Sparse PCA

The first principal component is $z_1 = X v_1$

$$z_1 = \underbrace{\begin{bmatrix} | & | & & | & | \\ f_1 & f_2 & \dots & f_{p-1} & f_p \\ | & | & & | & | \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{2p} \end{bmatrix}}_{v_1} = v_{11} f_1 + v_{21} f_2 + \dots + v_{p-1} f_{p-1} + v_{p1} f_p$$

What if we want to select a subset of $\{f_1 \dots f_p\}$ important for summarizing the information in X?

SCoTLASS by Joliffe et. al. (2003):

$$\text{maximize } v^T S v \quad \text{subject to } \begin{cases} \|v\|^2 = 1 \\ \|v\|_1 \le c \end{cases}$$

If $c \uparrow \infty \Rightarrow$ PCA. If $c \downarrow 0 \Rightarrow v = 0$.

$1 \le c \le \sqrt{p}$ as $\|v\|_2 \le \|v\|_1 \le \sqrt{p}\,\|v\|_2$

If $v = (1, 0 \dots, 0) \Rightarrow \|v\|_1 = 1$. If $v = (\frac{1}{\sqrt{p}}, \dots, \frac{1}{\sqrt{p}}) \Rightarrow \|v\|_1 = \sqrt{p}$

Power iteration method

<u>Step 1</u> $\tilde{v} = \frac{X^T X}{n-1} v$

<u>Step 2</u> $v = \frac{\tilde{v}}{\|\tilde{v}\|_2}$

$\Longleftrightarrow$ At iteration $t+1$:

$$v^{(t+1)} = \frac{X^T X v^{(t)}}{\|X^T X v^{(t)}\|_2}$$

Penalized matrix decomposition by Witten et al (2009):

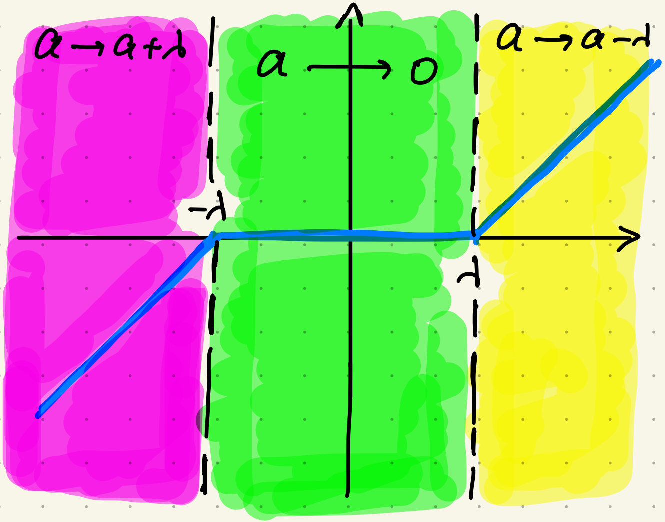
At iteration $t+1$: $v^{(t+1)} = \frac{S_\lambda(X^T X v)}{\|S_\lambda(X^T X v)\|_2}$

where $S_\lambda(a) = \text{sign}(a)(|a| - \lambda)_+$ is

soft-thesholding operator applied coordinatewise.

and λ is such that $\|v^{(t+1)}\|_1 = c$.

$$S_\lambda(a) = \operatorname{sign}(a)\,(|a| - \lambda)_+$$



Note that

- $S_\lambda(a) = a$ for $\lambda = 0$
- $S_\lambda(a/\alpha) = S_{\lambda\alpha}(a)$

$a \geq \lambda$ $\quad \operatorname{sign}(a) = 1,\ |a| = a > \lambda \Rightarrow S_\lambda(a) = 1\cdot(a-\lambda)$

$-\lambda \leq a \leq \lambda$ $\quad |a| < \lambda \Rightarrow |a| - \lambda < 0 \Rightarrow S_\lambda(a) = 0$

$a \leq -\lambda$ $\quad \operatorname{sign}(a) = -1,\ |a| = -a > \lambda \Rightarrow S_\lambda(a) = -1\cdot(-a-\lambda)$

$$= a + \lambda$$

Sparse SVD

Penalized matrix decomposition by Witten et al (2009):

Given $X \in \mathbb{R}^{n \times p}$ find $d \in \mathbb{R}$, $u \in \mathbb{R}^n$, $v \in \mathbb{R}^p$

$$\underset{u,d,v}{\text{minimize}} \; \|X - d u v^T\|_F \quad \text{subject to} \begin{cases} \|u\|_2 = 1 \\ \|u\|_1 \le c_1 \\ \|v\|_2 = 1 \\ \|v\|_1 \le c_2 \end{cases}$$

- If c_1 and c_2 are very large then u and v are singular vectors (see HW2)
- Why not just $\|u\|_1 = 1$ and $\|v\|_1 = 1$?

Let's denote by (i,j) the index with the largest $|X_{i,j}|$. Then $u = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ i \\ \\ \\ \end{matrix}$ $\quad v = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ j \\ \\ \\ \end{matrix}$ $\quad d = X_{ij}$

$$X = \begin{pmatrix} & | & \\ - & x_{ij} & - \\ & | & \end{pmatrix} \simeq \begin{pmatrix} 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & x_{ij} & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} = x_{ij} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} \\ \\ i \\ \\ \\ \end{matrix} \begin{array}{c} (0 \ldots 0 \; 1 \; 0 \ldots 0) \\ \quad\quad\; j \end{array}$$

- For fixed u you need to

  $$\underset{v}{\text{maximize}}\ u^T X v \quad \text{subject to} \begin{cases} \|v\|_2 = 1 \\ \|v\|_1 \le c_2 \end{cases}$$

- Lagrangian is

  $$\mathcal{L}(v, \lambda, \mu) = u^T X v - \lambda(\|v\|_1 - c_2) - \frac{\mu}{2}(v^T v - 1)$$

- One can show that optimal u and v are:

  $v = \dfrac{S_\lambda(X^T u)}{\mu}$ where $S_\lambda(a)$ is soft-thesholding

- To enforce $\|v\|_2 = 1$ we need

  $\mu = \|S_\lambda(X^T u)\|_2$ so $v = \dfrac{S_\lambda(X^T u)}{\|S_\lambda(X^T u)\|_2} = v(\lambda)$

- We need to find λ such that $\|v(\lambda)\|_1 \le c_2$

- If we don't have additional constraint $\|u\|_1 \leq c_1$ we will get sparse PCA.

At iteration $t+1$: <u>Step 1</u> $u^{(t+1)} = \dfrac{X v^{(t)}}{\|X v^{(t)}\|_2}$

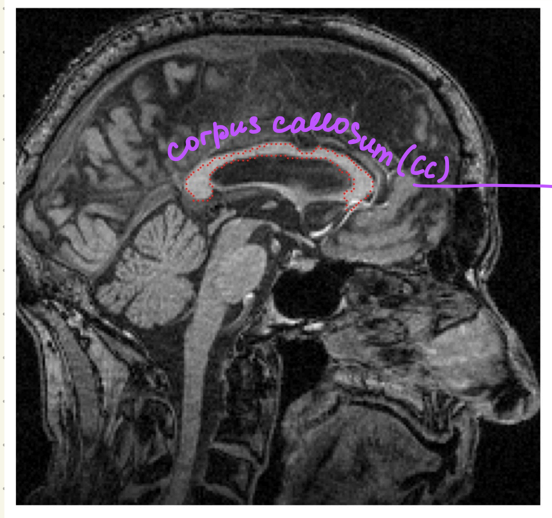
<u>Step 2</u> $v^{(t+1)} = \dfrac{S_\lambda(X^T u^{(t+1)})}{\|S_\lambda(X^T u^{(t+1)})\|_2} = \dfrac{S_{\lambda'}(X^T X v^{(t)})}{\|S_{\lambda'}(X^T X v^{(t)})\|_2}$

| $S_\lambda(X^T X v / \|Xv\|_2) = S_{\lambda'}(X^T X v)$ for $\lambda' = \lambda \cdot \|Xv\|_2$

- Note that if $\lambda = 0$ then $v^{(t+1)} = \dfrac{X^T X v^{(t)}}{\|X^T X v^{(t)}\|_2}$ that is, power iteration method.

- After finding u_1, d_1, v_1 you can appey method to $\tilde{X} = X - d_1 u_1 v_1^T$ and find $u_2, d_2, v_2 \ldots$

Sparse PCA on shapes (from ESLII)



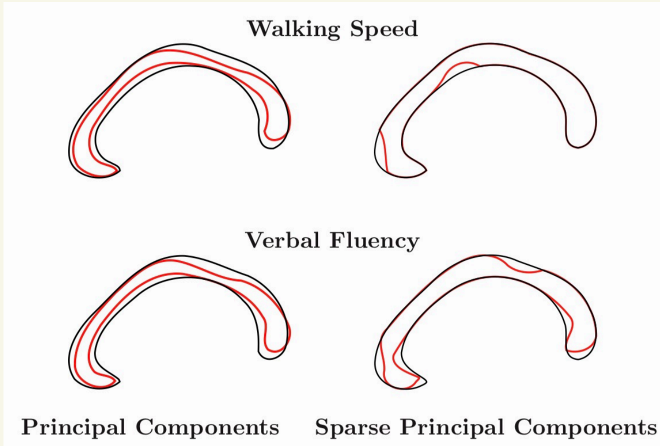
CC annotated with landmarks

$$\begin{pmatrix} x_1 & y_1 \\ \cdots & \\ x_N & y_N \end{pmatrix} \longrightarrow (x_1\, y_1 \dots x_N\, y_N)$$

$$X = \begin{pmatrix} \text{———} \\ \text{———} \\ \text{———} \\ \text{———} \\ \text{———} \end{pmatrix}$$ n = 569 elderly people

black: the mean CC shape

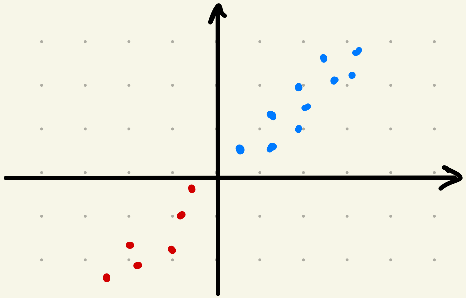
red: PC loading vectors

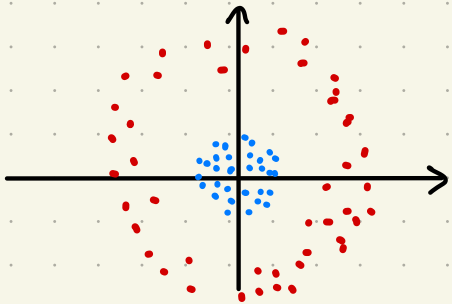


Kernel PCA

The main limitation of PCA is linearity

$$z_i = X v_i = \begin{pmatrix} | & & | \\ f_1 & \ldots & f_p \\ | & & | \end{pmatrix} \begin{pmatrix} v_{1i} \\ \vdots \\ v_{pi} \end{pmatrix} = f_1 v_{1i} + \ldots + f_p v_{pi}$$ (linear function of $f_1 \ldots f_p$)

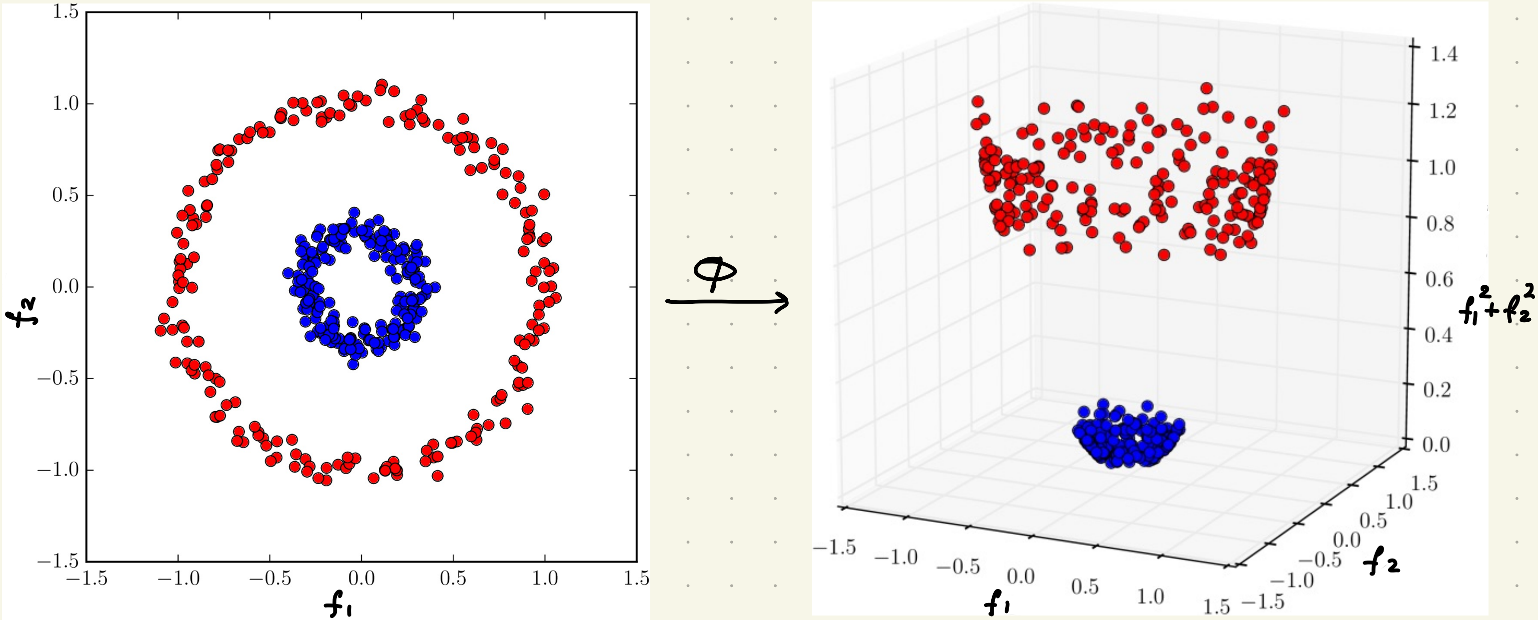




Idea: transform the feature space, each observation $x_i \in \mathbb{R}^p$ becomes $\Phi(x_i) \in \mathbb{R}^q$ (typically, $q > p$)

Then PCA for $\Phi(x_1) \ldots \Phi(x_n)$ is non-linear for $x_1 \ldots x_n$.

Example: $\Phi(f_1, f_2) = (f_1, f_2, f_1^2 + f_2^2)$



https://atiulpin.wordpress.com/2015/04/02/a-tutorial-on-kernel-principal-component-analysis/

Denote by $\Phi = \begin{pmatrix} -\Phi(x_1)^T- \\ \cdots\cdots \\ -\Phi(x_n)^T- \end{pmatrix} \in \mathbb{R}^{n\times q}$ the transformed data

and $K = \Phi\Phi^T \in \mathbb{R}^{n\times n}$ the inner product matrix,

i.e. $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$.

① Assume that Φ is centered.

We can compute PCs using K only (not Φ)!

- Eigenvectors of $S_\Phi = \frac{1}{N-1}\Phi^T\Phi$ live in the row-space of Φ, i.e. $v = \Phi^T\alpha$ for some $\alpha \in \mathbb{R}^n$.

$S_\Phi v = \frac{1}{N-1}\Phi^T(\Phi v) = \lambda v$

- α is an eigenvector of K.

$\frac{1}{N-1}\Phi\Phi^T\Phi\Phi^T\alpha = \lambda\Phi\Phi^T\alpha \Rightarrow K^2\alpha = \underbrace{(N-1)\lambda}_{\lambda'} K\alpha$

Solve $K\alpha = \lambda' \cdot \alpha$

- α should be normalized by the sqrt of the e. value

| $\|v\|^2 = \alpha^T \Phi \Phi^T \alpha = \alpha^T K \alpha = \lambda' \|\alpha\|^2 = 1 \Rightarrow \|\alpha\| = \frac{1}{\sqrt{\lambda'}}$

- The PC scores for Φ are just $K\alpha$

| $z = \Phi v = \Phi \Phi^T \alpha = K\alpha.$

- To project $\varphi(x)$ onto the PC direction v we need to know α and $\langle \varphi(x_i), \varphi(x) \rangle$ only.

| $v^T \varphi(x) = \alpha^T \Phi \cdot \varphi(x) = \alpha^T \begin{pmatrix} \langle \varphi(x_1), \varphi(x) \rangle \\ \vdots \\ \langle \varphi(x_n), \varphi(x) \rangle \end{pmatrix}$

② If Φ is not centered then replace K by $\tilde{K} = CKC$ where $C = I - \frac{11^T}{n}$.

| $\tilde{\Phi} = \Phi C$ so $\tilde{K} = \tilde{\Phi} \tilde{\Phi}^T = C \Phi \Phi^T C = CKC$

Kernel PCA

→ compute K

→ compute the top eigenvalue of K (λ')

→ find the top eigenvector of K (α) and scale $\|\alpha\| = \frac{1}{\sqrt{\lambda'}}$

→ find scores $z = K\alpha$.

KPCA relies only on $K(x,y) = \langle \varphi(x), \varphi(y) \rangle$, that is called kernel function.

Examples:

Quadratic kernel : $K(x,y) = (1 + \langle x,y \rangle)^2$

$K(x,y) = 1 + x_1^2 y_1^2 + \ldots + x_p^2 y_p^2 + 2x_1y_1 + \ldots + 2x_py_p = \langle \varphi(x), \varphi(y) \rangle$ for $\varphi(x) = \begin{pmatrix} 1 \\ x_1^2 \\ \vdots \\ x_p^2 \\ \sqrt{2}\,x_1 \\ \vdots \\ \sqrt{2}\,x_p \end{pmatrix}$

Polynomial kernel: $K(x,y) = (1 + \langle x,y \rangle)^d$

Radial kernel: $K(x,y) = e^{-\gamma \|x-y\|^2}$

Local dimension reduction methods

tSNE (t-distributed stochastic neighbor embedding)

Given points $x_1 \dots x_n \in \mathbb{R}^p$

- compute distances $\|x_i - x_j\|^2$
- compute probabilities p_{ij} of selecting neighbors (i,j) (use Gaussian distribution)

Given embedding $z_1 \dots z_n \in \mathbb{R}^q$ $(q < p)$

- compute distances $\|z_i - z_j\|^2$
- compute probabilities q_{ij} of selecting neighbors (i,j) (use t-distribution)

Find $z_1 \dots z_n$ such that p_{ij} and q_{ij} are "Similar" (use KL divergence)

Main parameter:

- perplexity, balances local and global attention

tSNE vs PCA:

\+ Non-linear, good for visualization, captures local neighbours

\- Slow, struggles with noise, less interpretable:

- no meaning of the tSNE coordinates and distances
- small distances are informative
- cluster sizes are not informative
- distances between clusters are not informative

UMAP (Uniform Manifold Approximation and Projection)

given points $x_1 \ldots x_n \in \mathbb{R}^p$

- constructs a weighted k-neighbour graph
  (use "fuzzy simplicial complex")

given embedding $z_1 \ldots z_n \in \mathbb{R}^q$ $(q < p)$

- constructs a weighted k-neighbour graph
  (use "fuzzy simplicial complex")

Find $z_1 \ldots z_n$ such that the graphs are "similar"
(use cross-entropy)

Two main parameters:

- n_neighbors, the number of nearest neighbours
- min_dist, how tightly UMAP packs neighbours

Comparing to tSNE, UMAP is faster and better at preserving more global structure

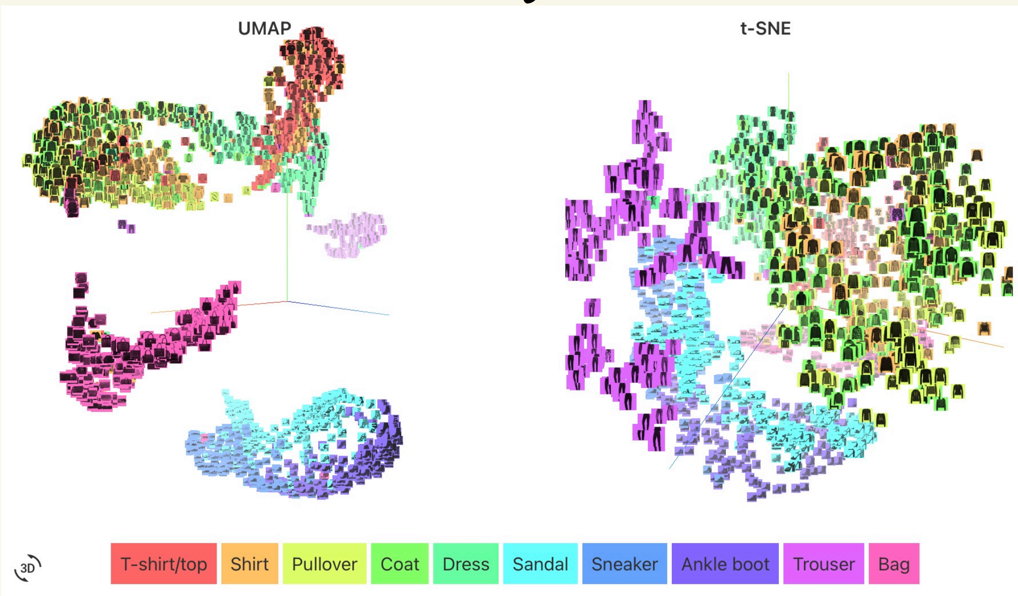


Figure 2: *Dimensionality reduction applied to the Fashion MNIST dataset. 28x28 images of clothing items in 10 categories are encoded as 784-dimensional vectors and then projected to 3 using UMAP and t-SNE.*

https://pair-code.github.io/understanding-umap/

Practical aspects:

Sometimes it's better to combine PCA & tSNE/UMAP.

- Filter the data
- Do PCA, reduce dimensionality and noise
- Plot with UMAP/tSNE, try various hyperparameters.

Perturbation theory for PCA

Given $S \in \mathbb{R}^{p\times p}$, consider $\hat{S} = S + E \in \mathbb{R}^{p\times p}$ where $E \in \mathbb{R}^{p\times p}$ is a symmetric noise matrix.

Denote the eigendecompositions by $S = U\Lambda U^T$, $\hat{S} = \hat{U}\hat{\Lambda}\hat{U}^T$.

Recall the definition of the spectral norm:

$$\|A\|_2 = \sqrt{\lambda_1(A^TA)} = \sigma_1(A).$$

<u>Thm</u> (Weyl's) $\max_{i=1..p} |\lambda_i - \hat{\lambda}_i| = \|\Lambda - \hat{\Lambda}\|_2 \leq \|E\|_2$, i.e.

eigenvalues are stable under perturbation.

$$\hat{\lambda}_1 = \max_{\|v\|=1} v^T \hat{S} v = \max_{\|v\|=1} (v^TSv + v^TEv) \leq$$

$$\leq \lambda_1 + \max_{\|v\|=1} |v^TEv| = \lambda_1 + \|E\|_2$$

$$|\hat{\lambda}_1 - \lambda| \leq \|E\|_2$$

Eigenvectors are not stable!

<u>Example</u>: $S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $E = \begin{pmatrix} 0 & \varepsilon \\ \varepsilon & 0 \end{pmatrix}$, $\hat{S} = \begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$

$u_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $d_1 = 1$; $u_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $d_2 = 1$

$\hat{u}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $d_1 = 1 + \varepsilon$; $\hat{u}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $d_2 = 1 - \varepsilon$

Measuring agreement between U and $\hat{U}$ is tricky.

Recall $U_{(r)} = (u_1 \dots u_r) \in \mathbb{R}^{p \times r}$ and $\hat{U}_{(r)} = (\hat{u}_1 \dots \hat{u}_r) \in \mathbb{R}^{p \times r}$

- Why not $\| U_{(r)} - \hat{U}_{(r)} \|_F$?

  If $d_1 = \dots = d_p$ then U is any orthogonal matrix.

- We need to measure the agreement between

  $\mathcal{U} = \mathrm{span}(u_1, \dots, u_r)$ and $\hat{\mathcal{U}} = \mathrm{span}(\hat{u}_1, \dots, \hat{u}_r)$

Principal angles between subspaces

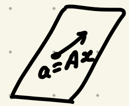
Consider $A, B \in \mathbb{R}^{n\times r}$ such that $A^TA = B^TB = I$.

Denote $\mathcal{A} = \text{span}(a_1,\dots,a_r)$ $\quad \mathcal{B} = \text{span}(b_1,\dots,b_r)$

Then principal angle between $\mathcal{A}$ and $\mathcal{B}$ is

$$\theta_1 = \angle(\mathcal{A},\mathcal{B}) = \arccos(\sigma_1(A^TB))$$

$$\angle(\mathcal{A},\mathcal{B}) = \min_{\substack{a\in\mathcal{A},\, b\in\mathcal{B}\\ \|a\|=\|b\|=1}} \arccos(a^Tb) = \min_{\substack{x,y\\ \|x\|=\|y\|=1}} \arccos(x^TA^TBy)$$



$\mathcal{A}$

The general statement is

$$A^TB = U \cos\Theta \, V^T \quad \text{where } \cos\Theta = \begin{pmatrix} \cos\theta_1 & & \\ & \ddots & \\ & & \cos\theta_r \end{pmatrix}$$

and $\theta_1 \dots \theta_r$ are called principal angles.

Distance Between Subspaces

Define the distance between $\mathcal{A}$ and $\mathcal{B}$ as

$d(\mathcal{A}, \mathcal{B}) = \|\sin\theta\|_F$

Let $P_A = AA^T$, $P_B = BB^T$ denote the projection operators and $A_\perp$ and $B_\perp$ are orthogonal complements.

Then $d(\mathcal{A}, \mathcal{B}) = \frac{1}{\sqrt{2}} \|P_A - P_B\|_F = \|A^T B_\perp\|_F$.

$$I = A^T A = A^T (BB^T + B_\perp B_\perp^T) A = U \cos^2\theta\, U^T + A^T B_\perp B_\perp^T A$$

$$A^T B_\perp B_\perp^T A = I - U\cos^2\theta\, U^T = U(I - \cos^2\theta) U^T = U \sin^2\theta\, U^T$$

$$tr(U \sin^2\theta\, U^T) = tr(\sin^2\theta) = \|\sin\theta\|_F^2 = tr(A^T B_\perp B_\perp^T A) = \|A^T B_\perp\|_F^2$$

$$\|P_A - P_B\|_F^2 = \|AA^T - BB^T\|_F^2 = tr(AA^TAA^T) - 2\,tr(AA^TBB^T) + tr(BB^TBB^T)$$

$$= r - 2\,tr(A^TBB^TA) + r = 2r - 2\,tr(I - A^TB_\perp B_\perp^T A) = 2\,tr(A^TB_\perp B_\perp^T A)$$

Davis-Kahan theory

Denote by δ the eigengap, i.e.

$$\delta = \min_{1 \leq i \leq r,\, r+1 \leq j \leq p} |\lambda_i(S) - \lambda_j(\tilde{S})| > 0$$

Then $$d(U_{(r)}, \hat{U}_{(r)}) \leq \frac{\|E\|_F}{\delta}.$$