http://dx.doi.org/10.1287/ics.2015.0016

14th INFORMS Computing Society Conference
Richmond, Virginia, January 11–13, 2015
pp. 212–225

Principal Component Analysis and Optimization: A Tutorial

Robert Reris and J. Paul Brooks
Systems Modeling and Analysis, Virginia Commonwealth University rerisra@vcu.edu, jpbrooks@vcu.edu

Abstract Principal component analysis (PCA) is one of the most widely used multivariate techniques in statistics. It is commonly used to reduce the dimensionality of data in order to examine its underlying structure and the covariance/correlation structure of a set of variables. While singular value decomposition provides a simple means for identification of the principal components (PCs) for classical PCA, solutions achieved in this manner may not possess certain desirable properties including robustness, smoothness, and sparsity. In this paper, we present several optimization problems related to PCA by considering various geometric perspectives. New techniques for PCA can be developed by altering the optimization problems to which principal component loadings are the optimal solutions.

Keywords principal component analysis, optimization, dimensionality reduction

A Brief History of Principal Components Analysis

In 1901, Karl Pearson published a paper titled "On Lines and Planes of Closest Fit to Systems of Points in Space" [26]. The paper is an exploration of ideas concerning an affine space that best fits a series of points, where the fit is measured by the sum of squared orthogonal distances from each point to the space. While Pearson was surely aware of the possibilities these ideas might present, he may not have been quite aware of just how impactful principal component analysis (PCA) has become in modern multivariate data analysis. It is now one of the most important and widely used statistical techniques in many various fields ranging from biology to finance.

A search of Google Scholar indicates about 1.5 million papers were published in the last ten years containing some mention of PCA. Conducting the same search for the decade ending 1990 yields about 164,000. As larger datasets become more pervasive and the 'big data' revolution continues its expansion, multivariate statistical methods such as PCA which aid in the efficiency with which data is stored and analyzed will undoubtedly continue to occupy an increasing proportion of attention in the research community.

While Pearson's work is most often considered to have laid the foundations of PCA, it may be argued that earlier ideas concerning the principal axes of ellipsoids and other quadratic surfaces also played a part in its development. Francis Galton [12] drew a connection between the principal axes and the "correlation ellipsoid". Beltrami [4] and Jordan [18] derived the singular value decomposition (SVD), which is directly related and is, in fact, the technique most often used to find principal components (PCs).

In the years following Pearson's famous paper, few works attempting to continue the development of his ideas in depth were published. Harold Hotelling's paper, "Analysis of a

Complex of Statistical Variables with Principal Components" was not published until thirty years later [15]. While Pearson's ideas were focused on the geometry of points and fitted subspaces, Hotelling was concerned with the idea of what might be referred to today as 'data reduction' - taking a set of variables and finding a smaller set of independent variables which determine the values of the original variables. He also presented ideas concerning ellipsoids of constant probability for multivariate normal distributions.

The ideas presented by Carl Eckart and Gale Young in their 1936 paper, "The Approximation of One Matrix By Another of Lower Rank" established a connection between PCs and the SVD [10]. They proved that the best low-rank approximation of a matrix in terms of minimizing the L_2 distance between points in this matrix and the original points is given by taking the diagonal matrix obtained in the decomposition and simply removing its smallest elements.

In this paper, we return to the geometric ideas concerning the derivation of PCs and present several different optimization problems for which PCs provide optimal solutions. We will show that PCs correspond to optimal solutions of rather dramatically different-looking optimization problems. Their simultaneous solution using PCs attests to the variety of useful properties of traditional PCA. New methods for PCA, including robust methods, can be derived by altering the original optimization problems.

In the next section, we review an application of PCA and illustrate it's versatility as a data analysis tool. Then, we discuss several optimization problems for which PCs are optimal solutions. We conclude by reviewing a sample of recently-proposed PCA approaches that may be viewed in terms of alternate formulations of the optimization problems.

1. Applying Principal Component Analysis

In this section, we use a single dataset to illustrate several of the many uses of PCA in a data analysis. We illustrate how PCA can be used for dimensionality reduction, ranking, regression, and clustering. The dataset, results of a series of road tests conducted by Motor Trend magazine in 1974, consists of 11 variables describing and quantifying various aspects of 32 automobiles from the 1973–74 model year [14]. Each observation of the dataset corresponds to a car model and each variable corresponds to a characteristic or road test result. The R code, interactive output, and comments for generating the results discussed here are available at `http://www.people.vcu.edu/~jpbrooks/pcatutorial` and `http://scholarscompass.vcu.edu/ssor_data/2`.

Let X denote the $n \times m$ dataset matrix whose rows correspond to automobile models and columns correspond to particular characteristics. For reasons explained below, we begin by centering and scaling the data by first subtracting the column-wise mean from each value and then dividing by the column-wise standard deviation. The output of PCA includes an $m \times m$ *rotation matrix* or *loadings matrix* V, an $n \times m$ matrix of *scores* Z, and the standard deviations of the projections along the PCs $\lambda_1, \ldots, \lambda_m$. The columns of the rotation matrix $v_1, \ldots, v_m$ define the PCs: the i^{th} PC defines the linear function $f(x) = v_i^T x$. The matrix Z contains the values of each PC function evaluated at each observation in the dataset: $Z = XV$.

Dimensionality reduction. Our dataset consists of 32 observations and 11 variables. For an analysis, the ratio of variables to observations is rather high. PCA can be used to reduce the dimensionality of the data by creating a set of derived variables that are linear combinations of the original variables. The values of the derived variables are given in the columns of the scores matrix Z. The scores for the first three PCs are plotted in Figure 1. Now that the dimensions of our dataset have been reduced from 11 to three, we can visualize relationships to see which models are similar to other models.

How much of the original information has been lost in this projection to three dimensions? The proportion of the information explained by the first three PCs is $(\sum_{k=1}^{3} \lambda_k^2)/(\sum_{k=1}^{m} \lambda_k^2)$.

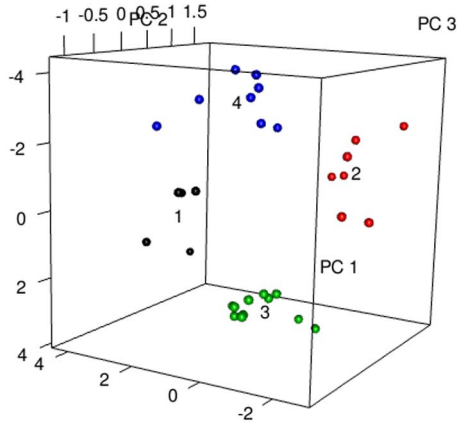

FIGURE 1. Plot of the scores of the automobile data on first three principal components. The colors of points indicate the membership of the points in clusters determined using cluster analysis. The numbers are plotted at the four cluster centroids.

Table 1 contains the standard deviations, proportion of variance explained, and cumulative proportion of variance explained for the eleven PCs. Approximately 60.0% of the variance of the data is captured by the first PC, and the second and third PCs account for 24.1% and 5.7%. So, with three suitably chosen variables we have accounted for about 90% of the information present in the original 11 variables.

Which variables in the original dataset contributed the most to this representation? We can interpret the loadings in the rotation matrix V. The loadings are given in Table 2. The columns of the rotation matrix have norm one. Their individual entries are the cosines of the angles formed by the original basis vectors and the loadings vectors. For example, $\cos(68.3°) = 0.37$ and $\cos(101.5°) = -0.20$, indicating the angles between the first eigenvector and the original basis vectors corresponding to cylinder quantity and 1/4 mile time are

TABLE 1. Principal component standard deviations, variance explained, and cumulative variance explained for the principal components for the Motor Trend dataset.

Output	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
S.D.	2.57	1.63	0.79	0.52	0.47	0.46	0.37	0.35	0.28	0.23	0.15
Var. Prop. (%)	60.0	24.1	5.7	2.5	2.0	1.9	1.2	1.1	0.7	0.4	0.2
Cum. Prop. (%)	60.0	84.2	89.9	92.3	94.4	96.3	97.5	98.6	99.3	99.8	100.0

TABLE 2. Principal component loadings for the Motor Trend dataset.

Variable	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}
Miles/gallon	−0.36	0.02	−0.23	−0.02	0.10	−0.11	0.37	−0.75	0.24	0.14	−0.12
No. cylinders	0.37	0.04	−0.18	−0.00	0.06	0.17	0.06	−0.23	0.05	−0.85	−0.14
Displacement	0.37	−0.05	−0.06	0.26	0.39	−0.34	0.21	0.00	0.20	0.05	0.66
Horsepower	0.33	0.25	0.14	−0.07	0.54	0.07	−0.00	−0.22	−0.58	0.25	−0.26
Rear axle ratio	−0.29	0.27	0.16	0.85	0.08	0.24	0.02	0.03	−0.05	−0.10	−0.04
Weight	0.35	−0.14	0.34	0.25	−0.08	−0.46	−0.02	−0.01	0.36	0.09	−0.57
1/4 mile time	−0.20	−0.46	0.40	0.07	−0.16	−0.33	0.05	−0.23	−0.53	−0.27	0.18
V/S	−0.31	−0.23	0.43	−0.21	0.60	0.19	−0.27	0.03	0.36	−0.16	0.01
Transmission	−0.23	0.43	−0.21	−0.03	0.09	−0.57	−0.59	−0.06	−0.05	−0.18	0.03
No. gears	−0.21	0.46	0.29	−0.26	0.05	−0.24	0.61	0.34	−0.00	−0.21	−0.05
No. carburetors	0.21	0.41	0.53	−0.13	−0.36	0.18	−0.17	−0.40	0.17	0.07	0.32

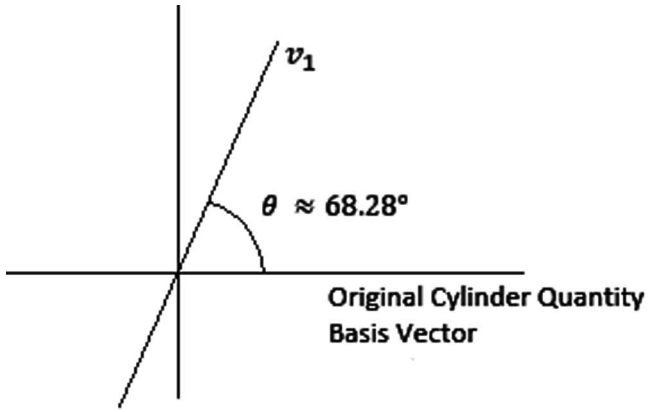

FIGURE 2. The angle between the first principal component loadings vector and the original cylinder quantity basis vector. ($\cos(68.3°) \approx 0.37$)

68.3° and 101.5°, respectively. Figure 2 illustrates this property for the number of cylinders. Combining this knowledge with that obtained from the score matrix yields more insights into the data. Models with larger scores on the first PC will likely correspond to those with more cylinders and carburetors and better fuel economy (lower values), as indicated by the sign and magnitudes of the loadings of these characteristics on the first component. Models with larger scores on the second PC will likely correspond to those with faster quarter mile times and a larger number of carburetors.

Note that the signs on the loadings and scores are not meaningful when considering one variable or observation at a time, but rather their signs relative to each other are. If two variables have opposite signs on one loadings vectors, then that PC reflects a trade-off between the two variables. If they have the same sign, then on that PC, the variables tend to increase and decrease together. Because the scores are given by inner products of the original data and the loadings vectors, the data that have positive scores will correspond to those with larger positive values for the variables with positive loadings and smaller (even negative) values for the variables with negative loadings.

Principal components regression. In addition to providing visualization capability, dimensionality reduction can also help with other downstream analyses. An example is *principal components regression* (PCR). The presence of collinearity between independent variables is detrimental to multiple linear regression (MLR) models derived by ordinary least squares [23]. It also causes a significant increase in the variance and instability of the coefficients, and minimizes the power of statistical tests of their significance.

For MLR, we seek an intercept β_0 and coefficients β to derive a model of the form

$$y = \beta_0 + X\beta. \tag{1}$$

where y is the dependent variable. If we first perform PCA, and regress on the derived variables using the scores, we get a model of the form

$$y = \beta_0 + Z\beta. \tag{2}$$

Hence the regression can be performed on the PC scores. Often PCs with the smallest standard deviations are eliminated until the usual metrics considered in regression such as adjusted R^2 or mean-squared-error reach a desirable level and once this happens, the coefficients obtained are transformed back onto the scale of the original variables. Since small PC standard deviations are associated with correlated independent variables, eliminating these PCs before performing the regression dampens the effects of multicollinearity. Consider

using PCR with the automobile data. In Table 1, the last six PCs provide only an additional 5% of explained variation, which indicates that there is likely multicollinearity in the data and that not all variables are needed to explain the variation in the data. One could take the first 3-5 PC scores and the original response and conduct an MLR.

There is a caveat when using PCR. While projecting points onto a smaller orthogonal subspace helps to ensure uncorrelated independent variables, there is no guarantee of a linear relationship between the principal components and the independent variable. Hence, after transformation, in choosing which components to retain it may be beneficial to consider the inclusion of components not solely based on their variance but also on the strength of their correlation with the response variable of interest.

Ranking. Suppose we wanted to combine the 11 automobile characteristics into a single measure that ranks automobile models. As we describe further below, the first PC is the direction along which the measurements are most varied in the variables that are most heavily loaded on it. The scores on the first PC are given in Table 3. Cars with large scores on the first PC tend to be those that have the most power and worst fuel economy, and those with negative scores tend to be smaller cars that get better gas mileage. As indicated by the loadings, it is along these factors that cars most distinguish themselves from one another. This can be confirmed by examining the loadings on the first component in Table 2. We might expect 'power' cars to have more cylinders, horsepower, and weight. We'd also expect these types of cars to have slower quarter mile time and get worse gas mileage. Hence we might associate the first component with 'power' and use the scores on this component to tell us exactly how much of a 'power car' each data point might be. A higher positive score indicates a power car while a lower negative score might indicate a smaller, possibly more fuel efficient car.

The other PCs may be examined similarly. Notice that the largest three loadings (in absolute value) on the first component are the smallest on the second, and are nearly zero. Hence, once the degree to which each data point is a 'power' car has been accounted for, loadings on this component indicate there is unique direction of significant variation along which data points contain a large amount of horsepower, rear axle ratio, transmission, number of gears, and number of carburetors while being lightweight and slow with a straight engine. Whatever quality one chooses to ascribe these characteristics, the scores on the second component would then tell us the strength of that particular quality in each data point, and so on for the other components.

Cluster analysis. We can use this dataset to observe clusters in the data. Refer again to Figure 1. Note that in cluster 4 (blue points), the Ford Pantera, Maserati, and Ferrari are clustered together. These are three high end, high performance sports cars. Cluster 3 (green points) contains a group of smaller vehicles, such as the Toyota Corolla, Honda Civic, and Fiat 128. Plotting the scores on the first two or three PCs is a common way to validate a cluster analysis, and can be used as a tool for defining clusters visually.

The relationship between PCA and cluster analysis has been formalized in an interesting way by Zha et al. [28] and Ding and He [9]. They show that the scores on the first $k-1$ principal components are the optimal solution to the continuous relaxation of a formulation for minimum sum of squares clustering for k clusters (where integer variables indicate cluster membership).

Total least squares regression. Finally, suppose we would like to build a model to predict one of the characteristics, miles per gallon, of a particular model using the other ten characteristics. An MLR via ordinary least squares would set the data from the miles per gallon column as the independent variable and minimize the sum of squared *vertical* distances of points to a fitted hyperplane. We can use the output of PCA for *total least squares regression* that minimizes the sum of squared *orthogonal* distances of points to the fitted hyperplane. Figure 3 illustrates geometrically the differences between total least

TABLE 3. Principal component scores on the first principal component for the Motor Trend dataset.

Car	Score on PC1 (z_1)
Mazda RX4	−0.65
Mazda RX4 Wag	−0.62
Datsun 710	−2.74
Hornet 4 Drive	−0.31
Hornet Sportabout	1.94
Valiant	−0.06
Duster 360	2.96
Merc 240D	−2.02
Merc 230	−2.25
Merc 280	−0.52
Merc 280C	−0.50
Merc 450SE	2.21
Merc 450SL	2.02
Merc 450SLC	2.11
Cadillac Fleetwood	3.84
Lincoln Continental	3.89
Chrysler Imperial	3.54
Fiat 128	−3.80
Honda Civic	−4.19
Toyota Corolla	−4.17
Toyota Corona	−1.87
Dodge Challenger	2.15
AMC Javelin	1.83
Camaro Z28	2.84
Pontiac Firebird	2.21
Fiat X1-9	−3.52
Porsche 914-2	−2.61
Lotus Europa	−3.33
Ford Pantera L	1.35
Ferrari Dino	−0.00
Maserati Bora	2.63
Volvo 142E	−2.38

squares regression and MLR using ordinary least squares for a dataset with two independent and one dependent variable.

MLR assumes that independent variables are measured without error, while in total least squares regression, both the independent and dependent variables have error. To use PCA for deriving the hyperplane for total least squares regression, take the linear function that is the last PC and set it equal to zero. In other words, the last PC loadings vector is orthogonal to the regression hyperplane. For the automobile dataset, we can take the eleventh PC from Table 2 as the coefficients of the regression hyperplane: $v_{11}^T x = 0$. If we divide the coefficients by 0.12, the negative of the coefficient for miles per gallon, we can get the following regression equation for miles per gallon:

Miles/gallon = −1.1 No. cylinders + 5.3 Displacement − 2.1 Horsepower − 0.3 Rear axle ratio − 4.5 Weight + 1.5 1/4 mile time + 0.1 V/S + 0.2 Transmission − 0.4 No. gears + 2.6 No. carburetors

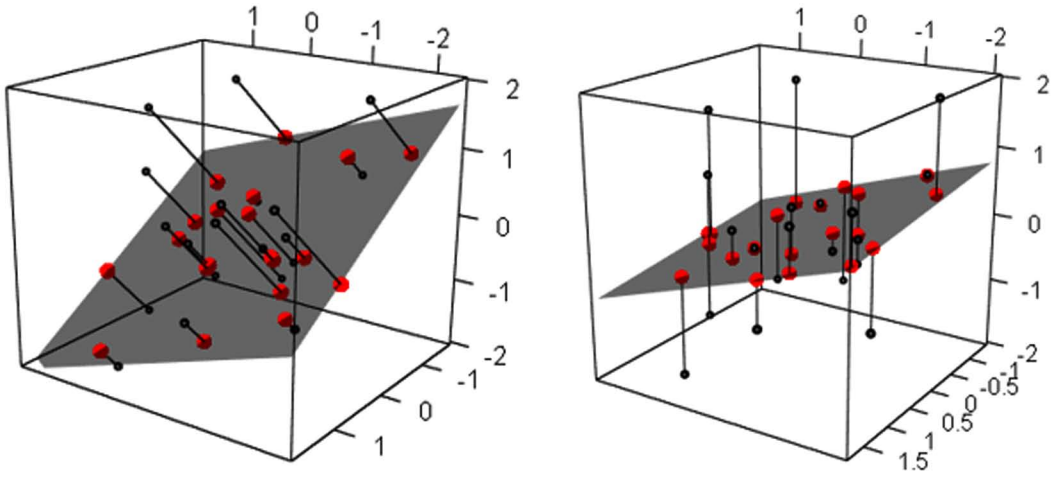

FIGURE 3. Total least squares (left) minimizes the sum of orthogonal distances of points to a fitted hyperplane, while multiple linear regression via ordinary least squares (right) minimizes the sum of vertical distances of points to a fitted hyperplane.

2. Principal Component Analysis and Optimization

In this section, we describe several ties between PCA and mathematical optimization. Much of what is developed here is described by Jolliffe [17], who includes discussions of further connections between PCA and optimization for both population-based and sample-based estimates. Our treatment of these subjects differs in that we focus on providing an explicit statement of the associated optimization problems based on the geometric properties of PCs for a sample of points.

Throughout this development, X is the $n \times m$ data matrix with rows x_i; q denotes the dimension of a subspace into which points are projected ($q \leq m$); Z is the $n \times q$ matrix of scores; Y is the $n \times m$ matrix of projected points in terms of the original dimensions; V is the $m \times q$ matrix comprised of the first q columns v_i, $i = 1, \ldots, q$ of the rotation matrix; and λ_i is the standard deviation of the projected points on the i^{th} PC. We indicate the variables in the optimization problems beneath the objective sense, and to maintain simplicity, we use the same notation for the variables of the problems and their optimal solutions given by PCA.

2.1. Subspace Estimation

PCA can be viewed as providing a series of best-fit subspaces of dimensions $q = 1, 2, \ldots, m-1$. The idea is that for a given set of data, latent structures exist that are out of view and redundancies are present which may be eliminated while retaining most if not all of the data's relevant information. After eliminating these redundancies, we can represent the data nearly as completely in a q-dimensional space as in the original m dimensions.

Consider finding a best-fit subspace for a given cloud of points. Subspaces contain the origin which is why we typically center the data by subtracting column-wise means. To find the best-fitting q-dimensional subspace where $q < m$, we can solve

$$\min_{V,Z} \sum_{i=1}^{n} \|x_i - V z_i\|_2^2, \tag{3}$$

s.t.

$$V^T V = I,$$

where I is the identity matrix. The problem minimizes the sum of squared distances of points to their projections in a q-dimensional subspace. The score z_i is the projection of x_i in the q-dimensional subspace, the quantity Vz_i is the projected point in terms of the original coordinates, and the columns of V define a basis for the subspace. An optimal solution is to set V to be the loadings for the first q PCs and $Z = XV$.

Note that any set of vectors spanning the subspace defined by the columns of V is also optimal for (3), so that the first q PCs are not a unique solution. If by solving (3) an oblique set of spanning vectors is obtained, a basis may be derived by an orthogonalization procedure. However, the basis may not be the same basis given by PCA, as any rotation of the basis vectors in the subspace V remains optimal.

Consider Figure 4 that depicts the best-fit two-dimensional subspace for a cloud of points in three dimensions. At optimality, the columns of V define a basis for the fitted plane, z_i gives the coordinates of the i^{th} projected point on the plane in terms of the basis, and y_i (given by $y_i = Vz_i$) is the projected point in terms of the original three coordinate axes.

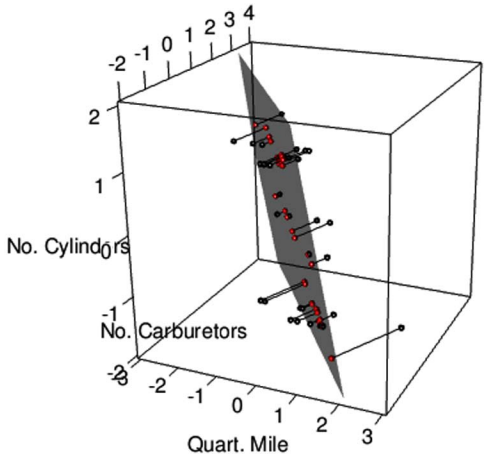

FIGURE 4. Plot of the number of cylinders, quarter mile time, and number of carburetors in the Motor Trend dataset, with their orthogonal projections (red points) onto the hyperplane of maximum variation with respects to these variables.

For any q, setting the columns of V to the loadings for the first q PCs and Z to the scores on the corresponding subspace gives an optimal solution for (3). Therefore, PCA provides optimal solutions to a series of optimization problems simultaneously: it solves (3) for $q = 1, \ldots, m$.

The loadings for the first q PCs and the corresponding scores are not unique solutions to (3). Any basis for the fitted subspaces would also be optimal.

Consider the following alternative optimization problem for finding a best-fit subspace.

$$\min_{Y} \sum_{i=1}^{n} \|x_i - y_i\|_2^2, \tag{4}$$

s.t.

$$\text{rank}(Y) \leq q.$$

In this formulation, we do not decompose the projected points to find the fitted subspace explicitly. Rather, the sum of squared distances of points x_i to their projections in terms of the original coordinates y_i is minimized while restricting the dimension of the subspace containing the projections to a value at most q. Setting $Y = XVV^T$ where the columns of

V are the loadings of first q PCs provides an optimal, but not necessarily unique, solution to (4).

While PCA provides a series of optimal solutions for (3) and (4) for any given q where $q < m$, solving these optimization problems does not necessarily provide the same results as PCA. In the next section, we describe constructive optimization-based approaches for finding the PC loadings.

2.2. Successive Directions of Maximum/Minimum Variation

Consider seeking a vector such that the variance of projections of points in X onto this one-dimensional subspace is maximized. Once we have found it, we may then seek a second vector orthogonal to the first which maximizes the variance in X 'left over' from the projection on the first vector. We can continue finding such vectors that "explain" the variance left in X after the projections onto the vectors found thus far. The variance of points projected on a vector v_1 is $\text{var}(Xv_1) = v_1^T S v_1$, where S is the sample covariance/correlation matrix given by $S = \frac{1}{n-1} X^T X$. The first optimization problem may therefore be written as:

$$\max_{v_1} v_1^T X^T X v_1, \tag{5}$$

s.t.

$$v_1^T v_1 = 1.$$

The only solution for v_1 satisfying the KKT conditions is the eigenvector of S associated with the largest eigenvalue which is equal to λ_1^2, the variance of the projections on the first PC; the optimal objective function value is λ_1^2. This solution, giving the loadings for the first PC, is optimal. The objective function may also be written $\sum_{i=1}^{n} (v_1^T x_i)^2 = \sum_{i=1}^{n} z_i^2$. The variance of the projections is the variance of the scores of points projected on v_1.

If the data variables are measured on different scales, then those with larger values will have higher variances. Therefore, it is common to normalize the columns by dividing by the column-wise standard deviations so that all variables are on a common scale. If X is centered but not scaled, then S is the covariance matrix. If X is our centered and scaled data, then S is the correlation matrix. For example, in our example dataset, the variance of the rear axle ratio variable is 15,360.8 while the variance of the cylinder variable is 3.2, and indication that scaling is warranted.

Perhaps a more pure geometric interpretation of (5) is to find a vector v_1 such that the sum of squared lengths of the projections of points in X onto v_1 is maximized. Also, note that the angle θ made by v_1 and each unit direction e_k satisfies $\cos\theta = v_1^T e_k$, so the k^{th} element of v_1 is the cosine of the angle that v_1 makes with the original coordinate axes. Figure 2 illustrates this fact for the number of cylinders in the automobile data.

After finding v_1, to find the second direction of maximum variation in the data left after projection onto v_1, we solve the following problem.

$$\max_{v_2} v_2^T X^T X v_2, \tag{6}$$

s.t.

$$v_2^T v_2 = 1,$$
$$v_1^T v_2 = 0.$$

Note that the second constraint is linear because we assume v_1 is known. To enforce the idea that v_2 captures the variation 'left over' after projection on v_1, this optimization problem requires that v_2 be orthogonal to v_1. We seek the vector that maximizes the lengths of projected points that is orthogonal to v_1. The only solution satisfying the KKT conditions for (6) is the eigenvector of S associated with the second largest eigenvalue which is equal to λ_2^2. This optimal solution gives the loadings for the second PC. The elements of v_2 give

the cosine of the angle that v_2 makes with the original coordinate axes. Carrying on, we can find the k^{th} PC loadings vector by solving

$$\max_{v_k} v_k^T X^T X v_k, \tag{7}$$

s.t.

$$v_k^T v_k = 1,$$
$$v_j^T v_k = 0,\ j = 1, \ldots, k-1.$$

The vectors v_j, $j = 1, \ldots, q$ form the first q columns of the rotation matrix given by PCA. Finding the PC loadings in this manner corresponds to a *forward* view of PCA because it finds successive directions of maximum variation in the data. The first q PC loadings also provide an optimal solution to the following problem

$$\max_{V} V^T X^T X V, \tag{8}$$

s.t.

$$V^T V = I.$$

One may also begin finding a vector v_m such that $\text{var}(v_m^T X)$ is minimized. The vector v_m gives the direction of minimum variation in the data and is the vector along which the sum of squared lengths of projections is minimized. The PC loadings for the m^{th} PC are optimal. We can find successive directions of minimum variation, and solve for the k^{th} PC as follows:

$$\min_{v_k} v_k^T X^T X v_k, \tag{9}$$

s.t.

$$v_k^T v_k = 1,$$
$$v_j^T v_k = 0, j = k+1, \ldots, m$$

The only vector satisfying the KKT conditions is the eigenvalue of S associated with the k^{th} smallest eigenvalue, λ_k^2. As with the forward approach, PC loadings provide an optimal solution to an analogous minimization problem to (8). This method of calculating PC loadings corresponds to a *backward* view of PCA. Such an approach can be useful when developing extensions of PCA [7].

The role of eigenvalues and eigenvectors of the covariance/correlation matrix in PCA portend the connection with SVD for calculations. To see this, recalling the fundamental theorem of linear algebra [27], the SVD of X is:

$$X = U \Lambda V^T \tag{10}$$

In this factorization, U and V are orthonormal matrices the columns of which are commonly referred to as the left and right singular vectors of X; U is $n \times n$ and V is $m \times m$. Finally, Λ is an $n \times m$ diagonal matrix containing the singular values of X. The spectral decomposition of the sample covariance/correlation matrix S is:

$$S = \frac{1}{n-1} X^T X = \frac{1}{n-1} \left(U \Lambda V^T\right)^T \left(U \Lambda V^T\right) = \frac{1}{n-1} V \Lambda^T \Lambda V^T$$

Of note here is the fact that the right singular vectors of X are equal to eigenvectors of S, and the eigenvalues of X are the square root of the eigenvalues of $(n-1)S$. Therefore, the right singular vectors of X are the columns of the rotation matrix V in PCA. Further, multiplying both sides of (10) by V on the right, $XV = U\Lambda$ which are the scores Z. By truncating the matrix $\Lambda^T \Lambda$ down to its first q columns, we obtain a matrix of points projected onto a q-dimensional orthogonal subspace. What Eckart and Young proved in 1936 [10] is that of all q-dimensional representations of X, none account for a greater proportion of the variance of X while being as 'close' to X in terms of the Euclidean distance between the original points and their representation in the reduced rank subspace.

2.3. Maximizing the Distances between Projected Points

For a specified dimension q where $q < m$, PCA also yields the subspace in which the sum of squared pairwise distances between points within it are maximized, formally stated as:

$$\max_{Z,V} \sum_{i=1}^{n} \sum_{k=1}^{n} \|z_i - z_k\|_2^2, \tag{11}$$

s.t.

$$\begin{aligned} Z &= XV, \\ V^T V &= 1. \end{aligned}$$

An optimal solution is to set V to be the loadings for the first q PCs, then set $Z = XV$. Consider again Figure 4. The best-fit two-dimensional subspace for the points is the one for which the projected points are most spread out on the plane.

As with (3) and (4), PCA provides an optimal solution for (11) for any given q where $q < m$, but the solution may not be unique. In other words, (11) is not a constructive approach for calculating all of the PC loadings.

Classical PCA, based on the L_2 norm, is quite powerful in the sense of providing a projection of data onto a lower dimensional subspace which satisfies numerous optimality criteria. Though this is not the way principal components are classically derived, we have presented several optimization problems to which PCA provides optimal solutions. Depending on the particular application, solving a single optimization problem may be sufficient. For example, one may only need the scores, *reconstructions* (projected points in terms of the original coordinates), or the subspace, and not each basis vector returned by PCA. In this way, each of the optimization problems presented here can be used as the basis for new methods for PCA. The next section describes proposed methods for PCA in this vein.

3. Reformulations for Alternative Objectives

As indicated by our Google Scholar search, much research has been conducted in creating new methods for PCA. Common goals are to increase the robustness of PCA to outlier observations, to create *sparse* PCs where the loadings have few nonzero entries, or to create *smooth* PCs where the loadings of certain variables are similar. In general, the equivalences among the optimization problems that we have presented break down when any of the optimization problems is modified. Therefore, the choice in modifying PCA should be made based on the particular analysis needed. In this section, we review a sample of alternative approaches for PCA and describe how they may be viewed in the context of the related optimization problems. Our review is by no means exhaustive but is intended to give the reader some ideas about how optimization may be used to create new methods for PCA.

We begin by reviewing three motivations for modifying traditional PCA. First, the subspaces generated by PCA are not robust to outliers. This is due to the fact that variance places too much emphasis on points that are far from the mean, and the sum of squared Euclidean distances places too much weight on observations that are not close to the density of the data. The optimal subspace can be skewed with basis vectors of the new subspace 'pulled' too far in the direction of outlying points. Using alternate norms in measuring distance can alleviate the effects of outliers. Second, the PC loadings for traditional PCA may be dense and therefore difficult to interpret. In our example in Section 1, each of the variables has a nonzero loading. If only miles per gallon and horsepower were allowed to have nonzero loadings, we would expect the first PC to have loadings for those variables with opposite signs as a reflection of the trade-off in horsepower and fuel economy. Penalizing the number of variables with nonzero PC loadings can help with sparsity and therefore interpretation of the loadings vectors. Third, we may desire a degree of smoothness in the loadings matrix. For example, if data are measured over time or in a spatial context, it may be desirable

to have the loadings for "neighboring" variables to be similar to facilitate interpretation. Penalizing differences between loadings matrix values can encourage smoothness.

Several investigators have proposed robust methods for PCA by using the L_1 norm with the optimization formulation (3). Ke and Kanade [19, 20] alter (3) by replacing the L_2 norm with the L_1 norm and minimizing the sum of L_1-norm distances of points to their projections in a fitted subspace. They propose L1-PCA a locally-converging heuristic procedure for deriving solutions by alternating fixing V and solving for Z, and fixing Z and solving for V. Brooks et al. [5] leverage the fact that when the fitted subspace is a hyperplane ($q = m-1$), the L_1-norm best-fit subspace can be found by fitting a series of L_1 linear regressions, each of which may be viewed as a linear program. To extend this property to a method for PCA called L1-PCA*, they find a basis spanning the projections in the $m-1$-dimensional subspace, find the L_1-norm best-fit $m-2$-dimensional hyperplane for the projected points, and so on until $q = 1$. Because subspaces of lower dimension are fitted as the algorithm progresses, it may be viewed as an example of backward PCA [7]. Park and Klabjan [25] propose a locally-convergent algorithm for minimizing the sum of L_1 norm distances to the L_2 projections of points onto the fitted subspace.

The objective function for (3) may be re-written as $\|X^T - VZ^T\|_2^2$ and viewed as a decomposition of X into a product of matrices V and Z. Many methods for alternative matrix decompositions have been proposed. Allen et al. [1] propose GMD, Generalized least squares Matrix Decomposition, and accompanying algorithms for datasets with structured variables or known two-way dependencies. This framework allows one to simultaneously control sparsity and smoothness of the PCs. The optimization problem that they investigate generalizes SVD. Before presenting their model, note that based on SVD we can rewrite the objective function of (3) as $\sum_{i=1}^{n} \|x_i - V\Lambda u_i\|_2^2$. The GMD problem is

$$\min \sum_{i=1}^{n} \|x_i - V\Lambda u_i\|_{Q,R}^2 \tag{12}$$

s.t.

$$\begin{aligned} U^T Q U &= I, \\ V^T R V &= I, \\ \operatorname{diag}(\Lambda) &\geq 0, \end{aligned}$$

where I are appropriately-sized identity matrices. The L_2 norm is replaced by the Q, R norm defined by $\|A\|_{Q,R} = \sqrt{\operatorname{tr}(QARA^T)}$. By adjusting Q and R, one can control sparsity and smoothness.

The approach of finding the projections of points into a fitted subspace as in (4) has inspired much interest in convex optimization. Candès et al. [6] and Goldfarb et al. [13] discuss procedures for finding approximate solutions to a modified problem where the sum of distances from points to their projections measured using the L_1 norm is minimized. Using the L_1 norm can impart robustness to outliers. The proposed solution methods use smooth approximations of the L_1 norm of the projections and of the rank of the projection matrix Y to convert the problem to one conducive to convex optimization. Balzano et al. [2] develop an online method for subspace estimation for incomplete data called GROUSE. Balzano and Wright [3] demonstrate local convergence under assumptions about the randomness in the streaming data.

Algorithms for a version of the successive direction of maximum variations is proposed by Markopoulos et al. [22]. They show that when the objective function for (5) is replaced with an L_1 norm analog, namely $\|Xv_1\|_1$, that the problem is NP-complete for general n and m. They give a globally-convergent $O(n^m)$ algorithm. Kwak [21] provides a fast locally-convergent algorithm for the same problem. The method is extended to a method for PCA by operating on the projections of points in the subspace orthogonal to the previous directions found. The globally- and locally-convergent techniques have been extended for solving the

L_1 norm analog of (8) [24, 22]; Park and Klabjan [25] also suggest a locally-convergent algorithm. Galpin and Hawkins [11] suggest an L_1 norm alternative formulations for (5), including one where the constraint $\|v_1\|_2 = 1$ is replaced with $\|v_1\|_1 = 1$.

D'Aspremont et al. [8] propose a version of (5) with a cardinality constraint on the number of nonzero entries in v_1 is added to ensure sparsity. They then propose an SDP relaxation for the resulting optimization problem. Solution techniques are explored by Iyengar et al. [16].

4. Conclusions

The purpose of this paper is to provide an optimization-based context for PCA methods and to help inspire the optimization community to develop new methods for PCA. We began with a brief history of PCA and an overview of some of the ways it is used. We also showed how classical PCA can be viewed as providing optimal solutions to several optimization problems simultaneously. The most common means for finding these solutions is the singular value decomposition of the covariance or correlation matrix, and if not for its simplicity and the straightforward way in which it may be used to find principal components, the robustness problems of classical PCA may not be so easily forgiven.

By 'reverse engineering' classical PCA to consider the optimization problems for which classical PCA finds optimal solutions, we can consider possibilities for tweaking them in such a way that 'better' solutions may be found that satisfy alternate criteria. However, alternate metrics come with their own sets of problems. Their solutions are often only achieved by iterative processes which must be solved via algorithms that are computationally complex. Globally optimal solutions are not guaranteed. Additionally, once we substitute alternate distance or information metrics into one of the optimization problem formulations, solutions that are optimal for this problem are no longer guaranteed to be optimal for all others. Many open problems with regard to the alternate formulations for PCA remain. In particular, the existence of polynomial-time globally-convergent algorithms for many of the proposed variations is unknown.

Nevertheless, it is arguable that pursuit of less computationally intensive algorithms that find optimal PCs under a set of more robust criteria is worth striving for. Viewing PCA as not just an eigenvalue decomposition problem, but the solution of multiple various optimization problem formulations can help to improve classical PCA.

5. Acknowledgments

This work was supported in part by NIH awards 2P60MD002256-06 and 8U54HD080784 and an award from The Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] G.I. Allen, L. Grosenick, and J. Taylor. A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159, 2014.

[2] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Proceedings of the 48th Annual Allerton Conference on Communicaton, Control and Computing*, 2010.

[3] L. Balzano and S.J. Wright. Local convergence of an algorithm for subspace identification from partial data. *CoRR*, arXiv:1306.3391, 2013.

[4] E. Beltrami. On bilinear functions. *University of Minnesota, Dept. of Computer Science, Technical Report*, 1990(TR_1990-37), 1990. Original pubishing date: 1873.

[5] J.P. Brooks, J.H. Dulá, and E.L. Boone. A pure L_1-norm principal component analysis. *Computational Statistics & Data Analysis*, 61:83–98, 2013.

[6] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

[7] J. Damon and J. S. Marron. Backwards principal component analysis and principal nested relations. *Journal of Mathematical Imaging and Vision*, 50(1-2):107–114, 2014.

[8] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G.R.G. Lanckreit. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

[9] C. Ding and X. He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

[10] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[11] J.S. Galpin and D.M. Hawkins. Methods of L_1-estimation of a covariance matrix. *Computational Statistics & Data Analysis*, 5(4):305–319, 1987.

[12] F. Galton. *Natural Inheritance*. Macmillan and Co., London, 1889.

[13] D. Goldfarb, S. Ma, and K. Scheinberg. Fast alternating linearization methods for minimizing the sum of two convex functions. *Mathematical Programming, Series A*, 141:349–382, 2013.

[14] H.V. Henderson and P.F. Velleman. Building multiple regression models interactively. *Biometrics*, 37:391–411, 1981.

[15] H. Hotelling. *Analysis of a Complex of Statistical Variables Into Principal Components*. Warwick and York, 1933.

[16] G. Iyengar, D.J. Phillips, and C. Stein. Approximating semidefinite packing problems. *SIAM Journal on Optimization*, 21(1):231–268, 2011.

[17] I.T. Joliffe. *Principal Component Analysis*. Springer-Verlag, New York, 2nd edition, 2002.

[18] C. Jordan. Mémoire sur les formes bilinéaires. *Journal de Mathematiques Pures et Appliquées, Deuxième série*, 19:35–54, 1874.

[19] Q. Ke and T. Kanade. Robust subspace computation using L1 norm. Technical Report CMU-CS-03-172, Carnegie Mellon University, Pittsburgh, PA, 2003.

[20] Q. Ke and T. Kanade. Robust L_1 norm factorization in the presence of outliers and missing data by alternative convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2005.

[21] N. Kwak. Principal component analysis based on L1-norm maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9):1672–1680, 2008. JID: 9885960; ppublish.

[22] P.P. Markopoulos, G.N. Karystinos, and D.A. Pados. Optimal algorithms for L_1-subspace signal processing. *CoRR*, abs/1405.6785, 2014.

[23] D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to Linear Regression*. Wiley, Hoboken, 2012.

[24] F. Nie, H. Huang, C. Ding, D. Luo, and H. Wang. Robust principal component analysis with non-greedy ℓ_1-norm maximization. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 1433–1438, 2011.

[25] Y.W. Park and D. Klabjan. Algorithms for L1 principal component analysis. Submitted, 2014.

[26] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, 1901.

[27] G. Strang. The fundamental theorem of linear algebra. *The American Mathematical Monthly*, 100:848–855, 1993.

[28] H. Zha, X. He, C. Ding, M. Gu, and H.D. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems 14*, pages 1057–1064. MIT Press, 2002.