**Mathematics of PCA – Added info**

To thoroughly describe PCA in an effective way, the treatment from different sources are combined here. Let us consider the case of a vector x of p number of variables. With , the variance of the linear function is maximized in PCA. The linear function,  which is uncorrelated with, can then be calculated to capture the remaining variance. Therefore the *k*-th linear function, , is calculated to have maximum variance and to be uncorrelated with *.* Consider the case where the vector of random variables *x* has a known covariance matrix *S*.  is an eigenvector of covariance matrix *S* corresponding to its *k*-th largest eigenvalue. If is chosen to have unit length (), then the variance ofis. To populate the first projection vectors  in, PCA finds maximum variance, such that

|  |  |
| --- | --- |
|  | (A.1) |

With the constraint of unit length of and maximum variance of, the method of Lagrange multipliers can be applied as

|  |  |
| --- | --- |
|  | (A.2) |

where *λ* is a Lagrange multiplier. Since differentiation gives the maximum value, equation (A.2) results in

|  |  |
| --- | --- |
|  | (A.3) |

where *Ip* is a (*p*×*p*) identity matrix. This is known as the problem of eigenstructure for the covariance matrix. To avoid a trivial null solution, (*S*- *λIp)* should be zero. *λ* and *α*1 should be an eigenvalue of *S* and the corresponding vector respectively. Therefore, the eigenvalue λ represents the variance because:

|  |  |
| --- | --- |
|  | (A.4) |

Since variance should be maximized in PCA, the eigenvalue *λ* must be as large as possible. The vector *α*1 is the eigenvector corresponding to the largest eigenvalue *λ1* of *S*. A graphical representation of the eigenvectors and eigenvalues and the assignment of PCs is shown in Figures A.1 and A.2. The second principal component maximizes the variance.

|  |  |
| --- | --- |
|  | (A.5) |

subject to the constraint, . Thus, it should be uncorrelated with . Using the method of Lagrange multipliers,

|  |  |
| --- | --- |
|  | (A.6) |

where *λ* andare Lagrange multipliers. The following relations result in. The vector *αk* is called the loadings for the *k*-th principal component (PC). The algorithms for calculation of principal components are mainly based on the factorization of matrices. Singular vector decomposition (SVD) and eigenvalue decomposition are the main techniques for factorization of matrices. For any (*I*×I) matrix A and *P* which are non zero orthonormal matrices, the eigenvalue problem can be expressed as

|  |  |
| --- | --- |
|  | (A.7) |

where  is an eigenvalue matrix and its components are . Then matrix A by eigenvalue decomposition is

|  |  |
| --- | --- |
|  | (A.8) |

Here, the property *P*T=*P*-1 was used from the fact that *P* is orthonormal. If a covariance matrix S of X is a matrix A, the data manipulation involves decomposition of the data matrix **X** into two matrices **V** and **U**, and **V** is orthonormal,

|  |  |
| --- | --- |
|  | (A.9) |

The columns of **U** are known as scores and those of **V** are called loadings. PCA is a technique to decompose eigenvalues of a covariance matrix, *S*, of a given data matrix. The loadings can be understood as the weights for each original variable when calculating the principal components. The matrix **U** contains the original data in a rotated coordinate system. The mathematical analysis involves finding these new “data” matrices **U** and **V**. The dimensions of **U** (i.e. its rank) that capture all the information of the entire data set of **X** (i.e. # of variables) is far less than that of **X** (ideally 2 or 3). One now compresses the **N** dimensional plot of the data matrix **X** into 2 or 3 dimensional plot of **U** and **V**. While the eigenvalues geometrically represent the length of each of the principal axes, i.e. scores, the eigenvectors of the covariance matrix represent the orientation of principal axes of the ellipsoid (i.e. loadings). By using just a few latent variables, the dimensionality of the original multivariate data sets are reduced and visualized by their projections in 2D or 3D with a minimal loss of information. Therefore, PCA is a process of dimensionally reduced mapping of a multivariate data set.

eigen3

**Figure A.1.** A graphical representation of the data points and their eigenvalues

eigen10 final

**Figure A.2.** Determination of two principal components (PC1 and PC2) in a new scaled coordinate, x1 and x2

References:

* Massart DL, Vandeginste BGM, Deming SN, Michotte Y, Kaufman L. Chemometrics: A Textbook. Amsterdam: Elsevier, 1988.
* Jolliffe IT. Principal Component Analysis. New York: Springer-Verlag, 2002.
* Davis JC. Statistics and Data Analysis in Geology. New York: John Wiley & Sons, 1986.
* Suh C. Informatics Aided Design of Crystal Chemistry. Engineering Science, vol. Ph.D.: Rensselaer Polytechnic Institute, 2005.