MACHINE LEARNING FOR SIGNAL PROCESSING

20-1-2025

Sriram Ganapathy
LEAP lab, Electrical Engineering, Indian Institute of Science,
sriramg@iisc.ac.in

Viveka Salinamakki, Varada R.
LEAP lab, Electrical Engineering, Indian Institute of Science

http://leap.ee.iisc.ac.in/sriram/teaching/MLSP25/

Decision Theory (PRML Chap. 1.5)

- Decision Theory
  - Inference problem
    - Finding the joint density
  - Decision problem $p(\mathbf{x}, \mathbf{t})$
    - Using the inference to make the classification or regression decision

DECISION PROBLEM - CLASSIFICATION

- ✓ Minimizing the mis-classification error
- ✓ Decision based on maximum posteriors

$$argmax_j \; p(C_j|\mathbf{x})$$

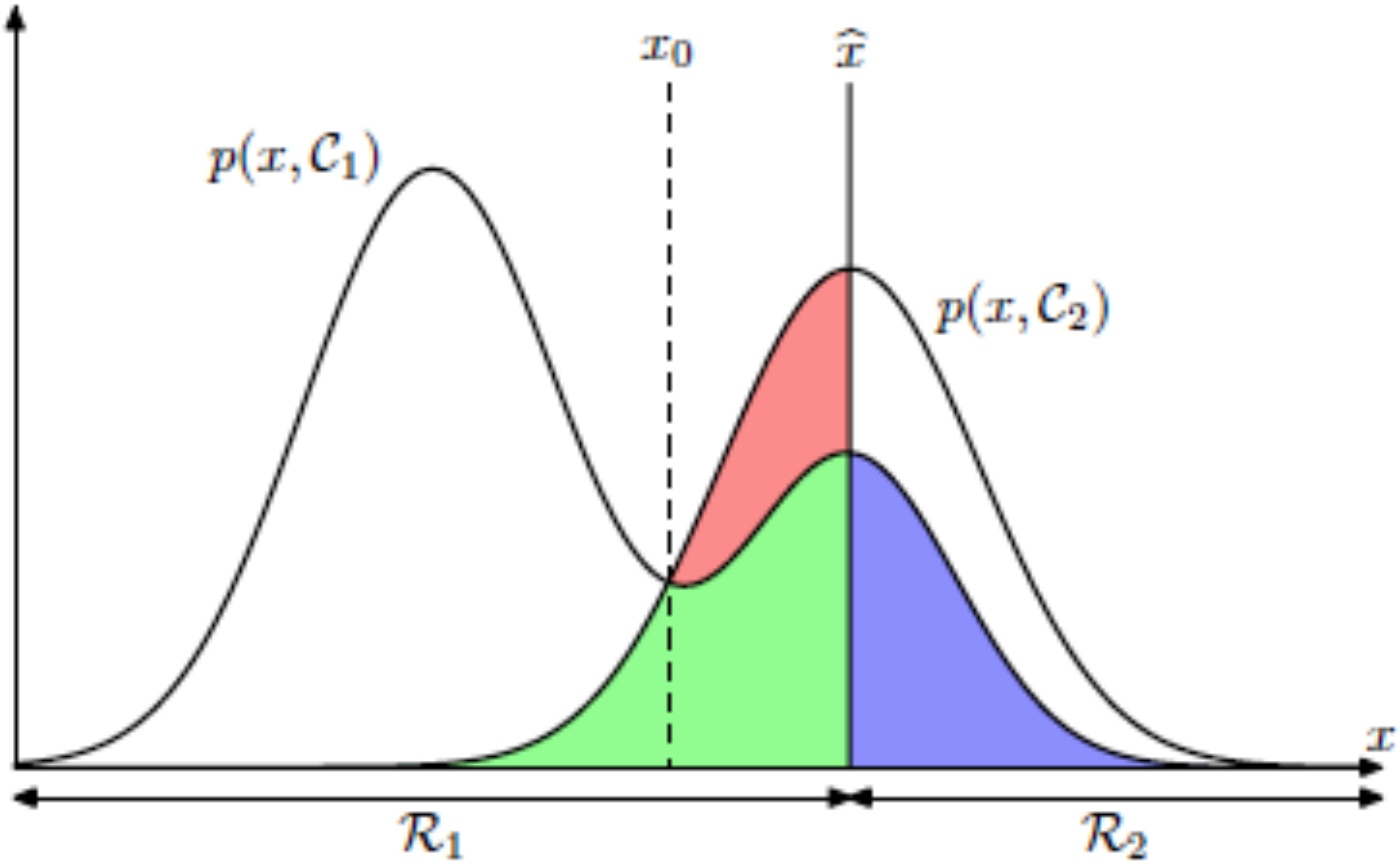
- ✓ Loss matrix
  - Minimizing the expected loss

$$argmax_j \; \sum_k L_{k,j} p(C_k|\mathbf{x})$$

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

LEAP

VISUALIZING THE MAX. POSTERIOR CLASSIFIER



APPROACHES FOR INFERENCE AND DECISION

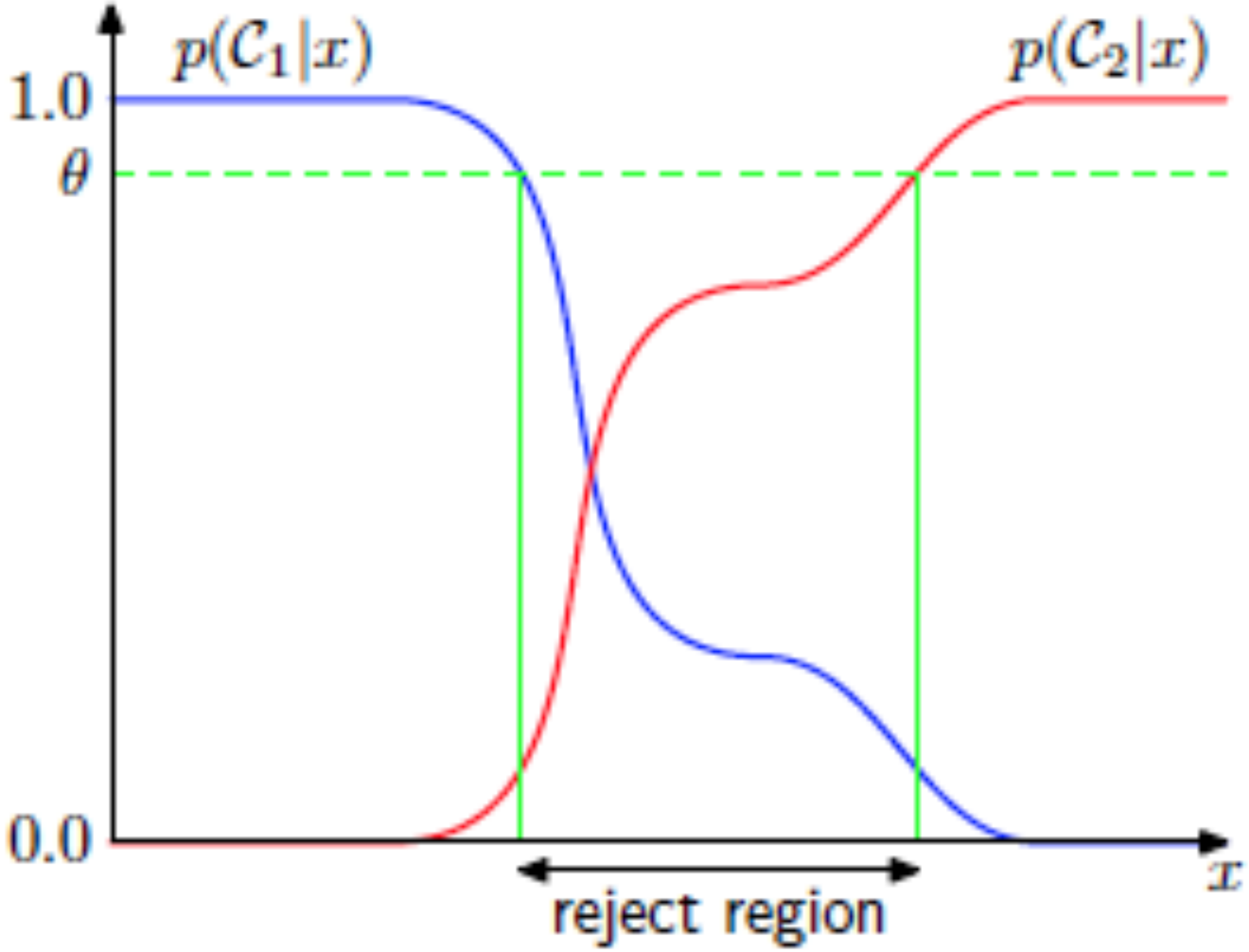
I. Finding the joint density from the data.

II. Finding the posteriors directly.

$$p(C_k|\mathbf{x}) \propto p(\mathbf{x}|C_k)p(C_k)$$

III. Using discriminant functions for classification.

ADVANTAGE OF POSTERIORS

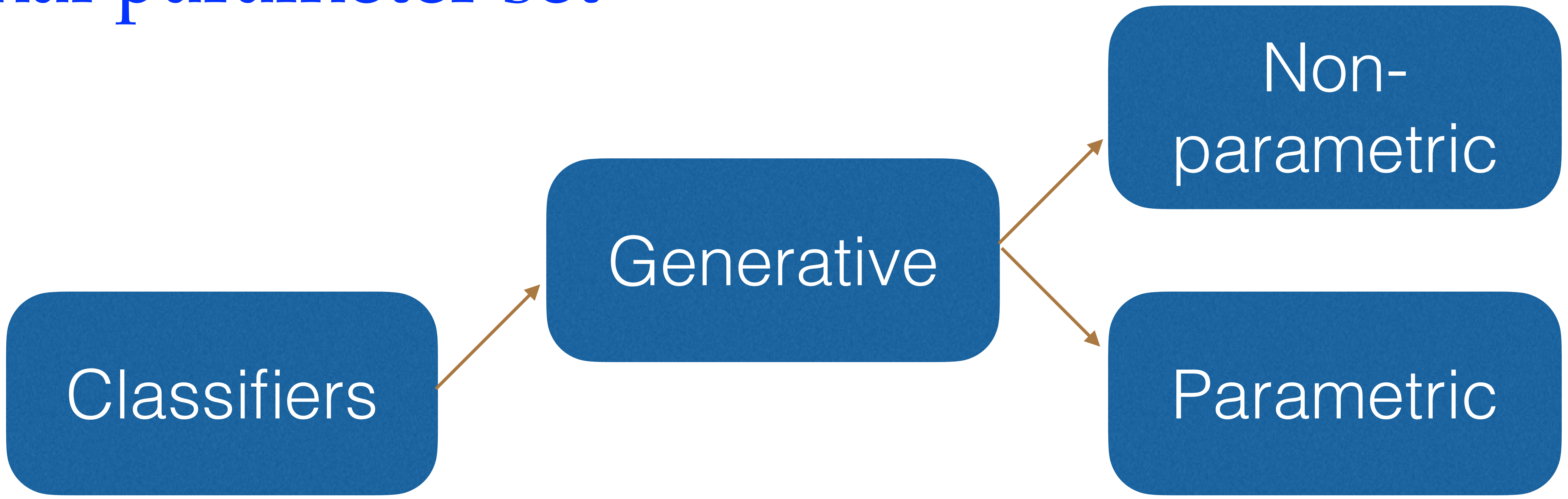


Decision Rule for Regression

- Minimum mean square error loss
- Solution is conditional expectation.

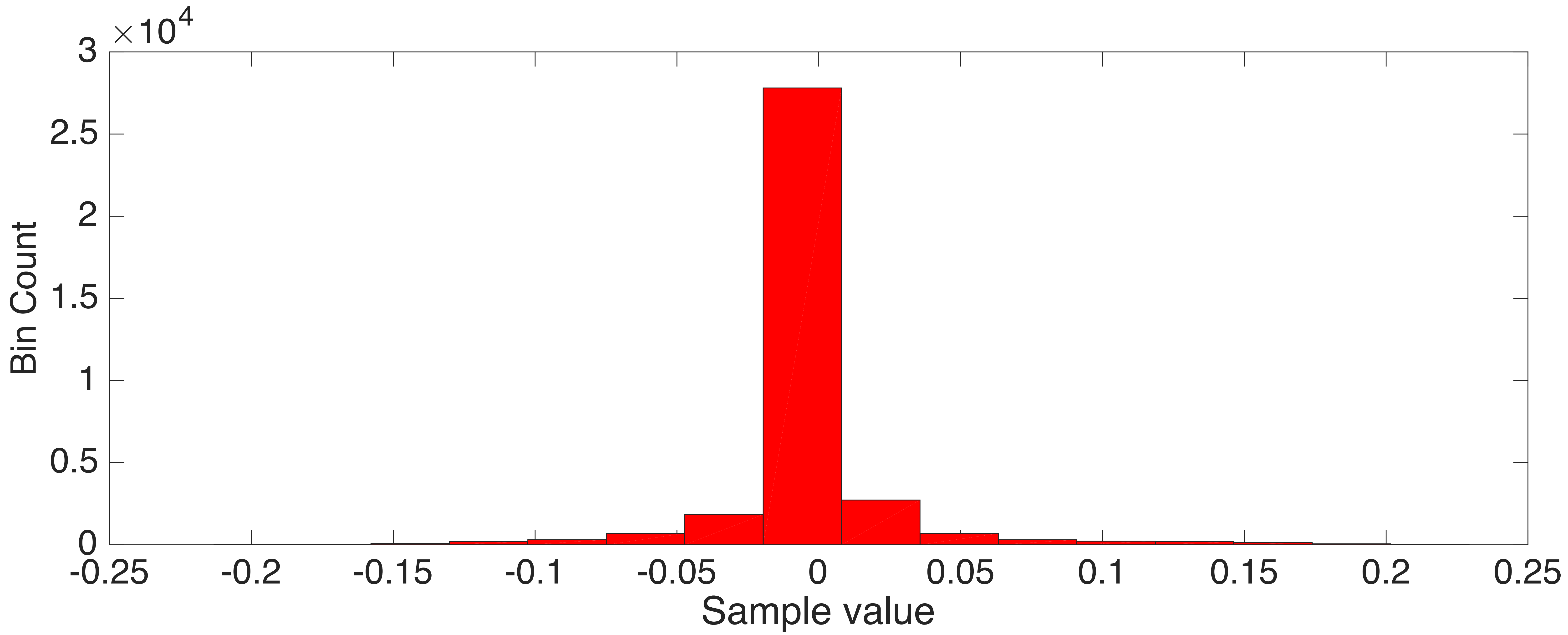
Generative Modeling

- Collection of probability distributions which are described by a finite dimensional parameter set



Non-parametric Modeling

- Non-parametric models do not specify an apriori set of parameters to model the distribution. Example - Histogram



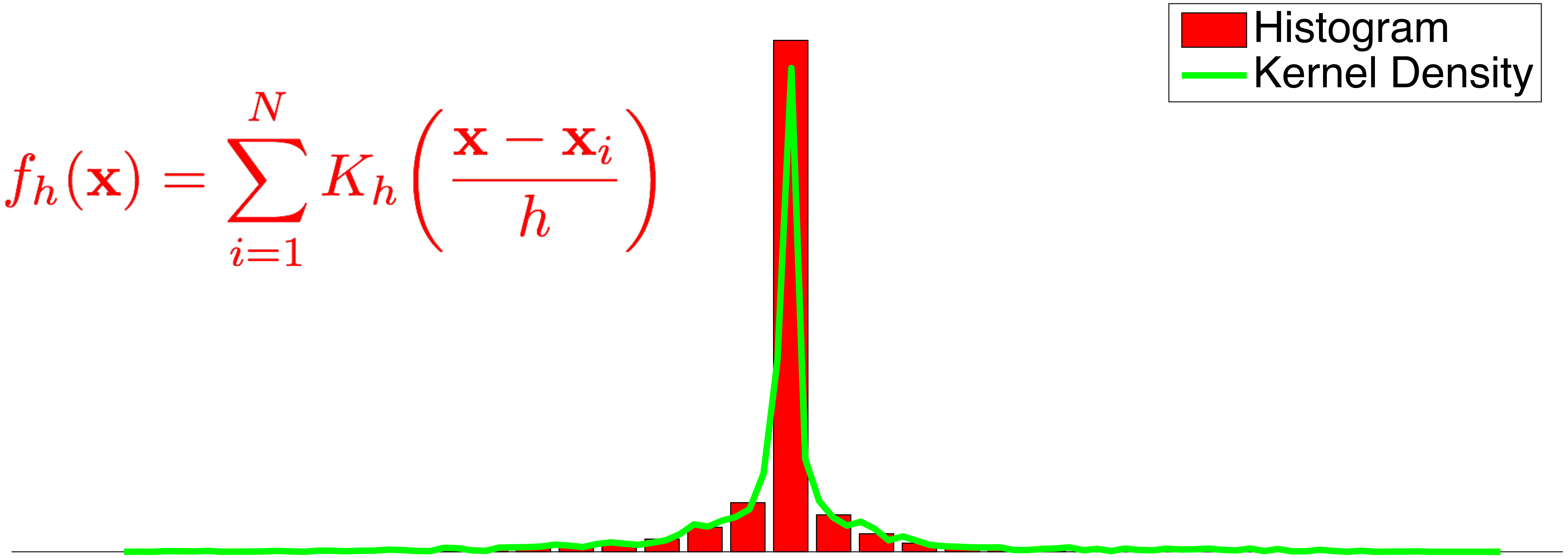
The density is not smooth and has block like shape.

Non-parametric Modeling

- Non-parametric models do not specify an apriori set of parameters to model the distribution.

  - Example - Kernel Density Estimators

$$f_h(\mathbf{x}) = \sum_{i=1}^{N} K_h\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$



Kernel is a smooth function which obeys certain properties

Non-parametric Modeling

- Non-parametric methods are dependent on number of data points
  - Estimation is difficult for large datasets.
- Likelihood computation and model comparisons are hard.
- Limited use in classifiers

Parametric Models (Chap 2 PRML)

- Collection of probability distributions which are described by a finite dimensional parameter set

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_K) \qquad P = \{P_{\boldsymbol{\theta}}\}$$

- Examples -
  - Poisson Distribution

$$p_\lambda(j) = \frac{\lambda^j}{j!} e^{-\lambda}$$

  - Bernoulli Distribution

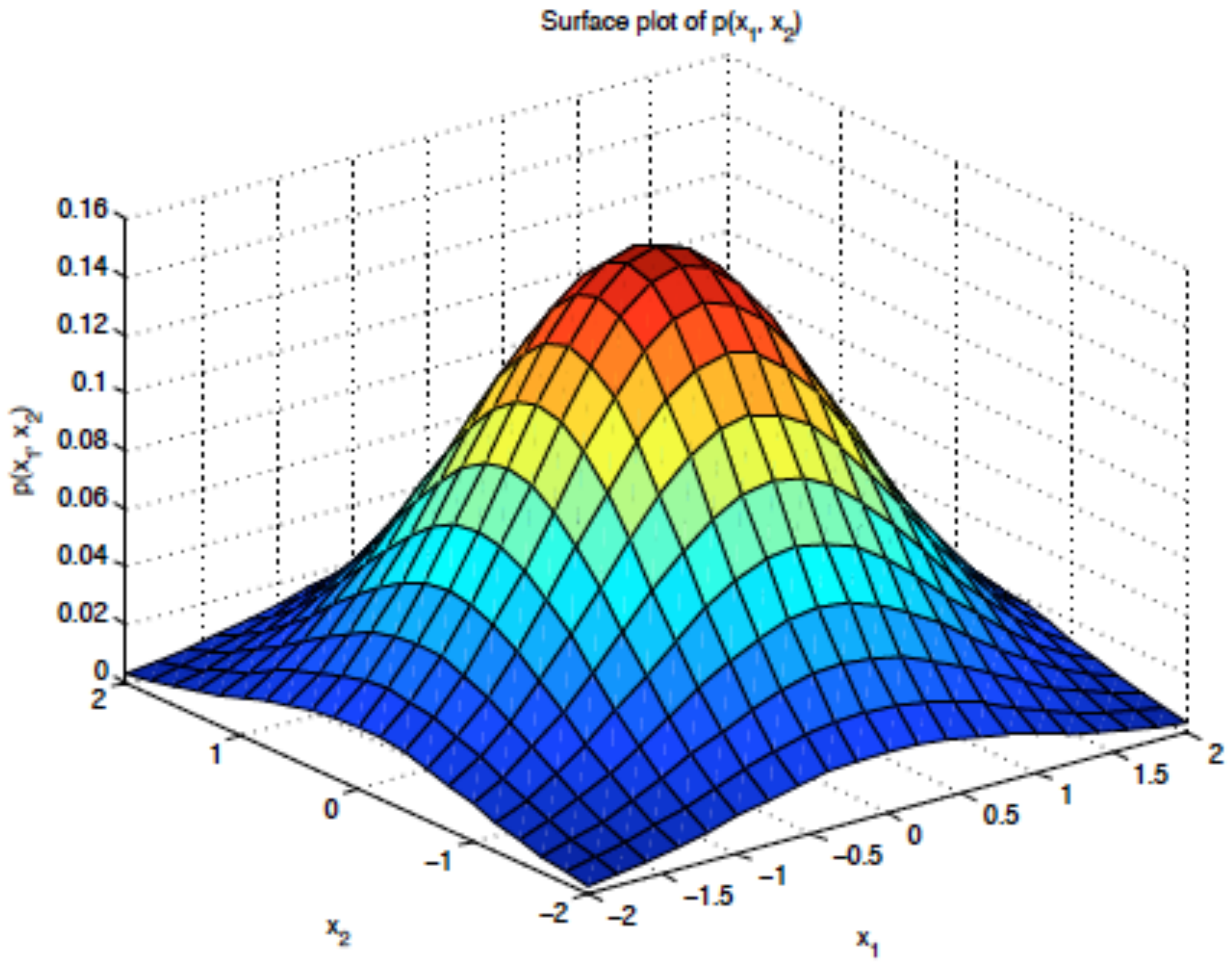
$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{x_i}$$

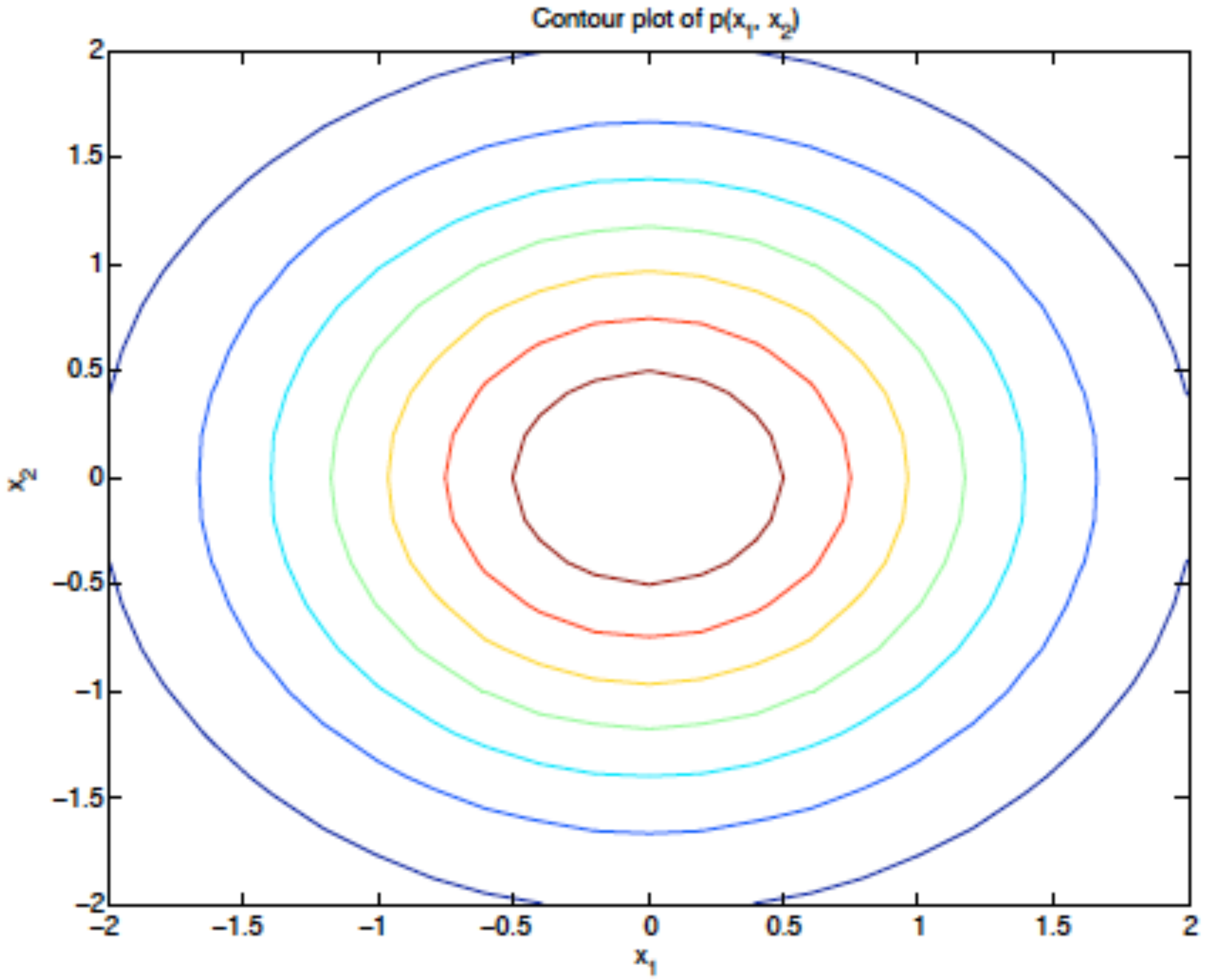
  - Gaussian Distribution

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

Gaussian Distribution

One of most widely used and well studied model



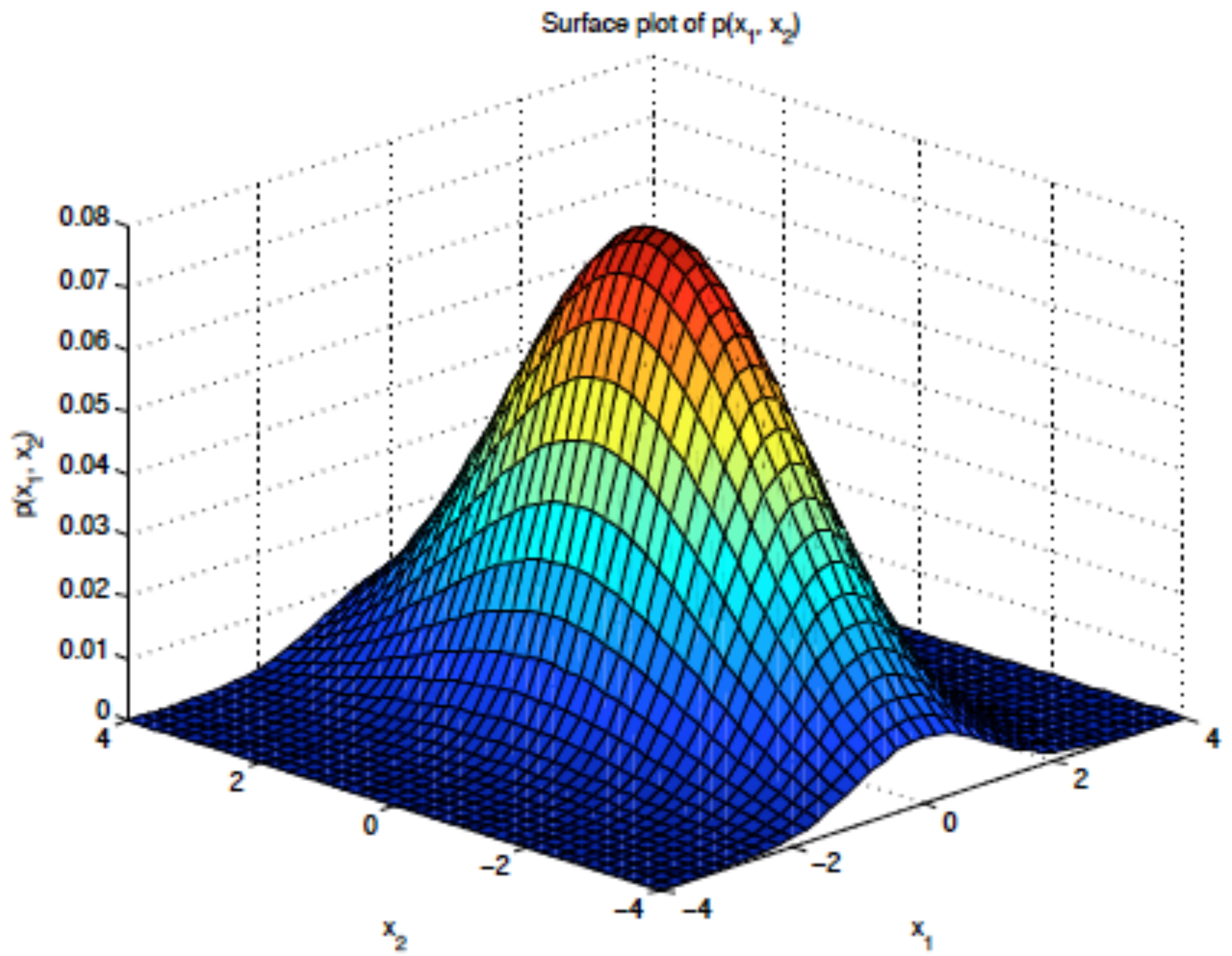


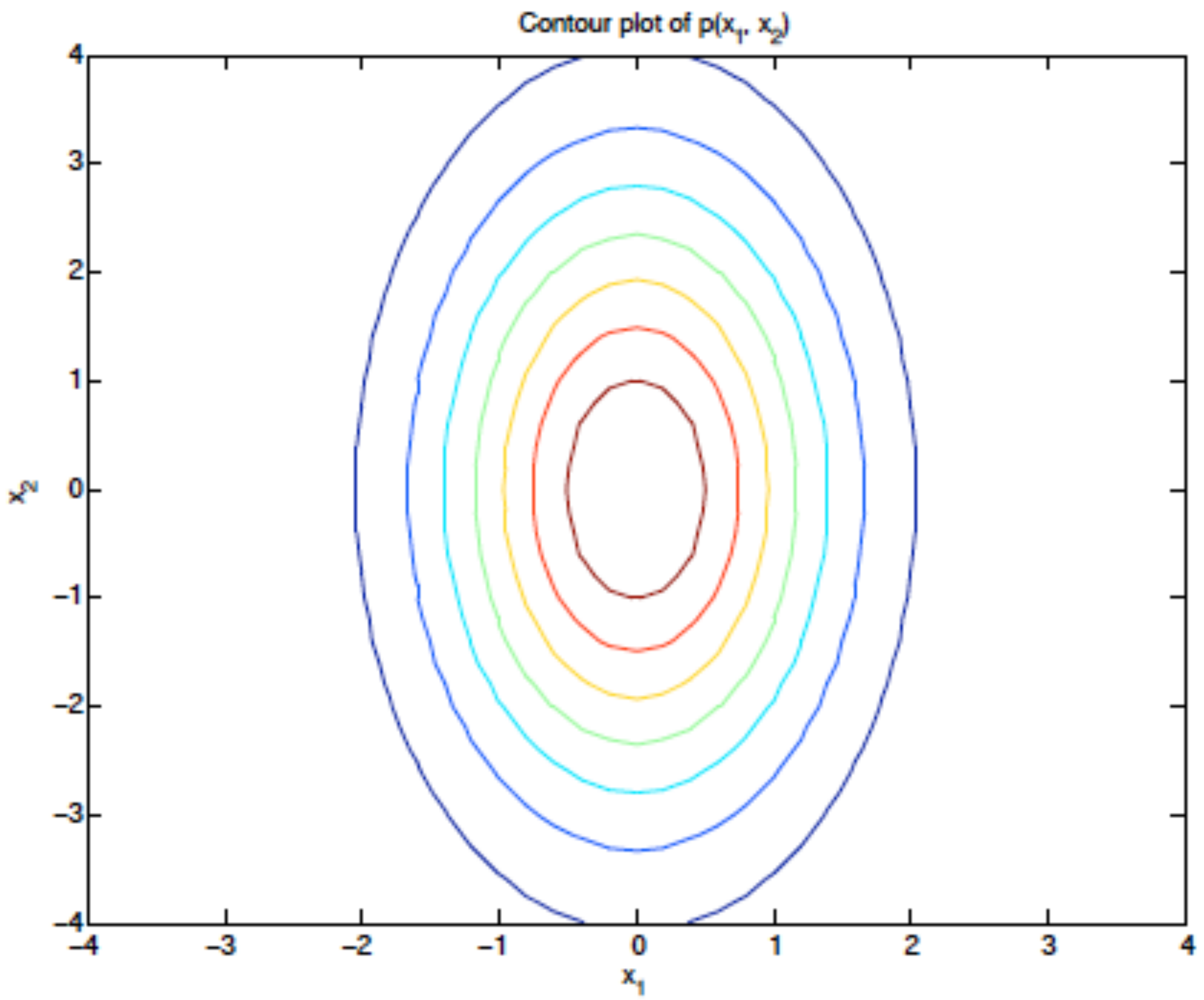
Points of equal probability lie on on contour

Diagonal Gaussian with Identical Variance

Gaussian Distribution

Insights into two dimensional Gaussian distribution

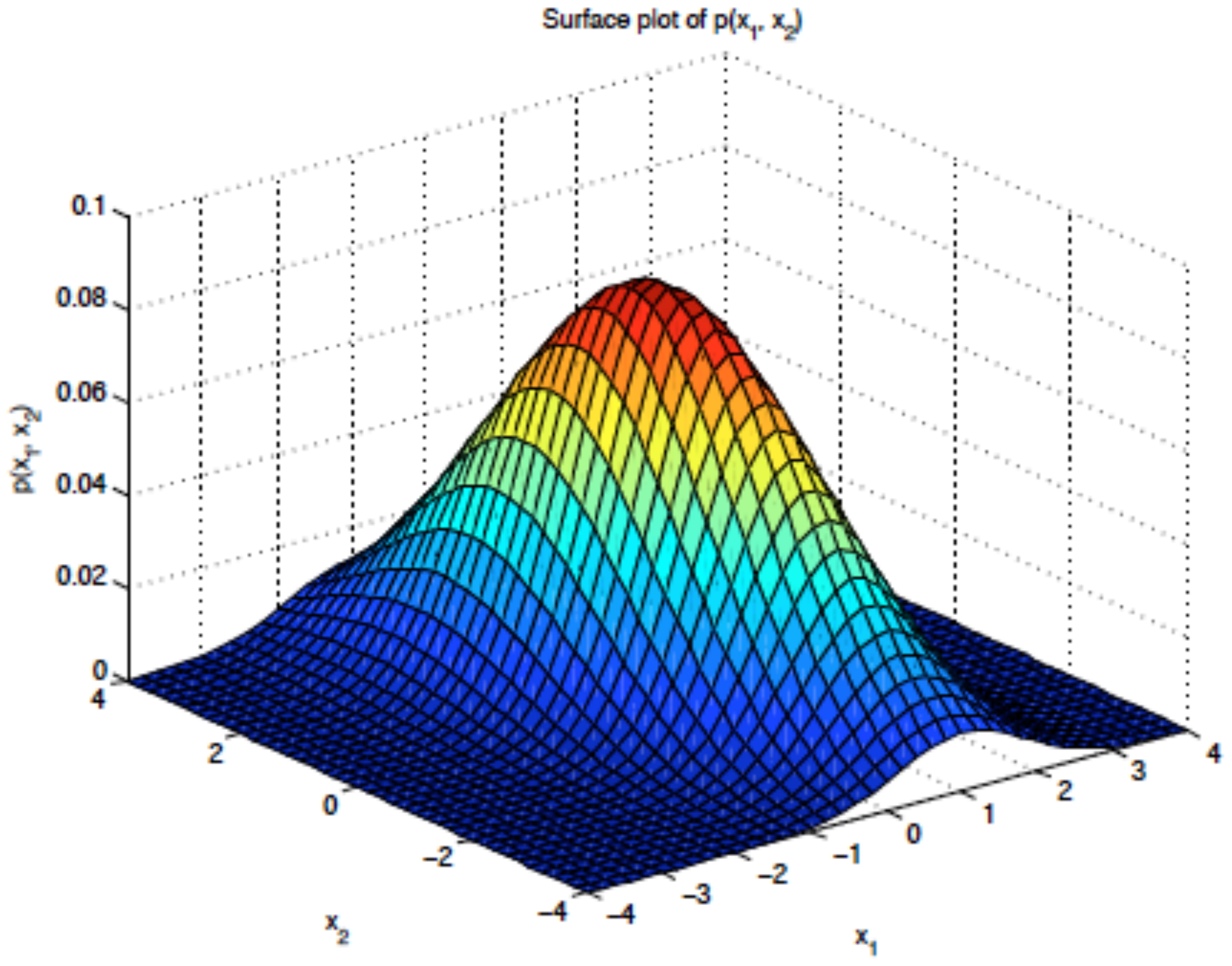


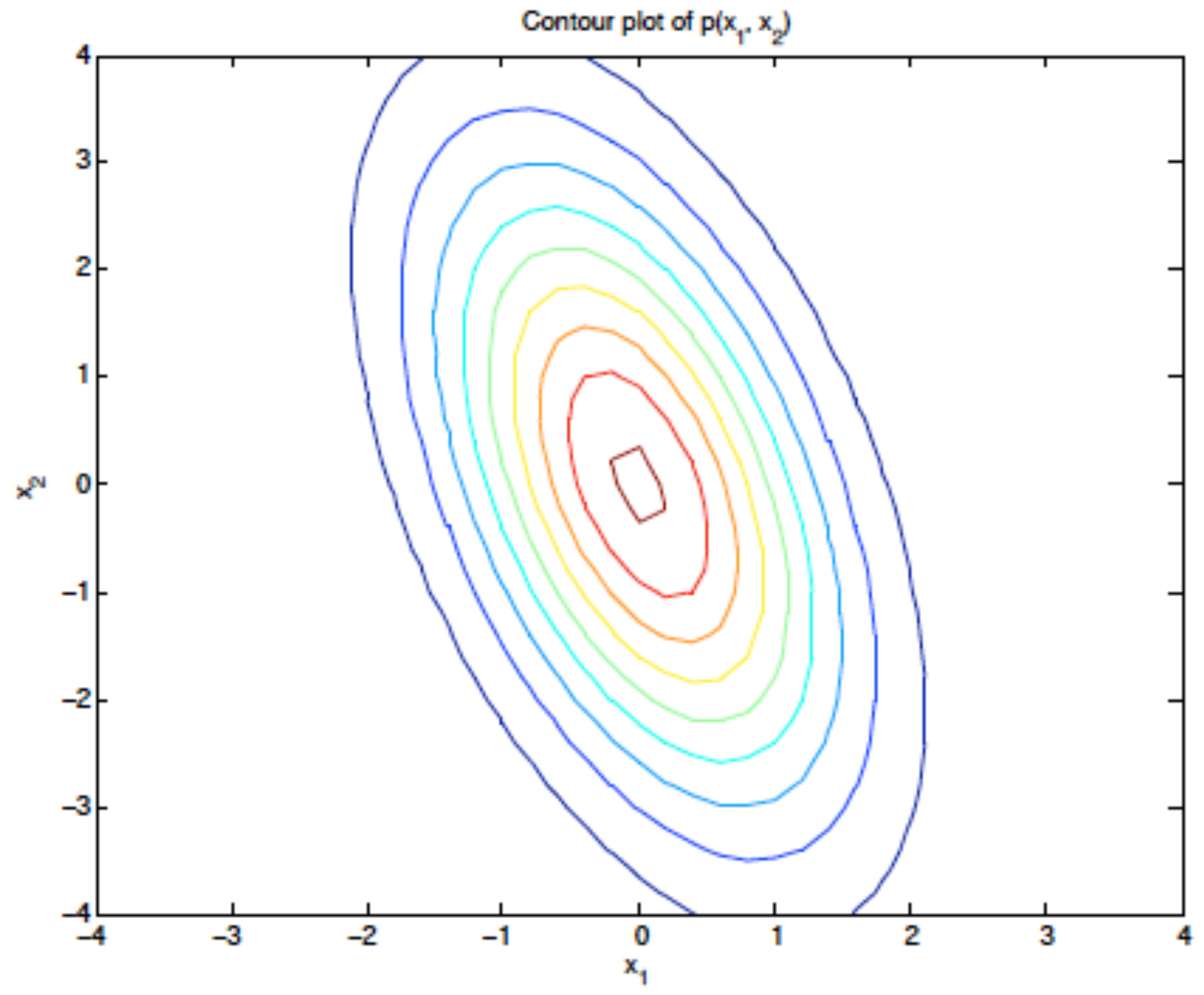


Diagonal Gaussian with different variance

Gaussian Distribution

Insights into two dimensional Gaussian Distribution

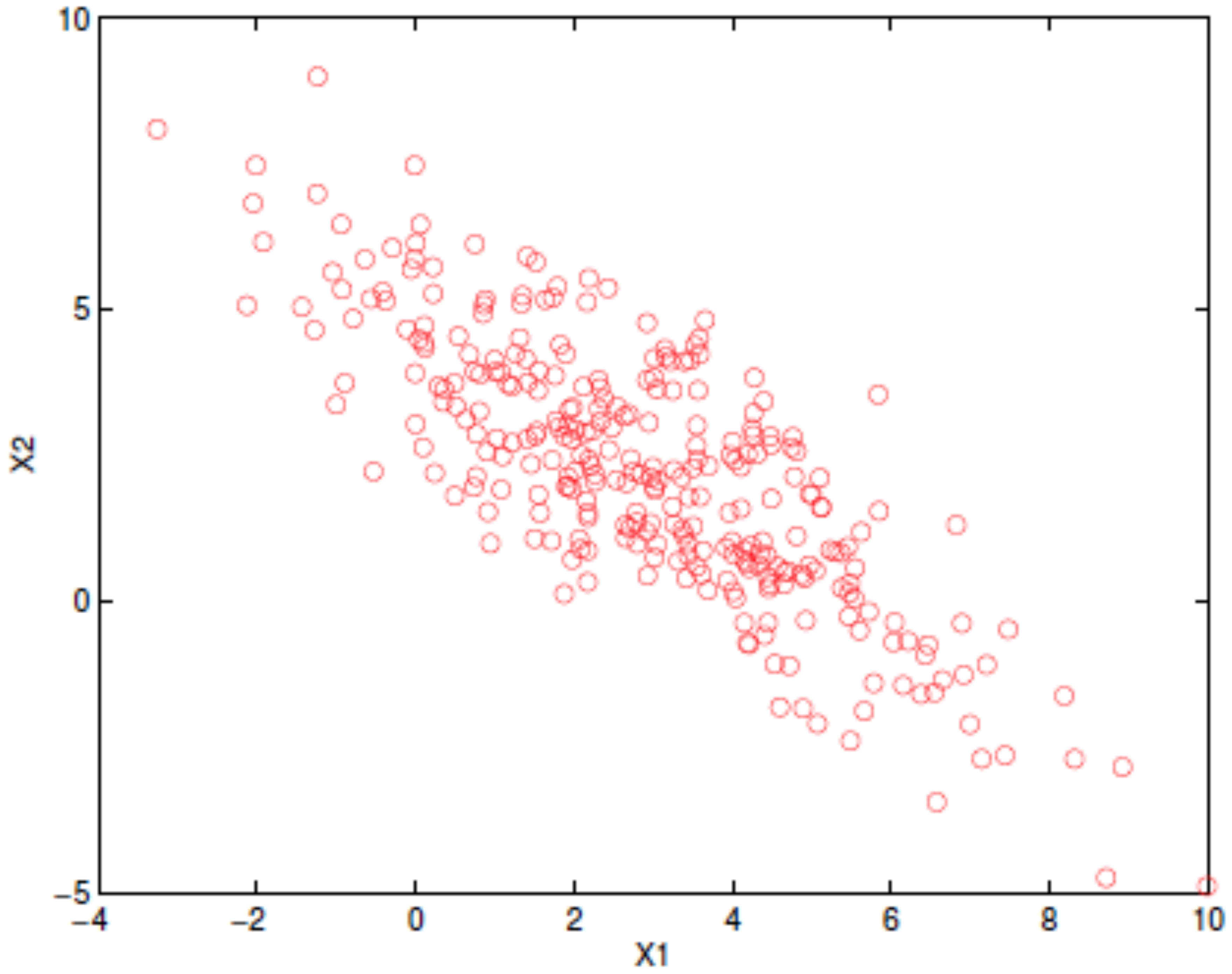


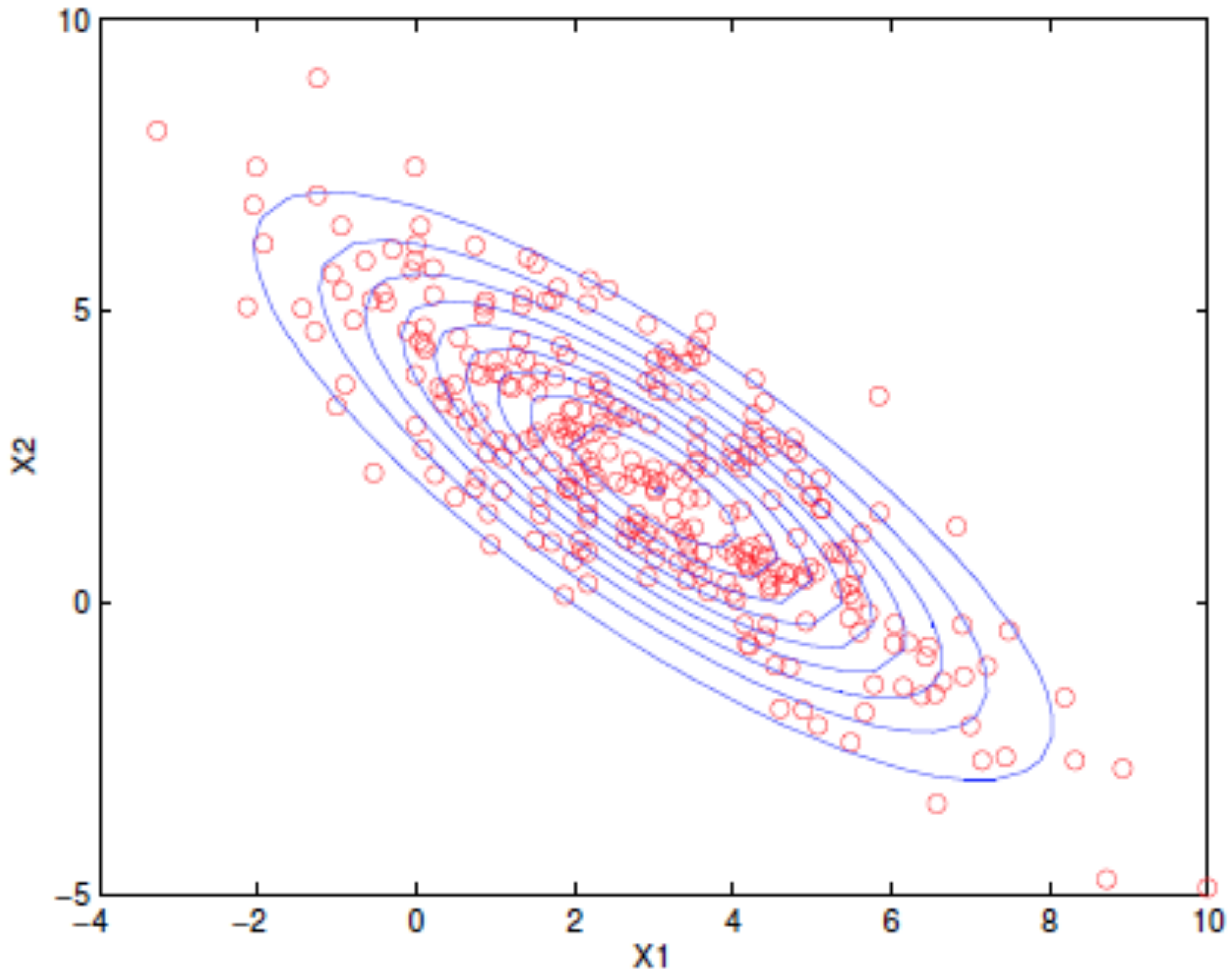


Full covariance Gaussian distribution

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

MLE

For the Gaussian distribution

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

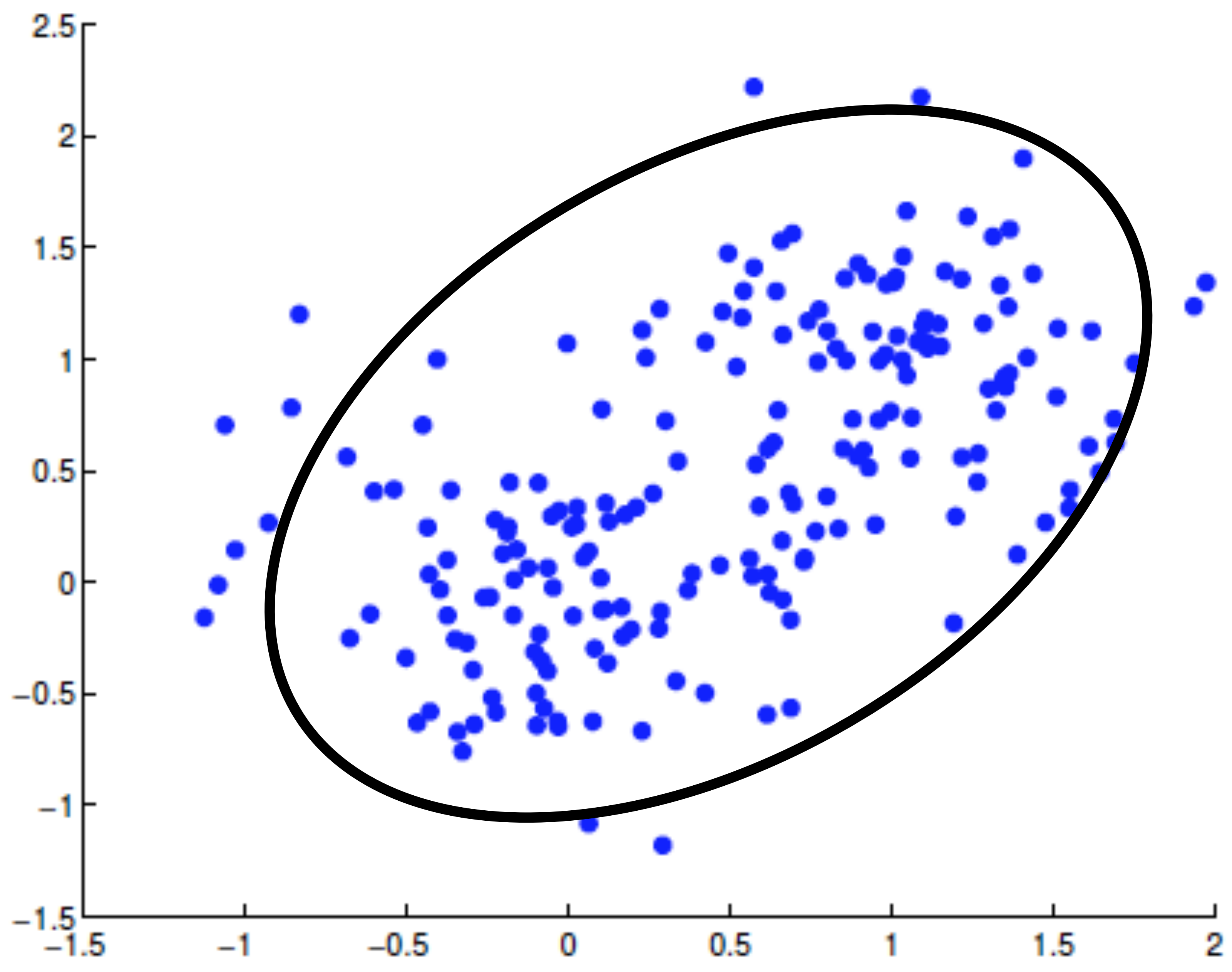
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i|\boldsymbol{\theta})$$

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

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

Gaussian Distribution

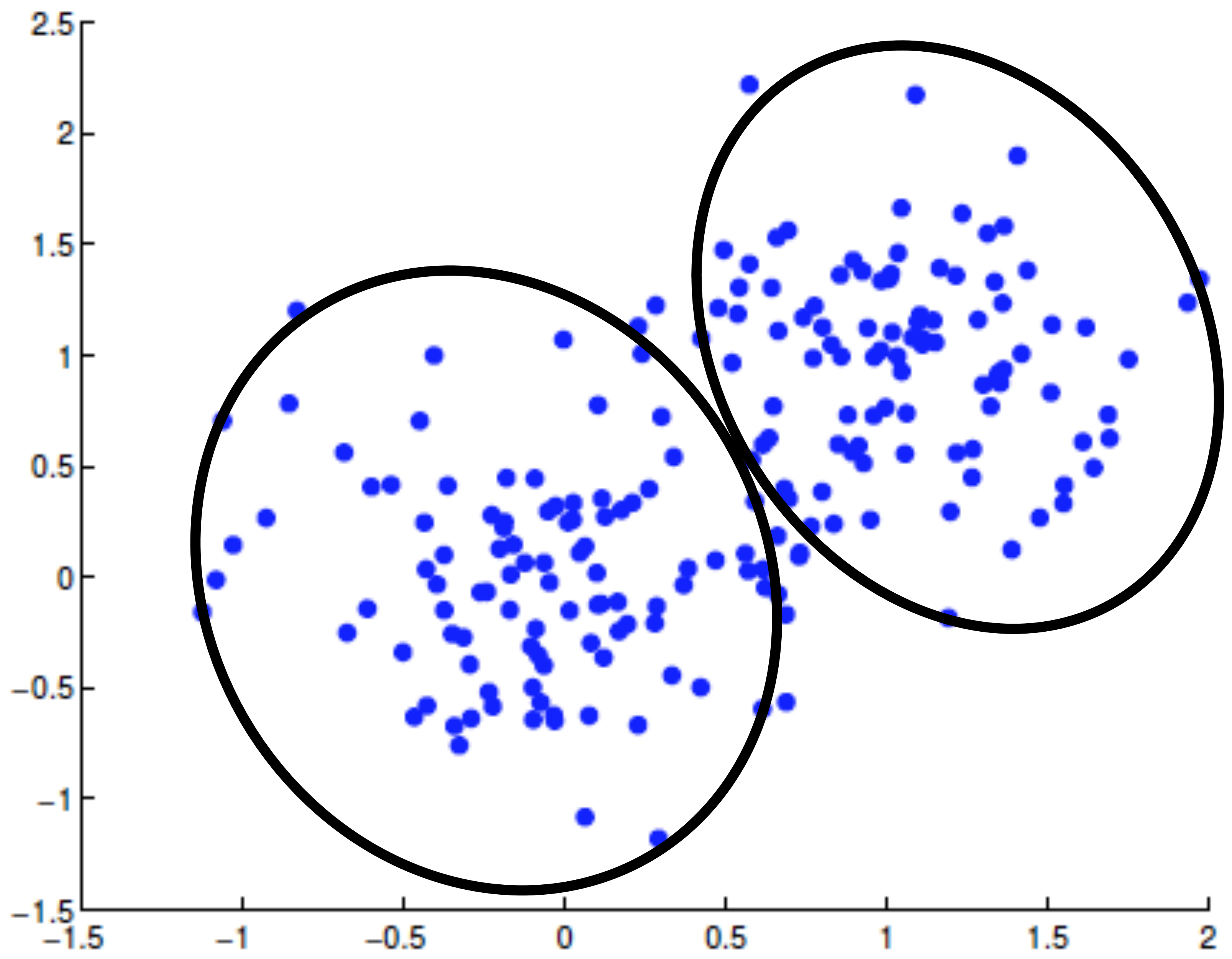
Often the data lies in clusters (2-D example)



Fitting a single Gaussian model may be too broad.

Gaussian Distribution

Need mixture models



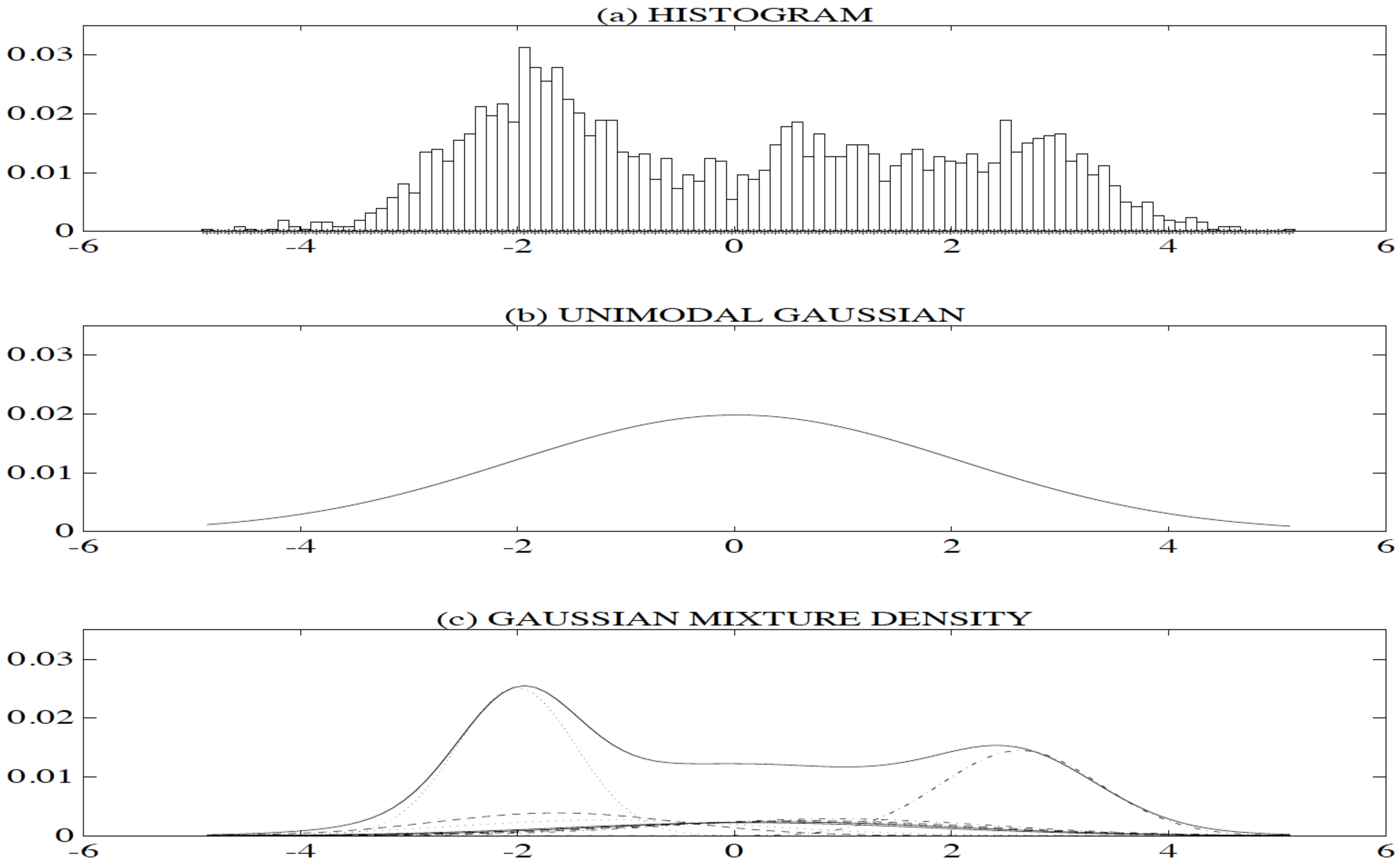
Can fit any arbitrary distribution.

Established 1911

Gaussian Distribution

Often the data lies in clusters

1-D example



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

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

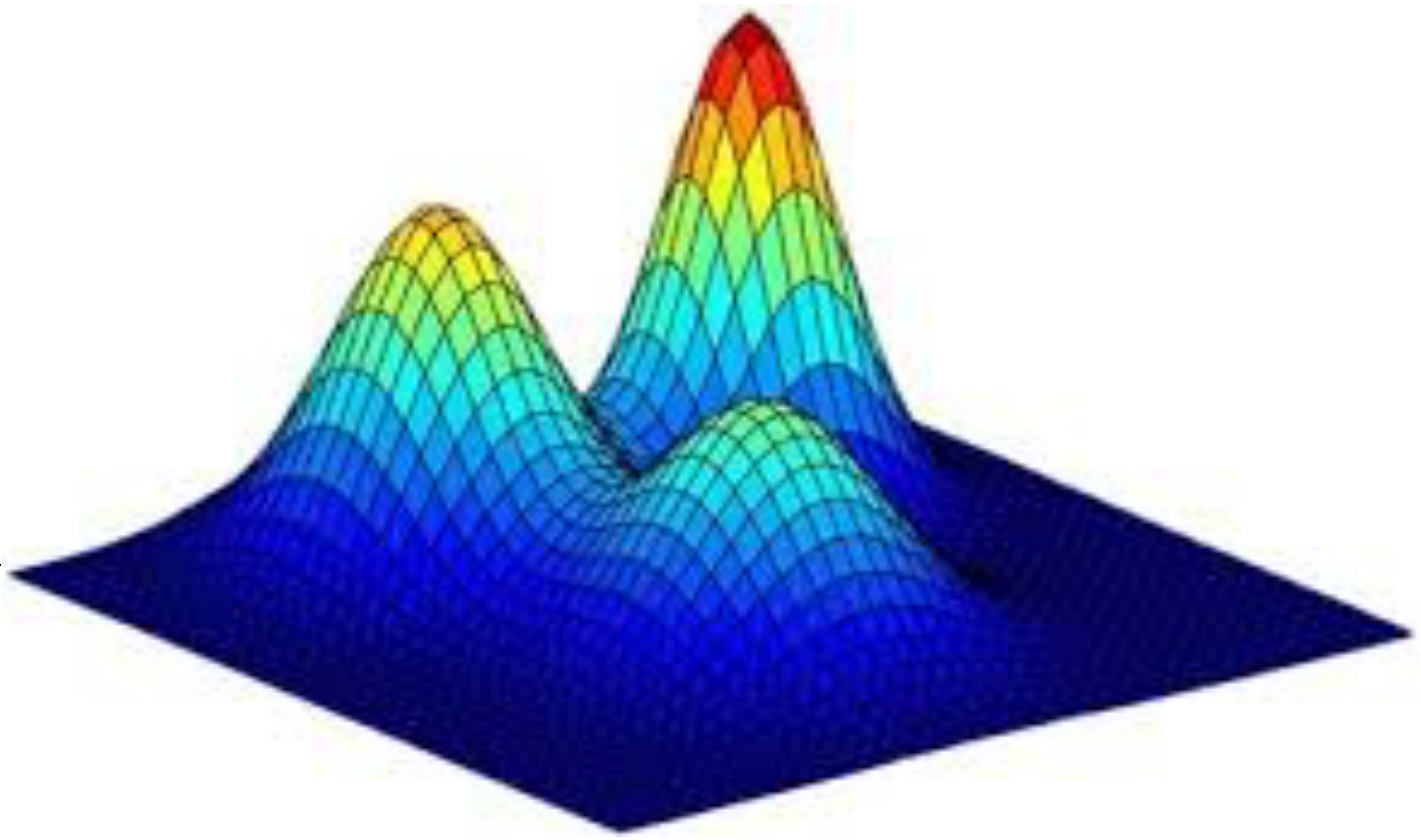
$$p(\mathbf{x}|\boldsymbol{\theta}_k) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_k|}} exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^* \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}-\boldsymbol{\mu}_k) \right\}$$

The weighting coefficients have the property

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

Gaussian Mixture Models

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



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 number of parameters is $KD^2 + KD + K$

MLE for GMM

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

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

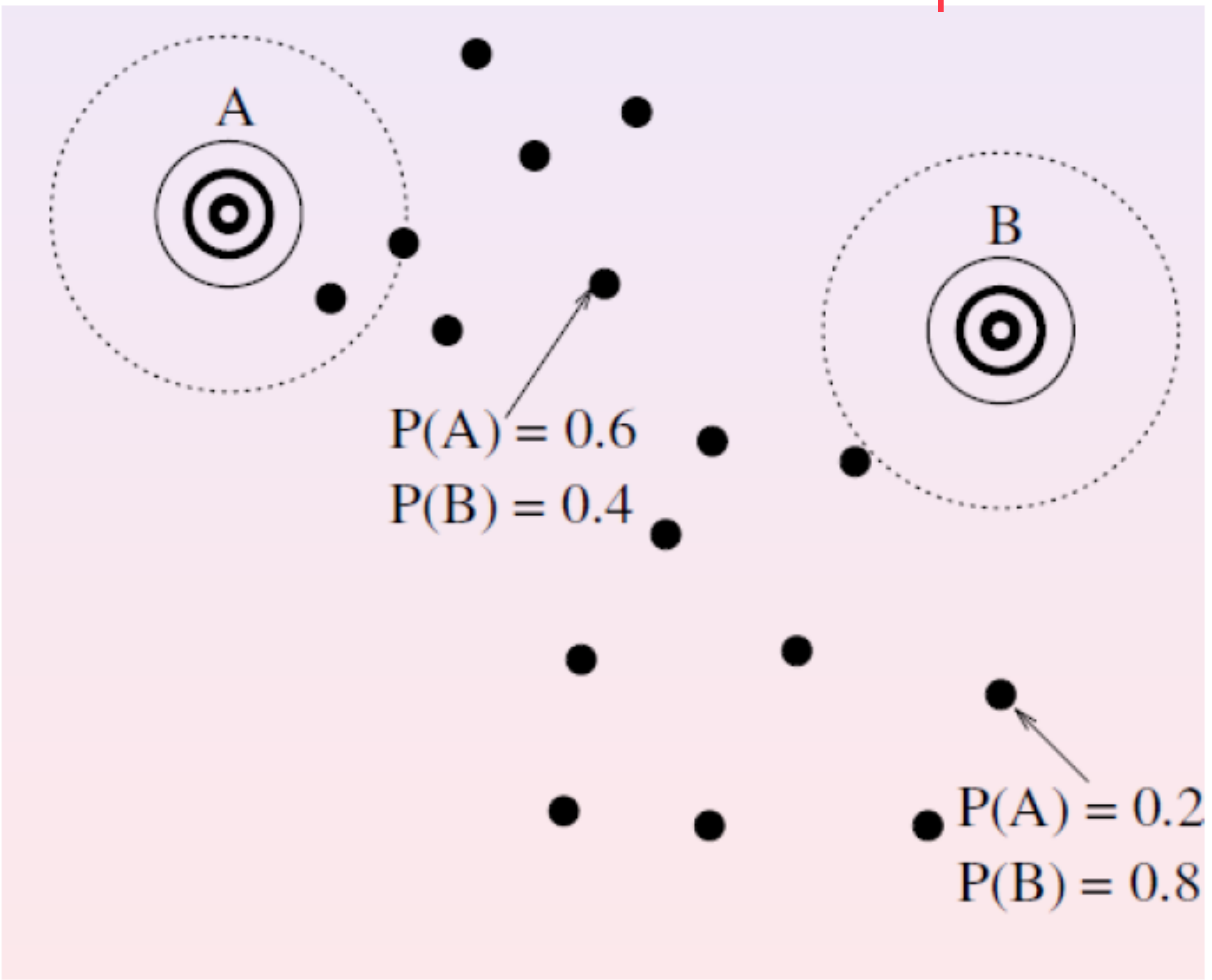
- Solving for the optimal parameters using MLE for GMM is not straight forward.
- Resort to the Expectation Maximization (EM) algorithm

EM Algorithm For GMMs

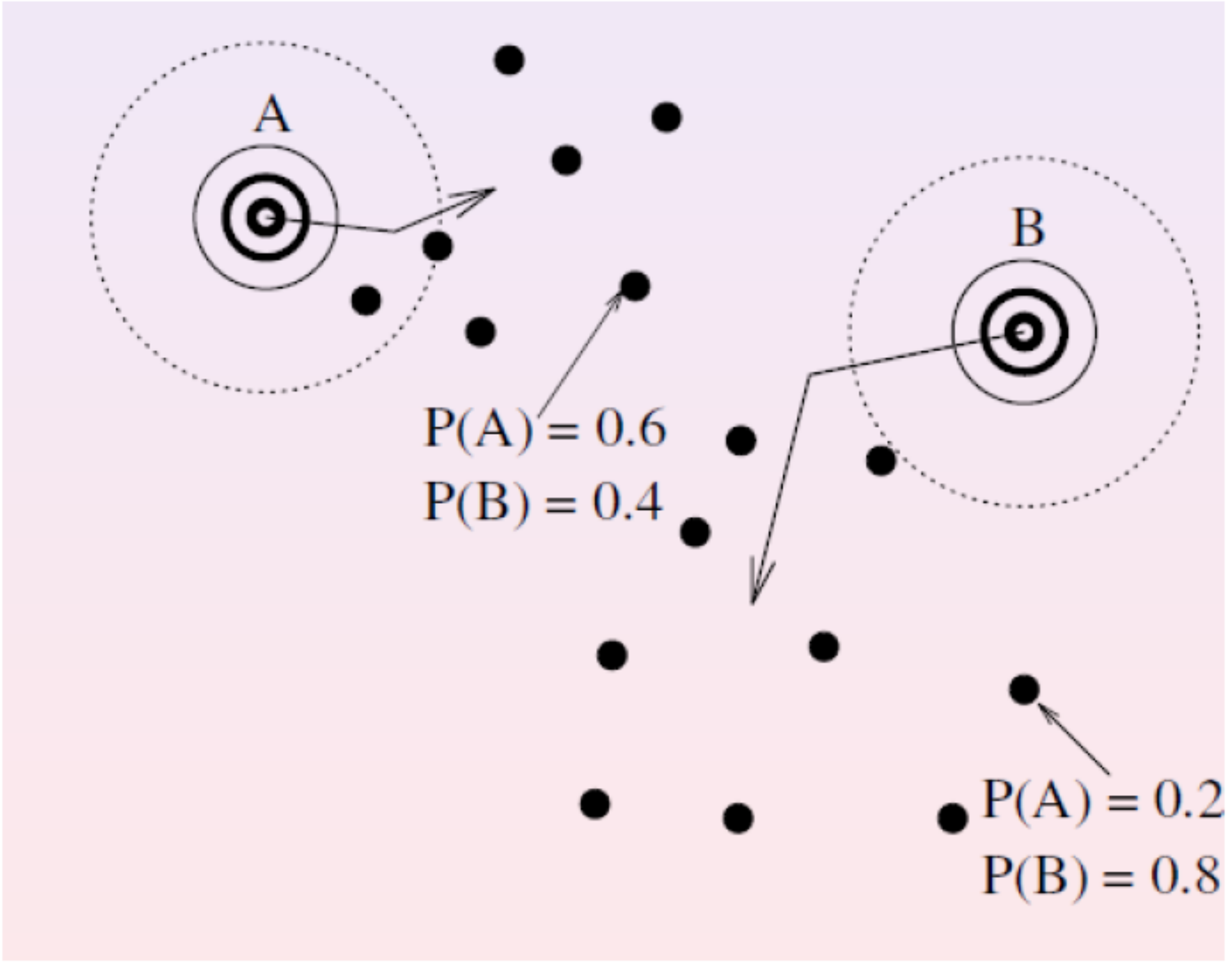
EM Algorithm for GMM

- The hidden variables will be the index of the mixture component which generated
- Re-estimation formulae

E-step



M-step



THANK YOU

Sriram Ganapathy and TA team
LEAP lab, C328, EE, IISc
sriramg@iisc.ac.in