

# How to Perform Principal Components Analysis in SAS?

Authored by  
**stats writer**

November 19, 2025

## RECOMMENDED CITATION

stats writer (2025). *How to Perform Principal Components Analysis in SAS?*.  
PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=96759>

Principal Components Analysis (PCA) is an indispensable statistical technique falling under the umbrella of dimensionality reduction. Its primary goal is to simplify complex datasets by transforming a large set of correlated variables into a smaller set of uncorrelated variables, known as principal components, while retaining as much information--or variance--as possible. This transformation is crucial in fields like machine learning, finance, and bioinformatics where high-dimensional data often complicates modeling and interpretation.

In essence, PCA works by identifying the directions (or axes) along which the variation in the data is maximal. The first principal component (PC1) accounts for the largest possible variance, and each succeeding component accounts for the largest remaining variance, always being orthogonal (uncorrelated) to the preceding component. By capturing the underlying structure and minimizing redundancy among variables, PCA provides a powerful method for data visualization, noise filtering, and improving the efficiency of subsequent analytical processes.

When working within the SAS environment, the execution of PCA is streamlined through the dedicated procedure, **PROC PRINCOMP**. This powerful tool handles all the necessary calculations, including standardizing variables, determining eigenvalues and eigenvectors, and computing the scores for the new principal components. The resulting output from this procedure offers a comprehensive view, detailing the relative importance of each component and providing standardized values for every observation based on these new components.

## Leveraging PROC PRINCOMP in SAS: Syntax and Parameters

The most straightforward and efficient way to conduct Principal Components Analysis within the robust SAS software is by utilizing the **PROC PRINCOMP** statement. This procedure is specifically designed for performing principal component analysis, factoring, and scoring. To initiate the analysis, you only need to specify the input dataset and the variables you wish to analyze. Furthermore, you can optionally specify output datasets to store the component scores and the detailed statistics, which is highly recommended for reproducibility and further exploration.

The basic syntax for employing PROC PRINCOMP is clean and follows standard SAS conventions. We typically use the **DATA=** option to reference the source dataset, the **OUT=** option to save the new dataset containing the principal component scores appended to the original data, and the **OUTSTAT=** option to capture vital statistics necessary for interpretation, such as correlations, means, and eigenvectors. Defining the variables to be included in the analysis is done using the required **VAR** statement.

Here is a generalized representation of the fundamental syntax structure required to execute PCA using **PROC PRINCOMP**:

```
proc princomp data=my_data out=out_data outstat=stats;  
var var1 var2 var3;  
run;
```

Each statement within the procedure serves a distinct and essential purpose in defining the parameters of the analysis and managing the resulting output files:

**data:** Specifies the name of the input dataset containing the variables upon which the PCA will be performed.

**out:** Designates the name for the new dataset that will be created. This output file includes all the original data variables augmented by the calculated principal component scores for each observation.

**outstat:** Instructs SAS to generate a dataset specifically dedicated to housing the statistical outputs, including means, standard deviations, the correlation coefficients matrix, the eigenvalues, and the eigenvectors (or scoring coefficients).

**var:** A mandatory statement that explicitly lists the set of quantitative variables from the input dataset that are to be used in the principal components analysis calculation.

## Preparing the Data: A Practical Example

To demonstrate the functionality of **PROC PRINCOMP**, we will first establish a sample dataset. This example involves metrics collected from 20 different basketball players, providing a simple yet illustrative multi-dimensional space where PCA can be applied effectively. The dataset includes three quantitative variables: **points** scored, **assists** made, and **rebounds** secured. These variables, while distinct, are likely correlated, making them excellent candidates for a Principal Components Analysis to uncover underlying performance dimensions.

We use standard SAS data steps, including the **DATA**, **INPUT**, and **DATALINES** statements, to construct this dataset titled `my\_data`. Following the creation of the dataset, a subsequent `PROC PRINT` step is included to verify the data integrity and display the initial observations, ensuring that the input variables are correctly loaded and ready for analysis. The structure of the data step is critical, as it defines the raw data that **PROC PRINCOMP** will process.

The following code block outlines the creation and initial display of our sample dataset containing the 20 observations for basketball players:

```
/*create dataset*/  
data my_data;  
input points assists rebounds;  
datalines;
```

```
22 8 4
29 7 3
10 4 12
5 5 15
35 6 2
8 3 10
10 4 8
8 4 3
2 5 17
4 5 19
9 9 4
7 6 4
31 5 3
4 6 13
5 7 8
8 8 4
10 4 8
20 4 6
25 8 8
18 8 3
;
run;
```

```
/*view dataset*/
proc print data=my_data;
```

The dataset generated confirms the structure, showing three core metrics (points, assists, rebounds) for each player. This prepared data is now the input for the central PCA procedure.

| Obs | points | assists | rebounds |
|-----|--------|---------|----------|
| 1   | 22     | 8       | 4        |
| 2   | 29     | 7       | 3        |
| 3   | 10     | 4       | 12       |
| 4   | 5      | 5       | 15       |
| 5   | 35     | 6       | 2        |
| 6   | 8      | 3       | 10       |
| 7   | 10     | 4       | 8        |
| 8   | 8      | 4       | 3        |
| 9   | 2      | 5       | 17       |
| 10  | 4      | 5       | 19       |
| 11  | 9      | 9       | 4        |
| 12  | 7      | 6       | 4        |
| 13  | 31     | 5       | 3        |
| 14  | 4      | 6       | 13       |
| 15  | 5      | 7       | 8        |
| 16  | 8      | 8       | 4        |
| 17  | 10     | 4       | 8        |
| 18  | 20     | 4       | 6        |
| 19  | 25     | 8       | 8        |
| 20  | 18     | 8       | 3        |

## Executing PCA and Interpreting Core Statistics

With the data successfully loaded into the SAS environment, we proceed to perform the core PCA using **PROC PRINCOMP**. In this specific application, we are instructing the procedure to analyze the variance explained by the three performance metrics: **points**, **assists**, and **rebounds**. We ensure that the results are stored in output datasets named `out\_data` (for scores) and `stats` (for statistical measures), allowing for robust post-analysis manipulation and interpretation.

The use of the **VAR** statement is critical here, explicitly selecting only the quantitative variables of interest. This ensures that the analysis focuses solely on the dimensions we intend to reduce. Since we are using the default settings of **PROC PRINCOMP**, the analysis is performed on the correlation matrix of the variables, which implicitly standardizes the data (mean of 0, standard deviation of 1) before calculating the principal components, preventing variables with larger magnitudes (like points) from disproportionately influencing the results.

The following SAS code executes the required analysis, using our prepared dataset `my\_data`:

```
/*perform principal components analysis*/  
proc princomp data=my_data out=out_data outstat=stats;  
var points assists rebounds;  
run;
```

The initial output generated by **PROC PRINCOMP** provides crucial descriptive statistics, which are essential for quality checking the data and understanding the basis of the component calculation. This includes the mean and standard deviation for each input variable, giving us insight into the data's central tendency and spread. More importantly, SAS displays the correlation matrix, which quantifies the linear relationship between every pair of input variables--a necessary check, as PCA performs best when the original variables exhibit multicollinearity.

## The PRINCOMP Procedure

|              |    |
|--------------|----|
| Observations | 20 |
| Variables    | 3  |

| Simple Statistics |             |             |             |
|-------------------|-------------|-------------|-------------|
|                   | points      | assists     | rebounds    |
| Mean              | 13.50000000 | 5.800000000 | 7.700000000 |
| StD               | 10.06034424 | 1.765159900 | 5.120649630 |

| Correlation Matrix |        |         |          |
|--------------------|--------|---------|----------|
|                    | points | assists | rebounds |
| points             | 1.0000 | 0.2341  | -.6232   |
| assists            | 0.2341 | 1.0000  | -.3855   |
| rebounds           | -.6232 | -.3855  | 1.0000   |

| Eigenvalues of the Correlation Matrix |            |            |            |            |
|---------------------------------------|------------|------------|------------|------------|
|                                       | Eigenvalue | Difference | Proportion | Cumulative |
| 1                                     | 1.85089366 | 1.05544499 | 0.6170     | 0.6170     |
| 2                                     | 0.79544867 | 0.44179099 | 0.2651     | 0.8821     |
| 3                                     | 0.35365768 |            | 0.1179     | 1.0000     |

| Eigenvectors |          |          |          |
|--------------|----------|----------|----------|
|              | Prin1    | Prin2    | Prin3    |
| points       | 0.603439 | -.478471 | 0.637908 |
| assists      | 0.460853 | 0.862106 | 0.210683 |
| rebounds     | -.650750 | 0.166848 | 0.740733 |

## Understanding Variance and Dimensionality Reduction

The central results of the PCA are found in the section detailing the eigenvalues and eigenvectors. The eigenvalues represent the variance explained by each principal component. Since we started with three input variables, we generate exactly three principal components (Prin1, Prin2, Prin3). The magnitude of the eigenvalue directly correlates with the component's importance; larger eigenvalues signify components that capture a greater proportion of the total variation present in the dataset.

The table titled **Eigenvalues of the Correlation Matrix** is the primary resource for determining

how many components should be retained for analysis--a process often guided by Kaiser's criterion (retaining components with eigenvalues greater than 1) or by assessing the cumulative percentage of variance explained. This table not only lists the raw eigenvalues but also the proportion of variance explained by each component and the cumulative percentage, allowing analysts to make informed decisions about the effective dimensionality reduction achieved.

Analyzing the variance explained for our basketball data reveals the following distribution of total variation across the three principal components:

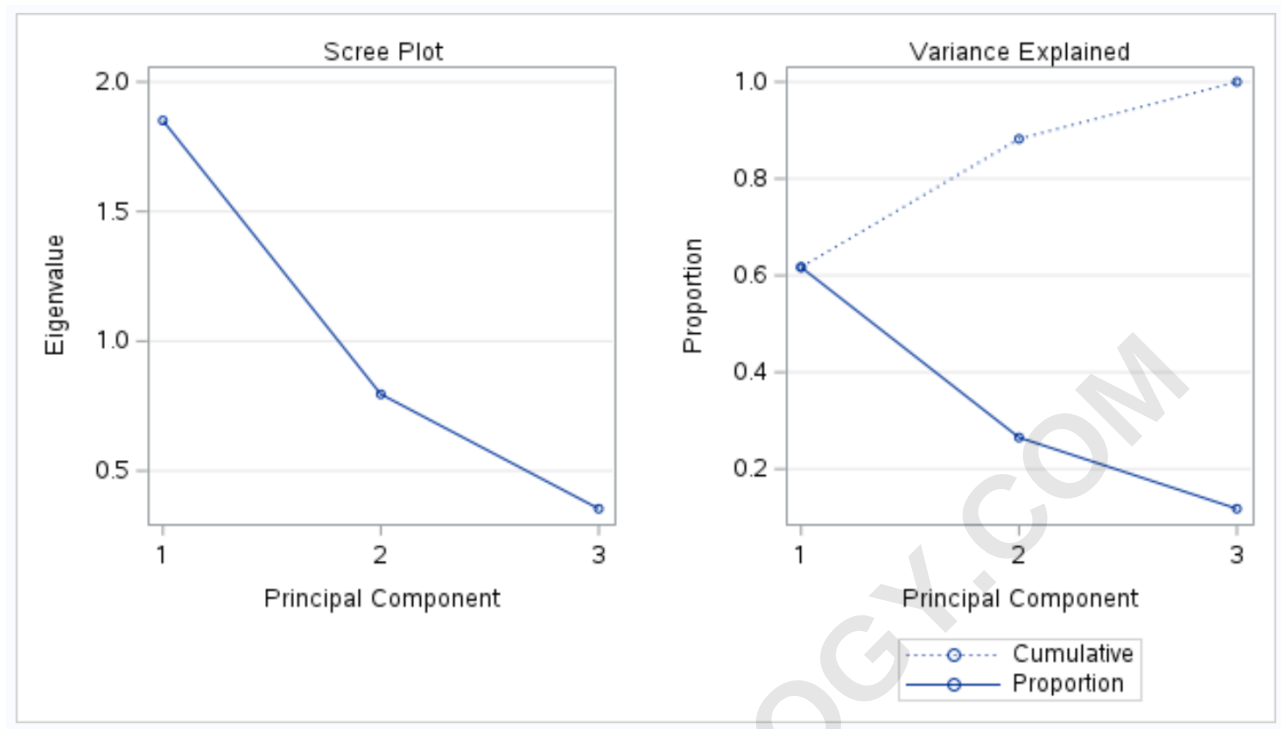
The first principal component (Prin1) is the most significant, accounting for **61.7%** of the total variation in the dataset. This component is the primary axis of variance.

The second principal component (Prin2) explains an additional **26.51%** of the total variation in the dataset.

The third principal component (Prin3) accounts for the remaining **11.79%** of the total variation in the dataset.

Crucially, when summing the percentages of the first two components ( $61.7\% + 26.51\%$ ), we find that they collectively explain **88.21%** of the total variance. This high cumulative percentage strongly suggests that we can effectively reduce the complexity of the dataset from three dimensions to two, sacrificing minimal informational content, thus achieving significant dimensionality reduction.

The procedure also generates two essential plots to visualize these findings: the **Scree Plot** and the **Variance Explained Plot**. The Scree Plot graphically displays the eigenvalues in descending order, helping identify the 'elbow' point where the marginal gain in variance explained diminishes rapidly. The Variance Explained Plot specifically visualizes the percentage contribution of each individual principal component (y-axis) against the component number (x-axis), providing a clear visual confirmation of the variance contribution breakdown calculated in the eigenvalues table.



## Visualizing PCA Results with a Biplot in SAS

After successfully calculating the principal components, the next critical step is visualizing the results to gain intuitive insights into how the original observations relate to the newly formed dimensions. The most effective visualization tool for PCA results is the biplot. A biplot displays both the observations (data points) and the variables (vectors) in a single two-dimensional plane, typically defined by the first two principal components (Prin1 and Prin2). This allows analysts to visually identify clusters of observations and understand how the original variables influence the creation of these components.

To generate a biplot in SAS, we leverage the results stored in the ``out_data`` dataset created by **PROC PRINCOMP**, which already contains the computed principal component scores (Prin1, Prin2, etc.) for every observation. We use the versatile **PROC SGPLOT** procedure to create the scatter plot. Before plotting, we must ensure we have a unique identifier for each observation, which we create by modifying the ``out_data`` to include an **obs** column representing the row number (``_n_``) of the original data.

The following code snippet first prepares the data by adding the observation index and then uses **PROC SGPLOT** with the **SCATTER** statement. We map Prin1 to the x-axis and Prin2 to the y-axis. The ``DATALABEL=obs`` option is crucial as it labels each data point with its corresponding observation number, facilitating direct linkage back to the original dataset for interpretation:

```
/*create dataset with column called obs to represent row numbers of original data*/
```

```
data biplot_data;
```

```
set out_data;
```

```
obs=_n_;
```

```
run;
```

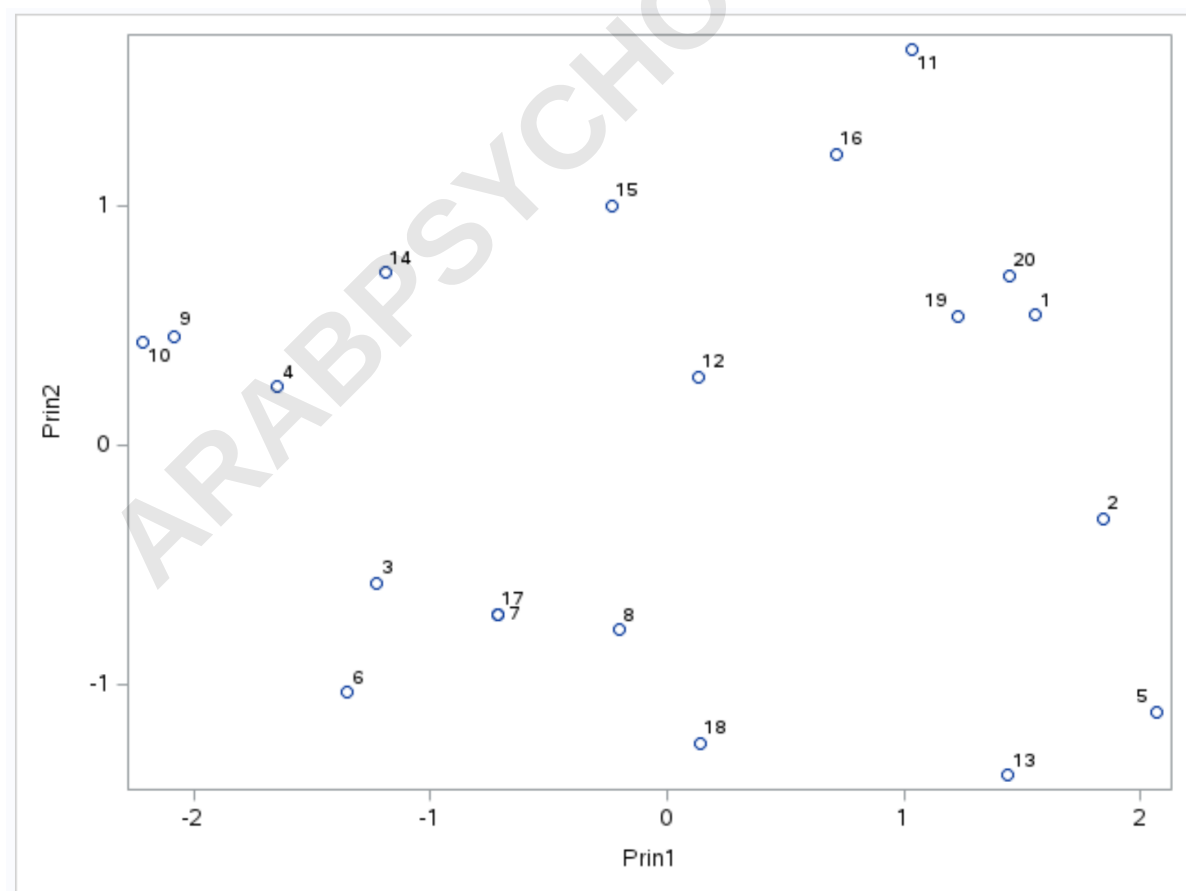
```
/*create biplot using values from first two principal components*/
```

```
proc sgplot data=biplot_data;
```

```
scatter x=Prin1 y=Prin2 / datalabel=obs;
```

```
run;
```

The resulting biplot visualization provides an immediate summary of the relationships between the basketball players based on their multivariate statistics. The x-axis represents the first principal component (Prin1), and the y-axis represents the second principal component (Prin2). Individual observations are plotted as labeled circles. Because the first two principal components account for 88.21% of the total variance, the spatial relationships observed in this two-dimensional plot are highly reliable representations of the relationships in the original three-dimensional space.



Observations positioned near each other on the plot share similar underlying characteristics across the variables **points**, **assists**, and **rebounds**. For instance, examining the far left side of the plot reveals that observations **#9** and **#10** are mapped in extremely close proximity. By cross-referencing these indices with our original dataset, we confirm that their statistical profiles are indeed similar:

**Observation #9:** 2 points, 5 assists, 17 rebounds

**Observation #10:** 4 points, 5 assists, 19 rebounds

The similarity in their values across the three variables, particularly their high rebound counts and low offensive output, explains their tight clustering in the negative direction of the principal components, illustrating the effective summarization of high-dimensional data achieved by PCA.

## Conclusion and Next Steps

Performing Principal Components Analysis in SAS utilizing the **PROC PRINCOMP** procedure is a robust and highly efficient method for achieving data reduction and gaining critical insight into the correlation structure of multivariate datasets. We successfully transformed the original three performance metrics (points, assists, rebounds) into two composite principal components that collectively account for nearly 90% of the total variability. This reduction simplifies subsequent modeling efforts while preserving the majority of the data's informational content.

The detailed output provided by **PROC PRINCOMP**--including descriptive statistics, the correlation matrix, and, most importantly, the eigenvalues and eigenvectors--allows for meticulous interpretation of the results. By examining the eigenvectors (component loadings), an analyst can determine the exact contribution of each original variable to the formation of the new principal components, thereby identifying the latent factors driving the overall variation in player performance.

Furthermore, the construction of the biplot provides an essential visual bridge between the complex statistical output and intuitive understanding. It confirms that observations with similar characteristics are grouped together in the reduced space, validating the effectiveness of the PCA transformation. Mastery of **PROC PRINCOMP** is a cornerstone for advanced data preparation and exploration in the SAS environment.

For those seeking to expand their analytical capabilities in SAS, exploring other multivariate techniques or specific modeling procedures is highly recommended. These further tutorials explain how to perform other common tasks in SAS: