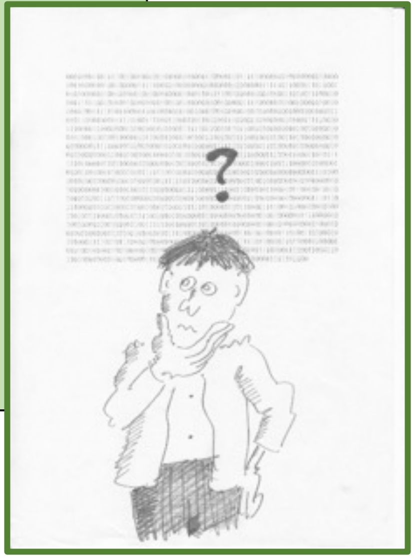
MLE, EM

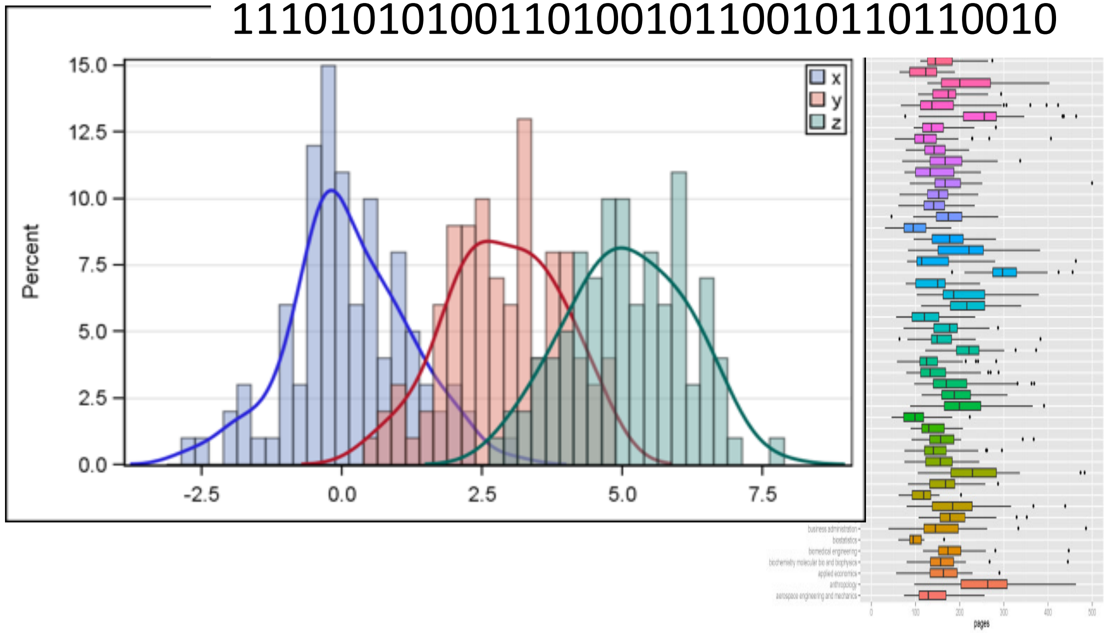
Statistics and data analysis

Leon Anavy
Zohar Yakhini

IDC, Herzeliya



001001110101010010101010010010001 0
101010001010111110101101001100100 1
111010101001101001011001011011001 0



Slides also by Ben Galili and Arik Shamir, IDC

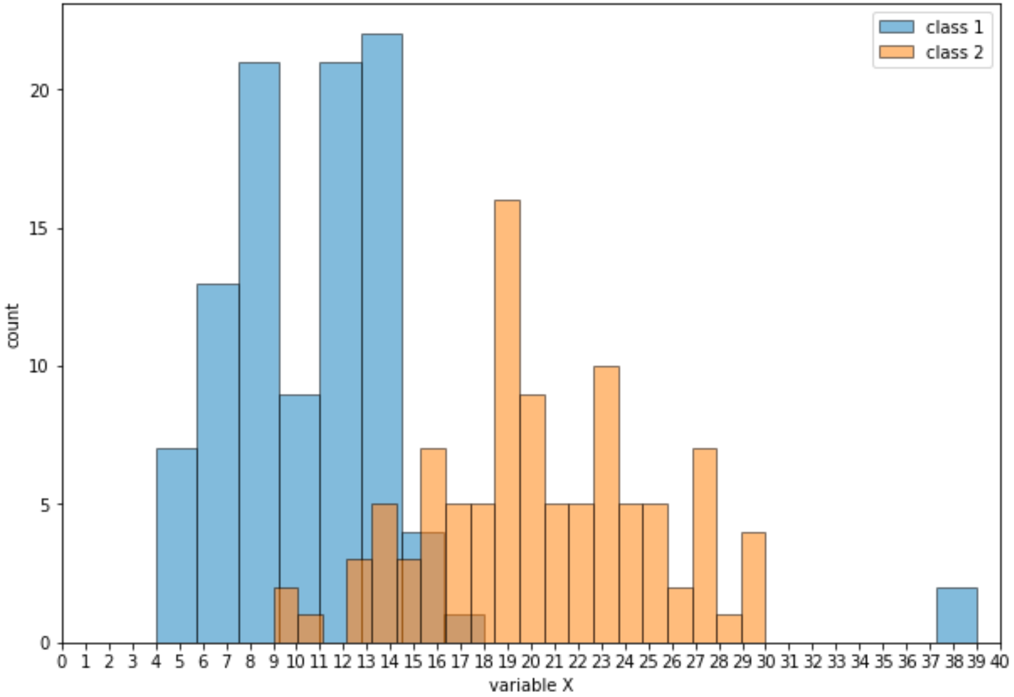
Motivation

- Given the following Dataset – 100 points in each class
- We want to apply a Bayesian classifier, using counting?
- Small recap/clarification:
  - Bayes classifier uses MAP:

$$argmax\left(P(x|A) * P(A), P(x|B) * P(B)\right)$$

  - Assuming the same class priors we get:

$$argmax\left(P(x|A), P(x|B)\right)$$

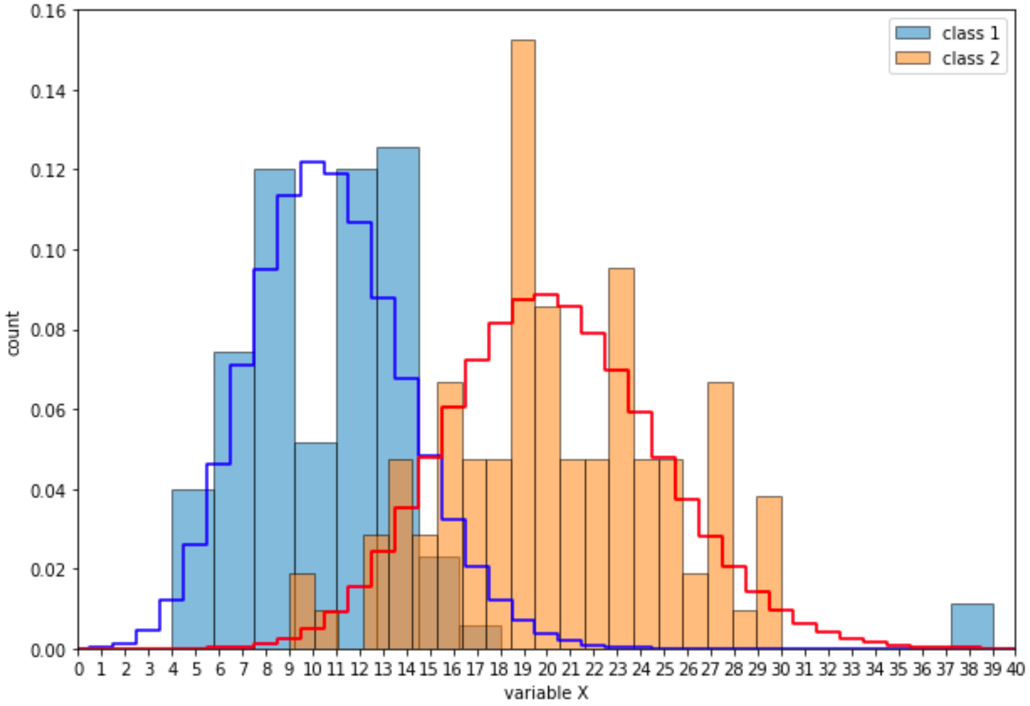


Motivation: Bayes classification

- Select distribution
  - Visually (in the future more principal approaches will b discussed)
- Select distribution parameters
  - Poisson – λ
  - Which λ is better for Class A? 8, 10.5 or 11.7?
  - How we will decide?
- MLE – a principled approach to finding the parameters
  - In this case the MLE finds the following λ:
    - Class A – 10.74
    - Class B – 20.31
  - What will be the prediction of new data point where x=38 according to MLE?

$$P(x|A) = Poisson(10.74).pmf(38) = 6.24 * 10^{-11}$$
$$P(x|B) = Poisson(20.31).pmf(38) = 1.42 * 10^{-4}$$

  - The prediction will be class B



MLE

- A straightforward approach to parameter estimation.
- Directly works in simple cases.
- Forms the basis for most parameter estimation approaches

Steps of MLE

- Given
  - A set of observed values D = $\{x_1, \ldots, x_n\}$
  - A model and a vector of parameters for this model, θ

- We define
  - The likelihood of the model given the data, L(θ|D), which we define to be P(D|θ)
  - Log-likelihood of the model given the data is often more useful and we write:
    L(θ)=log P(D|θ)

- In MLE we seek

$$\theta^* = \underset{\theta \in \Theta}{\arg max} \; L(\theta|D)$$

MLE for independent instances

- We often assume that the data instances are a result of independent identically distributed (i.i.d.) random variables.
  This is the same as assuming independent repeats of the same generation mechanism.

$$\theta^* = \underset{\theta \in \Theta}{\arg max} \; L(\theta|D) = \underset{\theta \in \Theta}{\arg max} \; \log \prod_i p(x_i|\theta)$$

$$= \underset{\theta \in \Theta}{\arg max} \; \sum_i \log p(x_i|\theta)$$

- Solving the optimization problem varies in its complexity, depending on the form of $p(x|\theta)$ – the (common) pdf of the underlying variables

An (very) easy case

- Assuming
  - A coin has a probability *p* of being heads, *1-p* of being tails.
  - Observation:
    We toss the coin until the first head.
    That is we are observing $K \sim Geo(p)$

- What is the value of p based on MLE, given that the first success was in the k-th toss?

Coin tossing

$$\mathrm{L}(\Theta) = \log \mathrm{P}(\mathrm{D}|\Theta) = \log\left((1-p)^{k-1}p^{1}\right)$$

$$= (k-1)\log(1-p) + \log p$$

$$\frac{d\mathrm{L}(\Theta)}{dp} = -\frac{k-1}{1-p} + \frac{1}{p}$$

$$\frac{d\mathrm{L}(\Theta)}{dp} = \frac{-kp+p+1-p}{(1-p)p} = 0$$

$$1 = \mathrm{kp}$$

$$\Rightarrow \quad \mathrm{p} = \frac{1}{\mathrm{k}}$$

Coin tossing – N independent experiments

$$\mathrm{L}(\Theta) = \log \mathrm{P}(\mathrm{D}|\Theta) = \sum_{i=1}^{N} \log\left((1-p)^{k_i-1}p^1\right)$$

$$= \sum_{i=1}^{N}\left((k_i - 1)\log(1-p) + \log p\right)$$

$$= \log(1-p)\left(\left(\sum_{i=1}^{N} k_i\right) - \mathrm{N}\right) + \mathrm{N}\log p$$

$$\frac{d\mathrm{L}(\Theta)}{dp} = -\frac{\sum_{i=1}^{N} k_i - N}{1-p} + \frac{N}{p}$$

$$\frac{d\mathrm{L}(\Theta)}{dp} = \frac{-p\sum_{i=1}^{N} k_i + Np + N - Np}{(1-p)p} = 0$$

$$N = p\sum_{i=1}^{N} k_i \rightarrow p = \frac{N}{\sum_{i=1}^{N} k_i} = \frac{1}{\frac{\sum_{i=1}^{N} k_i}{N}} = \frac{1}{\bar{k}}$$

Next easy case

- Assuming
  - A coin has a probability *p* of being heads, *1-p* of being tails.
  - Observation:
    We toss the coin N times, observing a set of Hs and Ts.

- What is the value of p based on MLE, given the observation?

Coin tossing … cont

What is missing in the equation??

$$\mathrm{L}(\Theta) = \log \mathrm{P}(\mathrm{D}|\Theta) = \log(p^m(1-p)^{N-m})$$
$$= m\log p + (N-m)\log(1-p)$$

$$\frac{d\mathrm{L}(\Theta)}{dp} = \frac{m}{p} - \frac{N-m}{1-p} = 0$$

$$\frac{d\mathrm{L}(\Theta)}{dp} = \frac{m - pm - pN + pm}{p(1-p)} = 0$$

$$\Rightarrow \quad p = \frac{m}{N}$$

Another simple example: Poisson

$$\mathrm{L}(\theta) = \log \mathrm{P}(\mathrm{D}|\theta)$$

$$= \log \prod_{j=1}^{n} e^{-\lambda} \frac{\lambda^{x_j}}{x_j!}$$

$$= \sum_{j=1}^{n} \log e^{-\lambda} \frac{\lambda^{x_j}}{x_j!}$$

$$= \sum_{j=1}^{n} \left(\log e^{-\lambda} + \log \lambda^{x_j} - \log x_j! \right)$$

$$= \sum_{j=1}^{n} \left(-\lambda + x_j \log \lambda - \log x_j! \right) = -n\lambda + \log \lambda \sum_{j=1}^{n} x_j - \sum_{j=1}^{n} \log x_j!$$

$$\frac{d\mathrm{L}(\theta)}{d\lambda} = -n + \frac{1}{\lambda} \sum_{j=1}^{n} x_j = 0$$

⇒ $$\frac{1}{n} \sum_{j=1}^{n} x_j = \lambda$$

Expectation Maximization (EM) Algorithm

- Iterative method for parameter estimation where layers of data are missing from the observation
- Dempster, Laird, Rubin,
  J of the Royal Stat Soc, 1977
- Many variations followed. Research into methodology and applications is very active
- Has two steps:
  Expectation (E) and Maximization (M)
- Applicable to a wide range of machine learning and inference tasks

The basic EM setting - summarized

C = (X, Z)

+ C: complete data (“augmented data”)
+ X: observed data (“incomplete” data)
+ Z: hidden data (“missing” data)

Randomly selecting one of two coins

There are two coins, with p_A and p_B .
One of the coins is selected, with w_A and w_B probabilities.
Then it is tossed 10 times.
We observe the results of many repeats of this exercise.
If we know which coin is tossed in each set then we can do MLE and get both the ps and the ws.
But we don't …
Lets see what EM can do here.

```
HHHHTHHHHH

THHHHHHHTH

HHHHHHHTHH

HHTHTTHHTT

HHTHHHHHTH

HTTHTHHHTT

HTTHTHHHHT

HTHHTHHHHT
```

The EM algorithm

- Consider a set of starting parameters
- Use these to “estimate” the values of the missing data, per observed data point
- Use the “complete” data to update all parameters (of both Z and $X|Z$)
- Repeat until convergence

EM: uncovering the coins …

🟠	HHHHTHHHHH	0.8	0.2
🟠	THHHHHHHTH		
🟠	HHHHHHHTHH		
🟢	HHTHTTHHTT		
🟠	HHTHHHHHTH		
🟢	HTTHTHHHTT		
🟢	HTTHTHHHHT		
🟠	HTHHTHHHHT		

Coin A responsibilities

Coin B responsibilities

1

Init p_A = 0.6
p_B = 0.5
ws are 0.5

2

Compute responsibilities

$$P_A(x_1) = w_A \binom{10}{9} 0.6^9 0.4^1 = 0.04$$

$$P_B(x_1) = w_B \binom{10}{9} 0.5^9 0.5^1 = 0.01$$

$$r(x_1, A) = \frac{0.04}{0.05} = 0.8$$

$$r(x_1, B) = \frac{0.01}{0.05} = 0.2$$

Note: use aposteriori estimates

EM: uncovering the coins …

🟠	HHHHTHHHHH	0.8	0.2
🟠	THHHHHHHTH	0.76	0.24
🟠	HHHHHHHTHH		
🟢	HHTHTTHHTT		
🟠	HHTHHHHHTH		
🟢	HTTHTHHHTT		
🟢	HTTHTHHHHT		
🟠	HTHHTHHHHT		

Coin A responsibilities

Coin B responsibilities

1

Init $p_A = 0.6$
$p_B = 0.5$
ws are 0.5

$$P_A(x_2) = w_A \binom{10}{8} 0.6^8 0.4^2 = 0.12$$

$$P_B(x_2) = w_B \binom{10}{8} 0.5^8 0.5^2 = 0.044$$

$$r(x_2, A) = \frac{0.12}{0.164} = 0.76$$

$$r(x_2, B) = \frac{0.044}{0.164} = 0.24$$

EM: uncovering the coins …

🟠	HHHHTHHHHH	0.8	0.2
🟠	THHHHHHHTH	0.76	0.24
🟠	HHHHHHHTHH	0.8	0.2
🟢	HHTHTTHHTT		
🟠	HHTHHHHHTH	0.76	0.24
🟢	HTTHTHHHTT		
🟢	HTTHTHHHHT		
🟠	HTHHTHHHHT		

Coin A responsibilities

Coin B responsibilities

1

Init $p_A = 0.6$
$p_B = 0.5$
ws are 0.5

EM: uncovering the coins …

🟠	HHHHTHHHHH	0.8	0.2
🟠	THHHHHHHTH	0.76	0.24
🟠	HHHHHHHTHH	0.8	0.2
🟢	HHTHTTHHTT	0.45	0.55
🟠	HHTHHHHHTH	0.76	0.24
🟢	HTTHTHHHTT	0.45	0.55
🟢	HTTHTHHHHT		
🟠	HTHHTHHHHT		

Coin A responsibilities

Coin B responsibilities

1

Init $p_A = 0.6$
$p_B = 0.5$
ws are 0.5

$$P_A(x_4) = w_A \binom{10}{5} 0.6^5 0.4^5$$

$$P_B(x_4) = w_B \binom{10}{5} 0.5^5 0.5^5$$

$$r(x_4, A) = 0.45$$

$$r(x_4, B) = 0.55$$

EM: uncovering the coins …

🟠	HHHHTHHHHH	0.8	0.2
🟠	THHHHHHHTH	0.76	0.24
🟠	HHHHHHHTHH	0.8	0.2
🟢	HHTHTTHHTT	0.45	0.55
🟠	HHTHHHHHTH	0.76	0.24
🟢	HTTHTHHHTT	0.45	0.55
🟢	HTTHTHHHHT	0.55	0.45
🟠	HTHHTHHHHT	0.64	0.36

1

Init $p_A = 0.6$
$p_B = 0.5$
ws are 0.5

2

Compute responsibilities

Coin A responsibilities

Coin B responsibilities

3

Compute new assignments:

$$New\ w_A = \frac{1}{N}\sum_{i=1}^{N} r(x_i, A)$$

$$New\ w_B = \frac{1}{N}\sum_{i=1}^{N} r(x_i, B)$$

$$New\ w_A = \frac{1}{8}\sum_{i=1}^{8} r(x_i, A) = \frac{5.2}{8} = 0.65$$

$$New\ w_B = \frac{1}{8}\sum_{i=1}^{8} r(x_i, B) = \frac{2.8}{8} = 0.35$$

EM: uncovering the coins …

3+4 Compute MLEs for the model parameters:

$$p_A = \frac{1}{(New\ w_A)N}\sum_{i=1}^{N} r(x_i, A)v(i)$$

$$p_B = \frac{1}{(New\ w_B)N}\sum_{i=1}^{N} r(x_i, B)v(i)$$

$$p_A = \frac{1}{5.2}\sum_{i=1}^{8} r(x_i, A)v(i) = 0.745$$

$$p_B = \frac{1}{2.8}\sum_{i=1}^{8} r(x_i, B)v(i) = 0.649$$

		$r(x_i,A)$	$r(x_i,B)$	Value observed at i: $v(i)$
🟠	HHHHTHHHHH	0.8	0.2	0.9
🟠	THHHHHHHTH	0.76	0.24	0.8
🟠	HHHHHHHTHH	0.8	0.2	0.9
🟢	HHTHTTHHTT	0.45	0.55	0.5
🟠	HHTHHHHHTH	0.76	0.24	0.8
🟢	HTTHTHHHTT	0.45	0.55	0.5
🟢	HTTHTHHHHT	0.55	0.45	0.6
🟠	HTHHTHHHHT	0.64	0.36	0.7

1 Init $p_A = 0.6$
$p_B = 0.5$
ws are 0.5

2 Compute responsibilities

3+4

EM: uncovering the coins …

🟠	HHHHTHHHHH	0.8	0.2	0.9
🟠	THHHHHHHTH	0.76	0.24	0.8
🟠	HHHHHHHTHH	0.8	0.2	0.9
🟢	HHTHTTHHTT	0.45	0.55	0.5
🟠	HHTHHHHHTH	0.76	0.24	0.8
🟢	HTTHTHHHTT	0.45	0.55	0.5
🟢	HTTHTHHHHT	0.55	0.45	0.6
🟠	HTHHTHHHHT	0.64	0.36	0.7

3+4

$$w_A = 0.65$$
$$w_B = 0.35$$

$$p_A = 0.745$$
$$p_B = 0.649$$

1

new p_A
new p_B
new ws

2

Compute responsibilities

$r(x_i,A)$

$r(x_i,B)$

Value observed at i: $v(i)$

3+4

Reiterate with the new model

The EM algorithm for two coins

- Consider a set of starting parameters, including the parameters of Z
- Use these to “estimate” the values of the missing data, per observed data point.
  - + Compute responsibilities using MAP
- Use the “complete” data to update all parameters (of both Z and X|Z)

$$New\ w_A = \frac{1}{N}\sum_{i=1}^{N} r(x_i, A) \qquad p_A = \frac{1}{(New\ w_A)N}\sum_{i=1}^{N} r(x_i, A)v(i)$$

$$New\ w_B = \frac{1}{N}\sum_{i=1}^{N} r(x_i, B) \qquad p_B = \frac{1}{(New\ w_B)N}\sum_{i=1}^{N} r(x_i, B)v(i)$$

- Repeat until convergence

EM for GMMs

- Step 1: Expectation (E-step)

  Evaluate the “responsibilities” of each data point to each Gaussian using the current parameters

- Step 2: Maximization (M-step)

  Re-estimate parameters (ws, µs and σs) using the existing “responsibilities”.
  That is – every data point, x, contributes to each Gaussian component, G_i, in proportion to its responsibility:
  $r(x,G_i)$.

Responsibilities for Gaussian mixtures

$$r(x,k) = \frac{w_k N(x|\mu_k\,,\sigma_k)}{\sum_{j=1}^{K} w_j N(x|\mu_j\,,\sigma_j)}$$

Parameter updates for Gaussian mixtures

$$New\ w_j = \frac{1}{N}\sum_{i=1}^{N} r(x_i, j)$$

Parameter updates for Gaussian mixtures

$$New\ \mu_k = \frac{1}{(New\ w_k)N} \sum_{i=1}^{N} r(x_i, k)x_i$$

$$(New\ \sigma_k)^2 = \frac{1}{(New\ w_k)N} \sum_{i=1}^{N} r(x_i, k)(x_i - New\ \mu_k)^2$$

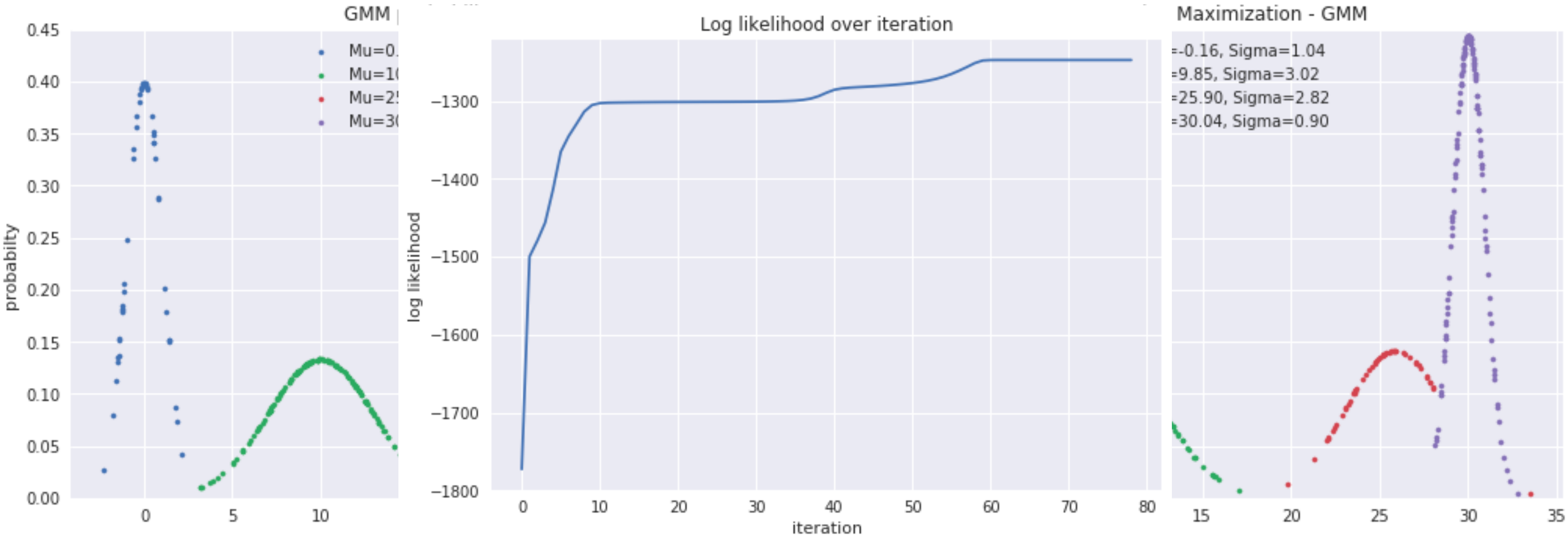
Log Likelihood for Gaussian mixtures

Likelihood of single data point: $L^t(x|\Theta) = \sum_k w_k N(x|\mu_k^t, \sigma_k^t)$

Likelihood of data: $L^t(X|\Theta) = \prod_{x \in X} L^t(x|\Theta)$

Log likelihood of data: $\log\left(L^t(X|\Theta)\right) = \sum_{x \in X} \log\left(L^t(x|\Theta)\right)$

Running example

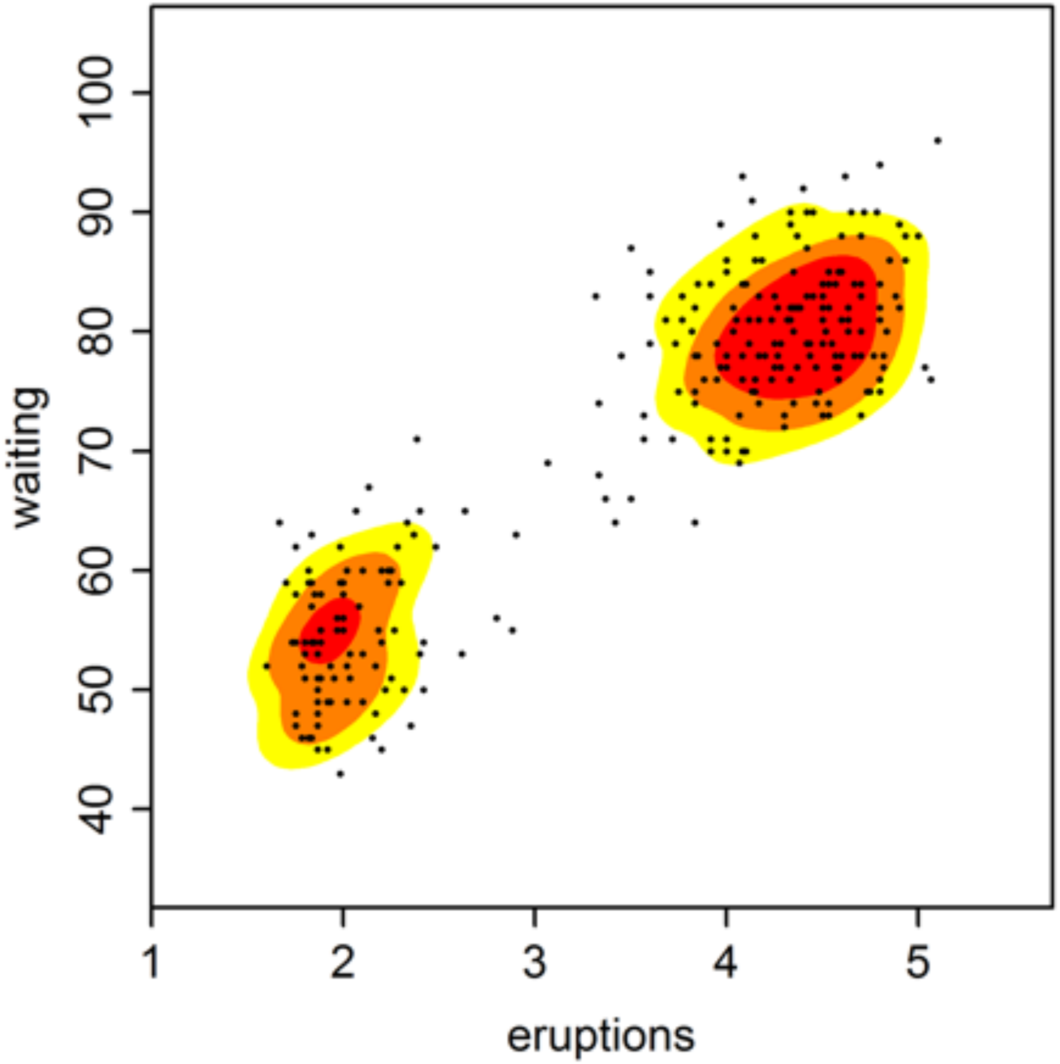


EM algorithm

- A general algorithm/framework for inference where observations are dependent on a hidden intermediate, Z
- Requires "specialization" to any given task or configuration.

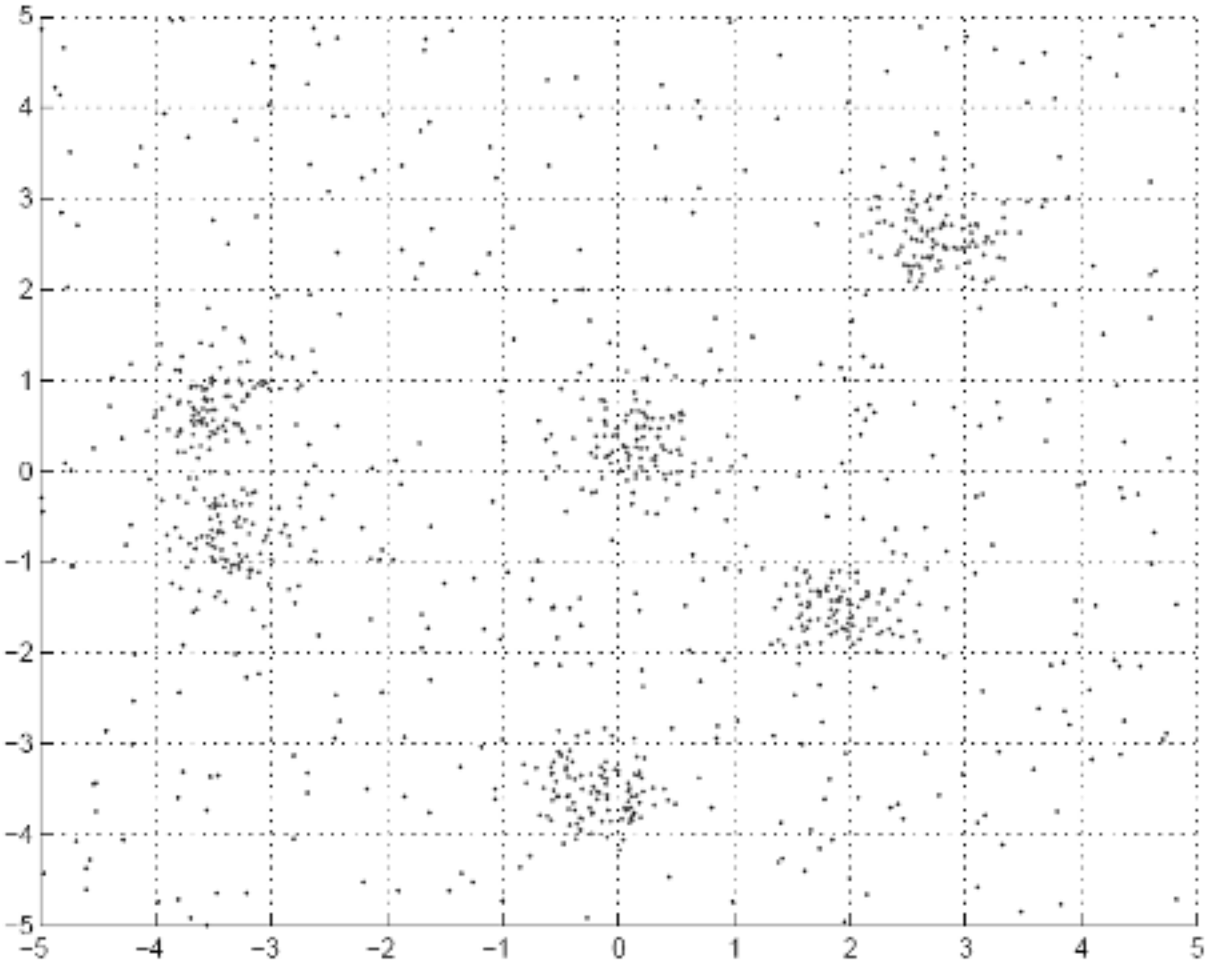
- Consider a set of starting parameters, including the parameters of Z
- Use these to compute responsibilities
- MLE:
  Use the "complete" data to update all parameters (of both Z and $X|Z$)
- Repeat until convergence

Old Faithful Revisited

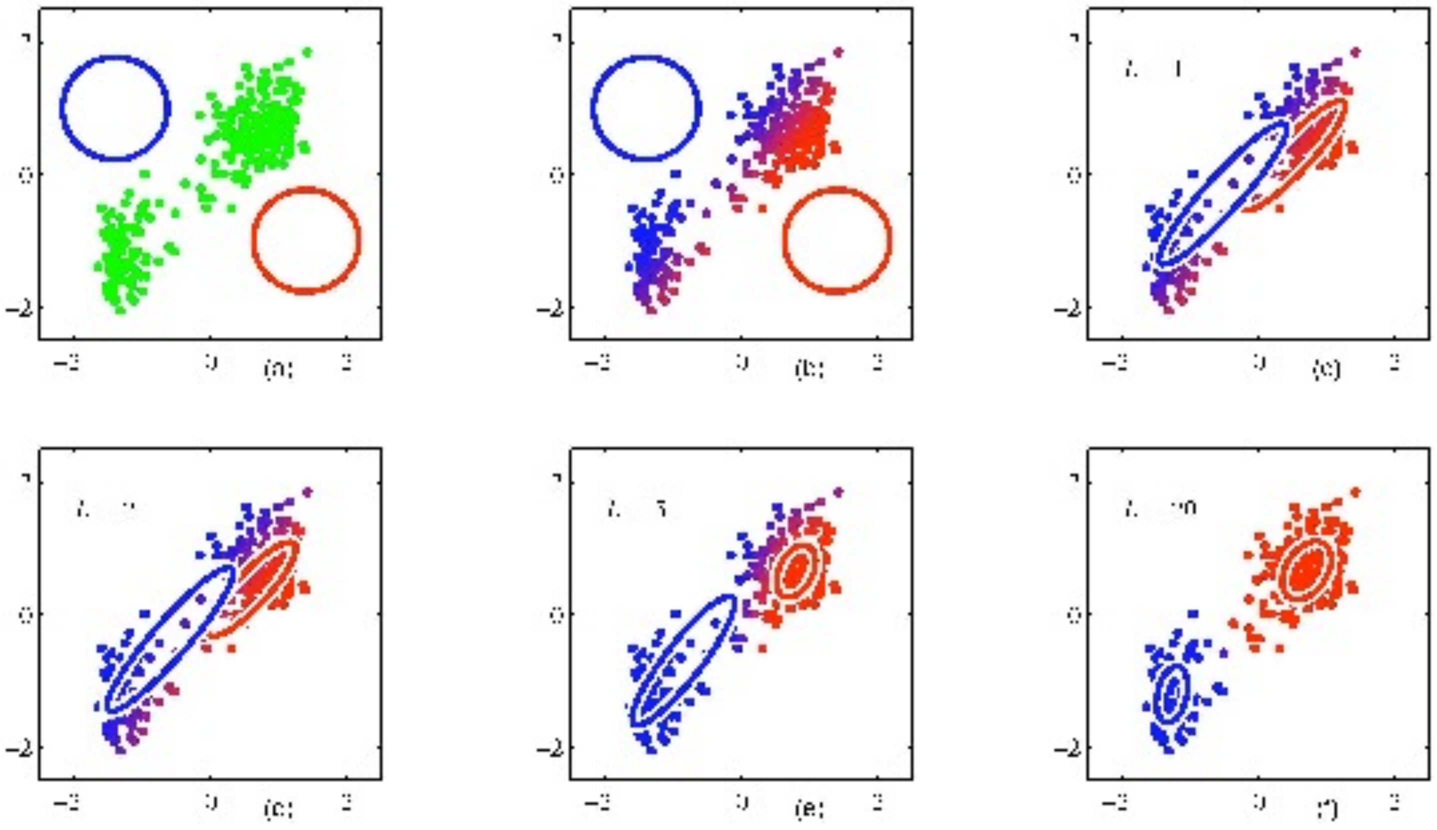


- What can we observe?
- What would we like to know in order to understand the entire system (and make predictions)?
- Gaussian Mixtures

Several underlying distributions



Visual example of EM for 2D GMMs



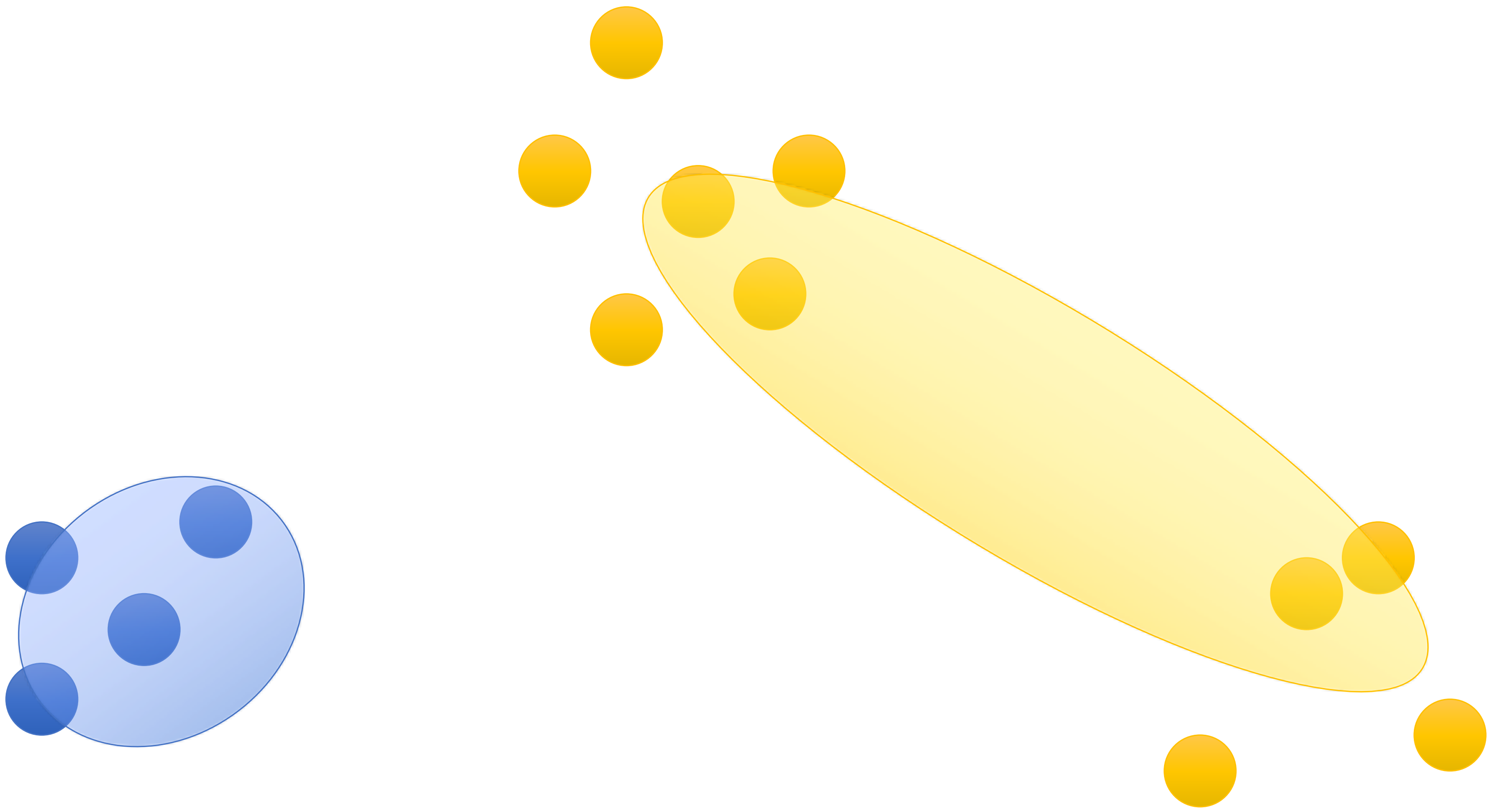
Multidimensional GMM EM

How many parameters do we need to infer?

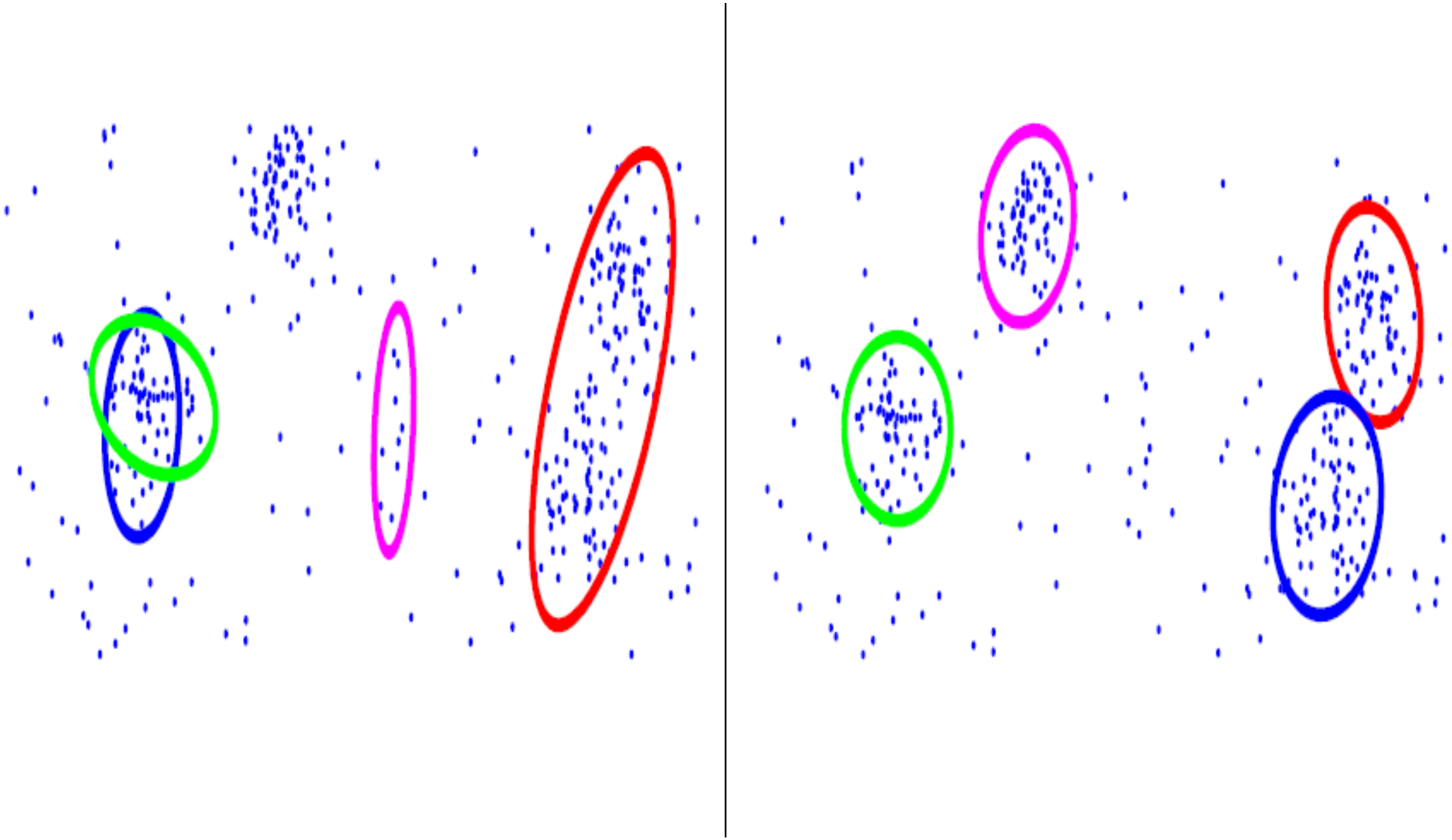
- k weights – w (actually $k - 1$ …)
- d means per Gaussian ($\vec{\mu}$)
- $\binom{d+1}{2}$ = d variances + $\binom{d}{2}$ covariances,

  per Gaussian

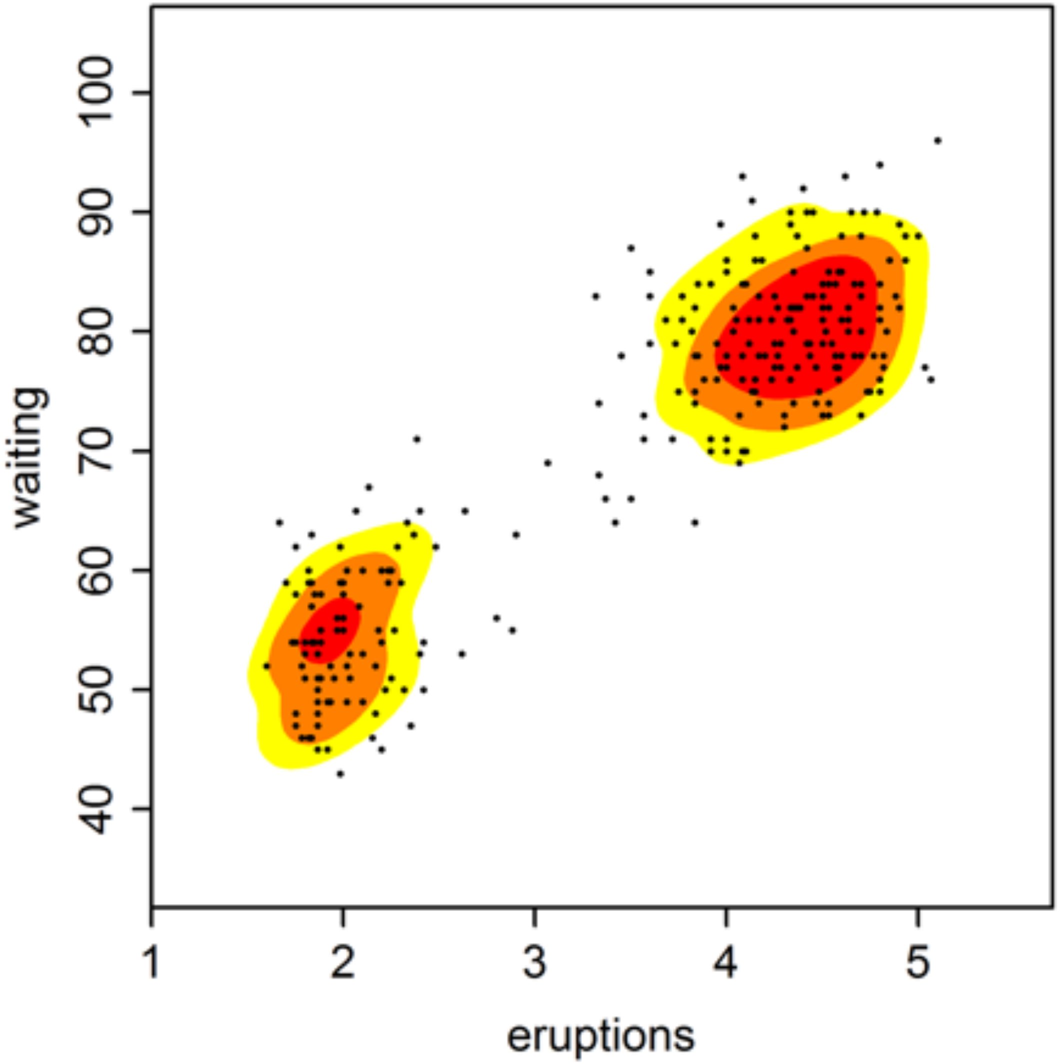
Incorrect Number of Gaussians



Local minima

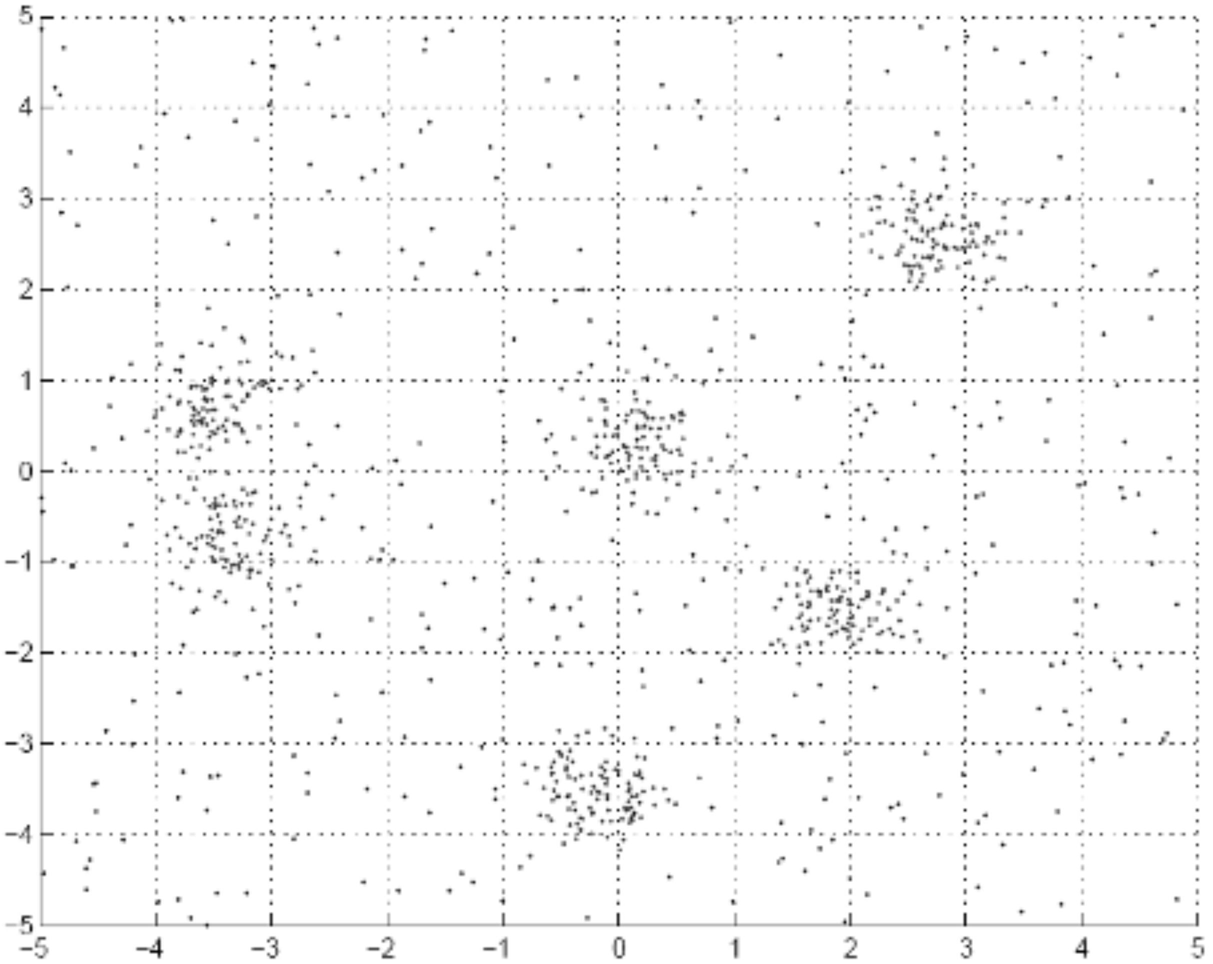


Old Faithful Revisited

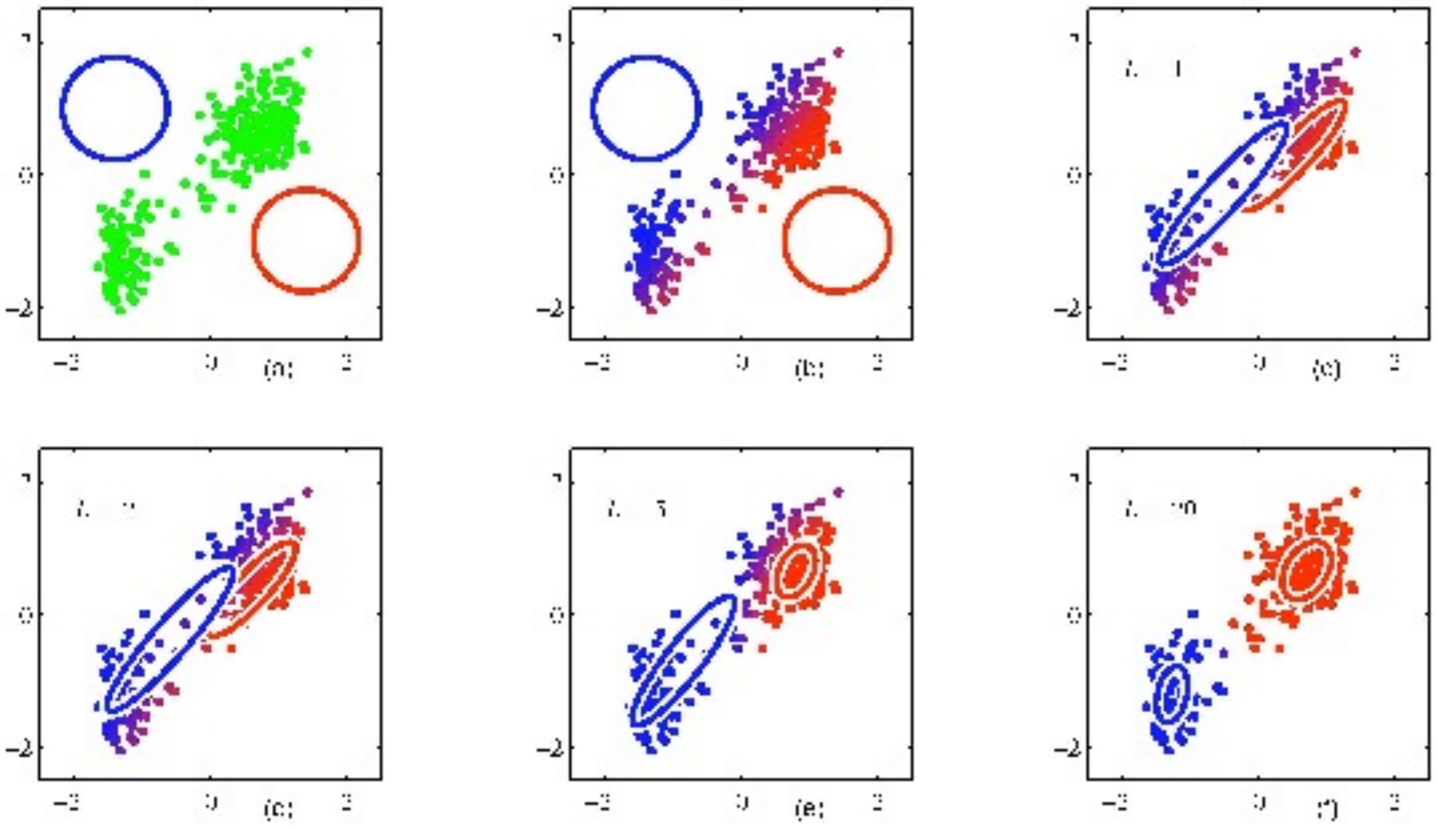


- What can we observe?
- What would we like to know in order to understand the entire system (and make predictions)?
- Gaussian Mixtures

Several underlying distributions



Visual example of EM for 2D GMMs



Summary

- MLE – an approach to estimating parameters of a distribution from observed data
- Uses an optimization of likelihood
- EM – an iterative approach to complex MLE, addressing the estimation of hidden parameters