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

Amélie Crépet, Hugo Harari-Kermadec, Jessica Tressou

▶ **To cite this version:**

Amélie Crépet, Hugo Harari-Kermadec, Jessica Tressou. Using empirical likelihood to combine data: Application to food risk assessment. Biometrics, 2009, 65 (1), pp.257-266. ⟨10.1111/j.1541-0420.2008.01051.x⟩. ⟨hal-00153249v4⟩

HAL Id: hal-00153249

https://hal.science/hal-00153249v4

Submitted on 27 Mar 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Biometrics 000, 1–24 DOI: 000
000 0000

Using Empirical Likelihood to combine data: Application to food risk assessment.

Amélie Crépet[1], Hugo Harari-Kermadec[2,3], and Jessica Tressou[1,4,*]

[1]INRA, UR1204, Mét@risk, AgroParisTech, 16 rue Claude Bernard, F75231 Paris, France

[2]INRA, UR1001, CORELA, 65 bd de Brandebourg, F94205 Ivry-sur-Seine, France

[3]CREST-LS, 3 Avenue Pierre Larousse, F92245 Malakoff, France

[4]HKUST-ISMT, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

email: jessica.tressou@agroparistech.fr

Summary: This paper introduces an original methodology based on Empirical Likelihood which aims at combining different food contamination and consumption surveys in order to provide risk managers with a risk measure taking into account all the available information. This risk index is defined as the probability that exposure to a contaminant exceeds a safe dose. It is naturally 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.

Key words: Euclidean Likelihood; Exposure to methyl mercury; Incomplete U-statistics; Risk index; Sea food consumption.

1. Introduction

Certain foods may contain varying amounts of chemicals, which may cause major health problems when accumulating inside the body in excessive doses. This paper focuses on the methyl mercury risk assessment due to the consumption of fish and other seafoods by the French population. Indeed, at high concentrations, this 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 are 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 (Davidson et al., 1995; Grandjean et al., 1997; National Research Council (NRC) of the National Academy of Sciences Price, 2000).

A commonly used measure of 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. For methyl mercury, the latest epidemiological results compiled by the Joint Expert Committee on Food Additives and Contaminants (FAO/WHO, 2003) yields a safe dose called *Provisional Tolerable Weekly Intake* (PTWI) of 1.6μg per week per kg of body weight (μg/w/kgbw in abbreviated form). 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. In France, two main consumption 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. 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, in order to correct for 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 Ridder and Moffitt (2006). It can be also linked to *meta-analysis* techniques mostly used in medical statistics (Edger and Smith, 1997; Hedges and Olkin, 1985). Other methods can be applied to incorporate side information, see Deville and Sarndal (1992); Hellerstein and Imbens (1999); Ireland and Kullback (1968). The methodology chosen in this paper is based on Empirical Likelihood techniques introduced by Owen (1988) as a powerful semi parametric inference method based on a data driven likelihood ratio function. Refer to Owen (2001) 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 Tressou (2005), 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 parametric methods are not 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 Qin (1993); Tsao

and Wu (2006), or Owen (2001) pages 51, 130 and 223-225). In addition, Qin (1994) use a semi-empirical likelihood ratio to construct confidence intervals for the difference of two sample means, one sample being parametrically modelled, and the other one remaining nonparametric. In these works, the problem is to estimate a parameter when two surveys are available for the same inference. Our concern is different: we consider several surveys of different nature, i.e., consumption and contamination surveys, that must be implemented together in order to obtain a risk index. The fact that alternative consumption surveys are available is only a secondary concern. Moreover, application of the classical methods to a concrete applied problem raises intractable difficulties in terms of computation in the context of food risk assessment. Indeed, data set lengths do not add but multiply, and the combination of, say, three 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 2 presents the data, introduces the notation and framework used in food risk assessment problems, and defines the Empirical Likelihood Program (**ELP**), which is difficult to solve due to the high nonlinearity of the parameter of interest. Section 3 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 Bertail and Tressou (2006). The practical computation of this solution in the multidimensional case is treated in Section 4 via incomplete U-statistics. An alternative "Euclidean Likelihood Program" is considered in Section 5, replacing the Kullback-Leibler distance involved in the **ELP** by the Euclidean distance. A validation of these methodologies using simulated datasets is given in Section 6. Finally, Section 7 gives an illustration of these methodologies on true datasets concerning methyl mercury exposure of the French population. The possible generalizations of these methodologies and the specific extensions in the case of food risk assessment are addressed

in Section 8. Technical details are postponed to the appendix, and proofs are available in the Web Appendix, or in Crépet et al. (2007).

2. Data, notation and framework

Our goal is to estimate the food risk index θ_d defined as the probability that exposure to a contaminant exceeds a tolerable dose d, when P products (or groups of products) are potentially contaminated. For this purpose, $R+P$ data sets are available: R (P-dimensional) data sets coming from R consumption surveys and describing the consummed quantities of P products, and P sets of contamination values. We assume that the R consumption surveys concern the same population and therefore that 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. To estimate θ_d in the case of methyl mercury and for the French population, we dispose of $R = 2$ consumption surveys (INCA and SECODIP) described in the next paragraph, and the tolerable dose d is the PTWI of 1.6 μg/w/kgbw. Firstly, contamination of fish and other sea products are considered coming from a single data set and therefore $P = 1$. Secondly, the proper contamination of each type of products ("Fish" on one hand and "Mollusks and shellfish" on the other hand) is considered so that $P = 2$.

2.1 *Data description*

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 (MAAPAR, 2002) and the French Research Institute for Exploitation of the Sea (IFREMER, 1998). These $L = 2832$ analytical data are expressed in terms of total mercury in mg/kg of fresh weight. Considering two groups of products, the data set sizes are $L_1 = 1541$ for "Fish" group and $L_2 = 1291$ for "Mollusks and shellfish" group. 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 (Claisse et al., 2001; Cossa et al., 1989). Adhering to international recommendations (GEMs/Food-WHO, 1995) 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 Bertail and Tressou (2006); Tressou (2006) for further discussions.

The INCA survey. The French "INCA" survey, carried out by CREDOC-AFFSA-DGAL (1999), 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). Since body weight of all individuals is available, "relative" consumptions are computed by dividing the amount consumed during the week by the body weight to obtain exposure expressed in the same unit as the PTWI.

SECODIP. The SECODIP panel for fish, from *TNS SECODIP* (`http://www.secodip.fr`), is composed of 3211 households surveyed over one year (1999). 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 acquisition 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.

Differences between the two consumption surveys.

[Table 1 about here.]

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. This is 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). 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, 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_0 to scale the SECODIP consumption to account for all these facts introducing an additional model constraint in the **ELP**, see Remark 4 in the next section.

2.2 *Notation and framework*

Notation. Let Q denote the random variable of the contamination of the considered product, with distribution $\mathcal{Q}$. Vector $\mathbf{q} = (q_1, \ldots, q_L)'$ is an i.i.d. sample of length L from $\mathcal{Q}$.

In the following, r is the consumption survey number and takes the value 1 for the INCA survey or 2 for the SECODIP panel. $C^{(r)}$ denotes the random variable of the "relative"

consumption with distribution $\mathcal{C}^{(r)}$. Vector $\mathbf{c}^{(r)} = \left(c_1^{(r)}, \ldots, c_{n_r}^{(r)}\right)$ is an i.i.d. sample of length n_r from $\mathcal{C}^{(r)}$.

When a single product is considered ($P = 1$), the probability that the exposure of one individual exceeds a dose d is $\theta_d^{(r)} = \Pr\left(QC^{(r)} > d\right)$, when using the consumption distribution $\mathcal{C}^{(r)}$. Therefore, the common risk index must be solution of the two equations $\theta = \theta_d^{(1)}$ and $\theta = \theta_d^{(2)}$, which can be written as follows:

$$\mathbb{E}_{\mathcal{D}^{(1)}}\left[\mathbb{1}_{\left(QC^{(1)}>d\right)} - \theta\right] = 0, \tag{1}$$

$$\mathbb{E}_{\mathcal{D}^{(2)}}\left[\mathbb{1}_{\left(QC^{(2)}>d\right)} - \theta\right] = 0, \tag{2}$$

where $\mathcal{D}^{(r)} = \mathcal{Q} \times \mathcal{C}^{(r)}$ is the joint distribution of the contamination and the consumption r.

*Empirical Likelihood Program (***ELP***).* The principle of the Empirical Likelihood method is to determine a set of weights for the data of maximal product under constraints given by the model. We define the set $\mathcal{P}$ of positive weights,

- $\mathbf{p}^{(r)} = (p_1^{(r)}, \ldots, p_{n_r}^{(r)})'$ with $\sum_{i=1}^{n_r} p_i^{(r)} = 1$ for the consumption sample r,
- $\mathbf{w} = (w_1, \ldots, w_L)'$ with $\sum_{l=1}^{L} w_l = 1$ for the contamination sample.

The Empirical Likelihood is given by

$$\max_{\left(\mathbf{p}^{(1)},\mathbf{p}^{(2)},\mathbf{w}\right)\in\mathcal{P}} \prod_{i=1}^{n_1} p_i^{(1)} \prod_{j=1}^{n_2} p_j^{(2)} \prod_{l=1}^{L} w_l, \tag{3}$$

under the following model constraints:

$$\mathbb{E}_{\widetilde{\mathcal{D}}^{(1)}}\left[\mathbb{1}_{\left(QC^{(1)}>d\right)} - \theta\right] = 0 \text{ and } \mathbb{E}_{\widetilde{\mathcal{D}}^{(2)}}\left[\mathbb{1}_{\left(QC^{(2)}>d\right)} - \theta\right] = 0, \tag{4}$$

where $\widetilde{\mathcal{D}}^{(r)}$ denotes the weighed joint discrete probability distribution of the contamination sample and the consumption sample r.

These model constraints on θ have an explicit expression, for $r = 1, 2$,

$$\sum_{l=1}^{L} \sum_{i=1}^{n_r} w_l p_i^{(r)} \mathbb{1}_{(q_l c_i^{(r)} > d)} - \theta = 0.$$

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 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 Bertail and Tressou (2006). 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 (Lee, 1990). 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 $\mathbb{E}_{\mathcal{D}}\left[\mathbb{1}_{\left(QC^{(r)}>d\right)}-\theta\right]$. The influence function $\psi^{(r)}$ at point $\left(q,c^{(r)}\right)$ is given by, for $r=1,2$,

$$\psi^{(r)}\left(q,c^{(r)}\right)=\mathbb{E}_{\mathcal{Q}}\left[\mathbb{1}_{(QC^{(r)}>d)}-\theta\middle|\, C^{(r)}=c^{(r)}\right]+\mathbb{E}_{\mathcal{C}^{(r)}}\left[\mathbb{1}_{(QC^{(r)}>d)}-\theta\middle|\, Q=q\right].$$

This functional of $\mathcal{D}^{(r)}$ can be estimated by its empirical counterpart $\widehat{\psi}^{(r)}$ given by

$$\begin{aligned}\widehat{\psi}^{(r)}\left(q,c^{(r)}\right) &= U_0\left(c^{(r)}\right)+U_1^{(r)}(q),\\ \text{where } U_0\left(c^{(r)}\right) &= \frac{1}{L}\sum_{l=1}^{L}\mathbb{1}_{(q_l c^{(r)}>d)}-\theta, \qquad (5)\\ \text{and}\quad U_1^{(r)}(q) &= \frac{1}{n_r}\sum_{i=1}^{n_r}\mathbb{1}_{(qc_i^{(r)}>d)}-\theta,\ \text{for } r=1,2. \qquad (6)\end{aligned}$$

$U_0\left(c^{(r)}\right)$ and the $U_1^{(r)}(q)$ are generalized U-statistics with kernel $\mathbb{1}_{(qc^{(r)}>d)}$ and degree 1, see Lee (1990). For simplicity, the dependence in n_r and L is implicit in the notation.

An approximate version of the model constraints (4) can now be written, for $r=1,2$,

$$\mathbb{E}_{\widetilde{\mathcal{D}}^{(r)}}\left[\widehat{\psi}^{(r)}\left(Q,C^{(r)}\right)\right]=0,\ \text{i.e.,}\ \sum_{i=1}^{n_r}p_i^{(r)}U_0\left(c_i^{(r)}\right)+\sum_{l=1}^{L}w_l U_1^{(r)}(q_l)=0.$$

The following theorem establishes the asymptotic convergence of the approximate version of the Empirical Likelihood when only one product is considered.

THEOREM 1: *Under hypotheses H1-H4 given in the appendix, the Empirical Likelihood*

Program involves solving the dual program

$$l_{n_1,n_2,L}(\theta) = \sup_{\substack{\lambda_1,\lambda_2,\gamma_1,\gamma_2,\gamma_a\in\mathbb{R} \\ n_1+n_2+L-\gamma_1-\gamma_2-\gamma_a=0}} \left[\begin{array}{c} \sum_{i=1}^{n_1} \ln\left\{\gamma_1 + \lambda_1 U_0\left(c_i^{(1)}\right)\right\} \\ + \sum_{j=1}^{n_2} \ln\left\{\gamma_2 + \lambda_2 U_0\left(c_j^{(2)}\right)\right\} \\ + \sum_{l=1}^{L} \ln\left\{\gamma_a + \lambda_1 U_1^{(1)}(q_l) + \lambda_2 U_1^{(2)}(q_l)\right\} \end{array} \right]. \quad (7)$$

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 at the true value θ_d of the parameter is asymptotically $\chi^2(1)$:*

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

The proof of these results is given in the Web Appendix B. In particular, it is demonstrated that the constraint $n_1 + n_2 + L - \gamma_1 - \gamma_2 - \gamma_a = 0$ naturally arises from the combination of the first order conditions related to the model constraints in the **ELP**.

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: The result remains true in the general case, that is for $P > 1$ or $R > 2$, with a log-likelihood ratio converging to $(P+1)^2\chi^2(1)$, see Crépet et al. (2007). However, some refinements are needed in practice to make it tractable, as detailed in the next section. Note that solving the **ELP** in the case $R = 1, P \geq 1$ is equivalent to building bootstrap confidence intervals as in Bertail and Tressou (2006).

Remark 2: 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 of Owen (2001).

Remark 3: These model constraints can be augmented by some calibration constraints

in order to correct for known bias of the surveys. These additional constraints 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 constraints of the form

$$\sum_{i=1}^{n_1} p_i^{(1)} Z_i^{(1)} = z_0, \qquad \sum_{j=1}^{n_2} p_j^{(2)} Z_j^{(2)} = z_0, \tag{8}$$

where $Z_i^{(1)}$ and $Z_j^{(2)}$ are vectors describing the belonging to specified sociodemographic categories in surveys 1 and 2, while 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 and Lawless (1994) and Owen (2001), Chapter 3, page 51. For example, the proportion of children (34%) in INCA survey is high compared to the national census (15%, INSEE, Institut National de la Statistique et des Etudes Economiques, 1999): 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 (8). The additional constraint is therefore $\sum_{i=1}^{n_1} p_i^{(1)} \mathbb{1}_{(3 \leq Z_i^{(1)} \leq 14)} = 0.15$, 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 (7) in the part concerning the first survey. It becomes

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

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

REMARK 4: The additional model constraint related to the scaling of the SECODIP data mentioned in Section 2.1 is $\sum_{i=1}^{n_1} p_i^{(1)} c_i^{(1)} = \alpha_0 \sum_{j=1}^{n_2} p_j^{(2)} c_j^{(2)}$, where α_0 is the scaling coefficient. It 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., $r_{n_1,n_2,L_1}(\theta_d,\alpha_0) \to (P+1)^2\chi^2(2)$. We then optimize on α_0 for each θ to get a profiled likelihood on θ.

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

In this section, we turn to the case of several potentially contaminated products. For the simplicity of the statement, we only consider the case of $P = 2$ products (or groups of products), with contamination samples $\mathbf{q}^{[a]} = \left(q_1^{[a]},\dots,q_{L_a}^{[a]}\right)'$ and $\mathbf{q}^{[b]} = \left(q_1^{[b]},\dots,q_{L_b}^{[b]}\right)'$. The consumption surveys are now 2-dimensional and the samples are denoted by $\mathbf{c}^{(r)} = \left(c_1^{(r)},\dots,c_{n_r}^{(r)}\right)'$, with $c_i^{(r)} = \left(c_{i,a}^{(r)}, c_{i,b}^{(r)}\right)$. Moreover, the set of weights $\mathcal{P}$ is now extended such that $\mathbf{w}$ is replaced with $\mathbf{w}^{[k]} = (w_1^{[k]},\dots,w_{L_k}^{[k]})'$ for $k = a, b$.

When one considers more than one potentially contaminated product, the computation of the different U-statistics defined in (5) and (6) becomes too heavy if the data sets are large. Indeed, one needs to compute at least $n_r L_a L_b$ 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) and (6). 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. For $r = 1$ or 2, the incomplete version of equation (5) for two products is given by

$$U_{\mathcal{B}_0}\left(c^{(r)}\right) = \frac{1}{B}\sum_{(l_a,l_b)\in\mathcal{B}_0} \mathbb{1}_{\left(q_{l_a}^{[a]}c_a^{(r)} + q_{l_b}^{[b]}c_b^{(r)} > d\right)} - \theta, \tag{9}$$

where the sum is taken over the set $\mathcal{B}_0$ of indexes (l_a, l_b), randomly chosen with replacement from $\{1,\dots,L_a\}\otimes\{1,\dots,L_b\}$, with size B.

The incomplete version of (6) for two products writes, when the first product contamination is fixed:

$$U_{\mathcal{B}_{-a}}^{(r)}(q^{[a]}) = \frac{1}{B}\sum_{(l_b,i)\in\mathcal{B}_a} \mathbb{1}_{\left(q^{[a]}c_{i,a}^{(r)} + q_{l_b}^{[b]}c_{i,b}^{(r)} > d\right)} - \theta, \tag{10}$$

where the sum is taken over the set $\mathcal{B}_{-a}$ of indexes (l_b, i) randomly chosen with replacement from $\{1, \ldots, L_b\} \otimes \{1, \ldots, n_r\}$, with size B. When one fixes the second product contamination, an equivalent formula is obtained by exchanging indexes a and b.

The approximate influence function is now given by

$$\widehat{\psi}_B^{(r)}\left(q^{[a]}, q^{[b]}, c^{(r)}\right) = U_{\mathcal{B}_0}\left(c^{(r)}\right) + U_{\mathcal{B}_{-a}}^{(r)}(q^{[a]}) + U_{\mathcal{B}_{-b}}^{(r)}(q^{[b]}).$$

The model constraints can then be written as follows:

$$\begin{aligned}
&\sum_{i=1}^{n_1} p_i^{(1)} U_{\mathcal{B}_0}\left(c_i^{(1)}\right) + \sum_{l_a=1}^{L_a} w_{l_a}^{[a]} U_{\mathcal{B}_{-a}}^{(1)}\left(q_{l_a}^{[a]}\right) + \sum_{l_b=1}^{L_b} w_{l_b}^{[b]} U_{\mathcal{B}_{-b}}^{(1)}\left(q_{l_b}^{[b]}\right) = 0, \qquad (11)\\
&\sum_{j=1}^{n_2} p_j^{(2)} U_{\mathcal{B}_0}\left(c_j^{(2)}\right) + \sum_{l_a=1}^{L_a} w_{l_a}^{[a]} U_{\mathcal{B}_{-a}}^{(2)}\left(q_{l_a}^{[a]}\right) + \sum_{l_b=1}^{L_b} w_{l_b}^{[b]} U_{\mathcal{B}_{-b}}^{(2)}\left(q_{l_b}^{[b]}\right) = 0.
\end{aligned}$$

The chosen size B must fulfill an additional assumption, H5 given in appendix, in order to ensure that the incomplete U-statistics are close enough to the complete ones, i.e., that their difference is of order $o(B^{-1/2})$ Lee (1990).

COROLLARY 1: *Under the assumptions H1-H5 given in the appendix, the likelihood ratio for two products, at the true value θ_d of the parameter, $r_{n_1,n_2,L_a,L_b}(\theta_d)$, is asymptotically $\chi^2(1)$:*

$$r_{n_1,n_2,L_a,L_b}(\theta_d) \xrightarrow{\mathcal{L}} 9\chi^2(1).$$

See the Web Appendix C for the proof. Note in particular that B, the size of the incomplete U-statistics, must go to infinity quicker than $\max\{n_1, n_2, L_a, L_b\}$. This result yields an $(1-\alpha)^{th}$ confidence interval for θ_d given, in the case $P = 2$, by

$$\left\{\theta : r_{n_1,n_2,L_a,L_b}(\theta) \leq 9\chi^2_{1-\alpha}(1)\right\}.$$

As before, these results remain true in the general case, that is for $P > 2$ (see Crépet et al., 2007, for details).

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 distribution on the sample described by the $\mathbf{p}^{(r)}$ and $\mathbf{w}^{[k]}$ (or equivalently by $\widetilde{\mathcal{D}}^{(1)} \times \widetilde{\mathcal{D}}^{(2)}$), and the observed data, i.e., the discrete distribution giving uniform weights to each observation ($1/n_r$ for consumption data, $r = 1, 2$ and $1/L_k$ for contamination data, $k = a, b$). Following the ideas of Owen (2001), we replace the Kullback-Leibler distance by the Euclidean distance, also called the χ^2 distance (recall that for two probabilities P and Q, with Q dominated by P, the Kullback-Leibler distance between P and Q is $\int \log(\frac{dQ}{dP} - 1)dP$, while the Euclidean distance is $\int (\frac{dQ}{dP} - 1)^2 dP$). When using the Euclidean distance in the case $P = 2$, the program is given by

$$\mathbf{l}_{n_1,n_2,L_a,L_b}(\theta) = \min_{\left(\mathbf{p}^{(1)},\mathbf{p}^{(2)},\mathbf{w}^{[a]},\mathbf{w}^{[b]}\right)\in\mathcal{P}} \frac{1}{2} \left\{ \begin{array}{c} \sum_{i=1}^{n_1} \left(n_1 p_i^{(1)} - 1\right)^2 + \sum_{j=1}^{n_2} \left(n_2 p_j^{(2)} - 1\right)^2 \\ + \sum_{l_a=1}^{L_a} \left(L_a w_{l_a}^{[a]} - 1\right)^2 + \sum_{l_b=1}^{L_b} \left(L_b w_{l_b}^{[b]} - 1\right)^2 \end{array} \right\}, \tag{12}$$

under the approximated model constraints (11) and the constraint that each set of weights sums to 1. We get a result equivalent to Corollary 1:

COROLLARY 2: *Under the assumptions H1-H5 given in the appendix, the Euclidean Likelihood ratio is asymptotically* $9\chi^2(1)$*:*

$$\mathbf{r}_{n_1,n_2,L_a,L_b}(\theta_d) = 2\left\{\mathbf{l}_{n_1,n_2,L_a,L_b}(\theta_d) - \inf_{\theta} \mathbf{l}_{n_1,n_2,L_a,L_b}(\theta)\right\} \xrightarrow{\mathcal{L}} 9\chi^2(1).$$

The proof of this result is given in Crépet et al. (2007).

The choice of the Euclidean distance is closely related to the Generalized Method of Moments (GMM), see Newey and Smith (2004); Bonnal and Renault (2004) for precisions on the links between Empirical Likelihood and GMM. Instead of logarithms, the optimization program (12) only involves quadratic terms and is then much easier to solve. 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 $p_i^{(1)}$, $p_j^{(2)}$, $w_{l_a}^{[a]}$ and $w_{l_b}^{[b]}$ are not forced to be positive. However, these weights are asymptotically nonnegative with probability one, see Bonnal and Renault (2004).

The gain in computation time is counter-balanced by a loss in adaptability to the data and to the constraints. 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 at the final stage, to get precise confidence regions and estimators. This is illustrated in the next section.

6. Validation on simulated data

In order to validate the proposed methodology, coverage probabilities of the 95% confidence interval resulting from Corollary 2 are assessed by simulation of known contamination and consumption distributions as in Bertail and Tressou (2006) and Tressou (2006). 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 as illustrated in the food risk application, 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 (their parameters were chosen to fit as much as possible the INCA dataset and the available contamination data). 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 defined 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 (the proportion of the food eaten at home is distributed according to a Beta distribution with mean 0.8 and variance $0.8(1 - 0.8)$) with mean 20% to account for consumption outside the home. Note that the only features that are not reproduced are the high proportion of children in sample 1 and the aggregation/disaggregation of consumptions within households. Then [Step 2] is repeated $S = 200$ times.

Results: We obtain a coverage probability of 95.5%. This validates the methodology and is comforted by the known robustness of Euclidean Likelihood, i.e., its coverage probability converges to the confidence level from above.

7. Methyl mercury risk assessment

7.1 *Results when considering one single food group*

We first merge all the seafoods into a single group. Consequently, the consumption data are contained in vectors $c^{(1)} = (c_1^{(1)}, \ldots, c_{3003}^{(1)})'$ and $c^{(2)} = (c_1^{(2)}, \ldots, c_{9588}^{(2)})'$ for the INCA and SECODIP surveys, respectively, where $c_i^{(r)}$ gives the total weekly consumption of fish and seafood for individual i in survey r. The contamination data are given by vector $q = (q_1, \ldots, q_{2832})'$, where q_j indicates the total mercury in mg/kg of fresh weight of the analysed

seafood item j. All contamination data are attributed to the total individual consumption of seafoods. Calculations can therefore be performed using the complete linearized U-statistics of degree $(1, 1)$.

[Figure 1 about here.]

Figure 1(a) shows the two 95% confidence regions for the couple of parameters $(\theta_{1.6}, \alpha_0)$. We compare the results obtained with and without the constraint on the proportion of children. The unconstrained confidence region for $(\theta_{1.6}, \alpha_0)$ is marked by a dashed line, the solid line corresponding to the constrained confidence region. We can see that the constraint makes the 2 surveys closer (α_0 is closer to 1, 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 to the INCA survey, $\mathbf{p}^{(1)}$, is modified so that children become less influent, which explains the risk reduction and the decrease of α_0.

Figure 1(b) shows the profiles of the Empirical Likelihood ratios ($r_{n_1,n_2,L_1}(\theta_{1.6})$). We get two profiles, the dashed line corresponds to the unconstrained case. The horizontal line gives the 95% level of the χ^2 distribution ($4\chi^2_{95\%}(1)$), 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 $\alpha^*_0 = 1.31$. 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 α_0-scaled SECODIP data (marked by the vertical dashed black line). Indeed, the best correction α_0 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 unconstrained Empirical likelihood.

[Figure 2 about here.]

7.2 *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 $L_2 = 1291$. Consequently, the contamination data are given by vectors $q^{[a]} = (q_1^{[a]}, \ldots, q_{1541}^{[a]})'$ and $q^{[b]} = (q_1^{[b]}, \ldots, q_{1291}^{[b]})'$ for "Fish" and "Mollusk and shellfish", respectively. The consumption data are given by vectors $c^{(1)} = (c_1^{(1)}, \ldots, c_{3003}^{(1)})'$ and $c^{(2)} = (c_1^{(2)}, \ldots, c_{9588}^{(2)})'$ for the INCA and SECODIP surveys, respectively, where now $c_i^{(r)} = (c_{i,a}^{(r)}, c_{i,b}^{(r)})$, with $c_{i,a}^{(r)}$ and $c_{i,b}^{(r)}$ providing the total weekly consumption of "Fish" and "Mollusk and shellfish", respectively, for individual i in survey r. Calculations are done by using incomplete U-statistics defined in equations (9) and (10) with a size $B = 10000$. Note that α_0 is here 2-dimensional.

The constrained 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 $\alpha_0^* = (1.8, 1.65)$. Figure 3 shows the profiles of the Empirical and Euclidean Likelihood ratios, both

with and 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 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. Unfortunately, this is not possible with the available SECODIP food nomenclature.

[Figure 3 about here.]

8. 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 the empirical likelihood ratio is asymptotically χ^2. 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 calibration constraints related to side information. When the Empirical Likelihood problems get complicated, Euclidean Likelihood can be helpfully used, at least to find first step estimators. A technical improvement of the present food risk assessment 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 a regression based method for the decomposition of household nutritional intakes into individual intakes accounting for outside consumptions, see also Allais

and Tressou (2007). 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_0 between SECODIP and INCA panels. We plan to explore this issue in future works.

From an applied point of view, we obtain 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.

Supplementary Materials

Web Appendices referenced in Sections 3 and 4 are available under the Paper Information link at the Biometrics website `http://www.biometrics.tibs.org`.

Acknowledgements

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.

References

Allais, O. and Tressou, J. (2007). Using decomposed household food acquisitions as inputs of a Kinetic Dietary Exposure ModelAvailable at `https://hal.archives-ouvertes.fr/hal-00139914`.

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

Bertail, P. and Tressou, J. (2006). Incomplete generalized U-statistics for food risk assessment. *Biometrics* **62,** 66–74.

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

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,**.

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

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.

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. Technical report, IFREMER, Nantes.

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

Crépet, A., Harari, H., and Tressou, J. (2007). Using empirical likelihood to combine data: Application to food risk assessment. Document de travail CREST. Available at `http://hal.archives-ouvertes.fr/hal-00153249`.

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.

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

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

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.

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

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.

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

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

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.

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

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. Technical report.

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

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.

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

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.

National Research Council (NRC) of the National Academy of Sciences Price (2000). Toxicological effects of methyl mercury. Technical report, National Academy Press, Washington, DC.

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

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

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

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,** 339–348.

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

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

Qin, J. S. (1994). Semi-empirical likelihood ratio confidence intervals ratio for the difference of 2 sample means. *Annals of the Institute of Statistical Mathematics* **46,** 117–126.

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

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.

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

Tressou, J. (2006). Non parametric modelling of the left censorship of analytical data in food risk exposure assessment. *Journal of the American Statistical Association* **101,** 1377–1386.

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,** 1349–1358.

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

WHO (1990). Methylmercury, Environmental Health Criteria 101. Technical report, Geneva, Switzerland.

Submitted October 2007.

Appendix Technical details

This appendix gives the technicals assumptions required in Theorem 1, and Corollaries 1 and 2.

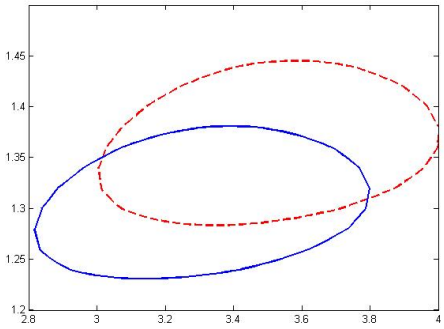
H1: The contamination sample $\mathbf{q}$ is i.i.d. and $U_0(C)$ has finite variance.

H2: The independent consumption samples $\mathbf{c}^{(1)}$ and $\mathbf{c}^{(2)}$ are i.i.d. and $\left(U_1^{(1)}(Q),\ U_1^{(2)}(Q)\right)'$ has finite invertible covariance matrix.

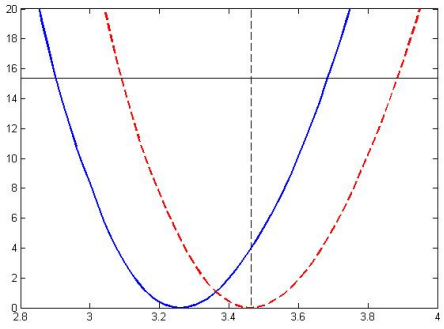
H3: The parameter θ verifies equations 1 and 2.

H4: The sample lengths n_1, n_2 and L go to infinity and their ratios are bounded.

H5: The incomplete statistic size B is such that $n_1 + n_2 + L_a + L_b = o(B)$.

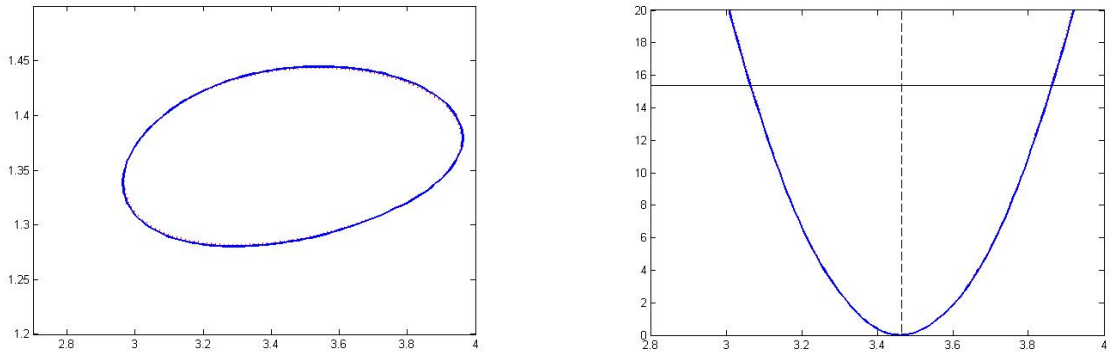


(a) Empirical Likelihood confidence regions
horizontal axis is $\theta_{1.6}$,
vertical axis is α_0



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

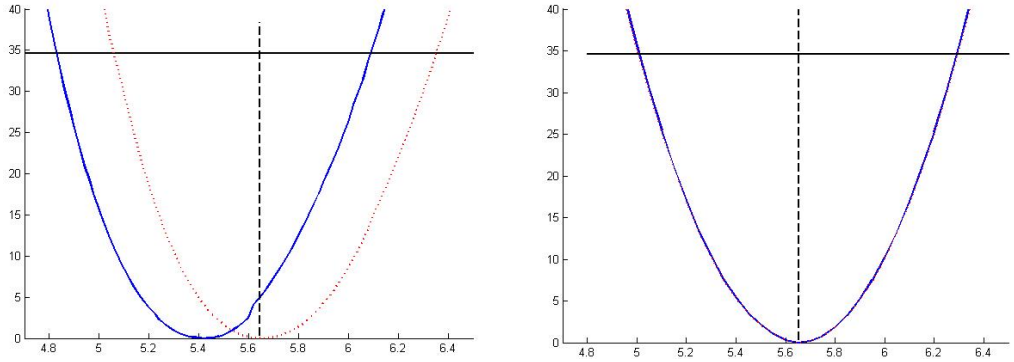
Figure 1. Empirical Likelihood for one product (solid with age constraint, dash without)



(a) Euclidean Likelihood confidence regions horizontal axis is $\theta_{1.6}$, vertical axis is α_0

(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, dash without)



(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

Table 1
Previous estimates of the methyl mercury food risk index $\theta_{1.6}$ (expressed in %, with basic 95% confidence intervals)

	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]