Master Mathématiques appliquées et statistique - Université de Bordeaux

Année 2023-2024

Projet d'expertise en Statistique et Probabilités – Parcours Modélisation Statistique et Stochastique et CMI ISI – 1ère année

Sujets à prendre en groupe de 3 étudiants.

Le projet dit d'expertise en statistique et probabilités se déroule pendant tout le 2ème semestre de Janvier à Mai. **En particulier son déroulement se poursuit à temps plein après la session d'examens du 2ème semestre**. Un groupe doit être formé de 3 étudiants qui travailleront en équipe sur le sujet qui leur a été affecté.

L'objectif de ce projet (sous la direction d'un ou plusieurs membres de l'équipe pédagogique) consiste en un travail d'initiation à la recherche en probabilités ou statistique de nature théorique et/ou appliqué. Il peut se baser sur la modélisation et la mise en place de méthodes numériques adaptées à la résolution d'une application concrète liée à une problématique d'analyse de données. Il peut comporter une part importante de programmation. Le projet donne lieu à la rédaction d'un mémoire, et s'achève par une soutenance orale au mois de Mai (la date exacte vous sera communiquée ultérieurement). Un pré-rapport et une soutenance à mi-parcours (fin Mars / début Avril) seront également demandés.

Les groupes doivent être formés et les sujets affectés pour le **lundi 29 janvier 2024** date de début du projet. La procédure pour l'attribution des sujets est la suivante :

- chaque groupe envoie un mail à `jeremie.bigot@u-bordeaux.fr` pour faire part du groupe (3 étudiants), du sujet choisi en préférence, du sujet choisi en deuxième choix, et du sujet choisi en troisième choix avant le **jeudi 25 janvier 2024**,
- les affectations finales seront notifiées par e-mail,
- les étudiants qui ne se sont pas manifestés seront affectés par défaut à un sujet par l'équipe enseignante.

La liste des sujets proposés est la suivante :

1. **Quantiles d'une mesure de probabilité en dimension supérieure ou égale à 2 et applications en statistique**
   Contact : Jérémie Bigot
   Mail : `jeremie.bigot@u-bordeaux.fr`

   Si μ est une mesure de probabilité (discrète ou absolument continue) en dimension 1, c'est à dire dont le support est inclus dans $\mathbb{R}$, alors son quantile q_α d'ordre $0 < \alpha < 1$ peut se définir facilement à l'aide de la formule
   $$q_\alpha = F_\mu^-(\alpha),$$
   où F_μ^- est l'inverse (généralisée) de la fonction de répartition F_μ de μ. De façon plus intuitive, en statistique descriptive et lorsque μ est discrète, on peut aussi définir q_α comme la valeur qui sépare les observations en une proportion α qui sont plus petites que q_α, et une proportion $1-\alpha$ qui sont plus grandes que q_α.

Supposons maintenant que μ soit une mesure de probabilité à support sur $\mathbb{R}^2$, par exemple une Gaussienne de moyenne nulle et de matrice de covariance l'identité, ou bien une mesure discrète associée aux observations du couple de variables $(taille, poids)$ dans une population.

Comment peut-on alors définir la notion de quantile dans ce contexte, étant donné que l'inverse de la fonction de répartition de μ n'est pas clairement définie et qu'il n'existe pas de relation d'ordre canonique sur $\mathbb{R}^2$?

Le but de ce projet est ainsi de s'intéresser à la généralisation de la notion de quantile en dimension 1 pour des mesures de probabilités à support sur $\mathbb{R}^d$ avec $d \geqslant 2$. Il est proposé de s'intéresser à de récentes contributions [1,2,3] en statistique sur la définition d'une notion de quantiles multivariés (qui étend la notion en dimension 1) à l'aide d'outils issus de la théorie du transport optimal de mesures de probabilités qui est un domaine de recherche actuellement un essor très important avec de nombreuses applications en science des données et apprentissage automatique.

Le travail consistera en :

— une lecture de l'article [1] qui introduit la notion de quantile par transport optimal. Vous serez bien sûr guidé dans la compréhension de ce travail de recherche, une source utile pourra être ce rapport (en français) de Master en Sciences Mathématiques de l'Université de Liège :

https://matheo.uliege.be/bitstream/2268.2/4964/4/memoire_TixhonStephanie.pdf

— une implémentation en R ou Python de la notion de quantile multivarié proposée dans cet article,

— une application à l'étude de données anthropométriques (relative à des caractéristiques de taille et forme du corps humain) dont on pourra trouver une description dans [3] et ici https://www.openlab.psu.edu/ansur2/

Références bibliographiques

[1] V. Chernozhukov, A. Galichon, M. Hallin, and M. Henry. Monge-Kantorovich depth, quantiles, ranks and signs. Annals of Statistics, 45(1) : 223-256, 2017.
[2] G. Carlier, V. Chernozhukov, and A. Galichon. Vector quantile regression : An optimal transport approach. Annals of Statistics, 44(3) :1165-1192, 2016.
[3] Gwendoline De Bie. Apprentissage sur l'espace des mesures : réseaux profonds invariants et régression de quantile. Thèse à l'Université Paris sciences et lettres, 2020.

2. Introduction aux Monte Carlo Tree search

Contact : Luis Fredes
Mail : luis.fredes@u-bordeaux.fr

Les algorithmes Monte Carlo Tree Seach sont utilisés dans quelques processus de décision, par exemple, dans les jeux comme le go, les échecs, entre autres. Ces algorithmes ont des applications notamment dans le domaine de réseaux de neurones où ils ont été utilisés dans les jeux de société.

Pour un jeu donné, l'idée consiste qu'à chaque étape du jeux on va explorer à l'aide de Monte Carlo les possibles résultats du jeu a base de simulations. Plus précisément, on simule, à partir d'une configuration du jeu, une suite de mouvements aléatoires, jusqu'à la fin du jeu, puis on garde les statistiques des scores finaux sur les simulations pour ensuite choisir le mouvement

avec le meilleur taux de réussite du jeu. Ici pour l'élection de comment jouer on utilise un coté heuristique, car en faisant comme ça on espère être placé sur les bonnes configurations du jeu (où l'on perd peu souvent).
Le but de ce travail est de se familiariser avec la théorie des Monte Carlo tree search et d'illustrer la résolution de certains problèmes à l'aide de ceux-ci.

Références bibliographiques
[1] Browne, C. B. et al (2012). A survey of monte carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in games.
[2] Fu, M. C. (2018). Monte Carlo tree search : A tutorial. In 2018 Winter Simulation Conference (WSC) (pp. 222-236). IEEE.
[3] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning : An introduction. MIT press.

3. **Comment définir et calculer une moyenne dans un ensemble non convexe**
Contact : RICHOU Adrien
Courriel : adrien.richou@u-bordeaux.fr

Imaginons que l'on cherche à calculer la moyenne spatiale de la population croate : le caractère fortement non-convexe de la Croatie fait que cette moyenne risque de se retrouver hors du pays ce qui n'est pas un résultat très acceptable. On peut alors se demander s'il est possible de définir une nouvelle notion de moyenne qui soit raisonnable (i.e. qui corresponde à la moyenne habituelle lorsque le pays est convexe par exemple) et qui reste toujours à l'intérieur du pays. Plus généralement on peut se demander comment définir naturellement l'espérance d'une variable aléatoire contrainte à rester à l'intérieur d'un domaine non convexe. Une réponse possible est de définir, pour une variable aléatoire X à valeur dans un ensemble $\mathcal{D} \subset \mathbb{R}^d$, l'espérance de X de la façon suivante

$$\tilde{\mathbb{E}}[X] \in \text{argmax}_{y \in \mathcal{D}} \mathbb{E}[d^2(y, X)]$$

avec $\mathbb{E}$ l'espérance usuelle et $d(x, y)$ la longueur du plus court chemin entre x et y restant à l'intérieur de $\mathcal{D}$. On appelle cette nouvelle moyenne, moyenne de Fréchet. Le but de ce projet est d'étudier une méthode numérique permettant de calculer en pratique cette moyenne de Fréchet et, si le temps le permet, calculer la moyenne de Fréchet de la population croate.

4. **Comment gérer optimalement, et de manière durable, une forêt**
Contact : Adrien Richou
Courriel : adrien.richou@u-bordeaux.fr

Supposons que l'on gère l'exploitation d'une forêt de manière durable. À un instant $t > 0$, on note X_t la quantité de bois exploitable dans cette forêt. Sachant que toute opération de coupe induit des frais fixes, on ne peut pas se permettre de couper du bois en permanence et on cherche donc à optimiser les instants où l'on effectue des coupes dans cette forêt. Pour répondre à cette question, on commence par modéliser la dynamique de X_t par ce que l'on appelle une équation différentielle stochastique, à savoir une équation différentielle à laquelle on ajoute du bruit. Ainsi, on suppose que

$$X_t = \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s,$$

avec $(W_t)_{t \geqslant 0}$ un mouvement brownien. L'intégrale $\int_0^t \sigma(X_s)dW_s$, appelée Intégrale d'Itô, étant

un objet mathématique difficile à appréhender, on se contentera de considérer des approximations en temps discret de X_t :

$$\tilde{X}_{(k+1)h} = b(\tilde{X}_{kh})h + \sigma(\tilde{X}_{kh})\sqrt{h}\varepsilon_{k+1},$$

avec $(\varepsilon_k)_{k\geqslant 1}$ des variables indépendantes de loi gaussienne centrée réduite et h un pas de temps petit.

Trouver les instants optimaux de coupe consiste à trouver un seuil optimal à partir duquel on choisit d'effectuer la coupe lorsque celui-ci est atteint. Il est possible de calculer ce seuil optimal théoriquement. Néanmoins celui-ci dépend de b, or en pratique on ne connaît pas parfaitement la dynamique de X, et donc b, ce qui rend cette solution inexploitable.

Une première solution pour contourner ce problème consiste à se donner une fenêtre de temps $[0, T]$ où l'on n'agit pas sur la forêt mais où l'on utilise l'observation de $(X_t)_{t\in[0,T]}$ pour estimer b. On utilise ensuite cette estimation de b pour avoir une estimation du seuil optimal que l'on exploite alors tout le reste du temps, après T. L'inconvénient de cette approche est que l'erreur d'estimation initiale aura un impact sur le long terme. Une autre approche consiste à faire de l'apprentissage par renforcement, en alternant des phases d'apprentissage (on observe juste la dynamique de X pour améliorer l'estimation de b) et des phases d'exploitation (on utilise notre estimation de b pour avoir une estimation du seuil optimal de coupe, et on applique cette stratégie de coupe approchée). Le but est alors de déterminer le temps optimal que l'on doit passer dans chaque phase.

Toute l'étude mathématique de ce problème se trouve dans l'article [1] disponible en ligne, en accès libre. Le but de ce projet est d'étudier numériquement le modèle proposé ci-dessus et d'illustrer l'optimalité des solutions proposées dans l'article de Christensen et Strauch.

Références bibliographiques

[1] *Nonparametric learning for impulse control problems*, Sören Christensen, Claudia Strauch, arXiv :1909.09528.

5. Algorithmes pour la constitution automatique de groupes d'étudiants pour des TER

Contact : Avalos-Fernandez Marta et Russon Dylan
Mail : marta.avalos-fernandez@u-bordeaux.fr, dylan.russon@u-bordeaux.fr

Les projets de groupe au sein d'un programme de master reflètent l'importance du travail collaboratif dans le monde professionnel en encourageant les étudiants à travailler conjointement pour résoudre des problèmes complexes. Les avantages de ces projets incluent la promotion de l'apprentissage collaboratif, la diversité des compétences, la préparation à la réalité du monde professionnel et le développement de compétences transversales. Cependant, pour tirer le meilleur parti de ces avantages, une gestion efficace est essentielle pour atténuer les inconvénients potentiels, tels que les conflits interpersonnels, les inégalités de contribution et les contraintes de temps. La constitution des groupes devrait viser un équilibre entre la diversité des compétences et la compatibilité des membres. Des équipes trop hétérogènes peuvent rencontrer des difficultés de communication, tandis que des équipes trop homogènes risquent de manquer de perspectives variées et innovantes. La personnalité des membres, leurs préférences de travail, ainsi que leurs compétences interpersonnelles, jouent également un rôle important dans la réussite d'un groupe. D'autre part, le choix du sujet à aborder dans le cadre du projet (ainsi que le choix de ses partenaires lorsque ceux-ci ont déjà eu l'occasion de se connaître) constituent également des facteurs de motivation et d'engagement.

Dans le contexte de l'enseignement en à distance via internet (e-learning), les défis peuvent être exacerbés. Les étudiants sont souvent géographiquement dispersés, ce qui peut compliquer la coordination, et la diversité culturelle peut entraîner des malentendus. De plus, les étudiants s'orientant vers l'e-learning ont le plus souvent des obligations familiales et professionnelles, ce qui peut rendre difficile leur participation régulière aux projets de groupe. Des outils sur le web tels que CATME Team-Maker ont été développés afin d'attribuer des étudiants à des équipes compatibles en fonction des caractéristiques des étudiants collectées à l'aide de questionnaires (par exemple, le style de leadership, l'âge, la filière, la disponibilité de l'horaire, ou encore les résultats académiques des étudiants) autour desquelles l'encadrant souhaite organiser ses équipes.

La question peut, en effet, être formulée sous la forme d'un problème d'optimisation discrète, comme dans le problème de formation d'équipe (Team Formation Problem - TFP) où l'on cherche à effectuer l'allocation de plusieurs individus correspondant à un ensemble de compétences en tant que groupe jugées nécessaires pour un travail efficace. Ou bien, elle peut être abordée dans le cadre du modèle d'appariement stable, où les individus d'une population expriment des préférences quant à être appariés avec des individus d'une autre population, résolu à l'aide d'algorithmes tels que l'algorithme de Gale-Shapley (comme dans le cas des admissions universitaires - Parcoursup, Mon Master - et la constitution de colocataires, ou encore, les sites de rencontres). Des études proposent de combiner l'objectif d'optimiser l'efficacité et la stabilité, ou encore, en tenant compte de critères d'équité. D'autres approches reposent sur du clustering, des réseaux d'inférence (notamment dans la formation de groupes dans des réseaux sociaux), la classification supervisée ou les algorithmes génétiques.

L'objectif de ce TER est de se familiariser avec la thématique, par une revue de la littérature, et d'implémenter plusieurs algorithmes illustrant différentes approches pour résoudre le problème de constitution de groupes d'étudiants. Ces approches devront prendre en compte différentes situations :

— Modalité d'enseignement en e-learning ou en face à face.
— Critères de compatibilité et/ou de diversité, recherche d'homogénéité ou d'hétérogénéité.
— Prise en compte ou non des souhaits des étudiants (quant au sujet du projet parmi les sujets proposés, quant à la constitution du groupe).
— Prise en compte de caractéristiques des étudiants par questionnaire ou un autre moyen.

Réfrences bibliographiques

[1] Nebojsa Gavrilovic, Tatjana Sibalija, Dragan Domazet. Design and implementation of discrete Jaya and discrete PSO algorithms for automatic collaborative learning group composition in an e-learning system. Applied Soft Computing, Volume 129, 2022,

[2] Moreno J, Sánchez JD, Pineda AF. A hybrid approach for composing groups in collaborative learning contexts. Heliyon. 2021, 10 ;7(6) :e07249.

[3] Lechuga, C.G., Doroudi, S. Three Algorithms for Grouping Students : A Bridge Between Personalized Tutoring System Data and Classroom Pedagogy. Int J Artif Intell Educ 33, 843–884 (2023).

[4] Vallès-Català T, Palau R (2023) Minimum entropy collaborative groupings : A tool for an automatic heterogeneous learning group formation. PLoS ONE 18(3) : e0280604.

[5] Bousalem, Zakaria & Qazdar, Aimad & El Guabassi, Inssaf. (2023). Cooperative Learning Groups : A New Approach Based on Students' Performance Prediction. International Journal of Online and Biomedical Engineering (iJOE). 19. 34-48.

[6] Victor-ikoh, Maudlyn & Catherine, Ogunmodimu. (2021). Students' Group Formation Using K-Means Clustering in Combination with a Heterogeneous Grouping Algorithm. American Journal of Engineering Research. 10. 01-09.

[7] Li X, Ouyang F, Chen W. Examining the effect of a genetic algorithm-enabled grouping method on collaborative performances, processes, and perceptions. J Comput High Educ. 2022 ;34(3) :790-819.

[8] Li, R., & Bringardner, J. (2022), Work in Progress : Using CATME in Team Development of One-Semester-Long Open-Ended First-Year Engineering Student Design Projects Paper presented at 2022 First-Year Engineering Experience, East Lansing, Michigan.

6. **Apprentissage statistique dans un réseau de capteurs et application à la reconstruction de la dynamique temporelle de stations de vélos en libre service**

Contact : Jérémie Bigot
Mail : `jeremie.bigot@u-bordeaux.fr`

Dans ce projet, il est proposé de s'intéresser à des observations qui sont collectées sous la forme du nombre de vélos disponibles au cours du temps dans l'ensemble des stations de vélos en libre service dans une ville qui propose ce type de service. Il s'agit du cadre de l'observation de données spatio-temporelles qui se retrouve dans de nombreux domaines d'applications, en particulier sur les plateformes de mise à disposition de données libres créées récemment par de nombreuses municipalités.

Dans ce projet, on s'intéresse à la situation où certains capteurs d'enregistrement des données ne fonctionnent pas (par exemple dans un souci d'économie d'énergie), ce qui entraîne des observations manquantes dans plusieurs stations d'une ville pendant une période temps donnée. On se propose alors de développer des méthodes qui permettent d'estimer le nombre de vélos disponibles (avec construction d'intervalles de confiance) dans ces stations où les capteurs sont en défaut à partir des observations disponibles dans les autres stations.

Pour cela, on utilisera une approche basée sur de la régression linéaire pénalisée à partir de données structurée en réseau spatial (c'est à dire un graphe dont les sommets sont les positions des stations et les arêtes représentent les connexions entre les stations). Une première partie du projet sera centrée sur l'étude de la structure du graphe des réseaux de capteurs. Dans un deuxième temps, on s'intéressera à la question de l'extension du modèle linéaire usuel pour l'interpolation lisse de données sur des graphes. Ce projet s'appuiera sur le package R `sand` pour l'analyse de données sur des réseaux, ainsi que sur vos connaissances en régression linéaire et algèbre linéaire pour la diagonalisation de matrices.

Références bibliographiques

[1] Kolaczyk, E.D. (2009). Statistical Analysis of Network Data : Methods and Models. Springer, New York.

[2] Kolaczyk, E.D. and Csardi, G. (2014). Statistical Analysis of Network Data with R. Springer, New York.

[3] Package R `sand` - `https://cran.r-project.org/web/packages/sand/index.html` et `https://github.com/kolaczyk/sand`

7. **Plus grandes valeurs propres de matrices aléatoires**

Contact : Delphine Féral
Mail : `dferal@u-bordeaux.fr`

Soit $N \geqslant 1$ un entier et $\mathbb{M}_N = (M_{i,j})_{1 \leqslant i,j \leqslant N}$ une matrice symétrique de taille $N \times N$ qui est supposée aléatoire i.e. constituée de variables aléatoires réelles (v.a.r.) telles que :
les coefficients $\{M_{i,j},\, 1 \leqslant i < j \leqslant N\} \cup \Big\{\sqrt{2}M_{i,i},\, 1 \leqslant i < j \leqslant N\Big\}$ sont des v.a.r. i.i.d., de même loi symétrique, de variance σ^2 et admettent un moment d'ordre 4.
La matrice d'intérêt est la matrice $\mathbb{X}_N := \dfrac{1}{\sqrt{N}}\mathbb{M}_N$, appelée ***matrice de Wigner d'ordre*** N**.** L'objectif est d'étudier le comportement du spectre de $\mathbb{X}_N$ lorsque la taille $N \to \infty$ (Voir le cours en Français [1]). On note $\lambda_1(\mathbb{X}_N) \geqslant \ldots \geqslant \lambda_N(\mathbb{X}_N)$ les N valeurs propres ordonnées de $\mathbb{X}_N$: ce sont des v.a.r. dépendantes.
Ici, nous étudions le comportement global et local du spectre lorsque $N \to \infty$.

En fait, au départ : les coefficients de la matrice étaient supposés être normalement distribués : le modèle de Wigner Gaussien réel est appelé ***GOE (Ensemble Orthogonal Gaussien)***, noté $\mathbb{X}_N^G$, et tel que les coefficients $\Big\{M_{i,j}^G, \sqrt{2}M_{i,i}^G,\, 1 \leqslant i \leqslant j \leqslant N\Big\}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

- **Le comportement global du spectre :** est lié à ***la loi du demi-cercle*** μ_σ ***de paramètre*** σ qui est une loi de probabitité portée par le compact $[-2\sigma, 2\sigma]$ et de densité :

  $$f_\sigma(x) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}\, 1_{[-2\sigma,2\sigma]}(x).$$

  Pour tout borélien B de $\mathbb{R}$, $\mu_N(B) := \dfrac{1}{N}\mathrm{Card}\big\{1 \leqslant k \leqslant N \,:\, \lambda_k(\mathbb{X}_N^G) \in B\big\}$ la proportion de valeurs propres de $\mathbb{X}_N^G$ dans B converge presque sûrement vers $\mu_\sigma(B) = \displaystyle\int_B f_\sigma(x)\,\mathrm{d}x$.
  Noter que ceci implique que lorsque $N \to \infty$, et avec une probabilité qui tend vers 1, toutes les valeurs propres de $\mathbb{X}_N^G$ restent dans l'intervalle $[-2\sigma, 2\sigma]$.

- **Localement, le bord droit du spectre :** la plus grande valeur propre $\lambda_1(\mathbb{X}_N^G)$ converge (p.s.) vers 2σ. De plus, elle fluctue selon une nouvelle loi (penser au TCL) dite ***loi de Tracy-Widom*** de fonction de répartition notée F_1 i.e. que pour tout réel t,

  $$\lim_{N\to\infty} \mathbb{P}\Big[N^{2/3}(\lambda_1(\mathbb{X}_N^G) - 2\sigma) \leqslant t\Big] = F_1(t).$$

Il est d'un grand intérêt de comprendre si (et jusqu'où) l'hypothèse de normalitÈ peut être affaiblie i.e. comprendre la sensibilité des résultats à la loi des coefficients. En 1991, M.L. Metha a conjecturé que le comportement asymptotique des grandes matrices aléatoires est ***universel*** (comme la Loi Forte de Grands Nombres et le TCL) : i.e. qu'il ne dépend pas de la loi particulière des coefficients de la matrice $(\mathbb{X}_N^G)$ mais seulement de leurs 4 premiers moments. Ainsi, les résultats précédents sur les matrices gaussiennes $(\mathbb{X}_N^G)$ s'étendraient à toutes les matrices de Wigner $(\mathbb{X}_N)$ définies ci-dessus. Aujourd'hui, il reste à prouver l'universalité de la loi de Tracy-Widom.

Dans ce TER, nous nous intéressons à deux variantes du modèle de Wigner :

(i) D'abord nous allons considérer des ***matrices de Wigner dites déformées*** définies par :

$$\mathbb{A}_N := \mathbb{X}_N + \mathbb{D}_N$$

où X_N est une matrice de Wigner comme ci-dessus et $\mathbb{D}_N$ est une matrice (non aléatoire) symétrique de rang 1 qui admet une valeur propre $\theta > 0$ (notez que ce modèle, outre son intérêt

purement théorique, trouve des applications en télécommunications numériques). L'objectif est alors de comprendre l'influence du paramètre θ (et des vecteurs propres de $\mathbb{D}_N$) sur le comportement asymptotique du spectre $\mathbb{A}_N$ (cf. par exemple [4] pour une introduction, et [3]).
Fortement étudié depuis 2001, ce modèle est aujourd'hui bien compris : **le premier objectif du mémoire** sera notamment d'illustrer les résultats connus à l'aide de simulations et l'élaboration de tests statistiques (à imaginer).

(ii) La seconde variante à étudier est ***le modèle dit de "Wigner à queues lourdes (et*** $\alpha = 4$***)"*** et sa version déformée. Des avancées sur la compréhension de ce modèle ont été récemment obtenues par S. Diaconu dans [2] et [3]. Des questions restent ouvertes, et ces travaux laissent penser - sans le justifier totalement - qu'il existe des liens intéressants avec la variante **(i)**.
Le deuxième objectif du TER sera d'essayer de comprendre et illustrer les résultats de S. Diaconu : vous serez bien sûr accompagnés et guidés dans la compréhension de ces travaux de recherche.
Le troisième objectif du TER sera d'étudier certaines des questions ouvertes en considérant en détails le cas où les coefficients suivent une loi explicite de type Paréto (sans 4ème moment).

Références :
[1] Capitaine M., Notes de Cours "Introduction aux grandes matrices aléatoires", https://www.math.univ-toulouse.fr/~capitain/M2RpolyWEB.pdf
[2] Diaconu S., More Limiting Distributions for Eigenvalues of Wigner Matrices, 2022, https://arxiv.org/abs/2203.08712
[3] Diaconu S., Finite Rank Perturbations of Heavy-Tailed Wigner Matrices, 2022, https://arxiv.org/pdf/2208.02756.pdf
[4] Rochet J., Notes "Matrices aléatoires et perturbations de rang fini", Octobre 2012 (Introduction Exposé de Magistère).

8. **Analyse et représentation de données issues d'une enquête linguistique historique en domaine occitan**
Contact : Alexandre Genadot
Mail : alexandre.genadot@u-bordeaux.fr

Nous nous intéresserons dans ce projet aux données issues d'une grande enquête linguistique menée dans le Sud-Ouest par Edouard Bourciez, alors professeur à l'université de Bordeaux, lors de l'hiver 1894-1895, voir ici : `https://www.math.u-bordeaux.fr/~agenadot/bourciezproject.html`

Il sera fourni aux étudiants un tableau de traduction de différents mots, une centaine disons, en différent lieux, 3000 environs, en graphie dite patoisante. Les objectifs de ce TER pourront être :

- de mener des statistiques descriptives sur ces données selon au moins deux axes. En premier lieu, on pourra produire des cartes de représentations de traits linguistiques spécifiques. On s'intéressera ensuite tout particulièrement au techniques de clustering pour des données qualititatives avec composante spatiale.

- de construire un outil simple de visualisation de ces données (avec shiny ou autre...).

Ce projet s'appuie donc sur les deux UE du M1 :
- Représentation de données et statistique multidimensionnelle (semestre 1),
- Statistique computationnelle pour l'exploration de données (semestre 2).

Avoir un intérêt pour la linguistique ou l'occitan peut éventuellement être un plus.