STATS 235: Modern Data Analysis
Mixture Models

Babak Shahbaba

Department of Statistics, UCI

Introduction

- In this lecture, we discuss a clustering method based on modeling the observed data as a "finite" mixture of Gaussians.
- This is also known as "soft K-means" clustering.
- A simple version of this method assumes that all Gaussians have the same covariance matrix that is diagonal.
- We can extend this method by allowing for different covariance matrices.
- Finally, we can generalize this method by taking the number of clusters into infinity.

Finite mixture models

- We present a mixture of K base distributions, $P_1, \ldots, P_K$, as follows:

$$P(x_i|\pi, \theta) = \sum_{k=1}^{K} \pi_k P_k(x_i|\theta_k)$$

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

- We mainly focus on mixture of Gaussians,

$$P(x_i|\pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

Finite mixture models

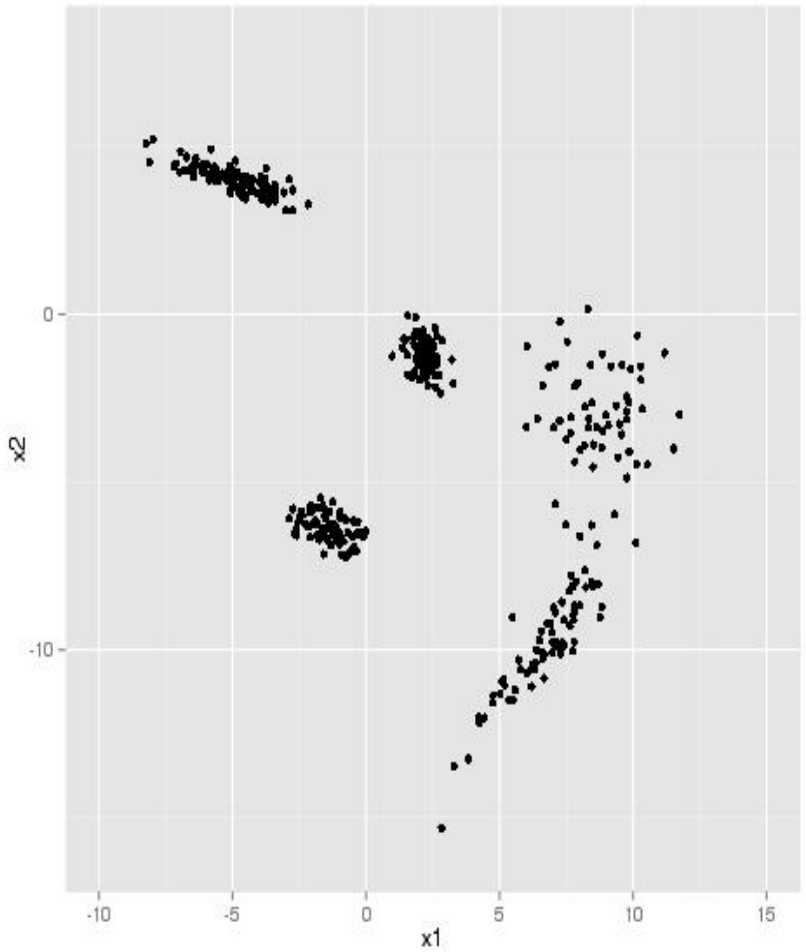
- We could of course maximize the log-likelihood function,

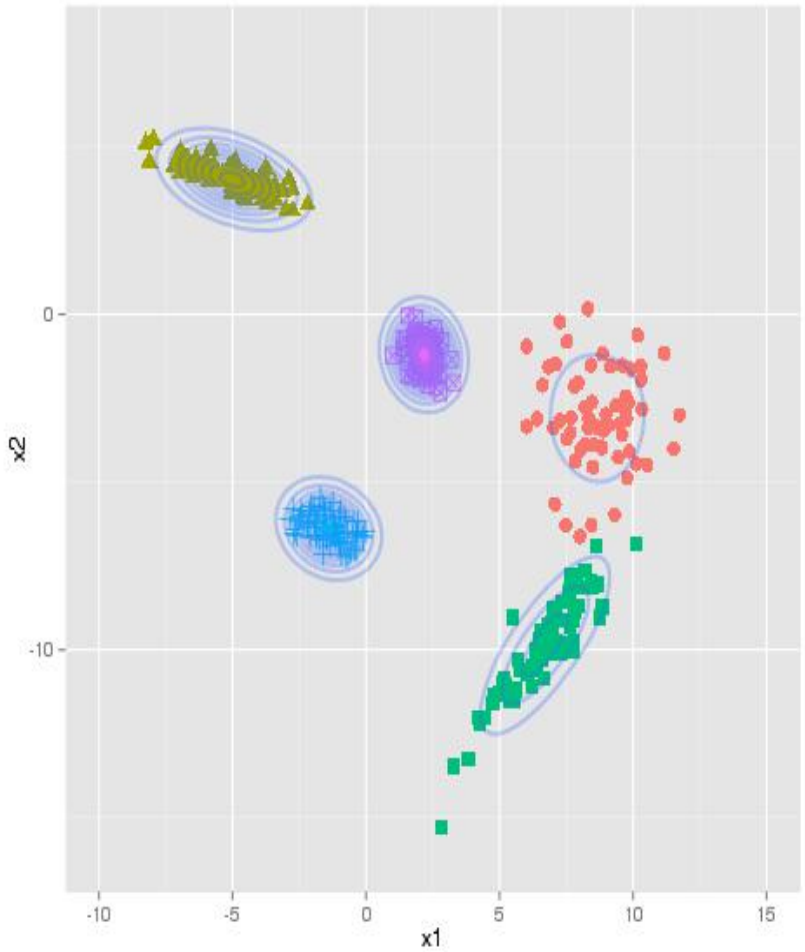
$$\sum_i \log(\sum_k \pi_k N(x_i|\mu_k, \Sigma_k))$$

  subject to the constraints that $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

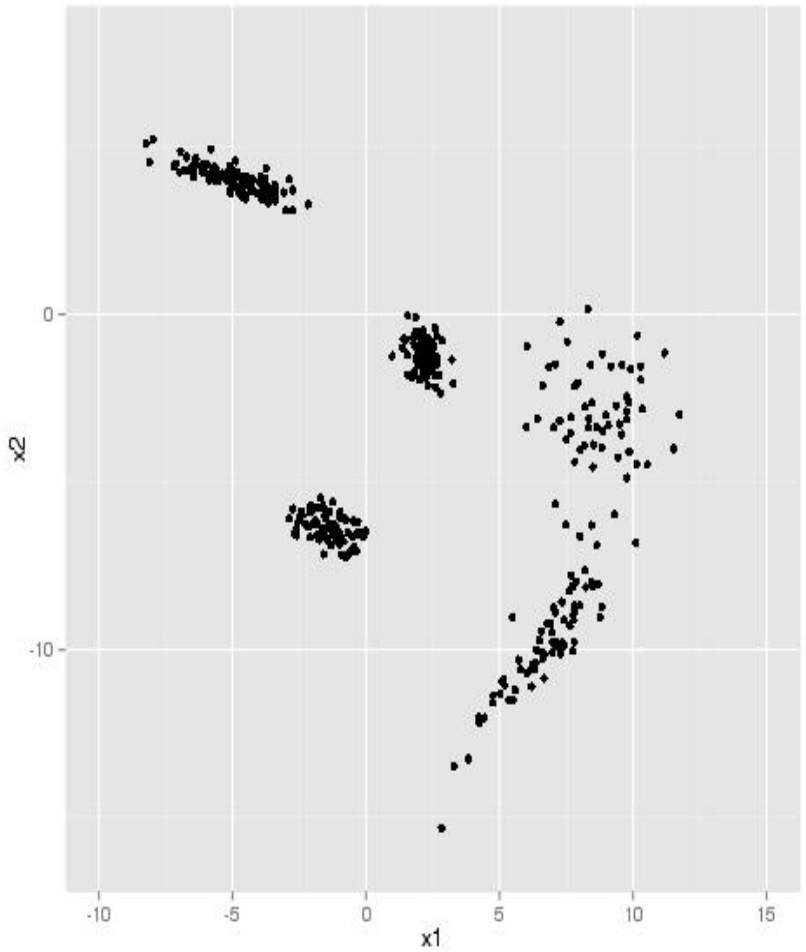
- However, in such cases, it is easier to use the data augmentation approach by introducing latent indicators and estimating the parameters using the *expectation-maximization* (EM) algorithm

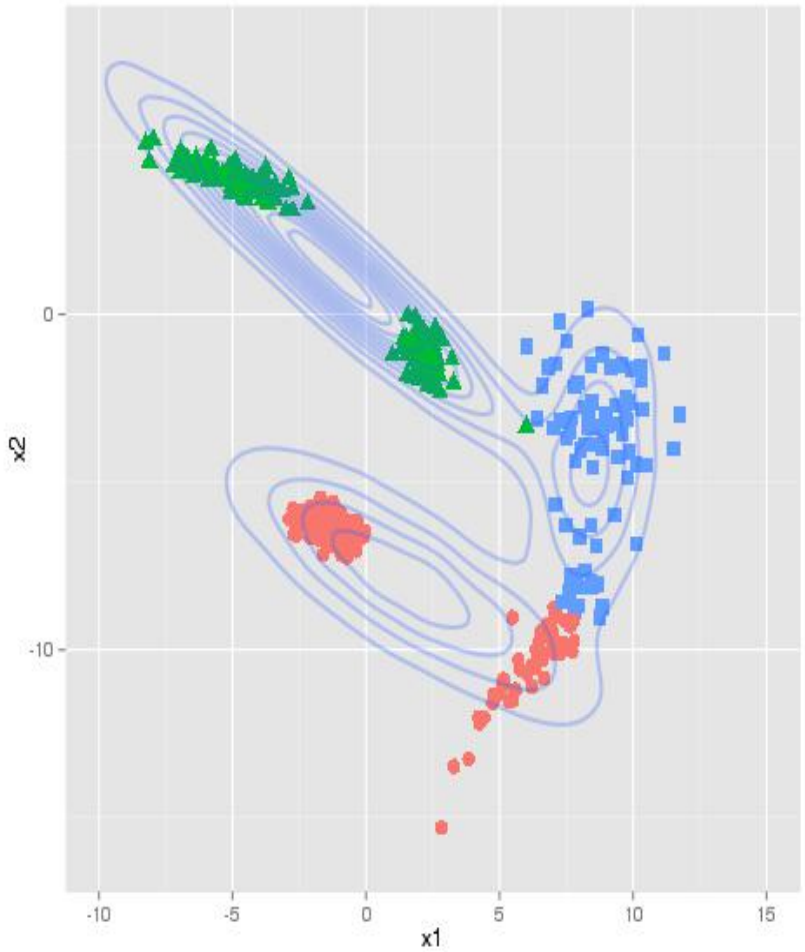
Example: Mixture of 5 Gaussians





Example: Mixture of 3 Gaussians





Dirichlet process mixture models

- We now discuss Dirichlet process mixture models, which are usually used for nonparametric density estimation and clustering.
- Ferguson (1973) introduced the Dirichlet process as a class of prior distributions for which the support is large, and the posterior distribution is manageable analytically.
- Antoniak (1974) proposed the idea of using a Dirichlet process as the prior for the mixing proportions of a simple distribution (e.g., Gaussian).
- First, let's review simple mixture distributions from the Bayesian point of view.

Mixture distributions

- Consider a random sample, $x_1, \ldots, x_n$, drawn independently from some unknown distribution.
- We can use a mixture of simple distributions to model the density of X:

$$P(x_i|\pi, \phi) = \sum_{k=1}^{K} \pi_k P_k(x_i|\phi_k)$$

$$0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

Here, π_k are the mixing proportion, and P_k is a simple distribution such as normal with $\phi = (\mu, \sigma)$.

Mixture distributions

- A common prior for π_k is a symmetric Dirichlet distribution

$$P(\pi_1, ..., \pi_K) = \frac{\Gamma(\gamma)}{\Gamma(\gamma/K)^K} \prod_{k=1}^{K} \pi_k^{(\gamma/K)-1}$$

- We assume ϕ_c are assumed to be independent under the prior with distribution G_0.

Finite mixtures

- For a finite K, we can represent the mixture model as follows:

$$\begin{aligned} \phi_k &\sim G_0 \\ \pi_1, \ldots, \pi_K &\sim Dirichlet(\gamma/K, \ldots, \gamma/K) \\ z_i|\pi_1, \ldots, \pi_K &\sim Discrete(\pi_1, \ldots, \pi_K) \\ x_i|z_i, \phi &\sim P_k(x_i|\phi_{k_i}) \end{aligned}$$

- By integrating over the Dirichlet prior, we obtain the following conditional distribution for z_i:

$$P(z_i = k|z_1, \ldots, z_{i-1}) = \frac{n_{ik} + \gamma/K}{i - 1 + \gamma}$$

where, n_{ik} represents the number of data points previously (i.e., before the i^{th}) assigned to component k.

Dirichlet process - Prior

- We might not know the number of components, K, *a priori*. We can assume $K \to \infty$. The result is a Dirichlet process.
- Consider a restaurant with infinite possibilities, i.e., infinite menu items and infinite number of tables.
- The rule is that people at the same table must order the same food.
- First customer comes and sits anywhere she wants and order whatever she wants. The second customer can sit with the first one (therefore, order the same food), or he can sit by himself and order something different, and so on.
- Note that tables with more customers are more attractive.

Computation

- We can derive Drichlet process mixture as the limit of finite mixture (i.e., $K \to \infty$).

$$\begin{aligned} P(z_i = k | z_1, \ldots, z_{i-1}) &\to \frac{n_k}{i - 1 + \gamma} \\ P(z_i \neq z_j, \forall j < i | z_1, \ldots, z_{i-1}) &\to \frac{\gamma}{i - 1 + \gamma} \end{aligned}$$

- Therefore, the conditional prior distribution of θ_i is

$$\theta_i | \theta_1, ..., \theta_{i-1} \sim \frac{1}{i - 1 + \gamma} \sum_{j<i} \delta(\theta_j) + \frac{\gamma}{i - 1 + \gamma} G_0$$

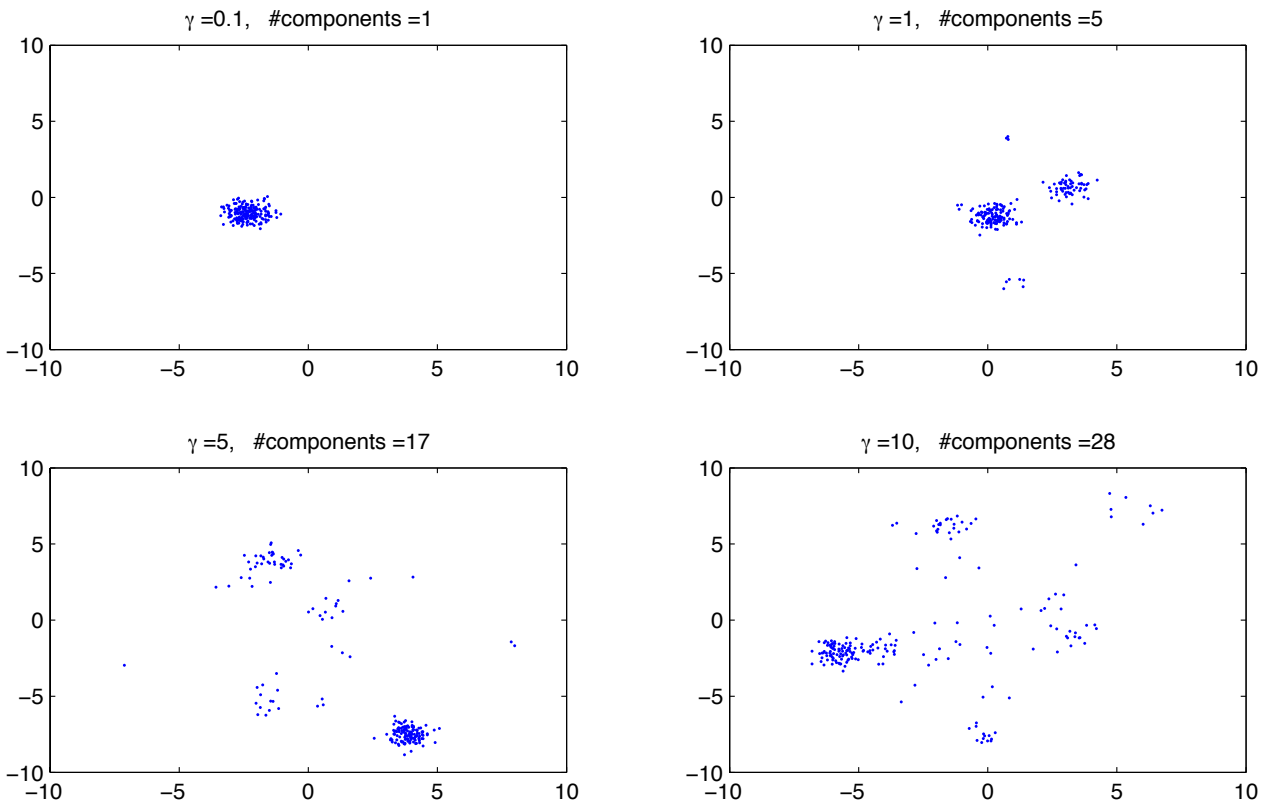
Computation

- The result is a Dirichlet process, which is a distribution over distributions.
- The Dirichlet process mixture of simple distributions has the following form:

$$\begin{aligned} x_i|\theta_i &\sim F(\theta_i) \\ \theta_i|G &\sim G \\ G &\sim \mathcal{D}(G_0, \gamma) \end{aligned}$$

Samples from Dirichlet process mixture prior

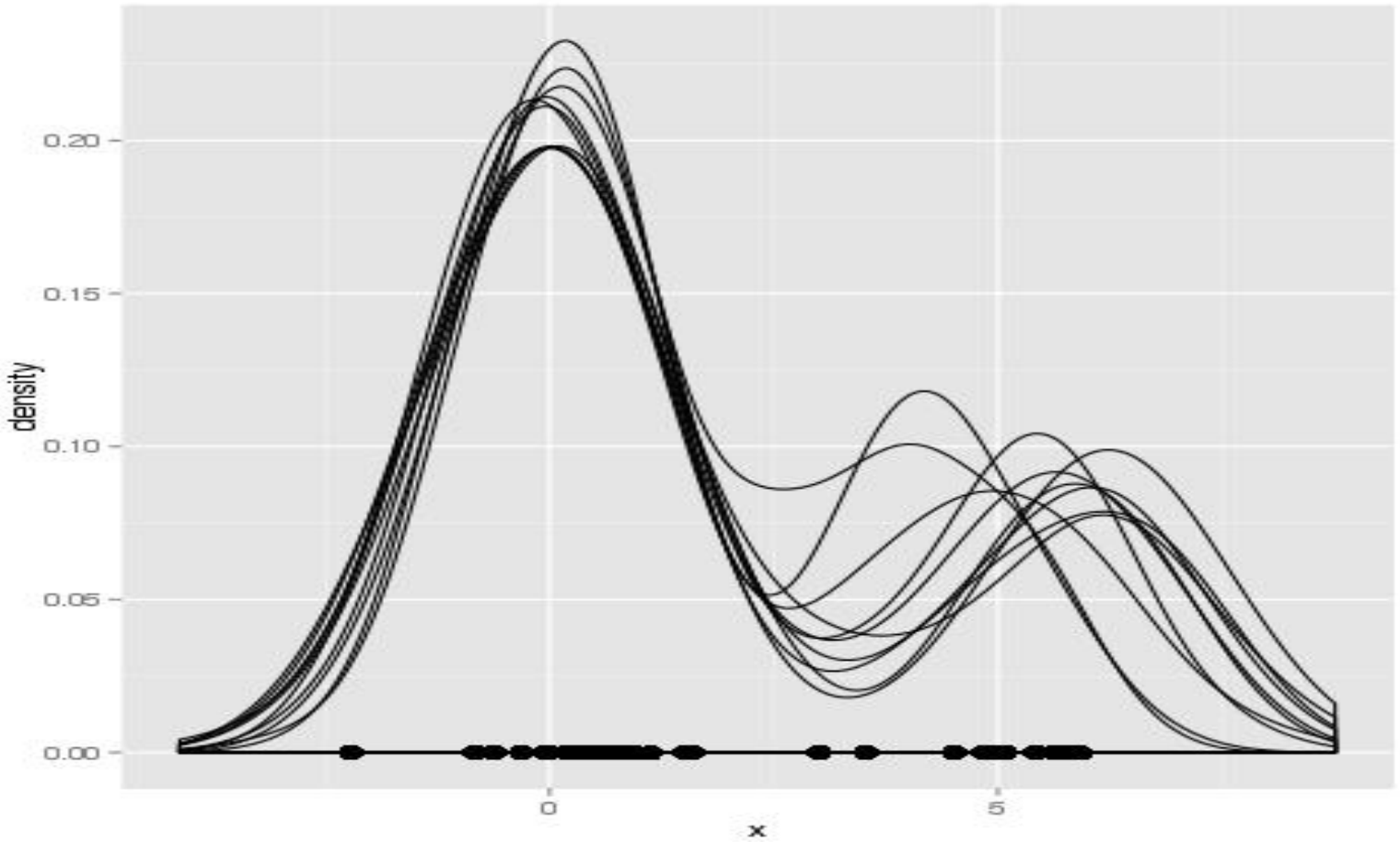
- The following graphs shows 4 different datasets ($n = 200$) sampled from Dirichlet process mixture priors with different γ.



Dirichlet process- Posterior

- Now, assume that our objective is to cluster customers according to their taste of food.
- We can use an MCMC algorithm for this purpose.
- After all customers are seated, we let them move around and change their place if they want.
- As before, a more crowded table tends to attract other customers.
- When this procedure converges, customers form few clusters.
- Neal (2000) discussed several MCMC algorithms for Ditichlet process mixtures.

DPM of univariate Gaussians



10 Posterior samples using the Gibbs algorithm of Escobar and West (1995).