Maximum Likelihood Under Incomplete Information: Toward a Comparison of Criteria

Inés Couso, Didier Dubois

Maximum de vraisemblance et information incomplète

Maximum likelihood under incomplete information

Inés Couso[1] Didier Dubois[2]

[1] Dep. Statistics, Universidad de Oviedo (España)
[2] IRIT Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 09, France

couso@uniovi.es dubois@irit.fr

Résumé :

La méthode du maximum de vraisemblance est bien connue et constitue l'un des piliers de l'inférence statistique. Elle suppose des données précises. Mais quand les données sont imprécises ou incomplètes, il n'est pas trivial de définir une fonction de vraisemblance. Cela dépend du problème qu'on veut traiter. Cette article discute plusieurs définitions possibles et les difficultés qu'elles peuvent engendrer.

Mots-clés :

Maximum de vraisemblance, données imprécises, fonctions de croyance

Abstract:

Maximum likelihood is a standard approach to computing a probability distribution that best fits a given dataset. However, when datasets are incomplete or contain imprecise data, depending on the purpose, a major issue is to properly define the likelihood function to be maximized. This paper compares several proposals in terms of their intuitive appeal, showing their anomalous behavior on examples.

Keywords:

Likelihood maximization, incomplete data, belief functions

1 Introduction

Edwards ([6], p. 9) rappelle qu'une fonction de vraisemblance est proportionnelle à la probabilité d'obtention des données observées, calculée selon un modèle probabiliste. L'axiome fondamental stipule que la probabilité d'observer l'un parmi deux résultats de tests statistiques est la somme des probabilités d'obtention de chacun de ces résultats. En particulier, le résultat d'une observation au sens d'Edwards est entendu comme un événement élémentaire (une réalisation), et non un événement au sens large. Seuls les événements élémentaires peuvent être observés. Par exemple, si on lance un dé, le résultat obtenu sera par exemple 1, 3 ou 5, et non "impair". Si on accepte ce point de vue, comment définir la fonction de vraisemblance quand les données sont imprécisément observées, ou incomplètes ? Pour répondre à cette question, il faut d'abord comprendre ce qu'est le résultat d'une épreuve statistique dans ce contexte. En particulier, si on s'intéresse à modéliser un certain phénomène aléatoire, les observations qu'on obtient alors ne nous informent pas directement sur la variable aléatoire représentant ce phénomène. A cause de l'interférence avec un processus d'observation imparfait, les observations seront ensemblistes. Pour tirer le meilleur parti des observations imprécises, il faut clarifier ce qu'on souhaite modéliser :

1. le phénomène aléatoire *au travers* du processus d'observation ;
2. ou le phénomène aléatoire *malgré* le processus d'observation imparfait.

Dans le premier cas, les observations imprécises sont considérees comme résultats d'épreuves statistiques, et on va construire alors la fonction de vraisemblance d'un ensemble aléatoire, dont les réalisations contiennent les réalisations précises mais mal connues de la variable aléatoire sous-jacente que nous voudrions modéliser. En fait, la plupart des auteurs envisagent l'autre point de vue, et considèrent que les vrais résultats sont les réalisations de cette variable aléatoire. Mais alors, on peut définir autant de fonctions de vraisemblance

qu'il y a de résultats précis conformes aux observations imprécises. Plusieurs approches pour contourner ce problème on été proposées dans la littérature. La plus répandue utilise l'algorithme EM, qui revient à construire un échantillon fictif, conforme aux observations imprécises, de la variable aléatoire mal observée, et à maximiser la vraisemblance par rapport à cet échantillon fictif. Dans cet article, on analyse cette approche en la replaçant dans le contexte de l'approche épistémique du raisonnement statistique décrite dans [1] et en la comparant avec d'autres propositions plus récentes de Denoeux [5], Hüllermeier [8], and Guillaume [7]. On ne considère pas ici le problème de l'imprécision due à un nombre trop faible d'observations précises (voir par exemple Serrurier and Prade [10]).

2 La variable aléatoire et sa mesure

Soit un ensemble d'éventualités Ω et une variable aléatoire $X : \Omega \to \mathcal{X}$ qui représente un certain phénomène répétable. Par simplicité, supposons que $\mathcal{X} = \{a_1, \ldots, a_m\}$ est fini. De plus, il y a un outil de mesure, qui suit une loi associée à une autre variable aléatoire Y à valeurs dans $\mathcal{Y}$ qui fournit des informations imprécises sur les réalisations de X. A savoir, chaque réalisation $Y(\omega)$ correspond à l'information $\Gamma(\omega) \in \wp(\mathcal{X})$, où $\Gamma : \Omega \to \wp(\mathcal{X})$, est une fonction multivoque qui représente la mesure (imprécise) de X. Donc, on suppose que X est une sélection de Γ, soit $X(\omega) \in \Gamma(\omega),\ \forall \omega \in \Omega$ [3]. Si $Im(\Gamma) = \{A_1, \ldots, A_r\}$ est l'ensemble image de Γ (la collection des résultats de la mesure) et $\mathcal{Y} = \{b_1, \ldots, b_r\}$, il y a une bijection entre $Im(\Gamma)$ et $\mathcal{Y}$ telle que :

$$Y(\omega) = b_j \text{ ssi } \Gamma(\omega) = A_j,\ j = 1, \ldots, r.$$

C-à-d, on peut identifier b_j au singleton $\{A_j\}$ dans l'ensemble $\wp(\mathcal{X})$ des parties de $\mathcal{X}$. On peut envisager deux façons de représenter l'information sur la distribution jointe du vecteur aléatoire (X, Y).

Le point de vue de la génération de l'imprécision. Dans ce cas, on insiste sur les réalisations de X et le processus d'"imprécisiation" qui fournit des observations imparfaites de X, à l'aide de la matrice $(M|\mathbf{p})$:

- $p_{.j|k.} = P(Y = b_j|X = a_k)$ est la probabilité d'observer A_j si le vrai résultat est a_k et
- $p_{k.} = P(X = a_k)$ est la probabilité que le vrai résultat est a_k.

Cette matrice determine la probabilité jointe qui modélise le phénomène aléatoire sous-jacent et lien entre ses réalisations et les observations incomplètes. Voici quelques exemples et leurs matrices caractéristiques :

- **Partition [4].** Supposons que $Im(\Gamma) = \{A_1, \ldots, A_r\}$ forme une partition de $\mathcal{X}$. Donc, on peut facilement calculer les probabilités $P(Y = b_j|X = a_k) = 1$ si $a_k \in A_j$ et 0 sinon, pour tous j, k.
- **Hypothèse du sur-ensemble [9].** $Im(\Gamma)$ coïncide avec $\wp(\mathcal{X}) \setminus \{\emptyset\}$. Pour tout $k = 1, \ldots, m$ il y a une constante c_k telle que $P(Y = b_j|X = a_k) = c_k$, si $A_j \ni a_k$ et 0 sinon. De plus pour chaque $k \in \{1, \ldots, m\}$ il y a 2^{m-1} sous-ensembles de $\mathcal{X}$ qui le contiennent. Donc la constante vaut $1/2^{m-1}$, soit : $P(Y = b_j|X = a_k) = \begin{cases} 1/2^{m-1} & \text{si } A_j \ni a_k \\ 0 & \text{sinon.} \end{cases}$ C'est une sorte d'hypothèse d'absence au hasard, où le processus d'imprécisiation est totalement aléatoire. On la présente souvent comme capturant l'idée d'absence d'information sur le processus de mesure, ce qui est contestable.

Le point de vue de la désambiguation On peut aussi caractériser la probabilité jointe de (X, Y) à l'aide de la distribution marginale de Y (la masse de Möbius $m(A_j) = P(y = b_j) = p_{.j},\ j = 1, \ldots, r$ d'une fonction de croyance) qui décrit les observations imprécises [3] et la probabilité chaque résultat $X = a_k$, conditionnelle à l'observation $\Gamma(\omega) = A_j$, pour tout $j = 1, \ldots, r$. La nouvelle matrice $(M'|\mathbf{p}')$ a pour coefficients :

- $b_{kj} = p_{k.|.j} = P(X = a_k|Y = b_j)$ est la

probabilité que la vraie valeur de X est a_k si on sait qu'elle est dans A_j ;

— $p_{.j} = P(\Gamma = A_j) = P(Y = b_j)$ est la probabilité que le processus de mesure fournit A_j.

Cette matrice détermine aussi la probabilité jointe de la variable aléatoire sous-jacente et le lien entre ses réalisations et les observations incomplètes. A savoir, le vecteur $(p_{.1}, \ldots, p_{.r})^T$ caractérise le processus de mesure et la matrice $M' = (p_{k.|.j})_{k=1,\ldots,m;j=1,\ldots,r}$ représente la probabilité conditionnelle de X sachant Y (observation). Voici un exemple :

— **Probabilité conditionnelle uniforme** Sous cette hypothèse, la probabilité (marginale) P_X induite par X est la transformée pignistique [12] de la fonction de croyance associée à la masse de Möbius m. La distribution conditionnelle s'écrit : $p_{k.|.j} = \frac{1}{\#A_j}$, if $a_k \in A_j$ et 0 sinon. Et la distribution marginale est : $p_{k.} = \sum_{j:A_j \ni a_k} \frac{1}{\#A_j} p_{.j}$.

3 Différentes fonctions de vraisemblance

Chaque matrice $(M|\mathbf{p})$ ou $(M'|\mathbf{p}')$ caractérise la distribution jointe de (X, Y). Pour chaque paire $(k, j) \in \{1, \ldots, m\} \times \{1, \ldots, r\}$, soit p_{kj} la probabilité jointe $p_{kj} = P(X = a_k, Y = b_j)$. Selon les notations utilisées ci-dessus, les marginales sur $\mathcal{X}$ et $\mathcal{Y}$ s'écrivent :

— $p_{.j} = \sum_{k=1}^{m} p_{kj}$ qui est la probabilité de $Y = b_j$, pour chaque $j = 1, \ldots, r$, et

— $p_{k.} = P(X = a_k) = \sum_{j=1}^{r} p_{kj}$ est la probabilité de $X = a_k$, pour tout k.

Supposons maintenant que cette distribution jointe est paramétrique, avec un (vecteur de) paramètre(s) θ à valeurs dans un ensemble Θ (au sens où $(M|\mathbf{p})$ et $(M'|\mathbf{p}')$ sont des fonctions de θ) qui la détermine complètement. Clairement, le nombre de composants of θ est au plus égal à la dimension des matrices, soit $\min\{m(r+1), r(m+1)\}$. Soit une suite de N copies iid de $Z = (X, Y)$, $\mathbf{Z} = ((X_1, Y_1), \ldots, (X_N, Y_N))$. On note $\mathbf{z} = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ un échantillon du vecteur (X, Y). Donc, $\mathbf{y} = (y_1, \ldots, y_N)$ est l'échantillon observé (une réalisation du vecteur $\mathbf{Y} = (Y_1, \ldots, Y_n)$), et $\mathbf{x} = (x_1, \ldots, x_N)$ représente un échantillon artificiel arbitraire de $\mathcal{X}$ pour la variable X latente non-observée, qui parcourt $\mathcal{X}^N$. Les échantillons $\mathbf{x}$ soit choisis tels que le nombre de répétitions n^{kj} de chaque paire $(a_k, b_j) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ dans l'échantillon est en accord avec les nombres $n_{.j}$ d'observations A_j. On note $\mathcal{X}^{\mathbf{y}}$ (resp. $\mathcal{Z}^{\mathbf{y}}$) l'ensemble des échantillons $\mathbf{x}$ (resp. échantillons joints complets $\mathbf{z}$) qui respectent cette condition. On peut considérer trois fonctions de log-vraisemblance différentes, selon qu'on se réfère à

l'échantillon observé : $L^{\mathbf{y}}(\theta) = \log \mathbf{p}(\mathbf{y}; \theta) = \log \prod_{i=1}^{N} p(y_i; \theta)$. En notant $p^{\theta}_{.j} = p(b_j; \theta)$, elle s'écrit aussi $\sum_{j=1}^{r} n_{.j} \log p^{\theta}_{.j}$. où $n_{.j}$ est le nombre de répétitions of A_j dans l'échantillon de taille N ;

l'échantillon caché : $L^{\mathbf{x}}(\theta) = \log \mathbf{p}(\mathbf{x}, \theta)$. Elle s'écrit aussi $\log \prod_{i=1}^{N} p(x_i; \theta) = \sum_{k=1}^{m} n_{k.} \log p^{\theta}_{k.}$, où $n_{k.}$ est le nombre d'occurrences de a_k dans l'échantillon fictif $\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}^{\mathbf{y}}$.

l'échantillon complet : $L^{\mathbf{z}}(\theta) = \log \mathbf{p}(\mathbf{z}, \theta) = \log \prod_{i=1}^{N} p(z_i; \theta)$. Elle s'écrit aussi $\sum_{k=1}^{m} \sum_{j=1}^{r} n_{kj} \log p^{\theta}_{kj}$ où n_{kj} est le nombre de répétitions de la paire (a_k, b_j) dans l'échantillon $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$.

Dans la suite, on compare quelques stratégies de maximisation de la vraisemblance, en présence d'informations imprécises $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$:

1. *L'estimation de maximum de vraisemblance (EMV) standard* : on calcule l'argument du maximum de $L^{\mathbf{y}}$ considéré comme une fonction de domaine Θ, soit : $\hat{\theta} = \arg\max_{\theta \in \Theta} L^{\mathbf{y}}(\theta) = \arg\max_{\theta \in \Theta} \prod_{j=1}^{r} (p^{\theta}_{.j})^{n_{.j}}$. Le résultat est une distribution sur $\wp(\mathcal{X})$ (masse de Möbius) pourvu qu'il n'y ait pas de connaissance supplémentaire reliant les distributions sur X et sur Y. L'algorithme EM [4] est une technique

itérative EMV basée sur le recours à une variable latente X, qui calcule un échantillon fictif pour X afin d'atteindre un maximum local de $L^{\mathbf{y}}$.

2. *La stratégie maximax* [8] : elle vise à trouver $(\mathbf{x}^*, \theta^*) \in \mathcal{X}^{\mathbf{y}} \times \Omega$ qui maximise $L^{\mathbf{z}}(\theta)$ ou $L^{\mathbf{x}}(\theta)$, par exemple, dans le premier cas : $(\mathbf{x}^*, \theta^*) = \arg\max_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}},\theta\in\Theta} L^{\mathbf{z}}(\theta)$, soit $\arg\max_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}},\theta\in\Theta} \prod_{k=1}^{m}\prod_{j=1}^{r}(p_{kj}^{\theta})^{n_{kj}}$. Le vecteur $\mathbf{x}^*$ est un échantillon fictif.

3. *La stratégie maximin* [7] : trouver $\theta_* \in \Theta$ qui maximise $L_-^{\mathbf{z}}(\theta) = \min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} L^{\mathbf{z}}(\theta) = \min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} \sum_{k=1}^{m}\sum_{j=1}^{r} n_{kj}\log p_{kj}^{\theta}$. Ou alternativement, on maximise $L_-^{\mathbf{x}}(\theta) = \min_{\mathbf{x}\in\mathcal{X}^{\mathbf{y}}} L^{\mathbf{x}}(\theta)$ C'est une approche robuste qui présuppose aussi un échantillon fictif $\mathbf{x}_*$.

4 Comparaison entre les fonctions de vraisemblance

Sous certaines conditions sur les matrices M and M', certaines des fonctions de vraisemblance ci-dessus peuvent coïncider. Nous donnons quelques exemples de résultats en ce sens. Une première question concerne le paramètre θ, qui définit la fonction de vraisemblance de la paire de variables (X, Y), donc celles de X et Y. Parfois, X, et Y dépendent de paramètres distincts θ_1, θ_2.

Définition 1. *On dira que le paramètre $\theta \in \Theta$ est* séparable *pour la matrice $(M|\mathbf{p})$ s'il peut être "séparé" en deux composantes (peut-être multidimensionelles) $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$ avec $\Theta = \Theta_1 \times \Theta_2$ de sorte que $p_{.j|k.}^{\theta}$ et $p_{j.}^{\theta}$ sont des fonctions de θ_1 et θ_2, respectivement.*

Proposition 1. *Si θ est séparable pour $(M|\mathbf{p})$ alors, étant donné un échantillon $\mathbf{x} \in \mathcal{X}^N$ et l'échantillon correspondant $\mathbf{z} \in (\mathcal{X}\times\mathcal{Y})^N$ induit par $\mathbf{x}$ and $\mathbf{y}$, on a $\arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta) \subseteq \arg\max_{\theta\in\Theta} L^{\mathbf{x}}(\theta)$.*

Définition 2. *On dira que le paramètre $\theta \in \Theta$ est* séparable *par rapport à la matrice $(M'|\mathbf{p}')$s'il peut être "séparé" en deux composantes (peut-être multidimensionelles) $\theta_3 \in \Theta_3$, $\theta_4 \in \Theta_4$ telles que $\Theta = \Theta_1 \times \Theta_2$ de sorte que $p_{k.|.j}^{\theta}$ et $p_{.j}^{\theta}$ sont des fonctions de θ_3 and θ_4, respectivement.*

Soit $\mathcal{Z}^{\mathbf{y}} = \{((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X}\times\mathcal{Y})^N : x_i \in G(y_i), i = 1, \ldots, N\}$ la collection d'échantillons complets compatibles avec l'échantillon observé $\mathbf{y} = (y_1, \ldots, y_N)$, avec $G(y_i) = A_j$ si $y_i = b_j$.

Proposition 2. *Si θ est séparable pour $(M'|\mathbf{p}')$, alors $\cup_{\mathbf{z}\in\mathcal{Z}^{\mathbf{y}}} \arg\max_{\theta\in\Theta} L^{\mathbf{z}}(\theta) \subseteq \arg\max_{\theta\in\Theta} L^{\mathbf{y}}(\theta)$.*

Voyons les conséquences de l'hypothèse de distribution conditionelle uniforme.

Proposition 3. *Si $\mathbf{y} \in \mathcal{Y}^N$ est l'échantillon observé et $\prod_{j=1}^{r}\prod_{k:n_{kj}\neq 0} p_{k.|.j}^{n_{kj}}$ a une valeur c qui ne dépend ni du choix de $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$, ni de θ, alors, pour tout $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ on a $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{y};\theta)$.*

Corollaire 1. *mètre Sous l'hypothèse de distribution conditionelle uniforme, pour tout $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$, $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{y};\theta)$, où c ne dépend ni du choix de $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ ni de θ.*

Considérons les hypothèses relatives au processus d'imprécisiation :

Proposition 4. *Si $\mathbf{y} \in \mathcal{Y}^N$ est l'échantillon observé et $\prod_{k=1}^{m}\prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}}$ a une valeur c qui ne dépend pas du choix de $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$, ni de θ, alors, pour tout $\mathbf{x} \in \mathcal{X}^{\mathbf{y}}$ et l'échantillon correspondant $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ on a $\mathbf{p}(\mathbf{z};\theta) = c\mathbf{p}(\mathbf{x};\theta)$.*

Corollaire 2. *Si l'une des conditions suivantes et satisfaite :*

- *$\{A_1, \ldots, A_r\}$ forme une partition of $\mathcal{X}$*
- *l'hypothèse du sur-ensemble tient*

alors $\prod_{k=1}^{m}\prod_{j:n_{kj}\neq 0} p_{.j|k.}^{n_{kj}}$ est une constante c, avec $c = 1$ dans le premier cas.

5 Comparaison entre méthodes d'estimation

Comparons les mérites et limitations de plusieurs stratégies. Ici on ne donne que quelques pistes à l'aide d'exemples.

Approches de type EM. Soit $\mathcal{P}^{\mathcal{X}^N}$ l'ensemble des mesures de probabilité P sur l'espace mesurable $(\mathcal{X}^N, \wp(\mathcal{X}^N))$. L'algorithme EM [4] cherche à maximiser la fonction $F : \mathcal{P}^{\mathcal{X}^N} \times \Theta \to \mathbb{R} : F(\mathbf{P}, \theta) = \mathbf{L}^{\mathbf{y}}(\theta) - \mathbf{D}(\mathbf{P}, \mathbf{P}(\cdot|\mathbf{y}; \theta))$, $\forall \mathbf{P} \in \mathcal{P}^{\mathcal{X}^N}$, $\theta \in \Theta$, où $\mathbf{p}(\mathbf{x}|\mathbf{y}; \theta) = \frac{\mathbf{p}(\mathbf{x},\mathbf{y};\theta)}{\mathbf{p}(\mathbf{y};\theta)}$, pourvu que $\mathbf{p}(\mathbf{y}; \theta) > 0$. De plus, $\mathbf{D}(\mathbf{P}, \mathbf{P}')$ est la divergence de Kullback-Leibler de $\mathbf{P}'$ à $\mathbf{P}$, $\sum_{\mathbf{x}\in\mathcal{X}^N} \mathbf{p}(\mathbf{x}) \log[\frac{\mathbf{p}(\mathbf{x})}{\mathbf{p}'(\mathbf{x})}]$, où $\mathbf{p}$ est la masse associée à $\mathbf{P}$. Il est alors clair que $L^{\mathbf{y}}(\theta) \geq F(\mathbf{P}, \theta)$ et que si $\mathbf{P} = \mathbf{P}(\cdot|\mathbf{y}; \theta)$, alors $F(\mathbf{P}, \theta) = \mathbf{L}^{\mathbf{y}}(\theta)$. Etant donnée la valeur $\theta^{(n-1)}$ obtenue à la $n-1$ème étape **M**, l'étape E calcule $\mathbf{P}(\cdot|\mathbf{y}; \theta^{(n-1)})$ (en fait, elle détermine un échantillon fictif $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$), et la $n$ème étape **M** trouve une valeur de θ qui maximise $F(\mathbf{P}(\cdot|\mathbf{y}; \theta^{(n-1)}), \theta)$, soit $L^{\mathbf{y}}(\theta)$ basée sur l'échantillon fictif $\mathbf{z}$. Donc l'algorithme EM trouve itérativement un modèle probabiliste paramétrique $\mathbf{P}^{\theta}$ et une distribution de probabilité $\mathbf{P}(\cdot|\mathbf{y}; \theta)$ sur $\mathcal{X}$, en accord avec les données $\mathbf{y}$, de façon à ce que la divergence de $\mathbf{P}^{\theta}$ à $\mathbf{P}(\cdot|\mathbf{y}; \theta)$ soit minimale [2]. $\mathbf{P}^{\theta}$ est une EMV pour l'échantillon fictif $\mathbf{z} \in \mathcal{Z}^{\mathbf{y}}$ en accord avec $\mathbf{P}(\cdot|\mathbf{y}; \theta)$, qui fournit la meilleure imputation de X en ce sens.

Exemple 1 ([4])**.** *Un échantillon de 197 animaux est réparti en 4 catégories, soient les données :*

$$n_{\cdot 1} = 125, n_{\cdot 2} = 18, n_{\cdot 3} = 20, n_{\cdot 4} = 34.$$

La première catégorie est un mélange de 2 sous-catégories, sans qu'on sache le nombre d'individus dans chacune d'elles. Par ailleurs, un modèle génétique de la population impose les restrictions suivantes sur la distribution entre les 5 catégories : $p_{11} = 0.5$, $p_{12} = p_4$, $p_2 = p_3$. *Avec un paramètre* π *:* $p_{12} = 0.25\pi = p_4$ *et* $p_2 = 0.25(1-\pi) = p_3$, *la vraisemblance associée à un échantillon* $\mathbf{x}$ *compatible avec l'échantillon observé* $\mathbf{y}$ *est de la forme :* $L^{\mathbf{x}}(\theta) = \log[0.5^{n_{11}}(0.25\pi)^{n_{12}}(0.25(1-\pi))^{18}(0.25(1-\pi))^{20}(0.25\pi)^{34}]$, *où* $n_{11} + n_{12} = 125$. *Partons de* $\pi^{(0)} = 1/2$, *comme Dempster et al.[4]. A la premère itération, l'étape E trouve :*

$$\hat{p}_{11}^{(0)} = \frac{125}{197}\frac{0.5}{0.5+0.125} = \frac{100}{197}$$

$$\hat{p}_{12}^{(0)} = \frac{125}{197}\frac{0.125}{0.5+0.125} = \frac{25}{197},$$

$$\hat{p}_2^{(0)} = \frac{18}{197},\ \hat{p}_3^{(0)} = \frac{20}{197},\ \hat{p}_4^{(0)} = \frac{34}{197}.$$

Cette distribution empirique est en accord avec l'échantillon observé. Elle partage $n_{\cdot 1}$ *en* $n_{11} = 100$ *et* $n_{12} = 25$). *Puis, l'étape M trouve l'EMV de* π *pour un échantillon fictif* $\mathbf{x}$ *induit par cette distribution empirique. Cet EMV vaut* $\pi^{(1)} = 0.608247423$. *Ce n'est pas un EMV pour l'échantillon observé. En fait, cet EMV pour l'échantillon observé est unique et vaut* $\pi^* = 0.6268214980$. *Il peut être atteint en itérant la procédure.*

Dans cet exemple, le paramètre $\theta = \pi$ relatif à Y détermine la distribution de X, et donc le maximum de $L^{\mathbf{y}}$ donne une distribution unique pour X. Mais il y a des situations où le résultat de l'algorithm EM est loin d'être unique [2] :

Exemple 2. *Supposons qu'on lance un dé et* X *décrit le résultat obtenu. La distribution de probabilité de* X *est un vecteur* $(p_1, \ldots, p_6) \in [0,1]^6$, *avec* $\sum_{i=1}^6 p_i = 1$. *Supposons qu'après chaque lancer, on nous dit soit que le résultat est plus petit ou égal à 3 (*A_1*) ou plus grand ou égal à 3 (*A_2*). Pour chaque lancer où le résultat* (X) *est 3, l'informateur doit décider entre* A_1 *or* A_2. *Supposons que la probabilité conditionnelle* $P(y_n = A_1|X_n = 3)$ *soit une constante* $\alpha \in [0,1]$ *pour tout lancer,* $n = 1, \ldots, N$. *On lance le dé* $N = 1000$ *fois et on nous dit* $n_{\cdot 1} = 300$ *fois que le résultat était plus petit ou égal à 3. Soit* θ *le vecteur* $(p_1, p_2, p_3, p_4, p_5; \alpha)$. *La fonction de vraisemblance basée sur l'échantillon observé* $\mathbf{y}$ *peut s'écrire :* $\mathbf{p}(\mathbf{y}; \theta) = (p_1+p_2+\alpha p_3)^{300} \cdot [1-(p_1+p_2+\alpha p_3)]^{700}$. *Cette fonction a son maximum pour tout vecteur* θ *satisfaisant la contrainte* $p_1 + p_2 + \alpha p_3 = 0.3$. *Si on emploie l'algorithme EM, on obtient un vecteur* θ *satisfaisant la contrainte ci-dessus dès la première itération après l'étape M. On obtiendrait un vecteur* $\theta^{(1)}$ *qui dépend du choix initial de* $\theta^{(0)}$. *Si on commence avec* $\theta^{(0)} = (\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}; \frac{1}{2})$, *on obtient* $\theta^{(1)} = (0.12, 0.12, 0.16, 0.2, 0.2; \frac{3}{8})$. *C'est*

aussi l'EMV de θ basé sur un échantillon (fictif) de 1000 lancers du dé où la répartition des résultats est 120, 120, 160, 200, 200, 200. Mais ce n'est pas la seule EMV basée sur l'échantillon observé.

Les stratégies maximax/maximin L'estimation parametrique basée sur les stratégies maximax or maximin ne coïncide pas en général avec l'EMV. De plus elles peuvent donner des imputations de X discutables.

Exemple 3. *Supposons qu'on lance une pièce de monnaie $N = 10$ fois, et que Pierre annonce 4 faces, 2 piles ne dit rien le reste du temps, sur les 4 autres lancers. Le paramètre θ est de la forme (p, α, β), où $p = P(X = h)$, $\alpha = P(\Gamma = \{h, t\}|X = h)$ et $\beta = P(\Gamma = \{h, t\}|X = t)$. Il détermine la distribution de probabilité jointe sur (X, Y) : $P(h, \{h\}) = (1 - \alpha)p; P(h, \{h, t\}) = \alpha p; P(t, \{t\}) = (1 - \beta)p; P(t, \{h, t\}) = \beta p$;, et 0 sinon. L'EMV de θ n'est pas unique. Elle est atteinte par tous les vecteurs $\theta = (p, \alpha, \beta) \in [0, 1]^3$ qui satisfont les contraintes : $(1-\alpha)p = 0.4$ et $(1-\beta)(1-p) = 0.2$, correspondant aux probabilités marginales $P(\Gamma = \{h\})$ and $P(\Gamma = \{t\})$ respectivement. En revanche, la stratégie maximax cherche une paire $(\theta^*; \mathbf{x}^*) = (p^*, \alpha^*, \beta^*; \mathbf{x}^*)$ qui maximise $L^{\mathbf{z}}(\theta)$. On peut voir que le tuple qui maximise $L^{\mathbf{z}}(\theta)$ est unique. Il correspond au vecteur de paramètres $\theta^* = (p^*, \alpha^*, \beta^*) = (0.8, 0.5, 0)$ et à l'échantillon où tous les résultats inconnus sont "face". Autrement dit,la stratégie maximax suppose que tous les résultats manquants sont les mêmes que les résultats les plus fréquemment observés ("face"). Dans ce cas, la probabilité de face est estimée par la fréquence correspondante (0.8). Selon cette stratégie, et sans avoir la moindre information sur le comportement de Pierre, on prédit que le résultat est en fait "face" chaque fois qu'il refuse de l'annoncer.*

Exemple 4. *Considérons à nouveau l'example 1. En notant $p_{12} = 0.25\pi = p_4$ and $p_2 = 0.25(1 - \pi) = p_3$, la matrice $(M'|\mathbf{p}')$ est de la forme*

$$\left(\begin{array}{ccccc|c} \frac{0.5}{0.5+0.25\pi} & \frac{0.25\pi}{0.5+0.25\pi} & 0 & 0 & 0 & 0.5 + 0.25\pi \\ 0 & 0 & 1 & 0 & 0 & 0.25(1-\pi) \\ 0 & 0 & 0 & 1 & 0 & 0.25(1-\pi) \\ 0 & 0 & 0 & 0 & 1 & 0.25\pi \end{array}\right)$$

et ne dépend que d'un seul paramètre. La famille $\{A_{11}, A_{12}, A_2, A_3, A_4\}$ forme une partition de $\mathcal{X}$ (les 5 catégories d'animaux). Donc, selon le Corollaire 2, $L^{\mathbf{x}}(\theta)$ et $L^{\mathbf{z}}(\theta)$ coïncident pour tout θ. Donc, les stratégies basées sur leurs maximisations respectives donnent les mêmes paires optimales. Voyons la stratégie maximax. Elle donne $(\mathbf{x}^*, \theta^*)$, avec $\theta^* = 0.4722222$. l'échantillon optimal correspondant $\mathbf{x}^*$ est le seul qui soit induit par la distribution empirique $(\frac{125}{197}, 0, \frac{18}{197}, \frac{20}{197}, \frac{34}{197})$. Donc, selon ce critère, les 125 résultats observés du groupe 1 sont dans le premier sous-groupe, encore un choix discutable. La stratégie maximin mène au choix opposé, soit $\mathbf{x}_* = (0, \frac{125}{197}, \frac{18}{197}, \frac{20}{197}, \frac{34}{197})$ car $0.25\pi < 0.5$. Le paramètre maximin optimal π_* est $159/197 = 0.8071$. Notons que ce paramètre n'est pas séparable pour M'. Il s'avère qu'aucune des valeurs $\pi^* = 0.4722222$, $\pi_* = 0.8071$ ne coïncide avec l'EMV de π relatif à $L^{\mathbf{y}}(\theta)$,soit $\hat{\pi} = 0.6268214980$.

Exemple 5. *Considérons le cas de la pièce dans l'exemple 3, et supposons de plus qu'on sache que : $\alpha = 1 - \alpha = 0.5$ and $\beta = 1 - \beta = 0.5$. Autrement dit, indépendamment du résultat (pile ou face) Pierre refuse de le donner avec une probabilité 0.5 (le comportement de Pierre ne dépend pas de ce résultat). C'est l'hypothèse du sur-ensemble [8] déjà mentionnée. Avec cette contrainte supplémentaire, l'EMV de $\theta = (p, 0.5, 0.5)$ est atteint pour $\hat{p} = 4/6 = 2/3$. Cette EMV fournit la même estimation que si l'on avait lancé la pièce six fois, car, comme conséquence de l'hypothèse du sur-ensemble ici, les quatre autres lancers ne jouent aucun rôle dans nos statistiques. En conséquence, la probabilité conditionnelle $P(X = h|\Gamma = \{h, t\})$ est supposée coïncider avec $P(X = h|\Gamma \neq \{h, t\})$ et avec $P(X = h) = p$.*

Cette probabilité est estimée à partir des six résultats observés, dont 4 sont "face" et le reste est "pile". En revanche, la stratégie maximax sans l'hypothèse du sur-ensemble nous amène à prendre en compte les lancers non observés en les supposant conformes aux résultats observés les plus fréquents, soit l'imputation de X correspondant à un jeu de données comportant 8 fois face et deux 2 fois pile.

L'approche maximin considère toutes les fonctions de log-vraisemblance $L_k^{\mathbf{x}}(p) = (4+k)\log p + (6-k)\log(1-p)$ avec $0 \leq k \leq 4$. Pour chaque valeur de p on doit trouver le jeu de données complet qui minimise $L^{\mathbf{x}}(p)$. Comme $L_k^{\mathbf{x}}(p)$ est de la forme $k \log \frac{p}{(1-p)} + a$, il est facile de voir que si $p < 1/2$, le minimum $L^-(p)$ est atteint pour $k = 4$, et si $p > 1/2$, il est atteint pour $k = 0$. Donc, c'est $8\log p + 2\log(1-p)$ si $p < 1/2$ et $4\log p + 6\log(1-p)$ sinon. On voit que $L^-(p)$ est croissant si $p < 1/2$ et décroissant si $p > 1/2$. Son maximum est atteint pour $p = 1/2$. Donc l'approche maximin est prudente au sens où on suppose que le lancer de pièce maximise l'entropie. On trouve une distribution uniforme (5 fois pile et 5 fois face), en accord avec les observations.

6 L'approche EM evidentielle

T. Denoeux décrit ce qu'il appelle l'algorithme EM Evidentiel (EEM) dans [5] quand le jeu de données est à la fois imprécis et incertain, et défini par une fonction de masse sur $\wp(\mathcal{X}^N)$. Dans le cas général, l'information disponible (incomplète) sur l'échantillon $\mathbf{x} \in \mathcal{X}^N$ est représenté par une fonction de masse μ, qu'on peut voir comme une distribution de probabilité sur $\mathcal{Y}^N$. On considère le cas particulier où μ possède un seul element focal $B \in \mathcal{X}^N$ (correspondant à un échantillon $\mathbf{y}_B$). L' approche EEM propose une expression de la fonction de vraisemblance sur $\mathcal{X}$, basée sur ce jeu de données B imprécis (équation 16 dans [5]) :

$$\lambda(\mathbf{y}_B;\theta) = P(\mathbf{x} \in B;\theta) = \sum_{\mathbf{x}\in B} \mathbf{p}(\mathbf{x};\theta).$$

En particulier, si on suppose que B est un produit cartésien de N ensembles de la famille $\{A_1,\ldots,A_r\}$, où $n_{\cdot j}$ est le nombre de répétitions de A_j dans $\mathbf{y}_B$, $\lambda(\mathbf{y}_B;\theta)$ peut alternativement s'écrire comme suit, en supposant l'indépendence entre les observations :

$$\lambda(\mathbf{y}_B;\theta) = \prod_{j=1}^{r} P(X \in A_j;\theta)^{n_{\cdot j}}.$$

Quand les données sont incertaines et modélisées par la fonction de masse μ, cette fonction de vraisemblance généralisée est de la forme [5] :

$$\sum_{B\subseteq\mathcal{X}^N} \mu(B)\lambda(\mathbf{y}_B;\theta) = \sum_{\mathbf{x}\in\mathcal{X}^N} \mathbf{p}(\mathbf{x};\theta) pl(\mathbf{x}).$$

où $pl(\mathbf{x}) = \sum_{\mathbf{x}\in B} \mu(B)$ est la fonction de contour induite par μ sur $\mathcal{X}^N$ [11]. L'algorithme EM Evidentiel est une variante de l'algorithme EM classique qui choisit une valeur de θ maximisant la fonction de vraisemblance généralisée $\lambda(\mathbf{y}_B;\theta)$.

Cependant la procédure EEM ne coïncide pas toujours avec une EMV au sens de $L^{\mathbf{y}}$. En effet

- Le critère EEM n'effectue pas une EMV relativement à l'espace $\mathcal{Y}$, car le résultat est une distribution sur $\mathcal{X}$.
- Ce critère n'est pas toujours dans l'esprit d'une fonction de vraisemblance sauf si le jeu de données imprécises forme une partition de $\mathcal{X}$. En fait, $P(X \in A_j)$ ne coïncide pas toujours avec $P(Y = b_j)$.

Comme signalé plus haut si $\mathcal{X}$ est l'ensemble des résultats attendus, l'ensemble A_j apparaissant dans $\mathbf{y}_B$ n'est pas une réalisation de Y, c'est un événement pour X. Donc $P(X \in A_j;\theta)$ n'est pas une fonction vraisemblance relative à $\mathcal{X}$ en accord avec la vision traditionnelle exposée par Edwards [6]. Si on veut considérer A_j comme une réalisation, on doit le voir comme un singleton sur l'ensemble des parties $\wp(\mathcal{X})$. Notons que même si $A_i \cap A_j \neq \emptyset$, on a $\{A_i\} \cap \{A_j\} = \emptyset$ (les réalisations distinctes sont toujours mutuellement exclusives).

On peut donc distinguer les cas suivants :

— Le cas où $Im(\Gamma)$ forme une partition de $\mathcal{X}$. Alors, $P(X \in A_j) = P(Y = A_j) = p_{.j}$, $\forall j = 1, \ldots, r$, et donc $\prod_{j=1}^{r} P(X \in A_j; \theta)^{n_{.j}}$ coïncide avec la fonction de vraisemblance $\mathbf{p}(\mathbf{y}_B; \theta)$.
— Le cas où les ensembles $A_1, \ldots, A_r$ se chevauchent. Alors, $\mathbf{p}(\mathbf{y}_B; \theta)$ et $\lambda(\mathbf{y}_B; \theta)$ peuvent ne pas coïncider.

Exemple 6. *Dans l'exemple 2, supposons qu'on nous annonce n_1 fois que le résultat est plus petit ou égal à 3, et $n_2 = N - n_1$ fois qu'il est plus grand ou égal à 3. Le critère EEM est $\lambda(\mathbf{y}_B; \theta) = (p_1+p_2+p_3)^{n_1} \cdot (p_3+p_4+p_5+p_6)^{n_2}$ avec $\sum_{i=1}^{6} p_i = 1$. on voit facilement que ce critère atteint son maximum ($\lambda(\mathbf{y}_B; \theta) = 1$) pour le vecteur θ tel que $p_3 = 1$. Mais une telle estimation de θ ne semble pas crédible. La fonction de vraisemblance standard sur $\mathcal{Y}$ est de la forme $\mathbf{p}(\mathbf{y}_B; \theta) = \pi^{n_1} \cdot (1 - \pi)^{n_2}$ avec $\pi = p_1 + p_2 + \alpha p_3$ (voir l'example 2).*

Il y a donc des situations où la méthode EEM risque de donner des résultats contre-intuitifs.

7 Conclusion

Ce travail suggère qu'il n'est pas trivial d'étendre l'EMV aux données incomplètes malgré l'existence de plusieurs propositions. En particulier, il est très discutable de reconstruire la distribution des variables non-observées quand les paramètres qui la définissent ne sont pas étroitement contraints par les paramètres des distributions qui gouvernent les variables observées. Le célèbre article sur EM [4] ne traite que les observations imprécises formant une partition de l'espace non-observé et propose un exemple didactique où un seul paramètre détermine la distribution jointe de X et Y. Il n'est pas facile d'adapter la procédure EM aux données incomplètes qui se chevauchent. En général, soit on applique une EMV standard sur $\mathcal{Y}$ aux observations imprécises vues comme des réalisations (on obtient une masse de Möbius sur $\mathcal{X}$), soit on ajoute une hypothèse qui revient à sélectionner une mesure de probabilité unique dans l'ensemble convexe (crédal) induit par cette fonction de masse. Chaque approche de l'EMV avec données imprécises propose son hypothèse. Comme le montrent les exemples, il est facile de trouver des cas où ces méthodes fournissent des solutions discutables : celle de l'algorithme EM [4] peut dépendre de la valeur initiale du paramètre, l'approche EEM [5] optimise un critère qui n'est pas toujours une fonction de vraisemblance, l'approche maximax [8] peut choisir une distribution trés déséquilibrée pour la variable latente, tandis que l'approche maximin robuste [7] favorise des distributions peu informatives. D'autres efforts sont nécessaires pour caractériser des classes of problèmes où une méthode d'estimation se justifie alors que les autres méthodes échouent.

Références

[1] I. Couso, D. Dubois, Statistical reasoning with set-valued information : Ontic vs. epistemic views. Int. J. of App. Reas., 55(7),1502-1518, 2014.

[2] I. Couso, D. Dubois, Belief revision and the EM algorithm, IPMU 2016, Proceedings, Part II, Springer, 279-290.

[3] A. P. Dempster. Upper and lower probabilités induced by a multivalued mapping. Ann. Math. Statistics, 38, 325-339, 1967.

[4] A.P. Dempster, N.M. Laird, D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B, 39, 1-38, 1977.

[5] T. Denoeux, Maximum Likelihood Estimation from Uncertain Data in the Belief Function Framework, IEEE Trans. Knowledge and Data Eng. 26, 119-130, 2013.

[6] A.W.F. Edwards. Likelihood. Cambridge University Press, 1972.

[7] R. Guillaume, D. Dubois, Robust parameter estimation of density functions under fuzzy interval observations, 9th ISIPTA Symposium, Pescara, Italy, 2015.

[8] E. Hüllermeier Learning from imprecise and fuzzy observations Int. J. App. Reas., 55(7) 1519-1534, 2014.

[9] E. Hüllermeier, W. Cheng : Superset Learning Based on Generalized Loss Minimization. ECML/PKDD (2), 260-275, 2015.

[10] M. Serrurier, H. Prade : An informational distance for estimating the faithfulness of a possibility distribution, viewed as a family of probability distributions, with respect to data. Int. J. App. Reas. 54(7) : 919-933, 2013.

[11] Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[12] P. Smets, Decision making in the TBM : the necessity of the pignistic transformation Int. J. App. Reas., 38,133-147, 2005.