Lecture 17: Clustering and density estimation

Professor Ilias Bilionis

Density estimation using Gaussian mixtures

PREDICTIVE
SCIENCE LABORATORY

Density estimation

Your are given n observations:

$$\mathbf{x}_{1:n} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$$

(inputs, features, …)

Problem: Assuming the observations are independent, find the probability density $p(\mathbf{x})$.

Mixtures of Gaussians

$\underline{\pi}, \underline{\mu}, \underline{\underline{\Sigma}}$ } must find

mean

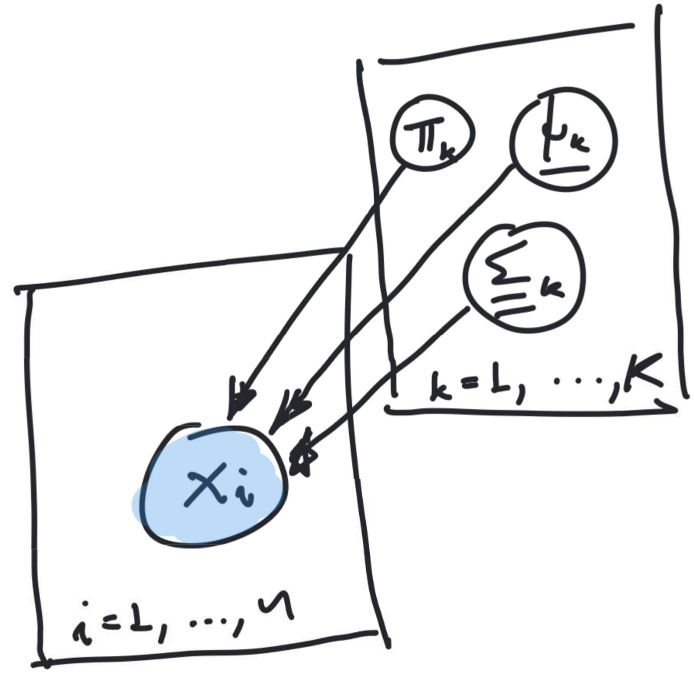
cov.

$$p(\underline{x} \mid \ldots) = \pi_1 N(\underline{x} \mid \underline{\mu}_1, \underline{\underline{\Sigma}}_1) + \pi_2 N(\underline{x} \mid \underline{\mu}_2, \underline{\underline{\Sigma}}_2) + \ldots + \pi_K N(\underline{x}, \underline{\mu}_K, \underline{\underline{\Sigma}}_K)$$

probability of sampling component 1

p. of sampling comp 2

restrictions: $\sum_{i=1}^{K} \pi_i = 1$, $\underline{\underline{\Sigma}}_i$ pos. def.



Training the model

Likelihood:

$$p(x_{1:n} \mid \underline{\pi}, \underline{\mu}, \underline{\underline{\Sigma}}) = \prod_{i=1}^{n} p(\underline{x}_i \mid \underline{\pi}, \underline{\mu}, \underline{\underline{\Sigma}})$$

$$= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k N(\underline{x}_i \mid \underline{\mu}_k, \underline{\underline{\Sigma}}_k) \right\}$$

log rules

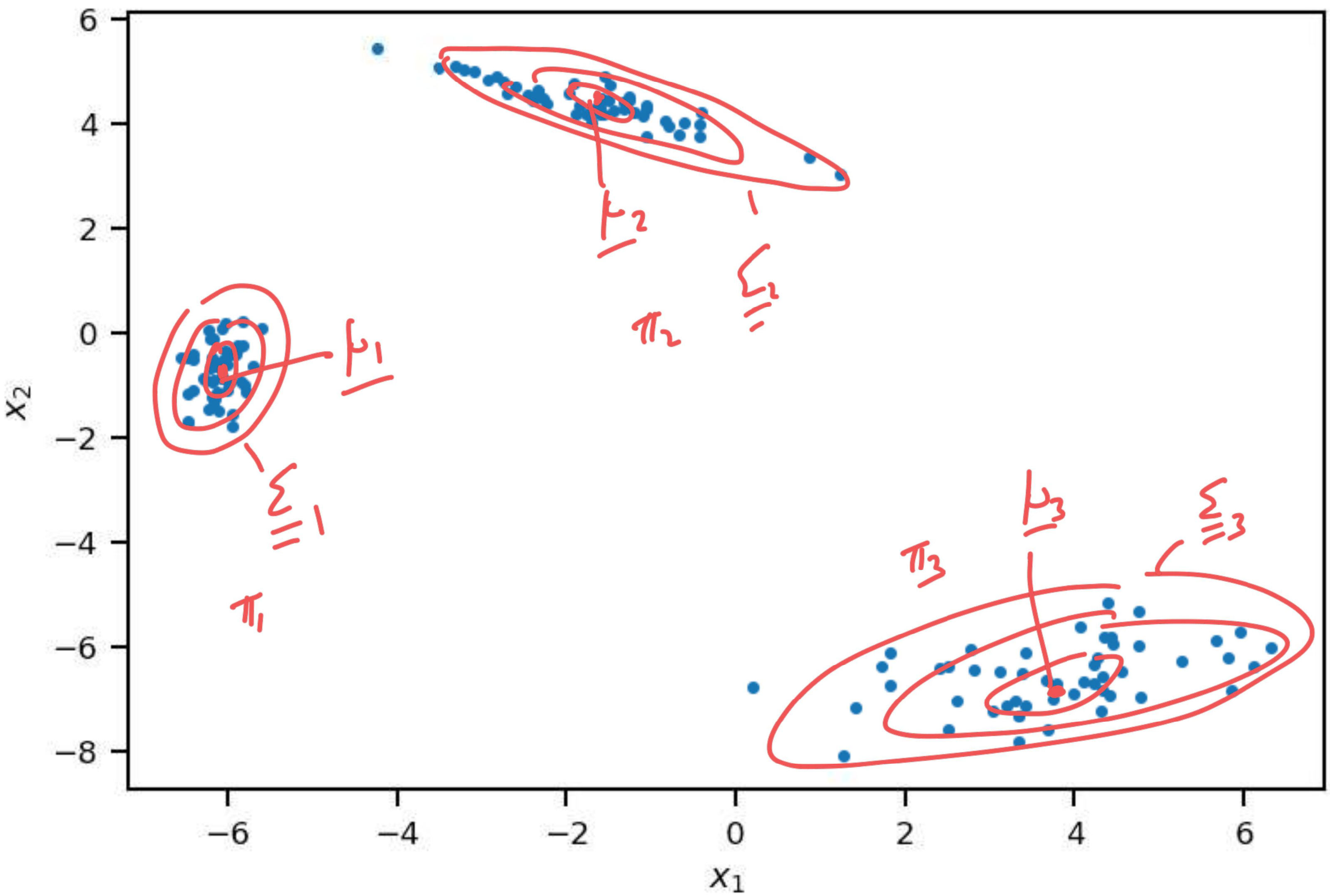
$$\max_{\underline{\pi}, \underline{\mu}, \underline{\underline{\Sigma}}} \log p(x_{1:n} \mid \underline{\pi}, \underline{\mu}, \underline{\underline{\Sigma}}) = \sum_{i=1}^{n} \log\left(\quad \ldots \quad \right)$$

Expectation-Maximization

Bishop 2006.

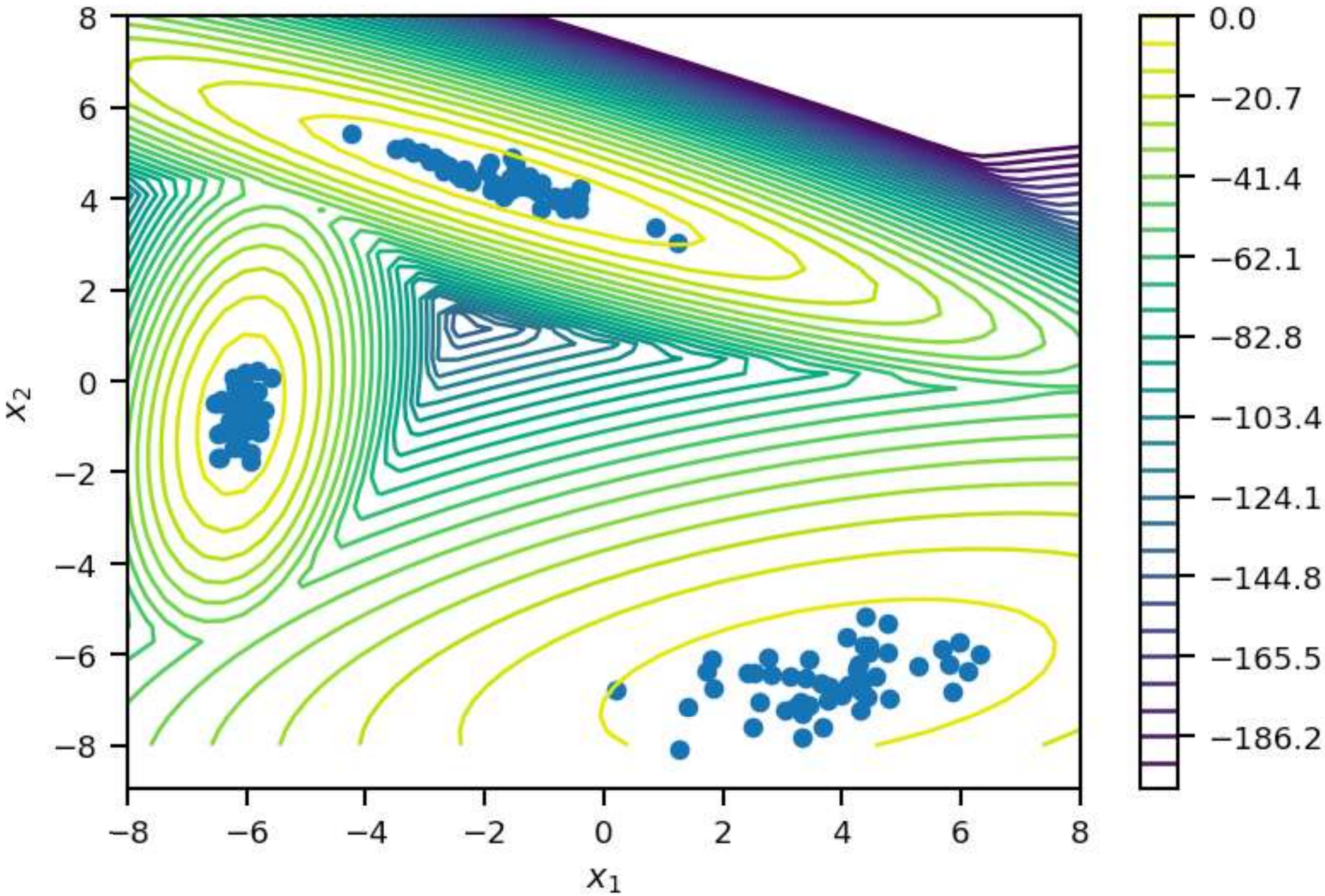
PREDICTIVE
SCIENCE LABORATORY

Example



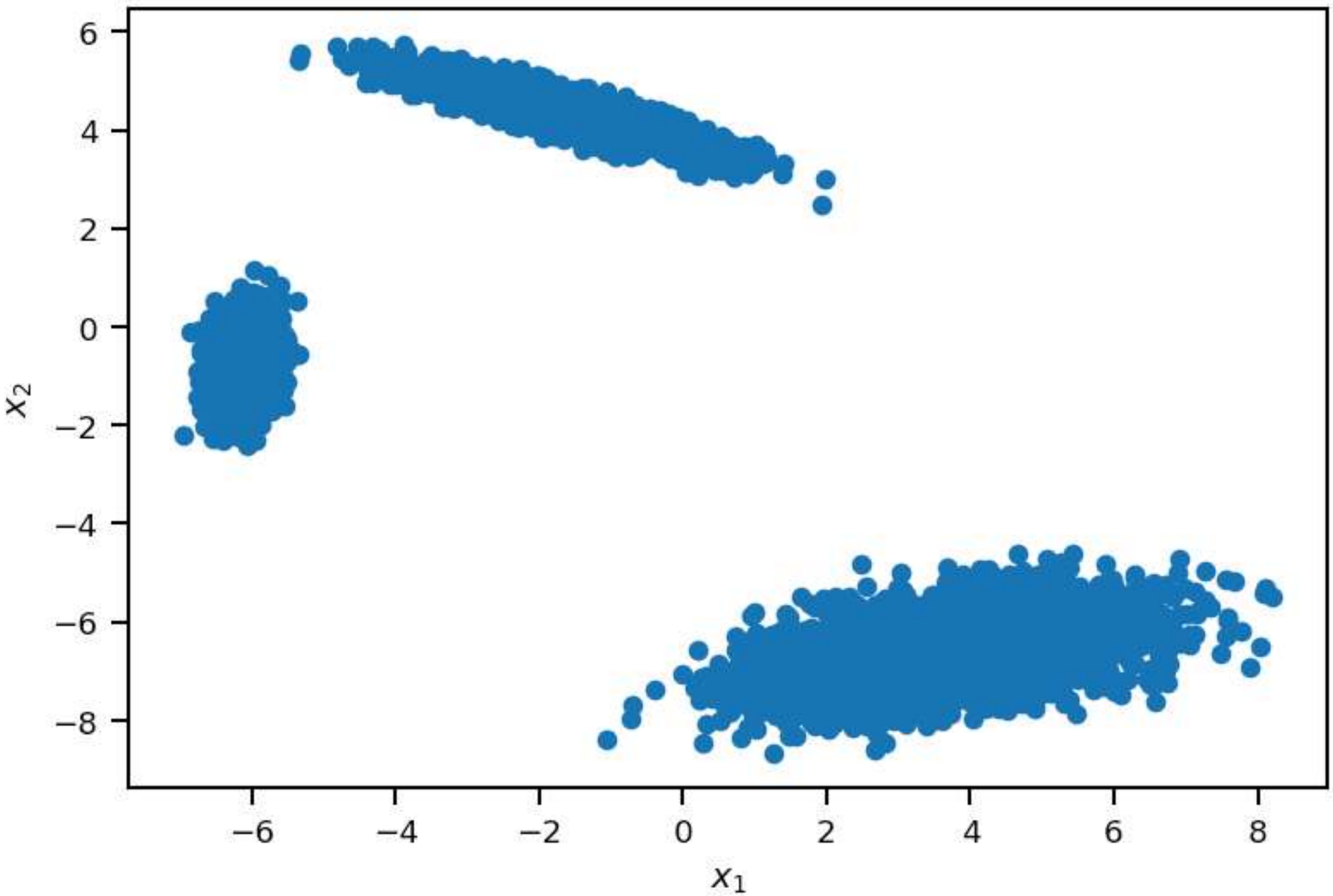
PREDICTIVE
SCIENCE LABORATORY

Log probability density: Density estimation with mixtures



PREDICTIVE
SCIENCE LABORATORY

You can sample from the density



PREDICTIVE
SCIENCE LABORATORY

You can use the model to do clustering

$$p(\underline{x} \text{ belongs to } k^{th} \text{ comp.} \mid \underline{\pi}, \underline{\Sigma}, \underline{\mu}) \propto \pi_k \cdot N(\underline{x} \mid \underline{\mu}_k, \underline{\underline{\Sigma}}_k)$$

$$= \frac{\pi_k N(\underline{x} \mid \underline{\mu}_k, \underline{\underline{\Sigma}}_k)}{\sum_{k'=1}^{K} \pi_{k'} N(\underline{x} \mid \underline{\mu}_{k'}, \underline{\underline{\Sigma}}_{k'})}$$

to make a hard-assignment to a cluster, take the $\arg\max_k (\cdot)$

PREDICTIVE SCIENCE LABORATORY