Example Distributions

Distribution	PDF or PMF	Mean	Variance
Bernoulli(p)	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$	np	$np(1-p)$
Geometric(p)	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, \ldots$	λ	λ
Uniform(a, b)	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Gaussian(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Two R.V.

Bivariate CDF

$F_{XY}(x,y) = P(X \leq x, Y \leq y)$

Bivariate PMF

$P_{XY}(x,y) = P(X=x, Y=y)$

Marginal

Given → $p(x,y)$ → captures max info

$$\iint p(x,y)\,dx\,dy = 1$$

$$p(x) = \sum_y p(x,y) = \int p(x,y)\,dy$$

$$p(y) = \sum_x p(x,y) = \int p(x,y)\,dx$$

distribution wrt x

$$p(x,y) = p(x)\,p(y|x)$$

$$= p(y)\,p(x|y)$$

Marginal Conditional

$$p(x)p(y|x) = p(y)p(x|y)$$

$$p(y|x) = \frac{p(y)\,p(x|y)}{p(x)}$$

$$= \frac{p(y)\,p(x|y)}{\sum_{y'} p(y')\,p(x|y')}$$

Two r.v. are independent if:

$$P_{XY}(x,y) = P_X(x)\,P_Y(y)$$

$$P_{Y|X}(x,y) = P_Y(y)$$

$$\mathrm{Cov}[X,Y] = E\big[(X - E(X))(Y - E(Y))\big]$$

$$= E\big[XY - X\,E(Y) - Y\,E(X) + E(X)E(Y)\big]$$

$$= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y)$$

$$= E(XY) - E(X)E(Y)$$

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2$$

$x \in \mathbb{R}^n$

$$p(x;\mu,\Sigma) = \frac{1}{\quad} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

normalizing constant $\rightarrow (2\pi)^{n/2}|\Sigma|^{1/2}$

$x \rightarrow$ space over which it is defined

$\mu, \Sigma \rightarrow$ parameters

$\Sigma \rightarrow$ cov matrix

Σ is always P.S.D.

Proof: $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ $P(X)$

$$\Sigma = \int_{\mathbb{R}^n} P(x)(x-\mu)(x-\mu)^T dx$$

$$\mu = E(X) = \int_{\mathbb{R}^n} P(x)\, x\, dx$$

then $v^T \Sigma v \geq 0$ as Σ

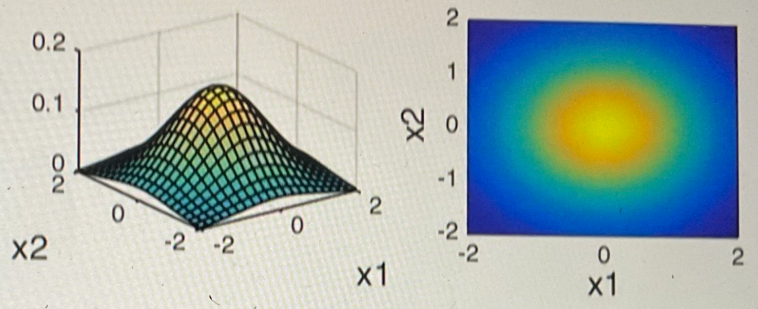
If Gaussian n dimension

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Multivariant Gaussian (Normal examples)

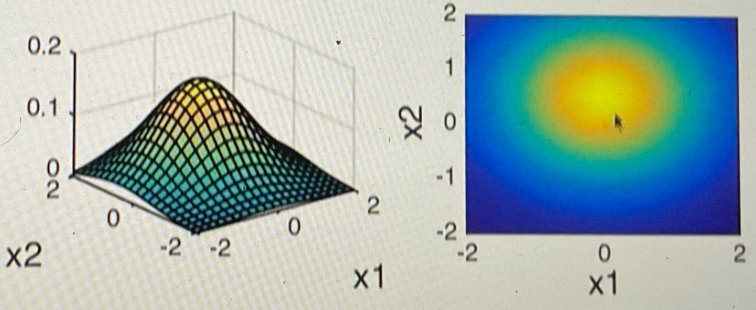
$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

$\mu = [0 \quad 0]^T$



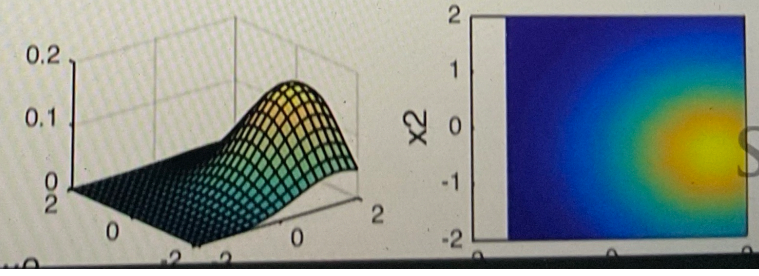
$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

$\mu = [0 \qquad 0.5]^T$



$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$

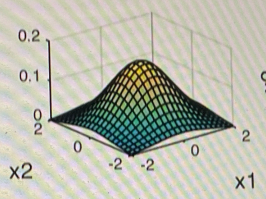
$\mu = [1.5 \qquad -0.5]^T$

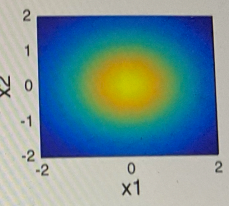


Stanf

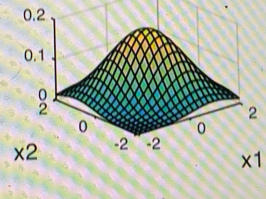
Multivariant Gaussian (Normal examples)

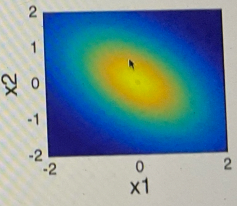
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



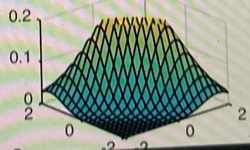


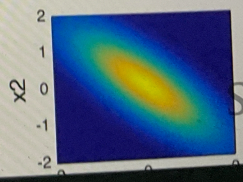
$$\Sigma = \begin{matrix} 1 & -0.5 \\ -0.5 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$





$$\Sigma = \begin{matrix} 1 & -0.8 \\ -0.8 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$

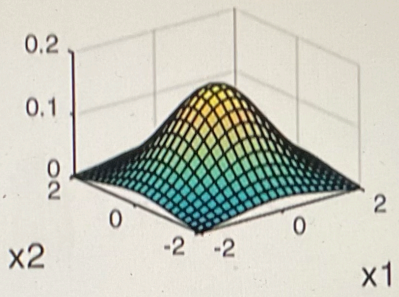


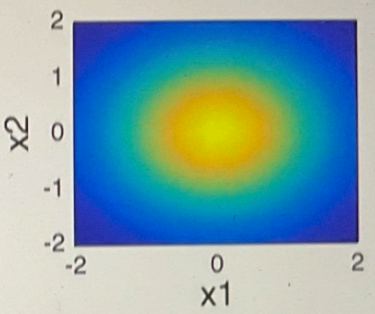


Stanf

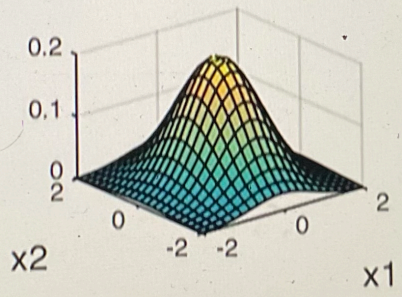
Multivariant Gaussian (Normal examples)

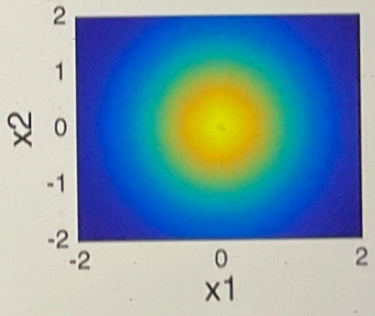
$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



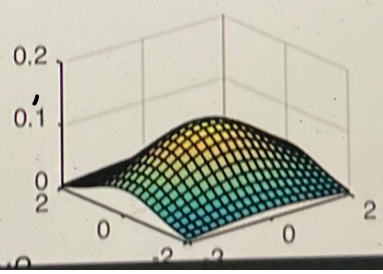


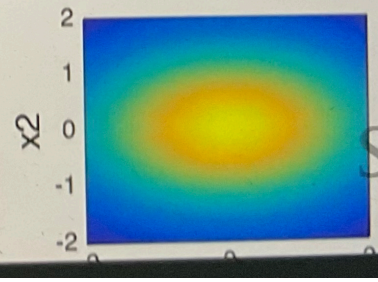
$$\Sigma = \begin{matrix} 0.6 & 0 \\ 0 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$



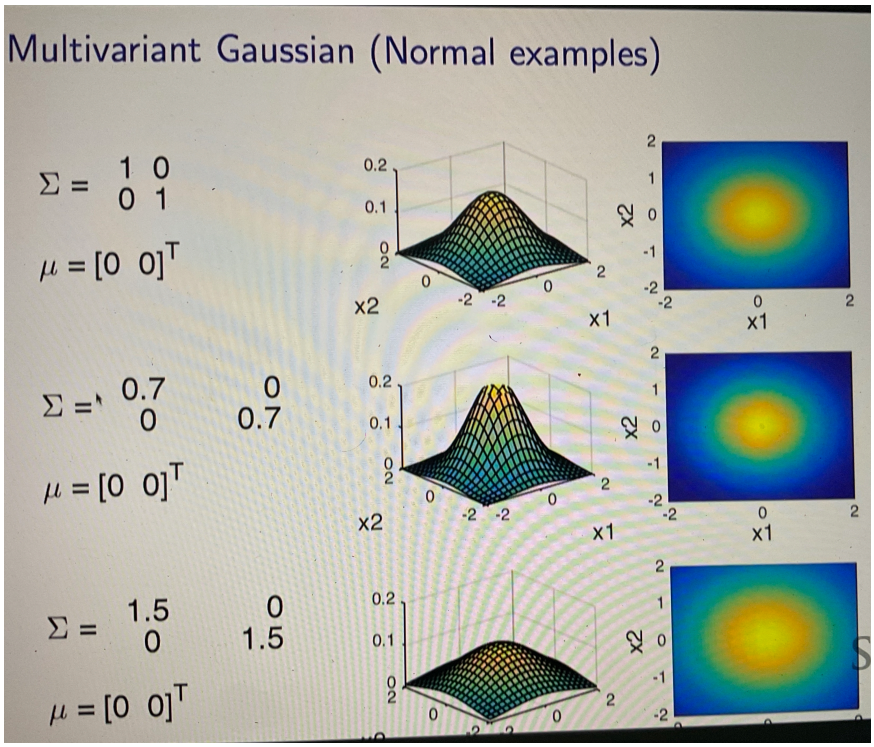


$$\Sigma = \begin{matrix} 2 & 0 \\ 0 & 1 \end{matrix} \qquad \mu = [0 \;\; 0]^T$$





Stanf



$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

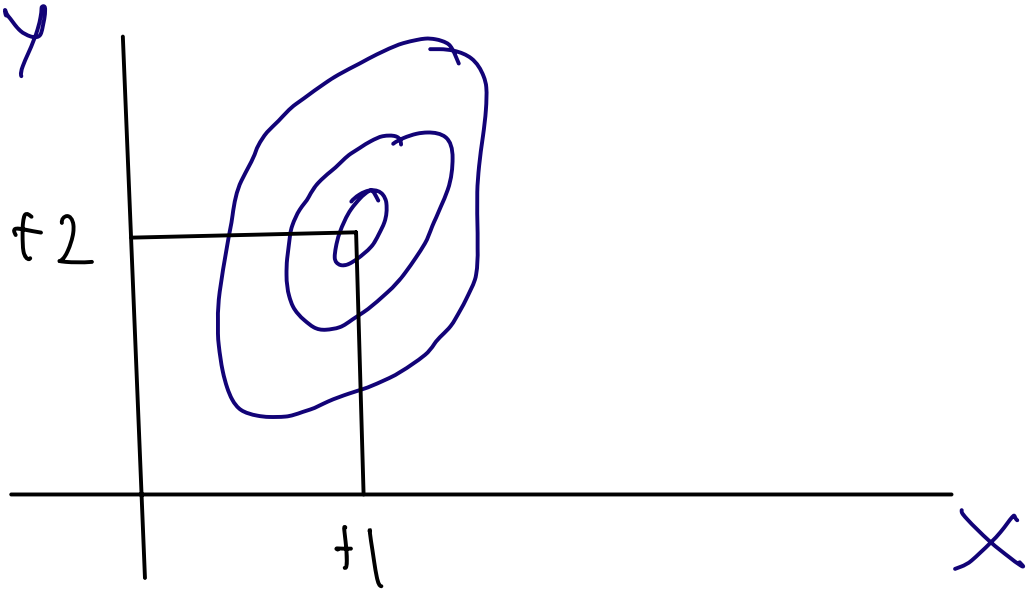
$$E(X \mid Y=y) = \sum_{x} \frac{x P(X=x, Y=y)}{P(Y=y)}$$

$E[X|Y)(Y) = E[X|Y=y]$ is a func^c of y so it is R.V.

$E[X]$ - constant

$E[X|Y]$ - R.V. over Y space (sample space of Y)

$E[X|Y=y]$ → function of y



$E[X|Y=y]$

LOTUS

$E[X] = E[E[X|Y]]$

$$E[E[X|Y]] = E\left[\sum_x x P(X=x|Y)\right]$$

$$= \sum_y \left(\sum_x x P(X=x|Y=y)\right) P(Y=y)$$

$$= \sum_y \sum_x x P(X=x, Y=y)$$

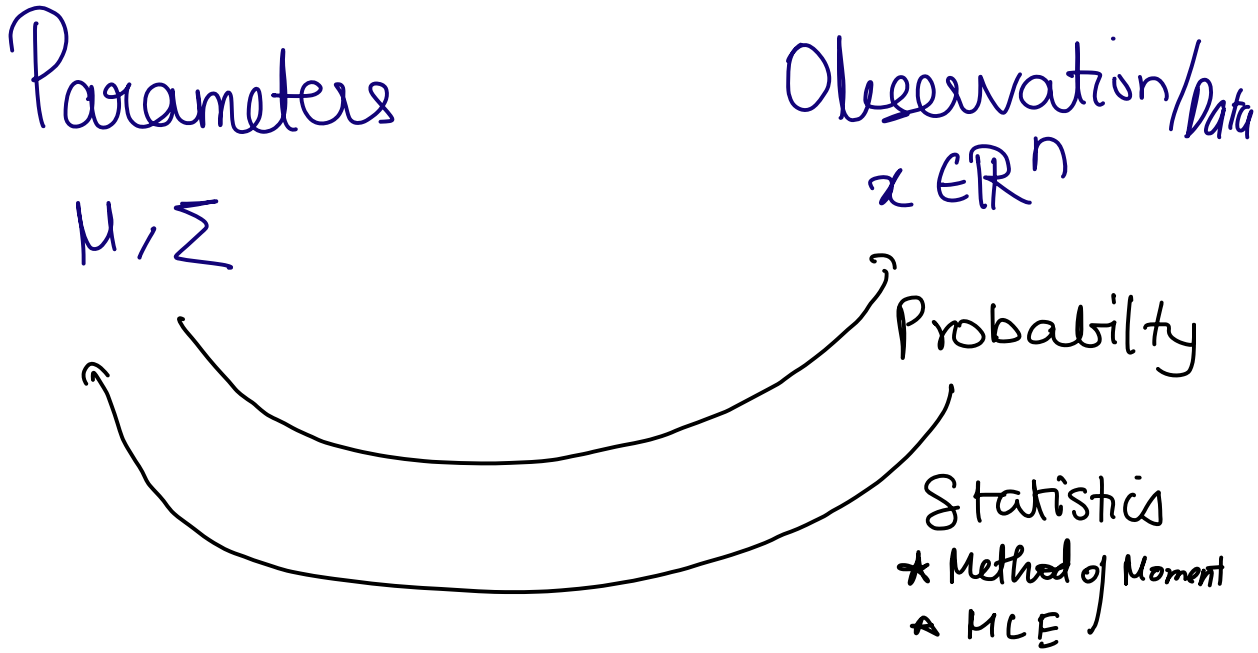
$$= \sum_x x \sum_y 1(X=x, Y=y)$$

$$= \sum_x x P(X=x)$$

$$= E[X]$$

$$P(a|b,c) = \frac{P(b|a,c)P(a|c)}{P(b|c)}$$

$$P(a|b) = \frac{P(b|a)\,P(a)}{P(b)}$$



Training Data

(x, y) → learn parameters (Stat)

Model → predict future data (Prob)

GAUSSIAN

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

$X = \{x^{(1)} \dots x^{(n)}\}$ $x^{(i)}$ → i^{th} example $\in \mathbb{R}^d$

i.i.d.

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

$$p(x^{(1)} \dots x^{(n)}) = \prod_{i=1}^{n} p(x^{(i)})$$

$$p(x^{(1)} \dots x^{(n)}; \mu, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu)^T \dots\right)$$

↑ given

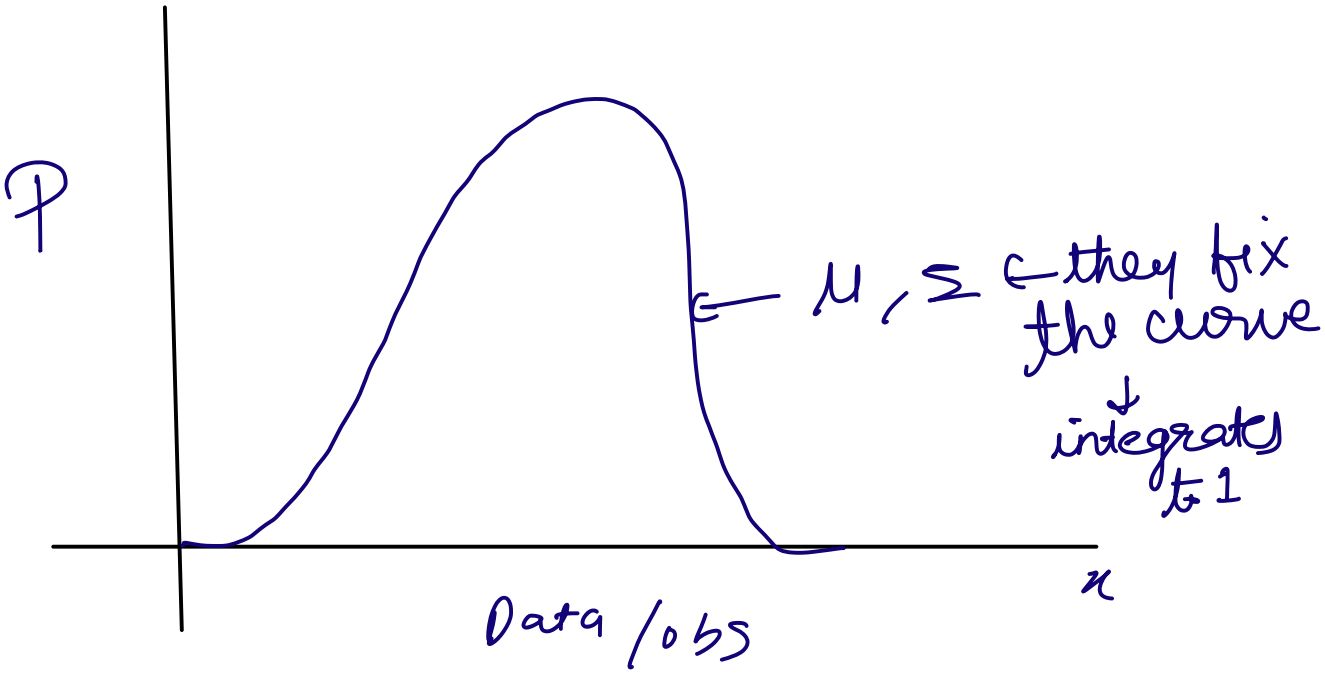
likelihood

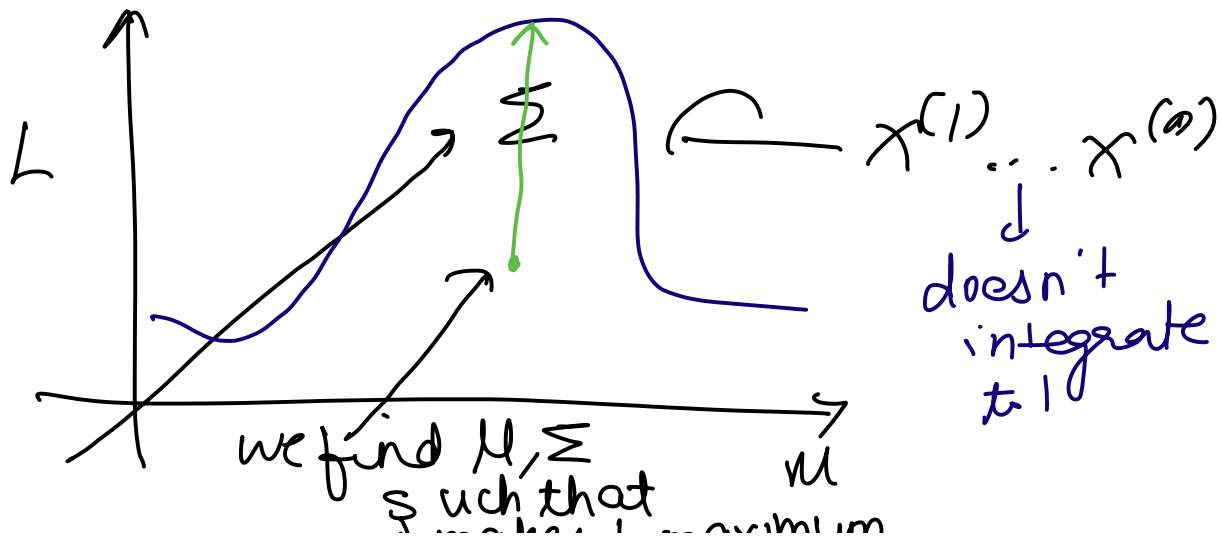
$$L(\mu, \Sigma; x^{(1)} \dots x^{(n)}) =$$

(Same expression, different interpretation)

probability of data given parameters

likelihood of parameters given data





θ makes L maximum:

$$L(\theta; X) = \prod_{i=1}^{n} L(\theta; x^{(i)})$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} L(\theta; x^{(i)})$$

$$= \arg\max_{\theta} \log \prod_{i=1}^{n} L(\theta; x^{(i)})$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log L(\theta; x^{(i)})$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \ell(\theta; x^{(i)})$$

GAUSSIAN

$$= \arg\max_{\mu, \Sigma} \sum_{i=1}^{n} \log \left[\frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) \right) \right]$$

$$\hat{\mu}, \hat{\Sigma} = \sum_{i=1}^{n} k - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu)$$

For μ

$$\nabla_\mu \left(\sum_{i=1}^{n} k - 0.5 \log|\Sigma| - 0.5 (x^{(i)} - \mu)^T \Sigma^{-1} (x^{(i)} - \mu) \right)$$

$$= \sum_{i=1}^{n} \left(0 - 0 - 0.5 \times 2 \times (x^{(i)} - \mu) \Sigma^{-1} \right)$$

$$\sum_{i=1}^{n} \left(\Sigma^{-1} x^{(i)} - \Sigma^{-1} \mu \right) = 0$$

$$n \Sigma^{-1} \mu = \sum_{i=1}^{n} \Sigma^{-1} x^{(i)}$$

$$\Sigma^{-1} \mu = \Sigma^{-1} \sum_{i=1}^{n} \frac{1}{n} x^{(i)}$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}$$

For Σ

$S = \Sigma^{-1}$ $\quad \nabla_S \ell = 0 \iff \nabla_{\Sigma^{-1}} \ell = 0$

$$\nabla_S \left(\sum_{i=1}^{n} \frac{1}{2} \log |S| - \frac{1}{2} (x^{(i)} - \mu)^T S (x^{(i)} - \mu) \right)$$

$$= \frac{1}{2} \left[n S^{-1} - \sum_{i=1}^{n} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \right]$$

$$\Sigma = S^{-1} = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

Use $\nabla_A x^T A x = x x^T$

Proof:

$\langle A, B \rangle = \operatorname{tr}(A^T B) = A : B$

$\langle A, BC \rangle = A : BC = AC^T : B$

$= B^T A : C$

$= C^T B^T : A^T$

$= BC : A$

$d(x^T A x)$

$d(X : AX)$

$d(XX^T : A)$

$XX^T : dA$

$\Sigma \approx E(X - E(X))$

$d(X^T AX)$

$d(X : AX)$

$dX : AX + X : AdX$

$dX : AX + A^T X : dX$

$dX : AX + A^T X$

Note: $L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y} | X; \theta)$