Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework

Alexis Boulin, Elena Di Bernardino, Thomas Laloë, Gwladys Toulemonde

Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework

Alexis Boulin [1] & Elena Di Bernardino [1] & Thomas Laloë [1] & Gwladys Toulemonde [2]

[1] *Laboratoire J.A. Dieudonné, UMR CNRS 7351, Université Côte d'Azur, Nice, France*
[2] *Univ Montpellier, CNRS, Inria, Montpellier, France*

Résumé. Modéliser la dépendance entre maxima est un sujet d'intérêt dans les domaines d'application d'analyse du risque. Dans cet objectif, la copule de valeurs extrêmes, caractérisée par le madogramme, peut être utilisée comme une description de la structure de dépendance. Concrètement, la famille des distributions à valeurs extrêmes est très riche et survient naturellement comme la limite composante par composante des maxima préalablement normalisés. Dans cette présentation, nous étudions l'estimation non paramétrique du madogramme lorsque les données sont absentes complètement au hasard. Nous fournissons un théorème de la limite centrale fonctionnelle pour l'estimateur considéré du madogramme, correctement normalisé, vers un processus Gaussien tendu pour lequel la fonction de covariance dépend des probabilités de perte de la donnée. L'expression explicite de la variance asymptotique est aussi donnée. Nos résultats sont illustrés dans une étude numérique lorsque la taille d'échantillon est finie.

Mots-clés. Madogramme, Copule de valeurs extrêmes, Absence complètement au hasard, Estimation non paramétrique.

Abstract. The modeling of dependence between maxima is an important subject in several applications in risk analysis. To this aim, the extreme value copula function, characterised via the madogram, can be used as a margin-free description of the dependence structure. From a practical point of view, the family of extreme value distributions is very rich and arises naturally as the limiting distribution of properly normalised componentwise maxima. In this talk, we investigate the nonparametric estimation of the madogram where data are completely missing at random. We provide the functional central limit theorem for the considered multivariate madrogram correctly normalized, towards a tight Gaussian process for which the covariance function depends on the probabilities of missing. Explicit formula for the asymptotic variance is also given. Our results are illustrated in a finite sample setting with a simulation study.

Keywords. Madogram, Extreme value copula, Missing Completely At Random (MCAR), Nonparametric estimation.

1 Introduction

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé où vit $\mathbf{X} = (X_1, \ldots, X_d)$ un vecteur aléatoire à valeurs dans l'espace mesurable $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ avec $d \geq 2$. Ce vecteur aléatoire est distribué selon F et ses marges sont notées $F_j(x) = \mathbb{P}\{X_j \leq x\}$ pour tout $x \in \mathbb{R}$ et $j \in \{1, \ldots, d\}$. Une fonction $C : [0,1]^d \to [0,1]$ est une copule s'il s'agit de la restriction en $[0,1]^d$ d'une fonction de distribution dont ses marges sont distribuées uniformément le long du segment $[0,1]$. Depuis les travaux de Sklar, il est de connaissance commune que chaque fonction de distribution F peut être décomposée comme $F(\mathbf{x}) = C(F_1(x_1), \ldots, F_d(x_d))$ pour $\mathbf{x} \in \mathbb{R}^d$ et la copule C est unique si les marges sont continues.

L'étude de la notion de copule dans le cadre des extrêmes mène à la copule de valeurs extrêmes définie de la façon suivante :

$$C(\mathbf{u}) = \exp\left(-\ell(-\ln(u_1), \ldots, -\ln(u_d))\right), \quad \mathbf{u} \in (0,1]^d, \tag{1}$$

où $\ell : [0,\infty)^d \to [0,\infty)$ est la fonction de dépendance caudale qui est une fonction convexe, homogène d'ordre un et qui satisfait $\max(z_1, \ldots, z_d) \leq \ell(z_1, \ldots, z_d) \leq z_1 + \cdots + z_d$. Notons $\Delta^{d-1} = \{\mathbf{w} \in [0,1]^d, w_1 + \cdots + w_d = 1\}$ le simplexe unité de $\mathbb{R}^d$. Par son caractère homogène, la fonction ℓ est caractérisée par la fonction de dépendance de Pickands qui correspond à la restriction de ℓ au simplexe unité Δ^{d-1}, à proprement dit

$$\ell(z_1, \ldots, z_d) = (z_1 + \cdots + z_d) A(w_1, \ldots, w_d), \quad w_j = \frac{z_j}{z_1 + \cdots + z_d}.$$

Inspiré par la notion de variogramme en géostatistique, le λ-madogramme fut défini par [3] afin de capturer les dépendances extrémales bivariées de processus spatiaux. Cette définition fut ensuite généralisée en plus grandes dimensions via le **w**-madogramme définie dans [5] :

$$\nu(\mathbf{w}) = \mathbb{E}\left[\bigvee_{j=1}^{d} \{F_j(X_j)\}^{1/w_j} - \frac{1}{d}\sum_{j=1}^{d} \{F_j(X_j)\}^{1/w_j}\right], \quad \mathbf{w} \in \Delta^{d-1}, \tag{2}$$

si $w_j = 0$ et $0 < u < 1$, alors $u^{1/w_j} = 0$ par convention. Cette quantité caractérise la fonction de dépendance de Pickands via la relation formulée par la Proposition 2.2 de [5],

$$A(\mathbf{w}) = \frac{\nu(\mathbf{w}) + c(\mathbf{w})}{1 - \nu(\mathbf{w}) - c(\mathbf{w})}, \quad c(\mathbf{w}) = d^{-1}\sum_{j=1}^{d} w_j/(1 + w_j).$$

Dans cette communication, les principaux résultats obtenus que nous proposons de présenter sont la définition d'un estimateur du **w**-madogramme en (2) dans le cadre de données manquantes et l'étude de son comportement asymptotique. Un cadre multivarié ($d \geq 2$) avec données manquantes et une structure de dépendance décrite par une copule de

valeurs extrêmes seront considérés dans cette présentation. Nous évoquerons un théorème de la limite centrale fonctionnelle qui fournit la convergence faible du processus empirique centré du madogramme multivarié vers un processus Gaussien tendu dont la fonction de covariance va dépendre des probabilités d'observations, l'énonciation de ce théorème peut être retrouvée en [1]. Lorsque la trajectoire de ce processus empirique est fixée, nous montrons dans la Proposition 1 que cette variable aléatoire converge vers une loi normale centrée dont il est possible de fournir une expression explicite de la variance asymptotique. Enfin, des résultats numériques permettront de vérifier nos contributions sur un échantillon de taille finie.

2 Cadre de données manquantes

On suppose ici que la copule C est de valeurs extrêmes définie en (1). On considère un échantillon $\mathbf{X}_1, \ldots, \mathbf{X}_n$ de $\mathbf{X}$. En présence de données manquantes, on n'observe pas complètement le vecteur $\mathbf{X}_i$ pour $i \in \{1, \ldots, n\}$. Pour formaliser, on considère $\mathbf{I}_i \in \{0,1\}^d$ indiquant, $\forall j \in \{1, \ldots, d\}$, $I_{i,j} = 0$ si $X_{i,j}$ n'est pas observé et $I_{i,j} = 1$ si $X_{i,j}$ est observé. On introduit alors le vecteur aléatoire $\tilde{\mathbf{X}}_i$ à valeurs dans l'espace produit $\bigotimes_{j=1}^{d}(\mathbb{R} \cup \{\texttt{NA}\})$ tel que

$$\tilde{X}_{i,j} = X_{i,j} I_{i,j} + \texttt{NA}(1 - I_{i,j}), \quad i \in \{1, \ldots, n\},\ j \in \{1, \ldots, d\}.$$

On suppose alors que nous observons un $2d$-uplet tel que :

$$(\mathbf{I}_i, \tilde{\mathbf{X}}_i), \quad i \in \{1, \ldots, n\}. \tag{3}$$

Pour chaque $i \in \{1, \ldots, n\}$, $\mathbf{I}_i$ est distribuée identiquement et indépendamment selon $\mathbf{I} = (I_1, \ldots, I_d)$ où I_j est distribuée selon une variable aléatoire de Bernoulli avec probabilité $p_j = \mathbb{P}\{I_j = 1 | \mathbf{X}\}$ pour $j \in \{1, \ldots, d\}$. On note p la probabilité d'observer complètement une réalisation de $\mathbf{X}$, autrement dit $p = \mathbb{P}\{I_1 = 1, \ldots, I_d = 1 | \mathbf{X}\}$. Concernant le mécanisme de données manquantes, nous formulons l'hypothèse suivante :

Hypothèse A. *Les vecteurs $\boldsymbol{I}$ et $\boldsymbol{X}$ sont indépendants, ainsi*

$$p_j = \mathbb{P}\{I_j = 1\},\ j \in \{1, \ldots, d\}, \quad p = \mathbb{P}\{I_1 = 1, \ldots I_d = 1\}.$$

Sous l'Hypothèse A, nous pouvons définir des estimateurs consistants des distributions marginales et jointes. Par convenance, nous écrivons $\{\tilde{\mathbf{X}}_i \leq \mathbf{x}_i\} = \{\tilde{\mathbf{X}}_{i,1} \leq x_1, \ldots, \tilde{\mathbf{X}}_{i,d} \leq x_d\}$,

$$\hat{F}_{n,j}(x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\tilde{X}_{i,j} \leq x\}} I_{i,j}}{\sum_{i=1}^{n} I_{i,j}},\ \forall x \in \mathbb{R}, \quad \hat{F}_n(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\tilde{\mathbf{X}}_i \leq \mathbf{x}\}} \Pi_{j=1}^{d} I_{i,j}}{\sum_{i=1}^{n} \Pi_{j=1}^{d} I_{i,j}},\ \forall \mathbf{x} \in \mathbb{R}^d. \tag{4}$$

Avec ces notations, nous rappelons l'estimateur de la copule hybride formulée par [8], *i.e.*

$$\hat{C}_n^{\mathcal{H}}(\mathbf{u}) = \hat{F}_n(\hat{F}_{n,1}^{\leftarrow}(u_1), \ldots, \hat{F}_{n,d}^{\leftarrow}(u_d)), \quad \mathbf{u} \in [0,1]^d,$$

où $F^{\leftarrow}$ correspond à l'inverse généralisé de la fonction F, *i.e.* $F^{\leftarrow}(u) = \inf\{x \in \mathbb{R} | F(x) \geq u\}$ avec $0 < u < 1$. L'estimateur de la copule hybride est une extension de l'estimateur classique non paramétrique de la copule $\hat{C}_n$ (voir, *e.g.* [4] pour une définition) permettant que les marges de l'estimateur ne soient pas nécessairement égales aux estimateurs des marges ce qui est une propriété très appropriée dans le cadre plus général de données manquantes. Nous formulons une hypothèse permettant de garantir la convergence du présent processus empirique $\sqrt{n}(\hat{C}_n^{\mathcal{H}} - C)$:

Hypothèse B.

1. *La fonction de distribution F a des marges continues $F_1, \ldots, F_d$.*
2. *Pour chaque $j \in \{1, \ldots, d\}$, les dérivées partielles $\dot{\ell}_j$ de ℓ par rapport à x_j existent et sont continues sur l'ensemble $\{x \in [0, \infty)^d : x_j > 0\}$.*

L'hypothèse B1 est classique dans la littérature des copules sans laquelle une quelconque convergence faible du processus $\sqrt{n}(\hat{C}_n - C)$ ne peut être espérée. L'hypothèse B2 a été formulé par [7] permettant de généraliser la condition formulée par [4] demandant initialement que les dérivées partielles de la copule existent et soient continues sur $[0, 1]^d$. Des conditions plus faibles sont toujours le fruit de nouvelles recherches, voir par exemple [2] où l'hypothèse que l'ensemble des points où les dérivées partielles de la copule existent et sont continues soit de mesure de Lebesgue 1 est demandée. Nous proposons à présent un estimateur du $\mathbf{w}$-madogramme en (2) dans le cadre de données manquantes complètement au hasard.

Definition 1. *Soit $(\boldsymbol{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$ un échantillon donné par l'Equation* (3)*, l'estimateur hybride du $\mathbf{w}$-madogramme en* (2) *est défini par*

$$\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) = \frac{1}{\sum_{i=1}^n \Pi_{j=1}^d I_{i,j}} \sum_{i=1}^n \left[\left(\bigvee_{j=1}^d \left\{ \hat{F}_{n,j}(\tilde{X}_{i,j}) \right\}^{1/w_j} - \frac{1}{d} \sum_{j=1}^d \left\{ \hat{F}_{n,j}(\tilde{X}_{i,j}) \right\}^{1/w_j} \right) \Pi_{j=1}^d I_{i,j} \right], \tag{5}$$

où $\hat{F}_{n,j}(x)$ sont définis en (4)*.*

Comme classiquement rencontré dans la littérature, voir *e.g.* [6], cet estimateur ne vérifie pas la condition $\hat{\nu}_n(\mathbf{e}_j) = (d-1)/2d$ où $\mathbf{e}_j$ est un vecteur de la base canonique pour $j \in \{1, \ldots, d\}$. Néanmoins, une correction préalable permettant d'y remédier est définie ci-dessous.

Definition 2. *Soit $(\boldsymbol{I}_i, \tilde{\mathbf{X}}_i)_{i=1}^n$ un échantillon donné en* (3) *et $\hat{\nu}_n^{\mathcal{H}}$ l'estimateur hybride en* (5)*. Considérons par $\lambda_1, \ldots, \lambda_d : \Delta^{d-1} \to \mathbb{R}$ des fonctions continues vérifiant $\lambda_j(\boldsymbol{e}_k) = \delta_{jk}$ (delta de Kronecker) pour $j, k \in \{1, \ldots, d\}$, on définit la version corrigée de l'estimateur*

hybride du $\mathbf{w}$*-madogramme par*

$$\hat{\nu}_n^{\mathcal{H}*}(\mathbf{w}) = \hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \sum_{j=1}^{d} \frac{\lambda_j(\mathbf{w})(d-1)}{d} \left[\frac{1}{\sum_{i=1}^{n} \Pi_{j=1}^{d} I_{i,j}} \sum_{i=1}^{n} \left(\left\{ \hat{F}_{n,j}(\tilde{X}_{i,j}) \right\}^{1/w_j} \Pi_{j=1}^{d} I_{i,j} \right) - \frac{w_j}{1+w_j} \right]. \quad (6)$$

Sous les hypothèses A et B, nous démontrons que le procesus $\sqrt{n}(\hat{\nu}_n^{\mathcal{H}} - \nu)$ convergent faiblement vers un processus Gaussien dans $\ell^\infty(\Delta^{d-1})$, l'espace des fonctions bornées sur le simplexe, voir Théorème 1 de [1]. Lorsque $\mathbf{w} \in \Delta^{d-1}$ est fixé, en utilisant le précédent résultat, nous pouvons établir la convergence en loi des variables aléatoires $\sqrt{n}(\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w}))$ vers une Gaussienne centrée dont l'expression fermée est donnée. Les mêmes conclusions peuvent être obtenues pour $\sqrt{n}(\nu^{\mathcal{H}*} - \nu)$.

Proposition 1. *Soit* $\mathbf{p} = (p_1, \ldots, p_d, p)$ *et* $\mathbf{w} \in \Delta^{d-1}$*, sous les hypothèses A et B, nous avons*

$$\sqrt{n}\left(\hat{\nu}_n^{\mathcal{H}}(\mathbf{w}) - \nu(\mathbf{w})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \mathcal{S}^{\mathcal{H}}(\mathbf{p}, \mathbf{w})\right), \sqrt{n}\left(\hat{\nu}_n^{\mathcal{H}*}(\mathbf{w}) - \nu(\mathbf{w})\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \mathcal{S}^{\mathcal{H}*}(\mathbf{p}, \mathbf{w})\right).$$

De plus, l'expression des variances asymptotiques a été obtenue (voir [1]).

3 Résultats numériques

Figure 1: $\mathcal{E}_n^{\mathcal{H}}$ (première colonne) et $\mathcal{E}_n^{\mathcal{H}*}$ (deuxième colonne) données par (7), fonction de $\mathbf{w}$, pour le modèle logistique avec $\theta = 2$. Les courbes de niveaux de la variance asymptotique (à gauche) et des contreparties empiriques (à droite) sont représentées.

Dans cette section, nous illustrons nos résultats concernant l'expression de la variance asymptotique à travers des simulations numériques. Pour ce faire, nous comparons la variance asymptotique en Proposition 1 avec des contreparties empiriques obtenues via des simulations de Monte-Carlo. Nous considérons le modèle logistique défini par la fonction de dépendance suivante :

$$A(w_1, \ldots, w_d) = \left(\sum_{j=1}^{d} w_j^{\theta} \right)^{1/\theta}, \quad \theta \in [1, \infty).$$

Pour l'expérience numérique, la version corrigée du **w**-madogramme est calculée en utilisant $\lambda_j(\mathbf{w}) = \mathbf{e}_j$ et pour $d = 3$. La contrepartie empirique de la variance asymptotique donnée en Proposition 1 est calculée pour chaque élément **w** d'un certain grillage du simplexe Δ^{d-1}. Pour cet élément **w**, $n_{iter} = 100$ échantillons de taille $n = 512$ sont générés du modèle Logistique. Un estimateur du **w**-madogramme est ainsi associé à chaque échantillon. Nous calculons alors la variance empirique de l'erreur normalisée, à proprement dit :

$$\mathcal{E}_n^{\mathcal{H}}(\mathbf{w}) := \widehat{Var}\left(\sqrt{n}\left(\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}}}(\mathbf{w}) - \nu(\mathbf{w})\right)\right), \quad \mathcal{E}_n^{\mathcal{H}*}(\mathbf{w}) := \widehat{Var}\left(\sqrt{n}\left(\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}*}}(\mathbf{w}) - \nu(\mathbf{w})\right)\right), \quad (7)$$

où $\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}}}$ et $\hat{\boldsymbol{\nu}}_{\boldsymbol{n}}^{\boldsymbol{\mathcal{H}*}}$ sont les vecteurs composés des n_{iter} estimateurs hybrides et corrigées (voir (5) et (6)) du **w**-madogramme, respectivement. Les résultats de cette expérience sont données par la Figure 1.

Dans les deux cas, hybride comme corrigé, nous observons une adéquation entre la forme explicite de la variance asymptotique donnée à la Proposition 1 avec la contrepartie empirique obtenue via simulation.

References

[1] A. Boulin, E. D. Bernardino, T. Laloë, and G. Toulemonde. Non-parametric estimator of a multivariate madogram for missing-data and extreme value framework, 2021.

[2] A. Bücher, J. Segers, and S. Volgushev. When uniform weak convergence fails: Empirical processes for dependence functions and residuals via epi- and hypographs. *The Annals of Statistics*, 42(4):1598 – 1634, 2014.

[3] D. Cooley, P. Naveau, and P. Poncet. *Variograms for spatial max-stable random fields*, pages 373–390. Springer New York, New York, NY, 2006.

[4] J.-D. Fermanian, D. Radulović, and M. Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860, 2004.

[5] G. Marcon, S. Padoan, P. Naveau, P. Muliere, and J. Segers. Multivariate nonparametric estimation of the pickands dependence function using bernstein polynomials. *Journal of Statistical Planning and Inference*, 183:1–17, 2017.

[6] P. Naveau, A. Guillou, D. Cooley, and J. Diebolt. Modelling pairwise dependence of maxima in space. *Biometrika*, 96(1):1–17, 2009.

[7] J. Segers. Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, 18(3):764–782, 2012.

[8] J. Segers. Hybrid copula estimators. *J. Statist. Plann. Inference*, 160:23–34, 2015.