Study EM and MCMC algorithms for Gaussian mixture model

Dongkuan Xu
College of IST, Penn State University

Gaussian mixture model (GMM)

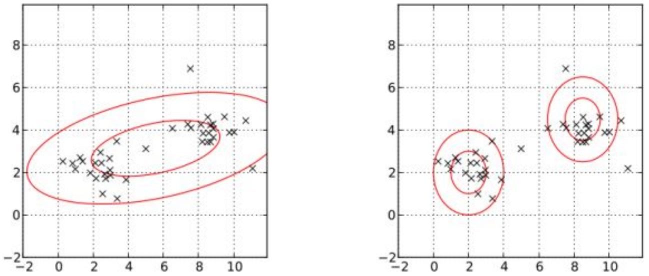
- Definition

$$p(\vec{x}) = \sum_{i=1}^{K} \phi_i \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i)$$

$$\mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^K |\Sigma_i|}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^{\mathrm{T}} \Sigma_i{}^{-1}(\vec{x}-\vec{\mu}_i)\right)$$

$$\sum_{i=1}^{K} \phi_i = 1$$

- Related work
  - GMM based on EM algorithm [1-3]
  - Bayesian GMM based on EM or MCMC [4-5]
  - GMM with unknown mixture number [6-7]
- Application
  - Data clustering
  - Image segmentation
  - Time series analysis
  - Genetics



Motivation: data could be multimodal

Problem 1: log-likelihood is hard to calculate when MLE

- Maximum likelihood estimation
    - Likelihood function

$$\mathcal{L}(\theta\,;x)$$

    - Maximum likelihood estimate

$$\hat{\theta} \in \{\underset{\theta\in\Theta}{\arg\max}\ \mathcal{L}(\theta\,;x)\}$$

- Log-likelihood is hard to calculate by direct derivative: summation operations in the logarithmic operations

$$\sum_{i=1}^{N} \log\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)\right\}$$

Solution: Expectation-maximization (EM) for GMM

- Log-likelihood function

$$\sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k) \right\}$$

- Estimation step

$$\gamma(i,k) = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i | \mu_j, \Sigma_j)}$$

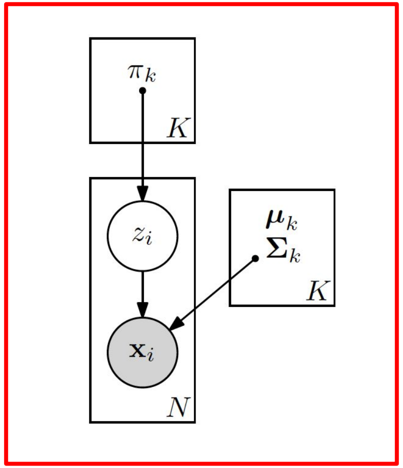
- Maximization step

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T$$

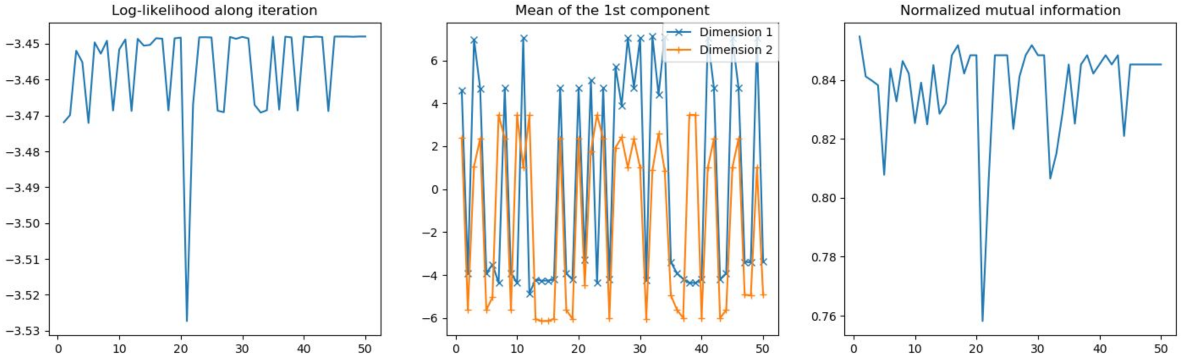
$$\pi_k = N_k / N$$

$$N_k = \sum_{i=1}^{N} \gamma(i,k)$$

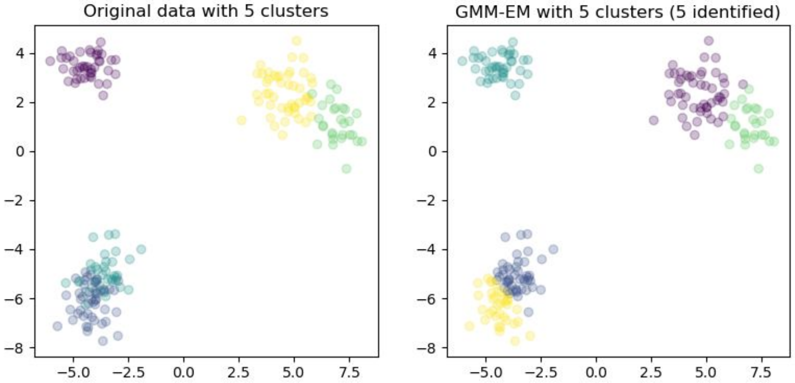


GMM with EM

- Stopping criteria
  - (1) Maximum iteration number of EM (2) A threshold about the gain on log-likelihood
- Running time per iteration: 0.00326 s
- Iteration: (1) Log-likelihood (2) Mean of the 1st component (3) Normalized mutual information (NMI)



- Clustering results:



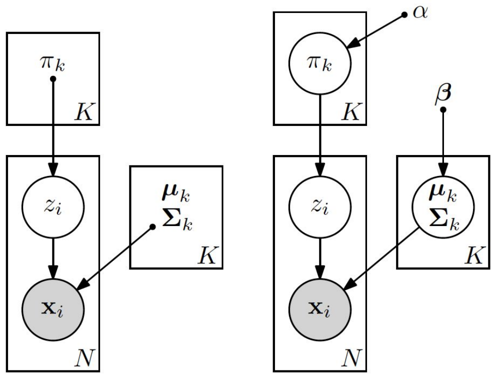
Problem 2: Too many free parameters for EM when high-dimensional

- ❑ Parameters of GMM

$$\mu_k = \frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)x_i \qquad \Sigma_k = \frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)(x_i-\mu_k)(x_i-\mu_k)^T \qquad \pi_k = N_k/N$$

- ❑ MLE approach based on EM may fail due to singularities or degeneracies
- ❑ A Bayesian approach alleviate these by treating $\Theta = (\boldsymbol{\pi}, \{\boldsymbol{\mu}_k\}, \{\boldsymbol{\Sigma}_k\})$ as random variables and working with distributions over Θ rather than point estimates
- ❑ Choose conjugate priors to marginalize over the parameters
  - ❑ Symmetric Dirichlet prior for mixture weights
  - ❑ Normal-inverse-Wishart (NIW) prior for component parameters

$$\begin{aligned}
\boldsymbol{\pi} &\sim \mathrm{Dir}\left(\alpha/K\mathbf{1}\right)\\
z_i &\sim \boldsymbol{\pi}\\
\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k &\sim \mathrm{NIW}(\mathbf{m}_0, \kappa_0, \nu_0, \mathbf{S}_0)\\
\mathbf{x}_i &\sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})
\end{aligned}$$



π_k
K
z_i
$\mathbf{x}_i$
N
$\boldsymbol{\mu}_k$ $\boldsymbol{\Sigma}_k$
K
α
π_k
K
β
z_i
$\mathbf{x}_i$
N
$\boldsymbol{\mu}_k$ $\boldsymbol{\Sigma}_k$
K

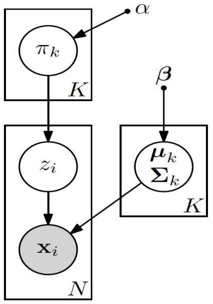
GMM with EM vs Bayesian GMM

Solution: Bayesian approach for GMM

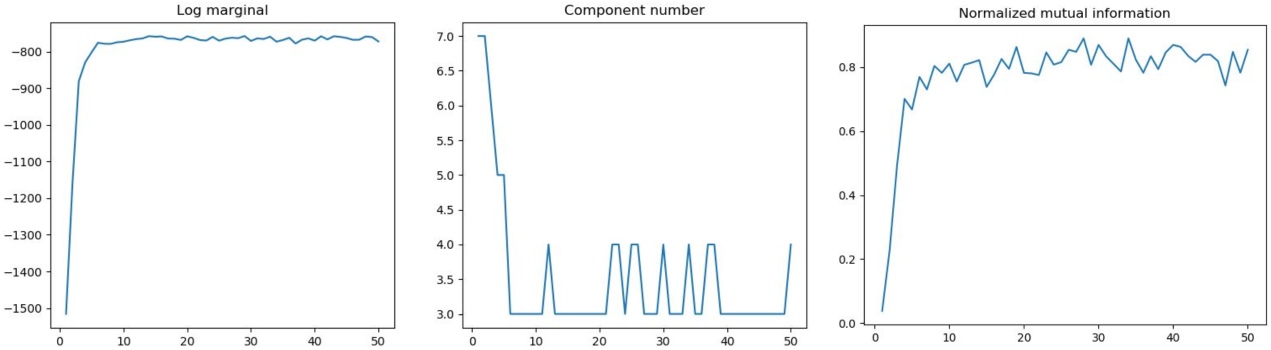
- Algorithm of Collapsed Gibbs sampler for GMM:

```
Choose an initial $\mathbf{z}$.
for $T$ iterations do                                                   ▷ Gibbs sampling iterations
    for $i = 1$ to $N$ do
        Remove $\mathbf{x}_i$'s statistics from component $z_i$.          ▷ Old assignment for $\mathbf{x}_i$
        for $k = 1$ to $K$ do                                          ▷ Every possible component
            Calculate $P(z_i = k|\mathbf{z}_{\backslash i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(z_i = k|\mathbf{z}_{\backslash i}, \boldsymbol{\alpha})\, p(\mathbf{x}_i|\mathcal{X}_{k\backslash i}, \boldsymbol{\beta})$.
        end for
        Sample $k_{\text{new}}$ from $P(z_i|\mathbf{z}_{\backslash i}, \mathcal{X}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ after normalizing.
        Add $\mathbf{x}_i$'s statistics to the component $z_i = k_{\text{new}}$.   ▷ New assignment for $\mathbf{x}_i$
    end for
end for
```

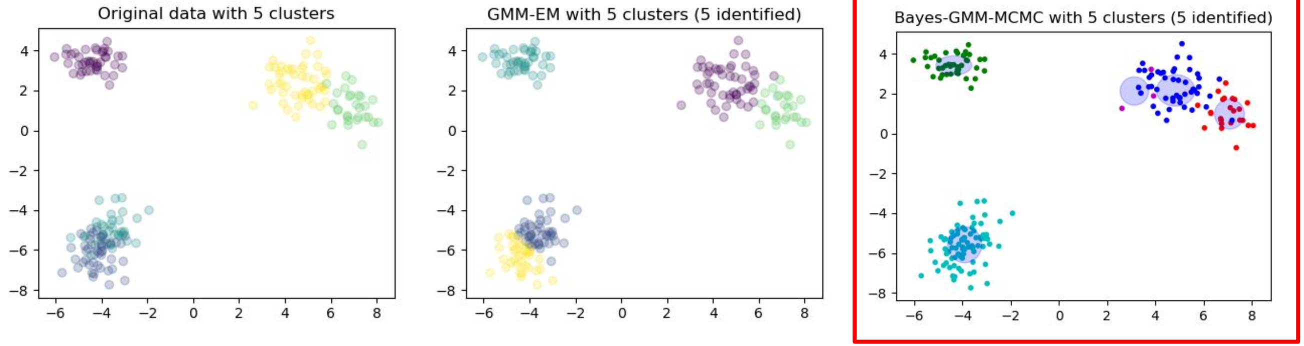
Bayesian GMM



- Stopping criteria:
    - (1) Maximum iteration number of EM
    - (2) A threshold about the gain of log marginal of data and component assignments: p(X, z) (**possible**)
- Running time per iteration: 0.00492 (**higher than the one of GMM-EM**)
- Iteration: (1) Log marginal of p(X, z) (2) Normalized mutual information (NMI) (3) Component number

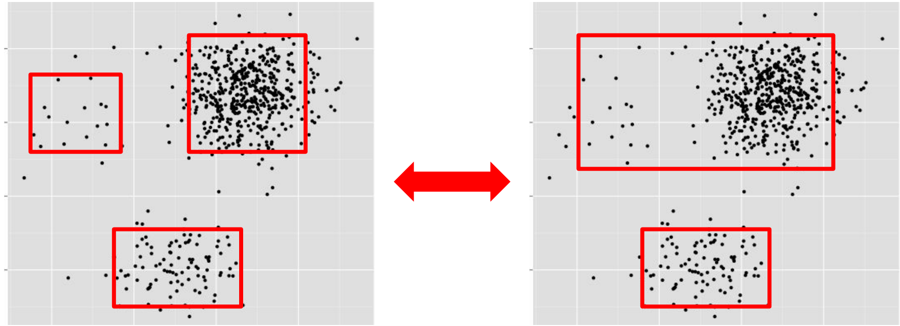


- Clustering result: **a little worse than GMM-EM**



Problem 3: GMM with unknown mixture number

- Unknown mixture number



- MLE, EM do not work

$$\sum_{i=1}^{N}\log\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(x_i|\mu_k,\Sigma_k)\right\}$$

log-likelihood

$$\gamma(i,k)=\frac{\pi_k\mathcal{N}(x_i|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(x_i|\mu_j,\Sigma_j)}$$

log-likelihood

- A number of possible solutions
  - reversible jump MCMC (Richardson and Green 1997, Gruet et al. 1999)
  - Bayes factors (Kass and Raftery 1995, Richardson and Green 1997)
  - entropy distance or K-L divergence (Mengersen and Robert 1996, Sahu and Cheng 2003)
  - birth-and-death processes (Stephens 2000a, Cappé et al. 2002)

Solution: Reversible jump MCMC for GMM with unknown mixture number

- ❏ Idea: birth and death moves
  - ❏ Add a new normal component in the mixture generated from the prior, or remove one component, according to the acceptance probability
- ❏ Assumption:
  - ❏ Competing models can be enumerable and represented as: $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$
  - ❏ Current state of Markov chain is: (k, θ_k)
- ❏ GMM with RJMCMC:
  - ❏ Propose a visit from current model $(k, \theta_k)$to next model $(\theta_{k'}, k')$
  - ❏ Accept the visit or not
  - ❏ Repeat proposing model visit until stopping criteria are met
  - ❏ Clustering based on standard GMM

Reversible jump MCMC for GMM with unknown mixture number

- Algorithm of GMM with RJMCMC:
  - (1) Initialize current model indicator k
  - (2) Propose a visit from mode M_k to model M_k ’ with probability J(k → k’)
  - (3) Sample parameter u of GMM from a proposal density $q(u|\theta_k, k, k')$
  - (4) Set $(\theta_{k'}, u') = g_{k,k'}(\theta_k, u)$, where $g_{k,k'}(\cdot)$is a bijection, where u and u ‘ play the role of matching the dimensions of both vectors
  - (5) The acceptance probability of the new model $(\theta_{k'}, k')$ the minimum between 1 and :

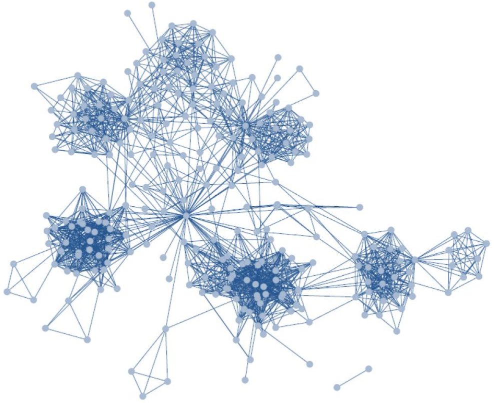
$$\underbrace{\frac{p(y|\theta_{k'}, k')p(\theta_{k'})p(k')}{p(y|\theta_k, k)p(\theta_k)p(k)}}_{\text{model ratio}} \underbrace{\frac{J(k' \to k)q(u'|\theta_{k'}, k', k)}{J(k \to k')q(u|\theta_k, k, k')} \left| \frac{\partial g_{k,k'}(\theta_k, u)}{\partial(\theta_k, u)} \right|}_{\text{proposal ratio}}$$

  - (6) Calculate MCMC standard error of parameter samples
  - (7) ite = ite + 1
  - (8) if ite < max or MCMC standard error > φ, then go to (2)
  - (9) output cluster assignment for all data samples

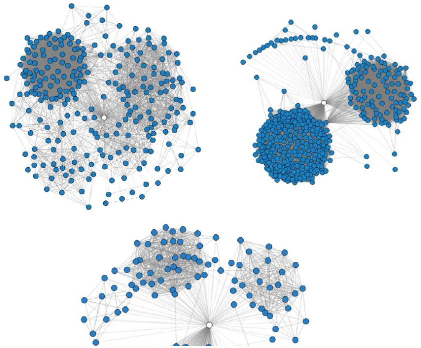
Network-structure data clustering

- Network data is ubiquitous
  - Web network
  - Social network
  - Biological network, etc.

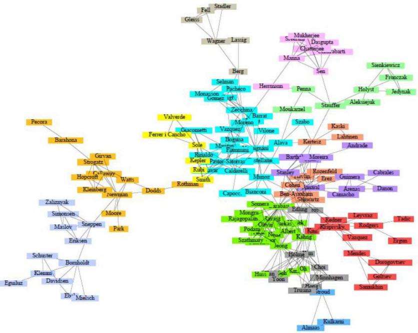
- Network clustering
  - Detect sub-networks that satisfy certain properties
  - Many connections within clusters and few connections across clusters



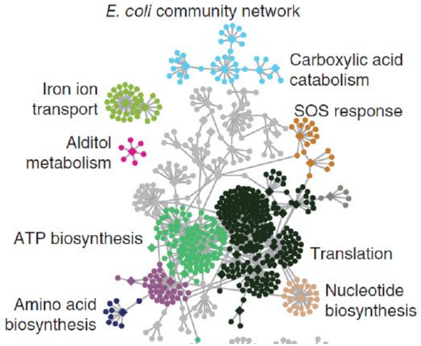
Web network



Social network



Co-author network



Gene network

Comparison between EM-GMM and Bayesian-GMM applied to network data clustering

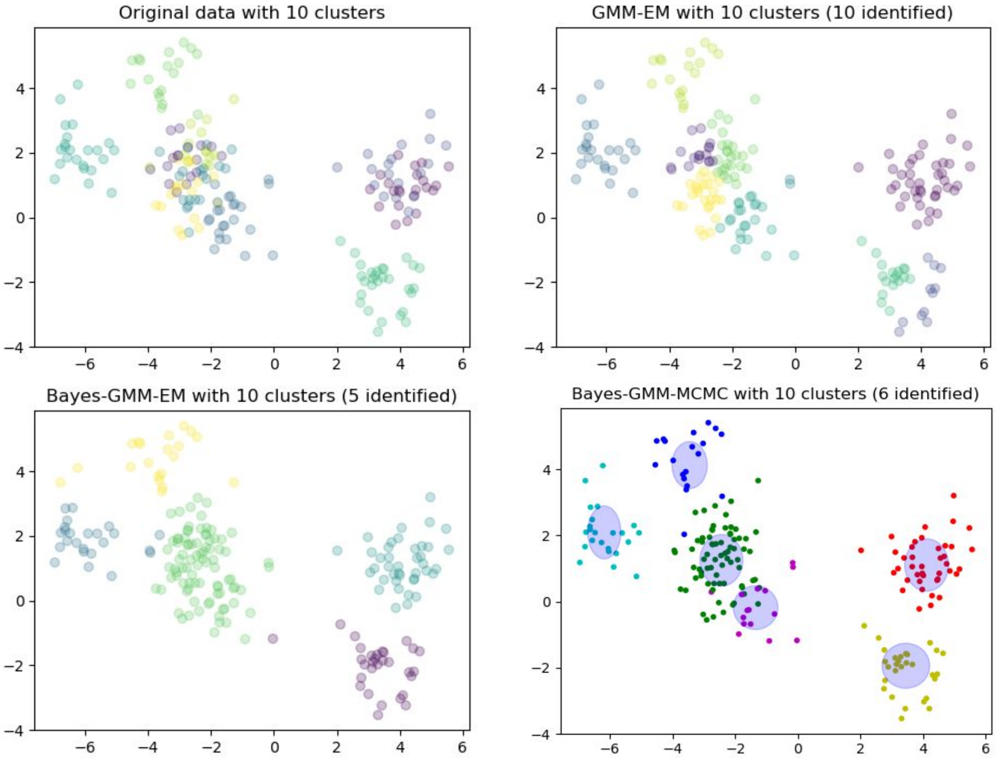
- ❏ Methods:
  - ❏ GMM based on EM
  - ❏ Bayesian GMM based on EM
  - ❏ Bayesian GMM based on MCMC

- ❏ Data set 1: Newsgroup20 (600 instances, 6 clusters, 600-dimension, sparse)

Method	GMM-EM	Bayes-GMM-EM	Bayes-GMM-MCMC
NMI	0.1902517	0.1325872	0.0450131
Identified components	6	6	2

Comparison between EM-GMM and Bayesian-GMM applied to network data clustering

- Data set 2: Synthetic (200 instances, 10 cluster, 2-dimension)


Original data with 10 clusters
GMM-EM with 10 clusters (10 identified)
Bayes-GMM-EM with 10 clusters (5 identified)
Bayes-GMM-MCMC with 10 clusters (6 identified)