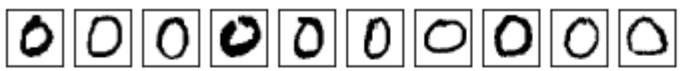
Decision Boundaries, Gaussian Estimation of Parameters

- Gaussian discriminant analysis
- ROC curves
- Gaussian estimation of parameters
- Sample mean
- Sample covariance

Gaussian Discriminant Analysis

$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)$ ML (maximum likelihood) estimate

- generative classifier: assume that x was generated by a class conditional distribution:



$p(\boldsymbol{x}|y=0) \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$

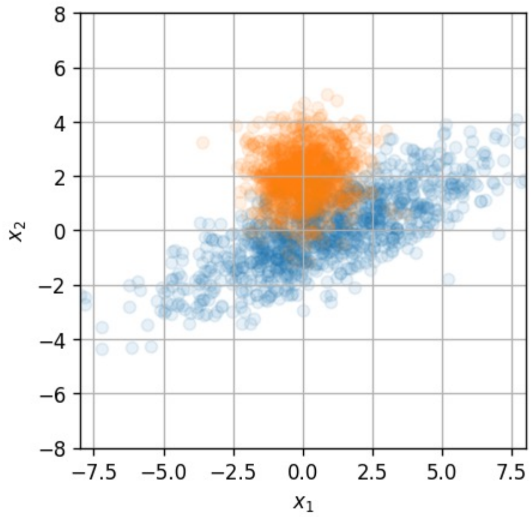
$p(\boldsymbol{x}|y=1) \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$

- given a new $\boldsymbol{x}$, we can evaluate $p(\boldsymbol{x}|y=0)$ vs. $p(\boldsymbol{x}|y=1)$

Gaussian Discriminant Analysis

$$\boldsymbol{x}|y=0 \sim \mathcal{N}\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}8 & 3\\3 & 2\end{bmatrix}\right)$$

$$\boldsymbol{x}|y=1 \sim \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}1 & 0.1\\0.1 & 1\end{bmatrix}\right)$$



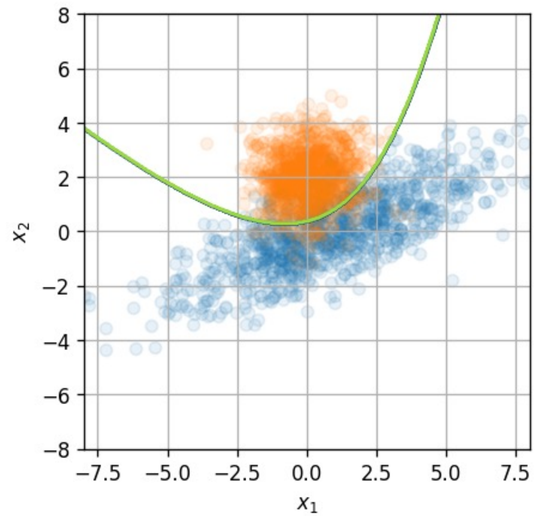
$$\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right) \quad \text{vs.} \quad \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)$$

- decision boundary has a quadratic form:

$$\boldsymbol{x}^T B\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{w} \lesseqgtr c$$

Gaussian Discriminant Analysis

$$\frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)}{\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)} \lesseqgtr 1$$



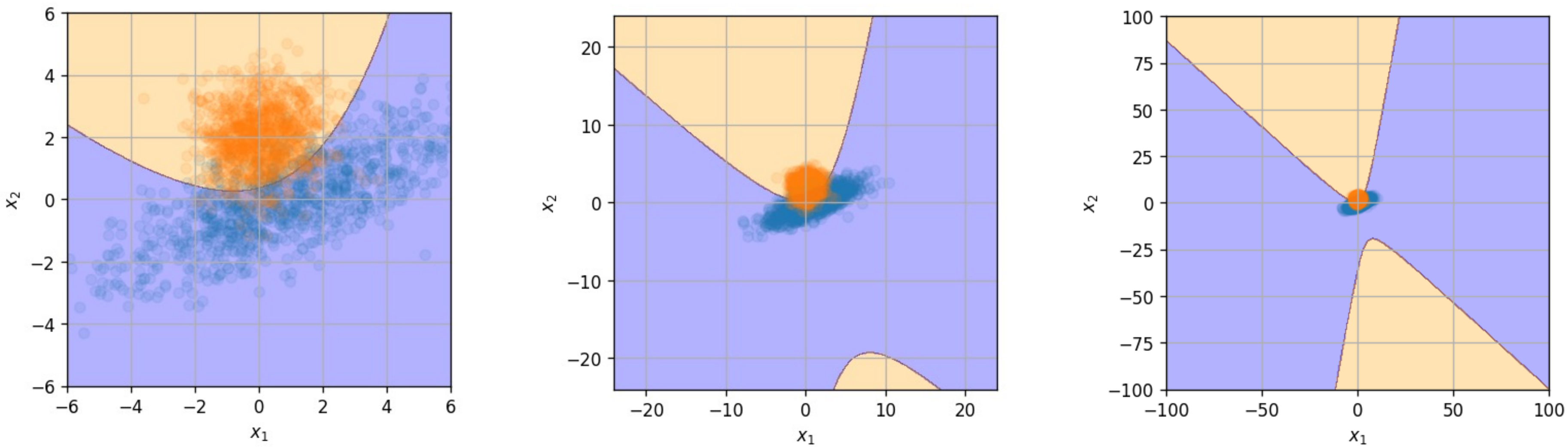
$$\log\left(|\boldsymbol{\Sigma}_1|\right)-\log\left(|\boldsymbol{\Sigma}_0|\right)-(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)+(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) \gtrless 0$$

$$c=\log\left(|\boldsymbol{\Sigma}_0|\right)-\log\left(|\boldsymbol{\Sigma}_1|\right)+\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0-\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1$$

$$\boldsymbol{w}=2(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0-\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1)$$

$$\boldsymbol{B}=\boldsymbol{\Sigma}_1^{-1}-\boldsymbol{\Sigma}_0^{-1}$$

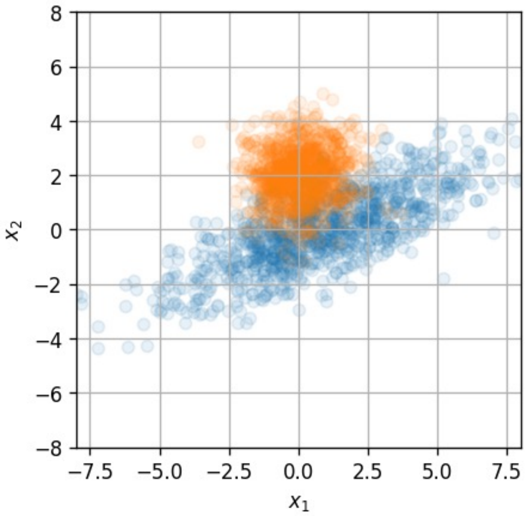
Quadratic Decision Boundary



Adding a non-uniform prior

$$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)$$ ML (maximum likelihood) estimate

$$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)p(y)$$ MAP (maximum a posteriori) estimate



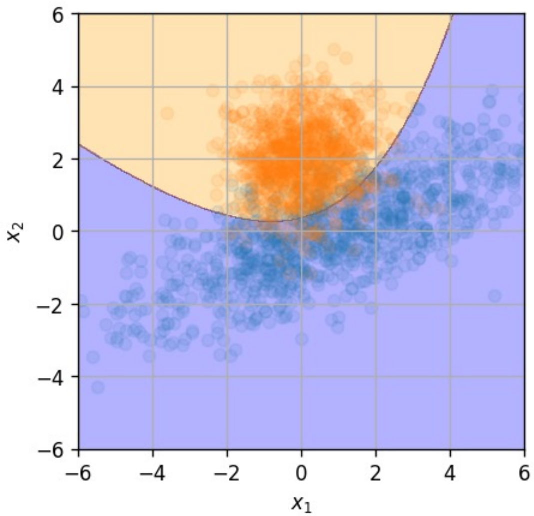
$$\frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)\right)p(y=0)}{\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right)p(y=1)} \lesseqgtr 1$$

$$\boldsymbol{x}^T B\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{w} \lesseqgtr c$$

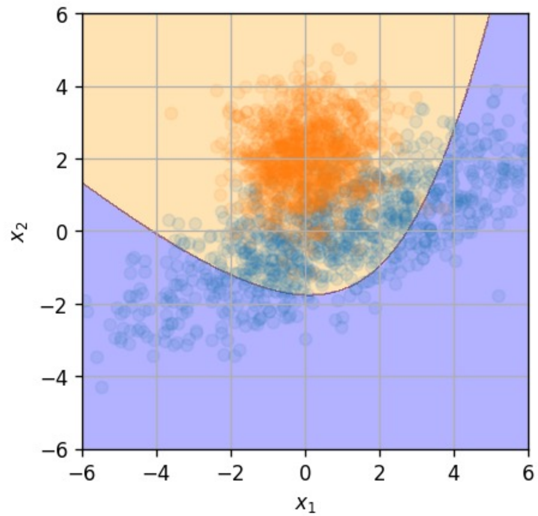
$$c = \ldots -2\log(p(y=0)) + 2\log(p(y=1))$$

Gaussian Discriminant Analysis

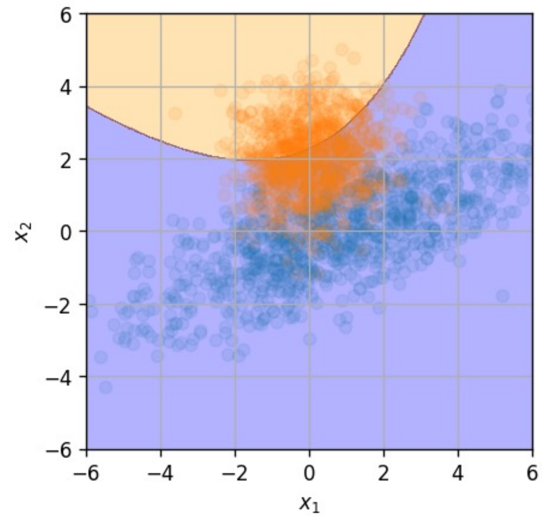
$p(y=0) = p(y=1)$



$p(y=0) < p(y=1)$

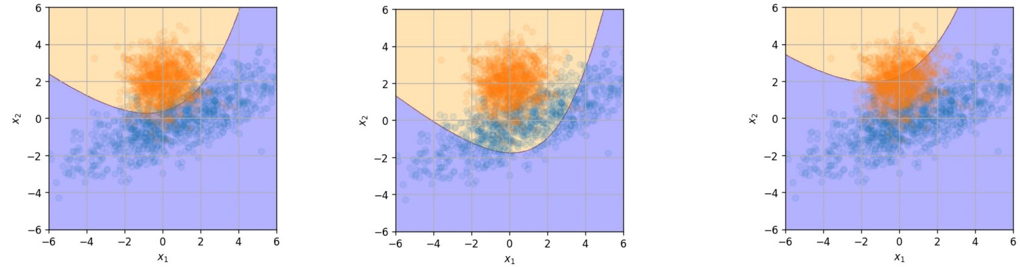


$p(y=0) > p(y=1)$



ROC (receiver operating characteristic) curve

$p(y=0) = p(y=1)$ $p(y=0) < p(y=1)$ $p(y=0) > p(y=1)$



miss prob

$\mathbb{P}(\hat{y} = 0 | y = 1)$

false alarm prob

$\mathbb{P}(\hat{y} = 1 | y = 0)$

Detection prob

$\mathbb{P}(\hat{y} = 1 | y = 1) = 1 -$ miss prob

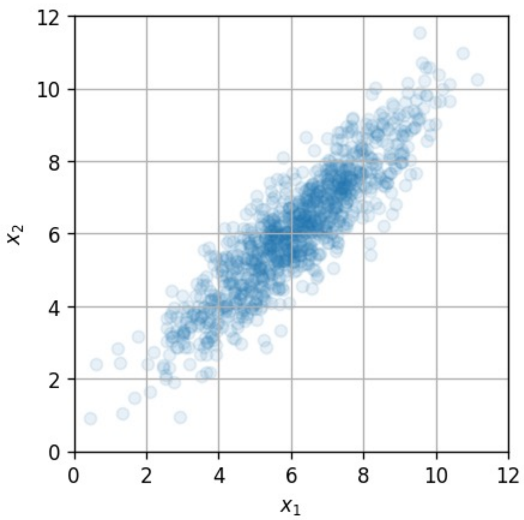
false alarm prob

$\mathbb{P}(\hat{y} = 1 | y = 0)$

Gaussian Estimation of Parameters

$\widehat{y} = \arg\max_y p(\boldsymbol{x}|y)$ ML (maximum likelihood) estimate

$\boldsymbol{x} \in \mathbb{R}^d$



- assumption is class conditional is normal: $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- given some data, how do we estimate $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$?
  - answer . . .

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum \boldsymbol{x}_i \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T$$

$$\boldsymbol{\mu} = E[\boldsymbol{x}] \qquad \boldsymbol{\Sigma} = E\left[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\right]$$

- assumption: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ are i.i.d.

Gaussian Estimation of Parameters

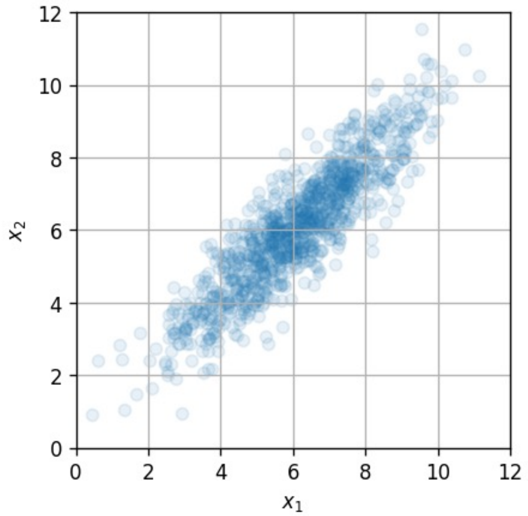
- given some data, how do we estimate $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$?

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- answer . . .

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum \boldsymbol{x}_i \qquad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}})^T$$

$\boldsymbol{x} \in \mathbb{R}^d$




```
print(np.shape(X))
```

```
(2, 1000)
```

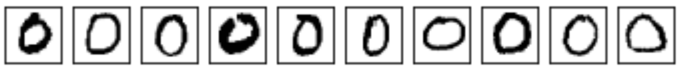
```
mu = np.mean(X, axis=1)     #take the mean
sigma = np.cov(X)           #wants each data point as a column
```

```
mu = [5.97 5.99]

sigma =
[[3.01  2.73]
 [2.73  3.01]]
```

Regularization

- in practice can have high dimensions (and limited training data)

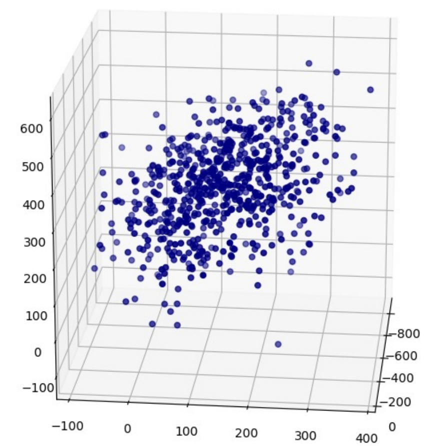


$d = 784$

$n = 5923$

n samples in of a d-dimensional vector

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T$$



- potential problem: Gaussian pdf involves computing $\widehat{\boldsymbol{\Sigma}}^{-1}$

$$\text{rank}(\widehat{\boldsymbol{\Sigma}}) \leq n$$

- before inverting, add to diagonal: $\widehat{\boldsymbol{\Sigma}} + \lambda \mathrm{I}$

$$(\widehat{\boldsymbol{\Sigma}} + \lambda \mathrm{I}) \text{ is invertible for } \lambda > 0$$

- a number of techniques for dealing with this – *graphical lasso*

Empirical Mean is the Maximum Likelihood Estimate

- if $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are i.i.d., then the maximum likelihood estimate for the mean is:

$$\widehat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{n} \sum \boldsymbol{x}_i$$

- proof:

$$L(\boldsymbol{\mu}) = \prod_i p_\mu(\boldsymbol{x}_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right)$$

$$\log L(\boldsymbol{\mu}) = -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{dn}{2} \log(2\pi) - \sum_i \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

approach: set $\nabla \log L(\boldsymbol{\mu}) = 0$ and solve for $\boldsymbol{\mu}$.

$$\nabla_\mu \log L(\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}) = 0$$

$$\Longrightarrow \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{n} \sum_i \boldsymbol{x}_i$$

$$(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

$$= \boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i - 2\boldsymbol{x}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^T$$

$$\nabla \boldsymbol{a}^T \boldsymbol{\mu} = \boldsymbol{a}$$

$$\nabla \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu} = (\boldsymbol{A} + \boldsymbol{A}^T) \boldsymbol{\mu}$$

$$\nabla_\mu = -2\boldsymbol{\Sigma}^{-1} \boldsymbol{x}_i + 2\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

Empirical Covariance is the Maximum Likelihood Estimate

- if $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are i.i.d., then the maximum likelihood estimate for the mean is:

$$\widehat{\boldsymbol{\Sigma}}_{\text{ML}} = \frac{1}{n} \sum (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^T$$

- proof:

$$L(\Sigma) = \prod_i p_\Sigma(\boldsymbol{x}_i) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right)$$

$$\log L(\boldsymbol{\Sigma}) = -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{dn}{2} \log(2\pi) - \sum_i \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

approach: define $\Phi = \Sigma^{-1}$, set $\nabla \log L(\Phi) = 0$ and solve for $\boldsymbol{\Sigma}$.

$$\nabla_\Phi \log L(\Phi) = \frac{n}{2} \Phi^{-1} - \frac{1}{2} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

$$\Longrightarrow \boldsymbol{\Sigma}_{\text{ML}} = \frac{1}{n} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

$$\nabla \log |\Phi^{-1}| = -(\Phi^{-1})^T$$

$$\begin{aligned} &(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \\ &\quad = \operatorname{tr}\left((\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right) \\ &\quad = \operatorname{tr}\left((\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \right) \end{aligned}$$

$$\nabla \operatorname{tr}(\boldsymbol{A}\Phi) = \boldsymbol{A}^T$$