Principal Component Analysis (PCA)

- $\{\underline{x}_i\}_{i=1}^N$ feature vectors i.i.d. distribution $p_{\underline{x}}(\underline{x})$
- $\underline{x}_i$ realization of random vector $\underline{X}_i$
- $\underline{X}_i$ has mean $\underline{\mu}$ and covariance matrix C
- subtract out $\underline{\mu}$ from all $\underline{x}_i$'s

Then representative feature vector $\underline{x}$ is a realization of random vector $\underline{X}$ with $E[\underline{X}] = 0$, $\text{Cov}(\underline{X}) = C$

Goal Find orthogonal matrix $W_{d\times d}$ (i.e., $W^T = W^{-1}$)
s.t. (i) $\tilde{\underline{X}} = W\underline{X}$ has uncorrelated features $\tilde{X}_i$
(ii) The first k features $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_k$ capture as much variance in data as possible for each $k = 1, 2, \ldots, d$.

$$\tilde{C} = \text{Cov}(\tilde{\underline{X}}) = \text{Cov}(W\underline{X}) \overset{\text{zero-mean } \underline{X}}{=} E[(W\underline{X})(W\underline{X})^T]$$
$$= W E[\underline{X}\underline{X}^T] W^T$$
$$= W C W^T$$

We want $\tilde{C}$ to be diagonal with the diagonal elements (variances of $\tilde{X}_i$'s) in decreasing order

Solution Eigen-decomposition of C

$$C = U D U^T$$

with $D = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \lambda_d \end{bmatrix}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

and $U = [\underline{u}_1\ \underline{u}_2 \cdots \underline{u}_d]$ s.t. $U^T U = I$

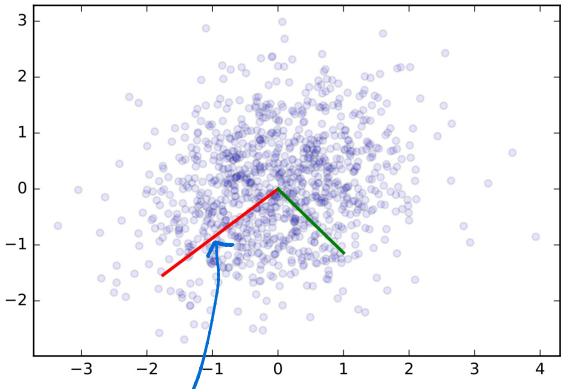
Now set $W = U^T$. Then,

$$\tilde{C} = \text{Cov}(\tilde{\underline{X}}) = \text{Cov}(U^T \underline{X}) = U^T C U$$

$$= U^T U D U^T U = D$$

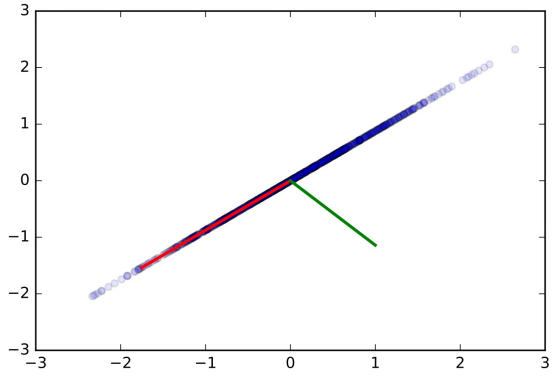
$\Rightarrow \tilde{\underline{X}}$ has uncorrelated features with $\text{Var}(\tilde{X}_i) = \lambda_i$

Also, $\text{Var}(\tilde{X}_1) \geq \text{Var}(\tilde{X}_2) \geq \cdots \geq \text{Var}(\tilde{X}_d) \geq 0$



max. variance

$d = 2$



Projection to 1-d

PCA can be used for dimensionality reduction by keeping only top k rows ($k < d$) of W

$$W = \begin{bmatrix} \underline{u}_1^T \\ \underline{u}_2^T \\ \vdots \\ \underline{u}_d^T \end{bmatrix}_{d \times d} \longrightarrow W_k = \begin{bmatrix} \underline{u}_1^T \\ \vdots \\ \underline{u}_k^T \end{bmatrix}_{k \times d}$$

$$\tilde{\underline{X}}_{k \times 1} = W_k \, \underline{X}_{d \times 1}$$

To project back into original space

$$\underbrace{W_k^T \tilde{\underline{X}}}_{d \times 1} = W_k^T W_k \underline{X}_{d \times 1} \approx \underline{X}$$

In practice, C is not known a priori but is estimated from data.

data matrix $\rightarrow X_{N \times d} = \begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_N^T \end{bmatrix}$

d typically $<< N$

Estimate of Covariance:

$$\hat{C}_{d \times d} = \frac{1}{N} \sum_{i=1}^{N} \underline{x}_i \underline{x}_i^T = \frac{1}{N} X^T X$$

To find PCA transform W, we can compute eigen decomposition of $\hat{C}$ as $U D U^T$ and set $W = U^T$

But since $\hat{C}$ is a scaled version of $X^T X$, we can also get W from SVD of X

$$X = U S V^T$$

$$X^T X = V S^2 V^T$$

$$\hat{C} = \frac{1}{N} X^T X = V \frac{S^2}{N} V^T$$

Set $W = V^T$

Variance (estimate) of i^{th} feature $= \frac{\sigma_i^2}{N}$.

Choosing Number of Principal Components $k < d$

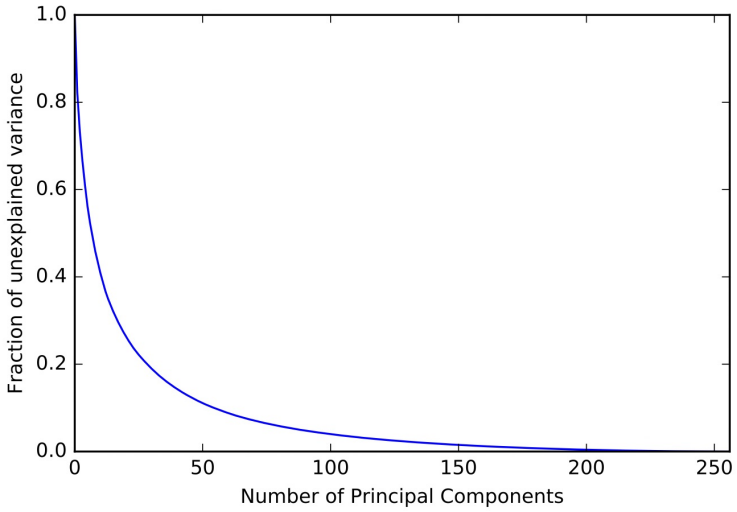
With dim. reduction from d to k,

$$\underset{d\times 1}{\underline{x}_i} \longrightarrow \underbrace{W_k \underline{x}_i}_{k\times 1}$$

Error in approximating $\underline{x}_i = \| \underline{x}_i - W_k^T W_k \underline{x}_i \|$

Cost function :

$$J(k) = \frac{1}{N} \sum_{i=1}^{N} \| \underline{x}_i - W_k^T W_k \underline{x}_i \|^2$$



"Scree" Plot

Look for knee in scree plot.