Review: PCA so far

$A_0 = \begin{pmatrix} | & & | \\ d_1 & \cdots & d_n \\ | & & | \end{pmatrix}$: $m \times n$ data matrix whose columns contain n samples (data points) $d_1, \ldots, d_n$ of m measurements each.

$A = \begin{pmatrix} | & & | \\ \bar{d}_1 & \cdots & \bar{d}_n \\ | & & | \end{pmatrix} = A - \begin{pmatrix} \mu_1 & \cdots & \mu_1 \\ \vdots & \ddots & \vdots \\ \mu_m & \cdots & \mu_m \end{pmatrix}$: μ_i = mean of row i (measurement i)

recentered data matrix obtained from A by subtracting the means of the measurements (rows)

$S = \frac{1}{n-1} AA^T = \frac{1}{n-1} \begin{pmatrix} (\text{row }1)\cdot(\text{row }1) & (\text{row }1)\cdot(\text{row }2) & \cdots \\ (\text{row }2)\cdot(\text{row }1) & (\text{row }2)\cdot(\text{row }2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$:

$m \times m$ covariance matrix containing the variances of the measurements on the diagonal:

$$\frac{1}{n-1}(\text{row } i)\cdot(\text{row } i) = \frac{1}{n-1}\left(\bar{x}_{1i}^2 + \cdots + \bar{x}_{ni}^2\right) = s_i^2$$

$\longrightarrow$ total variance is $s^2 = s_1^2 + \cdots + s_m^2 = \mathrm{Tr}(S)$

NB: total variance is just

$$s^2 = s_1^2 + \cdots + s_m^2 = \frac{1}{n-1}\left(\bar{x}_{11}^2 + \cdots + \bar{x}_{n1}^2\right) + \cdots + \frac{1}{n-1}\left(\bar{x}_{1m}^2 + \cdots + \bar{x}_{nm}^2\right)$$

$$= \frac{1}{n-1}(\text{sum of squares of all entries of } A)$$

$$= \frac{1}{n-1}\left(\|\bar{d}_1\|^2 + \cdots + \|\bar{d}_n\|^2\right)$$

For $u \in \mathbb{R}^m$, $\|u\| = 1$, the variance in the u-direction is

$$s(u)^2 = u^T S u = \frac{1}{n-1}\left[(\bar{d}_1 \cdot u)^2 + \cdots + (\bar{d}_n \cdot u)^2\right]$$

If σ_1^2 is the largest eigenvalue of S then this is maximized at a unit σ_1^2-eigenvector u_1 with maximum value σ_1^2.

u_1 is the direction of largest variance.

Eg: From last time:

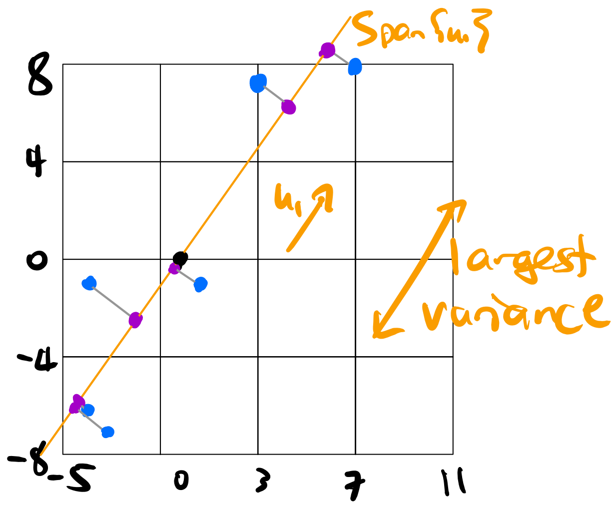
$$A_0 = \begin{pmatrix} x_1 & \cdots & x_n \\ y_1 & \cdots & y_n \end{pmatrix} = \begin{pmatrix} 8 & 1 & 12 & 6 & 1 & 2 \\ 15 & 2 & 16 & 7 & 7 & 1 \end{pmatrix} \qquad \mu_1 = 5, \quad \mu_2 = 8$$

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

$$S = \frac{1}{5} A A^T = \begin{pmatrix} 20 & 25 \\ 25 & 40 \end{pmatrix} \qquad s_1^2 = 20, \quad s_2^2 = 40 \qquad s^2 = 20 + 40 = 60$$

$\sigma_1^2 \approx 56.9$

$u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$



● $= \bar{d}_i$ ● = projection of ● onto Span{u_1}

So the direction of largest variance is u_1, and the variance in that direction is $\approx 56.9 > 20, 40$.

Our data points are "stretched out" most in the u_1-direction.

NB: Here's how I should (but won't) grade the final exam:

- Put the scores of each problem in an $m \times n$ matrix A_0
  ($m = \#$problems, $n = \#$students)
- Subtract row averages $(\mu_1, \ldots, \mu_m)$ to recenter
  $\rightsquigarrow$ matrix $A = \begin{pmatrix} | & & | \\ d_1 & \cdots & d_n \\ | & & | \end{pmatrix}$
- Compute the 1^{st} principal component u_1
- $D = \begin{pmatrix} D_1 \\ \vdots \\ D_m \end{pmatrix}$ $\quad D_j =$ max score on problem j
- The score for student i is

$$\frac{d_i \cdot u_1}{D \cdot u_1} \quad \text{(percent)}$$

This maximizes the standard deviation by reweighting the problems.

Relationship to SVD: Eigenvalues & eigenvectors of

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

compute the SVD of $\frac{1}{\sqrt{n-1}} A$ and $\frac{1}{\sqrt{n-1}} A^T$!

$$\frac{1}{\sqrt{n-1}} A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T \quad \& \quad \frac{1}{\sqrt{n-1}} A^T = \sigma_1 v_1 u_1^T + \cdots + \sigma_r v_r u_r^T$$

NB: the SVD of A is

$$A = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \cdots + \sqrt{n-1}\,\sigma_r u_r v_r^T$$

- $\sigma_1^2 \geq \cdots \geq \sigma_r^2 > 0$ are the nonzero eigenvalues of S

  NB the singular values of A are $\sqrt{n-1}\,\sigma_1, \ldots, \sqrt{n-1}\,\sigma_r$

- The trace of a square matrix is the sum of its eigenvals (HW11)

$$\Rightarrow \boxed{\text{total variance} = s^2 = \mathrm{Tr}(S) = \sigma_1^2 + \cdots + \sigma_r^2}$$

- $u_1, \ldots, u_r$ = orthonormal eigenvectors of S

  = left-singular vectors of $\frac{1}{\sqrt{n-1}} A$ (& of A)

  $S = \left(\frac{1}{\sqrt{n-1}}A\right)\left(\frac{1}{\sqrt{n-1}}A\right)^T$, not $\left(\frac{1}{\sqrt{n-1}}A\right)^T\left(\frac{1}{\sqrt{n-1}}A\right)$

- $v_i = \frac{1}{\sigma_i} \cdot \frac{1}{\sqrt{n-1}} A^T u_i$

  = right-singular vectors of $\frac{1}{\sqrt{n-1}} A$ (& of A)

We know that u_1 is the direction of largest variance.
What about $u_2, \ldots, u_r$?

QO with Extra Constraints:

- $s(u)^2 = u^T S u$ is maximized subject to $\|u\|=1$ at u_1 with $s(u_1)^2 = \sigma_1^2$
  $\rightarrow$ u_1 is the direction with largest variance
- $s(u)^2$ is maximized subject to $\|u\|=1$ and $u \perp u_1$ at u_2 with $s(u_2)^2 = \sigma_2^2$
  $\rightarrow$ u_2 is the direction with 2^{nd}-largest variance
- $s(u)^2$ is maximized subject to $\|u\|=1$ and $u \perp u_1, \dots, u \perp u_{i-1}$ at u_i with $s(u_i)^2 = \sigma_i^2$
  $\rightarrow$ u_i is the direction with i^{th}-largest variance

NB: if A has full row rank $(r=m)$ then

- $s(u)^2 = u^T S u$ is minimized subject to $\|u\|=1$ at u_r with $s(u_r)^2 = \sigma_r^2$
  $\rightarrow$ u_r is the direction with smallest variance

(If A does not have full row rank then $s(u)^2 = 0$ for any $u \in \mathrm{Nul}(A^T) \neq \{0\}$.)

The columns of $\sqrt{n-1}\,\sigma_i u_i v_i^T$ are the orthogonal projections of the columns of A onto Span$\{u_i\}$.

$$\Rightarrow A = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \cdots + \sqrt{n-1}\,\sigma_r u_r v_r^T$$

"breaks apart" your data points into principal components.

Def: Let A be a recentered data matrix with SVD

$$A = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \cdots + \sqrt{n-1}\,\sigma_r u_r v_r^T.$$

The i^{th} principal component of A is $\sqrt{n-1}\,\sigma_i u_i v_i^T$.

> The columns of the i^{th} principal component of A are the orthogonal projections of the columns of A onto Span$\{u_i\}$ = direction of i^{th}-largest variance.

Eg: In our example, $\frac{1}{\sqrt{6-1}} A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$

$\sigma_1^2 \approx 56.9$ $\qquad$ $\sigma_2^2 \approx 3.07$ $\qquad$ $S = \begin{pmatrix} 20 & 25 \\ 25 & 40 \end{pmatrix}$ $\quad s_1^2 = 20$, $s_2^2 = 40$

$u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$ $\qquad$ $u_2 \approx \begin{pmatrix} 0.828 \\ -0.561 \end{pmatrix}$

Total variance: $\sigma_1^2 + \sigma_2^2 = 56.9 + 3.1 = 60 = 20 + 40$

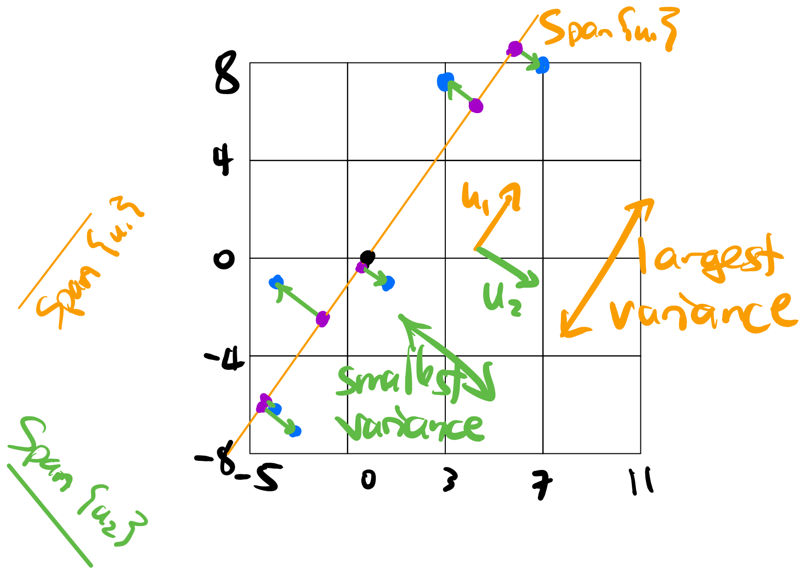
• $= \bar{d}_i$

• = columns of $\sqrt{5}\,\sigma_1 u_1 v_1^T$

= projections of • onto

↖ = columns of $\sqrt{5}\,\sigma_2 u_2 v_2^T$

= projections of • onto



NB: In this case, $s(u)^2$ is minimized at u_2 with minimum value σ_2^2 = smallest eigenvalue of S.

$$s(u_2)^2 = \frac{1}{n-1}\left[(d_1 \cdot u_2)^2 + \cdots + (d_n \cdot u_2)^2\right]$$

$$= \frac{1}{n-1}\left[\text{sum of squares of lengths of } \nwarrow\right]$$

Conclusion: The direction of largest variance is the line of best fit in the sense of orthogonal least squares, and the

$$(\text{error})^2 = (\text{sum of squares of lengths of } \nwarrow)$$

$$= (n-1)\, s(u_2)^2 = (n-1)\,\sigma_2^2$$

Subspace(s) of Best Fit

What happens in general $(m>2)$?

Def: Let V be a subspace of $\mathbb{R}^m$. The variance along V of our (recentered) data points $\bar{d}_1, \ldots, \bar{d}_n$ is

$$s(V)^2 = \frac{1}{n-1}\left(\|(\bar{d}_1)_V\|^2 + \cdots + \|(\bar{d}_n)_V\|^2\right).$$

orthogonal projections

NB: If $V = \mathrm{Span}\{u\}$ for u a unit vector then

$(\bar{d}_i)_V = (\bar{d}_i \cdot u)u$, so $\|(\bar{d}_i)_V\|^2 = (\bar{d}_i \cdot u)^2\|u\|^2 = (\bar{d}_i \cdot u)^2$,

so

$$\boxed{s(V)^2 = \frac{1}{n-1}\left[(\bar{d}_1\cdot u)^2 + \cdots + (\bar{d}_n \cdot u)^2\right] = s(u)^2}$$

Recall: if $u \perp v$ then $\|u+v\|^2 = \|u\|^2 + \|v\|^2$.

Taking $u = (\bar{d}_i)_V$ & $v = (\bar{d}_i)_{V^\perp}$ gives $\underbrace{\bar{d}_i = (\bar{d}_i)_V + (\bar{d}_i)_{V^\perp}}_{\text{orthogonal decomposition}}$

$$\Longrightarrow \|\bar{d}_i\|^2 = \|(\bar{d}_i)_V\|^2 + \|(\bar{d}_i)_{V^\perp}\|^2$$

Sum over all i:

For any subspace V,

$$s(V)^2 + s(V^\perp)^2 = \frac{1}{n-1}\left[\|\bar{d}_1\|^2 + \cdots + \|\bar{d}_n\|^2\right]$$

$$= (\text{total variance}) = \sigma_1^2 + \cdots + \sigma_r^2$$

(p. 1)

NB: $s(V^\perp)^2 = \frac{1}{n-1}\left(\|(\tilde{d}_1)_{V^\perp}\|^2 + \cdots + \|(d_n)_{V^\perp}\|^2\right)$

is $\frac{1}{n-1}\times$ the sum of the squares of the (orthogonal) distances of the $\tilde{d}_i$ to V.

Def: The d-space of best fit in the sense of orthogonal least squares is the d-dimensional subspace V minimizing $s(V^\perp)^2$. The error^2 is $s(V^\perp)^2$.

(in terms of distances it's $(n-1)s(V^\perp)^2$)

NB: Minimizing $s(V^\perp)^2$ means maximizing $s(V)^2$ since $s(V)^2 + s(V^\perp)^2 =$ total variance.

Thm: Let A be a centered data matrix with SVD

$$\frac{1}{\sqrt{n-1}}A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T.$$

The d-space of best fit to its columns is

$$V = \text{Span}\{u_1, \ldots, u_d\}.$$

The variance along V is $s(V) = \sigma_1^2 + \cdots + \sigma_d^2$ and the error^2 is $s(V^\perp)^2 = \sigma_{d+1}^2 + \cdots + \sigma_r^2$.

So you "split" the total variance $\sigma_1^2 + \cdots + \sigma_r^2 = s^2 = s(V)^2 + s(V^\perp)^2$ into the large part $s(V)^2 = \sigma_1^2 + \cdots + \sigma_d^2$ and the small part $s(V^\perp)^2 = \sigma_{d+1}^2 + \cdots + \sigma_r^2$.

Eg: The line of best fit is the first principal component $V = \text{Span}\{u_1\}$. The error$^2 = \sigma_2^2 + \cdots + \sigma_r^2$.

Eg: The plane of best fit is the span of the first 2 principal components: $V = \mathrm{Span}\{u_1, u_2\}$ $\text{error}^2 = \sigma_3^2 + \cdots + \sigma_r^2$

Eg: Suppose

$$A = 10u_1v_1^T + 8u_2v_2^T + .2u_3v_3^T + .1u_4v_4^T$$

Then A fits the plane $V = \mathrm{Span}\{u_1, u_2\}$ to a small $\text{error}^2 = .2^2 + .1^2$.

But A does not fit the line $L = \mathrm{Span}\{u_1\}$ well: the $\text{error}^2 = 8^2 + .2^2 + .1^2$.

Upshot: If $\sigma_1, \ldots, \sigma_d$ are much larger than $\sigma_{d+1}, \ldots, \sigma_r$ then your data closely fit the d-space

$$V = \mathrm{Span}\{u_1, \ldots, u_d\}$$

(but not a smaller subspace like $\mathrm{Span}\{u_1, \ldots, u_{d-1}\}$).

NB: This is all applied to the recentered data points. Your original data points $d_1, \ldots, d_n$ = columns of A fit the translated subspace

$$V + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} \quad \text{(add back the means)}.$$

See the Netflix problem on HW15.