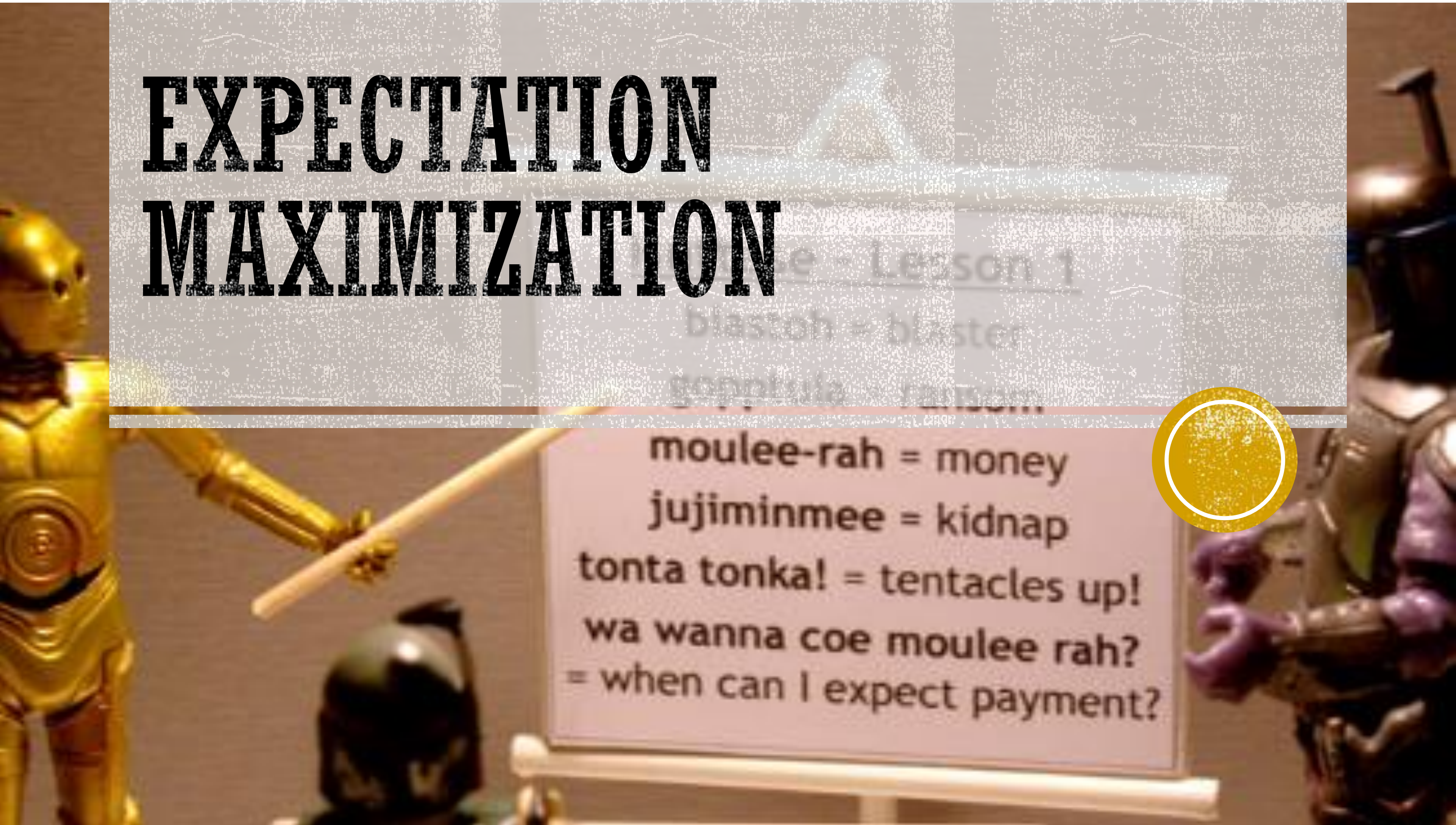
EXPECTATION MAXIMIZATION



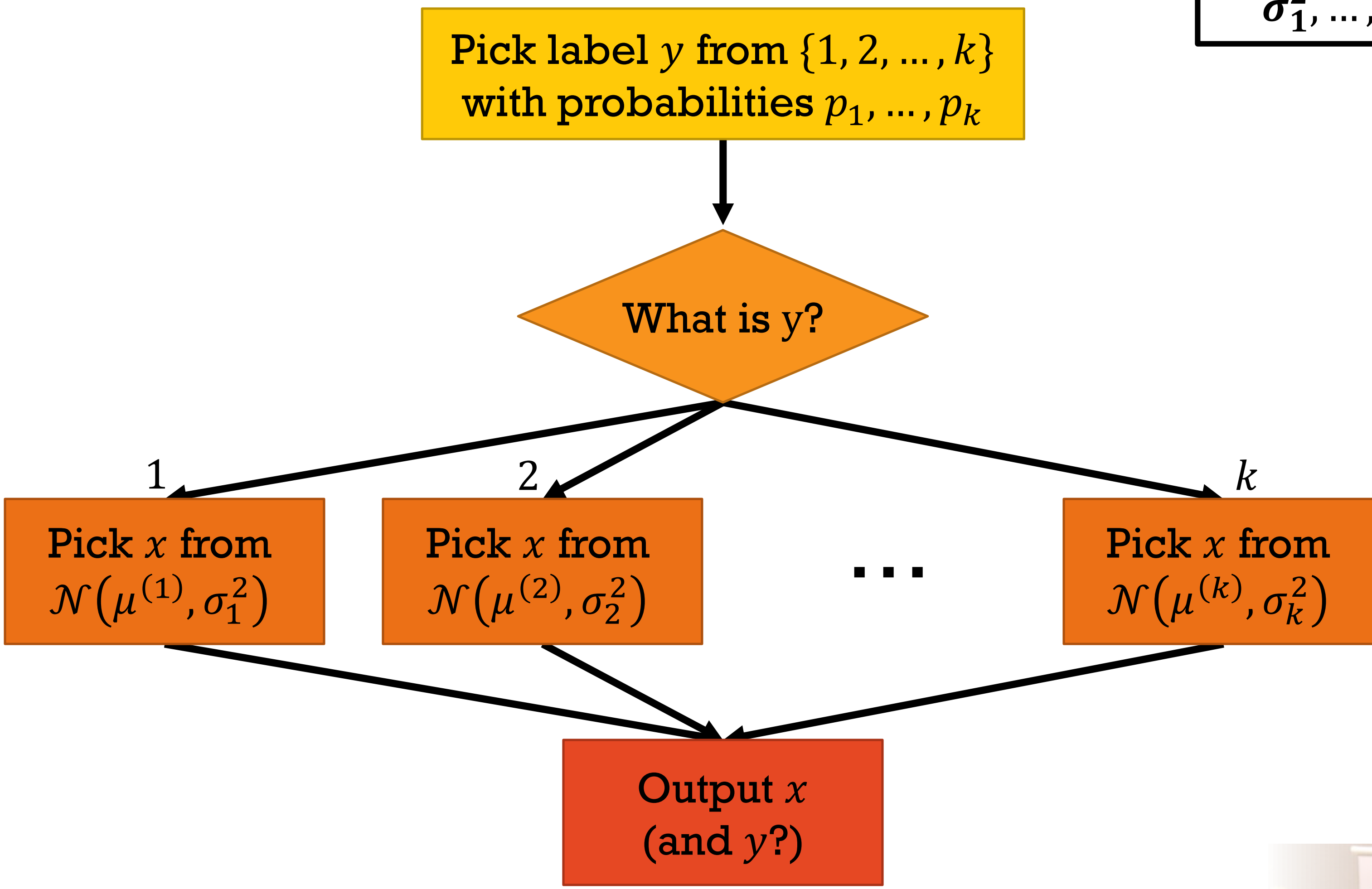
GENERATIVE MODEL

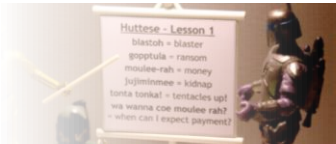
Model Parameters

$\boldsymbol{p_1, \ldots, p_k}$

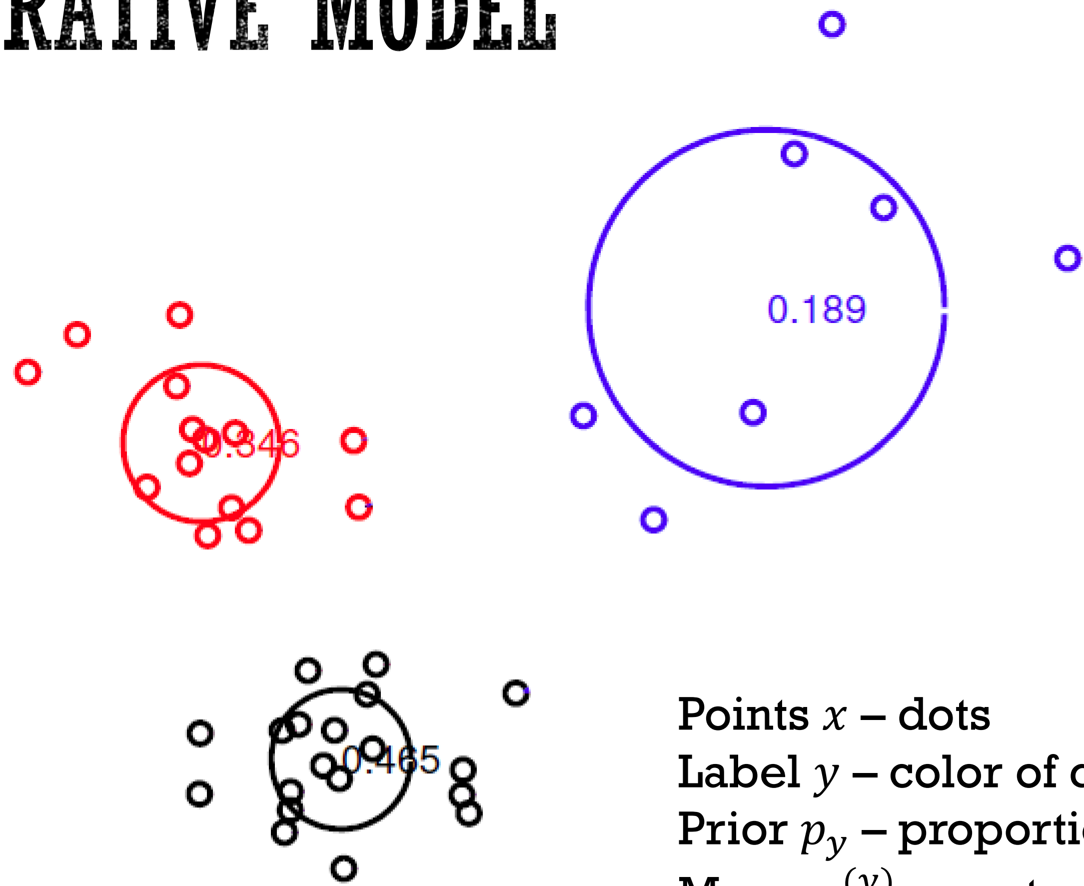
$\boldsymbol{\mu^{(1)}, \ldots, \mu^{(k)}}$

$\boldsymbol{\sigma_1^2, \ldots, \sigma_k^2}$

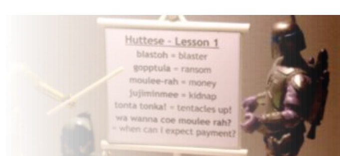




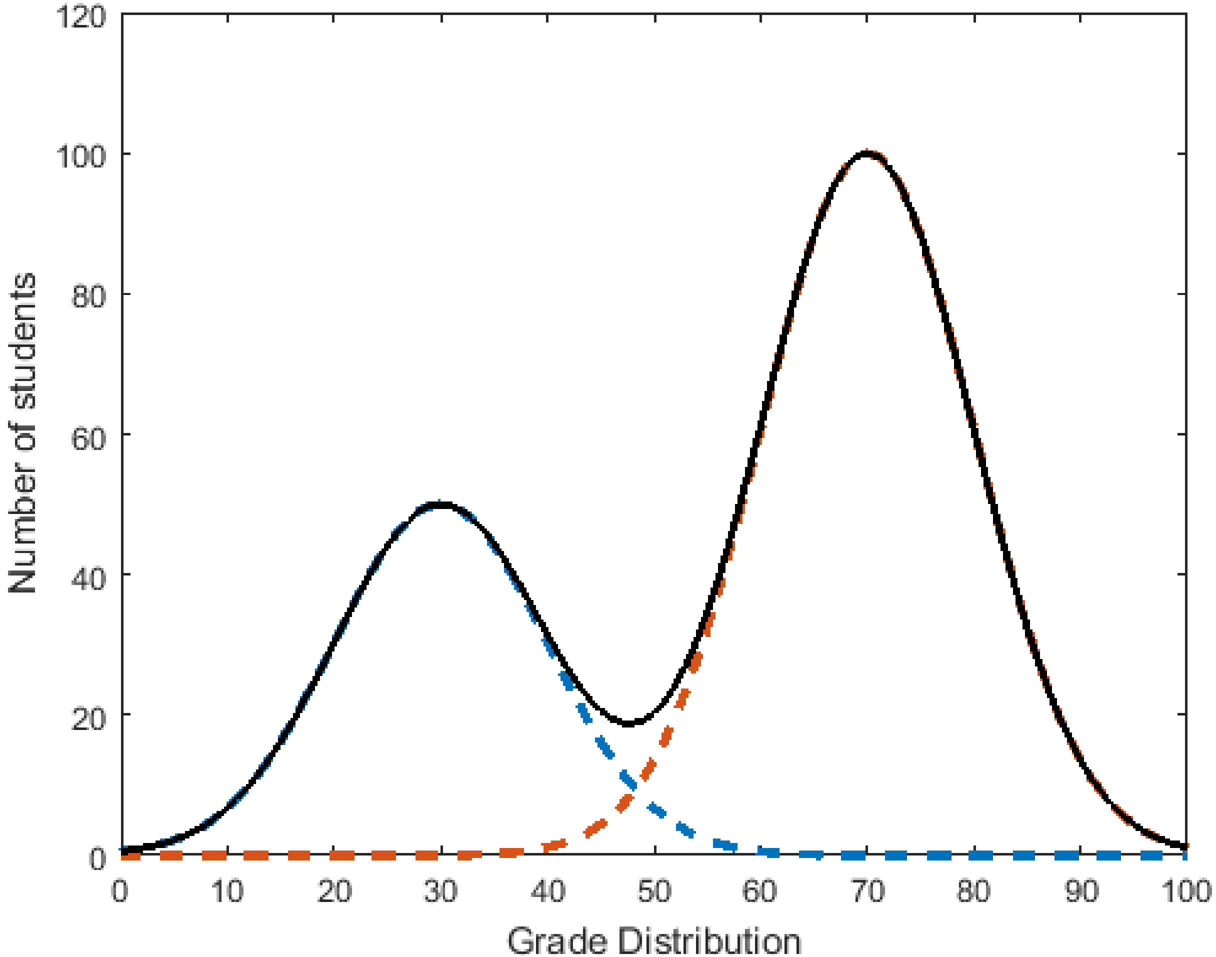
Generative Model

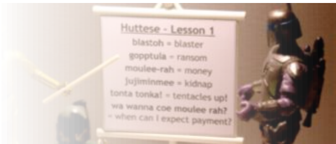


Points x – dots
Label y – color of dots
Prior p_y – proportion of dots
Mean $\mu^{(y)}$ – center of circle
Variance σ_y^2 – size of circle



GENERATIVE MODEL





Gaussian Mixture Models

Given training set $x^{(1)}, \ldots, x^{(N)}$, we wish to model the data by specifying the joint distribution of X with a latent variable Z:

$$p(x, z) = p(x|z)p(z).$$

Here $Z \sim \mathrm{Multinomial}(\pi)$, i.e.

$$P(Z = j) = \pi_j, \quad j = 1, \ldots, m,$$

and $\sum_{j=1}^{m} \pi_j = 1$, and we have

$$X \,\big|\, \{Z = j\} \sim \mathcal{N}(\mu_j, \Sigma_j).$$

- This means that the log-likelihood of the data is given by

$$
\begin{aligned}
\ell(\pi, \mu, \Sigma) = \sum_{i=1}^{N} \log p(x^{(i)}) &= \sum_{i=1}^{N} \log \sum_{j=1}^{m} p(x^{(i)} \mid Z = j) P(Z = j) \\
&= \sum_{i=1}^{N} \log \sum_{j=1}^{m} \pi_j \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)} - \mu_j, \Sigma_j^{-1}(x^{(i)} - \mu_j)\right\rangle} \right)
\end{aligned}
$$

- Unfortunately, there is no closed-form solution to optimizing this log-likelihood. Let's see why by examining the conditions we get when we try to optimize $\ell(\pi, \mu, \Sigma)$.

- Differentiating with respect to μ_k and setting to zero, we get

$$\sum_{i=1}^{N} \frac{\pi_k \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_j, \Sigma_j^{-1}(x^{(i)}-\mu_j)\right\rangle} \right) \Sigma^{-1} \left(x^{(i)} - \mu_k \right)}{\sum_{j=1}^{m} \pi_j \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_j, \Sigma_j^{-1}(x^{(i)}-\mu_j)\right\rangle} \right)} = 0$$

- We will denote

$$\gamma\left(z_k^{(i)}\right) := \frac{\pi_k \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_k, \Sigma_k^{-1}(x^{(i)}-\mu_k)\right\rangle} \right)}{\sum_{j=1}^{m} \pi_j \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_j, \Sigma_j^{-1}(x^{(i)}-\mu_j)\right\rangle} \right)},$$

and rearranging, we get

$$\mu_k = \frac{\sum_{i=1}^{N} x^{(i)} \gamma\left(z_k^{(i)}\right)}{\sum_{i=1}^{N} \gamma\left(z_k^{(i)}\right)}. \tag{1}$$

- Similarly, differentiating with respect to Σ_k^{-1} and setting to zero, we get

$$\Sigma_k = \frac{\sum_{i=1}^{N} \gamma\left(z_k^{(i)}\right)\left(x^{(i)} - \mu_k\right)\left(x^{(i)} - \mu_k\right)^T}{\sum_{i=1}^{N} \gamma\left(z_k^{(i)}\right)}. \tag{2}$$

- Since π satisfies the constraint $\sum_{j=1}^{m} \pi_j = 1$, we'll have to use Lagrange multipliers to optimize with respect to π. We have

$$\frac{\mathrm{d}}{\mathrm{d}\pi_k}\left[\sum_{i=1}^{N} \log \sum_{j=1}^{m} \pi_j \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_j, \Sigma_j^{-1}(x^{(i)}-\mu_j)\right\rangle}\right) - \lambda\left(\sum_{j=1}^{m} \pi_j - 1\right)\right] = 0,$$

- which implies that

$$\sum_{i=1}^{N} \frac{\left(C_k \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_k, \Sigma_k^{-1}(x^{(i)}-\mu_k)\right\rangle}\right)}{\sum_{j=1}^{m} \pi_j \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)}-\mu_j, \Sigma_j^{-1}(x^{(i)}-\mu_j)\right\rangle}\right)} = \lambda.$$

- Multiplying by π_k on both sides, we get

$$\sum_{i=1}^{N} \gamma\left(z_k^{(i)}\right) = \lambda \pi_k.$$

- Summing over k gives us $\lambda = N$, and thus

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \gamma\left(z_k^{(i)}\right). \tag{3}$$

As suggested in the notation, $\gamma\left(z_k^{(i)}\right)$ represents an important quantity related to the latent variable Z. Indeed, if we compute

$$P\left(Z = k \,|\, x^{(i)}\right) = \frac{p\left(x^{(i)} \,|\, Z = k\right) P(Z = k)}{p\left(x^{(i)}\right)}$$

using Bayes' rule, we get

$$\begin{aligned} P\left(Z = k \,| x^{(i)}\right) &= \frac{\pi_k \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)} - \mu_k, \Sigma_k^{-1}\left(x^{(i)} - \mu_k\right)\right\rangle}\right)}{\sum_{j=1}^{m} \pi_j \left(C_j \exp^{-\frac{1}{2}\left\langle x^{(i)} - \mu_j, \Sigma_j^{-1}\left(x^{(i)} - \mu_j\right)\right\rangle}\right)} \\ &= \gamma\left(z_k^{(i)}\right). \end{aligned}$$

As $\gamma\left(z_k^{(i)}\right)$ contains the parameters μ, Σ and π in a complex way, (1), (2) and (3) cannot be solved in closed-form. However, it suggests the following two-step algorithm:

- E-step: Set

$$\gamma_{t+1}\left(z_k^{(i)}\right) = P_{\mu(t),\Sigma(t),\pi(t)}\left(Z = k \,\middle|\, x^{(i)}\right)$$

- M-step: Set

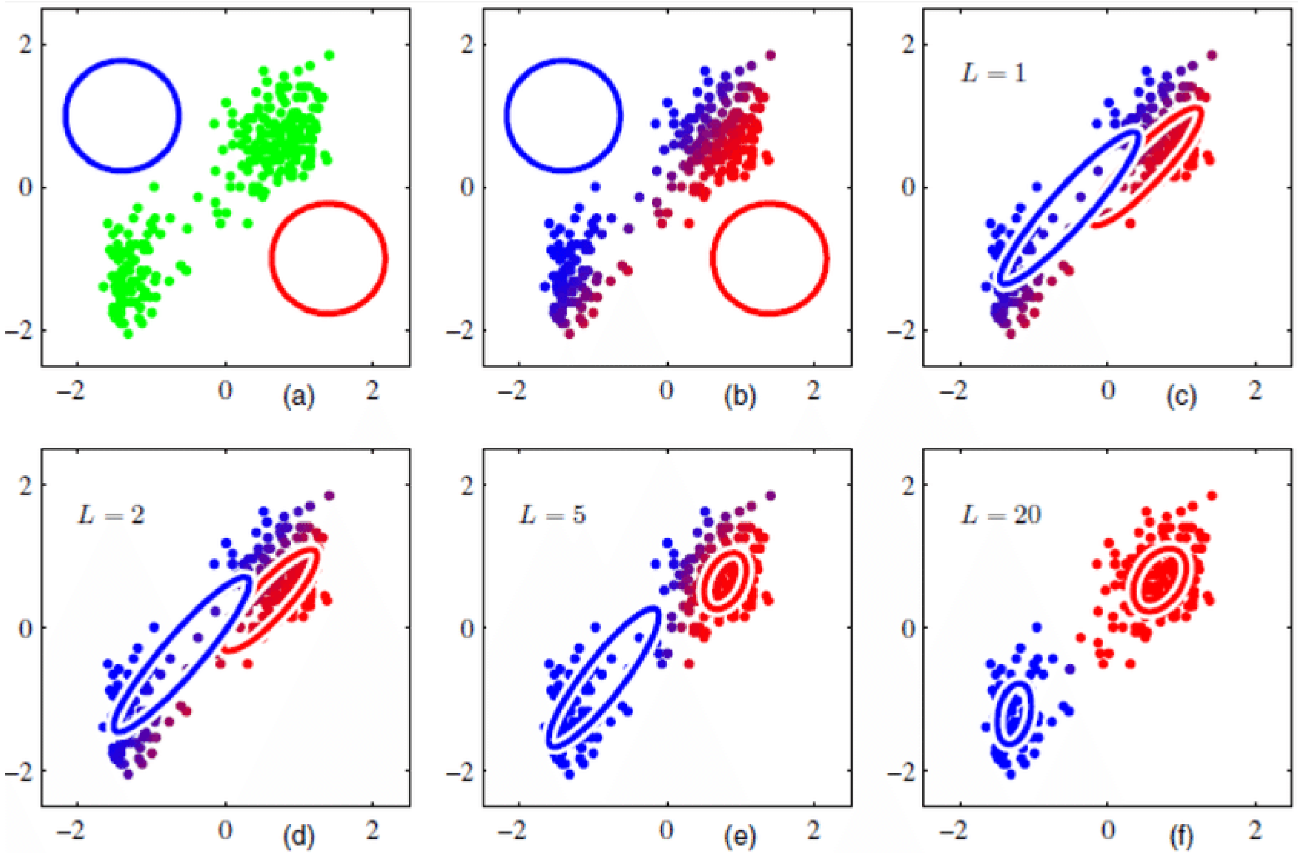
$$\pi_k(t+1) = \frac{1}{N}\sum_{i=1}^{N} \gamma_{t+1}\left(z_k^{(i)}\right)$$

$$\mu_k(t+1) = \frac{\sum_{i=1}^{N} x^{(i)}\gamma_{t+1}\left(z_k^{(i)}\right)}{\sum_{i=1}^{N} \gamma_{t+1}\left(z_k^{(i)}\right)}$$

$$\Sigma_k(t+1) = \frac{\sum_{i=1}^{N} \gamma_{t+1}\left(z_k^{(i)}\right)\left(x^{(i)} - \mu_k(t+1)\right)\left(x^{(i)} - \mu_k(t+1)\right)^T}{\sum_{i=1}^{N} \gamma_{t+1}\left(z_k^{(i)}\right)}.$$

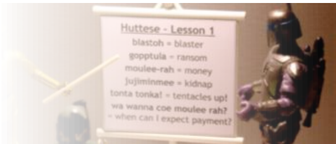
- Repeat until convergence.

EM on Old Faithful dataset



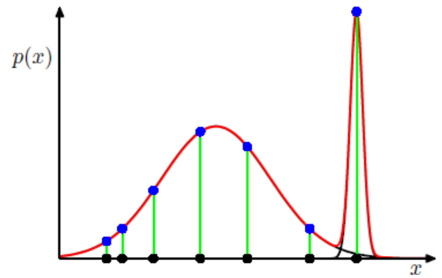
COMPARISON WITH K-MEANS

- Like k-means, EM clustering may get stuck in local minima.
- Unlike k-means, the local minima are more favorable because soft labels allow points to move between clusters slowly.



Potential problems with Gaussian mixture models

- Singularities may arise:



If the center μ_j of one of the Gaussians happens to coincide with some data point $x^{(i)}$, we have a contribution of the form

$$\frac{1}{\sqrt{2\pi}\sigma_j},$$

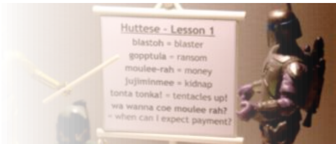
which can make the likelihood go to infinity if $\sigma_j \to 0$.

Potential problems cont.

- Identifiability problem:
  For any given maximum likelihood solution, a K-component mixture model will have a total of $K!$ equivalent solutions corresponding to the $K!$ different ways of assigning K sets of parameters to K components.

MODEL SELECTION

- By setting $p_{k+1} = 0$, we see that (mixture model with k clusters) contained in (mixture model with $k+1$ clusters).
- Therefore, likelihood for (mixture model with $k+1$ clusters) is greater or equal to that of (mixture model with k clusters).
- How to choose the right k and prevent over-/under-fitting?



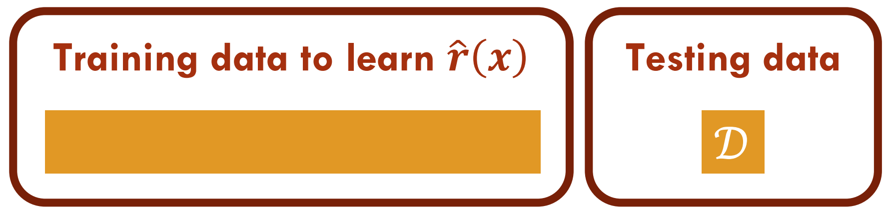
VALIDATION VS CROSS-VALIDATION

Method 1 (Simulation)

Estimate testing error using simple validation or cross-validation.

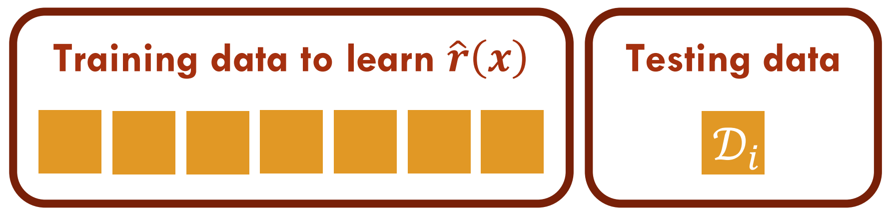
testing error

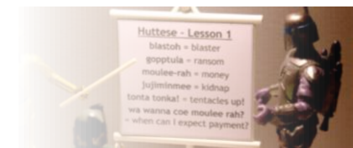
- $\hat{R}(\mathcal{D})$



***k*-fold cross-validation**.

- $\hat{R}_{\text{CV}} = \frac{1}{m}\sum_{i=1}^{m} \hat{R}(\mathcal{D}_i)$





BAYESIAN INFORMATION CRITERION

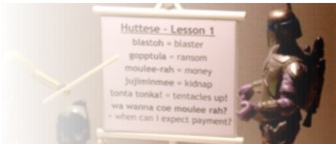
Method 2 (Marginal Likelihood)

Maximize the marginal likelihood integral. But computing this integral is tedious, so we approximate it using the BIC.

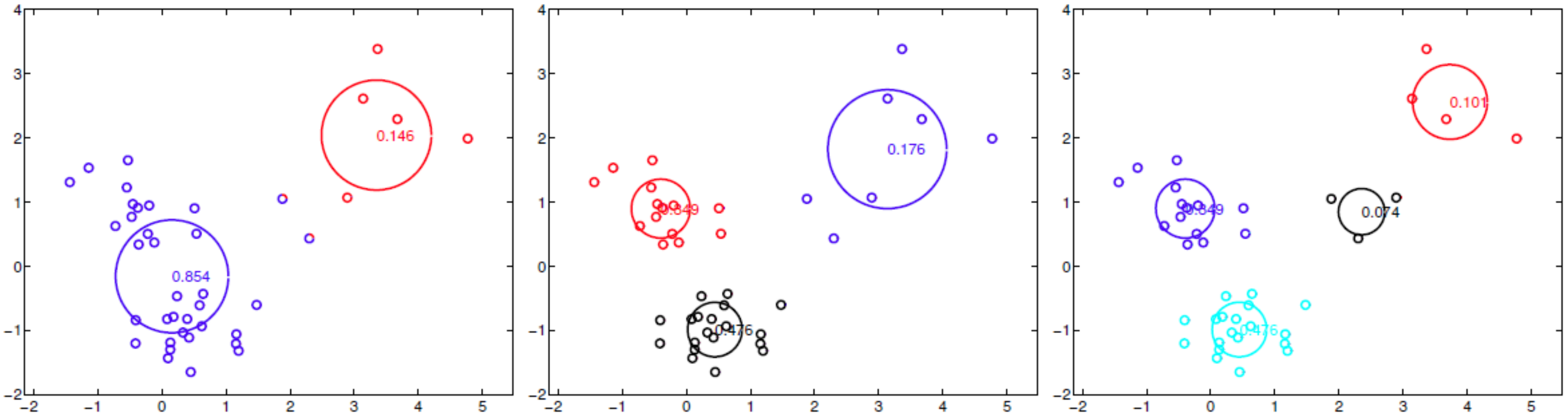
$$\mathrm{BIC}(\theta) = \mathcal{L}_n(\theta) - \frac{\text{\# of free params}}{2} \log n$$

For Gaussian mixtures, we have $k(d+2) - 1$ free parameters.

$$\mathrm{BIC}(\theta) = \mathcal{L}_n(\theta) - \frac{k(d+2)-1}{2} \log n$$



BAYESIAN INFORMATION CRITERION



$l(D;\hat{\theta}) = -118.25$ $\quad$ $l(D;\hat{\theta}) = -98.64$ $\quad$ $l(D;\hat{\theta}) = -94.11$

$BIC(D;\hat{\theta}) = -131.16$ $\quad$ $BIC(D;\hat{\theta}) = -118.93$ $\quad$ $BIC(D;\hat{\theta}) = -121.78$