Lecture 6 PCA and SVD

Note by Samion Suwito on 2/6/25

Last Time

Symmetric and PSD/PD matrices
$A \in \mathbb{S}^n \implies A = U\Lambda U^\intercal$ for diagonal Λ and orthogonal Λ
Orthonormal basis is a rotation of a basis and matrix A scales the basis

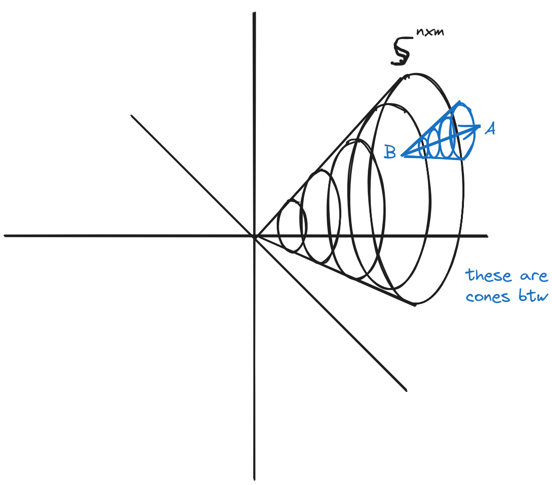
$\{x : x^\intercal Ax\}$ $A > 0$ Represented as $x = \sum \alpha_i u_i$

$$\begin{aligned} \left(\sum \alpha_i u_i\right)^\intercal x \left(\sum \alpha_i u_i\right) &\leq 1 \\ = \left(\sum \alpha_i u_i\right)^\intercal \left(\sum \alpha_i \lambda_i u_i\right) \\ = \sum \lambda_i \alpha_i^2 \leq 1 \end{aligned}$$

The norm balls are elipsces and the lengths are inversely proportional to λ_i^2

PSD makes form a (convex) cone

$\implies$ if $A, B \geq 0 \implies \theta A + (1-\theta)B \geq 0 \; \forall \theta \in [0,1]$
If $A \geq 0 \implies \alpha A \geq 0 \;\; \forall \alpha \geq 0$
The cone by the statements above define an *order* on set of PSD matrices called the lorner order. Namely for $A, B \geq 0$ if $A - B \geq 0$



Remark: A useful criterion for checking whether a matrix is PSD, is via "*Schur Complements*"

$$M = \begin{bmatrix} A & X \\ X^\intercal & B \end{bmatrix}$$

where B is invertible symmetric and A is symmetric.
Schur complement: $S := A - XBX^\intercal$.
Then $M > 0 \iff A > 0 \text{ and } S > 0$ and $M \geq 0 \iff A \geq 0 \text{ and } S \geq 0$.

Principal Component Analysis

Application to PCA:
PCA is an ”unsupervised” technique for data exploration/visualisation.

John Novembre: Map of Europe.

Sequenced the genomes of individuals in Europe. And was able to map the vectorised genomes to a map of Europe

Given centred data vectors $x^{(1)} \dots x^{(m)}$ $\left(\frac{1}{m}\sum_{i=1}^m x^{(i)} = 0\right)$
Goal: Find a ”direction” u (unit vector) that “best” explains data.

PCA approach:

$$\min_{u:\|u\|_2 \leq 1} \sum_{i=1}^m \|\pi_{\mathrm{Sp(u)}}(x^{(i)}) - x^{(i)}\|_2^2$$

Choose one directional subspace that minimises this total error.

$$= \sum_{i=1}^m \left(u^\intercal x^{(i)} (u - x^{(i)})^\intercal\right) \left(\left(u^\intercal x^{(i)}\right) u - x^{(i)}\right)$$

$$= \sum_{i=1}^m \left(u^\intercal x^{(i)}\right)^2 - 2\left(u^\intercal x^{(i)}\right)\left(u^\intercal x^{(i)}\right) + x^{(i)\intercal} x^{(i)}$$

$$\sum_{i=1}^m \|x^{(i)}\|_2^2 - u^\intercal \left(\sum_{i=1}^m x^{(i)} x^{(i)\intercal}\right) u$$

Therefore since the first part is constant the optimization problem of finding u is equivalent to

$$\max_{u:\|u\|_2 = 1} \; u^\intercal C u$$

Where $C = \sum x^{(i)} x^{(i)\intercal} = XX^\intercal$.

Solution is going to be eigenvector with maximum eigenvalue. Therefore optimal u is eigenvector of C corresponding to $\lambda_{\max}(C)$
To visualise *two dimensions* subtract $\pi_{\mathrm{Sp}(u^*)}(x^{(i)})$ from $x^{(i)}$ then repeat the procedure on new dataset $\tilde{x}^{(i)} = x^{(i)}\pi_{\mathrm{Sp}(u^*)}(x^{(i)})$ Next principal component turns out to be u^n corresp. Let $C = U\Lambda U^\intercal = \sum \lambda_i u^{(i)} u^{(i)\intercal}$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ to "visualise" data in 2-dim and then plot points $\left(u^{(1)\intercal} x^{(i)}, u^{(2)\intercal} x^{(i)}\right)_{i=1...m}$

Singular Value Decomposition

Question: What is generalisation of Spectral theorem for generic $A \in \mathbb{R}^{n \times m}$
SVD

Theorem

Every $A \in \mathbb{R}^{m \times n}$ can be represented as:

$$A = U\Sigma V^\intercal$$

Where U is an $m \times m$ orthogonal matrix, V is an $n \times n$ orthogonal matrix and Σ is an $m \times n$ and $\sigma_i > 0$ and r is rank(A)

$$\Sigma = \begin{bmatrix} \sigma_1 & \dots & 0 & \vdots \\ \vdots & \ddots & \vdots & 0 \\ 0 & \dots & \sigma_r & \vdots \\ \cdots & 0 & \cdots & 0 \end{bmatrix}$$

and $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$

"*Compact SVD*" = $A = U_r \Sigma_r V_r^\intercal$
Where $U_r =$ first r columns of U. V_r is first r columns of V_r and Σ_r is upper left most $r \times r$ of Σ.

$$[U_r \quad U_{n-r}] \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix} [V_r \quad V_{n-r}]^\intercal = U_r \Sigma_r V_r^\intercal$$

making compact SVD the same being full and also he doesn't care about which basis vectors you use.

Understanding the SVD

$A = U\Sigma V^\top = U_r \Sigma_r V_r^\top = \sum_{i=1}^r \sigma_i u_i v_i^\top$ where x_i is the ith column of X
$Av_i = \sigma_i u_i$
Then vectors v_i represent a basis in $\mathbb{R}^n$ which gets sent by A to $\mathbb{R}^m$ to $\sigma_i u_i$ where u_i is a basis of $\mathbb{R}^m$

Connection to Spectral Decomposition

$A^\top A$ for generic $A \in \mathbb{R}^{m\times n}$ is symmetric matrix. So Is $AA^\top$. Using SVD we can write $A^\top A$ as

$$V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma^\top \Sigma V^\top$$

And $\Sigma^\top \Sigma$ looks like $n \times n$ matrix

$$\begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_r^2 \end{bmatrix}$$

Giving the eigenvalues of $A^\top A$
Then $AA^\top = U\Sigma\Sigma^\top U^\top$ where $\Sigma\Sigma^\top$ is $m \times m$ diagonal ($\sigma_1^2, \dots, \sigma_r^2, 0, \dots$) which is the spectral decomposition of $AA^\top$

Idea behind proof of SVD:

$$v_1 = \arg\max_{v:\|v\|_2=1} \|Av\|_2$$

Write $u_1 = \frac{Av_1}{\sigma_1}$ and since we know $\|v\|_2 = 1$ where $\sigma_1 = \|Av_1\|_2$ Now write $A' = A - u_1 v_1^\top \sigma_1$, and repeat to find σ_2, v_2, u_2

Connection to Matrix Norms

Frobinias Norm =

$$\|A\|_F^2 = \sum_{ij} |A_{ij}|^2 = \operatorname{Tr}\left(A^\top A\right) = \sum_{i=1}^n \lambda_i \left(A^\top A\right) =$$

$$\sum_{i=1}^r \sigma_i^2$$

Also by construction also: $\|A\|_2 = \sigma_1$

Nuclear Norm

$$\|A\|_* = \sum_{j=1}^{r} \sigma_i$$

Other Interpretations of Singular Values

For $A \in \mathbb{R}^{n \times m}$, the “condition number”is $\kappa(A) = \frac{\sigma_1}{\sigma_n}$ and it measures sensitivity of solutions $Ax = b$ when b varies. How much x changes when b changes.