**CS655 – Machine Learning  
Week 5 Assignment – Dimensionality Reduction**

**Student Name: Zhiyang (Madeline) Lin**

**Class and Section...CS655**

**Total Points...60 points**

**1. Describe the four criteria for choosing how many components to extract. Explain the rationale for each.**

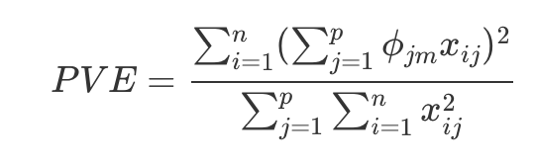
Four criteria for choosing how many components to extract are:

1) The Eigenvalue Criterion

It is the sum of the eigenvalues is equal to the number of variables entered into the PCA (Principal Component Analysis). The eigenvalues will range from greater than one to near zero. An eigenvalue of 1 means that the principal component would explain about one variable’s worth of the variability. The rationale for this is that each component should explain at least one variable’s worth of the variability; therefore, the eigenvalue criterion states that only components with eigenvalues greater than 1 should be retained.

2) The Proportion of Variance Explained Criterion (“PVE”)

The PVE identifies the optimal number of principal components (“PCs”) to keep based on the total variability that we would like to account for. The formula of PVE for the m-th PC is calculated as:



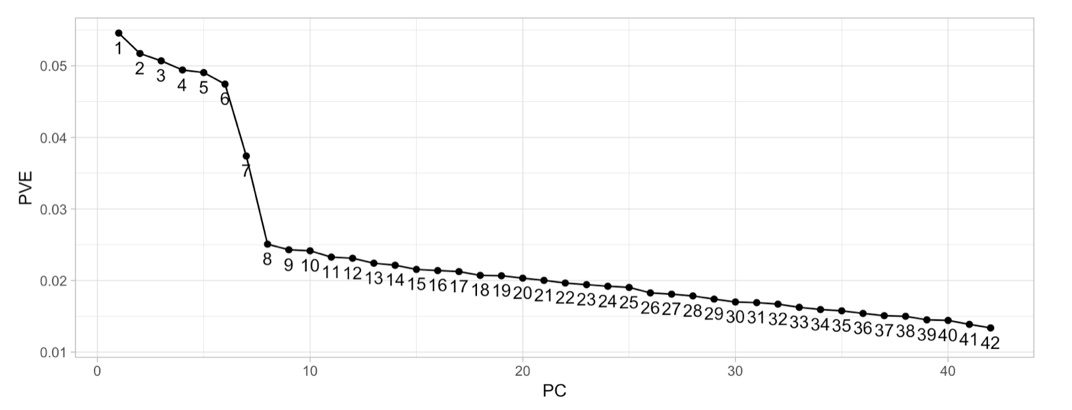
PVE criterion keeps all PCs that are above or equal to a pre-specified threshold of total variability explained. The reasonable amount of variability varies by application and the data being used. For instance, when the PCs are used for descriptive purposes only (i.e. customer profiling), the proportion of variability explained may be lower than otherwise. When the PCs are used as derived features for models downstream, the PVE should be as much as can conveniently be achieved, given any constraints.

3) The Minimum Communality Criterion

The minimum communality criterion states that enough components should be extracted so that the communalities for each of these variables exceeds a certain threshold. First, it identifies the set of variables are to be retained. Second, it calculates the communalities of these variables which should exceed a certain threshold.

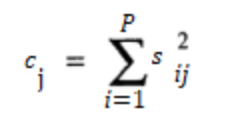
4) The Scree Plot Criterion

The scree plot shows the eigenvalues or PVE for each individual PC. Most scree plots look broadly similar in shape, starting high on the left, falling rather quickly, and then flattening out at some point. This is because the first component usually explains much of the variability, the next few components explain a moderate amount, and the latter components only explain a small fraction of the overall variability. The scree plot criterion looks for the “elbow” in the curve and selects all components just before the line flattens out. An example which shows PVE is shown as below.



**2. Explain the concept of communality, so that someone new to the field could understand it.**

Communality is the proportion of variance of a particular variable that is shared with other variables. It represents the overall importance of each of the variables in the PCA as a whole. A variable’s communality ranges from 0 to 1. A variable that does not have any unique variance at all (i.e. one with explained variance that is 100% a result of other variables) has a communality of 1. A variable with variance that is completely unexplained by any other variables has a communality of 0. The calculation of communality in PCA is the sum of squares of the component weights across all components. Specifically, it is the formula as below.



Where Cj = communality of the j-th variable, Sij = loading (or correlation) between the i-th component and the j-th variable.

Usually, communalities less than 0.5 are considered too low.

**3. Explain the difference between principal components analysis and factor analysis. What is a drawback of factor analysis?**

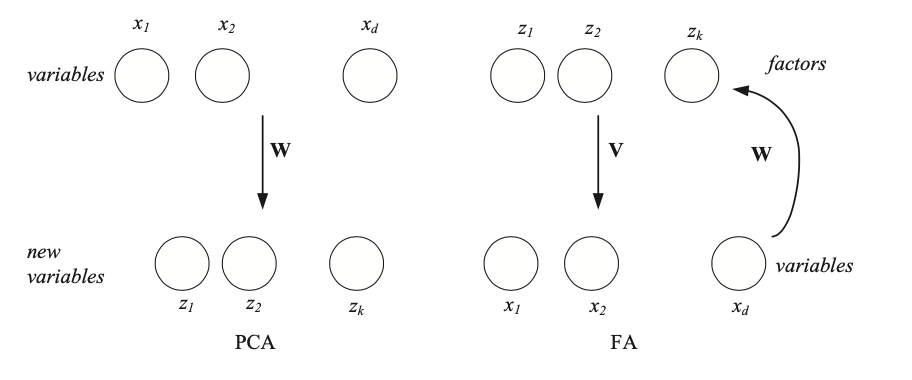
Principal components analysis (“PCA”) is an unsupervised learning method for dimensionality reduction. Its goal is to find a mapping from the inputs in the original d-dimensional space to a new (k<d)-dimensional space, with minimum loss of information.

Factor analysis (“FA”) is another unsupervised learning method for dimensionality reduction. Its goal is to ask whether the data that is observed can be explained by a smaller number of uncorrelated factors or latent variables.

The mathematics of FA and PCA are different.

PCA seeks to identify variables that are composites of the observed variables. We are trying to find a low-dimensional space such that when x is projected there so that the information loss is minimized

FA explicitly assumes the existence of latent factors underlying the observed data. We are trying to find a small number of factors z, which are combined to generate x :



The drawback of factor analysis is that it makes assumption that the input features are linearly related to one another. This means the factor analysis may not perform well on sets of features that are not linearly related. Also, factor analysis assumes bivariate normal between each pair of variables which can be problematic in situations where the input features are not normally distributed. Additionally, factor analysis does not handle categorical data, outliers, missing data well and cannot produce meaningful output if variable are not correlated.

**4. Explain why we perform factor rotation. Describe three different methods for factor rotation.**

Factor rotation is the repositioning of factors to a newer, more interpretable configuration by a set of mathematically unique and specific transformations. A rotation is to minimize the number of factors needed to explain each variable. Rotations are performed iteratively and on every pair of factors.

The reason why we perform factor rotation is to ease interpretability by simplifying the rows and columns of the matrix of factor loadings and optimize (maximize or minimize) the objective criterion simultaneously for all the factors that would be mathematically difficult.

Three different methods for factor rotation are as below.

Quartimax: a method which tends to rotate the axes so that the variables have high loadings for the first factor, and low loadings thereafter. It minimizes the number of factors needed to explain a variable.

Varixmax: a method which maximizes the variability in the loadings for the factors, with a goal of working toward the ideal column of zeroes and ones for each variable.

Equimax: a method to seek to compromise between simplifying the columns and the rows. It can be seen as a method sharpening some properties of varimax.

**5. Determine whether the following statements are true or false. If false, explain why the statement is false, and how one could alter the statement to make it true.**

**a. Positive correlation indicates that, as one variable increases, the other variable increases as well.**

True

Because a positive correlation is a relationship between two variables that tend to move in the same direction. A positive correlation exists when one variable tends to decrease as the other variable decreases, or one variable tends to increase when the other increases.

**b. Changing the scale of measurement for the covariance matrix, for example from meters to kilometers, will change the value of the covariance.**

True

Covariance is defined as the expected value of variations of two variables from their expected values. Simply speaking, covariance measures how much variables change together. The absolute value of covariance depends on the unit of variables. For example, if we change the unit of a variable from kilometer to meter unit, then the deviance from mean of 1 in kilometer units (km) is changed into 1,000 in meter units (m). The unit change makes huge difference in the value of covariance, even when the relationship of two variables is the same.

**c. The total variability in the data set equals the number of records.**

False

Variability refers to how spread out a group of data is. It is not the sum of the number of records.

To find the total variability in our group of data, we simply add up the deviation of each score from the mean. The average deviation of a score can then be calculated by dividing this total by the number of scores.

**d. The value of the ith principal component equals the ith eigenvalue divided by the number of variables.**

False

The value/variance for the ith principal component is equal to the ith eigenvalue.

**e. The second principal component represents any linear combination of the variables that accounts for the most variability in the data, once the first principal component has been extracted.**

False

The first principal component is the linear combination of x-variables that has maximum variance (among all linear combinations).  It accounts for as much variation in the data as possible.

The second principal component is the linear combination of x-variables that accounts for as much of the remaining variation as possible, with the constraint that the correlation between the first and second component is 0.