PGM - HW2 - Ayman CHAOUKI & Otmane SAKHI

1 Conditional independence and factorizations

Exercise 1.1

Let $X, Y \sim \mathcal{B}(\frac{1}{2})$ be two independent Bernoulli random variables and $Z = X + Y$, We also define:

$$T = \begin{cases} 0 \text{ if Z is even} \\ 1 \text{ otherwise} \end{cases}$$

Let's prove that $p \in \mathcal{L}(G)$
With the chain rule, $p(x, y, z, t) = p(x)p(y|x)p(z|x, y)p(t|z, x, y)$ when these conditional probabilities are defined.
If $x + y \neq z$ then $p(z|x, y) = 0$ and $p(x, y, z, t) = 0$, otherwise $p(z|x, y) = 1$ and $p(t|z, x, y) = p(t|z)$
because in that case we have $(\{Z = z\} \cap \{X + Y = z\} = \{Z = z\}) \Rightarrow p(z, x, y) = p(z)$
When all the conditional probabilities are defined $p(x, y, z, t) = p(x)p(y)p(z|x, y)p(t|z)$ ($p(y|x) = p(y)$ because $X \perp\!\!\!\perp Y$)
Let's prove that $X \perp\!\!\!\perp Y|T$
We have $p(X = 1, Y = 1|T = 1) = 0$, $p(X = 1|T = 1) = p(Y = 0) = \frac{1}{2}$ and $p(Y = 1|T = 1) = p(X = 0) = \frac{1}{2}$
$\Rightarrow p(X = 1, Y = 1|T = 1) \neq p(X = 1|T = 1)\, p(Y = 1|T = 1)$ and we have not $X \perp\!\!\!\perp Y|T$
Conclusion
We have then $p \in \mathcal{L}(G)$ but $X \perp\!\!\!\perp Y|T$

Exercise 1.2

a - Let $Z \sim \mathcal{B}(\pi)$ with $\pi \neq 0$ and X, Y two discrete random variable such that $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y|Z$
Let's prove that $X \perp\!\!\!\perp Z$ or $X \perp\!\!\!\perp Z$
Chain rule $\Rightarrow \forall x, y, z : p(x, y, z) = p(x)p(z|x)p(y|x, z) = p(x)p(z|x)p(y|z)$ because $X \perp\!\!\!\perp Y|Z$
Marginalizing over $Z \Rightarrow p(x, y) = p(x) \sum_z p(z|x)p(y|z) \Rightarrow p(x)p(y) = p(x) \sum_z p(z|x)p(y|z)$ Since $X \perp\!\!\!\perp Y$
We take $p(x) \neq 0 \forall x$ (if there is x s.t $p(x) = 0$ We define $X' = X$ on $X^{-1}(X(\Omega) \setminus \{x\})$), so we have
$\forall x, y : p(y) = \sum_z p(z|x)p(y|z)$ and we also have $\forall y : p(y) = \sum_z p(z)p(y|z)$
$\Rightarrow \forall x, y : \sum_z p(z)p(y|z) = \sum_z p(z|x)p(y|z)$ which leads to:

$$\begin{aligned} &\forall x, y : p(z = 1|x)p(y|z = 1) + (1 - p(z = 1|x))\, p(y|z) = \pi p(y|z = 1) + (1 - \pi)\, p(y|z = 0) \\ \iff &\forall x, y : (\pi - p(z = 1|x))\, (p(y|z = 1) - p(y|z = 0)) = 0 \end{aligned}$$

Suppose that Z and Y are not independent, then $\exists y : p(y|z = 1) \neq p(y|z = 0) \Rightarrow \forall x : p(z = 1|x) = \pi$ and
$\forall x : p(z = 0|x) = 1 - p(z = 1|x) = 1 - \pi = p(z = 0)$, which means that $X \perp\!\!\!\perp Z$ (1)
Suppose that Z and X are not independent, then $\exists x : p(z = 1|x) \neq \pi \Rightarrow \forall y : p(y|z = 1) = p(y|z = 0)$
Then $\forall y : \frac{p(y,z=1)}{\pi} = \frac{p(y,z=0)}{1-\pi} \Rightarrow \forall y : p(y, z = 1) = \pi (p(y, z = 0) + p(y, z = 1)) = \pi p(y)$ and we also have
$\forall y : p(y, z = 0) = p(y) - p(y, z = 1) = (1 - \pi)p(y)$ which means that $\forall y : p(y, z) = p(y)p(z)$, $Y \perp\!\!\!\perp Z$ (2)
Conclusion : From (1) and (2) We conclude that $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$

b - Let's disprove the previous statement in the general case :
We define$X \sim \mathcal{B}(\frac{1}{2})$, $T \sim Multinomial\left(\{-1, 0, 1\}, \left(\frac{1}{3}, \frac{4}{9}, \frac{2}{9}\right)\right)$, $Z = \begin{cases} 1 + T \text{ if } X = 0 \\ W \sim \mathcal{U}(\{0, 1, 2\}) \text{ otherwise} \end{cases}$, $Y = \begin{cases} U \sim \mathcal{B}\left(\frac{1}{3}\right) \text{ if } Z = 0 \\ V \sim \mathcal{B}\left(\frac{1}{2}\right) \text{ otherwise} \end{cases}$
With $U \perp\!\!\!\perp X$, $V \perp\!\!\!\perp X$ $U \perp\!\!\!\perp Z|X$ and $V \perp\!\!\!\perp Z|X$
We briefly check that $X \perp\!\!\!\perp Z$ and $Y \perp\!\!\!\perp Z$:
$p(z = 1) = p(z = 1|x = 0)p(x = 0) + p(z = 1|x = 1)p(x = 1) + p(z = 1|x = 2)p(x = 2) = \frac{7}{18}$ but $p(z = 1|x = 0) = \frac{4}{9}$, then $X \perp\!\!\!\perp Z$
$p(y = 0|z = 0) = \frac{2}{3}$ and $p(y = 0|z = 1) = p(y = 0|z = 2) = \frac{1}{2}$, then $p(y = 0|z = 0) \neq p(y = 0|z = 1)$ and $Y \perp\!\!\!\perp Z$
Now let's check that $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Y$:
$\forall x, y \in \{0, 1\} : p(x, y|z = 0) = p(x, u|z = 0) = \frac{p(x)p(u)p(z=0|x)}{p(z=0)}$ because $Y = U$ when $Z = 0$, $U \perp\!\!\!\perp X$ and $Z \perp\!\!\!\perp U|X$
$\forall x, y : p(x, y|z = 0) = p(x|z = 0)p(y|z = 0)$ because $p(u) = p(y|z = 0)$. And in the same fashion we have:
$\forall x, y \in \{0, 1\} : p(x, y|z \neq 0) = p(x, v|z \neq 0) = \frac{p(x)p(v)p(z\neq 0|x)}{p(z\neq 0)}$ because $Y = V$ when $Z \neq 0$, $V \perp\!\!\!\perp X$ and $Z \perp\!\!\!\perp V|X$
$\forall x, y : p(x, y|z \neq 0) = p(x|z \neq 0)p(y|z \neq 0)$ because $p(v) = p(y|z \neq 0)$. Then $\forall x, y, z : p(x, y|z) = p(x|z)p(y|z)$, conclusion $X \perp\!\!\!\perp Y|Z$
$\forall x, y \in \{0, 1\} : p(y|x) = \frac{p(y,x)}{p(x)} = \frac{\sum_z p(x)p(z|x)p(y|z)}{p(x)} = \sum_z p(z|x)p(y|z)$ because $X \perp\!\!\!\perp Y|Z$ and we have then
$\forall x, y \in \{0, 1\} : p(y|x) = p(z = 0|x)p(y|z = 0) + p(z = 1|x)p(y|z = 1) + p(z = 2|x)p(y|z = 2)$
We calculate $p(y|x)$ $\forall (x, y) \in \{0, 1\}^2$ since the involved probabilities are easily computed from the random variables definitions:
$p(y = 0|x = 0) = \frac{5}{9}$, $p(y = 0|x = 0) = 1 - p(y = 1|x = 0) = \frac{4}{9}$, $p(y = 0|x = 1) = \frac{5}{9}$ and $p(y = 1|x = 1) = 1 - p(y = 0|x = 1) = \frac{4}{9}$
Marginalizing w.r.t Z $\Rightarrow p(y = 0) = \frac{5}{9}$ and $p(y = 1) = \frac{4}{9}$, which means that $\forall x, y \in \{0, 1\} : p(y) = p(y|x)$, conclusion $X \perp\!\!\!\perp Y$
Conclusion : We have $X \perp\!\!\!\perp Y|Z$, $X \perp\!\!\!\perp Y$ but Z is not independent from X and Z is not independent from Y

2 Distributions factorizing in a graph

Exercise 2.1

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $i \to j \in \mathcal{E}$ is covered edge. We define $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ where $\mathcal{E}' = (\mathcal{E} \setminus \{i \to j\}) \cup \{j \to i\}$. We have:

$$p \in \mathcal{L}(\mathcal{G}) \iff p(x) = p(x_i|x_{\pi_i})\, p\left(x_j|x_{\pi_j}\right) \prod_{k \notin \{i,j\}} p(x_k|x_{\pi_k})$$

$$p \in \mathcal{L}(\mathcal{G}') \iff p(x) = p(x_i|x_{\pi_i}, x_j)\, p(x_j|x_{\pi_i}) \prod_{k \notin \{i,j\}} p(x_k|x_{\pi_k})$$

Now we calculate:

$$\begin{aligned}
\forall x_i, x_{\pi_i}, x_j, x_{\pi_j} : p(x_i|x_{\pi_i} p(x_j|x_{\pi_j})) &= \frac{p(x_i, x_{\pi_i})}{p(x_{\pi_i})} \frac{p(x_j, x_{\pi_j})}{p(x_{\pi_j})} \\
&= \frac{p(x_i, x_{\pi_i})}{p(x_{\pi_i})} \frac{p(x_j, x_{\pi_j})}{p(x_i, x_{\pi_i})} \text{ , Because } \pi_j = \pi_i \cup \{i\} \\
&= \frac{p(x_j, x_{\pi_j})}{p(x_{\pi_i})} \\
&= \frac{p(x_j, x_{\pi_i}, x_i)}{p(x_{\pi_i})} \\
&= \frac{p(x_j, x_{\pi_i}, x_i)}{p(x_j, x_{\pi_i})} \frac{p(x_j, x_{\pi_i})}{p(x_{\pi_i})} \\
&= p(x_i|x_{\pi_i}, x_j) p(x_j|x_{\pi_i})
\end{aligned}$$

Which means that $(p \in \mathcal{L}(\mathcal{G}) \iff p \in \mathcal{L}(\mathcal{G}'))$ and we conclude that $\mathcal{L}(\mathcal{G}) = \mathcal{L}(\mathcal{G}')$

Exercise 2.2

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a directed tree and $\mathcal{G}'(\mathcal{V}', \mathcal{E}')$ its corresponding undirected tree. $\mathcal{C}_{max}$ denotes the set of maximal cliques of $\mathcal{G}'$
Let's prove that : $\forall C \in \mathcal{C}_{max} : card(C) = 2$
We know that $\forall v :\in \mathcal{V}'$, v has a parent or a child, so $\forall C \in \mathcal{C}_{max} card(C) \geq 2$
Suppose that $\exists C \in \mathcal{C}_{max} : card(C) \geq 3$, then $\exists C' \subset C$ such that C' is a clique and $card(C') = 3$
C' would correspond then to either a loop or a v-structure in $\mathcal{G}$! **Contradiction**
We conclude that $\forall C \in \mathcal{C} : card(C) = 2$ and $\mathcal{C}_{max} = \{\{x_i, x_{\pi_i}\}, i \in \mathcal{V}\}$. We have then

$$p \in \mathcal{L}(\mathcal{G}') \iff p(x) = \frac{1}{Z} \prod_i \Psi_i(x_i, x_{\pi_i})$$

Where $\Psi_i \geq 0$ and Z a normalizing factor.
Let's prove that $\mathcal{L}(\mathcal{G}) \subset \mathcal{L}(\mathcal{G}')$
Let $p \in \mathcal{L}(\mathcal{G})$, then $p(x) = \prod_i p(x_i|x_{\pi_i})$, with $p(x_i|x_{\pi_i}) \geq 0$ and $\sum_i p(x_i|x_{\pi_i}) = 1$
We define $\forall i : \Psi_i(x_i, x_{\pi_i}) = p(x_i|x_{\pi_i})$ and $Z = 1$
We get $p(x) = \prod_i \Psi_i(x_i, x_{\pi_i}) \in \mathcal{L}(\mathcal{G}')$, **Conclusion :** $\mathcal{L}(\mathcal{G}) \subset \mathcal{L}(\mathcal{G}')$ (1)
Let's prove that $\mathcal{L}(\mathcal{G}') \subset \mathcal{L}(\mathcal{G})$
Let $p \in \mathcal{L}(\mathcal{G}')$, then $p(x) = \frac{1}{Z} \prod_i \Psi_i(x_i, x_{\pi_i})$ with $Z = \sum_x \prod_{x_i} \Psi_i(x_i, x_{\pi_i})$ and $\forall i : \Psi_i \geq 0$
Let n be the number of nodes, for $n = 1$ we have $p(x) = \frac{\Psi_1(x_1)}{\sum_{x_1} \Psi(x_1)}$, by putting $\Psi'(x_1) = \frac{\Psi_1(x_1)}{\sum_{x_1} \Psi(x_1)}$, we have $\sum_{x_1} \Psi'(x_1) = 1$
By induction, suppose that $\forall p \in \mathcal{L}(\mathcal{G})$ where $card(\mathcal{V}) = n$ we have $\forall i, \exists \Psi_i \geq 0 : p(x) = \prod_{i=1}^{n} \Psi_i(x_i, x_{\pi_i})$ and $\sum_{x_i} \Psi_i(x_i, x_{\pi_i}) = 1$
Let $p \in \mathcal{L}(\mathcal{G}')$ with $card(\mathcal{V}) = n + 1$: we have $p(x) = \frac{1}{Z} \prod_{i=1}^{n} \Psi_i(x_i, x_{\pi_i}).\Psi_{n+1}(x_{n+1}, x_{\pi_{n+1}})$
$Z = \sum_x \prod_{i=1}^{n+1} \Psi_i(x_i, x_{\pi_i}) = \left(\sum_{x_{n+1}} \Psi_{n+1}(x_{n+1}, x_{\pi_{n+1}})\right) (\sum_x \prod_{i=1}^{n} \Psi_i(x_i, x_{\pi_i}))$, notice that we took $n + 1$ as a leaf.
then $p(x) \frac{\sum_{x_{n+1}} \Psi_{n+1}(x_{n+1}, x_{\pi_{n+1}})}{\Psi_{n+1}\left(x_{n+1}, x_{\pi_{n+1}}\right)}$ factorizes in a tree of $card(\mathcal{V}) = n$, with the induction hypothesis we have
$\forall i, \exists \Psi'_i \geq 0 : p(x) \frac{\sum_{x_{n+1}} \Psi_{n+1}(x_{n+1}, x_{\pi_{n+1}})}{\Psi_{n+1}\left(x_{n+1}, x_{\pi_{n+1}}\right)} = \prod_{i=1}^{n} \Psi'_i(x_i, x_{\pi_i})$ where $\forall i : \sum_{x_i} \Psi'_i(x_i, x_{\pi_i}) = 1$, we set
$\Psi'_{n+1}(x_{n+1}, x_{\pi_{n+1}}) = \frac{\Psi_{n+1}\left(x_{n+1}, x_{\pi_{n+1}}\right)}{\sum_{x_{n+1}} \Psi_{n+1}(x_{n+1}, x_{\pi_{n+1}})}$, we get $\sum_{x_{n+1}} \Psi'_{n+1}(x_{n+1}, x_{\pi_{n+1}}) = 1$ and $p(x) = \prod_{i=1}^{n+1} \Psi'_i(x_i, x_{\pi_i}) \in \mathcal{L}(\mathcal{G})$
Conclusion : $\mathcal{L}(\mathcal{G}') \subset \mathcal{L}(\mathcal{G})$ (2)
From (1) and (2) we conclude that $\mathcal{L}(\mathcal{G}) = \mathcal{L}(\mathcal{G}')$

3 Implementation - Gaussian mixtures

Exercice 3a :

In the Kmean algorithm, We choose the initial cluster centers randomly from the data points distribution without replacement. Even with this random initialization, we converge in almost all cases to similar final centers and pretty close distortion measures. However, sometimes we draw points close to each other, pushing the algorithm to converge to a poor local minima. You can check the figures below for more insight.

Exercice 3b :

In the M step of the itetation $t+1$, and by defining $\tau_{i,t}^j = \tau_i^j(\theta_t)$ we need to maximize :

$$F_{t+1}(\Pi, \mu, \Sigma) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,t}^j \log(\pi_j) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,t}^j \left[\log(\frac{1}{(2\pi)^{\frac{k}{2}}}) + \log(\frac{1}{|\Sigma_j|^{\frac{1}{2}}}) + -\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\right]$$

We consider the case where the covariance matrices are proportional to the identity, $\Sigma_j = \sigma_j^2 I_d$ with $d=2$ in our problem. The covriance matrices parameters get reduced to $\boldsymbol{\sigma} = (\sigma_j^2)_{j\in 1,\dots,k}$ and the functional to maximize becomes written as :

$$F_{t+1}(\Pi, \mu, \boldsymbol{\sigma}) = \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,t}^j \log(\pi_j) + \sum_{i=1}^{n}\sum_{j=1}^{k} \tau_{i,t}^j \left[\log(\frac{1}{(2\pi)^{\frac{k}{2}}}) + \log(\frac{1}{\sigma_j^d}) - \frac{1}{2\sigma_j^2}(x_i-\mu_j)^T(x_i-\mu_j)\right]$$

- Π estimation: $\arg\max_\Pi F_{t+1}(\Pi, \mu, \boldsymbol{\sigma}) = \arg\max_\Pi \sum_{i=1}^n\sum_{j=1}^k \tau_{i,t}^j \log(\pi_j)$ which is the value that maximizes the loglikelihood of a multinomial distribution, thus we can directly write that : $\boldsymbol{\pi_{j,t+1} = \frac{1}{n}\sum_{i=1}^n \tau_{i,t}^j} \quad \forall j$

- μ estimation: $\arg\max_\mu F_{t+1}(\Pi, \mu, \boldsymbol{\sigma}) = \arg\max_\mu -\sum_{i=1}^n\sum_{j=1}^k \tau_{i,t}^j (x_i-\mu_j)^T(x_i-\mu_j)$. wich is concave on μ. Setting its gradient to zero leads to the following equations : $\sum_{i=1}^n \tau_{i,t}^j(x_i-\mu_{j,t+1}) = 0 \quad \forall j$, which gives us the estimator at t+1:

$$\boldsymbol{\mu_{j,t+1} = \frac{\sum_{i=1}^n \tau_{i,t}^j x_i}{\sum_{i=1}^n \tau_{i,t}^j}} \quad \forall j$$

- $\boldsymbol{\sigma}$ Estimation: $\arg\max_{\boldsymbol{\sigma}} F_{t+1}(\Pi, \mu_{t+1}, \boldsymbol{\sigma}) = \arg\max_{\boldsymbol{\sigma}} \sum_{i=1}^n\sum_{j=1}^k \tau_{i,t}^j \left[-d\log(\sigma_j) - \frac{1}{2\sigma_j^2}(x_i-\mu_{j,t+1})^T(x_i-\mu_{j,t+1})\right]$. We know that this functional is concave on it's natural parameter $-\frac{1}{\sigma^2}$ as it is an exponential family, so we can get the maximum by setting the gradient wrt to σ to zero:

$$\begin{aligned}
\nabla_{\boldsymbol{\sigma}} F_{t+1} = 0 &\iff \sum_{i=1}^n \tau_{i,t}^j \left[\frac{-d}{\sigma_{j,t+1}} + \frac{1}{\sigma_{j,t+1}^3}(x_i-\mu_{j,t+1})^T(x_i-\mu_{j,t+1})\right] = 0 \quad \forall j \\
&\iff \frac{1}{\sigma_{j,t+1}^2}\sum_{i=1}^n \tau_{i,t}^j (x_i-\mu_{j,t+1})^T(x_i-\mu_{j,t+1}) = d\sum_{i=1}^n \tau_{i,t}^j \quad \forall j \\
&\iff \boldsymbol{\sigma_{j,t+1}^2 = \frac{\sum_{i=1}^n \tau_{i,t}^j (x_i-\mu_{j,t+1})^T(x_i-\mu_{j,t+1})}{d\sum_{i=1}^n \tau_{i,t}^j}} \quad \forall \boldsymbol{j}
\end{aligned}$$

Exercice 3c :

For the general case, the estimator at the t+1 iteration of the covariance matrix is written as : $\boldsymbol{\Sigma_{j,t+1} = \frac{\sum_{i=1}^n \tau_{i,t}^j (x_i-\mu_{j,t+1})(x_i-\mu_{j,t+1})^T}{\sum_{i=1}^n \tau_{i,t}^j}}$ for all j, The results of our implementation can be seen in the figures below.

Exercice 3d :

To compare the models, we used the Normalized loglikelihood that is independent of the number of points in each dataset. We can clearly see from the data distribution/Kmeans convergence that the groupments are no circles, so the use of the isotropic assumption is not justified. The normalized Loglikelihoods show that the **General Gaussian Mixture model** outperforms the Isotropic model in the training set (it can be normal because we generaly have a better fit for models with less assumptions/more parameters) but also in the test set, which means that our model did not overfit and it indeed, matches the assumptions of our data distribution.

Normalized Loglikelihoods

Model	Train	Test
Isotropic	-5.3637	-5.4512
General	**-4.6915**	**-4.8535**

4 Figures

Kmeans

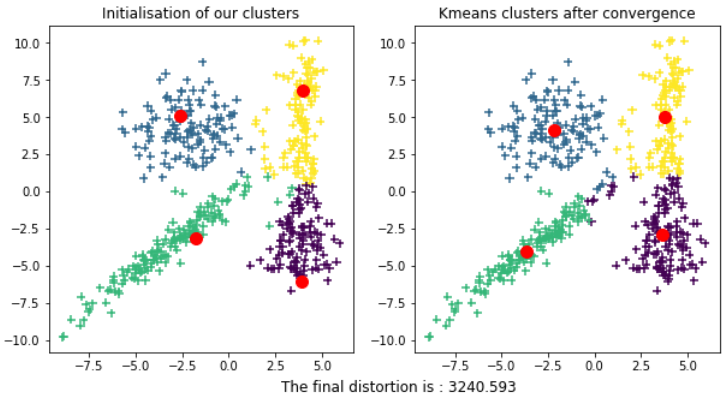

Figure 1: Kmeans with good initialization

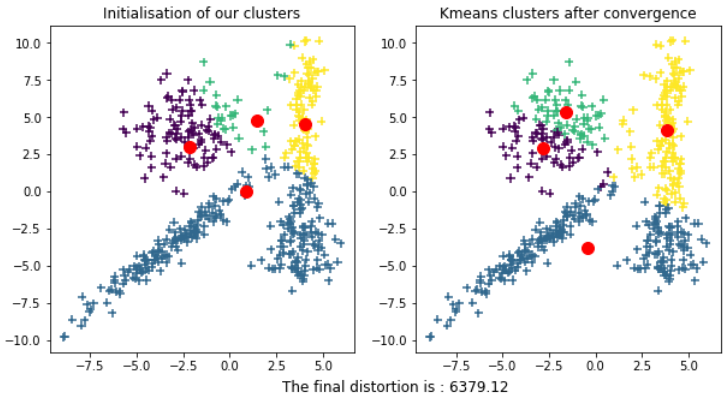

Figure 2: Kmeans with poor initialization

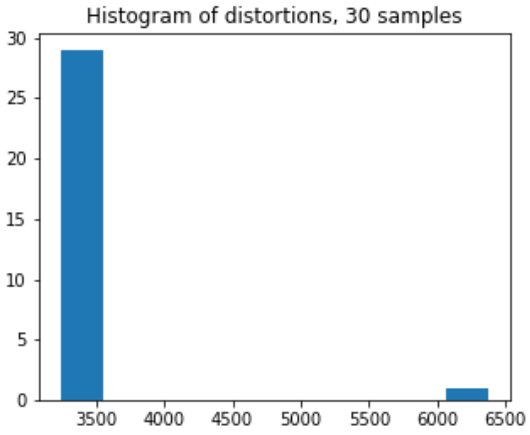

Figure 3: Kmeans distortion histogram

Gaussian Mixtures - EM

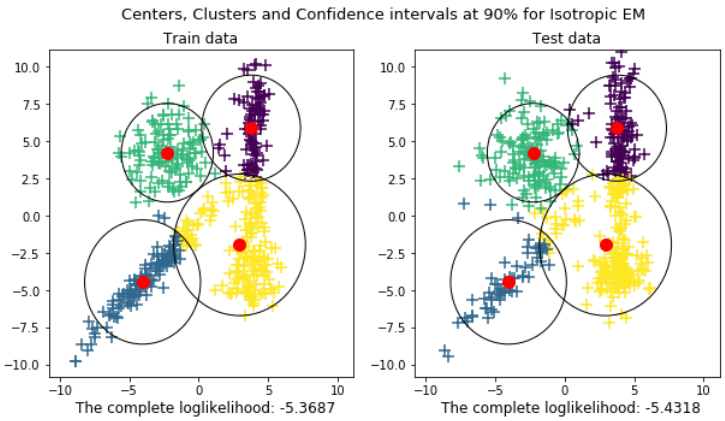

Figure 4: Isotropic GMM results

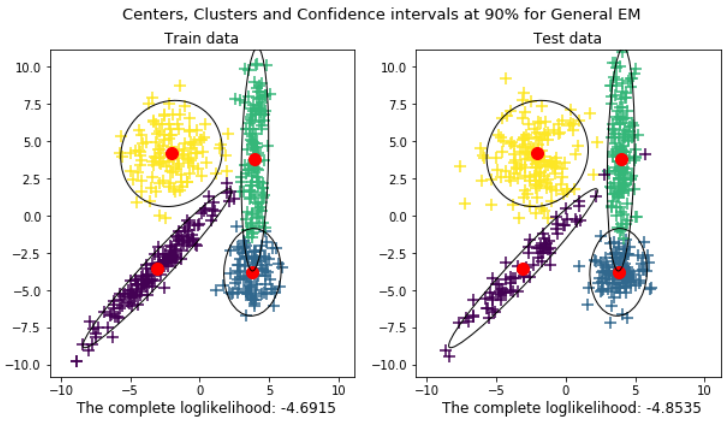

Figure 5: General GMM results