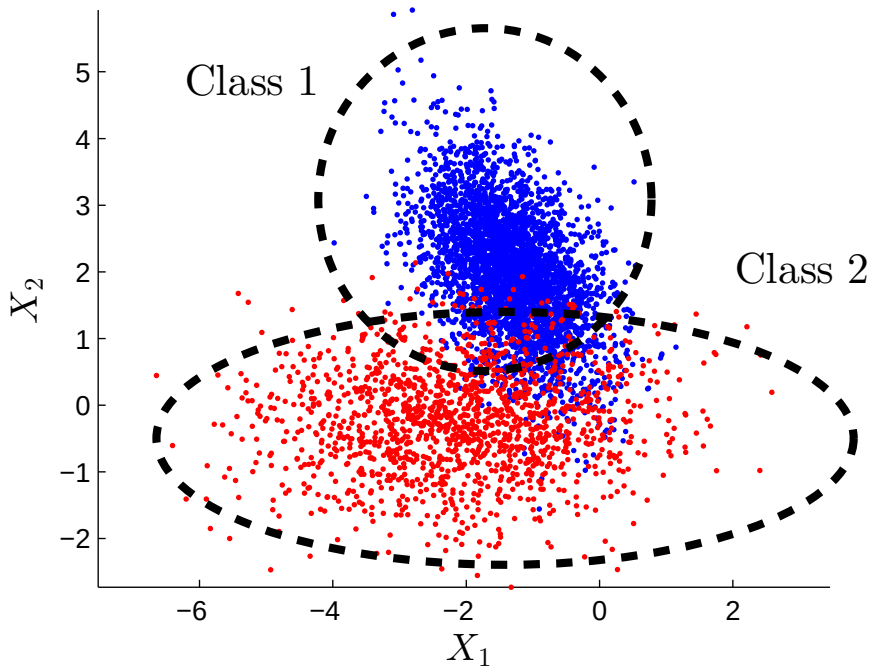
Generative vs. Discriminative Models, Maximum Likelihood Estimation, Mixture Models

Mihaela van der Schaar

Department of Engineering Science
University of Oxford

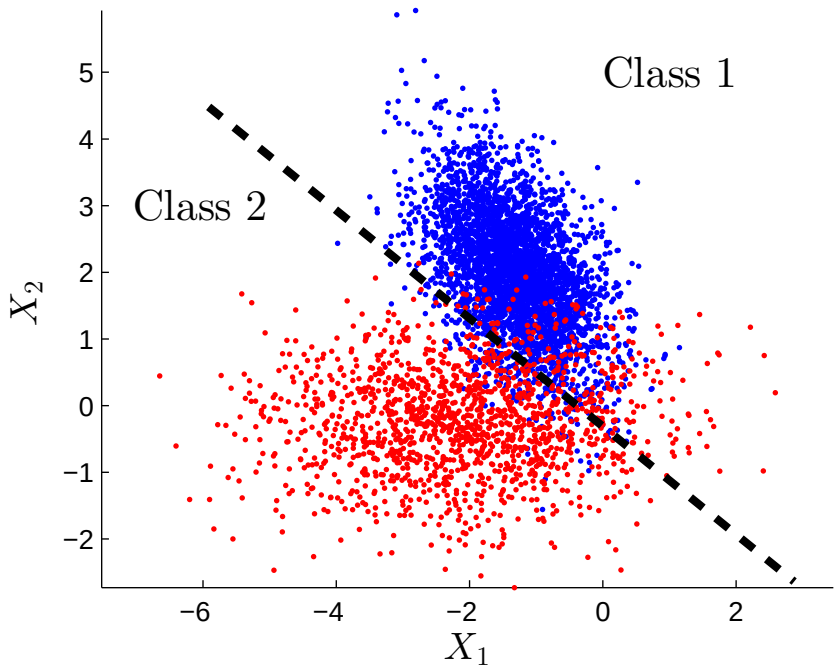
Generative vs Discriminative Approaches

- **Machine learning:** learn a (random) function that maps a variable X (feature) to a variable Y (class) using a (labeled) dataset $\mathcal{M} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
  - **Discriminative Approach:** learn $\mathbb{P}(Y|X)$.
  - **Generative Approach:** learn $\mathbb{P}(Y, X) = \mathbb{P}(Y|X)\mathbb{P}(X)$.



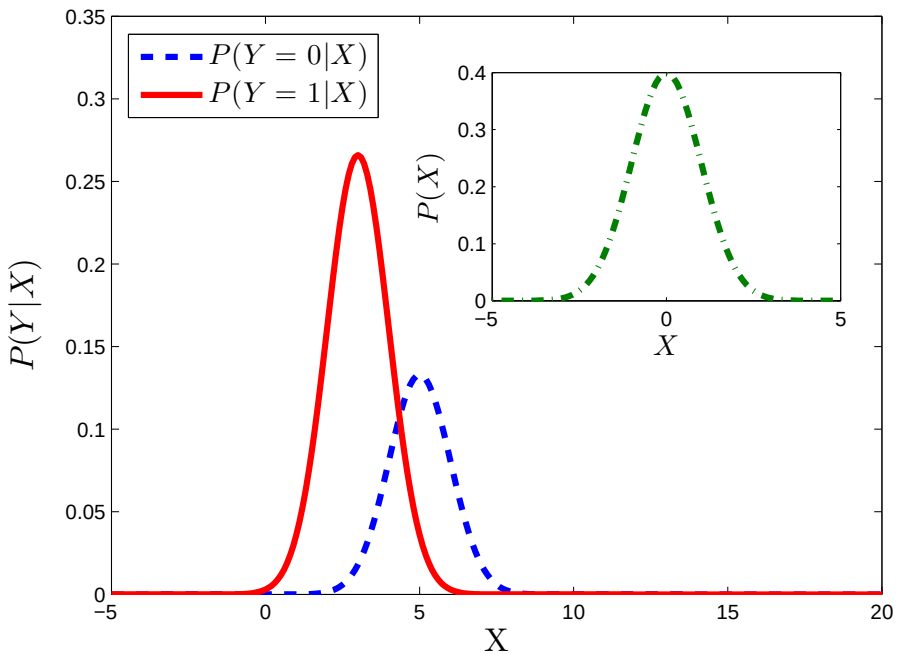
Generative vs Discriminative Approaches

- **Discriminative Approach:** Finds a good fit for $\mathbb{P}(Y|X)$ without explicitly modeling the generative process.
- **Example techniques:** K nearest neighbors, logistic regression, linear regression, SVMs, perceptrons, etc.
- **Example problem:** 2 classes, separate the classes.



Generative vs Discriminative Approaches

- **Generative Approach:** Finds a probabilistic model (a joint distribution $\mathbb{P}(Y, X)$) that explicitly models the distribution of both the features and the corresponding labels (classes).
- **Example techniques:** Naive Bayes, Hidden Markov Models, etc.



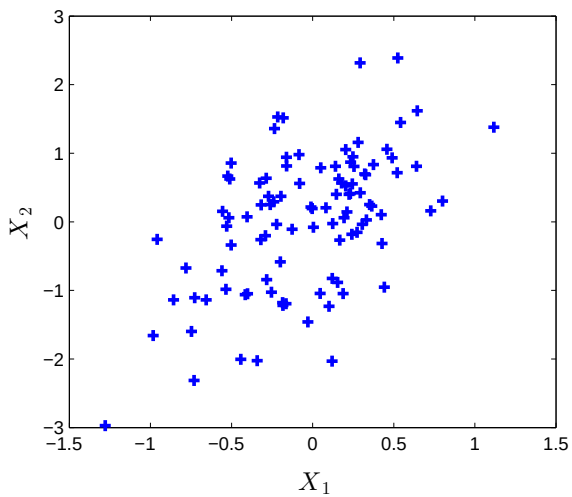
Generative vs Discriminative Approaches

- **Generative Approach:** Finds parameters that explain all data.
  - Makes use of all the data.
  - Flexible framework, can incorporate many tasks (e.g. classification, regression, survival analysis, generating new data samples similar to the existing dataset, etc).
  - Stronger modeling assumptions.
- **Discriminative Approach:** Finds parameters that help to predict relevant data.
  - Learns to perform better on the given tasks.
  - Weaker modeling assumptions.
  - Less immune to overfitting.

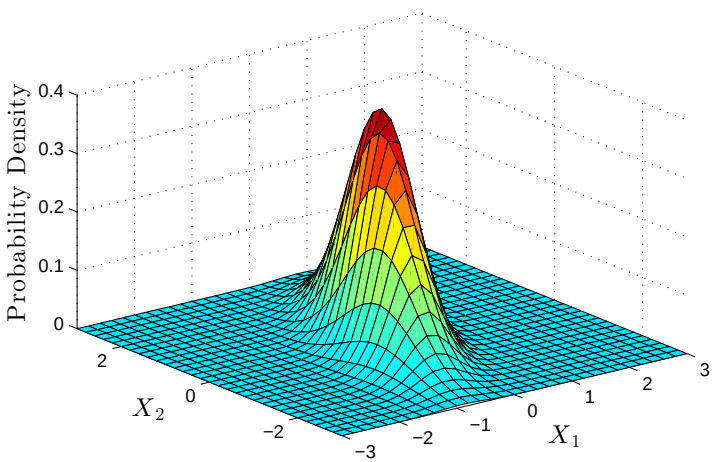
Problem Setup

- We are given a dataset $\mathcal{D} = (X_i)_{i=1}^n$ with n entries.
  - **Example:** X_i's are the annual incomes of n individuals picked randomly from a large population.
- **Goal:** estimate the **probability distribution** that describes the entire population from which the random samples $(X_i)_{i=1}^n$ are drawn.

What we observe: random samples drawn from a distribution



What we want to estimate: the distribution!



Models: Parametric Families of Distributions

- **Key to make progress:** restrict to a *parametrized* family of distributions!
- **Formalization becomes as follows:**
  - The dataset $\mathcal{D}$ comprise independent and identically distributed (iid) samples from a distribution $\mathbb{P}_\theta$ with a parameter θ, i.e.

$$X_i \sim \mathbb{P}_\theta, \ (X_1, X_2, \ldots, X_n) \sim \mathbb{P}_\theta^{\otimes n}.$$

  - The distribution $\mathbb{P}_\theta$ belongs to the family $\mathcal{P}$, i.e.

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\},$$

    where Θ is a **parameter space**.
  - **Estimating the distribution $\mathbb{P}_\theta \in \mathcal{P} \rightarrow$ estimating the parameter $\theta \in \Theta$!**

The Likelihood Function

How is the family of models $\mathcal{P}$ related to the dataset $\mathcal{D}$?

- **The likelihood function $\mathcal{L}_n(\theta, \mathcal{D})$:** is defined as

$$\mathcal{L}_n(\theta, \mathcal{D}) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i).$$

- Intuitively, $\mathcal{L}_n(\theta, \mathcal{D})$ quantifies how compatible is any choice of θ with the occurrence of $\mathcal{D}$.

Maximum Likelihood Estimator (MLE)

Given a dataset $\mathcal{D}$ of size n drawn from a distribution $\mathbb{P}_\theta \in \mathcal{P}$, the MLE estimate of θ is defined as

$$\hat{\theta}_n^* = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta, \mathcal{D}).$$

Why is $\hat{\theta}_n^*$ a good estimator for θ?

1. **Consistency:** the estimate $\hat{\theta}_n^*$ converges to θ *in probability*!

$$\hat{\theta}_n^* \xrightarrow{p} \theta.$$

2. **Asymptotic Normality:** can compute asymptotic *confidence intervals*!

$$\sqrt{n}\,(\hat{\theta}_n^* - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\theta)).$$

3. **Asymptotic Efficiency:** the asymptotic variance of $\hat{\theta}_n^*$ is, in fact, equal to the **Cramer-Rao** lower bound for the variance of a consistent, asymptotically normally distributed estimator!
4. **Invariance under re-parametrization:** If $g(.)$ is a continuous and continuously differentiable function, then the MLE of $g(\theta)$ is $g(\hat{\theta}_n^*)$.

- **See proofs in (Keener, Chapter 8).**

The Gaussian Family $\mathcal{P}$ and Parameter Space Θ

- The dataset $\mathcal{D} = (X_1, X_2, \ldots, X_n)$ is drawn from a distribution $\mathbb{P}_\theta^{\otimes n} = \prod_{i=1}^n \mathbb{P}_\theta(X_i)$, where

$$\mathbb{P}_\theta(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \cdot \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right),$$

  where $\theta = (\mu, \sigma)$.
- The parameter space Θ is

$$\Theta = \left\{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\right\}.$$

- The family $\mathcal{P}$ is the family of Gaussian distributions given by

$$\mathcal{P} = \left\{\frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{(X-\mu)^2}{2\sigma^2}} : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\right\}.$$

The Gaussian Likelihood Function

- The likelihood function is given by

$$\mathcal{L}(\theta, \mathcal{D}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\sqrt{2\pi}\,\sigma\right)^{-n} \cdot e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

- The MLE estimate $\hat{\theta}_n^* = (\hat{\mu}_n^*, \hat{\sigma}_n^*)$ is given by

$$(\hat{\mu}_n^*, \hat{\sigma}_n^*) = \underset{\mu\in\mathbb{R},\sigma\in\mathbb{R}^+}{\arg\max} \left(\sqrt{2\pi}\,\sigma\right)^{-n} \cdot e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

- It is usually more convenient to work with the **log-likelihood function**

$$\log\left(\mathcal{L}(\theta, \mathcal{D})\right) = \frac{-n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Finding $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ (I)

- The log(.) operation is monotonic, therefore

$$\arg\max_{\theta \in \Theta} \log(\mathcal{L}(\theta, \mathcal{D})) = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, \mathcal{D})$$

- Can solve the optimization problem $\arg\max_\theta \mathcal{L}(\theta, \mathcal{D})$ by equating the first derivative of $\log(\mathcal{L}(\theta, \mathcal{D}))$ with respect to θ and equating to zero (first-order condition), i.e.

$$\frac{\partial}{\partial \theta} \log(\mathcal{L}(\theta, \mathcal{D})) = 0.$$

- **What properties $\log(\mathcal{L}(\theta, \mathcal{D}))$ must have for the above method to work?**

Finding $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ (I)

- The log(.) operation is monotonic, therefore

$$\arg\max_{\theta \in \Theta} \log(\mathcal{L}(\theta, \mathcal{D})) = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta, \mathcal{D})$$

- Can solve the optimization problem $\arg\max_\theta \mathcal{L}(\theta, \mathcal{D})$ by equating the first derivative of $\log(\mathcal{L}(\theta, \mathcal{D}))$ with respect to θ and equating to zero (first-order condition), i.e.

$$\frac{\partial}{\partial \theta} \log(\mathcal{L}(\theta, \mathcal{D})) = 0.$$

- **What properties $\log(\mathcal{L}(\theta, \mathcal{D}))$ must have for the above method to work?**
  - **Concavity and log-concavity!**

Finding $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ (II)

- Note that $\theta = (\mu, \sigma)$ is vector-valued: the first-order condition becomes

$$\bigtriangledown_\theta \log(\mathcal{L}(\theta, \mathcal{D})) = \mathbf{0} \rightarrow \left[\frac{\partial \log(\mathcal{L}(\theta, \mathcal{D}))}{\partial \mu} \; \frac{\partial \log(\mathcal{L}(\theta, \mathcal{D}))}{\partial \sigma} \right]^T = \mathbf{0}$$

- By taking the first derivative with respect to μ and σ, we have that:

$$\frac{\partial}{\partial \mu} \left(\frac{-n \log(2\pi\sigma^2)}{2} - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \right) = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial \sigma} \left(\frac{-n \log(2\pi\sigma^2)}{2} - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \right) = \frac{-n}{\sigma} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^3}$$

Finding $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ (III)

- The MLE estimators are:
  - **Sample Mean:**

$$\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma^2} = 0 \rightarrow \hat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  - **Sample Variance:**

$$\frac{-n}{\sigma} + \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^3} = 0 \rightarrow (\hat{\sigma}_n^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_n^*)^2$$

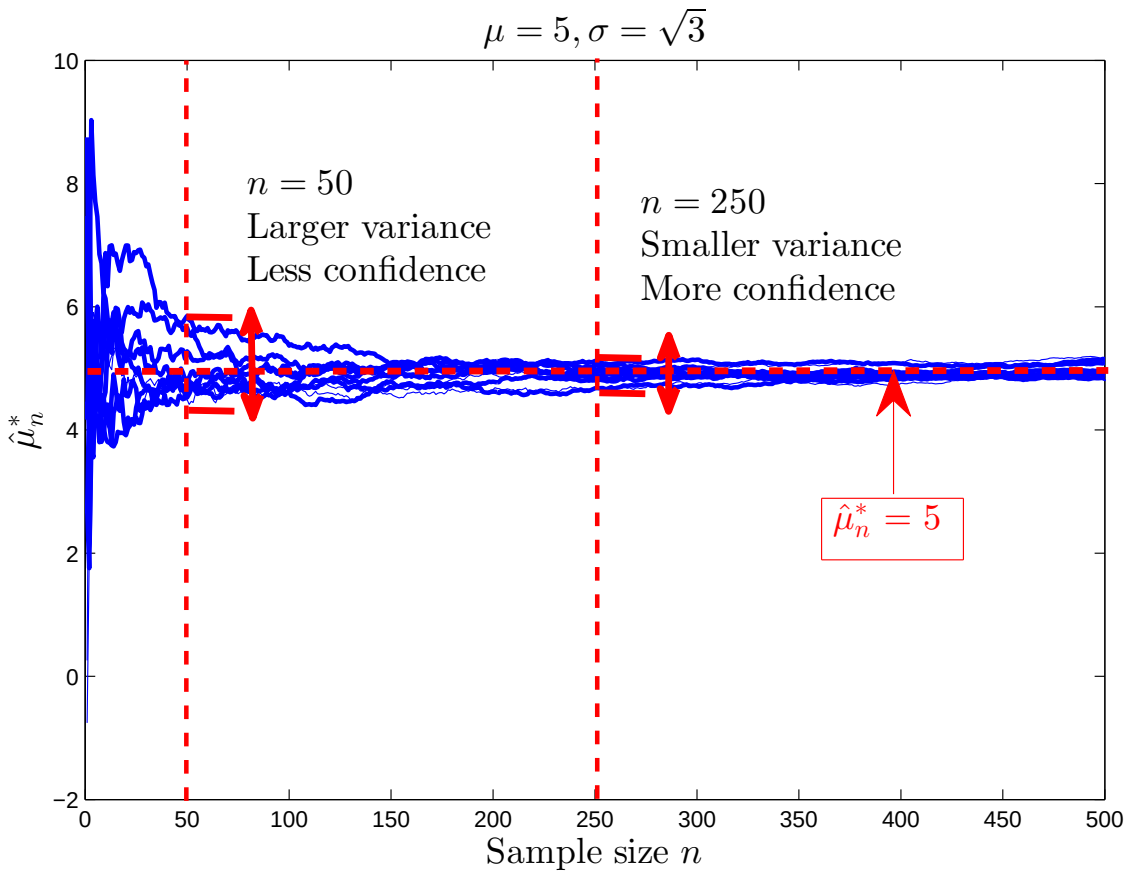
Finding $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$ (IV)

- **Exercise:** try to derive the MLE estimator when X is a *multivariate Gaussian* distribution:
  - The dataset $\mathcal{D} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, where $\mathbf{X}_i$ is M-dimensional, and $\mathbf{X}_i \sim \mathcal{N}(\mu, \mathbf{\Sigma})$.
  - The parameter space is $\Theta = \left\{(\mu, \mathbf{\Sigma}) : \mu \in \mathbb{R}^M, \mathbf{\Sigma} \succeq 0\right\}$
  - The multivariate Gaussian distribution is

$$\mathbb{P}_\theta(\mathbf{X}) = (2\pi)^{\frac{-M}{2}} \cdot |\mathbf{\Sigma}|^{\frac{-1}{2}} \cdot \exp\left(\frac{-1}{2}(\mathbf{X} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X} - \mu)\right)$$

What is our confidence in the estimates $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$?

- Depends on the sample size n: the larger n, the smaller the variances of $\hat{\mu}_n^*$ and $\hat{\sigma}_n^*$.



Confidence Sets

- Point estimators provide no quantification for uncertainty $\rightarrow$ need to introduce a measure of *confidence* in an estimate!

Confidence Sets

A $(1-\alpha)$ *confidence set* for a parameter θ is a subset of the parameter space, $\tilde{\Theta}(X_1, \ldots, X_n) \subset \Theta$, such that

$$\mathbb{P}(\theta \in \tilde{\Theta}) \geq 1 - \alpha.$$

- Confidence intervals are one-dimensional confidence sets.
- Because of the **asymptotic normality** of the general MLE estimates, we can compute **asymptotic confidence intervals**.
- **Normality** $\rightarrow$ compute confidence intervals.
- **Asymptotic normality** $\rightarrow$ compute asymptotic confidence intervals.

Example: Unknown Mean and Known Variance (I)

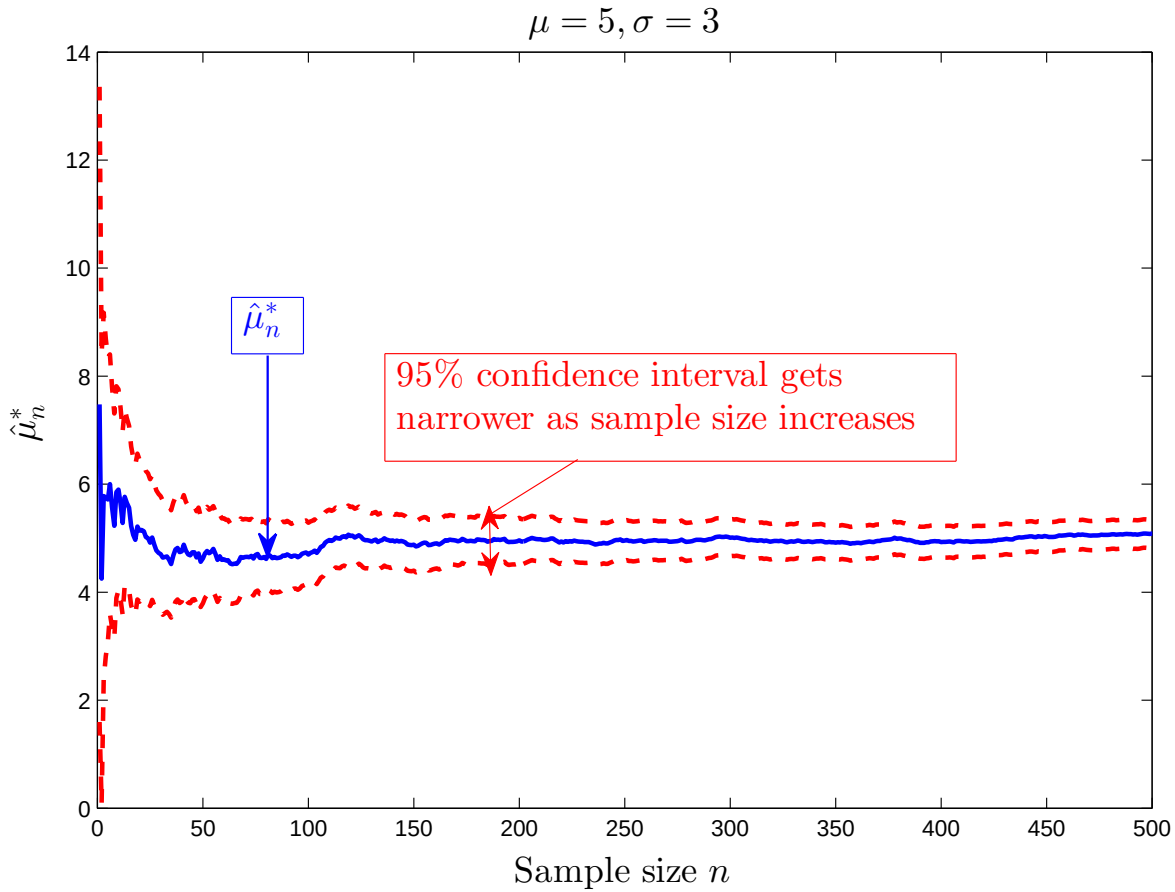
- Assume we know σ and want to estimate μ from $\mathcal{D}$
- The MLE estimate for μ is $\hat{\mu}_n^* = \frac{1}{n}\sum_{i=1}^n X_i$
- We know that $\frac{\sqrt{n}}{\sigma}(\hat{\mu}_n^* - \mu) \sim \mathcal{N}(0,1)$ **(central limit theorem)**
- We want to compute the confidence interval $\tilde{\Theta} = [\hat{\mu}_n^* - \tilde{\mu}, \hat{\mu}_n^* + \tilde{\mu}]$ **(symmetric normal distribution)**

Computing the Confidence Interval for μ

- Find γ for which $\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}(\hat{\mu}_n^* - \mu) \geq \gamma\right) = \frac{\alpha}{2} \rightarrow \gamma = Q^{-1}(\frac{\alpha}{2})$, where $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(\frac{-u^2}{2}\right) du$.
- $\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}|\hat{\mu}_n^* - \mu| \leq Q^{-1}(\frac{\alpha}{2})\right) = 1 - \alpha \iff \tilde{\mu} = \frac{\sigma}{\sqrt{n}}Q^{-1}(\frac{\alpha}{2})$

Example: Unknown Mean and Known Variance (II)

- The 95% confidence interval for the MLE mean estimate is $\tilde{\Theta} = [\hat{\mu}_n^* - 1.96\frac{\sigma}{\sqrt{n}}, \hat{\mu}_n^* + 1.96\frac{\sigma}{\sqrt{n}}]$



The Categorical Family $\mathcal{P}$ and the Parameter Space Θ

- Each data point takes a value from a finite set of values: $X_i \in \{1, 2, \ldots, r\}$.
- The probability that $X_i = j$ is given by $p_j \in [0, 1]$.
- The parameter of a categorical distribution is $\theta = (p_1, \ldots, p_r)$.
- The parameter space is the *simplex*
  $\Theta = \left\{(p_1, \ldots, p_r) : (p_1, \ldots, p_r) \in [0, 1]^r, \sum_{j=1}^r p_j = 1\right\}$.
- The probability mass function of the dataset $\mathcal{D}$:

$$\mathbb{P}_\theta(X_1, \ldots, X_n) = \prod_{i=1}^n \prod_{j=1}^r p_j^{\mathbf{1}_{\{x_i = j\}}}$$

$$= p_1^{n_1} \cdot p_2^{n_2} \ldots p_r^{n_r}, \; n_j = \sum_{i=1}^n \mathbf{1}_{\{x_i = j\}}.$$

Finding $\hat{\theta}_n^*$ (I)

- The log-likelihood function: $\log(\mathcal{L}(\theta, \mathcal{D})) = \sum_{j=1}^{r} n_j \cdot \log(p_j)$.
- The MLE estimate $\hat{\theta}_n^*$ is

$$\hat{\theta}_n^* = \arg\max_{p_1,\ldots,p_r} \sum_{j=1}^{r} n_j \log(p_j)$$

$$s.t. \sum_{j=1}^{r} p_j = 1.$$

Constrained optimization: **Not as easy as in the Gaussian case!**

Finding $\hat{\theta}_n^*$ (II): Method A

The Method of Lagrange Multipliers

- Maximize $\sum_{j=1}^{r} n_j \log(p_j) - \lambda\left(\sum_{j=1}^{r} p_j - 1\right)$ via the first-order condition.
- $\bigtriangledown_\theta \log(\mathcal{L}(\theta, \mathcal{D})) = 0 \rightarrow p_{j,n}^* = \lambda^{-1} n_j$.
- Since $\sum n_j = n$ and $\sum p_{j,n}^* = 1$, then $\lambda = n$, and

$$\hat{\theta}_n^* = \left[\frac{n_1}{n} \ldots \frac{n_r}{n}\right]^T.$$

- The MLE $\hat{\theta}_n^*$ is the empirical distribution function, which uniformly converges to the true probability mass function.
- This matches our expectations regarding the consistency of $\hat{\theta}_n^*$.

Finding $\hat{\theta}_n^*$ (III): Method B

An Information-Theoretic Approach

- We can reformulate the MLE optimization problem as follows

$$\begin{aligned}\hat{\theta}_n^* &= \arg\max_{p_1,\ldots,p_r} \sum_{j=1}^{r} n_j \log(p_j) = \arg\max_{p_1,\ldots,p_r} \sum_{j=1}^{r} \frac{n_j}{n} \log(p_j) \\ &= \arg\max_{p_1,\ldots,p_r} \sum_{j=1}^{r} q_j \log\left(\frac{p_j}{q_j}\right) + \sum_{j=1}^{r} q_j \log(q_j), \quad q_j = \frac{n_j}{n}. \\ &= \arg\max_{p_1,\ldots,p_r} \underbrace{\sum_{j=1}^{r} q_j \log\left(\frac{p_j}{q_j}\right)}_{\text{KL divergence}} = \arg\min_{\mathbf{p}} D(\mathbf{q}||\mathbf{p})\end{aligned}$$

- Since $D(\mathbf{q}||\mathbf{p}) \geq 0$, then the minimum is achieved at $\mathbf{q} = \mathbf{p}$, and we have $p_{j,n}^* = q_j = \frac{n_j}{n}$.

Confidence Intervals

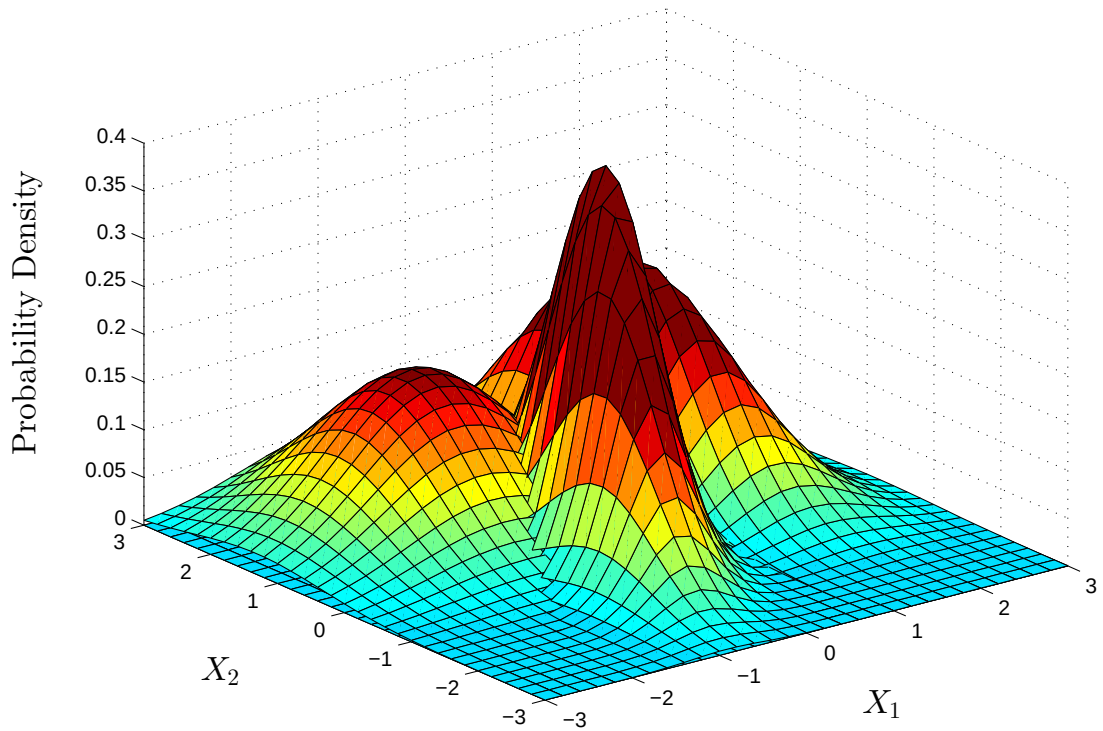
- Unlike the case of the Gaussian distribution, the MLE estimators are not normally distributed for any n.
- For large n, we can construct asymptotic confidence intervals using asymptotic normality.
- For arbitrary n, consider the case when $r = 2$ (Bernoulli distribution), we have $\theta = (p_1, p_2 = 1 - p_1)$. A **conservative** $(1 - \alpha)$ confidence interval for the parameter p_1 is

$$\tilde{\Theta} = \left[\frac{n_1}{n} - \frac{Q^{-1}(\alpha/2)}{2\sqrt{n}}, \frac{n_1}{n} + \frac{Q^{-1}(\alpha/2)}{2\sqrt{n}} \right]$$

- Derivation follows from the central limit theorem and bounding the variance of a Bernoulli random variable by $\sqrt{p_1(1 - p_1)} \leq \frac{1}{2}$.

Heterogeneous Populations

- In many applications, the data is sampled from K different populations with different parameters.
- **Example: Gaussian mixture with 3 components.**



Gaussian Mixture Models

- A K-component (univariate) Gaussian mixture model has the following parameters
  - The Gaussian parameters $\theta_k = (\mu_k, \sigma_k), 1 \leq k \leq K$.
  - The mixing proportion $\pi_k, \sum_{i=1}^{K} \pi_k = 1$.
- The dataset $\mathcal{D}$ has the following structure:

$$\mathcal{D} = ((X_1, Z_1), \ldots, (X_n, Z_n)).$$

  - Each variable X_i is drawn from a Gaussian distribution

$$X_i \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

  - The variable Z_i represent the membership of the i^{th} data point to one of the K components, and is drawn from a categorical distribution

$$Z_i \sim \text{Categorical}(\pi_1, \ldots, \pi_K).$$

Complete-Data and Marginal Likelihood Functions

Two possible scenarios:

- **If Z_i is observed:** then the *complete-data likelihood* function is **uni-modal**:

$$\mathcal{L}(\theta, \mathcal{D}) = \prod_{i=1}^{n} \pi_{z_i} \cdot \frac{1}{\sqrt{2\pi}\,\sigma_{z_i}} \cdot e^{-\frac{(x_i - \mu_{z_i})^2}{2\sigma_{z_i}^2}}$$

- **If Z_i is latent:** then the *marginal likelihood* function is **multi-modal**:

$$\mathcal{L}(\theta, \mathcal{D}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \cdot \frac{1}{\sqrt{2\pi}\,\sigma_k} \cdot e^{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}}$$

Hard problem $\rightarrow$ approximate solution using the EM algorithm (next lecture)!

Maximum Likelihood is a Frequentist Methods

A frequentist method...

1. Never uses or gives the probability of a hypothesis (no prior or posterior).
2. Depends on the likelihood for both observed and unobserved data.
3. Does not require a prior.
4. Tends to be less computationally intensive.

Frequentist measures such as p-values and confidence intervals are often used in current research practices since the 20^{th} century.

Bayesian Methods

On the other hand, a Bayesian method...

1. Assumes a prior: uses probabilities for both hypotheses and data.
2. Depends on the prior and likelihood of observed data.
3. Requires one to know or construct a subjective prior.
4. May be computationally intensive due to integration over many parameters.

Many recent advances in Bayesian methods! Read about: variational Bayesian methods, Markov Chain Monte Carlo methods, etc.

References

1. Robert W Keener, "Statistical theory: notes for a course in theoretical statistics," 2006.
2. Robert W Keener, "Theoretical Statistics: Topics for a Core Course," 2010.
3. Christopher Bishop, "Pattern Recognition and Machine Learning," 2007.