Inférence extrémale multivariée par mesures angulaires

Thomas Opitz

Inférence extrêmale multivariée par mesures angulaires

Thomas Opitz [1]

[1] *Institut de Mathématiques et de Modélisation, Université Montpellier II (thomas.opitz@math.univ-montp2.fr)*

Résumé. Nous traitons le problème de la validation de modèles fondés sur une propriété asymptotique de variation régulière multivariée. La modélisation suppose, après transformation des données en coordonnées pseudo-polaires définies à l'aide d'une fonction d'agrégation homogène pour le rayon et une norme pour l'angle, une queue de type Pareto pour le rayon et l'indépendance asymptotique entre rayons et angles. Ceci permet de représenter la structure de dépendance extrêmale par une mesure dite "angulaire". Dans ce contexte, nous définissons la notion de Loi Radiale de Pareto (LRP) et nous proposons des outils exploratoires et des tests statistiques afin de valider les hypothèses liées à cette approche et, le cas échéant, aux modèles paramétriques. Nous illustrons les outils en pratique à l'aide d'un jeu de données financières de rendements des indices boursiers DAX et CAC 40.

Mots-clés. données financières, extrêmes multivariés, indice de queue, loi de Pareto, qualité d'ajustement, variation régulière multivariée

Abstract. We tackle the problem of validating models motivated by the asymptotic property of multivariate regular variation. After transformation of data to pseudo-polar coordinates defined with respect to a radial aggregation function for the radius and a norm for the angle, we assume a radial Pareto tail and asymptotic independence between radii and angles, thus allowing to characterize the extremal dependence structure by a so-called angular measure. In this context, we define Radial Pareto Distributions and propose exploratory tools and statistical tests to validate the underlying assumptions of the approach and, if need be, of parametric models. A financial data set of DAX and CAC40 stock index returns serves to illustrate the tools in practice.

Keywords. financial data, multivariate extremes, Pareto distribution, goodness-of-fit, tail index, multivariate regular variation

1 Introduction

La théorie des valeurs extrêmes traite de l'inférence statistique pour évaluer, modéliser et prédire le comportement extrêmal de processus aléatoires. Ses applications se trouvent dans des domaines divers comme la climatologie, la finance, l'actuariat et l'analyse de réseaux de communication. Ce travail, dédié à l'analyse de la dépendance extrêmale dans un

cadre multivarié, concerne la validation de modèles fondés sur la propriété asymptotique de variation régulière multivariée. Cette approche est particulièrement adaptée à l'analyse de données à queue lourde pour lesquelles le maximum sur un ensemble d'observations peut jouer un rôle dominant parmi les observations, voire très dominant s'il est du même ordre de grandeur que la somme des observations. Justifiés asymptotiquement, les modèles utilisés en théorie des valeurs extrêmes permettent de faire des prédictions au-delà des plus grandes valeurs observées, c'est-à-dire pour des niveaux encore plus extrêmes.

Suivant le paradigme de la théorie des valeurs extrêmes, les approches inférentielles écartent les observations "ordinaires", ne travaillant que sur un échantillon de maxima par bloc ou un sous-ensemble d'observations dépassant un haut seuil. La modélisation d'extrêmes multivariés nécessite de trouver une correspondance entre la convergence des structures marginales et de dépendance et le choix du sous-ensemble extrême (maxima ou dépassements). Par conséquent, le choix de support pour les dépassements multivariés peut varier selon la finalité de la modélisation et selon la vitesse et la géometrie de convergence.

En approchant la distribution des données extrêmes par la mesure limite obtenue sous l'hypothèse de variation régulière multivariée, nous obtenons un modèle avec des queues de type Pareto pour les composantes marginales $X_i \geq 0, i = 1, ..., d$ et pour une variable agrégée $R = \text{rad}(\mathbf{X})$ définie à l'aide d'une fonction d'agrégation homogène $\text{rad}(t\mathbf{x}) = t^\beta \text{rad}(\mathbf{x})$ ($t > 0$ et $\mathbf{x} \geq 0$) d'ordre $\beta > 0$. Souvent, une norme est choisie pour rad. Nous appelons la valeur de cette fonction le *rayon* et la complétons avec l'*angle* $\mathbf{A} = \mathbf{X}/\|\mathbf{X}\|_{L^1}$ pour obtenir des coordonnées pseudo-polaires $(R, \mathbf{A}) \in [0, \infty) \times S_+$. Défini ainsi, ce modèle implique l'indépendance entre rayons et angles pour un support de dépassements radiaux d'un seuil $r_0 > 0$. Si $\delta > 0$ est l'indice de queue associé à $\mathbf{X}$ par la variation regulière multivariée (cf. Section 2), le modèle se caractérise par l'indice de queue $\delta/\beta > 0$ pour le rayon et par la mesure angulaire ρ sur S_+, vérifiant $\Pr(R \geq r, \mathbf{A} \in B) = r^{-\delta/\beta}\rho(B)$ pour $r > r_0$. L'angle $\mathbf{A}$ indique la contribution des composantes à l'événement agrégé $R = \text{rad}(\mathbf{X})$. Nous supposons que rad est chosie de telle sorte que $\rho(S_+) < \infty$.

Si l'échelle marginale des données empêche d'observer la variation régulière multivariée, une pré-transformation marginale permet de la faire apparaître. Cependant, l'interprétation des rayons et angles devient plus difficile à cause de cette échelle transformée.

Dans le cadre bivarié, la structure d'indépendance entre rayon et angle facilite l'analyse visuelle des données et la construction de tests statistiques. Des tests pour valider l'indépendance entre angles et rayons contre l'alternative d'une tendance dans les moments de la distribution des angles par rapport aux rayons sont proposés. Des tests d'ajustement de type Cramér-von-Mises sont aussi proposés pour la validation de modèles paramétriques cohérents avec une mesure angulaire empirique et pour l'adéquation à des LRP.

La monographie de Resnick (2007) donne une introduction à l'analyse de données à queue lourde. Beirlant et al. (2004) traite de l'inférence extrêmale dans un cadre plus vaste reposant sur la théorie des valeurs extrêmes classique.

Dans ce qui suit, nous rappelons la notion de variation régulière multivariée et nous définissons les LRP. Les nouveaux outils d'exploration et d'inférence sont ensuite présentés

de façon détaillée.

2 Variation régulière multivariée

Soit un vecteur aléatoire non-négatif $\mathbf{X} = (X_1, ..., X_d) \geq \mathbf{0}$ donné. Pour une norme $\|\cdot\|$, nous dénotons le *rayon* $R = \|\mathbf{X}\| > 0$ et l'*angle* $\mathbf{A} = \mathbf{X}/\|\mathbf{X}\|_{L^1} \in S_+ = \{\mathbf{x} \mid \mathbf{x} \geq 0, \sum_{i=1,...,d} x_i = 1\}$.

Le vecteur $\mathbf{X}$ est à variation régulière multivariée s'il existe une séquence $b_n \to \infty$ et une mesure limite non-triviale ν telles que

$$n\mathbb{P}(\mathbf{X}/b_n \in \cdot) \xrightarrow{v} \nu(\cdot) \ , \tag{1}$$

où "v" désigne la convergence vague (cf Resnick (2007)). La mesure limite ν est nécessairement homogène d'ordre $-\delta < 0$, avec δ l'*indice de queue* de $\mathbf{X}$. Il s'avère utile de passer dans un repère pseudo-polaire, résultant de la transformation $(r, \mathbf{a}) = T(\mathbf{x}) = (\|\mathbf{x}\|, \mathbf{x}/\|\mathbf{x}\|_{L^1})$. ν se factorise alors, $\nu(d(r, \mathbf{a})) = \delta r^{-\delta-1} dr \times \rho(d\mathbf{a})$, et la *mesure angulaire* ρ est une mesure de Radon finie. Réciproquement, une telle mesure est une mesure angulaire possible. Les queues marginales de la mesure ν vérifient $\nu(\{x_i > u_i\}) = c_i u_i^{-\delta}$, où c_i sont des constantes non-négatives. Dans la suite nous négligerons le cas $c_i = 0$ correspondant à une queue dominée par les autres. La mesure angulaire est sujette aux conditions de moments : $\int_{S_+} w_i^\delta \|\mathbf{w}\|^{-\delta} \rho(d\mathbf{w}) = c_i$, $i = 1, ..., d$.

Pour des réplications iid $\mathbf{X_i}$ de $\mathbf{X}$, le processus ponctuel $\{(R_i/b_n, \mathbf{A_i}), i = 1, 2, ..., n\}$ converge faiblement vers un processus de Poisson $\mathrm{PRM}(\alpha r^{-\alpha-1} dr \times \rho(d\mathbf{a}))$. Cette convergence est équivalente à la propriété (1). Alors

$$\{\mathbf{A_i} \mid R_i \geq c b_n\} \xrightarrow{v} \mathrm{PRM}(c^{-1}\rho(\cdot)) \quad \text{pour tout } c > 0. \tag{2}$$

3 Lois Radiales de Pareto (LRP)

En transformant la queue radiale d'une mesure limite ν en une mesure de probabilité, nous obtenons une LRP. Pour définir le rayon, nous admettons des fonctions d'agrégation homogène vérifiant $\mathrm{rad}(t\mathbf{x}) = t^\beta \mathrm{rad}(\mathbf{x})$ avec $\beta > 0$ pour tout $t > 0$ et $\mathbf{x} \geq 0$.

Définition 3.1 (Loi Radiale de Pareto). *Soit ρ une mesure de Radon non-nulle donnée sur le simplexe $S_+ = \{\mathbf{x} : \mathbf{x} \geq 0, \sum_{i=1}^d x_i = 1\}$. Nous appelons* Loi Radiale de Pareto *avec indice de queue $\alpha > 0$ la loi définie sur $\{(r, \mathbf{a}) \mid r \geq \rho(S_+)^{1/\alpha}, \mathbf{a} \in S_+\}$ par* $\mathrm{RP}(dr \times d\mathbf{a}) = \alpha r^{-\alpha-1} dr \times \rho(d\mathbf{a})$.

Nous utilisons la notation $\mathrm{RP}(\alpha, \rho)$. *Si le rayon r résulte de coordonnées Euclidéennes en utilisant la fonction d'agrégation* rad, *c.-à-d. $r = \mathrm{rad}(\mathbf{x})$, nous la dénotons* $\mathrm{RP}(\alpha, \mathrm{rad}, \rho)$.

3.1 Estimation

Dans les applications nous nous focalisons sur les événements extrêmes, en faisant l'hypothèse que les données sont issues d'un mélange avec comme composantes une LRP pour la queue et une composante censurée pour la partie centrale de la distribution. En remplaçant la convergence par une égalité dans (1) et en intégrant n et b_n dans ν, i.e. dans ρ, nous obtenons $\Pr((R, \mathbf{A}) \in d(r, \mathbf{a})) = \nu(d(r, \mathbf{a})) = (\alpha r^{-\alpha-1})\, dr \times \rho(d\mathbf{a})$. Pour $\mathbf{X_i}, i = 1, ..., n$ un échantillon et $r_0 > 0$ un seuil radial, nous considérons le sous-échantillon de dépassements radiaux en coordonnées pseudo-polaires $(R_i, \mathbf{A_i}), i = 1, ..., n_e < n$. Par une approche semi-paramétrique, nous estimons d'abord l'indice de queue α (estimateur de Hill) et procédons ensuite à l'estimation d'une mesure angulaire empirique qui attribue la masse $n^{-1} r_0^{\alpha}$ à chaque angle $\mathbf{A_i}$. Dans les approches paramétriques, la méthode de maximum de vraisemblance nous permet d'ajuster un modèle paramétrique de la mesure angulaire.

3.2 Analyse de l'indépendance rayon-angle

Nous transformons les n_e rayons R_i dépassant r_0 selon $F_{R|R \geq r_0}(R)$ vers une échelle uniforme $U_{R,i}$, soit par une transformation probabilité intégrale empirique, soit en utilisant l'indice de queue estimé $\widehat{\alpha}$ et $F_{R|R \geq r_0}(r) = 1 - (r_0^{-1} r)^{-\widehat{\alpha}}$.

Dans la suite, nous nous restreignons au cadre bivarié. S_+ est alors identifié à l'intervalle $[0, 1]$ et nous dénotons les angles $A_{1,i}$ par $A_i, i = 1, ..., n_e$. Le *diagramme à rayons uniformes* est obtenu en représentant les angles observés A_i contre les rayons uniformes $U_{R,i}$ pour $i = 1, ..., n_e$.

Pour tester l'indépendance angle-rayon contre une tendance des moments de A_i par rapport à R_i, la statistique $T_k = \sum_{i=1}^{n_e} A_i^k U_{R,i}$ est utilisée. Pour une tendance dans la variance, nous appliquons un estimateur local de la variance $v(A_i) = v(A_i \mid (A_1, ..., A_{n_e}), (U_{R,1}, ..., U_{R,n_e}))$, ce qui donne la statistique $T_v = \sum_{i=1}^{n_e} v(A_i) U_{R,i}$. Pour assurer des tests puissants et non-asymptotiques, nous testons conditionnellement aux angles observés. En calculant les statistiques de test pour un grand nombre de permutations aléatoires de $U_{R,i}$, nous obtenons facilement des quantiles approchés de la distribution sous H_0. Grâce à la structure simple des statistiques de test, le coût du calcul numérique n'est pas contraignant et une très bonne approximation des vrais quantiles est possible.

3.3 Tests d'ajustement

Nous définissons la statistique de type Cramér-von-Mises $T_{CvM1} = \int_0^1 (\widehat{\rho}_n(a) - \rho_{\widehat{\theta}}(a))^2 \rho_{\widehat{\theta}}(d\, a)$ en tant que distance entre une mesure angulaire empirique $\widehat{\rho}_n$ et une mesure candidate paramétrique $\rho_{\widehat{\theta}}$ afin de mesurer la qualité d'ajustement de cette dernière. Une expression

analytique de cette intégrale, similaire à la statistique classique de Cramér-von-Mises, nous permet une évaluation simple de quantiles numériques de T_{CvM1}. L'hypothèse nulle, justifiée par la convergence (2), est ici un processus binomial ou un processus de Poisson pour $\widehat{\rho}_n$.

Plus généralement, pour tester l'ajustement d'une LRP $\mathrm{RP}(\widehat{\alpha}, \mathrm{rad}, \widehat{\rho}_{\widehat{\theta}})$ à un jeu de données, nous transformons d'abord les rayons vers une échelle standard ($\alpha = 1$) et nous utilisons la fonction de répartition empirique $H_n(r,a) = n^{-1} \sum_{i=1}^{n_e} \mathbf{1}(A_i \leq a)\mathbf{1}(R_i \leq r)$. Ainsi,

$$T_{CvM2} = \int_{r_0}^{\infty} \int_{0}^{1} \left(H_n(r,a) - \widehat{\rho}_{\widehat{\theta}}(a) \left(\frac{1}{r_0} - \frac{1}{r} \right) \right)^2 \widehat{\rho}_{\widehat{\theta}}(da) \frac{dr}{r^2}$$

définit une statistique bivariée de type Cramér-von-Mises. L'expression explicite de cette statistique permet un calcul aisé de tout quantile.

Une géneralisation des notions introduites dans le cadre bivarié pour des données multivariées peut se faire à l'aide d'approches "par paires", par exemple en additionnant les statistiques de test pour toutes les paires de composantes du vecteur multivarié.

4 Application

Nous illustrons les notions introduites à l'aide d'un jeu de données de rendements hebdomadaire des indices boursiers DAX et CAC 40 entre 1990 et 2011 pour lequel une modélisation des pertes extrêmes est proposée. La Figure 1 rassemble une séléction de graphiques qui montrent les données, des diagrammes diagnostiques, le diagramme à rayons uniformes pour un sous-ensemble de données extrêmes et la mesure angulaire empirique cumulative correspondante. Tout en comportant beaucoup d'information, ces graphiques ne permettent pas de décider objectivement quant au choix d'un "bon" modèle. Nous montrerons comment les nouveaux outils proposés aident à pallier ce problème.

Bibliographie

[1] Beirlant, J. et al (2004), Statistics of extremes : Theory and applications, John Wiley & Sons Inc.

[2] Resnick, S.I. (2007), Heavy-tail phenomena : Probabilistic and statistical modeling, Springer.

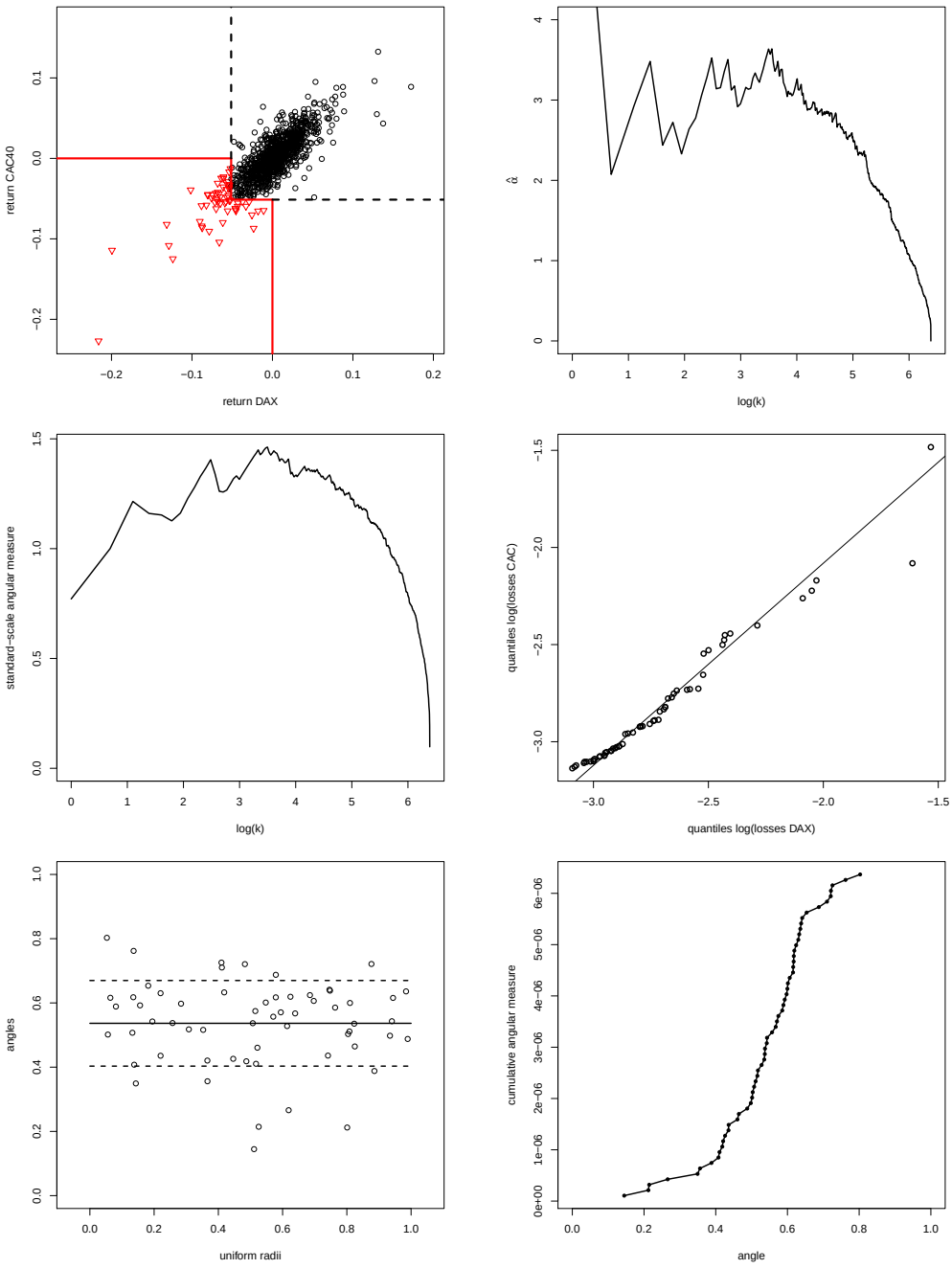

FIGURE 1 – Haut : Données avec un sous-ensemble de données extrêmes, Hill plot. Milieu : Diagramme exploratoire de stabilité de la mesure angulaire standard empirique, QQ-plot des dépassements marginaux pour les données extrêmes sélectionnées. Bas : Diagramme à rayons uniformes et mesure angulaire empirique cumulative pour les données sélectionnées.