HWK 2 - Graphical Models

Alejandro de la Concha (alexdavidhalo@gmail.com)

Tong ZHAO (tong.zhao@eleves.enpc.fr)

1 Conditional Independence and Factorizations

(1) Given a joint distribution $p \in \mathcal{L}(G)$, the implied factorization is:

$$p(x, y, z, t) = p(x)p(y)p(z|x, y)p(t|z)$$

In this model, $X \perp\!\!\!\perp Y|T$ does not hold for most of $p \in \mathcal{L}$. A and B is not d-seperated T because the chain (X, Z, Y) is a v-structure and T is a descendant of Z. As a counter-example, we consider a binary model (meaning that all the nodes take value from $\{0, 1\}$) and it follows the rules as below:

$$\begin{cases} p(X = 0) = p(Y = 0) = 0.5 \\ Z = 0 \ if \ X = Y, Z = 1 \ otherwise \\ T = Z \end{cases}$$

If we know $t = 1$, we have $x = y$. $p(x, y|t) = p(x|t) > p(x|t)p(y|t)$, so X and Y are not conditional independent given T.

(2a) It is true if Z is a binary variable. From the assumptions, we have:

$$\begin{aligned} P(X) = \frac{P(X, Y)}{P(Y)} &= \frac{1}{P(Y)} \sum_z P(X, Y|z)P(z) \\ &= \frac{1}{P(Y)} \sum_z P(X|z)P(Y|z)P(z) = \sum_z P(X|z)P(z|Y) \end{aligned}$$

So we have:

$$P(X = x_i) = P(X = i|z = 0)P(z = 0|Y) + P(X = i|z = 1)P(z = 1|Y) \tag{1}$$

The equation (1) should hold for all possible ys. We take randomly two values y_1 and y_2, and then we have:

$$\begin{aligned} P(X = x_i) &= P(X = x_i|Z = 0)P(Z = 0|Y = y_1) + P(X = x_i|Z = 1)P(Z = 1|Y = y_1) \\ &= P(X = x_i|Z = 0)P(Z = 0|Y = y_2) + P(X = x_i|Z = 1)P(Z = 1|Y = y_2) \end{aligned}$$

which means that:

$$\begin{aligned}&P(X=x_i|Z=0)\Big(P(Z=0|Y=y_1)-P(Z=0|Y=y_2)\Big)+\\&P(X=x_i|Z=1)\Big(P(Z=1|Y=y_1)-P(Z=1|Y=y_2)\Big)=0\end{aligned}$$

Since $P(Z=1|Y)+P(Z=0|Y)=1$, we can simplify the above equation as:

$$\Big(P(X=x_i|Z=0)-P(X=x_i|Z=1)\Big)\Big(P(Z=0|Y=y_1)-P(Z=0|Y=y_2)\Big)=0 \qquad (2)$$

The equation (2) holds for $\forall x_i \in X$ and $\forall y_1, y_2 \in Y$. So either item should be 0.

If $P(X=x_i|Z=0)-P(X=x_i|Z=1)=0$, we have $X \perp\!\!\!\perp Z$.

If $P(Z=0|Y=y_1)-P(Z=0|Y=y_2)=0$, we have:

$$\frac{P(Y=y_1|Z=0)}{P(Y=y_1)}=\frac{P(Y=y_2|Z=0)}{P(Y=y_2)}=\cdots=k \text{ (constant)}$$

$$\sum_{y\in Y}P(Y=y|Z=0)=k\sum_{y\in Y}P(Y=y)=k=1$$

So we have $P(Y=y|Z=0)=P(Y=y)$ holds for $\forall y \in Y$, which means that $Y \perp\!\!\!\perp Z$.

From the above proof, we deduce that if Z is a binary variable, the statement is true.

(2b) In general the statement is not true. Here we show a counter-example. Suppose that X takes values in $\{1,\dots,M\}$ and Y takes values in $\{1,\dots,N\}$. $Z=X+Yi$ is a complex number. Given Z, we have $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp Y$. However, X depends on Z and Y depends on Z.

2 Distributions Factorizing in a Graph

(1) First of all we show that G' is a DAG. If it exists a cycle after we flipped a covered edge $i \to j$, then the cycle must include the edge $j \to i$. (It is the only different edge between G and G') Suppose that the path of the cycle is: $y \to x \to a_1 \to \cdots \to a_n \to y$, we know that a_n is a parent of y. Since $x \to y$ in G is a covered edge, a_n is also a parent of x, which means that there is a cycle in G, namely $x \to a_1 \to \cdots \to a_n \to x$. It contradicts our assumption that G is a DAG.

We want to show that for any distribution p, $\mathcal{L}(G) = \mathcal{L}(G')$. By using the hypothesis of covered edge.

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

What's more, we know that:

$$p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i) = p(x_i, x_j|x_{\pi_i})$$
$$p(x_i|x_{\pi_i}, x_j)p(x_j|x_{\pi_i}) = p(x_i, x_j|x_{\pi_i})$$

So we deduce that $\mathcal{L}(G) = \mathcal{L}(G')$.

We can also show that G and G' have the same v-structures. Assume that there is a v-structure in G' which does not appear in G, we know that the v-structure must contain $y \to x$. Suppose that the other edge in the structure is $z \to x$, we have in consequence z is a parent of x while z and y are not adjacent in G, which contradicts our assumption of a covered edge. Hence G and G' share the same v-structures.

According to the theorem [Verma and Pearl, 1990], two dags are equivalent if and only if they have the same skeletons and the same v-structures. We deduce that $\mathcal{L}(G) = \mathcal{L}(G')$.

(2) From the directed tree G we construct a symmetrized graph $\tilde{G} = (V, \tilde{E})$, which is equivalent to its moralized graph $\bar{G}$. This is true since G has not v-structures, then it has at maximum a parent, which is by definition a cliques, then no edge is added during the moralization process. Furthermore, this means that $\bar{G}$ is equivalent G'.

According to the Proposition 4.25, If G is a DAG without any v-structure $\mathcal{L}(G) = \mathcal{L}(\tilde{G}) = \mathcal{L}(\bar{G})$. So we deduce that $\mathcal{L}(G) = \mathcal{L}(G')$.

3 Implementation-Gaussian Mixtures

(1) The K-means is sensible to the initial values we use, as the solution given by the algorithm is a local minimum for the problem of minimizing the distortion. Nevertheless, in this data set, the difference between using different initializations was not significant. The results are expected to change too if we use a different distortion measure, as we are solving a different optimization problem. In this case we used the L_∞ and the L_1 norm, but the clusters didn't change too much in this data set. You can find more details about the experiments in the code. However, the selection is shown to have an impact in the computational time of the algorithm . Both metrics, L_∞ and the L_1, improved the computational time with respect to the euclidean metric.

(2) The deduction of the E-step for the isotropic Gaussian Mixture models is the same as the one for the general Gaussian Mixture models. After that, the M-step consists in minimizing with respect to $\theta_t = (\pi_t, \mu_t, {\sigma^2}_t)$ the expression:

$$Q(x;\theta_t) = \sum_{i=1}^{N}\sum_{k=1}^{K} \log(\pi_{k,t})\tau_{i,k} + \sum_{i=1}^{N}\sum_{k=1}^{K} \tau_{i,k}(-\frac{D}{2}\log 2\pi - \frac{D}{2}\log {\sigma^2}_{k,t} - \frac{1}{2{\sigma^2}_{k,t}}(x_i - \mu_{k,t})^T(x_i - \mu_{k,t}))$$

In the lectures we derived that $\hat{\pi_{k,t}} = \frac{\sum_{i=1}^{N}\tau_{i,k}}{N}$ and $\hat{\mu_{k,t}} = \frac{\sum_{i=1}^{N}\tau_{i,k}x_i}{\sum_{i=1}^{N}\tau_{i,k}}$

A problem arises when we want to maximize with respect to ${\sigma^2}_{k,t}$. Take for example the case when we have just one cluster with a data point, then $(x - \mu_t)^T(x - \mu_t)) = 0$. In that case the M step expression goes to infinity as ${\sigma^2}_{k,t}$ goes to zero, so the maximization problem is badly defined. This is not the only case, in general, the EM algorithm diverges when, during the interactions, one of the clusters shrinks to just a small number of observations. Otherwise, we can derive with respect to ${\sigma^2}_{k,t}$ and get:

$$\frac{\partial Q(x;\theta_t)}{\partial \sigma_{k,t}} = \sum_{i=1}^{N} \pi_{i,k}(-\frac{D}{\sigma_k} + \frac{\|x_i - \mu_k\|^2}{\sigma_k^3})$$

So we have:

$$\hat{\sigma^2}_k = \frac{1}{D}\frac{\sum_{i=1}^{N}\tau_{i,k}\|x_i - \hat{\mu_{k,t}}\|^2}{\sum_{i=1}^{N}\tau_{i,k}}$$

We met this problem during the implementation, Bishop 2006 suggests to restart the mean and the variance during the interactions to get a local minima with good proprieties. We tried to implement this solution, but it didn't work. We solved the problem by choosing a convenient initialization.

(3) The formula for the EM estimator of the covariance matrix for a general Gaussian Mixture is:

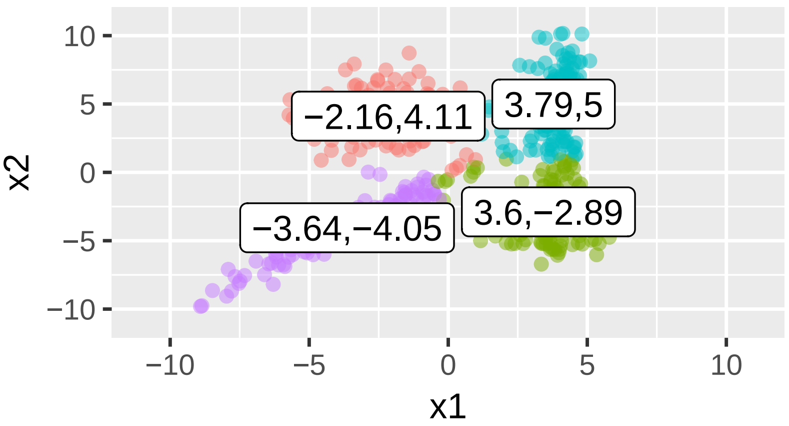
$$\hat{\sum_{j,t}} = \frac{\sum_{i=1}^{N}\tau_{(i,j)}(x_i - \mu_{j,t})(x_i - \mu_{j,t})^\intercal}{\sum_{i=1}^{N}\tau_{i,k}}$$

(4) The K-means algorithm finds a partition of the data in the mathematical sense that minimizes the distortion measure. For that reason, the clusters do not intersect between themselves and the regions defined are convex. In this data set, the initializations and the choice between the L_∞ and the euclidean norm slightly change in the centroids and clusters. We found the results of these methods are satisfying.

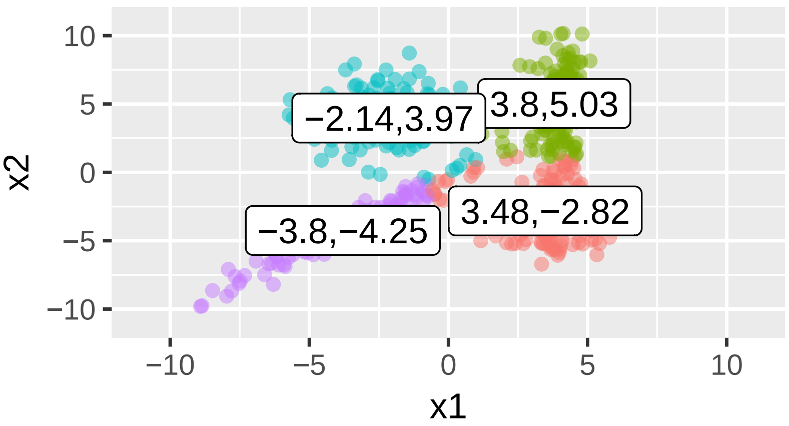
Both mixture models have the advantage compared to the K-means model that they don't assign just one label to each data point but a probability that it belongs to each cluster, for that reason they perform better when the clusters overlap, something happening frequently in the data set. The log likelihood table confirms what we deduce from the figure, the general Gaussian Mixture fits better the data set than the isotropic model. The reason is that the isotropic model has less parameters and forces the observations among clusters to be independent, something that geometrically seems not to be true. One problem, that could arise, given the flexibility of the general Gaussian Mixture is that it might overfit the data, but, given that the difference between the likelihood of the test and train set is not too big, we can conclude this problem doesn't happen. An additional detail to note is that, the log likelihood of both models decreases in the test set as expected.

In conclusion, we consider that the best model in this case was the general Mixture Gaussian model since its flexibility captures the features of the data without over-fitting. The local minima found for the isotropic Gaussian Mixture model doesn't show a better performance than the one of the K-means algorithm, we will rather prefer the K-means from both models.

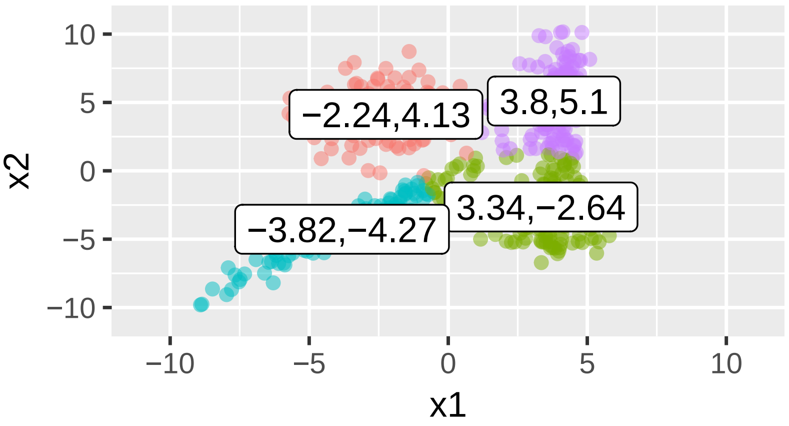
K−means with euclidean distance
and seed= 0



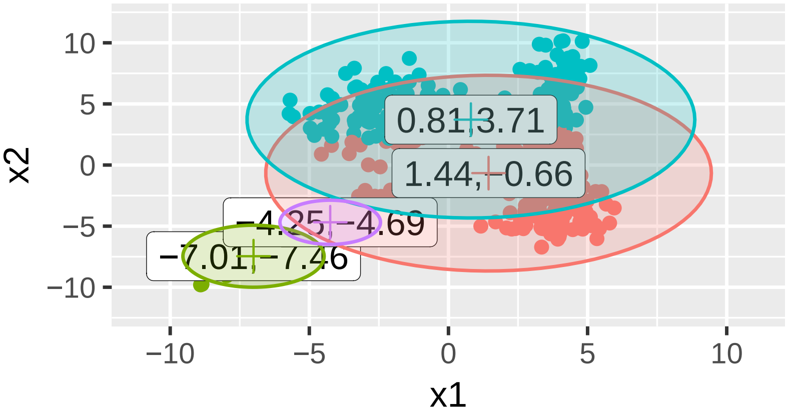
K−means with euclidean distance
and seed= 100



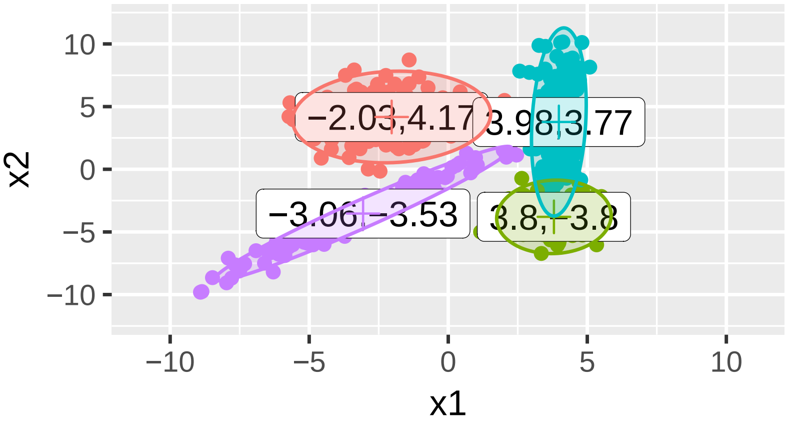
K−means with maximum distance
and seed= 0



EM algorithm for Mixture
of Gaussians with sigma=c*I



EM algorithm for
Mixture of Gaussians



	train	**test**
Mixture of Gaussians with sigma=c*I	−2721.61	−2764.14
Mixture of Gaussians	−2327.72	−2408.98