PCA So Far

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

$A = \begin{pmatrix} | & & | \\ \bar{d}_1 & \cdots & \bar{d}_n \\ | & & | \end{pmatrix}$ a centered data matrix

SVD: $\frac{1}{\sqrt{n-1}} A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$

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

- diagonal entries are $s_1^2, s_2^2, \ldots, s_m^2$ = variances of the measurements (rows)
- nonzero eigenvalues are $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2 > 0$ with orthonormal eigenvectors $u_1, u_2, \ldots, u_r$
- total variance

$$s^2 = \mathrm{Tr}(S) = s_1^2 + s_2^2 + \cdots + s_m^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2$$

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

$$s(u)^2 = u^T S u = \frac{1}{n-1}\left[(\bar{d}_1 \cdot u)^2 + (\bar{d}_2 \cdot u)^2 + \cdots + (\bar{d}_n \cdot u)^2\right]$$

$$= \frac{1}{n-1}\left[\text{sum of length}^2 \text{ of projections of } \bar{d}_i \text{ onto } \mathrm{Span}\{u\}\right]$$

- $s_i^2 = s(e_i)^2$
- The variance is maximized at $x = u_1$; the maximum value is $s(u_1)^2 = \sigma_1^2$.
- If A has FRR, the variance is minimized at $x = u_r$; the minimum value is $s(u_r)^2 = \sigma_r^2$.

More generally, we can maximize $s(u)^2$ subject to $\|u\|=1$ and $u \cdot u_1 = 0$. We get the direction of second-largest variance $= u_2$, with second-largest variance $= \sigma_2^2$ (=second-largest eigenvalue of S). And so on. See L23, near the end.

> σ_i^2 = i^{th}-largest variance
> u_i = direction with i^{th}-largest variance

Recall (L25): In the SVD

$$A = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,u_2 v_2^T + \cdots + \sqrt{n-1}\,u_r v_r^T$$

$\rightsquigarrow$ the columns of $\sqrt{n-1}\,u_i v_i^T$
= the orthogonal projections of the columns of A (the $\vec{d}_i$) onto $\text{Span}\{u_i\}$.

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

This is "the component of A in the direction of i^{th} largest variance", and the SVD says that A is decomposed as a sum of its principal components.

Running Example:

$$\vec{d}_i = \begin{pmatrix}3\\7\end{pmatrix}, \begin{pmatrix}-4\\-6\end{pmatrix}, \begin{pmatrix}7\\8\end{pmatrix}, \begin{pmatrix}1\\-1\end{pmatrix}, \begin{pmatrix}-4\\-1\end{pmatrix}, \begin{pmatrix}-3\\-7\end{pmatrix} \rightsquigarrow A = \begin{pmatrix}3 & -4 & 7 & 1 & -4 & 3\\ 7 & -6 & 8 & -1 & -1 & -7\end{pmatrix}$$

$$S = \frac{1}{6-1} AA^T = \begin{pmatrix}20 & 25\\ 25 & 40\end{pmatrix} \qquad \begin{matrix}s_1^2 = 20\\ s_2^2 = 40\end{matrix}$$

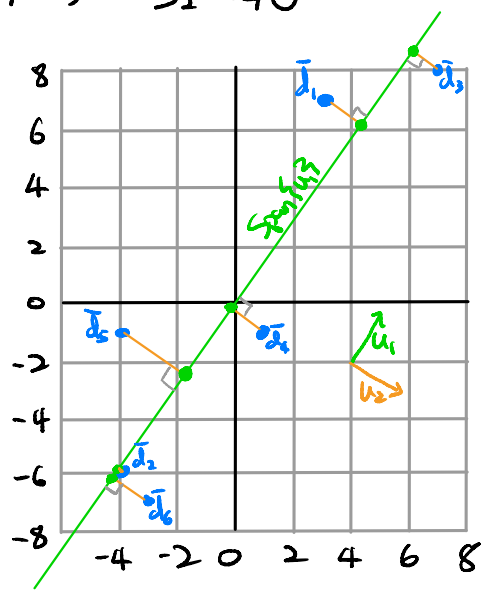
$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$ for

$\sigma_1^2 \approx 56.9 \qquad \sigma_2^2 \approx 3.1$

$$u_1 \approx \begin{pmatrix}.561\\.828\end{pmatrix} \qquad u_2 \approx \begin{pmatrix}.828\\-.561\end{pmatrix}$$

The variance is maximized in the u_1-direction, with maximum variance ≈ 56.9.

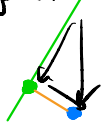
The variance is minimized in the u_2-direction, with minimum variance ≈ 3.1.



● = 1st principal comp.

／ = 2nd principal comp.

● = ● + ／ vs



NB: Total variance:

$$s^2 = \mathrm{Tr}(S) = 60 = s_1^2 + s_2^2 = 40 + 20$$
$$= \sigma_1^2 + \sigma_2^2 = 56.9 + 3.1$$

Saying that the variance in the u_2-direction is minimized means the sum of the squares of the lengths of the ⟍ is minimized. These are the orthogonal distances to $\text{Span}\{u_1\}$ = ⟋, so:

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

$$\text{error}^2 = \frac{1}{n-1}(\text{sum of length}^2 \text{ of } \diagdown) = \sigma_2^2.$$

The rest of this lecture is devoted to explaining this phenomenon, and understanding what happens when $m > 2$.

Subspace(s) of Best Fit / Matrix Approximations

As above,

$$A = \begin{pmatrix} | & & | \\ \bar{d}_1 & \cdots & \bar{d}_n \\ | & & | \end{pmatrix}$$ a centered data matrix $(m \times n)$

If $u \in \mathbb{R}^m$, $\|u\| = 1$, then

$s(u)^2 = \frac{1}{n-1}$ [sum of length^2 of projections of $\bar{d}_i$ onto Span$\{u\}$]

This definition makes sense if we replace Span$\{u\}$ with any subspace.

Def: Let V be a subspace of $\mathbb{R}^n$. The variance along V of our data points is

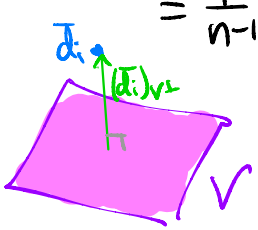
$$s(V)^2 = \frac{1}{n-1}\left[\|(\bar{d}_1)_V\|^2 + \|(\bar{d}_2)_V\|^2 + \cdots + \|(\bar{d}_n)_V\|^2\right]$$

orthogonal projections

NB: The way we've defined things, if $V = \text{Span}\{u\}$ then $s(u)^2 = s(\text{Span}\{u\})^2$

NB: $s(V^\perp)^2 = \frac{1}{n-1}\left[\|(\bar{d}_1)_{V^\perp}\|^2 + \|(\bar{d}_2)_{V^\perp}\|^2 + \cdots + \|(\bar{d}_n)_{V^\perp}\|^2\right]$

$= \frac{1}{n-1} \times$ the sum of the (orthogonal) distance^2 from the data points to V.



How to Compute $s(V)^2$:

$k = \dim V$

Find an orthonormal basis $\{w_1, w_2, \ldots, w_k\}$ for V.

Then $\boxed{s(V)^2 = s(w_1)^2 + s(w_2)^2 + \cdots + s(w_k)^2}$.

Proof: By the projection formula,

$$(\overline{d_i})_V = (\overline{d_i} \cdot w_1)w_1 + (\overline{d_i} \cdot w_2)w_2 + \cdots + (\overline{d_i} \cdot w_k)w_k$$

If we take the dot product

$$\|(d_i)_V\|^2 = (d_i)_V \cdot (d_i)_V$$

using the above formula and distribute, then $w_i \cdot w_j = 0$ for $i \neq j$ and $w_i \cdot w_i = 1$, so

$$\|(d_i)_V\|^2 = (\overline{d_i} \cdot w_1)^2 + (\overline{d_i} \cdot w_2)^2 + \cdots + (\overline{d_i} \cdot w_k)^2.$$

Therefore,

$$s(V)^2 = \frac{1}{n-1}\left[\|(d_1)_{V^\perp}\|^2 + \|(d_2)_{V^\perp}\|^2 + \cdots + \|(d_n)_{V^\perp}\|^2\right]$$

$$= \frac{1}{n-1}\Big[\boxed{(\overline{d_1} \cdot w_1)^2 + (\overline{d_1} \cdot w_2)^2 + \cdots + (\overline{d_1} \cdot w_k)^2} + \boxed{(\overline{d_2} \cdot w_1)^2 + (\overline{d_2} \cdot w_2)^2 + \cdots + (\overline{d_2} \cdot w_k)^2} + \cdots + \boxed{(\overline{d_n} \cdot w_1)^2 + (\overline{d_n} \cdot w_2)^2 + \cdots + (\overline{d_n} \cdot w_k)^2}\Big]$$

(sum by columns) $= s(w_1)^2 + s(w_2)^2 + \cdots + s(w_k)^2$

//

Eg: Let's compute $s(V)^2$ for $V=\mathbb{R}^m$ using the orthonormal basis $\{e_1, e_2, \dots, e_m\}$.

$$s(\mathbb{R}^m)^2 = s(e_1)^2 + s(e_2)^2 + \cdots + s(e_m)^2$$
$$= s_1^2 + s_2^2 + \cdots + s_m^2 = s^2$$

So $s(\mathbb{R}^m)^2 = s^2 = \mathrm{Tr}(S) =$ total variance.

Now we come to the fact that relates best fit and largest variance.

Theorem: If V is any subspace of $\mathbb{R}^m$, then

$$\boxed{s(V)^2 + s(V^\perp)^2 = s^2 = \text{total variance.}}$$

Proof: Choose an orthonormal basis $\{w_1, \dots, w_k\}$ for V ($k = \dim V$) and $w_{k+1}, \dots, w_m$ for $V^\perp$ ($m-k = \dim V^\perp$). Then $\{w_1, \dots, w_k, w_{k+1}, \dots, w_m\}$ is orthonormal since $w_i \cdot w_j = 0$ for $i \le k$ and $j > k$ because $w_i \in V$, $w_j \in V^\perp$. Hence $\{w_1, \dots, w_m\}$ is an orthonormal basis for $\mathbb{R}^m$, so

$$s(V)^2 + s(V^\perp)^2$$
$$= s(w_1)^2 + \cdots + s(w_k)^2 + s(w_{k+1})^2 + \cdots + s(w_m)^2$$
$$= s(\mathbb{R}^m)^2 = s^2.$$

//

Eg: Let's compute $s(V)^2$ for $V = \mathrm{Col}(A)$ using the orthonormal basis $\{u_1, u_2, \dots, u_r\}$.

$$s(\mathrm{Col}(A))^2 = s(u_1)^2 + s(u_2)^2 + \cdots + s(u_r)^2$$
$$= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2 = s^2.$$

This means $s(V^\perp)^2 = s(\mathrm{Nul}(A^T))^2 = 0$, which makes sense because the $\bar{d}_i$ all project to 0 in $\mathrm{Nul}(A^T) = \mathrm{Col}(A)^\perp$.

Recall that $s(V^\perp)^2 = \frac{1}{n-1} \cdot$ sum of distance^2 of $\bar{d}_i$ to V.

Def: For any dimension $k \le m$, the k-space of best fit for our data points (in the sense of orthogonal least squares) is the subspace V of dimension k that minimizes $s(V^\perp)^2 =$ error^2.

Since $s(V)^2 + s(V^\perp)^2 = s^2$ is independent of V, it follows that

$$\text{maximizing } s(V)^2 \equiv \text{minimizing } s(V^\perp)^2.$$

($s(V)^2$: variance along V; $\frac{1}{n-1}$ sum of orthogonal distance^2 to V)

In other words,

$$\text{the } k\text{-space of best fit} = \text{the } k\text{-space of largest variance.}$$

We know that $V_1 = \text{Span}\{u_1\}$ is the line of best fit, with $s(V_1)^2 = \sigma_1^2$ $s(V_1^\perp)^2 = \sigma_2^2 + \sigma_3^2 + \cdots + \sigma_r^2$.

To find the plane of best fit, we want to maximize the variance. The most we can increase $s(V_1)^2$ is by the second-largest variance σ_2^2: if $V_2 = \text{Span}\{u_1, u_2\}$ then

$$s(V_2)^2 = s(u_1)^2 + s(u_2)^2 = \sigma_1^2 + \sigma_2^2 \quad s(V_2^\perp)^2 = \sigma_3^2 + \cdots + \sigma_r^2.$$

Continuing in this way, the greedy algorithm gives the k-space of best fit:

$$V_k = \text{Span}\{u_1, u_2, \ldots, u_k\}$$

$$s(V_k)^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2 \quad s(V_k^\perp)^2 = \sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2$$

We've split the total variance $s^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2$ into the large part $= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_k^2$ and the small part $= \sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2 = \text{error}^2$.

Recall: Since the SVD of A is

$$A = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,\sigma_2 u_2 v_2^T + \cdots + \sqrt{n-1}\,\sigma_r u_r v_r^T,$$

the columns of

$$\sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,\sigma_2 u_2 v_2^T + \cdots + \sqrt{n-1}\,\sigma_k u_k v_k^T$$

are the orthogonal projections of the $\vec{d}_i$ onto $V_k = \text{Span}\{u_1, u_2, \ldots, u_k\}$.

So not only does the SVD compute the k-space of best fit for every k, it also finds the best approximations (=orthogonal projections) of your data points on those spaces!

More precisely, if

$$A_k = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,\sigma_2 u_2 v_2^T + \cdots + \sqrt{n-1}\,\sigma_k u_k v_k^T$$

then

$$A - A_k = \sqrt{n-1}\,\sigma_{k+1} u_{k+1} v_{k+1}^T + \cdots + \sqrt{n-1}\,\sigma_r u_r v_r^T$$

The columns are the orthogonal projections of the $\bar{d}_i$ onto $V_k^\perp$. The sum of the length^2 is

$$(n-1) \times \text{variance} = (n-1)\, s(V_k^\perp)^2 = (n-1)(\sigma_{k+1}^2 + \cdots + \sigma_r^2)$$

But if you add the length^2 of all the columns of a matrix, you just get the sum of the squares of the entries of that matrix!

$$B = \begin{pmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{pmatrix} = (b_{ij}) \rightsquigarrow (v_1 \cdot v_1) + \cdots + (v_n \cdot v_n) = \sum b_{ij}^2$$

So A_k is in fact the best approximation to A of rank k, in the sense of sum of squares of matrix entries!

Summary: Subspaces of Best Fit

A: $m \times n$ centered data matrix

SVD: $\frac{1}{\sqrt{n-1}} A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T$

For $k \leq m$, the k-space of best fit is

$$V_k = \text{Span}\{u_1, u_2, \ldots, u_k\} \text{ with}$$

$$\text{error}^2 = \sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2$$

The columns of

$$A_k = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,\sigma_2 u_2 v_2^T + \cdots + \sqrt{n-1}\,\sigma_k u_k v_k^T$$

are the orthogonal projections of the columns of A onto V_k.

This is the best rank-k approximation of A.

NB: This is the error^2 in the sense of variance:

$$s(V_k^\perp)^2 = \frac{1}{n-1}(\text{sum of distance}^2 \text{ of } \bar{d}_i \text{ to } V_k)$$

So the error^2 in terms of absolute distances is just $(n-1)s(V_k^\perp)^2$ = sum of squares of singular values of A (not $\frac{1}{\sqrt{n-1}} A$).

Eg: Suppose that

$$\frac{1}{\sqrt{n-1}}A = 10\,u_1v_1^T + 8u_2v_2^T + 0.2u_3v_3^T + 0.1\,u_4v_4^T.$$

- Best-fit line:

  $V_1 = \text{Span}\{u_1\}$ $\quad \text{error}^2 = 8^2 + 0.2^2 + 0.1^2 \approx 64$

- Best-fit plane:

  $V_2 = \text{Span}\{u_1, u_2\}$ $\quad \text{error}^2 = 0.2^2 + 0.1^2 \approx 0.04$

- Best-fit 3-space:

  $V_3 = \text{Span}\{u_1, u_2, u_3\}$ $\quad \text{error}^2 = 0.1^2 \approx 0.01$

Conclusion:

- Your data are not approximately collinear. The best-fit line is V_1, but error^2 is large.
- Your data are approximately coplanar. The best-fit plane is V_2, and error^2 is small.
- The 3-space of best-fit doesn't improve the error^2 by much over V_2 — adding u_3 means you're storing extra data with little increase in precision.

So the most useful approximation to the data is

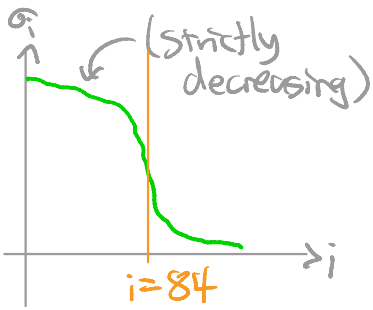
$$A \approx A_2 = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \sqrt{n-1}\,\sigma_2 u_2 v_2^T.$$

Upshot: If $\sigma_1, \sigma_2, \ldots, \sigma_k$ are much larger than $\sigma_{k+1}, \sigma_{k+2}, \ldots, \sigma_r$ then your data closely fit the k-space of best fit $V_k = \text{Span}\{u_1, \ldots, u_k\}$, and

$$A \approx A_k = \sqrt{n-1}\,\sigma_1 u_1 v_1^T + \cdots + \sqrt{n-1}\,\sigma_k u_k v_k^T$$

is a good approximation that doesn't waste a lot of memory.

When m is large, there could be thousands of singular values, so you'll see them plotted like this:



In this case, it looks like the data are approximately 84-dimensional.

(Exactly where you set your cutoff is determined by what p value you need...)

NB: If you want the best-fit k-space before recentering, just add back the means!

(Original data points $d_1, d_2, \ldots, d_n$) fit $V + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix}$ ← means

Original matrix $A_0 \approx A_k + \begin{pmatrix} \mu_1 & \cdots & \mu_1 \\ \vdots & & \vdots \\ \mu_m & \cdots & \mu_m \end{pmatrix}$