E9 205 Machine Learning for Signal Processing

ML, MAP, MMSE and Gaussian Modeling

28-08-2019

Instructor - Sriram Ganapathy (sriram@ee.iisc.ernet.in)

Teaching Assistant - Prachi Singh (prachisingh@iisc.ac.in).

Decision Theory (PRML Chap. 1.5)

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

Decision Problem - Classification

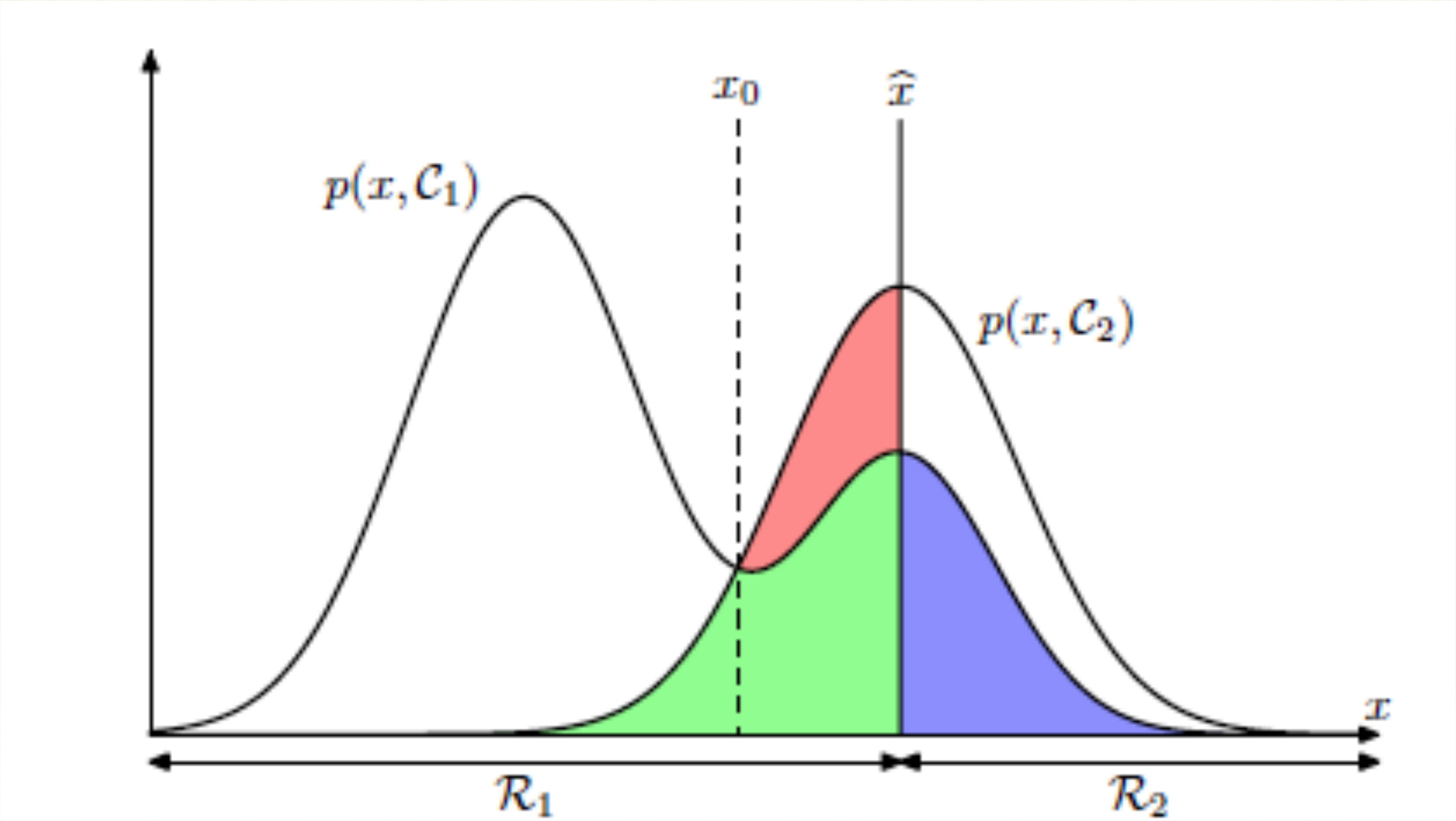
- Minimizing the mis-classification error
- Decision based on maximum posteriors

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

- Loss matrix
  - Minimizing the expected loss

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

Visualizing the Max. Posterior Classifier



Approaches for Inference and Decision

I. Finding the joint density from the data.

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

II. Finding the posteriors directly.

III. Using discriminant functions for classification.

Approaches for Inference and Decision

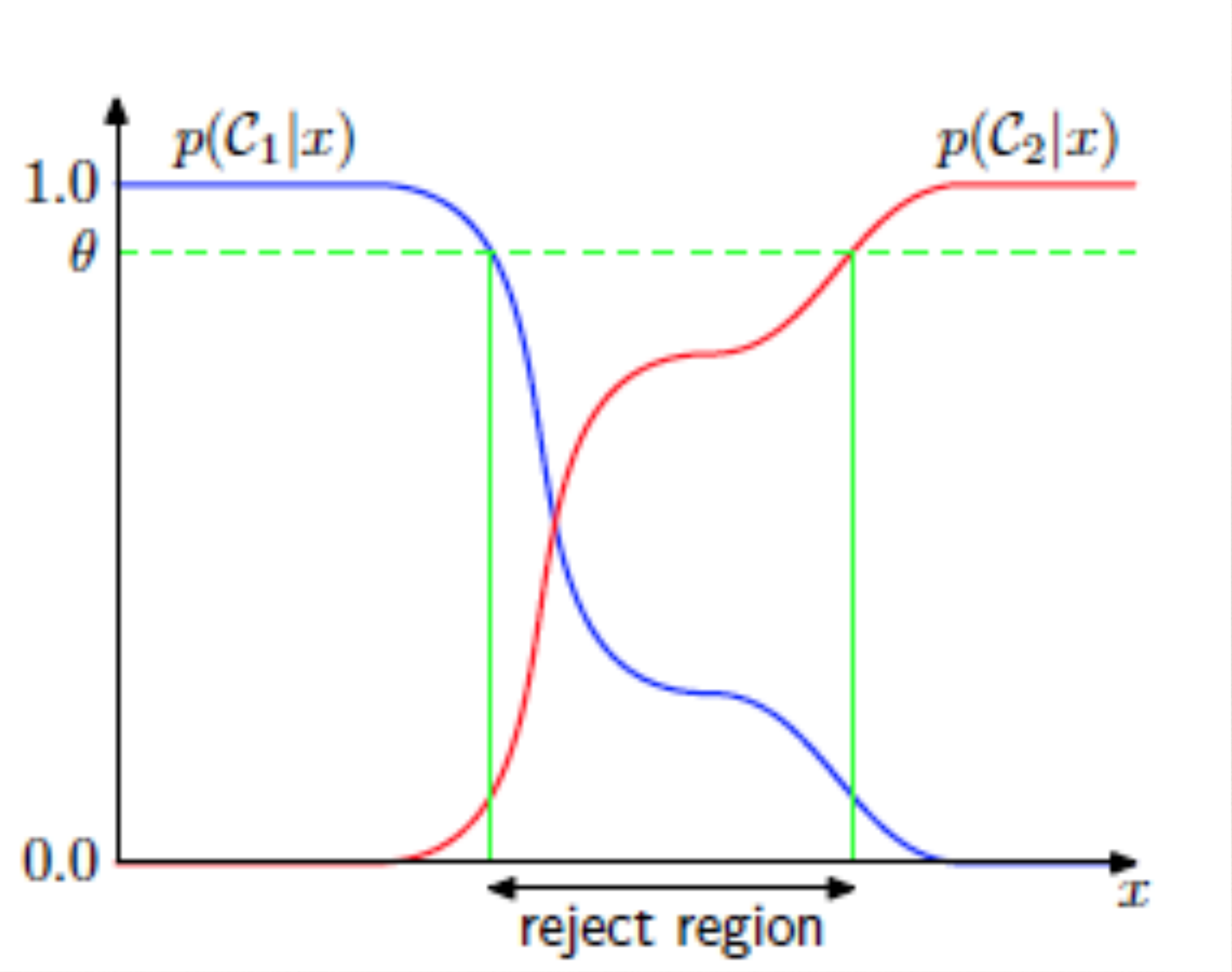
I. Finding the joint density from the data.

$$p(C_k|\mathbf{x}) \; \alpha \; p(\mathbf{x}|C_k)p(C_k)$$

II. Finding the posteriors directly.

III. Using discriminant functions for classification.

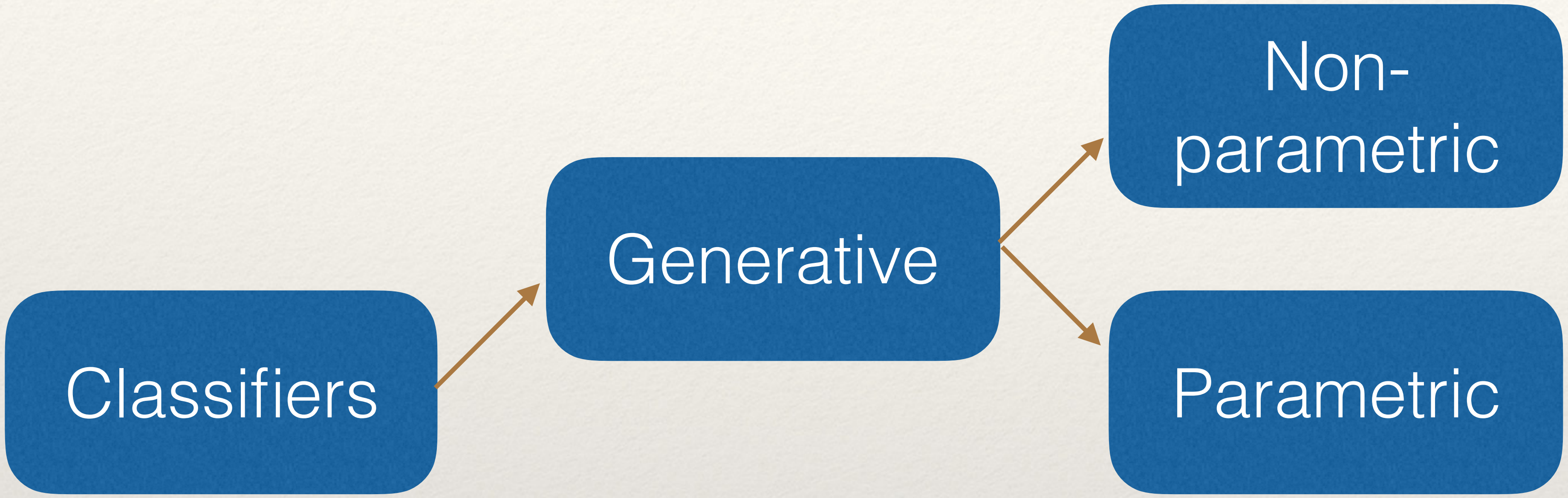
Advantage of Posteriors



Decision Rule for Regression

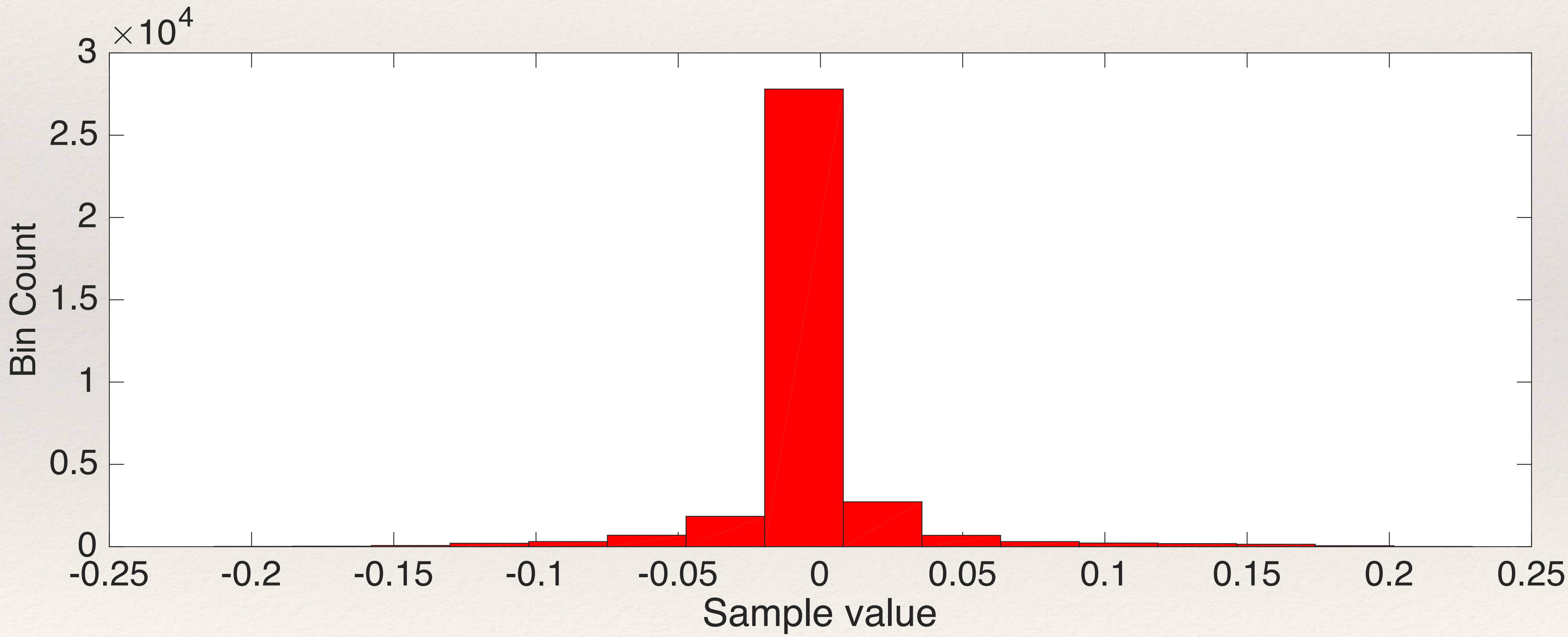
- Minimum mean square error loss
- Solution is conditional expectation.

Generative Modeling



Non-parametric Modeling

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

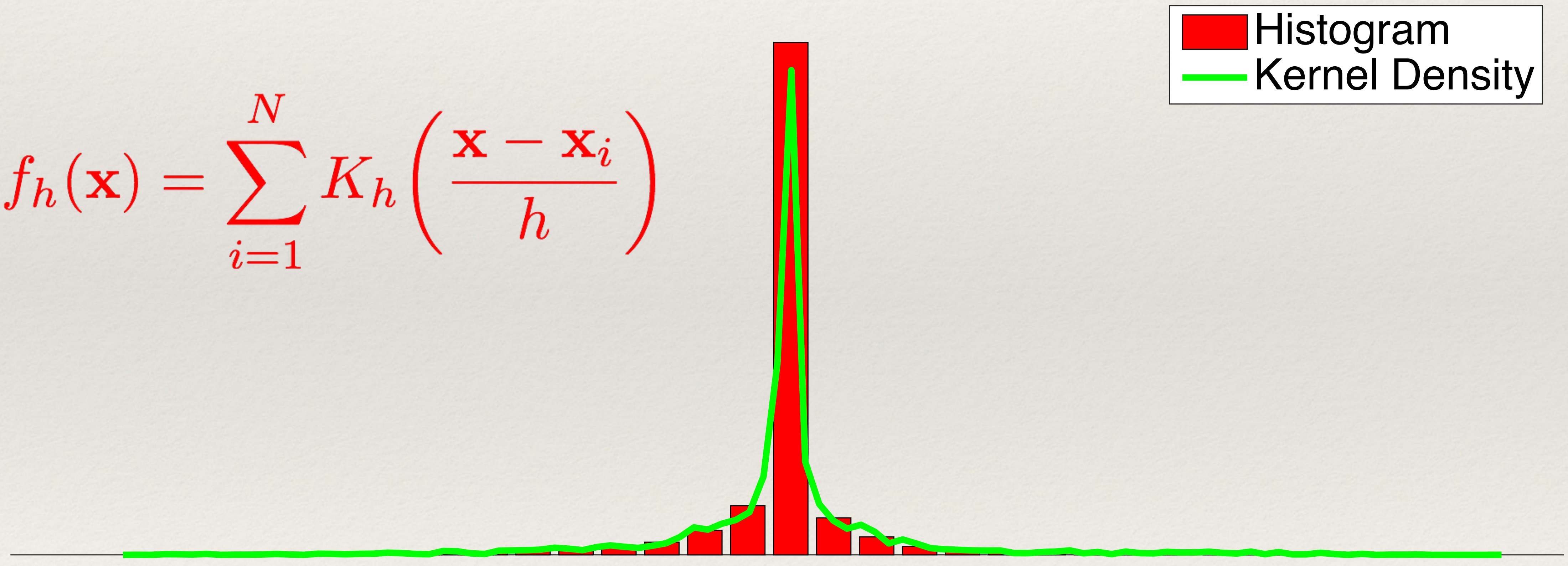


The density is not smooth and has block like shape.

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

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

- Examples -
  - Poisson Distribution

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

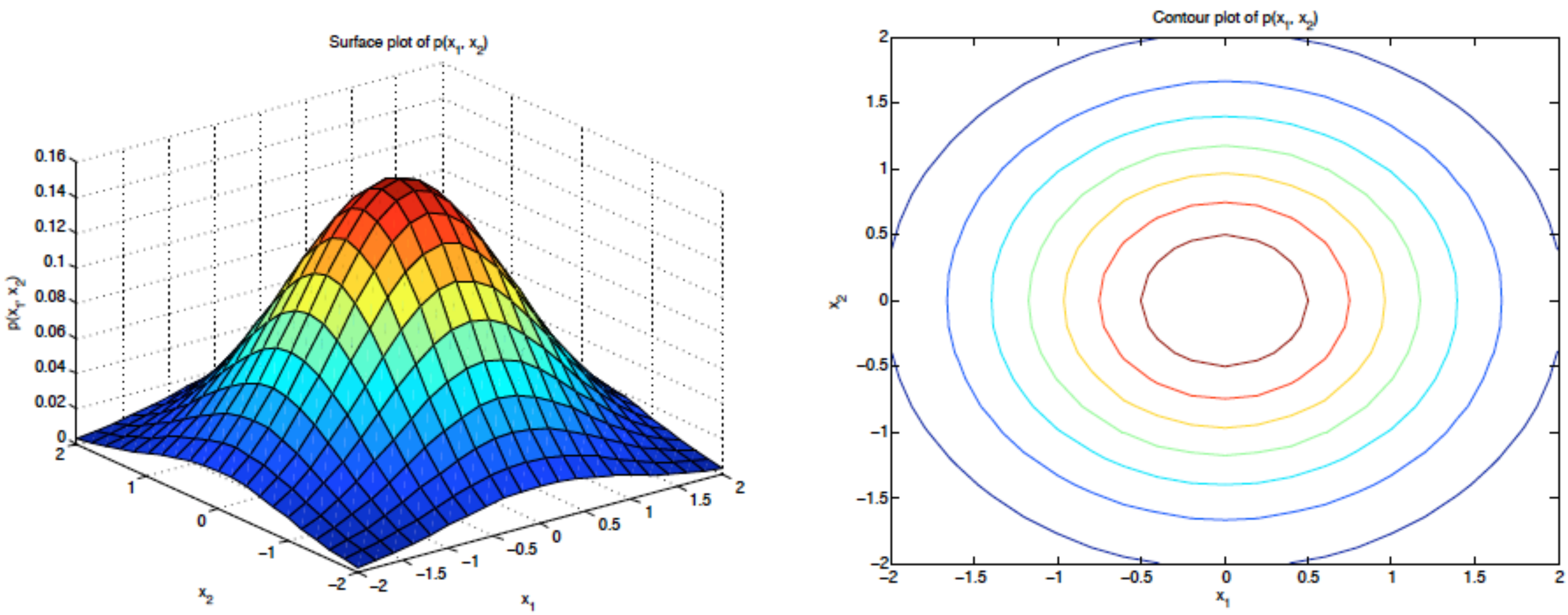
  - Bernoulli Distribution

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

  - Gaussian Distribution

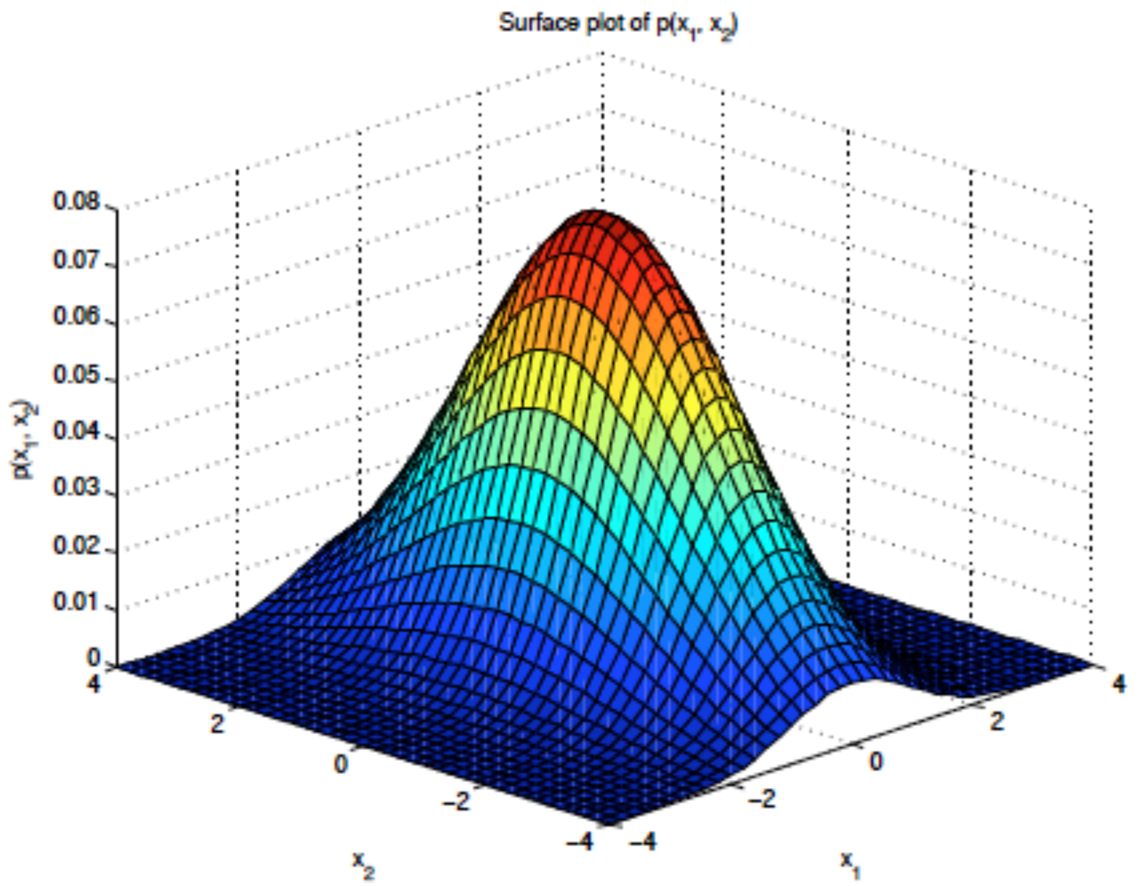
$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^*\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

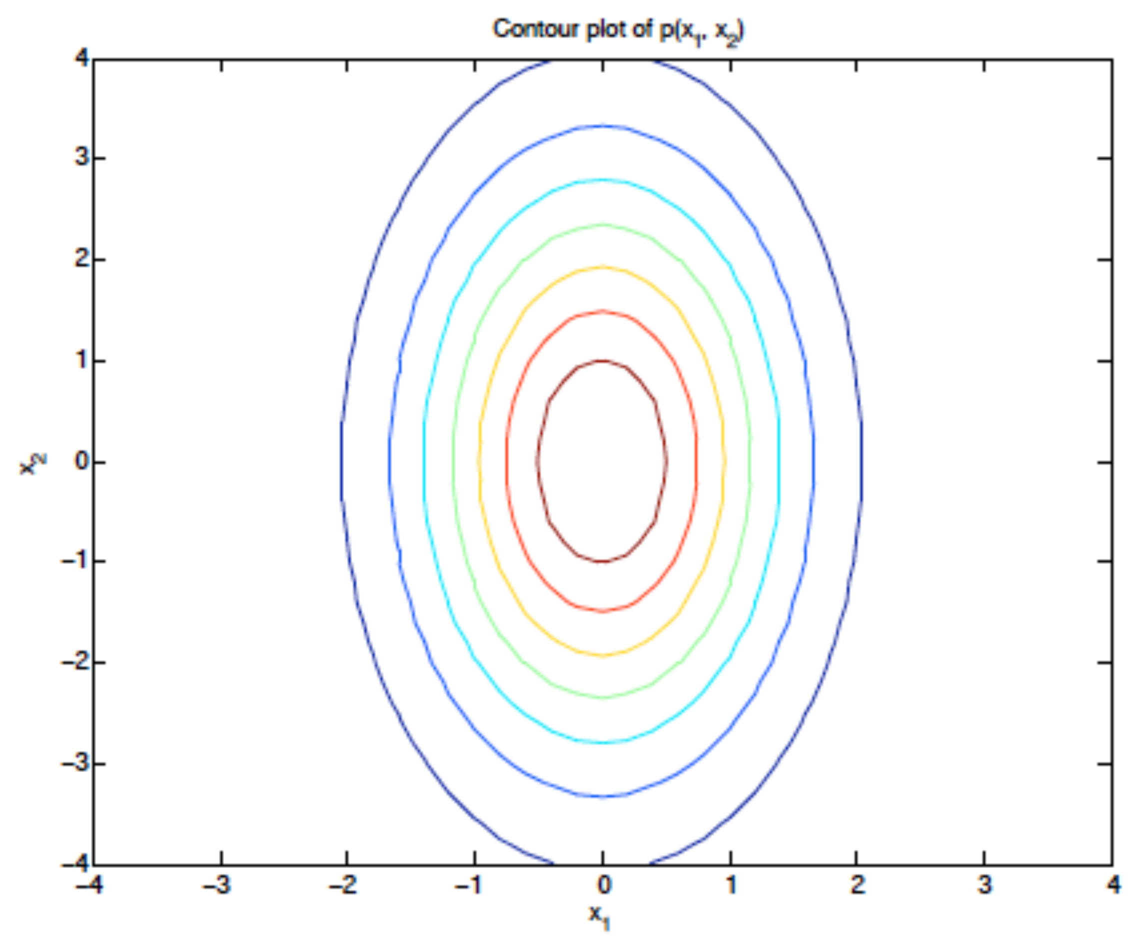
Gaussian Distribution



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

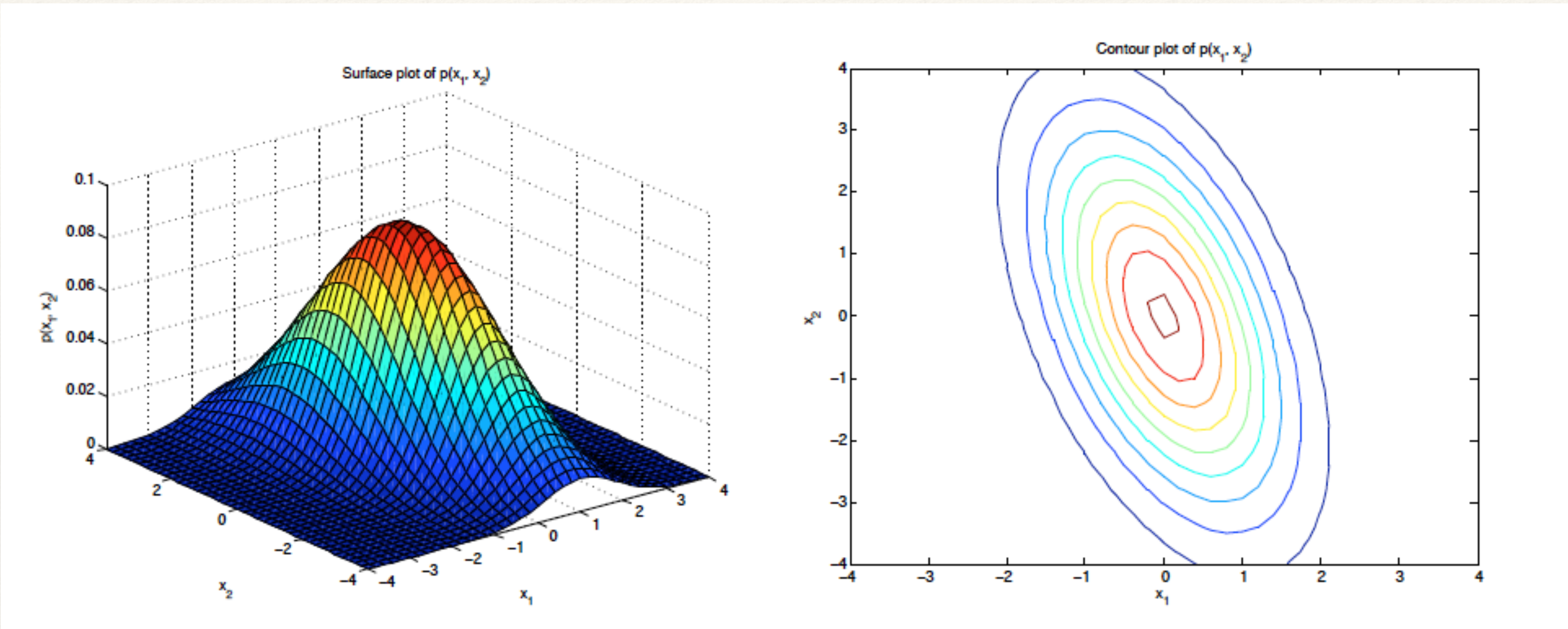
Gaussian Distribution





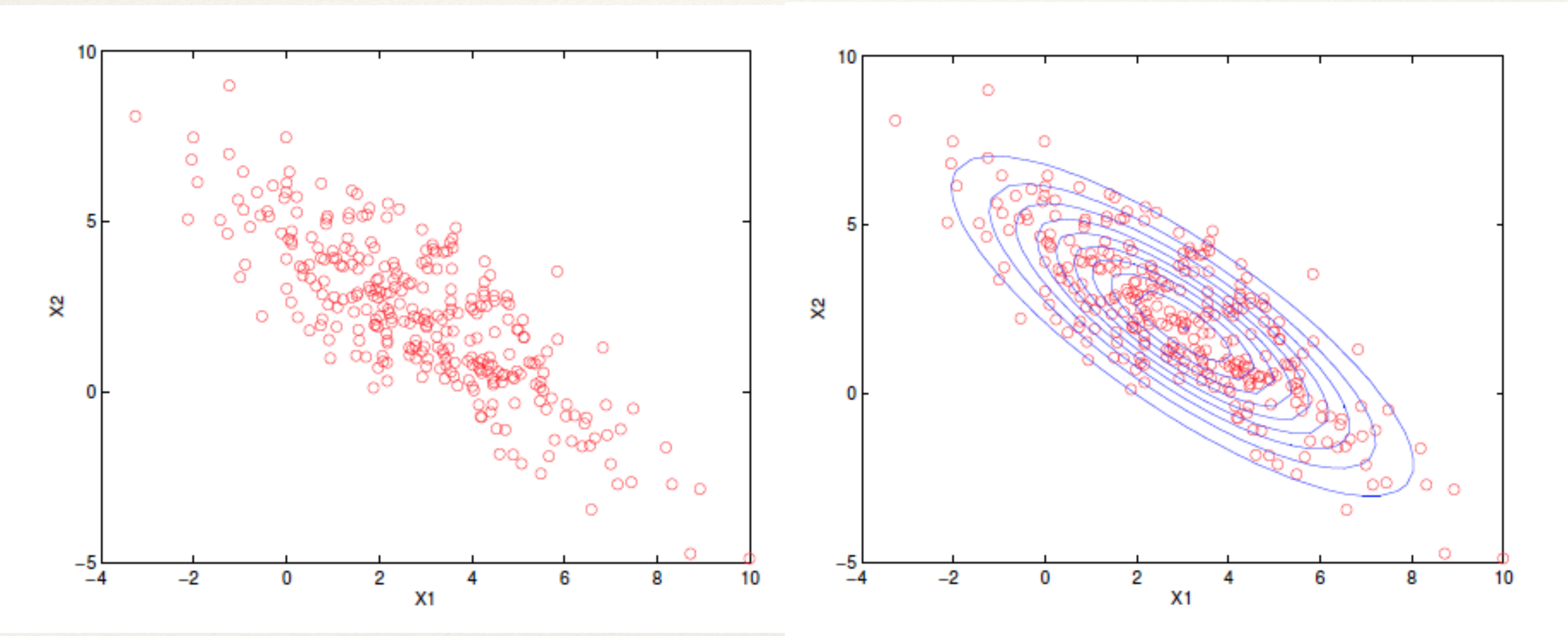
Diagonal Gaussian with different variance

Gaussian Distribution



Full covariance Gaussian distribution

Gaussian Distribution



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).

Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

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 $\frac{\partial \log L}{\partial \boldsymbol{\mu}} = 0$

Established 1911