# Théorème de Donsker pour des mesures empiriques locales sur des régions aléatoires

Benjamin Bobbia [1] & Clément Dombry [2] & Davit Varron [3]

[1] *École supérieur d'ingénieur Léonard de Vinci, 12 Avenue Léonard de Vinci 92 400 Courbevoie benjamin.bobbia@devinci.fr*

[2] *Université de Franche-Comté, 16 route de Gray 25 000 Besançon clement.dombry@univ-fcomte.fr*

[3] *Université de Franche-Comté, 16 route de Gray 25 000 Besançon davit.varron@univ-fcomte.fr*

**Résumé.** Les processus empiriques sont aujourd'hui des objets bien connus. Une des raisons qui a poussée au développement de l'étude des processus empiriques est qu'il est possible, ans de nombreuses modélisations, d'écrire les estimateurs comme images de mesures empiriques. Dans ce travail nous regardons le cas particulier des mesures empiriques locales, c'est-à-dire, la mesure empirique construite sur un sous-échantillon d'observations conditionnées à un être dans une certaine partie de l'espace. De nombreux résultats existent pour ce type de mesure, mais que peut-on-dire si la partie de l'espace en question dépend des données ? Il est possible de s'en sortir au prix d'une grande technicité et de conditions de régularité, le propos de ce travail et de présenter un cadre général pour l'étude de ces mesures empiriques particulières permettant d'obtenir des résultats asymptotiques à moindre cout (technique et conditions). Nous donnons alors un théorème de Donsker pour de telles mesures et nous présentons des exemples d'application comme la théorie des valeurs extrêmes ou les variations régulières multivariées.

**Mots-clés.** Mesure empirique locale, Extrêmes, théorème de Donsker.

**Abstract.** Nowadays, empirical processes are well known object. A reason that push forward theirs studies is that, in many modelisations, we can write the estimators as images of empirical measures. In this work we investigate the case of local empirical measures, that is the empirical measure built over a subsample with data conditionned to be in a certain area. There exist numerous results about such result, but what can we say if the previous area is data driven ? We can handle this situation with high technical cost and additional assumptions. The main aim of the present work is to present a general framework which allows to derive asymptotic results for this particular empirical measures with lower cost (in terms of technicality and assumptions). We give a Donsker-type theorem for such measures and present some domains of application as extreme values theory or multivariate regular variations.

**Keywords.** Local empirical measures, Extremes, Donsker's theorem.

Cette communication présente certains résultats issus de l'article en prépublication *A Donsker and Glivencko-Cantelli theorem for random measures linked to extreme value theory* qu'il est possible de trouver à l'adresse
**https://hal.archives-ouvertes.fr/hal-03402380/document**.

# 1 La mesure empirique locale et motivations

Considérons $P_0$ une mesure sur un espace $\mathcal{X}$ ainsi qu'une suite $(X_1, dots, X_n)$ un échantillon i.i.d de loi $P_0$. On définit alors la mesure empirique, avec pour but d'estimer $P_0$, par

$$\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

Ainsi, pour une fonction $\varphi$ suffisamment régulière, $\varphi(\mathbb{P}_n)$ est censée être proche de $\varphi(P_0)$ au sens de la convergence en probabilité. Il apparaît alors naturel de s'intéresser à cette quantité dans le cadre de l'estimation. Cependant dans certains contextes comme l'estimation par plus proche voisin ou la théorie des valeurs extrêmes, l'estimation met en jeu uniquement un sous-échantillon de $(X_1, \dots, X_n)$, on construit alors une mesure empirique, dite local, sur ce sous-échantillon d'intérêt. Cette mesure est définie comme suit. Soit $(A_n)_{n\geq 1}$ une suite d'ensembles tel que $p_n := P_0(A_n)$ tend vers 0, on définit alors

$$\mathbb{P}_n^{loc} := \frac{1}{card(\{i \text{ tel que } X_i \in A_n\})}\sum_{i=1}^{n} \mathbb{1}_{A_n}(X_i)\delta_{X_i}.$$

L'utilisation d'une telle mesure apparaît notamment dans l'approche peak-over-thresholds introduite par Balkema et de Haan (1974) en théorie des valeurs extrêmes. Cette approche consiste à utiliser uniquement les données supérieures à un certain seuil $y_n$ (tendant vers l'infini) pour obtenir des informations sur la queue de distribution. Dans ce cas notre mesure empirique locale s'écrit

$$\mathbb{P}_n^{ext} := \frac{1}{card(\{i \text{ tel que } X_i > y_n\})}\sum_{X_i > x_n} \delta_{X_i/y_n}.$$

Dans cette approche, il est souvent préférable d'utiliser comme seuil un point de l'échantillon, en particulier le $k$-ième plus grand que l'on note $X_{n-k:n}$ (avec $k = k_n$ convergeant vers l'infini mais moins vite que $n$). Ce qui revient à considérer une mesure empirique locale avec un ensemble $A_n$ dépendant des données, objet centrale de ce travail.

# 2 Résultats

Un des objectifs du présent travail est de fournir un cadre général dans lequel il sera pertinent d'étudier les mesures empiriques locales introduites précédemment.

**Hypothèse de mesure empirique conditionnelle.** *Soit $(P_n)_{n\geq 1}$ une suite de mesures aléatoires sur $\mathcal{X}$. On dit que $P_n$ satisfait la condition de mesure empirique conditionnelle s'il existe une variable aléatoire $U_n$ prenant ses valeurs dans ensemble $E$ et une famille de probabilité sur $\mathcal{X}$, $(Q_u)_{u\in E}$ telles que pour presque tout $u \in E$ on a, sachant que $U_n = u$,*

$$P_n \overset{\mathcal{L}}{=} \frac{1}{k_n}\sum_{i=1}^{k_n} \delta_{X_i^{(u)}}$$

*avec $k_n$ un entier et $(X_1^{(u)}, \ldots, X_{k_n}^{(u)})$ i.i.d de loi $Q_u$.*

Dans le contexte des extrêmes, il est possible de montrer que la suite de mesures empiriques locales $\mathbb{P}_n^{ext}$ satisfait cette hypothèse de mesure empirique conditionnelle avec $U_n = Y_{n-k_n:n}$ et $P_u(x) = P_0(x/u|X > u)$.
Sous cette hypothèse sur la famille $(P_n)_{n\geq 1}$ et lorsque la suite $U_n$ converge en loi vers un certain $U$ dans $E$, il est possible d'établir des résultats de convergence pour le processus empirique indexé par une classe de fonctions $\mathcal{F}$

$$\mathbb{G}_n(f) := \sqrt{k_n}(P_n(f) - P_U(f)), \quad f \in \mathcal{F},$$

où $P(f)$ désigne $\int f \mathrm{d}P$.

**Théorème 1.** *Sous des hypothèses classiques sur la classe $\mathcal{F}$ et de régularité sur l'application $u \mapsto \{f \mapsto P_u(f)\}$ on a*

$$\mathbb{G}_n \overset{\mathcal{L}}{\to} \mathcal{W}, \qquad \text{dans } l^\infty(\mathcal{F}),$$

*avec $\mathcal{W}$ le $P_U$-Pont Brownien indexé par $\mathcal{F}$.*

# 3 Exemples de domaines d'applications

Parmi les applications envisagées, cette section détail deux applications. Une à l'estimation de l'indice des valeurs extrêmes dans le cas des extrêmes univariés réels ainsi qu'une autre a un test d'ajustement de modèle pour une variable à variations régulières multivariées.

## 3.1 Théorie des valeurs extrêmes

Comme $\mathbb{P}_n^{ext}$ vérifie l'hypothèse de mesure empirique conditionnelle, le théorème précédent permet d'affirmer que pour une classe de fonction $\mathcal{F}$ appropriée, on a

$$\sqrt{k}(\mathbb{P}_n^{ext} - P_0) \overset{\mathcal{L}}{\to} \mathcal{W}_{P_0} \quad \text{dans } l^\infty(\mathcal{F}),$$

avec $\mathcal{W}_{P_0}$ le $P_0$ pont Brownien indexé par $\mathcal{F}$.
Ce résultat nous permet l'estimation du bien nommé paramètre appelé indice des valeurs

extrêmes $\gamma$ qui renseigne sur le comportement de la queue de distribution de $P_0$ (plus $\gamma$ est élevé, plus la queue de $P_0$ est lourde). En effet, pour les distributions à queues lourdes $\gamma$ peut être estimer par le célèbre estimateur de Hill (1975)

$$\hat{\gamma}_n := \frac{1}{k} \sum_{Xi > X_{n-k:n}} \log\left(\frac{X_i}{X_{n-k:n}}\right) = \varphi(\mathbb{P}_n^{ext}),$$

avec $\varphi : P \mapsto \int \log \mathrm{d}P$.

### 3.2 Variations régulières multivariées

Une variable aléatoire $Z \in \mathbb{R}^d$ est dite à variations régulières si

$$\mathbb{P}\left(\left(\frac{Z}{\|Z\|}, \frac{\|Z\|}{y}\right) \in . \mid \|Z\| > y\right) \underset{y\to\infty}{\longrightarrow} \mu(.),$$

avec $\mu$ de la forme $\sigma \otimes Pareto_\alpha$ où $\sigma$ désigne une mesure sur la sphère unité de $\mathbb{R}^d$.
Avec un échantillon $(Z_1, \ldots, Z_n$ et en notant $Y_i = \|Z_i\|$ on peut estimer $\mu$ par

$$\mathbb{P}_n^{reg} := \frac{1}{k} \sum_{Y_i > Y_{n-k:n}} \delta_{\left(\frac{Z_i}{\|Z_i\|}, \frac{\|Z_i\|}{Y_{n-k:n}}\right)}.$$

Une fois encore $\mathbb{P}_n^{reg}$ satisfait l'hypothèse de mesure empirique conditionnel avec $U_n = Y_{n-k:n}$ et $P_u(A) = \mathbb{P}\left(\left(\frac{Z}{\|Z\|}, \frac{\|Z\|}{u}\right) \in A \mid \|Z\| > u\right)$ pour $A \in \mathcal{A}$ une tribu sur $\mathbb{R}^d$.
Étant donnée un échantillon $(Z_1, \ldots, Z_n)$, on souhaite tester ici si effectivement $Z$ est à variations régulières, c'est-à-dire, tester l'hypothèse

$$H_0 : \mu \text{ est à marginales indépendantes avec seconde marginal } Pareto_\alpha.$$

On construit alors, pour une fonction $\varphi$ bien choisie la statistique de test de type Kolmogorov-Smirnov

$$T_n = \|\varphi(\mathbb{P}_n^{reg})\|_\infty.$$

Ainsi, sous $H_0$, le Théorème 1 fournit la convergence en loi vers un certain $T$ dont la loi peut être approchée par bootstrap.

# Bibliographie


Balkema, A. A. and de Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, 2(5):792-804.
Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics*, 3(5):119-131.