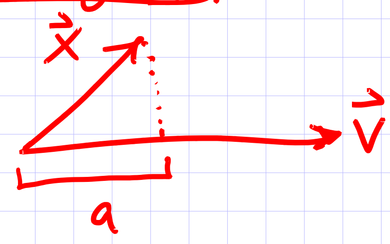
Lecture 18: 3/27/17

Recall: Projection



Length of projection of $\vec{x}$ onto $\vec{v}$

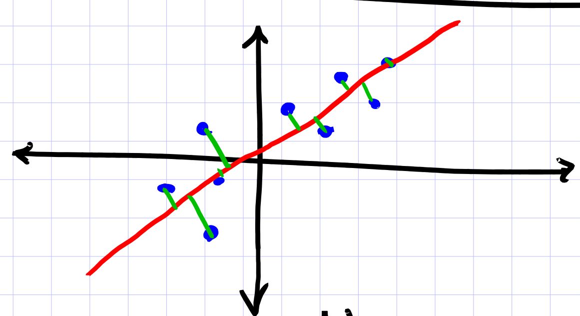
$$a = \vec{v}^T\vec{x} \quad \text{if } \|x\|_2 = 1 \text{ and } \|v\|_2 = 1$$

Vector representing that projection

$$a\vec{v} = (\vec{v}^T\vec{x})\vec{v}$$

★ For finding the first principal component...

Minimize the Reconstruction Error

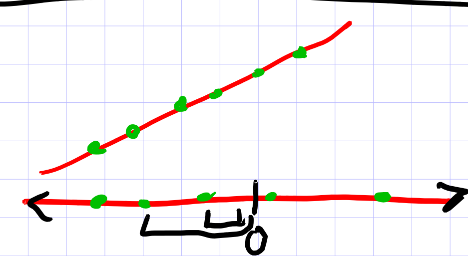


$$\hat{v} = \operatorname*{argmin}_{\vec{v}} \frac{1}{N}\sum_{i=1}^{N} \|\vec{x}^{(i)} - (\text{projection of } x^{(i)})\|_2^2$$

$$= \operatorname*{argmin}_{\vec{v}} \frac{1}{N}\sum_{i=1}^{N} \|\vec{x}^{(i)} - (\vec{v}^T x^{(i)})\,\vec{v}\|_2^2$$

such that $\|v\|_2 = 1$

Maximize the Variance



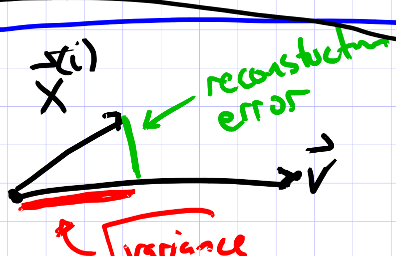
sum of the squares

$$\hat{v} = \operatorname*{argmax}_{\vec{v}} \frac{1}{N}\sum_{i=1}^{N} \left(\text{projection length of } x^{(i)}\right)^2$$

$$= \operatorname*{argmax}_{\vec{v}} \frac{1}{N}\sum_{i=1}^{N} \left(\vec{v}^T\vec{x}^{(i)}\right)^2$$

such that $\|v\|_2 = 1$

These two objectives give back the same $\hat{v}$. They are equivalent.



By Pythagorean Thm. min green is equivalent to max red.

Thm: The vector that maximizes variance is the eigenvector of Σ the sample covariance matrix with largest eigenvalue.

Recall: $\vec{v}$ is an eigenvector of Σ if $\Sigma\vec{v} = \lambda\vec{v}$ for some eigenvalue $\lambda \in \mathbb{R}$

Recall: The eigenvectors of a symmetric matrix are orthogonal to each other.

Σ is a symmetric matrix!

PCA Projection

$$u^{(i)} = \begin{bmatrix} \vec{v}_1^T \vec{x}^{(i)} \\ \vdots \\ \vec{v}_k^T \vec{x}^{(i)} \end{bmatrix}$$

$\vec{v}_1$ is the eigenvector of Σ w/ 1st largest eigenvalue
$\vec{v}_2$ is the eigenvector of Σ w/ 2nd largest eigenvalue
$\vdots$
$\vec{v}_k$ is the eigenvector of Σ w/ k^{th} largest eigenvalue

★ Collectively, we have the set of vectors $\vec{v}_1, \ldots, \vec{v}_k$ that minimize reconstruction error and are orthogonal to each other.

How do we find these eigenvectors?

1. Power Iteration (Von Meses iteration)
   - one at a time PCs
2. Singular Value Decomposition (SVD)
   - all the PCs at once
   - Σ is $M \times M$ $= \frac{1}{N} X^T X$
   - trick whereby we only need compute SVD of X, not of Σ
     (X: $N \times M$ matrix)
3. Approximate (Random) Methods
   - very efficient for large data