

Processus empirique basé sur des U -statistiques à deux échantillons

Davide Giraudo
en collaboration avec Herold Dehling et Olimjon Sharipov

Ruhr-Universität Bochum
Financé par le Grant SFB 823

Mars 2020

U -statistique empirique à un échantillon

La convergence de la U -statistique empirique à un échantillon définie par

$$\frac{1}{n^{3/2}} \sum_{1 \leq i \neq j \leq n} (\mathbf{1}\{g(X_i, X_j) \leq s\} - \mathbb{P}\{g(X_i, X_j) \leq s\})$$

a été traitée (Svetlana Borovkova, Robert Burton et Herold Dehling). La convergence a lieu dans $D[0, 1]$ vers un processus gaussien dont la fonction de covariance est explicite.

U-statistique à deux échantillons

La convergence de la U -statistique à deux échantillons de noyau h défini par

$$T_n(h, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n (h(X_i, X_j) - \mathbb{E}[h(X_i, X_j)])$$

a également été traitée (Herold Dehling, Roland Fried, Isabel Garcia et Martin Wendler) où $(X_i)_{i \geq 1}$ est strictement stationnaires et vérifie certaines conditions de dépendance.

La convergence a lieu dans $D[0, 1]$ vers un processus gaussien dont la fonction de covariance est explicite.

U -statistique empirique à deux échantillons

Pour $n \geq 1, 0 \leq t \leq 1, s \in \mathbb{R}$, on définit

$$e_n(s, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n \mathbf{1}\{g(X_i, X_j) \leq s\},$$

où $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ est une fonction mesurable et pour tout réel x , $[x]$ désigne l'unique entier k tel que $k \leq x < k + 1$.

Lien avec les U -statistiques à deux échantillons:

$$T_n(h, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n h(X_i, X_j) :$$

$e_n(s, t) = T_n(h_s, t)$ où $h_s(u, v) = \mathbf{1}\{g(X_i, X_j) \leq s\}$.

Objectifs (1)

Soit $(X_i)_{i \geq 1}$ une suite i.i.d. et

$$e_n(s, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n \mathbf{1}_{\{g(X_i, X_j) \leq s\}}.$$

Trouver une condition portant sur g et $(X_i)_{i \geq 1}$ telle que

$$e_n(s, t) - \mathbb{E}[e_n(s, t)] \rightarrow W(s, t)$$

en loi dans $D([-R, R] \times [0, 1])$ vers un processus $W(\cdot, \cdot)$ et fournir une description de W .

Objectifs (2)

Soit $(X_i)_{i \geq 1}$ une suite i.i.d. et

$$e_n(s, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n \mathbf{1}_{\{g(X_i, X_j) \leq s\}}.$$

Dans certains cas, la probabilité $\mathbb{P}\{g(X_i, X_j) \leq s\}$ est inconnue. Elle est donc remplacée par un estimateur. On cherche une condition portant sur g et $(X_i)_{i \geq 1}$ telle que

$$e_n(s, t) - \frac{[nt](n - [nt])}{n^{3/2}} \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \mathbf{1}_{\{g(X_i, X_j) \leq s\}} \rightarrow W'(s, t)$$

en loi dans $D([-R, R] \times [0, 1])$ vers un processus $W'(\cdot, \cdot)$ et à fournir une description de W' .

Résultat

Theorem (Dehling, G., Sharipov (2020+))

Soit $(X_i)_{i \geq 0}$ une suite i.i.d. et e_n défini par

$$e_n(s, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n \mathbf{1}\{g(X_i, X_j) \leq s\}$$

On suppose que g est symétrique et que pour tout $u \in \mathbb{R}$, la variable aléatoire $g(u, X_1)$ a une densité f_u et que $\sup_{u,x} f_u(x) < +\infty$. Alors pour tout $R > 0$,

$$e_n(s, t) - \mathbb{E}[e_n(s, t)] \rightarrow W(s, t) \text{ en loi dans } D([-R, R] \times [0, 1]),$$

où $(W(s, t), s \in \mathbb{R}, t \in [0, 1])$ est un processus gaussien centré, dont la covariance est donnée pour $0 \leq t \leq t' \leq 1$ et $s, s' \in \mathbb{R}$ par :

$$\text{Cov}(W(s, t), W(s', t')) = t(1-t')(1-2t+2t') c_{s,s'},$$

où $c_{s,s'} = \mathbb{E}[h_{1,s}(X_1) h_{1,s'}(X_1)]$ et

$$h_{1,s}(u) = \mathbb{P}\{g(u, X_1) \leq s\} - \mathbb{P}\{g(X_1, X_2) \leq s\}.$$

Exemple

Soit $g(u, v) = u + v$ et supposons que X_1 aie une densité f bornée par M . La densité de $g(u, X_1) = u + X_1$ est $f_u(x) = f(x - u)$. De plus,

$$h_{1,s}(u) = \mathbb{P}\{X_1 \leq s - u\} - \mathbb{P}\{X_1 + X_2 \leq s\}.$$

Le cas non-symétrique

Les hypothèses sont les suivantes

1. la variable aléatoire $g(u, X_1)$ a une densité $f_{1,u}$ et $\sup_{u,x} f_{1,u}(x) < +\infty$.
2. La variable aléatoire $g(X_1, v)$ a une densité $f_{2,v}$ et $\sup_{u,x} f_{2,v}(x) < +\infty$.

La covariance du processus (gaussien centré) limite est donnée pour $t \leq t'$ par

$$\begin{aligned} \text{Cov}(W(s, t), W(s', t')) &= t(1-t)(1-t') \mathbb{E}[h_{1,s}(X_0) h_{1,s'}(X_0)] \\ &\quad + t(1-t')(t'-t) \mathbb{E}[h_{2,s}(X_0) h_{1,s'}(X_0)] \\ &\quad + tt'(1-t') \mathbb{E}[h_{2,s}(X_0) h_{2,s'}(X_0)], \end{aligned}$$

où

$$\begin{aligned} h_{1,s}(u) &= \mathbb{P}\{g(u, X_1) \leq s\} - \mathbb{P}\{g(X_1, X_2) \leq s\}, \\ h_{2,s}(v) &= \mathbb{P}\{g(X_1, v) \leq s\} - \mathbb{P}\{g(X_1, X_2) \leq s\}. \end{aligned}$$

Exemple de fonctionnelle

Soit μ une mesure finie sur les boréliens de \mathbb{R} .

Alors

$$\sup_{0 \leq t \leq 1} \int_{\mathbb{R}} (e_n(s, t) - \mathbb{E}[e_n(s, t)])^2 d\mu(s) \rightarrow \sup_{0 \leq t \leq 1} \int_{\mathbb{R}} W(s, t)^2 d\mu(s).$$

Idée de démonstration : aperçu global

Les étapes principales sont les suivantes :

1. On considère le noyau $h_s: \mathbb{R}^2 \rightarrow \mathbb{R}$ défini par $h_s(u, v) = \mathbf{1}_{\{g(u, v) \leq s\}}$. La décomposition d'Hoeffding's donne: une U -statistique contenant une partie linéaire + un terme dit dégénéré.
2. On établit la convergence des lois fini-dimensionnelles de la partie linéaire vers celles du processus W .
3. Puis on démontre que le processus associé à la partie linéaire converge vers W dans $D([-R, R] \times [0, 1])$ pour tout $R > 0$.
4. Enfin, on montre que la contribution de la partie dégénérée est négligeable.

Décomposition d'Hoeffding classique

Soit $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ une fonction mesurable et $(X_i)_{i \geq 1}$ une suite i.i.d. Soit

$$U_n = \sum_{1 \leq i < j \leq n} h(X_i, X_j).$$

On définit $\theta := \mathbb{E}[h(X_1, X_2)]$,

$$h_1(x) = \mathbb{E}[h(x, X_1)] - \theta, \quad h_2(y) = \mathbb{E}[h(X_1, y)] - \theta,$$

$$h_3(x, y) = h(x, y) - h_1(x) - h_2(y) - \theta.$$

Alors

$$U_n = \binom{n}{2} \theta + \sum_{i=1}^{n-1} (n-i) h_1(X_i) + \sum_{j=2}^n (j-1) h_2(X_j) + \sum_{1 \leq i < j \leq n} h_3(X_i, X_j)$$

et

$$\mathbb{E}[h_3(X_i, X_j) \mid X_1, \dots, X_{j-1}] = \mathbb{E}[h_3(X_i, X_j) \mid X_{i+1}, \dots, X_n] = 0.$$

Décomposition d'Hoeffding

Soit $\theta_s := \mathbb{P}\{g(X_1, X_2) \leq s\}$, $h_s(u, v) = \mathbf{1}\{g(u, v) \leq s\}$

$$h_{1,s}(u) = \mathbb{P}\{g(u, X_1) \leq s\} - \theta_s, \quad h_{2,s}(v) = \mathbb{P}\{g(X_1, v) \leq s\} - \theta_s,$$

et

$$h_{3,s}(u, v) = h_s(u, v) - h_{1,s}(u) - h_{2,s}(v) - \theta_s.$$

Alors

$$e_n(s, t) - \mathbb{E}[e_n(s, t)] = R_n(s, t) + W_n(s, t),$$

où

$$R_n(s, t) = \frac{1}{n^{3/2}} \sum_{i=1}^{\lfloor nt \rfloor} \sum_{j=\lfloor nt \rfloor+1}^n h_{3,s}(X_i, X_j)$$

$$W_n(s, t) = \frac{n - \lfloor nt \rfloor}{n^{3/2}} \sum_{i=1}^{\lfloor nt \rfloor} h_{1,s}(X_i) + \frac{\lfloor nt \rfloor}{n^{3/2}} \sum_{j=\lfloor nt \rfloor+1}^n h_{2,s}(X_j).$$

Lois fini-dimensionnelles

Cramer-Wold : soient $0 = t_0 < t_1 < \dots < t_d \leq t_{d+1} = 1$, $s_k \geq 1$, $1 \leq k \leq d$. On doit montrer que

$$\sum_{k,\ell=1}^d a_{k,\ell} W_n(s_\ell, t_k) \rightarrow \sum_{k,\ell=1}^d a_{k,\ell} W(s_\ell, t_k) \text{ en loi.}$$

On exprime $\sum_{k,\ell=1}^d a_{k,\ell} W_n(s_\ell, t_k)$ comme une combinaison linéaire de variables aléatoires indépendantes à laquelle on applique un théorème limite central pour les tableaux de variables aléatoires.

Soient

$$I_{n,u} = \{i \in \mathbb{N} \mid [nt_{u-1}] + 1 \leq i \leq [nt_u]\}, 2 \leq u \leq d,$$

$$I_{n,1} = \{i \in \mathbb{N} \mid 1 \leq i \leq [nt_1]\} \text{ et } I_{n,d+1} = \{i \in \mathbb{N} \mid [nt_d] + 1 \leq i \leq n\}.$$

Alors l'égalité suivante a lieu:

$$\begin{aligned} \sum_{k,\ell=1}^d a_{k,\ell} W_n(s_\ell, t_k) &= \sum_{k,\ell=1}^d a_{k,\ell} \frac{n - [nt_k]}{n^{3/2}} \sum_{u=1}^{d+1} \mathbf{1}\{u \leq k\} \sum_{i \in I_{n,u}} h_{1,s_\ell}(X_i) + \\ &\quad \sum_{k,\ell=1}^d a_{k,\ell} \frac{[nt_k]}{n^{3/2}} \sum_{u=1}^{d+1} \mathbf{1}\{u \geq k+1\} \sum_{i \in I_{n,u}} h_{2,s_\ell}(X_i). \end{aligned}$$

Convergence de la partie linéaire : outil

Theorem (Davydov, Zitikis)

Soit ξ_n , $n \geq 1$ des processus stochastiques définis sur $[0, 1]^2$ et dont les trajectoires se trouvent dans $D([0, 1]^2)$. Supposons que :

1. $\xi_n \rightarrow \xi$ (fin. dim.) et ξ a des trajectoires continues ;
2. on peut exprimer $\xi_n = \xi_n^\circ - \xi_n^*$ où ξ_n° et ξ_n^* sont croissants, $\xi_n(0, 0) = 0$;
3. il existe des constantes $\gamma \geq \beta > 2$, $c \in (0, \infty)$ telles que pour tout $n \geq 1$, si $\|(s, t) - (s', t')\|_\infty \geq 1/n$, alors

$$\mathbb{E} [|\xi_n(s, t) - \xi_n(s', t')|^\gamma] \leq c \|(s, t) - (s', t')\|_\infty^\beta$$

4. la convergence en probabilité suivante a lieu :

$$\max_{1 \leq i, j \leq n} \left| \xi_n^* \left(\frac{i}{n}, \frac{j}{n} \right) - \xi_n^* \left(\frac{i-1}{n}, \frac{j}{n} \right) \right| + \max_{1 \leq i, j \leq n} \left| \xi_n^* \left(\frac{i}{n}, \frac{j}{n} \right) - \xi_n^* \left(\frac{i}{n}, \frac{j-1}{n} \right) \right| \rightarrow 0.$$

Alors $(\xi_n)_{n \geq 1}$ converge en loi vers ξ dans $D([0, 1]^2)$.

Convergence de la partie linéaire

1. Condition sur $\mathbb{E} [|\xi_n(s, t) - \xi_n(s', t')|^\gamma]$: inégalité de Rosenthal et

$$a_s(u) = \mathbb{P}\{g(u, X_1) \leq s\}, \quad b_s(v) = \mathbb{P}\{g(X_1, v) \leq s\},$$

$$\|a_s - a_{s'}\|_\infty \leq C |s - s'|, \quad \|b_s - b_{s'}\|_\infty \leq C |s - s'| \quad (*)$$

2. Condition

$$\max_{1 \leq i, j \leq n} \left| \xi_n^* \left(\frac{i}{n}, \frac{j}{n} \right) - \xi_n^* \left(\frac{i-1}{n}, \frac{j}{n} \right) \right| + \max_{1 \leq i, j \leq n} \left| \xi_n^* \left(\frac{i}{n}, \frac{j}{n} \right) - \xi_n^* \left(\frac{i}{n}, \frac{j-1}{n} \right) \right| \rightarrow 0,$$

Le processus ξ_n^* est donné par

$$\xi_n^*(s, t) := \frac{[nt]}{n^{3/2}} \sum_{i=1}^{[nt]} a_s(X_i) + \frac{[nt]}{n^{3/2}} \sum_{j=1}^{[nt]} b_s(X_j) + 2\theta_{-R+2Rs} \frac{[nt]}{n^{1/2}},$$

et on utilise à nouveau (*).

Traitement de la partie dégénérée (1)

Rappelons que

$$R_n(s, t) = \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n h_{3,s}(X_i, X_j),$$

où, par définition de $h_{3,s}$,

$$\|h_{3,s} - h_{3,s'}\|_{\infty} \leq C |s - s'|,$$

$$\mathbb{E}[h_{3,s}(X_i, X_j) \mid \sigma(X_1, \dots, X_{j-1})] = 0$$

$$\mathbb{E}[h_{3,s}(X_i, X_j) \mid \sigma(X_k, k \geq i+1)] = 0.$$

On doit montrer que

$$\sup_{-R \leq s \leq R} \sup_{0 \leq t \leq 1} |R_n(s, t)| \rightarrow 0 \text{ en probabilité.}$$

Traitement de la partie dégénérée (2)

Premier pas : supremum sur t pour un s fixé :

$$\sup_{0 \leq t \leq 1} |R_n(s, t)| = \frac{1}{n^{3/2}} \max_{1 \leq \ell \leq n-1} \left| \sum_{i=1}^{\ell} \sum_{j=\ell+1}^n h_s(X_i, X_j) \right|.$$

Soit $h_{i,j} := h_{3,s}(X_i, X_j)$ et pour n fixé, $S_\ell := \sum_{i=1}^{\ell} \sum_{j=\ell+1}^n h_{i,j}$, $1 \leq \ell \leq n-1$ et $S_0 = 0$. Alors pour $2 \leq \ell \leq n$,

$$\begin{aligned} S_\ell - S_{\ell-1} &= \sum_{i=1}^{\ell} \sum_{j=\ell+1}^n h_{i,j} - \sum_{i=1}^{\ell-1} \sum_{j=\ell}^n h_{i,j} \\ &= \sum_{i=1}^{\ell-1} \sum_{j=\ell+1}^n h_{i,j} + \sum_{j=\ell+1}^n h_{\ell,j} - \sum_{i=1}^{\ell-1} \sum_{j=\ell+1}^n h_{i,j} - \sum_{i=1}^{\ell-1} h_{i,\ell} \\ &= \sum_{j=\ell+1}^n h_{\ell,j} - \sum_{i=1}^{\ell-1} h_{i,\ell} \end{aligned}$$

donc

$$S_k = \sum_{\ell=1}^k (S_\ell - S_{\ell-1}) = \sum_{1 \leq i < \ell \leq k} h_{i,\ell} + \sum_{\ell=1}^k \sum_{j=\ell+1}^n h_{\ell,j}.$$

Traitement de la partie dégénérée (3)

Terme $\max_{1 \leq k \leq n-1} \sum_{1 \leq i < \ell \leq k} h_{i,\ell}$: on applique une inégalité exponentielle pour les U -statistiques dégénérées à noyau borné.

Terme $\max_{1 \leq k \leq n-1} \sum_{\ell=1}^k \sum_{j=\ell+1}^n h_{\ell,j}$: il s'agit également d'un maximum d'une U -statistique, mais la suite considérée est $(X_n, X_{n-1}, \dots, X_1)$ au lieu de (X_1, \dots, X_n) .

On déduit que

$$\mathbb{P} \left\{ \sup_{0 \leq t \leq 1} |R_n(s, t)| > \varepsilon \right\} \leq C \exp(-c\sqrt{n\varepsilon}),$$

où les constantes c et C sont universelles (en particulier, indépendantes de s , n et ε).

Puis on découpe $[-R, R]$ en "petits" intervalles et on utilise l'inégalité

$$\|h_{3,s} - h_{3,s'}\|_{\infty} \leq K |s - s'|.$$

Différent centrage

Theorem (Dehling, G., Sharipov (2020+))

Soit $(X_i)_{i \geq 0}$ une suite i.i.d. et e_n défini par

$$e_n(s, t) := \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} \sum_{j=[nt]+1}^n \mathbf{1}\{g(X_i, X_j) \leq s\}$$

On suppose g symétrique et pour tout $u \in \mathbb{R}$, la variable aléatoire $g(u, X_1)$ a une densité f_u et $\sup_{u,x} f_u(x) < +\infty$. Alors pour tout $R > 0$,

$$e_n(s, t) - \frac{[nt](n - [nt])}{n^{3/2}} \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \mathbf{1}\{g(X_i, X_j) \leq s\} \rightarrow W(s, t)$$

en loi dans $D([-R, R] \times [0, 1])$ où $(W(s, t), s \in \mathbb{R}, t \in [0, 1])$ est un processus gaussien centré, dont la covariance est donnée pour $0 \leq t \leq t' \leq 1$ et $s, s' \in \mathbb{R}$ par :

$$\begin{aligned} \text{Cov}(W(s, t), W(s', t')) &= (1 - 2t)(1 - t)(1 - 2t')(t(1 - t') + t'^2) \\ &\quad \mathbb{E}[h_{1,s}(X_1), h_{1,s'}(X_1)]. \end{aligned}$$

et

$$h_{1,s}(u) = \mathbb{P}\{g(u, X_1) \leq s\} - \mathbb{P}\{g(X_1, X_2) \leq s\}.$$

Perspectives

1. Trouver d'autres exemples de fonctionnelles pour lesquelles les théorèmes limites obtenues peuvent être utilisés.
2. Traiter le cas dépendant, par exemple celui des suites stationnaires β -mélangeantes.

Conjecture : les coefficients de mélange doivent tendre polynomialement vers 0, *i. e.*, il existe un q_0 tel que si $\beta(n) \leq n^{-q_0-\delta}$ pour un $\delta > 0$, alors (à l'expression de la fonction de covariance près) la même convergence a lieu que dans le cas i.i.d.

Difficulté principale : une inégalité de déviation pour les U -statistiques de données β -mélangeantes.