Principal Component Analysis

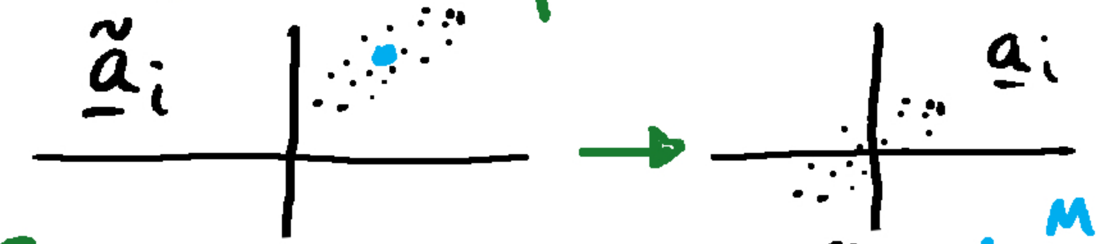
Objectives

- Define principal components
- Relate principal components to singular vectors
- Relate geometry of data matrix to singular vectors and singular values

PCA represents maximum "variance"

Data: $\underline{a}_i,\ i=1,2,\cdots M$ $(N \times 1)$ vectors, $\underline{A} = [\underline{a}_1\ \underline{a}_2 \cdots \underline{a}_M]$

– PCA assumes zero mean $\Rightarrow$ 1st step: center data



$$\underline{a}_i = \tilde{\underline{a}}_i - \frac{1}{M}\sum_{j=1}^{M} \tilde{\underline{a}}_j$$

– First principal component: direction $\underline{f}$ accounting for maximum variance in data, $\|\underline{f}\|_2^2 = 1$

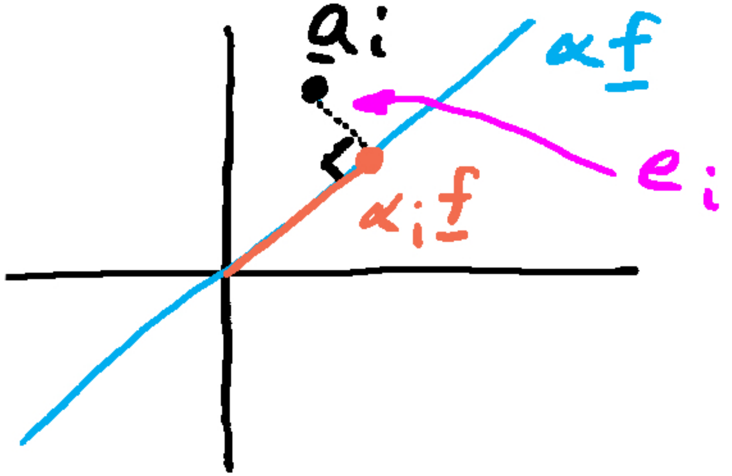
$$\max_{\|\underline{f}\|_2^2=1} \left\{ \frac{1}{M}\sum_{i=1}^{M} \|\alpha_i \underline{f}\|_2^2 \right\}$$

best line

$$\min_{\alpha_i} \|\underline{a}_i - \alpha_i \underline{f}\|_2^2 \Rightarrow \max_{\|\underline{f}\|_2^2=1} \left\{ \frac{1}{M}\sum_{i=1}^{M} |\underline{f}^T \underline{a}_i|^2 \right\}$$

$$\alpha_i = \underline{f}^T \underline{a}_i$$

$$\|\underline{f}^T \underline{A}\|_2^2 = \|\underline{A}^T \underline{f}\|_2^2$$



Principal Components are singular vectors

$$\max_{\|\underline{f}\|_2^2=1} \frac{1}{m}\underline{f}^T\underline{A}\,\underline{A}^T\underline{f} \Rightarrow \underline{A} = \underline{U}\,\underline{\Sigma}\,\underline{V}^T \qquad \underline{A}\,\underline{A}^T = \underline{U}\,\underline{\Sigma}\,\underline{V}^T\underline{V}\,\underline{\Sigma}\,\underline{U}^T$$

$$= \underline{U}\,\underline{\Sigma}^2\,\underline{U}^T$$

$$\max_{\|\underline{f}\|_2^2=1} \frac{1}{m}\underline{f}^T\underline{U}\,\underline{\Sigma}^2\,\underline{U}^T\underline{f} \Rightarrow \underline{f} = \underline{u}_1$$ (notes) "best line"

Variance associated w. 1^{st} PC $\frac{1}{m}\underline{u}_1^T\underline{A}\,\underline{A}^T\underline{u}_1 = \frac{\sigma_1^2}{m}$

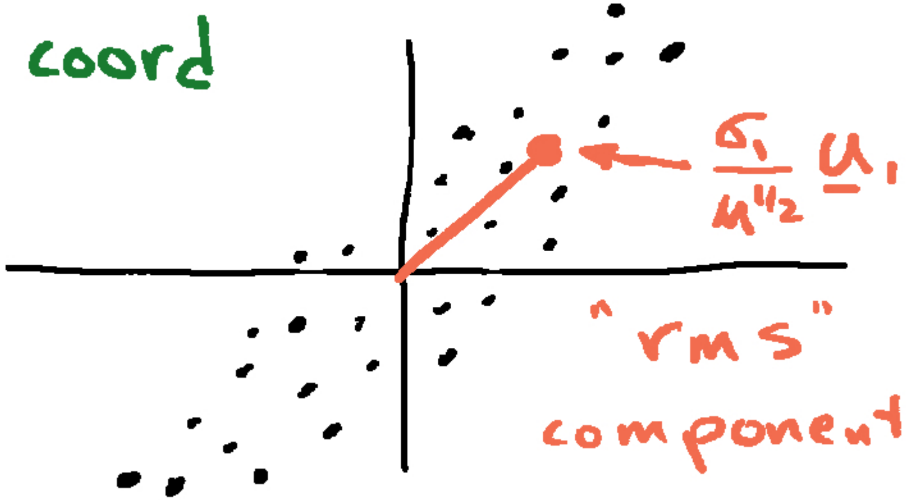
Coordinates of data: $\alpha_i = \underline{u}_1^T\underline{a}_i$, $\underline{\alpha}^T = [\alpha_1\ \alpha_2 \cdots \alpha_m]$

$$\underline{\alpha}^T = \underline{u}_1^T\underline{A}$$

root mean square coord

$$= \underline{u}_1^T\underline{U}\,\underline{\Sigma}\,\underline{V}^T \qquad \left(\frac{1}{m}\sum_{i=1}^{m}|\alpha_i|^2\right)^{1/2}$$

$$= \sigma_1\underline{V}_1^T \qquad = \frac{1}{m^{1/2}}\|\underline{\alpha}\|_2 = \frac{\sigma_1}{m^{1/2}}$$



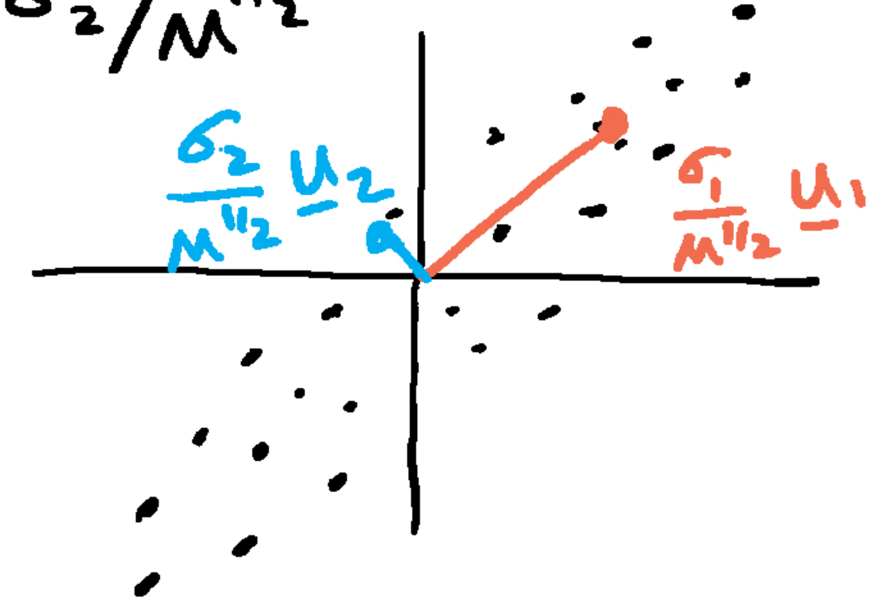
PC are singular vectors

2^{nd} PC: $\max_{\|\underline{g}\|_2^2=1,\ \underline{g}^T\underline{u}_1=0} \frac{1}{M}\sum_{i=1}^{M} |\underline{g}^T\underline{a}_i|^2$

$\Rightarrow \underline{g} = \underline{u}_2$ (2^{nd} left singular vector)

Variance associated w. 2^{nd} PC: $\frac{1}{M}\underline{u}_2^T\underline{A}\,\underline{A}^T\underline{u}_2 = \frac{\sigma_2^2}{M}$

RMS value of 2^{nd} PC coord: $\sigma_2/M^{1/2}$



k^{th} PC: $\underline{u}_k$

k^{th} PC Variance: $\frac{\sigma_k^2}{M}$

k^{th} PC coord RMS: $\frac{\sigma_k}{M^{1/2}}$

Summary $\underline{A} = \underline{U}\,\underline{\Sigma}\,\underline{V}^T$

- Left sing. vectors $\longleftrightarrow$ PC for columns of $\underline{A}$
- Sing. values $\longleftrightarrow$ ~ RMS value PC coords
- PC for rows of $\underline{A}$
  - use columns of $\underline{A}^T = \underline{V}\,\underline{\Sigma}\,\underline{U}^T$
  - Right sing. vectors $\underline{v}_i$ are PC
  - Sing. values ~ RMS value PC coords
- Eckhart - Young: SVD gives best rank r approximation to $\underline{A}$ $\quad \underline{A} \approx \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^T$

$$r = 1 \quad \underline{A} \approx \underline{u}_1 (\sigma_1 \underline{v}_1^T)$$