EECS 445

Introduction to Machine Learning

Gaussian Mixture Models

Prof. Kutty

Generative Models

- Why?
  - describes internal structure of the data
  - can also be used for classification, soft clustering, graphical models
- generative story with i.i.d. assumption

$$x^{(i)} \sim \mathrm{Distr}(x; \bar{\theta})$$

(identically distributed)

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

(independently distributed)

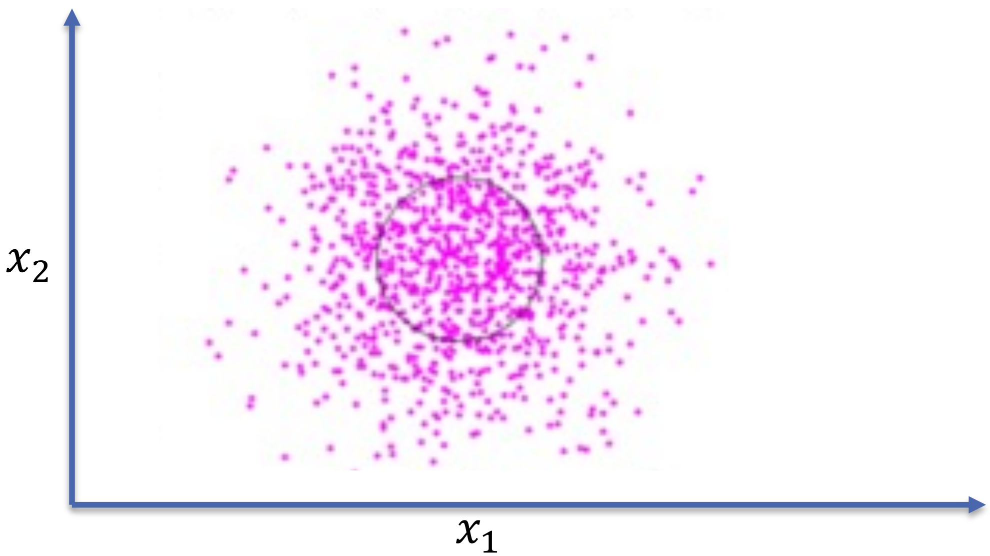
Determine distribution parameters $\bar{\theta}$

Multivariate Gaussian Distribution

Underlying Distribution for this (unlabeled) Dataset

for $\bar{x} \in \mathbb{R}^d$ $d \geq 2$

Example 1: Here $\bar{x} \in \mathbb{R}^2$



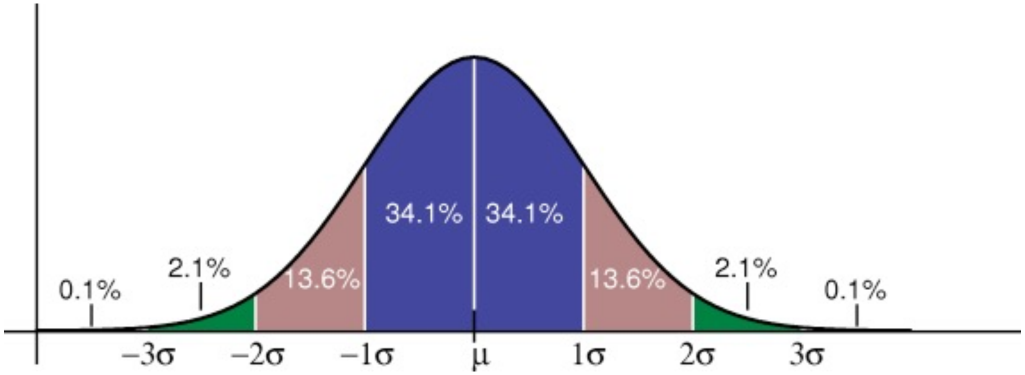
Example 2: Here $\bar{x} \in \mathbb{R}^4$

$x_1^{(i)}$	$x_2^{(i)}$	$x_3^{(i)}$	$x_4^{(i)}$
0.0002	10.052	8.602	227
1110	12.110	-805.1	-84.5
0.01	0.01	5292.01	837.1
710	-73610	8015.03	-2.503
-1120.09	11.01	1680	-5686
774.11	3.67	46.86	51.13
3.532	624	587.4	-3700

Gaussian (normal) Distribution

univariate Gaussian

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$



Multivariate Gaussian

d by 1 mean vector

d by d covariance matrix

$$\mathcal{N}(\bar{x}|\bar{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\bar{x}-\bar{\mu})^T\Sigma^{-1}(\bar{x}-\bar{\mu})]$$

d by 1 data

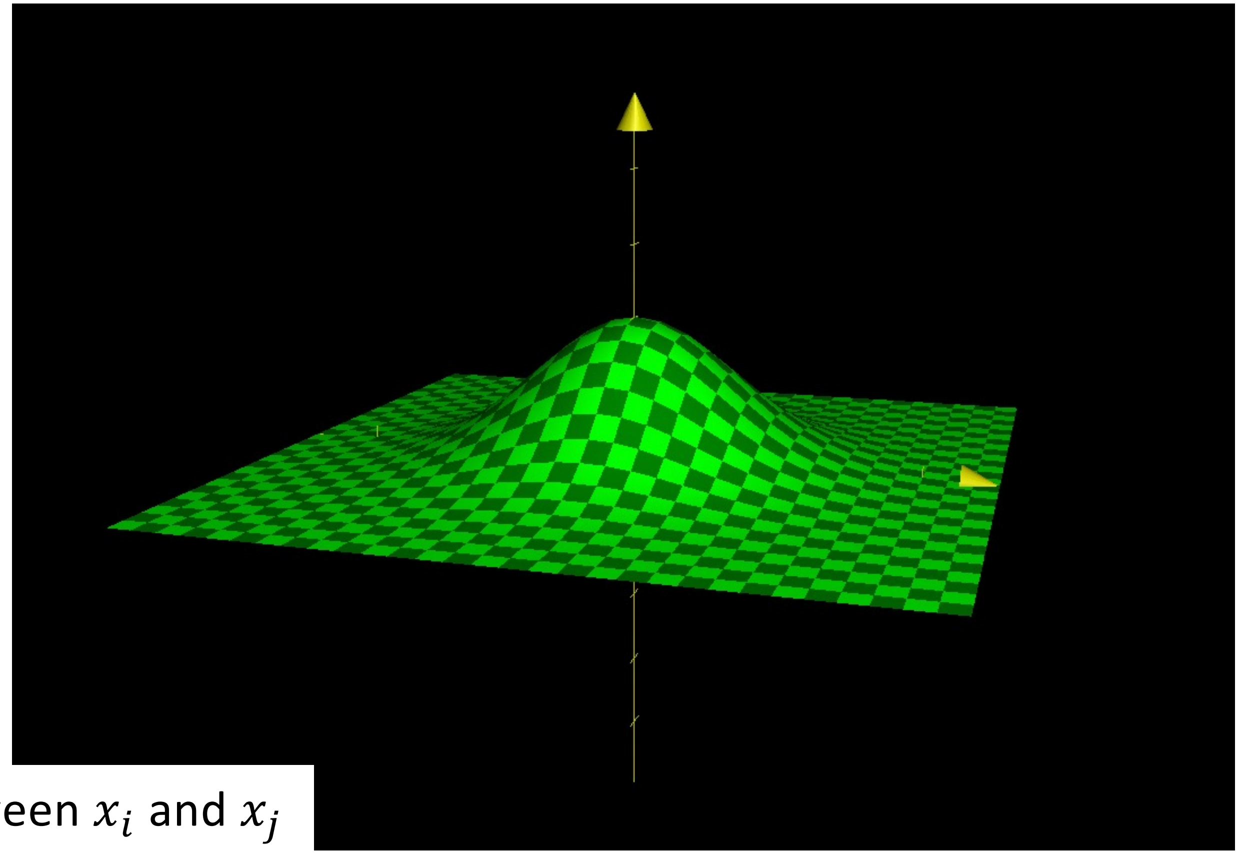
Multivariate Gaussian (normal) Distribution general form

d by 1 mean vector

d by d covariance matrix

$$\mathcal{N}(\bar{x}|\bar{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\bar{x}-\bar{\mu})^T\Sigma^{-1}(\bar{x}-\bar{\mu})]$$

d by 1 data



e.g., for d=2

$$\bar{\mu} = E\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

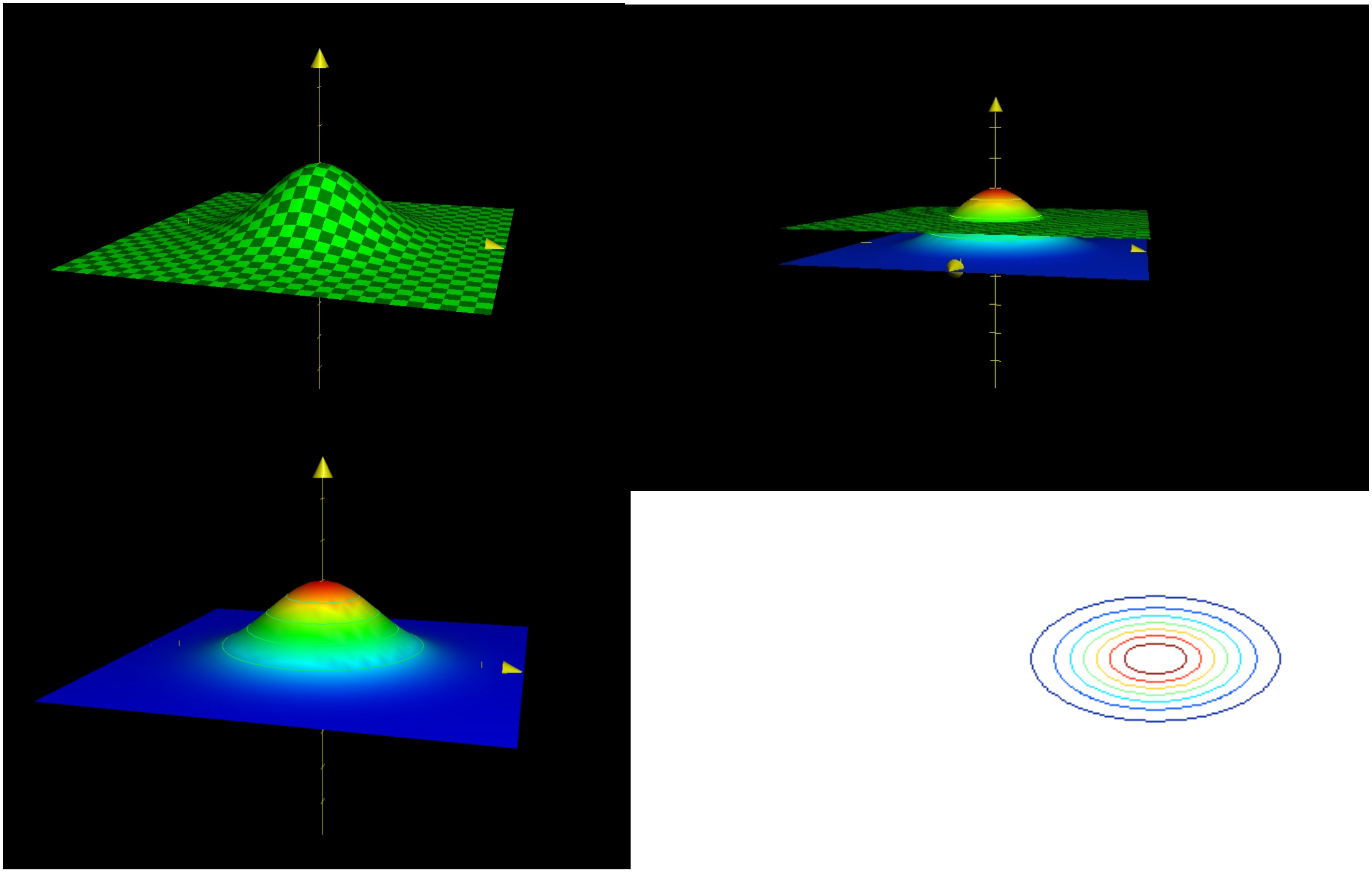
$$\Sigma = E[(\bar{x}-\bar{\mu})(\bar{x}-\bar{\mu})^T] = \begin{bmatrix} \quad \end{bmatrix}$$

Σ_{ij} measures the covariance between x_i and x_j

What does the pdf look like?

e.g., for d=2

[visualization 1](), [visualization 2]()



Contour Plots

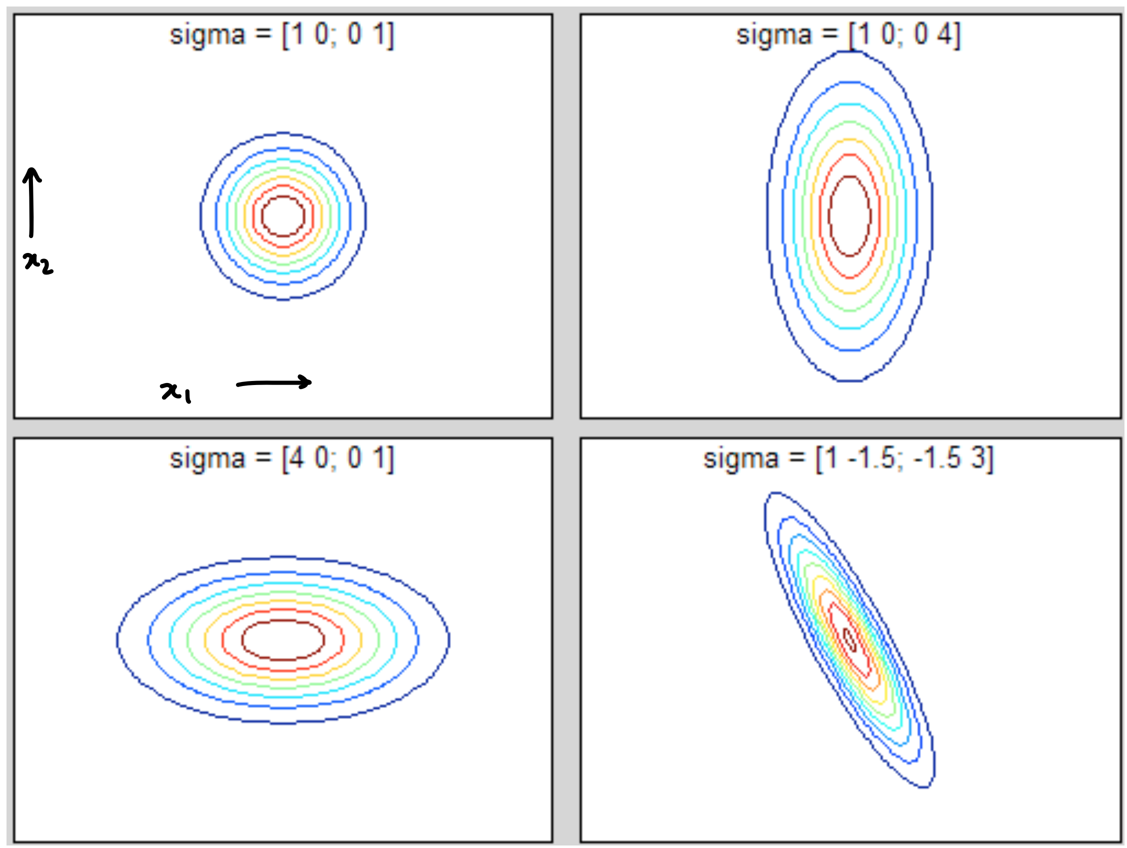
d by 1 mean vector

d by d covariance matrix

$$\mathcal{N}(\bar{x}|\bar{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\bar{x}-\bar{\mu})^T \Sigma^{-1} (\bar{x}-\bar{\mu})]$$

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

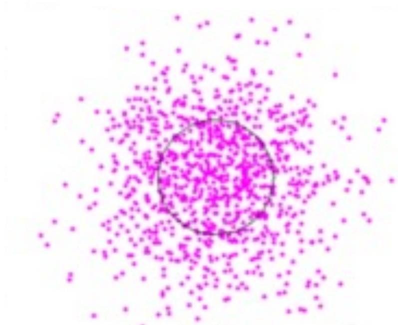
e.g., for d=2

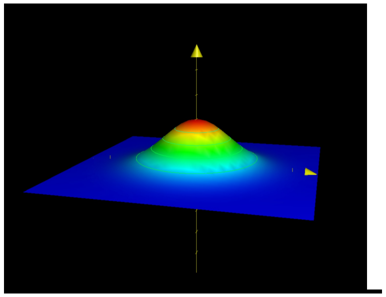


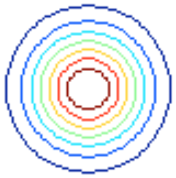
Spherical Gaussian Distribution

Maximum Likelihood Estimate

spherical Gaussian







d by 1 mean vector

d by d covariance matrix

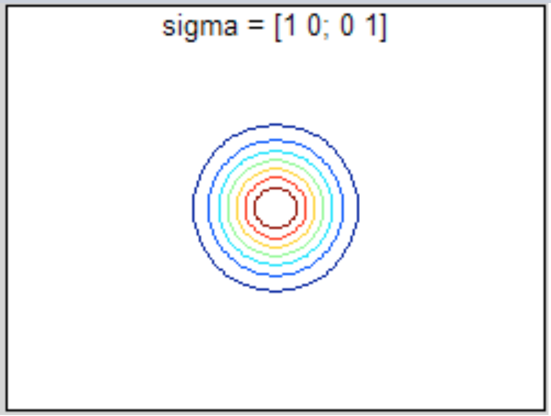
$$\mathcal{N}(\bar{x}|\bar{\mu},\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(\bar{x}-\bar{\mu})^T\Sigma^{-1}(\bar{x}-\bar{\mu})]$$

d dimensional identity matrix

Spherical Gaussian $\Sigma = \sigma^2\mathbf{I}_d$ has one free parameter

Likelihood of the Spherical Gaussian

$$\mathcal{N}(\bar{x}|\bar{\mu},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} exp^{-\frac{1}{2\sigma^2}||\bar{x}-\bar{\mu}||^2}$$



- Given $S_n = \{\bar{x}^{(i)}\}_{i=1}^{n}$ drawn iid according to $\mathcal{N}(\bar{x}|\bar{\mu},\sigma^2)$
- Want to maximize $p(S_n)$ wrt parameters $\bar{\theta} = (\bar{\mu},\sigma^2)$

$$p(S_n) = \prod_{i=1}^{n} p\left(\bar{x}^{(i)}\right) = \prod_{i=1}^{n}\left(\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\exp\left(-\frac{1}{2\sigma^2}\left\|\bar{x}^{(i)}-\bar{\mu}\right\|^2\right)\right)$$

Log Likelihood of the Spherical Gaussian

$$\mathcal{N}(\bar{x}|\bar{\mu},\sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} exp^{-\frac{1}{2\sigma^2}||\bar{x}-\bar{\mu}||^2}$$

$\ln(AB) = \ln A + \ln B$

$$l(S_n;\bar{\mu},\sigma^2) = \ln p(S_n) = \ln \prod_{i=1}^{n} p\left(\bar{x}^{(i)}\right)$$

$$= \ln \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left\|\bar{x}^{(i)} - \bar{\mu}\right\|^2\right)\right)$$

$$= \sum_{i=1}^{n} \ln \left(\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}} \exp\left(-\frac{1}{2\sigma^2}\left\|\bar{x}^{(i)} - \bar{\mu}\right\|^2\right)\right)$$

$\ln \frac{1}{A} = -\ln A$

$\ln A^c = c \ln A$

$$= \sum_{i=1}^{n} \left(\ln\left(\frac{1}{(2\pi\sigma^2)^{\frac{d}{2}}}\right) + \ln \exp\left(-\frac{1}{2\sigma^2}\left\|\bar{x}^{(i)} - \bar{\mu}\right\|^2\right)\right)$$

$$l(S_n;\bar{\mu},\sigma^2) = \sum_{i=1}^{n} \left(-\frac{d}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left\|\bar{x}^{(i)} - \bar{\mu}\right\|^2\right)$$

Spherical Gaussian: MLE of the mean $\bar{\mu}$

Data drawn iid $S_n = \{\bar{x}^{(i)}\}_{i=1}^{n}$

Log likelihood

$$l(S_n; \bar{\mu}, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{d}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2 \right)$$

$$\nabla_{\bar{\mu}} l(S_n; \bar{\mu}, \sigma^2) = \sum_{i=1}^{n} -\nabla_{\bar{\mu}} \frac{d}{2} \ln(2\pi\sigma^2) - \nabla_{\bar{\mu}} \left(\frac{1}{2\sigma^2} \left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2 \right)$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \nabla_{\bar{\mu}} \left(\quad \left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2 \right)$$

$$= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} 2\left(\bar{x}^{(i)} - \bar{\mu}\right)(-1) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\bar{x}^{(i)} - \bar{\mu}\right)$$

Set $\nabla_{\bar{\mu}} l(S_n; \bar{\mu}, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\bar{x}^{(i)} - \bar{\mu}\right) = 0$ and solve for $\bar{\mu}$.

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

Spherical Gaussian: MLE of the variance σ^2

Data drawn iid $S_n = \{\bar{x}^{(i)}\}_{i=1}^{n}$

Log likelihood

$$l(S_n; \bar{\mu}, \sigma^2) = \sum_{i=1}^{n} \left(-\frac{d}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2 \right)$$

$\frac{1}{2\pi v} \cdot 2\pi$

$$\frac{\partial l(S_n; \bar{\mu}, \sigma^2)}{\partial \sigma^2} = \sum_{i=1}^{n} -\frac{d}{2} \frac{\partial(\ln(2\pi v))}{\partial v} - \left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2 \frac{\partial \left(\frac{1}{2v} \right)}{\partial v}$$

let $v = \sigma^2$

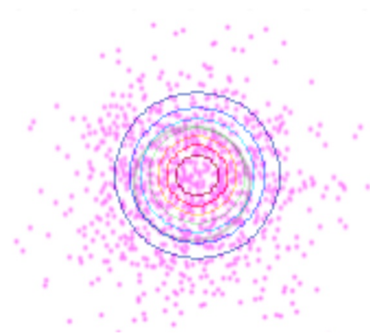
$$= \sum_{i=1}^{n} \left(-\frac{d}{2} \frac{1}{v} + \frac{\left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2}{2v^2} \right)$$

$$= -\frac{nd}{2v} + \sum_{i=1}^{n} \frac{\left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2}{2v^2}$$

Set $\frac{\partial l(S_n; \bar{\mu}, v)}{\partial v} = -\frac{nd}{2v} + \sum_{i=1}^{n} \frac{\left\| \bar{x}^{(i)} - \bar{\mu} \right\|^2}{2v^2} = 0$ and solve for v.

$$\sigma^2{}_{MLE} = \frac{\sum_{i=1}^{n} \left\| \bar{x}^{(i)} - \bar{\mu}_{MLE} \right\|^2}{nd}$$

MLE for the spherical Gaussian



- Given $S_n = \{x^{(i)}\}_{i=1}^{n}$ drawn iid

$$p(S_n) = \prod_{i=1}^{n} p\left(x^{(i)}\right)$$

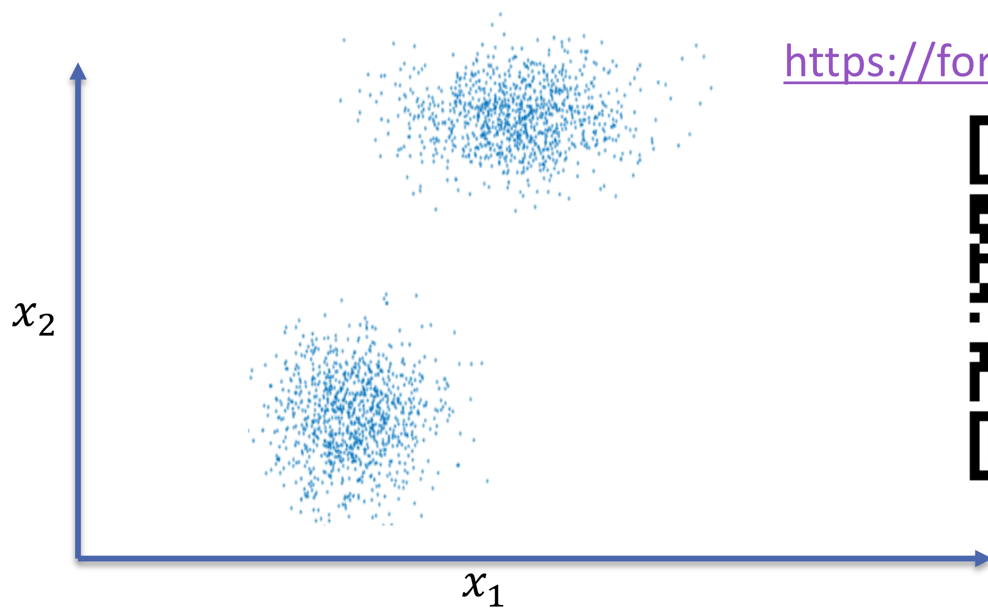
- Want to maximize $p(S_n)$ wrt μ

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

- Want to maximize $p(S_n)$ wrt σ^2

$$\sigma^2{}_{MLE} = \frac{\sum_{i=1}^{n} \left\|\bar{x}^{(i)} - \bar{\mu}_{MLE}\right\|^2}{nd}$$

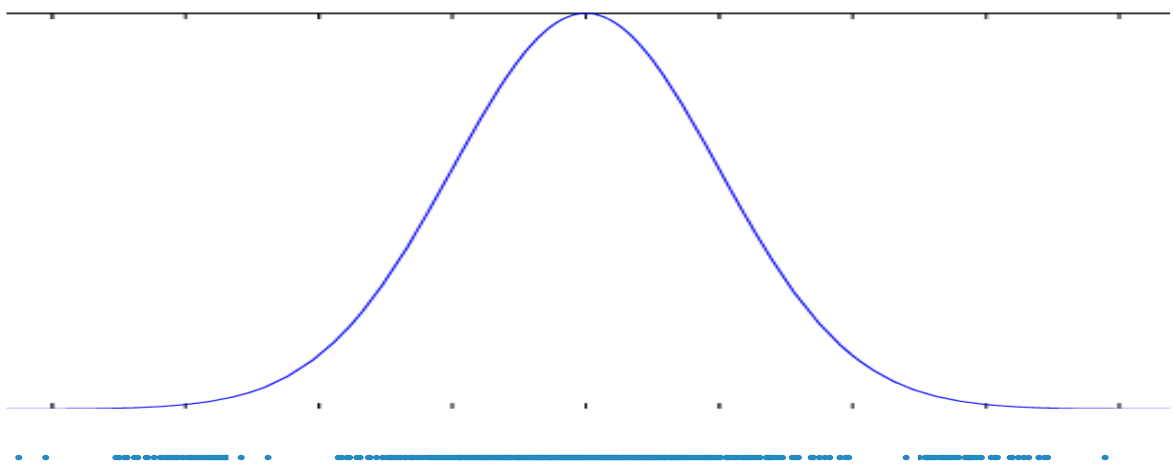
Mixture Distributions

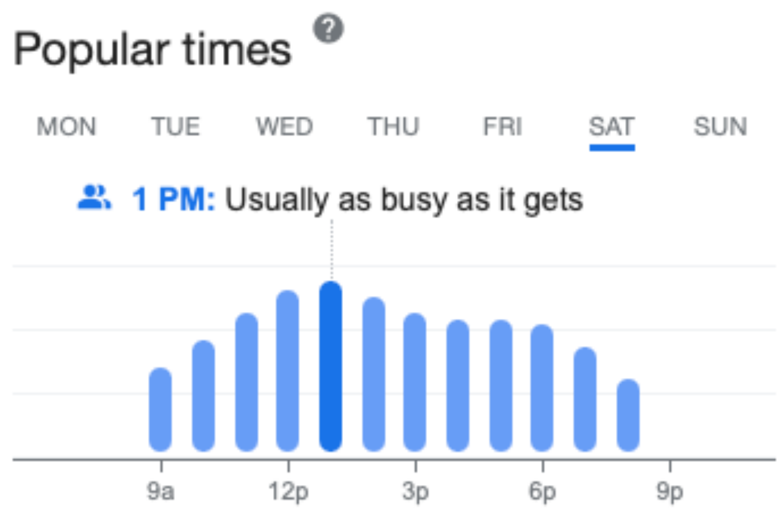


https://forms.gle/ffiBvNbPjHF8ghi77

Why Mixture of Distributions?

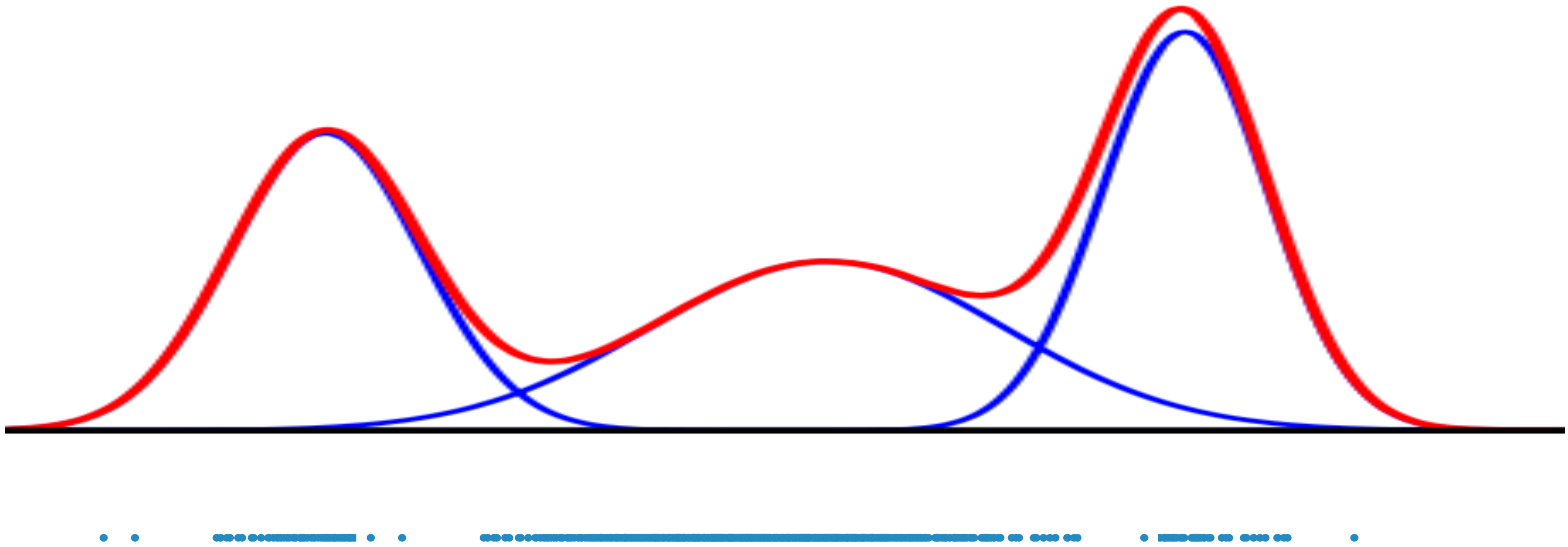
A single Gaussian is not a good fit for this dataset



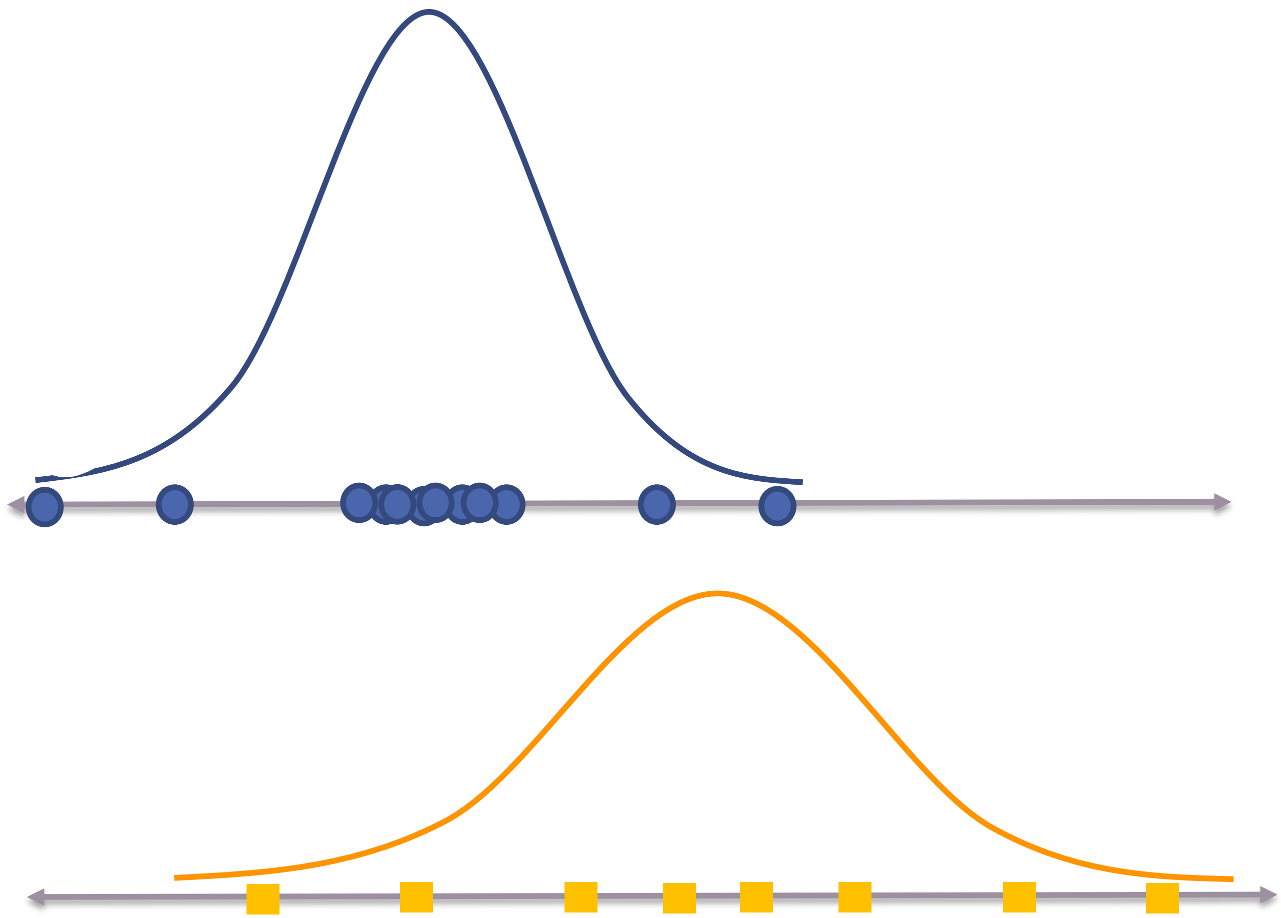


Mixture of Distributions

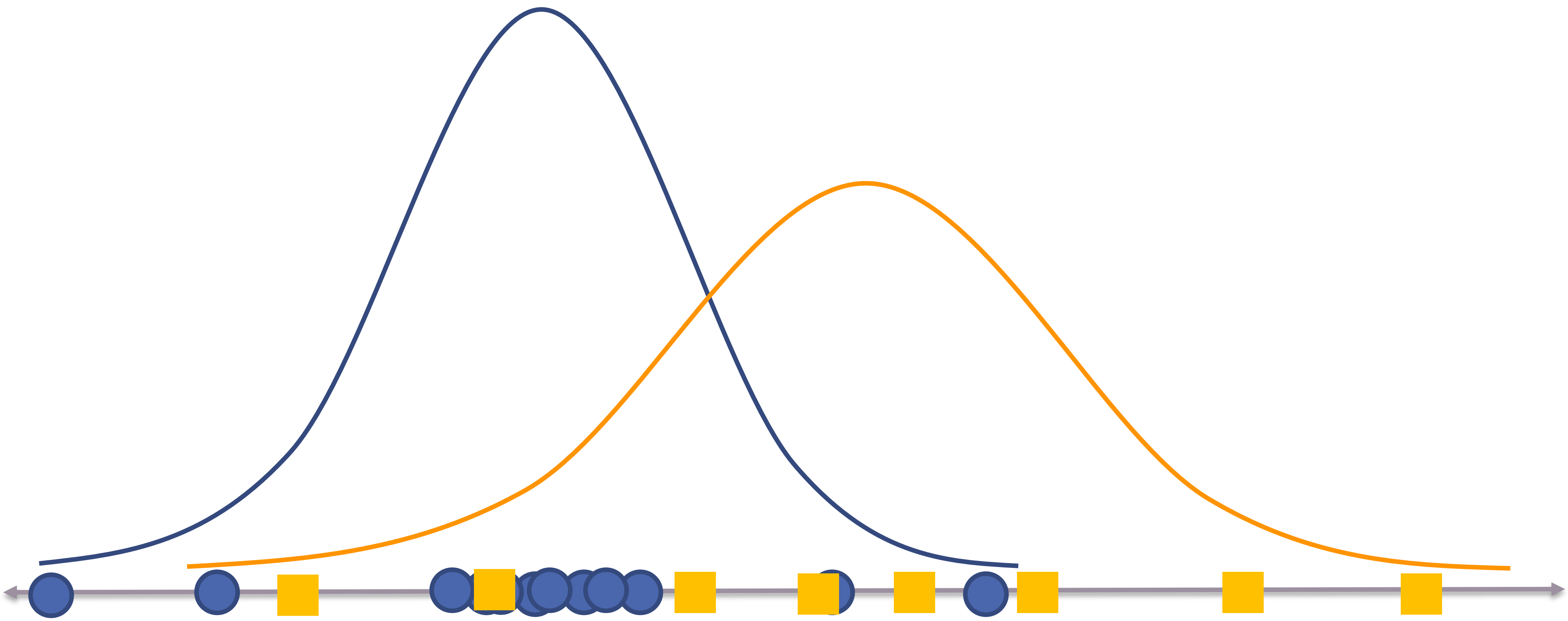
In this model each datapoint $\bar{x}^{(i)}$ is assumed to be generated from a mixture of k distributions.



MLE of a single Gaussian: intuition

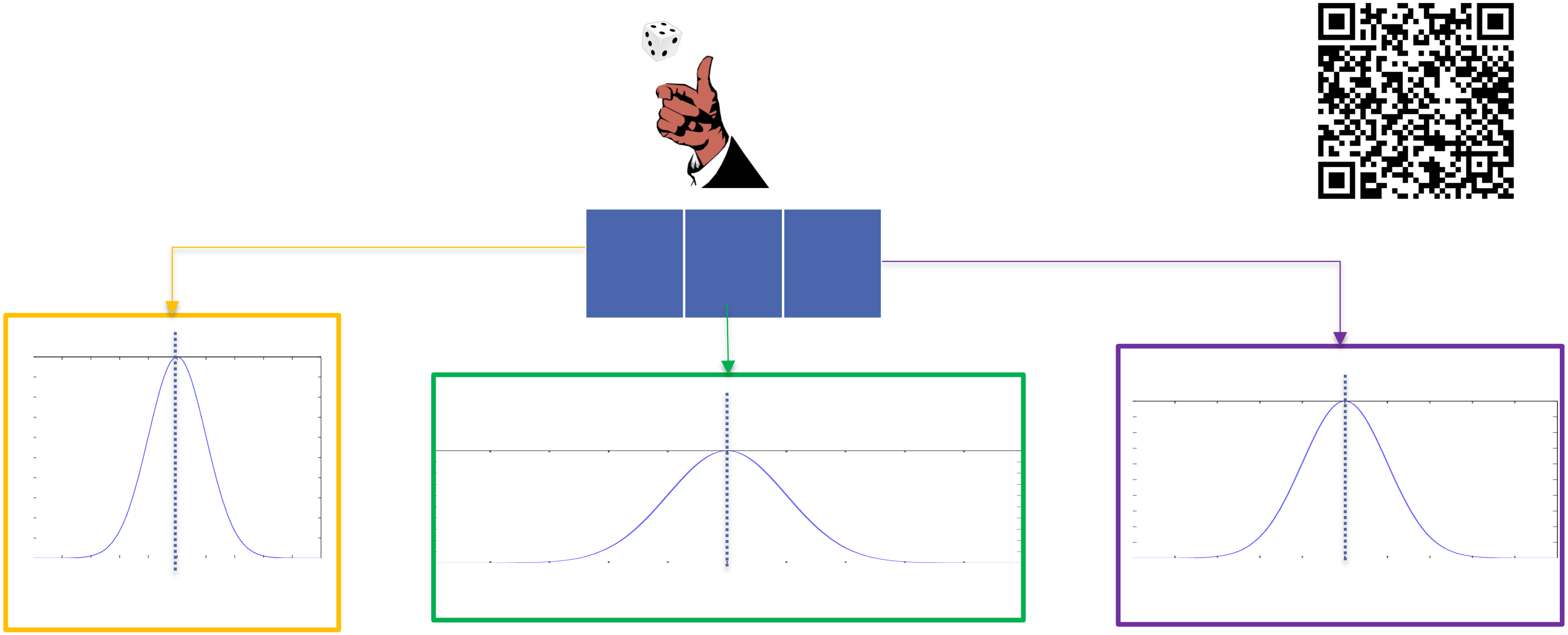


MLE of GMM with *known* labels: intuition



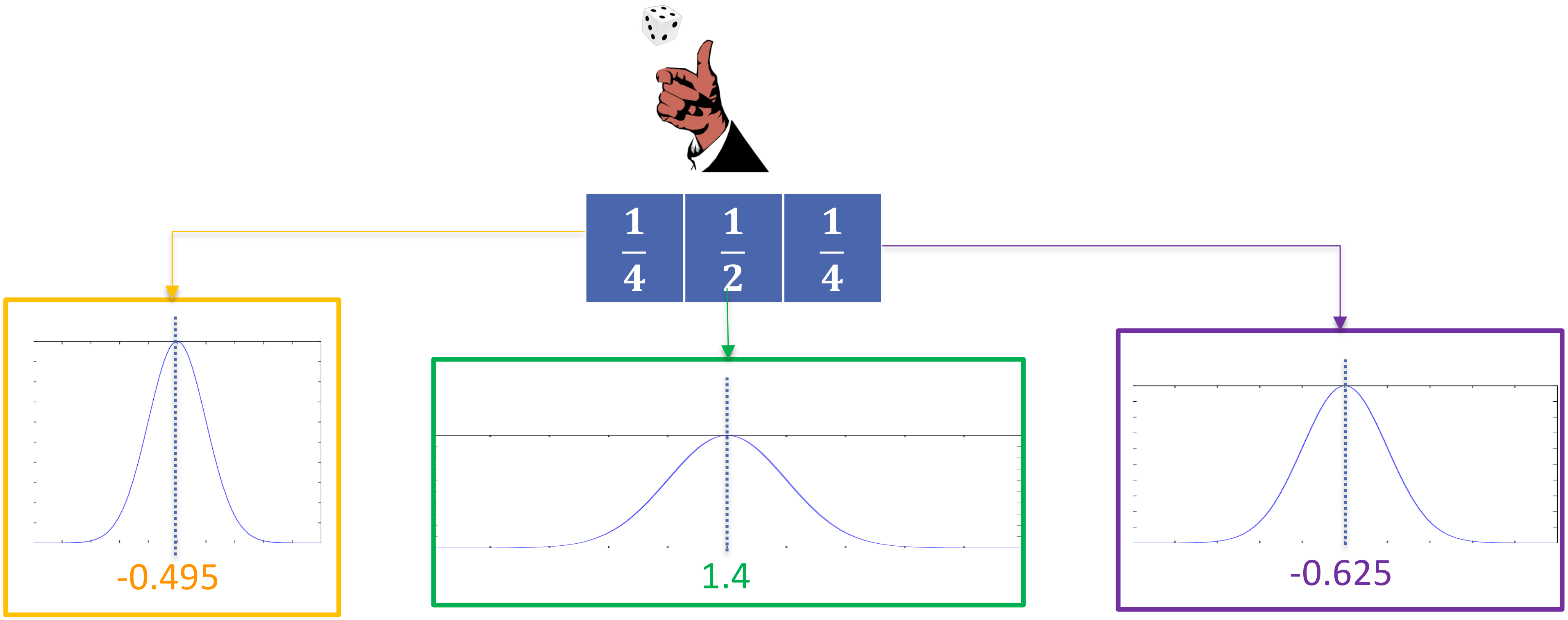
can also determine relative chance of each Gaussian

MLE of GMM with known labels: Example



$2.1, 0, 3.5, -1, 1.5, 2.5, -0.5, 0.05, 1, -2, 0, 1, -2, 1.1, -0.5, -0.03$

MLE of GMM with known labels: Example

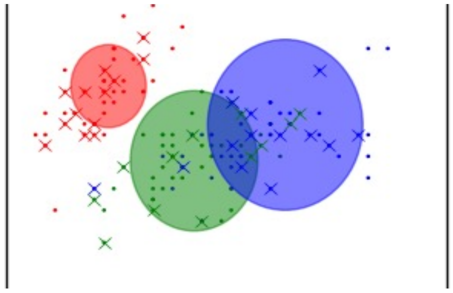


$2.1, 0, 3.5, -1, 1.5, 2.5, -0.5, 0.05, 1, -2, 0, 1, -2, 1.1, -0.5, -0.03$

MLE for GMMs with known labels

Define indicator function

$$\delta(j \mid i) = \begin{cases} 1 & \text{if } \bar{x}^{(i)} \text{ belongs to cluster } j \\ 0 & \text{otherwise} \end{cases}$$



Log likelihood objective

$$\ln \prod_{i=1}^{n} \Pr\left(\bar{x}^{(i)}, y^{(i)} \middle| \, \bar{\theta}\right)$$

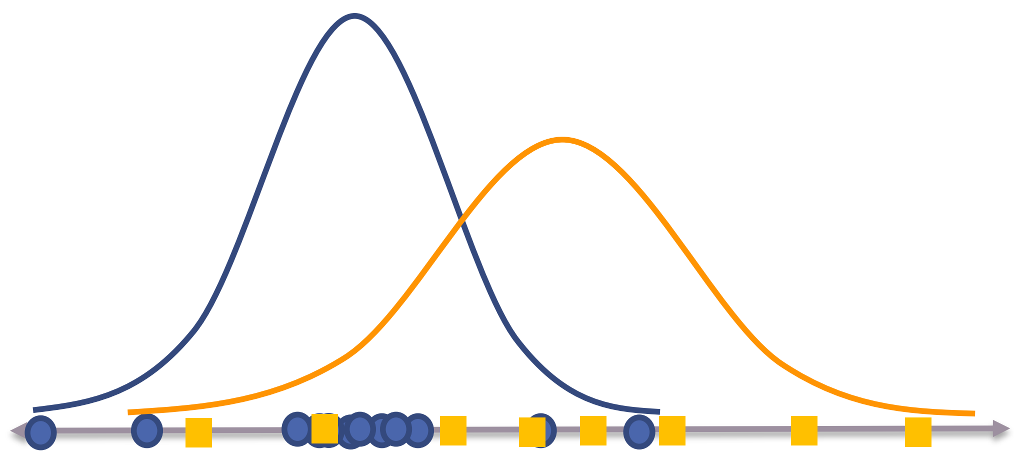
Log-Likelihood for GMMs with known labels

Product rule $Pr(A,B) = Pr(A|B)Pr(B)$

$$P(S_n) = \prod_{i=1}^{n} p(\bar{x}^{(i)}, y^{(i)})$$

$$= \prod_{i=1}^{n} p(\bar{x}^{(i)}|y^{(i)})p(y^{(i)})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{k} \delta(j \mid i)(N(\bar{x}^{(i)}|\bar{\mu}^{(j)}, \sigma_j^2)\gamma_j)$$



Maximum log likelihood objective

$$\ln P(S_n) = \ln \prod_{i=1}^{n} \sum_{j=1}^{k} \delta(j \mid i)(N(\bar{x}^{(i)}|\bar{\mu}^{(j)}, \sigma_j^2)\gamma_j)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} \delta(j \mid i) \ln (\gamma_j N(\overline{x}^{(i)}|\overline{\mu}^{(j)}, \sigma_j^2))$$

MLE for GMMs with known labels

Maximum log likelihood objective

$$\sum_{i=1}^{n}\sum_{j=1}^{k}\delta(j \mid i)\ln\left(\gamma_j N(\bar{x}|\bar{\mu}^{(j)},\sigma_j^2)\right)$$

MLE solution (given “cluster labels”):

Define $\hat{n}_j = \sum_{i=1}^{n}\delta(j \mid i)$ number of points assigned to cluster j

$\gamma_j = \frac{\hat{n}_j}{n}$ fraction of points assigned to cluster j

$\bar{\mu}^{(j)} = \frac{1}{\hat{n}_j}\sum_{i=1}^{n}\delta(j \mid i)\ \bar{x}^{(i)}$ mean of points in cluster j

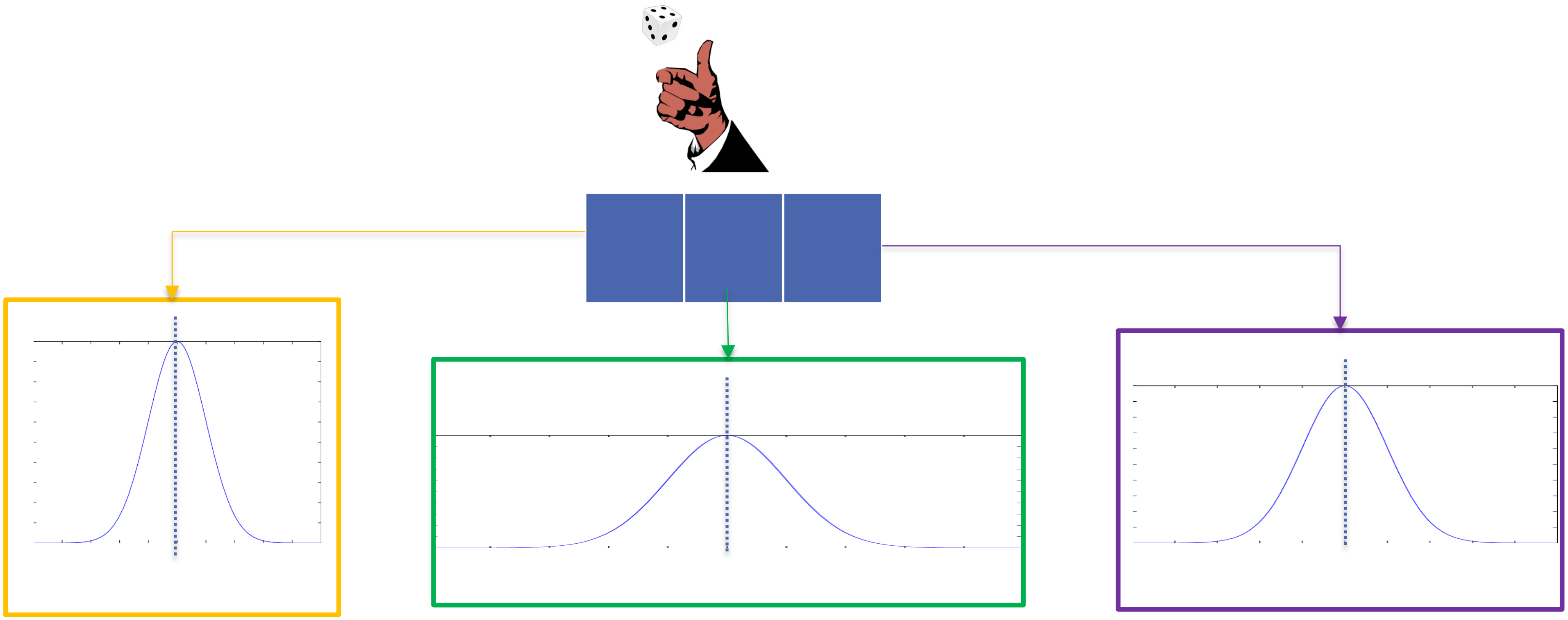
$\sigma_j^2 = \frac{1}{d\hat{n}_j}\sum_{i=1}^{n}\delta(j \mid i)\left\|\bar{x}^{(i)} - \bar{\mu}^{(j)}\right\|^2$ spread in cluster j

MLE for GMMs with known labels

Issue?

In general, $\delta(j|i)$ is unknown!

Parameters of GMMs



$$2.1, 0, 3.5, -1, 1.5, 2.5, -0.5, 0.05, 1, -2, 0, 1, -2, 1.1, -0.5, -0.03$$

Expectation Maximization for GMMs

- **E-step**: fix $\bar{\theta} = [\gamma_1, ..., \gamma_k, \bar{\mu}^{(1)}, ..., \bar{\mu}^{(k)}, \sigma_1^2, ..., \sigma_k^2]$

  softly assign points to clusters according to posterior prob

$$p(j|i) = \frac{\gamma_j N(\bar{x}^{(i)} \mid \bar{\mu}_j, \sigma_j^2)}{\sum_t \gamma_t N(\bar{x}^{(i)} \mid \bar{\mu}_t, \sigma_t^2)}$$

$0 \leq p(j|i) \leq 1$

Expectation Maximization for GMMs

- **M-Step**: optimizes each cluster separately given $p(j|i)$

$$\hat{n}_j = \sum_{i=1}^{n} p(j|i) \qquad \hat{\bar{\mu}}^{(j)} = \frac{1}{\hat{n}_j} \sum_{i=1}^{n} p(j|i)\bar{x}^{(i)}$$

$$\hat{\gamma}_j = \frac{\hat{n}_j}{n} \qquad \hat{\sigma}_j^2 = \frac{1}{d\hat{n}_j} \sum_{i=1}^{n} p(j|i)||\bar{x}^{(i)} - \hat{\bar{\mu}}^{(j)}||^2$$

Expectation Maximization for GMMs: M step (note correspondence with known labels)

if you knew the "soft" cluster assignment $p(j|i)$,

you could compute MLE parameters $\bar{\theta}$ as follows

MLE for GMM with known labels

$\hat{n}_j = \sum_{i=1}^{n} \delta(j \mid i)$

$$\hat{n}_j = \sum_{i=1}^{n} p(j|i)$$

effective number of points assigned to cluster j

$\gamma_j = \frac{\hat{n}_j}{n}$

$$\hat{\gamma}_j = \frac{\hat{n}_j}{n}$$

"fraction" of points assigned to cluster j

$\bar{\mu}^{(j)} = \frac{1}{\hat{n}_j} \sum_{i=1}^{n} \delta(j \mid i)\ \bar{x}^{(i)}$

$$\hat{\bar{\mu}}^{(j)} = \frac{1}{\hat{n}_j} \sum_{i=1}^{n} p(j|i) \bar{x}^{(i)}$$

weighted mean of points in cluster j

$\sigma_j^2 = \frac{1}{d\hat{n}_j} \sum_{i=1}^{n} \delta(j \mid i) \left\|\bar{x}^{(i)} - \bar{\mu}^{(j)}\right\|^2$

$$\hat{\sigma}_j^2 = \frac{1}{d\hat{n}_j} \sum_{i=1}^{n} p(j|i) ||\bar{x}^{(i)} - \hat{\bar{\mu}}^{(j)}||^2$$

weighted spread in cluster j