PCA & SPCA Tutorial

Dr. E. Garcia, admin@miislita.com

First Published: March 25, 2008
Last Updated: May 1, 2008
Uploaded: July 4, 2009

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

Note - To use this material as a classroom demonstration, you need EXCEL and any SVD calculator. The one at http://www.bluebit.gr/matrix-calculator/ is good enough. If you don't know/have EXCEL, please ask your instructor for alternatives. You can also write your own SVD program. Source code is listed in the references.

Introduction

Principal Component Analysis (PCA) is an exploratory tool designed by Karl Pearson in 1901 to identify unknown trends in a multidimensional data set **X**. The algorithm was introduced to psychologists 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). By providing a tutorial on PCA using SVD, students are familiarized with both matrix decomposition techniques.

A Reaction Equation Approach

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

X		→	**Y**
Y		→	$\mathbf{Y^T}$
1/(n – 1)	$\mathbf{Y^TY}$	→	**A**
A		→	$\mathbf{USV^T}$

The first three reactions mean center **X** across the origin and take dot products, normalized by **1/(n – 1)**. Computing $\mathbf{Y^TY}$ produces an array of sum of square deviations. Multiplying $\mathbf{Y^TY}$ by **1/(n – 1)** yields a matrix **A** where diagonal elements ($i = j$) are variances σ_{ij}^2 and non-diagonal elements ($i \neq j$) are co-variances $\sigma_i\sigma_j$. To simplify, **A** is frequently called the *covariance matrix* of **X**. However, keep in mind that **A** actually is a matrix of variances and covariances.

Computing PCA with SVD

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

When these eigenvectors multiply **Y,** coordinates are shifted and rotated until they end up aligned with vectors, termed now *basis vectors*. This is an *affine transformation* since it involves translation. Note that PCA now reexpresses the data as a linear combination of its basis vectors, **YV**. **V** columns ($\mathbf{V^T}$ rows) are found to produce the desired linear combinations. The first column of **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 arranging columns of **X** from left to right in descending order of variances before implementing PCA. Since diagonal elements of **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 **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 $\mathbf{YVV^T}$ and add the mean values that were removed.

A Practical Example

Let's 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). The goal is to identify which of these variables describe a pattern.

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 arrange columns of **X** in descending order of σ^2 values, from left to right. To get **Y**, we subtract μ values from each row.

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

A Note on covariance formulas - You can also use EXCEL's VAR and COVAR formulas to construct **A**, which simplifies all the calculations. However, a word of caution is in order. If your version of EXCEL uses in the denominator of the covariance formula **n** 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 for small **n**. The proper normalization for an unbiased estimator is **1/(n – 1)**. See Reference 3 footnote 5. For large **n** values (**n >> 1**), the error due to using **1/n** should be insignificant.

Figure 3 depicts the covariance matrix **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 **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*. Accordingly, 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 out of the three possible PCs; thus, **k = 2**. If using the Bluebit calculator, do this: Paste **A** into Bluebit 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 **U**, **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 **A**.

Step 4. Compute **V** from $\mathbf{V^T}$ 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
	PC1	**PC2**	**PC3**
YV =	-3.34	-2.87	-0.94
	-10.41	-0.33	-0.17
	10.40	-2.76	-1.15
	-9.03	-4.91	1.59
	7.37	-3.03	0.51
	5.66	-0.64	-1.03
	-12.95	6.32	-1.32
	7.35	0.67	1.07
	11.47	5.13	2.10
	16.04	1.84	-0.96
	-15.85	1.09	0.96
	-6.70	-0.50	-0.67

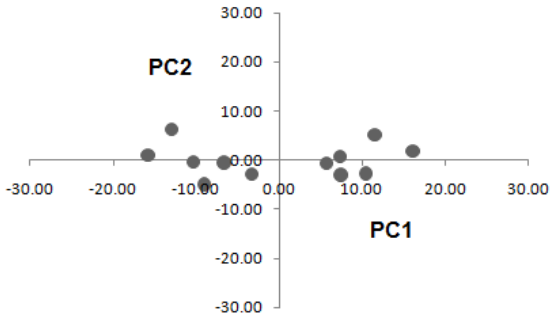

Figure 5. Transformation of the data set and visualization of the two dominant PCs.

Note that by computing **V** and **YV** computations are further simplified, with the added benefit that **YV** columns can be straightforward plotted in EXCEL, to get a visual representation of the principal components.

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. Thus, for certain clusters (e.g., non overlapping/self-intersecting clusters), the so-called K-Means Initial Centroid Problem can be addressed with PCA.

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. SPCA consists in transforming **X** values into z-scores prior to implementing PCA. In this case the matrix that undergoes SVD is not a covariance, but a correlation matrix. The SPCA version of the transformed data set is shown in Figure 6.

A	B	C	D
$Y_s V_s$ =	$PC1_s$	$PC2_s$	$PC3_s$
	-0.19	0.75	-0.04
	-1.42	0.21	0.05
	1.80	0.68	-0.06
	-1.69	0.06	-0.70
	0.91	0.10	-0.40
	1.09	0.42	0.12
	-1.49	-0.13	0.87
	0.74	-0.57	-0.19
	1.01	-1.55	-0.03
	2.52	-0.01	0.28
	-2.51	-0.35	-0.03
	-0.76	0.39	0.12

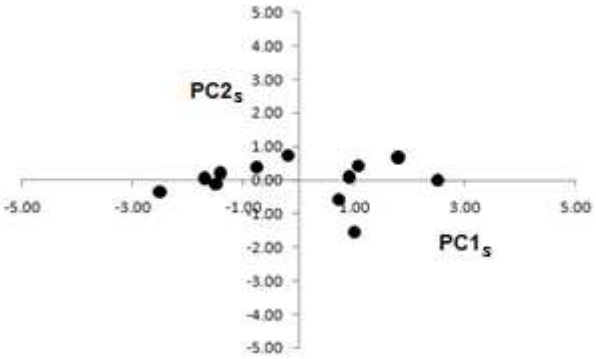

Figure 6. Principal components. The **s** subscripts of **Y** and **V** indicate that SPCA was used.

In Figure 7 we have plotted standardized ($\mathbf{Y_sV_s}$) vs. non-standardized (**YV**) principal components. In graph (a) PC1s vs PC1 points almost describe a straight line. In (b) and (c), points are tighter. This is expected since variance normalization further reduces the spreading of the data, hence noisy dimensions.

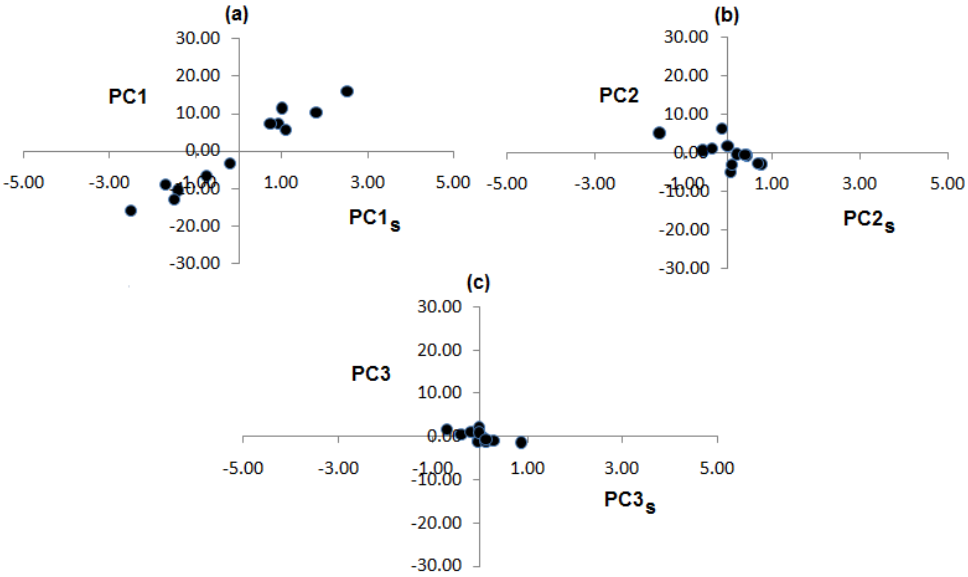

Figure 7. Standardized vs. non-standardized principal components.

Beyond the Covariance Matrix

According to Maydeu-Olivares and Bockenholt (8) and Mardia *et. al.* (9), a covariance matrix can be converted into a squared distance matrix Δ. The elements of this matrix are the squared distance between the i and j variables. These are obtained with the formula $\delta_{ij} = \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}$, where σ_i^2 and σ_j^2 are the variances of the i and j variables and σ_{ij} is the covariance between these. Figure 8 depicts the squared distance matrix Δ, calculated from Figure 3.

S	T	U	V
Δ =	**Weight**	**Height**	**Age**
Weight	0.00	27.46	58.08
Height	27.46	0.00	34.28
Age	58.08	34.28	0.00

Figure 8. Squared Distance Matrix Δ generated from covariance matrix.

As expected, smaller distances implies smaller dissimilarities and larger covariances. For this data set, these results confirm that Weight-Height changes are more related than Weight-Age or Height-Age changes.

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 revised version of the tutorial

- to help graduate students taking my *Search Engines Architecture* (7) course with a linear algebra review.
- because most tutorials on the topic discuss PCA, but ignore SPCA or the connection between PCA and SVD.
- since sometimes PCA is explained through unnecessary matrix manipulations, but no real examples.
- to incorporate some corrections and changes made in class, but not in the online version.

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 due to a common source of fluctuations, 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/
8. Maydeu-Olivares, A, Ulf Bockenholt, U., *Structural Equation Modeling of Paired-Comparison and Ranking Data*; Psychological Methods, Vol. 10, No. 3, 285–304 (2005).
   http://www.statmodel.com/download/maydeuolivbaockenholt.pdf
9. Mardia, K. V., Kent, J. T., Bibby, J. M.; *Multivariate Analysis*; London: Academic Press (1979).