ToolBox Documentation

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.

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

C is called the cavariance matrix because it's elements are the variance and covariance between the various dimensions in B[2], i.e. element c_{ij} is the covariance between the distributions along the ith and jth dimension. 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 corrosponding to the largest eigenvalue is the independent axis along which the data has the highest variance. The matrix V of eigenvectors is reffered to as the *loadings*. Finally, to find the amount of each principal component in the original samples, we find

$$T = BV. \tag{5}$$

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

$$B = U\Sigma V^T \tag{6}$$

it follows that

[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 $m \times m$.

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

$$\implies CV = V(\Sigma^T \Sigma). \tag{8}$$

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:

$$V = V \tag{9}$$

$$\Lambda = \Sigma^T \Sigma \tag{10}$$

$$T = BV = U\Sigma \tag{11}$$

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):

$$f_k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{n} \lambda_i} \tag{12}$$