

Quiz Partie 1 : modélisation statistique (corrigé)

Durée : 45 min

Nom : Prénom :

Justify all your answers briefly. ♦ *Justifiez brièvement toutes vos réponses.***Fundamentals** *Questions courtes*

Q1. (2 pts) $\mathbf{X} = (X_1, X_2)^\top \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = (12, 11)^\top$ and $\Sigma = \begin{pmatrix} 16 & 12 \\ 12 & 25 \end{pmatrix}$. What is the distribution of $D = X_1 - X_2$. Justify. ♦ *Quelle est la distribution de $D = X_1 - X_2$? Justifiez.*

$D = a^\top \mathbf{X}$ with $a = (1, -1)^\top$: every linear combination of a Gaussian vector is Gaussian (definition), with mean $a^\top \mu = 1$ and variance $a^\top \Sigma a = 16 + 25 - 2 \times 12 = 17$. So $D \sim \mathcal{N}(1, 17)$.

Grading (2 pts): Gaussian, because a linear combination of a Gaussian vector is Gaussian: 0.5. Mean = 1: 0.5. Variance = 17: 1 (0.5 if $a^\top \Sigma a$ is right but the arithmetic slips, e.g. 41 from forgetting -2Cov).

Q2. (1 pt) Fill in the blank. Let A and B be two random variables. If $(A, B)^\top$ is a **Gaussian vector**, then $\text{Cov}(A, B) = 0 \iff A$ and B are independent. ♦ *Complétez. Soient A et B deux variables aléatoires. Si $(A, B)^\top$ est un vecteur gaussien, alors $\text{Cov}(A, B) = 0 \iff A$ et B sont indépendantes.*

Grading (1 pt): All or nothing: “ (A, B) is a Gaussian vector” (or “jointly Gaussian”). “ A and B are Gaussian” gets 0: Gaussian marginals are not enough.

Q3. (2 pts) A bank describes $n = 40$ clients by $d = 60$ variables and fits a Gaussian model $\mathcal{N}(\mu, \Sigma)$ by maximum likelihood. What could go wrong when estimating μ ? When estimating Σ ? Propose two ways to fix it. ♦ *Une banque décrit $n = 40$ clients par $d = 60$ variables et ajuste un modèle gaussien $\mathcal{N}(\mu, \Sigma)$ par maximum de vraisemblance. Qu'est-ce qui pourrait mal se passer lors de l'estimation de μ ? Lors de l'estimation de Σ ? Proposez deux façons d'y remédier.*

$\hat{\mu}$: no problem, it is d averages, each of 40 values. Σ has $d(d+1)/2 = 1830$ parameters for 40 points. $\text{rank}(\hat{\Sigma}) \leq \min(n-1, d) = 39 < 60$: $\hat{\Sigma}$ is singular, so $\hat{\Sigma}^{-1}$, the density and the Mahalanobis distances do not exist (the likelihood has no maximizer). Fixes: (1) shrinkage $(1-\lambda)\hat{\Sigma} + \lambda\tau I$, $\tau = \text{tr}(\hat{\Sigma})/d$, invertible at the price of a bias; (2) reduce the dimension (PCA) before fitting. Also accepted: a diagonal model.

Grading (2 pts): $\hat{\mu}$ is fine (d averages of 40 values): 0.5. $\hat{\Sigma}$ singular, with a reason ($\text{rank} \leq n-1$, or 1830 parameters for 40 points): 0.5. First fix: 0.5. Second fix: 0.5 (shrinkage, PCA, diagonal model).

Q4. (1 pt) We reject H_0 when the p-value is small and do not reject it when the p-value is large. Does this mean that the p-value is the probability that H_0 is true? ♦ *On rejette H_0 quand la p-valeur est petite et on ne la rejette pas quand elle est grande. Cela signifie-t-il que la p-valeur est la probabilité que H_0 soit vraie ?*

No. The p-value is computed *assuming* H_0 is true: the probability, under H_0 , of a statistic at least as extreme as the one observed. H_0 is not random; a small p-value means the data are unlikely under H_0 , not that H_0 has a small probability.

Grading (1 pt): “No” alone: 0. Explanation (computed assuming H_0 ; probability of data at least this extreme): 1.

Situations *Mises en situation*

Q5. (8 pts) **Customer segments.** A retailer describes $n = 400$ customers by two features x_1, x_2 (visit frequency and basket size, rescaled). It fits Gaussian mixtures with full covariance matrices and $K = 1, \dots, 6$ components by EM. ℓ_K is the maximized log-likelihood, p_K the number of parameters, $\text{BIC}_K = -2\ell_K + p_K \log n$ (computed with $\log n \approx 6$). ♦ **Segmentation de clientèle.** Un admin d'un site e-commerce décrit $n = 400$ clients par deux variables x_1, x_2 (fréquence de visite et taille du panier, remises à l'échelle). Il ajuste par EM des mélanges gaussiens à matrices de covariance pleines, à $K = 1, \dots, 6$ composantes. ℓ_K est la log-vraisemblance maximisée, p_K le nombre de paramètres,

$\text{BIC}_K = -2\ell_K + p_K \log n$ (calculé avec $\log n \approx 6$).

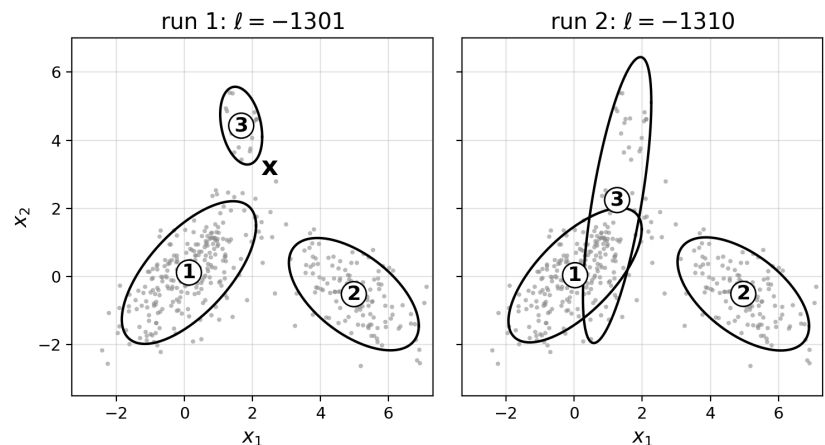
K	1	2	3	4	5	6
ℓ_K	-1620	-1360	-1301	-1299	-1298	-1295
p_K	5	11	17	23	29	35
BIC_K	3270	2786	2704	2736	2770	2800

- a) (3 pts) Why do we compute the BIC to choose K ? What does it mean to choose K by minimizing the BIC rather than by maximizing the likelihood? ♦ *Pourquoi calcule-t-on le BIC pour choisir K ? Que signifie choisir K en minimisant le BIC plutôt qu'en maximisant la vraisemblance?*

The likelihood always increases with K (-1620 up to -1295): maximizing it picks the largest K and overfits (components that fit a few points or noise). $\text{BIC} = -2\ell_K + p_K \log n$ adds a penalty that grows with the number of parameters: minimizing it looks for the best fit that the complexity can justify. Here $K = 3$ (2704): beyond it, ℓ_K gains at most 3 while each extra component costs 6 parameters, i.e. +36.

Grading (3 pts): Mention understanding of overfitting: ℓ_K always increases with K , so maximizing it overfits; $\text{BIC} = \text{fit} + \text{a penalty growing with the number of parameters}$; Minimum BIC, hence $K = 3$.

The retailer keeps $K = 3$. The figure shows two EM runs, same data, same $K = 3$, two different initializations. ♦ *Le distributeur garde $K = 3$. La figure montre deux exécutions d'EM : mêmes données, même $K = 3$, deux initialisations différentes.*



- b) (3 pts) In run 1, the new customer $\mathbf{x}$ is closest to the center of segment 3, yet EM assigns it to segment 1. How is that possible? ♦ *Dans l'exécution 1, le segment 3 est le plus proche du nouveau client $\mathbf{x}$, pourtant EM l'affecte au segment 1. Comment est-ce possible?*

$\gamma_k = \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k) / \sum_j \pi_j \mathcal{N}(\mathbf{x}; \mu_j, \Sigma_j)$: the assignment depends on the weights and the covariances, not on the distance to the centers. Weights: segment 3 holds far fewer customers than segment 1 (few points in its ellipse), so a small π_3 can outweigh a higher density under segment 3. Covariances: segment 3 is tight, so its density falls fast away from its center, while $\mathbf{x}$ lies along the elongated direction of segment 1. Either argument, well explained, earns credit. K-means (nearest center) would put $\mathbf{x}$ in segment 3.

Grading (3 pts): Responsibility formula $\gamma_k \propto \pi_k \mathcal{N}(\mathbf{x}; \mu_k, \Sigma_k)$ or mentioning the responsibilities / probabilities and that distances are not enough. One argument well explained (weights or covariances).

- c) (2 pts) Same algorithm, same data: why do the two runs differ? Which run do you keep? ♦ *Même algorithme, mêmes données : pourquoi les deux exécutions diffèrent-elles? Laquelle gardez-vous?*

EM only reaches a local maximum of the likelihood, which depends on the initialization. Keep run 1: higher log-likelihood ($-1301 > -1310$), and its ellipses match the three visible groups (in run 2, component 3 stretches from the small group down through segment 1). In practice: several initializations, keep the largest ℓ .

Grading (2 pts): EM reaches a local maximum that depends on the initialization: 1. Keep run 1 because its ℓ is higher: 1.

Q6. (6 pts) Electricity bills. The monthly electricity bill (MAD) of $n = 300$ households is regressed on the surface (m^2), the number of rooms, the number of occupants and an indicator “insulated” (1 if the house is insulated). The surface and the number of rooms have correlation 0.92. Gaussian linear model output: ♦ *Factures d’électricité.* La facture mensuelle d’électricité (MAD) de $n = 300$ ménages est régressée sur la surface (m^2), le nombre de pièces, le nombre d’occupants et un indicateur « isolé » (1 si le logement est isolé). La surface et le nombre de pièces ont une corrélation de 0,92. Sortie du modèle linéaire gaussien :

covariate / covariable	estimate	SE	95% CI / IC à 95%	p (t-test)
intercept	106.9	48.6	[11.3, 202.5]	0.028
surface	1.98	0.71	[0.58, 3.37]	0.0056
rooms / pièces	26.1	20.1	[−13.5, 65.7]	0.20
occupants	31.2	7.5	[16.5, 45.9]	$3.9 \cdot 10^{-5}$
insulated / isolé	−78.6	25.8	[−129.4, −27.8]	0.0025

a) (3 pts) Interpret the effects of “occupants” and “insulated” on the bill. ♦ *Interprétez les effets de « occupants » et de « isolé » sur la facture.*

All else being equal (same surface, rooms, insulation), one more occupant goes with a bill higher by 31 MAD per month on average; significant ($p = 3.9 \cdot 10^{-5}$, CI [16.5, 45.9] excludes 0). At equal surface, rooms and occupants, an insulated house pays on average 79 MAD less per month; significant ($p = 0.0025$, CI [−129.4, −27.8]).

Grading (3 pts): Per variable, 1.5: direction and size with units (0.5), “all else being equal” (0.5), significance from the p-value or the CI (0.5).

b) (3 pts) A colleague concludes: “rooms is not significant, so the number of rooms has no influence on the bill.” Yet regressing the bill on the number of rooms *alone* gives +75.6 MAD per room with $p \approx 7 \cdot 10^{-17}$. Explain these two results. ♦ *Un collègue conclut : « la variable pièces n’est pas significative, donc le nombre de pièces n’a aucune influence sur la facture. » Pourtant, la régression de la facture sur le nombre de pièces seul donne +75,6 MAD par pièce avec $p \approx 7 \cdot 10^{-17}$. Expliquez ces deux résultats.*

(1) Failing to reject is not accepting H_0 : the CI [−13.5, 65.7] is wide. (2) In the full model, the coefficient is the effect of one more room *at fixed surface*; alone, rooms also carries the effect of surface, which it tracks (correlation 0.92).

Grading (3 pts): Not significant $\neq$ no effect, the CI is wide. At fixed surface vs. alone, where rooms also carries the effect of surface. Correlation 0.92 inflates the SE, so the test has low power.