CONJOINT

A Mu Sigma University Publication, March 2016

Mu Sigma

In This Issue...

Abstract

Although most software packages classify Principal Component Analysis (PCA) under Factor Analysis, as a procedure for variable reduction, PCA is in essence rather different from Factor Analysis. This article discusses some details of these two techniques with a view to pointing out the difference.

Principal Components Analysis and Factor Analysis

- By T. Krishnan

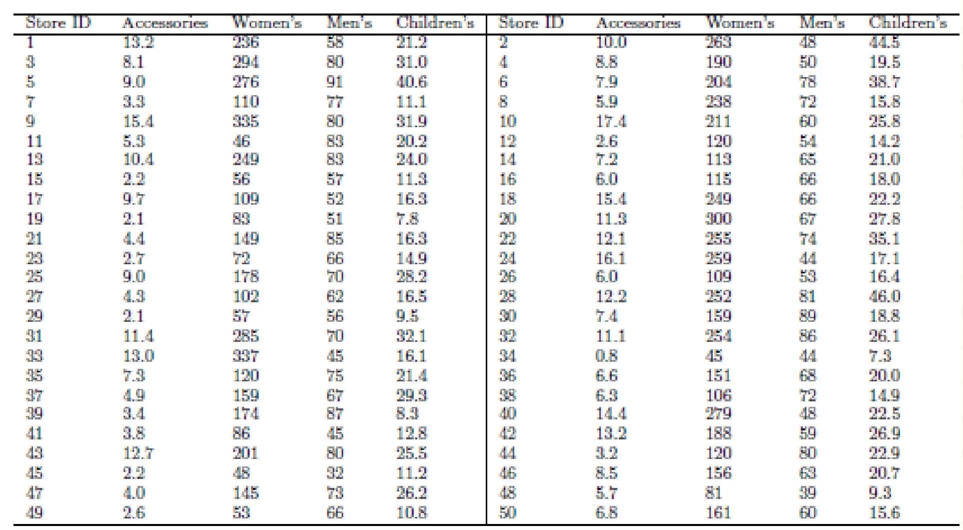
Common Theme of the Methods

Both Principal Component Analysis and Common Factor Analysis are methods to decompose a covariance or a correlation matrix with a view to reduce the dimensionality of the data. Although they are based on different mathematical models, they can be applied on the same data and more often than not produce similar results. The results generally consist of fewer components or factors than the number of variables in the data. This leads to useful interpretation of the components and factors, which can be used in further analysis like regression, ANOVA, discriminant analysis, cluster analysis, etc.

What are Principal Components?

Let us consider the following data set where we have the sales figures of the 50 stores of an apparel chain. The data consists of sales of the four departments—women's, men's, children's, and accessories, for a particular month in '0000 USD.

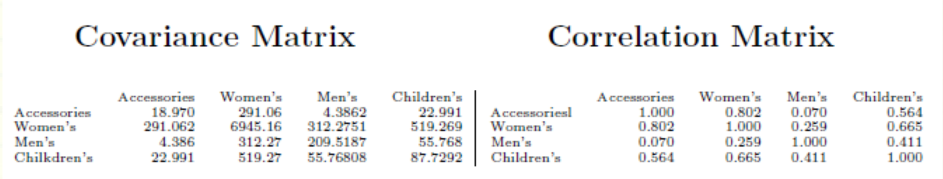
Data on Sales of Stores in Different Departments



Store ID	Accessories	Women's	Men's	Children's	Store ID	Accessories	Women's	Men's	Children's
1	13.2	236	58	21.2	2	10.0	263	48	44.5
3	8.1	294	80	31.0	4	8.8	190	50	19.5
5	9.0	276	91	40.6	6	7.9	204	78	38.7
7	3.3	110	77	11.1	8	5.9	238	72	15.8
9	15.4	335	80	31.9	10	17.4	211	60	25.8
11	5.3	46	83	20.2	12	2.6	120	54	14.2
13	10.4	249	83	24.0	14	7.2	113	65	21.0
15	2.2	56	57	11.3	16	6.0	115	66	18.0
17	9.7	109	52	16.3	18	15.4	249	66	22.2
19	2.1	83	51	7.8	20	11.3	300	67	27.8
21	4.4	149	85	16.3	22	12.1	255	74	35.1
23	2.7	72	66	14.9	24	16.1	259	44	17.1
25	9.0	178	70	28.2	26	6.0	109	53	16.4
27	4.3	102	62	16.5	28	12.2	252	81	46.0
29	2.1	57	56	9.5	30	7.4	159	89	18.8
31	11.4	285	70	32.1	32	11.1	254	86	26.1
33	13.0	337	45	16.1	34	0.8	45	44	7.3
35	7.3	120	75	21.4	36	6.6	151	68	20.0
37	4.9	159	67	29.3	38	6.3	106	72	14.9
39	3.4	174	87	8.3	40	14.4	279	48	22.5
41	3.8	86	45	12.8	42	13.2	188	59	26.9
43	12.7	201	80	25.5	44	3.2	120	80	22.9
45	2.2	48	32	11.2	46	8.5	156	63	20.7
47	4.0	145	73	26.2	48	5.7	81	39	9.3
49	2.6	53	66	10.8	50	6.8	161	60	15.6

In This Issue. . .

The covariance and correlation matrices of the four variables are as follows:



Covariance Matrix

	Accessories	Women's	Men's	Children's
Accessories	18.970	291.06	4.3862	22.991
Women's	291.062	6945.16	312.2751	519.269
Men's	4.386	312.27	209.5187	55.768
Chilkdren's	22.991	519.27	55.76808	87.7292

Correlation Matrix

	Accessories	Women's	Men's	Children's
Accessoriesl	1.000	0.802	0.070	0.564
Women's	0.802	1.000	0.259	0.665
Men's	0.070	0.259	1.000	0.411
Children's	0.564	0.665	0.411	1.000

The issue dealt with by principal components is a representation of the four dimensions (variables) in terms of a smaller number (one or more) of dimensions without losing much information. The notion of information used in this context is the variation among the observational units of the four variables. The four dimensions together have a variance of 7261.384 (sum of the diagonal elements of the covariance matrix given above). Can a large part of this total variance be captured by one or more combinations of the four dimensions? Let us simplify this question by asking for only linear combinations of the four dimensions represented by a general form:

$$Y_1 = \ell_1 Accessories + \ell_2 Women's + \ell_3 Men's + \ell_4 Children's$$

If we ask for only one combination, the question reduces to asking for $\ell_1, \ell_2, \ell_3, \ell_4$ such that Y_1 has the largest variance among all choices of $\ell_1, \ell_2, \ell_3, \ell_4$. But this question is ill formulated; for, if we find one such Y_1, then cY_1 for $c > 1$ will have $c^2 Var(Y_1)$ and so this variance can be made infinite. But then Y_1 and cY_1 represent the same variable in different units. Thus a meaningful formulation of the problem can be to constrain the $\ell_1, \ell_2, \ell_3, \ell_4$ in a way. One such way is to make them relative, that is by saying $\ell_1 + \ell_2 + \ell_3 + \ell_4 = 1$, in which case the above ill-posed nature of the problem will disappear. Another way is to make $\ell_1^2 + \ell_2^2 + \ell_3^2 + \ell_4^2 = 1$. Let us call such a linear combination a **normalized** one. Although both constraints are reasonable, the latter leads to mathematically and computationally elegant and convenient solutions.

In our example, the normalized linear combination with the largest variance is

$$0.0412 Accessories + 0.995 Women's + 0.0463 Men's$$

$$+0.075 Children's$$

with a variance of 7010.734. This is called the **first principal component** of the covariance matrix. It captures a proportion of 0.9655 of the total variance of 7261.384. This component is heavily weighted towards the Women's clothing variable, the reason being that Women's clothing has a large variance relative to the other variables. We shall discuss this issue of whether we should consider different variables with largely different variances for this analysis in a later section.

In situations where the first principal component does not capture an adequate amount of the overall variance, we might want another component. This component should ideally not contain information (variance) contained in the first component. A statistical way of formulating this idea is to ask for another component which is uncorrelated with the first one and among such has the largest variance. This is called the second principal component. If there are p dimesnions in the original data, we can thus extract p components with decreasing variance, each uncorrelated with all the earlier ones. These then will have a total variance the same as that of the original variables; the components will then be just a transformed version of the original variables. A useful situation is one where the few first components explain a large amount of the total variance.

The mathematical solution to this problem is the computation of eigenvalues and eigenvectors of the covariance or the correlation matrix. These eigenvalues will all be non-negative (0 will be an eigenvalue if and only if the matrix is singular). If the eigenvalues are arranged in decreasing order then they represent the variances of the first, second, ... components and the corresponding eigenvectors give the coefficients of the linear combinations representing the corresponding principal components.

In the example the second, third, and the fourth components are:

$$-0.0448 Accessories - 0.0588 Women's + 0.9769 Men's + 0.2007 Children's$$

$$0.07989 Accessories - 0.0676 Women's - 0.2005 Men's + 0.9741 Children's$$

$$-0.9949 Accessories + 0.0389 Women's - 0.0582 Men's + 0.0723 Children's$$

with variances 202.01, 42.12, and 6.17 respectively of the total 7261.38 with proportions 0.0278, 0.0058, and 0.00085 respectively.

The first component has positive signs for all the variables and so can be interpreted as an overall sales level, whereas the other components are differences of various kinds.

Covariance or Correlation

The correlation matrix is the covariance matrix of standardized variables. Thus all the variances are 1 as you notice in the correlation matrix above. If your variables are measured on very different scales, you may want to work with correlations rather than covariances. In that case the variances of the four components are:
2.480, 0.990, 0.357, and 0.173
respectively adding upto 4. The proportions explained are:
0.620, 0.247, 0.089, 0.043
respectively. The components are given by
Component 1: 0.536Accessories+ 0.583Women's+ 0.278Men's + 0.543Children's
Component 2: 0.418Accessories+ 0.188Women's −0.873Men's −0.167Chilkdren's
Component 3: −341Accessories −0.268Women's −0.378Men's + 0.818Children's
Component 4: 0.649Accessories −0.743Women's + 0.134Men's+ 0.000Children's

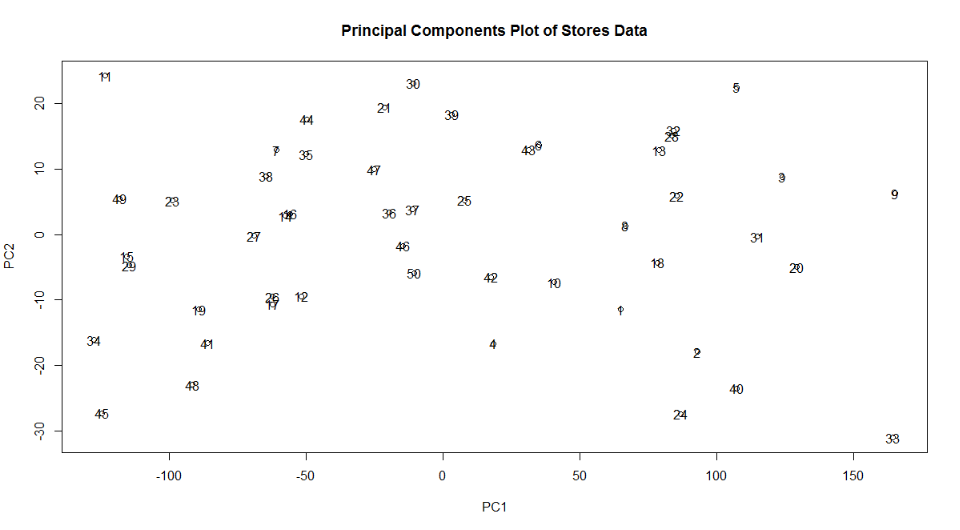
An explanation for the coefficients (loadings) is that each coefficient is the correlation between the component and the variable concerned. As before, the first component is a kind of sales level being a weighted average of the four sales figures. The other components are differences of various kinds. The first component explains only about 62% of variance and the first two together about 87%.

CONJOINT

A Mu Sigma University Publication, March 2016

Mu Sigma

For each case (store) in the data, scores based on each of the principal components can be computed. Since the first two components capture a great proportion of the data, a plot of the first two component scores for each case can be plotted as a scatter plot. This can sometimes help in finding clusters in the data.



Factor Analysis

We have seen how principal component analysis is a procedure for computing new variables that summarize variation in reduced space parsimoniously. The first principal component of the correlation matrix was
0.536Accessories+ 0.583Women's+ 0.278Men's + 0.543Children's

This equation is of the form:
Component = Linear combination of observed variables
Factor analysis turns this equation around:
Observed variable = Linear combination of Factors + Error
In mathematical terms the factor model can be written as

$$\boldsymbol{y} = \boldsymbol{\Lambda x} + \boldsymbol{z}$$

where $\boldsymbol{y}$ is a p-vector of observed variables, $\boldsymbol{x}$ is a k-vector ($k < p$) of latent (unobserved) variables, $\boldsymbol{z}$ is a p-vector of the so-called unique scores, $\boldsymbol{\Lambda}$ is a $p \times k$ matrix of factor loadings. It is assumed that $E(\boldsymbol{x}) = E(\boldsymbol{y}) = \underset{\sim}{0}, E(\boldsymbol{xx}^T) = \boldsymbol{\Phi}, E(\boldsymbol{zz}^T) = \boldsymbol{\Psi}$, a diagonal matrix. Although this looks like a linear regression model, it is not such a model since there are no unique observable or of factor scores or residuals to examine.

Factor analysts are less interested in prediction than in decomposing a covariance or a correlation matrix. Hence the fundamental equation of factor analysis is not in terms of the linear model stated above, but its quadratic form as:
Observed covariances = Factor covariances + Error Covariances

In This Issue...

The mathematical version of this is:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\Phi\mathbf{\Lambda}^T + \mathbf{\Psi}$$

where $\mathbf{\Sigma}$ is the $p \times p$ covariance matrix, $\mathbf{\Phi}$ is a $k \times k$ matrix of factor correlations and $\mathbf{\Psi}$ is a $p \times p$ matrix of unique/error variances. The diagonal elements of matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are hypothetical ones arrived at by the analysis and are known respectively as **communalities** and **specificities**. Thus factor analysis expresses variation within and relations among the observed variables partly as common variation among factors and partly as specific variation among random errors.

From the correlation or covariance matrix, factor loadings are estimated. This is called **initial factor extraction**. There are many methods for doing this, like generalized least squares and maximum likelihood.

The initial factors or the principal components do not necessarily give interpretable factors or components. Factors or components are **rotated** in order to make them interpretable. This is achieved by making the large loadings larger and the smaller ones smaller so that each variable is associated with a small number of factors or components. It is hoped that variables that load strongly with a given factor or component will have a clear meaning in the context of the data.

Generally, the initial factor analysis called Exploratory Factor Analysis is followed with Confirmatory Factor Analysis to confirm the hypothesized factor structure and to validate it by computing goodness-of-fit measures.

In the stores sales example, only one factor could be extracted since the number of parameters to be estimated will be too high for more than one factor. The factor scores are 0.818, 0.979, 0.262, 0.683 which are correlations of the four departmental sales with the latent factor extracted. This shows that the factor is an overall weighted summary of the four sales figures.

Principal Components versus Factor Analysis

One of the main differences between principal components and factors is that factors are indeterminate or latent whereas principal components are explicit. Unlike principal components, there are no natural factor scores for observational units. The reason is that in factor analysis there are more indeterminable parameters than observations. Interpretations have to be done, as done above, by considering factor weights as correlations of factors with the variables.

References

- Frank Walkey and Garry Welch (2010): *Demystifying Factor Analysis:: How It Works and How To Use It*. Xlibris, corp.
- Marjorie A. Pett, Nancy R. Lackey, and John J. Sullivan (2003): *Making Sense of Factor Analysis: The Use of Factor Analysis for Instrument Development in Health Care Research*. Sage Publications.