Principal Component Analysis for Dimensionality Reduction in Educational Research

Nonhlanhla Luphade, Manala Tyobeka

Abstract

In recent years, there has been high interest on how big data can be exploited to benefit the educational sector. This influx of information allows for better informed decision-making, however, it does comes with its own set of problems. The complexity of high dimensional data sets make it hard to process, visualise and interpret such sets. There is also the problem of multiple variables in the data set being highly correlated with one another, which results in the effects of certain variables being incorrectly estimated. Dimensionality reduction techniques have been developed to alleviate such problems while preserving most of the relevant information in the data. This paper attempts to scrutinize Principal Component Analysis (PCA), a method commonly used for dimensionality reduction in educational research. The conventional approach to formulating PCA is generally sensitive to outliers, missing data and non-linearity in the data set. This paper further explores alternative approaches to PCA (i.e. robust PCA, probabilistic PCA and kernel PCA) which enable the user to reduce the impact of either of these problems which may exist in the data set, and lastly determine whether these approaches have any kind of benefit to educational research. The results in this paper do not suggest an overall winning technique, rather, the results indicate the need for closer attention to be paid to the nature of data sets in order to select the most appropriate PCA method to improve the quality of results.

Keywords
Dimensionality Reduction — PCA — Robust PCA — Probabilistic PCA — Kernel PCA

Contents

1. Introduction

The collection of large and diverse datasets has created a great challenge in data analysis. These datasets tend to have significant amounts of redundancies (Hegde, 2016). To efficiently manipulate data represented in high-dimensional space and address the impact of redundant dimensions on the final results, we consider projecting the data onto a lower dimensional subspace using the characteristics of the original variables, without losing important information (Houari u. a., 2016).

Principle Component Analysis (PCA) is a technique commonly used in educational research for dimensionality reduction. It does this by finding linear orthogonal combinations of the input variables that are ordered by decreasing variance, while retaining the most valuable parts of the variables. However, this technique does have its shortfalls i.e. (1) it is sensitive to outliers, (2) it requires a complete input data matrix, and (3) it assumes that the relationship between variables in the input data matrix can be captured by means of linear projections.

This paper attempts to explore the performance of PCA on data describing secondary school students' achievement in Mathematics from two Portuguese schools, and compare the results to robust PCA, probabilistic PCA and kernel PCA. Three alternative PCA techniques which attempt to address the three shortfalls of standard PCA mentioned earlier, respectively.

2. Literature Review

This section of the paper considers PCA: details of how it works and the shortfalls of this technique. It further explores three alternative methods which attempt reduce the impact of certain issues a data set may present which may negatively affect results obtained in standard PCA. The alternative methods include robust PCA, probabilistic PCA and kernel PCA.

2.1 Standard PCA

The main linear technique for dimensionality reduction is Principal Component Analysis. PCA was developed by Hotelling (1933), as a method used for describing the variance-covariance structure of variables, by deriving a reduced set of orthogonal, linear projections of correlated variables (Izenman, 2008).

Suppose that X is a matrix of order $(n \times p)$, where the rows represent n samples, and the columns represent the p variables. Generally, n is larger than p. This is the framework in which we will be discussing PCA and other techniques in this paper. However, there is nothing heeding the number of variables from exceeding the number of samples, this is often the case in the analysis of microarray and chemometric problems (Hawkins u. a., 2001).

The p variables are responsible for reproducing the total variability in the data set. Oftentimes, much of the variability in the data set can be accounted for by k components, such that k is less than p. If that is the case, then the p variables can be replaced by the k components, resulting in the original data set of n samples measured on p variables being reduced to a data set of n samples being measured on k components (Johnson u. a., 2002).

The assumption made when formulating these k components, which are generally called principal components, is that they are particular linear combinations of the p variables. These linear combinations constitute a new coordinate system derived by rotating the original system. In this new coordinate system, the axes constitute the directions with maximum variability, as well as provide a more manageable interpretation of the variance-covariance structure of the data set (Johnson u. a., 2002).

The estimation of coefficients in each of the principal components depends on the variance-covariance structure of the p variables in the data set. Let C_{XX} be the variance-covariance structure matrix for X, then

$$C_{XX} = V\Lambda V^T,$$

where V is an orthogonal matrix of eigenvectors (i.e. each column is an eigenvector) and Λ is a diagonal matrix of eigenvalues, λ_i, in decreasing order along the diagonal. The eigenvectors are called the principal axes of the data (i.e. the direction of the axes after rotation which constitute the directions with maximum variability). PCA makes use of Singular Value Decomposition (SVD) to factorise C_{XX}. Let

$$X = USV^T,$$

where U is the orthogonal matrix of left-singular vectors (i.e. the unitary matrix), S is the diagonal matrix of singular values, and V is the orthogonal matrix of right-singular vectors (i.e. principal axes). From here it can be shown that:

$$C_{XX} = \frac{1}{n-1}(USV^T)^T USV^T = \frac{1}{n-1}VS^2V^T,$$

which indicates that the orthogonal matrix of right-singular vectors, V, contains the principal directions, and the singular values are related to the eigenvalues of the variance-covariance matrix such that: $\lambda_i = \frac{s_i^2}{n-1}$. The principal components are given by:

$$XV = USV^TV = US.$$

This follows, since the eigenvectors in V are orthogonal to one another.

PCA uses a least-squares error criterion as a measure of how well the variance-covariance structure can be reconstructed by the linear combinations of a set of p variables in the data set. This is as result of the *Eckart-Young Theorem* which states that:
"If $\mathbf{A}$ and $\mathbf{B}$ are both $(J \times K)$-matrices, and we plan on using B with reduced rank $r(\mathbf{B}) = b$ to approximate $\mathbf{A}$ with full rank $r(\mathbf{A}) = min(J,K)$, then

$$\lambda_j((\mathbf{A}-\mathbf{B})^T(\mathbf{A}-\mathbf{B})) \geq \lambda_{j+b}(\mathbf{A}\mathbf{A}^T),$$

with equality if,

$$\mathbf{B} = \sum_{i=1}^{b} \lambda_i^{1/2}\mathbf{u}_i\mathbf{v}_i^T,$$

where,

- $\lambda_i = \lambda(\mathbf{A}\mathbf{A}^T)$,
- $u_i = \mathbf{v}_i(\mathbf{A}\mathbf{A}^T)$,
- $v_i = \mathbf{v}_i(\mathbf{A}^T\mathbf{A})$.

Because the above choice of $\mathbf{B}$ provides a simultaneous minimisation for all eigenvalues λ_j, it follows that the minimum is achieved for difference functions of those eigenvalues, say, the trace or the determinant of $(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^T$. "(Izenman, 2008)"

As a result, SVD is highly sensitive to outliers. How does this effect PCA? In the case where there is a sufficiently outlying individual cell (i.e. an anomaly is the data set or a cell which has been incorrectly recorded), this individual cell may draw selected principal components towards itself (Hawkins u. a., 2001).

Besides being sensitive to outliers, another shortfall to the conventional approach to PCA is that it requires a complete data set (i.e. no missing values in the data set). This may not always be the case in reality, due to shortcomings in data collection processes (Hawkins u. a., 2001).

There is also the assumption that relationships in the data set are linearly separable when conducting PCA, this too may not always be the case. The remainder of this chapter explores possible solutions to these shortfalls in standard PCA.

2.2 Robust PCA

Robust PCA is an alternative approach to PCA, which makes use of a more robust SVD, as opposed to the standard SVD used in PCA, for the analysis of projecting the original data set onto a rotated set of axes which account for the maximum variability in the data set. Attempts have been made to construct a more robust SVD. We begin by exploring the efforts of Gabriel and Zamir of estimating a more robust SVD, and then conclude with an extension of Gabriel and Zamir's proposal by Hawkin.

Let X be a matrix. Gabriel und Zamir (1979) attempt to address the problem of the sensitivity to outliers of the SVD, using a method called Alternating Least-squares approach (ALS). This method is an iterative procedure which attempts to approximate matrices by other matrices of lower-rank. Let $(n \times p)$ be the dimensions of matrix X, ALS considers a least squares fitting subject to weights, w_{ij}. Fitting a matrix of lower-rank to X is equivalent to fitting by a matrix product, AB^T. Weights are assigned to rows and columns in the new matrices. These weights are inversely dependent on the row and column sums of the squared residuals. The weighted least squares is then used in each iterative step until convergence. The minimising criterion for this iterative process can be described as follows:

$$\phi(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{n}\sum_{j=1}^{p} w_{ij}(x_{ij} - a_i b_j^T)^2,$$

where a_i and b_j are the rows of matrix A and B, respectively. This method has been shown to reduce to effect of outliers by means of manipulating the weightings of the observations. To deal with matter of missing entries in the data, this method assigns zero weights.

Hawkins u. a. (2001) suggest that despite ALS succeeding in reducing the effect of outliers, ALS is still a least-squares procedure. Therefore, attempts could be made to further reduce the impact of outliers. The method used to do so is called ”AL1-SVD”. It is similar to ALS , in that this approach uses an iterative approach to approximate a matrix by other matrices of lower-rank. However, in attempt to generate a more robust SVD, the criterion in which the leading eigenvector-and-eigenvalue pairs have the property of minimising a Euclidean norm of the unexplained variation of the data matrix in ALS, is replaced by a criterion in which the pair has the property of minimising the L1 norm of the unexplained variation of the data matrix. This is possible because the L1 norm is generally more robust to outliers, the Euclidean norm tends to drastically inflate the influence of outliers (i.e. $||.||_1$ vs. $||.||_2$).

When compared to ALS and conventional formulations of the SVD, AL1-SVD returns eigenvectors which are generally not orthogonal to one another. Further, the eigenvalues may not follow a descending order which is common in ALS and conventional formulations of the SVD (Hawkins u. a., 2001). Therefore, careful attention needs to be paid when interpretting the output of a PCA which makes use of robust SVD.

2.3 Probabilistic PCA

Probabilistic PCA is a dimensionality reduction technique which describes a data structure by means of lower dimensional latent space. It attempts to formulate the standard PCA within a Gaussian latent variable model which is closely related to statistical factor analysis (Tipping und Bishop, 1999). Let x_i be a p-dimensional observation vector corresponding to a k-dimensional vector of latent variables, t_i. Assuming that the relationship is linear, by standard Factor Analysis we have that:

$$x_i = Wt_i + \mu_i + \varepsilon_i,$$

where W is a $(p \times k)$-dimensional matrix relating the two sets of variables, μ_i allows the model to have a non-zero mean, and ε_i is an error term.

Generally, we assume that $t_i \sim N(0,I)$, meaning that the latent variables are assumed to normally distributed with unit variance, and that each latent variable is independent of one another. Further, we assume that $\varepsilon_i \sim N(0,\Gamma)$. It then follows that the observations in the data set correspond to some Gaussian distribution, i.e. $x_i \sim N(\mu, WW^T + \Gamma)$. Using this, we can construct the probabilistic framework for PCA. Assume an isotropic Gaussian noise model, $N(0,\sigma^2 I)$, for ε_i. This implies that the distribution of the observations given some latent space, t_i can be given by:

$$x_i|t_i \sim N(Wt_i + \mu_i, \sigma^2 I),$$

where the marginal distribution of our observations is:

$$x_i \sim N(\mu_i, WW^T + \sigma^2 I).$$

In formulating the PCA within this probabilistic framework, principal components can be estimated using maximum likelihood estimation, as if they were parameters. This may be computationally expensive. Therefore, as an alternative, the estimation of parameters of the latent variable model can be done by means of an iterative, and computationally efficient algorithm: Expectation-Maximisation

(EM) algorithm, for effective PCA (Dempster u. a., 1977). Probabilistic PCA is not only useful for dimensionality reduction, but also enables PCA projections to be obtained when some data values are missing in the data set. By framing the PCA within a Gaussian probabilistic framework, we are able to estimate possible missing values in the data set.

2.4 Kernel PCA

Standard PCA attempts to reduce the dimensionality of a data set, by describing the variance-covariance matrix structure of a set of variables in the data set, using a reduced set of linear combinations of the variables (Schölkopf u. a., 1997). However, in the real world, linear projections of a set of variables in a data set may not sufficiently capture the nature of the variance-covariance structure. Kernel PCA provides a non-linear means of describing the variance-covariance structure a data set.

According to Vapnik-Chervonenkis theory, the mapping of data from its original data space to a higher dimensional space allows for improved classification power. However, high dimensional mapping can be computationally expensive. The kernel trick provides means in which to conduct this mapping whilst reducing expense on computational processes. The kernel trick is represented as follows:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j).$$

PCA is made non-linear by means of letting n observations, x_i, measured against p variables in the data set be mapped onto a non-linear feature space, $\phi(x_i)$. The dimensions of this feature space are such that they are greater than the dimensions of the original data space.

Using the kernel trick, for any new observation x, its projection onto the principal components is:

$$y_j(x) = \phi(x)^T v_j = \sum_{i=1}^{n} \alpha_{ji} K(x, x_i),$$

where,

- $v_j = \sum_{i=1}^{n} \alpha_{ji} \phi(x_i)$ is an eigenvector of the covariance matrix in the feature space,
- α_{ji} is a vector of coefficients corresponding to v_j.

The justification for such an approach is that in some cases, the input in a data set may not be linearly separable. However, when projected onto a higher dimensional feature space, it becomes easier to extract relevant information in the data set using PCA, since the non-linear relationships which may have existed appear almost linear in a higher dimensional feature space. Once the principal components are obtained from this higher dimensional feature space, the result will be non-linear in the original data space (Wang, 2012).

3. Exploratory Data Analysis

3.1 Data description

This project uses a complete data set obtained from Kaggle called ”student-mat”. The dataset has 395 observations measured over 33 attributes. The data describes student achievement in secondary education of two Portuguese schools in their Mathematics course. The data attributes include grades, demographic, social and school related features. It was collected using school reports and questionnaires.

The target attribute, ”G3”, describes what student got in their third and final term for Mathematics (out of 20). For this paper, the variable G3 was divided into five groups:

- **Group 1**: students that obtained above 16.
- **Group 2**: students that obtained between 13 and 16 inclusive.
- **Group 3**: students that obtained between 9 and 12 inclusive.
- **Group 4**: students that obtained between 5 and 8 inclusive.
- **Group 5**: students that obtained below 5.

3.2 Preparation and Exploration

More than half of the variables in the data set are categorical. These variables were encoded as dummy variables, because PCA requires a numerical input matrix to conduct the analysis. The conversion of these variables into dummy variables increased the number of variables in the data set to 42.

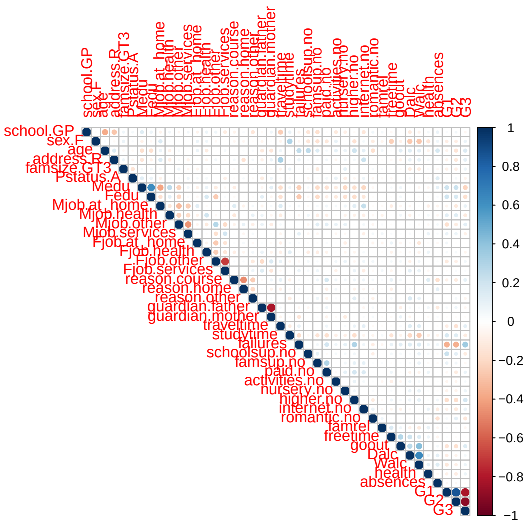

Figure 1. Correlation plot for student maths

Figure 1 indicates the correlations between all 42 variables in the new data set. From the plot, we see significant correlations between variables in the data set. PCA provides a means to deal with highly correlated variables in the data set. It does this by generating a set of orthogonal principal components. Variables which are correlated with one another are loaded onto the same principal component.

To perform PCA, the data set was randomly divided into a training set (70% of the observations) and a testing set (30% of the observations). Table 1 indicates that the distribution of the classes in the target variable of the training set is fairly similar to that of the original data set.

	1	2	3	4	5
Original	24	107	162	63	39
Training	13	75	117	43	28

Table 1. Distribution of classes in the target variable

The Mahalanobis distance calculates the distance of observations in the data set relative to the centroid (i.e. a point in the multivariate space where all means from all the variables intersect) when variables may be correlated with one another. It allows us to find possible outliers in the data set.

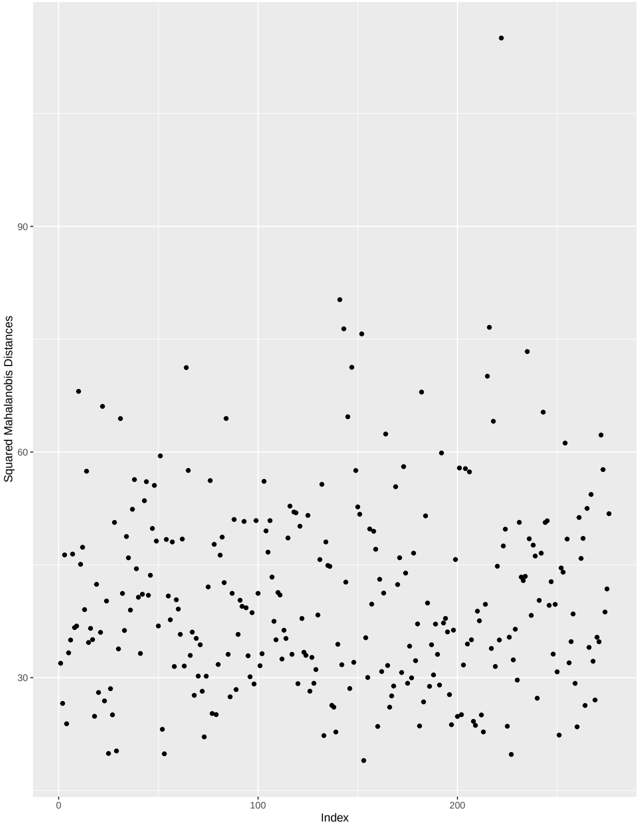

Figure 2. Mahalanobis squared distances of observations in training set

The larger the Mahalanobis distance is, the further away from the centroid an observation is. Figure 2 indicates that there is at least one significant outlier in the training set. PCA allows us to visualise possible outliers in the data set. However, in the presence of influential outliers, outliers may draw principal components towards themselves which may weaken the classification abilities of principal components on unseen data.

4. Methodology

The section of the paper describes the various PCA techniques employed, each technique's algorithm, expected output and how the performance of each method is measured.

4.1 Standard PCA

Let X be our training data matrix of dimensions (276×41), which describes student achievement in secondary education of two Portuguese schools in their Mathematics course. PCA is a linear dimensionality reduction technique that reduces the dimensions of an input data matrix X to k principal components using SVD. We use SVD to factorise X, such that X is the product of matrices of lower dimensions:

$$X = USV^T$$

where U is a unitary matrix of dimension $(276 \times r)$, S is a diagonal matrix of singular values of dimension $(r \times r)$, and V is the matrix containing the principal axes responsible for the rotation of the data matrix of dimension $(r \times 41)$. r is the rank of X, and since there exists correlation between variable (see figure 1), we can expect that $r < 41$.

From the decomposition of X, we can obtain the principal components which are given by:

$$XV = US$$

These principal components are linear combinations of the variables in the data matrix which explain the variation in X. PCA tries to load maximum possible information in the first principle component, then maximum remaining information in the second and so on (Wold u. a., 1987).

Using the first two principal components which explain the most variation in the data obtained from the PCA, we will be able to visualise the data in a scatter plot of the multidimensional projections of the data onto the components. From this scatter plot, we will be able to observe relationships and the nature of these relationships which may exist within the data set (i.e. the formation of clusters in data), which principal components are responsible for distinguishing between clusters, the amount of variance in the data set is captured by each principal component, and detecting the presence of outliers in the data set.

4.2 What if the data has outliers?

It is a well-known issue that the standard PCA is sensitive to outliers because the calculation of the sample mean and covariance can be significantly influenced by a small number of outliers (Fowler u. a., 2019). To mitigate this problem, Candès u. a. (2011) proposed robust PCA. This method is called robust PCA because of its ability to recover the underlying low-rank structure in the data more accurately when there are outliers in the data set.

Robust PCA is very similar to standard PCA. The main difference is that a more robust SVD is used to factorise X (the training set), such that X is the product of matrices of lower dimensions, as opposed to the conventional SVD. The approximation of a robust SVD is computed using an alternating L1 norm (as opposed to the more usual least squares Euclidean norm used in conventional SVD).

The decomposition of the X allows us to obtain the principal components. However, these components are not in descending order of amount of variation explained by each component. A principal component which explains more variance may follow a component which explains less variance. Therefore, careful attention needs to paid when selecting appropriate principal components.

Once the two principal components which explain the most variance in the data set are selected, we will be able to visualise data in a scatter plot of the multidimensional projections of the data on the components. The performance of robust PCA in visualising the data matrix will be compared to standard PCA.

4.3 What if the data has missing values?

PCA cannot be performed on a data matrix if it has missing values. To address this limitation, Tipping und Bishop (1999) proposed a probabilistic formulation of the PCA. This formulation incorporates Expectation-Maximisation approach for PCA with a probabilistic model. It formulates the standard PCA within a Gaussian latent variable model, such that the latent variables and the noise are normally distributed. Recall that the data set describing students' achievement in mathematics is complete. 39 students (i.e. approximately 10% of observations in the data set) were randomly selected, and readings from these observations were deleted. In so doing, we are able to observe whether the estimation of missing values in probabilistic PCA yields similar results to standard PCA.

Within this framework, we also obtain parameters. In fact, the principal axes are one of those parameters. Maximum likelihood estimation is the obvious choice for parameter estimation, however, due to there being no closed-form analytic solution for some of the parameters in the model, these values must be obtained by some iterative procedure - EM algorithm.

Once the first two principal components which explain the most variance in the data set are selected, we will be able to visualise data in a scatter plot of the multidimensional projections of the data on the components. The performance of probabilistic PCA in visualising the data matrix will be compared to standard PCA.

4.4 What if the data is linearly inseparable?

PCA performs linear transformations on a given data set, however, many real-world data sets are not linearly separable. If the data cannot be linearly transformed within the original data space, standard PCA will generally not be very helpful (García-González u. a., 2020). Kernel PCA is the generalisation of the standard PCA for non-linear dimensionality reduction, which better explains complicated spatial structures of high dimensional features (Wang, 2012).

Assume that the data set is linearly inseparable, and therefore cannot be separated effectively using a standard PCA. To mitigate this problem, the kernel trick method is employed (Schölkopf u. a., 1997). This method involves picking a kernel: the three commonly used kernels are Gaussian, polynomial and hyperbolic tangent. The selection of a kernel is usually based on some prior knowledge of the nature of the data set. In the absence of this knowledge, a cross-validation process would have to be conducted to select the most appropriate kernel and its hyper-parameters which maximise one's chances of achieving a certain goal, without overfitting on the training set.

Once the kernel is selected, we construct a normalised kernel matrix of the training set data. By doing this, we are able to map the data to higher dimensional space where the data is linearly separable. In this higher dimensional space, we are able to factorise the variance-covariance structure, i.e. solve the eigenvalue problem. The projection of any data point, x_j, onto the principal components is given by:

$$y_j = \sum_{i=1}^{n} \alpha_{ji} K(x_j, x_i), j = 1, ..., d$$

The standard PCA is extended to enable classification of non-linear data using the ”kernel trick”. By using this trick, we are able to project the original data from the training set into a higher dimensional space without sacrificing too much computational time (Wang, 2012).

For this project, kernel PCA is performed on the training set. A gaussian kernel of the form:

$$K(x_j, x_i) = exp(\sigma||x_j - x_i||^2)$$

will be used. To select the most appropriate sigma (i.e. a sigma which maximises the amount of variance captured by the principal components), a grid search is performed. A scatter plot of multidimensional projections of the data onto the first two principal components which explain the most variation in the data is plotted. The performance of the kernel PCA in visualising the data matrix will be compared to standard PCA

4.5 Further assessment of performance of PCA methods

All the algorithms mentioned in this section are dimensionality reduction methods. The main aim of these algorithms is to reduce the dimensions of the training set. To further evaluate the performance of the principal components, we need to test them on unseen data.

To do that, the PCA models were used to to reduce the dimensions of the test set. The performance of the PCA methods are the evaluated by using a kNN-classifier on the unobserved lower-rank test data. This means that we train a kNN-classifier on the lower-rank training set, which involves storing the lower-rank training set and class labels of the training samples. Thereafter, we use classifier to predict on the test set. The number clusters, k, is defined by the user. Since students were divided into five groups, k was chosen to be 5. The lower-rank test set is assigned to the nearest group. Since the data is continuous, Euclidean distance is used to measure nearness of an observation to a cluster.

The performance of robust PCA, probabilistic PCA and kernel PCA was compared to standard PCA. The performance measure used is the ability of these methods to correctly classify observations from the unseen test data.

5. Results & Discussion

As mentioned previously in the paper, PCA is a commonly used technique for dimension reduction in educational research. PCA makes use of SVD to decompose the structure of the data in such a way that still preserves relevant information in variance-covariance structure of a set of variables. In so doing, it assumes that this information can be efficiently captured by means of linear combinations of the set of variables (1). Performing SVD generally requires that the date set be complete (2), and that the presence of outliers be kept to a minimum, if not none, for more accurate results (3).

This section of the paper attempts to implement PCA, and compare its performance to PCA using robust SVD, PPCA and KPCA on the training set data. For the sake of comparison, the performance of these techniques will be measured based on their ability visualise the structure of the data set and nature of relationships which may exist using first two principal components which capture the highest variability in the data structure. These principal components will then be used to build kNN-classifiers to predict on unseen data, to observe which of these methods is more efficient in minimising the misclassifications.

5.1 Standard PCA

Standard PCA was implemented using the ‘prcomp()‘ function in R. The calculation for standard PCA is done by singular value decomposition of the data matrix, as opposed to an eigen-decomposition of the variance-covariance matrix.

Since the aim is to maximise the amount of variability in the data structure that is captured by the first two principal components, the PCA was conducted on variables in the training set which were neither centered nor scaled. The first two principal components captured 95.7% of the variability in the data structure. Whereas, the first two principal components obtained from PCA on centered data captured 83.2% of the variability, and PCA on scaled and centered data captured only 15.7% of the variability. Based on the amount of variability captured by PCA on the data matrix which was neither centered nor scaled, it is worth investigating how well these principal components separate the performance of students in mathematics in the third term.

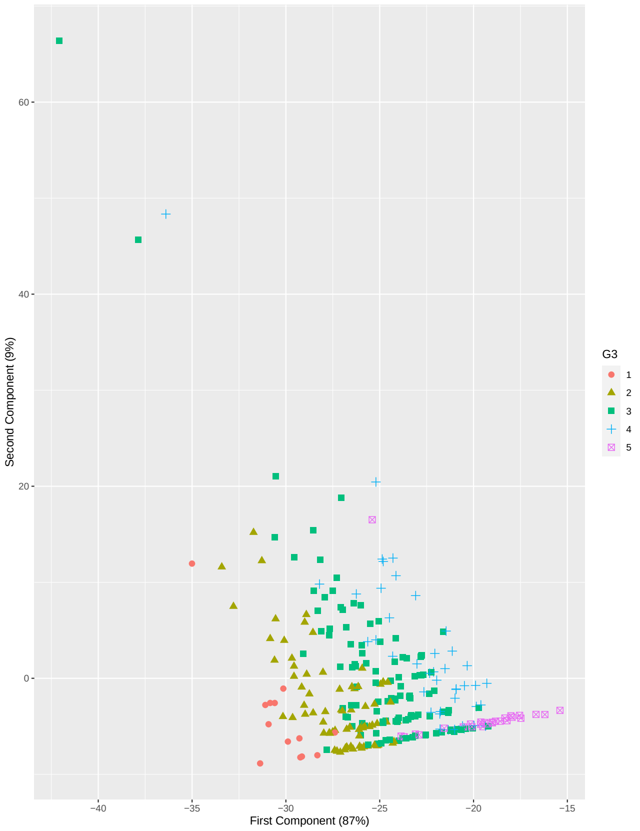

Figure 3. Scatterplot of observations against the first and second components found in PCA

Figure 3 indicates which clusters the observations in the data set fall in to. Along the first component there is some overlap between the clusters. However, we see that higher grades lie on the left-end of the first component, whereas students with lower grades lie on the right-end of the first component. As the first component increases, we notice a decrease in the grades. In the case of the second component, we see that it fails to distinguish between the varying clusters.

PCA is a useful technique for identifying outliers in multivariate data structures which cannot be visualised in 2-or-3-dimensional spaces. In figure 3, we note three outliers upper-most left-corner of the plot which have been assigned to clusters 3 and 4 (i.e. students with averages between 20% and 60% in the third term). In the event that these observations were correctly recorded, these observations may be students who may have shared similar backgrounds to students who performed exceptionally well in the third term, however, did not manage to get the grades in the third term to be classed in clusters 1 and 2 (i.e. students with grades between 60% and 100%).

5.2 Robust PCA

Robust PCA is implemented using the ‘pca()‘ function as a wrapper function, and indicating that the method to be used in this function is ”robustPCA” in R. This approach to PCA is robust to outliers. Further, it can also handle missing values. However, it is not meant to be used for missing value estimation. The scores calculated for missing values in the data set are set to zero.

Unlike standard PCA, robust PCA makes use of a robust SVD to get an accurate estimation of coefficients of the principal components for a data matrix with outliers. In the event of missing values in the data structure, one must expect the accuracy of the estimation to decrease.

As mentioned earlier, the order of eigenvalues of the robust SVD may not follow a descending order, which is common in conventional formulations of the standard SVD. Therefore, when attempting to reduce the dimensionality of a data structure, careful attention must be made when selecting principal components to ensure that the maximum variability in the data structure is captured. A robust PCA was conducted on the training set. The first and the third component obtained from robust PCA captured the most variability in the data set (i.e. 96.9% of the variability in the data set was captured in these two principal components).

The results in Figure 4 appear similar to that of standard PCA. Along the first component there is some overlap between the clusters. However, we see that observations of students who got higher grades for the third term lie on the right-end of the first component, whereas students with lower grades lie on the left-end of the first component. As the first component increases, we see an increase in the grades. The third component, however, fails to distinguish between the varying clusters.

Like standard PCA, we also see three outliers which have assigned to clusters 3 and 4, only this time they are in the top-right corner of the plot. It is worth noting, that in the presence of outliers, there is an improvement in the amount of variability captured by robust PCA compared to standard PCA. One would expect that as the number of outliers increases and/or magnitude of the distance of outliers from the remainder of the sample increases, the robust PCA would do a better job at capturing the variability of the data set.

5.3 Probabilistic PCA

Probabilistic PCA was implemented using the ‘pca()‘ function as a wrapper function, and indicating that the method to be used in this function is ”ppca” in R. The implementation of this function allows us to perform PCA on incomplete data, as the missing values in the data set can be estimated.

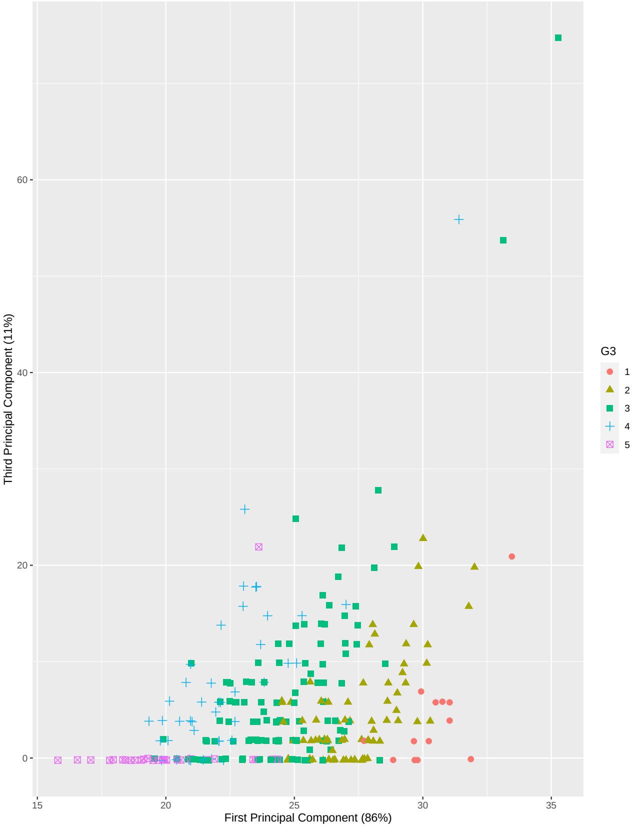

Figure 4. Scatterplot of observations against the first and second components found in Robust PCA

This function fills the initial matrix of coefficients for principal components (i.e. loadings matrix) with random numbers chosen from a Gaussian normal distribution. Therefore, we are required to run this function multiple times to ensure that the estimates for the loadings matrix have converged to the most accurate values. The amount of variability in the data set captured by the first two components is 83.2%.

The results in Figure 5 appear similar to that of standard PCA. Along the second component there is some overlap between the clusters. However, we see that with higher grades lie on the lower-end of the second component, whereas students with lower grades lie on the upper-end of the second component. Despite the first component capturing more variability in the data structure, the second component is more successful in distinguishing between the different clusters.

Recall that when this PPCA was constructed measurements for 10% of the sample were removed from the data set. Using PPCA, we were able to capture 83.2% of the variability in the data structure using the first two principal components, that is 12.5% less than standard PCA. It is clear, that using PPCA to estimate missing values in the data set does not necessarily do better than standard PCA on a complete data set, however, the estimation of missing values in PPCA does not significantly reduce the amount of

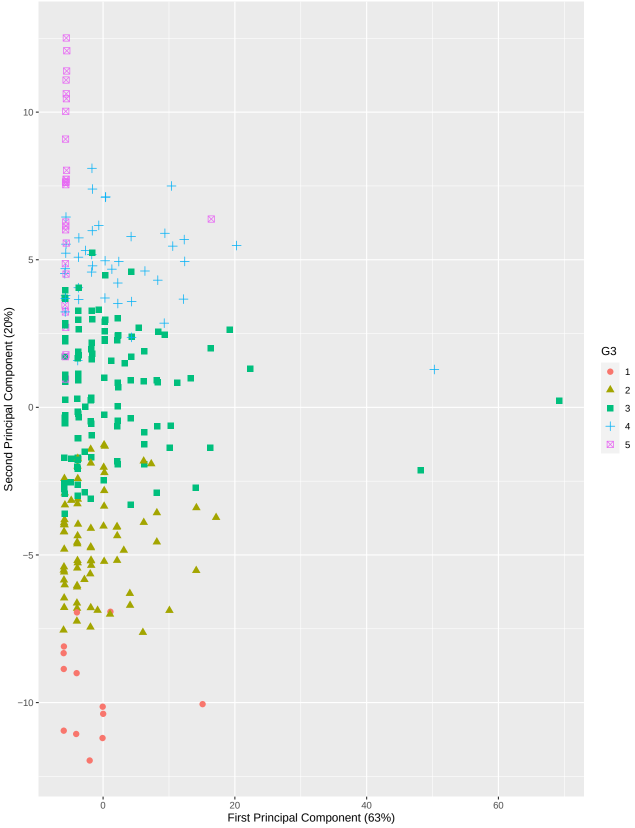

Figure 5. Scatterplot of observations against the first and second components found in PPCA

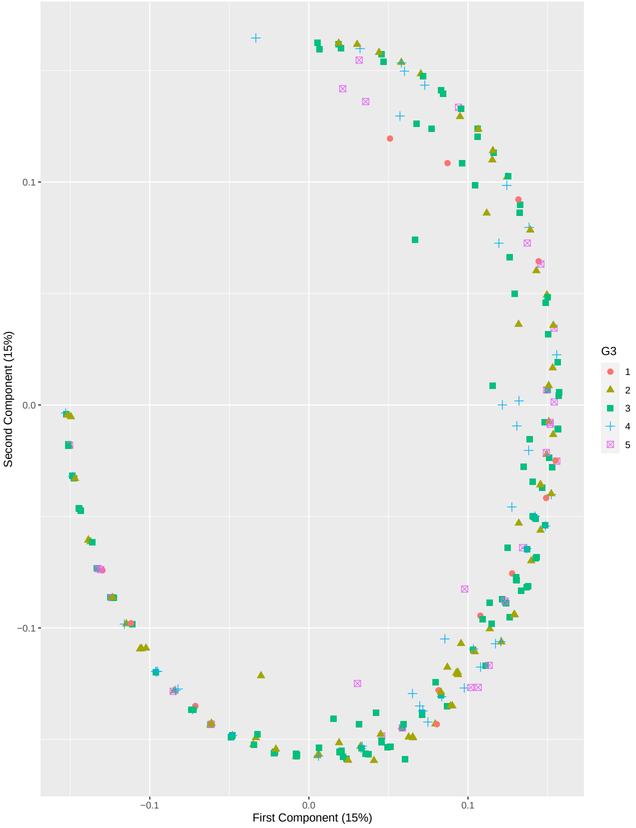

Figure 6. Scatterplot of observations against the first and second components found in KPCA

variability captured by the components.

5.4 Kernel PCA

Kernel PCA is implemented using the ‘kpca‘ function in R. KPCA is a nonlinear form of principal component analysis. It makes use of kernel to transform the data matrix to a higher dimensional space where PCA is performed. When transformed back to the original data space, the principal components are able to classify non-linear relationships in the data matrix.

The kernel function employed was the Radial Basis (i.e. Gaussian) function. A kind of grid search was conducted to determine the appropriate value for the hyper-parameter of this function which maximises the amount variability captured by the first two principal components:

Sigma	
100	10
1	0.1
0.001	0.0001
0.00001	0.0000001

Table 2. Values for grid search to determine appropriate inverse kernel width

From figure 6, we see that the first two components obtained from KPCA using a Gaussian kernel function captured 29.7% of the variation in the data matrix. From the plot, we see that the first two components failed to distinguish between the clusters in the data set. We are, therefore; within reason to conclude that perhaps the assumption of possible non-linearity does not hold in this instance. Hence, the poor performance of visualising the data matrix see in figure 6

5.5 Predictions on Unseen Data

We now look at the performance of the principal components selected in the techniques above on unseen data. Using the principal components derived from the training set, kNN classifiers we built for each techniques with the exception of KPCA.

Recall that for KPCA, the eigenvectors generated come from the kernel Hilbert space. As a result, unseen data points cannot be transformed when we do not have any knowledge of the explicit mapping function, ϕ, that is used. Instead, we conduct KPCA on the entire data set to obtain a transformed the data set, the classifier is built using the transformed training set and we observe the accuracy of the model on the transformed test set.

The accuracy of standard PCA is 66.39%, this is the reference point that we will use to compare results obtained in the other methods. The accuracy for each method can be seen in table 3:

Method	PCA	RPCA	PPCA	KPCA
Accuracy (in %)	66.39	63.87	75.63	41.18

Table 3. Test set classification performance of the two-dimensional results of all of the techniques.

5.6 Discussion

When using two-dimensions, robust PCA is able to capture more of the variation in the data set than standard PCA in the presence of outliers. By using the L1 norm as a minimising criterion for estimating loadings matrix, robust PCA reduces the impact of outliers whilst still being able to distinguish between the different clusters in the data set.

Recall that, even though the original data did not have any missing values, we decided to randomly remove values from 10% of observations in the data set. This was done with the intention of observing how PCA could be improved in the presence of missing data. Since standard PCA requires that missing values either be removed or imputed. By framing PCA with some probabilistic framework, we were able to estimate values for missing entries. Despite there being missing values, we saw that PPCA was able to account for a significant amount of variation in the data structure. Even though it was not as high as standard PCA, we were still able to distinguish between the different clusters in the data set.

Kernel PCA provided us a means to visualise the data set within non-linear framework. This was not very successful for this specific data set. However, this method is clearly worth considering when assumptions of linearity do not hold in the data set.

Based on the second half of the results which involved building kNN-classifiers on the principal components obtained from the various PCA techniques on unseen data. We see that robust PCA performs slightly worse-off than standard PCA in classifying observations from an unseen data set. In the scatterplots of training observations against the first and second components found in both robust PCA (4) and standard PCA (3), firstly, we see that there exists some overlap between the different groups in the data which negatively impacted both methods' ability to correctly classify observations from unseen observations. Secondly, we noted the presence of at least three possible outliers in the data set, which account for 1.09% of the data in the training set. Perhaps, there may have been some overfitting with the robust PCA resulting in this method performing worse off than standard PCA.

When comparing probabilistic PCA to standard PCA, we see that the probabilistic PCA results in a higher accuracy rate. This comes as a surprise, since there were missing values in the data set which were estimated. Perhaps, the performance of PCA could have been improved had we performed a grid search to determine an appropriate k (i.e. number of neighbours considered) on some validation set. However, it is worth noting how our assumption of a Gaussian distribution framework in which to conduct PCA did not significantly worsen the performance of the components on unseen data.

Kernel PCA did a poor job at assigning members in the test set to clusters. This comes as no surprise, since the assumption of non-linearity did not hold for this data set.

6. Conclusion

The aim of this project was to explores PCA as a dimensionality reduction method and other alternative approaches to PCA which enable the user to reduce the impact of problems which may exist in the data set and determine whether these approaches have any kind of benefit to educational research. The results of the project show that standard PCA is an effective method for dimensionality reduction. However, it also encourages users of this technique to pay closer attention to the nature of their data sets, what it is that they would like to achieve from the analysis, and what shortfalls in their data set might weaken the results obtained from standard PCA.

In the case of wanting to minimise the impact of outliers, robust PCA may be a more preferred alternative as opposed to simply just using standard PCA. In the presence of missing data, which is very common in real-life data perhaps using the probabilistic PCA may be more ideal when little information is available to inform removing or imputing data in order to conduct a standard PCA. Lastly, in the case where linear projections fail to capture the structure and relationships between variables in a data set, it is worth considering applying a Kernel PCA in order to reduce the dimensionality of the data structure.

References

[Candès u. a. 2011] Candès, Emmanuel J. ; Li, Xiaodong ; Ma, Yi ; Wright, John: Robust principal component analysis? In: *Journal of the ACM (JACM)* 58 (2011), Nr. 3, S. 1–37

[Dempster u. a. 1977] Dempster, Arthur P. ; Laird, Nan M. ; Rubin, Donald B.: Maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1977), Nr. 1, S. 1–22

[Fowler u. a. 2019] Fowler, JW ; Alpert, BK ; Joe, Y-I ; O'Neil, GC ; Swetz, DS ; Ullom, JN: A robust principal component analysis for outlier identification in messy microcalorimeter data. In: *Journal of Low Temperature Physics* (2019), S. 1–9

[Gabriel und Zamir 1979] Gabriel, K R. ; Zamir, Shmuel: Lower rank approximation of matrices by least squares with any choice of weights. In: *Technometrics* 21 (1979), Nr. 4, S. 489–498

[García-González u. a. 2020] GARCÍA-GONZÁLEZ, Alberto ; HUERTA, Antonio ; ZLOTNIK, Sergio ; DÍEZ, Pedro: A kernel Principal Component Analysis (kPCA) digest with a new backward mapping (pre-image reconstruction) strategy. In: *arXiv preprint arXiv:2001.01958* (2020)

[Hawkins u. a. 2001] HAWKINS, Douglas M. ; LIU, Li ; YOUNG, S S.: Robust singular value decomposition. In: *National Institute of Statistical Science Technical Report* 122 (2001)

[Hegde 2016] HEGDE, Vinayak: Dimensionality reduction technique for developing undergraduate student dropout model using principal component analysis through R package. In: *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* IEEE (Veranst.), 2016, S. 1–6

[Hotelling 1933] HOTELLING, Harold: Analysis of a complex of statistical variables into principal components. In: *Journal of educational psychology* 24 (1933), Nr. 6, S. 417

[Houari u. a. 2016] HOUARI, Rima ; BOUNCEUR, Ahcène ; KECHADI, M-Tahar ; TARI, A-Kamel ; EULER, Reinhardt: Dimensionality reduction in data mining: A Copula approach. In: *Expert Systems with Applications* 64 (2016), S. 247–260

[Izenman 2008] IZENMAN, Alan J.: *Modern multivariate statistical techniques*. Bd. 10. Springer, 2008. – 978–0 S

[Johnson u. a. 2002] JOHNSON, Richard A. ; WICHERN, Dean W. u. a.: *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2002

[Schölkopf u. a. 1997] SCHÖLKOPF, Bernhard ; SMOLA, Alexander ; MÜLLER, Klaus-Robert: Kernel principal component analysis. In: *International conference on artificial neural networks* Springer (Veranst.), 1997, S. 583–588

[Tipping und Bishop 1999] TIPPING, Michael E. ; BISHOP, Christopher M.: Probabilistic principal component analysis. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (1999), Nr. 3, S. 611–622

[Wang 2012] WANG, Quan: Kernel principal component analysis and its applications in face recognition and active shape models. In: *arXiv preprint arXiv:1207.3538* (2012)

[Wold u. a. 1987] WOLD, Svante ; ESBENSEN, Kim ; GELADI, Paul: Principal component analysis. In: *Chemometrics and intelligent laboratory systems* 2 (1987), Nr. 1-3, S. 37–52