Advanced Machine Learning
Generative Model

Yu Wang
Assistant Professor
Department of Computer Science
University of Oregon

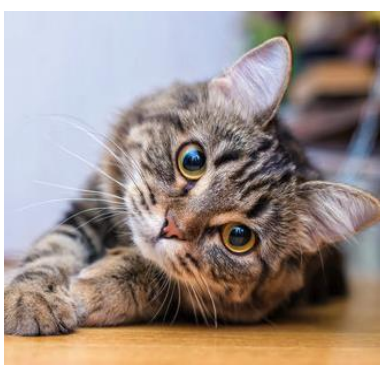
Data Distribution

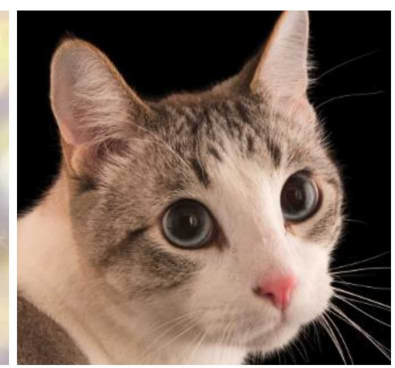
Dog





Cat



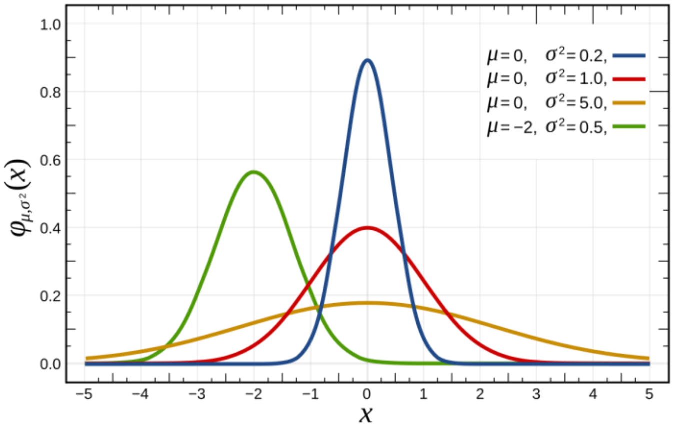

Data Distribution

Dog – P(Dog)

Cat – P(Cat)

1. There is no concrete image/shape of the dog, everyone can come up with one of your own choice

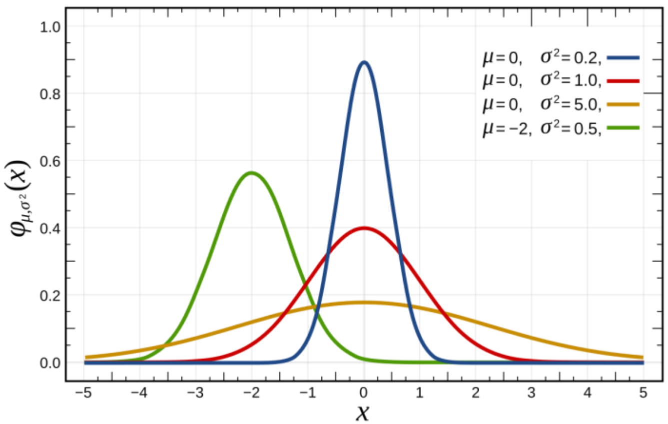
2. But somehow dog and cat image distributions are different



When you draw an image, you are actually sampling from a probability distribution!

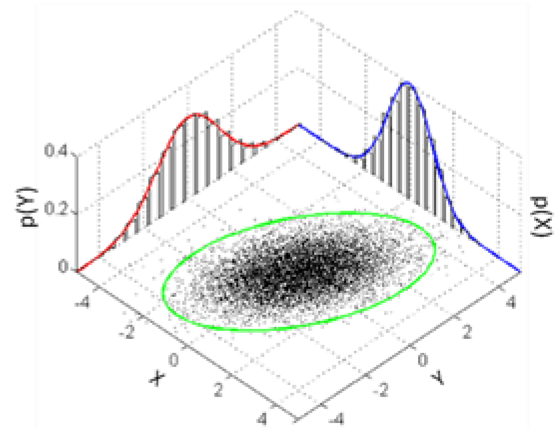
Data Distribution

1D Gaussian Distribution



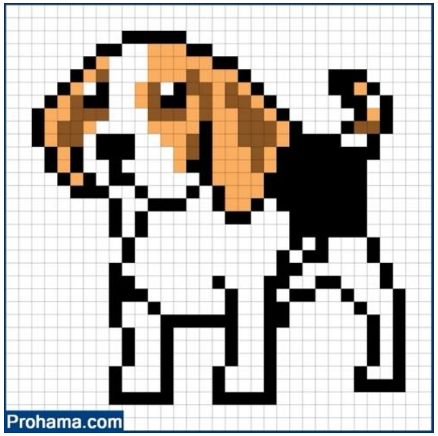
$\mathbb{R}$

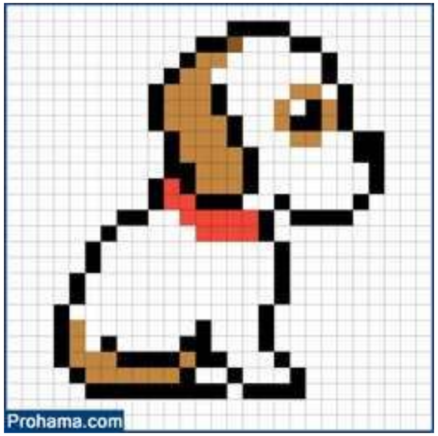
2D Gaussian Distribution



$\mathbb{R}^2$

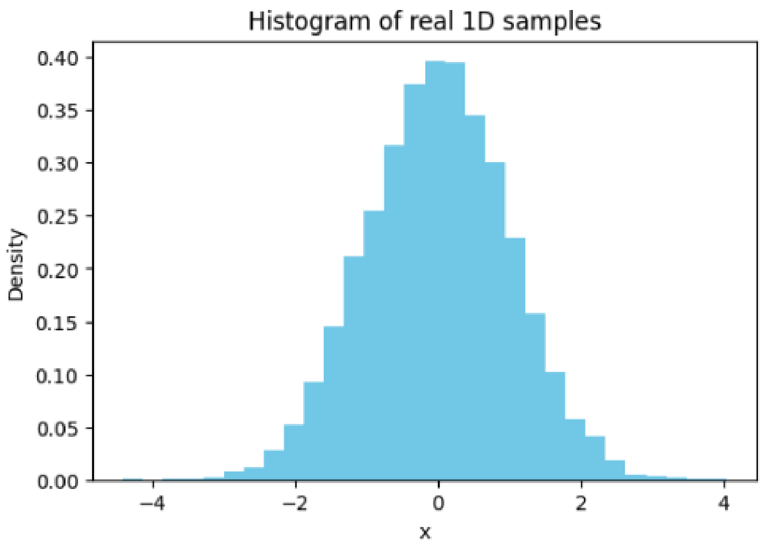
$\mathbb{R}^{256\times256}$

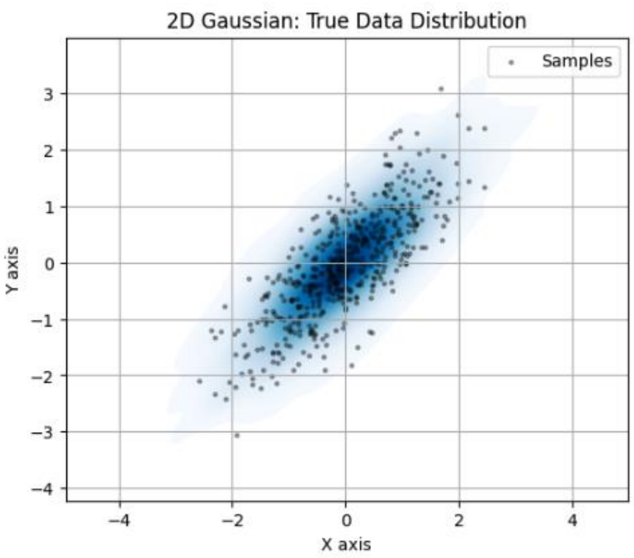




$\mathbb{R}^{256\times256}$

Data Distribution






```python
import numpy as np
import matplotlib.pyplot as plt

# Generate real 1D data samples from a Gaussian distribution N(mean, std^2)
mean1d, std1d = 0.0, 1.0
real_samples_1d = np.random.normal(mean1d, std1d, size=10000)

# Plot histogram of real samples
plt.figure(figsize=(6,4))
plt.hist(real_samples_1d, bins=30, density=True, color='skyblue')
plt.title("Histogram of real 1D samples")
plt.xlabel("x"); plt.ylabel("Density")
plt.show()
```

```python
# Generate real 2D data (Gaussian with correlation)
mean2d = [0.0, 0.0]
cov2d = [[1.0, 0.8],
         [0.8, 1.0]]  # covariance matrix with correlation 0.8
real_samples_2d = np.random.multivariate_normal(mean2d, cov2d, size=5000)
```

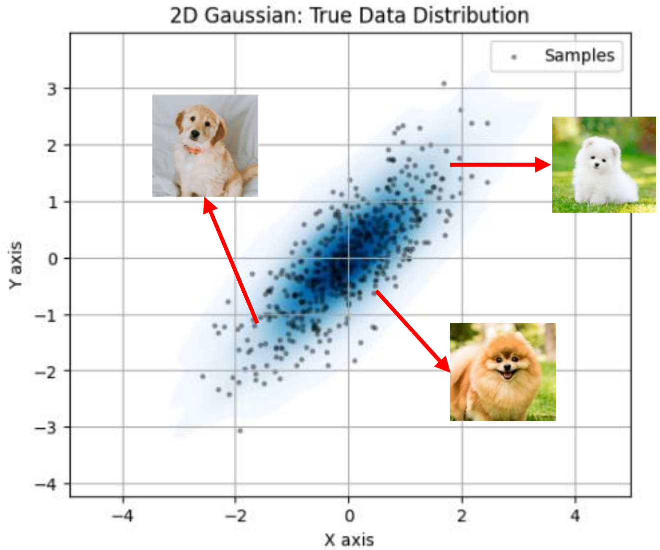
```python
import seaborn as sns

# Unpack real samples into x and y
x_real = real_samples_2d[:, 0]
y_real = real_samples_2d[:, 1]

# Plot: scatter with density contour
plt.figure(figsize=(6, 5))
sns.kdeplot(x=x_real, y=y_real, fill=True, cmap="Blues", thresh=0.01, levels=100)
plt.scatter(x_real[:500], y_real[:500], s=5, color="black", alpha=0.3, label="Samples")
plt.title("2D Gaussian: True Data Distribution")
plt.xlabel("X axis"); plt.ylabel("Y axis")
plt.legend()
plt.grid(True)
plt.axis("equal")
plt.show()
```

Estimating Data Distribution





Estimate Data Distribution



Generative Model

Probability distribution of the **objective** based on the **observed data**

$$\{x_i\}_{i=1}^{N} \xrightarrow[\text{Good Model}]{} P(x) \xrightarrow[\text{Good Data}]{} x$$

- **Machine Learning Methods**
  - Gaussian Kernel Density Estimation
  - Gaussian Mixture Models

Using existing function to estimate what you do not know that can best fit your observation

- **Deep Learning Methods**
  - Auto-Encoder (AE)
  - Variational AE (VAE)
  - Generative Adversarial Network
  - Diffusion Model

Using learnable function to estimate what you do not know that can best fit your observation

Generative Model

Probability distribution of the **objective** based on the **observed data**

$$\{x_i\}_{i=1}^{N} \xrightarrow{\text{Good Model}} P(x) \xrightarrow{\text{Good Data}} x$$

- **Machine Learning Methods**
  - Gaussian Kernel Density Estimation
  - Gaussian Mixture Models

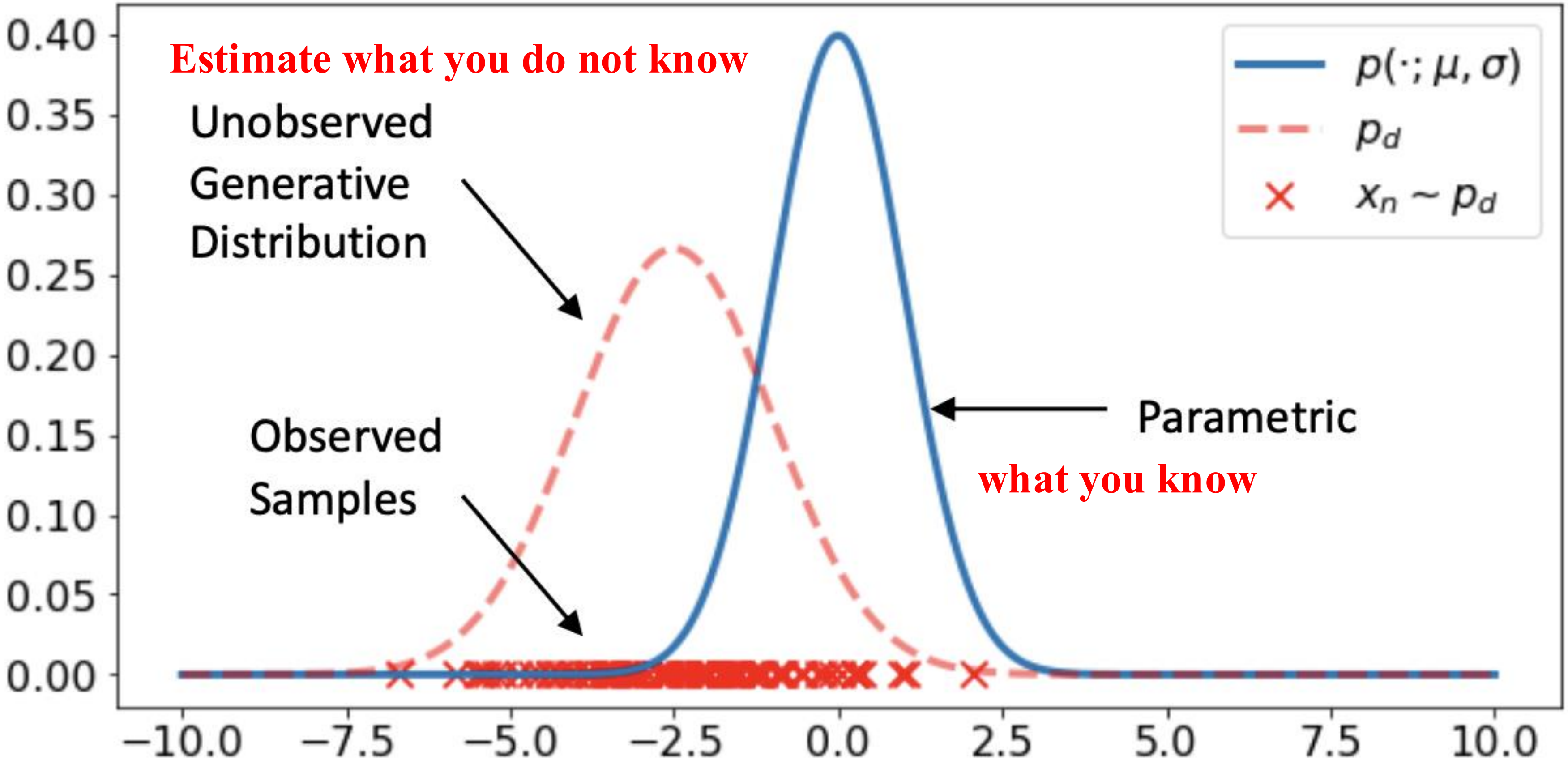
Using existing function to estimate what you do not know that can best fit your observation

- **Deep Learning Methods**
  - Auto-Encoder (AE)
  - Variational AE (VAE)
  - Generative Adversarial Network
  - Diffusion Model

Using learnable function to estimate what you do not know that can best fit your observation

Gaussian Kernel Density Estimation

Using existing function to estimate what you do not know that can best fit your observation

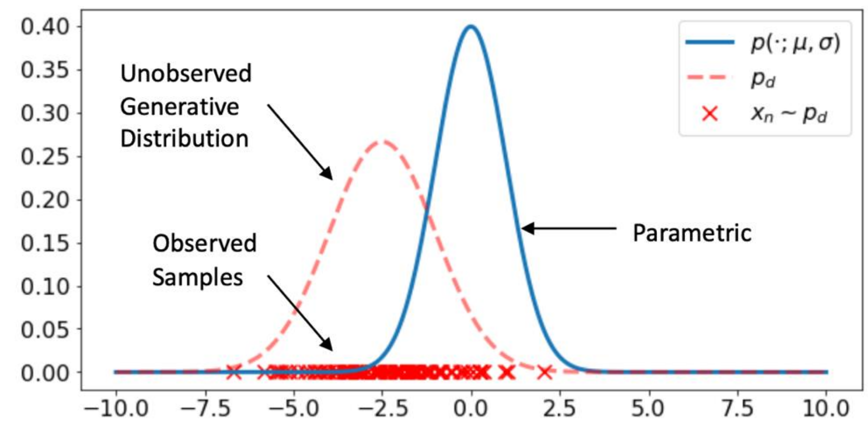


Gaussian Kernel Density Estimation

Using existing function to estimate what you do not know that can best fit your observation

What you know is Gaussian

$$p(x;\mu,\sigma) = N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



What you observe is a set of i.i.d samples from a Gaussian

$$\{x_i | x_i \sim P_d\}_{i=1}^{N}$$

How can we estimate the gaussian distribution?

What is μ, σ ?

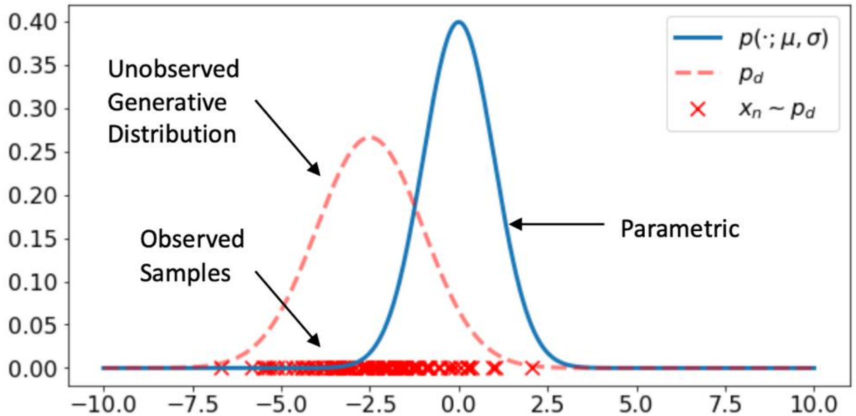
Gaussian Kernel Density Estimation

Using existing function to estimate what you do not know that can best fit your observation

$$\{x_i | x_i \sim P_d\}_{i=1}^{N}$$

$$p(x;\mu,\sigma) = N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

What is μ, σ ?



1- Maximum Likelihood:

$$\underset{\mu,\sigma}{\operatorname{argmax}}\, p(x_1, \ldots, x_N; \mu, \sigma) = \prod_{n=1}^{N} p(x_n; \mu, \sigma)$$

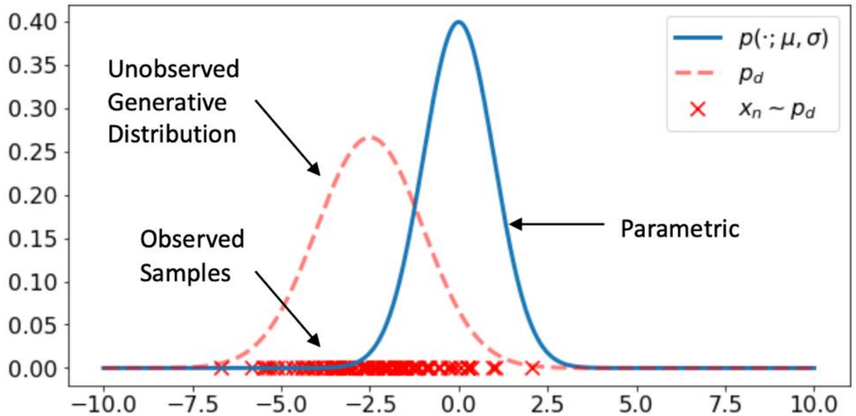
Gaussian Kernel Density Estimation

Using existing function to estimate what you do not know that can best fit your observation

$$\{x_i | x_i \sim P_d\}_{i=1}^{N}$$

$$p(x;\mu,\sigma) = N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

What is μ, σ ?



2 - Maximum Log-Likelihood:

$$\underset{\mu,\sigma}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} p(x_n;\mu,\sigma)\right) = \sum_{n=1}^{N} \log(p(x_n;\mu,\sigma))$$

Gaussian Kernel Density Estimation

Gaussian Kernel Density Estimation

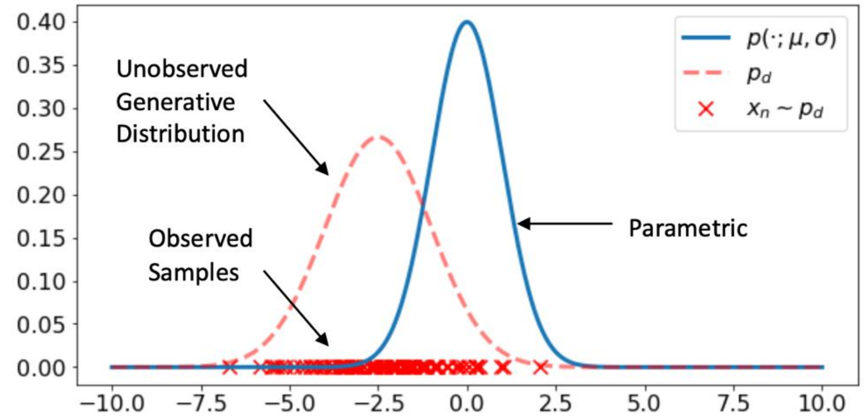
$$\{x_i | x_i \sim P_d\}_{i=1}^{N}$$

$$\downarrow$$

$$p(x;\mu,\sigma) = N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\downarrow$$

What is μ, σ ?



3- Minimizing Negative Log-Likelihood:

$$\underbrace{\operatorname*{argmin}_{\mu,\sigma} \sum_{n=1}^{N} \frac{\log(2\pi\sigma^2)}{2} + \frac{(x_n-\mu)^2}{2\sigma^2}}_{\text{L}} \longrightarrow \begin{cases} \dfrac{\partial L}{\partial \mu} = 0 \Rightarrow \mu_* = \dfrac{1}{N}\sum_{n=1}^{N} x_n \\ \dfrac{\partial L}{\partial \sigma} = 0 \Rightarrow \sigma_*^2 = \dfrac{1}{N}\sum_{n=1}^{N} (x_n - \mu^*)^2 \end{cases}$$

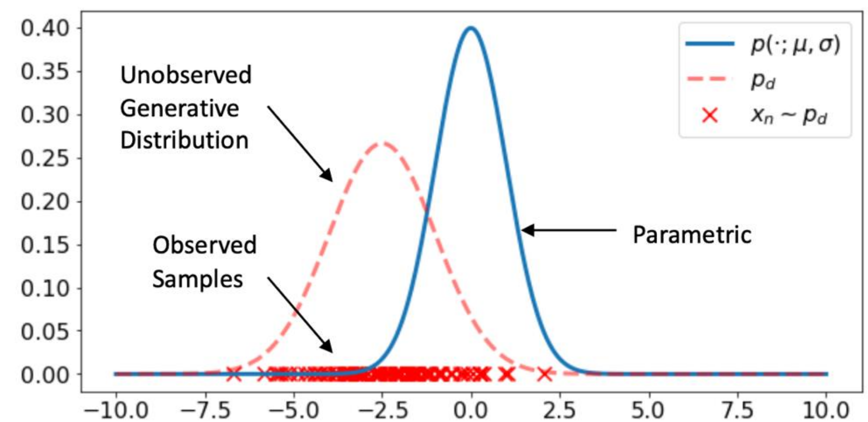
Gaussian Kernel Density Estimation

2 - Maximum Log-Likelihood:

$$\underset{\mu,\sigma}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} p(x_n;\mu,\sigma)\right) = \sum_{n=1}^{N} \log(p(x_n;\mu,\sigma))$$

Monte-Carlo Approximation

$$\int_{\mathbb{R}} p_d(x)\log(p(x;\mu,\sigma))\,dx \approx \frac{1}{N}\sum_{n=1}^{N}\log(p(x_n;\mu,\sigma))$$



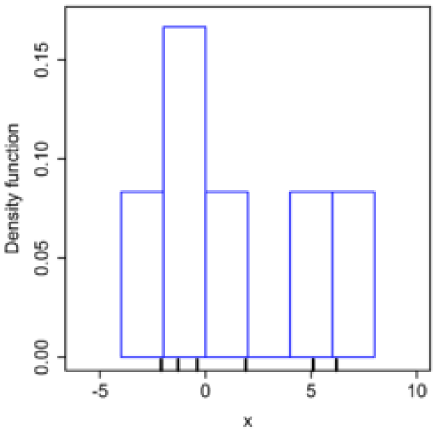
$$\underset{\mu,\sigma}{\operatorname{argmax}} \int_{\mathbb{R}} p_d(x)\log(p(x;\mu,\sigma))\,dx = \underset{\mu,\sigma}{\operatorname{argmax}} \int_{\mathbb{R}} p_d(x)\log(p(x;\mu,\sigma))\,dx - \int_{\mathbb{R}} p_d(x)\log(p_d(x))\,dx$$

$$= \underset{\mu,\sigma}{\operatorname{argmax}} \int_{\mathbb{R}} p_d(x)\log\left(\frac{p(x;\mu,\sigma)}{p_d(x)}\right)dx = \underset{\mu,\sigma}{argmin} \int_{\mathbb{R}} p_d(x)\log\left(\frac{p_d(x)}{p(x;\mu,\sigma)}\right)dx$$

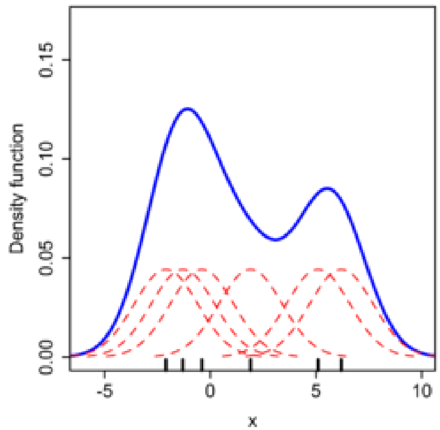
$$= \underset{\mu,\sigma}{\operatorname{argmin}}\, D_{KL}(p_d \,||\, p(\cdot;\mu,\sigma))$$

Maximize Log-Likelihood of parametric distribution = Minimize KL Divergence between the data distribution and the parametric distribution

Problem with Discrete Bin Density Estimation



6 dimension



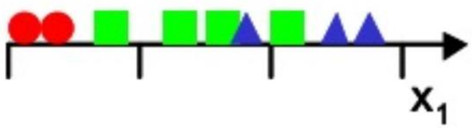
2 dimension

Non-smooth

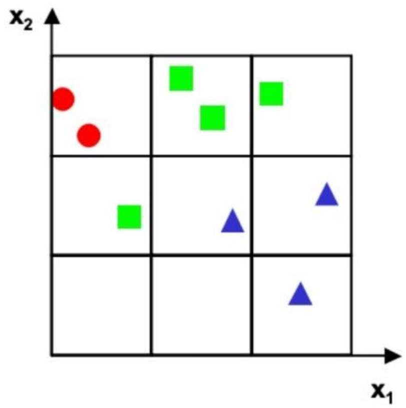
Curse of dimensionality

Problem with Discrete Bin Density Estimation

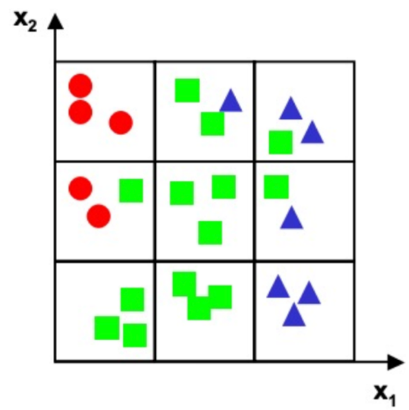
- We add a second feature.



Constant # examples



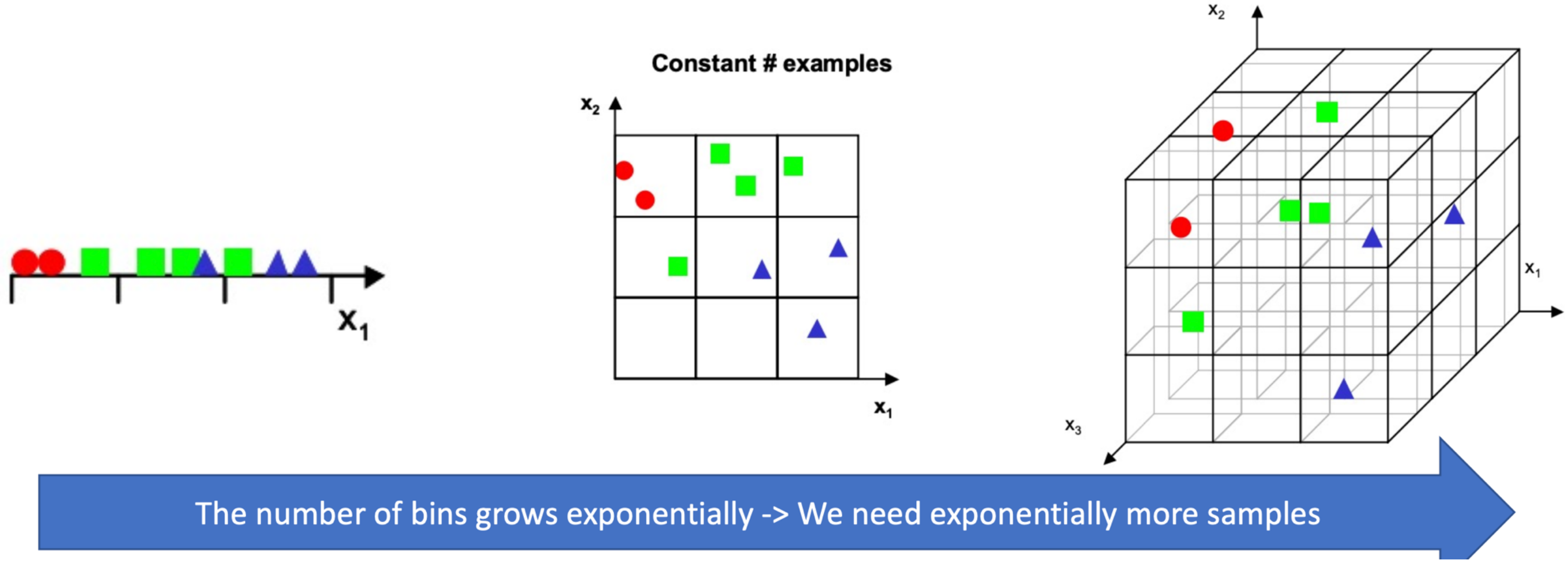
Constant density



- How many samples do we need if we wanted to keep the average density per segment constant?

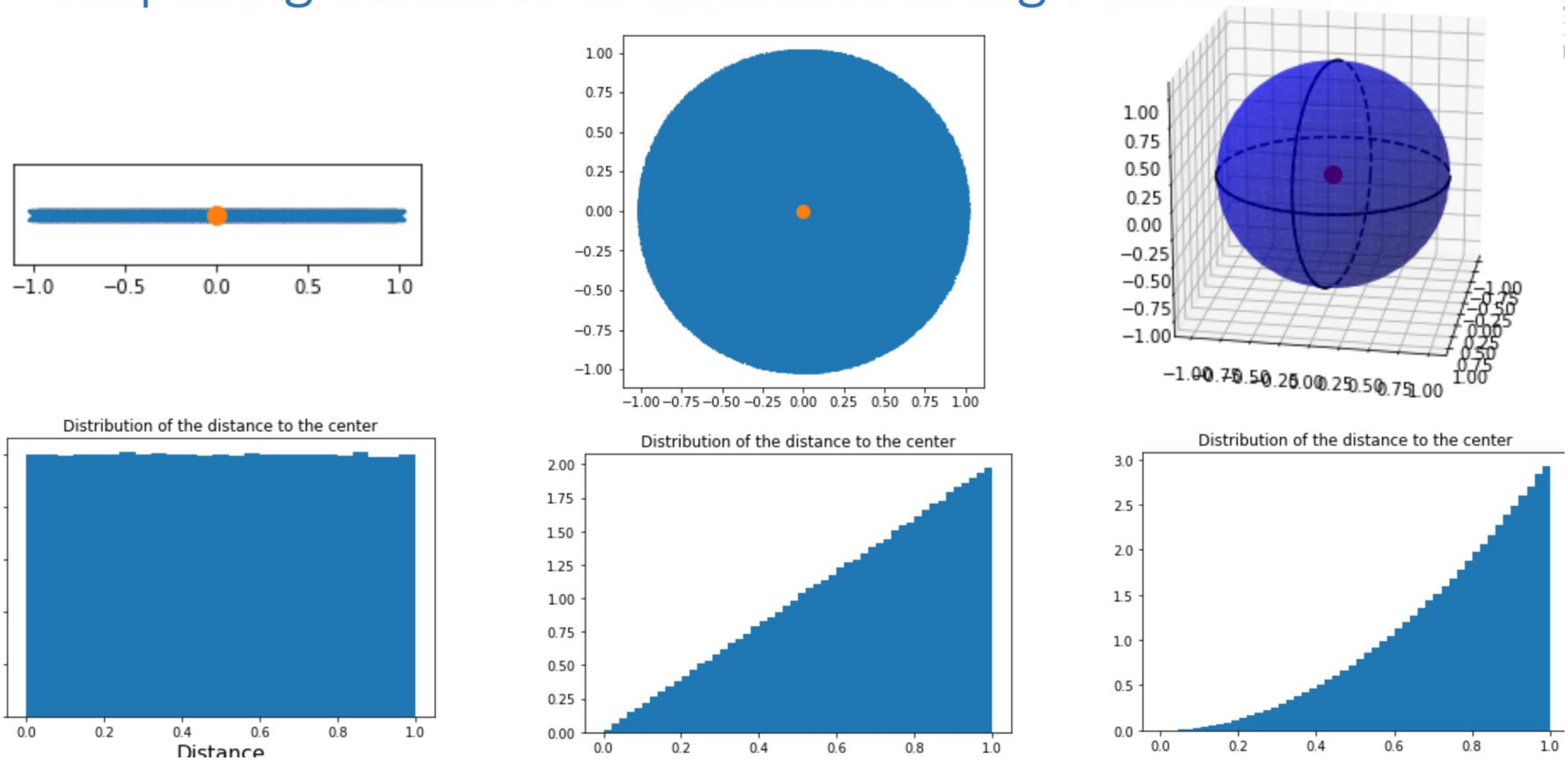
Problem with Discrete Bin Density Estimation

- Lets add a third feature:



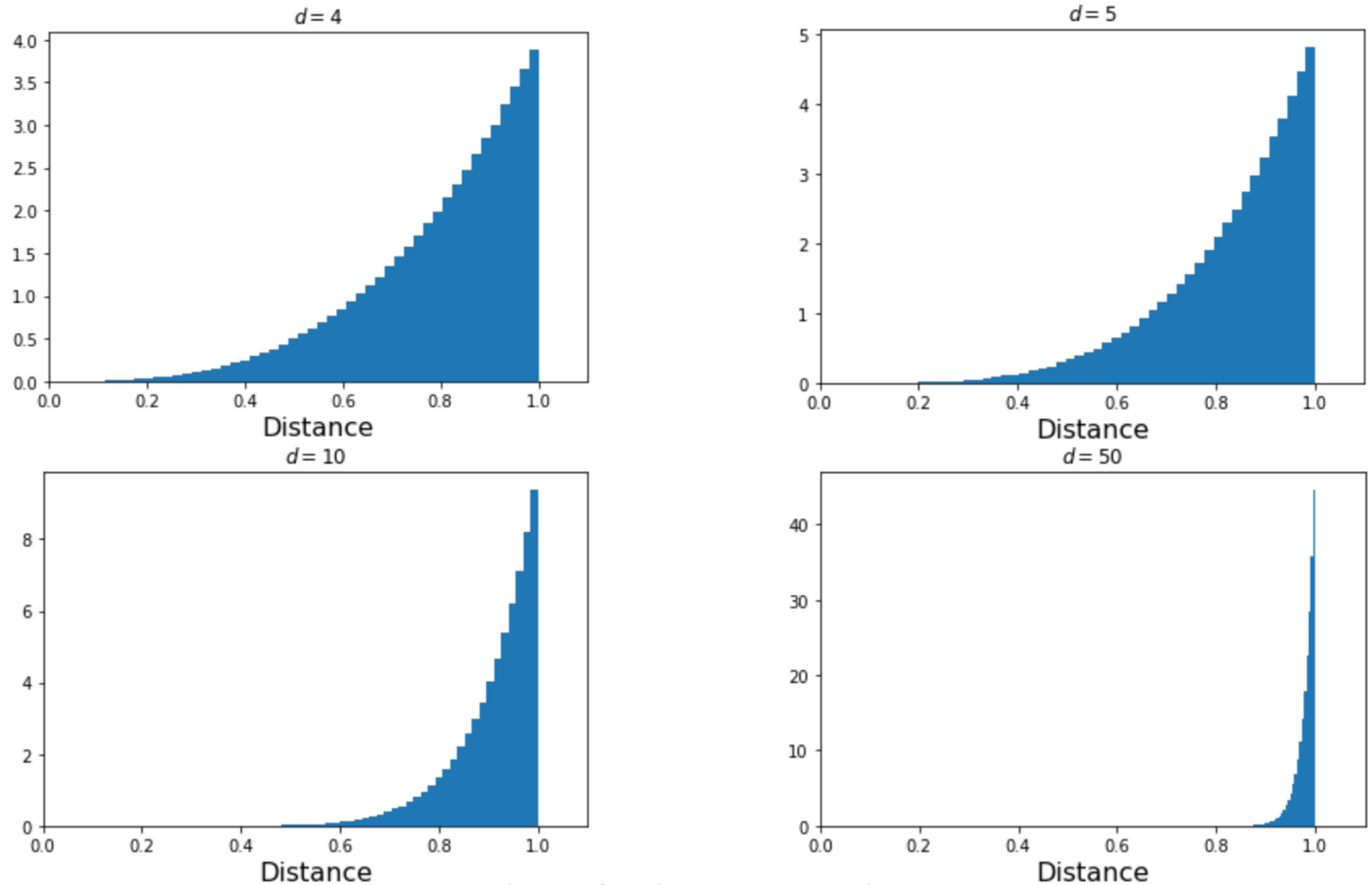
Problem with Discrete Bin Density Estimation

Surprising behavior of distances in high dimensions



Problem with Discrete Bin Density Estimation

Surprising behavior of distances in high dimensions



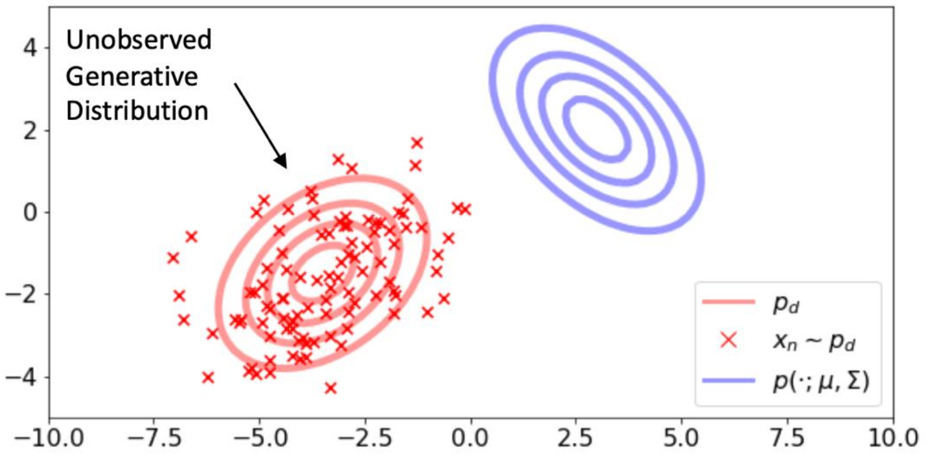
Gaussian Kernel Density Estimation - Example

[DEMO]

Gaussian Kernel Density Estimation

How about 2D Gaussian Dimension?

$$p(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}}$$

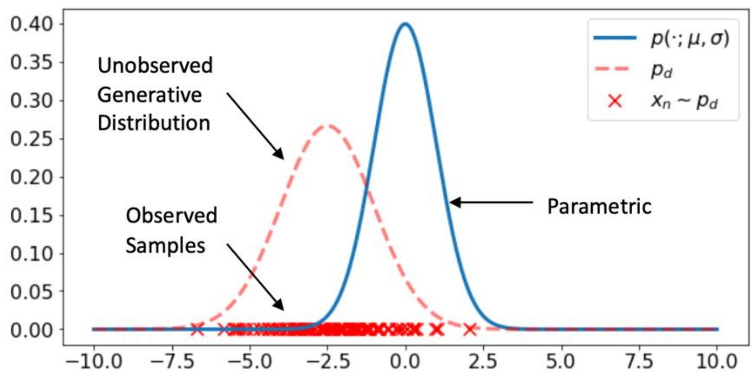


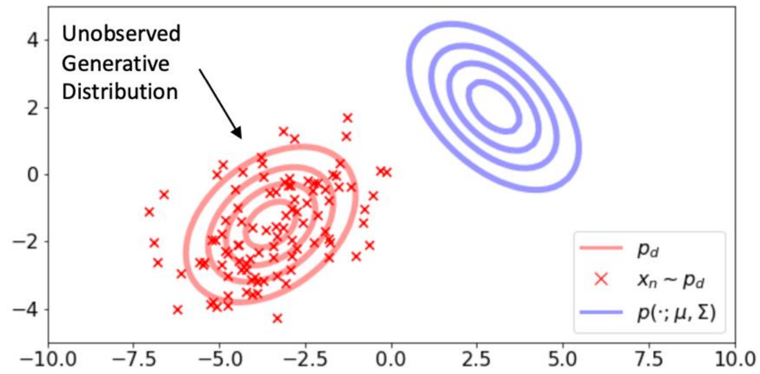
$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \mu_* = \frac{1}{N}\sum_{n=1}^{N} x_n$$

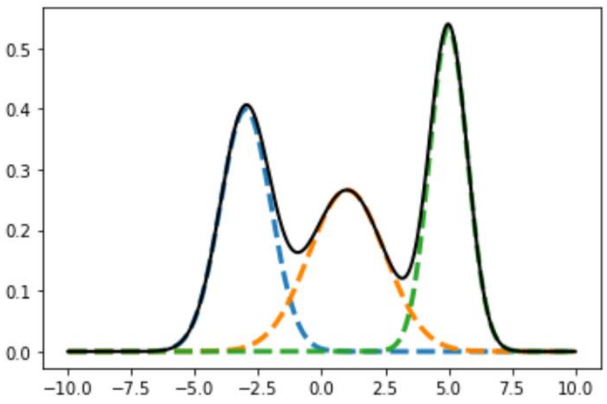
$$\frac{\partial L}{\partial \Sigma} = 0 \Rightarrow \Sigma_* = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu^*)(x_n - \mu^*)^T$$

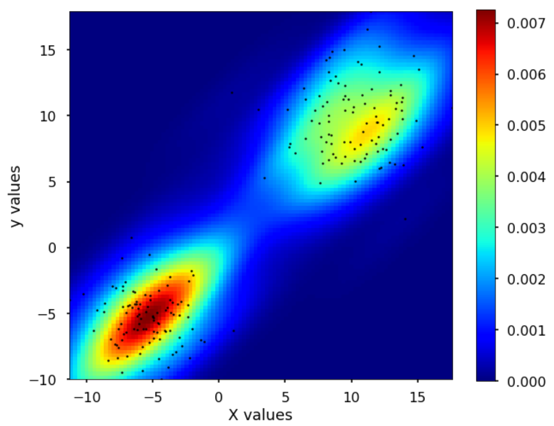
Any problem with such estimation?

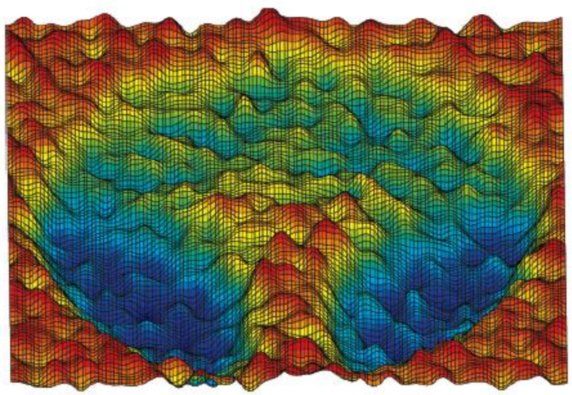
Gaussian Mixture Models





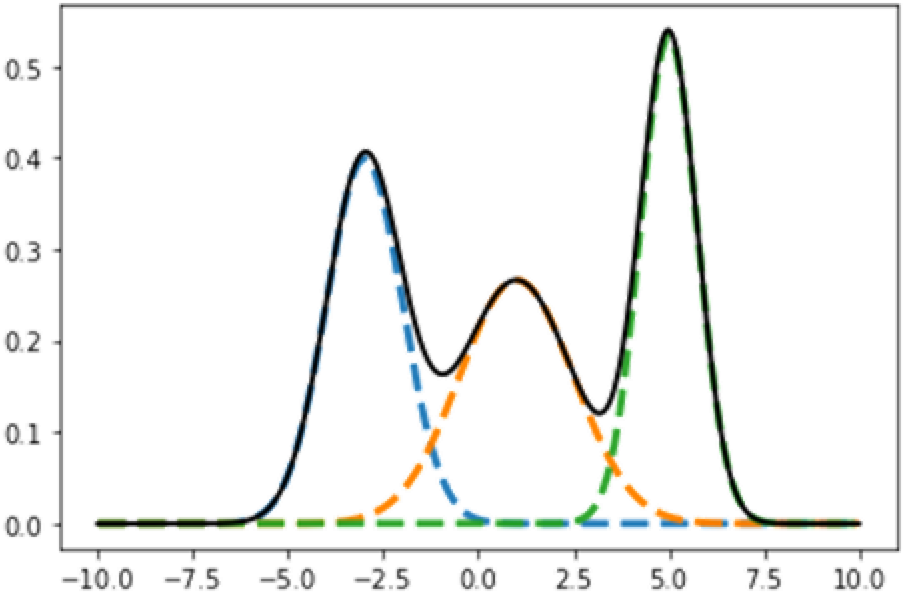






If using one existing function to estimate is not enough, then let's try more!

Gaussian Mixture Models



Assume we have N i.i.d samples from a mixture of Gaussians distribution, $\{x_n \sim p_d\}_{n=1}^{N}$.

- **How do we estimate the parameters of the mixture from the observed samples?**

$$p\left(x; [(\alpha_k, \mu_k, \sigma_k)]_{k=1}^{K}\right) = \sum_{k=1}^{K} \alpha_k N(x; \mu_k, \sigma_k) = \sum_{k=1}^{K} \frac{\alpha_k}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

Where $\alpha_k \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$, and $\sigma_k \geq 0$.

Gaussian Mixture Models

$$p(x; [(\alpha_k, \mu_k, \sigma_k)]_{k=1}^{K}) = \sum_{k=1}^{K} \alpha_k N(x; \mu_k, \sigma_k) = \sum_{k=1}^{K} \frac{\alpha_k}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

Where $\alpha_k \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$, and $\sigma_k \geq 0$.

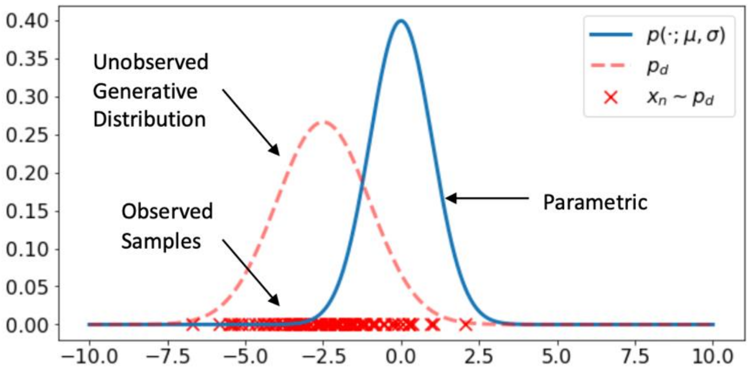
$$\underset{\mu,\sigma}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} p(x_n; \mu, \sigma)\right) = \sum_{n=1}^{N} \log(p(x_n; \mu, \sigma))$$

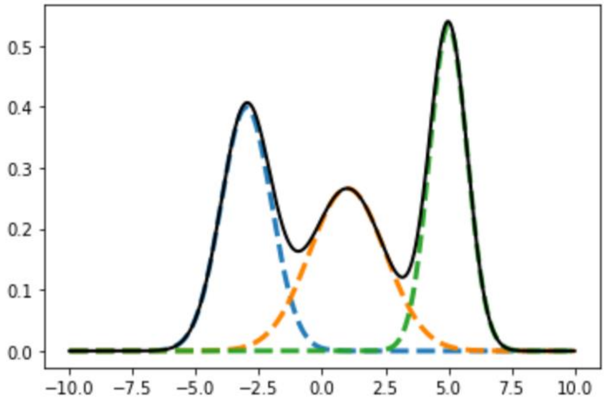
$$\sum_{n=1}^{N} Log(\sum_{k=1}^{K} p(x_n; u, \sigma)) = \sum_{n=1}^{N} Log(\sum_{k=1}^{K} \frac{\alpha_k}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_n-\mu_k)^2}{2\sigma_k^2}})$$

Non-convex and no closed form solution to update all parameters at the single epoch

Gaussian Mixture Models

If, for each sample, we knew which Gaussian it is sampled from the problem would have been solved! Meaning that we could have solved K maximum log-likelihoods to estimate parameters of each Gaussian independent of the others!

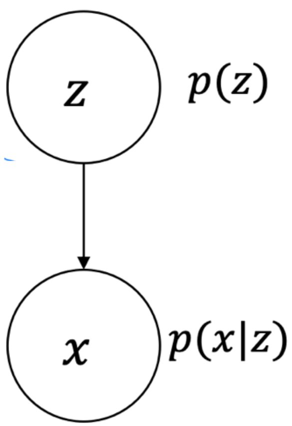




Gaussian Mixture Models

If, for each sample, we knew which Gaussian it is sampled from the problem would have been solved! Meaning that we could have solved K maximum log-likelihoods to estimate parameters of each Gaussian independent of the others!

$z = [z_1, \dots, z_K], z_k \in \{0,1\}$



$$p(x|z_k = 1) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

$$p(x) = \sum_{k=1}^{K} p(x|z_k = 1)p(z_k = 1) = \sum_{k=1}^{K} \frac{\alpha_k}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

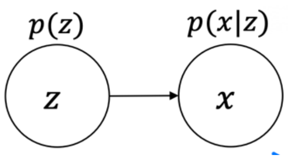
Gaussian Mixture Models

$$p(x) = \sum_z p(z, x)$$

$$p(x) = \int p(z, x)\, dz$$

$$\frac{d}{dx}\log(f(x)) = \frac{1}{f(x)} \cdot f'(x)$$

$$\frac{d}{d\theta}\log p(x) = \frac{d}{d\theta}\log\left(\sum_z p(z,x)\right) = \frac{\frac{d}{d\theta}\sum_z p(z,x)}{\sum_{z'} p(z',x)} = \frac{\sum_z \frac{d}{d\theta} p(z,x)}{\sum_{z'} p(z',x)}$$



$$\frac{\sum_z p(z,x)\frac{\frac{d}{d\theta}p(z,x)}{p(z,x)}}{\sum_{z'} p(z',x)} = \sum_z \left(\frac{p(z,x)}{\sum_{z'} p(z',x)}\right)\frac{d}{d\theta}\log(p(z,x)) = \sum_z p(z|x)\frac{d}{d\theta}\log(p(z,x))$$

$$\sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)p(z)) = \sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)) + \sum_z p(z|x)\frac{d}{d\theta}\log(p(z))$$

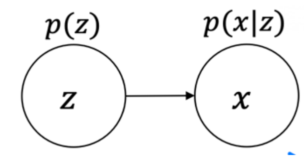
$p(z|x)$ **Soft assignment of data x to each Gaussian**

$\log(p(z))$ **Factuality of the latent distribution z**

Gaussian Mixture Models

$$p(x) = \sum_z p(z,x) \qquad p(x) = \int p(z,x)\,dz \qquad \frac{d}{dx}\log(f(x)) = \frac{1}{f(x)} \cdot f'(x)$$

$$\frac{d}{d\theta}\log p(x) = \frac{d}{d\theta}\log\left(\sum_z p(z,x)\right) = \frac{\frac{d}{d\theta}\sum_z p(z,x)}{\sum_{z'} p(z',x)} = \frac{\sum_z \frac{d}{d\theta} p(z,x)}{\sum_{z'} p(z',x)}$$



$$\frac{\sum_z p(z,x)\frac{\frac{d}{d\theta}p(z,x)}{p(z,x)}}{\sum_{z'} p(z',x)} = \sum_z \left(\frac{p(z,x)}{\sum_{z'} p(z',x)}\right)\frac{d}{d\theta}\log(p(z,x)) = \sum_z p(z|x)\frac{d}{d\theta}\log(p(z,x))$$

$$\sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)p(z)) = \sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)) + \sum_z p(z|x)\frac{d}{d\theta}\log(p(z))$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) = 0$$

$p(z|x)$ **Soft assignment of data x to each Gaussian**

Where $\alpha_k \geq 0, \sum_{k=1}^{K}\alpha_k = 1$

$\log(p(z))$ **Factuality of the latent distribution z**

Constrained Optimization

Gaussian Mixture Models

$$\max/\min f(x)$$

$$s.t.\, g_i(x) = 0, i = 1,2, \dots, m$$

$$\underset{\mu,\sigma}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} p(x_n; \mu, \sigma)\right) = \sum_{n=1}^{N} \log(p(x_n; \mu, \sigma))$$

Where $\alpha_k \geq 0$, $\sum_{k=1}^{K} \alpha_k = 1$, and $\sigma_k \geq 0$.

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

$$\nabla \mathcal{L}(x, \lambda) = 0$$

Karush–Kuhn–Tucker conditions

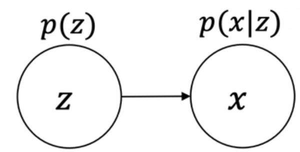
Gaussian Mixture Models

$$p(x) = \sum_z p(z,x)$$

$$p(x) = \int p(z,x)\,dz$$

$$\frac{d}{dx}\log(f(x)) = \frac{1}{f(x)} \cdot f'(x)$$

$$\frac{d}{d\theta}\log p(x) = \frac{d}{d\theta}\log\left(\sum_z p(z,x)\right) = \frac{\frac{d}{d\theta}\sum_z p(z,x)}{\sum_{z'} p(z',x)} = \frac{\sum_z \frac{d}{d\theta} p(z,x)}{\sum_{z'} p(z',x)}$$



$$\frac{\sum_z p(z,x)\frac{\frac{d}{d\theta}p(z,x)}{p(z,x)}}{\sum_{z'} p(z',x)} = \sum_z \left(\frac{p(z,x)}{\sum_{z'} p(z',x)}\right)\frac{d}{d\theta}\log(p(z,x)) = \sum_z p(z|x)\frac{d}{d\theta}\log(p(z,x))$$

$$\sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)p(z)) = \sum_z p(z|x)\frac{d}{d\theta}\log(p(x|z)) + \sum_z p(z|x)\frac{d}{d\theta}\log(p(z))$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) + \frac{d}{d\theta}\lambda\left(\sum_{k=1}^{3}\alpha_k - 1\right) = 0$$

$p(z|x)$ **Soft assignment of data x to each Gaussian** $\theta \in \{[(\alpha_k, \mu_k, \sigma_k)]_{k=1}^{K}, \lambda\}$

$\log(p(z))$ **Factuality of the latent distribution z**

Gaussian Mixture Models

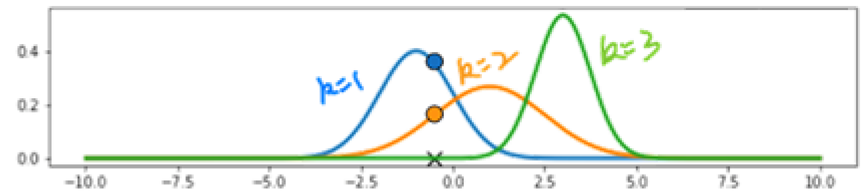
$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) + \lambda\sum_{k=1}^{3}\frac{d}{d\theta}\alpha_k = 0$$

$p(z|x)$ **Soft assignment of data x to each Gaussian**

$\log(p(z))$ **Factuality of the latent distribution z**

$$p(z_k|x_i) = \frac{p(x_i|z_k)p(z_k)}{p(x_i)} = \frac{p(x_i|z_k)p(z_k)}{\sum_{k=1}^{3} p(x_i|z_k)p(z_k)} = r_i^k$$

$$p(z_k) = \alpha_k$$



- $p(x|z)$: what is the probability of x_i coming from z.
- $p(z)$: what is the probability of z
- Our optimization is over $[(\alpha_k, \mu_k, \sigma_k)]_{k=1}^{K}$

Gaussian Mixture Models

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) + \frac{d}{d\theta}\lambda\,(\sum_{k=1}^{3}\alpha_k - 1) = 0$$

$$\theta \in \{[(\alpha_k,\mu_k,\sigma_k)]_{k=1}^{K},\lambda\}$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{\mathbf{z_1},z_2,z_3\}} r_i^k\frac{d}{du_1}\log(p(x_i|z_k)) = 0 \quad \sum_{i=1}^{N}\sum_{z_k\in\{z_1,\mathbf{z_2},z_3\}} r_i^k\frac{d}{du_2}\log(p(x_i|z_k)) = 0 \quad \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,\mathbf{z_3}\}} r_i^k\frac{d}{du_3}\log(p(x_i|z_k)) = 0$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{\mathbf{z_1},z_2,z_3\}} r_i^k\frac{d}{d\sigma_1}\log(p(x_i|z_k)) = 0 \quad \sum_{i=1}^{N}\sum_{z_k\in\{z_1,\mathbf{z_2},z_3\}} r_i^k\frac{d}{d\sigma_2}\log(p(x_i|z_k)) = 0 \quad \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,\mathbf{z_3}\}} r_i^k\frac{d}{d\sigma_3}\log(p(x_i|z_k)) = 0$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{\mathbf{z_1},z_2,z_3\}} r_i^k\frac{d}{d\alpha_1}\log(p(z_k)) + \frac{d}{d\alpha_1}\lambda\,(1-\sum_{k=1}^{3}\alpha_k) = 0$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,\mathbf{z_2},z_3\}} r_i^k\frac{d}{d\alpha_2}\log(p(z_k)) + \frac{d}{d\alpha_2}\lambda\,(1-\sum_{k=1}^{3}\alpha_k) = 0$$

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,\mathbf{z_3}\}} r_i^k\frac{d}{d\alpha_3}\log(p(z_k)) + \frac{d}{d\alpha_3}\lambda\,(1-\sum_{k=1}^{3}\alpha_k) = 0$$

$$p(z_k|x_i) = \frac{p(x_i|z_k)p(z_k)}{p(x_i)} = \frac{p(x_i|z_k)p(z_k)}{\sum_{k=1}^{3}p(x_i|z_k)p(z_k)} = r_i^k$$

$$p(z_k) = \alpha_k$$

Gaussian Mixture Models

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) + \frac{d}{d\theta}\lambda\left(\sum_{k=1}^{3}\alpha_k - 1\right) = 0$$

$$\theta \in \{[(\alpha_k,\mu_k,\sigma_k)]_{k=1}^{K},\lambda\}$$

$$\sum_{i=1}^{N} r_i^1\frac{d}{du_1}\log(p(x_i|z_1)) = 0 \qquad \sum_{i=1}^{N} r_i^2\frac{d}{du_2}\log(p(x_i|z_2)) = 0 \qquad \sum_{i=1}^{N} r_i^3\frac{d}{du_3}\log(p(x_i|z_3)) = 0$$

$$\sum_{i=1}^{N} r_i^1\frac{d}{d\sigma_1}\log(p(x_i|z_1)) = 0 \qquad \sum_{i=1}^{N} r_i^2\frac{d}{d\sigma_2}\log(p(x_i|z_2)) = 0 \qquad \sum_{i=1}^{N} r_i^3\frac{d}{d\sigma_3}\log(p(x_i|z_3)) = 0$$

$$\sum_{i=1}^{N} r_i^1\frac{d}{d\alpha_1}\log(p(z_1)) - \lambda = 0 \qquad \sum_{i=1}^{N} r_i^1\frac{1}{\alpha_1} - \lambda = 0 \qquad \alpha_1 = \sum_{i=1}^{N}\frac{r_i^1}{\lambda} = \sum_{i=1}^{N}\frac{r_i^1}{N} \qquad \sum_{k=1}^{3}\alpha_k = 1$$

$$\sum_{i=1}^{N} r_i^2\frac{d}{d\alpha_2}\log(p(z_2)) - \lambda = 0 \qquad \sum_{i=1}^{N} r_i^2\frac{1}{\alpha_2} - \lambda = 0 \qquad \alpha_2 = \sum_{i=1}^{N}\frac{r_i^2}{\lambda} = \sum_{i=1}^{N}\frac{r_i^2}{N} \qquad \sum_{k=1}^{3}\sum_{i=1}^{N} r_i^k/\lambda = 1$$

$$\sum_{i=1}^{N} r_i^3\frac{d}{d\alpha_3}\log(p(z_3)) - \lambda = 0 \qquad \sum_{i=1}^{N} r_i^3\frac{1}{\alpha_3} - \lambda = 0 \qquad \alpha_3 = \sum_{i=1}^{N}\frac{r_i^3}{\lambda} = \sum_{i=1}^{N}\frac{r_i^3}{N} \qquad N = \sum_{i=1}^{N}\sum_{k=1}^{3} r_i^k = \lambda$$

Gaussian Mixture Models

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) + \frac{d}{d\theta}\lambda\,(\sum_{k=1}^{3}\alpha_k - 1) = 0$$

$$\theta \in \{[(\alpha_k,\mu_k,\sigma_k)]_{k=1}^{K},\lambda\}$$

$$\sum_{i=1}^{N} r_i^1\frac{d}{d\alpha_1}\log(p(z_1)) - \lambda = 0 \qquad \sum_{i=1}^{N} r_i^1\frac{1}{\alpha_1} - \lambda = 0 \qquad \boxed{\alpha_1 = \sum_{i=1}^{N}\frac{r_i^1}{\lambda} = \sum_{i=1}^{N}\frac{r_i^1}{N}} \qquad \sum_{k=1}^{3}\alpha_k = 1$$

$$\sum_{i=1}^{N} r_i^2\frac{d}{d\alpha_2}\log(p(z_2)) - \lambda = 0 \qquad \sum_{i=1}^{N} r_i^2\frac{1}{\alpha_2} - \lambda = 0 \qquad \boxed{\alpha_2 = \sum_{i=1}^{N}\frac{r_i^2}{\lambda} = \sum_{i=1}^{N}\frac{r_i^2}{N}} \qquad \sum_{k=1}^{3}\sum_{i=1}^{N} r_i^k/\lambda = 1$$

$$\sum_{i=1}^{N} r_i^3\frac{d}{d\alpha_3}\log(p(z_3)) - \lambda = 0 \qquad \sum_{i=1}^{N} r_i^3\frac{1}{\alpha_3} - \lambda = 0 \qquad \boxed{\alpha_3 = \sum_{i=1}^{N}\frac{r_i^3}{\lambda} = \sum_{i=1}^{N}\frac{r_i^3}{N}} \qquad N = \sum_{i=1}^{N}\sum_{k=1}^{3} r_i^k = \lambda$$

E-step: Estimate the Latent Variable given the current Gaussian Parameter

Gaussian Mixture Models

$$\sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(x_i|z_k)) + \sum_{i=1}^{N}\sum_{z_k\in\{z_1,z_2,z_3\}} p(z_k|x_i)\frac{d}{d\theta}\log(p(z_k)) + \frac{d}{d\theta}\lambda\left(\sum_{k=1}^{3}\alpha_k - 1\right) = 0$$

$$\theta \in \{[(\alpha_k,\mu_k,\sigma_k)]_{k=1}^{K},\lambda\}$$

$$\sum_{i=1}^{N} r_i^1\frac{d}{du_1}\log(p(x_i|z_1)) = 0 \qquad \sum_{i=1}^{N} r_i^2\frac{d}{du_2}\log(p(x_i|z_2)) = 0 \qquad \sum_{i=1}^{N} r_i^3\frac{d}{du_3}\log(p(x_i|z_3)) = 0$$

$$\sum_{i=1}^{N} r_i^1\frac{d}{d\sigma_1}\log(p(x_i|z_1)) = 0 \qquad \sum_{i=1}^{N} r_i^2\frac{d}{d\sigma_2}\log(p(x_i|z_2)) = 0 \qquad \sum_{i=1}^{N} r_i^3\frac{d}{d\sigma_3}\log(p(x_i|z_3)) = 0$$

- **M**aximization Step: for fixed r_n^k solve the maximum log-likelihood to obtain optimal parameters:
  - Means: $\mu_k = \frac{1}{N_k}\sum_n r_n^k x_n$
  - Variances: $\sigma_k^2 = \frac{1}{N_k}\sum_n r_n^k (x_n - \mu_k)^2$
  - Covariances: $\Sigma_k = \frac{1}{N_k}\sum_n r_n^k (x_n - \mu_k)(x_n - \mu_k)^T$

$$p(z_k|x_i) = \frac{p(x_i|z_k)p(z_k)}{p(x_i)} = \frac{p(x_i|z_k)p(z_k)}{\sum_{k=1}^{3} p(x_i|z_k)p(z_k)} = r_i^k$$

Gaussian Mixture Models

- **E-step: Estimate the Latent Variable given the current Gaussian Parameter**

$$\alpha_1 = \sum_{i=1}^{N} \frac{r_i^1}{\lambda} = \sum_{i=1}^{N} \frac{r_i^1}{N}$$

$$\alpha_2 = \sum_{i=1}^{N} \frac{r_i^2}{\lambda} = \sum_{i=1}^{N} \frac{r_i^2}{N}$$

$$\alpha_3 = \sum_{i=1}^{N} \frac{r_i^3}{\lambda} = \sum_{i=1}^{N} \frac{r_i^3}{N}$$

$$p(z_k|x_i) = \frac{p(x_i|z_k)p(z_k)}{p(x_i)} = \frac{p(x_i|z_k)p(z_k)}{\sum_{k=1}^{3} p(x_i|z_k)p(z_k)} = r_i^k$$

```
# E-step: compute responsibilities for each point and cluster
# Compute Gaussian likelihoods for each cluster
inv_cov0 = torch.linalg.inv(cov[0])
inv_cov1 = torch.linalg.inv(cov[1])
det_cov0 = torch.linalg.det(cov[0])
det_cov1 = torch.linalg.det(cov[1])
# Mahalanobis distances for each point
diff0 = data - mu[0]  # shape (N,2)
diff1 = data - mu[1]
exp_term0 = torch.sum(diff0 @ inv_cov0 * diff0, dim=1)  # (N,)
exp_term1 = torch.sum(diff1 @ inv_cov1 * diff1, dim=1)
# Gaussian PDF for each point under each cluster
coeff0 = 1.0 / (2 * math.pi * torch.sqrt(det_cov0))
coeff1 = 1.0 / (2 * math.pi * torch.sqrt(det_cov1))
p0 = coeff0 * torch.exp(-0.5 * exp_term0)  # (N,)
p1 = coeff1 * torch.exp(-0.5 * exp_term1)  # (N,)
# Compute responsibilities (unnormalized weights)
weighted_p0 = pi[0] * p0
weighted_p1 = pi[1] * p1
total = weighted_p0 + weighted_p1
r0 = weighted_p0 / total    # (N,) responsibility for cluster 0
r1 = weighted_p1 / total    # (N,) responsibility for cluster 1
responsibilities = torch.stack([r0, r1], dim=1)  # shape (N, 2)
```

- **M**aximization Step: for fixed r_n^k solve the maximum log-likelihood to obtain optimal parameters:
  - Means: $\mu_k = \frac{1}{N_k}\sum_n r_n^k x_n$
  - Variances: $\sigma_k^2 = \frac{1}{N_k}\sum_n r_n^k (x_n - \mu_k)^2$
  - Covariances: $\Sigma_k = \frac{1}{N_k}\sum_n r_n^k (x_n - \mu_k)(x_n - \mu_k)^T$

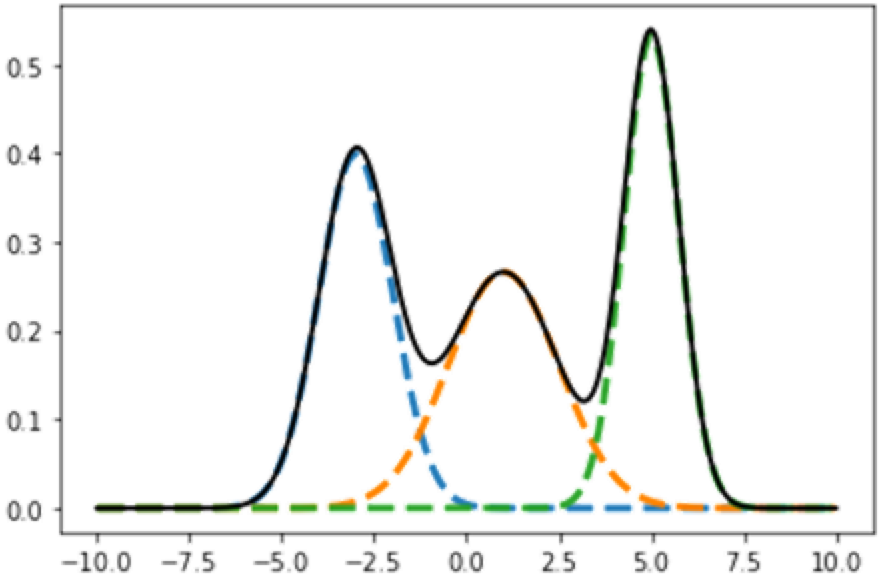
```
# M-step: re-calculate π, μ, Σ using the current responsibilities
N0 = r0.sum()
N1 = r1.sum()
pi = torch.tensor([N0 / N, N1 / N])
mu[0] = (r0.unsqueeze(1) * data).sum(dim=0) / N0
mu[1] = (r1.unsqueeze(1) * data).sum(dim=0) / N1
diff0 = data - mu[0]
diff1 = data - mu[1]
cov[0] = (r0.unsqueeze(1) * diff0).T @ diff0 / N0
cov[1] = (r1.unsqueeze(1) * diff1).T @ diff1 / N1
```

Gaussian Mixture Models – EM algorithm Example

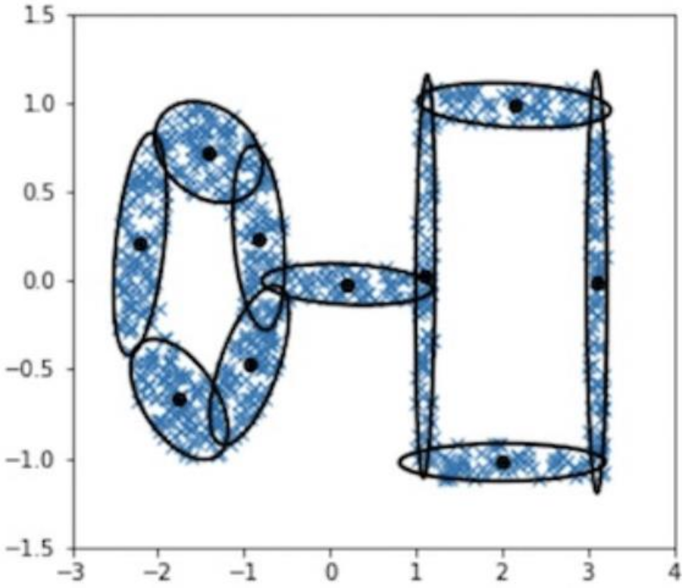
DEMO

Problems

Using **existing function** **to** **estimate what you do not know** **that** **can best fit your observation**



Three bumps but I just give you two different gaussian

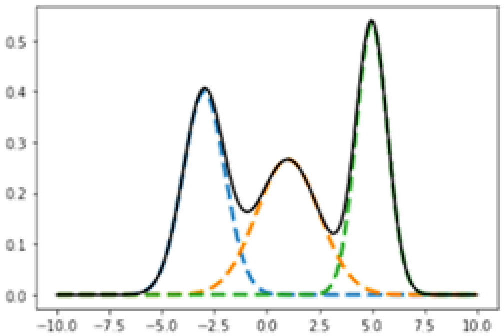


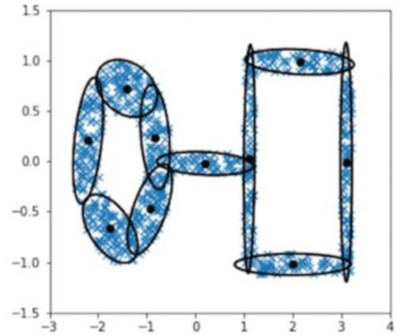
I just give you two different gaussian.

All you know for modeling what you do not know is fixed. But how do you know those fixed things is able to model the unknown thing?

Problems

Using existing function to estimate what you do not know that can best fit your observation





$\mathbb{R}^1, \mathbb{R}^2$

$\downarrow$

$\mathbb{R}^{256\times 256}$

What you have is some low-dimensional data
But what you want to model is some high-dimensional data, how it could be?