Cognitive Profiles of People Living with Dementia - PCA-Based Clustering Analysis

Hanlong Chen[1,2], Bruna Seixas-Lima[1], Howard Chertkow[1,3], Malcolm Binns[1,2]

[1]Rotman Research Institute, Baycrest

[2]Dalla Lana School of Public Health, Biostatistics Division, University of Toronto

[3]Institute of Medical Science, Temerty Faculty of Medicine, University of Toronto

Baycrest
Rotman Research Institute

Dalla Lana
School of Public Health

Background

- Despite increasing evidence that shows diverse cognitive patterns in dementia, existing clinical categories often overlook **subtle differences** in core domains such as **memory, language, attention, executive** and **visuospatial function**.

- Our objective is to find and classify **statistically distinct cognitive profiles** that contribute to both **clinical decision-making** and future research on **individualized interventions**.

Toronto Dementia Research Alliance (TDRA):

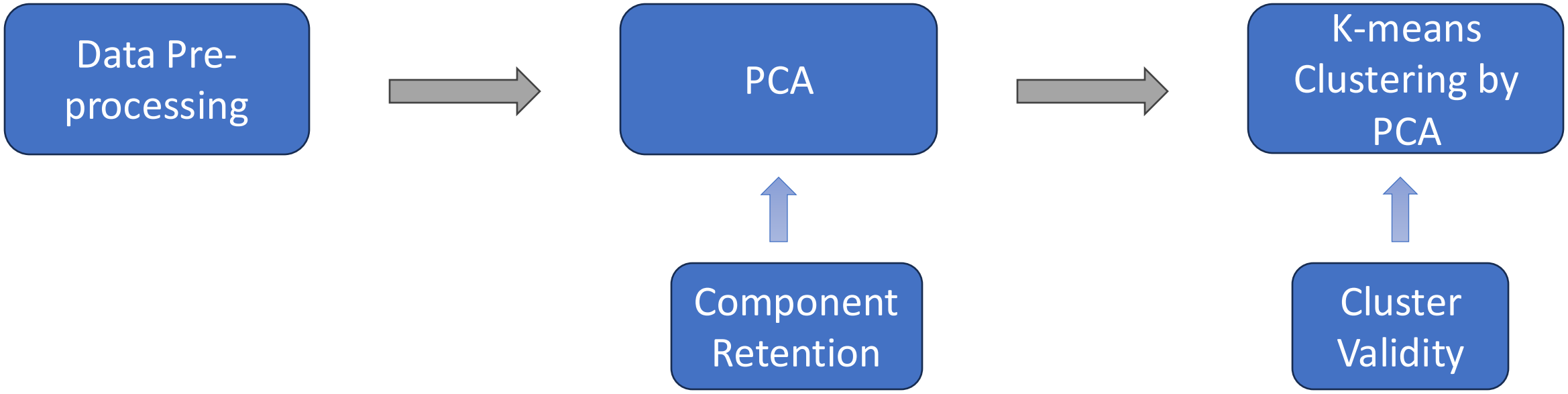
- Centralizes data to connect basic science with clinical research, aiming to better understand, prevent, and treat dementia from the following partners.

- The **dataset** for this research was obtained from the **TDRA**, a collaborative effort involving the institutions below.

Baycrest camh Ontario Shores Centre for Mental Health Sciences Sunnybrook Health Sciences Centre UNITY HEALTH TORONTO UHN

Method

- **Applied dimensionality-reduction** and **clustering framework** to define meaningful data-driven dementia subtypes in core cognitive domains statistically.
- **Compared different component retention strategies for Principal Component Analysis (PCA)** and **verified cluster validity** systematically.



Data Pre-processing

We analyzed **TDRA** data from **721** (reduced from 2394) individuals with dementia.

- Patient selection based on established research criteria.
- **Selection of total scores** to reduce redundancy and resolve multicollinearity.
- **Imputation of missing values** for the selected test scores using mean imputation as there are very few missing values.
- Identification and removal of multivariate outliers using the **Mahalanobis distance** method.

Each patient assessed through several neuropsychological tests from **Toronto Cognitive Assessment** (TorCA) reflecting

- **Memory**
- **Executive Function**
- **Attention/Working Memory**
- **Language**
- **Visuospatial Function**
- **Orientation**

Trial 1	2	3	4	5	6	7	8	9	Total Scores
1	1	1	1	1	1	1	1	1	9

Table 1: Example of the TorCA Dataset

Data Pre-processing

There are now **721 number of observations** and **25 variables (tests)** in the dataset for PCA.

Test Domains	
Language	Orientation
Memory	Attention/WM
Executive	Visuospatial

Test Name	Variable Name
Verbal Fluency	torca_vf_tot
Semantic Fluency	torca_sf_tot
MINT Naming Total Correct	torca_naming_tot_scr
Sentence Repetition Total	torca_srept_tot
Single Word Reading Total	torca_lang_swr_tot
Semantic Knowledge Total	torca_lang_sem_tot
Single Word Comprehension Total	torca_sw_compr_tot
Single Word Reading Comprehension Total	torca_swr_compr_tot
Sentence Comprehension Total Score	torca_sntc_compr_tot
Orientation Total	torca_orient_tot

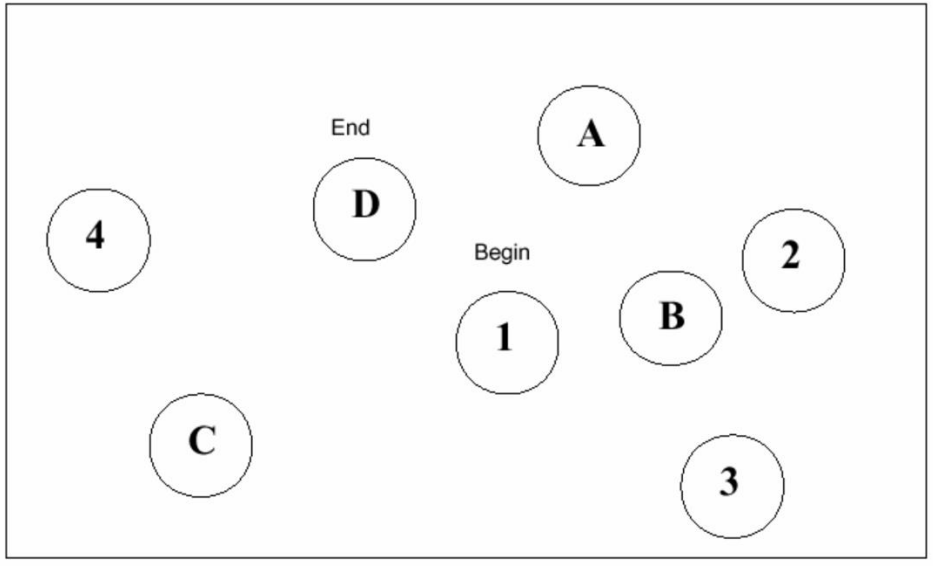
Test Name	Variable Name
CERAD Trials Total	torca_cerad_trial_tot
CERAD Delayed Recall Total	torca_cd_rcll_crrct
CERAD Delayed Recognition Total	torca_cd_rcg_crrct
Figure Recall Total	torca_frcl_tot
Figure Recognition Total	torca_frcg_tot
Serial 7's Total	torca_serial7_tot
Serial 3's Total	torca_serial3_tot
Longest Forward Digit Span	torca_ds_long_f_tot
Longest Backward Digit Span	torca_ds_long_b_tot
Trails A Total Score	torca_trails_num_tot
Trails B Total Score	torca_trails_ltr_tot
Similarities Total Score	torca_simil_tot
Alternating Sequences Total Score	torca_alt_seq_tot
Clock Drawing Total	torca_clock_tot
Benson Figure Copy Total	torca_fc_tot

Principal Component Analysis

Why PCA Prior to Clustering?

- **Dimensionality Reduction**
  - Fewer components while preserving variance
  - Mitigates "curse of dimensionality" for more robust clustering

- **Addresses Multicollinearity (Overlapping Domains over Tests) & Improves Interpretability**
  - Tests rarely measure isolated domains in real life
  - Example: **Trail B** assesses executive switching + visuospatial search



Principal Component Analysis

Why PCA Prior to Clustering?

Problems with Direct Clustering on 25 Raw Test Scores

- **Curse of Dimensionality**
  - Distance metrics become less meaningful in high dimensions
  - Clusters become sparse and poorly separated
  - Most algorithms perform poorly with dimensions > 10-15

Feature space	Algorithm	Silhouette (k=4)	R^2	CH
PC2 - PC5	k-means	~ 0.19	0.433	180
PC2 - PC5	Ward's	~ 0.18	0.366	136
25 total scores	k-means	~ 0.12	0.305	103
25 total scores	Ward's	~ 0.06	0.260	82

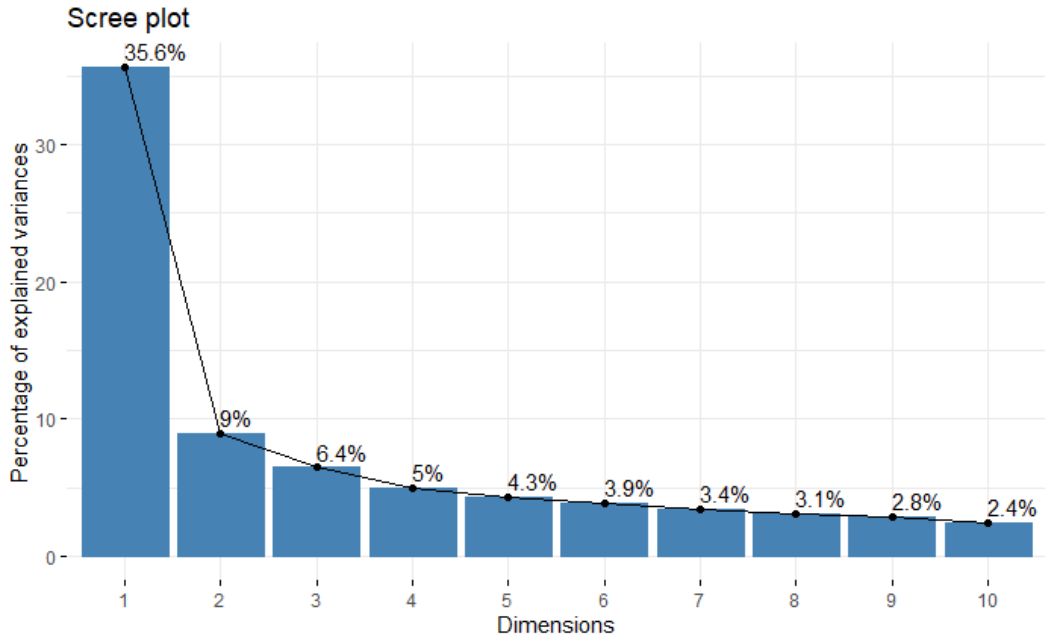
- PCA-based clustering achieves 40% better R^2 and 75% higher Calinski-Harabasz (CH) index than raw score clustering
- K-means clustering has the highest silhouette score (0.19), Calinski-Harabasz index (180) and explained approximately 43% of the total variance (R^2 = 0.433).

Principal Component Analysis

The **optimal number of principal components** to retain was determined by multiple established criteria

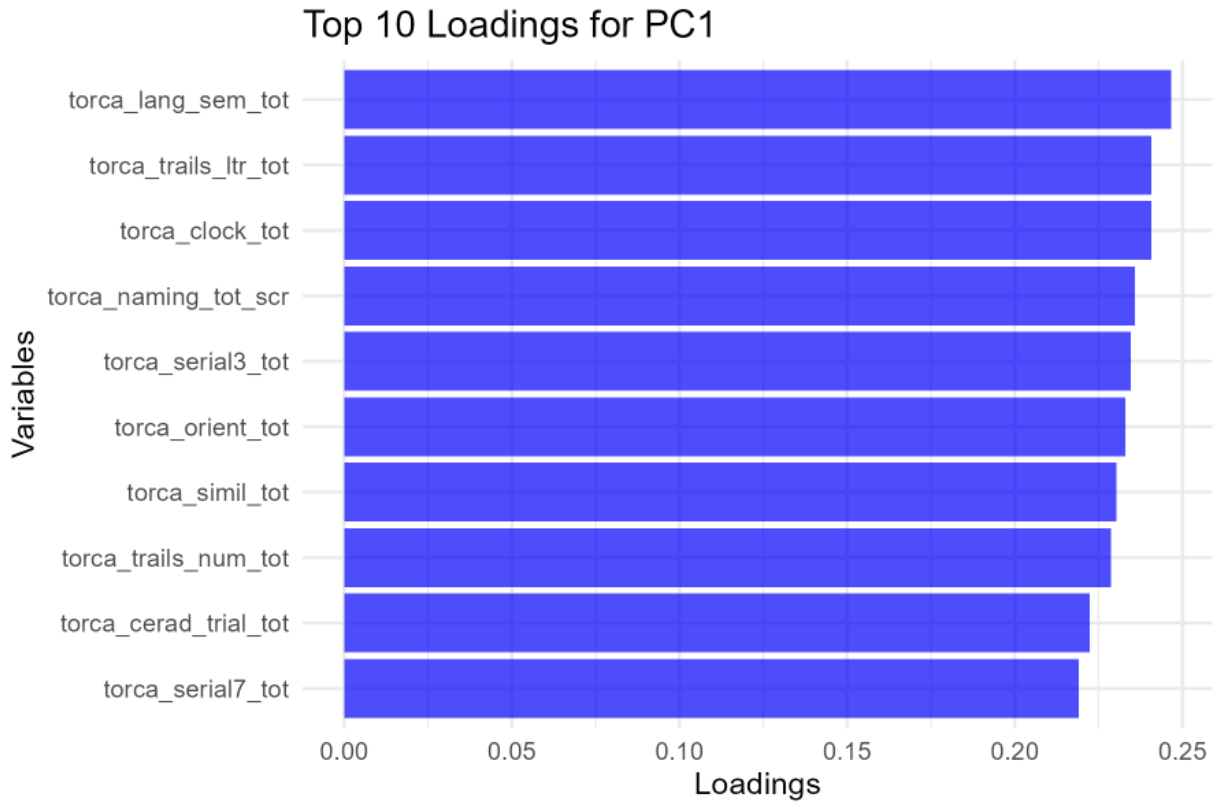
- Cattell's Scree Plot/Elbow Criterion
- Cumulative proportion of variance explained
- Kaiser Criterion
- Mean eigenvalue criterion
- Broken Stick
- Horn's Parallel Analysis
- Cross-validation

Retention Method	Number of Retained Components
Cumulative Porprotion	11
Kaiser Criterion	5
Mean Eigenvalue Criterion	5
Broken Stick	10
Parallel	3
Cross-Validation	about 5
Retained Component Criterion (RCC Package)	5
Jackknife	21
Bootstrap	22



- Based on a synthesis of these criteria, **five** principal components were retained (explained 60.2% of the total variance).
- The loadings from the PCA were examined to interpret the cognitive domains captured by each retained component.

Principal Component Analysis - Loadings

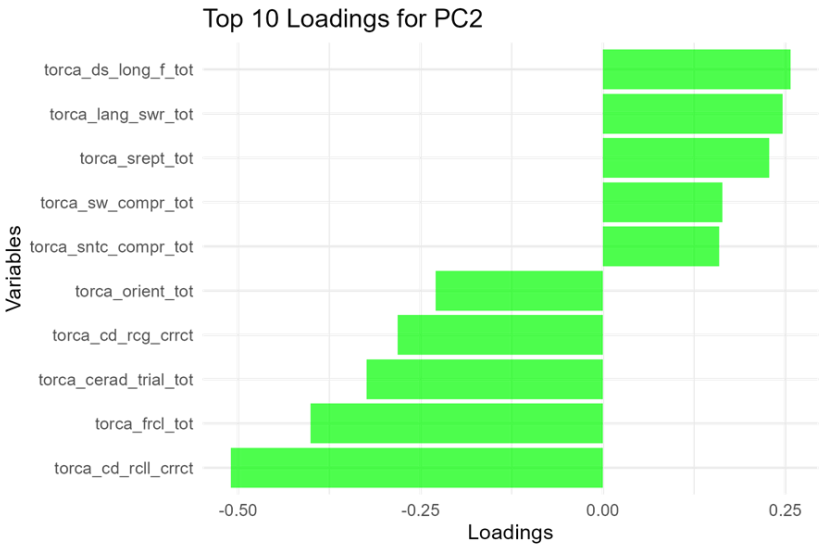


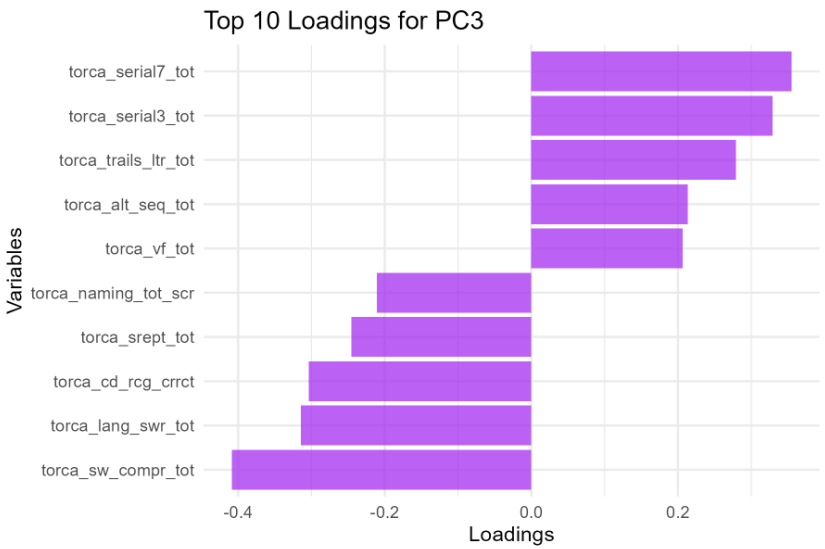
The loadings from the PCA were examined to interpret the cognitive domains captured by each component.

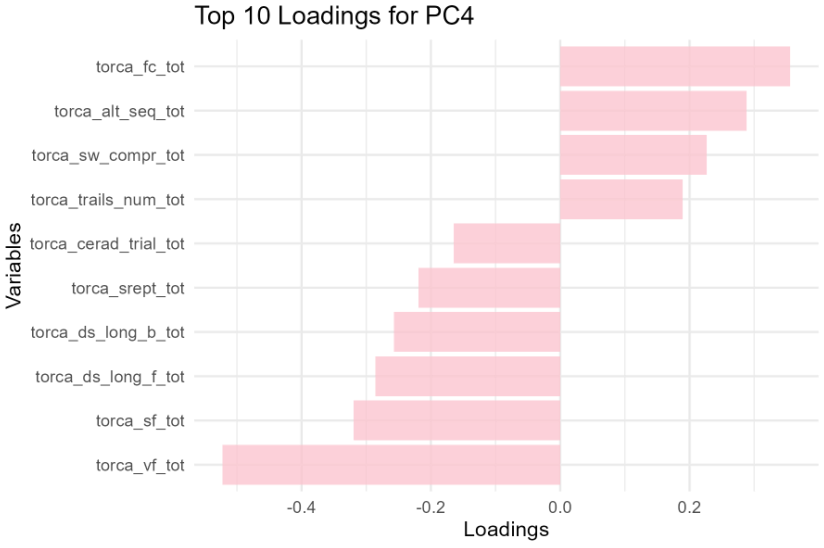
- PC1 was identified as representing **overall cognitive severity**

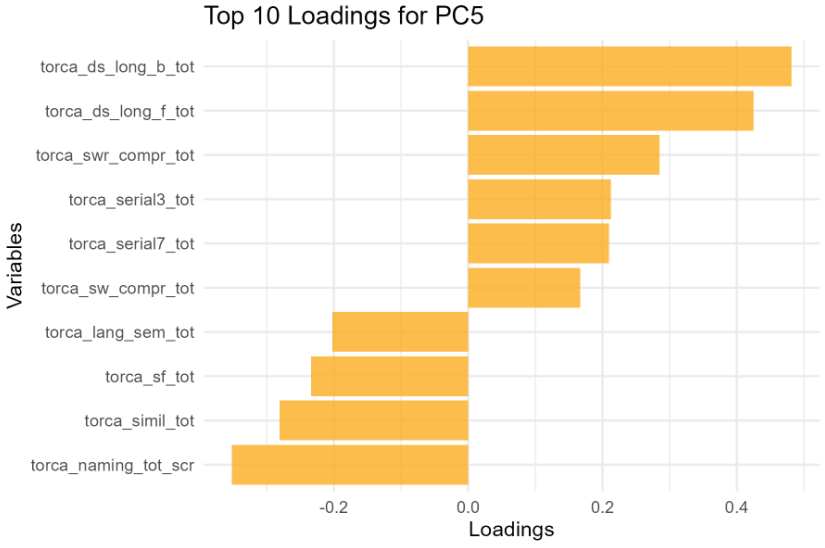
$$PC1\ Score = \beta_1 \times x_1 + \beta_2 \times x_2 + \cdots + \beta_{10} \times x_{10} + \ldots$$

Principal Component Analysis - Loadings



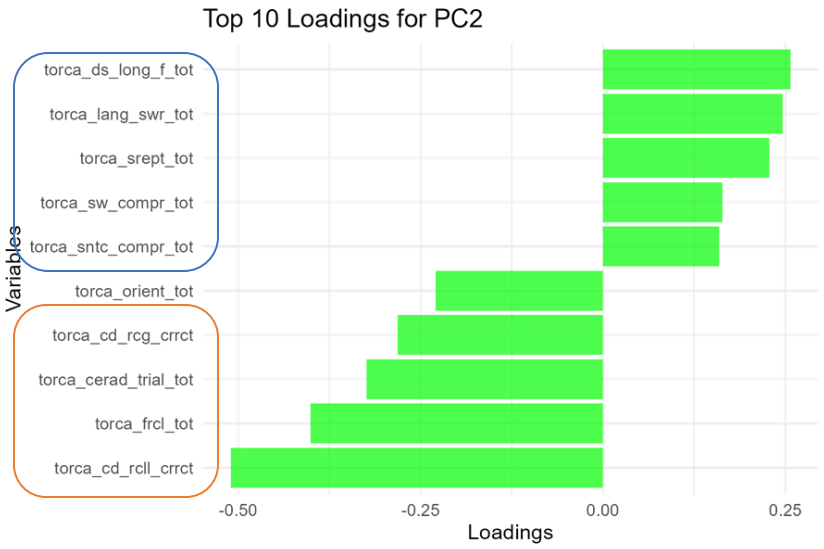


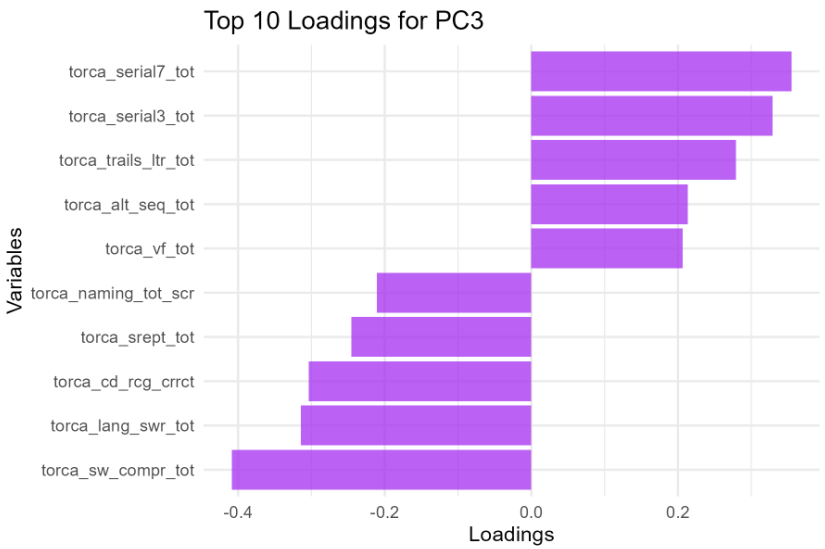


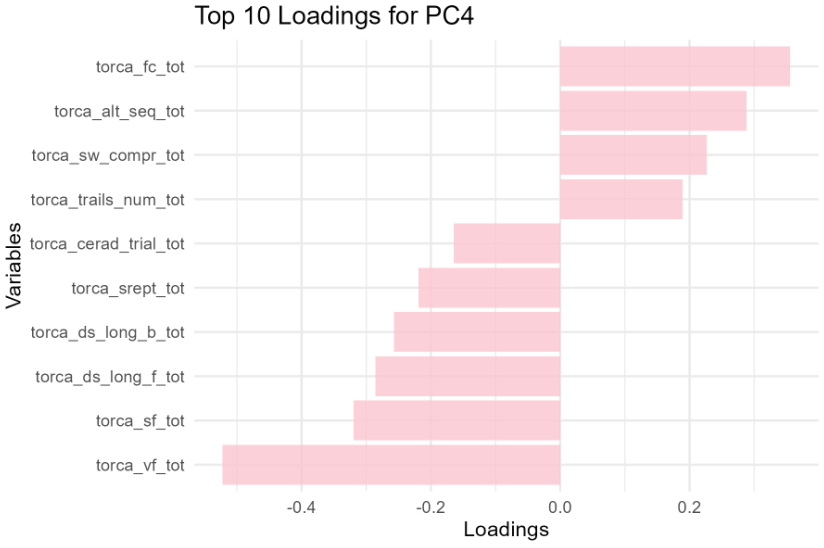


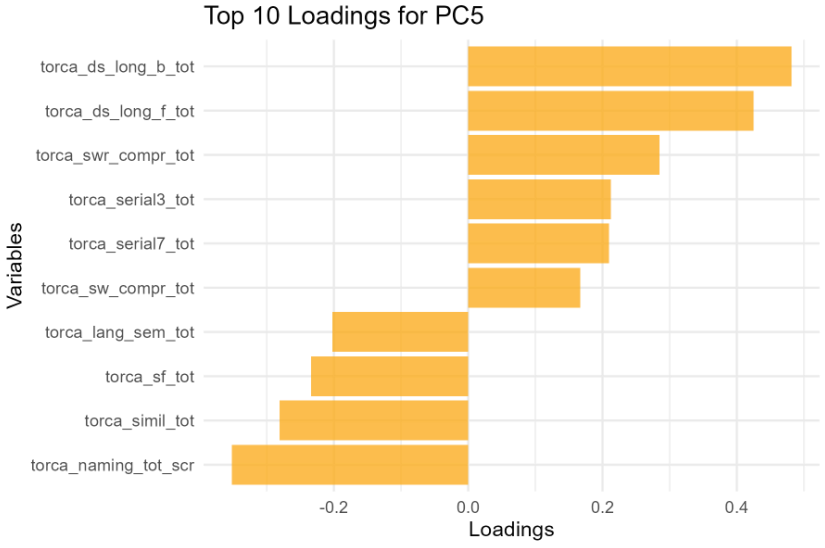
PC2 through PC5 captured patterns of **relative strengths and weaknesses** across various interrelated cognitive domains.

Principal Component Analysis - Loadings









PC2 through PC5 captured patterns of **relative strengths and weaknesses** across various interrelated cognitive domains.

Principal Component Analysis – Result

	Better	Worse
PC1	Overall Cognitive Severity	
PC2	**Attention + Language** Digit Span (F), Single Word Reading, Sentence Repetition	**Episodic Memory** CERAD Delayed Recall, Figure Recall
PC3	**Executive + Attention** Serial 3/7, Trails B, Alternating Sequence	**Language** Single Word Reading & Comprehension, Naming, Sentence Repetition
PC4	**Visuospatial + Executive** Figure Copy, Alternating Sequence, Trails A	**Verbal Fluency + Attention** Verbal Fluency, Semantic Fluency, Digit Span (F/B)
PC5	**Attention** Digit Span (F/B), Serial 3/7	**Semantic Language** Naming, Semantic Fluency, Semantic Knowledge

Cluster Analysis – K-means vs. Ward’s

Clustering Methods Comparison and Selection

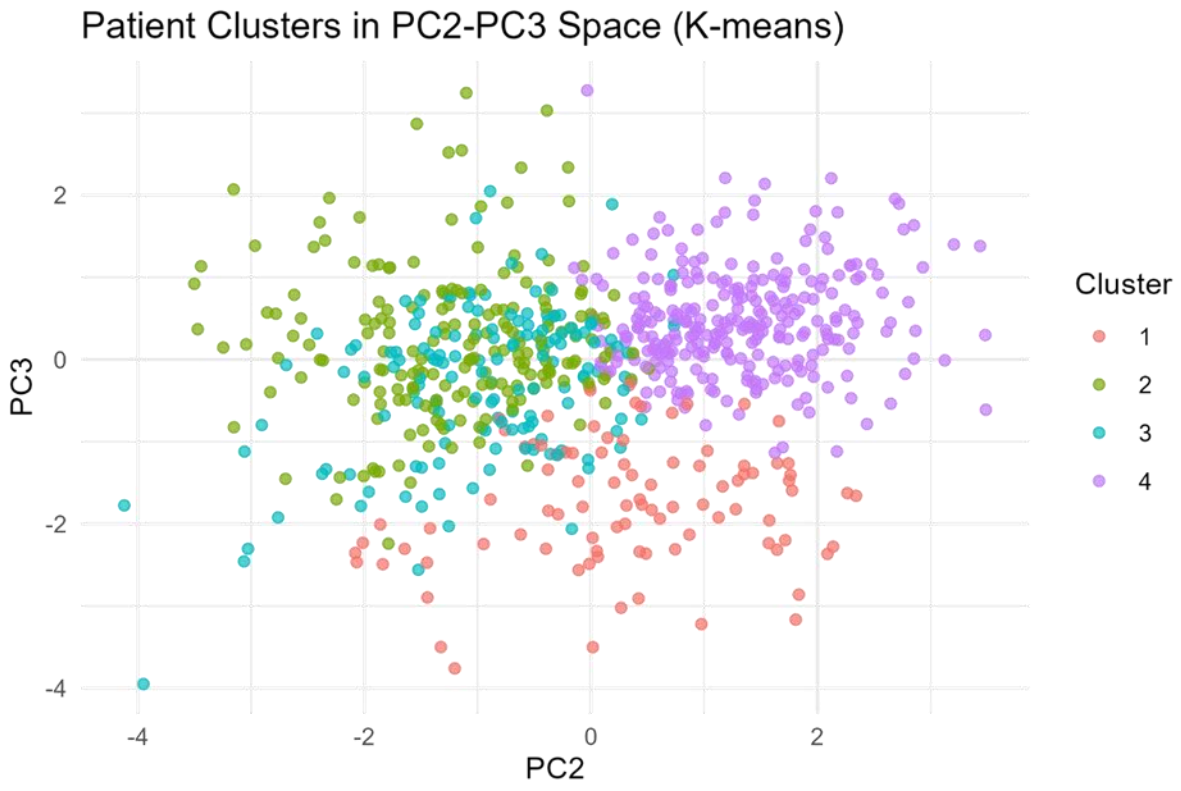
We applied both **K-means** clustering and **Ward's hierarchical** clustering method to the participant scores on **PC2-PC5**.

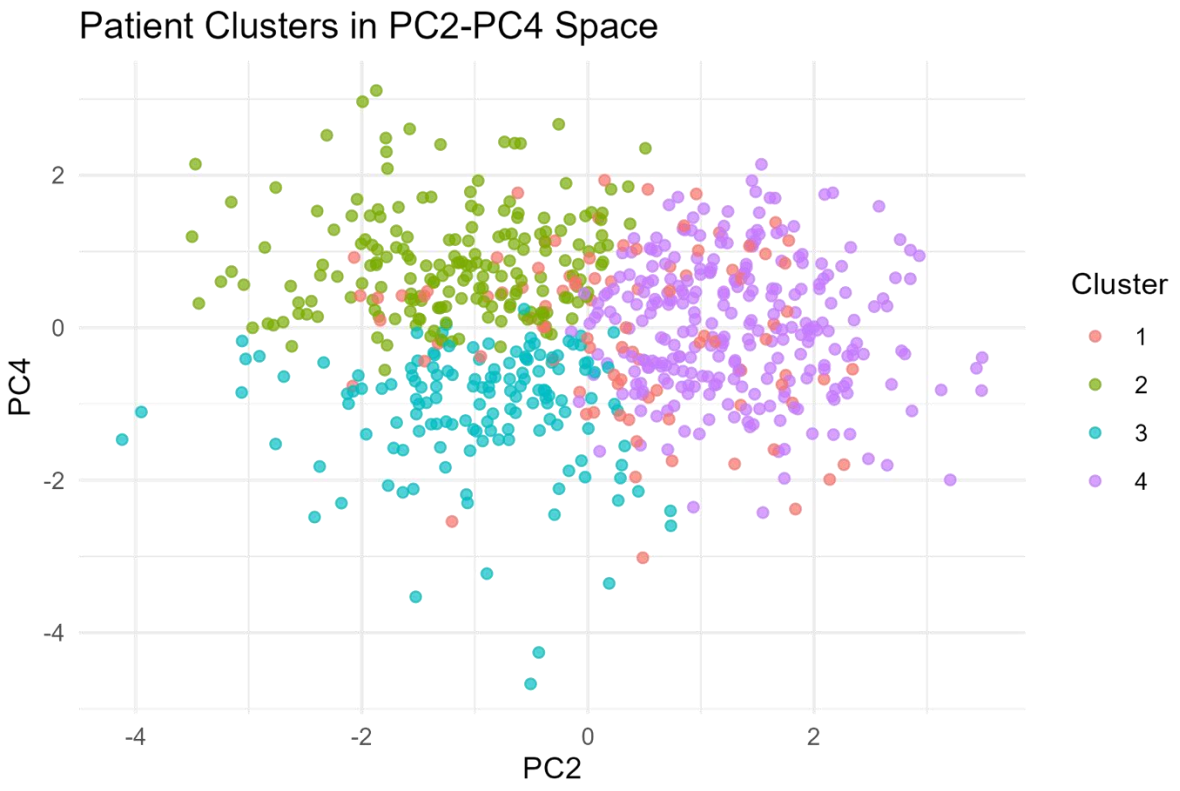
- These specific components were chosen as they reflect **relative cognitive profiles**, such as memory-language tradeoffs, executive function, rather than absolute levels of impairment.

- **PC1**, primarily reflecting global cognitive severity, was **intentionally excluded from** the clustering input to prevent it from dominating the clustering result and potentially obscuring more subtle, meaningful cognitive subtypes.

- **Both clustering** methods produced remarkably **similar profile for clusters**, the consistency between the K-means and Ward's methods enhance confidence in the robustness of the identified cognitive profiles.

Cluster Analysis – Scatter Plots



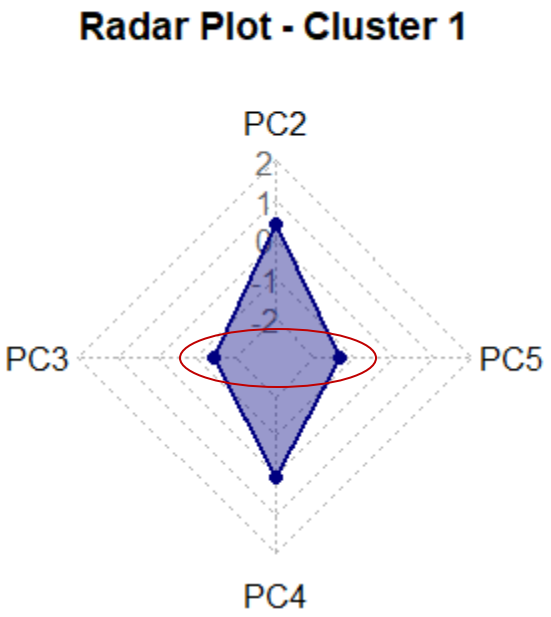


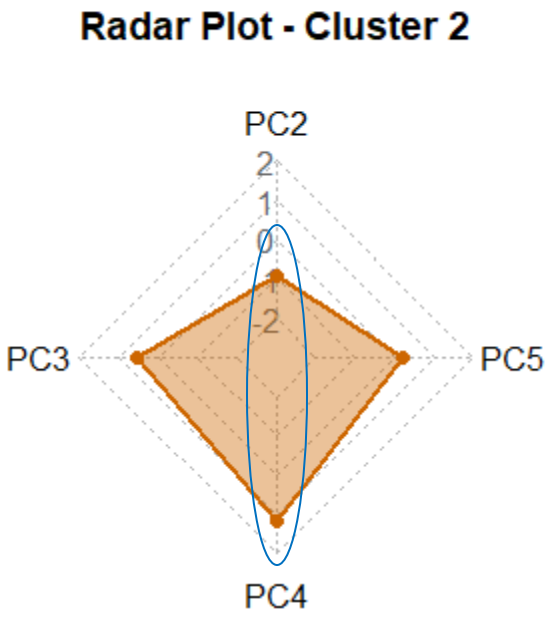
- PC2 splits **cluster 2,3** and **cluster 4**.
- PC3 splits **cluster 1** and **cluster 2,3,4**.
- PC4 splits **cluster 2** and **cluster 3**.

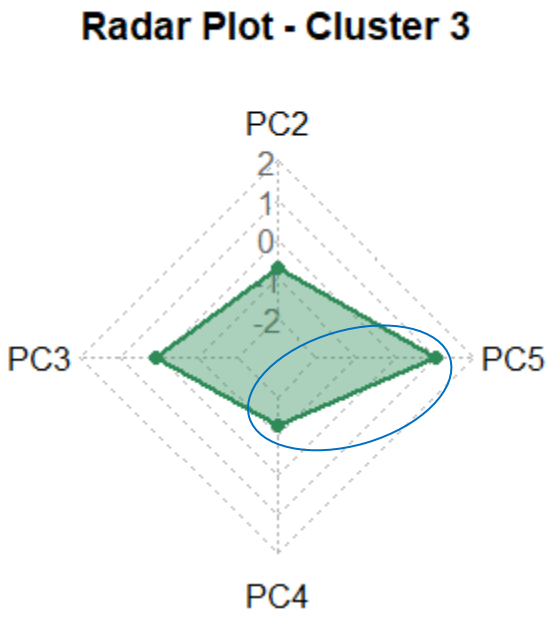
Cluster Analysis - Results

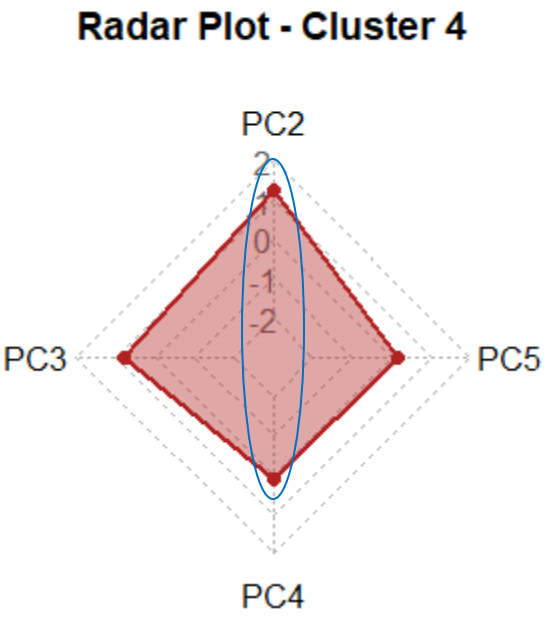
Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Avg Age	74.16	70.93	68.14	77.73
Cluster Size	92	194	145	278
Severity (PC1)	-2.37	0.33	1.03	0.28
PC2 Mean	0.32	-1.20	-0.96	1.31
PC3 Mean	-1.77	0.22	-0.29	0.46
PC4 Mean	-0.03	0.85	-1.07	0.02
PC5 Mean	-0.58	0.04	0.36	0.01
AD	34 (37%)	20 (10.3%)	19 (13.1%)	90 (32.4%)
MCI	21 (22.8%)	114 (58.8%)	88 (60.7%)	148 (53.2%)
V-MCI	11 (12%)	36 (18.6%)	24 (16.6%)	9 (3.2%)
VaD	14 (15.2%)	12 (6.2%)	8 (5.5%)	7 (2.5%)
Mixed	12 (13%)	8 (4.1%)	6 (4.1%)	24 (8.6%)

Cluster Analysis - Results









	Better	Worse
PC2	Attention + Language	Episodic Memory
PC3	Executive + Attention	Language
PC4	Visuospatial + Executive	Verbal Fluency + Attention
PC5	Attention	Semantic Language

- **Cluster 1:** Low PC3, PC5, indicating relatively worse in **Attention** and **Executive function**, relatively better in **Language**.
- **Cluster 2:** Low PC2, indicating relatively worse on **Attention and Language**, but relatively better in **Episodic Memory**. High PC4, relatively better in **Visuospatial** and **Executive functions**, worse in **Attention** and **Language (verbal fluency)**.
- **Cluster 3:** Low PC4, indicating relatively worse in **Visuospatial and Executive functions**, but better in **verbal fluency** and **Attention**. High PC5, better in **Attention** but worse in **semantic language**.
- **Cluster 4**: Low PC4, indicating slightly worse in **Visuospatial and Executive function**, but relatively better in **verbal fluency** and **Attention**. High PC2, indicating better **Language** and **Attention**, worse on **Episodic Memory.**

Cluster Analysis - Results

Cluster 1 (n = 92, age = 74.16)	
Key PC profile	Lowest PC1, PC3, PC5
Cognitive strengths	**General language** relatively preserved
Cognitive weaknesses	**General impairment**, relatively more severe **executive dysfunction**
Clinical interpretation	Older & More Severe mixed dementia phenotype
Cluster 2 (n = 194, age = 70.93)	
Key PC profile	Lowest PC2, Highest PC4
Cognitive strengths	Relatively better in **episodic memory** (delayed recall), visuospatial and executive functions
Cognitive weaknesses	Relatively worse in **attention**, **language** (verbal fluency)
Clinical interpretation	Early-Onset Alzheimer's Disease (EOAD)-like non-amnesic subtype
Cluster 3 (n = 145, age = 68.14)	
Key PC profile	Highest PC1, Lowest PC4, Highest PC5
Cognitive strengths	Relatively better in **attention** and **language** (verbal fluency)
Cognitive weaknesses	Relatively worse in **semantic language**, **visuospatial** and **executive functions**
Clinical interpretation	Younger, Language-advantaged subtype
Cluster 4 (n = 278, age = 77.73)	
Key PC profile	Highest PC2
Cognitive strengths	Relatively stronger in **language** and **attention**
Cognitive weaknesses	Relatively weaker in **episodic memory**
Clinical interpretation	Late-Onset Alzheimer's Disease (LOAD)-like amnesic subtype

Cluster Analysis - Conclusions

Severity and Diagnosis Relationship

- Clusters with **lower PC1 scores** (e.g., Cluster 1) tend to show higher percentages of **severe diagnoses (AD, VaD).**
- While clusters with **higher PC1 scores** (e.g., Cluster 2, 3) more frequently exhibit **milder diagnoses (MCI)**.

Age-Related Cognitive Patterns

- **PC2** showed a notable **variation with age (ρ = 0.36)**.
- The **oldest cohort** (Cluster 4, avg age **77.7 yrs**) has **highest PC2** scores, indicating better attention (working memory) relatively more impaired episodic memory.
- In contrast, a **younger group** (Cluster 2, avg age **70.9 yrs**) showed **lower PC2** scores and the opposing pattern of relative cognitive differences.
- These distinct **age-related profiles** help explain how domains may be differentially vulnerable depending on **EOAD** vs. **LOAD**.

Cluster Analysis - Results

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Avg Age	74.16	70.93	68.14	77.73
Cluster Size	92	194	145	278
Severity (PC1)	-2.37	0.33	1.03	0.28
PC2 Mean	0.32	-1.20	-0.96	1.31
PC3 Mean	-1.77	0.22	-0.29	0.46
PC4 Mean	-0.03	0.85	-1.07	0.02
PC5 Mean	-0.58	0.04	0.36	0.01
AD	34 (37%)	20 (10.3%)	19 (13.1%)	90 (32.4%)
MCI	21 (22.8%)	114 (58.8%)	88 (60.7%)	148 (53.2%)
V-MCI	11 (12%)	36 (18.6%)	24 (16.6%)	9 (3.2%)
VaD	14 (15.2%)	12 (6.2%)	8 (5.5%)	7 (2.5%)
Mixed	12 (13%)	8 (4.1%)	6 (4.1%)	24 (8.6%)

Acknowledgement

- **Dr. Malcolm Binns** [supervision]
- **Dr. Bruna Seixas Lima** [curation]
- **Dr. Howard Chertkow** [conceptualization]

Bibliography

[1] Greenacre, M., Groenen, P.J.F., Hastie, T. *et al.* " Principal component analysis." *Nat Rev Methods Primers* **2**, 100 (2022). https://doi.org/10.1038/s43586-022-00184-w

[2] Cangelosi, Richard, and Alain Goriely. "Component retention in principal component analysis with application to cDNA microarray data." *Biology direct 2* (2007): 1-21.

[3] Franklin, Scott B., et al. "Parallel analysis: a method for determining significant principal components." *Journal of Vegetation Science 6.1* (1995): 99-106.

[4] Jackson, J. E., and A. Edward. "User's guide to principal components. " *John Willey Sons. Inc., New York 40* (1991).

[5] Habes, M., Grothe, M. J., Tunc, B., McMillan, C., Wolk, D. A., & Davatzikos, C. (2020). Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods. *Biological psychiatry*, *88*(1), 70-82.

Thank You!

I am happy to answer any questions!

E-mail: sl.chen@mail.utoronto.ca

Mahalanobis Distance

ID	Age	Memory Score	Language Score
A	72	22	15
B	69	23	16
C	71	21	17
Z	75	22	25 ← Looks normal by each variable, but is multivariate outlier jointly

Mahalanobis Distance

2 . Compute the sample mean (μ) and covariance matrix (Σ) from A, B, C

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{M} \\ \bar{L} \end{bmatrix} = \begin{bmatrix} 22 \\ 16 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

(variances = 1; covariance = –0.5)

3 . Mahalanobis distance for Z

$$\mathbf{d} = \mathbf{x}_Z - \boldsymbol{\mu} = \begin{bmatrix} 22 - 22 \\ 25 - 16 \end{bmatrix} = \begin{bmatrix} 0 \\ 9 \end{bmatrix}$$

$$D_M^2 = \mathbf{d}^{\mathsf{T}} \, \Sigma^{-1} \, \mathbf{d} = 108 \quad \Longrightarrow \quad D_M = \sqrt{108} = 10.39$$

4 . Statistical decision

For **p = 2** variables and significance **α = 0.01**:

$$\chi^2_{2,\,0.99} = 9.21$$

$$D_M^2 = 108 \; > \; 9.21 \quad \Longrightarrow \quad \boxed{\text{Z is a multivariate outlier}}$$

Retention Criteria

Component	Eigenvalue	% Variance
PC1	**2.70**	67.5 %
PC2	**1.10**	27.5 %
PC3	0.14	3.5 %
PC4	0.06	1.5 %
Sum	**4.00**	100 %