INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2004-29

$g-$Divergence Empirique et Vraisemblance Empirique Généralisée

P. BERTAIL[1]
H. HARARI-KERMADEC[2]
D. RAVAILLE[3]

[1] CREST-LS.
[2] INRA-Corela, 65 Boulevard de Brandebourg, 94205 Ivry-sur-Seine Cédex, tél. : 01 49 59 69 58, fax : 01 46 59 69 90. Mail : harari@dptmaths.ens-cachan.fr
[3] ENS-CACHAN.

γ-Divergence empirique et vraisemblance empirique généralisée

Patrice BERTAIL, Hugo HARARI-KERMADEC et Denis RAVAILLE [1].

Résumé Dans cet article, nous généralisons les résultats obtenus avec la distance de Kullback (vraisemblance empirique) et les Cressie-Read (vraisemblance empirique généralisée) à des γ-divergences. Nous introduisons une famille Bartlett corrigeable, les Quasi-Kullback, barycentres de la distance de Kullback et du χ^2, qui ne sont pas du type Cressie-Read et qui possèdent d'intéressantes propriétés à distance finie. Nous concluons ce travail par des simulations de régions de confiance multidimensionnelles obtenues pour différentes divergences.

Abstract In this paper, we generalize the results obtained with the Kullback distance (corresponding to empirical likelihood) and Cressie-Read metrics (generalized empirical likelihood) to general γ-discrepancies, for some convex functions γ satisfying a few regularity properties. In particular, we introduce a new Bartlett correctable family of empirical discrepancies, the Quasi-Kullback, out of Cressie-Read family, which possess interesting finite sample properties. We conclude this work with some simulations in the multidimensional case for different discrepancies.

[1]P. Bertail : CREST-LS

H. HARARI-KERMADEC : INRA-Corela (65 Boulevard de Brandebourg 94205 Ivry-sur-Seine Cedex, tel : 01 49 59 69 58, fax : 01 49 59 69 90, harari@dptmaths.ens-cachan.fr) et CREST-LS,

D. RAVAILLE : ENS-CACHAN

1 Introduction

La méthode de vraisemblance empirique a été principalement introduite par Owen (1988, 1990, 2001), bien qu'on puisse la voir comme une extension des méthodes de calage (voir Deville & Särndal, 1992) utilisées depuis de nombreuses années en sondage notamment sous la forme (« model based likelihood ») introduite par Hartley & Rao (1968). Cette méthode de type non-paramétrique consiste à maximiser la vraisemblance d'une loi ne chargeant que les données, sous des contraintes satisfaites par le modèle (des contraintes de marges en sondage). Owen (1988, 1990) et de nombreux auteurs (voir Owen 2001 pour de nombreuses références) ont montré que l'on pouvait en effet obtenir dans ce cadre une version non-paramétrique du théorème de Wilks, à savoir la convergence du rapport de vraisemblance correctement renormalisé vers une loi du χ^2, permettant ainsi de réaliser des tests ou de construire des régions de confiance non-paramétriques pour certains paramètres du modèle. Cette méthode a été généralisée à de nombreux modèles économétriques, lorsque le paramètre d'intérêt est défini à partir de contraintes de moments (Qin & Lawless 1994, Newey & Smith 2003) et de manière générale est asymptotiquement valide pour tout paramètre multidimensionnel Hadamard différentiable (Bertail 2002, 2003). Elle se présente désormais comme une alternative à la méthode des moments généralisés.

Une interprétation possible de la méthode est de considérer celle-ci comme le résultat de la minimisation de la distance de Kullback entre la probabilité empirique des données $\mathbb{P}_n$ et une mesure (ou probabilité) $\mathbb{Q}$ dominée par $\mathbb{P}_n$ (ne chargeant donc que les points de l'échantillon), satisfaisant les contraintes, linéaires ou non, imposées par le modèle. L'utilisation de métriques différentes de la Kullback a été suggérée par Owen (1990) et de nombreux autres auteurs : parmi les métriques utilisées, on peut citer l'entropie relative étudiée par DiCiccio & Romano (1990) et Jing & Wood (1995) (qui a donné lieu à des développements en économétrie sous le nom de « Entropy econometrics », voir Golan et al. 1996) ou plus récemment les divergences de type Cressie-Read (Baggerly 1998, Corcoran 1998, Newey & Smith 2003, Bertail 2003) qui a donné lieu à des extensions économétriques sous le nom de « vraisemblances empiriques généralisées », bien que le caractère « vraisemblance » de la méthode soit perdue.

L'utilisation de métriques différentes de la Kullback pose à la fois des questions de généralisation et de choix des métriques en question. En particulier, on peut se demander :

1. Quels types de métriques permettent de conserver des propriétés similaires à la méthode originale de Owen (1988) ?

2. Il y a-t-il un avantage particulier à choisir une métrique plutôt qu'un autre, théoriquement ou algorithmiquement ?
3. Quelles sont les propriétés à distance finie de ces méthodes ?

L'objectif de ce travail est de répondre d'abord à la question 1 et de montrer que l'on peut obtenir par des arguments très simples des résultats généraux en remplaçant la distance de Kullback par une distance du type γ-divergence, pour toute fonction γ^* convexe satisfaisant certaines propriétés de régularité. Ces résultats ne sont pas spécifiques aux divergences de type Cressie-Read (invalidant ainsi une conjecture de Newey et Smith, 2003) et vont dans le sens des travaux obtenus indépendamment par Broniatowski & Kéziou (2003) pour des problèmes de tests paramétriques ou semiparamétriques. Nous montrons en particulier que les résultats obtenus sur les vraisemblances empiriques généralisées sont fortement liés, sous certaines conditions sur les fonctions γ^* considérées, aux propriétés de dualité convexe de ces métriques (cf. Rockafeller, 1970 et 1971), telles qu'elles sont étudiées par exemple par Borwein & Lewis (1991).

Nous discutons brièvement de la question 2 d'un point de vue de la théorie asymptotique en nous appuyant tout particulièrement sur les travaux de Mykland (1994), Baggerly (1998), Corcoran (1998) et Bertail (2002). D'un point de vue théorique, une des propriétés remarquables de la log-vraisemblance empirique est d'être, comme le log du rapport de vraisemblance dans les modèles paramétriques, corrigeable au sens de Bartlett, i.e. une correction explicite consistant à normaliser le log du rapport de vraisemblance par son espérance conduit à des régions de confiance possédant des propriétés au troisième ordre. On entend par là que l'erreur commise en utilisant la région de confiance asymptotique (i.e. ici la loi du χ^2) sur le niveau est de l'ordre de $\mathcal{O}(n^{-2})$. Cette propriété est en fait là encore essentiellement due aux propriétés de dualité convexe. Une lecture attentive de Corcoran (1998) montre que, parmi les divergence de type Cressie-Read, seule la vraisemblance empirique possède cette propriété mais que d'autres $\gamma-$divergences la possèdent également. Nous introduisons en particulier une famille de γ-divergences, barycentres de la distance de Kullback et du χ^2, qui sont Bartlett corrigeables (voir page 13). Une comparaison fine de ces γ-divergences nécessitent une analyse à l'ordre cinq i.e. jusqu'à l'ordre $\mathcal{O}(n^{-3})$ qui dépasse largement le cadre de cet article et dont on peut légitimement discuter l'intérêt.

Nous apportons quelques éléments de réponse à la question 3, en montrant que le comportement de ces statistiques est lié à celui des sommes autonormalisées dans le cadre des quasi-Kullback et par méthode de Monte-Carlo pour plusieurs divergences. Nous concluons ce travail par une étude par simu-

lations des zones de confiance (multidimensionnelles, $p = 2$) obtenues pour différentes divergences.

2 γ-divergences et dualité convexe.

Afin de généraliser la méthode de vraisemblance empirique, on rappelle quelques notions sur les γ-divergences (Csiszár 1967), dont nous donnerons quelques exemples (voir également Rockafeller 1970 ou Broniatowski & Kéziou 2003). Nous rappelons en annexe A quelques éléments de calcul convexe qui simplifient considérablement l'approche et les preuves. On pourra se référer à Rockafeller (1968, 1970 et 1971) et Liese & Vajda (1987) pour plus de précisions et un historique de ces métriques.

2.1 Cadre général

On considère un espace probabilisé $(\mathcal{X}, \mathcal{A}, \mathcal{M})$ où $\mathcal{M}$ est un espace de mesures signées et pour simplifier, $\mathcal{X}$ un espace de dimension finie muni de la tribu des boréliens. Le fait de travailler avec des mesures signées est fondamental comme nous le verrons dans les applications. Soit f une fonction mesurable définie de $\mathcal{X}$ dans $\mathbb{R}^p$. Pour toute mesure $\mu \in \mathcal{M}$, on note $\mu f = \int f(x)\mu(dx)$.

On utilise dans toute la suite la notation γ pour des fonctions convexes. On note $d(\gamma) = \{x \in \mathbb{R},\ \gamma(x) < \infty\}$ le domaine de γ et respectivement $\inf d(\gamma)$ et $\sup d(\gamma)$ les points terminaux de ce domaine. Pour toute fonction γ convexe, on introduit sa conjuguée convexe γ^* ou transformée de Fenchel-Legendre

$$\gamma^*(x) = \sup_{y \in \mathbb{R}} \{xy - \gamma(y)\} \ \forall\ x \in \mathbb{R}.$$

Nous ferons les hypothèses suivantes sur la fonction γ. Les hypothèses sur la valeur de γ en 0 correspondent essentiellement à une renormalisation (cf. Rao & Ren, 1991).

Hypothèses 2.1

(i) γ *est strictement convexe et* $d(\gamma) = \{x \in \mathbb{R},\ \gamma(x) < \infty\}$ *contient un voisinage de 0.*

(ii) γ *est deux fois différentiable sur un voisinage de 0.*

(iii) $\gamma(0) = 0$ *et* $\gamma^{(1)}(0) = 0$,

(iv) $\gamma^{(2)}(0) > 0$, *ce qui implique que* γ *admet un unique minimum en zéro.*

On a alors les propriétés classiques

Propriétés 2.1

(a) Par définition, γ^* est convexe et semi-continue inférieurement et de domaine de définition $d(\gamma^*)$ non vide si $d(\gamma)$ est non vide.

(b) Sous les hypothèses 2.1, la dérivée de γ est inversible et :

$$\gamma^*(x) = x.\gamma^{(1)\ -1}(x) - \gamma(\gamma^{(1)\ -1}(x)).$$

On en déduit $(\gamma^*)^{(1)} = \gamma^{(1)\ -1}$ *et* $(\gamma^*)^{(2)}(0) = \dfrac{1}{\gamma^{(2)}(0)}$.

Soit γ vérifiant les hypothèses (2.1). La γ-divergence associée à γ, appliquée à $\mathbb{Q}$ et $\mathbb{P}$, où $\mathbb{Q}$ (respectivement $\mathbb{P}$) est une mesure signée (respectivement une mesure signée positive), est définie par :

$$I_{\gamma^*}(\mathbb{Q},\mathbb{P}) = \begin{cases} \int_\Omega \gamma^*\left(\dfrac{d\mathbb{Q}}{d\mathbb{P}} - 1\right) d\mathbb{P} & \text{si } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{sinon.} \end{cases}$$

Ces pseudo-métriques introduites par Rockafellar (1968 et 1970) sont en fait des cas particuliers de « distances » convexes (Liese-Vajda, 1987). En tant que fonctionnelles sur des espaces de probabilité, elles sont également convexes et, vues comme des fonctionnelles sur des espaces de Orlicz (cf. Rao et Ren, 1991), elles satisfont des propriétés de dualités convexes (Rockafellar, 1971, Léonard, 2001). En particulier, l'intérêt des γ -divergences réside pour nous dans le théorème suivant (réécrit sous une forme simplifiée) du à Borwein & Lewis (1991) (voir également Léonard, 2001) qui résulte des propriétés des intégrales de fonctionnelles convexes.

Théorème 2.1 (Minimisation et Conjugaison)

Soit γ une fonction convexe partout finie et différentiable telle que $\gamma^* \geq 0$ et $\gamma^*(0) = 0$. Alors il vient

$$\inf_{\mathbb{Q}\in\mathcal{M}} (I_{\gamma^*}(\mathbb{Q},\mathbb{P})|\ (\mathbb{Q}-\mathbb{P})f = b_0) = \sup_{\lambda\in\mathbb{R}^p} (\lambda' b_0 - \int_\Omega \gamma(\lambda' f) d\mathbb{P}).$$

Si de plus, on a les contraintes de qualifications suivante, il existe $R \in \mathcal{M}$ telle que $Rf = b_0$ et

$$\inf d(\gamma^*) \ < \ \inf_\Omega \frac{dR}{d\mathbb{P}} \ \leq \ \sup_\Omega \frac{dR}{d\mathbb{P}} \ < \ \sup d(\gamma^*),$$

alors le sup *est atteint à droite en un point λ^* et l'* inf *à gauche en $\mathbb{Q}^*$ donné par*

$$\mathbb{Q}^* = (1 + \gamma^{(1)}((\lambda^*)' f)) \cdot \mathbb{P}.$$

2.2 Exemples

Nous donnons ici quelques exemples de γ-divergences qui sont utilisés pour généraliser la méthode de vraisemblance empirique.

2.2.1 Cressie-Read

Les distances les plus utilisées (Kullback, entropie relative, χ^2 et Hellinger) se regroupent dans la famille des Cressie-Read (voir Csiszár 1967 et Cressie & Read 1984). Le tableau 1 donne les fonctions γ et γ^* classiques, ainsi que leur domaine.

Divergences	α	γ_α		γ^*_α	
		$\gamma_\alpha(x)$	$d(\gamma_\alpha)$	$\gamma^*_\alpha(x)$	$d(\gamma^*_\alpha)$
entropie relative	1	$e^x - 1 - x$	$\mathbb{R}$	$(x+1)\log(x+1) - x$	$]-1, +\infty]$
Kullback	0	$-\log(1-x) - x$	$]-\infty, 1[$	$x - \log(1+x)$	$]-1, +\infty[$
Hellinger	0.5	$\dfrac{x^2}{2-x}$	$]-\infty, 2[$	$2(\sqrt{(x+1)} - 1)^2$	$]-1, +\infty[$
χ^2	2	$\dfrac{x^2}{2}$	$\mathbb{R}$	$\dfrac{x^2}{2}$	$\mathbb{R}$

TAB. 1 – Les principales Cressie-Read

Dans le cas général, les Cressie-Read s'écrivent

$$\gamma^*_\alpha(x) = \frac{(1+x)^\alpha - \alpha x - 1}{\alpha(\alpha-1)},\ \gamma_\alpha(x) = \frac{[(\alpha-1)x+1]^{\frac{\alpha}{\alpha-1}} - \alpha x - 1}{\alpha}$$

$$I_{\gamma^*_\alpha}(\mathbb{Q}, \mathbb{P}) = \frac{1}{\alpha(\alpha-1)} \int_\Omega \left[\left(\frac{d\mathbb{Q}}{d\mathbb{P}} \right)^\alpha - \alpha \left(\frac{d\mathbb{Q}}{d\mathbb{P}} - 1 \right) - 1 \right] d\mathbb{P}.$$

Si on suppose que $\mathbb{Q}$ est dominée par $\mathbb{P}$, et que $\mathbb{Q}(\Omega) = \mathbb{P}(\Omega)$, on peut simplifier l'écriture de l'intégrale $I_{\gamma^*_\alpha}(\mathbb{Q},\mathbb{P}) = \dfrac{1}{\alpha(\alpha-1)}\int_\Omega \left[\left(\dfrac{d\mathbb{Q}}{d\mathbb{P}}\right)^\alpha - 1\right] d\mathbb{P}$. On notera que cette forme simplifiée oblige à tenir compte de la contrainte supplémentaire sur la masse de $\mathbb{Q}$, ce qui n'est pas nécessaire si on travail avec la forme initiale.

2.2.2 Normes $\mathbb{L}^p$

Dans le cas des distances $\mathbb{L}^p$, correspondant à $\gamma^*(x) = |x|^p$, on peut poser

$$\gamma(x) = \begin{cases} \frac{p}{q}(\frac{|x|}{p})^q & \text{sur } x \in \mathbb{R}, \text{ si } p > 1 \\ \frac{-p}{q}(\frac{-x}{p})^q & \text{sur } x \in \mathbb{R}^*_-, \text{ si } 0 < p < 1 \end{cases} \text{ où } \frac{1}{p} + \frac{1}{q} = 1.$$

Il vient alors immédiatement $I_{\gamma^*}(\mathbb{Q},\mathbb{P}) = (\mathbb{1}_{p>1}(p) - \mathbb{1}_{0<p<1}(p))\mathcal{L}_p(\mathbb{Q},\mathbb{P})$, où $\mathcal{L}_p(\mathbb{Q},\mathbb{P}) = \int_\Omega \dfrac{|l-k|^p}{k^{p-1}} d\lambda$ si $\mathbb{Q} = l.\lambda$ et $\mathbb{P} = k.\lambda$.

Cependant, si $p \neq 2$, $\gamma^{(2)}(0) = 0$ et, comme nous le verrons plus loin, la fonctionnelle n'étant plus quadratique au voisinage de 0. Les lois limites ne sont dès lors plus du type χ^2 mais du type Chaos de Wiener.

3 Extension de la méthode de vraisemblance empirique aux γ-divergences.

L'objectif de ce chapitre est d'étendre la méthode de vraisemblance empirique à des γ-divergences autres que celle de Kullback ou les Cressie-Read, et de montrer en quoi les résultats obtenus par Owen (1990) et tous ceux récemment obtenus dans la littérature économétrique sont essentiellement liés aux propriétés de convexité de la fonctionnelle I_{γ^*}. Nous nous restreignons ici au cas de la moyenne multivariée pour simplifier l'exposition et les preuves, mais les résultats sont également valides pour des contraintes de moments plus générales en nombres finis.

3.1 Vraisemblance empirique

On considère une suite de vecteurs aléatoires $X, X_1, \cdots, X_n$ de $\mathbb{R}^p$, $n \geq 1$, indépendants et uniformément distribués de loi de probabilité P dans un espace de probabilité $\mathcal{P}$. On note Pr la probabilité sous la loi jointe $P^{\otimes n}$ de $(X_1, ..., X_n)$. On cherche alors à obtenir une région de confiance pour

$T(P) = \int X dP$ sous l'hypothèse que $V_P(X)$ est une matrice définie positive. Pour cela dans l'optique traditionnelle de Von Mises, on construit la probabilité empirique $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ avec $\delta_x(y) = \mathbb{1}_{\{y=x\}}$, qui est l'estimateur du maximum de vraisemblance non-paramétrique de P, dans le sens où elle maximise, parmi les lois possibles, la fonctionnelle $L(\mathbb{Q}) = \prod_{i=1}^n \mathbb{Q}(\{x_i\})$ où $\forall i \in \{1, ..., n\}$, $\{x_i\}$ représente le singleton $x_i = X_i(\omega)$, $\omega \in \Omega$. On ne s'intéresse donc ici qu'à l'ensemble $\mathcal{P}_n$ des probabilités $\mathbb{Q}$ dominées par $\mathbb{P}_n$, c'est-à-dire de la forme, $\mathbb{Q} = \sum_{i=1}^n q_i \delta_{X_i}$, $q_i \geq 0$, $\sum_{i=1}^n q_i = 1$.

On définit une région de confiance selon le principe de la vraisemblance empirique, comme suit

$$\begin{aligned} C_{\eta,n} &= \left\{ T(\mathbb{Q}) \,\middle|\, \mathbb{Q} \ll \mathbb{P}_n, \mathbb{Q} \in \mathcal{P}, R_n(\mathbb{Q}) = \frac{L(\mathbb{Q})}{L(\mathbb{P}_n)} \leq \eta \right\} \\ &= \left\{ T(\mathbb{Q}) \,\middle|\, \mathbb{Q} = \sum_{i=1}^n q_i \delta_{X_i}, q_i \geq 0, \sum_{i=1}^n q_i = 1,\ R_n(\mathbb{Q}) = \frac{\prod_{i=1}^n q_i}{\prod_{i=1}^n \frac{1}{n}} \leq \eta \right\}, \end{aligned}$$

où η est déterminé par la précision $1-\alpha$ que l'on veut atteindre pour la région de confiance : $\Pr(T(\mathbb{Q}) \in C_{\eta,n}) = 1-\alpha$. L'intérêt de la définition de $C_{\eta,n}$ vient de l'observation suivante de Owen (1988) :

$$\Pr(T(\mathbb{Q}) = \mu \in C_{\eta,n}) = \Pr(\eta_n(\mu) \geq \eta),$$

$$\text{avec } \eta_n(\mu) = \sup_{\mathbb{Q} \in \mathcal{P}_n, T(\mathbb{Q}) = \mu} R_n(\mathbb{Q}) = \frac{\sup_{\{\mathbb{Q} \in \mathcal{P}_n, T(\mathbb{Q}) = \mu\}} L(\mathbb{Q})}{\sup_{\mathbb{Q} \in \mathcal{P}_n} L(\mathbb{Q})},$$

qui s'interprète clairement comme un rapport de vraisemblance. Un estimateur de μ est alors donné en minimisant le critère $\eta_n(\mu)$

$$\widehat{\mu}_n = \arg\min_{\mu}(\eta_n(\mu))$$

Owen (1988, 1991 et 2001) a montré que $2\beta_n(\mu) = -2\log(\eta_n(\mu))$ converge vers une loi du $\chi^2(p)$. Ceci permet d'obtenir des intervalles de confiance asymptotiques. En effet, il vient $\Pr(\mu \in C_{\eta,n}) = \Pr(\beta_n(\mu) \leq -\log(\eta))$. On en déduit que pour $\eta = \exp(-\frac{\chi^2_{1-\alpha}}{2})$, $C_{\eta,n}$ est asymptotiquement de niveau $1-\alpha$. Or, on a clairement

$$\beta_n(\mu) = n \inf_{\{\mathbb{Q} \in \mathcal{P}_n, T(\mathbb{Q}) = \mu\}} \{K(\mathbb{Q}, \mathbb{P}_n,)\}.$$

Cette présentation suggère la généralisation suivante.

3.2 Minimisation empirique des γ-divergences

On définit désormais pour une fonction γ donnée, $\beta_n^\gamma(\mu)$ comme le minimum de la γ-divergence empirique associée, contrainte par la valeur μ de la fonctionnelle et la région de confiance C_{η,n,γ^*} correspondante soit

$$\beta_n^\gamma(\mu) = n \inf_{\{\mathbb{Q}<<P_n, T(\mathbb{Q})=\mu\}} \{I_{\gamma^*}(\mathbb{Q}, \mathbb{P}_n)\}$$
$$C_{\eta,n,\gamma^*} = \{\mu \mid \exists \mathbb{Q},\ T(\mathbb{Q}) = \mu,\ \mathbb{Q} \ll \mathbb{P}_n \text{ et } I_{\gamma^*}(\mathbb{Q}, \mathbb{P}_n) \leq \eta\}.$$

Nous expliquerons plus loin, pourquoi on n'impose pas que $\mathbb{Q}$ soit une probabilité mais plutôt une mesure signée dans $\mathcal{M}_n = \{\mathbb{Q},\ \mathbb{Q} \ll \mathbb{P}_n\}$. Ceci s'explique en partie par le théorème 2.1, qui donne des conditions d'existence de solutions seulement pour des mesures signées. Le fait de ne pas imposer que la mesure soit de masse 1 facilite l'optimisation, mais demande de reformuler la contrainte sur le paramètre recherché. En effet, on a $\mathbb{E}_\mathbb{Q}[\mu] = \mathbb{Q}(1) \cdot \mu$. On écrit alors le paramètre d'intérêt $T(\mathbb{Q}) = \frac{\mathbb{E}_\mathbb{Q}[X]}{\mathbb{Q}(1)}$.

Intuitivement, pour généraliser de la méthode de vraisemblance empirique, on considère la valeur empirique de la fonctionnelle définie par

$$M(R, \mu) = \inf_{\{\mathbb{Q} \ll R,\ T(\mathbb{Q})=\mu\}} \{I_{\gamma^*}(\mathbb{Q}, R)\}$$

pour $R \in \mathcal{P}$, i.e. la minimisation d'un contraste sous les contraintes imposées par le modèle. Si le modèle est vrai, i.e. $T(P) = \mu$ pour la probabilité P sous-jacente, alors on a clairement $M(P, \mu) = 0$. Un estimateur de $M(P, \mu)$ à μ fixé est simplement donné par l'estimateur plugin $M(P_n, \mu)$, qui n'est rien d'autre que $\beta_n^\gamma(\mu)/n$. Cet estimateur peut donc permettre de tester $M(P, \mu) = 0$ ou dans une approche duale de construire une région de confiance pour μ.

On suppose que γ satisfait les hypothèses suivantes :

Hypothèses 3.1 *(i) γ vérifie les hypothèses 2.1,*

(ii) La dérivée seconde de γ est minorée par $m > 0$ sur $d(\gamma) \cap \mathbb{R}^+ (\neq \emptyset)$.

Il est simple de vérifier que les fonctions et divergence données dans la partie précédente vérifient cette hypothèse supplémentaire. L'hypothèse (ii) est vérifiée en particulier lorsque $\gamma^{(1)}$ est elle-même convexe (entraînant $\gamma^{(2)}(x)$ croissante donc $\geq \gamma^{(2)}(0) > 0$ sur $\mathbb{R}^+$), ce qui est le cas pour toutes les divergences étudiées ici. Pour le cas de la moyenne et pour $\mathbb{Q}$ dans $\mathcal{M}_n$ (et donc de masse 1), on peut réécrire les contraintes de minimisation sous la forme

$$[T(\mathbb{Q}) - T(\mathbb{P}_n)].(X - \mu) = \mu - \bar{\mu}.$$

Il vient

$$\begin{aligned}\beta_n^\gamma(\mu) &= n \inf_{\mathbb{Q}\in\mathcal{M}_n,\, T(\mathbb{Q})=\mu} \{I_{\gamma^*}(\mathbb{Q},\mathbb{P}_n)\} \\ &= n \inf_{\mathbb{Q}\in\mathcal{M}_n} \{I_{\gamma^*}(\mathbb{Q},\mathbb{P}_n)|\ [T(\mathbb{Q})-T(\mathbb{P}_n)]\cdot(X-\mu)=\mu-\bar{\mu}\} \\ &= n \sup_{\lambda\in\mathbb{R}^p} \left\{ \lambda'(\mu-\bar{\mu}) - \int_\Omega \gamma(\lambda'(X-\mu)) d\mathbb{P}_n \right\}.\end{aligned}$$

On en déduit l'expression duale de $\beta_n^\gamma(\mu)$ qui permet de dériver les propriétés usuelles des vraisemblances empiriques ainsi que leur généralisation :

$$\beta_n^\gamma(\mu) = \sup_{\lambda\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \lambda'(\mu-X_i) - \sum_{i=1}^n \gamma(\lambda'(X_i-\mu)) \right\}. \qquad \text{(Dual)}$$

On a alors le

Théorème 3.1 *Si $X_1,\cdots,X_n$ sont des vecteurs aléatoires de $\mathbb{R}^p$, i.i.d. de loi P absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$, de moyenne μ et de variance $V_P(X)$ de rang q et si γ vérifie les hypothèses 3.1 alors, $\forall\ 0<\alpha<1$, et pour $\eta = 2\gamma^{(2)}(0)\chi_q^2(1-\alpha)$, C_{η,n,γ^*} est convexe et*

$$\begin{aligned}\lim_{n\to\infty} \Pr(\mu \notin C_{\eta,n,\gamma^*}) &= \lim_{n\to\infty} \Pr(\beta_n^\gamma(\mu)\geq\eta) \\ &= \Pr(Z \geq \frac{\eta}{2\gamma^{(2)}(0)}) = 1-\alpha.\ \textit{où}\ Z\sim\chi_q^2.\end{aligned}$$

Remarque 3.1 *Supposons que nous voulions mener le même raisonnement en y intégrant les contraintes $\mathbb{Q}(1)=1$ et $\mathbb{Q}\geq 0$, forçant la mesure à être une probabilité. Alors les contraintes de qualification peuvent ne jamais être vérifiées et le problème dual peut ne pas avoir de solutions. Par exemple, en prenant la divergence du χ^2, c'est-à-dire $\gamma(x)=\frac{x^2}{2}$, la contrainte supplémentaire conduit au problème de minimisation*

$$\min_{\{q_i\}\in\mathbb{R}^p} \left\{ \chi^2(\mathbb{P}_n,\mathbb{Q}) \left| \sum_{i=1}^n q_i X_i = \mu,\ \sum_{i=1}^n q_i = 1, q_i \geq 0 \right. \right\}.$$

Le calcul du Lagrangien correspondant montre facilement que cela implique $\mu = \frac{1}{n}\sum_{i=1}^n X_i = \bar{\mu}$: il n'existe donc pas de solution au problème.

Remarque 3.2 *$\gamma^{(2)}(0)\neq 0$ assure le comportement quadratique de la $\gamma-$divergence empirique au voisinage de 0 et explique son comportement χ^2. Si ce n'est pas le cas, la limite dépend des dérivées d'ordre supérieur de γ. Une telle étude peut faire l'objet de travaux ultérieurs. Dans tous nos exemples, les divergences sont normalisées de telle sorte que $\gamma^{(2)}(0)=1$.*

3.3 Vraisemblance empirique : la Kullback.

Dans le cas particulier, où $\gamma^*(x) = x - \log(1+x)$, et en imposant $q_i \geq 0$ (ce qui est de toute façon vérifié à cause du domaine de γ^*), la forme duale obtenue en (Dual) se réécrit $\beta_n(\mu) = \sup_{\lambda \in \mathbb{R}^p} \{\sum_{i=1}^n \log(1 + \lambda'(X_i - \mu))\}$. Comme le remarque Bertail (2002, 2003), cette quantité est elle-même un rapport de log-vraisemblance paramétrique indexée par le paramètre λ (pour tester $\lambda = 0$), qui peut également être vue comme une vraisemblance duale au sens de Mykland (1995). Il est donc immédiat d'obtenir dans ce cas, que le rapport de vraisemblance est asymptotiquement $\chi^2(p)$ pourvu que la variance de $(\mu - X_i)$ soit définie positive. En tant que vraisemblance paramétrique, elle est aussi Bartlett corrigeable, un point souligné à l'origine par Hall, DiCiccio et Romano (1991). Dans la représentation duale, la preuve de la correction au sens de Bartlett devient triviale. De manière générale pour une divergence quelconque, la forme duale n'est pas une vraisemblance. Néanmoins nous proposons plus loin une famille de divergences, les Quasi-Kullback, qui peuvent être Bartlett corrigeables car proche d'une vraisemblance dans leur forme duale.

3.4 Les Cressie-Read

Les résultats établis pour les Cressie-Read (voir Newey & Smith, 2003) dont on rappelle ici la forme

$$\gamma^*_\alpha(x) = \frac{(1+x)^\alpha - \alpha x - 1}{\alpha(\alpha - 1)}$$

s'obtiennent facilement en appliquant le théorème (3.1). Les hypothèses du théorème (2.1) sont vérifiées pour calculer la valeur de la vraisemblance empirique en un point μ dès qu'il existe une famille de poids $\{q_i\}_{1..n}$ avec (au pire) $\forall i,\ q_i > -1$, telle que $\sum_{i=1}^n q_i X_i = \mu \sum_{i=1}^n q_i$. Ceci traduit qu'il existe une solution au problèmes primal et dual au moins pour tous les points μ de l'enveloppe convexe des X_i. Pour les Cressie-Read autre que la distance du χ^2 on peut imposer $q_i > 0$. On peut alors appliquer le théorème 2.1 et l'on obtient l'expression de $\mathbb{Q}$

$$\mathbb{Q} = \sum_{i=1}^n q_i \delta_{X_i} = \sum_{i=1}^n \frac{1}{n} [1 + (\alpha - 1)\lambda'_n(X_i - \mu)]^{\frac{1}{\alpha - 1}} \delta_{X_i}$$

d'où l'on déduit la valeur du rapport de « vraisemblance » généralisée

$$I_{\gamma^*_\alpha}(\mathbb{Q}, \mathbb{P}_n) = \sum_{i=1}^n \frac{1}{n} \gamma^*_\alpha(nq_i - 1) = \sum_{i=1}^n \frac{(nq_i)^\alpha - \alpha n q_i + \alpha - 1}{n\alpha(\alpha - 1)}$$

On regroupe les valeurs des poids et des divergences pour les exemples usuels ($\alpha = 0, \frac{1}{2}, 1, 2$) dans le tableau 2.

Divergences	**poids**(q_i)	**minimum de la divergence**
entropie relative	$\frac{1}{n}\exp(\lambda_n'(X_i - \mu))$	$\sum q_i \log(nq_i) + 1 - \sum q_i$
Kullback	$\frac{1}{n(1 - \lambda_n'(X_i - \mu))}$	$-1 - \sum \frac{1}{n}\log(nq_i) + \sum q_i$
Hellinger	$\frac{4}{n(2 - \lambda_n'(X_i - \mu))^2}$	$2\sum \left(\sqrt{q_i} - \sqrt{\frac{1}{n}}\right)^2$
χ^2	$\frac{1}{n}(1 + \lambda_n'(X_i - \mu))$	$\sum \frac{(nq_i - 1)^2}{2n}$

TAB. 2 – Poids et rapports de vraisemblances pour les divergences usuelles

On notera que dans le cas particulier du χ^2, on peut calculer le multiplicateur de Lagrange et donc la valeur de la vraisemblance en un point μ. D'un point de vue algorithmique, c'est donc la plus simple. On obtient en effet par un calcul direct $\lambda_n = S_n^{-1}(\mu - \bar{\mu})$ avec $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu).(X_i - \mu)'$ d'où $q_i = \frac{1}{n}(1 + (\mu - \bar{\mu})' S_n^{-1}(X_i - \mu))$ et

$$I_{\gamma_2^*}(\mathbb{Q}, \mathbb{P}_n) = \sum_{i=1}^n \frac{(n \cdot q_i - 1)^2}{2n} = \frac{1}{2}(\mu - \bar{\mu})' S_n^{-1}(\mu - \bar{\mu}),$$

qui s'interprète directement comme une somme vectorielle autonormalisée.

3.5 Les Polylogarithmes

La famille des Cressie-Read ne contient pas toutes les γ-divergences pour lesquelles il existe une représentation duale explicite. Considérons par exemple

la famille basée sur les Polylogarithmes. Les Polylogarithmes, définis par :

$$Li_\alpha(x) = \sum_{k\geq 1} \frac{x^k}{k^\alpha}$$

sont liés aux fonctions Gamma Γ et zeta de Riemann ζ. Si, $\forall \alpha \in [-1; +\infty[$, on pose

$$h_\alpha(x) = 2^{\alpha-1}(Li_\alpha(x) - x) = \frac{x^2}{2} + 2^{\alpha-1} \sum_{k\geq 3} \frac{x^k}{k^\alpha}$$

on obtient une famille de fonctions définies sur $]-1, 1[$ (au moins) et qui vérifient toutes nos hypothèses 3.1. On a en particulier $h_0(x) = \dfrac{x^2}{2(1-x)}$, $h_1(x) = -\ln(1-x) - x = \gamma_0(x)$ et $h_2(x) = \int_0^x \int_0^t \dfrac{2ds}{1+e^{-s}} dt$. On peut expliciter la conjuguée convexe de h_0, qui correspond à la distance de Hellinger à une similitude près $h_0^*(x) = (\sqrt{1+x} - 1)^2$.

3.6 Quasi-Kullback

Nous introduisons maintenant une famille de divergences qui possèdent des propriétés de type Lipschitz intéressantes :

$$\forall \varepsilon \in [0; 1],\ \forall x \in]-\infty; 1[\ , \qquad K_\varepsilon(x) = \varepsilon \frac{x^2}{2} + (1-\varepsilon)(-x - \log(1-x)).$$

L'idée est de conserver les avantages de la distance de Kullback tout en évitant les problèmes algorithmiques liés au comportement du log au voisinage de 0. Cette famille vérifie nos hypothèses 3.1 et l'on peut expliciter K_ε^* :

$$K_\varepsilon^*(x) = -\frac{1}{2} + \frac{(2\varepsilon - x - 1)\sqrt{1 + x(x + 2 - 4\varepsilon)} + (x+1)^2}{4\varepsilon} - (\varepsilon - 1) \log \frac{2\varepsilon - x - 1 + \sqrt{1 + x(x + 2 - 4\varepsilon)}}{2\varepsilon}$$

de dérivée seconde $K_\varepsilon^{*(2)}(x) = \dfrac{1}{2\varepsilon} + \dfrac{2\varepsilon - x - 1}{2\varepsilon\sqrt{1 + 2x(1-2\varepsilon) + x^2}}$.

Ces expressions permettent de démontrer très facilement que $K_\varepsilon^{(2)}(x) \geq \varepsilon$ et $0 \leq K_\varepsilon^{*(2)} \leq 1/\varepsilon$. Algorithmiquement, on est alors assuré d'une meilleure convergence. On peut par ailleurs obtenir aisément le résultat suivant.

Théorème 3.2 *Sous les hypothèses du théorème 3.1, on a les inégalités suivantes* ***à n fixe*** *:*

$$\begin{aligned}\Pr(\mu \notin C_{\eta,n,\gamma^*}) &= \Pr(\beta_n^{\gamma}(\mu) \geq \eta) \\ &\leq \Pr\left(\frac{n}{2\varepsilon}(\mu-\bar{\mu})' S_n^{-1}(\mu-\bar{\mu}) \geq \eta\right)\end{aligned}$$

Si $X_i - \mu$ est de loi symétrique, alors sans aucune hypothèse de moments,

$$\Pr(\mu \notin C_{\eta,n,\gamma^*}) \leq 2\exp(-\eta\varepsilon).$$

Par ailleurs, si on choisit pour ε la suite $\varepsilon_n = \mathcal{O}(n^{-3/2}\log(n)^{-1})$, les Quasi-Kullback sont corrigeables au sens de Bartlett, jusqu'à l'ordre $\mathcal{O}(n^{-3/2})$.

Remarque 3.3 *Cette inégalité exponentielle est classique et remonte à des travaux de Efron (1969). La première partie du théorème implique que pour toute la classe des Quasi-Kullback, le comportement de la divergence empirique se ramène à l'étude d'une somme autonormalisée. L'études de cette quantité fait actuellement l'objet de nombreux travaux : voir Götze & Chistyakov (2003) et Jing & Wang (1999). Les bornes obtenues par ces auteurs sont meilleures que celle de Efron et donnent un contrôle plus précis de l'approximation par une loi du χ^2. Malheureusement ces bornes ne sont pas universelles et dépendent des moments d'ordres supérieurs ($\mathbb{E}|X_i|^3$ voire $\mathbb{E}|X_i|^{10/3}$). On a par exemple, si $\beta_{10/3} = \mathbb{E}|X|^{10/3} < \infty$, pour $p = 1$ et avec $A \in \mathbb{R}$ et $\theta \in]0,1[$, on a dans le cas général non-symétrique*

$$\Pr\left(\frac{n}{2}(\mu-\bar{\mu})' S_n^{-1}(\mu-\bar{\mu}) \geq \eta\varepsilon\right) = \Pr(\chi^2 \geq \eta\varepsilon) + A\beta_{10/3} n^{-1/2} e^{-\theta\eta\varepsilon}. \quad \text{(Dev)}$$

Ceci permet d'obtenir des intervalles de confiance uniformes pour les Quasi-Kullback sous l'hypothèse de bornitude de $\beta_{10/3}$ par une valeur M donnée.

Remarque 3.4 *Le choix de ε permettant la correction au sens de Bartlett n'est sans doute pas optimale mais permet de simplifier considérablement les preuves. Une lecture attentive des travaux de Corcoran (2001) qui donnent des conditions nécessaires de corrigeabilité au sens de Bartlett (pour des divergences ne dépendant pas de n, ce qui n'est pas le cas ici), permettent également de montrer que si ε est petit, alors la statistique est corrigeable au sens de Bartlett mais ne permettent pas précisément de calibrer ε. Nous conjecturons qu'une vitesse en $o(n^{-1})$ est suffisante. Ceci fera l'objet de travaux ultérieurs. Il convient de noter que les choix de ε selon que l'on veuille obtenir une bonne borne exponentielle (obtenue pour $\varepsilon = 1$ i.e. le cas χ^2) ou la correction au sens de Bartlett (correction asymptotique du biais) ne sont pas compatibles : il s'agit d'approches différentes.*

4 Simulations et comparaisons

Cette partie présente quelques résultats de simulations dans le cas multivarié, pour différentes métriques. Nous comparons les zones de confiance obtenues. Les simulations et les graphiques ont été réalisés à l'aide du logiciel Matlab : les algorithmes sont disponibles auprès des auteurs.

4.1 Exemple introductif : données uniformes

Les données sont ici des v.a. uniformes sur le carré $[0,1]^2$. On a choisi de représenter les zones de confiance à 90%, 95% et 99% pour les divergences de Kullback, d'Hellinger, du χ^2 et de l'entropie relative (figure 1).

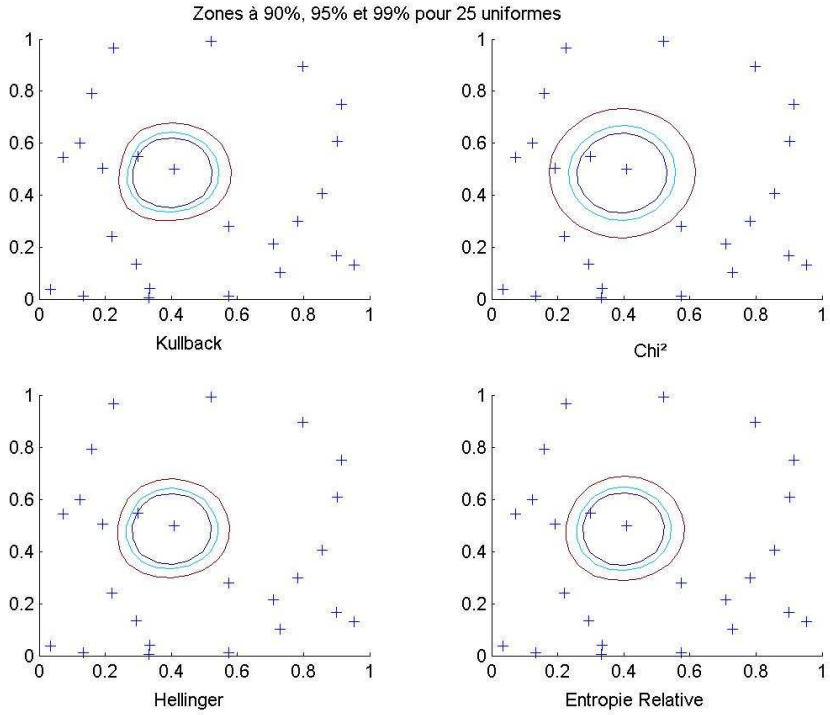

FIG. 1 – Zones de confiance pour 4 divergences avec les mêmes données

Naturellement, les zones grandissent avec la confiance et sont convexes. On remarque tous de suite que la figure obtenue pour le χ^2 se distingue : les zones sont circulaires et particulièrement grandes. Il s'agit en fait d'un cas particulier à bien des égards. La divergence du χ^2, point fixe de la conjugaison convexe, donne toujours des régions en ellipses, ici très proche de cercles car

nos données sont réparties indépendamment et de la même façon suivant des deux axes. On peut aussi remarquer que ces zones sortent de l'enveloppe convexe des observations pour le χ^2, et même du support de la loi, le carré $[0,1]^2$. C'est le cas pour des valeurs de α proche de 1 ou des données en petit nombre, comme pour la figure 3. Ceci tient au fait que certains poids sont négatifs. En effet, avec cette divergence, si l'on impose à la mesure $\mathbb{Q}$ d'être une probabilité, il n'existe pas de solution à la minimisation. Pour certaines valeurs de μ, pourtant hors de l'enveloppe convexe des données, on peut alors trouver des mesures $\mathbb{Q}$ telles que la divergence reste acceptable.

Il peut être judicieux de faire varier le nombre de données, pour observer la variation des surfaces en fonction de ce nombre.

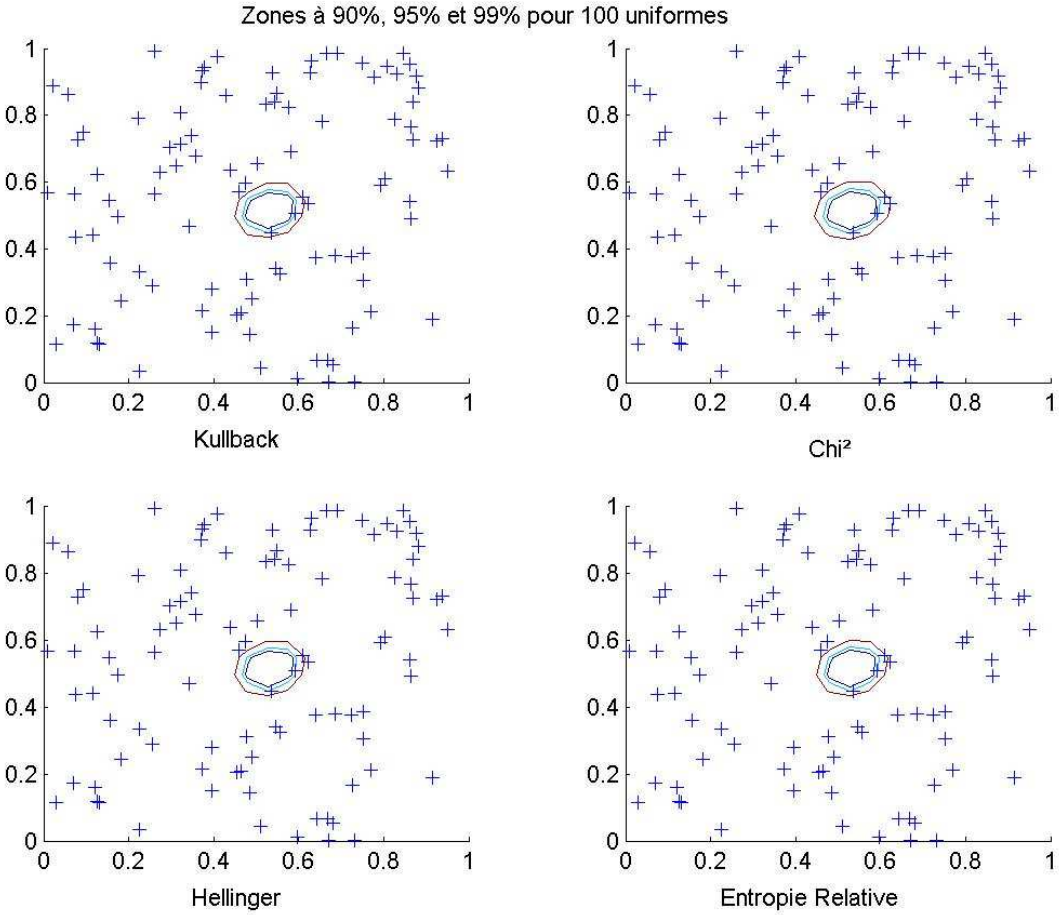

FIG. 2 – Zones de confiance pour 100 données uniformes

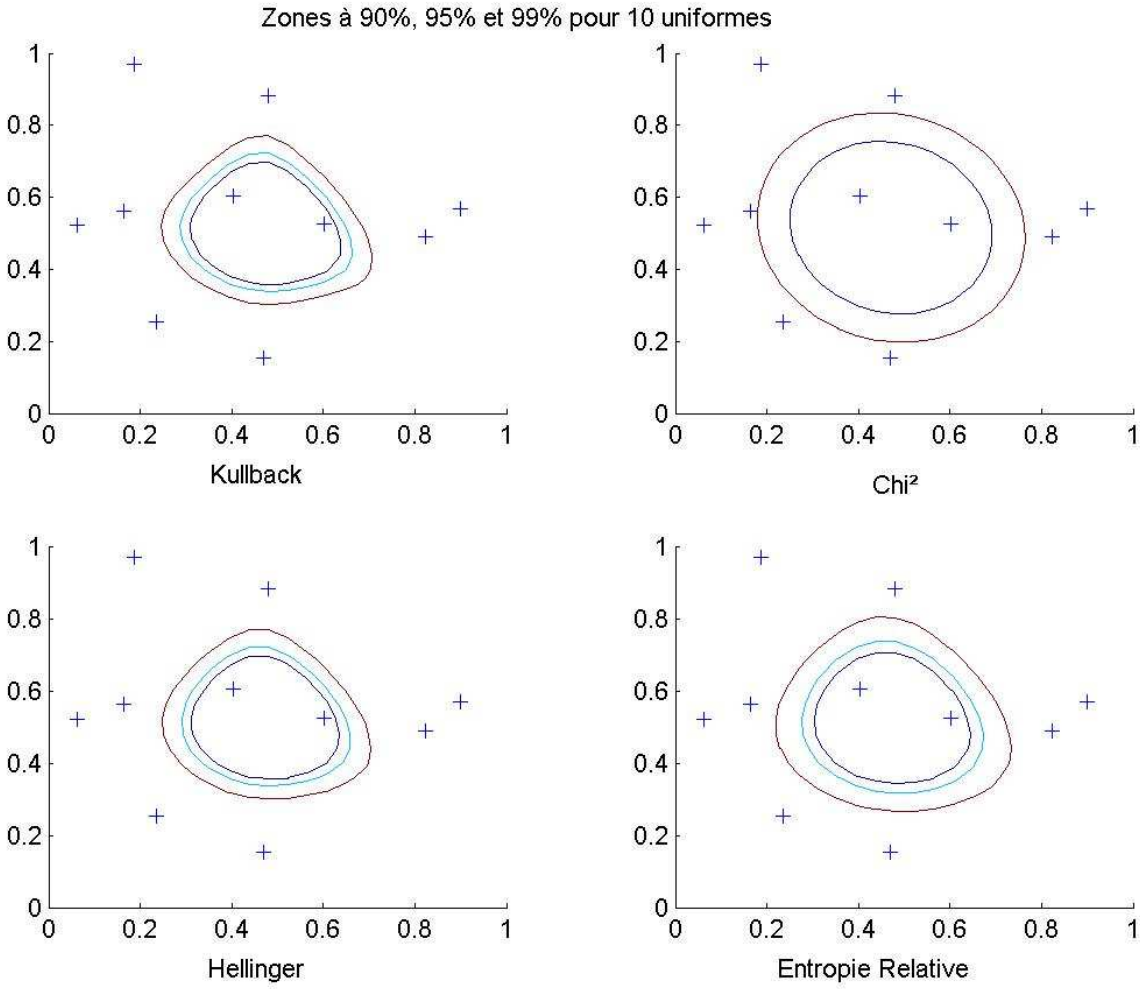

FIG. 3 – Zones de confiance pour 10 données uniformes

Pour les données en faible quantité, on note que les zones s'adaptent bien au données, comme c'est classique pour la vraisemblance empirique. Il y a bien sûr une particularité pour le χ^2, qui ne peut qu'adapter les axes de l'ellipse, et explose si les données sont en très faible quantité (dans la figure 3, la ligne de niveau correspondant à 99% est hors du cadre, et est donc invisible).

4.2 Sur le choix de la divergence

On peut étudier des données de tous types avec nos 4 divergences, pour essayer de mieux saisir leurs différences. Il apparaît sur nos simulations, ce

qui est conforme à la théorie (comportement asymptotique), que dès que le nombre de données est important (supérieur à 25), les surfaces sont fortement similaires. On a représenté dans la figure 4 les zones de confiance pour 50 données générées par 4 lois :

- gaussienne,
- exponentielle (de paramètre 1),
- Marshall-Olkin (ce copule a une dépendance entre les 2 coordonnées qui fait apparaître une courbe exponentielle) et
- copule (de type 't' et de paramètre 0.9 et 4, ce qui charge beaucoup la diagonale).

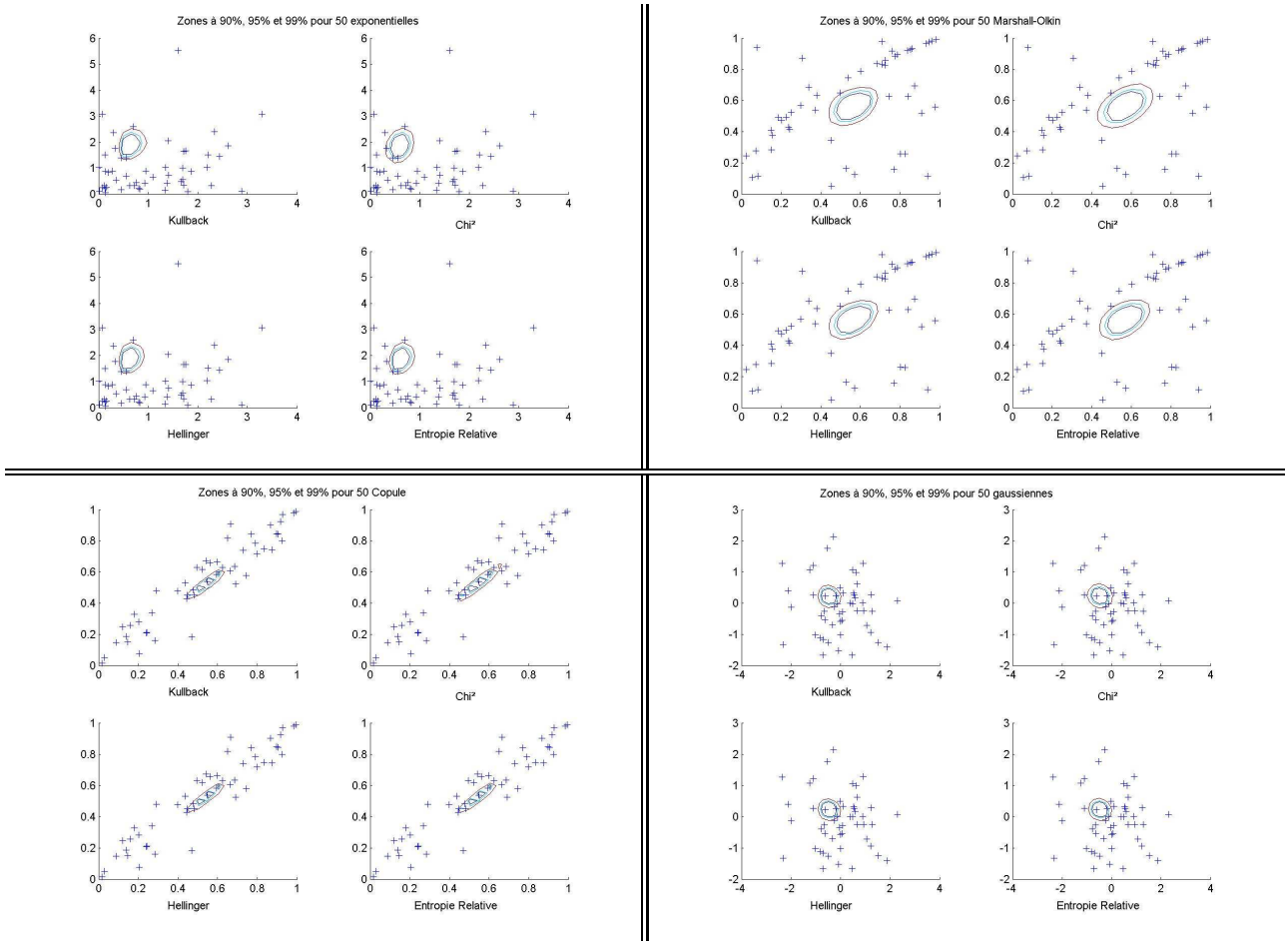

FIG. 4 – Zones de confiance pour différentes lois

Remarque 4.1 *Nos simulations de copules de Marshall-Olkin sont réalisées à partir de Embrechts, Lindskog & McNeil (2001), où l'on trouve une bonne introduction aux copules en général, celles concernant les t-copules utilisent*

le programme MATLAB de Peter Perkins, matlabcode for copulas.

Comme l'on dispose d'une formule de calcul directe de la vraisemblance pour le χ^2, on gagne beaucoup en vitesse de calcul en utilisant cette divergence pour effectuer des tests ou construire des régions de confiance.

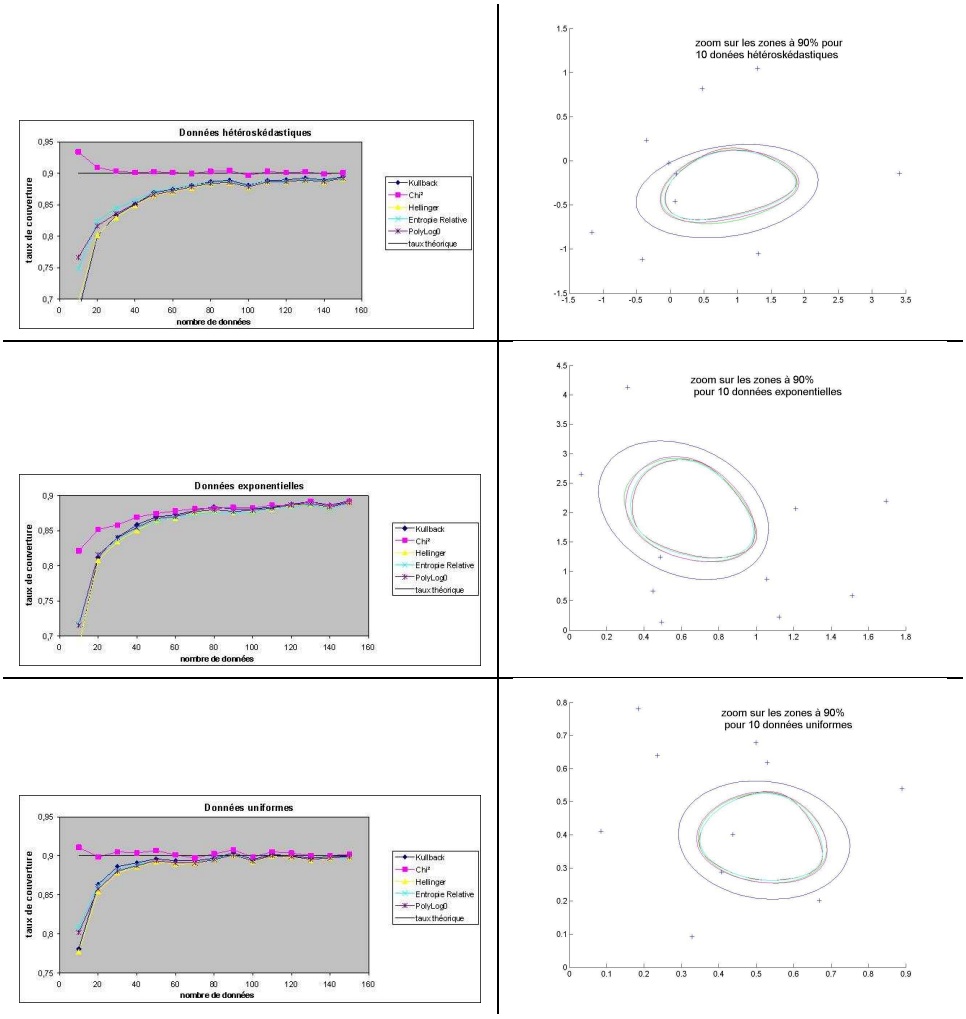

FIG. 5 – Évolution des taux de couverture

On a représenté dans la figure 5 l'évolution en fonction de n des niveaux de confiance estimé par une méthode de type Monte-Carlo (10000 répétitions des expériences) pour 5 distances (Kullback, χ^2, Hellinger, Entropie relative et PolyLog0) et 3 types de données (de mélange, exponentielles et uniformes). Les données que nous appelons « de mélange » ou « hétéroscédastiques » consistent en un produit d'une gaussienne centrée réduite sur $\mathbb{R}^2$ et d'une

uniforme sur $[0, 2]$. Le graphique donne un exemple de zones de confiance. Les taux de couverture sont bien meilleurs pour le χ^2, mais avec une surface significativement plus grande. On notera cependant que c'est cette méthode qui est la plus robuste, ce qui s'explique par le caractère autonormalisé de la somme dans sa version duale et par les résultats obtenus en 3.2.

5 Conclusion

Ce travail généralise la validité des raisonnements sur la vraisemblance empirique menés par Owen (1990) aux γ-divergences. On gagne un choix étendu de métriques pour une méthode qui ne suppose que peu de choses sur le modèle sous-jacent aux données. Ce travail propose des pistes pour choisir la divergence en fonction du nombre de données et pour éviter les problèmes d'implémentation, en particulier en introduisant la famille des Quasi-Kullback.

A Quelques éléments de calcul convexe.

Le lemme suivant simplifie la recherche des fonctions convexes dont on connaît la conjuguée et réciproquement (voir Borwein & Lewis 1991) :

Lemme A.1 (Involutivité de la conjugaison) *Quelle que soit γ convexe, il y a équivalence entre les assertions suivantes :*

(i) $\gamma = (\gamma^*)^*$,

(ii) γ *est fermée, c'est-à-dire que son graphe* $\mathcal{G} = \{(x, \gamma(x)),\ x \in \mathbb{R}\}$ *est fermé,*

(iii) γ *est semi-continue inférieurement.*

Grâce à ces méthodes, il est assez simple d'obtenir les tableaux 3 et 4 qui permettent de calculer les conjuguées convexes utilisées dans ce travail.

Fonction h	$f(x)$	$f(ax)$	$f(x+b)$	$af(x)$
Fonction h^*	$g(x) = f^*(x)$	$g\left(\frac{x}{a}\right)$	$g(x) - bx$	$ag\left(\frac{x}{a}\right)$
validité		$\forall a \neq 0$	$\forall b$	$\forall a > 0$

TAB. 3 – Propriétés élémentaires

$f = g^*$		$g = f^*$	
Fonction	$d(f)$	Fonction	$d(g)$
0	$[-1,1]$	$\mid x\mid$	$\mathbb{R}$
0	$[0,1]$	x^+	$\mathbb{R}$
$\frac{\lvert x\rvert^p}{p}\ \forall p>1$	$\mathbb{R}$	$\frac{\lvert x\rvert^q}{q}$ pour $\frac{1}{p}+\frac{1}{q}=1$	$\mathbb{R}$
$-\frac{\lvert x\rvert^p}{p}\ \forall p\in]0,1[$	$\mathbb{R}_+$	$-\frac{(-x)^q}{q}$	$-\mathbb{R}_+^*$
$\sqrt{(1+x^2)}$	$\mathbb{R}$	$-\sqrt{(1-x^2)}$	$[-1,1]$
$-\log(x)$	$\mathbb{R}_+^*$	$-1-\log(-y)$	$-\mathbb{R}_+^*$
$\log(\cos(x))$	$\left]-\frac{\pi}{2},\frac{\pi}{2}\right[$	$\frac{x}{\tan(x)}-\frac{1}{2}\log(1+x^2)$	$\mathbb{R}$
e^x	$\mathbb{R}$	$\begin{cases} x\log(x)-1 & \text{si } x>0 \\ 0 & \text{si } x=0 \end{cases}$	$\mathbb{R}_+$
$\log(1+e^x)$	$\mathbb{R}$	$\begin{cases} x\log(x)+(1-x)\log(1-x) & \text{si } x\in]0,1[\\ 0 & \text{si } x\in\{0,1\} \end{cases}$	$[0,1]$
$-\log(1-e^x)$	$\mathbb{R}_+^*$	$\begin{cases} x\log(x)-(1+x)\log(1+x) & \text{si } x>0 \\ 0 & \text{si } x=0 \end{cases}$	$\mathbb{R}_+$

TAB. 4 – Tableau des principales conjuguées convexes

B Démonstrations des résultats

B.1 Preuve du Théorème 3.1

La preuve suit les grandes lignes de Owen (1990). Tout d'abord, si $q < p$, on peut, par une nouvelle paramétrisation, se ramener à des données de taille q, ce qui permet de démontrer le résultat si on le prouve pour $q = p$. La convexité de C_{η,n,γ^*} découle immédiatement de celle de γ^* et de la linéarité de l'intégrale. Comme on a supposé les X_i de loi continue, les X_i sont distincts avec probabilité 1.

Pour démontrer le théorème 3.1, il nous faut nous assurer de l'existence de $\eta_n(\mu)$ et de $\beta_n^\gamma(\mu)$, au moins pour certains μ. D'après Owen (2000), chap. 11, p. 217 pour $\mathcal{S}$ la sphère des vecteurs unités de $\mathbb{R}^p$,

$$\inf_{\theta\in\mathcal{S}} \Pr[(X-\mu)'\theta > 0] > 0.$$

On pose alors $\varepsilon = \inf_{\theta\in\mathcal{S}} \Pr[(X-\mu)'\theta > 0]$ et on remarque que Glivenko-Cantelli nous donne

$$\sup_{\theta\in\mathcal{S}} [\Pr - \mathbb{P}_n][(X-\mu)'\theta > 0] \xrightarrow[n\to\infty]{\Pr \text{ p.s.}} 0$$

d'où l'on déduit $\inf_{\theta\in\mathcal{S}} \mathbb{P}_n[(X-\mu)'\theta > 0] > \frac{\varepsilon}{2}$ pour n assez grand. Ceci signifie qu'en prenant n assez grand, tout μ appartenant à l'enveloppe convexe des points formés par l'échantillon est admissible.

On rappelle également le lemme suivant : voir Owen (2001),

Lemme B.1 (Négligeabilité) *Soit $(Y_i)_{i=1}^n$ une suite de variables aléatoires indépendantes et identiquement distribuées et $\forall\ n \in \mathbb{N}$, $Z_n = \max_{i=1,..,n} |Y_i|$. Si $\mathbb{E}[Y_1^2] < \infty$, alors, en probabilité, $Z_n = o(n^{1/2})$ et $\frac{1}{n}\sum_{i=1}^n |Y_i|^3 = o(n^{1/2})$.*

Rappelons que le programme d'optimisation dual d'après (2.1) vaut

$$\beta_n^\gamma(\mu) = \sup_{\lambda\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \lambda'(\mu - X_i) - \sum_{i=1}^n \gamma(\lambda'(X_i - \mu)) \right\}. \qquad \text{(Dual)}$$

La condition au premier ordre impliquée par (Dual) nous permet d'affirmer que la dérivée par rapport à λ^j de la partie droite est nulle, pour $j \in \{1, ..., p\}$. Il vient les conditions suivantes

$$\begin{aligned} \forall\ j \in \{1,...,p\} \quad 0 &= \sum_{i=1}^n (X_i^j - \mu^j)[1 + \gamma^{(1)}(\lambda'(X_i - \mu))] \\ \text{donc} \quad 0 &= g(\lambda) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)\left[1 + \gamma^{(1)}(\lambda'(X_i - \mu))\right] \end{aligned}$$

On pose $Y_i = X_i - \mu$ et l'on note λ_n le λ réalisant le sup dans (Dual), on a donc $g(\lambda_n) = 0$. On pose $\lambda_n = \rho_n \theta_n$ avec $\rho_n \geq 0$ et $\|\theta_n\|_2 = 1$,

$$\begin{aligned} 0 &= \theta_n' g(\lambda_n) \\ -\theta_n' \bar{Y} &= \frac{1}{n}\sum_{i=1}^{n} \theta_n' Y_i \cdot \gamma^{(1)}(\lambda_n' Y_i). \end{aligned}$$

Un développement de Taylor de $\gamma^{(1)}$ au voisinage de 0 donne

$$\gamma^{(1)}(\rho_n \theta_n' Y_i) = \rho_n \theta_n' Y_i \cdot \gamma^{(2)}(\rho_n t_i),$$

avec t_i entre 0 et $\theta_n' Y_i$. On a alors sous l'hypothèse (ii) de 3.1

$$\begin{aligned} -\theta_n' \bar{Y} &= \rho_n \frac{1}{n}\sum_{i=1}^{n} (\theta_n' Y_i)^2 \cdot \gamma^{(2)}(\rho_n t_i) \\ &\geq \rho_n \frac{1}{n} \sum_{i:\theta_n' Y_i \geq 0}^{n} (\theta_n' Y_i)^2 \cdot \gamma^{(2)}(\rho_n t_i) \\ &\geq m\rho_n \frac{1}{n} \sum_{i:\theta_n' Y_i \geq 0}^{n} (\theta_n' Y_i)^2. \end{aligned}$$

Or, $\dfrac{1}{n} \displaystyle\sum_{i:\theta_n' Y_i \geq 0}^{n} (\theta_n' Y_i)^2$est minorée. En effet, en raisonnant par l'absurde, et en remarquant que θ_n prend ses valeurs dans un compact, on peut extraire une sous-suite telle que $\dfrac{1}{n} \displaystyle\sum_{i:\theta_n' Y_i \geq 0}^{n} (\theta_n' Y_i)^2 \to 0$ et $\theta_n \to \theta_0$. On a alors $\mathbb{E}_{\mathbb{P}}[(\theta_0' Y_i)^2 \mathbb{1}_{\theta_0' Y_i \geq 0}] = 0$, ce qui contredit que Σ est inversible.

Le théorème centrale limite implique que $-\theta_n' \bar{Y} = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$. Par suite, il vient $\|\lambda_n\|_2 = \rho_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$. On définit alors $\tilde{\lambda}_n = \lambda_n + \frac{S_n^{-1}(\bar{\mu}-\mu)}{\gamma^{(2)}(0)}$. En effectuant un développement de Taylor de γ en 0 dans l'expression de $g(\lambda_n)$, il vient

$$0 = \gamma^{(2)}(0) S_n \tilde{\lambda}_n + \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)\alpha_{i,n} \ ,$$

où, uniformément en i,

$$\|\alpha_{i,n}\| \leq B|\lambda_n'(X_i - \mu)| \leq B\|\lambda_n\| Z_n = o_{\mathbb{P}}(1),$$

car $Z_n = \max_{i=1,..,n} \|X_i - \mu\| = o_{\mathbb{P}}(n^{1/2})$ d'après le lemme B.1. Finalement, comme S_n est minorée et que $\bar{\mu} - \mu = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, on a $\tilde{\lambda_n} = o_{\mathbb{P}}(n^{-1/2})$.

De même, en effectuant un développement limité de γ autour de 0 dans l'expression de $\beta_n^{\gamma}(\mu)$, il vient

$$
\begin{aligned}
\beta_n^\gamma(\mu) &= -n.\lambda_n'(\bar{\mu}-\mu) - \textstyle\sum_{i=1}^n \gamma\left(\lambda_n'(X_i-\mu)\right) \\
&= -n.\lambda_n'(\bar{\mu}-\mu) - \textstyle\sum_{i=1}^n \left(\dfrac{(\lambda_n'(X_i-\mu))^2}{2}\gamma^{(2)}(0) + \tilde{\alpha}_{i,n} \right) \\
&= -n.\lambda_n'(\bar{\mu}-\mu) - \dfrac{\gamma^{(2)}(0)}{2}(n\lambda_n' S_n \lambda_n) - \textstyle\sum_{i=1}^n \tilde{\alpha}_{i,n} \\
&= -n\lambda_n'(\bar{\mu}-\mu) - \textstyle\sum_{i=1}^n \tilde{\alpha}_{i,n} \\
&\quad -n\dfrac{\gamma^{(2)}(0)}{2}\left(\tilde{\lambda}_n' S_n \tilde{\lambda}_n - \dfrac{2}{\gamma^{(2)}(0)}\tilde{\lambda}_n'(\bar{\mu}-\mu) + \dfrac{(\bar{\mu}-\mu)' S_n^{-1}(\bar{\mu}-\mu)}{\gamma^{(2)}(0)^2} \right) \\
&= \dfrac{n}{\gamma^{(2)}(0)}(\bar{\mu}-\mu)' S_n^{-1}(\bar{\mu}-\mu) - \textstyle\sum_{i=1}^n \tilde{\alpha}_{i,n} \\
&\quad -\dfrac{\gamma^{(2)}(0)}{2} n\tilde{\lambda}_n' S_n \tilde{\lambda}_n - \dfrac{n(\bar{\mu}-\mu)' S_n^{-1}(\bar{\mu}-\mu)}{2\gamma^{(2)}(0)} \\
&= \dfrac{n(\bar{\mu}-\mu)' S_n^{-1}(\bar{\mu}-\mu)}{2\gamma^{(2)}(0)} - \dfrac{\gamma^{(2)}(0)}{2} n\tilde{\lambda}_n' S_n \tilde{\lambda}_n - \textstyle\sum_{i=1}^n \tilde{\alpha}_{i,n}
\end{aligned}
$$

où $\|\tilde{\alpha}_{i,n}\| \leq \widetilde{B}|\lambda_n'(X_i-\mu)|^3$, pour $\widetilde{B} > 0$, en probabilité, ce qui donne :

$$
\|\sum_{i=1}^n \tilde{\alpha}_{i,n}\| \leq \widetilde{B}^3 \|\lambda_n\|^3 \sum_{i=1}^n \|(X_i-\mu)\|^3 = \mathcal{O}_{\mathbb{P}}(n^{-3/2}) \cdot n \cdot o_{\mathbb{P}}(n^{1/2}) = o_{\mathbb{P}}(1)
$$

De plus, comme on sait que $\lambda_n = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$, $\tilde{\lambda}_n = o_{\mathbb{P}}(n^{-1/2})$, $S_n = \mathcal{O}_{\mathbb{P}}(1)$, $\bar{\mu}-\mu = \mathcal{O}_{\mathbb{P}}(n^{-1/2})$ et $n(\bar{\mu}-\mu)' S_n^{-1}(\bar{\mu}-\mu) \xrightarrow[n\to\infty]{\text{en loi}} \chi^2(p)$, il vient

$$
2\gamma^{(2)}(0)\beta_n^\gamma(\mu) \xrightarrow[n\to\infty]{\text{en loi}} \chi^2(p).
$$

B.2 Preuve du Théorème 3.2

D'après l'égalité (Dual) donnant $\beta_n^\gamma(\mu)$ et en développant K_ε au voisinage de 0 on a

$$
\begin{aligned}
\beta_n^\gamma(\mu) &= \sup_{\lambda\in\mathbb{R}^p} \left\{ n\lambda'(\mu-\bar{\mu}) - \frac{1}{2}\sum_{i=1}^n (\lambda'(X_i-\mu))^2 K_\varepsilon^{(2)}(t_{i,n}) \right\} \\
\beta_n^\gamma(\mu) &\leq \sup_{\lambda\in\mathbb{R}^p} \left\{ n\lambda'(\mu-\bar{\mu}) - \frac{1}{2}\sum_{i=1}^n (\lambda'(X_i-\mu))^2 \varepsilon \right\},
\end{aligned}
$$

car $K_\varepsilon^{(2)} \geq \varepsilon$. Si l'on pose $l = \varepsilon\lambda$,

$$\sup_{\lambda\in\mathbb{R}^p}\left\{n\lambda'(\mu-\bar{\mu}) - \frac{1}{2}\sum_{i=1}^{n}(\lambda'(X_i-\mu))^2\varepsilon\right\} = \frac{1}{\varepsilon}\sup_{l\in\mathbb{R}^p}\left\{nl'(\mu-\bar{\mu}) - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)'ll'(X_i-\mu)\right\}.$$

Or ce sup est le même que dans le cas du χ^2. Il est atteint en $\lambda_n = S_n^{-1}(\mu-\bar{\mu})$ avec $S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu).(X_i-\mu)'$ et vaut $\frac{n}{2}(\mu-\bar{\mu})'S_n^{-1}(\mu-\bar{\mu})$ d'où

$$\beta_n^\gamma(\mu) \leq \frac{n}{2\varepsilon}(\mu-\bar{\mu})'S_n^{-1}(\mu-\bar{\mu})$$
$$\Pr(\mu \notin C_{\eta,n,\gamma^*}) \leq \Pr\left(\frac{n}{2}(\mu-\bar{\mu})'S_n^{-1}(\mu-\bar{\mu}) \geq \eta\varepsilon\right).$$

L'inégalité exponentielle est une conséquence directe de l'inégalité de Hoeffding, cf. Efron (1969).

Montrons maintenant la propriété de correction au sens de Bartlett. On note $\beta_n^\varepsilon(\mu)$ la valeur de n fois le sup dans le programme dual pour K_ε. $\beta_n^0(\mu)$ correspond alors à la vraisemblance empirique ($\gamma = K_0$) et $\beta_n^1(\mu)$ au χ^2 ($\gamma = K_1$). Soit $\mathbb{E}_n$ un estimateur de $\mathbb{E}[\beta_n^0(\mu)]/p$, on peut écrire

$$\begin{aligned}
T_n^\varepsilon = \frac{2\beta_n^\varepsilon(\mu)}{\mathbb{E}_n} &= \frac{2}{\mathbb{E}_n}\sup_{\lambda\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\lambda'(\mu-X_i) - K_\varepsilon(\lambda'(X_i-\mu))\right\} \\
&= \frac{2}{\mathbb{E}_n}\sup_{\lambda\in\mathbb{R}^p}\left\{\varepsilon\sum_{i=1}^{n}\lambda'(\mu-X_i) - K_1(\lambda'(X_i-\mu))\right. \\
&\qquad \left. +(1-\varepsilon)\sum_{i=1}^{n}\lambda'(\mu-X_i) - K_0(\lambda'(X_i-\mu))\right\} \\
&\leq \frac{2}{\mathbb{E}_n}\left\{\varepsilon\beta_n^1(\mu) + (1-\varepsilon)\beta_n^0(\mu)\right\} \\
T_n^\varepsilon &\leq T_n^0 + \varepsilon\left[T_n^1 - T_n^0\right],
\end{aligned}$$

d'où

$$\bar{F}_{T_n^\varepsilon}(\eta) \leq \bar{F}_{T_n^0+\varepsilon[T_n^1-T_n^0]}(\eta).$$

D'après les résultats de DiCiccio, Hall & Romano (1991),

$$\bar{F}_{T_n^0} = \Pr(\frac{2\beta_n^0(\mu)}{\mathbb{E}_n} \geq \cdot) = \bar{F}_{\chi^2} + \mathcal{O}(n^{-2})$$

et donc,

$$\begin{aligned}
\bar{F}_{T_n^0+\varepsilon[T_n^1-T_n^0]}(\eta) &= \Pr\{T_n^0+\varepsilon\left[T_n^1-T_n^0\right]\geq\eta, T_n^1-T_n^0\leq\varepsilon^{-1}n^{-3/2}\}\\
&\quad+\Pr\{T_n^0+\varepsilon\left[T_n^1-T_n^0\right]\geq\eta, T_n^1-T_n^0\geq\varepsilon^{-1}n^{-3/2}\}\\
&\leq \Pr\{T_n^0+n^{-3/2}\geq\eta\}+\Pr\{T_n^1-T_n^0\geq\varepsilon^{-1}n^{-3/2}\}\\
&\leq \bar{F}_{T_n^0}(\eta-n^{-3/2})+\Pr\{T_n^1-T_n^0\geq\varepsilon^{-1}n^{-3/2}\}\text{ (Bartlett pour }T_n^0)\\
&\leq \bar{F}_{\chi^2}(\eta-n^{-3/2})+\mathcal{O}(n^{-2})+\Pr\{T_n^1-T_n^0\geq\varepsilon^{-1}n^{-3/2}\}\\
&\leq \bar{F}_{\chi^2}(\eta)+\mathcal{O}(n^{-3/2})+\Pr\{T_n^1-T_n^0\geq\varepsilon^{-1}n^{-3/2}\}.
\end{aligned}$$

Si on prend ε de l'ordre de $n^{-3/2}\log(n)^{-1}$, le reste est de l'ordre de $\mathcal{O}(n^{-3/2})$ (en utilisant par exemple l'inégalité de déviation modérée (Dev) pour T_n^1 et le fait que T_n^0 est déjà corrigé au sens de Bartlett). La divergence est donc corrigeable au sens de Bartlett (au moins jusqu'à l'ordre $n^{-3/2}$).

Références

[1] Baggerly, K. A. Empirical likelihood as a goodness of fit measure. *Biometrika 85* (1998), 535–547.

[2] Bertail, P. Empirical likelihood in some semi-parametric models. *preprint CREST 12* (2002). revised for Bernoulli.

[3] Bertail, P. *Empirical likelihood in non and semi-parametric models.* M. Nikulin, 2003. to appear in *Semi-parametric models and applications.*

[4] Borwein, J. M., and Lewis, A. S. Duality relationships for entropy like minimization problem. *SIAM Journal on Computation and Optimization 29*, 2 (1991), 325–338.

[5] Broniatowski, M., and Kéziou, A. *Parametric estimation and tests through divergences.* PhD thesis, L.S.T.A., 2003.

[6] Chistyakov, G. P., and Götze, F. Moderate deviations for Student's statistic. *Theory of Probability & Its Applications 47*, 3 (2003), 415–428.

[7] Corcoran, S. A. Bartlett adjustment of empirical discrepancy statistics. *Biometrika 85*, 4 (1998), 967–972.

[8] Cressie, N., and Read, T. R. C. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B 46*, 3 (1984), 440–464.

[9] Csiszár, I. Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica 2* (1967), 299–318.

[10] Deville, J. C., and Särndal, C. E. Calibration estimators in survey sampling. *Journal of the American Statistical Association 87* (1992), 376–382.

[11] DiCiccio, T., Hall, P., and Romano, J. Empirical likelihood is bartlett-correctable. *Annals of statistics 19*, 2 (1991), 1053–1061.

[12] DiCiccio, T., and Romano, J. Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review 58* (1990), 59–76.

[13] Efron, B. Student's t-test under symmetry conditions. *Journal of american statistical society 64* (1969), 1278–1302.

[14] Embrechts, P., Lindskog, F., and McNeil, A. J. *Handbook of heavy tailed distributions in finance*. Elsevier, 2003, ch. Modelling dependence with copulas and applications to risk management. edited by Rachev ST.

[15] Golan, A., Judge, G., and Miller, D. *Maximum Entropy Econometrics*. Wiley, New York, 1996.

[16] Hartley, H. O., and Rao, J. N. K. A new estimation theory for sample surveys. *Biometrika 55* (1968), 547–557.

[17] Jing, B.-Y., and Wang, Q. An exponential nonuniform Berry-Esseen bound for self-normalized sums. *Annals of Probability 27*, 4 (1999), 2068–2088.

[18] Jing, B. Y., and Wood, A. T. A. Exponential empirical likelihood is not bartlett correctable. *Annals of Statistics 24* (1996), 365–369.

[19] Leonard, C. Minimizers of energy functionals. *Acta Mathematica Hungarica 93* (2001), 281–325.

[20] Liese, F., and Vajda, I. *Convex Statistical distance*. Teubner, Leipzig, 1987.

[21] Mykland, P. A. Bartlett type of identities. *Annals of Statistics 22* (1994), 21–38.

[22] Mykland, P. A. Dual likelihood. *Annals of Statistics 23* (1995), 396–421.

[23] Newey, W. K., and Smith, R. J. Higher order properties of GMM and generalized empirical likelihood estimators. Preliminary Version for Econometrica, 2002.

[24] Owen, A. B. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika 75*, 2 (1988), 237–249.

[25] Owen, A. B. Empirical likelihood ratio confidence regions. *Annals of Statistics 18* (1990), 90–120.

[26] Owen, A. B. *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, 2001.

[27] Qin, J., and Lawless, J. Empirical likelihood and general estimating equations. *Annals of Statistics 22*, 1 (1994), 300–325.

[28] Rao, M. M., and Ren, Z. D. *Theory of Orlicz Spaces*. Marcel Dekker, New York, 1991.

[29] Rockafellar, R. T. Integrals which are convex functionals. *Pacific Journal of Mathematics 24* (1968), 525–539.

[30] Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[31] Rockafellar, R. T. Integrals which are convex functionals (II). *Pacific Journal of Mathematics 39* (1971), 439–469.