PCA so far

$\vec{d}_1 + \cdots + \vec{d}_n = 0$

$A = \begin{pmatrix} | & & | \\ \vec{d}_1 & \cdots & \vec{d}_n \\ | & & | \end{pmatrix}$ $m \times n$ recentered data matrix, rank r

$S = \frac{1}{n-1} A A^T$ covariance matrix

- diagonal entries $s_1^2, \ldots, s_m^2$ are the measurement variances
- nonzero eigenvalues are $\sigma_1^2, \ldots, \sigma_r^2$ with orthonormal eigenvectors $u_1, \ldots, u_r$
- total variance $s^2 = \mathrm{Tr}(S) = s_1^2 + \cdots + s_m^2 = \sigma_1^2 + \cdots + \sigma_r^2$

trace = sum of eigenvals

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

- u_1 = direction of largest variance

  $\rightarrow$ maximizes $\frac{1}{n-1}\left[(\vec{d}_1 \cdot u_1)^2 + \cdots + (\vec{d}_n \cdot u_1)^2\right]$

  $= \frac{1}{n-1}$[sum of length^2 of orthogonal projections of $\vec{d}_1, \ldots, \vec{d}_n$ onto Span$\{u_1\}$]

  $= \frac{1}{n-1}$[sum of length^2 of cols of $\sqrt{n-1}\,\sigma_1 u_1 v_1^T$]

  $= \sigma_1^2$

- u_2 = direction of 2^{nd} largest variance

  $\rightarrow$ variance in u_2-direction is σ_2^2.

- etc.

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

Its columns are the projections of the data points onto $\mathrm{Span}\{u_i\}$ = direction of $i^{\underline{th}}$ largest variance:

$$\sqrt{n-1}\,\sigma_i u_i v_i^T = \begin{pmatrix} | & & | \\ (\bar{d}_1 \cdot u_i)u_i & \cdots & (\bar{d}_n \cdot u_i)u_i \\ | & & | \end{pmatrix}$$

The variance in the u_i-direction is

$$\sigma_i^2 = \frac{1}{n-1}\left[\text{sum of length}^2 \text{ of cols of } \sqrt{n-1}\,\sigma_i u_i v_i^T\right]$$

The SVD of A is

$$A = (\sqrt{n-1}\,\sigma_1)u_1 v_1^T + \cdots + (\sqrt{n-1}\,\sigma_r)u_r v_r^T$$
$$= \text{sum of the principal components.}$$

The $i^{\underline{th}}$ column of this equality says

$$\bar{d}_i = (\bar{d}_i \cdot u_1)u_1 + \cdots + (\bar{d}_i \cdot u_r)u_r$$
$$= (\text{projection of } \bar{d}_i \text{ onto } \mathrm{Span}\{u_1\})$$
$$+ \cdots + (\text{projection of } \bar{d}_i \text{ onto } \mathrm{Span}\{u_r\})$$

So the PCA decomposes the data points into principal components.

NB: Since $\{u_1,\ldots,u_r\}$ is an orthonormal basis for $V=\mathrm{Col}(A)$, the projection formula says

$$b_V = (b\cdot u_1)u_1 + \cdots + (b\cdot u_r)u_r$$

for $b\in\mathbb{R}^m$. But $\bar{d}_i = i^{\text{th}}$ col of A is in $\mathrm{Col}(A)$, so

$$\bar{d}_i = (\bar{d}_i)_V = (\bar{d}_i\cdot u_1)u_1 + \cdots + (\bar{d}_i\cdot u_r)u_r.$$

This is another way of thinking about the decomposition into principal components.

Eg: $A_0 = \begin{pmatrix} 8 & 1 & 12 & 6 & 1 & 2 \\ 15 & 2 & 16 & 7 & 7 & 1 \end{pmatrix}$

$A = \begin{pmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{pmatrix} = \begin{pmatrix} | & & | \\ \bar{d}_1 & \cdots & \bar{d}_6 \\ | & & | \end{pmatrix}$

$$\frac{1}{\sqrt{6-1}}A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$$

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

$u_1 \approx \begin{pmatrix} 0.561 \\ 0.828 \end{pmatrix}$ $\qquad$ $u_2 \approx \begin{pmatrix} 0.828 \\ -0.561 \end{pmatrix}$ $\qquad$ $s_2^2 = 40$

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

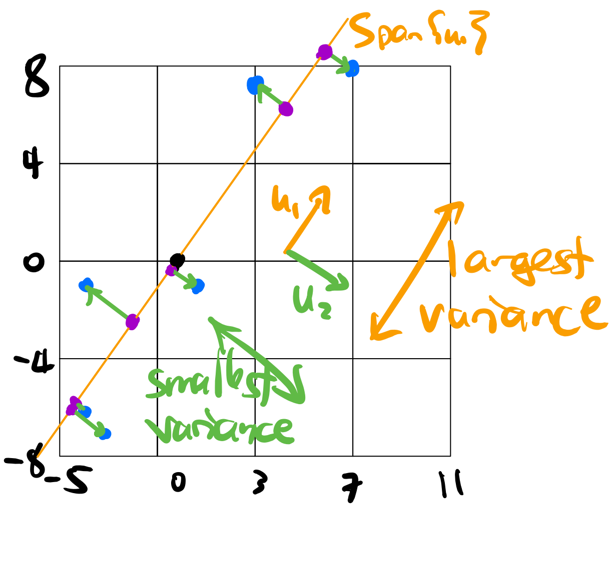
• $= \bar{d}_i$

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

$=$ projections of • onto Span$\{u_1\}$

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

$=$ projections of • onto Span$\{u_2\}$



In this case, the variance $s(u)^2 = u^T S u$ is minimized in the u_2-direction, since σ_2^2 is the smallest eigenvalue of S (0 is not an eigenvalue / A has full row rank).

The minimized quantity is

$$\sigma_2^2 = \frac{1}{n-1}\left[\text{sum of length}^2 \text{ of projections onto } \mathrm{Span}\{u_2\}\right]$$

But $\mathrm{Span}\{u_2\} = \mathrm{Span}\{u_1\}^\perp$, so projection onto $\mathrm{Span}\{u_2\}$ is $b_{V^\perp}$ for $V = \mathrm{Span}\{u_1\}$. So we've minimized

$$\sigma_2^2 = \frac{1}{n-1}\left[\text{sum of orthogonal distance}^2 \text{ from } \bar{d}_i \text{ to } \mathrm{Span}\{u_1\}\right]$$

This says $\mathrm{Span}\{u_1\}$ is the line of best fit in the sense of orthogonal least squares.

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

$$(\text{error})^2 = \frac{1}{n-1}(\text{sum of squares of lengths of } \nwarrow)$$
$$= s(u_2)^2 = \sigma_2^2$$

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

Fact: For any subspace V,

$$s(V)^2 + s(V^\perp)^2 = s_1^2 + \cdots + s_m^2 = \sigma_1^2 + \cdots + \sigma_r^2$$
$$= (\text{total variance})$$

Proof: 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 $\underbrace{\bar{d}_i = (\bar{d}_i)_V + (\bar{d}_i)_{V^\perp}}_{\text{orthogonal decomposition}}$

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

Sum over all i:

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

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

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

But $\|\bar{d}_i\|^2 = \bar{d}_i \cdot \bar{d}_i$ = sum of entries^2 of i^{th} column

$$\Rightarrow \frac{1}{n-1}\left[\|\bar{d}_1\|^2 + \cdots + \|\bar{d}_n\|^2\right] = \frac{1}{n-1}\left[\text{sum of all entries}^2 \text{ of } A\right]$$

On the other hand,

s_j^2 = variance of measurement j

$= \frac{1}{n-1}\left[\text{sum of entries}^2 \text{ of } j^{th} \text{ row of } A\right]$

So the total variance is

$$s^2 = s_1^2 + \cdots + s_m^2 = \frac{1}{n-1}\left[\text{sum of all entries}^2 \text{ of } A\right]$$

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

The total variance is also $\mathrm{Tr}(S) = \sigma_1^2 + \cdots + \sigma_r^2$. ✓

NB: $s(V^\perp)^2 = \frac{1}{n-1}\left(\|(\bar{d}_1)_{V^\perp}\|^2 + \cdots + \|(d_n)_{V^\perp}\|^2\right)$
is $\frac{1}{n-1} \times$ the sum of the squares of the (orthogonal) distances of the $\bar{d}_i$ 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 is fixed.

> The d-space of best fit is the d-space of largest variance!

We know how to find the line of best fit: $\mathrm{Span}\{u_1\}$.

What about the plane of best fit? It's $V = \mathrm{Span}\{u_1, u_2\}$.

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

Projection formula: $\{u_1, u_2\}$ is an o.n. basis for V, so

$$(\bar{d}_i)_V = (\bar{d}_i \cdot u_1)u_1 + (\bar{d}_i \cdot u_2)u_2 \quad \longleftarrow \text{orthogonal summands}$$

$$\Rightarrow \|\bar{d}_i\|_V^2 = (\bar{d}_i \cdot u_1)^2 + (\bar{d}_i \cdot u_2)^2$$

$$\Rightarrow s(V)^2 = \frac{1}{n-1}\left[(\bar{d}_1 \cdot u_1)^2 + (\bar{d}_1 \cdot u_2)^2 + \cdots + (\bar{d}_n \cdot u_1)^2 + (\bar{d}_n \cdot u_2)^2\right]$$

$$= \frac{1}{n-1}\left[(\bar{d}_1 \cdot u_1)^2 + \cdots + (\bar{d}_n \cdot u_1)^2\right] + \frac{1}{n-1}\left[(\bar{d}_1 \cdot u_2)^2 + \cdots + (\bar{d}_n \cdot u_2)^2\right]$$

$$= s(u_1)^2 + s(u_2)^2 = \sigma_1^2 + \sigma_2^2.$$

$$\Rightarrow \text{error}^2 = (\sigma_1^2 + \cdots + \sigma_r^2) - (\sigma_1^2 + \sigma_2^2) = \sigma_3^2 + \cdots + \sigma_r^2.$$

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

$$V_d = \mathrm{Span}\{u_1, \dots, u_d\}.$$

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

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

Upshot: The greedy algorithm will find the d-space of best fit by "peeling off" the remaining direction of largest variance d times.

Eg: The line of best fit is the first principal component $V_1 = \mathrm{Span}\{u_1\}$. The error$^2 = \sigma_2^2 + \cdots + \sigma_r^2$.

Eg: Suppose

$$\frac{1}{\sqrt{n-1}}A = 10u_1v_1^T + 8u_2v_2^T + .2u_3v_3^T + .1u_4v_4^T$$

Then A fits the plane $V_2 = \mathrm{Span}\{u_1, u_2\}$ to a small $\text{error}^2 = .2^2 + .1^2$.

But A does not fit the line $V_1 = \mathrm{Span}\{u_1\}$ well: the $\text{error}^2 = 8^2 + .2^2 + .1^2$.

On the other hand, A fits the 3-space $V_3 = \mathrm{Span}\{u_1, u_2, u_3\}$ even better: $\text{error}^2 = .1^2$

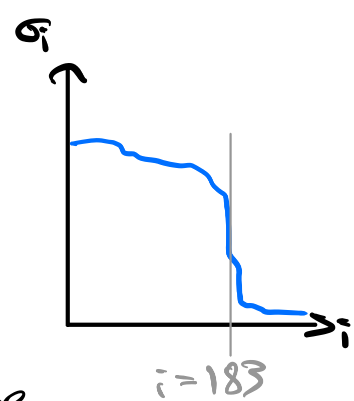
→ but this isn't much better than the plane of best fit.

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

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

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

You'll sometimes see singular values graphed in order (think: a big matrix has lots of singular values). This gives a way to visualize the dimensionality of your data. Eg this data set wants to lie on a 183-dimensional subspace.



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

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

See the Netflix problem on HW15.