

Principle Components Analysis

A Short Primer by Chris Simpkins, simpkins@cc

High-level Ideas

A PCA projection represents a data set in terms of the orthonormal eigenvectors of the data set's covariance matrix. A covariance matrix captures the correlation between variables in a data set. PCA finds the orthonormal eigenvectors of the covariance matrix as the basis for the transformed feature space. (Eigenvectors can be thought of as the "natural basis" for a given multi-dimensional data set.) Higher eigenvalues in the covariance matrix indicate lower correlation between the features in the data set. PCA projections seek uncorrelated variables.

Every data set has principle components, but PCA works best if data are Gaussian-distributed. For high dimensional data the Central Limit theorem allows us to assume Gaussian distributions.

Covariance Matrices

The variance of a single variable x is given by:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}$$

The variance of two variables, x and y , is given by:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n}$$

Covariance tells you how two variable vary together:

- If the covariance between two variables is positive, then as one variable increases the other will increase.
- If the covariance between two variables is negative, then as one variable increases the other will decrease.
- If the covariance between two variables is zero, then the two variables are completely independent of each other.

For a set of variables $\langle X_1, \dots, X_n \rangle$, such as the features of a data set, we can construct a matrix which represents the covariance between each pair of variables X_i and X_j where i and j are indexes of the feature vector.

$$\text{cov}(X) = \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{var}(X_n) \end{bmatrix}$$

Notice that:

- along the diagonal we have simply the variance of an individual variable, and
- the matrix is symmetric, that is, $cov(X_i, X_j) = cov(X_j, X_i)$.

Using the Covariance Matrix to find Principle Components

For PCA, we subtract the means \bar{X}_i from each x_i before constructing the covariance matrix so that each \bar{X}_i has a mean of zero. Subtracting the means allows us to rewrite the covariance matrix as the following matrix multiplication:

$$\Sigma = \frac{1}{n} \mathbf{X} \mathbf{X}^T$$

Then, by the spectral decomposition theory, we can factor the matrix above into:

$$\Sigma = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix of the eigenvalues of the covariance matrix ordered from highest to lowest:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

Then, the principal components are the row vectors of \mathbf{U}^T . We call \mathbf{U}^T the projection weight matrix \mathbf{W} and the transformed data matrix \mathbf{S} can be obtained from the original data matrix \mathbf{X} by:

$$\mathbf{S} = \mathbf{W} \mathbf{X}$$

Note that if we choose not to use eigenvectors that correspond to lower eigenvalues so that \mathbf{W} has fewer rows, then each \mathbf{s} will have lower dimensionality than its corresponding \mathbf{x} . Discarding these eigenvectors can be thought of as discarding noise from the data, since these eigenvectors represent highly correlated, and thus uninformative variables.

Quantifying the Variation in a PCA Transformation

The total variation in a PCA transformation of a data set is the sum of the eigenvalues of the covariance matrix. Since these eigenvalues are contained in $\mathbf{\Lambda}$,

$$\sum_{i=1}^n \text{Var}(PC_i) = \sum_{i=1}^n \lambda_i = \text{trace}(\mathbf{\Lambda})$$

So the fraction

$$\sum_{i=1}^k \frac{\lambda_i}{\text{trace}(\mathbf{\Lambda})}$$

gives the cumulative proportion of the variance explained by the first k principle components. (Recall that the trace of a matrix is the sum of the elements on its main diagonal.)

References

- A tutorial on Principle Components Analysis, Lindsay I Smith, http://csnet.otago.ac.nz/cosc453/student_tutorials/
- A tutorial on Principle Components Analysis, Jon Shlens <http://www.dgp.toronto.edu/~aranjan/tuts/pca.pdf>
- A survey of dimension reduction techniques, Imola K. Fodor, <http://www.llnl.gov/CASC/sapphire/pubs/148494.pdf>