Vraisemblance empirique généraliséeet estimation semi-paramétrique

Hugo Harari-Kermadec

UNIVERSITE PARIS X - NANTERRE
ECOLE DOCTORALE CONNAISSANCE, LANGAGE, MODÉLISATION

T H È S E

en vue de l'obtention du grade de
DOCTEUR DE L'UNIVERSITÉ PARIS X
Discipline : Mathématiques Appliquées et Applications des Mathématiques

présentée et soutenue publiquement par

Hugo HARARI-KERMADEC

le 5 décembre 2006

Titre de la thèse

Vraisemblance empirique généralisée et estimation semi-paramétrique

sous la direction de
Patrice BERTAIL

Composition du Jury

Rapporteurs

M. Yuichi Kitamura	Professeur, Yale University
M. Michel Broniatowski	Professeur, Université Paris VI

Examinateurs

Mme Dominique Picard	Professeur, Université Paris VII
M. Christian Léonard	Professeur, Université Paris X
M. Jean-Marc Robin	Professeur, Université Paris I
M. Patrice Bertail	Professeur, Université Paris X

Remerciements

Je souhaite tout d'abord remercier mon directeur de thèse, Patrice Bertail, qui a donné de son temps sans compter tout au long de ces trois années. Avec le soutien et la confiance de Patrice, la recherche a toujours été un travail agréable et captivant.

Je tiens à remercier Dominique Picard, pour sa bienveillance essentielle au départ de ce projet ; bienveillance qu'elle a bien voulu renouveler en acceptant de participer à mon jury. Je remercie également Michel Broniatowski et Yuichi Kitamura, qui ont bien voulu me faire l'honneur d'être les rapporteurs de cette thèse. Enfin, je remercie sincèrement Christian Léonard et Jean-Marc Robin d'avoir accepté de faire partie du jury.

Le soutien et la bonne humeur d'Emmanuelle Gauthèrat ont beaucoup compté dans la gaieté du travail au CREST, surtout lorsqu'il s'est agit de s'atteler aux travaux les moins gratifiants. J'en profite pour remercier Christian Robert pour la confiance qu'il m'a accordée. Merci également à tous les membres du LS, aux assistants de l'ENSAE, à Frédéric Pascal et à Pierre Neuvial pour nos discussions sur la recherche et l'enseignement en Statistique. Merci aussi aux informaticiens de l'ENSAE, qui ont permis que mes programmes, souvent maladroits, finissent par tourner.

Du côté de l'INRA, je tiens à remercier Jessica Tressou qui a ouvert la voie menant de l'ENSAE à l'INRA. Merci à Fabrice Etilé et Véronique Nichèle, pour leur aide en Econométrie. Merci aussi à Amélie Crépet et à Denis Ravaille, qui ont participé à certains des travaux de cette thèse ; à Pierre Combris et France Caillavet, ainsi qu'à l'ensemble des membres du CORELA, pour leur chaleureux accueil ; à David Delobel et à Christine Boizot pour leur soutien technique indispensable ; à Odile Bouffard et Béatrice de Vendomois pour leur aide dans la préparation des missions.

Merci à mes parents, Ben et Leo, et à mon frère, Diego, qui ont eu la patience d'écouter mes explications alambiquées. Leur confiance m'a permis de me lancer dans cette aventure et de tenir éloigné le découragement. Merci aussi à Mario Wschebor, pour sa bienveillance à mon égard ; à mes amis, à mes camarades et à mes étudiants, pour la vie partagée en dehors de la recherche.

Enfin et surtout, un immense merci à toi Mays, pour ton amour, ton soutien et ta confiance, renouvelés tout au long de ces 3 années mouvementées.

شكرا جزيلا يا حبيبي

Table des matières

Table des figures

Chapitre 1

Introduction

La VRAISEMBLANCE EMPIRIQUE est une méthode d'estimation inspirée de la méthode du maximum de vraisemblance usuelle, mais s'affranchissant du choix d'une famille paramétrique de lois. La méthode de vraisemblance empirique a été principalement introduite par Owen (1988, 1990, 2001), bien qu'on puisse la considérer comme une extension des méthodes de calage utilisées depuis de nombreuses années en sondage (voir Deville & Sarndal, 1992, 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. Owen a montré que l'on pouvait obtenir une version non-paramétrique du théorème de Wilks. Ce théorème établit la convergence du rapport de vraisemblance vers une loi du χ^2, permettant ainsi de réaliser des tests ou de construire des régions de confiance. Cette méthode a été généralisée à de nombreux modèles, lorsque le paramètre d'intérêt est défini à partir de contraintes de moments (Qin & Lawless, 1994, Newey & Smith, 2004). Les propriétés de la vraisemblance empirique en font une alternative à la méthode des moments généralisés.

Le chapitre introductif rappellera les principaux concepts de la vraisemblance empirique ainsi que ces propriétés les plus importantes puis nous présenterons les résultats obtenus au cours de cette thèse. Le corps de cette thèse se présente en deux parties. Dans la première partie, nous étudierons les propriétés théoriques de la vraisemblance empirique et de ses généralisations. Nous montrerons en effet au chapitre 2 que l'on peut replacer la méthode de vraisemblance empirique dans un cadre plus général, celui des divergences empiriques. Ce chapitre fait l'objet d'un article à paraître aux *Annales d'Économie et de Statistique*, en collaboration avec Patrice Bertail et Denis Ravaille. Le chapitre 3 présentera des résultats non asymptotiques originaux pour les divergences empiriques. L'article correspondant, rédigé avec Patrice Bertail et Emmanuelle Gauthèrat, est soumis. Nous proposerons au chapitre 4 une méthode pour traiter des données non indépendantes, lorsqu'elles sont décrites par une chaîne de Markov. Ce travail est en révision pour le numéro spécial de *Econometric Theory* consacré à la vraisemblance empirique. Dans la deuxième partie, nous appliquerons ces méthodes à des problèmes issus de l'analyse de la consommation alimentaire. Nous décrirons au chapitre 5 l'estimation d'un indice de risque d'exposition au méthylmercure par consommation des produits de la mer. Cette étude, menée en collaboration avec Amélie Crépet et Jessica Tressou, est soumise. Une deuxième application, exposée au chapitre 6, étudie l'effet des normes sociales sur le sur-poids et l'obésité en France. Ce travail, entreprit

avec Fabrice Étilé, est en révision en vue d'une publication dans *Health Economics*.

1.1 La vraisemblance empirique

Dans cette section, nous présentons une revue de la littérature traitant de la vraisemblance empirique. Une grande partie des résultats énoncés se trouvent dans le livre d'Owen (2001) *Empirical Likelihood*.

1.1.1 Le cas de l'espérance

Définition de la vraisemblance empirique

Soit une suite de vecteurs aléatoires $X, X_1, \cdots, X_n$ définis sur $(\Omega, \mathcal{A})$ et à valeur dans $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, indépendants et identiquement distribués de loi de probabilité $\mathbb{P}_0$. On pose $\Pr = \mathbb{P}_0^{\otimes n}$. On cherche à construire des régions de confiance pour l'espérance μ_0 de X. Nous commencerons par proposer la présentation classique de la vraisemblance empirique, avant d'en donner une interprétation plus probabiliste.

L'idée d'Owen (1988), dont nous reprenons ici la présentation originale, est de construire une vraisemblance en se donnant comme modèle l'ensemble des multinomiales ne chargeant que l'échantillon. On affecte ainsi un poids $q_i > 0$ à chacune des observations, la somme des poids étant fixée à 1. On définit ainsi une mesure de probabilité $\mathbb{Q} = \sum_{i=1}^n q_i \delta_{X_i}$. La vraisemblance $L(\mathbb{Q})$ du modèle multinomial s'écrit alors

$$L(\mathbb{Q}) = \prod_{i=1}^n q_i.$$

Remarquons que sous la contrainte $\sum_{i=1}^n q_i = 1$, L est maximale en $\mathbb{P}_n = \sum_{i=1}^n \frac{1}{n}\delta_{X_i}$, la probabilité empirique. On peut alors définir le rapport de vraisemblance

$$\mathcal{R}(\mathbb{Q}) = \frac{L(\mathbb{Q})}{L(\mathbb{P}_n)} = \prod_{i=1}^n nq_i.$$

Puisque plusieurs multinomiales peuvent avoir la même espérance, pour définir le rapport de vraisemblance en une valeur μ du paramètre, Owen (1990) propose de déterminer la multinomiale maximisant le rapport de vraisemblance sous la contrainte $\sum_{i=1}^n q_i X_i = \mu$. Ce problème peut alors s'écrire

$$R(\mu) = \sup_{q_1,\cdots,q_n} \left\{ \prod_{i=1}^n nq_i \middle| \sum_{i=1}^n q_i = 1,\ \sum_{i=1}^n q_i X_i = \mu,\ \forall i \in [|1,n|], q_i > 0 \right\}.$$

On peut utiliser la méthode du multiplicateur de Lagrange pour résoudre notre problème. Pour ce faire, on considère le log du produit des nq_i, ce qui linéarise l'optimisation et permet d'omettre la contrainte de positivité des q_i. Posons

$$\mathcal{L}_\mu = \sum_{i=1}^n \log(nq_i) - n\lambda \sum_{i=1}^n q_i(X_i - \mu) - n\gamma \left(\sum_{i=1}^n q_i - 1 \right).$$

Annulons la dérivée de $\mathcal{L}_\mu$ par rapport à q_i :

$$q_i^{-1} - n\lambda(X_i - \mu) - n\gamma = 0. \tag{1.1}$$

En multipliant (1.1) par q_i et en sommant sur i, on obtient : $\sum_{i=1}^{n}(1-n\lambda q_i(X_i-\mu)-n\gamma q_i) = 0$. En utilisant les contraintes, on a $n = n\gamma$ et donc $\gamma = 1$. On obtient alors l'expression de q_i grâce à (1.1) :

$$q_i = \frac{1}{n(1 + \lambda(X_i - \mu))},$$

où λ est donné par la contrainte $\mathbb{E}_{\mathbb{Q}}[X - \mu_0] = 0$, c'est-à-dire :

$$\sum_{i=1}^{n} \frac{X_i - \mu}{n(1 + \lambda(X_i - \mu))} = 0.$$

Le rapport de vraisemblance empirique permet de construire des régions de confiance asymptotiques pour μ_0. Pour ce faire, on définit une région de confiance $C_{\eta,n}$ par inversion du test de rapport de vraisemblance :

$$C_{\eta,n} = \left\{ \mu \,\middle|\, -2\log\left(R(\mu)\right) \leq \eta \right\}.$$

où η est un paramètre permettant de régler le niveau de confiance que l'on veut atteindre. On note $\beta_n(\mu)$ la statistique pivotale :

$$\beta_n(\mu) = -2\log\left(R(\mu)\right).$$

Théorème de convergence

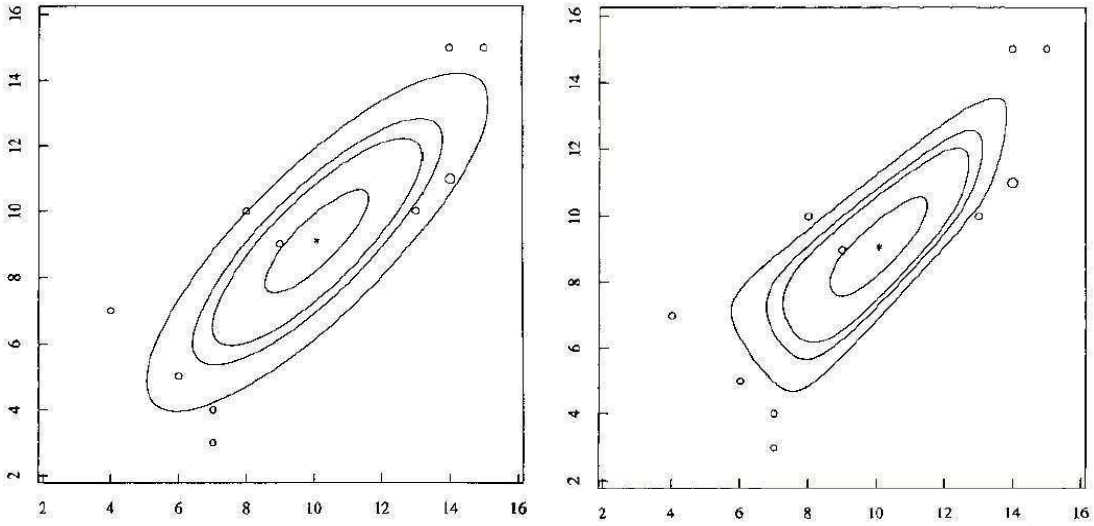
Owen (1988, 1990) a établi le théorème suivant :

Théoreme 1.1 (Owen) *Si $X_1, \cdots, X_n$ sont des vecteurs aléatoires de $\mathbb{R}^p$, indépendants et identiquement distribués, d'espérance μ_0 et de variance Σ de rang q, alors, $\forall\ 0 < \eta < 1$, $C_{\eta,n}$ est convexe et*

$$\Pr(\mu_0 \in C_{\eta,n}) = \Pr(\beta_n(\mu_0) \leq \eta) \xrightarrow[n\to\infty]{} F_{\chi^2_q}(\eta),$$

où $F_{\chi^2_q}(\cdot)$ est la fonction de répartition d'une distribution du χ^2_q.

Owen (1990) propose d'illustrer ce résultat en construisant des régions de confiance pour l'espérance par la méthode de vraisemblance empirique. On dispose d'un échantillon de onze données $X_1, \cdots, X_{11} \in \mathbb{R}^2$, concernant une population de canards dont on observe le comportement et le plumage. On construit des régions de confiance par la méthode de vraisemblance classique, en supposant les données gaussiennes, et par la méthode de vraisemblance empirique. On obtient alors les Figures 1.1-(a) et 1.1-(b).



(a) Vraisemblance paramétrique (b) Vraisemblance empirique
Régions à 50%, 90%, 95% et 99%.

FIG. 1.1 – Régions de confiance pour l'espérance de l'échantillon, extrait de Owen (1990)

L'hypothèse gaussienne resurgit sur la forme (elliptique) des régions de confiance paramétriques, sans qu'il soit possible de justifier cette hypothèse. Cet exemple illustre une propriété intéressante de la vraisemblance empirique : la forme des régions de confiance s'adapte à la géométrie des données. On observe également que les régions sont plus petites que pour la vraisemblance paramétrique. Cette différence de taille est générale et l'on peut facilement démontrer que la région de confiance à 100% pour la vraisemblance empirique est l'enveloppe convexe des points, alors qu'avec la méthode paramétrique on obtiendrait $\mathbb{R}^2$. Cependant, on verra dans le chapitre 3 qu'en fait les régions de confiance d'Owen sont trop petites au regard du taux de couverture et que le niveau de confiance est surestimé pour les petites valeurs de n.

Correction de Bartlett

DiCiccio et al. (1991) montre que la vraisemblance empirique est corrigeable au sens de Bartlett. Ceci signifie que l'on peut corriger la statistique $\beta_n(\mu)$ d'un facteur qui permet de faire disparaître le biais au premier ordre. Ceci assure un meilleur comportement à distance finie, puisque la différence avec le comportement asymptotique est réduit de $\mathcal{O}(n^{-1})$ à $\mathcal{O}(n^{-2})$. On peut énoncer ce résultat sous la forme suivante :

Théoreme 1.2 (Correction de Bartlett) *Soient $X_1, X_2, \ldots$ des vecteurs aléatoires de $\mathbb{R}^p$, indépendants et identiquement distribués, d'espérance μ_0 et de variance Σ de rang q. Si, de plus, on a $\mathbb{E}[\|X_1\|^8] < \infty$ et*

$$\limsup_{||t||\to\infty} |\mathbb{E}[\exp(it'X_1)]| < \infty,$$

alors, en posant $E_n = \mathbb{E}[\beta_n(\mu_0)]/q$, on a

$$\Pr\left(\frac{\beta_n(\mu_0)}{E_n} \leq \eta\right) = F_{\chi^2_q}(\eta) + \mathcal{O}(n^{-2}).$$

L'idée consiste à corriger $\beta_n(\mu_0)$ afin que son espérance soit égale à celle de la distribution du χ^2_q, c'est-à-dire q, ce qui annule le terme en n^{-1}. Le développement d'Edgeworth de $\Pr(\beta_n(\mu_0) \leq \eta)$ pour $p = q = 1$ permet de comprendre le fonctionnement de la correction de Bartlett. En effet, on a

$$\Pr(\beta_n(\mu_0) \leq \eta) = F_{\chi^2_1}(\eta) + \frac{a\sqrt{\eta}}{n}\frac{e^{-\eta/2}}{\sqrt{2\pi}} + \mathcal{O}(n^{-2}).$$

Or, le terme correctif E_n peut se développer en $E_n = 1 - a/n + \mathcal{O}(n^{-2})$ et

$$\begin{aligned} F_{\chi^2_1}\left(\eta(1-a/n)\right) &= F_{\chi^2_1}(\eta) - 2\int_{\sqrt{\eta(1-\frac{a}{n})}}^{\sqrt{\eta}} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt + \mathcal{O}(n^{-2}) \\ &= F_{\chi^2_1}(\eta) - 2\sqrt{\eta}\left(1 - \sqrt{1-\frac{a}{n}}\right)\frac{e^{-\eta/2}}{\sqrt{2\pi}} + \mathcal{O}(n^{-2}) \\ &= F_{\chi^2_1}(\eta) - \frac{a\sqrt{\eta}}{n}\frac{e^{-\eta/2}}{\sqrt{2\pi}} + \mathcal{O}(n^{-2}). \end{aligned}$$

Finalement,

$$\Pr\left(\frac{\beta_n(\mu_0)}{E_n} \leq \eta\right) = F_{\chi^2_1}\left(\eta(1-a/n)\right) + \frac{a\sqrt{\eta}}{n}\frac{e^{-\eta/2}}{\sqrt{2\pi}} + \mathcal{O}(n^{-2}) = F_{\chi^2_1}(\eta) + \mathcal{O}(n^{-2}).$$

Dans le cas $p = q = 1$, en posant $m_j = \mathbb{E}_{\mathbb{P}_0}||X - \mu_0||^j$ et $\gamma_j = \frac{m_j}{m_2^{j/2}}$, on peut montrer que

$$a = \frac{m_4}{2m_2^2} - \frac{m_3^2}{3m_2^3} = \frac{1}{2}\gamma_4 - \frac{1}{3}\gamma_3^2,$$

que l'on estime naturellement par sa contrepartie empirique : $\widehat{a} = \frac{\widehat{m}_4}{2\widehat{m}_2^2} - \frac{\widehat{m}_3^2}{3\widehat{m}_2^3}$. DiCiccio et al. (1991) montrent que l'ordre de l'approximation n'est pas perturbé lorsque l'on remplace E_n par son estimateur :

$$\Pr\left(\frac{\beta_n(\mu_0)}{1-\widehat{a}/n} \leq \eta\right) = F_{\chi^2_q}(\eta) + \mathcal{O}(n^{-2}).$$

L'efficacité concrète de la correction de Bartlett peut être illustrée par simulation : on estime le taux de couverture à distance finie de la région de confiance que l'on compare au niveau de confiance visé $(1-\alpha)$. La Figure 1.2 est obtenue en simulant des échantillons d'un mélange d'échelle (le produit d'une uniforme sur $[0;1]$ et d'une gaussienne centrée réduite sur $\mathbb{R}^6$). Pour chaque échantillon, on construit une région de confiance à 90% pour l'espérance suivant la méthode de la vraisemblance empirique, avec et sans correction de Bartlett. Grâce à la méthode de Monte-Carlo, on constate alors que le taux de couverture est amélioré par la correction de Bartlett : on se rapproche de $1-\alpha = 0.9$. Malheureusement, cette amélioration est loin d'être suffisante pour les petites valeurs de n.

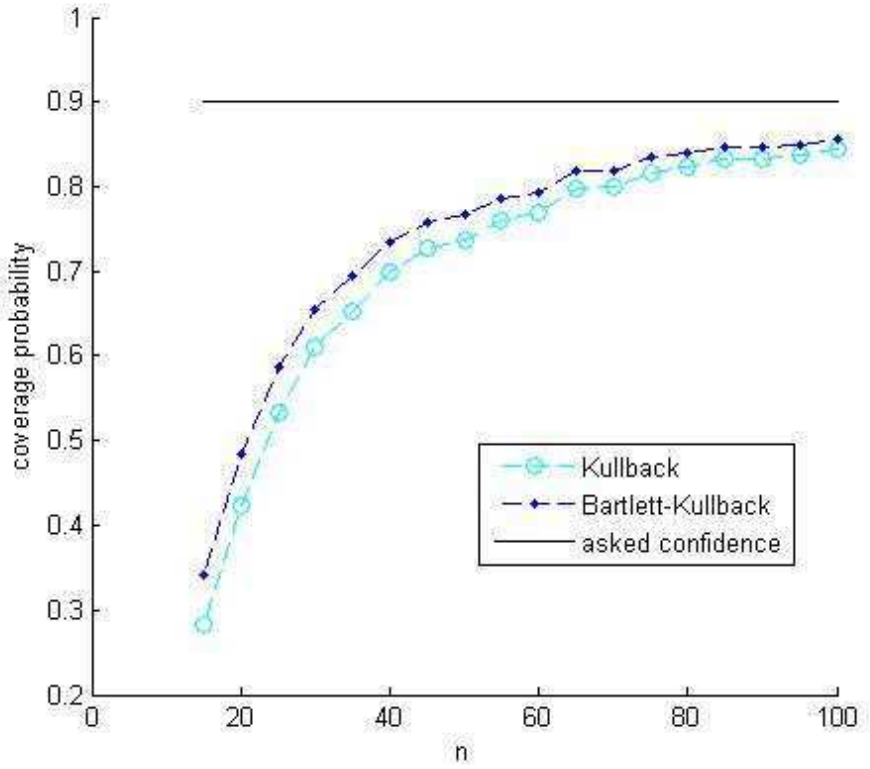


FIG. 1.2 – Taux de couverture à distance finie

Borne inférieure de l'erreur de première espèce

Ces considérations sur le comportement à distance finie soulignent les limites des résultats asymptotiques établis par le Théorème 1.1. On est alors amené à chercher des bornes à distance finie pour le niveau de confiance $\Pr(\beta_n(\mu_0) \leq \eta)$. Malheureusement, on ne peut construire une borne inférieure pour le taux de couverture des régions de confiance de la méthode de vraisemblance empirique (voir le chapitre 3).

En revanche, on peut obtenir une borne supérieure pour ce taux de couverture. En effet, comme le remarque Tsao (2004), les poids affectés aux données par la vraisemblance empirique étant toujours positifs et de somme égale à 1, les régions de confiance sont nécessairement incluses dans l'enveloppe convexe de l'échantillon, notée $\mathcal{H}(X_1, ..., X_n) \subset \mathbb{R}^p$:

$$C_{\eta,n} \subset \mathcal{H}(X_1, ..., X_n).$$

On peut donc majorer le taux de couverture par la probabilité que l'espérance appartienne à cette enveloppe convexe :

$$\forall \eta > 0,\ \Pr\left(\beta_n(\mu_0) \leq \eta\right) = \Pr(\mu_0 \in C_{\eta,n}) \leq \Pr\left(\mu_0 \in \mathcal{H}(X_1, ..., X_n)\right)$$

Tsao (2004) donne un majorant explicite pour cette borne pour $p \leq 2$:

$$\Pr\left(\mu_0 \in \mathcal{H}(X_1, ..., X_n)\right) \leq 1 - \frac{n}{2^{n-1}}$$

et conjecture que ce majorant se généralise pour p quelconque en

$$\Pr\left(\mu_0 \in \mathcal{H}(X_1, ..., X_n)\right) \leq 1 - \frac{1}{2^{n-1}} \sum_{0 \leq k < p} C_{n-1}^k.$$

Ces résultats corroborent notre commentaire de la Figure 1.2 : la vraisemblance empirique sous-évalue l'erreur de première espèce. En particulier, lorsque q/n n'est pas très petit, le taux de couverture des régions de confiance pour α petit est nécessairement inférieur au niveau nominal. Ceci motive la recherche d'alternatives à la vraisemblance empirique. Nous proposerons au chapitre 2 une généralisation de la vraisemblance empirique, les divergences empiriques, et nous montrerons que les régions de confiance construites avec certaines divergences empiriques peuvent s'étendre au-delà de l'enveloppe $\mathcal{H}$. On obtient ainsi des taux de couverture qui dépassent la borne supérieure de Tsao (2004). Pour certaines divergences empiriques, nous établirons au chapitre 3 des bornes *inférieures* pour le taux de couverture.

Vraisemblance empirique et divergence de Kullback

Dans l'optique de généraliser la méthode de vraisemblance empirique, il est utile d'introduire une seconde écriture de β_n faisant intervenir la divergence de Kullback. Soit $\mathbb{M}$ l'ensemble des mesures de probabilités sur $(\Omega, \mathcal{A})$. Pour $\mathbb{P}, \mathbb{P}'$ appartenant à $\mathbb{M}$, on note :

$$K(\mathbb{P}, \mathbb{P}') = \begin{cases} -\int \log\left(\frac{d\mathbb{P}}{d\mathbb{P}'}\right) d\mathbb{P}' & \text{si } \mathbb{P} \ll \mathbb{P}', \\ +\infty & \text{sinon.} \end{cases}$$

La statistique pivotale β_n intervenant dans le Théorème 1.1 peut s'écrire :

$$\begin{aligned}
\beta_n(\mu) &= -2 \log\left(\sup_{q_1, \cdots, q_n} \left\{ \prod_{i=1}^n n q_i \,\middle|\, \sum_{i=1}^n q_i X_i = \mu,\ \sum_{i=1}^n q_i = 1,\ \forall i \in [|1, n|], q_i > 0 \right\} \right) \\
&= -2n \sup_{q_1, \cdots, q_n} \left\{ \frac{1}{n} \sum_{i=1}^n \log(n q_i) \,\middle|\, \sum_{i=1}^n q_i X_i = \mu,\ \sum_{i=1}^n q_i = 1 \right\} \\
&= -2n \sup_{\mathbb{Q} \in \mathbb{M}} \left\{ \int \log\left(\frac{d\mathbb{Q}}{d\mathbb{P}_n}\right) d\mathbb{P}_n \,\middle|\, \mathbb{Q} \ll \mathbb{P},\ \mathbb{E}_{\mathbb{Q}}[X - \mu] = 0 \right\} \\
&= 2n \inf_{\mathbb{Q} \in \mathbb{M}} \left\{ K(\mathbb{Q}, \mathbb{P}_n) \,\middle|\, \mathbb{E}_{\mathbb{Q}}[X - \mu] = 0 \right\}.
\end{aligned}$$

Cette écriture de la vraisemblance empirique peut se comparer à la définition du maximum de vraisemblance paramétrique comme un minimum de contraste basé sur la divergence de Kullback. En effet, dans le cadre paramétrique, lorsque l'on se donne un ensemble de mesures $\mathbb{P}_\theta$ équivalentes à une mesure dominante ν, on peut définir

$$\gamma(\theta_1, \theta_2) = K(\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2}).$$

Définissons $U_n(\theta) = -\frac{1}{n} \sum_{i=1}^n \log \frac{d\mathbb{P}_\theta}{d\nu}$. En posant $\mathbb{P}_0 = \mathbb{P}_{\theta_0}$, on a d'après la loi des grands nombres :

$$U_n(\theta_1) - U_n(\theta_0) = \frac{1}{n} \sum_{i=1}^n \log \frac{d\mathbb{P}_0}{d\nu} - \frac{1}{n} \sum_{i=1}^n \log \frac{d\mathbb{P}_{\theta_1}}{d\nu} \xrightarrow[n \to \infty]{\text{p.s.}} \mathbb{E}_{\mathbb{P}_0} \log \frac{d\mathbb{P}_0}{d\nu} - \mathbb{E}_{\mathbb{P}_0} \log \frac{d\mathbb{P}_{\theta_1}}{d\nu}$$

On remarque alors que

$$\mathbb{E}_{\mathbb{P}_0} \log \frac{d\mathbb{P}_0}{d\nu} - \mathbb{E}_{\mathbb{P}_0} \log \frac{d\mathbb{P}_{\theta_1}}{d\nu} = \mathbb{E}_{\mathbb{P}_0} \log \frac{d\mathbb{P}_0}{d\mathbb{P}_{\theta_1}} = K(\mathbb{P}_{\theta_1}, \mathbb{P}_0) = \gamma(\theta_1, \theta_0).$$

$U_n(\theta_1) - U_n(\theta_0)$ converge donc vers $\gamma(\theta_1, \theta_0)$ presque sûrement. L'estimateur du minimum de contraste est alors donné par

$$\arg\inf_{\theta_1}\{U_n(\theta_1) - U_n(\theta_0)\} = \arg\inf_{\theta_1} U_n(\theta_1).$$

On peut alors interpréter la vraisemblance empirique comme une méthode de contraste pour le modèle des multinomiales dominées par $\mathbb{P}_n$ (qui dépend cette fois-ci de n), en définissant :

$$U_n(\mu) = -\inf_{\mathbb{Q} \ll \mathbb{P}_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log \frac{d\mathbb{Q}}{d\mathbb{P}_n} \middle| \mathbb{E}_{\mathbb{Q}}[X - \mu] = 0 \right\} = \inf_{\mathbb{E}_{\mathbb{Q}}[X-\mu]=0} K(\mathbb{Q}, \mathbb{P}_n)$$

qui n'est autre que $\beta_n(\mu)/2n$.

1.1.2 Équation de moments

Définitions

Afin de généraliser la méthode de vraisemblance empirique à l'estimation d'un paramètre θ_0 de $\mathbb{R}^p$, on utilise communément une équation de moments (ou équation d'estimation). On suppose alors que θ_0 peut être défini par :

$$\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0)] = 0$$

où m est une fonction régulière de $\mathcal{X} \times \mathbb{R}^p \to \mathbb{R}^r$. On estime alors naturellement θ_0 grâce au pendant empirique de l'équation de moments : $\mathbb{E}_{\mathbb{P}_n}[m(X, \bar{\theta})] = 0$. Cette équation n'a pas toujours de solution $\bar{\theta}$. Ceci est en particulier possible en cas de sur-identification, c'est-à-dire lorsque $p < r$. Ce cas, développé par la suite, permet de prendre en compte des contraintes portant sur l'échantillon, et d'ajouter ainsi de l'information aux données, issue d'autres échantillons ou de considérations théoriques. S'il existe une solution, on estime θ_0 par la valeur θ du paramètre la plus vraisemblable, c'est-à-dire celle qui vérifie l'équation de moments pour la mesure $\mathbb{Q}$ la plus proche possible de $\mathbb{P}_n$, au sens de la distance de Kullback. On mesure ainsi la vraisemblance d'un θ quelconque en adaptant β_n :

$$\beta_n(\theta) = 2n \inf_{\mathbb{Q} \in \mathbb{M}} \left\{ K(\mathbb{Q}, \mathbb{P}_n) \middle| \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0 \right\},$$

et

$$\beta_n(\widehat{\theta}) = \inf_{\theta \in \mathbb{R}^p} \{\beta_n(\theta)\} = 2n \inf_{(\theta, \mathbb{Q}) \in \mathbb{R}^p \times \mathbb{M}} \left\{ K(\mathbb{Q}, \mathbb{P}_n) \middle| \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0 \right\}.$$

On a alors pour rapport de vraisemblance

$$\beta_n(\theta) - \beta_n(\widehat{\theta}) = -2n \log \left(\frac{\sup_{\mathbb{Q} \in \mathbb{M}} \left\{ \prod_{i=1}^{n} q_i \middle| \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0 \right\}}{\sup_{(\theta, \mathbb{Q}) \in \mathbb{R}^p \times \mathbb{M}} \left\{ \prod_{i=1}^{n} q_i \middle| \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0 \right\}} \right)$$

et l'on peut construire la région de confiance

$$C_{\eta,n} = \left\{ \theta \middle| \beta_n(\theta) - \beta_n(\widehat{\theta}) \leq \eta \right\}.$$

Théorème de convergence pour la vraisemblance empirique

Comme dans le cas de l'espérance, on obtient un théorème de convergence, dû a Qin & Lawless (1994) :

Théoreme 1.3 (Qin & Lawless) *Soient X, X_1, ..., X_n une suite de vecteurs aléatoires, indépendants et identiquement distribués à valeurs dans $\mathcal{X}$. Soit θ_0 appartenant à $\mathbb{R}^p$ tel que $\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0)] = 0$, que $\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0) m(X, \theta_0)']$ soit définie positive et que le rang de $\mathbb{E}[\partial m(X, \theta_0)/\partial \theta]$ soit p. Supposons qu'il existe un voisinage V de θ_0 sur lequel $m(x, .)$ est C^2. Supposons de plus que $||m(., \theta)||^3$, $||\partial m(., \theta)/\partial \theta||$ et $||\partial^2 m(., \theta)/\partial \theta \partial \theta'||$ sont majorés sur V par une fonction intégrable. Alors, pour tout $\eta > 0$,*

$$\Pr(\theta_0 \in C_{\eta,n}) = \Pr(\beta_n(\theta_0) - \beta_n(\widehat{\theta}) \leq \eta) \xrightarrow[n \to \infty]{} F_{\chi^2_p}(\eta).$$

Corollaire 1.4 *Soit $\theta_0' = (\theta_{1,0}, \theta_{2,0})' \in \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$, et $\widehat{\theta}_2 = \arg\inf_{\theta_2 \in \mathbb{R}^{q_2}} \beta_n(\theta_{1,0}, \theta_2)$. Sous les hypothèses du Théorème 1.3,*

$$\beta_n(\theta_{1,0}, \widehat{\theta}_2) - \beta_n(\widehat{\theta}) \xrightarrow[n \to \infty]{\mathcal{L}} \chi^2_{q_1}.$$

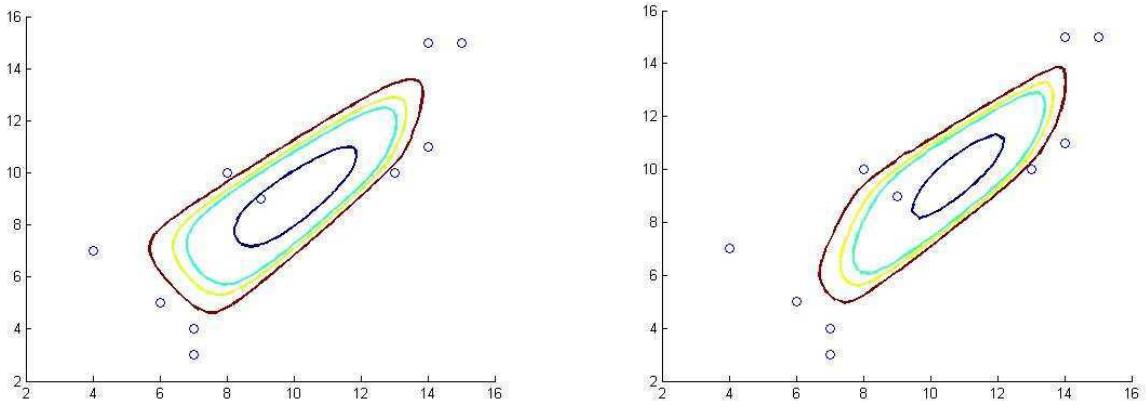
On peut donc construire une région de confiance de la même façon que précédemment pour une partie seulement du paramètre. Cette version de la vraisemblance empirique permet de rechercher un paramètre d'intérêt θ_1 en présence d'un paramètre de nuisance θ_2.

Pour illustrer ce théorème, on reprend l'exemple utilisé par Owen (1990). En plus des onze données bivariées $X_1, \cdots, X_{11}$, on dispose de l'âge A_i des canards observés et de l'espérance A_0 de l'âge au sein de l'espèce (estimé par des experts ou issu d'une enquête plus large). On peut alors redresser l'échantillon pour le contraindre à être représentatif en se qui concerne l'âge. L'équation de moments s'écrit

$$\mathbb{E}_{\mathbb{P}_0}[(X - \mu_0, A - A_0)'] = (0, 0)'.$$

On dispose donc de 3 contraintes pour estimer un paramètre bivarié μ_0. Si l'espérance A_0 est supérieure à la moyenne des âges A_i des canards observés, les pondérations données par la vraisemblance empirique, c'est-à-dire les q_i, chargerons d'avantages les canards âgés. Les

zones de confiance sont modifiées en conséquence, voir la Figure 1.3(b), et leur surface est réduite puisque l'on dispose de plus d'information.



(a) Non Contraint (b) Contraint

Régions à 50%, 90%, 95% et 99%.

FIG. 1.3 – Régions de confiance pour la moyenne de l'échantillon

Normalité asymptotique

Contrairement au cas de l'espérance pour lequel $\widehat{\theta} = \overline{X}$, la normalité asymptotique de l'estimateur $\widehat{\theta}$ n'est plus aussi triviale. Qin & Lawless (1994) ont établi le résultat suivant :

Théoreme 1.5 *Sous les hypothèses du Théorème 1.3, en posant*

$$D = \mathbb{E}\left[\frac{\partial m(.,\theta_0)}{\partial \theta}\right] \text{ et } M = \mathbb{E}[m(.,\theta_0)m(.,\theta_0)']^{-1},$$

on a la normalité asymptotique de $\widehat{\theta}$ *:*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, V) \text{ avec } V = (D'MD)^{-1}.$$

Soit $\widehat{\lambda}$ *le multiplicateur de Lagrange associé à la contrainte* $m(X, \widehat{\theta}) = 0$*, il est également asymptotiquement gaussien :*

$$\sqrt{n}\widehat{\lambda} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, U) \text{ avec } U = M\left(I_p - DVD'M\right).$$

De plus, $\widehat{\theta}$ *et* $\widehat{\lambda}$ *ne sont pas corrélés asymptotiquement.*

1.1.3 Équation de moments conditionnels

Soit (X_i, Z_i) des variables aléatoires définies sur $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}^{p'}$ de distribution jointe $\mathbb{P}_0 \in \mathbb{M}$. Dans la section précédente, on a toujours supposé que l'équation d'estimation était de la forme

$$\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0)] = 0.$$

Dans de nombreuses applications, en particulier en Économétrie, on dispose d'équation de moments conditionnels, de la forme :

$$\mathbb{E}_{\mathbb{P}_0}[m(X,\theta_0)|Z]=0,$$

pour Z une variable aléatoire liée à X. Cette contrainte est bien plus forte que la précédente. On peut bien sûr se ramener à la forme précédente en intégrant par rapport à Z, mais on perd alors beaucoup d'information. À l'inverse, on peut considérer que l'on a une équation d'estimation par valeur de Z, ce qui donne, en général, une infinité d'équations. La notion d'instrument sert à traiter ce problème classique, puisque l'on peut remarquer que pour toute fonction f,

$$\mathbb{E}_{\mathbb{P}_0}[m(X,\theta_0)\cdot f(Z)]=0,$$

et l'on est ramené à une équation d'estimation simple. La difficulté réside alors dans le choix de l'instrument $f(Z)$. Kitamura et al. (2004) introduisent la méthode « Smooth Empirical Likelihood (SEL) », qui est une modification (un lissage en un certain sens) de la vraisemblance empirique qui permet d'utiliser de façon efficiente des équations d'estimation présentées sous la forme de contraintes de moments conditionnels.

La méthode SEL consiste à affecter des pondérations w_{ij} pour lisser la contrainte conditionnelle en estimant par une méthode à noyau la densité de Z :

$$w_{ij}=\frac{\mathcal{K}\left((Z_i-Z_j)/h\right)}{\sum_j \mathcal{K}\left((Z_i-Z_j)/h\right)},$$

où $\mathcal{K}$ est le noyau et h la fenêtre. On résout alors à i fixé le programme d'optimisation suivant :

$$L_i(\theta)=\sup_{(q_{ij})_j}\left\{\sum_{j=1}^n w_{ij}\log(q_{ij})\middle|\sum_{j=1}^n q_{ij}=1,\ \sum_{j=1}^n q_{ij}m(X_j,\theta)=0\right\}$$

qui utilise la contrainte $\mathbb{E}_{\mathbb{P}_0}[m(X,\theta_0)|Z_i]=0$ en pondérant les observations X_j par les w_{ij}. On prend ainsi plus en compte les X_j pour lesquels Z_j est proche de la valeur de Z_i. Il reste ensuite à optimiser en θ la « vraisemblance » globale $L(\theta)=\sum_{i=1}^n L_i(\theta)$.

Si l'on pose $\mathbb{Q}_i=\sum_{j=1}^n q_{ij}\delta_{X_j}$, on peut réécrire la contrainte $\sum_{j=1}^n q_{ij}m(X_j,\theta)=0$ plus élégamment : $\mathbb{E}_{\mathbb{Q}_i}m(X,\theta)=0$. Posons également $\mathbb{W}_i=\sum_{j=1}^n w_{ij}\delta_{X_j}$. On remarque alors que pour chaque i, le programme de la méthode SEL peut s'écrire sous la forme d'une optimisation similaire à celle de la vraisemblance empirique classique, où l'on aurait remplacé la probabilité empirique usuelle $\mathbb{P}_n=\frac{1}{n}\sum_{j=1}^n\delta_{X_j}$, par une version locale lissée $\mathbb{W}_i$. Soit

$$\begin{aligned}\beta_{i,n}(\theta)&=2n\inf_{\mathbb{Q}_i\ll\mathbb{W}_i}\left\{K(\mathbb{Q}_i,\mathbb{W}_i)\middle|\mathbb{E}_{\mathbb{Q}_i}[m(X,\theta)]=0\right\}\\&=2n\inf_{\mathbb{Q}_i\ll\mathbb{W}_i}\left\{-\int\log\left(\frac{d\mathbb{Q}_i}{d\mathbb{W}_i}\right)d\mathbb{W}_i\middle|\mathbb{E}_{\mathbb{Q}_i-\mathbb{W}_i}[m(X,\theta)]=-\mathbb{E}_{\mathbb{W}_i}[m(X,\theta)]\right\}.\end{aligned}$$

En utilisant la méthode de Lagrange, on obtient le problème dual :

$$\begin{aligned}\beta_{i,n}(\theta) &= 2 \sup_{\lambda_i \in \mathbb{R}^r} \left\{ -\sum_{j=1}^{n} w_{ij} \log(1 + \lambda_i' m(X_i, \theta)) \right\} \\ &= 2\sum_{j=1}^{n} w_{ij} \log(w_{ij}) - 2\sum_{j=1}^{n} w_{ij} \log\left(\frac{w_{ij}}{1 + \lambda_i' m(X_i, \theta)} \right).\end{aligned}$$

On reconnaît alors dans le second terme le programme d'optimisation de la méthode SEL :

$$\begin{aligned}\beta_{i,n}(\theta) &= 2\sum_{j=1}^{n} w_{ij} \log(w_{ij}) - 2 \sup_{(q_{ij})_j} \left\{ \sum_{j=1}^{n} w_{ij} \log(q_{ij}) \middle| \sum_{j=1}^{n} q_{ij} = 1,\ \sum_{j=1}^{n} q_{ij} m(X_j, \theta) = 0 \right\} \\ &= 2\sum_{j=1}^{n} w_{ij} \log(w_{ij}) - 2L_i(\theta).\end{aligned}$$

Comme $\sum_{j=1}^{n} w_{ij} \log(w_{ij})$ est le supremum en θ de $L_i(\theta)$, $\beta_{i,n}(\theta)$ s'interprète comme le log d'un rapport de vraisemblance. On pose

$$\beta_n(\theta) = \sum_i \beta_{i,n}(\theta) = 2\sum_{i,j} w_{ij} \log(w_{ij}) - 2\sum_i L_i(\theta).$$

On obtient alors, sous des hypothèses techniques sur le noyau $\mathcal{K}$ et la fenêtre h, le résultat attendu :

$$\beta_n(\theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_q, \text{ où } q \text{ est le rang de } \frac{\partial m}{\partial \theta}(\cdot, \theta_0).$$

1.1.4 Échantillons multiples

Au chapitre 5, pour estimer le paramètre d'intérêt, on dispose de deux échantillons indépendants l'un de l'autre, $X^{(1)}, X_1^{(1)}, \cdots, X_{n_1}^{(1)}$ de loi $\mathbb{P}_1$ et $X^{(2)}, X_1^{(2)}, \cdots, X_{n_2}^{(2)}$ de loi $\mathbb{P}_2$. Chaque échantillon est i.i.d., et l'on suppose que les deux échantillons ont la même espérance $\mathbb{E}_{\mathbb{P}_1}(X^{(1)}) = \mathbb{E}_{\mathbb{P}_2}(X^{(2)}) = \mu_0 \in \mathbb{R}$. On cherche à utiliser l'information provenant des deux échantillons pour estimer μ_0. Pour ce faire, il est possible d'adapter la méthode de vraisemblance empirique.

Une première méthode

Une méthode générale a été brièvement étudié dans Owen (2001), page 223. Owen propose de maximiser la vraisemblance de deux mesures de probabilité $\mathbb{Q}_1 = \sum_{i=1}^{n_1} q_i^{(1)} \delta_{X_i^{(1)}}$ et $\mathbb{Q}_2 = \sum_{j=1}^{n_2} q_j^{(2)} \delta_{X_j^{(2)}}$ sous une contrainte faisant intervenir les deux jeux de poids simultanément :

$$R(\theta) = \max \left\{ \prod_{i=1}^{n_1} n_1 q_i^{(1)} \prod_{j=1}^{n_2} n_2 q_j^{(2)} \middle| \mathbb{E}_{\mathbb{Q}_1 \otimes \mathbb{Q}_2} h(X^{(1)}, X^{(2)}, \theta) = 0 \right\},$$

où $h : (\mathbb{R}^{p_1}, \mathbb{R}^{p_2}, \mathbb{R}^1) \to \mathbb{R}^1$. Owen énonce alors le théorème suivant :

Théoreme 1.6 (Owen) *Supposons que* $\mathbb{E}_{\mathbb{P}_1\otimes\mathbb{P}_2}[h(X^{(1)},X^{(2)},\theta_0)]=0$*, que* $\min\{n_1,n_2\}$ *tend vers l'infini et que la variance de* $h(X^{(1)},X^{(2)},\theta_0)$ *est finie et strictement positive. Si de plus*

$$\mathbb{E}\left[\mathbb{E}[h(X^{(1)},X^{(2)},\mu)|X^{(1)}]^2\right]>0 \text{ ou } \mathbb{E}\left[\mathbb{E}[h(X^{(1)},X^{(2)},\mu)|X^{(2)}]^2\right]>0,$$

alors

$$-2\log R(\theta_0)\xrightarrow[n\to\infty]{\mathcal{L}}\chi^2_1.$$

Pour appliquer ce théorème à notre problème, il faut trouver une fonction h telle que nos deux contraintes $\mathbb{E}_{\mathbb{P}_1}(X^{(1)})=\mu_0$ et $\mathbb{E}_{\mathbb{P}_2}(X^{(2)})=\mu_0$ sous la forme $\mathbb{E}_{\mathbb{P}_1\otimes\mathbb{P}_2}h(X^{(1)},X^{(2)},\mu_0)=0$. Malheureusement, on ne peut poser

$$h(X^{(1)},X^{(2)},\mu)=(X^{(1)}-\mu)\cdot(X^{(2)}-\mu).$$

En effet, dans ce cas les variances des deux espérances conditionnelles de h sont nulles. On peut par contre tester l'égalité des espérances, en posant $h(X^{(1)},X^{(2)},\theta)=X^{(1)}-X^{(2)}-\theta$, mais cette formulation ne permet pas d'estimer μ_0.

Une méthode adaptée

On considère également la vraisemblance : $\prod_{i=1}^{n_1}n_1q_i^{(1)}\prod_{j=1}^{n_2}n_2q_j^{(2)}$. L'idée est de maximiser la vraisemblance sous les deux contraintes données par notre modèle :

$$C(\mu)=\left\{\left(q_i^{(1)}\right)_{1\le i\le n_1},\left(q_j^{(2)}\right)_{1\le j\le n_2}\middle|\forall r\in\{1,2\},\ \sum_{i=1}^{n_r}q_i^{(r)}X_i^{(r)}=\mu,\ \sum_{i=1}^{n_r}q_i^{(r)}=1,\right\}.$$

À μ fixé,

$$L_{n_1,n_2}(\mu)=\sup_{C(\mu)}\left\{\prod_{i=1}^{n_1}n_1q_i^{(1)}\prod_{j=1}^{n_2}n_2q_j^{(2)}\right\}$$

peut s'écrire comme le produit de 2 programmes indépendants : $L_{n_1,n_2}(\mu)=L_{n_1}(\mu)L_{n_2}(\mu)$ avec pour $r=1,2$, $L_{n_r}(\mu)=\sup_{\left(q_i^{(r)}\right)}\left\{\prod_{i=1}^{n_r}n_rq_i^{(r)}\right\}$, chaque jeu de poids vérifiant ses contraintes respectives. En construisant le Lagrangien pour chaque optimisation séparément, on obtient

$$l_{n_r}(\mu)=-\log\left[L_{n_r}(\mu)\right]=\sup_{\lambda_r}\left\{\sum_{i=1}^{n_r}\log\left[1+\lambda_r'\left(X_i^{(r)}-\mu\right)\right]\right\}$$

où $\lambda_r\in\mathbb{R}^d$ est le multiplicateur de Lagrange associé à la contrainte $\sum_{i=1}^{n_r}q_i^{(r)}X_i^{(r)}=\mu$. Le programme de vraisemblance empirique est finalement équivalent à :

$$l_{n_1,n_2}(\mu)=-\log\left[L_{n_1,n_2}(\mu)\right]=\sup_{\lambda_1,\lambda_2}\left\{\sum_{i=1}^{n_1}\log\left[1+\lambda_1'\left(X_i^{(1)}-\mu\right)\right]+\sum_{j=1}^{n_2}\log\left[1+\lambda_2'\left(X_j^{(2)}-\mu\right)\right]\right\}.$$

On peut définir le rapport de vraisemblance, $r_{n_1,n_2}(\mu)=2\left[l_{n_1,n_2}(\widehat{\mu})-l_{n_1,n_2}(\mu)\right]$, où $\widehat{\mu}$ est l'estimateur de μ donnée par l'$\arg\sup_\mu l_{n_1,n_2}(\mu)$.

Théoreme 5.1
Soit $\left(X_i^{(1)}\right)_{1\leq i\leq n_1} \sim \mathbb{P}_1$ *i.i.d. et* $\left(X_j^{(2)}\right)_{1\leq j\leq n_2} \sim \mathbb{P}_2$ *i.i.d., deux échantillons indépendants de même espérance* $\mu_0 \in \mathbb{R}^d$. *Supposons, pour* $r = 1, 2$, *que la matrice de variance-covariance de* $X^{(r)} - \mu_0$ *est inversible et que de plus* $\mathbb{E}\left[||X^{(r)} - \mu||^3\right] < \infty$ *sur un voisinage de* μ_0. *Supposons enfin que* $\min(n_1, n_2)$ *tend vers l'infini et que* $\log\log\max(n_1, n_2) = o\left(\min(n_1, n_2)^{1/3}\right)$, *alors*

$$r_{n_1,n_2}(\mu_0) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_d.$$

Une région de confiance pour μ_0 est alors donnée par $\left\{\mu \,\middle|\, r_{n_1,n_2}(\mu) \leq F_{\chi^2_d}(1-\alpha)\right\}$. Ce résultat nous permet d'obtenir une méthode prenant en compte les différents échantillons de données de consommations disponibles dans le cadre d'une étude de risque alimentaire au chapitre 5.

1.2 Les divergences empiriques : une généralisation de la vraisemblance empirique

Comme nous l'avons souligné ci-dessus, la vraisemblance empirique peut s'interpréter comme une méthode de contraste basée sur la divergence de Kullback. Des méthodes similaires basées sur d'autres divergences ont été proposé dans la littérature. Parmi les divergences utilisées, on peut citer l'entropie relative (Kitamura & Stutzer, 1997) et la divergence du χ^2 (Hansen et al., 1996). Newey & Smith (2004) ont montré que l'idée peut être généralisée et que la méthode est valide pour toute divergence issue de la famille des Cressie-Read, qui contient l'entropie relative et les divergences de Kullback et du χ^2. Dans le chapitre 3, nous démontrerons que les résultats restent valides pour une classe bien plus large de divergences. Dans cette optique, nous introduisons un ensemble d'outils liés à la dualité convexe.

1.2.1 Divergences et dualité convexe

Afin de généraliser la méthode de vraisemblance empirique, on rappelle quelques notions sur les φ-divergences introduites par Csiszár (1967). On pourra se référer à Rockafellar (1968, 1970, 1971) et Liese & Vajda (1987) pour plus de précisions et un historique de ces métriques. Broniatowski & Kéziou (2004) utilisent également ces notions dans un cadre paramétrique.

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. Soit f une fonction mesurable définie de $\mathcal{X}$ dans $\mathbb{R}^p$ et R une mesure appartenant à $\mathcal{M}$, on note $Rf = \int f(x)R(dx)$.

On utilise dans toute la suite la notation φ pour des fonctions convexes. On note

$$d(\varphi) = \{x \in \mathbb{R} | \varphi(x) < \infty\}$$

le domaine de φ et respectivement $\inf d(\varphi)$ et $\sup d(\varphi)$ les points terminaux de ce domaine. Pour toute fonction φ convexe, on introduit sa conjuguée convexe φ^* ou transformée de Fenchel-Legendre

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

Nous ferons les hypothèses suivantes sur la fonction φ.

Hypothèses 1.1

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

(ii) *φ est deux fois différentiable sur un voisinage de 0,*

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

(iv) *$\varphi^{(2)}(0) > 0$, ce qui implique que φ admet un unique minimum en 0,*

(v) *La dérivée seconde de φ est minorée par $c > 0$ sur $d(\varphi) \cap \mathbb{R}^+ (\neq \emptyset)$.*

Les hypothèses sur la valeur de φ en 0 correspondent essentiellement à une renormalisation (cf. Rao & Ren, 1991). L'hypothèse (v) est moins générale et nous servira pour établir notre principal résultat, le Théorème 2.3 . Elle est vérifiée en particulier lorsque $\varphi^{(1)}$ est elle-même convexe (entraînant $\varphi^{(2)}(x)$ croissante donc $\varphi^{(2)}(x) \geq \varphi^{(2)}(0) > 0$ pour x dans $\mathbb{R}^+$), ce qui est le cas pour toutes les divergences couramment rencontrées. Elle n'est pas nécessaire pour la définition suivante.

La φ-divergence associée à φ, appliquée à $\mathbb{Q}$ et $\mathbb{P}$, où $\mathbb{Q}$ et $\mathbb{P}$ sont des mesures respectivement signée et positive, est définie par :

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

Ces pseudo-métriques introduites par Rockafellar (1968 et 1970) sont en fait des cas particuliers de « distances » convexes (Liese-Vajda, 1987). En particulier, l'intérêt des φ-divergences réside dans le théorème suivant (réécrit sous une forme simplifiée) dû à Borwein & Lewis (1991) (voir également Léonard, 2001).

Théoreme 1.7 (Minimisation et Conjugaison) *Soit φ une fonction convexe partout finie et différentiable telle que $\varphi^* \geq 0$ et $\varphi^*(0) = 0$. Soit $\mathbb{P}$ une mesure de probabilité discrète. Alors il vient*

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

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

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

alors il existe $\mathbb{Q}^\diamond$ et $\lambda^\diamond$ réalisant respectivement l' inf *et le* sup *et tels que*

$$\mathbb{Q}^\diamond = \left(1 + \varphi^{(1)}(\lambda^{\diamond\prime} m)\right) \mathbb{P}.$$

Ce théorème nous servira d'équivalent à la méthode de Lagrange utilisée pour l'optimisation dans le cas de la vraisemblance empirique. Il permet de passer d'une optimisation portant sur la mesure $\mathbb{Q}$ à une optimisation bien plus simple, portant sur le vecteur λ, qui joue le rôle du multiplicateur de Lagrange.

1.2.2 Extension de la méthode de vraisemblance empirique aux φ-divergences.

L'objectif du chapitre 2 est d'étendre la méthode de vraisemblance empirique 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_{φ^*}. Nous nous plaçons ici dans le cadre introduit dans la section 1.1.2. On dispose alors d'une suite de vecteurs aléatoires $X, X_1, \cdots, X_n$ de $\mathbb{R}^p$, $n \geq 1$, indépendants et identiquement distribués de loi de probabilité $\mathbb{P}_0$ dans un espace de probabilité $\mathcal{P}$. On note Pr la probabilité sous la loi jointe $\mathbb{P}_0^{\otimes n}$ de $(X_1, ..., X_n)$. On cherche à estimer un paramètre θ_0 défini par d'une équation de moments de la forme :

$$\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0)] = 0$$

où m est une fonction régulière de $\mathcal{X} \times \mathbb{R}^p$ dans $\mathbb{R}^r$ avec $r \geq p$.

Minimisation empirique des φ-divergences

Comme nous l'avons remarqué, la vraisemblance empirique peut s'interpréter comme une méthode de contraste, dont la fonction de contraste est basée sur la divergence de Kullback. Nous démontrerons au chapitre 2 que l'on peut remplacer la divergence de Kullback par une large classe de φ-divergences et obtenir ainsi une généralisation de la vraisemblance empirique. Pour une fonction φ donnée, nous proposons de définir désormais la statistique pivotale $\beta_n^{\varphi}(\theta)$ comme le minimum de la φ-divergence empirique associée, sous la contrainte de l'équation d'estimation :

$$\beta_n^{\varphi}(\theta) = 2n \inf_{\mathbb{Q} \in \mathcal{M}_n} \left\{ I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \,\middle|\, \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0 \right\},$$

où $\mathcal{M}_n$ est l'ensemble des mesures signées dominées par $\mathbb{P}_n$. La région de confiance C_{η,n,φ^*} correspondante s'écrit

$$C_{\eta,n,\varphi^*} = \left\{ \theta \,\middle|\, \beta_n^{\varphi}(\theta) \leq \eta \right\} = \left\{ \theta \,\middle|\, \exists \mathbb{Q} \in \mathcal{M}_n,\ \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0,\ 2nI_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \leq \eta \right\}.$$

Pour $\mathbb{Q}$ dans $\mathcal{M}_n$, on peut réécrire les contraintes de minimisation sous la forme

$$\mathbb{E}_{(\mathbb{Q}-\mathbb{P}_n)}[m(X, \theta)] = -\overline{m}_n(\theta),\ \text{où } \overline{m}_n(\theta) = \mathbb{E}_{\mathbb{P}_n}[m(X, \theta)].$$

Le Théorème 1.7 permet d'écrire

$$\begin{aligned}
\beta_n^{\varphi}(\theta) &= 2n \inf_{\mathbb{Q} \in \mathcal{M}_n} \left\{ I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \,\middle|\, \mathbb{E}_{\mathbb{Q}}[m(X, \theta)] = 0 \right\} \\
&= 2n \inf_{\mathbb{Q} \in \mathcal{M}_n} \left\{ I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \,\middle|\, \mathbb{E}_{(\mathbb{Q}-\mathbb{P}_n)}[m(X, \theta)] = -\overline{m}_n(\theta) \right\} \\
&= 2n \sup_{\lambda \in \mathbb{R}^p} \left\{ -\lambda' \overline{m}_n(\theta) - \int_{\Omega} \varphi(\lambda' m(X, \theta)) d\mathbb{P}_n \right\}.
\end{aligned}$$

On en déduit l'expression duale de $\beta_n^{\varphi}(\theta)$ qui permet de généraliser les propriétés usuelles de la vraisemblance empirique :

$$\beta_n^{\varphi}(\theta) = 2 \sup_{\lambda \in \mathbb{R}^p} \left\{ -\sum_{i=1}^{n} \lambda' m(X_i, \theta) - \sum_{i=1}^{n} \varphi(\lambda' m(X_i, \theta)) \right\}. \tag{Dual}$$

L'écriture duale permet d'établir le théorème suivant :

Théoreme 2.3 *Soient $X_1, \cdots, X_n$ des vecteurs aléatoires de $\mathbb{R}^p$, i.i.d. de loi $\mathbb{P}_0$ absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^p$. Soit θ_0 tel que $\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0)] = 0$ et que $\mathbb{E}_{\mathbb{P}_0}[m(X, \theta_0) m(X, \theta_0)']$ soit de rang q. Si φ vérifie les hypothèses 1.1 alors, quelque soit $\eta > 0$, C_{η,n,φ^*} est convexe et*

$$\Pr(\theta \in C_{\eta,n,\varphi^*}) = \Pr(\beta_n^{\varphi}(\theta) \leq \eta) \xrightarrow[n\to\infty]{} F_{\chi^2_q}\left(\frac{\eta}{\varphi^{(2)}(0)} \right).$$

Divergences Empiriques couramment utilisées

Dans le cas particulier où $\varphi^*(x) = x - \log(1+x)$, on obtient la vraisemblance empirique. D'autres fonctions φ sont couramment utilisées en Statistique et en Économétrie, comme alternatives à la vraisemblance empirique. Le Théorème 2.3 permet de donner un cadre commun à toutes ces méthodes, et d'en introduire de nouvelles.

Les divergences les plus utilisées (Kullback, entropie relative, χ^2 et Hellinger) se regroupent dans la famille des Cressie-Read (voir Csiszár, 1967, Cressie & Read, 1984).

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

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

La vraisemblance empirique a été généralisé à l'aide de cette famille de divergence par Newey & Smith (2004) sous le nom de « Generalized Empirical Likelihood ». On notera que dans le cas particulier du χ^2, c'est-à-dire $\varphi_\alpha^*(x) = \frac{x^2}{2}$, on peut calculer la valeur explicite du multiplicateur de Lagrange et donc la valeur de la vraisemblance en un point θ. D'un point de vue algorithmique, c'est donc la plus simple. Pour l'étude de l'estimateur basé sur le χ^2, très lié aux GMM (Generalized Method of Moments), on se référera à Bonnal & Renault (2004) ainsi qu'à Newey & Smith (2004). L'entropie relative (φ_1) conduit au KLIC (Kullback-Leibler Information Criterion) qui est différent du maximum de vraisemblance empirique et a été étudié sous ce nom par Kitamura & Stutzer (1997).

1.3 Quasi-vraisemblance empirique

1.3.1 Motivation

On s'intéresse dans les chapitres 2 et 3 à l'étude d'une famille particulière de fonctions, les Quasi-Kullback, qui permettent de construire des φ-divergences aux propriétés intéressantes.

La famille des Quasi-Kullback est constituée des barycentres des divergences de Kullback et du χ^2 :

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

On obtient alors la statistique pivotale $\beta_n^{K_\varepsilon^*}(\theta)$, plus simplement notée $\beta_n^\varepsilon(\theta)$:

$$\begin{aligned}\beta_n^\varepsilon(\theta) &= \sup_{\lambda\in\mathbb{R}^q}\left\{-n\lambda'\overline{m}_n(\theta) - \sum_{i=1}^n K_\varepsilon\left(\lambda' m(X_i,\theta)\right)\right\}\\ &= \sup_{\lambda\in\mathbb{R}^q}\left\{-n\varepsilon\left[\lambda'\overline{m}_n(\theta) + \frac{\lambda' S_n^2(\theta)\lambda}{2}\right] + (1-\varepsilon)\sum_{i=1}^n \log\left(1-\lambda' m(X_i,\theta)\right)\right\},\end{aligned}$$

où $S_n^2(\theta) = \mathbb{E}_{\mathbb{P}_n}[m(X,\theta)m(X,\theta)']$.

La famille des Quasi-Kullback vérifie nos hypothèses 1.1 et l'on obtient la convergence en loi de $\beta_n^\varepsilon(\theta_0)$ vers un χ^2 grâce au Théorème 2.3 . On peut expliciter K_ε^*, qui est finie quelque soit $x \in \mathbb{R}$ pour $\varepsilon > 0$:

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

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

L'idée est de concilier les avantages de la divergence de Kullback (l'aspect adaptatif des régions de confiance et la correction de Bartlett) et de la divergence du χ^2 (la robustesse et la simplicité algorithmique). Pour motiver l'étude de cette famille, nous donnons ici certaines propriétés intéressantes des Quasi-Kullback, démontrées aux chapitres 2 et 3.

- Conformément aux attentes, la forme des régions s'adapte d'autant plus aux données que ε est proche de 0, la valeur correspondant à la divergence de Kullback (Figure 1.5).
- On obtient la correction de Bartlett, pour ε petit (Théorème 2.4).
- Dès que $\varepsilon > 0$, les régions de confiance peuvent dépasser de l'enveloppe convexe des données, ce qui augmente fortement la robustesse de la méthode.
- On obtient de plus un contrôle à distance fini du niveau de la région de confiance, grâce à une inégalité exponentielle explicite (Théorèmes 3.4 et 3.5).
- Enfin, contrairement à la divergence de Kullback, la conjuguée convexe de K_ε est définie sur $\mathbb{R}$ et est relativement lisse, ce qui simplifie l'implémentation du problème d'optimisation.

Les différentes propriétés ci-dessus influencent de façon contradictoire le choix de ε, qu'il faut donc adapter à chaque problème précis.

1.3.2 Régions de confiance asymptotique et Quasi-Kullback

Les propriétés remarquables des Quasi-Kullback décrites ci-dessus appellent une étude plus complète, en particulier du point de vue du comportement non-asymptotique, peu traité

dans la littérature sur la vraisemblance empirique. On peut se demander comment le choix de ε influence le comportement à distance finie des régions de confiance, en particulier du point de vue du taux de couverture. Ce questionnement devient particulièrement critique pour des données multidimensionnelles dont la loi est éloignée de la gaussienne. Pour motiver et illustrer cette affirmation, on simule des données de mélange d'échelle, c'est-à-dire le produit d'une uniforme sur $[0;1]$ et d'une gaussienne standard sur $\mathbb{R}^6$.

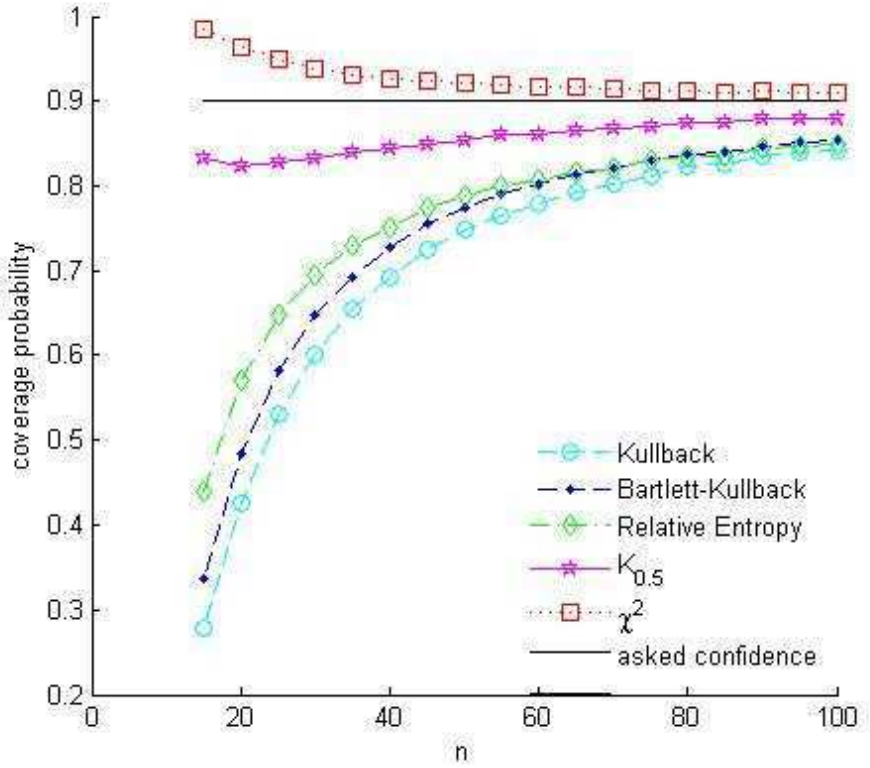


FIG. 1.4 – Taux de couverture pour différentes divergences

La Figure 1.4 présente les taux de couverture obtenus pour des régions de confiance à 90% par simulation de Monte-Carlo (100 000 répétitions), avec des divergences communément utilisées et une Quasi-Kullback (pour $\varepsilon = 0.5$). Asymptotiquement, toutes ces divergences sont équivalentes d'après le Théorème 2.3 . Pourtant, ces simulations montrent clairement que les comportements à distance finie sont très différents. La vraisemblance empirique, qui correspond à la divergence de Kullback, est loin du taux asymptotique lorsque l'échantillon est petit, même lorsqu'elle a été corrigée au sens de Bartlett. C'est le coût de l'adaptation de la forme des régions de confiance aux données, également payé par l'entropie relative. La divergence du χ^2 s'avère à l'inverse trop conservative, et procure des taux de couverture supérieurs au niveau de confiance visé. On observe que l'utilisation de la Quasi-Kullback permet de se placer « entre » la divergence de Kullback et le χ^2. La Figure 1.5 illustre ce comportement intermédiaire pour les régions de confiance, en reprenant l'échantillon d'Owen.

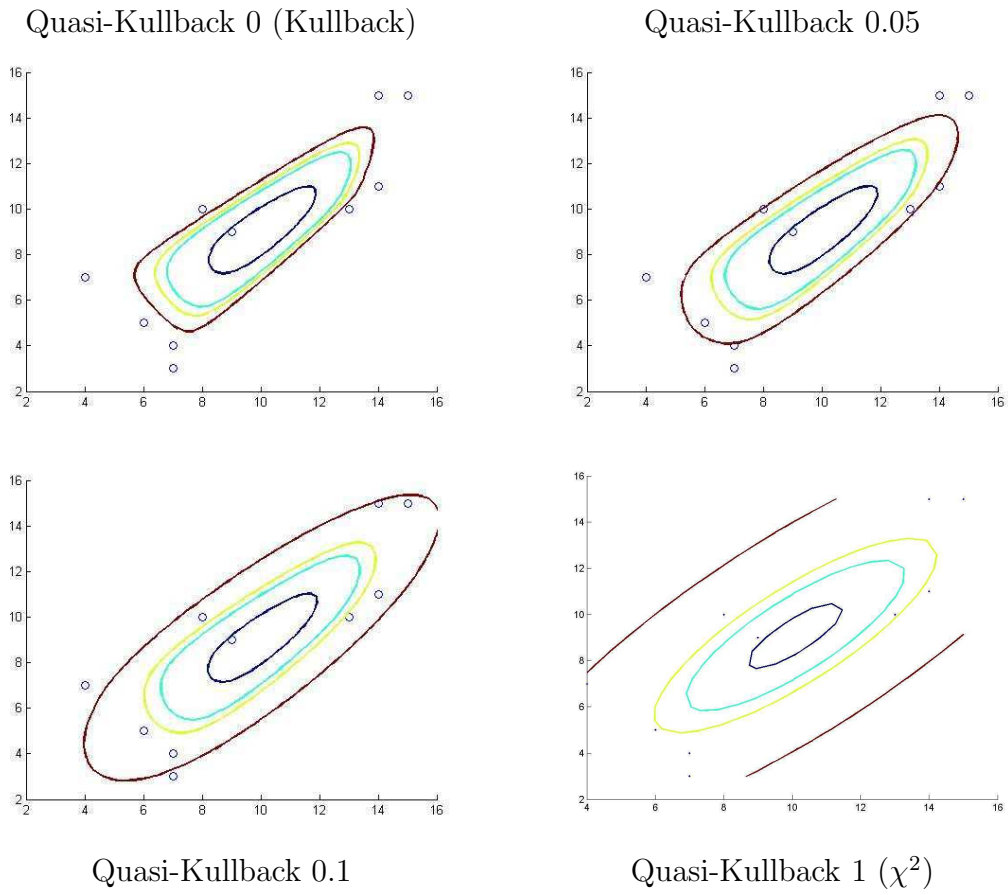


FIG. 1.5 – Régions de confiance pour les onze canards d'Owen, en fonction de ε

Par ailleurs, la fonction K^*_ε étant définie sur $\mathbb{R}$, les pondérations q_i peuvent prendre des valeurs négatives ou nulles sans faire exploser la divergence. On s'affranchie alors de la limitation à l'enveloppe convexe des données, dont les effets ont été soulignés par Tsao (2004) et que nous avons abordé au paragraphe 1.1.1, tout en conservant la possibilité de pratiquer une correction de Bartlett.

Correction de Bartlett

Puisque, pour ε suffisamment petit, les Quasi-Kullback se comportent comme la vraisemblance empirique, on peut espérer obtenir une correction de Bartlett. Nous montrerons dans le chapitre 2 le théorème suivant :

Théoreme 2.4 *Si on choisit pour ε une suite $\varepsilon_n = \mathcal{O}(n^{-3/2}\log(n)^{-1})$, les Quasi-Kullback sont corrigeables au sens de Bartlett, jusqu'à l'ordre $\mathcal{O}(n^{-3/2})$.*

On arrive, grâce à la correction de Bartlett pour les Quasi-Kullback, à améliorer notre taux de couverture des régions de confiance, comme l'illustre la Figure 1.6 .

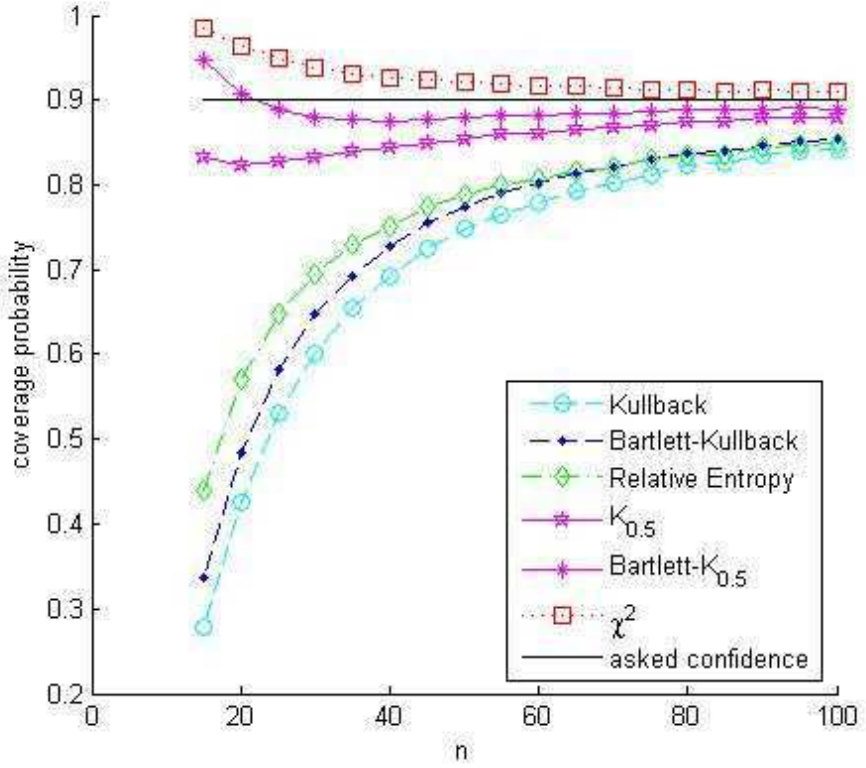


FIG. 1.6 – Taux de couverture et Quasi-Kullback

Sur nos simulations, il apparaît que l'introduction des Quasi-Kullback apporte un début de solution au problème de comportement à distance finie, puisque le courbe des taux de couverture obtenues se glisse entre celle du χ^2 et celle de la divergence de Kullback, et se trouve améliorée par la correction établie au Théorème 2.4.

Quasi-Kullback adaptative

Les simulations ci-dessus montrent que le choix de ε permet d'améliorer le taux de couverture des régions de confiance. En particulier, pour n petit, il semble peu raisonnable de choisir ε trop petit. À l'inverse, la théorie asymptotique, à travers la correction de Bartlett, nous incite à choisir des ε petits et diminuant avec n. L'idéal serait de choisir ε en fonction de n, de la dimension et du comportement des données.

Nous proposons au chapitre 3 une méthode de validation croisée pour calibrer ε. Bien sûr, cette méthode demande un temps de calcul important, mais il s'agit de la mettre en pratique dans des cas où n est relativement petit, lorsque les résultats asymptotiques sont trop approximatifs. Pour n grand, on est théoriquement incité à utiliser la divergence de Kullback avec correction de Bartlett, mais toutes les régions de confiance sont asymptotiquement équivalentes, et le temps de calcul est bien plus faible dans le cas de la divergence du χ^2.

1.3.3 Bornes multidimensionnelles explicites

Dans la section précédente, on s'est intéressé à l'influence du choix de ε sur les taux de couverture obtenus pour des régions de confiance **asymptotiques**. On a vu qu'en choisissant ε de façon adaptative, on peut corriger la différence entre le niveau asymptotique et le niveau à distance finie. Une autre approche, plus théorique, est de chercher à contrôler directement un niveau à distance finie. On cherche alors à obtenir des bornes pour l'erreur de première espèce. Évidemment, plus la borne est générale, plus elle est grossière. Nous démonterons au chapitre 3 le théorème suivant qui permettra d'obtenir ce type de bornes via des bornes sur les sommes autonormalisées :

Théoreme 3.6 *Si $X_1, \cdots, X_n$ sont des vecteurs aléatoires de $\mathcal{X}$, i.i.d. de loi $\mathbb{P}_0$. Soit m de $\mathcal{X} \times \mathbb{R}^p$ dans $\mathbb{R}^q$ telle que $\mathbb{E}[m(X_1, \theta_0)] = 0$. Si de plus $\mathbb{E}[m(X_1, \theta_0)m'(X_1, \theta_0)]$ est de rang q, alors quelque soit $n > q$ fixé, c'est-à-dire* ***non asymptotiquement****,*

$$\begin{aligned}\Pr(\theta_0 \notin C_{\eta,n,K_\varepsilon^*}) &= \Pr(\beta_n^\varepsilon(\theta_0) \geq \eta) \\ &\leq \Pr\left(\frac{n}{2}\overline{m}_n'(\theta_0)S_n^{-2}(\theta_0)\overline{m}_n(\theta_0) \geq \eta\varepsilon\right)\end{aligned}$$

où $\overline{m}_n(\theta_0)$ et $S_n^2(\theta_0)$ sont les moyennes empiriques des $m(X_i, \theta_0)$ et des $m(X_i, \theta_0)m'(X_i, \theta_0)$.

Sommes autonormalisées

Des bornes exponentielles ont été établies pour ce type de statistiques, dans le cas unidimensionnel. On peut les établir à partir de bornes de Berry-Esséen non-uniformes ou à partir de bornes de Cramer (Shao, 1997, Jing & Wang, 1999, Chistyakov & Götze, 2003, Jing et al., 2003). Néanmoins, à notre connaissance, on ne dispose de bornes exponentielles dont les constantes sont explicites que dans le cas d'une distribution symétrique, c'est-à-dire telle que Z et $-Z$ aient même loi. Nous rappelons ces résultats dont l'étude remonte aux résultats de Hoeffding (1963) et Efron (1969) avant d'être complétée par Pinelis (1994).

Théoreme 1.8 (Hoeffding) *Soit $Z_1, \cdots, Z_n$ un échantillon i.i.d. de vecteurs de $\mathbb{R}^q$ de loi centrée et symétrique. Posons $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$ et $S_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$. Alors, sans aucune hypothèse de moments,*

$$\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) \leq 2q\exp(-u/2q).$$

Cette borne à l'avantage d'être explicite, de formulation exponentielle et simple, ce qui permet de la réutiliser pour obtenir des bornes dans des cas plus généraux, comme c'est le cas au paragraphe suivant. On peut obtenir une meilleure borne, grâce à un résultat de Pinelis (1994) :

Théoreme 1.9 (Pinelis) *Sous les hypothèses du Théorème 1.8,*

$$\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) \leq \frac{2e^3}{9}\overline{F}_{\chi_q^2}(u),$$

où $\overline{F}_{\chi_q^2} = 1 - F_{\chi_q^2}$ est la fonction de survie d'un χ_q^2.

En dehors du cas symétrique, on sait depuis Bahadur & Savage (1956), qu'il est impossible d'obtenir des bornes exponentielles indépendantes de moments γ_k d'ordre élevé, avec $\gamma_k = \mathbb{E}[|S^{-1}Z_1|^k]$ et $S = \mathbb{E}[Z_1 Z_1']^{1/2}$ (voir également Romano & Wolf, 2000). Jing & Wang (1999) donnent par exemple une borne dans le cas unidimensionnel non symétrique. Supposons $\gamma_{10/3} < \infty$, alors il existe $A \in \mathbb{R}$ et $a \in]0,1[$ tels que

$$\Pr\left(n\overline{Z}_n^2 S_n^{-2} \geq u\right) \leq \overline{F}_{\chi_1^2}(u) + A\gamma_{10/3} n^{-1/2} e^{-au}. \tag{1.2}$$

Malheureusement, les constantes A et a ne sont pas explicites et la borne ne peut donc pas être utilisée en pratique. Pour établir des bornes explicites au chapitre 3, nous utiliserons une méthode de symétrisation due à Panchenko (2003), ainsi que des éléments de preuves de Bercu et al. (2002).

Théoreme 3.4 *Soit $(Z_i)_{i=1,\cdots,n}$ un échantillon i.i.d. de $\mathbb{R}^q$ de loi $\mathbb{P}$. Supposons S^2 inversible et $\gamma_4 < \infty$. On a alors les inégalités suivantes, **pour** $\mathbf{n > q}$ **fini** et pour tout $a > 1$:*

$$\begin{aligned}\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) &\leq \left\{2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}}\gamma_4^{-\widetilde{q}} e^{-\frac{n}{\gamma_4(q+1)}\left(1-\frac{1}{a}\right)^2}\right\} \\ &\leq \left\{2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n}{\gamma_4(q+1)}\left(1-\frac{1}{a}\right)^2}\right\}\end{aligned}$$

où $\widetilde{q} = \frac{q-1}{q+1}$ et $C(q) = \frac{(2e\pi)^{2\widetilde{q}}(q+1)}{2^{2/(q+1)}(q-1)^{3\widetilde{q}}} \leq \frac{(2e\pi)^2(q+1)}{(q-1)^{3\widetilde{q}}} \leq 18$. De plus, si $nq \leq u$,

$$\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) = 0,$$

L'idée est de se ramener au cas symétrique grâce au lemme de symétrisation de Panchenko (2003). Dans le cas multidimensionnel, ceci nécessite de contrôler la plus petite valeur propre de S_n^2. C'est ce contrôle qui introduit la seconde exponentielle dans la borne et le moment γ_4.

On peut également généraliser la borne obtenue au Théorème 1.9 au cas non symétrique, en utilisant les mêmes méthodes de symétrisations. On obtient alors le théorème suivant :

Théoreme 3.5 *Sous les hypothèses du Théorème 3.4, **pour** $\mathbf{n > q}$ **fini**, pour tout $a > 1$ et pour u tel que $2q(1+a) \leq u \leq nq$,*

$$\begin{aligned}\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) &\leq \frac{2e^3}{9\Gamma(\frac{q}{2}+1)}\left(\frac{u-q(1+a)}{2(1+a)}\right)^{\frac{q}{2}} e^{-\frac{u-q(1+a)}{2(1+a)}} + C(q)\left(\frac{n^3}{\gamma_4}\right)^{\widetilde{q}} e^{-\frac{n\left(1-\frac{1}{a}\right)^2}{\gamma_4(q+1)}} \\ &\leq \frac{2e^3}{9\Gamma(\frac{q}{2}+1)}\left(\frac{u-q(1+a)}{2(1+a)}\right)^{\frac{q}{2}} e^{-\frac{u-q(1+a)}{2(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n\left(1-\frac{1}{a}\right)^2}{\gamma_4(q+1)}}\end{aligned}$$

De plus, si $nq \leq u$,

$$\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) = 0,$$

On a alors le choix entre 2 bornes. Pour utiliser ces bornes, on va prendre l'inf en a, à u fixé, la borne la plus fine étant celle dont l'inf est le plus petit. La Figure 1.7 montre que

pour $q = 1$, c'est la borne du Théorème 3.4 qui est la plus précise, en plus d'être la plus simple. Par contre, pour $q \geq 2$, c'est le Théorème 3.5 qu'il convient utiliser.

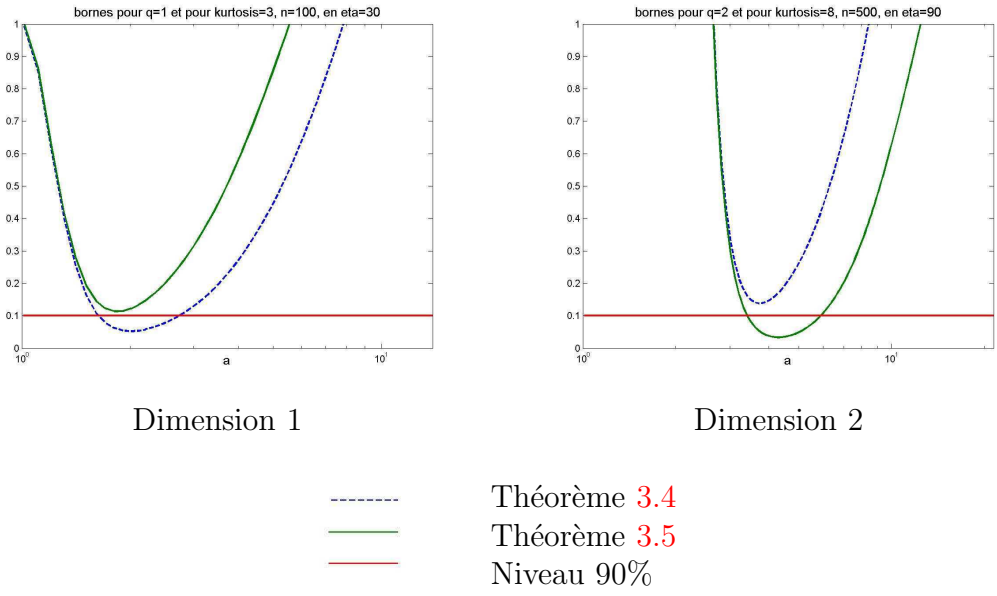


FIG. 1.7 – Comportement des bornes non asymptotiques en fonction de a

Bornes exponentielles explicites pour les Quasi-Kullback

On utilise ces bornes, intéressantes par elles-mêmes dans le cadre des sommes autonormalisées, pour minorer le taux de couverture des régions de confiance construites à l'aide des Quasi-Kullback, grâce au Théorème 3.6 . Dans le cas symétrique, la borne obtenue par Pinelis (1994) est la meilleure, pour pratiquement tout niveau u et quelque soit la dimension q. On applique donc le Théorème 1.9, plutôt que le Théorème 1.8, pour majorer notre erreur de première espèce. Dans le cas général, on utilise la borne du Théorème 3.4 pour $q = 1$ et celle du Théorème 3.5 pour $q \geq 2$.

Hjort et al. (2004) étudient la convergence de la méthode de vraisemblance empirique lorsque q croît avec n. Ils montrent en particulier que le Théorème de Wilks (la loi limite en χ^2_q) reste valide tant que $q = O(n^{\frac{1}{3}})$. Nos bornes montrent que l'on peut aller jusqu'à l'ordre $q = O\left(\frac{n}{log(n)}\right)$. Il est intéressant de remarquer dans ce contexte que la constante $C(q)$ reste bornée lorsque q diverge (voir le chapitre 3). On peut enfin remarquer que nos bornes ne sont pas valides pour $\varepsilon = 0$, c'est-à-dire dans le cas de la vraisemblance empirique. Ceci entrerait en contradiction avec les résultats de Tsao (2004), cités plus haut, qui donne une borne minorant le niveau d'erreur.

1.4 Généralisation aux chaînes de Markov

Dans le chapitre 4, nous cherchons à étendre la méthode de vraisemblance empirique à un cadre non i.i.d. . Kitamura (1997) a proposé une méthode inspirée du Bootstrap pour appliquer la vraisemblance empirique à des données faiblement dépendantes. Nous nous

inspirons d'un résultat plus récent sur le Bootstrap, de Bertail & Clémençon (2004a), qui améliore la vitesse de convergence de l'estimateur obtenu et qui s'applique à des données dont la dépendance est décrite par une chaîne de Markov. L'idée est d'utiliser la structure markovienne des données pour découper l'échantillon en blocs indépendants, auxquels nous pourrons appliquer la méthode de vraisemblance empirique.

1.4.1 Notations et définitions relatives aux chaînes de Markov

Le résultat de Bertail & Clémençon (2004a) est valable pour les chaînes de Markov possédant un atome ou pour les chaînes que l'on peut artificiellement étendre pour les rendre atomiques. Nous rappelons ici quelques notions relatives aux chaînes de Markov.

Une chaîne sur un espace E ($\mathbb{R}^d$ ou $\mathbb{Z}^d$ pour simplifier) est ψ-irréductible lorsque, quelque soit l'état initial x, si $\psi(A) > 0$ alors la chaîne visite A avec probabilité 1. On se donne une chaîne de Markov $X = (X_n)_{n\in\mathbb{N}}$ apériodique, ψ-irréductible, de probabilité de transition Π, et de distribution initiale ν. Pour tout ensemble $B \in \mathcal{E}$ et pour tout $n \in \mathbb{N}$, on a donc

$$X_0 \sim \nu \text{ et } \mathbb{P}(X_{n+1} \in B \mid X_0, ..., \ X_n) = \Pi(X_n, \ B) \text{ presque sûrement.}$$

Posons $\mathbb{P}_\nu$ et $\mathbb{P}_x$ (pour x dans E) les lois de X lorsque $X_0 \sim \nu$ et $X_0 = x$ respectivement. $\mathbb{E}_\nu[\cdot]$ est l'espérance sous $\mathbb{P}_\nu$ et $\mathbb{E}_x[\cdot]$ celle sous $\mathbb{P}_x$. Enfin, pour tout ensemble $B \in \mathcal{E}$, on définit l'espérance $\mathbb{E}_B[\cdot] = \mathbb{E}[\cdot|X_0 \in B]$.

On suppose également qu'il existe une mesure de probabilité μ sur E invariante pour X, c'est-à-dire telle que $\mu\Pi = \mu$, où $\mu\Pi(dy) = \int_{x\in E} \mu(dx)\Pi(x, dy)$ (μ est alors unique).

Considérons une fonction mesurable $m : E \times \mathbb{R}^p \to \mathbb{R}^p$ et un paramètre d'intérêt satisfaisant l'équation de moment

$$\mathbb{E}_\mu[m(X, \theta_0)] = 0$$

1.4.2 Vraisemblance empirique et chaîne de Markov

Pour un atome A de la chaîne X, on définit $\tau_A = \tau_A(1) = \inf\{k \geq 1, \ X_k \in A\}$ le temps d'atteinte de A et $\tau_A(j) = \inf\{k > \tau_A(j-1), \ X_k \in A\}$ les temps de retour successifs en A.

Pour estimer θ_0, plutôt que d'utiliser directement $m(X_i, \theta_0)$ comme dans le cas i.i.d., nous construisons des blocs de données

$$B_i = (X_{\tau_A(i)+1}, ..., \ X_{\tau_A(i+1)}).$$

Ces blocs permettent d'observer la structure de la chaîne avec sa mémoire. Nous appliquerons la méthode de vraisemblance empirique aux blocs plutôt qu'à la chaîne elle-même. Posons

$$M(B_i, \theta) = \sum_{k=\tau_A(i)+1}^{\tau_A(i+1)} m(X_k, \theta).$$

Certaines hypothèses sur la mémoire de la chaîne sont nécessaires. Soit $\kappa > 0$ et ν une mesure sur E, on définit les hypothèses suivantes sur le temps de retour au petit ensemble :

$$\mathbf{H0}(\kappa) : \ \mathbb{E}_A[\tau_A^\kappa] < \infty,$$
$$\mathbf{H0}(\kappa, \ \nu) : \ \mathbb{E}_\nu[\tau_A^\kappa] < \infty,$$

Soit également les conditions d'intégrabilité par « bloc » :

$$\mathbf{H1}(\kappa,\ m):\ \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A}||m(X_i,\theta_0)||\right)^{\kappa}\right]<\infty,$$
$$\mathbf{H1}(\kappa,\ \nu,\ m):\ \mathbb{E}_\nu\left[\left(\sum_{i=1}^{\tau_A}||m(X_i,\theta_0)||\right)^{\kappa}\right]<\infty.$$

Nous proposons alors un algorithme pour appliquer la méthode de φ-divergence empirique aux blocs.

1. Compter le nombre de visites à l'atome $l_n+1=\sum_{i=1}^n \mathbb{1}_{X_i\in A}$.
2. Diviser la chaîne $X^{(n)}=(X_1,....,X_n)$ en l_n+2 blocs correspondant au parcours de la chaîne entre deux visites à l'atome A,
$$B_0=(X_1,...,\ X_{\tau_A(1)}),\ B_1=(X_{\tau_A(1)+1},...,\ X_{\tau_A(2)}),...,$$
$$B_{l_n}=(X_{\tau_A(l_n)+1},...,\ X_{\tau_A(l_n+1)}),\ B^{(n)}_{l_n+1}=(X_{\tau_A(l_n+1)+1},...,\ X_n),$$
avec la convention $B^{(n)}_{l_n+1}=\emptyset$ lorsque $\tau_A(l_n+1)=n$.
3. On ne prend en compte ni le premier bloc B_0 ni le dernier $B^{(n)}_{l_n+1}$ (éventuellement vide si $\tau_A(l_n+1)=n$).
4. La log-vraisemblance $\beta_n(\theta)$ s'écrit :
$$\beta_n(\theta)=2\sup_{(q_j)_j}\left\{\log\left[\prod_{j=1}^{l_n}l_nq_j\right]\middle|\ \sum_{j=1}^{l_n}q_j\cdot M(B_j,\theta)=0,\ \sum_{j=1}^{l_n}q_j=1\right\}.$$
On l'évalue plus facilement en utilisant la forme duale
$$\beta_n(\theta)=2\sup_{\lambda\in\mathbb{R}^p}\left\{\sum_{j=1}^{l_n}\log\left[1+\lambda' M(B_j,\theta)\right]\right\}.$$
5. La région de confiance
$$C_{\eta,n}=\left\{\theta\,|\,\beta_n(\theta)\le\eta\right\}.$$
est alors de niveau α si l'on prend $\eta=\chi^2_p(1-\alpha)$.

Nous établirons au chapitre 4 le théorème suivant, assurant la validité de notre algorithme :

Théoreme 4.2 *Sous les hypothèses* ***H0***$(1,\nu)$, ***H0***(2) *et* ***H1***$(2,m)$, *si* $\mathbb{E}_A[M(B,\theta_0)M(B,\theta_0)']$ *est inversible alors*
$$\beta_n(\theta_0)\xrightarrow[n\to\infty]{\mathcal{L}}\chi^2_p$$
et donc
$$\mathbb{P}_\nu\left(\theta_0\in C_{\eta,n}\right)\xrightarrow[n\to\infty]{}1-\alpha.$$

Cette méthode peut-être étendue aux chaînes de Markov non atomiques, grâce à une méthode de construction d'un atome par extension de la chaîne introduite par Nummelin (1978). Nous étudions cette généralisation au chapitre 4. Nos résultats s'adaptent au cadre des divergences empiriques, sans difficulté supplémentaire.

1.5 Applications à l'étude de la consommation alimentaire

1.5.1 Risque alimentaire : contamination au mercure

Nous nous intéresserons au chapitre 5 à l'un des principaux problèmes en gestion du risque : la diversité des sources de données. En effet, il existe en France plusieurs jeux de données de consommation, mais ceci présentent des types différents (enquête de budget des ménage, carnet de consommation individuelle, rappel de 24 heures, questionnaire de fréquence) et utilisent différentes méthodologie statistiques (stratification, tirage aléatoire, méthode des quotas). À ces jeux de données de consommation viennent s'ajouter les données de contamination, qui donnent des taux de présence de contaminant dans les aliments. Il est donc essentiel de mettre au point des méthodes permettant de prendre en compte l'information contenue dans ces différents jeux de données en prenant en compte les différences de constitution.

On s'intéresse dans cette partie à l'estimation d'un indice de risque et à la construction d'un intervalle de confiance pour celui-ci. L'indice de risque θ_d mesure la probabilité que l'exposition à un contaminant, le méthylmercure dans notre application, dépasse la dose tolérable quotidienne d. La vraisemblance empirique, grâce au Théorème 5.1, permet de combiner plusieurs jeux de données de consommation et de contamination.

L'intérêt de l'utilisation de la vraisemblance empirique et qu'elle permet de recalculer des pondérations pour les données et donc de s'affranchir des effets des méthodes employées et de objectifs de chaque enquête. On obtient comme estimateur de l'indice de risque 3.27% avec comme intervalle de confiance à 95% [3.08%; 3.47%].

Notations

Nous présentons ici notre méthode dans le cas d'un contaminant présent dans P produits. Pour $k = 1, ..., P$, on note $\left(q_l^{[k]}\right)_{l=1,...,L_k}$ l'échantillon des données de contamination pour le produit k, supposé i.i.d. de distribution $\mathcal{Q}^{[k]}$. La probabilité empirique de la contamination du produit k est donc

$$\mathcal{Q}_{L_k}^{[k]} = \frac{1}{L_k} \sum_{l=1}^{L_k} \delta_{q_l^{[k]}}.$$

On dispose de deux échantillons multidimensionnels pour les consommations, identifiés par $r = 1, 2$. Ainsi, $\left(c_i^{(r)}\right)_{i=1,...,n_r}$ est l'échantillon P-dimensionnel des consommations de l'enquête r, i.i.d. de distribution $\mathcal{C}^{(r)}$. La probabilité empirique de la consommation pour l'enquête r est donc

$$\mathcal{C}_{n_r}^{(r)} = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{c_i^{(r)}}.$$

Pour l'enquête r, la probabilité que l'exposition d'un individu dépasse la dose tolérable d est donc $\theta_d^{(r)} = \Pr\left(D^{(r)} > d\right)$, où $D^{(r)} = \left(q^{[1]}, \cdots, q^{[P]}\right) \cdot c^{(r)}$.

Vraisemblance empirique

On se donne les multinomiales $\widetilde{\mathcal{C}}_{n_r}^{(r)} = \sum_{i=1}^{n_r} p_i^{(r)} \delta_{c_i^{(r)}}$, pour $r = 1, 2$, correspondant aux 2 enquêtes de consommation et $\widetilde{\mathcal{Q}}_{L_k}^{[k]} = \sum_{l=1}^{L_k} w_l^{[k]} \delta_{q_l^{[k]}}$, pour $k = 1, \ldots, P$, correspondant aux P échantillons de contamination. Elles vérifient $\sum_{i=1}^{n_r} p_i^{(r)} = 1$ et $\sum_{l=1}^{L_k} w_l^{[k]} = 1$. On définit les lois jointes $\widetilde{\mathcal{D}}_r = \prod_{k=1}^{P} \widetilde{\mathcal{Q}}_{L_k}^{[k]} \otimes \widetilde{\mathcal{C}}_{n_r}^{(r)}$.

Le paramètre d'intérêt, l'indice de risque, est définit pour chacune des enquêtes de consommation sous la forme d'une contrainte

$$\mathbb{E}_{\widetilde{\mathcal{D}}_r} \left\{ \mathbb{1}_{\sum_{k=1}^{P} q^{[k]} c_k^{(r)} > d} - \theta_d \right\} = 0, \tag{1.3}$$

pour $r = 1, 2$. Ces contraintes font intervenir des produits $p_i^{(r)} w_l^{[k]}$. On simplifie ces contraintes à l'aide de résultats sur les U-statistiques généralisées, en les linéarisant par décomposition de Hoeffding (Bertail & Tressou, 2004). On pose alors pour $c = (c_1, \cdots, c_P)'$ fixé,

$$U_0(c) = \frac{1}{\prod_{k=1}^{P} L_k} \sum_{\substack{1 \le l_k \le L_k \\ 1 \le k \le P}} \mathbb{1}_{\sum_{k=1}^{P} q_{l_k}^{[k]} c_k > d} - \theta_d, \tag{1.4}$$

et, pour $m = 1 \cdots P$ et $r = 1, 2$, à q_m fixé,

$$U_m^{(r)}(q_m) = \frac{1}{n_r \times \prod_{k \neq m} L_k} \sum_{i=1}^{n_r} \sum_{\substack{1 \le l_k \le L_k \\ k \neq m}} \mathbb{1}_{q_m c_{i,m}^{(r)} + \sum_{k \neq m} q_{l_k}^{[k]} c_{i,k}^{(r)} > d} - \theta_d. \tag{1.5}$$

On obtient alors une forme approchée des contraintes (1.3) ne faisant plus intervenir les produits $p_i^{(r)} w_l^{[k]}$:

$$\begin{aligned} \sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{k=1}^{P} \left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_k^{(1)}\left(q_{l_k}^{[k]}\right) \right] = 0, \\ \sum_{j=1}^{n_2} p_j^{(2)} U_0\left(c_j^{(2)}\right) + \sum_{k=1}^{P} \left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_k^{(2)}\left(q_{l_k}^{[k]}\right) \right] = 0. \end{aligned} \tag{1.6}$$

Une fois les contraintes linéarisées, on peut établir la convergence asymptotique du rapport de vraisemblance empirique.

Théoreme 5.2 *Soit P échantillons de contamination indépendants $\left(q_{l_k}^{[k]}\right)_{l_k=1}^{L_k}$ i.i.d., où k varie de 1 à P et deux échantillons de consommations, indépendants et P-dimensionnels, $\left(c_i^{(1)}\right)_{i=1}^{n_1}$ i.i.d. et $\left(c_j^{(2)}\right)_{j=1}^{n_2}$ i.i.d. d'indice de risque commun $\theta_d^{(1)} = \theta_d^{(2)} = \theta_{d,0} \in [0; 1]$. Supposons que, pour $r = 1, 2$, la variance de $U_0\left(c_i^{(r)}\right)$ est finie et que pour $k = 1, ..., P$, la matrice de variance-covariance de $\left(U_k^{(1)}\left(q_{l_k}^{[k]}\right), U_k^{(2)}\left(q_{l_k}^{[k]}\right)\right)'$ est finie et inversible. Si de*

plus n_1, n_2 et $(L_k)_{1\leq k\leq P}$ tendent vers l'infini et que leurs rapports restent bornés alors le programme de vraisemblance empirique s'écrit sous sa forme duale :

$$l_{n_1,n_2,L_1,\cdots,L_P}(\theta_d) = \sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\cdots,\gamma_{P+2}\in\mathbb{R} \\ n_1+n_2+\sum_{k=1}^{P} L_k=\sum_{\kappa=1}^{P+2}\gamma_\kappa}} \left\{ \begin{array}{l} \displaystyle\sum_{i=1}^{n_1} \log\left\{\gamma_1+\lambda_1 U_0\left(c_i^{(1)}\right)\right\} + \sum_{j=1}^{n_2} \log\left\{\gamma_2+\lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ \displaystyle + \sum_{k=1}^{P}\sum_{l_k=1}^{L_k} \log\left\{\gamma_{2+k}+\lambda_1 U_k^{(1)}\left(q_{l_k}^{[k]}\right)+\lambda_2 U_k^{(2)}\left(q_{l_k}^{[k]}\right)\right\} \end{array} \right\}. \quad (1.7)$$

L'estimateur associé est $\widehat{\theta}_d = \arg\sup_{\theta_d} l_{n_1,n_2,L_1,\cdots,L_P}(\theta_d)$. De plus, si l'on pose

$$r_{n_1,n_2,L_1,\cdots,L_P}(\theta_d) = 2l_{n_1,n_2,L_1,\cdots,L_P}\left(\widehat{\theta}_d\right) - 2l_{n_1,n_2,L_1,\cdots,L_P}(\theta_d),$$

alors

$$r_{n_1,n_2,L_1,\cdots,L_P}(\theta_{d,0}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi_1^2.$$

Ce théorème permet de construire un intervalle de confiance asymptotique de niveau α pour $\theta_{d,0}$:

$$\left\{\theta_d \,\middle|\, r_{n_1,n_2,L}(\theta_d) \leq F_{\chi_1^2}(1-\alpha)\right\}.$$

Vraisemblance euclidienne

Comme dans le reste de cette thèse, on peut rechercher des alternatives à la divergence de Kullback et à la vraisemblance empirique. Comme il s'agit ici de plusieurs grands jeux de données, le temps de calcul est un aspect important et l'on se tourne tout naturellement vers la divergence du χ^2. On remplace alors le programme de la vraisemblance empirique

$$\sup_{\left\{p_i^{(1)},\, p_j^{(2)},\, w_{l_k}^{[k]},k=1,..,P\right\}} \left(\sum_{i=1}^{n_1}\log p_i^{(1)} + \sum_{j=1}^{n_2}\log p_j^{(2)} + \sum_{k=1}^{P}\sum_{l_k=1}^{L_k}\log w_{l_k}^{[k]}\right)$$

par l'équivalent pour la divergence du χ^2 :

$$\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) = \min_{\left\{p_i^{(1)},\, p_j^{(2)},\, w_{l_k}^{[k]},k=1,..,P\right\}} \frac{1}{2}\left[\sum_{i=1}^{n_1}\left(n_1 p_i^{(1)}-1\right)^2 + \sum_{j=1}^{n_2}\left(n_2 p_j^{(2)}-1\right)^2 + \sum_{k=1}^{P}\sum_{l_k=1}^{L_k}\left(L_k w_{l_k}^{[k]}-1\right)^2\right], \quad (1.8)$$

sous les contraintes (1.6) et les contraintes sur la somme de chaque jeux de poids.

Nous établissons le théorème suivant :

Théoreme 5.4 *Sous les hypothèses du Théorème 5.2, la statistique pivotale*

$$\mathbf{r}_{n_1,n_2,L_1,\ldots,L_P}(\theta_{d,0}) = 2\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta_{d,0}) - 2\inf_{\theta}\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta)$$

est asymptotiquement χ_1^2 :

$$\mathbf{r}_{n_1,n_2,L_1,\ldots,L_P}(\theta_{d,0}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi_1^2.$$

Estimation de l'indice de risque pour le méthylmercure

On s'intéresse au risque d'exposition au méthylmercure dont la dose tolérable quotidienne est fixée à $1,6\mu$g. On dispose de 2 enquêtes de consommation. L'enquête INCA sur $n_1 = 3003$ individus pendant une semaine. Le panel est composé de 1985 adultes de plus de 15 ans et de 1018 enfants de 3 à 14 ans. Les enfants étant sur-représentés (34% au lieu de 15% pour le recensement) nous ajoutons au modèle une contrainte :

$$\mathbb{E}_{\widetilde{\mathcal{C}}_{n_1}^{(1)}}\left[\mathbb{1}_{3\leq Z_i^{(1)}\leq 14}\right] = 0.15,$$

où $Z_i^{(1)}$ est l'âge de l'individu i de l'enquête $r = 1$ (INCA).

La seconde enquête sur la consommation, SECODIP, est composée de 3211 ménages enquêtés sur un an. En divisant les ménages pour obtenir des consommations individuelles, on obtient $n_2 = 9588$ consommations individuelles. Le contaminant étudié, le méthylmercure, est présent dans 2 produits : le poisson et les fruits de mer. On dispose de $L_1 = 1541$ observation pour le niveau de contamination pour le poisson et de $L_2 = 1291$ observations pour les fruits de mer. Même linéarisées, les contraintes restent très lourdes : $U_1^{(2)}(q^{[2]})$ est une somme de $n_2 \times L_1 = 9588 \times 1541$ termes. On est donc obligé d'avoir recours à des U-statistiques incomplètes (voir chapitre 5). On obtient finalement comme intervalle de confiance pour l'indice de risque d'exposition au méthylmercure $[5.20\%; 5.64\%]$ et comme estimateur $\hat{\theta}_d = 5.43\%$. Ce résultat appelle une prise en compte de ce risque.

1.5.2 Économétrie : estimation sous contrainte de moment conditionnel

Au chapitre 6, suite à une étude socio-économique des déterminants de l'obésité, nous proposerons un modèle pour l'Indice de Masse Corporelle (IMC) idéal et nous chercherons à estimer en particulier le rôle de la norme sociale. On dispose des données de l'*Enquête sur les Conditions de Vie des Ménages*. Cette enquête porte sur près de 10 000 personnes réparties en 5000 ménages. Le questionnaire de l'année que nous utiliserons (2001) recueille des informations sur le poids et la taille des individus, ainsi qu'une mesure du poids idéal. Les données nécessaires à notre étude sont renseignées pour près de 4000 individus. Pour chaque individu, on obtient alors l'IMC réel et l'IMC idéal en divisant le poids réel et le poids idéal par la taille au carrée.

Pour étudier le rôle de la norme sociale, nous regroupons les individus en groupes sociaux et nous proposons un modèle explicatif pour l'IMC idéal W^* :

$$W^* = \alpha DW^* + \beta W + \delta H + \eta$$

où DW^* est la moyenne des IMC idéaux des autres individus du groupe, W est l'IMC réel et H regroupe les variables socio-démographiques ne servant pas à définir le groupe (revenu, lieu de résidence, état civil, ...). Le résidu η est supposé d'espérance nulle conditionnellement aux instruments, c'est-à-dire conditionnellement aux variables contenues dans H et à l'éducation Ed :

$$\mathbb{E}\left[W^* - \alpha DW^* + \beta W + \delta H \,\middle|\, H, Ed\right] = 0 \tag{1.9}$$

On est donc amené à utiliser la méthode de vraisemblance empirique valable dans le cadre d'une équation d'estimation conditionnelle introduite par Kitamura et al. (2004) et rappelée au paragraphe 1.1.3. Cette méthode est en fait mal adaptée à des jeux de données de taille importante, puisqu'elle fait intervenir n optimisations.

Un changement de divergence peut permettre de résoudre ce problème de temps de calcul. En effet, la méthode de Kitamura et al. (2004), reste valable pour d'autres divergences. En reprenant les notations du paragraphe 1.1.3, on pose :

$$\beta_{i,n}(\theta) = 2n \inf_{\mathbb{Q}_i \ll \mathbb{W}_i} \left\{ I_{\varphi^*}(\mathbb{Q}_i, \mathbb{W}_i) \middle| \mathbb{E}_{\mathbb{Q}_i}[m(X,\theta)] = 0 \right\},$$

et

$$\beta_n(\theta) = \sum_{i=1}^{n} \beta_{i,n}(\theta).$$

La convergence de $\beta_n(\theta_0)$ vers un χ^2 reste encore à démontrer dans ce cas général, mais a été démontrée par Smith (2005) pour les divergences I_{φ^*} de la famille des Cressie-Read. Bonnal & Renault (2004) ont étudié plus spécifiquement le cas de la divergence du χ^2, correspondant à la fonction $\varphi_2(x) = \varphi_2^*(x) = \frac{x^2}{2}$. L'avantage de cette divergence est que le multiplicateur de Lagrange λ_i est donné sous forme explicite. La valeur de $\beta_{i,n}(\theta)$ est par conséquent calculable sans procédure d'optimisation, coûteuse en temps de calcul, en tout point θ. On a en effet pour chaque programme d'optimisation, $\lambda_i = S_{i,w}^{-2}(\theta) g_{i,w}(\theta)$ avec

$$g_{i,w}(\theta) = \mathbb{E}_{\mathbb{W}_i} m(z_j, \theta) = \sum_{j=1}^{n} w_{ij} m(z_j, \theta),$$

$$S_{i,w}^2(\theta) = \mathbb{E}_{\mathbb{W}_i} m(z_j, \theta) m(z_j, \theta)' = \sum_{j=1}^{n} w_{ij} m(z_j, \theta) m(z_j, \theta)'$$

et donc, d'après le Théorème 3.1, l'optimum est atteint en $\mathbb{Q}_i^\diamond = \sum_{j=1}^n q_{ij} \delta_{X_j}$ avec

$$q_{ij}^\diamond = w_{ij} \left(1 + g_{i,w}'(\theta) S_{i,w}^{-2}(\theta) m(X_j, \theta)\right)$$

et

$$\beta_{i,n}(\theta) = 2n I_{\varphi_2^*}(\mathbb{Q}_i^\diamond, \mathbb{W}_i) = n g_{i,w}'(\theta) S_{i,w}^{-2}(\theta) g_{i,w}(\theta).$$

En utilisant cette divergence pour estimer les paramètres de notre modèle (1.9), on trouve alors que l'effet de la norme sociale est bien supérieur à celui de l'IMC réel : $\alpha = 0.90$ alors que $\beta = 0.19$. Ces deux coefficients sont significatifs à 99%. Le chapitre 6 présente les motivations économétriques du modèle étudié et discute ces résultats, en particulier en les comparant à ce qui serait obtenu avec les méthodes usuelles (moindres carrés, méthode des moments généralisée).

Première partie

Etude théorique : généralisations de la méthode de vraisemblance empirique

Chapitre 2

Divergence empirique et vraisemblance empirique généralisée

2.1 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 2004) et de manière générale est asymptotiquement valide pour tout paramètre multidimensionnel Hadamard différentiable (Bertail, 2004, 2006). 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 divergence de 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 la distance du χ^2 et les divergences de type Cressie-Read (Baggerly 1998, Corcoran 1998, Bonnal & Renault 2001, Newey & Smith 2004, Bertail 2006) 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 divergence de 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, d'un point de vue théorique ou algorithmique ?
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 φ-divergence, pour toute fonction φ^* 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, 2004, voir la remarque 2.1 ci-dessous) et vont dans le sens des travaux obtenus indépendamment par Bronia-

towski & Kéziou (2003) pour des problèmes de tests paramétriques ou semi-paramétriques. Nous montrons en particulier que les résultats obtenus sur les vraisemblances empiriques généralisées sont fortement liés, sous certaines conditions sur les fonctions φ^* 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 du point de vue de la théorie asymptotique, en nous appuyant tout particulièrement sur les travaux de Mykland (1994), Baggerly (1998), Corcoran (1998) et Bertail (2004). 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 divergences de type Cressie-Read, seule la vraisemblance empirique possède cette propriété mais que d'autres φ-divergences la possèdent également. Nous introduisons en particulier une famille de φ-divergences, barycentres de la distance de Kullback et du χ^2, qui sont Bartlett corrigeables (voir page 58). Une comparaison fine de ces φ-divergences nécessitent une analyse à l'ordre 5 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 simulations des zones de confiance (multidimensionnelles, $p = 2$) obtenues pour différentes divergences.

2.2 φ-Divergences et dualité convexe

Afin de généraliser la méthode de vraisemblance empirique, on rappelle quelques notions sur les φ-divergences (Csiszár, 1967), dont nous donnerons quelques exemples (voir également Rockafeller, 1970, ou Broniatowski & Kéziou, 2003). Nous rappelons en annexe 2.6 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.

2.2.1 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 $\mu f = \mathbb{E}_\mu[f] = \int f(x)\mu(dx)$.

On utilise dans toute la suite la notation φ pour des fonctions convexes. On note

$$d(\varphi) = \{x \in \mathbb{R},\ \varphi(x) < \infty\}$$

le domaine de φ et respectivement $\inf d(\varphi)$ et $\sup d(\varphi)$ les points terminaux de ce domaine. Pour toute fonction φ convexe, on introduit sa conjuguée convexe φ^* ou transformée de Fenchel-Legendre

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

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

Hypothèses 2.1

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

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

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

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

On a alors les propriétés classiques

Théoreme 2.1

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

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

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

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

Soit φ vérifiant les hypothèses 2.1. La φ-divergence associée à φ, 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 :

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

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 φ-divergences réside pour nous dans le théorème suivant (réécrit sous une forme simplifiée) dû à Borwein & Lewis (1991) (voir également Léonard, 2001) qui résulte des propriétés des intégrales de fonctionnelles convexes.

Théoreme 2.2 (Minimisation et Conjugaison) *Soit φ une fonction convexe partout finie et différentiable telle que $\varphi^* \geq 0$ et $\varphi^*(0) = 0$. Soit $\mathbb{P}$ une mesure de probabilité discrète. Alors il vient*

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

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

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

alors il existe $\mathbb{Q}^\diamond$ et $\lambda^\diamond$ réalisant respectivement l' inf *et le* sup *et tels que*

$$\mathbb{Q}^\diamond = \left(1 + \varphi^{(1)}(\lambda^{\diamond\prime} f)\right) \mathbb{P}.$$

2.2.2 Exemples

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

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 2.1 donne les fonctions φ et φ^* classiques, ainsi que leur domaine.

Divergences	α	φ_α		φ^*_α	
		$\varphi_\alpha(x)$	$d(\varphi_\alpha)$	$\varphi^*_\alpha(x)$	$d(\varphi^*_\alpha)$
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	$\frac{x^2}{2-x}$	$]-\infty, 2[$	$2(\sqrt{(x+1)} - 1)^2$	$]-1, +\infty[$
χ^2	2	$\frac{x^2}{2}$	$\mathbb{R}$	$\frac{x^2}{2}$	$\mathbb{R}$

TAB. 2.1 – Les principales Cressie-Read

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

$$\varphi_\alpha^*(x) = \frac{(1+x)^\alpha - \alpha x - 1}{\alpha(\alpha-1)},\ \varphi_\alpha(x) = \frac{\left[(\alpha-1)x+1\right]^{\frac{\alpha}{\alpha-1}} - \alpha x - 1}{\alpha}$$

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

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

2.3 Extension de la méthode de vraisemblance empirique aux φ-divergences

L'objectif de ce chapitre est d'étendre la méthode de vraisemblance empirique à des φ-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_{φ^*}. 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, de la forme :

$$\mathbb{E}_P[m(X,\theta)] = 0$$

où m est une fonction régulière de $\mathcal{X} \times \mathbb{R}^p$ dans $\mathbb{R}^r$ avec $r \geq p$. Nos résultats ne couvrent pas directement le cas de contraintes de moments conditionnels ni le cas d'un paramètre défini par une infinité de moments (comme c'est souvent le cas dans les modèles semiparamétriques). Pour des résultats dans cette direction pour des divergences particulières, on se référera à Bonnal & Renault (2001) et Kitamura (2004).

2.3.1 Vraisemblance empirique

On considère une suite de vecteurs aléatoires $X, X_1, \cdots, X_n$ de $\mathbb{R}^p$, $n \geq 1$, indépendants et identiquement 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 $\mu_0 = \mathbb{E}_P X = \int X dP$ 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 de probabilités, la fonctionnelle $L(\mathbb{Q}) = \prod_{i=1}^n \mathbb{Q}(\{x_i\})$ où $\forall i \in \{1, ..., n\}$, $\{x_i\}$ représente le singleton $x_i = X_i(\omega)$, pour $\omega \in \Omega$ fixé. 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 pour la moyenne μ_0, selon le principe de la vraisemblance empirique, comme suit

$$\begin{aligned} C_{\eta,n} &= \left\{ \mu \left| \mathbb{E}_{\mathbb{Q}}[X-\mu]=0, \mathbb{Q}\in\mathcal{P}_n, R_n(\mathbb{Q}) = \frac{L(\mathbb{Q})}{L(\mathbb{P}_n)} \leq \eta \right. \right\} \\ &= \left\{ \mu \left| \sum_{i=1}^n q_i(X_i-\mu)=0, q_i\geq 0, \sum_{i=1}^n q_i = 1,\ R_n(\mathbb{Q}) = \frac{\prod_{i=1}^n q_i}{\prod_{i=1}^n \frac{1}{n}} \leq \eta \right. \right\}, \end{aligned}$$

où η est déterminé par la confiance $1-\alpha$ que l'on veut atteindre : $\Pr(\mu_0 \in C_{\eta,n}) = 1-\alpha$. L'intérêt de la définition de $C_{\eta,n}$ vient de l'observation suivante de Owen (1988) :

$$\forall \mu \in \mathbb{R}^p, \Pr(\mu \in C_{\eta,n}) = \Pr(\eta_n(\mu) \geq \eta)$$

$$\text{avec } \eta_n(\mu) = \sup_{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} R_n(\mathbb{Q}) = \frac{\sup_{\{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0\}} L(\mathbb{Q})}{\sup_{\mathbb{Q}\in\mathcal{P}_n} L(\mathbb{Q})},$$

qui s'interprète clairement comme un rapport de vraisemblance. Un estimateur de μ_0 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

$$\Pr(\mu_0 \in C_{\eta,n}) = \Pr(\beta_n(\mu_0) \leq -\log(\eta)).$$

On en déduit que pour $\eta = \exp(-\frac{\chi^2_{1-\alpha}}{2})$, $C_{\eta,n}$ est asymptotiquement de niveau $1-\alpha$.

La statistique pivotale de la vraisemblance empirique, $\beta_n(\mu)$, peut s'interpréter directement comme la minimisation d'une divergence de Kullback, sous certaines contraintes empiriques sur les moments . En effet, on a

$$\begin{aligned} \beta_n(\mu) &= -\log(\eta_n(\mu)) = \inf_{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} -\log R_n(\mathbb{Q}) \\ &= \inf_{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} -\sum_{i=1}^n \log(nq_i) = n \inf_{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} \int -\log\left(\frac{d\mathbb{Q}}{d\mathbb{P}_n}\right) d\mathbb{P}_n. \end{aligned}$$

En utilisant $\int \left(\frac{d\mathbb{Q}}{d\mathbb{P}_n} - 1\right) d\mathbb{P}_n = 0$, on obtient

$$\begin{aligned} \beta_n(\mu) &= n \inf_{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} \int \left[\left(\frac{d\mathbb{Q}}{d\mathbb{P}_n} - 1\right) - \log\left(1 + \frac{d\mathbb{Q}}{d\mathbb{P}_n} - 1\right)\right] d\mathbb{P}_n \\ &= n \inf_{\mathbb{Q}\in\mathcal{P}_n, \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} K(\mathbb{Q}, \mathbb{P}_n). \end{aligned}$$

Cette présentation suggère la généralisation suivante.

2.3.2 Minimisation empirique des φ-divergences

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

$$\beta_n^\varphi(\mu) = n \inf_{\{\mathbb{Q}<<P_n, \mathbb{E}_\mathbb{Q}[X]=\mu\}} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n)\}$$
$$C_{\eta,n,\varphi^*} = \{\mu |\ \exists \mathbb{Q},\ \mathbb{E}_\mathbb{Q}[X-\mu] = 0,\ \mathbb{Q} \ll \mathbb{P}_n \text{ et } nI_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \leq \eta\}.$$

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 2.2, 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 prendre des précautions avec la contrainte sur le paramètre recherché. En effet, en imposant $\mathbb{E}_\mathbb{Q}[X-\mu] = 0$, on définit μ comme une moyenne renormalisée : $\mu = \frac{\mathbb{E}_\mathbb{Q}[X]}{\mathbb{Q}(1)}$.

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

$$M(R, \mu) = \inf_{\{\mathbb{Q} \ll R,\ \mathbb{E}_\mathbb{Q}[X-\mu]=0\}} \{I_{\varphi^*}(\mathbb{Q}, R)\}$$

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. $\mathbb{E}_P[X-\mu] = 0$ 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 $\beta_n^\varphi(\mu)/n$. 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 φ satisfait les hypothèses suivantes :

Hypothèses 2.2 *(i) φ vérifie les hypothèses 2.1,*

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

Il est simple de vérifier que les fonctions et divergences données dans la partie précédente vérifient cette hypothèse supplémentaire. L'hypothèse (ii) est vérifiée en particulier lorsque $\varphi^{(1)}$ est elle-même convexe (entraînant $\varphi^{(2)}(x)$ croissante donc $\geq \varphi^{(2)}(0) > 0$ 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 $\mathcal{M}_n$, on peut réécrire les contraintes de minimisation sous la forme

$$\mathbb{E}_{(\mathbb{Q}-\mathbb{P}_n)}[X-\mu] = \mu - \bar{\mu},\ \text{où } \bar{\mu} = \mathbb{E}_{\mathbb{P}_n}[X].$$

Il vient

$$\begin{aligned}
\beta_n^\varphi(\mu) &= n \inf_{\mathbb{Q} \in \mathcal{M}_n,\ \mathbb{E}_\mathbb{Q}[X-\mu]=0} \{I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n)\} \\
&= n \inf_{\mathbb{Q} \in \mathcal{M}_n} \left\{ I_{\varphi^*}(\mathbb{Q}, \mathbb{P}_n) \,\middle|\, E_{(\mathbb{Q}-\mathbb{P}_n)}[X-\mu] = \mu - \bar{\mu} \right\} \\
&= n \sup_{\lambda \in \mathbb{R}^p} \left\{ \lambda'(\mu - \bar{\mu}) - \int_\Omega \varphi(\lambda'(X-\mu)) d\mathbb{P}_n \right\}.
\end{aligned}$$

On en déduit l'expression duale de $\beta_n^\varphi(\mu)$ qui permet de généraliser les propriétés usuelles de la vraisemblance empirique à notre cadre plus large :

$$\beta_n^\varphi(\mu) = \sup_{\lambda\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \lambda'(\mu - X_i) - \sum_{i=1}^n \varphi(\lambda'(X_i - \mu)) \right\}. \tag{2.1}$$

Note 2.1 *L'égalité (2.1) invalide une conjecture formulée dans une version préliminaire de Newey & Smith (2004), qui stipule qu'une telle relation de dualité n'est valable et explicite que pour la famille des Cressie-Read. On obtient ici que l'opération consistant à minimiser toute φ-divergence équivaut à la recherche d'un pseudo maximum de vraisemblance (Generalized Empirical Likelihood, GEL, dans la terminologie de Newey & Smith). On introduit ci-dessous, dans le paragraphe 2.3.6, une famille de φ-divergences qui ne sont pas des Cressie-Read, pour lesquelles le programme (2.1) est équivalent à un GEL, avec une forme explicite pour φ.*

L'écriture (2.1) permet d'établir le

Théoreme 2.3 *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 φ vérifie les hypothèses 2.2 alors,*

$$2\varphi^{(2)}(0)\beta_n^\varphi(\mu_0) \xrightarrow[n\to\infty]{\text{en loi}} \chi^2(q)$$

et $\forall\ 0 < \alpha < 1$, et pour $\eta = \frac{\chi^2_{1-\alpha}(q)}{2\varphi^{(2)}(0)}$, C_{η,n,φ^*} est convexe et

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

Note 2.2 *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 $\varphi(x) = \frac{x^2}{2}$, la contrainte supplémentaire conduit au problème de minimisation*

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

Le calcul du Lagrangien correspondant montre facilement qu'il n'existe de solution qu'en $\mu = \frac{1}{n}\sum_{i=1}^n X_i = \bar{\mu}$, la « vraisemblance » vaut alors $+\infty$ partout ailleurs et les régions de confiance dégénèrent.

2.3.3 Vraisemblance empirique : la divergence de Kullback

Dans le cas particulier, où $\varphi^*(x) = x - \log(1 + x)$, on obtient automatiquement une probabilité : la présence du log dans l'expression de φ^* entraîne $q_i \geq 0$, et la contrainte $\sum_{i=1}^n q_i(X_i - \mu) = 0$ entraîne étonnement que $\sum_{i=1}^n q_i = 1$.

Par ailleurs, la forme duale donnée par l'équation (2.1) peut se réécrire

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

Comme le remarque Bertail (2004, 2006), 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.

2.3.4 Les Cressie-Read

Les résultats établis pour les Cressie-Read (voir Newey & Smith, 2004), dont on rappelle ici la forme $\varphi^*_\alpha(x) = \frac{(1+x)^\alpha - \alpha x - 1}{\alpha(\alpha-1)}$, s'obtiennent facilement en appliquant le Théorème 2.3. Les hypothèses du Théorème 2.2 sont vérifiées pour calculer la valeur de $\beta_n^{\varphi_\alpha}$ 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

$$\sum_{i=1}^{n} q_i (X_i - \mu) = 0.$$

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. On peut alors appliquer le Théorème 2.2 à φ_α et l'on obtient

$$\inf_{\mathbb{Q} \in \mathcal{M}_n,\ \mathbb{E}_{\mathbb{Q}}[X-\mu]=0} \{I_{\varphi^*_\alpha}(\mathbb{Q}, \mathbb{P}_n)\} = \frac{1}{n} \sup_{\lambda \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \lambda'(\mu - X_i) - \sum_{i=1}^{n} \varphi_\alpha(\lambda'(X_i - \mu)) \right\}$$

c'est-à-dire, en explicitant $I_{\varphi^*_\alpha}$ et φ_α,

$$\frac{1}{\alpha(\alpha-1)} \int_\Omega \left[\left(\frac{d\mathbb{Q}^\diamond}{d\mathbb{P}} \right)^\alpha - \alpha \left(\frac{d\mathbb{Q}^\diamond}{d\mathbb{P}} - 1 \right) - 1 \right] d\mathbb{P}$$
$$= \frac{-1}{n\alpha} \sum_{i=1}^{n} \left\{ \left[1 + (\alpha - 1)\lambda^{\diamond\,\prime}(X_i - \mu) \right]^{\frac{\alpha}{\alpha-1}} - 1 \right\}.$$

$\mathbb{Q}^\diamond$ est donné par

$$\mathbb{Q}^\diamond = \sum_{i=1}^{n} q_i^\diamond \delta_{X_i} = \sum_{i=1}^{n} \frac{1}{n} \left[1 + (\alpha - 1)\lambda^{\diamond\,\prime}(X_i - \mu) \right]^{\frac{1}{\alpha-1}} \delta_{X_i},$$

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

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

On regroupe les valeurs des poids et des divergences pour les exemples usuels ($\alpha = 0, 1/2, 1, 2$) dans le tableau 2.2.

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

TAB. 2.2 – Poids et rapports de vraisemblances pour les divergences usuelles

Comme mentionné par un rapporteur, l'entropie relative (φ_1) conduit en économétrie au critère KLIC (Kullback-Leibler Information Criterion) qui est différent du maximum de vraisemblance empirique. Il est été étudié sous ce nom par Kitamura & Stutzer (1997).

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 $\lambda_n = S_n^{-1}(\mu)(\mu - \bar{\mu})$ avec $S_n(\mu) = \frac{1}{n}\sum_{i=1}^n (X_i - \mu).(X_i - \mu)'$ d'où $q_i^\diamond = \frac{1}{n}(1 + (\mu - \bar{\mu})' S_n^{-1}(\mu)(X_i - \mu))$ et

$$I_{\varphi_2^*}(\mathbb{Q}^\diamond, \mathbb{P}_n) = \sum_{i=1}^n \frac{(n \cdot q_i^\diamond - 1)^2}{2n} = \frac{1}{2}(\mu - \bar{\mu})' S_n^{-1}(\mu)(\mu - \bar{\mu}),$$

qui s'interprète directement comme une somme vectorielle autonormalisée. Pour l'étude de l'estimateur basé sur le χ^2, très lié aux GMM (Generalized Method of Moments), on se

référera à Bonnal & Renault (2004) ainsi qu'à Newey & Smith (2004). L'essentiel de la différence entre les GMM et la φ-divergence du χ^2 tient au fait qu'ici, la matrice $S_n(\mu)$ est recalculée à chaque valeur du paramètre μ, alors que pour les GMM on estime dans un premier temps $S_n(\mu_0)$, par exemple par $S_n(\bar{\mu})$. On obtient alors

$$I_{\varphi_\alpha^*}(\mathbb{Q}^\diamond, \mathbb{P}_n) = \frac{1}{2}(\mu - \bar{\mu})' S_n^{-1}(\mu)(\mu - \bar{\mu})$$
$$GMM(\mu) = \frac{1}{2}(\mu - \bar{\mu})' S_n^{-1}(\bar{\mu})(\mu - \bar{\mu}).$$

Dans le cas de la moyenne et en dimension 1, on a $S_n(\bar{\mu}) + (\mu - \bar{\mu})^2 = S_n(\mu)$ et l'on obtient alors :

$$2I_{\varphi_\alpha^*}(\mathbb{Q}^\diamond, \mathbb{P}_n) = \frac{(\mu - \bar{\mu})^2}{S_n(\mu)} = \frac{(\mu - \bar{\mu})^2}{S_n(\bar{\mu}) + (\mu - \bar{\mu})^2}$$
$$2I_{\varphi_\alpha^*}(\mathbb{Q}^\diamond, \mathbb{P}_n) = \frac{(\mu - \bar{\mu})^2 / S_n(\bar{\mu})}{1 + (\mu - \bar{\mu})^2 / S_n(\bar{\mu})} = \frac{2GMM(\mu)}{1 + 2GMM(\mu)}.$$

2.3.5 Les Polylogarithmes

La famille des Cressie-Read ne contient pas toutes les φ-divergences dont la forme duale peut s'écrire explicitement. Considérons par exemple la suivante 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 $]-1, 1[$, qui vérifient toutes nos hypothèses 2.2. On a en particulier pour $x \in]-1, 1[$ $h_0(x) = \dfrac{x^2}{2(1-x)}$, $h_1(x) = -\log(1-x) - x = \gamma_0(x)$ 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 $h_0^*(x) = (\sqrt{1+x} - 1)^2$. On peut également souligner que

$$h_\infty(x) = \lim_{\alpha \to +\infty} h_\alpha(x) = \frac{x^2}{2} \mathbb{1}_{x \in]-1,1[} \text{ et donc } h_\infty^*(x) = \frac{x^2}{2} \mathbb{1}_{x \in]-1,1[} + \text{sign}(x)(x - 1/2) \mathbb{1}_{x \notin]-1,1[}$$

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 concilier les avantages de la divergence de Kullback (l'aspect adaptatif des régions de confiance et la correction de Bartlett) et de la divergence du χ^2 (la robustesse et la simplicité algorithmique). Nous réservons l'étude détaillée de cette famille à des travaux ultérieurs mais nous donnons ici certaines propriétés intéressantes des Quasi-Kullback :

- Conformément aux attentes, la forme des régions s'adaptent d'autant plus aux données que ε est proche de 0, la valeur correspondant à la divergence de Kullback.
- On obtient la correction de Bartlett, pour ε petit.
- Dès que $\varepsilon > 0$, les régions de confiance peuvent être plus grandes que l'enveloppe convexe des données, ce qui augmente fortement la robustesse.
- On peut obtenir de plus un contrôle à distance fini du niveau de la région de confiance, grâce à des inégalités exponentielles explicites. Ce contrôle est d'autant plus fin que ε est proche de 1.
- Enfin, contrairement à la divergence de Kullback, la conjuguée convexe de K_ε est définie sur $\mathbb{R}$ et est relativement lisse, ce qui simplifie l'implémentation du problème d'optimisation.

Malheureusement les différentes propriétés ci-dessus influencent de façon contradictoire le choix de ε, qu'il faut donc adapter à chaque problème précis. Une procédure basée sur la validation croisée pourrait s'avérer intéressante comme le montre une étude récente par simulation.

La famille des Quasi-Kullback vérifie nos hypothèses 2.2 et l'on obtient la convergence vers un χ^2_q grâce au Théorème 2.3. On peut expliciter K^*_ε, qui est finie quelque soit $x \in \mathbb{R}$ pour $\varepsilon > 0$:

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

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

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

Théoreme 2.4 *Sous les hypothèses du Théorème 2.3, on a :*

(a) $\forall n > 0$ *fixé, c'est-à-dire* ***non asymptotiquement****,*

$$\begin{aligned}\Pr(\mu_0 \notin C_{\eta,n,K^*_\varepsilon}) &= \Pr(\beta_n^{K_\varepsilon}(\mu_0) \geq \eta) \\ &\leq \Pr\left(\frac{n}{2}(\mu_0 - \bar{\mu})' S_n^{-1}(\mu_0 - \bar{\mu}) \geq \eta\varepsilon\right)\end{aligned}$$

(b) Si, $\mu_0 \in \mathbb{R}$ *et* $X_i - \mu_0$ *est de loi symétrique, alors sans aucune hypothèse de moments, d'après l'inégalité de Hoeffding,*

$$\Pr(\mu_0 \notin C_{\eta,n,K^*_\varepsilon}) \leq 2\exp(-\eta\varepsilon).$$

(c) De manière générale, si $X_i - \mu_0$ est de loi symétrique (au sens ou $-(X_i - \mu_0)$ a même loi que $X_i - \mu_0$), d'après les inégalités de Pinélis (1994), on a le contrôle

$$\Pr(\mu_0 \notin C_{\eta,n,K_\varepsilon^*}) \leq 2e^3/9 \Pr(\chi^2(q) \geq 2\eta\varepsilon)$$

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

Note 2.3 *La première partie (a) 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). L'inégalité exponentielle (b) dans le cas symétrique est classique en dimension 1 et remonte à des travaux de Efron (1969). Les inégalités de Pinélis (1994) permettent d'améliorer ce résultat et impliquent (c) à savoir que, dans le cas symétrique, l'erreur de première espèce des régions de confiance associées aux quasi-Kullback peut toujours être contrôlée à distance finie par la loi d'un $\chi^2(q)$. La borne est optimale pour $\varepsilon = 1$ i.e. pour le choix de la distance du χ^2. Par contre en terme de correction de Bartlett (asymptotique), (d) signifie qu'un choix de ε petit s'impose. 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.*

D'un point de vue pratique, lorsque la taille de l'échantillon n est grande, on aura plutôt tendance à choisir ε petit quitte à appliquer une correction de type Bartlett, alors que si n est petit, le choix du χ^2 et un contrôle exact sont plus appropriés. Un étude par simulation en cours montre qu'il peut être intéressant de déterminer ε par des techniques de validation croisée même si dans ce cas les propriétés théoriques sont moins claires. Nous établirons au chapitre 3 des bornes valable dans le cas non-symétrique. Nous proposerons également une méthode pour déterminer la valeur optimale de ε.

2.4 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.

2.4.1 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 2.1).

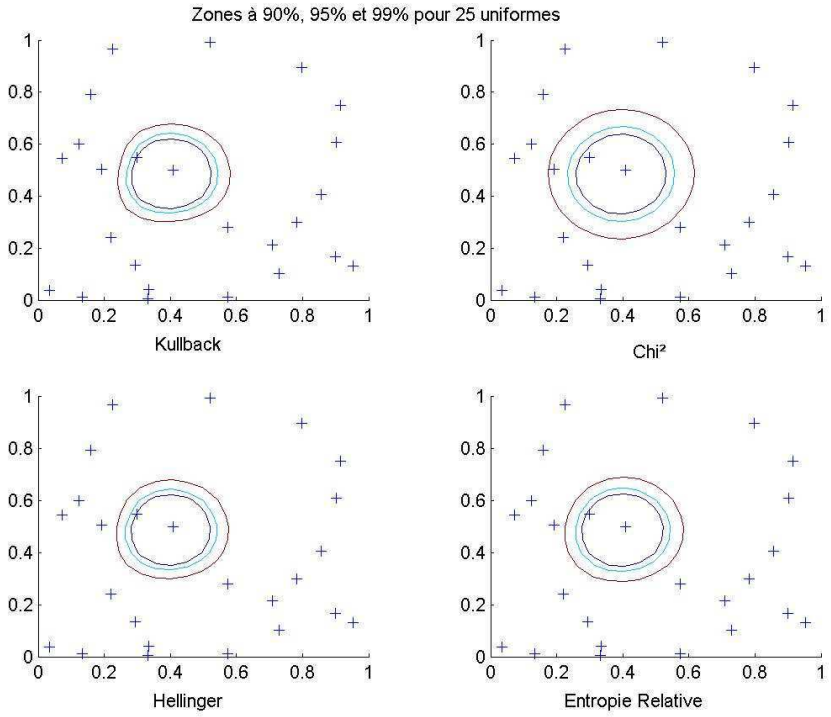


FIG. 2.1 – Zones de confiance pour 4 divergences avec les mêmes données

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 2.3. 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.

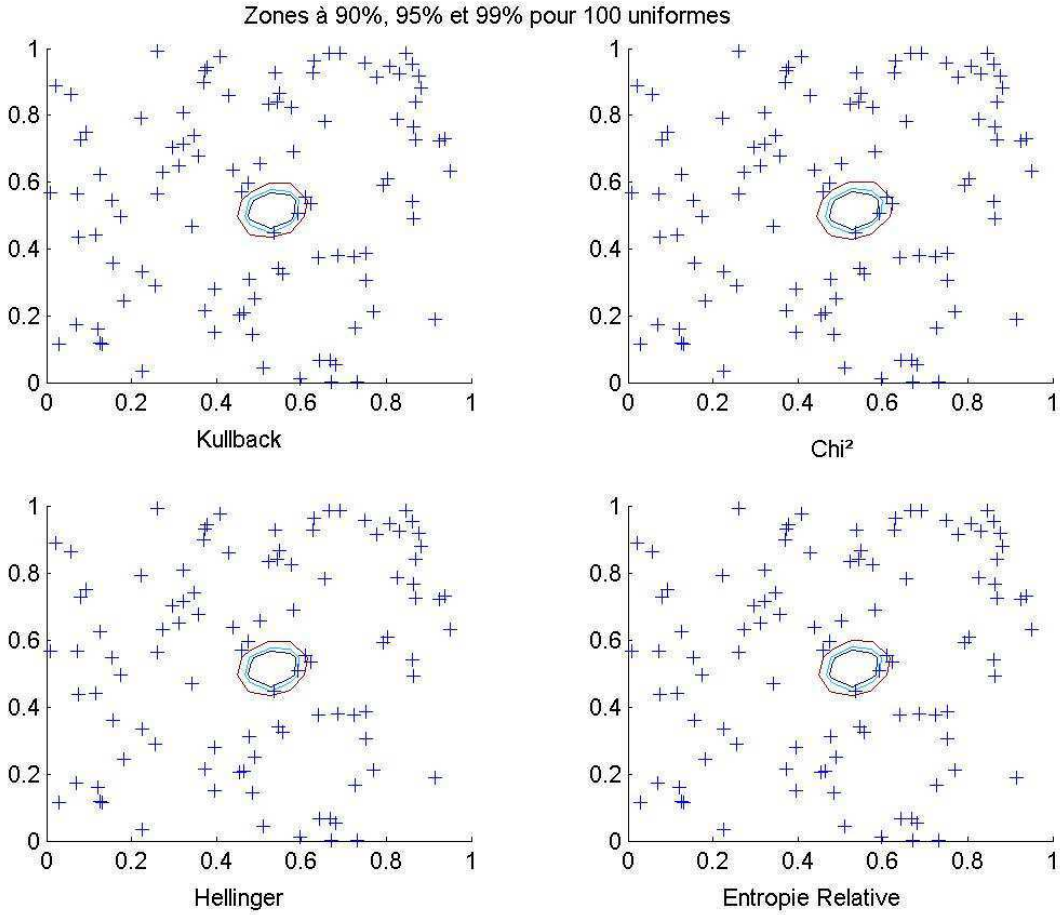


FIG. 2.2 – Zones de confiance pour 100 données uniformes

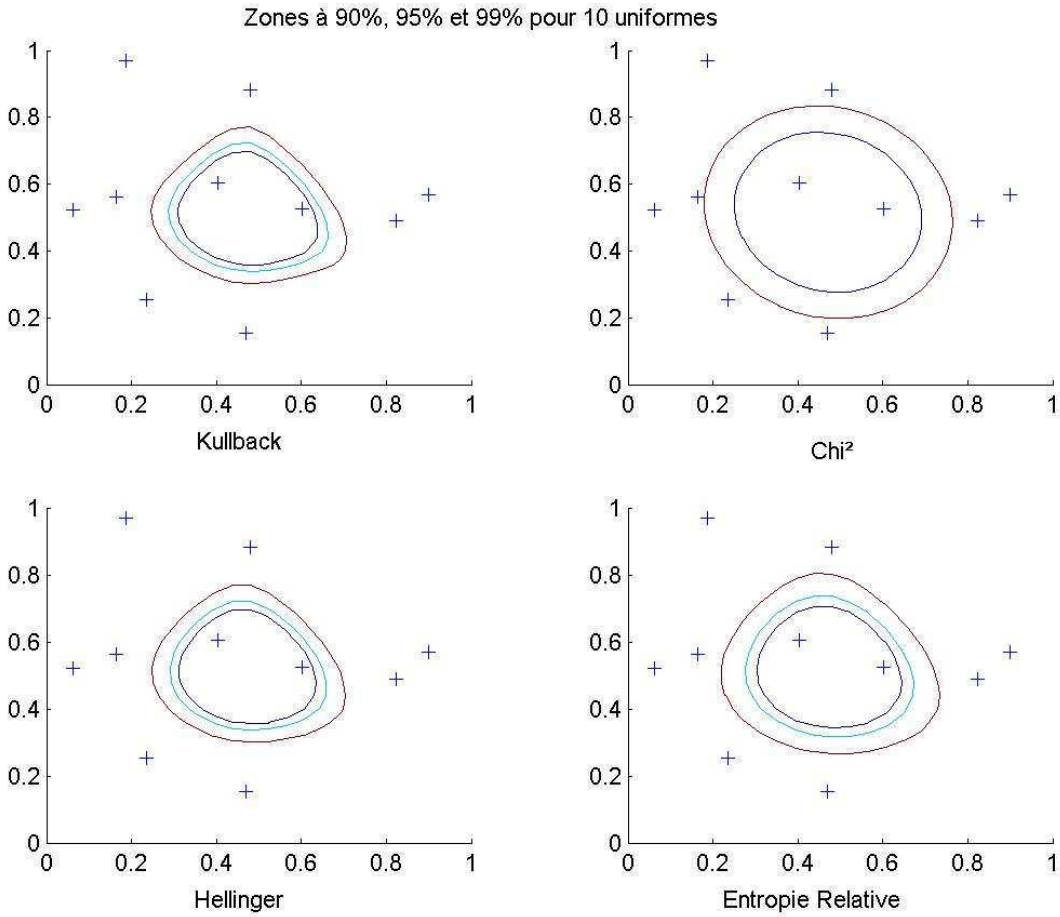


FIG. 2.3 – Zones de confiance pour 10 données uniformes

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 2.3, la ligne de niveau correspondant à 99% est hors du cadre, et est donc invisible).

2.4.2 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 2.4 les zones de confiance pour 50 données suivant une loi de Marshall-Olkin (ce copule a une dépendance entre les 2 coordonnées qui fait apparaître une courbe exponentielle) :

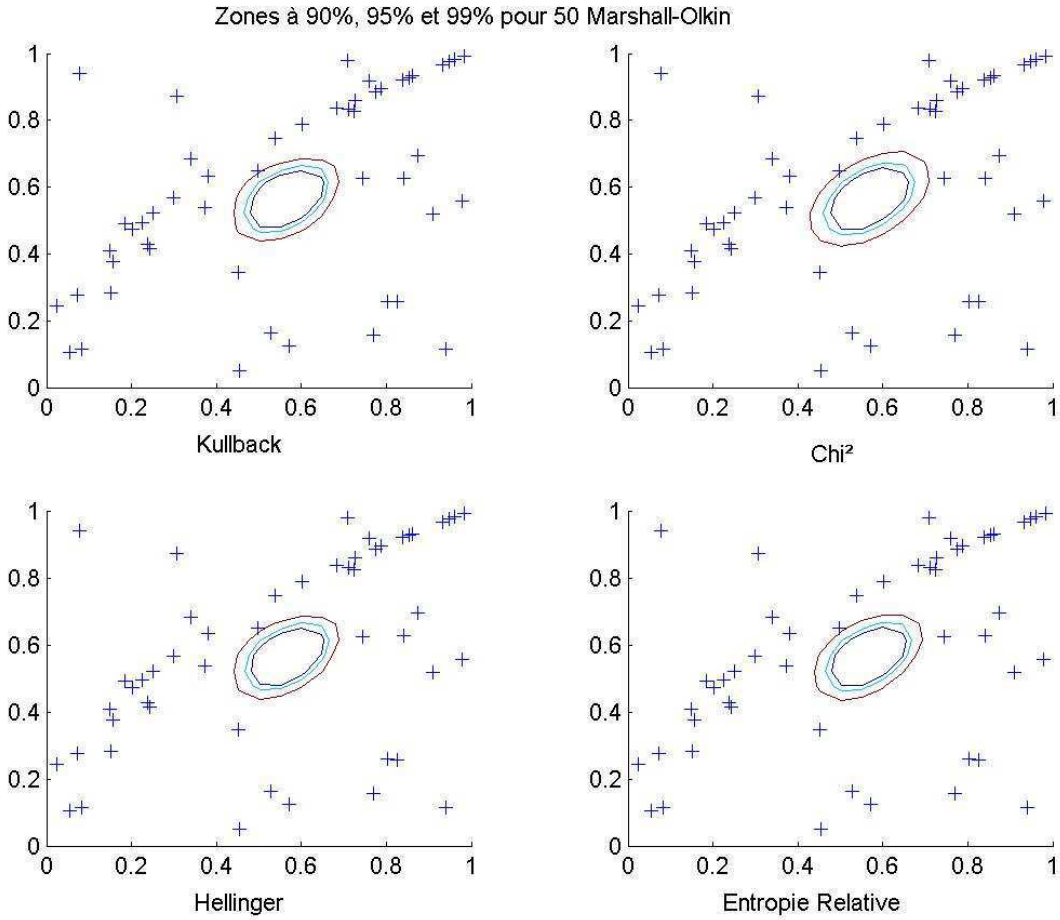


FIG. 2.4 – Zones de confiance pour un copule

Note 2.4 *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.

On a représenté dans la Figure 2.5 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.

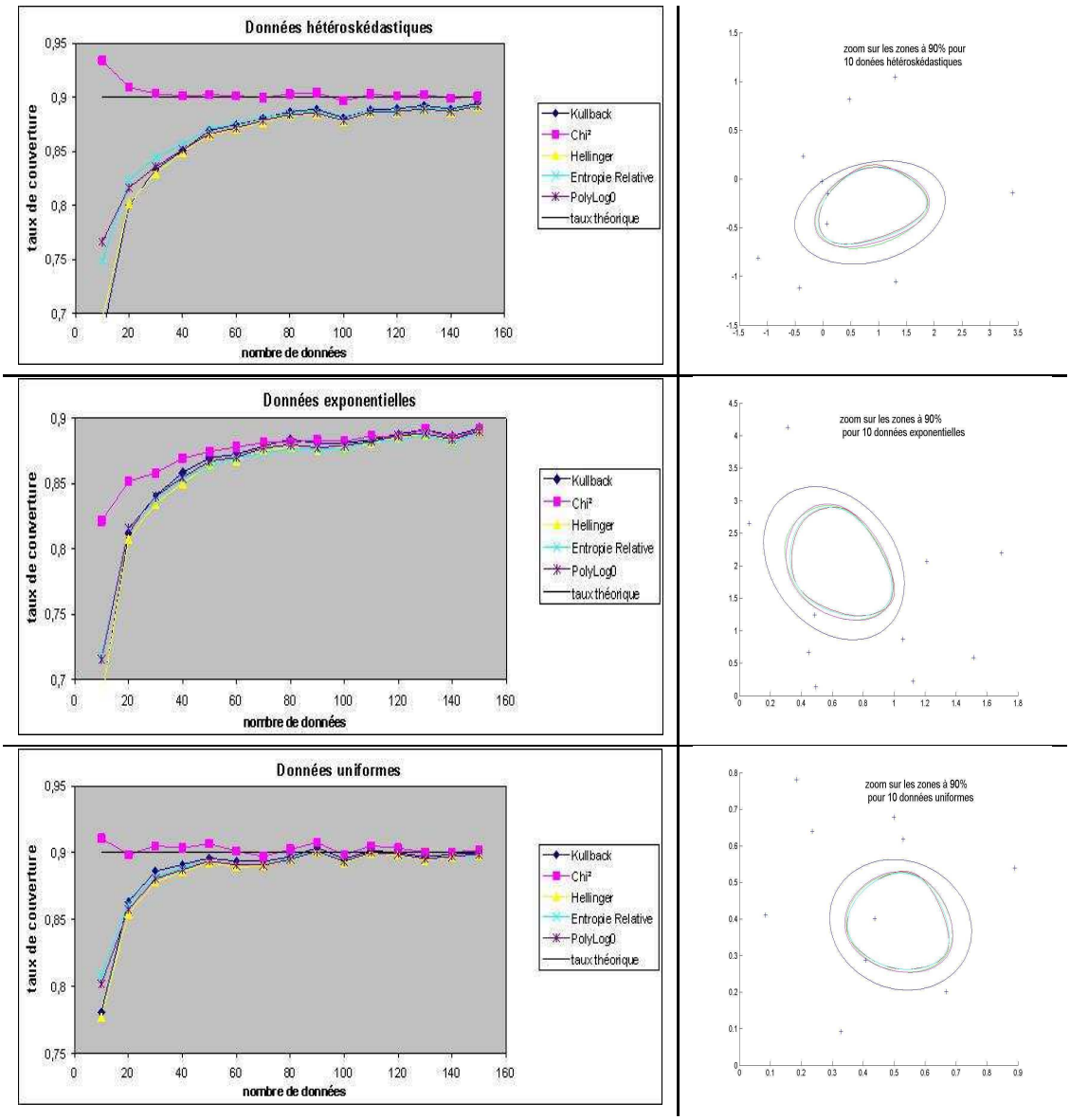


FIG. 2.5 – Évolution des taux de couverture

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 du Théorème 2.4.

2.5 Conclusion

Ce travail généralise la validité des raisonnements sur la vraisemblance empirique menés par Owen (1990) aux φ-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.

2.6 Annexe : Calcul convexe

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

Lemme 2.1 (Involutivité de la conjugaison) *Pour toute fonction φ convexe, les assertions suivantes sont équivalentes :*

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

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

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

Grâce à ces méthodes, il est assez simple d'obtenir les tableaux 2.3 et 2.4 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. 2.3 – Propriétés élémentaires

$f = g^*$		$g = f^*$	
Fonction	$d(f)$	Fonction	$d(g)$
0	$[-1,1]$	$\mid x\mid$	$\mathbb{R}$
0	$[0,1]$	x^+	$\mathbb{R}$
$\frac{\mid x\mid^p}{p}\ \forall p>1$	$\mathbb{R}$	$\frac{\mid x\mid^q}{q}$ pour $\frac{1}{p}+\frac{1}{q}=1$	$\mathbb{R}$
$-\frac{\mid x\mid^p}{p}\ \forall p\in]0,1[$	$\mathbb{R}_+$	$-\frac{(-x)^q}{q}$	$-\mathbb{R}_+^*$
$\sqrt{(1+x^2)}$	$\mathbb{R}$	$-\sqrt{(1-x^2)}$	$[-1,1]$
$-\log(x)$	$\mathbb{R}_+^*$	$-1-\log(-y)$	$-\mathbb{R}_+^*$
$\log(\cos(x))$	$\left]-\frac{\pi}{2},\frac{\pi}{2}\right[$	$\frac{x}{\tan(x)}-\frac{1}{2}\log(1+x^2)$	$\mathbb{R}$
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. 2.4 – Tableau des principales conjuguées convexes

2.7 Annexe : Preuves

2.7.1 Preuve du Théorème 2.3

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 $q = p$. Ceci revient à supposer que la matrice de variance-covariance (notée Σ) des X_i est inversible. La convexité de C_{η,n,φ^*} découle immédiatement de celle de φ^* 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 2.3, il nous faut nous assurer de l'existence de $\eta_n(\mu)$ et de $\beta_n^{\varphi}(\mu)$, au moins pour certains μ. D'après Owen (2001), 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 2.2 *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 (2.2) vaut

$$\beta_n^{\varphi}(\mu) = \sup_{\lambda\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \lambda'(\mu - X_i) - \sum_{i=1}^n \varphi(\lambda'(X_i - \mu)) \right\}. \tag{2.1}$$

La condition au premier ordre impliquée par (2.1) 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

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

On pose $Y_i = X_i - \mu$, Y_i est alors centrée et de même matrice de variance-covariance que X_i, Σ. On note λ_n le λ réalisant le sup dans (2.1), on a donc $g(\lambda_n) = 0$. On pose $\lambda_n = \rho_n\theta_n$ avec $\rho_n \geq 0$ et $\|\theta_n\|_2 = 1$,

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

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

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

avec t_i entre 0 et $\theta_n'Y_i$. On a alors sous l'hypothèse (ii) de 2.2

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

Or, $\frac{1}{n}\sum_{i:\theta_n'Y_i\geq 0}^{n}(\theta_n'Y_i)^2$ 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

$$\frac{1}{n}\sum_{i:\theta_n'Y_i\geq 0}^{n}(\theta_n'Y_i)^2 \xrightarrow[n\to\infty]{} 0 \text{ et } \theta_n \xrightarrow[n\to\infty]{} \theta_0.$$

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 $\tilde{\lambda}_n = \lambda_n + \frac{S_n^{-1}(\bar{\mu}-\mu)}{\varphi^{(2)}(0)}$. En effectuant un développement de Taylor de φ en 0 dans l'expression de $g(\lambda_n)$, il vient

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

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 2.2. 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 φ autour de 0 dans l'expression de $\beta_n^{\varphi}(\mu)$, il vient

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

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

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

2.7.2 Preuve du Théorème 2.4

Inégalité exponentielle

Pour alléger les notations, on note $\beta_n^{\varepsilon} = \beta_n^{K_\varepsilon}$. D'après l'égalité (2.1) donnant $\beta_n^{\varepsilon}(\mu)$ et en développant K_ε au voisinage de 0 on a

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

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

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

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

Correction de Bartlett

Montrons maintenant la propriété de correction au sens de Bartlett. $\beta_n^{\varepsilon}(\mu)$ est défini comme n fois le sup dans le programme dual écrit pour K_ε. $\beta_n^0(\mu)$ correspond alors à la

vraisemblance empirique ($\varphi = K_0$) et $\beta_n^1(\mu)$ au χ^2 ($\varphi = K_1$). 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),

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

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 $\mathcal{O}(n^{-3/2})$ (puisque T_n^1 est bornée en probabilité et 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 $n^{-3/2}$).

Chapitre 3

Empirical φ-discrepancies and quasi-empirical likelihood : exact exponential bounds

3.1 Introduction

Empirical likelihood is now a useful and classical method for testing or constructing confidence regions for the value of some parameters in non-parametric or semi-parametric models. It has been introduced and studied by Owen (1988, 1990), see Owen (2001) for a complete overview and exhaustive references. The now well-known idea of empirical likelihood consists in maximizing a profile likelihood supported by the data, under some model constraints. It can be seen as an extension of "model based likelihood" used in survey sampling when some marginal constraints are available (see (Hartley & Rao, 1968, Deville & Sarndal, 1992)). Owen and many followers have shown that one can get a useful and automatic non-parametric version of Wilks' theorem (stating the convergence of the log-likelihood ratio to a χ^2 distribution). Generalizations of empirical likelihood methods are available for many statistical and econometric models as soon as the parameter of interest is defined by some moment constraints (see (Qin & Lawless, 1994, Newey & Smith, 2004)). It can now be considered as an alternative to the generalized method of moments (GMM, see (Smith, 1997)). Moreover just like in the parametric case, this log-likelihood ratio is Bartlett-correctable. This means that an explicit correction leads to confidence regions with third order properties. The asymptotic error on the level is then of order $\mathcal{O}(n^{-2})$ instead of $\mathcal{O}(n^{-1})$ under some regularity assumptions (see (DiCiccio et al., 1991, Bertail, 2006)).

A possible interpretation of empirical log-likelihood ratio is to see it as the minimization of the Kullback divergence, say K, between the empirical distribution of the data $\mathbb{P}_n$ and a measure (or a probability measure) $\mathbb{Q}$ dominated by $\mathbb{P}_n$, under linear or non-linear constraints imposed on $\mathbb{Q}$ by the model (see (Bertail, 2006)). The use of other pseudo-metrics instead of the Kullback divergence K has been suggested by Owen (1990) and many other authors. For example, the choice of relative entropy has been investigated by DiCiccio & Romano (1990), Jing & Wood (1996), Kitamura & Stutzer (1997) and led to "Entropy econometrics" in the econometric field (see (Golan et al., 1996)). Related results may be found in the probabilistic literature about divergence or the method of entropy in mean (see (Csiszár, 1967, Liese & Vajda, 1987, Léonard, 2001b, Gamboa & Gassiat, 1996, Broniatowski & Kéziou, 2004)). More recently, some generalizations of the empirical likelihood method have also been obtained by using Cressie-Read discrepancies (Baggerly, 1998, Corcoran, 1998) and led to some econometric extensions known as "generalized empirical likelihood" (Newey & Smith, 2004), even if the "likelihood" properties and in particular the Bartlett-correctability in these cases are lost (Jing & Wood, 1996). Bertail et al. (2004) have shown that Owen's original method in the case of the mean can be extended to any regular convex statistical divergence or φ-discrepancy (where φ^* is a regular convex function) under weak assumptions. We call this method "empirical energy minimizers" by reference to the theoretical probabilistic literature on the subject (see (Léonard, 2001b)) and references therein).

However, the previous results (including Bartlett-correction) are all asymptotic results. A natural statistical issue is how the choice of φ^* influences the corresponding confidence regions and their coverage probability, for finite sample size n, in a multivariate setting.

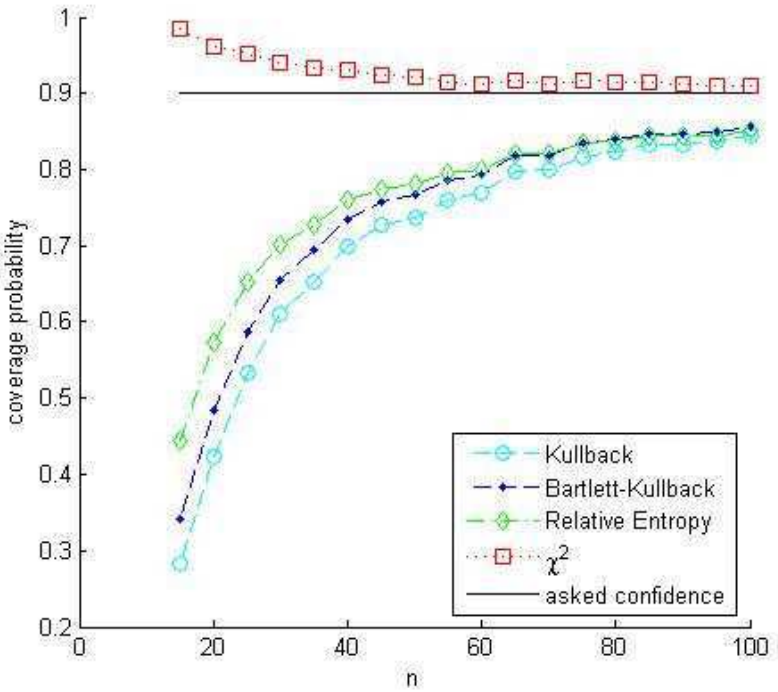


FIG. 3.1 – Coverage probability for different discrepancies

To illustrate this fact, we use different discrepancies to build confidence intervals for the mean of the product of a uniform r.v. with an independent standard gaussian r.v. (a scale mixture) on $\mathbb{R}^6$. The Figure (3.1) represents the coverage probability obtained by Monte-Carlo simulations (100 000 repetitions) for different divergences and different sample sizes n. Asymptotically, all these empirical energy minimizers are theoretically equivalent in the case of the mean (Bertail et al., 2004). However, this simulation clearly stresses their distinct behavior for small sample sizes. Empirical likelihood corresponding to K performs very badly for small sample size, even with a Bartlett-correction. However, the χ^2 divergence (leading to GMM type of estimators) tends to be too conservative. These problems tend to increase with the dimension of the parameter of interest. For very small sample size, Tsao (2004) obtained an exact upper bounds for the coverage probability of empirical likelihood for q, the parameter size, less than 2, which confirms our simulation results. It also sheds some doubt on the relevance of empirical likelihood when n is small compared to q.

One goal of this paper is to introduce and study a family of discrepancies for which we have a non-asymptotic control of the level of the confidence regions -a lower bound for the coverage probability- for any parameter size. The basic idea is to consider a family of barycenters of the Kullback divergence and the χ^2 divergence, called quasi-Kullback, defined by $(1-\varepsilon)K + \varepsilon\chi^2$ for $\varepsilon \in [0,1]$ and to minimize the dual expression of this divergence on the constraints. It can be seen as a quasi-empirical likelihood or a penalized empirical likelihood. The domain of the corresponding divergence is the whole real line making the algorithmic aspects of the problem much more tractable than for empirical likelihood when the number of constraints is large. Moreover, this approach allows us to keep the interesting

properties of both discrepancies. On the one hand, from an asymptotic point of view, we show that this method is still Bartlett-correctable for an adequate choice of ε, typically depending on n. Regions are still automatically shaped by the sample, as in the empirical likelihood case without the limitation stressed by Tsao (2004). On the other hand, for any fixed value of ε, it is possible to use the self-normalizing properties of the χ^2 divergence to obtain non-asymptotic exponential bounds for the error of the confidence intervals.

Exponential bounds for self-normalized sums have been obtained by several authors in the unidimensional case or can be derived from non-uniform Berry-Esséen or Cramer type bounds (Shao, 1997, Jing & Wang, 1999, Chistyakov & Götze, 2003, Jing et al., 2003). However, to our knowledge, non-asymptotic exponential bounds with **explicit constants** are only available for symmetric distribution (Hoeffding, 1963, Efron, 1969, Pinelis, 1994). In this paper, we obtain a generalization of this kind of bounds by using the symmetrization method developed by Panchenko (2003) as well as arguments taken from the literature on self-normalized process (see (Bercu et al., 2002)). Our bounds hold for any value of the parameter size q : one technical difficulty in this case is to obtain an explicit exponential bound for the smallest eigenvalue of the empirical variance. For this, we use chaining arguments from Barbe & Bertail (2004). These bounds are of interest in our quasi-empirical likelihood framework but also for self-normalized sums. A consequence of these results is that quasi-empirical likelihood allows to build asymptotic confidence intervals even if the number of constraints q grows as $o(n/log(n))$.

The layout of this paper is the following. In Part 2, we first recall some basic facts about convex integral functionals and their dual representation. As a consequence, we briefly state the asymptotic validity of the corresponding "empirical energy minimizers" in the case of M-estimators. We then focus in part 3 on a specific family of discrepancies, that we call quasi-Kullback divergences. These pseudo-distances enjoy several interesting convex duality and Lipschitz properties. This makes them an alternative method to empirical likelihood, easier to handle in practical situations. Moreover, for adequate choices of the weight ε, the corresponding empirical energy minimizers are shown to be Bartlett-correctable. In part 4, our main result claims that, for these discrepancies, it is possible to obtain exact asymptotic exponential bounds in a multivariate framework. A data-driven method for choosing the weight ε is also proposed. Part 5 gives some small sample simulation results and compares the confidence regions and their level for different discrepancies. The proofs of the main theorems are postponed. There, some lemmas are also of interest for self-normalized sums. Some additional details, discussions and simulations may be found in Bertail et al. (2005).

3.2 Empirical φ-discrepancy minimizers

3.2.1 Notations : φ-discrepancies and convex duality

We consider a measured space $(\mathcal{X}, \mathcal{A}, \mathcal{M})$ where $\mathcal{M}$ is a space of signed measures. It will be essential for applications to work with signed measures. Let f be a measurable function defined from $\mathcal{X}$ to $\mathbb{R}^r$, $r \geq 1$. For any measure $\mu \in \mathcal{M}$, we write $\mu f = \int f d\mu$ and if μ is a density of probability, $\mu f = \mathbb{E}_\mu(f(X))$. In the following, we consider φ, a convex function whose support $d(\varphi)$, defined as $\{x \in \mathbb{R},\ \varphi(x) < \infty\}$, is assumed to be non-void (φ is said to

be proper). We denote respectively $\inf d(\varphi)$ and $\sup d(\varphi)$, the extremes of this support. For every convex function φ, its convex dual or Fenchel-Legendre transform is given by

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

Recall that φ^* is then a semi-continuous inferiorly (s.c.i.) convex function. We define by $\varphi^{(i)}$ the derivative of order i of φ when it exists. From now on, we will assume the following assumptions for the function φ.

H1 φ is strictly convex and $d(\varphi)$ contains a neighborhood of 0 ;

H2 φ is twice differentiable on a neighborhood of 0 ;

H3 (renormalization) $\varphi(0) = 0$ and $\varphi^{(1)}(0) = 0$, $\varphi^{(2)}(0) > 0$, which implies that φ has an unique minimum at zero ;

H4 φ is differentiable on $d(\varphi)$, that is to say differentiable on $int\{d(\varphi)\}$, with right and left limits on the respective endpoints of the support of $d(\varphi)$, where $int\{.\}$ is the topological interior.

H5 φ is twice differentiable on $d(\varphi)\cap\mathbb{R}^+$ and, on this domain, the second order derivative of φ is bounded from below by $m > 0$.

Let φ satisfies the hypotheses **H1, H2, H3**. Then, the Fenchel dual transform φ^* of φ also satisfies these hypotheses. The φ-discrepancy I_{φ^*} between $\mathbb{Q}$ and $\mathbb{P}$, where $\mathbb{Q}$ is a signed measure and $\mathbb{P}$ a positive measure, is defined as follows :

$$I_{\varphi^*}(\mathbb{Q},\mathbb{P}) = \begin{cases} \int_{\mathcal{X}} \varphi^*\left(\frac{d\mathbb{Q}}{d\mathbb{P}} - 1\right) d\mathbb{P} & \text{if } \mathbb{Q} \ll \mathbb{P} \\ +\infty & \text{else.} \end{cases} \tag{3.1}$$

For details on φ-discrepancies or divergences Csiszàr (Csiszár, 1967) and some historical comments, see Rockafellar (1968, 1970, 1971), Liese & Vajda (1987), Léonard (2001a). It is easy to check that Cressie-Read discrepancies (Cressie & Read, 1984) fulfill these assumptions. Indeed, a Cressie-Read discrepancy can be seen as a φ-discrepancy, with φ^* given by :

$$\varphi^*_\kappa(x) = \frac{(1+x)^\kappa - \kappa x - 1}{\kappa(\kappa-1)},\ \varphi_\kappa(x) = \frac{[(\kappa-1)x+1]^{\frac{\kappa}{\kappa-1}} - \kappa x - 1}{\kappa}$$

for some $\kappa \in \mathbb{R}$. This family contains all the usual discrepancies, such as Relative Entropy ($\kappa \to 1$), Hellinger distance ($\kappa = 1/2$), the χ^2 ($\kappa = 2$) and the Kullback distance ($\kappa \to 0$).

For us, the main interest of φ-discrepancies lies on the following duality representation, which follows from results of Borwein & Lewis (1991) on convex functional integrals (see also (Léonard, 2001b, Broniatowski & Kéziou, 2004)).

Theorem 3.1 *Let $\mathbb{P} \in \mathcal{M}$ be a probability measure with a finite support and f be a measurable function on $(\mathcal{X},\mathcal{A},\mathcal{M})$. Let φ be a convex function satisfying assumptions **H1-H3**. If the following qualification constraint holds,*

$$Qual(\mathbb{P}) : \begin{cases} \exists \mathbb{T} \in \mathcal{M}, \mathbb{T}f = b_0 \ and \\ \inf d(\varphi^*) \ < \ \inf_{\mathcal{X}} \frac{d\mathbb{T}}{d\mathbb{P}} \ \leq \ \sup_{\mathcal{X}} \frac{d\mathbb{T}}{d\mathbb{P}} \ < \ \sup d(\varphi^*) \ \ \mathbb{P}-a.s., \end{cases}$$

then, we have the dual equality :

$$\inf_{\mathbb{Q}\in\mathcal{M}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P})|\ (\mathbb{Q}-\mathbb{P})f = b_0\} = \sup_{\lambda\in\mathbb{R}^r} \left\{\lambda' b_0 - \int_{\mathcal{X}} \varphi(\lambda' f) d\mathbb{P}\right\}. \tag{3.2}$$

*If φ satisfies **H4**, then the supremum on the right hand side of (3.2) is achieved at a point λ^* and the infimum on the left hand side at $\mathbb{Q}^*$ is given by*

$$\mathbb{Q}^* = (1+\varphi^{(1)}(\lambda^{*\prime} f))\mathbb{P}.$$

3.2.2 Empirical optimization of φ-discrepancies

Let $X_1, ...X_n$ be i.i.d. r.v.'s defined on $\mathcal{X} = \mathbb{R}^p$ with common probability measure $\mathbb{P} \in \mathcal{M}$. Consider the empirical probability measure $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_{X_i} is the Dirac measure at X_i. We will here consider that the parameter of interest $\theta \in \mathbb{R}^q$ is the solution of some M-estimation problem $\mathbb{E}_{\mathbb{P}} f(X,\theta) = 0$, where f is now on a regular differentiable function from $\mathcal{X}\times\mathbb{R}^q \to \mathbb{R}^r$. For simplicity, we now assume that f takes its value in $\mathbb{R}^q$, that is $r = q$ and that there is no over-identification problem. The over-identified case can be treated similarly by first reducing the problem to the strictly identified case (see (Qin & Lawless, 1994)).

For a given φ, we define, by analogy to Owen (1990, 2001), the quantity

$$\beta_n(\theta) = n \inf_{\{\mathbb{Q}\ll\mathbb{P}_n,\ \mathbb{E}_{\mathbb{Q}} f(X,\theta)=0\}} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n)\}$$

We define the corresponding random confidence region

$$\mathscr{C}_n(\eta) = \{\theta\in\mathbb{R}^q\ |\exists \mathbb{Q}\ll\mathbb{P}_n \text{ with } \mathbb{E}_{\mathbb{Q}} f(X,\theta) = 0 \text{ and } nI_{\varphi^*}(\mathbb{Q},\mathbb{P}_n)\le\eta\},$$

where $\eta = \eta(\alpha)$ is a quantity such that

$$\Pr(\theta\in\mathscr{C}_n(\eta)) = 1-\alpha+o(1).$$

Denote $\mathcal{M}_n = \{\mathbb{Q}\in\mathcal{M} \text{ with } \mathbb{Q}\ll\mathbb{P}_n\} = \{\mathbb{Q} = \sum_{i=1}^n q_i\,\delta_{X_i},\ (q_i)_{1\le i\le n}\in\mathbb{R}^n\}$. Considering this set of measures, instead of a set of probabilities, can be partially explained by Theorem 3.1. It establishes the existence of the solution of the dual problem for general signed measures, but in general not for probability measures.

The underlying idea of empirical likelihood and its extensions is actually a plug-in rule. Consider the functional defined by

$$M(\mathbb{P},\theta) = \inf_{\{\mathbb{Q}\in\mathcal{M},\ \mathbb{Q}\ll\mathbb{P},\ \mathbb{E}_{\mathbb{Q}} f(X,\theta)=0\}} I_{\varphi^*}(\mathbb{Q},\mathbb{P})$$

that is, the minimization of a contrast under the constraints imposed by the model. This can be seen as a projection of $\mathbb{P}$ on the model of interest for the given pseudo-metric I_{φ^*}. If the model is true at $\mathbb{P}$, that is, if $\mathbb{E}_{\mathbb{P}} f(X,\theta) = 0$ at the true underlying probability $\mathbb{P}$, then clearly $M(\mathbb{P},\theta) = 0$. A natural estimator of $M(\mathbb{P},\theta)$ for fixed θ is given by the plug-in estimator $M(\mathbb{P}_n,\theta)$, which is $\beta_n(\theta)/n$. This estimator can then be used to test $M(\mathbb{P},\theta) = 0$ or, in a dual approach, to build confidence region for θ by inverting the test.

For $\mathbb{Q}$ in $\mathcal{M}_n$, the constraints can be rewritten as $(\mathbb{Q}-\mathbb{P}_n)f(.,\theta) = -\mathbb{P}_n f(.,\theta)$. Using Theorem 3.1, we get the dual representation

$$\begin{aligned}\beta_n(\theta) &:= n \inf_{\mathbb{Q}\in\mathcal{M}_n} \{I_{\varphi^*}(\mathbb{Q},\mathbb{P}_n),\ (\mathbb{Q}-\mathbb{P}_n)f(.,\theta) = -\mathbb{P}_n f(.,\theta)\} \\ &= n \sup_{\lambda\in\mathbb{R}^q} \mathbb{P}_n\Big(-\lambda' f(.,\theta) - \varphi(\lambda' f(.,\theta))\Big). \end{aligned} \tag{3.3}$$

Notice that $-x-\varphi(x)$ is a strictly concave function and that the function $\lambda \to \lambda' f$ is also concave. The parameter λ can be simply interpreted as the Kuhn & Tucker coefficient associated to the original optimization problem. From this representation of $\beta_n(\theta)$, we can now derive the usual properties of the empirical likelihood and its generalization. In the following, we will also use the notations

$$\overline{f}_n = \frac{1}{n}\sum_{i=1}^n f(X_i,\theta),\ S_n^2 = \frac{1}{n}\sum_{i=1}^n f(X_i,\theta)f(X_i,\theta)' \text{ and } S_n^{-2} = (S_n^2)^{-1}.$$

The following theorem states that generalized empirical likelihood essentially behaves asymptotically like a self-normalized sum. Links to self-normalized sum for finite n will be investigated in paragraph 3.4.

Theorem 3.2 *Let X, $X_1,...,X_n$ be in $\mathbb{R}^p$, i.i.d. with probability $\mathbb{P}$ and $\theta\in\mathbb{R}^q$ such that $\mathbb{E}_{\mathbb{P}} f(X,\theta) = 0$. Assume that $S^2 = \mathbb{E}_{\mathbb{P}} f(X,\theta)f(X,\theta)'$ is of rank q and that φ satisfies the hypotheses **H1-H4**. Assume that the qualification constraints Qual$(\mathbb{P}_n)$ hold. For any α in $]0,1[$, set $\eta = \frac{\varphi^{(2)}(0)\chi_q^2(1-\alpha)}{2}$, where $\chi_q^2(.)$ is the χ^2 distribution quantile. Then $\mathscr{C}_n(\eta)$ is a convex asymptotic confidence region with*

$$\begin{aligned}\lim_{n\to\infty} \Pr(\theta\notin\mathscr{C}_n(\eta)) &= \lim_{n\to\infty} \Pr(\beta_n(\theta)\geq\eta) \\ &= \lim_{n\to\infty} \Pr\left(n\overline{f}_n' S_n^{-2}\overline{f}_n \geq \chi_q^2(1-\alpha)\right) \\ &= 1-\alpha.\end{aligned}$$

The proof of this theorem starts from the convex dual-representation and follows the main arguments of Bertail et al. (2004) and Owen (2001) for the case of the mean. It is left to the reader.

Note 3.1 *If φ is finite everywhere then the qualification constraints are not needed (this is for instance the case for the χ^2 divergence). However, in the case of empirical likelihood or the generalized empirical method introduced below, this actually simply puts some restriction on the θ which are of interest as noticed in the following examples.*

3.2.3 Two basic examples

Empirical likelihood and the Kullback discrepancy In the particular case

$$\varphi_0(x) = -x-\log(1-x) \text{ and } \varphi_0^*(x) = x-\log(1+x)$$

corresponding to the Kullback divergence $K(\mathbb{Q},\mathbb{P}) = -\int \log(\frac{d\mathbb{Q}}{d\mathbb{P}})d\mathbb{P}$, the dual program obtained in (3.3) becomes, for the admissible θ,

$$\beta_n(\theta) = \sup_{\lambda\in\mathbb{R}^q} \left(\sum_{i=1}^n \log(1+\lambda' f(X_i,\theta)) \right).$$

As a parametric likelihood indexed by λ, it is easy to show that $2\beta_n(\theta)$ is asymptotically $\chi^2(q)$ when $n\to\infty$, if the variance of $f(X,\theta)$ is definite. It is also Bartlett-correctable (DiCiccio et al., 1991). Using a duality point of view, the proof of the Bartlett-correctability is almost immediate, see Mykland (1995) and Bertail (2004, 2006). For a general discrepancy, the dual form is not a likelihood and may not be Bartlett-correctable, see DiCiccio et al. (1991) and Jing & Wood (1996).

Moreover, we necessarily have the $q_i's > 0$ and $\sum_{i=1}^n q_i = 1$, so that the qualification constraint essentially means that 0 belongs to the convex hull of the $f(X_i,\theta)$. Only the θ's which satisfy this constraint are of interest to us; asymptotically, this is by no mean a restriction, unless we have some very specific configuration of the data.

GMM and χ^2 discrepancy The particular case of the χ^2 discrepancy corresponds to $\varphi_2(x) = \varphi_2^*(x) = \frac{x^2}{2}$. $\beta_n(\theta)$ can be explicitly calculated. Indeed, we get easily that $\lambda = S_n^{-2}\overline{f}_n$ so that, by Theorem 3.1, the minimum is attained at $\mathbb{Q}^* = \sum_{i=1}^n q_i\delta_{X_i}$ with

$$q_i = \frac{1}{n}(1+\overline{f}_n' S_n^{-2} f(X_i,\theta))$$

and

$$I_{\varphi_2^*}(\mathbb{Q}^*,\mathbb{P}_n) = \sum_{i=1}^n \frac{(nq_{i,n}-1)^2}{2n} = \frac{1}{2}\overline{f}_n' S_n^{-2}\overline{f}_n,$$

which is exactly the square of a self-normalized sum which typically appears in the Generalized Method of Moments (GMM).

Notice that, in opposition to the Kullback discrepancy, we may charge positively some region outside of the convex hull of the points, yielding bigger (that is too conservative) confidence region. However, as noticed in the introduction, the results of Tsao (2004) shows that taking the convex hull of the points (the largest confidence region for empirical likelihood) may yield too narrow confidence regions, when n is small compared to q.

Note 3.2 *If S_n^2 is of rank $l < q$, write $S_n^2 = R'\begin{pmatrix}\Delta_n & 0\\ 0 & 0\end{pmatrix}R$, where Δ_n is invertible of rank l, $R = \begin{pmatrix}R_a\\ R_b\end{pmatrix}$ is an orthogonal matrix with $R_a \in \mathfrak{M}_{l,q}(\mathbb{R})$ and $R_b \in \mathfrak{M}_{q-l,q}(\mathbb{R})$. By straightforward arguments, the duality relationship still holds and becomes*

$$\beta_n(\theta) = n\sup_{\lambda\in\mathbb{R}^l}\left\{-\lambda' R_a\overline{f}_n - \frac{1}{2}\lambda'\Delta_n\lambda\right\} = \frac{1}{2}(R_a\overline{f}_n)'\Delta_n^{-1}(R_a\overline{f}_n).$$

Notice that $(R_a\overline{f}_n)(R_a\overline{f}_n)' = \Delta_n$. This means that if S_n^2 has rank $l<q$ we can always reduce the problem to the study of a self-normalized sum in $\mathbb{R}^l$ and that, from an algorithmic point of view this reduction is carried out internally by the optimization program. From now on, we will assume that S_n^2 is of rank $l = q$.

3.3 Quasi-Kullback and Bartlett-correctability

The main underlying idea of this section is that we want to keep the good properties of the Kullback discrepancy and to avoid some algorithmic problems linked with the behavior of the log of the Kullback discrepancy in the neighborhood of 0. For this, we will introduce family of divergences, the quasi-Kullback. This kind of discrepancies is actually currently used in the convex optimization literature (see for instance (Ausslender et al., 1999)) because the resulting optimization algorithm leads to efficient tractable interior point solutions when the number of constraint is large.

For $\varepsilon \in]0;1]$ and $x \in]-\infty;1[$ let,

$$K_\varepsilon(x) = \varepsilon\, x^2/2 + (1-\varepsilon)(-x - \log(1-x)).$$

We call the corresponding K_ε^*-discrepancy, the quasi-Kullback discrepancy. The parameter $\varepsilon > 0$ may be interpreted as a regularization parameter (proximal in term of convex optimization). This family fulfills our hypotheses **H1-H5**. Its Fenchel-Legendre transform K_ε^* has the following explicit expression, for all x in $\mathbb{R}$:

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

Note that the second order derivative of K_ε is bounded from below : $K_\varepsilon^{(2)}(x) \geq \varepsilon$. Moreover, the second order derivative of K_ε^* is bounded both from below and above : $0 \leq K_\varepsilon^{*(2)}(x) \leq 1/\varepsilon$. These controls ensure a quick and regular convergence of the algorithms based on such discrepancies. The corresponding "quasi-empirical likelihood" may be seen as a "regularized" empirical likelihood.

The following theorem establishes sufficient conditions on the regularization parameter ε to obtain the Bartlett-correctability of quasi-empirical likelihood.

Theorem 3.3 *Under the assumptions of Theorem 3.2, assume that $f(X,\theta)$ satisfies the Cramer condition : $\overline{\lim}_{||t||\to\infty}|\mathbb{E}_{\mathbb{P}}\exp(it'f(X,\theta))| < 1$, as well as the moment condition $\mathbb{E}_{\mathbb{P}}||f(X,\theta)||^s < \infty$, for $s > 8$.*
If $\varepsilon \doteq \varepsilon_n = \mathcal{O}\left(n^{-3/2}/\log(n)\right)$ then the quasi-empirical likelihood is Bartlett-correctable up to $\mathcal{O}(n^{-3/2})$.

This choice of ε is probably not optimal but considerably simplifies the proof. An attentive reading of Corcoran (1998) shows that, if ε is small enough, the statistic is Bartlett-correctable. Unfortunately, as our discrepancy depend on n, Corcoran's result cannot be applied directly and does not allow ε to be precisely calibrated. We conjecture that, at the cost of tedious calculations, the rate of ε_n in $o(n^{-1})$ is enough, at least to get Bartlett-correctability up to $o(n^{-1})$.

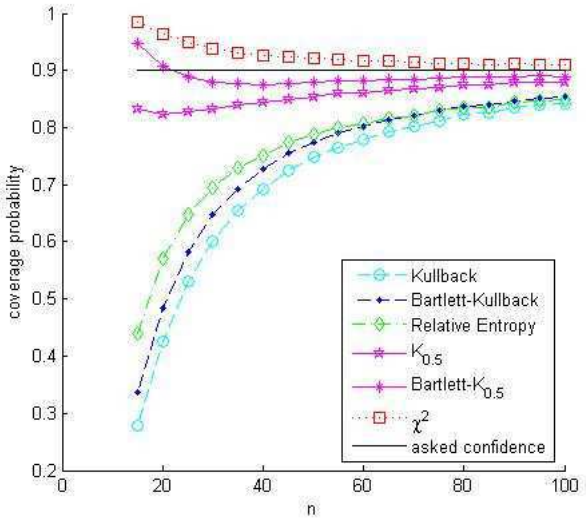


FIG. 3.2 – Coverage probabilities and Quasi-Kullback

Figure 3.2 illustrates the improvements coming from the use of Quasi-Kullback. It presents the coverage probabilities of the usual discrepancies given in the introduction, as well as the ones for Quasi-Kullback discrepancy (for a given value of $\varepsilon = 0.5$) on the same data. As expected, the Quasi-Kullback discrepancy leads to a confidence region with a coverage probability much closer to the targeted one, especially with a Bartlett adjustment even for an ε which is not close to 0.

3.4 Exponential bounds for self-normalized sums and quasi-empirical likelihood

Another interesting feature of quasi-Kullback discrepancies is that the control of the second order derivatives allows the behavior of $\beta_n(\theta)$ to be linked to that of self-normalized sums. We thus can get exponential bounds for the quantities of interest. Some of the bounds that we propose here for self-normalized sums are new and of interest by themselves. These bounds may be quite easily obtained in the symmetric case (that is for random variables having a symmetric distribution) and are well-known in the unidimensional case.

Self-normalized sums have recently given rise to an important literature : see for instance Jing & Wang (1999), Chistyakov & Götze (2003) or Bercu et al. (2002) for self-normalized processes. Unfortunately, except in the symmetric case, these bounds are not universal and depend on higher order moments, $\gamma_3 = \mathbb{E}|S^{-1}f(X_i,\theta)|^3$ or even an higher moment condition : $\gamma_{10/3} = \mathbb{E}|S^{-1}f(X_i,\theta)|^{10/3}$. Actually, uniform bounds in $\mathbb{P}$ are impossible to obtain, otherwise this would contradict Bahadur & Savage (1956)'s result on the non-existence of uniform confidence region over large class of probabilities, see Romano & Wolf (2000) for related results. For symmetric random variables, related inequalities for self-normalized sums have been obtained by Pinelis (1994), following Eaton (1974).

In the general non-symmetric case, for $q = 1$, if $\gamma_{10/3} < \infty$, for some $A \in \mathbb{R}$ and some $a \in]0, 1[$, the result of Jing & Wang (1999) leads to

$$\Pr\left(\frac{n}{2}\overline{f}_n^2/S_n^2 \geq \varepsilon\eta\right) = \chi_1^2(\varepsilon\eta) + A\gamma_{10/3}n^{-1/2}e^{-a\varepsilon\eta}. \tag{3.4}$$

However the constants A and a are not explicit and the bound is of no practical use. In the non-symmetric case our bounds are worse than (3.4) as far as the control of the approximation by a χ^2 distribution are concerned, but entirely explicit.

Theorem 3.4 *Let $(Z_i)_{i=1,\cdots,n}$ be an i.i.d. sample in $\mathbb{R}^q$ with probability $\mathbb{P}$. Let's define $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$, $S_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$ and $S^2 = \mathbb{E}_{\mathbb{P}} Z_1 Z_1'$. Suppose that S^2 is of rank q. Then the following inequalities hold,* ***for finite*** *$n > q$ and for $u < nq$,*

a) *if Z_1 has a symmetric distribution, without any moment assumption,*

$$\Pr\left(n\overline{Z}_n' S_n^{-2}\overline{Z}_n \geq u\right) \leq 2qe^{-\frac{u}{2q}}; \tag{3.5}$$

b) *for general distribution of Z_1 with kurtosis $\gamma_4 < \infty$, for any $a > 1$,*

$$\begin{aligned}\Pr\left(n\overline{Z}_n' S_n^{-2}\overline{Z}_n \geq u\right) &\leq 2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}}\gamma_4^{-\widetilde{q}} e^{-\frac{n}{\gamma_4(q+1)}\left(1-\frac{1}{a}\right)^2} \\ &\leq 2qe^{1-\frac{u}{2q(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n}{\gamma_4(q+1)}\left(1-\frac{1}{a}\right)^2}\end{aligned} \tag{3.6}$$

with $\widetilde{q} = \frac{q-1}{q+1}$, $\gamma_4 = \mathbb{E}_{\mathbb{P}}(\|S^{-1}Z_1\|_2^4)$ and $C(q) = \frac{(2e\pi)^{2\widetilde{q}}(q+1)}{2^{2/(q+1)}(q-1)^{3\widetilde{q}}} \leq \frac{(2e\pi)^2(q+1)}{(q-1)^{3\widetilde{q}}} \leq 18$.
Moreover for $nq \leq u$, we have

$$\Pr\left(n\overline{Z}_n S_n^{-2}\overline{Z}_n \geq u\right) = 0.$$

The proof is postponed to the Section 3.6.3 . Part a) in the symmetric multidimensional case follows by an easy but crude extension of Hoeffding (1963) (or (Efron, 1969, Eaton & Efron, 1970)). The exponential inequality (3.5) is classical in the unidimensional case. Other type of inequalities with suboptimal rate in the exponential have also been obtained by Major (2004).

In the general multidimensional framework, the main difficulty is actually to keep the self-normalized structure when symmetrizing the original sum. Another difficulty is to have a precise control of the behavior of the smallest eigenvalue of the normalizing empirical variance. The second term in the right hand side of inequality (3.6) is essentially due to this control. The crude bound obtained in part a) allows us to use a multidimensional extension of a symmetrization lemma by Panchenko (2003). However for $q > 1$, the bound of part a) is clearly not optimal. A better bound, which has not exactly an exponential form, has been obtained by Pinelis (1994). It essentially says that in the symmetric case the tail of the self-normalized sum can essentially be bounded by the tail of a χ^2 distribution (up to a constant equal to $2e^3/9$). This bounds gives the right behavior of the tail (in q) when n grows, which is not the case for a). However, in the unidimensional case a) still gives a better approximation than Pinelis (1994). It still can be used in the multidimensional case to get crude but exponential bounds.

For these reason, we will extend the results of Theorem 3.4 when using a χ^2 type of control. This essentially consists in extending lemma 1 of Panchenko (2003) to non exponential bound.

In the following, we denote f_q the density function of a χ^2_q law, which is given by $f_q(x) = \frac{1}{2^{q/2}\Gamma(q/2)} x^{q/2-1} e^{-\frac{x}{2}}$, with $\Gamma(p) = \int_0^{+\infty} x^{p-1}e^{-x}dx$. We denote $\overline{F}_q$ the survival function, $\overline{F}_q(x) = \int_x^{+\infty} f_q(y)dy$.

Theorem 3.5 *We use the same notations as in the Theorem 3.4. Then the following inequalities hold,* ***for finite*** $n > q$ *and for* $u < nq$,

a) *(Pinelis 1994) if* Z_1 *has a symmetric distribution, without any moment assumption,*

$$\Pr\left(n\overline{Z}'_n S_n^{-2}\overline{Z}_n \geq u\right) \leq \frac{2e^3}{9}\overline{F}_q(u), \tag{3.7}$$

b) *for general distribution of* Z_1 *with kurtosis* $\gamma_4 < \infty$, *for any* $a > 1$ *and for* $2q(1+a) \leq u$,

$$\begin{aligned}\Pr\left(n\overline{Z}'_n S_n^{-2}\overline{Z}_n \geq u\right) &\leq \frac{2e^3}{9\Gamma(\frac{q}{2}+1)}\left(\frac{u-q(1+a)}{2(1+a)}\right)^{\frac{q}{2}} e^{-\frac{u-q(1+a)}{2(1+a)}} + C(q)\left(\frac{n^3}{\gamma_4}\right)^{\widetilde{q}} e^{-\frac{n\left(1-\frac{1}{a}\right)^2}{\gamma_4(q+1)}} \\ &\leq \frac{2e^3}{9\Gamma(\frac{q}{2}+1)}\left(\frac{u-q(1+a)}{2(1+a)}\right)^{\frac{q}{2}} e^{-\frac{u-q(1+a)}{2(1+a)}} + C(q)\, n^{3\widetilde{q}} e^{-\frac{n\left(1-\frac{1}{a}\right)^2}{\gamma_4(q+1)}}\end{aligned} \tag{3.8}$$

Moreover, for $nq \leq u$, $\Pr\left(n\overline{Z}'_n S_n^{-2}\overline{Z}_n \geq u\right) = 0$.

Note 3.3 *In the best case, past studies give some bounds for n sufficiently large, without an exact value for "sufficiently large". Here, the bounds are valid for any n. All the constants are also explicit. This bound may also be used to give some ideas on the sample size needed to reach a given confidence level (as a function of q and* γ_4*).*

The following corollary implies that, for the whole class of quasi-Kullback discrepancies, the finite sample behavior of the corresponding empirical energy minimizers can be reduced to the study of a self-normalized sum.

Corollary 3.6 *Under the hypotheses of Theorem 3.2, the following inequalities hold,* ***for finite*** $n > q$, *for any* $\eta > 0$, *for any* $n \geq \frac{2\varepsilon\eta}{q}$,

$$\Pr(\theta \notin \mathscr{C}_n(\eta)) = \Pr(\beta_n(\theta) \geq \eta) \leq \Pr\left(n\overline{f}'_n S_n^{-2}\overline{f}_n \geq 2\varepsilon\eta\right). \tag{3.9}$$

Else if $n > \frac{2\varepsilon\eta}{q}$, $\Pr(\theta \notin \mathscr{C}_n(\eta)) = 0$.
Then, for $n > 2q$, *bounds (3.5-3.8) may be used with* $u = 2\varepsilon\eta$ *and* $Z_i = f(X_i, \theta)$.

Note 3.4 *In Hjort et al. (2004), convergence of empirical likelihood is investigated when q is allowed to increase with n. They show that convergence to a* χ^2 *distribution still holds when* $q = O(n^{\frac{1}{3}})$ *as n tends to infinity.*

Our bounds shows that even if $q = o\left(n/log(n)\right)$, *it is still possible to get asymptotically valid confidence intervals with our bounds. Notice that the constant* $C(q)$ *does not increase with q as can be seen on Figure 3.3.*

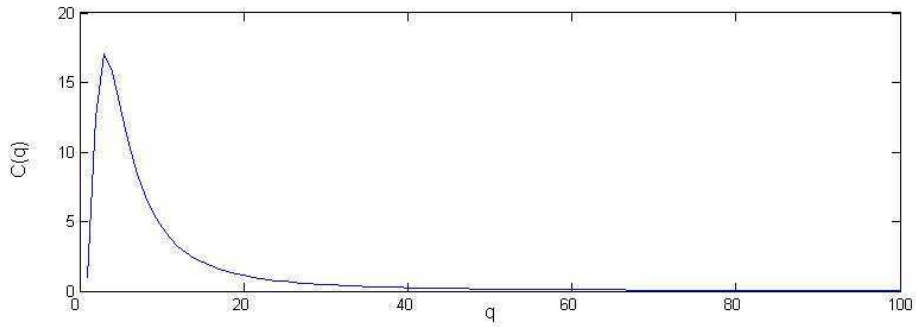


FIG. *3.3 – Value of $C(q)$ as a function of q*

A close examination of the bounds shows that essentially $q\gamma_4$ has to be small compared to n for practical use of these bounds. Of course practically γ_4 is not known, however one may use an estimator or an upper bound for this quantity to get some insight on a given estimation problem.

Notice that the bounds are non-informative when $\varepsilon \to 0$, which corresponds to empirical likelihood. Actually, it is not possible to establish an exponential bound for this case. If we were able to do so, for a sufficiently large η, we could control the confidence region built with empirical likelihood for any level $1-\alpha$. This would contradict the statements of Tsao (2004), which gives a lower bound for the attainable levels.

3.5 Discussion and simulation results

3.5.1 Non-asymptotic comparisons

Some previous simulations (see (Bertail et al., 2005)) show that, for small values of n, the values of η are quite high, leading to confidence regions that may be too conservative but that are very robust. In the following

- "Symmetric bound" corresponds to η obtained by inverting the Pinelis inequality in the symmetric case, that is the quantile of a χ^2.
- "NS", for "Non-symmetric", corresponds the η obtained by inverting the general exponential bounds, from Theorems 3.4b) or 3.5b).

Simulations show that the profile quasi-likelihood gets wider as ε increases. As a consequence, the asymptotic confidence intervals become wider. With the non-asymptotic bounds, the behavior of the corresponding confidence interval as ε increases is more delicate to understand. The profile likelihood gets wider but the η's corresponding to the symmetric bound and NS bounds decrease like $1/\varepsilon$. These two behaviors have contradictory effects on the confidence intervals $\mathscr{C}_n(\eta)$. It seems that the effect of the decrease of η dominates : the confidence intervals get smaller when ε increases. In higher dimension or for a smaller α, the two contradictory effects could be balanced.

In Figure 3.4, we build confidence regions for the mean of multi-dimensional ($q = 2$) data, for two sizes ($n = 500$ and 2000) and two distributions :

1) a couple of independent gaussian scale mixtures (that is realizations of $U * N$, where

U and N are respectively independent uniform r.v.'s on [0,1] and standard gaussian r.v.'s) and

2) the distribution $\frac{1}{100}\cdot\delta_{(10,10)} + \frac{0.81}{4}\sum_{a,a'=\pm1}\delta_{(a,a')} + \frac{0.09}{2}\sum_{a=\pm1}\big(\delta_{(a,10)}+\delta_{(10,a)}\big)$, that will be referred as discrete distribution d_1.

We give in Figure 3.4 the corresponding 90% confidence regions, using respectively the asymptotic approximation from Theorem 3.2, the symmetric bound from Theorem 3.5a) and the general bounds (NS) from Theorems 3.4b) and 3.5b) with the true kurtosis.

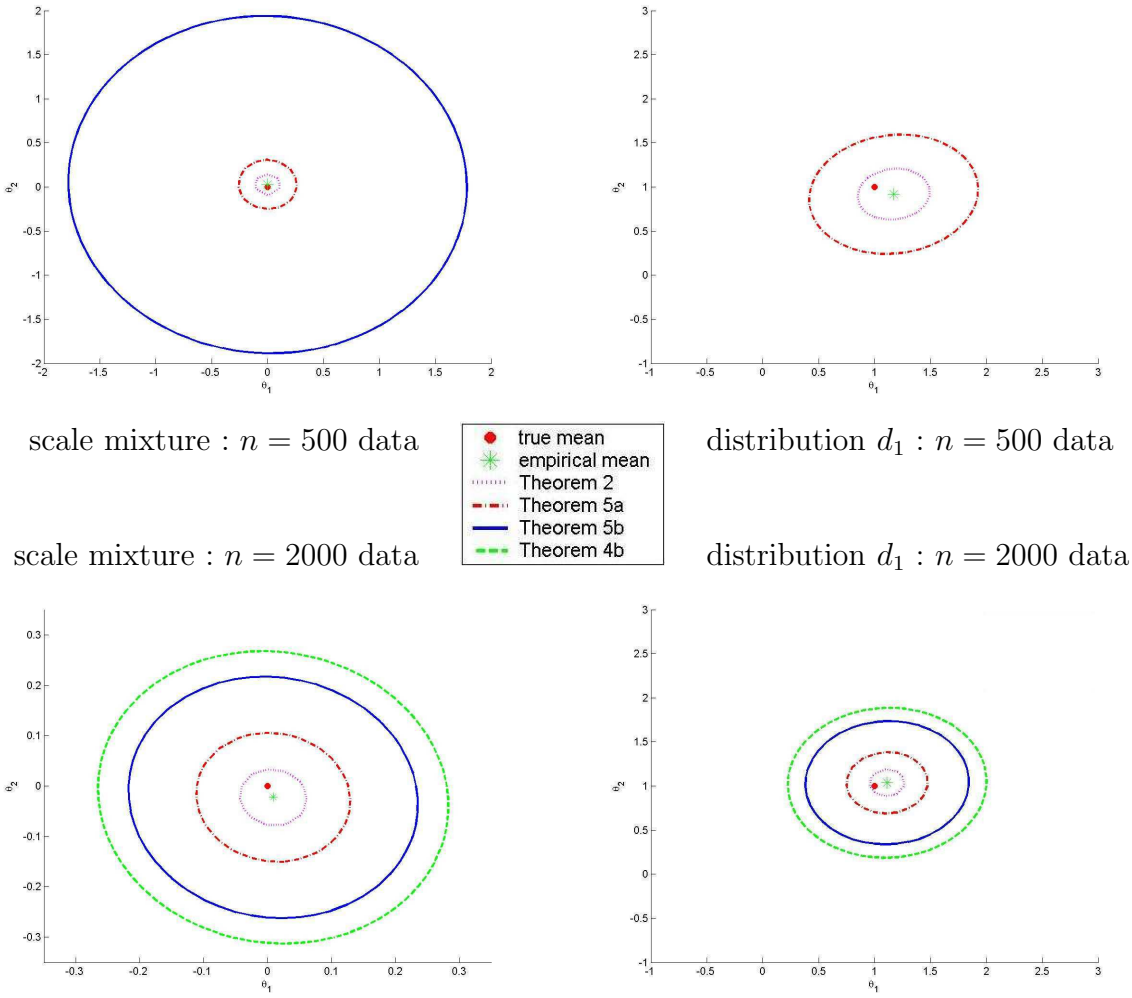


FIG. 3.4 – Confidence regions, for 2 distributions and 2 data sizes

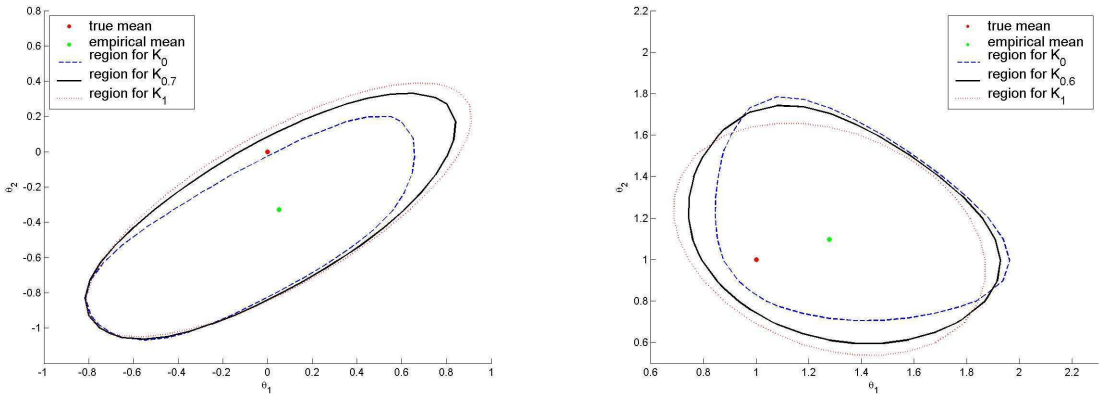
For small sample size, as expected, the confidence regions obtained with NS bounds are quite large (for our discrete data and $n = 500$, the regions are too large to be represented on the figure) with a coverage probability close to 1. On the contrary, the asymptotic confidence regions are small but when the distribution has a large γ_4, the coverage probability can be significantly smaller than the targeted level $1-\alpha$. Thus the use of NS bounds are essentially justified to protect oneself against exotic distributions.

3.5.2 Adaptative asymptotic confidence regions

Corollary 3.6 does not allow for a precise calibration of ε for finite sample size. Indeed, the finite exponential bounds essentially say that the bigger ε is (close to 1), the better the bound. This clearly advocates that, in term of our bound sizes, the χ^2 discrepancy leads to the best results. This is partially true in the sense that the χ^2 leads immediately to a self-normalized sum which has quite robust properties. However, it can be argued that, for regular enough distributions, the χ^2 discrepancy leads to confidence regions that are too conservative. The result on Bartlett-correctability suggests that the bias of the empirical minimizer for quasi-Kullback is smaller for very small values of ε (see also Newey and Smith (Newey & Smith, 2004) for argument in that direction). Choosing adequately ε could result in a better equilibrium and a compromise between coverage probability and the adaptation to the data.

From a practical point of view, several choices are possible for calibrating ε. A simple solution is simply to use cross-validation (either bootstrap, leave one-out or K-fold methods). Of course, this is very computationally-expensive but the use of a quasi-Kullback distance eases the convergence of the algorithms. It is not clear how the use of cross-validation and thus the use of an ε depending on the data will deteriorate the finite sample bounds.

The Figure 3.5 allows us to compare the asymptotic confidence regions built with the Kullback discrepancy (K_0), the χ^2 (K_1) and the Quasi-Kullback (K_ε) with ε chosen by cross-validation, for a parameter in $\mathbb{R}^2$. The algorithm leads to $\varepsilon \simeq 0.7$ for the scale mixture example and $\varepsilon \simeq 0.6$ for a standard exponential distribution .



scale mixture : 15 data exponential distribution : 25 data

FIG. 3.5 – Asymptotic confidence regions for data driven K_ε.

Figure 3.6 represents the coverage probability obtained by Monte-Carlo (25 000 repetitions) simulations of scale-mixture distribution with $q = 6$ for K_ε with data driven ε and some specific choice of ε, for different sample sizes n.

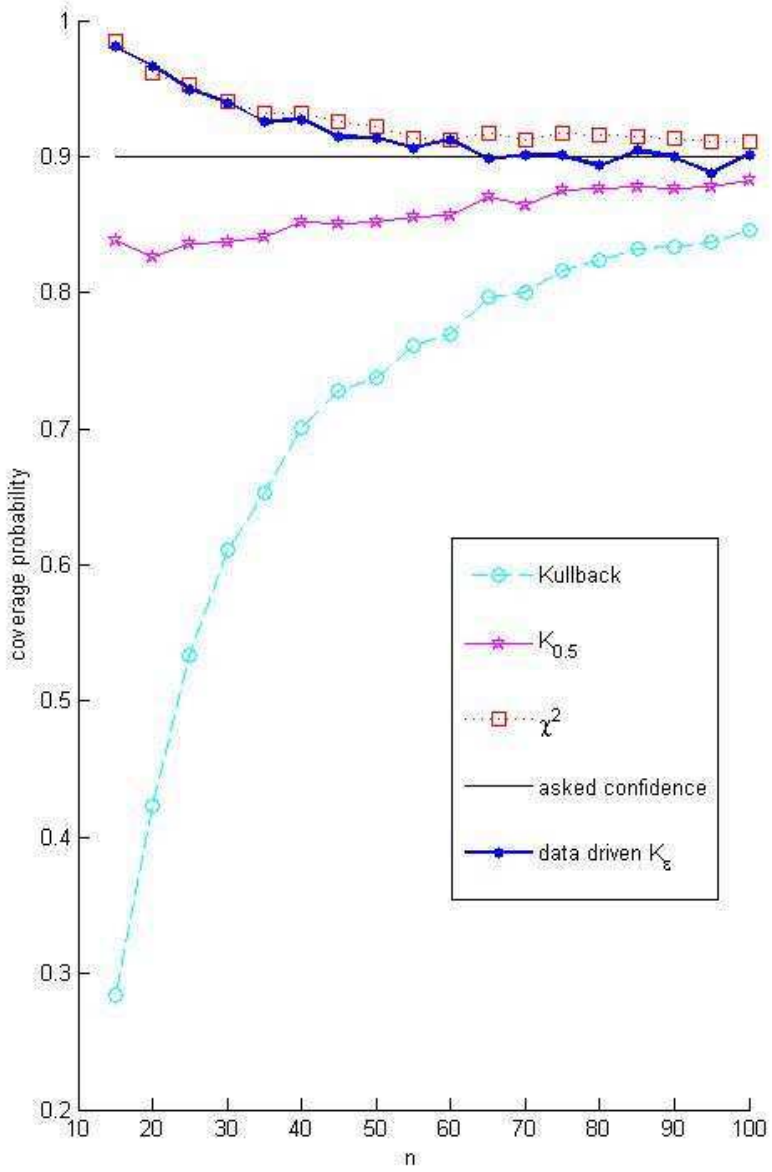


FIG. 3.6 – Coverage probability for different data sizes n for data-driven ε.

In multidimensional case ($q > 1$), with n finite, the volume of the confidence region for the quasi-Kullback divergence remains closed to the volume of the ellipsoid corresponding to the χ^2 divergence with a better coverage probability.

The adaptative value of ε decreases with n : over our 25 000 Monte-Carlo repetitions, the mean value of ε is 1 for $n = 15$ and $n = 20$. It decreases to 0.7 for $n = 100$.

For smooth distributions like our scale mixture, the coverage probability of the confidence

region constructed with the calibrated K_ε is close to the targeted one. Moreover, the region is small and adapts to the data. Note that when, for all values of ε, the cross-validation estimate of the coverage probability is smaller than the targeted confidence, the distribution may be "exotic". In such a case, the NS bound should be considered.

The simulations and graphics have been computed with Matlab : algorithms are available from the authors on request. The Monte-Carlo simulations of Figure 3.6 have been carried out on 18 computers with 2.5 GHz processors and took 18*200 hours of computation time.

3.6 Proofs of the main results

3.6.1 Proof of theorem 3.3

Write $\beta_n^\varepsilon(\theta)$ for the the value of n times the sup in the dual program (3.3) when $\varphi = K_\varepsilon$. $\beta_n^0(\theta)$ corresponds to the log likelihood ratio for Kullback discrepancy $\varphi = K_0$ and $\beta_n^1(\theta)$ corresponds to the minimization of the χ^2-divergence $\varphi = K_1$. Let $\mathbb{E}_n$ be either the true value of $\mathbb{E}[\beta_n^0(\theta)]/q$ or an estimator of this quantity such that empirical likelihood is Bartlett-correctable when standardized by this quantity. We denote

$$T_n^\varepsilon = \frac{2\beta_n^\varepsilon(\theta)}{\mathbb{E}_n}.$$

Then, using DiCiccio et al. (1991) (see also (Bertail, 2006)), under the Cramer condition and assuming $\mathbb{E}_{\mathbb{P}}||f(X,\theta)||^8 < \infty$, the Bartlett-correctability of T_n^0 implies that

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

where we denote $\bar{F}_Z(.) = \int_0^{+\infty} d\mathbb{P}(z)$, when $Z \sim \mathbb{P}$. This equality implies in particular that

$$\bar{F}_{T_n^0}(\eta - n^{-\frac{3}{2}}) = \bar{F}_{\chi^2(q)}(\eta) + \mathcal{O}(n^{-\frac{3}{2}}). \tag{3.10}$$

Now, we can write

$$\begin{aligned} T_n^\varepsilon &= \frac{2}{\mathbb{E}_n} \sup_{\lambda \in \mathbb{R}^q} \left\{ \sum_{i=1}^n \lambda' f(X_i,\theta) - \sum_{i=1}^n K_\varepsilon(\lambda' f(X_i,\theta)) \right\} \\ &\leq \frac{2}{\mathbb{E}_n} \left\{ \varepsilon\beta_n^1(\theta) + (1-\varepsilon)\beta_n^0(\theta) \right\}. \end{aligned}$$

In other words

$$T_n^\varepsilon \leq T_n^0 + \varepsilon \left[T_n^1 - T_n^0 \right].$$

This implies

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

We also have with (3.10)

$$\begin{aligned} \bar{F}_{T_n^0+\varepsilon[T_n^1-T_n^0]}(\eta) &\leq \Pr(T_n^0 + n^{-\frac{3}{2}} \geq \eta) + \Pr(|T_n^1 - T_n^0| \geq \varepsilon^{-1} n^{-\frac{3}{2}}) \\ &= \bar{F}_{T_n^0}(\eta - n^{-\frac{3}{2}}) + \Pr(|T_n^1 - T_n^0| \geq \varepsilon^{-1} n^{-\frac{3}{2}}) \\ &= \bar{F}_{\chi^2}(\eta) + \mathcal{O}(n^{-\frac{3}{2}}) + \Pr(|T_n^1 - T_n^0| \geq \varepsilon^{-1} n^{-\frac{3}{2}}). \end{aligned}$$

If we take ε of order $n^{-3/2}\log(n)^{-1}$, the last term in the right hand side of this inequality is of order $\mathcal{O}(n^{-3/2})$. This can be shown by using for example the moderate deviation inequality (3.4) for T_n^1 and the fact that T_n^0 is already Bartlett-correctable. It follows that the corresponding discrepancy is still Bartlett-correctable, at least up to the order $\mathcal{O}(n^{-3/2})$.

3.6.2 Some bounds for self-normalized sums

Lemma 3.1 (Extension of Panchenko, 2003 Corollary 1) *Let Γ be the unit circle of $\mathbb{R}^q$, $\Gamma = \{\lambda \in \mathbb{R}^q,\ \|\lambda\|_{2,q} = 1\}$. Let $(Z_i)_{1\le i\le n}$ and $(Y_i)_{1\le i\le n}$ be i.i.d. centered random vectors in $\mathbb{R}^q$ with $(Z_i)_{1\le i\le n}$ independent of $(Y_i)_{1\le i\le n}$. We denote $S_n^2 = \frac{1}{n}\sum_i^n Z_i Z_i'$, $S^2 = \mathbb{E}(Z_1 Z_1')$ and, for all random vector W : $S_{n,W}^2 = \frac{1}{n}\sum_i^n W_i W_i'$.*
If there exists $D > 0$ and $d > 0$ such that, for all $u \ge 0$,

$$\Pr\left(\sup_{\lambda\in\Gamma}\left(\frac{\sqrt{n}\lambda'(\overline{Z}_n - \overline{Y}_n)}{\sqrt{\lambda' S_{n,(Z-Y)}^2\lambda}}\right) \ge \sqrt{u}\right) \le De^{-du},$$

then, for all $u \ge 0$,

$$\Pr\left(\sup_{\lambda\in\Gamma}\frac{\sqrt{n}\lambda'\overline{Z}_n}{\sqrt{\lambda' S_n^2\lambda + \lambda' S^2\lambda}} \ge \sqrt{u}\right) \le De^{1-du}. \tag{3.11}$$

Proof : This proof is an extension of (Panchenko, 2003)'s Lemma 1 of Panchenko to the mutidimensional case. Denote

$$\begin{aligned} A_n(Z) &= \sup_{\lambda\in\Gamma}\sup_{b>0}\left\{\mathbb{E}_Y\left[4b(\lambda'(\overline{Z}_n - \overline{Y}_n) - b\lambda' S_{n,Z-Y}^2\lambda)|Z\right]\right\} \\ C_n(Z,Y) &= \sup_{\lambda\in\Gamma}\sup_{b>0}\left\{4b(\lambda'(\overline{Z}_n - \overline{Y}_n) - b\lambda' S_{n,Z-Y}^2\lambda)\right\}. \end{aligned}$$

By Jensen inequality, we have Pr-almost surely

$$A_n(Z) \le \mathbb{E}_Y[C_n(Z,Y)|Z]$$

and, for any convex function Φ, by Jensen inequality, we also get

$$\Phi(A_n(Z)) \le \mathbb{E}_Y[\Phi(C_n(Z,Y))|Z].$$

We obtain

$$\mathbb{E}_Z(\Phi(A_n(Z))) \le \mathbb{E}(\Phi(C_n(Z,Y))). \tag{3.12}$$

Now remark that

$$\begin{aligned} A_n(Z) &= \sup_{\lambda\in\Gamma}\sup_{b>0}\left\{4b\left(\lambda'\overline{Z}_n - b\lambda' S_n^2\lambda - b\lambda' S^2\lambda\right)\right\} \\ &= \sup_{\lambda\in\Gamma}\frac{\lambda'\overline{Z}_n}{\sqrt{\lambda' S_n^2\lambda + \lambda' S^2\lambda}} \end{aligned}$$

and

$$C_n(Z,Y) = \sup_{\lambda\in\Gamma} \frac{\lambda'(\overline{Z}_n - \overline{Y}_n)}{\sqrt{\lambda' S^2_{n,Z-Y}\lambda}}.$$

Now, notice that $\sup_{\lambda\in\Gamma} \frac{\lambda'\overline{Z}_n}{\sqrt{\lambda' S_n^2 \lambda}} > 0$ and apply the same arguments as Corollary 1's proof of Panchenko (2003) applied to inequality (3.12) to obtain the result. ■

Lemma 3.2 (Extension of Panchenko, 2003 Lemma 1) *Let ν and ξ, 2 r.v., such that, $\mathbb{E}(\nu) \leq \mathbb{E}(\xi)$ and, for $t > 0$,*

$$\Pr(\nu > t) \leq C\overline{F}_q(t)$$

then for $t \geq 2q$ we have

$$\Pr(\xi > t) \leq C\left(\frac{(t-q)}{2}\right)^{\frac{q}{2}} \frac{e^{-\frac{(t-q)}{2}}}{\Gamma(q/2+1)}.$$

Proof : We follow the lines of the proof of Panchenko's Lemma, with function Φ given by $\Phi(x) = \max(x - t + q; 0)$. Remark that $\Phi(0) = 0$ and $\Phi(t) = q$, then we have

$$\begin{aligned}\Pr(\xi \geq t) &\leq \frac{1}{\Phi(t)}\left(\Phi(0) + \int_0^{+\infty} \Phi'(x)\Pr(\nu \geq x)dx\right) \\ &\leq \frac{1}{q}\int_{t-q}^{+\infty} \overline{F}_q(x)dx.\end{aligned}$$

By integration by parts, we have

$$\int_{t-q}^{+\infty} \overline{F}_q(x)dx = \int_{t-q}^{+\infty} x f_q(x)dx - (t-q)\int_{t-q}^{+\infty} f_q(x)dx.$$

It follows by straightforward calculations (using $\Gamma(p) = (p-1)\Gamma(p-1)$) that for $t \geq 2q$,

$$\begin{aligned}\Pr(\xi \geq t) &\leq \frac{1}{q}\int_{t-q}^{+\infty} \overline{F}_q(x)dx = \overline{F}_{q+2}(t-q) - \frac{t-q}{q}\overline{F}_q(t-q) \\ &\leq \overline{F}_{q+2}(t-q) - \overline{F}_q(t-q) = \left(\frac{(t-q)}{2}\right)^{q/2} \frac{e^{-\frac{(t-q)}{2}}}{\Gamma(\frac{q}{2}+1)}.\end{aligned}$$

The last equality follows from using the recurrence relation 26.4.8 of Abramovitch & Stegun (1970, page 941). ■

We now extend a result of Barbe & Bertail (2004), which controls the behavior of the smallest eigenvalue of the empirical variance. In the following, for a given symmetric matrix A, we denote $\mu_1(A)$ its smallest eigenvalue.

Lemma 3.3 *Let $(Z_i)_{i=1,\cdots n}$ be i.i.d. random vectors in $\mathbb{R}^q$ with common mean 0. Denote $S^2 = \mathbb{E}(Z_1 Z_1')$ and $S_n^2 = \frac{1}{n}\sum_{i=1}^n Z_i Z_i'$, $0 < m_4 = \mathbb{E}(\|Z_1\|_2^4) < +\infty$ and $\widetilde{q} = \frac{q-1}{q+1}$. Then, for any $1 \leq q < n$ and $0 < u \leq \mu_1(S^2)$,*

$$\Pr\left(\mu_1(S_n^2) \leq u\right) \leq C(q)\,\frac{n^{3\widetilde{q}}\mu_1(S^2)^{2\widetilde{q}}}{m_4^{\widetilde{q}}} e^{-\frac{n(\mu_1(S^2)-u)^2}{m_4(q+1)}} \wedge 1,$$

with

$$C(q) = \pi^{2\widetilde{q}}(q+1)e^{2\widetilde{q}}(q-1)^{-3\widetilde{q}}2^{2\widetilde{q}-\frac{2}{q+1}} \tag{3.13}$$
$$\leq 4\pi^2(q+1)e^2(q-1)^{-3\widetilde{q}}. \tag{3.14}$$

Proof : This proof is adapted from the proof of Barbe & Bertail (2004) and makes use of some idea of Bercu et al. (2002). In the following, we denote by $\mathcal{S}_{q-1}$ the northern hemisphere of the sphere.

We first have by a truncation argument and applying Markov's inequality on the last term in the inequality (see the proof of (Barbe & Bertail, 2004) Lemma 4), for every $M > 0$, is less than

$$\Pr\left(\mu_1\left(\sum_{i=1}^n Z_i Z_i'\right) \leq t\right) \leq$$
$$\Pr\left(\inf_{v\in\mathcal{S}_{q-1}} \sum_{i=1}^n (v'Z_i)^2 \leq t, \sup_{i=1,\ldots,n} ||Z_i||_2 \leq M\right) + n\frac{m_4}{M^4} \tag{3.15}$$

We call I the first term on the right hand side of this inequality.
Notice that by symmetry of the sphere, we can always work with the northern hemisphere of the sphere rather than the sphere. Notice first, that, if $\sup_{i=1,\ldots,n} ||Z_i||_2 \leq M$, then for u, v in $\mathcal{S}_{q-1}$, we have

$$\left|\sum_{i=1}^n (v'Z_i)^2 - \sum_{i=1}^n (u'Z_i)^2\right| \leq 2n||u-v||M^2.$$

Thus if u and v are apart of $t\eta/(2nM^2)$ then $|\sum_{i=1}^n (v'Z_i)^2 - \sum_{i=1}^n (u'Z_i)^2| \leq \eta t$. Now let $N(\mathcal{S}_{q-1}, \varepsilon)$ be the smallest number of caps of radius ε centered at some points on $\mathcal{S}_{q-1}$ (for the $||.||_2$ norm) needed to cover $\mathcal{S}_{q-1}$ (the half sphere). Following the same arguments as Barbe & Bertail (2004), we have, for any $\eta > 0$,

$$I \leq N\left(\mathcal{S}_{q-1}, \frac{t\eta}{2nM^2}\right) \max_{u\in S_{q-1}} \Pr\left(\sum_{i=1}^n (u'Z_i)^2 \leq (1+\eta)t\right).$$

The proof is now divided in three steps, i) control of $N(\mathcal{S}_{q-1}, \frac{t\eta}{2nM^2})$ ii) control of the maximum over $\mathcal{S}_{q-1}$ of the last expression in I, iii) optimization over all the free parameters.
i) On the one hand, we have

$$N(\mathcal{S}_{q-1}, \varepsilon) \leq b(q)\varepsilon^{-(q-1)} \vee 1, \tag{3.16}$$

with, for instance, $b(q) \leq \pi^{q-1}$. Indeed, following Barbe & Bertail (2004), the northern hemisphere can be parameterized in polar coordinates, realizing a diffeomorphism with $S^{q-2}\times[0,\pi]$. Now proceed by induction, notice that for $q = 2$, $\mathcal{S}_{q-1}$, the half circle can be covered by $[\pi/2\varepsilon] \vee 1 + 1 \leq 2([\pi/2\varepsilon] \vee 1) \leq \pi/\varepsilon \vee 1$ caps of diameter 2ε, that is, we can choose the caps with their center on a $\varepsilon-$grid on the circle. Now, by induction we can cover the cylinder $S^{q-2}\times[0,\pi]$ with $[\pi/2\varepsilon\ (\pi)^{q-2}/\varepsilon^{q-2}] \vee 1 + 1 \leq \pi^{q-1}/\varepsilon^{q-1}$ intersecting cylinders which in turn can be mapped to region belonging to caps of radius ε, covering the whole sphere (this is

still a covering because the mapping from the cylinder to the sphere is contractive).
ii) On the other hand, for all $t > 0$, we have by exponentiation and Markov's inequality, and independence of (Z_i), for any $\lambda > 0$

$$\max_{u\in\mathcal{S}_{q-1}} \Pr\left(\sum_{i=1}^{n} u'Z_iZ_i'u \le t\right) \le e^{\lambda t} \max_{u\in\mathcal{S}_{q-1}} \left(\mathbb{E}\left[e^{-\lambda u'Z_1Z_1'u}\right]\right)^n.$$

Now, using the classical inequalities, $\log(x) \le x-1$ and $e^{-x}-1 \le -x+x^2/2$, both valid for $x > 0$, we have

$$\begin{aligned}\max_{u\in\mathcal{S}_{q-1}} \left(\mathbb{E}\left[e^{-\lambda u'Z_1Z_1'u}\right]\right)^n &\le \max_{u\in\mathcal{S}_{q-1}} \exp n\left(\mathbb{E}\left[e^{-\lambda u'Z_1Z_1'u}-1\right]\right)\\ &\le \max_{u\in\mathcal{S}_{q-1}} \exp n\left(-\lambda u'S^2u+\frac{\lambda^2}{2}m_4\right)\\ &= \exp\left(\frac{\lambda^2}{2}nm_4-\lambda n\mu_1(S^2)\right). \qquad (3.17)\end{aligned}$$

iii) From (3.17) and (3.16), we deduce that, for any $t > 0, \lambda > 0,\ \eta > 0$,

$$I \le b(q)\left(\frac{2nM^2}{t\eta}\right)^{q-1} e^{\lambda(1+\eta)t+\frac{\lambda^2}{2}nm_4-\lambda n\mu_1(S^2)}.$$

Optimizing the expression $\exp(-(q-1)\log(\eta)+\lambda\eta t)$ in $\eta > 0$, yields immediately, for any $t > 0$, any $M > 0$, any $\lambda > 0$

$$I \le b(q)\left(\frac{2enM^2\lambda}{q-1}\right)^{q-1} e^{\lambda\left(t-n\mu_1(S^2)\right)+n\lambda^2m_4/2}.$$

The infimum in λ in the exponential term is attained at $\lambda = \frac{\mu_1(S^2)-\frac{t}{n}}{m_4}$, provided that $0 < t < n\mu_1(S^2)$. Therefore, for these t and all $M > 0$, we get $\Pr(\mu_1(\sum_{i=1}^n Z_iZ_i') \le t)$ is less than

$$b(q)\left(\frac{2enM^2\mu_1(S^2)}{m_4(q-1)}\right)^{q-1} \exp\left(-\frac{n}{2m_4}\left(\mu_1(S^2)-\frac{t}{n}\right)^2\right)+n\frac{m_4}{M^4}.$$

We now optimize in $M^2 > 0$ and the optimum is attained at

$$M_*^2 = \left(\frac{2nm_4}{(q-1)b(q)}\right)^{\frac{1}{q+1}}\left(\frac{2en}{q-1}\frac{\mu_1(S^2)}{m_4}\right)^{-\frac{(q-1)}{q+1}} \exp\left(\frac{n(\mu_1(S^2)-\frac{t}{n})^2}{2m_4(q+1)}\right),$$

yielding the bound

$$\Pr\left(\mu_1\left(\sum_{i=1}^{n} Z_iZ_i'\right) \le t\right) \le \tilde{C}(q)\, n^{3\frac{q-1}{q+1}}\mu_1(S^2)^{\frac{2(q-1)}{q+1}}m_4^{-\frac{q-1}{q+1}} \exp\left(-\frac{n\left(\mu_1(S^2)-\frac{t}{n}\right)^2}{m_4(q+1)}\right),$$

with

$$\tilde{C}(q) = b(q)^{\frac{2}{q+1}}(q+1)e^{\frac{2(q-1)}{q+1}}(q-1)^{-3\frac{q-1}{q+1}}2^{\frac{2q-4}{q+1}}.$$

Using $b(q) \le \pi^{q-1}$ we obtained $C(q)$, which is bounded by the simpler bound (for large q this bound will be sufficient) $4\pi^2(q+1)e^2(q-1)^{-3\frac{q-1}{q+1}}$, using the fact that $m_4 \ge 1$.
The result of the Lemma follows by applying this inequality on inequality 3.15 with $t = nu$.
■

3.6.3 Proof of Theorem 3.4

Notice that we have always $\bar{Z}_n' S_n^{-2} \bar{Z}_n \leq q$. Indeed, there exists an orthogonal transformation O_n and a diagonal matrix $\Lambda_n^2 := diag[\hat{\mu}_j]_{1\leq j\leq q}$ with $\hat{\mu}_j > 0$ being the eigenvalues of S_n^2, such that $S_n^2 = O_n' \Lambda_n^2 O_n$. Now put $Y_{i,n} := [Y_{i,j,n}]_{1\leq j\leq q} = O_n Z_i$. It is easy to see that by construction the empirical variance of the $Y_{i,n}$ is

$$\frac{1}{n}\sum_{i=1}^{n} Y_{i,n} Y_{i,n}' = \frac{1}{n}\sum_{i=1}^{n} O_n Z_i Z_i' O_n' = O_n S_n^2 O_n' = \Lambda_n^2.$$

It also follows from this equality that, for all $j = 1, \cdots, q$, $\frac{1}{n}\sum_{i=1}^{n} Y_{i,j,n}^2 = \hat{\mu}_j$, and

$$\bar{Z}_n' S_n^{-2} \bar{Z}_n = \bar{Y}_n' \Lambda_n^{-2} \bar{Y}_n = \sum_{j=1}^{q} \left(\frac{1}{n}\sum_{i=1}^{n} Y_{i,j,n} \right)^2 / \hat{\mu}_j \leq q.$$

by Cauchy-Schwartz. So, for all $u > qn$

$$\Pr\left(\bar{Z}_n' S_n^{-2} \bar{Z}_n \geq \frac{u}{n} \right) = 0.$$

a) In the symmetric and unidimensional framework ($q = 1$), this bound follows from Hoeffding inequality (see (Efron, 1969, Edelman, 1986)). Consider now the symmetric multidimensional framework ($q > 1$). Let $\sigma_i, 1 \leq i \leq n$ be Rademacher random variables, independent from $(Z_i)_{1\leq i\leq n}$, $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = 1) = 1/2$. We denote $\sigma_n(Z) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sigma_i Z_i \right)$ and remark that $S_n^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_i Z_i Z_i' \sigma_i$. Since the Z_i's have a symmetric distribution, meaning that $-Z_i$ has the same distribution as Z_i, we can make a first symmetrization step :

$$\Pr\left(n \overline{Z}_n' S_n^{-2} \overline{Z}_n \geq u \right) = \Pr(\sigma_n(Z)' S_n^{-2} \sigma_n(Z) \geq u).$$

Now, we have

$$\begin{aligned} \sigma_n(Z)' S_n^{-2} \sigma_n(Z) &= \sigma_n(Y)' \Lambda_n^{-2} \sigma_n(Y) \\ &= \sum_{j=1}^{q} \left(\sum_{i=1}^{n} \sigma_i Y_{i,j,n} \right)^2 / \sum_{i=1}^{n} Y_{i,j,n}^2. \end{aligned}$$

It follows that

$$\begin{aligned} \Pr(\sigma_n(Z)' S_n^{-2} \sigma_n(Z) \geq u) &\leq \sum_{j=1}^{q} \Pr\left(\frac{|\sum_{i=1}^{n} \sigma_i Y_{i,j,n}|}{\sqrt{\sum_{i=1}^{n} Y_{i,j,n}^2}} \geq \sqrt{u/q} \right) \\ &\leq 2 \sum_{j=1}^{q} \mathbb{E} \Pr\left(\frac{\sum_{i=1}^{n} \sigma_i Y_{i,j,n}}{\sqrt{\sum_{i=1}^{n} Y_{i,j,n}^2}} \geq \sqrt{u/q} \,\middle|\, (Z_i)_{1\leq i\leq n} \right). \end{aligned}$$

Apply now Hoeffding inequality to each unidimensional self-normalized term in this sum to conclude.

b) The Z_i's are not anymore symmetric. Our first step is to control $\Pr(n\bar{Z}_n' S_n^{-2} \bar{Z}_n \geq t)$. Define

$$B_n = \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \overline{Z}_n \geq 0}} \left\{ \frac{\lambda' \overline{Z}_n}{\sqrt{\lambda' S_n^2 \lambda}} \right\} \text{ and } D_n = \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \overline{Z}_n \geq 0}} \left\{ \sqrt{1 + \frac{\lambda' S^2 \lambda}{\lambda' S_n^2 \lambda}} \right\}.$$

First of all, remark that the following events are equivalent

$$\left\{ n\overline{Z}_n' S_n^{-2} \overline{Z}_n \geq t \right\} = \left\{ B_n \geq \sqrt{\frac{t}{n}} \right\}.$$

and notice that

$$\Pr\left(B_n \geq \sqrt{\frac{t}{n}} \right) \leq \inf_{a>-1} \left\{ \Pr\left(B_n D_n^{-1} \geq \sqrt{\frac{t}{n(1+a)}} \right) + \Pr(D_n \geq \sqrt{1+a}) \right\}.$$

The control of the first term on the right side is obtained by applying part a) of Theorem 3.2 to $n^{1/2} \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \in \Gamma}} \frac{\lambda' \overline{Z}_n - \overline{Y}_n}{\sqrt{\lambda' S_n^{2,Z-Y} \tilde{\lambda}}}$. Then, by application of Lemma 3.1 and the previous remark, we get
$\sqrt{n} B_n D_n^{-1} \leq n^{1/2} \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \overline{Z}_n \geq 0}} \frac{\lambda' \overline{Z}_n}{\sqrt{\lambda' S_n^2 \tilde{\lambda} + \lambda' S^2 \tilde{\lambda}}}$, we have for all $t > 0$,

$$\Pr\left(B_n D_n^{-1} \geq \sqrt{\frac{t}{n}} \right) \leq 2qe^{1 - \frac{t}{2q}}.$$

The control of the second term is trivial and useless for $a \leq 0$. Whereas, for all $a > 0$, and all $t > 0$ we have

$$\begin{aligned} \left\{ D_n \geq \sqrt{a+1} \right\} &= \left\{ \sup_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \overline{Z}_n \geq 0}} \left(1 + \frac{\lambda' S^2 \lambda}{\lambda' S_n^2 \lambda} \right) \geq 1 + a \right\} \\ &= \left\{ \inf_{\substack{\|\lambda\|_{2,q}=1 \\ \lambda' \overline{Z}_n \geq 0}} \left(\lambda' S^{-1} S_n^2 S^{-1} \lambda \right) \leq \frac{1}{a} \right\} = \left\{ \mu_1(S^{-1} S_n^2 S^{-1}) \leq \frac{1}{a} \right\}. \end{aligned}$$

We now use Lemma 3.3 applied to the r.v.'s $(S^{-1} Z_i)_{i=1,\cdots,n}$. Note that here we have

$$m_4 = \mathbb{E}\|S^{-1} Z_1\|_2^4 = \gamma_4, \ \mathbb{E}(S^{-1} Z_1)^2 = Id_q, \ \mu_1(Id_q) = 1, \text{ and } u = \frac{1}{a}.$$

For all $1 < a$, we have,

$$\Pr(D_n > \sqrt{1+a}) \leq C(q) \left(\frac{n^3}{\gamma_4} \right)^{\tilde{q}} e^{-\frac{n}{(q+1)\gamma_4}(1-\frac{1}{a})^2}.$$

Since $\inf_{a>-1} \leq \inf_{a>1}$, we conclude that, for any $t > n$,

$$\Pr\left(B_n > \sqrt{\frac{t}{n}} \right) \leq \inf_{a>1} \left\{ 2qe \, e^{-\frac{t}{2q(1+a)}} + C(q) \left(\frac{n^3}{\gamma_4} \right)^{\tilde{q}} e^{-\frac{n}{(q+1)\gamma_4}(1-\frac{1}{a})^2} \right\}.$$

3.6.4 Proof of Theorem 3.5

Part a) is proved in Pinelis (1994). Now, the proof of part b) follows the same lines as the Theorem 3.4 combining Lemmas 3.1, 3.2 and 3.3.

3.6.5 Proof of corollary 3.6

Following the arguments of the remark of Theorem 3.2, we use the dual form and expand K_ε near 0. Then we get

$$\begin{aligned}\beta_n(\theta) &= \sup_{\lambda\in\mathbb{R}^q}\left\{-n\lambda'\overline{f}_n - \frac{1}{2}\sum_{i=1}^{n}(\lambda' f(X_i,\theta))^2 K_\varepsilon^{(2)}(t_{i,n})\right\}\\ &\leq \sup_{\lambda\in\mathbb{R}^q}\left\{-n\lambda'\overline{f}_n - \frac{1}{2}\sum_{i=1}^{n}(\lambda' f(X_i,\theta))^2 \varepsilon\right\}. \qquad (3.18)\end{aligned}$$

Indeed, by construction of the quasi-Kullback, we have $K_\varepsilon^{(2)} \geq \varepsilon$. If we write $l = -\varepsilon\lambda$, the right hand side of inequality (3.18) becomes

$$\frac{n}{\varepsilon}\sup_{l\in\mathbb{R}^q}\left\{l'\overline{f}_n - \frac{1}{2}l' S_n^2 l\right\} = \frac{n}{2\varepsilon}\overline{f}_n' S_n^{-2}\overline{f}_n.$$

Thus we immediately get

$$\Pr(\theta \notin \mathscr{C}_n(\eta)) \leq \Pr\left(\frac{n}{2}\overline{f}_n' S_n^{-2}\overline{f}_n \geq \eta\varepsilon\right).$$

Chapitre 4

Empirical likelihood and Markov chains

4.1 Introduction

4.1.1 Empirical likelihood for atomic Markov chains

Empirical Likelihood (EL), introduced by Owen (1988), is a powerful semi-parametric method. It can be used in a very general setting and leads to effective estimation, tests and confidence intervals. This method shares many good properties with the conventional parametric log-likelihood ratio : both statistics have χ^2 limiting distribution and are Bartlett correctable, meaning that the error can be reduced from $\mathcal{O}(n^{-1})$ to $\mathcal{O}(n^{-2})$ by a simple adjustment.

Owen's framework has been intensively studied in the 90's (see Owen, 2001, for an overview), leading to many generalizations and applications, but mainly for an i.i.d. setting. The case of weakly dependent processes has been studied in Kitamura (1997) under the name of *Block Empirical Likelihood* (BEL). This work is inspired by similitudes with the bootstrap methodology. Kitamura proposed to apply the empirical likelihood framework not directly on the data but on blocks of consecutive data, to catch the dependence structure. This idea, known as *Block Bootstrap* (BB) or blocking technique (in the probabilistic literature, see Doukhan & Ango Nze, 2004, for references) goes back to Kunsch (1989) in the bootstrap literature and has been intensively exploited in this fields (see Lahiri, 2003, for a survey). However, the BB performances have been questioned, see Götze & Kunsch (1996) and Horowitz (2003). Indeed it is known that the blocking technique distorts the dependence structure of the data generating process and its performance strongly relies on the choice of the block size. From a theoretical point of view, the assumptions used to prove the validity of the BB and of Kitamura's BEL are generally strong : one generally assumes that the time series is stationary and satisfies some strong-mixing properties. In addition, to have a precise control of the coverage probability of the confidence intervals, one has to assume that the strong mixing coefficients are exponentially decreasing (see Lahiri, 2003, Kitamura, 1997). Moreover the choice of the tuning parameter (the block size) may be quite difficult from a practical point of view.

In this paper, we focus on generalizing empirical likelihood to Markov chains. Questioning the restriction implied by the markovian setting is a natural issue. It should be mentioned that homogeneous Markov chain models cover a huge number of time-series models. In particular, a Markov chain can always be written in a non parametric way : $X_i = h(X_{i-1}, \cdots, X_{i-p}, \varepsilon_i)$, where $(\varepsilon_i)_{i\geq 0}$ is i.i.d. with density f and ε_i is independent of $(X_k)_{0\leq k<i}$, see Kallenberg (2002). Note that both h and f are unknown functions. Such representation explains why, provided that p is large enough, any time series of length n can be generated by a Markov chain, see Knight (1975). Note also that a Markov chain may not be necessarily strong-mixing, so that our method also covers cases for which BB and BEL may fail. For instance, the simple linear model $X_i = \frac{1}{2}(X_{i-1} + \varepsilon_i)$ with $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = 0) = \frac{1}{2}$ is not strong-mixing (see Doukhan & Ango Nze, 2004, for results on dependence in Econometrics).

Our approach is also inspired by some recent developments in the bootstrap literature on Markov chains : instead of choosing blocks of constant length, we use the Markov chain structure to choose some adequate cutting times and then we obtain blocks of various lengths. This construction, introduced in Bertail & Clémençon (2004a), catches better the structure

of the dependence. It is originally based on the existence of an atom for the chain i.e. an accessible set on which the transition kernel is constant (see Meyn & Tweedie, 1996). The existence of an atom allows to cut the chain into regeneration blocks, separated from each other by a visit to the atom. These blocks (of random lengths) are independent by the strong Markov property. Once these blocks are obtained, the *Regenerative Block-Bootstrap (RBB)* consists in resampling the data blocks to build new regenerative processes. The rate obtained by resampling these blocks ($\mathcal{O}(n^{-1+\varepsilon})$) is better than the one obtained for the Block Bootstrap ($\mathcal{O}(n^{-3/4})$) and is close to the classical rate $\mathcal{O}(n^{-1})$ obtained in the i.i.d. case, see Götze & Kunsch (1996) and Lahiri (2003).

These improvements suggest that a version of the empirical likelihood (EL) method based on such blocks could yield improved results in comparison to the method presented in Kitamura (1997). Indeed it is known that EL enjoys somehow the same properties in term of accuracy as the bootstrap but without any Monte-Carlo step. The main idea is to consider the renewal blocks as independent observations and to follow the empirical likelihood method. Such program is made possible by transforming the original problem based on moments under the stationary distribution into an equivalent problem under the distribution of the observable blocks (via Kac's Theorem). The advantages of the method proposed in this paper are at least twofold : first the construction of the blocks is automatic and entirely determined by the data : it leads to a unique version of the empirical likelihood program. Second there is not need to ensure stationarity nor any strong mixing condition to obtain a better coverage probability for the corresponding confidence regions.

4.1.2 Empirical discrepancies for Harris chains

Assuming that the chain is atomic is a strong restriction to this method. This hypothesis essentially holds for discrete Markov chains and queuing (or storage) systems returning to a stable state (for instance the empty queue) : see chapter 2.4 of Meyn & Tweedie (1996). However this method can be extended to the more general case of Harris chains. Indeed any chain having some recurrent properties can be extended to a chain possessing an atom which then enjoys some regenerative properties. Nummelin gives an explicit construction of such extension that we recall in section 4.4 (see Nummelin, 1978, Athreya & Ney, 1978)). In Bertail & Clémençon (2006a) an extension of the RBB procedure to general Harris chains based on the Nummelin' splitting technique is proposed (*the Approximate Regenerative Block-Bootstrap, ARBB*). One purpose of this paper is to prove that these approximatively regenerative blocks can also be used in the framework of empirical likelihood and lead to consistent results.

Empirical likelihood can be seen as a contrast method based on the Kullback discrepancy. To replace the Kullback discrepancy by some other discrepancy is an interesting problem which has led to some recent works in the i.i.d. case. Newey & Smith (2004) generalized empirical likelihood to the family of Cressie-Read discrepancies (see also Guggenberger & Smith, 2005). The resulting methodology, *Generalized Empirical Likelihood*, is included in the empirical φ-discrepancy method introduced by Bertail et al. (2005) (see also Bertail et al., 2004, Kitamura, 2006). It should be noticed that the constant length blocks procedure has been studied in the case of empirical euclidean likelihood by Lin & Zhang (2001). Our proposal is straightforwardly compatible with these generalizations, but we will not pursue this approach here.

4.1.3 Outline

The outline of the paper is the following. In section 4.2, notations are set out and key concepts of the Markov atomic chain theory are recalled. In section 4.3, we present how to construct regenerative data blocks and confidence regions based on these blocks. In section 4.4 the Nummelin splitting technique is shortly recalled and a framework to adapt the regenerative empirical likelihood method to general Harris chains is proposed. We essentially obtain consistent results but also briefly discuss higher order properties. In section 4.5, we propose some moderate sample size simulations.

4.2 Preliminary statement

4.2.1 Notation and definitions

For simplicity's sake we will keep essentially the same notations as Bertail & Clémençon (2006b). For further details and traditional properties of Markov chains, we refer to Revuz (1984) or Meyn & Tweedie (1996). We consider a space E (to simplify $\mathbb{R}^d$, $\mathbb{Z}^d$, or a subset of these spaces) endowed with a σ-algebra $\mathcal{E}$. Recall first that a chain is ψ-irreducible if for any starting state x in $\mathcal{E}$, the chain visits A with probability 1, as soon as $\psi(A) > 0$. This means that the chain visits all sets of positive ψ-measure. ψ may be seen as a dominating measure. To prevent one from unusual behavior of the chain, we consider in the following a chain $X = (X_i)_{i\in\mathbb{N}}$ which is aperiodic (it will not be cyclic) and ψ-irreducible. Let Π be the transition probability, and ν the initial probability distribution. For a set $B \in \mathcal{E}$ and $i \in \mathbb{N}$, we thus denote

$$X_0 \sim \nu \text{ and } \mathbb{P}(X_i \in B \mid X_0, ..., X_{i-1}) = \Pi(X_{i-1}, B) \text{ a.s. .}$$

In what follows, $\mathbb{P}_\nu$ and $\mathbb{P}_x$ (for x in E) denote the probability measure on the underlying probability space such that $X_0 \sim \nu$ and $X_0 = x$ respectively. $\mathbb{E}_\nu(\cdot)$ is the $\mathbb{P}_\nu$-expectation, $\mathbb{E}_x(\cdot)$ the $\mathbb{P}_x$-expectation, $\mathbb{1}_\mathcal{A}$ denotes the indicator function of the event $\mathcal{A}$ and $\mathbb{E}_\mathcal{A}(\cdot)$ is the expectation conditionally on $X_0 \in \mathcal{A}$.

A measurable set B is *recurrent* if, as soon as the chain hits the set B, the chain returns infinitely often to B. The chain is said *Harris recurrent* if it is ψ-irreducible and every measurable set with positive ψ-measure is recurrent. A probability measure μ on E is said invariant for the chain when $\mu\Pi = \mu$, where

$$\mu\Pi(dy) = \int_{x\in E} \mu(dx)\Pi(x, dy) .$$

An irreducible chain is said *positive recurrent* when it admits an invariant probability (it is then unique).

Notice that as defined, a Markov chain is generally non-stationary (if $\nu \neq \mu$) and may not be strong-mixing. The fully non-stationary case corresponding to the null recurrent case (including processes with unit roots) could actually be treated by using the arguments of Tjostheim (1990), but would considerably complicate the exposition and the notations.

4.2.2 Markov chains with an atom

Assume that the chain is ψ-irreducible and possesses an accessible atom, that is to say a set A , with $\psi(A) > 0$ such that the transition probability is constant on A ($\Pi(x, .) = \Pi(y, .)$ for all x, y in A). The class of atomic Markov chains contains not only chains defined on a countable state space but also many specific Markov models used to study queuing systems and stock models (see Asmussen, 1987, for models involved in queuing theory). In the discrete case, any recurrent state is an accessible atom : the choice of the atom is thus left to the statistician which can for instance use the mostly visited point. In many other situations the atom is determined by the structure of the model (for a random walk on $\mathbb{R}^+$, with continuous increment, 0 is the only possible atom).

Denote by $\tau_A = \tau_A(1) = \inf\{k \geq 1,\ X_k \in A\}$ the hitting time of the atom A (the first visit) and, for $j \geq 2$, denote by $\tau_A(j) = \inf\{k > \tau_A(j-1),\ X_k \in A\}$ the successive return times to A. The sequence $(\tau_A(j))_{j\geq 1}$ defines the successive times at which the chain forgets its past, called *regeneration times*. Indeed, the transition probability being constant on the atom, X_{τ_A+1} only depends on the information that X_{τ_A} is in A and not any more on the actual value of X_{τ_A} itself.

For any initial distribution ν, the sample path of the chain may be divided into blocks of random length corresponding to consecutive visits to A :

$$B_j = (X_{\tau_A(j)+1}, ...,\ X_{\tau_A(j+1)}).$$

The sequence of blocks $(B_j)_{1\leq j<\infty}$ is then i.i.d. by the strong Markov property (see Meyn & Tweedie, 1996). Notice that the block $B_0 = (X_1, ...,\ X_{\tau_A})$ is independent of the other blocks, but not with the same distribution, because its distribution strongly depends on the initial distribution ν.

Let $m : E\times\mathbb{R}^p \to \mathbb{R}^r$ be a measurable function and θ_0 be the true value of some parameter $\theta \in \mathbb{R}^p$ of the chain, given by an estimating equation on the invariant measure μ :

$$\mathbb{E}_\mu[m(X, \theta_0)] = 0. \tag{4.1}$$

For example, the parameter of interest can be a moment or the mean (in this case $\theta_0 = \mathbb{E}_\mu[X]$ and $m(X, \theta) = X - \theta$).

In this framework, Kac's Theorem, stated below (see Theorem 10.2.2 in Meyn & Tweedie, 1996) allows to write functionals of the stationary distribution μ as functionals of the distribution of a regenerative block.

Theorem 4.1 *The chain X is positive recurrent iff $\mathbb{E}_A(\tau_A) < \infty$. The (unique) invariant probability distribution μ is then the Pitman's occupation measure given by*

$$\mu(F) = \mathbb{E}_A\left[\sum_{i=1}^{\tau_A} \mathbb{1}_{X_i\in F}\right] / \mathbb{E}_A[\tau_A],\ \textit{for all } F \in \mathcal{E}.$$

In the following we denote

$$M(B_j, \theta) = \sum_{i=\tau_A(j)+1}^{\tau_A(j+1)} m(X_i, \theta)$$

so that we can rewrite the estimating equation (4.1) as :

$$\mathbb{E}_A[M(B_j, \theta_0)] = 0. \tag{4.2}$$

The power of Kac's Theorem and of the regenerative ideas is that the decomposition into independent blocks can be automatically used to obtain limit theorems for atomic chains. One may refer for example to Meyn & Tweedie (1996) for the Law of Large Numbers (LLN), Central Limit Theorem (CLT), Law of Iterated Logarithm, Bolthausen (1982) for the Berry-Esseen Theorem, Bertail & Clémençon (2004a) for Edgeworth expansions. These results are established under some hypotheses related to the distribution of the B_j's. Let $\kappa > 0$ and ν be a probability distribution on $(E, \mathcal{E})$. The following assumptions shall be involved throughout this article :

Return time conditions :

$$\begin{aligned} \mathbf{H0}(\kappa) &: \ \mathbb{E}_A[\tau_A^\kappa] < \infty, \\ \mathbf{H0}(\kappa,\ \nu) &: \ \mathbb{E}_\nu[\tau_A^\kappa] < \infty. \end{aligned}$$

When the chain is stationary and strong mixing, these hypotheses can be related to the rate of decay of α-mixing coefficients $\alpha(p)$, see Bolthausen (1982). In particular, the hypotheses are satisfied if $\sum_{j\geq 1} j^\kappa \alpha(j) < \infty$.

Block-moment conditions :

$$\begin{aligned} \mathbf{H1}(\kappa,\ m) &: \ \mathbb{E}_A\left[\left(\sum_{i=1}^{\tau_A} ||m(X_i, \theta_0)||\right)^\kappa\right] < \infty, \\ \mathbf{H1}(\kappa,\ \nu,\ m) &: \ \mathbb{E}_\nu\left[\left(\sum_{i=1}^{\tau_A} ||m(X_i, \theta_0)||\right)^\kappa\right] < \infty. \end{aligned}$$

Equivalence of these assumptions with easily checkable drift conditions may be found in Meyn & Tweedie (1996).

4.3 The regenerative case

4.3.1 Regenerative Block Empirical Likelihood algorithm

Let $X_1, \cdots, X_n$ be an observation of the chain X. If we assume that we know an atom A for the chain, the construction of the regenerative blocks is then trivial. Consider the empirical distribution of the blocks :

$$\mathbb{P}_{l_n} = \frac{1}{l_n} \sum_{j=1}^{l_n} \delta_{B_j},$$

where l_n is the number of complete regenerative blocks, and the multinomial distributions

$$\mathbb{Q} = \sum_{j=1}^{l_n} q_j \delta_{B_j}, \ \text{with } 0 < q_j < 1,$$

dominated by $\mathbb{P}_{l_n}$. To obtain a confidence region, we will apply Owen (1990)'s method to the blocks B_j using the likelihood associated to $\mathbb{Q}$, that is we are going to minimize the Kullback distance between $\mathbb{Q}$ and $\mathbb{P}_{l_n}$ under the condition (4.2). More precisely, the *Regenerative Block Empirical Likelihood* is defined in the next 4 steps :

Algorithm 4.1 (ReBEL - Regenerative Block Empirical Likelihood construction)

1. *Count the number of visits* $l_n + 1 = \sum_{i=1}^{n} \mathbb{1}_{X_i \in A}$ *to* A *up to time* n.
2. *Divide the observed trajectory* $X^{(n)} = (X_1, ..., X_n)$ *into* $l_n + 2$ *blocks corresponding to the pieces of the sample path between consecutive visits to the atom* A,

$$B_0 = (X_1, ..., \ X_{\tau_A(1)}), \ B_1 = (X_{\tau_A(1)+1}, ..., \ X_{\tau_A(2)}), ...,$$
$$B_{l_n} = (X_{\tau_A(l_n)+1}, ..., \ X_{\tau_A(l_n+1)}), \ B^{(n)}_{l_n+1} = (X_{\tau_A(l_n+1)+1}, ..., \ X_n),$$

   with the convention $B^{(n)}_{l_n+1} = \emptyset$ *when* $\tau_A(l_n + 1) = n$.
3. *Drop the first block* B_0 *and the last one* $B^{(n)}_{l_n+1}$ *(eventually empty when* $\tau_A(l_n + 1) = n$*).*
4. *Evaluate the empirical log-likelihood ratio* $r_n(\theta)$ *(practically on a grid of the set of interest) :*

$$r_n(\theta) = \sup_{(q_1,\cdots,q_{l_n})} \left\{ \log \left[\prod_{j=1}^{l_n} l_n q_j \right] \middle| \sum_{j=1}^{l_n} q_j \cdot M(B_j, \theta) = 0, \ \sum_{j=1}^{l_n} q_j = 1 \right\}.$$

   Using Lagrange arguments or convex duality, this can be more easily calculated as

$$r_n(\theta) = \sup_{\lambda \in \mathbb{R}^p} \left\{ \sum_{j=1}^{l_n} \log \left[1 + \lambda' M(B_j, \theta) \right] \right\}.$$

Note 4.1 (Small samples) *Eventually, if the chain does not visit* A, $l_n = -1$. *Of course the algorithm cannot be implemented and no confidence interval can be built. Actually, even when* $l_n \geq 0$, *the algorithm can be meaningless and at least a reasonable number of blocks are needed to build a confidence interval. In the positive recurrent case, it is known that* $l_n \sim n/\mathbb{E}_A[\tau_A]$ *a.s. and the length of each block has expectation* $\mathbb{E}_A[\tau_A]$. *Many regenerations of the chain should then be observed as soon as* n *is significantly larger than* $\mathbb{E}_A[\tau_A]$. *Of course, the next results are asymptotic, for finite sample consideration on empirical likelihood methods (in the i.i.d. setting), refer to Bertail et al. (2005).*

The next theorem states the asymptotic validity of ReBEL in the case $r = p$ (just-identified case). For this, we introduce the ReBEL confidence region defined as follows :

$$C_{n,\alpha} = \left\{ \theta \in \mathbb{R}^p \middle| \ 2 \cdot r_n(\theta) \leq F_{\chi^2_p}(1 - \alpha) \right\},$$

where $F_{\chi^2_p}$ is the distribution function of a χ^2 distribution with p degrees of freedom.

Theorem 4.2 *Let μ be the invariant measure of the chain, let $\theta_0 \in \mathbb{R}^p$ be the parameter of interest, satisfying $\mathbb{E}_\mu[m(X,\theta_0)] = 0$. Assume that $\mathbb{E}_A[M(B,\theta_0)M(B,\theta_0)']$ is of full-rank. Assume **H0**$(1,\nu)$, **H0**(2) and **H1**$(2,m)$, then*

$$2r_n(\theta_0) \xrightarrow{n\to\infty} \chi^2_p$$

and therefore

$$\mathbb{P}_\nu\left(\theta_0 \in C_{n,\alpha}\right) \to 1-\alpha.$$

The proof relies on the same arguments as the one for empirical likelihood based on i.i.d. data. This can be easily understood : our data, the regenerative blocks, are i.i.d. (see Owen, 1990, 2001). The only difference with the classical use of empirical likelihood is that the length of the data (i.e. the number of blocks) is a random value l_n. However, we have that $n/l_n \to \mathbb{E}_A(\tau_A)$ a.s. (see Meyn & Tweedie, 1996). The proof is given in the appendix.

Note 4.2 *Let's make some very brief discussion on the rate of convergence of this method. Bertail & Clémençon (2004a) shows that the Edgeworth expansion of the mean standardized by the empirical variance holds up to $\mathcal{O}_{\mathbb{P}_\nu}(n^{-1})$ (in opposition to what is expected when considering a variance built on fixed length blocks). It follows from their result that*

$$\mathbb{P}_\nu\left(2\cdot r_n(\theta_0) \le u\right) = F_{\chi^2_p}(u) + \mathcal{O}_{\mathbb{P}_\nu}(n^{-1})$$

This is already (without Bartlett correction) better than the Bartlett corrected empirical likelihood when one use fixed length blocks (see Kitamura, 1997). Actually, we expect, in this atomic framework, that a Bartlett correction would lead to the same result as in the i.i.d. case : $\mathcal{O}(n^{-2})$. However, to prove this conjecture, one should establish an Edgeworth expansion for the likelihood ratio (which can be derived from expansion for self normalized sums) up to order $\mathcal{O}(n^{-2})$ which is a very technical task. This is left for further works.

4.3.2 Estimation and the over-identified case

The properties of empirical likelihood proved by Qin & Lawless (1994) can be extended to our markovian setting. In order to state the corresponding results respectively on estimation, confidence region under over-identification ($r \ge p$) and hypotheses testing, we introduce the following additional assumptions. Assume that there exists a neighborhood V of θ_0 and a function N with $\mathbb{E}_\mu\left[N(X)\right] < \infty$, such that :

H2(a) $\partial m(x,\theta)/\partial\theta$ is continuous in θ and bounded in norm by $N(x)$ for θ in V.

H2(b) $\mathbb{E}_\mu[\partial m(X,\theta_0)/\partial\theta]$ is of full rank.

H2(c) $\partial^2 m(x,\theta)/\partial\theta\partial\theta'$ is continuous in θ and its norm can be bounded by $N(x)$, for θ in V.

H2(d) $\|m(x,\theta)\|^3$ is bounded by $N(x)$ on V.

Notice that **H2**(d) implies in particular the block moment condition **H1**$(3,m)$ since by Kac's Theorem

$$\mathbb{E}_\mu\left[\|m(X,\theta)\|^3\right] = \frac{\mathbb{E}_A\left[\sum_{i=1}^{\tau_A}\|m(X_i,\theta)\|^3\right]}{\mathbb{E}_A[\tau_A]} \le \frac{\mathbb{E}_A\left[\sum_{i=1}^{\tau_A} N(X_i)\right]}{\mathbb{E}_A[\tau_A]} \le \mathbb{E}_\mu\left[N(X)\right] < \infty.$$

Empirical likelihood provides a natural way to estimate θ_0 in the i.i.d. case (see Qin & Lawless, 1994). This can be straightforwardly extended to Markov chains. The estimator is the maximum empirical likelihood estimator defined by

$$\tilde{\theta}_n = \arg \inf_{\theta \in \Theta} \{r_n(\theta)\}.$$

The next theorem shows that, under natural assumptions on m and μ, $\tilde{\theta}_n$ is an asymptotically gaussian estimator of θ_0.

Theorem 4.3 *Assume that the hypotheses of Theorem 4.2 holds. Under the additional assumptions **H2**(a), **H2**(b) and **H2**(d), $\tilde{\theta}_n$ is a consistent estimator of θ_0. If in addition **H2**(c) holds, then $\tilde{\theta}_n$ is asymptotically gaussian :*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

with

$$\Sigma = \mathbb{E}_A[\tau_A] \left[\mathbb{E}_A \left[\frac{\partial M(B_1, \theta)}{\partial \theta} \right]' \mathbb{E}_A[M(B_1, \theta) M'(B_1, \theta)]^{-1} \mathbb{E}_A \left[\frac{\partial M(B_1, \theta)}{\partial \theta} \right] \right]^{-1}$$

Notice that all the terms involved in the expression of Σ can be easily estimated by replacing them by empirical sums over blocks. The corresponding estimator is straightforwardly asymptotically convergent by the LLN for Markov chains.

The case of over-identification ($r > p$) is an important feature, specially for econometric applications. In such a case, the statistic $2r_n(\tilde{\theta}_n)$ may be considered to test the moment equation (4.1) :

Theorem 4.4 *Under the assumptions of Theorem 4.3, if the moment equation (4.1) holds, then we have*

$$2r_n(\tilde{\theta}_n) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_{r-p}.$$

We now turn to a theorem equivalent to the Theorem 2 in the over-identified case. Then, the likelihood ratio statistic used to test $\theta = \theta_0$ must be corrected. We now define

$$W_{1,n}(\theta) = 2r_n(\theta) - 2r_n(\tilde{\theta}_n).$$

The ReBEL confidence region of nominal level $1 - \alpha$ in the over-identified case is now given by

$$C^1_{n,\alpha} = \left\{ \theta \in \mathbb{R}^p \,\middle|\, W_{1,n}(\theta) \leq F_{\chi^2_p}(1 - \alpha) \right\}.$$

Theorem 4.5 *Under the assumptions of Theorem 4.3, the likelihood ratio statistic for $\theta = \theta_0$ is asymptotically χ^2_p :*

$$W_{1,n}(\theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_p$$

and $C^1_{n,\alpha}$ is then an asymptotic confidence region of nominal level $1 - \alpha$.

To test a sub-vector of the parameter, one can also build the corresponding empirical likelihood ratio (see Qin & Lawless, 1994, Kitamura, 1997, Kitamura et al., 2004, Guggenberger & Smith, 2005). Let $\theta' = (\theta_1, \theta_2)'$ be in $\mathbb{R}^q \times \mathbb{R}^{p-q}$, where $\theta_1 \in \mathbb{R}^q$ is the parameter of interest and $\theta_2 \in \mathbb{R}^{p-q}$ is a nuisance parameter. Assume that the true value of the parameter of interest is θ_{10}. The empirical likelihood ratio statistic in this case becomes

$$W_{2,n}(\theta_1) = 2 \cdot \left(\inf_{\theta_2} r_n((\theta_1, \theta_2)') - \inf_{\theta} r_n(\theta) \right) = 2 \cdot \left(\inf_{\theta_2} r_n((\theta_1, \theta_2)') - r_n(\tilde{\theta}_n) \right),$$

and the empirical likelihood confidence region is given by

$$C^2_{n,\alpha} = \left\{ \theta_1 \in \mathbb{R}^q \,\middle|\, W_{2,n}(\theta_1) \leq F_{\chi^2_q}(1-\alpha) \right\}.$$

Theorem 4.6 *Under the assumptions of Theorem 4.3,*

$$W_{2,n}(\theta_{10}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_q$$

and $C^2_{n,\alpha}$ is then an asymptotic confidence region of nominal level $1-\alpha$.

4.4 The case of general Harris chains

4.4.1 Algorithm

As explained in the introduction, the splitting technique introduced in Nummelin (1978) allows to extend our algorithm to general Harris recurrent chains. The idea is to extend the original chain to a "virtual" chain with an atom. The splitting technique relies on the crucial notion of *small set*. Recall that, for a Markov chain valued in a state space $(E, \mathcal{E})$ with transition probability Π, a set $S \in \mathcal{E}$ is said to be *small* if there exist $q \in \mathbb{N}^*$, $\delta > 0$ and a probability measure Φ supported by S such that, for all $x \in S$, $A \in \mathcal{E}$,

$$\Pi^q(x, A) \geq \delta \Phi(A), \tag{4.3}$$

Π^q being the q-th iterate of Π. For simplicity, we assume that $q = 1$ (we can always rewrite the chain as a chain based on $(X_i, \cdots, X_{i+q-1})'$ for $q > 1$) and that Φ has a density ϕ with respect to some reference measure $\lambda(\cdot)$. Note that an accessible small set always exists for ψ-irreducible chains : any set $A \in \mathcal{E}$ such that $\psi(A) > 0$ actually contains such a set (see Jain & Jamison, 1967). For a discussion on the practical choice of the small set, see Bertail & Clémençon (2006b).

The idea to construct the split chain $\widetilde{X} = (X, W)$ is the following :

- if $X_i \notin S$, generate (conditionally to X_i) W_i as a Bernoulli random value, with probability δ.
- if $X_i \in S$, generate (conditionally to X_i) W_i as a Bernoulli random value, with probability $\frac{\delta\phi(X_{i+1})}{p(X_i, X_{i+1})}$,

where p is the transition density of the chain X. This construction essentially relies on the fact that under the minorization condition (4.3), $\Pi(x, A)$ may be written on S as a mixture : $\Pi(x, A) = (1-\delta)\frac{\Pi(x,A)-\delta\Phi(A)}{1-\delta} + \delta\Phi(A)$, which is constant (independent of the starting point x) when one picks the second component (see Meyn & Tweedie, 1996, Bertail & Clémençon, 2006b, for details).

When constructed this way, the split chain is an atomic Markov chain, with marginal distribution equal to the original distribution of X (see Meyn & Tweedie, 1996). The atom is then $A = S \times \{1\}$. In practice, we will only need to know when the split chain hits the atom, i.e. we only need to simulate W_i when $X_i \in S$.

The return time conditions are now defined as uniform moment condition over the small set :

$$\mathbf{H0}(\kappa): \ \sup_{x\in S} \mathbb{E}_x[\tau_S^\kappa] < \infty,$$
$$\mathbf{H0}(\kappa,\ \nu): \ \mathbb{E}_\nu[\tau_S^\kappa] < \infty.$$

The Block-moment conditions become :

$$\mathbf{H1}(\kappa,\ m): \ \sup_{x\in S} \mathbb{E}_x\left[\left(\sum_{i=1}^{\tau_S} \|m(X_i,\theta_0)\|\right)^\kappa\right] < \infty,$$
$$\mathbf{H1}(\kappa,\ \nu,\ m): \ \mathbb{E}_\nu\left[\left(\sum_{i=1}^{\tau_S} \|m(X_i,\theta_0)\|\right)^\kappa\right] < \infty.$$

Unfortunately, the Nummelin technique involves the transition density of the chain, which is of course unknown in a non parametric framework. An approximation p_n of this density can however be computed easily by using standard kernel methods. This leads us to the following version of the empirical likelihood program.

Algorithm 4.2 (Approximate regenerative block EL construction)

1. *Find an estimator p_n of the transition density (for instance a Nadaraya-Watson estimator).*
2. *Choose a small set S and a density ϕ on S and evaluate $\delta = \min_{x,y\in S}\left\{\frac{p_n(x,y)}{\phi(y)}\right\}$.*
3. *When X hits S, generate $\widehat{W}_i$ as a Bernoulli with probability $\delta\phi(X_{i+1})/p_n(X_i, X_{i+1})$. If $\widehat{W}_i = 1$, the approximate split chain $(X_i, \widehat{W}_i) = \widehat{X}_i$ hits the atom $A = S \times \{1\}$ and i is an approximate regenerative time. These times define the approximate return times $\widehat{\tau}_A(j)$.*
4. *Count the number of visits of A say $\hat{l}_n + 1 = \sum_{i=1}^n \mathbb{1}_{\widehat{X}_i \in A}$ up to time n.*
5. *Divide the observed trajectory $X^{(n)} = (X_1, ..., X_n)$ into $\hat{l}_n + 2$ blocks corresponding to the pieces of the sample path between approximate return times to the atom A,*

$$\widehat{B}_0 = (X_1, ..., \ X_{\widehat{\tau}_A(1)}), \ \widehat{B}_1 = (X_{\widehat{\tau}_A(1)+1}, ..., \ X_{\widehat{\tau}_A(2)}), ...,$$
$$\widehat{B}_{\hat{l}_n} = (X_{\widehat{\tau}_A(\hat{l}_n)+1}, ..., \ X_{\widehat{\tau}_A(\hat{l}_n+1)}), \ \widehat{B}^{(n)}_{\hat{l}_n+1} = (X_{\widehat{\tau}_A(\hat{l}_n+1)+1}, ..., \ X_n),$$

with the convention $\widehat{B}^{(n)}_{\hat{l}_n+1} = \emptyset$ when $\widehat{\tau}_A(\hat{l}_n + 1) = n$.

6. *Drop the first block* $\widehat{B}_0$, *and the last one* $\widehat{B}^{(n)}_{\hat{l}_n+1}$ *(eventually empty when* $\widehat{\tau}_A(\hat{l}_n+1)=n$*).*
7. *Define*

$$M(\widehat{B}_j,\theta)=\sum_{i=\widehat{\tau}_A(j)+1}^{\widehat{\tau}_A(j+1)} m(X_i,\theta).$$

Evaluate the empirical log-likelihood ratio $r_n(\theta)$ *(practically on a grid of the set of interest) :*

$$\hat{r}_n(\theta)=\sup_{(q_1,\cdots,q_{\hat{l}_n})}\left\{\log\left[\prod_{j=1}^{\hat{l}_n}\hat{l}_n q_j\right]\middle|\sum_{j=1}^{\hat{l}_n} q_j\cdot M(\widehat{B}_j,\theta)=0,\ \sum_{j=1}^{\hat{l}_n} q_j=1\right\}.$$

Using Lagrange arguments or convex duality, this can be more easily calculated as

$$\hat{r}_n(\theta)=\sup_{\lambda\in\mathbb{R}^p}\left\{\sum_{j=1}^{\hat{l}_n}\log\left[1+\lambda' M(\widehat{B}_j,\theta)\right]\right\}.$$

4.4.2 Main theorem

The practical use of this algorithm crucially relies on the preliminary computation of a consistent estimate of the transition kernel. We thus consider some conditions on the uniform consistency of the density estimator p_n. These assumptions are satisfied for the usual kernel or wavelets estimators of the transition density.

H3 For a sequence of nonnegative real numbers $(\alpha_n)_{n\in\mathbb{N}}$ converging to 0 as $n\to\infty$, $p(x,y)$ is estimated by $p_n(x,y)$ at the rate α_n for the mean square error when error is measured by the L^∞ loss over $S\times S$:

$$\mathbb{E}_\nu\left[\sup_{(x,y)\in S\times S}|p_n(x,x')-p(x,x')|^2\right]=\mathcal{O}_\nu(\alpha_n),\ \text{as } n\to\infty.$$

H4 The minorizing probability Φ is such that $\inf_{x\in S}\phi(x)>0$.

H5 The densities p and p_n are bounded over S^2 and $\inf_{x,y\in S} p_n(x,y)/\phi(y)>0$.

Since the choice of Φ is left to the statistician, one can use for instance the uniform distribution overs S, even if it may not be optimal to do so. in that case, **H4** is automatically satisfied. Similarly, it is not difficult to construct an estimator p_n satisfying the constraints of **H5**.

Results of the previous section can then be extended to Harris chains :

Theorem 4.7 *Let* μ *be the invariant measure of the chain, and* $\theta_0\in\mathbb{R}^p$ *be the parameter of interest, satisfying* $\mathbb{E}_\mu[m(X,\theta_0)]=0$*. Consider* $A=S\times\{1\}$ *an atom of the split chain,* τ_A *the hitting time of* A *and* $B=(X_1,\cdots,X_{\tau_A})$*. Assume hypotheses* ***H3****,* ***H5*** *and* ***H5*** *and suppose that* $\mathbb{E}_A[M(B,\theta_0)M(B,\theta_0)']$ *is of full rank.*

(a) *Assume* ***H0***$(4,\nu)$ *and* ***H0***(2) *as well as* ***H1***$(4,\nu,m)$ *and* ***H1***$(2,m)$*, then we have in the just-identified case* $(r=p)$ *:*

$$2\hat{r}_n(\theta_0) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_p$$

and therefore

$$\widehat{C}_{n,\alpha} = \left\{ \theta \in \mathbb{R}^p \,\middle|\, 2\cdot \hat{r}_n(\theta) \leq F_{\chi^2_p}(1-\alpha) \right\}.$$

is an asymptotic confidence region of level $1-\alpha$*.*

(b) *Under the additional assumptions* ***H2***(a)*,* ***H2***(b) *and* ***H2***(d)*,*

$$\hat{\theta} = \arg\inf_{\theta\in\Theta}\{\hat{r}_n(\theta)\}$$

is a consistent estimator of θ_0*. If in addition* ***H2***(c) *holds, then* $\sqrt{n}(\hat{\theta}-\theta_0)$ *is asymptotically normal.*

(c) *In the case of over-identification* $(r>p)$*, we have :*

$$\widehat{W}_{1,n}(\theta_0) = 2\hat{r}_n(\theta_0) - 2\hat{r}_n(\hat{\theta}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_p$$

and

$$\widehat{C}^1_{n,\alpha} = \left\{ \theta \in \mathbb{R}^p \,\middle|\, \widehat{W}_{1,n}(\theta) \leq F_{\chi^2_p}(1-\alpha) \right\},$$

is an asymptotic confidence region of level $1-\alpha$*. The moment equation (4.1) can be tested by using the following convergence in law :*

$$2\hat{r}_n(\hat{\theta}) \xrightarrow[n\to\infty]{\text{under (4.1)}} \chi^2_{r-p}.$$

(d) *Let* $\theta' = (\theta_1,\theta_2)'$*, where* $\theta_1 \in \mathbb{R}^q$ *and* $\theta_2 \in \mathbb{R}^{p-q}$*. Under the hypotheses* $\theta_1 = \theta_{10}$*,*

$$\widehat{W}_{2,n}(\theta_{10}) = 2\inf_{\theta_2}\hat{r}_n((\theta_{10},\theta_2)') - 2\hat{r}_n(\hat{\theta}) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_q$$

and then

$$\widehat{C}^2_{n,\alpha} = \left\{ \theta_1 \in \mathbb{R}^q \,\middle|\, \widehat{W}_{2,n}(\theta_1) \leq F_{\chi^2_q}(1-\alpha) \right\},$$

is an asymptotic confidence region of level $1-\alpha$ *for the parameter of interest* θ_1*.*

4.5 Some simulation results

4.5.1 Linear model with markovian residuals

To illustrate our methodology, and to compare with *Block Empirical Likelihood* (BEL Kitamura, 1997), we consider a dynamic model :

$$Y = \theta_0 Z + \varepsilon, \text{ with } \mathbb{E}[\varepsilon] = 0,$$

where Z and ε are independent Markov chains defined by :

$$Z_0 = 0 \text{ and } Z_i = \left(0.3 + 0.67\mathbb{1}_{Z_{i-1}^2 > 0.3}\right) Z_{i-1} + 0.3(u_i - 1), \tag{4.4}$$

the u_i being i.i.d. with exponential distribution of parameter 1 and

$$\varepsilon_0 = 0 \text{ and } \varepsilon_i = \left(0.97\mathbb{1}_{\varepsilon_{i-1}^2 > 10^{-4}}\right) \varepsilon_{i-1} + 0.03v_i, \tag{4.5}$$

the v_i being i.i.d. with distribution $\mathcal{N}(0,1)$. $X = (Y, Z)$ is then a Markov chain. Let μ be its invariant probability measure. The moment equation corresponding to the linear model is

$$\mathbb{E}_\mu[m(X, \theta_0)] = \mathbb{E}_\mu\left[(Y - \theta_0 Z)Z\right] = 0,$$

where $m\left((Y, Z), \theta\right) = (Y - \theta Z)Z$.

The chain Z has two types of behaviors : while it is smaller than 0.3, it behaves like an A.R. (with exponential innovations) with coefficient 0.3 ; if it gets bigger than 0.3, it behaves like an A.R. with coefficient 0.97. The chain ε is i.i.d. normal while it is smaller than 0.03 and is an A.R. with coefficient 0.97 otherwise. This leads to many small excursions and some large excursions, which are interesting for our method based on blocks of random lengths.

We suppose now that we observe $X = (Y, Z)$ and that we want to find a confidence interval for θ. Note here that the only assumption on Z and ε is that they are markovian of order 1. In practice, the laws of u and v are unknown, as well as the model of the dependence between Z_i and Z_{i-1} on one hand, and ε_i and ε_{i-1} on the other hand. Adjusting a markovian model to the Z_i may be possible but difficult without some previous knowledge. It is much more difficult to adjust a model to the ε_i since they are in practice unobserved, so that a parametric approach is unadapted here.

Figure 4.1, composed of 3 graphics, illustrates our methodology. The first graphic represents 300 realizations the 2 Markov chains, Z and ε. The second graphic represents $Y = \theta_0' Z + \varepsilon$ with $\theta_0 = 1$ and marks the estimated renewal times. As the chain X is 2-dimensional, the small set S is a product of 2 set : $S = S_Y \otimes S_Z$. The 2 sets have been chosen empirically to maximize the number of blocks. X is in S when both $Y \in S_Y$ and $Z \in S_Z$. On the graphics, this is realized when the trajectories of Y and Z are in between the 2 plain black lines. For i such that X_i hits on S, we generate a Bernoulli B_i, and if $B_i = 1$, i is a renewal time. The last graphic gives the confidence intervals built with ReBEL and BEL respectively, and the 95% confidence level. On this example, we can see that θ_0 is in the ReBEL confidence interval but not in the BEL's one.

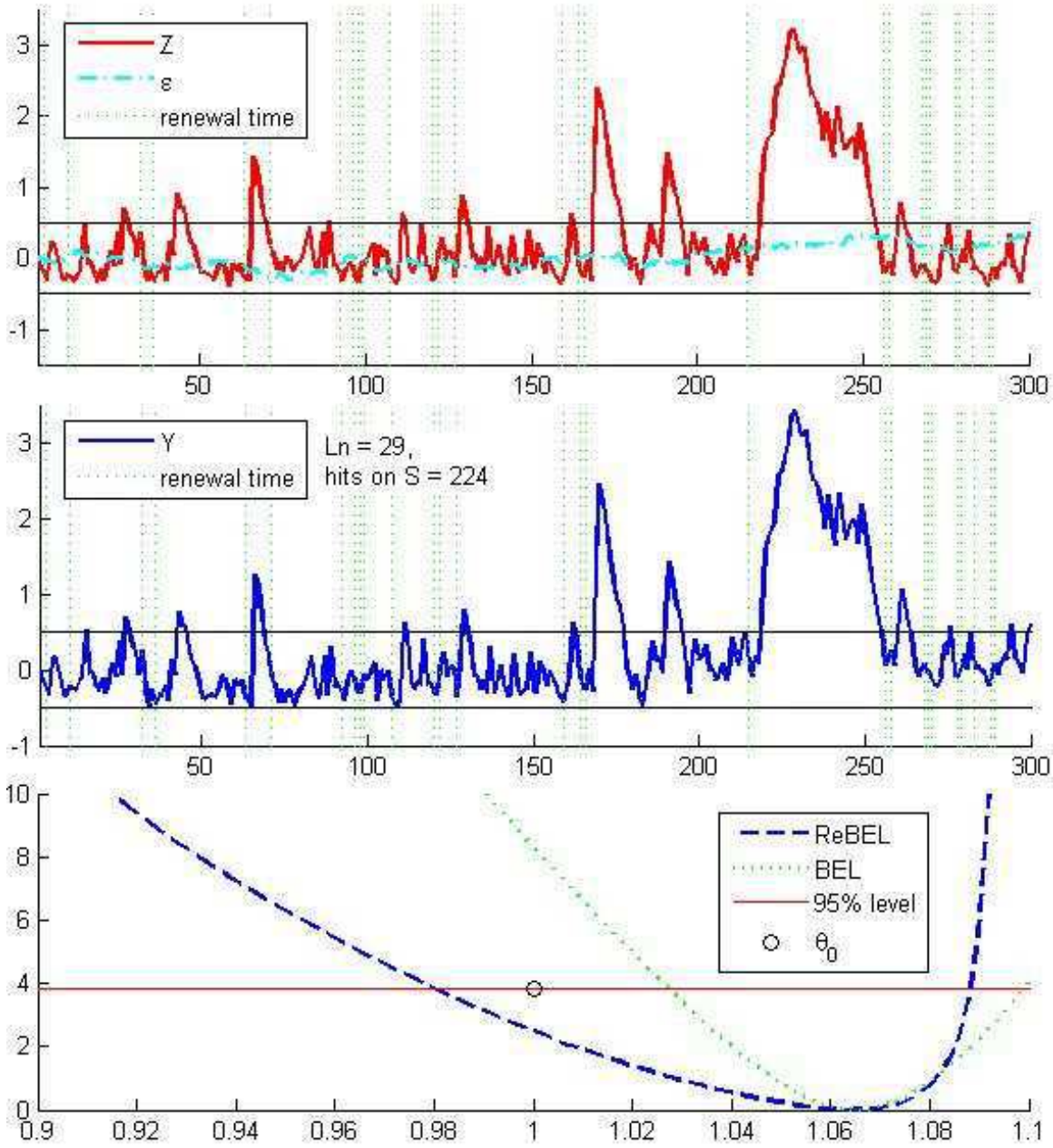


FIG. 4.1 – Trajectories and likelihoods, with 95% confidence level ($\theta_0 = 1$)

We now turn to a simulation with a parameter in $\mathbb{R}^2$. Let $Z_i = (Z_i^1, Z_i^2)'$ be a 2-dimensional chain, where Z_i^1 and Z_i^2 follow independently the dynamic (4.4). The residual ε remains 1-dimensional and follows the dynamic (4.5). The model we want to estimate is $Y = \theta_0' Z + \varepsilon$, and we choose $\theta_0 = (1, 1)'$. Figure 4.2 gives the trajectory of Y and the confidence regions built with ReBEL and BEL respectively.

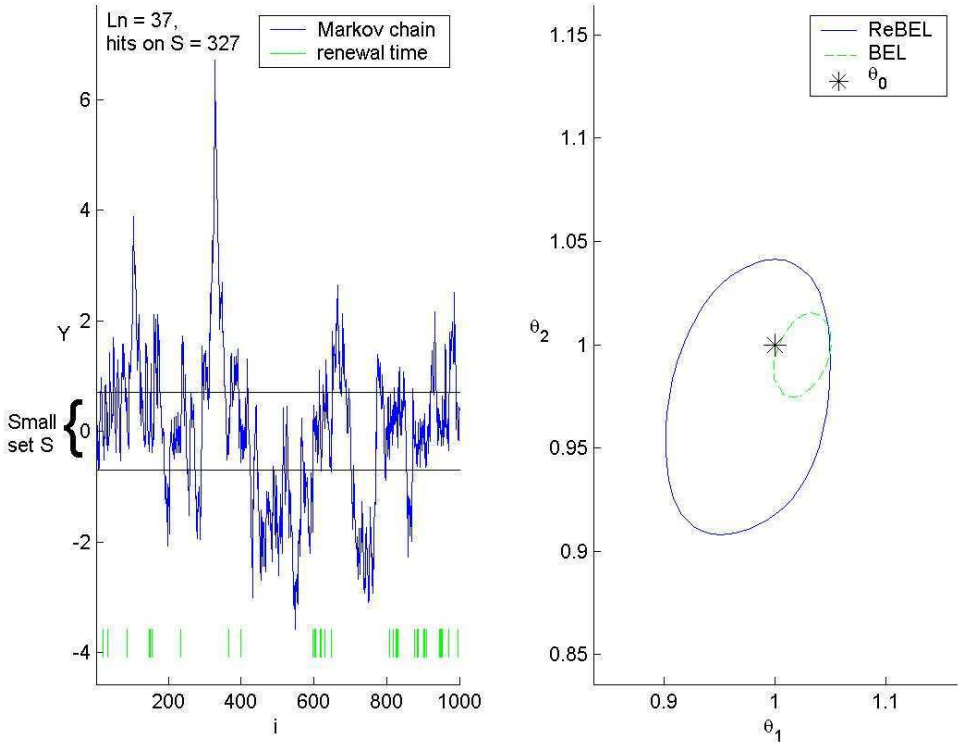


FIG. 4.2 – A trajectory for $\theta_0 = (1;1)'$ and 95% confidence regions for ReBEL and BEL

These simulation studies show that BEL may be too optimistic and can lead to too narrow confidence intervals. To support this point, in the next paragraph, we evaluate by Monte-Carlo simulations the coverage probability and the power of these regions against local alternatives.

4.5.2 Coverage probability and power

To study the small sample behavior of our method, we make a Monte-Carlo experiment (1500 repetition) and compare the coverage probabilities achieved by ReBEL and BEL algorithms. The BEL blocks are of lengths $n^{1/3}$, as suggested by Hall et al. (1995).

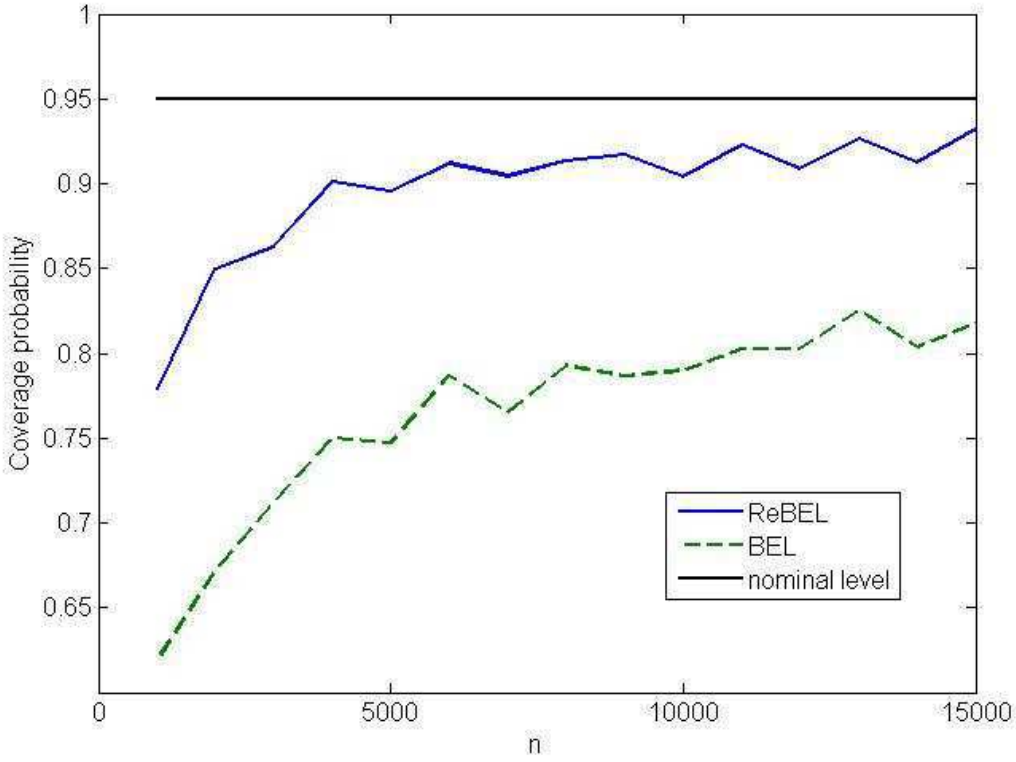


FIG. 4.3 – Coverage probabilities for ReBEL and BEL algorithms

As expected, ReBEL performs better in this framework, for small samples at least. Nevertheless, this is at the price of some loss on the power function. We give in Table 4.1 the false positive rate of the two procedures, i.e. the proportion of Monte-Carlo experiments for which $\theta_0(1 + n^{-1/2})$ is in the 95% confidence interval. The false positive rate of ReBEL is twice worse than BEL's.

n	ReBEL	BEL
1 000	0.17	0.09
5 000	0.29	0.12
10 000	0.27	0.14
15 000	0.28	0.14

TAB. 4.1 – ReBEL and BEL false positive rates against $\theta_0(1 + n^{-1/2})$ alternatives.

4.6 Proofs

4.6.1 Lemmas for the atomic case

Denote $Y_j = M(B_j, \theta_0)$, $\overline{Y} = 1/l_n \sum_{j=1}^{l_n} Y_j$ and define

$$S_{l_n}^2 = 1/l_n \sum_{j=1}^{l_n} M(B_j, \theta_0) M(B_j, \theta_0)' = 1/l_n \sum_{j=1}^{l_n} Y_j Y_j' \text{ and } S_{l_n}^{-2} = (S_{l_n}^2)^{-1}.$$

To demonstrate the Theorem 4.2, we need 2 technical lemmas.

Lemma 4.1 *Assume that $\mathbb{E}_A[M(B, \theta_0)M(B, \theta_0)']$ exists and is full-rank, with eigenvalues $\sigma_p \geq \cdots \geq \sigma_1 > 0$. Then, assuming **H0**$(1, \nu)$ and **H0**(1), we have*

$$S_{l_n}^2 \rightarrow_{\mathbb{P}_\nu} \mathbb{E}_A[M(B, \theta_0)M(B, \theta_0)'].$$

Therefore, for all $u \in \mathbb{R}^p$ with $\|u\| = 1$,

$$\sigma_1 + o_\nu(1) \leq u' S_{l_n}^2 u \leq \sigma_p + o_\nu(1).$$

Proof : The convergence of $S_{l_n}^2$ is a LLN for the sum of a random numbers of random variables, and is a straightforward corollary of the Theorem 6 of Teicher & Chow (1988) (chapter 5.2, page 131).

Lemma 4.2 *Assuming **H0**$(1, \nu)$, **H0**(2) and **H1**$(2, m)$, we have*

$$\max_{1 \leq j \leq l_n} ||Y_j|| = o_\nu(n^{1/2}).$$

Proof : By **H1**$(2, m)$,

$$\mathbb{E}_A \left[\left(\sum_{i=1}^{\tau_1} \|m(X_i, \theta_0)\| \right)^2 \right] < \infty,$$

and then,

$$\mathbb{E}_A[||Y_1||^2] = \mathbb{E}_A \left[\left\| \sum_{i=1}^{\tau_1} m(X_i, \theta_0) \right\|^2 \right] < \infty.$$

By Lemma A.1 of Bonnal & Renault (2004), the maximum of n i.i.d. real-valued random variables with finite variance is $o(n^{1/2})$. Let Z_n be the maximum of n independent copies of $||Y_1||$, Z_n is then such as $Z_n = o_\nu(n^{1/2})$. As l_n is smaller than n, $\max_{1 \leq j \leq l_n} ||Y_j||$ is bounded by Z_n and therefore, $\max_{1 \leq j \leq l_n} ||Y_j|| = o_\nu(n^{1/2})$.

4.6.2 Proof of Theorem 4.2

The likelihood ratio statistic $r_n(\theta_0)$ is the supremum over $\lambda \in \mathbb{R}^p$ of $\sum_{j=1}^{l_n} \log(1+\lambda' Y_j)$. The first order condition at the supremum λ_n is then :

$$1/l_n \sum_{j=1}^{l_n} \frac{Y_j}{1+\lambda_n' Y_j} = 0. \tag{4.6}$$

Multiplying by $\lambda_n$and using $1/(1+x) = 1 - x/(1+x)$, we have

$$1/l_n \sum_{j=1}^{l_n} (\lambda_n' Y_j) \left(1 - \frac{\lambda_n' Y_j}{1+\lambda_n' Y_j}\right) = 0, \text{ and then } \lambda_n' \overline{Y} = 1/l_n \sum_{j=1}^{l_n} \frac{\lambda_n' Y_j Y_j' \lambda_n}{1+\lambda_n' Y_j}.$$

Now we may bound the denominators $1+\lambda_n' Y_j$ by $1 + ||\lambda_n|| \max_j ||Y_j||$ and then

$$\lambda_n' \overline{Y} = 1/l_n \sum_{j=1}^{l_n} \frac{\lambda_n' Y_j Y_j' \lambda_n}{1+\lambda_n' Y_j} \geq \frac{\lambda_n' S_{l_n}^2 \lambda_n}{(1 + ||\lambda_n|| \max_j ||Y_j||)}.$$

Multiply both sides by the denominator, $\lambda_n' \overline{Y}(1 + ||\lambda_n|| \max_j ||Y_j||) \geq \lambda_n' S_{l_n}^2 \lambda_n$ or

$$\lambda_n' \overline{Y} \geq \lambda_n' S_{l_n}^2 \lambda_n - ||\lambda_n|| \max_j ||Y_j|| \lambda_n' \overline{Y}.$$

Dividing by $||\lambda_n||$ and setting $u = \lambda_n / ||\lambda_n||$, we have

$$u' \overline{Y} \geq ||\lambda_n|| \left[u' S_{l_n}^2 u - \max_j ||Y_j|| u' \overline{Y} \right]. \tag{4.7}$$

Now we control the terms in the [ ]. First, by Lemma 4.1, $u' S_{l_n}^2 u$ is bounded between $\sigma_1 + o_\nu(1)$ and $\sigma_p + o_\nu(1)$. Second, by Lemma 4.2, $\max_j ||Y_j|| = o_\nu(n^{1/2})$. Third, the CLT applied to the Y_j's gives $\overline{Y} = \mathcal{O}_\nu(n^{-1/2})$. Then, inequality (4.7) gives :

$$\mathcal{O}_\nu(n^{-1/2}) \geq ||\lambda_n|| \left[u' S_{l_n}^2 u - o_\nu(n^{1/2}) \mathcal{O}_\nu(n^{-1/2}) \right] = ||\lambda_n|| (u' S_{l_n}^2 u + o_\nu(1)),$$

and $||\lambda_n||$ is then $\mathcal{O}_\nu(n^{-1/2})$. Coming back to the first order condition 4.6 and using the equality $\frac{1}{1+x} = 1 - x + \frac{x^2}{1+x}$, we get :

$$0 = 1/l_n \sum_{j=1}^{l_n} Y_j \left(1 - \lambda_n' Y_j + \frac{(\lambda_n' Y_j)^2}{1+\lambda_n' Y_j}\right) = \overline{Y} - S_{l_n}^2 \lambda_n + 1/l_n \sum_{j=1}^{l_n} \frac{Y_j (\lambda_n' Y_j)^2}{1+\lambda_n' Y_j}.$$

The last term is $o_\nu(n^{-1/2})$ by Lemma A.2 of Bonnal & Renault (2004) and then

$$\lambda_n = S_{l_n}^{-2} \overline{Y} + o_\nu(n^{-1/2}).$$

Now, developing the log up to the second order,

$$2r_n(\theta_0) = 2 \sum_{j=1}^{l_n} \log(1+\lambda_n' Y_j) = 2 l_n \lambda_n' \overline{Y} - l_n \lambda_n' S_{l_n}^2 \lambda_n + 2 \sum_{j=1}^{l_n} \eta_j,$$

where the η_i are such that, for some finite $B > 0$ and with probability tending to 1, $|\eta_j| \leq B|\lambda_n' Y_j|^3$. Since, by Lemma 4.2, $\max_j ||Y_j|| = o_\nu(n^{1/2})$,

$$\sum_{j=1}^{l_n} \|Y_j\|^3 \leq n \max_j ||Y_j|| \left(\frac{1}{l_n} \sum_{j=1}^{l_n} \|Y_j\|^2 \right) = n o_\nu(n^{1/2}) \mathcal{O}_\nu(1) = o_\nu(n^{3/2})$$

from which we find

$$2 \sum_{j=1}^{l_n} \eta_j \leq B \|\lambda_n\|^3 \sum_{j=1}^{l_n} \|Y_j\|^3 = \mathcal{O}_\nu(n^{-3/2}) o_\nu(n^{3/2}) = o_\nu(1).$$

Finally,

$$2r_n(\theta_0) = 2l_n \lambda_n' \overline{Y} - l_n \lambda_n' S_{l_n}^2 \lambda_n + o_\nu(1) = l_n \overline{Y} S_{l_n}^{-2} \overline{Y} + o_\nu(1) \underset{\mathcal{L}}{\rightarrow} \chi_p^2.$$

This concludes the proof of Theorem 4.2.

4.6.3 Proof of Theorem 4.3

In order to prove Theorem 4.3, we use a result established by Qin & Lawless (1994).

Lemma 4.3 (Qin & Lawless, 1994) *Let $Z, Z_1, \cdots, Z_n \sim F$ be i.i.d. observations in $\mathbb{R}^d$ and a function $g : \mathbb{R}^d \times \mathbb{R}^p \rightarrow \mathbb{R}^r$ such that $\mathbb{E}_F[g(Z, \theta_0)] = 0$. Suppose that the following hypotheses hold :*

(1) $\mathbb{E}_F[g(Z, \theta_0) g'(Z, \theta_0)]$ is positive definite,

(2) $\partial g(z, \theta)/\partial \theta$ is continuous and bounded in norm by an integrable function $G(z)$ in a neighborhood V of θ_0,

(3) $||g(z, \theta)||^3$ is bounded by $G(z)$ on V,

(4) the rank of $\mathbb{E}_F[\partial g(Z, \theta_0)/\partial \theta]$ is p,

(5) $\dfrac{\partial^2 g(z, \theta)}{\partial \theta \partial \theta'}$ is continuous and bounded by $G(z)$ on V.

Then, the maximum empirical likelihood estimator $\tilde{\theta}_n$ is a consistent estimator of θ_0 and $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is asymptotically normal with mean zero.

Let's set

$$Z = B_1 = \left(X_{\tau_A(1)+1}, \cdots, X_{\tau_A(2)}\right) \in \bigcup_{n \in \mathbb{N}} \mathbb{R}^n$$

and $g(Z, \theta) = M(B_1, \theta)$. Expectation under F is then replaced by $\mathbb{E}_A$. Theorem 4.3 is a straightforward application of the Lemma 4.3 as soon as the assumptions hold.

By assumption, $\mathbb{E}_A[M(B, \theta_0) M(B, \theta_0)']$ is of full rank. This implies (1).

By **H2**(a), there is a neighborhood V of θ_0 and a function N such that, for all i between $\tau_A + 1$ and $\tau_A(2)$, $\partial m(X_i, \theta)/\partial \theta$ is continuous on V and bounded in norm by $N(X_i)$. $\partial M(B_1, \theta)/\partial \theta$ is then continuous as a sum of continuous functions and is bounded for θ in

V by $L(B_1) = \sum_{i=\tau_A(1)+1}^{\tau_A(2)} N(X_i)$. Since N is such that $\mathbb{E}_\mu[N(X)] < \infty$, we have by Kac's Theorem,

$$\mathbb{E}_A\left[\sum_{i=\tau_A(1)+1}^{\tau_A(2)} N(X_i)\right] / \mathbb{E}_A[\tau_A] = \mathbb{E}_A[L(B_1)]/\mathbb{E}_A[\tau_A] < \infty.$$

The bounding function $L(B_1)$ is then integrable. This gives assumption (2). Assumption (5) is derived from **H2**(c) by the same arguments.

By **H2**(d), $\|m(X_i,\theta)\|^3$ is bounded by $N(X_i)$ for θ in V, and then

$$\|M(B_1,\theta)\|^3 \leq \sum_{i=\tau_A(1)+1}^{\tau_A(2)} \|m(X_i,\theta)\|^3 \leq \sum_{i=\tau_A(1)+1}^{\tau_A(2)} N(X_i) = L(B_1).$$

Thus, $\|M(B_1,\theta)\|^3$ is also bounded by $L(B_1)$ for θ in V, and hypotheses (3) follows.

By Kac's Theorem,

$$\mathbb{E}_A[\tau_A]^{-1}\mathbb{E}_A[\partial M(B_1,\theta_0)/\partial\theta] = \mathbb{E}_\mu[\partial m(X_i,\theta_0)/\partial\theta],$$

which is supposed to be of full rank by **H2**(b). Thus $\mathbb{E}_A[\partial M(B_1,\theta_0)/\partial\theta]$ is of full rank and this gives assumption (4).

Under the same hypotheses, Theorem 2 and Corollaries 4 and 5 of Qin & Lawless (1994) hold. They give respectively our Theorems 4.5, 4.4 and 4.6.

4.6.4 Proof of Theorem 4.7

Let suppose that we know the real transition density p. The chain can then be split with the Nummelin technic as above. We get an atomic chain $\widetilde{X}$. Let's denote by B_j the blocks obtained from this chain. The Theorem (4.2) can then be applied to $Y_j = M(B_j,\theta_0)$.

Unfortunately, we do not know p, and then we can not use the Y_j. Instead, we have the vectors $\widehat{Y}_j = M(\widehat{B}_j,\theta_0)$, built on approximatively regenerative blocks. To prove the Theorem 4.7, we essentially need to control the difference between the two statistics $\overline{Y} = \frac{1}{l_n}\sum_{j=1}^{l_n} Y_j$ and $\widehat{Y} = \frac{1}{\hat{l}_n}\sum_{j=1}^{\hat{l}_n} \widehat{Y}_j$. This can be done by using Lemmas (5.2) and (5.3) in Bertail & Clémençon (2006a) : under **H0**$(4,\nu)$,

$$\left|\frac{\hat{l}_n}{n} - \frac{l_n}{n}\right| = \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2}) \tag{4.8}$$

and under **H1**$(4,\nu,m)$ and **H1**$(2,m)$,

$$\left\|\frac{\hat{l}_n}{n}\widehat{Y} - \frac{l_n}{n}\overline{Y}\right\| = \left\|\frac{1}{n}\sum_{j=1}^{\hat{l}_n}\widehat{Y}_j - \frac{1}{n}\sum_{j=1}^{l_n} Y_j\right\| = \mathcal{O}_{\mathbb{P}_\nu}(n^{-1}\alpha_n^{1/2}). \tag{4.9}$$

With some straightforward calculus,

$$\left\|\widehat{Y} - \overline{Y}\right\| \leq \frac{n}{\hat{l}_n}\left\|\frac{\hat{l}_n}{n}\widehat{Y} - \frac{l_n}{n}\overline{Y}\right\| + \left|\frac{l_n}{\hat{l}_n} - 1\right| \left\|\overline{Y}\right\|. \tag{4.10}$$

Since

$$\left| l_n - n/\mathbb{E}_A[\tau_A] \right| \xrightarrow[n\to\infty]{a.s.} 0,$$

with equation (4.8) and some calculus, we have

$$\frac{n}{\hat{l}_n} = \frac{n}{l_n}\left(1 + \frac{n}{l_n}\frac{\hat{l}_n - l_n}{n}\right)^{-1} = \mathcal{O}_{\mathbb{P}_\nu}(\mathbb{E}_A[\tau_A])\left(1 + \mathcal{O}_{\mathbb{P}_\nu}(\mathbb{E}_A[\tau_A])\mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2})\right)^{-1} = \mathcal{O}_{\mathbb{P}_\nu}(\mathbb{E}_A[\tau_A])$$

and

$$\left|\frac{l_n}{\hat{l}_n} - 1\right| = \frac{n}{\hat{l}_n}\frac{\hat{l}_n - l_n}{n} = \mathcal{O}_{\mathbb{P}_\nu}(\mathbb{E}_A[\tau_A])\mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2}) = \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2}).$$

By this and equation (4.10), we have :

$$\left\|\widehat{Y} - \overline{Y}\right\| \leq \mathcal{O}_{\mathbb{P}_\nu}(\mathbb{E}_A[\tau_A])\mathcal{O}_{\mathbb{P}_\nu}(n^{-1}\alpha_n^{1/2}) + \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2})\mathcal{O}_{\mathbb{P}_\nu}(n^{-1/2}) = \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2}n^{-1/2}). \quad (4.11)$$

Therefore

$$n^{1/2}\widehat{Y} = n^{1/2}\overline{Y} + n^{1/2}\left(\overline{Y} - \widehat{Y}\right) = n^{1/2}\overline{Y} + \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2}).$$

Using this and the CLT for the Y_i, we show that $n^{1/2}\widehat{Y}$ is asymptotically gaussian.

The same kind of arguments give a control on the difference between empirical variances. Consider

$$\widehat{S}^2_{\hat{l}_n} = \sum_{j=1}^{\hat{l}_n} \widehat{Y}_j\widehat{Y}_j' \text{ and } \widehat{S}^{-2}_{\hat{l}_n} = (\widehat{S}^2_{\hat{l}_n})^{-1}.$$

By Lemma (5.3) of Bertail & Clémençon (2006a) we have, under $\mathbf{H1}(4,\nu,m)$ and $\mathbf{H1}(2,m)$, $\left\|\frac{\hat{l}_n}{n}\widehat{S}^2_{\hat{l}_n} - \frac{l_n}{n}S^2_{l_n}\right\| = \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n)$, and then

$$\left\|\widehat{S}^2_{\hat{l}_n} - S^2_{l_n}\right\| \leq \frac{n}{\hat{l}_n}\left\|\frac{\hat{l}_n}{n}\widehat{S}^2_{\hat{l}_n} - \frac{l_n}{n}S^2_{l_n}\right\| + \left|\frac{l_n}{\hat{l}_n} - 1\right|\left\|S^2_{l_n}\right\| = \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n) + \mathcal{O}_{\mathbb{P}_\nu}(\alpha_n^{1/2}) = o_{\mathbb{P}_\nu}(1). \quad (4.12)$$

The proof of the Theorem (4.2) is then also valid for the approximated blocks $\widehat{B}_j$ and reduce to the study of the square of a self-normalized sum based on the pseudo-blocks. We have $\hat{r}_n(\theta_0) = \sup_{\lambda\in\mathbb{R}^p}\left\{\sum_{j=1}^{\hat{l}_n}\log\left[1+\lambda'\widehat{Y}_j\right]\right\}$. Let $\hat{\lambda}_n = -\widehat{S}^{-2}_{\hat{l}_n}\widehat{Y} + o_{\mathbb{P}_\nu}(n^{-1/2})$ be the optimum value of λ, we have :

$$2\hat{r}_n(\theta_0) = -2\hat{l}_n\hat{\lambda}_n'\widehat{Y} - \sum_{j=1}^{\hat{l}_n}(\hat{\lambda}_n'\widehat{Y}_j)^2 + o_{\mathbb{P}_\nu}(1) = \hat{l}_n\widehat{Y}'\widehat{S}^{-2}_{\hat{l}_n}\widehat{Y} + o_{\mathbb{P}_\nu}(1).$$

Using the controls given by equations (4.11) and (4.12), we get

$$2\hat{r}_n(\theta_0) = [l_n + \mathcal{O}_{\mathbb{P}_\nu}(n\alpha_n^{1/2})]\cdot\left[\overline{Y}' + \mathcal{O}_{\mathbb{P}_\nu}\left(\sqrt{\frac{\alpha_n}{n}}\right)\right]\cdot[S^{-2}_{l_n} + o_{\mathbb{P}_\nu}(1)]\cdot\left[\overline{Y} + \mathcal{O}_{\mathbb{P}_\nu}\left(\sqrt{\frac{\alpha_n}{n}}\right)\right] + o_{\mathbb{P}_\nu}(1).$$

Developing this product, the main term is $l_n\overline{Y}S^{-2}_{l_n}\overline{Y} \sim_{\mathbb{P}_\nu} 2r_n(\theta_0)$ and all other terms are $o_{\mathbb{P}_\nu}(1)$:

$$2\hat{r}_n(\theta_0) = l_n\overline{Y}S^{-2}_{l_n}\overline{Y} + o_{\mathbb{P}_\nu}(1) \underset{\mathcal{L}}{\to} \chi^2_p,$$

Results (b), (c) and (d) can be derived from the atomic case by using the same arguments.

Deuxième partie

Applications des méthodes de vraisemblance empirique

Chapitre 5

Using empirical likelihood to combine data, application to food risk assessment

Introduction

Empirical likelihood introduced by Owen (Owen, 1988, 1990) is a nonparametric inference method based on a data driven likelihood ratio function. The empirical likelihood techniques have been widely developed these last years. Refer to Owen (2001) book and the references therein for a complete bibliography on the topic. Like the bootstrap and the jackknife, empirical likelihood inference does not require the specification of a family of distributions for the data. Empirical likelihood can be thought as a bootstrap without resampling or as a likelihood without parametric assumptions. Likelihood methods are very effective. They can be used to find efficient estimators, and to build tests that have good power. Likelihood is also flexible : when data are incompletely observed, distorted or sampled with bias, likelihood methods can be used to offset or even correct these problems (Owen, 2001). Knowledge arising from other data can be incorporated via constraints under the form of estimating equations. Other methods can be applied to incorporate side information, like in survey sampling, Deville & Sarndal (1992), by weighting, Hellerstein & Imbens (1999) or by data combination, Moffitt & Ridder (2005). This issue can also be linked to an early paper by Ireland & Kullback (1968). Empirical likelihood has the advantage to combine the reliability of the nonparametric methods with the flexibility and the effectiveness of the likelihood approach.

A fundamental problem in risk assessment and particularly in food risk assessment is the diversity of data sources. We often have consumption data coming from different surveys (household budget panels, food dietary records, 24 hours recall and food frequency questionnaires) using different methodologies (stratified sampling, random sampling or quota methods) and analytical contamination data coming from different laboratories. An accurate estimation of a food risk index and its uncertainty are essential since the resulting confidence intervals for the risk index may serve as arguments for nutritional recommendations or new standards about the contamination of the food. The aim of this paper is to show how empirical likelihood can be used to combine different sources of data in order to estimate a food risk index.

In the first section, we recall that the flexibility of empirical likelihood allows to combine several independent sources of data. Owen generalizes Wilks' result, stating that likelihood ratio in parametric models are asymptotically χ^2, to nonparametric or semi-parametric models. This result allows to build confidence region for simple parameters. We extend this result to the problem of combining data. We focus on building confidence region for the common mean of two independent samples. Similar problems have been studied by Qin (1993), see also chapters 3, 6 and 11 of Owen (2001), pages 51, 130 and 223-225.

In the second section, our aim is to build confidence regions for a parameter of interest which is a food risk index, using the generalization of empirical likelihood to combine different sources of data. This risk index is defined as the probability that the exposure to a contaminant exceeds a safe dose d. Exposure to a contaminant that concerns P food products is calculated as the cross product between the $P-$dimensional vector of consumptions and the P contamination values. The safe dose d is called Provisional Tolerable Weekly Intake (PTWI) when consumption is expressed on a week and body weight basis. For our estimation problem, we have $P+2$ samples corresponding to the contaminations of the P products and 2 complementary consumption surveys. The optimization program of the empirical likelihood

is at first glance difficult to solve in this case, due to the high nonlinearity of the parameter of interest. Following Bertail (2006), a solution is to linearize the constraints defining the parameter of interest. The linearization consists in decomposing the nonlinear functional into a sum of independent influence functions using Hadamard differentiability arguments. Since our parameter of interest is also a generalized U-statistics (Bertail & Tressou, 2004), this linearization can be viewed as an Hoeffding decomposition. On the other hand, the high multidimensionality of the problem requires the use of incomplete U-statistics in the case of $P > 1$ (Lee, 1990). The asymptotic convergence to a χ^2 of the likelihood ratio calculated with this linearization and incomplete U-statistics is checked and the ideas of the proof are given in appendix 5.5.3.

In the third section, we apply our results to assess the risk due to the presence of methylmercury (MeHg) in fish and sea products. Indeed, at high concentrations, methylmercury, a well-known environmental toxic found in the aquatic environment, can cause lesions of the nervous system and serious mental deficiencies in infants whose mothers were exposed during pregnancy (WHO, 1990). There is also some concerns that methylmercury may give rise to retarded development or other neurological effects at lower levels of exposure, which are consistent with standard patterns of fish consumption (Davidson et al., 1995, Grandjean et al., 1997, National Research Council (NRC) of the national academy of sciences Price, 2000). The latest epidemiological results compiled by the Joint Expert Committee on Food Additives and Contaminants (FAO/WHO, 2003) yields a Provisional Tolerable Weekly Intake (PTWI) for methylmercury of 1.6 μg per week per kg of body weight. Methylmercury is mainly found in fish and fishery products. Other food products are therefore excluded to estimate human exposure in this paper.

The main result of this paper is to show how to improve the estimation of the probability that the French exposure exceeds the PTWI by combining the two consumption surveys available in France and the contamination data by empirical likelihood techniques. The resulting estimator of the probability of exceeding the methylmercury PTWI when eating sea products is 5.43%. A 95% confidence interval is given by $[5.20\%; 5.64\%]$.

5.1 Empirical likelihood as a tool for combining data

Suppose that we have two independent samples $\left(X_i^{(1)}\right)_{1\leq i\leq n_1}$ and $\left(X_j^{(2)}\right)_{1\leq j\leq n_2}$ which are respectively independent and identically distributed (i.i.d.) with distributions P_1 and P_2 with the same mean $\mathbb{E}_{P_1}(X^{(1)}) = \mathbb{E}_{P_2}(X^{(2)}) = \mu \in \mathbb{R}^d$. The empirical likelihood for these two samples is given by

$$\prod_{i=1}^{n_1} n_1 p_i^{(1)} \prod_{j=1}^{n_2} n_2 p_j^{(2)},$$

where $\mathcal{P} = \left\{ \left(p_i^{(1)}\right)_{1\leq i\leq n_1}, \left(p_j^{(2)}\right)_{1\leq j\leq n_2} \right\}$ is the set of weights related to $\left(X_i^{(1)}\right)_{1\leq i\leq n_1}$ and $\left(X_j^{(2)}\right)_{1\leq j\leq n_2}$, with constraints

$$0 \leq p_i^{(1)} \leq 1,\ 0 \leq p_j^{(2)} \leq 1,\ \sum_{i=1}^{n_1} p_i^{(1)} = 1,\ \sum_{j=1}^{n_2} p_j^{(2)} = 1.$$

The constraints on the positivity of the weights are forced as soon as log-likelihoods are considered. The weights being positives and summing to 1, none can be bigger than 1. Therefore, we only keep the constraints on the sums.

The idea now is to maximize this empirical likelihood product under the constraints provided by the model :

$$C(\mu) = \left\{ \mathcal{P} \left| \sum_{i=1}^{n_1} p_i^{(1)} X_i^{(1)} = \mu,\ \sum_{j=1}^{n_2} p_j^{(2)} X_j^{(2)} = \mu,\ \sum_{i=1}^{n_1} p_i^{(1)} = 1,\ \sum_{j=1}^{n_2} p_j^{(2)} = 1 \right. \right\}.$$

This constraints set $C(\mu)$ can be augmented by some estimating equations that would allow to incorporate some knowledge arising from other data or from the model under consideration. For example, the national census provides the margin distribution of the population according to different criteria (age, sex, region, profession) and could be integrated via estimating equations of the form

$$\sum_{i=1}^{n_1} p_i^{(1)} Z_i^{(1)} = z_0, \qquad \sum_{j=1}^{n_2} p_j^{(2)} Z_j^{(2)} = z_0, \tag{5.1}$$

where $Z_i^{(1)}$ and $Z_j^{(2)}$ are vectors describing the belonging to specified sociodemographic categories in surveys 1 and 2 and z_0 is the vector of the corresponding percentages of these categories based on the national census. The convergence results will not be affected by the introduction of such sociodemographic criteria, see Qin & Lawless (1994) and Owen (2001), chapter 3, page 51.

At any fixed μ,

$$L_{n_1,n_2}(\mu) = \sup_{\mathcal{P}\in C(\mu)} \left\{ \prod_{i=1}^{n_1} n_1 p_i^{(1)} \prod_{j=1}^{n_2} n_2 p_j^{(2)} \right\}$$

can be seen as the product of two independent likelihoods : $L_{n_1,n_2}(\mu) = L_{n_1}(\mu) L_{n_2}(\mu)$ with

$$L_{n_r}(\mu) = \sup_{\left(p_i^{(r)}\right)_{1\leq i\leq n_r}} \left\{ \prod_{i=1}^{n_r} n_r p_i^{(r)} \right\},$$

for $r = 1, 2$, each set of weights verifying its constraints. Using Kühn and Tücker's arguments on each optimization separately (Owen, 1990), we can write the log-likelihood

$$l_{n_r}(\mu) = -\log\left[L_{n_r}(\mu)\right] = \sup_{\lambda_r} \left\{ \sum_{i=1}^{n_r} \log\left[1 + \lambda_r'\left(X_i^{(r)} - \mu\right)\right] \right\}$$

where $\lambda_r \in \mathbb{R}^d$ is the Kühn and Tücker's vector associated to the constraint $\sum_{i=1}^{n_r} p_i^{(r)} X_i^{(r)} = \mu$. Finally the empirical likelihood program is equivalent to

$$l_{n_1,n_2}(\mu) = -\log\left[L_{n_1,n_2}(\mu)\right] = \sup_{\lambda_1,\lambda_2}\left\{\sum_{i=1}^{n_1}\log\left[1+\lambda_1'\left(X_i^{(1)}-\mu\right)\right]+\sum_{j=1}^{n_2}\log\left[1+\lambda_2'\left(X_j^{(2)}-\mu\right)\right]\right\}.$$

The likelihood ratio test statistic can be written $r_{n_1,n_2}(\mu) = 2\left[l_{n_1,n_2}(\widehat{\mu}_n) - l_{n_1,n_2}(\mu)\right]$, where $\widehat{\mu}_n$ is the estimator : $\widehat{\mu}_n = \arg\sup_\mu l_{n_1,n_2}(\mu)$.

Theorem 5.1 (Convergence to a χ^2)
Let $\left(X_i^{(1)}\right)_{1\le i\le n_1} \sim P_1$ *i.i.d. and* $\left(X_j^{(2)}\right)_{1\le j\le n_2} \sim P_2$ *i.i.d. be two independent samples with common mean* $\mu_0 \in \mathbb{R}^d$. *For* $r = 1, 2$, *assume that the variance-covariance matrix of* $X^{(r)} - \mu_0$ *is invertible and that* $\mathbb{E}\left[||X^{(r)}-\mu||^3\right] < \infty$ *on a neighborhood of* μ_0. *If, in addition,* $\min(n_1, n_2)$ *goes to* $+\infty$ *and that* $\log\log\max(n_1, n_2) = o(\min(n_1, n_2)^{1/3})$, *then*

$$r_{n_1,n_2}(\mu_0) \xrightarrow[n_1\to\infty]{n_2\to\infty} \chi^2(d).$$

A confidence interval for μ_0 is thus given by $\left\{\mu \,|\, r_{n_1,n_2}(\mu) \le \chi^2_{1-\alpha}(d)\right\}$, where $\chi^2_{1-\alpha}(d)$ is the $(1-\alpha)^{th}$ percentile of the χ^2 distribution with d degree of freedom. The proof of this theorem is postponed to the appendix, see 5.5.1. The existence of a third moment is needed to follow the main arguments of Qin & Lawless (1994). The control $\max(n_1, n_2)$ by a function of $\min(n_1, n_2)$ is required for the control of $l_{n_1,n_2}(0)$. Other assumptions are classical in empirical likelihood literature.

5.2 Generalization to the construction of confidence intervals for a food risk index

In this section, we want to estimate θ_d, the probability that exposure to a contaminant exceeds a tolerable dose d, when P products (or groups of products) are assumed to be contaminated taking into account different data sources. For this purpose, $P+2$ data sets are available : two data sets coming from two complementary consumption surveys and the P sets of contamination values. In the first section, we have 2 samples and a linear model constraint. In this section, we want to combine $P+2$ samples under 2 nonlinear model constraints. We assume that the 2 consumption surveys concern the same population. Therefore the probabilities that exposure to a contaminant exceeds a dose d, estimated with each consumption samples, are equal, and their common value is θ_d. Our aim is to estimate θ_d and to give a confidence interval.

5.2.1 Framework and notations

For $k = 1, ..., P$, $Q^{[k]}$ denotes the random variable for the contamination of product k, with distribution $\mathcal{Q}^{[k]}$. $\left(q_l^{[k]}\right)_{l=1,...,L_k}$ is a L_k-sample i.i.d. from $\mathcal{Q}^{[k]}$. Its empirical distribution

is

$$\mathcal{Q}_{L_k}^{[k]} = \frac{1}{L_k} \sum_{l=1}^{L_k} \delta_{q_l^{[k]}},$$

where $\delta_{q_l^{[k]}}(q) = 1$ if $q = q_l^{[k]}$ and 0 else.

$C^{(r)}$ denotes the P-dimensional random variable for the relative consumption vector in survey $r = 1, 2$, with distribution $\mathcal{C}^{(r)}$. Consumptions are "relative" consumptions in the sense that they are expressed in terms of individual body weight. $\left(c_{1,i}^{(r)} \dots c_{P,i}^{(r)}\right)_{1 \le i \le n_r} = \left(c_i^{(r)}\right)_{1 \le i \le n_r}$ is a n_r-i.i.d. sample from $\mathcal{C}^{(r)}$ for survey $r = 1, 2$. Their empirical distributions for survey $r = 1, 2$ are

$$\mathcal{C}_{n_r}^{(r)} = \frac{1}{n_r} \sum_{i=1}^{n_r} \delta_{c_i^{(r)}}.$$

The probability that the exposure of one individual exceeds a dose d is $\theta_d^{(r)} = \Pr\left(D^{(r)} > d\right)$, with $D^{(r)} = \sum_{k=1}^{P} Q^{[k]} C_k^{(r)}$ when using the survey r.

5.2.2 Empirical likelihood program

We define the sets of weights $\mathcal{P} = \left\{ \left(p_i^{(1)}\right)_{i=1}^{n_1}, \left(p_j^{(2)}\right)_{j=1}^{n_2}, \left\{ \left(w_{l_k}^{[k]}\right)_{l_k=1}^{L_k}, k = 1, \dots, P \right\} \right\}$ associated to the 2 samples of consumption and the P samples of contamination. The empirical likelihood is given by

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{k=1}^{P} \prod_{l_k=1}^{L_k} w_{l_k}^{[k]},$$

with 2 constraints on consumption weights : for $r = 1, 2$, $\sum_{i=1}^{n_r} p_i^{(r)} = 1$ and P constraints on contamination weights : $\forall 1 \le k \le P$, $\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} = 1$.

Let $\widetilde{\mathcal{Q}}_{L_k}^{[k]}$ denote a discrete probability measure dominated by $\mathcal{Q}_{L_k}^{[k]}$, that is $\widetilde{\mathcal{Q}}_{L_k}^{[k]} = \sum_{l=1}^{L_k} w_l^{[k]} \delta_{q_l^{[k]}}$ with $w_l^{[k]} > 0$ and $\sum_{l=1}^{L_k} w_l^{[k]} = 1$ for $k = 1, \dots, P$. In the same way, $\widetilde{\mathcal{C}}_{n_1}^{(1)}$ and $\widetilde{\mathcal{C}}_{n_2}^{(2)}$ are discrete probability measures dominated by $\mathcal{C}_{n_1}^{(1)}$ and $\mathcal{C}_{n_2}^{(2)}$, i.e. $\widetilde{\mathcal{C}}_{n_r}^{(r)} = \sum_{i=1}^{n_r} p_i^{(r)} \delta_{c_i^{(r)}}$ with $p_i^{(r)} > 0$ and $\sum_{i=1}^{n_r} p_i^{(r)} = 1$, $r = 1, 2$. $\mathbb{E}_{\widetilde{\mathcal{D}}_r}$ denotes the expectation under the joint discrete probability distribution $\widetilde{\mathcal{D}}_r = \prod_{k=1}^{P} \widetilde{\mathcal{Q}}_{L_k}^{[k]} \times \widetilde{\mathcal{C}}_{n_r}^{(r)}$, which is the reweighted joint discrete probability distribution of

the P contamination samples and the r^{th} consumption survey sample.

The model constraints can now be written, for $r = 1, 2$,

$$\mathbb{E}_{\widetilde{\mathcal{D}}_r}\left\{\mathbb{1}\left\{\sum_{k=1}^{P} Q^{[k]}C_k^{(r)} > d\right\} - \theta_d\right\} = 0, \tag{5.2}$$

Theses model constraints on θ_d have an explicit (but unpleasant) expression :

$$\text{for } r = 1, 2: \quad \theta_d = \theta_d^{(r)} = \sum_{i=1}^{n_r}\sum_{l_1=1}^{L_1}\cdots\sum_{l_k=1}^{L_k}\cdots\sum_{l_P=1}^{L_P} p_i^{(r)}\left(\prod_{j=1}^{P} w_{l_j}^{[j]}\right)\mathbb{1}\left\{\sum_{k=1}^{P} q_{l_k}^{[k]}c_{k,i}^{(r)} > d\right\}.$$

5.2.3 Linearization and approximated empirical likelihood

The preceding empirical likelihood program is difficult to solve, both from theoretical and practical points of view, because of the highly nonlinear form of the model constraints. The same problem already appears when studying the asymptotic behavior of the plug-in estimator of θ_d with only one consumption survey. One solution is to see θ_d as a generalized U-statistic and to linearize it using Hoeffding decomposition (Lee, 1990, Bertail & Tressou, 2004). More generally, a method is to linearize the constraints to solve the optimization problem. This linearization is asymptotically valid as soon as the parameter of interest is Hadamard differentiable, see Bertail (2006) for details. Linearization is made easier by considering the influence function of $\Psi_{\mathcal{D}}$=$\mathbb{E}_{\mathcal{D}}\left[\mathbb{1}\left\{\sum_{k=1}^{P} Q^{[k]}C_k^{(r)} > d\right\} - \theta_d\right]$, where $\mathcal{D}$ is the joint distribution of contaminations and consumptions. The influence function of $\Psi_{\mathcal{D}}$ at point $\left(q_1, \ldots, q_P, c^{(r)}\right)$ is, for $r = 1, 2$:

$$\begin{aligned}\Psi_{\mathcal{D}}^{(1)}\left(q_1, \ldots, q_P, c\right) = {} & \mathbb{E}_{\prod_{k=1}^{P} \mathcal{Q}_{L_k}^{[k]}}\left[\mathbb{1}_{\sum_{k=1}^{P} Q^{[k]}C_k^{(r)} > d} - \theta_d \middle| C^{(r)} = c\right] \\ & + \sum_{m=1}^{P} \mathbb{E}_{\mathcal{C}_{n_r}^{(r)} \times \prod_{k\neq m} \mathcal{Q}_{L_k}^{[k]}}\left[\mathbb{1}_{\sum_{k=1}^{P} Q^{[k]}C_k^{(r)} > d} - \theta_d \middle| Q^{[m]} = q_m\right].\end{aligned}$$

This functional of $\mathcal{D}$ can be estimated by its empirical counterpart $\Psi_{\widehat{\mathcal{D}}}^{(1)}$, where $\widehat{\mathcal{D}}$ denotes the empirical version of $\mathcal{D}$. $\Psi_{\widehat{\mathcal{D}}}^{(1)}$ can be written explicitly :

$$\Psi_{\widehat{\mathcal{D}}}^{(1)}\left[q_1, \ldots, q_P, c\right] = U_0\left(c\right) + U_1^{(r)}(q_1) + \ldots + U_m^{(r)}(q_m) + \ldots + U_P^{(r)}(q_P), \tag{5.3}$$

where

$$U_0\left(c\right) = \frac{1}{\prod_{k=1}^{P} L_k}\sum_{\substack{1\leq l_k\leq L_k \\ 1\leq k\leq P}} \mathbb{1}_{\sum_{k=1}^{P} q_{l_k}^{[k]}c_k > d} - \theta_d, \tag{5.4}$$

and, for $m = 1 \cdots P$ and $r = 1, 2$,

$$U_m^{(r)}(q_m) = \frac{1}{n_r \times \prod_{\substack{k=1 \\ k \neq m}}^{P} L_k} \sum_{i=1}^{n_r} \sum_{l_1=1}^{L_1} \cdots \sum_{l_{m-1}=1}^{L_{m-1}} \sum_{l_{m+1}=1}^{L_{m+1}} \cdots \sum_{l_P=1}^{L_P} \mathbb{1}\left\{ q_m c_{i,m}^{(r)} + \sum_{\substack{k=1 \\ k \neq m}}^{P} q_{l_k}^{[k]} c_{i,k}^{(r)} > d \right\} - \theta_d. \quad (5.5)$$

$U_0\left(c^{(r)}\right)$ and the $\left(U_m^{(r)}(q^{[m]})\right)_{m=1}^{P}$ are generalized U-statistics with kernel $\mathbb{1}_{\sum_{k=1}^{P} q^{[k]} c_k > d}$ and degree $(1, ..., 1) \in \mathbb{R}^P$, see Lee (1990). For simplicity, the dependence in $n_r, L_1, ..., L_P$ is not explicited in the notations.

An approximate version of the model constraints (5.2) can now be written :

$$\text{for } r = 1, 2 : \qquad \mathbb{E}_{\widetilde{\mathcal{D}}_r}\left[\Psi_{\widehat{\mathcal{D}}}^{(1)}\left(Q^{[1]}, \ldots, Q^{[P]}, C^{(r)}\right)\right] = 0,$$

that is

$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_k^{(1)}\left(q_{l_k}^{[k]}\right)\right] = 0,$$
$$\sum_{j=1}^{n_2} p_j^{(2)} U_0\left(c_j^{(2)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_k^{(2)}\left(q_{l_k}^{[k]}\right)\right] = 0.$$

The following theorem establishes the asymptotic convergence of the approximate version of the empirical likelihood when $P = 1$.

Theorem 5.2 *Assume that we have a contamination data* $(q_l)_{1 \leq l \leq L}$ *i.i.d. and* 2 *independent consumption samples* $\left(c_i^{(1)}\right)_{1 \leq i \leq n_1}$ *i.i.d. and* $\left(c_j^{(2)}\right)_{1 \leq j \leq n_2}$ *i.i.d. with common risk index :*

$$\theta_d^{(1)} = \theta_d^{(2)} = \theta_d \in \mathbb{R}.$$

Assume that for $r = 1, 2$, $U_0\left(c_1^{(r)}\right)$ *have finite variances and that* $\left(U_1^{(1)}(q_1),\ U_1^{(2)}(q_1)\right)'$ *has a finite invertible variance-covariance matrix. Assume also that* n_1, n_2 *and* L *go to infinity and that their ratios are bounded, then the empirical likelihood program consists in solving the dual program*

$$l_{n_1,n_2,L}(\theta_d) = \sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_3 \in \mathbb{R} \\ n_1+n_2+L-\gamma_1-\gamma_2-\gamma_3=0}} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \log\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right)\right\} + \sum_{j=1}^{n_2} \log\left\{\gamma_2 + \lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ + \sum_{l=1}^{L} \log\left\{\gamma_3 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right\}. \quad (5.6)$$

Define the maximum likelihood estimator associated to this quantity $\widehat{\theta}_d = \arg\sup_{\theta_d} l_{n_1,n_2,L}(\theta_d)$. *Define the* log*-likelihood ratio* $r_{n_1,n_2,L}(\theta_d) = 2\left[l_{n_1,n_2,L}\left(\widehat{\theta}_d\right) - l_{n_1,n_2,L}(\theta_d)\right]$, *then*

$$r_{n_1,n_2,L}(\theta_d) \to \chi^2(1).$$

The proof of these results is given in appendix 5.5.2. This theorem yields an $(1-\alpha)^{th}$ confidence interval for θ_d such that

$$\left\{\theta_d : r_{n_1,n_2,L}(\theta_d) \leq \chi^2_{1-\alpha}(1)\right\}.$$

From a practical point of view, the linearization of the constraints allows for a good convergence of the optimization algorithm (for instance by using a gradient descent method such as Newton-Raphson). The algorithmic aspects of empirical likelihood are discussed in chapter 12 from Owen (2001).

5.2.4 Extension to the case of several products by incomplete U-statistics

For $P > 1$, the computation of the different U-statistics defined in (5.4) and (5.5) becomes too heavy when the data sets are large (if L_k and/or n_r are large). Indeed, one needs to compute at least $n_r \prod_{k=1}^{P} L_k$ terms. To solve this problem, we proceed to an approximation by replacing the complete U-statistics by incomplete U-statistics. The properties of incomplete U-statistics are well described in Blom (1976) or Lee (1990).

Let us define the incomplete U-statistics associated to equations (5.4) and (5.5). For $r = 1$ or 2, the incomplete version of (5.4) is given by

$$U_{0,\mathcal{B}_0^{(r)}}\left(c^{(r)}\right) = \frac{1}{B_0^{(r)}} \sum_{(l_1 \ldots l_P) \in \mathcal{B}_0^{(r)}} \mathbb{1}\left\{\sum_{k=1}^{P} q_{l_k}^{[k]} c_k^{(r)} > d\right\} - \theta_d, \tag{5.7}$$

where the set $\mathcal{B}_0^{(r)}$ of size $B_0^{(r)}$ is a set of indexes $(l_1, \ldots, l_P)$, randomly chosen with replacement from $\bigotimes_{k=1}^{P} \{1, \ldots, L_k\}$.

For $m = 1, \ldots, P$, the incomplete version of (5.5) is given by

$$U^{(r)}_{m,\mathcal{B}_m^{(r)}}(q_m) = \frac{1}{B_m^{(r)}} \sum_{(l_1,\ldots,l_{m-1},l_{m+1},\ldots,l_P,i) \in \mathcal{B}_m^{(r)}} \mathbb{1}\left\{\sum_{k=1}^{m-1} q_{l_k}^{[k]} c_{i,k}^{(r)} + q_m c_{i,m}^{(r)} + \sum_{k=m+1}^{P} q_{l_k}^{[k]} c_{i,k}^{(r)} > d\right\} - \theta_d, \tag{5.8}$$

where the set $\mathcal{B}_m^{(r)}$ of size $B_m^{(r)}$ is a set of indexes $(l_1, \ldots, l_{m-1}, l_{m+1}, \ldots, l_P, i)$ that are randomly chosen with replacement from $\bigotimes_{\substack{k=1 \\ k \neq m}}^{P} \{1, \ldots, L_k\} \times \{1 \ldots n_r\}$.

In the following, we will use $B = B_0^{(r)} = B_m^{(r)}$, for $m = 1, ..., P$ and $r = 1, 2$. As soon as B is chosen such as $n_1 + n_2 + \sum_{k=1}^{P} L_k = o(B)$, the difference between the complete and the incomplete versions is of order $o(B^{-1/2})$.

The approximate influence function is now given by

$$\Psi_B^{(1)}\left(q_1, \ldots, q_P, c^{(r)}\right) = U_{0,\mathcal{B}_0^{(r)}}\left(c^{(r)}\right) + U^{(r)}_{1,\mathcal{B}_1^{(r)}}(q_1) + \ldots + U^{(r)}_{m,\mathcal{B}_m^{(r)}}(q_m) + \ldots + U^{(r)}_{P,\mathcal{B}_P^{(r)}}(q_P).$$

The model constraints becomes then

$$\sum_{i=1}^{n_1} p_i^{(1)} U_{0,\mathcal{B}_0^{(1)}}\left(c_i^{(1)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_{k,\mathcal{B}_k^{(1)}}^{(1)}\left(q_{l_k}^{[k]}\right)\right] = 0, \tag{5.9}$$

$$\sum_{j=1}^{n_2} p_j^{(2)} U_{0,\mathcal{B}_0^{(2)}}\left(c_j^{(2)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_{k,\mathcal{B}_k^{(2)}}^{(2)}\left(q_{l_k}^{[k]}\right)\right] = 0.$$

Corollary 5.3 *Assume that n_1, n_2 and $(L_k)_{1\leq k\leq P}$ go to infinity and that their ratios are bounded. Take B such as $n_1+n_2+\sum_{k=1}^{P} L_k = o(B)$. Then, under the hypothesis of Theorem 5.2, the likelihood ratio for P products, $r_{n_1,n_2,L_1,...,L_P}(\theta_d)$, is asymptotically $\chi^2(1)$:*

$$r_{n_1,n_2,L_1,...,L_P}(\theta_d) \to \chi^2(1).$$

See the appendix 5.5.3 for the proof. Note in particular that the cardinal B of the incomplete index set must go to infinity quicker than $\max\{n_1, n_2, L_1, ..., L_P\}$. As before, this yields an $(1-\alpha)^{th}$ confidence interval for θ_d such that

$$\left\{\theta_d : r_{n_1,n_2,L_1,...,L_P}(\theta_d) \leq \chi^2_{1-\alpha}(1)\right\}.$$

5.2.5 A faster alternative : Euclidean likelihood

The empirical likelihood program as written in this paper consists in minimizing the Kullback-Leibler distance between a multinomial on the sample $(\widetilde{\mathcal{D}}_1 \times \widetilde{\mathcal{D}}_2)$ and the observed data $(\mathcal{D}_1 \times \mathcal{D}_2)$. Following the ideas of Bertail et al. (2004), we replace the Kullback-Leibler distance by the Euclidean distance (also called the χ^2 distance). When using the Euclidean distance, the objective function becomes

$$\mathbf{l}_{n_1,n_2,L_1,...,L_P}(\theta_d) = \min_{\left\{p_i^{(1)},\, p_j^{(2)},\, w_{l_k}^{[k]},k=1,..,P\right\}} \frac{1}{2}\left[\sum_{i=1}^{n_1}\left(n_1 p_i^{(1)} - 1\right)^2 + \sum_{j=1}^{n_2}\left(n_2 p_j^{(2)} - 1\right)^2 + \sum_{k=1}^{P}\sum_{l_k=1}^{L_k}\left(L_k w_{l_k}^{[k]} - 1\right)^2\right], \tag{5.10}$$

under the approximated model constraints 5.9 and the constraint that each set of weights sum to 1. We get a result equivalent to corollary 5.3 :

Corollary 5.4 *Under the assumptions of Corollary 5.3, the statistic*

$$\mathbf{r}_{n_1,n_2,L_1,...,L_P}(\theta_d) = \mathbf{l}_{n_1,n_2,L_1,...,L_P}(\theta_d) - \inf_{\theta} \mathbf{l}_{n_1,n_2,L_1,...,L_P}(\theta)$$

is asymptotically $\chi^2(1)$.

The proof of this result is given in appendix 5.5.4.

The choice of this distance is closely related to the Generalized Method of Moments (GMM), see Newey & Smith (2004), Bonnal & Renault (2004) for precisions on the links between empirical likelihood and GMM. Instead of logarithms, the optimization program (5.10)

only involves quadratic terms and is then much easier to solve, as shown in appendix 5.5.4. This considerably decreases the computation time, making exploration easier and allowing to test different constraints and models.

A specificity of Euclidean distance is that the weights $p_i^{(1)}$, $p_j^{(2)}$ and $w_{l_k}^{[k]}$ cannot be forced to be positives. Under this constraint, automatically realized for Kullback-Leibler distance, $\mathbf{r}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d)$ is not defined for almost all $\theta_d \in \mathbb{R}$.

The gain in computation time is counter-balanced by a lost in adaptability to the data and to the constraints. Numerical results will be given in the applications for both the Kullback-Leibler and the Euclidean distances. Practical use of these methods shows that Euclidean distance can be used for initial exploration (looking for the most useful constraints for example) and to give first-step estimators. Empirical likelihood can then be used on the final stage, to get precise confidence regions and estimators. The first-step estimators given by Euclidean likelihood can be used as starting values for the empirical likelihood optimization. The following example, using large data sets and a complicated model, illustrates the interest of this strategy.

5.3 Application : Risk assessment for fish and sea product consumption

in this section, our previous theoretical results are used to assess the risk due to the presence of methylmercury (MeHg) in fish and sea products. The risk is characterized by the probability that the exposure to MeHg exceeds the safe dose of 1.6 μg of MeHg per week per kg of body weight, defined by toxicologists over the lifetime. Chronic exposure must be assessed and require reliable estimates of long-term food consumption. Nevertheless, for technical reasons, collecting individual food consumption on a long term is difficult. In France, two main data sets are available. The SECODIP panel collecting long-term household purchases (from 1989 to nowadays) allows the estimation of the chronic probability to be over the PTWI. Unfortunately data only record households' purchase. A first approximation consists in extrapolating household consumptions to individual ones by dividing households' purchase by the family size. The second source of data is the national INCA survey based on short-term consumptions (one-week), which allows to calculate the individual probability to exceed the PTWI on one week. Such a survey does not permit to evaluate precisely chronic exposure but only to extrapolate a one-week consumption for the life time consumption.

Some unpublished preliminary studies show that the use of INCA or SECODIP survey for the exposure estimation to methylmercury gives very different results. Those results are consistent with the literature showing that survey durations influence the percentage of consumers (due to infrequency of purchase) and the level of food intakes among consumers only (Lambe et al., 2000). Numerous methods have been proposed to extrapolate from short-term to long-term intake based on repeated short-term measures in the field of nutrition (see Hoffmann et al., 2002, Price et al., 1996). Another idea developed here, is to combine information from short and long-term consumption surveys using empirical likelihood. The variability of food contamination is also taken into account in our model.

5.3.1 Data description and specific features

Food consumption data

The French "INCA" survey ($r = 1$), carried out by CREDOC-AFSSA-DGAL (1999), records $n_1 = 3003$ individual consumptions during one week. The survey is composed 2 samples : 1985 adults aged 15 years or over and 1018 children aged between 3 to 14 years. The data were obtained during an 11-month period from consumption logs completed by the participants for a period of 7 consecutive days. National representativeness of each subsample (adults,children) was ensured by stratified sampling (region of residence, town size) and by the application of quotas (age, sex, individual professional/cultural category, household size). From this survey, 92 food items were selected with respect to fish or fishery products. This includes fish, fish farming, shellfish, mollusks, mixed dishes, soups and miscellaneous fishery products. Since body weight of all individuals is available, "relative" consumptions are computed by dividing the amount consumed during the week by the body weight.

The proportion of children (34%) in this survey is high compared to the national census (INSEE, 1999) (15%) : it is usually recommended to work on adults and children samples separately. In order to use the two subsamples, we correct this selection bias by adding a margin constraint on the proportion of children (aged between 3 and 14 years) as proposed in (5.1. The additional constraint is

$$\mathbb{E}_{\widetilde{\mathcal{C}}_{n_1}^{(1)}}\left[\mathbb{1}_{3\leq Z_i^{(1)}\leq 14}\right] = 0.15,$$

where $Z_i^{(1)}$ is the age of individual i in the survey $r = 1$ (INCA).

This modifies the form of the dual log-likelihood (5.6) in the part concerning the first survey. It becomes

$$\sum_{i=1}^{n_1} \log\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right) + \lambda_{\text{age}}\left(\mathbb{1}_{3\leq Z_i^{(1)}\leq 14} - 0.15\right)\right\},$$

where λ_{age} is the Kühn and Tücker coefficient associated to the "age" constraint.

The SECODIP panel for fish, from *TNS SECODIP* (http ://www.secodip.fr), is composed of 3211 households surveyed over one year (the 1999 year). In this panel, 24 food groups containing fish or sea products are retained. Individual consumption is created by inputting to each individual the household's purchase divided by the number of persons in the household. We also divide this result by 52 (number of weeks in a year) and 60 (mean body weight). This results into $n_2 = 9588$ individual relative week consumptions.

The differences between the two surveys have many explanations :

- the SECODIP panel is an Household Budget Survey. However Serra-Majem et al. (2003) found that, in general, results from Household Budget Surveys in Canada and Europe agree well with individual dietary data ;
- the SECODIP panel does not account for outside consumptions : members of the panel do not record purchases for outdoor consumptions ;
- the INCA survey is realized in a public health perspective. People could modify their consumption behavior during the survey week in favor of foods they assume to be "healthy" as fish.

All these arguments explain the higher fish consumption in INCA survey. We choose to introduce a coefficient α to scale the SECODIP consumption to account for all these facts introducing an additional model constraint

$$\mathbb{E}(C^{(1)}) = \alpha \mathbb{E}(C^{(2)}).$$

The coefficient α is estimated together with the risk index θ_d, leading to confidence regions for (θ_d, α) calibrated by a $\chi^2(2)$ distribution, i.e. $r_{n_1,n_2,L_1}(\theta_d, \alpha) \to \chi^2(2)$. We then optimize on α for each θ_d to get a profiled likelihood on θ_d.

Contamination data

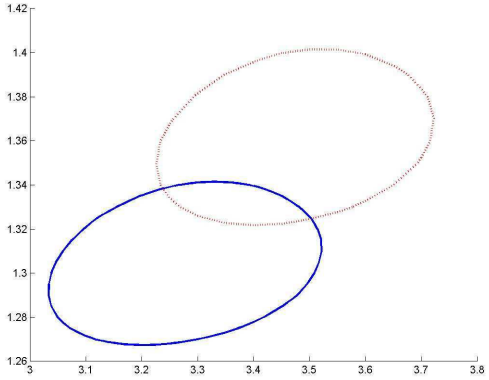
Food contamination data concerning fish and fishery products available on the French market were generated by accredited laboratories from official national surveys performed between 1994 and 2003 by the French Ministry of Agriculture and Fisheries MAAPAR (1998-2002) and the French Research Institute for Exploitation of the Sea (IFREMER, 1994-1998). These $L = 2832$ analytical data are expressed in terms of total mercury in mg/kg of fresh weight. Part of the mercury present in the sea can be transform by microbial activity in its organic form, methylmercury (MeHg), which is the dangerous form to human health. MeHg is present in sea-foods, the highest levels being found in predatory fishes, particularly those at the top of the aquatic food chain. According to Claisse et al. (2001), Cossa et al. (1989), methylmercury levels in fish and fishery products can be extrapolated from the mercury content. For this reason, conversion factors have been applied to the analytical data in order to obtain the corresponding methylmercury (MeHg) concentration in the different foods considered : 0.84 for fish, 0.43 for mollusk and 0.36 for shellfish.

Contamination data are frequently left censored because of the quantification limits of analytical methods. In our sample, we find 7% of censored data for which the levels of mercury were below some detection or quantification limit. We adhere to international recommendations (GEMs/Food-WHO, 1995) and replace the censored values with half the detection or quantification limit. We refer to Bertail & Tressou (2004), Tressou (2006) for further discussion on the impact of left censored level.

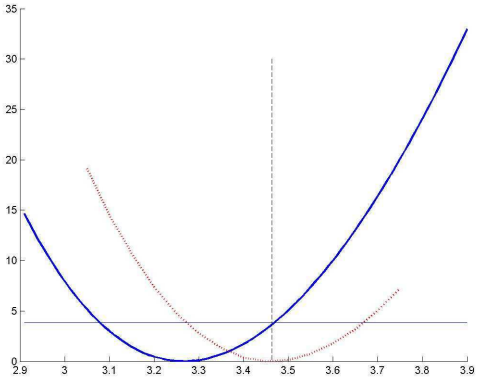
5.3.2 Results when considering one global sea product

We first merge all the products of interest into a single group, “sea products”. Any contamination data is attributed to the total individual consumption of sea products. Calculations can therefore be performed using the complete U-statistics of degree $(1, 1)$.

Figure 5.1(a) shows the two 95% confidence regions for the couple of parameters $(\theta_{1.6}, \alpha)$. We compare the results obtained with and without the constraint on the proportion of children. The unconstrained confidence region for $(\theta_{1.6}, \alpha)$ is marked by a dotted line, the solid line corresponding to the constrained confidence region. We can see that the constraint make the 2 surveys closer (α is smaller, the confidence region is translated to the bottom) and decrease the risk ($\theta_{1.6}$ is smaller, the confidence region is translated to the left). Children



(a) Empirical likelihood confidence region
horizontal axis is $\theta_{1.6}$, vertical axis is α



(b) Empirical likelihood ratio profile
horizontal axis is $\theta_{1.6}$, vertical axis is r_{n_1,n_2,L_1}

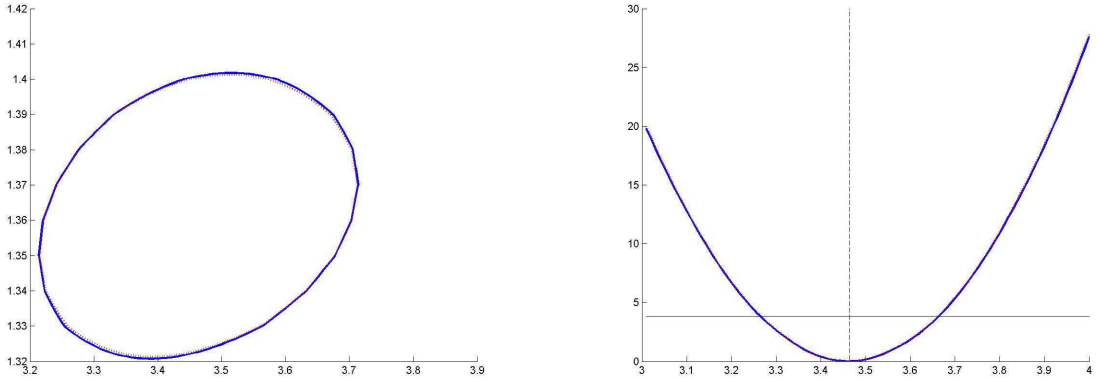
FIG. 5.1 – Empirical likelihood for one product (solid, with age constraint ; dot, without age constraint)

are known to be a more sensitive group to food exposure because of their higher relative consumptions : they eat more compared to their body weight than adults. When adding the age constraint, the discrete probability measure related to the INCA survey, the $\left(p_i^{(1)}\right)_{1\leq i\leq n_1}$, are modified so that children become less influent, which explains the risk reduction and the decrease of α.

Figure 5.1 (b) shows the profiles of the empirical likelihood ratios $(r_{n_1,n_2,L_1}(\theta_{1.6}))$. We get 2 profiles, the dotted line corresponds to the unconstrained case. The horizontal line gives the 95% level of the chi-square distribution $(\chi^2_{95\%}(1))$, limiting the confidence interval for the risk index. The 95% confidence interval for $\theta_{1.6}$ constraining INCA children proportion is [3.08% ; 3.47%] and the risk index estimator is $\theta^*_{1.6} = 3.27\%$. The optimal scaling parameter is $\alpha^* = 1.31$. This is an estimation of the factor to convert individual food purchases of sea products into individual consumptions of sea products.

When the constraint on age is ignored, the estimator of $\theta_{1.6}$ is the arithmetic mean of INCA survey and $\alpha-$scaled SECODIP data (marked by the vertical dotted black line). Indeed, the best correction α is when both means are equal and then the maximum of the likelihood for $\theta_{1.6}$ is this common value. The SECODIP data has then no effect on the value of the estimator but has an effect on the confidence interval : uncertainty is reduced thanks to the large sample of consumption values provided by the SECODIP data.

Euclidean likelihood : The Euclidean distance is not as sharp as the Kullback discrepancy, which is used in the empirical likelihood case. Moreover, the constraint on age being linear and only on the smaller consumption sample INCA, the associated term in the Euclidean likelihood is small in front of the risk index term, which is nonlinear and concerns both consumption samples INCA and SECODIP. The effect of the constraint is thus highly reduced : confidence regions as shown in Figure 5.2 (a) as well as profiles as shown in Figure 5.2 (b) are almost identical. They give results quite close to what is obtained with the constrained empirical likelihood.



(a) Euclidean likelihood confidence region horizontal axis is $\theta_{1.6}$, vertical axis is α

(b) Euclidean likelihood ratio profile horizontal axis is $\theta_{1.6}$, vertical axis is r_{n_1,n_2,L_1}

FIG. 5.2 – Euclidean likelihood for one product (solid, with age constraint ; dot, without age constraint)

5.3.3 Results when considering two products

Products are now grouped into two types of sea products, the first one is fish and the second one is mollusk and shellfish. We have $L_1 = 1541$ values of contamination for the fish group and $L_2 = 1291$ values for the second.Calculation are done using incomplete U-statistics defined in equations (5.7) and (5.8) with a size $B = 10000$. α is here 2-dimensional.

The confidence interval for the risk index is [5.20% ; 5.64%] and the estimator is given by $\theta^*_{1.6} = 5.43\%$. The correction factors on SECODIP data are $\alpha^*_1 = 1.8$ and $\alpha^*_2 = 1.65$. Figure (5.3) shows the profile of the empirical likelihood ratio. The probability calculated when products are considered as a single group is smaller than when products are gathered into two groups (see also Tressou et al., 2004). Consequently in order to improve this risk assessment, it would be interesting to go deeper in the food nomenclature of both surveys to create more groups but it is not possible with the available SECODIP food nomenclature.

5.4 Conclusion

This paper shows how empirical likelihood method can be generalized to combine different sources of data. We apply our theoretical results to assess the risk due to the presence of methylmercury in fish and sea products. We combine the two different main French consumption surveys and some French contamination data in order to estimate a food risk index. Results show that empirical likelihood is a powerful method to build confidence intervals for this risk index using all the available information.

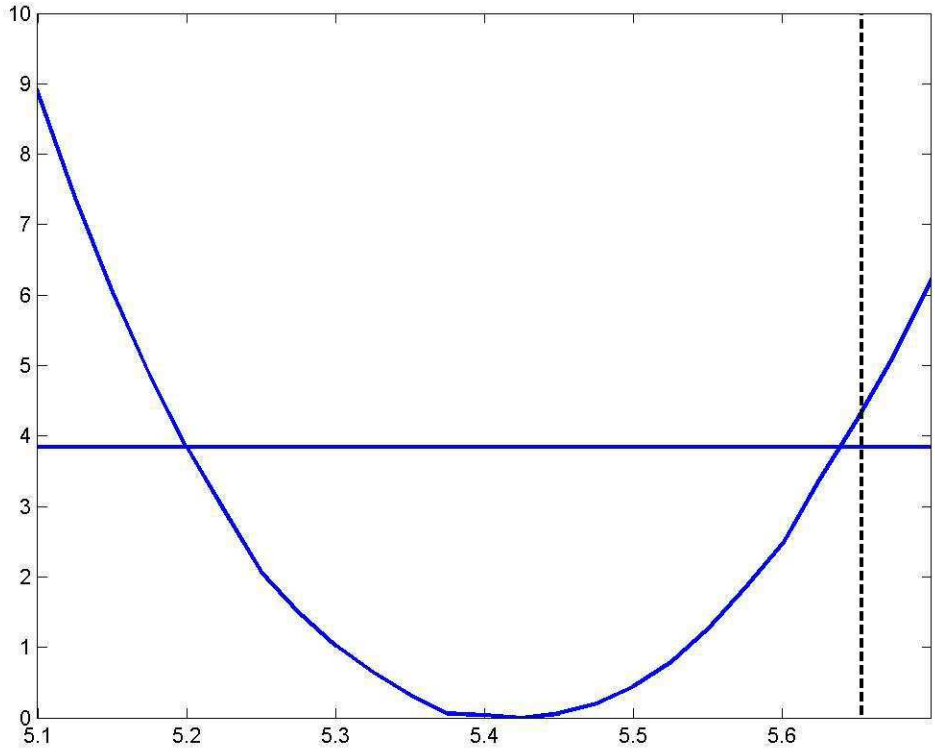


FIG. 5.3 – Empirical likelihood ratio profile for two products with age constraint (horizontal axis is $\theta_{1.6}$ and vertical axis is r_{n_1,n_2,L_1,L_2}

A technical improvement would consist in using a statistical method to disaggregate household purchases into individual "at home" consumptions and correct for the difference between "at home" and total food consumption. Chesher (1997) proposes such a method for the decomposition of household nutritional intakes into individual intakes accounting for outside consumptions. In an empirical likelihood program, this method would require the estimation of a great number of parameters which may cause optimization problems. This kind of methodology could however avoid the use of an ad hoc scaling parameter α between SECODIP and INCA panels. We plan to explore this issue in further works.

From an applied point of view, we obtain with different methods combining the available information that the probability to exceed the PTWI is of the order of 5%. This can be considered as an important risk at a population scale. It also motivates some further works to characterize the at-risk population.

Acknowledgments : *We thank Christine Boizot (INRA-CORELA) for the support she has provided in handling the SECODIP data as well as Jean-Charles Leblanc (AFSSA) for the contamination data. Many thanks also to Patrice Bertail (CREST-LS) for his careful reading of the manuscript. All errors remain ours.*

5.5 Proofs

5.5.1 Proof of Theorem 5.1

When $\mu \in \mathbb{R}$, the theorem can be linked to a special case of Chapter 11.4 of Owen (2001), with $h(X^{(1)}, X^{(2)}, \mu) = (X^{(1)} - \mu)(X^{(2)} - \mu)$.

From now on, we suppose, without lost of generality, that the true value of μ_0 is 0. Write $P = P_1 \otimes P_2$ and $n = \min(n_1, n_2)$. o and $\mathcal{O}$ are all taken in probability. The Lagrangian of our optimization program can be written :

$$\log\left(\prod_{i=1}^{n_1} n_1 p_i^{(1)} \prod_{j=1}^{n_2} n_2 p_j^{(2)}\right) - n_1\lambda_1' \sum_{i=1}^{n_1} p_i^{(1)}(X_i^{(1)} - \mu) - n_2\lambda_2' \sum_{j=1}^{n_2} p_j^{(2)}(X_j^{(2)} - \mu) - \alpha_1\left(\sum_{i=1}^{n_1} p_i^{(1)} - 1\right) - \alpha_2\left(\sum_{j=1}^{n_2} p_j^{(2)} - 1\right).$$

The values of the Lagrange Multipliers depend on μ. To emphasize this dependence for the most important multiplier, we will denote $\lambda_r(\mu)$. We note $\widehat{\mu}$ the point where the maximum of $\log\left(\prod_{i=1}^{n_1} n_1 p_i^{(1)} \prod_{j=1}^{n_2} n_2 p_j^{(2)}\right)$ is realized.

The proof follows the main arguments of Qin & Lawless (1994) : we first check that $\widehat{\mu}$ goes to zero at least at a slow rate ($n^{-1/3}$). Then we use this result to control $\lambda_r(\widehat{\mu})$. This gives us a better approximation of $\widehat{\mu}$ that is sufficient to obtain the asymptotic convergence.

A first control of $\widehat{\mu}$. As pointed previously, at any fixed μ, $l_{n_1,n_2}(\mu)$ can be studied as the sum of two independent suprema : $l_{n_1,n_2}(\mu) = l_{n_1}(\mu) + l_{n_2}(\mu)$. From the proof of Lemma 1 of Qin & Lawless (1994), we have that :

(i) there exists $c_r > 0$ such as, if $||\mu|| \geq n_r^{-1/3}$ then $l_{n_r}(\mu) \geq c_r n_r^{1/3}$,
(ii) $l_{n_r}(0) = \mathcal{O}(\log\log n_r)$,
(iii) if $||\mu|| \leq n_r^{-1/3}$ then $\lambda_r(\mu) = \mathcal{O}(n_r^{-1/3})$.

Using (i), if $||\mu|| \geq n^{-1/3}$, we get that $l_{n_1,n_2}(\mu) \geq cn^{1/3}$, where $c = \min(c_1, c_2)$. Using the equation (ii), $l_{n_1,n_2}(0) = \mathcal{O}(\log\log n_1 + \log\log n_2)$. Then, under the assumption that $\log\log\max(n_1, n_2) = o(n^{1/3})$ and for n large enough, $l_{n_1,n_2}(0) < cn^{1/3}$. Then the optimum of l_{n_1,n_2} is reached at a point $\widehat{\mu}$ such as $||\widehat{\mu}|| \leq n^{-1/3}$.

Control of $\lambda_r(\widehat{\mu})$. Following Owen (2001) Chapter 11.2, we have :

$$\lambda_r(0) = S_r^{-2}(0)(\overline{X}_r - 0) + o(n_r^{-1/2}), \tag{5.11}$$

with $\overline{X}_r = \frac{1}{n_r}\sum_{i=1}^{n_r} X_i^{(r)}$ and $S_r^2(\mu) = \frac{1}{n_r}\sum_{i=1}^{n_r}\left(X_i^{(r)} - \mu\right)\left(X_i^{(r)} - \mu\right)'$, $r = 1, 2$.

We need a similar expression for $\lambda_r(\widehat{\mu})$. As $||\widehat{\mu}|| \leq n^{-1/3}$, we already know by (iii) that $\lambda_r(\widehat{\mu}) = \mathcal{O}(n^{-1/3})$. This allows us to expand the constraint $\sum_{i=1}^{n_r} p_i^{(r)} X_i^{(r)} = \widehat{\mu}$ to obtain equations similar to (11.1) and (11.5) in Owen (2001),

$$0 = \frac{1}{n_r}\sum_{i=1}^{n_r}\frac{X_i^{(r)} - \widehat{\mu}}{1 + \lambda_r'(\widehat{\mu})\left(X_i^{(r)} - \widehat{\mu}\right)} = \overline{X}_r - \widehat{\mu} - S_r^2(\widehat{\mu})\lambda_r(\widehat{\mu}) + R \tag{5.12}$$

with $||R|| \leq \frac{1}{n_r}\sum_{i=1}^{n_r} ||X_i^{(r)} - \widehat{\mu}||^3 ||\lambda_r(\widehat{\mu})||^2 \left|\left|1 - \lambda_r'(\widehat{\mu})\left(X_i^{(r)} - \widehat{\mu}\right)\right|\right|^{-1}$.

Following the lines of the proof of Lemma (11.2) of Owen (2001), since $\mathbb{E}||X_i^{(r)} - \widehat{\mu}||^3 < \infty$, we have $\max_i ||X_i^{(r)} - \widehat{\mu}|| = o(n_r^{1/3})$ and therefore $\lambda_r'(\widehat{\mu})\left(X_i^{(r)} - \widehat{\mu}\right) = o(1)$. Then

$$||R|| = \mathcal{O}(1)\mathcal{O}(n_r^{-2/3})\mathcal{O}(1) = \mathcal{O}(n_r^{-2/3}) = o(n_r^{-1/2}).$$

By (5.12), we have then,

$$\lambda_r(\widehat{\mu}) = S_r^{-2}(\widehat{\mu})(\overline{X}_r - \widehat{\mu}) + o(n_r^{-1/2}). \tag{5.13}$$

A closer control of $\widehat{\mu}$. To find $\widehat{\mu}$, we take the derivative of the Lagrangian with respect to μ, we get

$$n_1\lambda_1(\widehat{\mu}) + n_2\lambda_r(\widehat{\mu}) = 0. \tag{5.14}$$

Replacing the expression of $\lambda_r(\widehat{\mu})$ given by (5.13) into (5.14), we get

$$n_1 S_1^{-2}(\widehat{\mu})(\overline{X}_1 - \widehat{\mu}) + n_2 S_2^{-2}(\widehat{\mu})(\overline{X}_2 - \widehat{\mu}) = o_P(n^{1/2}),$$

and

$$\widehat{\mu} = \left(n_1 S_1^{-2}(\widehat{\mu}) + n_2 S_2^{-2}(\widehat{\mu})\right)^{-1}\left(n_1 S_1^{-2}(\widehat{\mu})\overline{X_1} + n_2 S_2^{-2}(\widehat{\mu})\overline{X}_2\right) + o_P(n^{-1/2}). \tag{5.15}$$

To simplify the statement, we now write $S_r^{-2} = S_r^{-2}(0)$. Since we have $||\widehat{\mu}|| \leq n^{-1/3}$, we get

$$S_r^{-2}(\widehat{\mu}) = S_r^{-2} + o_P(1) \tag{5.16}$$

and finally we have the expression

$$\widehat{\mu} = \left(n_1 S_1^{-2} + n_2 S_2^{-2}\right)^{-1}\left(n_1 S_1^{-2}\overline{X}_1 + n_2 S_2^{-2}\overline{X}_2\right) + o_P(n^{-1/2}) = \mathcal{O}_P(n^{-1/2}).$$

Asymptotic behavior of r_{n_1,n_2}. By chapter 11.2 of Owen (2001), we have :

$$2l_{n_r}(0) = n_r \overline{X}_r' S_r^{-2} \overline{X}_r + o_P(1).$$

As $\widehat{\mu} = \mathcal{O}_P(n^{-1/2})$, a similar expression is also valid for $2l_{n_r}(\widehat{\mu})$

$$2l_{n_r}(\widehat{\mu}) = n_r(\overline{X}_r - \widehat{\mu})' S_r^{-2}(\widehat{\mu})(\overline{X}_r - \widehat{\mu}) + o_P(1)$$

and then by (5.16), we obtain

$$\begin{aligned} 2l_{n_r}(\widehat{\mu}) &= n_r(\overline{X}_r - \widehat{\mu})' S_r^{-2}(\overline{X}_r - \widehat{\mu}) + o_P(1) \\ &= 2l_{n_r}(0) - 2n_r \overline{X}_r' S_r^{-2} \widehat{\mu} + n_r \widehat{\mu}' S_r^{-2} \widehat{\mu} + o_P(1) \end{aligned}$$

Finally, it follows

$$\begin{aligned} r_{n_1,n_2}(0) &= 2\left[l_{n_1,n_2}(\widehat{\mu}) - l_{n_1,n_2}(0)\right] + o_P(1) \\ &= 2\left[l_{n_1}(\widehat{\mu}) - l_{n_1}(0)\right] + 2\left[l_{n_2}(\widehat{\mu}) - l_{n_2}(0)\right] + o_P(1) \\ &= \left[2\left(n_1 \overline{X}_1' S_1^{-2} + n_2 \overline{X}_2' S_2^{-2}\right) - \widehat{\mu}'\left(n_1 S_1^{-2} + n_2 S_2^{-2}\right)\right] \widehat{\mu} + o_P(1) \end{aligned}$$

Using the expression (5.15) of $\widehat{\mu}$, we get

$$\begin{aligned} \widehat{\mu}'\left(n_1 S_1^{-2} + n_2 S_2^{-2}\right) &= \left(n_1 \overline{X}_1' S_1^{-2} + n_2 \overline{X}_2' S_2^{-2}\right)\left(n_1 S_1^{-2} + n_2 S_2^{-2}\right)^{-1}\left(n_1 S_1^{-2} + n_2 S_2^{-2}\right) + o_P(n^{1/2}) \\ &= \left(n_1 \overline{X}_1' S_1^{-2} + n_2 \overline{X}_2' S_2^{-2}\right) + o_P(n^{1/2}), \end{aligned}$$

yielding

$$\begin{aligned} r_{n_1,n_2}(0) &= \left[2\left(n_1 \overline{X}_1' S_1^{-2} + n_2 \overline{X}_2' S_2^{-2}\right) - \left(n_1 \overline{X}_1' S_1^{-2} + n_2 \overline{X}_2' S_2^{-2}\right)\right] \widehat{\mu} + o_P(1) \\ &= \left(n_1 S_1^{-2} \overline{X}_1 + n_2 S_2^{-2} \overline{X}_2\right)' \left(n_1 S_1^{-2} + n_2 S_2^{-2}\right)^{-1} \left(n_1 S_1^{-2} \overline{X}_1 + n_2 S_2^{-2} \overline{X}_2\right) + o_P(1). \end{aligned}$$

The Lindeberg-Feller Central Limit Theorem gives the convergence of $r_{n_1,n_2}(0)$ to a $\chi^2(1)$. The only condition to check is that, for all $\varepsilon > 0$, as $n_1, n_2 \to \infty$,

$$B_{n_1,n_2} = \sum_{i=1}^{n_1} \mathbb{E}_P\left[Y_i^{(1)\prime} Y_i^{(1)} \mathbb{1}_{|Y_i^{(1)}|>\varepsilon}\right] + \sum_{j=1}^{n_2} \mathbb{E}_P\left[Y_j^{(2)\prime} Y_j^{(2)} \mathbb{1}_{|Y_j^{(2)}|>\varepsilon}\right] \to 0, \tag{5.17}$$

with $Y_i^{(r)} = \left(n_1 S_1^{-2} + n_2 S_2^{-2}\right)^{-1/2} S_r^{-2} X_i^{(r)}$. The Y_i and Y_j being i.i.d., we have

$$B_{n_1,n_2} = \mathbb{E}_P\left[n_1 Y_i^{(1)\prime} Y_i^{(1)} \mathbb{1}_{|Y_i^{(1)}|>\varepsilon}\right] + \mathbb{E}_P\left[n_2 Y_j^{(2)\prime} Y_j^{(2)} \mathbb{1}_{|Y_j^{(2)}|>\varepsilon}\right].$$

$Y_i^{(r)}$ is of order $\mathcal{O}_P(n^{-1/2})$ and therefor $\mathbb{1}_{|Y_i^{(1)}|>\varepsilon}$ tends to zero in probability, whereas $n_r Y_i^{(r)\prime} Y_i^{(r)}$ is asymptotically $\chi^2(1)$. By Slutsky's Lemma, the product tends to 0 and then $B_{n_1,n_2} \to 0$. This ends the proof.

5.5.2 Proof of Theorem 5.2

First, we consider the empirical likelihood optimization program for two consumption surveys and one food product. Recall that $U_0(c)$ and $U_1^{(r)}(q)$ are dependent of θ_d :

$$U_0(c) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\{q_l c > d\}} - \theta_d \text{ and } U_1^{(r)}(q) = \frac{1}{n_r}\sum_{i=1}^{n_r} \mathbb{1}_{\left\{qc_i^{(r)} > d\right\}} - \theta_d, \text{ for } r = 1, 2.$$

The program is to maximize

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{l=1}^{L} w_l, \tag{5.18}$$

under the constraints

$$\sum_{i=1}^{n_1} p_i^{(1)} = 1, \quad \sum_{j=1}^{n_2} p_j^{(2)} = 1, \quad \sum_{i=1}^{L} w_l = 1,$$
$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l) = 0, \quad \sum_{j=1}^{n_2} p_j^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l) = 0.$$

To carry out this optimization, we take the log of (5.18). This forces the weights to be positive. The difference between these constraints and the nonlinear ones defined in equation (5.2) is $o\left(N_r^{-1/2}\right)$ where $N_r = n_r + L$.

First approximation of the weights We need an approximation of the weights to control the order of the Lagrange Multipliers. In order to obtain such an approximation, we consider an easier program. As the expectation of $U_0\left(c_i^{(1)}\right)$, $U_0\left(c_j^{(2)}\right)$ and $U_1^{(r)}(q_l)$ are zero, we consider the likelihood $\prod_{i=1}^{n_1} \widetilde{p}_i^{(1)} \prod_{j=1}^{n_2} \widetilde{p}_j^{(2)} \prod_{l=1}^{L} \widetilde{w}_l$ under the additional constraints :

$$\sum_{i=1}^{n_1} \widetilde{p}_i^{(1)} U_0\left(c_i^{(1)}\right) = 0, \quad \sum_{j=1}^{n_2} \widetilde{p}_j^{(2)} U_0\left(c_j^{(2)}\right) = 0, \quad \sum_{l=1}^{L} \widetilde{w}_l U_1^{(r)}(q_l) = 0, \ r = 1, 2. \tag{5.19}$$

The constraints are thus splitted in two, each contraint concerning only one set of weights. The optimization program is therefor divided in 3 independent sub-programs, the 2 first on the $\widetilde{p}_i^{(r)}$'s being the classical empirical likelihood for the mean and the last one on the $\widetilde{w}_l$'s having 2 constraints. As done in Qin & Lawless (1994), Theorem 1, we have a control on the order of the optimal weights of each sub-program :

$$\widetilde{p}_i^{(r)} = \frac{1}{n_r} \frac{1}{1 + t_r U_0\left(c_i^{(r)}\right)} \qquad \text{with } t_r = \mathcal{O}(n_r^{-1/2})$$
$$\widetilde{w}_l = \frac{1}{L} \frac{1}{1 + (\tau_1, \tau_2)' \left(U_1^{(1)}(q_l), U_1^{(2)}(q_l)\right)} \qquad \text{with } \tau_r = \mathcal{O}(L^{-1/2}).$$

The optimum of this new program, which is given by the optimum on each of the 3 sub-programs, is smaller than (5.18), because we added constraints :

$$\prod_{i=1}^{n_1} \widetilde{p}_i^{(1)} \prod_{j=1}^{n_2} \widetilde{p}_j^{(2)} \prod_{l=1}^{L} \widetilde{w}_l \leq \prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{l=1}^{L} w_l.$$

This means that weights in (5.18), the $p_i^{(1)}$'s, $p_j^{(2)}$'s, and w_l's, are closer to $1/n_1$, $1/n_2$ and $1/L$ than the $\widetilde{p}_i^{(1)}$'s, $\widetilde{p}_j^{(2)}$'s, and $\widetilde{w}_l$'s. Notice that

$$\begin{aligned}\sum_{i=1}^{n_r} \left| \widetilde{p}_i^{(r)} - \frac{1}{n_r} \right| \left| U_0\left(c_i^{(r)}\right) \right| &= \frac{1}{n_r} \sum_{i=1}^{n_r} \left| \frac{1}{1 + t_r U_0\left(c_i^{(r)}\right)} - 1 \right| \left| U_0\left(c_i^{(r)}\right) \right| \\ &\leq |t_r| \frac{1}{n_r} \sum_{i=1}^{n_r} \left| U_0\left(c_i^{(r)}\right) \right|^2 + o(t_r) = \mathcal{O}\left(n_r^{-1/2}\right). \end{aligned} \tag{5.20}$$

Then, coming back to the original program (5.18), we have :

$$\begin{aligned}\left| \sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) \right| &\leq \left| \frac{1}{n_r} \sum_{i=1}^{n_r} U_0\left(c_i^{(r)}\right) \right| + \left| \sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) - \sum_{i=1}^{n_r} \frac{1}{n_r} U_0\left(c_i^{(r)}\right) \right| \\ &\leq \left| \frac{1}{n_r} \sum_{i=1}^{n_r} U_0\left(c_i^{(r)}\right) \right| + \sum_{i=1}^{n_r} \left| p_i^{(r)} - \frac{1}{n_r} \right| \left| U_0\left(c_i^{(r)}\right) \right| \\ &\leq \left| \frac{1}{n_r} \sum_{i=1}^{n_r} U_0\left(c_i^{(r)}\right) \right| + \sum_{i=1}^{n_r} \left| \widetilde{p}_i^{(r)} - \frac{1}{n_r} \right| \left| U_0\left(c_i^{(r)}\right) \right| \\ &= \mathcal{O}\left(n_r^{-1/2}\right), \end{aligned}$$

by standard CLT arguments on $U_0\left(c_i^{(r)}\right)$ and (5.20).

By similar arguments on w_l, we have

$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) = \mathcal{O}\left(n_1^{-1/2}\right), \quad \sum_{j=1}^{n_2} p_j^{(2)} U_0\left(c_j^{(2)}\right) = \mathcal{O}\left(n_2^{-1/2}\right),$$

$$\text{and } \sum_{l=1}^{L} w_l U_1^{(r)}(q_l) = \mathcal{O}\left(L^{-1/2}\right). \tag{5.21}$$

Lagrangian The optimization program (5.18) can be rewritten

$$\max_{w_l,\, \gamma_a\, ,p_i^{(r)},\, \gamma_r,\, \lambda_r} \mathbf{H}\left(w_l,\ \gamma_a\ , p_i^{(r)},\ \gamma_r,\ \lambda_r\right)$$

with :

$$\mathbf{H}\left(w_l,\ \gamma_a,\ p_i^{(r)},\ \gamma_r,\ \lambda_r\right) = \log\left(\prod_{i=1}^{n_1} p_i^{(1)} \prod_{i=1}^{n_2} p_i^{(2)} \prod_{l=1}^{L} w_l\right) - \gamma_1\left[\sum_{i=1}^{n_1} p_i^{(1)} - 1\right] - \gamma_2\left[\sum_{i=1}^{n_2} p_i^{(2)} - 1\right] - \gamma_a\left[\sum_{i=1}^{L} w_l - 1\right] - \lambda_1\left[\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l)\right] - \lambda_2\left[\sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l)\right].$$

Using $\frac{\partial \mathbf{H}}{\partial p_i^{(r)}} = \frac{1}{p_i^{(r)}} - \gamma_r - \lambda_r U_0\left(c_i^{(r)}\right) = 0$ and the similar expression for $\frac{\partial \mathbf{H}}{\partial w_l}$ gives that

$$p_i^{(r)} = \frac{1}{\gamma_r + \lambda_r U_0\left(c_i^{(r)}\right)} \text{ and } w_l = \frac{1}{\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)}. \tag{5.22}$$

Note that we also have

$$\sum_{i=1}^{n_r} p_i^{(r)} \frac{\partial \mathbf{H}}{\partial p_i^{(r)}} = n_r - \gamma_r - \lambda_r \sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_r^{(1)}\right) = 0 \tag{5.23}$$

and using the constraints, we get that

$$0 = \sum_{i=1}^{n_1} p_i^{(1)} \frac{\partial \mathbf{H}}{\partial p_i^{(1)}} + \sum_{i=1}^{n_2} p_i^{(2)} \frac{\partial \mathbf{H}}{\partial p_i^{(2)}} + \sum_{i=1}^{L} w_l \frac{\partial \mathbf{H}}{\partial w_l} = n_1 + n_2 + L - \gamma_1 - \gamma_2 - \gamma_a. \tag{5.24}$$

The problem (5.18) can be rewritten using (5.22) and (5.24) in the dual form

$$\sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_a \in \mathbb{R} \\ n_1+n_2+L-\gamma_1-\gamma_2-\gamma_a=0}} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \log\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right)\right\} + \sum_{j=1}^{n_2} \log\left\{\gamma_2 + \lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ + \sum_{l=1}^{L} \log\left\{\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right\}.$$

Furthermore, combining (5.23) with

$$\sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) = \mathcal{O}(n_r^{-1/2})$$

gives that $\gamma_r = n_r + v_r$, where v_r is given by $\lambda_r \cdot \mathcal{O}(n_r^{-1/2})$ and then

$$p_i^{(r)} = \frac{1}{n_r + v_r + \lambda_r U_0\left(c_i^{(r)}\right)} \quad \text{and} \quad w_l = \frac{1}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)}.$$

Let us consider the case of the w_l. Adapting Owen's proof, equation (5.21) for $r = 1$ combined with (5.22) yields for the $(w_l)_l$ constraint

$$\begin{aligned}\mathcal{O}\left(L^{-1/2}\right) &= \sum_{i=1}^{L} w_l U_1^{(1)}(q_l) = \sum_{i=1}^{L} \frac{U_1^{(1)}(q_l)}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)} \\ &= \sum_{i=1}^{L} \frac{U_1^{(1)}(q_l)}{L} - \frac{1}{L}\sum_{i=1}^{L} \frac{\left[-v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right] \cdot U_1^{(1)}(q_l)}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)}, \\ &= \overline{U_1^{(1)}} - \frac{\lambda_1}{L}\sum_{i=1}^{L} w_l \left[U_1^{(1)}(q_l)\right]^2 - \frac{\lambda_2}{L}\sum_{i=1}^{L} w_l U_1^{(1)}(q_l) U_1^{(2)}(q_l) + \frac{v_1 + v_2}{L}\sum_{i=1}^{L} w_l U_1^{(1)}(q_l),\end{aligned}$$

where $\overline{U_1^{(1)}} = L^{-1}\sum_{i=1}^{L} U_1^{(1)}(q_l)$. The last term is equivalent to $(v_1 + v_2)\mathcal{O}(L^{-3/2})$ and then can be included in $\mathcal{O}\left(L^{-1/2}\right)$

$$\overline{U_1^{(1)}} = \frac{\lambda_1}{L}\sum_{i=1}^{L} w_l \left[U_1^{(1)}(q_l)\right]^2 + \frac{\lambda_2}{L}\sum_{i=1}^{L} w_l U_1^{(1)}(q_l) U_1^{(2)}(q_l) + \mathcal{O}\left(L^{-1/2}\right)$$

Using Owen's arguments, we get

$$\overline{U_1^{(1)}} + \mathcal{O}\left(L^{-1/2}\right) = \frac{\lambda_1}{L}\overline{\left[U_1^{(1)}\right]^2} + \frac{\lambda_2}{L}\overline{U_1^{(1)}U_1^{(2)}}, \quad \overline{U_1^{(2)}} + \mathcal{O}\left(L^{-1/2}\right) = \frac{\lambda_2}{L}\overline{\left[U_1^{(2)}\right]^2} + \frac{\lambda_1}{L}\overline{U_1^{(1)}U_1^{(2)}},$$

where $\overline{\left[U_1^{(1)}\right]^2} = L^{-1}\sum_{i=1}^{L}\left[U_1^{(1)}(q_l)\right]^2$ and $\overline{U_1^{(1)}U_1^{(2)}} = L^{-1}\sum_{i=1}^{L} U_1^{(1)}(q_l)U_1^{(2)}(q_l)^2$. This can be rewritten :

$$\begin{pmatrix}\lambda_1 \\ \lambda_2\end{pmatrix} = L\begin{bmatrix}\overline{\left[U_1^{(1)}\right]^2} & \overline{U_1^{(1)}U_1^{(2)}} \\ \overline{U_1^{(1)}U_1^{(2)}} & \overline{\left[U_1^{(2)}\right]^2}\end{bmatrix}^{-1}\begin{pmatrix}\overline{U_1^{(1)}} + \mathcal{O}\left(L^{-1/2}\right) \\ \overline{U_1^{(2)}} + \mathcal{O}\left(L^{-1/2}\right)\end{pmatrix}. \tag{5.25}$$

As the empirical variance-covariance matrix convergences to the non-degenerated variance-covariance matrix $\mathbb{E}_{\mathbb{P}}\left[\left(U_1^{(1)}\ U_1^{(2)}\right)'\left(U_1^{(1)}\ U_1^{(2)}\right)\right]$ and as $\overline{U_1^{(1)}}$ and $\overline{U_1^{(2)}}$ are of order $\mathcal{O}(L^{-1/2})$, it follows that λ_1 and λ_2 are of order $\mathcal{O}(L^{1/2})$.

When considering $p_i^{(r)}$ instead of w_l, the calculus are easier and we get in a similar fashion

$$\lambda_r = n_r\left(\overline{\left[U_0^{(r)}\right]^2}\right)^{-1}\overline{U_0^{(r)}} + \mathcal{O}(n_r^{1/2}), \tag{5.26}$$

where $\overline{U_0^{(r)}} = n_r^{-1}\sum_{i=1}^{n_r} U_0\left(c_i^{(r)}\right)$ and $\overline{\left[U_0^{(r)}\right]^2} = n_r^{-1}\sum_{i=1}^{n_r}\left[U_0\left(c_i^{(r)}\right)\right]^2$.

Now that we control the size of λ_r at the optimum for both n_r and L with (5.26) and (5.25), the arguments of Owen (2001) chapter 11.4 and the proof of Qin & Lawless (1994) give the expected convergence of $r_{n_1,n_2,L}(\theta_d) = 2\left(l_{n_1,n_2,L}(\theta_d) - l_{n_1,n_2,L}(\widehat{\theta}_d)\right)$ to a $\chi^2_{(1)}$.

5.5.3 Proof of Corollary 5.3

The preceding arguments may be generalized to the case of P products. We give here a proof for $P = 2$. The incomplete U-statistics related to the contamination of the 2 products are denoted $U^{(r)}_{a,B}$ and $U^{(r)}_{b,B}$. The difference between the incomplete and the complete statistics are of order $\mathcal{O}(B^{-1/2})$, and then does not affect the asymptotic results. The program consists in maximizing

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{i=1}^{n_2} p_i^{(2)} \prod_{l=1}^{L_a} w_l^{[a]} \prod_{l=1}^{L_b} w_l^{[b]},$$

under the constraints :

$$\sum_{i=1}^{n_1} p_i^{(1)} = 1, \quad \sum_{i=1}^{n_2} p_i^{(2)} = 1, \quad \sum_{i=1}^{L_a} w_l^{[a]} = 1, \quad \sum_{i=1}^{L_b} w_l^{[b]} = 1,$$

$$\sum_{i=1}^{n_1} p_i^{(1)} U_{0,\mathcal{B}_0^{(1)}}\left(c_i^{(1)}\right) + \sum_{l=1}^{L_a} w_l^{[a]} U^{(1)}_{a,\mathcal{B}_a^{(1)}}\left(q_l^{[a]}\right) + \sum_{l=1}^{L_b} w_l^{[b]} U^{(1)}_{b,\mathcal{B}_b^{(1)}}\left(q_l^{[b]}\right) = 0,$$

$$\sum_{i=1}^{n_2} p_i^{(2)} U_{0,\mathcal{B}_0^{(2)}}\left(c_i^{(2)}\right) + \sum_{l=1}^{L_a} w_l^{[a]} U^{(2)}_{a,\mathcal{B}_a^{(2)}}\left(q_l^{[1]}\right) + \sum_{l=1}^{L_b} w_l^{[b]} U^{(2)}_{b,\mathcal{B}_b^{(2)}}\left(q_l^{[b]}\right) = 0.$$

with for $r = 1, 2$ and $k = a, b$, we can check with similar arguments that

$$\sum_{i=1}^{n_r} p_i^{(r)} U_{0,\mathcal{B}_0^{(r)}}\left(c_i^{(r)}\right) = \mathcal{O}\left(n_r^{-1/2}\right), \quad \sum_{l=1}^{L_k} w_l U^{(r)}_{k,\mathcal{B}_k^{(r)}}\left[q_l^{[k]}\right] = \mathcal{O}\left(L_k^{-1/2}\right).$$

We get as before for $r = 1, 2$ and $k = a, b$

$$p_i^{(r)} = \frac{1}{n_r + v_r + \lambda_r U_{0,\mathcal{B}_0^{(r)}}\left(c_i^{(r)}\right)} \quad \text{and} \quad w_l^{[k]} = \frac{1}{L_k + v_k + \lambda_1 U^{(1)}_{k,\mathcal{B}_k^{(1)}}\left(q_l^{[k]}\right) + \lambda_2 U^{(2)}_{k,\mathcal{B}_k^{(2)}}\left(q_l^{[k]}\right)},$$

with $v_1 + v_2 + v_a + v_b = 0$ and the proof follows the same lines as for 1 product.

5.5.4 Proof of Corollary 5.4

The objective function of the program is now

$$\frac{1}{2} \min_{\left\{p_i^{(1)}, p_i^{(2)}, w_{l_k}^{[k]}, k=1,..,P\right\}} \sum_{r=1}^{2} \sum_{i=1}^{n_r} \left(n_r p_i^{(r)} - 1\right)^2 + \sum_{k=1}^{P} \sum_{l_k=1}^{L_k} \left(L_k w_{l_k}^{[k]} - 1\right)^2.$$

We get then simpler expressions, which allow to reach explicit solutions for the weights.

For the sake of simplicity, we present the results for two consumptions surveys and one food product ($P = 1$), the optimization program can be rewritten

$$\frac{1}{2} \min_{\left\{p_i^{(1)}, p_i^{(2)}, w_l\right\}} \sum_{i=1}^{n_1} \left(n_1 p_i^{(1)} - 1\right)^2 + \sum_{i=1}^{n_2} \left(n_2 p_i^{(2)} - 1\right)^2 + \sum_{l=1}^{L} \left(L w_l - 1\right)^2,$$

under the constraints :

$$\sum_{i=1}^{n_1} p_i^{(1)} = 1, \quad \sum_{i=1}^{n_2} p_i^{(2)} = 1, \quad \sum_{i=1}^{L} w_l = 1,$$

$$\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l) = 0, \quad \sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l) = 0.$$

Define

$$\begin{aligned}\mathbf{H}(\cdot) = \frac{1}{2}\sum_{i=1}^{n_1}\left(n_1 p_i^{(1)} - 1\right)^2 + \frac{1}{2}\sum_{i=1}^{n_2}\left(n_2 p_i^{(2)} - 1\right)^2 + \frac{1}{2}\sum_{l=1}^{L}\left(L w_l - 1\right)^2 \\ - \lambda_1\left[\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l)\right] - \lambda_2\left[\sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l)\right] \\ - \gamma_1\left[\sum_{i=1}^{n_1} p_i^{(1)} - 1\right] - \gamma_2\left[\sum_{i=1}^{n_2} p_i^{(2)} - 1\right] - \gamma_a\left[\sum_{i=1}^{L} w_l - 1\right].\end{aligned}$$

Then the first order condition of the optimization program leads to

$$\partial\mathbf{H}/\partial p_i^{(r)} = n_r(n_r p_i^{(r)} - 1) - \gamma_r - \lambda_r U_0\left(c_i^{(r)}\right) = 0$$

so that we get $p_i^{(r)} = \dfrac{1}{n_r} + \dfrac{\gamma_r + \lambda_r U_0\left(c_i^{(r)}\right)}{n_r^2}$. As the weights sum to 1, we have

$$1 = \sum_{i=1}^{n_r} p_i^{(r)} = 1 + \frac{\gamma_r + \lambda_r \overline{U_0^{(r)}}}{n_r} \text{ so } \gamma_r = -\lambda_r \overline{U_0^{(r)}},$$

and finally

$$p_i^{(r)} = \frac{1}{n_r} + \lambda_r \frac{U_0\left(c_i^{(r)}\right) - \overline{U_0^{(r)}}}{n_r^2} \quad \text{and} \quad w_l = \frac{1}{L} + \lambda_1 \frac{U_1^{(1)}(q_l) - \overline{U_1^{(1)}}}{L^2} + \lambda_2 \frac{U_1^{(2)}(q_l) - \overline{U_1^{(2)}}}{L^2}.$$

The constraints can be rewritten

$$\begin{aligned}\overline{U_0^{(1)}} + \overline{U_1^{(1)}} + \lambda_1\left[\frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L}\right] + \lambda_2 \frac{Cov\left(U_1^{(1)}, U_1^{(2)}\right)}{L} = 0, \\ \overline{U_0^{(2)}} + \overline{U_1^{(2)}} + \lambda_2\left[\frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L}\right] + \lambda_1 \frac{Cov\left(U_1^{(1)}, U_1^{(2)}\right)}{L} = 0,\end{aligned}$$

where $\mathbb{V}$ and Cov denote the empirical variance operator, $\mathbb{V}(X) = \overline{(X^2)} - \left(\overline{X}\right)^2$, and the covariance operator, $Cov(X,Y) = \overline{(X \cdot Y)} - \overline{X} \cdot \overline{Y}$. These terms do not depend on θ_d.

Note that $\overline{U_0^{(r)}} = \overline{U_1^{(r)}}$ by definition of these U-statistics and write it $\overline{U^{(r)}}$. The optimum is then reached at

$$\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = -2 \begin{bmatrix} \frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L} & \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} \\ \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} & \frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}.$$

Thus the optimal value can be computed explicitly. Finally, replacing the values of the weights and the λ's in the optimization program, we get :

$$l(n_1, n_2, L) = \frac{4}{2} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}' \begin{bmatrix} \frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L} & \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} \\ \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L} & \frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}.$$

Case $P > 1$:
We also use this framework for the 2 surveys 2 products context. The form of the Euclidean likelihood is almost the same, with $\overline{U^{(r)}} := \overline{U_0^{(r)}} = \overline{U_1^{(r)}} = \overline{U_2^{(r)}}$ and we easily get by straightforward calculous

$$l(n_1, n_2, L_1, L_2) =$$
$$\frac{9}{2} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}' \begin{bmatrix} \frac{\mathbb{V}\left(U_0^{(1)}\right)}{n_1} + \frac{\mathbb{V}\left(U_1^{(1)}\right)}{L_1} + \frac{\mathbb{V}\left(U_2^{(1)}\right)}{L_2} & \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L_1} + \frac{Cov\left(U_2^{(1)},U_2^{(2)}\right)}{L_2} \\ \frac{Cov\left(U_1^{(1)},U_1^{(2)}\right)}{L_1} + \frac{Cov\left(U_2^{(1)},U_2^{(2)}\right)}{L_2} & \frac{\mathbb{V}\left(U_0^{(2)}\right)}{n_2} + \frac{\mathbb{V}\left(U_1^{(2)}\right)}{L_1} + \frac{\mathbb{V}\left(U_2^{(2)}\right)}{L_2} \end{bmatrix}^{-1} \begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix},$$

and the result follows.

Chapitre 6

Ideal body weight and social norm : evidence from French data using empirical discrepancies

6.1 Introduction

Trends in obesity have become a major health concern in France as well as in many developed and developing countries. According to the WHO health standards, individuals are overweight when their Body Mass Index (BMI), which equals their weight in kilograms divided by their height in meters squared, is over 25. In 1990, 29.7% of French adults aged over 15 were overweight against 37.5% in 2002 (OECD Health Data 2005). Overweight and obesity are associated with a number of comorbidities, such as heart diseases and diabetes, which represent an increasing financial burden. Although overweight and obesity are less prevalent in France than in the UK, the US or Greece, the medical cost of obesity was already 1 billion euros in 1991 (Detournay *et al.*, 2000). This suggests an obvious arena for policy intervention through taxes, subsidies and information (Boizot-Szantaï and Etilé, 2005). However, calibrating public intervention requires that social interactions in the consumer's weight control problem be identified, because they induce social multiplier effects (Clark and Oswald, 1998). In this perspective, this paper investigates social interactions through social norms.

We use the French survey "Enquête sur les Conditions de Vie des Ménages". This general survey covers roughly 10 000 people in 5000 households. In 2001, it included a detailed health interview of one person in each household with information on actual body weight and height, as well as a measure of ideal body weight. Using the latter, we ask whether ideal BMI predicts attitudes towards food, and whether social norms influence ideal BMI.

The economic setting is a microeconomic model, in which the individual utility has two arguments : (i) a private return function, which depends on food consumption and a numeraire good ; (ii) a loss function, which increases in the differences between actual BMI and various reference points. We consider three types of reference points : social norms sustained by preference interactions (Manski, 2000) ; habitual weight, which induces costs of adjustment ; an idiosyncratic bliss point.

We assume that the loss function represents satisfaction with weight, and we expect individuals to focus on this argument of their utility function when asked to declare their desired weight. Hence, a first key assumption is that ideal body weight maximises weight satisfaction. A second assumption is that costs of adjustment may be convex or concave, with a discontinuous first derivative around habitual weight (see Suranovic *et al.*, 1999, for a similar assumption regarding smoking cessation). Last, the loss function is the sum of the costs of deviation from the reference points. The model has two predictions : first, individuals with high marginal adjustment costs around habitual weight are more likely to declare their habitual weight as their ideal ; second, for individuals with low marginal adjustment costs around habitual weight, ideal weight is a weighted average of the latter, their social norm, and their idiosyncratic bliss point.

Investigating the role of social norms requires that groups of 'significant' others be appropriately defined, and that a measure of social norms in these groups be constructed. Using findings from social psychology, we argue that averaging ideal BMI in groups of 'significant' others yields a proxy measure of social norms on body shape. Gender, age and occupation are

used to define these groups, although education and occupation are found to be interchangeable. Individuals are 'significant' to each other if they have the same sex and occupation, and their age difference is lower than 5 years. A spatial weight matrix is then constructed to compute a proxy measure of social norms.

We are not able to propose a structural test of the model because marginal adjustment costs can not be identified. We just observe that, for about 40% of the sample, actual and ideal BMI are equals, so that actual BMI is a good proxy for habitual BMI. Those individuals who would like to gain weight (10% of the sample) are likely to have specific health conditions. For the remaining 50% - those individuals who want to slim -,we estimate an equation that specifies ideal BMI as a function of social norms, actual BMI (our proxy for habitual BMI) and individual characteristics (following the model's second prediction). We identify this equation by instrumenting actual BMI and our measure of social norms. We use the estimator Continuously updated GMM (CUP-GMM) proposed by Bonnal & Renault (2001), which is linked with the Smoothed Empirical Likelihood (SEL) estimator of Kitamura *et al.* (2004). It exploits optimally the moment conditions implied by the exclusion restrictions. The CUP-GMM estimator is easier to implement than the SEL estimator, because it is based on the χ^2 divergence, instead of the Kullback-Leibler divergence. We find that it is generally more efficient than the GMM when instruments are not weak.

There are two important results. First, simple regressions of attitudes toward foods on ideal BMI and social norm reveal that social norm has no effect *per se* while ideal BMI does, actual BMI being held constant. Second, the elasticity of individual ideal BMI to the social norm is lower than 0.9.

The paper is organized as follows. Section 2 discusses the main findings from the literature. Section 3 presents the data and explains why ideal body weight may be used to measure social norms. Section 4 develops the theoretical model. Section 5 develops the empirical specification and outlines the identification issues. Section 6 presents the new estimator. Estimation results follow in Section 7. Section 8 concludes.

6.2 Obesity and social norms

6.2.1 Key findings from the economic literature

Economic explanations of the obesity epidemic focuses essentially on the role of two factors : food prices, *i.e.* the price of calorie intakes, and sedentariness, *i.e.* the price of calorie expenditures. The full price of a meal has fallen since forty years, because the costs of primary food products have declined, as well as those of food preparation (see *inter alia* Cutler *et al.*, 2003). Using American micro data, Lakdawalla and Philipson (2002) find that reductions in food prices may account for 40% of the obesity epidemic in the U.S. While time spent on preparing meals and eating at home has not reduced in France, long-period time series show a decrease in the prices of energy-dense food relatively to the price of fruits and vegetables (Combris *et al.*, 2006, Warde *et al.*, 2006). Hence, the cost of an healthy diet is now much higher than that of a fat- and sugar-rich diet, which clearly contributes to the epidemic (Darmon *et al.,* 2004 ; Drewnowski et Darmon, 2004). Furthermore, as emphasized

by Philipson and Posner (1999), while most individuals were paid to exercise in the agricultural and industrial societies, this is no more the case in post-industrial societies with public welfare. Hence, the price of calorie expenditure has risen.[1] Yet, considering Lakdawalla and Philipson's results, a large part of the individual variance has still to be explained, and social norms are candidate variables for this. In this perspective, Burke and Heiland (2005) propose a dynamic model in which social norms on BMI for period t depend on the actual BMI distribution at time $t-1$. Social norms act as a social multiplier : a price decrease has a direct positive effect on calorie intakes, which moves the BMI distribution towards the right ; the social multiplier effect is produced by the subsequent increase in the social norm, which reduces the cost of over-eating for the individual. Using simulation methods, the authors find that the social multiplier effect increases the skewness of the BMI distribution, which is consistent with time trends observed in American data. Hence, studying the role of social norms is of particular importance.

Until the beginning of the nineties, norms were considered by mainstream economics as preference variables, and nuisance parameters for the empirical analysis of public policies. Yet, a number of norms are backed by explicit interactions on the labour and marriage markets. This is shown by a number of empirical papers, which define the medical threshold for overweight and obesity as the social norm. For instance, Averett and Korenman (1996) and Cawley (2004) find significant differences in economic status (income and hourly pay) for white obese women in America. Averett and Korenman show that this differential is essentially accounted for by differences in marriage probabilities and spouse's earnings. Using French data, Paraponaris *et al.* (2005) suggest that time spent in an unemployment spell is positively correlated with the body mass index.[2] Various sociological experiments have shown that overweight is considered by recruiters and supervisors as a signal for unobservable predispositions, such as laziness, lack of self-control (gluttony), an higher probability of illness etc. As long as these defaults are believed to be negatively correlated with productivity (the so-called 'halo' effect), an informational discrimination can arise, especially in the race for job positions that require self-control, dynamism, and leadership.[3] This may explain why the obesity wage penalty is more important in high income occupations than in low ones (Carr and Friedman, 2005).[4]

While overweight is penalized, thinness yields a number of benefits. New markets have developed upon the ideal of thinness, which has largely diffused in the western middle and

[1] Using aggregate American data, Cutler *et al.* (2003) cast some doubt on this explanation by showing that the share of the population in energy-demanding jobs is quite stable since twenty years. To our knowledge, accurate data on trends in energy expenditures are not available for France.

[2] However, this study fails to control for unobserved heterogeneity, and the presence of a "third factor" accounts for the negative weight-labour correlations in some social groups (Cawley, 2004)

[3] Veblen (1899) noted that "those members of respectable society who advocate athletic games commonly justify their attitude on this head to themselves and to their neighbours on the ground that these games serve as an invaluable means of development. They not only improve the contestant's physique, but it is commonly added that they also foster a manly spirit, both in the participants and in the spectators." Since the end of the nineteenth century, sport and many activities of body control are associated with positive spiritual values.

[4] An alternative explanation emphasizes that high income positions are generally proposed with employer-sponsored insurance. The obesity wage penalty would simply represent the employer's risk premium (Bhattacharya and Bundorf, 2005).

upper classes after World War II.[5] The high division of labour in the sector of entertainment renders possible for individuals from lower social classes to succeed, even if they have no other capital than their body. This is sometimes an incentive to deviate from the norms of one's own social background. Last, the medicalisation of obesity has also led individuals to consider their weight as a risk factor for health. One standard prediction of the health demand model, is that individuals with higher lifecycle wage profiles have higher incentives to remain healthy : sticking to the medical standards of thinness is an health investment.

6.2.2 Why social norms may produce endogenous effects ?

From our point of view, the empirical literature has not paid enough attention to the following subtle distinction between social norms on body weight. First, norms may arise from expectations interactions. A standard example is the statistical discrimination on the job market, whereby obese individuals are expected to be less productive because data available to human resources services show correlations between obesity and productivity (Manski, 2000). More generally, facing the same constraints, either in terms of body characteristics demanded on the labour market, or in terms of returns to health investments, individuals with similar occupations are likely to have similar perceptions of ideal body weight. Following Manski (1993), this is a correlated effect, which *in a sociological perspective* produces standards of behaviours rather than social norms.

For sociologists, norms emerge from preference interactions with 'significant' others, be they individuals with the same social status and from the same social class or not.[6] They consider that, in the long run, external constraints are internalized by the social group, in the sense that they induce institutionalized role expectations that are sustained by reactions of group's members.[7] For instance, while Bourdieu (1979) recognizes that social norms on body weight are generated by market mechanisms, because the body is a component of human capital,[8] he also emphasizes that they are internalized by individuals in social schemes of perceptions.[9] Following this view, social norms produce endogenous effects and not only

[5] Kersh and Morone (2002) state that thinness became a prevalent standard around 1890 in the U.S., before its medicalisation, and in close connection with Christian moral concern. But the 60s were probably the key moment (standards the 60s were certainly the moment of history, where controlling one's body (through birth control *inter alia)* became the norms : thinness was a symbol of freedom for middle and upper-classes women.

[6] To our knowledge, the notion of 'significant others' dates back to Festinger (1954), who proposed the hypothesis that individuals make permanent comparisons to the behaviours and the expectations of members of a reference group, which determine feelings of satisfaction.

[7] See Horne (2001) for sociological perspectives on norms and on the process of internalization (behavioural regularities becoming role expectations and ultimately values/preferences).

[8] *"The interest the different classes have in self-presentation, the attention they devote to it, their awareness of the profits it gives and the investment of time, effort, sacrifice and car which they actually put into it are proportionate to the chances of material or symbolic profit they can reasonably expect from it. More precisely, they depend on the existence of a labour market in which physical appearance may be valorized in the performance of the job itself or in professional relations ; and on the differential chances of access to this market and the sectors of this market in which beauty and deportment most strongly contribute to occupational value." (p. 202)*

[9] *"Tastes in food also depend on the idea each class has of the body and of the effects of food on the body, that is on its strength, health and beauty [...] whereas the working class are more attentive to the*

correlated effects. The latter do not generate a social multiplier, while the former do, which is important in a public policy perspective.

A key question is how to measure social norms. One solution is to rely on observations of norm-related sanctions. But regarding body shape, norms are generally enforced *via* the emotions triggered by other's or one's own view on oneself, and there are no formal enforcement mechanisms in the group (Amadieu, 2002). As in Stutzer and Lalive (2004), we rather use a proxy measure of "individual's beliefs about how one ought to behave".

6.3 Constructing a proxy measure of social norms

6.3.1 Ideal and actual BMIs in the data

This paper uses data from the survey "Enquête Permanente sur les Conditions de Vie des Ménages" EPCV2001), which was carried out by the INSEE (the French National Statistical Agency) in 2001. It contains information at both the household and the individual levels, and one randomly-drawn individual in each household answered a detailed health questionnaire. The starting sample consists of 5194 individuals in the same number of households. Given the presence of missing values, 3972 individuals are kept for the analysis. All variables are presented with descriptive statistics in Table A.1., Appendix A.

All measures of height and actual and ideal body weights are self-declared. We are not able to correct for declaration biases.[10] Throughout the paper, we use the BMI as a measure of body shape. BMI is a good predictor of overweight- and obesity-related morbidity. It also adjusts body weight for differences in height, which gives a more precise picture of individual's body shape than body weight alone. It thus takes into consideration both medical and aesthetic concerns with body weight. Our measure of ideal BMI is based on the following question *"What is the weight you would like to reach or keep ?"*. The distributions of actual and ideal BMIs are shown in Figures A.1. and A.2. of Appendix A.

Figure A.3. in Appendix A reports the distributions of the difference and the contrast between actual and ideal BMIs in the left and the right frames respectively. For 40% of the sample, actual and ideal body weights are equal. Only 6% of the sample wants to gain weight. They have peculiar characteristics. For instance, they are more prone to mental disorders and negative affects : 19% of them take a psychiatric treatment against 14% in the whole sample, 25% feel lonely (vs. 15%), etc. Third, more than 50% of the sample would like to loose weight. In the sample, the average ideal BMI is about 0.95 time the average actual BMI. For those who want to slim, this coefficient is about 0.9, which is the ratio factor that

strength of the (male) body than its shape, and tend to go for products that are both cheap and nutritious, the professions prefer products that are tasty, health-giving, light and not fattening.[...] It follows that the body is the most indisputable materialization of class taste, which it manifests in several ways. It does this first in the seemingly most natural features of the body, the dimensions (volume, height, weight) and shapes (round or square, stiff or supple, straight or curved), which express in countless ways a whole relation to the body, i.e., a way of treating it, caring for it, feeding it, maintaining it, which reveals the deepest dispositions of the habitus." (p. 190).

[10]Data with both self-reported and measured weights are not yet available for France. See Chou *et al.* (2001), Lakdawalla and Philipson (2002) and Cawley (2004) for correction procedures in US data.

is typically found for the *whole* population in American data (Burke and Heiland, 2005). Hence, there is probably less discrepancy between actual and ideal BMIs in France than in the US.

6.3.2 Using ideal BMI to measure social norms

The economics of social interaction often posits that norms operate through individual expectations of average behaviour. However, as noted by Elster (1989), social norms are "sustained by the feelings of embarrassment, anxiety, guilt and shame that a person suffers at the prospect of violating them.[...]. Social norms have a grip on the mind that is due to the strong emotion they can trigger". Some psychological studies have found that these feelings are associated with the discrepancies between the representations of attributes individuals actually possess and the attributes they would like to possess *or* 'significant' others believe they ought to possess (see for instance Tangney *et al.*, 1998). Hence, ideal BMI is a measure of both individual aspirations and *what ought to be*. Sociologists define norms as 'ought' statement that are based on shared representations of ideal attributes (Horne, 2001). Following these arguments, this paper constructs a measure of social norms on body shape by averaging the perceptions of ideal BMI in the group of 'significant' others.

For this view to be consistent, there are three conditions. First, the individual-specific social norm (*i.e.* the mean reference-group ideal BMI) should affect individual perceptions of ideal BMI. Second, the latter should have a causal effect on actual BMI. Third, conditionally on ideal and actual BMIs, norms must have no effect on food choices. Our data set does not allow us to test the second and third conditions, so that we essentially focus on the first one. However, Section 7 reports some results on the correlations between ideal BMI and attitudes towards food.

6.3.3 Defining appropriate groups of 'significant' others

As noted by Manski (1993), it is worth having some a priori knowledge of who are the 'significant' others. French sociologists supposes *a priori* that occupation and gender are the main variables that explain social differentiation in actual and ideal body shapes (Bourdieu, 1979, Regnier, 2006). Differences in food habits and foodways by occupation groups and cohorts have also been extensively described (Grignon and Grignon, 1999). Qualitative observations suggest that the lower social classes face a dissonance between the norms of eating and the standards of healthy eating, which partly explains their risk excess for overweight and obesity (see for instance Lhuissier, 2006). However, although gender and occupation are important, other sociodemographic characteristics such as age, education or localization may play some role.

A necessary condition for a social norm to exist in a social group is that its members' perceptions of ideal body weight are correlated. Individuals belong to the same social group if they are close in the social space spanned by the attributes that define the groups. These attributes $\{Q^1, .., Q^r, ..., Q^R\}$ are necessarily discrete, because a social group exists only insofar as they are entrance hurdles (Goblot, 1925). As a consequence, the Q^rs are not commensurable. We have to work separately on each Q^r, and to estimate auto-correlation in ideal

BMI for each Q^r distance metric. Let W^* denote the ideal BMI, and take the normalization $w^* = \frac{W^* - \overline{W^*}}{\sigma_{\overline{W^*}}}$, an interesting statistics is the auto-covariance at distance 0 for the whole sample :

$$\hat{F}(0, Q^r) = \frac{\sum_i \sum_j 1\{Q_i^r = Q_j^r, i \neq j\} w_i^* w_j^*}{\sum_i \sum_j 1\{Q_i^r = Q_j^r, i \neq j\}}. \tag{6.1}$$

Auto-correlation at distance 0 tells us whether individuals with similar characteristics Q^r tend to have similar ideal BMI. Hence, the higher is $\hat{F}(0, Q^r)$, the more likely it is that social norms matter.

We first compute $\hat{F}(0, Q^r)$ by gender for age, education and type of residential area separately. $\hat{F}(0, Q^r)$ is the highest for the age distance metric. Borrowing a procedure from Conley and Topa (2002), we then regress ideal BMI by gender on polynomials of age. Then, we examine the auto-correlation at distance 0 in the residuals. The results are reported in Figures B1 and B2 for men and women respectively. The black squares represent average point estimates. The upper and lower bounds of the 95% confidence regions are the 5^{th} and 95^{th} percentiles of bootstrap estimates (with 250 bootstrap replications). Auto-correlation at distance 0 is significantly positive for the education distance metric, both for men and for women (the first vertical line). It is not the case for men, when one considers occupation (the second vertical line in Figure B2).[11] However, we can repeat the procedure, and regress ideal BMI on age and occupation dummies, then take the residuals and examine their auto-correlation. There is no more auto-correlation for the education distance metric. At every stages, autocorrelation for the localization distance metric is not significant. Hence, using gender, age and occupation is sufficient to capture social clustering in individual perceptions of ideal body shape, although education and occupation are interchangeable. We thus stay in line with existing literature in sociology . Individuals are connected in the social space if : (i) they are of same sex ; (ii) they have no more than 5 years of age difference ; (iii) they have the same occupation.[12] Section 4 uses this definition to construct a spatial weight matrix.

Section 3 now proposes an economic setting to guide the empirical analysis.

6.4 An economic model of ideal body shape

This section connects the previous perspectives on ideal body shape and social norms with the economic literature on social interactions and adjustment costs. It relies on the conjecture that the structure of the questionnaire induces a framing effect : as ideal body weight is asked just after actual weight has been recalled, the discrepancy between actual and ideal BMIs is interpreted as a measure of satisfaction with weight or body shape.

More precisely, we consider an agent whose action set at time t is summarized by the consumptions of food F_t and non-food goods C_t. They affect the BMI W_t, through a weight

[11] As this is a two-step method, we should have bootstrapped the whole procedure. This is left for a future version of the paper.

[12] The age criterion is somewhat arbitrary. Ideally, the age "window" should vary by gender and cohort. This is left for a future version of the paper.

production equation :

$$W_t = w(F_{t-1}, C_{t-1}, W_{t-1}). \tag{6.2}$$

We assume that the utility function $U(.)$ is separable into weight satisfaction $WS(W_t)$ on the one hand, and satisfaction from other commodities (including food) $CS(F_t, C_t)$ on the other hand. Weight satisfaction is a function of actual BMI and various reference points. The distinction between one's own view and others' view on ideal body shape is represented by a partition of weight satisfaction into the private and social losses produced by the deviation from the various reference points. There are three types of reference points : social norms, habitual weight which produces adjustment costs, and idiosyncratic bliss points.

6.4.1 Social norms

Weight satisfaction is the sum of private and social returns. The latter are captured by $v(W_t; W_t^g)$ where W_t^g is the social norm. Following Lakdawalla and Philipson (2002) and Burke and Heiland (2005), social returns are functions of W_t only, are concave, and peaks at a reference weight W_t^g : the closer is W_t from W_t^g, the higher is individual's well-being. A key assumption is that W_t and W_t^g are complementary in the sub-utility function $v(.)$, so that utility is comparison-concave, and its concavity depends on the marginal cost of deviating from the reference point (Clark and Oswald, 1998). As in Akerlof (1997), we consider a quadratic loss function :

$$v(W_t; W_t^g) = -\frac{1}{2}\gamma^g \left(W_t - W_t^g\right)^2, \tag{6.3}$$

where γ^g is positive and might be specific to the social group g the individual belongs to.

6.4.2 Habitual weight and adjustment costs

Loosing weight is not easy mainly because body weight is produced by a set of consumption habits (foodways, exercise, smoking and drinking behaviours essentially), which induce adjustment costs. For instance, adopting healthy foodways requires a number of knowledge such as how to buy healthy food products, how to cook them properly etc. There is a selection effect, whereby many individuals are unable to change their habits on the long-term. There are also specific cognitive adjustment costs when one follows hypo-caloric slimming diets. During moments of high-awareness, dieters generally avoid answering to their basic caloric needs (as signalled by the sensations of hunger and satiety). In this case, the body interprets calorie restrictions as a threat, protects its fat reserves, and send signals that induce losses of control especially during moments of low awareness (Heatherton *et al.*, 1993, Basdevant, 1998).[13]

We thus introduce the adjustment cost function $\Gamma(.)$ to take these specific private returns into consideration. $\Gamma(.)$ is quadratic :

[13]Laboratory experiments that present food to dieters and non-dieters generally observe that tempting food alone does not defeat dieters' motivations, but it does so when it is associated with other external factors such as emotional arousal (Herman and Polivy, 2003). Lack of self-control may reinforce in individuals negative feelings about themselves, making a success in the future more unlikely.

$$\Gamma(W_t, W_t^h) = \begin{cases} -\left[\frac{1}{2}\gamma_2^{h-}(W_t - W_t^h)^2 + \gamma_1^{h-}(W_t - W_t^h)\right] & if \quad W_t \leq W_t^h, \\ -\left[\frac{1}{2}\gamma_2^{h+}(W_t - W_t^h)^2 + \gamma_1^{h+}(W_t - W_t^h)\right] & if \quad W_t > W_t^h, \end{cases} \tag{6.4}$$

$$= \begin{cases} -\left[\frac{1}{2}\gamma_2^{h-}\left(W_t - \left(W_t^h - \frac{\gamma_1^{h-}}{\gamma_2^{h-}}\right)\right)^2 - \frac{1}{2}\frac{\left(\gamma_1^{h-}\right)^2}{\gamma_2^{h-}}\right] & if \quad W_t \leq W_t^h, \\ -\left[\frac{1}{2}\gamma_2^{h+}\left(W_t - \left(W_t^h - \frac{\gamma_1^{h+}}{\gamma_2^{h+}}\right)\right)^2 - \frac{1}{2}\frac{\left(\gamma_1^{h+}\right)^2}{\gamma_2^{h+}}\right] & if \quad W_t > W_t^h. \end{cases}$$

The main characteristics of adjustment costs is that their right and left first derivatives may not be continuous at the habitual weight level $W_t^h : \gamma_1^{h+} \geq 0 \geq \gamma_1^{h-}$. For most individual, slimming only induces adjustment costs so that $\gamma_2^{h+} = \gamma_1^{h+} = 0$, and $\gamma_1^{h-} < 0$. Individuals with chronic illnesses such as cancer or AIDS have problems for gaining weight : for them, $\gamma_1^{h+} > 0$. Adjustment costs are asymmetric and may be convex, as in the problem of smoking quits (Suranovic *et al.* 1999). Convex adjustment costs are observed when a small decrease in calorie intakes - or equivalently in actual BMI - yields a dramatic loss of utility. Hence, the sign of γ_2^h is not fixed *a priori*.

6.4.3 Idiosyncratic reference points

A concave loss function $u(.)$ which peaks at an individual reference point W_t^p, captures the remaining idiosyncratic variations in ideal BMI :

$$u(W_t; W_t^p) = a(F_t, C_t) - \frac{1}{2}\gamma^p(W_t - W_t^p)^2. \tag{6.5}$$

Note that γ^p is positive.

6.4.4 A mesurement equation for ideal BMI

Given a static income constraint $\pi_F F_t + C_t = I_t$, where I_t is income, the consumer's utility at time t is :

$$\begin{aligned} V(F_t) &= U\left[CS(F_t, I - \pi_F F_t); WS(W_t)\right] \\ with \quad WS(W_t) &= -\frac{1}{2}\gamma^g\left(W_t - W_t^g\right)^2 - \frac{1}{2}\gamma^p(W_t - W_t^p)^2 + \Gamma(W_t, W_t^h), \end{aligned} \tag{6.6}$$

with respect to F_t. The following equivalence can be noted :

$$WS(W_t) = -\frac{1}{2}S\left(W_t - W_t^*\right)^2,$$

with

$$W_t^* = \frac{1}{S}\left[\gamma^p W_t^p + \gamma^g W_t^g + 1_{\{W_t \leq W_t^h\}}\left(\gamma_2^{h-}W_t^h - \gamma_1^{h-}\right) + 1_{\{W_t > W_t^h\}}\left(\gamma_2^{h+}W_t^h - \gamma_1^{h+}\right)\right], \tag{6.7}$$

$$S = \gamma^p + \gamma^g + 1_{\{W_t \leq W_t^h\}}\gamma_2^{h-} + 1_{\{W_t > W_t^h\}}\gamma_2^{h+},$$

where $1_{\{W_t \leq W_t^h\}}$ equals 1 if $W_t \leq W_t^h$ and 0 otherwise.

The key conjecture underlying the paper is that having to declare one's actual body weight just before one's ideal weight induces a framing effect : individuals focuses on weight satisfaction, which is $-\frac{1}{2}S\left(W_t - W_t^*\right)^2$ in the individual utility function, and the private returns from actions are separable in the pleasure of food consumption and the utility of body shape. Food consumption enters in $CS(.)$, while body shape enters in $WS(.)$. As such, when asked to declare their ideal body weight, individuals focus on weight satisfaction : ideal BMI maximises weight satisfaction, and equation (6.7) can be considered as a measurement equation that links the ideal BMI to the main preference parameters.

There are four situations (the first and second cases are illustrated by the Figures C.1. and C.2. in Appendix C.). Let $W_t^{ref} = \frac{\gamma^p}{\gamma^p+\gamma^g}W_t^p + \frac{\gamma^g}{\gamma^p+\gamma^g}W_t^g$:

– If $W_t^{ref} < W_t^h$ and $-(\gamma^p+\gamma^g)(W_t^h - W_t^{ref}) < \gamma_1^{h-}$ then the marginal cost of slimming at W_t^h is lower than its marginal benefit in terms of conformity to the reference weight. Ideal BMI is then

$$\begin{aligned} W_t^* &= \frac{1}{S}\left[\gamma^p W_t^p + \gamma^g W_t^g + \gamma_2^{h-} W_t^h - \gamma_1^{h-}\right], \\ S &= \gamma^p + \gamma^g + \gamma_2^{h-}. \end{aligned} \tag{6.8}$$

– If $W_t^{ref} < W_t^h$ and $0 \geq -(\gamma^p+\gamma^g)(W_t^h - W_t^{ref}) > \gamma_1^{h-}$ then the marginal cost of slimming is greater than the marginal benefit of conformity so that ideal BMI is simply habitual BMI.

– Symmetrically, when $W_t^{ref} > W_t^h$ and $-(\gamma^p+\gamma^g)(W_t^h - W_t^{ref}) \geq \gamma_1^{h+}$ then :

$$\begin{aligned} W_t^* &= \frac{1}{S}\left[\gamma^p W_t^p + \gamma^g W_t^g + \gamma_2^{h+} W_t^h - \gamma_1^{h+}\right], \\ S &= \gamma^p + \gamma^g + \gamma_2^{h+}. \end{aligned} \tag{6.9}$$

– When $W_t^{ref} > W_t^h$ and $\gamma_1^{h+} > -(\gamma^p+\gamma^g)(W_t^h - W_t^{ref}) \geq 0$ then ideal and habitual BMIs are equal.

Since actual weight is the best measure for habitual weight at time t, the shape of adjustment costs around habitual weight explains why ideal and actual BMIs are equal for many individuals.[14] Before estimating (6.8) in Section 7, Sections 5 and 6 present the econometric methods.

6.5 Econometric specification

6.5.1 Model specification

The theoretical model suggests a two-steps strategy for the empirical analysis : in the first step, model the inequality conditions that lead to equations (6.8) and (6.9) ; in a second step, estimate these linear equations with a correction for the selection bias. Unfortunately, it is

[14] Another obvious reason is that there are measurement errors ('heaping' effects for instance) in answers about ideal body weight.

well-known that the first-step selection equation is identified only if there are some variables that determine selection but do not enter equations (6.8) and (6.9). The selection conditions derived from the model do not provide such variables. Moreover, absent a good measure of W_t^p, the structural model can not be estimated. In the current version of the paper, we focus on the sub-sample of individuals who want to slim. We start from the following linear specification for ideal BMI, which is the structural equation for the sub-sample :

$$W_t^* = \alpha^g W_t^g + \beta^g W_t^h + (1 - \alpha^g - \beta^g) W_t^p - \gamma_1^{h-}, \tag{6.10}$$

$$\alpha^g = \frac{\gamma^g}{\gamma^p + \gamma^g + \gamma_2^{h-}}, \beta^g = \frac{\gamma_2^h}{\gamma^p + \gamma^g + \gamma_2^{h-}}. \tag{6.11}$$

Let Q be the variables that define group membership (age, gender and occupation), and $\Psi(Q)$ the interaction dummy that defines social group membership. $\mathbb{E}(W_t^*|\Psi(Q) = g)$ is our measure of W_t^{*g}. Dropping the time index (we only have cross-section data), a measurement equation is therefore :

$$W^{*g} = k_1 \mathbb{E}(W^*|\Psi(Q) = g) + k_2 + \eta^{*g}, \tag{6.12}$$

where η^{*g} is a random measurement error term with mean zero, and k_1 and k_2 are parameters for the structural measurement errors. While equation (6.10) will be estimated on the sub-sample of individuals who want to slim, the expectation $\mathbb{E}(W^*|\Psi(Q) = g)$ is taken over the whole sample.

The best predictor of W^h for the econometrician is actual weight W, so that a measurement equation for W^h is :

$$W^h = k_3 W + k_4 + \eta^{*h}, \tag{6.13}$$

where η^{*h} is a measurement error with mean zero.

The idiosyncratic reference point W^p is a function of observable and unobservable individual characteristics. Let H represents individual variables that could affect perceptions of ideal body shape (such as income, area of residence, marital status etc.) ; $\mathbb{E}(H|\Psi(Q))$ are contextual effects, whereby individuals have similar perceptions due to the average within-group observable characteristics[15]. We assume that :

$$W^p = \overline{\delta}\mathbb{E}(H|\Psi(Q)) + \delta H + \eta^{*p}.$$

The measurement biases $k_4 + \eta^{*h}$ and $k_2 + \eta^{*g}$ are specified as linear functions of observable or unobservable individual or group-level characteristics, so that a general econometric counterpart of equation (6.10) is for individual i :

$$W_i^* = \sum_{g=1}^{G} \alpha^g 1\{i \in g\} k_1 \mathbb{E}(W^*|\Psi(Q) = g) + \sum_{g=1}^{G} \beta^g 1\{i \in g\} k_3 W_i + \overline{\delta}\mathbb{E}(H_i|\Psi(Q_i)) + \delta H_i + \eta_i,$$

where η_i is an error-term with mean zero. It captures the effect of unobservable individual characteristics. When it is correlated with Q ($\mathbb{E}(\eta_i|Q_i) \neq 0$), there are correlated effects,

[15]Imagine for instance that there is some segregation by social class on the marriage market. Then, ideal body weight in this social class may depend on the average rate of singles, which determines somewhat how competitive is the class-specific marriage market.

whereby agents in the same group behave similarly because they have "similar unobserved characteristics or face similar institutional environments" (Manski, 1993). Obviously, the impact of market-based social norms depicted in Section 2 may be interpreted in terms of correlated effects when they only represent external economic constraints that are not internalized by individuals as preference parameters.

We have defined the social norm W^{*g} as the mean reference-group ideal BMI. For each individual i, it will be estimated by averaging W^* over the other group's members.[16] Using results from Section 3, we construct a $N \times N$ spatial weights matrix D, which specifies for each observation i the set of her neighbours. Basically, this is a matrix of 0 and 1, and $d_{ij} = 1$ if i and j are members of the same social group (and $d_{ii} = 0$ by convention). We standardize the elements of this matrix by $\sum_j d_{ij}$. Then, a $N \times 1$ vector $\hat{W}^{*g}$ containing in row i a first-stage estimates of W^{*g} for individual i is constructed as : $\hat{W}^{*g} = DW^*$. Note that even if the estimation sub-sample will not include individuals whose ideal and actual BMIs are equal, these observations will be used to construct D. Equation (6.10) becomes in matrix notation :

$$W^* = \overline{\alpha} DW^* + \overline{\beta} W + \overline{\delta} \mathbb{E}(H|\Psi(Q)) + \delta H + \eta, \tag{6.14}$$

where DW^* equals $\mathbb{E}(W^*|\Psi(Q))$, $\overline{\alpha} = k_1\alpha$, $\overline{\beta} = k_3\beta$.

6.5.2 Identification issues

The specification raises several identification issues (*cf.* the discussion in Manski, 1993). First, the α^g are not identified, because by definition W^{*g} takes only one value for group g, so that there is no within group variation of social interaction effects. Hence, we have to impose the restriction that $\forall g, \alpha^g = \alpha$ and as a consequence $\forall g, \quad \beta^g = \beta$.

Second, taking expectations of (6.14) with respect to $\Psi(Q)$ reveals that DW^* is a function of $\mathbb{E}(H|\Psi(Q))$ (and $\mathbb{E}(\eta|\Psi(Q))$ if there are correlated effects). The collinearity between these variables implies that the effect of $\mathbb{E}(W^*|\Psi(Q))$ is not identified without further assumptions. As it is often the case in models of social interactions, we will assume that there are no contextual effects : $\overline{\delta} = 0$. The model becomes :

$$W^* = \overline{\alpha} DW^* + \overline{\beta} W + \delta H + \eta. \tag{6.15}$$

Third, following Manski, the model is identified if $\mathbb{E}(\eta|\Psi(Q)) = 0$ (no correlated effects). As this is unlikely to be the case (at least because contextual effects are omitted), we have to instrument $DW^* = \mathbb{E}(W^*|\Psi(Q))$. Actual weight might also be endogenous ($\mathbb{E}(\eta|W) \neq 0$), because unobservable individual characteristics such as tendencies to cognitive restrictions have simultaneously a negative effect on ideal body weight and a positive effect on actual weight. Actual weight will be instrumented by the education levels, assuming that conditionally to actual weight education has no direct effect on individual's ideal body weight. We

[16] i is excluded to avoid an obvious source of endogeneity.

will use the same set of instruments for the social norm. Since there are four education levels for two variables, the model is formally over-identified.

To estimate (6.15), two methods are implemented. As a baseline estimator, we use a Generalized Method of Moments (GMM) that exploits orthogonality conditions between the set of instruments and the residuals. We also propose a method from the empirical likelihood literature, that constructs from the instruments an optimal set of moment conditions : the Continuously updated GMM (CUP-GMM) method proposed by Bonnal & Renault (2001). This method has two key advantages : estimations are more precise and asymptotic normality has not to be used for computing p-values. Section 6 below presents the empirical likelihood method.

6.6 Econometric method

6.6.1 Empirical likelihood estimators

Suppose we have an i.i.d. data set $Y_1, \cdots, Y_n$ whose p.d.f. $f_{\theta,\tau}(y)$ depends on parameters θ and τ. The latter is a nuisance parameter and θ is the parameter of interest. To find θ, a parametric likelihood method can be applied, if one is willing to assume that the p.d.f belongs to a specified parametric distribution family. The likelihood of the data reads :

$$v(\theta, \tau) = \prod_{i=1}^{n} f_{\theta,\tau}(Y_i). \tag{6.16}$$

The estimator of (θ, τ) is $(\tilde{\theta}, \tilde{\tau}) = \arg\max\{\log v(\theta, \tau)\}$. One key problem here is that the econometrician has to impose strong restrictions on the distribution of the data. For instance, to estimate linear equations such as $Z = X\theta + \varepsilon$, one can take $f_{\theta,\tau}(y) = \phi\left(\frac{y-x\theta}{\sigma}\right)$ where $y = (z, x)$, ϕ is the standard normal p.d.f. and σ is a nuisance parameter. For the sake of simplicity, we suppose that X is nonrandom.

Empirical likelihood has been designed as a means of relaxing these restrictions when θ has to satisfy the following moment condition : $\mathbb{E}[m(Y, \theta)] = 0$. Actually, the more general choice for $f_{\theta,\tau}(y)$ is the multinomial density, because it has as many degrees of freedom as there are observations :

$$f_{\theta,\tau}(y) = \begin{cases} q_i & \text{if } \exists i,\ y = Y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{6.17}$$

with $0 < q_i < 1$ and $\sum q_i = 1$, the corresponding probability is $Q = \sum_{i=1}^{n} q_i \delta_{Y_i}$, where δ_Y is the Dirac measure at Y. Here, we have $\tau = (q_1, q_2, ..., q_n)$ and the dependence of $f_{\theta,\tau}(y)$ in θ appears through the moment equation(s). This lead to the Empirical Likelihood $V(\theta)$,

which is defined as a function of θ only (Owen, 2001) :

$$\begin{aligned} V(\theta) &= \max_{\tau} \{V(\theta,\tau) \quad | \quad \mathbb{E}[m(Y,\theta)] = 0\}, \\ &= \max_{\sum_{i=1}^n f_{\theta,\tau}(Y_i) m(Y_i,\theta)=0} \prod_{i=1}^n f_{\theta,\tau}(Y_i), \\ &= \max_{\sum_{i=1}^n q_i m(Y_i,\theta)=0} \prod_{i=1}^n q_i. \end{aligned} \tag{6.18}$$

The estimator is then given by $\widehat{\theta} = \text{argmax}\{\log V(\theta)\}$.

To test the moment condition, one can build a parametric likelihood ratio :

$$r(\theta_0) = \frac{\max_\tau \left\{v(\theta,\tau) \quad \middle| \quad \int m(Y,\theta_0) f_{\theta_0,\tau}(y) = 0\right\}}{\max_{(\theta,\tau)} \{v(\theta,\tau)\}} \tag{6.19}$$

which is asymptotically χ^2_d, where d is the dimension of m. The equivalent for empirical likelihood is

$$\begin{aligned} R(\theta_0) &= \frac{\max_\tau \{V(\theta,\tau) \,|\textstyle\sum_{i=1}^n q_i m(Y_i,\theta_0) = 0\}}{\max_{(\theta,\tau)} \{V(\theta,\tau)\}}, \\ &= \frac{V(\theta_0)}{n^{-n}}. \end{aligned} \tag{6.20}$$

Again this ratio is asymptotically χ^2_d. If the parametric likelihood ratio is to be used as a test of the moment condition under the specification of the p.d.f., the empirical likelihood ratio yields a confidence region for θ_0 under the moment condition. θ is then in the confidence region with level $1-\alpha$ if $R(\theta)$ is smaller than the $1-\alpha$-quantile of a χ^2_d. The urge difference is that no parametric assumption has to be done on the p.d.f. $f_{\theta,\tau}$. Furthermore, additional moment conditions can be easily handled in this framework. Estimators, confidence intervals or tests can be efficiently achieved by using empirical likelihood.

The main technical difficulty, evaluating the empirical likelihood $V(\theta)$ at any given θ, is resolved by a Lagrangian program :

$$\begin{aligned} \log V(\theta) &= \max_{\tau} \{\log V(\theta,\tau) \quad | \quad \mathbb{E}[m(Y,\theta)] = 0\}, \\ &= \max_{\substack{\sum_{i=1}^n q_i m(Y_i,\theta)=0 \\ \sum_{i=1}^n q_i - a = 0}} \left\{\log\left(\prod_{i=1}^n q_i\right)\right\}, \\ &= \max_{\lambda,\mu} \sum_{i=1}^n \log(q_i) - n\lambda \sum_{i=1}^n q_i m(Y_i,\theta) - \mu\left(\sum_{i=1}^n q_i - 1\right), \end{aligned} \tag{6.21}$$

which lead to $q_i^* = \frac{1}{n(\lambda^* m(Y_i,\theta)+1)}$ where λ^* is the optimal multiplier and depends on θ. To find $\widehat{\theta}$, one has then to maximize over θ

$$\log V(\theta) = \sum_{i=1}^n \log\left(\frac{1}{n(\lambda^* m(Y_i,\theta)+1)}\right). \tag{6.22}$$

At this point, it is interesting to note that

$$\begin{aligned} \log(R(\theta)) &= n\log(n) + \log V(\widehat{\theta}), && (6.23)\\ &= n\log(n) + \log\left(\prod_{i=1}^{n} q_i^*\right), && (6.24)\\ &= K(Q^*, P_n) \end{aligned}$$

where K is the Kullback-Leibler discrepancy, P_n is the empirical probability measure on the data set $(\sum_{i=1}^n \frac{1}{n}\delta_{Y_i})$ and $Q^* = \sum_{i=1}^n q_i^*\delta_{Y_i}$. is the multinomial probability measure obtained after the optimization of (6.21). Bertail *et al.* (2006b) show that the asymptotic convergence property of $2\left(n\log(n) + \log V(\widehat{\theta})\right)$ to a χ^2_d, i.e. of $2K(Q^*, P_n)$, is conserved if we change the discrepancy. Bertail *et al.* (2006a) propose to use a penalized version of the Kullback-Leibler discrepancy, the Quasi-Kullback :

$$K_\varepsilon(Q^*, P_n) = (1-\varepsilon)\left(n\log(n) + \log\left(\prod_{i=1}^{n} q_i\right) + \sum_{i=1}^{n}(nq_i - 1)\right) + \frac{\varepsilon}{2}\sum_{i=1}^{n}(nq_i-1)^2, \quad (6.25)$$

where ε is set by bootstrap, to improve the confidence level of the region. Notice that choosing $\varepsilon = 1$ leads to $K_1(Q^*, P_n) = \frac{1}{2}\sum_{i=1}^n (nq_i - 1)^2$, which is almost equivalent to the Generalized Method of Moments, see Bertail *et al.* (2006a) or Bonnal and Renault (2001).

6.6.2 Estimating conditional moment by smooth empirical likelihood

This paper is interested in the estimation of linear equations, which can be rewritten as conditional moment equations. Consider the equation $y = X\theta + \varepsilon$, where ε is an error term. Suppose that some of the r.h.s. variables are endogenous, and that we have a set of instruments Z for these variables. Then, the set of exclusion restrictions that identifies θ is :

$$\mathbb{E}[y - X\theta | Z] = 0. \quad (6.26)$$

Note that one use more often the less stringent condition $\mathbb{E}[(y - X\theta)Z] = 0$, which is a necessary but not sufficient condition for (6.26) to hold. A number of econometric articles are devoted to the search for an optimal moment condition derived from (6.26) : what is the best function $h(.)$ such that the condition $\mathbb{E}[(y - X\theta)h(Z)] = 0$ gives the most efficient estimates of θ ?

Kitamura *et al.* (2004) introduced the Smooth Empirical Likelihood (SEL), modifying (smoothing in some sense) the empirical likelihood program to use efficiently conditional moments conditions. Suppose that the data is composed of observations (y_i, X_i, Z_i) and that the previous moment condition $\mathbb{E}[m(Y, \theta)] = 0$ is replaced by a conditional moment condition, taking the more general form $\mathbb{E}[m(y, X, \theta)|Z]$, where $\theta \in \mathbb{R}^q$ is the parameter of interest and m is some regular differentiable function from $\mathcal{X}\times\mathbb{R}^q \to \mathbb{R}^r$. The problem is that the empirical counterpart of the conditional moment condition is : $\forall i, \quad \mathbb{E}[m(y, X, \theta)|Z_i] = 0$. Unfortunately, conditionally to Z_i one observes only one couple (y, X) so that $\mathbb{E}[m(y, X, \theta)|Z_i]$ has no

direct empirical counterpart. Kitamura *et al.* propose to use the observations j that are close to i for a metric distance on Z, which smoothes the conditional equation in some sense. For using observations j in the neighbourhood of i, one may compute weights w_{ij} on the couples (Z_i, Z_j) using a kernel Φ :

$$w_{ij} = \frac{\Phi\left(\frac{Z_i - Z_j}{h}\right)}{\sum_j \Phi\left(\frac{Z_i - Z_j}{h}\right)} \tag{6.27}$$

where h is an appropriate bandwith.

In a second step, one has to solve for each i the optimization program :

$$L_i(\theta) = \sup_{(q_{ij})_j} \left\{ \sum_{j=1}^n w_{ij} \log(q_{ij}) \middle| \sum_{j=1}^n q_{ij} = 1,\ \sum_{j=1}^n q_{ij} m(y_j, X_j, \theta) = 0 \right\} \tag{6.28}$$

and then to optimize $L(\theta) = \sum_{i=1}^n L_i(\theta)$ with respect to θ.

We remark that, for each i, Kitamura's program is equivalent to the following optimization where $\mathbb{W}_i = \frac{1}{n}\sum_{j=1}^n w_{ij}\delta_{y_j,X_j}$ replace the usual empirical probability measure $\frac{1}{n}\sum_{j=1}^n \delta_{y_j,X_j}$:

$$\begin{aligned}
\beta_{i,n}(\theta) &= n \inf_{\mathbb{Q}_i m(y,X,\theta)=0} \{K(\mathbb{Q}_i, \mathbb{W}_i)\}, && (6.29)\\
&= n \inf_{\mathbb{Q}_i m(y,X,\theta)=0} \left\{ \mathbb{W}_i(1) - \mathbb{Q}_i(1) - \int \log\left(\frac{d\mathbb{Q}_i}{d\mathbb{W}_i}\right) d\mathbb{W}_i \right\}, && (6.30)\\
&= \sup_{\lambda \in r} \left\{ -\lambda' \sum_{j=1}^n w_{ij} m(y_j, X_j, \theta) - \sum_{j=1}^n w_{ij} \log\left(1 + \lambda' m(y_j, X_j, \theta)\right) \right\}, && (6.31)\\
&= \sum_{j=1}^n w_{ij} \log(w_{ij}) - \sup_{\substack{\sum_{j=1}^n q_{ij}=1 \\ \sum_{j=1}^n q_{ij} m(y_j,X_j,\theta)=0}} \left\{ \sum_{j=1}^n w_{ij} \log(q_{ij}) \right\}, && (6.32)\\
&= \sum_{j=1}^n w_{ij} \log(w_{ij}) - L_i(\theta). && (6.33)
\end{aligned}$$

As $\sum_{j=1}^n w_{ij}\log(w_{ij})$ is the supremum over θ of $L_i(\theta)$, $\beta_{i,n}(\theta)$ is actually the likelihood ratio. SEL is then re-interpreted as the minimization of the Kullback discrepancy between the probability $(\mathbb{Q}_i)$ of $(y, X)|Z_i$ and a first step kernel estimator $(\mathbb{W}_i)$.

Here again, SEL method can be modified as for Empirical Likelihood, by replacing K with the χ^2 discrepancy :

$$\beta_{i,n}(\theta) = n \inf_{\mathbb{Q}_i m(Z,\theta)=0} \{K_1(\mathbb{Q}_i, \mathbb{W}_i)\} = \inf_{\substack{\sum_{j=1}^n q_{ij}=1 \\ \sum_{j=1}^n q_{ij} m(Z_j,\theta)=0}} \sum_{j=1}^n \frac{(q_{ij} - w_{ij})^2}{2w_{ij}}. \tag{6.34}$$

See Bonnal & Renault (2001) for a complete study.

6.7 Results

Central to this paper is the idea that social norms do not affect behaviours directly, but through individual perceptions of ideal body weight. Hence, before estimating our model of ideal body weight, we show that ideal BMI predicts some attitudes towards food.

6.7.1 Ideal body weight as a predictor of attitudes towards food

Economists are often reluctant to use subjective variables, because they are not revealed by observable actions : what are the incentives for telling the truth about one's own 'ideal' BMI ? One way to answer this question is to examine whether the discrepancy between actual and ideal BMI, *i.e.* weight satisfaction, predicts attitudes towards food. Since we only have cross-section data, we are not able to produce a causal analysis, as for instance Clark and Georgellis (2004), who show that past job satisfaction predicts current job quits. Here, prediction means correlation.

The data contains information on some of the consumer's behaviours and attitudes. The interesting variables are : the subjective health (coded in 4 levels), the perception of the diet quality (4 levels), the frequency of consumption of low fat or sugar products, the restrictions on food products that are motivated by concerns with their fat or sugar content, the last day consumptions of carbohydrated drinks and alcohol (see Table A.1. in Appendix A). Our intuition is that, if ideal BMI matters, then individuals who feel overweight should eat healthier products (with less fat and sugar for instance). When asked to evaluate the quality of their diet, they should perceive it as rather unbalanced. Last, we should observe negative correlations between the deviation from ideal BMI and subjective health.

Controlling for a number of variables (income, education, etc.), we have estimated probit and ordered probit models of the attitude variables, with 5 key explanatory variables : DIFFNEG, DIFF0, DIFFPOS, the logarithm of the BMI and the logarithm of height. Denote actual and ideal BMIs respectively by W and W^*, then we define DIFFNEG, DIFF0, DIFFPOS as :

$$DIFFNEG = \begin{cases} -ln(W/W^*) & if \quad W < W^*, \\ 0 & otherwise. \end{cases} \tag{6.35}$$

$$DIFF0 = \begin{cases} 1 & if \quad W = W^*, \\ 0 & otherwise. \end{cases} \tag{6.36}$$

$$DIFFPOS = \begin{cases} ln(W/W^*) & if \quad W > W^*, \\ 0 & otherwise. \end{cases} \tag{6.37}$$

These three variables intend to capture asymmetric effects of deviations from ideal body shape.

Table D.1. in Appendix D reports the results for the 5 key explanatory variables.[17] The regression results are clear : given actual BMI, those individuals who feel overweight (an actual BMI greater than their ideal BMI) also feel in worse subjective health (see the

[17] Full results available upon request from the authors.

coefficient on DIFFPOS). They perceive their diet as being less healthy. They are also more likely to declare restrictions on unhealthy food. They declare consuming more often sugar-free products. Their last day consumption of sodas and alcohol is lower. The picture is also fairly consistent for those who feel underweight. They also feel in worse subjective health, tend to perceive their diet as less healthy, are less likely to declare restrictions or to declare consuming light products. However, their last-day consumptions of alcohol or sodas tend to be lower.

The discrepancy between actual and ideal body weights is correlated with a number of behaviours and attitudes. Ideal body weight captures information about individual preferences that are not captured by actual weight. Moreover, given weight satisfaction and actual weight, the correlations with our measure of social norm are always insignificant. These are suggestive evidence that, if social norms affect behaviours, this is through individual perceptions and not directly.

6.7.2 The effect of social norms on ideal body weight

Tables D2 in Appendix D reports the main results. The econometric method is indicated in the first row. QK-SEL means that we use the SEL estimator with a χ^2 divergence. For each regression, the point-estimates of the coefficients are reported, as well as their p-values. All BMI variables are in logarithm.

The first column (OLS / 1) produces OLS results from a specification, which controls only for the H and the education variables. Men have an higher ideal BMI than women (+11%), as expected, and the age effect is increasing concave. The occupation dummies are often not significant, but workers, farmers and some fractions of office workers and associate professionals seems to have an higher ideal BMI than executives. Hence, the social gradient is not so clear as found by Regnier (2006) with the same data set but different estimation sample and methods.

Last, those individuals in the lowest education groups have an higher ideal BMI (+2,6%). This effect disappears when actual BMI is introduced as an explanative variable, as is shown by the OLS results in the second column (OLS / 2). Here, the gender effects is divided by half, the age effect is no more significant, and the social gradient is somewhat reversed (an exception being the farmers). These results emphasize that choosing education as an instrument may be interesting, because it is known to be correlated with the actual BMI, and conditionnally to the latter it is not correlated with the ideal BMI.

The age trend and the sex and occupation dummies are dropped in the third specification (OLS / 3). They are replaced by the social norm ($\mathbb{E}(W^*|\Psi(Q))$ where W^* is in logarithm). With an OLS estimator the 'elasticity' of the ideal BMI to the social norm is 0.365.[18] Specification 3 is re-estimated with a GMM estimator in column 4, using three education dummies to instrument actual and ideal BMIs (GMM / 3). As shown by the partial R^2 and the F-test for the significance of the excluded instruments in OLS regressions of the endogenous variables, the instruments are significantly correlated with the endogenous variables (especially

[18]Note that the social norm is an average of the logarithms of ideal BMIs. Hence, we do not have a true elasticity, unless we define the social norm as as the geometric mean of the ideal BMIs.

the actual BMI), and the p-value of the Hansen's over-identification test is correct. The results in column 5 are produced with the QK-SEL estimator (QK-SEL / 3). It is clearly more efficient than the standard GMM for the key variables of the model : it produces a confidence region that is smaller for the social norm and the actual BMI. Note that the p-values are not smaller for all variables (see for instance SINGFAM). Column 5's results show that the social norm has a strong and very significant effect on individual perceptions of ideal BMI. Individual perceptions are a weighted average of the social norm and the actual BMI, with respective coefficients of 0.9 and 0.2.

Individual characteristics play a minor role here, because actual BMI captures most of their influence. However, we find a significant positive income effect, and negative effect of single-parenthood. Assumin that there are no measurement errors or contextual effects, the coefficients of the H variables in Table D2 represents within-group variations in the idiosyncratic reference point (W^p in equation (6.10)). In this case our results imply that, in any given social group, single-parents have a lower ideal body weight. One appealing explanation is in terms of value on the re-marriage market : thinness may compensate for the presence of children. Conversely, an higher income may compensate for overweight.

6.7.3 Discussion

This subsection briefly discusses the results.

First, we have found a remarkably high elasticity of ideal BMI to the social norm (about 0.9). But this is clearly an upper bound for the true elasticity in the whole population, because the sub-sample of individuals for which, according to the model, social norms matter, was selected. Indeed, the structural model implies that the elasticity is 0 for those with high marginal adjustment costs, whose weight satisfaction is actually maximised. This is consistent with the estimation results displayed in Table D1, which show that their food attitudes are not clearly oriented (see the coefficients on $DIFF0$) : they consume less carbohydrated drinks and alcohol, but they are also less likely to adopt restrictions and to eat light or sugar-free products.

Second, the magnitude of the estimates are unlikely to be due to a weakness of the instrument set. Actually, it is well-known that poor instruments leads to the same bias as the OLS estimates. Here, the instrumental variable estimate of the elasticity is much higher than its OLS counterpart.

Third, the specification assumes that the contextual variables DH do not impact the ideal BMI. They may also be an instrument for DW^*, but only if they are uncorrelated with the omitted group-specific unobservable factors $\mathbb{E}(\eta|\Psi(Q))$. To test this assumption, specification 3 is estimated in Table D2's Column 6 using education and DH as the excluded instruments (GMM / 4). The over-identification restrictions are clearly rejected. Hence, some contextual effects are present, and their omission biases the estimates for the H variables.

Fourth, is it possible to identify the impact of the variables Q ? Actually, this could be done only by relying on non-linearities between some control function $f(Q)$ for the effects of Q, and $\mathbb{E}(W^*|\Psi(Q))$. We consider the following specification :

$$W^* = \overline{\alpha} DW^* + \overline{\beta} W + \delta H + \overline{\rho} f(Q) + \epsilon \tag{6.38}$$

where $f(Q)$ is a vector of variables including gender, occupation, age and age squared as in specifications 1 and 2. The estimation results for this fourth specification are presented in Table D3. Whatever the method, the social norm elasticity is lower. However, the partial R^2 in the first-step regressions are very low, when one uses only education to instrument actual and ideal BMIs (Columns 2 and 3's results in Table D3). This is also the case, but for actual BMI only, when one uses the contextual variables for the instrumentation (see the Columns 4 and 5's results in Table D3). The regression results show that, when there are not many instruments, the QK-SEL estimator is much more efficient than the GMM, as it was observed previously for specification 3. Using only education to instrument the social norm, the elasticity estimated by the GMM is insignificant (Column 3), while the point estimates is almost the same (about 0.38) but is significant for the QK-SEL estimator. The gain of using the QK-SEL estimator disappears when the contextual effects are also used as instruments. Moreover, in this case, the over-identifcation restrictions are not rejected, which means that the introduction of the $f(Q)$ variables is sufficient to control for the omitted contextual variables. That the instruments are weak is reflected in the coefficient of actual BMI, which is almost the same as in the OLS estimates of Specifications 2 and 3. This explains why the elasticity of ideal BMI to the social norm differs from the estimate in Table D2, and is biased towards the OLS estimates of specification 3.

Table A.1. – Name and variable definition (N=3972)

Name	Definition	Mean	Standard deviation
WEIGHT	Self-declared actual body weight	68.90	14.12
IDEALWEIGHT	Ideal body weight	65.58	12.02
HEIGHT	Height (in m)	1.67	0.09
BMI, W	Self-declared Body weight in kg divided by height in meters squared	24.61	4.19
IDEALBMI, W*	3972	23.39	3.16
PHYSADVICE	Answer "yes" to « According to you, [does] your usual general practitioner (…) gives you advices on your lifestyle –food, exercise, tobacco and alcohol-?	54.9%	
VPREV2	Has had more than 1 preventive check-up in the last 12 months	19.9%	
INSURMAX	Full insurance coverage	94.0%	
SEX	=1 if male, = 0 otherwise	43.0%	
AGE	Age	51.19	17.66
INCMIN	Minimum yearly net household income adjusted by the number of consumption unit (OECD scale), in 2001 FF	135867.3	83298.3
EDUCATION LEVELS			
EDUC1	No qualification or Primary school only (NOQUAL, CEP)	35.0%	
EDUC2	Secondary general or vocational education ,first cycle only (CAP, BEPC)	32.5%	
EDUC3	Secondary general, technical or vocational education, second cycle. (Bac)	12.2%	
EDUC4	Qualification higher than the Baccalaureat (Bac2, Bac3+)	20.4%	
STANDARD OCCUPATIONAL CLASSIFICATION (SOC)			
FARMERS	Farmers and farm managers	5.1%	
OWNERS	Business owners (mainly in skilled trades occupations)	6.7%	
EXECUTIVES	Public and private sectors executives (include managers in the public and industrial sectors, professionals in the private sector, upper categories in the teaching, culture and media sectors) (reference).	11.6%	
MIDPUB1	Teachers, professional occupations in the health and social welfare sectors (nurses, community workers, etc.)	8.1%	
MIDPUB2	Associate professional and technical occupations in the public sector (police officer etc.)	6.1%	
MIDPRIV	Associate professional and technical occupations in the private sector (technician etc.)	5.9%	
EMPPUB	Administrative, secretarial and personal service occupations in the public sector without management responsibilities (nursing assistant, policeman etc.)	11.0%	
EMPPRIV1	Administrative, secretarial and personal service occupations in the private sector without management responsibilities (company secretary etc)	9.8%	
EMPPRIV2	Sale and customer service occupations in the private sector without management responsibilities (retail cashiers etc.)	11.6%	
SKWORK1	Skilled worker	11.0%	
SKWORK2	Transport and mobile machine drivers and operatives	3.1%	
UNSKWORK	Elementary occupation	10.1%	
REGION			
REGION1	Ile-de-France (reference)	17.2%	
REGION2	Nord, Champagne-Ardennes, Lorraine, Alsace	16.8%	
REGION3	Pays de Loire, Bretagne, Centre, Limousin, Aquitaine, Poitou-	24.4%	

	Charente		
REGION4	Bourgogne, Franche-Comté, Rhône-Alpes, Auvergne, Midi-Pyrénées, Languedoc	25.8%	
REGION5	PACA, Corse	8.0%	
REGION6	Picardie, Normandie	7.9%	
	URBAN UNIT		
STRAT1	Rural area	26.5%	
STRAT2	Small towns	16.8%	
STRAT3	Middle towns	13.4%	
STRAT4	Big towns	28.6%	
STRAT5	Paris (reference)	14.6%	
	MARITAL STATUS		
SINGFAM	Single parent family	6.7%	
COUPLECH2	Couple with at least two children	18%	
COUPLECH1	Couple with one children	13.1%	
COUPLENOCH	Couple without children	29.7%	
SINGLE	Single without children (never been in couple, separated or divorced)	23.9%	
WIDOWED	Widowed	13.8%	
	ATTITUDES AND BEHAVIOURS		
EXERCISE	Frequency of exercise = Never or less than once in a month	65.3%	
	Between one and three times in a month	5.5%	
	Once a week at least	12.2%	
	Several times in a week	17.0%	
SMOKE	Regular smoker	26.3%	
SUBJHEALTH	Subjective health status = poor	9.8%	
	Fair	26.8%	
	Good	47.4%	
	Very good	16.0%	
SUBJDIET	Perception of diet quality = unbalanced	4.1%	
	Not well balanced	19.2%	
	Fairly balanced	43.9%	
	Well balanced	32.8%	
LIGHTFAT	Consume products light in fat = rarely or never	49.6%	
	Sometimes	19.1%	
	At least once a week	8.5%	
	Every days	22.7%	
FREESUGAR	Consume products light in sugar = rarely or never	65.3%	
	Sometimes	11.9%	
	At least once a week	5.8%	
	Every days	17.7%	
RESTRICT	Avoid consuming some tasty food products when they are too rich in fat or sugar	43.6%	
SODA	Had drunken at least one glass of a carbo-hydrated drink in the last day	39.0%	
ALCOHOL	Had drunken at least one glass of alcohol in the last day	50.9%	

Figure A1.

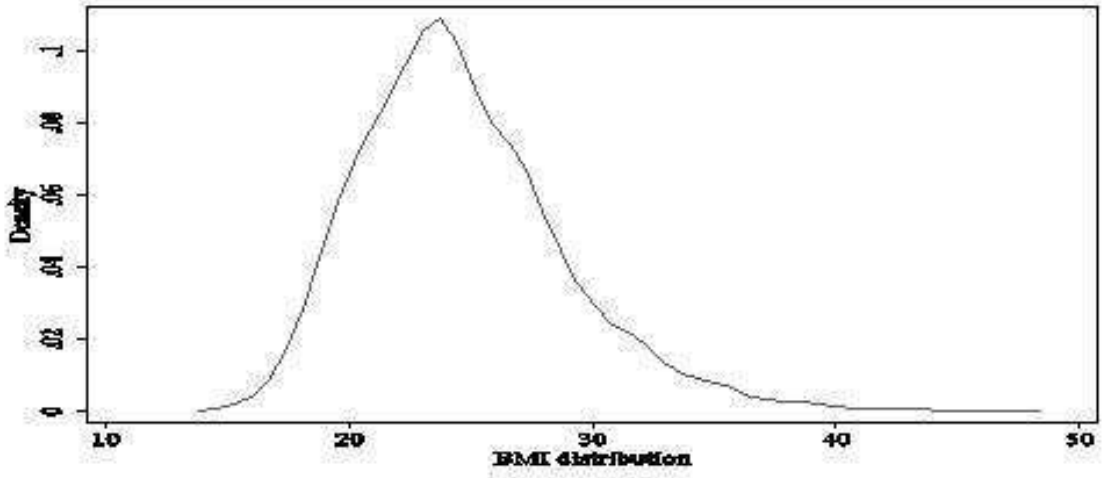


Figure A2.

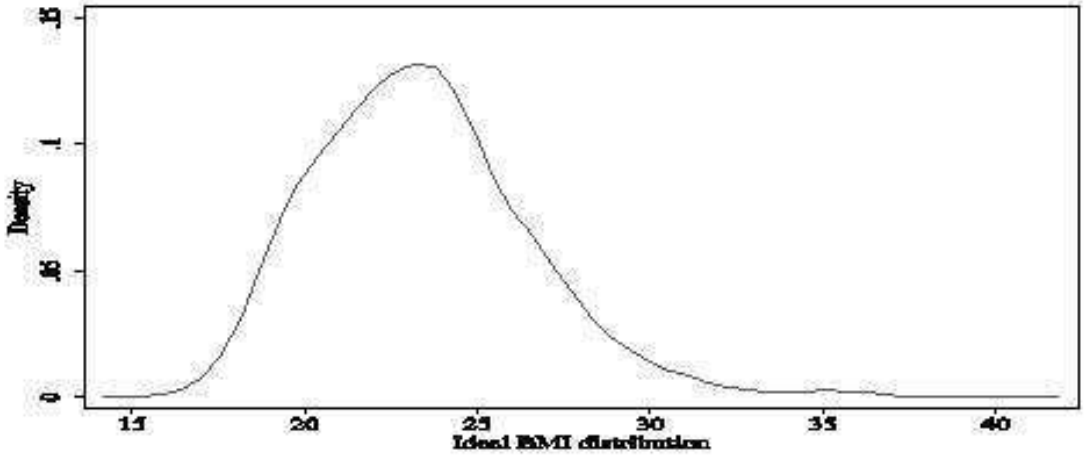


Figure A3

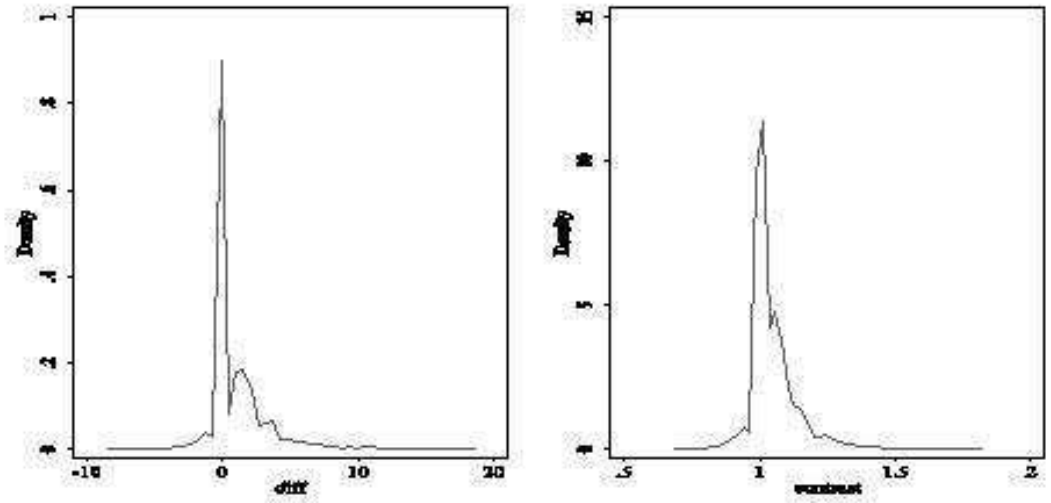


Figure B1 – Residual auto-correlation at distance 0 for ideal BMI – Men

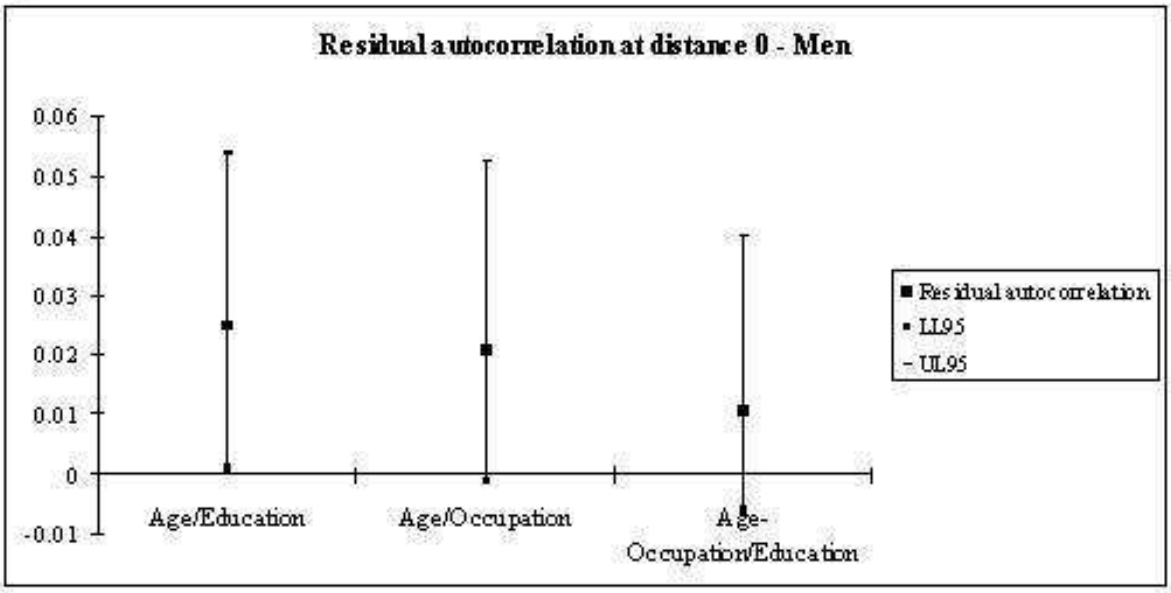


Figure B2 – Residual auto-correlation at distance 0 for ideal BMI – Women

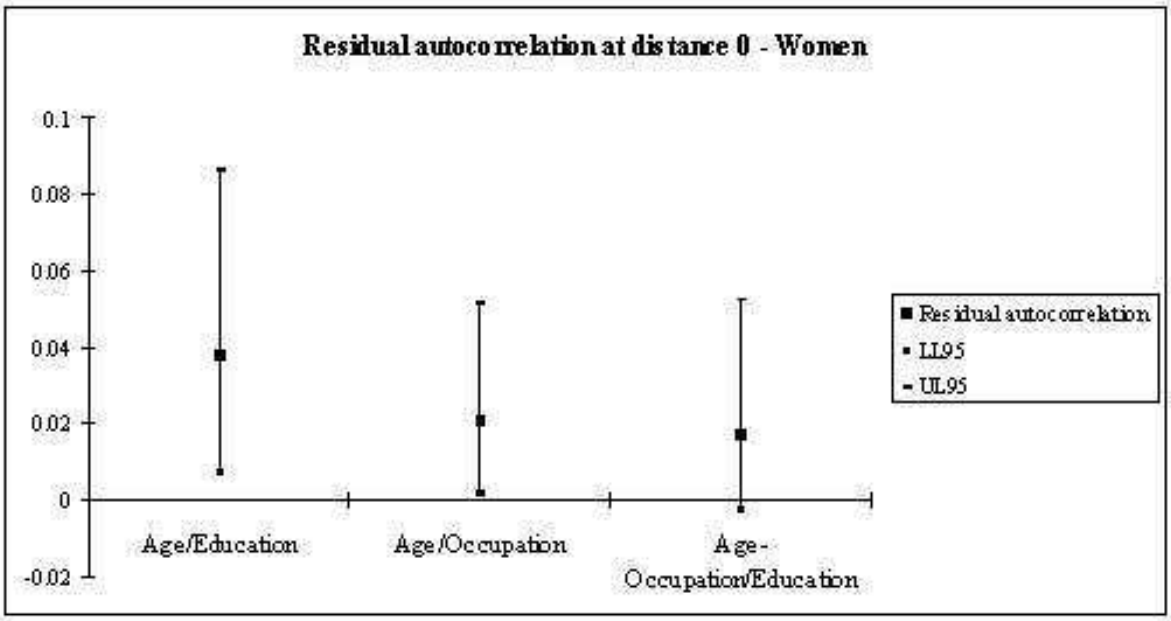


Figure C.1. - Case n°1

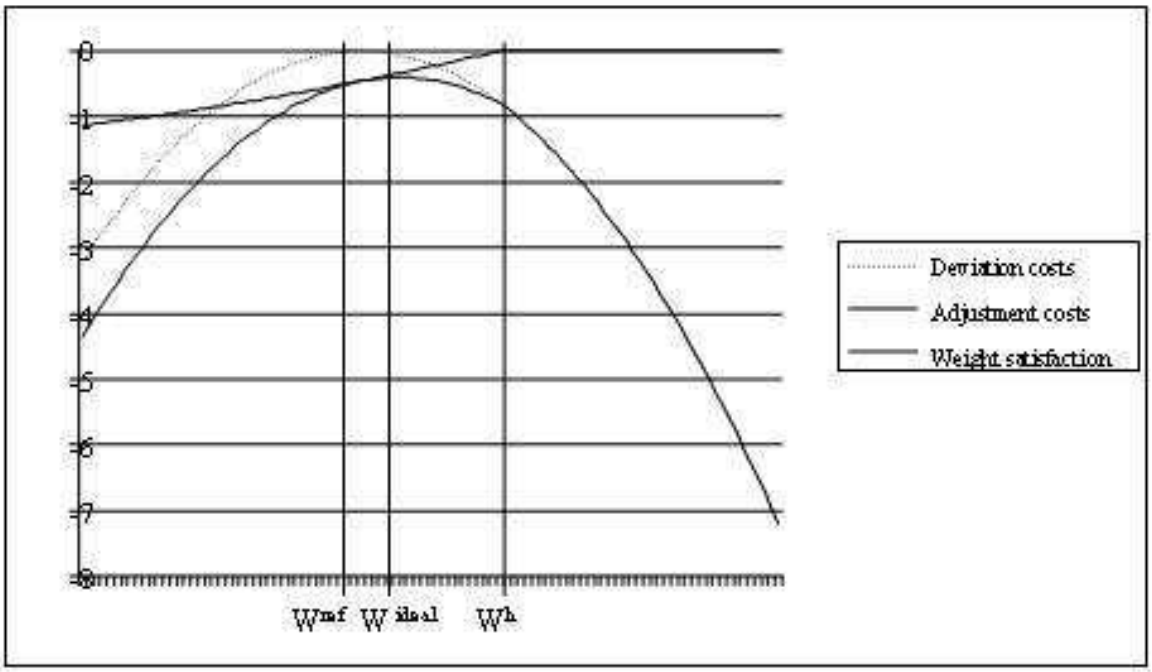
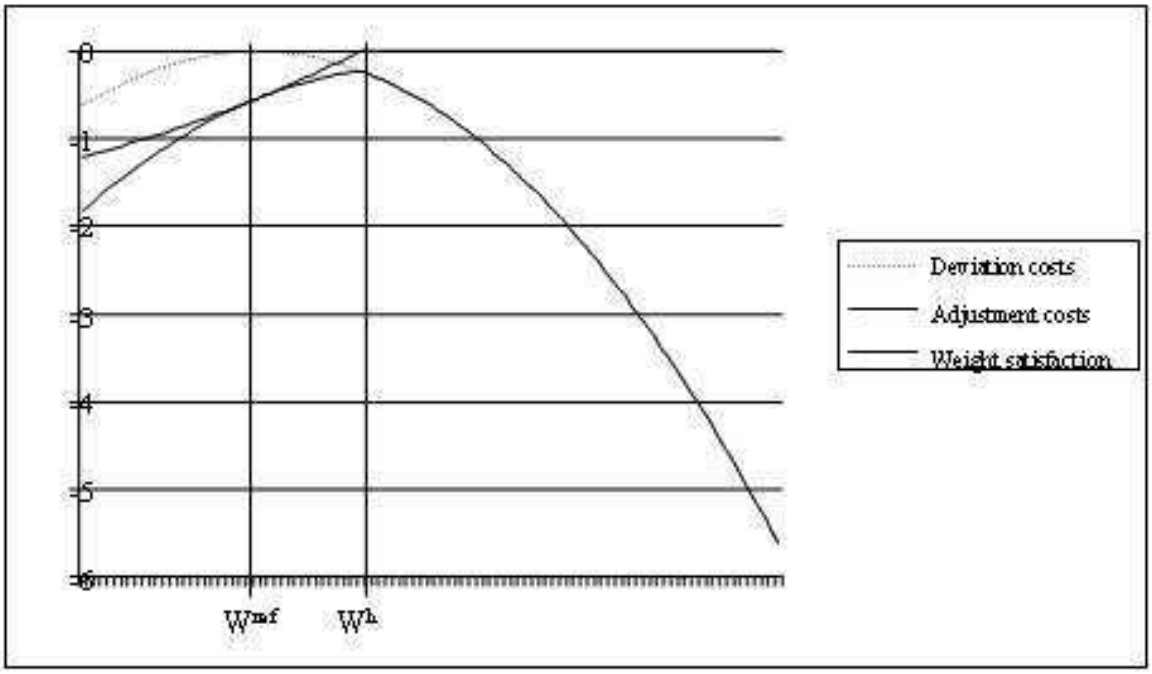


Figure C.2. - Case n°2



*Notes for all tables: *=significant at the 10% level, **=significant at the 5% level, ***=significant at the 1% level*

Table D.1. – Ideal Body Weight as a predictor of food attitudes

	SUBJHEALTH	**SUBJDIET**	**LIGHTFAT**	**FREESUGAR**	**RESTRICT**	**SODA**	**ALCOHOL**
Model	*Ordered probit*				*Probit*		
DIFFNEG	-4.949***	-2.140***	-5.530***	-3.267***	-4.559***	-2.590**	-2.016**
	(0.827)	(0.787)	(1.013)	(1.070)	(1.021)	(1.034)	(0.952)
DIFF0	-0.004	-0.017	-0.308***	-0.263***	-0.276***	-0.177***	-0.224***
	(0.047)	(0.047)	(0.050)	(0.054)	(0.056)	(0.059)	(0.057)
DIFFPOS	-0.922**	-2.232***	0.070	0.911**	0.952**	-1.706***	-1.756***
	(0.388)	(0.387)	(0.403)	(0.423)	(0.460)	(0.516)	(0.485)
Log(Social norm: E(W*|Q))	0.593	-0.991	-0.169	-1.349*	-0.006	-1.131	-1.176
	(0.686)	(0.684)	(0.728)	(0.793)	(0.812)	(0.846)	(0.825)
Log(Actual BMI: W)	-0.271	-0.257	0.385**	0.802***	-0.013	-0.132	-0.166
	(0.166)	(0.166)	(0.175)	(0.187)	(0.196)	(0.206)	(0.199)
Log(HEIGHT)	-0.215	-0.145	0.935*	0.819	0.798	0.809	0.686
	(0.465)	(0.466)	(0.488)	(0.530)	(0.549)	(0.570)	(0.556)
Controls	Log(INCMIN), SEX, AGE/10, $(AGE/10)^2$, FARMERS, OWNERS, MIDPUB1, MIDPUB2, MIDPRIV, EMPPUB, EMPPRIV1, EMPPRIV2, SKWORK1, SKWORK2, UNSKWORK,EDUC1-EDUC3, COUPLECH1, COUPLENOCH, SINGLE, SINGFAM, WIDOWED, STRAT1-STRAT4, REGION2-REGION5						

Table D2 – The determinants of ideal BMI (Dependent variable : log (W*); N=2121).

Method / Specification	**OLS / 1**		**OLS / 2**		**OLS / 3**		**GMM / 3**		**QK –SEL / 3**		**GMM / 3**	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Log(Social norm)					0.365***	0.000	0.906	0.248	0.904***	0.000	0.531***	0.000
Log(Actual BMI)			0.636***	0.000	0.641***	0.000	0.195	0.729	0.194***	0.000	0.486***	0.000
SEX	0.110***	0.000	0.050***	0.000								
AGE/10	0.053***	0.000	0.006	0.197								
$(AGE/10)^2$	-0.003***	0.001	0.000	0.436								
FARMERS	0.041***	0.005	0.012	0.109								
OWNERS	0.004	0.724	-0.008	0.200								
MIDPUB1	0.008	0.425	-0.007	0.185								
MIDPUB2	0.008	0.463	0.001	0.876								
MIDPRIV	0.023	0.058	0.002	0.695								
EMPPUB	0.025*	0.015	-0.002	0.735								
EMPPRIV1	-0.002	0.814	-0.014***	0.010								
EMPPRIV2	0.015	0.181	-0.010	0.076								
SKWORK1	0.023*	0.043	-0.006	0.297								
SKWORK2	0.034*	0.030	0.002	0.803								
UNSKWORK	0.021	0.085	-0.015**	0.015								
EDUC1	0.026***	0.002	0.003	0.503								
EDUC2	0.012	0.122	0.001	0.831								
EDUC3	0.005	0.529	-0.003	0.502								
Log(INCMIN)	-0.008	0.146	0.003	0.325	0.009***	0.000	0.001	0.930	0.002***	0.000	0.005	0.114
COUPLECH1	0.000	0.971	-0.001	0.756	-0.005	0.218	-0.006	0.331	-0.01	0.400	-0.005	0.233
COUPLENOCH	0.003	0.674	-0.003	0.394	-0.003	0.474	-0.007	0.595	0.012*	0.034	-0.003	0.502
SINGLE	-0.002	0.793	-0.003	0.387	-0.003	0.410	-0.006	0.429	-0.003	0.699	-0.004	0.321
SINGFAM	-0.037***	0.000	-0.016***	0.001	-0.017***	0.001	-0.023**	0.013	-0.02	0.404	-0.019***	0.003
WIDOWED	0.028***	0.006	0.006	0.263	0.006	0.163	0.006	0.683	0.012	0.352	0.009	0.121
Other control variables *H* : constant, STRAT1-STRAT4, REGION2-REGION5, log(HEIGHT)												
Excluded instruments							EDUC1-EDUC3				EDUC1-3, E(H|Ψ(Q))	
Significance of the excluded instruments, social norm (first-step regression)							F-test, p-value: 0.000; partial R^2: 0.1356				partial R^2: 0.6035	
Significance for the excluded instruments, actual BMI (first-step regression)							F-test, p-value: 0.000; partial R^2: 0.044				partial R^2: 0.0994	
Hansen's over-identification test of excluded instruments (GMM estimator only)							p-value: 0.878				p-value: 0.000	

Table D3 – The determinants of ideal BMI (Dependent variable : log (W*); N=2121)

Method / specification	**OLS / 4**		**GMM / 4**		**QK –SEL / 4**		**GMM / 4**		**QK –SEL / 4**	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
Log(Social norm)	0.137***	0.005	0.388	0.492	0.382***	0.000	0.372***	0.000	0.414***	0.000
Log(Actual BMI)	0.637***	0.000	0.705***	0.000	0.705***	0.000	0.647***	0.000	0.641***	0.000
SEX	0.036***	0.000	0.005	0.922	0.005	0.142	0.012	0.220	0.001	0.737
AGE/10	-0.002	0.682	-0.024	0.504	-0.021***	0.000	-0.020**	0.017	-0.039***	0.000
$(AGE/10)^2$	0.001*	0.070	0.002	0.323	0.002***	0.000	0.002***	0.002	0.003***	0.000
FARMERS	0.004	0.589	-0.016	0.646	-0.017	0.512	-0.012	0.148	-0.044	0.286
OWNERS	-0.011*	0.062	-0.021	0.193	-0.017	0.318	-0.018***	0.001	-0.019	0.362
MIDPUB1	-0.007	0.165	-0.009	0.101	-0.011	0.226	-0.008	0.139	0.023*	0.078
MIDPUB2	-0.002	0.784	-0.007	0.465	-0.013	0.344	-0.007	0.148	-0.017	0.346
MIDPRIV	-0.001	0.887	-0.010	0.463	-0.014	0.476	-0.007	0.216	0.011	0.653
EMPPUB	-0.007	0.195	-0.022	0.344	-0.023**	0.011	-0.018***	0.003	0.003	0.793
EMPPRIV1	-0.015***	0.002	-0.021**	0.012	-0.030***	0.000	-0.019***	0.000	-0.026**	0.010
EMPPRIV2	-0.013	0.015	-0.025	0.147	-0.023**	0.012	-0.020***	0.001	-0.024**	0.022
SKWORK1	-0.011	0.060	-0.026	0.264	-0.019*	0.077	-0.020***	0.003	-0.016	0.239
SKWORK2	-0.006	0.458	-0.028	0.455	-0.027	0.489	-0.022***	0.009	-0.011	0.818
UNSKWORK	-0.021	0.001	-0.041	0.175	-0.04***	0.001	-0.036***	0.000	-0.045***	0.005
Log(INCMIN)	0.002	0.379	0.004	0.253	0.005***	0.000	0.003	0.359	0.008***	0.000
COUPLECH1	-0.002	0.656	-0.003	0.509	-0.005	0.503	-0.003	0.371	-0.017*	0.070
COUPLENOCH	-0.004	0.275	-0.007	0.234	-0.011***	0.003	-0.007	0.071	0.004	0.399
SINGLE	-0.004	0.287	-0.005	0.267	-0.006	0.187	-0.006	0.103	-0.035***	0.000
SINGFAM	-0.015***	0.001	-0.011	0.177	-0.013	0.423	-0.013**	0.044	-0.021	0.251
WIDOWED	0.004	0.510	-0.003	0.795	-0.003	0.762	-0.001	0.939	0.009	0.407
Other control variables : constant, STRAT1-STRAT4, REGION2-REGION5, log(HEIGHT)										
Excluded instruments (number)			EDUC1-EDUC3 (3)				EDUC1-EDUC3, $E(H\|\Psi(Q))$ (18)			
Significance of the excluded instruments, social norm			F-test, p-value: $3e^{-4}$; partial R^2: 0.008				F-test, p-value: 0.000; partial R^2: 0.418			
Significance for the excluded instruments, actual BMI			F-test, p-value: 0.015; partial R^2: 0.005				F-test, p-value: 0.236; partial R^2: 0.010			
Hansen's over-identification test of excluded instruments (GMM estimator only)			p-value: 0.392				p-value: 0.494			

Bibliographie

M. Abramovitch & L. A. Stegun. *Handbook of Mathematical Tables*. National Bureau of Standards, Washington, DC, 1970.

G. A. Akerlof. Social distance and social decisions. *Econometrica*, 65 :1005–1027, 1997.

O. Allais & V. Nichèle. Markov switching aids models : an application to french beef, poultry and fish consumption. Working Paper, INRA, CORELA, 2003.

J-F. Amadieu. *Le poids des apparences*. Odile Jacob, Paris, 2002.

D. W. K. Andrews & M. M. A. Schafgans. Semiparametric estimation of a sample selection model. Working Paper n°1119, Cowles Foundation, Yale University, 1996.

S. Asmussen. *Applied Probabilities and Queues*. Wiley, 1987.

K.B. Athreya & P. Ney. A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.*, 245 :493–501, 1978.

A. Ausslender, M. Teboulle, & S. Ben-Tiba. Logarithm-quadratic proximal method for variational inequalities. *Computational Optimization and Applications*, 12 :31–40, 1999.

S. Averett & S. Korenman. The economic reality of *The Beauty Myth*. *The Journal of Human Resources*, 31 :304–330, 1996.

K. A. Baggerly. Empirical likelihood as a goodness of fit measure. *Biometrika*, 85 :535–547, 1998.

R. R. Bahadur & L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, 27 :1115–1122, 1956.

J. Banks, R. W. Blundell, & A. Lewbel. Quadratic engel curves and consumer demand. *Review of Economics and Statistics*, 79(4) :527–539, 1997.

P. Barbe & P. Bertail. Testing the global stability of a linear model. Working Paper n°46, CREST, 2004.

O. E. Barndorff-Nielsen & D. R. Cox. Bartlett adjustments to the likelihood ratio statistic and the distribution of the maximum likelihood estimator. *Journal of the Royal Statistical Society, Series B*, 46 :483–495, 1984.

O. E. Barndorff-Nielsen & P. Hall. On the level-error after bartlett adjustement of the likelihood ratio statistic. *Biometrika*, 75 :374–378, 1988.

M. S. Bartlett. Approximate confidence intervals. *Biometrika*, 40 :12–19, 1953.

A. Basdevant. Sémiologie et clinique de la restriction alimentaire. *Cahiers de Nutrition et de Diététique*, 33 :235–241, 1998.

V. Bentkus & F. Götze. Optimal rates of convergence in the CLT for quardatic forms. *Annals of Probability*, 24(1) :466–490, 1996.

B. Bercu, E. Gassiat, & E. Rio. Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *Annals of Probability*, 30(4) :1576–1604, 2002.

P. Bertail. Empirical likelihood in some nonparametric and semiparametric models. In M. S. Nikulin, N. Balakrishnan, M. Mesbah, & N. Limnios, editors, *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, Statistics for Industry and Technology. Birkhauser, 2004.

P. Bertail. Empirical likelihood in some semi-parametric models. *Bernoulli*, 12(2) :299–331, 2006.

P. Bertail & S. Clémençon. Edgeworth expansions for suitably normalized sample mean statistics of atomic Markov chains. *Prob. Th. Rel. Fields*, 130 :388–414, 2004a.

P. Bertail & S. Clémençon. Note on the regeneration-based bootstrap for atomic Markov chains. To appear in *Test*, 2004b.

P. Bertail & S. Clémençon. Regenerative block bootstrap for Markov chains. *Bernoulli*, 12 (4) :689–712, 2006a.

P. Bertail & S. Clémençon. Approximate regenerative block-bootstrap for Markov chains : second-order properties. In *Compstat 2004 Proceedings*. Physica Verlag, 2004c.

P. Bertail & S. Clémençon. Regeneration-based statistics for Harris recurrent Markov chains. In P. Bertail, P. Doukhan, & P. Soulier, editors, *Dependence in Probability and Statistics*, volume 187 of *Lecture Notes in Statistics*. Springer, 2006b.

P. Bertail & J. Tressou. Incomplete generalized U-Statistics for food risk assessment. *Biometrics*, 2004. Accepted.

P. Bertail, H. Harari-Kermadec, & D. Ravaille. φ-Divergence empirique et vraisemblance empirique généralisée. To appear in *Annales d'Économie et de Statistique*, 2004.

P. Bertail, E. Gauthérat, & H. Harari-Kermadec. Exponential bounds for quasi-empirical likelihood. Working Paper n°34, CREST, 2005.

M. Bertrand. Consommation et lieux d'achat des produits alimentaires en 1991. *INSEE Résultats*, 54-55, 1993.

J. Bhattacharya & M. K. Bundorf. The incidence of the healthcare costs of obesity. Working Paper n°11303, National Bureau of Economic Research, 2005.

P. Bickel, C. Klaassen, Y. Ritov, & J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

G. Blom. Some properties of incomplete u-statistics. *Biometrika*, 63 :573–580, 1976.

C. Boizot-Szantaï & F. Etilé. The food prices / body mass index relationship : theory and evidence from a sample of french adults. XIth International Congress of the European Association of Agricultural Economists, Copenhaguen, August 24-27, 2005.

E. Bolthausen. The Berry-Esseen theorem for strongly mixing Harris recurrent Markov chains. *Z. Wahr. Verw. Gebiete*, 60 :283–289, 1982.

H. Bonnal & É. Renault. Minimum chi-square estimation with conditional moment restrictions. Working Paper, C.R.D.E., 2001.

H. Bonnal & É. Renault. On the efficient use of the informational content of estimating equations : Implied probabilities and euclidean empirical likelihood. Working Paper n°2004s-18, Cahiers scientifiques (CIRANO), 2004.

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

P. Bourdieu. *La distinction*. Editions de Minuit, Paris, 1979.

F. Bourguignon, P.-A. Chiappori, & P. Rey. *Théorie micro-économique : l'équilibre concurrentiel*, volume 1. Fayard les savoirs, 1992.

M. Broniatowski & A. Kéziou. Parametric estimation and tests through divergences. Working Paper, L.S.T.A., Université Paris VI, 2004.

M A. Burke & F. Heiland. Social dynamics of obesity. Working Paper n°40, Center on Social and Economic Dynamics, 2005.

D. Carr & M. Friedman. Is obesity stigmatizing ? body weight, perceived discrimination, and psychological well-being in the united states. *Journal of Health and Social Behavior*, 46 :244–259, 2005.

J. Cawley. The impact of obesity on wages. *The Journal of Human Resources*, 39 :451–474, 2004.

G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3) :305–334, 1987.

J. Chen, S.-Y. Chen, & J. N. K. Rao. Empirical likelihood confidence intervals for the mean of a population containing many zero values. *Canadian Journal of Statistics*, 31(1) :53–68, 2003.

S. Chen & S. Khan. Estimation of a nonparametric censored regression model. Unpublished Manuscript, 2000.

S.-Y. Chen & J. Qin. Empirical likelihood-based confidence intervals for data with possible zero observations. *Statistics & Probability Letters*, 65(1) :23–37, 2003.

A. Chesher. Diet revealed ? : Semiparametric estimation of nutrient intake-age relationships. *Journal of the Royal Statistical Society A*, 160(3) :389–428, 1997.

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

D. Claisse, D. Cossa, G. Bretaudeau-Sanjuan, G. Touchard, & B. Bombled. Methylmercury in molluscs along the french coast. *Marine pollution bulletin*, 42 :329–332, 2001.

A. E. Clark & Y. Georgellis. Kahneman meets the quitters : Peak-end behaviour in the labour market. mimeo DELTA/PSE, 2004.

A. E. Clark & A. J. Oswald. Comparison-concave utility and following behaviour in social and economic settings. *Journal of Public Economics*, 70 :133–155, 1998.

S. Clémençon. Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the nummelin splitting technique. *Stat. Prob. Letters*, 55 : 227–238, 2001.

P. Combris, F. Etilé, & L-G. Soler. Alimentation et santé : changer les comportements de consommation ou mieux réguler l'offre alimentaire. In I. Proust, editor, *Désirs et peurs alimentaires au XXIème siècle*, pages 203–261. Dalloz, Paris, 2006.

T. G. Conley & G. Topa. Socio-economic distance and spatial patterns in uemployment. *Journal of Applied Econometrics*, 17 :303–327, 2002.

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

D. Cossa, D. Auger, B. Averty, M. Lucon, P. Masselin, J. Noel, & J. San-Juan. Atlas des niveaux de concentration en métaux métalloïdes et composés organochlorés dans les produits de la pêche côtière française. Technical report, IFREMER, Nantes, 1989.

S. R. Cosslett. Efficient semiparametric estimation of censored and truncated regressions via a smoothed self-consistency equation. *Econometrica*, 72(4) :1277–1293, 2004.

CREDOC-AFSSA-DGAL. *Enquête INCA (individuelle et nationale sur les consommations alimentaires)*. Lavoisier, Paris, TEC&DOC edition, 1999. (Coordinateur : J.L. Volatier).

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

I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2 :299–318, 1967.

D. Cutler, E. Glaeser, & J.M. Shapiro. Why have americans become more obese ? *Journal of Economic Perspectives*, 17 :93–118, 2003.

N. Darmon, E. L. Ferguson, & A. Briend. A cost constraint alone has adverse effects on food selection and nutrient density : An analysis of human diet by linear programming. *Journal of Nutrition*, 132 :3764–3771, 2002.

N. Darmon, A. Briend, & A. Drewnowsky. Energy-dense diets are associated with lower costs : a community study of french adults. *Public Health Nutrition*, 7 :21–27, 2004.

P.W. Davidson, G. Myers, C.Cox, C. F. Shamlaye, T. Clarkson, D.O. Marsh, M.A. Tanner, M. Berlin, J. Sloane-Reves, E. Cernichiari, O. Choisy, A. Choi, & T. W. Clarkson. Longitudinal neurodevelopmental study of seychellois children following in utero exposure to mehg from maternal fish ingestion : Outcomes at 19-29 months. *Neurotoxicology*, 16 : 677–688, 1995.

A. Deaton & J. Muellbauer. An almost ideal demand system. *American Economic Review*, 70(3) :312–326, 1980a.

A. Deaton & J. Muellbauer. *Economics and consumer behavior*. Cambridge University Press, 1980b.

B. Detournay, F. Fagnani, M. Phillippo, C. Pribi, M. A. Charles, C. Sermand, A. Basdevant, & E. Eschwege. Obesity morbidity and health care costs in france : an analysis of the 1991-1992 medical care household. *International Journal of Obesity*, 24 :151–155, 2000.

J. C. Deville & C. E. Sarndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87 :376–382, 1992.

T. DiCiccio & J. Romano. Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review*, 58 :59–76, 1990.

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

S. Donald, G. Imbens, & W. K. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Econometrics*, 117 :55–93, 2003.

P. Doukhan & P. Ango Nze. Weak dependence, models and applications to econometrics. *Econometric Theory*, 20(6) :995–1045, 2004.

A. Drewnowsky & N. Darmon. Replacing fats and sweets with vegetables and fruits - a question of cost. *American Journal of Public Health*, 94(9) :1555–1559, 2004.

Goblot E. *La barrière et le niveau. Étude sociologique sur la bourgeoisie française moderne*. Presses Universitaires de France, Paris, 1925.

M. L. Eaton. A probability inequality for linear combinations of bounded random variables. *Annals of Statistics*, 2 :609–614, 1974.

M. L. Eaton & B. Efron. Hotelling's t^2 test under symmetry conditions. *Journal of american statistical society*, 65 :702–711, 1970.

D. Edelman. Bounds for a nonparametric t table. *Biometrika*, 73 :242–243, 1986.

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

J. H. J. Einmahl & I. W. McKeague. Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2) :267–290, 2003.

J. Elster. Social norms and the economic theory. *Journal of Economic Perspective*, 3 :99–117, 1989.

P. Embrechts, F. Lindskog, & A. J. McNeil. Handbook of heavy tailed distributions in finance. chapter Modelling dependence with copulas and applications to risk management. North-Holland, Amsterdam, 2003.

FAO/WHO. Evaluation of certain food additives and contaminants _for methylmercury. Sixty first report of the Joint FAO/WHO Expert Committee on Food Additives, Technical Report Series, WHO, Geneva, Switzerland, 2003.

L. Festinger. A theory of social comparison processes. *Human Relations*, pages 114–140, 1954.

F. Gamboa & E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *Annals of Statistics*, 25(1) :328–350, 1996.

GEMs/Food-WHO. Reliable evaluation of low-level contamination of food, workshop in the frame of GEMS/Food-EURO. Technical report, Kulmbach, Germany, 26-27 May 1995, 1995.

A. L. Gibbs & F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3) :419–436, 2002.

A. Golan, G. Judge, & D. Miller. *Maximum Entropy Econometrics*. Wiley, New York, 1996.

A. Golan, J. Perloff, & E. Shen. Estimating a demand system with nonnegativity constraints : Mexican meat demand. *Review of Economics and Statistics*, 83(3) :541–550, 2001.

C. Gourieroux. Econometrie des variables qualitatives. 6 :185–220, 1989.

P. Grandjean, P. Weihe, R. White, F. Debes, S. Araki, K. Yokoyama, K. Murata, N. Sorensen, R. Dahl, & P. Jorgensen. Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicology Teratology*, 19 :417–428, 1997.

C. Grignon & Ch. Grignon. Long-term trends in food consumption : a french portrait. *Food and Foodways*, 8 :151–174, 1999.

F. Götze & H. R. Kunsch. Second order correctness of the blockwise bootstrap for stationary observations. *Annals of Statistics*, 24 :1914–1933, 1996.

P. Guggenberger & R. J. Smith. Generalized empirical likelihood estimators and tests under weak, partial and strong identification. *Econometric Theory*, 21 :667–709, 2005.

P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer, 1992.

P. Hall & J. Horowitz. Bootstrap critical values for tests based on generalized-method-of-moments estimators. *Econometrica*, 64 :891–916, 1996.

P. Hall, J. Horowitz, & B.-Y. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82 :561–574, 1995.

L. P. Hansen, J. Heaton, & A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3) :262–280, 1996.

H. O. Hartley & J. N. K. Rao. A new estimation theory for sample surveys. *Biometrika*, 55 :547–557, 1968.

T. F. Heatherton, J. Polivy, C.P. Herman, & R. F. Baumeister. Self-awareness, task failure, and disinhibition : How attentional focus affects eating. *Journal of Personality*, 61 :49–61, 1993.

J. K. Hellerstein & G. Imbens. imposing moment restrictions from auxiliary data by weighting. *the review of Econometrics and Statistics*, 81(1) :1–14, 1999.

C. P. Herman & J. Polivy. Dieting as an exercise in behavioral economics. In G. Loewenstein, D. Read, & R. F. Baumeister, editors, *Time and Decision*, pages 459–489. Russell Sage Foundation, New-York, 2003.

N. L. Hjort, I. W. McKeague, & I. Van Keilegom. Extending the scope of empirical likelihood. Working Paper n°0414, Institut de Statistique, UCL, 2004.

W. Hoeffding. Probability inequalities for sums of bounded variables. *Journal of the American Statistical Association*, 58 :13–30, 1963.

K. Hoffmann, H. Boeingand, A. Dufour, J. L. Volatier, J. Telman, M. Virtanen, W. Becker, & S. De Henauw. Estimating the distribution of usual dietary intake by short-term measurements. *European Journal of Clinical Nutrition*, 56 :53–62, 2002.

C. Horne. Sociological perspectives on the emergence of norms. In M. Hechter & K-D. Opp, editors, *Social Norms*, pages 3–34. Russell Sage Foundation, New-York, 2001.

J. Horowitz. The bootstrap in econometrics. *Statistical Science*, 18(2) :211–218, 2003.

J. L. Horowitz. Semiparametric and nonparametric estimation of quantal response models. In G. Maddala, C. Rao, & H. Vinod, editors, *Econometrics*, volume 11 of *Handbook of Statistics*, pages 45–72. North-Holland, Amsterdam, 1993.

J. L. Horowitz & S. Lee. Semiparametric methods in applied econometrics : do the models fit the data ? *Statistical Modelling*, 2 :3–22, 2002.

IFREMER. Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO), 1994-1998.

INSEE. Enquête insee, institut national de la statistique et des Études Économiques, la situation démographique en 1999. mouvement de la population et enquête emploi de janvier 1999, 1999.

C. T. Ireland & S. Kullback. contingency tables with given marginals. *biometrika*, 55(1) : 179–188, 1968.

J. Jain & B. Jamison. Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.*, 8 :9–40, 1967.

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

B. Y. Jing & A. T. A. Wood. Exponential empirical likelihood is not bartlett correctable. *Annals of Statistics*, 24 :365–369, 1996.

B.-Y. Jing, Q.-M. Shao, & Q. Wang. Self-normalized Cramér-type large deviations for independent random variables. *Annals of Probability*, 31(4) :2167–2215, 2003.

O. Kallenberg. *Foundations of modern probability*. Springer-Verlag, New York, 2002. Second edition.

R. Kersh & J. Morone. How the personal becomes political : Prohibitions, public health and obesity. Working Paper n°9, Campbell Public Affairs Institute, 2002.

Y. Kitamura. Empirical likelihood methods in econometrics : theory and practice. Working Paper n°1569, Cowles Foundation discussion paper, 2006.

Y. Kitamura. Empirical likelihood methods with weakly dependent processes. *Annals of Statistics*, 25(5) :2084–2102, 1997.

Y. Kitamura & M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65(4) :861–874, 1997.

Y. Kitamura, G. Tripathi, & H. Ahn. Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, 72(6) :1667–1714, 2004.

R. W. Klein & R. H. Spady. An efficient semiparametric estimator for binary response models. *Econometrica*, 61(2) :387–421, 1993.

F. Knight. A predictive view of continuous time processes. *Annals of Probability*, 3 :573–596, 1975.

H. R. Kunsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17 :1217–1241, 1989.

S. N. Lahiri. *Resampling methods for dependent Data*. Springer, 2003.

D. Lakdawalla & T. Philipson. The growth of obesity and technological change : a theoretical and empirical examination. Working Paper n°8946, National Bureau of Economic Research, 2002.

J. Lambe, J. Kearney, C. Leclercq, H.F.J. Zunft, S. De Henauw, C.J.E. Lamberg-Allardt, A. Dunne, & M.J. Gibney. The influence of survey duration on estimates of food intakes and its relevance for public health nutrition and food safety issues. *European Journal of Clinical Nutrition*, 53 :166–173, 2000.

S. Lecocq. Econométrie des systèmes de demande. Working Paper, INRA, CORELA, 2000.

A. J. Lee. *U-Statistics : Theory and Practice*, volume 110 of *Statistics : textbooks and monographs*. Marcel Dekker, Inc, New York, USA, 1990.

L.-F. Lee. Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics*, 65(2) :381–428, 1995.

C. Léonard. Convex conjugates of integral functionals. *Acta Mathematica Hungarica*, 93(4) : 253–280, 2001a.

C. Léonard. Minimization of energy functionals applied to some inverse problems. *Applied mathematics and optimization*, 44(3) :273–297, 2001b.

C. Léonard. Minimizers of energy functionals. *Acta Mathematica Hungarica*, 93(4) :281–325, 2001c.

C. Léonard. Minimizers of energy functional under not very integrable constraints. *Journal of convex analysis*, 10(1) :63–88, 2003a.

C. Léonard. Convex optimization problem arising from probabilistic questions. *Preprint CMAP Ecole Polytechnique*, 506, 2003b. www.cmap.polytechnique.fr/preprint/.

A. Lewbel. Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics*, 97 :145–177, 2000.

A. Lhuissier. Education alimentaire en milieu populaire : des normes en concurrence. *Journal des Anthropologues*, 106-107 :61–76, 2006.

F. Liese & I. Vajda. *Convex Statistical distance*. Teubner, Leipzig, 1987.

L. Lin & R. Zhang. Blockwise empirical euclidean likelihood for weakly dependent processes. *Statistics and Probability Letters*, 53(2) :143–152, 2001.

MAAPAR. Résultats des plans de surveillance pour les produits de la mer. Ministère de l'Agriculture, de l'Alimentation, de la Pêche et des Affaires Rurales, 1998-2002.

P. Major. A multivariate generalization of hoeffding's inequality. Arxiv preprint math. PR/0411288, 2004.

V. K. Malinovskii. *On some asymptotic relations and identities for Harris recurrent Markov Chains*, pages 317–336. 1985.

V. K. Malinovskii. Limit theorems for Harris Markov chains i. *Theory Prob. Appl.*, 31 : 269–285, 1987.

V. K. Malinovskii. Limit theorems for Harris Markov chains ii. *Theory Prob. Appl.*, 34 : 252–265, 1989.

C. F. Manski. Identification of endogenous social effects : The reflection problem. *Review of Economic Studies*, 60 :531–542, 1993.

C. F. Manski. Economic analysis of social interactions. *Journal of Economic Perspectives*, 14 :115–136, 2000.

S. P. Meyn & R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1996.

R. Moffitt & G. Ridder. The econometrics of data combination. In J. Heckman & E. Leamer, editors, *Handbook of Econometrics*, volume 6. North-Holland, Amsterdam, 2005.

P. A. Mykland. Bartlett type of identities. *Annals of Statistics*, 22 :21–38, 1994.

P. A. Mykland. Dual likelihood. *Annals of Statistics*, 23 :396–421, 1995.

National Research Council (NRC) of the national academy of sciences Price. Toxicological effects of methyl mercury. Technical report, National academy press, Washington, DC., 2000.

W. K. Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2) : 99–135, 1990.

W. K. Newey. Efficient estimation of tobit models under conditional symmetry. In W. A. Barnett, J. Powell, & G. Tauchen, editors, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, chapter 11, pages 3–39. Cambridge University Press, New York, 1991.

W. K. Newey & R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1) :219–255, 2004.

V. Nichèle. Les systèmes de demande : théorie et mise en oeuvre. *Master of Paris 1 Lecture's notes*, 83(3) :541–550, 2003.

E. Nummelin. A splitting technique for Harris recurrent chains. *Z. Wahrsch. Verw. Gebiete*, 43 :309–318, 1978.

E. Nummelin. *General irreducible Markov chains and non negative operators*. Cambridge University Press, 1984.

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

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

A. B. Owen. Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18 :90–120, 1990.

A. B. Owen. Empirical likelihood for linear models. *Annals of Statistics*, 19(4) :1725–1747, 1991.

D. Panchenko. Symmetrization approach to concentration inequalities for empirical processes. *Annals of Probability*, 31(4) :2068–2081, 2003.

A. Paraponaris, B. Saliba, & B. Ventelou. Obesity, weight status and employability : Empirical evidence from a french national survey. *Economics and Human Biology*, 3 :241–258, 2005.

T. J. Philipson & R. A. Posner. The long-run growth in obesity as a function of technological change. Working Paper n°7423, National Bureau of Economic Research, 1999.

I. Pinelis. Probabilistic problems and Hotelling's t^2 test under a symmetry condition. *Annals of Statistics*, 22(1) :357–368, 1994.

J. L. Powell. Symmetrically trimmed least squares estimation for tobit models. *Econometrica*, 54(6) :1435–1460, 1986.

P.S. Price, C.L. Curry, P.E. Goodrum, M.N. Gray, J.I. McCrodden, N.W. Harrington, H. Carlson-Lynch, & R.E. Keenan. Monte carlo modeling of time-dependent exposures using a microexposure event approach. *Risk Analysis*, 16(3) :339–348, 1996.

J. Qin. Empirical likelihood in biased sample problems. *Annals of Statistics*, 21(3) :1182–1196, 1993.

Y. S. Qin & J. Lawless. Empirical likelihood and general estimating equations. *Annals of Statistics*, 22(1) :300–325, 1994.

M. M. Rao & Z. D. Ren. *Theory of Orlicz Spaces*. Marcel Dekker, New York, 1991.

D. Revuz. *Markov Chains*. North-Holland, 1984.

F. Régnier. Obésité, corpulence et souci de minceur : inégalités sociales en france et aux etats-unis. *Cahiers de Nutrition et Diététique*, 41(2) :97–103, 2006.

R. T. Rockafellar. Integrals which are convex functionals. *Pacific Journal of Mathematics*, 24 :525–539, 1968.

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

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

J. P. Romano & M. Wolf. Finite sample nonparametric inference and large sample efficiency. *Annals of Statistics*, 28(3) :756–778, 2000.

L. Serra-Majem, D. MacLean, L. Ribas, D. Brule, W. Sekula, R. Prattala, R. Garcia-Closas, A. Yngve, & M. Lalondeand A. Petrasovits. Comparative analysis of nutrition data from national, household, and individual levels : results from a who-cindi collaborative project in canada, finland, poland, and spain. *Journal of Epidemiology and Community Health*, 57 :74–80, 2003.

Q.-M. Shao. Self-normalized large deviations. *Annals of Probability*, 25(1) :285–328, 1997.

R. J. Smith. Efficient information theoretic inference for conditional moment restrictions. Working Paper n°14/05, CeMMAP, 2005.

R. J. Smith. Alternative semi-parametric likelihood approaches to generalised method of moments estimation. *Economic Journal*, 107(441) :503–519, 1997.

W. L. Smith. Regenerative stochastic processes. *Proc. Royal Stat. Soc., Serie A*, 232 :6–31, 1955.

A. Stutzer & R. Lalive. The role of social work norms in job searching and subjective well-being. *Journal of the European Economic Association*, 2 :696–719, 2004.

S. M. Suranovis, R. S. Goldfarb, & T. C. Leonard. An economic theory of cigarette addiction. *Journal of Health Economics*, 18 :1–29, 1999.

J. P. Tangney, P. M. Niedenthal, M. V. Covert, & D. H. Barlow. Are shame and guilt related to distinct self-discrepancies ? a test of higgins's (1987) hypotheses. *Journal of Personality and Social Psychology*, 75 :256–268, 1998.

H. Teicher & Y. S. Chow. *Probability Theory : Independence, Interchangeability, Martingales*. Springer-Verlag, New York, 1988. Second edition.

Y. Thibaud & J. Noël. Evaluation des teneurs en mercure, methyl mercure et sélénium dans les poissons et coquillages des côtes françaises de la méditerranée. Technical report, Rapp. DERO 89-09, IREMER, Nantes, 1989.

T. S. Thompson. Some efficiency bounds for semiparametric discrete choice models. *Journal of Econometrics*, 58(1-2) :257–274, 1993.

H. Thorisson. *Coupling, Stationarity and Regeneration*. Springer, 2000.

D. Tjostheim. Non-linear time series and Markov chains. *Advances in Applied Probability*, 22(3) :587–611, 1990.

J. Tressou. Non parametric modelling of the left censorship of analytical data in food risk exposure assessment, 2006. Working paper.

J. Tressou, A. Crépet, P. Bertail, M. H. Feinberg, & J. C. Leblanc. Probabilistic exposure assessment to food chemicals based on extreme value theory. application to heavy metals from fish and sea products. *Food and Chemical Toxicology*, 42(8) :1349–1358, 2004.

M. Tsao. Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *Annals of Statistics*, 32(3) :1215–1221, 2004.

M. Van Der Laan. *Efficient and Inefficient Estimation in Semiparametric Models*. PhD thesis, Dpt of Statistics, Berkeley, 1967.

T. Veblen. *Theory of the Leisure Class*. Bookseller, New York, 1899.

A. Warde, S-L. Cheng, W. Olsen, & D. Southerton. Changes in the practice of eating : a comparative analysis of time-use. To appear in *Acta Sociologica*, 2006.

WHO. Methylmercury, environmental health criteria 101. Technical report, Geneva, Switzerland, 1990.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9 :60–62, 1938.

S. T. Yen. Working wives and food away from home : The box-cox double hurdle model. *American Journal of Agricultural Economics*, 75 :884–895, 1993.