Compromis performance-complexité pour les statistiques en grande dimension

Romain COUILLET[1],Florent CHATELAIN[2],Nicolas LE BIHAN[2]

[1]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

florent.chatelain@gipsa-lab.fr, romain.couillet@univ-grenoble-alpes.fr, nicolas.le-bihan@gipsa-lab.fr

Résumé – Cet article introduit un cadre de matrices aléatoires pour l'analyse du compromis entre performance et complexité dans une classe d'algorithmes d'apprentissage automatique, sous un régime de données de grande dimension. Plus précisément, nous analysons les propriétés spectrales de $K \odot B \in \mathbb{R}^{n\times n}$, pour une matrice aléatoire à noyau $K \in \mathbb{R}^{n\times n}$ sur laquelle on applique un masque de parcimonie $B \in \{0,1\}^{n\times n}$: cela réduit le nombre des K_{ij} à évaluer, et fait baisser la complexité, tout en affaiblissant la puissance de l'inférence statistique sur K. De manière surprenante nous montrons que, sous des hypothèses réalistes, les performances ne sont que marginalement altérées.

Abstract – This paper introduces a random matrix framework to analyze the trade-off between performance and complexity in a class of machine learning algorithms, under a high-dimensional data regime. More precisely, we analyse the spectral properties of K_{ij}, for a random matrix with kernel $K_{ij} \in \mathbb{R}^{n\times n}$ on which we apply a sparsity mask $B \in \{0,1\}^{n\times n}$: this reduces the number of K_{ij} to be evaluated, and lowers the complexity, while weakening the power of statistical inference on K. Under realistic assumptions, we demonstrate a surprisingly low performance decay even under severe sparsity levels.

1 Introduction

L'augmentation des volumes des données et le besoin présupposé de concevoir des algorithmes autonomes pour les traiter exercent une forte pression sur l'agilité des techniques d'apprentissage automatique et la compréhension des limites fondamentales du traitement des données de grande dimension.

En supposant de grandes dimension p et taille n de données $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ (formellement, $n, p \to \infty$ avec $p/n \to c \in (0, \infty)$), la théorie des matrices aléatoires a récemment montré que les noyaux non linéaires $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ [1, 2, 3] et les matrices à fonction d'activation [4], omniprésentes en apprentissage machine, présentent un comportement simple et tractable dans la limite.

Les matrices à noyau $K \in \mathbb{R}^{n\times n}$ au cœur des algorithmes d'apprentissage sont néanmoins coûteuses à évaluer, à stocker, mais surtout à exploiter (inversion, extraction de vecteurs propres, etc.). L'objectif de l'article est d'évaluer les conséquences théoriques d'une réduction (éventuellement drastique) du nombre d'entrées de K à évaluer, en termes de performances. Plus précisément, en introduisant un masque aléatoire B qui élimine une proportion $1 - \varepsilon$ des entrées de K, nous découvrons le spectre limite du noyau 'creux' $K \odot B$ et étudions le phénomène de transition de phase lorsque $K \odot B$ opère dans un cadre de classification spectrale non supervisée.

2 Modèle

Soit $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ une collection de n échantillons de données de dimension p , modélisé par :

$$X = Z + \sqrt{n}\mu v^{\mathsf{T}} \tag{1}$$

où $Z \in \mathbb{R}^{p\times n}$ a des entrées i.i.d. $N(0,1)$, $\mu \in \mathbb{R}^p$ et $v \in \mathbb{R}^n$ sont des vecteurs déterministes, indépendants de p et n et tel que $\|\mu\|$ est constant, $\|v\|^2 = 1$ et $\limsup_n \max_{1\le i\le n}\{\sqrt{n}v_i^2\} = 0$.

De plus, définissons la matrice binaire symétrique $B \in \{0,1\}^{n\times n}$, qui servira de masque appliqué à la matrice de noyau canonique $\frac{1}{p}X^{\mathsf{T}}X$, avec, pour $1 \le i < j \le n$,

$$B_{ij} \sim \mathrm{Bern}(\varepsilon)$$

et $B_{ji} = B_{ij}$, et pour $1 \le i \le n$, $B_{ii} = b \in \{0,1\}$.

L'objet central de cet article est la *matrice de noyau creuse*

$$S = \frac{1}{p}X^{\mathsf{T}}X \odot B.$$

D'après la définition de B, en moyenne, une proportion $1 - \varepsilon$ des entrées hors-diagonale de S est mise à zéro (donc en pratique non évaluée), de sorte que $1 - \varepsilon$ joue le rôle d'exhausteur de sparsité ; quant aux entrées diagonales, elles sont soit toutes maintenues (si $b = 1$), soit mises à zéro (si $b = 0$).

L'objectif de cet article est de fournir une description du comportement spectral, et plus fondamentalement (a) de l'existence d'une valeur propre isolée dominante $\hat{\lambda}$ dans le spectre de

S et (b) de la corrélation entre le vecteur propre $\hat{v}$ associé à $\hat{\lambda}$ et le vecteur de population v, en fonction du rapport limite c, du rapport signal/bruit $\|\mu\|^2$, et du paramètre de parcimonie ε.

À cette fin, nous supposons p et n grands, ou mathématiquement que $p, n \to \infty$ de telle sorte que $p/n \to c \in (0, \infty)$. Nous rappelons en outre que $\|\mu\|$ est fixe par rapport à p, n, que $\|v\| = 1$, et que $\limsup_n \max_{1 \le i \le n} \{\sqrt{n} v_i^2\} = 0$.

En s'appuyant sur les outils de la théorie des matrices aléatoires (en particulier les outils développés dans [5]), le comportement spectral en grande dimension de S est accessible via une analyse approfondie de sa *résolvante*

$$Q(z) \equiv (S - zI_p)^{-1} \tag{2}$$

définie pour tout $z \in \mathbb{C} \setminus \mathrm{Sp}(S)$ avec $\mathrm{Sp}(S)$ l'ensemble des valeurs propres de S. La résolvante $Q(z)$ devient l'objet central de nos principaux résultats techniques, décrits ci-après.

3 Résultats principaux

3.1 Équivalent déterministe et spectre limite

Notre principal résultat fournit un *équivalent déterministe* $\bar{Q}(z)$ pour la résolvante $Q(z)$ définie dans (2), c'est-à-dire que $\bar{Q}(z)$ est déterministe et telle que, pour toutes les suites de matrices déterministes $A \in \mathbb{R}^{n \times n}$ et de vecteurs $a, b \in \mathbb{R}^n$ de normes bornées (par rapport à n), avec probabilité 1,

$$\frac{1}{n} \operatorname{tr} A(Q(z) - \bar{Q}(z)) \to 0, \quad a^\mathsf{T} \left(Q(z) - \bar{Q}(z)\right) b \to 0.$$

On note ce résultat $Q(z) \leftrightarrow \bar{Q}(z)$ qui nous permet, comme on le montrera par la suite, de transférer la plupart des propriétés spectrales de $Q(z)$ (et donc de S) à $\bar{Q}(z)$.

Theorem 3.1 (Equivalent déterministe de $Q(z)$)**.** *Considérons la résolvante $Q(z) = \left(\frac{X^\mathsf{T} X}{n} \odot B - zI_p\right)^{-1}$, pour $X \in \mathbb{R}^{p \times n}$ avec $X = Z + M$ où Z a des entrées i.i.d. de moyenne nulle et de variance unité, et $M = \mu v^\mathsf{T}$. Sous les hypothèses et notations de la Section 2, lorsque $p, n \to \infty$,*

$$\begin{aligned} Q(z) \leftrightarrow \bar{Q}(z) &\equiv m(z) \left(I_n + \frac{\|\mu\|^2 \varepsilon m(z)}{c + \varepsilon m(z)} v v^\mathsf{T}\right)^{-1} \\ &= m(z) I_n - \frac{\|\mu\|^2 \varepsilon m(z)^2 v v^\mathsf{T}}{c + \varepsilon m(z)(1 + \|\mu\|^2)} \end{aligned} \tag{3}$$

où $m(z)$ est la transformée de Stieltjes ($m(z) = \int (t - z)^{-1} \nu(dt)$) de la mesure spectrale limite presque sûre $\nu = \lim_n \frac{1}{n} \sum_{\lambda \in \mathrm{Sp}(S)} \delta_\lambda$ de S, et est l'unique solution analytique complexe de l'équation fonctionnelle

$$z = b - \frac{1}{m(z)} - \frac{\varepsilon}{c} m(z) + \frac{\varepsilon^3 m(z)^2}{c(c + \varepsilon m(z))}. \tag{4}$$

La preuve de ce théorème n'est pas présentée ici.

Avant d'exploiter le Théorème 3.1, quelques remarques s'imposent. Observons tout d'abord que $\bar{Q}(z)$ prend la forme d'une perturbation de la matrice identité pondérée par la matrice rang-1 pondérée vv^T. Conformément aux modèles de matrices aléatoires [6, 7], cette forme prédit l'existence possible d'une valeur propre dominante isolée $\hat{\lambda}$ dans le spectre de S, de vecteur propre $\hat{v}$ aligné dans une certaine mesure à v. Ceci est établi en Section 3.2. Par ailleurs, des implications intéressantes du Théorème 3.1 peuvent être tirées dans la limite où $\varepsilon \to 0$ ou 1.

Remark 1 (Limites de Marčenko-Pastur et du demi-cercle). Lorsque $\varepsilon = 1$, avec $z' = z + 1 - b$, l'équation (4) devient

$$z' m_b(z')^2 + (cz' + 1 - c) m_b(z') + c = 0$$

où $m_b(z') \equiv m(z' + b - 1)$ est la transformée de Stieltjes de la mesure $\nu(\cdot + b - 1)$. Nous retrouvons ici l'équation définissant la transformée de Stieltjes de la distribution de Marčenko-Pastur [8] pour la variable $z + 1 - b$. En particulier, la mesure limite ν a pour support $[(1 - \sqrt{1/c})^2 + b - 1, (1 + \sqrt{1/c})^2 + b - 1]$.

Si au contraire $\varepsilon \ll 1$, on peut montrer à partir du Corollaire 3.2.1 plus bas que l'équation (4) devient

$$z - b + \frac{1}{m(z)} + \frac{\varepsilon}{c} m(z) = O_\varepsilon(\varepsilon^2). \tag{5}$$

Ce qui, en notant $z' = \sqrt{c/\varepsilon}(z - b)$, conduit à

$$m_{b,\varepsilon}(z')^2 + z' m_{b,\varepsilon}(z') + 1 = O_\varepsilon(\varepsilon^{\frac{3}{2}}) \tag{6}$$

où $m_{b,\varepsilon}(z') = \sqrt{\varepsilon/c}\, m(\sqrt{\varepsilon/c} z' + b)$ pour $m_{b,\varepsilon}$ la transformée de Stieltjes de la mesure décalée et pondérée $\nu((\cdot - b)\sqrt{c/\varepsilon})$. Nous retrouvons ainsi la transformée de Stieltjes de la loi du demi-cercle de Wigner [9]. En particulier, au premier ordre en ε, le support limite de ν est $[-2\sqrt{\varepsilon/c} + b, 2\sqrt{\varepsilon/c} + b]$.

La Remarque 1 prédit donc que la mesure limite ν du spectre du noyau creux $S = K \odot B$ évolue de la loi de Marčenko-Pastur, typique de la matrice de Gram $X^\mathsf{T} X$ vers la mesure du demi-cercle, typique de la matrice symétrique B à entrées indépendantes (à symétrie près). Il est intéressant de noter que le terme négligeable en ε dans (6) est d'ordre $O_\varepsilon(\varepsilon^{\frac{3}{2}})$, suggérant une convergence rapide vers le comportement en demi-cercle, dès que ε est éloigné de 1. La Figure 1 confirme visuellement cette observation en affichant les comportements limites pour $\varepsilon \in \{0.1, 0.5, 0.9\}$. La figure prédit également la possibilité, pour $c < 1$, d'un état transitoire où le support de ν est divisé en deux composantes connexes.

Après avoir établi le Théorème 3.1 nous pouvons maintenant évaluer les conditions exactes sous lesquelles le signal μv^T peut être récupéré à partir de S et, sous ces conditions, la qualité de l'estimation du vecteur d'information v.

3.2 Transition de phase, valeurs propres et vecteurs propres isolés

Cette section établit (i) la condition sur $\|\mu\|$ sous laquelle la plus grande valeur propre $\hat{\lambda}$ de S s'isole (et devient donc informative) et, dans ce cas, (ii) la limite ζ de l'alignement $\hat{\zeta} \equiv$

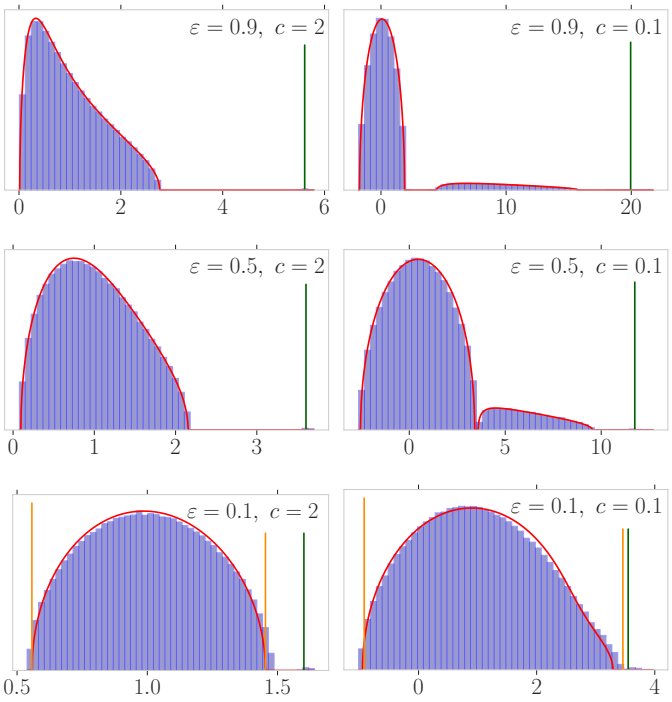

FIGURE 1 – Distribution empirique des valeurs propres de $\frac{1}{p}X^\mathsf{T}X \odot B$, pour $b = 1$, $n = 1000$ et $v = [1_{\frac{n}{2}}^\mathsf{T}, -1_{\frac{n}{2}}^\mathsf{T}]^\mathsf{T}$. (ligne du **haut**) $\varepsilon = 0.9$; (ligne du **milieu**) $\varepsilon = 0.5$; (ligne du **bas**) $\varepsilon = 0.1$. (colonne **gauche**) $p = 2000$ ($c = 2$) et $\|\mu\|^2 = 3$; (colonne **droite**) $p = 100$ ($c = 0.1$) et $\|\mu\|^2 = 1$. (**Trait rouge**) Densité spectrale limite théorique ν calculée par inversion numérique de la transformée de Stieltjes $m(z)$ du Théorème 3.1. [1] (**Trait vert**) position théorique du spike isolé du Théorèmé 3.2. (**Trait orange**) approximation des limites du support pour ε petit comme donné dans le Corollaire 3.2.1.

$|\hat{v}^\mathsf{T}v|^2$ entre le vecteur propre $\hat{v}$ associé à $\hat{\lambda}$ et le vecteur propre de population v de la matrice d'information $(\mu v^\mathsf{T})^\mathsf{T}(\mu v^\mathsf{T})$.

À cette fin, nous exploitons le Théorème 3.1 en remarquant que, selon l'intégrale de Cauchy,

$$|\hat{v}^\mathsf{T}v|^2 = \frac{-1}{2\pi\imath}\oint_{\mathcal{C}_{\hat{\lambda}}} v^\mathsf{T}Q(z)v\,dz \simeq \frac{-1}{2\pi\imath}\oint_{\mathcal{C}_{\lambda}} v^\mathsf{T}\bar{Q}(z)v\,dz \quad (7)$$

pour $\mathcal{C}_x$ un contour complexe suffisamment petit et orienté positivement entourant x, et λ la limite (si définie) de $\hat{\lambda}$ lorsque $p, n \to \infty$; l'approximation n'est vraie que si $\hat{\lambda}$ reste effectivement isolé de toutes les autres valeurs propres de S. L'expression *déterministe* du membre droit de (7) peut être évaluée explicitement et conduit au résultat suivant.

Theorem 3.2 (Spectre isolé)**.** *Definissons les fonctions*

$$F(x) = x^4 + 2x^3 + \left(1 - \frac{c}{\varepsilon}\right)x^2 - 2cx - c$$

$$G(x) = b + \frac{\varepsilon}{c}(1+x) + \frac{1}{1+x} + \frac{\varepsilon}{x(1+x)}$$

et notons Γ la plus grande solution de $F(\Gamma) = 0$. Alors, sous les hypothèses de la Section 2, lorsque $p, n \to \infty$, avec probabilité 1, la plus grande valeur propre $\hat{\lambda}$ de S et son vecteur propre associé $\hat{v}$ vérifient

$$\hat{\lambda} \to \lambda = \begin{cases} G(\|\mu\|^2) & ,\ \|\mu\|^2 > \Gamma \\ G(\Gamma) & ,\ \|\mu\|^2 \le \Gamma \end{cases}$$

$$\hat{\zeta} \equiv |\hat{v}^\mathsf{T}v|^2 \to \zeta = \begin{cases} \frac{F(\|\mu\|^2)}{\|\mu\|^2(1+\|\mu\|^2)^3} & ,\ \|\mu\|^2 > \Gamma \\ 0 & ,\ \|\mu\|^2 \le \Gamma. \end{cases}$$

Le Théorème 3.2 assure la présence d'une valeur propre dominante isolée $\hat{\lambda}$ de S et un alignement non trivial $\hat{\zeta}$ entre le vecteur propre correspondant $\hat{v}$ et le vecteur d'information v du modèle, si et seulement si $\|\mu\|^2 > \Gamma$. Si au contraire $\|\mu\|^2 \le \Gamma$, alors $\hat{\lambda}$ converge vers le bord droit E_ν^+ du support $[E_\nu^-, E_\nu^+]$ de ν, de sorte que

$$E_\nu^+ = b + \frac{\varepsilon}{c}(1+\Gamma) + \frac{1}{1+\Gamma} + \frac{\varepsilon}{\Gamma(1+\Gamma)}.$$

Il est intéressant de noter que les valeurs-limites λ et ζ ont des expressions explicites, tandis que le seuil Γ reste implicite (du moins, il prend la forme peu pratique d'une racine d'un polynôme d'ordre 4). Lorsque $\varepsilon \ll 1$ cependant, la valeur de Γ devient accessible.

Corollary 3.2.1 (Approximation des petits ε)**.** *Avec les notations du Théorème 3.2, dans la limite des petits ε,*

$$\Gamma = \sqrt{\frac{c}{\varepsilon}} - 1 + \varepsilon + O(\varepsilon^{\frac{3}{2}}), \quad E_\nu^\pm = b \pm 2\sqrt{\frac{\varepsilon}{c}} + \frac{\varepsilon^2}{c} + O(\varepsilon^{\frac{5}{2}}).$$

La précision de ces estimations est illustrée en Figure 1.

3.3 Compromis performance-complexité

En écartant une proportion ε des entrées de K, le coût de calcul de $K \odot B$ est réduit d'un facteur ε par rapport au coût de calcul de K. Par ailleurs, pour estimer le vecteur propre principal $\hat{v}$ de $K \odot B$, on peut naturellement recourir à la méthode de la *puissance* qui exécute la procédure $v^{t+1} = \tilde{v}^{t+1}/\|\tilde{v}^{t+1}\|$ avec $\tilde{v}^{t+1} = (K \odot B)v^t$ pour tous les $t \ge 0$ pour un certain v^0 arbitraire jusqu'à convergence. Pour tout n grand, le produit $(K \odot B)v^t$ a un coût de calcul d'ordre $O(n^2\varepsilon)$ et donc encore réduit d'un facteur ε.

4 Application en clustering spectral

Notre analyse des noyaux creux trouve plusieurs applications immédiates, illustrées ici. Une application directe est celle du 'regroupement spectral à noyau creux' de données gaussiennes de grande taille $x_i \sim \mathcal{N}(\pm\mu, I_p)$ où $\sqrt{n}v_i \in \{\pm 1\}$ tient compte du signe de $\mathbb{E}[x_i]$.

Les performances asymptotiques du noyau creux $\frac{1}{p}X^\mathsf{T}X \odot B$ sont décrites ci-dessous et comparées à la technique standard de sous-échantillonnage des données de proportion ε. Dans cette application, nous supposons que $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ modélise un ensemble de données généré par un modèle de mélange gaussien à deux classes, avec $x_i \sim \mathcal{N}(v_i\mu, I_p)$.

Nous pouvons alors écrire $X = Z + \sqrt{n}\mu v^\mathsf{T}$ où $\sqrt{n}v \in \{-1, 1\}^n$. Nous imposons en outre que $\sum_i v_i = 0$, c'est-à-dire que chaque classe est de même taille.

D'après l'idée de l'algorithme de regroupement spectral [10] de noyau $K = \{\frac{1}{p} x_i^\mathsf{T} x_j\}_{i,j=1}^n = \frac{1}{p} X^\mathsf{T} X$, le vecteur propre dominant $\hat{\mathbf{v}}$ de $X^\mathsf{T} X$ devrait être aligné à v. De plus, par la symétrie complète du modèle de classes, l'estimation naturelle $\hat{\mathcal{C}}_i$ de la classe $\mathcal{C}_i$ de x_i est directement donnée par $\mathrm{sgn}(\hat{\mathbf{v}}_i)$. En travaillant plutôt sur le noyau creux $\frac{1}{p} X^\mathsf{T} X \odot B$, et donc sur le vecteur propre dominant $\hat{v}$ plutôt que $\hat{\mathbf{v}}$ (les deux étant égaux lorsque $\varepsilon = 1$ et $b = 1$), la performance correspondante du regroupement spectral creux est donné comme suit.

Theorem 4.1 (Performance du regroupement spectral creux)**.** *Soit $\hat{\mathcal{C}}_i = \mathrm{sgn}(\hat{v}_i)$ la classe estimée $\mathcal{C}_i$ du vecteur x_i, avec la convention $v_1 \hat{v}_1 > 0$ pour les vecteurs propres. Alors, avec probabilité 1,*

$$\frac{1}{n} \sum_{i=1}^n \delta_{\{\mathcal{C}_i = \hat{\mathcal{C}}_i\}} = Q\left(\sqrt{\zeta/(1-\zeta)}\right) + o(1)$$

où ζ est donné au Théorème 3.2 et $Q(x) = \frac{1}{2\pi} \int_x^\infty e^{-t^2/2} dt$.

La Figure 2 illustre le Théorème 4.1 en comparant les performances théoriques et simulées pour des ε et $\|\mu\|^2$ variables. Elle confirme la chute soudaine de la précision de classification en dessous du seuil de transition de phase et montre que les asymptotiques prédites sont déjà assez précises dans ce cadre modérément grand de $n = 200$, $p = 800$. Une approche

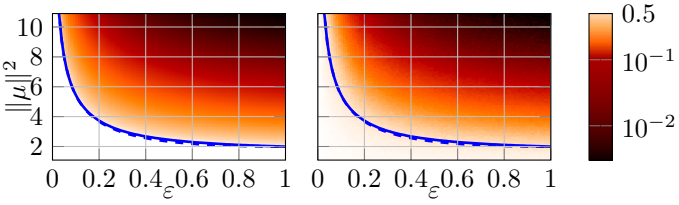

FIGURE 2 – Performance du regroupement spectral en fonction d'ε (abscisses) et $\|\mu\|^2$ (ordonnées) pour $c = 4$, $n_1 = n_2 = n/2$ et $n = 200$. (**Gauche**) limites asymptotiques du Théorème 4.1 ; (**droite**) simulations moyennées sur 100 réalisations. (**Bleu**) Transition de phase théorique pour $\|\mu\|^2 = \Gamma$ obtenue par le Théorème 3.2. (**Tirets bleus**) Approximation pour $\varepsilon \ll 1$ d'après le Corollaire 3.2.1.

alternative pour réduire la complexité de calcul du regroupement spectral consiste à sous-échantillonner $n_s < n$ vecteurs $X_s \in \mathbb{R}^{p \times n_s}$ de l'ensemble des données X et à effectuer le regroupement spectral sur la matrice résultante $\frac{1}{p} X_s^\mathsf{T} X_s$. En prenant $n_s = \lceil \varepsilon n \rceil$, la complexité, en utilisant une méthode de puissance, est réduite d'un facteur $O(\varepsilon^2)$. Pour réduire la complexité d'un regroupement spectral à n dimensions, on peut alors prendre $n_s = n/m$ ($\varepsilon = 1/m$) et effectuer m regroupements spectraux parallèles, chacun de complexité réduite de $1/m^2$: finalement, de manière similaire à la méthode du noyau creux, cela réduit le coût global d'un facteur $\varepsilon = 1/m$.

Cependant, cette dernière procédure perd le bénéfice de la 'redondance' inhérente aux données issues de la même classe, que les méthodes à noyau exploitent [3]. Ceci est tout préjudiciable à ses performances. En effet, la précision asymptotique $|\hat{v}_s^\mathsf{T} v_s|$, avec $v_s \in \mathbb{R}^{n_s}$ le sous-ensemble normalisé de v sur les n_s indices sélectionnés et $\hat{v}_s \in \mathbb{R}^{n_s}$ le vecteur propre dominant de $\frac{1}{p} X_s^\mathsf{T} X_s$, découle du Théorème 3.2 en prenant dans l'énoncé du théorème : 1) $\varepsilon = 1$ et 2) $c \to c/\varepsilon$ où $\varepsilon = n_s/n$ devient le taux de sous-échantillonnage. En utilisant l'indice s dans la suite pour désigner le cas de sous-échantillonnage, cela donne

$$\begin{aligned} F_s(x) &= x^4 + 2x^3 + \left(1 - \frac{c}{\varepsilon}\right) x^2 - \frac{2cx}{\varepsilon} - \frac{c}{\varepsilon} \qquad (8) \\ &= (x+1)^2 \left(x + \sqrt{\frac{c}{\varepsilon}}\right)\left(x - \sqrt{\frac{c}{\varepsilon}}\right) \end{aligned}$$

dont la plus grande racine $\Gamma_s = \sqrt{c/\varepsilon}$ est la transition de phase classique de Marčenko-Pastur [6]. On obtient donc

$$|\hat{v}_s^\mathsf{T} v_s|^2 \to \zeta_s = \frac{\max\{F_s(\|\mu\|^2), 0\}}{\|\mu\|^2(1+\|\mu\|^2)^3} = \frac{\max\{\|\mu\|^4 - c/\varepsilon 0\}}{\|\mu\|^2(1+\|\mu\|^2)}.$$

Notons que, pour $x > 0$, $F(x) - F_s(x) = c(2x+1)(\varepsilon^{-1} - 1) \geq 0$ avec égalité seulement si $\varepsilon = 1$, de sorte que (i) $\Gamma < \Gamma_s$: la transition de phase du noyau creux se produit à des rapports signal/bruit $\|\mu\|^2$ inférieurs, et (ii) $\zeta > \zeta_s$: l'alignement asymptotique est plus grand pour le noyau creux. Les deux méthodes ne sont équivalentes que si $\varepsilon = 1$, tandis que le gain du noyau creux est accru dans le régime de faible densité.

La comparaison avec la transition de phase obtenue pour le noyau creux selon le Corollaire 3.2.1 ($\Gamma = \sqrt{c/\varepsilon} - 1 + O(\varepsilon)$) montre un gain d'ordre 1 sur la transition de phase du rapport signal/bruit lorsque ε est petit. La Figure 3, à comparer à la Figure 2, illustre clairement ce résultat. Mais la Figure 3 révèle un message plus fondamental : ici, pour c petit, la transition de phase du noyau creux a une forme de "plateau", ce qui suggère fortement que des niveaux très sévères de parcimonie peuvent être effectuées sans aucune perte de performance.

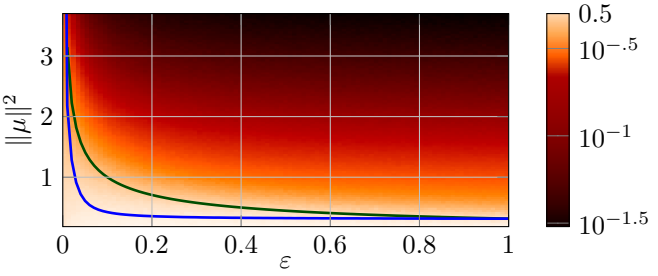

FIGURE 3 – Performance de classification du ε-sous-échantillonnage en fonction de ε (abcisse) et $\|\mu\|^2$ (ordonnée) pour $c = 0.1$, $n_1 = n_2 = n/2$ et $n = 1000$. Simulations réalisées sur 100 tirages. (**Vert**) Transition de phase du ε-sous-échantillonnage ($\|\mu\|^2 = \sqrt{c/\varepsilon}$). (**Bleu**) Transition de phase du noyau perforé (d'après le Théorème 3.2).

Références

[1] N. El Karoui, *The spectrum of kernel random matrices*. The Annals of Statistics,38((1), 2010.

[2] X. Cheng, and A. Singer, *The spectrum of random inner-product kernel matrices*, Random Matrices : Theory and Applications, 2(04), 2013.

[3] R. Couillet, and F. Benaych-Georges, *Kernel spectral clustering of large dimensional data*, Electronic Journal of Statistics, 10(1), 2016.

[4] , Z. Liao, and R. Couillet, *On the Spectrum of Random Features Maps of High Dimensional Data*, ICML, 2018.

[5] , L.A. Pastur, and M. Shcherbina, *Eigenvalue distribution of large random matrices*, 171, 2011.

[6] , J. Baik, and J.W. Silverstein, *Eigenvalues of large sample covariance matrices of spiked population models*, Journal of multivariate analysis, 97(6), 2006.

[7] , F. Benaych-Georges, and R.R. Nadakuditi, *The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices*, Advances in Mathematics, 227(1), 2011.

[8] , V.A. Marcenko, and L.A. Pastur, *Distribution of eigenvalues for some sets of random matrices*, Mathematics of the USSR-Sbornik,1(4), 1967.

[9] , E.P. Wigner, *Characteristic Vectors of Bordered Matrices With Infinite Dimensions*, Annals of Mathematics, 62(3), 1955.

[10] , U. Von Luxburg, *A tutorial on spectral clustering*, Statistics and computing, 17(4), 2007.