MACHINE LEARNING FOR SIGNAL PROCESSING

15-1-2025

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

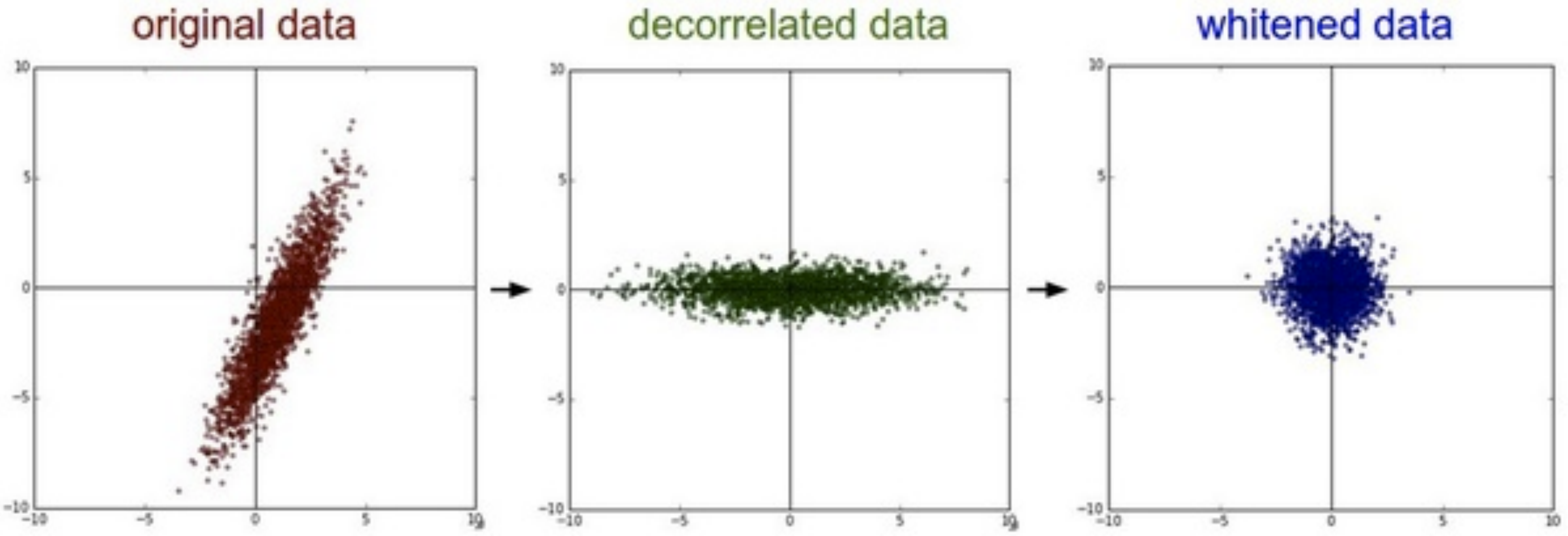
Viveka Salinamakki, Varada R.
LEAP lab, Electrical Engineering, Indian Institute of Science

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

PRINCIPAL COMPONENT ANALYSIS

- Reducing the data $\mathbf{x}_n$ of dimension D to lower dimension
- Projecting the data into subspace which preserves maximum data variance
  - ✓ Maximize variance in projected space $M < D$
- Equivalent formulated as minimizing the error between the original and projected data points.

WHITENING VS DECORRELATIONS



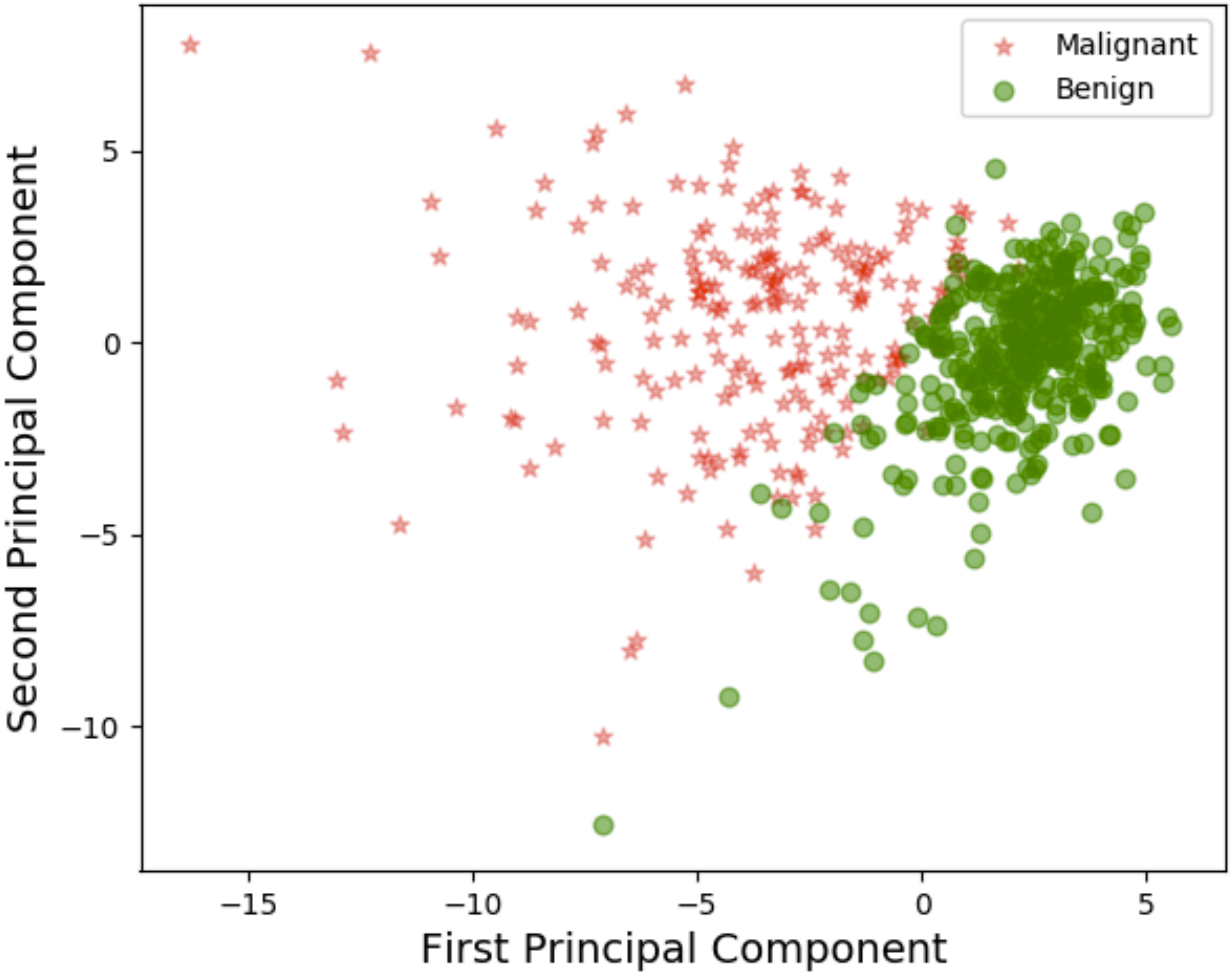
APPLICATION

- Wisconsin Cancer dataset (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
- 569 participants
- 212 (M) 357 (B)
- 30 features —> digitized image of a fine needle aspirate (FNA) of a breast mass. The features describe characteristics of the cell nuclei present in the image.

Raw Features

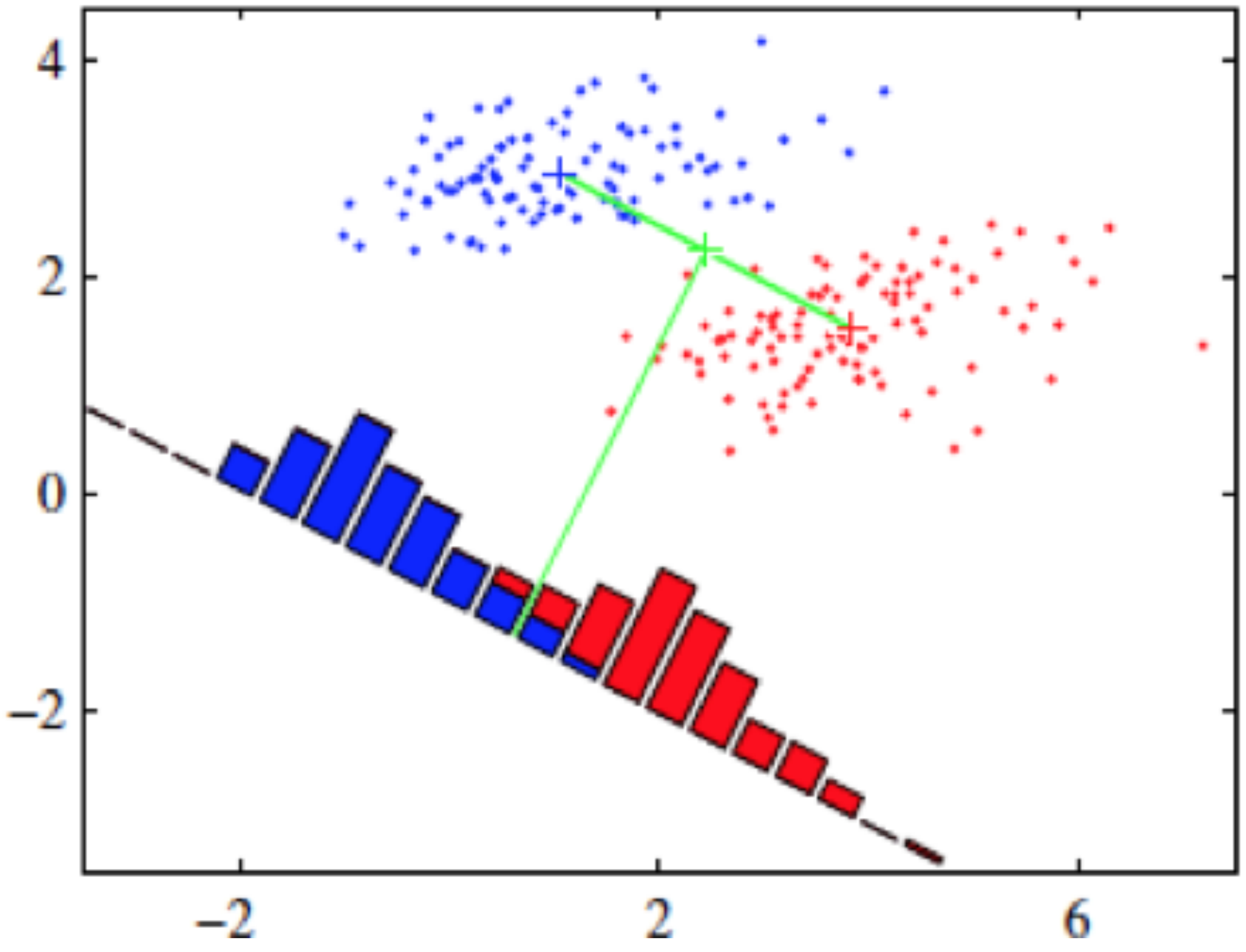
INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

PCA



Malignant
Benign
Second Principal Component
First Principal Component
5
0
−5
−10
−15
−10
−5
0
5

Without the Within Class Factor



LINEAR DISCRIMINANT ANALYSIS

- Generalized Eigenvalue problem

Find a linear transform $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with a criterion which maximizes the class separation

- Maximize the between class distance in the projected space while minimizing the within class covariance

$$J = \frac{\mathbf{w}^T \boldsymbol{S}_b \mathbf{w}}{\mathbf{w}^T \boldsymbol{S}_w \mathbf{w}}$$

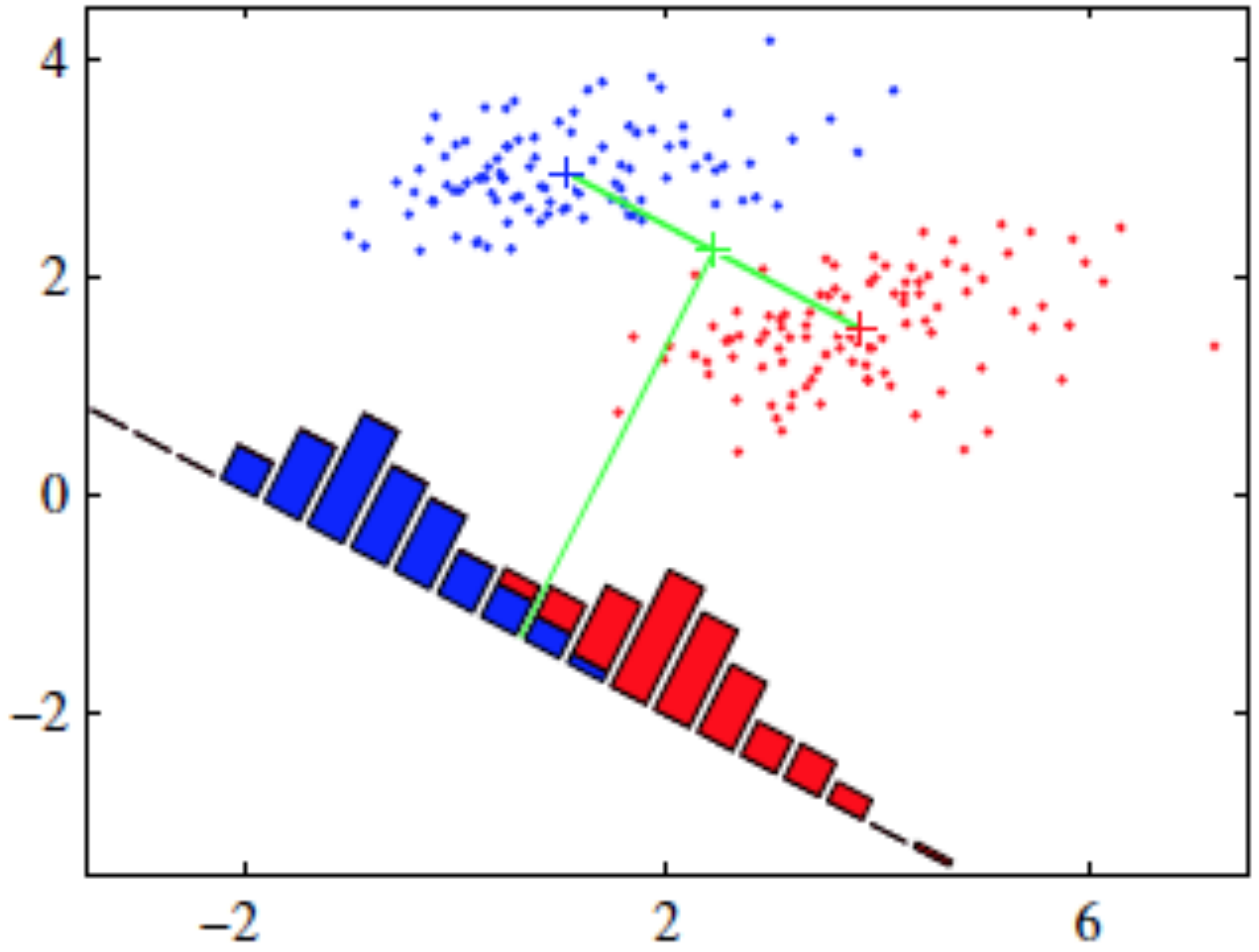
$$\boldsymbol{S}_b = \sum_{k=1}^{K} N_k(\mathbf{m}_k - \boldsymbol{m})(\mathbf{m}_k - \boldsymbol{m})^T \quad \boldsymbol{S}_w = \sum_{k=1}^{K} \sum_{n \in C_k} (\mathbf{x}_n - \boldsymbol{m}_k)(\mathbf{x}_n - \boldsymbol{m}_k)^T$$

Eigen analysis of $\boldsymbol{S}_w^{-1} \boldsymbol{S}_b$

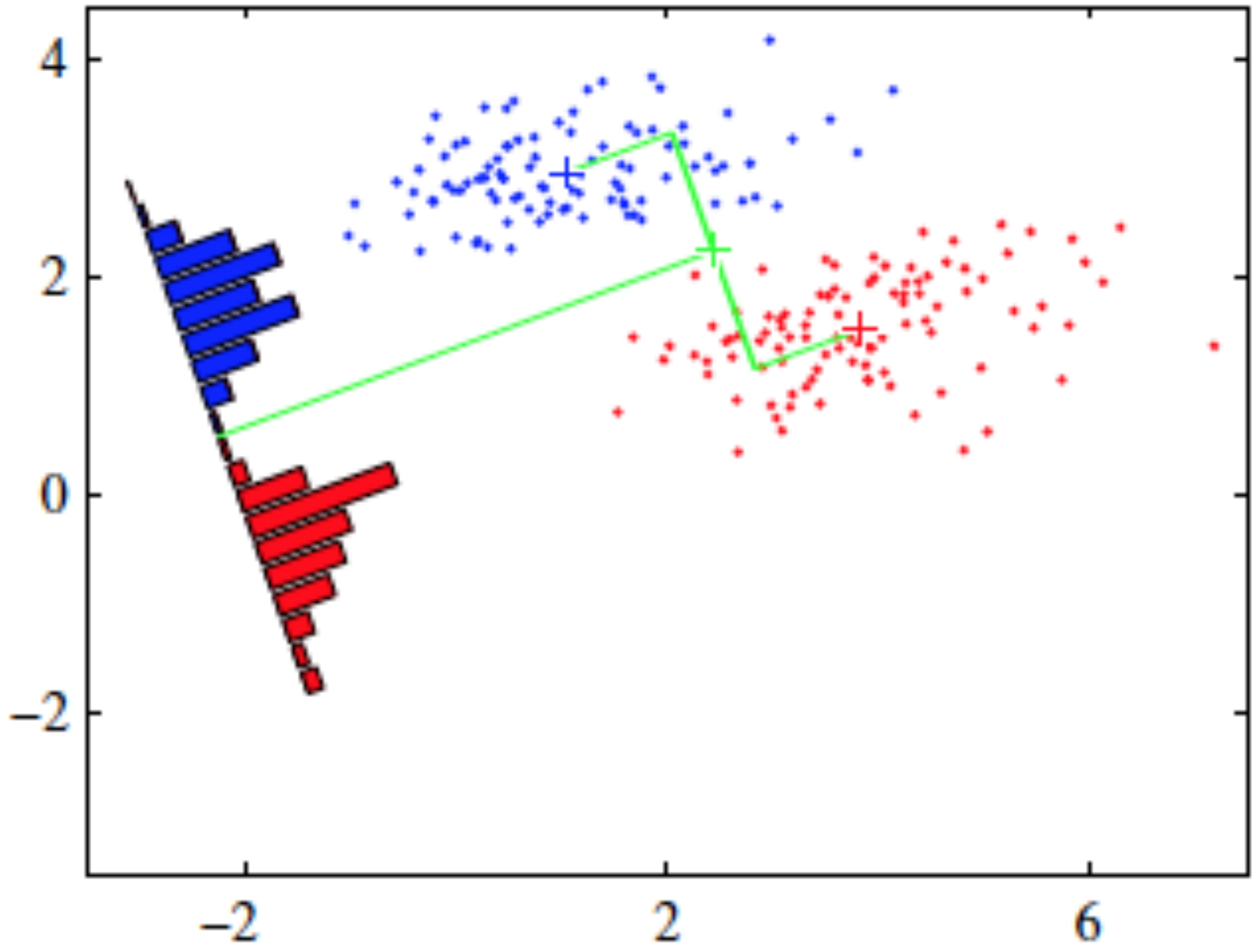
PRML - C. Bishop (Sec. 4.1.4, Sec. 4.1.6)

LINEAR DISCRIMINANT ANALYSIS

Projecting on line joining means

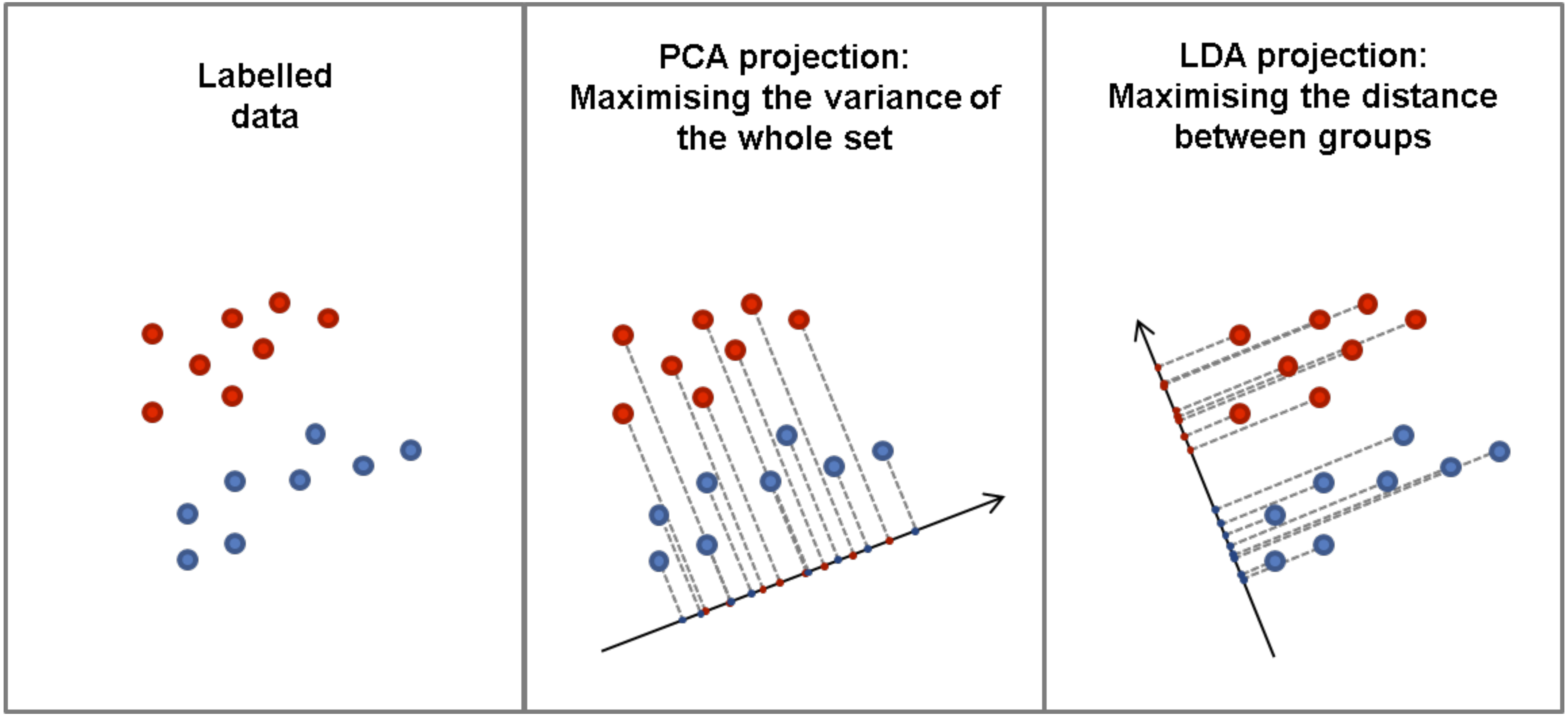


Fisher Discriminant

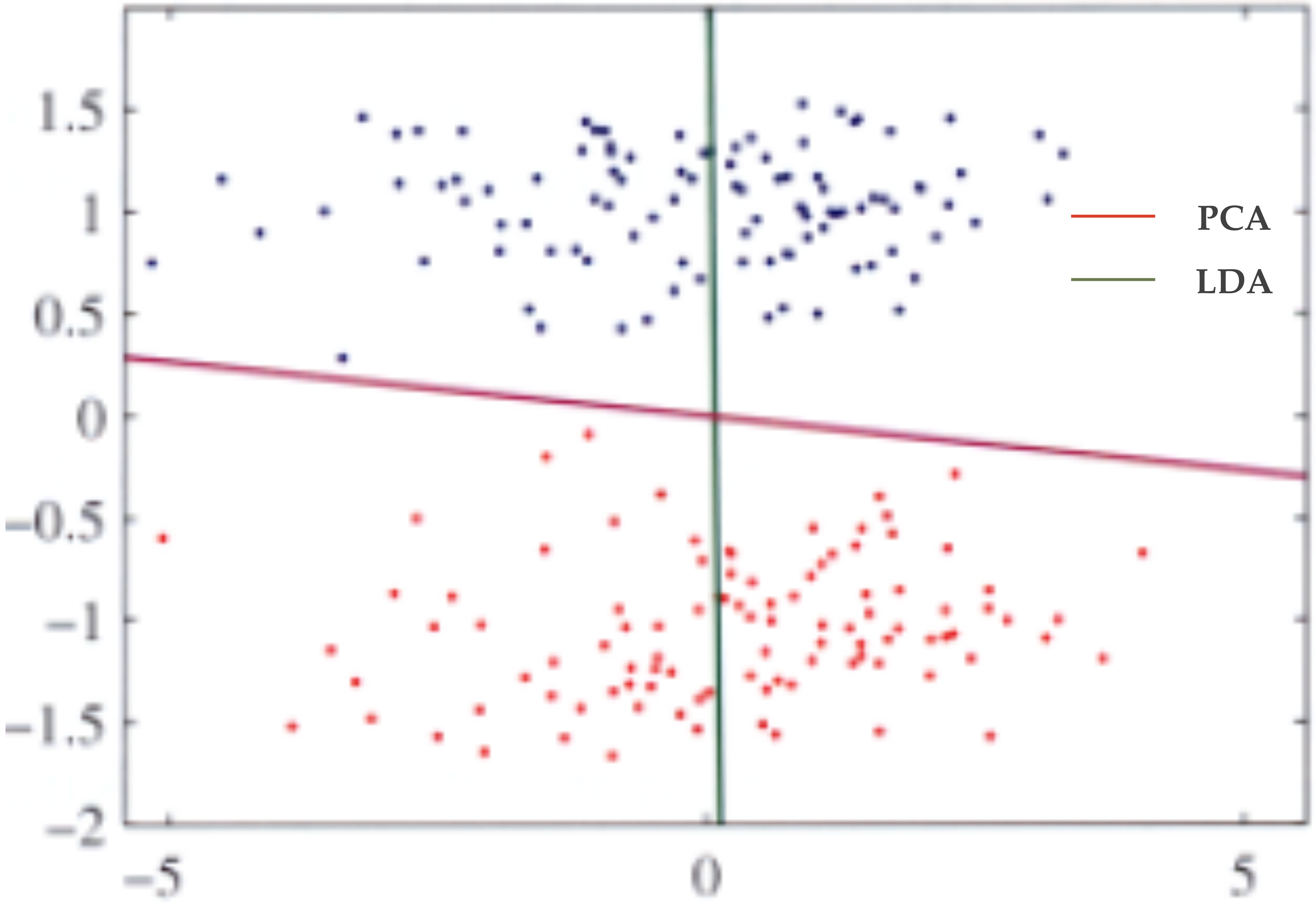


PRML - C. Bishop (Sec. 4.1.4, Sec. 4.1.6)

PCA VERSUS LDA



Linear Discriminant Analysis



PRML - C. Bishop (Sec. 4.1.4, Sec. 4.1.6)

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

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

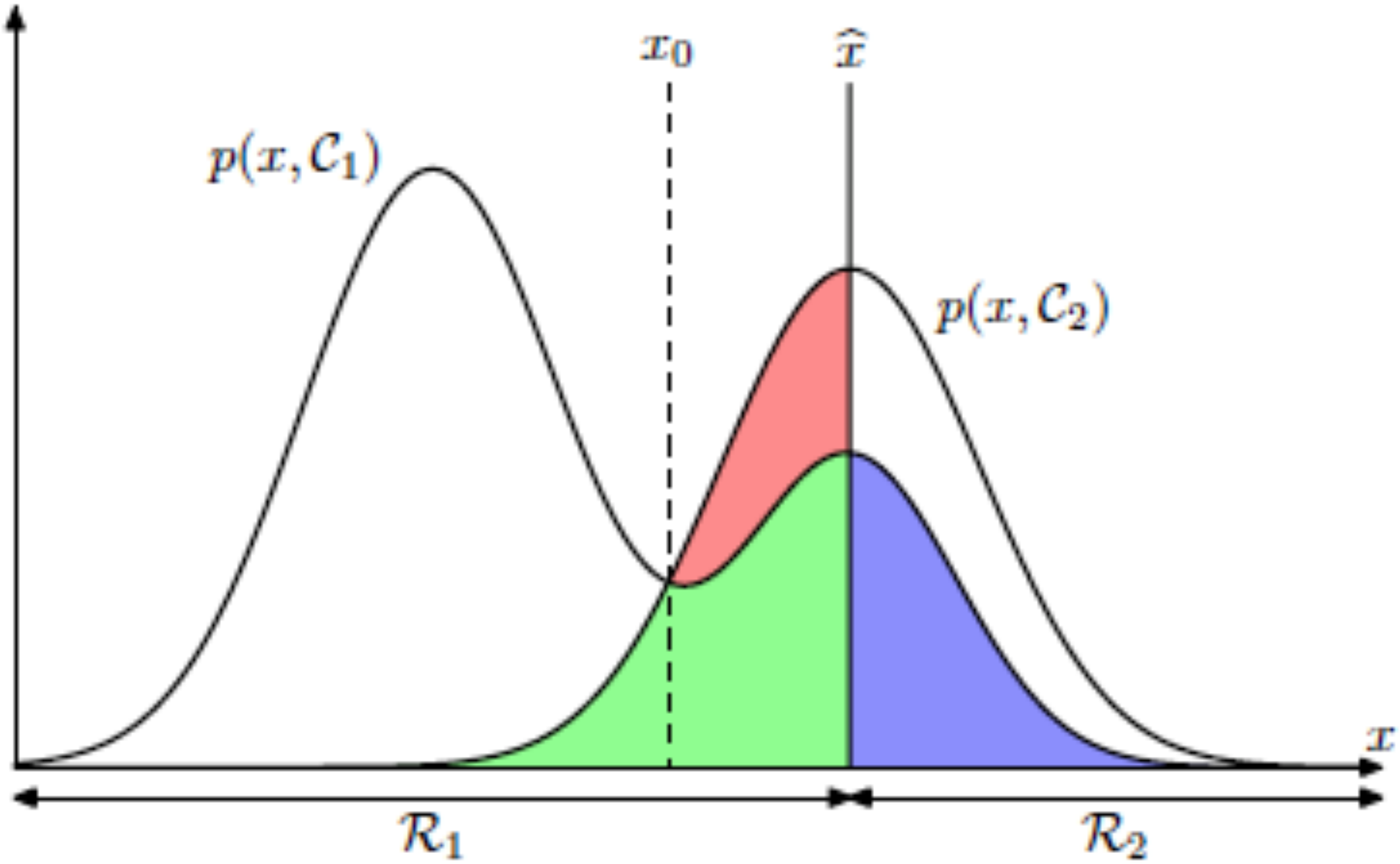
- ✓ Loss matrix
  - Minimizing the expected loss

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

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

LEAP

Visualizing the Max. Posterior Classifier



Approaches for Inference and Decision

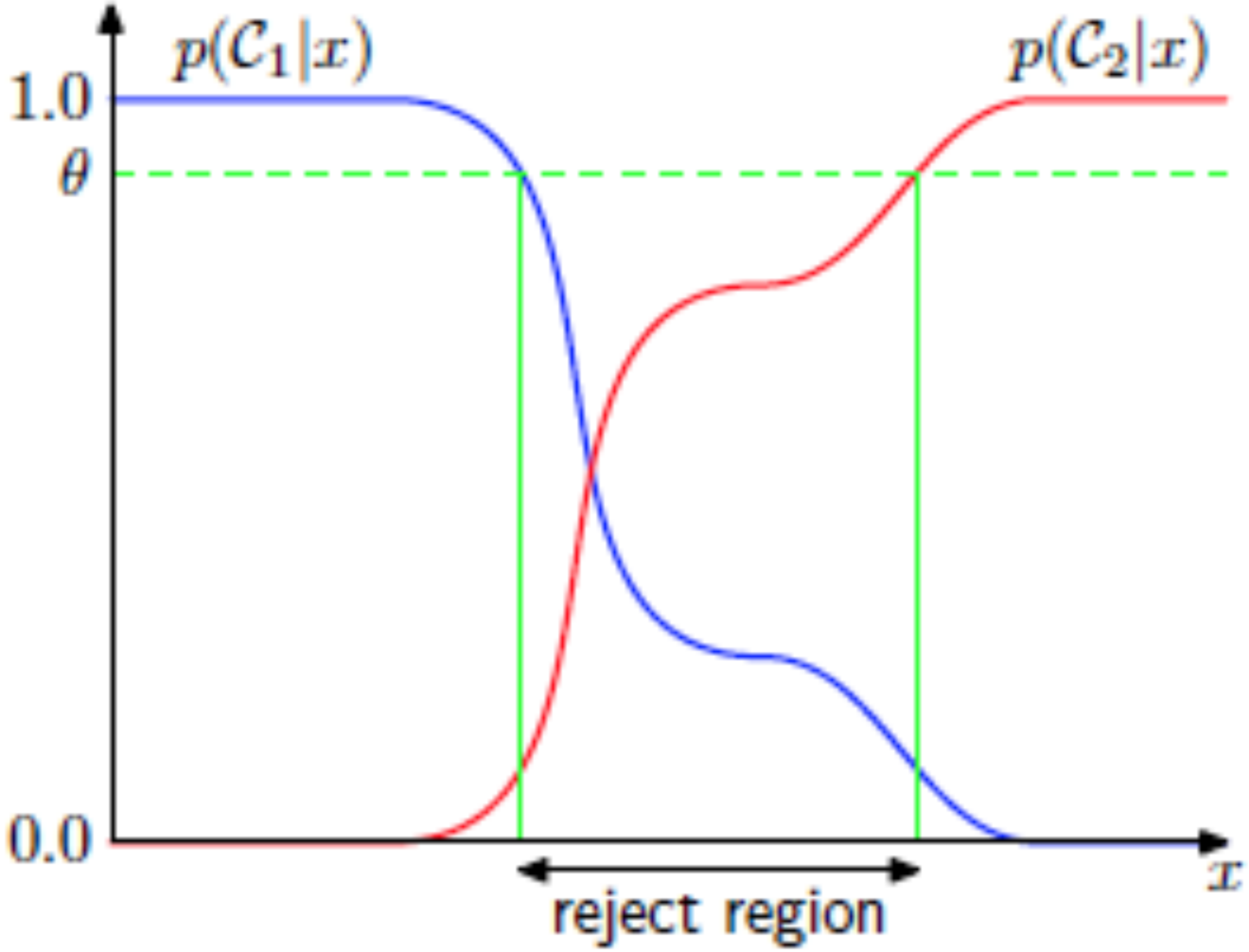
I. Finding the joint density from the data.

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

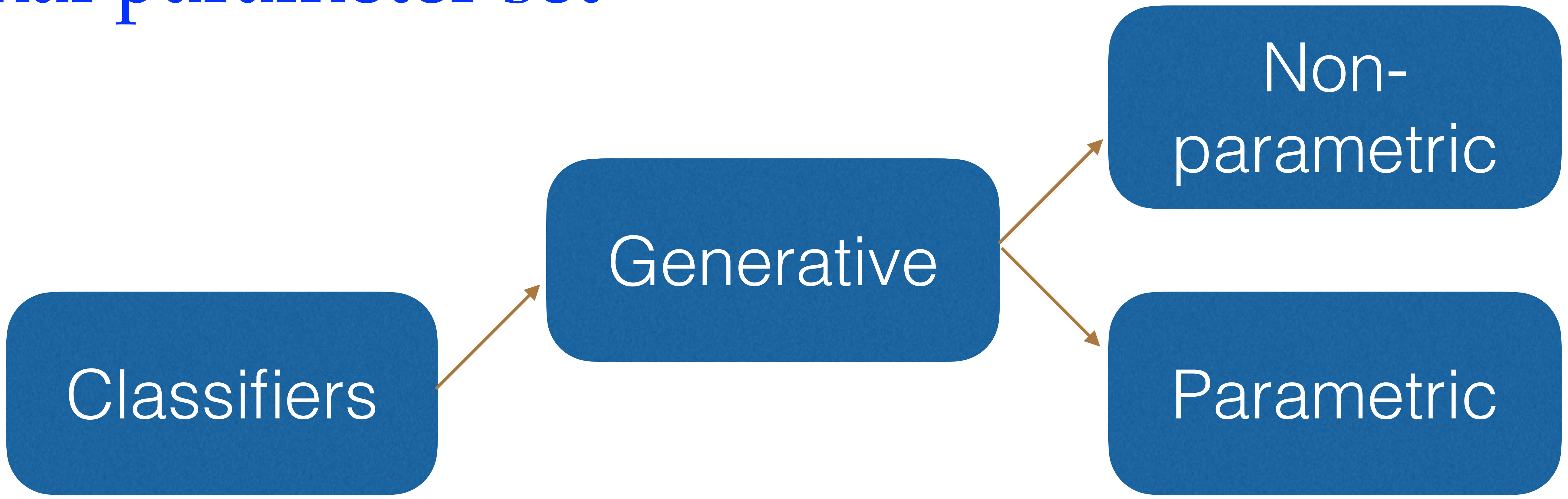


Decision Rule for Regression

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

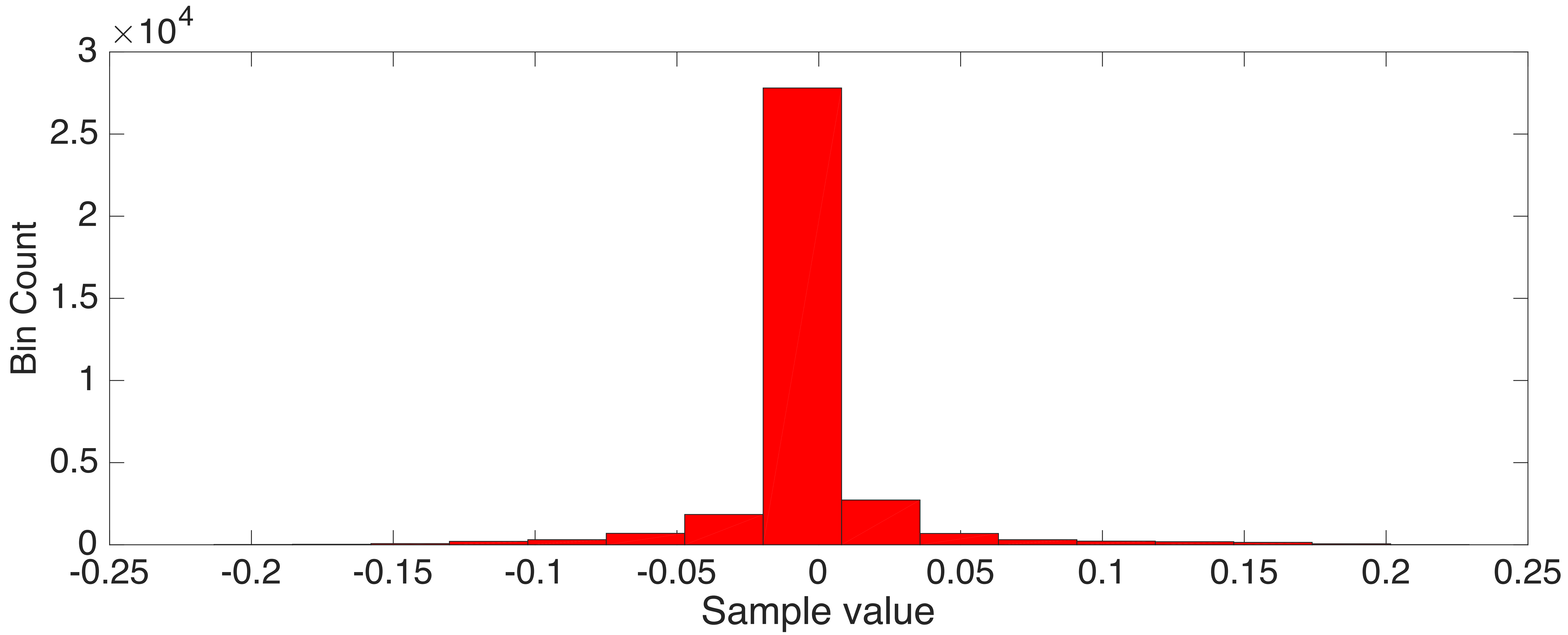
Generative Modeling

- 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



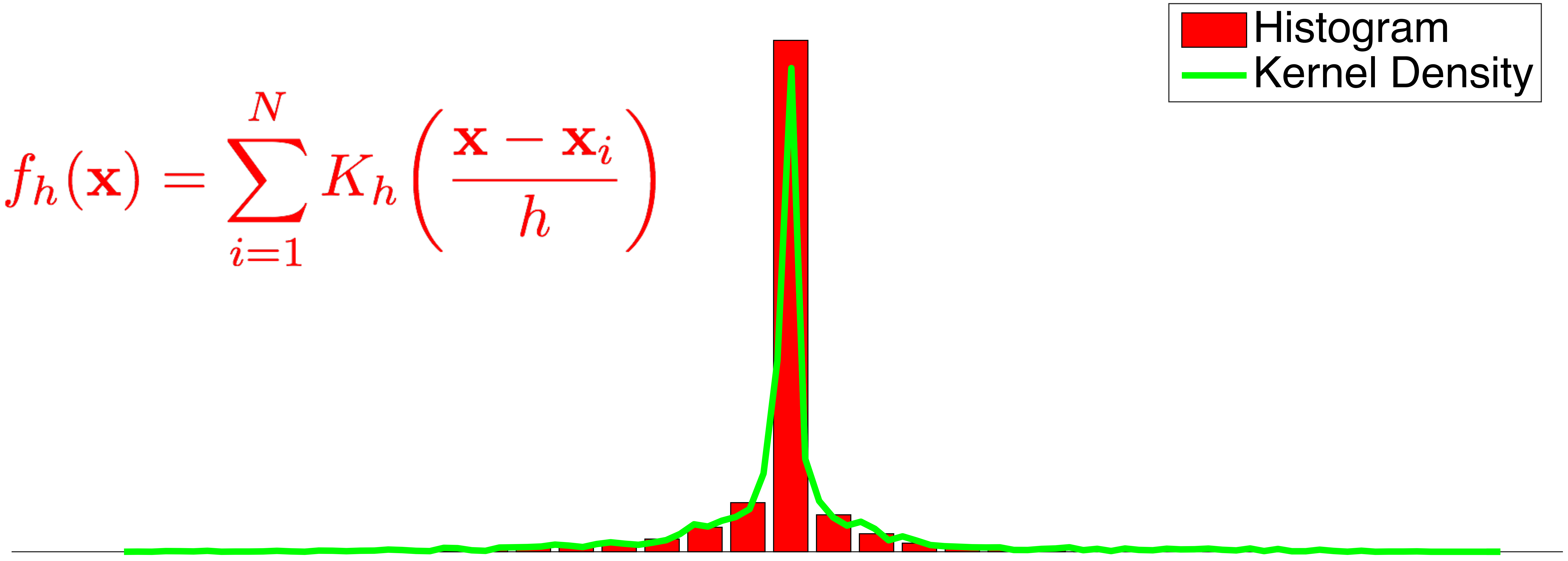
The density is not smooth and has block like shape.

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

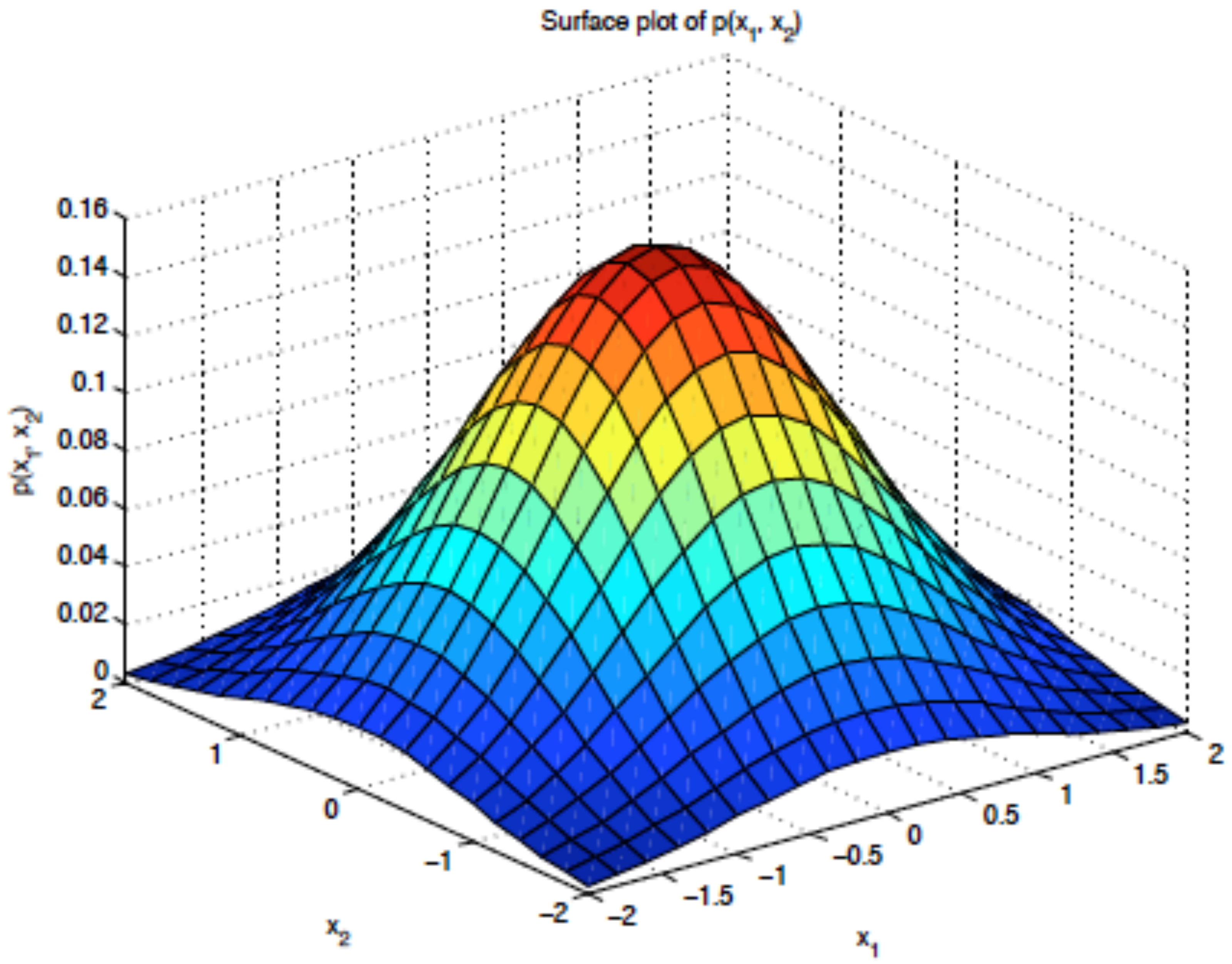
$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{x_i}$$

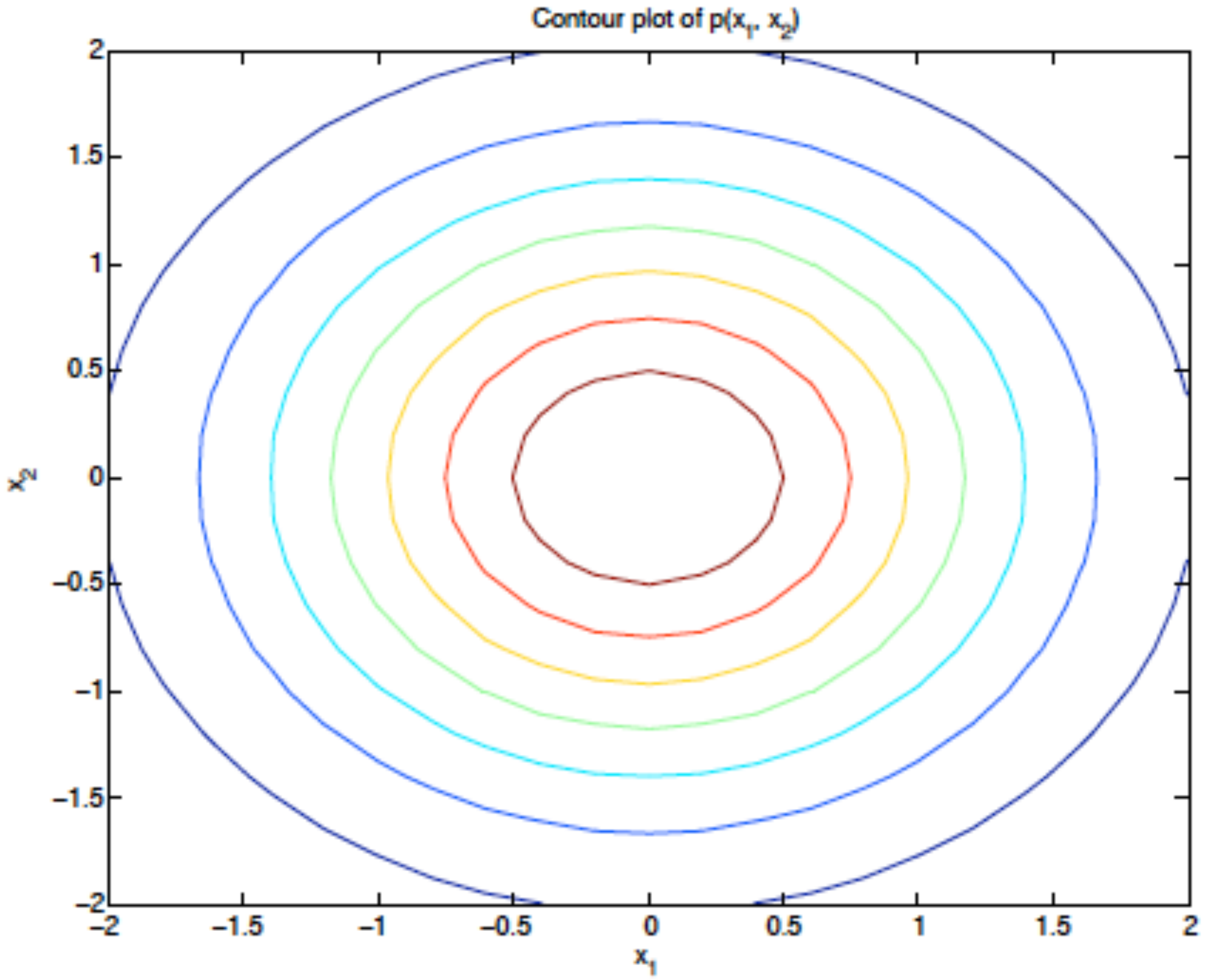
  - Gaussian Distribution

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

Gaussian Distribution

One of most widely used and well studied model

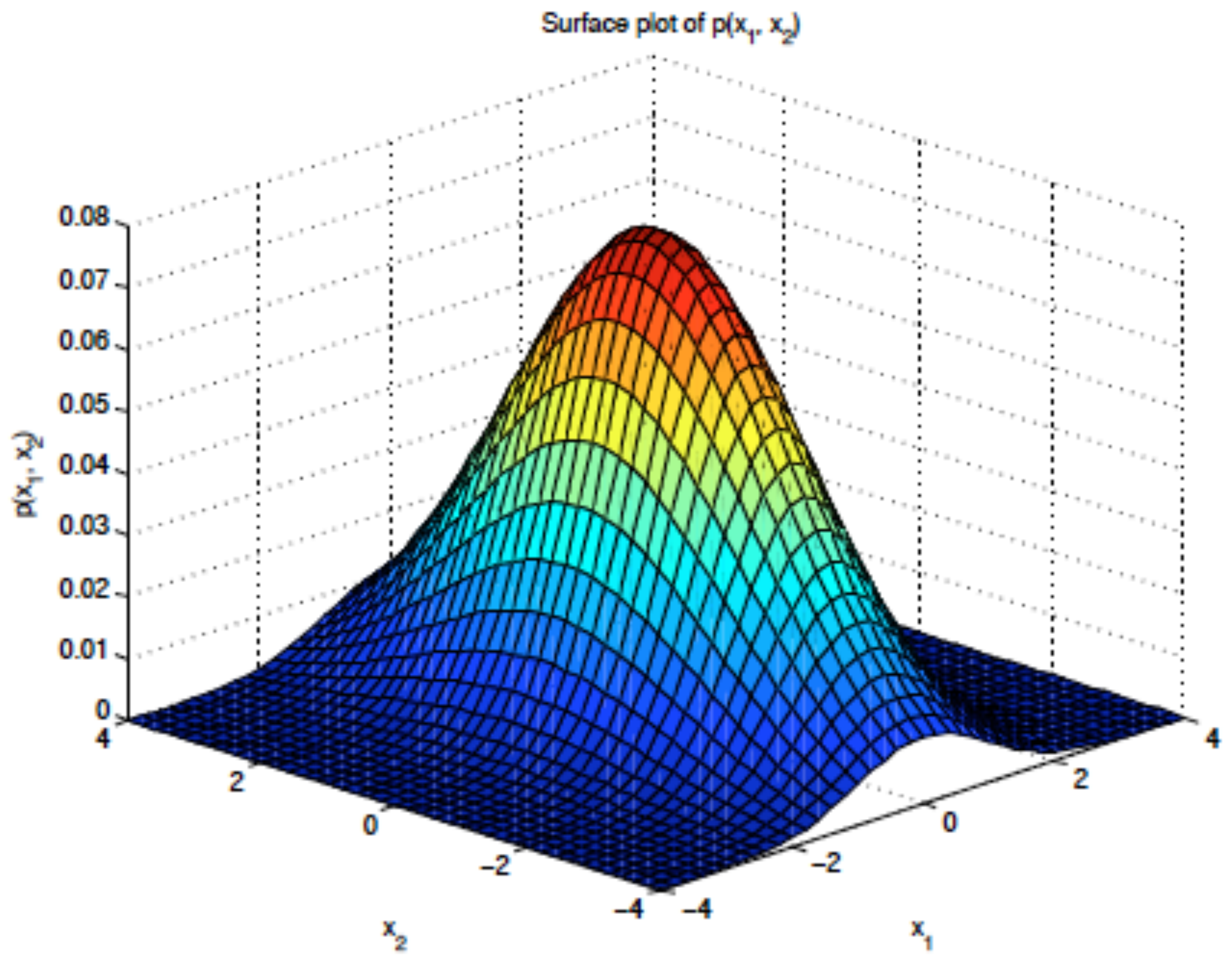


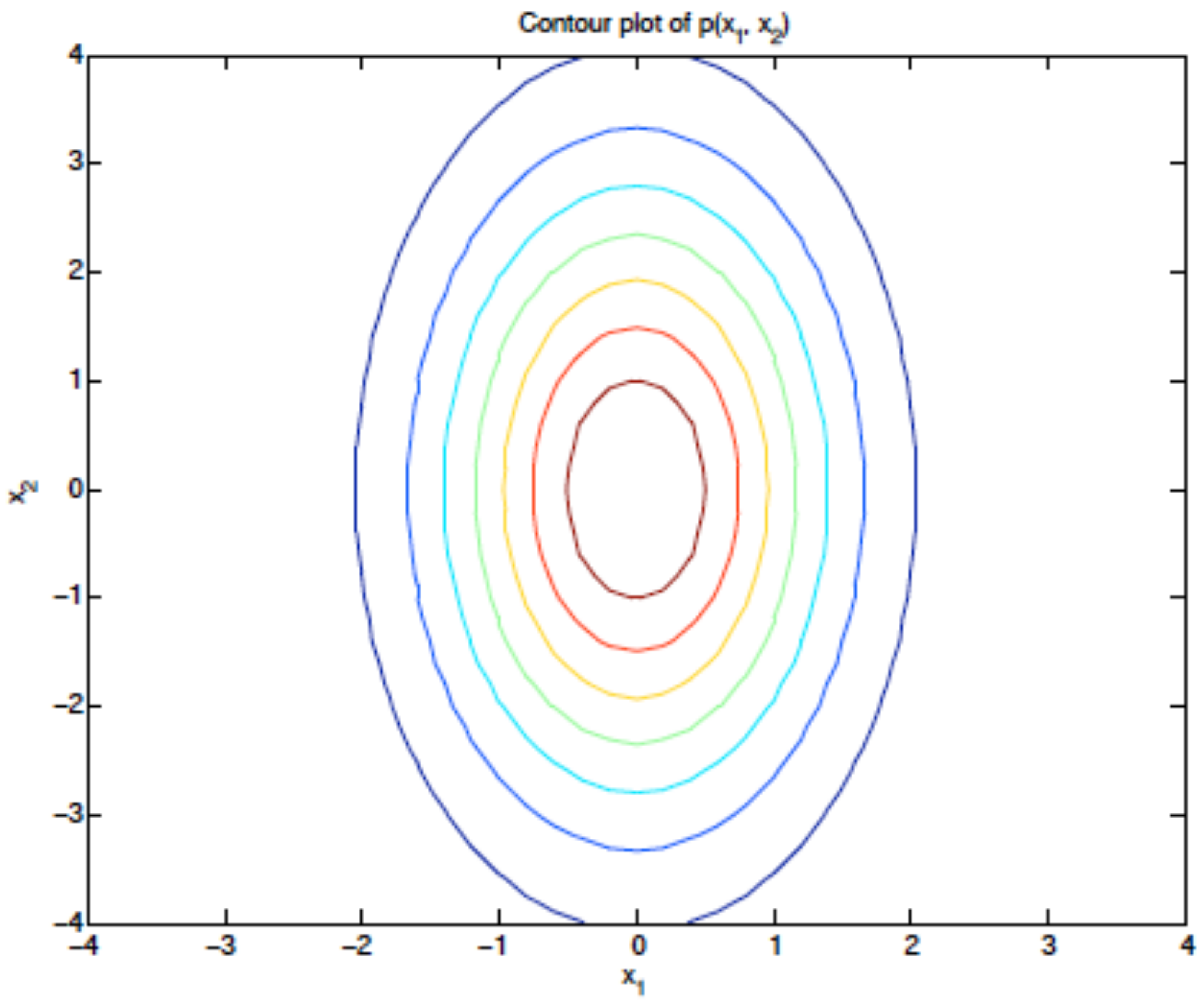


Points of equal probability lie on on contour
Diagonal Gaussian with Identical Variance

Gaussian Distribution

Insights into two dimensional Gaussian distribution

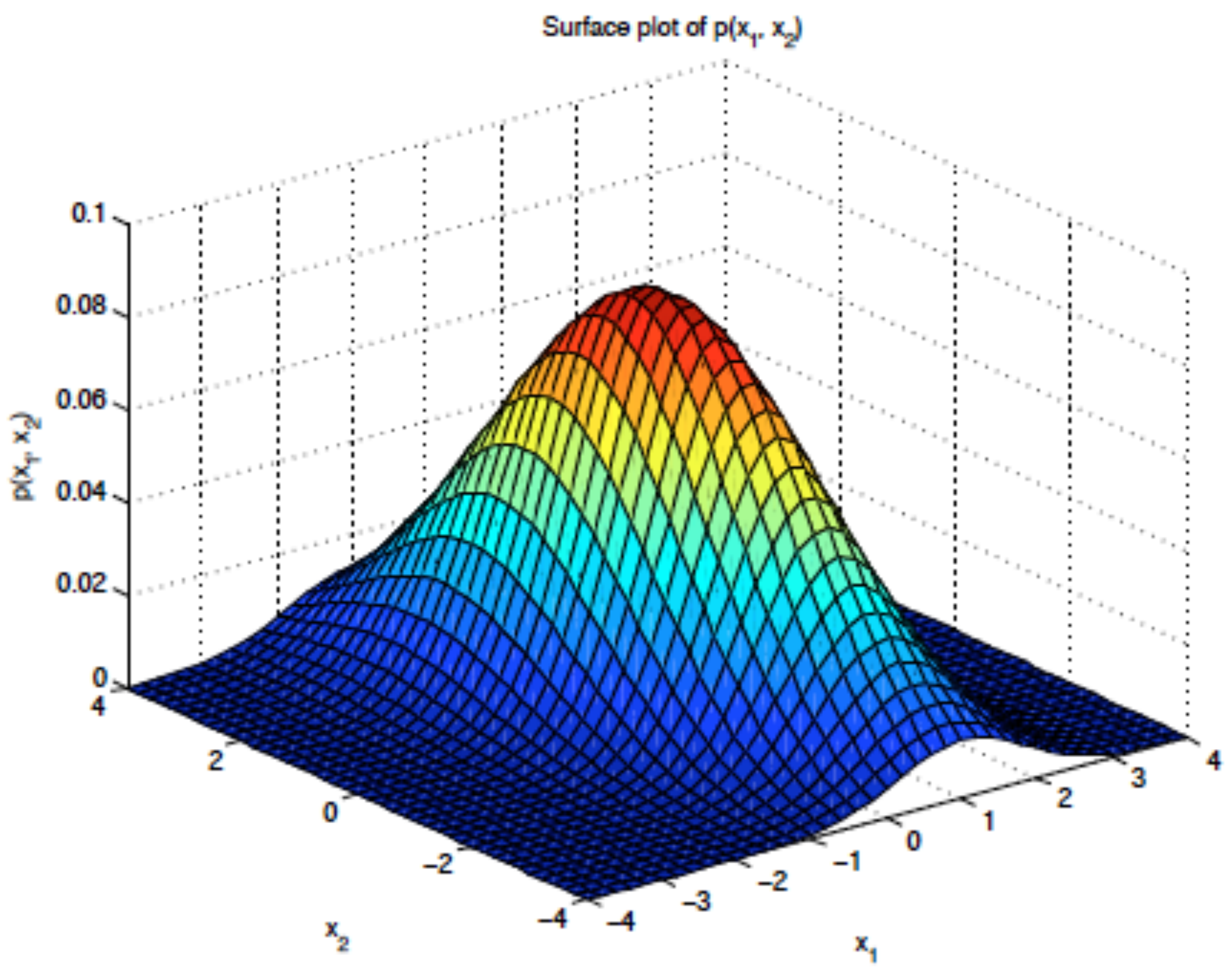


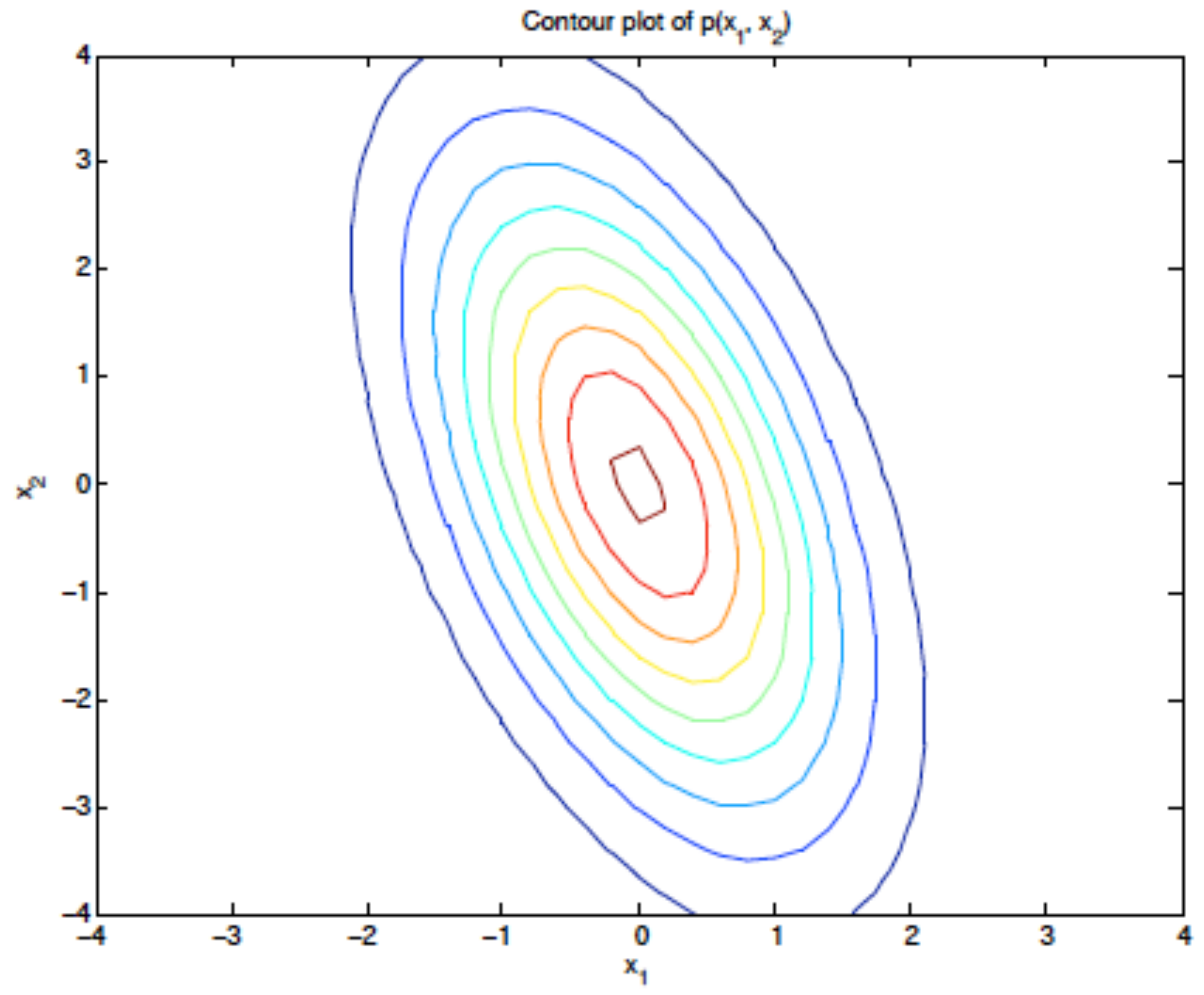


Diagonal Gaussian with different variance

Gaussian Distribution

Insights into two dimensional Gaussian Distribution

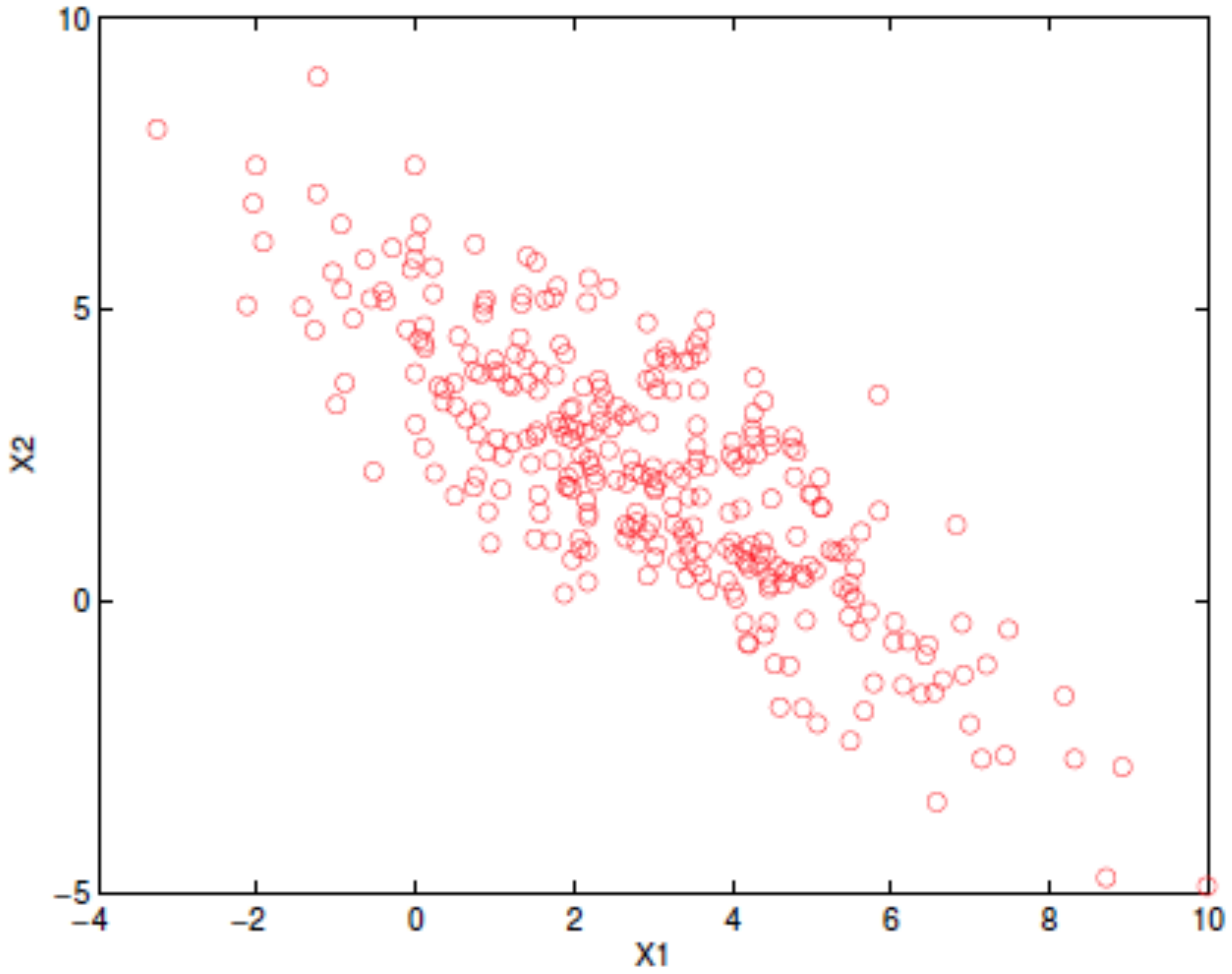


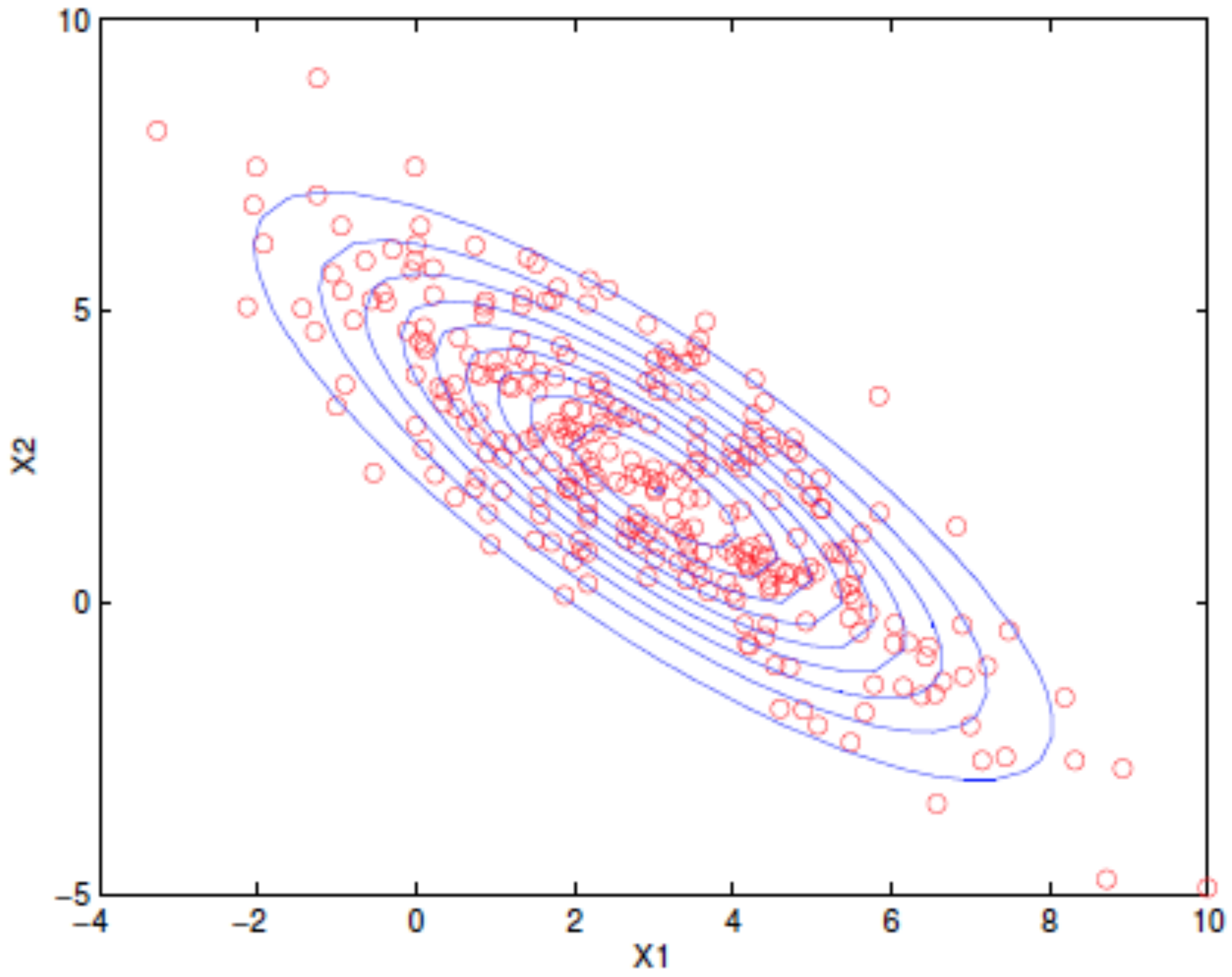


Full covariance Gaussian distribution

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

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

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

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

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

THANK YOU

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