

PRINCIPAL COMPONENT ANALYSIS

By:

Dr. Akash Asthana

Assistant Professor,

Dept. of Statistics,

University of Lucknow, Lucknow

PRINCIPAL COMPONENT ANALYSIS

- ✘ It is a dimension reduction technique.
- ✘ In some situations the measurements are taken over a large number of variables.
- ✘ But it is not possible to deal with a large number of variables.
- ✘ Therefore instead of these large number of variables their linear combinations, which are linearly independent and orthonormal also, are used which can explain maximum possible variation in the data.
- ✘ These linear combinations are called as principal components.

PRINCIPAL COMPONENT ANALYSIS

- ✘ Transforming the original vector variable to the vector of principal components amounts to a rotation of coordinate axes to a new coordinate system that has inherent statistical properties.
- ✘ The set of principal components yields a convenient set of coordinates, and the accompanying variances of the components characterize their statistical properties.
- ✘ The method of principal components is used to find the linear combinations with large variance.

PRINCIPAL COMPONENT ANALYSIS

- ✘ Let X be a random vector of p variables having variance-covariance matrix Σ . Without loss of generality the mean vector of X can be taken as 0.
- ✘ The main objective of Principal Component Analysis is to obtain the linear combinations of X vector in a manner that the variance of the combination is maximum.
- ✘ Let the linear combination of X is $\beta'X$.
- ✘ Then $V(\beta'X) = \beta'\Sigma\beta$. (1)
- ✘ As these linear combinations are orthonormal we will maximize this variance under the condition $\beta'\beta = 1$.

PRINCIPAL COMPONENT ANALYSIS

- ✘ For the purpose we define the function:

$$\varphi = \beta' \Sigma \beta - \lambda(\beta' \beta - 1)$$

Where λ is Lagrange's multiplier (a scalar quantity).

- ✘ For maximum variance: $\frac{\partial \varphi}{\partial \beta'} = 0$

- ✘ It will give the equation: $\Sigma \beta - \lambda \beta = 0$

or $(\Sigma - \lambda I)\beta = 0$ (2)

- ✘ This equation will have a solution if $(\Sigma - \lambda I)$ will be singular i.e. $|\Sigma - \lambda I| = 0$. (3)

- ✘ In other words we can say that λ is the characteristic root of Σ and β be the characteristic vector of Σ .

PRINCIPAL COMPONENT ANALYSIS

- ✘ Also from equation (1), we get: $\beta' \Sigma \beta = \lambda \beta' \beta = \lambda$.
- ✘ Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ be the characteristic vectors of the matrix Σ .
- ✘ As the Linear combination must have maximum variance we take $\lambda = \lambda_1$ and $\beta^{(1)}$ be the characteristic vector associated with it.
- ✘ Therefore first principal component for X matrix is given by $U_1 = \beta^{(1)'} X$ where $\beta^{(1)}$ is the characteristic vector associated with the maximum characteristic root λ_1 of Σ .
- ✘ Let us consider another principal component $\beta' X$ of X which have maximum variance (lower than U_1) and is uncorrelated with $\beta^{(1)'} X$.
- ✘ Then we would have:
$$V(\beta' X) = \beta' \Sigma \beta; \text{Cov}(\beta' X, \beta^{(1)'} X) = \beta' \Sigma \beta^{(1)} = 0; \beta' \beta = 1 \quad (4)$$
- ✘ Now for obtaining the second principal component we maximize $\beta' \Sigma \beta$ under the conditions $\beta' \beta = 1$ and $\beta' \Sigma \beta^{(1)} = 0$.

PRINCIPAL COMPONENT ANALYSIS

- ✘ Now we define a function:

$$\varphi_1 = \beta' \Sigma \beta - \lambda(\beta' \beta - 1) - 2\nu \beta' \Sigma \beta^{(1)}$$

Where λ and ν are the Lagrange's multipliers (Scalars).

- ✘ On maximizing the φ with respect to β' we get:

$$2\Sigma\beta - 2\lambda\beta - 2\nu\Sigma\beta^{(1)} = 0 \quad (5)$$

- ✘ Pre-multiplying it by $\beta^{(1)'}$ we get

$$2\beta^{(1)'} \Sigma \beta - 2\lambda \beta^{(1)'} \beta - 2\nu \beta^{(1)'} \Sigma \beta^{(1)} = 0$$

$$\text{or } -2\nu\lambda = 0$$

- ✘ As λ can't be we get $\nu = 0$. putting it in (5) we get $(\Sigma - \lambda I)\beta = 0$.
- ✘ Which again show that β is the characteristic vector of matrix Σ and λ is its Eigen root.
- ✘ We can define this principal component as $U_2 = \beta^{(2)'} X$, where $\beta^{(2)}$ is the solution of equation (2) for the $\lambda = \lambda_2$.

PRINCIPAL COMPONENT ANALYSIS

- ✘ Proceeding in same manner at $(k+1)^{\text{th}}$ stage we get following conditions:

$$V(\beta'X) = \beta' \Sigma \beta \quad (6)$$

$$\beta' \beta = 1 \quad (7)$$

$$\text{Cov}(\beta'X, \beta^{(i)'}X) = \beta' \Sigma \beta^{(i)} \quad \forall i = 1, 2, \dots, k \quad (8)$$

- ✘ Proceeding in same manner and using Lagrange's multiplier we can define the function $\varphi_{(k+1)}$ as:

$$\varphi_{(k+1)} = \beta' \Sigma \beta - \lambda(\beta' \beta - 1) - 2 \sum_{i=1}^k v_i \beta' \Sigma \beta^{(i)} \quad (9)$$

- ✘ On maximizing the equation (9) with respect to β' we again get that β is the characteristic vector of the matrix Σ corresponding to $(k+1)^{\text{th}}$ largest characteristic root ($\lambda_{(k+1)}$) of it.
- ✘ Also the variance of the $(k+1)^{\text{th}}$ principal component is $\lambda_{(k+1)}$.
- ✘ $V(\beta^{(k+1)'}X) = \beta^{(k+1)'} \Sigma \beta^{(k+1)} = \lambda_{(k+1)} \beta^{(k+1)'} \beta^{(k+1)} = \lambda_{(k+1)}$ by using equations (1) and (2).

PRINCIPAL COMPONENT ANALYSIS

✘ In matrix form we can define:

$$\boldsymbol{\beta} = (\beta^{(1)} \ \beta^{(2)} \ \dots \ \beta^{(p)})' \text{ and } \Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

✘ Then we can define principal components as:

“ Let \mathbf{X} be a p -component random vector having mean 0 and variance-covariance matrix Σ . Then there exist an orthogonal linear transformation $\mathbf{U} = \boldsymbol{\beta}'\mathbf{X}$ such that the covariance matrix of \mathbf{U} is Λ , where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ are the roots of equation (3). The k^{th} column of $\boldsymbol{\beta}$, $\beta^{(k)}$ satisfies the equation (2). The k^{th} component of \mathbf{U} , $U_k = \beta^{(k)'}\mathbf{X}$ has maximum variance of all normalized linear combinations uncorrelated with $U_1, U_2, \dots, U_{(k-1)}$.”

PRINCIPAL COMPONENT ANALYSIS

- ✘ In general situation the variance-covariance matrix Σ is unknown. Therefore for obtaining the principal components its MLE is used and thus obtained principal components are called as MLE of principal components.
- ✘ As we know that for a square matrix of order m there will be at most m characteristic roots. Therefore for a p component matrix X one can obtain at most p -principal components.

STEPS FOR PRINCIPAL COMPONENT ANALYSIS:

1. First transform the matrix of all variables under consideration to a matrix X such that mean of X will be 0.
2. Obtain the Variance-covariance matrix of X , Σ (or its MLE) under the assumption that X is Normally Distributed.
3. Obtain the Characteristic roots of Σ and arrange them in descending order ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$).
4. For each distinct Eigen root obtain Eigen vector.
5. Normalize these Eigen vectors dividing these by their norms ($\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(p)}$).
6. Then obtain the principal components by multiplying these β_i 's with X (i.e. $\beta^{(1)}X, \beta^{(2)}X, \dots, \beta^{(p)}X$)
7. In the situation if the unit of measurements for variables are not same it is better to use correlation matrix in place of variance-covariance matrix.

PRINCIPAL COMPONENT ANALYSIS: AN EXAMPLE USING SPSS

- ✘ For performing Principal Component Analysis (PCA) using SPSS Following steps are used.
- ✘ Click on Analyze → Dimension Reduction → Factor

The screenshot displays the IBM SPSS Statistics Data Editor interface. The 'Analyze' menu is open, and the path 'Dimension Reduction' > 'Factor' is highlighted. The main window shows a list of variables with their names and types. The 'Factor' dialog box is also visible, showing the 'Factor' option selected under 'Dimension Reduction'.

Name	Type
1 manufact	String
2 model	String
3 sales	Numeric
4 resale	Numeric
5 type	Numeric
6 price	Numeric
7 engine_5	Numeric
8 horsepower	Numeric
9 wheelbae	Numeric
10 width	Numeric
11 length	Numeric
12 curb_wgt	Numeric
13 fuel_cap	Numeric
14 mpg	Numeric
15 inakes	Numeric
16 zresale	Numeric
17 ztype	Numeric
18 zprice	Numeric
19 zengine_	Numeric
20 zhorsepo	Numeric
21 zwheelba	Numeric
22 zwidth	Numeric
23 zlength	Numeric
24 zcurb_wg	Numeric

EXAMPLE (CONTD.)

- ✘ It will open the factor analysis window put all the variables required for PCA in variable box. Then click on Extraction.

The screenshot displays the IBM SPSS Statistics Data Editor interface. The main window shows a list of variables with columns for Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure, and Role. A 'Factor Analysis' dialog box is open in the foreground, showing a list of variables on the left and a 'Variables' list on the right. The 'Variables' list includes Price in thousand, Engine size (angl), Horsepower (hor), Wheelbase (whe), Width (width), Length (length), and Curb weight (curb). The 'Extraction' button is highlighted in the dialog box.

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role	
1	manufact	String	13	0	Manufacturer	None	None	13	Left	Nominal	Input
2	model	String	17	0	Model	None	None	17	Left	Nominal	Input
3	sales	Numeric	11	3	Sales in thousa...	None	None	8	Right	Scale	Input
4	resale	Numeric	11	3	4-year resale va...	None	None	8	Right	Scale	Input
5	type	Numeric	11	0	Vehicle	None	None	8	Right	Scale	Input
6	price	Numeric	11	3	Price in	None	None	8	Right	Scale	Input
7	engine_s	Numeric	11	1	Engine	None	None	8	Right	Scale	Input
8	horsepow	Numeric	11	0	Horsepo	None	None	8	Right	Scale	Input
9	wheelbas	Numeric	11	1	Wheelba	None	None	8	Right	Scale	Input
10	width	Numeric	11	1	Width	None	None	8	Right	Scale	Input
11	length	Numeric	11	1	Length	None	None	8	Right	Scale	Input
12	curb_wgt	Numeric	11	3	Curb we	None	None	8	Right	Scale	Input
13	fuel_cap	Numeric	11	1	Fuel ca	None	None	8	Right	Scale	Input
14	mpg	Numeric	11	0	Fuel eff	None	None	8	Right	Scale	Input
15	lnsales	Numeric	8	2	Log-tra	None	None	8	Right	Scale	Input
16	zresale	Numeric	11	5	Zscore:	None	None	8	Right	Scale	Input
17	ztype	Numeric	11	5	Zscore:	None	None	8	Right	Scale	Input
18	zprice	Numeric	11	5	Zscore:	None	None	8	Right	Scale	Input
19	zengine_	Numeric	11	5	Zscore:	None	None	8	Right	Scale	Input
20	zhorsepo	Numeric	11	5	Zscore: Horse...	None	None	8	Right	Scale	Input
21	zwheelba	Numeric	11	5	Zscore: Wheel...	None	None	8	Right	Scale	Input
22	zwidth	Numeric	11	5	Zscore: Width	None	None	8	Right	Scale	Input
23	zlength	Numeric	11	5	Zscore: Length	None	None	8	Right	Scale	Input
24	zcurb_wg	Numeric	11	5	Zscore: Curb w...	None	None	8	Right	Scale	Input

EXAMPLE (CONTD.)

- ✘ On clicking Extraction window will be open. Click on Correlation matrix and then fixed number of factors put the number of variables in the analysis in the box shown.

The screenshot shows the IBM SPSS Statistics Data Editor interface. A data table is visible in the background with columns for Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure, and Role. The 'Factor Analysis: Extraction' dialog box is open in the foreground. The 'Method' is set to 'Principal components'. Under the 'Analyze' section, the 'Correlation matrix' radio button is selected. Under the 'Extract' section, the 'Fixed number of factors' radio button is selected, and the 'Factors to extract' field is set to 0. The 'Maximum iterations for convergence' is set to 25. The 'Continue', 'Cancel', and 'Help' buttons are visible at the bottom of the dialog box.

EXAMPLE (CONTD.)

- ✘ Click on continue then click on Scores → Save as variable → Display factor score coefficient matrix.

The screenshot displays the IBM SPSS Statistics Data Editor interface. The main window shows a list of variables with columns for Name, Type, Width, Decimals, Label, Values, Missing, Columns, Align, Measure, and Role. A 'Factor Analysis' dialog box is open, and the 'Factor Analysis: Factor Scores' sub-dialog box is also open. The 'Save as variables' checkbox is checked, and the 'Display factor score coefficient matrix' checkbox is also checked. The 'Method' section has 'Regression' selected. The 'Factor Scores' sub-dialog box has buttons for 'Descriptives...', 'Extraction', 'Rotation', 'Scores', and 'Options...'. The 'Factor Analysis' dialog box has 'Continue', 'Cancel', and 'Help' buttons. The 'Data View' and 'Variable View' tabs are visible at the bottom left.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	manufact	String	13	0	Manufacturer	None	None	13	Left	Nominal	Input
2	model	String	17	0	Model	None	None	17	Left	Nominal	Input
3	sales	Numeric	11	3	Sales in thousa...	None	None	8	Right	Scale	Input
4	resale	Numeric	11	3	4-year resale va...	None	None	8	Right	Scale	Input
5	type	Numeric	11	0	Vehicle type	None	None	8	Right	Scale	Input
6	price	Numeric	11	3	Price in thousa...	None	None	8	Right	Scale	Input
7	engine_s	Numeric	11	1	Engine size	None	None	8	Right	Scale	Input
8	horsepow	Numeric	11	0	Horsepower	None	None	8	Right	Scale	Input
9	wheelbas	Numeric	11	1	Wheelbase	None	None	8	Right	Scale	Input
10	width	Numeric	11	1	Width	None	None	8	Right	Scale	Input
11	length	Numeric	11	1	Length	None	None	8	Right	Scale	Input
12	curb_wgt	Numeric	11	3	Curb weight	None	None	8	Right	Scale	Input
13	fuel_cap	Numeric	11	1	Fuel capacity	None	None	8	Right	Scale	Input
14	mpg	Numeric	11	0	Fuel efficiency	None	None	8	Right	Scale	Input
15	lnsales	Numeric	8	2	Log-transformed sales	None	None	8	Right	Scale	Input
16	zresale	Numeric	11	5	Zscore: 4-year resale	None	None	8	Right	Scale	Input
17	ztype	Numeric	11	5	Zscore: Vehicle type	None	None	8	Right	Scale	Input
18	zprice	Numeric	11	5	Zscore: Price	None	None	8	Right	Scale	Input
19	zengine_	Numeric	11	5	Zscore: Engine size	None	None	8	Right	Scale	Input
20	zhorsepo	Numeric	11	5	Zscore: Horsepower	None	None	8	Right	Scale	Input
21	zwheelba	Numeric	11	5	Zscore: Wheelbase	None	None	8	Right	Scale	Input
22	zwidth	Numeric	11	5	Zscore: Width	None	None	8	Right	Scale	Input
23	zlength	Numeric	11	5	Zscore: Length	None	None	8	Right	Scale	Input
24	zcurb_wg	Numeric	11	5	Zscore: Curb weight	None	None	8	Right	Scale	Input
25	zfuel_ca	Numeric	11	5	Zscore: Fuel capacity	None	None	8	Right	Scale	Input
26	zmpg	Numeric	11	5	Zscore: Fuel efficiency	None	None	8	Right	Scale	Input

EXAMPLE (CONTD.)

- ✘ On applying the PCA using SPSS five tables are generated out of which three are important for explaining PCA.
- ✘ Table – 1: Total Variance Explained

Component	Eigen Value	% of Variance	Cumulative %	Component	Eigen Value	% of Variance	Cumulative %
1	5.804	64.490	64.490	6	0.155	1.719	96.491
2	1.517	16.860	81.349	7	0.139	1.547	98.038
3	0.623	6.918	88.267	8	0.114	1.266	99.305
4	0.338	3.757	92.025	9	0.063	0.695	100.000
5	0.247	2.747	94.772				

EXAMPLE (CONTD.)

- ✘ This table shows that nine principal components are generated against the 9 variables. Second column represents the Eigen value of correlation matrix. Third and fourth column shows the Percentage of variance and cumulative percentage of variance explained by these components.
- ✘ As we know that Eigen values are arranged in the descending order. In this example we can see that this process is followed. For first component the Eigen value is 5.804 which is largest among all the Eigen value. It also shows that the variance of first principal component is 5.804. The % of variance is 64.490 which is also equal to cumulative %. It shows that first principal component explain the 64.490% of total variation in the data. For second component cumulative % is 81.439, which shows that both, first and second component collectively explain the 81.439% of total variation in the data.

EXAMPLE (CONTD.)

✦ Table 2: Component Score Coefficient Matrix

	1	2	3	4	5	6	7	8	9
Price in thousands	0.105	-0.457	0.233	0.822	0.713	-0.399	-0.226	0.153	-1.773
Engine size	0.152	-0.160	0.183	-0.665	-0.838	1.068	0.742	0.779	-1.603
Horsepower	0.133	-0.351	0.435	-0.022	-0.370	-0.233	0.396	-0.660	2.765
Wheelbase	0.124	0.388	0.183	0.655	-0.110	-0.811	1.385	1.115	0.200
Width	0.143	0.159	0.282	-1.026	1.260	-0.510	-0.031	-0.181	-0.143
Length	0.126	0.337	0.545	0.351	-0.552	0.279	-1.344	-1.120	-0.652
Curb weight	0.159	0.025	-0.353	0.190	0.304	0.819	-1.261	1.616	1.367
Fuel capacity	0.149	0.078	-0.605	0.337	0.418	0.989	0.915	-1.578	-0.010
Fuel efficiency	-0.0146	0.070	0.654	0.289	0.725	1.570	0.522	0.278	0.517

EXAMPLE (CONTD.)

- ✘ The Second table shows the coefficients for each variable in the corresponding principal components.
- ✘ It actually represents the Eigen vectors for the correlation matrix of variables.
- ✘ First Principal component can be written as:

$$U_1 = 0.105(\text{Price in Thousand}) + 0.152(\text{Engine Size}) + 0.133(\text{Horse Power}) + 0.124(\text{Wheel Base}) + 0.143(\text{Width}) + 0.126(\text{Length}) + 0.159(\text{Curb Weight}) + 0.149(\text{Fuel Capacity}) + (-0.146)(\text{Fuel Efficiency})$$

- ✘ In the same manner other Principal components can be written.

EXAMPLE (CONTD.)

✦ Table 3: Component score coefficient matrix.

Component	1	2	3	4	5	6	7	8	9
1	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
6	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000

✦ This table shows that all the principal components are linearly independent of each other and are normalized.

REFERENCES

1. Anderson TW, An introduction to Multivariate Statistical Analysis, 3rd Edition, John Wiley & Sons Inc., New Jersey.
2. Malhotra NK, Birks DF, Marketing Research an Applied Approach, 4th Edition, Prentice Hall, New Delhi.
3. Johnson RA, Wichern DW, Applied Multivariate Statistical Analysis, 3rd Edition, Prentice Hall, New Delhi.
4. Morrison DF, Multivariate Statistical Methods, 2nd Edition, McGraw Hill Publication, India.