PCA Additional Information

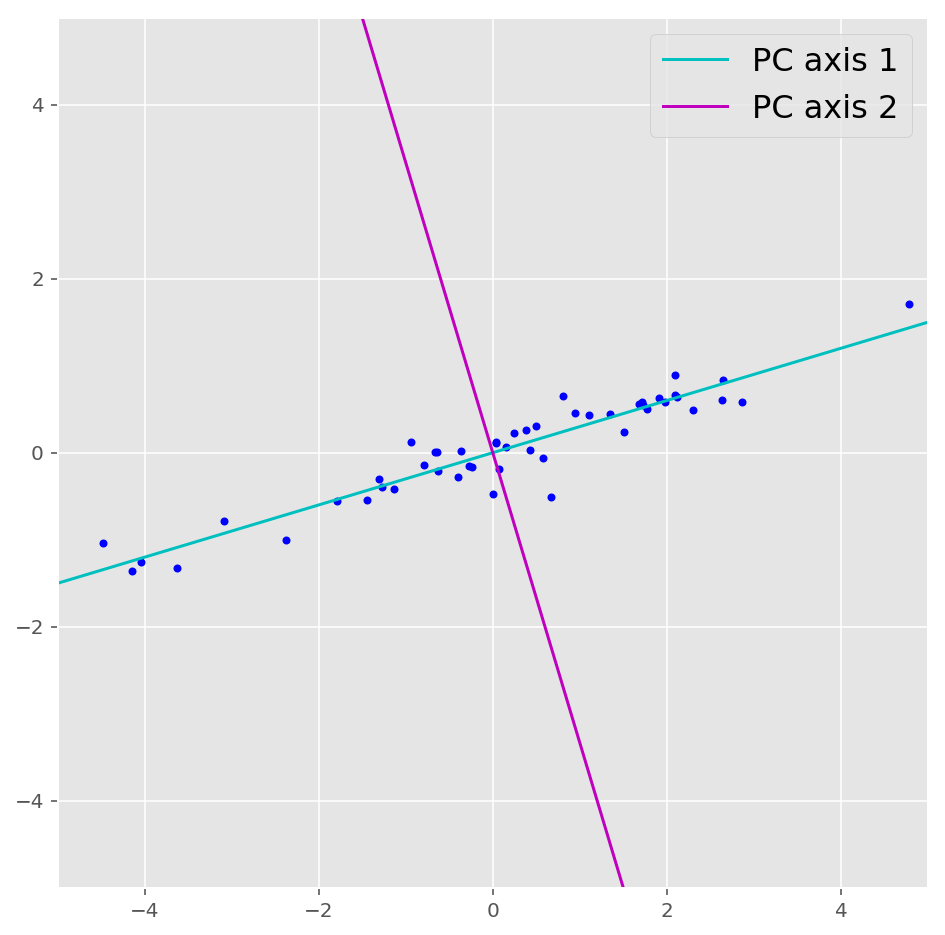
KEY POINTS:

1. Our raw/original data exists in an N-dimensional space where N is number of raw/original features. This assumes that all features are numerical or have been converted from categorical ('apple', 'pear') to numerical.
2. Our original coordinate system is one in which each feature lies along its own axis. The axis are orthogonal, e.g., they are separated by 90 degrees.
3. One could transform ("project") the data onto any one of an infinite number of other coordinate systems. This could be a new coordinate system with the same dimensionality. For example, a 2-D traditional x-y coordinate system versus one in which the axes are 45 degrees off from those traditional axes. Under such a transformation, the data points haven't really moved--we're just using a different set of numbers to describe their location in space--numbers along the axes of the new coordinate system rather than along the old.

DEFINITIONS:

|  |  |
| --- | --- |
| PCA | PCA finds a new coordinate system of the same dimensionality in which most of the variance of the data occurs in a small number of dimensions (though data will likely vary at least somewhat in all the dimensions). |
| Subspace | any lower-dimensional space of the "full" data space. If our original data has a feature for which all the values are identical (variance=0), we know that is it a meaningless features and we could discard it, meaning our data now lives in an N-1 dimensional subspace of the original space. More generally, this has axes that differ from those of the original sub-space. PCA can help use determine what those axes should be. |
| Dimensionality  Reduction | generally, is any method used to reduce the number of dimensions of our data. This could mean simply discarding some of the original features or could mean transforming our data into a new set or coordinates determined by PCA, in which we use only a subset of the PCA axes. |

GRAPHS:

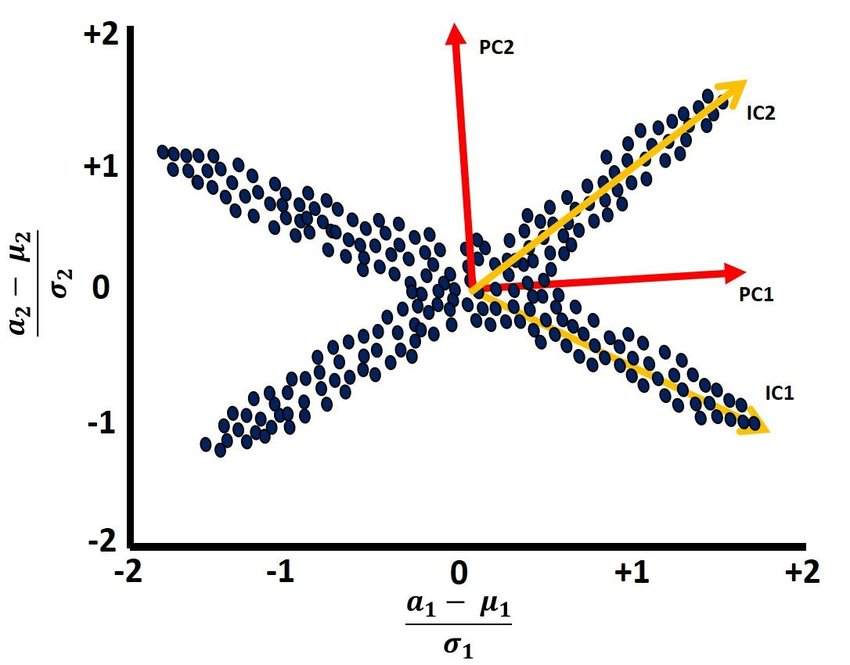


**Figure 1:** PCA determines the new, cyan and magenta axes. Most of the variance of the data occurs along the cyan axis. We can reduce the dimensionality of the data, with very little loss of information, by transforming it into a sub-space that includes only the PC 1 dimension (discard the PC 2 dimension).

A graph with a line and a line

Description automatically generated

**Figure**2: Variance along an axis may be driven by separation of data clusters as seen here. But often it is driven simply by a more distributed range of data values as in Figure 1.



**Figure 3:** PCA axes (red) are always orthogonal. Due to this fact, in the above figure, PCA axes do not align well with the orientations of the spread of the data (subjectively). ICA axes (yellow) can align, since they do not have to be orthogonal. However, unlike in this example, PCA typically does a good job, and is much easier to perform computationally versus ICA.