Data Mining and Machine Learning

Statistical Modelling (2)

Peter Jančovič

UNIVERSITY OF BIRMINGHAM

Objectives

- In part 1 of this topic we
  - Reviewed univariate Gaussian PDF
  - Introduced multivariate Gaussian PDF
  - Introduced maximum likelihood (ML) estimation of Gaussian PDF parameters

- In this part, we will
  - Introduce Gaussian Mixture Models (GMMs)
  - Introduce ML estimation of GMM parameters

Fitting a Gaussian PDF to Data

- Suppose $y = y_1, \ldots, y_t, \ldots, y_T$ is a set of T data values
- For a Gaussian PDF p with mean μ and variance σ, define:

$$p(y \mid \mu, \sigma) = \prod_{t=1}^{T} p(y_t \mid \mu, \sigma)$$

- The ‘best fitting’ Gaussian maximises $p(y|\mu,\sigma)$
- Maximising $p(y|\mu,\sigma)$ with respect to μ,σ is called <u>Maximum Likelihood (ML)</u> estimation of μ,σ

$$\mu = \frac{1}{T}\sum_{t=1}^{T} y_t, \quad \sigma = \frac{1}{T}\sum_{t=1}^{T}(y_t - \mu)^2$$

Multi-modal distributions

- In practice the distributions of many naturally occurring phenomena do not follow the simple bell-shaped Gaussian curve
- For example, if the data arises from several difference sources, there may be several distinct peaks (e.g. distribution of heights of adults)
- These peaks are the modes of the distribution and the distribution is called multi-modal

Gaussian Mixture PDFs

- Gaussian Mixture PDFs, or Gaussian Mixture Models (GMMs) used to model multi-modal and other non-Gaussian distributions.

- A GMM is just a weighted average of several Gaussian PDFs, called the <u>component</u> PDFs

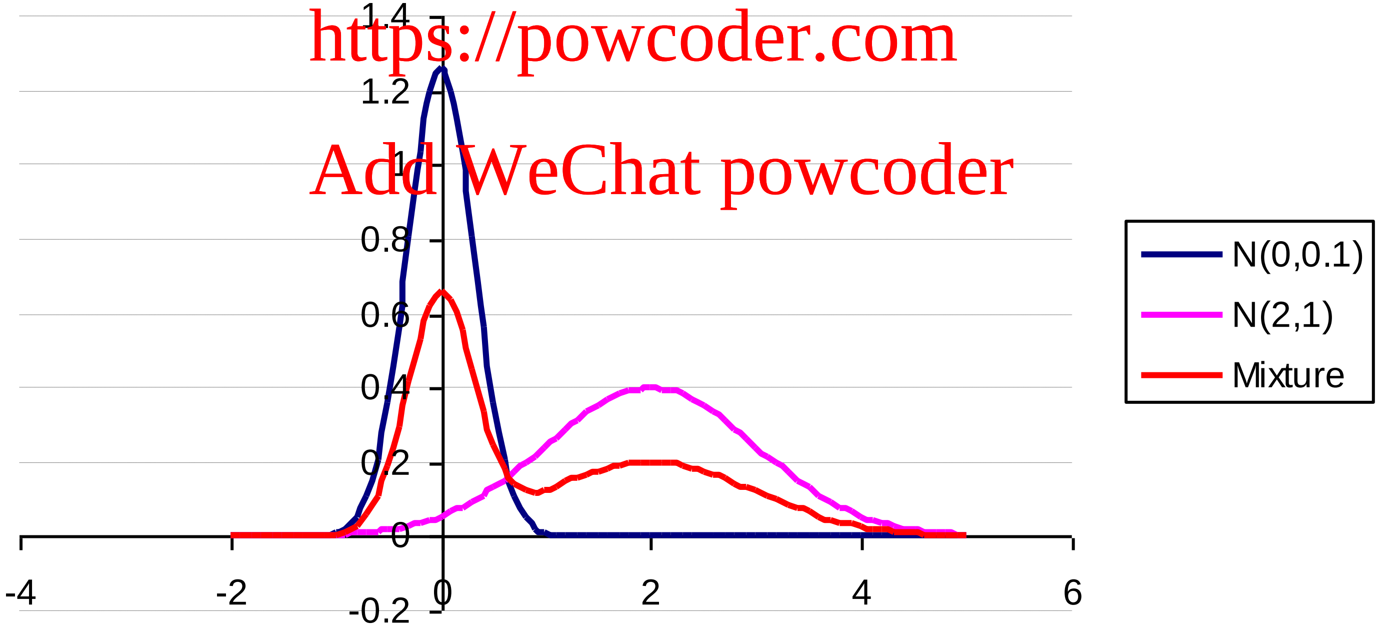
- For example, if p_1 and p_2 are Gaussian PDFs, then

$$p(y) = w_1 p_1(y) + w_2 p_2(y)$$

defines a 2 component Gaussian mixture PDF

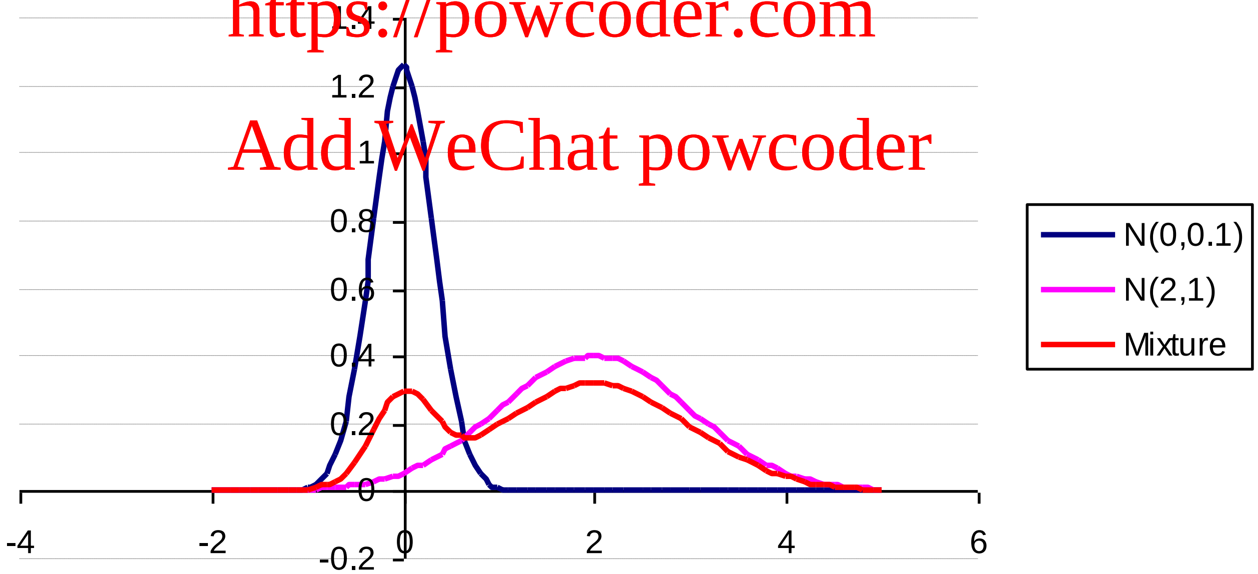
Gaussian Mixture - Example

- 2 component mixture model
  - Component 1: $\mu=0$, $\sigma=0.1$
  - Component 2: $\mu=2$, $\sigma=1$
  - $w_1 = w_2 = 0.5$



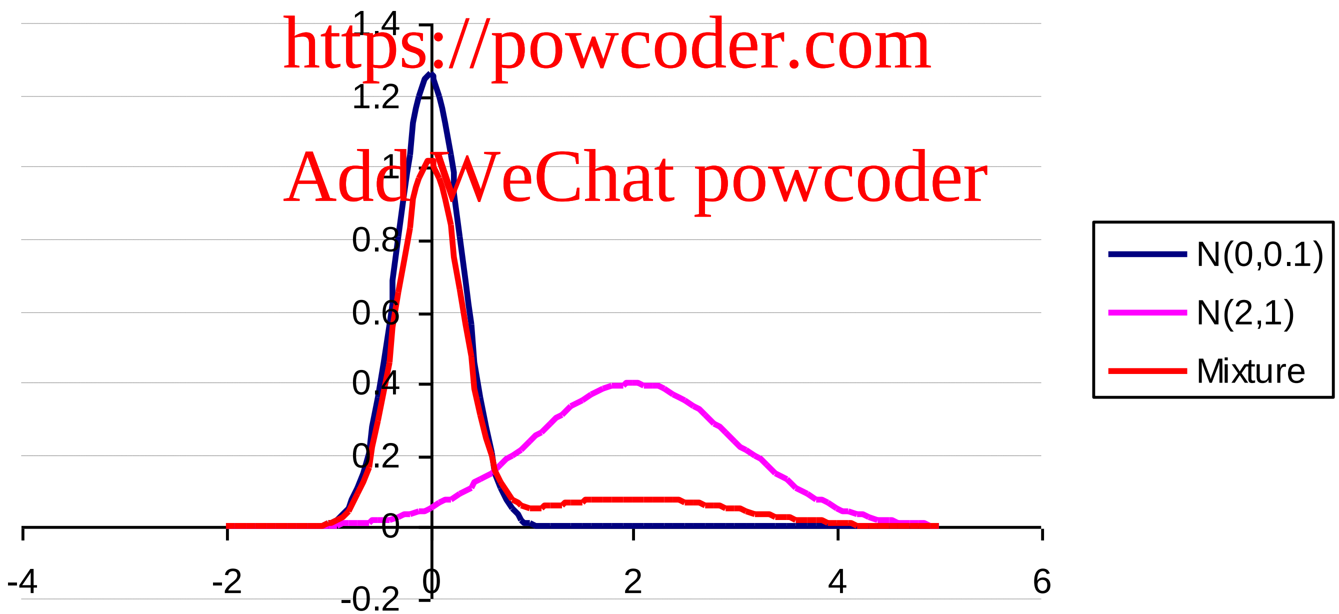
Example 2

- 2 component mixture model
  - Component 1: $\mu=0$, $\sigma=0.1$
  - Component 2: $\mu=2$, $\sigma=1$
  - $w_1 = 0.2$ $w_2=0.8$



Example 3

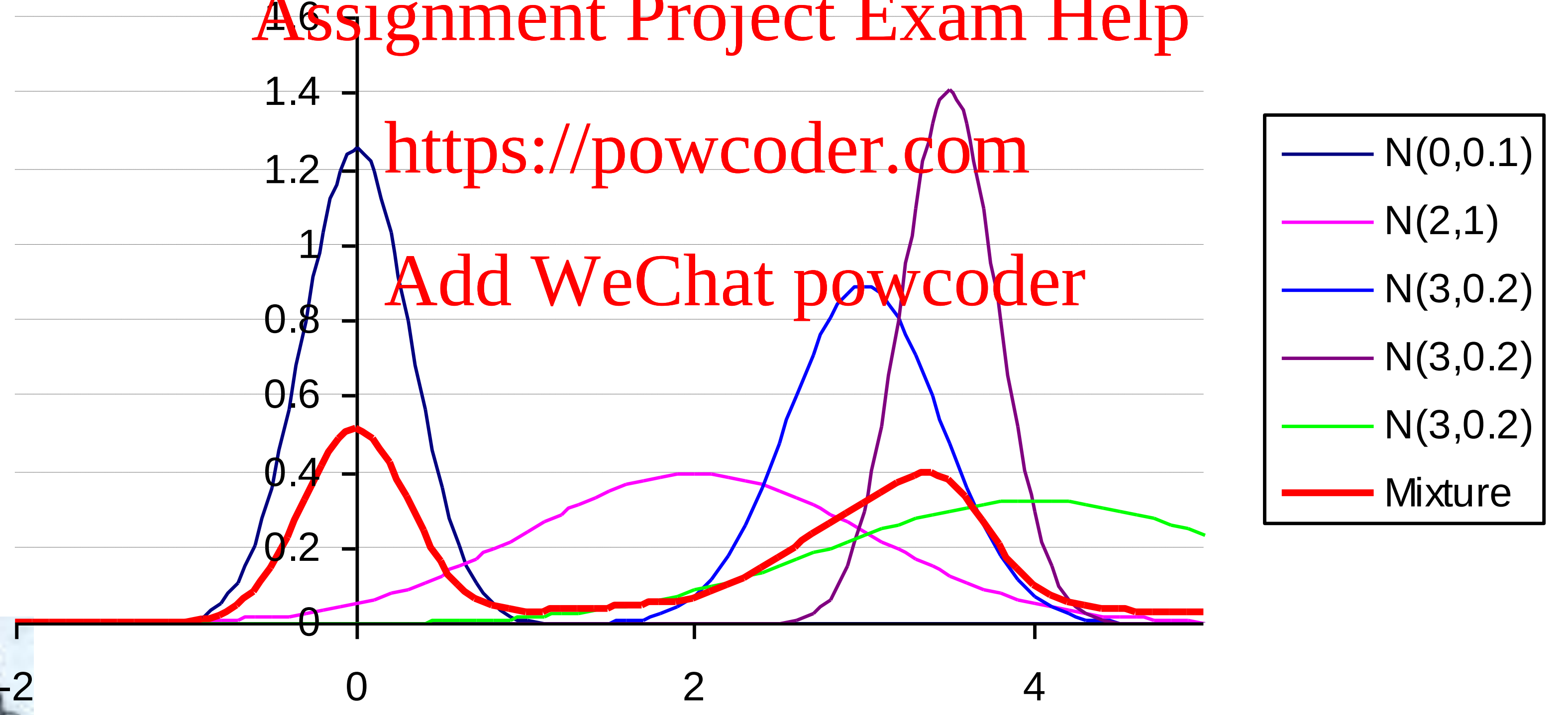
- 2 component mixture model
  - Component 1: $\mu=0$, $\sigma=0.1$
  - Component 2: $\mu=2$, $\sigma=1$
  - $w_1 = 0.2$ $w_2=0.8$



UNIVERSITY OF BIRMINGHAM

Example 4

- 5 component Gaussian mixture PDF



UNIVERSITY OF BIRMINGHAM

Gaussian Mixture Model

- In general, an M component Gaussian mixture PDF is defined by:

$$p(y) = \sum_{m=1}^{M} w_m p_m(y)$$

where each p_m is a Gaussian PDF and

$$0 \le w_m \le 1, \sum_{m=1}^{M} w_m = 1$$

Relationship with Clustering

- Both model data using a set of centroids / means
- In clustering there is no parameter that specifies the ‘spread’ of a cluster. In a GMM component this is done by the covariance matrix.
- In clustering we assign a sample to the closest centroid. In a GMM a sample is assigned to all components with varying probability.

Estimating the parameters of a Gaussian mixture model

- A Gaussian Mixture Model with M components has
  - M means: $\mu_1, \ldots, \mu_M$
  - M variances: $\sigma_1, \ldots, \sigma_M$
  - M mixture weights: $w_1, \ldots, w_M$
- Given $y = y_1, \ldots, y_T$, how do we estimate these parameters?
- I.e. how do we find a maximum likelihood estimate of $\mu_1, \ldots, \mu_M$, $\sigma_1, \ldots, \sigma_M$, $w_1, \ldots, w_M$?

Parameter Estimation

- If we knew which component each sample y_t came from, then parameter estimation would be easy:
  - Set μ_m to be the average value of the samples which belong to the m^{th} component
  - Set σ_m to be the variance of the samples which belong to the m^{th} component
  - Set w_m to be the proportion of samples which belong to the m^{th} component
- But we <u>don't</u> know which component each sample belongs to

The E-M Algorithm

- Step 1:

  Choose number of GMM components, M, and initial GMM parameters:

$$\mu_1^{(0)},\ldots,\mu_M^{(0)},\sigma_1^{(0)},\ldots,\sigma_M^{(0)},w_1^{(0)},\ldots,w_M^{(0)}$$

The E-M Algorithm

- Step 2: For each sample y_t and each GMM component m calculate $P(m|y_t)$ using Bayes theorem and current set of parameters (see next slide)

- Step 3: Define the new estimate of GMM parameters, ${\mu_m}^{(1)}$ and ${\sigma_m}^{(1)}$ as:

$$ {\mu_m}^{(1)} = \frac{1}{P_i} \sum_{t=1}^{T} P(m|y_t) y_t \quad \text{where} \quad P_i = \sum_{t=1}^{T} P(m|y_t) $$

$$ {\sigma_m}^{(1)} = \frac{1}{P_i} \sum_{t=1}^{T} P(m|y_t) \left(y_t - {\mu_m}^{(1)}\right)^2 $$

REPEAT
(Step 2 and 3)

E-M continued

Calculate from m^{th} Gaussian component

m^{th} weight

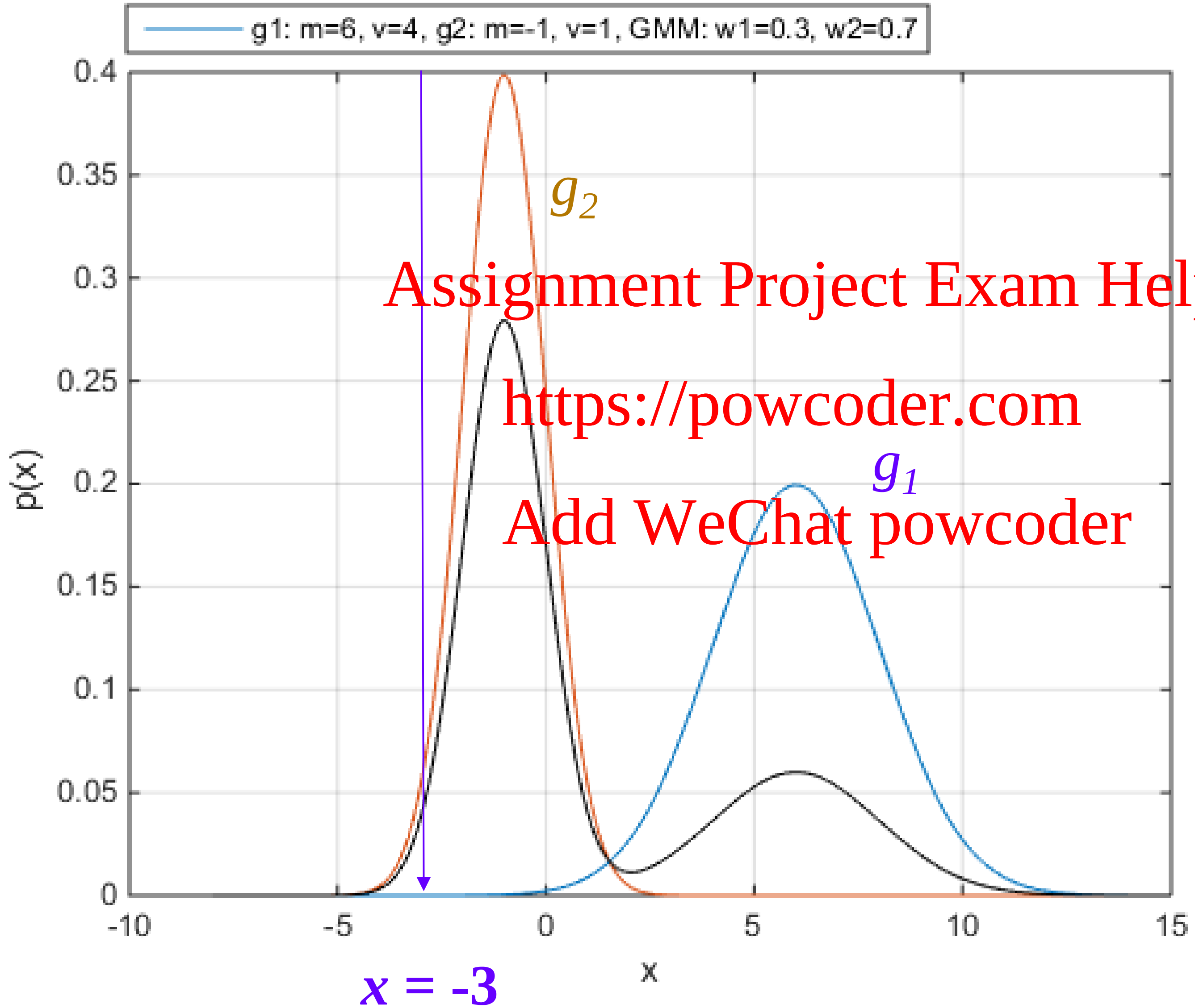
- From Bayes' theorem:

$$\boxed{P(m \mid y_t)} = \frac{p(y_t \mid m)P(m)}{p(y)} = \frac{\boxed{p_m(y_t)}\boxed{w_m}}{\boxed{\sum_{k=1}^{M} p_k(y_t) w_k}}$$

This is a measure of how much y_t 'belongs to' the m^{th} component

Sum over all components

Example



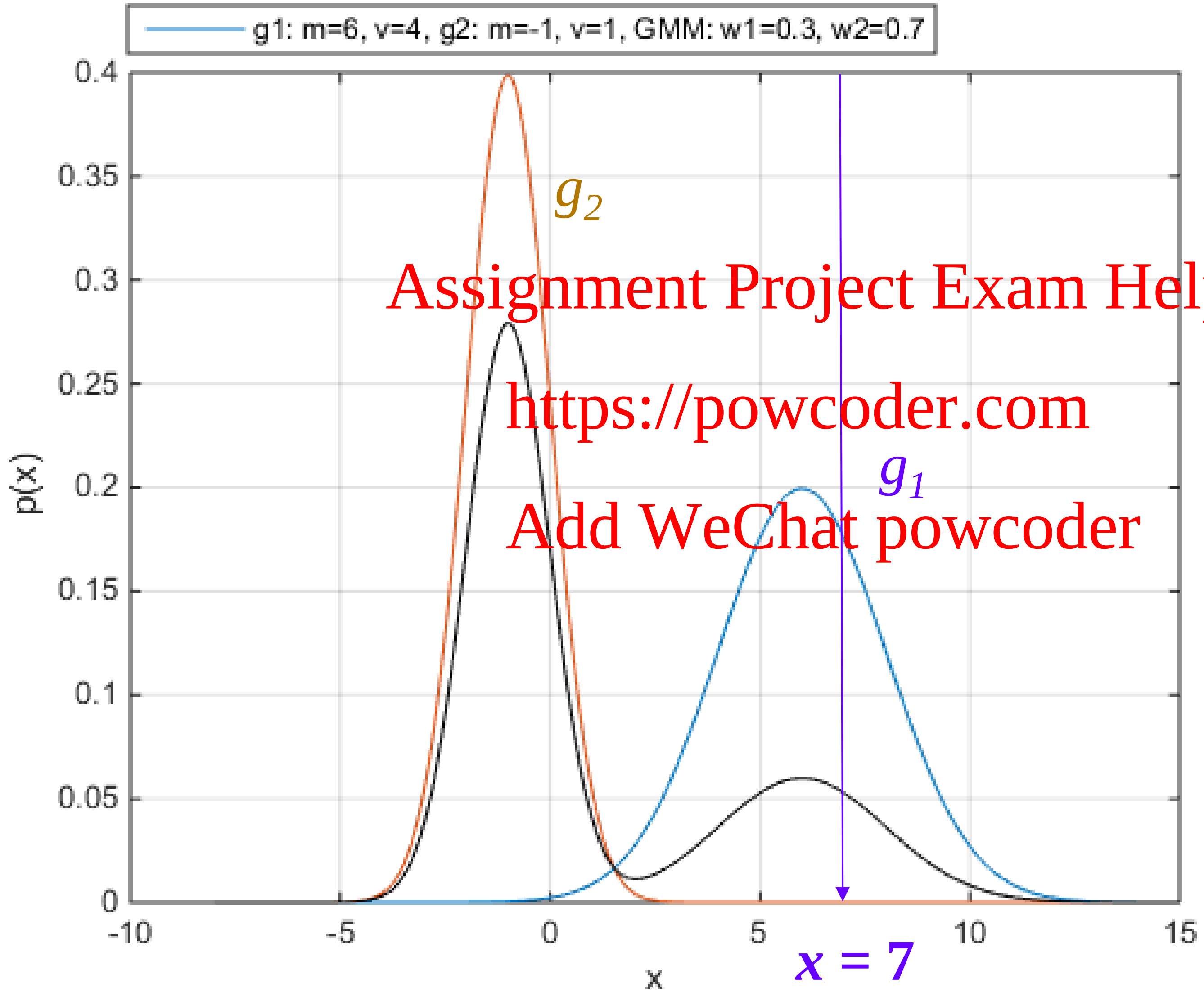
$g_1(x) \approx 0$

$g_2(x) = 0.054$

$$P(1 \mid x) \approx \frac{0 \times 0.3}{0 \times 0.3 + 0.054 \times 0.7} = 0$$

$P(2 \mid x) \approx 1$

Example



$$g_1(x) = 0.176$$

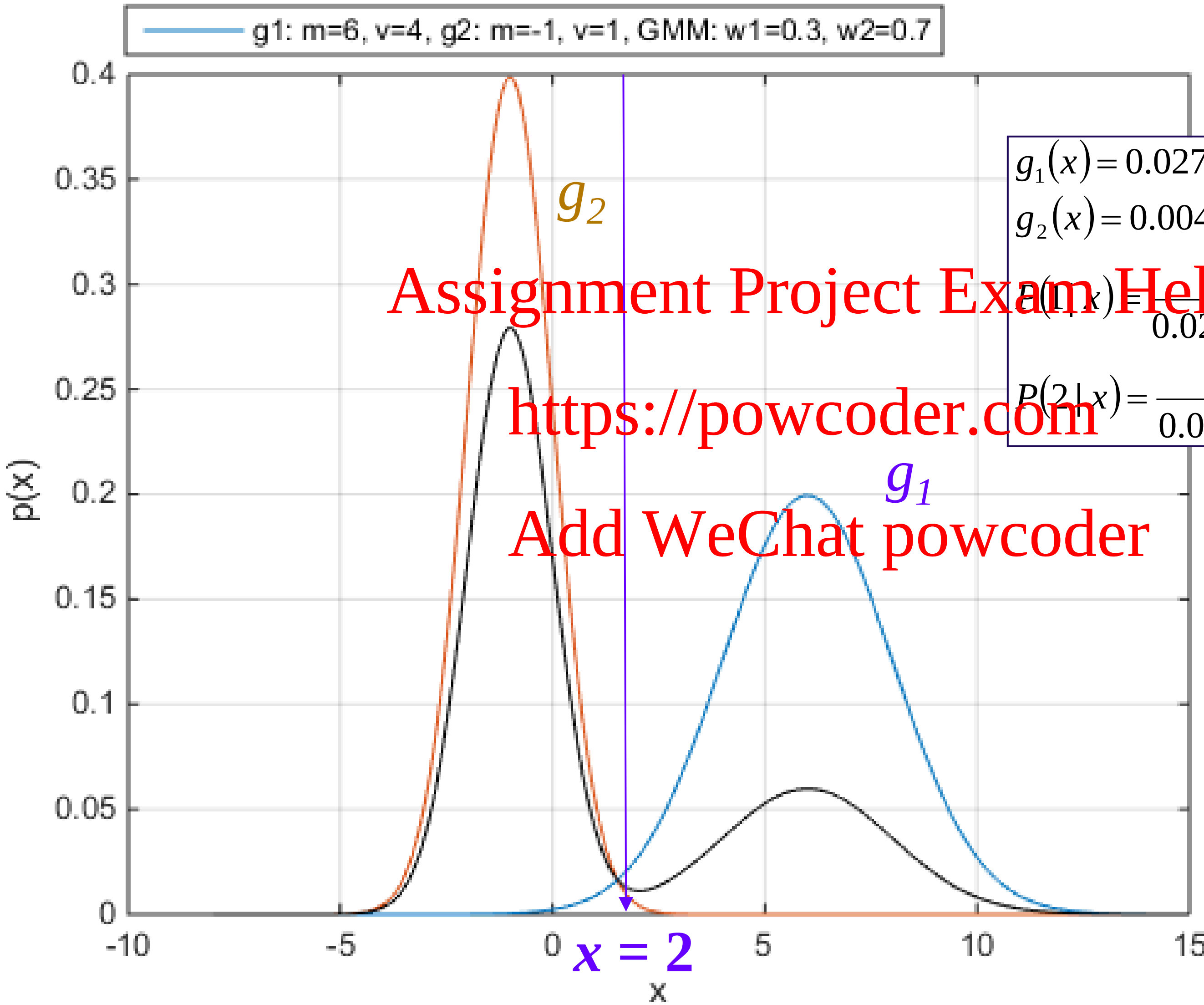
$$g_2(x) \approx 0$$

$$P(1 \mid x) \approx \frac{0.176 \times 0.3}{0.176 \times 0.3 + 0 \times 0.7} = 1$$

$$P(2 \mid x) \approx 0$$

UNIVERSITY OF BIRMINGHAM

Example



$$g_1(x) = 0.027$$
$$g_2(x) = 0.004$$
$$P(1 \mid x) = \frac{0.027 \times 0.3}{0.027 \times 0.3 + 0.004 \times 0.7} = 0.723$$
$$P(2 \mid x) = \frac{0.004 \times 0.7}{0.027 \times 0.3 + 0.004 \times 0.7} = 0.277$$

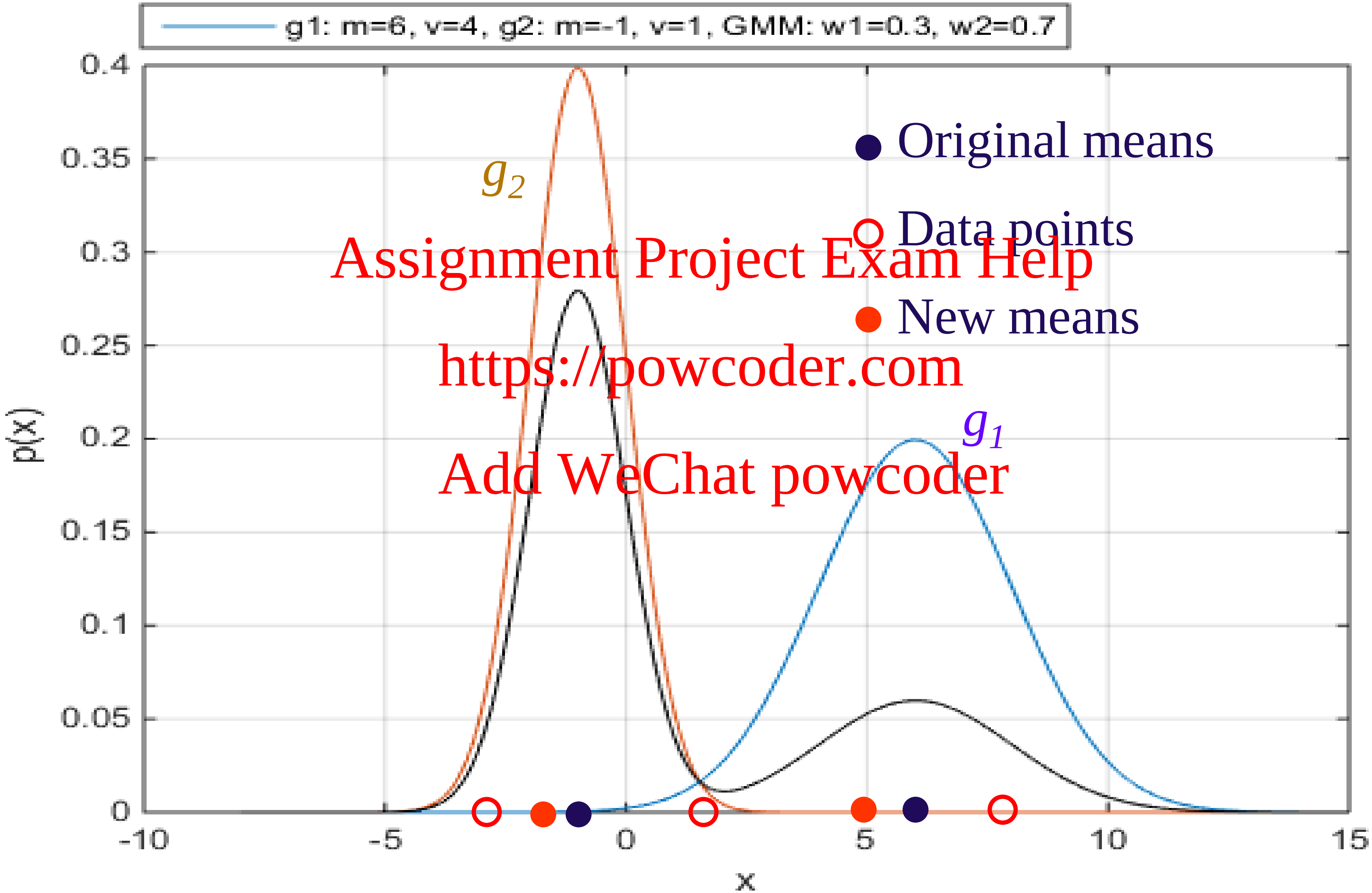
Example (continued)

- So, given these initial estimates of g_1 and g_2, and data points $X = \{x_1, x_2, x_3\} = \{-3, 2, 7\}$, the new values of μ_1 and μ_2 are:

$$\mu_1 = \frac{0\times x_1 + 0.723\times x_2 + 1\times x_3}{0+0.723+1} = \frac{0\times(-3)+0.723\times 2+1\times 7}{1.723} = 4.9$$

$$\mu_2 = \frac{1\times x_1 + 0.277\times x_2 + 0\times x_3}{1+0.277+0} = \frac{1\times(-3)+0.277\times 2+0\times 7}{1.277} = -1.92$$

Example



g1: m=6, v=4, g2: m=-1, v=1, GMM: w1=0.3, w2=0.7
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-10
-5
0
5
10
15
x
p(x)
g_2
g_1
Original means
Data points
New means

E-M and *k*-means clustering

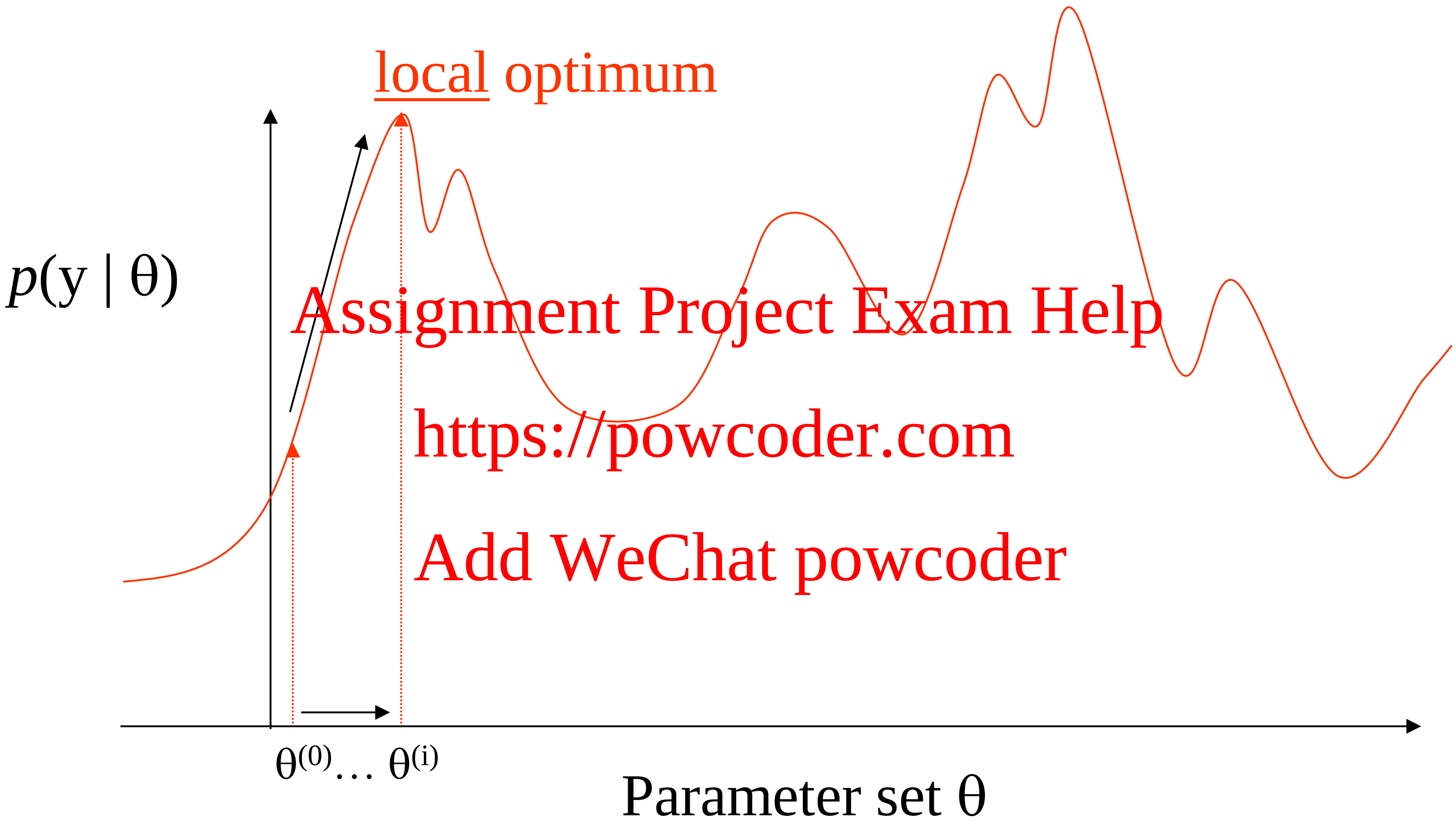
- Consider:
  - Estimating GMM component means in E-M, and
  - Estimating centroids in *k*-means clustering
- Notation
  - GMM component means $\mu_1,\ldots, \mu_M$
  - Cluster centroids $c_1,\ldots, c_M$
- Given a sample y
  - E-M: Calculate $P(m \mid y)$ for each GMM component m
  - K-means: Calculate $d(c_m,y)$ for each centroid c_m
- Reestimation
  - E-M: For each m, allocate $P(m|y_t)y_t$ to reestimation of μ_m
  - K-means: Allocate all of y_t to the closest centroid ($\min\{d(c_m,y_t)\}$)

E-M and *k*-means clustering

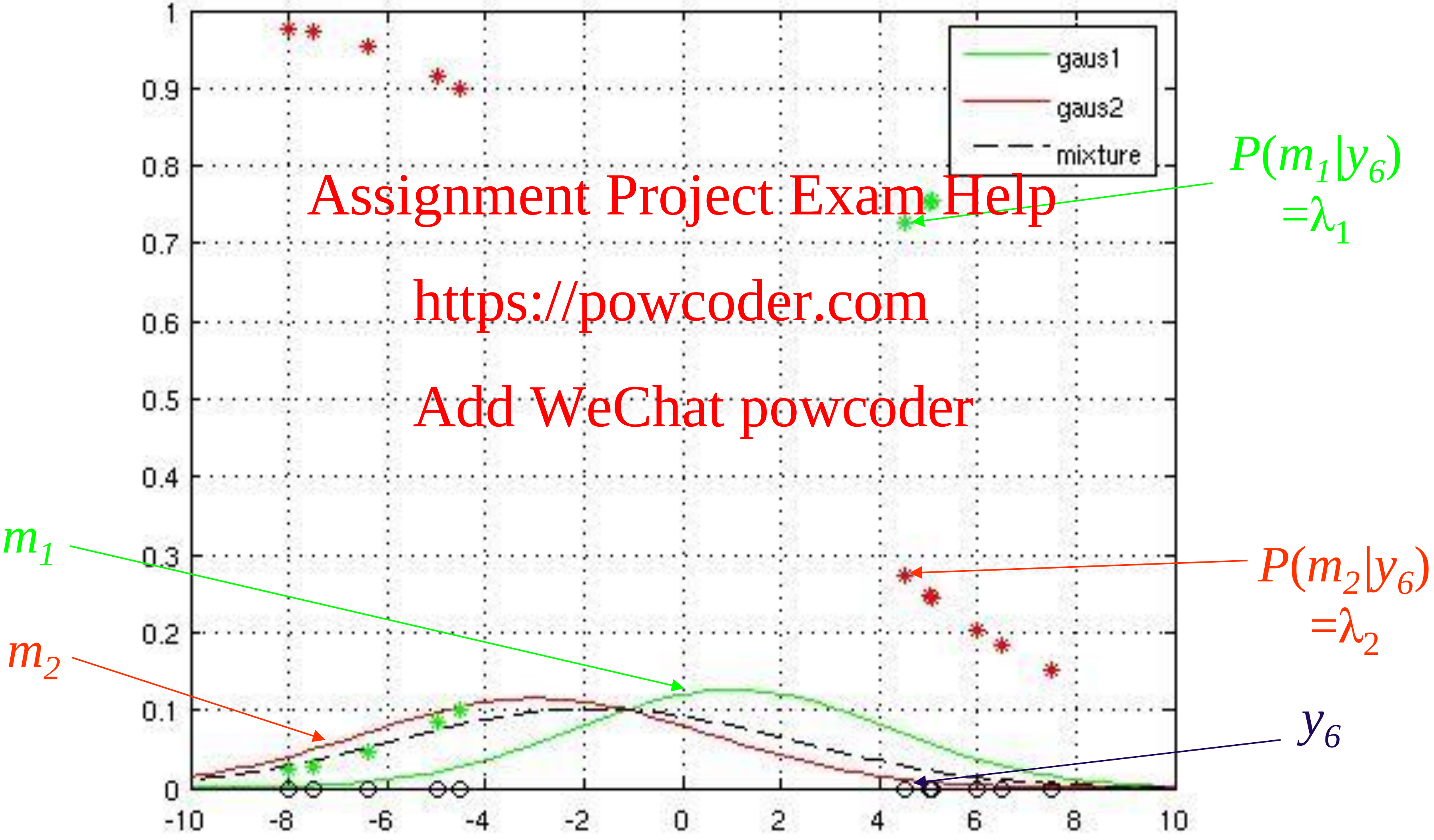
- In some implementations of E-M, y is used <u>only</u> to reestimate the mean μ_m for the most probable GMM component m (i.e. $\max\{P(m|y)\}$)

- If the GMM component variances are all equal, and all of the component weights w_m are equal, then the following are equivalent:
  - $m = \mathrm{argmin}\{d(y, c_m)\}$ (c_m is closest centroid to y)
  - $m = \mathrm{argmax}\{P(m|y)\}$ (i.e. m is the most probable GMM component)

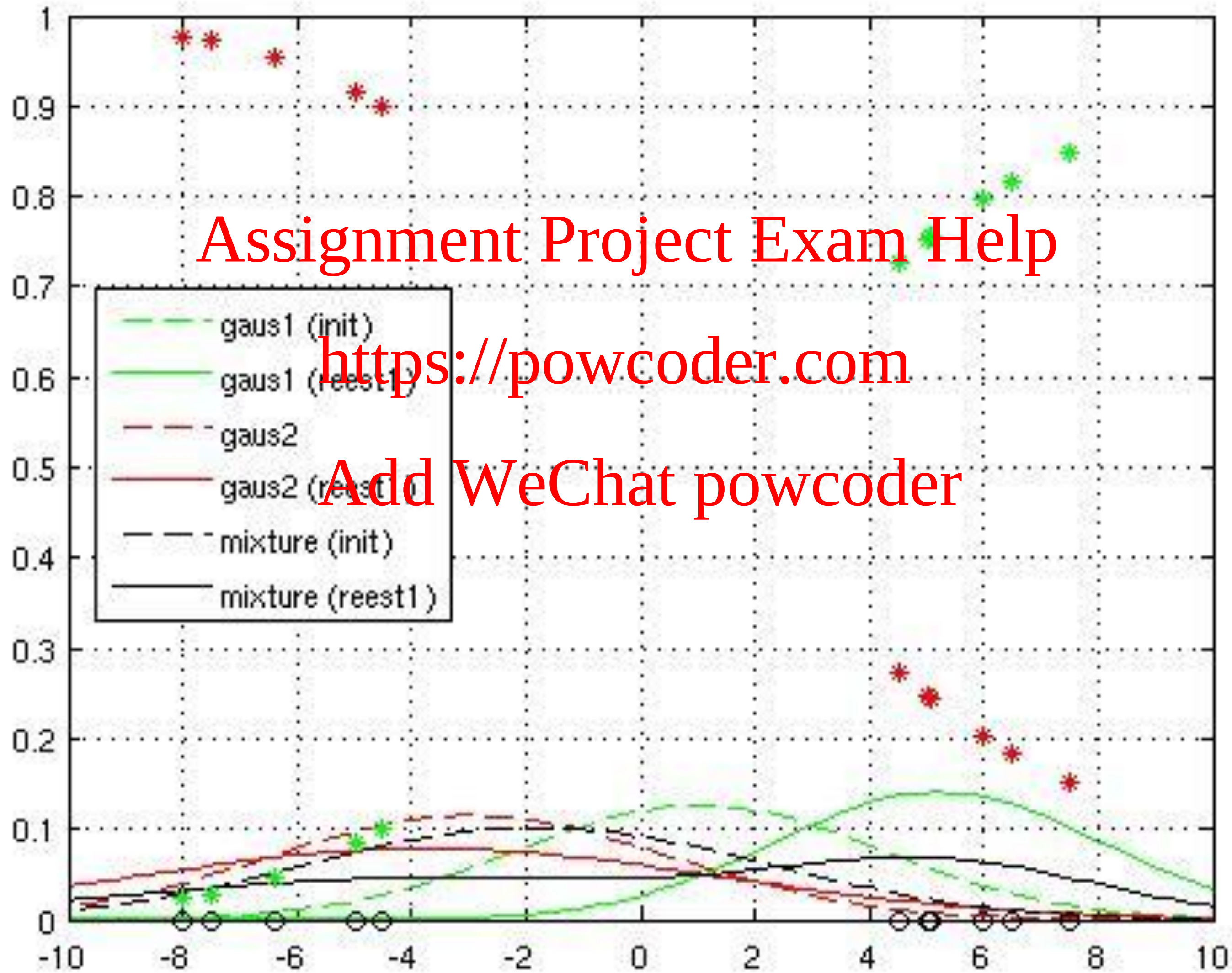
The E-M algorithm



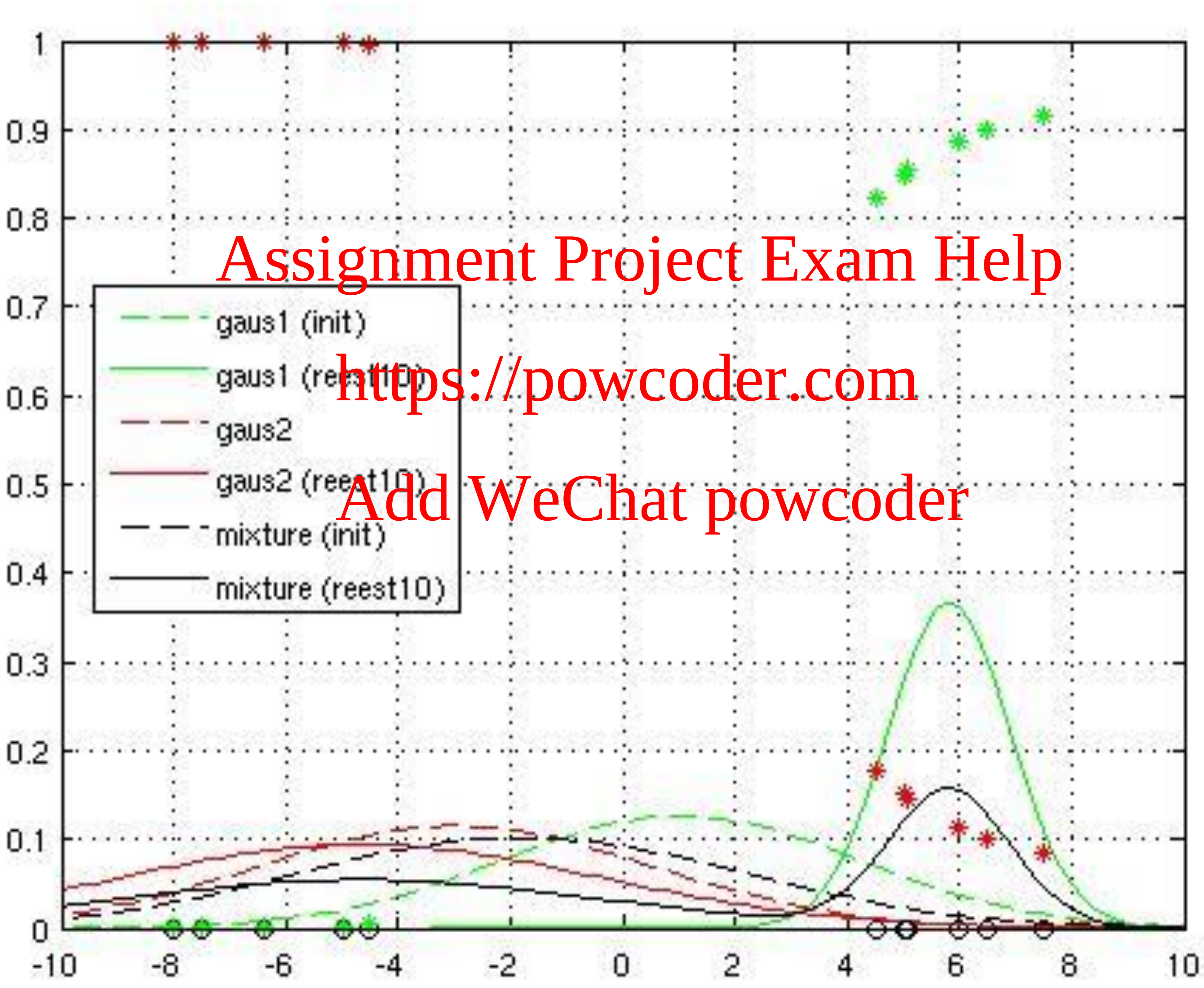
Example – initial model



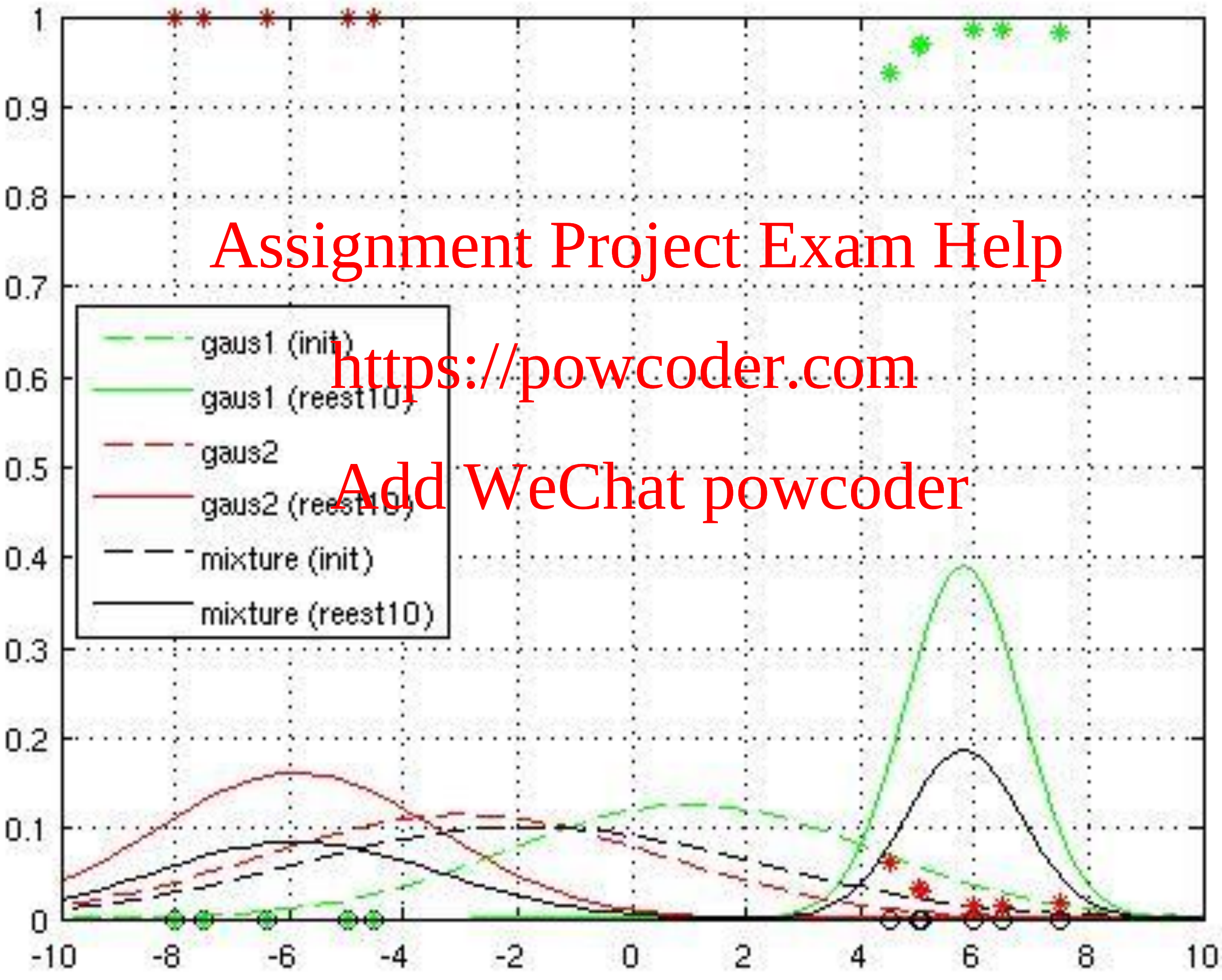
Example – after 1st iteration of E-M



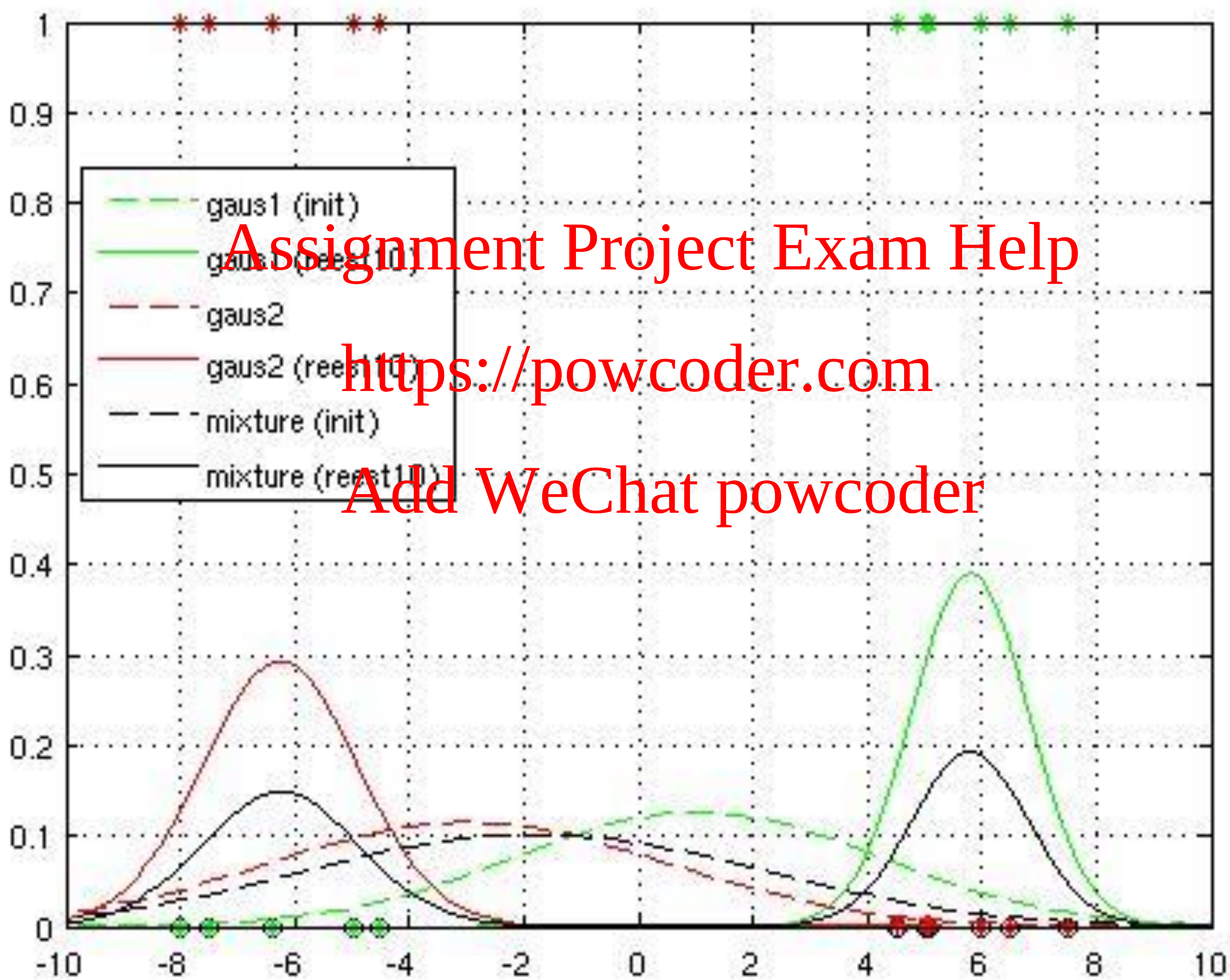
Example – after 2nd iteration of E-M



Example – after 4th iteration of E-M



Example – after 10th iteration of E-M



Summary

- Gaussian mixture PDFs (GMMs)
- Maximum likelihood (ML) parameter estimation – the E-M algorithm
- Comparison of E-M for GMMs with k-means clustering