Machine Learning for Sensory Signals

Gaussian and Mixture Gaussian Models

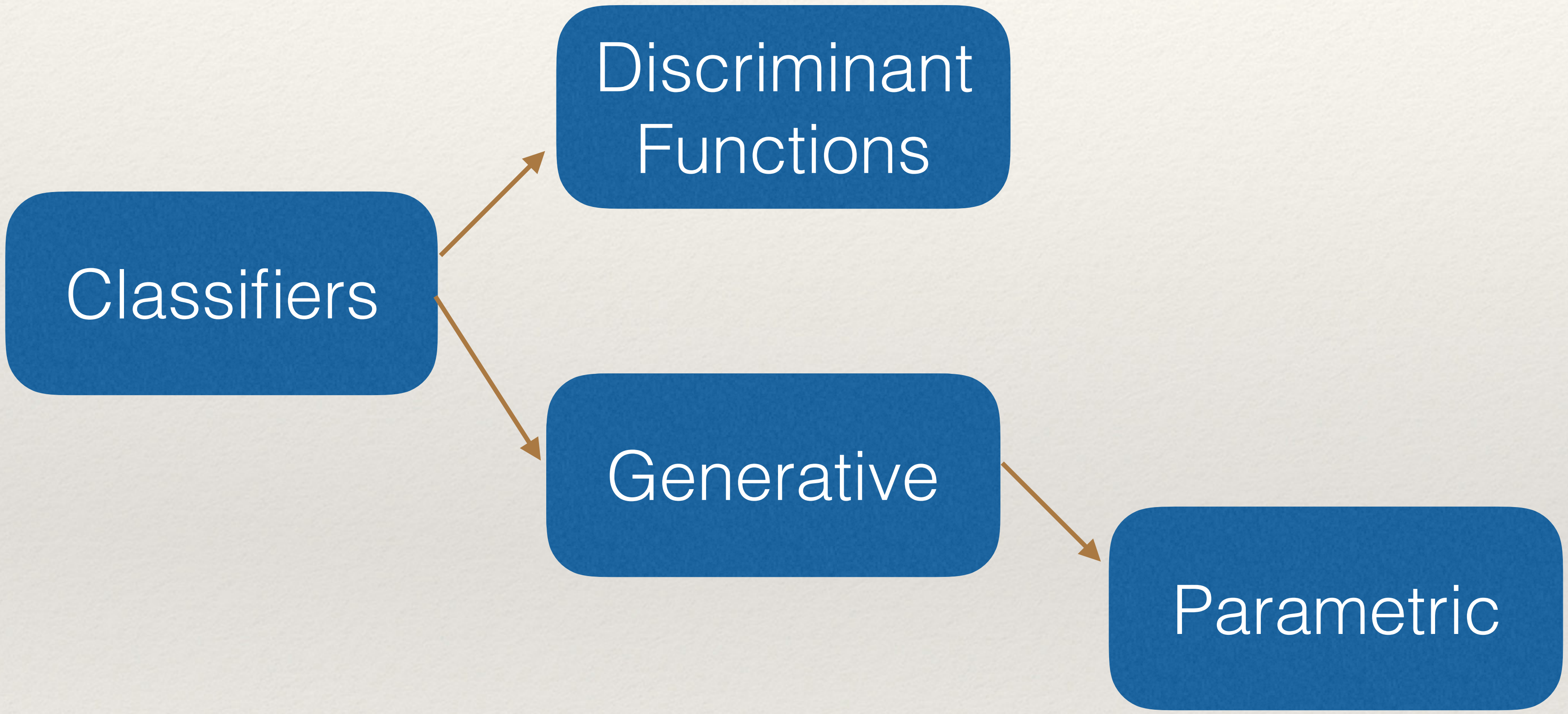
23-02-2017

Established 1911

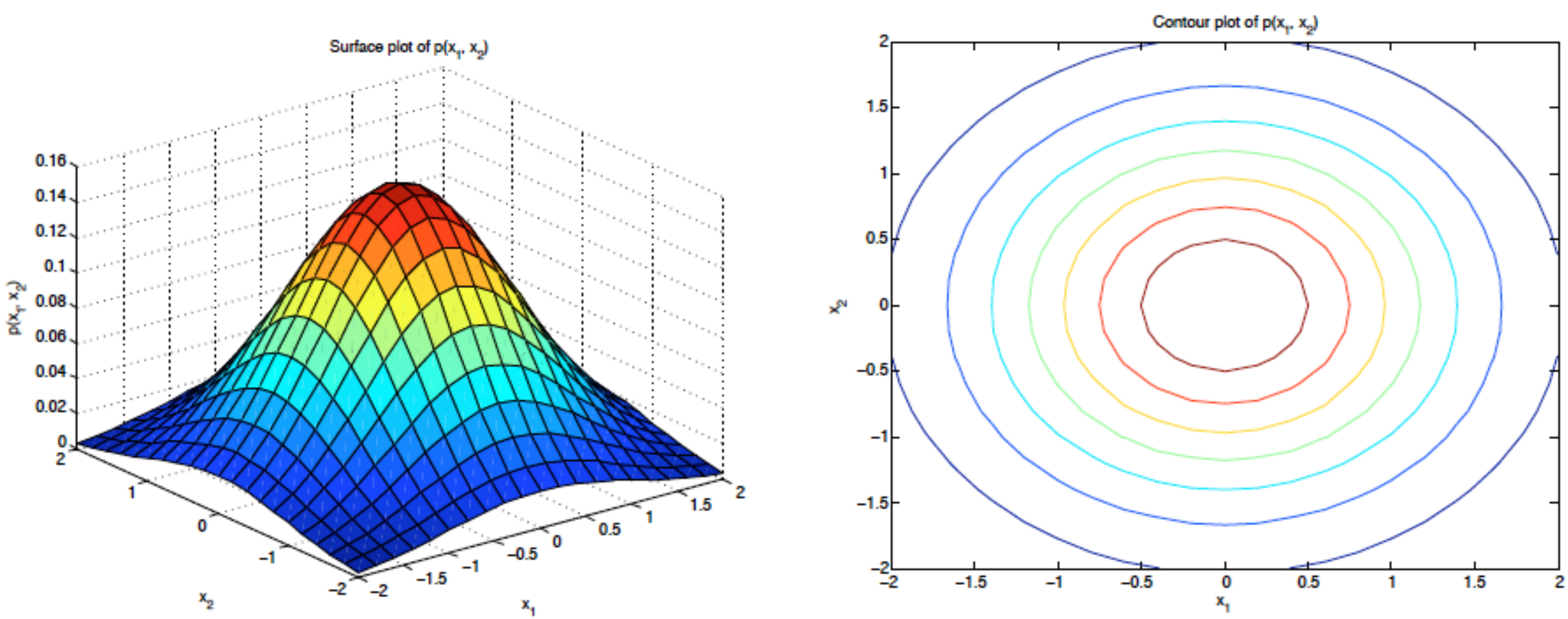
INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Classifier Types

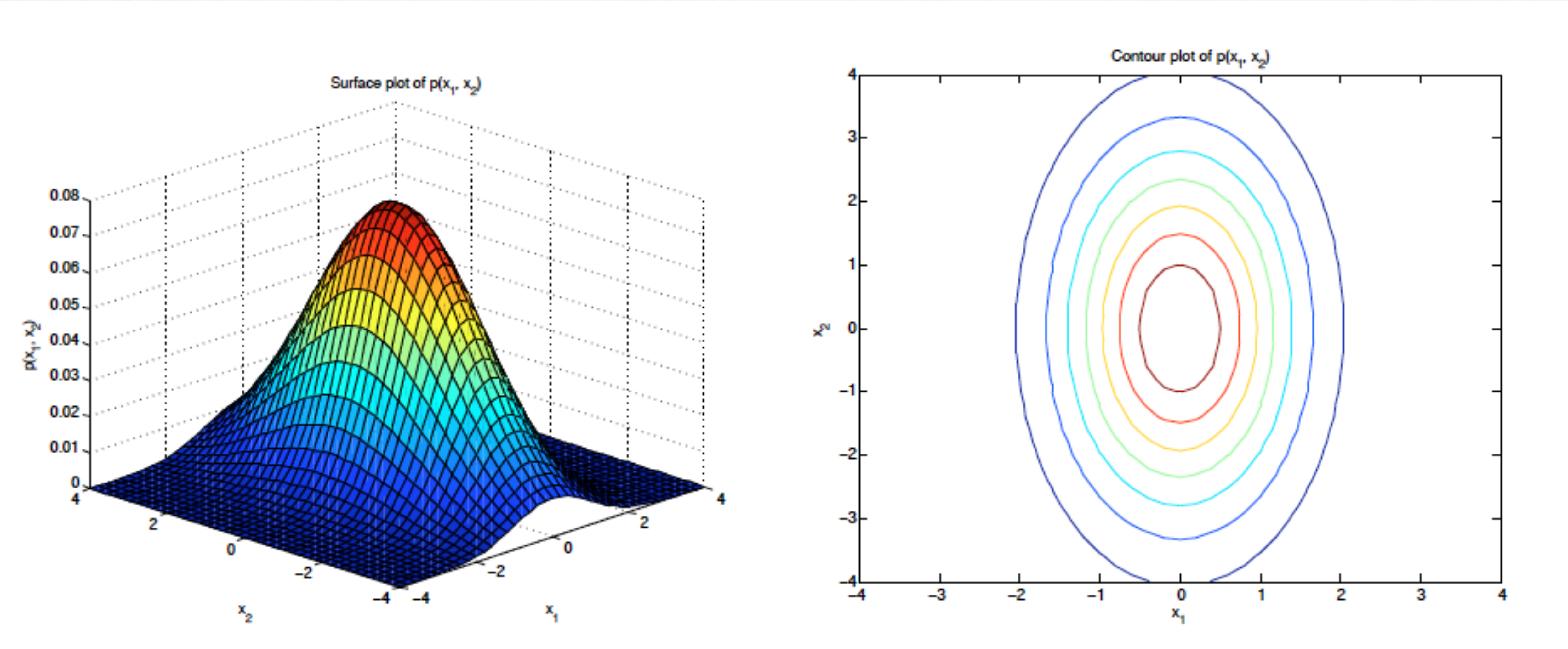


Gaussian Distribution



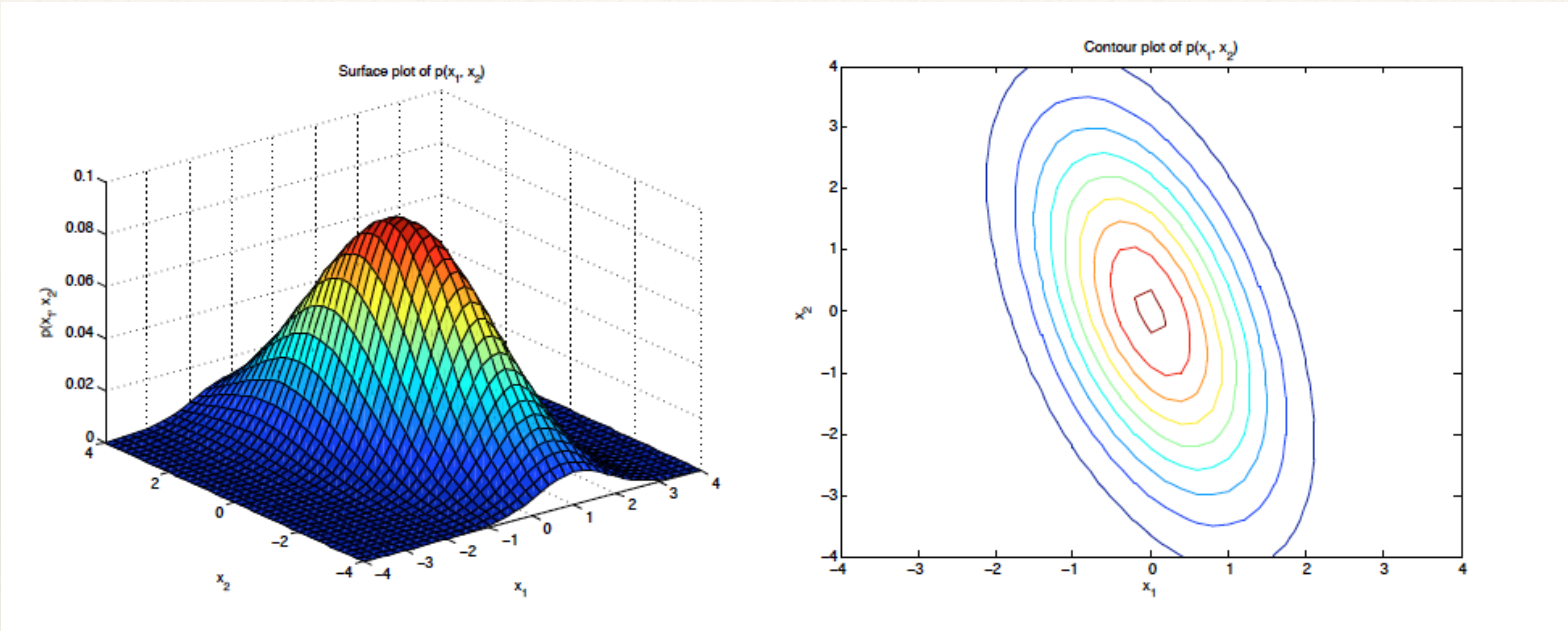
Points of equal probability lie on on contour
Diagonal Gaussian with Identical Variance

Gaussian Distribution



Diagonal Gaussian with different variance

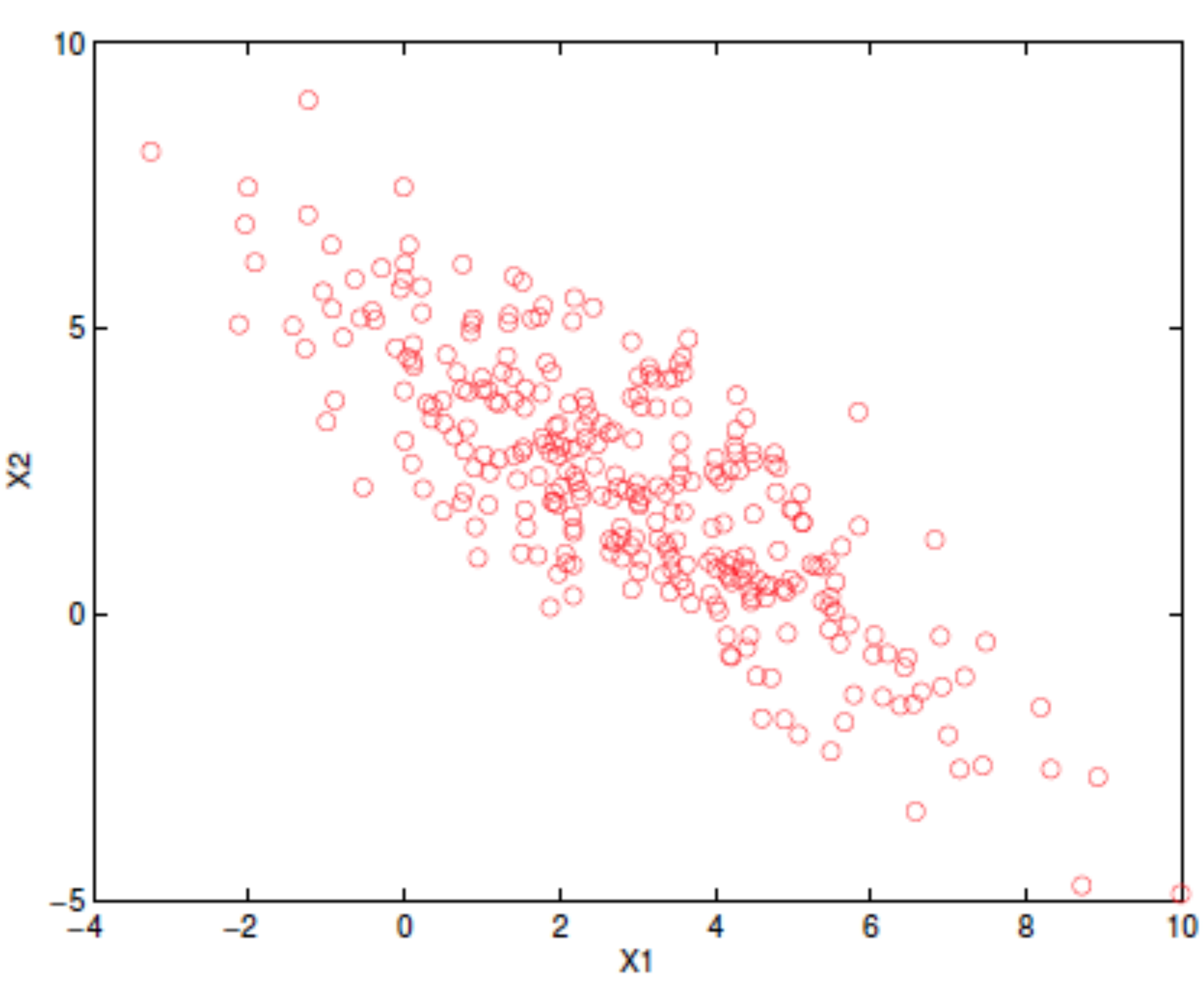
Gaussian Distribution

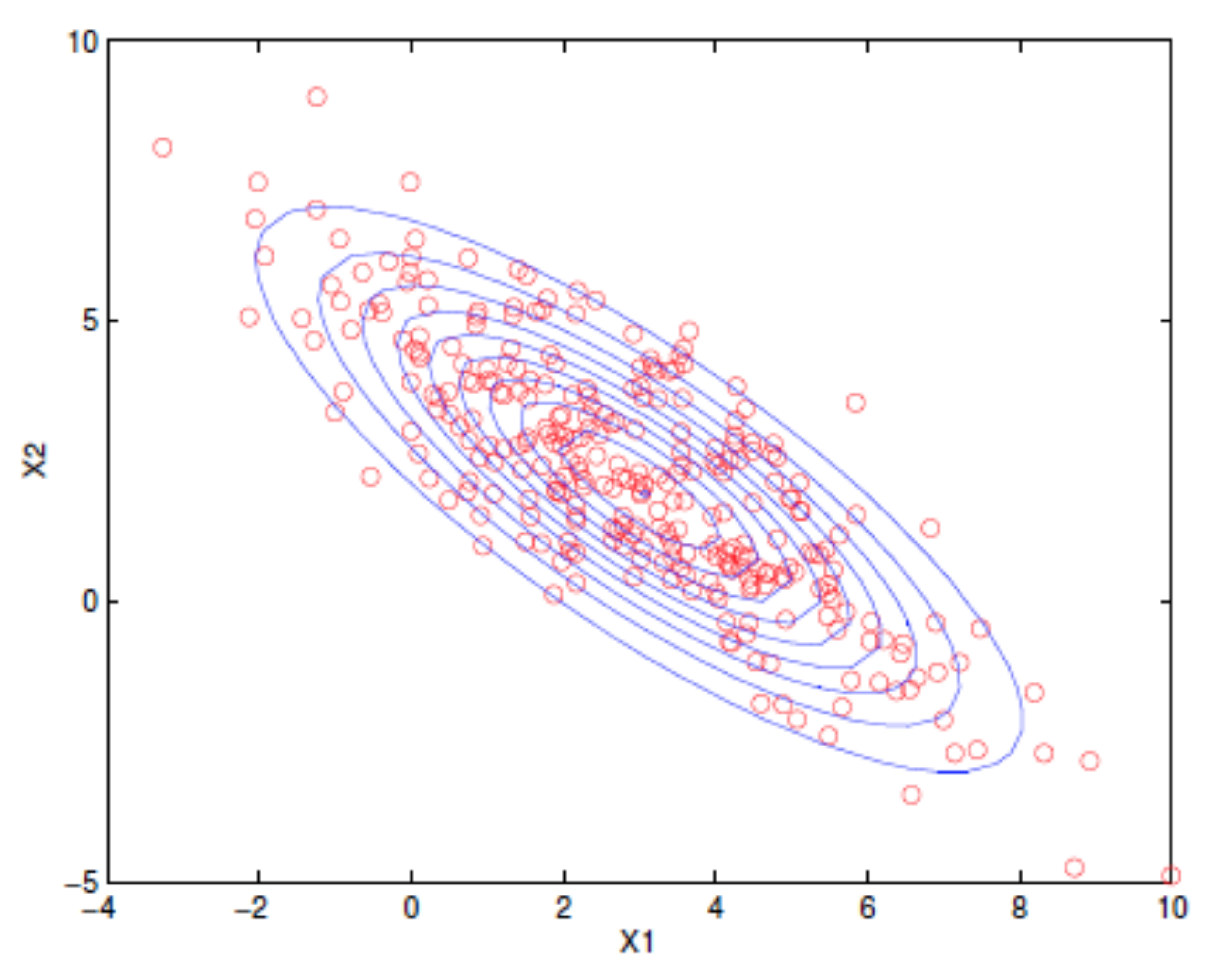


Full covariance Gaussian distribution

Gaussian Distribution

Fitting the data with a Gaussian Model





Finding the parameters of the Model

- The Gaussian model has the following parameters

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Total number of parameters to be learned for D dimensional data is $D^2 + D$
- Given N data points $\{\mathbf{x}_i\}_{i=1}^{N}$ how do we estimate the parameters of model.
  - Several criteria can be used
  - The most popular method is the maximum likelihood estimation (MLE).

MLE

Define the likelihood function as $L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta})$

The maximum likelihood estimator (MLE) is

$$\boldsymbol{\theta}^* = arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

The MLE satisfies nice properties like

- Consistency (convergence to true value)
- Efficiency (has the least Mean squared error).

MLE

For the Gaussian distribution

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^* \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \right\}$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta})$$

$$\log L(\boldsymbol{\theta}) = -\frac{ND}{2} - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{N}\Big((\mathbf{x}_i-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\Big)$$

To estimate the parameters $\dfrac{\partial \log L}{\partial \boldsymbol{\mu}} = 0$

Established 1911

MLE

Using matrix differentiation rules, for a symmetric matrix $\mathbf{A}$

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x} \qquad \boldsymbol{\mu}^* = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

Using matrix differentiation rules for log determinant and trace

$$\frac{\partial \log(|\mathbf{A}|)}{\partial \mathbf{A}} = 2\mathbf{A}^{-1} - diag(\mathbf{A}^{-1})$$

$$\frac{\partial tr(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B} + \mathbf{B}^T - diag(\mathbf{B})$$

$$\boldsymbol{\Sigma}^* = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu}^*)(\mathbf{x}_i - \boldsymbol{\mu}^*)^T$$

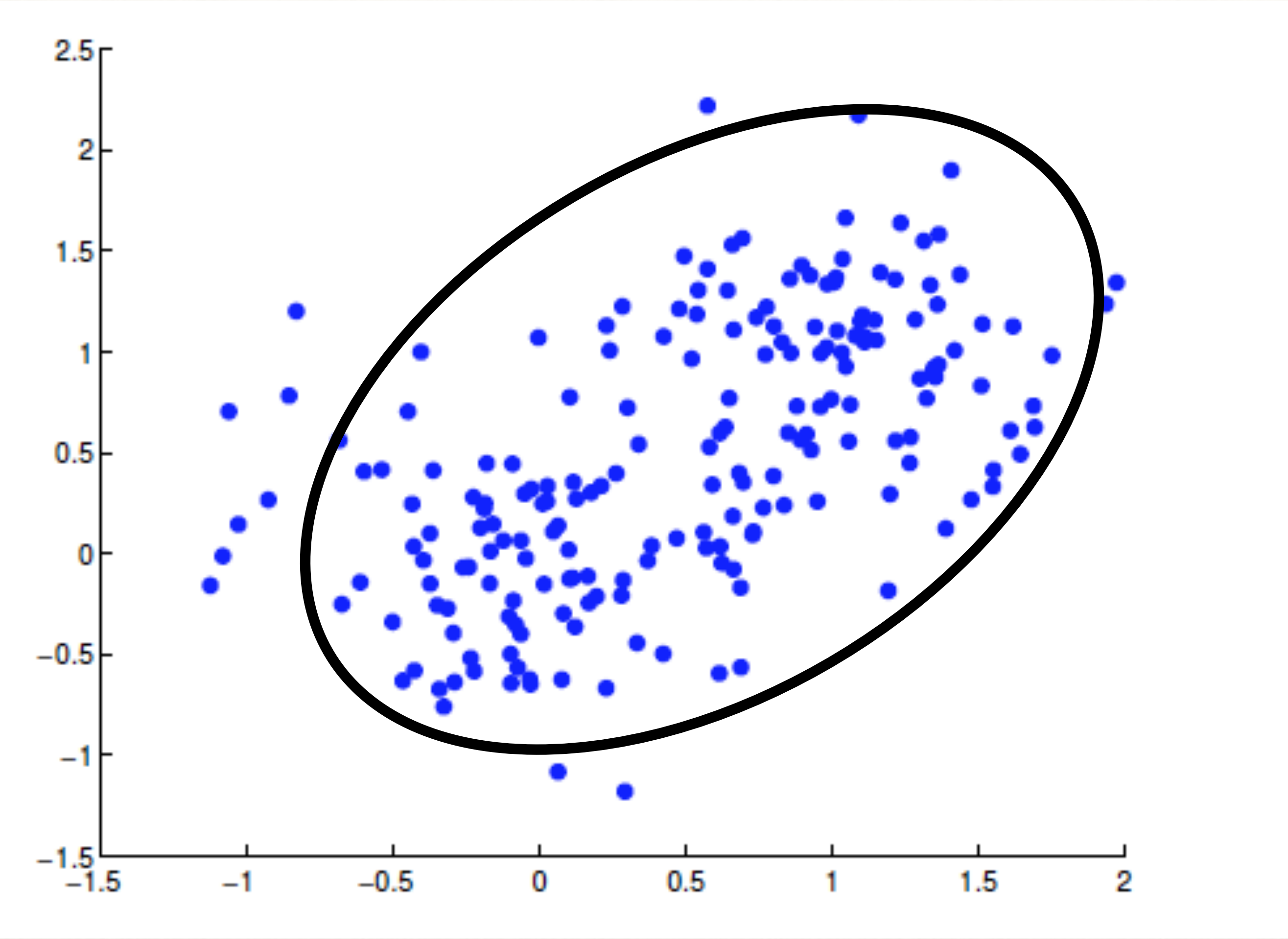
Established 1911

Sample mean and Sample Covariance

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Issues with Gaussian Distribution

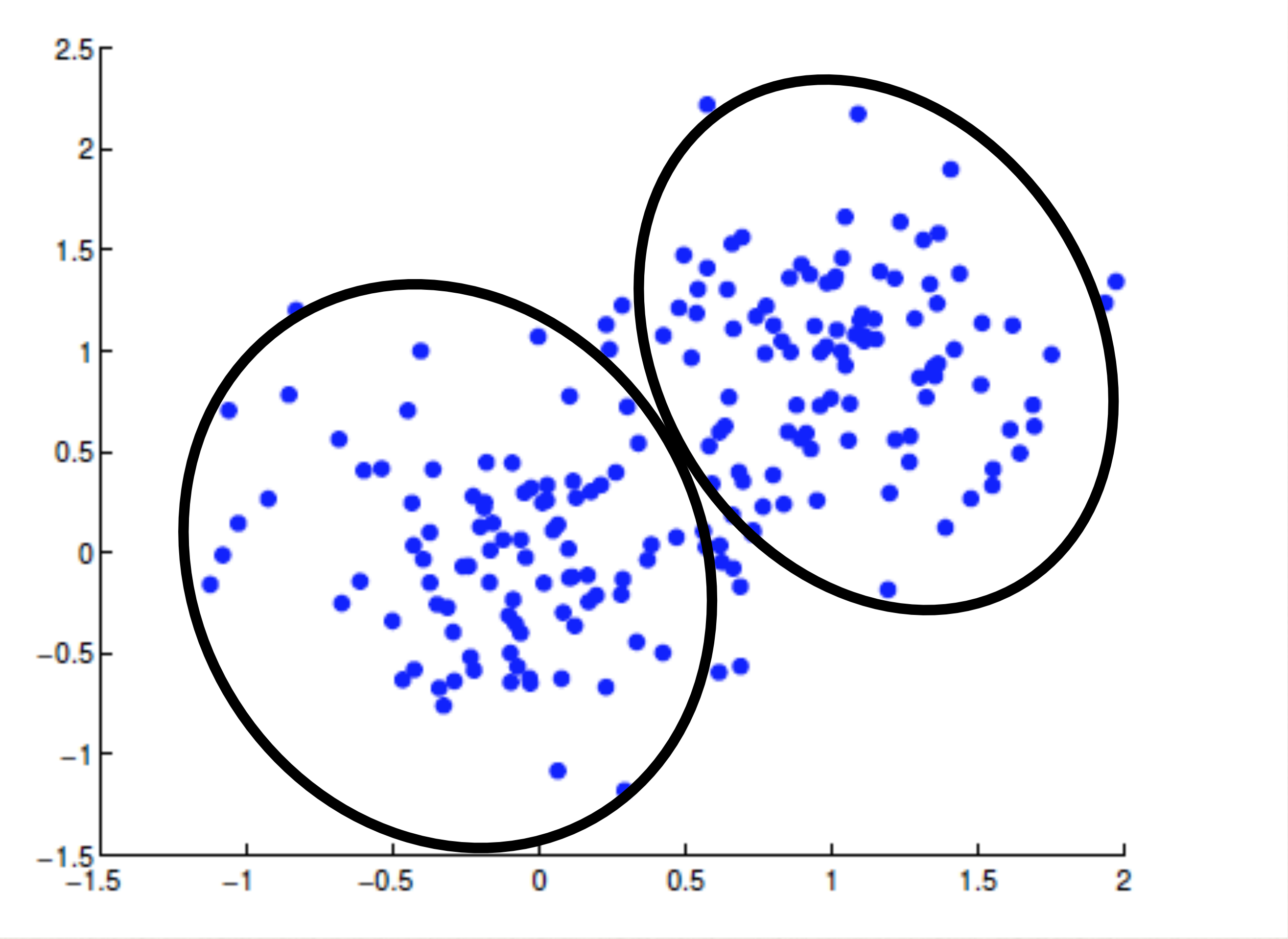
Often the data lies in clusters (2-D example)



Fitting a single Gaussian model may be too broad.

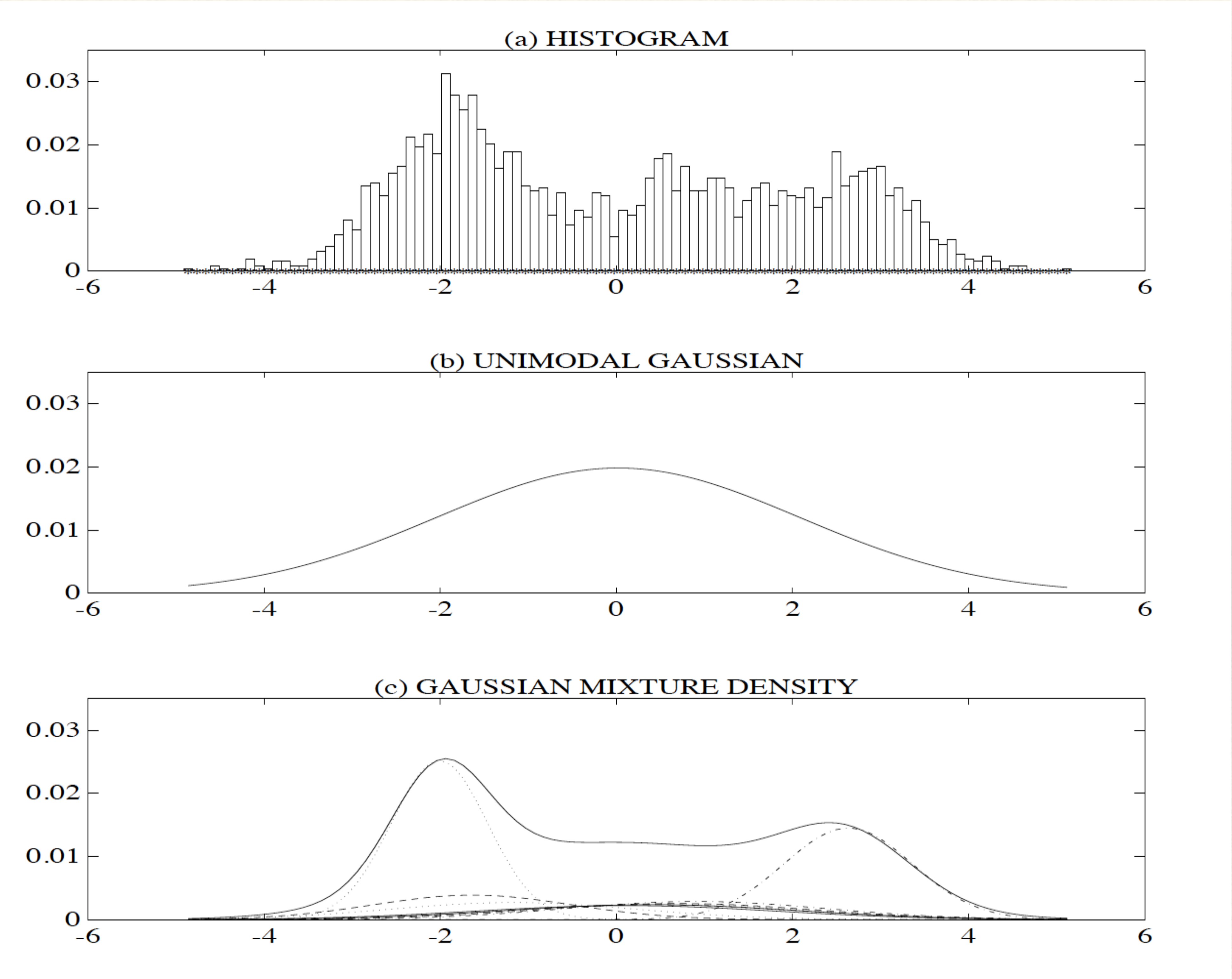
Issues with Gaussian Distribution

Need mixture models



Can fit any arbitrary distribution.

Issues with Gaussian Distribution



Summary

- The Gaussian model - parametric distributions
- Simple and useful properties.
- Can model unimodal (single peak distributions)
- MLE gives intuitive results
- Issues with Gaussian model
  - Multi-modal data
  - Not useful for complex data distributions
- Need for mixture models

Gaussian Mixture Models

A Gaussian Mixture Model (GMM) is defined as

$$p(\mathbf{x}|\mathbf{\Theta}) = \sum_{k=1}^{K} \alpha_k p(\mathbf{x}|\boldsymbol{\theta}_k)$$

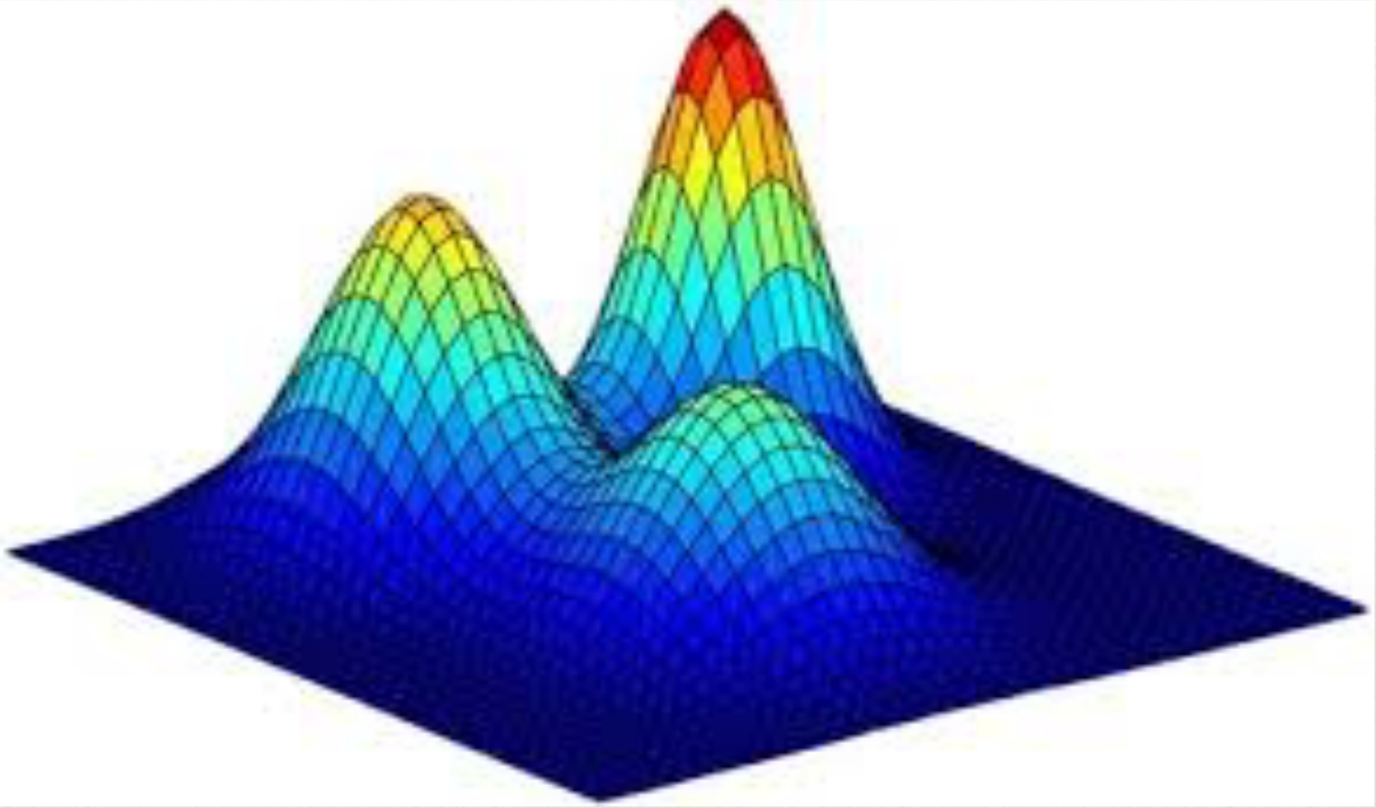
$$p(\mathbf{x}|\boldsymbol{\theta}_k) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{\Sigma}_k|}} exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^* \mathbf{\Sigma}_k^{-1} (\mathbf{x}-\boldsymbol{\mu}_k) \right\}$$

The weighting coefficients have the property

$$\sum_{k=1}^{K} \alpha_k = 1$$

Gaussian Mixture Modeling

- Properties of GMM
  - Can model multi-modal data.
  - Identify data clusters.
  - Can model arbitrarily complex data distributions



The set of parameters for the model are

$$\boldsymbol{\Theta}_k = \{\alpha_k, \boldsymbol{\theta}_k\}_{k=1}^{K} \qquad \boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$$

- The log-likelihood function over the entire data in this case will have a logarithm of a summation

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \alpha_k p(\mathbf{x}_i | \boldsymbol{\theta}_k)\right)$$

- Solving for the optimal parameters using MLE for GMM is not straight forward.

- Resort to the Expectation Maximization (EM) algorithm

MLE for GMM