# **Principal Component Analysis**

**Justification for selecting principal components:**

**For Scree Plot**

In general, the number of components chosen depends on the variance threshold or elbow point. Between 90-95% variance retention is common.

<https://dataheadhunters.com/academy/dissecting-eigenvectors-their-role-in-dimensionality-reduction/>

The point at which the Y axis of eigenvalues or total variance explained creates an "elbow" will generally indicate how many PCA components that we want to include.

<https://www.ibm.com/topics/principal-component-analysis>

The scree plot criterion looks for the “elbow” in the curve and selects all components just before the line flattens out. (In the PCA literature, the plot is called a ‘Scree’ Plot because it often looks like a ‘scree’ slope, where rocks have fallen down and accumulated on the side of a mountain.)

<https://sanchitamangale12.medium.com/scree-plot-733ed72c8608>

The “PCA 101” best practice is to plot the singular values in the “scree” plot and look for an “elbow” where the significance of components stops dropping off rapidly, and which captures a large fraction of the variance.

<https://druce.ai/2022/05/PCA>

A widely applied approach is to decide on the number of principal components by examining a scree plot. By eyeballing the scree plot, and looking for a point at which the proportion of variance explained by each subsequent principal component drops off.

<https://www.geo.fu-berlin.de/en/v/soga-py/Advanced-statistics/Multivariate-Approaches/Principal-Component-Analysis/PCA-the-basics/Choose-Principal-Components/index.html#:~:text=Another%20simple%20approach%20to%20decide,%3D%22_blank%22%7D>

**For Setting Threshold**

Another simple approach to decide on the number of principal components is to set a threshold, say 80%, and stop when the first k components account for a percentage of total variation greater than this threshold.

<https://www.geo.fu-berlin.de/en/v/soga-py/Advanced-statistics/Multivariate-Approaches/Principal-Component-Analysis/PCA-the-basics/Choose-Principal-Components/index.html#:~:text=Another%20simple%20approach%20to%20decide,%3D%22_blank%22%7D>

The common way of selecting the Principal Components to be used is to set a threshold of explained variance, such as 80%, and then select the number of components that generate a cumulative sum of explained variance as close as possible of that threshold.

<https://towardsdatascience.com/pca-102-should-you-use-pca-how-many-components-to-use-how-to-interpret-them-da0c8e3b11f0>

A good strategy is to choose the number of dimensions for which the cumulative explained variance exceeds a threshold, e.g., 0.95 (95%).

<https://www.baeldung.com/cs/pca>

You can choose the number of principal components that maximizes or optimizes your desired metric.

<https://www.linkedin.com/advice/3/how-can-you-determine-optimal-number-principal-4opaf>

When applying PCA, a threshold (x% of the total variance in the data) is chosen by the user, which defines the amount of information preserved and determines the number of retained principal components.

<https://doras.dcu.ie/17940/1/ISB2013_draft_final.pdf>

**For Kaiser’s Rule**

The Kaiser’s rule (Kaiser-Guttman criterion) is a widely used method to evaluate the maximum number of linear combinations to extract from the data set. According to that rule only those principal components are retained, whose variances exceed 1.

<https://www.geo.fu-berlin.de/en/v/soga-py/Advanced-statistics/Multivariate-Approaches/Principal-Component-Analysis/PCA-the-basics/Choose-Principal-Components/index.html#:~:text=The%20Kaiser%27s%20rule%20>

The Kaiser-Guttman rule states that components based on eigenvalues greater than 1 should be retained. This is based on the notion that, since the sum of the eigenvalues is p, an eigenvalue larger than 1 represents an ``above average’’ component.

<https://www.statpower.net/Content/312/R%20Stuff/PCA.html>

Heuristic procedures included: retaining components with eigenvalues (@ls) > 1 (i.e., Kaiser-Guttman criterion);

<https://www.jstor.org/stable/1939574>

In PCA the Kaiser criterion drops the components, for which the eigenvalues are less than 1 (when the data is standardized). Greater than 1 eigenvalue suggests that the corresponding component explains more variance than a single variable, given that a variable accounts for a unit of variance (Beavers, 2013).

<https://www.diva-portal.org/smash/get/diva2:896127/FULLTEXT01.pdf>

Kaiser's criterion suggests that you should only keep the factors that have eigenvalues greater than one, and discard the rest.

<https://www.linkedin.com/advice/0/what-advantages-disadvantages-using-kaisers>

**For interpretation**

But, why do we multiply P times X? The rows of P are called principal components. Each column of Y contains elements that are a dot product between the jth row of P and the ith column of X.Hence, Y contains projections of the original features onto the space spanned by our principal components, which are unit vectors. In other words, multiplying P times X means projecting X onto the space spanned by the rows of P.

<https://medium.com/analytics-vidhya/pca-a-linear-transformation-f8aacd4eb007#:~:text=PCA%20as%20a%20Linear%20Transformation,-Let%27s%20start%20from&text=With%20PCA%20we%20apply%20an,by%20P%20as%20shown%20below>

To interpret each component, we must compute the correlations between the original data and each principal component.

<https://online.stat.psu.edu/stat505/lesson/11/11.4>

PCA does not discard any samples or characteristics (variables). Instead, it reduces the overwhelming number of dimensions by constructing principal components (PCs). PCs describe variation and account for the varied influences of the original characteristics.

<https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/>

Reducing the number of dimensions can increase the dataset’s manageability and computational efficiency. By identifying the principal components that explain the most variation in the data, PCA reduces redundant information by creating a set of entirely uncorrelated components.

Creating a lower-dimensional representation of a high-dimensional dataset can help analysts visualize and understand underlying relationships in the data. Towards this end, analysts frequently use the 1st and 2nd principal components as the X and Y axes to graph the data in two dimensions and identify clusters.

<https://statisticsbyjim.com/basics/principal-component-analysis/>

A PCA plot is a scatter plot created by using the first two principal components as axes. The first principal component (PC1) is the x-axis, and the second principal component (PC2) is the y-axis. The scatter plot shows the relationships between observations (data points) and the new variables (the principal components). The position of each point shows the values of PC1 and PC2 for that observation.

The direction and length of the plot arrows indicate the loadings of the variables, that is, how each variable contributes to the principal components. If a variable has a high loading for a particular component, it is strongly correlated with that component. This can highlight which variables have a significant impact on data variations.

The number of principal components that remain after applying PCA can help you interpret the data output. The first principal component explains the most data variance, and each later component accounts for less variance. Thus, the number of components can indicate the amount of information retained from the original dataset. Fewer components after applying PCA could mean that you didn’t capture much data variation. More components indicate more data variation, but the results may be harder to interpret.

<https://www.ibm.com/topics/principal-component-analysis>