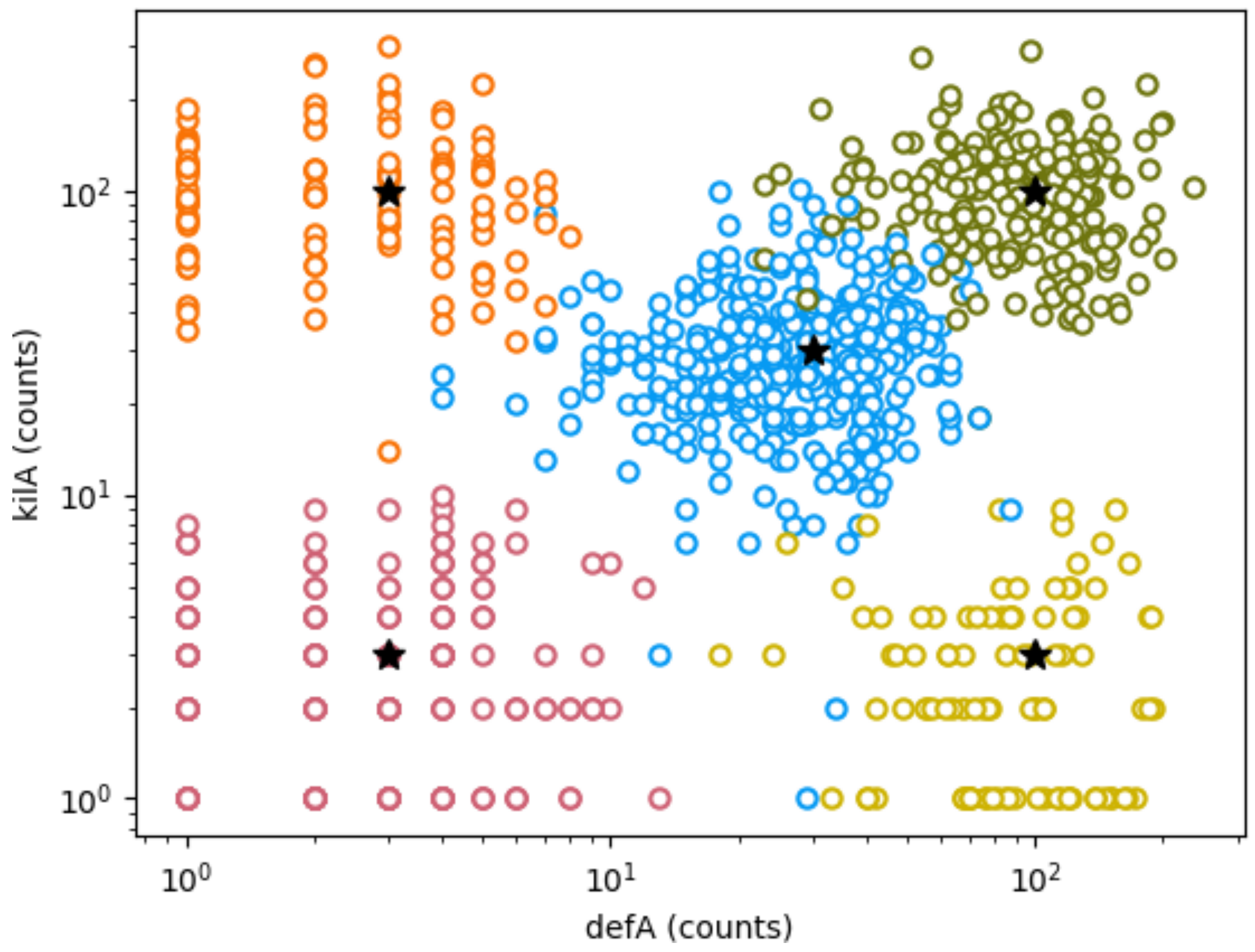
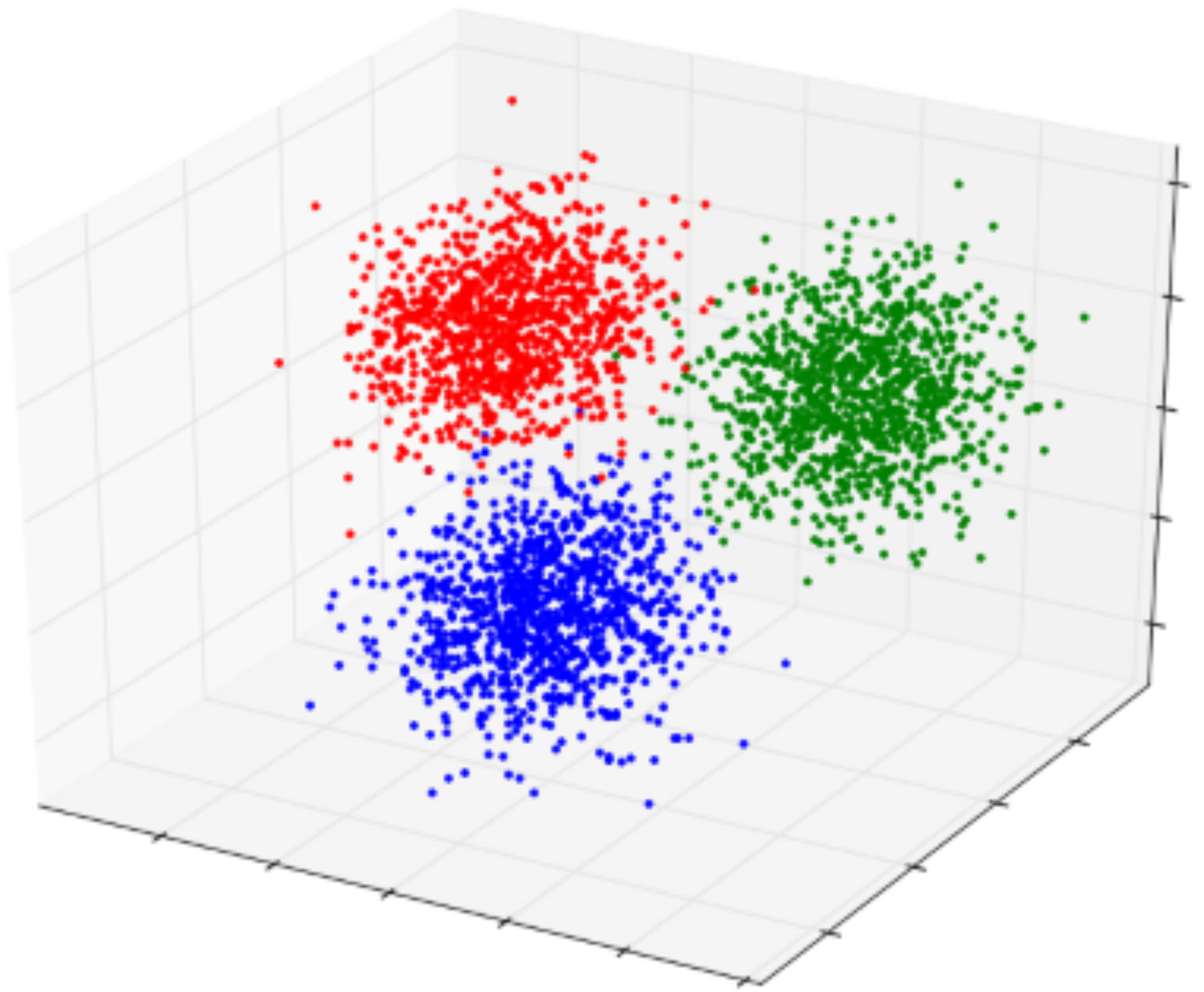
Week 11: PCA

What is Principal Component Analysis and SVD

Higher Dimensional Data





N=2

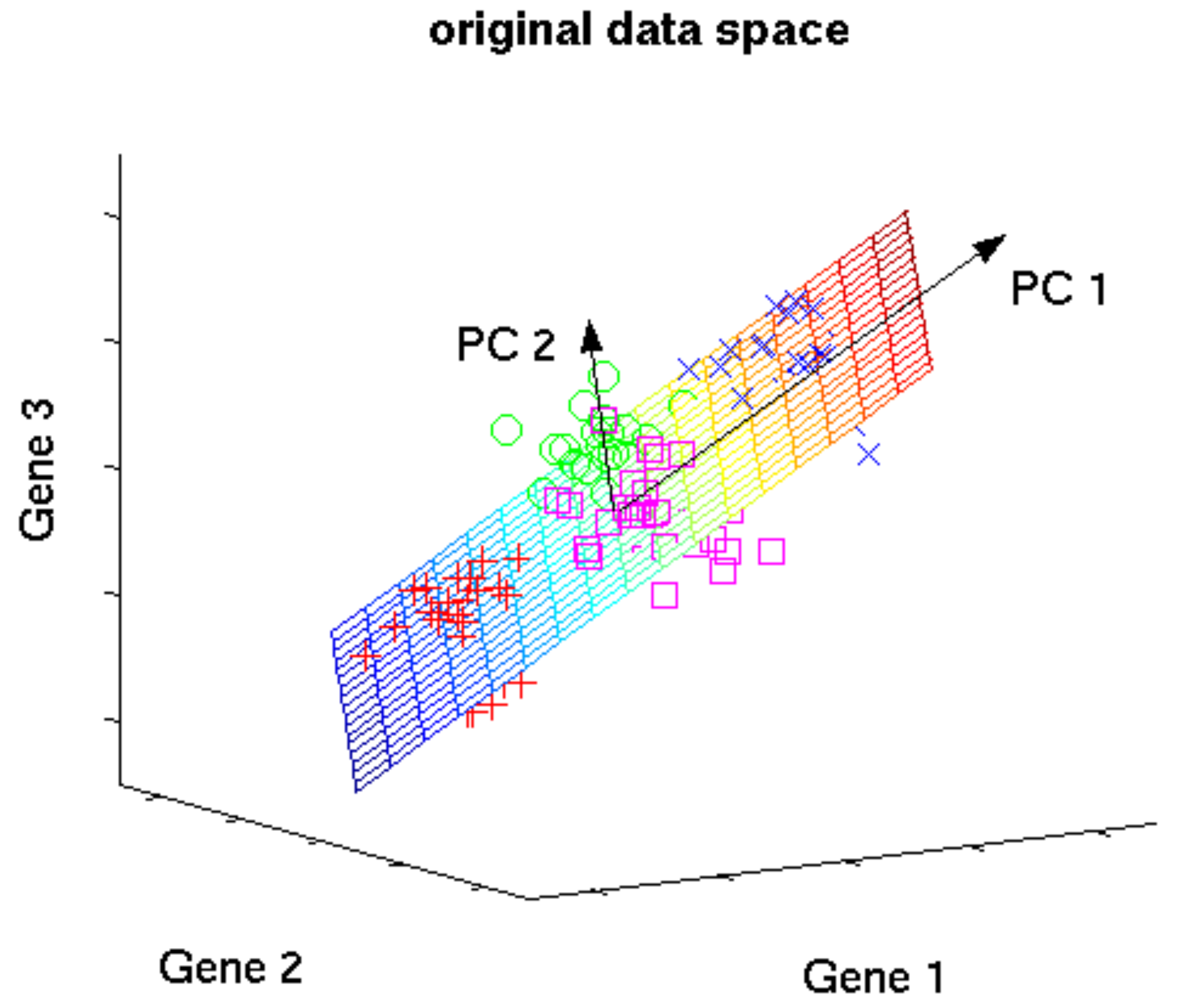
N=3

N=4

(Each point is a single cell (observation), we are working in ‘gene space’

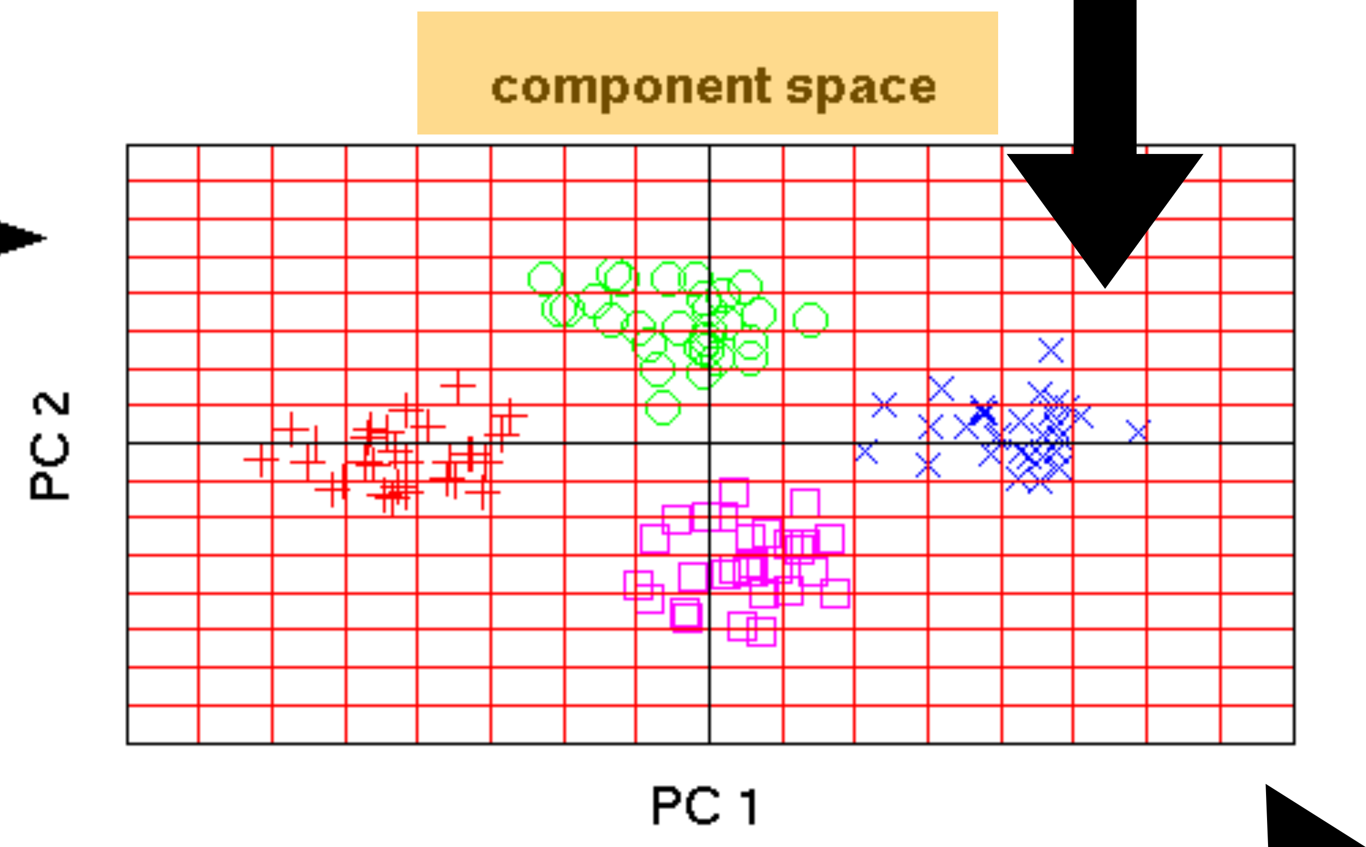
PCA: A rotation

Multidimensional Gaussian with its
2 PC's marked

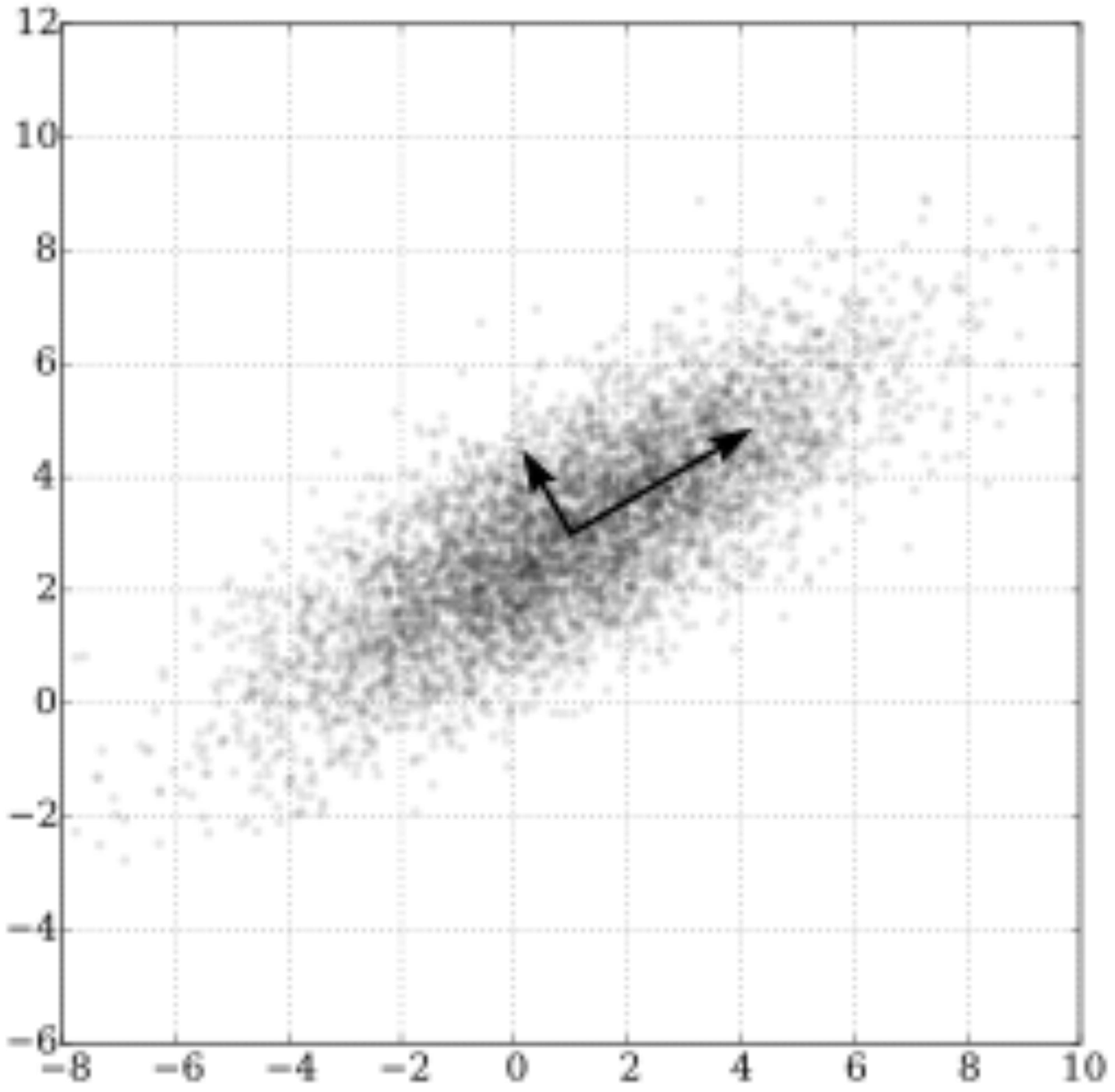


PCA

PC: data along new vector



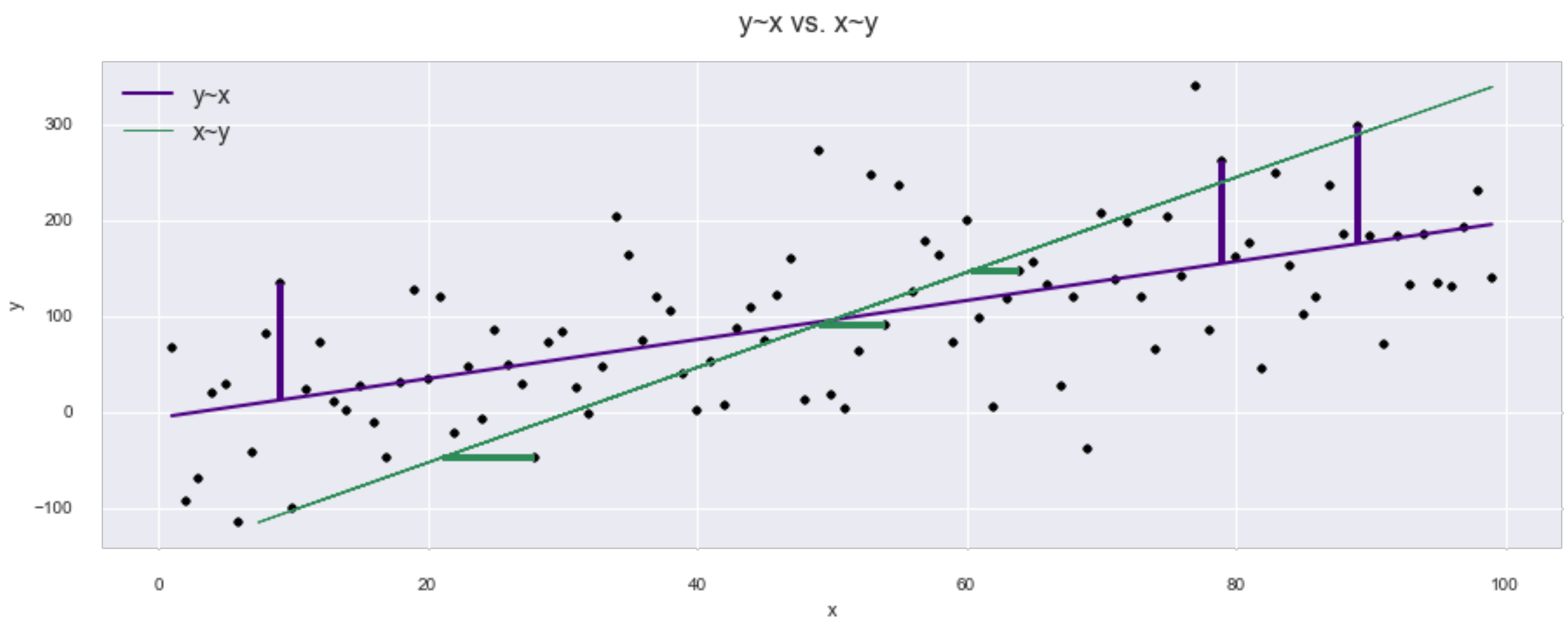
Eigen Vector

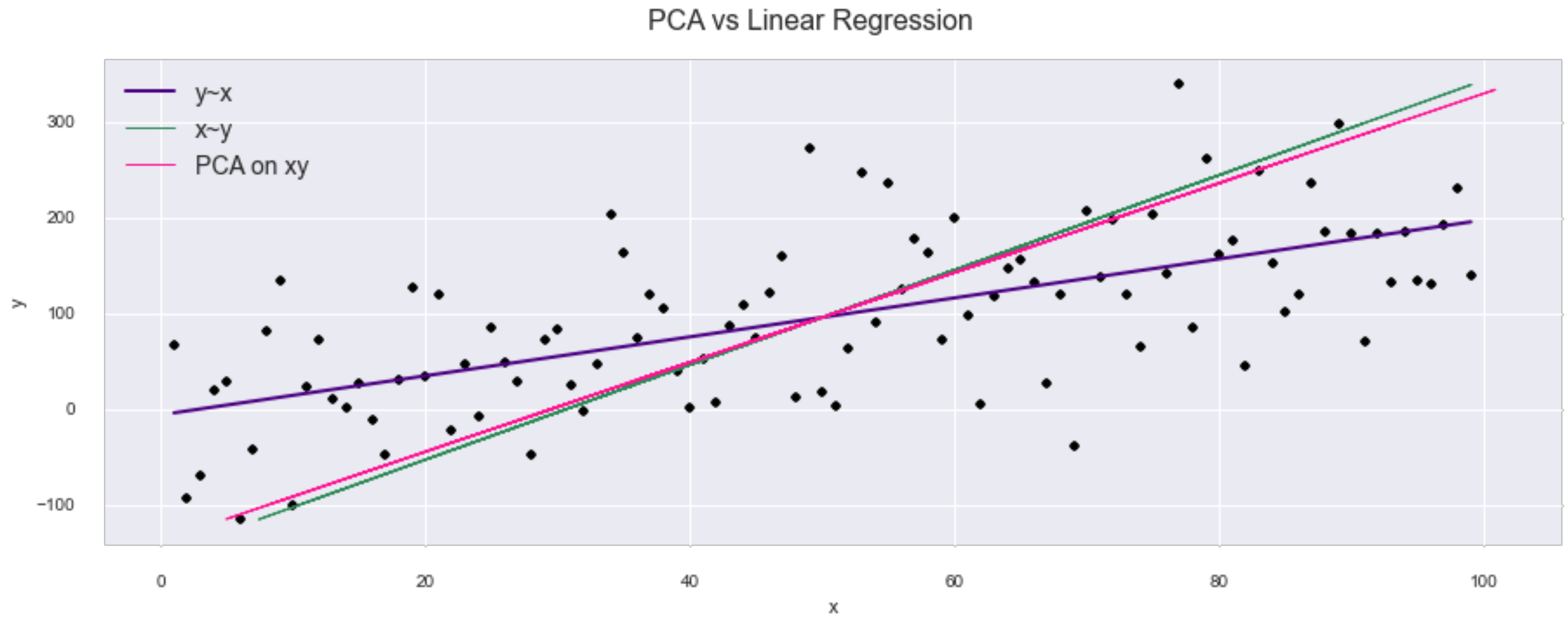


Transforms data to a new coordinate system where each axes captures maximum possible variance. We are now in 'component space'.

(Note: Variables covary which gives us information)

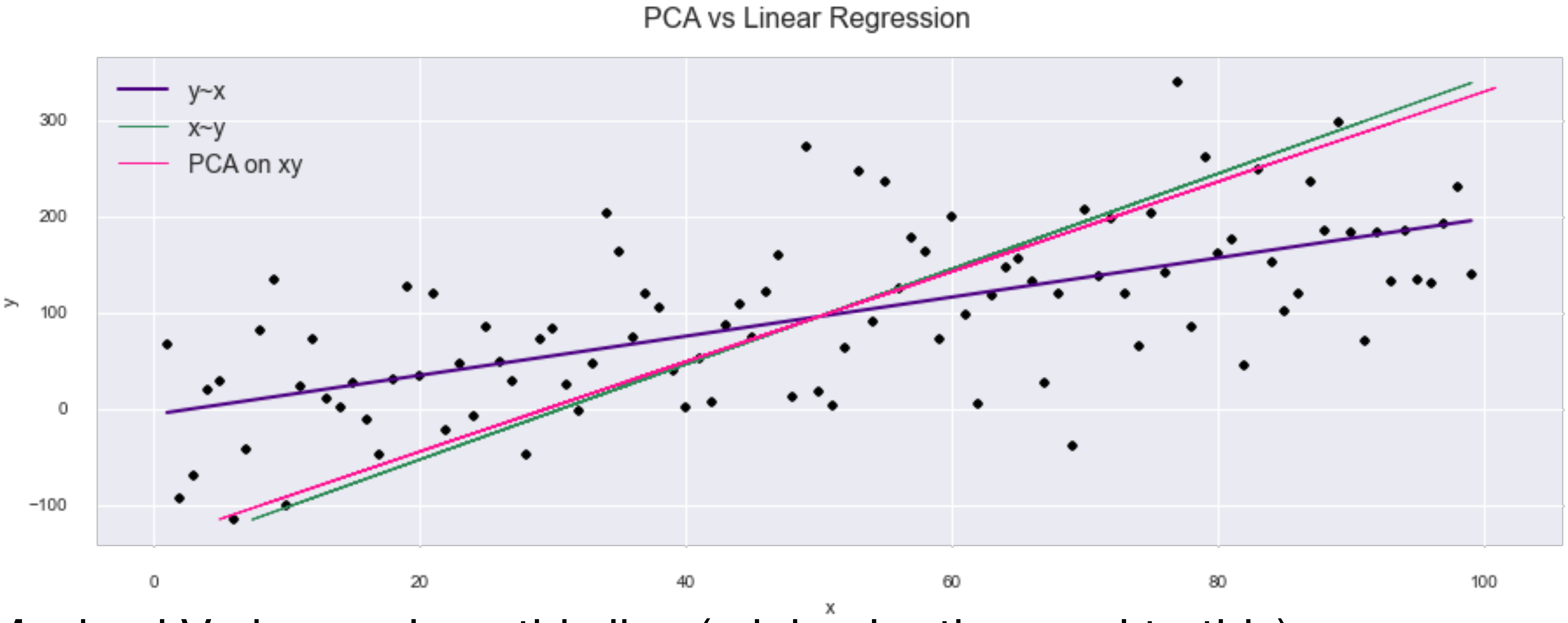
Linear Regression (OLS) vs PC





https://shankarmsy.github.io/posts/pca-vs-lr - Know Thy Data

What about covariance across this line?

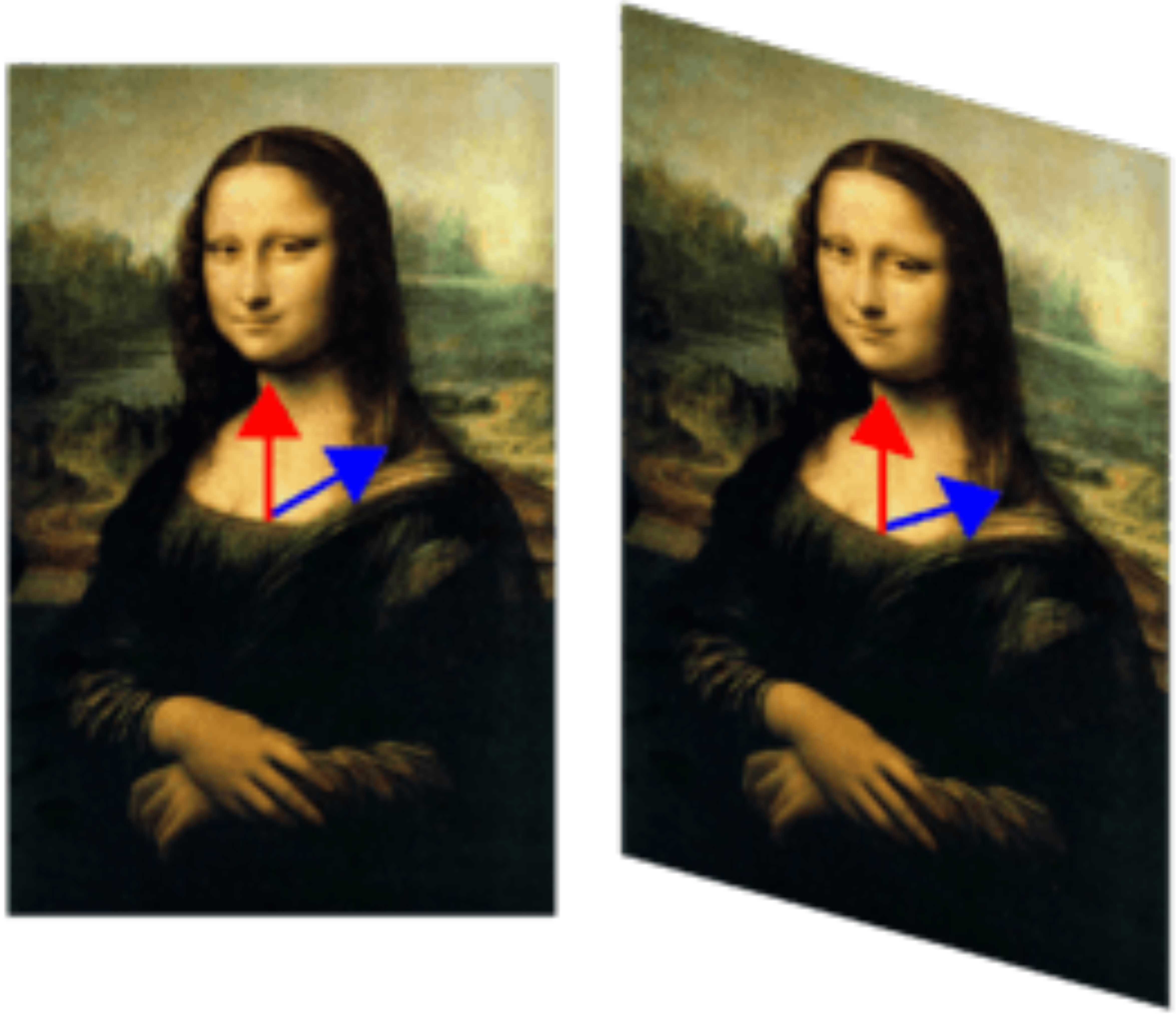


Maximal Variance along this line (minimal orthogonal to this)

No covariance left along this line (if there is non zero covariance, we can still rotate more to get a better line of best fit, we have not gotten maximal variance along line yet)

https://shankarmsy.github.io/posts/pca-vs-lr - Know Thy Data

What is an Eigen vector?



Eigen vectors of a covariance matrix provide the direction along which variance varies the most. This is exactly what we want!

Eigen-vectors represent the axes of our new coordinate space. They will no longer change direction in a linear transformation

The associated Eigen Values tell us how much variance that particular eigen vector captures.

https://kids.kiddle.co/Eigenvalues_and_eigenvectors

Eigen Vectors and Covariance Matrices

Data Matrix: **X** (n points in p dimensions)
(n cells and p genes)

$$\Sigma = \begin{bmatrix} \mathrm{cov}(1,1) & \mathrm{cov}(1,2) & \dots & \mathrm{cov}(1,p) \\ \mathrm{cov}(2,1) & \mathrm{cov}(2,2) & \dots & \mathrm{cov}(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(p,1) & \mathrm{cov}(p,2) & \dots & \mathrm{cov}(p,p) \end{bmatrix}$$

P dimensional data (for ex: p genes)

$$\mathbf{\Sigma} = \mathbf{W \Lambda W^T}$$

(Eigendecomposition)

Finds eigen values **(W)** with zero covariances

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

λ_i is eigen value

Single Value Decomposition and rank

Assume only n observations, but p variables where n<p. Can we fit it?

$$\text{rank of X } r \leq \min(n, p)$$

Can apply SVD only on cantered data

$$x^*_{ij} = x_{ij} - \bar{\mathbf{x}}_j$$

Can decompose it now as:

$$\mathbf{X}^* = \mathbf{USW}^\top.$$

$$\mathbf{W}^\top_{r \text{ x } p}$$

Captures the r principal components, eigen vectors of our covariance matrix

Plotting data in our lower dimensional space

Assume q PC's are kept

$$\mathbf{Y}_q = \mathbf{X}^*\mathbf{W}_q$$

Data in the new component space

We are rotating the data by our new vectors in **W**