What is PCA, really?

Kendrick Kay

Principal components analysis (PCA)

- PCA is a widely used, powerful technique.
- It is fast to compute and based on sound principles.

- But it can be mysterious/opaque to the uninitiated.
- And unless you know what you are doing, there are some tricky details.

Principal components analysis (PCA)

- PCA is a linear dimensionality reduction technique based on eigenvalue decomposition of a matrix

- In short:
  - You start with data arranged as a 2D matrix:
    subjects (neurons, voxels, units) x features (attributes)
  - You calculate the top several principal components (PCs). PCs are just linear combinations of features. The first PC explains maximal variance in the data, the second PC is orthogonal to the first and explains a maximal amount of the remaining variance, and so on. Often, you try to look at the PCs and interpret them.
  - You look at the fall-off of the singular values (to see how many PCs you need to capture most of the variance in the data).
  - You look at the loading of your subjects on the PCs to understand the primary ways in which your subjects vary.

Details on PCA

- $X = U*S*V^T$ [this is singular value decomposition (SVD)]
- For example:
  - X is 100 subjects x 10 features. It contains your data.
  - U is 100 x 10. The *n*th column consists of loadings on the *n*th PC (one loading for each subject). Each column of U is 100 subjects x 1.
  - S is 10 x 10 with singular values along the diagonal.
  - V is 10 x 10. The *n*th column is the *n*th PC. Each column of V is 10 features x 1.
- Important properties:
  - U and V are orthonormal (each column is unit length and has a dot product of 0 with every other column) (i.e., $U^TU = I$, $V^TV = I$).
  - S has non-zero elements (s_i) only along the diagonal.
  - The order of the columns of U, S, and V matter. (The PCs come in decreasing order of variance explained.)
  - $U(:,1:n)*S(1:n,1:n)*V(:,1:n)^T$ is a rank-n "reconstruction" of the original data matrix.
  - The percent variance explained by each PC is given by s_i^2 normalized by the sum of all s_i^2, multiplied by 100.
  - There is a sign ambiguity in the PCs (if you sign flip a column of V, you can compensate by sign flipping the corresponding column of U).

Some further details

- Note that $X^TX = (V*S*U^T)*(U*S*V^T) = V*S^2*V^T$
  - Thus, SVD on either X or X^TX yields the same singular vectors in V. (However, the singular values in the latter case will be the square of the singular values in the former case. If you compute on X^TX, you should not square the eigenvalues when calculating percent variance explained.)
- Note that X*V is just rotating your data to a new coordinate system given by the columns of V.
  - In particular, notice that $X*V = (U*S*V^T)*V = U*S$

Normalization issues

- Before performing PCA on your data, you should think carefully about the mean and variance of each column and each row of your data matrix
- PCA is often performed on a “correlation matrix”, that is, each column of X is mean-subtracted and unit-length-normalized and then X^TX is computed. This is matrix of correlation coefficients, on which SVD can then be performed. This method of preparation is sensible **if** you really don’t care about the units of your features nor the mean of each feature. (The covariance matrix of the data is similar to the correlation matrix: there is just an overall difference in overall scaling...)
- However, you don’t have to perform the mean subtraction and unit-length normalization. Whether you should or not depends on your analysis goals, and whether you do or not **will** make a difference to the results and the interpretation.

What is PCA good for?

- Dimensionality reduction
  - Instead of having to interpret and/or compute on a large number of features, you can often reduce things down to just a handful of features.
- Helps focus interpretation of complex data on the “main” components
- Helps summarize a set of data
- Helps visualize your data (since it is hard to visualize more than 3 dimensions at a time)
- Helps assess how “redundant/correlated” your features are

- **Caution/limitation:** There is no guarantee that PCs are “matched” to latent structure in your data; the PCs are guaranteed to be deeply meaningful only in the case of data that are distributed as a multivariate Gaussian.

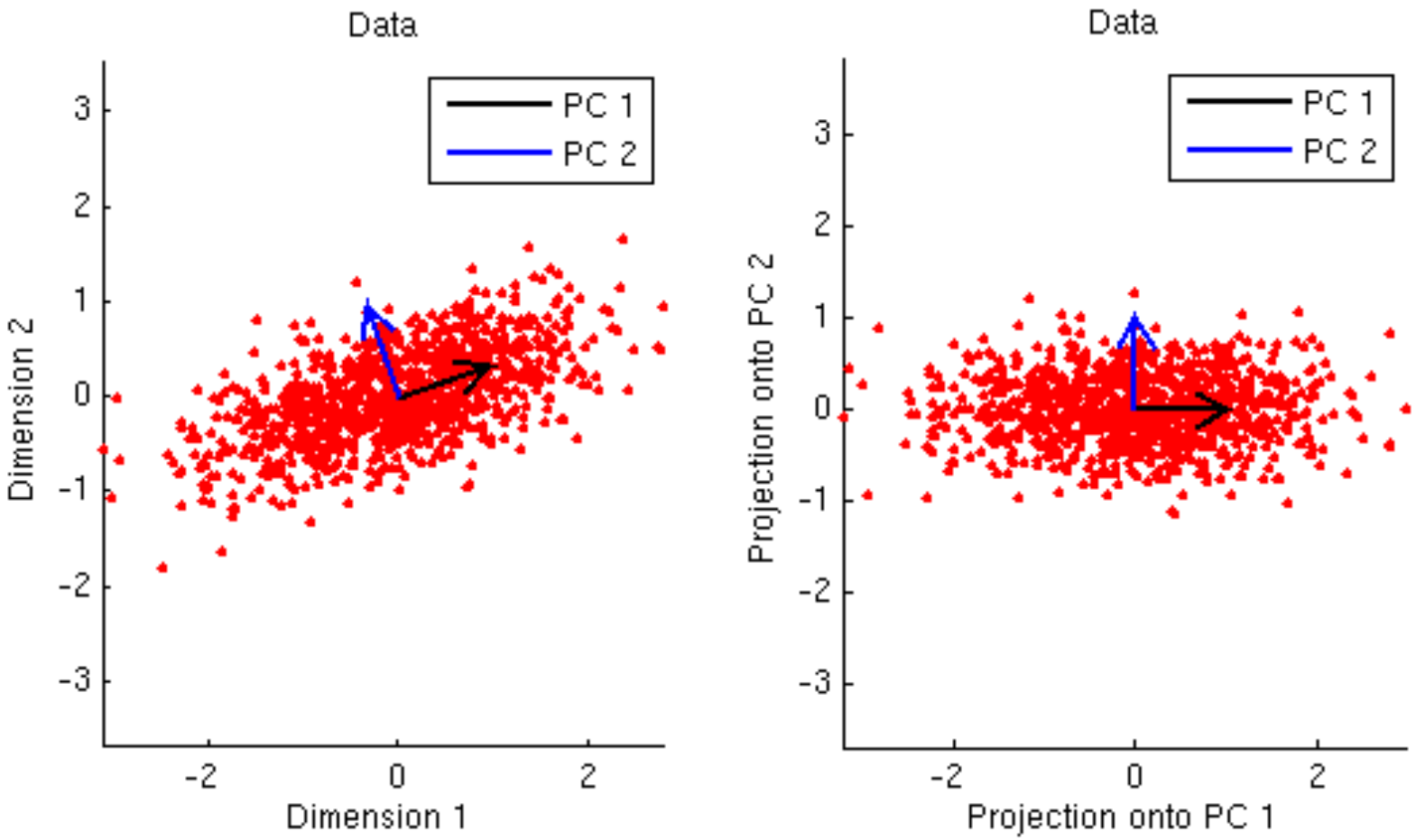
Conceptually, what is PCA doing?

- PCA is just performing a rotation of your data
- PCA finds a special orthonormal basis for your data
- PCA is iterative variance explanation
- PCA is fitting a multivariate Gaussian to your data
- PCA finds low-rank approximations of your data
- PCA is a simple autoencoder
- PCA whitens (decorrelates) your data
  - In the original data, features may be correlated, but after PCA, the loadings are now uncorrelated.

More info

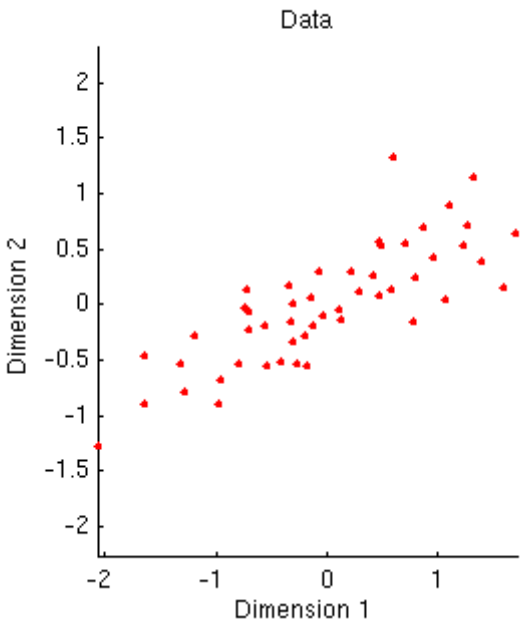
- http://randomanalyses.blogspot.com/2012/01/principal-components-analysis.html
- And many other resources on the Internet, of course...

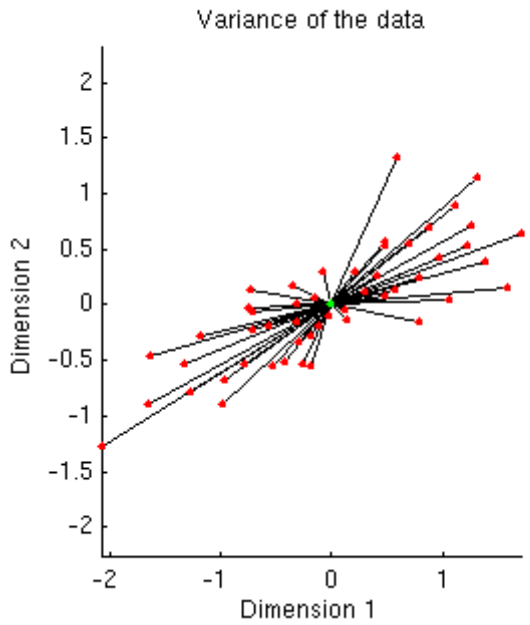
Some illustrations



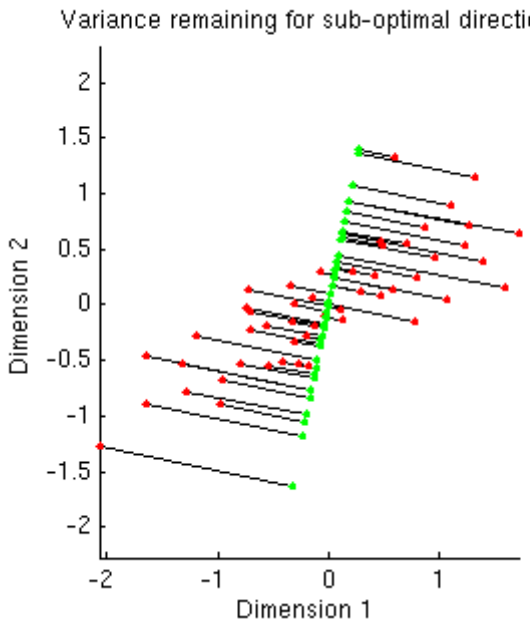
PCA is just a rotation

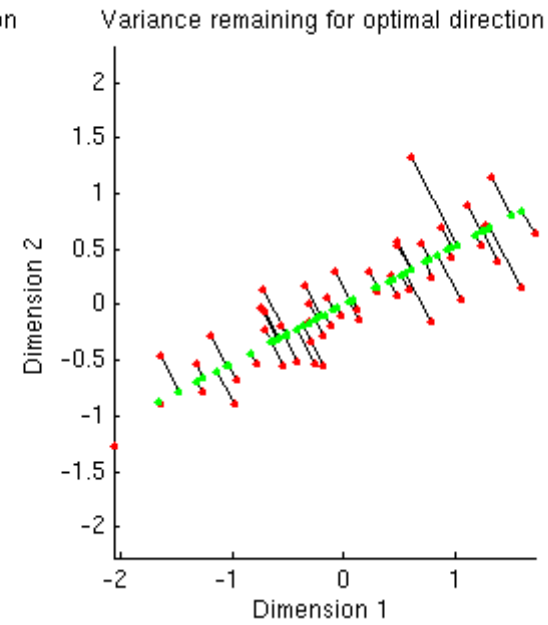
Some illustrations





Variance is sum of squared distances





Principal components point in directions of maximal variance