Mathematical basis for PCA:

Singular value decomposition (SVD)

Eckart & Young (39.)

If B has rank M

then $\| A - B \| \geq \| A - A_M \|$

$$A_M = \sum_{j=1}^{M} \sigma_i \, \vec{u}_i \, \vec{v}_i^T$$

where $\vec{u}_i$ are the right eigenvectors of A

$\vec{v}_i$ are the left eigenvectors of A

σ_i are the singular values of A

These definitions follow from the SVD

SVD

$A_{n \times n}$

$$A V = V \Lambda$$

$$V = (\vec{v}_1 \cdots \vec{v}_n)$$

$$\Lambda = \mathrm{diag}(\lambda_i)$$

$$A = V \Lambda V^{-1}$$

Analogue to Diagonalization of square matrices

↳ but for rectangular matrices.

SVD

Analysis for rectangular matrices:

Given $A_{m\times n} \in \mathbb{R}^{m\times n}$

$$\boxed{A_{m\times n} V_{n\times n} = U_{m\times m} \Sigma_{m\times n}}$$

$$V_{n\times n} = \left(\vec{v}_1 \ \cdots \ \vec{v}_n\right)$$

$$U_{m\times m} = \left(\vec{u}_1 \ \cdots \ \vec{u}_m\right)$$

$$\Sigma_{m\times n} = \left(\begin{array}{c|c} \begin{matrix}\sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r\end{matrix} & 0 \\ \hline 0 & 0 \end{array}\right)$$

Singular values.

$$\min(m,n) \geq r$$

---

r is the rank of A (at most $\min(m,n)$ but could be smaller.)

Reduced form of the SVD:

$$A_{m\times n} = U_{m\times m} \Sigma_{m\times n} V^T_{n\times n} = \sum_{j=1}^{r} \sigma_j \vec{u}_j \vec{v}_j^T = \underbrace{(U_r)}_{m\times r} \underbrace{\Sigma_r}_{(r\times r)} \underbrace{(V_r^T)}_{r\times n}$$

Truncations of U and V taking the first r vectors $\vec{u}_j$ and $\vec{v}_j$.

What are these U and V?

Definition of SVD

$$AV = U\Sigma \qquad VV^T = I \qquad UU^T = I$$

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

$$\underset{n\times n}{(A^TA)} = V\Sigma^T U^T U \Sigma V^T = V(\Sigma^T\Sigma)V^T$$

$$\underset{m\times m}{(AA^T)} = U\Sigma V^T V \Sigma^T U^T = U(\Sigma\Sigma^T)U^T$$

$V_{n\times n}$ contains the eigenvectors of (A^TA)

$U_{m\times m}$ contains the eigenvectors of (AA^T)

$\Sigma^T\Sigma$ and $\Sigma\Sigma^T$ contain the non-zero eigenvalues of A^TA and AA^T

Singular values $\sigma_i = \sqrt{\text{eigenvectors of } (A^TA) \text{ and } (AA^T)}$

Left and right singular vectors come in pairs with the same σ_k^2

$$(A^TA)\,\vec{v}_k = \sigma_k^2\,\vec{v}_k$$
$$(AA^T)\,\vec{u}_k = \sigma_k^2\,\vec{u}_k$$

with the same σ_k^2

It follows from the equality that:

$$\vec{u}_k = \frac{A\vec{v}_k}{\sigma_k}$$

Check: $(AA^T)\vec{u}_k = AA^T\frac{A\vec{v}_k}{\sigma_k} = \sigma_k^2\frac{A\vec{v}_k}{\sigma_k} = \sigma_k^2\vec{u}_k$ ✓

Eckart-Young:

"Given $A_{m\times n}$, find $B_{m\times n}$ of rank at most k that is close to A"

i.e., $\|A - B\|$ small.

↑ Matrix norm.

Matrix Norms:

- **Induced norms:** e.g. l_2 norm ≡ Spectral norm

$$\|A\|_2 = \max_{\vec{x}} \frac{\|A\vec{x}\|}{\|\vec{x}\|} = \sigma_1$$

↙ largest singular value of A

These are vector norms

- **Element-wise norms:**

e.g. Frobenius norm:

$$\|A\|_F^2 = \sum_{ij} |a_{ij}|^2 = Tr(A^TA) =$$

$$= Tr(U\Sigma^2U^T) = Tr(\Sigma^2) = \sum_{j=1}^{r}\sigma_j^2$$

Equivalent to the vector norm of $\|vec(A)\|$

(1) E-Y for spectral norm:

Statement: If $\text{rank}(B) \le k$ then $\|A-B\| = \max_{\vec{x}} \frac{\|(A-B)\vec{x}\|}{\|\vec{x}\|}$

$$\ge \frac{\|(A-A_k)\vec{x}\|}{\|\vec{x}\|}$$

where $A_k = U_k \Sigma_k V_k$

Proof: (*) $\|(A-A_k)\vec{x}\| = \sigma_{k+1}$ since $A_k = U_k \Sigma_k V_k$ where these are from cutting

$\exists\ \vec{x} \neq 0_V$ / $B\vec{x} = 0_V$ and $\vec{x} = \sum_{j=1}^{k+1} c_j \vec{v}_j$

given that $\begin{cases} \dim[\ker B] \ge n-k \\ \Rightarrow \{\vec{v}_j\}_{j=1}^{k+1} \cap \ker B \neq \emptyset \end{cases}$

Then we have: $\|(A-B)\vec{x}\|^2 = \|A\vec{x}\|^2 = \sum_{j=1}^{k+1} c_j^2 \sigma_j^2$

$$\ge \left(\sum_{j=1}^{k+1} c_j^2\right) \sigma_{k+1}^2 = \|\vec{x}\|^2 \sigma_{k+1}^2$$

(**) $\|(A-B)\vec{x}\|^2 \ge \|\vec{x}\|^2 \sigma_{k+1}^2$ ✓

where the $\{\sigma_1, \sigma_2, \dots, \sigma_k, \sigma_{k+1}, \dots \sigma_r\}$ are ordered in decreasing order.

Hence: From (*) and (**) $\frac{\|(A-B)\vec{x}\|^2}{\|\vec{x}\|^2} \ge \sigma_{k+1}^2 = \|A-A_k\|^2$ ✓

(2) E-Y for Frobenius norm:

Constructive proof:

Find B that minimises

$$\min_B \|A - B\|_F^2$$

where B of rank k or less

(I) Any B of rank k or less can be written as:

$$B = \underset{m\times k}{C}\ \underset{k\times n}{R}$$

where

$$C^TC = D$$

$$RR^T = I$$

This is from SVD (for example)

(II) Consider E and minimize over C, R:

$$E = \|A - CR\|_F^2$$

Matrix calculus

$$\frac{\partial E}{\partial C} = 2(CR - A)R^T \quad (1)$$

$$\frac{\partial E}{\partial R} = 2(R^TC^T - A^T)C \quad (2)$$

At the minimum $\Rightarrow$

$$CRR^T = AR^T \quad (1)$$

$$R^TC^TC = A^TC \quad (2)$$

Combining (1) & (2) we get

$$(A^TA)R^T = R^TD$$

$$R^T = V_k$$

$$(AA^T)C = CD$$

$$C = U_k.$$

SVD gives best CR ✓

And the error is then:

$$E = \|A - CR\|_F^2 = \sum_{j=k+1}^{r} \sigma_j^2$$

Connection with PCA

$$\underset{N\times p}{Y^T} = \begin{pmatrix} x_1^{(1)} & \cdots & x_p^{(1)} \\ \vdots & & \\ x_1^{(N)} & \cdots & x_p^{(N)} \end{pmatrix} = \begin{pmatrix} \vec{y}_1^{\,T} \\ \vdots \\ \vec{y}_N^{\,T} \end{pmatrix}$$

$\vec{y}_i \in \mathbb{R}^p \qquad i = 1, \dots, N$

$\vec{1}_{N\times 1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ vector of ones.

Let: $\underset{1\times N}{\vec{1}^T}\,\underset{N\times p}{Y^T} = \left(\sum_{i=1}^{N} x_1^{(i)} \quad \cdots \quad \sum_{i=1}^{N} x_p^{(i)} \right)_{1\times p}$

$$\frac{1}{N}\,\vec{1}\left(\vec{1}^T \cdot Y^T\right) = \begin{pmatrix} \langle x_1 \rangle_i & \cdots & \langle x_p \rangle_i \\ \vdots & & \\ \langle x_1 \rangle_i & \cdots & \langle x_p \rangle_i \end{pmatrix}_{N\times p}$$

Then define:

$$\underset{N\times p}{\tilde{Y}^T} = Y^T - \frac{1}{N}\,1 1^T Y^T = \underbrace{\left(I - \frac{1}{N}\,1 1^T\right)}_{\text{Centering matrix}} Y^T$$

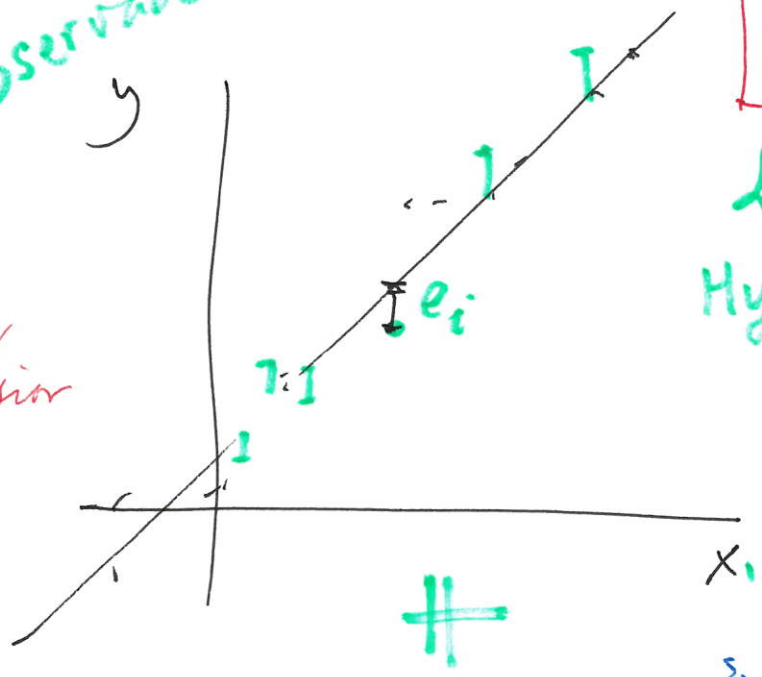
It then follows that $\left(\tilde{Y}\tilde{Y}^T\right)_{p\times p} = C_X$ sample covariance matrix

Using the SVD of $\tilde{Y}$ which is the centered data matrix.

So we conclude that PCA just considers the SVD of the centered data matrix $\tilde{Y}$, which is equivalent to obtaining the eigenvectors of the covariance matrix $C_X = (\tilde{Y}\tilde{Y}^T)$

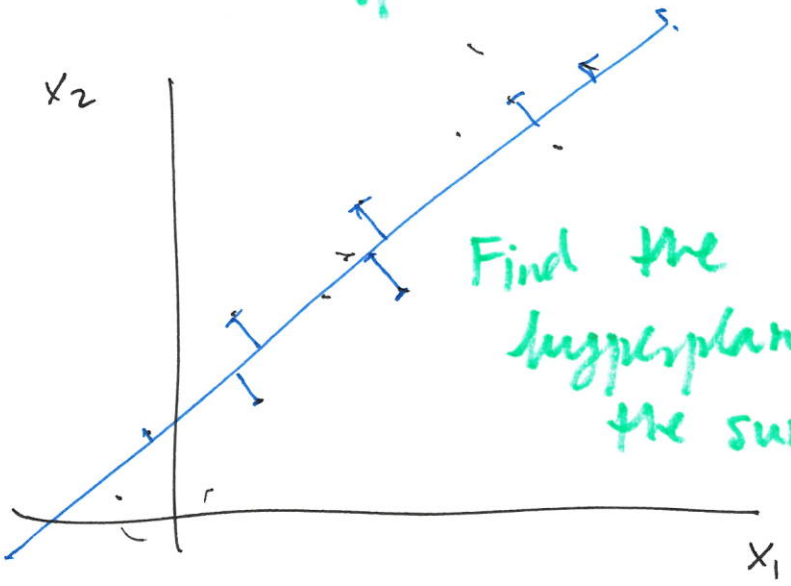
Geometric picture of PCA/SVD

observable:



(A) Linear regression

Linear regression:
Hyperplane that minimise the sum of squared error



(B) PCA

PCA

Find the hyperplane such that the sum of distances that are normal to the plane are minimised.

The hyperplanes we obtain are not the same for linear regression and PCA.

PCA through SVD is close to total least squares or proper orthogonal decompositions.