	6.0239	52.559
comp 6	1.662	5.1950	57.754
comp 7	1.475	4.610	62.363
comp 8	1.426	4.457	66.820
comp 9	1.228	3.837	70.657
comp 10	1.077	3.366	74.023
comp 11	1.009	3.152	77.174
comp 12	0.939	2.936	80.110
comp 13	0.821	2.564	82.674
comp 14	0.744	2.324	84.998
comp 15	0.686	2.144	87.142
comp 16	0.653	2.040	89.182
comp 17	0.576	1.800	90.983
comp 18	0.486	1.520	92.503
comp 19	0.452	1.413	93.915
comp 20	0.372	1.162	95.077
comp 21	0.318	0.994	96.071
comp 22	0.294	0.919	96.991
comp 23	0.238	0.744	97.735
comp 24	0.172	0.537	98.273
comp 25	0.151	0.471	98.743
comp 26	0.101	0.315	99.059
comp 27	0.090	0.281	99.339
comp 28	0.084	0.261	99.600
comp 29	0.048	0.151	99.752
comp 30	0.038	0.120	99.871
comp 31	0.027	0.085	99.957
comp 32	0.014	0.043	100.00

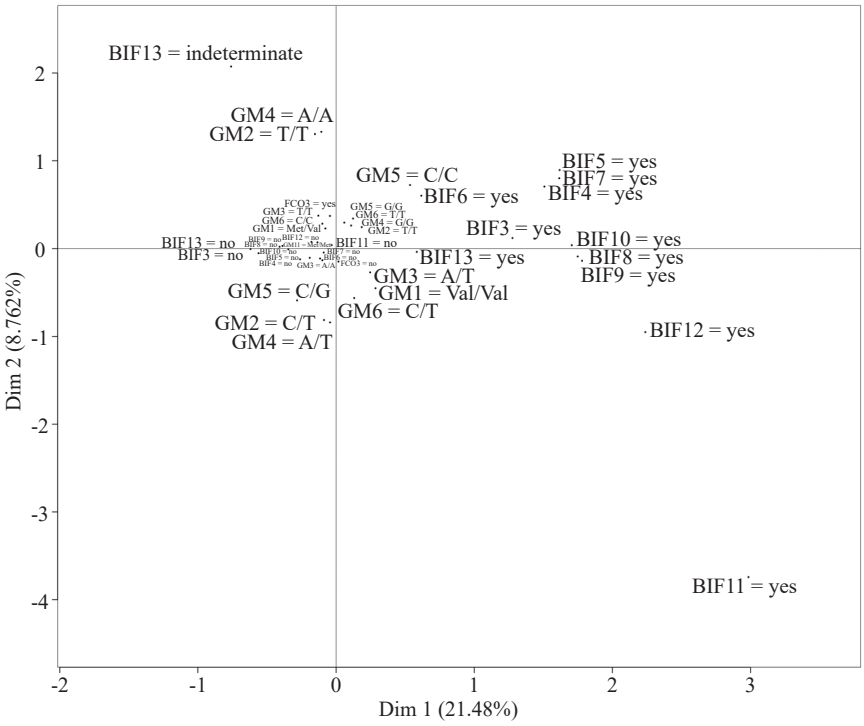
Table 4. Case I: The contribution levels for all variables to each PC

	BIF1	BIF2	FCO1	FCO2	FCO4	FCO5	FCO6	BIF3	BIF4
dim 1	0.6174	0.4916	0.3566	0.1602	0.0528	0.2437	0.2960	0.7162	0.3929
dim 2	0.0166	0.0771	0.0044	0.0067	0.1463	0.0124	0.0173	0.0064	0.0863
dim 3	0.0339	0.0349	0	0	0.0038	0.0349	0.0478	0.00002	0.0081
dim 4	0.0214	0.0591	0.1757	0.1362	0.2248	0.2286	0.2407	0.0089	0.2685
dim 5	0.0372	0.0026	0.0034	0	0.0325	0.0916	0.1016	0.0059	0.0083
dim 6	0.1070	0.0202	0.1154	0.1703	0.0880	0.0224	0.0047	0.0037	0.0687
dim 7	0.0070	0.0290	0.0220	0.1131	0.0477	0.0135	0.0240	0.0241	0.0029
dim 8	0.0241	0.0057	0.0616	0.0890	0.0129	0.0253	0.0244	0.0020	0.0034
dim 9	0.0061	0.0571	0.0062	0.0140	0.0525	0.1129	0.0981	0.0079	0.0077
dim 10	0.0026	0.0125	0.0012	0.0289	0.0507	0.0010	0.0059	0.0009	0.0009
dim 11	0.0039	0.0233	0.0006	0.0356	0.0250	0.0648	0.0265	0.0048	0.0052
	BIF5	BIF6	BIF7	BIF8	BIF9	BIF10	BIF11	BIF12	BIF13
dim 1	0.3080	0.0714	0.1450	0.7151	0.6911	0.5895	0.0948	0.3986	0.3639
dim 2	0.0326	0.0681	0.0357	0.0019	0.0042	0.0003	0.1490	0.0721	0.0460
dim 3	0.0203	0	0.0023	0.0058	0.0140	0.0010	0.0499	0.0508	0.0481
dim 4	0.1398	0.0213	0.2948	0.0039	0	0.0657	0.0213	0.0094	0.0577
dim 5	0.0016	0.0521	0.0309	0.0197	0.0212	0.0397	0.0253	0.0852	0.0480
dim 6	0.1404	0.1927	0.0299	0.0192	0.0184	0.0103	0.0683	0.0825	0.0253
dim 7	0.0333	0.1311	0.0252	0.0001	0.0008	0.0016	0.0118	0.0072	0.0654
dim 8	0.0209	0.0334	0.0449	0.0267	0.0298	0.0005	0.0129	0.0139	0.0034
dim 9	0.0106	0.1132	0.0095	0.0203	0.0276	0.0068	0.2099	0.0175	0.0034
dim 10	0.0053	0.0482	0.0048	0.0604	0.0413	0	0.0180	0.0193	0.1943
dim 11	0.0149	0.0263	0.1276	0.0055	0.0015	0	0.0145	0.0013	0.2722
	FCO3	GM1	GM2	GM3	GM4	GM5	GM6		
dim 1	0.0008	0.0297	0.0207	0.0283	0.0075	0.0702	0.0115		
dim 2	0.0549	0.0773	0.6483	0.0780	0.6842	0.2550	0.1612		
dim 3	0.0091	0.0233	0.8551	0.0435	0.8645	0.2410	0.2946		
dim 4	0.0951	0.0252	0.0205	0.0121	0.0340	0.1847	0.1690		
dim 5	0.0119	0.2857	0.0365	0.2610	0.0245	0.3201	0.3811		
dim 6	0.0510	0.0317	0.0522	0.1712	0.0961	0.0143	0.0584		
dim 7	0.3549	0.1322	0.0244	0.2375	0.0111	0.0832	0.0721		
dim 8	0.0048	0.4499	0.0017	0.4557	0.0012	0.0121	0.0218		
dim 9	0.0019	0.0507	0.0575	0.1530	0.0602	0.1126	0.0107		
dim 10	0.2155	0.0356	0.0149	0.0100	0.0021	0.2333	0.0606		
dim 11	0.0022	0.1312	0.0294	0.0125	0.0174	0.0520	0.1105		

(a) Scree plot



(b) Levels



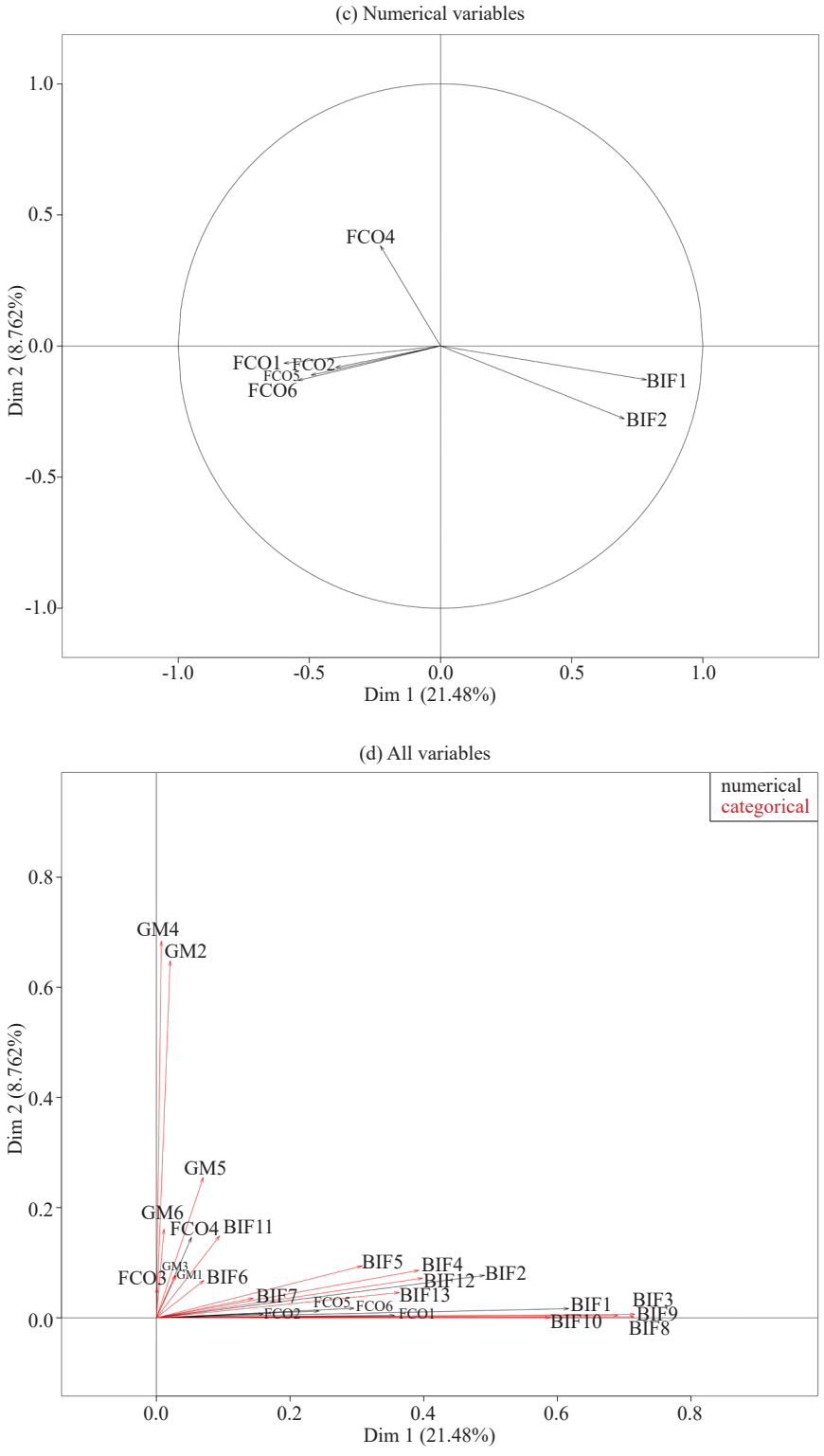


Figure 1. Case I: (a) Variations in principal components; (b) Results for the levels of the qualitative variables; (c) Results for the quantitative variables; (d) Results for the square loadings

5.2 *Case II: Assessing the intricate interconnections among TBI biomarkers through PCA*

By identifying "Brain Injury Fingerprints" (BIFs) and "Functional Connectivity Outcomes" (FCOs), PCAmix may adequately investigate the complex relationship between structural brain damage and functional recovery in traumatic brain injury (TBI). This method could identify latent patterns in the data, identifying hidden correlations between neuroimaging signs and clinical outcomes. The eigenvalues and proportions of each principal component (PC) are shown in Table 5, which provides an overview of the local PCAmix implementation outcomes. The first 7 PCs explain more than 75% of the total variance, the contributions of all variables to each one of those PCs are presented in Table 6. Figure 2(a) and 2(b) display the PCs along with their associated variances. PC_1 accounts for 33.7 % of the total variance and is predominantly influenced by structural brain injury. Categories such as "BIF3 = yes" and "BIF8 = yes," indicative of severe brain damage, exhibit a strong correlation with PC_1. Conversely, PC_2, which accounts for 11.9% of the variance, is more significantly influenced by functional recovery. Variables such as "BIF4 = yes" and "BIF7 = yes," linked to improved functional outcomes, exhibit a positive correlation with PC_2.

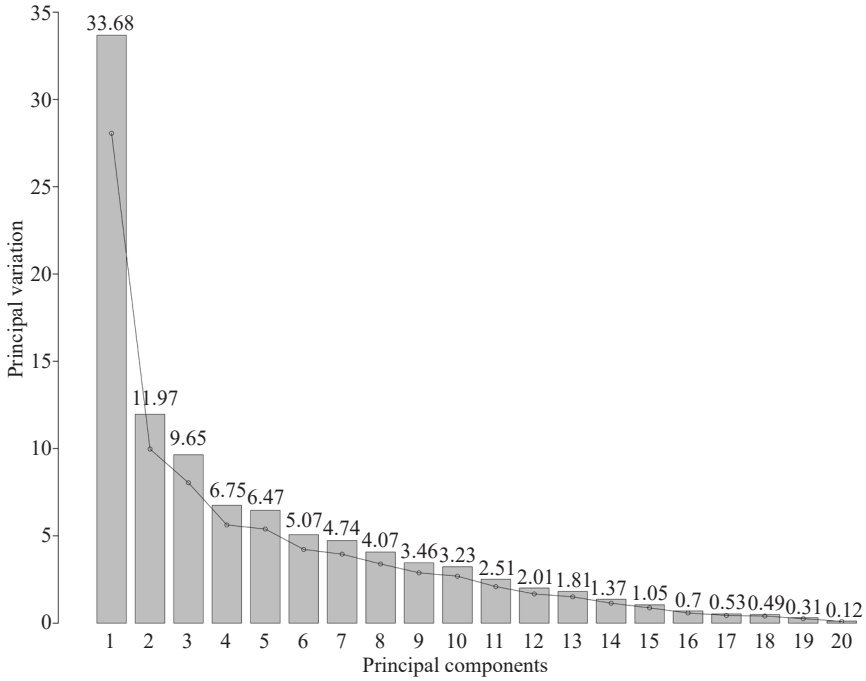
Table 5. Case II: the contribution of each principle component (PC)

	Eigenvalue	Proportion	Cumulative
comp 1	6.736	33.678	33.678
comp 2	2.394	11.970	45.648
comp 3	1.929	9.646	55.294
comp 4	1.351	6.754	62.048
comp 5	1.294	6.472	68.520
comp 6	1.014	5.070	73.590
comp 7	0.948	4.741	78.332
comp 8	0.815	4.075	82.406
comp 9	0.693	3.465	85.871
comp 10	0.646	3.229	89.100
comp 11	0.501	2.507	91.607
comp 12	0.402	2.012	93.619
comp 13	0.362	1.808	95.426
comp 14	0.274	1.370	96.796
comp 15	0.210	1.051	97.847
comp 16	0.141	0.703	98.550
comp 17	0.106	0.532	99.082
comp 18	0.098	0.491	99.573
comp 19	0.061	0.306	99.879
comp 20	0.024	0.121	100.000

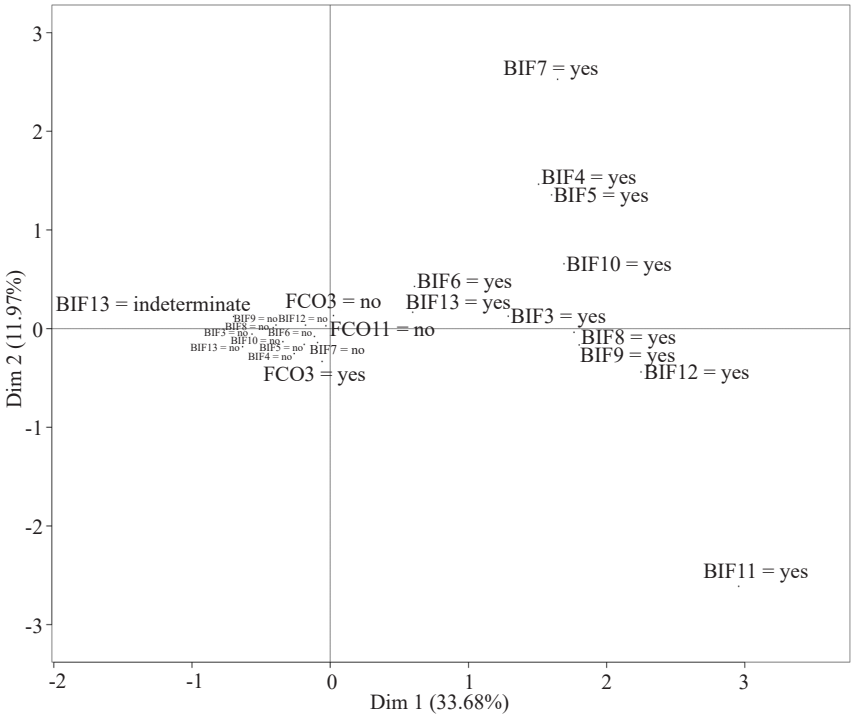
Table 6. Case II: the contribution levels for all variables

	BIF1	BIF2	FCO1	FCO2	FCO4	FCO5	FCO6	BIF3	BIF4	BIF5
dim 1	0.6330	0.4905	0.3455	0.1540	0.0493	0.2421	0.2881	0.7290	0.3929	0.3018
dim 2	0.0119	0.1218	0.1017	0.0771	0.3818	0.2305	0.2309	0.0070	0.3703	0.2162
dim 3	0.1723	0.1089	0.1233	0.1436	0.0042	0.0926	0.1336	0.0020	0.0764	0.1428
dim 4	0.0216	0.0004	0.1361	0.3621	0.0209	0.2360	0.2128	0.0005	0.0001	0.0142
dim 5	0.0241	0.0469	0.0055	0.0122	0.2365	0.0105	0.0035	0.0251	0.0140	0.0557
dim 6	0.0002	0.0001	0.0004	0.0687	0.0019	0.0005	0.0005	0.0001	0.0003	0.0005
dim 7	0.0390	0.0560	0.0222	0.0020	0.0169	0.0739	0.0425	0.0045	0.0001	0.0796
	BIF6	BIF7	BIF8	BIF9	BIF10	BIF11	BIF12	BIF13	FCO3	
dim 1	0.0696	0.1503	0.7274	0.7071	0.5793	0.0929	0.4022	0.3792	0.0014	
dim 2	0.0342	0.3548	0.0003	0.0060	0.0873	0.0726	0.0154	0.0304	0.0440	
dim 3	0.1407	0.0043	0.0061	0.0067	0.0010	0.3420	0.2428	0.0037	0.1822	
dim 4	0.0269	0.0007	0.0495	0.0466	0.0030	0.0112	0.0527	0.0579	0.0976	
dim 5	0.4114	0.0650	0.0149	0.0183	0.0176	0.0325	0.0081	0.0001	0.2924	
dim 6	0.0002	0.0041	0.0054	0.0057	0.0003	0.0051	0.0049	0.9028	0.0124	
dim 7	0.0474	0.3171	0.0713	0.0391	0.0028	0.0033	0.0606	0.0345	0.0354	

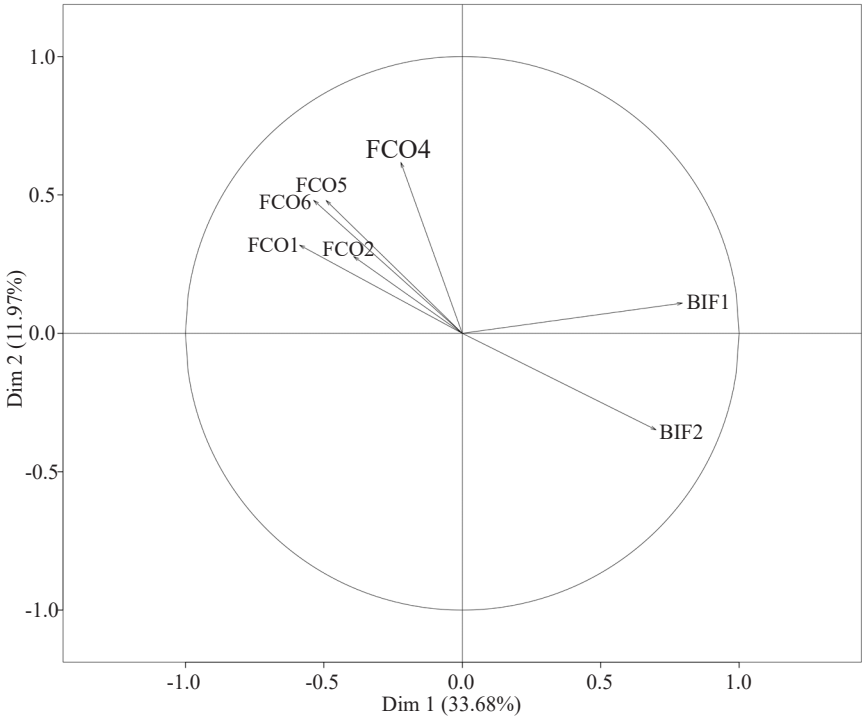
(a) Scree plot



(b) Levels



(c) Numerical variables



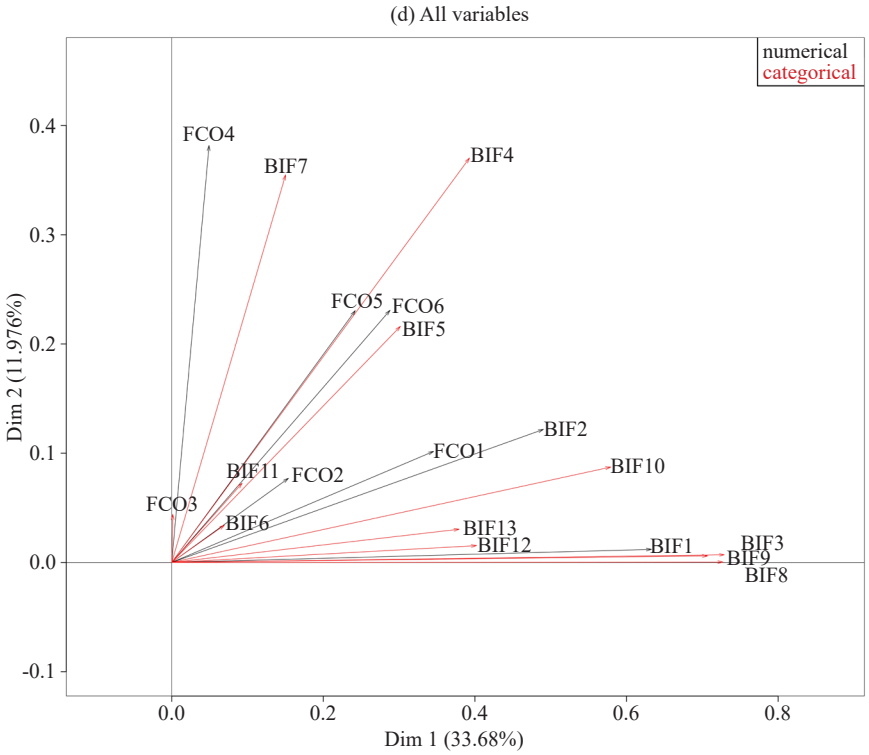


Figure 2. Case II: (a) Variations in principal components; (b) Results for the levels of the qualitative variables; (c) Results for the quantitative variables; (d) Results for the square loadings

Figure 2(c) elucidates the relationship between structural and functional components. BIF1 and BIF2, indicators of significant structural damage, exhibit a strong correlation with PC_1. In contrast, "FCO1" through "FCO6," which denote different dimensions of functional recovery, are strongly associated with PC_2. Figure 2(d) clearly illustrates the distinction between structural and functional components. PC_1 predominantly reflects structural injury, with "BIF1," "BIF2," "BIF3," "BIF8," and "BIF9" exhibiting clustering. PC_2 is more closely associated with functional recovery, with "FCO1" to "FCO6" and specific structural measures such as "BIF10" and "BIF11" forming a distinct cluster.

To sum up this case, PC_1 primarily captures the dimension of structural brain injury, as evidenced by the strong clustering of "BIF1" and "BIF9." Higher scores on PC_1 indicate more severe brain injury. Conversely, PC_2 reflects functional recovery outcomes, with "FCO2" and "FCO4" being key contributors. Higher scores on PC_2 are associated with better recovery. The inverse relationship between PC_1 and PC_2 suggests that severe structural damage is linked to poorer functional outcomes. However, the presence of independent variance in PC_2 indicates that other factors, such as genetic or environmental factors, may also influence recovery.

5.3 *Case III: Unveiling the Genetic Underpinnings of TBI Recovery with PCAmix*

In Case III, PCAmix focuses on the interplay between genetic markers (GMs) and functional outcomes (FCOs), isolating these factors to explore their influence on recovery without the confounding effects of structural brain injuries. By analyzing the correlation between genetic variants and functional outcomes, this approach can potentially identify specific genetic predispositions that may influence recovery trajectories. This could lead to the development of personalized treatment strategies based on an individual's genetic profile. Table 7 presents a summary of the PCA findings, including the eigenvalues and the percentage associated with each principal component. Over 75% of the overall variation is accounted for by the first 8 principal components; the contributions of all variables to each one of those PCs are presented in Table 8. Figure 3(a) presents the PCs and their corresponding variances. The first principal component, accounting for 16.96% of the variance, is primarily driven by functional recovery outcomes. Variables like "FCO5" and "FCO6" strongly align with PC_1, indicating their significant contribution to recovery. In contrast, the

second principal component, explaining 14.70% of the variance, is more influenced by genetic factors. Genetic markers such as "GM2 = T/T" and "GM4 = A/A" are strongly correlated with PC_2, suggesting their significant role in shaping individual recovery trajectories, as shown in Figure 3(b).

Figure 3(c) further illuminates the relationship between functional outcomes and genetic markers. "FCO1" to "FCO6" are closely linked to PC_1, emphasizing their contribution to functional recovery. Genetic markers, while influencing PC_2, have a more limited impact on PC_1, suggesting that functional outcomes are primarily driven by non-genetic factors. Figure 3(d) clearly visualizes the separation between genetic and functional factors. PC_1 is dominated by functional outcomes, with "FCO5" and "FCO6" being the primary drivers. PC_2, on the other hand, is primarily influenced by genetic markers, with "GM2" and "GM4" contributing significantly to the variance.

In summary, PC_1 primarily reflects functional recovery outcomes. "FCO5" and "FCO6", key indicators of functional recovery, strongly correlate with PC_1, suggesting that this component captures the shared variance among these functional measures. The genetic influence on PC_1 is secondary, indicating that genetic factors may play a less significant role in driving functional recovery. In contrast, PC_2 is dominated by genetic variation. Genetic markers like "GM2 = T/T" and "GM4 = A/A" have a strong influence on PC_2, suggesting that these genetic variants may play a crucial role in shaping individual differences in recovery trajectories. The minimal overlap between PC_1 and PC_2 suggests that genetic and functional factors may operate through distinct, yet interconnected, pathways in influencing TBI outcomes.

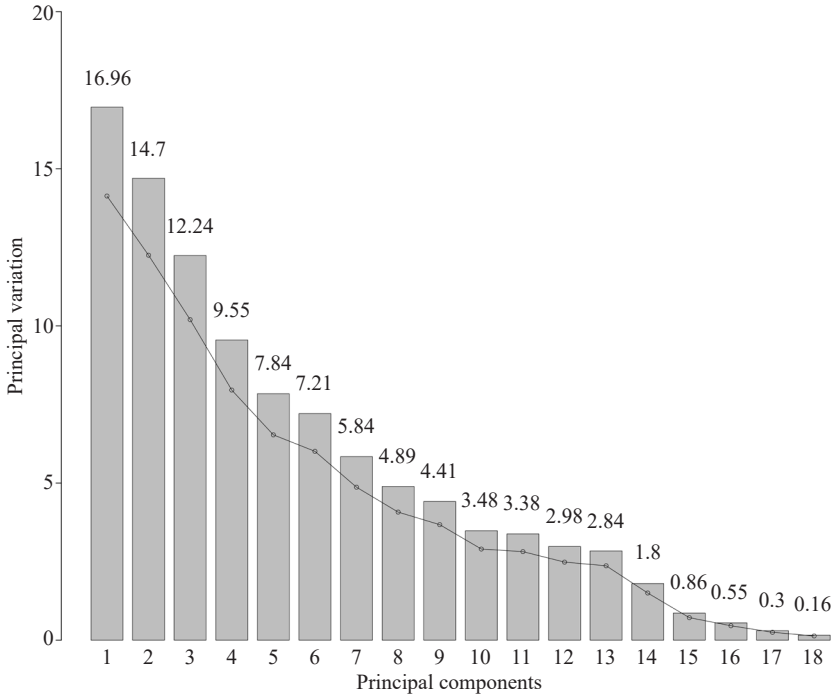
Table 7. Case III: the contribution of each principle component (PC)

	Eigenvalue	Proportion	Cumulative
comp 1	3.053	16.962	16.962
comp 2	2.646	14.703	31.665
comp 3	2.202	12.235	43.900
comp 4	1.718	9.546	53.446
comp 5	1.411	7.840	61.286
comp 6	1.298	7.214	68.500
comp 7	1.052	5.844	74.344
comp 8	0.879	4.885	79.229
comp 9	0.794	4.413	83.642
comp 10	0.626	3.479	87.122
comp 11	0.609	3.381	90.502
comp 12	0.536	2.979	93.481
comp 13	0.511	2.837	96.318
comp 14	0.324	1.797	98.116
comp 15	0.156	0.864	98.980
comp 16	0.099	0.552	99.532
comp 17	0.055	0.304	99.836
comp 18	0.029	0.164	100.000

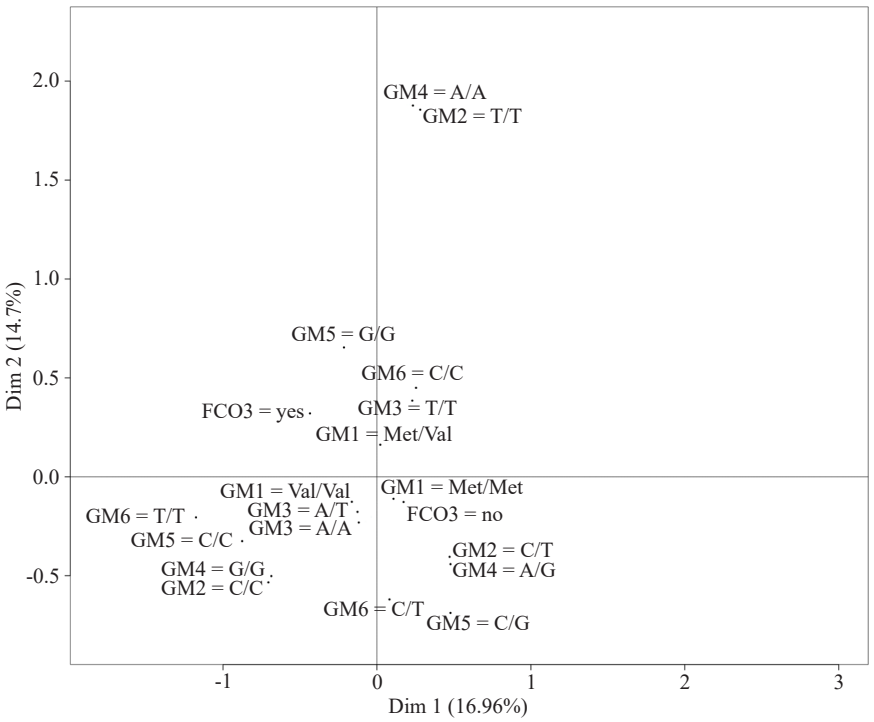
Table 8. Case III: the contribution levels for all variables

	FCO1	FCO2	FCO4	FCO5	FCO6	FCO3
dim 1	0.4430	0.2708	0.1422	0.5030	0.5800	0.0752
dim 2	0.0242	0.0212	0.0542	0.0042	0.0035	0.0408
dim 3	0.2034	0.1341	0.2382	0.0226	0.0287	0.0000
dim 4	0.0058	0.0014	0.0059	0.1260	0.1227	0.0018
dim 5	0.0065	0.0907	0.0006	0.0316	0.0332	0.3676
dim 6	0.0364	0.1230	0.0562	0.1184	0.1064	0.1468
dim 7	0.0054	0.0127	0.0103	0.0472	0.0117	0.0002
dim 8	0.0013	0.0002	0.0045	0.0005	0.0003	0.0371
	GM1	GM2	GM3	GM4	GM5	GM6
dim 1	0.0108	0.2953	0.0280	0.2829	0.1966	0.2253
dim 2	0.0192	0.8643	0.0794	0.8814	0.4125	0.2417
dim 3	0.1449	0.4924	0.0793	0.5856	0.1655	0.1077
dim 4	0.4085	0.0165	0.2893	0.0087	0.3224	0.4092
dim 5	0.2796	0.0140	0.4952	0.0160	0.0067	0.0694
dim 6	0.3161	0.0139	0.3302	0.0058	0.0379	0.0074
dim 7	0.1026	0.0113	0.1472	0.0109	0.4185	0.2739
dim 8	0.2056	0.1191	0.0064	0.0537	0.0471	0.4035

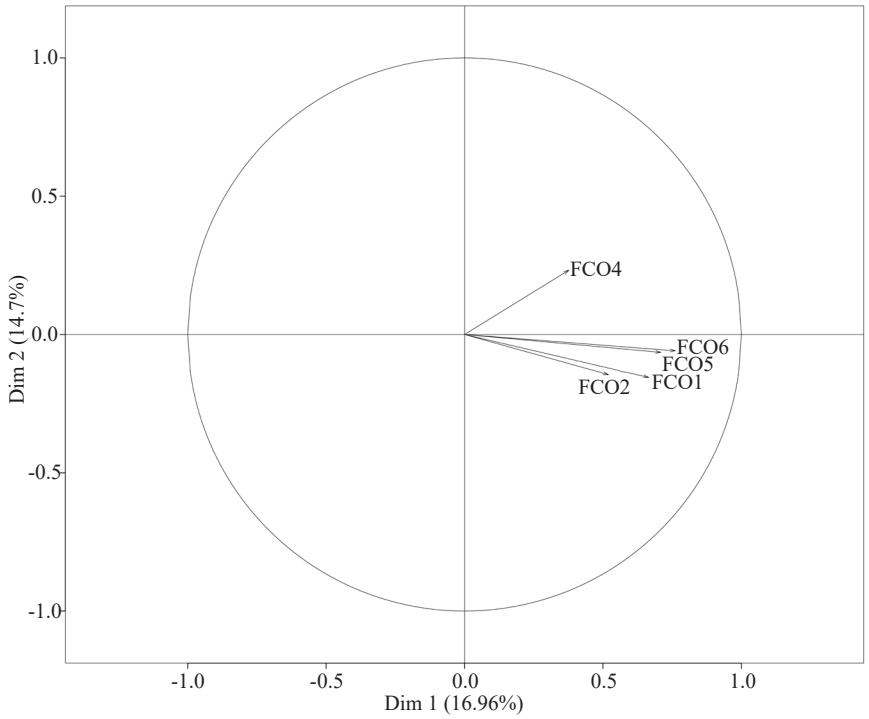
(a) Scree plot



(b) Levels



(c) Numerical variables



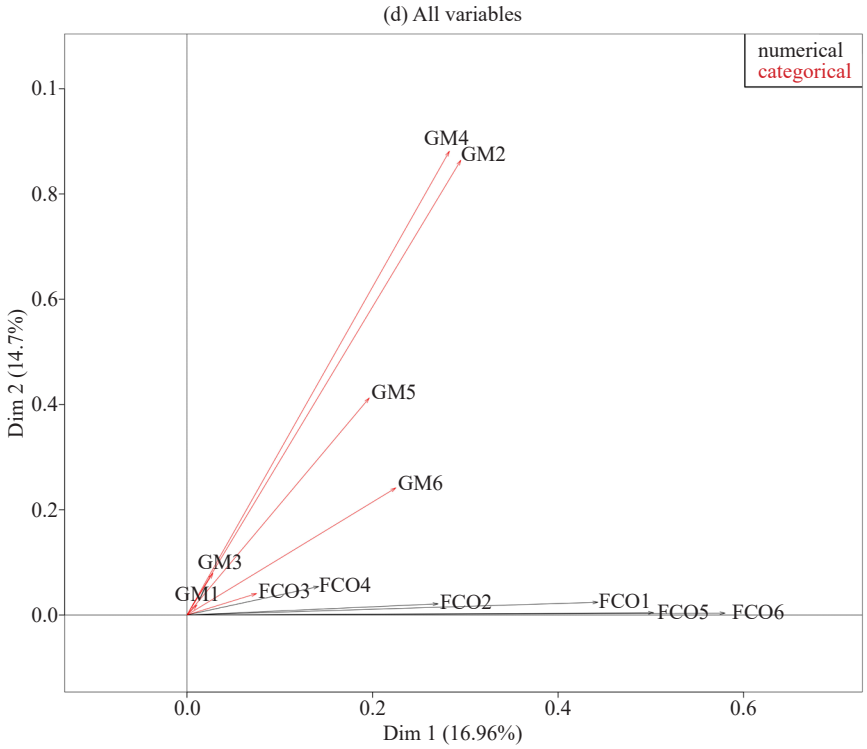


Figure 3. Case III: (a) Variations in principal components.; (b) Results for the levels of the qualitative variables; (c) Results for the quantitative variables; (d) Results for the square loadings

5.4 *Summary of global and local PCAmix results*

Case I: The global PCAmix combines variables from Brain Imaging Findings (BIF), Functional and Cognitive Outcomes (FCO), and Genetic Markers (GM), aiming to capture the overall variance structure across these domains. Brain imaging findings (BIF) were the primary contributors to variance, emphasizing their critical role in predicting functional and cognitive outcomes. Functional outcomes (FCO) were inversely related to imaging findings, with severe injuries predicting poorer recovery. Genetic markers (GM) showed an independent influence, suggesting that genetic variations uniquely contribute to resilience and recovery, separate from imaging findings. However, this approach is less effective at uncovering specific associations or interactions between variables in distinct subdomains. For instance, the relationships between genetic markers and recovery outcomes are diluted in the global model due to the dominance of imaging-related variance. The global PCAmix offers a holistic but generalized view, prioritizing the largest sources of variance while overshadowing weaker, potentially meaningful cross-domain relationships.

Case II: Assessing the Intricate Interconnections Among TBI Biomarkers Through PCA allows a more focused exploration of the relationship between brain imaging findings (BIF) and functional and cognitive outcomes (FCO). This results in more apparent clustering and differentiation between these variables, which was less apparent in the global analysis. The second case improves specificity by highlighting how specific brain imaging findings, such as midline shift or cisternal compression, relate to recovery metrics like GOSE scores or PTSD. Furthermore, it reveals some additional insights, like the impact of "BIF" variables on both short-term (e.g., GOSE at 3 months) and long-term recovery (e.g., GOSE at 6 months), showing patterns not emphasized in the global model. By excluding "GMs", this analysis avoids the confounding influence of genetic markers, which might not directly influence imaging variables. This allows for focused insights into injury-outcome relationships, which are clinically actionable for predicting recovery trajectories.

Case III: Unveiling the Genetic Underpinnings of TBI Recovery with PCAmix focuses on the relationship between genetic markers (GM) and functional and cognitive outcomes (FCO), offering insights into how genetic variability might influence recovery. This level of specificity is absent in the global PCAmix, where genetic markers are marginalized due to their relatively small contribution to the total variance. The third study case helps uncover patterns in how genetics

influence recovery from brain injury, like the association of "GM1" and "GM6" with PTSD scores and neurocognitive performance, indicating a potential role of dopaminergic pathways in emotional and cognitive recovery. Such specificity is invaluable for personalized treatment approaches. Moreover, the analysis identifies potential clusters or subgroups based on genetic variability, hinting at individualized recovery trajectories. This case enables a deeper understanding of the "GMs" and "FCOs" relationship and reveals precise genotype-phenotype relationships, offering valuable insights for personalized recovery strategies and genetic research.

6. Conclusions

This study demonstrates the utility of PCA as a powerful technique for dimensionality reduction and data interpretation by selecting ordered and uncorrelated PCs. Using the PCAmix method, we effectively analyzed mixed data comprising both numerical and categorical variables, assigning equal importance in the final components without analyzing each type separately. Our findings identified three key domains influencing brain injury outcomes: brain imaging findings (BIF) assessing structural brain injury, functional and cognitive outcomes (FCO) evaluating recovery, and genetic markers (GM).

Global PCAmix is an ideal approach for initial broad exploration, capturing dominant variance trends across all domains. However, it lacks the precision required to disentangle specific relationships, making it less effective at uncovering associations or interactions between variables in distinct subdomains. In contrast, Unveiling the Genetic Underpinnings and Biomarkers of TBI Recovery with PCAmix analyses offers greater specificity and clarity, focusing on relationships within subdomains by limiting the scope to selected variable groups. This avoids the issue of dominant variables overshadowing smaller but significant contributions, enhances differentiation between variables within a domain, and provides focused interpretations by addressing specific research questions. Hidden patterns and relationships that may be obscured in the global analysis become evident in Unveiling some important factors, making them more effective for targeted research questions, while the global analysis offers a high-level summary.

These results emphasize the multifactorial nature of brain injury outcomes, where interactions between structural damage, genetic factors, and functional recovery collectively shape patient prognosis. This comprehensive approach could guide targeted interventions or personalized medicine strategies. PCAmix provides a robust framework for analyzing such complex interactions, balancing the need for both broad overviews and detailed subdomain insights.

Conflict of interest

The authors declare no conflicts of interest with the manuscript's content, either entirely or partially.

References

[1] Murtagh F, Heck A. *Multivariate Data Analysis*. Springer Science and Business Media; 2012.

[2] Beckett C, Eriksson L, Johansson E, Wikström C. Multivariate data analysis (MVDA). In: Schlindwein WS, Gibson M. (eds.) *Pharmaceutical Quality by Design: A Practical Approach*. John Wiley and Sons; 2018. p.201-225.

[3] Grentzelos C, Caroni C, Barranco-Chamorro I. A comparative study of methods to handle outliers in multivariate data analysis. *Computational and Mathematical Methods*. 2021; 3(3): e1129.

[4] Mardia KV, Kent JT, Taylor CC. *Multivariate Analysis*. John Wiley & Sons; 2024.

[5] Everitt B, Hothorn T. *An Introduction to Applied Multivariate Analysis with R*. Springer, New York; 2011.

[6] Backhaus K, Erichson B, Gensler S, Weiber R, Weiber T. Multivariate analysis. *Springer Books*. 2021; 10: 978-397.

[7] Berkenkemper S, Klinken S, Kleinebudde P. Multivariate data analysis to evaluate commonly used compression descriptors. *International Journal of Pharmaceutics*. 2023; 637: 122890.

[8] Jolliffe IT, Cadima J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*. 2016; 374(2065): 20150202.

[9] Greenacre M, Groenen PJ, Hastie T, d'Enza AI, Markos A, Tuzhilina E. Principal component analysis. *Nature*

Reviews Methods Primers. 2022; 2(1): 100.
[10] Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports*. 2022; 12(1): 14683.
[11] Gewers FL, Ferreira GR, Arruda HFD, Silva FN, Comin CH, Amancio DR, et al. Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys (CSUR)*. 2021; 54(4): 1-34.
[12] Brunette M, Bourke R, Hanewinkel M, Yousefpour R. Adaptation to climate change in forestry: A multiple correspondence analysis (MCA). *Forests*. 2018; 9(1): 20.
[13] Husson F, Josse J. *Multiple Correspondence Analysis*. Visualization and Verbalization of Data; 2014.
[14] Greenacre M. *Correspondence Analysis in Practice*. Chapman and Hall/CRC; 2017.
[15] Florensa D, Godoy P, Mateo J, Solsona F, Pedrol T, Mesas M, et al. The use of multiple correspondence analysis to explore associations between categories of qualitative variables and cancer incidence. *IEEE Journal of Biomedical and Health Informatics*. 2021; 25(9): 3659-3667.
[16] Olugbara CT, Letseka M, Olugbara OO. Multiple correspondence analysis of factors influencing student acceptance of massive open online courses. *Sustainability*. 2021; 13(23): 13451.
[17] Moschidis S, Markos A, Thanopoulos AC. "Automatic" interpretation of multiple correspondence analysis (MCA) results for nonexpert users, using R programming. *Applied Computing and Informatics*. 2022; 1-12.
[18] Abdi H, Valentin D. Multiple correspondence analysis. In: Salkind NJ. (ed.) *Encyclopedia of Measurement and Statistics*. SAGE Publications, Thousand Oaks; 2007. p.1-13.
[19] Severeyn E, La Cruz A, Puente C, Velásquez J, Huerta M. Unveiling the barriers to e-business adoption in venezuelan public enterprises: A multiple correspondence analysis approach. In: *2024 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE; 2024. p.1-6.
[20] Duval J, Inglis D, Almila AM. Multiple correspondence analysis. In: *The SAGE Handbook of Cultural Sociology*. Sage Publications; 2016. p.255-271.
[21] Zhou L, Liu Z, Liu F, Peng J, Zhou T. Nonlinear canonical correspondence analysis and its application. *Scientific Reports*. 2023; 13(1): 7518.
[22] Jolliffe IT. *Principal Component Analysis for Special Types of Data*. Springer, New York; 2002.
[23] Bouwmans T, Javed S, Zhang H, Lin Z, Otazo R. On the applications of robust PCA in image and video processing. *Proceedings of the IEEE*. 2018; 106(8): 1427-1457.
[24] Kurita T. Principal component analysis (PCA). In: *Computer Vision: A Reference Guide*. Cham: Springer International Publishing; 2021. p.1013-1016.
[25] Rodríguez-Heras JD, Cabeza-Marchena JA, Nieto-Ramos LM, Márquez-Castillo AE, Garizabal-Donado LE. Principal component analysis method application for inventory related decisions-making. *Procedia Computer Science*. 2024; 241: 558-563.
[26] Holland S. *Principal Component Analysis (PCA)*. University of Georgia; 2019.
[27] Sanguansat P. *Principal Component Analysis: Engineering Applications*. BoD-Books on Demand; 2012.
[28] De Leeuw J, Van Rijckevorsel J. *HOMALS and PRINCALS-Some Generalizations of Principal Components Analysis*. Data Analysis and Informatics; 1980.
[29] Kiers HAL. *Simple Structure in Component Analysis Techniques for Mixtures of Qualitative and Quantitative Variables*. Psychometrika; 1991.
[30] Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate Analysis of Mixed Data: The R Package PCAmixdata. *arXiv:1411.4911*. 2014. Available from: https://doi.org/10.48550/arXiv.1411.4911.