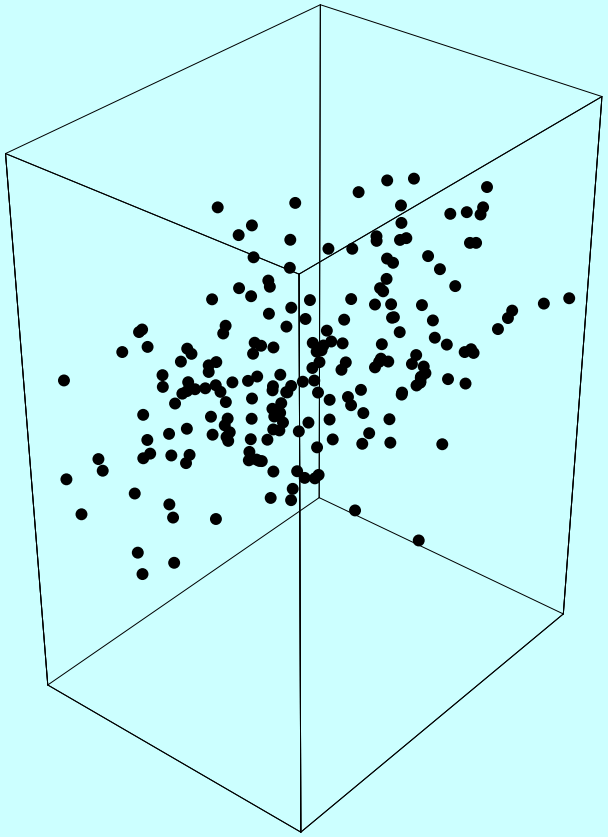
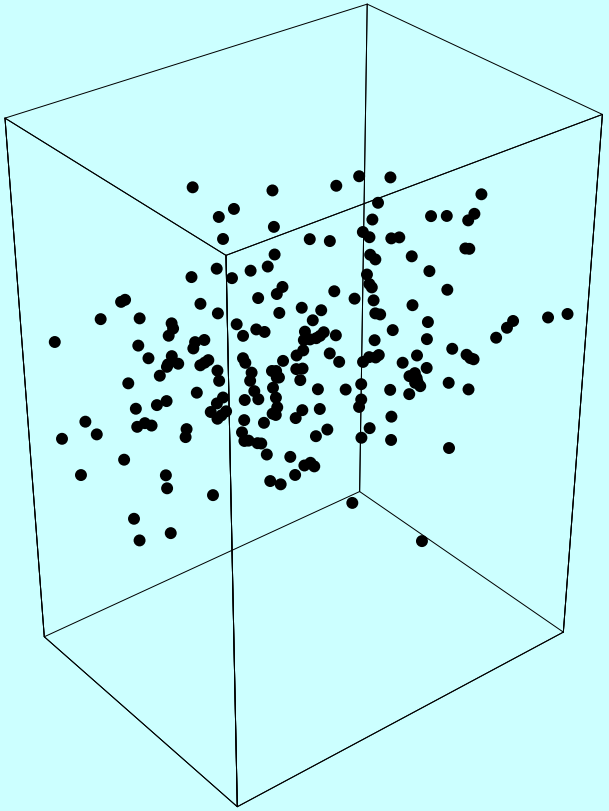
Principal (*not* principle) components and SVD

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Multivariate normal distribution





Multivariate normal distribution

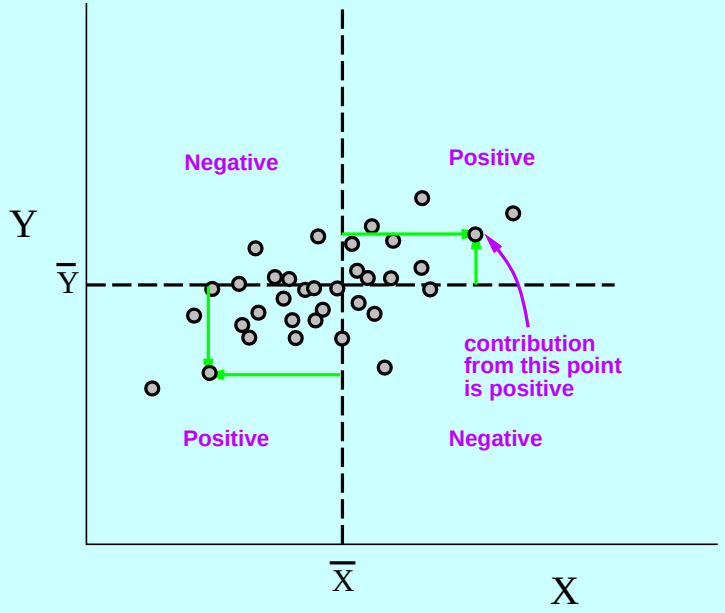
If we have a set of m variables, and n observations, and they follow a joint normal distribution, This is characterized by its means (one for each of the m variables) and an $m \times m$ covariance matrix. The covariances for two variables X and Y is the expectation of the product of their deviations from their means:

$$\mathrm{Cov}\,(X, Y) \;=\; \mathbb{E}\left[(X - \mu_X)\,(Y - \mu_Y)\right]$$

If covariances are positive, the two variables are positively correlated, if negative, negatively correlated. In fact, the correlation coefficient is $\mathrm{Cov}\,(X, Y)/(\sigma_X \sigma_Y)$, the covariance of the standardized variables.

Similarly, in a sample

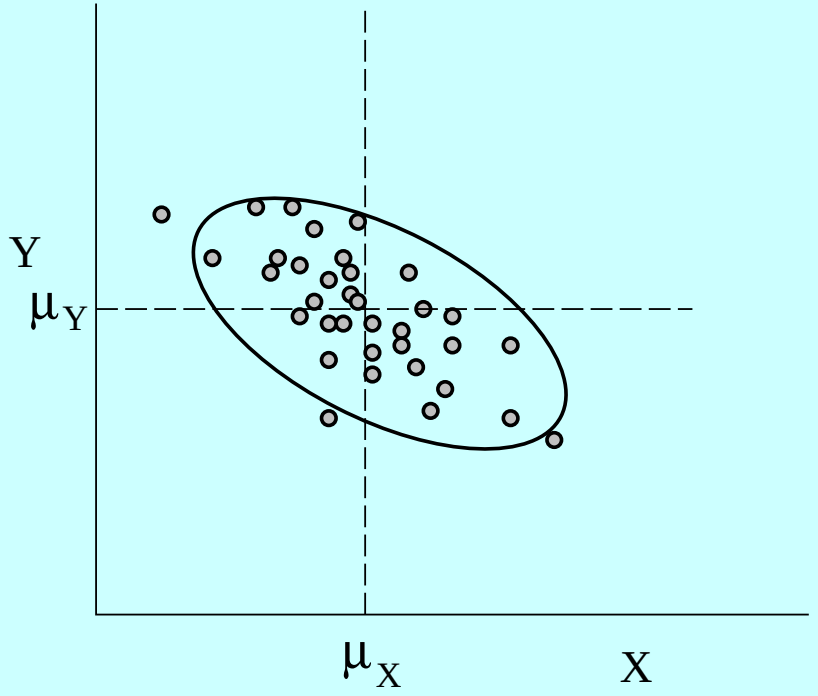
A covariance is the expectation (in a distribution) or sample average of the product of the departure of X from its mean, times the departure of Y from its mean: $(X - \bar{X})(Y - \bar{Y})$



For a positive relation, the product is mostly from the upper-right and lower-left quadrants (when the origin is placed at the means) and is positive.

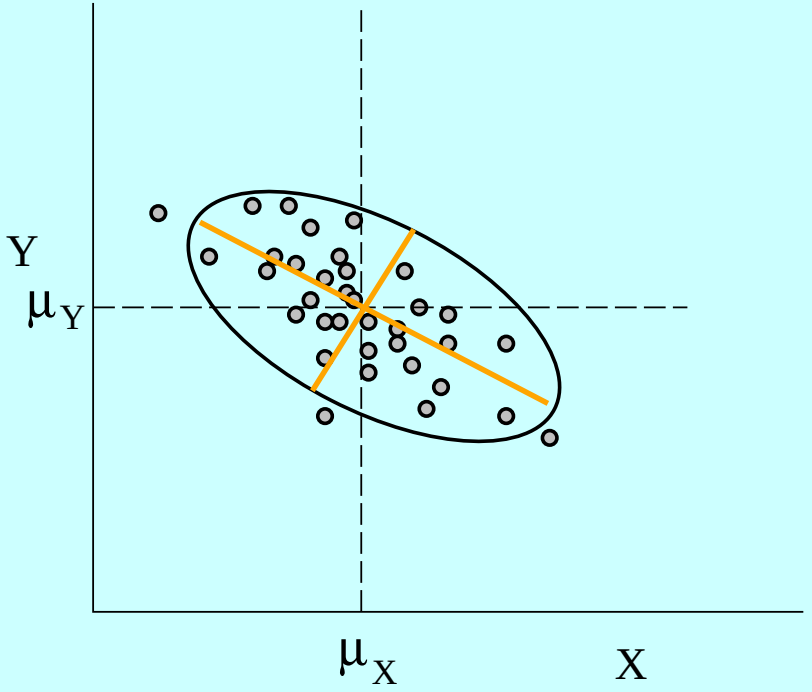
The covariance of X with itself is of course the average of $(X - \bar{X})^2$, which is the variance.

Multivariate normal distribution II

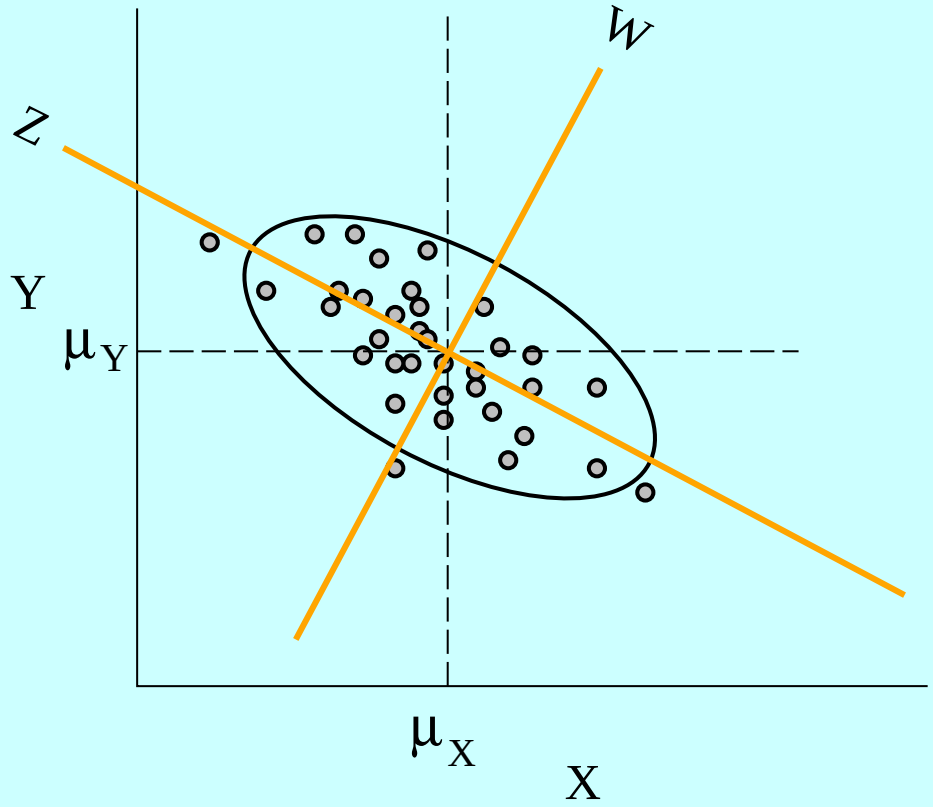


The contours of equal density of a multivariate normal are ellipsoids. In three dimensions they are watermelon-shaped (or egg-shaped). In two dimensions (which is all we can plot here) they are ellipses:

Ellipses (or ellipsoids) have axes

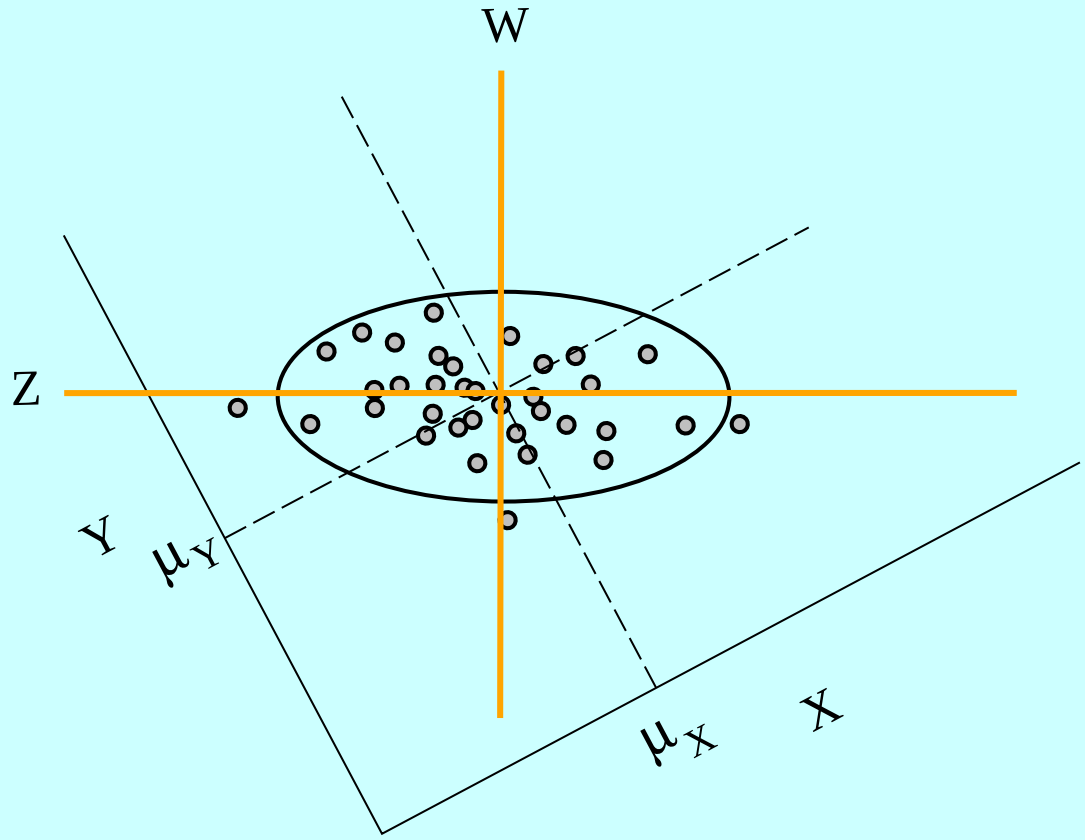


Principal components



We can find new axes in the space that make the variables uncorrelated. These have different variances. They are the ~~*principle*~~ *principal components*. This is shown in 2 dimensions but we can find all the axes. The number of them is the smaller of the number of variables (m) and one less than the number of data points ($n - 1$).

Principal components



We can find new axes in the space that make the variables uncorrelated. These have different variances. They are the *principal components*. This is shown in 2 dimensions but we can find all the axes. The number of them is the smaller of the number of variables (m) and one less than the number of data points ($n-1$).

Principle (no! “principal”) components

Princip*al* components are a way of re-describing your data. If we take all the principal components we are simply choosing a new set of axes to plot them on. It is thus mostly a descriptive technique.

Just as three points in space always lie on a plane (and two points in a line), there is a set of axes in $\min(m, n-1)$ dimensions which describe a linear subspace of the original space; the coordinates in these axes describe all the points exactly.

This is often misunderstood! For example if we measure gene expressions of 5,000 loci (by the way, pronounced “low sigh”, not “low key”, unless you’re being finicky about Latin) in 83 individuals, you don’t have 5000 dimensions, just 82. You can’t ask and answer 5000 questions that have independent answers.

Properties of ~~principle~~ principal components

It is often helpful to plot your points against the coordinates of only the first few axes. This loses some of the information.

The principal components fit the data points as closely as possible (in a least squares sense) for that number of axes. We want to choose those principal components first that have the biggest variances.

For example, the data were in a somewhat flat pancake-like cloud, they would be well-fit by two axes, that lie in the panckake. If they lay in a long cucumber-like cloud, one axis would do well.

But if the cloud is curved, linear methods like this won't work well – there are methods such as *nonmetric multidimensional scaling* available instead. All of these are methods of *ordination*.

Principal components are often called PC1, PC2, etc., in order of decreasing variances. The technique is also sometimes called Singular Value Decomposition (SVD), just to cater to mathematical snobbery.

The math of principal components

Given data matrix $\mathbf{X}$ with each row a data point, and each column a variable,

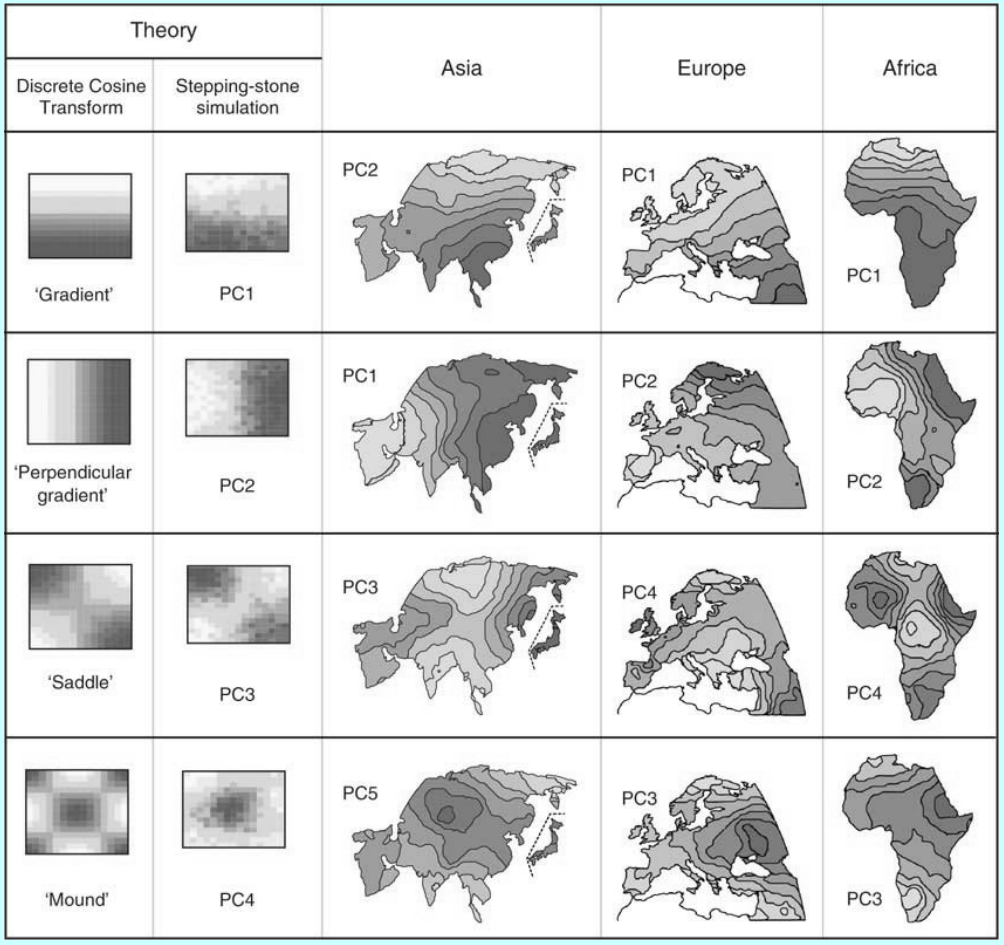
- Compute the covariance matrix of $\mathbf{X}$, call it C. The R function is `cov`.
- Take the eigenvalues and eigenvectors of $\mathbf{C}$, so that

$$\mathbf{C} = \mathbf{U}\,\Lambda\,\mathbf{U}^T$$

  The R function is either `svd` or `eigen`.
- The coordinates of the data point $\mathbf{x}$ on the PCs is given by the columns of $\mathbf{U}$, each multiplied by the square root of its eigenvalue, PC1 being the column with the largest eigenvalue.
- The variance of each principal component is its eigenvalue.

We will do this in the exercise.

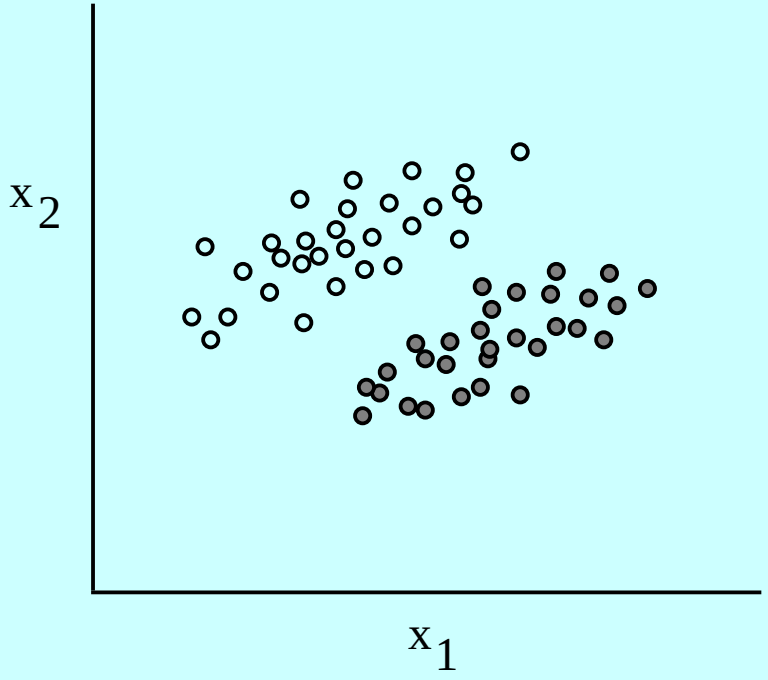
PCs may not imply unusual processes



Right three columns: PCs of continental gene frequencies, plotted on maps, computed in papers by Cavalli-Sforza, Menozzi, and Piazza, *The History and Geography of Human Genes*, 1994.

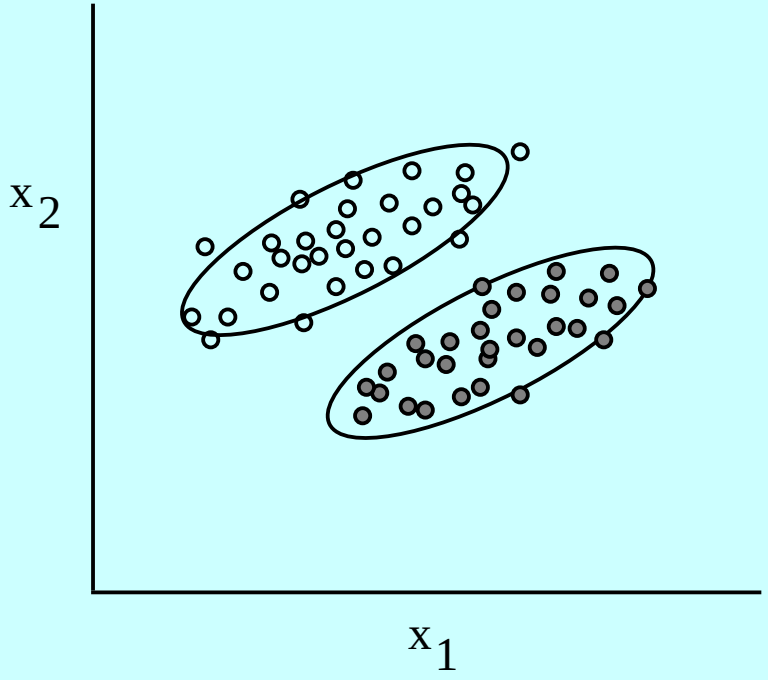
Left two columns: what are expected if gene frequencies were under only genetic drift and local migration for a long time. Computed by Novembré and Stephens *Nature Genetics* 2008.

Principal components aren't the only possible axes



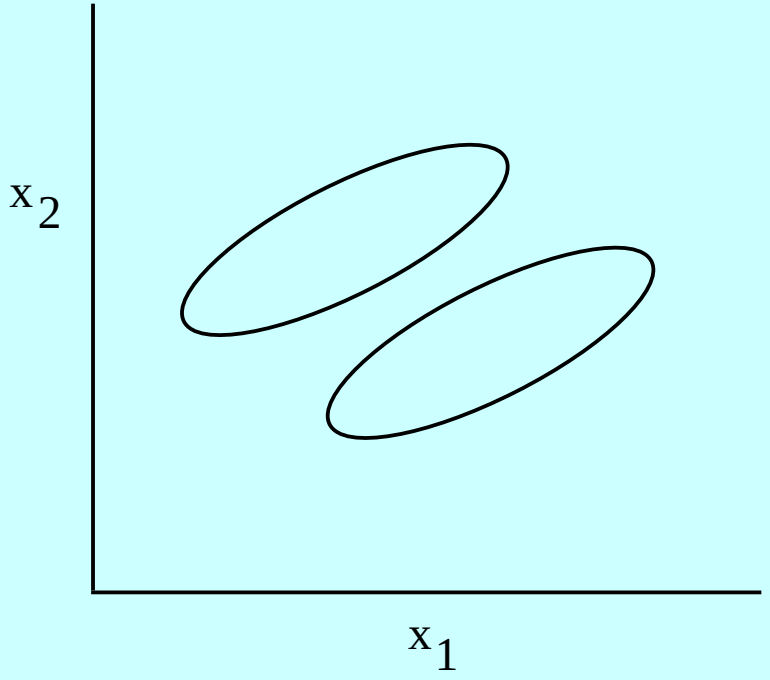
Suppose we have two populations of points, with multivariate normal distributions that are the same except that the means are different. Here are (imaginary) data that we have collected for the two groups.

Principal components aren't the only possible axes



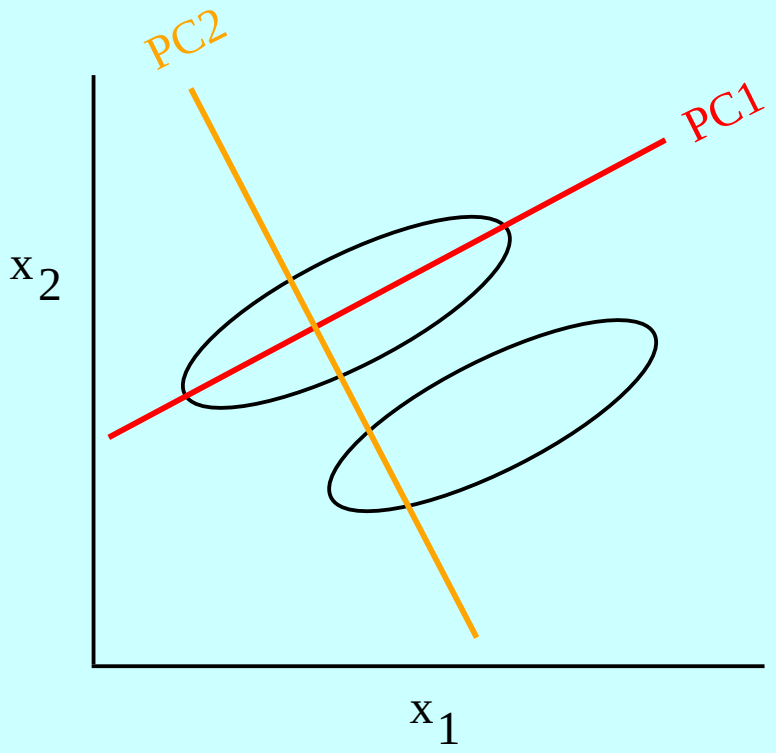
We can consider the bivariate normal distribution for each group – it has contours of its density function which are ellipses.

Principal components aren't the only possible axes



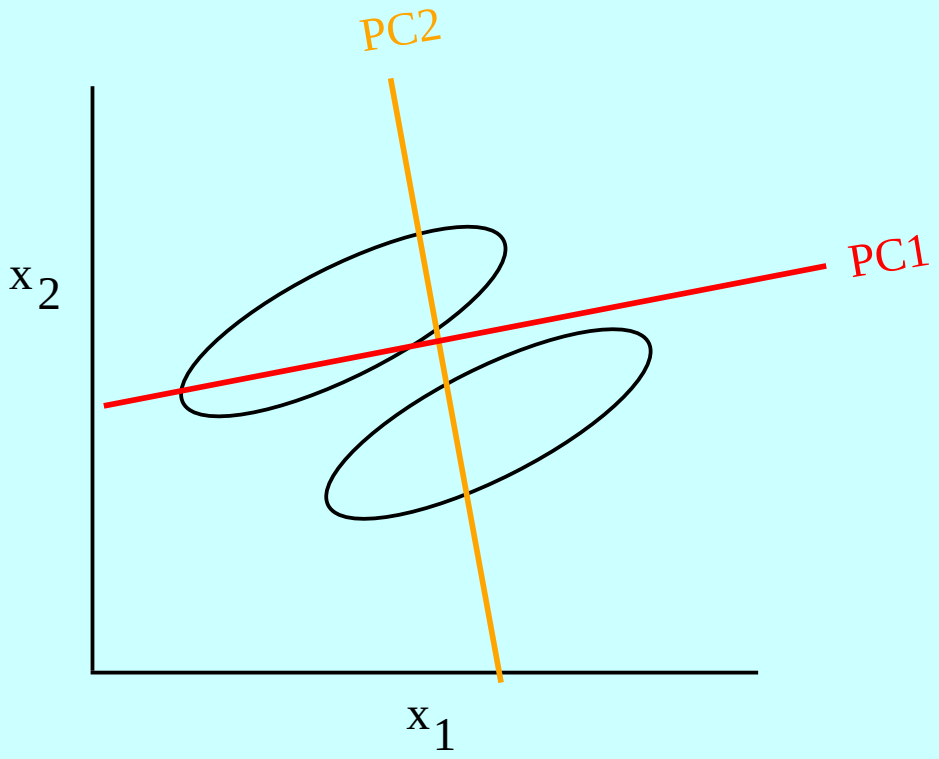
Let's represent each group by one of its density contours, to make it easy to see.

Principal components aren't the only possible axes



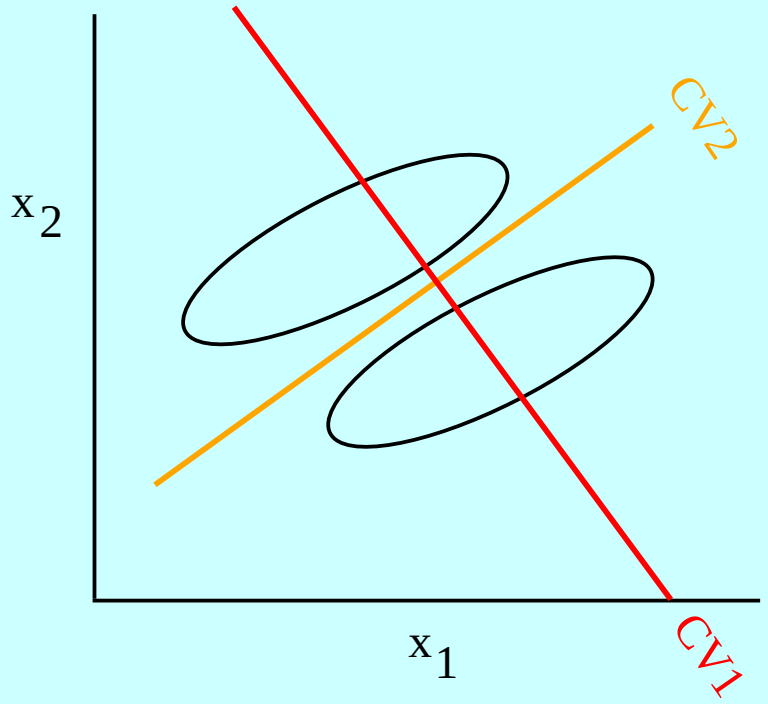
If we compute the principal components for a single group, they are the same for both groups, and give these axes.

Principal components aren't the only possible axes



If we just compute principal components from the overall distribution, it tilts a little more horizontally (in this case).

Principal components aren't the only possible axes



But if we choose axes that have the difference between groups be as large as possible measured in within-group standard deviations, we get "discriminant analysis" or "canonical variates" (a linear version of what machine learning people (mis)call "Support Vector Machines").

Principal Coordinates

Given the set of distances among points, we can come up with a placement of the points in space that has those distances (by triangulating). From those we can calculate principal components.

Any rotation or translation of the set of points will just rotate of translate the PCs along with them. So their exact location doesn't matter, in some sense. Working from the distances can get you the PCs without the step that determines the locations of the points.

Doing this from a table of distances is called "principal coordinates". It gets you the same PCs, but it lacks the interpretation of the axes in terms of the original variables (because you got rid of those).

The tip of the iceberg

So there are all sorts of intriguing methods such as

- Discriminant Analysis: Finding an axis, and a cutoff point on that axis, that does as good a job as possible of deciding in which of two groups a point lies.
- Canonical Variates: for several groups of points, finding axes that make between-group variance relative to within-group variance as big as possible.
- The bizarrely-named Support Vector Machines (no, machine learning folks, you do not “learn” a result, you *infer* it. It might be wrong, you know!) This is a nonmetric technique intended to handle curved spaces while doing something like discriminant analyses.
- Nonmetric Multidimensional Scaling. Similar intentions to SVM with a very different machinery.
- Cluster Analysis. Mostly descriptive, with very few convincing valid statistical models, but very important as an alternative to continuous scales.