Cohérence de grandes matrices aléatoires

Distribution asymptotique dans un cadre gaussien dépendant par bandes

Maxime Boucher [1*] & Didier Chauveau [2*] & Marguerite Zani [3*]

[1] *maxime.boucher@univ-orleans.fr,* [2] *didier.chauveau@univ-orleans.fr,*
[3] *marguerite.zani@univ-orleans.fr*

[*] *Institut Denis Poisson, Université d'Orléans, Collegium Sciences et Techniques, bâtiment de mathématiques, Rue de Chartres B.P.6759, 45067 Orléans CEDEX 2, FRANCE.*

Résumé. Dans cet exposé, on introduit la τ-cohérence, notée $L_{n,\tau}$, d'une matrice aléatoire X_n de taille $n \times p$, avec p très grand devant n, définie comme étant le maximum en valeur absolue du coefficient de corrélation empirique de Pearson calculé sur des paires de colonnes de X_n séparées par un entier τ. On s'intéresse au cas où les ligne de X_n sont des observations indépendantes de loi normale dans $\mathbb{R}^p$, centrée et de matrice de covariance réduite Σ. En particulier, on suppose que Σ est définie par bande : une bande centrale de corrélation, une bande de transition asymptotiquement nulle et une partie extérieure d'indépendance. On montre, en utilisant la méthode de Chen-Stein, que sous certaines hypothèses, la τ-cohérence, correctement corrigée, admet une distribution asymptotique de Gumbel. On visualise cette convergence via des simulations Monte-Carlo utilisant le calcul sur GPU et des découpages pour palier le problème des tailles de matrices et des temps de calculs.

Mots-clés. Matrice aléatoire, cohérence, corrélation de Pearson, méthode de Chen-Stein, asymptotique, grandes dimensions, parcimonies, simulation GPGPU.

Abstract. In this presentation, we introduce the quantity $L_{n,\tau}$, called τ-coherence of a $n \times p$ random matrix where p is greater than n. It is defined to be the largest magnitude of the Pearson correlation coefficients between pair of columns of the random matrix separated by an integer τ. In our study, lines of the observation matrix are i.i.d observations of a p-dimensional centered and reduced Gaussian vector with Σ as correlation matrix. We suppose that Σ is divided into three bands: a central band with correlation coefficients, a transition band with asymptotically null coefficients, and an outside part with null coefficients. Using the Chen-Stein method, and under sufficient hypotheses, we can show that the τ-coherence, with correction terms, has an asymptotic Gumbel behaviour. We construct Monte-Carlo simulations using splitting technics and GPGPU to handle large matrices and computing times.

Keywords. Random matrices, coherence, Chen-Stein method, Pearson correlation, asymptotic, sparsity, high dimension, GPGPU computing.

1 Étude théorique de la τ-cohérence

1.1 Modèle

On se propose d'étudier la distribution asymptotique de la τ-cohérence d'une matrice aléatoire. On se donne une matrice $\mathbf{X}$ d'observations de taille $n \times p$ où n et p seront très grands avec $n << p$. Chaque ligne de $\mathbf{X}$ est donc une observation de dimension p, et les lignes seront indépendantes et identiquement distribuées. On définit la τ-cohérence comme étant la valeur :

$$L_{n,\tau} := \max_{1 \leq i < j \leq p, |i-j| \geq \tau} |\rho_{ij}|,$$

où ρ_{ij} est le coefficient de corrélation empirique de Pearson entre les colonnes i et j de $\mathbf{X}$ et $\tau \in \mathbb{N}^*$. En particulier, on se place dans le cadre gaussien : c'est-à-dire que l'on suppose que chaque ligne est une observation d'un p-vecteur gaussien. On a donc le modèle suivant :

$$\left(X_k^1, X_k^2, \ldots, X_k^p\right)_{1 \leq k \leq n} \overset{i.i.d}{\sim} \mathcal{N}_p\left(0, \Sigma\right),$$

où Σ est la matrice de covariance réduite que l'on suppose définie en bande comme suit, avec τ et K deux entiers :

$$\Sigma_{k,j} = \begin{cases} r_{kj} & \text{if} \quad |k-j| < \tau \\ \epsilon_n & \text{if} \quad \tau \leq |k-j| \leq \tau + K \\ 0 & \text{if} \quad \tau + K < |k-j| \end{cases} .$$

Autrement dit, on suppose que deux composantes proches du vecteur (en terme d'indice) sont corrélées, alors que si elles sont suffisamment éloignés elles sont indépendantes. Dans notre modèle, on généralise celui de [Cai et Jiang, 2011] en ajoutant une bande de transition avec des coefficients $\epsilon_n \underset{n \to +\infty}{\longrightarrow} 0$. Cela permet de considérer une décorrélation progressive entre deux composantes du vecteur, au fur et à mesure qu'elles s'éloignent l'une de l'autre. La distribution asymptotique de L_n a été décrite dans le cas où toutes les observations sont des entrées gaussiennes indépendantes dans [Cai et Jiang, 2012]. On peut également citer [Shao et Zhou, 2014] où la convergence en loi de $L_{n,\tau}$ est étudiée sans hypothèse de normalité.

1.2 Résultat

Commençons par définir

$$\forall \delta \in]0,1[, \; \Gamma_{p,\delta} = \{k \in \{1, \ldots, p\} \text{ tel que } |r_{kj}| > 1 - \delta \text{ pour } j \in \{1, \ldots, p\} \text{ et } k \neq j\}.$$

Nous avons le résultat suivant, présenté dans [Boucher *et al.*, 2021] :

Théorème 1.2.1 *Soit n un entier non nul et $p = p_n$ une suite d'entiers tels que $p \underset{n\to+\infty}{\longrightarrow} +\infty$. On se donne une suite de réels $(\epsilon_n)_{n\geq 1} \in]-1,1[$. Supposons les conditions suivantes :*

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

Hyp 2 : $\tau = o(p_n^t)$ *quand* $n \to +\infty$ *pour tout* $t > 0$.

Hyp 3 : $\exists \delta \in]0,1[$ *tel que* $|\Gamma_{p,\delta}| = o(p_n)$.

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

Hyp 5 : $K = K(n) = \mathcal{O}(p_n^\nu)$ *pour* $0 < \nu < c(\gamma,\delta) < 1$, *où* $c(\gamma,\delta)$ *est une constante explicite dépendant uniquement de* γ *et* δ

Sous ces conditions, on peut montrer que :

$$nL_{n,\tau}^2 - 4\log(p_n) + \log(\log(p_n)) \underset{n\to+\infty}{\xrightarrow{\mathcal{L}}} Z \tag{1}$$

où Z est une variable aléatoire qui admet pour fonction de répartition $F(y) = exp\left(-\frac{1}{\sqrt{8\pi}}e^{-\frac{y}{2}}\right)$ pour tout $y \in \mathbb{R}$.

En regardant la définition de la matrice Σ, combinée à l'hypothèse 3, on voit que nos composantes peuvent être corrélées si elles sont proches. En revanche, le nombre de composantes fortement corrélées est faible. De plus, avec les ϵ_n, la corrélation de transition est asymptotique nulle. Cela signifie que pour des n suffisamment grands, tous les coefficients ϵ_n passent sous le seuil $1-\delta$. Donc on ne rajoute pas de fortes corrélations pour n grand. Cependant, on gagne en généralité par rapport au modèle de [Cai et Jiang, 2011] puisque notre bande contenant les ϵ_n peut croître vers $+\infty$ plus rapidement que τ.

1.3 Idée de la preuve

Pour étudier la distribution de la τ-cohérence, il est judicieux de choisir une nouvelle variable aléatoire. Ainsi, comme dans [Cai et Jiang, 2011], on considère une variable aléatoire intermédiaire $\tilde{V}_{n,\tau} := \underset{|i-j|\geq\tau}{max}\left|\sum_{k=1}^{n} X_k^i X_k^j\right|$. L'idée étant de décrire le comportement asymptotique de cette dernière pour ensuite revenir à la τ-cohérence grâce au résultat suivant :

Lemme 1.3.1 *Soit τ et K deux entiers. Soit $\mathbf{X}$ une matrice d'observation de taille (n,p) où $(X^1, X^2, \ldots, X^p)$ sont les p colonnes de $\mathbb{R}^n$. Soit $L_{n,\tau}$, la τ-cohérence de $\mathbf{X}$ et $\tilde{V}_{n,\tau} = \max_{1\leq k<j\leq p, |k-j|\geq\tau} |{}^tX^kX^j|$. On suppose que $\log(p_n) = o(n^{\frac{1}{3}})$ quand $n \to +\infty$. Alors,*

$$\frac{n^2L_{n,\tau}^2 - \tilde{V}_{n,\tau}^2}{n} \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} 0 \tag{2}$$

Pour décrire la loi asymptotique de $\tilde{V}_{n,\tau}$, on utilise la méthode de Chen-Stein dont on rappelle l'énoncé (on renvoie également vers [Arratia *et al.*, 1989]):

Lemme 1.3.2 (Chen-Stein method) *Soit I un ensemble d'indices. Soit $\alpha \in I$ et B_α un sous-ensemble de I (i.e. pour tout α, $B_\alpha \subset I$). Soit $(\eta_\alpha)_{\alpha \in I}$ des variables aléatoires. Pour un $t \in \mathbb{R}$ fixé, on définit $\lambda := \sum_{\alpha \in I} \mathbb{P}(\eta_\alpha > t)$. Alors,*

$$\left| \mathbb{P}\left(\max_{\alpha \in I} (\eta_\alpha) \leq t \right) - e^{-\lambda} \right| \leq \min\left(1, \frac{1}{\lambda}\right) . (b_1 + b_2 + b_3) \tag{3}$$

où

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

Nous appliquons ce résultat pour $\eta_\alpha = \rho_{ij}$ et pour $I = \{(i,j) \in [\![1,p]\!]^2 : |i-j| \geq \tau\}$. Cela fait donc apparaître la variable $\tilde{V}_{n,\tau}$. Ensuite, il ne nous reste plus qu'à calculer λ_n dont la limite nous donnera la fonction de répartition asymptotique et pour finir contrôler les $(b_i)_{i=1,2,3}$ en montrant qu'ils sont bien tous trois asymptotiquement nuls. La difficulté principale est de gérer le coefficient b_2 qui prend en compte la dépendance entre les colonnes de $\mathbf{X}$.

Notre approche consiste à considérer une nouvelle variable aléatoire $\tilde{V}'_{n,\tau}$ qui est le maximum de la même quantité mais sur un ensemble plus petit que I en retirant de l'étude toutes les quantités de corrélation trop forte. On peut montrer que les deux variables sont bien équivalentes en probabilité pour enfin appliquer la méthode de Chen-Stein à cette dernière.

2 Simulation de la τ-cohérence par utilisation du GPGPU

Nous étudions ici empiriquement la distribution asymptotique de $L_{n,\tau}$ dans le cas où la matrice d'observation $\mathbf{X}$ est grande avec $n << p$. D'un point de vue pratique, lorsque p est suffisamment grand, la matrice de corrélation issue de $\mathbf{X}$ ne tient plus dans la mémoire vive d'un ordinateur classique (on considère par exemple le cas où pour $n = 4000$, on a $p = 44000$ et donc la matrice de corrélation de taille $p \times p$ nécessite plus de $10Go$ de mémoire). De plus, quand n et p augmentent, les temps de simulations augmentent. Ces deux aspects obligent à repenser les techniques de simulations.

On procède alors de la manière suivante : on découpe la matrice $\mathbf{X}$ en paquets de colonnes, dont la taille est fixé par l'utilisateur. Ensuite, on calcule toutes les matrices de corrélations issues de ces paquets de colonnes, ainsi que les matrices de corrélations calculées sur deux paquets de colonnes distincts. Avec ce découpage, on peut reconstruire la matrice de corrélation cible de dimension $p \times p$ en éliminant le problème de mémoire. De plus, nous effectuons nos calculs de matrices de corrélations sur une GPU. Cela nous permet de réduire drastiquement les temps de calculs. Ces deux techniques nous permettent d'obtenir des échantillons de Monte-Carlo pour la τ-cohérence en un temps raisonnable pour des n et p grands (par exemple, pour $n = 4000$ on a $p = 44000$). On peut alors visualiser la convergence en loi apparaissant dans le théorème présenté. On montre également les gains de temps en fonction de p.

References

[Arratia *et al.*, 1989] Arratia, R., Goldstein, L. et Gordon, L. (1989). Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25.

[Boucher *et al.*, 2021] Boucher, M., Chauveau, D. et Zani, M. (2021). Coherence of high-dimensional random matrices in a gaussian case : application of the chen-stein method.

[Cai et Jiang, 2011] Cai, T. T. et Jiang, T. (2011). 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.

[Cai et Jiang, 2012] Cai, T. T. et Jiang, T. (2012). Phase transition in limiting distributions of coherence of high-dimensional random matrices. *J. Multivariate Anal.*, 107:24–39.

[Shao et Zhou, 2014] Shao, Q.-M. et Zhou, W.-X. (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *Ann. Probab.*, 42(2):623–648.