TELECOM Paris
2021/2022

MDI720
Anne Sabourin

EXAMEN FINAL : 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 sauf mention explicite du contraire. Des points seront attribués pour la présentation.

Questions générales

1) Soit A une matrice non aléatoire dans $\mathbb{R}^{q\times p}$, X un vecteur aléatoire dans $\mathbb{R}^p$ et μ un vecteur non aléatoire dans $\mathbb{R}^q$. Que vaut la matrice de covariance $\mathbb{C}\mathrm{ov}(AX+\mu)$ en fonction de $\mathbb{C}\mathrm{ov}(X)$?

2) Soit Σ une matrice de covariance de dimension n et soit $\lambda > 0$ un réel. Montrer que la matrice $S_\lambda = \Sigma + \lambda \mathrm{I}_n$ (I_n est la matrice identité) est toujours inversible et calculer son inverse en faisant apparaitre dans le résultat :
   - le réel λ, et
   - une matrice orthogonale V formée de vecteurs propres de Σ, et
   - une matrice diagonale L formée des valeurs propres de Σ.

3) Soient $Y_1, Y_2, \ldots, Y_n$ *i.i.d.* tels que $s^2 = \mathbb{E}[Y_1^2] < \infty$. On cherche à estimer $\mu_0 = E[Y_1] > 0$. On considères les deux estimateurs concurrents :
   - $\widehat{\mu}_1 = \frac{1}{n}\sum_1^n Y_i$
   - $\widehat{\mu}_2 = \frac{\lambda}{n}\sum_1^n Y_i$, avec $\lambda > 0$ fixé.

   On se demande quel estimateur choisir.
   (a) Donner le biais et la variance de chaque estimateur en fonction des grandeurs inconnues μ_0, s^2 et de λ.
   (b) Pour quelles valeurs de λ le risque quadratique de $\widehat{\mu}_2$ est-il inférieur à celui de $\widehat{\mu}_1$? (la réponse doit faire intervenir μ_0, s^2).

Moindres carrés Dans cette partie on note $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ et $Z = (\mathbf{1}_n, X) \in \mathbb{R}^{n\times(p+1)}$, avec $X \in \mathbb{R}^{n\times p}$ une matrice de prédicteurs (déterministe) et $\mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ le prédicteur constant. On suppose que $p < n$ et que Z est de rang $p+1$. On note $\widehat{\theta}$ l'estimateur des moindres carrés ordinaires, $\bar{Y}_n = \frac{1}{n}\sum Y_i \mathbf{1}_n$ et $\widehat{Y} = Z\widehat{\theta}$. On suppose aussi le modèle $Y = Z\theta^* + \epsilon$ avec $\mathbb{C}\mathrm{ov}(\epsilon) = \sigma^2 I_n$

4) Rappelez (sans justifier) l'interprétation de $\widehat{Y}$ comme une projection orthogonale d'un certain vecteur sur un certain espace, que l'on précisera. Exprimer (en justifiant) une relation d'égalité reliant les quantités $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$, $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$ et $\sum_{i=1}^n (\widehat{Y}_i - \bar{Y}_n)^2$.

5) On s'intéresse à l'erreur d'estimation d'un seul paramètre θ_j, $j \in \{1, \ldots, p\}$. En partant de l'expression explicite de $\widehat{\theta}$, qu'on ne cherchera pas à redémontrer, calculez le risque quadratique $\mathbb{E}\left((\widehat{\theta}_j - \theta_j^*)^2\right)$ où $\widehat{\theta}_j$ est la j^{eme} composante du vecteur $\widehat{\theta}$ estimé avec toutes les covariables.

6) Dans cette questions on considère que les colonnes Z_j de la matrice Z sont orthogonales deux à deux, $Z_j^\top Z_k = 0$ pour $j \neq k$. On considère l'estimateur alternatif $\widehat{\theta}'_j$ obtenu en utilisant seulement la j^{eme} variable pour l'estimation, c'est-à-dire en minimisant (par rapport à θ_j) l'erreur $\|Y - Z_j\theta_j\|_2^2$, où Z_j est la j^{eme} colonne de Z. Donnez l'expression de $\widehat{\theta}'_j$ et comparez-la à celle de $\widehat{\theta}_j$ (sous les mêmes hypothèses).

Tests, intervalle de confiance

7) Dans le cadre du modèle linéaire de la partie précédente, on suppose en plus que le vecteur des bruits ϵ est Gaussien de matrice de covariance $\sigma^2 \mathrm{I}_n$.
   (a) On s'intéresse à la variance des bruits σ^2. Rappelez l'expression de l'estimateur sans biais $\widehat{\sigma}^2$ vu en cours pour cette quantité. Quelle est la loi de $(n-p-1)\widehat{\sigma}^2/\sigma^2$? (pas de justification demandée).
   (b) En déduire un intervalle de confiance pour σ^2 construit à partir de $\widehat{\sigma}^2$ et des quantiles 0.025 et 0.975 de la loi de probabilité de la question (a). Quelle est le niveau de confiance de cet intervalle ?
8) On suppose maintenant que la variance des bruits σ^2 est connue : $\sigma^2 = 25$.
   (a) Rappelez (sans justifier) la loi suivie par l'estimateur des moindres carrés $\widehat{\theta}$.
   (b) On veut tester l'hypothèse nulle $H_0 : \theta_j = 0$ pour un certain $j \in \{0, \dots, p\}$ fixé. Déduire de (a) un test d'hypothèse au niveau 95%. On précisera en particulier la statistique de test et la région de rejet associée en fonction des quantiles d'une loi classique (on ne demande pas de donner la valeur numérique des quantiles). On choisira une région de rejet centrée autour de 0.

Ridge et Lasso. On considère dorénavant le modèle linéaire $Y = X\theta^\star + \epsilon$ avec $\mathbb{C}\mathrm{ov}(\epsilon) = \sigma^2\mathrm{I}_n$, $X \in \mathbb{R}^{n\times p}$ la matrice des covariables et $\theta^\star \in \mathbb{R}^p$ le paramètre inconnu. Ce modèle sans intercept est justifié après recentrage des covariables et de la cible.

9) On considère l'estimateur ridge $\widehat{\theta}^{\mathrm{r}}_\lambda = \mathrm{argmin}_{\theta\in\mathbb{R}^p} \left\{\|Y - X\theta\|^2 + \lambda\|\theta\|_2^2\right\}$. Etablir l'expression explicite de $\widehat{\theta}^{\mathrm{r}}_\lambda$ en fonction des données du problème (Y, X, λ).
10) Donnez un inconvénient et deux bonnes raisons d'utiliser l'estimateur $\widehat{\theta}^{\mathrm{r}}_\lambda$ avec un λ strictement positif, en justifiant à partir de notions de biais, de variance et de stabilité numérique. (on ne demande pas de faire les calculs pour justifier la réponse, 2 lignes maximum par inconvénient ou avantage)
11) Soit $y \in \mathbb{R}$ et $\alpha > 0$ un réel. Calculer la solution du problème de minimisation dans $\mathbb{R}$:

$$\mathrm{argmin}_{z\in\mathbb{R}}(y-z)^2 + 2\alpha|z|,$$

autrement dit déterminer le minimiseur $\widehat{z}$ en distinguant les cas $y > \alpha$, $y < -\alpha$ et $-\alpha < y < \alpha$. Que se passe -t-il lorsque $\alpha \to \infty$ alors que y est fixe ?
12) On s'intéresse à l'algorithme de descente par coordonnées pour le calcul de la solution de l'estimateur Lasso. c'est une méthode itérative de recherche de la solution. On fixe tour à tour $p-1$ coordonnées de θ et on optimise la coordonnée restante. Supposons qu'on optimise la p^{eme} coordonnée à l'itération t : on fixe $\theta_1^t, \dots, \theta_{p-1}^t$ et on considère

$$\widehat{\theta}^{t+1}_{\lambda,p} = \mathrm{argmin}_{\theta_p\in\mathbb{R}} \|Y - \sum_{j=1}^{p-1}\theta_j^t X_j - \theta_p X_p\|_2^2 + 2\lambda\Big(\sum_{j=1}^{p-1}|\theta_j^t| + |\theta_p|\Big),$$

où X_j est la j^{eme} colonne de la matrice X.

Montrez que ce problème se réécrit sous la forme du problème de la question (11) avec un certain choix de α et de z à déterminer en fonction de Y, X et $(\theta_1^t, \ldots, \theta_p^t)$. On introduira le résidu partiel :

$$r_p^t = Y - \sum_{j=1}^{p-1} \theta_j^t X_j.$$

En déduire la solution $\widehat{\theta}_p^{t+1}$ cherchée.