High Dimensional Data Analysis Project 2026

Analysis of Smartphone Addiction among AIMS Students

Jean Baptiste Moute
MAIRAME SAMBA NIANG
NDEYE BINTA NDIAYE
Gaolatlhe Angelah Kgato

African Institute for Mathematical Sciences, AIMS-Senegal

Supervised by Dr Sophie Dabo

January 22, 2026

Presentation Outline

AIMS African Institute for Mathematical Sciences SENEGAL

Introduction to High-Dimensional Data Analysis

Context

This section introduces the fundamental properties, procedures, key concepts, and basic rules of:

- **CA** (Correspondence Analysis)
- **MCA** (Multiple Correspondence Analysis)
- **SVD** (Singular Value Decomposition)

Pedagogical Approach

Emphasis on practical application: how to use CA, MCA, and SVD in real data analysis scenarios.

AIMS African Institute for Mathematical Sciences SENEGAL

Singular Value Decomposition (SVD) I

Mathematical Foundation

SVD decomposes any matrix $X_{n\times p}$ into three matrices: $X = U\sum V^T$:

- $U_{n\times n}$: left singular vectors (orthonormal)
- $\sum_{n\times p}$: diagonal matrix of singular values ($\sigma_1 \geq \sigma_2 \geq ... \geq 0$)
- $V_{p\times p}$: right singular vectors (orthonormal)

SVD: Practical Applications

Why SVD Matters in Practice

1. **Dimensionality Reduction**: Keep only top k components
2. **Noise Filtering**: Small singular values often represent noise
3. **Data Compression**: Store U_k, $\sum_k$, V_k instead of full X
4. **Foundation for PCA**: PCA uses SVD on centered data

AIMS African Institute for Mathematical Sciences SENEGAL

Correspondence Analysis (CA) I

What is CA?

CA is a dimension reduction technique for **contingency tables** (cross-tabulations of categorical variables).

- Visualizes associations between row and column categories
- Similar to PCA but for categorical data
- Based on chi-squared distances

Correspondence Analysis (CA) II

Mathematical Setup

Given a contingency table N with row totals r_i and column totals c_j:

- Correspondence matrix: $\mathrm{P} = \frac{1}{n}\mathrm{N}$ (joint probabilities)
- Row profiles: $\frac{N_{ij}}{N_i}$ (conditional distribution given row)
- Column profiles: $\frac{N_{ij}}{N_j}$ (conditional distribution given column)

Multiple Correspondence Analysis (MCA) I

Extension of CA

MCA extends CA to analyze **multiple categorical variables** simultaneously.

- Input: data matrix with n individuals and Q categorical variables
- Analyzes relationships between all variable categories
- MCA transforms categorical variables into a complete disjunctive (indicator) matrix and then applies SVD to this matrix under appropriate normalization.

MCA: Practical Implementation I

When to Use MCA

- Survey data with multiple-choice questions
- Customer segmentation with categorical features
- Pattern discovery in categorical datasets
- Complement to clustering for categorical data

Relationship: SVD, PCA, CA, and MCA

Method	Data Type	Matrix Decomposed
SVD	Any	X (raw data)
PCA	Quantitative	$X_{centered}$ or X_{scaled}
CA	Contingency table	Standardized residuals
MCA	Multiple categorical	Indicator matrix Z

Project Objectives

Problem Statement

Analyze smartphone addiction among AIMS students and identify predictive factors for the time spent on the phone.

- Dataset: **369 observations** (after duplicate removal)
- Studied variables: Phone, SocialNetworks, Happiness, Walk, InstagramRatio
- Approaches: PCA, Multiple Linear Regression, Model Selection

AIMS African Institute for Mathematical Sciences SENEGAL

Descriptive Statistics

Dataset Overview (N=369)

Variable	Mean	Std. Deviation	Min	Max
Phone (min)	125.89	132.64	1.0	1440.0
SocialNetworks	80.41	104.62	1.0	1440.0
Happiness	7.46	1.62	1.0	10.0
Walk (km)	6.21	3.24	1.0	20.0
InstagramRatio	1.17	3.20	0.0	51.33

Table: Statistical summary of main variables

Observations

- High variability in Phone usage (1-1440 min)
- High average happiness score (7.46/10)

Correlation Matrix

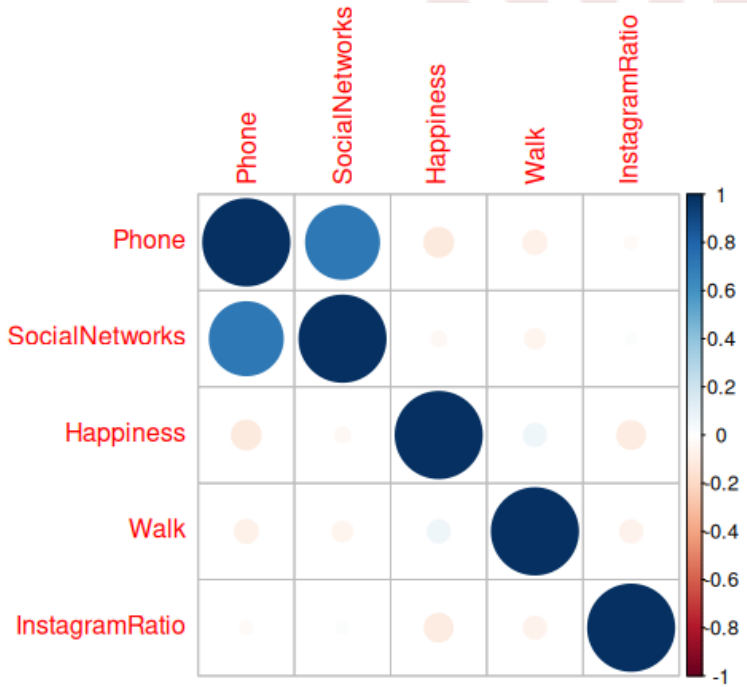

Figure: Correlations between variables

Interpretation

- Strong positive correlation between **Phone** and **SocialNetworks**
- Weaker correlations with **Happiness** and **Walk**

Scree Plot

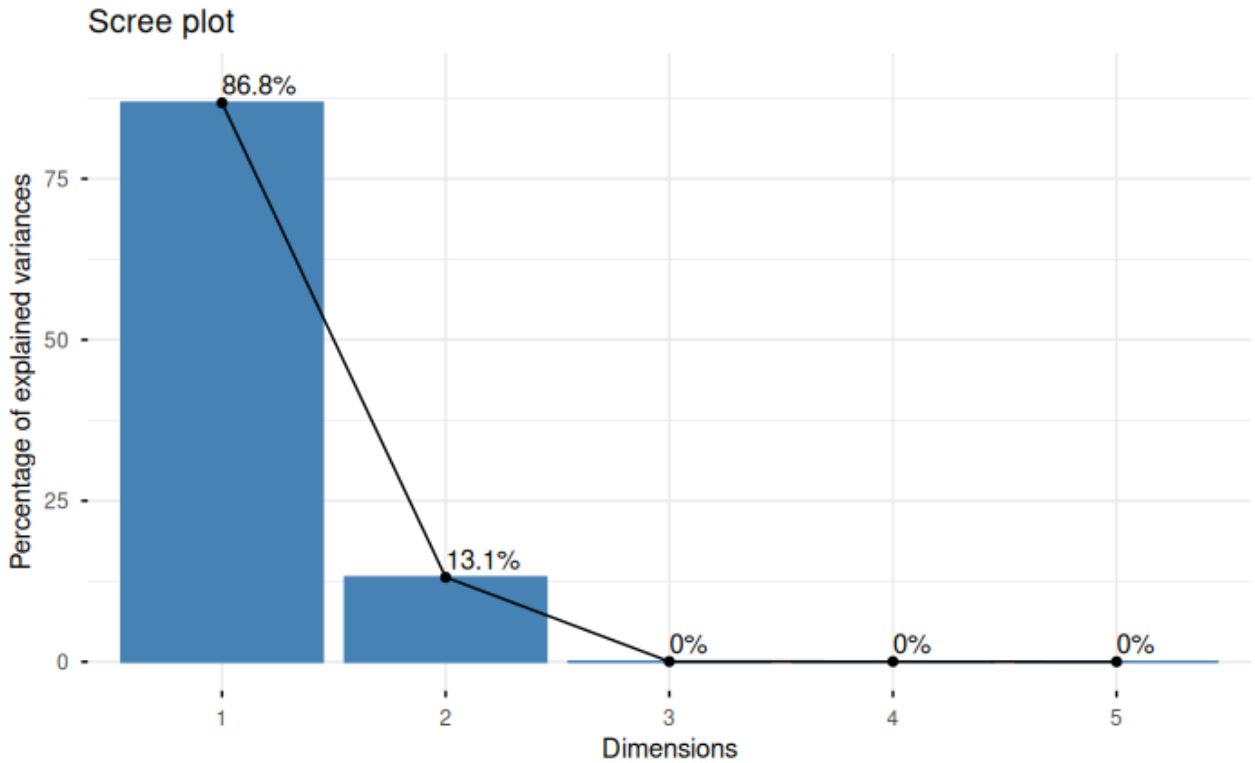

Figure: Percentage of explained variance per component

PCA Results: Eigenvalues

Eigenvalues and Variance

Components	Eigenvalue	Variance (%)	Cumulative (%)
comp.1	24790.62	86.80	86.80
comp.2	3746.35	13.11	99.91
comp.3	11.06	0.0387	99.95
comp.4	9.55	0.033	99.99
comp.5	2.526	0.008	100.00

Kaiser Criterion

According to the Kaiser criterion (Eigenvalue > 1), we retain 2 principal components. These two components explain **99.91%** of the total variance, indicating that the original data can be well summarized in a two-dimensional space.

AIMS African Institute for Mathematical Sciences SENEGAL

Variables Correlation

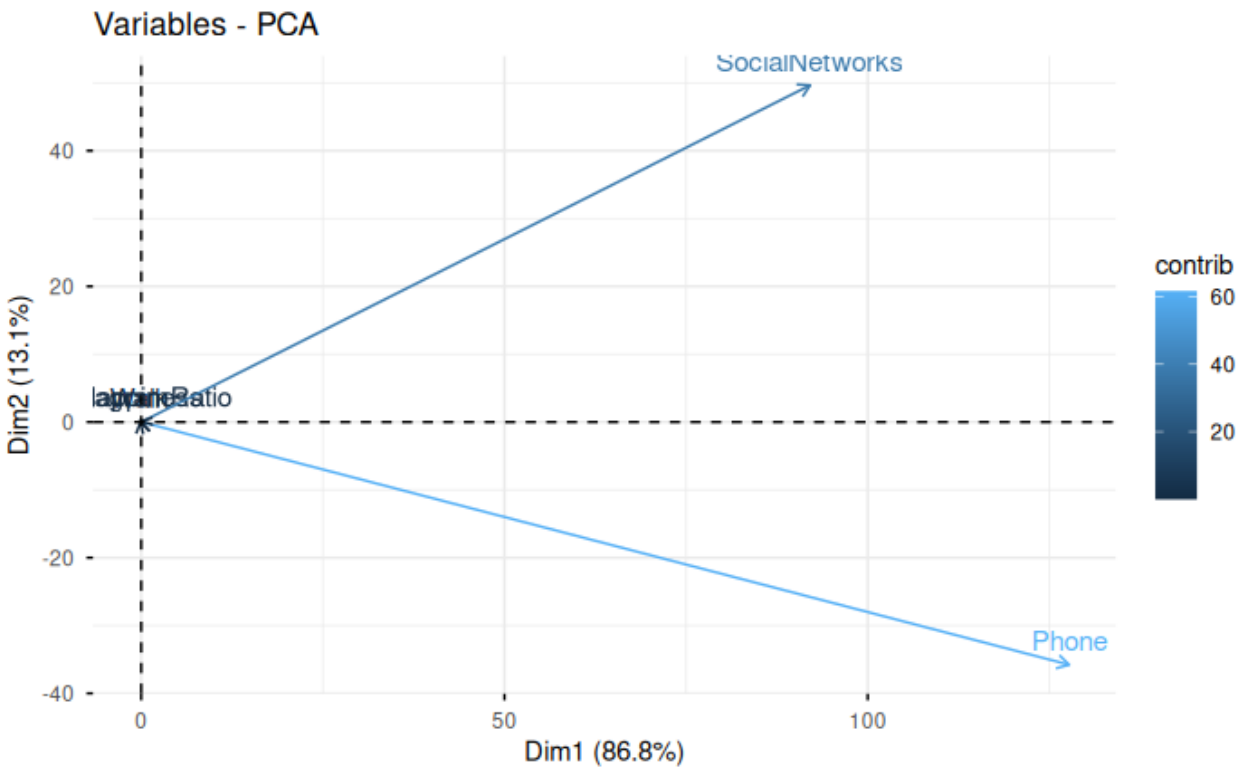

Figure: Projection of variables into the first two components

Interpretation of Components

Variable	Dim.1 (Contrib %)	Dim.2 (Contrib %)
Phone	48.58	1.09
SocialNetworks	47.23	2.03
Happiness	2.28	33.86
Walk	1.81	22.06
InstagramRatio	0.09	40.96

Table: Contribution of variables to the principal axes

Individuals Projection

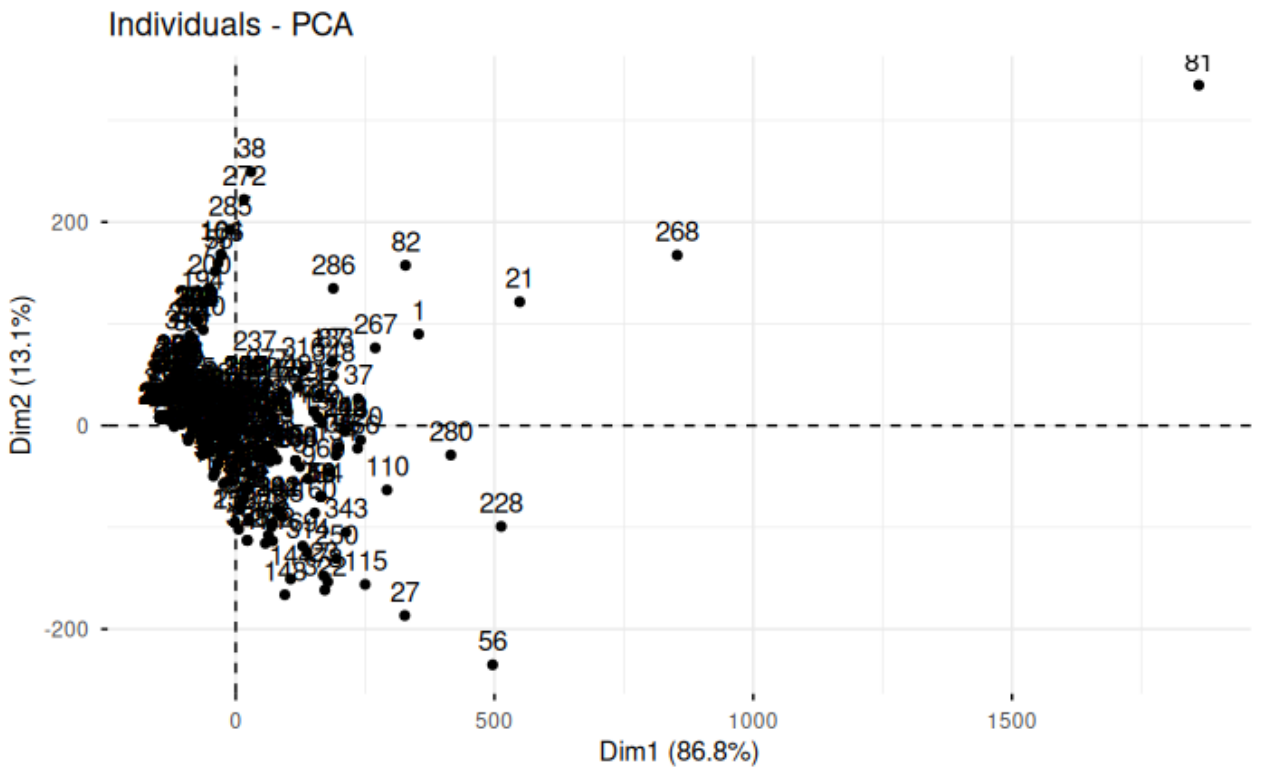

Figure: Distribution of 369 students in the principal component space

Regression Modeling: Summary I

Model Performance

- Linear model for **Phone usage**
- Adjusted $R^2 = 0.55$
- Model highly significant (Fisher p-value $< 2.2 \times 10^{-16}$)

Key Effects

- **SocialNetworks**: strong positive effect
- **Happiness**: significant negative effect
- **High Instagram usage**: lower Phone time
- Walk: no significant impact

AIMS African Institute for Mathematical Sciences SENEGAL

Regression Modeling: Summary II

Illustrative Result

At identical profiles, a low Instagram user spends **about 53 minutes more** on the phone.

Model Selection: Summary I

Selection Method

- Backward stepwise regression
- Removal of non-significant variables ($p > 0.05$)
- Model comparison using **AIC**

Final Selected Model

$$\text{Phone} = \beta_0 + \beta_1\,\text{SocialNetworks} + \beta_2\,\text{Happiness} + \beta_3\,\text{HighInstagram}$$

AIMS African Institute for Mathematical Sciences SENEGAL

Model Selection: Summary II

Why This Model?

- **Lower AIC**: 3761.47
- **Adjusted** R^2: 0.5487
- All variables statistically significant
- Fewer variables $\Rightarrow$ better interpretability

Final Predictions: Summary

Student Profile

- SocialNetworks = 30 min/day
- Happiness = 8 / 10

Key Result

For identical profiles:

- High Instagram user: **38.6 min/day**
- Low Instagram user: **78.0 min/day**
- Difference: **+39 minutes per day**

AIMS African Institute for Mathematical Sciences SENEGAL

Conclusion: Implications and Perspectives

Practical Recommendations

- Promote well-being to reduce screen time
- Raise awareness of social media overuse
- Encourage structured and conscious usage

Future Work

- Understand behavioral mechanisms of HighInstagram users
- Include longitudinal and behavioral data

AIMS African Institute for Mathematical Sciences SENEGAL

Thank you for your attention!

Any questions?

Reference: "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman.