Weak convergence of U-statistics on a row-column exchangeable matrix

Tâm Le Minh (Université Paris-Saclay, MIA-Paris)

Stéphane Robin, Sophie Donnet, François Massol

13/06/2022

53èmes Journées de Statistique de la Société Française de Statistique

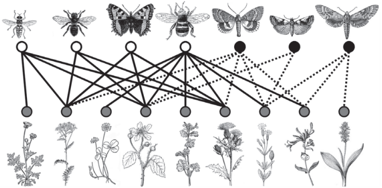
RCE matrices

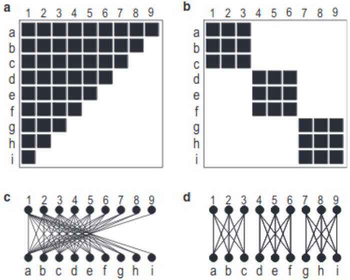
Definition

A matrix Y is row-column exchangeable (RCE) if for any permutations σ_1 and σ_2 of $\mathbb{N}$:

$$Y \overset{\mathcal{D}}{=} (Y_{\sigma_1(i),\sigma_2(j)})_{i\geq 1, j\geq 1}$$

Motivation : exchangeable bipartite networks





$(X_1, X_2, \ldots)$ array of i.i.d. random variables, h a symmetric function

$$U_n^h = \binom{n}{r}^{-1} \sum_{1 \le i_1 < \ldots < i_r \le n} h(X_{i_1}, \ldots, X_{i_r}).$$

Theorem (Hoeffding, 1948)

Let

- $\theta := \mathbb{E}[h(X_1, \ldots, X_r)]$,
- $V := \mathrm{Cov}\big(h(X_1, X_2, \ldots, X_r), h(X_1, X_{r+1}, \ldots, X_{2r-1})\big)$.

If $\mathbb{E}[h(X_1, \ldots, X_r)^2] < \infty$, then

$$\sqrt{n}(U_n^h - \theta) \xrightarrow[n \to \infty]{\mathcal{D}} \mathcal{N}(0, V).$$

U-statistics

Quadruplet kernel :

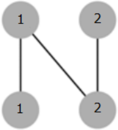
$$h(Y_{\{i_1,i_2;j_1,j_2\}}) = h\left(\begin{bmatrix} Y_{i_1j_1} & Y_{i_1j_2} \\ Y_{i_2j_1} & Y_{i_2j_2} \end{bmatrix}\right)$$

U-statistics on a matrix of size $m \times n$

$$U^h_{m,n} = \left[\binom{m}{2}\binom{n}{2}\right]^{-1} \sum_{i_1<i_2}^{m} \sum_{j_1<j_2}^{n} h(Y_{\{i_1,i_2;j_1,j_2\}})$$

Example : motif frequencies

$$\begin{aligned} h(Y_{\{1,2;1,2\}}) = {} & Y_{11}Y_{12}Y_{21}(1-Y_{22}) + Y_{21}Y_{22}Y_{11}(1-Y_{12}) \\ & + Y_{12}Y_{11}Y_{22}(1-Y_{21}) + Y_{22}Y_{21}Y_{12}(1-Y_{11}) \end{aligned}$$



Outline

1. Main result

2. Application

Framework

Sequence of dimensions :

- $c \in \,]0, 1[$,
- $m_N := 2 + \lfloor c(N+1) \rfloor$, $n_N := 2 + \lfloor (1-c)(N+1) \rfloor$,
- $U_N^h := U_{m_N, n_N}^h$.

$\rightsquigarrow$ At step N, there are $N+4$ nodes in the network.

Decreasing filtration :

- $\mathcal{F}_N := \sigma\big((U_{k,l}^h, k \geq m_N, l \geq n_N)\big)$, $\mathcal{F}_\infty := \bigcap_{N=1}^{\infty} \mathcal{F}_N$.

Property

$\mathcal{F}_\infty \subset ... \subset \mathcal{F}_N \subset \mathcal{F}_{N-1} \subset ... \subset \mathcal{F}_0$.

Main result

Theorem 1

For RCE models, if $\mathbb{E}[h(Y_{\{1,2;1,2\}})^2] < \infty$, then

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

where

- $U_\infty^h = \mathbb{E}[h(Y_{\{1,2;1,2\}})|\mathcal{F}_\infty]$,
- W is a random variable with characteristic function $\phi(t) = \mathbb{E}[\exp(-\frac{1}{2}t^2 V)]$ (gaussian mixture),
- $V = \frac{4}{c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{1,3;3,4\}})\big|\mathcal{F}_\infty\big) + \frac{4}{1-c}\mathrm{Cov}\big(h(Y_{\{1,2;1,2\}}), h(Y_{\{3,4;1,3\}})\big|\mathcal{F}_\infty\big)$.

Outline of the proof

Eagleson & Weber, 1978 : weak convergence of sums of reverse martingale differences.

$$Z_{NK} := \sqrt{N}(U_K^h - U_{K+1}^h)$$

$\rightsquigarrow \sum_{K=N}^{\infty} Z_{NK} = \sqrt{N}(U_N^h - U_\infty^h)$

3 steps :

1. $(U_N, \mathcal{F}_N)$ is a reverse martingale : for each N, $U_N^h = \mathbb{E}[U_{N-1}^h | \mathcal{F}_N]$,
2. there exists $V > 0$ such that $\sum_{K=N}^{\infty} \mathbb{E}[Z_{NK}^2 | \mathcal{F}_{K+1}] \xrightarrow[N\to\infty]{\mathbb{P}} V$,

   (asymptotic variance),
3. for any $\epsilon > 0$, $\sum_{K=N}^{\infty} \mathbb{E}[Z_{NK}^2 \mathbb{1}_{\{|Z_{NK}|>\epsilon\}} | \mathcal{F}_{K+1}] \xrightarrow[N\to\infty]{\mathbb{P}} 0$

   (conditional Lindeberg condition).

Asymptotically normal case

Theorem 2

In addition to the assumptions of Theorem 1, if U_∞^h and V are constant with $V > 0$, then

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

où

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

U_∞^h and V are constant if Y is a dissociated RCE array.

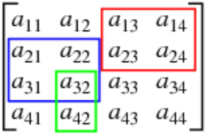
Dissociated arrays

Definition

Y is a dissociated array

$\Leftrightarrow$

For any $(m, n) \in \mathbb{N}^2$, $(Y_{ij})_{i \leq m, j \leq n}$ and $(Y_{ij})_{i > m, j > n}$ are independent.



submatrices

Key idea of the proof

Aldous-Hoover representation theorem

Let $\alpha, (\xi_i)_{1\leq i<\infty}$, $(\eta_j)_{1\leq j<\infty}$ and $(\zeta_{ij})_{1\leq i<\infty, 1\leq j<\infty}$ be arrays of i.i.d. random variables.

- If Y is RCE, then $Y \stackrel{\mathcal{D}}{=} Y^*$ where $Y^*_{ij} = f(\alpha, \xi_i, \eta_j, \zeta_{ij})$.
- If Y is RCE and dissociated, then $Y \stackrel{\mathcal{D}}{=} Y^*$ where $Y^*_{ij} = f(\xi_i, \eta_j, \zeta_{ij})$.

In the general RCE case, $\sigma(\alpha) \subset \mathcal{F}_\infty$, which explains the mixture in the limiting distribution.

Degenerate case

If $V = 0$, then the U-statistic is degenerate and

$$\sqrt{N}(U_N^h - U_\infty^h) \xrightarrow{\mathbb{P}} 0.$$

There exists $2 \leq d \leq 4$ and a random variable W such that

$$N^{\frac{d}{2}}(U_N^h - U_\infty^h) \xrightarrow[N\to\infty]{\mathcal{L}} W.$$

But the distribution of W is not explicit.

Outline

A RCE model

Bipartite Expected Degree Distribution model

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

where

- $\lambda = \mathbb{E}[Y_{ij}]$
- $\int f = \int g = 1, \int f^k = F_k, \int g^k = G_k.$

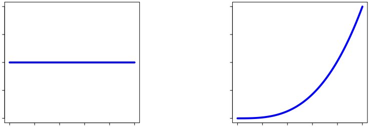
The BEDD model is :

- a model with latent variables for the expected degrees of the nodes,
- RCE and dissociated,
- identified by a quadruplet of nodes.

Example : Estimation of F_2

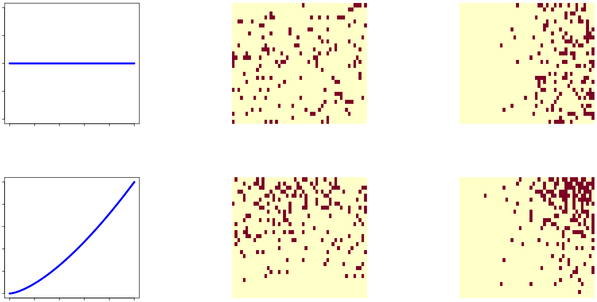
$g_0(v) =$ $g(v) =$

$f \equiv 1 \Leftrightarrow F_2 = 1$
$f \not\equiv 1 \Leftrightarrow F_2 > 1$



$f_0(u) =$

$f(u) =$



Example : Estimation of F_2

Step 1 : Choice of the kernels

- $h_1(Y_{\{1,2;1,2\}}) = Y_{11}Y_{12}$, $\mathbb{E}h_1 = \lambda^2 F_2$
- $h_2(Y_{\{1,2;1,2\}}) = Y_{11}Y_{22}$, $\mathbb{E}h_2 = \lambda^2$

Step 2 : Asymptotic normality

$\widehat{\theta}_N := U_N^{h_1} / U_N^{h_2}$

$$\sqrt{\frac{N}{V^{h_1}}} U_N^{h_2} \left(\widehat{\theta}_N - F_2\right) \xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0,1)$$

A consistent estimator for V^{h_1} is sufficient to build asymptotic confidence intervals.

Example : Estimation of F_2

V^{h_1} is derived from Theorem 2 :

$$V^{h_1} = \frac{\lambda^4}{c}(F_4 - F_2^2) + \frac{4\lambda^4}{1-c} F_2^2 (G_2 - 1)$$

Define more kernels to estimate G_2 and F_4 :

- $h_3(Y_{\{1,2;1,2\}}) = Y_{11} Y_{21}$, $\mathbb{E} h_3 = \lambda^2 G_2$
- $h_4(Y_{\{1,2;1,2\}}) = (Y_{11}^2 - Y_{11})(Y_{12}^2 - Y_{12})$, $\mathbb{E} h_4 = \lambda^4 F_4 G_2$

Step 3 : Consistent estimator of V^{h_1}

$$\widehat{V}_N^{h_1} = \frac{1}{c}\left[\frac{U_N^{h_4}(U_N^{h_2})^2}{(U_N^{h_3})^2} - (U_N^{h_1})^2\right] + \frac{4}{1-c}(U_N^{h_1})^2 \left[\frac{U_N^{h_3}}{U_N^{h_2}} - 1\right].$$

Conclusion

U-statistics can be used to perform statistical inference on bipartite networks :

- estimation,
- confidence intervals,
- statistical testing,
- network comparison.

Any row-column exchangeable model can be used (stochastic block models, graphons, ...).

Bibliography

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

Eagleson, G., & Weber, N. (1978). Limit theorems for weakly exchangeable arrays. In : *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 84, No. 1, pp. 123-130). Cambridge University Press.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19, 273-325.

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

Thank you for your attention !