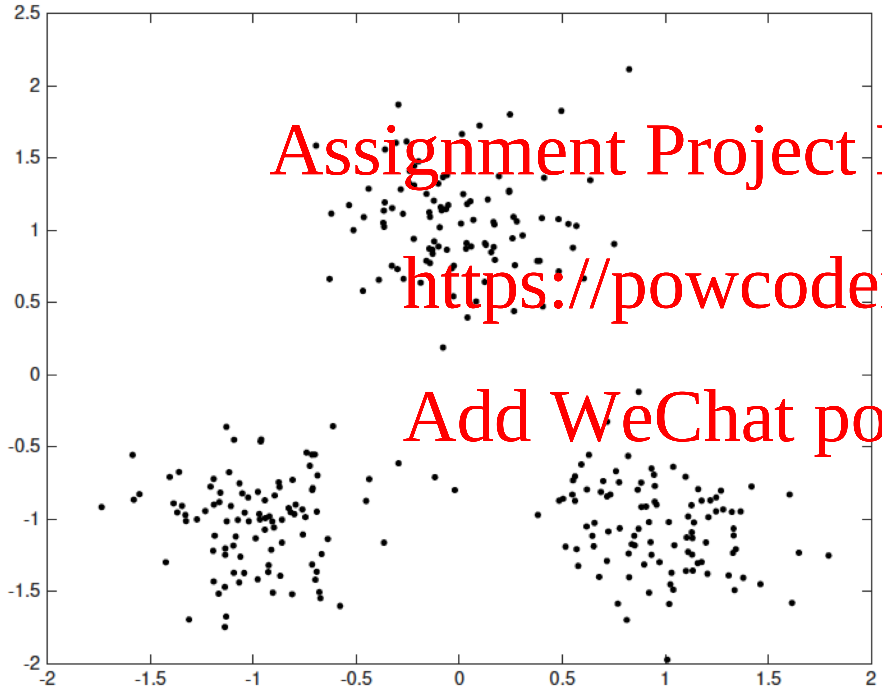
Probabilistic modelling



How could have this data been generated?

Can you fit models to the data?

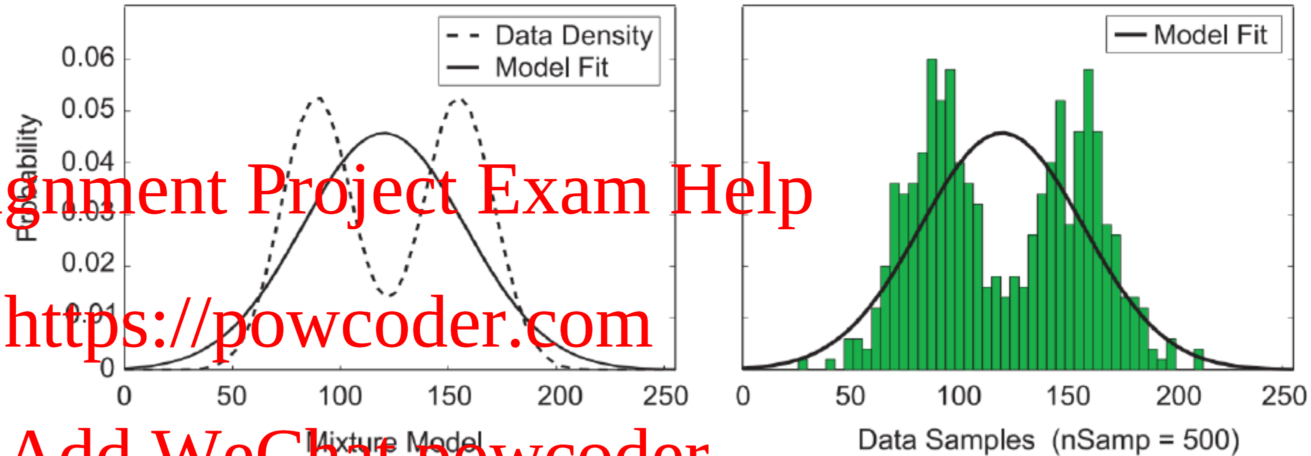
UNIVERSITY OF BIRMINGHAM

Gaussian mixture models (GMMs)

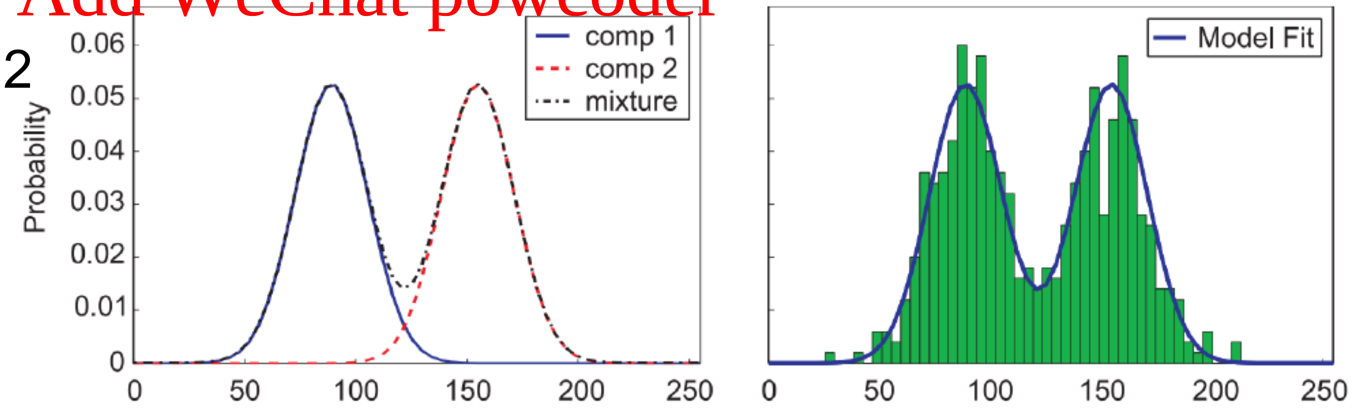
- Assume data was generated by a set of Gaussian distributions.
- The probability density is a mixture of them.
- Find the parameters of the Gaussian distributions and how much each distribution contributes to the data.
- This is a mixture model of Gaussian.

Visualizing GMMs - 1D Gaussians

If you fit one Gaussian



Now we try a GMM with 2 Gaussians
(each contribute 50%)



UNIVERSITY OF BIRMINGHAM

Generative Models

- In supervised learning, we model the joint distribution

$$p(\mathbf{x}, z) = p(\mathbf{x}|z)p(z)$$

- In unsupervised learning, we do not have labels z, we model

$$p(\mathbf{x}) = \sum_z p(\mathbf{x}, z) = \sum_z p(\mathbf{x}|z)p(z)$$

Hidden/Latent variables

UNIVERSITY OF BIRMINGHAM

Gaussian mixture models (GMMs)

- A GMM represents a distributions as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

- with π_k the mixing coefficients, where

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{and} \quad \pi_k \geq 0 \quad \forall k$$

- GMM is a density estimator.
- GMM is universal approximators of densities (if you have enough Gaussians)

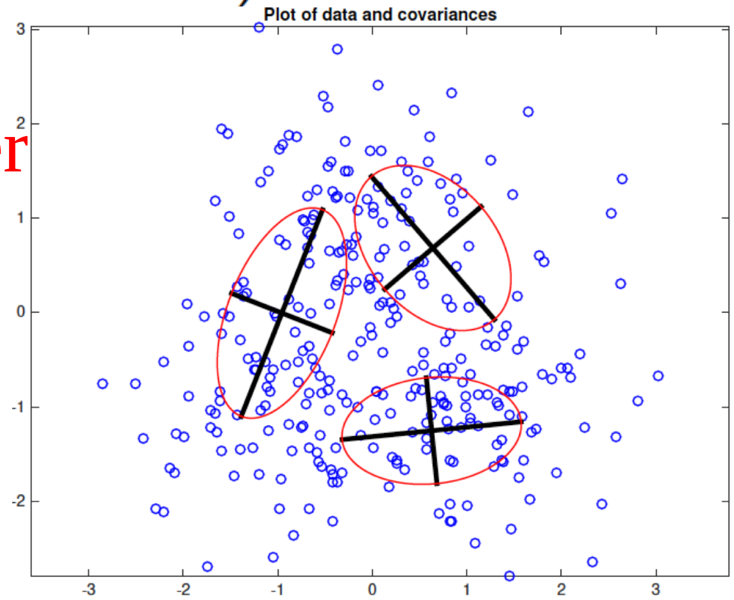
UNIVERSITY OF
BIRMINGHAM

Fitting GMMs: Maximum likelihood and EM

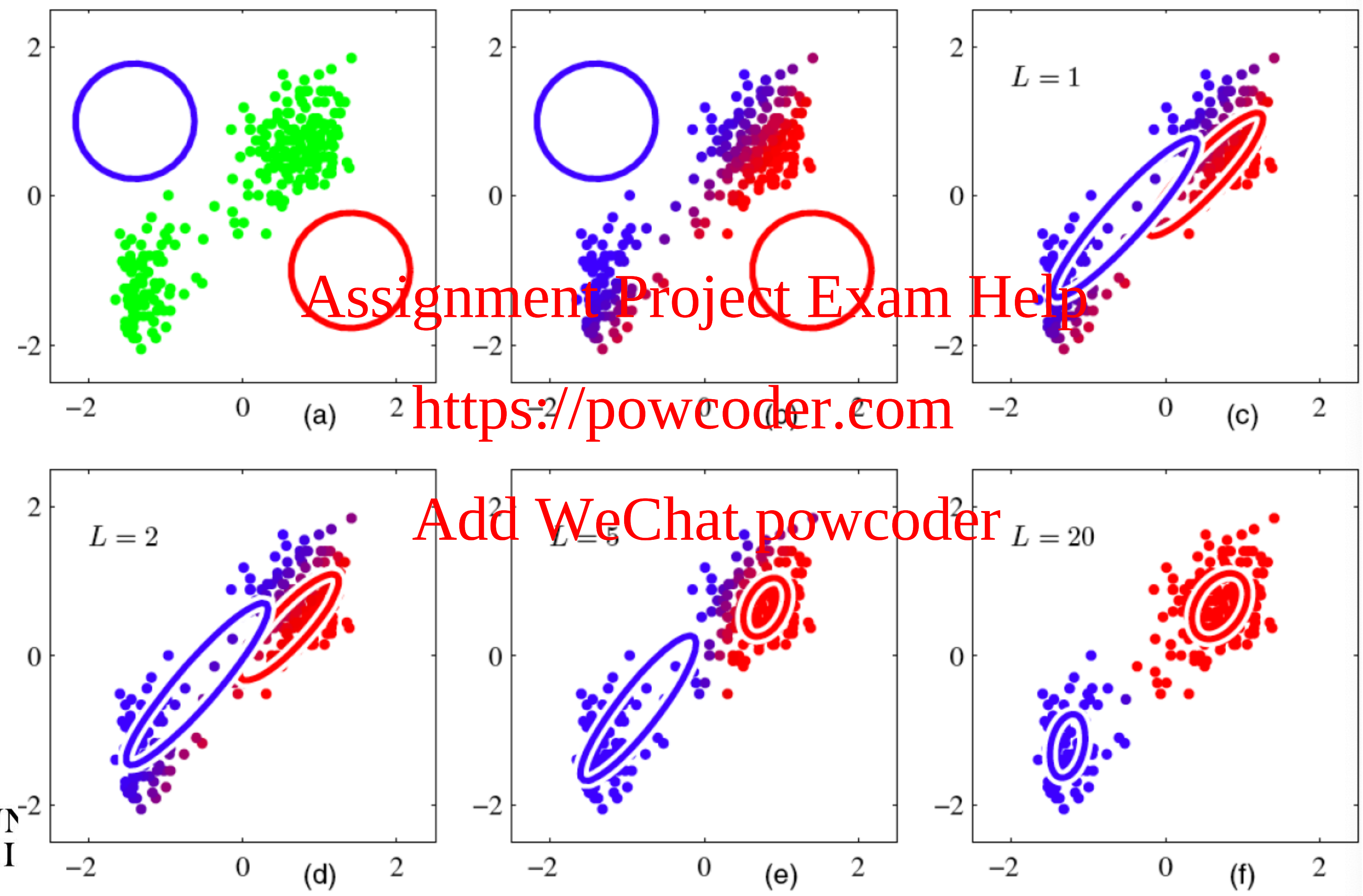
- To have a model best fit data, we need to maximize the (log) likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}^{(n)}|\mu_k, \Sigma_k) \right)$$

- Expectation
  if we knew π_k, μ and Σ, we can get "soft" Z_k
  $P(z^{(n)}|x)$ - responsibility
- Maximization
  if we know Z_k, we can get π_k, μ and Σ



UNIVERSITY OF
BIRMINGHAM



UN
BI

Expectation-Maximization (EM Algorithm)

An optimization process that alternates between 2 steps:

1. E-step: compute the posterior probability over z given the current model.

$$\gamma_k^{(n)} = p(z^{(n)}|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}^{(n)}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}^{(n)}|\mu_j, \Sigma_j)}$$ responsibility

- Which Gaussian generate each data point with how much possibility?

Expectation-Maximization (EM Algorithm)

2. M-step: Assuming data was really generated this way, change the parameters of each Gaussian to maximize the probability that it would generate the data it is currently responsible for.

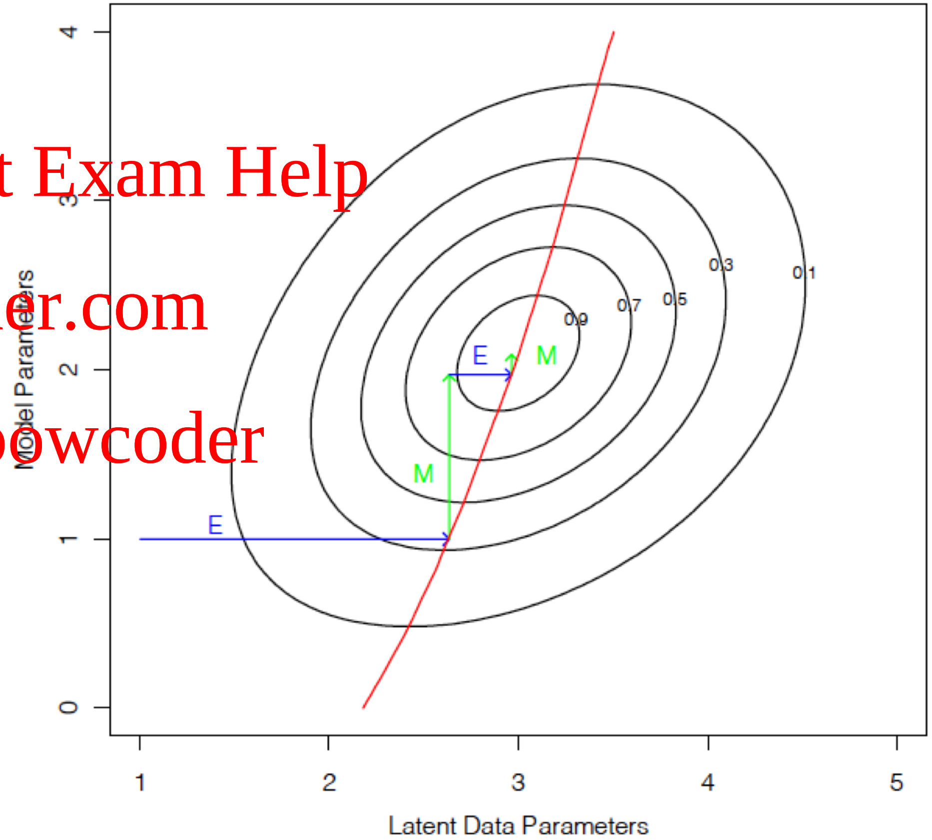
$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_k^{(n)}\mathbf{x}^{(n)}$$

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_k^{(n)}(\mathbf{x}^{(n)}-\mu_k)(\mathbf{x}^{(n)}-\mu_k)^T$$

$$\pi_k = \frac{N_k}{N} \quad \text{with} \quad N_k = \sum_{n=1}^{N}\gamma_k^{(n)}$$

Expectation-Maximization (EM Algorithm)

- A general algorithm for optimizing many latent variable models (not just for GMMs).
- Iteratively computes a lower bound then optimizes it.
- Converges but maybe to a local minima.
- Can use multiple restarts.



Elements of Statistical Learning (2nd edition)

UNIVERSITY OF BIRMINGHAM

Summary

- Clustering

  group similar data points

  need a distance measure
- Agglomerative hierarchical clustering

  successively merges similar groups of points

  build a dendrogram (binary tree)

  different ways to measure distance between clusters.
- GMM using EM

  build a generative model based on Gaussian distributions

  need to pre-define k (number of clusters)

  Using EM to find the best fit of the model.