# Raking-Ratio empirical process

Mickael Albertus
Supervisor: Philippe Berthet

International Workshop Postdam–Toulouse
March 13–15, 2019

# Introduction

# Notation

Let :

- $X_1, \ldots, X_n$ i.i.d. with unknown law $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$ ;

# Notation

Let :

- $X_1, \ldots, X_n$ i.i.d. with unknown law $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$ ;
- $\mathcal{A}^{(N)} = \{A_1^{(N)}, \ldots, A_{m_N}^{(N)}\}$ a partition of $\mathcal{X}$ s.t. we know $P(A_i^{(N)})$ ;

# Notation

Let :

- $X_1, \ldots, X_n$ i.i.d. with unknown law $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$ ;
- $\mathcal{A}^{(N)} = \{A_1^{(N)}, \ldots, A_{m_N}^{(N)}\}$ a partition of $\mathcal{X}$ s.t. we know $P(A_i^{(N)})$ ;
- $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ the empirical measure indexed by $\mathcal{F}$ ;

# Notation

Let :

- $X_1, \ldots, X_n$ i.i.d. with unknown law $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$ ;
- $\mathcal{A}^{(N)} = \{A_1^{(N)}, \ldots, A_{m_N}^{(N)}\}$ a partition of $\mathcal{X}$ s.t. we know $P(A_i^{(N)})$ ;
- $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ the empirical measure indexed by $\mathcal{F}$ ;
- $\alpha_n(f) = \sqrt{n}(\mathbb{P}_n(f) - P(f))$ the empirical process indexed by $\mathcal{F}$.

# Notation

Let :

- $X_1, \dots, X_n$ i.i.d. with unknown law $P$ on a measurable space $(\mathcal{X}, \mathcal{A})$ ;
- $\mathcal{A}^{(N)} = \{A_1^{(N)}, \dots, A_{m_N}^{(N)}\}$ a partition of $\mathcal{X}$ s.t. we know $P(A_i^{(N)})$ ;
- $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ the empirical measure indexed by $\mathcal{F}$ ;
- $\alpha_n(f) = \sqrt{n}(\mathbb{P}_n(f) - P(f))$ the empirical process indexed by $\mathcal{F}$.

If $\mathcal{F}$ is a Donsker class, $\alpha_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathbb{G}$ in $l^\infty(\mathcal{F})$ where $\mathbb{G}$ is the $P$-brownian bridge, *i.e* the Gaussian process with covariance

$$\mathrm{Cov}(\mathbb{G}(f), \mathbb{G}(g)) = P(fg) - P(f)P(g).$$

# Plan

1. Raking-ratio method

# Plan

1. Raking-ratio method
2. Extension 1: auxiliary information learning

# Plan

1. Raking-ratio method
2. Extension 1: auxiliary information learning
3. Extension 2: re-sampling method with auxiliary information

# Raking-ratio method

# Raking-Ratio method

**Literature:** Deming/Stephan, Sinkhorn, Ireland/Kullback.

**Description:**

| | $A_1^{(2)}$ | $A_2^{(2)}$ | $A_3^{(2)}$ | $\mathbb{P}_n[\mathcal{A}^{(1)}]$ | $P[\mathcal{A}^{(1)}]$ |
|---|---|---|---|---|---|
| $A_1^{(1)}$ | 0.2 | 0.25 | 0.1 | 0.55 | 0.52 |
| $A_2^{(1)}$ | 0.1 | 0.2 | 0.15 | 0.45 | 0.48 |
| $\mathbb{P}_n[\mathcal{A}^{(2)}]$ | 0.3 | 0.45 | 0.25 | 1 | |
| $P[\mathcal{A}^{(1)}]$ | 0.31 | 0.4 | 0.29 | | |

We have a table of frequencies whose margins do not correspond to known margins. The algorithm proposes to correct this

# Raking-Ratio method

| | $A_1^{(2)}$ | $A_2^{(2)}$ | $A_3^{(2)}$ | $\mathbb{P}_n^{(1)}[\mathcal{A}^{(1)}]$ | $P[\mathcal{A}^{(1)}]$ |
|---|---|---|---|---|---|
| $A_1^{(1)}$ | 0.189 | 0.236 | 0.095 | 0.52 | 0.52 |
| $A_2^{(1)}$ | 0.11 | 0.21 | 0.16 | 0.48 | 0.48 |
| $\mathbb{P}_n^{(1)}[\mathcal{A}^{(2)}]$ | 0.299 | 0.446 | 0.255 | 1 | |
| $P[\mathcal{A}^{(2)}]$ | 0.31 | 0.4 | 0.29 | | |

The totals for each line are first corrected by applying a rule of three. Each cell is multiplied by the ratio of the expected total of each line on the total of each line.

# Raking-Ratio method

| | $A_1^{(2)}$ | $A_2^{(2)}$ | $A_3^{(2)}$ | $\mathbb{P}_n^{(2)}[\mathcal{A}^{(1)}]$ | $P[\mathcal{A}^{(1)}]$ |
|---|---|---|---|---|---|
| $A_1^{(1)}$ | 0.196 | 0.212 | 0.108 | 0.516 | 0.52 |
| $A_2^{(1)}$ | 0.114 | 0.188 | 0.182 | 0.484 | 0.48 |
| $\mathbb{P}_n^{(2)}[\mathcal{A}^{(2)}]$ | 0.31 | 0.4 | 0.29 | 1 | |
| $P[\mathcal{A}^{(2)}]$ | 0.31 | 0.4 | 0.29 | | |

The same reasoning is applied to correct the totals for each column.
These last two operations are repeated in a loop.

# Raking-Ratio method

| | $A_1^{(2)}$ | $A_2^{(2)}$ | $A_3^{(2)}$ | $\mathbb{P}_n^{(\infty)}[\mathcal{A}^{(1)}]$ | $P[\mathcal{A}^{(1)}]$ |
|---|---|---|---|---|---|
| $A_1^{(1)}$ | 0.199 | 0.212 | 0.109 | 0.52 | 0.52 |
| $A_2^{(1)}$ | 0.111 | 0.188 | 0.181 | 0.48 | 0.48 |
| $\mathbb{P}_n^{(\infty)}[\mathcal{A}^{(2)}]$ | 0.31 | 0.4 | 0.29 | 1 | |
| $P[\mathcal{A}^{(2)}]$ | 0.31 | 0.4 | 0.29 | | |

Very quickly the algorithm stabilizes. Totals are the expected totals. For this example it took only 7 iterations.

**Remark:** we can rake on more than two partitions!

# Notation of Raking-Ratio method

In turn $N$ the algorithm does:

$$p^{(N+1)}(A) = \sum_{j=1}^{m_{N+1}} p^{(N)}(A \cap A_j^{(N+1)}) \frac{P(A_j^{(N+1)})}{p^{(N)}(A_j^{(N+1)})}.$$

# Notation of Raking-Ratio method

In turn $N$ the algorithm does:

$$p^{(N+1)}(A) = \sum_{j=1}^{m_{N+1}} p^{(N)}(A \cap A_j^{(N+1)}) \frac{P(A_j^{(N+1)})}{p^{(N)}(A_j^{(N+1)})}.$$

We define the raked empirical measure $\mathbb{P}_n^{(N)}$ to be $\mathbb{P}_n^{(0)} = \mathbb{P}_n$ and

$$\mathbb{P}_n^{(N+1)}(f) = \sum_{j=1}^{m_{N+1}} \mathbb{P}_n^{(N)}(f \mathbf{1}_{A_j^{(N+1)}}) \frac{P(A_j^{(N+1)})}{\mathbb{P}_n^{(N)}(A_j^{(N+1)})}.$$

In particular, $\mathbb{P}_n^{(N+1)}(A_j^{(N+1)}) = P(A_j^{(N+1)}), \forall j = 1, \ldots, m_{N+1}$.

# Notation of Raking-Ratio method

Let $\alpha_n^{(N)}(f) = \sqrt{n}(\mathbb{P}_n^{(N)}(f) - P(f))$ the raked empirical process.

$$\alpha_n^{(N+1)}(f) = \sum_{j \leqslant m_{N+1}} \frac{P(A_j^{(N+1)})}{\mathbb{P}_n^{(N)}(A_j^{(N+1)})} \left( \alpha_n^{(N)}(f \mathbf{1}_{A_j^{(N+1)}}) - \mathbb{E}[f | A_j^{(N+1)}] \alpha_n^{(N)}(A_j^{(N+1)}) \right)$$

with $\mathbb{E}[f|A] = \frac{P(f\mathbf{1}_A)}{P(A)}$.

In particular, $\alpha_n^{(N+1)}(A_j^{(N+1)}) = 0, \quad \forall j = 1, \dots, m_{N+1}$.

**Remark:** $\alpha_n^{(N)}$ is no more centered.

# Raking-Ratio method

## Goals

- Properties of $\alpha_n^{(N)}(\mathcal{F})$ ;

# Raking-Ratio method

## Goals

- Properties of $\alpha_n^{(N)}(\mathcal{F})$ ;
- Weak convergence in $\ell^\infty(\mathcal{F})$ of $\alpha_n^{(N)}(\mathcal{F})$ when $n \to +\infty$ towards a centered Gaussian process $\mathbb{G}^{(N)}(\mathcal{F})$ ;

## Goals

- Properties of $\alpha_n^{(N)}(\mathcal{F})$ ;
- Weak convergence in $\ell^\infty(\mathcal{F})$ of $\alpha_n^{(N)}(\mathcal{F})$ when $n \to +\infty$ towards a centered Gaussian process $\mathbb{G}^{(N)}(\mathcal{F})$ ;
- Variance of $\mathbb{G}^{(N)}(f)$: is it lower than that of $\mathbb{G}$? If a loop is performed with the Raking-Ratio method, does the variance decrease with each loop turn?

# Raking-Ratio method

## Goals

- Properties of $\alpha_n^{(N)}(\mathcal{F})$ ;
- Weak convergence in $\ell^\infty(\mathcal{F})$ of $\alpha_n^{(N)}(\mathcal{F})$ when $n \to +\infty$ towards a centered Gaussian process $\mathbb{G}^{(N)}(\mathcal{F})$ ;
- Variance of $\mathbb{G}^{(N)}(f)$: is it lower than that of $\mathbb{G}$? If a loop is performed with the Raking-Ratio method, does the variance decrease with each loop turn?
- If we rake only two partitions, what's the limit of $\alpha_n^{(N)}(\mathcal{F})$ as $n, N \to +\infty$?

# Theoretical results

## Law of iterated logarithm

If $\mathcal{F}$ satisfies some entropy conditions then for all $N_0 \in \mathbb{N}$,

$$\limsup_{n\to+\infty} \sqrt{\frac{n}{LLn}} \max_{0\leqslant N\leqslant N_0} ||\mathbb{P}_n^{(N)} - P||_{\mathcal{F}} \leqslant \sqrt{2}\sigma_{\mathcal{F}} \prod_{N=1}^{N_0} \left(1 + \frac{M}{\delta_N}\right) \text{ a.s.,}$$

where

- $\delta_N = \min_{j\leqslant m_N} P(A_j^{(N)})$ ;
- $\sigma_{\mathcal{F}}^2 = \sup_{\mathcal{F}} \mathrm{Var}(f)$ ;
- $M = ||f||_{\mathcal{F}}$.

Recall that

$$\limsup_{n\to+\infty} \sqrt{\frac{n}{LLn}} ||\mathbb{P}_n - P||_{\mathcal{F}} \leqslant \sqrt{2}\sigma_{\mathcal{F}}.$$

# Theoritical results

## Talagrand inequality

If $\mathcal{F}$ satisfies some entropy conditions then for all $N_0 \in \mathbb{N}$ and $t > t_0$,

$$\mathbb{P}\left(\max_{0 \leqslant N \leqslant N_0} ||\alpha_n^{(N)}||_{\mathcal{F}} > t\right) \leqslant D_1 \exp(-D_2 t^2),$$

or

$$\mathbb{P}\left(\max_{0 \leqslant N \leqslant N_0} ||\alpha_n^{(N)}||_{\mathcal{F}} > t\right) \leqslant D_1 t^{\nu} \exp(-D_2 t^2),$$

for some $D_1, D_2, \nu > 0$.

# Theoritical results

**Weak convergence of $\alpha_n^{(N)}$**

Under some entropy conditions on $\mathcal{F}$,

$$(\alpha_n^{(0)}, \dots, \alpha_n^{(N_0)}) \xrightarrow[n \to +\infty]{\mathcal{L}} (\mathbb{G}^{(0)}, \dots, \mathbb{G}^{(N_0)}) \quad \text{in} \quad \ell^\infty(\mathcal{F}^{N_0} \to \mathbb{R}^{N_0})$$

with $\mathbb{G}^{(N)}$ the Gaussian process defined by

$$\mathbb{G}^{(0)} = \mathbb{G} \quad \text{and} \quad \mathbb{G}^{(N+1)}(f) = \mathbb{G}^{(N)}(f) - \sum_{j=1}^{m_{N+1}} \mathbb{E}[f | A_j^{(N+1)}] \mathbb{G}^{(N)}(A_j^{(N+1)})$$

# Theoritical results

## Weak convergence of $\alpha_n^{(N)}$

Under some entropy conditions on $\mathcal{F}$,

$$(\alpha_n^{(0)}, \dots, \alpha_n^{(N_0)}) \xrightarrow[n \to +\infty]{\mathcal{L}} (\mathbb{G}^{(0)}, \dots, \mathbb{G}^{(N_0)}) \quad \text{in} \quad \ell^\infty(\mathcal{F}^{N_0} \to \mathbb{R}^{N_0})$$

with $\mathbb{G}^{(N)}$ the Gaussian process defined by

$$\mathbb{G}^{(0)} = \mathbb{G} \quad \text{and} \quad \mathbb{G}^{(N+1)}(f) = \mathbb{G}^{(N)}(f) - \sum_{j=1}^{m_{N+1}} \mathbb{E}[f|A_j^{(N+1)}]\mathbb{G}^{(N)}(A_j^{(N+1)})$$

Recall that

$$\alpha_n^{(N+1)}(f) = \sum_{j \leqslant m_{N+1}} \frac{P(A_j^{(N+1)})}{\mathbb{P}_n^{(N)}(A_j^{(N+1)})} \left( \alpha_n^{(N)}(f \mathbf{1}_{A_j^{(N+1)}}) - \mathbb{E}[f|A_j^{(N+1)}]\alpha_n^{(N)}(A_j^{(N+1)}) \right)$$

# Main tool

## Spirit of strong approximation

$$
\begin{tikzcd}
X_n \arrow[r, "\mathcal{L}"] \arrow[d, leftrightarrow, "\overset{\mathcal{L}}{=}"'] & X \arrow[d, leftrightarrow, "\overset{\mathcal{L}}{=}"] \\
Y_n \arrow[r, "p.s."'] & Y
\end{tikzcd}
$$

Results: KMT, Berthet-Mason.

# Main tool

**Strong approximation of $\alpha_n^{(N)}(\mathcal{F})$**

Under some entropy conditions on $\mathcal{F}$ we can construct on the same probability space $X_1, \ldots, X_n$ and a version $\mathbb{G}_n^{(N)}$ of $\mathbb{G}^{(N)}$ such that for large $n$,

$$\mathbb{P}\left(\max_{0\leqslant N\leqslant N_0} ||\alpha_n^{(N)} - \mathbb{G}_n^{(N)}||_{\mathcal{F}} > Cv_n\right) \leqslant \frac{1}{n^2},$$

with $v_n \to 0$.

By Borell-Cantelli,

$$\max_{0\leqslant N\leqslant N_0} ||\alpha_n^{(N)} - \mathbb{G}^{(N)}||_{\mathcal{F}} = O_{\text{p.s.}}(v_n).$$

# Consequence of strong approximation

## Berry-Esseen bound

Under some entropy conditions on $\mathcal{F}$,

$$\max_{0\leqslant N\leqslant N_0}\sup_{f\in\mathcal{F}}\sup_{x\in\mathbb{R}}\left|\mathbb{P}(\alpha_n^{(N)}(f)\leqslant x)-\mathbb{P}(\mathbb{G}^{(N)}(f)\leqslant x)\right|\leqslant Cv_n.$$

# Consequence of strong approximation

## Berry-Esseen bound

Under some entropy conditions on $\mathcal{F}$,

$$\max_{0 \leqslant N \leqslant N_0} \sup_{f \in \mathcal{F}} \sup_{x \in \mathbb{R}} \left| \mathbb{P}(\alpha_n^{(N)}(f) \leqslant x) - \mathbb{P}(\mathbb{G}^{(N)}(f) \leqslant x) \right| \leqslant C v_n.$$

## Bias and variance estimation

Under some entropy conditions on $\mathcal{F}$, there exists $C > 0$ such that

$$\limsup_{n \to +\infty} \frac{\sqrt{n}}{v_n} \max_{0 \leqslant N \leqslant N_0} \sup_{f \in \mathcal{F}} \left| \mathbb{E}[\mathbb{P}_n^{(N)}(f)] - P(f) \right| \leqslant C,$$

$$\limsup_{n \to +\infty} \frac{n}{v_n} \sup_{f \in \mathcal{F}} \left| \mathrm{Var}(\mathbb{P}_n^{(N)}(f)) - \frac{1}{n} \mathrm{Var}(\mathbb{G}^{(N)}(f)) \right| \leqslant C.$$

# Raking-Ratio results

We denote

- $\mathbb{E}[f|\mathcal{A}^{(k)}] = (\mathbb{E}[f|A_1^{(k)}], \dots, \mathbb{E}[f|A_{m_k}^{(k)}])^t$ ;
- $\mathbb{G}[\mathcal{A}^{(k)}] = (\mathbb{G}(A_1^{(k)}), \dots, \mathbb{G}(A_{m_k}^{(k)}))^t$ ;
- $(\mathsf{P}_{\mathcal{A}^{(k)}|\mathcal{A}^{(l)}})_{i,j} = P(A_j^{(k)}|A_i^{(l)})$.

**Expression of $\mathbb{G}^{(N)}$**

For all $N \in \mathbb{N}^*$ and $f \in \mathcal{F}$ it holds

$$\mathbb{G}^{(N)}(f) = \mathbb{G}(f) - \sum_{k=1}^{N} \Phi_k^{(N)}(f)^t \cdot \mathbb{G}[\mathcal{A}^{(k)}]$$

where

$$\Phi_k^{(N)}(f) = \mathbb{E}[f|\mathcal{A}^{(k)}] + \sum_{\substack{1 \leqslant L \leqslant N-k \\ k < l_1 < \dots < l_L \leqslant N}} (-1)^L \mathsf{P}_{\mathcal{A}^{(l_1)}|\mathcal{A}^{(k)}} \mathsf{P}_{\mathcal{A}^{(l_2)}|\mathcal{A}^{(l_1)}} \dots \mathsf{P}_{\mathcal{A}^{(l_L)}|\mathcal{A}^{(l_L-1)}} \cdot \mathbb{E}[f|\mathcal{A}^{(l_L)}].$$

# Raking-Ratio results

We denote $(\mathrm{Var}((X_1,\dots,X_n)^t))_{i,j} = \mathrm{Cov}(X_i, X_j)$

**Variance and covariance of $\mathbb{G}^{(N)}$**

For all $N \in \mathbb{N}^*$ and $f, g \in \mathcal{F}$ it holds

$$\mathrm{Var}(\mathbb{G}^{(N)}(f)) = \mathrm{Var}(\mathbb{G}(f)) - \sum_{k=1}^{N} \Phi_k^{(N)}(f)^t \cdot \mathrm{Var}(\mathbb{G}[\mathcal{A}^{(k)}]) \cdot \Phi_k^{(N)}(f)$$

$$\begin{aligned}\mathrm{Cov}(\mathbb{G}^{(N)}(f), \mathbb{G}^{(N)}(g)) &= \mathrm{Cov}(\mathbb{G}(f), \mathbb{G}(g)) \\ &\quad - \sum_{k=1}^{N} \mathrm{Cov}\left(\Phi_k^{(N)}(f)^t \cdot \mathbb{G}[\mathcal{A}^{(k)}], \Phi_k^{(N)}(g)^t \cdot \mathbb{G}[\mathcal{A}^{(k)}]\right)\end{aligned}$$

# Raking-Ratio results

**Corollary 1**

For any $N \in \mathbb{N}$ and $f \in \mathcal{F}, \mathrm{Var}(\mathbb{G}^{(N)}(f)) \leqslant \mathrm{Var}(\mathbb{G}(f))$.

For any $\{f_1, \ldots, f_m\} \in \mathcal{F}, \Sigma_m - \Sigma_m^{(N)}$ is positive definite with

$$\Sigma_n^{(N)} = \mathrm{Var}((\mathbb{G}^{(N)}(f_1), \ldots, \mathbb{G}^{(N)}(f_m))^t),$$
$$\Sigma_n = \mathrm{Var}((\mathbb{G}(f_1), \ldots, \mathbb{G}(f_m))^t).$$

# Raking-Ratio results

## Corollary 1

For any $N \in \mathbb{N}$ and $f \in \mathcal{F}$, $\mathrm{Var}(\mathbb{G}^{(N)}(f)) \leqslant \mathrm{Var}(\mathbb{G}(f))$.

For any $\{f_1, \ldots, f_m\} \in \mathcal{F}$, $\Sigma_m - \Sigma_m^{(N)}$ is positive definite with

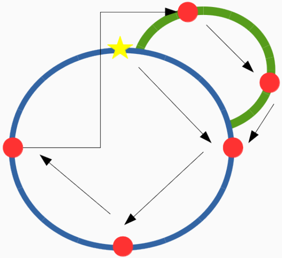
$$\Sigma_n^{(N)} = \mathrm{Var}((\mathbb{G}^{(N)}(f_1), \ldots, \mathbb{G}^{(N)}(f_m))^t),$$
$$\Sigma_n = \mathrm{Var}((\mathbb{G}(f_1), \ldots, \mathbb{G}(f_m))^t).$$

## Corollary 2

Let $N_0, N_1 \in \mathbb{N}$ s.t. $N_1 \geqslant 2N_0$ and

$$\mathcal{A}^{(N_0 - i)} = \mathcal{A}^{(N_1 - i)}, \quad \forall 0 \leqslant i < N_0.$$

Then for all $f \in \mathcal{F}$,
$\mathrm{Var}(\mathbb{G}^{(N_1)}(f)) \leqslant \mathrm{Var}(\mathbb{G}^{(N_0)}(f))$.



# Results for 2 margins

We note $\mathcal{A} = \mathcal{A}^{(2)} = \{A_1, \ldots, A_{m_1}\}$ and $\mathcal{B} = \mathcal{A}^{(1)} = \{B_1, \ldots, B_{m_2}\}$.

## Expression of $\mathbb{G}^{(N)}$

Let $N \in \mathbb{N}$ and $f \in \mathcal{F}$. Then for $m \in \mathbb{N}$,

$$\mathbb{G}^{(2m)}(f) = \mathbb{G}(f) - \left(S_{1,even}^{(m-1)}(f)\right)^t \mathbb{G}[\mathcal{A}] - \left(S_{2,even}^{(m-2)}(f)\right)^t \mathbb{G}[\mathcal{B}]$$

$$\mathbb{G}^{(2m+1)}(f) = \mathbb{G}(f) - \left(S_{1,odd}^{(m-1)}(f)\right)^t \mathbb{G}[\mathcal{A}] - \left(S_{2,odd}^{(m-1)}(f)\right)^t \mathbb{G}[\mathcal{B}]$$

# Results for 2 margins

We note $\mathcal{A} = \mathcal{A}^{(2)} = \{A_1, \ldots, A_{m_1}\}$ and $\mathcal{B} = \mathcal{A}^{(1)} = \{B_1, \ldots, B_{m_2}\}$.

## Expression of $\mathbb{G}^{(N)}$

Let $N \in \mathbb{N}$ and $f \in \mathcal{F}$. Then for $m \in \mathbb{N}$,

$$\mathbb{G}^{(2m)}(f) = \mathbb{G}(f) - \left(S_{1,even}^{(m-1)}(f)\right)^t \mathbb{G}[\mathcal{A}] - \left(S_{2,even}^{(m-2)}(f)\right)^t \mathbb{G}[\mathcal{B}]$$

$$\mathbb{G}^{(2m+1)}(f) = \mathbb{G}(f) - \left(S_{1,odd}^{(m-1)}(f)\right)^t \mathbb{G}[\mathcal{A}] - \left(S_{2,odd}^{(m-1)}(f)\right)^t \mathbb{G}[\mathcal{B}]$$

with $S_{1,even}^{(N)}(f) = \sum_{k=0}^{N} (\mathsf{P}_{\mathcal{B}|\mathcal{A}}\mathsf{P}_{\mathcal{A}|\mathcal{B}})^k (\mathbb{E}[f|\mathcal{A}] - \mathsf{P}_{\mathcal{B}|\mathcal{A}}\mathbb{E}[f|\mathcal{B}])$

$$S_{2,odd}^{(N)}(f) = \sum_{k=0}^{N} (\mathsf{P}_{\mathcal{A}|\mathcal{B}}\mathsf{P}_{\mathcal{B}|\mathcal{A}})^k (\mathbb{E}[f|\mathcal{B}] - \mathsf{P}_{\mathcal{A}|\mathcal{B}}\mathbb{E}[f|\mathcal{A}])$$

$$S_{2,even}^{(N)}(f) = S_{2,odd}^{(N)}(f) + (\mathsf{P}_{\mathcal{A}|\mathcal{B}}\mathsf{P}_{\mathcal{B}|\mathcal{A}})^{N+1}\mathbb{E}[f|\mathcal{B}]$$

$$S_{1,odd}^{(N)}(f) = S_{1,even}^{(N)}(f) + (\mathsf{P}_{\mathcal{B}|\mathcal{A}}\mathsf{P}_{\mathcal{A}|\mathcal{B}})^{N+1}\mathbb{E}[f|\mathcal{A}]$$

Recall that

$$\mathbb{G}^{(N)}(f) = \mathbb{G}(f) - \sum_{k=1}^{N} \Phi_k^{(N)}(f)^t \cdot \mathbb{G}[\mathcal{A}^{(k)}]$$

## Results for 2 margins

**Hypothesis**

Matrices $\mathsf{P}_{\mathcal{A}|\mathcal{B}}\mathsf{P}_{\mathcal{B}|\mathcal{A}}$ and $\mathsf{P}_{\mathcal{B}|\mathcal{A}}\mathsf{P}_{\mathcal{A}|\mathcal{B}}$ are ergodic.

# Results for 2 margins

Hypothesis

Matrices $\mathsf{P}_{\mathcal{A}|\mathcal{B}}\mathsf{P}_{\mathcal{B}|\mathcal{A}}$ and $\mathsf{P}_{\mathcal{B}|\mathcal{A}}\mathsf{P}_{\mathcal{A}|\mathcal{B}}$ are ergodic.

Convergence of $S_{i,even}^{(N)}(f), S_{i,odd}^{(N)}(f)$

$S_{i,even}^{(N)}(f), S_{i,odd}^{(N)}(f)$ for $i = 1, 2$ converge respectively towards $S_{i,even}(f), S_{i,odd}(f)$. They verify the relations:

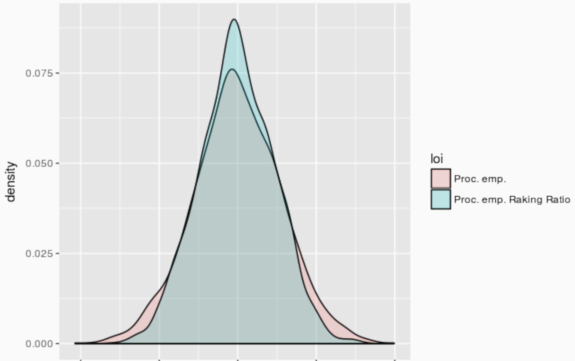
$$S_{1,odd}(f) = S_{1,even}(f) + \begin{pmatrix} \mathbb{E}[f] \\ \vdots \\ \mathbb{E}[f] \end{pmatrix}, \qquad S_{2,even}(f) = S_{2,odd}(f) + \begin{pmatrix} \mathbb{E}[f] \\ \vdots \\ \mathbb{E}[f] \end{pmatrix}.$$

# Results for 2 margins

## Convergence of $\mathbb{G}^{(N)}$

The sequence of process $(\mathbb{G}^{(N)})_N$ converges in distribution when $N \to +\infty$ to the centered Gaussian process $\mathbb{G}^{(\infty)}$ indexed by $\mathcal{F}$ and defined by

$$\mathbb{G}^{(\infty)}(f) = \mathbb{G}(f) - S_{1,even}(f)^t \cdot \mathbb{G}[\mathcal{A}] - S_{2,odd}(f)^t \cdot \mathbb{G}[\mathcal{B}].$$



# Extension 1: auxiliary information learning

## Motivation

We suppose that the auxiliary information is given by an estimate of the probability of belonging to a set of several partitions.

The auxiliary information is given by

$$\mathbb{P}'_N[\mathcal{A}^{(N)}] = (\mathbb{P}_n(A_1^{(N)}), \ldots, \mathbb{P}'_N(A_{m_N}^{(N)})),$$

a multinomial distribution with $n_N > 0$ trials and event probabilities

$$P[\mathcal{A}^{(N)}] = (P(A_1^{(N)}), \ldots, P(A_{m_N}^{(N)})).$$

# Introduction

## Motivation

We suppose that the auxiliary information is given by an estimate of the probability of belonging to a set of several partitions.

The auxiliary information is given by

$$\mathbb{P}'_N[\mathcal{A}^{(N)}] = (\mathbb{P}_n(A_1^{(N)}), \dots, \mathbb{P}'_N(A_{m_N}^{(N)})),$$

a multinomial distribution with $n_N > 0$ trials and event probabilities

$$P[\mathcal{A}^{(N)}] = (P(A_1^{(N)}), \dots, P(A_{m_N}^{(N)})).$$

## Goal

We study the raking-ratio empirical process which uses $\mathbb{P}'_N[\mathcal{A}^{(N)}]$ instead of $P[\mathcal{A}^{(N)}]$.

## Definition of $\widetilde{\mathbb{P}}_n^{(N)}(\mathcal{F})$

We define the $N$-th raking-ratio empirical measure with auxiliary information learning $\widetilde{\mathbb{P}}_n^{(N)}(\mathcal{F})$ as the same way as $\mathbb{P}_n^{(N)}(\mathcal{F})$: $\widetilde{\mathbb{P}}_n^{(0)}(f) = \mathbb{P}_n(f)$ and for $N \geqslant 1$,

$$\widetilde{\mathbb{P}}_n^{(N)}(f) = \sum_{j=1}^{m_N} \frac{\mathbb{P}'_N(A_j^{(N)})}{\widetilde{\mathbb{P}}_n^{(N-1)}(A_j^{(N)})} \widetilde{\mathbb{P}}_n^{(N-1)}(f 1_{A_j^{(N)}}).$$

Notice that

$$\widetilde{\mathbb{P}}_n[\mathcal{A}^{(N)}] = \left(\widetilde{\mathbb{P}}_n^{(N)}(A_1^{(N)}), \ldots, \widetilde{\mathbb{P}}_n^{(N)}(A_{m_N}^{(N)})\right) = \mathbb{P}'_N[\mathcal{A}^{(N)}].$$

Recall that

$$\mathbb{P}_n^{(N)}(f) = \sum_{j=1}^{m_N} \frac{\mathbb{P}_n(A_j^{(N)})}{\mathbb{P}_n^{(N-1)}(A_j^{(N)})} \mathbb{P}_n^{(N-1)}(f 1_{A_j^{(N)}}).$$

**Definition of $\widetilde{\alpha}_n^{(N)}(\mathcal{F})$**

We define the $N$-th raking-ratio empirical process with estimated auxiliary information by

$$\widetilde{\alpha}_n^{(N)}(f) = \sqrt{n}(\widetilde{\mathbb{P}}_n^{(N)}(f) - P(f)).$$

Notice that $\alpha_n^{(N)}(A_j^{(N)}) \neq 0$.

# Main result

**Strong approximation of $\alpha_n^{(N)}(\mathcal{F})$**

Under some entropy conditions on $\mathcal{F}$ we can construct on the same probability space $X_1, \dots, X_n$ and a version $\mathbb{G}_n^{(N)}$ of $\mathbb{G}^{(N)}$ such that for large $n$,

$$\mathbb{P}\left( \max_{0 \leqslant N \leqslant N_0} ||\widetilde{\alpha}_n^{(N)} - \mathbb{G}_n^{(N)}||_{\mathcal{F}} > C\left( v_n + \sqrt{\frac{n\log(n)}{n_{(N_0)}}} \right) \right) \leqslant \frac{1}{n^2},$$

with $v_n \to 0$ and $n_{(N_0)} = \min_{N \leqslant N_0} n_N$.

# Extension 2: re-sampling method with auxiliary information

## Notation

Bootstrap is a statistical method for re-sampling. It replaces $P$ by $\mathbb{P}_n$.

A general way to define the bootstrap is to multiply $f(X_i)$ by a random variable $Z_i$ such that $\mathbb{E}[Z_i|X_i] = 1$ and $\mathrm{Var}(Z_i) = 1$.

# Introduction

## Notation

Bootstrap is a statistical method for re-sampling. It replaces $P$ by $\mathbb{P}_n$.

A general way to define the bootstrap is to multiply $f(X_i)$ by a random variable $Z_i$ such that $\mathbb{E}[Z_i|X_i] = 1$ and $\mathrm{Var}(Z_i) = 1$.

We define the bootstrapped empirical measure and process:

$$\mathbb{P}_n^*(f) = \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i f(X_i), \qquad \alpha_n^*(f) = \sqrt{n}(\mathbb{P}_n^*(f) - \mathbb{P}_n(f)).$$

# Introduction

## Notation

Bootstrap is a statistical method for re-sampling. It replaces $P$ by $\mathbb{P}_n$.

A general way to define the bootstrap is to multiply $f(X_i)$ by a random variable $Z_i$ such that $\mathbb{E}[Z_i|X_i] = 1$ and $\mathrm{Var}(Z_i) = 1$.

We define the bootstrapped empirical measure and process:

$$\mathbb{P}_n^*(f) = \frac{1}{\sum_{i=1}^n Z_i} \sum_{i=1}^n Z_i f(X_i), \qquad \alpha_n^*(f) = \sqrt{n}(\mathbb{P}_n^*(f) - \mathbb{P}_n(f)).$$

## Goal

- Make the strong approximation of $\alpha_n^*$ to $\mathbb{G}^*$, a $P$-Brownian bridge independent of $\mathbb{G}$ ;
- Bootstrap the Raking-Ratio empirical process to simulate its distribution.

# Strong approximation of the bootstrapped empirical process

**Strong approximation of $\alpha_n^*$**

Under some entropy conditions on $\mathcal{F}$ we can construct on the same probability space $(X_n, Z_n)$ and $(\mathbb{G}_n,, \mathbb{G}_n^*)$ of $P$-Brownian bridge such that for large $n$,

$$\mathbb{P}\left(\{||\alpha_n - \mathbb{G}_n||_{\mathcal{F}} > Cv_n\} \bigcup \{||\alpha_n^* - \mathbb{G}_n^*||_{\mathcal{F}} > Cv_n\}\right) \leqslant \frac{1}{n^2},$$

with $v_n \to 0$ depends on the entropy of $(\mathcal{F}, P)$.

## Goal

How can we adapt the bootstrap method to simulate the distribution of the Raking-Ratio empirical process?

$\mathbb{P}_n^{*(0)} = \mathbb{P}_n^*$ and

$$\mathbb{P}_n^{*(N+1)}(f) = \sum_{j=1}^{m_{N+1}} \mathbb{P}_n^{*(N)}(f\mathbb{1}_{A_j^{(N+1)}}) \frac{\mathbb{P}_n(A_j^{(N+1)})}{\mathbb{P}_n^{*(N)}(A_j^{(N+1)})},$$

$$\alpha_n^{*(N)}(f) = \sqrt{n}(\mathbb{P}_n^{*(N)}(f) - \mathbb{P}_n(f)).$$

# Bootstrap and Raking-Ratio

## Goal

How can we adapt the bootstrap method to simulate the distribution of the Raking-Ratio empirical process?

$\mathbb{P}_n^{*(0)} = \mathbb{P}_n^*$ and

$$\mathbb{P}_n^{*(N+1)}(f) = \sum_{j=1}^{m_{N+1}} \mathbb{P}_n^{*(N)}(f \mathbb{1}_{A_j^{(N+1)}}) \frac{\mathbb{P}_n(A_j^{(N+1)})}{\mathbb{P}_n^{*(N)}(A_j^{(N+1)})},$$

$$\alpha_n^{*(N)}(f) = \sqrt{n}(\mathbb{P}_n^{*(N)}(f) - \mathbb{P}_n(f)).$$

## Result

$\alpha_n^{*(N)} \to \mathbb{G}^{*(N)}$ in $\ell^\infty(\mathcal{F})$ and $\mathbb{G}^{*(N)}$ has the same distribution as $\mathbb{G}^{(N)}$.

Thank you for your attention!

Questions?