
EXAM : Modèle linéaire

La durée de l'examen est 3 heures. Les calculatrices, téléphones portables et ordinateurs sont interdits. Pour chaque étudiant, une feuille A4 recto-verso est autorisée. Chaque réponse doit être justifiée. Des points seront attribués pour la présentation.

Questions générales

- 1) Soit $X \in \mathbb{R}^p$ un vecteur aléatoire de covariance $\Sigma \in \mathbb{R}^{p \times p}$. La matrice Σ admet-elle toujours une décomposition spectrale de la forme UDU^T où $D \in \mathbb{R}^p$ est diagonale? Si oui, on spécifiera les propriétés et dimension de U . On suppose que tous les éléments de D sont strictement positifs. A l'aide de U et D , donner A tel que $\text{cov}(AX) = I_p$.
- 2) Soit Y_1, Y_2, \dots, Y_n i.i.d. tel que $\mathbb{E}[Y_1^2] < \infty$. On cherche à estimer $\mu_0 = E[Y_1]$. Soit $(w_i)_{1 \leq i \leq n}$ un vecteur de poids positifs et déterministes. Donner $\hat{\mu}$ qui minimise $\sum_{i=1}^n w_i(Y_i - \mu)^2 + \lambda \mu^2$? Donner son biais et sa variance, pour tout $n > 1$. Pour quelle valeur de λ le biais est-il le plus petit? Pour quelle valeur de λ la variance est-elle la plus petite?
- 3) Quelle est la projection orthogonale du vecteur $\mathbf{y} = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ sur $\text{Vect}(\mathbf{1}_n)$, avec $\mathbf{1}_n = (1, \dots, 1)^T \in \mathbb{R}^n$? Peut-on retrouver ce projecteur à l'aide du problème d'optimisation de la question précédente?
- 4) Soit Y_1, Y_2, \dots, Y_n i.i.d. tel que $m_2 = \mathbb{E}[Y_1^2] < \infty$ et $\mathbb{E}[Y_1] = 0$. Soit $\hat{\mu} = n^{-1} \sum_{i=1}^n Y_i$. Exprimer $E[\hat{\mu}^2]$ en fonction de m_2 et n . En déduire un estimateur sans biais de m_2 .
- 5) La fonction $x \mapsto \max(|x| - 1, 0)$, définie sur \mathbb{R} , est-elle convexe? Est-elle concave? Admet-elle un argument minimum (argmin) ou maximum (argmax) unique?

Moindres carrés : $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ et $X = (\mathbf{1}_n, \tilde{X}) \in \mathbb{R}^{n \times (p+1)}$, $\mathbf{1}_n = (1, \dots, 1)^T \in \mathbb{R}^n$.

- 6) On suppose avoir calculé la matrice $X^T X \in \mathbb{R}^{(p+1) \times (p+1)}$. Que vaut le coefficient $(1, 1)$ de cette matrice? Expliquer comment retrouver, uniquement à l'aide des coefficients de cette matrice, la valeur de la variance empirique de la k -ième variable.
- 7) Soit $\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|_2^2$ et $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)^T = X\hat{\theta}$.

(a) Montrer que

$$\min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|_2^2 \leq \min_{\theta_0 \in \mathbb{R}} \|Y - \mathbf{1}_n \theta_0\|_2^2$$

(b) Montrer que

$$\arg \min_{\theta_0 \in \mathbb{R}} \|Y - \mathbf{1}_n \theta_0\|_2^2 = \bar{y} = n^{-1} \sum_{i=1}^n Y_i$$

(c) En déduire que le coefficient de détermination

$$R^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{y})^2} \leq 1$$

8) On suppose que X est de rang plein et que $\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|_2^2$. On suppose aussi le modèle dit de “fixed-design” suivant $Y = X\theta^* + \epsilon$ avec $\text{cov}(\epsilon) = I_n \sigma^2$ et X déterministe. Soit $(Y_{n+1}, x_{n+1}^T) \in \mathbb{R} \times \mathbb{R}^{p+1}$ des nouveaux points selon le modèle précédent indépendant de Y . Donner la valeur prédite \hat{Y}_{n+1} au point x_{n+1} par l’estimateur des moindres carrés. Exprimer la variance de l’erreur de prédiction, i.e., $\text{var}(Y_{n+1} - \hat{Y}_{n+1})$, en fonction de X , x_{n+1} et σ^2 (on prendra en compte le caractère aléatoire de Y_{n+1} et du vecteur Y).

9) Soit n un entier pair. Donner une formule explicite du problème $\arg \min_{\theta \in \mathbb{R}^{p+1}} \frac{1}{2} (Y - X\theta)^T \Omega (Y - X\theta)$ pour une matrice $\Omega = \text{diag}(w_1, \dots, w_n)$ telle que $w_i = 1$ si i est pair et 0 si i est impair. Donner une condition équivalente à l’unicité des solutions.

10) Donner une façon équivalente de calculer la solution des moindres carrés $\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^{p+1}} \|Y - X\theta\|_2^2$ en utilisant une regression sans intercept. On prendra soin de bien introduire toutes les notations nécessaires.

Tests, intervalle de confiance et Bootstrap. On suppose toujours $Y = (Y_1, \dots, Y_n)^T \in \mathbb{R}^n$ et $X = (1_n, \tilde{X}) \in \mathbb{R}^{n \times (p+1)}$. On note $\theta^* = (\theta_0^*, \dots, \theta_p^*)^T \in \mathbb{R}^{p+1}$.

11) Soit $(X_i)_{i=1, \dots, n}$ i.i.d. tel que $X_1 \sim \mathcal{N}(0, \sigma^2)$. On note $\hat{\mu} = n^{-1} \sum_{i=1}^n X_i$ et $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n X_i^2$. Donner la loi de $\hat{\mu}$, de $n\hat{\sigma}^2$ et de $\sqrt{n}\hat{\mu}/\hat{\sigma}$. On veut tester si la vraie moyenne de X_1 est 0 à partir de la statistique $\sqrt{n}\hat{\mu}/\hat{\sigma}$. Donner une région de rejet valide pour un test de niveau 95% en fonction des quantiles d’une loi connue.

12) Soient Z_1, \dots, Z_n des variables aléatoires i.i.d. d’espérance μ et de variance finie. Soit $\hat{R} = \sqrt{n}(\hat{\mu} - \mu)$ avec $\hat{\mu} = n^{-1} \sum_{i=1}^n Z_i$. Donner la loi limite de \hat{R} et préciser un intervalle de confiance pour μ valide asymptotiquement. Écrire un pseudo code permettant de construire un intervalle de confiance par bootstrap pour μ .

Ridge. A partir de maintenant $X \in \mathbb{R}^{n \times p}$. On note ici $\hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^p} \frac{1}{2} \|Y - X\theta\|_2^2 + \frac{\lambda}{2} \|\theta\|_2^2$ l’estimateur Ridge, où $\lambda > 0$.

13) Exprimer la variance de $\hat{\theta}_n$ et montrer que cette dernière est plus petite que la variance des moindres carrés. On prendra le temps d’introduire clairement les quantités impliquées.

14) On rappelle la SVD complète $X = USV^T$ où $U \in \mathbb{R}^{n \times n}$ et $V \in \mathbb{R}^{p \times p}$.

(a) Montrer que

$$(X^T X + \lambda I_p)^{-1} X^T = X^T (X X^T + \lambda I_n)^{-1}$$

(indication : on pourra utiliser la SVD complète de X ou exprimer le problème précédent en un problème équivalent plus simple)

- (b) Si $n \ll p$, donner une formule alternative pour calculer l'estimateur Ridge avec moins d'opérations. Comparer le nombre d'opérations nécessaires.

LASSO. Ici encore $X \in \mathbb{R}^{n \times p}$.

- 15) Soit $\Omega = \text{diag}(w_1, \dots, w_n)$ avec $w_i > 0$ pour tout i . Exprimer (en justifiant) l'algorithme de descente par coordonnée du problème suivant

$$\hat{\theta}_n = \arg \min_{\theta \in \mathbb{R}^p} (Y - X\theta)^T \Omega (Y - X\theta) + 2\lambda \|\theta\|_1$$

(indication : on exprimera le problème initial à l'aide d'un problème plus simple faisant intervenir la fonction $\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} (z - x)^2 + 2\lambda|x|$)

- 16) (a) Montrer que 0 est solution du problème du LASSO $\min_{\theta} \|Y - X\theta\|_2^2 + 2\lambda\|\theta\|_1$ si et seulement si

$$Y^T X\theta \leq \theta^T X^T X\theta/2 + \lambda\|\theta\|_1, \quad \forall \theta \in \mathbb{R}^p$$

- (b) Démontrer que pour tout vecteur $u = (u_1, \dots, u_K)^T$, $v = (v_1, \dots, v_K)^T$ on a $|u^T v| \leq \|u\|_1 \max_{1 \leq k \leq K} |v_k|$
- (c) Montrer que $Y^T X\theta \leq \lambda_{max} \|\theta\|_1$ avec $\lambda_{max} = \max_{k=1, \dots, p} |X_k^T Y|$
- (d) En déduire que si $\lambda \geq \lambda_{max}$ alors 0 est une solution du Lasso.
- (e) En déduire que si $\lambda \geq \lambda_{max}$, alors 0 est l'unique solution du Lasso. On pourra commencer par considérer $\theta \neq 0$ tel que

$$Y^T X\theta = \theta^T X^T X\theta/2 + \lambda\|\theta\|_1,$$

puis les 2 cas : $\theta \in \ker(X)$ et son contraire.