BAYESIAN GAUSSIAN MIXTURE MODEL

David Esparza Alba
Ritsumeikan University

Basic Concepts

- First, we have to remember some important notions of our probability lessons.

- Consider we have two independents events, A and B, as shown in the figure.

Basic Concepts

$$P(A \cup B) = P(A) + P(B)$$

$$P(A \cap B) = P(A)P(B)$$

$$P(A^c) = 1 - P(A)$$

$$P(B^c) = 1 - P(B)$$

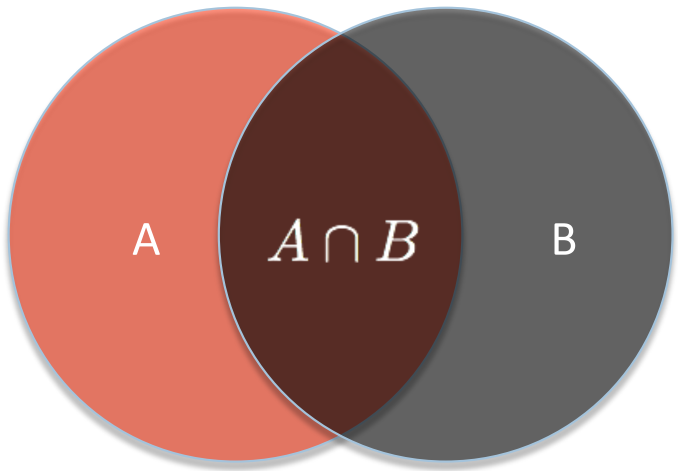
Basic Concepts

- What happen if these two sets are not independent?

Basic Concepts

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(B \cap A) = P(A|B)P(B) = P(B|A)P(A)$$



Bayes Theorem

- Bayes Theorem relates the conditional and marginal probabilities of events A and B, provided that the probability of B is not equal to zero.

- Prove:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We also know, that:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(B|A)P(A)$$

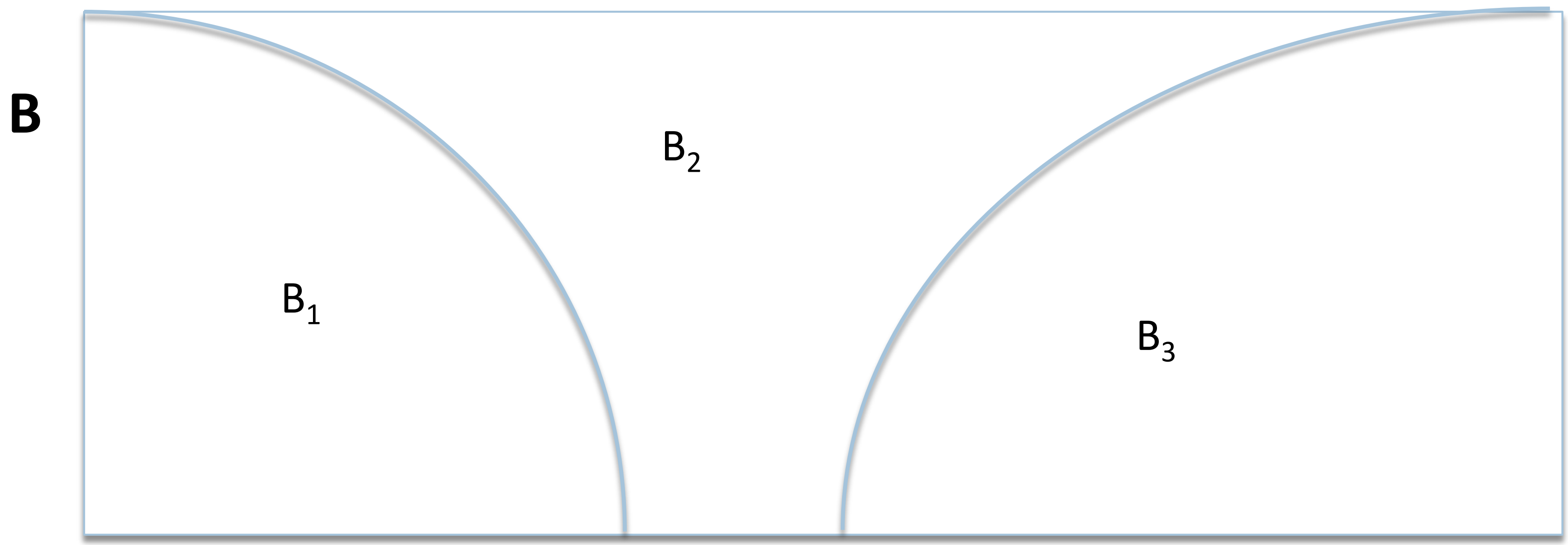
Then:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem

- Consider a probability space B, formed by the union of n different disjoints events, B = {B_1, ..., B_n}. The probability of B is given by:

$$P(B) = P(B \cap B_1) + ... + P(B \cap B_n)$$



Example: $P(B) = P(B \cap B_1) + P(B \cap B_2) + P(B \cap B_3)$

Bayes Theorem

- We can rewrite the last equation to:

$$
\begin{aligned}
P(B) &= P(B \cap B_1) + \ldots + P(B \cap B_n) \\
&= \sum_{i=1}^{n} P(B \cap B_i) \\
&= \sum_{i=1}^{n} P(B|B_i)P(B_i)
\end{aligned}
$$

- Then

$$
\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A \cap B)}{\sum_{i=1}^{n} P(B|B_i)P(B_i)}
\end{aligned}
$$

Bayes Theorem

- Bayes theorem says that:
- Where: $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A|B)$ is the posterior probability

$P(B|A)$ is called the likelihood

$P(A)$ is called the prior

$P(B)$ is a normalizing constant

Bayes Theorem

- According to Bayes Theorem we have the following:

$$posterior \propto likelihood \times prior$$
$$posterior = likelihood \times prior + const$$

- This relation is very important in the analysis of Bayesian statistics, and represent the basis for the Bayesian Gaussian Mixture Model which will be explained shortly.

- When the posterior is from the same family of the prior, we say the prior is a ***conjugate prior***.

Conjugate Prior Table

- Following there is a table of posterior and conjugate prior distributions relationships.

Posterior Distribution	**Model Parameter**	**Conjugate Prior**
Multinomial	$\mathbf{p}$	Dirichlet
Normal with unknown mean	μ	Normal
Normal with known mean	σ^2	Inverse Gamma
Multivariate Normal with unknown mean	$\boldsymbol{\mu}$	Multivariate Normal
Multivariate Normal with known mean	$\boldsymbol{\Sigma}$	Inverse-Wishart
Multivariate Normal	$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	Normal-Inverse-Wishart

Table 8.1

Probability Distributions

Discrete	Continuous
$p(x)$ – probability density	$f(x)$ – probability density
$p(x) \geq 0$	$f(x) \geq 0$
$\sum_x p(x) = 1$	$\int_{-\infty}^{\infty} f(x)dx = 1$
$P(x) = \sum_{x_i \leq x} p(x)$ - cumulative distribution	$F(x) = \int_{-\infty}^{x} f(x)dx$ - cumulative distribution
$E[g(x)] = \sum_x p(x)g(x)$	$P(a \leq X \leq b) = \int_a^b f(x)dx$
	$E[g(x)] = \int_{-\infty}^{\infty} f(x)g(x)dx$

$$Var(x) = E[x^2] - E[x]^2$$

Table 8.2

$$Cov(x, y) = E[xy] - E[x]E[y]$$

VARIATIONAL INFERENCE

Ritsumeikan University

David Esparza Alba

Variational Inference

- The main goal of variational inference is to estimate a set of parameters and latent variables denoted by **Z**, given only a set of observed data **X**. This is called, the Posterior Distribution:

Variational Inference

- The main goal of variational inference is to estimate a set of parameters and latent variables denoted by **Z**, given only a set of observed data **X**. This is called, the Posterior Distribution:

$$p(Z|X) = ?$$

Variational Inference

- The main goal of variational inference is to estimate a set of parameters and latent variables denoted by **Z**, given only a set of observed data **X**. This is called, the Posterior Distribution:

$$p(Z|X) = ?$$

- The objective is to find a variational distribution *q(**Z**)* that approximates the posterior distribution *p(**Z**|**X**)*.

$$p(Z|X) \approx q(Z)$$

Variational Inference

- How to know q(Z) is good enough?

- The Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions.

$$\begin{aligned} D_{KL}(q||p) &= \int_Z q(Z) \ln \frac{q(Z)}{p(Z|X)} dZ \\ &= \int_Z q(Z) \ln \frac{q(Z)}{p(Z,X)} dZ + \ln p(X) \end{aligned} \tag{8.1}$$

- We can define the Kullback-Leibler divergence as:

$$D_{KL}(q||p) = \ln p(X) - \mathcal{L}(q) \tag{8.2}$$

Variational Inference

- Then

$$\ln p(X) = D_{KL}(q||p) + \mathcal{L}(q) \tag{8.3}$$

- Where $\mathcal{L}(q)$ is called the "Lower Bound" and is defined by

$$\mathcal{L}(q) = -\int_Z q(Z) \ln \frac{q(Z)}{p(Z,X)} dZ \tag{8.4}$$

- Again, what we want to do is to make the difference between q(Z) and p(Z|X) small as possible, which is equal to minimize the Kullback-Leibler divergence, or maximize the Lower Bound.

Variational Inference

- Suppose the Z is divided into M disjoint groups, in that case, the variational distribution will be given by

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i) = \prod_{i=1}^{M} q_i \tag{8.5}$$

- Replacing this into equation (8.4), we have that the lower bound can be expressed as

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left[\ln p(X, Z) - \sum_i \ln q_i \right] \\
&= \int q_j \left[\int \ln p(X, Z \prod_{i \neq j} q_i dZ_i \right] - \int q_j \ln q_j dZ_j + const \\
&= \int q_j \ln \widetilde{p}(X, Z_j) dZ_j - \int q_j \ln q_j dZ_j + const
\end{aligned} \tag{8.6}$$

Variational Inference

- Where the new distribution $\widetilde{p}(X, Z_j)$ is defined by the relation

$$\ln \widetilde{p}(X, Z_j) = \mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right] + const \qquad (8.7)$$

- Then

$$\mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right] = \int \ln p(Z, X) \prod_{i \neq j} q_i dZ_i \qquad (8.8)$$

- Consider equation (8.6) as a negative Kullback_Leibler divergence between $q_j(Z_j)$ and $\widetilde{p}(X, Z_j)$, which reach its minimum when

$$q_j(Z_j) = \widetilde{p}(X, Z_j)$$

Variational Inference

- Then, the general expression for the optimal solution is expressed by

$$\begin{aligned}\ln q_j^*(Z_j) &= \widetilde{p}(X, Z_j) \\ &= \mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right] + const\end{aligned} \tag{8.9}$$

- If we take the exponential of both sides and normalize, we have

$$q_j^*(Z_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right]\right) dZ_j} \tag{810}$$

- In practice is often more convenient to work with equation (8.9) and then normalize (when required).

Variational Inference

- Then, the general expression for the optimal solution is expressed by

$$\begin{aligned} \ln q_j^*(Z_j) &= \widetilde{p}(X, Z_j) \\ &= \mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right] + const \end{aligned} \tag{8.9}$$

- If we take the exponential of both sides and normalize, we have

$$q_j^*(Z_j) = \frac{\exp\left(\mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right]\right)}{\int \exp\left(\mathbb{E}_{i \neq j}\left[\ln p(X, Z)\right]\right) dZ_j} \tag{810}$$

- In practice is often more convenient to work with equation (8.9) and then normalize (when required).

BAYESIAN GAUSSIAN MIXTURE MODEL

David Esparza Alba

Ritsumeikan University

Bayesian Gaussian Mixture Model (BGMM)

- For the GMM, the probability density function is given by:
- Where:

$$f(x|\pi, \mu, \Sigma) = \sum_{j=1}^{K} \pi_j \mathcal{N}(x|\mu_j, \Sigma_j) \quad (8.11)$$

K - Number of Gaussian functions

π_i - Mixing proportion of the i^{th} Gaussian

μ_i - Mean of the i^{th} Gaussian

Σ_i - Covariance matrix of the i^{th} Gaussian

Bayesian Gaussian Mixture Model (BGMM)

- The problem consists in estimate the set of parameters θ for each Gaussian function, then using the Bayes theorem, we can write the problem as:

$$f(\theta|x) \propto f(x|\theta)p(\theta)$$

- Where

$f(\theta|x)$ is the posterior distribution

$f(x|\theta)$ is the likelihood function

$f(\theta)$ is the prior distribution

Bayesian Gaussian Mixture Model (BGMM)

- The problem consists in estimate the set of parameters θ for each Gaussian function, then using the Bayes theorem, we can write the problem as:

- Where $$f(\theta|x) \propto f(x|\theta)p(\theta)$$

$f(\theta|x)$ is the posterior distribution

$f(x|\theta)$ is the likelihood function ← **Gaussian Mixture**

$f(\theta)$ is the prior distribution ← **See table xxx**

Bayesian Gaussian Mixture Model (BGMM)

- Knowing that the mixing proportions follow a multinomial distributions, then:

$$\sum_{i=1}^{K} \pi_i = 1 \tag{8.12}$$

- Based on table XXX, we are going to use a Dirichlet distribution as a prior of the mixing proportions.

$$p(\pi) = Dir(\alpha_0 * I_K)$$
$$p(\pi) = Dir(\alpha_0, ..., \alpha_0)$$

Bayesian Gaussian Mixture Model (BGMM)

- The conjugate prior over each covariance matrix is an Inverse-Wishart distribution.

$$p(\Sigma_k^{-1}) = p(\Lambda_k) = \mathcal{W}(W_0, v_0)$$

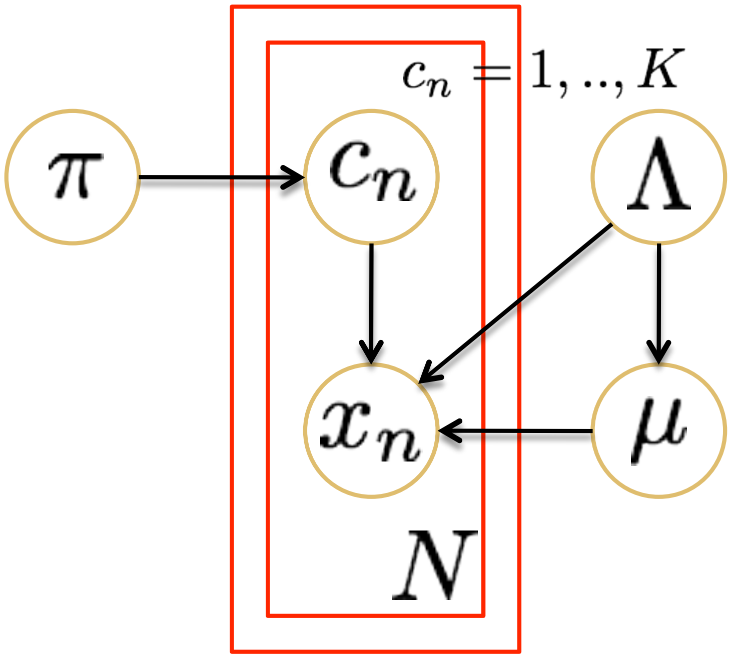
- The conjugate prior over the mean of each Gaussian is a multivariate Normal distribution.

$$p(\mu_k) = \mathcal{N}(m_0, (\beta_0\Lambda_k)^{-1})$$

The values α_0, W_0, v_o, m_0 and β_0 are called "hyperparameters"

Bayesian Gaussian Mixture Model (BGMM)

- The graph representation of our Gaussian model is given by the following diagram.



$$P(X, C, \pi, \mu, \Lambda) = p(X|C, \mu, \Lambda)p(\mu|\Lambda)p(\Lambda)p(C|\pi)p(\pi) \quad (8.13)$$

Bayesian Gaussian Mixture Model (BGMM)

- Now we are going to define each one of the terms in equation (8.13).

$$p(X|C,\mu,\Lambda) = \prod_{i=1}^{N}\prod_{j=1}^{K}\mathcal{N}(x_i|\mu_j,\Lambda_j^{-1})^{c_{ij}} \tag{8.14}$$

$$p(C|\pi) = \prod_{i=1}^{N}\prod_{j=1}^{K}\pi_j^{c_{ij}} \tag{8.15}$$

$$p(\pi) = Dir(\pi|\alpha_0) = C(\alpha_0)\prod_{j=1}^{K}\pi_j^{\alpha_0-1} \tag{8.16}$$

$$\begin{aligned} p(\mu,\Lambda) &= p(\mu|\Lambda)p(\Lambda) \\ &= \prod_{j=1}^{K}\mathcal{N}(\mu_j|m_0,(\beta_0\Lambda_j)^{-1})\mathcal{W}(\Lambda_j|W_0,v_0) \end{aligned} \tag{8.17}$$

- Where $C(\alpha_0)$ is the normalization constant for the Dirichlet distribution.

Bayesian Gaussian Mixture Model (BGMM)

- The main goal is to obtain the values of the parameters according to their posterior distributions, using a variational distribution q(θ).

- We can factorize the latent variables and the parameters from the variational distribution, so that:

$$q(C, \pi, \mu, \Lambda) = q(C)q(\pi, \mu, \Lambda) \tag{8.18}$$

- According to equation (8.9), we know that the optimal solution is given by:

$$\ln q_j^*(\theta_j) = \mathbb{E}_{i \neq j}[\ln p(X, \theta)] + const$$

Bayesian Gaussian Mixture Model (BGMM)

- The posterior distribution for the latent variables is given by:

$$\ln q^*(C) = \mathbb{E}_{\pi,\mu,\Lambda}[\ln p(X, C, \pi, \mu, \Lambda)] + const \quad (8.19)$$

- Considering only the terms that contains C from equation xxx, we have

$$\begin{aligned}\ln q^*(C) &= \mathbb{E}_{\pi}[\ln p(C|\pi)] + \mathbb{E}_{\mu,\Lambda}[\ln p(X|C, \mu, \Lambda)] + const \\ &= \sum_{i=1}^{N}\sum_{j=1}^{K} c_{ij} \ln \rho_{ij} + const \end{aligned} \quad (8.20)$$

where

$$\ln \rho_{ij} = \mathbb{E}[\ln \pi_j] + \frac{1}{2}\mathbb{E}[\ln |\Lambda_j|] - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\mathbb{E}_{\mu_j,\Lambda_j}[(x_i - \mu_j)^T \Lambda_j (x_i - \mu_j)] \quad (8.21)$$

Bayesian Gaussian Mixture Model (BGMM)

- Taking the exponential of both sides of xxx, we obtain:

$$q^*(C) \propto \prod_{i=1}^{N} \prod_{j=1}^{K} \rho_{ij}^{c_{ij}} \tag{8.22}$$

- Normalizing this distribution, we obtain

$$q^*(C) = \prod_{i=1}^{N} \prod_{j=1}^{K} r_{ij}^{c_{ij}} \tag{8.23}$$

where

$$r_{ij} = \frac{\rho_{ij}}{\sum_{h=1}^{K} \rho_{ih}} \tag{8.24}$$

Bayesian Gaussian Mixture Model (BGMM)

- Before to obtain the posterior distributions, is convenient to define three statistics of the observed data set.

$$N_j = \sum_{i=1}^{N} r_{ij} \tag{8.25}$$

$$\overline{x}_j = \frac{1}{N_j} \sum_{i=1}^{N} r_{ij} x_i \tag{8.26}$$

$$S_j = \frac{1}{N_j} \sum_{i=1}^{N} r_{ij} (x_i - \overline{x}_j)(x_i - \overline{x}_j)^T \tag{8.27}$$

Bayesian Gaussian Mixture Model (BGMM)

- Now we are going to proceed to find the factor $q(\pi, \mu, \Lambda)$ of the variational distribution.

- The log of the optimized factor is given by:

$$\ln q^*(\pi, \mu, \Lambda) = \mathbb{E}_C[\ln p(X, C, \pi, \mu, \Lambda)] + const$$

$$= \ln p(\pi) + \sum_{j=1}^{K} \ln p(\mu_j, \Lambda_j) + \mathbb{E}_C[\ln p(C|\pi)] + \sum_{j=1}^{K} \sum_{i=1}^{N} \mathbb{E}[c_{ij}] \ln \mathcal{N}(x_i|\mu_j, \Lambda_j^{-1}) + const \tag{8.28}$$

- We can see we have terms involving π and terms involving μ and Λ. Then we can factorize the factor in the following way:

$$q(\pi, \mu, \Lambda) = q(\pi) q(\mu, \Lambda)$$
$$= q(\pi) \prod_{j=1}^{K} q(\mu_j, \Lambda_j) \tag{8.29}$$

Bayesian Gaussian Mixture Model (BGMM)

- The next step is to find the variational posterior distribution for each one of the parameters, let's begin with the parameter π.

- Indentifying only the terms involving π from equation (8.28), we obtain

$$\begin{aligned}
\ln q^*(\pi) &= \ln p(\pi) + \mathbb{E}_C[\ln p(C|\pi)] + const \\
&= \ln\left(C(\alpha_0)\prod_{j=1}^{K}\pi_j^{\alpha_0-1}\right) + \mathbb{E}_C\left[\ln\left(\prod_{j=1}^{K}\prod_{i=1}^{N}\pi_j^{c_{ij}}\right)\right] + const \\
&= (\alpha_0 - 1)\sum_{j=1}^{K}\ln\pi_j + \sum_{j=1}^{K}\sum_{i=1}^{N}\mathbb{E}[c_{ij}]\ln\pi_j + const \\
&= (\alpha_0 - 1)\sum_{j=1}^{K}\ln\pi_j + \sum_{j=1}^{K}\sum_{i=1}^{N}r_{ij}\ln\pi_j + const \qquad (8.30)
\end{aligned}$$

Bayesian Gaussian Mixture Model (BGMM)

- Taking the exponential in both sides of equation (8.30), we have

$$
\begin{aligned}
q^*(\pi) &= \prod_{j=1}^{K} \pi_j^{\alpha_0-1} \prod_{j=1}^{K}\prod_{i=1}^{N} \pi_j^{r_{ij}} + const \\
&= \prod_{j=1}^{K} \pi_j^{\alpha_0-1} \prod_{j=1}^{K} \pi_j^{\sum_{i=1}^{N} r_{ij}} + const \\
&= \prod_{j=1}^{K} \pi_j^{\alpha_0-1} \prod_{j=1}^{K} \pi_j^{N_j} + const \\
&= \prod_{j=1}^{K} \pi_j^{\alpha_0+N_j-1} + const
\end{aligned}
\tag{8.31}
$$

Bayesian Gaussian Mixture Model (BGMM)

- With this, we can see the variational posterior distribution q* (π) is also a Dirichlet distribution, just as the prior.

$$p^*(\pi) \propto Dir(\pi|\alpha) \tag{8.32}$$

- Where α is a vector of size K, and components α_j given by

$$\boxed{\alpha_j = \alpha_0 + N_j} \tag{8.33}$$

Bayesian Gaussian Mixture Model (BGMM)

- Now we are going to find the variational posterior distribution of the parameter μ. For this we have to Identify the terms that contain the parameter μ from the equation (8.28).

$$
\begin{aligned}
\ln q^*(\mu) &= \sum_{j=1}^{K} \mathbb{E}_{\Lambda_j}\left[\ln p(\mu_j, \Lambda_j)\right] + \sum_{j=1}^{K}\sum_{i=1}^{N} \mathbb{E}[c_{ij}] \mathbb{E}_{\Lambda_j}\left[\ln \mathcal{N}(x_i|\mu_j, \Lambda_j^{-1})\right] + const \\
&= \sum_{j=1}^{K} \mathbb{E}_{\Lambda_j}\left[\ln(p(\mu_j|\Lambda_j)p(\Lambda_j))\right] + \sum_{j=1}^{K}\sum_{i=1}^{N} r_{ij} \mathbb{E}_{\Lambda_j}\left[\ln \mathcal{N}(x_i|\mu_j, \Lambda_j^{-1})\right] + const \\
&= \sum_{j=1}^{K} \mathbb{E}_{\Lambda_j}\left[\ln p(\mu_j|\Lambda_j)\right] + \sum_{j=1}^{K}\sum_{i=1}^{N} r_{ij} \mathbb{E}_{\Lambda_j}\left[\ln \mathcal{N}(x_i|\mu_j, \Lambda_j^{-1})\right] + const \\
&= \sum_{j=1}^{K} \mathbb{E}_{\Lambda_j}\left[\ln \mathcal{N}(\mu_j|m_0, (\beta_0\Lambda_j)^{-1})\right] + \sum_{j=1}^{K}\sum_{i=1}^{N} r_{ij} \mathbb{E}_{\Lambda_j}\left[\ln \mathcal{N}(x_i|\mu_j, \Lambda_j^{-1})\right] + const \\
&= \sum_{j=1}^{K} \ln\exp\left(-\frac{\beta_0}{2}(\mu_j - m_0)^T \mathbb{E}[\Lambda_j](\mu_j - m_0)\right) + \sum_{j=1}^{K}\sum_{i=1}^{N} \ln\exp\left(-\frac{1}{2} r_{ij}(x_i - \mu_j)^T \mathbb{E}[\Lambda_j](x_i - \mu_j)\right) + c \\
&= \sum_{j=1}^{K} -\frac{\beta_0}{2}(\mu_j - m_0)^T \mathbb{E}[\Lambda_j](\mu_j - m_0) - \sum_{j=1}^{K}\sum_{i=1}^{N} \frac{1}{2} r_{ij}(x_i - \mu_j)^T \mathbb{E}[\Lambda_j](x_i - \mu_j) + const \qquad (8.34)
\end{aligned}
$$

Bayesian Gaussian Mixture Model (BGMM)

- For the case in which D = 1, we have

$$\ln q^*(\mu) = \sum_{j=1}^{K} -\frac{\beta_0}{2}\mathbb{E}[\lambda_j](\mu_j - m_0)^2 - \sum_{j=1}^{K}\sum_{i=1}^{N}\frac{1}{2}r_{ij}\mathbb{E}[\lambda_j](x_i - \mu_j)^2 + const \qquad (8.35)$$

where

$$\lambda_j = \frac{1}{\sigma_j^2}$$

- Taking the exponential in both sides of the equation, we obtain

$$\begin{aligned}
q^*(\mu) &= \prod_{j=1}^{K}\exp\left(-\frac{\beta_0}{2}\mathbb{E}[\lambda_j](\mu_j - m_0)^2\right)\prod_{j=1}^{K}\prod_{i=1}^{N}\exp\left(-\frac{1}{2}r_{ij}\mathbb{E}[\lambda_j](x_i - \mu_j)^2\right) + const \\
&= \prod_{j=1}^{K}\exp\left(-\frac{\beta_0}{2}\mathbb{E}[\lambda_j](\mu_j - m_0)^2\right)\prod_{j=1}^{K}\exp\left(-\frac{1}{2}\mathbb{E}[\lambda_j]\sum_{i=1}^{N}r_{ij}(x_i - \mu_j)^2\right) + const \\
&= \prod_{j=1}^{K}\exp\left(-\frac{\beta_0}{2}\mathbb{E}[\lambda_j](\mu_j - m_0)^2 - \frac{1}{2}\mathbb{E}[\lambda_j]\sum_{i=1}^{N}r_{ij}(x_i - \mu_j)^2\right) + const \\
&= \prod_{j=1}^{K}\exp\left[-\frac{1}{2}\mathbb{E}[\lambda_j]\left(\beta_0(\mu_j - m_0)^2 + \sum_{i=1}^{N}r_{ij}(x_i - \mu_j)^2\right)\right] + const \qquad (8.36)
\end{aligned}$$

Bayesian Gaussian Mixture Model (BGMM)

- Expanding the binomials we get

$$
\begin{aligned}
q^*(\mu) &= \prod_{j=1}^{K} \exp\left[-\frac{1}{2}\mathbb{E}[\lambda_j]\left(\beta_0(\mu_j - m_0)^2 + \sum_{i=1}^{N} r_{ij}(x_i - \mu_j)^2\right)\right] + const \\
&= \prod_{j=1}^{K} \exp\left[-\frac{1}{2}\mathbb{E}[\lambda_j]\left(\beta_0(\mu_j^2 - 2\mu_j m_0 + m_0^2) + \sum_{i=1}^{N} r_{ij}(x_i^2 - 2x_i\mu_j + \mu_j^2)\right)\right] + const \\
&= \prod_{j=1}^{K} \exp\left[-\frac{1}{2}\mathbb{E}[\lambda_j]\left(\beta_0(\mu_j^2 - 2\mu_j m_0 + m_0^2) + \sum_{i=1}^{N} r_{ij}x_i^2 - 2\mu_j \sum_{i=1}^{N} r_{ij}x_i + \mu_j^2 \sum_{i=1}^{N} r_{ij}\right)\right] + const \\
&= \prod_{j=1}^{K} \exp\left[-\frac{1}{2}\mathbb{E}[\lambda_j]\left(\beta_0(\mu_j^2 - 2\mu_j m_0 + m_0^2) + \sum_{i=1}^{N} r_{ij}x_i^2 - 2\mu_j N_j \bar{x}_j + \mu_j^2 N_j\right)\right] + const \\
&= \prod_{j=1}^{K} \exp\left[-\frac{1}{2}\mathbb{E}[\lambda_j]\left(\mu_j^2(\beta_0 + N_j) - 2\mu_j(\beta_0 m_0 + N_j\bar{x}_j) + \left(\beta_0 m_0^2 + \sum_{i=1}^{N} r_{ij}x_i^2\right)\right)\right] + const \\
&= \prod_{j=1}^{K} \exp\left[-\frac{1}{2}(\beta_0 + N_j)\mathbb{E}[\lambda_j]\left(\mu_j^2 - 2\mu_j \frac{\beta_0 m_0 + N_j\bar{x}_j}{\beta_0 + N_j} + \frac{\beta_0 m_0^2 + \sum_{i=1}^{N} r_{ij}x_i^2}{\beta_0 + N_j}\right)\right] + const
\end{aligned}
$$

- Completing the square we obtain

$$
q^*(\mu) = \prod_{j=1}^{K} \exp\left[-\frac{1}{2}(\beta_0 + N_j)\mathbb{E}[\lambda_j]\left(\mu_j - \frac{\beta_0 m_0 + N_j\bar{x}_j}{\beta_0 + N_j}\right)^2\right] + const \qquad (8.37)
$$

Bayesian Gaussian Mixture Model (BGMM)

- We can see that the variational posterior distribution $q^*(\mu_j)$ is a normal distribution, as expected, because remember we are using a Normal distribution as a prior.

- According to equation, the updated hyperparameters obtained are:

$$\boxed{\begin{aligned} m_j &= \frac{\beta_0 m_0 + N_j \overline{x}_j}{\beta_0 + N_j} \\ \beta_j &= \beta_0 + N_j \end{aligned}}$$

$$m_j = \frac{\beta_0 m_0 + N_j \overline{x}_j}{\beta_0 + N_j} \tag{8.38}$$

$$\beta_j = \beta_0 + N_j \tag{8.39}$$

Bayesian Gaussian Mixture Model (BGMM)

- The variational posterior distribution for the covariance each covariance matrix q*(Λ_j) is a Wishart distribution with updated hyperparameters

$$\boxed{\begin{aligned} W_j^{-1} &= W_0^{-1} + N_j S_j + \tfrac{\beta_0 N_j}{\beta_0 + N_j}(\overline{x}_j - m_0)(\overline{x}_j - m_0)^T \\ v_j &= v_0 + N_j \end{aligned}}$$

(8.40)

(8.41)

- The procedure to obtain these updated hyperparameters is left as an exercise to the reader.

Bayesian Gaussian Mixture Model (BGMM)

- The last step consists in obtaining all the expectations included in equation (8.21) which are evaluated to give

$$\boxed{\begin{aligned} &\mathbb{E}_{\mu_j,\Lambda_j}[(x_i-\mu_j)^T\Lambda_j(x_i-\mu_j)] = D\beta_j^{-1} + v_j(x_i-m_j)^T W_j(x_i-m_j) \\ &\mathbb{E}[\ln|\Lambda_j|] = \sum_{i=1}^{D}\psi\left(\frac{v_j+1-i}{2}\right) + D\ln 2 + \ln|W_j| \\ &\mathbb{E}[\ln|\pi_j|] = \psi(\alpha_j) - \psi(\widehat{\alpha}) \end{aligned}}$$

(8.42)

(8.43)

(8.44)

where

$$\widehat{\alpha} = \sum_{i=1}^{K}\alpha_i$$

$$\psi(a) = \frac{d}{da}\ln\Gamma(a)$$

RESULTS

Ritsumeikan University

David Esparza Alba

Results

- The first step consists in the definition of the hyperparameters used in the model, which values were selected empirically.

$$K = 3$$

$$\alpha_0 = 1$$

$$W_0 = 0.01I$$

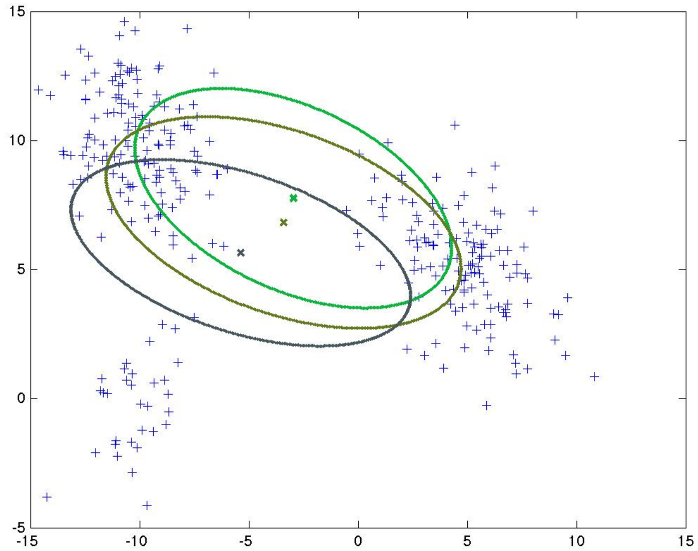
$$v_0 = d$$

$$m_0 = \bar{x}$$

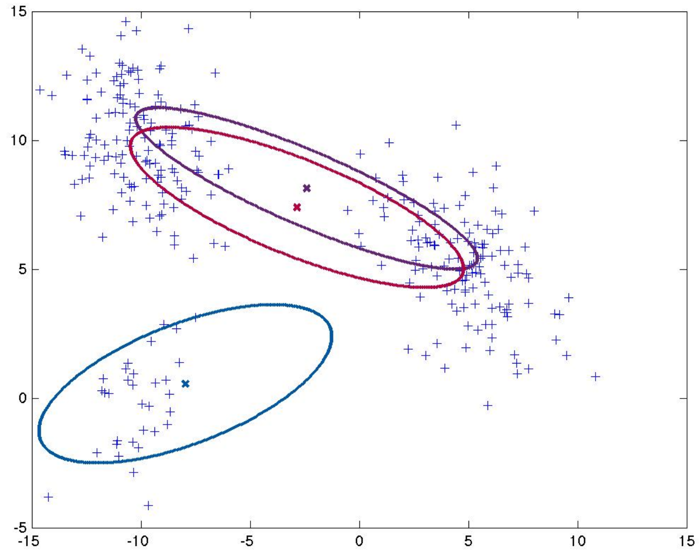
$$\beta_0 = 1$$

Results

BGMM after 1 iteration

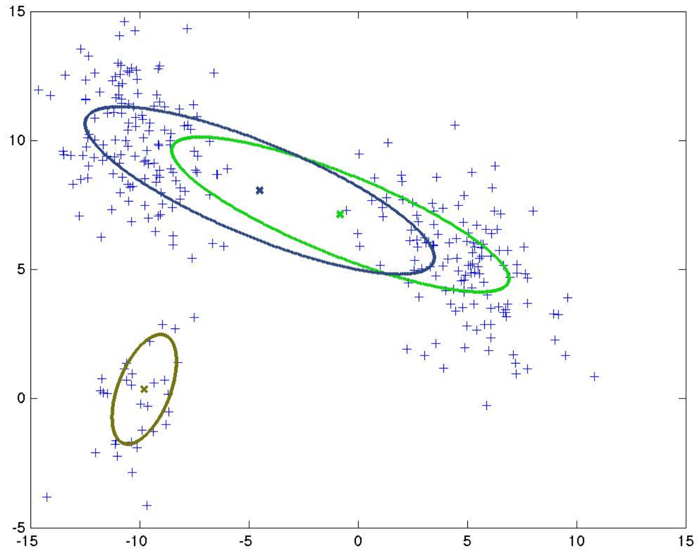


BGMM after 5 iterations

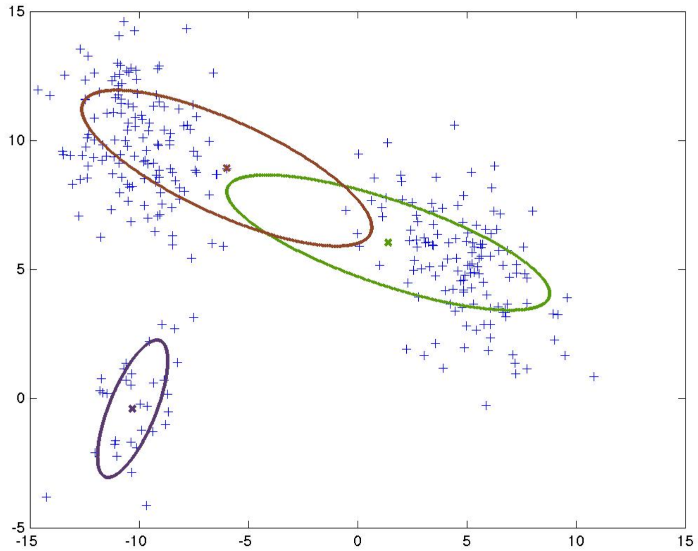


Results

BGMM after 15 iterations

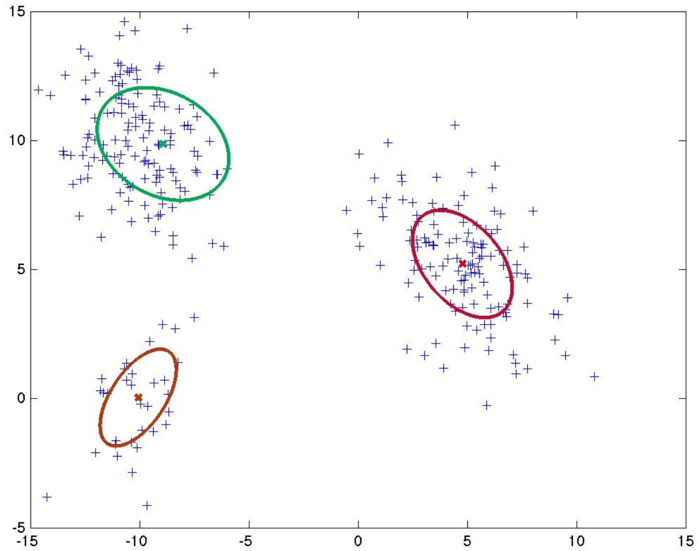


BGMM after 20 iterations

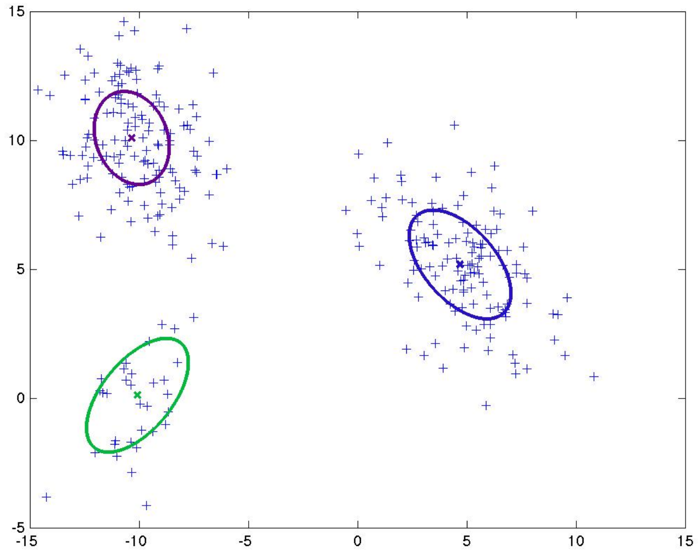


Results

BGMM after 25 iterations



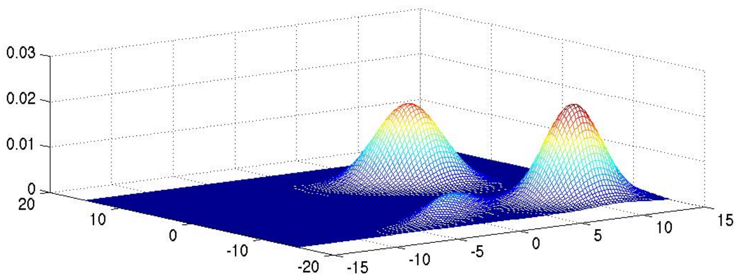
BGMM after 50 iterations

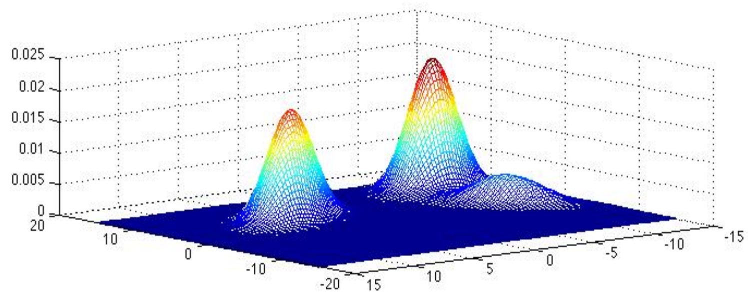


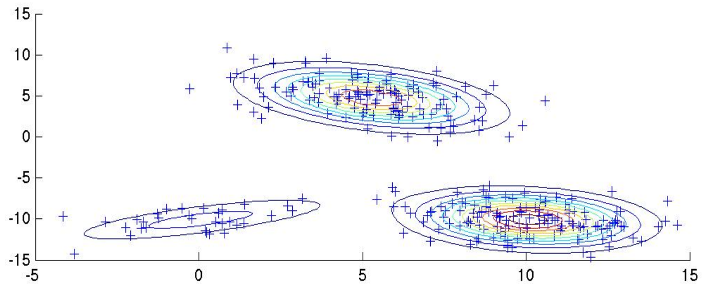
Results

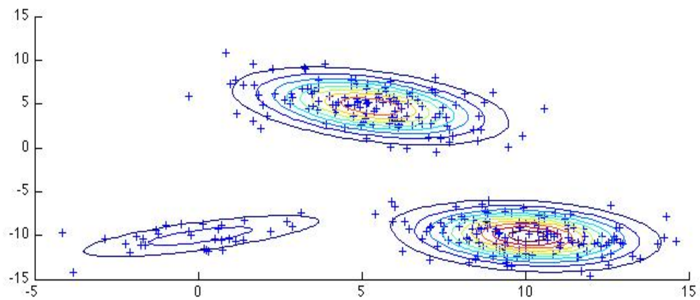
BGMM after 50 iterations

BGMM after 50 iterations

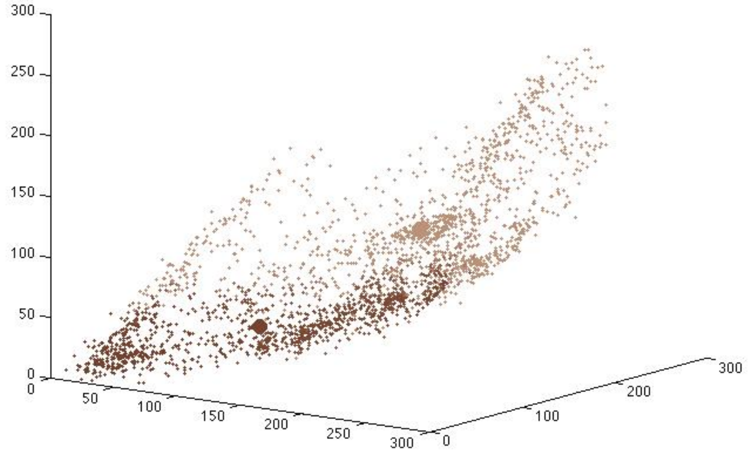




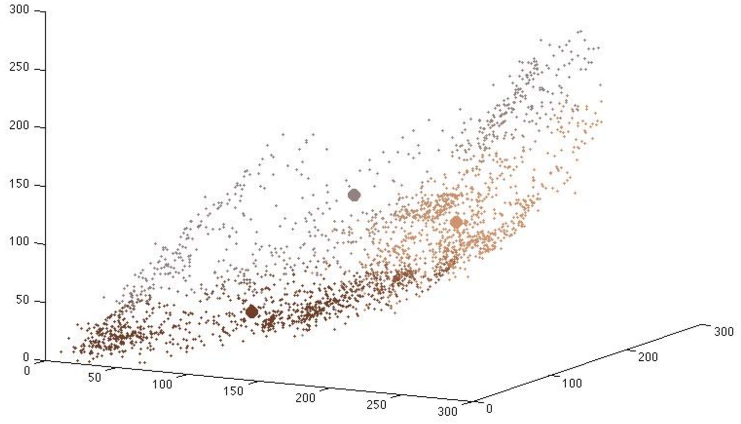




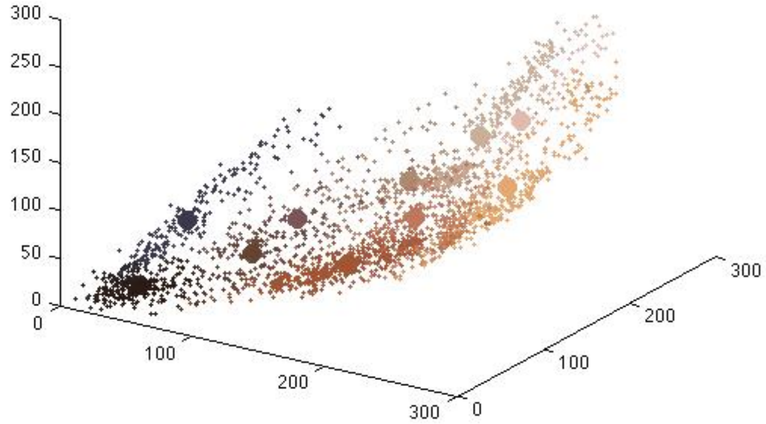
Results

BGMM with K = 2 Lena	BGMM with K = 2 Lena
	

Results

BGMM with K = 3 Lena	BGMM with K = 3 Lena
	

Results

BGMM with K = 10 Lena	BGMM with K = 10 Lena
	

Conclusions

Ritsumeikan University

David Esparza Alba

Conclusions

- It is convenient to use conjugate priors to update the parameters, because the process becomes easier.

- There is no need to estimate the expectation log-likelihood function, like in the EM algorithm.

- The use of a variational distribution make the calculus and math more complex. Other methods like "Sampling Methods" are suggested.

- Like in the EM, the BGMM needs a initial number of Gaussian fuctions, but computational is faster.