Aix*Marseille université

Année 2024-25

Master 2 informatique

Parcours IAAA, UE MIA

Feuille d'exercices numéro 2

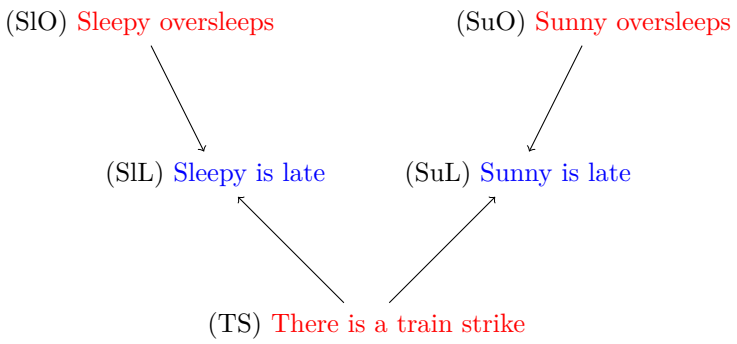
Statistique et probabilités.

1 Calcul probabiliste

Exercice 1 Propriétés élémentaires. ($\star$)

1. Prouvez que si A et B sont des événements indépendants, $P(A|B) = P(A)$.
2. On considère une fonction f de $\mathbf{R}$ vers $\mathbf{R}$. Quelles conditions doit-elle vérifier pour être une fonction de répartition valide ? Une densité de probabilité valide ?
3. Représentez graphiquement la fonction de masse, la densité de probabilité et la fonction de répartition pour une variable aléatoire X uniforme sur $[0, 1]$.
4. Même question pour X valant 1 avec probabilité .5, 2 avec probabilité .3 et 4 avec probabilité .2.
5. Un coffre A contient 100 pièces d'or. Un coffre B contient 60 pièces d'or et 40 pièces d'argent. Vous choisissez un coffre aléatoirement selon une loi uniforme et tirez une pièce aléatoirement selon une loi uniforme dans ce coffre. Si la pièce est en or, quelle est la probabilité que vous ayez choisi le coffre A ?

Exercice 2 Calcul probabiliste. ($\star\star$)



Considérons le modèle graphique représenté ci-dessus. SlO, SuO, SlL, SuL and TS sont des variables aléatoire binaires prenant leurs valeurs dans $\{0, 1\}$. Dans cet exercice nous allons essayer de déterminer ce qui peut être inféré sur les variables latentes (en rouge) à partir de l'observation des variables observées (en bleu).

Nous faisons l'hypothèse que si au moins un des deux évènements "Sleepy oversleeps" et "There is a train strike" a lieu, alors 'Sleepy is late' a lieu également (avec probabilité 1). De façon similaire, si au moins un des deux évènements 'Sunny oversleeps' et 'There is a train strike' a lieu, alors 'Sunny is late' a lieu également (avec probabilité 1). Nous pouvons l'écrire plus formellement, de la manière suivante :

$$P(SlL = 1|SlO = a, TS = b) = a \vee b,$$

et :

$$P(SuL = 1|SuO = a, TS = b) = a \vee b,$$

pour tout a, b dans $\{0, 1\}$. Le symbole $\vee$ représente le connecteur logique *ou* (inclusif) de $\{0, 1\}$ vers $\{0, 1\}$.

On note $l = P(SlO = 1)$, $u = P(SuO = 1)$ et $t = P(TS = 1)$.

1. Donner la factorisation de $P(SlL, SuL, SlO, SuO, TS)$ d'après le modèle graphique représenté ci-dessus.

   Solution : D'après le modèle graphique representé ci-dessus :

   $$P(SlL, SuL, SlO, SuO, TS) = P(SlL \mid SlO, TS)P(SuL \mid SuO, TS)P(SlO)P(SuO)P(TS).$$

2. La distribution de probabilité $P(SlL, SuL, SlO, SuO, TS)$ est-elle entièrement déterminée si les valeurs de l, u et t sont données ?

   Solution : SlO, SuO et TS étant des variables binaires, leur distribution est entièrement déterminée par la donnée de, respectivement, l, u et t. Comme SlL et SuL sont des variables binaires $P(SlL \mid SlO, TS)$ et $P(SuL \mid SuO, TS)$ sont entièrement déterminées respectivement par les formules $P(SlL = 1 \mid SlO = a, TS = b) = a \vee b$ et $P(SuL = 1 \mid SuO = a, TS = b) = a \vee b$.

   En appliquant la formule obtenue à la question précédente pour la distribution de probabilité $P(SlL, SuL, SlO, SuO, TS)$, on conclut donc que cette distribution est bien entièrement déterminée par la donnée de l, u et t.

3. Calculer $P(TS = 1|SlL = 1)$ en fonction de l, u et t.

Solution : On a :

$$P(TS=1 \mid SlL=1) = \frac{P(TS=1, SlL=1)}{P(SlL=1)} = \frac{\sum_{SuL,SuO,SlO} P(SlL=1, SuL, SlO, SuO, TS=1)}{\sum_{SuL,SuO,SlO,TS} P(SlL=1, SuL, SlO, SuO, TS)}.$$

Or :

$$\sum_{SuL,SuO,SlO} P(SlL=1, SuL, SlO, SuO, TS=1) = tAB,$$

avec :

$$\begin{aligned} A &:= \sum_{SuL,SuO} P(SuL \mid SuO, TS=1) P(SuO) \\ &= 0 + \sum_{SuO} P(SuO) = 1 \end{aligned}$$

et :

$$B := \sum_{SlO} P(SlL=1 \mid SlO, TS=1) P(SlO) = \sum_{SlO} P(SlO) = 1.$$

Donc : $\sum_{SuL,SuO,SlO} P(SlL=1, SuL, SlO, SuO, TS=1) = t.$

Solution : (continuée)

D'où :

$$\sum_{SuL,SuO,SlO,TS} P(SlL=1,SuL,SlO,SuO,TS) = \sum_{SuL,SuO,SlO} P(SlL=1,SuL,SlO,SuO,TS=0)+t.$$

Or :

$$\sum_{SuL,SuO,SlO} P(SlL=1,SuL,SlO,SuO,TS=0) = (1-t)CD,$$

avec :

$$C := \sum_{SuL,SuO} P(SuL \mid SuO, TS=0)P(SuO)$$

$$= P(SuL=0 \mid SuO=0,TS=0)P(SuO=0) + P(SuL=1 \mid SuO=1,TS=0)P(SuO=1) + 0 + 0$$

$$= 1-u+u = 1$$

et :

$$D := \sum_{SlO} P(SlL=1 \mid SlO, TS=0)P(SlO)$$

$$= 0 + p(SlO=1) = l.$$

Donc :

$$\sum_{SuL,SuO,SlO} P(SlL=1,SuL,SlO,SuO,TS=0) = (1-t)l.$$

Et finalement :

$$P(TS=1 \mid SlL=1) = \frac{t}{t+l-tl}.$$

4. Calculer $P(SlO=1|SlL=1)$ en fonction de l, u et t.

Solution : Similar computations lead to :

$$P(SlO=1 \mid SlL=1) = \frac{l}{t+l-tl}.$$

5. Calculer $P(TS=1|SlL=1,SuL=1)$ en fonction de l, u et t.

Solution : Similar computations lead to :

$$P(TS=1 \mid SlL=1,SuL=1) = \frac{t}{t+lu-ltu}.$$

6. Calculer $P(SlO=1|SlL=1,SuL=1)$ en fonction de l, u et t.

Solution : Similar computations lead to :

$$P(SlO = 1 \mid SlL = 1, SuL = 1) = \frac{l(t + u - tu)}{t + lu - ltu}.$$

7. Supposer à présent que $l = 0.5$, $t = 0.1$ et que l'évenement 'Sleepy is late' a lieu. Quel évènement est alors le plus probable : 'There is a train strike' ou 'Sleepy overslept' ?

Solution :

$$P(TS = 1 \mid SlL = 1) = \frac{t}{t + l - tl} = 0.1/(0.1 + 0.5 - 0.05) = 2/11 \approx .18$$

$$P(SlO = 1 \mid SlL = 1) = \frac{l}{t + l - tl} = 0.5/(0.1 + 0.5 - 0.05) = 10/11 \approx .91$$

L'évenement le plus probable est 'Sleepy overslept'.

8. Même question si on suppose en plus que $u = 0.01$ et que l'évènement 'Sunny is late' est également observé.

Solution :

$$P(TS = 1 \mid SlL = 1, SuL = 1) = \frac{t}{t + lu - ltu} = 200/209 \approx .96$$

$$P(SlO = 1 \mid SlL = 1, SuL = 1) = \frac{l(t + u - tu)}{t + lu - ltu} = 109/209 \approx .52$$

L'évenement le plus probable est à présent 'There is a train strike'.

9. Que se passe-t-il si on prend $l = 0.5$, $t = 0.1$ et $u = 0.2$?

Solution : On a :

$$P(TS = 1 \mid SlL = 1, SuL = 1) = \frac{t}{t + lu - ltu} = 10/19 \approx .53$$

$$P(SlO = 1 \mid SlL = 1, SuL = 1) = \frac{l(t + u - tu)}{t + lu - ltu} = 14/19 \approx .74$$

Si 'Sunny oversleeps' est suffisamment probable relativement à 'There is a train strike', l'observation qu'à la fois Sunny et Sleepy sont en retard ne rend pas la probabilité qu'il y ait eu un grève de train plus élevée que la probabilité que Sleepy ne se soit pas réveillé à l'heure.

2 Calcul de moments

Exercice 3 Preuve de propriétés vues en cours. ($\star\star$)

1. Montrez que $\mathrm{Var}(X) = E[X^2] - E[X]^2$
2. Prouvez la loi de l'espérance totale.
3. Prouvez la loi de la variance totale.

Exercice 4 Espérance et variance d'estimateurs classiques. ($\star\star$)

Soit n un entier naturel et $X_1, X_2, \ldots, X_n$ des variables aléatoires réelles mutuellement indépendantes. Par souci de simplicité, on suppose que tous les moments des ces variables aléatoires existent. On note

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

et

$$\hat{V} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

1. Montrez que
$$\hat{V} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \hat{\mu}^2$$

Solution : On développe les carrés et on regroupe les termes en trois sommes, puis on factorise et on reconnaît la définition de la moyenne empirique :

$$\begin{aligned}
\hat{V} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n} 2\hat{\mu}X_i + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\hat{\mu}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) + \hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\hat{\mu}\hat{\mu} + \hat{\mu}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2.
\end{aligned}$$

2. Exprimez l'espérance de $\hat{\mu}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \ldots, X_n$ (prouvez votre résultat)

Solution : Notons $\mu_i := \mathbf{E}[X_i]$ pour $i = 1 \ldots n$. Par linéarité de l'ésperance :

$$\mathbf{E}[\hat{\mu}] := \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu_i.$$

On suppose à présent que $X_1, X_2, \ldots, X_n$ suivent toute la même distribution (elles sont donc i.i.d. puisqu'on a déjà supposé qu'elles étaient indépendantes).

3. Répondez à nouveau à la question précédente dans ce cadre plus simple.

Solution : Notons $\mu := \mathbf{E}[X_i]$ pour $i = 1 \dots n$.

On a à présent :

$$\mathbf{E}[\hat{\mu}] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

4. Exprimez la variance de $\hat{\mu}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \dots, X_n$ (prouvez votre résultat)

Solution : Notons $\sigma^2 := \mathbf{Var}[X_i]$ pour $i = 1 \dots n$. On a :

$$\begin{aligned}
\mathbf{Var}[\hat{\mu}] &= \mathbf{E}[\hat{\mu}^2] - \mathbf{E}[\hat{\mu}]^2 \\
&= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] - \mu^2 \\
&= \mathbf{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j\right] - \mu^2.
\end{aligned}$$

Par linéarité de l'espérance, on obtient :

$$\mathbf{Var}[\hat{\mu}] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{E}[X_i X_j] - \mu^2.$$

Pour $i \neq j$, les variables X_i et X_j sont indépendantes et on obtient $\mathbf{E}[X_i X_j] = \mathbf{E}[X_i]\mathbf{E}[X_j] = \mu^2$. Pour $i = j$, on a : $\mathbf{E}[X_i X_j] = \mathbf{E}[X_i]^2 = \sigma^2 + \mu^2$. Au final :

$$\begin{aligned}
\mathbf{Var}[\hat{\mu}] &= \frac{1}{n^2}\left(n(\sigma^2 + \mu^2) + n(n-1)\mu^2\right) - \mu^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

5. Exprimez l'espérance de $\hat{V}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \dots, X_n$ (prouvez votre résultat)

Solution : Par linéarité de l'espérance :

$$\begin{aligned}\mathbf{E}[\hat{V}] &= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \hat{\mu}^2\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i^2] - \mathbf{E}[\hat{\mu}^2].\end{aligned}$$

Or on a vu dans la réponse à la question précédente que $\mathbf{E}[X_i^2] = \mu^2 + \sigma^2$ et que $\mathbf{E}[\hat{\mu}^2] = \mu^2 + \sigma^2/n$. Donc :

$$\mathbf{E}[\hat{V}] = \sigma^2(1 - 1/n) = \frac{n-1}{n}\sigma^2.$$

6. Exprimez la variance de $\hat{V}$ comme une fonction de l'espérance et des moments centrés des variables $X_1, X_2, \ldots, X_n$ (prouvez votre résultat)

Exercice 5 Loi de l'espérance totale. ($\star$)

Un téléphone kadok tient en moyenne $12h$ avec la batterie A, mais seulement $8h$ avec la batterie B. La batterie A se trouve dans 80% des téléphones kadok, le reste étant muni de la batterie B. Si vous achetez un téléphone kadok, combien d'heure vous attendez-vous à ce qu'il tienne ?

3 Application à des problèmes d'estimation

Exercice 6 Calculs de biais, variance, risque. ($\star$)

1. Calculer le biais et la variance de l'estimateur de la moyenne empirique pour un échantillon i.i.d.

Solution : Soient $X_1, X_2, \ldots, X_n$ des variables aléatoires i.i.d. de moyenne μ et de variance σ^2. L'estimateur de la moyenne empirique est donné par :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Biais :

$$\text{Biais}(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \mu = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] - \mu = \mu - \mu = 0.$$

Donc, l'estimateur est non biaisé.

Variance :

$$\text{Var}(\hat{\mu}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) + 0 = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

2. Calculer le biais de l'estimateur de la variance suivant pour un échantillon i.i.d. : $\hat{V} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$

Solution : L'estimateur $\hat{V}$ est l'estimateur de la variance avec le dénominateur n. Nous cherchons $\text{Biais}(\hat{V}) = \mathbb{E}[\hat{V}] - \sigma^2$.

Calcul de $\mathbb{E}[\hat{V}]$:

$$\mathbb{E}[\hat{V}] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \hat{\mu})^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - 2\mathbb{E}[\hat{\mu}X_i] + \mathbb{E}[\hat{\mu}^2]$$

Or

(a) $\sigma^2 = \text{Var}[X_i] = \mathbb{E}[X_i^2] - \mu^2$, donc $\mathbb{E}[X_i^2] = \mu^2 + \sigma^2$.

(b) $\mathbb{E}[\hat{\mu}X_i] = \frac{1}{n}\sum_j \mathbb{E}[X_iX_j] = \frac{n-1}{n}\mu^2 + \frac{1}{n}\mathbb{E}[X_i^2] = \frac{n-1}{n}\mu^2 + \frac{\sigma^2}{n} + \frac{\mu^2}{n} = \mu^2 + \frac{\sigma^2}{n}$.

(c) $\mathbb{E}[\hat{\mu}^2] = \frac{1}{n^2}\sum_i\sum_j \mathbb{E}[X_iX_j] = \frac{1}{n^2}\sum_i \mathbb{E}[X_i^2] + \frac{1}{n^2}\sum_{i\neq j}\mu^2$
$= \frac{n}{n^2}(\mu^2 + \sigma^2) + \frac{n(n-1)}{n^2}\mu^2 = \mu^2 + \frac{\sigma^2}{n}$.

Donc

$$\mathbb{E}[\hat{V}] = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma^2 - \frac{\sigma^2}{n}\right) = \sigma^2\left(1 - \frac{1}{n}\right).$$

Biais :

$$\text{Biais}(\hat{V}) = \mathbb{E}[\hat{V}] - \sigma^2 = \sigma^2\left(1 - \frac{1}{n}\right) - \sigma^2 = -\frac{\sigma^2}{n}.$$

Donc, l'estimateur est biaisé vers le bas de $-\frac{\sigma^2}{n}$.

3. Proposez un estimateur non-biaisé de la variance pour un échantillon i.i.d.

Solution : Un estimateur non biaisé de la variance est obtenu en utilisant le dénominateur $n-1$:

$$\tilde{V} := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

Vérification du caractère non biaisé :

$$\mathbb{E}[\tilde{V}] = \mathbb{E}\left[\frac{n-1}{n}\hat{V}\right] = \frac{n-1}{n}\mathbb{E}[\hat{V}] = \sigma^2.$$

Ainsi, $\tilde{V}$ est un estimateur non biaisé de la variance.

4. Comparez le risque quadratique moyen des deux estimateurs de la variance.

5. Pouvez-vous donner un estimateur avec un risque plus faible que les deux considérés jusqu'ici ?

Exercice 7 Estimation par maximum de vraisemblance des paramètres d'une loi Gaussienne multivariée. ($\star$)

Supposons qu'on observe un échantillon i.i.d. $x_1, x_2, \ldots, x_n \in (\mathbf{R}^d)^n$ de loi gaussienne multivariée $\mathcal{N}(\mu^*, \Sigma^*)$, pour un vecteur $\mu \in \mathbf{R}^d$ et une matrice $\Sigma^* \in \mathbf{S}_d$, où $\mathbf{S}_d$ est l'ensemble des matrices à coefficients réels symétriques définies positives de taille d pas d. On cherche à estimer μ^* et Σ^* à partir de l'observation de $x_1, x_2, \ldots, x_n$.

On définit la *vraisemblance* d'un couple de paramètres $(\mu, \Sigma) \in \mathbf{R}^d \times \mathbf{S}_d$ comme :

$$\ell(\mu, \Sigma) := p(x_1, x_2, \ldots, x_n; \mu, \Sigma),$$

où p corresponds à la densité de probabilité de $x_1, x_2, \ldots, x_n$.

On considère l'estimateur du *maximum de vraisemblance* pour μ^*, Σ^* défini par

$$\hat{\mu}, \hat{\Sigma} \in \arg\max_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \ell(\mu, \Sigma).$$

1. Montrer que

$$\hat{\mu}, \hat{\Sigma} \in \arg\max_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \log(\ell(\mu, \Sigma)).$$

Solution : Soit $\hat{\mu}, \hat{\Sigma} \in \arg\max_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \ell(\mu,\Sigma)$. Soit $(\mu,\Sigma) \in \mathbf{R}^d \times \mathbf{S}_d$. On a $\ell(\mu,\Sigma) \leq \ell(\hat{\mu},\hat{\Sigma})$. Or log est une fonction croissante, donc $\log(\ell(\mu,\Sigma)) \leq \log(\ell(\hat{\mu},\hat{\Sigma}))$. On en déduit que $\sup_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \log(\ell(\mu,\Sigma)) \leq \log(\ell(\hat{\mu},\hat{\Sigma}))$ et comme $(\hat{\mu},\hat{\Sigma}) \in \mathbf{R}^d\times\mathbf{S}_d$:

$$\hat{\mu}, \hat{\Sigma} \in \arg\max_{(\mu,\Sigma)\in\mathbf{R}^d\times\mathbf{S}_d} \log(\ell(\mu,\Sigma)).$$

2. Donner une expression (la plus simple que vous pouvez) pour $\log(\ell(\mu,\Sigma)))$ en utilisant la formule donnant la densité d'une loi gaussienne multivariée non dégénérée.

Solution : L'échantillon étant considéré i.i.d., on a :

$$\begin{aligned}
\log(\ell(\mu,\Sigma))) &= \log\left(\prod_{i=1}^{n} p(x_i;\mu,\Sigma)\right) \\
&= \sum_{i=1}^{n} \log\left(p(x_i;\mu,\Sigma)\right) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\right) \\
&= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log(|\Sigma|) - \sum_{i=1}^{n}\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu).
\end{aligned}$$

3. On suppose que la matrice de covariance empirique $\frac{1}{n}(x_i-\mu)(x_i-\mu)^T$ est inversible et on admet que $\log(\ell(\mu,\Sigma)))$ est de classe C^1 et admet un unique maximum sur $\mathbf{R}^d \times \mathbf{S}_d$ en un point où son gradient s'annule. Calculer une expression explicite pour $(\hat{\mu},\hat{\Sigma})$ en fonction de $x_1, x_2, \ldots, x_n$. Vous pouvez utiliser les identités de calcul différentiel matriciel suivantes sans les démontrer :

$$\frac{\partial u^T M^{-1} v}{\partial M} = -(M^{-1})^T u v^T (M^{-1})^T,$$

$$\frac{\partial \det(M)}{\partial M} = \det(M)(M^{-1})^T,$$

où M est une matrice inversible et u et v sont des matrices colonnes de dimension compatible avec M (par exemple si M est de taille n par n, u et v sont de taille n par 1).

Solution : Commencons par calculer les gradients par rapport à μ et par rapport à Σ.

$$\begin{aligned}\nabla_\mu \log(\ell(\mu,\Sigma))) &= 0 + 0 + \nabla_\mu \left(-\sum_{i=1}^{n} \frac{1}{2}(x_i-\mu)\Sigma^{-1}(x_i-\mu) \right) \\ &= -\sum_{i=1}^{n} \frac{1}{2}\nabla_\mu((x_i-\mu)^T\Sigma^{-1}(x_i-\mu)).\end{aligned}$$

Or pour Σ symmétrique, $(x_i-\mu)^T\Sigma^{-1}(x_i-\mu) = x_i^T\Sigma^{-1}x_i - 2\mu^T\Sigma^{-1}x_i + \mu^T\Sigma^{-1}\mu$ et $\nabla_\mu(x_i^T\Sigma^{-1}\mu) = \Sigma^{-1}x_i$ et $\nabla_\mu(\mu^T\Sigma^{-1}\mu) = 2\Sigma^{-1}\mu$.
Donc :

$$\begin{aligned}\nabla_\mu \log(\ell(\mu,\Sigma))) &= -\sum_{i=1}^{n} \frac{1}{2}(-2\Sigma^{-1}x_i + 2\Sigma^{-1}\mu) \\ &= \Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu) \right).\end{aligned}$$

Pour le gradient par rapport à Σ, en utilisant la règle de dérivation pour les fonctions composées de plusieurs variables (par le biais des matrices jacobiennes) et les identités de calcul différentiel données ci-dessus (et la symmétrie de l'inverse d'une matrice symmétrique inversible) on obtient :

$$\begin{aligned}\nabla_\Sigma \log(\ell(\mu,\Sigma))) &= 0 - \frac{n}{2}J_{\log}(|\Sigma|)J_{|.|}(\Sigma) + \sum_{i=1}^{n}\frac{1}{2}\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\Sigma^{-1} \\ &= -\frac{n}{2}\frac{1}{|\Sigma|}|\Sigma|\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T \right)\Sigma^{-1} \\ &= \frac{1}{2}\Sigma^{-1}\left(\left(\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T \right)\Sigma^{-1} - nI \right).\end{aligned}$$

Solution : (Continuée)

On cherche à présent $(\hat{\mu},\hat{\Sigma}) \in \mathbf{R}^d \times \mathbf{S}_d$ qui annulent ces gradients. En utilisant le fait que toute matrice de $\mathbf{S}_d$ est inversible, on obtient que $\nabla_\mu \log(\ell(\hat{\mu},\hat{\Sigma}))) = 0$ implique $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\nabla_\Sigma \log(\ell(\hat{\mu},\hat{\Sigma}))) = 0$ implique $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$. On vérifie bien que, réciproquement, ce choix de $\hat{\mu}$ et $\hat{\Sigma}$ annule les gradients et que $(\hat{\mu},\hat{\Sigma}) \in \mathbf{R}^d \times \mathbf{S}_d$. Au final, on a obtenu que l'estimateur du maximum de vraisemblance pour la moyenne d'une loi Gaussienne multivariée est simplement la moyenne empirique usuelle et celui pour la covariance est simplement la covariance empirique usuelle (au moins dans le cas où cette dernière est inversible).

4. Montrer que le couple $(\hat{\mu}, \hat{\Sigma})$ ainsi obtenu forme une statistique suffisante pour le couple de paramètres (μ^*, Σ^*).

Solution : On va montrer qu'on peut écrire $p(x_1, \ldots, x_n; \mu^*, \Sigma^*)$ avec une expression qui ne dépend de $x_1, \ldots, x_n$ que par le biais de $\hat{\mu}$ et $\hat{\Sigma}$.

Pour simplifier la notation on note simplement (μ, Σ) pour (μ^*, Σ^*) dans la suite.

On a :

$$p(x_1, \ldots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{nd/2}(\sqrt{|\Sigma|})^n} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu)\right).$$

Réécrivons la seule partie où les x_i apparaissent en fonction de $\hat{\mu}$ et $\hat{\Sigma}$:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu) &= \sum_{i=1}^{n}\left((x_i - \hat{\mu})^T\Sigma^{-1}(x_i - \hat{\mu}) - \hat{\mu}^T\Sigma^{-1}\hat{\mu} + \mu^T\Sigma^{-1}\mu + 2(\hat{\mu}^T\Sigma^{-1}x_i - \mu^T\Sigma^{-1}x_i)\right) \\ &= n(\mu^T\Sigma^{-1}\mu + \hat{\mu}^T\Sigma^{-1}\hat{\mu} - 2\mu^T\Sigma^{-1}\hat{\mu}) + \sum_{i=1}^{n}(x_i - \hat{\mu})^T\Sigma^{-1}(x_i - \hat{\mu})\end{aligned}$$

Et :

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \hat{\mu})^T\Sigma^{-1}(x_i - \hat{\mu}) &= \mathbf{Tr}\left(\sum_{i=1}^{n}(x_i - \hat{\mu})^T\Sigma^{-1}(x_i - \hat{\mu})\right) \\ &= \sum_{i=1}^{n}\mathbf{Tr}\left((x_i - \hat{\mu})^T\Sigma^{-1}(x_i - \hat{\mu})\right) \\ &= \sum_{i=1}^{n}\mathbf{Tr}\left((x_i - \hat{\mu})(x_i - \hat{\mu})^T\Sigma^{-1}\right) \\ &= \mathbf{Tr}\left(\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T\Sigma^{-1}\right) \\ &= \mathbf{Tr}\left(n\hat{\Sigma}\Sigma^{-1}\right) = n\mathbf{Tr}\left(\hat{\Sigma}\Sigma^{-1}\right).\end{aligned}$$

Exercice 8 Analyse des propriétés basiques d'un estimateur. ($\star\star$)

Supposons qu'on observe un échantillon $x_1, x_2, \ldots, x_n \in (\mathbf{R}^d)^n$, i.i.d. de distribution P. Soit $d : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}$ une fonction mesurant la 'dissimilarité' entre deux points de $\mathbf{R}^d$. On suppose que d est symmétrique, c'est à dire que $d(x_1, x_2) = d(x_2, x_1)$ pour tout choix de x_1, x_2. Nous cherchons à estimer la dissimilarité moyenne entre deux points tirés aléatoirement et indépendamment suivant la distribution P :

$$\delta(P, d) := \mathbf{E}_{a,b\sim P\otimes P}[d(a, b)]$$

sur la base de $x_1, x_2, ..., x_n$ (la notation $a, b \sim P \otimes P$ signifie que a et b sont deux échantillons tirés

indépendamment de la loi P). On suppose que $\mathbf{E}_{a,b\sim P\otimes P}[d(a,b)^2] < +\infty$.

Considérons l'estimateur :

$$\hat{\delta}(x_1, x_2, \ldots, x_n) := \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} d(x_i, x_j).$$

1. Quel est le biais de $\hat{\delta}$?

Solution : Par linéarité de l'espérance :

$$\begin{aligned} b(\hat{\delta}) &:= \mathbf{E}[\hat{\delta}] - \delta \\ &= \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbf{E}[d(x_i, x_j)] - \delta \end{aligned}$$

Comme on a toujours $i \neq j$ et que $x_1, \ldots, x_n$ sont i.i.d de loi P, on a toujours $\mathbf{E}[d(x_i, x_j)] = \delta$. Donc :

$$\begin{aligned} b(\hat{\delta}) &= \left(\frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \delta \right) - \delta \\ &= \delta - \delta \\ &= 0. \end{aligned}$$

$\hat{\delta}$ est donc un estimateur non biaisé de δ.

2. Quelle est la variance de $\hat{\delta}$? On l'exprimera en fonction de $\sigma_1^2 = \mathrm{Var}_{x_1\sim P}\mathbf{E}_{x_2\sim P}[d(x_1, x_2)]$ et $\sigma_2^2 = \mathrm{Var}_{(x_1,x_2)\sim P\otimes P}[d(x_1, x_2)]$.

Solution : On peut, par exemple, appliquer la formules donnant la variance d'une quantité multipliée par une constante et la formule donnant la variance d'une somme en fonction des covariances des termes :

$$\mathbf{Var}(\hat{\delta}) := \frac{1}{\binom{n}{2}^2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=1}^{n} \sum_{l=k+1}^{n} \mathbf{Cov}(d(x_i, x_j), d(x_k, x_l))$$

On distingue trois cas :

— Si $\{i, j\} \cap \{k, l\} = \emptyset$, alors $\mathbf{Cov}(d(x_i, x_j), d(x_k, x_l)) = 0$.

— Sinon, si $|\{i, j\} \cap \{k, l\}| = 1$, alors par symétrie de d,

$$\begin{aligned}
\mathbf{Cov}(d(x_i, x_j), d(x_k, x_l)) &= \mathbf{Cov}_{a,b,c\sim P\otimes P\otimes P}(d(a, b), d(a, c)) \\
&= \mathbf{E}_{a,b,c\sim P\otimes P\otimes P}(d(a, b), d(a, c)) - [\mathbf{E}_{a,b\sim P\otimes P}(d(a, b))]^2 \\
&= \mathbf{E}_{a\sim P}[\mathbf{E}_{b,c\sim P\otimes P}[(d(a, b), d(a, c))]] - [\mathbf{E}_{a,b\sim P\otimes P}(d(a, b))]^2 \\
&= \mathbf{E}_{a\sim P}[\mathbf{E}_{b\sim P}[(d(a, b)]\mathbf{E}_{c\sim P}[d(a, c))]] - [\mathbf{E}_{a,b\sim P\otimes P}(d(a, b))]^2 \\
&= \mathbf{E}_{a\sim P}[\mathbf{E}_{b\sim P}[(d(a, b)]^2] - [\mathbf{E}_{a,b\sim P\otimes P}(d(a, b))]^2 \\
&= \sigma_1^2.
\end{aligned}$$

— Sinon, $\{i, j\} = \{k, l\}$ et $\mathbf{Cov}(d(x_i, x_j), d(x_k, x_l)) = \mathbf{Var}(d(x_i, x_j)) = \sigma_2^2$.

De plus, $\{i, j\} = \{k, l\}$ pour exactement $\binom{n}{2}$ quadruplets parmi les $\binom{n}{2}^2$ quadruplets considérés et $|\{i, j\} \cap \{k, l\}| = 1$ pour exactement :

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} n - i - 1 + n - j + i - 1 + j - 2 = 2(n - 2)\binom{n}{2}$$

quadruplets parmi les $\binom{n}{2}^2$ quadruplets considérés (pour chaque choix de i et j on a $n - i - 1$ cas avec $k = i$ et $l \neq j$, $n - j$ cas avec $k = j$ et $l \neq i$, $i - 1$ cas avec $l = i$ et $k \neq j$ et $j - 2$ cas avec $l = j$ et $k \neq i$).

Au final, on obtient :

$$\mathbf{Var}(\hat{\delta}) := \binom{n}{2}^{-1} [2(n - 2)\sigma_1^2 + \sigma_2^2].$$

3. Prouver l'inégalité de Markov : soit X est un variable aléatoire réelle positive, de moyenne

finie μ et soit t un réel strictement positif, alors :

$$p(X \geq t) \leq \frac{\mu}{t}.$$

Solution : Notons P la loi de X. Par définition :

$$\mu = \mathbf{E}[X] = \int X dP$$

Soit $t > 0$. Comme X est positive, on a toujours : $X \geq \mathbf{1}_{X \geq t} t$, où $\mathbf{1}_{X \geq t}$ est la fonction qui vaut 0 si $X < t$ et 1 sinon. On en déduit par monotonie de l'espérance que :

$$\mu \geq \int \mathbf{1}_{X \geq t} t dP = t \int \mathbf{1}_{X \geq t} dP = t p(X \geq t)$$

Donc :

$$p(X \geq t) \leq \frac{\mu}{t}.$$

4. Utiliser l'inégalité de Markov pour prouver l'inégalité de Chebyshev : soit X est un variable aléatoire réelle de moyenne finie μ et de variance finie σ^2 et soit t un réel strictement positif, alors :

$$p(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Solution : On considére $Y = (X - \mu)^2$. Y est positive et d'espérance égale à la variance de X (par définition), qui est finie et égale à σ^2 par hypothèse. Soit $t > 0$. $t^2 > 0$, on peut donc appliquer l'inégalité de Markov à Y en t^2. On obtient :

$$p((X - \mu)^2 \geq t^2) \leq \frac{\sigma^2}{t^2}.$$

Or $(X - \mu)^2 \geq t^2$ est équivalent à $|X - \mu| \geq t$, donc les évènements $\{(X - \mu)^2 \geq t^2\}$ et $\{|X - \mu| \geq t\}$ sont identiques et $p((X - \mu)^2 \geq t^2) = p(|X - \mu| \geq t)$. On obtient donc finalement :

$$p(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

5. Utiliser l'inégalité de Chebyshev et les résultats des deux premières questions pour montrer que $\hat{\delta}$ est un estimateur *faiblement consistant* de δ, c'est à dire qu'on a $\hat{\delta} \rightarrow_p \delta$, c'est à dire que pour tout $\epsilon > 0$:

$$\lim_{n \to +\infty} p(|\hat{\delta}(x_1, \ldots x_n) - \delta| \geq \epsilon) = 0.$$

Solution : $\hat{\delta}$ est un estimateur faiblement consistant de δ si et seulement si on a $\hat{\delta} \rightarrow_p \delta$, c'est à dire que pour tout $\epsilon > 0$:

$$\lim_{n \to +\infty} p(|\hat{\delta}(x_1, \dots x_n) - \delta| \geq \epsilon) = 0.$$

Soit $\epsilon > 0$ et n un entier supérieur ou égal à un. $\hat{\delta}(x_1, \dots, x_n)$ vérifie les hypothèses d'application pour l'inégalité de Chebyshev, qu'on applique en $t = \epsilon$ pour obtenir (en utilisant les résultats des deux premières questions pour exprimer l'espérance et la variance de $\hat{\delta}$) :

$$p((|\hat{\delta}(x_1, \dots x_n) - \delta| \geq \epsilon) \leq \frac{\binom{n}{2}^{-1}[2(n-2)\sigma_1^2 + \sigma_2^2]}{\epsilon^2}.$$

Le membre de droite de cette inégalité tend vers 0 quand n tend vers $+\infty$ (il est en $O(1/n)$) et le membre de gauche est positif (puisqu'il s'agit d'une probabilité). On peut donc appliquer le théorème des gendarmes pour conclure que :

$$\lim_{n \to +\infty} p(|\hat{\delta}(x_1, \dots x_n) - \delta| \geq \epsilon) = 0.$$