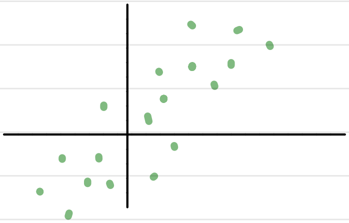
Singular Value Decomposition

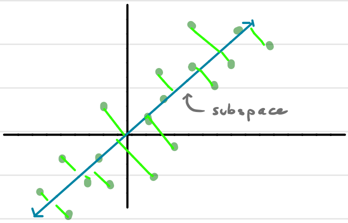
Principal axes: let $y_1 \dots y_n \subset \mathbb{R}^p$.



The data are p-dimensional, but the point cloud may be "close" to a lower-dimensional (linear) space.

Goal: ① Find the "best" low dimensional subspace
② Find the approximation $\tilde{y}_i$ of y_i for $i = 1 \dots n$.

Intuitively:



Start with question ② first.

Let $L \subset \mathbb{R}^p$ be an r-dimensional subspace
Let $B = [b_1 \dots b_r]$ be an orthonormal basis for L

This means ① $\forall\, x \in L \;\; \exists\, \underset{\sim}{a} \in \mathbb{R}^r : \; x = a_1 b_1 + \dots + a_r b_r$ $\{$ B is a basis
$= B\, a$

② $b_j^T b_j = 1, \; b_j^T b_k = 0, \; B^T B = I_r$ $\{$ B is orthonormal
$B \in \mathcal{V}_{r,p}$

Task: for $y_i \in \mathbb{R}^p$, find $\hat{y}_i \in L$ to minimize $\|y_i - \hat{y}_i\|^2$.

Solution: $\hat{y}_i \in L \iff \hat{y}_i = Ba$ some $a \in \mathbb{R}^r$

$$\|y_i - \hat{y}_i\|^2 = \|y_i - Ba\|^2 = (y_i - Ba)^T (y_i - Ba)$$

$$= y_i^T y_i - 2 y_i^T B a + a^T B^T B a$$

$$= a^T a - 2 a^T B^T y_i + c$$

Minimize in a with calculus: $\nabla_a \|y_i - Ba\|^2 = 2a - 2B^T y_i$

optimal a satisfies $\quad a = B^T y_i$

(Take 2nd deriv to show a min)

So $\hat{y}_i = Ba = BB^T y_i$ (compare to $\hat{y}$ in regression)

(BB^T is the proj. matrix onto L, the linear space spanned by the r columns of B)

Matrix form:

$$Y = \begin{pmatrix} y_1^T & \text{---} \\ \vdots & \\ y_n^T & \text{---} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

$$\hat{Y} = \begin{pmatrix} y_1^T BB^T \\ \vdots \\ y_n^T BB^T \end{pmatrix} = \underbrace{\begin{pmatrix} y_1^T B \\ \vdots \\ y_n^T B \end{pmatrix}}_{n \times r} \underbrace{B^T}_{r \times p} \in \mathbb{R}^{n \times p}$$

$$= YBB^T$$

Now solve problem 1 Find optimal subspace / optimal B

Task: minimize $\|Y - YBB^T\|^2$ in $B \in \mathbb{R}^{p\times r}$; $B^TB = I_r$

Simpler task: min $\|Y - Ybb^T\|^2$ in $b \in \mathbb{R}^p$; $b^Tb = 1$

Solution: $\|Y - Ybb^T\|^2 = \|Y(I - bb^T)\|^2$

$$= tr\left(Y(I-bb^T)(I-bb^T)Y^T\right)$$
$$= tr\left(Y(I-bb^T)Y^T\right)$$
$$= tr(YY^T) - tr(Ybb^TY^T)$$
$$= \|Y\|^2 - b^TY^TYb$$

$tr(S) = \Sigma s_{ii}$
$tr(A^TA) = \Sigma\Sigma a_{ij}^2$

Minimizing in b means maximizing b^TY^TYb in b.

Try taking a derivative: $\frac{\partial}{\partial b} b^TY^TYb = 2Y^TYb = 0 \iff b = 0$

forgot $b^Tb = 1$: $\frac{\partial}{\partial b}\left[b^TY^TYb - \lambda(b^Tb - 1)\right] = 0$

$$\iff 2Y^TYb - \lambda 2b = 0 \iff Y^TYb = \lambda b$$

$\Rightarrow$ critical point when b is an eigenvector of Y^TY.

$\Rightarrow$ critical points of $RSS(b) = \|Y - Ybb^T\|^2$ are λ-vecs of Y^TY

<u>Spectral Decomposition Thm:</u> Let $S \in \mathbb{R}^{p \times p}$ be symmetric, real.

Then $S = V \Lambda V^T$ for some $V \in \mathbb{R}^{p \times p} : V^T V = V V^T = I_p$

some $\Lambda = \mathrm{diag}(\lambda_1, \ldots \lambda_p)$, $\lambda_1 \geq \cdots \geq \lambda_p$

<u>Corollary:</u> Let $V = \begin{bmatrix} v_1 & \cdots & v_p \\ | & & | \end{bmatrix}$. Then v_j is an evec of S, with eval λ_j.

<u>Proof:</u> $S v_j = V \Lambda V^T v_j = V \Lambda \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = V \begin{bmatrix} 0 \\ 0 \\ \lambda_j \\ 0 \end{bmatrix} = \lambda_j v_j$.

<u>Important result:</u> Let $S = Y^T Y$. Then $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

(so $Y^T Y$ is positive semidefinite)

<u>Exercise:</u> Prove result.

Q: Which evec gives optimal approximation?
A: Minimizer of $\|Y - Ybb^T\|$, maximizer of $b^T Y^T Y b$.

Let $v_1 \dots v_p$ be evecs of $Y^T Y$ w/ evals $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$

then $v_j^T Y^T Y v_j = v_j^T [\lambda_j v_j]$ (v_j is evec w/ eval λ_j)

$= \underline{\underline{\lambda_j}}$

In other words, $\hat{Y} = Y v_1 v_1^T$ is the <u>best</u> one dim approx to Y.

Similarly, $\hat{Y} = Y BB^T$, $B = \begin{bmatrix} v_1 & \cdots & v_r \\ | & & | \end{bmatrix}$ is <u>best</u> r-dim approx.

<u>Proof</u>:
1) induction
2) matrix differentials
3) brute-force with Lagrange multipliers

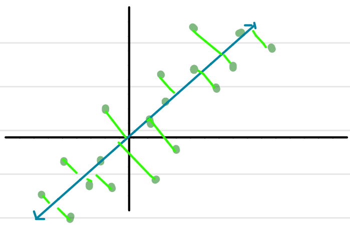
<u>In summary</u>:

<u>Goal</u>: Approx $\underset{\sim}{x}_i$ as $\underset{\sim}{x}_i \approx a_{i1} \underset{\sim}{b}_1 + \dots + a_{ir} \underset{\sim}{b}_r$, some <u>common</u> $\underset{\sim}{v}_1 \dots \underset{\sim}{v}_r$
$= B \underset{\sim}{a}_i$

$$\text{Approx } Y = \begin{pmatrix} \underset{\sim}{x}_1^T - \\ \vdots \\ \underset{\sim}{x}_n^T - \end{pmatrix} \approx \begin{pmatrix} \underset{\sim}{a}_1^T B^T \\ \vdots \\ \underset{\sim}{a}_n^T B^T \end{pmatrix} = \begin{bmatrix} \underset{\sim}{a}_1^T - \\ \vdots \\ \underset{\sim}{a}_n^T - \end{bmatrix} B^T = \underset{n \times r}{A} \; \underset{r \times p}{B^T}$$

<u>Result</u>: A minimizer of $\|Y - AB^T\|^2$ is
(princ axes) $B = [v_1 \cdots v_r]$ $\quad v_j = \text{evec}_j (Y^T Y)$
(princ coord) $A = YB$ = Coef of projection of rows of Y onto evecs of $Y^T Y$.

Recall picture:



Singular value decomposition:

evecs of Y^TY

Best $\begin{cases} p\text{-dim approx to rows of } Y \\ \text{rank-}p \text{ approx to } Y \end{cases}$ is $\hat{Y} = YVV^T$ $(VV^T = I)$

We can write $\hat{Y} = FV^T$, $F = YV \in \mathbb{R}^{n \times p}$

What are properties of F?

$$F^TF = V^TY^TYV = V^T[V\Lambda V^T]V = \Lambda \quad p \times p \text{ diagonal}$$

Claim: $F^TF = \Lambda \Rightarrow F = \underset{n \times p}{U}\Lambda^{1/2}$, some $U \in \mathbb{R}^{n \times p}$: $U^TU = I_p$.

Exercise: Prove claim.

This means $\hat{Y}$ can be written as
$$\hat{Y} = FV^T = U\Lambda^{1/2}V^T = UDV^T$$

where $U^TU = V^TV = I_p$, $D = \text{diag}(d_1, \dots, d_p)$, $d_1 \geq \dots \geq d_p \geq 0$.

This representation is called the singular value decomposition (SVD) of $\hat{Y}$.

Thm (SVD): Let $Y \in \mathbb{R}^{n \times p}$, $n \geq p$. Then Y can be written as

$Y = UDV^T$, where
1. $U \in \mathbb{R}^{n \times p}$, $U^T U = I_p$
2. $V \in \mathbb{R}^{p \times p}$, $V^T V = I_p$
3. $D = \text{diag}(d_1 \ldots d_p)$ $d_1 \geq \cdots \geq d_p \geq 0$.

Comments

* $\underset{\sim}{u}_1 \ldots \underset{\sim}{u}_p$ are the left sing. vecs,
* $\underset{\sim}{v}_1 \ldots \underset{\sim}{v}_p$ are the right sing vecs
* $d_1 \ldots d_p$ are the singular values
* $Y^T Y = V D U^T U D V^T = V D^2 V^T = V \Lambda V^T$ (think: col. cov)
  - columns of V are evecs of $Y^T Y$
  - D^2 gives evals of $Y^T Y$
* $Y Y^T = U D V^T V D U^T = U D^2 U^T = U \Lambda U^T$ (think: row cov)
  - colums of U are evecs of $Y Y^T$
  - D^2 are evals of $Y Y^T$
  - $Y Y^T$ also has $n-p$ additional evecs, with evals $= 0$.

$\Rightarrow \text{evec}(Y Y^T) = [U \; U^\perp]$, $\quad Y Y^T U = U \Lambda$

$Y Y^T U^\perp = 0 = U^\perp \cdot 0$

Reduced-rank matrix approximation

"Vector" view: data are $y_1 \dots y_n \subset \mathbb{R}^p$

approximate y_i with $\hat{y}_i = B\underset{\sim}{a}_i = b_1 a_{i1} + \dots + b_p a_{ip}$

$\Rightarrow$ B is the subspace in which $y_1 \dots y_n$ vary

$\Rightarrow$ variation in $y_1 \dots y_n \subset \mathbb{R}^p$ rep by var. in $\underset{\sim}{a}_1 \dots \underset{\sim}{a}_n \subset \mathbb{R}^r$.

Matrix view: data are $Y \in \mathbb{R}^{n \times p}$

approximate $\underset{n \times p}{Y}$ with $\hat{Y} = \underset{n \times r}{A} \underset{r \times p}{B^T}$

np numbers $\qquad < nr + rp = r(n+p)$ numbers

Matrix rank

Let $X \in \mathbb{R}^{m \times n}$. Write $X = \begin{bmatrix} x_{,1} & x_{,2} & \cdots & x_{,n} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_{1,} & — \\ x_{2,} & — \\ \vdots & \\ x_{m,} & — \end{bmatrix}$

$\text{rank}(X) = \dim(\text{span}(x_{,1}, \dots x_{,n}))$

$= \dim(\text{span}(x_{1,}, \dots x_{m,}))$

so $\text{rank}(X) \leq \min(m,n)$

$=$ smallest r for which $\exists\, A \in \mathbb{R}^{m \times r}$, $B \in \mathbb{R}^{n \times r}$, with $X = AB^T$.

$= \underset{m\times 1}{a_1} \underset{1 \times n}{b_1^T} + \dots + \underset{m \times 1}{a_r} \underset{1 \times n}{b_r^T}$

Thm (rank 1 matrix): Let $X = \underset{\sim}{a}\underset{\sim}{b}^T = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \\ a_2 b_1 & \ddots & \end{bmatrix}$. Then $\text{rank}(X) \leq 1$

Ex: Prove, using 1st 2 definitions of rank.