Probabilistic Graphical Models - Homework Assignment 2

Chia-Man Hung

November 21, 2016

1 Entropy and Mutual Information

1.(a)

$$H(X) = -\sum_{x \in \chi} p(x) log p(x) \tag{1.1}$$

Since $p(x) \le 1$, $-p(x)logp(x) \ge 0$. This implies $H(X) \ge 0$. The equality holds if and only if $-p(x)logp(x) = 0, \forall x \in \chi$.
In conclusion, $H(X) \ge 0$ with equality only when X is constant.

1.(b)
Denote by p the distribution of X and q the uniform distribution on χ, i.e. $q(x) = \frac{1}{k}$, where $k = Card(\chi)$.

$$\begin{aligned} D(p||q) &= -\sum_{x \in \chi} p(x) log q(x) - (-\sum_{x \in \chi} p(x) log p(x) \\ &= -\sum_{x \in \chi} p(x) log q(x) - H(X) \\ &= log k - H(X) \end{aligned} \tag{1.2}$$

1.(c)
From the previous question, we have $H(x) = log k - D(p||q)$. We also know that $D(p||q) \ge 0$. Thus,

$$H(X) \le log k \tag{1.3}$$

2.(a)
The mutual information

$$\begin{aligned} I(X_1,X_2) &= \sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1,x_2) log \frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)} \\ &= D(p_{1,2}||p_1p_2) \geq 0 \end{aligned} \tag{1.4}$$

2.(b)

$$\begin{aligned} I(X_1,X_2) = & \sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1,x_2) log \frac{p_{1,2}(x_1,x_2)}{p_1(x_1)p_2(x_2)} \\ &= -\sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1,x_2) log p_1(x_1) - \sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1,x_2) log p_2(x_2) \\ &\quad -(-\sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1,x_2) log p_{1,2}(x_1,x_2)) \\ &= -\sum_{x_1\in\chi_1} p_1(x_1) log p_1(x_1) - \sum_{x_2\in\chi_2} p_2(x_2) log p_2(x_2) \\ &\quad -(-\sum_{(x_1,x_2)\in\chi_1\times\chi_2} p_{1,2}(x_1,x_2) log p_{1,2}(x_1,x_2)) \\ &= H(X_1) + H(X_2) - H(X_1,X_2) \end{aligned} \tag{1.5}$$

2.(c)
By combining the two previous answers, we obtain

$$H(X_1,X_2) \leq H(X_1) + H(X_2) \tag{1.6}$$

2 Conditional independence and factorizations

1.

$$\begin{aligned} X \perp\!\!\!\perp Y|Z &\Leftrightarrow p(x,y|z) = p(x|z)p(y|z) \quad \forall z, p(z) > 0 \\ &\Leftrightarrow \frac{p(x,y,z)}{p(z)} = \frac{p(x,z)}{p(z)}\frac{p(y,z)}{p(z)} \quad \forall z, p(z) > 0 \\ &\Leftrightarrow \frac{p(x,y,z)}{p(y,z)} = \frac{p(x,z)}{p(z)} \quad \forall (y,z), p(y,z) > 0 \\ &\Leftrightarrow p(x|y,z) = p(x|z) \quad \forall (y,z), p(y,z) > 0 \end{aligned} \tag{2.1}$$

2.

$$p(x,y,z,t) = p(x)p(y)p(z|x,y)p(t|z) \tag{2.2}$$

X and Y is not d-separated by T as the chain (X, Z, Y) is not blocked at Z. D-separation is a necessary and sufficient condition for the conditional dependency. We conclude that $X \perp\!\!\!\perp Y|T$ does not hold for any $p \in L(G)$.

3.(a)

If Z is a binary variable, then following statement holds.

If X ⫫ Y | Z and X ⫫ Y then (X ⫫ Z or Y ⫫ Z).

$$
\begin{aligned}
X \perp\!\!\!\perp Y | Z &\Rightarrow p(x,y|z) = p(x|z)p(y|z) \quad \forall z, p(z) > 0 \\
&\Rightarrow \frac{p(x,y,z)}{p(z)} = \frac{p(x,z)}{p(z)} \frac{p(y,z)}{p(z)} \quad \forall z, p(z) > 0 \\
&\Rightarrow p(x,y,z)p(z) = p(x,z)p(y,z) \\
&\Rightarrow p(z|x,y)p(x,y)p(z) = p(z|x)p(x)p(z|y)p(y) \\
&\Rightarrow p(z|x,y)p(z) = p(z|x)p(z|y) as X \perp\!\!\!\perp Y \\
&\Rightarrow (1-p(z|x,y))(1-p(z)) = (1-p(z|x))(1-p(z|y)) \\
&\Rightarrow p(z) + p(z|x,y) = p(z|x) + p(z|y) \\
&\Rightarrow (p(z) + p(z|x,y))^2 = (p(z|x) + p(z|y))^2 \\
&\Rightarrow p(z) - p(z|x,y) = p(z|x) - p(z|y) or\, p(z) - p(z|x,y) = -p(z|x) + p(z|y) \\
&\Rightarrow p(z) = p(z|x) or\, p(z) = p(z|y) \\
&\Rightarrow X \perp\!\!\!\perp Z or\, Y \perp\!\!\!\perp Z
\end{aligned} \tag{2.3}
$$

3.(b)
This statement is not true in general.
Counter-example:
$X \in \{0,1\}, Y \in \{0,1\}, Z \in \{0,1,2\}$ with the following joint probability.

$p(X=0, Y=0, Z=0) = \frac{1}{16}$
$p(X=0, Y=1, Z=0) = \frac{1}{16}$
$p(X=1, Y=0, Z=0) = \frac{3}{16}$
$p(X=1, Y=1, Z=0) = \frac{3}{16}$

$p(X=0, Y=0, Z=1) = \frac{3}{32}$
$p(X=0, Y=1, Z=1) = \frac{1}{32}$
$p(X=1, Y=0, Z=1) = \frac{3}{32}$
$p(X=1, Y=1, Z=1) = \frac{1}{32}$

$p(X=0, Y=0, Z=2) = \frac{1}{32}$
$p(X=0, Y=1, Z=2) = \frac{3}{32}$
$p(X=1, Y=0, Z=2) = \frac{1}{32}$
$p(X=1, Y=1, Z=2) = \frac{3}{32}$

Then,

$p(X=0,Y=0)=\frac{3}{16}$
$p(X=0,Y=1)=\frac{3}{16}$
$p(X=1,Y=0)=\frac{5}{16}$
$p(X=1,Y=1)=\frac{5}{16}$

$p(X=0,Z=0)=\frac{1}{8}$
$p(X=0,Z=1)=\frac{1}{8}$
$p(X=0,Z=2)=\frac{1}{8}$
$p(X=1,Z=0)=\frac{3}{8}$
$p(X=1,Z=1)=\frac{1}{8}$
$p(X=1,Z=2)=\frac{1}{8}$

$p(Y=0,Z=0)=\frac{1}{4}$
$p(Y=0,Z=1)=\frac{3}{16}$
$p(Y=0,Z=2)=\frac{1}{16}$
$p(Y=1,Z=0)=\frac{1}{4}$
$p(Y=1,Z=1)=\frac{1}{16}$
$p(Y=1,Z=2)=\frac{3}{16}$

We have $X \perp\!\!\!\perp Y | Z$ and $X \perp\!\!\!\perp Y$ while X is not independent from Z and Y is not independent from Z either.

3 Distributions factorizing in graph

1.
Let G = (V, E) be a DAG. $\pi_j = \pi_i \cup \{i\}$. Let G' = (V, E'), with $E' = (E \backslash \{i \to j\}) \cup \{j \to i\})$. Prove that L(G) = L(G').

Let $p(x) \in L(G)$.

$$\begin{aligned}
p(x) &= p(x_i|x_{\pi_i(G)})p(x_j|x_{\pi_j(G)}) \prod_{k \neq i,j} p(x_k|x_{\pi_k(G)}) \\
&= p(x_i|x_{\pi_i(G)})p(x_j|x_i, x_{\pi_i(G)}) \prod_{k \neq i,j} p(x_k|x_{\pi_k(G)}) \\
&= p(x_j, x_i|x_{\pi_i(G)}) \prod_{k \neq i,j} p(x_k|x_{\pi_k(G)}) \\
&= p(x_i|x_j, x_{\pi_i(G)})p(x_j|x_{\pi_i(G)}) \prod_{k \neq i,j} p(x_k|x_{\pi_k(G)}) \\
&= p(x_i|x_{\pi_i(G')})p(x_j|x_{\pi_j(G')}) \prod_{k \neq i,j} p(x_k|x_{\pi_k(G')})
\end{aligned} \tag{3.1}$$

We obtain $p(x) \in L(G')$, i.e. $L(G) \subset L(G')$.

By symmetry, we also have $L(G') \subset L(G)$. Thus, $L(G') = L(G)$.

2.
Let G b a directed tree and G' its corresponding undirected tree. By the definition of a directed tree, G does not contain any v-structures, yielding

$$\forall i \in V, |\pi_i| \in \{0, 1\} \tag{3.2}$$

Given $p(x) \in L(G)$,

$$p(x) = \prod_{i \in V} p(x_i | x_{\pi_i}) \tag{3.3}$$

Thus, G does not contain any clique of size greater than 2. Therefore, p can be factorized in G', i.e. $p(x) \in L(G')$. This leads to $L(G) \subset L(G')$.

Let's prove the opposite direction.
Assume that $p(x) \in L(G')$. The factorization of an undirected graph is written as

$$p(x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \tag{3.4}$$

where

$$Z = \sum_x \prod_{c \in C} \psi_c(x_c) \tag{3.5}$$

G' is a tree, so it does not contain any clique of size greater than 2.

We can restrict cliques to be maximal cliques.

$$\begin{aligned} p(x) &= \frac{1}{Z} \prod_{c \in C_{max}} \psi_c(x_c) \\ &= \frac{1}{Z} \prod_{i \in V} \psi_i(x_i, x_{\pi_i}) \end{aligned} \tag{3.6}$$

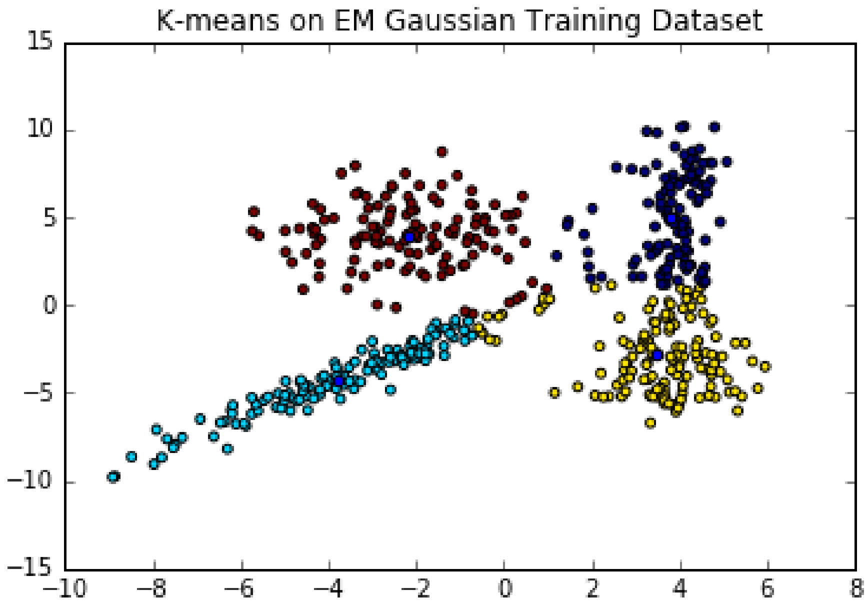
In fact, Z is a normalization term. We can split it into a product of several terms such that $\forall i \in V, \sum_{x_i} \psi_i(x_i, x_{\pi_i}) = 1$. The definition 4.1 in Lecture 4 is then verified. We have $p(x) \in L(G)$, i.e. $L(G') \subset L(G)$.

In conclusion, $L(G) = L(G')$.

4 Implementation - Gaussian mixtures

(a)

Figure 4.1: K-means algorithm. After 10 iterations. The blue points are the centers.



(b) Special case

In the case where the covariance matrices are proportional to the identity matrix, the updates of μ and π do not change. For the update of the covariance matrices, we need to maximize

$$f(\mu,\sigma^2) = -\sum_{i=1}^{n}\sum_{j=1}^{k}\tau_i^j(\frac{d}{2}log(2\pi) + \frac{d}{2}log\sigma^2 + \frac{1}{2\sigma^2}(x_i-\mu_j)^T(x_i-\mu_j))) \tag{4.1}$$

with respect to σ^2.

$$\nabla_{(\sigma^2)_i} f(\mu,\sigma^2) = -\sum_{i=1}^{n}\tau_i^j(\frac{d}{2(\sigma^2)_i} - \frac{1}{2((\sigma^2)_i)^2}(x_i-\mu_j)^T(x_i-\mu_j)) = 0, i = 1,...,K. \tag{4.2}$$

$$(\sigma^2)_i = \frac{\sum_{i=1}^{n}\tau_i^j(x_i-\mu_j)^T(x_i-\mu_j)}{d\sum_{i=1}^{n}\tau_i^j} \tag{4.3}$$

$$\Sigma_i = (\sigma^2)_i I_d \tag{4.4}$$

Figure 4.2: Special - Contours of the Gaussians at different iterations: 0, 1, 2.

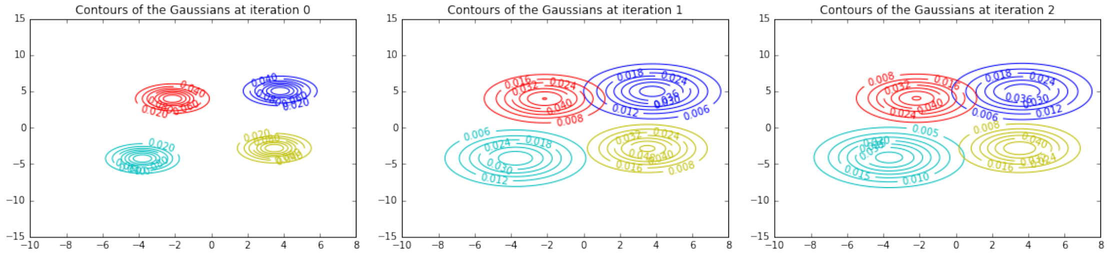


Figure 4.3: Special - Contours of the Gaussians at different iterations: 3, 5, 10.

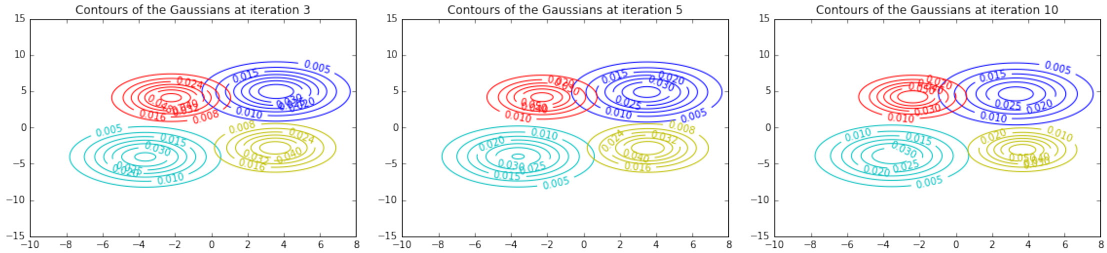


Figure 4.4: Special - Contours of the Gaussians at different iterations: 20, 30.

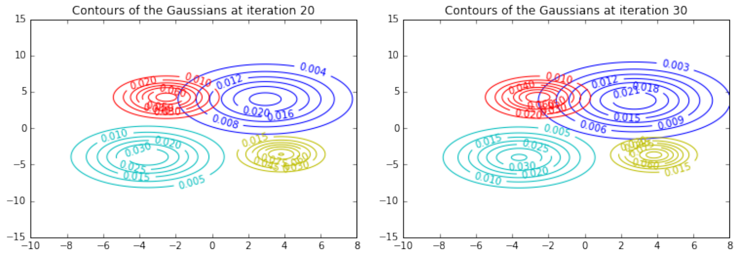
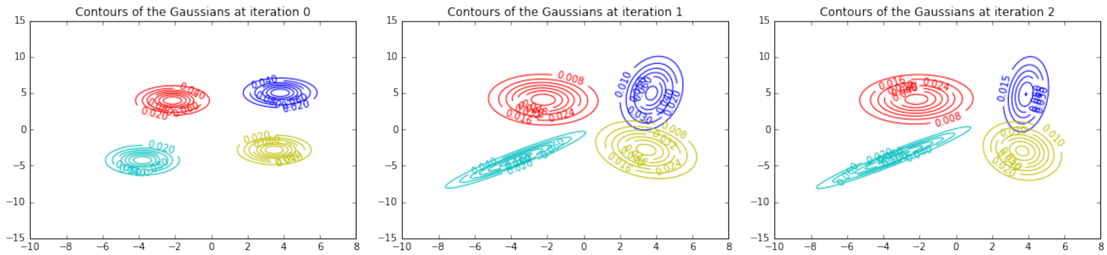


Figure 4.5: General - Contours of the Gaussians at different iterations: 0, 1, 2.



(c) General case

(d) Comparison and log-likelihood

Figure 4.6: General - Contours of the Gaussians at different iterations: 3, 5, 10.

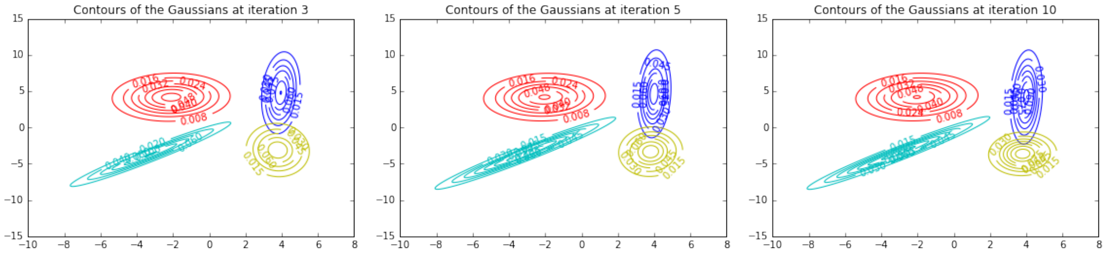


Figure 4.7: Special case: Log-likelihood. Left: training dataset / Right: test dataset.

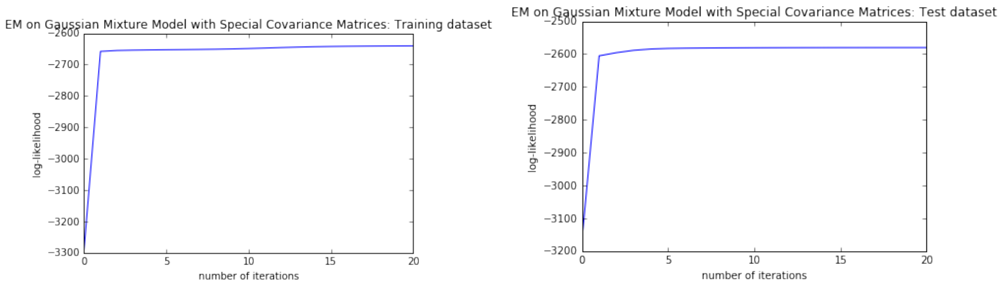
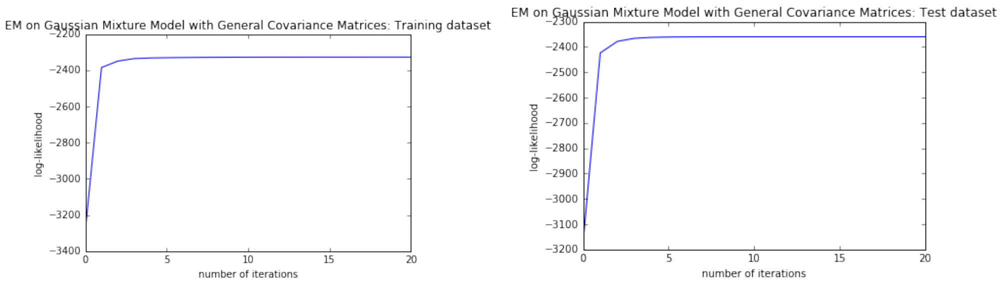


Figure 4.8: General case: Log-likelihood. Left: training dataset / Right: test dataset.



As expected, the log-likelihood increases with respect to the number of iterations. It converges to a larger value in the general case than in the special case where the covariance matrices are proportional to the identity matrix.