

PCA and SPCA Tutorial

Dr. Edel Garcia, admin@miislita.com

Copyright 2008 E. Garcia
First Published: March 25, 2008
Last Updated: April 16, 2008

Keywords: PCA, SPCA, SVD, principal component analysis, covariance matrix, correlation matrix

Abstract: This is a tutorial on PCA (Principal Component Analysis) and SPCA (Standardized PCA).

Introduction

Principal Component Analysis (PCA) is an exploratory tool designed to identify unknown trends in a multidimensional data set \mathbf{X} . The algorithm was introduced in 1933 by H. Hotelling (1), hence sometimes it is called Hotelling's Transform (1). However, today we know that implementing PCA is the equivalent of applying Singular Value Decomposition (SVD) on the covariance matrix of a data set (2, 3).

Assume that \mathbf{X} is an array of n observations x_{ij} (rows) occurring in $j, j+1, \dots, k$ dimensions (columns). Assume that we subtract the mean μ_j from the observations so that a new data set \mathbf{Y} with zero mean is obtained. Implementing PCA via SVD then reduces to computing the following *reaction equations*:

$$\begin{array}{lcl} \mathbf{X} & \rightarrow & \mathbf{Y} \\ \mathbf{Y} & \rightarrow & \mathbf{Y}^T \\ \mathbf{1}/(n-1) \quad \mathbf{Y}^T \mathbf{Y} & \rightarrow & \mathbf{A} \\ \mathbf{A} & \rightarrow & \mathbf{USV}^T \end{array}$$

The first three reactions mean center \mathbf{X} across the origin and take dot products. Computing $\mathbf{Y}^T \mathbf{Y}$ produces an array of sum of square deviations. Multiplying this array by $\mathbf{1}/(n-1)$ transforms its diagonal elements into variances σ^2 and non-diagonal elements into co-variances. Thus, a variance-covariance matrix \mathbf{A} is obtained. To simplify, \mathbf{A} is frequently called the *covariance matrix* of \mathbf{X} .

\mathbf{A} is then decomposed with SVD; i.e., $\mathbf{A} = \mathbf{USV}^T$. These terms are defined as follows. \mathbf{V}^T is the transpose of \mathbf{V} and \mathbf{S} is a diagonal matrix that stores singular values (i.e., $\lambda_1, \dots, \lambda_{i+1}, \dots, \lambda_k$). \mathbf{U} and \mathbf{V} are orthogonal matrices. Their column vectors are the so-called *left* and *right* eigenvectors of \mathbf{A} .

When these eigenvectors multiply \mathbf{Y} the result is an affine transformation. Essentially, coordinates are shifted and rotated until they end up aligned with vectors, termed now *basis vectors*. In this sense, PCs are linear combinations of the original axes.

\mathbf{V} columns (\mathbf{V}^T rows) are found to produce the desired linear combinations. The first column of \mathbf{V} corresponds to the largest PC, the second column corresponds to the second largest PC, and so on. These define the direction in which the variability of the original data set is maximized.

Although optional, we can reflect such ordering in the original data set by sorting columns from left to right in descending order of variances before implementing PCA. Since diagonal elements of \mathbf{A} will inherit this ordering, this can be used for double checking variance calculations. In addition, a plot of the first two PCs displays the transformations associated with the first two columns of \mathbf{X} . In principle, variances, eigenvalues, and eigenvectors should follow this ordering.

Once identified, PCs can be used for other studies like Residual Analysis, K-Means, K-Medoids, etc. To get the old data back, we compute \mathbf{YV}^T and add the mean values that were removed.

Important Note - To use this tutorial as a classroom demonstration, you need EXCEL and any SVD calculator. The one at <http://www.bluebit.gr/matrix-calculator/> is good enough. You can also write your own SVD program (2, 3). If you don't know/have EXCEL, please ask your instructor for alternatives.

Problem

Apply PCA to the data set **X** given in Figure 1. This consists of measurements of weight in pounds, height in inches, and age in years for 12 nutritionally deficient children (5).

Age	Weight	Height
8	64	57
10	71	59
6	53	49
11	67	62
8	55	51
7	58	50
10	77	55
9	57	48
10	56	42
6	51	42
12	76	61
9	68	57

Figure 1. Multidimensional data set **X**.

Step 1. Firstly, we compute μ values. With these we compute standard deviation σ and variance σ^2 values. Next, we sort columns of **X** in descending order of σ^2 values. To get **Y**, we subtract μ values from each row of **X**.

A	B	C	D	E	F	G	H
X =	Weight	Height	Age	Y =	Weight	Height	Age
	64	57	8		1.25	4.25	-0.83
	71	59	10		8.25	6.25	1.17
	53	49	6		-9.75	-3.75	-2.83
	67	62	11		4.25	9.25	2.17
	55	51	8		-7.75	-1.75	-0.83
	58	50	7		-4.75	-2.75	-1.83
	77	55	10		14.25	2.25	1.17
	57	48	9		-5.75	-4.75	0.17
	56	42	10		-6.75	-10.75	1.17
	51	42	6		-11.75	-10.75	-2.83
	76	61	12		13.25	8.25	3.17
	68	57	9		5.25	4.25	0.17
μ =	62.75	52.75	8.83				
σ =	8.99	6.82	1.90				
σ^2 =	80.75	46.57	3.61				

Figure 2. **X** and its **Y** representation.

Step 2. Next, compute the covariance matrix as $\mathbf{A} = (1/n - 1)Y^T Y$.

You can also use EXCEL's VAR and COVAR formulas to construct \mathbf{A} , which simplifies all the calculations. However, a word of caution is in order. If your version of EXCEL uses n in the denominator of the covariance formula instead of $n - 1$ you need to correct the results by multiplying covariances times $n/(n - 1)$. This is important as $1/n$ provides a biased estimation of variance especially for small n . The proper normalization for an unbiased estimator is $1/(n - 1)$.

Figure 3 depicts the covariance matrix \mathbf{A} after the corrections. Note how diagonal elements inherit the variance ordering of Figure 2.

J	K	L	M
A =	Weight	Height	Age
Weight	80.75	49.93	13.14
Height	49.93	46.57	7.95
Age	13.14	7.95	3.61

Figure 3. Covariance matrix \mathbf{A} .

Covariance tells whether changes in any two variables move together. Consider two variables x and y . Positive covariance means that high values of y are associated with high values of x . Negative covariance means that high values of y are associated with low values of x . Zero covariance means that there is no association between x and y . Figure 3 suggests for nutritionally-deficient children that **Weight-Height** changes are more related than **Weight-Age** changes or **Height-Age** changes.

Step 3. To visualize if there is a hidden pattern in the data we apply SVD to the covariance matrix and do a rank k approximation. In this example, we want to retain the first two dominant PCs; thus, $k = 2$. Evidently, for $k > 3$ a visual representation is not possible.

If using the Bluebit calculator, paste \mathbf{A} into this tool and check *Singular Value Decomposition*. From the pull-down menus, select *Values are delimited by Tabs* and *Show results using 2 decimal digits*. Click *Calculate* button. You should be able to get the \mathbf{U} , \mathbf{S} , and \mathbf{V}^T matrices shown in Figure 4.

U

-0.81	0.56	-0.18
-0.58	-0.82	0.02
-0.13	0.12	0.98

S

118.48	0.00	0.00
0.00	11.03	0.00
0.00	0.00	1.43

V^T

-0.81	-0.58	-0.13
0.56	-0.82	0.12
-0.18	0.02	0.98

Figure 4. SVD results obtained from \mathbf{A} .

Step 4. Compute **V** and **YV** and plot the first two columns of **YV**.

	A	B	C	D
V=	-0.81	0.56	-0.18	
	-0.58	-0.82	0.02	
	-0.13	0.12	0.98	
YV=	-3.37	-2.88	-0.95	
	-10.46	-0.36	-0.21	
	10.44	-2.72	-1.09	
	-9.09	-4.94	1.55	
	7.40	-3.00	0.55	
	5.68	-0.62	-0.99	
	-13.00	6.28	-1.37	
	7.39	0.70	1.11	
	11.55	5.18	2.15	
	16.12	1.90	-0.87	
	-15.93	1.04	0.89	
	-6.74	-0.52	-0.69	

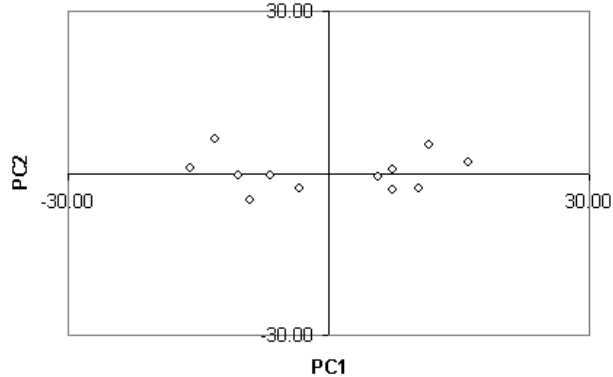


Figure 5. Transformation of the data set and visualization of the two dominant PCs.

Figure 5 indicates two things:

1. as expected, points are closer to PC1 than to PC2.
2. two distinct clusters are formed.

If we are interested in clustering points with a K-Means or K-Medoids algorithm we can choose points from PC1 as initial centroids.

Improving Results with SPCA

PCA has severe shortcomings. It can fail if the data is non-Gaussian or time-dependent. Suppose we want to apply PCA to images taken from a satellite at different time intervals. If some features change between scenes some principal components as the signal-to-noise ratio might also change. A single PCA will then depend on spectral and spatial features. The largest PCs might carry the most important information about scene variations, but they may not necessarily carry the information of interest (6).

To improve this situation we need to find a way to normalize the influence of each variable, enhancing the influence of variables with small variance and reducing the influence of variables with high variance. In this way, the different time variance patterns are extracted from a time series more effectively. This is what Standardized Principal Component Analysis (SPCA) does.

In SPCA, we transform **X** values into z-scores, defined as $z_{ij} = (x_{ij} - \mu_j) / \sigma_j$ such that a new matrix **Z** is obtained. For each dimension, we subtract μ_j and divide by σ_j values. Next we compute $\mathbf{Z}^T \mathbf{Z}$. The result is a correlation matrix. The following reaction equations summarize the construction of **B**.

$$\begin{array}{lcl}
 \mathbf{X} & \rightarrow & \mathbf{Z} \\
 \mathbf{Z} & \rightarrow & \mathbf{Z}^T \\
 \mathbf{Z}^T \mathbf{Z} & \rightarrow & \mathbf{B}
 \end{array}$$

Figure 6 shows the construction and analysis of the correlation matrix **B**. To indicate that results were obtained from **B** we have appended a (**B**) subscript to the data.

I	J	K	L	M	N	O	P
B =	Weight	Height	Age	V_(B) =	-0.60	0.06	0.80
Weight	10.61	8.06	8.90		-0.53	0.71	-0.46
Height	8.06	8.97	7.67		-0.59	-0.70	-0.39
Age	8.90	7.67	10.63				
ZV_(B) =	-0.15	0.76	0.00				
	-1.40	0.28	0.07				
	1.82	0.59	-0.03				
	-1.68	0.19	-0.69				
	0.91	0.07	-0.40				
	1.10	0.36	0.14				
	-1.49	-0.10	0.88				
	0.70	-0.59	-0.23				
	0.57	-0.55	-0.06				
	2.50	-0.15	0.26				
	-2.51	-0.22	-0.03				
	-0.73	0.42	0.15				

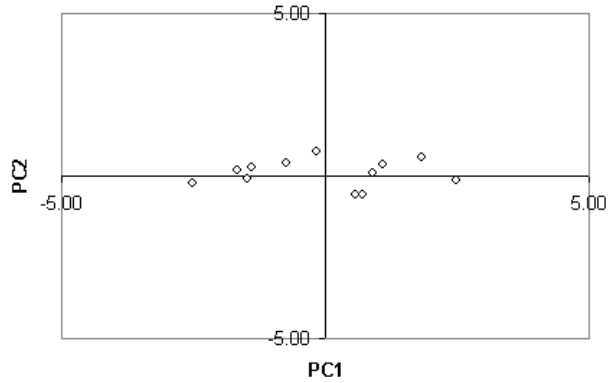


Figure 6. Correlation matrix **B** and a plot of its two dominant principal components.

Compared with Figure 5, points are distributed over smaller scales. The two clusters now look closer to one another and to the dominant principal component, PC1. To better compare the principal components of **A** and **B**, in Figure 7 we have plotted both.

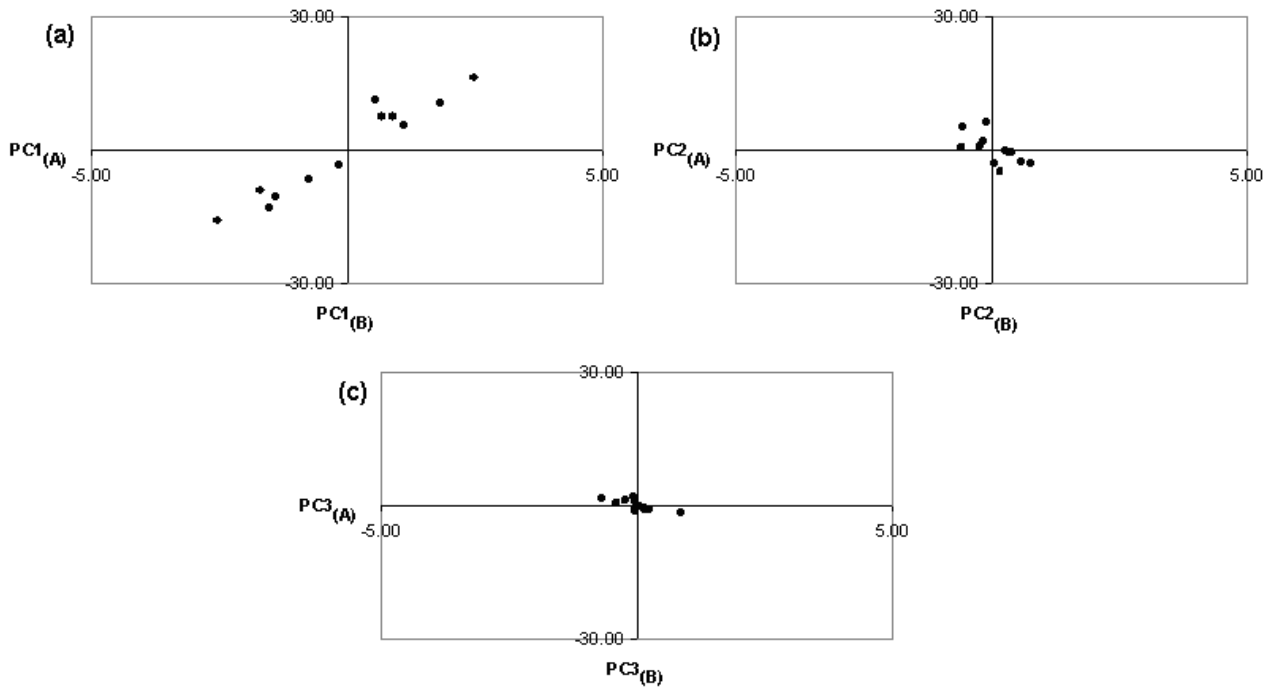


Figure 7. Plots of pairs of PCs from covariance (**A**) and correlation (**B**) matrices.

Notice that PC1 points tend to be proportionally distributed and that the less dominant the PCs, the tighter the points. This is expected since variance normalization reduces even more noisy dimensions.

Conclusion

Principal Component Analysis (PCA) is a discovery tool designed to identify unknown trends in a multidimensional data set. Implementing PCA is the equivalent of applying Singular Value Decomposition (SVD) on the covariance matrix. The algorithm is easy to understand since it is based on rather basic statistical and linear algebra concepts. However, it can fail if the major assumptions used (linearity and Gaussian data) are not applicable.

I wrote this tutorial for two reasons:

1. to help graduate students taking my *Search Engines Architecture* (7) course with a general overview and review of linear algebra concepts.
2. because most tutorials discuss PCA, but ignore SPCA.

Unlike PCA, SPCA equalizes dissimilar variations in the data set by using a correlation matrix instead of a covariance matrix. In general, a correlation matrix is recommended over a covariance matrix when the data is time-dependent, when variances are rather extreme, or when different units are used.

References

1. Hotelling, H., *Analysis of a complex of statistical variable into principal components*; J. Educ. Psych., vol. 24, 417-441 (1933).
2. Smith, L.; *A tutorial on Principal Components Analysis*; (2002).
http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
3. Shlens, J.; *A tutorial on Principal Component Analysis*; (2003)
http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf
4. Roden, J., Trout, D., King, B.; *A Tutorial on PCA Interpretation using CompClust* (2005).
http://woldlab.caltech.edu/compclust/pca_interpretation_tutorial.pdf
5. Kleinbaum, Kupper, Muller; *Applied Regression Analysis and Other Multivariable Methods*; (1988).
<http://www.biostat.jhsph.edu/~amanicha/BiostatII/labs/lab8.pdf>
6. Behrens, R.; *Change Detection Analysis with Spectral Thermal Imagery*; Naval PostGraduate School, Monterey, California; Thesis (1998).
http://www.nps.edu/Faculty/Olsen/Student_theses/Behrens_Thesis.pdf
7. Garcia, E., *Search Engines Architecture* (2008).
<http://irthoughts.wordpress.com/category/search-engines-architecture-course/>