Aix*Marseille université

Master 2 informatique

Parcours IAAA, UE MIA

Feuille d'exercices numéro 3

Année 2025-26

Probabilités et maximum de vraisemblance.

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 ($\star$)

On considère à présent une variable aléatoire binaire Z et deux variables aléatoires réelles X et Y telles que $P(X, Y, Z) = P(X|Z)P(Y|Z)P(Z)$ avec

1. $P(Z = 1) = 2/3$
2. $P(X|Z = 0) = \mathcal{N}(X; 0, 1)$
3. $P(X|Z = 1) = \mathcal{N}(X; 1, 1)$
4. $P(Y|Z = 0) = \mathcal{N}(Y; 0, 1)$
5. $P(Y|Z = 1) = \mathcal{N}(Y; -1, 1)$

où

$$\mathcal{N}(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

est la densité d'une loi normale de moyenne μ et de variance σ^2. On suppose qu'on observe $Y = y_0$. Calculez $P(X|Y = y_0)$.

Solution : On souhaite calculer la probabilité conditionnelle $P(X|Y = y_0)$.

Étape 1 : Connection avec la loi jointe

En utilisant la définition d'une probabilité conditionnelle, on peut écrire :

$$P(X|Y = y_0) = \frac{P(X, Y = y_0)}{P(Y = y_0)}$$

où $P(X, Y = y_0)$ s'obtient en marginalisant la loi jointe par rapport à Z pour $Y = y_0$ et un X arbitraire :

$$P(X, Y = y_0) = \sum_{z \in \{0,1\}} P(X, Y = y_0, Z = z)$$

et $P(Y = y_0)$ s'obtient à partir de $P(X|Y = y_0)$ en marginalisant sur X :

$$P(Y = y_0) = \int_{-\infty}^{\infty} P(X, Y = y_0)\, dX$$

Remarquez qu'en utilisant la définition des probabilités conditionnelles, on a réussi à connecter le calcul de la quantité désirée à de simples opérations de marginalisation de la loi jointe, plus le calcul d'expression algébriques simples (ici un quotient). Cette approche est toujours possible en calcul probabiliste et je vous encourage à la suivre systématiquement.

(solution continuée ci-après)

Solution : (suite de la solution)

Étape 2 : Calcul de $P(X, Y = y_0)$

En utilisant le fait que $P(X, Y, Z) = P(X|Z)P(Y|Z)P(Z)$, on obtient :

$$P(X, Y = y_0) = \sum_{z \in \{0,1\}} P(X|Z = z)P(Y = y_0|Z = z)P(Z = z)$$

On dispose des informations suivantes :

- $P(Z = 1) = \frac{2}{3}$, donc $P(Z = 0) = \frac{1}{3}$
- $P(X|Z = 0) = \mathcal{N}(X; 0, 1)$
- $P(X|Z = 1) = \mathcal{N}(X; 1, 1)$
- $P(Y = y_0|Z = 0) = \mathcal{N}(y_0; 0, 1)$
- $P(Y = y_0|Z = 1) = \mathcal{N}(y_0; -1, 1)$

On peut maintenant insérer ces termes dans l'expression pour $P(X, Y = y_0)$ et obtenir :

$$P(X, Y = y_0) = \frac{1}{3}\mathcal{N}(X; 0, 1)\mathcal{N}(y_0; 0, 1) + \frac{2}{3}\mathcal{N}(X; 1, 1)\mathcal{N}(y_0; -1, 1)$$

(solution continuée ci-après)

Solution : (suite de la solution)

Étape 3 : Calcul de $P(Y = y_0)$

Ensuite, on calcule $P(Y = y_0)$ en marginalisant sur X l'expression qu'on vient d'obtenir :

$$\begin{aligned} P(Y = y_0) &= \int_{-\infty}^{+\infty} P(X, Y = y_0) dX \\ &= \int_{-\infty}^{+\infty} \left(\frac{1}{3}\mathcal{N}(X; 0, 1)\mathcal{N}(y_0; 0, 1) + \frac{2}{3}\mathcal{N}(X; 1, 1)\mathcal{N}(y_0; -1, 1) \right) dX \\ &= \frac{1}{3}\mathcal{N}(y_0; 0, 1) \int_{-\infty}^{+\infty} \mathcal{N}(X; 0, 1) dX + \frac{2}{3}\mathcal{N}(y_0; -1, 1) \int_{-\infty}^{+\infty} \mathcal{N}(X; 1, 1) dX \end{aligned}$$

par linéarité de l'intégrale.

Comme une densité de probabilité somme toujours à 1, c'est aussi le cas pour une densité Gaussienne et donc :

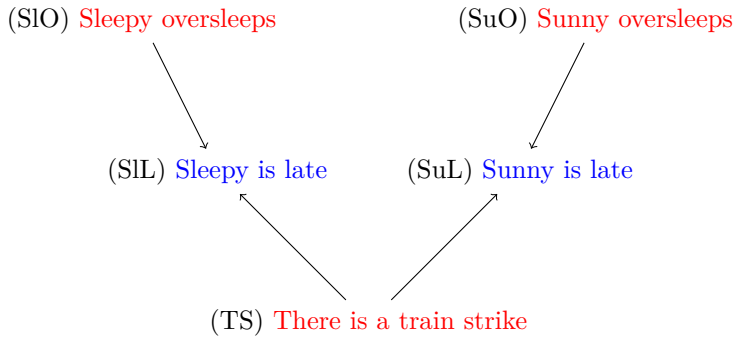
$$P(Y = y_0) = \frac{1}{3}\mathcal{N}(y_0; 0, 1) + \frac{2}{3}\mathcal{N}(y_0; -1, 1).$$

Étape 4 : Conclusion

L'expression finale obtenue pour $P(X|Y = y_0)$ est :

$$P(X|Y = y_0) = \frac{\frac{1}{3}\mathcal{N}(X; 0, 1)\mathcal{N}(y_0; 0, 1) + \frac{2}{3}\mathcal{N}(X; 1, 1)\mathcal{N}(y_0; -1, 1)}{\frac{1}{3}\mathcal{N}(y_0; 0, 1) + \frac{2}{3}\mathcal{N}(y_0; -1, 1)}.$$

Exercice 3 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 representé 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 :

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

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.

Exercice 4 Calcul probabiliste ($\star$)

On considère quatre variables aléatoires X_1, X_2, Z_1, Z_2, dont la loi jointe s'écrit $p(X_1, X_2, Z_1, Z_2) = p(X_1|Z_1)p(X_2|Z_2)p(Z_2|Z_1)p(Z_1)$, avec :

— Z_1 binaire, $p(Z_1 = 1) = p$.

— Z_2 binaire, $p(Z_2 = 1|Z_1 = 0) = q_1$ et $p(Z_2 = 1|Z_1 = 1) = q_2$

— $p(X_1 = x|Z_1 = 0) \sim \mathcal{N}(x; 0, 1)$

— $p(X_1 = x|Z_1 = 1) \sim \mathcal{N}(x; 1, 1)$

— $p(X_2 = x|Z_2 = 0) \sim \mathcal{N}(x; 0, 1)$

— $p(X_2 = x|Z_2 = 1) \sim \mathcal{N}(x; -1, 1)$

où $\mathcal{N}(x; \mu, \sigma^2)$ donne la densité d'une loi gaussienne de moyenne μ et de variance σ^2 en x.

1. On suppose qu'on observe $X_1 = x_1$. Donnez une expression pour $p(X_2|X_1 = x_1)$ en fonction de x_1, p, q_1 et q_2.
2. On suppose à présent qu'on observe $X_1 = 0$ et $X_2 = 0$. Donnez une expression pour $p(X_1 = 0, X_2 = 0)$ en fonction de p, q_1 et q_2, calculez son gradient par rapport à p, q_1 et q_2 et donnez les valeurs de p, q_1 et q_2 pour lesquels ce gradient s'annule.
3. Que est l'intérêt de ce calcul ?

Exercice 5 Calcul probabiliste ($\star$)

On considère quatre variables aléatoires X_1, X_2, Z_1, Z_2, dont la loi jointe s'écrit $p(X_1, X_2, Z_1, Z_2) = p(X_1|Z_1)p(X_2|Z_2)p(Z_1)p(Z_2)$, avec pour $i \in 1, 2$:

— $p(Z_i = 0) = p$, $p(Z_i = 1) = q$, $p(Z_i = 2) = 1 - p - q$

— $p(X_i = x|Z_i = 0) \sim \mathcal{N}(x; 0, 1)$

— $p(X_i = x|Z_i = 1) \sim \mathcal{N}(x; 1, 1)$

— $p(X_i = x|Z_i = 2) \sim \mathcal{N}(x; -1, 1)$

où $\mathcal{N}(x; \mu, \sigma^2)$ donne la densité d'une loi gaussienne de moyenne μ et de variance σ^2 en x.

1. On suppose qu'on observe $X_1 = x_1$. Donnez une expression pour $p(Z_1|X_1 = x_1)$ en fonction de x_1, p, et q.
2. On suppose qu'on observe $X_2 = x_2$. Donnez une expression pour $p(Z_1|X_2 = x_2)$ en fonction de x_2, p, et q.
3. On suppose à présent qu'on observe $X_1 = 0$ et $X_2 = 0$. Donnez une expression pour $p(X_1 = 0, X_2 = 0)$ en fonction de p et q.
4. Trouvez les valeurs de p et q qui maximisent $p(X_1 = 0, X_2 = 0)$ (justifiez votre réponse).
5. Que est l'intérêt de ce calcul ?

2 Maximum de vraisemblance

Exercice 6 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 :

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