Convergence en loi des U-statistiques sur une matrice échangeable ligne-colonne

Tâm Le Minh[1]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA-Paris, 75005, Paris, France, tam.le-minh@inrae.fr*

Résumé. Les U-statistiques sont utilisées pour estimer les paramètres d'une population en moyennant une fonction d'un sous-ensemble sur tous les sous-ensembles de cette population. Notre travail porte sur une population formée par les valeurs d'une matrice échangeable ligne-colonne. On considère les U-statistiques issues de fonctions sur des quadruplets, c'est-à-dire des sous-matrices de taille 2×2. On démontre un résultat de convergence faible pour ces U-statistiques et on établit un Théorème Central Limite dans le cas où la matrice est dissociée. Les matrices échangeables ligne-colonne sont une représentation naturelle des réseaux bipartites échangeables, nous appliquons donc les résultats à l'inférence statistique pour les réseaux.

Mots-clés. U-statistiques, échangeabilité ligne-colonne, Théorème Central Limite, réseaux bipartites

Abstract. U-statistics are used to estimate a population parameter by averaging a function of a subsample over all the subsamples of the population. In our work, the population we are interested in is formed by the entries of a row-column exchangeable matrix. We consider U-statistics derived from functions of quadruplets, i.e. submatrices of size 2×2. We prove a weak convergence result for these U-statistics in the general case and we establish a Central Limit Theorem when the matrix is also dissociated. Since row-column exchangeable matrices are an actual representation for exchangeable bipartite networks, we apply these results to statistical inference in network analysis.

Keywords. U-statistics, row-column exchangeability, Central Limit Theorem, bipartite network

1 Introduction

1.1 U-statistiques

Les U-statistiques sont une classe de statistiques qui généralisent le concept de moyenne empirique pour des fonctions à plusieurs variables. Par exemple, soit $X_1, X_2, ...$ une série de variables aléatoires indépendantes et identiquement distribuées. Soit h une fonction de r variables. Alors sur un échantillon de taille n, la U-statistique associée à h est

$$U_n = \left[r! \binom{n}{r} \right]^{-1} \sum_{1 \leq i_1 \neq ... \neq i_r \leq n} h(X_{i_1}, ..., X_{i_r}).$$

Si $\mathbb{E}[h(X_1, ..., X_r)] = \theta$, alors U_n est un estimateur sans-biais et de variance réduite de θ sur cet échantillon,: $\mathbb{E}[U_n] = \theta$ et $\mathbb{V}[U_n] \leq \mathbb{V}[h(X_1, ..., X_r)]$. Hoeffding (1948) a montré que la loi de

$\sqrt{n}(U_n - \theta)$ convergeait vers une loi normale centrée de variance connue, ce qui en fait un estimateur asymptotiquement normal. Les U-statistiques peuvent donc être utilisées pour l'estimation et la construction d'intervalles de confiance de paramètres.

Nandi et Sen (1963), puis Eagleson et Weber (1978) étendent ce résultat aux cas où les variables $X_1, X_2, ...$ sont échangeables, c'est-à-dire que pour toute permutation finie σ,

$$(X_1, X_2, ...) \overset{\mathcal{L}}{=} (X_{\sigma(1)}, X_{\sigma(2)}, ...).$$

Ils montrent que les U-statistiques de variables échangeables convergent en loi vers un mélange infini de gaussiennes, puis identifient les cas où ce mélange se réduit à une gaussienne simple, ce qui établit un Théorème Central Limite (TCL).

1.2 Matrices RCE

Certains types de données, tels que les réseaux d'interaction bipartites et les tables de présence-absence s'inscrivent souvent dans un cadre d'échangeabilité plus complexe. En effet, ces données peuvent se présenter sous la forme d'une matrice rectangulaire Y de taille $m \times n$, où les lignes et les colonnes représentent deux types d'entités différentes en interaction, comme m insectes et n plantes dans les réseaux plantes-pollinisateurs, ou m espèces et n sites dans les tableaux de présence-absence. Y_{ij} désigne donc l'interaction entre l'entité i de type 1 et l'entité j de type 2. Si les entités sont échangeables au sein de chaque type, alors la matrice est dite échangeable ligne-colonne (row-column exchangeable, RCE). Elles vérifient la propriété suivante, pour tout couple de permutations finies $\Phi = (\sigma_1, \sigma_2)$,

$$\Phi Y \overset{\mathcal{L}}{=} Y,$$

avec $\Phi Y := (Y_{\sigma_1(i)\sigma_2(j)})_{1 \leq i \leq m, 1 \leq j \leq n}$.

1.3 Cadre d'étude

Nous considérons des fonctions de quadruplets, c'est-à-dire des matrices de taille 2×2 que nous notons

$$Y_{\{i_1,i_2;j_1,j_2\}} := \begin{pmatrix} Y_{i_1 j_1} & Y_{i_1 j_2} \\ Y_{i_2 j_1} & Y_{i_2 j_2} \end{pmatrix}.$$

Sans perte de généralité, nous pouvons considérer que ces fonctions h sont symétriques, on parle alors de noyaux, telles que $h(Y_{\{1,2;1,2\}}) = h(Y_{\{2,1;1,2\}}) = h(Y_{\{1,2;2,1\}}) = h(Y_{\{2,1;2,1\}})$. La U-statistique associée au noyau h sur la matrice Y s'écrit alors

$$U^h_{m,n} = \binom{m}{2}^{-1} \binom{n}{2}^{-1} \sum_{\substack{1 \leq i_1 < i_2 \leq m \\ 1 \leq j_1 < j_2 \leq n}} h(Y_{\{i_1,i_2;j_1,j_2\}}). \tag{1}$$

La structure de symétrie des quadruplets est plus complexe que celles des variables qui sont étudiées dans la littérature des U-statistiques. Nous démontrons donc que les U-statistiques de quadruplets convergent bien en loi vers un mélange de gaussiennes et nous identifions précisément les cas où ce résultat devient un TCL.

2 Résultats

À présent, nous considérons Y une matrice RCE infinie, i.e. pour tout (m, n), la sous-matrice formée par les m premières lignes et les n premières colonnes de Y est RCE suivant la définition du paragraphe 1.2. Afin de déduire un résultat asymptotique, nous construisons une suite de dimensions $m_N \times n_N$ définie par

$$\begin{cases} m_N = 2 + \lfloor c(N+1) \rfloor \\ n_N = 2 + \lfloor (1-c)(N+1) \rfloor \end{cases}$$

La suite de U-statistiques considérée est $U_N^h := U_{m_N,n_N}^h$ définie par la formule (1).

De cette manière, U_N^h est calculée sur un nombre m_N de lignes et un nombre n_N de colonnes croissant à la même vitesse $O(N)$, et $c \in]0,1[$ est un facteur de forme qui rend la matrice rectangulaire.

2.1 Théorème général

Théorème 2.1. *Soient* $\mathcal{F}_N = \sigma\big((U_{kl}^h, k \geq m_N, l \geq n_N)\big)$ *et* $\mathcal{F}_\infty := \bigcap_{N=1}^\infty \mathcal{F}_N$. *On pose* $U_\infty^h := \mathbb{E}[h(Y_{\{1,2;1,2\}})|\mathcal{F}_\infty]$ *et*

$$V = \frac{4}{c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{1,3;3,4\}})\big|\mathcal{F}_\infty\big) + \frac{4}{1-c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{3,4;1,3\}})\big|\mathcal{F}_\infty\big).$$

Si $\mathbb{E}[h(Y_{\{1,2;1,2\}})^2] < \infty$ *et* $V \neq 0$, *alors*

$$\sqrt{N}(U_N^h - U_\infty^h) \xrightarrow[N\to\infty]{\mathcal{D}} W,$$

où W *est une variable aléatoire de fonction caractéristique* $\phi(t) = \mathbb{E}[\exp(-\frac{1}{2}t^2 V)]$ *(mélange infini de gaussiennes).*

Ce théorème découle du résultat de convergence des sommes de différences de martingales inverses d'Eagleson et Weber (1978). Ici, les différences de martingales inverses sont les $Z_{NK} := \sqrt{N}(U_K^h - U_{K+1}^h)$ et leur somme donne bien $\sum_{K=N}^\infty Z_{NK} = \sqrt{N}(U_K^h - U_\infty^h)$. La démonstration comporte trois étapes :

1. d'abord, on montre que Z_{NK} est bien une différence de martingales inverses, i.e. U_N^h est une martingale inverse par rapport à $\mathcal{F}_N$, i.e. $\mathbb{E}[U_N^h|\mathcal{F}_{N+1}] = U_{N+1}^h$ pour tout N,
2. puis, on montre l'existence et l'expression d'une variable aléatoire V ("variance asymptotique") telle que
   $\sum_{K=N}^\infty \mathbb{E}[Z_{NK}^2|\mathcal{F}_{K+1}] \xrightarrow[N\to\infty]{\mathbb{P}} V$,
3. enfin, on vérifie la condition de Lindeberg conditionnelle, i.e. pour tout $\epsilon > 0$,
   $\sum_{K=N}^\infty \mathbb{E}[Z_{NK}^2 \mathbb{1}_{\{|Z_{NK}|>\epsilon\}}|\mathcal{F}_{K+1}] \xrightarrow[N\to\infty]{\mathbb{P}} 0$.

2.2 Un Théorème Central Limite

On remarque que si V est constant, alors on obtient un TCL. On a démontré que c'était le cas si la matrice infinie Y est dissociée, c'est-à-dire que pour tout (m,n), $(Y_{ij})_{1\leq i\leq m,1\leq j\leq n}$ est indépendante de $(Y_{ij})_{m<i<\infty,n<j<\infty}$.

Théorème 2.2. *En plus des hypothèses du Théorème 2.1, si Y est dissociée, alors U_∞^h et V sont constants et*

$$\sqrt{N}(U_N^h - U_\infty^h) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,V),$$

Plus précisément,

1. $U_\infty^h = \mathbb{E}[h(Y_{\{1,2;1,2\}})]$,
2. $V = \frac{4}{c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{1,3;3,4\}})\big) + \frac{4}{1-c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{3,4;1,3\}})\big)$.

De manière plus générale, d'après le théorème de représentation d'Aldous-Hoover (Aldous, 1981), les modèles RCE dissociés correspondent aux modèles qui peuvent s'écrire :

$$Y \overset{\mathcal{L}}{=} Y^*,$$

où pour tout i, j, $Y_{ij}^* = \phi(\xi_i, \eta_j, \zeta_{ij})$ avec ϕ une fonction quelconque et les $\xi_i, \eta_j, \zeta_{ij}$ sont des variables aléatoires i.i.d.. Ces modèles sont ceux qui peuvent être représentés par un graphon (ou $W-$graphe). Cela identifie donc la classe de modèles pour lesquels le Théorème 2.2 s'applique.

3 Application à l'analyse de réseaux

Le TCL est notamment intéressant quand U_N est un estimateur. Les applications les plus directes rentrant dans le cadre des matrices RCE sont les réseaux bipartites. Nous étudions un exemple d'inférence à l'aide de U-statistiques sur des réseaux d'interaction.

3.1 Le modèle WBEDD

Nous considérons le modèle WBEDD (weighted bipartite expected degree distribution), qui est une extension échangeable du modèle à suite de degrés attendus de Chung et Lu (2002) pour les graphes pondérés bipartites. Le modèle EDD suppose que les espérances des poids (nombres de liens) des nœuds sont tirés dans une distribution. Il s'écrit ainsi :

$$\begin{aligned} \xi_i, \eta_j &\overset{iid}{\sim} \mathcal{U}[0,1] \\ Y_{ij} \mid \xi_i, \eta_j &\sim \mathcal{L}(\lambda f(\xi_i) g(\eta_j)), \end{aligned}$$

où f et g sont des fonctions définies sur $[0,1]$ avec $\int f = \int g = 1$ régissant respectivement la distribution des espérances des poids des nœuds en ligne et en colonne, λ est l'espérance de la densité du réseau et $\mathcal{L}(\mu)$ est une loi de probabilité d'espérance μ. On remarque que ce modèle est effectivement RCE.

Le modèle est adapté aux réseaux écologiques bipartites. En effet, pour l'étude des réseaux, les écologues utilisent souvent des modèles nuls, c'est-à-dire des modèles générant des réseaux aléatoires

sous une hypothèse nulle $\mathcal{H}_0$ à tester. Pour réaliser leur test, ils comparent une statistique d'intérêt calculée sur le réseau observé par rapport à sa distribution dans les réseaux générés par les modèles nuls. Parmi les modèles nuls les plus répandus, beaucoup génèrent des réseaux avec une distribution de degrés fixée (souvent celle qui est observée). Le modèle WBEDD ressemble à ces modèles dans la mesure où il génère des réseaux à partir d'hypothèses sur les poids des nœuds, les poids étant analogues aux degrés pour les réseaux pondérés. La différence principale est que l'on ne tire pas directement les poids, mais en fait l'espérance de ces poids dans une distribution, indépendemment pour chaque nœud.

3.2 Un test statistique

La fonction f détermine donc la distribution des poids des nœuds en ligne. Par exemple, pour des réseaux d'interaction écologiques, une fonction f non constante signifie qu'il y a un déséquilibre dans les rôles des espèces en ligne, en particulier entre des espèces généralistes (interagissant avec beaucoup d'espèces) et des espèces spécialistes (interagissant avec peu d'espèces). Pour identifier cet effet dans les réseaux observés, on peut réaliser un test de l'hypothèse $\mathcal{H}_0 : f \equiv 1$ contre $\mathcal{H}_1 : f \not\equiv 1$. Pour cet exemple, on dira que la distribution $\mathcal{L}(\mu)$ du modèle WBEDD est la loi de Poisson $\mathcal{P}(\mu)$, adaptée aux données de comptage.

Sous $\mathcal{H}_0$, $F_2 := \int f^2 = 1$ donc $\lambda^2 F_2 = \lambda^2$. Comme précédemment annoncé, en posant $h_1(Y_{\{i_1,i_2;j_1,j_2\}}) = (Y_{i_1j_1}Y_{i_1j_2} + Y_{i_2j_1}Y_{i_2j_2})/2$, on a $\mathbb{E}[h_1(Y_{\{i_1,i_2;j_1,j_2\}})] = \lambda^2 F_2 = \lambda^2$. L'application du théorème donne pour la U-statistique associée :

$$\sqrt{\frac{N}{V^{h_1}}}(U_N^{h_1} - \lambda^2) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

où $V^{h_1} = \lambda^4 c^{-1}(F_4 - F_2^2) + 4\lambda^4(1-c)^{-1}F_2(G_2 - 1) = 4\lambda^4(1-c)^{-1}(G_2 - 1)$ sous $\mathcal{H}_0$, avec les notations $F_4 := \int f^4$, $G_2 := \int g^2$.

On estime λ^2 par la U-statistique associée à $h_2(Y_{[i_1,i_2;j_1,j_2]}) = (Y_{i_1j_1}^2 Y_{i_2j_2}^2 + Y_{i_1j_2}^2 Y_{i_2j_1}^2)/2$ et $\lambda^2 G_2$ par celle qui est associée à $h_3(Y_{[i_1,i_2;j_1,j_2]}) = \frac{1}{2}(Y_{i_1j_1}Y_{i_2j_1} + Y_{i_1j_2}Y_{i_2j_2})$. En définissant

$$\hat{V}_N^{h_1} := \frac{4}{1-c}(U_N^{h_2})^2 \left[\frac{U_N^{h_3}}{U_N^{h_2}} - 1\right]$$

puis en appliquant Slutsky, on obtient

$$T_N := \sqrt{\frac{N}{\hat{V}_N^{h_1}}}(U_N^{h_1} - U_N^{h_2}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Ce résultat définit des intervalles de confiance asymptotiques pour T_N qui sera la statistique de test. Ces intervalles de confiances sont alors utilisés en guise d'intervalles d'acceptation du test.

Références

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4), 581-598.

Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99,15879–15882.

Eagleson, G. K. and Weber, N. C. (1978). Limit theorems for weakly exchangeable arrays. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 84, 123–130. Cambridge University Press.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 293–325.

Le Minh, T. (2021). U-statistics on bipartite exchangeable networks. *arXiv preprint arXiv:2103.12597.*

Nandi, H. and Sen, P. (1963). On the properties of U-statistics when the observations are not independent: Part two unbiased estimation of the parameters of a finite population. *Calcutta Statistical Association Bulletin*, 12, 124–148.