Exercise 1.1

For any $p \in \mathcal{L}(G)$, $p(x,y,z,t) = p(t|z)p(z|x,y)p(x)p(y)$.
$X \perp Y \mid T$ does not necessarily hold for any $p \in \mathcal{L}(G)$. Indeed if X and Y are i.i.d Rademacher variables, $Z := \mathbb{1}_{X=Y}$ and $T := Z$, then $P(X=1, Y=-1|T=1) = P(X=1, Y=-1|X=Y) = 0$, whereas $P(X=1|T=1)P(Y=-1|T=1) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$, hence $X \not\perp Y \mid T$.

Exercise 1.2

a.

Let us prove that the claim is true. Let $x \in \operatorname{Supp} X$ and $y \in \operatorname{Supp} Y$. Note that

$$\begin{aligned}
P_X(x)P_Y(y) &\overset{(1)}{=} P_{(X,Y)}(x,y) \\
&\overset{(2)}{=} P_{(X,Y)}(x,y|Z=1)P_Z(1) + P_{(X,Y)}(x,y|Z=0)P_Z(0) \\
&\overset{(3)}{=} P_X(x|Z=1)P_Y(y|Z=1)P_Z(1) + P_X(x|Z=0)P_Y(y|Z=0)P_Z(0) \\
&\overset{(4)}{=} \frac{P_Z(1|X=x)P_X(x)}{P_Z(1)}\frac{P_Z(1|Y=y)P_Y(y)}{P_Z(1)}P_Z(1) + \frac{P_Z(0|X=x)P_X(x)}{P_Z(0)}\frac{P_Z(0|Y=y)P_Y(y)}{P_Z(0)}P_Z(0) \\
&= P_X(x)P_Y(y)\left(\frac{P_Z(1|X=x)P_Z(1|Y=y)}{P_Z(1)} + \frac{(1-P_Z(1|X=x))(1-P_Z(1|Y=y))}{1-P_Z(1)}\right)
\end{aligned}$$

(1): X and Y are independent
(2): Law of total probability
(3): X and Y are conditionally independent given Z
(4): Bayes' theorem

Since x and y are in the supports of X and Y, we may simplify by $P_X(x)P_Y(y)$ and get

$$1 = \frac{P_Z(1|X=x)P_Z(1|Y=y)}{P_Z(1)} + \frac{(1-P_Z(1|X=x))(1-P_Z(1|Y=y))}{1-P_Z(1)}$$

which, after a bit of algebra, rewrites as $0 = (P_Z(1) - P_Z(1|X=x))(P_Z(1) - P_Z(1|Y=y)) \quad (*)$

If $\forall x \in \operatorname{Supp} X,\ P_Z(1) - P_Z(1|X=x) = 0$ then Z and X are independent and we are done. Otherwise, there exists some $x_0 \in \operatorname{Supp} X$ such that $P_Z(1) - P_Z(1|X=x_0) \neq 0$. Setting $x = x_0$ in $(*)$ yields $P_Z(1) - P_Z(1|Y=y)$ for all $y \in \operatorname{Supp} Y$, hence Y and Z are independent.

b.

Without the assumption that Z is binary, the claim does not hold. One can simply consider $Z = (X,Y)$. Let us prove first that $X \perp Y \mid (X,Y)$. Let $x_1, x_2 \in \operatorname{Supp} X$ and $y_1, y_2 \in \operatorname{Supp} Y$. Then

$$P_{(X,Y)}(x_1, y_1|(X,Y) = (x_2,y_2)) = \mathbb{1}_{x_1=x_2}\mathbb{1}_{y_1=y_2}$$

and

$$P_X(x_1|(X,Y)=(x_2,y_2)) \cdot P_Y(y_1|(X,Y)=(x_2,y_2)) = \frac{\mathbb{1}_{x_1=x_2}P_Y(y_2)}{P_Y(y_2)}\frac{\mathbb{1}_{y_1=y_2}P_Y(y_2)}{P_Y(y_2)} = \mathbb{1}_{x_1=x_2}\mathbb{1}_{y_1=y_2}$$

Furthermore, it is clear that $X \not\perp (X,Y)$ and $Y \not\perp (X,Y)$

Exercise 2.1

Let $n = |V|$, π_k be the set of parents of vertex k in G and σ_k be the set of parents of k in G'. Consider $p \in \mathcal{L}(G)$. Then

$$\forall x, p(x) = \prod_{k=1}^{n} p(x_k|x_{\pi_k}) = p(x_i|x_{\pi_i})p(x_j|x_{\pi_j}) \prod_{k \notin \{i,j\}} p(x_k|x_{\pi_k}) \overset{(1)}{=} p(x_i|x_{\pi_i})p(x_j|x_{\pi_i}, x_i) \prod_{k \notin \{i,j\}} p(x_k|x_{\pi_k})$$
$$\overset{(2)}{=} p(x_j|x_{\pi_i})p(x_i|x_{\pi_i}, x_j) \prod_{k \notin \{i,j\}} p(x_k|x_{\pi_k}) \overset{(3)}{=} p(x_j|x_{\sigma_j})p(x_i|x_{\sigma_i}) \prod_{k \notin \{i,j\}} p(x_k|x_{\sigma_k})$$

(1) comes from the assumption $\pi_j = \pi_i \cup \{i\}$. (2) stems from the following algebra

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

(3) follows from the edge covering assumption: since $\pi_j = \pi_i \cup \{i\}$ and i is a no longer a parent of j in G', it must be that $\sigma_j = \pi_i$ and since j is a parent of i in G', $\sigma_i = \pi_i \cup \{j\}$.
Hence $\forall x, p(x) = \prod_{k=1}^{n} p(x_k|x_{\sigma_k})$, thus $p \in \mathcal{L}(G')$ and $\mathcal{L}(G) \subset \mathcal{L}(G')$. The reverse inclusion is obtained directly by switching G with G' and i with j.

Exercise 2.2

Let us prove first that $\mathcal{L}(G) \subset \mathcal{L}(G')$. Since G is a directed tree, each vertex (except the root) has exactly one parent. Besides, cliques in G' have at most 2 elements: indeed if there were a clique with more than 3 elements, there would either be a cycle or a v-structure in G. Cliques in G' are therefore of the form {root}, {vertex other than root} and {parent, child}.
Consider $p \in \mathcal{L}(G)$: $p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i})$ and suppose WLOG that the root is the vertex with index 1. Then $p(x) = p(x_1|x_\emptyset) \prod_{i=2}^{n} p(x_i|x_{\pi_i}) = \psi_1(x_1) \prod_{i=2}^{n} \psi_i(x_i) \prod_{i=2}^{n} \psi_i(x_i, x_{\pi_i})$ where

$$\begin{aligned} \psi_1(x_1) &:= p(x_1|x_\emptyset) \\ \forall i \geq 2, \psi_i(x_i) &:= 1 \\ \psi_i(x_i, x_{\pi_i}) &:= p(x_i|x_{\pi_i}) \end{aligned}$$

Thus p factorises in G' and $p \in \mathcal{L}(G')$.

The reverse inclusion is harder to prove.

3.a

We run many random initializations around the mean of the training dataset. The results depicted on the graph below support to some extent the bypothesis that the K-Means algorithm often converges to local minima.

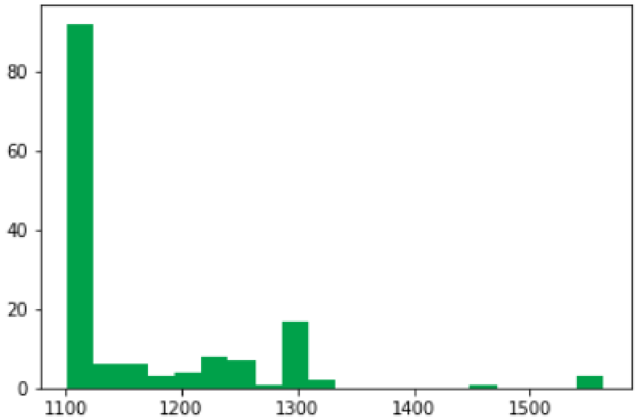


Figure 1: Distorsions distribution after 1500 random initializations

3.b

Since $\Sigma_j = \sigma_j^2 I_d$, the function to be maximized at step t is

$$(\pi, \mu, \sigma^2) \mapsto \sum_{i=1}^{n}\sum_{j=1}^{K} \tau_i^j \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{K} \tau_i^j \left(\log(\frac{1}{\sigma_j^d}) - \frac{1}{2}\frac{\|x_i - \mu_j\|^2}{\sigma_j^2} \right)$$

The function is separable in π and (μ, σ^2), so we may maximize separately the first and the second summand. Maximizing with respect to π has been done in the lecture notes: $\boxed{\forall j, \pi_{j,t+1} = \frac{1}{n}\sum_{i=1}^{n} \tau_i^j}$. It remains to minimize $(\mu, \sigma^2) \mapsto \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{K} \tau_i^j \left(d\log(\sigma_j) + \frac{1}{2}\frac{\|x_i - \mu_j\|^2}{\sigma_j^2} \right)$. For a fixed σ^2, this function is the separable sum of convex functions of the μ_j. Equating the gradient with respect to μ_j to 0 yields $\frac{1}{\sigma_j^2}\sum_{i=1}^{n} -2 \cdot \frac{\tau_i^j}{2}(x_i - \mu_j) = 0$, hence $\boxed{\forall j, \mu_{j,t+1} = \frac{\sum_{i=1}^{n} \tau_i^j x_i}{\sum_{i=1}^{n} \tau_i^j}}$. These optimal values do not depend on σ, so it remains to minimize $\sigma \mapsto \sum_{i=1}^{n}\sum_{j=1}^{K} \tau_i^j \left(d\log(\sigma_j) + \frac{1}{2}\frac{\|x_i - \mu_{j,t+1}\|^2}{\sigma_j^2} \right)$. This is the separable sum of functions of the σ_j^2, hence each summand may be minimized separately. Computing derivatives and studying their signs shows that $\boxed{\forall j, \sigma_j^2 = \frac{1}{d}\frac{\sum_{i=1}^{n} \tau_i^j \|x_i - \mu_{j,t+1}\|^2}{\sum_{i=1}^{n} \tau_i^j}}$.

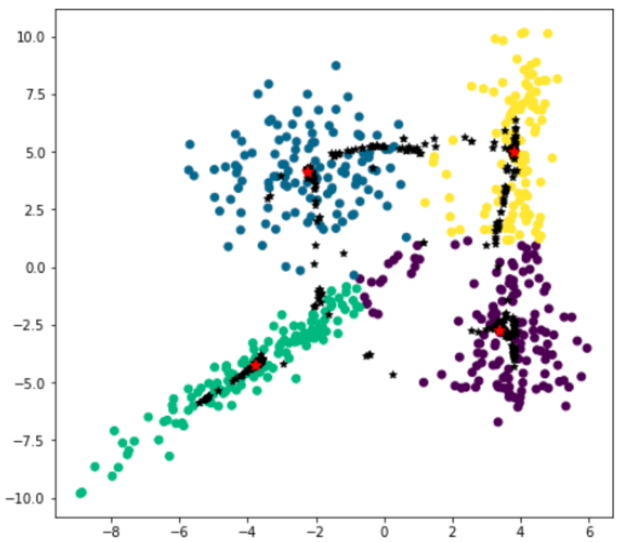
3.c

In the general case with no assumption on Σ, it can be proved that

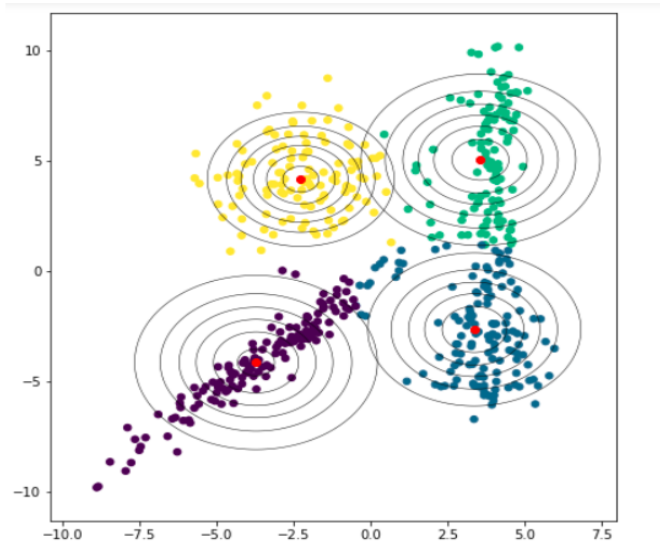
$$\Sigma_{j,t+1} = \frac{\sum_{i=1}^{n} \tau_i^j (x_i - \mu_{j,t+1})(x_i - \mu_{j,t+1})^T}{\sum_{i=1}^{n} \tau_i^j}$$

3.d

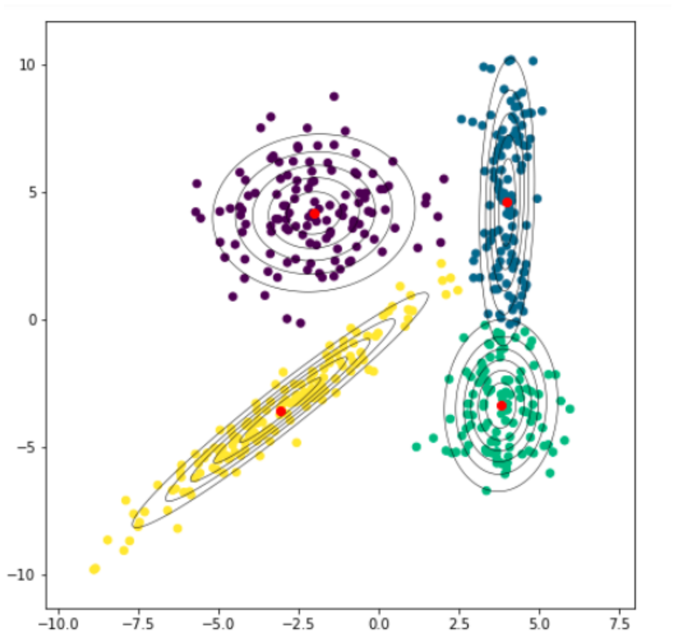
For the isotropic model, the likelihood on the training set and test set are respectively -2653.634 and -2651.575. Meanwhile, for the general model, the likelihood on the training set and test set are respectively -2332.262 and -2421.27. Thus estimates of the general model are more accurate than the restricted isotropic model. Moreover likekihoods on both training and test dataset are close.



(a) K-Means



(b) EM Isotropic



(c) EM General