Psych 253
Advanced Statistical Modeling

Eigenstuffs: Linear Algebra & the Principal Axes of Data

Daniel Yamins
Wu Tsai Neurosciences Institute
Departments of Psychology and Computer Science
Stanford Artificial Intelligence Laboratory
Stanford University

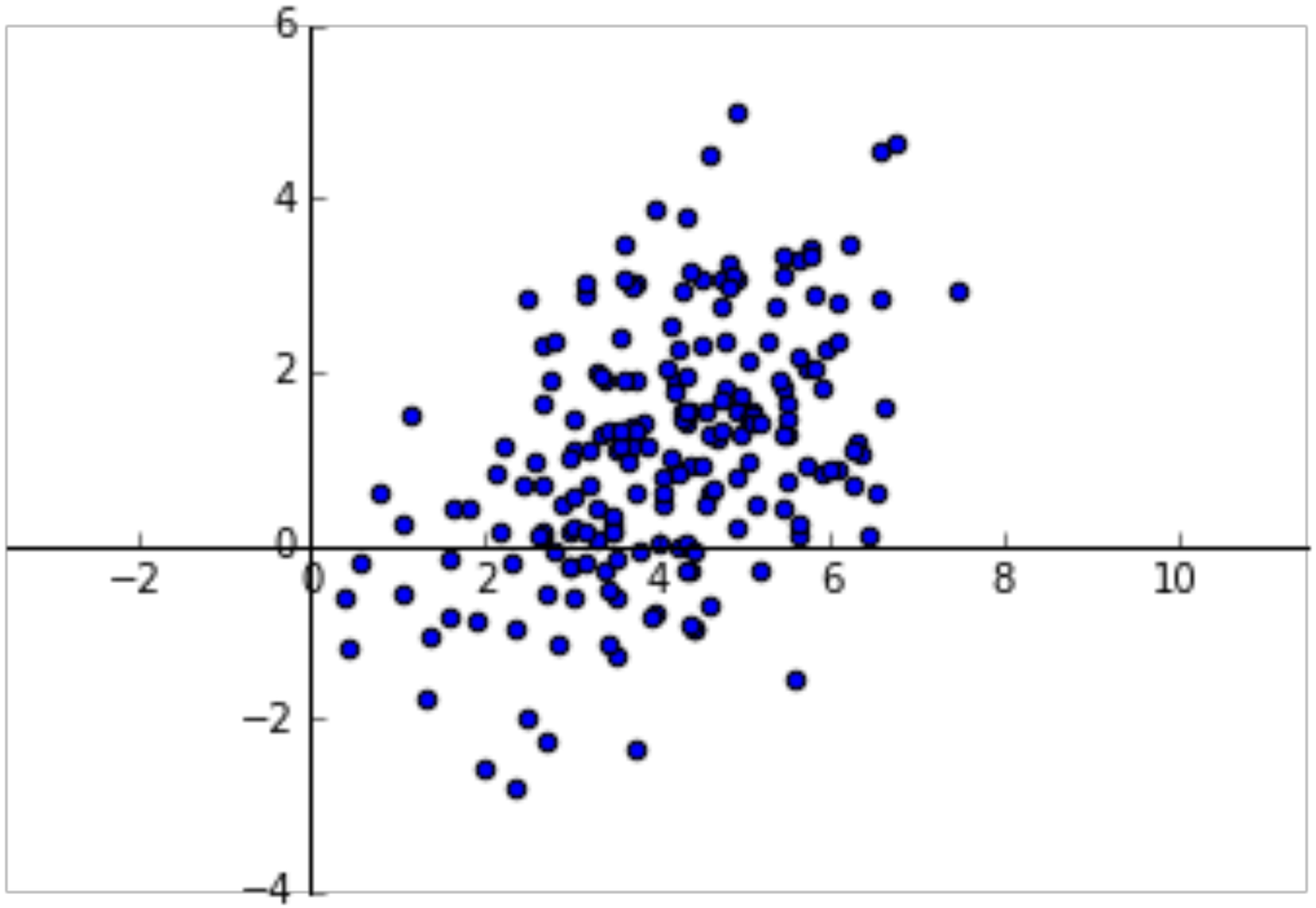
Russ Poldrack
Department of Psychology
Stanford University

Recover Principle Axes of Data

[Ipynb: Principal Axes in Data]

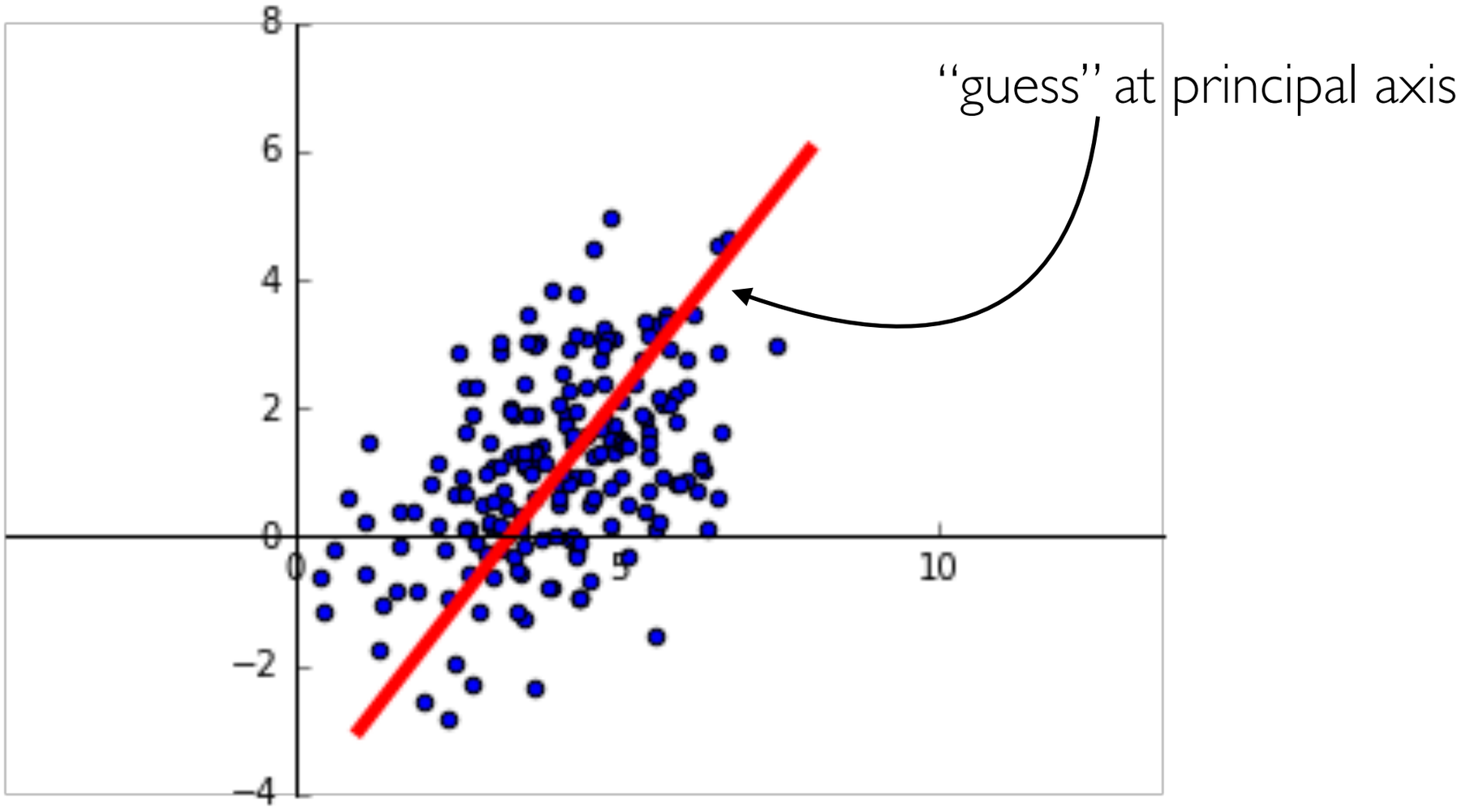
Recover Principle Axes of Data

Some data:



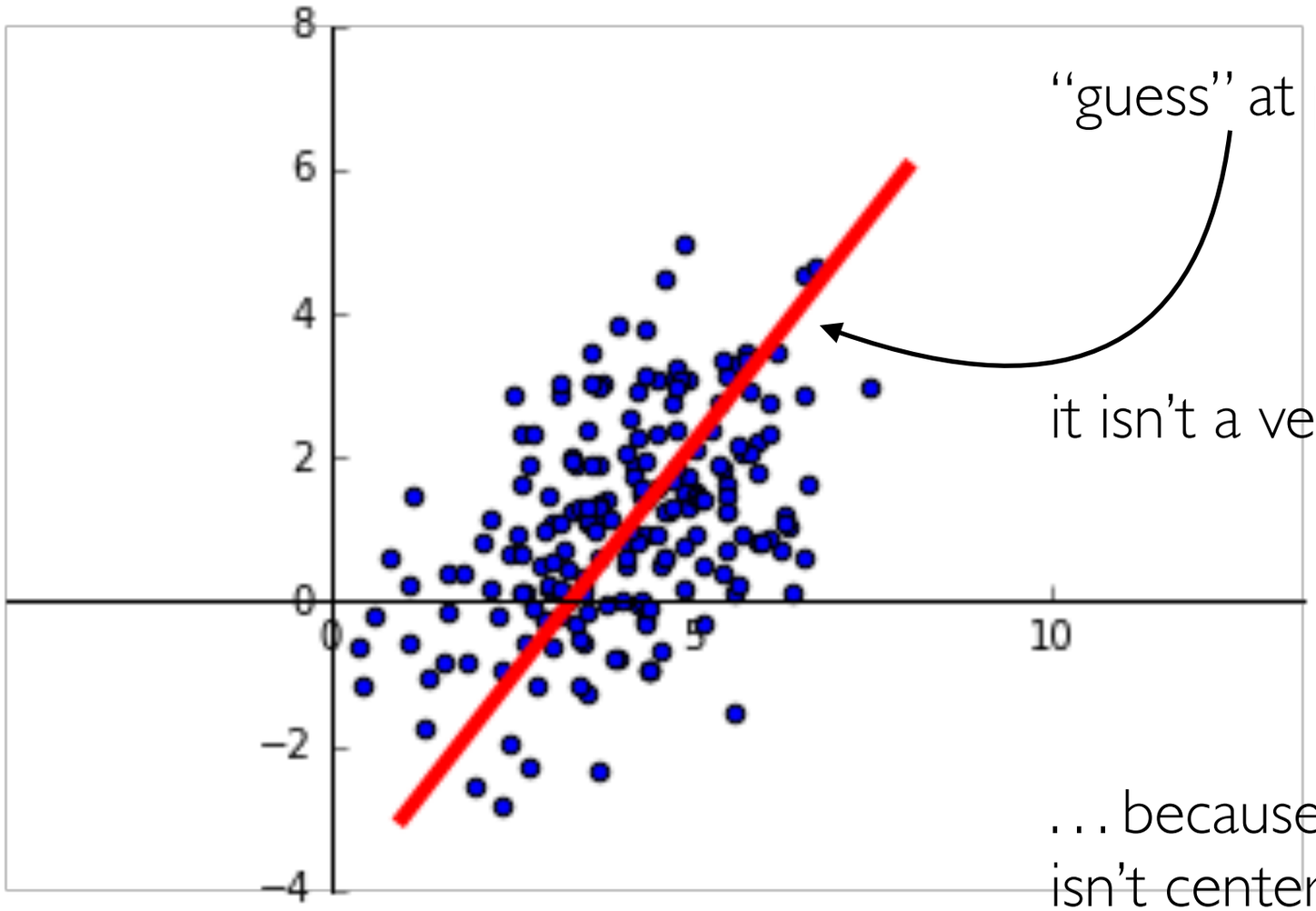
Recover Principle Axes of Data

Some data:



Recover Principle Axes of Data

Some data:



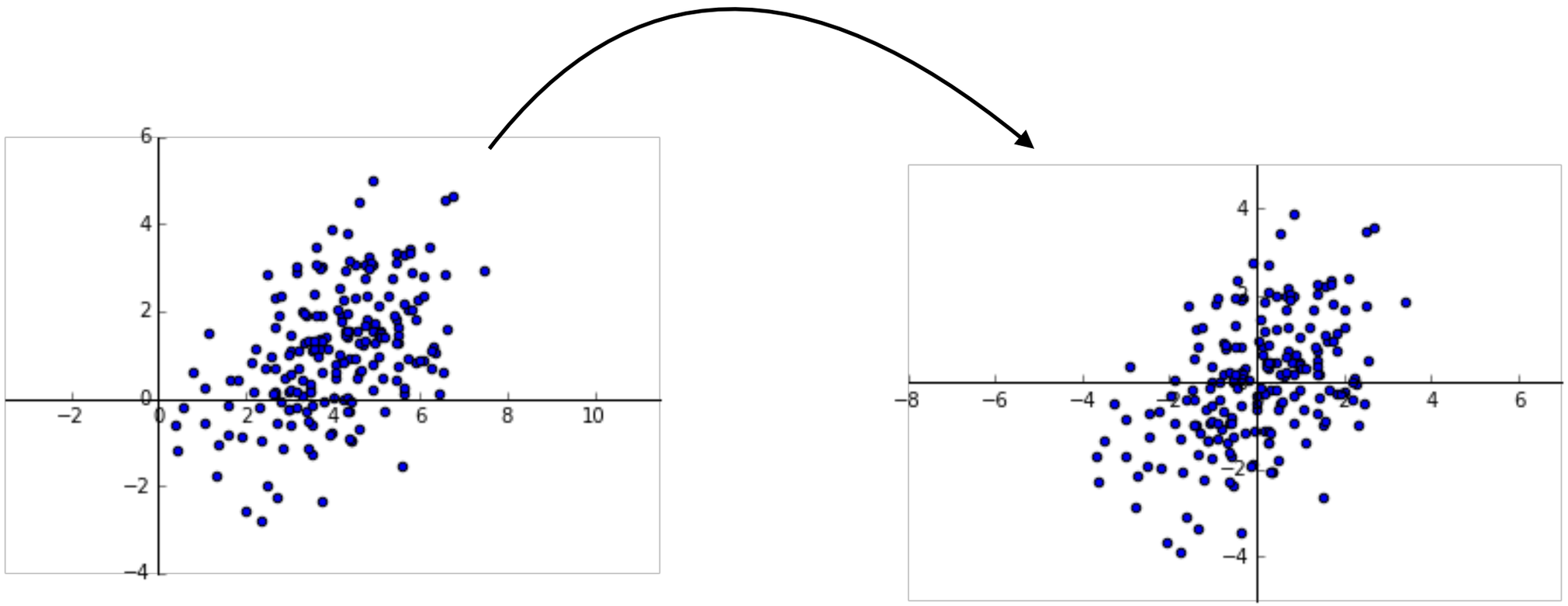
"guess" at principal axis

it isn't a vector ….

… because the data isn't centered

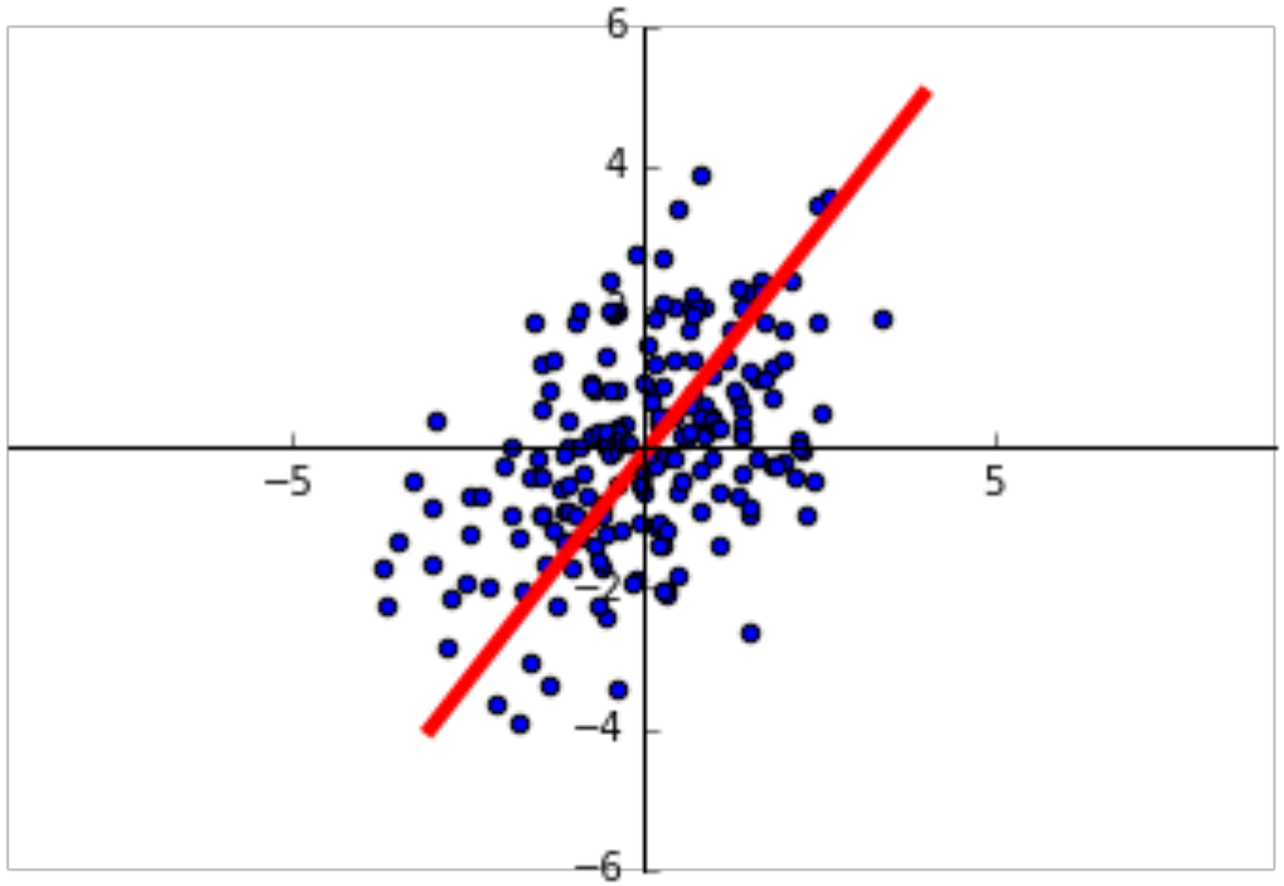
Recover Principle Axes of Data

Center the data: `Xctr = X - X.mean(0)`



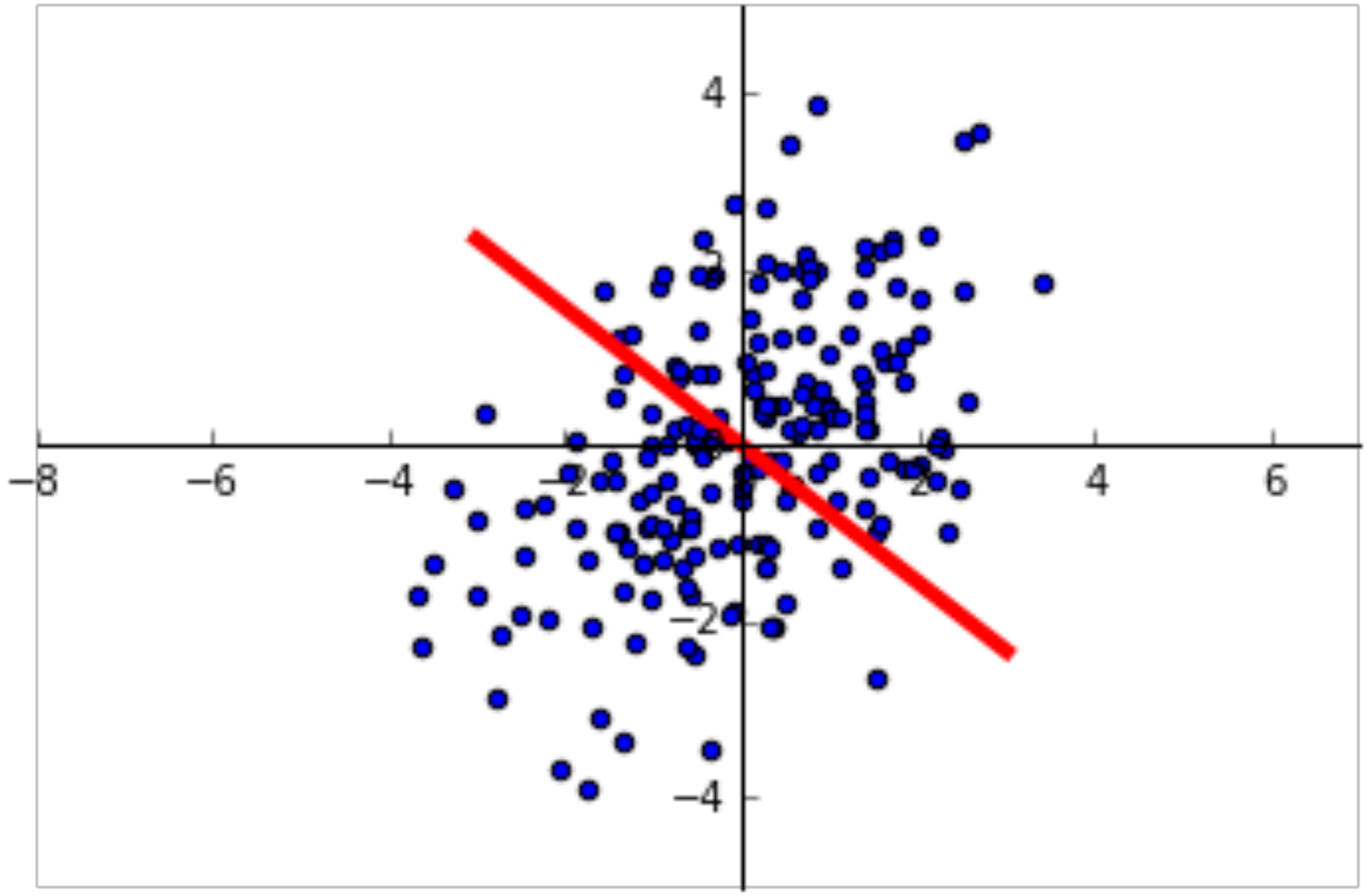
Recover Principle Axes of Data

Now the principal axis is a vector — goes through **`(0,0)`**



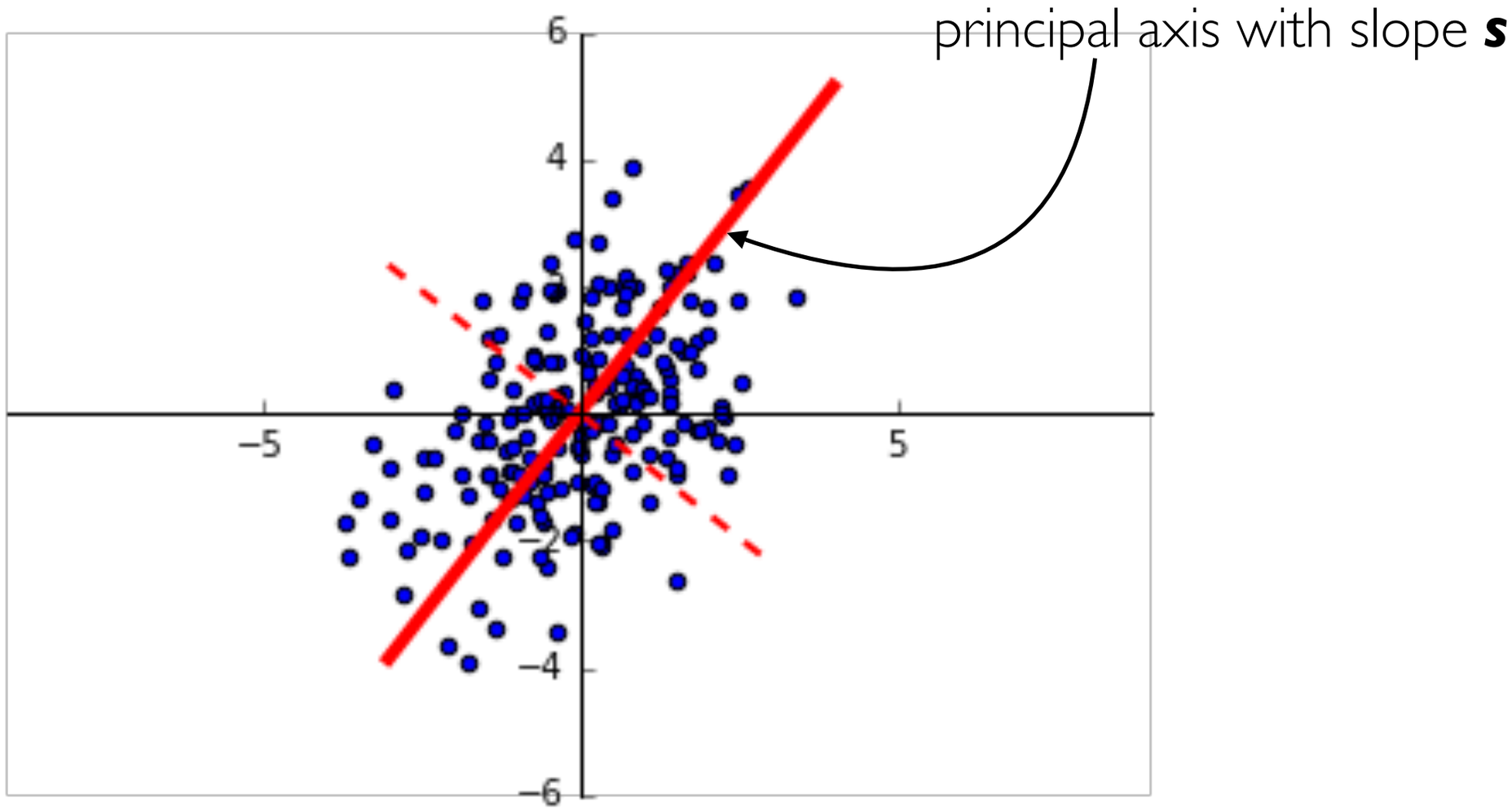
Recover Principle Axes of Data

Here's the axis orthogonal to the principal axis:



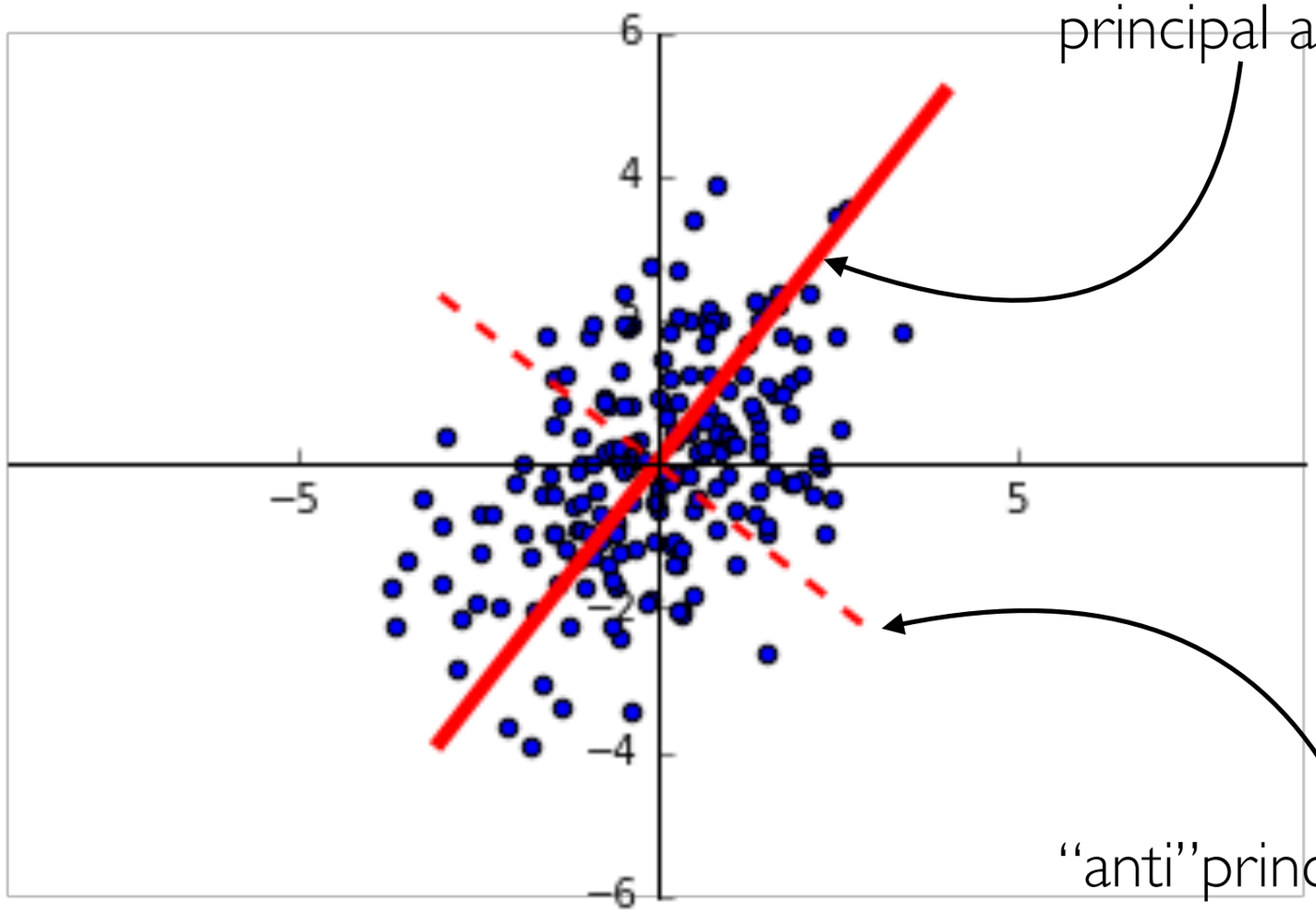
Recover Principle Axes of Data

Plotting both at the same time:



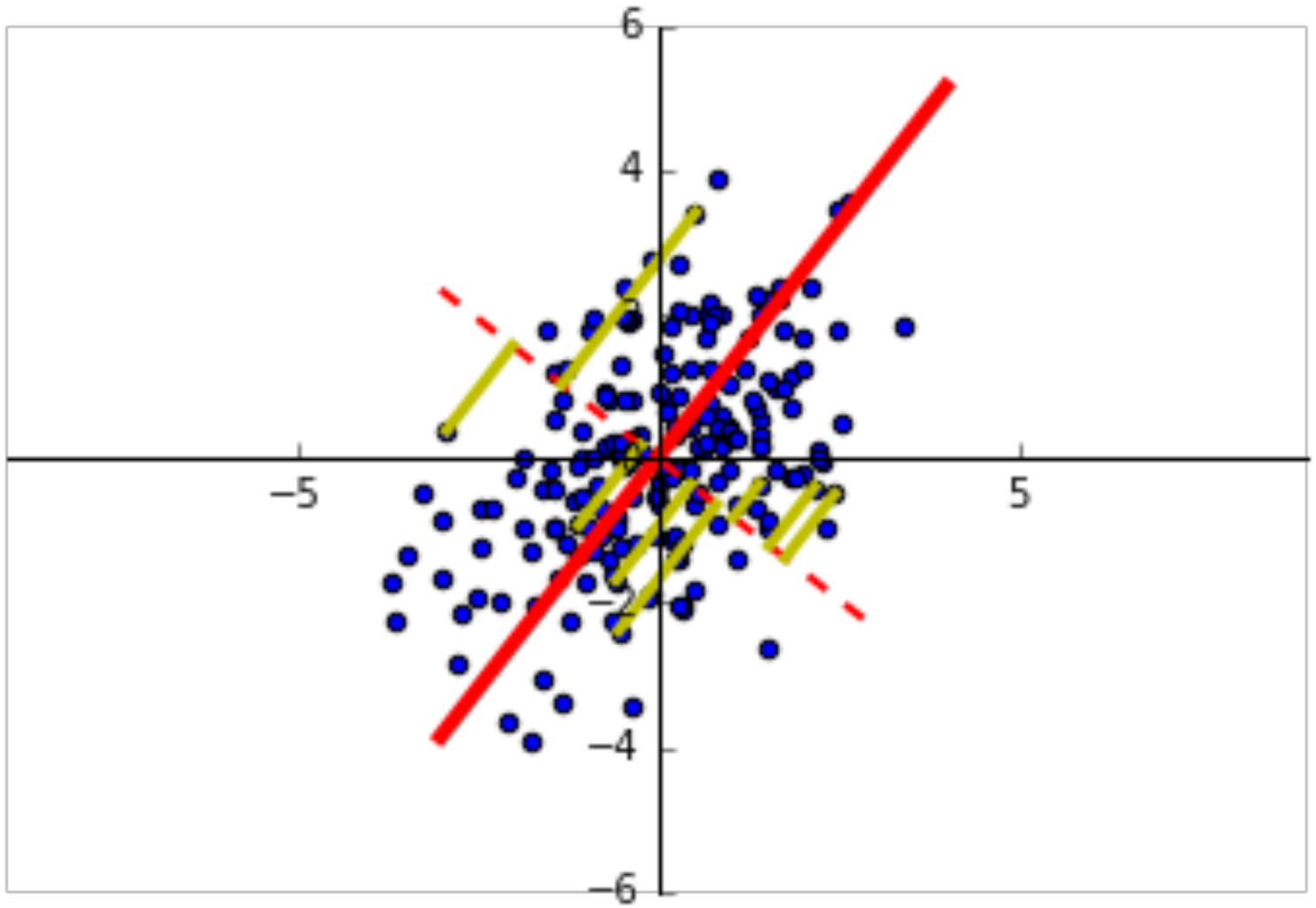
Recover Principle Axes of Data

Plotting both at the same time:



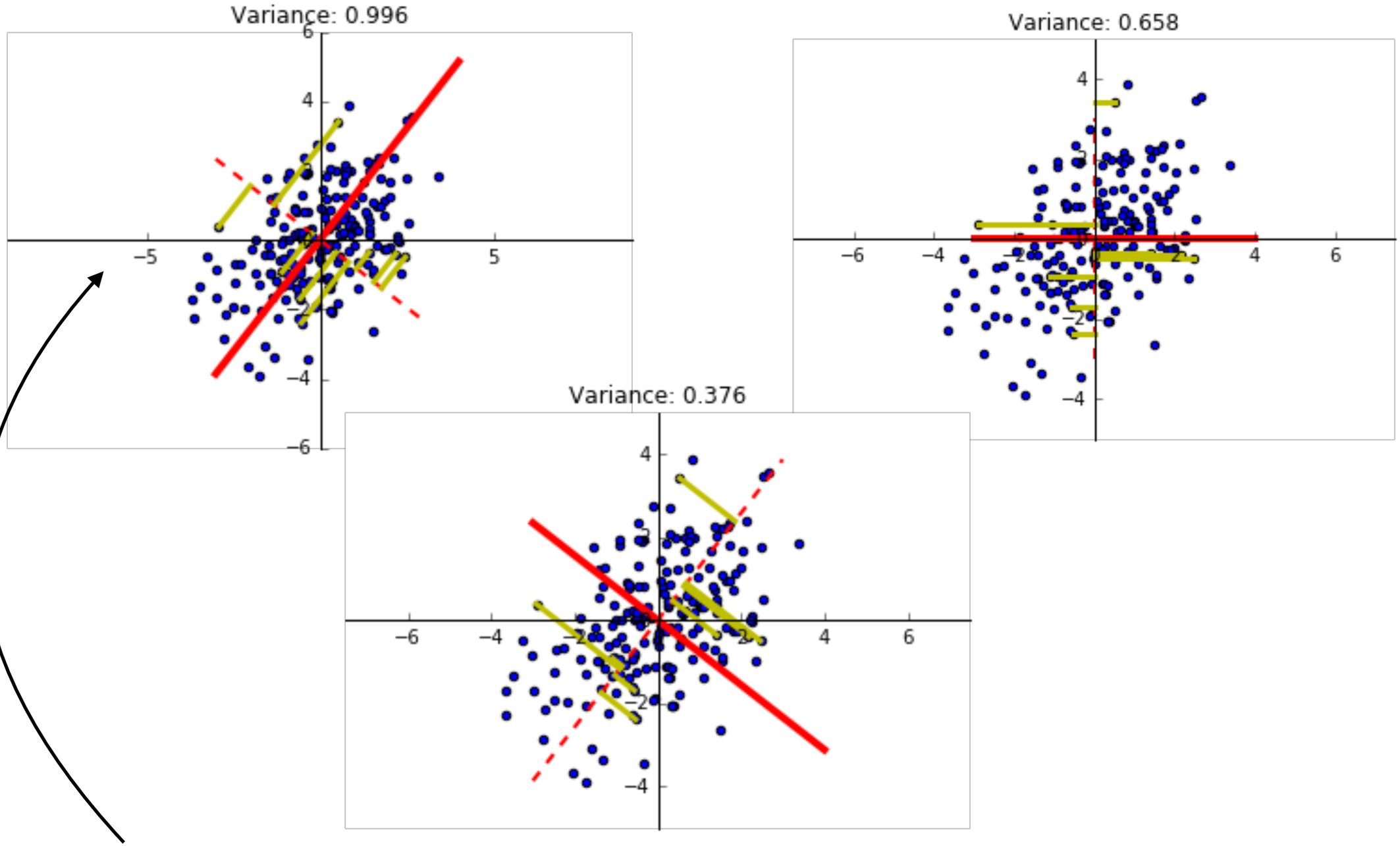
Recover Principle Axes of Data

Distances of points to (hyperplane) orthogonal to the principal axis:



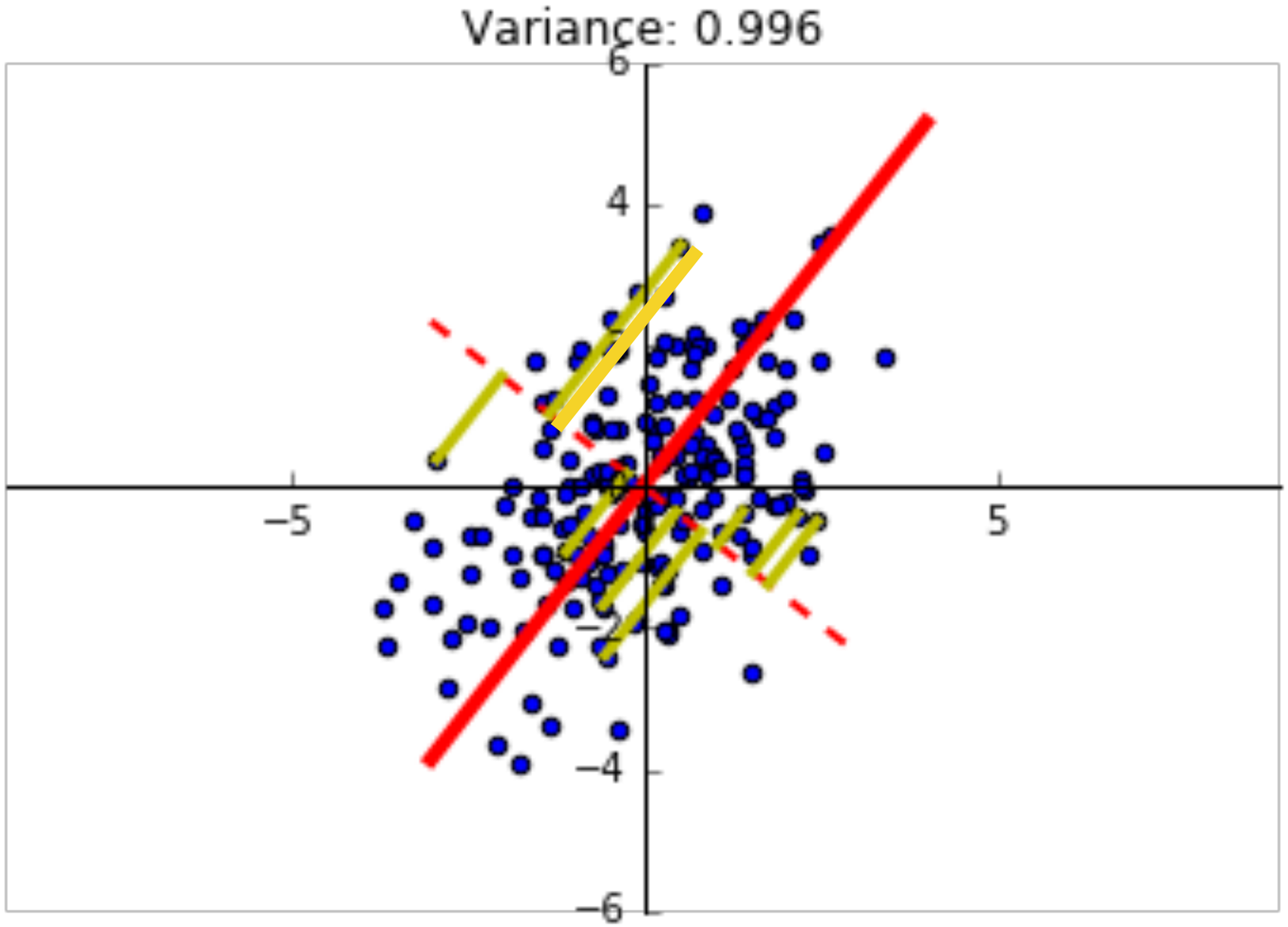
Recover Principle Axes of Data

Distances of points to (hyperplane) orthogonal to various axes:

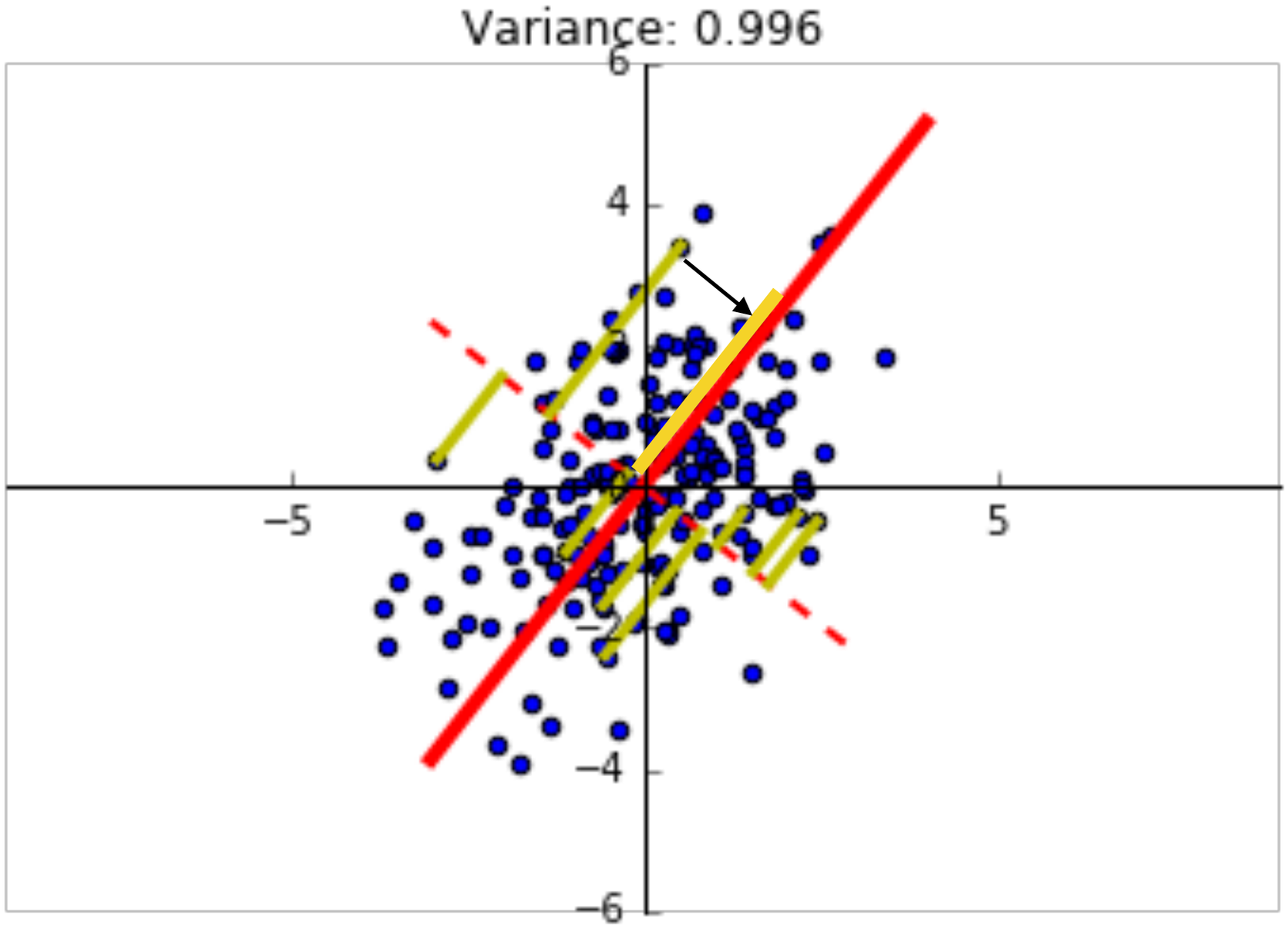


Principle axis maximizes the variance of these distances

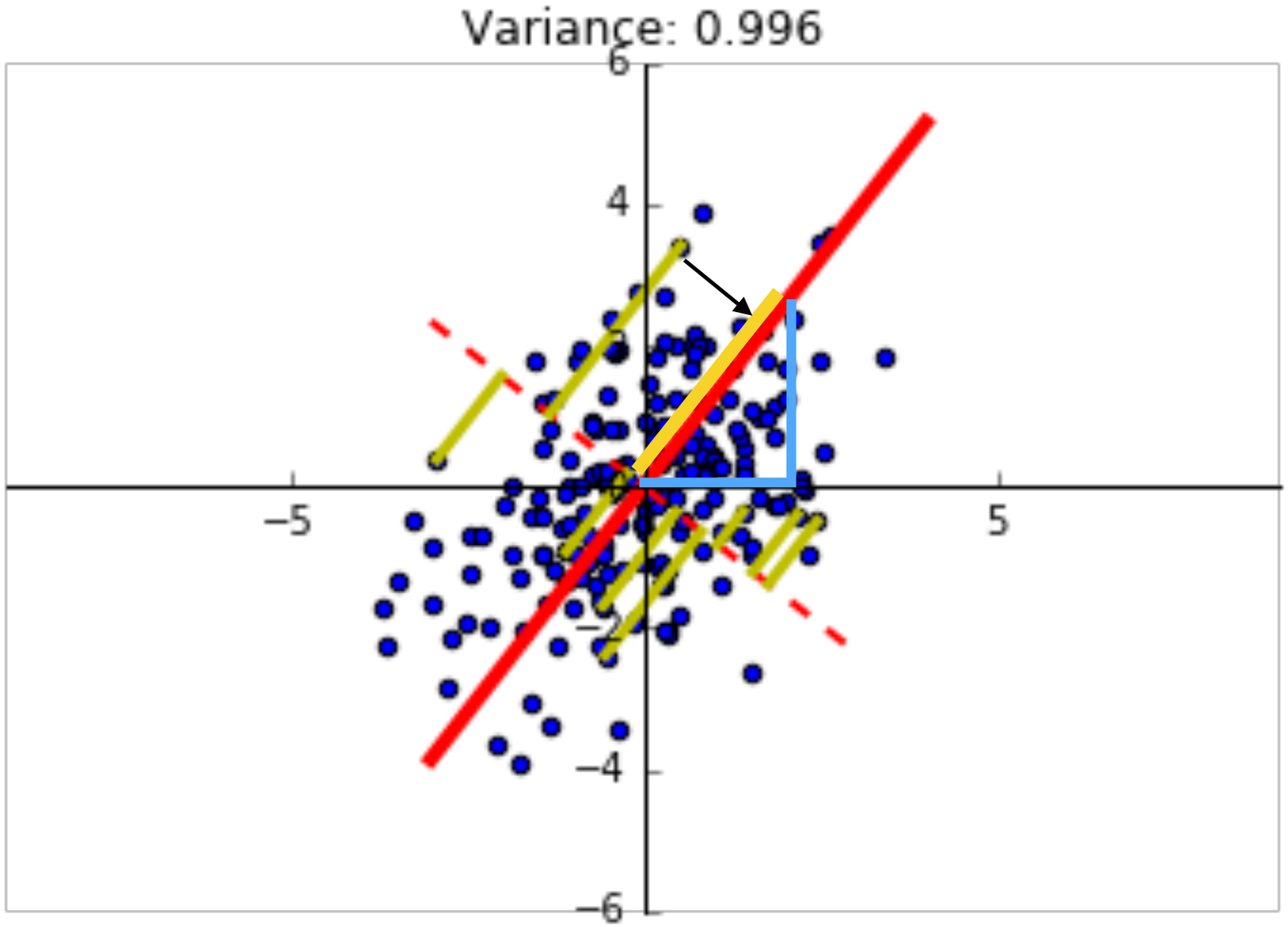
Recover Principle Axes of Data



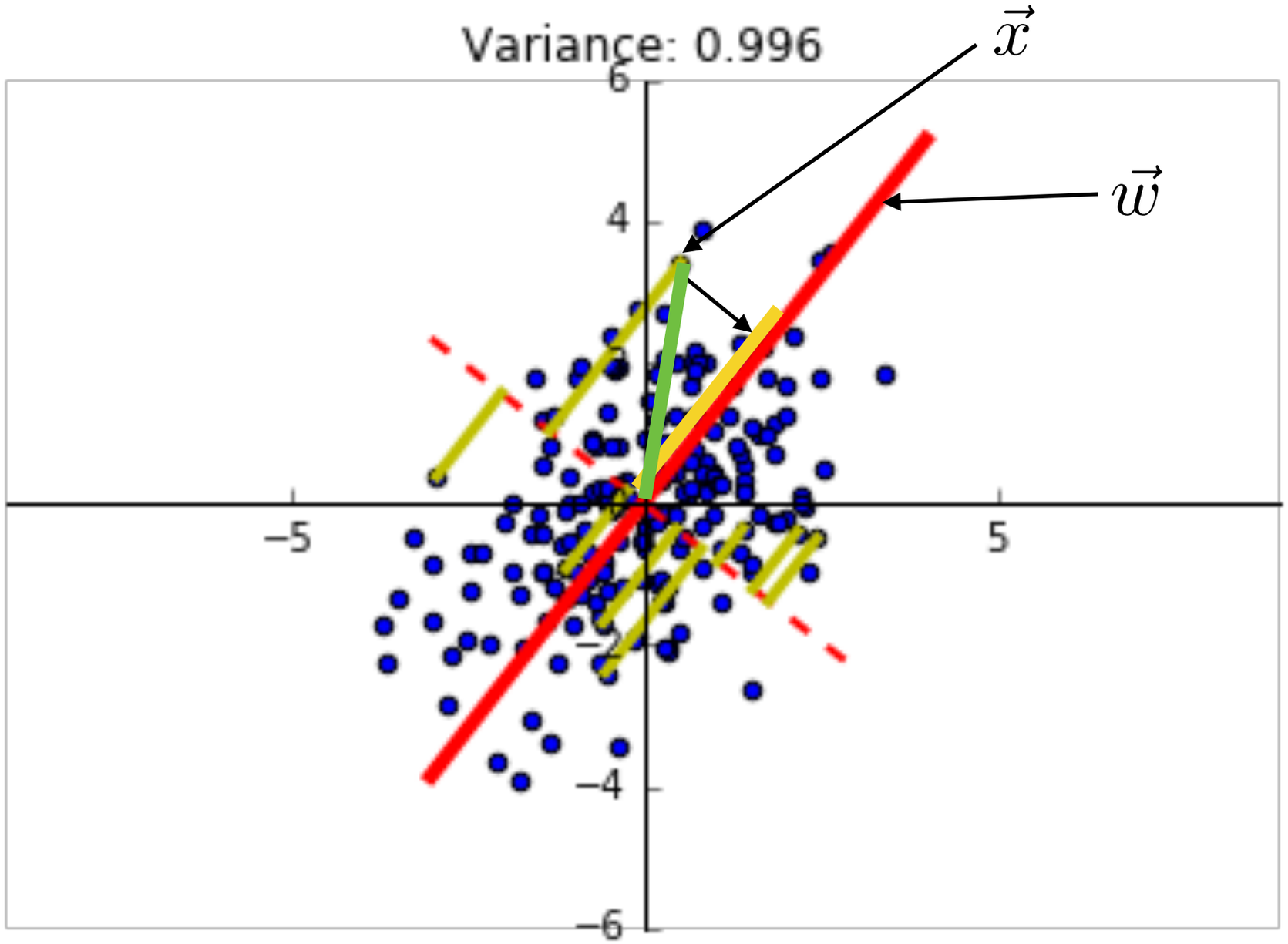
Recover Principle Axes of Data



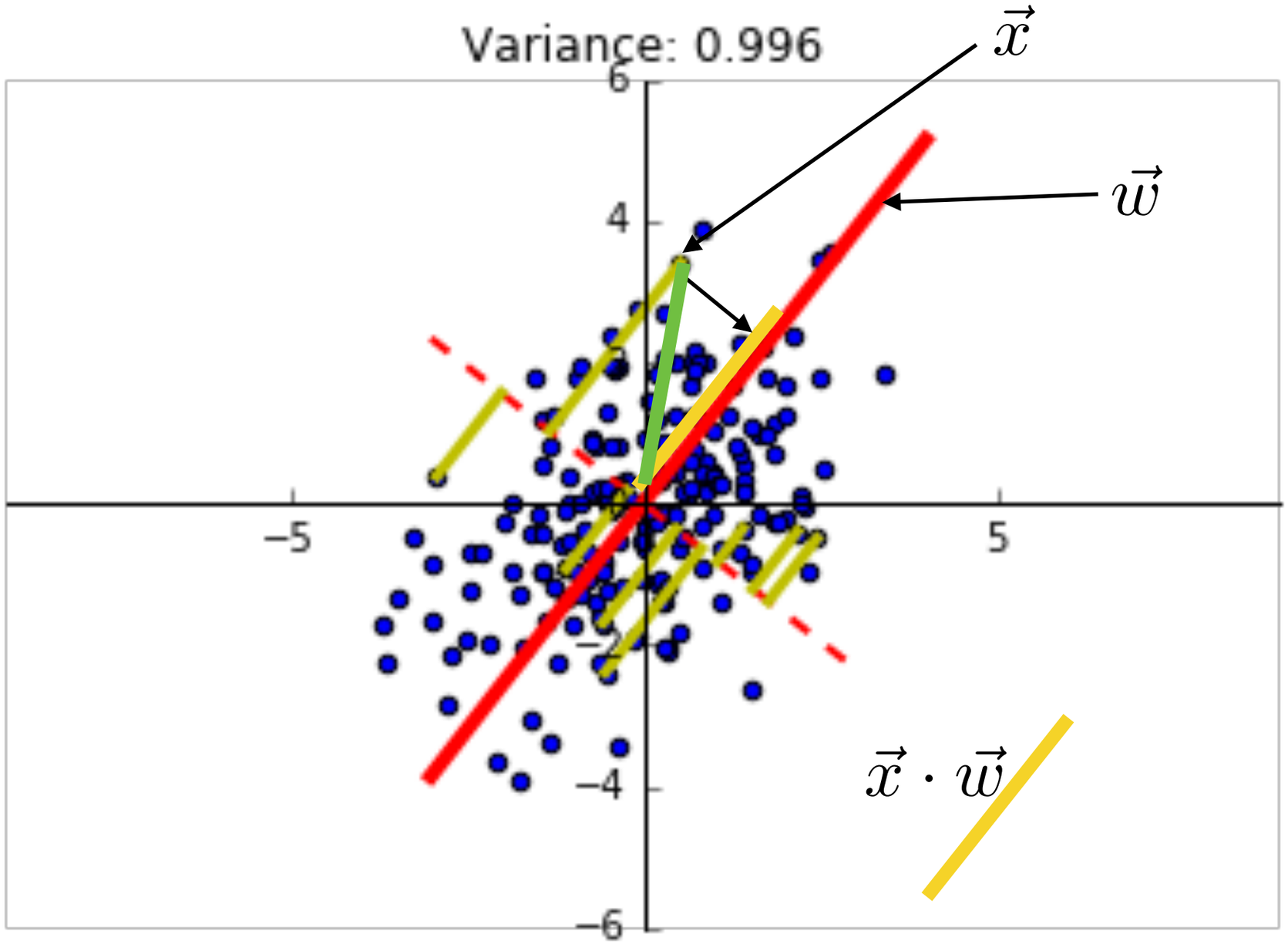
Recover Principle Axes of Data



Recover Principle Axes of Data

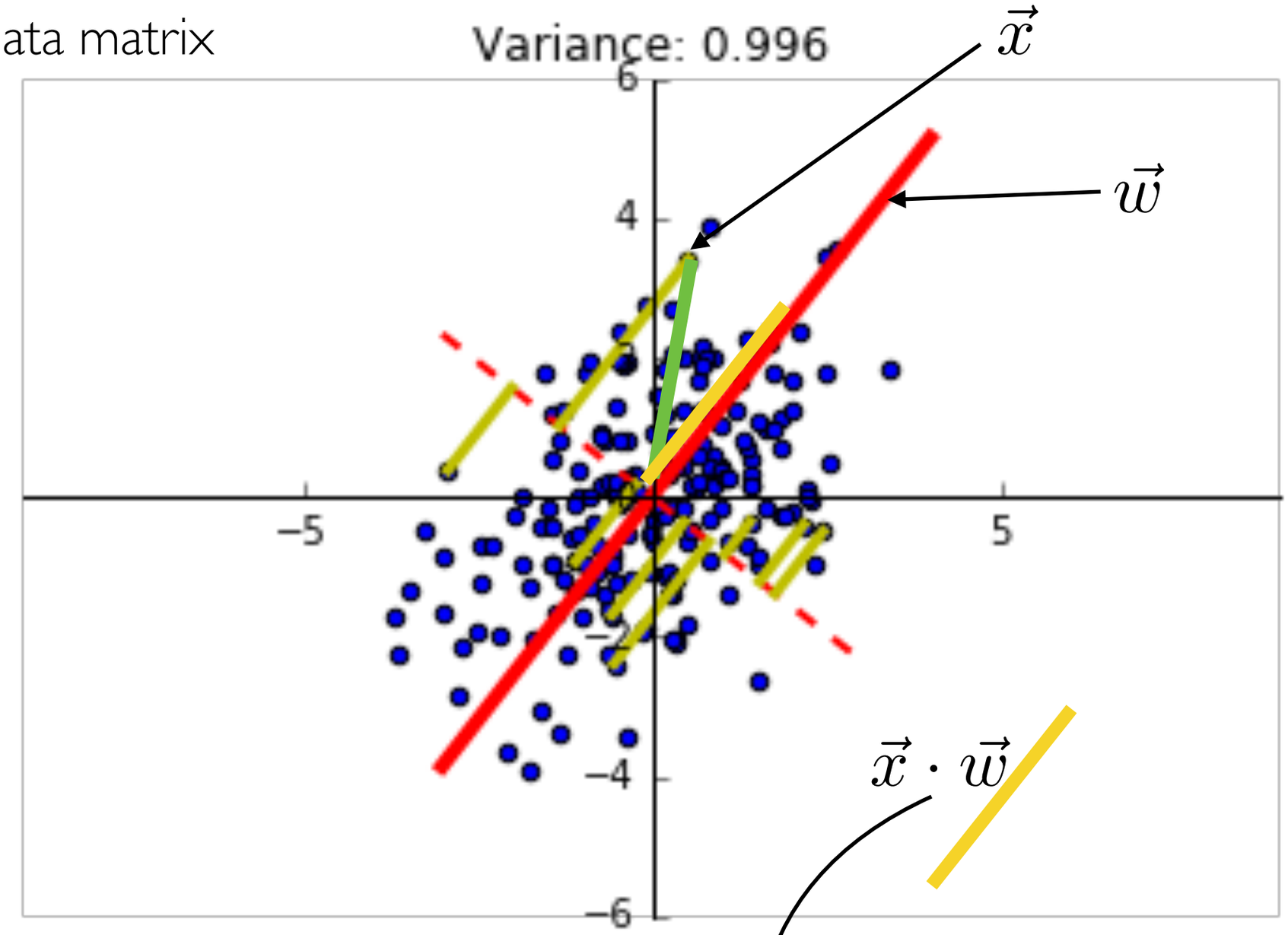


Recover Principle Axes of Data



Recover Principle Axes of Data

X = data matrix



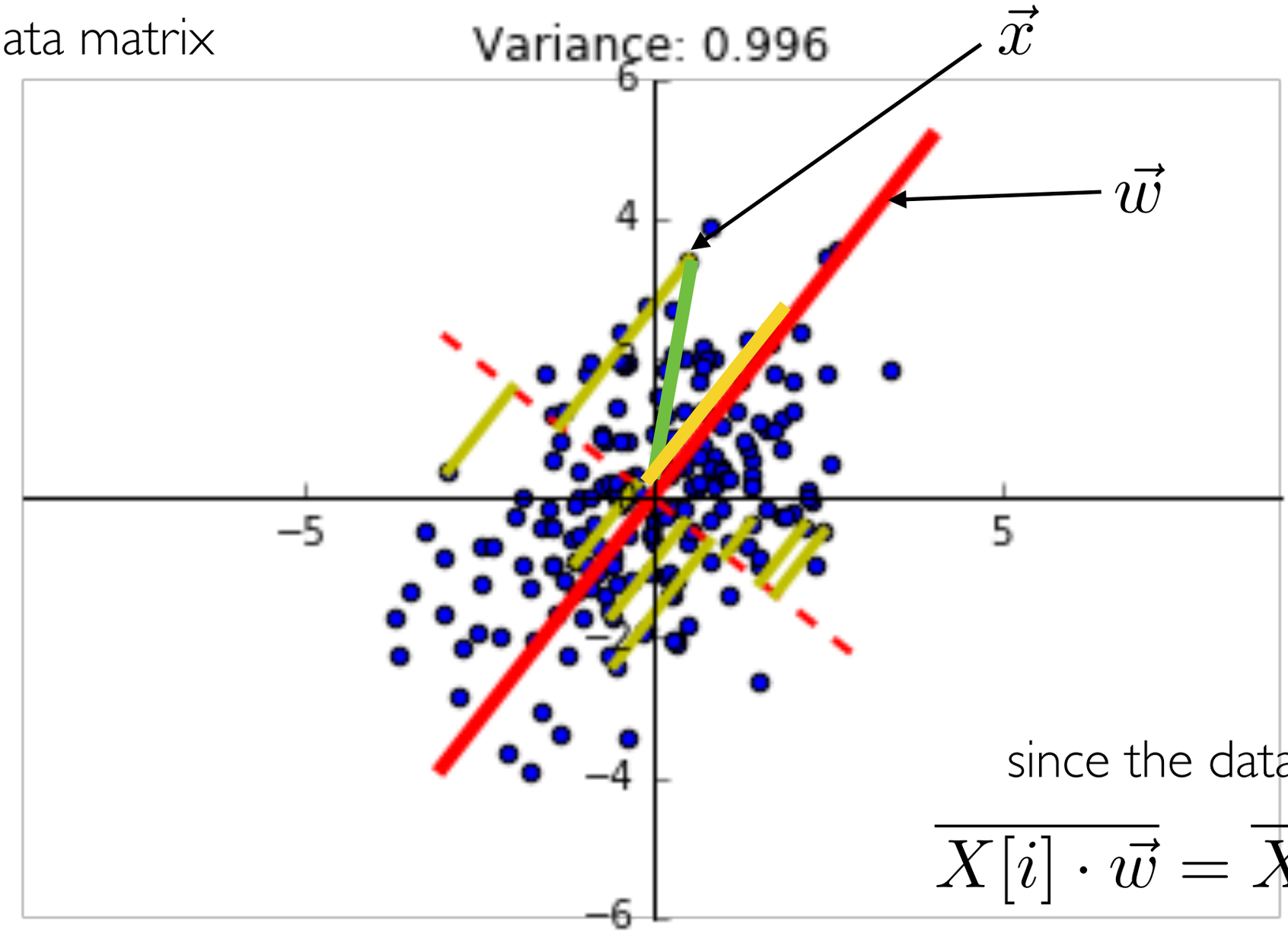
This mean is over all the data

$$var(X, \vec{w}) = \frac{1}{N} \sum_{i=1}^{N} (X[i] \cdot \vec{w} - \overline{X[i] \cdot \vec{w}})^2$$

This is a sum over the blue dots (the data)

Recover Principle Axes of Data

X = data matrix



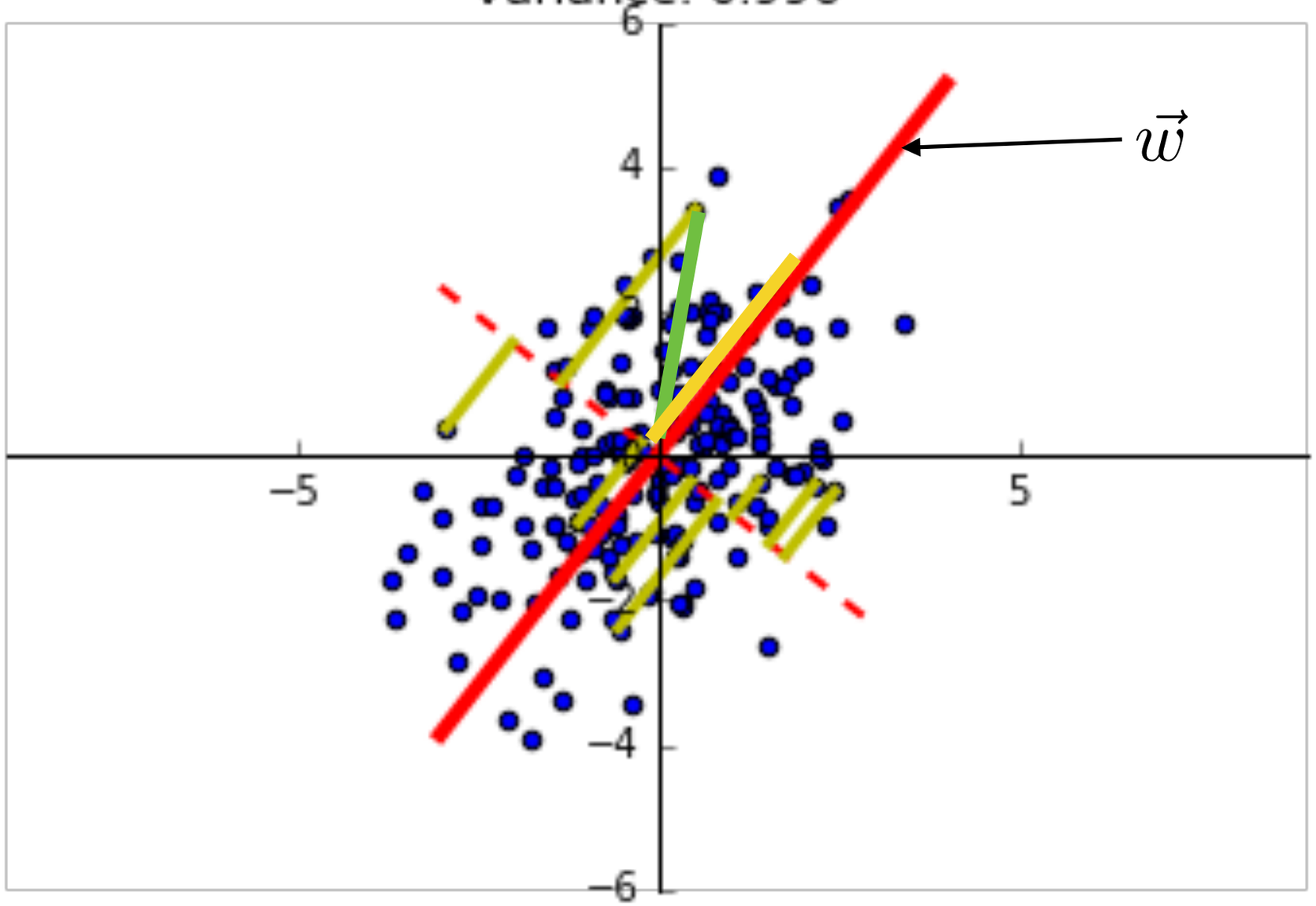
since the data is centered!

$$\overline{X[i] \cdot \vec{w}} = \overline{X[i]} \cdot \vec{w} = 0$$

$$var(X, \vec{w}) = \frac{1}{N} \sum_{i=1}^{N} (X[i] \cdot \vec{w} - \overline{X[i] \cdot \vec{w}})^2$$

Recover Principle Axes of Data

X = data matrix

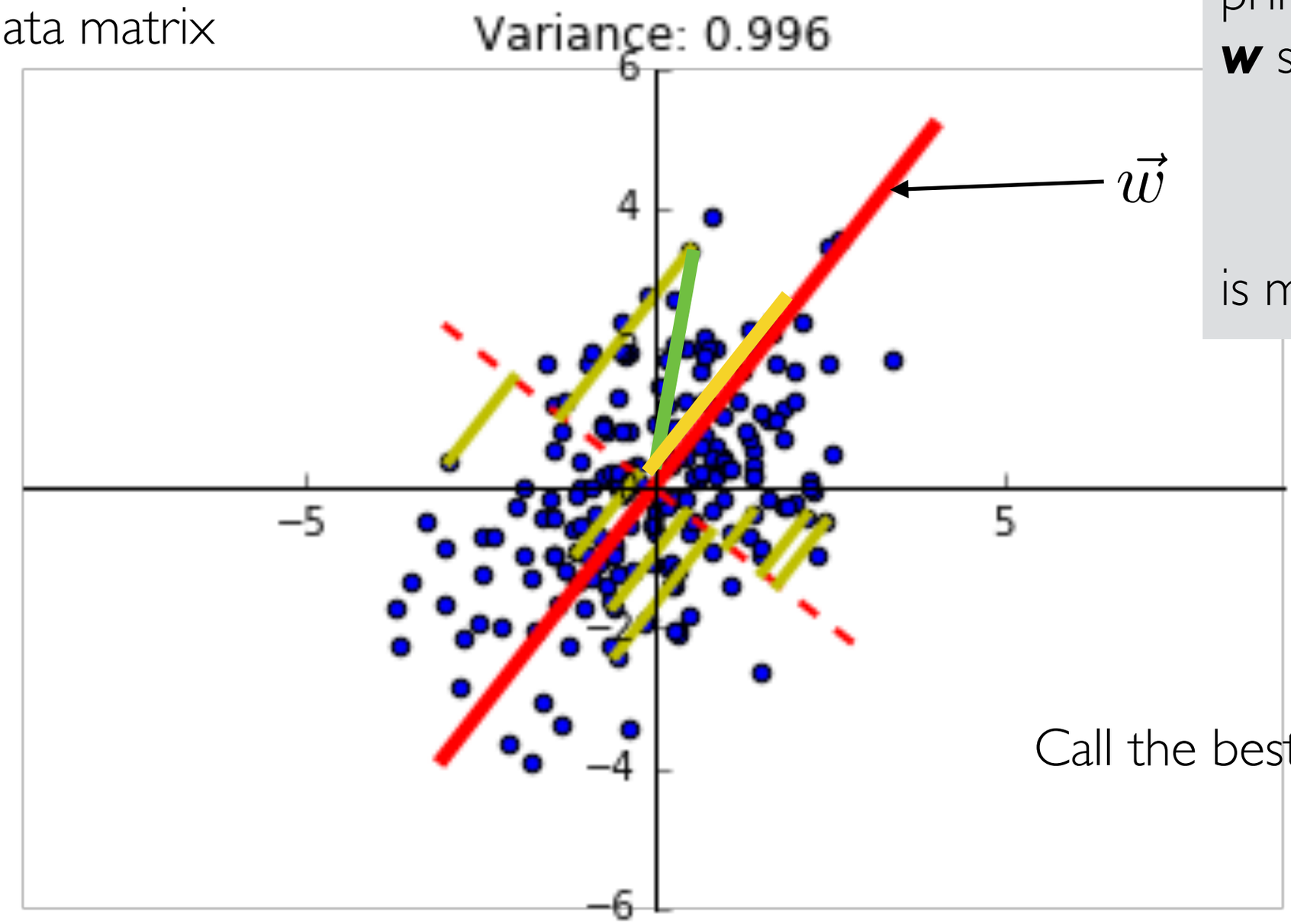


so:

$$var(X, \vec{w}) = \frac{1}{N} \sum_{i=1}^{N} (X[i] \cdot \vec{w})^2 = ||X \cdot \vec{w}||^2$$

Recover Principle Axes of Data

X = data matrix



principal axis:
w such that:

$$||X \cdot \vec{w}||^2$$

is maximized

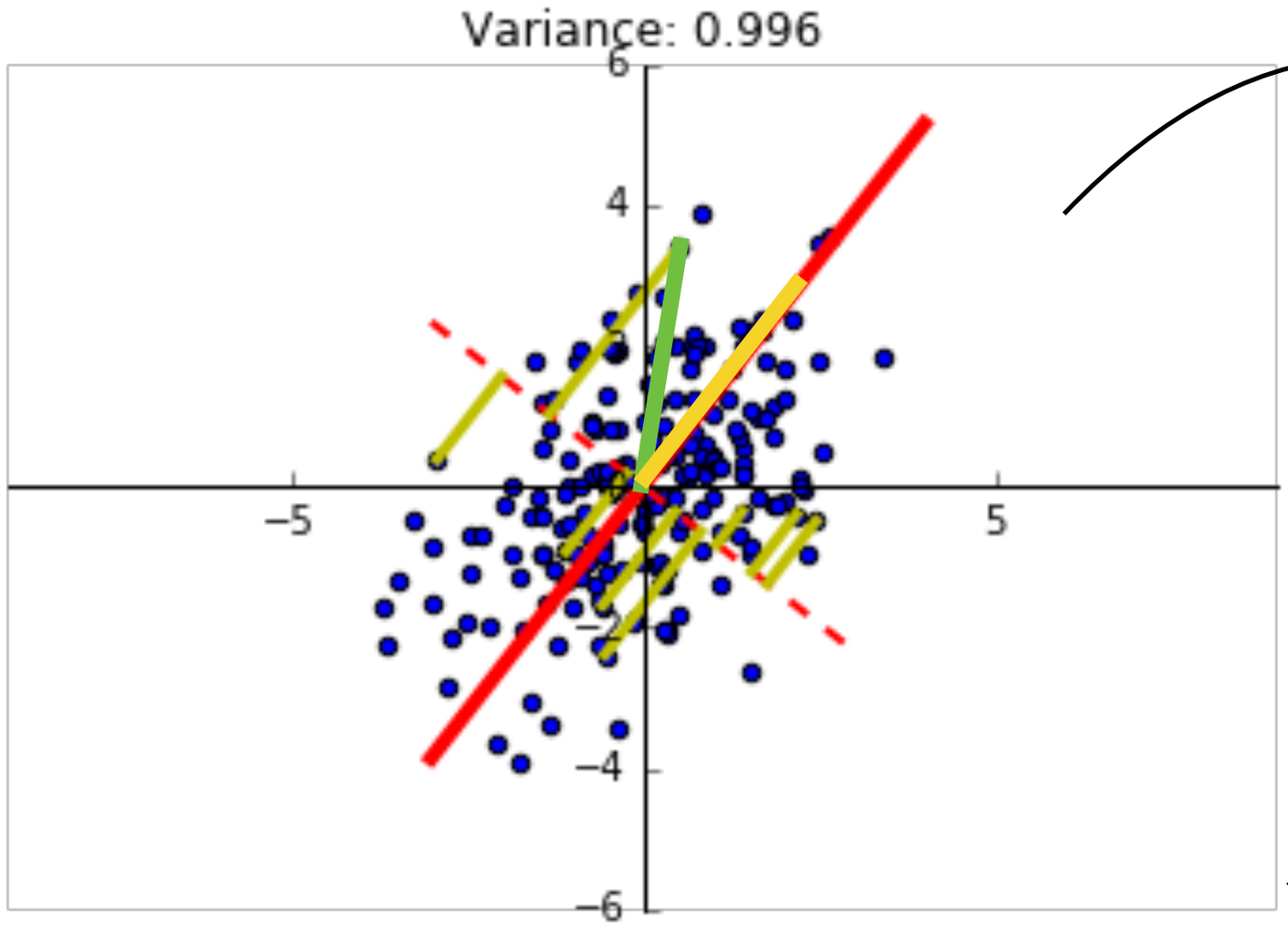
Call the best ***w_1***

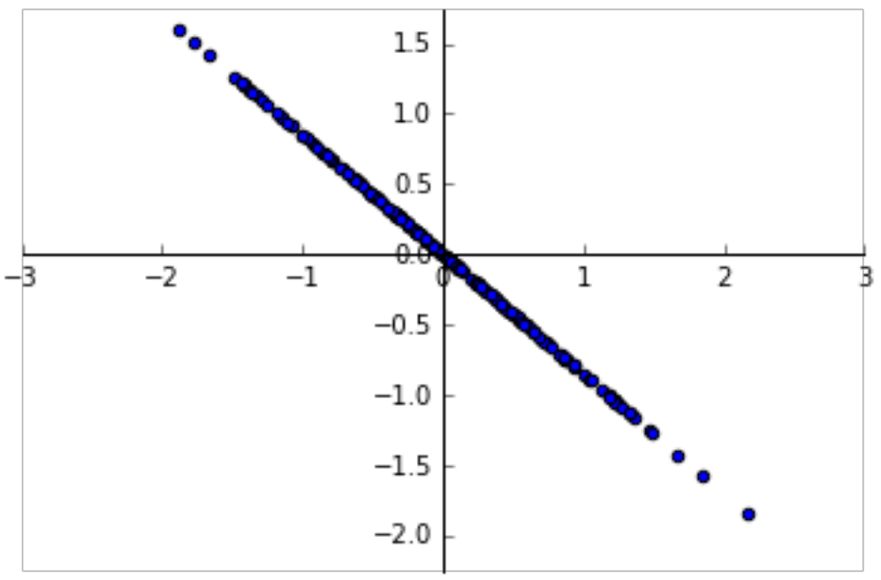
so:

$$var(X, \vec{w}) = \frac{1}{N} \sum_{i=1}^{N} (X[i] \cdot \vec{w})^2 = ||X \cdot \vec{w}||^2$$

Recover Principle Axes of Data

projecting the data onto the orthogonal hyperplane to the principal axis:



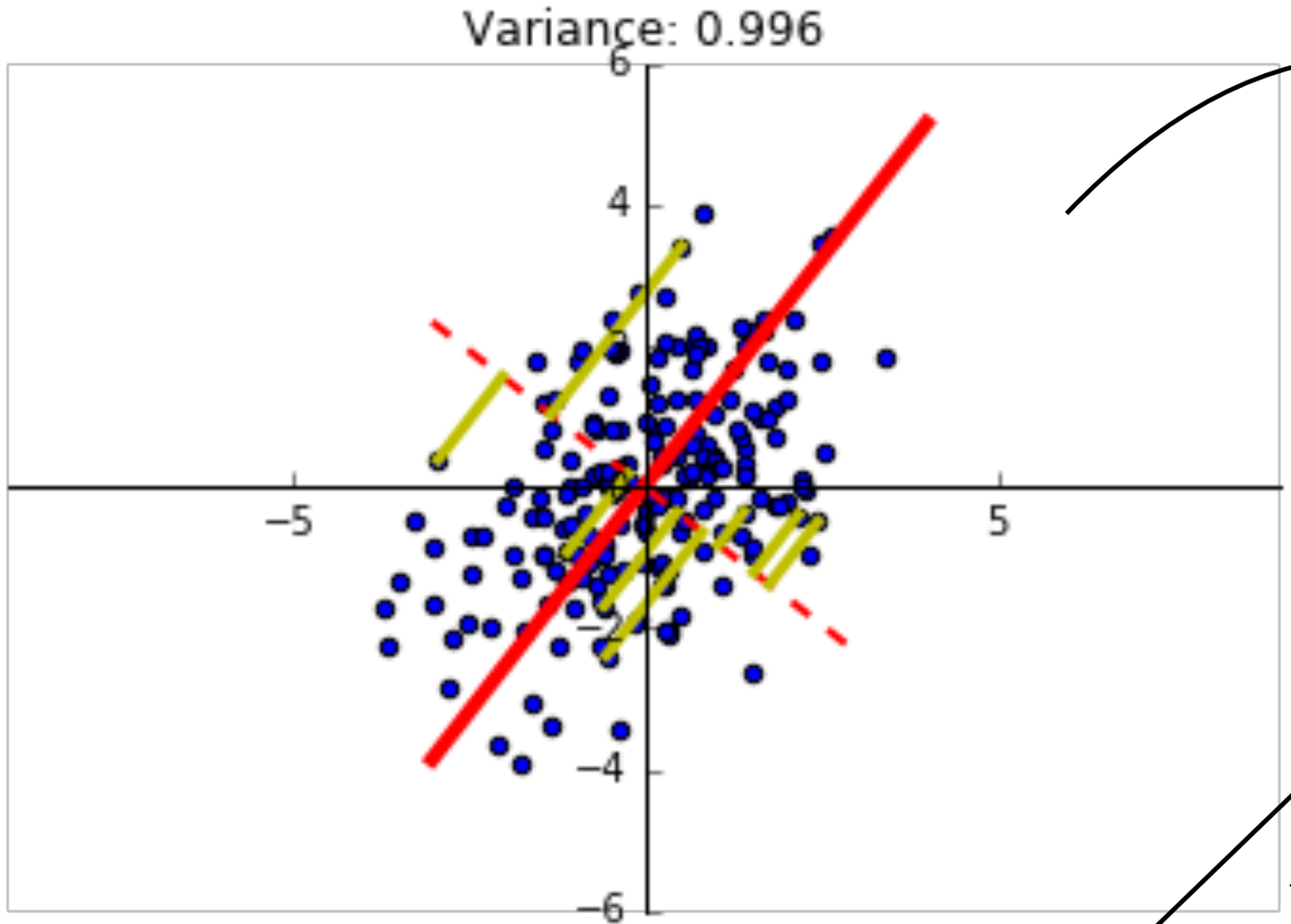


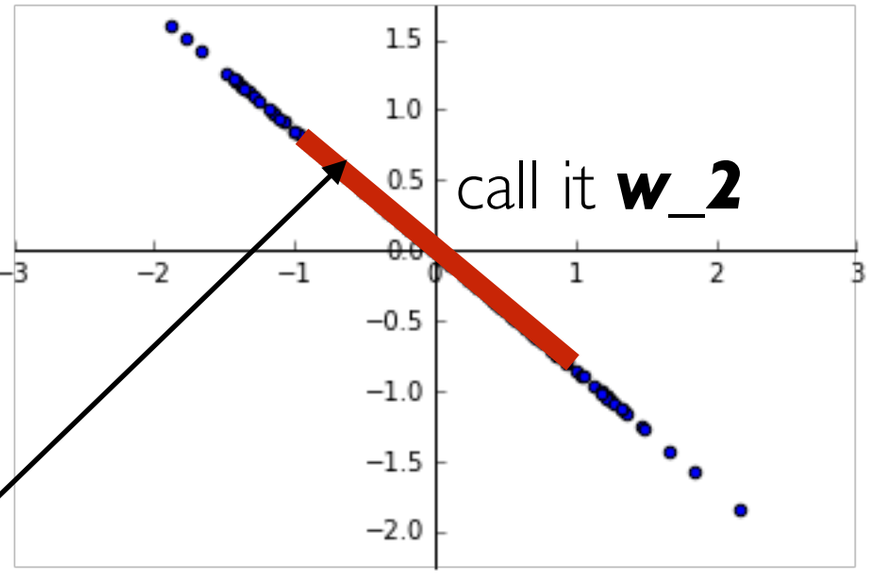
$$X_{proj} = X - (X \cdot \vec{w}_1)\vec{w}_1$$

↑ method of "deflation"

Recover Principle Axes of Data

projecting the data onto the orthogonal hyperplane to the principal axis:





call it ***w_2***

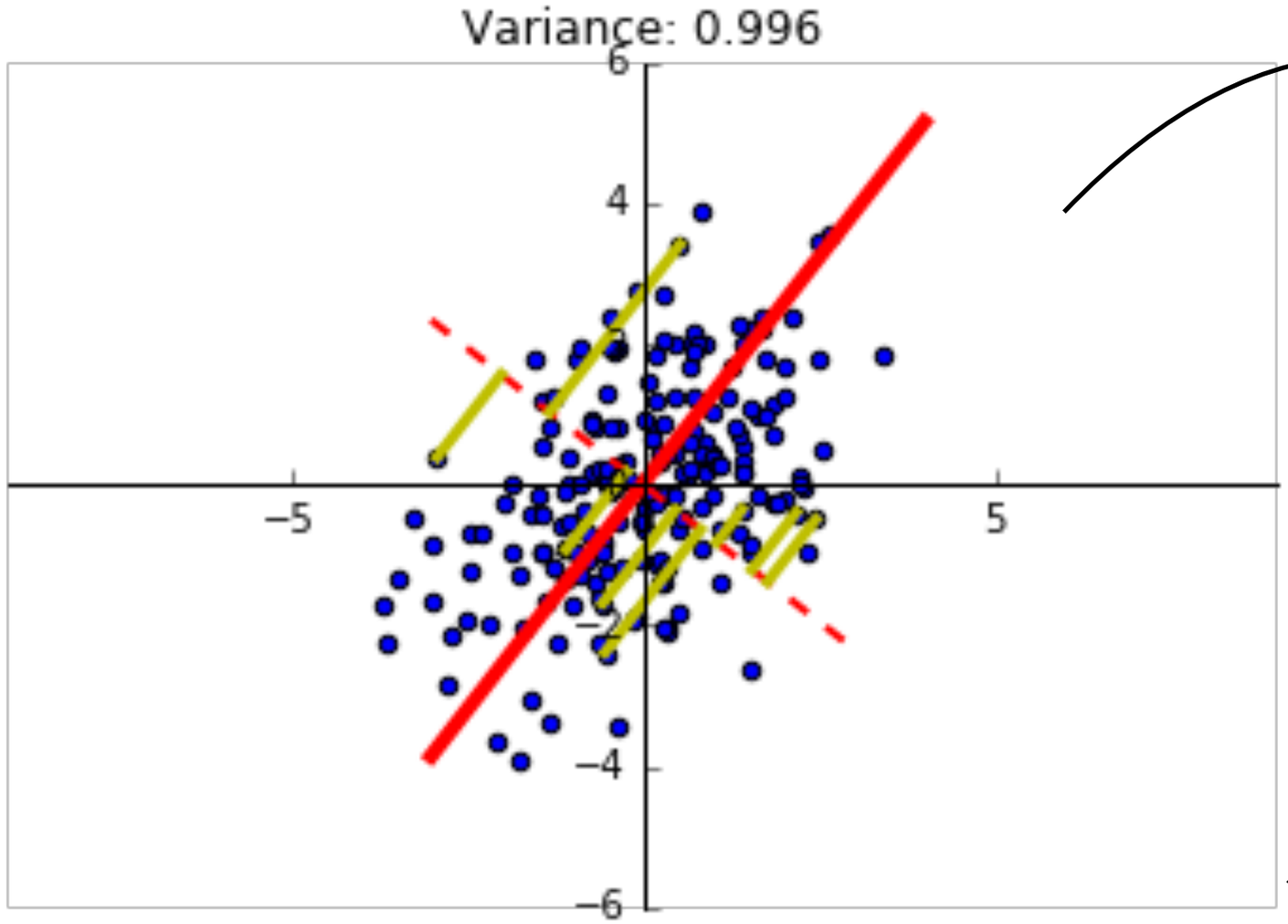
$$X_{proj} = X - (X \cdot \vec{w}_1)\vec{w}_1$$

↑ method of "deflation"

Now, do it again to find principle axis of X_{proj}

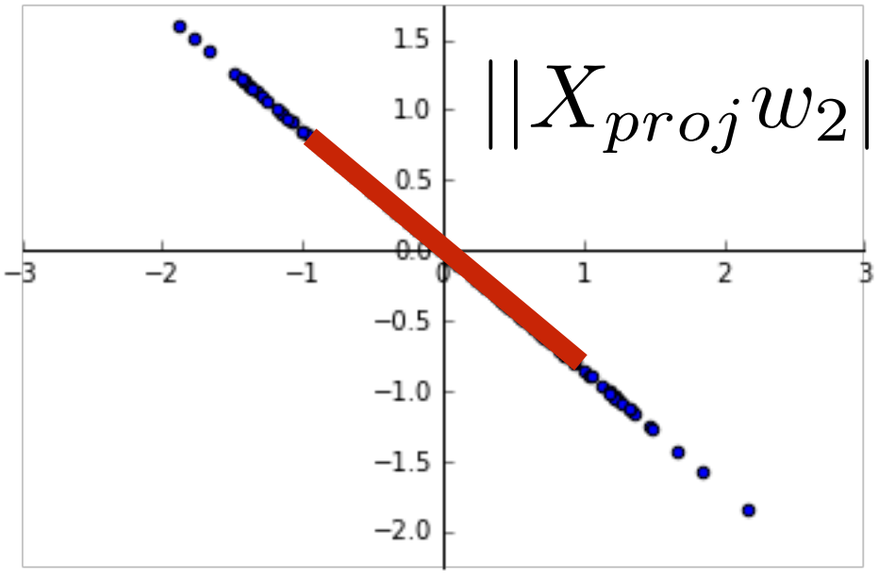
Recover Principle Axes of Data

projecting the data onto the orthogonal hyperplane to the principal axis:



w_2

maximizes



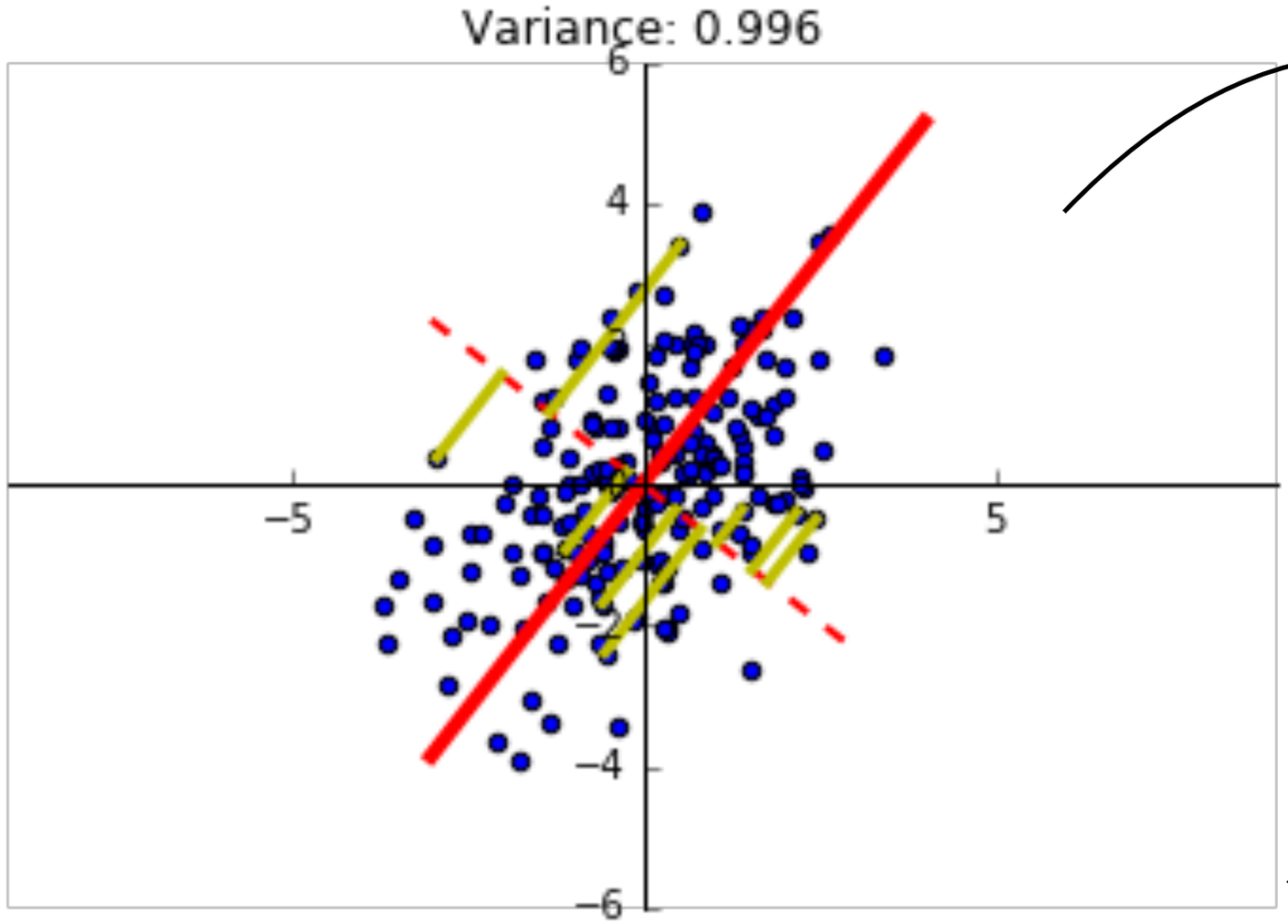
$$||X_{proj} w_2||^2$$

$$X_{proj} = X - (X \cdot \vec{w}_1)\vec{w}_1$$

↑ method of "deflation"

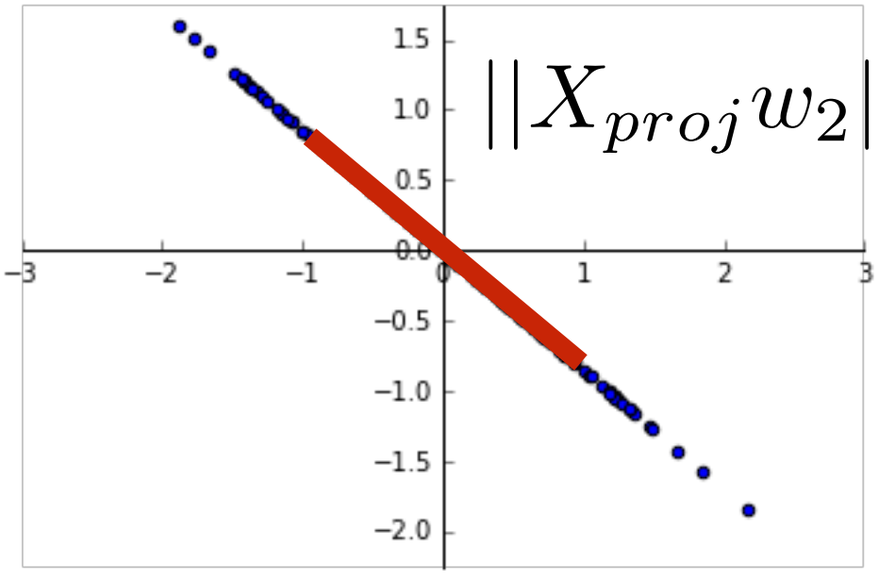
Recover Principle Axes of Data

projecting the data onto the orthogonal hyperplane to the principal axis:



w_2

maximizes



$$||X_{proj} w_2||^2$$

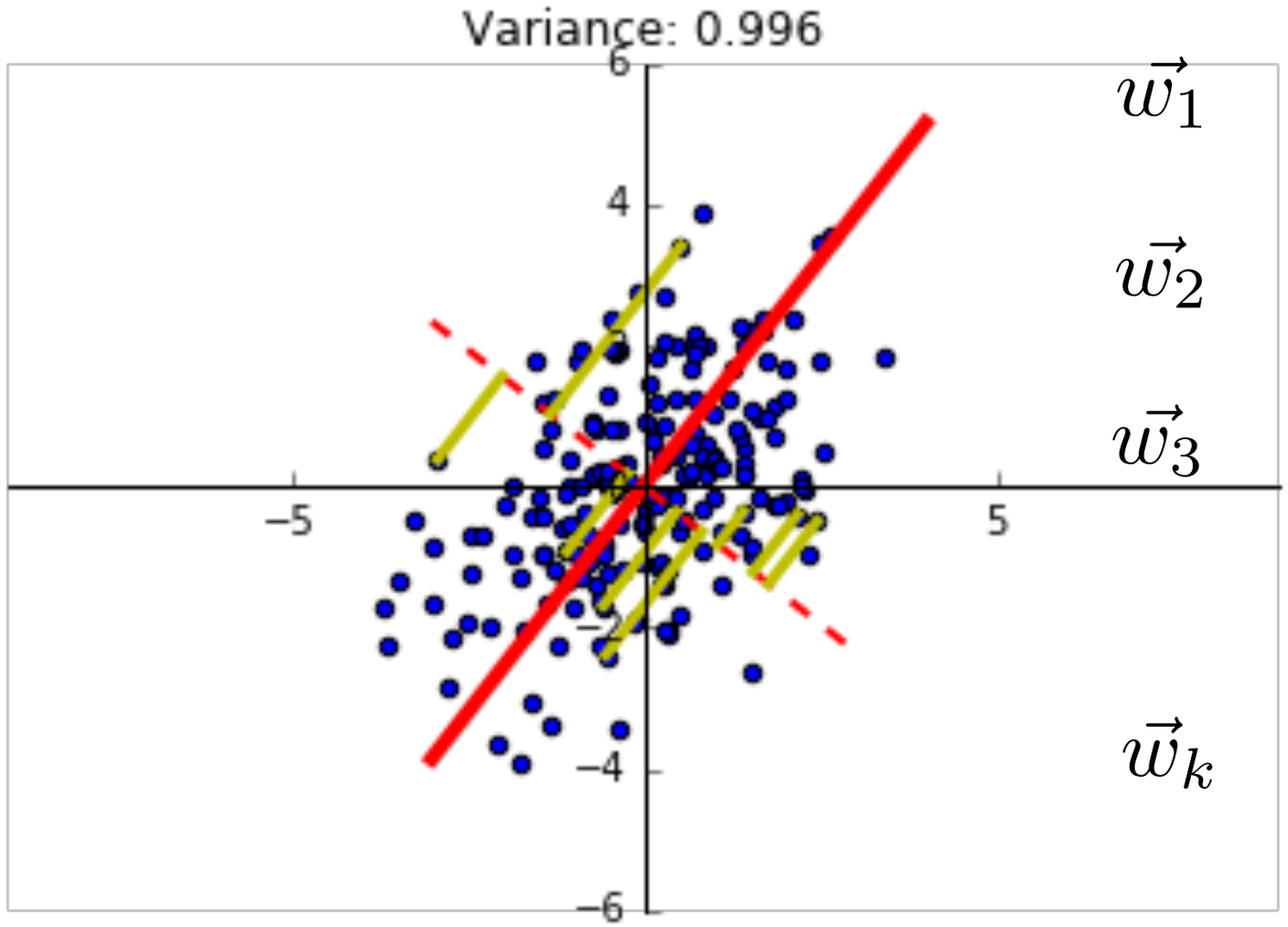
$$X_{proj} = X - (X \cdot \vec{w}_1)\vec{w}_1$$

$$X_{proj} \cdot \vec{w}_2 = (X - (X \cdot \vec{w})\vec{w}_1) \cdot w_2 = X \cdot \vec{w}_2 - (X \cdot \vec{w})(\vec{w}_1 \cdot \vec{w}_2) = X \cdot \vec{w}_2$$

0 since $\vec{w}_1 \perp \vec{w}_2$

Recover Principle Axes of Data

projecting the data onto the orthogonal hyperplane to the principal axis:



$\vec{w}_1$ principal axis of X

$\vec{w}_2$ principal axis of $X - (X \cdot \vec{w}_1)\vec{w}_1$

$\vec{w}_3$ principal axis of $X - (X \cdot \vec{w}_1)\vec{w}_1 - (X \cdot \vec{w}_2)\vec{w}_2$

.

.

.

$\vec{w}_k$ principal axis of:

$$X - \sum_{i<k} (X \cdot \vec{w}_i)\vec{w}_i$$

principal axis decomposition of X

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

$$||X\vec{w}_1||^2 =$$

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

$$||X\vec{w}_1||^2 = \vec{w}_1^T \cdot X^T \cdot X \cdot \vec{w}_1 =$$

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

$$||X\vec{w}_1||^2 = \vec{w}_1^T \cdot X^T \cdot X \cdot \vec{w}_1 = \vec{w}_1^T \cdot \text{cov} \cdot \vec{w}_1$$

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

$$||X\vec{w}_1||^2 = \vec{w}_1^T \cdot X^T \cdot X \cdot \vec{w}_1 = \vec{w}_1^T \cdot \text{cov} \cdot \vec{w}_1$$

cov is a symmetric matrix, so it has a diagonalization:

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

$$||X\vec{w}_1||^2 = \vec{w}_1^T \cdot X^T \cdot X \cdot \vec{w}_1 = \vec{w}_1^T \cdot \text{cov} \cdot \vec{w}_1$$

cov is a symmetric matrix, so it has a diagonalization:

$$\text{cov} = U^T \cdot D \cdot U$$

where: U is orthonormal: $U^T \cdot U = \text{Identity}$

D is diagonal

Recover Principle Axes of Data

Want better interpretation of $\vec{w}_1$

$$\text{cov} = X^T \cdot X$$

$$||X\vec{w}_1||^2 = \vec{w}_1^T \cdot X^T \cdot X \cdot \vec{w}_1 = \vec{w}_1^T \cdot \text{cov} \cdot \vec{w}_1$$

cov is a (real) symmetric matrix, so it has a (real-valued) diagonalization:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

write $U \cdot \vec{w}_1$ in terms of *its* components

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

write $U \cdot \vec{w}_1$ in terms of *its* components

Now:

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \vec{w}_1^T \cdot U^T \cdot \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{bmatrix} \cdot U \cdot \vec{w}_1$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

write $U \cdot \vec{w}_1$ in terms of *its* components

So:

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \vec{w}_1^T \cdot U^T \cdot \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

write $U \cdot \vec{w}_1$ in terms of *its* components

So:

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & d_n \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

write $U \cdot \vec{w}_1$ in terms of *its* components

So:

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \cdot \begin{bmatrix} d_1 b_1 \\ b_2 d_2 \\ \vdots \\ d_n b_n \end{bmatrix}$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

write $U \cdot \vec{w}_1$ in terms of *its* components

Thus:

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \cdot \begin{bmatrix} d_1 b_1 \\ b_2 d_2 \\ \vdots \\ d_n b_n \end{bmatrix} = \sum_{i=1}^{n} d_i b_i^2$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \sum_{i=1}^{n} d_i b_i^2$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \sum_{i=1}^{n} d_i b_i^2$$

maximized when: $b_1 = 1$

$$b_i = 0 \text{ for } i > 1$$

assuming: $||\vec{w}_1||^2 = \sum_i b_i^2 = 1$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \sum_{i=1}^{n} d_i b_i^2$$

maximized when: $b_1 = 1$

$$b_i = 0 \text{ for } i > 1$$

Thus:

$$\vec{w}_1 = U^T \cdot \vec{e}_1$$

where $\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

assuming: $||\vec{w}_1||^2 = \sum_i b_i^2 = 1$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \sum_{i=1}^{n} d_i b_i^2$$

maximized when: $b_1 = 1$

$$b_i = 0 \text{ for } i > 1$$

Thus:

$$\vec{w}_1 = U^T \cdot \vec{e}_1$$

where $\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

assuming: $||\vec{w}_1||^2 = \sum_i b_i^2 = 1$

$$||X \cdot w_1||^2 = \vec{w}_1^T \cdot \text{cov} \cdot \vec{w}_1 = d_1$$

Recover Principle Axes of Data

$$U \cdot \vec{w}_1 = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\vec{w}_1^T \cdot \mathbf{cov} \cdot \vec{w}_1 = \sum_{i=1}^{n} d_i b_i^2$$

maximized when: $b_1 = 1$

$$b_i = 0 \text{ for } i > 1$$

assuming: $||\vec{w}_1||^2 = \sum_i b_i^2 = 1$

And in general:

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$d_i = ||X \cdot \vec{w}_i||^2$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

$$\text{cov} \cdot \vec{w}_i =$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i = U^T \cdot (D \cdot \vec{e}_i)$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i = U^T \cdot (D \cdot \vec{e}_i) = d_i(U^T \cdot \vec{e}_i)$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

scalar "dilation"

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i = U^T \cdot (D \cdot \vec{e}_i) = d_i(U^T \cdot \vec{e}_i) = d_i \vec{w}_i$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

scalar "dilation"

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i = U^T \cdot (D \cdot \vec{e}_i) = d_i(U^T \cdot \vec{e}_i) = d_i \vec{w}_i$$

So $\boldsymbol{w_i}$, $\boldsymbol{d_i}$ are solutions of:

$$(\text{cov} - \lambda I) \cdot \vec{w} = 0$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

scalar "dilation"

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i = U^T \cdot (D \cdot \vec{e}_i) = d_i(U^T \cdot \vec{e}_i) = d_i \vec{w}_i$$

identity matrix of the right shape

So $\boldsymbol{w_i}$, $\boldsymbol{d_i}$ are solutions of:

$$(\text{cov} - \lambda I) \cdot \vec{w} = 0$$

Recover Principle Axes of Data

Starting with:

$$\text{cov} = U^T \cdot \begin{bmatrix} d_1 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{bmatrix} \cdot U$$

$$\vec{w}_i = U^T \cdot \vec{e}_i$$

$$\vec{e}_i = U \cdot \vec{w}_i$$

scalar "dilation"

$$\text{cov} \cdot \vec{w}_i = (U^T \cdot D \cdot U) \cdot \vec{w}_i = U^T \cdot (D \cdot \vec{e}_i) = d_i (U^T \cdot \vec{e}_i) = d_i \vec{w}_i$$

So $\boldsymbol{w_i}$, $\boldsymbol{d_i}$ are solutions of:

identity matrix of the right shape

$$(\text{cov} - \lambda I) \cdot \vec{w} = 0$$

Eigenvalues ($\boldsymbol{d_i}$'s) of **cov**

Eigenvectors ($\boldsymbol{w_i}$'s) of **cov**

Recover Principle Axes of Data

How to actually find w_1 for any matrix $\boldsymbol{A}$?

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) So: $A \cdot \vec{x} = \sum_i a_i (A \cdot \vec{w}_i) = \sum_i d_i a_i \vec{w}_i$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} =$$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i)^2}}$$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i)^2}} = \frac{d_1^k (a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots)}{d_1^k \sqrt{a_1^2 + \left(\frac{d_2}{d_1}\right)^k a_2^2 + \ldots}}$$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i)^2}} = \frac{d_1^k (a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots)}{d_1^k \sqrt{a_1^2 + \left(\frac{d_2}{d_1}\right)^k a_2^2 + \ldots}}$$

5) But:

$$\left(\frac{d_i}{d_1}\right)^k \to 0 \text{ as } k \to \infty$$

for i > 1

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i)^2}} = \frac{d_1^k (a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots)}{d_1^k \sqrt{a_1^2 + \left(\frac{d_2}{d_1}\right)^k a_2^2 + \ldots}}$$

(the $\left(\frac{d_2}{d_1}\right)^k$ term → 0; the $\left(\frac{d_2}{d_1}\right)^k$ term → 0)

5) But:

$$\left(\frac{d_i}{d_1}\right)^k \to 0 \text{ as } k \to \infty \qquad = \frac{d_1^k a_1 \vec{w}_1}{d_1^k \sqrt{a_1^2}} = \vec{w}_1$$

for i > 1

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i \ldots}} = \frac{d_1^k \left(a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots\right)}{\sqrt{\ldots a_2^2}}$$

Thus no matter what the initial guess for x is … this quantity

… converges to the first eigenvector:

5) But:

$$\left(\frac{d_i}{d_1}\right)^k \to 0 \text{ as } k \to \infty$$

for i > 1

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum a_i \vec{w}_i$

3) In general: [illegible]

This simple algorithm for finding the principle eigenvector is called the Power Method

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i \text{[illegible]}}} = \frac{d_1^k (a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots)}{\text{[illegible]}}$$

Thus no matter what the initial guess for x is … this quantity

… converges to the principal axis:

5) But:

$$\left(\frac{d_i}{d_1}\right)^k \rightarrow 0 \text{ as } k \rightarrow \infty$$

for i > 1

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

Recover Principle Axes of Data

*almost all

How to actually find w_1 for any matrix ***A***?

Separate consequence: ***All**** linear iterative processes converge to a single dimension — the principal axis of the single-step process.

1) Start with ran

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum a_i \vec{w}_i$

This simple algorithm for finding the principle eigenvector is called the Power Method

3) In general:

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i \ldots}} = \frac{d_1^k \left(a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots\right)}{\sqrt{\ldots}}$$

Thus no matter what the initial guess for x is … this quantity

… converges to the principal axis:

5) But:

$$\left(\frac{d_i}{d_1}\right)^k \to 0 \text{ as } k \to \infty$$

for i > 1

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

Recover Principle Axes of Data

We've seen that for all ***x***:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

When does this fail, actually?

Recover Principle Axes of Data

We've seen that for all $\boldsymbol{x}^*$: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

Recover Principle Axes of Data

How to actually find w_1 for any matrix ***A***?

1) Start with random $\vec{x}$ (non-zero)

2) Write $\vec{x}$ in an eigenbasis: $\vec{x} = \sum_i a_i \vec{w}_i$

3) In general: $A^k \cdot \vec{x} = \sum_i a_i (A^k \cdot \vec{w}_i) = \sum_i d_i^k a_i \vec{w}_i$

4) Normalizing:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} = \frac{\sum_i d_i^k a_i \vec{w}_i}{\sqrt{\sum_i (d_i^k a_i)^2}} = \frac{d_1^k (a_1 \vec{w}_1 + \left(\frac{d_2}{d_1}\right)^k a_2 \vec{w}_2 + \ldots)}{d_1^k \sqrt{a_1^2 + \left(\frac{d_2}{d_1}\right)^k a_2^2 + \ldots}}$$

5) But:

$$\left(\frac{d_i}{d_1}\right)^k \to 0 \text{ as } k \to \infty$$

for i > 1

6) Thus:

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

$$A \cdot \left(\sum_{i=2}^{n} a_i \vec{w}_i \right)$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

$$A \cdot \left(\sum_{i=2}^{n} a_i \vec{w}_i \right) = \sum_{i=2}^{n} a_i A \cdot \vec{w}_i$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

$$A \cdot \left(\sum_{i=2}^{n} a_i \vec{w}_i \right) = \sum_{i=2}^{n} a_i A \cdot \vec{w}_i = \sum_{i=2}^{n} a_i d_i \vec{w}_i$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

$$A \cdot \left(\sum_{i=2}^{n} a_i \vec{w}_i \right) = \sum_{i=2}^{n} a_i A \cdot \vec{w}_i = \sum_{i=2}^{n} a_i d_i \vec{w}_i \in \text{span}(\vec{w}_1)^\perp$$

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

$$A \cdot \left(\sum_{i=2}^{n} a_i \vec{w}_i \right) = \sum_{i=2}^{n} a_i A \cdot \vec{w}_i = \sum_{i=2}^{n} a_i d_i \vec{w}_i \in \text{span}(\vec{w}_1)^\perp$$

Why is this OK though basically?

Recover Principle Axes of Data

We've seen that for all ***x****: *almost all

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \to \vec{w}_1 \text{ as } k \to \infty$$

When does this fail, actually?

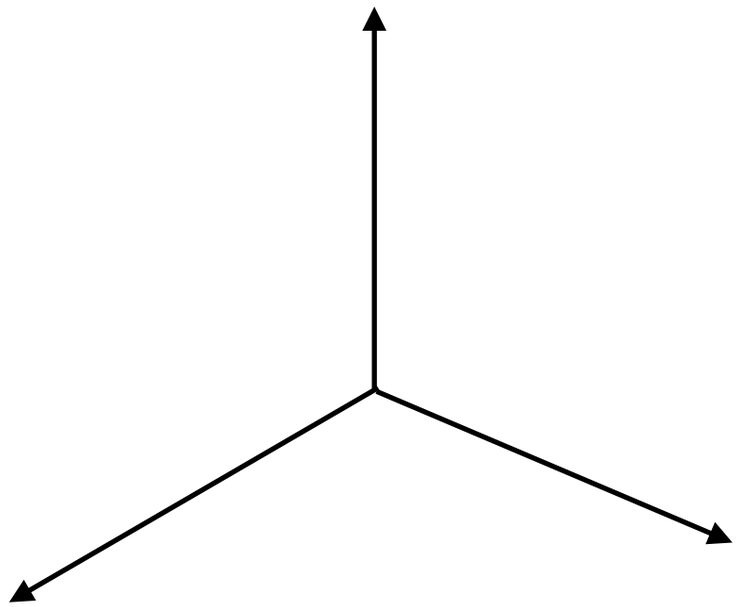
$$x \in \text{span}(\vec{w}_2, \vec{w}_3, \ldots, \vec{w}_n) = \text{span}(\vec{w}_1)^\perp$$

$$A \cdot \left(\sum_{i=2}^{n} a_i \vec{w}_i \right) = \sum_{i=2}^{n} a_i A \cdot \vec{w}_i = \sum_{i=2}^{n} a_i d_i \vec{w}_i \in \text{span}(\vec{w}_1)^\perp$$

Why is this OK though basically? Because the problematic set is measure 0.

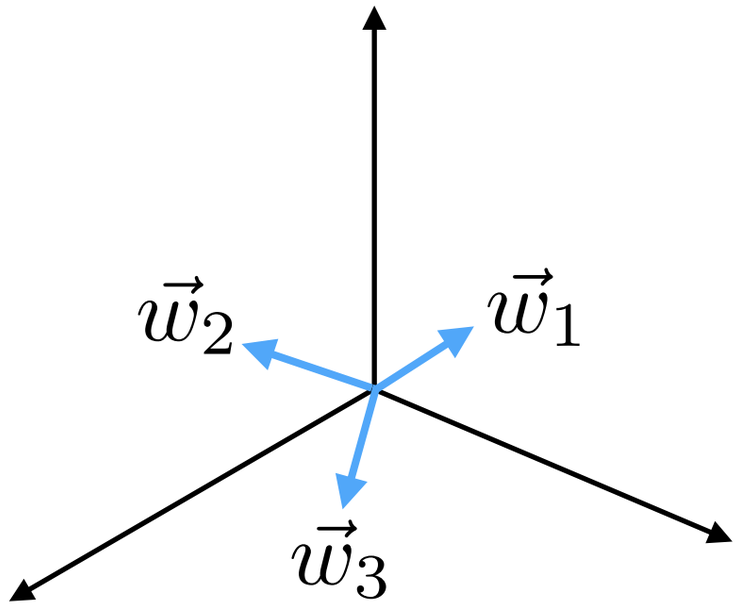
Recover Principle Axes of Data

Because the problematic set is measure 0.



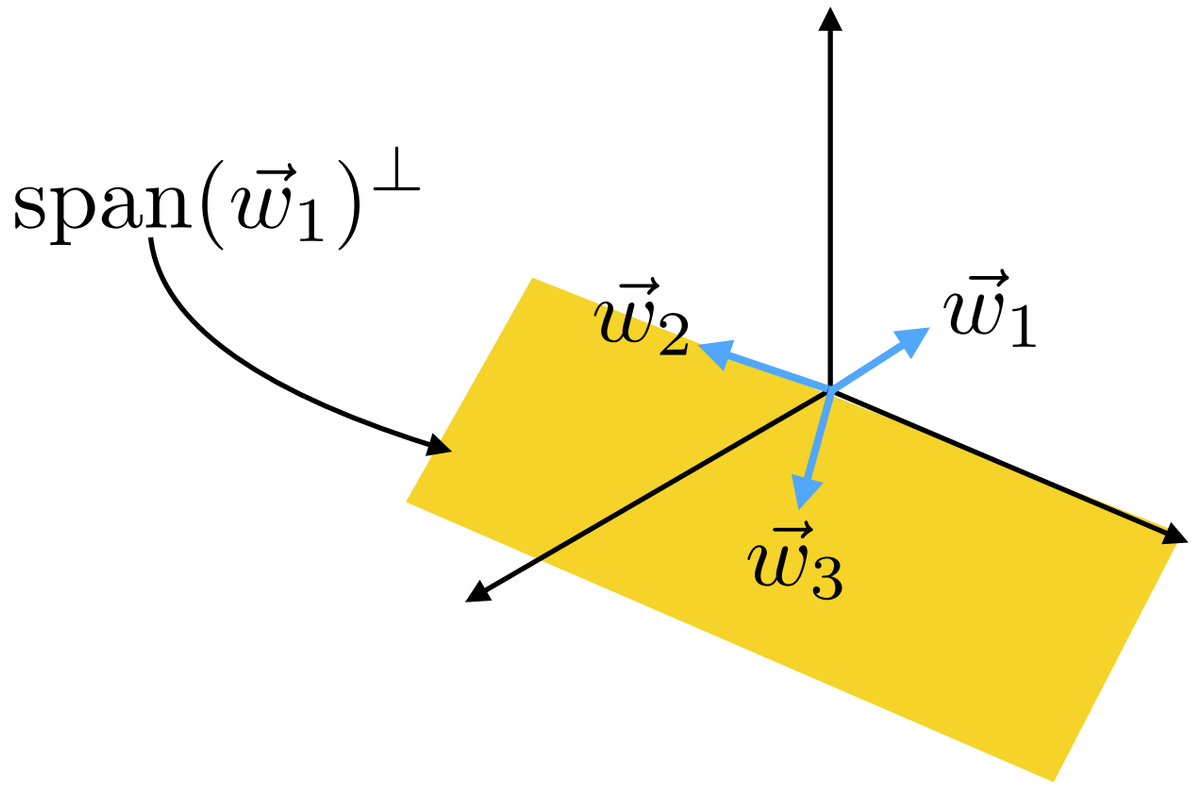
Recover Principle Axes of Data

Because the problematic set is measure 0.



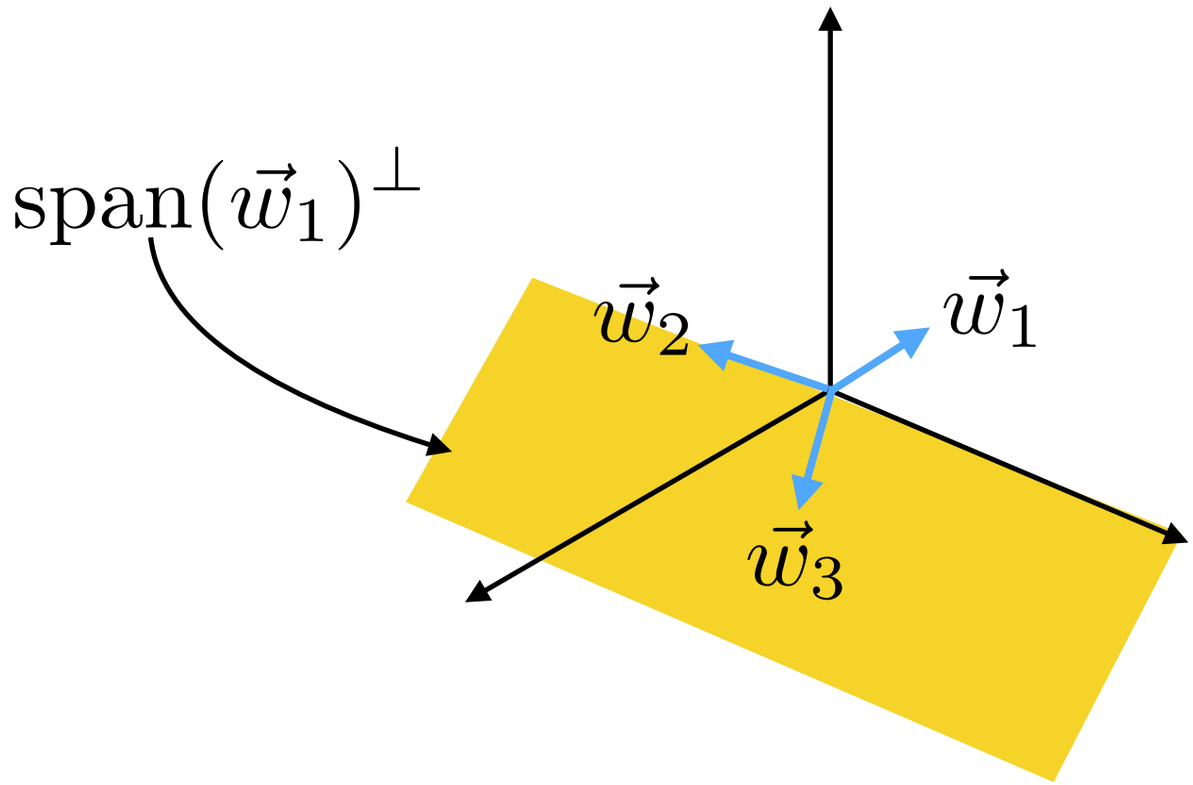
Recover Principle Axes of Data

Because the problematic set is measure 0.



Recover Principle Axes of Data

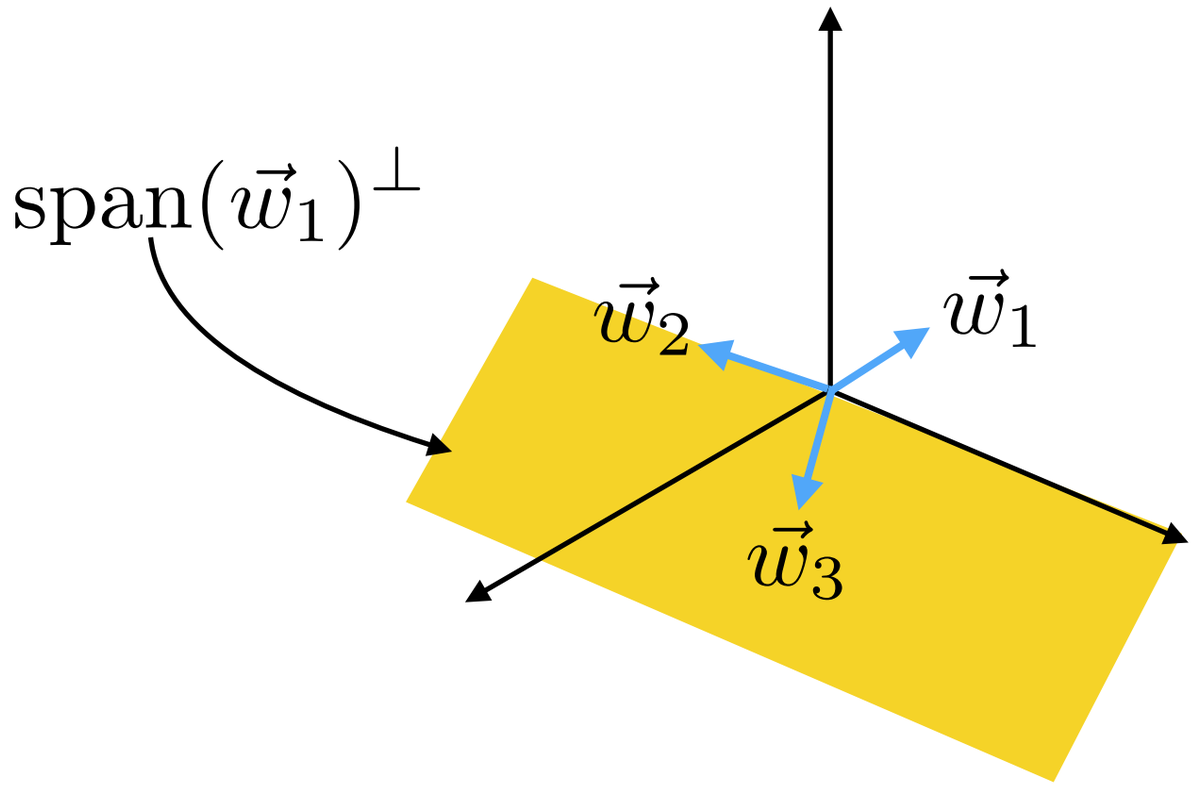
Because the problematic set is measure 0.



Almost any random choice of $\boldsymbol{x}$ will not lie in this plane…

Recover Principle Axes of Data

Because the problematic set is measure 0.



Almost any random choice of ***x*** will not lie in this plane…

… of course, deflating down to $x - (x \cdot \vec{w}_1)w_1$ ***will*** lie in the plane…

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$A \cdot \vec{w}_1 = d_1 \vec{w}_1 \quad \text{(by definition of eigenvector)}$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$A \cdot \vec{w}_1 = d_1 \vec{w}_1 \quad \text{(by definition of eigenvector)}$$

So:

$$\vec{w}_1^T \cdot A \cdot \vec{w}_1 = \vec{w}_1^T \cdot d_1 \vec{w}_1 = d_1 ||\vec{w}_1||^2$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$A \cdot \vec{w}_1 = d_1 \vec{w}_1 \quad \text{(by definition of eigenvector)}$$

So:

$$\vec{w}_1^T \cdot A \cdot \vec{w}_1 = \vec{w}_1^T \cdot d_1 \vec{w}_1 = d_1 ||\vec{w}_1||^2$$

Solving for d_1:

$$d_1 = \frac{\vec{w}_1^T \cdot A \cdot \vec{w}_1}{||\vec{w}_1||^2} = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

since w_1 is of unit norm

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

$$(A - (A \cdot \vec{w}_1) w_1^T) \cdot \vec{w}_1$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

$$(A - (A \cdot \vec{w}_1) w_1^T) \cdot \vec{w}_1 = A \cdot \vec{w}_1 - d_1 \vec{w}_1 \cdot \vec{w}_1^T \cdot \vec{w}_1$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

$$(A - (A \cdot \vec{w}_1) w_1^T) \cdot \vec{w}_1 = A \cdot \vec{w}_1 - d_1 \vec{w}_1 \cdot \boxed{\vec{w}_1^T \cdot \vec{w}_1}$$

??

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

$$(A - (A \cdot \vec{w}_1) w_1^T) \cdot \vec{w}_1 = A \cdot \vec{w}_1 - d_1 \vec{w}_1 \cdot \boxed{\vec{w}_1^T \cdot \vec{w}_1}$$

$= 1$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

$$(A - (A \cdot \vec{w}_1) w_1^T) \cdot \vec{w}_1 = A \cdot \vec{w}_1 - d_1 \vec{w}_1 \cdot \boxed{\vec{w}_1^T \cdot \vec{w}_1} \quad = 1$$

$$= d_1 \vec{w}_1 - d_1 \vec{w}_1 = 0$$

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A - (A \cdot \vec{w}_1) w_1^T = A - d_1 \vec{w}_1 \cdot \vec{w}_1^T$$

$$(A - (A \cdot \vec{w}_1) w_1^T) \cdot \vec{w}_1 = A \cdot \vec{w}_1 - d_1 \vec{w}_1 \cdot \boxed{\vec{w}_1^T \cdot \vec{w}_1} \quad = 1$$

$$= d_1 \vec{w}_1 - d_1 \vec{w}_1 = 0$$

$A_2 \equiv A - (A \cdot \vec{w}_1) w_1^T$ has "smooshed" top eigenvector / eigenvalue to 0

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A_2 = A - (A \cdot \vec{w}_1) w_1^T$$ ← method of "deflation"

Recover Principle Axes of Data

How to actually find w_1?

$$\frac{A^k \cdot \vec{x}}{||A^k \vec{x}||} \rightarrow \vec{w}_1 \text{ as } k \rightarrow \infty$$

How to actually find d_1?

$$d_1 = \vec{w}_1^T \cdot A \cdot \vec{w}_1$$

having found w_1 via above method

How to keep going to find next eigenvalues / eigenvectors?

$$A_2 = A - (A \cdot \vec{w}_1) w_1^T \quad \leftarrow \text{method of "deflation"}$$

Then apply the Power method to $\boldsymbol{A_2}$ get $\boldsymbol{w_2}$ and $\boldsymbol{d_2}$, and keep going iteratively…

$$\frac{A_2^k \cdot \vec{x}}{||A_2^k \vec{x}||} \rightarrow \vec{w}_2 \text{ as } k \rightarrow \infty$$

Recover Principle Axes of Data

[Ipynb: The Power Method]

Analysis of Eigenspectrum

Recall:

$$||X \cdot w_1||^2 = w_1^T \cdot X^T \cdot X \cdot w_1 = w_1^T \cdot cov(X) \cdot w_1 = w_1^T d_1 w_1 = d_1$$

Analysis of Eigenspectrum

Recall:

$$||X \cdot w_1||^2 = w_1^T \cdot X^T \cdot X \cdot w_1 = w_1^T \cdot cov(X) \cdot w_1 = w_1^T d_1 w_1 = d_1$$

And in general:

$$||X \cdot w_i||^2 = d_i$$

Analysis of Eigenspectrum

Recall:

$$||X \cdot w_1||^2 = w_1^T \cdot X^T \cdot X \cdot w_1 = w_1^T \cdot cov(X) \cdot w_1 = w_1^T d_1 w_1 = d_1$$

And in general:

$$||X \cdot w_i||^2 = d_i$$

But also recall that the quantity

$$||X \cdot w_i||^2$$

measures the amount of variance explained by i-th principal axis.

Analysis of Eigenspectrum

Recall:

$$||X \cdot w_1||^2 = w_1^T \cdot X^T \cdot X \cdot w_1 = w_1^T \cdot cov(X) \cdot w_1 = w_1^T d_1 w_1 = d_1$$

And in general:

$$||X \cdot w_i||^2 = d_i$$

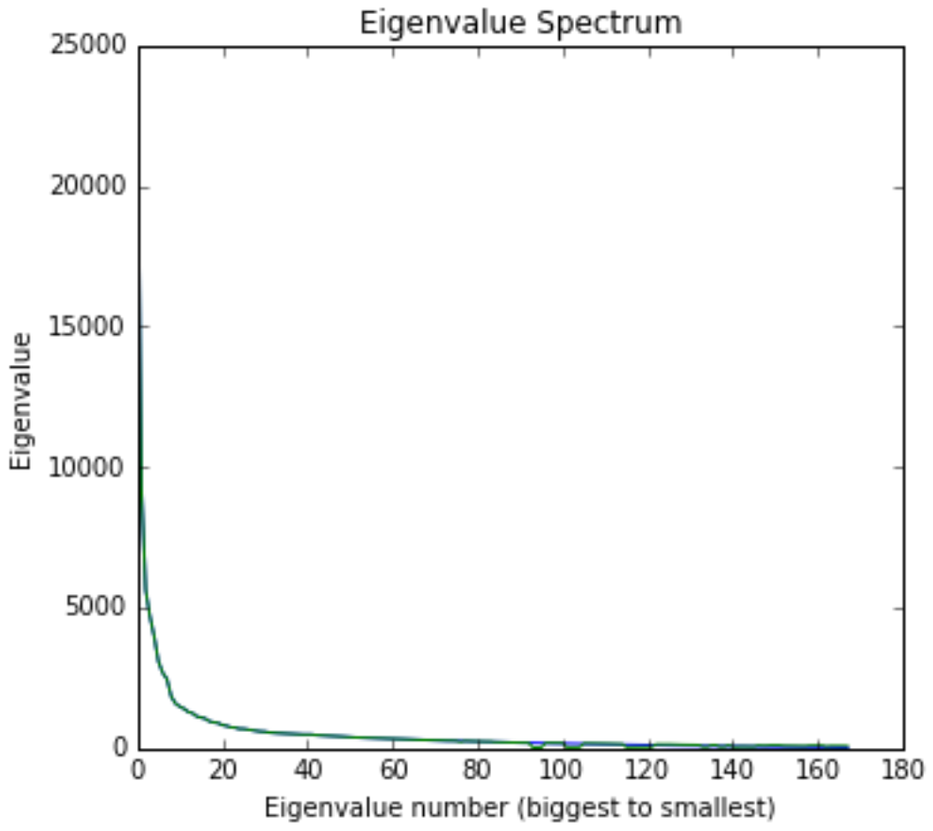
But also recall that the quantity

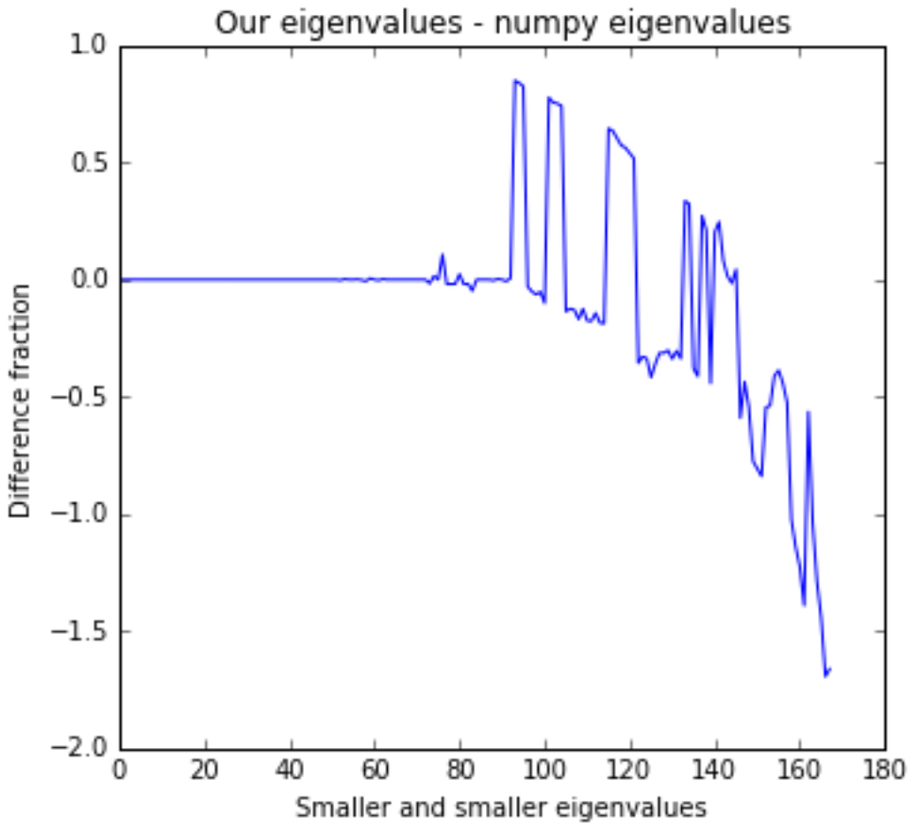
$$||X \cdot w_i||^2$$

measures the amount of variance explained by i-th principal axis.

Key point: amount of variance explained by i-th principal axis = i-th lagest eigenvalue of the data covariance matrix.

Recover Principle Axes of Data





Analysis of Eigenspectrum

Actually, just as d_1 measure the amount of variance explained by the first principal axis

$$d_1 + d_2$$

measures the amount of variance explained by the first two principal axis

Analysis of Eigenspectrum

Actually, just as d_1 measure the amount of variance captured by the first principal axis

$$d_1 + d_2$$

measures the amount of variance captured by the first two principal axis

In fact: $\sum_{i \leq k} d_i$ = amount of variance captured by first ***k*** principal axes

Analysis of Eigenspectrum

Actually, just as d_1 measure the amount of variance captured by the first principal axis

$$d_1 + d_2$$

measures the amount of variance captured by the first two principal axis

In fact: $\sum_{i \leq k} d_i$ = amount of variance captured by first ***k*** principal axes

= variance of the data captured by projection onto the subspace spanned by the first ***k*** eigenvectors

Thus:

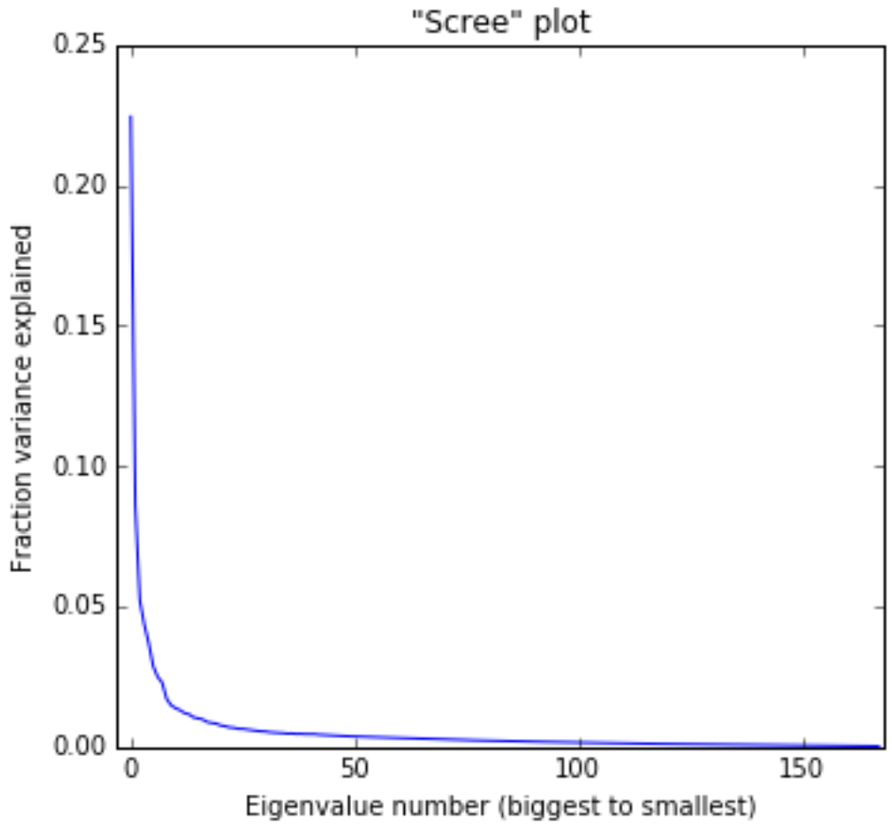
$$\frac{\sum_{i \leq k} d_i}{\sum_{i=1}^{n} d_i}$$ = *fraction* of total variance explained by first ***k*** principal axes

Analysis of Eigenspectrum

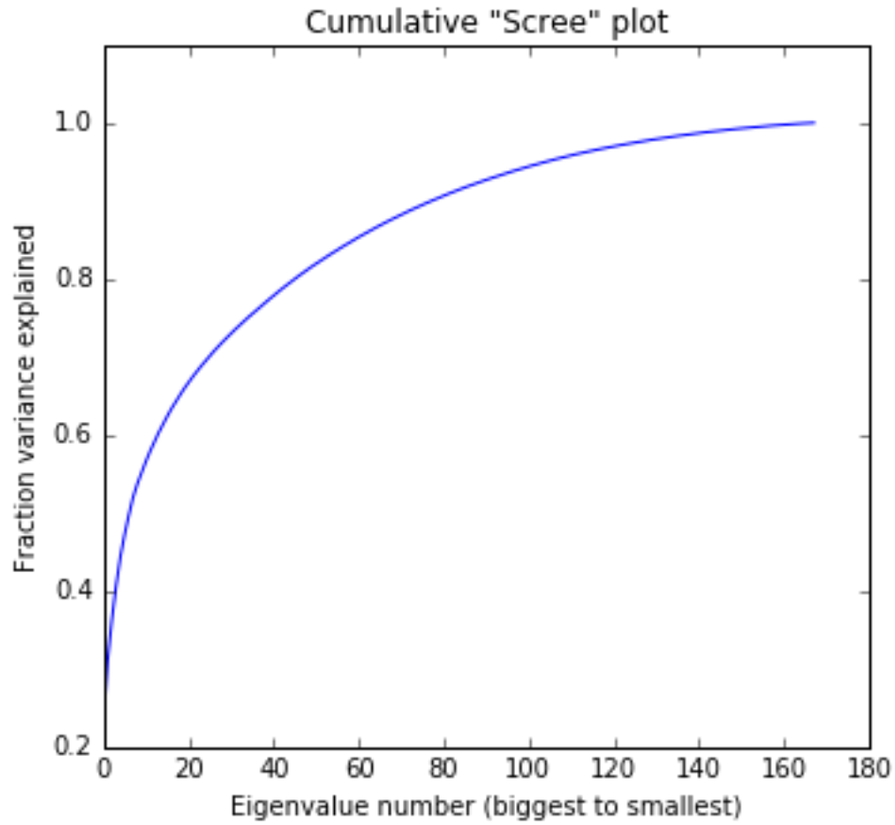
[Ipynb: scree plots]

Analysis of Eigenspectrum

This neural data is pretty intrinsically high dimensional …



$$\frac{d_i}{\sum_{i=1}^{n} d_i}$$



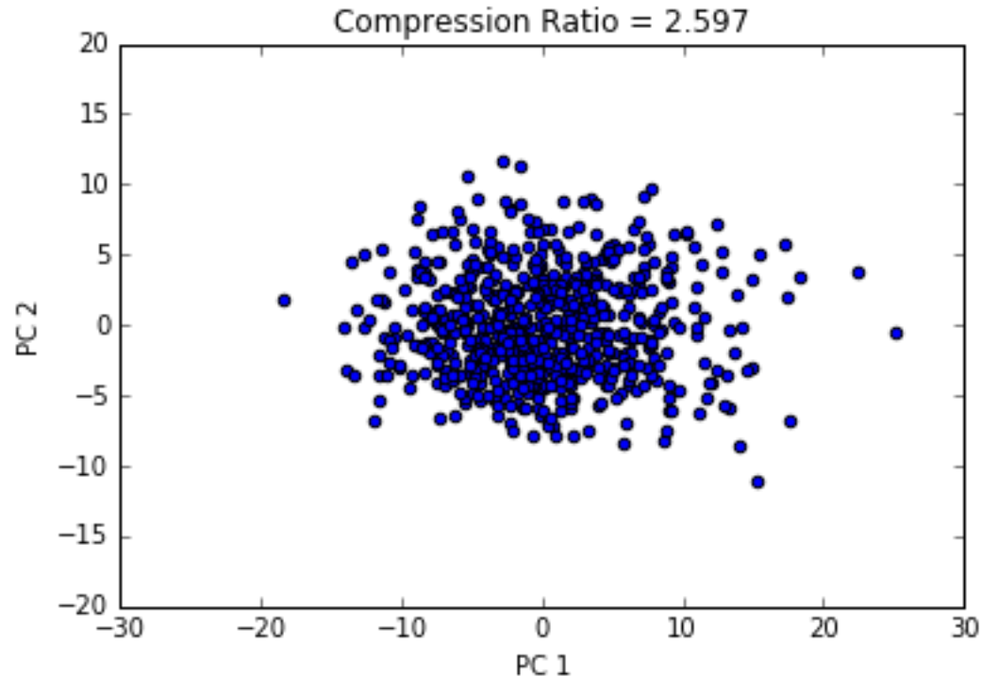
$$\frac{\sum_{i \leq k} d_i}{\sum_{i=1}^{n} d_i}$$

Analysis of Eigenspectrum

[Ipynb: compression ratio]

Analysis of Eigenspectrum

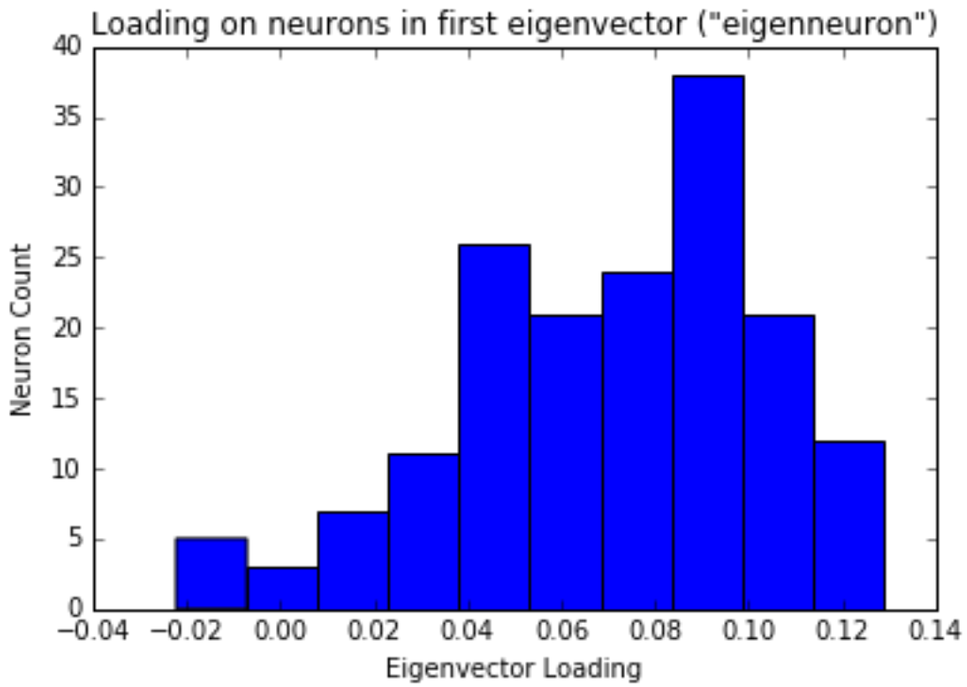
$$\frac{d_i}{d_{i+1}} = \textit{compression ratio} \text{ of i-th eigenvector}$$



Analysis of Eigenspectrum

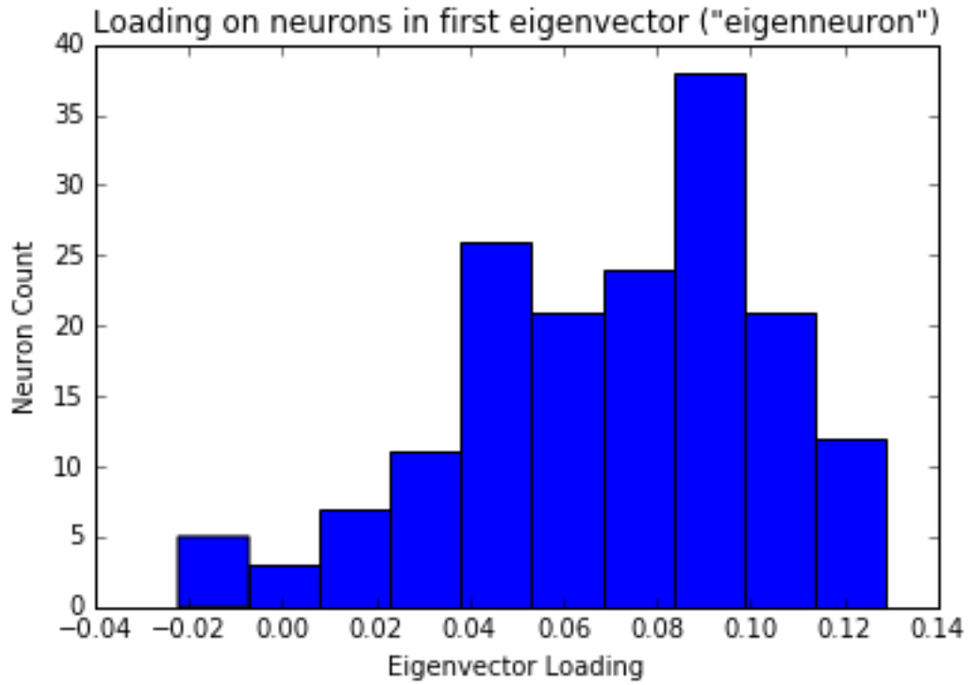
[Ipynb: Analysis of eigenvectors]

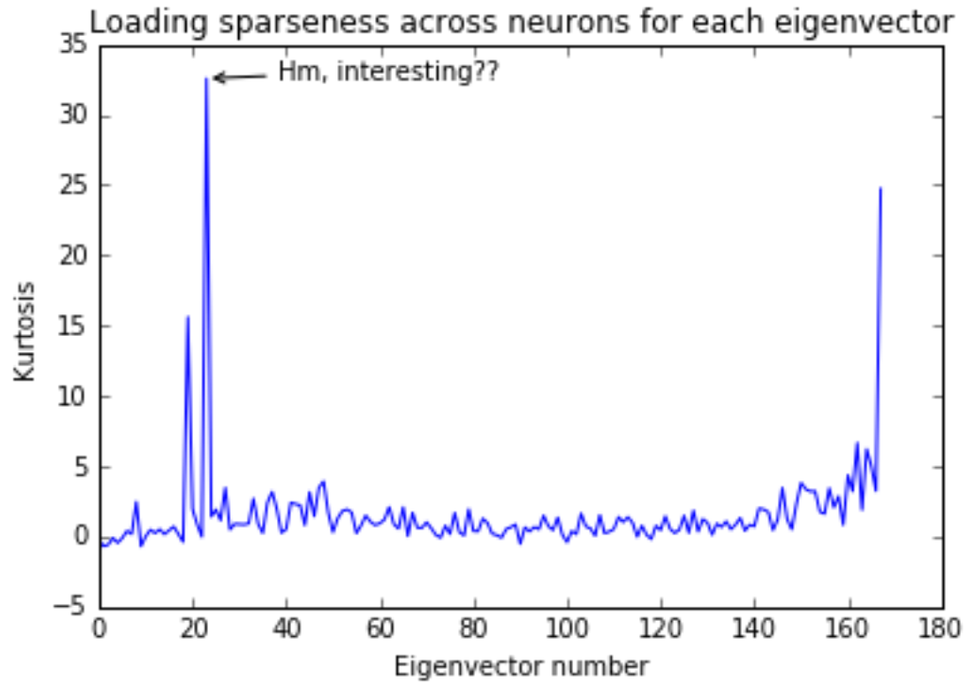
Analysis of Eigenspectrum



This first eigenvector is pretty well-distributed across neurons….
…let's look at this for all neurons

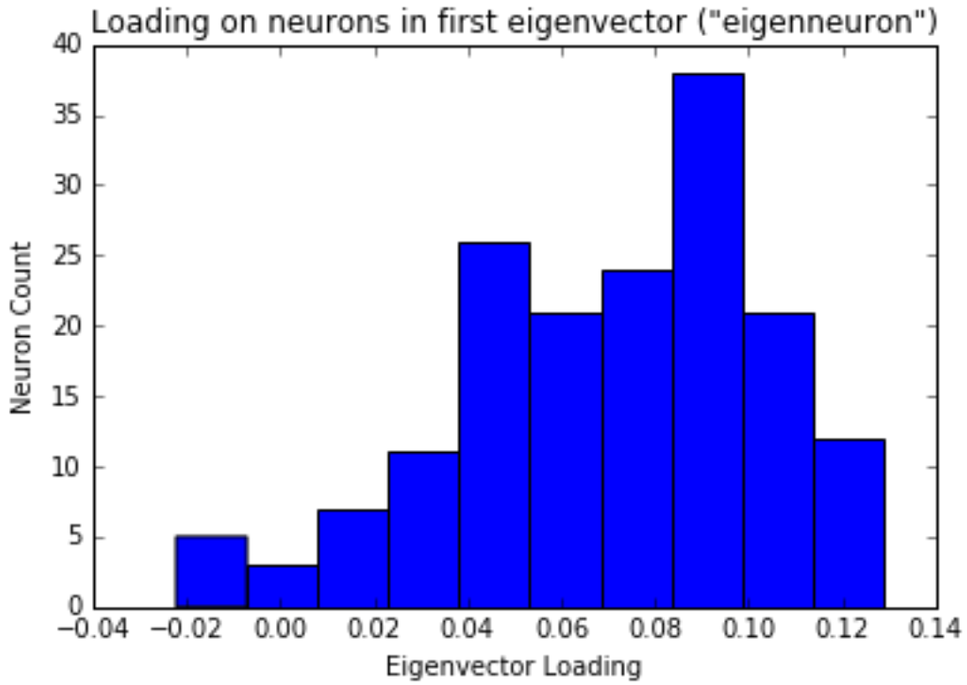
Analysis of Eigenspectrum

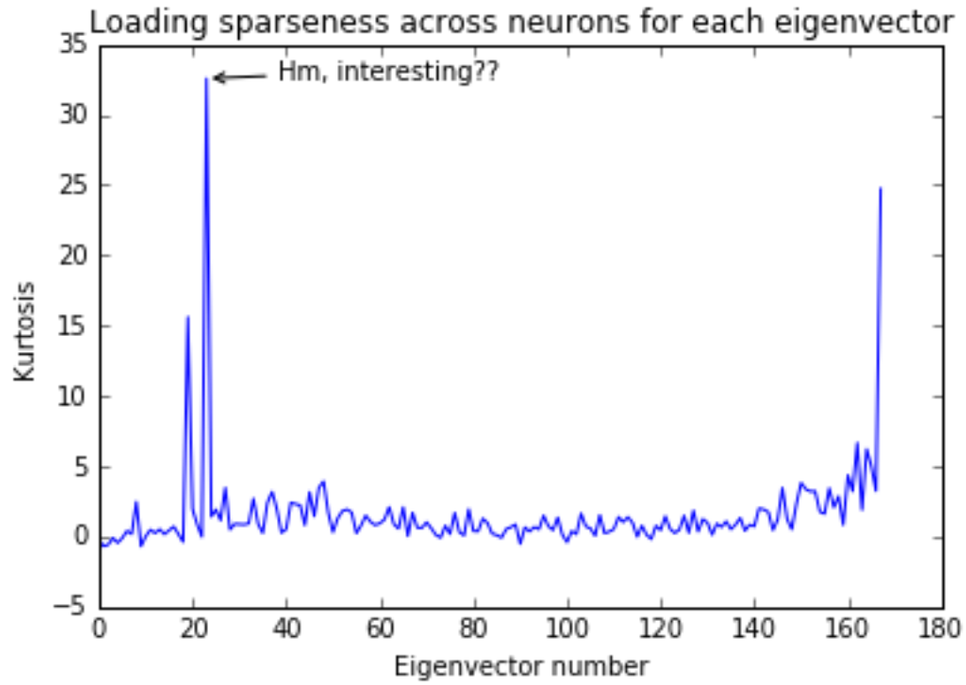


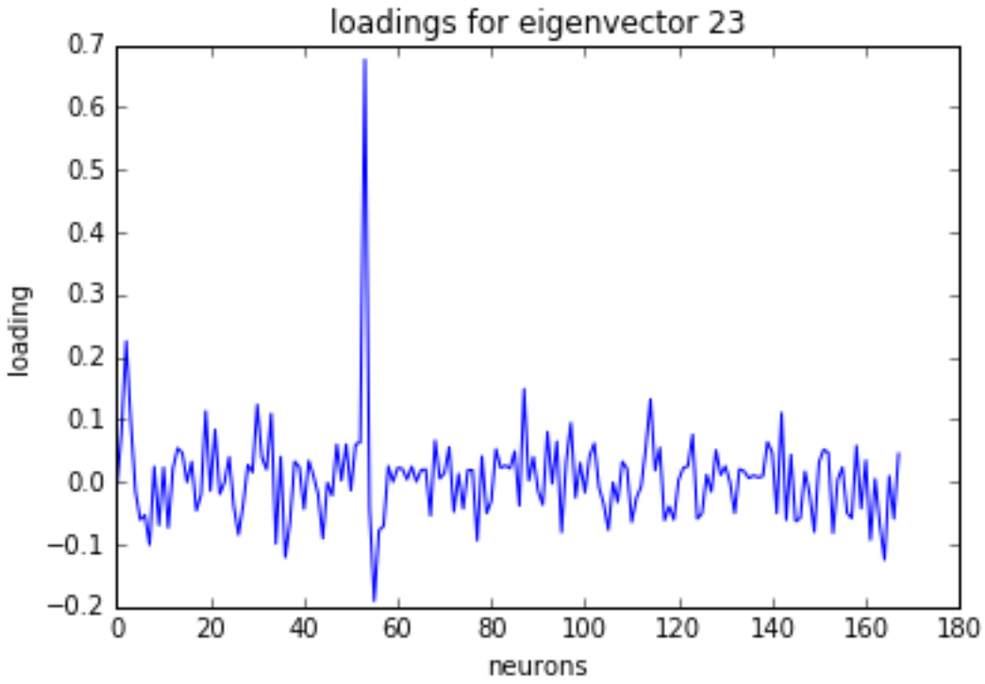


Interesting — most eigenvectors look essentially normally-distributed in terms of their sparsity (e.g. kurtosis around 0), but there's a few that are very sparse! E.g. one or a few neurons are especially heavily weighted in those eigenvectors.
Let's look at the sparsest one (eigenvector 23).

Analysis of Eigenspectrum

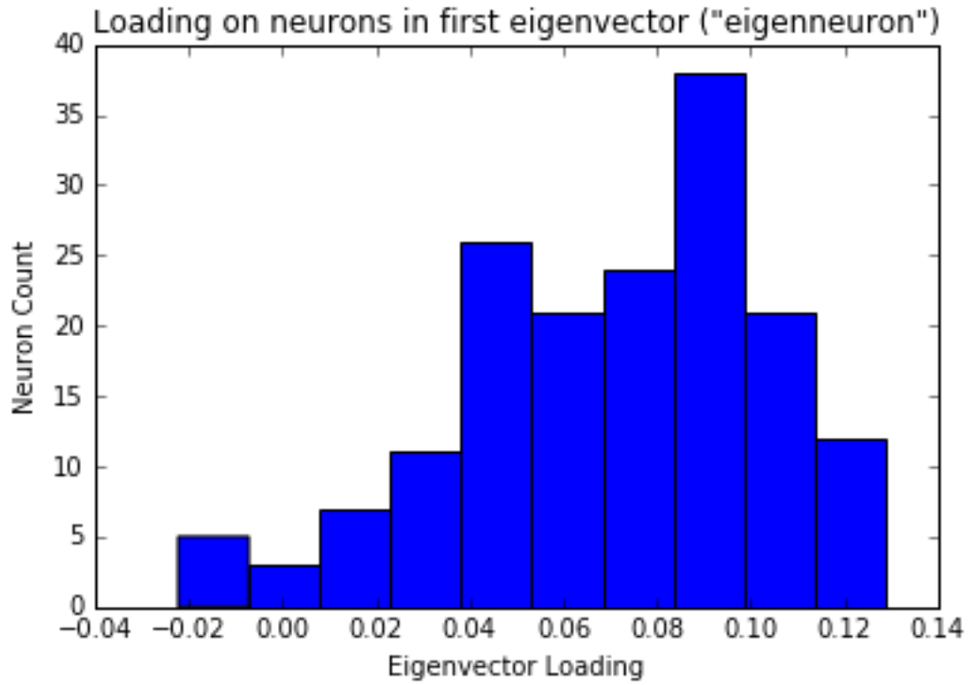


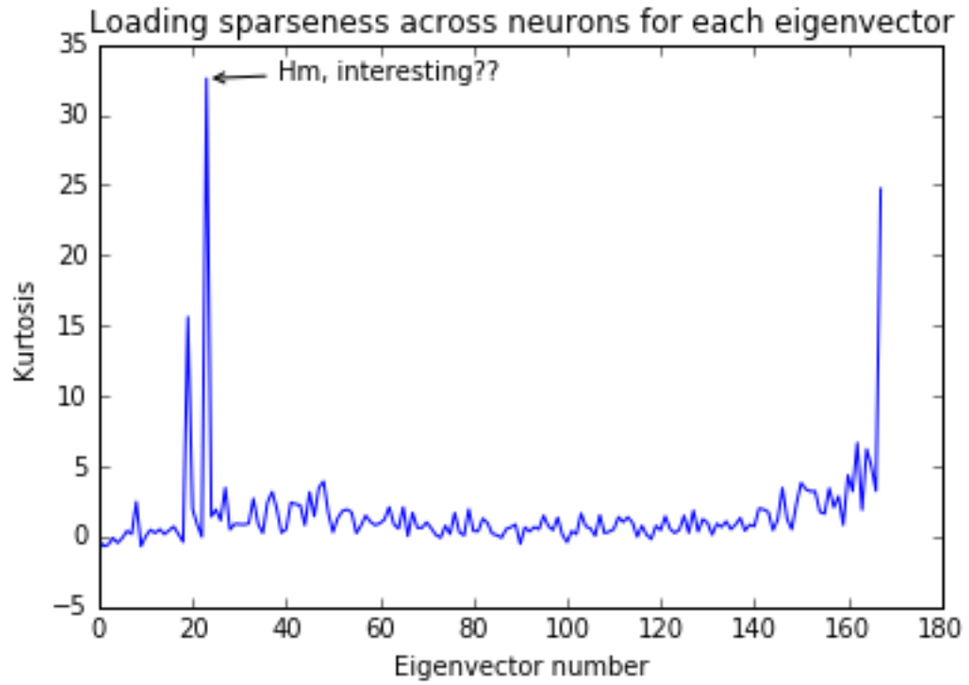




Ok it loads mostly on neuron 53.
What does neuron 53 do?

Analysis of Eigenspectrum





What does neuron 58 do? It's a little hard to tell