Principal Component Analysis

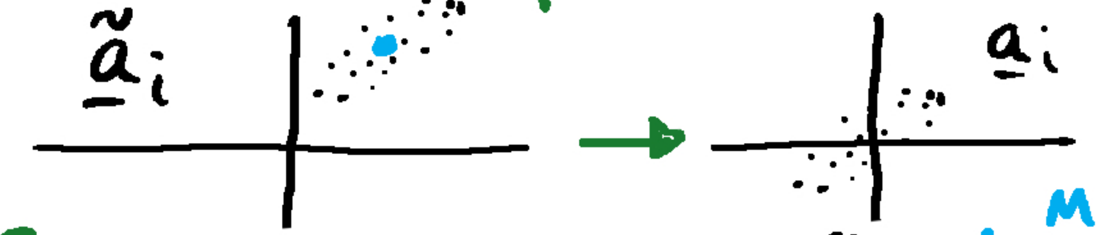
Objectives

- 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

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

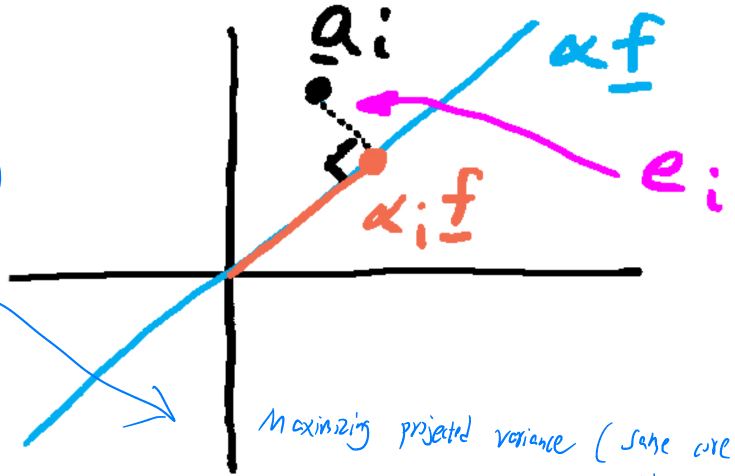
a_i projects along f

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

(minimizing reconstruction error)

(minimize the distance)

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



Maximizing projected variance (same core idea with minimizing the distance)

What PCA tries to do?

We want to use minimal information to preserve the most features possible for raw data.

∴ Variance here ≈ information

(only if the information spread, the distinction between each feature can therefore be preserved)

Principal Components are singular vectors

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

2-norm squared of $A^Tf = \max \|A^Tf\|_2^2$

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

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

The best direction of f when $f = u_1$

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

projection of a_i onto u_1 direction

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

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

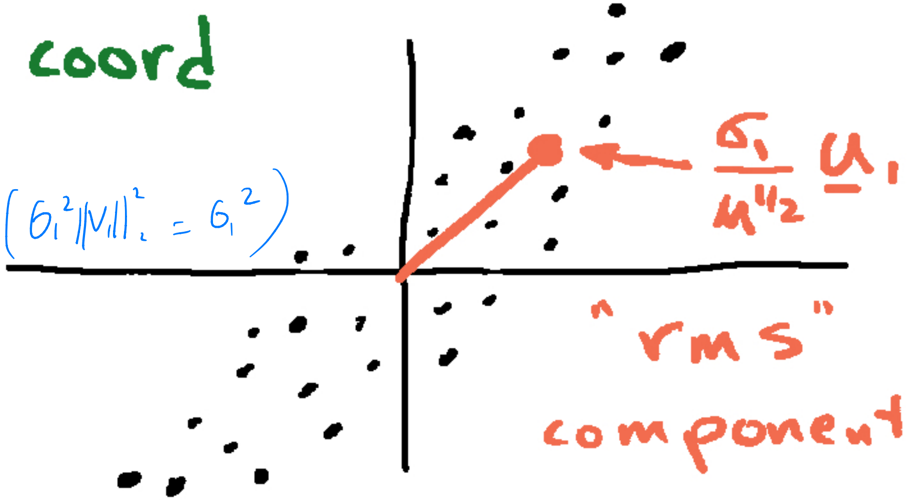
$$= \underline{u}_1^T \underline{U}\,\underline{\Sigma}\,\underline{V}^T$$

$$= \sigma_1 \underline{v}_1^T$$

root mean square coord

$$\left(\frac{1}{M}\sum_{i=1}^{M} |\alpha_i|^2\right)^{1/2} = \frac{1}{M^{1/2}}\|\underline{\alpha}\|_2 = \frac{\sigma_1}{M^{1/2}}$$

$(\sigma_1^2 \|v_1\|_2^2 = \sigma_1^2)$



PC are singular vectors

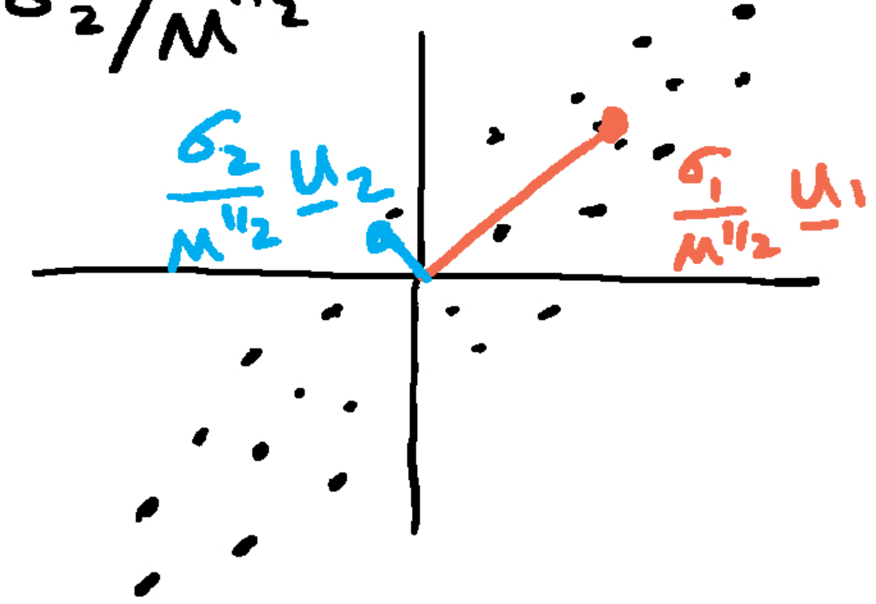
2nd PC: $\displaystyle\max_{\|\underline{g}\|_2^2=1,\ \underline{g}^T\underline{u}_1=0} \frac{1}{M}\sum_{i=1}^{M}|\underline{g}^T\underline{a}_i|^2$

Orthogonal to the first component

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

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

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



kth PC: $\underline{u}_k$

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

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

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

- Left sing. vectors $\longleftrightarrow$ PC for columns of $\underline{A}$
- Sing. values $\longleftrightarrow$ ~ RMS value PC coords

proportional to

- 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 $\underline{A}$ $\quad \underline{A} \approx \sum_{k=1}^{r} \sigma_k \underline{u}_k \underline{v}_k^T$

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

Principal Component Analysis
Proof: Left Singular Vector is the First Principal Component

Let $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_M \end{bmatrix}$ be an N-by-M ($N \geq M$) matrix with columns $\boldsymbol{a}_i$. The expression $\boldsymbol{f}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{f}$ represents the sum of squares of the elements of the vector $\boldsymbol{A}^T\boldsymbol{f}$, whose elements are the inner product between $\boldsymbol{f}$ and each column of $\boldsymbol{A}$. The solution to the problem

$$\max_{||\boldsymbol{f}||_2^2=1} \sum_{i=1}^{M} |\boldsymbol{f}^T\boldsymbol{a}_i|^2 = \max_{||\boldsymbol{f}||_2^2=1} \boldsymbol{f}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{f}$$

gives the direction $\boldsymbol{f}$ containing the maximum variability or variance across the columns of $\boldsymbol{A}$, that is, the direction that best fits the set of vectors $\boldsymbol{a}_i, i = 1, 2, \ldots, M$. The vector $\boldsymbol{f}$ is called the first principal component of the data $\boldsymbol{a}_i, i = 1, 2, \ldots, M$. Let the $\boldsymbol{A}$ have singular value decomposition $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ where $\boldsymbol{U}, \boldsymbol{V}$ are square matrices, that is, the full singular value decomposition.

Theorem:

$$\max_{||\boldsymbol{f}||_2^2=1} \boldsymbol{f}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{f} = \sigma_1^2$$

is obtained by setting $\boldsymbol{f} = \boldsymbol{u}_1$, the left singular vector corresponding to the largest singular value.

Proof: Substitute the singular value decomposition for $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ to write

$$\boldsymbol{f}^T\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{f} = \boldsymbol{f}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^T\boldsymbol{f} = \boldsymbol{f}^T\boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^T\boldsymbol{f}$$

where the second equality follows from the orthonormality of the columns of $\boldsymbol{V}$: $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$. Now let $\boldsymbol{z} = \boldsymbol{U}^T\boldsymbol{f}$ and note that $||\boldsymbol{z}||_2^2 = \boldsymbol{f}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{f} = \boldsymbol{f}^T\boldsymbol{f} = ||\boldsymbol{f}||_2^2$ because the right singular vectors in $\boldsymbol{U}$ are orthonormal. That is, $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$. Hence we may rewrite the maximization problem as

$$\max_{||\boldsymbol{z}||_2=1} \boldsymbol{z}^T\boldsymbol{\Sigma}^2\boldsymbol{z} = \max_{\sum_{i=1}^M z_i^2=1} \sum_{j=1}^{M} \sigma_i^2 z_i^2$$

where z_i is the i^{th} element of $\boldsymbol{z}$.

The unit norm constraint on $\boldsymbol{z}$ implies that an increase in the magnitude of any one element z_k must be offset by a decrease in the magnitudes of the other elements of $\boldsymbol{z}$. Since $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$, the best strategy is to allocate all of the unit energy in $\boldsymbol{z}$ to z_1. To see this, consider the case where $M = 2$ and $\sigma_1 > \sigma_2$. In this case we require $z_2^2 = 1 - z_1^2$ and

$$\sum_{i=1}^{M} \sigma_i^2 z_i^2 = \sigma_1^2 z_1^2 + \sigma_2^2(1 - z_1^2) = z_1^2(\sigma_1^2 - \sigma_2^2) + \sigma_2^2$$

Since $\sigma_1^2 - \sigma_2^2 > 0$, this quantity is maximized by choosing z_1^2 to be as large or possible, that is, $z_1^2 = 1$. Any energy allocated to z_2 is multiplied by σ_2^2, which results in a lower value than had that same energy been allocated to z_1 because $\sigma_1 > \sigma_2$. A similar argument applies to the case $M > 2$. The strategy that maximizes $\sum_{i=1}^{p} \sigma_i^2 z_i^2$ allocates all of the unit energy in $\boldsymbol{z}$ to z_1, so $\boldsymbol{z} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}$. The orthonormality of the columns of $\boldsymbol{U}$ thus imply that $\boldsymbol{f} = \boldsymbol{U}^T \boldsymbol{z}$ or $\boldsymbol{f} = \boldsymbol{u}_1$ where $\boldsymbol{u}_1$ is the first column of $\boldsymbol{U}$, the first left singular vector of $\boldsymbol{A}$.