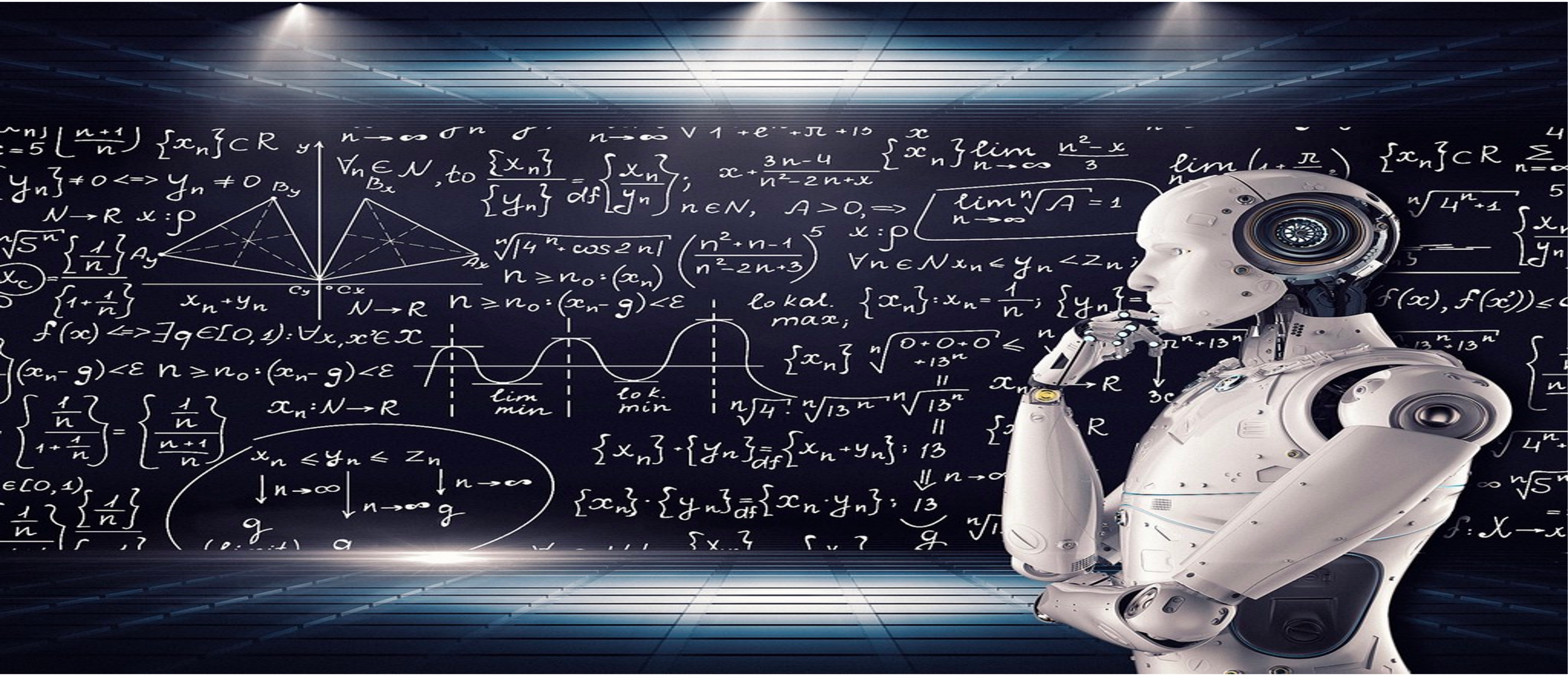
Gaussian Mixture Models



Refresher: Expectation of GMMs

- Calculate $E\big(g(X)\big)$ given that the probability of a value $\text{g}(x)$ is given by $f(x) = \sum_i w_i f_i(x)$

$$E\big(g(X)\big) = \int_{-\infty}^{\infty} g(x)f(x)dx = \int_{-\infty}^{\infty} g(x)\left(\sum_i w_i f_i(x)\right)dx$$

$$= \sum_i w_i \left(\int_{-\infty}^{\infty} g(x)f_i(x)dx\right) = \sum_i w_i E\big(g(X_i)\big)$$

$$E(X) = \sum_i w_i E(X_i)$$

Refresher: Variance of GMMs

$$Var(X) = E\left((X-\mu)^2\right) = \sum_i w_i E\left((X_i-\mu)^2\right)$$

$$= \sum_i w_i E\left(\left((X_i-\mu_i)+(\mu_i-\mu)\right)^2\right)$$

$$= \sum_i w_i E\left((X_i-\mu_i)^2 + 2(X_i-\mu_i)(\mu_i-\mu) + (\mu_i-\mu)^2\right)$$

$$= \sum_i w_i E\left((X_i-\mu_i)^2\right) + \sum_i 2w_i(\mu_i-\mu)E(X_i-\mu_i) + \sum_i w_i(\mu_i-\mu)^2$$

$$= \sum_i w_i E\left((X_i-\mu_i)^2\right) + \sum_i w_i(\mu_i-\mu)^2$$

Exercise 1 – Gaussian Mixtures

- Fred, Mel and Sid are repair technicians who work for Randobezeq
  - Fast Fred takes time which is distributed as N(40, 16) to repair a telephone line failure
  - Medium Mel takes time which is distributed as N(45, 36) for the same task.
  - Slow Sid takes time which is distributed as N(50, 64) for the same task.

- Fred is due to arrive to repair your phone at 11AM tomorrow. How confident can you be that you will be done by 11:45?

- $P(X \leq 45) = F_{fred}(45) \cong 0.89$

Exercise 1 – Gaussian Mixtures

- When a customer in North Randomistan orders a repair, there is a 40% chance Fred will do the work and 30% each that Mel or Sid will do the work.

- What is the distribution of the duration of repair in North Randomistan?

$$f(x) = 0.4 \cdot f_{fred}(x) + 0.3 \cdot f_{mel}(x) + 0.3 \cdot f_{sid}(x)$$

- Let Φ denote the CDF of a standard normal random variable. Use Φ to express the CDF of the duration of a repair in North Randomistan.
  - Fast Fred: N(40, 16)
  - Medium Mel: N(45, 36)
  - Slow Sid: N(50, 64)

$$F(x) = 0.4 \cdot \Phi\left(\frac{x-40}{4}\right) + 0.3 \cdot \Phi\left(\frac{x-45}{6}\right) + 0.3 \cdot \Phi\left(\frac{x-50}{8}\right)$$

Exercise 1 – Gaussian Mixtures

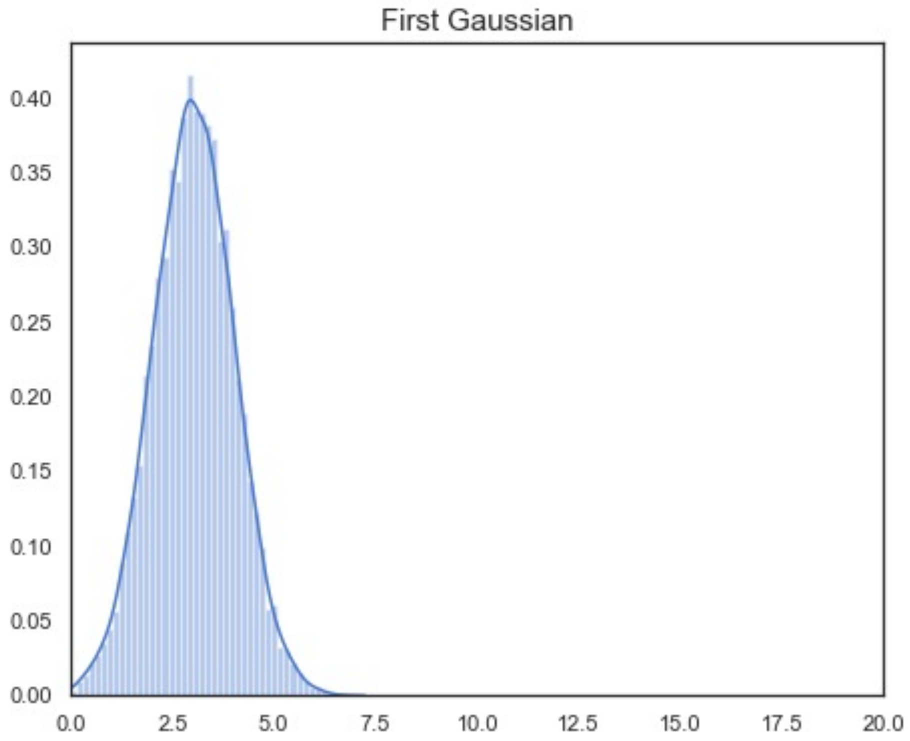
- If the repair starts at 11AM, what is the earliest time for which the customer can assume, at a 95% certainty, that the repair will be already done?

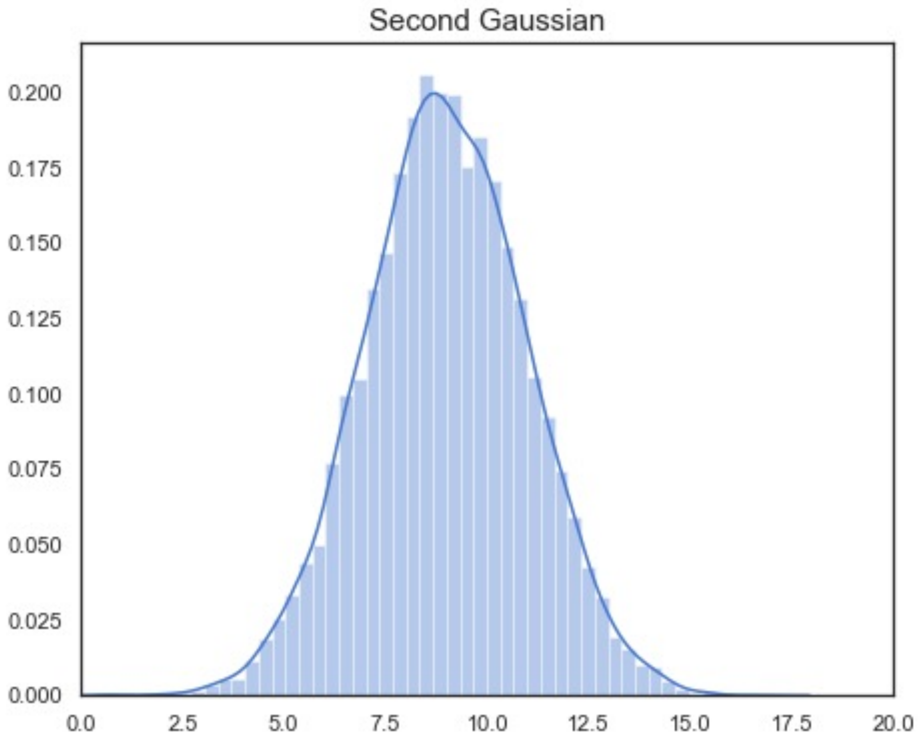
- If is possible to solve $0.95 = 0.4 \cdot \Phi\left(\frac{x-40}{4}\right) + 0.3 \cdot \Phi\left(\frac{x-45}{6}\right) + 0.3 \cdot \Phi\left(\frac{x-50}{8}\right)$ for x but it is significantly easier to use Python.

```
prob = 0.4 * fred.cdf(time) + 0.3 * mel.cdf(time) + 0.3 * sid.cdf(time)
while prob < 0.95:
    time +=1
    prob = 0.4 * fred.cdf(time) + 0.3 * mel.cdf(time) + 0.3 * sid.cdf(time)
```

Gaussian Mixtures vs Sum of Gaussians

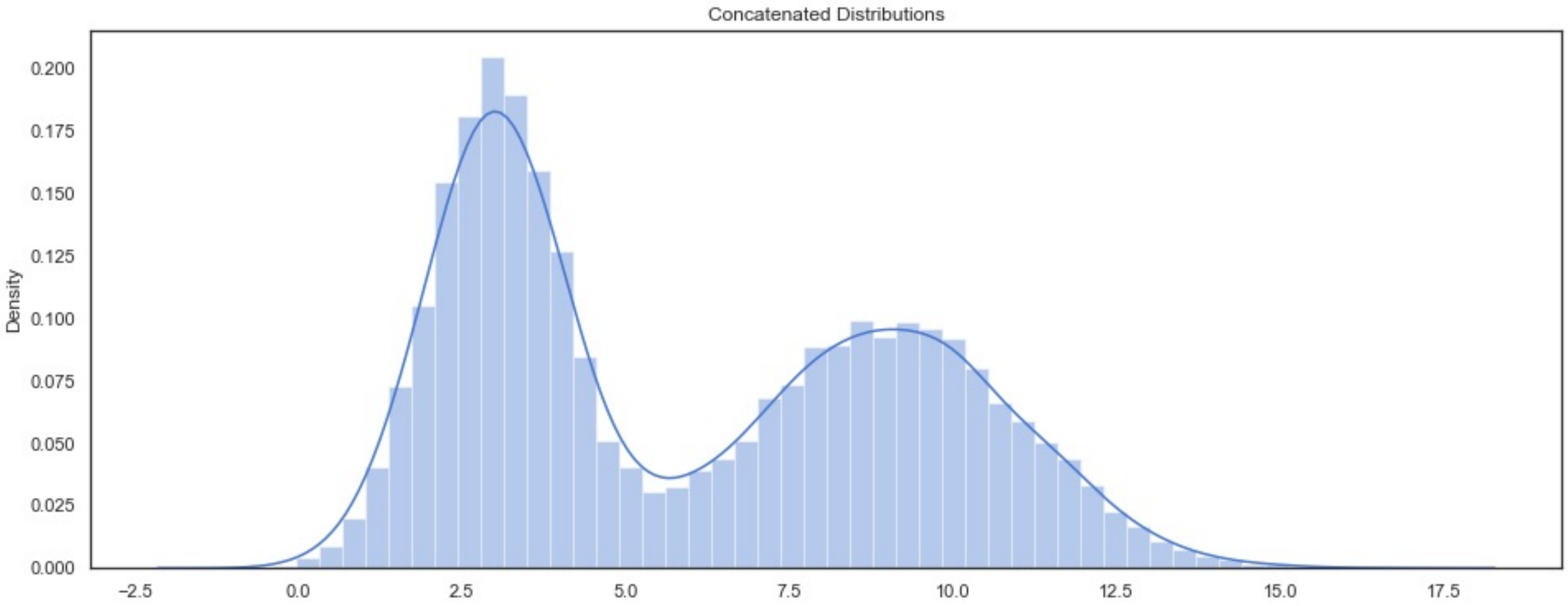
- Consider two distributions (arrays) sampled from two different Gaussians





Gaussian Mixtures

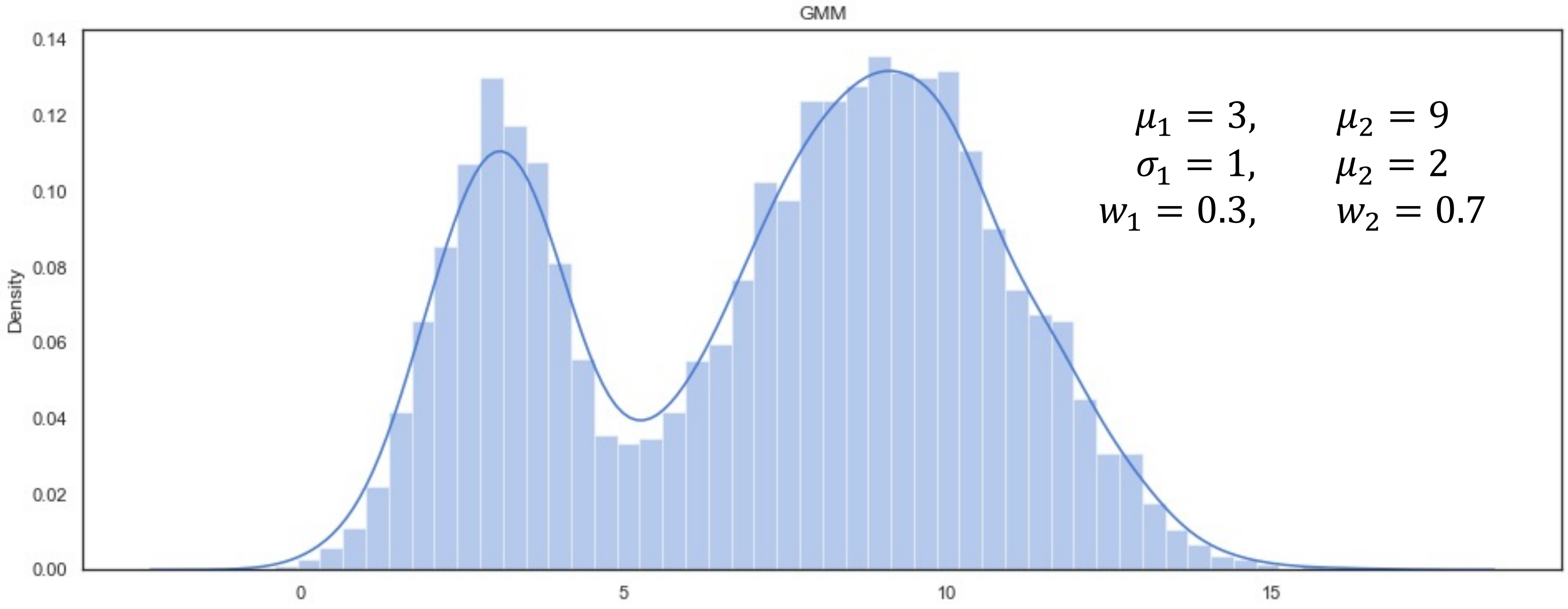
- What will happen if we concatenate the arrays?



- Is this a GMM?

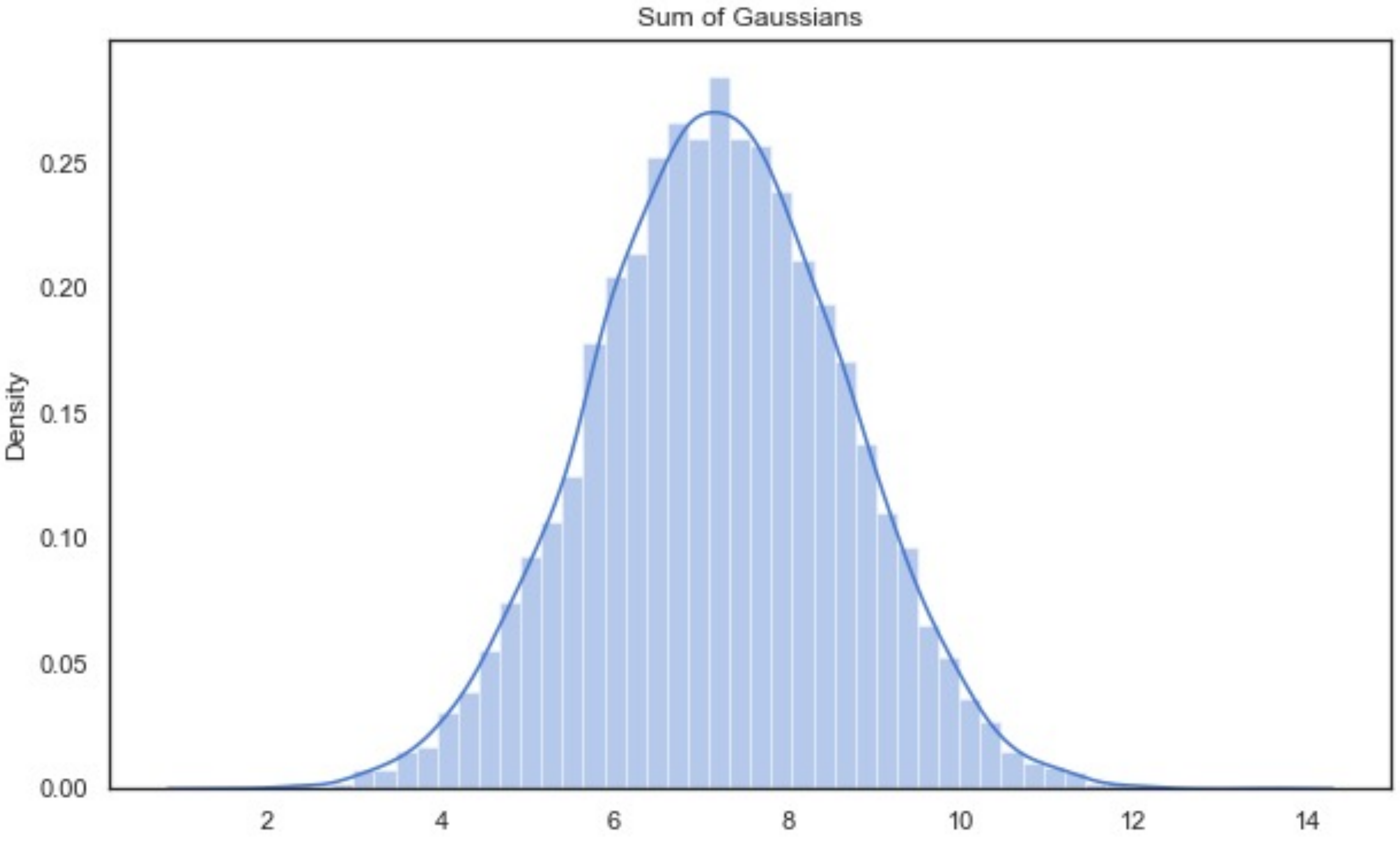
Gaussian Mixtures

- Every Gaussian in a GMM has 3 parameters: μ_i, σ_i and w_i.
- When we generate data from a GMM, we first choose the Gaussian from which we sample. Each Gaussian has probability w_i for getting picked.
- Next, we sample that Gaussian and repeat this process as much as we need.



Sum of Gaussians

- What will happen if take the weighted sum of the distributions?



Parameter Estimation

- How can we estimate the parameters (μ_i, σ_i^2) of each distribution?

- Sum of Gaussians:
  - $\mu_{sum} = w_1 \cdot \mu_1 + w_2 \cdot \mu_2$ - Theoretical formula
  - $\mu_{sum} = w_1 \cdot mean(arr_1) + w_2 \cdot mean(arr_2)$ - Formula
  - $\mu_{sum} = mean(arr_{sum})$ - Data

  - $\sigma_{sum}^2 = w_1^2 \cdot \sigma_1^2 + w_2^2 \cdot \sigma_2^2$ - Theoretical formula
  - $\sigma_{sum}^2 = w_1^2 \cdot variance(arr_1) + w_2^2 \cdot variance(arr_2)$ - Formula
  - $\sigma_{sum}^2 = variance(arr_{sum})$ - Data

Parameter Estimation - GMM

- $\mu_{GMM} = w_1 \cdot \mu_1 + w_2 \cdot \mu_2$ - Theoretical formula
- $\mu_{GMM} = w_1 \cdot mean(arr_1) + w_2 \cdot mean(arr_2)$ - Formula
- $\mu_{GMM} = mean(arr_{concat})$ - Data

- $\sigma_{GMM}^2 = \sum_i w_i E\left((X_i - \mu_i)^2\right) + \sum_i w_i (\mu_i - \mu)^2$ - Theoretical formula
- $\sigma_{GMM}^2 = w_1 \cdot variance(arr_1) + w_2 \cdot variance(arr_2) + w_1(mean(arr_1) - mean(arr_{concat}))^2 + w_2(mean(arr_2) - mean(arr_{concat}))^2$ - Formula
- $\sigma_{GMM}^2 = variance(arr_{concat})$ - Data