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

n° 2007-20

Using Empirical Likelihood to Combine Data : Application to Food Risk Assessement

A. CRÉPET[*]
H. HARARI-KERMADEC[*]
J. TRESSOU[*]

Les documents de travail ne reflètent pas la position de l'INSEE et n'engagent que leurs auteurs.

Working papers do not reflect the position of INSEE but only the views of the authors.

[*] INRA Mét@risk, INRA CORELA and CREST, LS.

[*]

[*] The third author is visiting Hong Kong University of Science and Technology, her research is supported in part by Hong Kong RGC Grand #601906.

Using empirical likelihood to combine data: Application to food risk assessment.

Amélie Crépet, Hugo Harari-Kermadec and Jessica Tressou*†

September 19, 2007

Abstract

This paper introduces an original methodology based on empirical likelihood which aims at combining different contamination and consumptions surveys in order to provide risk managers with a risk measure taking account of all the available information. This risk index is defined as the probability that exposure to a contaminant exceeds a safe dose. It is expressed as a non linear functional of the different consumption and contamination distributions, more precisely as a generalized U-statistic. This non linearity and the huge size of the data sets make direct computation of the problem unfeasible. Using linearization techniques and incomplete versions of the U-statistic, a tractable "approximated" empirical likelihood program is solved yielding asymptotic confidence intervals for the risk index. An alternative "Euclidean likelihood program" is also considered, replacing the Kullback-Leibler distance involved in the empirical likelihood by the Euclidean distance. Both methodologies are tested on simulated data and applied to assess the risk due to the presence of methyl mercury in fish and other seafoods.

Résumé

Dans cet article, nous proposons une méthode construite à partir de la vraisemblance empirique qui permet de combiner différents jeux de données, des enquêtes de consommation et des mesures de contamination, dans le cadre de l'estimation d'un risque alimentaire. L'indice de risque pertinent dans ce cadre est la probabilité que l'exposition au contaminant considéré dépasse la dose tolérable. Cet indice est donné par une fonction non linéaire des distributions des consommations et des contaminations qui prend la forme d'une U-statistique. Ce problème de non linéarité, ajouté à la grande taille des jeux de données, rend impossible l'évaluation directe de l'indice de risque. On obtient alors un indice approché à l'aide de U-statistiques incomplètes et de techniques de linéarisation. On obtient alors des intervalles de confiance asymptotique pour l'indice de risque. Nous proposons également une méthode alternative basée sur la vraisemblance empirique Euclidienne, ce qui revient à remplacer la divergence de Kullback-Leibler utilisée par la vraisemblance empirique par la distance Euclidienne. Ces deux méthodes sont validées par des simulations puis appliquées pour évaluer le risque dû à la présence de mé-mercure dans le poisson et dans d'autres produits de la mer.

*INRA Mét@risk, INRA CORELA and CREST LS

†The third author is visiting Hong Kong University of Science and Technology, her research is supported in part by Hong Kong RGC Grant #601906.

Introduction

Certain foods may contain varying amounts of chemicals such as methyl mercury (present in sea food), dioxins (in poultry, meat) or mycotoxins (in cereals, dried fruits, etc.), which may cause major health problems when accumulating inside the body in excessive doses. A commonly used measure of such chronic risks related to the presence of chemical contaminants in food is the probability that the contaminant intake/exposure exceeds a safe dose determined by international experts' committee based on experimental and/or epidemiological studies. A fundamental problem when estimating this food risk index is the diversity of data sources and the scarcity of *good* databases. First, the assessment is most of the time conducted from consumption and contamination data independently available since measuring the exposure directly over long periods of time is not feasible. Moreover, information on the consumption behavior of a given population is obtained through different types of survey (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 also come from different laboratories. Yet, an accurate estimation of the food risk index is crucial since the resulting confidence intervals may serve as arguments for nutritional recommendations or establishment of new standards on the contamination of the food. It is therefore necessary to develop a methodology to build such a confidence interval combining all the available data and side information, such as the main differences between the surveys, known biases or censorship, etc. Data combination is useful in many domains and have been considered from an econometric/economist point of view in (35). It can be also linked to *meta-analysis* techniques mostly used in medical statistics (13; 14; 18). Other methods can be applied to incorporate side information, see (12; 19; 23). The methodology chosen in this paper is based on empirical likelihood techniques introduced by Art B. Owen in (29; 30) as a powerful semiparametric inference method based on a data driven likelihood ratio function. Refer to Owen's book (31) and the references therein for a complete bibliography on the topic. Empirical likelihood is very well adapted to our estimation problem. Indeed, as explained in (37), due to the correlations among the different quantities and the presence of numerous null consumptions, fitting a parametric model to (multidimensional) consumption data is difficult and nonparametric methods are mostly recommended. Moreover, the estimation of the food risk index should include all the available sources of information about consumption and contamination. This kind of estimation problem has already been studied from a theoretical point of view (combination of independent samples for the estimation of their common mean, see (33; 40), or (31) pages 51, 130 and 223-225) but its application to a concrete applied problem raises intractable difficulties in term of computation in the context of food risk assessment. Indeed, data set lengths do not add but multiply, and the combination of say 3 data sets of length 1000 yields a billion triplets. We propose a solution based on U-statistics to handle this difficulty.

The outline of the paper is as follows. Section 1 introduces the framework and notation used in food risk assessment problems and defines the Empirical Likelihood Problem (ELP) which is difficult to solve due to the high nonlinearity of the parameter of interest. Section 2 states the first main result to approximate the ELP solution using linearization techniques, noticing that the food risk index is a generalized U-statistic that can be simplified through its Hoeffding decomposition (see 4). The practical computation of this solution in the muldimensional case is treated in section 3 via incomplete U-statistics. An alternative "Euclidean likelihood program" is considered in section 4, replacing the Kullback-Leibler distance involved in the ELP by the Euclidean distance. Finally, section 5 gives an illustration of these methodologies on true datasets concerning methyl mercury exposure of the French population as well as a validation of these methodologies using simulated datasets. The possible generalizations of these methodologies and the specific extensions in the case of food risk assessment are addressed in section 6. Technical proofs are postponed to an appendix section.

1 Framework and notation

Our goal is 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. For this purpose, $P+R$ data sets are available: R P-dimensional data sets coming from R complementary consumption surveys and the P sets of contamination values. We assume that the R consumption surveys concern the same population but present some specificities calling for adequate calibrations. 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 give a confidence interval for θ_d using empirical likelihood techniques. In the following, we will set R to 2 for simplicity of exposition.

Notation For $k = 1, ..., P$, $Q^{[k]}$ denotes the random variable for the contamination of product k, with distribution $\mathcal{Q}^{[k]}$. $\left(q_l^{[k]}\right)_{l=1,\dots,L_k}$ is an i.i.d. sample of length L_k 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 otherwise.

In the following, r is the consumption survey number and takes the value 1 or 2. $\left(C_1^{(r)}, \dots, C_P^{(r)}\right) = C^{(r)}$ denotes the P-dimensional random variable for the "relative" consumption vector[1], with distribution $\mathcal{C}^{(r)}$. $\left(c_{1,i}^{(r)} \dots c_{P,i}^{(r)}\right)_{1\leq i\leq n_r} = \left(c_i^{(r)}\right)_{1\leq i\leq n_r}$ is an i.i.d. sample of length n_r from $\mathcal{C}^{(r)}$. Its empirical distribution is

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

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.

Empirical likelihood program We define the sets of weights,

$$\mathcal{P} = \left\{\left(p_i^{(1)}\right)_{1\leq i\leq n_1}, \left(p_j^{(2)}\right)_{1\leq j\leq n_2}, \left\{\left(w_{l_k}^{[k]}\right)_{1\leq l_k\leq L_k}\right\}_{1\leq k\leq P}\right\},$$

associated to the 2 samples of consumption and the P samples of contamination. The empirical likelihood is given by

$$\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{k=1}^{P}\prod_{l_k=1}^{L_k} w_{l_k}^{[k]},$$

[1] Consumptions are "relative" consumptions in the sense that they are expressed in terms of individual body weight. This way, the individual exposure can be compared to the safe dose called PTWI, see section 5 for details.

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: $\forall 1 \le k \le P$, $\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} = 1$, and the following model constraints.

Model constraints 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 $k = 1, \ldots, P$. 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 reweighed joint discrete probability distribution of the P contamination samples and the r^{th} consumption survey sample.

The model constraints can now be written, for $r = 1, 2$ and for $\theta \in]0, 1[$,

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

Theses model constraints on θ have an explicit (but unpleasant) expression: for $r = 1, 2$,

$$\sum_{i=1}^{n_r} \sum_{l_1=1}^{L_1} \cdots \sum_{l_k=1}^{L_k} \cdots \sum_{l_P=1}^{L_P} p_i^{(r)} \left(\prod_{j=1}^{P} w_{l_j}^{[j]} \right) \mathbb{1} \left\{ \sum_{k=1}^{P} q_{l_k}^{[k]} c_{k,i}^{(r)} > d \right\} - \theta = 0.$$

2 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, see (4). One solution is to see this plug-in estimator as a generalized U-statistic and to linearize it using Hoeffding decomposition, see Lee's book, (25). 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 (2) for details. Linearization is made easier by considering the influence function of $\Psi_{\mathcal{D}} = \mathbb{E}_{\mathcal{D}} \left[\mathbb{1} \left\{ \sum_{k=1}^{P} Q^{[k]} C_k^{(r)} > d \right\} - \theta \right]$, 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 $r = 1, 2$:

$$\begin{aligned} \Psi_{\mathcal{D}}^{(1)} (q_1, \ldots, q_P, c) = {} & \mathbb{E}_{\prod_{k=1}^{P} \mathcal{Q}_{L_k}^{[k]}} \left[\mathbb{1}_{\sum_{k=1}^{P} Q^{[k]} C_k^{(r)} > d} - \theta \middle| C^{(r)} = c \right] \\ & + \sum_{m=1}^{P} \mathbb{E}_{\mathcal{C}_{n_r}^{(r)} \times \prod_{k \neq m} \mathcal{Q}_{L_k}^{[k]}} \left[\mathbb{1}_{\sum_{k=1}^{P} Q^{[k]} C_k^{(r)} > d} - \theta \middle| Q^{[m]} = q_m \right]. \end{aligned}$$

This functional of $\mathcal{D}$ can be estimated by its empirical counterpart $\Psi^{(1)}_{\widehat{\mathcal{D}}}$, where $\widehat{\mathcal{D}}$ denotes the empirical version of $\mathcal{D}$. $\Psi^{(1)}_{\widehat{\mathcal{D}}}$ can be written explicitly:

$$\Psi^{(1)}_{\widehat{\mathcal{D}}}\left[q_1, \ldots, q_P, c\right] = U_0\left(c\right) + U_1^{(r)}(q_1) + \ldots + U_m^{(r)}(q_m) + \ldots + U_P^{(r)}(q_P), \tag{2}$$

where

$$U_0\left(c\right) = \left(\prod_{k=1}^{P} L_k\right)^{-1} \sum_{\substack{1\leq l_k\leq L_k\\ 1\leq k\leq P}} \mathbb{1}\left\{\sum_{k=1}^{P} q_{l_k}^{[k]} c_k > d\right\} - \theta, \tag{3}$$

and, for $m = 1\cdots P$ and $r = 1, 2$,

$$U_m^{(r)}(q_m) = \left(n_r \times \prod_{\substack{k=1\\ k\neq m}}^{P} L_k\right)^{-1} \sum_{\Lambda^{(r)}_{[-m]}} \mathbb{1}\left\{q_m c_{i,m}^{(r)} + \sum_{\substack{k=1\\ k\neq m}}^{P} q_{l_k}^{[k]} c_{i,k}^{(r)} > d\right\} - \theta, \tag{4}$$

where the sum is taken over the set $\Lambda^{(r)}_{[-m]}$ of all indexes $(i, l_1, \ldots, l_{m-1}, l_{m+1}, \ldots, l_P)$, i.e. fixing the contamination of food m.

$U_0\left(c^{(r)}\right)$ and the $\left(U_m^{(r)}(q^{[m]})\right)_{m=1}^{P}$ are generalized U-statistics with kernel $\mathbb{1}\left\{\sum_{k=1}^{P} q^{[k]} c_k > d\right\}$ and degree $(1, ..., 1) \in \mathbb{R}^P$, see (25). For simplicity, the dependence in $n_r, L_1, ..., L_P$ is implicit in the notation.

An approximate version of the model constraints (1) can now be written:

$$\text{for } r = 1, 2: \qquad \mathbb{E}_{\widetilde{\mathcal{D}}_r}\left[\Psi^{(1)}_{\widehat{\mathcal{D}}}\left(Q^{[1]}, \ldots, Q^{[P]}, C^{(r)}\right)\right] = 0,$$

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 only one product is considered ($P = 1$, $L_1 = L$). The result remains true in the general case, $P > 1$, but needs some refinements to be tractable in practice as detailed in next section.

Theorem 1. *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 involves solving the dual* *program with log likelihood function $l_{n_1,n_2,L}(\theta_d)$ given by*

$$\sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_3\in\mathbb{R}\\ n_1+n_2+L-\gamma_1-\gamma_2-\gamma_3=0}} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \ln\left\{\gamma_1+\lambda_1 U_0\left(c_i^{(1)}\right)\right\} \\ +\sum_{j=1}^{n_2} \ln\left\{\gamma_2+\lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ +\sum_{l=1}^{L} \ln\left\{\gamma_3+\lambda_1 U_1^{(1)}(q_l)+\lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right\}. \tag{5}$$

Define the maximum likelihood estimator associated to this quantity

$$\hat{\theta} = \arg\sup_{\theta} l_{n_1,n_2,L}(\theta).$$

Then, the log-*likelihood ratio*

$$r_{n_1,n_2,L}(\theta_d) = 2\left[l_{n_1,n_2,L}\left(\hat{\theta}\right) - l_{n_1,n_2,L}(\theta_d)\right] \to 4\chi^2(1).$$

The proof of these results is given in appendix A.1. This theorem yields an $(1-\alpha)^{th}$ confidence interval for θ_d given by

$$\left\{\theta : r_{n_1,n_2,L}(\theta) \leq 4\chi^2_{1-\alpha}(1)\right\}.$$

Remark 1. *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 in (31).*

Remark 2. *This model constraints 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 marginal distribution of the population according to different criteria (age, sex, region, profession) and could be integrated via estimating equations of the form*

$$\sum_{i=1}^{n_1} p_i^{(1)} Z_i^{(1)} = z_0, \qquad \sum_{j=1}^{n_2} p_j^{(2)} Z_j^{(2)} = z_0, \tag{6}$$

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 (34) and (31), chapter 3, page 51.

3 Extension to the case of several products by incomplete U-statistics

For $P > 1$, the computation of the different U-statistics defined in (3) and (4) 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 (5) or (25).

Let us define the incomplete U-statistics associated to equations (3) and (4). For simplicity, the sizes of the incomplete U-statistics are fixed to the same constant B, which should be chosen greater than the size of the different data sets involved. B is chosen such as $n_1 + n_2 + \sum_{k=1}^P L_k = o(B)$ in order that the difference between the complete and the incomplete versions is of order $o(B^{-1/2})$, (25). For $r = 1$ or 2, the incomplete version of equation (3) is given by

$$U_{\mathcal{B}_0^{(r)}}\left(c^{(r)}\right) = B^{-1} \sum_{\mathcal{B}_0^{(r)}} \mathbb{1}\left\{\sum_{k=1}^{P} q_{l_k}^{[k]} c_k^{(r)} > d\right\} - \theta, \tag{7}$$

where the sum is taken over the set $\mathcal{B}_0^{(r)}$ of indexes $(l_1, \ldots, l_P)$, randomly chosen with replacement from $\bigotimes_{k=1}^{P} \{1, \ldots, L_k\}$, with size B.

For $m = 1, \ldots, P$, the incomplete version of (4) is given by

$$U_{\mathcal{B}_m^{(r)}}(q_m) = B^{-1} \sum_{\mathcal{B}_m^{(r)}} \mathbb{1}\left\{\sum_{k=1}^{m-1} q_{l_k}^{[k]} c_{i,k}^{(r)} + q_m c_{i,m}^{(r)} + \sum_{k=m+1}^{P} q_{l_k}^{[k]} c_{i,k}^{(r)} > d\right\} - \theta, \tag{8}$$

where the sum is taken over the set $\mathcal{B}_m^{(r)}$ of indexes $(l_1, \ldots, l_{m-1}, l_{m+1}, \ldots, l_P, i)$ randomly chosen with replacement from $\bigotimes_{\substack{k=1 \\ k \neq m}}^{P} \{1, \ldots,\ L_k\} \times \{1 \ldots n_r\}$, with size B.

The approximate influence function is now given by

$$\Psi_B^{(1)}\left(q_1, \ldots, q_P, c^{(r)}\right) = U_{\mathcal{B}_0^{(r)}}\left(c^{(r)}\right) + U_{\mathcal{B}_1^{(r)}}(q_1) + \ldots + U_{\mathcal{B}_m^{(r)}}(q_m) + \ldots + U_{\mathcal{B}_P^{(r)}}(q_P).$$

The model constraints can then be written as follows.

$$\sum_{i=1}^{n_1} p_i^{(1)} U_{\mathcal{B}_0^{(1)}}\left(c_i^{(1)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_{\mathcal{B}_k^{(1)}}\left(q_{l_k}^{[k]}\right)\right] = 0, \tag{9}$$

$$\sum_{j=1}^{n_2} p_j^{(2)} U_{\mathcal{B}_0^{(2)}}\left(c_j^{(2)}\right) + \sum_{k=1}^{P}\left[\sum_{l_k=1}^{L_k} w_{l_k}^{[k]} U_{\mathcal{B}_k^{(2)}}\left(q_{l_k}^{[k]}\right)\right] = 0.$$

Corollary 1. *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 assumptions of Theorem 1, the likelihood ratio for P products, $r_{n_1,n_2,L_1,\ldots,L_P}(\theta_d)$, is asymptotically $\chi^2(1)$:*

$$r_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) \rightarrow (P+1)^2 \chi^2(1).$$

See the appendix A.2 for the proof. Note in particular that B, the size of the incomplete U-statistics, 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 given by

$$\left\{\theta : r_{n_1,n_2,L_1,\ldots,L_P}(\theta) \leq (P+1)^2 \chi^2_{1-\alpha}(1)\right\}.$$

4 A faster alternative: Euclidean likelihood

The empirical likelihood program as written in this paper consists in minimizing the Kullback-Leibler distance between a multinomial distribution 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 (3), we replace the Kullback-Leibler distance by the Euclidean distance (also called the χ^2 distance). When using the Euclidean distance, the objective function $\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta)$ becomes

$$\min_{\left\{\substack{p_i^{(1)},\, p_j^{(2)},\\ w_{l_k}^{[k]},k=1,..,P}\right\}} \frac{1}{2}\begin{bmatrix} \sum_{i=1}^{n_1}\left(n_1 p_i^{(1)}-1\right)^2 \\ +\sum_{j=1}^{n_2}\left(n_2 p_j^{(2)}-1\right)^2 \\ +\sum_{k=1}^{P}\sum_{l_k=1}^{L_k}\left(L_k w_{l_k}^{[k]}-1\right)^2 \end{bmatrix}, \qquad (10)$$

under the approximated model constraints (9) and the constraint that each set of weights sums to 1. We get a result equivalent to Corollary 1:

Corollary 2. *Under the assumptions of Corollary 1, the statistic*

$$\mathbf{r}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) = 2\left[\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta_d) - \inf_{\theta}\mathbf{l}_{n_1,n_2,L_1,\ldots,L_P}(\theta)\right]$$

is asymptotically $(P+1)^2\chi^2(1)$.

The proof of this result is given in appendix A.3.

The choice of this distance is closely related to the Generalized Method of Moments (GMM), see (28; 6) for precisions on the links between empirical likelihood and GMM. Instead of logarithms, the optimization program (10) only involves quadratic terms and is then much easier to solve, as shown in appendix A.3. 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 $w_{l_k}^{[k]}$ are not forced to be positives. However, these weights are asymptotically nonnegative with probability one, see (6).

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. Section 5 illustrates the interest of this strategy in large data sets and a complicated model.

5 Application: Methyl mercury Risk Assessment

In this section, the proposed methodologies based on empirical and euclidean likelihood are applied to methyl mercury risk assessment in the French population. Indeed, at high concentrations, methyl mercury, 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 (41). There is also some concerns that methyl mercury may give rise to retarded development or other neurological effects at lower levels of exposure, which are consistent with standard patterns of fish consumption (11; 17; 27). The latest epidemiological results compiled by the Joint Expert Committee on Food Additives and Contaminants (15) yields a safe dose called *Provisional Tolerable Weekly Intake* (PTWI) for methyl mercury of 1.6 μ g per week per kg of body weight. Methyl mercury is mainly found in fish and other sea foods. Other foods are therefore excluded to estimate human exposure in this paper. 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. The INCA survey records detailed individual food consumption but only on a seven-day basis. We present these data sets together with the contamination data in section 5.1. Then, a validation on simulated data is proposed in section 5.2 following the main features of the actual data sets. Results are shown in section 5.3 and 5.4 considering one single food group ($P = 1$), or two food groups ($P = 2$) respectively.

5.1 Data description and specific features

Contamination data Food contamination data concerning fish and other seafoods 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 (26) and the French Research Institute for Exploitation of the Sea (21). These $L = 2832$ analytical data are expressed in terms of total mercury in mg/kg of fresh weight.

Considering two food groups ("Fish" on one hand and "Mollusks and shellfish" on the other hand), the data set sizes are $L_1 = 1541$ and $L_2 = 1291$. To extrapolate methyl mercury levels from the mercury content, the dangerous form to human health, conversion factors have been applied to the analytical data as 0.84 for fish, 0.43 for mollusk and 0.36 for shellfish, (8; 9).

Adhering to international recommendations (16) the 7% of left censored values, i.e contamination levels below some detection or quantification limit, were replaced with half the detection or quantification limit. Refer to (4; 38) for further discussions.

The INCA survey The French "INCA" survey ($r = 1$), carried out by (10), records $n_1 = 3003$ individual consumptions during one week. The survey is composed of 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 seafood. 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 (15%, (22)): 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 (2). The additional constraint is $\mathbb{E}_{\widetilde{\mathcal{C}}^{(1)}_{n_1}}\left[\mathbb{1}_{3\leq Z_i^{(1)}\leq 14}\right] = 0.15$, 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) in the part concerning the first survey. It becomes

$$\sum_{i=1}^{n_1} \ln \left\{ \gamma_1 + \lambda_1 U_0 \left(c_i^{(1)} \right) + \lambda_{\text{age}} \left(\mathbb{1}_{3 \le Z_i^{(1)} \le 14} - 0.15 \right) \right\},$$

where λ_{age} is the Kühn and Tücker coefficient associated to the "age" constraint.

SECODIP 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 seafoods are retained. Individual consumption is created by inputting to each individual the household's purchase divided by the number of persons in the household, which is a current practice in food risk assessment based on household aquisition data. 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.

Table 1: Basic percentile 95% confidence intervals for MeHg risk (expressed in %)

	INCA	SECODIP
One single product	3.47 [3.06 ; 3.86]	2.24 [1.91 ; 2.57]
Two products	5.68 [4.85 ; 6.40]	2.10 [1.66 ; 2.55]

Differences between the two surveys Some unpublished preliminary studies and basic confidence interval computations of Table 1 show that the use of INCA or SECODIP survey for the exposure estimation to methyl mercury gives 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 (24). 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 (20; 32). These works are based on INCA type data and do not use the available information from SECODIP type data. However, the differences between the two surveys have many explanations:

- the SECODIP panel is an Household Budget Survey. However (36) 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

$$\mathbb{E}(C^{(1)}) = \alpha_0 \mathbb{E}(C^{(2)}).$$

The coefficient α_0 is estimated together with the risk index θ_d, leading to confidence regions for (θ_d, α_0) calibrated by a $\chi^2(2)$ distribution, i.e. $r_{n_1,n_2,L_1}(\theta_d, \alpha_0) \to \chi^2(2)$. We then optimize on α for each θ to get a profiled likelihood on θ.

5.2 Validation on simulated data

In order to validate the proposed methodology, coverage probabilities of the 95% confidence interval resulting from corollary 2 is assessed by simulation of known contamination and consumption distributions as in (4) and (38). We choose to validate the methodology based on the Euclidean likelihood only because solving the Empirical likelihood program takes 2 to 4 hours for large data sets (in the application, we take $B = 10000$). It is therefore difficult to repeat this optimization a large number of times in order to validate the confidence level. Fortunately, the Euclidean likelihood is asymptotically equivalent to the Empirical likelihood and considerably quicker to implement.

The algorithm is as follows:

[Step 1] Define some true distributions of consumption and contaminations and approximate by a Monte Carlo simulation the parameter of interest θ_d.

[Step 2] Reproduce the observed sampling scheme from the true distributions defined in Step 1 and obtain the CI from corollary 2.

Repeat Step 2 S times and check whether the true value of θ_d from Step 1 belongs or not to the CI of Step 2.

For [Step 1], we choose a multivariate log normal distribution for consumption and Gamma distributions for the P contamination distributions[2]. A Monte Carlo simulation of size $1,000,000$ yields a true value of $\theta_{d=1.6} = 0.0529$. In [Step 2], two samples of consumption data are randomly selected from the multivariate log normal distribution determined in [Step 1], one with size $n_1 = 3003$, the other with size $n_2 = 9588$. Then the censorship mechanism is reproduced: the data are first diminished by a random factor with mean 20% to account for consumption outside the home[3].[4] Then [Step 2] is repeated $S = 200$ times.

Results: We obtain a coverage probability of 95.5%. This validate the methodology and is comforted by the know tendency of Euclidean likelihood to be robust, i.e. with the coverage probability converging to the confidence from above.

5.3 Results when considering one global seafood group

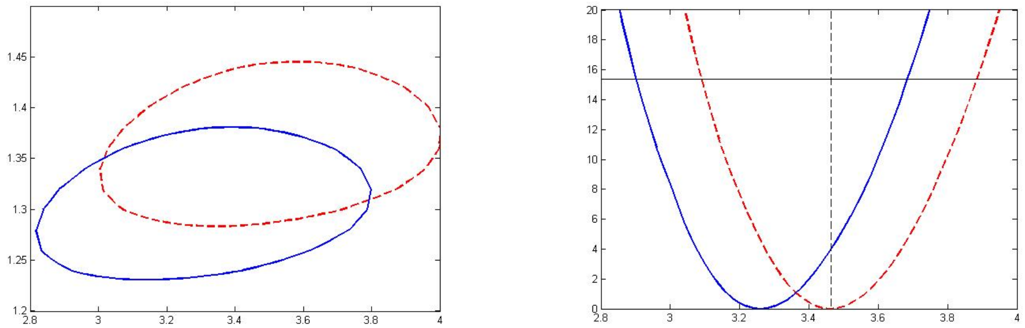
We first merge all the seafoods into a single group. Any contamination data is attributed to the total individual consumption of seafoods. Calculations can therefore be performed using the complete U-statistics of degree $(1, 1)$.

Figure 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 makes 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 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

[2] Their parameters were chosen to fit as much as possible the INCA dataset and the available contamination data.

[3] The proportion of the food eaten at home is distributed according to a Beta distribution with mean 0.8 and variance $0.8(1 - 0.8)$

[4] The only features that are not reproduced are the high proportion of children in sample 1 and the aggregation/disggregation of consumptions within households.



(a) Empirical likelihood confidence regions
horizontal axis is $\theta_{1.6}$,
vertical axis is α

(b) Empirical likelihood ratio profiles
horizontal axis is $\theta_{1.6}$,
vertical axis is $r_{n_1,n_2,L_1}(\theta_{1.6})$

Figure 1: Empirical likelihood for one product (solid with age constraint, dot without)

to the INCA survey, the $\left(p_i^{(1)}\right)_{1\le i\le n_1}$ are modified so that children become less influent, which explains the risk reduction and the decrease of α.

Figure 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 [2.90%; 3.68%] and the risk index estimator is $\theta^*_{1.6} = 3.26\%$. The optimal scaling parameter is $\alpha^* = 1.31$. This is an estimation of the factor to convert individual food purchases of seafoods into individual consumptions of seafoods.

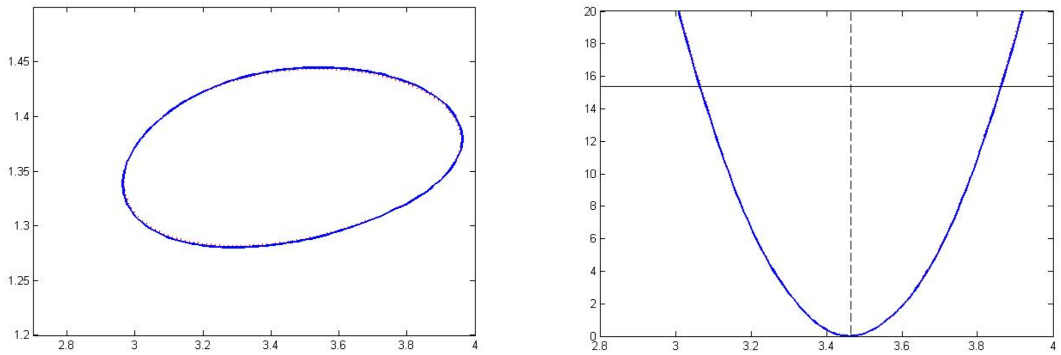
When the constraint on age is ignored, the estimator of $\theta_{1.6}$ is the arithmetic mean of INCA survey and $\alpha-$scaled SECODIP data (marked by the vertical dotted black line). Indeed, the best correction α is when both means are equal and then the maximum of the likelihood for $\theta_{1.6}$ 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 2(a) as well as profiles as shown in Figure 2(b) are almost identical. They give results quite close to what is obtained with the constrained empirical likelihood.

5.4 Results when considering two products

Seafoods are now clustered into two groups: the first one is “Fish” and the second one is “Mollusk and shellfish”. Recall that $L_1 = 1541$ and $L_2 = 1291$. Calculation are done using incomplete U-statistics defined in equations (7) and (8) with a size $B = 10000$. α is here 2-dimensional.

The constraint empirical likelihood confidence interval for the risk index is [4.83%; 6.09%] and the estimator is $\theta^*_{1.6} = 5.43\%$. The correction factors on SECODIP data are $\alpha^*_1 = 1.8$ and $\alpha^*_2 = 1.65$. Figure 3 shows the profiles of the empirical and euclidean likelihood ratios, both with and

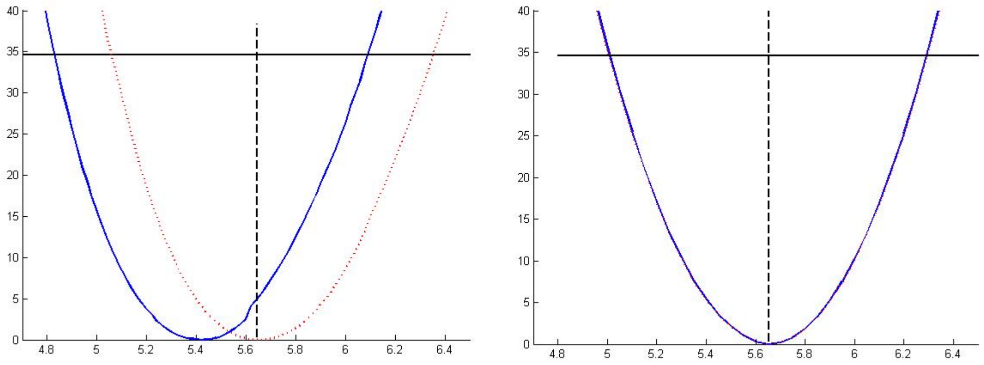


(a) Euclidean likelihood confidence regions horizontal axis is $\theta_{1.6}$, vertical axis is α

(b) Euclidean likelihood ratio profiles horizontal axis is $\theta_{1.6}$, vertical axis is $r_{n_1,n_2,L_1}(\theta_{1.6})$

Figure 2: Euclidean likelihood for one product (solid with age constraint, dot without)

without age constraint. The probability calculated when seafoods are considered as a single group is smaller than when seafoods are gathered into two groups, see also (39). 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. Unfortunately this is not possible with the available SECODIP food nomenclature.



(a) Empirical likelihood ratio profiles (b) Euclidean likelihood ratio profiles horizontal axis are $\theta_{1.6}$ and vertical axis are $r_{n_1,n_2,L_1}(\theta_{1.6})$

Figure 3: Empirical and Euclidean likelihood ratio profiles for two products

6 Discussion

This paper shows how empirical likelihood method can be generalized to combine different sources of data with particular focus on food risk assessment. Yet the methodology is general: if a parameter of interest can be written as a Hadamard differentiable functional of the distributions of random variables for which observations are available then the Approximate Empirical Likelihood Problem has a solution and asymptotic convergence of the likelihood ratio to a chi square distribution was shown. Moreover, when the parameter of interest can be written as a U-statistic, incomplete U-statistics can further be used to compute the associated confidence interval. We demonstrated on simulated data the efficiency of our methodology as far as a food risk index is concerned. Natural extensions could consider more consumption surveys or several contamination data sets, multiplying the number of model constraints and eventually the number of estimating equations referring to side information. The more the Empirical Likelihood Problem gets complicated, the more useful is the Euclidean Likelihood at least to find first step estimators. A technical improvement of the present model 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. (7) proposes a regression based method for the decomposition of household nutritional intakes into individual intakes accounting for outside consumptions, see also (1). In an empirical likelihood program, this kind of 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 future 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.

A Appendix: proofs

A.1 Proof of Theorem 1

First, we consider the empirical likelihood optimization program for two consumption surveys and one food product. Recall that $U_0\left(c\right)$ and $U_1^{(r)}(q)$ are dependent of θ: $U_0\left(c\right) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{q_l c > d} - \theta$ and $U_1^{(r)}(q) = \frac{1}{n_r}\sum_{i=1}^{n_r} \mathbb{1}_{qc_i^{(r)} > d} - \theta$, for $r = 1, 2$.

The program **ELP** is to maximize $\prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{l=1}^{L} w_l$,
under the constraints : $\sum_{i=1}^{L} w_l = 1$, and for $r = 1, 2$, $\sum_{i=1}^{n_r} p_i^{(r)} = 1$, and $\sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) + \sum_{l=1}^{L} w_l U_1^{(r)}(q_l) = 0$.

To carry out this optimization, we take the ln of the **ELP** objective function. This forces the weights to be positive. The difference between these constraints and the nonlinear ones defined in equation (1) is $o(N_r^{-1/2})$ 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

$$\prod_{i=1}^{n_1}\widetilde{p}_i^{(1)}\prod_{j=1}^{n_2}\widetilde{p}_j^{(2)}\prod_{l=1}^{L}\widetilde{w}_l \text{ under the additional constraints:}$$

$$\text{for } r=1,2\ \sum_{i=1}^{n_r}\widetilde{p}_i^{(r)}U_0\left(c_i^{(r)}\right)=0 \text{ and } \sum_{l=1}^{L}\widetilde{w}_lU_1^{(r)}(q_l)=0. \tag{11}$$

The constraints are thus split in two, each constraint concerning only one set of weights. The optimization program is therefore divided in 3 independent sub-programs, the two first ones 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 (34), Theorem 1, we have a control on the order of the optimal weights of each sub-program:

$$\begin{aligned}\widetilde{p}_i^{(r)}&=1/n_r\left(1+t_rU_0\left(c_i^{(r)}\right)\right)^{-1} &&\text{with } t_r=\mathcal{O}(n_r^{-1/2})\\ \widetilde{w}_l&=1/L\left(1+(\tau_1,\ \tau_2)'\left(U_1^{(1)}(q_l),\ U_1^{(2)}(q_l)\right)\right)^{-1} &&\text{with } \tau_r=\mathcal{O}(L^{-1/2}).\end{aligned}$$

The optimum of this new program, which is given by the optimum on each of the 3 sub-programs, is smaller than the **ELP** one, because we added constraints:

$$\prod_{i=1}^{n_1}\widetilde{p}_i^{(1)}\prod_{j=1}^{n_2}\widetilde{p}_j^{(2)}\prod_{l=1}^{L}\widetilde{w}_l\le\prod_{i=1}^{n_1}p_i^{(1)}\prod_{j=1}^{n_2}p_j^{(2)}\prod_{l=1}^{L}w_l.$$

This means that the weights in **ELP** – 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

$$\begin{aligned}\left|\sum_{i=1}^{n_r}p_i^{(r)}U_0\left(c_i^{(r)}\right)-\sum_{i=1}^{n_r}\frac{1}{n_r}U_0\left(c_i^{(r)}\right)\right|&\le\sum_{i=1}^{n_r}\left|p_i^{(r)}-\frac{1}{n_r}\right|\left|U_0\left(c_i^{(r)}\right)\right|\\ \le\sum_{i=1}^{n_r}\left|\widetilde{p}_i^{(r)}-\frac{1}{n_r}\right|\left|U_0\left(c_i^{(r)}\right)\right|&=\frac{1}{n_r}\sum_{i=1}^{n_r}\left|\frac{1}{1+t_rU_0\left(c_i^{(r)}\right)}-1\right|\left|U_0\left(c_i^{(r)}\right)\right|\\ &\le|t_r|\frac{1}{n_r}\sum_{i=1}^{n_r}\left|U_0\left(c_i^{(r)}\right)\right|^2+o(t_r)=\mathcal{O}\left(n_r^{-1/2}\right).\end{aligned} \tag{12}$$

Then, coming back to the original **ELP** program, we have:

$$\left|\sum_{i=1}^{n_r}p_i^{(r)}U_0\left(c_i^{(r)}\right)\right|\le\left|\frac{1}{n_r}\sum_{i=1}^{n_r}U_0\left(c_i^{(r)}\right)\right|+\left|\sum_{i=1}^{n_r}p_i^{(r)}U_0\left(c_i^{(r)}\right)-\sum_{i=1}^{n_r}\frac{1}{n_r}U_0\left(c_i^{(r)}\right)\right|=\mathcal{O}\left(n_r^{-1/2}\right),$$

by standard CLT arguments on $U_0\left(c_i^{(r)}\right)$ and (12). By similar arguments on w_l, we have, for $r=1,2$

$$\sum_{i=1}^{n_r}p_i^{(r)}U_0\left(c_i^{(r)}\right)=\mathcal{O}\left(n_1^{-1/2}\right) \text{ and } \sum_{l=1}^{L}w_lU_1^{(r)}(q_l)=\mathcal{O}\left(L^{-1/2}\right). \tag{13}$$

Lagrangian

The **ELP** program can be rewritten as the following maximum:

$$\max_{w_l,\, \gamma_a\, ,p_i^{(r)},\, \gamma_r,\, \lambda_r} \mathbf{H}\left(w_l,\ \gamma_a\ ,p_i^{(r)},\ \gamma_r,\ \lambda_r\right), \text{ where:}$$

$$\begin{aligned}\mathbf{H}\left(w_l,\ \gamma_a,\ p_i^{(r)},\ \gamma_r,\ \lambda_r\right) &= \ln\left(\textstyle\prod_{i=1}^{n_1} p_i^{(1)} \prod_{i=1}^{n_2} p_i^{(2)} \prod_{l=1}^{L} w_l\right) - \gamma_a\left[\textstyle\sum_{i=1}^{L} w_l - 1\right] \\ &\quad - \textstyle\sum_{r=1}^{2}\left\{\gamma_r\left[\sum_{i=1}^{n_r} p_i^{(r)} - 1\right] - \lambda_r\left[\sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) + \sum_{l=1}^{L} w_l U_1^{(r)}(q_l)\right]\right\}.\end{aligned}$$

Using $\partial\mathbf{H}/\partial p_i^{(r)} = 1/p_i^{(r)} - \gamma_r - \lambda_r U_0\left(c_i^{(r)}\right) = 0$ and the similar expression for $\partial\mathbf{H}/\partial w_l$ gives that

$$p_i^{(r)} = \left(\gamma_r + \lambda_r U_0\left(c_i^{(r)}\right)\right)^{-1} \text{ and } w_l = \left(\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right)^{-1}. \tag{14}$$

Note that we also have

$$\sum_{i=1}^{n_r} p_i^{(r)} \frac{\partial\mathbf{H}}{\partial p_i^{(r)}} = n_r - \gamma_r - \lambda_r \sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_r^{(1)}\right) = 0 \tag{15}$$

and using the constraints, we get that

$$0 = \sum_{i=1}^{n_1} p_i^{(1)} \frac{\partial\mathbf{H}}{\partial p_i^{(1)}} + \sum_{i=1}^{n_2} p_i^{(2)} \frac{\partial\mathbf{H}}{\partial p_i^{(2)}} + \sum_{i=1}^{L} w_l \frac{\partial\mathbf{H}}{\partial w_l} = n_1 + n_2 + L - \gamma_1 - \gamma_2 - \gamma_a. \tag{16}$$

The **ELP** problem can be rewritten using (14) and (16) in the dual form

$$\sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_a\in\mathbb{R} \\ n_1+n_2+L-\gamma_1-\gamma_2-\gamma_a=0}} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \ln\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right)\right\} + \sum_{j=1}^{n_2} \ln\left\{\gamma_2 + \lambda_2 U_0\left(c_i^{(2)}\right)\right\} \\ + \sum_{l=1}^{L} \ln\left\{\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right\}.$$

Furthermore, combining (15) with $\sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) = \mathcal{O}(n_r^{-1/2})$ gives that

$$\gamma_r = n_r + v_r \text{ with } v_r = \lambda_r \cdot \mathcal{O}(n_r^{-1/2})$$

and then

$$p_i^{(r)} = \left(n_r + v_r + \lambda_r U_0\left(c_i^{(r)}\right)\right)^{-1}, \text{ and } w_l = \left(L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right)^{-1}.$$

Let us consider the case of the w_l. Adapting Owen's proof, equation (13) for $r = 1$ combined with (14) yields a constraint given by

$$\begin{aligned}\mathcal{O}\left(L^{-1/2}\right) &= \sum_{i=1}^{L} w_l U_1^{(1)}(q_l) = \sum_{i=1}^{L} \frac{U_1^{(1)}(q_l)}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)} \\ &= \sum_{i=1}^{L} \frac{U_1^{(1)}(q_l)}{L} - \frac{1}{L} \sum_{i=1}^{L} \frac{\left[-v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right] \cdot U_1^{(1)}(q_l)}{L - v_1 - v_2 + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)} \\ &= \overline{U_1^{(1)}} - \frac{\lambda_1}{L} \sum_{i=1}^{L} w_l \left[U_1^{(1)}(q_l)\right]^2 - \frac{\lambda_2}{L} \sum_{i=1}^{L} w_l U_1^{(1)}(q_l) U_1^{(2)}(q_l) + \frac{v_1 + v_2}{L} \sum_{i=1}^{L} w_l U_1^{(1)}(q_l),\end{aligned}$$

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 $\mathcal{O}\left(L^{-1/2}\right)$. We get

$$\overline{U_1^{(1)}} = \frac{\lambda_1}{L}\sum_{i=1}^{L} w_l \left[U_1^{(1)}(q_l)\right]^2 + \frac{\lambda_2}{L}\sum_{i=1}^{L} w_l U_1^{(1)}(q_l)U_1^{(2)}(q_l) + \mathcal{O}\left(L^{-1/2}\right).$$

Using Owen's arguments, we obtain

$$\overline{U_1^{(1)}} + \mathcal{O}\left(L^{-1/2}\right) = \frac{\lambda_1}{L}\overline{\left[U_1^{(1)}\right]^2} + \frac{\lambda_2}{L}\overline{U_1^{(1)}U_1^{(2)}} \text{ and } \overline{U_1^{(2)}} + \mathcal{O}\left(L^{-1/2}\right) = \frac{\lambda_2}{L}\overline{\left[U_1^{(2)}\right]^2} + \frac{\lambda_1}{L}\overline{U_1^{(1)}U_1^{(2)}},$$

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:

$$\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} = L\begin{bmatrix}\overline{\left[U_1^{(1)}\right]^2} & \overline{U_1^{(1)}U_1^{(2)}}\\ \overline{U_1^{(1)}U_1^{(2)}} & \overline{\left[U_1^{(2)}\right]^2}\end{bmatrix}^{-1}\begin{pmatrix}\overline{U_1^{(1)}} + \mathcal{O}\left(L^{-1/2}\right)\\ \overline{U_1^{(2)}} + \mathcal{O}\left(L^{-1/2}\right)\end{pmatrix}. \tag{17}$$

As the empirical variance-covariance matrix convergences to the non-degenerated variance-covariance matrix $\mathbb{E}_{\mathbb{P}}[(U_1^{(1)},U_1^{(2)})'(U_1^{(1)},U_1^{(2)})]$ 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

$$\lambda_r = n_r\left(\overline{[U_0^{(r)}]^2}\right)^{-1}\overline{U_0^{(r)}} + \mathcal{O}(n_r^{1/2}), \tag{18}$$

where $\overline{U_0^{(r)}} = n_r^{-1}\sum_{i=1}^{n_r} U_0(c_i^{(r)})$ and $\overline{[U_0^{(r)}]^2} = n_r^{-1}\sum_{i=1}^{n_r}\left[U_0(c_i^{(r)})\right]^2$.

Now that we control the size of λ_r at the optimum for both n_r and L with (17) and (18), the ln can be expanded around zero, and the dominant terms are the same as for the Euclidean likelihood, which is considered here after. This gives the expected convergence:

$$r_{n_1,n_2,L}(\theta_d) = 2\left(l_{n_1,n_2,L}(\theta_d) - l_{n_1,n_2,L}(\hat{\theta})\right) \xrightarrow[n\to\infty]{\mathcal{L}} 4\chi^2(1).$$

A.2 Proof of Corollary 1, case $P > 1$

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 $U_{a,B}^{(r)}$ and $U_{b,B}^{(r)}$. 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 :

$$\sum_{i=1}^{L_a} w_l^{[a]} = 1,\ \sum_{i=1}^{L_b} w_l^{[b]} = 1 \text{ and for } r = 1,2\ :\ \sum_{i=1}^{n_r} p_i^{(r)} = 1,$$

$$\text{and } \sum_{i=1}^{n_r} p_i^{(r)} U_{0,\mathcal{B}_0^{(r)}}(c_i^{(r)}) + \sum_{l=1}^{L_a} w_l^{[a]} U^{(r)}_{a,\mathcal{B}_a^{(r)}}(q_l^{[a]}) + \sum_{l=1}^{L_b} w_l^{[b]} U^{(r)}_{b,\mathcal{B}_b^{(r)}}(q_l^{[b]}) = 0.$$

For $r = 1, 2$ and $k = a, b$, we can check as above that

$$\sum_{i=1}^{n_r} p_i^{(r)} U_{0,\mathcal{B}_0^{(r)}}\left(c_i^{(r)}\right) = \mathcal{O}\left(n_r^{-1/2}\right), \text{ and } \sum_{l=1}^{L_k} w_l U^{(r)}_{k,\mathcal{B}_k^{(r)}}\left[q_l^{[k]}\right] = \mathcal{O}\left(L_k^{-1/2}\right).$$

We get therefore for $r = 1, 2$ and $k = a, b$:

$$\begin{aligned} p_i^{(r)} &= \left(n_r + v_r + \lambda_r U_{0,\mathcal{B}_0^{(r)}}\left(c_i^{(r)}\right)\right)^{-1} \text{ and} \\ w_l^{[k]} &= \left(L_k + v_k + \lambda_1 U^{(1)}_{k,\mathcal{B}_k^{(1)}}\left(q_l^{[k]}\right) + \lambda_2 U^{(2)}_{k,\mathcal{B}_k^{(2)}}\left(q_l^{[k]}\right)\right)^{-1}, \end{aligned}$$

with $v_1 + v_2 + v_a + v_b = 0$ and the proof follows the same lines as for 1 product.

A.3 Euclidean likelihood (Proof of Corollary 2)

The objective function of the program is now

$$\frac{1}{2} \min_{\left\{p_i^{(1)},\, p_i^{(2)},\, w_{l_k}^{[k]}\right\}} \sum_{r=1}^{2} \sum_{i=1}^{n_r} \left(n_r p_i^{(r)} - 1\right)^2 + \sum_{k=1}^{P} \sum_{l_k=1}^{L_k} \left(L_k w_{l_k}^{[k]} - 1\right)^2.$$

We get then simpler expressions, which allow to reach explicit solutions for the weights.

Closed expression of 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

$$\frac{1}{2} \min_{\left\{p_i^{(1)},\, p_i^{(2)},\, w_l\right\}} \sum_{i=1}^{n_1} \left(n_1 p_i^{(1)} - 1\right)^2 + \sum_{i=1}^{n_2} \left(n_2 p_i^{(2)} - 1\right)^2 + \sum_{l=1}^{L} (L w_l - 1)^2,$$

under the constraints :

$$\sum_{i=1}^{L} w_l = 1 \text{ and for } r = 1, 2 \; : \; \sum_{i=1}^{n_r} p_i^{(r)} = 1, \text{ and } \sum_{i=1}^{n_r} p_i^{(r)} U_0\left(c_i^{(r)}\right) + \sum_{l=1}^{L} w_l U_1^{(r)}(q_l) = 0.$$

Define the corresponding Lagrangian

$$\begin{aligned} \mathbf{H}(\cdot) = {} & \frac{1}{2} \sum_{i=1}^{n_1} \left(n_1 p_i^{(1)} - 1\right)^2 + \frac{1}{2} \sum_{i=1}^{n_2} \left(n_2 p_i^{(2)} - 1\right)^2 + \frac{1}{2} \sum_{l=1}^{L} (L w_l - 1)^2 \\ & - \lambda_1 \left[\sum_{i=1}^{n_1} p_i^{(1)} U_0\left(c_i^{(1)}\right) + \sum_{l=1}^{L} w_l U_1^{(1)}(q_l)\right] - \lambda_2 \left[\sum_{i=1}^{n_2} p_i^{(2)} U_0\left(c_i^{(2)}\right) + \sum_{l=1}^{L} w_l U_1^{(2)}(q_l)\right] \\ & - \gamma_1 \left[\sum_{i=1}^{n_1} p_i^{(1)} - 1\right] - \gamma_2 \left[\sum_{i=1}^{n_2} p_i^{(2)} - 1\right] - \gamma_a \left[\sum_{i=1}^{L} w_l - 1\right]. \end{aligned}$$

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 $p_i^{(r)} = 1/n_r + \left[\gamma_r + \lambda_r U_0(c_i^{(r)})\right]/n_r^2$. As the weights sum to 1, we have

$$1 = \sum_{i=1}^{n_r} p_i^{(r)} = 1 + \left(\gamma_r + \lambda_r \overline{U_0^{(r)}}\right)/n_r, \text{ and then } \gamma_r = -\lambda_r \overline{U_0^{(r)}}.$$

Finally, we get

$$p_i^{(r)} = \frac{1}{n_r} + \frac{\lambda_r}{n_r^2}\left[U_0(c_i^{(r)}) - \overline{U_0^{(r)}}\right] \text{ and } w_l = \frac{1}{L} + \frac{\lambda_1}{L^2}\left[U_1^{(1)}(q_l) - \overline{U_1^{(1)}}\right] + \frac{\lambda_2}{L^2}\left[U_1^{(2)}(q_l) - \overline{U_1^{(2)}}\right].$$

Asymptotic distribution of the U-statistics
The constraints can be rewritten, for $r = 1, 2$:

$$\begin{aligned}
\overline{U_0^{(1)}} + \overline{U_1^{(1)}} + \lambda_1\left[\frac{\mathbb{V}(U_0^{(1)})}{n_1} + \frac{\mathbb{V}(U_1^{(1)})}{L}\right] + \lambda_2 \frac{\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})}{L} &= 0,\\
\overline{U_0^{(2)}} + \overline{U_1^{(2)}} + \lambda_2\left[\frac{\mathbb{V}(U_0^{(2)})}{n_2} + \frac{\mathbb{V}(U_1^{(2)})}{L}\right] + \lambda_1 \frac{\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})}{L} &= 0,
\end{aligned}$$

where $\mathbb{V}$ and $\mathbb{C}ov$ denote the empirical variance operator, $\mathbb{V}(X) = \overline{(X^2)} - \left(\overline{X}\right)^2$, and the covariance operator, $\mathbb{C}ov(X,Y) = \overline{(X \cdot Y)} - \overline{X} \cdot \overline{Y}$. These terms do not depend on θ.

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

$$\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = -2\begin{bmatrix} \frac{\mathbb{V}(U_0^{(1)})}{n_1} + \frac{\mathbb{V}(U_1^{(1)})}{L} & \frac{\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})}{L} \\ \frac{\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})}{L} & \frac{\mathbb{V}(U_0^{(2)})}{n_2} + \frac{\mathbb{V}(U_1^{(2)})}{L} \end{bmatrix}^{-1}\begin{pmatrix} \overline{U^{(1)}} \\ \overline{U^{(2)}} \end{pmatrix}.$$

The optimal value can thus be computed explicitly. Finally, replacing the values of the weights and the λ's in the optimization program, we get:

$$l_{n_1,n_2,L}(\theta) = \frac{4}{2} Y(\theta)' M^{-1} Y(\theta), \text{ where } Y(\theta) = \begin{pmatrix} \sqrt{N_1}\,\overline{U^{(1)}} \\ \sqrt{N_2}\,\overline{U^{(2)}} \end{pmatrix}$$

$$\text{and } M = \begin{bmatrix} \frac{N_1}{n_1}\mathbb{V}(U_0^{(1)}) + \frac{N_1}{L}\mathbb{V}(U_1^{(1)}) & \frac{\sqrt{N_1 N_2}}{L}\mathbb{C}ov(U_1^{(1)}, U_1^{(2)}) \\ \frac{\sqrt{N_1 N_2}}{L}\mathbb{C}ov(U_1^{(1)}, U_1^{(2)}) & \frac{N_2}{n_2}\mathbb{V}(U_0^{(2)}) + \frac{N_2}{L}\mathbb{V}(U_1^{(2)}) \end{bmatrix}.$$

$\overline{U^{(r)}} = \frac{1}{n_1 L}\sum_{i,l} \mathbb{1}_{q_l c_i^{(r)} > d} - \theta$ is a generalized U-statistic with kernel $\mathbb{1}_{q_l c_i^{(r)} > d} - \theta$ and of degree $(1,1)$. The CLT for U-statistics ensures that, with $N_r = n_r + L$, $n_r/N_r \to \eta_r$, and $L/N_r \to \beta_r$,

$$\sqrt{N_r}\,\overline{U^{(r)}} \xrightarrow[n_r, L \to \infty]{\mathcal{L}} \mathcal{N}(\theta_d - \theta, S_r^2),$$

where $S_r^2 = \frac{1}{\eta_r}\mathbb{V}\left[\psi_{\mathcal{C}}\right] + \frac{1}{\beta_r}\mathbb{V}\left[\psi_Q\right]$ and where $\psi_{\mathcal{C}}$ and ψ_Q are the gradients of order 1 of the U-statistic. We consider now the asymptotic covariance C_{12} of these two statistics i.e. the limit of $\sqrt{N_1 N_2}\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})$. To calculate C_{12}, we set $X_{il}^{(r)} = \mathbb{1}_{q_l c_i^r > d} - \theta$, and we have

$$Y(\theta) = \left(\begin{array}{c} \sqrt{N_1}\,\overline{U^{(1)}} \\ \sqrt{N_2}\,\overline{U^{(2)}} \end{array} \right) = \left(\begin{array}{c} \frac{\sqrt{N_1}}{n_1 L}\sum_{il} X_{il}^{(1)} \\ \frac{\sqrt{N_2}}{n_2 L}\sum_{il} X_{il}^{(2)} \end{array} \right).$$

As $\mathbb{E}\left[X_{il}^{(1)} X_{jk}^{(2)}\right] = 0$ for $l \neq k$, we have

$$\begin{aligned} \frac{\sqrt{N_1 N_2}}{n_1 n_2 L^2}\mathbb{C}ov(\sum_{il} X_{il}^{(1)}, \sum_{il} X_{il}^{(2)}) &= \frac{\sqrt{N_1 N_2}}{n_1 n_2 L^2}\mathbb{E}\left[\sum_{iljk} X_{il}^{(1)} X_{jk}^{(2)}\right] \\ &= \frac{\sqrt{N_1 N_2}}{n_1 n_2 L^2}\mathbb{E}\left[\sum_{ilj} X_{il}^{(1)} X_{jl}^{(2)}\right] = \frac{\sqrt{N_1 N_2}}{L} v_1^{(12)}, \end{aligned}$$

where $v_1^{(12)} = \mathbb{E}[X_{il}^{(1)} X_{jl}^{(2)}]$. Therefore, $C_{12} = (\beta_1 \beta_2)^{-1/2} v_1^{(12)}$ and $Y(\theta)$ is asymptotically a gaussian vector,

$$\text{with mean } \left(\begin{array}{c} \theta_d - \theta \\ \theta_d - \theta \end{array} \right) \text{ and variance } M_\infty = \left[\begin{array}{cc} S_1^2 & C_{12} \\ C_{12} & S_2^2 \end{array} \right].$$

Convergence of the pivotal statistic
We must now show that M is a convergent estimator for M_∞. By classical results on U-statistics, we have

$$\widehat{S}_r^2 = \frac{N_r}{n_r}\mathbb{V}(U_0^{(r)}) + \frac{N_r}{L}\mathbb{V}(U_1^{(r)}) \to S_r^2.$$

Let's show that $M_{12} = \frac{\sqrt{N_1 N_2}}{L}\mathbb{C}ov(U_1^{(1)}, U_1^{(2)}) \to C_{12}$. Let $i \leq n_1$ and $j \leq n_2$,

$$M_{12} = \frac{\sqrt{N_1 N_2}}{L}\mathbb{C}ov(U_1^{(1)}, U_1^{(2)}) = \frac{\sqrt{N_1 N_2}}{L^2}\sum_l X_{il}^{(1)} X_{jl}^{(2)} - \frac{\sqrt{N_1 N_2}}{L^3}\left(\sum_l X_{il}^{(1)}\right)\left(\sum_l X_{jl}^{(2)}\right).$$

By the LLN,

$$\frac{1}{L}\sum_l X_{il}^{(1)} X_{jl}^{(2)} \to v_1^{(12)} \text{ and } \frac{1}{L}\sum_l X_{il}^{(r)} \to 0$$

and then

$$M \to M_\infty.$$

To establish the convergence of $\mathbf{r}_{n_1,n_2,L}(\theta_d) = 2\left[\mathbf{l}_{n_1,n_2,L}(\theta_d) - \inf_\theta \mathbf{l}_{n_1,n_2,L}(\theta)\right]$, we consider $\hat{\theta}$, the minimiser of $\mathbf{l}_{n_1,n_2,L}(\theta)$, i.e. of $Y(\theta)' M^{-1} Y(\theta)$. Write $Y(\theta) = Z - \theta 1_2$. The first order condition gives : $-1_2' M^{-1} Z + \hat{\theta} 1_2' M^{-1} 1_2 = 0$ and then $\hat{\theta} = \frac{1_2' M^{-1} Z}{\theta 1_2' M^{-1} 1_2}$. Thus

$$\begin{aligned} \mathbf{r}_{n_1,n_2,L}(\theta_d) &= 4Y(\theta_d)' M^{-1} Y(\theta_d) - 4Y(\hat{\theta})' M^{-1} Y(\hat{\theta}) \\ &= 4\left(Z' M^{-1} Z - 2\theta_d 1_2' M^{-1} Z + \theta_d^2 1_2' M^{-1} 1_2\right) \\ &\quad - 4\left(Z' M^{-1} Z - 2\hat{\theta} 1_2' M^{-1} Z + (\hat{\theta})^2 1_2' M^{-1} 1_2\right) \end{aligned}$$

$$= 4\left[-2\theta_d 1_2' M^{-1} Z + \theta_d^2 1_2' M^{-1} 1_2 + 2\hat{\theta} 1_2' M^{-1} Z - (\hat{\theta})^2 1_2' M^{-1} 1_2\right].$$

By the first order condition, $1_2' M^{-1} Z = \hat{\theta} 1_2' M^{-1} 1_2$ and then

$$\begin{aligned}\mathbf{r}_{n_1,n_2,L}(\theta_d) &= 4\left[-2\theta_d \hat{\theta} 1_2' M^{-1} 1_2 + \theta_d^2 1_2' M^{-1} 1_2 + (\hat{\theta})^2 1_2' M^{-1} 1_2\right] \\ &= 4(\hat{\theta} - \theta_d)^2 1_2' M^{-1} 1_2 = 4\frac{\left(1_2' M^{-1}(Z - 1_2\theta_d)\right)^2}{1_2' M^{-1} 1_2}.\end{aligned}$$

$M^{-1/2}(Z - 1_2\theta_d)$ is asymptotically a standard gaussian vector. $\mathbf{r}_{n_1,n_2,L}(\theta_d)$ is then twice the square of a weighted mean of two independent standard gaussians, and then

$$\mathbf{r}_{n_1,n_2,L}(\theta_d) \xrightarrow{\mathcal{L}} 4\chi^2(1).$$

Case $P > 1$
We also use this framework for the 2 surveys 2 products context ($P = 2$). 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 calculus

$$l_{n_1,n_2,L_1,L_2}(\theta) = (2+1)^2 \left(\overline{U^{(1)}}, \overline{U^{(2)}}\right) A \left(\overline{U^{(1)}}, \overline{U^{(2)}}\right)'$$

$$\text{where } A = \begin{bmatrix} \frac{\mathbb{V}(U_0^{(1)})}{n_1} + \frac{\mathbb{V}(U_1^{(1)})}{L_1} + \frac{\mathbb{V}(U_2^{(1)})}{L_2} & \frac{\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})}{L_1} + \frac{\mathbb{C}ov(U_2^{(1)}, U_2^{(2)})}{L_2} \\ \frac{\mathbb{C}ov(U_1^{(1)}, U_1^{(2)})}{L_1} + \tfrac{\mathbb{C}ov(U_2^{(1)}, U_2^{(2)})}{L_2} & \frac{\mathbb{V}(U_0^{(2)})}{n_2} + \frac{\mathbb{V}(U_1^{(2)})}{L_1} + \frac{\mathbb{V}(U_2^{(2)})}{L_2} \end{bmatrix}^{-1}$$

and the result follows.

References

[1] Allais, O. and Tressou, J. (2006). Using decomposed household food acquisitions as inputs of a Kinetic Dietary Exposure Model. Available at `https://hal.archives-ouvertes.fr/hal-00139914`.

[2] Bertail, P. (2006). Empirical likelihood in some semi-parametric models. *Bernoulli 12*, 299–331.

[3] Bertail, P., Harari-Kermadec, H., and Ravaille, D. (2006). γ-divergence empirique et vraisemblance empirique généralisée. *Annales dEconomie et de Statistique*. In press.

[4] Bertail, P. and Tressou, J. (2006). Incomplete generalized U-statistics for food risk assessment. *Biometrics* **62**, 1. In press.

[5] Blom, G. (1976). Some properties of incomplete U-statistics. *Biometrika 63*, 573–580.

[6] Bonnal, H. and Renault, E. (2004). On the efficient use of the informational content of estimating equations: Implied probabilities and euclidean empirical likelihood. *Cahiers scientifiques (CIRANO) 2004s-18.*

[7] Chesher, A. (1997). Diet revealed?: Semiparametric estimation of nutrient intake-age relationships. *Journal of the Royal Statistical Society A* **160**, 3, 389–428.

[8] Claisse, D., Cossa, D., Bretaudeau-Sanjuan, G., Touchard, G., and Bombled, B. (2001). Methylmercury in molluscs along the French coast. *Marine Pollution Bulletin 42*, 329–332.

[9] Cossa, D., Auger, D., Averty, B., Lucon, M., Masselin, P., Noel, J., and San-Juan, J. (1989). Atlas des niveaux de concentration en métaux métalloïdes et composés organochlorés dans les produits de la pêche côtière franaise. Tech. rep., IFREMER, Nantes.

[10] CREDOC-AFFSA-DGAL. (1999). *Enquête INCA (individuelle et nationale sur les consommations alimentaires)*, TEC&DOC ed. Lavoisier, Paris. (Coordinateur : J.L. Volatier).

[11] Davidson, P., Myers, G., Cox, C., Shamlaye, C. F., Clarkson, T., Marsh, D., Tanner, M., Berlin, M., Sloane-Reves, J., Cernichiari, E., Choisy, O., Choi, A., and Clarkson, T. W. (1995). Longitudinal neurodevelopmental study of Seychellois children following in utero exposure to MeHg from maternal fish ingestion: Outcomes at 19-29 months. *Neurotoxicology 16*, 67–688.

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

[13] Edger, M. and Smith, G. (1997a). Meta-analysis. potentials and promise. *BMJ 315*, 1371–1374.

[14] Edger, M. and Smith, G. (1997b). Meta-analysis: principles and procedures. *BMJ 315*, 1533–1537.

[15] FAO/WHO. (2003). 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.

[16] GEMs/Food-WHO. (1995). Reliable evaluation of low-level contamination of food, workshop in the frame of GEMS/Food-EURO. Tech. rep., Kulmbach, Germany, 26-27 May 1995.

[17] Grandjean, P., Weihe, P., White, R., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sorensen, N., Dahl, R., and Jorgensen, P. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicology Teratology 19*, 41–428.

[18] Hedges, L. and Olkin, I. (1985). *Statistical methods for meta-analysis.* Academic Press, Orlando, FL.

[19] Hellerstein, J. K. and Imbens, G. (1999). Imposing moment restrictions from auxiliary data by weighting. *The review of Econometrics and Statistics* **81**, 1, 1–14.

[20] Hoffmann, K., Boeingand, H., Dufour, A., Volatier, J. L., Telman, J., Virtanen, M., Becker, W., and Henauw, S. D. (2002). Estimating the distribution of usual dietary intake by short-term measurements. *European Journal of Clinical Nutrition 56*, 53–62.

[21] IFREMER. (1994-1998). Résultat du réseau national d'observation de la qualité du milieu marin pour les mollusques (RNO).

[22] INSEE, Institut National de la Statistique et des Etudes Economiques. (1999). La situation démographique en 1999 - Mouvements de la population et enquête emploi de janvier 1999. Tech. rep.

[23] Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55**, 1, 179–188.

[24] Lambe, J., Kearney, J., Leclercq, C., Zunft, H., Henauw, S. D., Lamberg-Allardt, C., Dunne, A., and Gibney, M. (2000). 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*, 16–173.

[25] Lee, A. J. (1990). *U-Statistics: Theory and Practice.* Statistics: textbooks and monographs, Vol. **110**. Marcel Dekker, Inc, New York, USA.

[26] MAAPAR. (1998-2002). 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.

[27] National Research Council (NRC) of the National Academy of Sciences Price. (2000). Toxicological effects of methyl mercury. Tech. rep., National Academy Press, Washington, DC.

[28] Newey, W. K. and Smith, R. J. (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators. *Econometrica* **72**, 1, 219–255.

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

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

[31] Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall/CRC.

[32] Price, P., Curry, C., P.E.Goodrum, M.N.Gray, McCrodden, J., N.W.Harrington, Carlson-Lynch, H., and Keenan, R. (1996). Monte carlo modeling of time-dependent exposures using a microexposure event approach. *Risk Analysis* **16**, 3, 339–348.

[33] Qin, J. (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics* **21**, 3, 1182–1196.

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

[35] Ridder, G. and Moffitt, R. (2006). *Handbook of econometrics*, Heckman and Leamer ed. Elsevier, North-Holland, Amsterdam, Chapter The econometrics of data combination. See `http://www-rcf.usc.edu/~ridder/Wpapers/comsamp7nov03.pdf`.

[36] Serra-Majem, L., MacLean, D., Ribas, L., Brule, D., Sekula, W., Prattala, R., Garcia-Closas, R., Yngve, A., and Petrasovits, M. L. A. (2003). 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.

[37] Tressou, J. (2005). Méthodes statistiques pour l'évaluation du risque alimentaire. Ph.D. thesis, Université Paris X. Available at `http://tel.archives-ouvertes.fr/tel-00139909`.

[38] Tressou, J. (2006). Non parametric modelling of the left censorship of analytical data in food risk exposure assessment. Working paper.

[39] Tressou, J., Crépet, A., Bertail, P., Feinberg, M. H., and Leblanc, J. C. (2004). 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.

[40] Tsao, M. and Wu, C. (2006). Empirical likelihood inference for a common mean in the presence of heteroscedasticity. *Canadian Journal of Statistics* **34**, 1, 45–59.

[41] WHO. (1990). Methylmercury, Environmental Health Criteria 101. Tech. rep., Geneva, Switzerland.