Principal Component Analysis (PCA)

Major application of SVD to statistics & data analysis.
↳ interpretation?

n samples of m measurements each
⇝ $m \times n$ matrix, cols = samples.

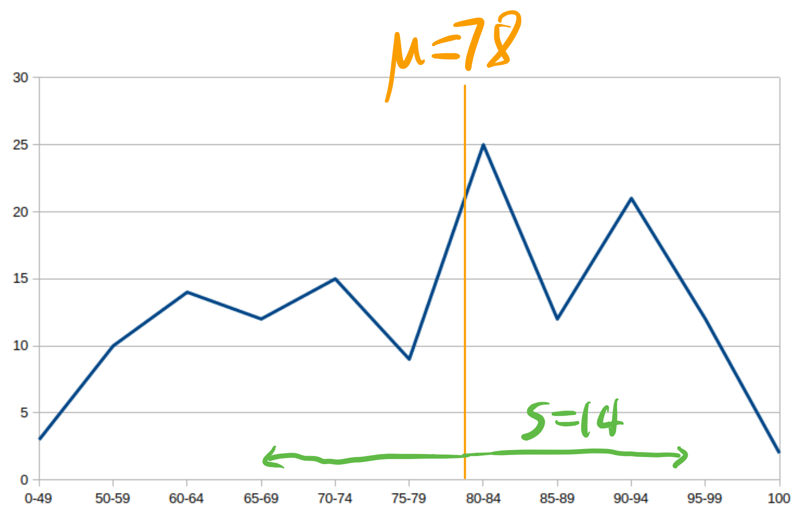
One measurement: eg. midterm scores $x_1, \ldots, x_n$

Mean / average: $\mu = \frac{1}{n}(x_1 + \cdots + x_n)$

Variance: $s^2 = \frac{1}{n-1}\left[(x_1-\mu)^2 + \cdots + (x_n-\mu)^2\right]$

Standard deviation: $s = \sqrt{\text{variance}}$

Tells you how "spaced out" the samples are:
$\approx 68\%$ of samples are within $\pm s$ of the mean.



Whence this formula? Take stats!

Two Measurements: eg. math & history scores per student.

$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ $\quad x_i = \text{math} \quad y_i = \text{history}$

Mean math score: $\mu_1 = \frac{1}{n}(x_1 + \cdots + x_n)$

Mean history score: $\mu_2 = \frac{1}{n}(y_1 + \cdots + y_n)$

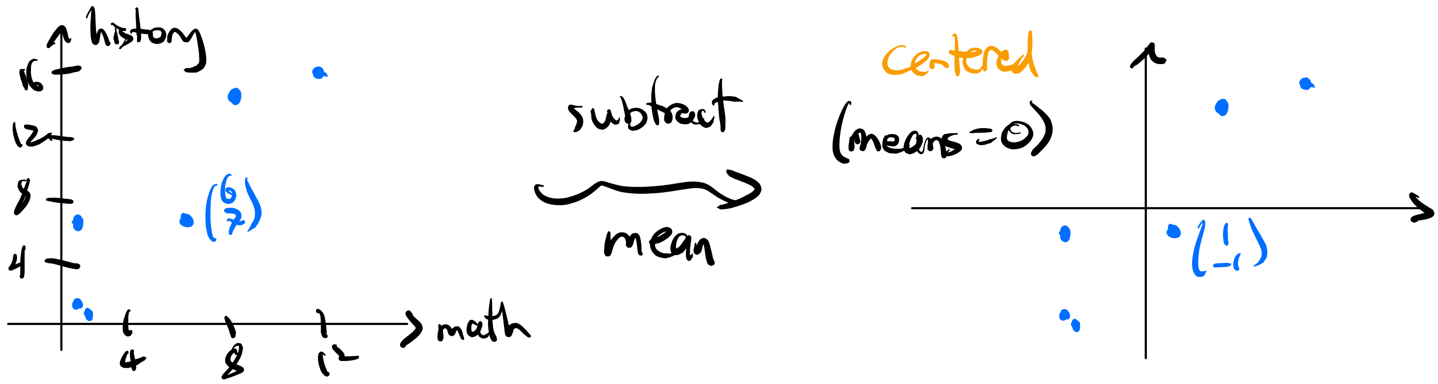
Variance math: $s_1^2 = \frac{1}{n-1}(\bar{x}_1^2 + \cdots + \bar{x}_n^2) \quad \bar{x}_i = x_i - \mu_1$

Variance history: $s_2^2 = \frac{1}{n-1}(\bar{y}_1^2 + \cdots + \bar{y}_n^2) \quad \bar{y}_i = y_i - \mu_2$

Total variance: $T = s_1^2 + s_2^2$

Eg: Scores $\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} 8 \\ 15 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 12 \\ 16 \end{pmatrix}, \begin{pmatrix} 6 \\ 7 \end{pmatrix}, \begin{pmatrix} 1 \\ 7 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ $\quad \mu_1 = 5, \ \mu_2 = 8$

subtract: $\begin{pmatrix} \bar{x}_i \\ \bar{y}_i \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \begin{pmatrix} -4 \\ -6 \end{pmatrix}, \begin{pmatrix} 7 \\ 8 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \begin{pmatrix} -4 \\ -1 \end{pmatrix}, \begin{pmatrix} -3 \\ -7 \end{pmatrix}$



Store in matrices:

$$A_0 = \begin{pmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{pmatrix} = \begin{pmatrix} 8 & 1 & 12 & 6 & 1 & 2 \\ 15 & 2 & 16 & 7 & 7 & 1 \end{pmatrix}$$

$$A = \begin{pmatrix} \bar{x}_1 & \cdots & \bar{x}_n \\ \bar{y}_1 & \cdots & \bar{y}_n \end{pmatrix} = \begin{pmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{pmatrix}$$

Covariance Matrix:

$$S = \frac{1}{n-1} AA^T = \frac{1}{n-1}\begin{pmatrix} \text{dot products} \\ \text{of rows} \end{pmatrix} = \frac{1}{n-1}\begin{pmatrix} \bar{x}_1^2 + \cdots + \bar{x}_n^2 & \bar{x}_1\bar{y}_1 + \cdots + \bar{x}_n\bar{y}_n \\ \text{same} & \bar{y}_1^2 + \cdots + \bar{y}_n^2 \end{pmatrix}$$

So $(1,1)$ entry is $s_1^2 = \frac{1}{n-1}(\bar{x}_1^2 + \cdots + \bar{x}_n^2)$

$(2,2)$ entry is $s_2^2 = \frac{1}{n-1}(\bar{y}_1^2 + \cdots + \bar{y}_n^2)$

$(1,2)$ is called covariance of rows 1 & 2.

- positive: $\bar{x}_i$ & $\bar{y}_i$ usually have the same sign: above-average math goes with above-average history (& vice-versa)
- negative: $\bar{x}_i$ & $\bar{y}_i$ usually have the opposite signs: above-average math goes with below-average history (& vice-versa)

In our case: $S = \frac{1}{5} AA^T = \begin{pmatrix} 20 & 25 \\ 25 & 40 \end{pmatrix}$ $\quad s_1^2 = 20 \quad s_2^2 = 40$

covariance $= 25 > 0$

SVD: say $\sigma_1^2 \geq \sigma_2^2$ are the eigenvals of S.

$\Rightarrow$ σ_1, σ_2 are the singular values of $\frac{1}{\sqrt{n-1}} A$

$\left(\frac{1}{\sqrt{n-1}} A\right)\left(\frac{1}{\sqrt{n-1}} A\right)^T = \frac{1}{n-1} AA^T = S$

Total variance is:

$$T = s_1^2 + s_2^2 = \mathrm{Tr}(S) = \sigma_1^2 + \sigma_2^2$$ ← Tr(S) = sum of eigenvals

much bigger

In our case: $\sigma_1 \approx 7.54 \quad \sigma_2 \approx 1.75$

$$T = \sigma_1^2 + \sigma_2^2 = 60 = 20 + 40 = s_1^2 + s_2^2$$

Let v_1, v_2 = unit eigenvectors of S
= right-singular vectors of $\frac{1}{\sqrt{n-1}} A^T$ (and A^T)

Then:

v_1 maximizes $\|A^T v\|$ subject to $\|v\| = 1$

HW12 #11(b)

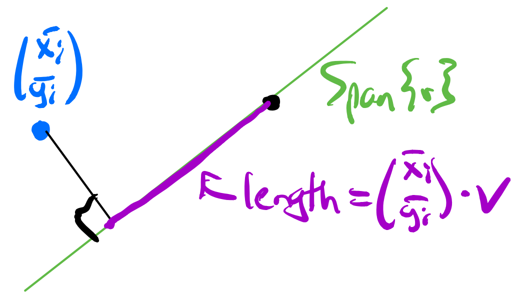
What is $A^T v$ for $\|v\| = 1$?

check: mean $= 0$

$$A^T v = \begin{pmatrix} \bar{x}_1 & \bar{y}_1 \\ \vdots & \\ x_n & \bar{y}_n \end{pmatrix} v = \begin{pmatrix} \binom{\bar{x}_1}{\bar{y}_1} \cdot v \\ \vdots \\ \binom{\bar{x}_n}{\bar{y}_n} \cdot v \end{pmatrix}$$

projection of $\binom{\bar{x}_i}{\bar{y}_i}$ onto $\text{Span}\{v\}$ is $\left[\binom{\bar{x}_i}{\bar{y}_i} \cdot v\right] v$ b/c $v \cdot v = 1$

$\binom{\bar{x}_i}{\bar{y}_i} \cdot v$ = "amount of $\binom{\bar{x}_i}{\bar{y}_i}$ in the v-direction"



So $\left\|\frac{1}{\sqrt{n-1}} A^T v\right\|^2 = \frac{1}{n-1}\left(\left[\binom{\bar{x}_1}{\bar{y}_1} \cdot v\right]^2 + \cdots + \left[\binom{\bar{x}_n}{\bar{y}_n} \cdot v\right]^2\right)$

= variance of the amount of the cols of A in the v-direction.

Maximize $\|A^T v\|^2$: v_1 is the direction of largest variance

NB $\left\|\frac{1}{\sqrt{n-1}} A^T v_1\right\| = \sigma_1$ = first singular value of $\frac{1}{\sqrt{n-1}} A^T$

σ_1^2 = variance in v_1-direction

$\frac{1}{\sqrt{n-1}} A^T v_1 = \sigma_1 u_1$

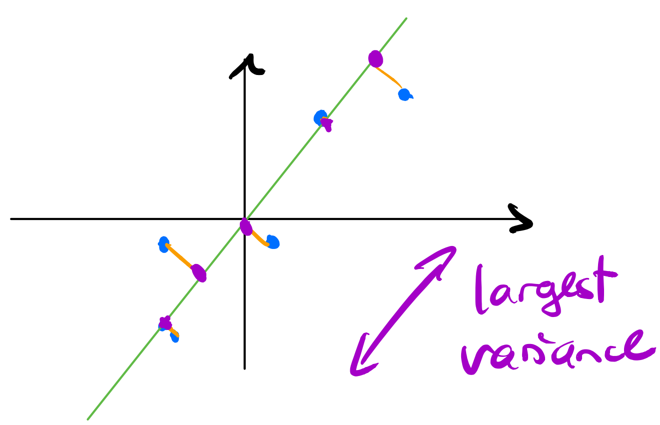
$v_1 = 1^{st}$ principal component.

In our case:

$$v_1 \approx \begin{pmatrix} .560 \\ .828 \end{pmatrix}$$

Largest variance: $\sigma_1^2 \approx 56.9$

(math-only variance: $s_1^2 \approx 20$)



Let $u_i = \frac{1}{\sigma_i} \cdot \frac{1}{\sqrt{n-1}} A^T v_i$ = left-singular vectors of $\frac{1}{\sqrt{n-1}} A^T$

SVD of $\frac{1}{\sqrt{n-1}} A$ is

$$\frac{1}{\sqrt{n-1}} A = \sigma_1 v_1 u_1^T + \sigma_2 v_2 u_2^T$$

1st summand:

$$\sigma_1 v_1 u_1^T = \sigma_1 v_1 \left(\frac{1}{\sigma_1} \frac{1}{\sqrt{n-1}} A^T v_1\right)^T = v_1 \left(\frac{1}{\sqrt{n-1}} A^T v_1\right)^T$$

$$= \frac{1}{\sqrt{n-1}} \left(\left[\binom{\bar{x}_1}{\bar{y}_1} \cdot v_1\right] v_1, \ldots, \left[\binom{\bar{x}_n}{\bar{y}_n} \cdot v_1\right] v_1 \right)$$

= projections of columns ($\times \frac{1}{\sqrt{n-1}}$) onto Span$\{v_1\}$ (purple dots in the picture)

Likewise, $\sigma_2 v_2 u_2^T$ = projections of columns ($\times \frac{1}{\sqrt{n-1}}$) onto Span$\{v_2\}$ = Span$\{v_1\}^\perp$ (orange lines in the picture)

Interpretation:

- Total variance is 60
- $\sigma_1^2 \approx 56.9$ is variance in the v_1-direction
- $\sigma_2^2 \approx 3.1$ is variance in the v_2-direction.

This says: students do $.828/.560 \approx 1.48$ times better at history than math; the standard deviation from this rule is $\sigma_2 \approx 1.75$.

PCA Reference Sheet:

- A_0: $m \times n$ matrix: cols are samples, rows are measurements
- A: subtract means of rows from rows of A_0
- $S = \frac{1}{n-1} A A^T$: Covariance matrix
  - → (i,i) entry is the variance s_i^2 of i^{th} row
  - → $Tr(S) = s_1^2 + \cdots + s_m^2$ is the total variance
  - → (i,j) entry is the covariance of rows i & j.
- Nonzero eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2$ of S
  - → singular values $\sigma_1, \ldots, \sigma_r$ of $\frac{1}{\sqrt{n-1}} A$
  - → Total variance is $s_1^2 + \cdots + s_m^2 = Tr(S) = \sigma_1^2 + \cdots + \sigma_r^2$
- Right singular vectors $v_1, \ldots, v_r$ of A^T: principal components
  - → v_1 is the direction with the most variance

    σ_1^2 = variance in v_1-direction
  - → v_2 is the direction with the most variance subject to $v \cdot v_1 = 0$; σ_2^2 = variance in v_2-direction

    ⋮
  - → v_n is the direction with the least variance
- $u_i = \frac{1}{\sqrt{n-1}} \frac{1}{\sigma_i} A^T v_i$
  - → $\sigma_1 v_1 u_1^T$ = projections of cols $\times \frac{1}{\sqrt{n-1}}$ onto Span$\{v_1\}$
  - → $\sigma_1 v_1 u_1^T + \sigma_2 v_2 u_2^T$ = projections $\times \frac{1}{\sqrt{n-1}}$ onto Span$\{v_1, v_2\}$

  etc.
- If (say) $\sigma_3, \sigma_4, \ldots, \sigma_r$ are small then

  $\frac{1}{\sqrt{n-1}} A \approx \sigma_1 v_1 u_1^T + \sigma_2 v_2 u_2^T$

  $\Rightarrow$ data almost lie on Span$\{v_1, v_2\}$

PCA detects all linear relations among your data!