Statistique mathématique – examen

M1 Mathématiques Fondamentales, Université Paris-Saclay
2024-2025

Vendredi 18 avril 2025
9h – 12h

Une feuille A4 recto-verso manuscrite est autorisée en tant que support. La calculatrice n'est pas autorisée.

Avant de commencer :

- *Le sujet comporte deux problèmes totalement indépendants. Il n'est pas attendu de traiter l'intégralité du sujet pour avoir la note maximale.*
- *Des résultats pourront être admis d'une question sur l'autre, à la condition de l'écrire clairement.*
- *Les questions marquées d'une astérisque sont des questions bonus, dans laquelle toute trace de recherche même non aboutie pourra rapporter des points. Elles sont bien sûr facultatives.*
- *Une attention particulière sera portée aux questions d'interprétation, à la rigueur et à la précision de la rédaction.*

Bon courage, et faites vous plaisir !

Problème I – Modèle linéaire avec bruits corrélés

On se place dans un modèle linéaire *non supposé gaussien*, où Y, vecteur aléatoire de $\mathbb{R}^n$, s'écrit

$$Y = X\theta + \varepsilon, \tag{1}$$

avec $X \in \mathbb{R}^{n\times p}$ une matrice déterministe, $\theta \in \mathbb{R}^p$ un vecteur de paramètres inconnus, et ε un vecteur aléatoire de $\mathbb{R}^n$ centré. *Contrairement au cas standard, on suppose ici que le bruit ε a une matrice de covariance $\Sigma \in \mathbb{R}^{n\times n}$ supposée symétrique définie positive, qui n'est pas forcément de la forme $\sigma^2 I_n$.* On suppose que Σ est connue.

1. Montrer que sous ces nouvelles hypothèses, on a toujours :

$$\text{le modèle (1) est identifiable en } \theta \iff X^TX \text{ inversible}\,.$$

Dans toute la suite de cet exercice, nous nous plaçons sous les hypothèses d'identifiabilité de la question précédente. On note :

$$\hat{\theta} := \arg\min_{\theta\in\mathbb{R}^p} \|Y - X\theta\|^2,$$

avec $\|\cdot\|$ norme euclidienne usuelle sur $\mathbb{R}^n$. On rappelle qu'il est unique par hypothèse, et dans cet exercice il est appelé *estimateur des moindres carrés ordinaire.*

2. Rappeler la forme close de $\hat{\theta}$, et déterminer l'espérance puis la matrice de covariance de $\hat{\theta}$ en fonction de θ, Σ et X.

Notons $\|\cdot\|_\Sigma$ la norme euclidienne sur $\mathbb{R}^n$ définie pour tout $u \in \mathbb{R}^n$ par $\|u\|_\Sigma^2 = u^T\Sigma^{-1}u$. On considère le nouveau problème d'optimisation :

$$\arg\min_{\theta\in\mathbb{R}^p} \|Y - X\theta\|_\Sigma^2\,. \tag{2}$$

3. Montrer que $X^T\Sigma^{-1}X$ est symétrique définie positive. *On pourra noter $S \in \mathbb{R}^{n\times n}$ la matrice symétrique définie positive telle que $S^2 = \Sigma^{-1}$ (dont on admettra l'existence et l'unicité).*

4. Prouver que le problème (2) admet une unique solution appelée *estimateur des moindres carrés généralisé*, notée $\hat{\theta}_g$, qui vaut

$$\hat{\theta}_g = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y\,.$$

5. Déterminer l'espérance puis la matrice de covariance de $\hat{\theta}_g$ en fonction de θ, Σ et X.

6. On note $\preceq$ l'ordre partiel sur les matrices symétriques défini dans le cours. En utilisant la matrice S définie à la question 3., montrer que

$$X(X^T\Sigma^{-1}X)^{-1}X^T \preceq \Sigma$$

puis en déduire

$$\mathrm{Var}(\hat{\theta}_g) \preceq \mathrm{Var}(\hat{\theta})\,.$$

Pour la première inégalité, on pourra prendre $v^TX(X^T\Sigma^{-1}X)^{-1}X^Tv$ pour $v \in \mathbb{R}^n$, écrire $v = SS^{-1}v$ et reconnaître une certaine projection orthogonale.

7. A quel théorème du cours le dernier résultat de la question 6 vous fait-il penser ? En s'inspirant de celui-ci, formulez une généralisation de ce théorème au cas où les bruits sont corrélés. *On ne demande aucune démonstration.*

8. On rappelle que Σ est connue. On suppose dans cette question que le modèle est gaussien, i.e. que $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Proposez un intervalle de confiance de probabilité de couverture $1 - \alpha$ pour θ_1, le premier coefficient de θ. On cherchera un intervalle de diamètre le plus petit possible (bien qu'il ne soit pas demandé de prouver une quelconque optimalité).

9⋆. On suppose dans cette question que le modèle est gaussien, que la matrice Σ est inconnue mais qu'on a $\|\Sigma\|_{\text{op}} \leq R^2$ avec $R > 0$. Comment pourrait-on répondre à la question 8 ?

10⋆. On ne suppose plus que le modèle est gaussien. Si Σ est inconnue, comment pourrait-on l'estimer ?

Problème II – Plantons des étoiles

"Il faut porter encore en soi un chaos, pour pouvoir mettre au monde une étoile dansante"
(Nietzsche)

Dans ce problème, pour tout $n \geq 1$, $[n]$ désigne l'ensemble $\{1, \dots, n\}$. On note $\mathbb{G}_n$ l'ensemble des graphes sur l'ensemble de sommets $V = [n]$, non orientés, simples et sans boucle[1]. Quand le graphe $g \in \mathbb{G}_n$ dont on parle est clair dans le contexte, on note $u \sim v$ (resp. $u \not\sim v$) si $u \in [n]$ et $v \in [n]$ sont reliés (resp. non reliés) dans g par une arête.

Pour $g \in \mathbb{G}_n$, on dit qu'un couple $\mathcal{E} = (u, W)$ où $u \in [n]$ et $W \subseteq [n] \setminus \{u\}$ est une *$k-$étoile de G centrée en u*, et l'on note $\mathcal{E} \in G$, si $|W| = k$ et si

$$\forall w \in W, \quad u \sim w \,.$$

Nous allons considérer des graphes aléatoires[2] . Pour $n \geq 1$, et deux paramètres $p \in [0, 1]$ et $k \geq 0$, *où p et k dépendent éventuellement de n*, on note :

- $\mathcal{G}(n, p)$ la loi d'un graphe aléatoire G dans $\mathbb{G}_n$ où, indépendamment pour toute paire $\{u, v\}$ de sommets avec $u \neq v \in [n]$, $u \sim v$ (resp. $u \not\sim v$) avec probabilité p (resp. $1 - p$).
- $\mathcal{G}(n, p, k)$ le modèle où G est obtenu de la façon suivante. On tire d'abord $H \sim \mathcal{G}(n, p)$, puis, indépendamment de H, on tire un noeud $u \in [n]$ uniformément au hasard, et un sous ensemble $W \subset [n] \setminus \{u\}$ de cardinal k uniformément au hasard. On note $\mathcal{E}(G) = (u, W)$ et on forme le graphe G en ajoutant à H les arêtes nécessaires de sorte que $\mathcal{E}(G) = (u, W)$ soit une $k-$étoile de G. *On dit qu'on a planté la $k-$étoile $\mathcal{E}(G)$ dans G.*

Notons que par définition, $\mathcal{G}(n, p, 0) = \mathcal{G}(n, p)$. Le but de ce problème est de détecter la présence de $k-$étoiles plantées dans des graphes, autrement dit, d'étudier le test d'hypothèses suivant :

$$\mathcal{H}_0 : G \sim \mathcal{G}(n, p) \quad \text{contre} \quad \mathcal{H}_1 : G \sim \mathcal{G}(n, p, k) \,. \tag{$\star$}$$

Lorsque les paramètres n, p, k sont clairs dans le contexte, on note $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) la distribution de G sous $\mathcal{H}_0$ (resp. sous $\mathcal{H}_1$).

A – Premières questions

[1] c'est-à-dire que toutes les arêtes de tels graphes sont de la forme $\{u, v\}$ avec $u, v \in [n]$ et $u \neq v$.

[2] si cela vous semble loin du contexte du cours, observez que l'on peut parfaitement encoder un graphe de $\mathbb{G}_n$ par un vecteur de $\mathbb{R}^{\binom{n}{2}}$ (et même de $\{0, 1\}^{\binom{n}{2}}$). Un graphe aléatoire peut ainsi être vu comme un vecteur aléatoire.

1. En donnant des arguments succints et informels, à votre avis, comment la difficulté du problème du test d'hypothèses ($\star$) varie :
   - à p fixé, lorsque k augmente ?
   - à k fixé, lorsque p augmente ?

Dans toute la suite, on considère le régime où p dépend de n et s'écrit $p = p(n) = \lambda/n$ avec λ réel strictement positif.

2. Pour une telle paramétrisation de p, le modèle $\mathcal{G}(n,p)$ est dit *de degré moyen d'ordre constant en n*. Justifier cette appellation.

Pour tout $1 \leq k \leq n$, on note $\mathcal{E}(n,k)$ l'ensemble des $k-$étoiles du graphe complet[3] de taille n.

3. Montrer que
$$|\mathcal{E}(n,k)| = n\binom{n-1}{k}.$$

B – Calcul de L

4. On rappelle que $p = p(n) = \lambda/n$. Montrer que pour tout graphe $g \in \mathbb{G}_n$,
$$\mathbb{P}_1(G = g) = \frac{1}{|\mathcal{E}(n,k)|}\left(\frac{\lambda}{n}\right)^{-k}\mathbb{P}_0(G = g)\sum_{\mathcal{E}\in\mathcal{E}(n,k)}\mathbb{1}_{\mathcal{E}\in g}\,.$$
On pourra commencer par calculer $\mathbb{P}_0(G = g)$, puis écrire
$$\mathbb{P}_1(G = g) = \sum_{\mathcal{E}\in\mathcal{E}(n,k)}\mathbb{P}_1(G = g\,|\,\mathcal{E}(G) = \mathcal{E})\mathbb{P}_1(\mathcal{E}(G) = \mathcal{E})\,.$$

Pour tout graphe $g \in \mathbb{G}_n$, on note
$$L(g) := \frac{\mathbb{P}_1(G = g)}{\mathbb{P}_0(G = g)}$$
et $X_k(g)$ le nombre de $k-$étoiles de g.

5. Déduire de la question précédente que pour tout graphe $g \in \mathbb{G}_n$,
$$L(g) = \frac{X_k(g)}{\mathbb{E}_0[X_k(G)]}\,.$$

6. Si l'on fixe un niveau α, quelle est la forme d'un test ϕ uniformément plus puissant pour le problème ($\star$) ? *On ne demande pas de calculer les constantes qui apparaissent, mais juste de donner sa forme globale.* Au vu de la question 5, en quoi ce test ϕ a-t-il une expression intuitive pour répondre au problème ($\star$) ?

Dans toute la suite, on dit que

- *La détection forte d'une $k-$étoile est réalisable* si pour tout $n \geq 1$ il existe un test $\phi_n : \mathbb{G}_n \to \{0,1\}$ tel que
$$\mathbb{P}_0(\phi_n = 1) + \mathbb{P}_1(\phi_n = 0) \underset{n\to+\infty}{\longrightarrow} 0\,.$$

[3] pour rappel, le graphe complet de taille n est le graphe de $\mathbb{G}_n$ dans lequel $u \sim v$ pour tous $u \neq v \in [n]$.

- *La détection faible d'une $k-$étoile est impossible* si pour toute famille de tests $(\phi_n)_{n\geq 1}$ avec $\phi_n : \mathbb{G}_n \to \{0,1\}$, on a

$$\mathbb{P}_0(\phi_n = 1) + \mathbb{P}_1(\phi_n = 0) \underset{n\to+\infty}{\longrightarrow} 1\,.$$

C – Un résultat positif

7. Montrer que

$$\mathbb{E}_0[X_k(G)] \leq n\frac{\lambda^k}{k!}\,.$$

8. Soit $\varepsilon > 0$. On suppose que $n \geq 3$ et que

$$k = k(n) \geq (1+\varepsilon)\frac{\log n}{\log\log n}\,.$$

Montrer qu'alors la détection forte d'une $k-$étoile est réalisable pour un test ϕ_n très simple que vous expliciterez.

On commencera par montrer que sous ces conditions, $\mathbb{P}_0(X_k(G) \geq 1) \underset{n\to\infty}{\longrightarrow} 0$.

D – Une inégalité informationnelle

Soient P_0 and P_1 deux mesures de probabilités définies sur un espace au plus dénombrable Γ. On notera, $P_0(\gamma) = P_0(\{\gamma\})$ pour tout $\gamma \in \Gamma$, et de même pour P_1. La *distance en variation totale entre P_0 et P_1* est définie par

$$d_{\mathrm{TV}}(P_0, P_1) := \sup_{B\subseteq\Gamma}(P_1(B) - P_0(B)).$$

9. Montrer que le sup dans la définition ci-dessus est atteint pour $B = B_{\mathrm{opt}}$ où

$$B_{\mathrm{opt}} := \{\gamma \in \Gamma,\, P_1(\gamma) > P_0(\gamma)\}$$

et en déduire que

$$d_{\mathrm{TV}}(P_0, P_1) = \frac{1}{2}\sum_{\gamma\in\Gamma}|P_1(\gamma) - P_0(\gamma)|\,.$$

10. Montrer que, pour X variable aléatoire à valeurs dans Γ,

$$\inf_{\phi}[P_0(\phi(X) = 1) + P_1(\phi(X) = 0)] = 1 - d_{\mathrm{TV}}(P_0, P_1),$$

où l'inf est pris sur toutes les fonctions mesurables $\phi : \Gamma \to \{0,1\}$. Donner une interprétation de ce résultat.

11. On suppose que $P_0(\gamma) > 0$ pour tout $\gamma \in \Gamma$. Pour tout $\gamma \in \Gamma$ on note

$$\ell(\gamma) := \frac{P_1(\gamma)}{P_0(\gamma)}\,.$$

Soit X une variable aléatoire à valeurs dans Γ. On note E_0 l'espérance relative à P_0. En utilisant la question 9, montrer que

$$2d_{\mathrm{TV}}(P_0, P_1) \leq \sqrt{E_0[\ell^2(X)] - 1}\,.$$

E – Un résultat négatif

Soit $\varepsilon > 0$. On suppose dans toute cette partie que $n \geq 3$ et que

$$k = k(n) \leq (1-\varepsilon)\frac{\log n}{\log\log n}\,.$$

12. Rappelons que X_k est définie dans la partie B. En écrivant

$$\mathbb{E}_0[X_k(G)^2] = M_1 + M_2 + M_3$$

avec

$$M_1 := \sum_{\mathcal{E}=(u,W)\in\mathcal{E}(n,k)} \sum_{\mathcal{E}'=(u,W')\in\mathcal{E}(n,k)} \mathbb{P}_0(\mathcal{E}\in G, \mathcal{E}'\in G),$$

$$M_2 := \sum_{\mathcal{E}=(u,W)\in\mathcal{E}(n,k)} \sum_{\substack{\mathcal{E}'=(u',W')\in\mathcal{E}(n,k)\\ u'\in W}} \mathbb{P}_0(\mathcal{E}\in G, \mathcal{E}'\in G),$$

et

$$M_3 := \sum_{\mathcal{E}=(u,W)\in\mathcal{E}(n,k)} \sum_{\substack{\mathcal{E}'=(u',W')\in\mathcal{E}(n,k)\\ u'\notin W}} \mathbb{P}_0(\mathcal{E}\in G, \mathcal{E}'\in G),$$

montrer que

$$\mathbb{E}_0[X_k(G)^2] = \mathbb{E}_0[X_k(G)]^2(1+o(1))\,,$$

où le $o(1)$ est pris lorsque $n \to +\infty$.

On montrera notamment que $M_1 = o(\mathbb{E}_0[X_k(G)]^2)$ et $M_2 = o(\mathbb{E}_0[X_k(G)]^2)$.

13. En déduire que la détection faible d'une $k-$étoile est impossible.

On pourra utiliser les questions 10 et 11.

14*. Avez-vous une idée d'application de ce problème ?

Grâce aux questions 8 et 13, nous avons établi que la difficulté du problème $(\star)$ change brutalement selon que $k \geq (1+\varepsilon)k^*(n)$ ou que $k \leq (1-\varepsilon)k^*(n)$, avec $k^*(n) = \frac{\log n}{\log\log n}$. Ce phénomène est communément appelé *transition de phase* et $k^*(n)$ est appelé *seuil critique*.

15*. A votre avis, que devient ce seuil critique $k^*(n)$ si on étudie le problème $(\star)$ cette fois-ci avec $p = p_0$ fixé constant dans $]0,1[$?

Fin du sujet.