

TP n°1 : Analyse Spectrale et Inférence

Le compte-rendu doit être rédigé en LaTeX et être le plus concis possible. Toutes les figures doivent être commentées.

On considère dans tout le TP le modèle $\frac{1}{n}X_N^*R_NX_N$ où $X_N \in \mathbb{C}^{N \times n}$ a des entrées i.i.d. centrées de variance unité. On supposera ici que $R_N = \text{diag}(\lambda_1^R I_{N_1}, \dots, \lambda_K^R I_{N_K})$ où $N_i/N \rightarrow c_i$ et $\lambda_1^R < \dots < \lambda_K^R$. On supposera aussi $N/n \rightarrow c \in (0, \infty)$.

Séparation exacte du spectre

On effectue ici des observations sur la répartition exacte des valeurs propres de $\frac{1}{n}X_N^*R_NX_N$.

1. Prenez plusieurs exemples de matrices X_N telles que $\mathbb{E}[|X_{ij}|^4] < \infty$. Vérifiez les propriétés de “no eigenvalue outside de support” et de séparation exacte du spectre pour une taille assez grande de N, n : à savoir, aucune valeur propre ne doit s’échapper asymptotiquement du spectre et le nombre de valeurs propres dans chaque composante connexe est en proportion exacte des multiplicité N_i .
2. Prenez maintenant X_{ij} tel que $\mathbb{E}[|X_{11}|^4] = \infty$ (on pourra prendre des distributions de type Student). Que peut-on dire de la séparation exacte ?

Graphes de $x(\tilde{t})$

On dénote $\tilde{\mathcal{F}}$ la loi limite des valeurs propres de $\frac{1}{n}X_N^*R_NX_N$.

1. Tracez $f(x) = \frac{1}{\pi} \lim_{y \rightarrow 0} \Im[\tilde{t}(x + iy)]$ pour tout $x > 0$. On pourra se contenter de prendre y petit et d’utiliser un algorithme du point fixe.
2. Tracez, comme vu en cours, la fonction $x(\tilde{t})$ pour $\tilde{t} \notin \{0, -1/\lambda_1^R, \dots, -1/\lambda_K^R\}$.
3. En jouant sur les paramètres c, c_i et λ_i^R , apportez des remarques sur les graphes de f et $x(\cdot)$ observés.

Estimation des λ_i^R

A l’aide des Exercices 3 et 4 de la feuille d’exercices n°5, nous allons estimer ici les valeurs des λ_i^R , pour $i = 1, \dots, K$. On supposera que les rapports c_i sont connus.

1. On se place tout d’abord dans un scénario où le support de f se décompose en K clusters exactement. Déterminez par simulations un ensemble de paramètres satisfaisant cette condition.
2. Grâce aux calculs effectués en cours (et dans les Exercices 3 et 4 mentionnés plus tôt),
 - (a) Implémentez l’algorithme d’estimation de chacun des λ_i^R . On notera $\hat{\lambda}_{i,N}^R$ l’estimateur de $\lambda_{i,N}^R$ pour une taille N du jeu de données.
 - (b) Tracez ensuite l’histogramme des $\hat{\lambda}_{i,N}^R$ pour un grand nombre de réalisations de X_N . Que remarquez-vous sur la forme des histogrammes ? Suggérez l’existence d’une loi limite pour les $\hat{\lambda}_{i,N}^R$.
 - (c) A partir de cet histogramme, évaluez l’erreur quadratique moyenne (MSE) des $\hat{\lambda}_{i,N}^R$, i.e., $\mathbb{E}[|\hat{\lambda}_{i,N}^R - \lambda_i^R|^2]$, et tracez $\mathbb{E}[|\hat{\lambda}_{i,N}^R - \lambda_i^R|^2]$ en fonction de N (en conservant c constant) pour chacun des i . Que remarquez-vous ? Affinez les propriétés de la loi limite discutée plus haut.
 - (d) Comparez les performances à une méthode “naïve” d’estimation des λ_i^R qui se baserait sur le fait que $n \gg N$. Commentez les résultats.
3. On se place désormais dans le cas $K = 2$. Déterminez par simulations une paramétrisation critique c_0 de c donnant lieu à une séparation du spectre de $\tilde{\mathcal{F}}$ en deux clusters ou en un seul.

4. Grâce aux calculs effectués en cours (et dans les Exercices 3 et 4),
- (a) En fixant N cette fois-ci, tracez la courbe de $\mathbb{E}[|\hat{\lambda}_{i,N}^R - \lambda_i^R|^2]$ pour différentes valeurs de n , de telle manière que c_0 soit contenu dans l'ensemble des N/n ainsi tracé. Que remarque-t-on ? Commentez.
 - (b) On prend ici $c > c_0$ (de sorte à avoir un seul cluster dans le spectre de $\tilde{\mathcal{F}}$). Proposez une méthode alternative basée sur l'estimation de λ_i^R et $(\lambda_i^R)^2$ permettant d'estimer λ_1^R et λ_2^R . On notera $\check{\lambda}_{i,N}^R$ les nouveaux estimés (un problème devrait apparaître que nous commenterons plus bas).
 - (c) Tracez à nouveau la courbe $\mathbb{E}[|\hat{\lambda}_{i,N}^R - t_i|^2]$ pour différentes valeurs de n et comparez à la courbe $\mathbb{E}[|\check{\lambda}_{i,N}^R - t_i|^2]$. Commentez.
 - (d) Quelles sont les limitations importantes de cette nouvelle méthode ? Peut-on passer outre ?