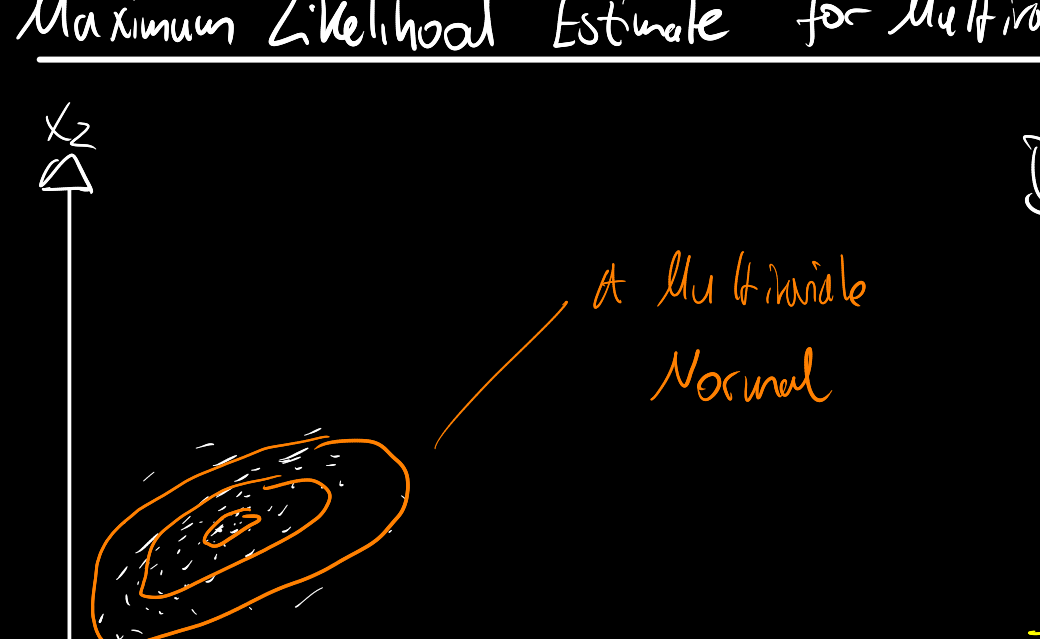
Maximum Likelihood Estimate for Multivariate Normal



$D = \{ (0.3, 1.7), (0.2, 1.6) \dots \}$ ($x^{[0]}$, $x^{[1]}$)

$\Downarrow$

$\underline{\mu}$ & $\underline{\underline{\Sigma}}$

$\Rightarrow$ MLE

$\underline{X}_i \sim \mathcal{N}(\underline{X}_i ; \underline{\mu}, \underline{\underline{\Sigma}})$ defined in $\mathbb{R}^K$ (not use D to avoid confusion with dataset)

$\underline{\mu}$: mean (vector) $\in \mathbb{R}^K$

$\underline{\underline{\Sigma}}$: covariance (matrix) $\in \mathbb{R}^{K \times K}$

$\underline{\underline{\Sigma}} > 0$ → sym. positive definite ($\underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}^T$)

e.g. $K=2$ $\underline{\underline{\Sigma}} = \begin{bmatrix} 1.7 & 0.3 \\ 0.3 & 1.5 \end{bmatrix}$ (covariances, variances)

$$\mathcal{N}(\underline{X}_i ; \underline{\mu}, \underline{\underline{\Sigma}}) = \frac{1}{\sqrt{\det(\underline{\underline{\Sigma}}) \cdot (2\pi)^K}} \exp\left(-\frac{1}{2}(\underline{X}_i - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{X}_i - \underline{\mu})\right)$$

Likelihood:

$$\mathcal{L}(D; \underline{\mu}, \underline{\underline{\Sigma}}) \overset{i.i.d.}{=} \prod_{i=0}^{N-1} p(\underline{X}_i = \underline{x}^{[i]})$$

$$= \prod_{i=0}^{N-1} \mathcal{N}(\underline{X}_i = \underline{x}^{[i]} ; \underline{\mu}, \underline{\underline{\Sigma}})$$

$$= \prod_{i=0}^{N-1} \frac{1}{\sqrt{\det(\underline{\underline{\Sigma}}) \cdot (2\pi)^K}} \exp\left(-\frac{1}{2}(\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}^{[i]} - \underline{\mu})\right)$$

$$= \det(\underline{\underline{\Sigma}})^{-\frac{N}{2}} \cdot (2\pi)^{-\frac{NK}{2}} \exp\left(-\frac{1}{2}\sum_{i=0}^{N-1} (\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}^{[i]} - \underline{\mu})\right)$$

don't confuse ($\sum$ and $\underline{\underline{\Sigma}}$)

Log-Likelihood

$$\ell(D; \underline{\mu}, \underline{\underline{\Sigma}}) = \log(\mathcal{L}(D; \underline{\mu}, \underline{\underline{\Sigma}}))$$

$$= -\frac{N}{2}\log\det(\underline{\underline{\Sigma}}) - \frac{NK}{2}\log(2\pi) - \frac{1}{2}\sum_{i=0}^{N-1} (\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} (\underline{x}^{[i]} - \underline{\mu})$$

Maximum Likelihood Estimate

$$\underline{\mu}^*, \underline{\underline{\Sigma}}^* = \underset{\substack{\underline{\mu} \in \mathbb{R}^K \\ \underline{\underline{\Sigma}} \in \mathbb{R}^{K \times K} \\ \underline{\underline{\Sigma}} > 0}}{\arg\max} \; \ell(D; \underline{\mu}, \underline{\underline{\Sigma}})$$

($\underline{\underline{\Sigma}} > 0$) — arises naturally *

Take derivative & set to zero

* More detail on this (was not 100% in the video): our MLE does not guarantee positive definiteness, although it guarantees symmetry. In order to be on the safe side one can use an called a "shrunk" covariance.

Essentially, this is: $\underline{\underline{\Sigma}}_{shrunk} = (1-\alpha)\underline{\underline{\Sigma}}_{MLE} + \alpha \frac{tr(\underline{\underline{\Sigma}}_{MLE})}{K} \cdot \underline{\underline{I}}$

which generates more diagonal dominance

($\underline{\mu}$)

$$\frac{\partial \ell}{\partial \underline{\mu}} = -\frac{1}{2}\sum_{i=0}^{N-1}\left(\underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu}) + (\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1}\right) \overset{!}{=} \underline{0}$$

$(\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1} = \underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu})$ because $\underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}^T$ ↳ $\underline{\underline{\Sigma}}^{-1} = \underline{\underline{\Sigma}}^{-T}$

Small correction to the video (at ~ 14:40)

→ the shown equality is not correct (left hand side is a row vector, right hand side is a column vector)

→ Problem is in the derivative. Correct would be according to formula (108) of the matrix cookbook:

$$\frac{\partial(\underline{x}^T \underline{\underline{B}} \underline{x})}{\partial \underline{x}} = \underline{\underline{B}}\underline{x} + \underline{\underline{B}}^T\underline{x}$$

in our case

$$\frac{\partial}{\partial \underline{\mu}}\left(-\frac{1}{2}\sum_{i=0}^{N-1}(\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu})\right)$$

$$= -\frac{1}{2}\sum_{i=0}^{N-1}\left(\underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu}) + \underline{\underline{\Sigma}}^{-T}(\underline{x}^{[i]} - \underline{\mu})\right)$$

$\underline{\underline{\Sigma}}^{-T} = \underline{\underline{\Sigma}}^{-1}$ due to symmetry

$$= -\frac{1}{2}\sum_{i=0}^{N-1} 2 \cdot \underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu}) \overset{!}{=} \underline{0} \quad | \cdot (-1)$$

$$= \underline{\underline{\Sigma}}^{-1}\sum_{i=0}^{N-1}(\underline{x}^{[i]} - \underline{\mu}) \overset{!}{=} \underline{0} \quad | \underline{\underline{\Sigma}}$$

$$\hookrightarrow \sum_{i=0}^{N-1}(\underline{x}^{[i]} - \underline{\mu}) \overset{!}{=} \underline{0}$$

$$\sum_{i=0}^{N-1}\underline{x}^{[i]} - \underbrace{\sum_{i=0}^{N-1}\underline{\mu}}_{N\underline{\mu}} \overset{!}{=} 0$$

$$\Leftrightarrow \boxed{\underline{\mu} = \frac{1}{N}\sum_{i=0}^{N-1}\underline{x}^{[i]}}$$

($\underline{\underline{\Sigma}}$)

$$\ell \overset{+}{=} -\frac{N}{2}\log\det(\underline{\underline{\Sigma}}) - \frac{1}{2}\sum_{i=0}^{N-1}\underbrace{(\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu})}_{\text{scalar}}$$

$\underline{\underline{\Sigma}}^{-1}$ does not commute

Grammar: commute inside trace

$tr(\underline{\underline{A}}\,\underline{\underline{B}}) = tr(\underline{\underline{B}}\,\underline{\underline{A}})$

in general: $\underline{\underline{A}}\,\underline{\underline{B}} \neq \underline{\underline{B}}\,\underline{\underline{A}}$

$$= -\frac{N}{2}\log\det(\underline{\underline{\Sigma}}) - \frac{1}{2}\sum_{i=0}^{N-1} tr\left((\underline{x}^{[i]} - \underline{\mu})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}^{[i]} - \underline{\mu})\right)$$

$$= -\frac{N}{2}\log\det(\underline{\underline{\Sigma}}) - \frac{1}{2}\sum_{i=0}^{N-1} tr\left(\underbrace{(\underline{x}^{[i]} - \underline{\mu})(\underline{x}^{[i]} - \underline{\mu})^T}_{\text{outer product}} \underline{\underline{\Sigma}}^{-1}\right)$$

$$\underline{\underline{S}} = (\underline{x}^{[i]} - \underline{\mu})(\underline{x}^{[i]} - \underline{\mu})^T = [\dots] \in \mathbb{R}^{K \times K}$$

↳ Matrix Cookbook

$$\frac{\partial \det(\underline{\underline{A}})}{\partial \underline{\underline{A}}} = \det(\underline{\underline{A}}) \cdot \underline{\underline{A}}^{-T} \qquad \frac{\partial\, tr(\underline{\underline{A}}\,\underline{\underline{B}}^{-1})}{\partial \underline{\underline{B}}} = -(\underline{\underline{B}}^{-1}\underline{\underline{A}}\,\underline{\underline{B}}^{-1})^T$$

$$\frac{\partial \ell}{\partial \underline{\underline{\Sigma}}} = -\frac{N}{2}\frac{1}{\det(\underline{\underline{\Sigma}})} \cdot \det(\underline{\underline{\Sigma}})\,\underline{\underline{\Sigma}}^{-T} + \frac{1}{2}\sum_{i=0}^{N-1}(\underline{\underline{\Sigma}}^{-1}\underline{\underline{S}}\,\underline{\underline{\Sigma}}^{-1})^T \overset{!}{=} \underline{\underline{0}}$$

($\underline{\underline{\Sigma}}^{-T} = \underline{\underline{\Sigma}}^{-1}$)

$$= -\frac{N}{2}\underline{\underline{\Sigma}}^{-1} + \frac{1}{2}\sum_{i=0}^{N-1}\underline{\underline{\Sigma}}^{-T}\underline{\underline{S}}^T\underline{\underline{\Sigma}}^{-T} \overset{!}{=} \underline{\underline{0}} \quad | \cdot 2$$

$$= -N\underline{\underline{\Sigma}}^{-1} + \sum_{i=0}^{N-1}\underline{\underline{\Sigma}}^{-1}\left((\underline{x}^{[i]} - \underline{\mu})(\underline{x}^{[i]} - \underline{\mu})^T\right)^T\underline{\underline{\Sigma}}^{-1} \overset{!}{=} \underline{\underline{0}}$$

$$= -N\underline{\underline{\Sigma}}^{-1} + \underline{\underline{\Sigma}}^{-1}\sum_{i=0}^{N-1}\left((\underline{x}^{[i]} - \underline{\mu})(\underline{x}^{[i]} - \underline{\mu})^T\right)\underline{\underline{\Sigma}}^{-1} \overset{!}{=} \underline{\underline{0}}$$

$\underline{\underline{\Sigma}} \,|\, \underline{\underline{\Sigma}}$

$$= -N\underline{\underline{\Sigma}} + \sum_{i=0}^{N-1}(\underline{x}^{[i]} - \underline{\mu})(\underline{x}^{[i]} - \underline{\mu})^T \overset{!}{=} \underline{\underline{0}}$$

$$\boxed{\underline{\underline{\Sigma}} = \frac{1}{N}\sum_{i=0}^{N-1}(\underline{x}^{[i]} - \underline{\mu})(\underline{x}^{[i]} - \underline{\mu})^T}$$ (MLE)

$[\dots]^{[i]} \in \mathbb{R}^{K \times K}$

data matrix

$$\underline{\underline{X}} = \begin{bmatrix} — \underline{x}^{[0]} — \\ — \underline{x}^{[1]} — \\ \vdots \\ — \underline{x}^{[N-1]} — \end{bmatrix} \in \mathbb{R}^{N \times K}$$

different way to express the dataset

K dim (≙ features), N samples

Mean along side rows $\underline{\mu} = [\dots\dots] \in \mathbb{R}^K$

Centered data matrix

$$\tilde{\underline{\underline{X}}} = \underline{\underline{X}} - \begin{bmatrix} — \underline{\mu} — \\ — \underline{\mu} — \\ \vdots \\ — \underline{\mu} — \end{bmatrix} \in \mathbb{R}^{N \times K}$$

$$\tilde{\underline{\underline{X}}} = \underline{\underline{X}} - \underline{1}_N \underline{\mu}^T$$

$\underline{1}_N = [1, 1, \dots, 1]^T$

$$\underline{\underline{\Sigma}}_{MLE} = \frac{1}{N}\tilde{\underline{\underline{X}}}^T\tilde{\underline{\underline{X}}} \in \mathbb{R}^{K \times K}$$

$= \frac{1}{N} \cdot$ (K × N)(N × K)

$= \frac{1}{N} \cdot$ (K × K)