In Silico Comparative Evaluation of Classical and Robust Dimension Reduction for Psychological Assessment

by

Anna Theresa Santiago

A thesis submitted in conformity with the requirements for the degree of Master of Science
Department of Public Health Sciences
University of Toronto

© Copyright by Anna Theresa Santiago 2018
In Silico Comparative Evaluation of Classical and Robust Dimension Reduction for Psychological Assessment

Anna Theresa Santiago

Master of Science

Department of Public Health Sciences
University of Toronto

2018

Abstract

The classic exploration of correlated multivariable psychological assessment data employs dimension reduction of the original $p$ variables to a lower $q$-dimensional space through principal component analysis (PCA). Standard criteria in the selection of the number of $q$ dimensions may be affected by non-normal distributions and atypical observations. Performance of PCA was compared to robust dimension reduction techniques including Grid Projection Pursuit and robust covariance estimation (ROBPCA) in the accurate identification of the predetermined number of $q$-dimensions according to selection criteria for the cumulative proportion of variation, Kaiser-Guttman eigenvalue rule, and quantitative localization of Cattell’s scree plot ‘elbow’ via segmented regression. Based on in silico benchmarking using various simulated multivariate conditions, ROBPCA was consistent in correctly identifying the true $q$-dimensions. PCA can be improved through the adjustment of values for the cumulative proportion of variation and eigenvalue thresholds. Grid Projection Pursuit was the lowest performer among the techniques under simulated conditions.
Acknowledgments

This study would not have been possible without the support of many individuals. I am indebted to my primary thesis supervisor Malcolm A. Binns for his leadership, insights on writing, and for illuminating me on various aspects of statistical analysis particularly segmented regression. I would like to express my gratitude to my co-supervisor Wendy Y. Lou for the valuable academic advice, research funding, and the windows of opportunity that she has opened for me. My special thanks are extended to the other thesis committee members including Deirdre Dawson, Sandra Gardner, Alex Kiss, and the oral defense chair Susan Bondy. Assistance provided by the staff of the Division of Biostatistics and the Graduate Office of the Dalla Lana School of Public Health was greatly appreciated. I also wish to thank staff, students, professors, and researchers whom I have interacted and worked with at the University of Toronto, Baycrest’s Rotman Research Institute, and the Ontario Neurodegenerative Disease Research Initiative for the many learning experiences. My special thanks to the university for the financial support I have received through the Drs. Jeanne Ikeda Douglas and Foch A. Douglas Graduate Student Endowment Fund, University of Toronto Fellowship for the Public Health Sciences, and the School of Graduate Studies Conference Grant. I wish to thank my family, Rona, Paolo, Jena, Michael, Jem, and my late parents Ihnas and Lory for the inspiration; also, my in-laws Art, Airie, Anne, John, and their families for their help and prayers. I dedicate this work to my husband Pius and my children Aidan, Adriana, and Adam who constitute my life’s purpose. Finally, I thank the Lord God Who is the $A$ and the $\Omega$—the Beginning and the End—from Whom all things were made and without Whom I am nothing.
# Table of Contents

ACKNOWLEDGMENTS ........................................................................................................ iii

TABLE OF CONTENTS .................................................................................................... iv

LIST OF TABLES ........................................................................................................ vi

LIST OF FIGURES ......................................................................................................... vii

CHAPTER 1 INTRODUCTION ............................................................................................ 1

1.1 DIMENSION REDUCTION IN PSYCHOLOGY .............................................................. 1

1.2 CLASSICAL PRINCIPAL COMPONENT ANALYSIS (PCA) .......................................... 2

1.2.1 Singular Value Decomposition (SVD) of the Data Matrix .................................... 2

1.2.2 Selection of the Number of Components to Retain ............................................. 3

1.2.3 Effect of Atypical Observations on PCA ............................................................ 5

1.3 ROBUST DIMENSION REDUCTION ......................................................................... 5

1.3.1 Robust Statistical Approaches to PCA ............................................................... 5

1.3.2 Projection Pursuit (PP) ....................................................................................... 6

1.3.3 Robust Covariance Estimation .......................................................................... 7

1.4 THESIS OBJECTIVES .............................................................................................. 10

1.5 THESIS CONTRIBUTION .......................................................................................... 10

CHAPTER 2 METHODOLOGY ......................................................................................... 11

2.1 CONTEXT: NEUROPSYCHOLOGICAL AND PSYCHOSOCIAL ASSESSMENT OF PATIENTS WITH ACQUIRED BRAIN INJURY (ABI) ...... 11

2.1.1 Classical PCA on Mean-Centered and Scaled Data ............................................. 15

2.1.2 Grid Projection Pursuit and ROBPCA on Real Data ........................................... 21

2.2 EMPIRICAL BENCHMARKING .................................................................................. 24

2.3 APPLIED COMPARATIVE EVALUATION OF CLASSICAL PCA, GRID PROJECTION PURSUIT, AND ROBPCA ........................................... 25

2.3.1 Cumulative Proportion of Variation Threshold ................................................ 26

2.3.2 Eigenvalue Threshold ......................................................................................... 27

2.3.3 Segmented Regression on the Eigenvalue Scree Plot ........................................... 28

2.4 SIMULATION OF MULTIVARIATE CORRELATED DATA ....................................... 28

CHAPTER 3 RESULTS ..................................................................................................... 33

3.1 PERFORMANCE BASED ON CUMULATIVE PROPORTION OF VARIATION θ ............................................................... 33

3.2 PERFORMANCE BASED ON EIGENVALUES D2 > 1 AND D3 < 1 ................................ 36
3.3  PERFORMANCE BASED ON SCREE PLOT BREAKPOINT $\Gamma = 3$ .............................................................................................................................................. 40

CHAPTER 4 DISCUSSION .......................................................................................................................................................................................... 45

4.1  REPLICABILITY OF DIMENSION REDUCTION SOLUTIONS IN REAL DATA.............................................................................................................. 45
4.2  SIMULATION STUDIES IN THE SELECTION OF Q DIMENSIONS FOR PCA .................................................................................................... 46
4.3  APPRAISAL OF BENCHMARKING RESULTS ............................................................................................................................................. 47

CHAPTER 5 CONCLUSION AND RECOMMENDATIONS ........................................................................................................................................ 49

REFERENCES ...................................................................................................................................................................................................... 50

APPENDIX TABLE A: DESCRIPTIVE STATISTICS OF ABI DATA SET ......................................................................................................................... 55
APPENDIX B: R SCRIPT FOR SIMULATED DATA SEQUENCE B ................................................................................................................................. 57
APPENDIX C: R SCRIPT FOR TECHNIQUE PERFORMANCE BASED ON EIGENVALUE THRESHOLDS FOR PCA ON DATA SEQUENCE B ................................................................................................................................. 66
List of Tables

Table 2.1 Classical correlation matrix for variables representing each assessment in the ABI data. ..........13

Table 2.2 Parameter specifications for assigned Poisson variables for each data sequence. ..........30

Table 2.3 Parameter specifications for assigned continuous variables for each data sequence. ..........30

Table 3.1 Technique performance according to cumulative proportion of variation θ=0.80 and θ values for 95% and 100% success. ..........34

Table 3.2 Results of technique performance based on the θ thresholds. ..........37

Table 3.3 ROBPCA tally of iterations based on number of resulting PCs according to data sequence. ..........41
List of Figures

Figure 2.1 Boxplots of data on variables for self-ordered pointing, locus of control, and general self-efficacy assessments. ..........12

Figure 2.2 Parallel boxplots of data on Shipley, PHQ, and MPAI assessments for individual variables representing each assessment. ..........12

Figure 2.3 Parallel boxplots of data on HVLT, D-KEFS, BRIEF-A, CHIP, and SF-36 assessments for individual variables according to increasing medians. ..........14

Figure 2.4 Cumulative proportion of variation plot for classical PCA on the ABI data set. ..........15

Figure 2.5 Scree plot for classical PCA on the ABI data set. The elbow may be considered at components 6, 8, or 12. Fourteen (14) components have eigenvalues ≥ 1. ..........16

Figure 2.6 Various types of outlying observations that may affect classical PCA. (from Varmuza and Filzmoser, 2009 p. 79). Point 1 is an orthogonal outlier, point 3 is a good leverage point, and point 2 is a bad leverage point. ..........17

Figure 2.7 Diagnostic plot for classical PCA showing orthogonal outliers. ..........19

Figure 2.8 Biplot for the first two components from the 12-dimensional PCA solution. ..........20

Figure 2.9 Cumulative proportion of variation plot for Grid Projection Pursuit and ROBPCA. ..........21

Figure 2.10 Scree plot for Grid PP and ROBPCA. ..........22

Figure 2.11 Diagnostic plots for Grid PP and ROBPCA. ..........22

Figure 2.12 Diagnostic plots for Grid PP and ROBPCA. ..........23

Figure 2.13 General methodological framework. ..........26

Figure 2.14 Representative histograms with overlaid density plots of simulated variables. ..........31

Figure 2.15 Scheme of multivariate data simulation leading to comparative evaluation of PCA, Grid PP, and ROBPCA. The process is looped 100 times for each of three multivariate data conditions. ..........32
**Figure 3.1** Sequence B histograms and empirical cumulative distribution plots for % success based on $\theta = 0.80$ (blue vertical line and value) and alternative cut-off $\theta$ (red vertical line and value) for 95% success.

**Figure 3.2** Trend of the cumulative proportion of variation values ($\theta$) for 95% and 100% successes according to dimension reduction technique across simulated data sequences.

**Figure 3.3** Boxplots of the distributions of $d_2$ and $d_3$ for each technique across data sequences.

**Figure 3.4** Success probability plots for the performance of the three techniques based on the Kaiser-Guttman eigenvalue $\geq 1$ rule for a 2-component solution using the simulated data set sequence 1. The range of observed eigenvalues (vertical dashed lines and values) under maximum success conditions are identified.

**Figure 3.5** Range plots of eigenvalues for maximum success probability.

**Figure 3.6** Comparison of success probabilities for $\gamma = 3$ as the component breakpoint.

**Figure 3.7** Scree plots with overlaid fitted regression models.

**Figure 3.8** MSE distribution for various $\gamma$.

**Figure 3.9** Trend of Mean $\Delta R^2$ for the segmented model at $\gamma = 3$ and a simple linear model.
Chapter 1
Introduction

Psychological research focused on the cognitive and behavioral effects of brain injury and disease involves the use of assessment tools to collect information. Inherent structural features of multivariate psychological assessment data include high-dimensionality and the presence of inter-correlated variables. These features are accommodated by computational techniques such as principal component analysis (PCA), which facilitate the exploration of underlying structural relationships and the extraction of dominant attributes in the data. The underlying multivariate data structure is primarily defined by the covariance matrix, which is considered to be “the key to multivariate statistics (Hampel et al., 1986).” PCA, as a classical multivariate technique, depends on adherence to statistical assumptions such as multivariate normality and the absence of influential observations. In the clinical setting, psychological tests are constructed to be meaningful yet tend to be non-normally distributed or skewed. It is also common to encounter observations which are atypical, particularly when data are collected from impaired human subjects such as those with brain injury. In the selection of important components in PCA, skewed distributions and anomalous observations may influence results. A pragmatic route is to consider robust alternatives to classical PCA when aiming for stable and correct outcomes when dealing with less than ideal, but more real-life data structures.

1.1 Dimension Reduction in Psychology

Batteries of psychological assessments are often used to measure complex human attributes such as “memory, attention, language, executive function, and spatial cognition (Maroof, 2012)” as well as understand the “psychology of the individual…whether in clinical, forensic, educational, counselling, health, coaching or occupational settings (Coaley, 2010).” Due to the multivariable nature of psychological assessments, data reduction techniques that simplify available information into a more manageable structure are often applied prior to formal analyses. Data reduction methods, including factor analysis and principal component analysis, ultimately assist in the enhanced understanding of neurologic conditions.

Exploratory factor analysis (EFA) is a common dimension reduction technique applied in psychology research which aims to obtain a reduced set of latent (or unobserved) factors from a
larger set of variables. EFA is similar to PCA in terms of the mathematical extraction of latent factors, such that statistical software for psychology research embeds PCA within factor analysis functions (Dugard et al., 2010; Jolliffe, 2002). The primary distinction between EFA and PCA is that EFA presumes that underlying attributes exist in the data and that these attributes are revealed through the linear transformation of the data resulting to latent factors. Furthermore, in EFA the original assessment variables are conveyed as a linear combination of latent factors, whereas in PCA, principal components are determined as linear combinations of the observed assessment variables (Schumacker, 2016).

1.2 Classical Principal Component Analysis (PCA)

Karl Pearson and Harold Hotelling were the pioneering proponents of PCA as a dimension reduction technique whereby the acquisition of a few linear combinations of the original multivariate data is achieved through the maximization of variance, hence resulting to a “parsimonious summarization” (Mardia et al., 1979) of the data with minimal information loss. The linear combinations are designated as principal components (PCs), “which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe, 2002).” A linear transformation of the original data set can be achieved through eigenvector decomposition of the original data matrix such as through singular value decomposition (SVD). This process is followed by the selection of a reduced number of PCs based on variance contribution as the basis of interestingness.

1.2.1 Singular Value Decomposition (SVD) of the Data Matrix

Let the data set be denoted as a matrix $X$ with $n \times p$ dimensions, whereby the number of participants is represented as $n$ rows, variables as $p$ number of columns and the $x_{i,j}$ matrix elements represent individual data points. To illustrate:

$$
X (n \times p) = \begin{bmatrix}
    x_{11} & \cdots & x_{1p} \\
    \cdots & \cdots & \cdots \\
    x_{n1} & \cdots & x_{np}
\end{bmatrix}.
$$

The data matrix $X$ is pre-processed through mean-centering and scaling. Mean-centering is achieved by subtracting the column-wise means from each column data point, while scaling involves the use of column-wise standard deviations as the divisor for each column data point to
attain standardization of each column to have zero mean and unit variance (Varmuza & Filzmoser, 2009; Abdi & Williams, 2010).

Classical PCA begins with the decomposition of $X$ into the product of three matrices:

$$X = T_0 \cdot D \cdot P^T,$$

whereby $T_0$ is an orthogonal $n \times n$ matrix containing the eigenvectors of $XX^T$ also known as the normalized scores, $D$ is a diagonal matrix of $p \times p$ dimensions containing singular values representing the standard deviations of the PCs, and $P^T$ is an orthogonal $p \times p$ matrix composed of the eigenvectors of $X^TX$ also referred to as the transposed loading matrix. Eigenvalues from the SVD of the standardized data matrix $X$ comprise the variance explained by each PC and are derived as the squared elements of the diagonal matrix $D$. The dimension reduction step from the original $p$ dimensions to the reduced $q$ dimensions involves the identification of the number of important PCs to retain.

### 1.2.2 Selection of the Number of Components to Retain

The predominant decision in dimension reduction is the choice of the number of $q$ components to retain from the original $p$ variables ($q < p$) that sufficiently and parsimoniously captures the essence of the data with minimal loss of information. There are ad hoc rules-of-thumb to support this decision using reference thresholds for cumulative proportions of variation and eigenvalues.

Retention of components that contribute to a cumulative percentage total of variation of at least 70% up to 90% is common practice in routine classical PCA (Jolliffe, 2002; Everitt & Hothorn, 2011). Maximization of variance is an aim of PCA so that preservation of as much of the total variation with the smallest possible value of $q$ is generally desired. While the 70-90% cut-off range is deemed sensible, adjusting the cutoff beyond this range may depend on the practical details of a specific data set. Jolliffe (2002) states that in the reduction of the original $p$-dimensions to reduced $m$-dimensions (which this thesis designates as $q$),

"...a value greater than 90% will be appropriate when one or two PCs represent very dominant and rather obvious sources of variation. Here the less obvious structures beyond these could be of interest, and to find them a cut-off higher than 90% may be necessary. Conversely, when $p$ is very large choosing $m$
corresponding to 70% may give an impractically large value of $m$ for further analyses. In such cases the threshold should be set somewhat lower.”

The Kaiser-Guttman rule of retaining components with variances or eigenvalues $\geq 1$ is another heuristic approach in the selection of $q$. This rule is used often applied to PCA involving the decomposition of correlation matrices and various covariance matrices such as those based on mean-centered and scaled data matrices. The rule is grounded on the assumption that if all variables $p$ are independent, then the PCs are the same as the variables in terms of having unit variance ($=1$). Therefore, PCs with variances $< 1$ are discarded on the basis that these possess less information than any of the original variables, whereas PCs with variances $> 1$ are retained since these capture most of the original information in the data (Jackson, 1993; Jolliffe, 2002). Furthermore, Jolliffe (2002) supplements that

“If the data set contains groups of variables having large within-group correlations, but small between group correlations, then there is one PC associated with each group whose variance is $> 1$, whereas any other PCs associated with the group have variances $< 1$. Thus, the rule will generally retain one, and only one, PC associated with each such group of variables, which seems to be a reasonable course of action for data of this type.”

A more subjective criterion in the selection of the number of important components is Cattell’s (1966) scree plot test to observe for the “elbow” or point in which the plot slope changes from “steep” on the left to “not steep” to the right. The scree plot is further described by Varmuza and Filzmoser (2009) to be similar to “the profile of a mountain: after a steep slope a more flat region appears that is built by fallen, deposited stones (called scree).” Cattell (1966) proposed that the point after the steep slope may be defined as the location of either the last component to be retained or excluded; where inclusion of the elbow point was preferred. Jolliffe (2002) mentioned that both inclusion and exclusion of the “elbow” in the retention of components were observed in practice.

Mardia et al. (1979) identified limitations in the use of either Kaiser’s rule or Cattell’s test whereby the former “tends to include too few components” and the latter is likely to “include too many components” and that a trade-off is commonplace. To reiterate Jolliffe (2002), these rules may hinge on the idiosyncrasies of any particular data set.
1.2.3 Effect of Atypical Observations on PCA

Deviations in a multivariate data set may include gross errors such as input mistakes or miscalculations, model distribution assumptions usually based on the central limit theorem, and unsuspected serial correlations that violate independence. Implementation of data quality control and pre-processing efforts to deal with gross errors do not guarantee that fitted models are free from the influence of atypical observations. In the case of multivariate methods in which marginal distributions of variables and correlation structures are also approximated, methods that can accommodate influential data points are relied upon (Hampel et al., 1986). Classical multivariate techniques that are based on normal-theory use test statistics which are functions of the sample covariance matrix. In the case of PCA, variance maximization and matrix decomposition are procedures which are sensitive to atypical data points that impair the estimation of PCs, thus compelling data analysts to explore robust options (Varmuza and Filzmoser, 2009).

1.3 Robust Dimension Reduction

1.3.1 Robust Statistical Approaches to PCA

Critchley (1985) asserts the need to “robustify” PCA since it is critically influenced by observations such as a sufficiently large outlier that can define a PC and provide artificial information. Robust statistics may be categorized under the realm of parametric statistics with accommodations for deviations from the standard statistical assumptions of normality. Robust statistics can incorporate additional parameters that supplement a given parametric model to approximately satisfy statistical assumptions. In the presence of mild deviations from parametric assumptions, such as in the case of high-quality data without outliers, the use of classical statistical methods is not catastrophic and robustness may be less of a concern. Hampel et al. (1986) exhorts that “… robustness is only one aspect of good data analysis, and that other aspects like model choice, interpretation of outliers, interpretation of the main results, and so on are at least equally important.”

A robust PCA, therefore, provides a dimension reduction solution that is not as affected by the presence of atypical observations. There are diverse robust PCA approaches that propose the use of alternative matrix decomposition methods or involve more stable measures of interestingness
such as robust estimates of the data variability (Xu & Caramanis, 2012). Two useful classes of technique that can tackle dimensionality reduction in a robust manner include projection pursuit and robust covariance estimation.

1.3.2 Projection Pursuit (PP)

The idea of projecting multivariate data onto low-dimensional space (1-2 dimensions) by maximizing a measure of usefulness called a projection index was introduced by Friedman and Tukey (1974). The original projection index presented by Friedman and Tukey is a function of the data that is a “product of a robust measure of scale (trimmed standard deviation) with a measure of clumpiness (a weighted count of the number of close pairs)” (Huber, 1985). A PP approach to dimension reduction involves the sequential estimation of PCs. If the projection index is the variance, then the result is equivalent to classical PCA (Croux et al., 2007). An appeal of PP is that it can assist in the visualization of patterns in high-dimensional data onto a humanly perceptible plane and can be applied in situations in which the number of variables is more numerous than observations \((p > n)\). In the simplest case where \(p=2\), then dimension reduction through PP involves the search for the one-dimensional vector that maximizes the projection index. For instance, if a PP method for PCA is performed for \(x_1, \ldots, x_n\) observations, \(p=2\) variables, and variance \((s^2)\) as the projection index, then the first PC is the unit vector \(a_1\) that is denoted as

\[
a_1 = \arg \max_{\|a\|=1} s^2(a^tx_1, \ldots, a^tx_n). \tag{1.2}
\]

PP in a robust sense maximizes the median absolute deviation (MAD) as the projection index where the MAD for a variable is given by

\[
MAD = \text{median}(|X_i - \text{median}(X)|) \tag{1.3}
\]

Various PP algorithms have been developed to investigate robust measures of usefulness or produce computationally efficient ways of searching for the best directions to project the data. This thesis employs the Grid projection pursuit (Grid PP) algorithm by Croux et al. (2007) in the acquisition of robust PCs via PP. Broadly, Grid PP sequentially searches for two-dimensional spaces that maximize the projection index within a unit circle. This technique has been demonstrated to capture data variability, be computationally efficient, and be applicable to data
in which the number of variables is greater than the number of observations \((p > n)\). Again using the example of \(p=2\) for \(x_1, \ldots, x_n\) observations and a projection index of interest \((S)\), the first PC is the vector defined by the coordinates of a point on the unit circle that maximizes the angle function \(\theta\) denoted by \(\theta \rightarrow S(\cos(\theta), \sin(\theta))\) over the interval \([-\frac{\pi}{2}, \frac{\pi}{2}]\). In the case of \(p>2\), the maximization of \(\theta\) is a pattern of 2-dimensional optimizations beginning with the identification of the first direction \(\hat{a}\), which corresponds to the variable with the largest \(S\). Prior to the search algorithm, the \(S\) values of all variables are arranged in decreasing order such that \(S_j = S(x_{1j}, \ldots, x_{nj})\) for \(j = 1, \ldots, p\). By denoting \(e_1 = (x_{11}, \ldots, x_{n1})\) as the first and \(e_p = (x_{1p}, \ldots, x_{np})\) as the last canonical basis vectors or directions representing PCs, it can be assumed that \(S(e_1) \geq S(e_2) \geq \cdots \geq S(e_p)\). These canonical basis vectors may also be assigned as the elements of the score matrix \(T_0\) from the SVD of the data matrix \(X\) (equation 1.1). As described by Croux et al. (2007),

“For \(j = 1, \ldots, p\)

(a) Maximize the objective function in the plane spanned by \(e_j\) and \(\hat{a}\), by a grid search of the function \(\theta \rightarrow S(\cos \theta \hat{a} + \sin \theta e_j)\), where the angle \(\theta\) is restricted to the interval \([-\frac{\pi}{(2^j)}, \frac{\pi}{(2^j)}]\). Denote \(\theta_0\) the angle where the maximum is attained over all grid points.

(b) Update \(\hat{a} \leftarrow \cos \theta_0 \hat{a} + \sin \theta_0 e_j\).”

After updating the first direction, the process is repeated over \(p-1\) planes such that the \(j\)th grid search updates the \(j\)th coordinate of \(\hat{a}\). The first few cycles are aimed at finding the best angle over which the data should be projected and the later cycles are meant to fine-tune the solution.

1.3.3 Robust Covariance Estimation

Replacement of the classical covariance matrix by a scatter matrix based on robust estimators showed promise except that certain estimators are either cumbersome to compute or have low efficiency (Hubert et al., 2008). A stable and robust estimator of multivariate location and scatter is the minimum covariance determinant (MCD) and is effectively computed by the FAST-MCD algorithm by Rousseeuw and Driessen (1999). The MCD estimates for location \((\hat{\mu}_{MCD})\) and scatter \((\hat{\Sigma}_{MCD})\) as defined by Hubert and Debruyne (2010) are as follows:

\[
\hat{\mu}_{MCD} = \frac{\sum_{i=1}^{n} w(d_i^2)x_i}{\sum_{i=1}^{n} w(d_i^2)} \quad (1.4)
\]
\[ \hat{\Sigma}_{MCD} = c_1 \frac{1}{n} \sum_{i=1}^{n} W(d_i^2) (x_i - \hat{\mu}_{MCD}) (x_i - \hat{\mu}_{MCD})^T \] (1.5)

where \( c_1 \) is a consistency factor described by Croux and Haesbroeck (1999) which is used as a multiplier to achieve Fisher consistency of \( \hat{\Sigma}_{MCD} \) under an elliptically symmetric \( p \)-dimensional normal distribution for a specified fraction of the data (outlying) that is discarded in the calculation of the MCD;

with the robust distance,

\[ d_i = \sqrt{\left( x - \hat{\mu}_0 \right)^T \hat{\Sigma}_0^{-1} \left( x - \hat{\mu}_0 \right)} \] (1.6)

which is the robust version of the multivariate Mahalanobis distance metric (distance of observation from the center of the multivariate data cloud); and a default weight function

\[ W(d^2) = I(d^2 \leq \chi^2_{p,0.975}) \] (1.7)

where \( \chi^2_{p,0.975} \) denotes the 0.975-quantile (\( \alpha \)-quantile) of the chi-squared distribution with \( p \)-dimensions as the degrees of freedom.

Broadly, the MCD is obtained through an iterative process involving calculation of the covariance matrix determinant (\( det \)) of several subsets of the data (\( h \)) with the aim of arriving at a subset with the smallest determinant. It is from this trimmed data where dimension reduction is performed. Calling to mind the determinant as the generalized variance, it is obtained as the product of the eigenvalues of the covariance matrix with values ranging from 0-1. A zero determinant is indicative of completely redundant variables such that the covariance matrix is defined to be singular. On the other hand, a determinant of 1 indicates complete orthogonality or absence of correlation among variables (Harlow, 2014).

In dimension reduction for psychological data where groups of variables could represent subtests of an assessment tool, it is ideal to have a determinant that is enough to indicate the presence of related yet distinct information. Whereas robust PCA using the MCD is highly efficient for moderate dimensions (up to 100), a drawback is that it is problematic to apply in situations wherein variables are substantially more numerous than observations\( (p > n) \). As noted by Hubert & Debruyne (2010), “the MCD estimator can only be computed when \( h > p \), otherwise
the covariance matrix of any \( h \)-subset will be singular…To avoid the curse of dimensionality it is however recommended that \( n > 5p \).

To accommodate very high-dimensional data, a hybrid technique combining projection pursuit and FAST-MCD was developed by Hubert et al. (2005). Robust PCA via the ROBPCA algorithm is initiated through projection pursuit as the pre-processing step. Data points are projected in many one-dimensional/univariate directions whereby on each direction “the univariate MCD estimator of location and scale is computed, and for every data point its standardized distance to that center is measured” (2008). The \( h \)-subset which is within an acceptable distance from the center (least outlying) is kept and its covariance matrix is used to arrive at PCs. The choice of \( q \) number of PCs to retain is decided upon and the data points are projected onto the \( q \)-dimensional subspace. From the \( q \)-dimensional subspace, the MCD scatter matrix and the resulting eigenvectors from this scatter matrix become the \( q \) robust PCs. The computational efficiency of ROBPCA and its ability to properly recognize outliers makes it an ideal tool for outlier detection aside from dimension reduction.
1.4 Thesis Objectives

In general, this thesis aims to compare performance of the classical PCA solution to Grid PP and ROBPCA in reducing the dimensionality of multivariate psychological assessment data. The specific objectives are:

• To compare the different aspects of dimension reduction including rule-of-thumb criteria for determining the number of important components according to cumulative percentage of total variance and eigenvalue thresholds and interpretation of components based on component loadings on real data;

• To evaluate the use of a two-segment regression model of the eigenvalue scree plot as a supplementary criterion for the selection of important components; and

• To compare the consistency of results when applied to simulated multivariate data with count and continuous variables between symmetric and skewed distributions.

1.5 Thesis Contribution

In this thesis, we examine the performance of PCA, Grid PP, and ROBPCA in the determination of $q$ according to standard thresholds. Subsequently, alternative cut-off values for the cumulative proportion of variation and eigenvalue rules-of-thumb are proposed. Quantitative localization of the scree plot elbow is compared across the dimension reduction techniques through the evaluation of fitted segmented regression models. Guided by a benchmarking framework, performance measures facilitate the comparative assessment of classical PCA, Grid PP, and ROBPCA when applied to varying multivariate data distributions.
Chapter 2
Methodology

2.1 Context: Neuropsychological and Psychosocial Assessment of Patients with Acquired Brain Injury (ABI)

The springboard for this *in silico* comparison of classical and robust dimension reduction methods is a data set of complete baseline neuropsychological and psychosocial assessment measures obtained from 84 ABI patients recruited for an occupational health study at Baycrest Health Sciences. The ABI data set includes 61 variables derived from 11 neuropsychological and psychosocial assessment tools (Appendix Table A), namely Shipley test, Self-Ordered Pointing, Hopkins Verbal Learning Test (HVLT), Delis-Kaplan Executive Function System (D-KEFS), Patient Health Questionnaire (PHQ), Behaviour Rating Inventory of Executive Function – Adult (BRIEF-A), Coping with Health Injuries and Problems (CHIP), Locus of Control, Mayo-Portland Adaptability Inventory (MPAI), General Self-Efficacy (GSE), and Short Form Health Survey (SF-36).

Univariate data visualization shows heterogeneity in both distribution and units of measurement for individual variables (Figures 2.1, 2.2, and 2.3). A mixture of symmetric, right-skewed, and left-skewed variables can be observed based on boxplots. As seen on Figure 2.3, variables representing subscales of the same assessment tool are arranged according to increasing medians. This visual format also reveals the variation of distributions from heavily right-skewed variables to those which are heavily left-skewed. Since skewness is a measure of symmetry and a depiction of normality, remedial measures involving transformation of notable asymmetrical variables are often implemented as part of data pre-processing. In this thesis, we harness the ABI data in its unadulterated form to serve as a unique environment to explore the performance of PCA, Grid PP, and ROBPCA. Of particular interest is how these three dimension reduction techniques hold up in the presence of continuous measures and scaled scores that typify psychological assessment data.
Naïve examination of the associations between representative variables with interesting data distributions was achieved through the construction of a correlation matrix based on the classical Pearson correlation coefficient (Table 2.1).

While there are more robust correlation coefficients that are better suited for non-normally distributed data, inspection of coefficients from the original data served as the motivation for the design of the correlation matrix in section 2.4. The variables `sopprer`, `preloc`, and `pregse` were included as single variable representatives for `Self-Ordered Pointing`, `Locus of Control`, and `GSE`. In relation to the assessments represented in Figure 2.2, `shipprer` for raw `Shipley` score, `gad7pre` for `PHQ` General Anxiety Disorder 7 measure, and `MPAI_Rad_pre` for `MPAI` adjustment subscale score were selected. Designated variables chosen to represent `HVLT`, `D-`
**KEFS, BRIEF-A, CHIP, and SF-36** include `hvtoprep` for total percentile score, `towerrv` for Tower rule violations per item ratio, `gecprep` for Global Executive Composite, `prechiii` for inconsistency index, and `presf36_rp` for Role-Physical survey score, respectively.

### Table 2.1 Classical correlation matrix for variables representing each assessment in the ABI data.

<table>
<thead>
<tr>
<th></th>
<th>sopprer</th>
<th>preloc</th>
<th>pregse</th>
<th>shipprer</th>
<th>gad7pre</th>
<th>MPAI_Rad_pre</th>
<th>hvtoprep</th>
<th>towerrv</th>
<th>gecprep</th>
<th>prechiii</th>
<th>presf36_rp</th>
</tr>
</thead>
<tbody>
<tr>
<td>sopprer</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>preloc</td>
<td>-0.016</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pregse</td>
<td>-0.142</td>
<td>0.502</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shipprer</td>
<td>-0.071</td>
<td>0.088</td>
<td>0.011</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gad7pre</td>
<td>-0.184</td>
<td>-0.270</td>
<td>-0.149</td>
<td>-0.227</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPAI_Rad_pre</td>
<td>0.020</td>
<td>-0.348</td>
<td>-0.243</td>
<td>-0.061</td>
<td>0.664</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hvtoprep</td>
<td>0.110</td>
<td>-0.101</td>
<td>-0.043</td>
<td>-0.311</td>
<td>-0.060</td>
<td>0.006</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>towerrv</td>
<td>-0.303</td>
<td>-0.019</td>
<td>-0.016</td>
<td>0.300</td>
<td>0.060</td>
<td>0.006</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gecprep</td>
<td>0.081</td>
<td>-0.235</td>
<td>-0.141</td>
<td>0.003</td>
<td>0.463</td>
<td>0.555</td>
<td>-0.081</td>
<td>-0.045</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>prechiii</td>
<td>-0.050</td>
<td>-0.024</td>
<td>0.118</td>
<td>-0.148</td>
<td>0.002</td>
<td>-0.102</td>
<td>-0.182</td>
<td>0.067</td>
<td>0.013</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>presf36_rp</td>
<td>0.002</td>
<td>0.172</td>
<td>0.079</td>
<td>-0.012</td>
<td>-0.201</td>
<td>-0.442</td>
<td>-0.151</td>
<td>0.182</td>
<td>-0.281</td>
<td>0.142</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 2.3 Parallel boxplots of data on HVLT, D-KEFS, BRIEF-A, CHIP, and SF-36 assessments for individual variables according to increasing medians.
2.1.1 Classical PCA on Mean-Centered and Scaled Data

Application of classical PCA on the ABI data set provides insight on the challenges posed on the investigator with regards to the choice of the number of important components to retain according to standard criteria. The primary goal in PCA’s dimension reduction decision is to keep the first few components that explain the majority of data variability. In practice, a cumulative proportion of variation of about more than 70% or up to 90% are considered substantial. Setting the threshold mid-way at 80% leads to 12 components as seen in Figure 2.4. On the combined use of Kaiser-Guttman’s eigenvalue $\geq 1$ rule and Cattell’s scree plot test in Figure 2.5, up to 14 components can be retained. A scree plot elbow may be visually assigned to lie at either component 6, 8 or 12, which are inconsistent with the Kaiser-Guttman rule. Moreover, Figure 2.5 necessitates a more objective approach to define the scree plot elbow point. In this situation, concurrent consideration of the cumulative proportion of variation, Kaiser-Guttman rule, and Cattell’s scree plot test presents a $q$ selection dilemma.

Figure 2.4 Cumulative proportion of variation plot for classical PCA on the ABI data set. Twelve (12) components contribute to about 80% of the variability in the data.
Diagnostic evaluation of results in terms of the presence of outlying observations with respect to the $q$-dimensional PCA space is focused on identification of atypical data points in the form of orthogonal outliers and leverage points. An approach is to project the data on the $q$-dimensional PCA space followed by the calculation of individual score and orthogonal distance values. Regular observations are points that lie within the $q$-dimensional PCA space with a distribution that can be visualized as an ellipse (Figure 2.6). Orthogonal outliers are points which lie outside the $q$-dimensional PCA space such as the point labeled 1 in Figure 2.6. Consequently, these points are not distinguishable from regular observations upon visual inspection of the projected data onto $q$ dimensions. Leverage points are observations that have very high score distances, such that these points are far from the center of the multivariate point cloud. Good leverage points, such as point 3 in Figure 2.6, lie within the PCA space and add stability to the $q$-dimensional solution since these observations contribute to the maximization of the variability of
the data. A bad leverage point is one that has large score and orthogonal distances and “levers” the PCA space resulting to shifting or destabilization.

Figure 2.6 Various types of outlying observations that may affect classical PCA. (from Varmuza and Filzmoser, 2009 p. 79). Point 1 is an orthogonal outlier, point 3 is a good leverage point, and point 2 is a bad leverage point.

The score distance \( (DS) \), similar to the Mahalanobis distance mentioned in section 1.3.3, is measured with respect to the center of the PCA space. Calculation of the \( DS \) for every \( i \)th data point can be achieved by following a modified version of the formula by Varmuza and Filzmoser (2009),

\[
DS_i = \left[ \sum_{k=1}^{q} \frac{t_{ik}^2}{v_k} \right]^{1/2},
\]  

(2.1)

where \( q \) is the number of components that form the PCA space, \( t_{ik} \) are the elements of the score matrix \( T \), and \( v_k \) is the variance of the \( k \)th component. Values of the score matrix \( T \) are determined as the product of the normalized score matrix \( T_0 \) and diagonal matrix of singular values \( D \) (equation 1.1), thus expressed as

\[
T = T_0 \cdot D.
\]

(2.2)

For a multivariate normal data set, the sum of squared score distances is approximately a chi-square distribution \( \chi^2_q \), with \( q \) degrees of freedom and a standard \( DS \) cut-off equal to the 97.5%
Quantile $\sqrt{\chi^2_{q,0.975}}$. Observations with $DS$ values above the cut-off are mapped as leverage points.

The orthogonal distance (OD) for every $i$th observation measures the perpendicular gap between the observation and the PCA space and is calculated by the following formula:

$$OD_i = \left\| x_i - P \cdot t_i^T \right\|$$  \hspace{1cm} (2.3)

where $x_i$ is the $i$th observation of the centered data matrix, $P$ is the loading matrix using $q$ PCs, and $t_i^T$ is the transposed score vector of the $i$th observation for $q$ PCs. An OD cut-off value suggested by Varmuza and Filzmoser (2009) is equal to

$$\left[ \text{median} \left( OD^{2/3} \right) + \text{MAD} \left( OD^{2/3} \right) \cdot z_{0.975} \right]^{2/3}$$  \hspace{1cm} (2.4)

where $z_{0.975}$ is the 97.5% quantile (=1.96) based on the z-distribution or standard normal.

A diagnostic plot or outlier map for PCA on the ABI data set involves plotting $OD_i$ versus $DS_i$ for all observations based on a particular $q$-dimensional solution. As an example for $q=12$, Figure 2.7 illustrates the absence of leverage points and presence of some orthogonal outliers, which are identified based on observation number. Even though bad leverage points are not present, the occurrence of orthogonal outliers could still potentially lever the dimension reduction results leading to a PCA space that is very different from the true $q$-dimensional PCA solution, particularly with respect to the component interpretation.

Interpretation of the resulting components relies on the examination of the PCA scores and loadings. The biplot in Figure 2.8 is a projection of the data based on a 2-dimensional coordinate system composed of the first two PCs. It can be observed that the biplot is a juxtaposition of two graphs: a scatterplot called the score plot and a plot of vectors radiating from the origin called the loading plot.
Figure 2.7 Diagnostic plot for classical PCA showing orthogonal outliers. Cut-off values for the score and orthogonal distances are demarcated as dashed lines. Rectangular areas above and to the right of the cut-off regions define various types of atypical observations.

In the score plot, each data point is projected according to its corresponding score values for components 1 and 2 from the score matrix $T$. Points on the score plot may be identified based on observation number if there is an interest in participant clustering with respect to the 2 components.

The loading plot includes vectors representing the original $p$ psychological assessment variables with coordinates corresponding to values from the loading matrix $P$ for components 1 and 2. The magnitude or length of a vector indicates the variance of the variable it represents and thus signifies the influence of that variable on the component axis. Correlation between two variables
can be observed with respect to the angle between two vectors, such that a smaller angle of separation is indicative of greater correlation.

Figure 2.8 therefore shows that variables from Shipley, HVLT and D-KEFS influence component 1. The inverse relationship between the variables on the positive side of the origin and the D-KEFs variables on the negative side can be further investigated through specific subject-matter contextualization. Component 2 on the vertical axis is influenced by variables from PHQ, BRIEF-A, and SF-36. As in component 1, the inverse relationships can also be examined in the light of psychological assessment and brain function. Other variables with lesser magnitudes on components 1 and 2 can be examined through biplots that explore the data with reference to a 2-dimensional coordinate system that involves other components of the $q$-dimensional PCA space.

Figure 2.8 Biplot for the first two components from the 12-dimensional PCA solution.
2.1.2 Grid Projection Pursuit and ROBPCA on Real Data

In the presence of skewed variables and orthogonal outliers, both Grid PP and ROBPCA were performed on the mean-centered and scaled data set for the purpose of comparison and to further motivate the thesis methodology. In the determination of important components based on a variance contribution of 80%, Grid PP suggests a 22-component solution which is markedly larger than classical PCA and ROBPCA’s proposed 4-component result (Figure 2.9).

![Cumulative proportion of variation plot for Grid Projection Pursuit and ROBPCA.](image)

Figure 2.9 Cumulative proportion of variation plot for Grid Projection Pursuit and ROBPCA.

Figure 2.10 depicts a different predicament in terms of the scree plot elbow and eigenvalue threshold criteria. The scree plot elbow for Grid PP can be located at component 9, while ROBPCA clearly shows an obvious pivot at component 4. The eigenvalue ≥ 1 threshold, Grid PP and ROBPCA present 31 and 10 components, respectively. The selection process for the reduced $q$-dimensional space is burdened by the different solutions provided by PCA, Grid PP, and ROBPCA on the ABI data set and using the same rule-of-thumb criteria.
Upon diagnostic evaluation of Grid PP and ROBPCA based on the $q$-dimensional solutions according to 80% cumulative proportion of variation, like classical PCA, both techniques present orthogonal outliers and do not show any leverage points (Figure 2.11). However, the individual orthogonal outliers are different for Grid PP and ROBPCA, which may be attributed to the 22-dimensional space covered by Grid PP compared to the 4-dimensional space for ROBPCA.
To further accentuate the impact of the choice of $q$ dimensions, biplot interpretation of the components resulting from Grid PP and ROBPCA are quite different. As shown in Figure 2.12, the coordinate plane for components 1 and 2 in Grid PP is very dissimilar to the plane occupied by components 1 and 2 in ROBPCA and this distinction is noticeable based on the spread of the scores as well as orientation of loading vectors. Component 1 for both Grid PP and ROBPCA show heavy loadings for the D-KEFS variables. In Grid PP, component 2 is dominated by HVLT variables whereas in ROBPCA, this axis is governed by variables under SF-36, BRIEF-A, and PHQ.

For component interpretation, there is a similarity in the groups of psychological assessments that dominate components 1 and 2 in ROBPCA and classical PCA. However, close inspection of ROBPCA and classical PCA biplots reveal differences in the direction of vectors. This could be attributed to the differences in the orientation of the two-dimensional plane spanned by components 1 and 2 when we consider $q=4$ for ROBPCA and $q=12$ for classical PCA.

![Figure 2.12 Diagnostic plots for Grid PP and ROBPCA.](image)

Given the dimension reduction quandary in this specific psychological assessment battery, *in silico* experimentation was executed to compare the performance of classical PCA, Grid PP, and ROBPCA under various multivariate data parameters.
2.2 Empirical Benchmarking

The underlying theoretical framework of this thesis is analogous to Hothorn et al.’s (2005) framework for comparing the performance of competing statistical learning algorithms in solving the same research problem. The framework compares candidate algorithms aimed at solving a learning sample of \( n \) observations denoted as \( L = \{z_1, ..., z_n\} \). Each candidate algorithm is assumed to have two general steps composed of an initial model fitting that generates a function \( a(\cdot | L) \) followed by the computation of objects of interest from the fitted model. A problem-specific performance measure denoted by the function \( p(a, L) \) is used to assess \( a(\cdot | L) \) in each of the candidate algorithms.

The thesis methodology is a benchmarking experiment that primarily compares the performance of classical PCA, Grid PP, and ROBPCA in presenting with the correct \( q \)-dimensional solution given various data distributions. Specific performance measures for the criteria used in determining \( q \) are based on success probabilities in relation to statistical power. Akin to null hypothesis testing whereby the power of a statistical test measures the probability that the test will reject the false null hypothesis for various possible values of the parameter of interest, let the power function of a statistical test \( \delta \) be denoted as \( \pi(\Phi|\delta) \) for each value of the parameter of interest \( \Phi \) that is an element of the parameter space \( \Omega \). The linkage between the benchmarking performance measure function \( p(a, L) \) with the power function can be expressed through the following relationship:

\[
p(a, L) = \pi(\Phi|\delta) = \Pr(X \in C|\Phi) \text{ for } \Phi \in \Omega,
\]

(2.1)

Such that:

- \( \Phi \) denotes the true \( q \)-dimensional solution;
- \( \delta \) denotes the criterion for the selection of important components in dimension reduction;
- \( X \) represents the value of the object of interest (e.g. eigenvalue); and
- \( C \) denotes the set of values that fall within the rule-of-thumb cut-off or threshold for a specific criterion.

The indicator function of \( C \), denoted by \( I_C(\cdot) \) is defined by

\[
I_C(X) = \begin{cases} 
1 & \text{if } X \in C \\
0 & \text{if } X \notin C.
\end{cases}
\]

(2.2)
Let the indicator function $I_{\theta}(X_i)$ for the $i$th iteration be a random variable $Y_i = I_{C}(X_i)$ such that the sequence $Y_1, Y_2, ..., Y_{100}$ are independent and identically distributed (i.i.d.) $n=100$ Bernoulli trials with parameter $p$ whereby $\Pr(Y = 1) = p$ and $\Pr(Y = 0) = 1 - p$ for some $p \in [0, 1]$. Thus the sum $Y = Y_1 + Y_2 + \cdots + Y_{100}$ is used as the criterion-specific success probability for a dimension reduction technique that has a binomial distribution with parameters $n$ and $p$.

$$\text{Success Probability} = Y = Y_1 + Y_2 + \cdots + Y_{100} \quad (2.3)$$

### 2.3 Applied Comparative Evaluation of Classical PCA, Grid Projection Pursuit, and ROBPCA

The methodological groundwork of this thesis is the object-oriented programming framework for robust multivariate analysis by Todorov and Filzmoser (2009) embedded in the rr cov package entitled Scalable Robust Estimators with High Breakdown Point version 1.4-3 and executed in RStudio version 1.0.136 (2016). This comprehensive robust multivariate package for the R programming language provides the ideal environment to compare classical and robust multivariate statistical analyses driven by the following stated goals:

1. To provide the end-user with a flexible and easy access to newly developed robust methods for multivariate data analysis;
2. To allow the programming statisticians an extension by developing, implementing and testing new methods with minimum effort, and
3. To guarantee the original developers and maintainer of the packages a high level of maintainability (Todorov and Filzmoser, 2009).

As stated in sections 1.2 and 1.3, the reduction of the original $p$ variables into $q$ linear combinations while preserving much of the data variability was achieved through the calculation of eigenvalues and eigenvectors from data matrix decomposition. This thesis designed a methodological pipeline (Figure 2.13) that facilitated the application of an empirical benchmarking experiment for specific measurable criteria $\delta$ corresponding to the cumulative proportion of variation (section 2.3.1), eigenvalue (section 2.3.2), and scree plot elbow localization by segmented regression (section 2.3.3).
In section 2.4 we present the construction of three kinds of simulated data sets with various parameter attributes. Labeled as sequence 1, 2, and 3, the simulated data sets were designed to depict symmetric, skewed, and heterogeneous multivariate data, respectively. Also in section 2.4, a common correlation matrix for all sequences was constructed such that two intercorrelated groups result to a true $q$-dimensional solution $Q=2$. PCA, Grid PP, and ROBPCA were then evaluated in terms of consistency in producing the correct solution under various data conditions.

The greater part of the comparative evaluation used on-demand calculations and summaries on the cumulative proportion of variation, eigenvalues, and component loadings extracted from the objects pca, grid, and hub through the R slot notations @ and $. Base R accessors such as [] for subsets and [[]] for single elements were also used to extract information from objects for further calculation or summarization. Annotated R code can be found in Appendix B for reference. Performance of each dimension reduction technique was evaluated after 100 iterations to determine consistency.

![Figure 2.13](image)

**Figure 2.13** General methodological framework: (1) Parameters of the marginal distributions of variables and correlation matrix observed from baseline psychological assessment of acquired brain injury patients served as the setting for the generation of simulated data. (2) Simulation of multivariate data assumed Poisson and continuous variables with various parameter values to depict multivariate normal and non-normal conditions. (3) Computational implementation is through a loop integrating simulated data generation and extraction of dimension reduction solutions. (4) Performance of each dimension reduction technique was compared based on success probabilities according to rule-of-thumb thresholds and consistency of solutions.

### 2.3.1 Cumulative Proportion of Variation Threshold

Using 80% ($\theta=0.80$) as the rule-of-thumb cumulative proportion threshold for retaining important components, the success probability of attaining the true $q$-dimensional solution $Q$ in
100 iterations was evaluated based on the distribution of $X$ which represents the value of the cumulative proportion of variation for $q$. Let the performance measure function $p(a, L)$ be expressed as

$$
\pi(\Phi|\delta) = Pr(q = Q|\theta = 0.80) = Pr(X \geq 0.80)
$$

Such that for one iteration, the indicator function of $\theta$ given by $I_\theta(X)$ is defined by

$$
I_\theta(X) = \begin{cases} 
1 & \text{if } X \geq 0.80 \\
0 & \text{if } X < 0.80.
\end{cases}
$$

Following $Y_i = I_c(X_i)$ as stipulated in section 2.2, the success probability for 100 iterations follows equation 2.3.

Graphical summarization was achieved through histograms and empirical cumulative distribution function plots. Based on the empirical cumulative distribution, the success probability for a two-component solution based on the 0.80 cut-off was determined as 1-error probability. Complementary to this criterion is the calculation of thresholds where 95% success for $q=Q$ was achieved.

### 2.3.2 Eigenvalue Threshold

Evaluation of the success probability for the true $q$-dimensional solution $Q$ using the Kaiser-Guttman eigenvalue $\geq 1$ rule was based on the eigenvalues $d$ for the last component to retain under $Q (d_q)$ and next component $(d_{q+1})$ such that

$$
Pr(q = Q) = Pr[(d_q \geq 1) \cap (d_{q+1} < 1)].
$$

This probability ensures that success is defined based on exactly $Q$ components, no more, no less. The distributions of values for $d_q$ and $d_{q+1}$ for each dimension reduction technique across various multivariate data conditions were summarized through boxplots. The ranges of eigenvalues to achieve maximum success for each simulated data condition were determined and visualized through density plots and range plots. Annotated R code for the evaluation of technique performance based on eigenvalue thresholds are presented in Appendix C.
2.3.3 Segmented Regression on the Eigenvalue Scree Plot

A quantifiable means to localize Cattell’s scree plot “elbow” was performed by fitting two-segment regression models for the eigenvalue as a function of the components according to the following model:

\[ Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i - \gamma) \cdot I(X_i - \gamma) + \epsilon_i \]  \hspace{1cm} (2.7)

with the following notational definitions:

- \( Y_i \) as the eigenvalue of the \( i \)th component;
- \( X_i \) as the \( i \)th component;
- \( \gamma \) as the assumed elbow or component breakpoint where \( \gamma \in \{2, 3, 4, 5\} \);
- \( \beta_0 \) as the intercept at \( X_i = 0 \);
- \( \beta_1, \beta_2 \) as the slopes before and after the breakpoint \( \gamma \);
- \( I(\cdot) \) as the indicator function for \( X_i - \gamma \), such that
  \[ I(X_i - \gamma) = \begin{cases} 0 & \text{if } X_i < \gamma \\ 1 & \text{if } X_i \geq \gamma \end{cases} \]
  ; and
- \( \epsilon_i \sim N(0, \sigma^2) \) as the error term.

For the true \( q \)-dimensional solution, the ideal component breakpoint \( \gamma \) should result to a minimum mean squared error (MSE). The proportion of iterations in which the ideal \( \gamma \) presents with the smallest MSE value was used to evaluate the consistency of classical PCA, Grid projection pursuit, and ROBPCA in producing \( Q \). Coupled with this metric is the comparison of segmented models under the ideal \( \gamma \) based on the adjusted coefficient of determination \( R^2 \). The incremental change in \( R^2 \) obtained between a simple linear model of the scree plot and a segmented model with the ideal \( \gamma \) was calculated using \( \Delta R^2 \) based on the expression:

\[ \Delta R^2 = R_{adj}^2 - R_{adj}^2 \hspace{1cm} (2.8) \]

wherein \( R_{adj}^2 \) and \( R_{adj}^2 \) represent the adjusted \( R^2 \) values for the segmented and simple linear models, respectively.

2.4 Simulation of Multivariate Correlated Data

Implementation of multivariate correlated data simulation according to a specified correlation matrix and variable parameters was achieved through the PoisNonNor version 1.1 package in
RStudio. The simulation process commenced with the construction of a correlation matrix $R$ for 6 hypothetical variables intended to produce two interrelated groups, namely variables 1, 3, 5 and variables 2, 4, 6, such that $Q=2$. Pearson correlation coefficients used to design $R$ mimic the range of bivariate correlations observed in the ABI data as seen in Table 2.1 for the paired variables $shipper—presf36_rp$ ($r = -0.01$), $towerrv—gecprep$ ($r = 0.05$), and $preloc—pregse$ ($r = 0.50$). The correlation matrix was constructed to be positive definite and is structured as follows:

$$R = \begin{pmatrix}
1 & -0.01 & 0.50 & -0.01 & 0.50 & -0.01 \\
-0.01 & 1 & 0.05 & 0.50 & -0.01 & 0.50 \\
0.50 & 0.05 & 1 & -0.01 & 0.50 & -0.01 \\
-0.01 & 0.50 & -0.01 & 1 & 0.05 & 0.50 \\
0.50 & -0.01 & 0.50 & 0.05 & 1 & -0.01 \\
-0.01 & 0.50 & -0.01 & 0.50 & -0.01 & 1
\end{pmatrix}.$$  

Variables 1-3 were designed to be Poisson distributed, while variables 4-6 were assigned as continuous variables. Three kinds of multivariate correlated data sets labeled as bell-shaped (B), skewed (Sk), and heterogenous (H) were simulated according to various parameters (Tables 2.2 and 2.3). Representative univariate distribution plots are displayed in Figure 2.14. The intention of the simulation process is to achieve authenticity with respect to the ABI data such that Poisson-distributed variables that mimic score data are combined with different continuous data.

Each simulated data sequence was of $100 \times 6$ dimensions or 100 rows and 6 columns; each of which was subjected to classical PCA, Grid PP, and ROBPCA. A total of 100 iterations of simulation and dimension reduction were performed for each data sequence (Appendix B). Simulation runtime for 100 iterations per data sequence is \(~465\) seconds (\(~8\ minutes) on a Dell 64-bit Windows 7 Professional, Intel® Core™ i5-4460S CPU @ 2.90 GHz, 8-GB RAM unit. Figure 2.15 summarizes the methodological pipeline of this thesis from data simulation to comparative evaluation of the dimension reduction techniques of interest.
Table 2.2 Parameter specifications for assigned Poisson variables for each data sequence.

<table>
<thead>
<tr>
<th>Data Sequence</th>
<th>λ for Poisson Variables 1-3</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Var1</td>
<td>Var2</td>
<td>Var3</td>
<td></td>
</tr>
<tr>
<td>Bell-Shaped (B)</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Skewed (Sk)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Heterogeneous (H)</td>
<td>0.8</td>
<td>1</td>
<td>1.2</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3 Parameter specifications for assigned continuous variables for each data sequence.

<table>
<thead>
<tr>
<th>Data Sequence</th>
<th>Parameters</th>
<th>Continuous Variables 4-6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Var4</td>
</tr>
<tr>
<td>B</td>
<td>Mean (µ)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Variance (σ²)</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Skewness (γ₁)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Kurtosis (γ²)</td>
<td>3</td>
</tr>
<tr>
<td>Sk</td>
<td>Mean (µ)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Variance (σ²)</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>Skewness (γ₁)</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>Kurtosis (γ²)</td>
<td>3</td>
</tr>
<tr>
<td>H</td>
<td>Mean (µ)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Variance (σ²)</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Skewness (γ₁)</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Kurtosis (γ²)</td>
<td>4</td>
</tr>
</tbody>
</table>
Figure 2.14 Representative histograms of simulated variables.
Figure 2.15 Scheme of multivariate data simulation leading to comparative evaluation of PCA, Grid PP, and ROBPCA. The process is looped 100 times for each of three multivariate data conditions.
Chapter 3
Results

As a recapitulation, the dimension reduction techniques of interest were evaluated based on the successful presentation of the intrinsic dimensionality $Q=2$ across three simulated data sequences representing bell-shaped (B), skewed (Sk), and heterogenous (H) multivariate data. The correlation matrix designed in Section 2.4 implanted 2 groups of inter-correlated variables that defines $Q=2$. Inspection of loadings of the first two components for the first five iterations per technique showed similar values for variables within a correlated group, but the range of values vary between iterations. Performance of PCA, Grid PP, and ROBPCA was measured in accordance to the proportion of iterations in which the following were true: cumulative proportion of variation of the 2nd component was at least $\theta = 0.80$; eigenvalues of the 2nd and 3rd components were $d_2 \geq 1$ and $d_3 < 1$, respectively; and the segmented regression model where the breakpoint $\gamma = 3$ presented with the minimum MSE and adjusted $R^2$ indicating a better fit than a simple linear model (large $\Delta R^2$). The proportion of successful iterations out of a total of 100 was interpreted as the probability of success.

3.1 Performance Based on Cumulative Proportion of Variation $\theta$

Success based on the variance contribution of at least 0.80 for $Q=2$ was exclusive to ROBPCA. Neither PCA nor Grid PP was able to achieve a minimum of $\theta = 0.80$ for 100 iterations in each simulated data sequence. Table 3.1 columns 4 and 5 show that for PCA and Grid PP, the alternative $\theta$ values for achieving 95% and 100% successes were below the minimum threshold of at least 0.70 (or 70% cumulative proportion of variation) stated in literature. The results in Table 3.1 are summarized from empirical cumulative distribution plots that indicate the error probability based on the 80% threshold and the resulting cumulative proportion of variation threshold for 95% success. As seen in Figure 3.1 for the bell-shaped simulated data sequence, performance is gauged in terms of the frequency of occurrence of at least 0.80 (or 80%) variance contribution of the second component. The empirical cumulative distribution function (ecdf) plots in Figure 3.1 demonstrate that for PCA and Grid PP, two components did not capture at least $\theta = 0.80$ (blue vertical line) cumulative proportion of variation under simulated data sequence B and this was recorded as 0.00% success in Table 3.1. For ROBPCA, 74.23% indicate the error probability and therefore a 25.77% success was recorded under data sequence.
B. The $\theta$ value for 5% error probability or 95% success was determined for each technique based on the $ecdf$ plot, while the $\theta$ value for 100% success was taken from the minimum cumulative proportion of variation observed.

**Table 3.1** Technique performance according to cumulative proportion of variation $\theta=0.80$ and $\theta$ values for 95% and 100% success.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Simulated Data Sequence</th>
<th>Success Probability at $\theta = 0.80$ for 2 PCs</th>
<th>$\theta$ at 95% Success Probability for 2 PCs</th>
<th>$\theta$ at Maximum Success Probability for 2 PCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>B</td>
<td>0.00</td>
<td>0.6204</td>
<td>0.6096</td>
</tr>
<tr>
<td></td>
<td>Sk</td>
<td>0.00</td>
<td>0.6143</td>
<td>0.6047</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>0.00</td>
<td>0.6296</td>
<td>0.6149</td>
</tr>
<tr>
<td>Grid PP</td>
<td>B</td>
<td>0.00</td>
<td>0.5402</td>
<td>0.4872</td>
</tr>
<tr>
<td></td>
<td>Sk</td>
<td>0.00</td>
<td>0.4768</td>
<td>0.4640</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>0.00</td>
<td>0.4862</td>
<td>0.4657</td>
</tr>
<tr>
<td>ROBPCA</td>
<td>B</td>
<td>25.77</td>
<td>0.7161</td>
<td>0.6768</td>
</tr>
<tr>
<td></td>
<td>Sk</td>
<td>10.28</td>
<td>0.7180</td>
<td>0.7087</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>33.16</td>
<td>0.7015</td>
<td>0.6795</td>
</tr>
</tbody>
</table>
These results imply that PCA and Grid PP tend to overestimate the true value of $q$ when $\theta = 0.80$ under simulated data conditions indicating more PCs than targeted. As a consequence, these techniques promote the inclusion of more PCs than necessary when standard variance contribution cut-offs are used. Consequently, to enhance the performance of PCA, decreasing the threshold to $\theta=0.60$ (60% cumulative proportion of variation) could result to maximum success probability for $Q=2$ even under the most asymmetrical data sequence.

In Figure 3.2, it is apparent that the true $Q=2$ solution was best explained by ROBPCA followed by PCA and least by Grid PP. Based on ROBPCA’s performance, 95% success probability can be achieved at $\theta=0.70$ under various simulated data settings.
3.2 Performance Based on Eigenvalues $d_2 \geq 1$ and $d_3 < 1$

PCA outperformed Grid PP and ROBPCA in the determination of the true 2-component solution based on the Kaiser-Guttman eigenvalue $\geq 1$ rule. In the experiment, PCA was 100% successful in identifying $Q = 2$ for all iterations across all data sequences whereby exactly 2 components have eigenvalues $\geq 1$ (Table 3.2). For PCA as seen in Figure 3.3, it is clear that the eigenvalues of the second component $d_2$ were greater than 1 and that the eigenvalues of the third component $d_3$ were below 1; thus, $Q=2$ was achieved under this rule.

ROBPCA was the second best performer, with unsuccessful iterations for the more asymmetric simulated data. According to the distributions of $d_2$ and $d_3$ for ROBPCA in Figure 3.3, we observe that the eigenvalues for the second component $d_2$ fell below the $d = 1$ cut-off under the Sk and H data conditions. These results are reflected in Table 3.2 as 99% success probabilities under skewed and heterogeneous conditions for ROBPCA.
Grid PP was the least successful in correctly identifying \( Q=2 \) for all data sequences, particularly in the Sk data sequence wherein none of the iterations produced the expected result. Even under more symmetric data conditions, the technique led to 44% success. It is noteworthy to mention that 5% success was achieved when the data sequence was heterogeneous. In general, Grid PP yielded high eigenvalues for both \( d_2 \) and \( d_3 \) compared to PCA and ROBPCA and this can be visualized in Figures 3.3 and 3.5.

The ranges of eigenvalues for maximum success are summarized in Table 3.2 column 4 to determine possible alternative eigenvalue cut-offs for each of the techniques. It is of interest to look for flexibility in the eigenvalue threshold that would yield intrinsic dimensionality and work consistently with other \( q \)-selection rules-of-thumb in contrast to a stringent single-value cut-off (Figure 3.4). Accordingly, eigenvalue thresholds that are greater than 1 may be corrective for overestimation of \( q \). As seen in Figure 3.5, increasing the eigenvalue cut-off to 1.5 for classical PCA may be considered to accommodate asymmetric and heterogeneous data. For Grid PP, the eigenvalue threshold may be adjusted to at least 2.0. However, an eigenvalue cut-off=2 for Grid PP may still overestimate \( q \) under asymmetric distributions when referring to Figure 3.3 wherein the third component yielded \( d_3 \approx 2 \) in some iterations for the Sk and H sequences. The Kaiser-Guttman rule can still be applied to ROBPCA under different data environments.

### Table 3.2 Results of technique performance based on the \( d \) thresholds.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Simulated Data Sequence</th>
<th>Success Probability for 2 PCs ((d_2 \geq 1 \text{ and } d_3 &lt; 1))</th>
<th>( d ) for Maximum Success [range]</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>B</td>
<td>100.00</td>
<td>([0.90, 1.42])</td>
</tr>
<tr>
<td></td>
<td>Sk</td>
<td>100.00</td>
<td>([0.82, 1.46])</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>100.00</td>
<td>([0.86, 1.52])</td>
</tr>
<tr>
<td>Grid PP</td>
<td>B</td>
<td>44.00</td>
<td>([1.53, 1.58])</td>
</tr>
<tr>
<td></td>
<td>Sk</td>
<td>0.00</td>
<td>([2.13, 2.20])</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>5.00</td>
<td>([2.13, 2.13])</td>
</tr>
<tr>
<td>ROBPCA</td>
<td>B</td>
<td>100.00</td>
<td>([0.90, 1.14])</td>
</tr>
<tr>
<td></td>
<td>Sk</td>
<td>99.00</td>
<td>([0.88, 1.06])</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>99.00</td>
<td>([0.88, 1.00])</td>
</tr>
</tbody>
</table>
Figure 3.3 Boxplots of the distributions of $d_2$ and $d_3$ for each technique across data sequences.

Figure 3.4 Success probability plots for the performance of the three techniques based on the Kaiser-Guttman eigenvalue ≥ 1 rule for a 2-component solution using the simulated data set sequence 1. The range of observed eigenvalues (vertical dashed lines and values) under maximum success conditions are identified.
Figure 3.5 Range plots of eigenvalues for maximum success probability.
3.3 Performance Based on Scree Plot Breakpoint $\gamma = 3$

Under the intrinsic dimensionality $Q=2$, Cattell’s scree plot elbow is expected to lie on the third component which is naturally interpreted as the breakpoint $\gamma = 3$. In this case, a segmented model comprising of a marked difference in slopes on either side of $\gamma = 3$ is a better fit for the scree plot when $Q$ is true. PCA presented with the highest success across all data sequences in identifying the segmented model with scree plot breakpoint at $\gamma = 3$ as the best fit for a 2-component solution (Figure 3.6). The success trend for PCA increased as the multivariate data condition became more heterogeneous. ROBPCA follows as the next best technique, but decreased in performance under data sequences Sk and H. Since ROBPCA involves an initial projection pursuit process and acquires subsets of the original data with the minimum covariance determinant (MCD), ROBPCA outcomes resulted in either 3-PCs or 4-PCs prior to dimension reduction (Table 3.3). The success performance of ROBPCA was determined for iterations with a 4-PC result, since a segmented model with breakpoint $\gamma = 3$ can only be fitted and evaluated in this specific case. Grid PP also performed poorly under this criterion, with substantial decrease in success as the data became more heterogeneous.
Figure 3.6 Comparison of success probabilities for $\gamma = 3$ as the component breakpoint (Note: Success for ROBPCA is conditional on a 4-PC result)

Table 3.3 ROBPCA tally of iterations based on number of resulting PCs according to data sequence.

<table>
<thead>
<tr>
<th>Data Sequence</th>
<th>Number of Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3-PC Result</td>
</tr>
<tr>
<td>B</td>
<td>29</td>
</tr>
<tr>
<td>Sk</td>
<td>9</td>
</tr>
<tr>
<td>H</td>
<td>33</td>
</tr>
</tbody>
</table>
In the scree plots summarizing the fitted values for the segmented models for all iterations (Figure 3.7), the $\gamma = 3$ breakpoint is visually apparent in all data sequences for PCA and ROBPCA. In Grid PP, the scree plot and fitted models approach a line as the data sequence progresses. Thus, quantitative localization via segmented regression for Grid PP is as challenging as qualitative visual inspection of a scree plot elbow.

Upon further examination of how the segmented model at $\gamma = 3$ is a better fit for the scree plot compared to a simple linear model, it can be inferred that PCA is the most appropriate technique to use when utilizing this criterion. Based on the distribution of MSE values for PCA as seen in Figure 3.8, the lowest values were observed for $\gamma = 3$ across all data sequences. In addition, the higher mean $\Delta R^2$ values for PCA compared to ROBPCA and Grid PP in Figure 3.9 reveals that the segmented model at $\gamma = 3$ explained the eigenvalue versus component relationship in PCA and this is consistent across all data sequences.

**Figure 3.7** Scree plots (solid black) with overlaid fitted regression models for various breakpoints ($\gamma = 2$, red; $\gamma = 3$, blue; $\gamma = 4$, green; $\gamma = 5$, pink).
Figure 3.8 MSE distribution for various $\gamma$. 
Figure 3.9 Trend of Mean $\Delta R^2$ for the segmented model at $\gamma = 3$ and a simple linear model.
Chapter 4
Discussion

4.1 Replicability of Dimension Reduction Solutions in Real Data

Lanning (1996) reiterated Everett’s (1983) suggestion that “under certain conditions, replicability provides an answer to the question of the number of dimensions to retain in component analysis.” Whereas rules-of-thumb criteria exist, decisions on the selection of important components to retain tend to be ad hoc in nature for a researcher working on a set of psychological assessment measures from a unique set of participants. In this case, replicability of dimension reduction solutions is encumbered by the characteristics and numbers of participants in the data set which are the main drivers of the measurement values from each assessment variable. Since classic dimension reduction methods such as PCA rely on similarities in variances to generate combinations of the original variables that form principal components, the covariance matrix of a multivariate correlated data set is unlike any other even when the same psychological assessment tools are used. The inherent peculiarities and complexities of any specific multivariate psychology data set may be difficult to replicate in practice.

Replicability has been explored in the estimation of the optimal number of PCs to retain from high-dimensional data through model validation methods. In chemometrics, Varmuza and Filzmoser (2009) promote the use of cross validation of the variance contribution of PCs from different subsets of the original data set. Yet, the $q$-selection process proceeds following rule-of-thumb criteria, for instance, at least 70% or up to 90% cumulative proportion of variation. Resampling methods such as bootstrap PCA for very high-dimensional data ($p > 1$ million) has been investigated to estimate sampling variability in PC scores and eigenvalues of a data set. Nonetheless, the theoretical properties of bootstrap PCA results are not well-understood and require further study under simulation (Fisher et al., 2016). Thus, this thesis opted for the recourse provided by a simulation experiment motivated by real data.
4.2 Simulation Studies in the Selection of $q$ Dimensions for PCA

The controlled data environment provided by simulation facilitates a better means to evaluate the performance of dimension reduction methods in properly identifying the intrinsic dimensionality (true $q$ dimensions) of multivariate data under various conditions. Simulation studies focusing on the accuracy of $q$-selection criteria in PCA were primarily prompted by recognition of the fact that overestimation and underestimation of $q$ are implications of the evidently ad hoc nature of existing rules-of-thumb. Zwick and Velicer (1982; 1986) evaluated the consistency of the following component selection rules specifically for PCA, namely Kaiser’s eigenvalue $\geq 1$, Cattell’s scree plot, Bartlett’s test, Velicer’s minimum average partial (MAP), and Horn’s parallel analysis. A population correlation matrix and several sample correlation matrices were synthetically generated and mean differences in the values of each rule were used to evaluate consistency. They have concluded that the Cattell’s scree plot remains useful and the exclusive use of Kaiser’s eigenvalue $\geq 1$ is not recommended. In this thesis, we have observed the challenge of using Cattell’s scree plot test in the ABI data when the “elbow” is visually ambiguous. With respect to the Kaiser-Guttman rule, we have discovered that increasing the minimum eigenvalue threshold to at least 1.5 may be suitable in conditions similar to this thesis’s motivating data set.

Modelling methods on the eigenvalue as a function of the PCs has been used as $q$ selection criteria for PCA under simulated conditions. Better recognized in the biological sciences as the broken stick (BS) model introduced by Frontier (1976), it is a stopping rule in numerical ecology. The BS model is determined by randomly dividing a “stick of unit length into the same number of pieces as there are PCA axes…the pieces are then put in order of decreasing length and compared to the eigenvalues (Borcard et al., 2011).” There exists a table of eigenvalues based on the broken stick distribution, which can be plotted and juxtaposed on the scree plot (Jackson, 1993).

Similar to the $q$-selection focus of this thesis, Cangelosi and Goriely (2007) applied modified BS models on simulated microarray data to determine the number of components to retain by overlaying a barplot of their designed BS distribution for each component on the eigenvalue scree plot. Where the scree plot lies above the BS distribution signifies the importance of the components.
In contrast to the intended use of BS for the eigenvalue scree plot but similar in implementation to this thesis, Peres-Neto et al. (2003) applied BS models to the squared loadings of PCs to assess the significance of loadings in PCA under various simulated data scenarios. It is evident that the broken stick model for component selection in PCA, while seemingly synonymous in name to breakpoint analysis, is different from the authentic segmented regression applied in this thesis whereby model fit is evaluated on the basis of the MSE and $R^2$.

4.3 Appraisal of Benchmarking Results

Given the specific correlation matrix and data attributes of the simulated data used in the benchmarking experiment, ROBPCA was consistently robust in identifying the correct $q$-dimensional solution and boosted when a cumulative proportion of variation cut-off is at 70%. In the cumulative proportion of variation threshold of 80%, none of the techniques showed success. Nevertheless, ROBPCA was able to explain the greatest variance contribution across simulated data conditions. We have shown that ROBPCA was 95% successful when the cumulative proportion of variation is set to at least 70%, which falls within the threshold range used in practice (70-90%). For the Kaiser-Guttman eigenvalue $\geq 1$ cut-off and quantitative localization of the scree plot breakpoint at $y = 3$, ROBPCA came in close second to PCA. Conversely, if the cumulative proportion of variation threshold is decreased to 60% and eigenvalue cut-off is increased to about 1.5, PCA may be as competitive as ROBPCA.

These discoveries rouse the idea that PCA can perform well in the presence of minimal orthogonal outliers and skewed variables when rules-of-thumb are adjusted post-benchmarking. Varmuza and Filzmoser (2009) stated that PCA remains a reliable dimension reduction tool “if the data distribution is elliptically symmetric around the center” and when variables are not highly skewed. Referring to the diagnostic plots for classical and robust dimension reduction applied to the real data set (Figures 2.7 and 2.11), majority of the observations are categorized as typical and while orthogonal outliers are present, the orthogonal distances are not remarkably large. If the distribution of the multivariate point cloud for this situation is visualized, it may still be reckoned as elliptically symmetric. The absence of any leverage points and the existence of orthogonal outliers located proximal to the centroid of the multivariate point cloud, aggregate all data points within the intrinsic dimensionality of the data distribution. However, the fact remains
that PCA did not explain as much variability as ROBPCA in the expected 2-component solution under the naïve use of standard selection criteria.

The development of ROBPCA is founded on the awareness that certain heavily influential outliers, though minimal in proportion to the rest of the data, can still disturb the intrinsic dimensionality of the data. Thus, the iterative acquisition of subsets of the original data that results to the MCD is a reflection of Hubert et al.’s (2008) “combinatorial viewpoint” that “it is more feasible to search for sufficiently many ‘good’ data points than to find all the ‘bad’ data points.” The preliminary dimension reduction step in ROBPCA leads to a number of initial components that is less than the original p-dimensions as opposed to PCA or Grid PP wherein linear transformation of the original p-dimensional data initially results to the same number of PCs from which the selection of the reduced q dimensions is performed. In section 3.3, ROBPCA’s performance in localizing the correct scree plot breakpoint at \( y = 3 \) based on regression measures, though at par with PCA, was influenced by the lesser number of initial PCs. On a related note, implementation of the benchmarking experiment for this portion of the methodological framework required the development of additional script (Appendix B) to accommodate the varying number of initial PCs produced in ROBPCA. Upon visual inspection of ROBPCA scree plots across data sequences, \( y = 3 \) is still unmistakably perceptible even though the shift in slopes of the two segments is not as pronounced as in PCA.

For Grid PP as the PP algorithm of interest, the multivariate data conditions and evaluation criteria used to compare its performance to PCA and ROBPCA showed that it is the least successful. About 45%-60% cumulative proportion of variation was observed for \( q=2 \) and scree plot segmented models for \( y = 3 \) are similar to simple linear models when data are more heterogeneous. These signify that under Grid PP there may be a tendency to overestimate \( q \) and retain more PCs than necessary. Among the projection pursuit techniques, Croux et al. (2007) has proven that Grid PP is an “omnibus” algorithm “that can serve to maximize any projection index” with varying efficiency depending on the degree of outlier contamination present in the data and the dimensions of the data matrix. Furthermore, Grid PP was presented to be more accurate when \( n \leq p \) and tends to yield very high contributions to explained variability in terms of eigenvalues. This is apparent in Grid PP’s eigenvalues under skewed and heterogeneous conditions in which \( d > 2 \), which could be attributable to projection index maximization.
Chapter 5
Conclusion and Recommendations

In the application of dimension reduction to highly correlated and non-normal multivariate psychological assessment data, inadvertent underestimation or overestimation of $q$ may be committed under naïve use of rules-of-thumb. ROBPCA is the ideal technique to use when standard criteria and threshold values are implemented in the selection of a reduced $q$-dimensional solution. In the combined implementation of 70\% cumulative proportion of variation, Kaiser-Guttman eigenvalue $\geq 1$ rule, and segmented regression localization of Cattell’s scree plot elbow, ROBPCA correctly reveals intrinsic dimensionality under various levels of data heterogeneity.

The potential biased estimation of $q$ in classical PCA for multivariate data with asymmetrical distributions and atypical observations may be circumvented by adjusting the cumulative proportion of variation threshold to 60\% and eigenvalue cut-off to $\sim 1.5$. In addition, Cattell’s scree plot elbow supplemented with the use of minimum MSE determination for quantitative localization of the best scree plot breakpoint could enhance the consistency of PCA.

Grid PP was demonstrated to be least effective under the specific performance criteria and multivariate data conditions explored in this thesis. It tends to overestimate $q$ based on at least 40\% cumulative proportion of variation, eigenvalue $> 2$, and inconspicuous scree plot elbow for the true $q$-dimensional solution in the face of asymmetric distributions. This algorithm may be further investigated under different data matrix dimensions using different scale measures in silico.

A heuristic approach through empirical benchmarking is recommended to serve as an auxiliary procedure to determine the appropriateness of rules-of-thumb criteria in the selection of $q$ when performing dimension reduction to specific multivariate data sets.
References


## Appendix Table A: Descriptive Statistics of ABI Data Set

<table>
<thead>
<tr>
<th>Psychological Assessment</th>
<th>Variables</th>
<th>Description</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Shipley Test: Evaluates crystallized ability and fluid cognitive ability. (Kaya et al., 2012)</td>
<td>shipprer</td>
<td>Shipley Raw Score</td>
<td>28.690</td>
<td>6.718</td>
<td>-0.758</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td>shippret</td>
<td>Shipley T Score</td>
<td>47.024</td>
<td>12.848</td>
<td>-0.496</td>
<td>-0.304</td>
</tr>
<tr>
<td></td>
<td>shipprep</td>
<td>Shipley Percentile</td>
<td>45.497</td>
<td>32.132</td>
<td>0.066</td>
<td>-1.379</td>
</tr>
<tr>
<td>(2) Self-Ordered Pointing: Assesses frontal lobe pathology via working memory. (Ross et al., 2007)</td>
<td>sopprer</td>
<td>Raw Score</td>
<td>4.131</td>
<td>2.216</td>
<td>0.559</td>
<td>0.391</td>
</tr>
<tr>
<td>(3) Hopkins Verbal Learning Test (HVLT): Evaluates verbal learning and memory. (Benedict et al., 1998)</td>
<td>hvtoprer</td>
<td>Total Raw Score</td>
<td>20.310</td>
<td>6.705</td>
<td>-0.189</td>
<td>-0.382</td>
</tr>
<tr>
<td></td>
<td>hvtoprep</td>
<td>Total Percentile</td>
<td>21.017</td>
<td>28.108</td>
<td>0.066</td>
<td>-0.222</td>
</tr>
<tr>
<td></td>
<td>hvtopret</td>
<td>Total T Score</td>
<td>35.619</td>
<td>13.223</td>
<td>0.464</td>
<td>-0.945</td>
</tr>
<tr>
<td></td>
<td>hvdepre</td>
<td>Delay Raw Score</td>
<td>6.345</td>
<td>3.511</td>
<td>-0.317</td>
<td>-1.091</td>
</tr>
<tr>
<td></td>
<td>hvdepre</td>
<td>Delay Percentile</td>
<td>22.777</td>
<td>27.585</td>
<td>0.927</td>
<td>-0.548</td>
</tr>
<tr>
<td></td>
<td>hvdepret</td>
<td>Delay T Score</td>
<td>36.012</td>
<td>13.578</td>
<td>0.204</td>
<td>-1.388</td>
</tr>
<tr>
<td></td>
<td>hvrepre</td>
<td>Retention Raw Score</td>
<td>72.690</td>
<td>31.269</td>
<td>-1.092</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>hvreprep</td>
<td>Retention Percentile</td>
<td>32.427</td>
<td>31.895</td>
<td>0.532</td>
<td>-1.261</td>
</tr>
<tr>
<td></td>
<td>hvrepret</td>
<td>Retention T Score</td>
<td>40.798</td>
<td>13.989</td>
<td>-0.106</td>
<td>-1.241</td>
</tr>
<tr>
<td></td>
<td>hvdspre</td>
<td>Discrimination Raw Score</td>
<td>9.429</td>
<td>2.490</td>
<td>0.559</td>
<td>-1.189</td>
</tr>
<tr>
<td></td>
<td>hvdsprep</td>
<td>Discrimination Percentile</td>
<td>33.331</td>
<td>30.024</td>
<td>0.380</td>
<td>-1.372</td>
</tr>
<tr>
<td></td>
<td>hvdspret</td>
<td>Discrimination T Score</td>
<td>41.262</td>
<td>13.432</td>
<td>-0.363</td>
<td>-1.189</td>
</tr>
<tr>
<td>(4) Delis-Kaplan Executive Function System (D-KEFS): Measures executive functioning through a set of tasks in game-like format. (Homack et al., 2005) Trail-Making Conditions: (2) Number Sequencing; (3) Letter Sequencing; (4) Number-Letter Sequencing; and (5) Motor Speed.</td>
<td>tra2prer</td>
<td>Trail-Making Condition 2 Raw Score</td>
<td>67.607</td>
<td>35.772</td>
<td>0.832</td>
<td>-0.298</td>
</tr>
<tr>
<td></td>
<td>tra2pres</td>
<td>Trail-Making Condition 2 Scaled Score</td>
<td>6.464</td>
<td>4.337</td>
<td>-0.032</td>
<td>-1.514</td>
</tr>
<tr>
<td></td>
<td>tra2prep</td>
<td>Trail-Making Condition 2 Percentile</td>
<td>28.081</td>
<td>29.423</td>
<td>0.592</td>
<td>-1.057</td>
</tr>
<tr>
<td></td>
<td>tra3prer</td>
<td>Trail-Making Condition 3 Raw Score</td>
<td>73.619</td>
<td>41.439</td>
<td>0.709</td>
<td>-0.868</td>
</tr>
<tr>
<td></td>
<td>tra3pres</td>
<td>Trail-Making Condition 3 Scaled Score</td>
<td>6.321</td>
<td>4.327</td>
<td>-0.030</td>
<td>-1.535</td>
</tr>
<tr>
<td></td>
<td>tra3prep</td>
<td>Trail-Making Condition 3 Percentile</td>
<td>27.164</td>
<td>29.273</td>
<td>0.632</td>
<td>-1.111</td>
</tr>
<tr>
<td></td>
<td>tra4prer</td>
<td>Trail-Making Condition 4 Raw Score</td>
<td>158.976</td>
<td>67.700</td>
<td>0.057</td>
<td>-1.687</td>
</tr>
<tr>
<td></td>
<td>tra4pres</td>
<td>Trail-Making Condition 4 Scaled Score</td>
<td>6.012</td>
<td>4.381</td>
<td>0.114</td>
<td>-1.545</td>
</tr>
<tr>
<td></td>
<td>tra4prep</td>
<td>Trail-Making Condition 4 Percentile</td>
<td>25.592</td>
<td>28.713</td>
<td>0.681</td>
<td>-0.961</td>
</tr>
<tr>
<td></td>
<td>tra5prer</td>
<td>Trail-Making Condition 5 Raw Score</td>
<td>56.357</td>
<td>36.334</td>
<td>1.320</td>
<td>0.868</td>
</tr>
<tr>
<td></td>
<td>tra5pres</td>
<td>Trail-Making Condition 5 Scaled Score</td>
<td>7.964</td>
<td>3.870</td>
<td>-0.684</td>
<td>-0.842</td>
</tr>
<tr>
<td></td>
<td>tra5prep</td>
<td>Trail-Making Condition 5 Percentile</td>
<td>37.415</td>
<td>28.708</td>
<td>0.041</td>
<td>-1.456</td>
</tr>
<tr>
<td></td>
<td>cond4av23</td>
<td>Difference Conditions 4 - Ave(2 and 3)</td>
<td>88.363</td>
<td>44.443</td>
<td>0.544</td>
<td>-0.592</td>
</tr>
<tr>
<td></td>
<td>cond4minus5</td>
<td>Difference Condition 4 - Condition 5</td>
<td>102.619</td>
<td>58.365</td>
<td>0.476</td>
<td>-1.207</td>
</tr>
<tr>
<td></td>
<td>letprer</td>
<td>Letter Fluency Raw Score</td>
<td>28.571</td>
<td>13.660</td>
<td>0.096</td>
<td>-0.398</td>
</tr>
<tr>
<td></td>
<td>letpret</td>
<td>Letter Fluency Scaled Score</td>
<td>7.702</td>
<td>3.936</td>
<td>0.133</td>
<td>-0.519</td>
</tr>
<tr>
<td></td>
<td>letprep</td>
<td>Letter Fluency Percentile</td>
<td>32.274</td>
<td>30.632</td>
<td>0.731</td>
<td>-0.762</td>
</tr>
<tr>
<td></td>
<td>bwerts</td>
<td>Tower Total Achievement</td>
<td>6.607</td>
<td>2.794</td>
<td>-0.103</td>
<td>-0.656</td>
</tr>
<tr>
<td></td>
<td>bwert1m</td>
<td>Tower Mean First Move Time</td>
<td>5.457</td>
<td>4.551</td>
<td>1.662</td>
<td>2.892</td>
</tr>
<tr>
<td></td>
<td>bwerrv</td>
<td>Tower Rule Violations per Item</td>
<td>0.547</td>
<td>0.767</td>
<td>2.145</td>
<td>5.142</td>
</tr>
<tr>
<td>(5) Patient Health Questionnaire (PHQ): Anxiety and Depression Questionnaire; from 0-21 for anxiety and 0-27 for depression, higher scores indicate higher levels of anxiety and depression (Kroenke et al., 2016).</td>
<td>gad7pre</td>
<td>Anxiety Scale</td>
<td>5.048</td>
<td>4.366</td>
<td>0.758</td>
<td>-0.111</td>
</tr>
<tr>
<td></td>
<td>phq9pre</td>
<td>Depression Scale</td>
<td>5.655</td>
<td>4.356</td>
<td>0.633</td>
<td>-0.476</td>
</tr>
</tbody>
</table>

*(1) Shipley Test: Evaluates crystallized ability and fluid cognitive ability. (Kaya et al., 2012)*

*(2) Self-Ordered Pointing: Assesses frontal lobe pathology via working memory. (Ross et al., 2007)*

*(3) Hopkins Verbal Learning Test (HVLT): Evaluates verbal learning and memory. (Benedict et al., 1998)*

*(4) Delis-Kaplan Executive Function System (D-KEFS): Measures executive functioning through a set of tasks in game-like format. (Homack et al., 2005) Trail-Making Conditions: (2) Number Sequencing; (3) Letter Sequencing; (4) Number-Letter Sequencing; and (5) Motor Speed.*

*(5) Patient Health Questionnaire (PHQ): Anxiety and Depression Questionnaire; from 0-21 for anxiety and 0-27 for depression, higher scores indicate higher levels of anxiety and depression (Kroenke et al., 2016).*
<table>
<thead>
<tr>
<th>Psychological Assessment</th>
<th>Variables</th>
<th>Description</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6) Behaviour Rating Inventory of Executive Function-Adult (BRIEF-A): Obtains information about an adult's perception of her/his executive function or self-regulation in daily life. (Rouel et al., 2016)</td>
<td>gecpre</td>
<td>Global Executive Composite (GEC)</td>
<td>116.226</td>
<td>24.757</td>
<td>0.363</td>
<td>-0.386</td>
</tr>
<tr>
<td></td>
<td>gecpret</td>
<td>GEC T Score</td>
<td>48.762</td>
<td>11.170</td>
<td>0.300</td>
<td>-0.938</td>
</tr>
<tr>
<td></td>
<td>gecprep</td>
<td>GEC Percentile</td>
<td>67.464</td>
<td>15.682</td>
<td>0.490</td>
<td>-0.134</td>
</tr>
<tr>
<td></td>
<td>birep</td>
<td>Behaviour Regulation Index</td>
<td>124.679</td>
<td>25.634</td>
<td>0.348</td>
<td>-0.418</td>
</tr>
<tr>
<td></td>
<td>mipre</td>
<td>Metacognition Index</td>
<td>60.488</td>
<td>12.917</td>
<td>0.410</td>
<td>-0.162</td>
</tr>
<tr>
<td></td>
<td>sumpre</td>
<td>Summary Score</td>
<td>76.095</td>
<td>24.956</td>
<td>-1.121</td>
<td>0.063</td>
</tr>
<tr>
<td>(7) Coping with Health Injuries and Problems (CHIP): Self-report instrument that measures coping reactions and strategies used by individuals with health problems. (Endler et al., 1998)</td>
<td>prehiddc</td>
<td>Distractive Coping</td>
<td>25.917</td>
<td>6.022</td>
<td>-0.224</td>
<td>-0.730</td>
</tr>
<tr>
<td></td>
<td>prechipc</td>
<td>Palliative Coping</td>
<td>26.000</td>
<td>5.357</td>
<td>0.348</td>
<td>-0.177</td>
</tr>
<tr>
<td></td>
<td>prehiic</td>
<td>Instrumental Coping</td>
<td>31.083</td>
<td>5.800</td>
<td>-0.180</td>
<td>-0.713</td>
</tr>
<tr>
<td></td>
<td>prehiip</td>
<td>Emotional Preoccupation</td>
<td>24.940</td>
<td>8.369</td>
<td>-0.057</td>
<td>-1.025</td>
</tr>
<tr>
<td></td>
<td>prehiii</td>
<td>Inconsistency Index</td>
<td>8.226</td>
<td>2.971</td>
<td>0.545</td>
<td>-0.466</td>
</tr>
<tr>
<td>(8) Locus of Control: Personality questionnaire on an individual's perceived control (Peterson, 2003)</td>
<td>preloc</td>
<td>Locus of Control Score</td>
<td>50.749</td>
<td>9.912</td>
<td>-0.195</td>
<td>-0.818</td>
</tr>
<tr>
<td>(9) Mayo-Portland Adaptability Inventory (MPAI): Specifically designed to evaluate deficits in individuals with acquired brain injury. (Bellon et al., 2012)</td>
<td>MPAI_Rad_pre</td>
<td>Adjustment Subscale</td>
<td>13.321</td>
<td>7.583</td>
<td>0.378</td>
<td>-0.689</td>
</tr>
<tr>
<td></td>
<td>MPAI_Rp_pre</td>
<td>Participation Subscale</td>
<td>10.821</td>
<td>4.939</td>
<td>0.080</td>
<td>-0.980</td>
</tr>
<tr>
<td>(10) General Self-Efficacy (GSE): Measures perceived ability to deal with demanding circumstances. (Romppel et al., 2013)</td>
<td>pregse</td>
<td>General Self-Efficacy Score</td>
<td>29.869</td>
<td>4.307</td>
<td>0.125</td>
<td>-0.159</td>
</tr>
<tr>
<td>(11) Short Form Health Survey (SF-36): Self-reported questionnaire on health-related quality of life; from 0-100, higher scores indicate better health-related quality of life (Fawkes, 2013)</td>
<td>presb6_pf</td>
<td>Physical Functioning</td>
<td>20.512</td>
<td>5.509</td>
<td>0.115</td>
<td>-1.133</td>
</tr>
<tr>
<td></td>
<td>presb6_rp</td>
<td>Role Physical</td>
<td>5.595</td>
<td>1.614</td>
<td>0.391</td>
<td>-1.487</td>
</tr>
<tr>
<td></td>
<td>presb6_bp</td>
<td>Bodily Pain</td>
<td>8.896</td>
<td>2.667</td>
<td>-0.283</td>
<td>-1.249</td>
</tr>
<tr>
<td></td>
<td>presb6_gh</td>
<td>General Health</td>
<td>18.090</td>
<td>3.801</td>
<td>0.022</td>
<td>-0.613</td>
</tr>
<tr>
<td></td>
<td>presb6_vt</td>
<td>Vitality</td>
<td>15.048</td>
<td>4.045</td>
<td>-0.173</td>
<td>-0.360</td>
</tr>
<tr>
<td></td>
<td>presb6_sf</td>
<td>Social Functioning</td>
<td>7.488</td>
<td>2.220</td>
<td>-0.650</td>
<td>-0.383</td>
</tr>
<tr>
<td></td>
<td>presb6_re</td>
<td>Role Emotional</td>
<td>4.952</td>
<td>1.241</td>
<td>-0.621</td>
<td>-1.317</td>
</tr>
<tr>
<td></td>
<td>presb6_mh</td>
<td>Mental Health</td>
<td>23.238</td>
<td>3.941</td>
<td>-0.445</td>
<td>-0.338</td>
</tr>
</tbody>
</table>
# Appendix B: R Script for Simulated Data Sequence B

# SEQUENCE 1 - Multivariate Normal Case

library(PoisNonNor)
packageVersion("PoisNonNor") # [1] '1.1'
library(rrcov)
packageVersion("rrcov") # [1] '1.3.11'
library(psych)
packageVersion("psych") # [1] '1.6.4'

## Parameters for Data Simulation

# Poisson lambdas for variables 1-3
lamvec <- c(24, 24, 24)

# correlation matrix (based on raw data)
cmat <- matrix(c(
  1, -0.01, 0.5, -0.01, 0.5, -0.01,
  -0.01, 1, 0.05, 0.5, -0.01, 0.5,
  0.5, 0.05, 1, -0.01, 0.5, -0.01,
  -0.01, 0.5, -0.01, 1, 0.05, 0.5,
  0.5, -0.01, 0.5, 0.05, 1, -0.01,
  -0.01, 0.5, -0.01, 0.5, -0.01, 1), nrow=6, byrow=TRUE)
det(cmat)

# means and variances of variables 4-6
mean.vec <- c(0, 0, 0)
variance.vec <- c(24, 24, 24)

# matrix of skewness and kurtosis of variables 4-6
rmat <- matrix(c(0, 3, 0, 3, 0, 3), byrow = TRUE, ncol=2)

# number of observations/rows
norow <- 100

## Loop through Simulated Data and Classic PCA

R <- 100
corsim <- c(); sumsim <- c()
load.pc1 <- c(); load.pc2 <- c(); load.pc3 <- c()
z2.mse <- c(); z3.mse <- c(); z4.mse <- c(); z5.mse <- c()
load.pc1g <- c(); load.pc2g <- c(); load.pc3g <- c()
z2.mse.g <- c(); z3.mse.g <- c(); z4.mse.g <- c(); z5.mse.g <- c()
load.pc1h <- c(); load.pc2h <- c(); load.pc3h <- c()
z2.mse.h <- c(); z3.mse.h <- c(); z4.mse.h <- c(); z5.mse.h <- c()
m.pca.eig1 <- c(); m.pca.eig2 <- c(); m.pca.eig3 <- c()
m.pca.eig4 <- c(); m.pca.eig5 <- c(); m.pca.eig6 <- c()
m.g.eig1 <- c(); m.g.eig2 <- c(); m.g.eig3 <- c()
m.g.eig4 <- c(); m.g.eig5 <- c(); m.g.eig6 <- c()
m.h.eig1 <- c(); m.h.eig2 <- c(); m.h.eig3 <- c()
m.h.eig4 <- c(); m.h.eig5 <- c(); m.h.eig6 <- c()
pca.cp <- c(); grid.cp <- c(); hub.cp <- c()
pca.eig <- c(); grid.eig <- c(); hub.eig <- c()
hub.numPCs<- c()

##Loop proper
system.time(
  for(i in 1:R) {
    simulated <- RNG_P_NN(lamvec, cmat, rmat, norow, mean.vec, variance.vec)
    corsim[[i]] <- cor(simulated) #correlation matrices per "R"
    sumsim[[i]] <- describe(simulated) #simulated data summary per "R"
  }
)

# Classical PCA
pca = PcaClassic(simulated, scale=TRUE)
imp = summary(pca)
a = data.frame(imp@importance)
cmprop = subset(a[,3]) #extract row on cumulative proportions
eigen = pca@eigenvalues
m = data.frame(pca@eigenvalues)
m.pca.eig1[i] = m[1,] #eigenvalue of PC1
m.pca.eig2[i] = m[2,] #eigenvalue of PC2
m.pca.eig3[i] = m[3,] #eigenvalue of PC3
m.pca.eig4[i] = m[4,] #eigenvalue of PC4
m.pca.eig5[i] = m[5,] #eigenvalue of PC5
m.pca.eig6[i] = m[6,] #eigenvalue of PC6

load = data.frame(pca@loadings)
load.pc1[[i]] <- load[,1] #Vector of loadings for PC1
load.pc2[[i]] <- load[,2] #Vector of loadings for PC2
load.pc3[[i]] <- load[,3] #Vector of loadings for PC3

PC = c(1,2,3,4,5,6)

# two-segment model for classical PCA
z2 <- lm(eigen~PC + I((PC-2)*ifelse(PC>=2,1,0)))
z3 <- lm(eigen~PC + I((PC-3)*ifelse(PC>=3,1,0)))
z4 <- lm(eigen~PC + I((PC-4)*ifelse(PC>=4,1,0)))
z5 <- lm(eigen~PC + I((PC-5)*ifelse(PC>=5,1,0)))

z2.mse[i] = mean(z2$residuals^2)
z3.mse[i] = mean(z3$residuals^2)
z4.mse[i] = mean(z4$residuals^2)
z5.mse[i] = mean(z5$residuals^2)

# Grid Projection Pursuit
grid = PcaGrid(simulated, scale=TRUE)
imp.g = summary(grid)
b = data.frame(imp.g@importance)
cmprop.g = subset(b[,3])
eigen.g = grid@eigenvalues
m.g = data.frame(grid@eigenvalues)
m.g.eig1[i] = m.g[1,] #eigenvalue of PC1
m.g.eig2[i] = m.g[2,] #eigenvalue of PC2
m.g.eig3[i] = m.g[3,] #eigenvalue of PC3
m.g.eig4[i] = m.g[4,] #eigenvalue of PC4
m.g.eig5[i] = m.g[5,] #eigenvalue of PC5
m.g.eig6[i] = m.g[6,] #eigenvalue of PC6
load.g = data.frame(grid@loadings)
load.pc1g[[i]] <- load.g[,1]
load.pc2g[[i]] <- load.g[,2]
load.pc3g[[i]] <- load.g[,3]

#two-segment model for Grid Projection Pursuit
z2.g <- lm(eigen.g~PC + I((PC-2)*ifelse(PC>=2,1,0)))
z3.g <- lm(eigen.g~PC + I((PC-3)*ifelse(PC>=3,1,0)))
z4.g <- lm(eigen.g~PC + I((PC-4)*ifelse(PC>=4,1,0)))
z5.g <- lm(eigen.g~PC + I((PC-5)*ifelse(PC>=5,1,0)))

z2.mse.g[i] = mean(z2.g$residuals^2)
z3.mse.g[i] = mean(z3.g$residuals^2)
z4.mse.g[i] = mean(z4.g$residuals^2)
z5.mse.g[i] = mean(z5.g$residuals^2)

#ROBPCA
hub = PcaHubert(simulated, scale=TRUE)
imp.h = summary(hub)
c = data.frame(imp.h@importance)
cmprop.h = subset(c[3,])
eigen.h = hub@eigenvalues
m.h = data.frame(hub@eigenvalues)
load.h = data.frame(hub@loadings)
load.pc1h[[i]] <- load.h[,1]
load.pc2h[[i]] <- load.h[,2]
load.pc3h[[i]] <- load.h[,3]

colseigen.h <- length(eigen.h)

for(j in 1:colseigen.h) {

#two-segment model for ROBPCA
if (colseigen.h == 3) {
  PC2 = c(1,2,3)
  z2.h <- lm(eigen.h~PC2 + I((PC2-2)*ifelse(PC2>=2,1,0)))
  m.h.eig1[i] = m.h[1,] #eigenvalue of PC1
  m.h.eig2[i] = m.h[2,] #eigenvalue of PC2
  m.h.eig3[i] = m.h[3,] #eigenvalue of PC3
}
else if (colseigen.h == 4) {
  PC2 = c(1,2,3,4)
  z2.h <- lm(eigen.h~PC2 + I((PC2-2)*ifelse(PC2>=2,1,0)))
z3.h <- lm(eigen.h~PC2 + I((PC2-3)*ifelse(PC2>=3,1,0)))
z3.mse.h[i] = mean(z3.h$residuals^2)
  m.h.eig1[i] = m.h[1,] #eigenvalue of PC1
  m.h.eig2[i] = m.h[2,] #eigenvalue of PC2
  m.h.eig3[i] = m.h[3,] #eigenvalue of PC3
  m.h.eig4[i] = m.h[4,] #eigenvalue of PC4
}
else if (colseigen.h == 5)
{
  PC2 = c(1,2,3,4,5)
  z2.h <- lm(eigen.h~PC2 + I((PC2-2)*ifelse(PC2>=2,1,0)))
  z3.h <- lm(eigen.h~PC2 + I((PC2-3)*ifelse(PC2>=3,1,0)))
  z3.mse.h[i] = mean(z3.h$residuals^2)
  z4.h <- lm(eigen.h~PC2 + I((PC2-4)*ifelse(PC2>=4,1,0)))
  z4.mse.h[i] = mean(z4.h$residuals^2)
  m.h.eig1[i] = m.h[1,] #eigenvalue of PC1
  m.h.eig2[i] = m.h[2,] #eigenvalue of PC2
  m.h.eig3[i] = m.h[3,] #eigenvalue of PC3
  m.h.eig4[i] = m.h[4,] #eigenvalue of PC4
  m.h.eig5[i] = m.h[5,] #eigenvalue of PC5
}
else if (colseigen.h == 6)
{
  PC2 = c(1,2,3,4,5,6)
  z2.h <- lm(eigen.h~PC2 + I((PC2-2)*ifelse(PC2>=2,1,0)))
  z3.h <- lm(eigen.h~PC2 + I((PC2-3)*ifelse(PC2>=3,1,0)))
  z3.mse.h[i] = mean(z3.h$residuals^2)
  z4.h <- lm(eigen.h~PC2 + I((PC2-4)*ifelse(PC2>=4,1,0)))
  z4.mse.h[i] = mean(z4.h$residuals^2)
  z5.h <- lm(eigen.h~PC2 + I((PC2-5)*ifelse(PC2>=5,1,0)))
  z5.mse.h[i] = mean(z5.h$residuals^2)
  m.h.eig1[i] = m.h[1,] #eigenvalue of PC1
  m.h.eig2[i] = m.h[2,] #eigenvalue of PC2
  m.h.eig3[i] = m.h[3,] #eigenvalue of PC3
  m.h.eig4[i] = m.h[4,] #eigenvalue of PC4
  m.h.eig5[i] = m.h[5,] #eigenvalue of PC5
  m.h.eig6[i] = m.h[6,] #eigenvalue of PC6
}
z2.mse.h[i] = mean(z2.h$residuals^2)

#Cumulative Percentage of Total Variation for 2 PCs
pca.cp[i] <- cmprop[,2]
grid.cp[i] <- cmprop.g[,2]
hub.cp[i] <- cmprop.h[,2]

#Eigenvalue of PC3
pca.eig[i] <- m[3,]
grid.eig[i] <- m.g[3,]
hub.eig[i] <- m.h[3,]

hub.numPCs[i] = ncol(imp.h@importance) #Number of PCs resulting from ROBPCA
#OUTPUT AND SUMMARIES
outseq1 <- cbind.data.frame(pca.cp, grid.cp, hub.cp, pca.eig, grid.eig, hub.eig, hub.numPCs)
write.csv(outseq1, file="~/outseq1.csv")
cpeig <- read.csv("~/outseq1.csv")
summary(cpeig)

#Distributions of Cumulative Percentages of Total Variation for 2 PCs
ecdfCP.pca <- ecdf(cpeig$pca.cp)
inv_ecdf.pca <- function(ecdfCP.pca){
  x <- environment(ecdfCP.pca)$x
  y <- environment(ecdfCP.pca)$y
  approxfun(y, x)
} g.pca <- inv_ecdf.pca(ecdfCP.pca); g.pca(0.05)

ecdfCP.grid <- ecdf(cpeig$grid.cp)
inv_ecdf.grid <- function(ecdfCP.grid){
  x <- environment(ecdfCP.grid)$x
  y <- environment(ecdfCP.grid)$y
  approxfun(y, x)
} g.grid <- inv_ecdf.grid(ecdfCP.grid); g.grid(0.05)

ecdfCP.hub <- ecdf(cpeig$hub.cp)
inv_ecdf.hub <- function(ecdfCP.hub){
  x <- environment(ecdfCP.hub)$x
  y <- environment(ecdfCP.hub)$y
  approxfun(x, y)
} g.hub <- inv_ecdf.hub(ecdfCP.hub); g.hub(0.05)

#Hubert Error Probability at theta=0.80
ecdf.hub <- function(ecdfCP.hub){
  x <- environment(ecdfCP.hub)$x
  y <- environment(ecdfCP.hub)$y
  approxfun(x, y)
} g.hubERR <- ecdf.hub(ecdfCP.hub)
g.hubERR(0.8)

par(pty="s")
par(mfrow=c(2,3), mai=c(0.7, 0.6, 0.2, 0.1))
hist(cpeig$pca.cp, xlim=c(0.40,0.90), ylim=c(0,35),
  xlab="Cumulative Prop. 2nd Component", main="Classical PCA")
hist(cpeig$grid.cp, xlim=c(0.40,0.90), ylim=c(0,35), breaks=10,
  xlab="Cumulative Prop. 2nd Component", main="Grid Projection Pursuit")
hist(cpeig$hub.cp, xlim=c(0.40,0.90), ylim=c(0,35),
  xlab="Cumulative Prop. 2nd Component", main="ROBPCA")
plot(ecdf(cpeig$pca.cp), xlim=c(0.40,0.90), ylab="Error Probability",
  xlab="", main="Classical PCA")
  abline(h=0.05, col="red", lty=2); abline(v=g.pca(0.05), col="red")
text(0.55,0.1, paste(round(g.pca(0.05),4)), col="red")
  abline(v=0.80, col="blue")
plot(ecdf(cpeig$grid.cp), xlim=c(0.40,0.90), ylab="Error Probability",
  xlab="", main="Grid Projection Pursuit")
  abline(h=0.05, col="red", lty=2); abline(v=g.grid(0.05), col="red")
#Distributions of Eigenvalues of 3rd PC
ecdfeig.pca <- ecdf(cpeig$pca.eig)

ecdfeig.grid <- ecdf(cpeig$grid.eig)

ecdfeig.hub <- ecdf(cpeig$hub.eig)

dev.off()
# Classic PCA
par(pty="s")
par(mfrow=c(1,2), mai=c(0.7, 1, 0.2, 0.1))

# For Summarized Scree

eigs.pca <- cbind.data.frame(m.pca.eig1, m.pca.eig2, m.pca.eig3, m.pca.eig4, m.pca.eig5, m.pca.eig6)
write.csv(eigs.pca, file="C:/Users/Anna/OneDrive/THESISR/2017/24MAR2017/SEQ1/SEQ1_eigsPCA.csv")

z2fit.pca <- lm(colMeans(eigs.pca)~PC + I((PC-2)*ifelse(PC>=2,1,0)))
z3fit.pca <- lm(colMeans(eigs.pca)~PC + I((PC-3)*ifelse(PC>=3,1,0)))
z4fit.pca <- lm(colMeans(eigs.pca)~PC + I((PC-4)*ifelse(PC>=4,1,0)))
z5fit.pca <- lm(colMeans(eigs.pca)~PC + I((PC-5)*ifelse(PC>=5,1,0)))

plot(PC, colMeans(eigs.pca), type="o", main="Classical PCA", ylim=c(0,2.5), xlab="Component", ylab="Eigenvalue", lwd=2)
lines(PC, z2fit.pca$fitted.values, lty=2, lwd=2, col="red")
lines(PC, z3fit.pca$fitted.values, lty=2, lwd=2, col="blue")
lines(PC, z4fit.pca$fitted.values, lty=2, lwd=2, col="green")
lines(PC, z5fit.pca$fitted.values, lty=2, lwd=2, col="violet")
text(2,1.2,"2", col="red"); text(3, 0.45,"3", col="blue")
text(4,0.3,"4", col="green"); text(5, 0.1,"5", col="violet")

mses <- cbind.data.frame(z2.mse, z3.mse, z4.mse, z5.mse)
write.csv(mses, file="C:/SEQ1/SEQ1_mses.csv")

mse.pca <- read.csv("C:/SEQ1/SEQ1_mses.csv")
summary(mse.pca) # since csv adds observation variable then there are 5 columns

barplot(table(apply(mse.pca,1,which.min)), ylim=c(0,100), main="Classical PCA", col="gray85", xlab=expression(paste(gamma," component breakpoint at min. MSE")), ylab="Success Probability")
text(0.7,96,"94", col="blue"); text(1.9,8,"6", col="blue")
table(apply(mse.pca,1,which.min)) # Get Probability Values for text above
dev.off()

# Grid Projection Pursuit
par(pty="s")
par(mfrow=c(1,2), mai=c(0.7, 1, 0.2, 0.1))

# For Summarized Scree

eigs.grid <- cbind.data.frame(m.g.eig1, m.g.eig2, m.g.eig3, m.g.eig4, m.g.eig5, m.g.eig6)
write.csv(eigs.grid, file="~/SEQ1/SEQ1_eigsGRID.csv")

z2fit.g <- lm(colMeans(eigs.grid)~PC + I((PC-2)*ifelse(PC>=2,1,0)))
z3fit.g <- lm(colMeans(eigs.grid)~PC + I((PC-3)*ifelse(PC>=3,1,0)))
z4fit.g <- lm(colMeans(eigs.grid)~PC + I((PC-4)*ifelse(PC>=4,1,0)))
z5fit.g <- lm(colMeans(eigs.grid)~PC + I((PC-5)*ifelse(PC>=5,1,0)))

plot(PC, colMeans(eigs.grid), type="o", main="Grid Projection Pursuit", ylim=c(0,3), xlab="Component", ylab="Eigenvalue", lwd=2)
lines(PC, z2fit.g$fitted.values, lty=2, lwd=2, col="red")
lines(PC, z3fit.g$fitted.values, lty=2, lwd=2, col="blue")
lines(PC, z4fit.g$fitted.values, lty=2, lwd=2, col="green")
lines(PC, z5fit.g$fitted.values, lty=2, lwd=2, col="violet")
text(2,1.5,"2", col="red"); text(3, 0.9,"3", col="blue")
text(4,0.5,"4", col="green"); text(5, 0.2,"5", col="violet")

mses.g <- cbind.data.frame(z2.mse.g, z3.mse.g, z4.mse.g, z5.mse.g); mses.g
write.csv(mses, file="~/SEQ1/SEQ1_msesGRID.csv")

mse.grid <- read.csv("~/SEQ1/SEQ1_msesGRID.csv")
summary(mse.grid)

barplot(table(apply(mse.grid,1,which.min)), ylim=c(0,100),
main="Grid Projection Pursuit", col="gray85",
xlab= expression(paste(gamma,"  ,component breakpoint at min. MSE")),
ylab="Success Probability")
text(0.7,96,"94", col="blue"); text(1.9,8,"6", col="blue")
table(apply(mse.grid,1,which.min)) #get Probability values

#ROBPCA
summary(z2.mse.h)
summary(z3.mse.h)
summary(z4.mse.h) #NULL
summary(z5.mse.h) #NULL
###Summarize resulting PCs for all reps
table(hub.numPCs) #get frequencies (Three-22, Four-78)

par(pty="s")
par(mfrow=c(1,3), mai=c(0.7, 0.7, 0.2, 0.1))

barplot(table(hub.numPCs), ylab="Frequency", xlab="Total Number of Components",
ylim=c(0,100), main="ROBPCA Results", col="azure")
text(0.7,25,"22", col="blue"); text(1.9,81,"78", col="blue")

#For Summarized Scree

eigs.hub <- cbind.data.frame(m.h.eig1, m.h.eig2, m.h.eig3,
c(m.h.eig4, rep(NA,length(m.h.eig1)-length(m.h.eig4)))
); eigs.hub
write.csv(eigs.hub, file="~/SEQ1/SEQ1_eigsHUB.csv")

PC.h <- c(1,2,3,4)
z2fit.h <- lm(colMeans(eigs.hub, na.rm=TRUE)~PC.h + I((PC.h-2)*ifelse(PC.h>=2,1,0)))
z3fit.h <- lm(colMeans(eigs.hub, na.rm=TRUE)~PC.h + I((PC.h-3)*ifelse(PC.h>=3,1,0)))

#For plotting with 4 resulting PCs (since last result only had 3)
plot(PC.h, colMeans(eigs.hub, na.rm=TRUE), type="o", main="ROBPCA Scree Plot",
ylim=c(0,3),
xlab="Component", ylab="Eigenvalue", lwd=2, axes=FALSE)
axis(side=1, at=seq(1,4,by=1))
axis(side=2, at=seq(0,2.5,by=0.5))
lines(PC.h, z2fit.h$fitted.values, lty=2, lwd=2, col="red")
lines(PC.h, z3fit.h$fitted.values, lty=2, lwd=2, col="blue")
text(3,1.2,"2", col="red"); text(3, 0.5,"3", col="blue")

mses.h <- cbind(z2.mse.h, c(z3.mse.h, rep(NA,length(z2.mse.h)-
length(z3.mse.h))))}; mses.h
write.csv(mses.h, file="~/SEQ1/SEQ1_msesROBPCA.csv")
mse.hub <- read.csv("~/SEQ1/SEQ1_msesROBPCA.csv")
mse.hub<-na.omit(mse.hub); mse.hub#delete rows with NA

barplot(table(apply(mse.hub,1,which.min)), ylim=c(0,80),
       main="ROBPCA (4-component results)", col="gray85",
       xlab= expression(paste(gamma, " ,component breakpoint at min. MSE")),
       ylab="Success Frequency")
       text(0.7,15,"12 (15.38%)", col="blue"); text(1.9,69,"66 (84.62%)",
       col="blue")

       table(apply(mse.hub,1,which.min)) #Get probability values

sessionInfo()
Appendix C: R Script for Technique Performance Based on Eigenvalue Thresholds for PCA on Data Sequence B

```r
# EIGENVALUE THRESHOLDS
### CLASSICAL PCA
# SEQ1
pca.eig1 <- read.csv("~/SEQ1/eigsPCA.csv")
# Success Pr of 2 Pcs
sum(ifelse(pca.eig1$m.pca.eig2>=1 & pca.eig1$m.pca.eig3<1,1,0))
out.pca1 <- matrix(NA, nrow=100, ncol=2)
for(i in 1:100){
  out.pca1[i,] <- t(c(i/50, mean(ifelse(pca.eig1$m.pca.eig3<i/50 & i/50<pca.eig1$m.pca.eig2,1,0))))
}
summary(out.pca1) # maximum success (V2) = 1
summary(subset(out.pca1, out.pca1[,2] == 1)) # range of V1 values with maximum success
plot(out.pca1[,1], out.pca1[,2], main="Classical PCA SEQ1",
     xlab=expression(~Lambda), ylab="Success Probability")
lines(supsmu(out.pca1[,1], out.pca1[,2]))
abline(v=0.9, lty=3); text(0.9, 0.2, "0.90")
abline(v=1.34, lty=3); text(1.34, 0.2, "1.420")
write.csv(out.pca1, "~/SEQ1/out.pca1.csv")
```