Mathematics ToolBox

Joseph Raso

January 2021

Contents

1 Principal Component Analysis

1.1 Theory

PCA is a "statistical interpretetion of the singular value decomposition". It is a bedrock tool for discovering the axes along which a data set varies the most. Starting from a data matrix X, with each row corresponding to one of n samples and each column corresponding to one of m dimensions[1],

$$X = \begin{bmatrix} — & x_1 & — \\ — & x_2 & — \\ & \cdots & \\ — & x_n & — \end{bmatrix} \tag{1}$$

we first find the mean sample

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

and subtract it off the data matrix:

$$B = X - \bar{X} = \begin{bmatrix} — & x_1 & — \\ — & x_2 & — \\ & \cdots & \\ — & x_n & — \end{bmatrix} - \begin{bmatrix} — & \bar{x} & — \\ — & \bar{x} & — \\ & \cdots & \\ — & \bar{x} & — \end{bmatrix} \tag{3}$$

Now B the same data matrix (i.e. same distribution), but shifted to be mean-zero.

Next, now that the data is mean-zero, it's pretty trivial to compute the covariance between the dimensions of B as the covariance matrix C.

$$C = \frac{B^T B}{n} \tag{4}$$

C is called the covariance matrix because it's elements are the variance and covariance between the various dimensions in B,[2] i.e. for element C_{ij},

$$C_{ij} = \frac{b_i \cdot b_j}{n} = \frac{1}{n}\sum_{k=1}^{n}(b_i)_k(b_j)_k = \mathbb{E}[b_i \cdot b_j] = \mathrm{Cov}(b_i, b_j) \tag{5}$$

Eigenvalue decomposition of C therefor yields:

1. A dimensional basis V in which the dimensional distributions are independent (have covariance 0); i.e. the basis in which C is diagonal.
2. The diagonalized covariance Λ, with diagonal elements $\lambda_1, \lambda_2, ...\lambda_m$ which are equal to the variance of the data along the corresponding eigen-axis.

It follows that the eigenvector corresponding to the largest eigenvalue is the independent axis along which the data has the highest variance. The matrix V of eigenvectors is referred to as the *loadings*. Finally, to find the amount of each principal component in the original samples, we find

$$T = BV. \tag{6}$$

[1] This is the transpose of the setup for singular value decomposition.

[2] Note the dimensions: B is n samples by m dimensions just like X, and consequently C is $m \times m$.

Connection to SVD If we look at the singular value decomposition of B into left eigenvectors U and right eigenvectors V,

$$B = U\Sigma V^T \tag{7}$$

it follows that

$$C = B^T B = (V\Sigma^T U^T)(U\Sigma V^T) = V(\Sigma^T \Sigma)V^T \tag{8}$$

$$\implies CV = V(\Sigma^T \Sigma). \tag{9}$$

Since the elements of the quare matrix $\Sigma^T\Sigma$ are equal to the variences along the corrosponding right eigenvectors V, $\Sigma^T\Sigma = \Lambda$, giving the right eigenvectors V the same values as in the PCA case. We can immediately then see that the loadings V, variances Λ and principal-component transformed data T is immediately available from an SVD breakdown of B:

$$V = V \tag{10}$$

$$\Lambda = \Sigma^T \Sigma \tag{11}$$

$$T = BV = U\Sigma \tag{12}$$

Dimensionality Reduction PCA can be used to sift out components dimensions in the data that are unimportant, allowing the size of the data set to be reduced to a set of axis that are most meanigful.

Deciding which components to keep is done by looking at the fraction f_k, where for the first k eigenvalues (for n total, sorted by magnitude):

$$f_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_i} \tag{13}$$

Taking a representative sample If we wanted to know what a representatince sample looked like somwhere along one of the principal components, say at a particular value p of the kth component, we would look at the component-space sample (row vector)

$$t_p; (t_p)_i = p \cdot \delta_{i,k} \quad \implies \quad t_p = [0, 0, ..., p, ..., 0, 0] \tag{14}$$

and find the corresponding real-space sample b_p by

$$t_p = b_p V \tag{15}$$

$$\implies \quad b_p = t_p V^{-1} \tag{16}$$

$$= p \cdot v_k^{-1} \tag{17}$$

where v_k^{-1} is the kth row of V^{-1}.

If we were curious about some combination of values along principal components, we could use equation 16 to to translate this back into the data space, i.e.

$$(t_p)_i = p_i \quad \implies \quad b_p = t_p V^{-1} = \sum_{i=1}^{n} p_k \cdot v_k^{-1} \tag{18}$$

In either case, to establish how this relates to the actual data, we have to add back the average sample.

$$x_p = b_p + barx \tag{19}$$

1.2 Tools

1.2.1 *(class)* PrincipalComponents.PCA(X)

Object representing the principal component analysis of input X.

Parameters:

X	*(ndarray)* The data matrix, with rows as samples, and columns as dimensions. (Is not required to be mean-zero.)

Attributes:

X	*(ndarray)* The original input data.
xbar	*(ndarray)* the mean sample (row) of X.
B	*(ndarray)* The data X, shifted to mean-zero.
C	*(ndarray)* The covariance matrix of B, normalized to the number of samples (rows of X).
s	*(ndarray)* Variance along each principal component, sorted by magnitude, normalized to the number of samples.
V	*(ndarray)* Loadings of each principal component as column vectors, sorted by magnitude of the variance (same order as v), oriented to put the first dimension of the first principal component in the positive direction.
T	*(ndarray)* The data X, rotated into the PCA basis.