Modèles graphiques probabiliste - DM2

Olivier Jais-Nielsen & Jérémy Rapin

Mardi 17 novembre 2010

1 Distributions factorisables dans un graphe

1.1 Propositions

1.1.1 Inversion d'une arrête couverte

Soit $p \in \mathscr{L}(G)$. On a :

$$p(x_1,...,x_n) = \prod_{k=1,...,n} p\left(x_k | x_{\pi_k}\right)$$

Pour tout $k = 1,...,n$, soit π'_k l'ensemble des parents de k dans G. On a :

$$\begin{cases} \pi'_k = \pi_k & \forall k \neq i,j \\ \pi'_i = \pi_i \cup \{j\} & \\ \pi'_j = \pi_j - \{i\} = \pi_i & \end{cases}$$

On a donc :

$$p(x_1,...,x_n) = p(x_i|x_{\pi_i})\, p\left(x_j|x_{\pi_j}\right) \prod_{\substack{k=1,...,n \\ k\neq i,j}} p\left(x_k|x_{\pi'_k}\right)$$

et

$$\begin{aligned} p(x_i|x_{\pi_i})\, p\left(x_j|x_{\pi_j}\right) &= p(x_i|x_{\pi_i})\, p(x_j|x_{\pi_i},x_i) \\ &= \frac{p(x_i,x_{\pi_i})\, p(x_j,x_{\pi_i},x_i)}{p(x_{\pi_i})\, p(x_{\pi_i},x_i)} \\ &= \frac{p(x_j,x_{\pi_i},x_i)}{p(x_{\pi_i})} \\ &= \frac{p(x_{\pi_i},x_j)\, p(x_j,x_{\pi_i},x_i)}{p(x_{\pi_i})\, p(x_{\pi_i},x_j)} \\ &= p(x_j|x_{\pi_i})\, p(x_i|x_{\pi_i},x_j) \\ p(x_i|x_{\pi_i})\, p\left(x_j|x_{\pi_j}\right) &= p\left(x_j|x_{\pi'_j}\right) p\left(x_i|x_{\pi'_i}\right) \end{aligned}$$

On a donc bien :

$$p(x_1,...,x_n) = \prod_{k=1,...,n} p\left(x_k | x_{\pi'_k}\right)$$

Finalement, $p \in \mathscr{L}(G')$ et donc $\mathscr{L}(G) \subset \mathscr{L}(G')$.

On remarque enfin que les rôles de i et j étant symétriques, on montre de manière identique que $\mathscr{L}(G') \subset \mathscr{L}(G)$.

D'où $\mathscr{L}(G) = \mathscr{L}(G')$.

1.1.2 Symétrisation d'un arbre sans structure en V

Soit $p \in \mathscr{L}(G)$. On a :

$$p(x_1,...,x_n) = \prod_{k=1,...,n} p\left(x_k | x_{\pi_k}\right)$$

Comme il n'y a pas de structure en V, les ensembles de parents dans G sont soit vides (pour les indices dans $I \subset \{1,...,n\}$) soit des singletons (pour les indices dans $J = \{1,...,n\} - I$). Pour tout $j \in J$, on note $\pi_j = \{l(j)\}$. Alors :

$$p(x_1,...,x_n) = \prod_{i \in I} p(x_i) \prod_{j \in J} p\left(x_j | x_{l(j)}\right)$$

Ainsi, si $\mathscr{C}$ est l'ensemble des cliques de G', on a :

$$\mathscr{C} = \{\{1\},...,\{n\}\} \cup \bigcup_{j \in J} \{\{j, l(j)\}\}$$

On pose pour tout $c \in \mathscr{C}$:

$$\begin{cases} \Psi_c(x_c) = 1 \geq 0 & \text{si } c = \{j\} \text{avec } j \in J \\ \Psi_c(x_c) = p(x_i) \geq 0 & \text{si } c = \{i\} \text{avec } i \in I \\ \Psi_c(x_c) = p\left(x_j | x_{l(j)}\right) \geq 0 & \text{si } c = \{j, l(j)\} \text{avec } j \in J \end{cases}$$

On a alors :

$$\prod_{c \in \mathscr{C}} \Psi_c(x_c) = \prod_{i \in I} p(x_i) \prod_{j \in J} p\left(x_j | x_{l(j)}\right)$$
$$\prod_{c \in \mathscr{C}} \Psi_c(x_c) = p(x_1,...,x_n)$$

De plus, si $Z = \sum_{x_1,...,x_n} \prod_{c \in \mathscr{C}} \Psi_c(x_c)$, alors :

$$Z = \sum_{x_1,...,x_n} p(x_1,...,x_n) = 1$$

Donc $p(x_1,...,x_n) = \frac{1}{Z} \prod_{c \in \mathscr{C}} \Psi_c(x_c)$.

On a donc $p \in \mathscr{L}(G')$ et $\mathscr{L}(G) \subset \mathscr{L}(G')$.

Soit $p \in \mathscr{L}(G')$. On a :

$$p(x_1,...,x_n) = \frac{1}{Z}\prod_{c\in\mathscr{C}}\Psi_c(x_c)$$
$$= \frac{1}{Z}\prod_{i\in I}\Psi_i(x_i)\prod_{j\in J}\Psi_j(x_j)\prod_{j\in J}\Psi_{j,l(j)}\left(x_{j,l(j)}\right)$$

Soit, pour tout $k = 1,...,n$:

$$\begin{cases} f_k(x_k) = \frac{\Psi_k(x_k)}{A_k} \geq 0 & \text{si } k \in I \text{avec } A_k = \sum_{x_k}\Psi_k(x_k) \\ f_k\left(x_k, x_{l(k)}\right) = \frac{\Psi_k(x_k)\Psi_{k,l(k)}\left(x_{k,l(k)}\right)}{B_k} \geq 0 & \text{si } k \in J \text{avec } B_k = \sum_{x_k}\Psi_k(x_k)\Psi_{k,l(k)}\left(x_{k,l(k)}\right) \end{cases}$$

Soit :

$$q(x_1,...,x_n) = \prod_{i\in I} f_i(x_i)\prod_{j\in J} f_j\left(x_j, x_{l(j)}\right)$$

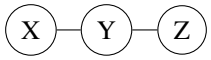
Comme pout tout $i \in I$, $\sum_{x_i} f_k(x_i) = 1$ et pour tout $j \in J$, $\sum_{x_j} f_j\left(x_j, x_{l(j)}\right) = 1$, alors q est une distribution de probabilité (qui se factorise sur le graphe G). Mais $q = \frac{Z}{\prod_{i\in I}A_i\prod_{j\in J}B_j}p$ donc $p \propto q$ et donc $p = q$ et enfin $p \in \mathscr{L}(G)$.

D'où $\mathscr{L}(G') \subset \mathscr{L}(G)$ et $\mathscr{L}(G) = \mathscr{L}(G')$.

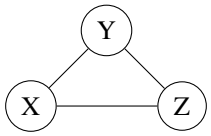
1.2 Plus petit graphe non-orienté sans équivalent DAG

Un graphe à un noeud est à la fois orienté et non orienté et un graphe à deux noeud non orienté est équivalent à tout graphe à deux noeuds dont il est le symétrisé.

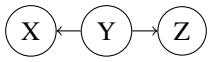
Soit un graphe à trois noeuds non orienté. Il est donc d'une des deux formes suivantes :



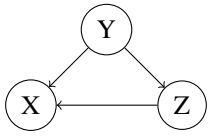
ou



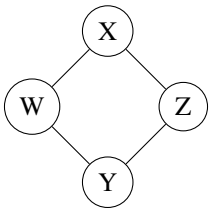
Le premier graphe factorise les distributions de X, Y et Z vérifiant $X \perp\!\!\!\perp Z|Y$. Il est donc équivalent au graphe orienté suivant :



Le second contient une seule clique maximale et décrit donc la totalité des distributions de X, Y et Z. Il est donc équivalent à, par exemple :



Le plus petit graphe non-orienté sans équivalent DAG a donc au moins 4 noeuds. On considère le graphe suivant :



Comme $\{W,Z\}$ sépare X et Y, les distributions de X, Y, Z et W factorisées par ce graphe vérifient :

$$X \perp\!\!\!\perp Y|W,Z$$

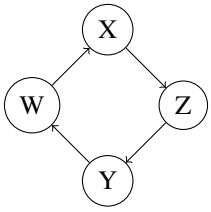
De même, elles vérifient aussi:

$$W \perp\!\!\!\perp Z|X,Y$$

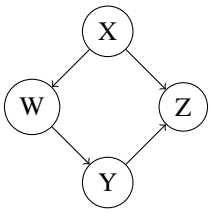
Si un DAG équivalent existe, il doit donc vérifier les propriétés suivantes :

- $\{X\}$ et $\{Y\}$ sont d-séparés par $\{W,Z\}$
- $\{W\}$ et $\{Z\}$ sont d-séparés par $\{X,Y\}$

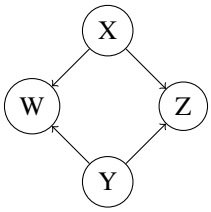
On suppose qu'il existe un tel DAG. Comme $\{W\}$ et $\{Z\}$ sont d-séparés par $\{X,Y\}$, il est d'une des 4 formes suivantes (à une permuttation $X \leftrightarrow Y$ et $W \leftrightarrow Z$ près):



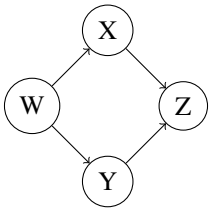
ou



ou



ou



Cependant, dans les 3 premiers cas, la chaîne $X \rightarrow Z \leftarrow Y$ présente une structure en V et n'est donc pas bloquée par $\{W,Z\}$. $\{X\}$ et $\{Y\}$ ne sont donc pas d-séparés par $\{W,Z\}$.

Dans le dernier cas, le graphe est cyclique et ce n'est donc pas un DAG.

Il n'existe donc pas de DAG équivalent au graphe non orienté considéré.

2 d-séparation

2.1 d-séparation et séparation usuelle

Si un DAG ne comporte pas de structure en V, alors une chaîne γ bloquée par C est caractérisée par :

$$\exists d \in \gamma, d \in C$$

Une chaîne est donc bloquée par C si et seulement si elle passe par C.

La notion de d-séparation dans ce DAG est alors équivalente à la séparation usuelle dans le graphe symmétrisé.

2.2 d-séparation d'un graphe moralisé

On suppose que A et B sont séparés par S dans G_M, graphe moralisé du DAG G. Soit γ une chaîne de G reliant A et B.

Si γ ne contient pas de structure en V, γ étant un chemin reliant A et B dans G_M, il existe donc $d \in \gamma$ tel que $d \in S$. Comme il n'y a pas de structure en V en d, la chaîne γ est donc bloquée dans G par S.

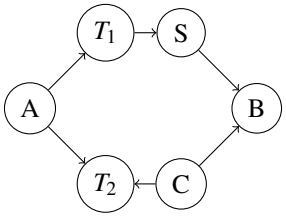
Si γ contient des structures en V, alors la moralisation crée un sous-chemin $\gamma' \subset \gamma$ (en reliant directement les parents des noeuds centraux des structures en V) reliant A et

B dans G_M. Il existe donc $d \in \gamma'$ (et donc $d \in \gamma$) tel que $d \in S$. Comme précedemment, la chaîne γ est donc bloquée dans G par S.

Donc toute chaîne reliant A et B dans G est bloquée par S. Ainsi, A et B sont d-séparés par S dans G.

2.3 Transitivité de la d-séparation

On considère le graphe suivant :



Soit γ une chaîne reliant A et B.

Si $\gamma = A \to T_1 \to S \to B$. Alors, en posant $d = S$, on a $d \in \{S\} \cap \gamma$ et il n'y a pas de structure en V en d. Donc γ est bloquée par S.

Sinon, $\gamma = A \to T_2 \leftarrow C \to B$. Alors, en posant $d = T_2$, on a $d \in \gamma - \{S\}$ et la chaîne forme une structure en V centrée sur d et d n'a pas de descendants. Donc γ est bloquée par S.

On en déduit que A et B sont d-séparés par S.

Soit γ une chaîne reliant A et S.

Si $\gamma = A \to T_1 \to S$. Alors, en posant $d = T_1$, on a $d \in T \cap \gamma$ (avec $T = \{T_1, T_2\}$) et il n'y a pas de structure en V en d. Donc γ est bloquée par T.

Sinon, $\gamma = A \to T_2 \leftarrow C \to B \leftarrow S$. Alors, en posant $d = B$, on a $d \in \gamma - T$ et la chaîne forme une structure en V centrée sur d et d n'a pas de descendants. Donc γ est bloquée par T.

On en déduit que A et S sont d-séparés par T.

Soit $\gamma = A \to T_2 \leftarrow C \to B$. γ est une chaîne reliant A et B. $\gamma \cap T = \{T_2\}$ et la chaîne forme une structure en V centrée sur T_2. De plus, l'unique autre élement intérieur de γ est C qui n'est pas centré en une structure en V. Donc γ n'est pas bloquée par T. On a donc que A et B ne sont pas d-séparés par T.

2.4 Exemples

$X_{\{1,2\}} \perp\!\!\!\perp X_4 | X_3$?

Les deux chaînes reliant $\{1,2\}$ et 4 sont $1 \to 8 \leftarrow 4$ et $2 \to 8 \leftarrow 4$. Elles contiennent chacune une structure en V dont le centre et ses descendant ne sont pas dans $\{3\}$. Elles sont donc bloquées par 3. On a donc bien $X_{\{1,2\}} \perp\!\!\!\perp X_4 | X_3$.

$X_{\{1,2\}} \perp\!\!\!\perp X_4 | X_5$?

Les deux mêmes chaînes précédentes ne sont pas bloquées car 5 fait partie des descendants de 8. On n'a donc pas $X_{\{1,2\}} \perp\!\!\!\perp X_4 | X_5$.

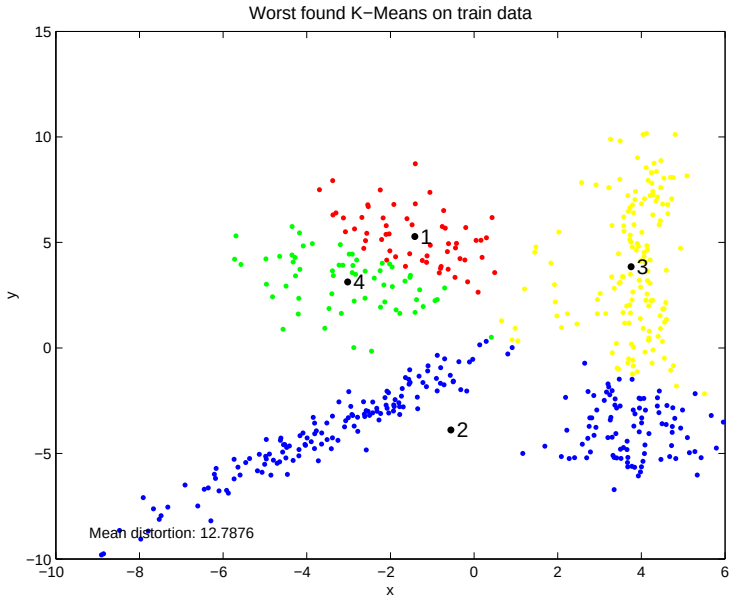
$X_1 \perp\!\!\!\perp X_6 | X_{\{2,4,7\}}$?

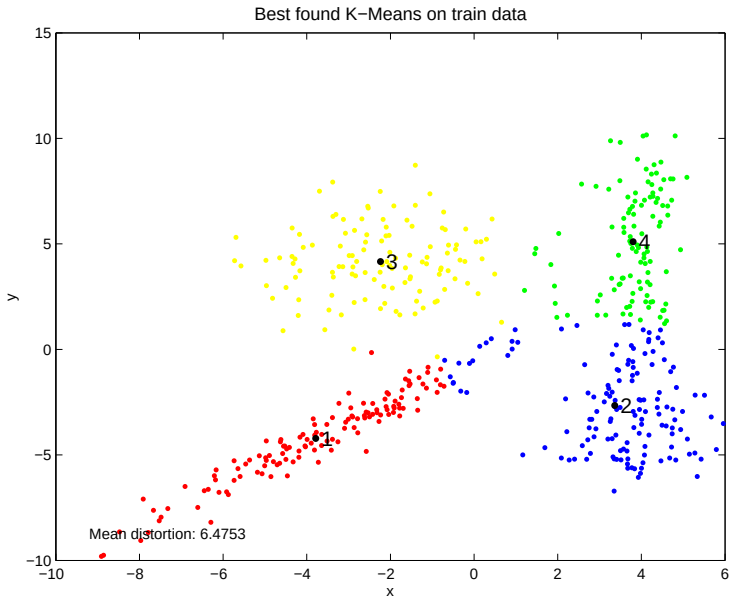
L'unique chaîne reliant 1 et 6 est $1 \to 8 \to 7 \to 6$ et contient donc $7 \in \{2,4,7\}$ donc est bloquée par $\{2,4,7\}$. On a donc bien $X_1 \perp\!\!\!\perp X_6 | X_{\{2,4,7\}}$.

3 Implémentation - Mélange de gaussiennes

Dans les différentes figures suivantes, chaque cluster est représenté pour une couleur. Les centres sont représentés par un point noir avec à sa droite le numéro de cluster associé.

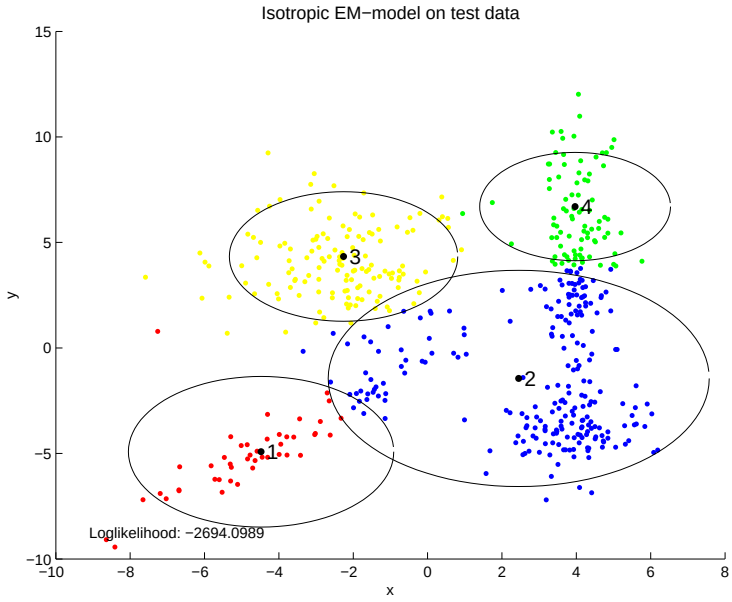
3.1 K-Means

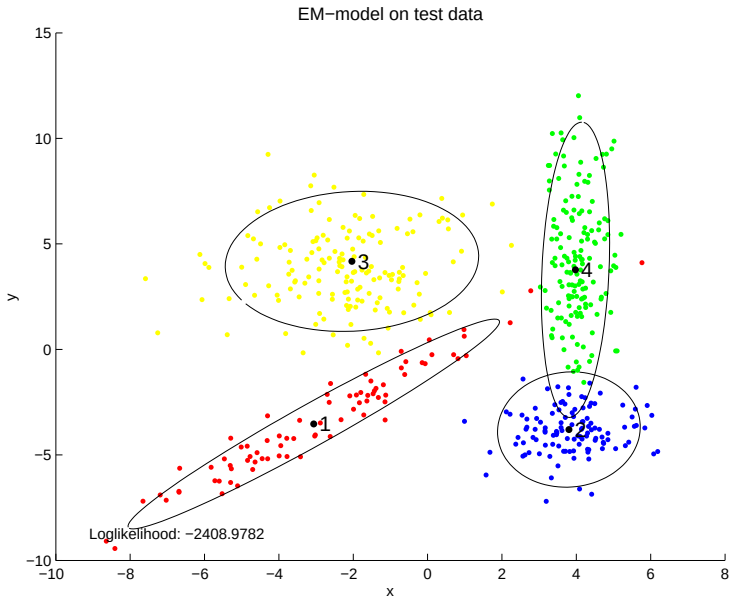

Worst found K−Means on train data
Mean distortion: 12.7876
1
2
3
4
x
y
15
10
5
0
−5
−10
−10
−8
−6
−4
−2
0
2
4
6



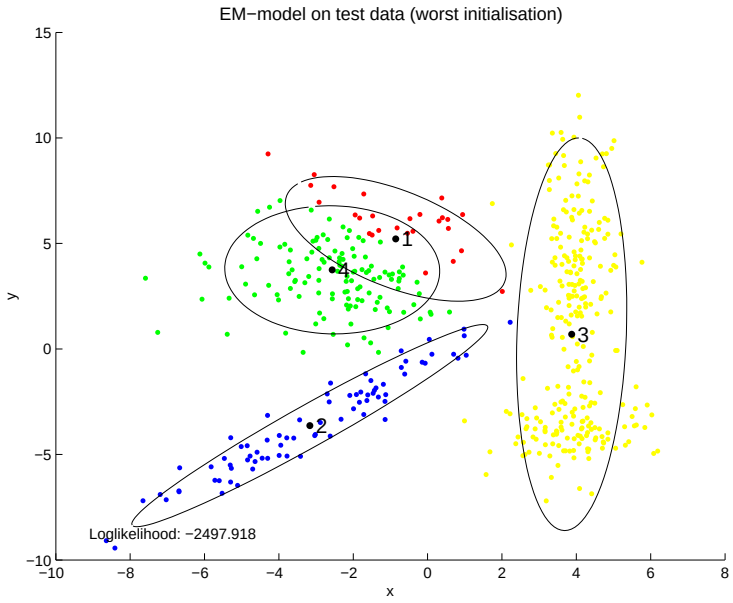
N'ayant pas d'information permettant d'initialiser au mieux le K-Means, on l'initialise avec des labels aléatoires. Cependant, l'algorithme peut converger vers des minimums locaux, comme on peut le voir sur les figures suivantes. Il convient donc de répéter plusieurs fois l'algorithme avec des initialisations différentes afin d'avoir plus de chances de trouver le minimum global.

3.2 E-M


Isotropic EM−model on test data
Loglikelihood: −2694.0989
x
y
1
2
3
4


EM−model on test data
Loglikelihood: −2408.9782
x
y
1
2
3
4

3.2.1 Importance de l'initialisation

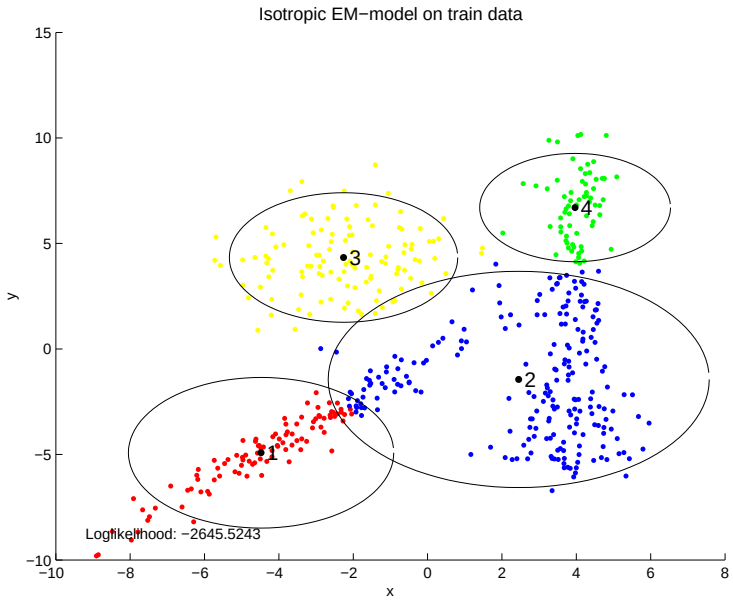


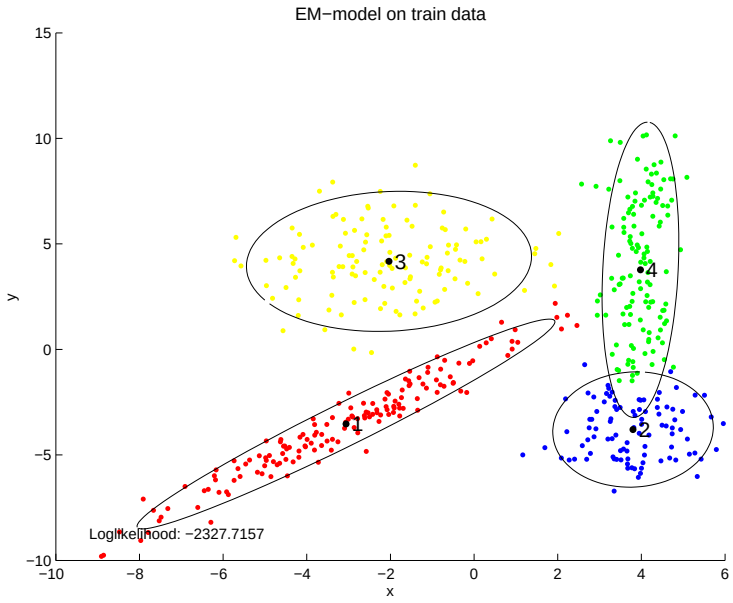
On peut observer une très grande différence entre l'algorithme EM avec une initialisation correcte, et le même algorithme avec une initialisation incorrecte (ici un minimum local du K-Means très éloigné du minimum global). On peut donc se rendre compte que l'initialisation pour l'EM est d'une grande importance.

3.2.2 Convergence

Suite à une erreur de codage au début, le critère d'arrêt était que la log-vraisemblance restait strictement égale entre 2 itérations. Ceci n'est pas sensé marcher pour l'algorithme EM, la convergence n'étant pas une convergence discrète. Cependant, cela marchait. En effet, nous avons remarqué que passé un certain point, la log-vraisemblance se mettait à osciller alors qu'elle n'est sensée être uniquement croissante. Nous avons d'abord pensé à une erreur, sans en trouver, puis nous nous sommes rendus compte que les valeurs des oscillations étaient discrétisées (k x 4.54 x 10-13, où k est un entier). La seule explication valable est qu'à ce moment nous atteignons les limites de la précision sous Matlab, les oscillations sont ainsi causées par les erreurs de calcul découlant du manque de précision, et le critère d'arrêt était toujours atteint tôt ou tard "par hasard" vu que l'on oscillait autour d'un même point, avec des valeurs discrètes. Avec une précision infinie, il n'y aurait pas eu d'oscillation, et le critère d'arrêt n'aurait jamais été atteint. Nous sommes ainsi revenu à un critère d'arrêt en précision par rapport à la log-vraisemblance, et n'avons plus observé d'oscillations.

3.2.3 Comparaison des algorithmes EM et EM isotropiques





L'algorithme EM est ici un mélange de gaussiennes, et on appellera EM isotropique le mélange de gaussiennes "isotropiques" dans le sens où les gaussiennes ne priviligient aucune direction (covariance proportionnelle à l'identité). On remarque très bien que l'algorithme EM normal semble plus précis, qu'il "colle" plus au données. Ainsi la log-vraisemblance de l'algorithme EM est plus élevée que celle de l'EM isotropique. Une autre façon de voire les choses est que l'on a plus de degrés de liberté dans l'algorithme EM, ce qui nous permet de se rapprocher d'autant plus de la distribution empirique et donc de faire augmenter la log-vraisemblance.

3.2.4 Comparaison des log-vraisemblance sur les données test et sur les données d'entraînement

Dans les deux cas on peut remarquer que la log-vraisemblance est plus faible pour les données test que pour les données d'entraînement. Il aurait été surprenant de trouver le contraire, les modèles étant spécifiquement conçus pour optimiser la log-vraisemblance des données d'entraînement.