

# Contents

<b>1</b>	<b>Principal Component Analysis (PCA)</b>	<b>2</b>
1.1	Introduction . . . . .	2
<b>2</b>	<b>High Dimensionality for Data</b>	<b>3</b>
<b>3</b>	<b>Data Reduction</b>	<b>3</b>
3.1	Variable Redundancy . . . . .	4
3.2	Data Reduction in Exploratory Graphics . . . . .	6
<b>4</b>	<b>Principal Component Analysis</b>	<b>6</b>
<b>5</b>	<b>Communality</b>	<b>7</b>
5.1	Communalities . . . . .	7
5.2	PCA Terminology . . . . .	7
<b>6</b>	<b>Principal Component Analysis</b>	<b>8</b>
6.1	Mathematical background of PCA . . . . .	8
<b>7</b>	<b>Orthogonal versus Oblique Solutions</b>	<b>9</b>
<b>8</b>	<b>Testing PCA Analysis Validity</b>	<b>9</b>
8.1	Sampling adequacy (KMO Statistic) . . . . .	9
8.2	Bartlett's Test for Sphericity . . . . .	9
<b>9</b>	<b>Review of Important Definitions</b>	<b>11</b>

# 1 Principal Component Analysis (PCA)

- Is a variable reduction technique
- Is used when variables are highly correlated
- Reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables
- Is a large sample procedure
- The number of components extracted is equal to the number of observed variables in the analysis.
  - \* The first principal component identified accounts for most of the variance in the data.
  - \* The second component identified accounts for the second largest amount of variance in the data and is uncorrelated with the first principal component and so on.
- The total amount of variance in PCA is equal to the number of observed variables being analyzed. In PCA, observed variables are standardized, (e.g., mean=0, standard deviation=1).
- Components accounting for maximal variance are retained while other components accounting for a trivial amount of variance are not retained. Eigenvalues indicate the amount of variance explained by each component. Eigenvectors are the weights used to calculate components scores.

## 1.1 Introduction

- Principal component analysis is appropriate when you have obtained measures on a number of (possibly) correlated observed variables and wish to develop a smaller number of artificial uncorrelated variables called **principal components** that will account for most of the variance in the observed variables.
- The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The principal components may then be used as predictor or criterion variables in subsequent analyses.
- Traditionally, principal component analysis is performed can be performed on raw data, on the symmetric **Covariance matrix** or on the symmetric **Correlation matrix**. (The covariance matrix contains scaled sums of squares and cross products. A correlation matrix is like a covariance matrix but first the variables, i.e. the columns, have been standardized.)
- If raw data is used, the procedure will create the original correlation matrix or covariance matrix, as specified by the user. If the correlation matrix is used, the variables are standardized and the total variance will equal the number of variables used in the analysis

(because each standardized variable has a variance equal to 1). If the covariance matrix is used, the variables will remain in their original metric. However, one must take care to use variables whose variances and scales are similar. Unlike **factor analysis**, which analyzes the common variance, the original matrix in a principal components analysis analyzes the total variance. Also, principal components analysis assumes that each original measure is collected without measurement error.

## 2 High Dimensionality for Data

- Advances in data collection and storage capabilities during the past decades have led to an information overload in most sciences. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis.
- Such datasets, in contrast with smaller, more traditional datasets that have been studied extensively in the past, present new challenges in data analysis.
- Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation. The dimension of the data is the number of variables that are measured on each observation.

## 3 Data Reduction

- Data Reduction or Dimensionality Reduction pertains to analytic methods (typically multivariate exploratory techniques such as Factor Analysis, Multidimensional Scaling, Cluster Analysis, Canonical Correlation, or Neural Networks) that involve reducing the dimensionality of a data set by extracting a number of underlying factors, dimensions, clusters, etc., that can account for the variability in the (multidimensional) data set.
- For example, in poorly designed questionnaires, all responses provided by the participants on a large number of variables (scales, questions, or dimensions) could be explained by a very limited number of "trivial" or artifactual factors. For example, two such underlying factors could be: (1) the respondent's attitude towards the study (positive or negative) and (2) the "social desirability" factor (a response bias representing a tendency to respond in a socially desirable manner).

### 3.1 Variable Redundancy

A specific (but fictitious) example of research will now be presented to illustrate the concept of variable redundancy introduced earlier. Imagine that you have developed a 7-item measure of job satisfaction. The *instrument* is reproduced here:

Please respond to each of the following statements by placing a rating in the space to the left of the statement. In making your ratings, use any number from 1 to 7 in which 1=strongly disagree and 7=strongly agree.

----- 1. My supervisor treats me with consideration.

----- 2. My supervisor consults me concerning important decisions that affect my work.

----- 3. My supervisors give me recognition when I do a good job.

----- 4. My supervisor gives me the support I need to do my job well.

----- 5. My pay is fair.

----- 6. My pay is appropriate, given the amount of responsibility that comes with my job.

----- 7. My pay is comparable to the pay earned by other employees whose jobs are similar to mine.

Notice that items 1-4 all deal with the same topic: the employees satisfaction with their supervisors. In this way, items 1-4 are somewhat redundant to one another. Similarly, notice that items 5-7 also all seem to deal with the same topic: the employees satisfaction with their pay. Empirical findings may further support the notion that there is redundancy in the seven items. Assume that you administer the questionnaire to 200 employees and compute all possible correlations between responses to the 7 items. The resulting fictitious correlations are reproduced in the table below.

When correlations among several variables are computed, they are typically summarized in the form of a correlation matrix, such as the one reproduced in the previous table. This is an appropriate opportunity to review just how a correlation matrix is interpreted.

The rows and columns of the table correspond to the seven variables included in the analysis: Row 1 (and column 1) represents variable 1, row 2 (and column 2) represents variable 2, and so forth. Where a given row and column intersect, you will find the correlation between the two corresponding variables.

- For example, where the row for variable 2 intersects with the column for variable 1, you find a correlation of 0.75; this means that the correlation between variables 1 and 2 is 0.75. The correlations show that the seven items seem to hang together in two distinct groups.
- First, notice that items 1-4 show relatively strong correlations with one another. This could be because items 1-4 are measuring the same construct. In the same way, items 5-7 correlate strongly with one another (a possible indication that they all measure the same construct as well).

Correlations among Seven Job Satisfaction Items

Variable	Correlations						
	1	2	3	4	5	6	7
1	1.00						
2	.75	1.00					
3	.83	.82	1.00				
4	.68	.92	.88	1.00			
5	.03	.01	.04	.01	1.00		
6	.05	.02	.05	.07	.89	1.00	
7	.02	.06	.00	.03	.91	.76	1.00

Note : N = 200.

Figure 1: Questionnaire

- Even more interesting, notice that items 1-4 demonstrate very weak correlations with items 5-7. This is what you would expect to see if items 1-4 and items 5-7 were measuring two different constructs.
- Given this apparent redundancy, it is likely that the seven items of the questionnaire are not really measuring seven different constructs; more likely, items 1-4 are measuring a single construct that could reasonably be labelled *satisfaction with supervision* while items 5-7 are measuring a different construct that could be labelled *satisfaction with pay*.
- If responses to the seven items actually displayed the redundancy suggested by the pattern of correlations, it would be advantageous to somehow reduce the number of variables in this data set, so that (in a sense) items 1-4 are collapsed into a single new variable that reflects the employees satisfaction with supervision, and items 5-7 are collapsed into a single new variable that reflects satisfaction with pay.
- We could then use these two new artificial variables (rather than the seven original variables) as predictor variables in multiple regression, or in any other type of analysis.
- In essence, this is what is accomplished by principal component analysis: it allows you to reduce a set of observed variables into a smaller set of artificial variables called principal components. The resulting principal components may then be used in subsequent analyses.

## Variable Reduction Procedure

Principal component analysis is a variable reduction procedure. It is useful when you have obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, you believe that it should be possible to reduce the observed variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables.

### 3.2 Data Reduction in Exploratory Graphics

The term can also refer in **exploratory graphics** to “Data Reduction by unbiased decreasing of the sample size”. This type of Data Reduction is applied in exploratory graphical data analysis of extremely large data sets. The size of the data set can obscure an existing pattern (especially in large line graphs or scatterplots) due to the density of markers or lines. Then, it can be useful to plot only a representative subset of the data (so that the pattern is not hidden by the number of point markers) to reveal the otherwise obscured but still reliable pattern.

## 4 Principal Component Analysis

- PCA is defined as the orthogonal linear transformation that transforms data to a new co-ordinate system, such that the greatest variance by any projection of the projection of the data comes to lie on the first co-ordinate (called the first principal component).

$$Y^T = X^T W = V \Sigma$$

where  $V \Sigma W^T$  is the singular value decomposition (SVD) of the data matrix  $X^T$

- This is a Technique used to reduce multidimensional data sets to lower dimensions for analysis. PCA is also known as KARhunen Loeve Transform, Hotelling Transform, and Proper Orthogonal Decomposition.
- It is mostly used as tool in Exploratory Data Analysis and used for predictive models.
- PCA involves calculation of Eigenvalue Deomposition/ Single Value Decomposition for a data set, usually mean centering the data for each attribute.
- It is mathematically defined as an orthogonal linear transformation that transforms the data to a new co-ordinate system, such that the greatest variance by any projection of the data comes to lie on the first co-ordinate (called the first principal component) the second greates variance ont eh second princiapl component anf so on.

## 5 Communalities

Communality refers to the total amount of variance an original variable shares with all other variables included in the analysis. This is the proportion of each variable's variance that can be explained by the principal components. (It is denoted as  $h^2$  and can be defined as the sum of squared factor loadings).

**Initial** - By definition, the initial value of the communality in a principal components analysis is 1.

**Extraction** - The values in this column indicate the proportion of each variable's variance that can be explained by the principal components. Variables with high values are well represented in the common factor space, while variables with low values are not well represented. They are the reproduced variances from the number of components that you have saved. You can find these values on the diagonal of the reproduced correlation matrix.

### 5.1 Communalities

## Principal Component Analysis

### 5.2 PCA Terminology

- PC loadings are correlation coefficients between the PC scores and the original variables.
- PC loadings measure the importance of each variable in accounting for the variability in the PC. It is possible to interpret the first few PCs in terms of 'overall' effect or a 'contrast' between groups of variables based on the structures of PC loadings.
- high correlation between PC1 and a variable indicates that the variable is associated with the direction of the maximum amount of variation in the dataset.
- More than one variable might have a high correlation with PC1. A strong correlation between a variable and PC2 indicates that the variable is responsible for the next largest variation in the data perpendicular to PC1, and so on.
- if a variable does not correlate to any PC, or correlates only with the last PC, or one before the last PC, this usually suggests that the variable has little or no contribution to the variation in the dataset. Therefore, PCA may often indicate which variables in a dataset are important and which ones may be of little consequence. Some of these low-performance variables might therefore be removed from consideration in order to simplify the overall analyses.

## Characteristics of Principal Components

The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables. Under typical conditions, this means that the first component will be correlated with at least some of the observed variables. It may be correlated with many.

The second component extracted will have two important characteristics. First, this component will account for a maximal amount of variance in the data set that was not accounted for

by the first component. Again under typical conditions, this means that the second component will be correlated with some of the observed variables that did not display strong correlations with component 1.

The second characteristic of the second component is that it will be uncorrelated with the first component. Literally, if you were to compute the correlation between components 1 and 2, that correlation would be zero.

The remaining components that are extracted in the analysis display the same two characteristics: each component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components, and is uncorrelated with all of the preceding components. A principal component analysis proceeds in this fashion, with each new component accounting for progressively smaller and smaller amounts of variance (this is why only the first few components are usually retained and interpreted). When the analysis is complete, the resulting components will display varying degrees of correlation with the observed variables, but are completely uncorrelated with one another.

## 6 Principal Component Analysis

### 6.1 Mathematical background of PCA

- Principal Component Analysis is a linear **dimensionality reduction** technique, which identifies orthogonal directions of maximum variance in the original data, and projects the data into a lower-dimensionality space formed of a sub-set of the highest-variance components (Bishop, 1995).
- The mathematical technique used in PCA is called **eigen analysis**. Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables.
- Software packages compute solutions for these weights by using a special type of equation called an **eigenequation**. The weights produced by these eigenequations are optimal weights in the sense that, for a given set of data, no other set of weights could produce a set of components that are more successful in accounting for variance in the observed variables.
- The weights are created so as to satisfy a principle of least squares that is similar (but not identical) to the principle of least squares used in multiple regression.

Remarks

- The words **linear combination** refer to the fact that scores on a component are created by adding together scores on the observed variables being analyzed.
- **Optimally weighted** refers to the fact that the observed variables are weighted in such a way that the resulting components account for a maximal amount of variance in the data set.



## 7 Orthogonal versus Oblique Solutions

- This course will discuss only principal component analysis that result in **orthogonal solutions**. An orthogonal solution is one in which the components remain uncorrelated (orthogonal means uncorrelated).
- It is possible to perform a principal component analysis that results in correlated components. Such a solution is called an **oblique solution**. In some situations, oblique solutions are superior to orthogonal solutions because they produce cleaner, more easily-interpreted results.
- However, oblique solutions are also somewhat more complicated to interpret, compared to orthogonal solutions. For this reason, we will focus only on the interpretation of orthogonal solutions

## 8 Testing PCA Analysis Validity

### 8.1 Sampling adequacy (KMO Statistic)

- Measured by the **Kaiser-Meyer-Olkin (KMO)** statistics, sampling adequacy predicts if the analyses are likely to perform well, based on correlation and partial correlation. KMO can also be used to assess which variables to drop from the model because they are too multi-collinear.
- There is a KMO statistic for each individual variable, and their sum is the KMO overall statistic. KMO varies from 0 to 1.0 and KMO overall should be 0.60 or higher to proceed with PCA analysis.
- Values below 0.5 imply that PCA may not be appropriate. (*Approach to overcoming this: If it is not, drop the **indicator variables** with the lowest individual KMO statistic values, until KMO overall rises above 0.60.*)
- Kaiser-Meyer-OlkinTo compute KMO overall, the numerator is the sum of squared correlations of all variables in the analysis (except the 1.0 self-correlations of variables with themselves, of course). The denominator is this same sum plus the sum of squared partial correlations of each variable  $i$  with each variable  $j$ , controlling for others in the analysis. The concept is that the partial correlations should not be very large if one is to expect distinct factors to emerge from PCA analysis.

### 8.2 Bartlett's Test for Sphericity

- Bartlett's measure tests the null hypothesis that the original correlation matrix is an identity matrix.
- For PCA and factor analysis to work we need some relationships between variables and if the R- matrix were an identity matrix then all correlation coefficients would be zero.
- Therefore, we want this test to be significant (i.e. have a significance value less than 0.05).

- A significant test tells us that the correlation matrix is not an identity matrix; therefore, there are some relationships between the variables we hope to include in the analysis. For these data, Bartlett's test is highly significant ( $p < 0.001$ ), and therefore PCA analysis is appropriate.

## 9 Review of Important Definitions

- An observed variable can be measured directly, is sometimes called a measured variable or an indicator or a manifest variable.
- A principal component is a linear combination of weighted observed variables. Principal components are uncorrelated and orthogonal.
- A latent construct can be measured indirectly by determining its influence to responses on measured variables. A latent construct could also be referred to as a factor, underlying construct, or unobserved variable.
- Factor scores are estimates of underlying latent constructs.
- Unique factors refer to unreliability due to measurement error and variation in the data.
- Principal component analysis minimizes the sum of the squared perpendicular distances to the axis of the principal component while least squares regression minimizes the sum of the squared distances perpendicular to the x axis (not perpendicular to the fitted line).
- Principal component scores are actual scores.
- Eigenvectors are the weights in a linear transformation when computing principal component scores. Eigenvalues indicate the amount of variance explained by each principal component or each factor.
- Orthogonal means at a 90 degree angle, perpendicular. Oblique means other than a 90 degree angle.
- An observed variable **loads** on a factor if it is highly correlated with the factor, has an eigenvector of greater magnitude on that factor.
- Communalities are the variance in observed variables accounted for by a common factor. Communality is more relevant to EFA than PCA.