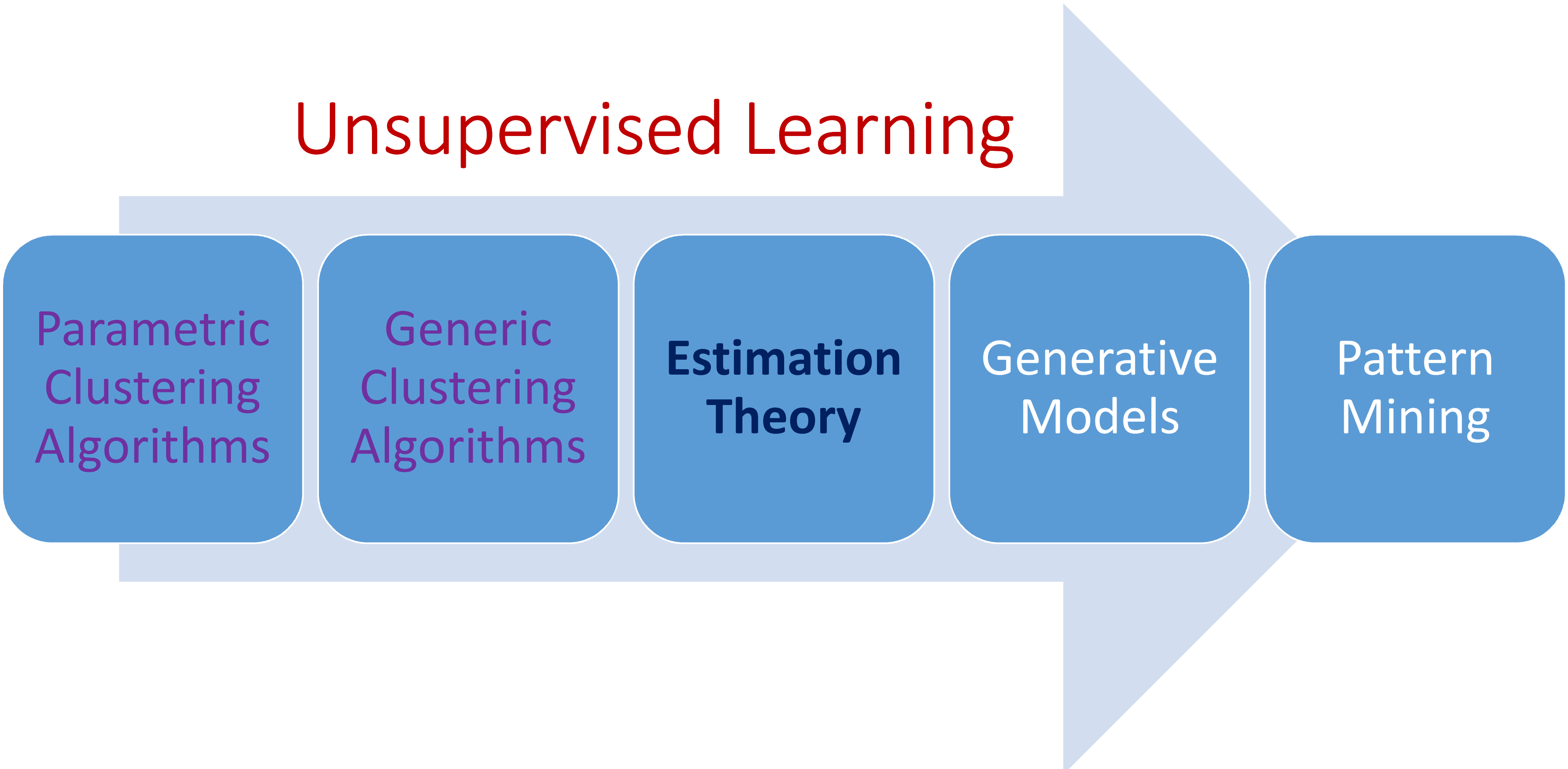
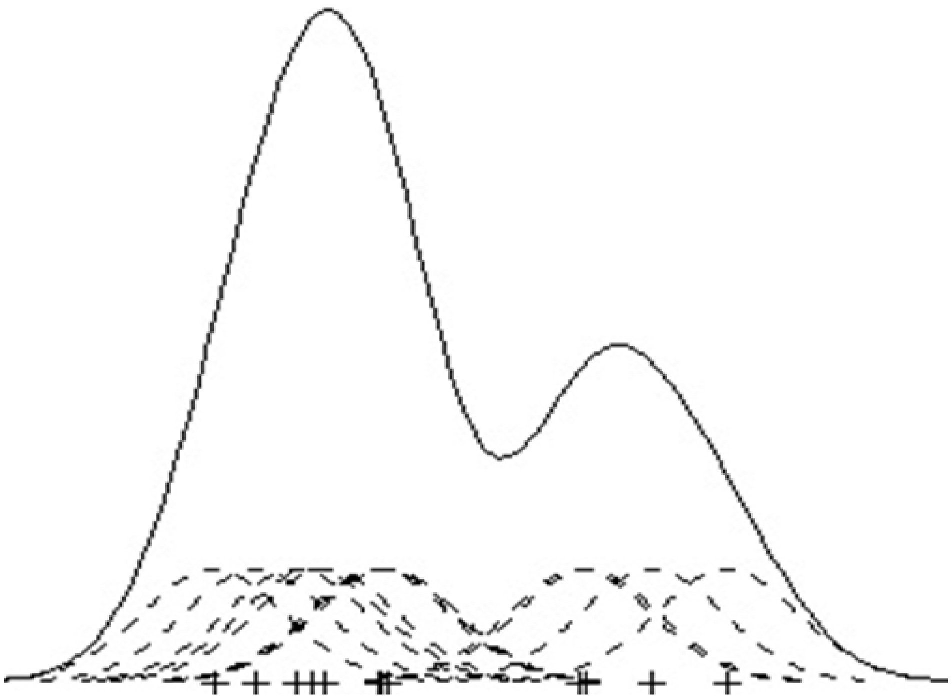
Parametric Estimation: Estimators, MLE & Mixture Models

Prithwijit Guha
Dept. of EEE, IIT Guwahati

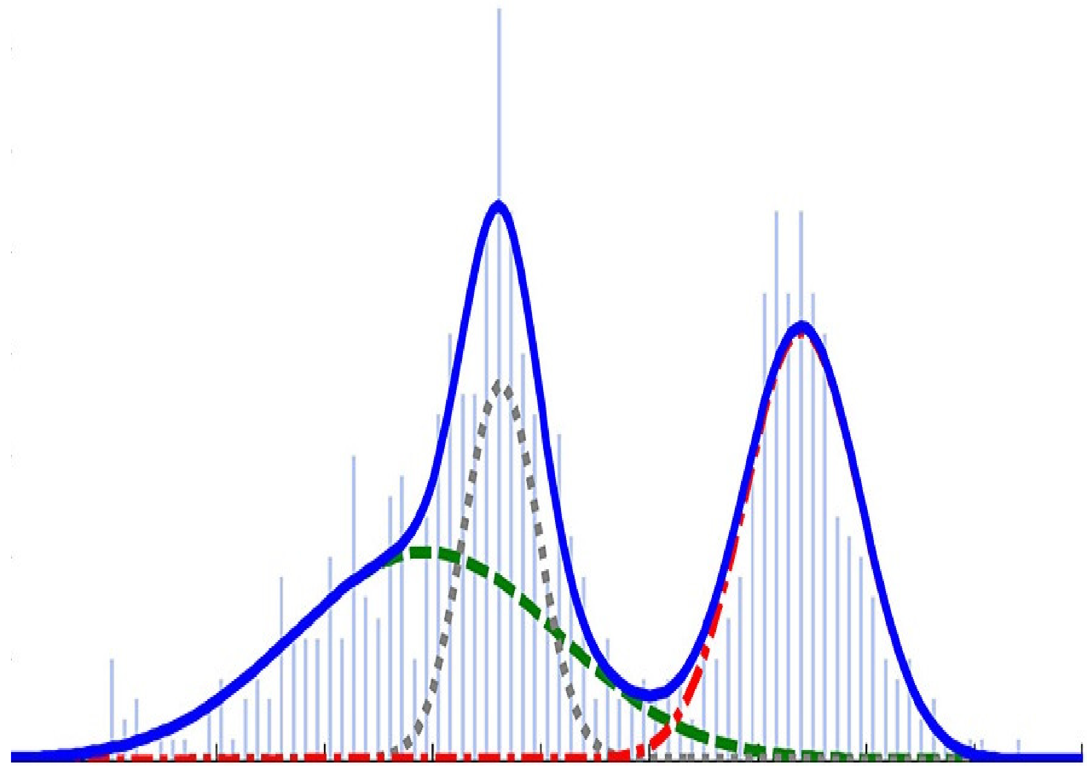


Unsupervised Learning
Parametric Clustering Algorithms
Generic Clustering Algorithms
Estimation Theory
Generative Models
Pattern Mining

Estimation Theory



Non-Parametric Estimation



Parametric Estimation

Estimation Theory

- Introduction to Estimators ($\hat{\theta}$)
- Bias and Variance
- Analysis of Mean ($\hat{\mu}$) and Variance ($\widehat{\sigma^2}$) Estimators
- Maximum Likelihood Estimation
- Learning Mixture Models
- Gaussian Mixture Models

Estimators

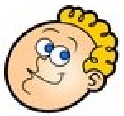
An Estimator is a Rule for Computing an Estimate of a certain Quantity from an Observed Data

An Estimator is a Statistic that Estimates some Fact about the Population. Estimator can be seen as a Rule to Create the Estimate.

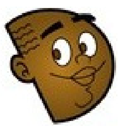
Sample Mean is an Estimate of Population Mean. The Quantity being Estimated is called the Estimand.

Experiments & Samples

$$x_1^{(1)} \quad x_2^{(1)} \quad x_3^{(1)} \quad \dots \quad x_i^{(1)} \quad \dots \quad x_{n-2}^{(1)} \quad x_{n-1}^{(1)} \quad x_n^{(1)}$$



$$\boxed{x_1^{(2)} \quad x_2^{(2)} \quad x_3^{(2)} \quad \dots \quad x_i^{(2)} \quad \dots \quad x_{n-2}^{(2)} \quad x_{n-1}^{(2)} \quad x_n^{(2)}}$$



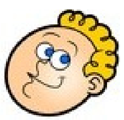
$$x_1^{(3)} \quad x_2^{(3)} \quad x_3^{(3)} \quad \dots \quad x_i^{(3)} \quad \dots \quad x_{n-2}^{(3)} \quad x_{n-1}^{(3)} \quad x_n^{(3)}$$

$$\vdots$$

$$x_1^{(j)} \quad x_2^{(j)} \quad x_3^{(j)} \quad \dots \quad x_i^{(j)} \quad \dots \quad x_{n-2}^{(j)} \quad x_{n-1}^{(j)} \quad x_n^{(j)}$$

$$\vdots$$

Experiments & Samples

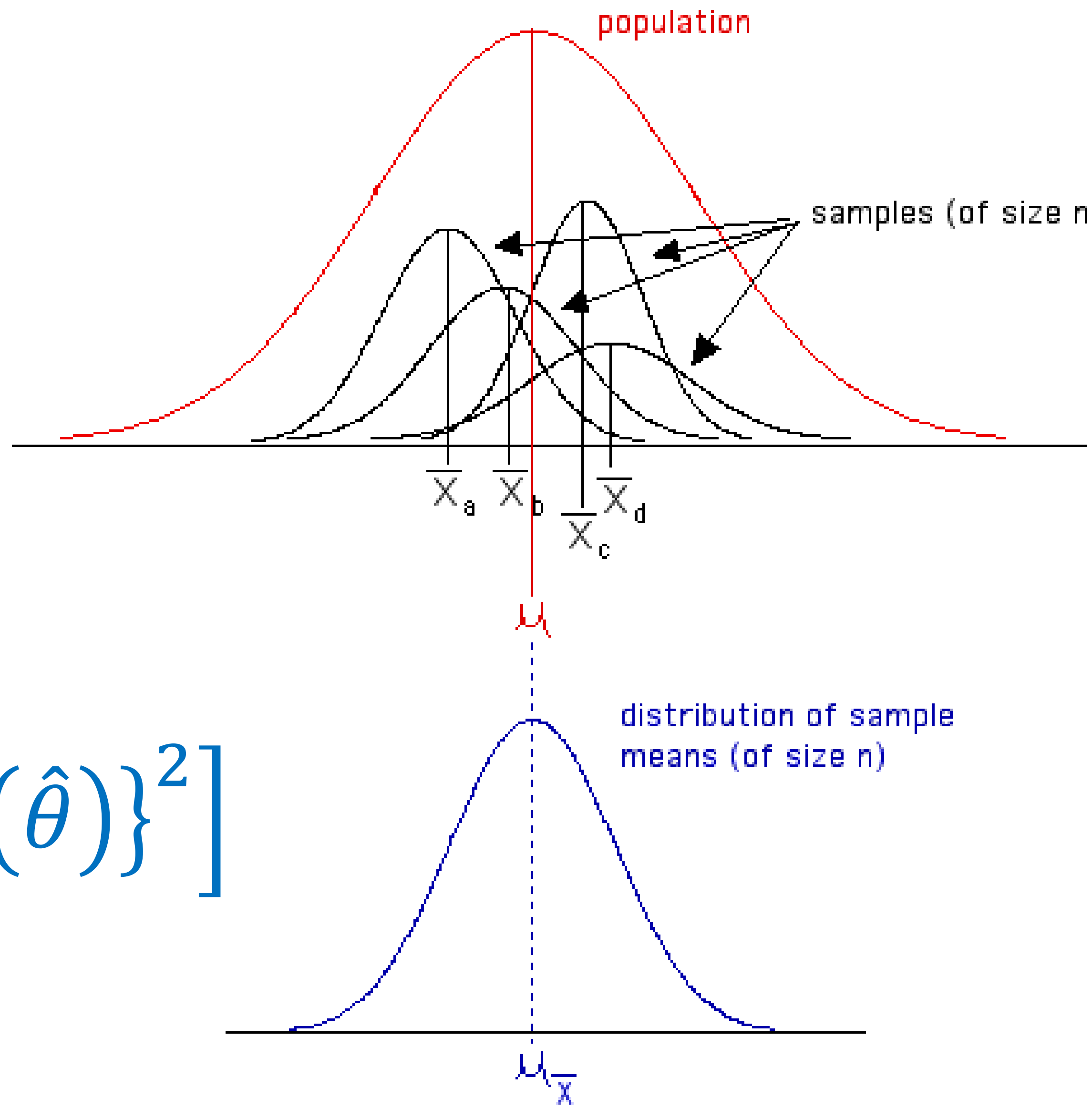
	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	…	$x_i^{(1)}$	…	$x_{n-2}^{(1)}$	$x_{n-1}^{(1)}$	$x_n^{(1)}$
	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	…	$x_i^{(2)}$	…	$x_{n-2}^{(2)}$	$x_{n-1}^{(2)}$	$x_n^{(2)}$
					$\vdots$				
	$x_1^{(j)}$	$x_2^{(j)}$	$x_3^{(j)}$	…	$x_i^{(j)}$	…	$x_{n-2}^{(j)}$	$x_{n-1}^{(j)}$	$x_n^{(j)}$
					$\vdots$				
	x_1	x_2	x_3		x_i		x_{n-2}	x_{n-1}	x_n

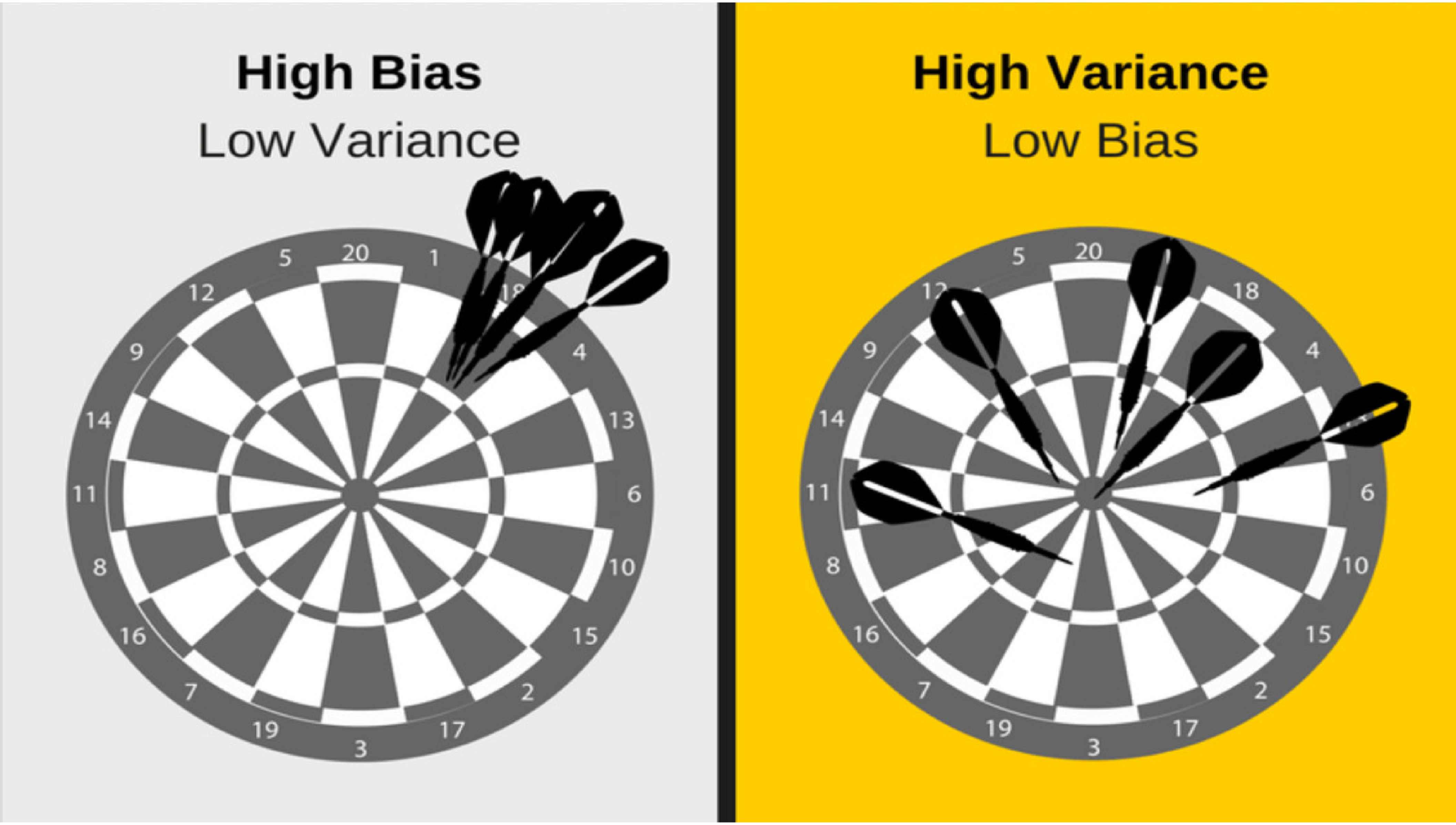
Random Variables (Outcomes in Columns)

Bias & Variance

$$Bias(\hat{\theta}) = \theta - E(\hat{\theta})$$

$$Var(\hat{\theta}) = E\left[\left\{\hat{\theta} - E(\hat{\theta})\right\}^2\right]$$




High Bias
Low Variance
High Variance
Low Bias

Population Mean & Variance

$$\{x_1, x_2, \dots x_i, \dots x_{n-1}, x_n\}$$

Samples Drawn in Any Experiment

$$\forall i \; E(x_i) = E(x) = \mu$$

$$\forall i \; E\left(x_i^2\right) = E(x^2)$$

$$\sigma^2 = E[(x-\mu)^2] = E(x^2) - \mu^2$$

Sample Mean & Variance Estimator

$$\{x_1, x_2, \dots x_i, \dots x_{n-1}, x_n\}$$

Samples Drawn in Any Experiment

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2$$

Bias: Mean Estimator

$$Bias(\hat{\mu}) = \mu - E(\hat{\mu})$$

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n} \times (n\mu) = \mu$$

$$Bias(\hat{\mu}) = 0$$

Sample Mean $(\hat{\mu})$ is an Unbiased Estimator

Variance: Mean Estimator

$$Var(\hat{\mu}) = E[(\hat{\mu} - E(\hat{\mu}))^2] = E[(\hat{\mu} - \mu)^2] = E(\hat{\mu}^2) - \mu^2$$

$$E(\hat{\mu}^2) = E\left[\left\{\frac{1}{n}\sum_{i=1}^{n} x_i\right\}\left\{\frac{1}{n}\sum_{j=1}^{n} x_j\right\}\right] = \frac{1}{n^2} E\left[\left\{\sum_{i=1}^{n} x_i^2\right\} + \left\{\sum_{\substack{i,j \\ i \neq j}} x_i x_j\right\}\right]$$

$$E(\hat{\mu}^2) = \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left(x_i^2\right) + \sum_{\substack{i,j \\ i \neq j}} E(x_i)E(x_j)\right]$$

Variance: Mean Estimator

$$E(\hat{\mu}^2) = \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left(x_i^2\right) + \sum_{\substack{i,j \\ i \neq j}} E(x_i)E(x_j)\right] = \frac{nE(x^2) + (n^2 - n)\mu^2}{n^2}$$

$$E(\hat{\mu}^2) = \frac{E(x^2) - \mu^2}{n} + \mu^2$$

$$\boxed{E(\hat{\mu}^2) = \frac{\sigma^2}{n} + \mu^2}$$

Variance: Mean Estimator

$$E(\hat{\mu}^2) = \frac{\sigma^2}{n} + \mu^2$$

$$Var(\hat{\mu}) = E(\hat{\mu}^2) - \mu^2 = \frac{\sigma^2}{n}$$

$$\lim_{n \to \infty} Var(\hat{\mu}) = \lim_{n \to \infty} \frac{\sigma^2}{n} = 0$$

Variance of Sample Mean Estimator Becomes Smaller for Larger Sample Sizes

Bias: Variance Estimator

$$Bias(\hat{v}) = \sigma^2 - E(\hat{v})$$

$$E(\hat{v}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{\mu}^2\right] = \frac{1}{n}\sum_{i=1}^{n} E\left(x_i^2\right) - E\left(\hat{\mu}^2\right)$$

$$E(\hat{v}) = \frac{1}{n} \times \{nE(x^2)\} - \left(\frac{\sigma^2}{n} + \mu^2\right) = \{E(x^2) - \mu^2\} - \frac{\sigma^2}{n}$$

$$E(\hat{v}) = \sigma^2 - \frac{\sigma^2}{n} = \left(\frac{n-1}{n}\right)\sigma^2$$

Bias: Variance Estimator

$$E(\hat{v}) = \left(\frac{n-1}{n}\right)\sigma^2$$

$$Bias(\hat{v}) = \sigma^2 - E(\hat{v}) = \frac{\sigma^2}{n}$$

$$\lim_{n\to\infty} Bias(\hat{v}) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0$$

Sample Variance ($\hat{v}$) is an Asymptotically Unbiased Estimator

Bias: Variance Estimator

$$E(\hat{v}) = \left(\frac{n-1}{n}\right)\sigma^2 \Rightarrow \left(\frac{n}{n-1}\right)E(\hat{v}) = E\left(\frac{n}{n-1}\hat{v}\right) = \sigma^2$$

$$\hat{v}_{new} = \frac{n}{n-1}\hat{v} = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

$$Bias(\hat{v}_{new}) = \sigma^2 - E(\hat{v}_{new}) = 0$$

Sample Variance ($\hat{v}_{new}$) is an Unbiased Estimator

Likelihood & Log-Likelihood

$$\boldsymbol{D} = \{x_1, x_2, \ldots x_i, \ldots x_{n-1}, x_n\}$$

i.i.d.

$$P(\boldsymbol{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} P(x_i \mid \boldsymbol{\theta})$$

$$\boxed{\boldsymbol{L}(\boldsymbol{\theta}) = \ln P(\boldsymbol{D} \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln P(x_i \mid \boldsymbol{\theta})}$$

Parameter Estimation: Gaussian Distribution

$$\boldsymbol{D} = \{x_1, x_2, \dots x_i, \dots x_{n-1}, x_n\}$$

The dataset ***D*** is drawn from a Gaussian Distribution with mean μ and variance $v = \sigma^2$. The most likely distribution parameters $\hat{\mu}_{MLE}$ and $\hat{v}_{MLE}$ need to estimated from available data.

$$P(x \mid \mu, v) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}}$$

Parameter Estimation: Gaussian Distribution

$$\boldsymbol{D} = \{x_1, x_2, \ldots x_i, \ldots x_{n-1}, x_n\}$$

i.i.d.

$$\boldsymbol{L}(\boldsymbol{\theta}) = \ln P(\,\boldsymbol{D} \mid \mu, v\,) = \sum_{i=1}^{n} \ln P(x_i \mid \mu, v)$$

$$P(x \mid \mu, v) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}}$$

Parameter Estimation: Gaussian Distribution

$$\boldsymbol{L}(\mu, v) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln v - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \boldsymbol{L}(\mu, v)}{\partial \mu} = -\frac{1}{2v}\sum_{i=1}^{n} 2(x_i - \mu)(-1) = \frac{1}{v}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial \boldsymbol{L}(\mu, v)}{\partial \mu} = 0 \quad \Rightarrow \quad \boxed{\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

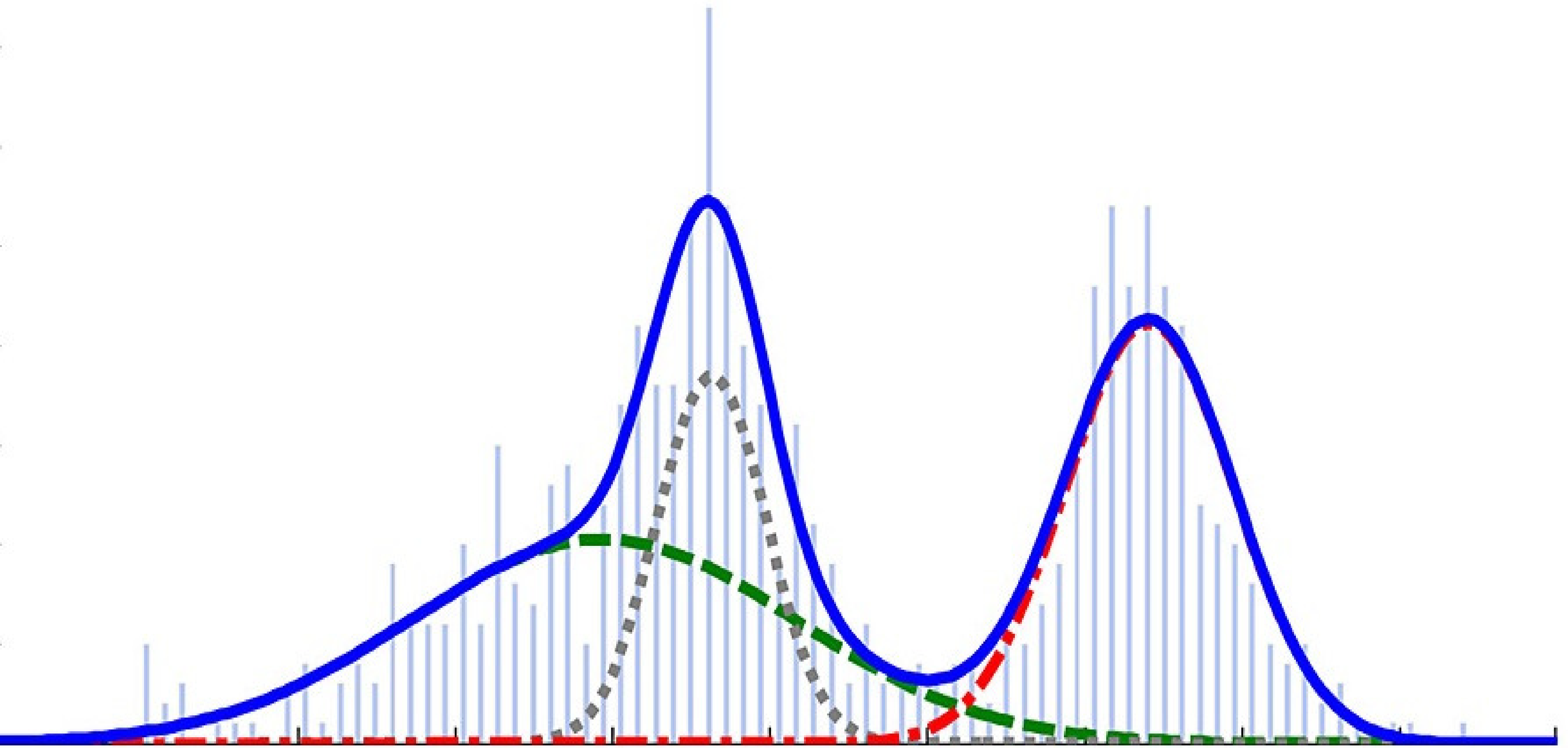
Parameter Estimation : Gaussian Distribution

$$\boldsymbol{L}(\mu, v) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln v - \frac{1}{2v}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \boldsymbol{L}(\mu, v)}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \boldsymbol{L}(\mu, v)}{\partial \mu} = 0 \Rightarrow \boxed{\hat{v}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2}$$

Learning Mixture Models



The Mixture Model

$$\boldsymbol{D} = \{x_1, x_2, \dots x_i, \dots x_{n-1}, x_n\}$$

i.i.d.

$$P(x \mid \boldsymbol{\theta}) = \sum_{j=1}^{m} P\left(x \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j\right) P(\boldsymbol{\omega}_j)$$

Evaluating Gradient of $L(\boldsymbol{\theta})$

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln P(x_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left\{ \sum_{j=1}^{m} P(x_i \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j) P(\boldsymbol{\omega}_j) \right\}$$

$$\nabla_{\boldsymbol{\theta}_r} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{P(\boldsymbol{\omega}_r)}{P(x_i \mid \boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}_r} P(x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r)$$

Assumption: $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_k$ are Functionally Independent

Introducing the Posterior

$$P(\boldsymbol{\omega}_j \mid x_k, \boldsymbol{\theta}) = \frac{P(x_k \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j)P(\boldsymbol{\omega}_j)}{P(x_k \mid \boldsymbol{\theta})}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}_r} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{P(\boldsymbol{\omega}_r)}{P(x_i \mid \boldsymbol{\theta})} \boldsymbol{\nabla}_{\boldsymbol{\theta}_r} P(x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r)$$

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}_r} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{P(\omega_r \mid x_i, \boldsymbol{\theta}) \boldsymbol{\nabla}_{\boldsymbol{\theta}_r} P(x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r)}{P(x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r)}$$

Gradient of Log-Likelihood

$$\nabla_{\boldsymbol{\theta}_r} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{P(\,\omega_r \mid x_i, \boldsymbol{\theta}\,)\nabla_{\boldsymbol{\theta}_r} P(\,x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r\,)}{P(\,x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r\,)}$$

$$\boxed{\nabla_{\boldsymbol{\theta}_r} L(\boldsymbol{\theta}) = \sum_{i=1}^{n} P(\,\omega_r \mid x_k, \boldsymbol{\theta}\,)\nabla_{\boldsymbol{\theta}_r} \ln P(\,x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r\,)}$$

Gaussian Mixture Models (GMM)

$$P(x \mid \boldsymbol{\theta}) = \sum_{j=1}^{m} P\left(x \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j \right) P(\boldsymbol{\omega}_j)$$

$$P\left(x \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j \right) = \frac{1}{\sqrt{2\pi v_j}} e^{-\frac{(x-\mu_j)^2}{2v_j}}$$

GMM: Component Mean

$$\ln P(\, x_i \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j \,) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln v_j - \frac{\left(x_i - \mu_j\right)^2}{2v_j}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_r} = \sum_{i=1}^{n} P(\, \boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta} \,) \frac{\partial}{\partial \mu_r} \{\ln P(\, x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r \,)\}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_r} = \sum_{i=1}^{n} P(\, \boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta} \,) \left(\frac{x_i - \mu_r}{v_r}\right)$$

GMM: Component Mean

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_r} = \sum_{i=1}^{n} P(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}) \left(\frac{x_i - \mu_r}{v_r} \right)$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \mu_r} = 0 \quad \Longrightarrow \quad \boxed{\hat{\mu}_r = \frac{\sum_{i=1}^{n} x_i P(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta})}{\sum_{i=1}^{n} P(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta})}}$$

GMM: Component Variance

$$\ln P\left(x_i \mid \boldsymbol{\omega}_j, \boldsymbol{\theta}_j \right) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln v_j - \frac{\left(x_i - \mu_j\right)^2}{2v_j}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial v_r} = \sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta} \right) \frac{\partial}{\partial v_r}\{\ln P\left(x_i \mid \boldsymbol{\omega}_r, \boldsymbol{\theta}_r \right)\}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial v_r} = \sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta} \right) \left\{ -\frac{1}{2v_r} + \frac{(x_i - \mu_r)^2}{2v_r^2} \right\}$$

GMM: Component Variance

$$\frac{\partial L(\boldsymbol{\theta})}{\partial v_r} = \sum_{i=1}^{n} P(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}) \left\{ -\frac{1}{2v_r} + \frac{(x_i - \mu_r)^2}{2v_r^2} \right\}$$

$$\frac{\partial L(\boldsymbol{\theta})}{\partial v_r} = 0 \quad \Rightarrow \quad \boxed{\hat{v}_r = \frac{\sum_{i=1}^{n} (x_i - \mu_r)^2 P(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta})}{\sum_{i=1}^{n} P(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta})}}$$

GMM: Mean & Variance Update

$$\hat{\mu}_r^{(t+1)} = \frac{\sum_{i=1}^{n} x_i P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)} \right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)} \right)}$$

$$\hat{v}_r^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{ x_i - \mu_r^{(t)} \right\}^2 P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)} \right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid x_i, \boldsymbol{\theta}^{(t)} \right)}$$

$x \in \mathbb{R}^1$

GMM: Mean & Covariance Update

$$\widehat{\boldsymbol{\mu}}_r^{(t+1)} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)} \right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)} \right)}$$

$$\widehat{\boldsymbol{C}}_r^{(t+1)} = \frac{\sum_{i=1}^{n} \left\{ \boldsymbol{x}_i - \boldsymbol{\mu}_r^{(t)} \right\} \left\{ \boldsymbol{x}_i - \boldsymbol{\mu}_r^{(t)} \right\}^T P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)} \right)}{\sum_{i=1}^{n} P\left(\boldsymbol{\omega}_r \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(t)} \right)}$$

$\boldsymbol{x} \in \mathbb{R}^d$

GMM: Applications

- Parametric Estimation of Distribution from Data
- Data Analytics using Estimated Distribution
- GMM Likelihoods used for Classification
- Posteriograms as Features Constructed from Data
- Numerous Applications in Different Domains

Summary

- Introduction to Estimators ($\hat{\theta}$)
- Bias and Variance
- Analysis of Mean ($\hat{\mu}$) and Variance ($\widehat{\sigma^2}$) Estimators
- Maximum Likelihood Estimation
- Learning Mixture Models
- Gaussian Mixture Models

Thank You