Paper 5110-2020

Principal Component Analysis Demystified

Caroline Walker, Warren Rogers LLC

ABSTRACT

Have you used or thought of using Principal Component Analysis (PCA) as a feature extraction method in your machine learning pipelines, but wished for a better understanding of what a principal component is and how it's obtained? We take a deep dive into a small dimensional data set, present a visual explanation of the role played by eigenvalues and eigenvectors when PCA is applied, and illustrate how what you start with leads to what you end with, what the advantages are, and what could get lost along the way.

INTRODUCTION

Principal Component Analysis (PCA), a dimensionality reduction technique, has become a widely used feature extraction method in machine learning pipelines. PCA provides a means of transforming an existing feature set into a set of new, linearly uncorrelated features. These new features are obtained via linear transformation of the original features, and are referred to as principal components. Standard PCA output includes a metric for assessing which principal components can be removed to reduce the dimensionality of the data set while maximizing the amount of information retained. Often this reduction is done automatically, as part of the procedure output.

VISUALIZING PCA

In practice, PCA is typically applied to data sets with many features and yields greatest benefit when there is redundancy in the form of linear correlation between some or all of those features. Such high dimensional data sets do not lend themselves well to visualization. With the aim of developing an intuitive understanding of the mechanisms of PCA we will work with an example data set consisting of just two features, F1 and F2, with high linear correlation[1].

Figure 1 shows our small example data set plotted with respect to the two features F1 and F2. As the plot illustrates, these features have a positive linear correlation. It will be useful for our later discussions to note the variance of these original features. Feature F1 has variance 1.468 and feature F2 has variance 0.716, making the total variance present in this feature set 2.184.

Since this is a two-dimensional data set, our aim in performing PCA will be to reduce the data set to just one dimension. We seek a single new feature which captures as much of the total variation from the original feature set as possible- we can think of this as maximizing the amount of information that will be retained by the new feature.

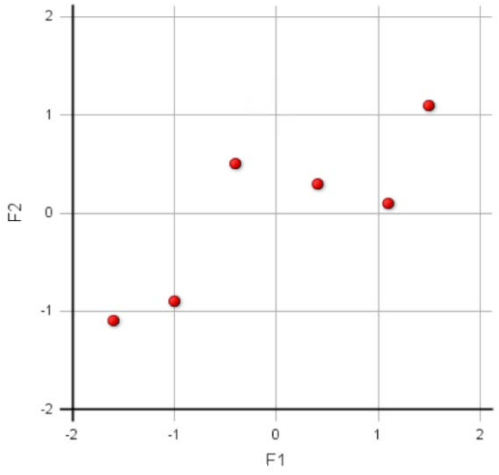

Figure 1.

[1] Note the values of both features have been mean centered.

Graphically, we can visualize any potential new feature as a line in the F1-F2 space. In Figure 2A, line A illustrates one such candidate feature, let's call this feature A. Projecting the data points from the two-dimensional F1-F2 space onto this line, as shown in Figures 2B and 2C, illustrates the values feature A will yield for each of the existing points in the data set.

For any potential new feature, if the projected points are closely clustered along the new feature line, then the new feature has not captured much of the information from the original data set and the transformation provided by the new feature will not differentiate well between data points. That is the case with feature A.

Let's see if we can find a feature that performs better than A. Figures 3A-3C show another potential new feature, feature B. The line representing feature B is a much better trend line for the F1-F2 data points, and consequently the points projected onto this line are more dispersed than we saw with feature A.

Figure 4 compares the transformed data values provided by features A and B. The values of feature A have variance 0.328. Recall our original feature set had total variance 2.184, so feature A has captured roughly 15% of the variability from our original feature set. This is

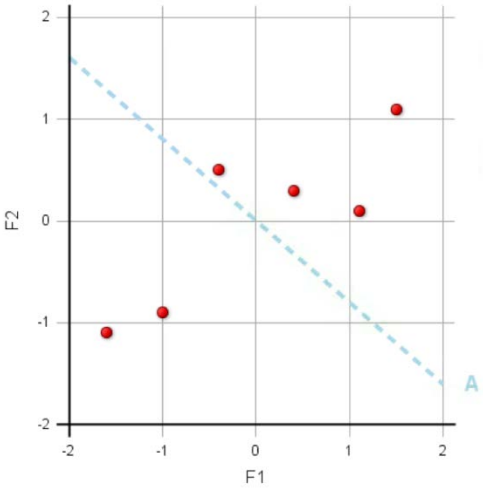

Figure 2A.

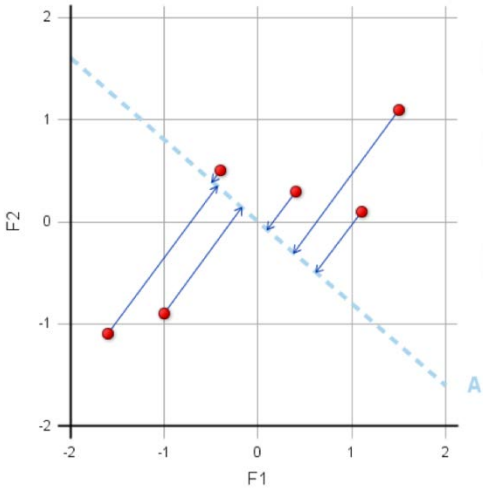

Figure 2B.

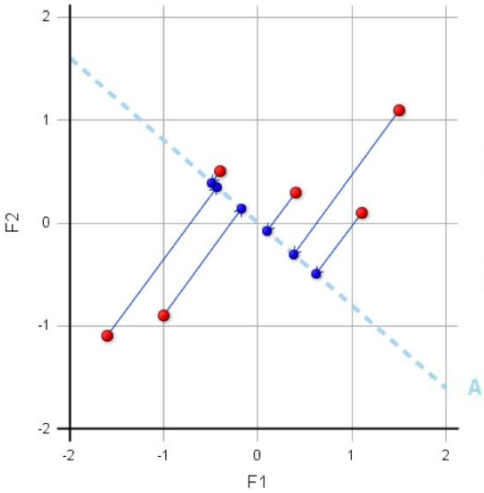

Figure 2C.

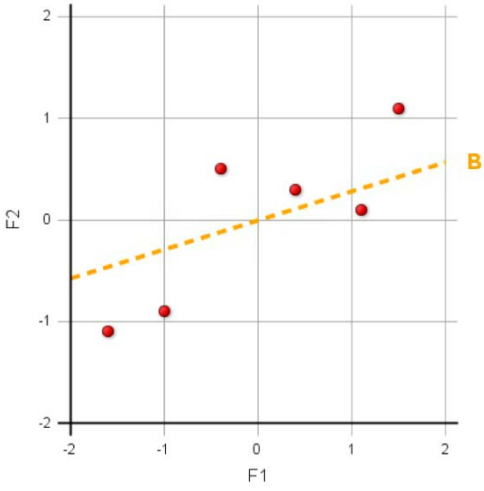

Figure 3A.

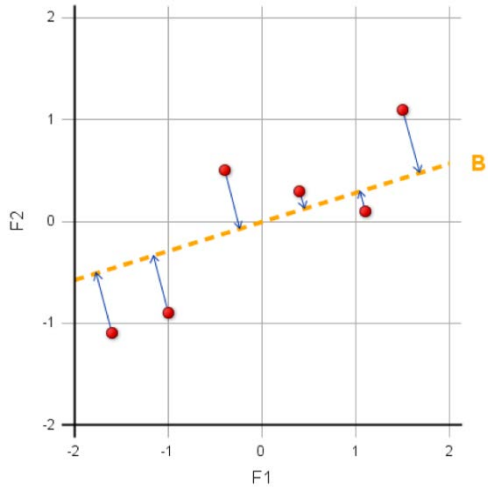

Figure 3B.

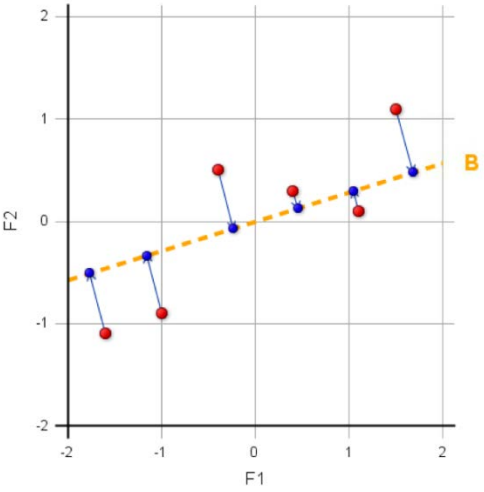

Figure 3C.

not particularly good. We could capture more variance in a single dimension by simply dropping either of our original features, rather than replacing both with feature A.

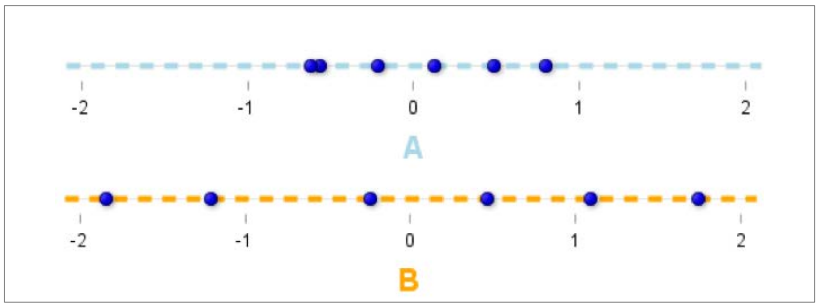

Figure 4.

In contrast, the values of feature B have variance 1.869. Feature B has captured roughly 86% of the variability that was present in our original data set. As a single feature, feature B provides a better means of discriminating between data points than either feature F1 or F2 alone. But is feature B the best feature possible, or can we do better? Fortunately we do not need to rely on trial and error to find the answer.

EIGENVECTORS, EIGENVALUES, AND THE SEARCH FOR THE BEST NEW FEATURE

PCA identifies the optimal solution[2] via the steps below:

1. Calculate the covariance[3] matrix of the original feature set.
2. Find the eigenvectors and eigenvalues of this covariance matrix.
3. Order the eigenvectors according the magnitude of their associated eigenvalues. For notational convenience, call the largest eigenvalue λ_1, the next largest eigenvalue λ_2, etc. The eigenvectors associated with these eigenvalues will be called v_1, v_2, etc.
4. The eigenvector v_1 will show the direction of maximum variance within the data set.

Why does this work? The covariance matrix provides a summary of how the values of the features relate across different observations in the data set. This matrix captures important (though not complete) information about the shape of the data cloud in the feature space. Specifically, the first eigenvector of this matrix identifies the direction of greatest 'stretch' within the feature set, and the eigenvalue describes the magnitude of that stretch. When the data points are projected onto the span of this vector, the variance of the projected points will be maximized.

Returning to our example data set from Figure 1, we can implement this method to find the best new feature for our data- the first principal component.

The covariance matrix for features F1 and F2 is:

$$\begin{bmatrix} 1.468 & 0.868 \\ 0.868 & 0.716 \end{bmatrix} \quad \textbf{(5.1)}$$

The eigenvectors of this matrix, though not always explicitly stated in the output of PCA algorithms, can be obtained via standard linear algebra software. They are also included in the output when running the PRINCOMP procedure in SAS® 9.4. Expressed as unit vectors, they are shown with their corresponding eigenvalues in 5.2 below.

[2] Mathematical proofs can be found in [1], [2], and [3].

[3] In some circumstances use of the correlation matrix may be preferable, see [1].

$$\begin{bmatrix} 0.836 \\ 0.549 \end{bmatrix} \lambda_1 = 2.038$$

(5.2)

$$\begin{bmatrix} -0.549 \\ 0.836 \end{bmatrix} \lambda_2 = 0.146$$

Figure 6 shows these eigenvectors plotted in the F1-F2 feature space. Vector v1 indicates the direction of maximum variance for our data set. Projecting the data points onto the line defined by this vector will give us the new feature we are looking for, the feature which maximizes the variance of the transformed data points. This feature is the first principal component.

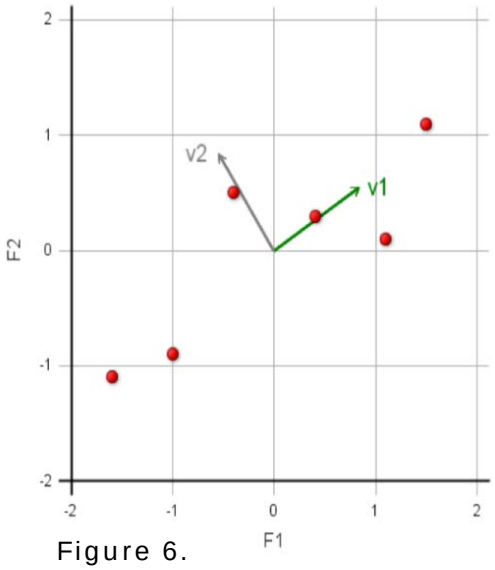

Figure 6.

In Figure 7 we can see the line defined by this first principal component (PC1) and the projection of the data onto that line. As expected, these values are widely dispersed. They have variance 2.038. This single feature has captured roughly 93% of the variance from the original two-dimensional data set.

Notice that the variance captured by this first principal component (2.038) is the same as the value of the first eigenvalue (shown in 5.2). This is not a coincidence. For each eigenvector of the covariance matrix, the eigenvalue indicates the variance of the data along that dimension. In other words, the eigenvalue tells us how much additional variance will be captured by the new feature set if the principal component described by that vector is retained. This fact becomes quite useful when performing PCA on data sets in higher dimensions.

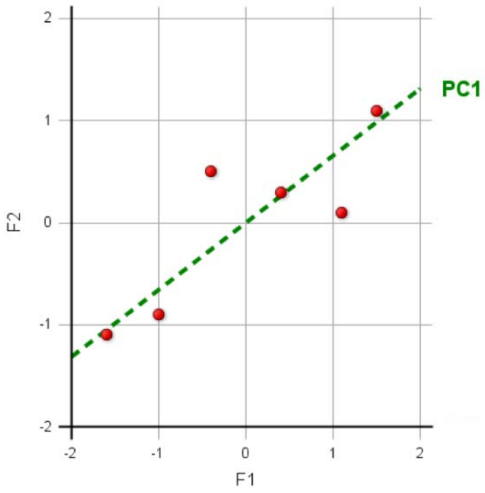

Figure 7A.

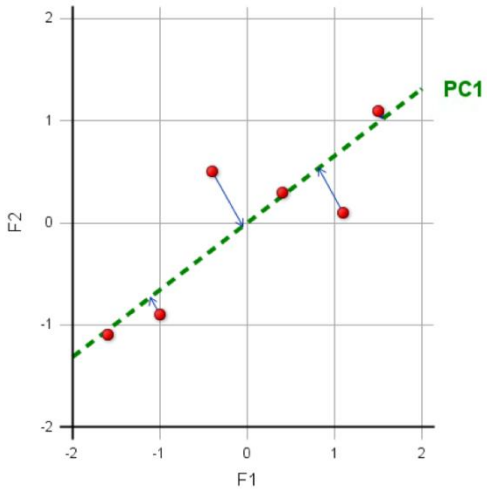

Figure 7B.

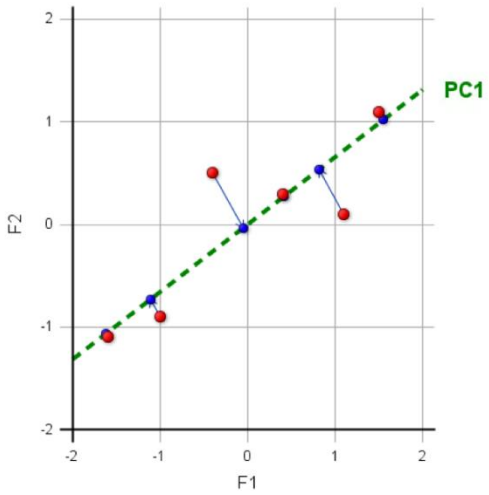

Figure 7C.

PCA IN HIGHER DIMENSIONS

The process we have illustrated here in two dimensions can easily be extended to higher dimensions. In those cases, we begin with a set of *n* features and our goal is to replace them with a set of *m* new, linearly uncorrelated, features, where $n > m$. The new features selected will be the principal components described by the first *m* eigenvectors of the covariance matrix. When choosing a value for *m*, it can be helpful to look at the eigenvalues of the correlation matrix to determine how much variance each principal component would contribute to the new feature set.

The plot in Figure 8 shows the values of the eigenvalues ordered from largest to smallest for a ten-dimensional feature set. Sharp elbows in plots such as these indicate points where the inclusion of additional new features (principal components) yield sharply diminished returns.

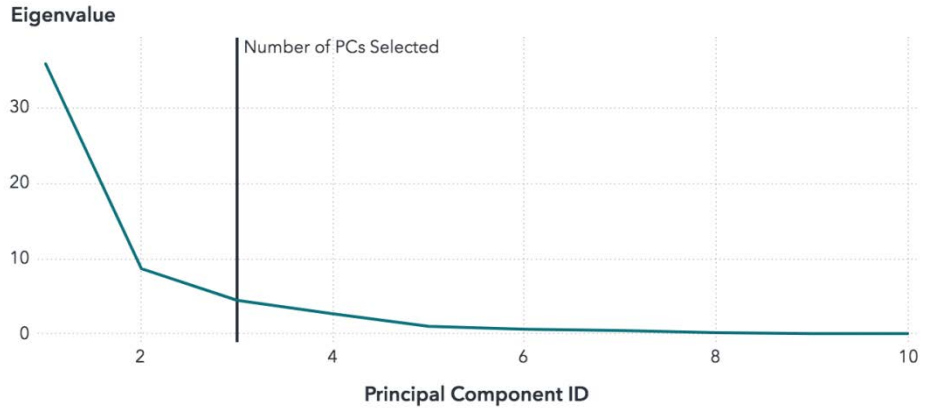

Figure 8.

A similarly useful plot, shown in Figure 9 organizes the same information in a slightly different way. Here the vertical axis shows the fraction of the original variance that will be captured by the new feature set, based on the number of principal components retained.

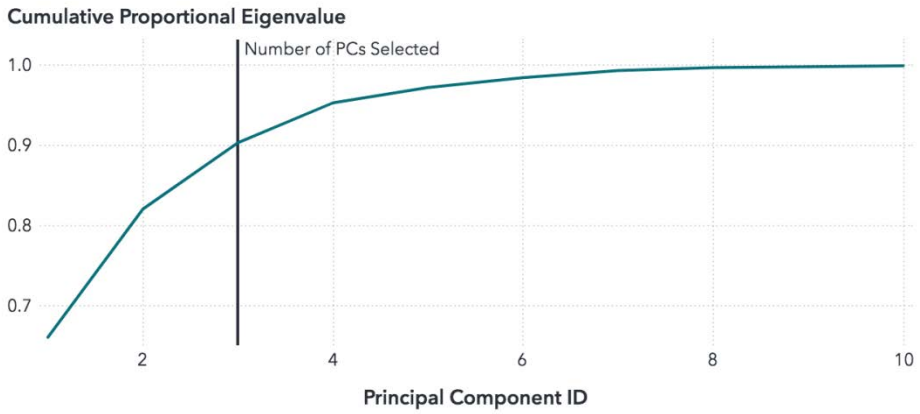

Figure 9.

From these two plots we can tell that the first two principal components capture considerably more variation than any of the subsequent principal components, and that retaining three principal components in the reduced feature set will capture around 90% of the data's original variance.

Most implementations of PCA provide users with flexible options for controlling the number of principal components retained. At the start, the user may specify a desired number of components, a desired percent variance to achieve, or allow the algorithm to identify the optimal number based on various other stopping criteria.

RECONSTRUCTION ERROR

Before concluding our discussion of PCA it is worth taking a moment to examine not just what was accomplished via PCA but also what was potentially lost. We have reduced the dimensionality of our data set by replacing our original features with a smaller set of new features. Although we sought features which maximized the total variance retained, some variance was eliminated, and also, possibly, some information.

For the two-dimensional example we considered, we can visualize the potential information lost by returning to the plot from Figure7C, reproduced here for convenience (Figure 10). It is clear that two points in particular are not as well represented by the new feature PC1, these are the two red points that lie farthest from the line PC1. The blue lines connecting these points to PC1 are a measure of the error associated with the one-dimensional representation of these points in PC1.

In this context, we are measuring error as the perpendicular distance between the points and the line, as compared to a regression context where error is measured in terms of vertical distance. The error we are interested in here is the information lost when the values of F1 and F2 are reconstructed from PC1. The reconstructed values will lie on the line PC1, while the actual values of F1 and F2 are (in most cases) some distance from the line. We need to capture the error in both the F1 and F2 directions and so the perpendicular distance is appropriate. The term *reconstruction error* is used to refer to the mean square of these error distances over all points in the data set. It is a measure of how much information may potentially be lost when the dimensionality reduction is applied.

How concerned should we be about the information loss represented by this reconstruction error? The hope of PCA is that the variance exhibited in these error directions can safely be considered noise, not information, and that removing it from the data set via dimensionality reduction will not hurt (and may in fact improve) the quality of the model results.

It should also be reassuring to know that, through a lovely mathematical symmetry, selecting the feature set which maximizes the variance retained is equivalent to selecting the set that minimizes the reconstruction error. The two goals are one in the same, and so PCA achieves them both[4].

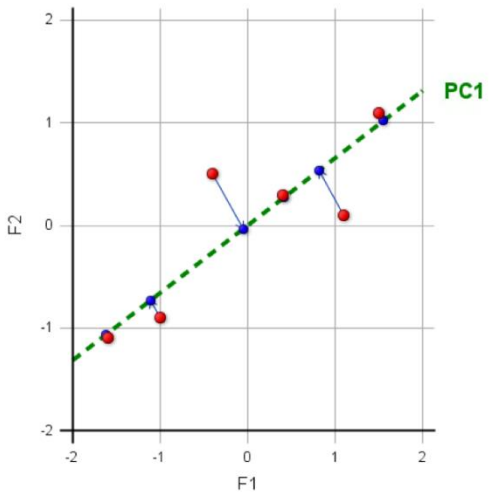

Figure 10.

[4] Proof provided in [3].

DISCUSSION & CONCLUSION

In considering the use of PCA it is important to keep in mind that this single method of dimensionality reduction is by no means a panacea. See Shlens [2] for an interesting example of a data set in which PCA would fail. PCA achieves optimality subject to a few key constraints: the features in the new feature set are linear combinations of the original features, they are linearly uncorrelated with one another, and they are defined with the aim of capturing maximal variance from the original data. These criteria will not be ideal for every application. Still, examples abound in which dimensionality reduction via PCA achieves exciting results (for those unfamiliar with 'eigenfaces', the subject is worth a look [4]).

In this brief presentation we have touched upon some main ideas important to PCA, but we certainly have not covered all that is worth examining on the topic. Hopefully this introduction may provide a useful foundation upon which to build further understanding. For readers interested in exploring PCA more deeply Joliffe [1], Shlens [2], and Wiskott [3] provide varied, and far more in depth, perspectives on the topic.

REFERENCES

1. Jolliffe, I.T. 2002. *Principal Component Analysis*, 2nd ed. New York, NY: Springer.
2. Shlens, Jonathon. 2014 "A Tutorial on Principal Component Analysis" https://arxiv.org/pdf/1404.1100.pdf
3. Wiskott, Laurenz. 2013 "Lecture Notes on Principal Component Analysis" http://cs233.stanford.edu/ReferencedPapers/LectureNotes-PCA.pdf
4. Eigenfaces. Retrieved February 27, 2019, https://en.wikipedia.org/wiki/Eigenface

RECOMMENDED RESOURCES

- "Principal Component Analysis Explained Visually" by Victor Powell and Lewis Lehe offers an interactive visualization of PCA in both 2D and 3D. http://setosa.io/ev/principal-component-analysis/

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Caroline Walker
Warren Rogers LLC
cwalker@warrenrogers.com