Application de la théorie des grandes matrices aléatoires à l'apprentissage pour les mégadonnées

Romain COUILLET[1]

[1]L2S, CentraleSupélec, CNRS, Université PSud, Gif-sur-Yvette, France*
romain.couillet@centralesupelec.fr

Résumé – Cet article présente les derniers résultats dans le domaine de la théorie des matrices aléatoires pour la compréhension et l'amélioration des méthodes et algorithmes d'apprentissage automatisé. Un exposé bref de ces résultats sera proposé, notamment dans le contexte des méthodes dites spectrales (méthodes à noyaux, classification non supervisée ou semi-supervisée, détection de communautés sur les graphes, machines à vecteurs de support) ainsi que dans le contexte des réseaux de neurones.

Abstract – This article introduces the recent findings in the field of random matrix theory for the understanding as well as the improvement of statistical machine learning methods and algorithms. A brief overview of these results will be proposed, notably in spectral methods (kernel approaches, clustering, community detection on graphs, semi-supervised learning and SVM) as well as neural network considerations.

1 Introduction

Le récent regain d'intérêt, dans quasiment l'ensemble de la communauté du traitement du signal, pour les méthodes d'apprentissage (particulièrement des réseaux de neurones profonds) ouvre aujourd'hui la voie à de nouvelles questions théoriques concernant une nouvelle compréhension de ces méthodes. En effet, à l'ère des mégadonnées (le dit "big data"), l'apprentissage automatisé doit être effectué sur des données nombreuses et de grandes tailles. Comme nous le verrons par la suite, nombre des méthodes standard (notamment les méthodes à noyaux) sont issues de l'hypothèse de données nombreuses mais de petites tailles, et les intuitions liées à ces méthodes (souvent heuristiques) ne sont plus valables en grandes dimensions.

L'objectif de cet article est d'effectuer un tour d'horizon, à l'aide de l'outil de la théorie des matrices aléatoires, de la compréhension nouvelle qu'apporte l'analyse en grandes dimensions des méthodes classiques d'apprentissage. Nous vérifierons notamment la perte de consistance de beaucoup de ces méthodes que nous parviendrons à améliorer grâce à cette analyse, donnant lieu parfois à de nouvelles méthodes originales dont la consistance est assurée dans la double asymptotique sur les taille et nombre de données.

En petite dimension, les méthodes d'apprentissage souffrent notamment d'une impossibilité d'analyse chronique liée à la non linéarité des structures impliquées. Assez étonnamment, dans le régime des matrices aléatoires de grandes tailles, un effet de concentration de la mesure survient qui permet de briser cette difficulté et de donner lieu à une compréhension parfois complète de méthodes restées jusque-là très heuristiques. Nous avons en particulier bon espoir que les études préliminaires discutées ici puissent ouvrir la voie à un nouveau paradigme d'analyse et de perfectionnement de l'apprentissage automatisé pour les mégadonnées.

*Ce travail est soutenu par le projet ANR RMT4GRAPH (ANR-14-CE28-0006).

2 Détection de communautés sur les graphes réalistes

Un premier exemple d'étude menée récemment par nos soins [1, 2] concerne l'analyse de méthodes de détections de communautés sur les grands graphes. Ce domaine très à la mode, de par ses implications pratiques à fort impact économique notamment (publicités ciblées, détection de groupes dans des grands réseaux sociaux), suscite particulièrement l'intérêt des mathématiciens et théoriciens des graphes qui ont développé et prouvé la consistance de nombreuses méthodes rapides et puissantes de classification [3, 4]. Cependant, un dénominateur commun à ces études est qu'elles reposent presque toutes sur le "modèle jouet" dit modèle stochastique par bloc (SBM). Sous cette hypothèse, les n nœuds du graphe appartiennent à l'une de k communautés (ou groupes). La probabilité de connexion des nœuds i et j, qui appartiennent aux groupes g_i et g_j, respectivement, est dénotée $C_{g_i g_j}$. S'ensuit le modèle matriciel aléatoire de la matrice d'adjacence $A \in \mathbb{R}^{n\times n}$ du graphe satisfaisant $A_{ij} \sim \text{Bernoulli}(C_{g_i g_j})$.

Ce modèle donne lieu à un certain nombre d'intuitions et conclusions précises importantes. Notamment, lorsque $n \to \infty$ de telle façon que $C_{g_i g_j} = O(1)$ (générant des

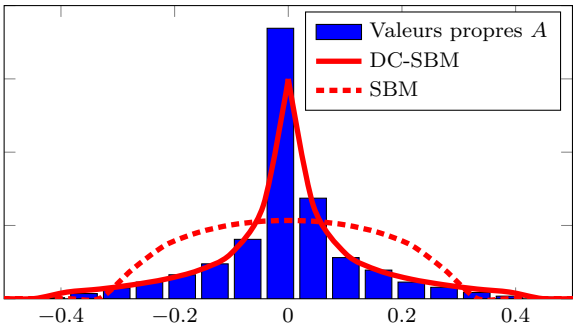

FIG. 1: Valeurs propres de A pour un graphe réaliste (PolBlogs) et lois limites des modèles SBM et DC-SBM.

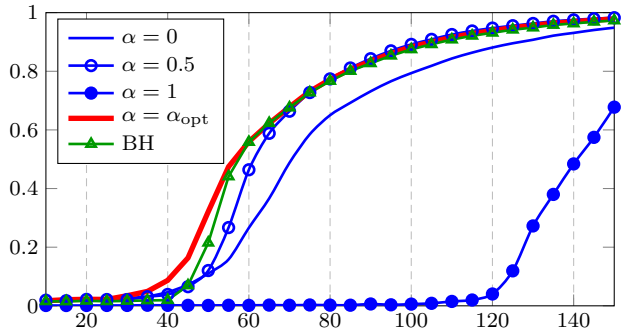

FIG. 2: Taux de classification des méthodes spectrales en fonction de la "distance" entre communautés.

graphes denses), alors un régime de classification asymptotique non trivial (à savoir qui donne lieu à une classification limite ni impossible, ni parfaite) apparaît lorsque $|C_{ab} - C_{cd}| = O(n^{-\frac{1}{2}})$ pour tous groupes a, b, c, d. Dans ce cas, une transition de phase est même observée (et démontrée) lorsque les C_{ab} satisfont une certaine inégalité: lorsqu'ils sont trop proches, aucune communauté ne peut être détectée asymptotiquement, mais au delà d'un seuil, une détection non nulle devient possible.

L'hypothèse du SBM est mathématiquement utile mais en aucun cas pratique, en ce sens qu'elle ne tient pas compte de l'hétérogénéité possible des degrés des nœuds. Un modèle bien plus approprié, de notre point de vue, est le modèle DC-SBM qui corrige les degrés en supposant que $A_{ij} \sim \text{Bernoulli}(q_i q_j C_{g_i g_j})$ avec $q_i > 0$ une probabilité intrinsèque pour le nœud i de se connecter à tout autre nœud du graphe, indépendamment de la structure de communauté. La figure 1 confirme la bien meilleure adéquation du modèle DC-SBM par rapport au modèle SBM dans le cas du graphe réel PolBlogs, à travers l'illustration de la loi limite des valeurs propres évaluée pour chaque modèle et comparée à l'histogramme des valeurs propres de A. Sous cette hypothèse, peu étudiée par les mathématiciens, les conclusions sont cependant moins immédiates.

Dans notre étude [1], nous avons tout d'abord démontré l'inconsistance des méthodes spectrales basées sur A pour le modèle DC-SBM. Ces méthodes spectrales consistent en une identification des groupes à travers une classification des entrées des vecteurs propres dominants de A. Nous avons alors démontré également dans [1] que ces méthodes spectrales ne sont efficaces que si A est remplacée par la matrice $L_1 = D^{-1}(A - dd^{\mathsf{T}}/(2m))D^{-1}$, avec D la matrice diagonale des degrés du graphe ($D_{ii} = \sum_j A_{ij}$), d le vecteur des éléments diagonaux de D et $m = \frac{1}{2}\sum_i d_i$. La difficulté technique ici consiste à lever la dépendance non linéaire intervenant maintenant dans les entrées de la matrice L_1, du fait de la dépendance entre D et A; cette opération est effectuée via une approximation asymptotique de L_1 par une matrice $\tilde{L}_1$ (et qui satisfait donc $\|L_1 - \tilde{L}_1\| \xrightarrow{\text{p.s.}} 0$) non observable mais dont les propriétés matricielles sont bien plus accessibles que celles de L_1.

Dans un second article [2], nous avons alors compris que L_1 avait souvent un spectre de valeurs propres qui masquait les communautés, induisant donc une contradiction sévère entre la nécessité de travailler avec L_1 au lieu de A et la capacité de chacune de ces matrices à isoler les informations des communautés. Nous avons ainsi considéré une matrice plus générale, définie par

$$L_\alpha \equiv D^{-\alpha}\left(A - \frac{dd^{\mathsf{T}}}{2m}\right)D^{-\alpha}.$$

Nous avons alors démontré que, pour chaque $\alpha \in \mathbb{R}$, il existe un seuil au delà duquel il est possible d'obtenir une classification non triviale, non pas à partir des vecteurs propres u_i de L_α mais à partir des vecteurs renormalisés $D^{\alpha-1}u_i$ (on retrouve en particulier ici le résultat selon lequel $\alpha = 1$ permet une classification consistante à partir des u_i). Comme ce seuil dépend de la valeur prise par α, il existe un choix optimal α_{opt} pour lequel $L_{\alpha_{\text{opt}}}$ apporte les meilleures performances. Nous avons démontré alors que α_{opt} dépend uniquement de la distribution empiriques des q_i et peut être estimé de manière consistante uniquement grâce au vecteur d, par le biais d'un algorithme somme toute peu coûteux mais reposant essentiellement sur la théorie des matrices aléatoires. La figure 2 présente les performances des algorithmes de classification pour différentes valeurs de α, y compris la valeur optimale α_{opt} (ici égale à $0,27$), ainsi que pour la méthode aujourd'hui populaire du Bethe Hessian (BH) [5]; cette figure confirme l'avantage substantiel d'un choix pertinent pour α.

3 Méthodes à noyaux

Les méthodes à noyaux consistent en un ensemble d'outil de classification et plus généralement de traitement de données. Pour de telles données $x_1, \ldots, x_n \in \mathbb{R}^p$, ces méthodes sont basées sur une matrice $K = \{\kappa(x_i, x_j)\}_{i,j=1}^n$ pour une certaine fonction $\kappa(x, y)$ souvent prise égale à $f(\|x_i - x_j\|^2)$ ou $f(x_i^{\mathsf{T}} x_j)$ pour un choix donné de f. Dans un cadre de classification non supervisée, les méthodes spectrales consisteront alors à extraire les vecteurs propres

dominants de K [6]; pour les approches d'apprentissage semi-supervisé, il sera plutôt question d'appliquer un algorithme de marche aléatoire ou de propagation d'étiquettes sur le graphe d'adjacence donné par K [7]; enfin, pour les machines à vecteurs de support, on s'intéressera à un problème d'optimisation centré sur K [8].

Dans tous ces cas de figure, l'analyse des performances des algorithmes est particulièrement compliquée, du fait notamment de la non-linéarité précisément induite par le noyau. Dans une série d'études [9, 10, 11], nous nous sommes placés dans le cadre d'un mélange gaussien, en ce sens que les données x_i appartiennent à k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ avec l'hypothèse que $x_i \in \mathcal{C}_a \Leftrightarrow x_i \sim \mathcal{N}(\mu_a, C_a)$. De même qu'en Section 2, nous avons également imposé que les μ_a et C_a soient suffisamment "proches" (dans un sens décrit précisément dans ces travaux) de sorte que les classes restent asymptotiquement non-trivialement observables. Sous ces hypothèses, nous avons alors démontré que, si $n, p \to \infty$ avec $p/n \to c \in (0, \infty)$, alors pour toute paire (x_i, x_j), $\|x_i - x_j\|^2 \to \tau$ pour un certain $\tau > 0$ qui ne dépend pas des classes des x_i, x_j. Ce phénomène surprenant, issu des asymptotiques en grande dimension, rend possible la linéarisation de la matrice à noyau K, avec $K(x_i, x_j) = f(\|x_i - x_j\|^2)$. On obtient ainsi l'approximation

$$\|K - \tilde{K}\| \xrightarrow{\text{p.s.}} 0$$

dans la limite des n, p large, avec $\tilde{K}$ une matrice qui s'exprime en fonction des paramètres μ_a et C_a des classes mais aussi en fonction des dérivées $f(\tau), f'(\tau), f''(\tau)$ de la fonction f. Une conclusion intéressante est notamment qu'un polynôme d'ordre 2 ayant les mêmes dérivées en τ que f conduira asymptotiquement aux mêmes performances des algorithmes à noyaux basés sur K. De ce résultat, il est permis notamment de déduire des choix de fonctions f ayant des propriétés privilégiées pour remplir des tâches spécifiques. Notamment, nous avons remarqué l'étonnante propriété qui prédit que, pour des données normalisées de classes ayant les mêmes moyennes mais des covariances distinctes, le choix $f'(\tau) = 0$ est optimal et conduit même à un changement radical de régime en ce sens qu'on peut assurer une classification asymptotique parfaite sous cette hypothèse [12].

Mais un résultat encore plus surprenant vient de la comparaison de cette étude théorique avec le cas pratique de données réelles. Lorsque les $x_i \in \mathbb{R}^p$ sont maintenant des images de dimension $\sqrt{p} \times \sqrt{p}$ vectorisées, nous avons constaté des similarités étonnantes avec les comportements attendus dans le cas de mélanges gaussiens, comme en témoigne la figure 3 sur les données de la base MNIST.

4 Réseaux de neurones

La dernière étude que nous présentons [14] concerne une première analyse des réseaux de neurones aléatoires de type "extreme learning machine" [15], aussi liés aux dites

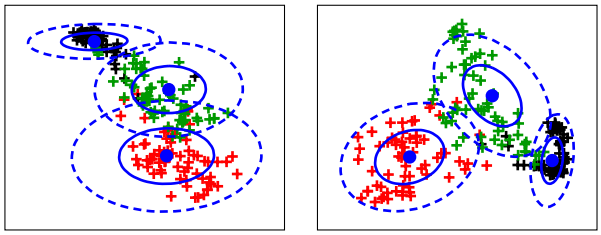

FIG. 3: Classification non supervisée de 3 classes d'images MNIST [13] (64 images par classes représentées en couleur), et centroids/écarts types prévus par la théorie sur un mélange gaussien (en bleu). Gauche: 1er/2e vecteurs propres; Droite: 2e/3e vecteurs propres.

"random feature maps" popularisées dans [16]. L'idée ici est de considérer un réseau de neurones à une seule couche cachée dont la matrice de connexion entre entrées et couche cachée est choisie aléatoirement et fixée. La matrice de sortie est alors obtenue par régression linéaire régularisée. Ce réseau de neurones se modélise comme:

$$\hat{Y} = B^\mathsf{T} \sigma(WX)$$

où $X = [x_1, \ldots, x_T] \in \mathbb{R}^{p \times T}$ sont les T données d'entrée, $W \in \mathbb{R}^{n \times p}$ est la matrice aléatoire de connexion entre l'entrée et les neurones, $\sigma(\cdot)$ est la fonction d'activation non linéaire qui opère entrée par entrée et B est la matrice de régression (régularisée par un coût ℓ_2) obtenue par

$$B \equiv \frac{1}{T} \sigma(WX) \left(\frac{1}{T} \sigma(WX)^\mathsf{T} \sigma(WX) + \gamma I_T \right)^{-1} Y^\mathsf{T}$$

où $Y = [y_1, \ldots, y_T] \in \mathbb{R}^{d \times T}$ est le vecteur de sorties associées lors de la phase d'entrainement aux entrées $X = [x_1, \ldots, x_T]$ et $\gamma > 0$ est un paramètre de calibrage utilisé afin d'éviter le sur-apprentissage.

Dans ce contexte, la compréhension des performances asymptotiques du réseau de neurones passe par une analyse de la matrice $G = \frac{1}{T} \sigma(WX)^\mathsf{T} \sigma(WX)$. Nous sommes parvenus à caractériser finement cette matrice dans [14]. Ici, contrairement aux travaux évoqués dans la Section 3, il n'est pas possible de linéariser la fonction $\sigma(\cdot)$ (qui nous placerait alors dans un cas réellement trivial). Les outils classiques de la théorie des matrices aléatoires sont également a priori inadaptés car ne considèrent que des modèles de matrices à entrées linéairement dépendantes ou indépendantes. Heureusement, nous pouvons exploiter ici le fait que les lignes de $\sigma(WX)$ restent indépendantes (ce qui n'est pas le cas des colonnes) et pouvons par ailleurs exploiter un phénomène de concentration de la mesure sur chacune des lignes. Grâce à cette double propriété, nous avons pu prouver que la matrice G se comporte asymptotiquement de manière similaire à une matrice de covariance empirique de covariance $\Phi \equiv \mathrm{E}_w[\sigma(X^\mathsf{T} w)\sigma(w^\mathsf{T} X)]$ avec $w \sim \mathcal{N}(0, I_p)$. Ce résultat permet alors d'exploiter des outils connus de la théorie des matrices aléatoires afin d'anticiper les performances du réseau de neurones,

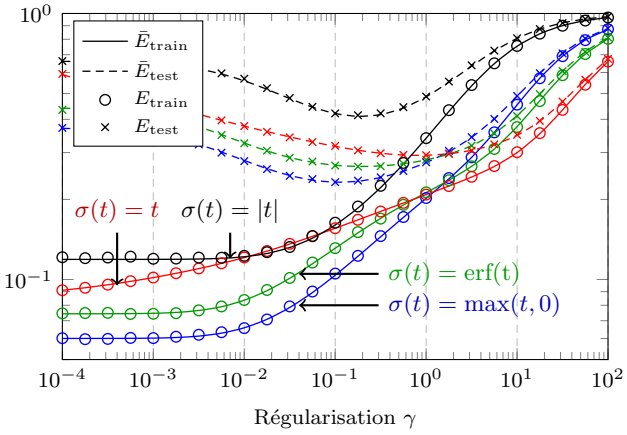

FIG. 4: Erreurs quadratiques pendant l'entrainement (E_{train}) et le test (E_{test}) d'images de la base MNIST, comparées aux approximations théoriques ($\bar{E}_{\text{train}}$ et $\bar{E}_{\text{test}}$).

comme le démontre la figure 4 qui présente les performances en termes d'erreur quadratique moyenne, pendant les phases d'entrainement et de test, d'un tel réseau de neurones pour la classification de deux classes d'images MNIST (à travers une régression vers les valeurs 1 et -1).

5 Conclusion

Cet article donne un bref aperçu des potentialités de la théorie des matrices aléatoires à comprendre et améliorer les algorithmes standard de l'apprentissage automatisé. Il ouvre également la voie, de par les multiples intuitions tirées de ces études préliminaires, à un meilleur usage qualitatif mais également quantitatif de ces méthodes. À l'heure où les outils d'apprentissage profond très mal compris prennent le pas sur les méthodes expertes au sein du traitement du signal et des données, il paraît essentiel de promouvoir ces nouvelles approches et d'exploiter au maximum les messages qu'elles peuvent délivrer sur des modèles d'apprentissage plus complexes.

References

[1] H. Tiomoko Ali and R. Couillet, "Performance analysis of spectral community detection in realistic graph models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'16)*, 2016.

[2] H. Tiomoko Ali and R. Couillet, "Random matrix improved community detection in heterogeneous networks," in *Proc. IEEE Asilomar Conference on Signals, Systems, and Computers*, (Pacific Grove, CA, USA), 2016.

[3] C. Bordenave, M. Lelarge, and L. Massoulié, "Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs," in *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pp. 1347–1357, IEEE, 2015.

[4] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20935–20940, 2013.

[5] A. Saade, F. Krzakala, and L. Zdeborová, "Spectral clustering of graphs with the bethe hessian," in *Advances in Neural Information Processing Systems*, pp. 406–414, 2014.

[6] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[7] K. Avrachenkov, P. Gonçalves, A. Mishenin, and M. Sokol, "Generalized optimization framework for graph-based semi-supervised learning," in *Proceedings of SIAM Conference on Data Mining (SDM 2012)*, vol. 9, SIAM, 2012.

[8] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, and T. Van Gestel, *Least squares support vector machines*, vol. 4. World Scientific, 2002.

[9] R. Couillet and F. Benaych-Georges, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[10] R. C. Zhenyu Liao, "A large dimensional analysis of least squares support vector machines," *Journal of Machine Learning Research*, 2017.

[11] X. Mai and R. Couillet, "The counterintuitive mechanism of graph-based semi-supervised learning in the big data regime," in *(submitted to) IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, (New Orleans, USA), 2017.

[12] R. Couillet and A. Kammoun, "Random matrix improved subspace clustering," in *2016 Asilomar Conference on Signals, Systems, and Computers*, 2016.

[13] Y. LeCun, C. Cortes, and C. Burges, "The MNIST database of handwritten digits," 1998.

[14] C. Louart, Z. Liao, and R. Couillet, "A random matrix approach to neural networks," *arXiv preprint arXiv:1702.05419*, 2017.

[15] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 2, pp. 513–529, 2012.

[16] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, pp. 1177–1184, 2007.