Cohérence de grandes matrices aléatoires. Théorèmes limites et applications.

Maxime Boucher

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE MATHÉMATIQUES, INFORMATIQUE, PHYSIQUE THÉORIQUE ET INGÉNIERIE DES SYSTÈMES

Institut Denis Poisson

Thèse présentée par :

Maxime BOUCHER

soutenue le : **18 Juin 2021**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : **Mathématiques**

Cohérence de grandes matrices aléatoires. Théorèmes limites et applications

Thèse dirigée par :

Didier CHAUVEAU	Professeur, Université d'Orléans
Marguerite ZANI	Professeure, Université d'Orléans

RAPPORTEURS :

Christophe BIERNACKI	Professeur, Université de Lille 1
Ghislaine GAYRAUD	Professeure, Université de Technologie de Compiègne

---

JURY :

Christophe BIERNACKI	Professeur, Université de Lille 1
Didier CHAUVEAU	Professeur, Université d'Orléans
Romain COUILLET	Professeur, Université de Grenoble-Alpes
Laurent DELSOL	Maître de Conférences, Université d'Orléans
Ghislaine GAYRAUD	Professeure, Université de Technologie de Compiègne Présidente de jury
Marguerite ZANI	Professeure, Université d'Orléans

Cohérence de grandes matrices aléatoires. Théorèmes limites et applications

Maxime BOUCHER

Institut Denis Poisson, FRANCE
UMR CNRS 7013

maxime.boucher@univ-orleans.fr

18 Juin 2021

Remerciements

Il est acquis que la thèse n'est pas l'affaire d'une seule personne, il est temps de rendre à César, ce qui est à César.

Avant tout, j'aimerais adresser mes plus sincères remerciements à mes directeurs de thèse : Didier Chauveau et Marguerite Zani. Je pense pouvoir dire sans aucun doute possible qu'on ne serait pas réuni aujourd'hui à la soutenance sans vous. Merci pour votre passion, votre énergie, votre rigueur et toutes les qualités qui font de chacun de vous d'excellent directeur de thèse. Didier, tu m'as permis d'aller plus loin dans la partie numérique, d'aborder des thématiques nouvelles auxquelles je n'aurais jamais osé. Ceci m'a permis d'approfondir mes connaissances et d'améliorer mes compétences ! Marguerite, tu m'as donné l'occasion d'aborder la recherche avec une autonomie et un appui sans faille. Je pense que je n'oublierai jamais les calculs de cette thèse sur lesquels nous avons passé (beaucoup !) de temps. Vous avez toujours été là pour répondre à mes questions, mes doutes, mes complications. Pour cela je vous en remercie infiniment ! Il serait compliqué de mettre à l'écart le contexte sanitaire que nous traversons. Pour cela, d'un point de vu plus personnel, je pense qu'il est important de vous adresser mes remerciements les plus chaleureux. Au-delà du professionnel, vous avez répondu présent lorsque la situation pesait sur mes motivations, mes doutes. Marguerite, faire équipe avec toi pour le TD de statistiques a été comme un phare dans cette période obscure, et c'est en grande partie grâce à toi que je peux me sentir fier du travail accompli. Merci beaucoup.

Je tiens à remercier chaleureusement Christophe Biernacki et Ghislaine Gayraud pour avoir accepté de rapporter cette thèse, d'avoir accepté en cette période plus que surchargée professionnellement. Merci pour la lecture attentive de ce mémoire, et d'y avoir consacré du temps. Vos remarques m'ont permis, et me permettront d'approfondir mes travaux dans un avenir proche.

Je suis très heureux de compter Romain Couillet parmi mon jury, ainsi que Laurent Delsol. Laurent, il serait inconvenant de ma part de ne pas rappeler ici que si j'arrive à bout de ma thèse, c'est parce qu'en fin de licence 3, tu m'as proposé un projet de Statistiques et que tu as su m'orienter au mieux par la suite. Nos discussions ont également joué un rôle important au cours de mon cursus. Je t'en remercie !

Le bon déroulement de cette thèse tient également à l'accueil reçu au laboratoire. Je remercie tous ces membres pour leurs disponibilités, leurs discussions. Merci à toi Diarra pour nos échanges. Il m'est impossible de ne pas remercier l'équipe Com' du laboratoire qui m'a donné l'occasion de participer à des événements de vulgarisations scientifiques que j'affectionne beaucoup. Merci à Philippe, Magali, Julien, Laurent, Noémie, Anne et également à Fanny et Sabrina. Grâce à vous, nous avons pu monter un Escape Game mathématique. Ce fut une très belle expérience ! Je reviendrais dans 10 ans pour la mise à jour des énigmes !

Je tiens à adresser mes remerciements à Marie-Laurence, Marie-France et Anne pour avoir été présente pour toutes les questions concernant l'administratif et de m'avoir aidé grandement dans ce cadre où je suis loin d'être performant ! Anne, merci pour ces années où tu as toujours eu une oreille attentive. C'est également grâce à toi que j'en suis ici aujourd'hui. Merci infiniment Anne !

Un petit mot pour mes camarades. Nous avons eu peu de temps pour se rencontrer, ce fut agréable de vous compter parmis nous : Tran, Grégoire, Alexis, Rita, Rana, Ouassim, Léo. Merci Hongwei, tu as soutenu avant moi, tu as donc gagné cette course ! Enfin, merci à toi Noémie ! Que ce soit pour l'équipe qu'on formait pour l'escape game, les entraînements de présentation, les discussions pour nous débloquer, ou tout simplement nos échanges, nos chants endiablés, ... merci beaucoup !

Sur un plan plus personnel, je tiens à remercier ma famille. Mes parents pour être venu aujourd'hui pour ce grand jour, et pour m'avoir permis de faire mes études (au prix de formulaire de mathématique un peu partout dans la maison ...), de m'avoir suivi jusqu'à aujourd'hui. Un grand merci à mes grands-parents pour avoir toujours été là aussi, à m'appeler régulièrement en me demandant si j'étais "au bahut", et pour avoir suivi cette thèse également en me demandant comment cela avançait, ce que je faisais.

Merci à mes amies de toujours Coralie et Émilie S pour être toujours à mes côtés quoi qu'il arrive ! Je suis tellement heureux de vous compter parmi nous en ce jour. Votre aide à réellement était vitale pour moi. Merci à Emilie L et Estelle ! Nous nous sommes rencontrés en faisant des maths pour nos études, 5 ans après nous faisons toujours des maths ensemble (et quelques jeux bien sûrs !) ! Je suis vraiment reconnaissant de vous avoir à mes côtés ! Un grand merci à Estelle qui a relu la totalité de ce manuscrit et qui a été une aide incroyable pour améliorer la qualité de rédaction de ce manuscrit. Merci beaucoup Estelle !

Pour finir, je te remercie Steven. Ces dernières lignes sont pour toi. Merci infiniment d'avoir été présent, de m'avoir soutenu, poussé dans les bonnes directions. Tu as été mon pilier. Tu m'as permis de ne pas lâcher mon cap et de m'accrocher. Merci d'avoir relu ce manuscrit et de m'avoir aider à l'améliorer, de m'écouter exposer pour éclaircir mes idées, ... Merci infiniment pour tout !

Table des matières

Table des figures

Introduction

Cette thèse est dédiée à l'étude de matrices aléatoires de grandes dimensions. Cette thématique a beaucoup évolué ces dernières années ([BS10], [AGZ10]). Que ce soit théoriques ou appliqués, les développements sont très variés : physique ([For10]), statistique ([Joh01],[Joh08]), économie ([And91]), traitement d'image et du signal ([Don06], [CT05], [CRT06b], [CRT06a]), intelligence artificielle ([BSQ18]), inférence en grandes dimensions ([BS96, CT07, BJYZ09, CWX10a, CZZ10, BRT09]), détection de signaux avec des matrices sparses de grandes dimensions ([BG16]). Les premiers travaux portaient sur l'étude spectrale de ces matrices aléatoires : [Wig58], [BS10], [Meh04], [BC12]. L'acquisition de toujours plus d'informations a conduit à une étude de matrices toujours plus grandes. On parle à présent de Big Data.

D'un point de vue statistique, le phénomène de Big Data est régulièrement décrit comme étant l'observation d'un nombre raisonnable de variables sur un nombre d'individus très important. Ce phénomène se modélise alors par des matrices d'observations possédant un nombre de lignes n (individus) beaucoup plus important que le nombre de colonnes p (caractères) : $n >> p$. Des études statistiques récentes étudient notamment ce qu'il se passe pour $n \to +\infty$ afin d'avoir des informations toujours plus précises sur les caractères.

Dans cette thèse, nous considérons un cas un peu différent. Nous nous intéressons à une matrice qui résume l'observation d'un nombre de caractères p beaucoup plus important que le nombre d'individus n sur lequel on les observe : $n << p$. On a alors une matrice d'observation rectangulaire dans le sens de la largeur (plutôt que dans le sens de la hauteur dans le format classique). De plus, nous allons considérer dans ce manuscrit que ce nombre p de variables tend vers l'infini en même temps et plus rapidement que n. Ce contexte a été étudié par exemple dans [CG15], [CZ16b]. Cette modélisation est motivée par les différentes études génétiques qui connaissent un essor ces dernières décennies. Dans ces applications, on dispose de matrices d'observations où les variables sont des caractères génétiques, le nombre d'individu d'où proviennent ces observations étant plus faible. Nous pouvons citer [PDLLR19], [ZLLT20], [BGLL15], [CCZ16] en rapport avec ces thèmes. Les matrices de grandes dimensions sont régulièrement supposées parcimonieuses

(sparses) contenant beaucoup de coefficients nuls. On peut citer les travaux de Donoho et Huo [DH01] où l'on cherche à reconstruire un signal k-sparse en grandes dimensions (avec au moins k coefficients non nuls). Cette reconstruction est possible dès lors que la matrice de reconstruction (de grande taille) vérifie le critère MIP (*Mutual Incoherence Property*) faisant intervenir les coefficients de corrélations.

Cette thèse étudie la structure de ces grandes matrices aléatoires à travers le comportement des corrélations. Plus précisément, on étudie la cohérence qui correspond au maximum, en valeur absolue, des corrélations empiriques en dehors d'une bande ou de blocs autour de la diagonale. Nous étudions des modèles où les matrices de covariances sont sparse : elles sont composées de nombreux coefficients nuls en dehors d'une bande ou en dehors de blocs diagonaux (des modèles semblables à [DH01]).

On rappelle quelques notions de base sur les coefficients de corrélations. On définit le coefficient de corrélation de Pearson entre deux variables aléatoires X et Y, quand il existe, par :

$$r_{XY} = \frac{cov(X,Y)}{\sqrt{var\,(X)}\sqrt{var\,(Y)}} \tag{1}$$

On considère à présent que l'on dispose de deux échantillons de taille n, notés $(X_1,\dots,X_n)$ et $(Y_1,\dots,Y_n)$. On définit alors le coefficient empirique de corrélation de Pearson par :

$$\rho_{XY} = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)\left(Y_i-\overline{Y_n}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X_n}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y_n}\right)^2}} \tag{2}$$

où $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est la moyenne empirique de l'échantillon $(X_1,\dots,X_n)$ et $\overline{Y_n}$ est celle du second échantillon. Si $\mathbb{E}[X]$ et $\mathbb{E}[Y]$ sont connues, alors on définit :

$$\tilde{\rho}_{XY} = \frac{\sum_{i=1}^{n}\left(X_i-\mathbb{E}[X]\right)\left(Y_i-\mathbb{E}[Y]\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\mathbb{E}[X]\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i-\mathbb{E}[Y]\right)^2}} \tag{3}$$

Quand les couples $(X_i,Y_i)_{1\leqslant i\leqslant n}$ forment un n-échantillon de la loi du couple (X,Y), alors les coefficients ρ_{XY} et $\tilde{\rho}_{XY}$ sont des estimateurs consistants de r_{XY}, c'est-à-dire :

$$\rho_{XY} \xrightarrow[n\to+\infty]{p.s.} r_{XY} \qquad \text{et} \qquad \tilde{\rho}_{XY} \xrightarrow[n\to+\infty]{p.s.} r_{XY} \tag{4}$$

Dans ce manuscrit, nous considérons un modèle statistique où les n observations indépendantes et identiquement distribuées d'un p-vecteur aléatoire sont résumées dans une matrice d'observation, de dimensions $(n \times p)$, notée $\mathbb{X}_n = \left(X_i^j\right)_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$. On note X^j la j^{ieme} colonne de $\mathbb{X}_n$.

On calcule alors ρ_{ij}, les coefficients de corrélations empiriques de Pearson comme définis dans (2) entre chaque couple de colonne (X^i, X^j) de la matrice $\mathbb{X}_n$. Ils sont rassemblés dans la matrice symétrique de corrélation empirique, notée R_n, définie par :

$$R_n = (\rho_{ij})_{1\leqslant i,j\leqslant p} = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \dots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \dots & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \dots & 1 \end{pmatrix}. \tag{5}$$

On introduit alors la cohérence :

Définition 0.0.1. *On appelle cohérence, notée L_n, le maximum en valeurs absolue des coefficients de corrélations empiriques ρ_{ij} hors des termes diagonaux :*

$$L_n = \max_{1\leqslant i<j\leqslant p} |\rho_{ij}| . \tag{6}$$

Dans le cas où les espérances sont connues, on définit de façon similaire la matrice de corrélation empirique $\tilde{R_n} = (\tilde{\rho_{ij}})_{1\leqslant i,j\leqslant p}$ et la cohérence, notée $\tilde{L_n}$, par :

$$\tilde{L_n} = \max_{1\leqslant i<j\leqslant p} |\tilde{\rho}_{ij}| . \tag{7}$$

De façon similaire, on définit la τ-cohérence :

Définition 0.0.2. *Soit $\tau \in \mathbb{N}^*$. On définit la τ-cohérence comme étant le maximum, en valeur absolue, des coefficients de corrélations empiriques à l'extérieur de la bande centrale de la matrice R_n de largeur τ. On la note $L_{n,\tau}$:*

$$L_{n,\tau} = \max_{|i-j|\geqslant\tau} |\rho_{ij}| \tag{8}$$

Cette thèse est composée de quatre chapitres. Dans le premier chapitre, nous introduisons la méthode de Chen-Stein [Che75], [Ste72], [AGG89]. Cette dernière permet de

donner une estimation de la probabilité d'un extremum de variables aléatoires faiblement dépendantes par une loi de Poisson. En particulier, pour une suite de variables aléatoires $(\rho_k)_{k\in I}$ où I est un ensemble d'indices, la méthode de Chen-Stein permet d'obtenir, pour un réel t donné :

$$\left|\mathbb{P}\left(\max_{k\in I}(\rho_k)\leqslant t\right)-\exp(-\lambda)\right|\leqslant b_1+b_2+b_3$$

où $\lambda:=\sum_{k\in I}\mathbb{P}(\rho_k>t)$ et les quantités b_1, b_2 et b_3 prennent en compte les dépendances entre les variables aléatoires $(\rho_k)_{k\in I}$. Nous exposons également les travaux de T. Cai et T. Jiang [CJ11a], [CJ12] dans lesquels ils ont appliqué cette méthode à la cohérence dans un cadre gaussien. Ils montrent notamment que la cohérence admet une distribution limite de Gumbel dans un cadre dépendant et dans un cas où la dimension p croît de manière exponentielle vers l'infini : $n=o(\log(p))$ quand $n\to+\infty$ (dans [CJ12]). Ils considèrent, dans [CJ11a], un cas gaussien dépendant où on suppose que la matrice de covariance du modèle gaussien est structurée par bande : une bande centrale contenant des coefficients non nuls et des coefficients nuls partout ailleurs. Ils supposent que la largeur de la bande centrale, notée τ, vérifie $\tau=o(p^t)$ pour tout $t>0$. Ils décrivent alors le comportement asymptotique de $L_{n,\tau}$ dans ce cas et quand la dimension p suit le régime $\log(p)=o\left(n^{1/3}\right)$.

Dans le chapitre 2, nous généralisons les travaux de Cai et al. En effet, il nous semble plus adapté, pour des applications sur des données, de considérer une transition entre la dépendance et l'indépendance. Cette transition se matérialise par une bande intermédiaire dans la matrice de covariance contenant des coefficients asymptotiquement nuls. Ces coefficients ε_n peuvent aussi être vus comme du bruit additionnel. En particulier, nous considérons un modèle gaussien de dimension p dont la matrice de covariance $\Sigma=(\sigma_{kj})_{1\leqslant k,j\leqslant p}$ est définie de la manière suivante :

$$\sigma_{k,j}=\begin{cases} r_{kj}\sigma_k\sigma_j & \text{si} \quad |k-j|<\tau \\ \epsilon_n\sigma_k\sigma_j & \text{si} \quad \tau\leq|k-j|\leq\tau+K \\ 0 & \text{si} \quad \tau+K<|k-j| \end{cases}. \tag{9}$$

où les coefficients σ_i^2 sont les variances des composantes du p-vecteur gaussien, r_{kj} est la corrélation entre la k^{ieme} et la j^{ieme} composante et la suite $(\varepsilon_n)_{n\geqslant 0}$ est une suite de nombres réels dans $[-1,1]$ telle que $\lim_{n\to+\infty}(\varepsilon_n)=0$. On suppose que la bande intermédiaire de largeur K vérifie $K>>\tau$ mais la matrice reste sparse. Nous montrons dans ce chapitre, en appliquant la méthode de Chen-Stein, que pour $\log(p)=o\left(n^{1/3}\right)$, et sous réserve des bonnes conditions sur les coefficients (ε_n) et (r_{ij}) la τ-cohérence convenablement normalisée vérifie la convergence en loi :

$$nL_{n,\tau}^2 - 4\log(p) + \log\log(p) \xrightarrow[n\to+\infty]{\mathcal{L}} Z \tag{10}$$

où Z est une variable aléatoire réelle dont la fonction de distribution est définie sur $\mathbb{R}$ par :

$$F(y) = \exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{1}{2}y\right)\right).$$

Dans le chapitre 3, nous présentons un modèle de covariance par bloc. La matrice de covariance est donc définie de manière diagonale par bloc. Cette étude est motivée par de possibles applications en génétiques (voir [BGLL15]). Des données possédant cette structure apparaissent également dans des recherches impliquant l'étude de voiture autonome (voir [BSQ18]). En notant $\Sigma = (\sigma_{ij})_{1\leqslant i,j\leqslant p}$ la matrice de covariance, et I_r l'ensemble des indices (i,j) appartenant aux blocs diagonaux, la structure de Σ s'écrit :

$$\sigma_{ij} = \begin{cases} r_{ij}\sigma_i\sigma_j & \text{si} \quad (i,j) \in I_r \\ 0 & \text{si} \quad (i,j) \notin I_r \end{cases}. \tag{11}$$

La matrice Σ reste essentiellement composée de 0. Nous montrons, en utilisant à nouveau la méthode de Chen-Stein, et en conservant l'hypothèse $\log(p) = o\left(n^{1/3}\right)$, que sous réserve de conditions sur les coefficients $(r_{ij})_{1\leqslant i,j\leqslant p}$, on peut décrire le comportement asymptotique de la cohérence adaptée à ce modèle et montrer que la convergence en loi (10) reste valable.

Enfin, dans le chapitre 4, on s'intéresse à une méthode de simulation d'échantillons de τ-cohérence dans le cadre du chapitre 2 (modèle de covariance par bandes). Dans ce chapitre, nous mettons en exergue le problème que pose les données Big Data : l'impossibilité de stocker entièrement une matrice de corrélation empirique de dimension $p \times p$ dès lors que p est grand. Nous montrons également que les temps de calculs de matrices de corrélations qui deviennent rapidement trop grand pour pouvoir étudier le comportement asymptotique de la τ-cohérence pour des n suffisamment grands. Cette problématique est également présentée dans [KUrNT11], [CDOR09]. Pour palier cet état de fait, nous proposons une technique de simulation de Monte-Carlo reposant sur deux principes. Le premier est un découpage de la matrice de corrélation empirique par blocs qui eux peuvent tenir dans la mémoire vive d'un ordinateur classique. De plus, afin de pouvoir obtenir un nombre d'échantillons Monte-Carlo suffisamment important (pour pouvoir étudier la convergence en loi), nous effectuons les calculs de corrélations en GPGPU (*General-Purpose computing on Graphics Processing Units*). Cette manière de procéder nous permet d'économiser un temps de simulation important (de l'ordre d'un facteur 20) et d'obtenir des échantillons

de τ-cohérence pour des n (et donc des p) suffisamment grands. Les techniques de programmation en GPU se développent depuis quelques années, nous pouvons citer dans ce sens [LDFF15], [ES18].

Les axes de recherches ouverts suite à nos travaux sont nombreux.

D'un point de vue computationnel, dans le chapitre 4, nous présentons une méthode de simulation basée sur l'utilisation d'une GPU. On peut se demander s'il y a une possibilité de combiner du calcul GPU avec du calcul parallèle afin de pouvoir gagner toujours plus de temps de calcul.

D'un point de vue modélisation, le modèle présenté dans le chapitre 2 peut être généralisé à nouveau en considérant que les coefficients (ε_n) peuvent se situer ailleurs dans la matrice, en dehors de la bande centrale. De la même manière, pour le modèle de covariance par blocs présenté dans le chapitre 3, l'étape suivante serait de considérer des coefficients non nuls (ε_n) situés en dehors des blocs. Ces deux généralisations supplémentaires sont un moyen d'avoir des modèles toujours plus adaptés aux données.

Une suite directe à nos travaux serait également l'application à des données réelles. On peut envisager la création d'un test d'hypothèse concernant la structure de covariance : a-t-elle une structure par bandes ou par blocs ? Dans la même optique, on pourrait par la suite construire un test d'hypothèse spécifiquement sur le paramètre τ (pour le modèle par bandes), ou sur la largeur du plus grand bloc (pour le modèle du chapitre 3), dans le but d'obtenir plus d'informations sur ces paramètres au vu des données. De façon plus générale, la cohérence peut-être un outil utile pour des études portant sur des données génétiques (de la classification par exemple en testant sur des données si le modèle a bien une structure de covariance par blocs).

Parmi les différentes perspectives, on peut également envisager une approche bayésienne en supposant que l'on dispose d'une loi a priori sur les coefficients de corrélations non nuls (r_{kj}) dans la bande centrale ou dans les blocs diagonaux. On pourrait alors déterminer la loi a posteriori de $L_{n,\tau}$ à partir de la loi a priori supposée pour les coefficients (r_{kj}). Ceci permettrait d'avoir un modèle plus général encore, et plus facilement applicable à des données.

Enfin, nous ne pouvons pas évoquer les perspectives sans considérer la dimension de nos matrices. En effet, dans ce manuscrit, l'hypothèse principale est que la dimension p vérifie $\log(p) = o(n^{1/3})$. Nous souhaitons pouvoir aller plus loin et considérer par exemple $\log(p) = o(n)$. Dans le chapitre 1, nous évoquons les travaux [CJ12] concernant la cohé-

rence pour des dimensions de p beaucoup plus grandes où l'hypothèse d'indépendance est la clé de voûte. Nous souhaitons pouvoir inclure de la dépendance dans un cas de très grandes dimensions.

Chapitre 1

Méthode de Chen-Stein

Sommaire

Dans ce chapitre, on s'intéresse à l'approximation d'événements faiblement dépendants par une loi de Poisson. Cette méthode a été initialement présentée par Chen en 1975 [Che75]. Elle permet de majorer la distance en variation totale entre une loi donnée et la loi de Poisson.

Les techniques présentées dans Chen s'inspirent des travaux de Stein, ([Ste72]), qui propose une majoration de l'erreur pour l'approximation normale d'une somme de variables aléatoires dépendantes. Chen a adapté cette majoration dans le cas d'une approximation par une loi de Poisson. De plus, il a montré que l'approximation par la loi de Poisson était possible en contrôlant uniquement les moments d'ordres 1 et 2. Plus tard, dans [AGG89], une réécriture de la méthode de Chen-Stein a été proposée dans l'objectif de clarifier et aussi de généraliser aux dimensions supérieures $d \geqslant 1$.

Dans le chapitre suivant, nous verrons qu'à l'aide de cette méthode, nous pouvons décrire le comportement de la τ-cohérence convenablement normalisée. A noter qu'elle est également utilisée par Cai et Jiang dans [CJ11a] et [CJ12] mais aussi par Shao dans [SZ14]. Aussi nous pouvons citer de nombreux travaux où l'utilisation de l'approximation

par une loi de Poisson a été utilisée et adaptée, par exemple dans [Cou07], [VA08],[Pec12], [TSCD08].

Pour présenter la méthode de Chen-Stein, nous allons découper ce chapitre en deux parties : la première où nous présenterons l'opérateur de Chen-Stein qui est à la base de la méthode et qui propose une caractérisation de la loi de Poisson de paramètre $\lambda > 0$; la seconde partie, où nous présentons le théorème de Chen-Stein qui fournit l'approximation.

1.1 Opérateur de Chen-Stein

Dans toute cette partie, on note Z une variable aléatoire de loi de Poisson de paramètre $\lambda > 0$.

Définition 1.1.1. *Soit $\mathcal{F} := \{f : \mathbb{N} \to \mathbb{R}\}$ l'ensemble des applications de $\mathbb{N}$ vers $\mathbb{R}$. Soit $\lambda > 0$. Soit T et S deux opérateurs définis par :*

$$T : \begin{pmatrix} \mathcal{F} & \to & \mathcal{F} \\ f & \mapsto & \forall \omega \in \mathbb{N},\ (Tf)(\omega) = \omega f(\omega) - \lambda f(\omega+1) \end{pmatrix}, \tag{1.1}$$

et

$$S : \begin{pmatrix} \mathcal{F} & \to & \mathcal{F} \\ f & \mapsto & \forall \omega \in \mathbb{N},\ (Sf)(\omega+1) = \dfrac{-1}{\lambda \mathbb{P}(Z=\omega)} \mathbb{E}\left[f(Z)\mathbf{1}_{Z \leqslant \omega}\right] \end{pmatrix}. \tag{1.2}$$

Nous posons également $Sf(0) = 0$.

Lemme 1.1.1. *Les deux opérateurs S et T sont réciproques, c'est-à-dire que pour toute application $h \in \mathcal{F}$, $T(Sh) = S(Th) = h$*

Démonstration. Commençons par $T(Sh) = h$. Soit $\omega \geq 0, \omega \in \mathbb{N}$,

$$\begin{aligned}
T(Sh)(\omega) &= \omega(Sh)(\omega) - \lambda(Sh)(\omega+1) \\
&= \omega\left(\frac{-1}{\lambda\mathbb{P}(Z=\omega-1)}\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega-1}]\right) - \lambda\left(\frac{-1}{\lambda\mathbb{P}(Z=\omega)}\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega}]\right) \\
&= \frac{-\omega}{\lambda}\frac{1}{\mathbb{P}(Z=\omega-1)}\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega-1}] + \frac{1}{\mathbb{P}(Z=\omega)}\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega}] \\
&= \frac{-\omega!}{\lambda^\omega}e^\lambda\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega-1}] + \frac{\omega!}{\lambda^\omega}e^\lambda\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega}] \\
&= \frac{\omega!}{\lambda^\omega}e^\lambda\left(\mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega}] - \mathbb{E}[h(Z)\mathbf{1}_{Z\leqslant\omega-1}]\right) \\
&= \frac{\omega!}{\lambda^\omega}e^\lambda h(\omega)\mathbb{E}[\mathbf{1}_{Z=\omega}] = \frac{\omega!}{\lambda^\omega}e^\lambda h(\omega)\mathbb{P}(Z=\omega) = h(\omega)
\end{aligned}$$

A présent, pour $S(Tf)(\omega+1)$. Soit $\omega \geq 0$,

$$\begin{aligned}
S(Tf)(\omega+1) &= \frac{-1}{\lambda \mathbb{P}(Z=\omega)} \mathbb{E}[(Tf)(Z)\mathbf{1}_{Z\leqslant\omega}] \\
&= \frac{-1}{\lambda \mathbb{P}(Z=\omega)} \mathbb{E}[(Zf(Z)-\lambda f(Z+1))\mathbf{1}_{Z\leqslant\omega}] \\
&= \frac{-1}{\lambda \mathbb{P}(Z=\omega)} \sum_{k=0}^{\omega} \mathbb{E}[(kf(k)-\lambda f(k+1))\mathbf{1}_{Z=k}] \\
&= \frac{-\omega!}{\lambda^{\omega+1}} \sum_{k=0}^{\omega} (kf(k)-\lambda f(k+1)) \frac{\lambda^k}{k!} \\
&= \frac{-\omega!}{\lambda^{\omega+1}} \sum_{k=0}^{\omega} kf(k)\frac{\lambda^k}{k!} + \frac{\omega!}{\lambda^{\omega+1}} \sum_{k=0}^{\omega} \lambda f(k+1)\frac{\lambda^k}{k!} \\
&= \frac{-\omega!}{\lambda^{\omega+1}} \sum_{k=1}^{\omega} kf(k)\frac{\lambda^k}{k!} + \frac{\omega!}{\lambda^{\omega}} \sum_{k=0}^{\omega} f(k+1)\frac{\lambda^k}{k!} \\
&= \frac{-\omega!}{\lambda^{\omega+1}} \sum_{j=0}^{\omega-1} f(j+1)\frac{\lambda^{j+1}}{j!} + \frac{\omega!}{\lambda^{\omega}} \sum_{k=0}^{\omega-1} f(k+1)\frac{\lambda^k}{k!} + \frac{\omega!}{\lambda^{\omega}} f(\omega+1)\frac{\lambda^{\omega}}{\omega!} \\
&= f(\omega+1).
\end{aligned}$$

□

Sur ces deux opérateurs, on s'intéresse principalement à T. En effet, grâce à lui, on peut donner une caractérisation de la loi de Poisson de paramètre $\lambda > 0$. Ce résultat est donné dans le lemme qui suit :

Lemme 1.1.2. *Soit Q une variable aléatoire à valeurs dans $\mathbb{N}$. On a :*

$$\text{"Pour toute fonction bornée } f,\ \mathbb{E}[(Tf)(Q)] = 0\text{"} \Leftrightarrow Q \sim \mathcal{P}(\lambda)$$

Démonstration. On peut montrer ce résultat en démontrant les deux sens de l'équivalence. Supposons que l'on dispose de $Q \sim \mathcal{P}(\lambda)$. Soit f une fonction bornée :

$$\begin{aligned}
\mathbb{E}[(Tf)(Q)] &= \sum_{k=0}^{+\infty}(Tf)(k)\mathbb{P}(Q=k) \\
&= \sum_{k=0}^{+\infty} kf(k)\mathbb{P}(Q=k) - \sum_{k=0}^{+\infty}\lambda f(k+1)\mathbb{P}(Q=k) \\
&= \sum_{k=1}^{+\infty}\frac{f(k)\lambda^k}{(k-1)!}e^{-\lambda} - \sum_{k=0}^{+\infty}\frac{f(k+1)\lambda^{k+1}}{k!}e^{-\lambda} \\
&= \sum_{j=0}^{+\infty}\frac{f(j+1)\lambda^{j+1}}{j!}e^{-\lambda} - \sum_{k=0}^{+\infty}\frac{f(k+1)\lambda^{k+1}}{k!}e^{-\lambda} \\
&= 0
\end{aligned}$$

Réciproquement, supposons que nous avons, pour toute fonction bornée f,

$$\mathbb{E}\left[(Tf)(Q)\right] = 0.$$

L'hypothèse est vraie pour n'importe quelle fonction bornée. En particulier pour la fonction suivante :

$$\text{Soit } k \in \mathbb{N}, k \geqslant 1, \text{ et pour } n \in \mathbb{N},\ \tilde{f}(n) = \mathbf{1}_{n=k}$$

On a alors :

$$\begin{aligned}
\mathbb{E}[(T\tilde{f})(Q)] = 0 &\Leftrightarrow \mathbb{E}[Z\tilde{f}(Q) - \lambda\tilde{f}(Q+1)] = 0 \\
&\Leftrightarrow \mathbb{E}[Z\mathbf{1}_{Q=k} - \lambda\mathbf{1}_{Q+1=k}] = 0 \\
&\Leftrightarrow k\mathbb{P}(Q=k) = \lambda\mathbb{P}(Q=k-1) \\
&\Leftrightarrow \mathbb{P}(Q=k) = \frac{\lambda}{k}\mathbb{P}(Q=k-1).
\end{aligned}$$

En utilisant une récurrence, on montre rapidement que l'on a, pour tout $k \in \mathbb{N}$:

$$\mathbb{P}(Q=k) = \frac{\lambda^k}{k!}\mathbb{P}(Q=0).$$

Puis, la formule des probabilités totales nous impose :

$$\sum_{k=0}^{+\infty}\mathbb{P}(Q=k) = 1 \Rightarrow \mathbb{P}(Q=0) = e^{-\lambda}.$$

Finalement, on a bien $Q \sim \mathcal{P}(\lambda)$. □

L'opérateur T est donc un outil supplémentaire pour caractériser la loi de Poisson. Il est utilisé dans la méthode de Chen-Stein.

1.2 Méthode de Chen-Stein

Nous commençons par introduire plusieurs notations, et rappeler la définition de la variation totale. Soit :

- I un ensemble d'indices arbitraires.
- α un élément de I.
- X_α une variable aléatoire suivant une loi de Bernoulli telle que :

$$p_\alpha := \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

- W la variable aléatoire définie par :

$$W = \sum_{\alpha \in I} X_\alpha. \tag{1.3}$$

- $\lambda = \mathbb{E}[W] = \sum_{\alpha \in I} p_\alpha$.

Nous faisons l'hypothèse : $0 < \lambda < +\infty$.

Afin de parler de convergence d'une distribution de probabilité vers celle de la loi de Poisson, nous devons introduire la métrique pour laquelle on obtient le résultat de convergence.

Definition 1.2.1. *Distance en variation totale.*
On appelle variation totale entre les distributions de variables aléatoires W et Z à valeurs entières la quantité :

$$\|\mathcal{L}(W) - \mathcal{L}(Z)\| = \sup_{\|h\|=1} |\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]|,$$

où $h : \mathbb{N} \to \mathbb{R}$ et $\|h\| = \sup_{k \in \mathbb{N}} |h(k)|$ est la norme de h et où $\mathcal{L}(W)$ désigne la loi de la variable aléatoire W.

On aura convergence de la distribution $\mathcal{L}(W)$ vers $\mathcal{L}(Z)$ si la quantité $\|\mathcal{L}(W) - \mathcal{L}(Z)\|$ tend vers 0. La méthode de Chen-Stein fournit une majoration de la variation totale et

ainsi, il suffit de montrer que les quantités impliquées dans la majoration sont asymptotiquement nulles pour avoir convergence vers la loi de Poisson.

Pour chaque $\alpha \in I$, on suppose que l'on a choisi un ensemble $B_\alpha \subset I$ tel que $\alpha \in B_\alpha$. En fait, il faut voir l'ensemble B_α comme un voisinage de dépendance pour α. C'est-à-dire que les variables X_i sont dépendantes de X_α dès lors que $i \in B_\alpha$ et les variables seront indépendantes sinon.

Nous introduisons de nouvelles notations :

- $b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta$.
- $b_2 = \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} p_{\alpha\beta}$ où $p_{\alpha\beta} = \mathbb{E}[X_\alpha X_\beta]$.
- $s_\alpha = \mathbb{E}[|\mathbb{E}[X_\alpha - p_\alpha | \sigma(X_\beta : \beta \in I \backslash B_\alpha)]|]$.
- $s'_\alpha = \mathbb{E}[|\mathbb{E}[X_\alpha - p_\alpha | \sum_{\beta \in I \backslash B_\alpha} X_\beta]|]$.
- $b_3 = \sum_{\alpha \in I} s_\alpha$ et $b'_3 = \sum_{\alpha \in I} s'_\alpha$.

De façon heuristique, la méthode de Chen-Stein nous dit que si les quantités b_1, b_2 et b_3 sont petites, alors le nombre total d'événements W suit approximativement une loi de Poisson de paramètre λ. De façon plus rigoureuse, nous avons le théorème suivant :

Théorème 1.2.1. *Théorème de Chen-Stein.*
Soit W une variable aléatoire définie dans 1.3. Soit $Z \sim \mathcal{P}(\lambda)$, en particulier, on a $\mathbb{E}[Z] = \mathbb{E}[W] = \lambda$. Alors :

1. $\|\mathcal{L}(W) - \mathcal{L}(Z)\| \leqslant 2[(b_1 + b_2)\frac{1-e^{-\lambda}}{\lambda} + b'_3 \ \min(1, \frac{1.4}{\sqrt{\lambda}})] \leqslant 2(b_1 + b_2 + b_3)$.
2. $\left|\mathbb{P}(W = 0) - e^{-\lambda}\right| \leqslant \dfrac{1 - e^{-\lambda}}{\lambda}(b_1 + b_2 + b_3) < (b_1 + b_2 + b_3)\min(1, \dfrac{1}{\lambda})$.

Remarques 1. *La méthode de Chen-Stein est une approximation par une loi de Poisson d'événements faiblement dépendant. Nous voyons ici que le terme "faiblement dépendant" se retrouve dans les quantités b_1, b_2 et b_3. En effet, si les variables aléatoires sont toutes indépendantes, on constate par exemple que $b_3 = 0$. De plus, dans un cadre indépendant, les quantités b_1 et b_2 sont égales. Alors que sous l'hypothèse de dépendance, on constate que la quantité b_2, qui traduit la corrélation en étant une probabilité de couple, est d'autant plus élevée que les corrélations sont grandes.*

Pour la démonstration du théorème 1.2.1, nous avons besoin du lemme suivant :

Lemme 1.2.1. *On fixe $h \in \mathcal{F}$ tel que pour tout $w \in \mathbb{N}$, $w \geqslant 0$, $h(w) \in [0,1]$. Soit $f := S\left(h(.) - \mathbb{E}\left[h(Z)\right]\right)$ et $\Delta(f)(w) = f(w+1) - f(w)$. On a alors,*

$$\|\Delta f\| \leqslant \frac{1}{\lambda}\left(1 - e^{-\lambda}\right) \text{ et } \|f\| \leqslant \min\left(1, 1.4\lambda^{-1/2}\right) \tag{1.4}$$

Nous renvoyons à [AGG89] pour la démonstration de ce lemme.

Démonstration. Nous allons d'abord montrer le premier point et montrer que le second découle du premier pour le choix d'une fonction h particulière.
Montrons que :

$$\|\mathcal{L}(W) - \mathcal{L}(Z)\| \leqslant 2[(b_1+b_2)\tfrac{1-e^{-\lambda}}{\lambda} + b_3' min(1, \tfrac{1.4}{\sqrt{\lambda}})] \leqslant 2(b_1+b_2+b_3).$$

Soit $h : \mathbb{N} \to \mathbb{R}$ telle que $\|h\| = 1$. Soit $\overline{h}$ définie par $\overline{h}(.) = h(.) - \mathbb{E}[h(Z)]$. On pose $f = S(\overline{h})$. En l'occurrence, on a $Tf = T\left(S\overline{h}\right) = \overline{h}$.
Posons les variables suivantes :

- $V_\alpha = \sum_{\beta \in I \backslash B_\alpha} X_\beta$.
- $W_\alpha = W - X_\alpha$.

On peut d'ores et déjà remarquer que $V_\alpha \leqslant W_\alpha \leqslant W$ car on somme respectivement des indicatrices sur des ensembles emboités : $I \backslash B_\alpha \subset I \backslash \{\alpha\} \subset I$.
On a également deux égalités :

1. $X_\alpha f(W) = X_\alpha f(W_\alpha + 1)$. En effet, on a :

   $$X_\alpha f(W) = X_\alpha f(W_\alpha + X_\alpha).$$

   Si $X_\alpha = 0$, on a : $0f(W) = 0f(W_\alpha + 0) \Leftrightarrow 0 = 0$.
   Et si $X_\alpha = 1$, on a : $1f(W) = f(W_\alpha + 1) \Leftrightarrow f(W_\alpha + 1) = f(W_\alpha + 1)$.
   Ainsi pour toutes les valeurs de X_α, on retrouve bien l'égalité.

2. $f(W_\alpha + 1) - f(W + 1) = X_\alpha[f(W_\alpha + 1) - f(W_\alpha + 2)]$. La démonstration de cette égalité se fait également de la même façon que la précédente, il suffit de regarder les cas $X_\alpha = 0$ ou 1.

Intéressons nous à présent à la quantité : $\mathbb{E}[h(W) - h(Z)]$.

$$\begin{aligned}
\mathbb{E}[Tf(W)] &= \mathbb{E}[h(W) - h(Z)] = \mathbb{E}[Wf(W) - \lambda f(W+1)] = \mathbb{E}[\sum_{\alpha \in I} X_\alpha f(W) - \sum_{\alpha \in I} p_\alpha f(W+1)] \\
&= \sum_{\alpha \in I} \mathbb{E}[X_\alpha f(W) - p_\alpha f(W+1)] = \sum_{\alpha \in I} \mathbb{E}[X_\alpha f(W_\alpha + 1) - p_\alpha f(W+1)] \\
&= \sum_{\alpha \in I} \mathbb{E}[p_\alpha f(W_\alpha + 1) - p_\alpha f(W+1)] + \sum_{\alpha \in I} \mathbb{E}[X_\alpha f(W_\alpha + 1) - p_\alpha f(W_\alpha + 1)] \\
&= \sum_{\alpha \in I} \mathbb{E}[p_\alpha (f(W_\alpha + 1) - f(W+1))] + \sum_{\alpha \in I} \mathbb{E}[(X_\alpha - p_\alpha) f(W_\alpha + 1)].
\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Tf(W)] &= \sum_{\alpha\in I}\mathbb{E}[p_\alpha X_\alpha[f(W_\alpha+1)-f(W_\alpha+2)]]+\sum_{\alpha\in I}\mathbb{E}[(X_\alpha-p_\alpha)f(W_\alpha+1)]\\ &= \sum_{\alpha\in I}\mathbb{E}[p_\alpha X_\alpha\{f(W_\alpha+1)-f(W_\alpha+2)\}]+\sum_{\alpha\in I}\mathbb{E}[(X_\alpha-p_\alpha)\{f(W_\alpha+1)-f(V_\alpha+1)\}]\\ &\quad+\sum_{\alpha\in I}\mathbb{E}[(X_\alpha-p_\alpha)f(V_\alpha+1)].\end{aligned}$$

D'autre part, on a :

$$\begin{aligned}\sum_{\alpha\in I}\mathbb{E}[p_\alpha X_\alpha\{f(W_\alpha+1)-f(W_\alpha+2)\}] &= \sum_{\alpha\in I}\mathbb{E}[p_\alpha X_\alpha\Delta f(W_\alpha+1)]\\ &\leqslant \sum_{\alpha\in I}p_\alpha\mathbb{E}[X_\alpha\|\Delta f\|]\\ &\leqslant \|\Delta f\|\sum_{\alpha\in I}p_\alpha^2.\end{aligned}$$

Et également :

$$\begin{aligned}\sum_{\alpha\in I}\mathbb{E}[(X_\alpha-p_\alpha)f(V_\alpha+1)] &= \sum_{\alpha\in I}\mathbb{E}[\mathbb{E}[(X_\alpha-p_\alpha)f(V_\alpha+1)|\sum_{\beta\in I\setminus B_\alpha}X_\beta]]\\ &\leqslant \sum_{\alpha\in I}\mathbb{E}[\mathbb{E}[(X_\alpha-p_\alpha)\|f\||\sum_{\beta\in I\setminus B_\alpha}X_\beta]]\\ &\leqslant \|f\|\sum_{\alpha\in I}\mathbb{E}[|\mathbb{E}[(X_\alpha-p_\alpha)|\sum_{\beta\in I\setminus B_\alpha}X_\beta]|]\\ &\leqslant \|f\|b_3'.\end{aligned}$$

Il nous reste la deuxième espérance $\sum_{\alpha\in I}\mathbb{E}[(X_\alpha-p_\alpha)\{f(W_\alpha+1)-f(V_\alpha+1)\}]$ à majorer. Commençons par remarquer que la variable W_α peut s'écrire :

$$W_\alpha=\sum_{\beta\in I,\beta\neq\alpha}X_\beta=\sum_{\beta\in B_\alpha,\beta\neq\alpha}X_\beta+\sum_{\beta\in I\setminus B_\alpha,\beta\neq\alpha}X_\beta=\sum_{\beta\in B_\alpha,\beta\neq\alpha}X_\beta+V_\alpha$$

Pour effectuer la majoration, on écrit la quantité $f(W_\alpha+1)-f(V_\alpha+1)$ comme une somme télescopique de $card(B_\alpha)-1$ termes où chacun de ces termes est de la forme :

$$f(U_\beta+X_\beta)-f(U_\beta),$$

où U_β est une variable qui change à chaque étape. De plus, X_β est une variable de Bernoulli, donc $X_\beta\in\{0,1\}$ et ainsi,

$$f\left(U_\beta+X_\beta\right)-f\left(U_\beta\right)=X_\beta\left[f\left(U_\beta+1\right)-f\left(U_\beta\right)\right].$$

D'après l'égalité (2), on a :

$$f(U_\beta + X_\beta) - f(U_\beta) = X_\beta[f(U_\beta + 1) - f(U_\beta)] \leqslant X_\beta \|\Delta f\|.$$

Finalement, en réintroduisant ces résultats, on obtient :

$$\begin{aligned}
\sum_{\alpha\in I} \mathbb{E}\left[(X_\alpha - p_\alpha)\{f(W_\alpha + 1) - f(V_\alpha + 1)\}\right] &= \sum_{\alpha\in I} \mathbb{E}\left[(X_\alpha - p_\alpha) \sum_{\beta\in B_\alpha, \beta\neq\alpha} [f(U_\beta + X_\beta) - f(U_\beta)]\right] \\
&\leqslant \sum_{\alpha\in I} \sum_{\beta\in B_\alpha,\beta\neq\alpha} \mathbb{E}\left[(X_\alpha + p_\alpha)X_\beta \|\Delta f\|\right] \\
&\leqslant \|\Delta f\| \sum_{\alpha\in I} \sum_{\beta\in B_\alpha,\beta\neq\alpha} \mathbb{E}\left[X_\alpha X_\beta + p_\alpha X_\beta\right] \\
&\leqslant \|\Delta f\| \sum_{\alpha\in I} \sum_{\beta\in B_\alpha,\beta\neq\alpha} (p_{\alpha\beta} + p_\alpha p_\beta)
\end{aligned}$$

On a donc montré que :

$$\begin{aligned}
\mathbb{E}[Tf(\overline{h})(W)] &= \mathbb{E}[h(W) - h(Z)] \\
&\leqslant \|\Delta f\| \sum_{\alpha\in I} p_\alpha^2 + \|\Delta f\| \sum_{\alpha\in I} \sum_{\beta\in B_\alpha,\beta\neq\alpha} (p_{\alpha\beta} + p_\alpha p_\beta) + \|f\| b_3' \\
&\leqslant \|\Delta f\| b_1 + \|\Delta f\| b_2 + \|f\| b_3.
\end{aligned}$$

Enfin, grâce au lemme 1.2.1, nous savons que :

$$\|\Delta f\| \leqslant \tfrac{1-e^{-\lambda}}{\lambda} \text{ et } \|f\| \leqslant \min\left(1, \tfrac{1.4}{\sqrt{\lambda}}\right).$$

On a donc :

$$\mathbb{E}\left[h(W) - h(Z)\right] \leqslant \frac{1 - e^{-\lambda}}{\lambda}(b_1 + b_2) + \min\left(1, \frac{1.4}{\sqrt{\lambda}}\right) b_3$$

$$\mathbb{E}[h(W) - h(Z)] \leqslant 2(b_1 + b_2 + b_3)$$

$$\|\mathcal{L}(W) - \mathcal{L}(Z)\| \leqslant 2(b_1 + b_2 + b_3).$$

Montrons que :

$$|\mathbb{P}(W = 0) - e^{-\lambda}| \leqslant \frac{1 - e^{-\lambda}}{\lambda}(b_1 + b_2 + b_3) < (b_1 + b_2 + b_3) \min\left(1, \frac{1}{\lambda}\right).$$

On obtient facilement ce résultat en considérant la fonction $h : \omega \mapsto \mathbf{1}_{\omega=0}$. En effet :

$$\mathbb{E}[h(W) - h(Z)] = \mathbb{P}(W = 0) - \mathbb{P}(Z = 0) = \mathbb{P}(W = 0) - e^{-\lambda}$$

avec $Z \sim \mathcal{P}(\lambda)$ □

Dans la section suivante, ainsi que dans le chapitre suivant, nous allons utiliser cette méthode pour décrire le comportement asymptotique de la τ-cohérence. Montrons comment T. Cai et al. l'utilisent.

1.3 Application de la méthode de Chen-Stein pour la cohérence

Nous avons évoqué dans ce qui précède que la méthode de Chen-Stein est utilisée pour décrire le comportement de la τ-cohérence et de la cohérence. Nous allons présenter les travaux exposés dans [CJ12] et [CJ11a] dans cette section en considérant deux cas : le cas sphérique indépendant et le cas gaussien dépendant. Nous généraliserons ce dernier dans le chapitre suivant.

1.3.1 Loi sphérique

Définition 1.3.1. *Soit X un vecteur aléatoire de $\mathbb{R}^n$. On dit que X suit une loi sphérique si pour toute matrice orthogonale H :*

$$X \overset{\mathcal{L}}{=} XH$$

On notera $X \sim SP(n)$ pour signifier que X est de distribution sphérique dans $\mathbb{R}^n$.

Remarques 2. *On constate que si $X \sim SP(n)$, alors $\mathbb{E}[X] = 0_{\mathbb{R}^n}$. Il suffit pour cela de choisir la matrice $H = -I_n$ qui est bien une matrice orthogonale.*

Exemple 1.3.1. *Nous pouvons citer parmi les lois sphériques :*
- *la loi normale multidimensionnelle centrée de $\mathbb{R}^n$ avec pour matrice de corrélation la matrice unitaire I_n.*
- *le mélange fini de lois gaussiennes n-dimensionnelles centrées de covariances $\sigma_k^2 I_n$: soit Y un vecteur aléatoire de $\mathbb{R}^n$ suivant cette loi, la densité de Y peut s'écrire :*

$$f_Y(x) = \sum_{k=1}^{K} \omega_k f_k(x)$$

où $\sum_{k=1}^{K} \omega_k = 1$ *et* f_k *est la densité de la loi* $\mathcal{N}_n(0, \sigma_k^2 I_n)$ *(voir [MP00] pour plus de détails sur les modèles de mélange).*

— *la loi de Student de* $\mathbb{R}^n$ *à* m *degrés de libertés.*

1.3.2 Cohérence pour le cas sphérique

Modèle

Dans l'étude présentée dans [CJ12], on considère que l'on dispose d'une matrice d'observations, que l'on notera $\mathbb{X}_n$, de dimension $n \times p$. Cette matrice est construite comme étant p vecteurs colonnes indépendants et de loi sphérique dans $\mathbb{R}^n$. On a donc

$$\mathbb{X}_n = \left(X^1, \ldots, X^p\right)$$

avec, pour tout $k = 1, \ldots, p$:

$$X^k \overset{i.i.d}{\sim} SP(n).$$

L'intérêt principal de ce modèle est de considérer des valeurs de p qui peuvent être très grandes. On suppose notamment ici que $p = p_n \underset{n \to +\infty}{\longrightarrow} +\infty$ tel que $p >> n$. Cela signifie que pour n grand, la matrice d'observation est rectangulaire avec sa largeur qui est d'autant plus grande que n est grand.

Remarques 3. *On constate que dans le cas sphérique, toutes les composantes de la matrice sont indépendantes. On peut donc voir la matrice d'observation de deux façons : comme une collection de p colonnes indépendantes suivant des lois sphériques de* $\mathbb{R}^n$*, ou bien comme n réalisations indépendantes d'un p-vecteur suivant une loi sphérique dans* $\mathbb{R}^p$*. Dans l'article [CJ12], le premier point de vue est privilégié. Nous verrons dans le cas dépendant que nous utiliserons la seconde interprétation.*

On calcule alors les corrélations empiriques ρ_{ij} de Pearson, définis dans (2) entre chaque couple de colonne (X^i, X^j) de la matrice $\mathbb{X}_n$. Ils sont rassemblés dans la matrice symétrique de corrélation empirique R_n.

Distribution asymptotique de L_n

Dans le cas sphérique, la distribution asymptotique de la cohérence, définie dans la formule (6), a été décrite selon trois régimes différents : régime sous-exponentiel, exponentiel et sur-exponentiel. Avant de les décrire, nous introduisons les hypothèses nécessaires :

Hypothèse 1.3.1 (Hypothèses A). *On suppose que :*
- *Les colonnes* $(X^1, ..., X^p)$ *sont des vecteurs aléatoires de* $\mathbb{R}^n$ *suivant une loi sphérique commune pouvant dépendre de* n.
- $\mathbb{P}(X^1 = 0) = 0$

On a alors les trois résultats suivant pour la cohérence L_n :

Théorème 1.3.1 (Régime sous-exponentiel). *On suppose que les hypothèses A sont vraies. On suppose que* $\lim_{n\to+\infty} [p_n] = +\infty$ *telle que* $\lim_{n\to+\infty} \frac{log(p_n)}{n} = 0$. *Alors,*

1. $L_n \xrightarrow[n\to+\infty]{\mathbb{P}} 0$
2. $n\log(1-L_n^2) + 4\log(p_n) - \log\log(p_n) \xrightarrow[n\to+\infty]{\mathcal{L}} Z$ *où* Z *est une variable aléatoire réelle qui admet pour fonction de répartition la fonction* $F(y) = 1 - \exp\left(-Ke^{y/2}\right)$ *pour tout* $y \in \mathbb{R}$ *et où* $K = \frac{1}{\sqrt{8\pi}}$.

Théorème 1.3.2 (Régime exponentiel). *On suppose que les hypothèses A sont vraies. On suppose que* $\lim_{n\to+\infty} [p_n] = +\infty$ *telle que* $\lim_{n\to+\infty} \frac{log(p_n)}{n} = \beta \in]0, +\infty[$. *Alors,*

1. $L_n \xrightarrow[n\to+\infty]{\mathbb{P}} \sqrt{1-\exp(-4\beta)}$
2. $n\log(1-L_n^2) + 4\log(p_n) - \log\log(p_n) \xrightarrow[n\to+\infty]{\mathcal{L}} Z$ *où* Z *est une variable aléatoire réelle qui admet pour fonction de répartition la fonction* $F(y) = 1 - \exp\left(-K(\beta)e^{(y+8\beta)/2}\right)$ *pour tout* $y \in \mathbb{R}$ *et où* $K(\beta) = \sqrt{\frac{\beta}{2\pi(1-\exp(-4\beta))}}$.

Théorème 1.3.3 (Régime sur-exponentiel). *On suppose que les hypothèses A sont vraies. On suppose que* $\lim_{n\to+\infty} [p_n] = +\infty$ *telle que* $\lim_{n\to+\infty} \frac{log(p_n)}{n} = +\infty$. *Alors,*

1. $L_n \xrightarrow[n\to+\infty]{\mathbb{P}} 1$ *et* $\frac{n}{\log(p_n)}\log(1-L_n^2) \xrightarrow[n\to+\infty]{\mathbb{P}} -4$
2. $n\log(1-L_n^2) + \frac{4n}{n-2}\log(p_n) - \log(n) \xrightarrow[n\to+\infty]{\mathcal{L}} Z$ *où* Z *est une variable aléatoire réelle qui admet pour fonction de répartition la fonction* $F(y) = 1 - \exp\left(-Ke^{y/2}\right)$ *pour tout* $y \in \mathbb{R}$ *et où* $K = \sqrt{\frac{1}{2\pi}}$.

Éléments de preuve dans le cas sphérique : utilisation de la méthode de Chen-Stein

Nous renvoyons à [CJ12] pour la preuve complète. Ici, nous allons expliciter comment intervient la méthode de Chen-Stein. Pour pouvoir l'appliquer, on commence par définir

l'ensemble d'indice I par :

$$I = \{(i,j) \in [\![1,p]\!] : 1 \leqslant i < j \leqslant p\}.$$

Il correspond aux couples (i,j) de la matrice R_n décrivant la partie triangulaire supérieure. Notamment, on constate que c'est sur cet ensemble que nous calculons la cohérence L_n.

Ensuite, ils considèrent l'événement $\{|\rho_{ij}| > t\}$. On fait alors le lien avec la méthode en posant $\alpha = (i,j)$ et les variables aléatoires $X_\alpha = \mathbf{1}_{|\rho_{ij}|>t}$. On a alors $W = \sum_{\alpha \in I} X_\alpha$. La méthode de Chen-Stein procure alors le résultat suivant :

$$\left|\mathbb{P}(W=0) - e^{-\lambda}\right| \leqslant b_{1,n} + b_{2,n} + b_{3,n}.$$

Or, $\mathbb{P}(W=0) = \mathbb{P}(L_n \leqslant t)$. Il reste ensuite à calculer λ qui correspond à la fonction de distribution asymptotique de L_n, et montrer que les coefficients de majorations $b_{1,n}$, $b_{2,n}$ et $b_{3,n}$ sont nuls asymptotiquement.

On rappelle que pour le cas sphérique, on applique la méthode de Chen-Stein directement à la cohérence. En effet, l'indépendance joue un rôle central pour montrer que les majorations tendent bien vers 0.

1.3.3 Cohérence pour le cas gaussien dépendant

Dans cette section, on présente les travaux antérieurs de T. Cai et T. Jiang, exposés dans [CJ11a]. Ils introduisent notamment une dépendance entre les colonnes de la matrice d'observation en se plaçant dans un cas gaussien et où la dimension p peut toujours croître vers l'infini en étant plus grande que n, mais moins rapidement que dans le cas sphérique.

Modèle

On suppose que l'on dispose d'un vecteur gaussien, de dimension p, noté $(X^1, \ldots, X^p) \sim \mathcal{N}(0, \Sigma)$ où $\Sigma \in \mathbb{R}^{p \times p}$. On suppose que la matrice Σ est construite par bande. Soit τ un entier non nul. $\Sigma = (\sigma_{ij})_{1 \leqslant i,j \leqslant p}$ vérifie :

$$\sigma_{ij} = \begin{cases} r_{ij}\sigma_i\sigma_j & \text{si} \quad |i-j| < \tau \\ 0 & \text{si} \quad \tau \leqslant |i-j| \end{cases}, \tag{1.5}$$

où les coefficients $r_{ij} \in [-1,1]$ sont des coefficients de corrélations non nuls. La matrice Σ est donc construite avec une bande centrale non nulle. On suppose donc qu'une composante du vecteur $(X^1, \ldots, X^p)$ est corrélée avec les autres composantes proches d'elle

(proche en terme d'indice dans le vecteur).

On se donne un n-échantillon de ce vecteur que l'on regroupe dans une matrice d'observation $\mathbb{X}_n \in \mathbb{R}^{n\times p}$. Cette fois, on a l'indépendance entre chaque ligne de notre matrice dûe à l'échantillonnage. Mais nous n'avons plus celle entre toutes les colonnes. De la matrice $\mathbb{X}_n$, on calcule la matrice de corrélation empirique, toujours notée $R_n = (\rho_{ij})_{1\leqslant i,j\leqslant p}$. On calcule alors la τ-cohérence, définie dans 0.0.2. Nous rappelons sa définition :

Rappel 1.3.1. *Soit $\tau \in \mathbb{N}^*$. On définit la τ-cohérence comme étant le maximum, en valeur absolue, des coefficients de corrélations empiriques à l'extérieur de la bande centrale de la matrice R_n de largeur τ. On la note $L_{n,\tau}$:*

$$L_{n,\tau} = \max_{|i-j|\geqslant\tau} |\rho_{ij}| . \tag{1.6}$$

Distribution limite de la τ-cohérence

Pour décrire la distribution asymptotique de la τ-cohérence, T.Cai et T.Jiang utilisent, comme dans le cas sphérique, la méthode de Chen-Stein. Dans ce cas, il n'y a plus d'indépendance entre toutes les composantes de la matrice. Afin de pouvoir gérer cette dépendance, la méthode n'est pas appliquée directement à la τ-cohérence mais à une variable intermédiaire plus simple. Nous appliquerons également cette technique dans le chapitre 2 de cette thèse.

Leur résultat se place toujours dans un cas de grandes dimensions dans le sens où p tend vers l'infini plus rapidement que n.

Définition 1.3.2. *Soit $\delta \in]0,1[$. On définit l'ensemble :*

$$\Gamma_{p,\delta} = \{1 \leqslant i \leqslant p : \exists j \in [\![1,p]\!], |r_{ij}| > 1-\delta, j \neq i\} \tag{1.7}$$

Théorème 1.3.4. *Supposons, pour $n \to +\infty$, que :*

1. *$p = p_n \to +\infty$ tel que $\log(p) = o\left(n^{\frac{1}{3}}\right)$*
2. *$\tau = o(p^t)$ pour tout $t > 0$*
3. *Il existe un $\delta \in]0,1[$ tel que $|\Gamma_{p,\delta}| = o(p)$*

Alors, dans le modèle considéré,

$$nL_{n,\tau}^2 - 4\log(p) + \log\log(p) \xrightarrow[n\to+\infty]{\mathcal{L}} Z \tag{1.8}$$

où Z a la fonction de distribution $F(y) = \exp\left(-\frac{1}{\sqrt{8\pi}}e^{-y/2}\right)$ pour tout $y \in \mathbb{R}$.

1.3.4 Discussion

On constate que le résultat asymptotique pour $L_{n,\tau}$ suppose que nous sommes dans un cadre de faible dépendance. En effet, l'hypothèse 3 suggère que le nombre de lignes où nous pouvons trouver un coefficient de corrélation plus élevé que le niveau $1-\delta$ est négligeable devant p. De façon heuristique, on suppose que nos variables ne sont pas trop corrélées. De plus, la bande centrale où nous pouvons trouver des coefficients non nuls reste négligeable par rapport à la matrice entière pour n (et donc p) grand. Même si τ peut croître vers ∞, il a un comportement du type logarithmique là où p peut croître de façon exponentielle.

On remarque également que la fonction de distribution asymptotique correspond à une loi de Gumbel. On rappelle la définition de celle-ci :

Définition 1.3.3. *Soit $\mu \in \mathbb{R}$ et $\beta > 0$. La variable aléatoire X suit une loi de Gumbel sur $\mathbb{R}$, de paramètres μ (position) et β (échelle), si la fonction de répartition de X est donnée par, pour tout $x \in \mathbb{R}$:*

$$F(x) = \exp\left(-\exp\left(\frac{\mu - x}{\beta}\right)\right).$$

Classiquement, la loi de Gumbel apparaît lorsqu'on s'intéresse à la loi du maximum de variables aléatoires indépendantes. En constatant que la τ-cohérence estime le maximum des coefficients de la matrice de corrélation empirique où nous avons des 0 théoriquement (et donc indépendance puisque nous sommes dans le cas gaussien), on comprend alors pourquoi $L_{n,\tau}$, convenablement normalisée, admet cette loi. En particulier, dans ce cas on a $\mu = -\log(8\pi)$ et $\beta = 2$.

Au regard du modèle de la matrice Σ, on constate que nous avons deux possibilités pour les composantes du vecteur gaussien de taille p : être corrélées ou être indépendantes. On a donc une frontière nette entre les deux. On peut alors se poser la question de savoir quel est l'impact sur la distribution asymptotique si l'on considère une bande intermédiaire sur Σ où les composantes seraient, à n fixé, corrélées mais indépendantes lorsque $n \to +\infty$.

Dans le chapitre suivant, nous étudions cette généralisation en imposant des conditions sur ces coefficients de corrélations asymptotiquement nuls.

Chapitre 2

Coherence of high-dimensional random matrices in a Gaussian case.

Ce chapitre reprend l'article en soumission écrit en collaboration avec Didier Chauveau et Marguerite Zani.

Résumé :
Cet article est dédié à l'étude de la τ-cohérence d'une matrice d'observations de dimension $(n \times p)$. La τ-cohérence est définie comme le plus grand coefficient, en valeur absolue, de la matrice de corrélation empirique associée aux observations. Nous utilisons la méthode de Chen-Stein pour démontrer la convergence en loi de la τ-cohérence, convenablement normalisée, vers une loi de Gumbel. On se place dans le cas où les observations proviennent d'un modèle gaussien dépendant. Nous supposons que la matrice de covariance du modèle est définie par bandes. Nous présentons également des simulations numériques pour mettre en évidence les problèmes que posent les grandes dimensions de nos matrices. Nous illustrons numériquement la convergence en loi de la τ-cohérence à l'aide de réplications Monte-Carlo en utilisant une stratégie HPC de découpage des matrices de corrélations de grandes tailles.

This chapter corresponds to the article written in colaboration with Didier Chauveau et Marguerite Zani.

Abstract :
This paper study the τ-coherence of a $(n \times p)$-observation matrix. The τ-coherence is defined as the largest magnitude outside a diagonal bandwith of size τ of the empirical correlation coefficients associated to our observations. Using the Chen-Stein method we derive the limiting law of the normalized coherence and show the convergence towards a Gumbel distribution when observations come from a specific dependant Gaussian case. We assume that the covariance matrix of the model has a band structure. On the other side, we provide numerical consideration highlighting issues from the high dimension hypotheses. We

numerically illustrate the asymptotic behaviour of the coherence with Monte-Carlo replications using a HPC splitting strategy for high dimensional correlation matrices.

Sommaire

2.1 Introduction

Random matrix theory has known a huge amount of breakthroughs for these last twenty years. Developments have been made in theoretical fields as well as in various applied domains. Among these applications, one can cite high-energy physics (e.g. [For10] on log–gases), electronic engineering (signal and imaging, see [Don06, CT05, CRT06b, CRT06a]), statistics (see [Joh01, Joh08]). Earlier works on random matrices were focused on spectral analysis of eigenvalues and eigenvectors (see [Wig58] or [Meh04, BS10], see also [BC12] and references therein). For a reference on random matrices theory, see [BS10, Meh04, AGZ10].

In statistics more particularly, random matrices are useful for inference in a high dimensional framework. One can think about high dimensional regression, hypothesis testing for high dimension parameters, inference for large covariance matrices. See e.g. [BS96, CT07, BJYZ09, CWX10a, CZZ10, BRT09]. In these contexts, the dimension p is much bigger than the sample size n.

We will be focusing here on the covariance structure of a certain type of random matrices. More precisely, we will be examining the coherence of random matrices with bandwise covariance of size $\tau > 1$. Let us define the model.

Let $(X^1, X^2, \ldots, X^p)$ be a p – dimensional Gaussian random vector with mean $\mu =^t (\mu^1, \cdots, \mu^p)$ in $\mathbb{R}^p$ and covariance matrix $\Sigma = (\sigma_{k,j})_{kj} \in \mathbb{R}^p \times \mathbb{R}^p$. For $n \in \mathbb{N}^*$, we consider a random sample $(X_i^1, X_i^2, \ldots, X_i^p)_{i=1,\ldots,n}$ issued from $(X^1, X^2, \ldots, X^p)$, arranged in a (n, p)-matrix $\mathbb{X}$. It means that each row can be seen as an individual and each column

as a character. We will write $\mathbb{X}^k$ the k^{th} column of $\mathbb{X}$. We are interested in the correlation terms of $\mathbb{X}$. In our model, p is much larger than n. The classical empirical Pearson's correlation coefficient is defined by :

$$\rho_{kj} = \frac{\sum_{i=1}^{n}\left(X_i^k - \overline{\mathbb{X}^k}\right)\left(X_i^j - \overline{\mathbb{X}^j}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i^k - \overline{\mathbb{X}^k}\right)^2}\sqrt{\sum_{i=1}^{n}\left(X_i^j - \overline{\mathbb{X}^j}\right)^2}} = \frac{< \mathbb{X}^k - \overline{\mathbb{X}^k}\mathbf{1}_n, \mathbb{X}^j - \overline{\mathbb{X}^j}\mathbf{1}_n >}{\parallel \mathbb{X}^k - \overline{\mathbb{X}^k}\mathbf{1}_n \parallel . \parallel \mathbb{X}^j - \overline{\mathbb{X}^j}\mathbf{1}_n \parallel}, \tag{2.1}$$

where $\mathbf{1}_n$ is the identical 1–vector in $\mathbb{R}^n$:

$$\mathbf{1}_n =^t (1, 1, \ldots, 1) \in \mathbb{R}^n$$

and $\parallel x \parallel$ stands for the the Euclidian norm of the vector x and $\overline{\mathbb{X}^k}$ is the empirical mean of the k^{th} column $\mathbb{X}^k$:

$$\overline{\mathbb{X}^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

or, equivalently if the mean μ is known,

$$\tilde{\rho}_{kj} = \frac{\sum_{i=1}^{n}\left(X_i^k - \mu\right)\left(X_i^j - \mu\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i^k - \mu\right)^2}\sqrt{\sum_{i=1}^{n}\left(X_i^j - \mu\right)^2}} = \frac{< \mathbb{X}^k - \mu^k, \mathbb{X}^j - \mu^j >}{\parallel \mathbb{X}^k - \mu^k \parallel . \parallel \mathbb{X}^j - \mu^j \parallel} \tag{2.2}$$

The empirical correlation coefficients $\rho_{k,j}$ are arranged in a (p,p)-matrix R_n (resp. $\tilde{R}_n$) which is the empirical correlation matrix of $\mathbb{X}$.

Definition 2.1.1. *With the notations above, we can define the largest magnitude of the off–diagonal terms of R_n and $\tilde{R}_n$:*

$$L_n = \max_{1\leq k<j\leq p} |\rho_{kj}|, \quad \tilde{L}_n = \max_{1\leq k<j\leq p} |\tilde{\rho}_{kj}| \tag{2.3}$$

The quantity $\tilde{L}_n$ is defined as the coherence of the matrix $\mathbb{X}_n$. With a slight abuse of terminology, we will call both L_n and $\tilde{L}_n$ coherence of $\mathbb{X}_n$.

The notion of coherence has first appeared in signal theory as an indicator of the sparsity of a matrix. More precisely, it is involved in the so-called Mutual Incoherence Property (MIP), which can be explained as follows : a mesurement (n,p) matrix X is used to recover a k–sparse signal β via linear mesurements $y = X\beta$ using a recovery algorithm. The condition

$$(2k-1)\tilde{L}_n < 1$$

ensures the exact recovery of β when β has at most k non zero entries. For details on this approach, see Donoho and Huo [DH01], Fuchs [Fuc04], Cai, Wang and Xu [CWX10b], and references therein.

Another domain where covariance and correlation matrices are highly used is statistic theory, for example testing $\Sigma = I$ against $\Sigma \neq I$. This issue has been considered in the case where n and p are of same order (i.e. $n/p \to \gamma \in (0, \infty)$) by Johnstone in the Gaussian case [Joh01], and Péché in the sub-Gaussian case [Pé09]. The test statistic relies – according to PCA methods – on the largest eigenvalue of the empirical covariance matrix $\lambda_{max}(\hat{\Sigma}_n)$. The asymptotic distribution of this maximum eigenvalue is the Tracy–Widom law.

However, testing $\Sigma = I$ can seem too restrictive, and one think about independance versus non independence in terms of correlation matrix i.e. testing $R_n = I$ against $R_n \neq I$. According to previous results, one could think about using $\lambda_{max}(R_n)$. However, even if Tracy–Widom law is conjectured in this case (see [Jia04b], and see also [HM18] for a study on the i.i.d. case), one choose to study instead the coherence as a test statistic. Jiang in [Jia04a] first adressed this problem and showed strong consistency of L_n and limit distribution of L_n^2 in the case where n and p are of the same order. Moments assumptions in [Jia04a] and dimension for p were substantially improved by a series of papers : Li and Rosalsky [LR06], Zhou [Zho07], Liu, Lin and Shao [LLS08], Li, Liu and Rosalsky [LLR10], Li, Qi and Rosalsky [LQR12]. In [CJ12] the authors consider the limiting distribution of the coherence in a spherical case. See also [CZ16a] for studies on the differential correlation matrices in high dimensional context.

Lately Cai and Jiang [CJ11a] (see also the supplement [CJ11b]) considered "ultra-high dimensions" i.e. p as large as e^{n^β}. In this paper, they also present a variant of the coherence, the so–called τ–coherence aimed to test wether the covariance Σ has a given bandwidth $\tau > 1$, where $\tau = 1$ would be a special case. We define it below :

Definition 2.1.2. *For any integer $\tau \geq 1$, we define the τ–coherence as :*

$$L_{n,\tau} = \max_{|k-j| \geqslant \tau} |\rho_{kj}| \tag{2.4}$$

In [CJ11a] strong laws and convergence of distributions of $L_{n,\tau}$ are given as well. Recently, Shao and Zhou [SZ14] studied coherence and τ coherence relaxing the normal hypothesis, improving assumptions on the moments of the entries and on the dimension p.

Our purpose is to generalize this model : we assume that $\Sigma = (\sigma_{kj})_{1 \leqslant k,j \leqslant p}$ is defined as follows :

$$\sigma_{k,j} = \begin{cases} \gamma_{kj}\sigma_k\sigma_j & \text{if} \quad |k-j| < \tau \\ \epsilon_n\sigma_k\sigma_j & \text{if} \quad \tau \leq |k-j| \leq \tau + K \\ 0 & \text{if} \quad \tau + K < |k-j| \end{cases} . \tag{2.5}$$

So, Σ is divided into three parts : a central band of size $\tau \in \mathbb{N}$, an outside part with null coefficients and a transitional bandwidth of size $K \in \mathbb{N}$. We have, for all $k \in [\![1;p]\!]$, $\sigma_k > 0$; for all $k, j \in [\![1;p]\!]$, $\gamma_{k,j} \in [-1,1]$ and $(\epsilon_n)_{n\geqslant 1}$ is a sequence of real numbers in $[-1,1]$ such that $\lim\limits_{n\to+\infty} |\epsilon_n| = 0$. To be more precise, the construction of Σ suggests that if we take two p-vectors which are close (in term of index, for example X^1 and X^2), they will be correlated. If they are far enough one to another, they will be independent. But, if they are not so close and not so far, their correlation will decrease to zero when n goes to infinity. This generalization of the model of [CJ11a] seemed to us more realistic for real data . Later, we will see that both τ and K may depend of n and may go to infinity under sufficient hypotheses. Without loss of generality, we can assume that $\mu = 0_{\mathbb{R}^p}$ and for all $k \in [\![1,p]\!]$, $\sigma_k = 1$.

All previously cited studies are highly related to the Chen-Stein method which we also use here. It relies on a Poisson approximation of weakly dependent events. For references on this method, see [AGG89]

This paper is organized as follows : section 2.2 presents the main results, further on section 2.3 gives some simulation results for our model, section 2.4 is devoted to the proof of the main result whereas the last section 2.5 gather technical results and proofs of technical lemmas.

2.2 Main result

We focus on correlations not too big, i.e. not too close to 1 or -1. Hence we define the following set :

Definition 2.2.1. *For any $\delta \in]0,1[$ we define by*

$$\Gamma_{p,\delta} = \{k \in [\![1;p]\!] : |r_{kj}| > 1-\delta \text{ for some } j \in [\![1;p]\!] \text{ and } k \neq j\},$$

where $(r_{kj})_{p\times p}$ is the correlation matrix issued from $\Sigma = (\sigma_{kj})_{p\times p}$.

The main result of the paper is the following Theorem :

Théorème 2.2.1. *Let n be an integer, $p = p_n$ a sequence such that $p_n \underset{n\to+\infty}{\longrightarrow} +\infty$. Let $(\epsilon_n)_{n\in\mathbb{N}^*}$ be a sequence of real number in $]-1,1[$. Let us assume the following conditions :*

Hyp 1 : $\log(p_n) = o(n^{\frac{1}{3}})$ *as* $n \to +\infty$

Hyp 2 : $\tau = \tau(n) = o(p_n^t)$ *as* $n \to +\infty$ *for any* $t > 0$.

Hyp 3 : $\exists \delta \in]0,1[$ *such that* $|\Gamma_{p,\delta}| = o(p_n)$ *where* $|\cdot|$ *denotes the size of the set.*

Hyp 4 : $\epsilon_n \sim \gamma\sqrt{\frac{\log(p_n)}{n}}$ *as* $n \to +\infty$ *and* $\gamma \in]-2+\sqrt{2}, 2-\sqrt{2}[$

Hyp 5 : $K = K(n) = \mathcal{O}\left(p_n^{\nu}\right)$ *where* $\nu \in]0, c(\gamma,\delta)[$
and $c(\gamma,\delta) = min\left(\frac{1}{3}(\frac{1}{2}\gamma^2 - 2|\gamma| + 1), \frac{\delta^2(2-\delta)^2}{36}\right)$.
Under these conditions, we can show that :

$$nL_{n,\tau}^2 - 4\log(p_n) + \log(\log(p_n)) \xrightarrow[n\to+\infty]{\mathcal{L}} Z \tag{2.6}$$

where Z *has the cdf* $F(y) = e^{-\frac{1}{\sqrt{8\pi}}e^{-\frac{y}{2}}}$ *for all* $y \in \mathbb{R}$.

We can observe that it is the same distribution as in [CJ11a]. The band in the covariance matrix which contains terms ε_n is a smooth transition between the central band and the external part of the matrix with null terms. This model is an illustration of vanishing dependence when the components are too far one from each other . In [CJ11a], the central band of the matrix is as large as τ with the condition of theorem 2.2.1. If we suppose $K = cst < +\infty$, we boil down to the same model. Indeed, we have a new number $\tilde{\tau} = \tau + K$ wich is the new width of the non–null bandwidth, and $\tilde{\tau}$ is still such that :

$$\forall t > 0,\ \tilde{\tau} = o(p^t) \text{ as } n \to +\infty \tag{2.7}$$

2.3 Numerical aspects

In this section, we provide some simulated examples to illustrate the behavior of our asymptotic result in practical simulations (n and p both large but finite). For this, we use the R Statistical Software [R C20]. A difficulty comes from the fact that in our context, we have to compute correlations of large matrices. We need Gaussian observation matrices of size $n \times p$ with $\log(p) = o\left(n^{\frac{1}{3}}\right)$. It means that for a large n, for example $n = 4000$, we will have $p \approx 45000$ taking $p = \left[\exp\left(n^{\frac{1}{3.5}}\right)\right]$ in our simulations (where $[x]$ is the integer part of x). For each $(n \times p)$-observation matrix, we have to compute the $(p \times p)$-correlation matrix to compute the τ-coherence. For the range of p that we consider, we can observe the evolution of the size of the $(p \times p)$-matrix in Gb according to n in Figure 2.1. For example, with $n = 4000$ and $p = 44112$, we have, for correlation stored in double, a $14.5Gb$ $(p \times p)$-matrix which is very large for a common computer. We must find a way to compute the τ-coherence without loading the entire $(p \times p)$-correlation matrix in the computer memory (RAM).

The idea is to generate the $(n \times p)$-observation matrix by packets of columns. Each packet will have a size $(n \times Tb)$ where Tb is choosen by the user. With these packets of columns, we compute all correlation blocks of size $(Tb \times Tb)$ between each pair of packets

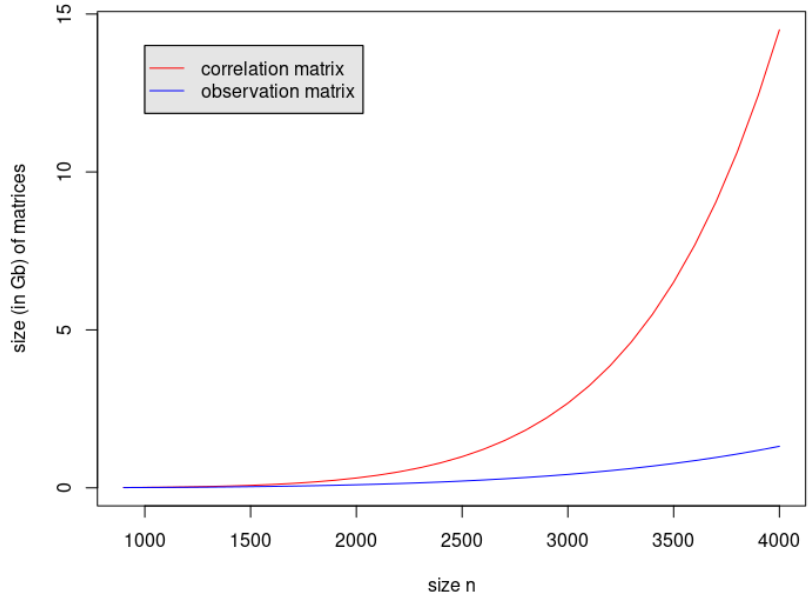

FIGURE 2.1 – Size of $(p \times p)$-correlation matrix and $(n \times p)$-observation matrix according to n with $p = \left[\exp(n^{1/3.5})\right]$ for real numbers stored as double precision numbers.

of columns. In that way, we must choose a size Tb in order to have two blocks fitting simultaneously in the computer memory. Then, we can compute the τ-coherence by taking the largest coefficient in absolute value in our block paying attention to wether the block corresponds to the central band (with bandwith τ) or not.

Using this strategy, we can generate correlation matrices even if p is very large, so we are able to study the limiting distribution of the τ-coherence. In that way, to illustrate our theorem, we consider the following parameters :

$$p = \left[\exp\left(n^{1/3.5}\right)\right] \ | \ \tau = 5 * [\log(p)] \ | \ K = 10 * \left[n^{1/10}\log(p)\right] \ | \ \varepsilon_n = 0.1 * \sqrt{\frac{\log(p)}{n}}$$

Our purpose here is to simulate a sample of τ-coherence by a Monte-Carlo procedure in order to compare its empirical distribution with the asymptotique one. We thus run $R = 200$ replications of the following procedure, simulating R times the matrices of observations and computing the correlations per blocks. For each replications, we generate an observation matrix $\mathbf{X}$ of size $(n \times p)$ using the following numerical scheme :

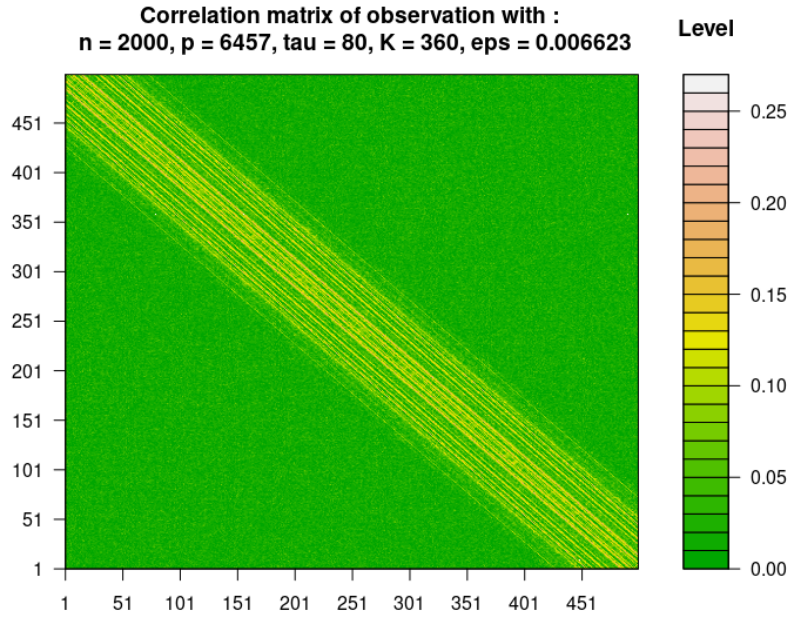

FIGURE 2.2 – Level plot of the correlation's structure with observations : zoom on the square $1:500$

$$\forall i \in [\![1,n]\!], \forall j \in [\![1,p]\!], X_i^j = \sum_{k=j}^{j+K-1} \varepsilon_n Y_i^k + \sum_{k=j+K}^{j+K+2\tau} r_k Y_i^k + \sum_{j+K+2\tau+1}^{j+2\tau+2K} \varepsilon_n Y_i^k$$

where all coefficients $(r_k)_{1\leqslant k\leqslant 1+2\tau}$ are real numbers in $[-1,1]$ (we take $r_1, \dots r_{1+2\tau} \overset{i.i.d}{\sim} \mathcal{U}_{[-1,1]}$ in the simulation), X_i^j is the coefficient of $\mathbf{X}$ on the i^{th} line and the j^{th} column and all random variable $Y_i^k \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ arranged in a $(n \times (p+2\tau+2K))$-matrix $\mathbf{Y}$. We highlight the fact that Y is quite larger than X. This numerical scheme is inspired by time series model.

We can observe that we generate data following our model and obtain an observation matrix associated to a correlation matrix with a band structure in Figure 2.2. We recognize a central band with non-null coefficients. In fact, we also notice that the transition band with ε_n' coefficients is not really recognizable but this is due to the fact that those coefficients are decreasing fastly to 0 when n goes to infinity (for instance, here, we have $\varepsilon_n \approx 0.007$ not different from 0 in the color scale).

With this observation matrix, we can use our procedure to compute the τ-coherence.

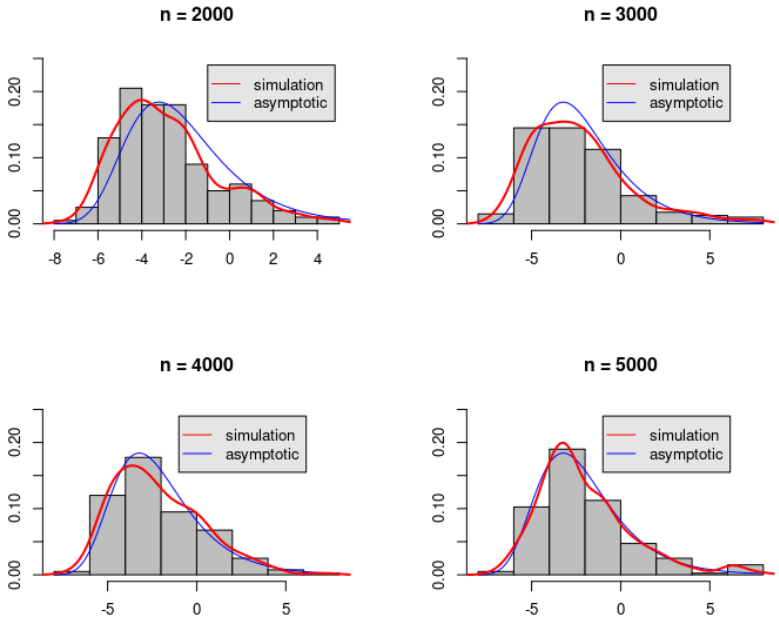

FIGURE 2.3 – Histograms and Kernel density estimates for $n = 2000, 3000, 4000, 5000$ and for $R = 200$ replications

After running R replications, we obtain a sample of τ-coherence. In Figure 2.3, we see that for n large enough, the sample distribution seems to approximate the limiting one.

Precisely, we compare the estimated density of the sample (in red) with the asymptotic density (in blue) which is defined by $f(x) = \frac{1}{2\sqrt{8\pi}} \exp\left(-\frac{1}{2}y - \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{2}y\right)\right)$ for all $x \in \mathbb{R}$. Also, in order to observe the convergence, we study numerically the distance between the sample and asymptotic distribution. We use the Kolmogorov, $\mathbb{L}^2$ and the Total Variation norms. We remind, respectively, the definition of these norms :

$$d_{KS}(\hat{f}, f) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|, \tag{2.8}$$

$$d_2(\hat{f}, f) = \int \left|\hat{f}(x) - f(x)\right|^2 dx, \tag{2.9}$$

$$d_{TV}(\hat{f}, f) = \frac{1}{2} \int \left|\hat{f}(x) - f(x)\right| dx. \tag{2.10}$$

We observe, in the results displayed in Figure 2.4, that the difference between both distributions decreases to 0 when n is increasing.

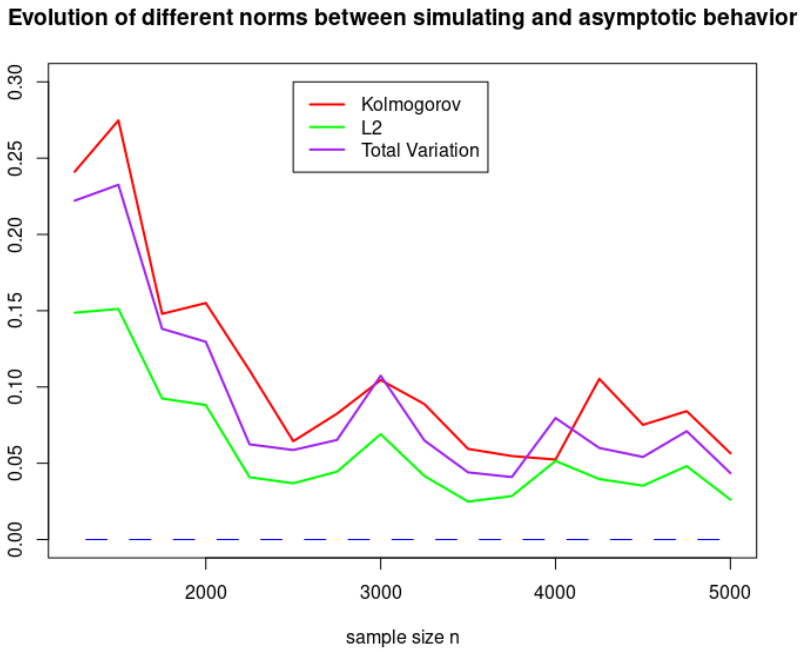

FIGURE 2.4 – Evolution of Kolmogorov, $\mathbb{L}^2$ and Total Variation norm between simulating and asymptotic behavior

These results provide numerical evidence that our limiting distribution is adequate. We also higlight the fact that the procedure we proposed here allows to compute τ-coherence corresponding to any large matrix X arising in actual (big) data experiments. However, this procedure is not very efficient if it is done with a classical programming. For example, computing only one replication for $n = 4000$ (and so $p = 44112$), requires about 90 min to obtain the value of one τ-coherence. In order to obtain more usable (i.e. fast) codes in perspective of real-size applications, we are currently exploring HPC strategies to compute correlation blocks using GPGPU computation. We are very confident into the use of GPU to reduce simulation's time.

2.4 Proof of the main result

In this part, we will describe the proof of our main result. First, we would like to higlight the fact that, as we said, we will apply the Chen-Stein method. But, we do not apply it directly to the τ-coherence. It will be more efficient to use the Chen-Stein method to a new easier to handle random variable.

First of all, we introduce many notation which will be used along this paper.

2.4.1 Notations

— $I = \{(k,j) \in [\![1,p]\!]^2 :\ 1 \leqslant k < j \leqslant p\}$
— $I_\tau = \{(k,j) \in I : |k-j| < \tau\}$
— $I_K = \{(k,j) \in I : \tau \leq |k-j| \leq \tau + K\}$
— $I_0 = \{(k,j) \in I : |k-j| > \tau + K\}$
— $E_\delta = \{(k,j) \in I : k \in \Gamma_{p,\delta} \text{ or } j \in \Gamma_{p,\delta}\}$
— $\Lambda_p^\tau = \{(k,j) \in I : |k-j| < \tau \text{ and } \max_{1\leqslant k\neq q, j\neq q\leqslant p} (|r_{kq}|, |r_{jq}|) \leqslant 1-\delta\}$
— $\Lambda_p^K = \{(k,j) \in I : \tau \leq |k-j| \leq \tau + K \text{ and } \max_{1\leqslant k\neq q, j\neq q\leqslant p} (|r_{kq}|, |r_{jq}|) \leqslant 1-\delta\}$
— $\Lambda_p^0 = \{(k,j) \in I : |k-j| > \tau + K \text{ and } \max_{1\leqslant k\neq q, j\neq q\leqslant p} (|r_{kq}|, |r_{jq}|) \leqslant 1-\delta\}$

With all these different sets, we can write three different partitions of the set I :

1. $I = I_\tau \cup I_K \cup I_0$
2. $I = E_\delta \cup \Lambda_p^\tau \cup \Lambda_p^K \cup \Lambda_p^0$
3. $I_0 \cup I_K = \Lambda_p^K \cup \Lambda_p^0 \cup \left[E_\delta \cap \overline{I_\tau}\right]$

The following Lemma gives the sizes of these three sets :

Lemma 2.4.1. *With the previous notations*

$$|I_\tau| = (\tau - 1)\left(\frac{2p-\tau}{2}\right) \tag{2.11}$$

$$|I_K| = (K+1)\left(\frac{2p-K-2\tau}{2}\right) \tag{2.12}$$

$$|I_0| = \frac{(p-\tau-K-1)(p-\tau-K)}{2} \tag{2.13}$$

2.4.2 Auxiliary variables

Now we introduce an auxiliary random variable which will be more convenient to handle in the Chen–Stein method. Let

$$V_{n,\tau} = \max_{1\leqslant k<j\leqslant p, |k-j|\geqslant\tau} |{}^tX^kX^j| = \max_{\alpha=(k,j)\in I_0\cup I_K} \left|{}^tX^kX^j\right| \tag{2.14}$$

Proposition 2.4.1. *Under the assumptions of Theorem 2.2.1, we have*

$$\frac{n^2L_{n,\tau}^2 - V_{n,\tau}^2}{n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \tag{2.15}$$

The proof of this Proposition is postponed to section 2.5.

Hence to study the asymptotic behaviour of $L_{n,\tau}$, it is enough to study the limiting distribution of $V_{n,\tau}$. To do so, we use another slightly different random variable defined by :

$$V'_{n,\tau} = \max_{\alpha \in \Lambda_p^0 \cup \Lambda_p^K} (Z_\alpha) \tag{2.16}$$

where the index $\alpha = (k, j)$ and $Z_\alpha = Z_{kj} = \left|{}^t X^k X^j\right|$.

The two variables $V_{n,\tau}$ and $V'_{n,\tau}$ are linked by the following inequalities :

Proposition 2.4.2. *Let*

$$a_n(y) = \sqrt{4n\log(p_n) - n\log\log(p_n) + ny} \text{ with } y \in \mathbb{R} \tag{2.17}$$

We have :

$$\mathbb{P}\left(V'_{n,\tau} > a_n(y)\right) \leqslant \mathbb{P}\left(V_{n,\tau} > a_n(y)\right) \leqslant \mathbb{P}\left(V'_{n,\tau} > a_n(y)\right) + o(1) \tag{2.18}$$

Démonstration. (For seek of simplicity in the remaining of the paper we will denote $a_n(y)$ by a_n).
To proove this result, we need the two following technical results which proofs are postponed to section 2.4.

Lemma 2.4.2. *Let a_n be as in formula* (2.17).
Then,

$$\mathbb{P}_0 := \mathbb{P}\left(\left|{}^t X^1 X^{\tau+K+2}\right| > a_n\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{p_n^2}\left(1 + o(1)\right) = \mathcal{O}_{n\to+\infty}\left(\frac{1}{p_n^2}\right) \tag{2.19}$$

Lemma 2.4.3. *Let a_n be as in formula* (2.17) *and let us define $c_\gamma := \frac{1}{2}\gamma^2 - 2|\gamma| + 2$ with γ defined in theorem 2.2.1*
Then, for any $d \in [0; c_\gamma[$ and as $n \to +\infty$:

$$\mathbb{P}_K := \mathbb{P}\left(\left|{}^t X^1 X^{\tau+1}\right| > a_n\right) = o\left(p_n^{-d}\right) \tag{2.20}$$

According to the partition $I_0 \cup I_K = \Lambda_p^0 \cup \Lambda_p^K \cup \left(E_\delta \cap \overline{I_\tau}\right)$,

$$\begin{aligned}
\mathbb{P}\left(V_{n,\tau} > a_n\right) &= \mathbb{P}\left(\max_{\alpha=(k,j)\in I_0\cup I_K} \left|{}^t X^k X^j\right| > a_n\right) \\
&\leqslant \mathbb{P}\left(V'_{n,\tau} > a_n\right) + \mathbb{P}\left(\max_{\alpha=(k,j)\in E_\delta\cap\overline{I_\tau}} \left|{}^t X^k X^j\right| > a_n\right) \\
&\leqslant \mathbb{P}\left(V'_{n,\tau} > a_n\right) + \sum_{\alpha=(k,j)\in E_\delta\cap\overline{I_\tau}} \mathbb{P}\left(\left|{}^t X^k X^j\right| > a_n\right) \\
&\leqslant \mathbb{P}\left(V'_{n,\tau} > a_n\right) + \sum_{\alpha\in\left[E_\delta\cap\overline{I_\tau}\right]\cap I_K} \mathbb{P}\left(Z_\alpha > a_n\right) + \sum_{\alpha\in\left[E_\delta\cap\overline{I_\tau}\right]\cap I_0} \mathbb{P}\left(Z_\alpha > a_n\right)
\end{aligned}$$

All variables Z_α having same distributions in the different sets above, we have

$$\begin{aligned}\mathbb{P}\left(V_{n,\tau} > a_n\right) &\leqslant \mathbb{P}\left(V'_{n,\tau} > a_n\right) + \left|\left[E_\delta \cap \overline{I_\tau}\right] \cap I_K\right| \mathbb{P}\left(Z_{1,\tau+1} > a_n\right) + \left|\left[E_\delta \cap \overline{I_\tau}\right] \cap I_0\right| \mathbb{P}\left(Z_{1,\tau+K+2} > a_n\right) \\ &\leqslant \mathbb{P}\left(V'_{n,\tau} > a_n\right) + |I_K|\,\mathbb{P}_K + |E_\delta|\,\mathbb{P}_0\end{aligned}$$

We can use the following result

Lemma 2.4.4.

$$|E_\delta| \leqslant 2p_n|\Gamma_{p,\delta}|$$

And from assumption 3 of theorem 2.2.1, we have :

$$|E_\delta| = o(p_n^2) \tag{2.21}$$

Now, we need to prove that $|I_K|\,\mathbb{P}_K + |E_\delta|\,\mathbb{P}_0 \underset{n\to+\infty}{\longrightarrow} 0$. First, from lemma 2.4.2 and (2.21),

$$\begin{aligned}|E_\delta|\,\mathbb{P}\left(Z_{1,\tau+K+2} > a_n\right) &\underset{n\to+\infty}{\sim} |E_\delta|\,\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}\frac{1}{p_n^2} \\ &\underset{n\to+\infty}{=} o(p_n^2)\frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}\frac{1}{p_n^2} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}o(1) \underset{n\to+\infty}{\longrightarrow} 0\end{aligned}$$

Secondly, using the lemma 2.4.1 (more precisely eq. (2.12)), lemma 2.4.3 we have :

$$|I_K|\,\mathbb{P}_K \underset{n\to+\infty}{\longrightarrow} 0 \quad\Leftrightarrow\quad \nu < c_\gamma - 1 \tag{2.22}$$

$$\tag{2.23}$$

and this is fullfilled from assumptions on theorem 2.2.1. Finally, we obtain :

$$|I_K|\,\mathbb{P}_K + |E_\delta|\,\mathbb{P}_0 \underset{n\to+\infty}{\longrightarrow} 0 \tag{2.24}$$

Then,

$$\mathbb{P}\left(V_{n,\tau} > a_n\right) \leqslant \mathbb{P}\left(V'_{n,\tau} > a_n\right) + o(1) \tag{2.25}$$

Also, it is easy to see that :

$$\mathbb{P}\left(V'_{n,\tau} > a_n\right) \leqslant \mathbb{P}\left(V_{n,\tau} > a_n\right) \tag{2.26}$$

□

Remarks 1. *We keep in mind here that the main constraint is $p_nK\mathbb{P}_K \to 0$ when $n \to \infty$ which leads to*

$$\nu < \frac{1}{2}\gamma^2 - 2|\gamma| + 1$$

hence it also implies

$$\gamma \in [-2,2] \text{ is such that } \frac{1}{2}\gamma^2 - 2|\gamma| + 1 > 0 \iff \gamma \in]-2+\sqrt{2}; 2-\sqrt{2}[$$

2.4.3 Chen–Stein method for $V'_{n,\tau}$

So now, we focus on the asymptotic behaviour of $V'_{n,\tau}$. For that purpose, we apply the Chen-Stein method. We remind this result in the following lemma (it can be found in [AGG89]) :

Lemma 2.4.5. *The Chen-Stein Method*
Let $\mathcal{I}$ be a set of index. Let $\alpha \in \mathcal{I}$ and B_α a set of subset of $\mathcal{I}$ (i.e. for all α, $B_\alpha \subset \mathcal{I}$). Let η_α be random variables. For a given $t \in \mathbb{R}$, we define $\lambda := \sum_{\alpha \in \mathcal{I}} \mathbb{P}(\eta_\alpha > t)$. Then,

$$\left| \mathbb{P}\left(\max_{\alpha \in \mathcal{I}} (\eta_\alpha) \leqslant t \right) - e^{-\lambda} \right| \leqslant \min\left(1, \frac{1}{\lambda}\right) . (b_1 + b_2 + b_3) \tag{2.27}$$

where

- $b_1 = \sum_{\alpha \in \mathcal{I}} \sum_{\beta \in B_\alpha} \mathbb{P}(\eta_\alpha > t)\mathbb{P}(\eta_\beta > t)$
- $b_2 = \sum_{\alpha \in \mathcal{I}} \sum_{\alpha \neq \beta \in B_\alpha} \mathbb{P}(\eta_\alpha > t, \eta_\beta > t)$
- $b_3 = \sum_{\alpha \in \mathcal{I}} \mathbb{E}\left[\left| \mathbb{E}[\mathbf{1}_{\eta_\alpha > t} | \sigma(\eta_\beta, \beta \in \mathcal{I} \backslash B_\alpha)] - \mathbb{E}[\mathbf{1}_{\eta_\alpha > t}] \right| \right]$

As we said, this method is an approximation of weekly dependent events by a Poisson law which is represented by the quantity $e^{-\lambda}$ (corresponding to $\mathbb{P}(Z = 0)$, Z having a Poisson law $\mathcal{P}(\lambda)$). We need to find weekly dependent events to have b_1, b_2 and b_3 small (even null or asymptotically null).

In our case, notations are :

- $\Lambda = \Lambda_p^0 \cup \Lambda_p^K$.
- $\alpha = (k, j) \in \Lambda$.
- $B_\alpha = B_{kj} = \{(u, v) \in \Lambda : |k - u| < \tau + K, |j - v| < \tau + K \text{ and } (k, j) \neq (u, v)\}$.
- $\eta_\alpha = Z_\alpha = Z_{kj} = \left| {}^t X^k X^j \right| = \left| \sum_{i=1}^{n} X_i^k X_i^j \right|$.
- $\lambda_n = \sum_{\alpha \in \Lambda} \mathbb{P}(Z_\alpha > a_n)$.
- $b_{1,n} = \sum_{\alpha \in \Lambda} \sum_{\beta \in B_\alpha} \mathbb{P}(Z_\alpha > a_n)\mathbb{P}(Z_\alpha > a_n)$
- $b_{2,n} = \sum_{\alpha \in \Lambda} \sum_{\alpha \neq \beta \in B_\alpha} \mathbb{P}(Z_\alpha > a_n, Z_\beta > a_n)$.
- $b_{3,n} = \sum_{\alpha \in \Lambda} \mathbb{E}\left[\left| \mathbb{E}\left[\mathbf{1}_{Z_\alpha > a_n} | \sigma\left(Z_\beta, \beta \in \Lambda \backslash B_\alpha\right) \right] - \mathbb{E}\left[\mathbf{1}_{Z_\alpha > a_n} \right] \right| \right]$

First of all, we compute λ_n to assure it converges (as $n \to +\infty$) to a finite value. Then, we compute $b_{1,n}$, $b_{2,n}$ and $b_{3,n}$. Let us start with a preliminar lemma.

Lemma 2.4.6. *Considering the previous notations, with straightforward computations we obtain the following results :*

- $|\Lambda_p^0| \sim p_n^2/2$ *as* $n \to +\infty$
- $|B_{ij}| \leqslant 8(\tau + K)p_n \sim 8Kp_n$ *as* $n \to +\infty$
- $|\Lambda_p^K| \leqslant |I_K|$

Computation of λ_n

According to the Chen-Stein method and using the fact that random variables have the same law when indices are in the same set, we have

$$\lambda_n = \sum_{\alpha\in\Lambda_p^0\cup\Lambda_p^K} \mathbb{P}\left(Z_\alpha > a_n\right) = \sum_{\alpha\in\Lambda_p^0} \mathbb{P}\left(Z_\alpha > a_n\right) + \sum_{\alpha\in\Lambda_p^K} \mathbb{P}\left(Z_\alpha > a_n\right) = |\Lambda_p^0|.\mathbb{P}_0 + |\Lambda_p^K|.\mathbb{P}_K$$

According to theorem 2.2.1, lemma 2.4.3 and lemma 2.4.6, we have

$$\lim_{n\to+\infty} |\Lambda_p^K|.\mathbb{P}_K = 0$$

while we have, according to lemma 2.4.2 and lemma 2.4.6,

$$|\Lambda_p^0|.\mathbb{P}_0 \underset{+\infty}{\sim} p_n^2 \frac{1}{p_n^2}\frac{1}{\sqrt{8\pi}}e^{y/2} = \frac{1}{\sqrt{8\pi}}e^{y/2}.$$

Finally, we obtain :

$$\lim_{n\to+\infty} (\lambda_n) = \frac{1}{\sqrt{8\pi}}e^{y/2} \tag{2.28}$$

This quantity appears in the distribution function of the asymptotic behaviour. Let us move now to the computation of $b_{1,n}$.

Computation of $b_{1,n}$

We add some notations :
— $B_\alpha^0 := B_\alpha \cap \Lambda_p^0$ and $|B_\alpha^0| \leqslant |B_\alpha| \leqslant 8(\tau+K)p_n$
— $B_\alpha^K := B_\alpha \cap \Lambda_p^K$ and $|B_\alpha^K| \leqslant K^2$
— $\mathbb{P}_\alpha := \mathbb{P}\left(Z_\alpha > a_n\right)$

As used above, for α taken in a same set, the values of the probabilities are the same. Then, we have :

$$\begin{aligned}
b_{1,n} &= \sum_{\alpha\in\Lambda_p^0\cup\Lambda_p^K}\sum_{\beta\in B_\alpha} \mathbb{P}_\alpha\mathbb{P}_\beta \\
&= \sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in B_\alpha^0} \mathbb{P}_\alpha\mathbb{P}_\beta + \sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in B_\alpha^K} \mathbb{P}_\alpha\mathbb{P}_\beta + \sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^0} \mathbb{P}_\alpha\mathbb{P}_\beta + \sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^K} \mathbb{P}_\alpha\mathbb{P}_\beta \\
&= \sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in B_\alpha^0} (\mathbb{P}_0)^2 + \sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in B_\alpha^K} \mathbb{P}_0\mathbb{P}_K + \sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^0} \mathbb{P}_K\mathbb{P}_0 + \sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^K} (\mathbb{P}_K)^2 \\
&= |\Lambda_p^0|.|B_\alpha^0|.(\mathbb{P}_0)^2 + |\Lambda_p^0|.|B_\alpha^K|\mathbb{P}_0\mathbb{P}_K + |\Lambda_p^K|.|B_\alpha^0|\mathbb{P}_K\mathbb{P}_0 + |\Lambda_p^K|.|B_\alpha^K|.(\mathbb{P}_K)^2
\end{aligned}$$

At this point, we need to check that $\lim_{n\to+\infty}(b_{1,n}) = 0$, so will focus particulary on :

1. $|\Lambda_p^0|.|B_\alpha^0|.\left(\mathbb{P}_0\right)^2$:

$$|\Lambda_p^0|.|B_\alpha^0|.\left(\mathbb{P}_0\right)^2 \sim \frac{1}{2}p_n^2.|B_\alpha^0|.\left(\mathbb{P}_0\right)^2 \leqslant 4(\tau+K)p_n^3.\mathcal{O}\left(\frac{1}{p_n^4}\right) = \mathcal{O}\left(p_n^{\nu-1}\right) \quad (2.29)$$

$$(2.30)$$

From assumptions on ν we have $\lim_{n\to+\infty}\left[|\Lambda_p^0|.|B_\alpha^0|.\left(\mathbb{P}_0\right)^2\right] = 0$

2. $|\Lambda_p^0|.|B_\alpha^K|\mathbb{P}_0\mathbb{P}_K$:

$$\begin{aligned}|\Lambda_p^0|.|B_\alpha^K|\mathbb{P}_0\mathbb{P}_K &\sim \frac{1}{2}p_n^2|B_\alpha^K|\mathbb{P}_0\mathbb{P}_K \leqslant \frac{1}{2}K^2p_n^2\mathbb{P}_0\mathbb{P}_K \leqslant \mathcal{O}\left(p_n^{2+2\nu}\right)\mathcal{O}\left(p_n^{-2}\right)\mathbb{P}_K \\ &\leqslant \mathcal{O}\left(p_n^{2\nu}\mathbb{P}_K\right)\end{aligned}$$

According to lemma 2.4.3, we will have $\lim_{n\to+\infty}[p_n^{2\nu}\mathbb{P}_K] = 0$ iff $2\nu < c_\gamma$ which is true from hypothesis 5 in theorem 2.2.1. Then, we obtain :

$$\lim_{n\to+\infty}\left[|\Lambda_p^0|.|B_\alpha^K|\mathbb{P}_0\mathbb{P}_K\right] = 0$$

3. $|\Lambda_p^K|.|B_\alpha^0|\mathbb{P}_K\mathbb{P}_0$: We use the same principle of computation than previsouly :

$$|\Lambda_p^K|.|B_\alpha^0|\mathbb{P}_K\mathbb{P}_0 \leqslant p_nK|B_\alpha^0|\mathbb{P}_K\mathbb{P}_0 \leqslant 8p_nK(\tau+K)p_n\mathbb{P}_K\mathbb{P}_0 = \mathcal{O}\left(p_n^{2\nu}\mathbb{P}_K\right) \quad (2.31)$$

So, from previous assumptions on ν, we have :

$$\lim_{n\to+\infty}\left[|\Lambda_p^K|.|B_\alpha^0|\mathbb{P}_K\mathbb{P}_0\right] = 0\,.$$

4. $|\Lambda_p^K|.|B_\alpha^K|.\left(\mathbb{P}_K\right)^2$: We have :

$$|\Lambda_p^K|.|B_\alpha^K|.\left(\mathbb{P}_K\right)^2 \leqslant pK^3\left(\mathbb{P}_K\right)^2 = \mathcal{O}\left(p_n^{1+3\nu}\right)\left(\mathbb{P}_K\right)^2 \quad (2.32)$$

According to lemma 2.4.3, if $1+3\nu < 2C_\gamma$, we have

$$\lim_{n\to+\infty}\left[|\Lambda_p^K|.|B_\alpha^K|.\left(\mathbb{P}_K\right)^2\right] = 0\,.$$

To conclude, we finally obtain :

$$\lim_{n\to+\infty}[b_{1,n}] = 0\,.$$

Remarks 2. *The main constraint here is* $p_nK^3\left(\mathbb{P}_K\right)^2 \to 0$ *which is true from condition* $p_nK\mathbb{P}_K \to 0$ *of remark 1.*

Computation of $b_{2,n}$

The computation of $b_{2,n}$ is the most technical part. As we did for the computation of $b_{1,n}$, we will divide this computation into four parts (according on which set we are). We remind the definition of $b_{2,n}$:

$$b_{2,n} = \sum_{\alpha\in\Lambda_p^0\cup\Lambda_p^K}\sum_{\beta\in B_\alpha}\mathbb{P}\left(Z_\alpha > a_n, Z_\beta > a_n\right) .$$

For this part,we introduce new notations :

— $\mathbb{P}_{\alpha\beta} := \mathbb{P}\left(Z_\alpha > a_n, Z_\beta > a_n\right)$

— $\mathbb{P}_{0i} := \mathbb{P}_{\alpha\beta}\mathbf{1}_{\alpha\in\Lambda_p^0}\mathbf{1}_{\beta\in\Omega_i}$ where Ω_i will be a subset of index and i an integer.

— $\mathbb{P}_{Ki} := \mathbb{P}_{\alpha\beta}\mathbf{1}_{\alpha\in\Lambda_p^K}\mathbf{1}_{\beta\in\Omega_i}$ where Ω_i will be a subset of index and i an integer.

To show that $\lim_{n\to+\infty}[b_{2,n}] = 0$, we will divide it into four sums, each one being the sum of the same probability on a given set of indices. Then, we have :

$$b_{2,n} = \underbrace{\sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in B_\alpha^0}\mathbb{P}_{\alpha\beta}}_{:=Q_1} + \underbrace{\sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in B_\alpha^K}\mathbb{P}_{\alpha\beta}}_{:=Q_2} + \underbrace{\sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^0}\mathbb{P}_{\alpha\beta}}_{:=Q_3} + \underbrace{\sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^K}\mathbb{P}_{\alpha\beta}}_{:=Q_4} \tag{2.33}$$

Computation of Q_1 :

First, we define some subset of indices. In particular, we have :

1. $\Omega_1 := \{(u,v)\in\Lambda_p^0 : i-u<\tau \text{ and } j-v<\tau\}$ and $|\Omega_1|\leqslant\tau^2$
2. $\Omega_2 := \{(u,v)\in\Lambda_p^0 : i-u<\tau \text{ and } \tau<j-v<\tau+K\}$ and $|\Omega_2|\leqslant\tau K$
3. $\Omega_3 := \{(u,v)\in\Lambda_p^0 : \tau<i-u<\tau+K \text{ and } j-v<\tau\}$ and $|\Omega_3|\leqslant\tau K$
4. $\Omega_4 := \{(u,v)\in\Lambda_p^0 : i-u<\tau \text{ and } \tau+K\leqslant j-v\}$ and $|\Omega_4|\leqslant\tau\left(p_n-\tau-K\right)\leqslant\tau p_n$
5. $\Omega_5 := \{(u,v)\in\Lambda_p^0 : \tau+K\leqslant i-u \text{ and } j-v<\tau\}$ and $|\Omega_5|\leqslant\tau\left(p_n-\tau-K\right)\leqslant\tau p_n$
6. $\Omega_6 := \{(u,v)\in\Lambda_p^0 : \tau<i-u<\tau+K \text{ and } \tau<j-v<\tau+K\}$ and $|\Omega_6|\leqslant K^2$
7. $\Omega_7 := \{(u,v)\in\Lambda_p^0 : \tau<i-u<\tau+K< \text{ and } \tau+K\leqslant j-v\}$ and $|\Omega_7|\leqslant K\left(p_n-\tau-K\right)\leqslant Kp_n$
8. $\Omega_8 := \{(u,v)\in\Lambda_p^0 : \tau+K\leqslant i-u \text{ and } j-v<\tau\}$ and $|\Omega_8|\leqslant K\left(p_n-\tau-K\right)\leqslant Kp_n$

We have, :

$$Q_1 \leqslant 4\sum_{i=1}^{8}\sum_{\alpha\in\Lambda_p^0}\sum_{\beta\in\Omega_i}\mathbb{P}_{\alpha\beta} \tag{2.34}$$

Then, we use the fact that on each subset, random variables have the same law, we have :

$$\begin{aligned} Q_1 &\leqslant \left|\Lambda_p^0\right| . |\Omega_1| \, \mathbb{P}_{01} + \left|\Lambda_p^0\right| . |\Omega_2| \, \mathbb{P}_{02} + \left|\Lambda_p^0\right| . |\Omega_3| \, \mathbb{P}_{03} + \left|\Lambda_p^0\right| . |\Omega_4| \, \mathbb{P}_{04} \quad (2.35) \\ &+ \left|\Lambda_p^0\right| . |\Omega_5| \, \mathbb{P}_{05} + \left|\Lambda_p^0\right| . |\Omega_6| \, \mathbb{P}_{06} + \left|\Lambda_p^0\right| . |\Omega_7| \, \mathbb{P}_{07} + \left|\Lambda_p^0\right| . |\Omega_8| \, \mathbb{P}_{08} \quad (2.36) \end{aligned}$$

So, we just have to show that each part will have a null limit when n is going to infinity.

Lemma 2.4.7. *Using the previous notations, we have, as $n \to +\infty$:*

$$\left|\Lambda_p^0\right| . |\Omega_1| \, \mathbb{P}_{01} \to 0 \quad (2.37)$$

Démonstration. We have :

$$\left|\Lambda_p^0\right| . |\Omega_1| \, \mathbb{P}_{01} \leqslant \left|\Lambda_p^0\right| \tau^2 \mathbb{P}_{01} \sim \frac{1}{2} p_n^2 \tau^2 \mathbb{P}_{01} = o\left(p_n^{2+2t} \mathbb{P}_{01}\right) \text{ for any } t > 0$$

where we use the lemma 2.4.6 for the equivalent. We can write :

$$\begin{aligned} \mathbb{P}_{01} &= \mathbb{P}\left(\left| \sum_{k=1}^{n} u_k^1 u_k^2 \right| > a_n, \left| \sum_{k=1}^{n} u_k^3 u_k^4 \right| > a_n \right) \quad (2.38) \\ & \quad (2.39) \end{aligned}$$

where $(u_k^1, u_k^2, u_k^3, u_k^4)_{1 \leqslant k \leqslant n} \overset{i.i.d}{\sim} \mathcal{N}_4(0, \Sigma_4)$ and

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & r_1 & 0 \\ 0 & 1 & 0 & r_2 \\ r_1 & 0 & 1 & 0 \\ 0 & r_2 & 0 & 1 \end{pmatrix},$$

where coefficients r_1, r_2 are from the correlation matrix (r_{kj}). From Lemma 6.11 of [CJ11a], focusing on equation (131), we know that

$$\mathbb{P}_{01} \leqslant \mathcal{O}\left(p_n^{-2b^2+\varepsilon_1}\right) + \mathcal{O}\left(p_n^{-2-2c^2+\varepsilon_2}\right) \text{ as } n \to +\infty \quad (2.40)$$

for any $\varepsilon_1, \varepsilon_2 > 0$ and where $a = \frac{1+(1-\delta)^2}{2}$, $b = \frac{a}{(1-\delta)^2}$ and $c = \frac{1-a}{3}$ for $\delta \in]0,1[$. By construction $b^2 - 1 > 0$, hence for a well-chosen t such that $t < b^2 - 1$, there exists $\varepsilon_1(\delta) > 0$ such that we have :

$$\varepsilon_1 < 2b^2 - 2 - 2t \, . \quad (2.41)$$

Analogously, we can find ε_2 such that :

$$\varepsilon_2 < 2\left(c^2 - t\right) . \tag{2.42}$$

Since $\tau = o(p^t)$ for any $t > 0$, we have

$$\left|\Lambda_p^0\right| . \left|\Omega_1\right| \mathbb{P}_{01} \to 0 \text{ as } n \to +\infty . \tag{2.43}$$

□

Lemma 2.4.8. *Using previous notations, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_2\right| \mathbb{P}_{02} \to 0 \tag{2.44}$$

Démonstration. We have :

$$\left|\Lambda_p^0\right| . \left|\Omega_2\right| \mathbb{P}_{02} \quad \leqslant \quad \tau K \left|\Lambda_p^0\right| \mathbb{P}_{02} \sim \frac{1}{2} p_n^2 \tau K \mathbb{P}_{02} = \mathcal{O}\left(\tau p_n^{2+\nu} \mathbb{P}_{02}\right)$$

where we use the lemma 2.4.6 for the equivalence above. In this proof, we almost have the same case than in the proof of lemma 2.4.7. In fact, the only difference is the matrix Σ_4 which is now

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & r & 0 \\ 0 & 1 & 0 & \varepsilon_n \\ r & 0 & 1 & 0 \\ 0 & \varepsilon_n & 0 & 1 \end{pmatrix} ,$$

where r is a coefficient from the matrix (r_{kj}). So, by the same method we have

$$p_n^{2+\nu} \mathbb{P}_{02} \to 0 \tag{2.45}$$

iff $\varepsilon_1 < 2b^2 - 2 - \nu$ and $\varepsilon_2 < 2c^2 - \nu$ where we still have $b = \frac{1+(1-\delta)^2}{2(1-\delta)^2}$ and $c = \frac{1-(1-\delta)^2}{6}$. Moreover we can show that $b^2 - 1 > c^2$. Then, if $\nu < 2c^2$ (fullfilled by assumptions in theorem 2.2.1), and from $\tau = o\left(p_n^t\right)$ for any $t > 0$, we have :

$$\left|\Lambda_p^0\right| . \left|\Omega_2\right| \mathbb{P}_{02} \to 0 . \tag{2.46}$$

□

Lemma 2.4.9. *Using notations previously introduced, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_3\right| \mathbb{P}_{03} \to 0 \tag{2.47}$$

Démonstration. This proof is exactly the same than for lemma 2.4.8 except that the matrix become

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & \varepsilon_n & 0 \\ 0 & 1 & 0 & r \\ \varepsilon_n & 0 & 1 & 0 \\ 0 & r & 0 & 1 \end{pmatrix} .$$

In particular, we obtain the same condition on ν. □

Lemma 2.4.10. *Using notations previously introduced, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_4\right| \mathbb{P}_{04} \to 0 \tag{2.48}$$

Démonstration. We have :

$$\left|\Lambda_p^0\right| . \left|\Omega_4\right| \mathbb{P}_{04} \leqslant \tau p_n \left|\Lambda_p^0\right| \mathbb{P}_{04} \sim \tau p_n^3 \mathbb{P}_{04} \tag{2.49}$$

Now, the correlation matrix is $\Sigma_4 = \begin{pmatrix} 1 & 0 & r & 0 \\ 0 & 1 & 0 & 0 \\ r & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$.
Thanks to the lemma 6.9 in [CJ11a], proved in the supplementary paper, we obtain $\mathbb{P}_{04} = \mathcal{O}\left(p_n^{-4+\varepsilon}\right)$ for any $\varepsilon > 0$. Then, we have $p_n^3 \tau \mathbb{P}_{04} = \mathcal{O}\left(\frac{\tau}{p_n^{1-\varepsilon}}\right)$ which tends to 0 as $n \to \infty$ since $\tau = o\left(p_n^t\right)$ for any $t > 0$. □

Lemma 2.4.11. *Using notations previously introduced, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_5\right| \mathbb{P}_{05} \to 0\,. \tag{2.50}$$

Démonstration. This proof is exactly the same than for lemma 2.4.10 considering the correlation matrix

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & r \\ 0 & 0 & 1 & 0 \\ 0 & r & 0 & 1 \end{pmatrix}$$

□

Lemma 2.4.12. *Using notations previously introduced, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_6\right| \mathbb{P}_{06} \to 0 \tag{2.51}$$

Démonstration. This proof is exactly the same than for lemma 2.4.8 except that the matrix become

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & \varepsilon_n & 0 \\ 0 & 1 & 0 & \varepsilon_n \\ \varepsilon_n & 0 & 1 & 0 \\ 0 & \varepsilon_n & 0 & 1 \end{pmatrix}$$

In particular, we have :

$$\left|\Lambda_p^0\right| . \left|\Omega_6\right| \mathbb{P}_{06} \leqslant K^2 \left|\Lambda_p^0\right| \mathbb{P}_{06} \sim \mathcal{O}\left(p_n^{2+2\nu} \mathbb{P}_{06}\right) \tag{2.52}$$

with Σ_4 as correlation matrix for the 4-uplet in $\mathbb{P}_{06}$. As for lemma 2.4.8, we have the following conditions

$$\varepsilon_1 < 2b^2 - 2 - 2\nu \text{ and } \varepsilon_2 < 2\left(c^2 - \nu\right) \tag{2.53}$$

which is summarized in $\nu < c^2$, and which is true considering theorem 2.2.1. Then, we obtain the desired result :

$$\left|\Lambda_p^0\right| . \left|\Omega_6\right| \mathbb{P}_{06} \to 0 \tag{2.54}$$

□

Lemma 2.4.13. *Using notations previously introduced, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_7\right| \mathbb{P}_{07} \to 0 \tag{2.55}$$

Démonstration. This proof is exactly the same than for lemma 2.4.10 considering the correlation matrix

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & \varepsilon_n & 0 \\ 0 & 1 & 0 & 0 \\ \varepsilon_n & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} .$$

□

Lemma 2.4.14. *Using notations previously introduced, we have :*

$$\left|\Lambda_p^0\right| . \left|\Omega_8\right| \mathbb{P}_{08} \to 0 \tag{2.56}$$

Démonstration. This proof is exactly the same than for lemma 2.4.10 considering the correlation matrix

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \varepsilon_n \\ 0 & 0 & 1 & 0 \\ 0 & \varepsilon_n & 0 & 1 \end{pmatrix} .$$

□

Remarks 3. *The main constraint here for Q_1 is $\nu < c^2$.*

Computation of Q_2 :

For this case, we will divide the computation into two parts. Indeed, we will consider two cases : when α is close to the set Λ_p^K and when it is not. For that purpose, we introduce the following sets :

$$I_{0,I} = \{(i,j) \in [\![1,p]\!], i < j \text{ and } \tau + K < j - i < \tau + 4K\} \text{ and } \Lambda_{p,I}^0 = \Lambda_p^0 \cap I_{0,I}$$

and

$$I_{0,II} = I_0 \backslash I_{0,I} \text{ and } \Lambda^0_{p,II} = \Lambda^0_p \cap I_{0,II}\,.$$

We can write :

$$Q_2 := \sum_{\alpha\in\Lambda^0_p}\sum_{\beta\in B^K_\alpha} \mathbb{P}_{\alpha\beta} = \sum_{\alpha\in\Lambda^0_{p,I}}\sum_{\beta\in B^K_\alpha} \mathbb{P}_{\alpha\beta} + \sum_{\alpha\in\Lambda^0_{p,II}}\sum_{\beta\in B^K_\alpha} \mathbb{P}_{\alpha\beta}\,.$$

Now, we look at the sum on $\Lambda^0_{p,I}$. We notice that on this set, the probability $\mathbb{P}_{\alpha\beta}$ is issued from a Gaussian vector with correlation matrix

$$\Sigma_4 = \begin{pmatrix} 1 & 0 & r_1 & r_2 \\ 0 & 1 & r_3 & r_4 \\ r_1 & r_3 & 1 & \varepsilon_n \\ r_2 & r_4 & \varepsilon_n & 1 \end{pmatrix}$$

where $|r_i| \leqslant 1-\delta$ for all $i \in \{1,2,3,4\}$. At least, coefficients $(r_i)_i$ can be replaced here by ε_n according to the position of the indice in both sets $\Lambda^0_{p,I}$ and B^K_α. But, we know that $\lim\limits_{n\to+\infty}(\varepsilon_n) = 0$ then for n large enough, we still have $|r_i| \leqslant 1-\delta$. Now, using Cauchy-Schwarz inequality, we write :

$$\mathbb{P}_{\alpha\beta} = \mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\mathbf{1}_{Z_\beta>a_n}\right] \leqslant \sqrt{\mathbb{E}\left[\mathbf{1}^2_{Z_\alpha>a_n}\right]\mathbb{E}\left[\mathbf{1}^2_{Z_\beta>a_n}\right]} \leqslant \sqrt{\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\right]\mathbb{E}\left[\mathbf{1}_{Z_\beta>a_n}\right]} = \sqrt{\mathbb{P}_\alpha\mathbb{P}_\beta}$$

Now, we use the fact that $\alpha \in \Lambda^0_{p,I} \subset \Lambda^0_p$ and $\beta \in B^K_\alpha \subset I_K$, then :

$$\mathbb{P}_{\alpha\beta} \leqslant \sqrt{\mathbb{P}_0\mathbb{P}_K}$$

which is true for any $|r_i| \leqslant 1$ then, $\sup\limits_{|r_i|\leqslant 1, i=1,\dots,4} \mathbb{P}_{\alpha\beta} \leqslant \sqrt{\mathbb{P}_0\mathbb{P}_K}$. At this point, remembering using their definition that $|\Lambda^0_{p,I}| \leqslant 3Kp$ and $|B^K_\alpha| \leqslant K^2$, and using $\mathbb{P}_0 = \mathcal{O}\left(p^{-2}\right)$, we have :

$$\sum_{\alpha\in\Lambda^0_{p,I}}\sum_{\beta\in B^K_\alpha} \mathbb{P}_{\alpha\beta} \leqslant 3K^3 p\sqrt{\mathbb{P}_0\mathbb{P}_K} \leqslant 3K^3\mathbb{P}_K^{1/2}\mathcal{O}(1) = \mathcal{O}\left(p^{3\nu}\mathbb{P}^{1/2}\right) \text{ as } n \to +\infty$$

.

Now, using lemma 2.4.3, we have :

$$p^{3\nu}\mathbb{P}^{1/2} \underset{n\to+\infty}{\longrightarrow} 0 \Leftrightarrow 3\nu < \frac{1}{2}\left(\frac{1}{2}\gamma^2 - 2|\gamma| + 2\right) \Leftrightarrow \nu < \frac{1}{6}c_\gamma \tag{2.57}$$

which is true according to assumptions of theorem 2.2.1.

Now, let us focuse on the computation of $\Lambda^0_{p,II}$. For that purpose, we introduce four subsets :

— $\Omega_1^2 := \{(u,v) \in B_\alpha^K : u - i < \tau \text{ and } j - v > \tau + K\}$ and $|\Omega_1^2| \leqslant K\tau$
— $\Omega_2^2 := \{(u,v) \in B_\alpha^K : \tau + K < u - i \text{ and } j - v < \tau\}$ and $|\Omega_2^2| \leqslant K\tau$
— $\Omega_3^2 := \{(u,v) \in B_\alpha^K : \tau \leqslant u - i \leqslant \tau + K \text{ and } j - v > \tau + K\}$ and $|\Omega_2^2| \leqslant K^2$
— $\Omega_4^2 := \{(u,v) \in B_\alpha^K : \tau + K < u - i \text{ and } \tau \leqslant j - v \leqslant \tau + K\}$ and $|\Omega_4^2| \leqslant K^2$

We have :

$$\sum_{\alpha\in\Lambda^0_{p,II}} \sum_{\beta\in B_\alpha^K} \mathbb{P}_{\alpha\beta} \leqslant 4\sum_{i=1}^{4} \sum_{\alpha\in\Lambda^0_{p,II}} \sum_{\beta\in\Omega_i^2} \mathbb{P}_{\alpha\beta}$$

.

In order to consider all these subset, we have the four next lemmas :

Lemma 2.4.15. *Considering the same notations as previously, if the probability* $\mathbb{P}_{\alpha\beta}$ *is issued from a Gaussian vector with covariance matrix* $\Sigma_4 = \begin{pmatrix} 1 & 0 & r_1 & x \\ 0 & 1 & 0 & 0 \\ r_1 & 0 & 1 & \varepsilon_n \\ x & 0 & \varepsilon_n & 1 \end{pmatrix}$ *where* $x \in \{\varepsilon_n, 0\}$*, then :*

$$\sum_{\alpha\in\Lambda^0_{p,II}} \sum_{\beta\in\Omega_1^2} \mathbb{P}_{\alpha\beta} \leqslant \mathcal{O}\left(p^{t+\nu+\epsilon}\mathbb{P}_K^{1/2}\right) \tag{2.58}$$

for any $t > 0$ *and any* $\epsilon > 0$*.*

Démonstration. In order to prove this result, we observe that in this case, for all $k \geqslant 1$, u_k^2 is independent with $\{u_k^1, u_k^3, u_k^4\}$. It means that conditionally at u_k^1, we have independence between Z_{12} and Z_{34}. In consequence, using Cauchy-Schwarz, we obtain :

$$\begin{aligned}
\mathbb{P}_{\alpha\beta} &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_{12}>a_n}\mathbf{1}_{Z_{34}>a_n}|u_k^1, k=1,\dots,n\right]\right] && (2.59)\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_{12}>a_n}|u_k^1, k=1,\dots,n\right]\mathbb{E}\left[\mathbf{1}_{Z_{34}>a_n}|u_k^1, k=1,\dots,n\right]\right] && (2.60)\\
&\leqslant \sqrt{\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_{12}>a_n}|u_k^1, k=1,\dots,n\right]^2\right]\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_{34}>a_n}|u_k^1, k=1,\dots,n\right]^2\right]} && (2.61)
\end{aligned}$$

Now, because u_k^1 is independent of u_k^2, we can use lemma 6.7 from [CJ11a] and we have :

$$\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_{12}>a_n}|u_k^1, k=1,\dots,n\right]^2\right] = \mathcal{O}\left(p^{-4+\epsilon}\right)$$

for any $\epsilon > 0$. And on the other side, we have :

$$\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_{34}>a_n}|u_k^1, k=1,\dots,n\right]^2\right] \leqslant \mathbb{P}_K$$

To finish, precising that $|\Lambda_{p,II}^0| \leqslant p^2$ and $|\Omega_1^2| \leqslant K\tau$, and writing $K = \mathcal{O}(p^\nu)$, we have the desired result :

$$\sum_{\alpha\in\Lambda_{p,II}^0}\sum_{\beta\in\Omega_1^2}\mathbb{P}_{\alpha\beta} \leqslant \mathcal{O}\left(p^{t+\nu+\epsilon}\mathbb{P}_K^{1/2}\right) \tag{2.62}$$

□

Now, from lemma 2.4.15 we have the condition :

$$t+\nu+\epsilon < \frac{1}{2}\left(\frac{1}{2}\gamma^2 - 2|\gamma| + 2\right),$$

which can be fullfilled from condition in (eq. (2.57)) and for a well-choosed $t > 0$ and $\epsilon > 0$. Finally we obtain, with our condition on ν that :

$$\lim_{n\to+\infty}\left[\sum_{\alpha\in\Lambda_{p,II}^0}\sum_{\beta\in\Omega_1^2}\mathbb{P}_{\alpha\beta}\right] = 0\,.$$

For the other subset Ω_i^2 for $i = 2, 3, 4$, we will use the same method. Indeed we notice that respectively for Ω_2^2, Ω_3^2 and Ω_4^2, the covariance matrices involved are respectively :

$$\Sigma_4^2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & x & r \\ 0 & x & 1 & \varepsilon_n \\ 0 & r & \varepsilon_n & 1 \end{pmatrix}, \Sigma_4^3 = \begin{pmatrix} 1 & 0 & \varepsilon_n & x \\ 0 & 1 & 0 & 0 \\ \varepsilon_n & 0 & 1 & \varepsilon_n \\ x & 0 & \varepsilon_n & 1 \end{pmatrix}, \Sigma_4^4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & x & \varepsilon_n \\ 0 & x & 1 & \varepsilon_n \\ 0 & \varepsilon_n & \varepsilon_n & 1 \end{pmatrix} \tag{2.63}$$

For each case, we use the fact that we always have u_k^2 (or u_k^1) which is independent from the other three random variables. Also, in order to use the lemma **??** 2.4.15, we notice that by construction $Z_{12} = Z_{21}$. Then, for cases Ω_3^2 and Ω_4^2, we just have to consider the Gaussian vector $(u_k^2, u_k^1, u_k^3, u_k^4)$ instead of $(u_k^1, u_k^2, u_k^3, u_k^4)$. In that way, all matrices have

the same form than in lemma 2.4.15 and similar upper-bound for the probability. Now, we use the upper-bound of subsets Ω_i^2. More precisely, for Ω_3^2 and Ω_4^2 :

$$\sum_{\alpha\in\Lambda_{p,II}^0}\sum_{\beta\in\Omega_3^2}\mathbb{P}_{\alpha\beta}\leqslant|\Lambda_{p,II}^0|.|\Omega_3^2|\mathbb{P}_K^{1/2}\mathcal{O}\left(p^{-2+\epsilon}\right)=\mathcal{O}\left(p^{2\nu+\epsilon}\mathbb{P}_K^{1/2}\right)\text{ for any }\epsilon>0$$

It means that we need to have, according to lemma 2.4.3 :

$$2\nu<\frac{1}{2}\left(\frac{1}{2}\gamma^2-2|\gamma|+2\right)\Leftrightarrow\nu<\frac{1}{4}\left(\frac{1}{2}\gamma^2-2|\gamma|+2\right). \tag{2.64}$$

With this condition

$$\lim_{n\to+\infty}\left[\sum_{\alpha\in\Lambda_{p,II}^0}\sum_{\beta\in\Omega_3^2}\mathbb{P}_{\alpha\beta}\right]=0$$

and

$$\lim_{n\to+\infty}\left[\sum_{\alpha\in\Lambda_{p,II}^0}\sum_{\beta\in\Omega_4^2}\mathbb{P}_{\alpha\beta}\right]=0$$

To finish, the case Ω_2^2 leads to the exactly same result than for Ω_1^2 because of the upper-bound of $|\Omega_2^2|$ which is the same than for $|\Omega_1^2|$. Then, we have :

$$\lim_{n\to+\infty}\left[\sum_{\alpha\in\Lambda_{p,II}^0}\sum_{\beta\in B_\alpha^K}\mathbb{P}_{\alpha\beta}\right]=0$$

and then

$$\lim_{n\to+\infty}[Q_2]=0$$

Remarks 4. *The main constaint here is* $\nu<\frac{1}{6}c_\gamma$.

Computation of Q_3 :

We focuse here on $Q_3=\sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^0}\mathbb{P}_{\alpha\beta}$. Once more, we will considere differents subsets for β according to its place into B_α^0. More precisely, let's define :

— $\Omega_1^3=\{(u,v)\in\Lambda_p^0 : i-u<\tau \text{ and } v-j<\tau\}$ and $|\Omega_1^3|\leqslant\tau^2$.
— $\Omega_2^3=\{(u,v)\in\Lambda_p^0 : i-u<\tau \text{ and } \tau\leqslant v-j\leqslant\tau+K\}$ and $|\Omega_2^3|\leqslant\tau K$.
— $\Omega_3^3=\{(u,v)\in\Lambda_p^0 : i-u<\tau \text{ and } \tau+K<v-j\}$ and $|\Omega_3^3|\leqslant\tau p$.
— $\Omega_4^3=\{(u,v)\in\Lambda_p^0 : \tau\leqslant i-u\leqslant\tau+K \text{ and } v-j<\tau\}$ and $|\Omega_4^3|\leqslant\tau K$.
— $\Omega_5^3=\{(u,v)\in\Lambda_p^0 : \tau+K<i-u \text{ and } v-j<\tau\}$ and $|\Omega_5^3|\leqslant\tau p$.

— $\Omega_6^3 = \{(u,v) \in \Lambda_p^0 : \tau \leqslant i-u \leqslant \tau+K \text{ and } \tau \leqslant v-j \leqslant \tau+K\}$ and $|\Omega_6^3| \leqslant K^2$.
— $\Omega_7^3 = \{(u,v) \in \Lambda_p^0 : \tau \leqslant i-u \leqslant \tau+K \text{ and } \tau+K < v-j\}$ and $|\Omega_7^3| \leqslant Kp$.
— $\Omega_8^3 = \{(u,v) \in \Lambda_p^0 : \tau+K < i-u \text{ and } \tau \leqslant v-j \leqslant \tau+K\}$ and $|\Omega_8^3| \leqslant Kp$.

Then, we have :

$$Q_3 \leqslant \sum_{i=1}^{8} \sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_i^3} \mathbb{P}_{\alpha\beta}$$

For Q_3, we use the computation of Q_2. Indeed, covariance matrices which are involved in the computation fo Q_3 are similar than for Q_2. The similarity comes from the fact that we exchange the role between α and β. More precisely, for Q_2 we had $\alpha \in \Lambda_p^0$ and $\beta \in B_\alpha^K$ and now we have $\alpha \in \Lambda_p^K$ and $\beta \in B_\alpha^0$. So, covariance matrices here will have the same structure than in Q_2 exchanging columns $\{1,2\}$ and columns $\{3,4\}$.

We notice that :

$$|\Omega_1^3|, |\Omega_2^3|, |\Omega_4^3| \leqslant |\Omega_6^3| \leqslant K^2 \tag{2.65}$$

Using the fact that $B_\alpha^0 \subset \Lambda_p^0$, we have :

$$\sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_6^3} \mathbb{P}_{\alpha\beta} \leqslant \sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_6^3} \mathbb{P}_0 \leqslant pK.K^2\mathbb{P}_0 = \mathcal{O}\left(p^{3\nu-1}\right) \text{ as } n \to +\infty$$

However, thanks to the condition in eq. (2.57), we have $\nu < \frac{1}{3}\left(\frac{1}{4}\gamma^2 - |\gamma| + 1\right)$. But, $\gamma \in]-2+\sqrt{2}, 2-\sqrt{2}[$. Then, $\frac{1}{3}\left(\frac{1}{4}\gamma^2 - |\gamma| + 1\right) \in]\frac{1}{6}, \frac{1}{3}[$. It leads, in particular, that $\nu < \frac{1}{3}$ and then :

$$\lim_{n \to +\infty} \left[\sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_6^3} \mathbb{P}_{\alpha\beta} \right] = 0$$

And so, thanks to equation eq. (2.65) :

$$\lim_{n \to +\infty} \left[\sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_1^3} \mathbb{P}_{\alpha\beta} \right] = \lim_{n \to +\infty} \left[\sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_2^3} \mathbb{P}_{\alpha\beta} \right] = \lim_{n \to +\infty} \left[\sum_{\alpha \in \Lambda_p^K} \sum_{\beta \in \Omega_4^3} \mathbb{P}_{\alpha\beta} \right] = 0 \tag{2.66}$$

Now, we look at the case when β belongs to $\Omega_3^3, \Omega_5^3, \Omega_7^3$ and Ω_8^3. We start by describing each covariance matrix involved. We note $x \in \{\varepsilon_n, 0\}$ and r a correlation coefficient such as $|r| \leqslant 1-\delta$. We have :

- for $\alpha \in \Lambda_p^K, \beta \in \Omega_3^3$, we have $\Sigma_4^3 = \begin{pmatrix} 1 & \varepsilon_n & r & 0 \\ \varepsilon_n & 1 & x & 0 \\ r & x & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

- for $\alpha \in \Lambda_p^K, \beta \in \Omega_5^3$, we have $\Sigma_4^3 = \begin{pmatrix} 1 & \varepsilon_n & 0 & x \\ \varepsilon_n & 1 & 0 & r \\ 0 & 0 & 1 & 0 \\ x & r & 0 & 1 \end{pmatrix}$
- for $\alpha \in \Lambda_p^K, \beta \in \Omega_7^3$, we have $\Sigma_4^3 = \begin{pmatrix} 1 & \varepsilon_n & \varepsilon_n & 0 \\ \varepsilon_n & 1 & x & 0 \\ \varepsilon_n & x & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$
- for $\alpha \in \Lambda_p^K, \beta \in \Omega_8^3$, we have $\Sigma_4^3 = \begin{pmatrix} 1 & \varepsilon_n & 0 & x \\ \varepsilon_n & 1 & 0 & \varepsilon_n \\ 0 & 0 & 1 & 0 \\ x & \varepsilon_n & 0 & 1 \end{pmatrix}$.

We observe that each time above one variable u_k^3 or u_k^4 is independent from the three other ones. Then we can use the same method than for Q_2 (conditioning with u_k^3 when u_k^4 is independent of the other ones or the contrary if it is u_k^3 which is independent).Then, by Cauch-Schwarz, we obtain the same upper-bound for $\mathbb{P}_{\alpha\beta}$. To show that we obtain the desired convergence, we study here the worst case. That is to say, using the fact that $|\Omega_i^3| \leqslant Kp$ for $i = 3, 5, 7, 8$, we can write :

$$\sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in\Omega_7^3}\mathbb{P}_{\alpha\beta} \leqslant |\Lambda_p^K|.|\Omega_7^3|\mathbb{P}_K^{1/2}\mathcal{O}\left(p^{-2+\epsilon}\right) \leqslant \mathcal{O}\left(p^{2\nu+\epsilon}\mathbb{P}_K\right) \text{ as } n \to +\infty$$

Then, we have exactly the same condition on ν that for equation eq. (2.64). It means that

$$\lim_{n\to+\infty}\left[\sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in\Omega_7^3}\mathbb{P}_{\alpha\beta}\right] = 0$$

and it induces that

$$\lim_{n\to+\infty}[Q_3] = 0$$

↪ **Computation of Q_4 :**

This last quantity is simplier because we can write :

$$Q_4 = \sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^K}\mathbb{P}_{\alpha\beta} \leqslant \sum_{\alpha\in\Lambda_p^K}\sum_{\beta\in B_\alpha^K}\mathbb{P}_K = \left|\Lambda_p^K\right|.\left|B_\alpha^K\right|\mathbb{P}_K \leqslant pK^3\mathbb{P}_k = \mathcal{O}\left(p^{1+3\nu}\mathbb{P}_K\right) \text{ as } n \to +\infty$$

So, to have $\lim_{n\to+\infty}(Q_4) = 0$, we must have, according the lemma lemma 2.4.3 :

$$1+3\nu<\frac{1}{2}\gamma^2-2|\gamma|+2 \quad \Leftrightarrow \quad 3\nu<c_\gamma-1$$

Under this assumption on ν we have

$$\lim_{n\to+\infty}[Q_4]=0$$

Finally, we have the desired result :

$$\lim_{n\to+\infty}[b_{2,n}]=0$$

.

Remarks 5. *The constraint on ν is $\nu<\frac{1}{3}(c_\gamma-1)$.*

Computation of $b_{3,n}$

We have :

$$\begin{aligned} b_{3,n} &= \sum_{\alpha\in\Lambda_p^0\cup\Lambda_p^K}\mathbb{E}\left[\left|\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}|\sigma\left(Z_\beta,\beta\in(\Lambda_p^0\cup\Lambda_p^K)\backslash B_\alpha\right)\right]-\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\right]\right|\right] \\ &= \sum_{\alpha\in\Lambda_p^0}\mathbb{E}\left[\left|\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}|\sigma\left(Z_\beta,\beta\in(\Lambda_p^0\cup\Lambda_p^K)\backslash B_\alpha\right)\right]-\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\right]\right|\right] \\ &+ \sum_{\alpha\in\Lambda_p^K}\mathbb{E}\left[\left|\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}|\sigma\left(Z_\beta,\beta\in(\Lambda_p^0\cup\Lambda_p^K)\backslash B_\alpha\right)\right]-\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\right]\right|\right] \end{aligned} \tag{2.67}$$

The first term of the RHS above is 0 from the choice of B_α. Hence

$$\begin{aligned} b_{3,n} &= \sum_{\alpha\in\Lambda_p^K}\mathbb{E}\left[\left|\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}|\sigma\left(Z_\beta,\beta\in(\Lambda_p^0\cup\Lambda_p^K)\backslash B_\alpha\right)\right]-\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\right]\right|\right] && (2.68)\\ &\leqslant \sum_{\alpha\in\Lambda_p^K}\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}|\sigma\left(Z_\beta,\beta\in(\Lambda_p^0\cup\Lambda_p^K)\backslash B_\alpha\right)\right]\right]+\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{Z_\alpha>a_n}\right]\right] && (2.69)\\ &\leqslant 2\left|\Lambda_p^K\right|\mathbb{P}_K && (2.70)\\ & && (2.71) \end{aligned}$$

According to the hypotheses in theorem 2.2.1, as we saw for the computation of λ_n, we have

$$\lim_{n\to+\infty}\left[\left|\Lambda_p^K\right|\mathbb{P}_K\right]=0$$

Finally, we obtain :

$$\lim_{n\to+\infty}\left[b_{3,n}\right]=0$$

Remarks 6. *Gathering remark 1 to 5 and noticing that on* $\gamma\in]-2+\sqrt{2},2-\sqrt{2}[$ *we have* $\frac{1}{3}(c_\gamma-1)<\frac{1}{6}c_\gamma$ *we get the final assumptions of theorem 2.2.1 for* ν*.*

□

2.5 Proofs of technical results

2.5.1 Proof of Proposition 2.4.1

We recall here the basic definition of tightness :

Definition 2.5.1. *Let* (u_n) *be a random sequence. We say that* (u_n) *is a tight sequence if :*

$$\forall\epsilon>0,\exists K>0,\ \sup_{n\geq1}\left(\mathbb{P}\left(|u_n|\geq K\right)\right)<\epsilon \tag{2.72}$$

Lemma 2.5.1. *Let* τ *be an integer. We define* $|||A|||=\max_{1\leq i<j\leq p,|i-j|>\tau}|A_{ij}|$ *for a* $(p\times p)$*-matrix. Let* $\mathbf{X}$ *be a random* (n,p)*-matrix where* $(X^1,X^2,\ldots,X^p)$ *are the* p *columns in* $\mathbb{R}^n$ *and* R_n *the empirical correlation matrix of* $\mathbf{X}$*. Let's define, for any* $k\in[\![1,p]\!]$ *:*

- $h_k=\frac{1}{\sqrt{n}}\parallel X^k-\overline{X^k}\mathbf{1}_n\parallel$ *with* $\overline{X^k}=\frac{1}{n}\sum_{i=1}^n X_i^k$
- $c_{n,1}=\max_{1\leqslant k\leqslant p}|h_k-1|$
- $c_{n,3}=\min_{1\leqslant k\leqslant p}h_k$
- $c_{n,4}=\max_{1\leqslant k\leqslant p}|\overline{X^k}|$
- $V_{n,\tau}=\max_{1\leqslant k<j\leqslant p,|k-j|\geqslant\tau}|{}^tX^kX^j|=\max_{1\leqslant k<j\leqslant p,|k-j|\geqslant\tau}|\sum_{i=1}^n X_i^kX_i^j|$

Then,

$$|||nR_n-{}^t\mathbf{X}\mathbf{X}|||\leqslant\frac{c_{n,1}^2+2c_{n,1}}{c_{n,3}^2}V_{n,\tau}+n\left(\frac{c_{n,4}}{c_{n,3}}\right)^2 \tag{2.73}$$

Démonstration. Let $\Delta_n := |nL_{n,\tau} - V_{n,\tau}|$ for any $n \geqslant 1$. We have :

$$\left|n^2 L_{n,\tau}^2 - V_{n,\tau}^2\right| = |nL_{n,\tau} - V_{n,\tau}| \,.\, |nL_{n,\tau} + V_{n,\tau}| \leqslant \Delta_n.(\Delta_n + 2V_{n,\tau}) \tag{2.74}$$

Here, we assume that the limiting distribution of $V_{n,\tau}$ is proved. Then, we have :

$$\frac{V_{n,\tau}}{\sqrt{n\log(p_n)}} \xrightarrow[n\to+\infty]{\mathbb{P}} 2 \tag{2.75}$$

Indeed, we obtained using the Chen-Stein method for $V'_{n,\tau}$:

$$\mathbb{P}\left(V'_{n,\tau} \leqslant a_n\right) = \exp\left(-\lambda_n\right) + o\left(1\right) \text{ as } n \to +\infty \tag{2.76}$$

Thanks to lemma 2.4.2, we have, for n large enough :

$$\mathbb{P}\left(V_{n,\tau} \leqslant a_n\right) = \exp\left(-\lambda_n\right) + o\left(1\right) \text{ as } n \to +\infty \tag{2.77}$$

This leads us to the asymptotic behaviour of $V_{n,\tau}$ which is :

$$\frac{1}{n}V_{n,\tau}^2 - 4\log\left(p_n\right) + \log\log\left(p_n\right) \xrightarrow[n\to+\infty]{\mathcal{L}} Z \tag{2.78}$$

where Z has the same cdf than in theorem 2.2.1. Then, we can write :

$$\frac{1}{n\log\left(p_n\right)}V_{n,\tau}^2 - 4 \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \tag{2.79}$$

And finally, we obtain eq. (2.75).

Now, we can notice :

$$\Delta_n \leqslant |||nR_n - {}^t\mathbf{X}\mathbf{X}||| \overset{(*)}{\leqslant} \frac{c_{n,1}^2 + 2c_{n,1}}{c_{n,3}^2}V_{n,\tau} + n\left(\frac{c_{n,4}}{c_{n,3}}\right)^2 \tag{2.80}$$

where $(*)$ refers to lemma 2.5.1. Using definition 2.5.1, we have that the sequence $\left(c_{n,1}\sqrt{\frac{n}{\log(p_n)}}\right)_{n\geqslant 1} := \left(c'_{n,1}\right)_{n\geqslant 1}$ and the sequence $\left(c_{n,4}\sqrt{\frac{n}{\log(p_n)}}\right)_{n\geqslant 1} := \left(c'_{n,4}\right)_{n\geqslant 1}$ are both tight sequence. In that way, both sequences $c_{1,n} = c'_{n,1}\sqrt{\frac{\log(p_n)}{n}}$ and $c_{4,n} = c'_{4,n}\sqrt{\frac{\log(p)}{n}}$ are tight too (from Hyp 1 in theorem 2.2.1, $\frac{\log(p_n)}{n} \to 0$ when $n \to +\infty$).

So,

$$\begin{aligned}
\Delta_n &\leqslant \frac{\frac{\log(p_n)}{n}c_{n,1}'^2 + 2\sqrt{\frac{\log(p_n)}{n}}c_{n,1}'}{c_{n,3}^2}V_{n,\tau} + n\frac{\log(p_n)}{n}\left(\frac{c_{n,4}'}{c_{n,3}}\right)^2 \\
\frac{\Delta_n}{\log(p_n)} &\leqslant \frac{\frac{1}{n}c_{n,1}'^2 + 2\sqrt{\frac{1}{n}}c_{n,1}'}{c_{n,3}^2}V_{n,\tau} + \left(\frac{c_{n,4}'}{c_{n,3}}\right)^2 \\
\frac{\Delta_n}{\log(p_n)} &\leqslant \frac{V_{n,\tau}}{\sqrt{n\log(p_n)}}\frac{c_{n,1}'^2\sqrt{\frac{\log(p_n)}{n}} + 2c_{n,1}'}{c_{n,3}^2} + \left(\frac{c_{n,4}'}{c_{n,3}}\right)^2
\end{aligned}$$

With this inequality, we notice that the sequence $(\Delta_n')_{n\geqslant 1} := \left(\frac{\Delta_n}{n}\right)_{n\geqslant 1}$ is tight.
So,

$$\begin{aligned}
\frac{\left|n^2L_{n,\tau}^2 - V_{n,\tau}^2\right|}{n} &\leqslant \frac{\Delta_n}{n}\left(\Delta_n + 2V_{n,\tau}\right) = \frac{\Delta_n'\log(p_n)}{n}\left(\Delta_n'\log(p_n) + 2V_{n,\tau}\right) \\
&\leqslant 2\sqrt{\frac{\left(\log(p_n)\right)^3}{n}}\left(\Delta_n'\sqrt{\frac{\log(p_n)}{n}} + V_{n,\tau}'\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \text{ because of 1 in theorem 2.2.1}
\end{aligned}$$

□

2.5.2 Proof of Lemma 2.4.2

Lemma 2.5.2. *We consider the following hypotheses :*

1. $\xi_1, \ldots, \xi_n$ *i.i.d random variables such that* $\mathbb{E}[\xi_1] = 0$ *and* $\mathbb{E}[\xi_1^2] = 1$.
2. $\exists t_0 > 0$, $\exists \alpha \in]0,1]$ *such that* $\mathbb{E}\left[e^{t_0|\xi_1|^\alpha}\right] < +\infty$.
3. $(p_n)_{n\in\mathbb{N}^*}$ *such that* $p_n \underset{n\to+\infty}{\longrightarrow} +\infty$ *and* $\log(p_n) = o_{n\to+\infty}\left(n^{\frac{\alpha}{2+\alpha}}\right)$
4. $(y_n)_{n\geqslant 1}$ *such that* $y_n \underset{n\to+\infty}{\longrightarrow} y > 0$

Then,

$$\mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\sum_{k=1}^{n}\xi_k \geqslant y_n\right) \underset{n\to+\infty}{\sim} \frac{1}{y\sqrt{2\pi}}p_n^{-\frac{1}{2}y_n^2}\sqrt{\log(p_n)}^{-1} \tag{2.81}$$

Démonstration. This lemma is the lemma 6.8 from [CJ11a] and it's proved in the supplement of this paper. □

Démonstration. Here, we will use the lemma 2.5.2. Indeed, we can check all the hypotheses of this lemma. First we write :

$$\mathbb{P}\left(\left|{}^tX^1X^{\tau+K+2}\right| > a_n\right) = \mathbb{P}(\left(\left|\sum_{i=1}^{n} X_i^1 X_i^{\tau+K+2}\right| > a_n\right) \tag{2.82}$$

Now, if we define $\xi_i = X_i^1 X_i^{\tau+K+2}$, we have :

1. $\mathbb{E}[\xi_i] \overset{\perp\!\!\!\perp}{=} \mathbb{E}[X_i^1]\mathbb{E}[X_i^{\tau+K+2}] = 0 \times 0 = 0$ where the independence come from the sample.
2. $\mathbb{E}[\xi_i^2] \overset{\perp\!\!\!\perp}{=} \mathbb{E}[(X_i^1)^2]\mathbb{E}[\left(X_i^{\tau+K+2}\right)^2] = 1 \times 1 = 1$
3. For $t_0 = \frac{1}{2}$ and $\alpha = 1$, we have :

$$\mathbb{E}[e^{t_0|\xi_1|^\alpha}] = \mathbb{E}[e^{\frac{|X_1^1 X_1^{\tau+K+2}|}{2}}] \leqslant \mathbb{E}[e^{\frac{1}{2}\left(X_1^1\right)^2}]\mathbb{E}[e^{\frac{1}{2}\left(X_1^{\tau+K+2}\right)^2}] < +\infty$$

4. We have $w_n := \frac{a_n}{\sqrt{n\log(p)}} \underset{n\to+\infty}{\longrightarrow} \sqrt{4} = 2 > 0$
5. According to the hypothesis 1 from theorem 2.2.1 : $\log(p_n) = o(n^{\frac{1}{3}})$ as $n \underset{n\to+\infty}{\longrightarrow} +\infty$

So we have all hypothesis to apply the lemma 2.5.2, and then :

$$\begin{aligned}
\mathbb{P}\left(\left|{}^tX^1X^{\tau+K+2}\right| > a_n\right) &= \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\left|{}^tX^1X^{\tau+K+2}\right| > \frac{a_n}{\sqrt{n\log(p_n)}}\right) \\
&= \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\left|\sum_{i=1}^{n} X_i^1 X_i^{\tau+K+2}\right| > \frac{a_n}{\sqrt{n\log(p_n)}}\right) \\
&= \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\sum_{i=1}^{n} X_i^1 X_i^{\tau+K+2} > \frac{a_n}{\sqrt{n\log(p_n)}}\right) \\
&\quad + \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\sum_{i=1}^{n} X_i^1 X_i^{\tau+K+2} < -\frac{a_n}{\sqrt{n\log(p_n)}}\right) \\
&= \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\sum_{i=1}^{n} X_i^1 X_i^{\tau+K+2} > w_n\right) \\
&\quad + \mathbb{P}\left(-\frac{1}{\sqrt{n\log(p_n)}}\sum_{i=1}^{n} X_i^1 X_i^{\tau+K+2} > w_n\right)
\end{aligned}$$

From lemma 2.5.2 we have

$$\begin{aligned}
\mathbb{P}\left(\left|{}^tX^1X^{\tau+K+2}\right| > a_n\right) &= \frac{1}{2\sqrt{2\pi}}p_n^{-\frac{1}{2}w_n^2}\frac{1}{\sqrt{\log(p_n)}}(1+o(1))+\frac{1}{2\sqrt{2\pi}}p_n^{-\frac{1}{2}w_n^2}\frac{1}{\sqrt{\log(p_n)}}(1+o(1)) \\
&= \frac{1}{\sqrt{2\pi}}p_n^{-\frac{1}{2}w_n^2}\frac{1}{\sqrt{\log(p_n)}}(1+o(1)) \\
&= \frac{1}{2\pi\log(p_n)}e^{-\frac{1}{2}\frac{a_n^2}{n\log(p_n)}-\frac{1}{2}\log\log(p_n)}(1+o(1)) \\
&= \frac{1}{\sqrt{2\pi}}e^{-2\log(p_n)+\frac{1}{2}\log\log(p_n)-\frac{1}{2}y-\frac{1}{2}\log\log(p_n)}(1+o(1)) \\
&= \frac{1}{p_n^2\sqrt{2\pi}}e^{-\frac{1}{2}y}(1+o(1)) \\
&\underset{n\to+\infty}{\sim} \frac{1}{p_n^2\sqrt{2\pi}}e^{-\frac{1}{2}y}
\end{aligned} \tag{2.83}$$

□

2.5.3 Proof of Lemma lemma 2.4.3

We remind that $\mathbb{P}_K := \mathbb{P}\left(|{}^tX^1X^{\tau+1}| > a_n\right)$. We define :

- $\mathbb{P}_K^+ := \mathbb{P}\left({}^tX^1X^{\tau+1} > a_n\right)$
- $\mathbb{P}_K^- := \mathbb{P}\left({}^tX^1X^{\tau+1} < -a_n\right)$
- $\xi_k := X_k^1X_k^{\tau+1}$
- $w_k := \dfrac{1}{\sqrt{1+\varepsilon_n^2\left(\xi_k-\varepsilon_n\right)}}$.

Notice that $(\xi_k)_{k\geqslant 1}$ are independent due to the independence between each line of $\mathbf{X}_n$. First we compute $\mathbb{E}\left[\xi_k\right] = \varepsilon_n$ and $\mathrm{var}\left(\xi_k\right) = 1+\varepsilon_n^2$. So, $\mathbb{E}\left[w_k\right] = 0$ and $\mathrm{var}\left(w_k\right) = 1$.We will apply the lemma 2.5.2 with w_k. Then,

$$\mathbb{P}_K^+ = \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\sum_{k=1}^{n}w_k > \underbrace{\frac{a_n-n\varepsilon_n}{\sqrt{(1+\varepsilon_n^2)n\log(p_n)}}}_{:=z_n}\right) \tag{2.84}$$

Thanks to hypotheses of theorem 2.2.1, we have $\lim\limits_{n\to+\infty}\left[z_n\right] := z = 2-\gamma > 0$. Then,

$$\begin{aligned}
\mathbb{P}_K^+ &\sim \frac{1}{z\sqrt{2\pi}} p_n^{-\frac{1}{2}z_n^2} \sqrt{\log(p_n)}^{-1} \\
&\sim \frac{1}{z\sqrt{2\pi}} \exp\left[-\frac{1}{2} z_n^2 \log(p_n) - \frac{1}{2}\log\log(p_n)\right] \\
&\sim \frac{1}{z\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(a_n - n\varepsilon_n)^2}{(1+\varepsilon_n^2)\, n\log(p_n)} \log(p_n) - \frac{1}{2}\log\log(p_n)\right] \\
&\sim \frac{1}{z\sqrt{2\pi}} \exp\left[\frac{-2}{1+\varepsilon_n^2}\log(p_n)\left(1 + \frac{\varepsilon_n^2}{4}\frac{\log\log(p_n)}{\log(p_n)} + \frac{y}{4}\frac{1}{\log(p_n)} + \frac{n\varepsilon_n^2}{4\log(p_n)} - \frac{\varepsilon_n}{2}\frac{a_n}{\log(p_n)}\right)\right]
\end{aligned}$$

With our hypotheses on ε_n, we have :

- $\dfrac{-2}{1+\varepsilon_n^2}\log(p_n) \underset{n\to+\infty}{\longrightarrow} -\infty$
- $\dfrac{\varepsilon_n^2}{4}\dfrac{\log\log(p_n)}{\log(p_n)} \underset{n\to+\infty}{\longrightarrow} 0$
- $\dfrac{y}{4}\dfrac{1}{\log(p_n)} \underset{n\to+\infty}{\longrightarrow} 0$
- $\dfrac{n\varepsilon_n^2}{4\log(p_n)} \underset{n\to+\infty}{\longrightarrow} \dfrac{1}{4}\gamma^2$ because $\varepsilon_n \sim \gamma\sqrt{\frac{\log(p_n)}{n}}$.
- $\dfrac{\varepsilon_n}{2}\dfrac{a_n}{\log(p_n)} \underset{n\to+\infty}{\longrightarrow} \gamma$ because $a_n \sim 2\sqrt{n\log(p_n)}$

$$p_n^a \mathbb{P}_K^+ \sim \frac{1}{z\sqrt{2\pi}} \exp\left[\left(a - 2 - \frac{1}{2}\gamma^2 + 2\gamma + o(1)\right)\log(p_n)\right] \tag{2.85}$$

Finally, for $\gamma \in]-2, 2[$:

$$\lim_{n\to+\infty}\left[p_n^a \mathbb{P}_K^+\right] = 0 \Leftrightarrow a < 2 + \frac{1}{2}\gamma^2 - 2\gamma \tag{2.86}$$

With the same ideas, we have :

$$\mathbb{P}_K^- = \mathbb{P}\left(\frac{1}{\sqrt{n\log(p_n)}}\sum_{k=1}^n w_k < -\frac{a_n + n\varepsilon_n}{\sqrt{(1+\varepsilon_n^2)n\log(p_n)}}\right) \tag{2.87}$$

$$= \mathbb{P}\left(\frac{-1}{\sqrt{n\log(p_n)}}\sum_{k=1}^n w_k > \underbrace{\frac{a_n + n\varepsilon_n}{\sqrt{(1+\varepsilon_n^2)n\log(p_n)}}}_{:=\tilde{z_n}}\right) \tag{2.88}$$

Thanks to lemma 2.5.2 and because $\lim_{n\to+\infty}[\tilde{z_n}] := \tilde{z} = 2+\gamma > 0$, we have :

$$\mathbb{P}_K^- \sim \frac{1}{\tilde{z}\sqrt{2\pi}} \exp\left[\frac{-2}{1+\varepsilon_n^2}\log(p_n)\left(1+\frac{\varepsilon_n^2}{4}\frac{\log\log(p_n)}{\log(p_n)}+\frac{y}{4}\frac{1}{\log(p_n)}+\frac{n\varepsilon_n^2}{4\log(p_n)}+\frac{\varepsilon_n}{2}\frac{a_n}{\log(p_n)}\right)\right] \tag{2.89}$$

$$p_n^b\mathbb{P}_K^- \sim \frac{1}{\tilde{z}\sqrt{2\pi}} \exp\left[\left(b-2-\frac{1}{2}\gamma^2-2\gamma+o(1)\right)\log(p_n)\right] \tag{2.90}$$

And finally, for $\gamma \in]-2,2[$:

$$\lim_{n\to+\infty}\left[p_n^b\mathbb{P}_K^-\right] = 0 \Leftrightarrow b < 2+\frac{1}{2}\gamma^2+2\gamma \tag{2.91}$$

To conclude, observing that

$$\min\left(2+\frac{1}{2}\gamma^2+2\gamma, 2+\frac{1}{2}\gamma^2-2\gamma\right) = \frac{1}{2}\gamma^2-2|\gamma|+2 := c_\gamma,$$

combining eq. (2.91) and eq. (2.86), we obtain, for all $d \in [0; c_\gamma[$ and as $n \to +\infty$:

$$\mathbb{P}_K := \mathbb{P}\left(\left|{}^tX^1X^{\tau+1}\right| > a_n\right) = o\left(p_n^{-d}\right) . \tag{2.92}$$

Chapitre 3

Cohérence pour un modèle de covariance par blocs

Sommaire

Dans le chapitre précédent, nous avons généralisé le modèle présenté dans [CJ11a] en ajoutant à la bande de covariance non nulle de largeur τ autour de la diagonale une bande intermédiaire asymptotiquement nulle, de taille $K >> \tau$. Pour certaines applications, par exemple en génétique, la matrice de covariance peut être décrite suivant un autre modèle : les covariances sont non nulles sur des blocs autour de la diagonale, par exemple dans [PDLLR19]. Motivés par ces applications, nous nous sommes intéressés à l'étude de la cohérence dans ce cas. Nous gardons la problématique de matrice d'observation $(n \times p)$ de grandes dimensions avec $p(n)$ tel que $p(n) \underset{n\to+\infty}{\longrightarrow} +\infty$ et $p(n) >> n$. Nous noterons $p = p(n)$ dans la suite pour plus de clarté.

3.1 Description du modèle

Considérons le p-vecteur aléatoire

$$\left(X^1, \ldots, X^p\right) \sim \mathcal{N}_p\left(\mu, \Sigma\right)$$

où la matrice de covariance $\Sigma = (\sigma_{ij}) \in \mathbb{R}^p \times \mathbb{R}^p$. On suppose que Σ est définie par blocs suivant le modèle ci-dessous. Soient :

- $N = N(n) \in \mathbb{N}$: le nombre de blocs.
- $(l_k)_{k \geqslant 0} \in (\mathbb{N}^*)^{\mathbb{N}}$ une suite de nombre entier strictement positifs où chaque l_k représente la dimension du k^{ieme} bloc. Pour simplifier, on pose $l_0 = 0$.
- Pour $k \in \{1, \ldots, N\}$, on définit l'ensemble :

$$S_k = \{(i,j) \in [\![1,p]\!]^2 : \sum_{u=0}^{k-1} l_u + 1 \leqslant i < j \leqslant \sum_{u=0}^{k} l_u\}$$

  Cet ensemble constituera le k^{ieme} bloc de corrélation dans la matrice de corrélation.
- On pose $I_r = \bigcup_{k=1}^{N} S_k$ (l'ensemble des blocs ayant des corrélations non nulles).
- On pose $I_0 = \{(i,j) \in [\![1,p]\!]^2 : 1 \leqslant i < j \leqslant p \text{ et } (i,j) \notin I_r\}$ (l'ensemble d'indices correspondant aux coefficients de corrélations théoriques nuls).

Plus précisément, la matrice de covariance Σ vérifie la structure :

$$\sigma_{ij} = \begin{cases} r_{ij}\sigma_i\sigma_j & \text{si} \quad (i,j) \in I_r \\ 0 & \text{si} \quad (i,j) \in I_0 \end{cases} . \tag{3.1}$$

Remarques 4. *Sans perte de généralité, on considère dorénavant que $\mu = 0_{\mathbb{R}^p}$ et que la matrice de covariance est réduite à la matrice de corrélation (autrement dit $\sigma_i = 1$ pour tout $i \in \{1, \ldots, p\}$).*

Constatons que nous ne considérons pas le même modèle de covariance que dans le chapitre précédent. On rappelle que précédemment, nous considérions une matrice avec une structure en bande. Cela se traduit de la manière suivante : si on considère deux composantes du p-vecteur gaussien, à savoir X^i et X^j, alors elles seront corrélées dès lors que la distance en terme d'indice (c'est-à-dire $|i-j|$) est faible et indépendante si cette distance est trop élevée. Cette fois-ci, pour notre modèle en bloc, on considère que les composantes du vecteur forment des paquets. Les composantes faisant parties d'un paquet sont corrélées entre-elles. Mais chaque variable est indépendante de celle appartenant à un autre groupe. Nous exposons dans la figure 3.1 la structure générale de la matrice de covariance/corrélation :

On peut d'ores et déjà constater certaines relations entre nos paramètres :

1. $\sum_{u=0}^{N} l_u = p$.
2. $\sum_{u=0}^{N} l_u^2 \leqslant p^2$: la surface occupée par les blocs est inférieure à la surface totale de Σ.

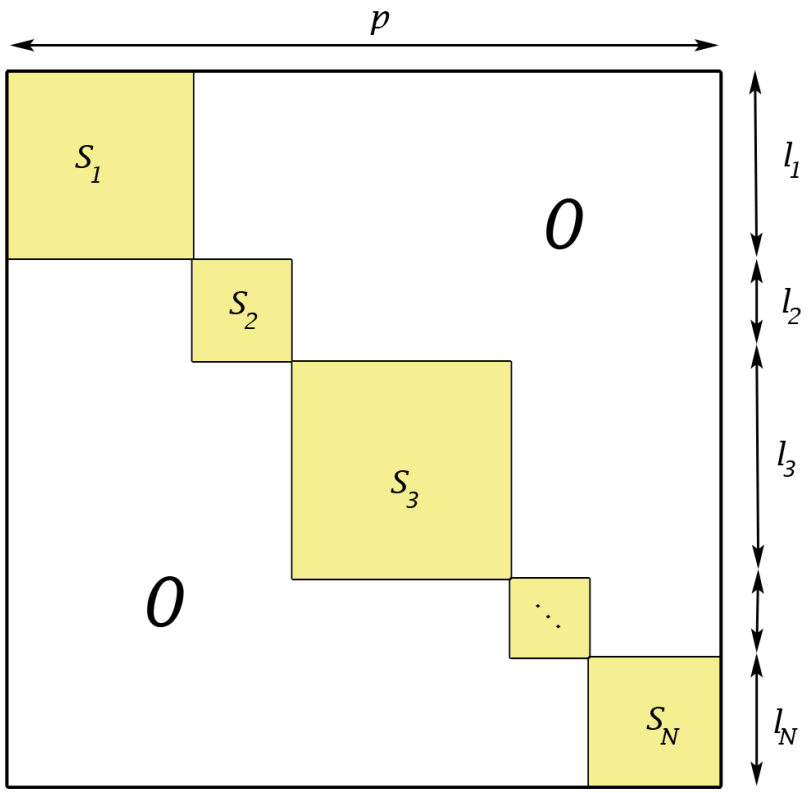

FIGURE 3.1 – Structure de la matrice Σ de covariance par blocs.

On suppose que l'on observe de façon indépendante et identiquement distribué un n-échantillon de ce p-vecteur gaussien. On résume cet échantillon dans la matrice d'observation $\mathbb{X}_n \in \mathbb{R}^{n\times p}$. On considère une nouvelle fois que $p >> n$ suit un régime que nous préciserons. On calcule la matrice de corrélation empirique $R_n = (\rho_{ij})_{1\leqslant i,j\leqslant p}$ associée à $\mathbb{X}_n$. On s'intéresse à la cohérence définie dans le chapitre 1 adaptée à notre modèle.

Definition 3.1.1. *On définit la cohérence pour le modèle de corrélation en bloc comme étant le maximum en valeurs absolues des coefficients de R_n en dehors des blocs. Autrement dit, avec les notations précédentes :*

$$L_{n,N} = \max_{(i,j)\in I_0} |\rho_{ij}|$$

Remarques 5. *Par analogie avec l'étude précédente, on appelle cette quantité la N-cohérence.*

Nous pouvons maintenant introduire notre théorème pour décrire le comportement de $L_{n,N}$ quand $n \to +\infty$.

3.2 Distribution asymptotique de la cohérence

Nous nous plaçons dans un contexte où les corrélations ne sont pas trop élevées. Pour cela, on définit le seuil $\delta \in]0,1[$ et l'ensemble :

$$\Gamma_{p,\delta} = \{k \in [\![1;p]\!] : |r_{kj}| > 1-\delta \text{ pour } j \in [\![1;p]\!] \text{ et } k \neq j\}.$$

Le résultat principal est le théorème suivant :

Théorème 3.2.1. *Soit n un entier naturel positif. Soit $p = p_n$ une suite telle que $p_n \to +\infty$ quand $n \to +\infty$. Supposons que :*

1. $\log(p_n) = o(n^{\frac{1}{3}})$ *quand* $n \to +\infty$.
2. $\exists \delta \in]0,1[$ *tel que* $|\Gamma_{p,\delta}| = o(p_n)$
3. $\max\limits_{1 \leqslant k \leqslant N} (l_k) = \mathcal{O}\left(p_n^{\nu}\right)$ *pour* $\nu < \frac{1}{36}\delta^2 (2-\delta)^2 < 1$

Dans ces conditions, on obtient :

$$nL_{n,N}^2 - 4\log(p_n) + \log(\log(p_n)) \xrightarrow[n\to+\infty]{\mathcal{L}} Z, \tag{3.2}$$

où Z possède comme fonction de distribution $F(y) = e^{-\frac{1}{\sqrt{8\pi}}e^{-\frac{y}{2}}}$ pour tout $y \in \mathbb{R}$.

Remarques 6. *Nous avons $\sum\limits_{u=0}^{N} l_u = p$ et donc $N \to +\infty$ quand $n \to +\infty$. De plus, pour n grand, on constate que la matrice de corrélation sera asymptotiquement constituée de nombreux petits blocs (en terme de largeur devant la dimension p).*

Corollaire 3.2.1. *La statistique $L_{n,N}$ vérifie, sous les conditions du théorème 3.2.1 :*

$$L_{n,N} \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \qquad \text{et} \qquad \sqrt{\frac{n}{\log(p)}} L_{n,N} \xrightarrow[n\to+\infty]{\mathbb{P}} 2 \tag{3.3}$$

Remarques 7. *Le corollaire 3.2.1 nous renseigne sur la vitesse de convergence de $L_{n,N}$ vers 0. Cette vitesse est de l'ordre de $\sqrt{\frac{\log(p)}{n}}$.*

3.3 Discussion

Pour ce modèle, nous constatons que la loi limite pour la N-cohérence normalisée reste la même que pour le modèle gaussien avec une structure de corrélations par bandes. Ceci est expliqué par l'hypothèse d'avoir des 0 en dehors des blocs de corrélations. On remarque qu'il existe certains cas pour lesquels le modèle par bloc coïncide avec le modèle par bandes. En effet, dans le cas où la largeur du plus grand bloc de corrélation vérifie :

$$\forall t > 0, \quad \max_{1 \leqslant k \leqslant N} (l_k) = o\left(p^t\right),$$

alors le modèle par blocs peut être considéré comme un cas particulier du modèle de T. Cai et al. Le paramètre τ de leur modèle étant alors $\max_{1 \leqslant k \leqslant N} (l_k)$.

Notre résultat permet d'aller plus loin. Là où le paramètre τ peut croître vers l'infini avec une vitesse logarithmique, nous supposons que le paramètre $\max_{1 \leqslant k \leqslant N} (l_k) = \mathcal{O}\left(p^\nu\right)$ avec $0 < \nu < \frac{1}{36}\delta^2 (2-\delta)^2$. La condition est donc plus générale : la plus grande largeur de bloc peut croître vers l'infini avec une puissance de p.

Ce modèle de corrélation par bloc est motivé par de possibles applications, en génétique par exemple. On constate la même structure en bloc dès que l'on s'intéresse aux gènes d'un individu. Citons par exemple les travaux [PDLLR19], ou encore les travaux [BSQ18] appliqués aux voitures autonomes. Afin d'être plus proche d'un modèle basé sur des données, la prochaine étape serait de considérer des coefficients non nuls, disposé en dehors des blocs.

3.4 Démonstration du théorème : application de la méthode de Chen-Stein

Pour prouver ce théorème, nous utilisons à nouveau la méthode de Chen-Stein. Comme dans le chapitre 3 : nous allons considérer une variable auxiliaire à laquelle nous appliquerons la méthode. Par la suite, nous montrerons que l'approximation par une loi de Poisson est pertinente en contrôlant les majorations et en calculant la fonction de distribution asymptotique.

3.4.1 Calculs préliminaires

Intéressons-nous aux cardinaux des ensembles précédemment introduits :

Lemme 3.4.1. *Nous avons, pour l'ensemble I_0 :*

$$|I_0| \leqslant \frac{1}{2}\left(1 - \frac{1}{N}\right) p^2$$

Démonstration. De façon évidente, on a :

1. Pour $k \in \{1, \ldots, N\}$, $|B_k| = \frac{1}{2} l_k (l_k - 1)$
2. $|I_r| = \frac{1}{2} \sum_{k=1}^{N} l_k (l_k - 1)$
3. $|I_0| = \frac{1}{2} \left(p^2 - \sum_{k=1}^{N} l_k^2 \right)$

Ainsi, grâce à l'inégalité de Cauchy-Schwarz, on obtient :

$$\sum_{k=1}^{N} l_k \leqslant \sqrt{\sum_{k=1}^{N} l_k^2} \sqrt{\sum_{k=1}^{N} 1} \Leftrightarrow \frac{1}{N} p^2 \leqslant \sum_{k=1}^{N} l_k^2 \leqslant p^2 \tag{3.4}$$

□

3.4.2 Utilisation d'une variable intermédiaire

Nous utilisons à nouveau une variable intermédiaire de la même forme que pour le modèle précédent. A savoir, nous introduisons :

$$V_{n,N} = \max_{(i,j) \in I_0} \left| \sum_{k=1}^{n} X_k^i X_k^j \right|$$

L'idée étant à nouveau de faire le lien entre $V_{n,N}$ et $L_{n,N}$. Pour ce faire, on constate que le résultat présenté dans le lemme 2.4.1 reste valable. Il est important de constater que ce lemme repose en partie sur la dimension p et de l'hypothèse $\log(p_n) = o\left(n^{1/3}\right)$ quand $n \to +\infty$. Considérant à nouveau ce régime pour p, on a alors une nouvelle fois :

$$\frac{n^2 L_{n,N}^2 - V_{n,N}^2}{n} \xrightarrow[n \to +\infty]{\mathbb{P}} 0 \tag{3.5}$$

Puisque nous sommes dans un cas où les corrélations ne sont pas trop élevées (les corrélations dépassant un seuil $1 - \delta \in]0, 1[$ sont en nombres limités), nous pouvons à nouveau réduire l'ensemble d'indices sur lequel nous calculons le maximum définissant $V_{n,N}$. En l'occurrence, nous introduisons les deux ensembles :

$$\Lambda_p^0 = \{(k, j) \in I_0 : \max_{1 \leqslant k \neq q, j \neq q \leqslant p} (|r_{kq}|, |r_{jq}|) \leqslant 1 - \delta\},$$

et

$$E_\delta = \{(i, j) \in I_0 : i \in \Gamma_{p,\delta} \text{ ou } j \in \Gamma_{p,\delta}\},$$

Nous pouvons quantifier leurs cardinaux dans le lemme suivant :

Lemme 3.4.2. *1.* $|E_\delta| = o\left(p^2\right)$

2. $\left|\Lambda_p^0\right| \sim |I_0|$ *quand* $n \to +\infty$.

Démonstration. 1. Grâce à la condition 2) du théorème 3.2.1, quand $n \to +\infty$:

$$|E_\delta| = 2p\left|\Gamma_{p,\delta}\right| = 2p.o\left(p\right) = o\left(p^2\right) \tag{3.6}$$

2. Dans un premier temps, nous pouvons écrire la partition suivante :

$$I_0 = \left(E_\delta \cap I_0\right) \cup \Lambda_p^0. \tag{3.7}$$

Ceci implique alors, pour les cardinaux, et en rappelant que $\Lambda_p^0 \subset I_0$:

$$|I_0| - |E_\delta| \leqslant \left|\Lambda_p^0\right| \leqslant |I_0|. \tag{3.8}$$

En utilisant le lemme 3.4.1, ainsi que l'hypothèse du théorème 3.2.1 contrôlant la plus grande largeur de bloc :

$$0 \leqslant \frac{1}{p^2}\sum_{k=1}^{N} l_k^2 \leqslant \frac{1}{p^2}\sum_{k=1}^{N}\left[\max_{1\leqslant u\leqslant N}\left(l_u\right) l_k\right] = \max_{1\leqslant u\leqslant N}\left(l_u\right)\frac{1}{p^2}\sum_{k=1}^{N} l_k = \frac{1}{p}\max_{1\leqslant u\leqslant N}\left(l_u\right) \tag{3.9}$$

Ainsi, avec l'hypothèse sur $\max\limits_{1\leqslant u\leqslant N}\left(l_u\right)$, on a :

$$\frac{1}{p}\max_{1\leqslant u\leqslant N}\left(l_u\right) = \mathcal{O}\left(p^{\nu-1}\right), \tag{3.10}$$

avec $\nu < 1$ d'où

$$\sum_{k=1}^{N} l_k^2 = o\left(p^2\right) \text{ quand } n \to +\infty,$$

Ce qui entraîne :

$$|I_0| = \frac{1}{2}\left(p^2 - \sum_{k=1}^{N} l_k^2\right) \sim \frac{1}{2}p^2$$

quand $n \to +\infty$. Finalement, pour l'équation 3.8, on obtient :

$$1 - \frac{|E_\delta|}{|I_0|} \leqslant \frac{\left|\Lambda_p^0\right|}{|I_0|} \leqslant 1. \tag{3.11}$$

Et puisque $|E_\delta| = o\left(p^2\right)$, on obtient bien le résultat souhaité.

□

A présent, on peut définir la variable aléatoire notée $V'_{n,N}$ par :

$$V'_{n,N} = \max_{(i,j)\in\Lambda_p^0} \left| \sum_{k=1}^{n} X_k^i X_k^j \right|.$$

Dans notre contexte, nous pouvons toujours étudier le comportement de $V'_{n,N}$ pour revenir ensuite à $V_{n,N}$ via le résultat suivant toujours vrai étant donné nos conditions :

Lemme 3.4.3. *Soit $(a_n)_{n\leqslant 1}$ une suite de réel définie par :*

$$a_n(y) = \sqrt{4n\log(p) - n\log\log(p) + ny}$$

On a, pour $n \to +\infty$:

$$\mathbb{P}\left(V'_{n,N} > a_n(y)\right) \leqslant \mathbb{P}\left(V_{n,N} > a_n(y)\right) \leqslant \mathbb{P}\left(V'_{n,N} > a_n(y)\right) + o(1) \tag{3.12}$$

Démonstration. Pour plus de simplicité, nous noterons $a_n(y) = a_n$ dans la suite. Dans un premier temps, nous savons que :

$$\mathbb{P}\left(V'_{n,N} > a_n\right) \leqslant \mathbb{P}\left(V_{n,N} > a_n\right).$$

Dans un second temps, nous pouvons écrire :

$$\begin{aligned}
\mathbb{P}\left(V_{n,N} > a_n\right) &\leqslant \mathbb{P}\left(V'_{n,N} > a_n\right) + \mathbb{P}\left(\max_{(i,j)\in(E_\delta\cap I_0)} \left|\sum_{k=1}^{n} X_k^i X_k^j\right| > a_n\right) \\
&\leqslant \mathbb{P}\left(V'_{n,N} > a_n\right) + \sum_{(i,j)\in(E_\delta\cap I_0)} \mathbb{P}\left(\left|\sum_{k=1}^{n} X_k^i X_k^j\right| > a_n\right).
\end{aligned}$$

Or, sur l'ensemble $(E_\delta \cap I_0)$, les couples (X^i, X^j) sont de même loi (gaussienne indépendante centrée réduite), et de façon analogue au chapitre précédent, on note :

$$\forall \alpha = (i,j) \in \Lambda_p^0, \quad \mathbb{P}_0 = \mathbb{P}\left(\left|\sum_{k=1}^{n} X_k^i X_k^j\right| > a_n\right).$$

On a vu que :

$$\mathbb{P}_0 = \mathcal{O}\left(p^{-2}\right)$$

Ainsi donc, le dernier terme vérifie :

$$\sum_{(i,j)\in(E_\delta\cap I_0)} \mathbb{P}\left(\left|\sum_{k=1}^{n} X_k^i X_k^j\right| > a_n\right) \leqslant |E_\delta|\,\mathbb{P}_0 = o(1).$$

On obtient alors le résultat souhaité. □

Nous pouvons donc à présent appliquer la méthode de Chen-Stein à la variable $V'_{n,N}$.

3.4.3 Méthode de Chen-Stein pour $V'_{n,N}$

Nous renvoyons à la présentation du résultat d'approximation par une loi de Poisson qui a été présenté dans 1.2.1. On obtient alors, pour $V'_{n,N}$:

$$\left|\mathbb{P}\left(V'_{n,N} \leqslant a_n\right) - \exp\left(-\lambda_n\right)\right| \leqslant b_{1,n} + b_{2,n} + b_{3,n}, \tag{3.13}$$

avec les notations suivantes :

- $Z_{ij} = \left|\sum_{k=1}^{n} X_k^i X_k^j\right|$
- $\lambda_n = \sum_{(i,j)\in\Lambda_p^0} \mathbb{P}\left(Z_{ij} > a_n\right)$
- $B_{ij} := \{(u,v) \in \Lambda_p^0 : |i-u| \leqslant \max_{1\leqslant k\leqslant N}(l_k) \text{ et } |j-v| \leqslant \max_{1\leqslant k\leqslant N}(l_k) \text{ et } (i,j)\neq(u,v)\}$
- $b_{1,n} = \sum_{(i,j)\in\Lambda_p^0} \sum_{(u,v)\in B_{uv}} \mathbb{P}\left(Z_{ij} > a_n\right)\mathbb{P}\left(Z_{uv} > a_n\right)$
- $b_{2,n} = \sum_{(i,j)\in\Lambda_p^0} \sum_{(u,v)\in B_{uv}} \mathbb{P}\left(Z_{ij} > a_n, Z_{uv} > a_n\right)$
- $b_{3,n} = \sum_{(i,j)\in\Lambda_p^0} \mathbb{E}\left[\left|\mathbb{P}\left(Z_{ij} > a_n | \sigma\left(Z_{uv}, (u,v) \in I_0\backslash B_{ij}\right)\right) - \mathbb{P}\left(Z_{ij} > a_n\right)\right|\right]$

Nous allons montrer que les termes majorants sont bien asymptotiquement nuls après avoir calculé la limite de λ_n ce qui nous donnera la convergence en loi de notre variable $V'_{n,N}$.

Lemme 3.4.4. *Suivant les notations précédemment introduites, pour $\alpha = (i,j) \in \Lambda_p^0$:*

$$|B_\alpha| \leqslant \left(2\max_{1\leqslant k\leqslant N}(l_k) + 1\right)^2 \leqslant \left(3\max_{1\leqslant k\leqslant N}(l_k)\right)^2 \tag{3.14}$$

Remarques 8. *Géométriquement, le voisinage B_α pour un $\alpha \in \Lambda_p^0$ forme un carré de longueur $2 \max_{1 \leqslant k \leqslant N} (l_k) + 1$. Avec ce voisinage, on prend en compte toutes les corrélations possibles entre deux couples de variables dans Λ_p^0.*

a) Calcul et limite de λ_n

En utilisant le fait que tous les couples de variables (X^i, X^j) sont de même loi quand $(i,j) \in I_0$, on obtient :

$$\lambda_n = \sum_{(i,j) \in \Lambda_p^0} \mathbb{P}\left(Z_{ij} > a_n\right) = \left|\Lambda_p^0\right| . \mathbb{P}_0 \tag{3.15}$$

L'équivalent du cardinal de Λ_p^0 nous donne, pour $n \to +\infty$:

$$\left|\Lambda_p^0\right| \mathbb{P}_0 \sim \frac{1}{2} p^2 \mathbb{P}_0 \tag{3.16}$$

Or, d'après le lemme 2.4.2, nous savons que, pour $y \in \mathbb{R}$ et quand $n \to +\infty$:

$$\mathbb{P}_0 \sim \frac{1}{p^2} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y\right)$$

D'où

$$\lim_{n \to +\infty} \lambda_n = \frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{2} y\right).$$

b) Contrôle de $b_{1,n}$

Pour le calcul de $b_{1,n}$, il suffit à nouveau d'utiliser l'égalité des probabilités sur Λ_p^0 (et également que $B_{ij} \subset \Lambda_p^0$). Ainsi, pour $n \to +\infty$:

$$
\begin{aligned}
b_{1,n} &= \sum_{(i,j)\in\Lambda_p^0} \sum_{(u,v)\in B_{ij}} \mathbb{P}\left(Z_{ij} > a_n\right) \mathbb{P}\left(Z_{uv} > a_n\right) = \sum_{(i,j)\in\Lambda_p^0} \sum_{(u,v)\in B_{ij}} \mathbb{P}_0^2 && (3.17)\\
&\leqslant \sum_{(i,j)\in\Lambda_p^0} |B_{ij}|\, \mathbb{P}_0^2 && (3.18)\\
&\leqslant \sum_{(i,j)\in\Lambda_p^0} \left(3 \max_{1\leqslant k\leqslant N}(l_k)\right)^2 \mathbb{P}_0^2 && (3.19)\\
&\leqslant \left|\Lambda_p^0\right| \left(3 \max_{1\leqslant k\leqslant N}(l_k)\right)^2 \mathbb{P}_0^2 && (3.20)
\end{aligned}
$$

En utilisant l'équivalent pour le cardinal de $\left|\Lambda_p^0\right| \sim \frac{1}{2}p^2$ on obtient :

$$
\left|\Lambda_p^0\right| \left(3 \max_{1\leqslant k\leqslant N}(l_k)\right)^2 \mathbb{P}_0^2 \sim \frac{1}{2}p^2 \left(3 \max_{1\leqslant k\leqslant N}(l_k)\right)^2 \mathbb{P}_0^2 = \frac{9}{2} \left(p \max_{1\leqslant k\leqslant N}(l_k)\mathbb{P}_0\right)^2 \qquad (3.21)
$$

A présent, puisque $\max_{1\leqslant k\leqslant N}(l_k) = \mathcal{O}\left(p^\nu\right)$, on a :

$$
p \max_{1\leqslant k\leqslant N}(l_k)\mathbb{P}_0 = \mathcal{O}\left(p^{1+\nu-2}\right) = \mathcal{O}\left(p^{\nu-1}\right)
$$

Et puisque $\nu < 1$, on a bien

$$
\lim_{n\to+\infty} b_{1,n} = 0
$$

c) Contrôle de $b_{2,n}$

Pour le calcul de cette quantité, notre méthode est semblable à celle utilisée pour le calcul du $b_{2,n}$ pour le modèle de corrélation par bande. On doit ici gérer les corrélations possibles entre deux couples de variables (X^i, X^j) et (X^u, X^v) pour un $(i,j) \in \Lambda_p^0$ et $(u,v) \in B_{ij}$. Nous savons, par l'appartenance des indices à Λ_p^0, que $X_k^i \perp\!\!\!\perp X_k^j$ et $X_k^u \perp\!\!\!\perp X_k^v$. En revanche, il est tout à fait possible que les couples d'indices (i,u) et/ou (j,v) tombent dans un bloc de corrélation B_k. Il est donc difficile d'identifier exactement des zones dans le voisinage où les probabilités sont les mêmes pour tout $(i,j) \in \Lambda_p^0$. En revanche, on peut identifier deux cas :

1. Le couple (i,j) est "proche" des blocs diagonaux.

2. Le couple (i,j) est "éloigné" des blocs diagonaux.

Pour différencier ces deux cas, nous allons découper l'ensemble Λ_p^0 en deux sous-ensembles :

1. $\Lambda_{p,I}^0 := \{(i,j) \in \Lambda_p^0 : |i-j| \leqslant 3 \max_{1 \leqslant k \leqslant N}(l_k)\}$ pour les indices "proches" des blocs diagonaux.
2. $\Lambda_{p,II}^0 := \{(i,j) \in \Lambda_p^0 : |i-j| > 3 \max_{1 \leqslant k \leqslant N}(l_k)\}$ pour les indices éloignés des blocs diagonaux.

L'idée est la suivante : lorsque nous sommes sur $\Lambda_{p,I}^0$, on peut difficilement identifier quelles variables sont corrélées entre elles. En revanche, l'ensemble lui-même est de cardinal raisonnable devant p^2. En revanche, sur $\Lambda_{p,II}^0$, les structures de covariances mises en jeux sont identifiables et en mesure de contrôler les cardinaux d'ensemble. De façon heuristique, dans un cas nous avons un ensemble très corrélé de cardinal faible et pour l'autre, nous avons de faibles corrélations pour un ensemble de cardinal plus grand.

Remarques 9. *Par construction :*

1. $\left|\Lambda_{p,I}^0\right| \leqslant 3 \max_{1 \leqslant k \leqslant N}(l_k) p$
2. $\left|\Lambda_{p,II}^0\right| \leqslant \frac{1}{2} p^2$

En effectuant ce découpage, nous pouvons écrire :

$$\begin{aligned} b_{2,n} &= \sum_{(i,j) \in \Lambda_p^0} \sum_{(u,v) \in B_{ij}} \mathbb{P}\left(Z_{ij} > a_n, Z_{uv} > a_n\right) \\ &= \sum_{(i,j) \in \Lambda_{p,I}^0} \sum_{(u,v) \in B_{ij}} \mathbb{P}\left(Z_{ij} > a_n, Z_{uv} > a_n\right) + \sum_{(i,j) \in \Lambda_{p,II}^0} \sum_{(u,v) \in B_{ij}} \mathbb{P}\left(Z_{ij} > a_n, Z_{uv} > a_n\right) \end{aligned}$$

Dans un premier temps, nous avons :

$$\begin{aligned}
\sum_{(i,j)\in\Lambda_{p,I}^0}\sum_{(u,v)\in B_{ij}}\mathbb{P}\left(Z_{ij}>a_n,Z_{uv}>a_n\right) &\leqslant \sum_{(i,j)\in\Lambda_{p,I}^0}\sum_{(u,v)\in B_{ij}}\mathbb{P}\left(Z_{ij}>a_n\right)\\
&= \sum_{(i,j)\in\Lambda_{p,I}^0}|B_{ij}|\,\mathbb{P}\left(Z_{ij}>a_n\right)\\
&\leqslant \sum_{(i,j)\in\Lambda_{p,I}^0}\left(3\max_{1\leqslant k\leqslant N}(l_k)\right)^2\mathbb{P}\left(Z_{ij}>a_n\right)\\
&\leqslant \left(3\max_{1\leqslant k\leqslant N}(l_k)\right)^2\mathbb{P}_0\left|\Lambda_{p,I}^0\right|\\
&\leqslant 27p\left(\max_{1\leqslant k\leqslant N}(l_k)\right)^3\mathbb{P}_0
\end{aligned}$$

D'après les hypothèses du théorème 3.2.1, ainsi que l'ordre de grandeur de $\mathbb{P}_0$, nous avons :

$$27p\left(\max_{1\leqslant k\leqslant N}(l_k)\right)^3\mathbb{P}_0=\mathcal{O}\left(p^{1+3\nu-2}\right) \tag{3.22}$$

Nous avons donc une limite nulle pour ce terme dès lors que :

$$3\nu-1<0\Leftrightarrow\nu<\frac{1}{3}$$

Dans un second temps, nous avons :

$$\sum_{(i,j)\in\Lambda_{p,II}^0}\sum_{(u,v)\in B_{ij}}\mathbb{P}\left(Z_{ij}>a_n,Z_{uv}>a_n\right)$$

Pour contrôler cette quantité, nous allons découper notre voisinage B_{ij} en quatre parties, définies de la façon suivante :

1. $\Omega_1:=\{(u,v)\in\Lambda_p^0 : i-\max\limits_{1\leqslant k\leqslant N}(l_k)\leqslant u\leqslant i \text{ et } j\leqslant v\leqslant j+\max\limits_{1\leqslant k\leqslant N}(l_k) \text{ et } (u,v)\neq(i,j)\}$
2. $\Omega_2:=\{(u,v)\in\Lambda_p^0 : i\leqslant u\leqslant i+\max\limits_{1\leqslant k\leqslant N}(l_k) \text{ et } j\leqslant v\leqslant j+\max\limits_{1\leqslant k\leqslant N}(l_k) \text{ et } (u,v)\neq(i,j)\}$
3. $\Omega_3:=\{(u,v)\in\Lambda_p^0 : i\leqslant u\leqslant i+\max\limits_{1\leqslant k\leqslant N}(l_k) \text{ et } j-\max\limits_{1\leqslant k\leqslant N}(l_k)\leqslant v\leqslant j \text{ et } (u,v)\neq(i,j)\}$
4. $\Omega_4:=\{(u,v)\in\Lambda_p^0 : i-\max\limits_{1\leqslant k\leqslant N}(l_k)\leqslant u\leqslant i \text{ et } j-\max\limits_{1\leqslant k\leqslant N}(l_k)\leqslant v\leqslant j \text{ et } (u,v)\neq(i,j)\}$

De façon heuristique, nous avons découpé notre voisinage "carré" en quatre : une partie nord-est (Ω_1), une partie sud-est (Ω_2), une partie sud-ouest (Ω_3) et une partie nord-ouest (Ω_4). Sur ces quatre ensembles, nous montrons que les lois de probabilités gaussiennes ont pour paramètres des matrices de corrélations de même structure. Ce découpage est présenté dans la figure 3.2

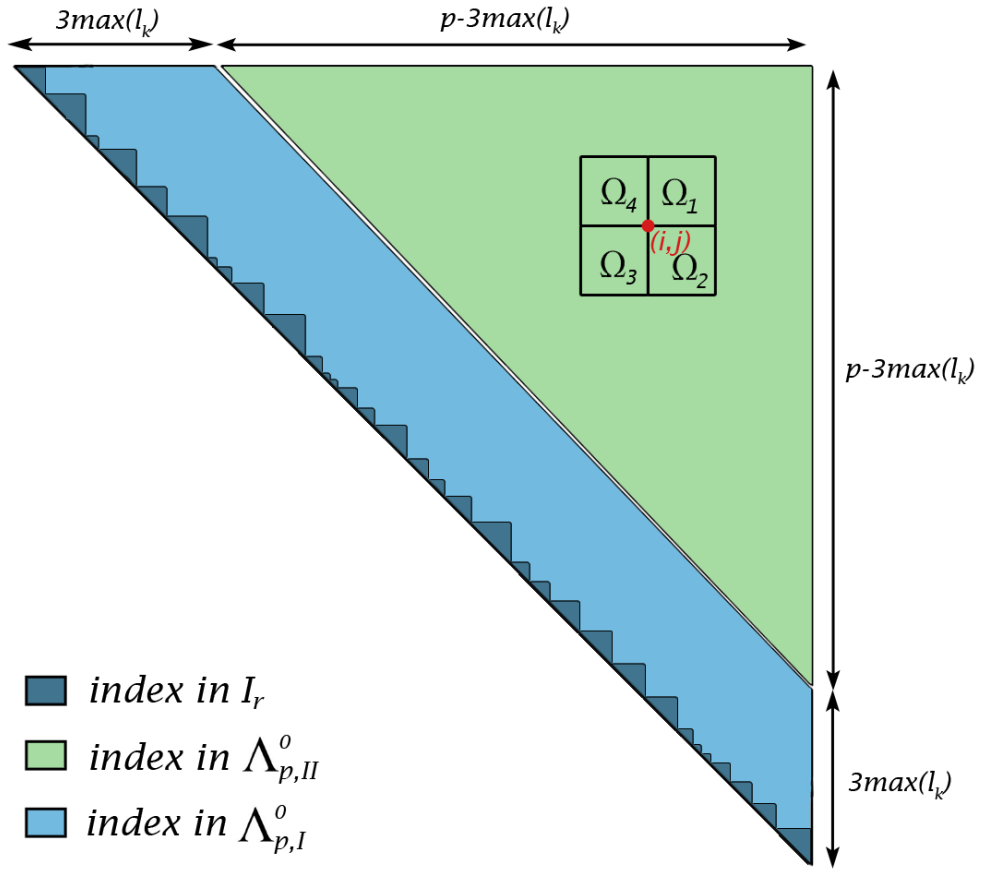

FIGURE 3.2 – Illustration du découpage du voisinage en quatre sous-parties.

Considérons le cas Ω_1 : Nous avons $(i,j) \in \Lambda^0_{p,II}$ et $(u,v) \in \Omega_1 \subset \Lambda^0_{p,II}$. Nous savons alors que le vecteur $\left(X^i_k, X^j_k, X^u_k, X^v_k\right) \overset{i.i.d}{\sim} \mathcal{N}\left(0, \Sigma^1_4\right)$ avec Σ^1_4 de la forme

$$\Sigma^1_4 = \begin{pmatrix} 1 & 0 & \times & \times \\ 0 & 1 & \times & \times \\ \times & \times & 1 & 0 \\ \times & \times & 0 & 1 \end{pmatrix}$$

Par construction de Ω_1, il est possible que les couples (i,u) et (j,v) tombent dans des

blocs de corrélations. Nous avons alors, dans le pire des cas des corrélations r_1 et r_2 (avec $|r_1|, |r_2| \leqslant 1-\delta$) respectivement pour (i,j) et (u,v). Ce qui donne :

$$\Sigma_4^1 = \begin{pmatrix} 1 & 0 & r_1 & \times \\ 0 & 1 & \times & r_2 \\ r_1 & \times & 1 & 0 \\ \times & r_2 & 0 & 1 \end{pmatrix}$$

Pour connaître le coefficient de corrélation entre X_k^i et X_k^v, comparons les indices : nous avons,

$$v - i = v - u + u - i$$

or, $u - i \in [\![-\max_{1\leqslant k\leqslant N}(l_k), 0]\!]$ et $v - u > 3\max_{1\leqslant k\leqslant N}(l_k)$ puisque $v-u > j-i$ par construction et $(i,j) \in \Lambda_{p,II}^0$. Donc $v - i > 2\max_{1\leqslant k\leqslant N}(l_k)$. Nous rappelons que le bloc de corrélation le plus important sur la diagonale est de largeur $\max_{1\leqslant k\leqslant N}(l_k)$. Il est donc impossible d'avoir (i,v) dans l'un d'eux. De la même manière, nous avons pour X_k^j et X_k^u :

$$j - u = j - i + \underbrace{i - u}_{>0} > j - i > 3\max_{1\leqslant k\leqslant N}(l_k)$$

Finalement, la matrice de corrélation pour Ω_1 est de la forme :

$$\Sigma_4^1 = \begin{pmatrix} 1 & 0 & r_1 & 0 \\ 0 & 1 & 0 & r_2 \\ r_1 & 0 & 1 & 0 \\ 0 & r_2 & 0 & 1 \end{pmatrix}$$

Nous utilisons exactement la même méthodologie pour montrer que la forme de la matrice de corrélation est la même pour Ω_2, Ω_3 et Ω_4. A présent, nous pouvons majorer la probabilité de couple. Pour cela, nous utilisons les travaux de Cai et Jiang, en particulier la démonstration du lemme 6.11 présent dans [CJ11b]. Nous en exposons les grandes lignes à présent.

Considérons un vecteur gaussien $(u^1, u^2, u^3, u^4) \sim \mathcal{N}(0, \Sigma_4^1)$. On introduit deux variables gaussiennes u^5 et u^6 centrées réduites et indépendantes. On a alors une égalité en loi :

$$\left(u^1, u^2, u^3, u^4\right) \overset{\mathcal{L}}{=} \left(u^1, u^2, r_1 u^1 + r_1' u^5, r_2 u^2 + r_2' u^6\right)$$

où $r_1' = \sqrt{1-r_1^2} > 0$ et $r_2' = \sqrt{1-r_2^2} > 0$.

On peut alors utiliser cette égalité pour étudier la probabilité en écrivant :

$$\mathbb{P}\left(|Z_{12}| > a_n, |Z_{34}| > a_n\right) = \mathbb{P}\left(|Z_{12}| > a_n, \left|\sum_{k=1}^{n}(r_1u^1 + r_1'u^5)(r_2u^2 + r_2'u^6)\right| > a_n\right) \tag{3.23}$$

De manière analogue à [CJ11b], on obtient :

$$\mathbb{P}\left(|Z_{12}| > a_n, |Z_{34}| > a_n\right) \leqslant \mathbb{P}\left(Z_{12} > ba_n\right) + 2\mathbb{P}\left(Z_{12} > a_n, Z_{16} > ca_n\right) + 2\mathbb{P}\left(Z_{12} > a_n\right)\mathbb{P}\left(Z_{56} > ca_n\right)$$

avec $b = \frac{1+(1-\delta)^2}{2(1-\delta)^2} > 1$ et $c = \frac{1-(1-\delta)^2}{6} > 0$.

A présent, puisque :

$$\forall \epsilon_1 > 0,\ \mathbb{P}\left(Z_{12} > ba_n\right) = \mathcal{O}\left(p^{-2b^2+\epsilon_1}\right) \tag{3.24}$$

$$\forall \epsilon_2 > 0,\ \mathbb{P}\left(Z_{12} > a_n, Z_{16} > ca_n\right) = \mathcal{O}\left(p^{-2-2c^2+\epsilon_2}\right) \tag{3.25}$$

$$\forall \epsilon_3 > 0,\ \mathbb{P}\left(Z_{12} > a_n\right)\mathbb{P}\left(Z_{56} > ca_n\right) = \mathbb{P}_0\mathcal{O}\left(p^{-2c^2+\epsilon_3}\right) = \mathcal{O}\left(p^{-2-2c^2+\epsilon_3}\right) \tag{3.26}$$

Donc pour contrôler,

$$\begin{aligned}\Lambda_{p,II}^0 \left|B_{ij}\right| \mathbb{P}\left(|Z_{12}| > a_n, |Z_{34}| > a_n\right) &\leqslant 9p^2\left(\max_{1\leqslant k\leqslant N}(l_k)\right)^2 \mathbb{P}\left(|Z_{12}| > a_n, |Z_{34}| > a_n\right) \\ &\leqslant \mathcal{O}\left(p^{2+2\nu}\right)\mathbb{P}\left(|Z_{12}| > a_n, |Z_{34}| > a_n\right)\end{aligned}$$

Au final,

$$\sum_{(i,j)\in\Lambda_{p,II}^0}\ \sum_{(u,v)\in B_{ij}} \mathbb{P}\left(Z_{ij} > a_n, Z_{uv} > a_n\right) \leqslant \mathcal{O}\left(p^{2+2\nu}\right)\mathbb{P}\left(|Z_{12}| > a_n, |Z_{34}| > a_n\right)$$

Ainsi, pour avoir une limite nulle au $b_{2,n}$, il nous faut :

$$2 + 2\nu - 2b^2 < 0 \Leftrightarrow \nu < b^2 - 1 = \left(\frac{1}{2} + \frac{1}{2(1-\delta^2)}\right)^2 - 1 \tag{3.27}$$

$$2 + 2\nu - 2 - 2c^2 < 0 \Leftrightarrow \nu < c^2 \tag{3.28}$$

Remarquons que $c^2 < b^2 - 1$, on a alors la condition pour ν :

$$\nu < \frac{(2-\delta)^2\delta^2}{36}$$

d) Contrôle de $b_{3,n}$

Par construction des voisinages, nous avons

$$b_{3,n} = 0$$

En effet si on sélectionne un couple $(u, v) \in \Lambda_p^0$ en dehors du voisinage B_α pour un $\alpha = (i, j) \in \Lambda_p^0$, alors les couples de variables aléatoires (X^i, X^j) et (X^u, X^v) seront indépendants. On a fait en sorte qu'il y ait au moins un écart de $\max(l_k)$ entre chaque indice i, j, u et v. Il n'y a donc aucune paire deux à deux qui tombent dans un bloc de corrélation. On a alors la non-corrélation et donc l'indépendance en vertu du cas gaussien considéré.

e) Distribution limite de $V'_{n,N}$ et de $V_{n,N}$

Nous avons montré précédemment que $b_{1,n}$, $b_{2,n}$ et $b_{3,n}$ sont nuls asymptotiquement. On a alors :

$$\lim_{n\to+\infty} \left[\mathbb{P}\left(V'_{n,N} \leqslant a_n\right)\right] = \exp\left(-\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{1}{2}y\right)\right)$$

A l'aide du lemme 3.4.3, on peut écrire :

$$\mathbb{P}\left(V'_{n,N} \leqslant a_n\right) \geqslant \mathbb{P}\left(V_{n,N} \leqslant a_n\right) \geqslant \mathbb{P}\left(V'_{n,N} \leqslant a_n\right) + o(1)$$

Ce qui donne, par encadrement :

$$\lim_{n\to+\infty}\left[\mathbb{P}\left(V_{n,N}\leqslant a_n\right)\right]=\exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{1}{2}y\right)\right)$$

Autrement dit, en explicitant a_n, nous avons :

$$\frac{1}{n}V_{n,N}^2-4\log\left(p\right)+\log\log\left(p\right)\xrightarrow[n\to+\infty]{\mathcal{L}}Z$$

où Z admet pour fonction de répartition la fonction $F(y)=\exp\left(-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{1}{2}y\right)\right)$

f) Retour à $L_{n,N}$

De la distribution asymptotique de $V_{n,N}$, on en déduit celle de $L_{n,N}$. Pour cela, on utilise l'équation 3.5 :

$$\frac{n^2L_{n,N}^2-V_{n,N}^2}{n}\xrightarrow[n\to+\infty]{\mathbb{P}}0 \tag{3.29}$$

Heuristiquement, la loi de $nL_{n,N}^2$ et de $\frac{1}{n}V_{n,N}^2$ est la même pour des n suffisamment grands. On en déduit alors la distribution en loi de $L_{n,N}$ qui est celle présentée dans le théorème 3.2.1 :

$$nL_{n,N}^2-4\log(p_n)+\log(\log(p_n))\xrightarrow[n\to+\infty]{\mathcal{L}}Z$$

3.4.4 Preuve du corollaire

Pour démontrer la convergence en probabilité de $L_{n,N}$, on utilise celle en loi. En effet, supposons que :

$$nL_{n,N}^2-4\log(p_n)+\log(\log(p_n))\xrightarrow[n\to+\infty]{\mathcal{L}}Z$$

Nous pouvons donc utiliser le lemme de Slutsky qui nous donne :

$$\left(nL_{n,N}^2-4\log(p_n)+\log(\log(p_n)),\frac{1}{n}\right)\xrightarrow[n\to+\infty]{\mathcal{L}}(Z,0)$$

En appliquant la fonction continue, de $\mathbb{R}^2$ dans $\mathbb{R}$, qui à (x, y) associe xy, on obtient :

$$L_{n,N}^2 - 4\frac{\log(p_n) + \log(\log(p_n)}{n} \xrightarrow[n\to+\infty]{\mathcal{L}} 0$$

Et donc

$$L_{n,N}^2 - 4\frac{\log(p_n) + \log(\log(p_n)}{n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0$$

A présent, puisque nous supposons que $\log(p) = o\left(n^{1/3}\right)$ pour $n \to +\infty$, on a bien :

$$L_{n,N}^2 \xrightarrow[n\to+\infty]{\mathbb{P}} 0 \text{ et } L_{n,N} \xrightarrow[n\to+\infty]{\mathbb{P}} 0$$

En procédant de la même façon, on obtient :

$$\frac{n}{\log(p)} L_{n,N}^2 \xrightarrow[n\to+\infty]{\mathbb{P}} 4$$

Chapitre 4

Simulation de la τ-cohérence en grande dimension

Sommaire

Dans le chapitre 2, nous avons décrit le comportement asymptotique de la τ-cohérence dans un cadre gaussien dépendant où la matrice d'observation était une très grande matrice de taille $(n \times p)$ avec $n << p$ et $\log(p) = o\left(n^{\frac{1}{3}}\right)$. Dans ce chapitre, on se propose de décrire une méthode de simulation pour pouvoir étudier numériquement le comportement asymptotique de cette statistique. Nous devons donc créer un processus de simulation qui puisse calculer une matrice de corrélation de taille $(p \times p)$ avec p qui prend des valeurs de plus en plus grandes. On se rend compte notamment que rapidement, la dimension p est suffisamment grande pour que la matrice de corrélation ne tienne plus dans la RAM d'un ordinateur classique. Il faut donc repenser les méthodes classiques de simulations. Pour cela, nous avons choisi d'orienter nos travaux sur l'utilisation du calcul GPGPU *General-Purpose computing on Graphics Processing Units,* pour gérer notre problème de dimension. Plus précisément, on développe deux stratégies de simulation : d'une part, on découpe nos matrices en blocs de tailles raisonnables, ceci afin de pouvoir étudier des matrices de grandes dimensions sans avoir à les stocker complètement. D'autre part, on

utilise le calcul en GPU afin de gagner en efficacité de temps de simulation, ceci afin de pouvoir faire des études statistiques pour des n et p grands.

Ce chapitre est organisé de la façon suivante : dans un premier temps, nous allons mettre en lumière les problèmes que posent les grandes dimensions. Ensuite, nous présenterons un schéma numérique et une méthode de calcul GPU pour générer un échantillon de τ-cohérence suivant les hypothèses du théorème du chapitre 2. Enfin, dans une dernière partie, nous donnerons les résultats obtenus par notre méthode et nous montrerons en quoi l'utilisation du GPU permet de gérer des problèmes en grandes dimensions.

4.1 Problématique

On se place dans le même contexte que celui présenté dans le chapitre 2. Nous le rappelons ici. On considère un vecteur gaussien de dimension p, noté

$$\left(X^1, \ldots, X^p\right) \sim \mathcal{N}_p\left(0, \Sigma\right)$$

Où $\Sigma = (\sigma_{ij})_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{p \times p}$ est la matrice de covariance/corrélation dont la structure est donnée par :

$$\sigma_{ij} = \begin{cases} r_{ij} & \text{si} \quad |i-j| < \tau \\ \epsilon_{ij} & \text{si} \quad \tau \leq |i-j| \leq \tau + K \\ 0 & \text{si} \quad \tau + K < |i-j| \end{cases} . \tag{4.1}$$

où les coefficients r_{ij} sont fixes dans $[-1, 1]$, et les coefficients (ϵ_{ij}) sont asymptotiquement nuls.

On tire un échantillon de taille n de ce p-vecteur gaussien, disposé dans une matrice d'observation, notée $\mathbb{X}_n$, de taille $n \times p$. On a donc une indépendance entre chaque ligne de $\mathbb{X}_n$ dûe à l'échantillonnage.

On s'intéresse à la τ-cohérence, définie par $L_{n,\tau} = \max_{|i-j| \geqslant \tau} |\rho_{ij}|$ où ρ_{ij} est le coefficient empirique de Pearson définie dans (2). On rappelle que pour notre théorème 2.2.1, on se place dans un cas où le nombre de colonnes p de la matrice d'observation est beaucoup plus important que la taille n de l'échantillon. On rappelle que notre résultat concerne l'asymptotique pour $n \to +\infty$ (et donc $p \to +\infty$ plus rapidement). Cela signifie d'un point de vue pratique que l'on considère un modèle où le nombre de variables observées

est plus grand que le nombre d'individus sur lesquels on les observe. Cette considération est pertinente par exemple en génétique lorsque l'on considère tous les gènes d'un groupe d'individus.

On met donc à jour le problème suivant : pour pouvoir calculer la τ-cohérence, nous avons besoin d'abord de calculer la matrice de corrélation empirique associée à $\mathbb{X}_n$. Cette matrice est de taille $p \times p$. On comprend que la dimension p de cette matrice devient rapidement trop grande pour pouvoir charger la matrice entièrement dans l'espace mémoire d'un ordinateur classique.

Plus précisement, on suppose dans le théorème 2.2.1 que la dimension p vérifie la condition $\log(p) = o\left(n^{1/3}\right)$ quand $n \to +\infty$. On peut s'intéresser alors à l'espace mémoire qu'occupera une matrice de taille $p \times p$ en fonction du régime de p. Dans la suite, nous utiliserons le régime $\log(p) = n^{\frac{1}{3.5}}$. Nous avons vu dans le résultat présenté dans 2.2.1 que la dimension p vérifie le régime $\log(p) = n^{1/3}$ pour n$\to +\infty$. Avec notre choix de p, nous nous approchons donc du régime critique.

Dans le graphique 4.1, on constate bien la croissance exponentielle du nombre p à gauche, ce qui entraîne une croissance très rapide également de l'espace mémoire nécessaire pour stocker une matrice de taille $p \times p$. On remarque également que pour des n de l'ordre de 4000 à 5000, la matrice d'observation de taille $n \times p$ nécessite une place en mémoire beaucoup plus faible. Cela signifie que pour des n raisonnables, on peut générer une matrice de taille $n \times p$ entièrement dans la mémoire. Il faut en revanche repenser la façon de calculer la matrice de corrélation sans avoir à la stocker complètement.

Au-delà du problème que pose la taille des matrices (et donc la gestion de l'espace mémoire), nous devons prendre en compte le temps de calcul pour simuler la matrice $\mathbb{X}_n$ et celui pour calculer la matrice de corrélation associée, avec la fonction `cor(.)` usuelle de R. Pour visualiser cet état de fait, on présente dans le tableau ci-dessous l'évolution du temps pour calculer la matrice de corrélation à partir d'une matrice d'observation $n \times p$ quelconque (ne suivant pas notre modèle) en faisant évoluer p. Pour exposer ces résultats, nous avons pris des valeurs de n successives entre 125 et 3000 avec un pas de 125. Nous avons choisi $p = \left[\exp\left(n^{1/3.5}\right)\right]$. Pour chaque valeurs de n (et donc de p), nous avons créé une matrice de taille $(n \times p)$ contenant uniquement des variables normales centrées réduites indépendantes. Puis nous avons calculé la matrice de corrélation de dimension $(p \times p)$. Nous avons récupéré le temps (en secondes) nécessaire pour la simulation de la matrice d'observation puis celui pour le calcul de la matrice de corrélation.

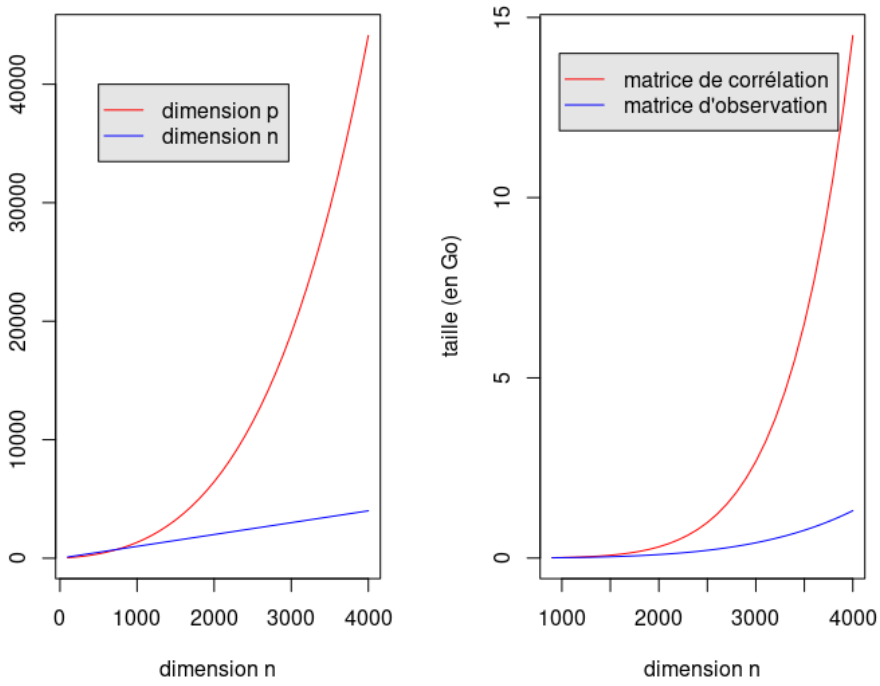

FIGURE 4.1 – Évolution de la dimension p en fonction de la taille n pour $p = [\exp(n^{1/3.5})]$ (à gauche) et évolution de la taille d'une $(n \times p)$-matrice et d'une $(p \times p)$-matrice (à droite)

n	500	750	1000	1250	1500	1750	2000	2250	2500	2750
p	366	756	1335	2144	3231	4650	6457	8719	11505	14894
Obs.(en s)	0.010	0.035	0.077	0.151	0.297	0.476	0.757	1.245	2.005	2.722
Corr.(en s)	0.081	0.296	1.113	3.600	9.814	23.767	53.683	110.345	215.760	394.620

On constate très nettement que le temps total (qui est la somme des deux) et majoritairement occupé par le temps de calcul de la matrice de corrélation. On constate que ce dernier devient rapidement trop important si l'on utilise simplement une fonction de calcul de corrélations d'un logiciel de statistiques (par exemple la fonction `cor(.)` du logiciel R). Afin d'obtenir des échantillons de τ-cohérence suffisamment grand (par réplications de Monte-Carlo), nous devons répéter un grand nombre de fois le calcul d'une matrice de corrélation de taille $p \times p$. On comprend rapidement que pour des n grands, le temps devient pénalisant. On doit donc également repenser notre méthode de simulation pour pouvoir gérer à la fois des tailles de matrices importantes mais aussi en un temps réduit.

La suite de ce chapitre est construite de la manière suivante : nous commençons par décrire le modèle de simulation que nous utilisons en montrant notamment pourquoi il

correspond bien à la structure de corrélation de notre modèle, puis nous expliciterons notre stratégie de découpage pour la matrice d'observation et celle de corrélation, comment nous en obtenons la τ-cohérence pour pouvoir effectuer des réplications de Monte-Carlo. Pour finir, nous exposerons nos résultats pour montrer dans un premier temps que les simulations tendent à confirmer notre théorème en pratique, puis nous exposerons également les gains en terme de temps de calcul apportés par l'utilisation de calculs en GPGPU.

4.2 Méthode de calcul de la τ-cohérence

4.2.1 Modèle de simulation

Pour créer une matrice d'observation $\mathbb{X}_n = \left(X_i^j\right)_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ de taille $n \times p$ vérifiant la structure de covariance 4.1, on considère le modèle, pour $1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant p$:

$$X_i^j = \sum_{k=j}^{j+2\tau+2K} C_{k-j+1} Y_i^k \tag{4.2}$$

où

$$\underline{C} = (C_1, \ldots, C_{1+2\tau+2K}) = \left(\underbrace{\varepsilon_n, \ldots, \varepsilon_n}_{K \text{ fois}}, \underbrace{r_1, r_2, \ldots, r_{1+2\tau}}_{1+2\tau \text{ termes}}, \underbrace{\varepsilon_n, \ldots, \varepsilon_n}_{K \text{ fois}} \right)$$

et avec

$$\left(Y_i^k, i = 1, \ldots, n, k = 1, \ldots, p + 2\tau + 2K\right) \overset{i.i.d}{\sim} \mathcal{N}_1(0, 1)$$

La quantité X_i^j est la composante de la i^{eme} ligne et de la j^{eme} colonne de $\mathbb{X}_n$. Dans nos simulations, nous prendrons

$$(r_k)_{1\leqslant k\leqslant 1+2\tau} \overset{i.i.d}{\sim} \mathcal{U}_{[-1,1]}$$

On constate donc que pour créer la matrice $\mathbb{X}_n$, nous allons utiliser une matrice, notée $\mathbb{Y}_n$ de taille $n \times (p+2\tau+2K)$ constituée de variables aléatoires gaussiennes normales centrées indépendantes. On peut voir le modèle de simulation comme une moyenne mobile pondérée de largeur $1 + 2\tau + 2K$.

On commence par remarquer qu'on a bien indépendance entre deux lignes de la matrice $\mathbb{X}_n$ grâce à l'indépendance entre les lignes de $\mathbb{Y}_n$. Regardons les propriétés de notre modèle :

Lemme 4.2.1. *Pour le modèle de simulation (4.2), on a :*

1. $\mathbb{E}\left[X_i^j\right] = 0$
2. $var\left(X_i^j\right) = 2K\varepsilon_n^2 + \sum_{k=1}^{1+2\tau+2K} r_k^2$

Démonstration. Le fait que les variables $\left(X_i^j\right)$ soient centrées repose uniquement sur le fait que les $\left(Y_i^j\right)$ sont centrées. De plus, par indépendance entre les $\left(Y_i^j\right)$, on a :

$$var(X_i^j) = \sum_{k=j}^{j+2\tau+2K} C_{k-j+1}^2 = \sum_{k=1}^{1+2\tau+2K} C_k^2 = \sum_{k=1}^{K} C_k^2 + \sum_{k=K+1}^{K+2\tau+1} C_k^2 + \sum_{k=2+2\tau+K}^{1+2\tau+2K} C_k^2 = \varepsilon^2 K + \sum_{k=1}^{1+2\tau} r_k^2 + \varepsilon^2 K$$

□

Maintenant, nous vérifions la structure de covariance de notre modèle :

Lemme 4.2.2 (Modèle de covariance)**.** .
Pour $1 \leqslant u < v \leqslant p$*, on a :*

$$cov\left(X_i^u, X_i^v\right) = \sum_{m=1}^{1+2\tau+2K-(v-u)} C_m C_{m+v-u} \mathbb{1}_{v-u \leqslant 2\tau+2K} \tag{4.3}$$

De plus, selon les valeurs de $v-u$*, on peut préciser :*
$cov\left(X_i^u, X_i^v\right) =$

$$\begin{cases} 2\varepsilon^2\left(K-(v-u)\right) + \varepsilon_n\left[\sum_{k=1}^{v-u} r_k + \sum_{k=2+2\tau-(v-u)}^{1+2\tau} r_k\right] + \sum_{k=1}^{1+2\tau-(v-u)} r_k r_{k+v-u} & si \quad 1 \leqslant v-u \leqslant K-1 \\ & et \quad 1 \leqslant v-u \leqslant 2\tau \\ 2\varepsilon^2\left(K-(v-u)\right) + 2\varepsilon_n \sum_{k=1}^{1+2\tau} r_k & si \quad 1 \leqslant v-u \leqslant K-1 \\ & et \quad v-u = 1+2\tau \\ 2\varepsilon^2\left(K-(v-u)\right) + 2\varepsilon_n \sum_{k=1}^{1+2\tau} r_k + \varepsilon^2\left(v-u-(1+2\tau)\right) & si \quad 1 \leqslant v-u \leqslant K-1 \\ & et \quad v-u \geqslant 2+2\tau \\ \sum_{k=1}^{1+2\tau-(v-u)} r_k r_{k+v-u} + \varepsilon \sum_{k=1+2\tau-(v-u)+1}^{1+2\tau} r_k & si \quad K \leqslant v-u \leqslant K+2\tau \\ & et \quad v-u \leqslant 2\tau \\ \varepsilon \sum_{k=1}^{1+2\tau} r_k & si \quad K \leqslant v-u \leqslant K+2\tau \\ & et \quad v-u \geqslant 2\tau+1 \\ \varepsilon^2\left(1+2\tau+2K-(v-u)\right) & si \quad K+2\tau+1 \leqslant v-u \leqslant 2\tau+2K \\ 0 & si \quad v-u \geqslant 1+2\tau+2K \end{cases} \tag{4.4}$$

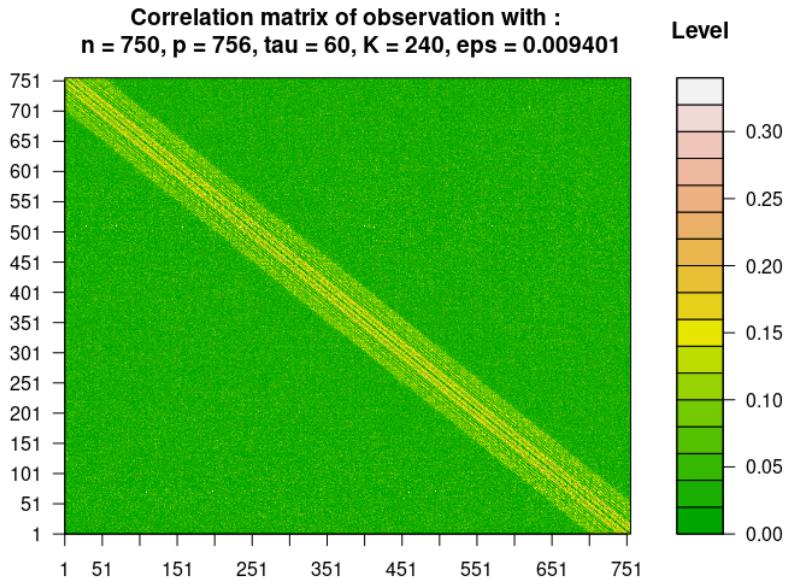

FIGURE 4.2 – Niveaux de corrélations pour une matrice d'observation suivant notre modèle.

On constate que nous obtenons une bande centrale de corrélations non nulles de largeur 2τ. En effet, dès lors que $v-u \leqslant 2\tau$, la covariance est alors constituée de quantités du type $\sum_k r_k r_{k+v-u}$ qui ne sont pas asymptotiquement nulles. Dans les autres cas, pour $v-u \geqslant 2\tau+1$, on constate que nous avons des coefficients multipliés par ε ou ε^2 qui correspondent à la bande intermédiaire. Constatons que cette bande est alors de longueur $2K$. Nous devons donc prendre en compte la structure des bandes dans notre modèle de simulation, notamment pour bien récupérer les bons coefficients qui interviennent dans le calcul de la τ-cohérence.

De plus, dans la présentation de notre modèle, nous avons beaucoup de possibilités en comparant $v-u$ avec les différents paramètres τ et K. Toutes ces conditions apparaissent notamment du fait que notre modèle nous indique que $\tau < K$ mais nous n'avons pas la comparaison, pour des n faibles, entre K et 2τ. Nous savons en revanche, que pour des n suffisamment grands, $K > 2\tau$.

Nous avons simulé une matrice d'observation pour notre régime de p suivant notre modèle et nous avons calculé la matrice de corrélation empirique associée à cette matrice. Nous exposons dans la figure 4.2 les niveaux de corrélations. On constate bien que les corrélations suivent la structure par bande désirée.

A présent, nous allons exposer une première stratégie de découpage sur les matrices d'observations.

4.2.2 Découpage de la matrice d'observation

Ce choix à été motivé par la raison suivante : pour des n raisonnables (dans le sens où on observe une convergence asymptotique), la matrice d'observation de taille $n \times p$ peut toujours être contenue dans un ordinateur classique. De plus, nous allons voir dans la suite que nous exploitons le produit matriciel. En effet, dans des logiciels tels que R, les codes où les calculs sont vectorisés sont plus efficaces en terme de temps. On se voit donc obligé de créer des matrices beaucoup plus grandes que $\mathbb{X}_n$. En particulier, on introduit la matrice $\mathbb{A}$ de taille $(p+2\tau+2K) \times p$ et, comme nous l'avons déjà introduit, la matrice $\mathbb{Y}_n$ de taille $n \times (p+2\tau+2K)$. La matrice $\mathbb{A}$ contient les coefficients de pondérations et est définie par :

$$\mathbb{A} = \begin{pmatrix}
C_1 & 0 & 0 & 0 & \dots & \dots & 0 \\
C_2 & C_1 & 0 & 0 & \dots & \dots & \vdots \\
C_3 & C_2 & C_1 & 0 & \dots & \dots & \vdots \\
\vdots & C_3 & C_2 & C_1 & \ddots & \dots & \vdots \\
\vdots & \vdots & C_3 & C_2 & \ddots & \dots & 0 \\
\vdots & \vdots & \vdots & C_3 & \ddots & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \dots & C_1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \dots & C_2 \\
C_{2\tau+2K} & \vdots & \vdots & \vdots & \vdots & \dots & C_3 \\
C_{1+2\tau+2K} & C_{2\tau+2K} & \vdots & \vdots & \vdots & \dots & \vdots \\
0 & C_{1+2\tau+2K} & C_{2\tau+2K} & \vdots & \vdots & \dots & \vdots \\
0 & 0 & C_{1+2\tau+2K} & C_{2\tau+2K} & \vdots & \dots & \vdots \\
\vdots & \vdots & 0 & C_{1+2\tau+2K} & \ddots & \dots & \vdots \\
\vdots & \vdots & \vdots & 0 & \ddots & \dots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \dots & C_{2\tau+2K} \\
0 & 0 & 0 & 0 & 0 & \dots & C_{1+2\tau+2K}
\end{pmatrix}$$

La matrice $\mathbb{Y}_n$ est la matrice contenant les variables gaussiennes i.i.d. centrées réduites : $\mathbb{Y}_n = \left(Y_i^j\right)_{ij}$ où les Y_i^j sont des variables aléatoires indépendantes telles que $Y_i^j \sim \mathcal{N}(0,1)$. Nous avons alors le résultat :

$$\mathbb{X}_n = \mathbb{Y}_n \mathbb{A} \tag{4.5}$$

L'idée est de découper la matrice $\mathbb{X}_n$ en paquets de colonnes, tous de même taille. Nous noterons à partir de maintenant T_b et N_b, deux nombres entiers positifs représentant respectivement la taille des paquets de colonnes et le nombre de paquets de colonnes.

Remarques 10. *On sait que p est un entier (nombre de colonnes). En revanche, il n'y a aucune raison que pour le régime choisi pour celui-ci, entraîne que $p = N_b \times T_b$. Si ce n'est pas le cas, nous devrions prendre en compte le reste de la division euclidienne de p par N_b et regarder comment traiter les colonnes restantes. Nous avons choisi ici d'effectuer un remaniement des variables afin que le nombre p soit bien un multiple de N_b dans le régime choisit par l'utilisateur. Pour ce faire, on choisit p selon un régime de croissance par rapport à n (nous choisirons dans les résultats $\log(p) = n^{\frac{1}{3.5}}$). Ensuite, nous appliquons un algorithme rapide 4.1. Ce petit programme, nous permet d'avoir des tailles de blocs comme un multiple de τ ce qui est plus pratique pour le calcul de la τ-cohérence. Pour cela, on utilise simplement le reste de la division euclidienne de T_b par τ.*

Listing 4.1 – Code permettant d'ajuster les paramètres.

```
ajust=function(p,tau,Nb){
        Tb=p%/%Nb
        if(Tb%%tau!=0){
                Tb = Tb + tau - Tb%%tau;
        }
        p = Nb*Tb
        return(list(p=p,tau=tau,Nb=Nb,Tb=Tb))
}
```

Puisque le nombre de colonne de $\mathbb{A}$ est celui de $\mathbb{X}_n$, on découpe la matrice $\mathbb{A}$ en N_b paquets de T_b colonnes. Notons $(\mathbb{A}^1, \ldots, \mathbb{A}^{N_b})$ où $\mathbb{A}^j$ est le j^{ieme} paquet et $\mathbb{A}^j$ est de taille $(p + 2\tau + 2K) \times T_b$. De la même façon, on note $\mathbb{X}_n = \left(\mathbb{X}^1, \ldots, \mathbb{X}^{N_b}\right)$ où $\mathbb{X}^j$ est le j^{ieme} paquet de colonne de $\mathbb{X}_n$. Il est donc de dimension $n \times T_b$. On a alors :

$$\text{Pour tout } j \in \{1, \ldots, N_b\}\ ,\ \mathbb{X}^j = \mathbb{Y} * \mathbb{A}^j \tag{4.6}$$

La création de la matrice d'observation se fait donc en N_b étapes. Afin de gagner un maximum de place en terme de mémoire, nous stockons chacun de ces paquets dans un fichier binaire. Cela nous oblige à tenir compte de cette structure au moment où nous récupérerons ces paquets pour le calcul de la corrélation.

A présent, nous pouvons passer à l'étape des calculs de corrélations et à celui de la τ-cohérence.

4.2.3 Reconstruction de la matrice de corrélation

À cette étape, nous disposons des fichiers binaires qui contiennent l'ensemble des observations. Pour rappel, chaque paquet est de dimension $n \times T_b$ et sont au nombre de N_b. Nous allons calculer alors les corrélations de chaque paquet puis nous allons croiser les corrélations pour reconstruire complètement la matrice R_n. Pour plus de clarté sur les différents blocs, nous allons écrire la matrice R_n de la manière suivante :

$$R_n = \begin{pmatrix} B_{1,1} & B_{1,2} & B_{1,3} & \dots & B_{1,N_b} \\ B_{2,1} & B_{2,2} & B_{2,3} & \dots & B_{2,N_b} \\ B_{3,1} & B_{3,2} & B_{3,3} & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \\ B_{N_b,1} & B_{N_b,2} & B_{N_b,3} & \ddots & B_{N_b,N_b} \end{pmatrix} \tag{4.7}$$

Chaque bloc $B_{i,j}$ est de dimension $T_b \times T_b$. On a donc besoin que ce bloc puisse tenir dans la RAM.

Remarques 11. *Par symétrie de la matrice R_n, nous calculerons uniquement les blocs correspondant à la matrice triangulaire supérieure de R_n (les blocs en rouge dans 4.7).*

Les calculs des blocs $B_{i,j}$ s'effectuent de la manière suivante :

- Pour i allant de 1 à N_b :
  1. On récupère le fichier binaire correspondant au paquet de colonnes $\mathbb{X}^i$.
  2. On calcule la covariance de ce paquet :
  $$cov\left(\mathbb{X}^i\right)$$
  On a alors une matrice de taille $T_b \times T_b$.
  3. On récupère les variances qui occupent la diagonale, et que l'on stocke dans le vecteur
  $$D \leftarrow diag\left(cov\left(\mathbb{X}^i\right)\right)$$
  4. On calcule les inverses des écart-types :
  $$D \leftarrow \sqrt{D}^{-1}$$
  5. On réduit la matrice de covariance pour obtenir celle de corrélation :
  $$B_{i,i} \leftarrow D * cov\left(\mathbb{X}^i\right) * D$$
  On rappelle que $B_{i,i}$ correspond à $cor\left(\mathbb{X}^i\right)$.

6. Pour j allant de $i+1$ à N_b :
   (a) On charge le paquet de colonnes $\mathbb{X}^j$.
   (b) On récupère les écarts-types des variables du bloc j, que l'on inverse et que l'on stocke dans le vecteur : D^j. Pour ce faire, on utilise une simple boucle sur les colonnes de $\mathbb{X}^j$ où on calcule pour chacune d'entre elles l'écart-type avec la fonction $sd(.)$ de R.
   (c) On calcule la covariance entre les paquets $\mathbb{X}^i$ et $\mathbb{X}^j$:

$$cov\left(\mathbb{X}^i, \mathbb{X}^j\right)$$

   (d) On renormalise pour obtenir la corrélation et donc on multiplie par D à droite et D^j à gauche.

$$B_{i,j} \leftarrow D * cov\left(\mathbb{X}^i, \mathbb{X}^j\right) * D^j$$

   $B_{i,j}$ correspond bien à $cor\left(\mathbb{X}^i, \mathbb{X}^j\right)$

Pour reconstruire la matrice R_n, nous n'utilisons pas la fonction `cor(·)` classique de R. En effet, nous devons gérer le cas où p (et donc R_n) est très grand. Très rapidement, au delà du problème de mémoire que cela entraîne, la fonction `cor(·)` demande un temps assez important de calcul si on l'utilise normalement, c'est-à-dire en CPU. A cette étape, deux choix sont possibles : utiliser du calcul parallèle "classique" en utilisant les "coeurs" de la machine (typiquement de 4 à 20), ou bien du calcul GPGPU utilisant les "coeurs" de la GPU (plus nombreux). Dans les deux cas, le logiciel R propose des librairies adaptées. Nous avons fait le choix d'envoyer plutôt le calcul de la corrélation en GPU (*Graphical Processing Units*). Cela nous permet de gagner beaucoup de temps de calcul. Dans la section suivant, nous exposerons les résultats de gain de temps entre la manière classique (CPU) et celle en GPU. Cependant, dans le package R que nous utilisons, à savoir *gpuR*, la fonction `cor(.)` n'est pas programmée, mais celle de `cov(.)` l'est. D'où notre choix. De plus, quitte à opter pour la GPU, nous envoyons tous les produits matricielles en GPU. Y compris ceux pour la création de la matrice d'observation $\mathbb{X}_n$.

Remarques 12. *Le choix du nombre N_b est libre pour l'utilisateur. Il faut simplement que ce dernier soit choisi de tel sorte que l'on puisse stocker un bloc $B_{i,j}$, deux vecteurs de variances et les deux paquets $\mathbb{X}^i$ et $\mathbb{X}^j$ simultanément. De manière large, on peut donc choisir N_b de telle façon que l'on puisse stocker deux blocs de corrélations simultanément.*

4.2.4 Calcul de la τ-cohérence

À présent, on s'intéresse au calcul de la τ-cohérence. Puisque nous avons reconstruit la matrice R_n avec des blocs $B_{i,j}$, il suffit de raisonner sur les blocs pour en retirer la τ-cohérence. On rappelle une nouvelle fois que celle-ci est définie par :

$$L_{n,\tau} = \max_{|i-j|\geqslant\tau} |\rho_{ij}|$$

On va donc prendre le maximum des coefficients de R_n en dehors de la bande centrale de largeur τ. Pour rappel, on a choisi la taille T_b de telle sorte qu'elle soit un multiple de τ. Pour récupérer $L_{n,\tau}$, nous allons mettre des 0 dans les blocs là où l'on ne doit pas calculer le maximum. Il y a donc deux cas à traiter :

1. Si on est sur un bloc diagonal : $B_{i,i}$ avec $1 \leqslant i \leqslant N_b$. Alors il suffit d'appliquer le petit algorithme :

   Listing 4.2 – boucle pour le calcul de la τ-cohérence sur un bloc diagonal

```
for (u in 1:Tb){
        for (v in 1:Tb){
                if (abs(v-u)<tau){
                        B[u,v]=0
                }
        }
}
```

   où $B[u, v]$ est ici la quantité à la ligne u, colonne v du bloc $B_{i,i}$. Puis, on récupère le maximum du bloc $\max(B_{i,i})$

2. Si on est sur un bloc de la forme $B_{i,i+1}$, avec $i \leqslant N_b - 1$: il faut retirer dans le calcul du maximum les termes qui font partie de la bande centrale. Ceux-ci correspondent aux coefficients contenus dans le coin inférieur gauche du bloc. On met alors à 0 ces quantités à l'aide du code suivant :

   Listing 4.3 – boucle pour le calcul de la τ-cohérence sur un bloc juxtaposé à la diagonale

```
for( v in 1:(tau-1) ){
        for( u in (Tb-tau+v+1):Tb ){
                B[u,v]<-0
        }
}
```

Sur les autres blocs $B_{i,j}$ ne rentrant pas dans les cas spécifiques précédent, il suffit de récupérer le maximum des coefficients de corrélation. À la sortie de l'algorithme, on calcule le maximum entre tous les maximums de chaque bloc ce qui nous donne la τ-cohérence $L_{n,\tau}$ de la matrice R_n.

4.3 Résultats

4.3.1 Visualisation de la convergence en loi

Pour observer numériquement la convergence en loi, nous effectuons des réplications de Monte-Carlo pour créer un échantillon de taille R de τ-cohérence. Pour bien la visualiser, nous allons considérer des tailles n d'échantillons de plus en plus grandes. De plus, nous souhaitons observer un régime pour p qui s'approche au maximum de la condition du théorème (i.e. $\log(p_n) = o(n^{1/3})$ quand $n \to +\infty$). Nous résumons notre modèle de simulation et nos paramètres ici :

1. Nous considérons les valeurs de n suivantes :

$$n = 1250, 1500, 1750, 2000, \ldots, 4750, 5000$$

2. Nous considérons les paramètres suivants pour notre modèle :

$$p = \left[\exp\left(n^{\frac{1}{3.5}}\right)\right] \ , \tau = 50 * [\log(p)] \ , K = 100 * \left[n^{1/10} * \log(p)\right] \ , \varepsilon_n = 0.1 * \sqrt{\frac{log(p)}{n}}$$

   où [.] désigne la partie entière.

3. Pour chaque valeurs de n, nous effectuons $R = 200$ réplications de Monte-Carlo. C'est-à-dire que pour $i = 1, \ldots, R$:
   (a) On génère la matrice d'observation $\mathbb{X}_n$ par découpage.
   (b) On reconstruit la matrice de corrélation par bloc et on récupère la valeur de la τ-cohérence pour cette matrice d'observation $\mathbb{X}_n$.
   (c) On sauvegarde cette donnée, notée $L_{n,\tau}^{(i)}$.

Ainsi, pour n fixé, on dispose de l'échantillon de taille R de Monte-Carlo :

$$L_{n,\tau}^{(1)}, \ldots, L_{n,\tau}^{(R)}$$

On effectue la normalisation de notre échantillon afin de construire la variable qui intervient dans notre résultat de convergence asymptotique :

$$\text{Pour } i = 1, \ldots, R : \quad Z_n^{(i)} = n\left(L_{n,\tau}^{(i)}\right)^2 - 4\log(p) + \log\log(p) \tag{4.8}$$

Ce qui nous donne un R échantillon pour observer la convergence en loi :

$$\left(Z_n^{(1)}, \ldots, Z_n^{(R)}\right).$$

Pour notre problème, la variable aléatoire asymptotique est de loi connue (dont on connaît la fonction de distribution). A l'occasion du chapitre 2, nous avons précisé qu'il s'agissait d'une loi de Gumbel. Cette loi apparaît notamment lorsque l'on s'intéresse au maximum de variables aléatoires indépendantes. La τ-cohérence ici n'est pas exactement le maximum de variable indépendante. Nous avons une partie de nos variables qui sont corrélées d'un coefficient ε_n. Donc à n fixé, nos variables sont dépendantes. Cependant, puisqu'asymptotiquement, $\varepsilon_n \to 0$, le résultat de convergence (et l'apparition de la loi de Gumbel) montre qu'il y a bien une indépendance asymptotique entre nos variables intervenant dans le calcul de $L_{n,\tau}$.

Dans notre cas, on note Z la variable asymptotique de notre théorème. Nous avons alors $Z \sim Gumbel\,(\mu, \beta)$ avec $\mu = -\log(8\pi)$ et $\beta = 2$ (voir définition 1.3.3). En particulier, nous savons que :

- La fonction de répartition de Z est donnée, pour tout $x \in \mathbb{R}$ par :

$$F(x) = \exp\left[-\frac{1}{\sqrt{8\pi}}\exp\left(-\frac{x}{2}\right)\right]$$

- Nous avons, pour les moments d'ordres 1 et 2 :

$$\mathbb{E}[Z] = \mu + \beta\gamma = -\log(8\pi) + 2\gamma$$

où γ est la constante d'Euler-Mascheroni, et :

$$var\,(Z) = \frac{\pi^2}{6}\beta^2 = \frac{2}{3}\pi^2$$

- Nous allons également nous intéresser aux quantiles de la loi de Gumbel. En particulier nous avons :

$$mediane(Z) = \mu - \beta\log\log(2) = -\log(8\pi) - 2\log\log(2)$$

Et en générale, le quantile à l'ordre $\alpha \in]0,1[$ de la loi de $Gumbel\,(\mu, \beta)$, noté q_α et vérifiant $F(q_\alpha) = \alpha$, est donné par :

$$q_\alpha = \mu - \beta\log(-\log(\alpha)) = -\log(8\pi) - 2\log(-\log(\alpha))$$

Ainsi, pour visualiser la convergence nous allons nous appuyer sur :

1. L'observation des histogrammes de notre échantillon $\left(Z_n^{(1)}, \ldots, Z_n^{(R)}\right)$ comparé avec la densité de la loi de Z.
2. L'observation de l'estimation des quantiles à l'ordre $\alpha = 0.1, 0.2, 0.3, \ldots, 0.9$ en fonction de n.
3. L'observation de différentes distances entre la loi de notre échantillon et celle de Z. En particulier, nous utiliserons la distance de Kolmogorov, la distance $\mathbb{L}^2$ et la distance en variation totale.

↪ Visualisation des histogrammes : Nous avons obtenu, en construisant des échantillons Monte-Carlo de taille $R = 200$, les histogrammes présentés dans la Figure 4.3. Sur chacun des histogrammes, nous avons représenté, en bleu, la densité asymptotique, à savoir celle de Z. Elle est définie pour tout $x \in \mathbb{R}$, par :

$$f(x) = \frac{1}{2\sqrt{8\pi}} \exp\left(-\frac{1}{2}x - \frac{1}{\sqrt{8\pi}} \exp\left[-\frac{1}{2}x\right]\right)$$

Nous avons également ajouté aux histogrammes l'estimateur à noyau de la densité, disponible sous R via la fonction *density*, tracé en rouge. On constate que la distribution de nos échantillons se rapproche de celle de Z au fur et à mesure que n augmente. Remarquons que la convergence est lente. Cette vitesse de convergence lente est connue : [LLS08], [SZ14]. L'écart est visuellement encore important pour $n = 4000$.

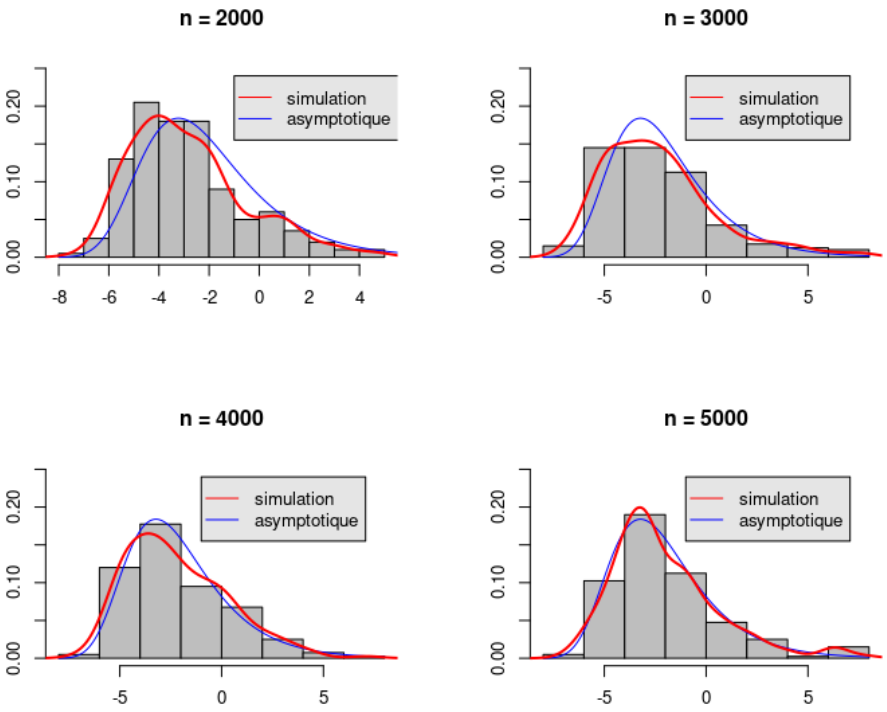

FIGURE 4.3 – Histogrammes des R-échantillons Monte-Carlo en fonction de n.

Regardons à présent ce qu'il se passe avec les quantiles empiriques et asymptotiques.

↪ Convergence des quantiles :

Ici, nous nous servons des quantiles de la loi de Z que nous avons noté q_α. Cette fois-ci, nous traçons en fonction de n, les valeurs des quantiles empiriques en fonction de n. Nous

utilisons la fonction *quantile* de R. Pour chaque valeur de $\alpha = 0.1, \ldots, 0.9$, nous traçons en bleu la valeur du quantile théorique de Z et en rouge la valeur du quantile empirique en fonction de n. Ces résultats sont présentés dans la Figure 4.4.

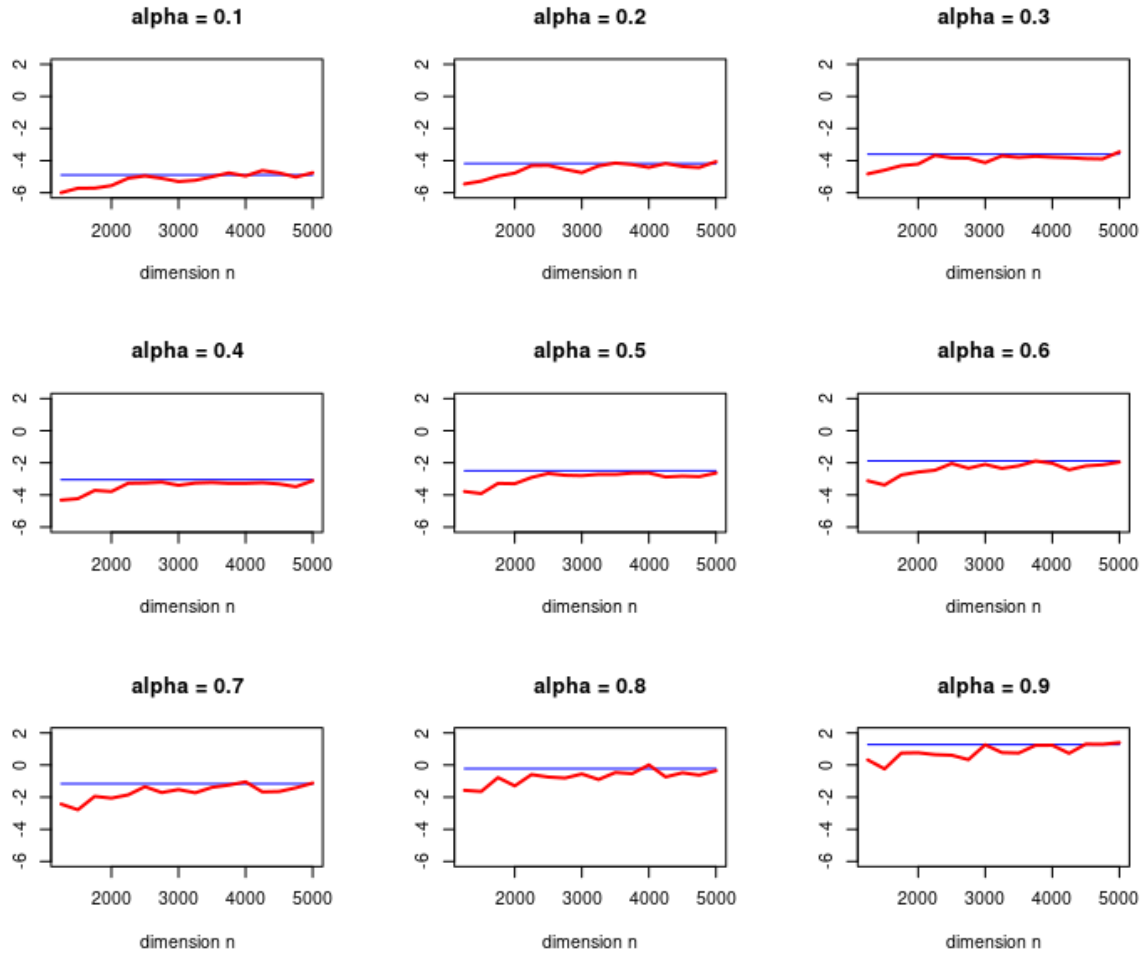

FIGURE 4.4 – Evolution des estimateurs des quantiles à l'ordre $\alpha = 0.1, \ldots, 0.9$ pour nos échantillons de τ-cohérence en fonction de n.

D'un point de vue global, on constate bien la convergence des quantiles. Remarquons encore une fois que pour n assez grand (aux alentours de $n = 4000$), on constate des différences entre l'estimation et la valeur cible. Il faut donc avoir des n suffisamment grand pour "attraper" convenablement la distribution asymptotique. Nous avons effectué notre étude jusqu'à $n = 5000$. En effet, tout en utilisant le package `gpuR`, le temps de calcul augmente.

Pour finir, nous allons étudier l'évolution de quelques distances entre les distributions de nos échantillons et celle de Z.

↪ Étude de distances entre la loi de notre échantillon et celle de Z :

Nous nous intéressons à trois distances, introduites dans le chapitre 2. Nous les rappelons ici. Il s'agit de la distance de Kolmogorov (notée d_{KS}), la distance en norme $\mathbb{L}^2$ (notée d_2) et celle en variation totale (notée d_{TV}) :

$$d_{KS}(\hat{f}, f) = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|, \tag{4.9}$$

où F_n est la fonction de répartition empirique de notre échantillon et F est la fonction de répartition de Z,

$$d_2(\hat{f}, f) = \int \left| \hat{f}(x) - f(x) \right|^2 dx, \tag{4.10}$$

où $\hat{f}$ est l'estimateur de la densité construit sur l'échantillon et f la densité de Z,

$$d_{TV}(\hat{f}, f) = \frac{1}{2} \int \left| \hat{f}(x) - f(x) \right| dx. \tag{4.11}$$

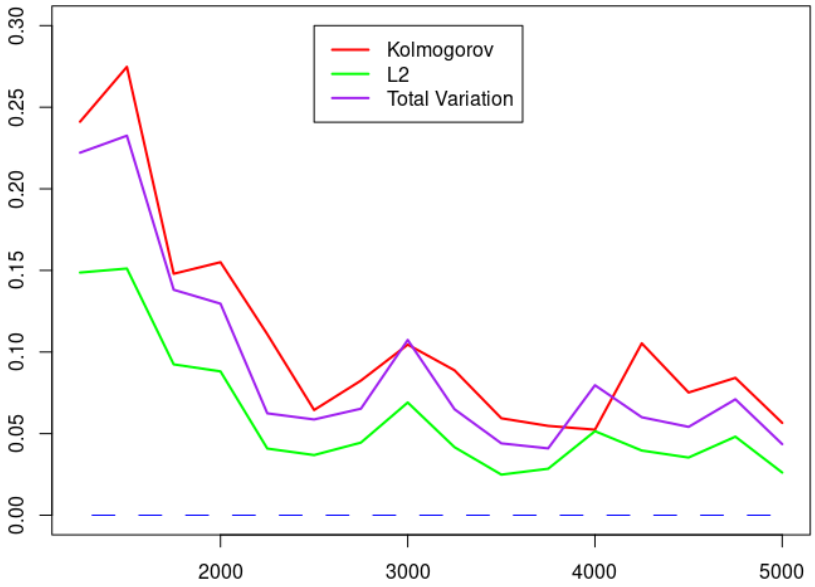

FIGURE 4.5 – Evolution des distances $\mathbb{L}^2$, Kolmogorov et Variation Totale en fonction de n.

Nous calculons les valeurs de ces trois distances en fonction de n, à chaque fois sur un échantillon de taille $R = 200$. Nous obtenons les résultats présentés en Figure 4.5. On constate bien la décroissance des distances au fur et à mesure que n croît vers $+\infty$. Remarquons que la décroissance "ralentit" pour n assez grand. On constate donc numériquement que cette convergence est lente et qu'il faut viser de grandes valeurs de n pour obtenir une adéquation convenable à la distribution asymptotique.

Au final, nous avons bien obtenu numériquement la convergence présentée dans le théorème 2.2.1. À présent, nous allons exposer en quoi l'utilisation du calcul en GPU a été une méthode pertinente pour nos simulations.

4.3.2 Utilisation de calculs en GPU

Au début de ce chapitre, nous avons présenté notre méthode de simulation qui consiste à découper nos matrices d'observations et de corrélations afin de pouvoir les étudier sans les charger complètement dans la mémoire. Ces matrices devenant beaucoup trop grandes (en terme de mémoire RAM). Notre schéma de simulation repose donc majoritairement sur du calcul matriciel. De façon générale, en utilisant le logiciel R, il est préférable d'utiliser cette méthode de calcul qui est plus rapide qu'un calcul semblable reposant sur des boucles `for`.

Très rapidement, on constate que les temps de simulation de notre schéma deviennent trop importants. En particulier, le calcul de la covariance entre nos paquets de colonnes de $\mathbb{X}_n$. Nous avons alors eu l'idée d'utiliser les calculs en GPU afin de pouvoir économiser du temps de simulation. Pour ce faire, nous avons utilisé le package *gpuR* de *R* (https ://CRAN.R-project.org/package=gpuR). Celui-ci fournit une interface R afin de pouvoir utiliser une GPU (l'utilisateur n'a ainsi pas besoin de coder en CUDA, le langage de programmation sur une GPU). Dans ce package, on trouve notamment la fonction `cov(.)` qui renvoit le calcul de la matrice de covariance programmé en GPU. Nous pouvons alors l'utiliser pour gagner du temps là où nous ne pouvions pas en gagner de façon classique.

De façon très claire, on constate que le temps gagné en utilisant un code GPU est très important. Cela nous permet donc de pouvoir créer des échantillons Monte-Carlo de τ-cohérence de tailles suffisantes en un temps raisonnable. En particulier, nous présentons dans la figure 4.6 le temps de simulation nécessaire pour obtenir une réalisation de $L_{n,\tau}$ pour les paramètres :

$$p = \left[\exp\left(n^{\frac{1}{3.5}}\right)\right] \ , \tau = 50 * [\log(p)] \ , K = 100 * \left[n^{1/10} * \log(p)\right] \ , \varepsilon_n = 0.1 * \sqrt{\frac{\log(p)}{n}}$$

Afin que la comparaison soit la plus juste possible, nous utilisons le même nombre de paquets N_b pour la simulation en CPU et en GPU (ce découpage étant adapté au fur et à mesure des valeurs de n et de p). Nous obtenons alors la Figure 4.6. Remarquons que pour $n = 4000$, le temps CPU est de 9731.586 s (soit 162.1931 min) alors que le temps GPU pour le même n est de 699.063 s (soit 11.65105 min). On a divisé le temps de simulation par 13.92.

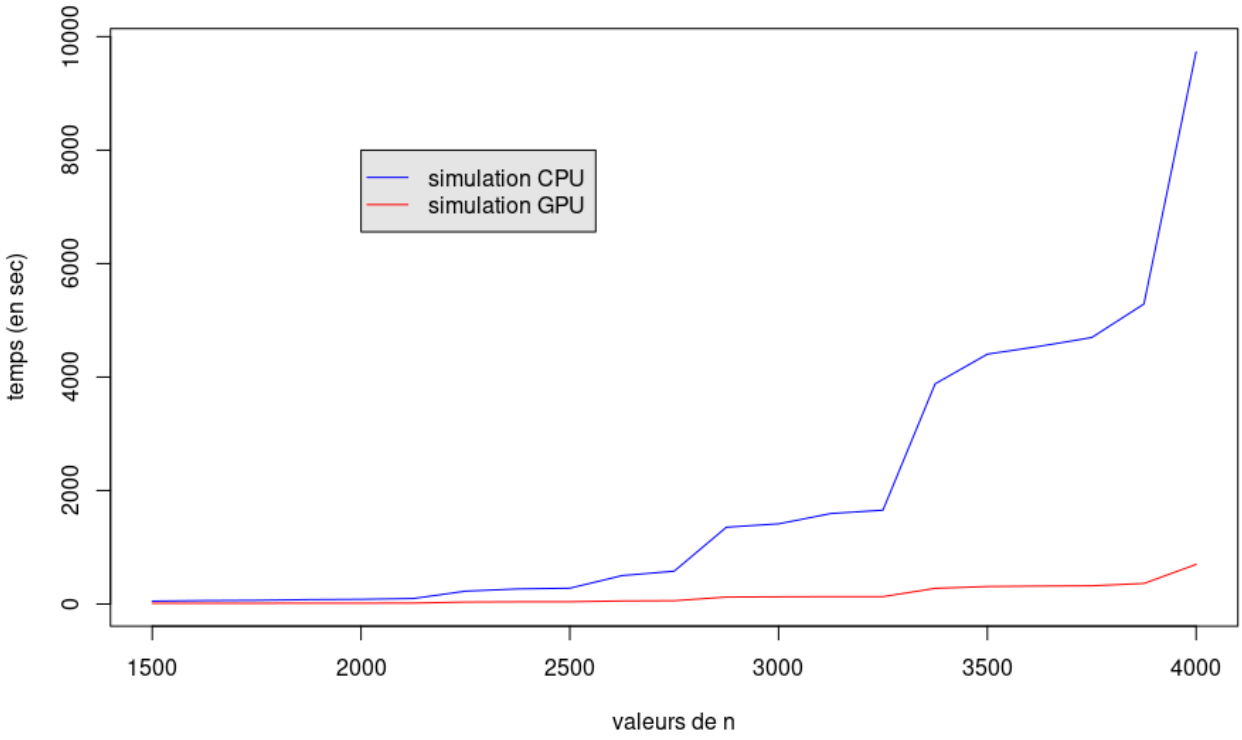

FIGURE 4.6 – Évolution du temps de simulation d'une réalisation de la τ-cohérence en fonction des valeurs de n pour la méthode de simulation CPU et GPU

Dans notre méthode pour calculer la τ-cohérence, nous découpons la matrice de corrélation en blocs. Afin de regarder l'impact du nombre de bloc, nous simulons une valeur de $L_{n,\tau}$ pour des valeurs de paramètres fixes, à savoir :

$$n = 4000 \qquad p = 10000 \qquad \tau = 200 \qquad \varepsilon = 1/n \qquad K = 800$$

Pour voir l'impact du nombre de découpages, nous considérons pour ces paramètres les valeurs de N_b suivantes :

$$N_{b,1} = 2 \qquad N_{b,2} = 10 \qquad N_{b,3} = 20 \qquad N_{b,4} = 25$$

Pour chaque valeurs de N_b, nous avons répété 7 simulations avec les mêmes paramètres. Nous avons ensuite fait les moyennes des temps obtenus, présentées dans le tableau suivant, en seconde :

	$N_{b,1} = 2$	$N_{b,2} = 10$	$N_{b,3} = 20$
GPU	29.27571	36.55586	73.583
CPU	315.905	409.1457	716.6927

On constate que pour un petit nombre de découpage N_b, le temps de simulation en moyenne est plus faible. De manière générale, on cherche à minimiser les échanges entre la GPU et le disque dur où les fichiers binaires sont stockés. On cherche donc à sélectionner un nombre de paquets le plus petit possible. Remarquons que cela est également valable pour la programmation en CPU (on observe à nouveau un gain de temps important en utilisant le calcul en GPU).

Annexe A

Notations fondamentales, résultats intermédiaires

Sommaire

A.1 Notations pour les études asymptotiques

Au cours de cette thèse, nous utilisons plusieurs notations liées aux asymptotiques : o, $\mathcal{O}$ et $\sim$. Nous rappelons leurs définitions :

Definition A.1.1. *On note* $u_n \underset{n\to+\infty}{\sim} v_n$ *ou* $u_n \sim v_n$ *quand* $n \to +\infty$ *si :*

$$\lim_{n\to+\infty} \frac{u_n}{v_n} = 1$$

On dit alors que u_n *est équivalent à* v_n *quand* $n \to +\infty$.

Definition A.1.2. *On note* $u_n = o_{n\to+\infty}(v_n)$ *ou* $u_n = o(v_n)$ *quand* $n \to +\infty$ *si :*

$$\lim_{n\to+\infty} \frac{u_n}{v_n} = 0$$

On dit alors que u_n *est négligeable devant* v_n *quand* $n \to +\infty$.

Definition A.1.3. *On note* $u_n = \mathcal{O}_{n\to+\infty}(v_n)$ *ou* $u_n = \mathcal{O}(v_n)$ *s'il existe une constante* $C > 0$ *telle que, quand* $n \to +\infty$ *:*

$$\left|\frac{u_n}{v_n}\right| \leqslant C$$

On dit alors que u_n *est dominée par* v_n *quand* $n \to +\infty$*.*

A.2 Lemmes provenant des travaux de T. Cai et al.

Les lemmes suivant proviennent de [CJ11a] et [CJ11b] et correspondent respectivement aux lemmes 6.9 et 6.11.

Le lemme suivant est utilisé pour pour les calculs de Q_1 du chapitre 2. Nous renvoyons à [CJ11b] pour la preuve de ce lemme.

Lemme A.2.1. *Soit* $(u_k^1, u_k^2, u_k^3, u_k^4)_{1\leqslant k\leqslant n}$ *une suite indépendante et identiquement distribuée de vecteurs aléatoires suivant la loi* $\mathcal{N}_4(0, \Sigma^*)$ *avec*

$$\Sigma^* = \begin{pmatrix} 1 & 0 & r & 0 \\ 0 & 1 & 0 & 0 \\ r & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ et } |r| \leqslant 1.$$

Soit $a_n = \sqrt{4n\log(p) - n\log\log(p) + ny}$ *où* $y \in \mathbb{R}$*. Supposons que* $n \to +\infty$*,* $p(n) \to +\infty$ *tel que* $\log(p) = o\left(n^{1/3}\right)$*.*
Alors, pour tout $\omega > 0$

$$\sup_{|r|\leqslant 1} \mathbb{P}\left(\left|\sum_{k=1}^n u_k^1 u_k^2\right| > a_n, \left|\sum_{k=1}^n u_k^3 u_k^4\right| > a_n\right) = \mathcal{O}\left(p^{-4+\omega}\right). \tag{A.1}$$

Le lemme suivant intervient dans nos démonstrations du chapitre 2.

Lemme A.2.2. *Soit* $(u_k^1, u_k^2, u_k^3, u_k^4)_{1\leqslant k\leqslant n}$ *une suite indépendante et identiquement distribuée de vecteurs aléatoires suivant la loi* $\mathcal{N}_4(0, \Sigma^*)$ *avec*

$$\Sigma^* = \begin{pmatrix} 1 & 0 & r_1 & 0 \\ 0 & 1 & 0 & r_2 \\ r_1 & 0 & 1 & 0 \\ 0 & r_2 & 0 & 1 \end{pmatrix} \text{ et } |r_1| \leqslant 1,\ |r_2| \leqslant 1.$$

Soit $a_n = \sqrt{4n\log(p) - n\log\log(p) + ny}$ *où* $y \in \mathbb{R}$. *Supposons que* $n \to +\infty$, $p \to +\infty$ *tel que* $\log(p) = o\left(n^{1/3}\right)$.
Alors, pour tout $\delta \in]0,1[$, *il existe un* $\omega_0 = \omega(\delta) > 0$ *tel que, pour* $n \to +\infty$ *:*

$$\sup_{|r_1|,|r_2|\leqslant 1-\delta} \mathbb{P}\left(\left|\sum_{k=1}^{n} u_k^1 u_k^2\right| > a_n, \left|\sum_{k=1}^{n} u_k^3 u_k^4\right| > a_n\right) = \mathcal{O}\left(p^{-2-\omega_0}\right). \tag{A.2}$$

Nous utilisons ce lemme A.2.2, à la fois dans le modèle de corrélations en bandes et celui par blocs. En particulier, nous utilisons la majoration de la probabilité. Nous exposons les éléments principaux de la démonstration.

Démonstration du Lemme A.2.2. Le principe de cette preuve repose sur l'utilisation d'un changement de variable. Soit

$$X = (u_k^1, u_k^2, u_k^3, u_k^4) \text{ et } Y = (u_k^1, u_k^2, r_1 u_k^1 + r_1' u_k^5, r_2 u_k^2 + r_2' u_k^6)$$

où $r_1' = \sqrt{1 - r_1^2}$, $r_2' = \sqrt{1 - r_2^2}$ et où u_k^5, u_k^6 sont indépendantes de toutes les autres variables. Alors on peut montrer facilement, en utilisant les fonctions caractéristiques, que

$$X \overset{\mathcal{L}}{=} Y \sim \mathcal{N}_4\left(0, \Sigma^*\right)$$

On garde la notation $Z_{ij} = \left|\sum_{k=1}^{n} u_k^i u_k^j\right|$. En majorant de manière classique, et en bornant les coefficients $|r_1|, |r_2| \leqslant 1 - \delta$ pour $\delta \in]0,1[$ on a :

$$\left|\sum_{k=1}^{n} \left(r_1 u_k^1 + r_1' u_k^5\right)\left(r_2 u_k^2 + r_2' u_k^6\right)\right| \leqslant (1-\delta)^2 Z_{12} + 3\max\left(Z_{16}, Z_{25}, Z_{56}\right)$$

Posons :

$$a = \tfrac{1}{2}\left(1 + (1-\delta)^2\right) \qquad b = \tfrac{a}{(1-\delta)^2} > 1 \qquad c = \tfrac{1}{3}(1-a) > 0$$

On constate alors que si $Z_{12} \leqslant b a_n$ et $\max\left(Z_{16}, Z_{25}, Z_{56}\right) \leqslant c a_n$, alors :

$$\left|\sum_{k=1}^{n} \left(r_1 u_k^1 + r_1' u_k^5\right)\left(r_2 u_k^2 + r_2' u_k^6\right)\right| \leqslant a_n$$

Ainsi, en utilisant également l'égalité en loi entre X et Y, et l'hypothèse d'indépendance entre les variables :

$$\begin{aligned} & \mathbb{P}\left(\left|\sum_{k=1}^{n} u_k^1 u_k^2\right| > a_n, \left|\sum_{k=1}^{n} u_k^3 u_k^4\right| > a_n\right) \\ = \quad & \mathbb{P}\left(\left|\sum_{k=1}^{n} u_k^1 u_k^2\right| > a_n, \left|\sum_{k=1}^{n} \left(r_1 u_k^1 + r_1' u_k^5\right)\left(r_2 u_k^2 + r_2' u_k^6\right)\right| > a_n\right) \\ \leqslant \quad & \mathbb{P}\left(Z_{12} > b a_n\right) + 2\mathbb{P}\left(Z_{12} > a_n, Z_{16} > c a_n\right) + \mathbb{P}\left(Z_{12} > a_n, Z_{56} > c a_n\right) \end{aligned}$$

À présent, en utilisant à nouveau les résultats de T. Cai et al [CJ11b] (équation (123) et (126)), on sait que :

$$\mathbb{P}\left(Z_{12} > b a_n\right) = \mathcal{O}\left(p^{-2b^2+\omega_1}\right) \text{ pour tout } \omega_1 > 0 \text{ et pour } n \to +\infty \tag{A.3}$$

$$\mathbb{P}\left(Z_{12} > a_n, Z_{16} > c a_n\right) = \mathcal{O}\left(p^{-2-2c^2+\omega_2}\right) \text{ pour tout } \omega_2 > 0 \text{ et pour } n \to +\infty \tag{A.4}$$

$$\mathbb{P}\left(Z_{12} > a_n, Z_{56} > c a_n\right) = \mathcal{O}\left(p^{-2-2c^2+\omega_3}\right) \text{ pour tout } \omega_3 > 0 \text{ et pour } n \to +\infty \tag{A.5}$$

On en déduit alors le résultat du lemme A.2. □

Nous utilisons de façon systématique, dans les chapitres 3 et 4, les ordres de grandeur de (A.3) à (A.5)

A.3 Éléments de preuves supplémentaires pour le modèle de corrélation par bandes

Nous apportons ici quelques éléments supplémentaires aux différentes preuves provenant du modèle de corrélations par bandes. Pour appuyer le propos, nous illustrons les découpages des différents voisinages. En particulier, on expose le découpage du voisinage pour le calcul de la quantité Q_1 dans la figure A.1, celui pour le calcul du Q_2 (en particulier pour le découpage de Λ_p^0 en $\Lambda_{p,I}^0$ et $\Lambda_{p,II}^0$) dans A.2 et A.3. Enfin, on illustre le découpage du voisinage B_α^0 pour le calcul du Q_3.

Preuve du lemme 2.4.8. Pour montrer

$$\left|\Lambda_p^0\right| . \left|\Omega_2\right| \mathbb{P}_{02} \to 0,$$

commençons par expliciter la matrice de corrélation intervenant pour la loi normale du vecteur de dimension 4 dont $\mathbb{P}_{02}$ est issue.

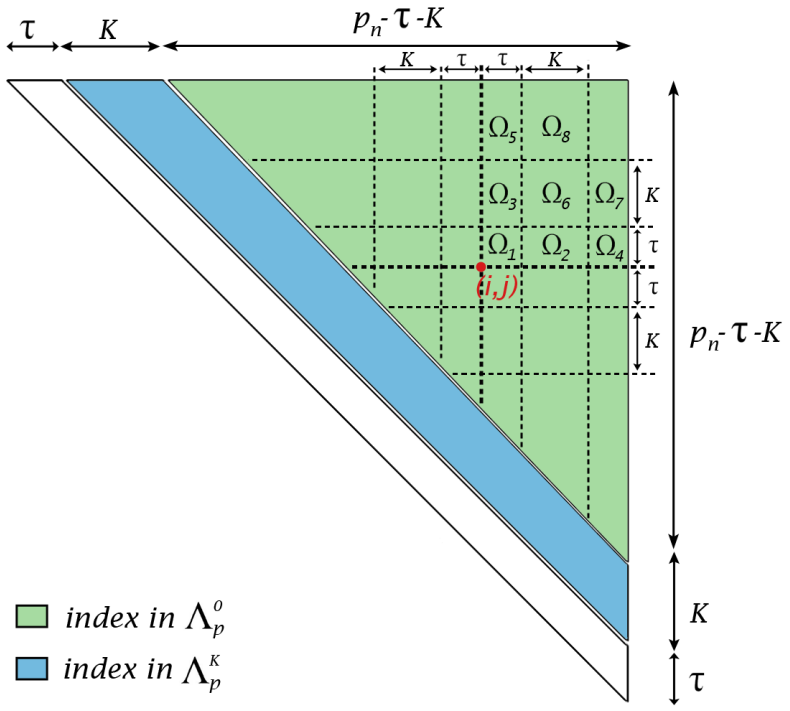

FIGURE A.1 – Découpage du voisinage pour le calcul de Q_1 (voir 2.4.3)

Soit $\alpha = (i,j) \in \Lambda_p^0$ et $\beta = (u,v) \in \Omega_2$. Par la définition de Ω_2, nous savons que la matrice de corrélation du vecteur $\left(X_k^i, X_k^j, X_k^u, X_k^v\right)$ est de la forme :

$$\Sigma = \begin{pmatrix} 1 & 0 & r_1 & \times \\ 0 & 1 & \times & r_2 \\ r_1 & \times & 1 & 0 \\ \times & r_2 & 0 & 1 \end{pmatrix}$$

Or, dans ce cas précis, nous avons :

$$v - i = \underbrace{v - j}_{>0} + \underbrace{j - i}_{>\tau+K} > \tau + K \tag{A.6}$$

D'où

$$cor\left(X_k^i, X_k^v\right) = 0. \tag{A.7}$$

De la même manière, nous avons :

$$j - u = j - i + i - u > \tau + K + 0 \Rightarrow cor\left(X_k^j, X_k^u\right) = 0 \tag{A.8}$$

Ainsi, on a la matrice de corrélation pour ce cas :

$$\Sigma = \begin{pmatrix} 1 & 0 & r_1 & 0 \\ 0 & 1 & 0 & r_2 \\ r_1 & 0 & 1 & 0 \\ 0 & r_2 & 0 & 1 \end{pmatrix}$$

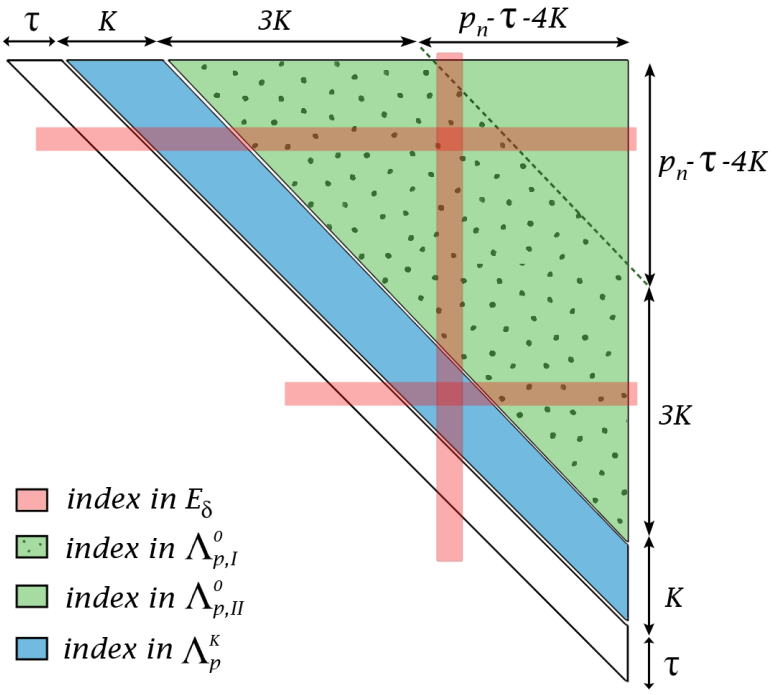

FIGURE A.2 – Découpage de l'ensemble Λ_p^0 en $\Lambda_{p,I}^0$ et $\Lambda_{p,II}^0$.

Nous sommes dans un cas similaire à celui du lemme A.2.2. En effet, nous savons que $\varepsilon_n \to +\infty$ quand $n \to +\infty$. Il existe donc un rang $n_0 \in \mathbb{N}$ tel que pour tout $n \geqslant n_0$, $|\varepsilon_n| \leqslant 1 - \delta$. On peut donc contrôler la probabilité $\mathbb{P}_{02}$ de la même manière que celle du lemme A.2.2. On a alors, pour $\mathbb{P}_{02}$, et en utilisant $|\Lambda_p^0| \sim p^2$ et $|\Omega_2| \leqslant \tau K$, en gardant les mêmes notations a, b et c de la preuve du lemme A.2.2 et avec les formules A.3 et A.5 :

$$\begin{aligned}
\tau K p^2 \mathbb{P}_{02} &= \tau \mathcal{O}\left(p^{2+\nu}\right) \mathbb{P}_{02} && \text{(A.9)}\\
&\leqslant \tau \mathcal{O}\left(p^{2+\nu-2b^2+\omega_1}\right) + \tau \mathcal{O}\left(p^{2+\nu-2-2c^2+\omega_2}\right) && \text{(A.10)}\\
&\leqslant \tau \mathcal{O}\left(p^{2+\nu-2b^2+\omega_1}\right) + \tau \mathcal{O}\left(p^{\nu-2c^2+\omega_2}\right) && \text{(A.11)}\\
& && \text{(A.12)}
\end{aligned}$$

Pour avoir la convergence vers 0, puisque $\tau = o\left(p^t\right)$ pour tout $t > 0$, il faut que ν vérifie deux conditions :

$$2b^2 - 2 - \nu > 0 \text{ et } 2c^2 - \nu > 0$$

Cela nous donne une condition sur ν qui dépend du δ puisque les quantités b et c en dépendent. On rappelle que $2b^2 - 2 > 0$ et $2c^2 > 0$. À présent, on compare, selon les valeurs de $\delta \in]0, 1[$ quelle condition est la plus restrictive. Pour cela, on pose :

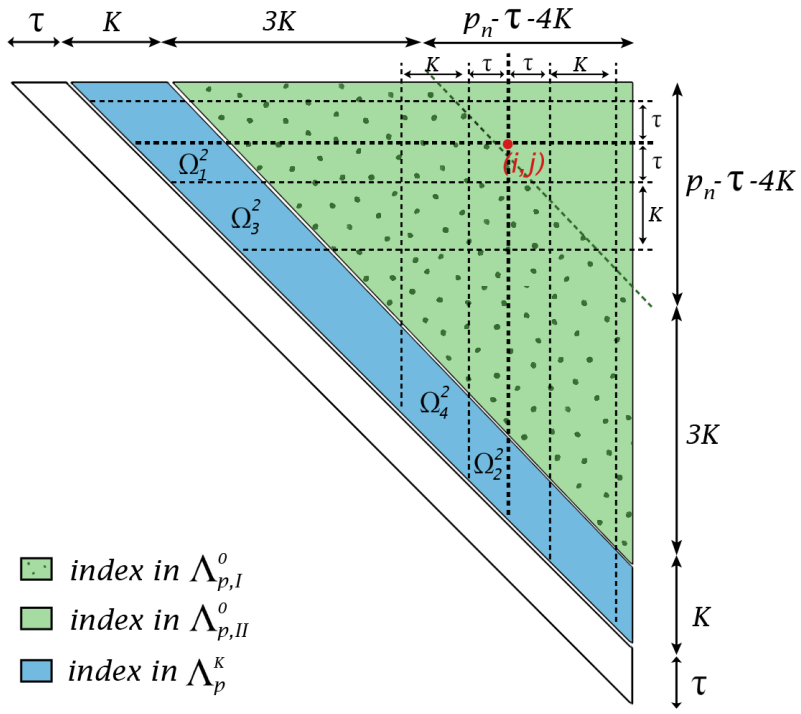

FIGURE A.3 – Découpage du voisinage sur Λ_p^K pour le calcul de Q_2.

$$\text{Pour tout } \delta \in]0,1[,\ h(\delta) = b^2 - 1 - c^2 = \tfrac{1}{2}(1-\delta)^{-2} + \tfrac{1}{2} - 1 - \left(\tfrac{1}{6} - \tfrac{1}{6}(1-\delta)^2\right)^2$$

En particulier, on a, pour $0 < \delta < 1$:

$$h'(\delta) = \frac{1}{(1-\delta)^3}\left[1 + \frac{1}{(1-\delta)^2}\right] - \frac{1}{9}(1-\delta)\delta(2-\delta)$$

Puisque $0 < \delta < 1$, on a $h'(\delta) > 0$ en constatant que $\frac{1}{(1-\delta)^3}\left[1 + \frac{1}{(1-\delta)^2}\right] > 1$ et $0 < \frac{1}{9}(1-\delta)\delta(2-\delta) < 1$. Donc h est croissante sur $]0,1[$ et $h(0) = 0$ d'où $h(\delta) > 0$ pour tout $0 < \delta < 1$. On en conclut :

$$b^2 - 1 > c^2$$

Ainsi, la condition la plus restrictive pour le ν dans ce cas est :

$$\nu < 2c^2 = \tfrac{1}{18}\delta^2(2-\delta)^2$$

Avec cette condition, $2b^2 - 2 - \nu > 0$ et $2c^2 - \nu > 0$ et donc il existe $t > 0$, $\omega_1(\delta) > 0$ et $\omega_2(\delta) > 0$ tels que :

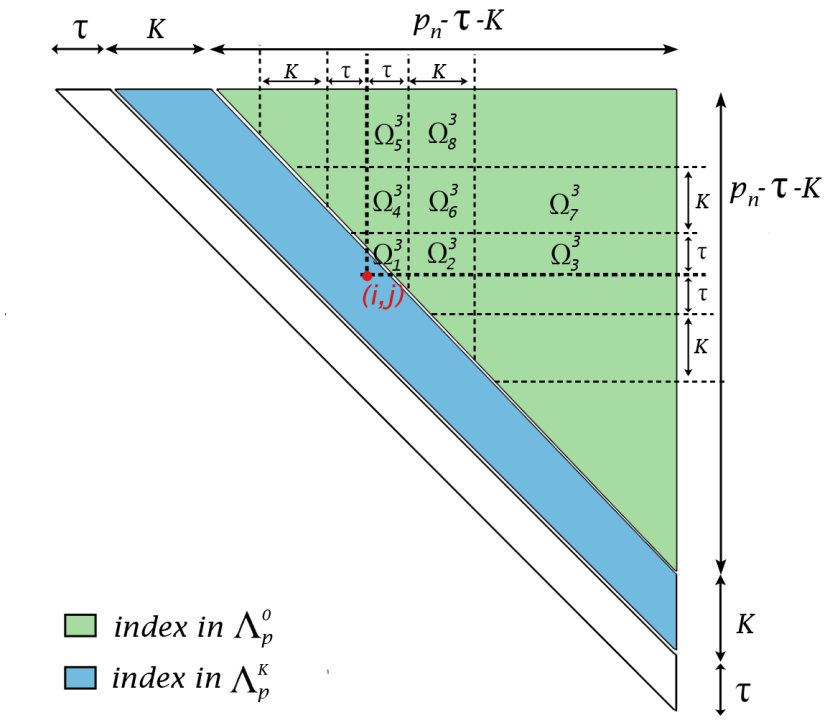

FIGURE A.4 – Découpage du voisinage B_α^0 pour le calcul de Q_3.

$$\tau\mathcal{O}\left(p^{2+\nu-2b^2+\omega_1}\right) \longrightarrow 0 \text{ quand } n \to +\infty$$

et

$$\tau\mathcal{O}\left(p^{\nu-2c^2+\omega_2}\right) \longrightarrow 0 \text{ quand } n \to +\infty$$

Et donc :

$$\tau K p^2 \mathbb{P}_{02} \underset{n\to+\infty}{\longrightarrow} 0$$

□

Preuve du lemme 2.4.9. Ici nous traitons la quantité $|\Lambda_p^0|.|\Omega_3|.\mathbb{P}_{03}$. La méthode est très semblable au cas précédent du lemme 2.4.8. En effet, regardons la matrice de covariance impliquée dans le calcul, celle du vecteur gaussien de dimension 4 qui intervient dans la construction de $\mathbb{P}_{03}$: soit $\alpha = (i,j) \in \Lambda_p^0$ et $\beta = (u,v) \in \Omega_3$. Par la définition de Ω_3, la

matrice est de la forme :

$$\Sigma = \begin{pmatrix} 1 & 0 & r_1 & \times \\ 0 & 1 & \times & r_2 \\ r_1 & \times & 1 & 0 \\ \times & r_2 & 0 & 1 \end{pmatrix}$$

En comparant les indices (i, v) et (j, u), on a :

$$v - i = \underbrace{v - j}_{>0} + \underbrace{j - i}_{>\tau+K} > \tau + K \text{ et } j - u = j - i + i - u > \tau + K + 0$$

Ceci implique d'après notre modèle :

$$cor\left(X_k^i, X_k^v\right) = 0 \text{ et } cor\left(X_k^j, X_k^u\right) = 0.$$

Ainsi, la matrice de covariance est bien de la même forme que pour le lemme 2.4.8, à savoir, avec $|r_1|, |r_2| \leqslant 1$

$$\Sigma = \begin{pmatrix} 1 & 0 & r_1 & 0 \\ 0 & 1 & 0 & r_2 \\ r_1 & 0 & 1 & 0 \\ 0 & r_2 & 0 & 1 \end{pmatrix}$$

On peut donc à nouveau utiliser la majoration provenant de la preuve du lemme A.2.2. De plus, puisque les cardinaux sont les mêmes que pour le lemme 2.4.8, on conclut directement sur la convergence

$$\left|\Lambda_p^0\right| . \left|\Omega_3\right| \mathbb{P}_{03} \underset{n\to+\infty}{\longrightarrow} 0$$

sous condition que $\nu < \frac{1}{18}\delta^2(2-\delta)^2$, c'est-à-dire la même condition qui apparaît dans la preuve précédente. □

Remarques 13. *Pour la preuve du théorème 2.2.1, il y a également le lemme 2.4.12 où la structure de la matrice de covariance est similaire aux précédentes. En effet, pour ce lemme, on montre, de manière similaire aux lemmes 2.4.8 et 2.4.9, que la matrice de corrélation impliquée dans le calcul de* $\mathbb{P}_{06}$ *est :*

$$\Sigma = \begin{pmatrix} 1 & 0 & \varepsilon_n & 0 \\ 0 & 1 & 0 & \varepsilon_n \\ \varepsilon_n & 0 & 1 & 0 \\ 0 & \varepsilon_n & 0 & 1 \end{pmatrix}.$$

Pour ce lemme, on utilise donc la même majoration de la probabilité que pour les preuves précédentes, et on obtient avec une approximation des cardinaux $|\Lambda_p^0| \sim \frac{1}{2}p^2$ *et* $|\Omega_6| \leqslant K^2$*, ainsi que l'hypothèse* $K = \mathcal{O}\left(p^\nu\right)$*, que :*

$$|\Lambda_p^0|.|\Omega_6|.\mathbb{P}_{06} \leqslant \mathcal{O}\left(p^{2\nu+2}\mathbb{P}_{06}\right)$$

On utilise donc les équations (A.3), (A.4) et (A.5) pour majorer $\mathbb{P}_{06}$*. On obtient alors les conditions suivantes pour la convergence nulle de* $|\Lambda_p^0|.|\Omega_6|.\mathbb{P}_{06}$ *:*

$$2b^2 - 2 - 2\nu > 0 \text{ et } c^2 - \nu > 0 \Rightarrow \nu < c^2$$

En effet, au cours de la preuve du lemme 2.4.8, nous avons montré que $c^2 < b^2 - 1$. *D'où, la condition sur* ν :

$$\nu < c^2 = \frac{1}{36}\delta^2 (2-\delta)^2 \tag{A.13}$$

Remarquons que cette condition est la plus restrictive pour ν *concernant le paramètre* δ *du modèle.*

À présent, apportons quelques précisions aux preuves des lemmes 2.4.10, 2.4.11, 2.4.13 et 2.4.14. La différence avec ce qui précède ne concerne que la majoration des cardinaux des ensembles. Focalisons-nous sur la preuve du lemme 2.4.13 (nous préciserons les différences pour les autres).

Preuve du lemme 2.4.13. Commençons par identifier la matrice de corrélation intervenant dans le calcul de $\mathbb{P}_{07}$. Nous avons, de par la définition de Λ_p^0 et Ω_7 :

$$\Sigma = \begin{pmatrix} 1 & 0 & \varepsilon_n & \times \\ 0 & 1 & \times & 0 \\ \varepsilon_n & \times & 1 & 0 \\ \times & 0 & 0 & 1 \end{pmatrix}$$

De plus, pour $\alpha = (i,j) \in \Lambda_p^0$ et $\beta = (u,v) \in \Omega_7$, nous avons les comparaisons :

$$v - i = \underbrace{v-j}_{>\tau+K} + \underbrace{j-i}_{>\tau+K} > 2(\tau+K) \text{ et aussi } j - u = \underbrace{j-i}_{>\tau+K} + \underbrace{i-u}_{>0} > \tau + K$$

D'où :

$$\Sigma = \begin{pmatrix} 1 & 0 & \varepsilon_n & 0 \\ 0 & 1 & 0 & 0 \\ \varepsilon_n & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

En particulier, cette forme de matrice de corrélation correspond à celle du lemme A.2.1. On peut donc donner un ordre de grandeur pour la probabilité $\mathbb{P}_{07}$, pour tout $\omega > 0$:

$$\mathbb{P}_{07} = \mathcal{O}\left(p^{-4+\omega}\right)$$

On a alors, puisque $|\Omega_7| \leqslant Kp$:

$$|\Lambda_p^0|.|\Omega_7|\mathbb{P}_{07} \leqslant pK|\Lambda_p^0|.\mathbb{P}_{07}$$

Et, en utilisant l'équivalent de $|\Lambda_p^0|$ pour $n \to +\infty$:

$$pK|\Lambda_p^0|.\mathbb{P}_{07} \sim \frac{1}{2}Kp^3\mathbb{P}_{07}$$

L'ordre de grandeur de $\mathbb{P}_{07}$ nous indique alors, pour tout $\omega > 0$ et pour $n \to +\infty$:

$$Kp^3\mathbb{P}_{07} = \mathcal{O}\left(p^{\nu+3-4+\omega}\right) = \mathcal{O}\left(p^{\nu-1+\omega}\right) \tag{A.14}$$

Or, puisque par définition, $\nu < 1$, on a bien la convergence vers 0 de cette quantité.

Remarques 14. *Pour les lemmes 2.4.10, 2.4.11 et 2.4.14, nous avons les mêmes matrices de corrélations (à une permutation près dans les lignes). Nous avons donc le même ordre de grandeur pour chacune des probabilités, à savoir* $\mathbb{P}_{04}$, $\mathbb{P}_{05}$, $\mathbb{P}_{08}$. *Puis les majorations des cardinaux qui interviennent sont soit identiques, soit plus faibles. Le cas que nous avons démontré ici plus en détail englobe donc les autres.*

□

Annexe B

Programmation R pour les simulations Monte-Carlo d'un échantillon de τ-cohérence suivant le modèle de covariance par bandes

Listing B.1 – Librairie nécessaire

```
library(gpuR)
```

Listing B.2 – Fonction d'ajustement

```
ajust=function(p,tau,Nb){
  Tb=p%/%Nb
  if(Tb%%tau!=0){
  Tb = Tb + tau - Tb%%tau;
  }
  p = Nb*Tb
  return(list(p=p,tau=tau,Nb=Nb,Tb=Tb))
}
```

Listing B.3 – Fonction de création des observations suivant le modèle de covariance par bandes

```
creat_mat_Obs_new_GPU=function(n,p,tau,K,eps,Nb,ctx_id=1){
  Tb=p%/%Nb
  X=matrix(NA,ncol=p,nrow=n)
  R=runif(1+2*tau,min=-1,max=+1)
  alpha=c(rep(eps,K),R,rep(eps,K)); S=sum(alpha^2);
  Y=matrix(rnorm(n*(p+2*tau+2*K)),ncol=p+2*tau+2*K,nrow=n)
  Y_GPU=gpuMatrix(Y,ctx_id=ctx_id)
  for(j in 1:Nb){
        cat("␣j␣=␣",j, "␣/␣",Nb, "␣\n")
        A=matrix(0,ncol=Tb,nrow=p+2*tau+2*K)
```

```
        for(u in 1:Tb){
          A[((j-1)*Tb+u):((j-1)*Tb+u+2*tau+2*K),u]=alpha/S
        }
        A_GPU=gpuMatrix(A,ctx_id=ctx_id)
        X_GPU=Y_GPU%*%A_GPU
        V=X_GPU[]
        X[1:n,((j-1)*Tb+1):(j*Tb)]=V
  }
  return(X)
}
```

Listing B.4 – Fonction de calcul de la τ-cohérence pour le modèle de covariance par bandes

```
tau_coher_GPU_final=function(n,p,K,tau,eps,Nb,indic=TRUE,ctx_id=1){
  cat( "␣\n\n␣MISE␣EN␣PLACE␣ET␣CREATION␣DES␣OBSERVATIONS␣:␣\n\n")
  cat("Reajustement␣des␣parametres␣pour␣un␣decoupage␣plus␣facile␣")
  res = ajust(p,tau,Nb);
  cat('␣TERMINE\n')
  p = res$p ; tau = res$tau ; Nb = res$Nb ; Tb = res$Tb ;
  cat('Creation␣des␣observations␣')
  X = creat_mat_Obs_new_GPU(n,p,tau,K,eps,Nb,ctx_id=1)
  cat("␣TERMINE\n")
  cat('Decoupage␣et␣sauvegarde␣de␣la␣matrice␣en␣paquet␣binaire␣')
  for( i in 1:Nb){
        name=paste("Obs_",i,".dat",sep="")
        C = X[,((i-1)*Tb+1):(i*Tb)] ; C = as.vector(C) ;
        writeBin(object = C, con = name)
  }
  cat('␣TERMINE\n\n')
  m = 0
  tau=2*tau+1
  cat('␣CALCUL␣DE␣LA␣TAU␣-␣COHERENCE␣:\n\n')
  for( i in 1:Nb){
    if(indic==TRUE){
          cat("␣i␣=␣",i,"␣/␣",Nb,"\n",sep="")
        }
    cat("Recuperation␣des␣donnees␣binaires␣")
        name=paste("Obs_",i,".dat",sep="")
        Xdiag=readBin(con = name, "double",n = n*Tb)
        Xdiag=matrix(Xdiag,ncol=Tb,byrow=FALSE)
        cat("␣TERMINE\n")
        cat("Envoie␣de␣la␣matrice␣de␣donnee␣en␣GPU␣")
        Xdiag_GPU=gpuMatrix(Xdiag,ctx_id=ctx_id)
        cat("␣TERMINE\n")
        cat('Calcul␣de␣la␣covariance␣du␣bloc␣diagonal␣(GPU)␣')
        C_GPU=cov(Xdiag_GPU)
```

```
34          cat('␣TERMINE\n')
35          cat("Recuperation␣de␣la␣covariance␣en␣CPU␣")
36          Cr=C_GPU[]
37          cat("␣TERMINE\n")
38          cat('Renormalisation␣du␣bloc␣diagonal␣')
39          D=diag(Cr); D=sqrt(D); D=D^(-1);
40          D1=D%*%t(D);
41          Cr=C_GPU[]*D1
42          cat('␣TERMINE\n')
43          cat('Boucle␣pour␣le␣calcul␣de␣la␣tau-coherence␣')
44          for (u in 1:dim(Cr)[1]){
45            for (v in 1:dim(Cr)[1]){
46              if (abs(v-u)<tau){
47                Cr[u,v]=0
48                  }
49            }
50          }
51          cat('␣TERMINE\n')
52          m = max(abs(Cr),m);
53          if( i != Nb ){
54            for (j in (i+1):Nb){
55              if(indic==TRUE){
56                    cat("␣----␣j␣=␣",j,"␣/␣",Nb,"\n")
57                  }
58                  if (j == (i+1)){
59                    cat("Recuperation␣des␣donnees␣binaires␣")
60                    name=paste("Obs_",j,".dat",sep="")
61                    Xbis=readBin(con = name, "double",n = n*Tb)
62                    Xbis=matrix(Xbis,ncol=Tb,byrow=FALSE)
63                    cat("␣TERMINE\n")
64                    cat("Envoie␣de␣la␣matrice␣de␣donnee␣en␣GPU␣")
65                    Xbis_GPU=gpuMatrix(Xbis,ctx_id=ctx_id)
66                    cat("␣TERMINE\n")
67                    cat('Calcul␣de␣la␣covariance␣du␣bloc␣diagonal␣(GPU)␣')
68                    C_GPU=cov(Xdiag_GPU,Xbis_GPU)
69                    cat('␣TERMINE\n')
70                    cat("Calculs␣pour␣renormalisation␣")
71                    U=rep(NA,ncol(Xbis))
72                    for (i in 1:ncol(Xbis)){
73                      U[i]=sd(Xbis[,i])
74                    }
75                    DXbis=(U)^(-1);
76                    DXbis=D%*%t(DXbis)
77                    cat("␣TERMINE\n")
78                    cat("Recuperation␣covariance␣CPU␣+␣Renormalisation␣")
79                    C=C_GPU[]*DXbis
```

```
                    cat("␣TERMINE\n")
                    cat("Boucle␣pour␣calcul␣de␣la␣tau␣-␣coherence␣")
                    for( v in 1:(tau-1) ){
                      for( u in (Tb-tau+v+1):Tb ){
                        C[u,v]<-0
                      }
                    }
                    cat( "␣TERMINE\n")
                    m = max(abs(C),m);
                  }
                  else{
                    cat("Recuperation␣des␣donnees␣binaires␣")
                    name=paste("Obs_",j,".dat",sep="")
                    Xbis=readBin(con = name, "double",n = n*Tb)
                    Xbis=matrix(Xbis,ncol=Tb,byrow=FALSE)
                    cat("␣TERMINE\n")
                    cat("Envoie␣de␣la␣matrice␣de␣donnee␣en␣GPU␣")
                    Xbis_GPU=gpuMatrix(Xbis,ctx_id=ctx_id)
                    cat("␣TERMINE\n")
                    cat('Calcul␣de␣la␣covariance␣du␣bloc␣diagonal␣(GPU)␣')
                    C_GPU=cov(Xdiag_GPU,Xbis_GPU)
                    cat('␣TERMINE\n')
                    cat("Calculs␣pour␣renormalisation␣")
                    U=rep(NA,ncol(Xbis))
                    for (i in 1:ncol(Xbis)){
                      U[i]=sd(Xbis[,i])
                    }
                    DXbis=(U)^(-1);
                    DXbis=D%*%t(DXbis)
                    cat("␣TERMINE\n")
                    cat("Recuperation␣covariance␣CPU␣+␣Renormalisation␣")
                    C=C_GPU[]*DXbis
                    cat("␣TERMINE\n")
                    m = max(abs(C),m);
                  }
              }
          }
    }
    cat('\n\nFIN␣DE␣SIMULATION\n\n')
    return(list(tau_coh=m,n=n,p=p,tau=tau,K=K,eps=eps,Nb=Nb,Tb=Tb))
}
```

Le code suivant fournit un exemple de simulation d'un échantillon de τ-cohérence :

Listing B.5 – Exemple de code de simulation pour l'étude numérique de la distribution asymptotique

de $L_{n,\tau}$

```
set.seed(753)
n=5000;
p=floor(exp(n^(1/3.5)));
tau=50*floor(log(p));
K=100*floor(n^(1/10)*log(p));
eps=0.1*sqrt(log(p)/n);
Nb=15;
ctx_id=1
R=200;
taucoh_regime1_5000=rep(NA,R)

for(i in 1:R){
  cat('
␣␣␣␣␣␣######################################################################
␣␣␣␣␣␣#####␣␣Iteration␣i␣=',i,'␣/␣',R,'␣|␣n␣=␣',n,'␣|␣',date(),'
␣␣␣␣␣␣######################################################################\n')
  res=tau_coher_GPU_final(n,p,K,tau,eps,Nb,indic=TRUE, ctx_id=1)
  taucoh_regime1_5000[i]=res$tau_coh
  n_regime1_5000=res$n
  p_regime1_5000=res$p
  tau_regime1_5000=res$tau
  K_regime1_5000=res$K
  eps_regime1_5000=res$eps
  Tb_regime1_5000=res$Tb
  save(taucoh_regime1_5000,n_regime1_5000,p_regime1_5000,
                tau_regime1_5000, K_regime1_5000,eps_regime1_5000,
                Tb_regime1_5000, file="echantillon_regime1_5000.RData")
}
```

Bibliographie

[AGG89] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations : the Chen-Stein method. *Ann. Probab.*, 17(1) :9–25, 1989.

[AGZ10] G.W. Anderson, A. Guionnet, and O. Zeitouni. An introduction to random matrices. In *Cambridge Studies in Advanced Mathematics*, volume 118. Cambridge University Press, Cambridge, 2010.

[And91] Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3) :817–858, 1991.

[BC12] Charles Bordenave and Djalil Chafaï. Around the circular law. *Probab. Surv.*, 9 :1–89, 2012.

[BG16] Cristina Butucea and Ghislaine Gayraud. Sharp detection of smooth signals in a high-dimensional sparse matrix with indirect observations. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4) :1564 – 1591, 2016.

[BGLL15] Anna Bonnet, Elisabeth Gassiat, and Céline Lévy-Leduc. Heritability estimation in high dimensional linear mixed models https ://arxiv.org/abs/1404.339, 2015.

[BJYZ09] Zhidong Bai, Dandan Jiang, Jian-Feng Yao, and Shurong Zheng. Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Stat.*, 37(6B) :3822–3840, 2009.

[BRT09] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.*, 37(4) :1705–1732, 2009.

[BS96] Zhidong Bai and Hewa Saranadasa. Effect of high dimension : by an example of a two sample problem. *Stat. Sin.*, 6(2) :311–329, 1996.

[BS10] Zhidong Bai and Jack W. Silverstein. *Spectral analysis of large dimensional random matrices. 2nd ed.* Dordrecht : Springer, 2nd ed. edition, 2010.

[BSQ18] Vincent Brault, Adeline Samson, and Jean-Charles Quinton. Modélisation statistique pour détecter des séquences vidéos similaires : application aux véhicules autonomes. In *Congrès de la Société Mathématique de France - SMF 2018*, Lille, France, June 2018. Société Mathématique de France.

[CCZ16] Tianxi Cai, T. Tony Cai, and Anru Zhang. Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, 111(514) :621–633, 2016. PMID : 28042188.

[CDOR09] Dar-Jen Chang, Ahmed H. Desoky, Ming Ouyang, and Eric C. Rouchka. Compute pairwise manhattan distance and pearson correlation coefficient of data points with gpu. In *2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing*, pages 501–506, 2009.

[CG15] T. Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression : Minimax rates and adaptivity https ://arxiv.org/abs/1506.05539, 2015.

[Che75] Louis H. Y. Chen. Poisson approximation for dependent trials. *Ann. Probab.*, 3(3) :534–545, 06 1975.

[CJ11a] T. Tony Cai and Tiefeng Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.*, 39(3) :1496–1525, 2011.

[CJ11b] T. Tony Cai and Tiefeng Jiang. Supplement to "limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices". *DOI :10.1214/11-AOS879SUPP*, 2011.

[CJ12] T. Tony Cai and Tiefeng Jiang. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *J. Multivariate Anal.*, 107 :24–39, 2012.

[Cou07] Olivier Couronné. Poisson approximation for large clusters in the supercritical fk model https ://arxiv.org/abs/0705.3781, 2007.

[CRT06a] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Robust uncertainty principles : exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2) :489–509, 2006.

[CRT06b] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.*, 59(8) :1207–1223, 2006.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12) :4203–4215, 2005.

[CT07] Emmanuel Candès and Terence Tao. The Dantzig selector : statistical estimation when p is much larger than n. (With discussions and rejoinder). *Ann. Stat.*, 35(6) :2313–2404, 2007.

[CWX10a] T. Tony Cai, Lie Wang, and Guangwu Xu. Shifting inequality and recovery of sparse signals. *IEEE Trans. Signal Process.*, 58(3) :1300–1308, 2010.

[CWX10b] Tony Tony Cai, Lie Wang, and Guangwu Xu. Stable recovery of sparse signals and an oracle inequality. *IEEE Trans. Inf. Theory*, 56(7) :3516–3522, 2010.

[CZ16a] T. Tony Cai and Anru Zhang. Inference for high-dimensional differential correlation matrices. *J. Multivariate Anal.*, 143 :107–126, 2016.

[CZ16b] T. Tony Cai and Anru Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data https ://arxiv.org/abs/1605.04358, 2016.

[CZZ10] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.*, 38(4) :2118–2144, 2010.

[DH01] David L. Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory*, 47(7) :2845–2862, 2001.

[Don06] D.L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52 :1289–1306, 2006.

[ES18] Taban Eslami and Fahad Saeed. Fast-gpu-pcc : A gpu-based technique to compute pairwise pearson's correlation coefficients for time series data—fmri study. *High-Throughput*, 7(2), 2018.

[For10] P.J. Forrester. Log–gases and random matrices. In *London Mathematical Society Monographs Series*, volume 34. Princeton University Press, Princeton, 2010.

[Fuc04] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Theory*, 50(6) :1341–1344, 2004.

[HM18] Johannes Heiny and Thomas Mikosch. Almost sure convergence of the largest and smallest eigenvalues of high-dimensional sample correlation matrices. *Stochastic Processes Appl.*, 128(8) :2779–2815, 2018.

[Jia04a] Tiefeng Jiang. The asymptotic distributions of the largest entries of sample correlation matrices. *Ann. Appl. Probab.*, 14(2) :865–880, 2004.

[Jia04b] Tiefeng Jiang. The limiting distributions of eigenvalues of sample correlation matrices. *Sankhyā*, 66(1) :35–48, 2004.

[Joh01] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*, 29(2) :295–327, 2001.

[Joh08] Iain M. Johnstone. Multivariate analysis and Jacobi ensembles : largest eigenvalue, Tracy-Widom limits and rates of convergence. *Ann. Stat.*, 36(6) :2638–2716, 2008.

[KUrNT11] Ekasit Kijsipongse, Suriya U-ruekolan, Chumpol Ngamphiw, and Sissades Tongsima. Efficient large pearson correlation matrix computing using hybrid mpi/cuda. In *2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 237–241, 2011.

[LDFF15] Mario Levorato, Lúcia Drummond, Yuri Frota, and Rosa Figueiredo. A GPU-accelerated local search algorithm for the Correlation Clustering problem. In *Proceedings of the Brazilian Symposium on Operations Research*, Porto de Galinhas, PE, Brazil, August 2015. SOBRAPO - Brazilian Society of Operations Research.

[LLR10] Deli Li, Wei-Dong Liu, and Andrew Rosalsky. Necessary and sufficient conditions for the asymptotic distribution of the largest entry of a sample correlation matrix. *Probab. Theory Relat. Fields*, 148(1-2) :5–35, 2010.

[LLS08] Wei-Dong Liu, Zhengyan Lin, and Qi-Man Shao. The asymptotic distribution and Berry-Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *Ann. Appl. Probab.*, 18(6) :2337–2366, 2008.

[LQR12] Deli Li, Yongcheng Qi, and Andrew Rosalsky. On Jiang's asymptotic distribution of the largest entry of a sample correlation matrix. *J. Multivariate Anal.*, 111 :256–270, 2012.

[LR06] Deli Li and Andrew Rosalsky. Some strong limit theorems for the largest entries of sample correlation matrices. *Ann. Appl. Probab.*, 16(1) :423–447, 2006.

[Meh04] M.L. Mehta. Random matrices. In *Pure and Applied Mathematics (Amsterdam)*, volume 142. Elsevier/Academic Press, Amsterdam, 2004.

[MP00] G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics, New York, 2000.

[Pé09] Sandrine Péché. Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probab. Theory Relat. Fields*, 143(3-4) :481–516, 2009.

[PDLLR19] Marie Perrot-Dockès, Céline Lévy-Leduc, and Loïc Rajjou. Estimation of large block structured covariance matrices : Application to "multi-omic" approaches to study seed quality https ://arxiv.org/abs/1806.10093, 2019.

[Pec12] Giovanni Peccati. The Chen-Stein method for Poisson functionals, https ://hal.archives-ouvertes.fr/hal-00654235. 18 pages ; some small typos, in particular in the proof of Theorem 5.1, have been corrected., April 2012.

[R C20] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[Ste72] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2 : Probability Theory*, pages 583–602, Berkeley, Calif., 1972. University of California Press.

[SZ14] Qi-Man Shao and Wen-Xin Zhou. Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *Ann. Probab.*, 42(2) :623–648, 03 2014.

[TSCD08] Narjiss Touyar, Sophie S. Schbath, Dominique Cellier, and Hélène Dauchel. Poisson approximation for the number of repeats in a stationary Markov chain. *Journal of Applied Probability*, 45(2) :440–455, 2008.

[VA08] Nicolas Vergne and Miguel Abadi. Poisson approximation for search of rare words in DNA sequences. *Alea : Estudos Neolatinos*, 4 :223 – 244, 2008.

[Wig58] Eugene P. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. Math. (2)*, 67 :325–327, 1958.

[Zho07] Wang Zhou. Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Trans. Am. Math. Soc.*, 359(11) :5345–5363, 2007.

[ZLLT20] Wencan Zhu, Céline Lévy-Leduc, and Nils Ternès. A variable selection approach for highly correlated predictors in high-dimensional genomic data https ://arxiv.org/abs/2007.10768, 2020.

Maxime BOUCHER

Cohérence de grandes matrices aléatoires. Théorèmes limites et applications.

Résumé :
Cette thèse concerne l'étude de la τ-cohérence d'une matrice d'observations aléatoires de grande taille $(n \times p)$ où $p >> n$ et avec p le nombre de variables observées sur n individus. La τ-cohérence est alors définie comme étant le maximum, en valeur absolue, des coefficients de la matrice de corrélation empirique associée, en dehors d'une bande centrale de largeur τ. Le premier chapitre est consacré à la présentation de la méthode de Chen-Stein qui permet l'approximation d'événements faiblement dépendants par une loi de Poisson et à la présentation des travaux de T. Cai et T. Jiang concernant la cohérence. Les deuxième et troisième chapitres sont consacrés à l'étude du comportement asymptotique de la τ-cohérence dans le cas où les observations proviennent d'un modèle gaussien et où la matrice de covariance possède une structure par bandes (chapitre 2) ou par blocs (chapitre 3). Dans le chapitre 4, nous présentons une méthode de simulation par réplications Monte-Carlo pour étudier numériquement la distribution asymptotique de la τ-cohérence. Nous utilisons des stratégies de découpage de nos matrices et des techniques HPC telles que le calcul en GPU pour, d'une part pouvoir calculer des corrélations sur des matrices trop grandes pour être stockées, et d'autre part réduire le temps de calcul. Nous présentons en annexe de ce manuscrit des éléments de preuves supplémentaires.

Mots clés : cohérence, matrices de grandes dimensions, corrélations, méthode de Chen-Stein, gaussien, programmation GPGPU, matrices aléatoires, parcimonies.

Coherence of high-dimensional random matrices. Limit theorems and applications.

Abstract :
This thesis focuses on the study of the τ-coherence of an high-dimensional $(n \times p)$-observation matrix with $p >> n$, where n is the number of individuals and p the number of variables. The τ-coherence is defined as the largest magnitude of the entries of the empirical correlation matrix outside a central band (with a bandwith τ). The first chapter is devoted to the presentation of the Chen-Stein method, which is an approximation of weakly dependent events by a Poisson distribution, and to some bibliography concerning coherence. The second and third chapter focus on the limiting behaviour of τ-coherence in a case where the observations are assumed to be Gaussian with bandwise (resp. blockwise) covariance in Chapter 2 (resp. Chapter 3). In the last chapter, we propose a Monte-Carlo simulation procedure allowing us to study numerically the limiting distribution of the τ-coherence for large (Big Data) matrices. We use a splitting strategy of our matrices and HPC method such as GPGPU computation in order to, from one side, being able to compute correlation matrices even if they are too large to be loaded in a computer, and on the other side, to reduce computation time. Finally the appendix is devoted to some technical results.

Keywords : coherence, high-dimensional matrices, correlations, Chen-Stein method, Gaussian, GPGPU, random matrices, sparsity.

UNIVERSITE D'ORLEANS

Institut Denis Poisson
Rue de Chartres B.P. 6759
45067 ORLEANS CEDEX 2

INSTITUT ORLÉANS · TOURS Mathématiques & Physique Théorique DENIS POISSON