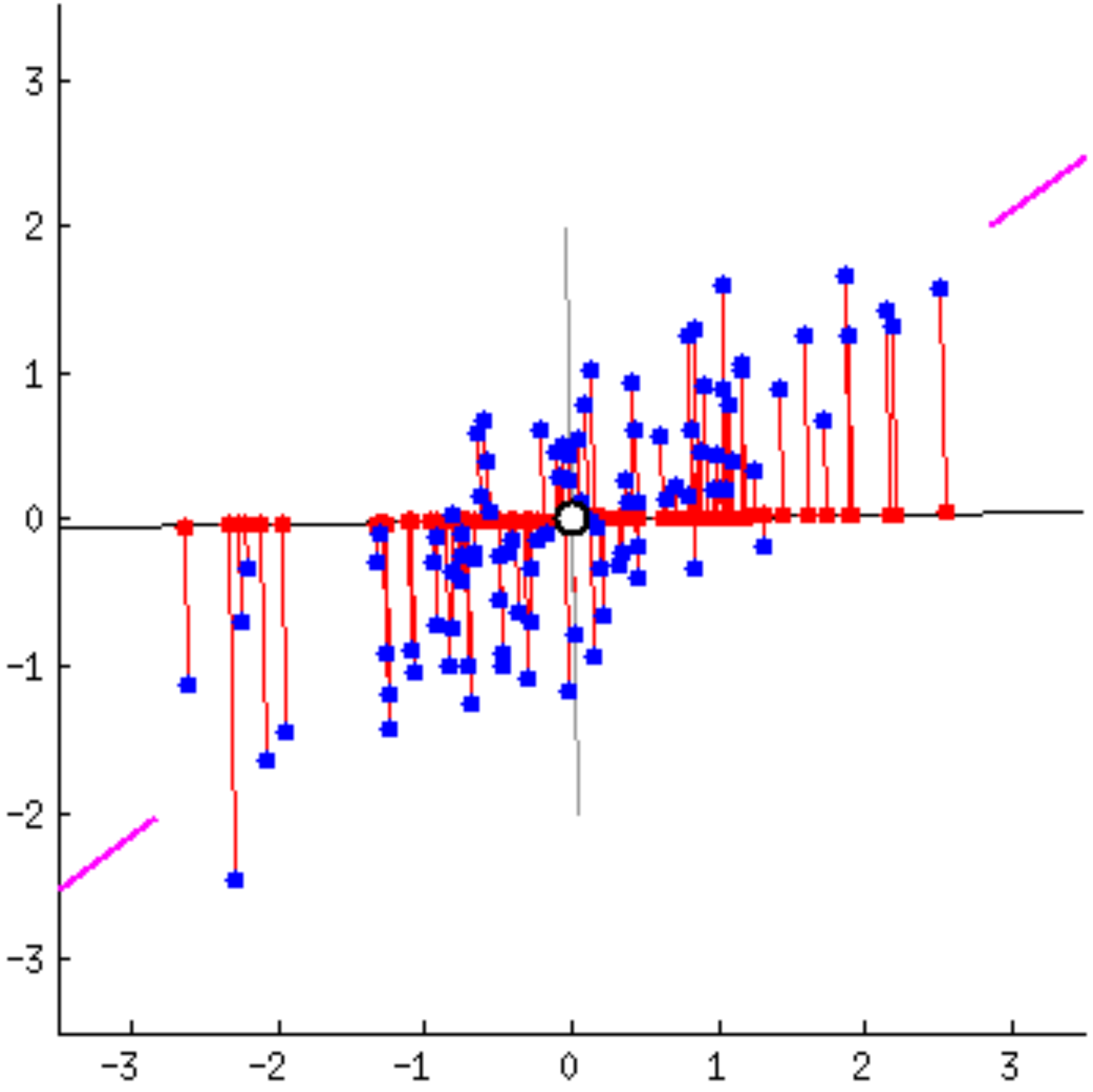
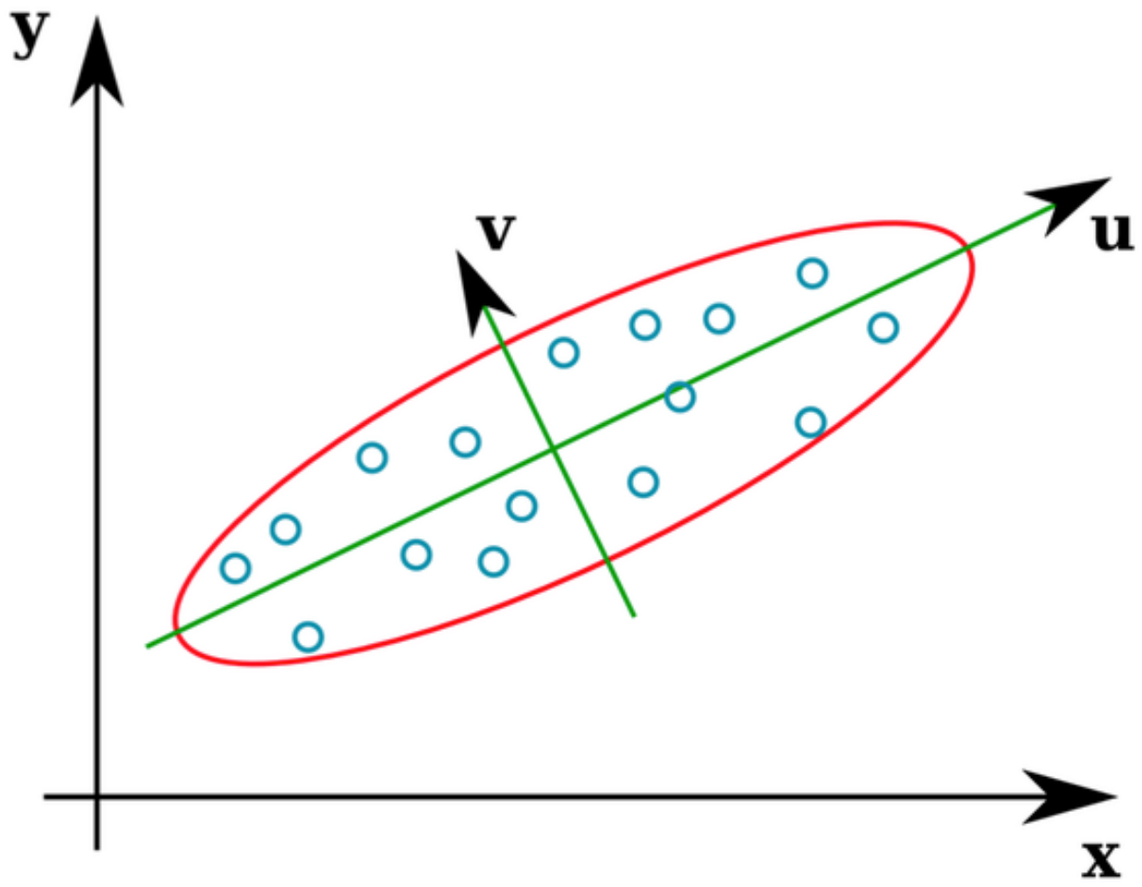
Principal Component Analysis

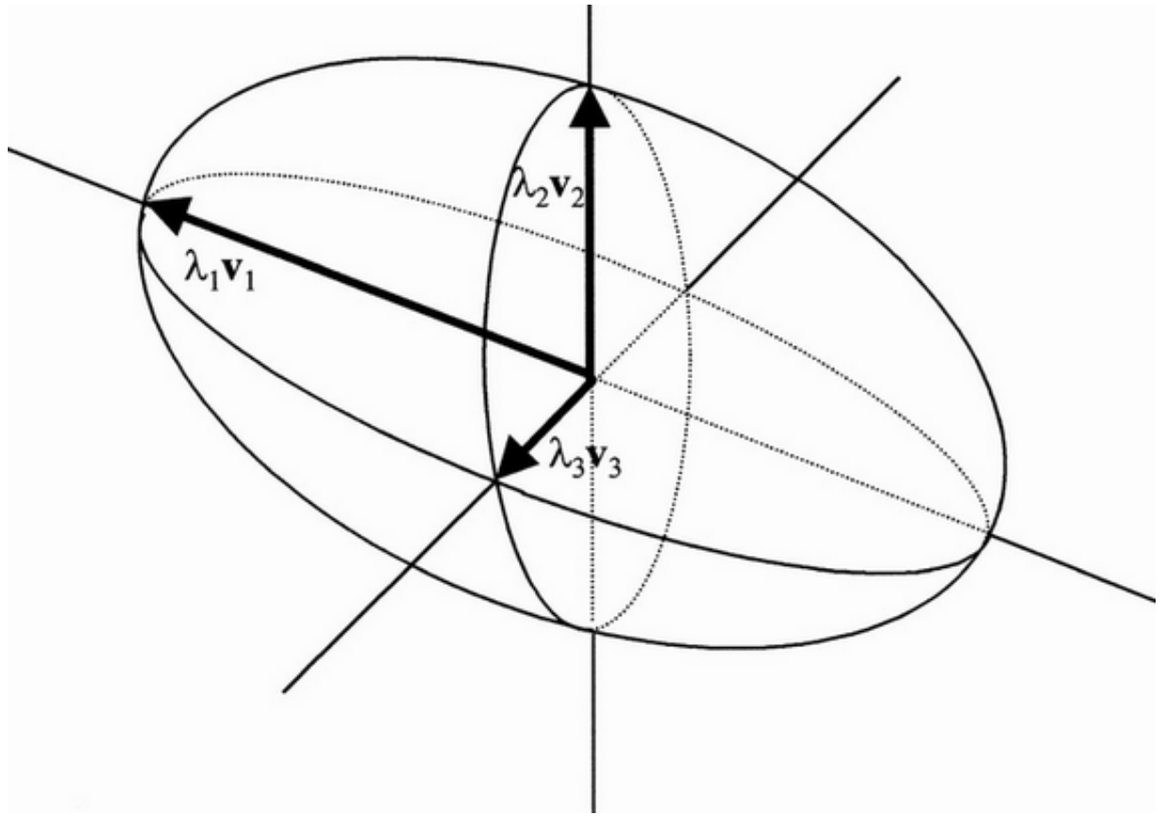
Dimensionality Reduction

On which line will the data points have the most spread?



http://i.stack.imgur.com/Q7HIP.gif





- Each data sample is a 2-dimensional point with coordinates x, y.
- The eigenvectors of the covariance matrix of these data samples are the vectors **u** and **v**;
  - **u**, longer arrow, is the first eigenvector
  - **v**, the shorter arrow, is the second. (The eigenvalues are the length of the arrows.)
  - the first eigenvector points (from the mean of the data) in the direction in which the data varies the most in Euclidean space, and the second eigenvector is orthogonal (perpendicular) to the first.

PCA: Main steps

1. Obtain **covariance matrix**

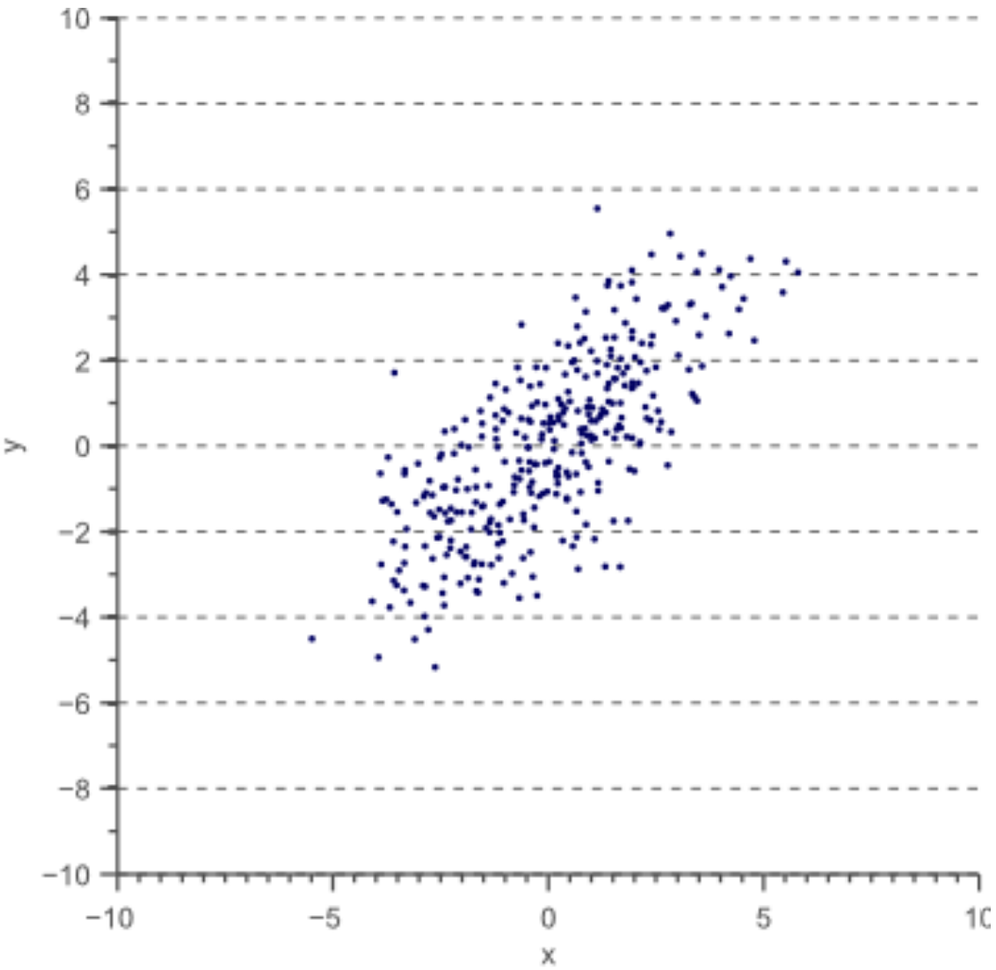
2. Obtain **eigenvalues** by solving the function $\det(A - \lambda I) = 0$

3. Obtain the **eigenvector** by solving for matrix X such that $[A - \lambda I][X] = 0$

4. Obtain the new **coordinates** of each data point in the direction of eigenvectors.

Variance and its limits

- Variance can only be used to explain the spread of data in the directions parallel to the axes of the feature space



- For this data, we could calculate the variance in the x-direction and the variance in the y-direction.

  - However, the horizontal spread and the vertical spread of the data does **not** explain the clear diagonal correlation.

Covariance Matrix

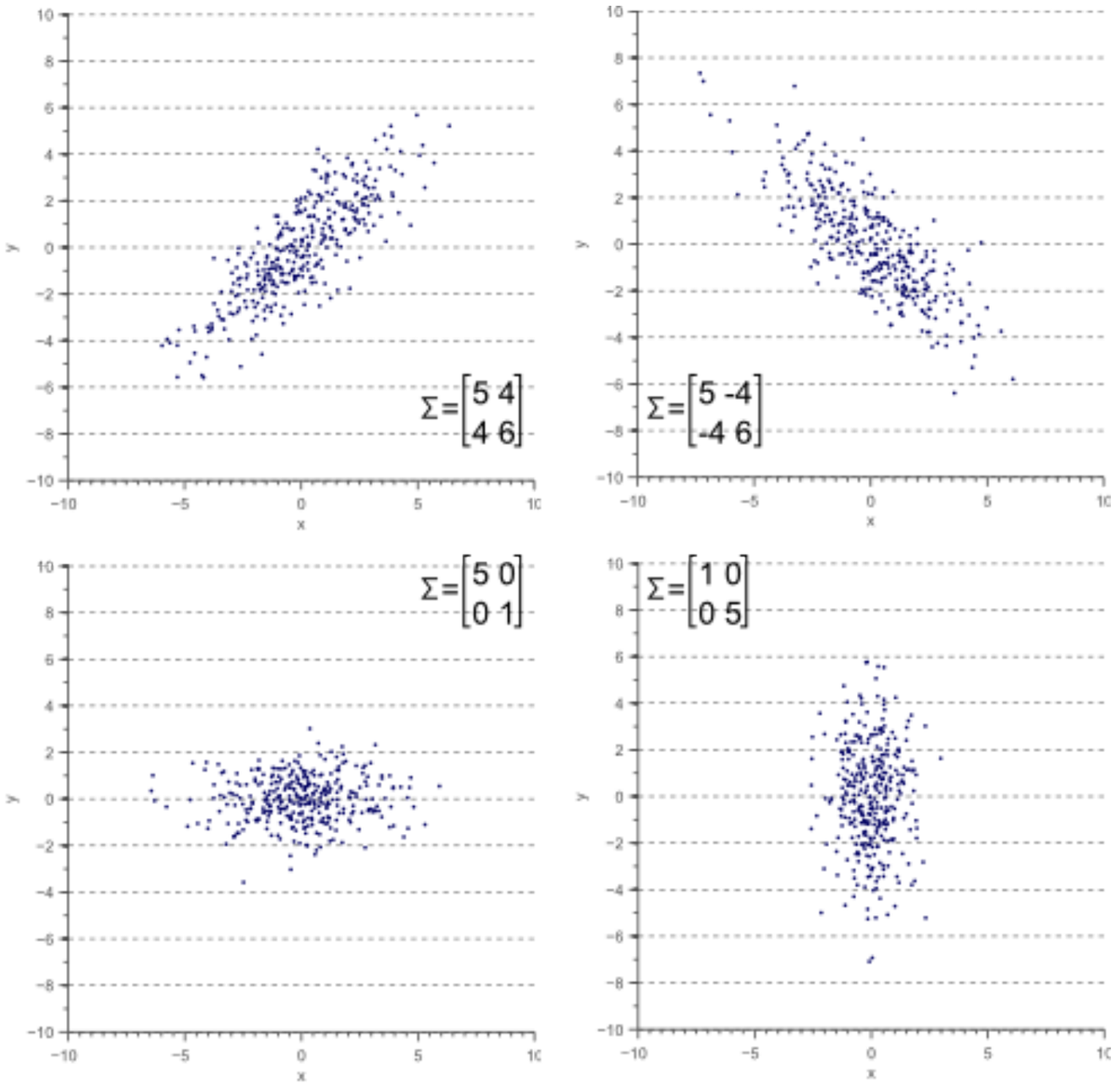
$$\Sigma = \begin{bmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{bmatrix}$$

- If x is positively correlated with y, y is also positively correlated with x. In other words, we can state that var(x,y) = var(y,x).

- the covariance matrix is always a symmetric matrix

- with the variances on its diagonal and the covariances off-diagonal.

- Diagonal spread is captured by the covariance, while axis-aligned spread is captured by the variance.



Eigen-decomposition of a covariance matrix

- Covariance matrix defines both the spread (variance), and the orientation (covariance) of our data.
- To represent the covariance matrix with a vector and its magnitude, try to find the vector that points into the direction of the largest spread of the data, and whose magnitude equals the spread (variance) in this direction.

Eigen-decomposition of a covariance matrix

- The largest eigenvector of the covariance matrix always points into the direction of the *largest variance of the data*, and the magnitude of this vector equals the corresponding **eigenvalue**.
- The second largest eigenvector is always orthogonal to the largest eigenvector, and points into the direction of the second largest spread of the data.