Covariance Matrix

- Recall that for a r.variable X with $\mu = \mathbb{E}X$, $\quad Var(X) = \mathbb{E}(X-\mu)^2 = \mathbb{E}X^2 - \mu^2$.
- Generalize the notion of variance for a random <u>vector</u> $X = (X_1, \dots, X_n)$. Squaring ?
- Two ways to multiply vectors in $\mathbb{R}^n$:
  1. Inner product: $x^T y = \langle x, y\rangle = \sum_1^n x_i y_i \in \mathbb{R}$
  2. Outer product: $x y^T = [x_i y_j]_{i,j=1}^n = x \otimes y \in \mathbb{R}^{n \times n}$.

1. $\boxed{Var(X) := \mathbb{E}(X-\mu)^T (X-\mu)} = \mathbb{E}\|X-\mu\|_2^2 = \sum_1^n Var(X_i)$.

   More informative is

2. **Def** The <u>covariance matrix</u> of a r.vector $X = (X_1, \dots, X_n)$ with $\mu = \mathbb{E}X$ is

$$Cov(X) := \mathbb{E}(X-\mu)(X-\mu)^T = \mathbb{E}XX^T - \mu\mu^T$$

- Entries:

$$Cov(X)_{ij} = \mathbb{E}(X-\mu)_i (X-\mu)_j = \mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j) = Cov(X_i, X_j)$$

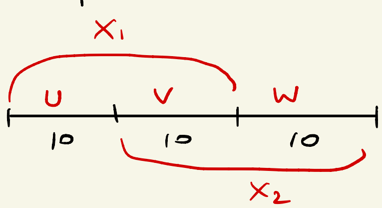
  Diagonal entries:

$$Cov(X)_{ii} = Cov(X_i, X_i) = Var(X_i)$$

- For $X \in \mathbb{R}^2$, $\quad Cov(X) = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_2, X_1) & Var(X_2) \end{bmatrix}$.

Ex: Flip a coin 30 times. $\quad X_1 = \#$heads in the first 20 flips, $\quad X = (X_1, X_2)$

$X_2 = \#$heads in the last 20 flips.



$X_1, X_2 \sim Binom(20, \frac{1}{2})$.

$Var(X_1) = Var(X_2) = 20 \cdot \frac{1}{2}(1 - \frac{1}{2}) = 5$.

$$Cov(X_1, X_2) = Cov(U+V, V+W)$$

$$= \underbrace{Cov(U,V)}_{=0} + \underbrace{Cov(V,V)}_{= Var(V)} + \underbrace{Cov(U,W)}_{=0} + \underbrace{Cov(V,W)}_{=0}$$

$$= 10 \cdot \tfrac{1}{2}(1-\tfrac{1}{2}) = 2.5 \quad V \sim Binom(10, \tfrac{1}{2})$$

$$\Rightarrow \quad Cov(X) = \begin{bmatrix} 5 & 2.5 \\ 2.5 & 5 \end{bmatrix}.$$

- $\boxed{Cov(X) \text{ is a } \underbrace{d \times d \text{ symmetric}}_{\text{Obvious}} \ \underbrace{\text{PSD matrix}}_{\text{Proof}}}$.

  $Cov(X) = \Sigma$

$\forall v \in \mathbb{R}^d$: $\quad v^T \Sigma v = v^T \mathbb{E}(X-\mu)(X-\mu)^T v = \mathbb{E}\, \underbrace{v^T(X-\mu)}_{=\langle X-\mu, v\rangle} \underbrace{(X-\mu)^T v}_{=\langle X-\mu, v\rangle} = \mathbb{E}\langle X-\mu, v\rangle^2 \qquad (*)$

Eigenvalues & Eigenvectors of $\mathrm{Cov}(X)$

- Eigenvalues of a symmetric PSD matrix $\Sigma = \mathrm{Cov}(X)$ are real, nonnegative

$$\lambda_1(\Sigma) \ge \lambda_2(\Sigma) \ge \cdots \ge \lambda_d(\Sigma) \ge 0$$

Prop $\lambda_1(\Sigma) = \max_{|v|=1} \mathrm{Var}(\langle X, v\rangle)$, $\quad v_1(\Sigma) = \operatorname{argmax}_{|v|=1} \mathrm{Var}(\langle X, v\rangle)$

Proof follows from the variational characterization:

$$\lambda_1(\Sigma) = \max_{|v|=1} \underbrace{v^T \Sigma v}_{\parallel \ (*)\, p.57}, \qquad v_1(\Sigma) = \operatorname{argmax}_{|v|=1} v^T \Sigma v$$

$$\mathbb{E}\langle X-\mu, v\rangle^2 = \mathrm{Var}(\langle X, v\rangle)$$

$\square$

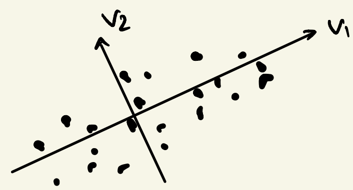
$\mathrm{Var}(X) = \mathbb{E}\|X-\mu\|_2^2 = \sum_1^n \mathrm{var}(\langle X, v_i\rangle)$

↑note

Interpretation: The $\underbrace{\text{direction } v = v_1(\Sigma)}$ explains the most variation of X.

equivalently, the linear combination $\langle X, v\rangle = \sum_{i=1}^d X_i v_i$

Remark: $v_2(\Sigma)$ explains the most of remaining variance (among directions $\perp v_1(\Sigma)$)



Ex (p.57): $\lambda_1(\Sigma) = 7.5 \quad v_1(\Sigma) = \begin{bmatrix}1\\1\end{bmatrix} \Rightarrow X_1 + X_2$ explains most (gives most overlap)

$\lambda_2(\Sigma) = 2.5 \quad v_2(\Sigma) = \begin{bmatrix}1\\-1\end{bmatrix} \Rightarrow X_1 - X_2$ explains least (gives least overlap)

Principal Component Analysis (PCA): $X \longmapsto \begin{bmatrix}\langle X, v_1\rangle \\ \langle X, v_2\rangle\end{bmatrix}$ projection $\mathbb{R}^d \to \mathbb{R}^2$

PCA In order to reduce dimension of X from d to k, project X onto the subspace spanned by the top k eigenvectors of $\mathrm{Cov}(X)$. "Principal components" of X.

This subspace explains the variability of X the best among all k-dimensional subspaces.

Advantages & disadvantages

General Multivariate Normal Distribution $N(\mu, \Sigma)$

- Recall standard normal $Z \sim N(0, I_n)$: pdf $\boxed{f_Z(z) = \frac{1}{(2\pi)^{n/2}} e^{-\|z\|_2^2/2}}$, $z \in \mathbb{R}^n$ (*)

Def (General normal)
$X = (X_1, \dots, X_n) \in \mathbb{R}^n$ is a normal r.vector if X is an affine transformation of a standard normal r.vector $Z \sim N(0, I_n)$, i.e.
$$X = AZ + \mu \qquad \text{for some invertible } A \in \mathbb{R}^{n \times n},\ \mu \in \mathbb{R}^n.$$

- $\mathbb{E}X = \mu$,
- $\mathrm{Cov}(X) = \mathbb{E}(AZ)(AZ)^T = \mathbb{E}\left[AZZ^TA^T\right] = A \cdot \underbrace{\mathbb{E}[ZZ^T]}_{= I_d} \cdot A^T = \boxed{AA^T =: \Sigma}$

- pdf? Assume $\mu = 0$ for simplicity

$$\forall B \in \mathcal{B}(\mathbb{R}^n): \quad P\{\underset{AZ}{\underset{\parallel}{X}} \in B\} = P\{Z \in A^{-1}(B)\} = \int_{A^{-1}(B)} f_Z(z)\, dz.$$

Change of var:
$x = Az$
$dx = |A| dz$
determinant

$$= \int_B \underline{f_Z(A^{-1}x)\, |A|^{-1}}\, dx$$

$$\Rightarrow f_X(x) = f_Z(A^{-1}x)\,|A|^{-1} \overset{(*)}{=} \frac{1}{(2\pi)^{n/2}|A|} \exp\left(-\|A^{-1}x\|_2^2/2\right)$$

- Let's express this in terms of $\Sigma = AA^T$.

$$|A| = \sqrt{|AA^T|} = \sqrt{|\Sigma|}.$$

det(AB) = det(A) det(B)

$$\|A^{-1}x\|_2^2 = (A^{-1}x)^T(A^{-1}x) = x^T(A^{-1})^T A^{-1}x = x^T(AA^T)^{-1}x = x^T\Sigma^{-1}x$$

$$\Rightarrow f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x^T\right)$$

- General case, for $\mu \neq 0$: apply the shift $\Rightarrow$

$$\boxed{f_X(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)}, \quad x \in \mathbb{R}^n \qquad (**)$$

- Compare to pdf of $X \sim N(\mu, \sigma^2)$ in 1D:
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-\mu)^2/2\sigma^2\right).$$