Apprentissage statistique, estimation de densité et prédiction linéaire

par :

Jaouad Mourtada[1] *— ENSAE/CREST*

1 Introduction

Dans ce texte, nous considérons des problèmes de prévision statistique (on parle aussi d'apprentissage supervisé). L'objectif consiste à prédire une certaine quantité d'intérêt appelée *réponse*, à partir de *variables* connues. Il y a bien sûr plusieurs façons d'appréhender ce problème, en fonction de la nature des variables et de la réponse, ainsi que du mécanisme qui les lie et des informations disponibles a priori. Par exemple, lorsque la réponse dépend des variables par un mécanisme connu, comme dans le cas de phénomènes naturels décrits par des lois précises, il est possible d'effectuer des prédictions à partir de règles définies a priori. En statistiques, la construction de bonnes règles de prédiction se fonde sur l'exploitation de jeux de données contenant de nombreux exemples, c'est-à-dire des observations de variables et de réponses associées. Il s'agit alors d'identifier des corrélations entre les variables et les réponses au sein des observations disponibles, afin d'effectuer de bonnes prédictions sur de nouvelles données.

Le problème de la prévision est intimement lié à celui de l'estimation (qui consiste à approcher des quantités moyennes ou globales associées à une population, à partir de l'observation d'échantillons issus de cette population), qui est au cœur de la théorie statistique. L'apprentissage statistique a en particulier émergé comme un sujet d'étude à part entière avec les travaux fondateurs de Vapnik et Chervonenkis [48] sur la classification au début des années 1970. Dans ce texte, nous aborderons certaines variantes de ce problème, en mettant l'accent sur les classes linéaires de prédicteurs, ainsi que sur le problème de l'estimation de densité.

1. jaouad.mourtada@ensae.fr

En Section 2, nous introduisons le problème de l'apprentissage statistique, en discutant quelques notions générales illustrées dans les sections suivantes. Le cas de la régression linéaire est examiné en détail en Section 3. Nous verrons en particulier que la difficulté du problème est caractérisée par une certaine quantité, le levier statistique, qui quantifie la sensibilité des prédictions. Cela permettra de mettre en évidence une propriété d'extrémalité asymptotique de la loi normale en grande dimension. Nous discuterons également l'obtention de bornes supérieures, qui se ramène à l'étude de la queue inférieure et des moments inférieurs de matrices aléatoires. Enfin, la Section 4 est consacrée à un autre problème d'apprentissage statistique, à savoir l'estimation de densité conditionnelle. Après avoir relevé des faiblesses de l'estimation par maximum de vraisemblance dans le cas où la loi des données s'écarte du modèle, nous décrivons un estimateur alternatif, qui corrige les prédictions de l'estimateur de vraisemblance en fonction d'une forme de levier à partir d'échantillons « fictifs ». Nous appliquons cet estimateur au modèle linéaire gaussien puis au modèle logistique, pour lesquels l'estimateur admet de meilleures garanties théoriques que l'estimateur du maximum de vraisemblance, tout en restant explicitement calculable par optimisation convexe.

2 Apprentissage statistique et risque minimax

En apprentissage statistique supervisé, l'objectif est de trouver une bonne façon de prédire une réponse $y \in \mathcal{Y}$ à partir d'une variable $x \in \mathcal{X}$. Ainsi, soient $\mathcal{X}, \mathcal{Y}$ et $\widehat{\mathcal{Y}}$ des espaces mesurables, correspondant respectivement à l'espace des variables, des réponses et des prédictions[2]. Afin de quantifier la qualité d'une prédiction en fonction de la valeur de la réponse, on se donne une *fonction de perte* $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbf{R}$; ainsi, $\ell(\hat{y}, y)$ mesure l'« erreur » de la prédiction $\hat{y} \in \widehat{\mathcal{Y}}$ lorsque la réponse est $y \in \mathcal{Y}$. Un *prédicteur* (ou : *fonction de prédiction*) est simplement une fonction (mesurable) $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ associant une prédiction $f(x)$ à chaque réalisation de la variable $x \in \mathcal{X}$.

Afin de donner un sens au problème, il reste à spécifier la relation entre la variable x et la réponse y. Dans le cadre statistique, on suppose que le couple variable-réponse est la réalisation d'une variable aléatoire (X, Y) de loi jointe P sur $\mathcal{X} \times \mathcal{Y}$. On mesure alors la qualité d'un prédicteur $f : \mathcal{X} \to \widehat{\mathcal{Y}}$ par son *risque*

2. On a souvent $\widehat{\mathcal{Y}} = \mathcal{Y}$, comme dans le cas de la régression traité en Section 3, mais pour l'estimation de densité en Section 4, $\widehat{\mathcal{Y}}$ sera distinct de $\mathcal{Y}$.

(perte moyenne)

$$L(f) = L_P(f) = \mathbb{E}\ell(f(X), Y) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) P(\mathrm{d}\, x, \mathrm{d}\, y)\,. \tag{1}$$

Le prédicteur ayant le plus faible risque, appelé *prédicteur de Bayes*, est caractérisé par la loi conditionnelle $P_{Y|X}$: presque sûrement,

$$f_{\text{Bayes}}(X) = \operatorname*{arg\,min}_{\hat{y}\in\widehat{\mathcal{Y}}} \mathbb{E}[\ell(\hat{y}, Y)|X] = \operatorname*{arg\,min}_{\hat{y}\in\widehat{\mathcal{Y}}} \int_{\mathcal{Y}} \ell(\hat{y}, y) P(\mathrm{d}\, y|X)\,. \tag{2}$$

Dans le problème d'apprentissage, la loi P des données, et donc le risque L et le prédicteur de Bayes, sont inconnus. On dispose en revanche d'un échantillon i.i.d. de loi P, soit $D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$. L'objectif consiste alors, étant donné D_n, à choisir un bon prédicteur $\hat{f}_n = f_n(D_n)$, où f_n est une fonction de l'ensemble $(\mathcal{X} \times \mathcal{Y})^n$ de jeux de données de taille n vers l'ensemble $\mathcal{F}(\mathcal{X}, \widehat{\mathcal{Y}}) = \widehat{\mathcal{Y}}^{\mathcal{X}}$ des prédicteurs. Notons que le prédicteur $\hat{f}_n$, et donc son risque $L(\hat{f}_n)$, est aléatoire, car dépendant de l'échantillon D_n. Nous considérerons ici surtout l'espérance $\mathbb{E}[L(\hat{f}_n)]$ du risque, mais il est aussi important de contrôler les queues de $L(\hat{f}_n)$, c'est-à-dire d'obtenir des bornes de risque en forte probabilité.

Une façon classique [47, 16, 10] d'évaluer la performance d'une procédure de prédiction $\hat{f}_n$ consiste à comparer son risque à celui du meilleur prédicteur au sein d'une classe de référence $\mathcal{F} \subset \mathcal{F}(\mathcal{X}, \widehat{\mathcal{Y}})$, uniformément sur une famille $\mathcal{P}$ de lois de probabilités P possibles. Formellement, on définit l'*excès de risque* du prédicteur $\hat{f}_n$ par rapport à la classe $\mathcal{F}$ sous la loi P par :

$$\mathcal{E}_P(\hat{f}_n; \mathcal{F}) = \mathbb{E}_P\big[L_P(\hat{f}_n)\big] - \inf_{f\in\mathcal{F}} L_P(f)\,. \tag{3}$$

L'excès de risque maximal d'une règle $\hat{f}_n$ par rapport à la classe $\mathcal{F}$ sur la famille de lois $\mathcal{P}$ est donc $\mathcal{E}(\hat{f}_n; \mathcal{F}, \mathcal{P}) = \sup_{P\in\mathcal{P}} \mathcal{E}_P(\hat{f}_n; \mathcal{F})$. Enfin, étant données une fonction de perte $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \longrightarrow \mathbf{R}$, une classe $\mathcal{F}$ de prédicteurs $\mathcal{X} \longrightarrow \widehat{\mathcal{Y}}$, une famille $\mathcal{P}$ de lois sur $\mathcal{X} \times \mathcal{Y}$, et une taille d'échantillon n, la difficulté du problème de prédiction/d'apprentissage défini par $(\ell, \mathcal{F}, \mathcal{P}, n)$ peut être mesurée par l'*(excès de) risque minimax*, c'est-à-dire l'excès de risque maximal de la meilleure procédure $\hat{f}_n$ possible :

$$\mathcal{E}^*_n(\ell, \mathcal{F}, \mathcal{P}) = \inf_{\hat{f}_n} \sup_{P\in\mathcal{P}} \mathcal{E}_P(\hat{f}_n; \mathcal{F})\,. \tag{4}$$

Cette définition appelle quelques commentaires. L'approche minimax est bien établie en statistiques au sein de la théorie de la décision. Il est important de noter que le risque minimax conduit à considérer des garanties uniformes

(sur une classe $\mathcal{P}$ de lois de probabilités), à un nombre d'échantillons n fixé. Ceci contraste avec une approche ponctuelle et asymptotique, faisant tendre n vers l'infini tout en fixant la loi P. L'avantage de l'approche minimax est son uniformité (qui assure un risque contrôlé dans le pire des cas), mais aussi le fait qu'elle fournit une notion d'optimalité et de « meilleure garantie possible », et permet de quantifier celle-ci en fonction des paramètres du problème (soit n, $\mathcal{P}$ et $\mathcal{F}$). L'inconvénient de cette approche est son caractère pessimiste (qui conduit à considérer la pire loi $P \in \mathcal{P}$), mais cette limitation peut être partiellement levée en considérant la question de l'*adaptation*, c'est-à-dire en envisageant simultanément plusieurs classes $\mathcal{P}$ plus ou moins complexes. Pour plus de détails sur cette approche, en particulier dans le cadre de l'estimation non paramétrique, nous renvoyons à l'ouvrage [45].

À ce stade, il peut être tentant de prendre pour $\mathcal{P}$ l'ensemble $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ de toutes les lois de probabilités sur $\mathcal{X} \times \mathcal{Y}$ (ou un sous-ensemble très riche et peu restrictif), et pour $\mathcal{F}$ l'ensemble $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ de tous les prédicteurs $\mathcal{X} \longrightarrow \mathcal{Y}$. Malheureusement, si $\mathcal{X}$ est infini (par exemple $\mathcal{X} = [0,1]$), il n'est pas possible d'obtenir des garanties non triviales avec ce choix : l'excès de risque minimax (4) est alors minoré par une constante indépendante de n (voir par exemple [16, Chapitre 7]). Cela tient au fait que, sans hypothèse additionnelle, la meilleure fonction $f^* : \mathcal{X} \longrightarrow \mathcal{Y}$ peut être arbitrairement complexe, de sorte qu'un échantillon de taille n ne contient que peu d'information sur les valeurs de f^* sur l'entièreté du domaine $\mathcal{X}$.

De cette obstruction, il découle qu'afin d'obtenir des garanties informatives, il est nécessaire de restreindre soit la classe $\mathcal{P}$ des lois de probabilités considérées, soit la classe $\mathcal{F}$ de prédicteurs de comparaison. Le premier choix correspond à une approche de modélisation statistique, qui consiste à supposer que la loi P appartient à un ensemble de lois restreint ou structuré (par exemple, dépendant d'un nombre de paramètres petit devant le nombre d'échantillons n). Cette approche est classique et utile, mais elle ne donne des garanties que lorsque les hypothèses sont satisfaites. Une autre approche, adoptée notamment par Vapnik et Chervonenkis [48], consiste à faire peu d'hypothèses contraignantes sur la loi P des données, mais à restreindre la classe de référence $\mathcal{F}$. Cette approche est plus générale et moins restrictive que la première, et peut-être aussi plus facile à justifier : en effet, le statisticien ne contrôle pas la vraie loi P des données, mais il peut choisir la classe $\mathcal{F}$ de référence. Ces deux approches ne sont toutefois pas incompatibles, et il est courant de restreindre à la fois $\mathcal{F}$ et $\mathcal{P}$ à des degrés divers. De plus, nous verrons dans les sections suivantes que la difficulté du problème dépend souvent peu de l'approche choisie. En revanche, le cas de l'estimation de densité en Section 4 illustrera le fait qu'il est parfois nécessaire de considérer

des procédures différentes dans le cadre de la seconde approche afin d'obtenir des garanties optimales.

3 Régression linéaire et queue inférieure de matrices aléatoires

Dans cette section, nous considérons un cas particulier du problème de prédiction défini en Section 2, à savoir la régression linéaire avec design aléatoire. Cela conduit à étudier certaines propriétés de matrices aléatoires. Sauf mention explicite du contraire, les résultats de cette section sont issus de l'article [39].

Nous considérons ici le cas de la *perte carrée* : $\widehat{\mathcal{Y}} = \mathcal{Y} = \mathbf{R}$, et $\ell_{\mathrm{sq}} : \mathbf{R} \times \mathbf{R} \longrightarrow \mathbf{R}$ est définie par $\ell_{\mathrm{sq}}(\hat{y}, y) = (\hat{y} - y)^2$. En outre, la classe de référence $\mathcal{F}$ est supposée être un espace vectoriel de fonctions $\mathcal{X} \longrightarrow \mathbf{R}$ de dimension $d \geq 1$. Quitte à effectuer un changement de variables, on peut supposer que $\mathcal{X} = \mathbf{R}^d$ et que $\mathcal{F}$ est la classe $\mathcal{F}_{\mathrm{lin}} = \{f_\theta : \theta \in \mathbf{R}^d\}$ des fonctions linéaires $f_\theta(x) = \langle \theta, x \rangle$ (où $\langle \cdot, \cdot \rangle$ désigne le produit scalaire usuel sur $\mathbf{R}^d$). On suppose que $\mathbb{E}Y^2 < +\infty$ et $\mathbb{E}\|X\|^2 < +\infty$, auquel cas pour tout $\theta \in \mathbf{R}^d$, le risque

$$L(f_\theta) = \mathbb{E}(\langle \theta, X \rangle - Y)^2$$

est fini. De plus, soit $\Sigma = \mathbb{E}XX^\top$ la matrice de covariance de X. Quitte à se restreindre à un sous-espace de $\mathbf{R}^d$, on peut supposer que Σ est inversible, auquel cas il existe un unique paramètre $\theta^* \in \mathbf{R}^d$ minimisant le risque $L(f_\theta)$, donné par $\theta^* = \Sigma^{-1}\mathbb{E}[YX]$. On définit l'*erreur* $\varepsilon = Y - \langle \theta^*, X \rangle$, de sorte que $\mathbb{E}\varepsilon X = 0$. Enfin, pour tout $\theta \in \mathbf{R}^d$,

$$L(f_\theta) - \inf_{f \in \mathcal{F}_{\mathrm{lin}}} L(f) = \|\Sigma^{1/2}(\theta - \theta^*)\|^2 = \|\theta - \theta^*\|_\Sigma^2 .$$

Si la régression linéaire est un problème classique, des progrès récents en concentration de la mesure et en analyse non-asymptotique de matrices aléatoires [43, 32, 1, 44, 33, 41] permettent un traitement non-asymptotique fin [23, 46, 4, 7, 27, 41, 34, 8, 39].

Risque minimax, loi des leviers et bornes inférieures. Nous considérons ici le risque minimax pour la régression linéaire, en s'intéressant en particulier à sa dépendance en la loi P_X de la variable X dans $\mathbf{R}^d$. Afin de simplifier les énoncés et d'obtenir des résultats plus précis, nous considérons le cas *bien spécifié*. On pose alors, pour toute loi P_X sur $\mathbf{R}^d$ telle que $\Sigma = \mathbb{E}XX^\top$ existe et est inversible

et tout $\sigma > 0$,

$$\mathcal{P}(P_X, \sigma^2) = \left\{ P_{(X,Y)} \,:\, Y = \langle \theta^*, X \rangle + \varepsilon,\ \theta^* \in \mathbf{R}^d,\ \mathbb{E}[\varepsilon | X] = 0,\ \mathbb{E}[\varepsilon^2 | X] \le \sigma^2 \right\}.$$

Le terme *bien spécifié* renvoie ici au fait que, sous $\mathcal{P}(P_X, \sigma^2)$, la meilleure fonction de prédiction $f_{\mathrm{Bayes}}(x) = \mathbb{E}[Y|X = x]$ (cas particulier de (2)) est linéaire.

Un estimateur classique est celui des *moindres carrés*, qui minimise l'erreur sur le jeu de données D_n : en notant

$$\widehat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top \tag{5}$$

la *matrice de covariance empirique*, et en supposant cette matrice inversible, on a

$$\hat{\theta}_n^{\mathrm{mc}} = \operatorname*{arg\,min}_{\theta \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \theta, X_i \rangle)^2 = \widehat{\Sigma}_n^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i \,. \tag{6}$$

Définition 1

La loi P_X sur $\mathbf{R}^d$ est dite *non dégénérée* si, pour tout hyperplan $H \subset \mathbf{R}^d$, on a $P_X(H) = \mathbb{P}(X \in H) = 0$. Pour tout $n \ge d$, cela revient à dire que $\widehat{\Sigma}_n$ est inversible p.s., c'est-à-dire que l'estimateur des moindres carrés $\hat{\theta}_n^{\mathrm{mc}}$ est bien défini p.s.

Proposition 1. *Si $n < d$ ou si P_X est dégénérée, alors l'excès de risque minimax $\mathcal{E}_n^*(\ell_{\mathrm{sq}}, \mathcal{F}_{\mathrm{lin}}, \mathcal{P}(P_X, \sigma^2))$ est infini. Sinon, le risque minimax vaut :*

$$\mathcal{E}_n^*(P_X, \sigma^2) = \mathcal{E}_n^*(\ell_{\mathrm{sq}}, \mathcal{F}_{\mathrm{lin}}, \mathcal{P}(P_X, \sigma^2)) = \frac{\sigma^2 \cdot \mathbb{E}\,\mathrm{Tr}(\Sigma^{1/2} \widehat{\Sigma}_n^{-1} \Sigma^{1/2})}{n} \,. \tag{7}$$

De plus, le risque minimax est atteint par l'estimateur des moindres carrés $\hat{\theta}_n^{\mathrm{mc}}$.

La Proposition 1 se démontre par des techniques classiques (minoration du risque minimax par le risque bayésien et convergence dominée). Tout d'abord, notons que le risque minimax (7) est invariant par transformation linéaire de X (ce qui est attendu par invariance du problème), nous supposons donc à partir de maintenant que X est isotrope : $\Sigma = \mathbb{E} X X^\top = I_d$. Le fait que le risque minimax soit infini lorsque X est non dégénéré ou $n < d$ provient d'une obstruction élémentaire : sous ces conditions, avec probabilité positive, les observations $X_1, \dots, X_n$ n'engendrent pas tout l'espace $\mathbf{R}^d$. Dans ce cas, il est impossible d'estimer la composante du paramètre orthogonale à l'espace engendrée par les X_i ; en revanche, cette composante peut-être arbitrairement grande, de sorte que l'on

commet ainsi une erreur qui peut être arbitrairement élevée[3]. Notons enfin que, même lorsque l'excès de risque minimax (sur la classe linéaire) est infini, en particulier lorsque $n < d$, il est possible d'obtenir un faible risque minimax sur des sous-ensembles de $\mathbf{R}^d$ comme les boules ℓ^p, sous certaines conditions.

Le risque minimax pouvant être infini pour certaines lois P_X, il est naturel de se demander à quel point il peut être faible, et pour quelles lois P_X. La Proposition 1 donne une première borne inférieure qualitative : le risque minimax est infini si $n < d$. Par ailleurs, il est possible d'obtenir une borne inférieure quantitative pour $n \geq d$, valable pour toute loi de X, à partir de (7) : par convexité de l'application $A \mapsto \mathrm{Tr}(A^{-1})$ sur le cône des matrices positives, et comme $\mathbb{E}[\widehat{\Sigma}_n] = I_d$,

$$\mathcal{E}_n^*(P_X, \sigma^2) = \frac{\sigma^2 \cdot \mathbb{E}\,\mathrm{Tr}(\widehat{\Sigma}_n^{-1})}{n} \geq \frac{\sigma^2 \cdot \mathrm{Tr}(\mathbb{E}[\widehat{\Sigma}_n]^{-1})}{n} = \frac{\sigma^2 d}{n}\,.$$

On peut comparer cette borne inférieure au risque minimax pour la loi gaussienne $P_X = \mathcal{N}(0, I_d)$; dans ce cas, la matrice $\widehat{\Sigma}_n^{-1}$ suit une loi de Wishart inverse, dont l'espérance est connue. On a alors (par exemple, [5]) :

$$\mathcal{E}_n^*(\mathcal{N}(0, I_d), \sigma^2) = \frac{\sigma^2 d}{n - d - 1}\,. \tag{8}$$

Ainsi, la borne inférieure de $\sigma^2 d/n$ est dans ce cas du bon ordre de grandeur lorsque $n \gtrsim d$, et même équivalente au risque minimax lorsque $d/n \longrightarrow 0$. En revanche, lorsque $d, n \longrightarrow \infty$ avec $d/n \longrightarrow \gamma \in (0, 1)$ (on parle parfois de régime asymptotique de grande dimension, voir par exemple [18]), les deux valeurs limites ne coïncident pas. Il est en fait possible de raffiner la borne inférieure précédente, en utilisant le fait que $\widehat{\Sigma}_n$ est une somme de matrices i.i.d. de rang 1.

Afin d'énoncer le résultat, nous avons besoin d'une définition supplémentaire. Étant donnés $X_1, \dots, X_n$ (tels que $\widehat{\Sigma}_n$ est inversible) et $x \in \mathbf{R}^d$, on note

$$h_n(x) = \left\langle \left(\sum_{i=1}^n X_i X_i^\top + x x^\top \right)^{-1} x, x \right\rangle \in [0, 1) \tag{9}$$

le *levier* de x au sein du jeu de données $X_1, \dots, X_n, x$. Le levier de x quantifie l'influence de la réponse y associée à x sur la prédiction de l'estimateur des moindres carrés : plus précisément, en notant $\hat{\theta}_n^{\mathrm{mc}}(x, y)$ l'estimateur des moindres carrés sur le jeu de données $D_n \cup \{(x, y)\}$, la prédiction $\langle \hat{\theta}_n^{\mathrm{mc}}(x, y), x \rangle$ en x est une fonction affine de y dont la pente vaut $h_n(x)$. Enfin, on note $h_{n+1} = h_n(X_{n+1})$ où X_{n+1} est un échantillon indépendant de loi P_X.

3. Cette intuition ne suffit pas tout à fait à conclure rigoureusement, car l'espace engendré par les X_i, à l'inverse de θ^*, est aléatoire.

Proposition 2. *Si $n \geq d$, alors pour toute loi P_X non dégénérée, on a :*

$$\mathcal{E}_n^*(P_X, \sigma^2) = \mathbb{E}\Big[\frac{h_{n+1}}{1-h_{n+1}}\Big] \geq \frac{\sigma^2 d}{n-d+1}. \tag{10}$$

La Proposition 2 indique que le risque minimax est déterminé par la loi des leviers h_{n+1} (dans un échantillon de taille $n+1$ tiré selon P_X). Intuitivement, plus les leviers sont « dispersés » ou inhomogènes, plus le risque minimax est élevé (par convexité de la fonction $x \mapsto x/(1-x)$). Cela correspond au fait que la difficulté de prédire en un point x est caractérisée par le levier $h_n(x)$, la présence de points ayant un fort levier implique donc que le problème est intrinsèquement plus difficile.

La borne inférieure (10), valable pour toute loi P_X, est proche du risque (8) obtenu dans le cas gaussien. En particulier, dans le régime $n, d \longrightarrow \infty$ avec $d/n \longrightarrow \gamma \in (0,1)$, les deux quantités convergent vers le même nombre $\sigma^2\gamma/(1-\gamma)$. Ainsi, la loi gaussienne est asymptotiquement la loi la plus favorable de X en grande dimension. De plus, la caractérisation de la Proposition 2 permet d'expliquer cela : si pour une suite de lois $(P_X^{(n)})$ sur $\mathbf{R}^{d_n}$ avec $d_n \longrightarrow \gamma$, le risque minimax converge la valeur minimale $\sigma^2\gamma/(1-\gamma)$, alors la loi du levier h_{n+1} d'un échantillon converge vers une constante (égale à γ). C'est donc le cas de la loi gaussienne, pour laquelle tous les points ont asymptotiquement le même levier γ.

Il serait intéressant de déterminer la loi extrémale P_X minimisant le risque minimax $\mathcal{E}_n^*(P_X, \sigma^2)$, de manière non-asymptotique avec $n \geq d \geq 1$ fixés. Il n'est pas difficile de montrer, à partir de la formule (7) et de la convexité de $A \mapsto \mathrm{Tr}(A^{-1})$, que la loi uniforme sur la sphère $\sqrt{d}S^{d-1}$ (où S^{d-1} est la sphère de $\mathbf{R}^d$ pour la norme ℓ^2) minimise $\mathcal{E}_n^*(P_X, \sigma^2)$ au sein des lois invariantes par rotation (qui contient la loi gaussienne). Cependant, il n'est pas trivial de démontrer qu'une loi P_X extrémale est invariante par rotation ; il suffirait pour cela (par un argument d'invariance) de montrer que la fonction $P_X \mapsto \mathcal{E}_n^*(P_X, \sigma^2)$ est convexe, mais cela conduit à vérifier des inégalités matricielles non triviales.

Notons pour finir qu'une version plus générale de la borne inférieure (10) est disponible [38, Chapitre 8]. Celle-ci affirme que, pour tout vecteur aléatoire X isotrope ($\mathbb{E}XX^\top = I_d$), tous $n \geq d$ et $\lambda > 0$, on a :

$$\frac{1}{d}\,\mathbb{E}\,\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\big\} \geq \frac{-(1-d+\lambda n) + \sqrt{(1-d+\lambda n)^2 + 4\lambda d n}}{2\lambda d}. \tag{11}$$

De plus, cette inégalité est asymptotiquement fine : si $d, n \longrightarrow \infty$ avec $d/n \longrightarrow \gamma \in (0,1)$, alors les deux termes de l'inégalité (11) convergent vers la même fonction de λ, γ. Il s'agit d'une conséquence de la loi de Marchenko-Pastur [36] décrivant le spectre asymptotique de matrices de covariances de vecteurs gaussiens.

La borne inférieure (11) affirme que la loi de Marchenko-Pastur, bien que non valide pour des vecteurs aléatoires X isotropes quelconques, fournit une borne inférieure dans le cas général. Notons enfin que le membre de gauche de (11) est aussi relié à une quantité statistique pour la régression linéaire, à savoir le risque optimal (bayésien) lorsque le paramètre θ^* est aléatoire et tiré selon la loi $\mathcal{N}(0, (\lambda n)^{-1} I_d)$.

Queue inférieure de matrices aléatoires et bornes de risque. L'expression (7) indique la dépendance du risque minimax en la loi de X, à travers la quantité $\mathbb{E}\,\mathrm{Tr}(\widehat{\Sigma}_n^{-1})$. En particulier, majorer le risque minimax revient à majorer $\mathbb{E}\,\mathrm{Tr}(\widehat{\Sigma}_n^{-1})$, donc à contrôler la queue *inférieure* de $\widehat{\Sigma}_n$. Il est important de relever qu'il n'est pas nécessaire de contrôler la queue supérieure de $\widehat{\Sigma}_n$, par exemple en majorant la norme d'opérateur $\|\widehat{\Sigma}_n - I_d\|_{\mathrm{op}}$ ou la plus grande valeur propre $\lambda_{\max}(\widehat{\Sigma}_n)$, mais seulement la trace $\mathrm{Tr}(\widehat{\Sigma}_n^{-1})$ et la plus petite valeur propre $\lambda_{\min}(\widehat{\Sigma}_n)$. Cette distinction est importante, car il est possible de contrôler la queue inférieure d'une matrice aléatoire sous des hypothèses bien plus faibles que la queue supérieure. Ce fait, établi dans une suite de travaux récents [44, 41, 33], tient au fait que la matrice $XX^\top$ est toujours positive, donc bornée à gauche, bien que des idées techniques supplémentaires soient nécessaires pour exploiter cette intuition.

Le résultat suivant, dû à Oliveira [41], permet d'illustrer ce fait. Soit X une variable aléatoire avec $\mathbb{E}XX^\top = I_d$, c'est-à-dire que $\mathbb{E}\langle\theta, X\rangle^2 = \|\theta\|^2$ pour tout $\theta \in \mathbf{R}^d$. On suppose que $\mathbb{E}\|X\|^4 < +\infty$, et l'on pose

$$\kappa_* = \sup\left\{\mathbb{E}\langle\theta, X\rangle^4 \,:\, \theta \in \mathbf{R}^d, \|\theta\| \le 1\right\}, \tag{12}$$

de sorte que $\|\langle\theta, X\rangle\|_{L^4} \le \kappa_*^{1/4} \|\langle\theta, X\rangle\|_{L^2}$ pour tout $\theta \in \mathbf{R}^d$. Par exemple, si $X \sim \mathcal{N}(0, I_d)$, on a $\kappa_* = 3$.

Théorème 1 : Oliveira [41]

Sous les hypothèses précédentes, on a pour tout $\delta > 0$:

$$\mathbb{P}\left(\lambda_{\min}(\widehat{\Sigma}_n) \le 1 - 9\kappa_*^{1/2}\sqrt{\frac{d + 2\log(2/\delta)}{n}}\right) \le \delta. \tag{13}$$

Il découle de résultats classiques sur les matrices aléatoires que cette borne est optimale (aux constantes près) dans le cas gaussien, pour $\delta \in (0, e^{-cn})$. En revanche, elle est valable sans hypothèses de moments fortes (seulement 4 moments), et donc aussi pour des variables à queues lourdes. Sous ces hypothèses,

on ne peut pas espérer un contrôle similaire de la queue supérieure de $\widehat{\Sigma}_n$, puisqu'il existe une loi telle que $\kappa_* \le c$ et $\mathbb{E}\|\widehat{\Sigma}_n - I_d\|_{\mathrm{op}} \gtrsim c^{-1} d/\sqrt{n}$ (prendre $X = UZ$ avec $Z \sim \mathcal{N}(0, I_d)$ et U scalaire indépendante de Z avec $\mathbb{E}U^2 = 1$ mais à queue lourde sous contrainte de moment d'ordre 4, et minorer $\mathbb{E}\|\widehat{\Sigma}_n - I_d\|_{\mathrm{op}} \ge \mathbb{E}\|\widehat{\Sigma}_n\|_{\mathrm{op}} - 1 \ge \mathbb{E}\max_{i\le n}\|X_i\|^2 - 1$).

Le risque minimax (7) fait intervenir $\mathbb{E}\,\mathrm{Tr}(\widehat{\Sigma}_n^{-1}) \ge \mathbb{E}\lambda_{\min}(\widehat{\Sigma}_n)^{-1}$, et donc des moments inférieurs de $\widehat{\Sigma}_n$. La borne (13) ne suffit pas à contrôler de telles quantités, car il est nécessaire de contrôler la queue inférieure

$$\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \le t) \tag{14}$$

pour tout $t \in (0, 1-C\sqrt{d/n})$. La borne (13) fournit un contrôle précis de la queue (14) pour $t \in (c, 1 - C\sqrt{d/n})$, mais pas pour $t \in (0, c)$ (où c, C sont des constantes absolues). Cela n'est pas surprenant, car le contrôle de $\lambda_{\min}(\widehat{\Sigma}_n)$ dans ce régime requiert des conditions différentes sur la loi de X; par exemple, la loi P_X uniforme sur $\{-1,1\}^d$ satisfait $\kappa_* = O(1)$, mais est dégénérée (au sens de la Définition 3) de sorte que $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) = 0) > 0$ et $\mathcal{E}_n^*(P_X, \sigma^2) = +\infty$. Afin d'obtenir un contrôle sur la queue (14) pour $t \in (0, c)$, une version quantitative de l'hypothèse 3 de non-dégénérescence est requise.

Hypothèse 1 (Petite boule)**.** Il existe des constantes $c > 0$ et $\alpha \in (0, 1]$ telles que, pour tout hyperplan $H \subset \mathbf{R}^d$ et tout $t \in (0, 1)$,

$$\mathbb{P}(\mathrm{dist}(X, H) \le t) \le (ct)^\alpha . \tag{15}$$

De manière équivalente, $\mathbb{P}(|\langle \theta, X\rangle| \le t) \le ct$ pour tout $\theta \in S^{d-1}$.

Le résultat suivant décrit la meilleure borne possible sur la queue inférieure (14) pour t petit, et montre que l'hypothèse 1 est nécessaire pour obtenir un tel contrôle.

Proposition 3. *Soit $d \ge 2$, et X un vecteur aléatoire dans $\mathbf{R}^d$ tel que $\mathbb{E}XX^\top = I_d$. Alors, pour tout $t \le 1$,*

$$\sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X\rangle| \le t) \ge 0.16 \cdot t,$$
$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \le t\big) \ge (0.025 \cdot t)^{n/2} .$$

Enfin, s'il existe des constantes $c_1, c_2 > 0$ telles que $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \le t) \le (c_1 t)^{c_2 n}$, alors X satisfait l'hypothèse 1 avec $c = \sqrt{c_1}$ et $\alpha = 2c_2$.

Il reste à montrer que l'hypothèse 1 est également suffisante pour obtenir la queue inférieure souhaitée. C'est ce qu'affirme l'énoncé suivant :

Théorème 2

Si X satisfait l'hypothèse 1 et si $n \geq 6d/\alpha$, alors en notant $C = 3c^4 e^{1+9/\alpha}$, on a pour tout $t \in (0,1)$

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leq t\big) \leq (Ct)^{\alpha n/6}. \tag{16}$$

La preuve de ce résultat est inspirée de celle du Théorème 3 par Oliveira [41], avec certaines différences spécifiques au cas où t est arbitrairement petit. Tout d'abord, $\lambda_{\min}(\widehat{\Sigma}_n)$ s'exprime naturellement comme l'infimum d'un processus stochastique :

$$\lambda_{\min}(\widehat{\Sigma}_n) = \inf_{\theta \in \mathbf{R}^d} \left\{ \langle \widehat{\Sigma}_n \theta, \theta \rangle = \frac{1}{n} \sum_{i=1}^{n} \langle \theta, X_i \rangle^2 \right\}. \tag{17}$$

Il serait alors naturel de raisonner de la façon suivante (qui fonctionne effectivement lorsque X possède des queues très légères, comparables à celles d'une loi gaussienne). Tout d'abord, on considère un sous-ensemble fini $A \subset S^{d-1}$ tel que tout élément de S^{d-1} est à distance au plus t de A ; on peut alors choisir A de sorte que $|A| \leq (3/t)^d$. À partir de l'hypothèse 1 et par des arguments classiques sur les sommes de variables indépendantes, on montre que pour tout $\theta \in A$, $\mathbb{P}(n^{-1} \sum_{i=1}^n \langle \theta, X_i \rangle^2 \leq 10t) \leq (c_1 t)^{c_2 n}$. Par une borne d'union sur A, on en déduit que $\mathbb{P}(\inf_{\theta \in A} n^{-1} \sum_{i=1}^n \langle \theta, X_i \rangle^2 \leq 10t) \leq (3/t)^d (c_1 t)^{c_2 n} \leq (c_3 t)^{c_4 n}$ pourvu que $n \gtrsim d$. Enfin, on étend la borne inférieure sur A à S^{d-1} par approximation.

Cet argument ne fonctionne pas ici, car la dernière étape d'approximation fait naturellement intervenir la plus grande valeur propre de $\widehat{\Sigma}_n$: en effet, étant donné $\theta \in S^{d-1}$ et $\theta' \in A$ tel que $\|\theta - \theta'\| \leq t$, on utilise que $|\langle \widehat{\Sigma}_n \theta, \theta \rangle - \langle \Sigma_n \theta', \theta' \rangle| \lesssim \lambda_{\max}(\widehat{\Sigma}_n) \cdot t$. Malheureusement, l'hypothèse 1 (qui n'implique l'existence d'aucun moment d'ordre $2+\varepsilon$) est bien trop faible pour obtenir un contrôle satisfaisant sur $\lambda_{\max}(\widehat{\Sigma}_n)$ pour $n \gtrsim d$, comme indiqué précédemment. Afin de contourner cette difficulté, l'approximation par une discrétisation finie A de S^{d-1} doit être remplacée par un autre argument. L'idée consiste à approcher, pour tout θ, la quantité $\langle \widehat{\Sigma}_n \theta, \theta \rangle$ par sa moyenne sur un petit voisinage de θ, à savoir une calotte sphérique centrée en θ. Il est alors possible de contrôler uniformément les moyennes locales par une technique spécifique, à savoir l'inégalité dite « PAC-Bayésienne » introduite par McAllester [37] dans le contexte de l'apprentissage statistique, et développée par Audibert et Catoni [9, 2] dans le cadre de l'estimation robuste. Le gain est que le terme d'approximation est une moyenne uniforme sur toutes les directions possibles (car la perturbation est centrée en θ et « isotrope »), et fait donc intervenir la quantité $\frac{1}{d} \operatorname{Tr}(\widehat{\Sigma}_n)$ au lieu de $\lambda_{\max}(\widehat{\Sigma}_n)$. La première quantité peut alors se contrôler facilement par un argument additionnel de troncature de X.

Ces résultats permettent d'obtenir des bornes pour la régression linéaire. Tout d'abord, afin d'obtenir des bornes fines, nous introduisons un autre paramètre de kurtosis que κ_* :

$$\tilde{\kappa} = \frac{\mathbb{E}\|X\|^4}{d^2}\,. \tag{18}$$

On a toujours $\tilde{\kappa} \le \kappa^*$, mais il est possible que $\tilde{\kappa} \ll \kappa^*$. En effet, si X est uniforme sur l'ensemble $\{\sqrt{d}e_j \,:\, 1 \le j \le d\}$ (où $(e_j)_{1\le j\le d}$ est la base usuelle de $\mathbf{R}^d$), on a $\tilde{\kappa} = 1$ tandis que $\kappa_* = d$.

Théorème 3

Pour toute loi P_X satisfaisant l'hypothèse 1 et telle que $\tilde{\kappa} < \infty$, en notant $C = 28c^4e^{1+9/\alpha}$ on a pour $n \ge \max(6d, 12\log(12\alpha^{-1}))/\alpha$:

$$\frac{\sigma^2 d}{n} \le \mathcal{E}_n^*(P_X, \sigma^2) \le \frac{\sigma^2 d}{n}\Big(1 + \frac{C\tilde{\kappa}\, d}{n}\Big)\,. \tag{19}$$

Dans le cas général mal spécifié (où $P \notin \mathcal{P}(P_X, \sigma^2)$), en notant $\theta^* = \arg\min_{\theta\in\mathbf{R}^d} L(f_\theta)$ et en supposant que

$$\chi = \mathbb{E}\big[(Y - \langle\theta^*, X\rangle)^4\|X\|^4\big]/d^2$$

est fini, on a pour $n \ge \max(96, 6d)/\alpha$,

$$\mathcal{E}_P(\hat{\theta}_n^{\mathrm{mc}}) \le \frac{1}{n}\mathbb{E}\big[(Y - \langle\theta^*, X\rangle)^2\|X\|^2\big] + 276C^2\sqrt{\kappa_*\chi}\Big(\frac{d}{n}\Big)^{3/2}\,. \tag{20}$$

Dans les équations (19) et (20), on vérifie par des calculs heuristiques asymptotiques (en faisant tendre $n \to \infty$ et en fixant P) que ces bornes sont fines dans ce régime, et décrivent donc bien le risque minimax et le risque dans le cas mal spécifié. Notons que la borne (19) dans le cas bien spécifié, qui repose sur une analyse directe, fait apparaître la constante de kurtosis $\tilde{\kappa}$, tandis que la borne (20) dans le cas mal spécifié, qui utilise le Théorème 3, fait apparaître la constante plus élevée κ_*.

4 Estimation de densité et régression logistique

Estimation de densité conditionnelle. Dans cette section, nous considérons un autre problème d'apprentissage statistique, à savoir l'*estimation de densité conditionnelle*. Dans ce problème, on cherche non pas (comme dans le cas de la

régression avec perte carrée) à effectuer une prédiction *ponctuelle* de la réponse (en cherchant à prédire une valeur proche), mais plutôt une prédiction *probabiliste* attribuant des probabilités aux différentes valeurs possibles de la réponse, en cherchant à attribuer la probabilité la plus élevée possible à celle-ci.

Formellement, soient $\mathcal{X}$ et $\mathcal{Y}$ deux ensembles (espaces mesurables), et μ une mesure de référence sur $\mathcal{Y}$. L'espace $\widehat{\mathcal{Y}}$ des prédictions (voir la Section 2) est ici l'ensemble des densités de probabilités sur $\mathcal{Y}$ par rapport à μ. On définit la *perte logarithmique* (aussi appelée *perte entropique* ou *logistique*) $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \longrightarrow \mathbf{R}$ par :

$$\ell(h, y) = -\log h(y) \qquad (h \in \widehat{\mathcal{Y}},\ y \in \mathcal{Y}). \tag{21}$$

Cette fonction de perte est couramment utilisée en théorie de l'information, car elle admet une interprétation en termes de codage et de compression de données [14, Chapitre 5]. En effet, si $\mathcal{Y}$ est un ensemble fini et h une mesure de probabilités sur $\mathcal{Y}$, il existe une façon de coder chaque élément $y \in \mathcal{Y}$ avec $\lceil -\log_2 h(y) \rceil$ bits, ce qui établit une correspondance entre les « codes » sur $\mathcal{Y}$ et les densités de probabilités. En ignorant la partie entière (ce qui change peu de choses dans le cas d'ensembles $\mathcal{Y}$ riches), la perte logarithmique (21) correspond donc à la longueur du code donné par h associé à $y \in \mathcal{Y}$.

Un prédicteur est ici une fonction $f : \mathcal{X} \longrightarrow \widehat{\mathcal{Y}}$ associant à tout $x \in \mathcal{X}$ une densité $f(x) \in \widehat{\mathcal{Y}}$ sur $\mathcal{Y}$, c'est-à-dire une *densité conditionnelle*. Pour $x \in \mathcal{X}$ et $y \in \mathcal{Y}$, on note $f(y|x) = f(x)(y)$, de sorte que la perte de la densité conditionnelle f sur l'échantillon $z = (x, y)$ s'écrit $\ell(f, z) = -\log f(y|x)$, et si $(X, Y) \sim P$, le risque de f vaut $L(f) = -\mathbb{E} \log f(Y|X)$. Le choix de la mesure de référence μ n'affecte le risque $L(f)$ que par une constante additive indépendante de f, de sorte que la différence

$$L(g) - L(f) = \mathbb{E} \log \left(\frac{f(Y|X)}{g(Y|X)} \right)$$

ne dépend pas du choix de μ. En notant $P_X^f = f \cdot (P_X \otimes \mu)$ la loi $f(y|x) P_X(\mathrm{d}\, x) \mu(\mathrm{d}\, y)$ sur $\mathcal{X} \times \mathcal{Y}$, et en supposant que la loi conditionnelle de Y sachant X admette une densité conditionnelle f^* par rapport à μ (*i.e.*, $P = P_X^{f^*}$), on a pour tout f,

$$L(f) - L(f^*) = \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{f^*(y|x)}{f(y|x)} \right) f^*(y|x) P_X(\mathrm{d}\, x) \mu(\mathrm{d}\, y) = \mathrm{KL}(P, P_X^f) \geq 0 \tag{22}$$

où $\mathrm{KL}(P, Q) = \int \log(\frac{\mathrm{d}\, P}{\mathrm{d}\, Q})\, \mathrm{d}\, P$ est la *divergence de Kullback-Leibler* (ou *entropie relative*) entre P et Q. Il en découle que le prédicteur de Bayes (2) f_{Bayes} n'est autre que la densité conditionnelle de Y sachant X, et que l'excès de risque est donné par la divergence de Kullback-Leibler.

Comme en Section 2, on se donne une classe $\mathcal{F}$ de prédicteurs (ici, de densités conditionnelles), c'est-à-dire un *modèle statistique* (conditionnel). Deux approches sont alors possibles. La première consiste à supposer que la vraie densité conditionnelle de Y sachant X appartient à $\mathcal{F}$; on dit alors que le modèle est *bien spécifié*. La seconde approche, moins restrictive, consiste à faire peu d'hypothèses sur la loi de $Y|X$ et à chercher une densité conditionnelle dont les performances prédictives sont proches de la meilleure densité de $\mathcal{F}$. On dit alors que le modèle est *mal spécifié*, et c'est à cette seconde configuration que nous nous intéresserons.

L'estimateur du maximum de vraisemblance. L'estimateur le plus classique est l'*estimateur du maximum de vraisemblance* (EMV), qui minimise l'erreur empirique :

$$\hat{f}_n^{\text{emv}} = \arg\min_{f\in\mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f, (X_i, Y_i)) \right\} = \arg\max_{f\in\mathcal{F}} \left\{ \prod_{i=1}^{n} f(Y_i|X_i) \right\}. \tag{23}$$

En d'autres termes, l'EMV choisit la densité qui maximise la probabilité/densité conditionnelle attribuée aux données (aux réponses).

Pour simplifier, l'EMV tend à avoir de bonnes performances lorsque le modèle $\mathcal{F}$ est bien spécifié et suffisamment peu complexe (par rapport à la taille n de l'échantillon). Par exemple, si $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$ est paramétrée de façon suffisamment régulière par un paramètre $\theta \in \Theta \subset \mathbf{R}^d$ et si $n \gg d$, l'EMV admet (sous certaines conditions) un excès de risque d'ordre $O(d/n)$, ce qui est optimal.

En revanche, même pour une classe $\mathcal{F}$ de la forme précédente, les performances de l'EMV peuvent se dégrader significativement dans le cas mal spécifié, et l'excès de risque peut être bien plus élevé que $O(d/n)$. De plus, cette sensibilité au caractère mal spécifié n'est pas spécifique à l'EMV, mais est souvent partagée par tout estimateur qui sélectionne un élément $\hat{f}_n$ de la classe de référence $\mathcal{F}$. Cette obstruction, essentiellement due à un défaut de convexité, est bien connue dans le cadre de l'agrégation de modèles [10, 28]. Nous verrons des exemples de ce phénomène dans le cas du modèle linéaire gaussien et du modèle logistique.

Agrégation par mélange bayésien. Une approche classique en estimation de densité (et en agrégation de modèles) est l'agrégation par mélange bayésien. L'idée consiste à « convexifier » le problème, en se plaçant sur l'espace des mesures de probabilités sur la classe $\mathcal{F}$. Plus précisément, en notant $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, on se donne une mesure de probabilités π sur l'espace Θ des paramètres, appelée *loi a priori*. On définit alors pour $i = 0, \dots, n$ le postérieur $\hat{\rho}_i$, qui est la loi

de probabilités sur Θ telle que

$$\frac{\mathrm{d}\,\hat{\rho}_i}{\mathrm{d}\,\pi}(\theta) = \frac{\prod_{j\le i} f_\theta(Y_j|X_j)}{\int_\Theta \prod_{j\le i} f_\vartheta(Y_j|X_j)\pi(\mathrm{d}\,\vartheta)}\,. \tag{24}$$

Le prédicteur agrégé $\hat{f}_n$, introduit et analysé par [3, 10, 50] (voir aussi [49, 35]), est donné par la moyenne des postérieurs prédictifs :

$$\hat{f}_n(y|x) = \frac{1}{n+1}\sum_{i=0}^{n}\int_\Theta f_\theta(y|x)\hat{\rho}_i(\mathrm{d}\,\theta)\,. \tag{25}$$

L'agrégation par mélange bayésien admet des garanties théoriques intéressantes : elle satisfait des bornes d'excès de risque qui ne se dégradent pas dans le cas mal spécifié. Dans le cas où f_θ dépend de manière suffisamment régulière de $\theta \in \Theta \subset \mathbf{R}^d$, en choisissant par exemple π uniforme sur Θ (en supposant Θ borné), on peut montrer que la procédure (25) satisfait une borne d'excès de risque en espérance de la forme $\mathcal{E}(\hat{f}_n) \lesssim d\log(n\cdot\mathrm{diam}(\Theta))/n$, où $\mathrm{diam}(\Theta)$ est le diamètre de Θ. Une telle garantie est remarquable, car valable sous des hypothèses faibles. En revanche, cette approche ne permet pas de traiter le cas où Θ est non borné, comme pour le modèle linéaire gaussien (voir plus bas) ; de plus, la borne contient un terme en $\log n$ supplémentaire, par rapport à une borne idéale en $O(d/n)$. Surtout, d'un point de vue computationnel, $\hat{f}_n$ requiert d'évaluer les postérieurs (24). Ainsi, le calcul approché du dénominateur de (24) se ramène à une tâche d'échantillonnage, ce qui est relativement coûteux.

Un estimateur alternatif. Nous décrivons une autre procédure pour l'estimation de densité conditionnelle, introduite dans [40] dont sont issus les résultats suivants. Cette procédure minimise une borne générale de risque fondée sur des arguments d'échangeabilité et de stabilité. Ce type d'argument a été utilisé dans de nombreux contextes en apprentissage statistique, voir par exemple [48, 17, 25].

Étant donné un jeu de données i.i.d. D_n et un échantillon « virtuel » $(x,y) \in \mathcal{X}\times\mathcal{Y}$, on note $\hat{f}_n^{(x,y)}$ l'EMV sur le jeu de données augmenté $D_n \cup (x,y)$.

Théorème 4

Pour tout estimateur $\hat{g}_n$ et toute loi P sur $\mathcal{X}\times\mathcal{Y}$, on a :

$$\mathcal{E}(\hat{g}_n) \le \mathbb{E}_{D_n,X}\Big[\sup_{y\in\mathcal{Y}}\Big\{\ell(\hat{g}_n(X),y) - \ell(\hat{f}_n^{(X,y)}(X),y)\Big\}\Big]\,. \tag{26}$$

La borne (26) est minimisée en prenant $\hat{g}_n = \tilde{f}_n$, où

$$\tilde{f}_n(y|x) = \frac{\hat{f}_n^{(x,y)}(y|x)}{\int_{\mathcal{Y}} \hat{f}_n^{(x,y')}(y'|x)\mu(\mathrm{d}\, y')}\,. \tag{27}$$

De plus, dans ce cas la borne (26) s'écrit :

$$\mathcal{E}(\tilde{f}_n) \le \mathbb{E}_{D_n,X}\Big[\log\Big(\int_{\mathcal{Y}} \hat{f}_n^{(X,y)}(y|X)\mu(\mathrm{d}\, y)\Big)\Big]\,. \tag{28}$$

Nous allons appliquer l'estimateur (27) et étudier les bornes correspondantes, pour deux modèles classiques : le modèle linéaire gaussien (pour une réponse appartenant à $\mathcal{Y} = \mathbf{R}$) et le modèle logistique (pour une réponse binaire, soit $\mathcal{Y} = \{-1, 1\}$).

Modèle linéaire gaussien. On suppose ici que $\mathcal{X} = \mathbf{R}^d$ et que $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, où $f_\theta(y|x) = \exp(-(y - \langle\theta, x\rangle)^2/2)$ est la densité de la loi gaussienne $\mathcal{N}(\langle\theta, x\rangle, 1)$ par rapport à la mesure de Lebesgue $\mu(\mathrm{d}\, y) = \mathrm{d}\, y/\sqrt{2\pi}$. Dans ce cas, la perte logarithmique s'écrit

$$\ell(f_\theta, (x, y)) = -\log f_\theta(y|x) = \frac{1}{2}(y - \langle\theta, x\rangle)^2\,, \tag{29}$$

ce qui coïncide (au facteur 1/2 près) avec la perte de la régression linéaire considérée en Section 3. En particulier, $L(f_\theta) = \mathbb{E}(Y - \langle\theta, X\rangle)^2$, et l'EMV $\hat{\theta}_n^{\mathrm{emv}}$ est simplement l'estimateur des moindres carrés $\hat{\theta}_n^{\mathrm{mc}}$. Ainsi, pour les estimateurs de la forme $\hat{f}_n = f_{\hat{\theta}_n} \in \mathcal{F}$ (qui choisissent une densité appartenant à la classe $\mathcal{F}$), le problème de l'estimation de densité est équivalent à celui de la régression linéaire.

Il découle alors des Propositions 1 et 2 que, dans le cas *bien spécifié*, l'estimateur minimax parmi les estimateurs à valeurs dans $\mathcal{F}$ est l'EMV $\hat{f}_n^{\mathrm{emv}}(y|x) = \mathcal{N}(\langle\hat{\theta}_n^{\mathrm{mc}}, x\rangle, 1)$, dont le risque vaut (lorsque $n \ge d$ et P_X est non dégénérée, avec les notations de la Section 3)

$$\mathcal{E}(\hat{f}_n^{\mathrm{emv}}) = \frac{\mathbb{E}[\mathrm{Tr}(\Sigma^{1/2}\widehat{\Sigma}_n^{-1}\Sigma^{1/2})]}{2n} = \frac{1}{2}\,\mathbb{E}\Big[\frac{h_{n+1}}{1 - h_{n+1}}\Big]\,.$$

Toujours dans le cas bien spécifié, il est possible de considérer le risque minimax parmi tous les estimateurs $\hat{f}_n$ (pouvant prendre des valeurs hors de $\mathcal{F}$). Dans ce cas, on montre de manière similaire que le risque minimax est infini si $n < d$ ou si P_X est dégénérée, et que sinon le risque minimax (bien spécifié) vaut

$$\mathcal{E}_n^*(\ell, \mathcal{F}, P_X \otimes \mathcal{F}) = -\frac{1}{2}\log\big(1 - h_{n+1}\big) \le \frac{1}{2}\mathbb{E}\log\Big(1 + \mathrm{Tr}(\Sigma^{1/2}\widehat{\Sigma}_n^{-1}\Sigma^{1/2})\Big)\,, \tag{30}$$

qui est également caractérisé par la loi du levier h_{n+1} et contrôlé par la queue inférieure de $\widehat{\Sigma}_n$. Lorsque $n \gg d$ et sous les conditions du Théorème 3, les deux expressions ci-dessous sont essentiellement équivalentes et d'ordre $d/(2n)$. Notons cependant que si $P_X = \mathcal{N}(0, \Sigma)$ et dans le régime asymptotique $n, d \longrightarrow \infty$ avec $d/n \longrightarrow \gamma \in (0, 1)$, le risque optimal (30) converge vers $-\frac{1}{2}\log(1-\gamma) < \frac{1}{2}\gamma/(1-\gamma)$, et l'on obtient donc un gain à considérer des estimateurs non restreints à $\mathcal{F}$ [38, Chapitre 8].

En revanche, la performance de l'EMV est sensible au caractère mal spécifié du modèle, c'est-à-dire à la loi de l'erreur $\varepsilon = Y - \langle \theta^*, X\rangle$ où $\theta^* = \arg\min_{\theta \in \mathbf{R}^d} L(f_\theta)$. (Le cas bien spécifié correspond au cas où ε est indépendante de X et de loi $\mathcal{N}(0,1)$.) En effet, le Théorème 3 donne dans ce cas que (sous des hypothèses convenables) $\mathcal{E}(\hat{\theta}_n^{\mathrm{emv}}) \lesssim \mathbb{E}[\varepsilon^2 \|\Sigma^{-1/2} X\|^2]/n$, qui peut être arbitrairement plus élevé que d/n lorsque $\mathbb{E}\,\varepsilon^2 \|\Sigma^{-1/2} X\|^2 \gg d$.

Théorème 5

L'estimateur $\tilde{f}_n$ défini par (27) s'écrit, pour le modèle linéaire gaussien,

$$\tilde{f}_n(y|x) = \mathcal{N}\big(\langle \hat{\theta}_n^{\mathrm{mc}}, x\rangle, (1 + \langle (n\widehat{\Sigma}_n)^{-1} x, x\rangle)^2\big) = \mathcal{N}\big(\langle \hat{\theta}_n^{\mathrm{mc}}, x\rangle, (1 - h_n(x))^{-2}\big) \quad (31)$$

où $h_n(x)$ est le levier de x, cf. (9). De plus, il satisfait la borne suivante, sans hypothèse sur la loi de $Y|X$:

$$\mathcal{E}(\tilde{f}_n) \le -\mathbb{E}\big[\log(1 - h_{n+1})\big] \le \frac{\mathbb{E}\,\mathrm{Tr}(\Sigma^{1/2}\widehat{\Sigma}_n^{-1}\Sigma^{1/2})}{n}. \quad (32)$$

La forme de l'estimateur $\tilde{f}_n$ s'interprète naturellement : la prédiction de la loi conditionnelle de Y associée à une valeur de $x \in \mathbf{R}^d$ est une gaussienne, mais dont la variance est corrigée par le levier de x au sein du jeu de données. Ainsi, les valeurs de x admettant un levier $h_n(x)$ élevé conduiront à des densités plus étalées ; cela est raisonnable, car nous avons vu en Section 3 que pour de tels points x, les prédictions de l'estimateur des moindres carrés seront plus incertaines. Cela suggère une interprétation générale de l'estimateur $\tilde{f}_n$ défini par (27), au-delà du cas gaussien : cet estimateur fondé sur des réponses y « fictives » calibre les prédictions en fonction de la sensibilité de l'EMV à la valeur de y en x, c'est-à-dire d'une notion de « levier » de x. De plus, la quantité (28) intervenant dans la borne de risque correspond à une notion de levier pour le problème de l'estimation de densité conditionnelle.

Du point de vue théorique, l'intérêt du Théorème 4 est que la borne (32) *ne dépend pas de la loi conditionnelle de Y sachant X*, c'est-à-dire de la loi de l'erreur

ε. De plus, cette borne valable dans le cas général mal spécifié est au plus deux fois la borne optimale (30) (et donc au plus deux fois la borne de l'EMV) dans le cas bien spécifié, quelle que soit la loi P_X de X. En particulier, sous les conditions du Théorème 3, cette borne est d'ordre $O(d/n)$ dans le cas mal spécifié, tandis que l'EMV peut se dégrader arbitrairement dans ce cas.

Notons que des bornes complémentaires d'excès de risque sur des boules ℓ^2 de $\mathbf{R}^d$ sont satisfaites par une variante régularisée de $\tilde{f}_n$, mais nous les omettons ici.

Modèle logistique. Nous traitons maintenant le cas du modèle logistique. Ici, on a toujours $\mathcal{X} = \mathbf{R}^d$, mais la réponse est binaire : $\mathcal{Y} = \{-1, 1\}$. Le modèle logistique est donné par $\mathcal{F} = \{f_\theta : \theta \in \mathbf{R}^d\}$, où f_θ est la densité sur $\{-1, 1\}$ (par rapport à la mesure de comptage) donnée par $f_\theta(1|x) = 1 - f_\theta(-1|x) = \sigma(\langle\theta, x\rangle)$, où $\sigma : \mathbf{R} \to \mathbf{R}$ est la fonction sigmoïde $\sigma(u) = e^u/(1 + e^u)$. Il s'agit du modèle « linéaire » standard lorsque y est binaire, à l'instar du modèle gaussien lorsque y est réel.

Pour des raisons techniques, nous ne considérons pas le modèle logistique $\mathcal{F}$ complet, mais seulement une boule ℓ^2 de rayon B, soit $\mathcal{F}_B = \{f_\theta : \|\theta\| \le B\}$. On suppose (et c'est la seule hypothèse) que X est borné presque sûrement, soit $\|X\| \le R$ p.s.

Dans ces conditions, l'estimateur le plus naturel est l'EMV $\hat{f}_n^{\text{emv}}$ sur $\mathcal{F}_B$. Sous ces hypothèses, des résultats classiques impliquent que l'excès de risque de l'EMV est d'au plus $O(\min(BR/\sqrt{n}, de^{BR}/n))$. Cette borne n'est cependant pas satisfaisante : la quantité $BR/\sqrt{n}$ décroît seulement en $1/\sqrt{n}$ (au lieu de $1/n$), tandis que la quantité de^{BR}/n est bien en d/n, mais avec un facteur exponentiel e^{BR} prohibitif. On peut se demander si l'EMV admet une meilleure borne, mais ce n'est pas le cas : sans hypothèse supplémentaire, tout estimateur prenant ses valeurs au sein du modèle logistique $\mathcal{F}$ peut avoir un excès de risque d'ordre $\Theta(\min(BR/\sqrt{n}, e^{BR}/n))$ [26].

La méthode d'agrégation par mélange bayésien (25) peut être appliquée à ce problème, et cette procédure admet une garantie améliorée en $O(d\log(BRn)/n)$ [29, 19]. En revanche, cette procédure est coûteuse à implémenter, car elle requiert d'effectuer de l'échantillonnage selon le postérieur.

Dans ce cas, la version régularisée de l'estimateur $\tilde{f}_n$ admet une forme simple. Étant donné l'échantillon D_n, $\lambda > 0$ ainsi qu'une paire (x, y), on pose

$$\hat{\theta}_\lambda^{(x,y)} = \operatorname*{arg\,min}_{\theta\in\mathbf{R}^d} \left\{ \frac{1}{n+1} \left(\sum_{i=1}^{n} \ell(f_\theta, (X_i, Y_i)) + \ell(f_\theta, (x, y)) \right) + \lambda\|\theta\|^2 \right\}. \tag{33}$$

Étant donné $x \in \mathbf{R}^d$, la prédiction $\tilde{f}_{\lambda,n}(y|x)$ vaut pour $y \in \{-1, 1\}$:

$$\tilde{f}_{\lambda,n}(y|x) = \frac{\sigma(y\langle\hat{\theta}_\lambda^{(x,y)}, x\rangle)e^{-\lambda\|\hat{\theta}_\lambda^{(x,y)}\|^2}}{\sigma(\langle\hat{\theta}_\lambda^{(x,1)}, x\rangle)e^{-\lambda\|\hat{\theta}_\lambda^{(x,1)}\|^2} + \sigma(-\langle\hat{\theta}_\lambda^{(x,-1)}, x\rangle)e^{-\lambda\|\hat{\theta}_\lambda^{(x,-1)}\|^2}}\,. \tag{34}$$

En particulier, le calcul de (34) se réduit au calcul de $\hat{\theta}_\lambda^{(x,y)}$ pour $y = \pm 1$, ce qui revient à résoudre deux problèmes de minimisation convexe de la forme (33). Les problèmes de minimisation convexe étant peu coûteux numériquement en comparaison aux problèmes d'échantillonnage, on obtient un gain computationnel.

Théorème 6

Pour toute loi P telle que $\|X\| \le R$ presque sûrement, l'estimateur (34) avec $\lambda = R^2/(n+1)$ satisfait, pour tout $B > 0$,

$$\mathcal{E}(\tilde{f}_{\lambda,n}; \mathcal{F}_B) \le \frac{3\,d + B^2R^2}{n}\,. \tag{35}$$

Ainsi, l'estimateur $\tilde{f}_{\lambda,n}$ contourne la borne inférieure en $\min(BR/\sqrt{n}, e^{BR}/n)$ pour les estimateurs restreints à la classe $\mathcal{F}$, en remplaçant la dépendance exponentielle en la norme par une dépendance quadratique. En particulier, dans le régime où $BR = O(\sqrt{d})$ (qui est naturel en dimension d), on obtient une borne $O(d/n)$.

Notons qu'il est possible d'obtenir une dépendance encore plus faible (logarithmique) en la norme via l'estimateur de mélange bayésien, au prix d'une procédure plus complexe à calculer. Une autre limitation de la garantie du Théorème 4 (ainsi que de celle de l'agrégation par mélange bayésien) est qu'elle ne contrôle que l'espérance de l'excès de risque, et pas ses déviations. Il serait intéressant d'obtenir une procédure calculable efficacement qui admette des bornes d'excès de risque favorables en forte probabilité.

5 Conclusion

Dans ce résumé, nous avons considéré deux problèmes d'apprentissage statistique, à savoir la régression linéaire et l'estimation de densité conditionnelle.

Pour la régression linéaire, la difficulté du problème est caractérisée par la loi des *leviers* des variables $X_1, \dots, X_n$. Ceci implique que la loi gaussienne est la plus favorable en grande dimension. La procédure la plus naturelle pour l'estimation de densité conditionnelle, à savoir l'estimateur du maximum de vraisemblance,

est satisfaisante dans le cas bien spécifié mais peut se dégrader significativement dans le cas mal spécifié. Il est cependant possible de corriger cet estimateur en le « régularisant » à partir d'échantillons « fictifs », ce qui revient à calibrer ses prédictions à partir d'une notion de levier. Pour les modèles linéaire gaussien et logistique (ainsi que quelques autres), l'estimateur ainsi obtenu admet de meilleures garanties que l'EMV dans le cas mal spécifié, tout en restant calculable par optimisation convexe.

Références

[1] R. Adamczak, A. Litvak, A. Pajor et N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2) :535-561, 2010.

[2] J.-Y. Audibert et O. Catoni. Robust linear least squares regression. *Annals of Statistics*, 39(5) :2766-2794, 2011.

[3] A. R. Barron. Are Bayes rules consistent in information ? In *Open Problems in Communication and Computation*, pages 85-91. Springer, 1987.

[4] L. Birgé et P. Massart. Minimum contrast estimators on sieves : exponential bounds and rates of convergence. *Bernoulli*, 4(3) :329-375, 1998.

[5] L. Breiman et D. Freedman. How many variables should be entered in a regression equation ? *J. American Statistical Association*, 78(381) :131-136, 1983.

[7] A. Caponnetto et E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3) :331-368, 2007.

[8] O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv:1603.05229*, 2016.

[9] O. Catoni. *PAC-Bayesian Supervised Classification : The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics, 2007.

[10] O. Catoni. *Statistical Learning Theory and Stochastic Optimization : Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*. Springer-Verlag, 2004.

[14] T. M. Cover et J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[16] L. Devroye, L. Györfi et G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, tome 31 de *Applications of Mathematics*. Springer-Verlag, 1996.

[17] L. Devroye et T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Information Theory*, 25(2) :202-207, 1979.

[18] N. El Karoui. Random matrices and high-dimensional statistics : beyond covariance matrices. In *Proceedings of the ICM*, pages 2875-2894, Rio, 2018.

[19] D. Foster, S. Kale, H. Luo, M. Mohri et K. Sridharan. Logistic regression : the importance of being improper. In *31st Conference on Learning Theory*, 2018.

[23] L. Györfi, M. Kohler, A. Krzyzak et H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2002.

[25] D. Haussler, N. Littlestone et M. K. Warmuth. Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115(2) :248-292, 1994.

[26] E. Hazan, T. Koren et K. Y. Levy. Logistic regression : Tight bounds for stochastic and online optimization. In *27th Conference on Learning Theory*, 2014.

[27] D. Hsu, S. M. Kakade et T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3) :569-600, 2014.

[28] A. Juditsky, P. Rigollet et A. B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36(5) :2183-2206, 2008.

[29] S. Kakade et A. Ng. Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems 17*, pages 641-648, 2005.

[32] V. Koltchinskii et K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1) :110-133, 2017.

[33] V. Koltchinskii et S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *IMRN*, 2015(23) :12991-13008, 2015.

[34] G. Lecué et S. Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3) :1520-1534, 2016.

[35] N. Littlestone et M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2) :212-261, 1994.

[36] V. A. Marchenko et L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4) :507-536, 1967.

[37] D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3) :355-363, 1999.

[38] J. Mourtada. *Contributions à l'apprentissage statistique : estimation de densité, agrégation d'experts et forêts aléatoires*. Institut polytechnique de Paris, 2020.

[39] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *arXiv:1912.10754*, 2019.

[40] J. Mourtada et S. Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *arXiv:1912.10784*, 2019.

[41] R. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3) :1175-1194, 2016.

[43] M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1) :60-72, 1999.

[44] N. Srivastava et R. Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *Annals of Probability*, 41(5) :3081-3111, 2013.

[45] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

[46] A. B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, Lecture Notes in Artificial Intelligence, pages 303-313. Springer, 2003.

[47] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[48] V. Vapnik et A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, 1974.

[49] V. Vovk. A game of prediction with expert advice. *Journal of Computer and System Sciences*, 56(2) :153-173, 1998.

[50] Y. Yang. Mixing strategies for density estimation. *Annals of Statistics*, 28(1) :75-87, 2000.

Jaouad Mourtada

Depuis septembre 2020, Jaouad Mourtada est maître de conférence au département de statistique de l'ENSAE/CREST. Ses recherches se situent à l'intersection des statistiques et de la théorie de l'apprentissage. Il s'est particulièrement intéressé à la compréhension de la complexité des problèmes de prédiction et d'estimation. Ses intérêts de recherche actuels tournent principalement autour de questions relatives à la théorie de l'apprentissage statistique et à la statistique robuste.

Email : jaouad.mourtada@ensae.fr
Site web : https://jaouadmourtada.github.io