Principal components analysis

Herman Kamper

2023-02

PCA minimizes the reconstruction loss

Let's first consider the case where we use only a single principle component:

$$
\begin{aligned}
J(\mathbf{w}_1) &= \sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2 \\
&= \sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \mathbf{w}_1 z_1^{(n)}||^2 \\
&= \sum_{n=1}^{N} (\mathbf{x}^{(n)} - \mathbf{w}_1 z_1^{(n)})^\top (\mathbf{x}^{(n)} - \mathbf{w}_1 z_1^{(n)}) \\
&= \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - 2 z_1^{(n)} \mathbf{w}_1^\top \mathbf{x}^{(n)} + (z_1^{(n)})^2 \mathbf{w}_1^\top \mathbf{w}_1 \right] \\
&= \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - 2 (z_1^{(n)})^2 + (z_1^{(n)})^2 \right] \\
&= \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - (z_1^{(n)})^2 \right] \\
&= c - \sum_{n=1}^{N} (z_1^{(n)})^2 = c - N \hat{\sigma}_{z_1}^2
\end{aligned}
$$

So maximizing the sample variance of z_1 is the same as minimizing the reconstruction error.

Let's now consider the case with $M > 1$ principle components:

$$\begin{aligned}
J(\mathbf{W}) &= \sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2 \\
&= \sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \mathbf{W}\mathbf{z}^{(n)}||^2 \\
&= \sum_{n=1}^{N} (\mathbf{x}^{(n)} - \mathbf{W}\mathbf{z}^{(n)})^\top (\mathbf{x}^{(n)} - \mathbf{W}\mathbf{z}^{(n)}) \\
&= \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - 2(\mathbf{z}^{(n)})^\top \mathbf{W}^\top \mathbf{x}^{(n)} + (\mathbf{z}^{(n)})^\top \mathbf{W}^\top \mathbf{W}\mathbf{z}^{(n)}\right] \\
&= \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - 2(\mathbf{z}^{(n)})^\top \mathbf{z}^{(n)} + (\mathbf{z}^{(n)})^\top \mathbf{z}^{(n)}\right] \\
&= \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - (\mathbf{z}^{(n)})^\top \mathbf{z}^{(n)}\right] \\
&= c - \sum_{n=1}^{N} (\mathbf{z}^{(n)})^\top \mathbf{z}^{(n)} = c - N \sum_{m=1}^{M} \hat{\sigma}_{z_m}^2
\end{aligned}$$

Proportion of variance explained

Let's look at one method that helps us decide how many principle components M to use.

A metric known as the *total variance* (more strictly the *total sample variance*) gives an idea of the variation in data by summing the sample variance over each dimension. Assuming the data has been mean-normalized, we can calculate the total variance of the data prior to performing PCA:

$$\sum_{d=1}^{D} \hat{\sigma}_{x_d}^2 = \sum_{d=1}^{D} \frac{1}{N} \sum_{n=1}^{N} (x_d^{(n)})^2 = \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{x}^{(n)}||^2$$

Similarly, we can calculate the total variance after projecting the data using PCA:

$$\sum_{m=1}^{M} \hat{\sigma}_{z_m}^2 = \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} (z_m^{(n)})^2 = \sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} (\mathbf{w}_m^\top \mathbf{x}^{(n)})^2 = \frac{1}{N} \sum_{n=1}^{N} ||\mathbf{W}^\top \mathbf{x}^{(n)}||^2$$

(You can also do this separately for each of the principal components to get the proportion of variance explained by that single component. Also note that the sample variance for the m^{th} component is something you would have already calculated while performing PCA: it's exactly the same as the m^{th} largest eigenvalue—see the PCA derivation.)

We can now express the total variance in the projected data as a proportion of the total variance in the data prior to projection:

$$\text{PVE} \triangleq \frac{\text{Total variance in projected data}}{\text{Total variance in original data}} = \frac{\sum_{m=1}^{M} \hat{\sigma}_{z_m}^2}{\sum_{d=1}^{D} \hat{\sigma}_{x_d}^2} = \frac{\sum_{n=1}^{N} ||\mathbf{W}^\top \mathbf{x}^{(n)}||^2}{\sum_{n=1}^{N} ||\mathbf{x}^{(n)}||^2}$$

This is referred to as the *proportion of variance explained* (PVE).

If our PCA projection captures all of the variation in the original data, then you will have a PVE of 100% (this happens when the data is lying on a linear manifold).

You can also show that

$$\frac{\text{Averaged squared error}}{\text{Total variance}} = \frac{\frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2}{\frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}^{(n)}||^2} = 1 - \text{PVE}$$

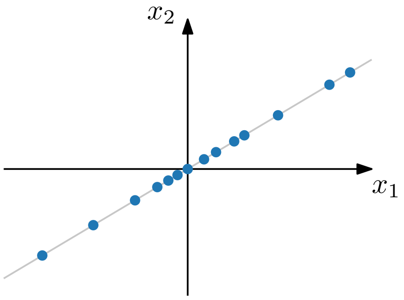
I.e., if PVE $= 1$ then we have a zero averaged squared error. So this gives us some additional insight into the PVE by viewing it in terms of the reconstruction error.

How can we use the PVE to decide on the number of principle components M? Concretely, by looking at the PVE while varying the number of components M, we might get to a point where increasing M further only gives us negligible improvements in PVE. We can choose M by looking for this type of "elbow" in a plot with PVE on the y-axis and different values of M on the x-axis.

Test yourself

Question: PVE when performing PCA on two-dimensional data

After mean normalization, we have a dataset such as the one below. Although the data is two-dimensional ($D = 2$), the data lives on a one-dimensional manifold. We apply PCA to the data. From the first definition of the PVE given above, show that the PVE will be 100% when applying PCA with $M = 1$. Treat x_1 and x_2 and z as random variables, i.e. show that the PVE will be 100% in all cases where we have two-dimensional data that lives on a linear one-dimensional manifold.



Answer:

Let's denote RV's explicitly here with capital letters. You will need $\text{var}[cX] = c^2\text{var}[X]$. So we are in a setting where we have RV's X_1 and X_2 and we assume they have been mean-normalized. The one RV is perfectly dependent on the other: $X_2 = aX_1$.

The total variance before projection:

$$\text{var}[X_1] + \text{var}[X_2] = \text{var}[X_1] + a^2\text{var}[X_1] = (1 + a^2)\text{var}[X_1]$$

The PCA projection Z recovers the dependence perfectly. You can convince yourself that this means our single projection vector will be:

$$\mathbf{w} = \frac{1}{\sqrt{1+a^2}} \begin{bmatrix} 1 \\ a \end{bmatrix}$$

(The vector is normalized to ensure that $||\mathbf{w}|| = 1$.) Our projection RV can then be calculated as follows:

$$Z = w_1 X_1 + w_2 X_2 = \frac{1}{\sqrt{1+a^2}} X_1 + \frac{a^2}{\sqrt{1+a^2}} X_1 = \sqrt{1+a^2} X_1$$

This means that the variance of the projection can be calculated as follows:

$$\text{var}[Z] = \text{var}[\sqrt{1+a^2} X_1] = (1+a^2)\text{var}[X_1]$$

I.e., $\text{var}[X_1] + \text{var}[X_2] = \text{var}[Z]$. This means that the PVE would be 100%.

Question: Reconstruction view of PVE

From the definition of PVE, prove that

$$\frac{\text{Averaged squared error}}{\text{Total variance}} = \frac{\frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2}{\frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}^{(n)}||^2} = 1 - \text{PVE}$$

Answer:

As part of the derivation where we showed that PCA minimizes the reconstruction loss, we found that

$$\sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2 = \sum_{n=1}^{N} \left[(\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - (\mathbf{z}^{(n)})^\top \mathbf{z}^{(n)} \right] = \sum_{n=1}^{N} (\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - N \sum_{m=1}^{M} \hat{\sigma}_{z_m}^2$$

This means that

$$\begin{aligned}
\frac{\frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2}{\frac{1}{N}\sum_{n=1}^{N} ||\mathbf{x}^{(n)}||^2} &= \frac{\sum_{n=1}^{N} ||\mathbf{x}^{(n)} - \hat{\mathbf{x}}^{(n)}||^2}{\sum_{n=1}^{N} (\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)}} \\
&= \frac{\sum_{n=1}^{N} (\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)} - N \sum_{m=1}^{M} \hat{\sigma}_{z_m}^2}{\sum_{n=1}^{N} (\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)}} \\
&= 1 - \frac{N \sum_{m=1}^{M} \hat{\sigma}_{z_m}^2}{\sum_{n=1}^{N} (\mathbf{x}^{(n)})^\top \mathbf{x}^{(n)}} \\
&= 1 - \frac{N \sum_{m=1}^{M} \hat{\sigma}_{z_m}^2}{N \sum_{d=1}^{D} \hat{\sigma}_{x_d}^2} = 1 - \text{PVE}
\end{aligned}$$

Acknowledgements

The discussion of proportion of variance explained is largely based on two sources: Section 10.2.3 in ISL and a lecture in Andrew Ng's Coursera machine learning course.