Apprentissage semi-supervisé sur les graphes pour les données de grande dimension

Xiaoyi MAI[1], Romain COUILLET[1,2]

[1]Laboratoire des Signaux et Systèmes, CentraleSupélec, Université Paris-Saclay, 3 rue Joliot Curie, 91192 Gif-Sur-Yvette

[2]Chaire DataScience GSTATS, GIPSA-lab, Université Grenoble–Alpes, 11 rue des Mathématiques, 38400 St Martin d'Hères

xiaoyi.mai@l2s.centralesupelec.fr, romain.couillet@centralesupelec.fr

Résumé – Nous montrons que, pour des données de grande dimension, la régularisation laplacienne, méthode classique en apprentissage semi-supervisé, bénéficie nullement de l'apport de données non-étiquetées, même nombreuses. A tel point qu'un simple regroupement spectral (non-supervisé) peut surpasser la méthode, remettant en question la nature même de l'apprentissage semi-supervisé. L'article présente une solution au moyen d'une régularisation recentrée, dont les performances théoriques sont corroborées sur données réelles.

Abstract – For large dimensional data, the Laplacian regularization, a classical graph-based semi-supervised learning method, is proved to have negligible learning gain from high unlabelled data. Consequently, the method is outperformed by a mere non-supervised spectral clustering, questioning the very purpose of semi-supervised learning. The article presents a solution based on a centered regularization approach, whose theoretical performances are corroborated on real data.

1 Introduction

L'apprentissage semi-supervisé (SSL) est une méthode d'apprentissage automatique qui exploite conjointement des données étiquetées et non-étiquetées. Comme l'étiquetage des données est un processus manuel qui prend beaucoup de temps, contrairement à la collecte peu coûteuse de données, l'apprentissage semi-supervisé vise en particulier à améliorer la précision de la classification, en utilisant une grande quantité de données non-étiquetées en association avec quelques données étiquetées. De toute évidence, combiner données étiquetées et non-étiquetées devrait permettre à l'apprentissage semi-supervisé de toujours être meilleur que les apprentissages supervisé ou non-supervisé pris de manière isolée. Cependant, en raison de la difficulté de combiner proprement les informations des données étiquetées et non-étiquetées, il a été montré [2, Chapitre 4] que de nombreuses techniques classiques d'apprentissage semi-supervisé n'atteignent pas cet objectif, ce qui rend les méthodes d'apprentissage semi-supervisé peu populaires.

Dans cet article, nous nous intéressons à la régularisation laplacienne [5], méthode classique de classification semi-supervisée sur graphe, en connexion avec la propagation des étiquettes, la marche aléatoire et les réseaux électriques. Nous prouvons, au moyen des matrices aléatoires, que dans le régime des grandes et nombreuses données, l'ajout de données non-étiquetées n'améliore pas les performances. Par conséquent, en grande dimension, la régularisation laplacienne est moins performante qu'un simple regroupement spectral (non-supervisé) lorsque de nombreux échantillons non-étiquetés sont disponibles. Ce problème d'inconsistence de la méthode vis-à-vis des données non-étiquetées est étroitement lié au phénomène de "concentration des distances" : les distances entre vecteurs de données de grande dimension ont tendance à devenir indiscernables, comme déjà souligné dans [3, 1]. Nous proposons ici un nouvel algorithme semi-supervisé sur les graphes qui permet un apprentissage consistent des données non-étiquetées en grande dimension, tout en assurant une exploitation efficace des données étiquetées.

2 Régularization Laplacienne

2.1 Prémilinaires

Nous commençons cette section en rappelant les bases de l'apprentissage sur les graphes, avant de discuter le comportement de la régularisation laplacienne sur les données de grande dimension. Considérons un ensemble $\{x_1, \ldots, x_n\} \in \mathbb{R}^p$ de données de dimension p appartenant à l'une des deux classes d'affinité $\mathcal{C}_1$ ou $\mathcal{C}_2$. Dans les méthodes basées sur des graphes, les données $x_1, \ldots, x_n$ sont représentées par des sommets d'un graphe, sur lequel une matrice de poids W est calculée par

$$W = \{w_{ij}\}_{i,j=1}^n = \left\{ h\left(\frac{1}{p}\|x_i - x_j\|^2\right)\right\}_{i,j=1}^n$$

pour une certaine fonction h positive décroissante, de sorte que les vecteurs voisins x_i et x_j soient reliés avec un poids important w_{ij}, à l'image de leur similarité en tant qu'échantillons de données. Une fonction typique pour définir w_{ij} est la fonction

gaussienne $w_{ij} = e^{-\|x_i - x_j\|^2/t}$. La connectivité de l'échantillon x_i est mesurée par son degré $d_i = \sum_{j=1}^n w_{ij}$; la matrice diagonale $D \in \mathbb{R}^{n\times n}$ ayant les d_i comme éléments diagonaux est appelée matrice de degrés.

L'approche par graphe suppose que les points de données appartenant au même groupe d'affinité sont "proches" au sens graphique. En d'autres termes, si $f \in \mathbb{R}^n$ est un vecteur de "signal" (ou de scores) d'appartenance à $\mathcal{C}_1$ ou $\mathcal{C}_2$ pour l'ensemble des échantillons $x_1, \ldots, x_n$, il est supposé varier peu entre ses indices i et j lorsque w_{ij} porte une grande valeur. L'hypothèse de lissage du graphe est généralement définie comme minimisant une pénalité de la forme

$$\frac{1}{2}\sum_{i,j=1}^n w_{ij}(f_i - f_j)^2 = f^\mathsf{T} L f$$

où $L = D - W$ est la matrice laplacienne du graphe. Il existe d'autres variantes de la pénalité de lissage impliquant des formes normalisées de la laplacienne, comme la laplacienne normalisée symétrique $L_s = I_n - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ ou la laplacienne normalisée de la marche aléatoire $L_r = I_n - WD^{-1}$.

En apprentissage semi-supervisé, on dispose de $n_{[l]}$ observations étiquetées $\{(x_1, y_1), \ldots, (x_{n_{[l]}}, y_{n_{[l]}})\}$ avec $y_i \in \{-1, 1\}$ l'étiquette de classe de x_i ainsi que de $n_{[u]}$ données non-étiquetées $\{x_{n_{[l]}+1}, \ldots, x_n\}$. Pour que le signal f soit en accord avec la classe des données étiquetées, la régularisation laplacienne impose des scores déterministes aux points étiquetés de f, par exemple en imposant $f_i = y_i$ pour tout x_i étiqueté. La formulation mathématique du problème devient alors

$$\min_{f\in\mathbb{R}^n} f^\mathsf{T} L f \quad \text{tel que} \quad f_i = y_i, \quad 1 \le i \le n_{[l]}. \tag{1}$$

En écrivant

$$f = \begin{bmatrix} f_{[l]} \\ f_{[u]} \end{bmatrix},\ L = \begin{bmatrix} L_{[ll]} & L_{[lu]} \\ L_{[ul]} & L_{[uu]} \end{bmatrix},$$

ce problème d'optimisation convexe à contraintes d'égalité sur $f_{[l]}$ se résoud en annulant la dérivée de la fonction de perte par rapport à $f_{[u]}$, et donne la solution explicite

$$f_{[u]} = -L_{[uu]}^{-1} L_{[ul]} f_{[l]}.$$

L'étape de décision consiste alors à affecter l'échantillon x_i non-étiqueté à $\mathcal{C}_1$ (resp., $\mathcal{C}_2$) si $f_i < 0$ (resp., $f_i > 0$).

Cette méthode est souvent nommée "régularisation laplacienne" car elle découvre les scores de classes des données non-étiquetées $f_{[u]}$ en les régularisant par la matrice laplacienne et par les contraintes des signaux de classe connus $f_{[l]}$. Il est néanmoins régulièrement observé en pratique que l'utilisation d'autres régularisateurs, tels que $f^\mathsf{T} L_s f$ ou $f^\mathsf{T} L_r f$, permet d'obtenir de meilleurs résultats de classification. Afin d'intégrer tous ces algorithmes de régularisation laplacienne dans un cadre commun, nous définissons $L^{(a)} = I - D^{-1-a} W D^a$ comme la matrice laplacienne $a-$normalisée. Remplacer L par $L^{(a)}$ dans (1) permet alors d'obtenir

$$f_{[u]} = -\left(L_{[uu]}^{(a)}\right)^{-1} L_{[ul]}^{(a)} f_{[l]} \tag{2}$$

où $f_{[l]} = y_{[l]}$ avec $y_{[l]}$ le vecteur d'étiquettes composé des y_i pour $1 \le i \le n_{[l]}$. On retrouve les solutions des laplaciennes standard L, symétrique L_s et de la marche aléatoire L_r respectivement pour $a = 0$, $a = -1/2$ and $a = -1$.

2.2 Apprentissage en grande dimension

Comme dans l'analyse de [4], nous adoptons ici le modèle de mélange présenté sous l'Hypothèse 1 pour lequel le problème d'apprentissage est "asymptotiquement non-trivial" pour des données de grande dimension, à savoir :

Hypothèse 1. *Les échantillons de données $x_1, \ldots, x_n \in \mathbb{R}^p$ sont des observations i.i.d. du modèle génératif tel que, pour $k \in \{1, 2\}$, $\mathbb{P}(x_i \in \mathcal{C}_k) = \rho_k$ et*

$$x_i \in \mathcal{C}_k \Leftrightarrow x_i \sim \mathcal{N}(\mu_k, C_k).$$

avec $\|C_k\| = O(1)$, $\|C_k^{-1}\| = O(1)$, $\|\mu_2 - \mu_1\| = O(1)$, $\mathrm{tr}(C_1 - C_2) = O(\sqrt{p})$ et $\mathrm{tr}(C_1 - C_2)^2 = O(\sqrt{p})$.

Les ratios $c_0 = \frac{n}{p}$, $c_{[l]} = \frac{n_{[l]}}{p}$ et $c_{[u]} = \frac{n_{[u]}}{p}$ sont uniformément bornés dans $(0, +\infty)$ pour p arbitrairement grand.

Sous l'Hypothèse 1, il est facilement démontré que, pour tout $i, j \in \{1, \ldots, n\}$, nous avons la surprenante convergence (un témoin de la malédiction de la dimension)

$$\frac{1}{p}\|x_i - x_j\|^2 = \tau + o_p(1), \quad \tau \equiv \frac{1}{p}\mathrm{tr}(\rho_1 C_1 + \rho_2 C_2) \tag{3}$$

indépendamment de la classe des x_i. Ainsi, pour tout i,

$$f_i = (c_{[l]}/c_0)(\rho_2 - \rho_1) + o_p(1) \tag{4}$$

de sorte que toutes les données non-étiquetées seront affiliées à la classe contenant le plus de données étiquetées, à moins de normaliser les scores déterministes des données étiquetées de sorte qu'ils soient équilibrés par classe, à savoir

$$f_{[l]} = \left(I_{n_{[l]}} - \frac{1}{n_{[l]}} 1_{n_{[l]}} 1_{n_{[l]}}^\mathsf{T}\right) y_{[l]}. \tag{5}$$

En utilisant (5), nous éliminons le terme dominant $\rho_2 - \rho_1$ de (4) et il s'agit alors d'étudier les termes d'ordre plus petit de f_i.

Proposition 1. *Sous l'Hypothèse 1, pour h trois-fois continûment dérivable dans un voisinage de τ, l'entrée $i > n_{[l]}$ du vecteur des scores des données non-étiquetées $f_{[u]}$ (défini par* (2) *avec $f_{[l]}$ donné par* (5)*) satisfait*

$$\sqrt{p} f_i = 2\rho_1\rho_2 \frac{c_{[l]}}{c_0}(1 + a)\frac{h'(\tau)}{h(\tau)}\frac{\mathrm{tr}C_2 - \mathrm{tr}C_1}{\sqrt{p}} + o_p(1).$$

La proposition ci-dessus indique que même avec $f_{[l]}$ équilibré par (5), le problème d'affectation de toutes les données non-étiquettées dans la même classe persiste (si $\mathrm{tr}C_2 \neq \mathrm{tr}C_1$) pour toutes les régularisations laplaciennes de la forme $L^{(a)} = I - D^{-1-a} W D^a$, à moins que $a \simeq -1$. En prenant $a = -1$, on affine alors l'expression de f_i comme suit :

Proposition 2. *Sous les conditions et notations de la Proposition 1 avec $a = -1$, pour $i > n_{[l]}$ et $x_i \in \mathcal{C}_k$,*

$$pf_i = g_i + o_p(1)$$

où $g_i \sim \mathcal{N}((-1)^k(1-\rho_k)m, \sigma_k^2)$ avec $r_k = \frac{\sigma_k^2}{m^2}$ une function strictement décroissante de $c_{[l]}$, mais indépendante de $c_{[u]}$.

Malgré la performance "raisonnable" de classification pour $a = -1$ (au contraire d'autres régularisations), le résultat-clé de la Proposition 2 est l'incapacité de la laplacienne régularisée à exploiter des données *non-étiquetées* supplémentaires, même en nombre considérable. Par contre, l'addition de données *étiquetées* est toujours bénéfique. Ce résultat va dans le sens de nombreuses remarques sur les limitations connues de l'apprentissage semi-supervisé mais explique ici pourquoi la régularisation laplacienne peut parfois être surpassée par des méthodes purement non-supervisées (telles que le regroupement spectral), comme observé dans la Figure 1.

3 Méthode Proposée

3.1 Motivation et algorithme

Evidemment, un apprentissage semi-supervisé est efficace si les scores f_i des données non-étiquetées exploitent à la fois les similarités w_{ij} croisées entre données étiquetées et non-étiquetées et les similarités entre données non-étiquetées. De ce point de vue, la solution $f_{[u]}$ donnée par (2) peut être découpée en deux opérations : les étiquettes connues se propagent tout d'abord vers les points non-étiquetés par l'action $s_{[u]} = -L_{[ul]}^{(a)} f_{[l]} = \left(D^{-1-a}WD^a\right)_{[ul]} f_{[l]}$, puis ce signal $s_{[u]}$ est "affiné" par multiplication à gauche par $L_{[uu]}^{(a)-1}$ dans le but d'exploiter les informations globales du sous-graphe des données non-étiquetées. Toutefois, en raison de la concentration des distances (caractérisée par (3)), nous observons que

$$L_{[uu]}^{(a)-1} s_{[u]} \simeq s_{[u]} + \frac{1}{n_{[l]}}(1_{n_{[u]}}^\mathsf{T} s_{[u]}) 1_{n_{[u]}}, \tag{6}$$

ce qui signifie qu'en grande dimension, l'opération $L_{[uu]}^{(a)-1} s_{[u]}$ ne fait qu'amplifier le signal constant $1_{n_{[u]}}$, sans effet sur les données. Ici réside la raison centrale de l'inefficacité de la régularisation laplacienne dans l'exploitation des informations non-étiquetées.

Pour combattre l'effet délétère de la concentration des distances à la source du problème, nous proposons d'utiliser une matrice de poids recentrée $\hat{W} \in \mathbb{R}^{n\times n}$ de la forme

$$\hat{W} = PWP \quad \text{avec} \quad P \equiv I_n - \frac{1}{n} 1_n 1_n^\mathsf{T}. \tag{7}$$

La propriété principale de $\hat{W}$ est d'être orthogonale au vecteur 1_n qui dominait jusqu'ici le comportement asymptotique de $L_{[uu]}^{(a)-1}$, avec (6) pour résultat. Par ailleurs, l'opération de recentrage préserve les distances moyennes entre similarités intra- et inter-classes induites par W et n'affecte donc pas les informations nécessaires à la classification. Cependant, $\hat{W}$ comporte maintenant des éléments positifs et négatifs, rendant l'optimisation de la pénalité de lissage (1) non nécessairement convexe. Ce problème est résolu en imposant la norme de f. Le problème d'optimisation devient alors :

$$\min_{f_{[u]} \in \mathbb{R}^{n_{[u]}}} -f^\mathsf{T} \hat{W} f \quad \text{tel que} \quad \|f_{[u]}\|^2 = n_{[u]} e^2. \tag{8}$$

La solution du problème nécessite l'introduction d'un multiplicateur lagrangien α associé à la contrainte de norme $\|f_{[u]}\|^2 = n_{[u]}e^2$. En nommant $\hat{f}_{[u]}$ la solution de (8), on obtient :

$$\hat{f}_{[u]} = \left(\alpha I_{n_{[u]}} - \hat{W}_{[uu]}\right)^{-1} \hat{W}_{[ul]} f_{[l]} \tag{9}$$

où α est déterminé par $\alpha > \|\hat{W}_{[uu]}\|$ et $\|\hat{f}_{[u]}\|^2 = n_{[u]}e^2$.

Remarque 1 (Recentrage de W)**.** *Au contraire d'une simple opération de recentrage de tous les éléments de W par $h(\tau)$, l'opération $\hat{W} = PWP$ a l'avantage supplémentaire d'équilibrer les poids positifs et négatifs résultants, à savoir que, pour tout $i \in \{1, \ldots, n\}$, $d_i = \sum_{j=1}^n w_{ij} = 0$. De cette façon, nous éliminons le risque d'un comportement instable de $f_{[u]}$ causé par des degrés potentiellement négatifs.*

3.2 Performance en grande dimension

Nous donnons dans cette section les propriétés théoriques de la régularization recentrée qui soutiennent son utilisation. D'abord, le résultat montrant que la méthode proposée permet un vrai apprentissage semi-supervisé pour les données de grande dimension est présenté dans Proposition 3.

Proposition 3. *Sous les conditions de Propostion 1, soit $\hat{f}_{[u]}$ défini par* (9)*, $f_{[l]}$ donné par* (5)*), $\alpha \in (\hat{W}_{[uu]}, \infty)$. Alors, pour $i > n_{[l]}$ et $x_i \in \mathcal{C}_k$,*

$$\hat{f}_i = \hat{g}_i + o_p(1)$$

où $\hat{g}_i \sim \mathcal{N}((-1)^k(1-\rho_k)\hat{m}, \hat{\sigma}_k^2)$ avec $\hat{r}_k = \frac{\hat{\sigma}_k^2}{\hat{m}^2}$ une function strictement décroissante de $c_{[l]}$ et de $c_{[u]}$.

En plus d'apprendre efficacement des données étiquettées et non-étiquettées en grande dimension, la regularization recentrée a une supériorité garantie sur la régularization Laplacienne, comme déclaré dans Proposition 4.

Proposition 4. *Sous les conditions et notations de Proposition 2 et de Proposition 3* [1]*,*

$$\lim_{\alpha \to +\infty} \hat{r}_k = r_k$$

Conformément aux résultats théoriques, on observe en Figure 1 que la classification par régularisation laplacienne n'augmente pas avec le nombre des données non-étiquetées, et se trouve dépassée par le regroupement spectral pour $c_{[u]}$ large. La régularisation recentrée ne souffre pas de cette limitation et domine les deux méthodes.

1. Plus précisément, nous demandons ici que $\hat{f}_{[u]}$ soit donée par (9) où $\hat{W} = PWP$ avec $W_{ij} = \hat{h}(\frac{1}{p}\|x_i - x_j\|^2)$ pour $\hat{h}$ satisfaisant que $\hat{h}''(\tau)/\hat{h}'(\tau) = h''(\tau)/h'(\tau) - h'(\tau)/h(\tau)$.Cette condition peut être enlevée quand $C_1 = C_2$.

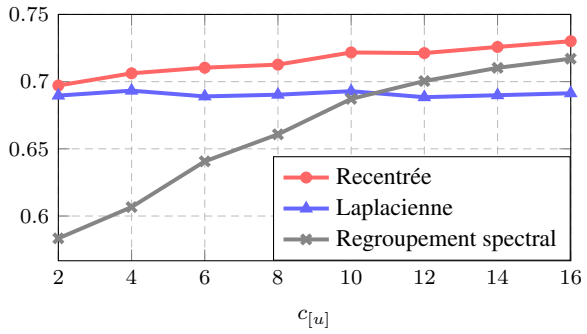

FIGURE 1 – Exatitude en fonction de $c_{[u]}$ pour les données gaussiennes avec $p = 80$, $h(t) = e^{-t}$. Moyennée sur $50000/n_{[u]}$ itérations.

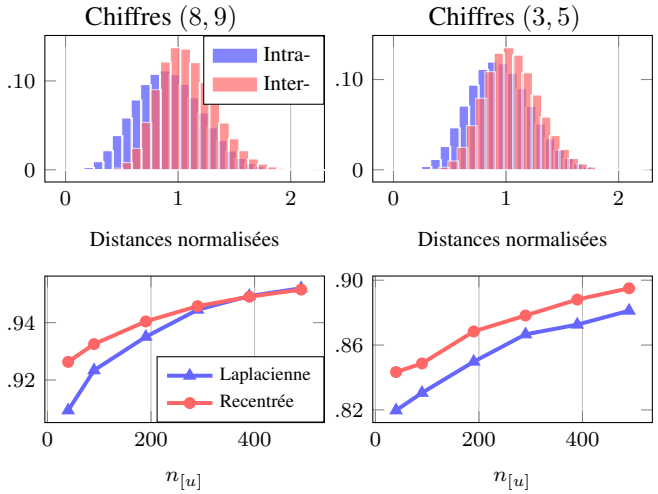

FIGURE 2 – En haut : distribution des distances normalisées pour les données MNIST à 2 classes. En bas : exatitude en fonction de $n_{[u]}$ avec $n_{[l]}$=10, moyennée sur 1000 itérations.

4 Étude numérique

Pour confirmer la puissance de $\hat{W}$, nous testons les deux méthodes sur les données de la base MNIST. Pour être exhaustifs, les résultats sont obtenus sur les graphes les plus performants pour les deux méthodes [2]. Les hyperparamètres des algorithmes (a pour la régularisation laplacienne et α pour la régularisation recentrée) sont optimisés par réglage fin.

La Figure 2 montre qu'une grande précision de classification est facilement obtenue avec les données MNIST, même avec l'approche laplacienne classique, mais moindre que la méthode proposée. L'avantage de l'algorithme proposé est également plus perceptible sur la tâche de classification pour laquelle la différence entre les distances intra- et inter-classes est moins marquée. Pour illustrer davantage l'impact de la concentration des distances, la Figure 3 présente des situations dans lesquelles le problème d'apprentissage devient plus difficile en présence d'un bruit additif. Le phénomène de concentration des distances est ainsi plus accentué et résulte en des gains de performance considérables. Par ailleurs, lorsque les informations de similarité sont sérieusement perturbées par le bruit additif, l'effet anticipé de saturation des performances de la régularisation laplacienne lorsque $n_{[u]}$ augmente est très marqué. Tous ces résultats suggère que la régularisation recentrée est une solution privilégiée dans toutes ces situations, mais est particulièrement bénéfique dans la situation difficile d'une différence subtile entre similarités intra- et inter-classes.

2. Sélectionnés parmi les graphes les plus utilisés, dont KNN avec $k \in \{2^1, \ldots, 2^q\}$ voisins, pour q le plus grand entier tel que $2^q < n$, et le graphe gausien $w_{ij} = e^{-\|x_i - x_j\|^2/\sigma^2}$ avec σ la distance moyenne entre les données.

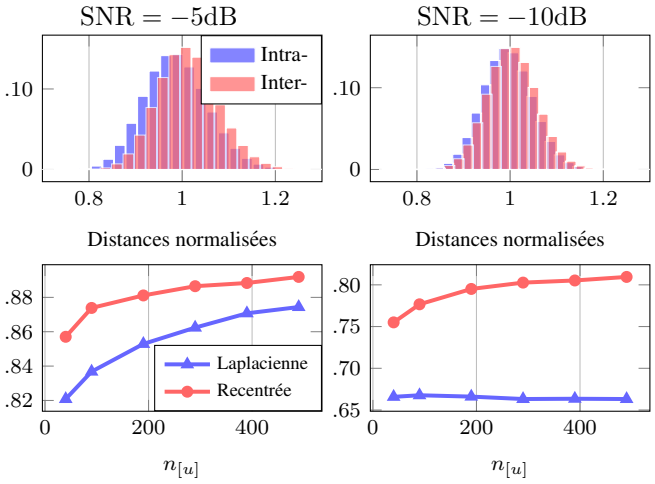

FIGURE 3 – En haut : distribution des distances normalisées pour les données MNIST $(8, 9)$ bruitées. En bas : exatitude en fonction de $n_{[u]}$ avec $n_{[l]}$=10, moyennée sur 1000 itérations.

Références

[1] Fabrizio Angiulli. On the behavior of intrinsically high-dimensional spaces : Distances, direct and reverse nearest neighbors, and hubness. *Journal of Machine Learning Research*, 18(170) :1–60, 2018.

[2] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

[3] Damien Francois, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7) :873–886, 2007.

[4] Xiaoyi Mai and Romain Couillet. A random matrix analysis and improvement of semi-supervised learning for large dimensional data. *The Journal of Machine Learning Research*, 19(1) :3074–3100, 2018.

[5] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.