2. PCA computes c orthonormal vectors which provide a basis for the normalized input data. These are unit vectors such that each point in a direction perpendicular to the others. These vectors are referred to as the *principal components*. The input data are a linear combination of principal components.

3. The principal components are sorted in order of decreasing "significance" or strength. The principal components basically serve as a new set of axes for the data, providing important information about variance. That is, the sorted axes are such that the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on. Figure 7.4 shows the first two principal components, $Y_1$ and $Y_2$, for the given set of data originally mapped to the axes $X_1$ and $X_2$. This information helps identify groups or patterns within the data.

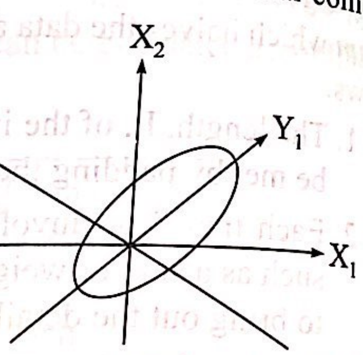

**Fig. 7.4** Principal components analysis. $Y_1$ and $Y_2$ are the first two principal components for the given data.

4. Since the components are sorted according to decreasing order of "significance", the size of the data can be reduced by eliminating the weaker components, *i.e.*, those with low variance. Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

PCA is computationally inexpensive, can be applied to ordered and unordered attributes, and can handle sparse data and skewed data. Multidimensional data of more than two dimensions can be handled by reducing the problem to two dimensions.

In comparison with wavelet transforms for data compression, PCA tends to be better at handling sparse data, while wavelet transforms are more suitable for data of high dimensionality.

### 7.4.2 Principal Comp

Principal components analysis is a technique used to reduce multidimensional data sets to lower dimensions for analysis. Depending on the field of application, it is also named the **discrete Karhunen-Loveve transform,** the **Hotelling transform** or **proper orthogonal decomposition (POD).**

PCA is mostly used as a tool in a exploratory data analysis and for making predictive models. PCA involves the calculation of the eigenvalue decomposition or singular value decomposition of a data set, usually after mean centering the data for each attribute. The results of PCA are usually discussed in terms of component scores and loadings.

Let the data to be compressed consist of N tuples or data vectors, from k dimensions. Principal Component Analysis or PCA searches for *c k*-dimensional orthogonal vectors that can be best used to represent the data, where *c*<< N. The original data are thus projected onto a much smaller space. PCA can be used to perform dimensionality reduction. The basic procedure is as follows:

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.