UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1
SCIENCES & GÉOGRAPHIE

N° attribué par la bibliothèque

|_|_|_|_|_|_|_|_|_|_|

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER

Discipline : Mathématiques Appliquées

Présentée et soutenue publiquement
par
VÉRONIQUE VENDITTI

Le 3 juillet 1998

ASPECTS DU PRINCIPE DE MAXIMUM D'ENTROPIE EN MODÉLISATION STATISTIQUE

Co-directeurs de thèse
Gilles Celeux et Claudine Robert

COMPOSITION DU JURY

Mr Jean-Louis Soler,	Président
Mme Catherine Huber,	Rapporteur
Mr Henri Caussinus,	Rapporteur
Mme Claudine Robert,	Co-directeur de thèse
Mr Gilles Celeux,	Co-directeur de thèse
Mr Jacques Demongeot	

Remerciements

Je tiens à remercier en tout premier lieu Mr Gilles Celeux d'avoir accepté, il y a quelques années de cela, que je débute une thèse sous sa direction. De l'action Systol au projet IS2, j'ai ainsi suivi la vie et l'évolution d'un projet de recherche INRIA, ainsi que son intégration au sein de la communauté scientifique de Grenoble. J'ai découvert grâce à Mr Gilles Celeux le monde de la recherche dans les meilleures conditions possibles, au point de vue matériel bien sûr, mais surtout au point de vue de l'encadrement, de l'écoute et de l'attention qu'il a porté à mon travail. Je le remercie de m'avoir fait profiter de sa grande connaissance de la statistique et de l'analyse des données.

Mme Claudine Robert a co-dirigé ce travail de thèse. Elle m'a fait découvrir, avec patience et indulgence, le principe de maximum d'entropie qui constitue la base de ce travail. Avoir été au fil de ces années "sa thésarde", a été pour moi un honneur et aussi tout simplement un plaisir. Je la remercie pour sa disponibilité et son attention face à mes multiples et répétées questions, et pour l'investissement qu'elle a consenti dans cette thèse. Au delà de la direction rigoureuse de mon travail, Mme Claudine Robert m'a fait bénéficier de sa connaissance très fine de l'analyse statistique ainsi que de son expérience en matière d'enseignement. Je tiens à lui exprimer ma profonde gratitude pour tout ce qu'elle m'a appris, en statistique bien sûr, mais aussi en méthodes de travail, et surtout en rigueur scientifique.

Je remercie Mr Henri Caussinus et Mme Catherine Huber d'avoir accepté de rapporter sur mon travail de thèse. Leur intérêt pour ce travail et leurs judicieuses remarques m'ont permis d'en affiner la rédaction, et m'ont poussée à aller plus avant dans certaines réflexions.

Mr Jean-Louis Soler a accepté de présider mon jury de thèse et je l'en remercie.

En parallèle à mon travail de thèse, j'ai découvert le plaisir (et les difficultés) d'enseigner. Je remercie les responsables d'enseignement qui m'ont fait confiance et m'ont donné la responsabilité des divers cours que j'ai assurés durant ces années. Merci également à mes étudiants de m'avoir laissé croire que je pouvais les guérir de leur allergie à la statistique.

Merci enfin à tous ceux qui ont supporté mes crises d'angoisse, de déprime, de découragement, et autres manifestations classiques et cycliques de l'état de stress quasi permanent (qu'il soit justifié ou non) dans lequel se trouvent les thésards de toutes disciplines. Je remercie donc tous mes amis et voisins, mes correspondants "mailistiques", les buveurs de thé et/ou de café de l'INRIA, les thésards de statistique de la tour IRMA, mes très patients collègues de bureau (surtout Isa, Catherine et Henri), et toute l'équipe du projet IS2, qu'ils soient permanents ou qu'ils aient été de passage à Grenoble.

Je ne pourrais jamais faire comprendre à ma mère ce que signifie $-\sum_x p_x \ln p_x$, ni pourquoi toutes ces équations m'importaient tellement, mais je veux ici la remercier du plus profond de mon cœur pour son soutien et son affection tout au long de mes années d'étude. S'il n'y a ne serait-ce qu'une phrase dont je puisse être fière dans cette thèse, je la dédie à mon père.

Et merci au hasard de m'avoir fait devenir un jour la voisine de bureau de Frank.

Table des matières

Avant-propos

La notion d'entropie apparaît dans de nombreux domaines (physique statistique, théorie de l'information, systèmes dynamiques, sciences sociales, ...). Ce terme se retrouve même dans le langage courant comme un synonyme de “désorganisation” et la maximisation de l'entropie se comprend comme une “diminution de l'information” ou une “augmentation du désordre”.

Nous nous proposons dans cette thèse de préciser la définition d'un principe de modélisation fondé sur l'utilisation d'une entropie définie comme une fonctionnelle sur l'ensemble des lois de probabilités. Nous avons pour but d'en donner une présentation facilement abordable et compréhensible, et de le justifier en dehors de toute approche subjective. Cette définition étant posée, nous chercherons à la confronter à l'approche habituelle des modèles classiques. Nous tâcherons également d'apporter des éléments de réponse à la question de savoir si un tel principe peut être utilisé en pratique comme un outil de modélisation à part entière.

Ainsi, ce travail s'articule autour de la présentation de ce principe, de son utilisation d'un point de vue théorique et d'une application à des données médicales.

Dans un premier chapitre, nous présenterons le principe de maximum d'entropie d'un point de vue théorique général (**Choix d'observables et nature des niveaux de contraintes dans l'application du principe de maximum d'entropie**). Notre approche se différencie des présentations habituelles de l'utilisation de la notion d'entropie en modélisation statistique. Nous définissons en effet ce principe à partir d'un théorème de concentration qui lui assure directement une justification. Nous mettrons d'autre part en avant de manière précise les particularités de ce principe et nous décrirons de quelle façon nous concevons son application.

Dans cette approche de la modélisation statistique il ne s'agit pas, comme cela est fait habituellement, de choisir directement la structure statistique qui semble le mieux correspondre au phénomène aléatoire étudié. En effet, il s'agit en premier lieu de traduire l'information que l'on possède sur ce phénomène au niveau de la loi des variables qui le représentent. Cette connaissance est formalisée par l'association entre des fonctions de ces variables (les *observables*) et les niveaux de leurs espérances (les *niveaux de contraintes*). Nous poserons dans ce travail deux principes de choix de ces éléments de connaissance en nous limitant, au moins dans un premier temps, à des observables les plus simples possibles en nombre et en nature. Nous nous sommes appuyés sur ces choix de parcimonie et de simplicité dans notre lecture théorique de ce principe et dans son application à des données réelles.

Après avoir décidé d'une information, il faut choisir une loi parmi l'ensemble des lois de probabilités qui répondent aux contraintes ainsi établies. Dans le cas d'une unique variable réelle, on peut par exemple considérer que les deux premiers moments constituent l'essentiel de l'information qu'apporterait un échantillon du phénomène. Cela revient à choisir parmi les lois dont l'espérance et la variance correspondent à des valeurs imposées a priori ou estimées (de manière subjective ou à partir de réalisations du phénomène). La loi correspondant à cette information élémentaire est alors la loi normale d'espérance et variance fixées par l'information.

Une manière de voir cette information revient à dire que disposer d'éléments de connaissance supplémentaires, c'est réduire l'incertitude qui pèse sur la détermination du modèle le plus probable, puisque le nombre de contraintes que la loi doit vérifier augmente d'autant plus. Les présentations habituelles laissent alors le choix entre un grand nombre de mesures de cette incertitude, désignées par le terme générique d'entropie. Nous préférons baser notre propos sur un théorème qui assure la concentration des lois empiriques dans le voisinage de la loi maximisant une certaine fonction définie sur l'ensemble des lois de probabilités, en l'occurence l'entropie de Shannon. Ce théorème justifie donc de définir le principe de modélisation suivant : on choisit, parmi les lois vérifiant certaines contraintes que l'on a jugées essentielles, la loi qui maximise l'entropie de Shannon. Ce principe est alors validé puisque la loi choisie est, en un certain sens, la plus probable pour une information donnée. C'est cette information qu'il faudra modifier si le modèle obtenu n'est finalement pas jugé adéquat.

Nous détaillerons une démonstration concise et rapide de la forme des lois issues de l'application de ce principe. Les propositions que nous présentons ont pour but d'apporter un éclairage précis sur la nature de ces lois, en évitant en particulier de différencier le type des variables mises en jeu. Notre présentation ne nécessite plus l'utilisation et la généralisation à tous les types de variables, discrètes ou continues, de la méthode des multiplicateurs de Lagrange habituellement utilisée pour maximiser sous contraintes l'entropie.

Notre manière d'appréhender ce principe de modélisation se traduit enfin par l'utilisation de niveaux de contraintes empiriques dans l'écriture et la résolution des équations permettant d'estimer les paramètres des lois obtenues. Nous montrerons alors que le mode de calcul des paramètres des lois de maximum d'entropie qu'impose ce choix se trouve être la méthode de maximum de vraisemblance. Nous apportons donc une justification à ce principe d'estimation, au sens où les équations qu'il définit découlent naturellement de l'application du principe de maximum d'entropie avec utilisation de niveaux exclusivement empiriques. Nous préciserons de plus les liens existant entre différentes lectures des modèles paramétrés et leur relecture en tant que modèle de maximum d'entropie.

Nous présenterons enfin divers exemples d'informations élémentaires définies à partir de variables soit de même type, soit de types différents et mèlant éventuellement une information empirique et une information a priori de certains niveaux de contraintes. Nous verrons comment de telles informations permettent de retrouver des modèles classiques dont la forme pourra alors être relue comme une conséquence du choix d'une famille d'observables.

Dans un deuxième chapitre, nous présenterons une utilisation du principe de maximum d'entropie comme outil de relecture de divers modèles de discrimination (**Le principe de maximum d'entropie : un nouveau regard sur les modèles de discrimination classiques**).

Dans une première partie, nous détaillerons les familles d'observables qui permettent de retrouver l'expression, en tant que modèles de maximum d'entropie, de différents modèles de discrimination (analyse discriminante linéaire, quadratique, modèle d'indépendance conditionnelle, régression logistique). Cette utilisation du principe de maximum d'entropie illustre son intérêt didactique. Une telle approche permet en effet de présenter un modèle comme l'expression correspondant naturellement à un certain résumé des données, et non comme une structure statistique choisie parmi une infinité de structures possibles. Cela fournit une manière de caractériser et de classifier ces modèles, puisqu'il sera possible de les comparer en terme de familles d'observables.

Nous montrerons de plus certaines limites du principe de maximum d'entropie qui ne permet pas de résoudre le problème du choix des probabilités a priori d'appartenance aux groupes quand celles-ci sont inconnues. Bien que l'application du principe de maximum d'entropie en fournisse une expression analytique à partir d'une information n'imposant aucune connaissance sur ces probabilités, il ne propose en effet aucune alternative aux solutions habituellement utilisées (en particulier prendre comme loi du groupe la loi équirépartie ou les fréquences observées).

Dans une deuxième partie, nous nous attacherons plus particulièrement à la méthode de régression logistique, couramment utilisée par exemple dans le domaine médical.

Ce modèle impose la linéarité du logit de la probabilité conditionnelle d'appartenance à un groupe de population (malades/non malades) en fonction des valeurs des variables explicatives (symptômes). On considère habituellement que les équations d'estimation par maximum de vraisemblance des paramètres de ce modèle n'ont pas à prendre en compte la nature de l'échantillon utilisé. Nous avons montré par un contre-exemple que cette équivalence des schémas d'échantillonnage n'est qu'asymptotique. Nous préférons donc parler de *procédure de régression logistique*, qui associe une certaine écriture des probabilités conditionnelles et une méthode donnée d'estimation des paramètres.

Nous préciserons d'autre part comment cette procédure peut être définie comme un modèle de maximum d'entropie. S'ils sont associés à des niveaux exclusivement empiriques, les observables élémentaires lui correspondant caractérisent en fait une situation de régression où les données sont exploitées de la manière la plus synthétique possible. La méthode d'estimation des paramètres ne pose alors plus de problème d'interprétation puisqu'il s'agit des équations découlant naturellement de l'application du principe de maximum d'entropie et dont la forme ne dépend pas du schéma d'échantillonnage à l'origine des données.

Dans un troisième chapitre, nous présenterons une application à des données réelles des diverses notions sur lesquelles se basent les résultats précédents (simplicité des observables, choix de l'entropie de Shannon, utilisation des niveaux de contraintes empiriques, ...) (**Un exemple de construction de modèles de maximum d'entropie**). Nous avons utilisé une série de données représentant des durées de séjour dans le service de gériatrie du CHU de Grenoble. La distribution empirique de ces données a la particularité de ne correspondre à aucune loi théorique usuelle. Différents modèles de maximum d'entropie ont été déterminés en faisant évoluer les observables utilisés, en nombre et en nature. Nous avons donc testé différentes familles d'observables, dans un ordre de complexité croissant pour étudier le comportement des lois de maximum d'entropie obtenues.

Chapitre I

Choix d'observables et nature des niveaux de contraintes dans l'application du principe de maximum d'entropie

1 Introduction

Nous nous plaçons dans le cadre général de la modélisation d'un phénomène aléatoire mettant en jeu un vecteur $\mathbf{X}$ constitué de J composantes $(X_1, \ldots, X_J)$, à valeurs dans un espace de nature quelconque $\mathcal{X}$. La modélisation statistique consiste à déterminer une loi de probabilité P de $\mathbf{X}$ qui décrive au mieux la réalité du phénomène. On peut choisir de traduire la connaissance dont on dispose au sujet de celui-ci par diverses contraintes au niveau de cette loi P. La détermination d'un modèle devra alors être en accord avec les contraintes ainsi imposées.

• On appelle **élément de connaissance** tout couple (φ, a) où φ est une fonction de $\mathcal{X}$ dans $\mathbf{R}$ et a un nombre réel. Les fonctions φ traduiront par exemple des moments de certaines variables, des liaisons que l'on juge essentielles entre certaines variables, des transformations par logarithme, des fonctions indicatrices de certaines valeurs particulières. On admet que ces fonctions permettent de traduire la connaissance que l'on a de certaines caractéristiques du phénomène au travers des équations correspondant à l'écriture de leurs espérances :

$$E_P(\varphi) = a.$$

Nous reviendrons sur l'utilisation de ces équations au paragraphe 5, p. 36.

• On appelle **information** de taille M, pour un problème de modélisation donné, un ensemble de M éléments de connaissance. On note $\mathcal{I}$ cet ensemble :

$$\mathcal{I} = \{(\varphi_m, a_m), m \in \{1, \ldots, M\}\}.$$

• Chaque fonction φ_m est appelée **observable** et chaque réel a_m est le **niveau de contrainte** associé. Le vocabulaire utilisé est issu de la physique et des théories sur les grandes déviations ([Ellis 85]).

• On note $\mathcal{M}(\mathcal{X})$ l'ensemble des lois de probabilité sur $\mathcal{X}$ et $\mathcal{M}_{\mathcal{I}}$ l'ensemble de celles vérifiant les contraintes associées à l'information :

$$\mathcal{M}_{\mathcal{I}} = \{P \in \mathcal{M}(\mathcal{X})\ ,\ E_P(\varphi_m) = a_m\ ,\ m \in \{1, \ldots, M\}\}.$$

La modélisation se traduira donc par le choix d'une loi P parmi toutes les lois dans $\mathcal{M}_{\mathcal{I}}$ pour lesquelles les espérances des fonctions φ_m, $m \in \{1, \ldots, M\}$, sont fixées (on "travaille à espérances des φ_m fixées"). Nous choisirons le plus souvent des contraintes linéaires.

Nous avons pour but dans ce texte de répondre aux points suivants :

– l'information doit permettre de déterminer à la fois la structure statistique recherchée et les équations permettant de calculer les paramètres qu'elle met en jeu. Suivant quels principes doit-on formaliser cette information?

– les théorèmes classiques associés au principe de maximum d'entropie (PME) sont des théorèmes d'existence des lois de maximum d'entropie. Mais comment déterminer *en pratique* la forme de ces lois?

Les paragraphes 2, p. 17, et 3, p. 19, de ce chapitre présentent le cadre d'application le plus simple du PME. Nous y rappelons la justification théorique de ce principe par la propriété de concentration des lois empiriques au voisinage de la loi de maximum d'entropie, pour une information donnée.

Dans le paragraphe 4, p. 24, nous démontrons que dans le cas d'un espace continu ($\mathbf{R}^J$) ou discret ($\mathbf{Z}^J$) et dans certains cas mixtes, s'il existe un état de Gibbs satisfaisant aux contraintes de modélisation choisies, il s'agit de l'unique loi de maximum d'entropie. Cette proposition suffit dans de nombreux cas pour mettre en pratique le PME : pour une famille de contraintes données, on regarde s'il existe un état de Gibbs les satisfaisant. C'est à partir de l'application de ce résultat que nous retrouverons tout au long de ce document les modèles les plus usuels de la statistique.

Plus précisément, un tel état de Gibbs définit une structure statistique exponentielle faisant apparaître autant de paramètres que de contraintes utilisées. On désigne ces paramètres sous le terme de multiplicateurs par analogie avec les multiplicateurs de Lagrange apparaissant dans le cas discret.

Il est alors assez naturel de s'interroger sur le lien entre les équations du PME, qui sont les équations à résoudre pour calculer ces multiplicateurs, et les équations des méthodes les plus classiques d'estimations statistiques des paramètres d'une telle structure (principe du maximum de vraisemblance, méthode des moments, ...).

Cela nous amène à différencier deux types de situations de modélisation. Dans les cas les plus fréquents, on ne dispose que d'un jeu de données pour estimer les paramètres du modèle. Mais on peut avoir à notre disposition quelques informations numériques a priori (on peut par exemple inclure dans les contraintes que l'espérance d'une loi est nulle pour des raisons de symétrie). Les conséquences de ces connaissances quant à l'application du PME sont détaillées dans le paragraphe 5, p. 36.

Le dernier paragraphe 6, p. 46, présente quelques illustrations des particularités du PME au travers de cas simples. On retrouve alors l'expression de modèles classiques dont l'efficacité trouve une nouvelle explication par la propriété de concentration des lois de maximum d'entropie.

2 Stratégie de modélisation

Nous n'avons pas encore défini le principe qui nous permettra de déterminer un modèle parmi ceux possibles. Mais nous pouvons déjà proposer quelques règles concernant le choix du contenu de l'information.

Il apparaît en pratique que les observables sont le plus souvent des polynômes si $\mathcal{X}$ est un sous-espace de $\mathbf{R}^J$, et des fonctions indicatrices si $\mathcal{X}$ est discret. Ces observables sont les plus efficaces, au sens où ils permettent de retrouver les expressions de modèles classiques, et les plus utilisables, puisqu'ils aboutissent à des calculs réalisables. En effet, si les observables sont de nature plus complexe, la structure du modèle et le calcul des paramètres le seront d'autant plus. Nous choisissons donc de nous limiter à des observables de nature simple, à moins de pouvoir soigneusement justifier l'utilisation d'observables plus compliqués. Nous définissons ainsi un **principe de simplicité**.

Ainsi, nous fonderons toute modélisation sur une information *aussi simple que possible*, au moins dans une première étape. Si on juge que le modèle obtenu répond au problème posé (classement, ajustement aux données ou tout autre critère établi avant la modélisation), il n'est évidemment pas nécessaire de modifier les éléments de connaissance utilisés. Dans le cas contraire, on modifiera le contenu de l'information pour remédier à une incompatibilité aux données ou une incapacité à répondre au problème. Par exemple, si l'information se base sur la connaissance des moments d'ordres successifs, on ajoutera des moments d'ordre supérieurs; on pourrait également ajouter à l'information les valeurs de certains modes s'ils sont essentiels pour caractériser le phénomène et si l'information minimale n'a pas permis de les retrouver dans le modèle, ...

Nous préférons donc parler d'information plutôt que d'information a priori comme on le fait classiquement, car dans ce contexte on part d'une information minimale qui sera éventuellement modifiée jusqu'à ce que le modèle soit jugé suffisamment adéquat. Mais, s'il s'avère nécessaire d'ajouter des observables, nous chercherons toujours à limiter la taille et la complexité de l'information.

Nous retrouvons donc dans cette manière de faire l'idée d'un **principe de parcimonie**. En effet, un modèle sera d'autant plus compliqué et demandera d'autant plus aux données (comme nous le verrons au paragraphe 5, p. 36) que le nombre d'observables sera important. C'est sur la base de ces deux principes que nous avons choisi les observables utilisés dans l'application que nous présentons au chapitre III, p. 159.

▷ **Exemple**

Dans le cas d'une unique variable réelle X, une première information peut se fonder sur la connaissance des deux premiers moments. Ils représentent en effet les paramètres de position et de dispersion les plus simples que l'on peut définir pour décrire une distribution. L'information minimale est donc basée, dans ce cas, sur les observables $\varphi_1(x) = x$ et $\varphi_2(x) = x^2$ (notés plus simplement x et x^2). La donnée des niveaux a_1 et a_2 des moments $E_P(X)$ et $E_P(X^2)$ complètera cette information. Le modèle obtenu à partir de cette information pourra donc être considéré comme le modèle le plus parcimonieux et le plus simple, au sens où les observables qui le caractérisent sont seulement au nombre de deux et sont de nature élémentaire. (Nous verrons de plus que le principe que nous allons définir ne donne pas de solution à partir du seul observable x.)

3 Rappel : résultats de concentration

Nous allons voir à présent qu'il existe une fonctionnelle qui nous permet d'atteindre une loi telle que les distributions empiriques que l'on pourra observer en soient très proches, en un sens que nous allons définir.

3.1 Concentration des lois empiriques

3.1.1 Notations

• On note Ω l'ensemble des suites d'éléments de $\mathcal{X}$:

$$\Omega = \{\omega = (\boldsymbol{x}_n)_{n\in\mathbf{N}^*}, \boldsymbol{x}_n \in \mathcal{X}\}.$$

• On associe à tout élément ω de Ω la suite $(\rho_N)_{N\in\mathbf{N}^*}$ des distributions empiriques, pour laquelle on note :

$$\rho_N(\boldsymbol{x}) = \frac{1}{N}\sum_1^N \delta_{\boldsymbol{x}_n}(\boldsymbol{x}) \ , \ \forall \boldsymbol{x} \in \mathcal{X},$$

où $\delta_{\boldsymbol{x}_n}$ est la mesure de Dirac au point $\boldsymbol{x}_n$ de la suite $\omega = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \ldots)$.

• On considère une information $\mathcal{I}$ de taille M. On note $\mathcal{M}_\mathcal{I}$ l'ensemble des lois satisfaisant aux contraintes et $\mathcal{M}_{\mathcal{I}_\alpha}$ l'ensemble de celles qui y satisfont à α près ($\alpha > 0$) :

$$\begin{aligned} \mathcal{M}_\mathcal{I} &= \{P, P \in \mathcal{M}(\mathcal{X}), E_P(\varphi_m) = a_m \ , \ m \in \{1, \ldots, M\}\} \\ \mathcal{M}_{\mathcal{I}_\alpha} &= \{P, P \in \mathcal{M}(\mathcal{X}), |E_P(\varphi_m) - a_m| \leq \alpha \ , \ m \in \{1, \ldots, M\}\} \end{aligned}$$

• On note S la fonction qui à toute probabilité P sur $\mathcal{X}$ associe les valeurs suivantes, selon le type du vecteur $\mathbf{X}$:

– pour $\mathcal{X}$ fini ou dénombrable :

$$S(P) = -\sum_x p_x \ln p_x,$$

où $p_x = P(\boldsymbol{x})$ pour $\boldsymbol{x} \in \mathcal{X}$;

- pour $\mathcal{X} \subseteq R^J$:
  - si $dP = f d\boldsymbol{x}$: (on note indifféremment $S(P)$ ou $S(f)$)

$$S(P) = -\int_{\mathcal{X}} f(\boldsymbol{x}) \ln f(\boldsymbol{x}) d\boldsymbol{x} = -E(\ln f) \ ;$$

  - sinon : $S(P) = -\infty$;
- pour $X = (Y, T)$, Y à valeurs dans $\mathcal{Y} \subseteq \mathcal{R}^{\mathcal{J}'}$, $J' < J$, T à valeurs dans $\mathcal{T}$ fini ou dénombrable. On note P_T la loi marginale de T et Q_X^t la loi conditionnelle de X sachant $T = t$, $t \in \mathcal{T}$.

$$S(P) = E_{P_T}\left(S\left(Q_X^T\right)\right) + S\left(P_T\right).$$

3.1.2 Propriété de concentration

Nous allons utiliser la fonctionnelle S pour choisir une loi particulière parmi celles qui vérifient les contraintes imposées par l'information.

Pour cela, on munit $\mathcal{M}(\mathcal{X})$, l'ensemble des lois de probabilité sur $\mathcal{X}$, de la topologie de la convergence étroite. On suppose de plus que l'ensemble $\mathcal{M}_{\mathcal{I}}$ est non-vide et contient un élément pour lequel S est de valeur finie.

On peut montrer alors qu'il existe un élément P^* (respectivement P_α^*) qui maximise la fonctionnelle S sur le convexe fermé $\mathcal{M}_{\mathcal{I}}$ (respectivement $\mathcal{M}_{\mathcal{I}_\alpha}$) et qui vérifie les propriétés suivantes :

Proposition 1 *Avec les notations et hypothèses précédentes :*

1. *La suite (P_α^*) converge vers P^* quand α tend vers 0.*
2. *Soit un voisinage ouvert quelconque $\mathcal{V}(P^*)$ de P^* pour la topologie étroite. Si $(\rho_N)_N$ est dans $\mathcal{M}_{\mathcal{I}_\alpha}$, la probabilité pour que cette suite ne soit pas dans $\mathcal{V}(P^*)$ est exponentiellement petite en N :*

$$\forall \mathcal{V}(P^*)\ ,\ \exists \alpha_0 > 0\ ,\ \exists \nu > 0\ ,\ \exists N_0 :$$

$$\forall \alpha \leq \alpha_0\ ,\ \forall N \geq N_0\ ,\ \Pr(\rho_N \notin \mathcal{V}(P^*) \mid \rho_N \in \mathcal{M}_{\mathcal{I}\alpha}) \leq e^{-N\nu}.$$

La loi P^* est l'unique élément de $\mathcal{M}_{\mathcal{I}}$ qui possède ces propriétés, dont l'association est désignée par le terme de propriété de concentration*.*

Nous ne reviendrons pas ici sur la démonstration de cette proposition, basée sur des résultats de grandes déviations, dont on trouvera les détails dans [Jaynes 82, Robert 90, Robert 91].

La loi P^* qui maximise S a donc, si elle existe, la propriété de concentrer les lois empiriques satisfaisant "presque" les contraintes considérées. Dans l'hypothèse où ces contraintes décrivent au mieux le phénomène, les distributions empiriques seront le plus probablement une image de la loi maximisant S. De plus, S est l'*unique* fonctionnelle vérifiant cette propriété.

La propriété de concentration nous paraît donc la meilleure manière de définir et de justifier le principe de modélisation que nous allons à présent énoncer.

3.2 Énoncé d'un principe de maximum d'entropie "minimaliste"

S se trouve être la fonctionnelle désignée sous le terme d'**entropie de Shannon**. Nous avons privilégié dans ce texte une approche évitant l'introduction de la notion générale d'entropie statistique (*"information-theoretic entropy"*) et les interprétations qui peuvent lui être associées. S étant la seule fonctionnelle qui vérifie la propriété de concentration des lois empririques, il n'est pas nécessaire d'interpréter son expression pour justifier son utilisation. (Nous verrons succinctement au paragraphe suivant de quelle manière la notion générale d'entropie peut être abordée.)

> Nous appelerons **principe de maximum d'entropie (PME)** le principe de modélisation qui consiste à modéliser l'incertitude par la loi P^* de $\mathcal{M}(\mathcal{X})$ qui maximise l'entropie de Shannon S parmi les lois qui vérifient les contraintes définies à partir d'une information donnée $\mathcal{I} = \{(\varphi_m, a_m), m \in \{1, \ldots, M\}\}$.

▷ **Remarques**

Nous ne considérons a priori aucune caractéristique particulière de l'espace de travail $\mathcal{X}$. La mesure de référence sur cet espace est alors naturellement la mesure de comptage dans le cas discret et la mesure de Lebesgue dans le cas d'un sous-espace de $\mathbf{R}^J$. Il est possible de généraliser la proposition précédente dans le cas de mesures de référence autres que celles-ci ([Robert 90, Robert 91]). Ce n'est pas notre propos dans ce travail, où nous ne traduisons l'information sur le phénomène étudié qu'au travers des observables. Si une mesure de référence autre que la mesure de comptage ou la mesure de Lebesgue a été explicitement choisie sur l'ensemble $\mathcal{X}$, un principe équivalent peut être défini à partir d'une entropie prenant en compte cette mesure (entropie de Kullback).

La propriété de la loi de maximum d'entropie de concentrer les lois empiriques justifie d'une part l'emploi de la fonctionnelle S dans la recherche d'un modèle et fournit d'autre part une explication de l'efficacité pratique remarquable des modèles obtenus par application du PME. Ainsi, tout modèle classique redéfini comme modèle de maximum d'entropie trouve une nouvelle justification de son intérêt et de sa force. Mais cette justification se fait en associant à chaque modèle une famille donnée d'observables. Si le modèle ne se révèle pas adéquat pour décrire la réalité du phénomène, les modifications devront intervenir sur la nature de l'information considérée.

3.3 Autres approches du principe de maximum d'entropie

Il est possible de fournir à la fonctionnelle S une interprétation en terme *d'information moyenne.*

En effet, dans le cas d'un espace $\mathcal{X}$ discret, **l'information au sens de Shannon** apportée par une réalisation $\boldsymbol{x}$ de $\boldsymbol{X}$ est $\ln(1/p_x) = -\ln p_x$. L'information moyenne est donc donnée par $-\sum_x p_x \ln p_x$, qui est l'expression de S dans le cas discret. À la base de l'utilisation de cette fonctionnelle, nous trouvons l'idée que la réalisation d'un événement rare (en dehors de toute considération sur son sens ou ses conséquences) apporte plus d'information sur le phénomène que la réalisation d'un événement fréquent.

D'autres considérations intuitives peuvent s'ajouter à celle-ci et constituent un ensemble de propriétés auxquelles devra répondre une mesure de l'information. En particulier, l'information apportée par la réalisation d'événements indépendants se doit d'être la somme des informations apportées par chacun d'eux. D'autres propriétés peuvent s'associer à celle-ci. On citera entre autres : la continuité de la fonctionnelle sur l'ensemble des valeurs (p_x) ; le fait que la valeur maximale de la fonctionnelle augmente quand le nombre de réalisations possibles augmente ; le fait que le maximum soit atteint pour la loi uniforme ; la nullité de la fonctionnelle si une valeur possible est de réalisation certaine ; ... (cf. [Kapur 89] pour une liste plus détaillée).

La définition de l'information proposée par Shannon ([Shannon 48]) vérifie l'ensemble de ces propriétés. En fait, il est possible de montrer que toute fonctionnelle qui répond à ces notions est de la forme $-k \sum_x p_x \ln p_x$ ([Khinchin 57]).

Une des propriétés de S est qu'elle atteint son maximum quand toutes les valeurs possibles de $\boldsymbol{X}$ apportent la même information. Cela s'interprète de manière intuitive par le fait qu'en l'absence de toute connaissance, il paraîtrait arbitraire de privilégier certaines valeurs par rapport à d'autres au sens de l'information qu'apporteraient leurs réalisations. Ainsi, si on traduit la connaissance du phénomène par diverses contraintes (autres que la connaissance du nombre de réalisations possibles), la loi qui maximise S aura tendance à “répartir au mieux” l'information apportée par les différentes valeurs possibles de $\boldsymbol{X}$.

Parmi les lois satisfaisant à diverses contraintes traduisant la connaissance du phénomène, on décide alors de choisir pour modéliser l'incertitude sur $\mathcal{X}$ celle qui maximise l'information moyenne ([Shannon 48, Shannon & Weaver 57]).

Ces considérations se généralisent au cas où $\mathcal{X}$ est un espace continu. On considère alors l'information apportée par l'appartenance des réalisations de $\mathbf{X}$ à des pavés bornés de $\mathcal{X}$.

Le terme d'**entropie** a été utilisé pour désigner la fonctionnelle S, pour éviter d'utiliser le terme trop général d'information et par analogie avec le vocabulaire de la thermodynamique (cf. la remarque de Von Neumann proposant le terme d'entropie pour désigner la fonctionnelle S, en réponse à Shannon qui cherchait un autre terme que celui d'information : *"Firstly, you have got the same expression* $-\sum p_x \ln p_x$ *as is used for entropy in thermodynamics and secondly and more importantly, since even after one hundred years, nobody understands what entropy is, if you use the word entropy you will always win in an argument!"* [Tribus 79]).

Le principe de maximum d'entropie se retrouve donc par cette approche informative qui se décrit de manière plus intuitive.

Les propriétés mathématiques de la fonctionnelle S ont été largement étudiées pour justifier et interpréter son utilisation dans un but de modélisation statistique ([Khinchin 57, Aczel, Forte & Ng 74, Kapur & Kesavan 87]). D'autres fonctionnelles sur $\mathcal{M}(\mathcal{X})$ ont été définies par la suite. Elles sont toutes considérées comme des mesures de "l'incertitude" associée à chaque loi de probabilité sur $\mathcal{X}$. Le principe à la base de la recherche d'un modèle est toujours de choisir la loi qui vérifie certaines contraintes et qui maximise cette "incertitude". Cette loi est alors interprétée comme "la plus probable", "la moins biaisée", ... Toutes ces interprétations sont intuitives et n'appuient pas l'utilisation du PME sur des propriétés probabilistes du modèle obtenu.

En fait, il se peut en pratique que la maximisation d'autres fonctionnelles que S pour une même série de contraintes donne des résultats assez proches de l'application du PME tel que nous l'avons défini, si ces fonctionnelles sont basées sur des fonctions concaves admettant leur maximum au même point que S. Mais S est l'unique entropie qui permet d'atteindre une loi qui vérifie la propriété de concentration ([Ellis 85]). Ainsi, notre approche évite l'arbitraire du choix de la fonctionnelle "mesurant" l'information apportée par les réalisations de $\mathbf{X}$ et ne nécessite pas de justifier son utilisation par la théorie de l'information.

Signalons enfin que l'utilisation du PME est parfois suggérée dans le cadre de la statistique bayésienne pour déterminer la loi a priori du paramètre expérimentalement invariable ([Geyer & Thompson 92]). Dans cette optique particulière, la propriété de concentration ne peut pas être invoquée comme justification de l'utilisation du PME. Cette propriété suppose en effet que l'on pourra obtenir expérimentalement un jeu d'observations des variables étudiées, ce qui ne peut pas être le cas dans ce cadre bayésien. Une vision informative de ces lois a priori est alors nécessaire pour justifier leurs utilisations.

4 Détermination des lois de maximum d'entropie

Il s'agit à présent de résoudre le problème de la recherche d'une loi maximisant une fonctionnelle donnée et vérifiant certaines contraintes. On peut utiliser alors deux types d'approches :

- soit on choisit de maximiser le lagrangien associé à ces contraintes ([Kapur 89, Kapur & Kesavan 87, Golan *et al.* 96, Good 63, Seidenfeld 87]). Cette maximisation ne pose pas de problème dans le cas de variables discrètes, mais l'extension au cas continu nécessite des considérations théoriques supplémentaires ([Lisman & Van Zuylen 72, Gokhale 74]) qui sont en fait rarement détaillées ;
- soit, de manière plus élégante, on utilise des propriétés de la *I-divergence* de Kullback-Leibler ([Kullback 59]). La première démonstration du *minimum discrimination information theorem* qui précise la forme des lois minimisant la *I-divergence* pour une loi de référence donnée, apparaît dans [Kullback 59] (avec référence à [Kullback 54] et [Sanov 57]). Une nouvelle preuve de ce théorème a été présentée dans [Kullback & Kharait 65], puis reprise et étendue dans [Csiszár 75].

Nous présentons ici une démonstration qui reprend, avec d'autres notations, le schéma de cette preuve.

4.1 États de Gibbs et lois de maximum d'entropie

Nous considérons dans un premier temps un vecteur $\boldsymbol{X}$ constitué de variables de même type (toutes discrètes ou toutes réelles), et dans un deuxième temps un vecteur de type mixte.

4.1.1 Variables de même nature

On utilise les notations détaillées au 3.1.1, p. 19. On désigne de plus par f_P l'application qui, pour tout élément P de $\mathcal{M}(\mathcal{X})$, fait correspondre à un élément $\boldsymbol{x}$ de $\mathcal{X}$ sa probabilité $P(\mathbf{X} = \boldsymbol{x})$ (si $\mathcal{X}$ est un espace discret) ou la valeur de la densité de P en $\boldsymbol{x}$ (si $\mathcal{X}$ est un ouvert de $\mathbf{R}^J$).

On considère une information de taille M fondée sur les éléments de connaissance $(\varphi_m, a_m)_{m\in\{1,\ldots,M\}}$. On note de plus φ_0 la fonction constante égale à 1 et $a_0 = 1$. Pour simplifier les notations et démonstrations de ce paragraphe, on considère le couple (φ_0, a_0) comme un élément de connaissance supplémentaire dans $\mathcal{I}$. On note

$\mathcal{I}_0 = \{(\varphi_m, a_m), m \in \{0, \dots, M\}\}$. Toute loi vérifiant les contraintes imposées par l'information $\mathcal{I}_0$ est une loi de $\mathcal{M}_\mathcal{I}$. La recherche de la loi de maximum d'entropie se fait donc toujours dans $\mathcal{M}_\mathcal{I}$, que l'on intègre explicitement ou non la contrainte de normalisation dans l'information.

Proposition 2 *S'il existe un élément P^* dans $\mathcal{M}_\mathcal{I}$, l'ensemble des lois vérifiant les contraintes imposées par une information $\mathcal{I}$ de taille M, d'entropie finie, qui soit une loi de Gibbs, c'est à dire qui soit tel que :*

$$f_{P^*}(\boldsymbol{x}) = \exp\left(\sum_{m=0}^{M} \lambda_m \varphi_m(\boldsymbol{x})\right) \ , \ \boldsymbol{x} \in \mathcal{X} \ , \ (\lambda_0, \dots, \lambda_M) \in \mathbf{R}^{M+1} \ , \tag{I.1}$$

alors cette loi P^* est l'unique loi qui réalise le maximum de l'entropie de Shannon sur $\mathcal{M}_\mathcal{I}$:

$$S(P^*) = \max\{S(P), P \in \mathcal{M}_\mathcal{I}\}.$$

Démonstration

Pour toute loi P dans $\mathcal{M}_\mathcal{I}$, on peut décomposer $S(P)$ comme suit :

$$S(P) = -E_P\left(\ln\left(\frac{f_P}{f_{P^*}}\right)\right) - E_P(\ln f_{P^*}) \ .$$

Or, d'après la forme de I.1, on a :

$$\begin{aligned} E_P(\ln f_{P^*}) &= \sum_{m=0}^{M} \lambda_m E_P(\varphi_m) \\ &= \sum_{m=0}^{M} \lambda_m a_m. \end{aligned}$$

Il suffit donc pour maximiser $S(P)$ de maximiser l'expression $-E_P\left(\ln\left(f_P/f_{P^*}\right)\right)$, c'est-à-dire l'expression $E_P\left(\ln\left(f_P^*/f_P\right)\right)$.

En appliquant l'inégalité de Jensen à la fonction ln, on obtient l'inégalité suivante :

$$E_P\left(\ln\left(\frac{f_P^*}{f_P}\right)\right) \leq \ln\left(E_P\left(\frac{f_P^*}{f_P}\right)\right) .$$

Or, on a : $E_P(f_{P^*}/f_P) = 1$.

On obtient finalement l'inégalité :

$$E_P\left(\ln\left(\frac{f_P^*}{f_P}\right)\right) \leq 0.$$

Cette inégalité est stricte dans le cas où la fraction f_P/f_{P^*} n'est pas presque partout constante.

$E_P\left(\ln\left(f_P^*/f_P\right)\right)$ n'atteint donc son maximum que dans le cas où f_P/f_{P^*} est presque partout constante. Or, dans ce cas, le fait que cette espérance soit égale à 0 implique que $f_P/f_{P^*} = 1$ presque partout, et donc que $P = P^*$.

P^* est donc la *seule* loi qui maximise l'entropie S sur $\mathcal{M}_\mathcal{I}$.

□

Pour retrouver une écriture plus classique de la loi P^*, on note Z la constante de normalisation. Ainsi, on peut écrire I.1 sous la forme suivante :

$$f_{P^*}(\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{m=1}^{M}\lambda_m\varphi_m(\boldsymbol{x})\right), \tag{I.2}$$

où $Z = \exp(-\lambda_0)$. (Le couple (φ_0, a_0) n'est en effet pas spécifiquement un élément de connaissance.)

Ainsi, pour une information donnée, la loi de Gibbs définie à partir de la famille d'observables $(\varphi_m)_{m\in\{1,\ldots,M\}}$ et d'une famille de M paramètres $(\lambda_m)_{m\in\{1,\ldots,M\}}$ qui vérifient les contraintes associées à l'information est *l'unique loi maximisant l'entropie de Shannon sous ces contraintes.*

▷ **Remarque**

Nous garderons par la suite la terminologie de **multiplicateurs** pour les paramètres $(\lambda_1, \ldots, \lambda_M)$ de la loi de maximum d'entropie, bien que la démonstration présentée ici n'utilise pas la méthode des multiplicateurs de Lagrange.

4.1.2 Variables de natures différentes

Considérons à présent le cas d'un vecteur $\boldsymbol{X}$ composé de J' variables $\boldsymbol{T} = (T_1, \ldots, T_{J'})$, à valeurs dans un espace discret $\mathcal{T}$ et de J'' variables $\boldsymbol{Y} = (Y_1, \ldots, Y_{J''})$, à valeurs dans $\mathbf{R}^{J''}$: $\boldsymbol{X} = (\boldsymbol{T}, \boldsymbol{Y})$.

On désigne par P_T la loi marginale de $\boldsymbol{T}$, et on note $P_T(\boldsymbol{t}) = p_t$ et par Q_Y^t la loi de $\boldsymbol{Y}$ conditionnellement à $\boldsymbol{T} = \boldsymbol{t}$, $\boldsymbol{t} \in \mathcal{T}$ (cette loi ne se définit que dans le cas où p_t est différent de 0). Si elle existe, on note f_Y^t la densité de Q_Y^t.
On note de plus E_{P_T} l'espérance suivant la loi de T et $E_{f_Y^t}$ l'espérance suivant la loi de $\boldsymbol{Y}$ conditionnellement à $\boldsymbol{T} = \boldsymbol{t}$.

La loi conjointe de $\boldsymbol{X}$ peut donc se décomposer sous la forme $(P_T, (Q_Y^t)_{t\in\mathcal{T}})$.

Nous utiliserons ici non seulement des observables $(\varphi_1, \ldots, \varphi_M)$ portant indifféremment sur des variables des deux types, mais aussi une famille d'observables particulière permettant de rechercher la loi de maximum d'entropie à loi de $\boldsymbol{T}$ fixée :

$$\mathcal{I} = \{(\mathbf{1}_{\{t\}\times \mathbf{R}^{J''}}, p_t^*), \boldsymbol{t} \in \mathcal{T}\} \bigcup \{(\varphi_1, a_1), \ldots, (\varphi_M, a_M)\}.$$

Les contraintes correspondant aux observables indicatrices s'écrivent en effet :

$$E_P(\mathbf{1}_{\{t\}\times \mathbf{R}^{J''}}) = p_t = p_t^*,$$

pour toute valeur possible du vecteur aléatoire $\boldsymbol{T}$. La loi de ce vecteur est donc fixée par l'information choisie. Cette information permet de conserver la linéarité des contraintes sur la loi conjointe.

$\mathcal{M}_\mathcal{I}$ désigne toujours l'ensemble des lois de $\boldsymbol{X}$ qui vérifient les contraintes.

On cherchera donc la loi d'entropie maximum $P^*_{T,Y}$ à loi marginale P^*_T de $\mathbf{T}$ fixée. La propriété suivante décrit la forme des lois conditionnelles recherchées :

Proposition 3 *Avec les notations précédentes :*

S'il existe dans $\mathcal{M}_\mathcal{I}$ une loi $P^*_{T,Y}$ telle que la densité de la loi de $\mathbf{Y}$ conditionnellement à $\boldsymbol{T} = \boldsymbol{t}$ s'écrive, pour tout t dans $\mathcal{T}$:

$$f_Y^{*t}(\boldsymbol{y}) = \frac{1}{Z(\boldsymbol{t})} \exp\left(\sum_{m=1}^{M} \lambda_m \varphi_m(\boldsymbol{t}, \boldsymbol{y})\right), \tag{I.3}$$

alors la loi $P^*_{T,Y}$ est l'unique loi de maximum d'entropie pour l'information $\mathcal{I}$ considérée.

Démonstration

Par définition (cf. 3.1.1, p. 19), la décomposition de l'entropie S fait apparaître la loi marginale des variables discrètes et les lois conditionnelles des variables réelles :

$$S(P_{T,Y}) = E_{P_T}(S(Q_Y^T)) + S(P_T).$$

La loi de $\boldsymbol{T}$ est fixée par les niveaux de contraintes $(p_t^*)_{t\in\mathcal{T}}$. On recherche donc les densités conditionnelles $(f_Y^t)_{t\in\mathcal{T}}$ qui maximisent $E_{P_T^*}(S(f_Y^T))$ sur $\mathcal{M}_\mathcal{I}$:

$$E_{P_T^*}(S(f_Y^T)) = \max_{\mathcal{M}_\mathcal{I}} \left\{E_{P_T^*}(S(f_Y^T))\right\}.$$

Les contraintes imposées par les éléments de connaissance $(\varphi_1, a_1), \ldots, (\varphi_M, a_M)$ sont les suivantes :

$$\begin{aligned} a_m &= \sum_t p_t^* \int_{\mathbf{R}^{J''}} \varphi_m(\boldsymbol{t}, \boldsymbol{y}) f_Y^t(\boldsymbol{y}) d\boldsymbol{y} \\ &= \sum_t p_t^* E_{f_Y^t}(\varphi_m(\boldsymbol{t}, .)). \end{aligned}$$

Il faut ajouter à ces équations les contraintes de normalisation portant sur chaque loi conditionnelle :

$$\forall \boldsymbol{t} \in \mathcal{T} \ , \ \int_{\mathbf{R}^{J''}} f_Y^t(\boldsymbol{y}) d\boldsymbol{y} = 1 \ .$$

On reprend la décomposition utilisée dans la démonstration précédente :

$$\begin{aligned}
E_{P_T}(S(f_Y^T)) &= \sum_t p_t^* S(f_Y^t) \\
&= \sum_t p_t^* \left(-E_{f_Y^t}(\ln \frac{f_Y^t}{f_Y^{*t}}) - E_{f_Y^t}(\ln f_Y^{*t}) \right) \\
&= -\sum_t p_t^* \left(E_{f_Y^t} \left(\ln \frac{f_Y^t}{f_Y^{*t}} \right) \right) - \sum_t p_t^* \left(\sum_m \lambda_m E_{f_Y^t}(\varphi_m(\boldsymbol{t}, .)) \right) \\
&\quad + \sum_t p_t^* \ln Z(\boldsymbol{t}) \\
&= -\sum_t p_t^* \left(E_{f_Y^t} \left(\ln \frac{f_Y^t}{f_Y^{*t}} \right) \right) - \sum_m \lambda_m \left(\sum_t p_t^* E_{f_Y^t}(\varphi_m(\boldsymbol{t}, .)) \right) \\
&\quad + \sum_t p_t^* \ln Z(\boldsymbol{t}) \\
&= -\sum_t p_t^* \left(E_{f_Y^t} \left(\ln \frac{f_Y^t}{f_Y^{*t}} \right) \right) - \sum_m \lambda_m a_m + \sum_t p_t^* \ln Z(\boldsymbol{t})
\end{aligned}$$

Maximiser $E_{P_T}(S(f_Y^T))$ revient donc à minimiser $\sum_t p_t^* \left(E_{f_Y^t} \left(\ln \frac{f_Y^t}{f_Y^{*t}} \right) \right)$.

On utilise à nouveau l'inégalité de Jensen. Les niveaux de contraintes p_t^* sont des réels positifs, on obtient donc finalement l'inégalité suivante :

$$\sum_t p_t^* E_{f_Y^{*t}} \left(\frac{f_Y^t}{f_Y^{*t}} \ln \left(\frac{f_Y^t}{f_Y^{*t}} \right) \right) \geq 0.$$

Le minimum de cette somme est donc atteint quand chacune des espérances y apparaîssant est nulle.

Cela implique alors que chacune des expressions $(f_Y^t / f_Y^{*t}) \ln(f_Y^t / f_Y^{*t})$ est nulle presque partout, c'est-à-dire $f_Y^t / f_Y^{*t} = 1$ presque partout.

Les lois conditionnelles $(Q_Y^t)_{t \in \mathcal{T}}$ qui permettent d'atteindre le minimum de l'entropie S (en les associant à la loi marginale de $\boldsymbol{T}$ imposée par l'information) sont donc bien les lois de densité f_Y^{*t}.

□

▷ **Remarque**

Par analogie avec le cas précédent, on désignera par **états de Gibbs** les lois conditionnelles de densité f_Y^{*t}, $\boldsymbol{t} \in \mathcal{T}$. Les multiplicateurs obtenus par application du PME sont ici les paramètres $(\lambda_1, \ldots, \lambda_M)$.

4.2 Équations du PME

Nous désignerons comme **équations du PME** les équations provenant de l'écriture des contraintes imposées par $\mathcal{I}$ aux paramètres de la loi de Gibbs correspondant aux observables $(\varphi_m)_{m\in\{1,\ldots,M\}}$:

$$E_{P^*}(\varphi_m) = a_m \ , \ m \in \{1,\ldots,M\},$$

où $f_{P^*}(\boldsymbol{x}) = 1/Z \exp\left(\sum_{m=1}^{M} \lambda_m \varphi_m(\boldsymbol{x})\right)$.

Les équations du PME sont donc :

$$\int_{\mathcal{X}} \frac{1}{Z} \varphi_j(\boldsymbol{x}) \exp\left(\sum_{m=1}^{M} \lambda_m \varphi_m(\boldsymbol{x})\right) d\boldsymbol{x} = a_j \ , \ j \in \{1,\ldots,M\}, \tag{I.4}$$

où $Z = \int_{\mathcal{X}} \exp(\sum_m \lambda_m \varphi_m(\boldsymbol{x}))d\boldsymbol{x}$. (Le signe $\int$ désigne aussi l'intégrale par rapport à la mesure de comptage dans le cas fini ou dénombrable.)

Pour une information de taille M, il y a donc toujours M équations du PME à résoudre pour obtenir les multiplicateurs $(\lambda_1,\ldots,\lambda_M)$ à partir des niveaux de contraintes $(a_1,\ldots,a_M)$, quelle que soit la nature de ces niveaux. Ainsi, si tous les niveaux de contraintes sont connus a priori, il n'y a pas besoin d'utiliser de jeux de données pour obtenir les valeurs des paramètres du modèle. Si tout ou partie des niveaux de contraintes ne sont pas connus a priori, il sera alors nécessaire de disposer d'un jeu de données pour estimer ces niveaux et résoudre le système de M équations du PME.

▷ **Exemple**

Considérons le cas d'une unique variable X sur $\mathbf{R}$ ($J = 1$) et d'une information fondée sur les deux premiers moments : $\mathcal{I} = \{(x, a_1), (x^2, a_2), a_1 \in \mathbf{R}, a_2 \in \mathbf{R}^+\}$. Nous pouvons montrer qu'il existe un état de Gibbs de la forme suivante :

$$f_{P^*}(x) = \frac{1}{Z} \exp(\lambda_1 x + \lambda_2 x^2) \ , \ x \in \mathbf{R},$$

qui est la loi d'entropie maximum pour certaines valeurs des paramètres (λ_1, λ_2).

Les équations du PME s'écrivent en effet :

$$\left\{ \begin{array}{lcl} a_1 & = & \dfrac{1}{Z} \displaystyle\int_x x \exp(\lambda_1 x + \lambda_2 x^2) dx \\ a_2 & = & \dfrac{1}{Z} \displaystyle\int_x x^2 \exp(\lambda_1 x + \lambda_2 x^2) dx \end{array} \right.$$

Il n'y aura donc un état de Gibbs satisfaisant aux contraintes qu'à la condition que les niveaux de contraintes vérifient : $a_2 \geq a_1^2$.

Or, on sait que toute loi dont la densité est proportionnelle à $\exp(\lambda_1 x + \lambda_2 x^2)$ est une loi de Gauss de moyenne a_1 et de variance $(a_2 - a_1^2)$.

Par identification, on obtient l'expression des multiplicateurs et de la constante de normalisation en fonction des niveaux de contraintes :

$$\lambda_1 = \frac{a_1}{a_2 - a_1^2}\ ,\ \lambda_2 = -\frac{1}{2(a_2 - a_1^2)}\ ,\ Z = \exp\left(\frac{a_1^2}{2(a_2 - a_1^2)}\right)\sqrt{2\pi}\sqrt{a_2 - a_1^2}.$$

Dans ce cas simple, si la condition d'existence est vérifiée, les multiplicateurs sont donc des fonctions explicites des niveaux de contraintes (ce qui n'est pas le cas en général).

Nous considérons que la connaissance de l'espérance et du moment d'ordre deux (ou de la variance) est ici la connaissance minimale dont on doit disposer pour choisir un modèle. Ainsi, pour une telle information, on considèrera donc que X suit une loi normale $\mathcal{N}(a_1, a_2 - a_1^2)$. En ce sens, le modèle gaussien peut être considéré comme un modèle *parcimonieux*, puisqu'il correspond à une information mettant en jeu une famille d'observables minimale.

▷ **Remarques**

Une information fondée uniquement sur l'observable x ne permet pas d'appliquer la proposition 2 précédente. En effet, il n'y a pas dans ce cas d'état de Gibbs dans $\mathcal{M}_\mathcal{I}$ car une fonction proportionnelle à $\exp(\lambda x)$ n'est pas intégrable.

On peut en fait montrer aisément qu'il n'existe pas de loi maximisant S dans ce cas. En effet, l'entropie d'une loi de Gauss dans $\mathcal{M}_\mathcal{I}$ de moyenne a et de variance σ^2 est : $S(P) = \ln\sigma + \ln\sqrt{2\pi} + 1/2$. On peut donc trouver dans $\mathcal{M}_\mathcal{I}$ des lois qui possèdent une entropie aussi grande que l'on veut en choisissant σ suffisamment grand.

Par contre, si la variable X est à valeurs dans $\mathbf{R}^+$, la loi issue du PME pour l'information $\mathcal{I} = \{(x, a), a \in \mathbf{R}\}$ est alors la loi exponentielle de densité : $f(x) = 1/a \exp(-x/a)$.

Ce dernier exemple illustre l'importance du choix de l'espace de travail $\mathcal{X}$. L'existence d'un état de Gibbs satisfaisant les contraintes dépend pour une grande part de la nature de cet espace.

Ainsi, la nature des variables impose souvent comme espace de travail l'ensemble $\mathbf{R}^+$ (taille, poids, taux de cholestérol, ...). Si le seul observable x ne constitue finalement pas une information suffisamment étendue pour décrire ce type de phénomène, on introduit l'observable x^2 dans l'information.

Il est alors possible de montrer qu'un état de Gibbs vérifiant les contraintes $E(X) = a_1$ et $E(X^2) = a_2$ n'existe sur $\mathbf{R}^+$ qu'à la condition que : $a_1^2 < a_2 < 2a_1^2$ ([Dowson & Wragg 73]). La loi est alors de densité $f(x) = 1/Z \exp(\lambda_1 x + \lambda_2 x^2)$, où la constante de normalisation Z s'exprime en fonction des multiplicateurs, de la densité de la loi normale centrée-réduite et de sa fonction de répartition.

Mais, dans de très nombreux cas, la loi ainsi obtenue sera très voisine d'une loi de Gauss sur $\mathbf{R}$. Le maniement du modèle gaussien étant particulièrement simple et aisé, on choisira alors de revenir à $\mathbf{R}$ comme espace de travail. On pourra agir ainsi dès que, par exemple, les niveaux de contraintes vérifient $a_1 - k\sqrt{a_2 - a_1^2} \geq 0$, où k est un entier de l'ordre de 3 à 6 (ce qui correspond à une loi normale $\mathcal{N}(a_1, a_2 - a_1^2)$ chargeant très peu $\mathbf{R}^-$). La condition d'existence précédente est alors vérifiée pour tout entier k supérieur à 1.

Les équations du PME découlent donc automatiquement de l'application du principe. Il n'est pas nécessaire de considérer un principe d'estimation particulier pour obtenir les paramètres $(\lambda_1, \ldots, \lambda_M)$ d'une loi de maximum d'entropie : leurs valeurs proviennent de la résolution d'un système comprenant autant d'équations que de multiplicateurs.

Les expressions des multiplicateurs en fonction des niveaux de contraintes sont explicites dans certains cas (cf. paragraphe 6, p. 46). On désignera alors par le terme d'**estimateurs du maximum d'entropie** (EME) les valeurs des multiplicateurs imposées par les équations du PME, bien qu'il s'agisse de résolution d'équations plutôt que d'estimation proprement dite.

L'expression de la loi d'entropie maximum ne dépend des niveaux de contraintes qu'au travers des équations permettant de déterminer les valeurs des multiplicateurs. Nous verrons par la suite comment traiter les niveaux de contraintes qui apparaissent dans ces équations.

Notons dès à présent que les lois issues du PME admettent au moins deux paramètrages possibles :

- d'une part, par les **multiplicateurs** $(\lambda_1, \ldots, \lambda_M)$, qui sont les paramètres de la loi de maximum d'entropie au sens classique, puisqu'ils apparaîssent directement dans son expression.

- d'autre part, par les **niveaux de contraintes** $(a_1, \ldots, a_M)$, qui interviennent directement dans l'écriture des contraintes et indirectement dans celle de la loi de maximum d'entropie.

4.3 Un cas particulier de recherche des lois de PME : le cas continu

La propriété 2 que nous avons présentée au paragraphe précédent nous permet d'affirmer que, s'il existe un état de Gibbs dans $\mathcal{M}_{\mathcal{I}}$, il représente l'unique solution au problème de maximisation sous contraintes imposé par l'information prise en compte.

Dans le cas de variables discrètes, la structure exponentielle de la loi de maximum d'entropie peut se comprendre comme une conséquence de l'application de la méthode des multiplicateurs de Lagrange. Pour comprendre l'origine de la forme exponentielle de la loi de maximum d'entropie dans le cas de variables continues, nous allons envisager le cas d'un espace $\mathcal{X}$ pour lequel on peut démontrer assez facilement que, s'il existe une loi de maximum d'entropie, il s'agit bien d'un état de Gibbs.

Nous nous plaçons à présent dans le cas où $\mathcal{X}$ est un *ouvert borné de* $\mathbf{R}^J$. $\mathcal{X}$ a donc la propriété d'être un espace topologique polonais (métrisable, séparable, complet).

On note toujours $\mathcal{M}(\mathcal{X})$ l'espace des lois de probabilités sur $\mathcal{X}$, muni de la topologie étroite (associée aux fonctions continues bornées). $\mathcal{M}(\mathcal{X})$ est alors, lui aussi, un espace polonais.

Les contraintes considérées ici sont basées sur M observables qui sont des *fonctions continues bornées* $(\varphi_m)_{m\in\{1,\dots,M\}}$ sur $\mathcal{X}$. Les niveaux de contraintes associés sont donnés par les réels $(a_1,\dots,a_M)$. On note toujours $\mathcal{M}_\mathcal{I}$ l'ensemble des lois répondant aux contraintes (y compris la contrainte de normalisation correspondant à $\varphi_0 \equiv 1$ et $a_0 = 1$).

On note I la **fonction d'information** définie sur $\mathcal{M}(\mathcal{X})$ par :

$$I(P) = -S(P)\,,\ P \in \mathcal{M}(\mathcal{X})\,.$$

On notera indifféremment $I(P)$ ou $I(f_P)$ l'information associée à la loi P, si P admet une densité f_P par rapport à μ la mesure de Lebesgue ($dP = f_P dx$). La nature du support $\mathcal{X}$ imposée ici assure que I est égale, à une constante près, à l'information de Kullback avec la loi uniforme sur $\mathcal{X}$ comme mesure de référence.

4.3.1 Existence

Proposition 4 *Avec les notations précédentes :*

Si $\mathcal{M}_\mathcal{I}$ est non-vide, alors il existe une unique loi P^* qui réalise le minimum de I :

$$\textit{Si } \mathcal{M}_\mathcal{I} \neq \emptyset\ ,\ \textit{alors } \exists\, P^*, I(P^*) = \min\{I(P), P \in \mathcal{M}_\mathcal{I}\}\,.$$

Démonstration

$\mathcal{X}$ est un espace polonais. On sait que pour ce type d'espace [Azencott 80], la fonction I possède entre autres propriétés d'être :

– positive

– nulle en $P = \mu$ uniquement

– convexe sur $\mathcal{M}(\mathcal{X})$

– s.c.i sur $\mathcal{M}(\mathcal{X})$

– inf-compact sur $\mathcal{M}(\mathcal{X})$ (i.e pour tout $\lambda \in \mathbf{R}$, $\{P,\ I(P) \leq \lambda\}$ est compact)

Or, $\mathcal{M}_{\mathcal{I}}$ est l'image réciproque d'un point de $\mathbf{R}^{M+1}$ dans $\mathcal{M}(\mathcal{X})$, qui est muni de la topologie étroite. $\mathcal{M}_{\mathcal{I}}$ est donc un fermé de $\mathcal{M}(\mathcal{X})$.

Si I est infinie pour tout élément de $\mathcal{M}_{\mathcal{I}}$, une mesure quelconque répond au problème. Sinon, on note P_0 une des mesures dans $\mathcal{M}_{\mathcal{I}}$ telle que $I(P_0) < +\infty$.

L'intersection de $\mathcal{M}_{\mathcal{I}}$ (fermé) et du sous-ensemble $\{P \in \mathcal{M}_{\mathcal{I}}\ ,\ I(P) \leq I(P_0)\}$ (compact de $\mathcal{M}_{\mathcal{I}}$) est compact.

Une fonction s.c.i admettant un minimum sur un compact, il existe donc un élément de $\mathcal{M}_{\mathcal{I}}$ qui réalise le minimum de la fonction d'information, c'est-à-dire qui maximise l'entropie S. Cet élément est unique car I est strictement convexe sur l'ensemble où elle prend des valeurs finies.

□

4.3.2 Expression

Nous allons à présent démontrer que P^* est un état de Gibbs.

Lemme 1 *Soient $(u_i)_{i\in\{1,\dots,n\}}$, n formes linéaires sur un espace vectoriel F. Si une forme linéaire u s'annule sur $\bigcap_1^n \ker(u_i)$, alors u est combinaison linéaire des $(u_i)_{i\in\{1,\dots,n\}}$.*

Démonstration

Ce résultat s'obtient grâce à une démonstration par récurrence.

– Si $n = 1$: le lemme est un résultat classique. Soit u_1 et u_2 deux formes linéaires telles que $\ker u_1 \subset \ker u_2$.

  – Si $\ker u_1 \neq \ker u_2$: dans ce cas, u_2 est identiquement nulle et $u_2 = 0u_1$.

  – Si $\ker u_1 = \ker u_2$: si $\ker u_1 = F$, alors $u_1 = u_2 \equiv 0$. Sinon, soit F' un supplémentaire de $\ker u_1$ dans F : $F = F' \oplus \ker u_1 = F' \oplus \ker u_2$. Les restrictions de u_1 et u_2 à F' sont des bijections linéaires de F' dans $\mathbf{R}$. Elles sont donc proportionnelles. On a finalement : $u_2 = \alpha u_2$.

– Hypothèse de récurrence : le lemme est vrai jusqu'au rang n-1.

- Pour le rang n :
  - Premier cas : si les $(u_i)_{i\in\{1,\ldots,n\}}$ sont dépendants, on peut trouver une famille de réels $(\alpha_i)_{i\in\{1,\ldots,n\}}$ telle que $u_n = \sum_{i=1}^{n-1} \alpha_i u_i$ (quitte à renuméroter les formes linéaires).

    On a alors $\bigcap_1^{n-1} \ker(u_i) \subseteq \ker(u_n)$. Soit $\bigcap_1^{n-1} \ker(u_i) = \bigcap_1^{n} \ker(u_i)$.

    Par application de l'hypothèse, u est combinaison linéaire des $(u_i)_{i\in\{1,\ldots,n-1\}}$.
  - Deuxième cas : les $(u_i)_{i\in\{1,\ldots,n\}}$ sont linéairement indépendants.

    Alors : $\forall\ k \in \{1,\ldots,n\}$, $\exists\ e_k \in \bigcap_{i\neq k} \ker(u_i)$ et $u_k(e_k) \neq 0$. (Autrement, pour k donné, u_k s'annulerait sur $\bigcap_{i\neq k} \ker(u_i)$ et par application de l'hypothèse de récurrence, serait combinaison linéaire des $n-1$ formes restantes). Soit $\overline{e}_k = e_k/u_k(e_k)$.

    On peut exprimer x quelconque dans F sous la forme :

$$x = u_1(x)\overline{e}_1 + \ldots + u_n(x)\overline{e}_n + \xi_x.$$

    On a donc : $u_k(x) = u_k(x)u_k(\overline{e}_k) + u_k(\xi_x) \Rightarrow u_k(\xi_x) = 0, \forall k \in \{1,\ldots,n\}$, $\forall x \in F$. Ainsi, ξ_x appartient à $\bigcap_1^n \ker(u_i)$ et par hypothèse on a : $u(\xi_x) = 0, \forall x \in F$. Finalement, on a : $u(x) = \sum_1^n u(\overline{e}_i)u_i(x) + u(\xi_x) = \sum_1^n \alpha_i u_i(x)$, $\forall x \in F$.

$\square$

Proposition 5 *Avec les notations précédentes :*

Si $\mathcal{M}_{\mathcal{I}}$ est non-vide et si la loi P^* qui réalise le minimum de I admet une densité f_{P^*} telle que $\inf f_{P^*} > 0$ *et* $\sup f_{P^*} < +\infty$, *alors il s'agit d'un état de Gibbs.*

Démonstration

On note $\mathcal{D}(\mathcal{X})$, l'ensemble des fonctions C^∞ sur $\mathcal{X}$.

On considère Γ_{M+1} le sous-espace de $\mathcal{D}'(\mathcal{X})$, ensemble des distributions sur $\mathcal{D}(\mathcal{X})$, engendré par les distributions $< \varphi_1, . >, \ldots, < \varphi_M, . >$ et par $< \varphi_0, . >$, où φ_0 est la fonction constante égale à 1. On note de plus K_{M+1} l'intersection des noyaux des distributions $(< \varphi_m, . >)_{m\in\{0,\ldots,M\}}$.

Remarquons en premier lieu que K_{M+1} ne peut pas être réduit à 0. En effet, dans le cas contraire, on aurait $\Gamma_{M+1}^\perp = \{0\}$. On pourrait donc écrire $(\Gamma_{M+1}^\perp)^\perp$ des deux manières suivantes :

$$\begin{aligned} (\Gamma_{M+1}^\perp)^\perp &= \{0\}^\perp = \mathcal{D}'(\mathcal{X}) \text{ de dimension infinie}\,, \\ \text{et } (\Gamma_{M+1}^\perp)^\perp &= \Gamma_{M+1} \text{ de dimension finie.} \end{aligned}$$

Cela assure l'existence d'au moins une fonction γ, C^∞ sur $\mathcal{X}$, telle que :

$$\int_{\mathcal{X}} \varphi_m(\boldsymbol{x})\gamma(\boldsymbol{x})d\boldsymbol{x} = 0 \ , \ m \in \{0, \ldots, M\}.$$

Toute fonction γ vérifie en particulier $\int_{\mathcal{X}} \gamma(\boldsymbol{x})d\boldsymbol{x} = 0$. Ainsi, toute fonction de la forme $f_{P^*} + t\gamma$, $t \in \mathbf{R}$, peut être considérée comme une densité de probabilité.

Nous allons montrer à présent que la distribution $< f_{P^*}, . >$ s'annule sur K_{M+1}.

On définit la fonction ψ de $[-\alpha, \alpha]$ dans $\mathbf{R}$, qui à t associe $I(f_{P^*} + t\gamma)$, où $\alpha > 0$ est tel que :

$$\forall t \text{ avec } |t| \leq \alpha \ , \ \forall \ \boldsymbol{x} \in \mathcal{X} \ , \ f_{P^*}(\boldsymbol{x}) + t\gamma(\boldsymbol{x}) > 0 \ .$$

(L'existence d'un tel réel α est assuré par le fait que $\inf f_{P^*} > 0$.)

La fonction ψ est alors dérivable pour tout $|t| \leq \alpha$. En utilisant le théorème de dérivation sous le signe intégrale, on obtient :

$$\psi'(t) = \int_{\mathcal{X}} [1 + \ln(f_{P^*}(\boldsymbol{x}) + t\gamma(\boldsymbol{x}))]\gamma(\boldsymbol{x})d\boldsymbol{x} \ ,$$

ce qui donne, pour $t = 0$:

$$\psi'(0) = \int_{\mathcal{X}} [1 + \ln f_{P^*}(\boldsymbol{x})]\gamma(\boldsymbol{x})d\boldsymbol{x} \ .$$

On sait que :

$$\forall m \in \{0, \ldots, M\}, \ \left\{ \begin{array}{lcl} \displaystyle\int_{\mathcal{X}} \varphi_m(\boldsymbol{x})\gamma(\boldsymbol{x})d\boldsymbol{x} & = & 0 \\ \displaystyle\int_{\mathcal{X}} \varphi_m(\boldsymbol{x})f_{P^*}(\boldsymbol{x})d\boldsymbol{x} & = & a_m \ , \ (a_0 = 1) \end{array} \right.$$

Donc $(f_{P^*} + t\gamma)$ vérifie bien les M contraintes correspondant aux observables $(\varphi_1, \ldots, \varphi_M)$ ainsi que la contrainte de normalité.

Or, par hypothèse, la fonction I atteint son minimum en f_{P^*}. La fonction ψ doit donc être minimum en $t = 0$. Ainsi, on en conclut que $\psi'(0) = 0$.

Finalement, $< 1 + \ln f_{P^*}, . >$ s'annule donc pour toute fonction γ qui appartient à K_{M+1}.

D'après le lemme 1, il existe donc une famille de $M + 1$ réels $(\mu_m)_{m \in \{0, \ldots, M\}}$ tels que :

$$< 1 + \ln f_{P^*}, . >= \sum_{m=0}^{M} \mu_m < \varphi_m, . >,$$

soit :

$$f_{P^*}(\boldsymbol{x}) = \exp\left(\sum_{m=0}^{M} \lambda_m \varphi_m(\boldsymbol{x})\right) \ , \ \forall \boldsymbol{x} \in \mathcal{X},$$

où $\lambda_0 = (\mu_0 - 1)$ et $\lambda_m = \mu_m$, $m \in \{1, \ldots, M\}$.

La densité de la loi P^* définit donc bien un état de Gibbs sur $\mathcal{X}$.

$\square$

5 Principe de maximum d'entropie, méthode du maximum de vraisemblance et méthode des moments

L'utilisation du PME dans la recherche d'un modèle statistique se caractérise par le fait qu'il nécessite de spécifier en premier lieu, non pas la structure statistique, mais une famille de fonctions des variables mises en jeu. Nous avons vu que ces fonctions représentent la connaissance que l'on a du phénomène au travers de contraintes alors imposées sur le modèle.

Les équations issues de ces contraintes mettent en jeu une première famille de paramètres : les *niveaux de contraintes*. Dans l'idéal, ces niveaux sont tous connus a priori. Si ce n'est pas le cas, il s'agit de fournir des approximations des niveaux inconnus. Nous proposons ici de prendre comme niveaux estimés les *moyennes empiriques des observables correspondants*. Une autre possibilité serait d'utiliser des estimateurs robustes de ces niveaux.

Les lois obtenues par application du PME sont de type exponentiel (s'il existe un état de Gibbs dans $\mathcal{M}_{\mathcal{I}}$) et mettent en jeu une deuxième famille de paramètres, en même nombre que la taille de l'information : les *multiplicateurs*. Les multiplicateurs sont solutions des équations du PME et dépendent des niveaux de contraintes (a priori ou estimés) au travers de ces équations (cf. 4.2).

5.1 Traitement des deux familles de paramètres

Nous nous retrouvons donc confrontés au problème de l'estimation des paramètres d'une loi de probabilité. Cette estimation concerne les multiplicateurs, mais leur détermination dans le cadre de l'application du PME dépend de la connaissance des niveaux de contraintes.

L'approche habituelle d'un état de Gibbs atteint par le PME consisterait à utiliser un principe d'estimation pour obtenir des valeurs des multiplicateurs à partir d'un jeu de données.

L'application de la **méthode du maximum de vraisemblance (MMV)** ([Gourieroux & Monfort 89]) à une loi de même structure que la loi de maximum d'entropie, en supposant que les données disponibles constituent un échantillon de cette loi, permet d'obtenir les équations auxquelles doivent répondre les estimations des paramètres de cette loi. Nous montrons dans ce paragraphe que ces équations sont les mêmes que celles du PME, dans le cas où l'information utilisée ne met en jeu que des niveaux de contraintes tous estimés, à partir du même échantillon, par les moyennes empiriques des observables.

5.1.1 Niveaux de contraintes et méthode des moments

Les niveaux de contraintes apparaissant dans l'information $\mathcal{I}$ peuvent être soit fournis de manière théorique, soit estimés subjectivement, soit estimés empiriquement.

Le premier cas correspond à la traduction de réalités physiques, biologiques, ... (une contrainte de symétrie d'une variable X par rapport à l'origine se traduit, en respectant le principe de parcimonie, par l'élément de connaissance $(x, 0)$). Le deuxième cas correspond à la traduction de la connaissance subjective du phénomène par un expert (un médecin donnera par exemple une estimation subjective de la probabilité d'apparition d'un symptôme comme un résumé des nombreux cas qu'il a étudiés, bien qu'aucune base de données n'ait été établie). Le troisième cas correspond à l'utilisation d'un jeu de réalisations du phénomène en estimant les niveaux de contraintes par les moyennes empiriques des observables leur correspondant.

La donnée de contraintes théoriques étant rare et la validité des contraintes subjectives étant discutable, la situation la plus fréquente est celle où les niveaux de contraintes sont *estimés sur un jeu de données* ([Choong *et al.* 96, Berger *et al.* 96]). (Nous reviendrons sur l'utilisation de niveaux théoriques pour un cas particulier au paragraphe 6, p. 46.)

Ainsi, à partir de N réalisations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ du vecteur aléatoire $\mathbf{X}$, on note $\overline{\varphi}_m$ le niveau de contrainte de l'observable φ_m “estimé” par la moyenne empirique de cet observable :

$$\overline{\varphi}_m = \frac{1}{N} \sum_{1}^{N} \varphi_m(\boldsymbol{x}_i) \ , \ m \in \{1, \ldots, M\}.$$

On notera $\overline{\mathcal{I}}$ l'information fondée sur les mêmes observables que $\mathcal{I}$ mais qui sont *tous* associés à des niveaux de contraintes empiriques :

$$\overline{\mathcal{I}} = \{(\varphi_m, \overline{\varphi}_m), m \in \{1, \ldots, M\}\}. \tag{I.5}$$

L'emploi de cette information, que nous désignerons sous le terme d'**information empirique**, se justifie par la première partie de la propriété de concentration. Prenons le cas d'échantillons simulés à partir de la loi de maximum d'entropie correspondant à une information donnée $\mathcal{I}$: cette propriété assure que la loi de maximum d'entropie obtenue à partir de l'information empirique correspondante $\overline{\mathcal{I}}$ sera très proche de la loi dont on est parti pour obtenir l'échantillon.

Dans le cas particulier où les observables sont définis comme des *polynômes des variables mises en jeu*, les équations du PME aboutissent donc aux mêmes résultats que l'utilisation de la **méthode des moments** pour une loi de même expression que l'état de Gibbs obtenu (s'il existe). Les équations du PME consistent en effet dans ce cas à écrire l'égalité entre certains moments théoriques et empiriques, puisqu'elles ont la forme suivante :

$$E_{P^*}(\varphi_m) = \overline{\varphi}_m \ , \ m \in \{1, \ldots, M\}.$$

Rappelons que la méthode des moments consiste à déterminer les valeurs des m paramètres d'une structure statistique donnée en égalisant les m premiers moments théoriques aux moments empiriques. Cela revient à résoudre un système comprenant autant d'équations qu'il y a de paramètres dans le modèle. Mais dans le cadre proposé par Pearson, le lien entre un principe qui aurait conduit à spécifier la structure statistique et cette estimation des paramètres n'est pas explicité.

$\triangleright$ **Remarque**

Dans le cas d'un vecteur aléatoire composé de variables discrètes $\boldsymbol{T}$ et de variables réelles $\boldsymbol{Y}$, nous avons utilisé dans la proposition 3 une information particulière qui fixe la loi du vecteur $\boldsymbol{T}$. L'information empirique impose alors de rechercher la loi de maximum d'entropie à loi marginale de $\boldsymbol{T}$ fixée par la loi *empirique* de ce vecteur.

5.1.2 Multiplicateurs et estimateurs du maximum de vraisemblance

Nous allons voir à présent à quel mode de calcul des paramètres de la loi du PME conduit l'utilisation systématique de niveaux de contraintes empiriques.

Nous supposerons ici que l'information $\mathcal{I}$ considérée permet d'obtenir une loi maximisant l'entropie.

Seul un changement des observables $(\varphi_m)_{m\in\{1,\ldots,M\}}$ peut modifier la forme de la loi P^* issue de l'application du PME. Il est donc équivalent d'utiliser l'information $\mathcal{I}$ ou l'information $\overline{\mathcal{I}}$ en ce qui concerne la *forme de la loi obtenue*, qui restera un état de Gibbs. Mais cette loi dépend des niveaux de contraintes au travers des multiplicateurs.

On notera $\mathcal{S}_M$ la structure statistique $\{\mathcal{X}, \mathcal{B}(\mathcal{X}), P_\nu, \nu \in \Theta\}$, Θ étant un ouvert de $\mathbf{R}^M$, où la densité d'une loi P_ν est donnée par :

$$f_{P_\nu}(\boldsymbol{x}) = \frac{1}{W} \exp\left(\sum_{m=1}^{M} \nu_m \varphi_m(\boldsymbol{x})\right) \ , \ \boldsymbol{x} \in \mathcal{X}. \tag{I.6}$$

L'utilisation exclusive des niveaux empiriques permet de lier les équations obtenues par l'application du PME et celles issues de la MMV dans l'étude d'une telle loi P_ν comme le montre la proposition suivante :

Proposition 6 *Les équations du PME associées à l'information* $\overline{\mathcal{I}} = \{(\varphi_m, \overline{\varphi}_m), m \in \{1, \ldots, M\}\}$ *sont les équations du maximum de vraisemblance pour le vecteur des paramètres* $\nu = (\nu_1, \ldots, \nu_M)$ *d'une loi* P_ν *de la structure statistique* $\mathcal{S}_M$*, en considérant* $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *comme un échantillon de* P_ν*.*

Démonstration

Les équations du PME associées à $\overline{\mathcal{I}}$ s'écrivent de la manière suivante :

$$E_{P^*}(\varphi_m) = \frac{1}{N}\sum_{i=1}^{N}\varphi_m(\boldsymbol{x}_i)\ ,\ m \in \{1,\ldots,M\}. \tag{I.7}$$

On ajoute à ces M équations celle qui correspond à la constante de normalisation Z :

$$Z = \int_{\mathcal{X}} \exp\left(\sum_{1}^{M}\lambda_m\varphi_m(\boldsymbol{x})\right) d\boldsymbol{x}\ .$$

(Le signe $\int$ désigne aussi l'intégrale par rapport à la mesure de comptage dans le cas discret fini ou dénombrable.)

Considérons à présent l'estimation par maximum de vraisemblance des paramètres de la densité d'une loi P_ν issue de la structure statistique $\mathcal{S}_M$:

$$f_{P_\nu}(\boldsymbol{x}) = \frac{1}{W}\exp\left(\sum_{m=1}^{M}\nu_m\varphi_m(\boldsymbol{x})\right)\ ,\ \boldsymbol{x} \in \mathcal{X}\ .$$

La constante de normalisation étant fonction des paramètres $\nu = (\nu_1,\ldots,\nu_M)$, on note :

$$W = e^{\psi(\nu)},$$

où $\psi(\nu) = \ln\left(\int_x \exp\left(\sum_{m=1}^{M}\nu_m\varphi_m(\boldsymbol{x})\right) dx\right)$.

La log-vraisemblance d'un échantillon $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$ de cette loi P_ν s'écrit de la manière suivante :

$$\begin{aligned}\ln L(\underline{\boldsymbol{x}},\nu) &= \sum_{i=1}^{N}\ln f_{P_\nu}(\boldsymbol{x}_i)\\ &= -N\psi(\nu) + \sum_i\left(\sum_m \nu_m\varphi_m(\boldsymbol{x}_i)\right)\\ &= -N\psi(\nu) + \sum_m \nu_m\left(\sum_i \varphi_m(\boldsymbol{x}_i)\right)\\ &= -N\left(\psi(\nu) - \sum_m \nu_m\overline{\varphi}_m\right),\end{aligned}$$

où on retrouve l'expression de $\overline{\varphi}_m$ la moyenne empirique de φ_m sur le jeu de données.

Par annulation de la dérivée de cette log-vraisemblance selon chaque paramètre ν_m, on obtient les équations suivantes:

$$\frac{\partial\psi(\nu)}{\partial\nu_m} = \overline{\varphi}_m\ ,\ m \in \{1,\ldots,M\}\ .$$

Or, suivant un résultat classique, on a :

$$\frac{\partial\psi(\nu)}{\partial\nu_m} = E_{P_\nu}(\varphi_m) \ , \ m \in \{1, \ldots, M\} \ .$$

Les équations du maximum de vraisemblance pour la loi P_ν sont donc :

$$E_{P_\nu}(\varphi_m) = \overline{\varphi}_m \ , \ m \in \{1, \ldots, M\} \ .$$

D'autre part, les équations des constantes de normalisation sont les mêmes pour la loi P^* et la loi P_ν.

□

Les **équations du PME** et les **équations du maximum de vraisemblance** pour une loi issue d'une structure exponentielle faisant apparaître M paramètres sont donc les mêmes. Mais dans la démarche habituelle aboutissant à l'utilisation de la MMV, il est nécessaire de poser l'hypothèse que le jeu de données constitue un échantillon de la loi dont les paramètres sont estimés.

▷ Remarque

La proposition 6 établit donc clairement l'équivalence entre PME et MMV au niveau du traitement des paramètres de la loi de maximum d'entropie que sont les multiplicateurs. Cette équivalence est souvent présentée comme un fait établi (*"the principle [PME] is more fundamental than the classical maximum likelihood principle"*, [Akaike 77]), mais correspond en fait à d'autres approches que celles présentées ici :

- dans le cas d'un espace discret, il y a égalité entre l'entropie de la loi dont chaque composante est estimée par maximum de vraisemblance et la log-vraisemblance calculée pour cette loi ([Dutta 66]) ;

- cette même égalité se retrouve, pour tout type de variables, si on cherche les paramètres d'une loi répondant à une série de contraintes, par la minimisation de sa distance de Kullback à la loi empirique observée sur un échantillon ([Kapur 89, Kapur & Kesavan 87]) ;

- maximiser la vraisemblance pour obtenir des estimations des paramètres d'une loi est équivalent à minimiser une "entropie de l'échantillon" pour cette loi, qu'elle soit issue du PME ou non ([Kapur & Kesavan 87]).

- L'utilisation de la MMV pour estimer les multiplicateurs *ou* les niveaux de contraintes d'une loi de maximum d'entropie correspondant à une information donnée, conduit à prendre comme estimateurs du maximum de vraisemblance de ces derniers les moyennes empiriques des observables. On "injecte" ensuite ces niveaux estimés dans les équations du PME, mais le lien entre celles-ci et les équations du maximum de vraisemblance *pour les multiplicateurs* n'est pas précisé ([Kapur 89, Kapur & Kesavan 87]).

La proposition présentée ici s'assimile à l'approche apparaissant brièvement dans [Kullback 59] (p. 94). Nous avons en fait clairement établi que la MMV appliquée à une loi appartenant à une structure statistique exponentielle de type $\mathcal{S}_M$ pour laquelle la statistique $\overline{\varphi} = (\overline{\varphi}_1, \ldots, \overline{\varphi}_M)$ est exhaustive, consiste à écrire les équations du PME à partir des observables $(\varphi_1, \ldots, \varphi_M)$ associés aux niveaux de contraintes empiriques $(\overline{\varphi}_1, \ldots, \overline{\varphi}_M)$.

5.2 Renforcement mutuel des trois principes

• En faisant apparaître la MMV comme étant équivalente au PME *avec estimation empirique des niveaux de contraintes*, ces deux principes se renforcent mutuellement.

- Les estimateurs du maximum de vraisemblance (EMV) bénéficient de la justification théorique que constitue la propriété de concentration. En effet, une loi issue de la structure $\mathcal{S}_M$ dont les valeurs des paramètres sont données par la MMV est une loi de maximum d'entropie. Cette loi possède donc la propriété de concentrer les lois empiriques dans son voisinage. Cela aide à comprendre l'efficacité pratique reconnue de l'estimation par maximum de vraisemblance.

- Le PME, basé sur une information $\overline{\mathcal{I}}$, se trouve doté d'une justification heuristique. Les équations du PME peuvent en effet s'interpréter comme le fait d'avoir choisi les valeurs des multiplicateurs qui maximisent la probabilité de l'échantillon observé dans le cas discret, ou de choisir une densité qui maximise la distance de Kullback à la distribution empirique dans le cas continu.

• PME et MMV diffèrent néanmoins dans leur nature. Notre vision du PME propose de définir des observables qui conduisent à une structure statistique exponentielle. On s'intéresse ensuite à l'estimation d'une des deux familles de paramètres $(\lambda_1, \ldots, \lambda_M)$ ou $(a_1, \ldots, a_M)$. La MMV suppose d'emblée une structure statistique donnée, sans s'interroger sur sa construction, et se propose d'en estimer les paramètres.

- La modélisation par PME *à partir d'une information* $\overline{\mathcal{I}}$, consiste donc plus précisément à décider de ne retenir d'un jeu de données que les moyennes $(\overline{\varphi}_1, \ldots, \overline{\varphi}_M)$. Le problème consistera à calculer ensuite les valeurs des paramètres $(\lambda_1, \ldots, \lambda_M)$ du modèle à partir des observables.

- Dans l'optique de la MMV, on choisit en premier lieu une structure statistique de type $\mathcal{S}_M$ par exemple. En supposant que le jeu de données disponible est constitué d'observations indépendantes d'une même loi de cette structure, il apparaît ensuite que pour identifier les paramètres $(\lambda_1, \ldots, \lambda_M)$, il suffit de ne conserver des données que les moyennes empiriques $(\overline{\varphi}_1, \ldots, \overline{\varphi}_M)$.

En ce sens, on peut dire que le PME renverse la démarche classiquement exposée dans les cours de statistique, dans laquelle on part d'une structure statistique pour en découvrir ensuite les statistiques exhaustives (cf. par exemple [Tassi 89]).

• Par contre, ces deux principes ne se définissent que par rapport à une mesure de référence donnée sur $\mathcal{X}$. (Rappelons que nous avons décidé ici de nous baser exclusivement sur les mesures de Lebesgue ou de comptage.)

• Il y a équivalence entre les équations du PME et celles obtenues par la méthode des moments pour une loi de la structure $\mathcal{S}_M$ dans le cas où les fonctions $(\varphi_m)_{m\in\{1,\ldots,M\}}$ sont des polynômes. Dans ce cas, l'application de la méthode des moments fournit les mêmes estimations que l'application de la MMV. Le PME basé sur l'information $\overline{\mathcal{I}}$ I.5 établit donc un pont entre la vision de l'estimation de Pearson et celle de Fisher.

▷ **Exemple**

Dans le cas d'une variable réelle X, nous avons vu que le modèle correspondant à une information que nous pouvons qualifier de minimale est le modèle gaussien.

En utilisant des niveaux de contraintes empiriques, la loi obtenue pour décrire le phénomène est alors la loi normale de paramètres la moyenne $\overline{\boldsymbol{x}}$ et la variance s^2 empiriques.

La moyenne et le moment d'ordre deux ne sont donc plus ici les statistiques exhaustives que l'on doit calculer parce que l'on a imposé comme modèle un modèle gaussien. Ce sont les deux valeurs dont on décide qu'elles suffisent à résumer les données et qui nous amènent ensuite à considérer comme modèle une loi de Gauss (les paramètres finalement choisis pour caractériser le modèle sont alors $(\overline{\boldsymbol{x}}, s^2)$, et non pas les multiplicateurs de l'état de Gibbs ou les niveaux de contraintes).

5.3 Diverses approches de la modélisation par les états de Gibbs

Nous allons à présent résumer les liens qui existent entre les différentes manières d'atteindre un modèle appartenant à une structure $\mathcal{S}_M$, selon que l'on choisisse une série de contraintes de modélisation ou que l'on impose directement la forme du modèle.

5.3.1 Choix du résumé des données ou choix de la structure statistique

En utilisant exclusivement les niveaux de contraintes empiriques, on impose en fait de rechercher un modèle à partir du choix d'un résumé des données.

Du choix du résumé au choix de la structure

Pour compléter l'étude des liens entre la vision de la modélisation statistique par le PME et par les approches classiquement proposées, nous introduisons à présent le **principe de Gauss**. Ce principe d'estimation de densité a été défini et étudié principalement dans les années 1965-70. Il se base sur une "généralisation" du lien entre moments empiriques et estimateurs du maximum de vraisemblance des paramètres de la distribution de la loi normale ([Aczel 66, Mc Bride 68, Campbell 70]).

Il consiste à chercher une loi P vérifiant une série de contraintes du type $E_P(\varphi_m) = a_m$, $m \in \{1, \ldots, M\}$, et qui soit telle que les estimateurs du maximum de vraisemblance des réels a_m soient les moyennes empiriques $\overline{\varphi}_m$, $m \in \{1, \ldots, M\}$.

Ce principe se base donc d'une part sur le choix de diverses contraintes (c'est-à-dire le choix d'une famille d'éléments de connaissance comme dans le PME), et impose d'autre part que le meilleur résumé des données *au sens de la MMV* soit les moyennes empiriques des fonctions considérées dans l'écriture des contraintes. Mais ce principe ne propose pas de moyen pour atteindre le modèle répondant à ces conditions.

Il est en fait possible de démontrer que ce principe aboutit à une loi appartenant à une structure $\mathcal{S}_M$, c'est-à-dire à une loi de maximum d'entropie associée à une information de type I.5 ([Campbell 70]).

Du choix de la structure au choix du résumé

MMV et méthode des moments correspondent à une deuxième approche où la forme de la structure est choisie avant toute considération sur les données disponibles. On recherche *ensuite* le meilleur résumé des données à utiliser pour déterminer les paramètres de la loi.

Dans le cadre de la structure $\mathcal{S}_M$, la MMV aboutit à l'utilisation des moments empiriques. La méthode se base sur le principe de vraisemblance selon lequel la fonction de vraisemblance contient toute l'information sur les paramètres apportée par l'échantillon. (Pour un historique de la MMV et de la polémique ayant opposé Pearson et Fisher, on pourra se rapporter à [Edwards 72, Armatte 88].)

5.3.2 Tableau récapitulatif

Nous pouvons donc résumer par le tableau I.1, p. 44, les différentes manières d'atteindre un état de Gibbs dont les paramètres doivent répondre aux équations du PME associées à l'information empirique $\overline{\mathcal{I}}$ de taille M.

Il apparaît en fait trois approches essentielles :

- choisir en premier lieu la structure statistique $\mathcal{S}_M$ qui met en jeu la statistique exhaustive $(\overline{\varphi}_1, \ldots, \overline{\varphi}_M)$, puis appliquer la MMV pour déterminer des valeurs des paramètres (*démarche classique*) ;

Choix d'une famille d'observables φ_m, $m \in \{1, \ldots, M\}$

$\Downarrow$

Déterminer P qui vérifie les contraintes :
$E_P(\varphi_m) = a_m$, $m \in \{1, \ldots, M\}$

$\Downarrow$	$\Downarrow$
On choisit $a_m = \overline{\varphi}_m$	**Principe de Gauss**
$\Downarrow$	(Imposer que les EMV des a_m
Principe de maximum d'entropie	soient les $\overline{\varphi}_m$)
$\Downarrow$	$\Downarrow$

État de Gibbs P^* tel que $E_{P^*}(\varphi_m) = \overline{\varphi}_m$, $m \in \{1, \ldots, M\}$

$\Uparrow$	$\Uparrow$
Méthode du maximum de vraisemblance	**Méthode des moments**
	(φ_m polynômes en $\boldsymbol{x}$)
$\Uparrow$	$\Uparrow$

Choix d'une structure statistique exponentielle mettant en jeu M paramètres : $\mathcal{S}_M = \{\mathcal{X}, \mathcal{B}(\mathcal{X}), P_\nu, \nu \in \Theta\}$, Θ ouvert de $\mathbf{R}^M$ $f_{P_\nu}(\boldsymbol{x}) \propto \exp\left(\sum_m \nu_m \varphi_m(\boldsymbol{x})\right)$

TAB. I.1 – *Les différentes approches d'un même modèle*

– choisir en premier lieu les contraintes $E(\varphi_m) = a_m$, $m \in \{1, \dots, M\}$, puis rechercher la loi qui assure d'estimer les paramètres a_m par les moyennes empiriques $\overline{\varphi}_m$ (*principe de Gauss* ou *méthode des moments*, si les fonctions φ_m sont polynomiales) ;

– choisir en premier lieu M observables $(\varphi_1, \dots, \varphi_M)$ et choisir de résumer le jeu de données par les quantités $\overline{\varphi}_m$. Nous proposons ensuite d'appliquer le PME à partir de l'information empirique $\overline{\mathcal{I}}$ (*PME*).

▷ **Remarque**

On pourrait être tenté d'utiliser une information mélant niveaux de contraintes empiriques et niveaux de contraintes choisis a priori, avant même toute considération sur l'échantillon. On risque alors d'obtenir une loi de maximum d'entropie dont les caractéristiques rentreraient en contradiction avec l'information apportée par l'échantillon.

Ce "danger" peut s'illustrer simplement en considérant une population constituée de deux groupes dont on peut avoir une idée assez juste de la répartition, et sur laquelle on étudie une variable dont on peut fournir une approximation de certaines caractéristiques. Prenons l'exemple de la taille des hommes et des femmes (on note X cette taille) dans la population française.

Si on dispose d'un échantillon totalement aléatoire (échantillon de mélange), tout niveau empirique sera peu éloigné de la valeur a priori que l'on aurait pu choisir. Si ce n'est pas le cas, cela signifie que cette valeur a priori a été mal choisie. La proportion d'hommes et de femmes dans un échantillon de ce type sera par exemple toujours plus ou moins proche de l'équirépartition.

Par contre, on peut disposer d'un échantillon pour lequel on a contrôlé la répartition hommes/femmes. Il devient alors impossible d'imposer certains a priori sur la distribution de la taille, au risque de rentrer en contradiction avec la distribution empirique. Si on impose 90% d'hommes dans l'échantillon, on ne peut pas imposer que la loi de X soit d'espérance $1.60\,m$ par exemple. Inversement, si on impose que la taille moyenne dans l'échantillon soit très petite ($\overline{x} < 1.60\,m$ par exemple), on ne peut pas imposer que la répartition hommes/femmes obtenue par PME soit équirépartie. Dans ces deux cas, la loi obtenue par PME risquerait de ne correspondre à aucune population : ni à celle correspondant à une loi de mélange "théorique", ni à celle correspondant à la loi empirique.

Prenons encore l'exemple de l'étude d'une variable Z désignant la fièvre dans une population donnée, constituée de personnes malades et non malades, dont on possède un échantillon de mélange. Imposer $E(Z) = 37^o$ est valide si la population présente une proportion "raisonnable" de malades. Mais cette contrainte ne peut plus être imposée si la population est en fait issue d'un service hospitalier traitant des maladies dont un des symptômes est une forte fièvre.

6 Formalisme des calculs et exemples

Pour résumer le formalisme d'application du PME pour une information donnée, nous utiliserons un tableau construit sur le modèle du tableau I.2, p. 46. Nous y ferons apparaître les observables, les niveaux de contraintes estimés, l'expression de l'état de Gibbs maximisant l'entropie S (s'il est possible de prouver son existence), et les équations à résoudre pour obtenir les valeurs des multiplicateurs.

Ce formalisme nous permettra d'écrire de manière "automatique" la recherche de la loi de maximum d'entropie dans les cas où la proposition 2, p. 25, est vérifiée.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
φ_1	$\overline{\varphi}_1 = \frac{1}{N}\sum_{i=1}^{N} \varphi_1(\boldsymbol{x}_i)$	λ_1
$\vdots$	$\vdots$	$\vdots$
φ_M	$\overline{\varphi}_M = \frac{1}{N}\sum_{i=1}^{N} \varphi_M(\boldsymbol{x}_i)$	λ_M

État de Gibbs

$$f_{P^*}(\boldsymbol{x}) = \frac{1}{Z}\exp(\sum_{m=1}^{M} \lambda_m \varphi_m(\boldsymbol{x}))\ ,\ \boldsymbol{x} \in \mathcal{X}$$

Équations du PME

$$\frac{1}{N}\sum_{i=1}^{N} \varphi_m(\boldsymbol{x}_i) = E_{P^*}(\varphi_m)\ ,\ m \in \{1, \ldots, M\}$$

Constante de normalisation

$$Z = \int \exp(\sum_{m=1}^{M} \lambda_m \varphi_m(\boldsymbol{x}))$$

TAB. I.2 – *Tableau récapitulatif: recherche d'une loi de PME*

Nous présentons à présent quelques situations qui mettent en jeu des informations de natures diverses pour lesquelles il existe effectivement un état de Gibbs satisfaisant aux contraintes alors imposées. Nous verrons dans ces exemples comment le PME permet en fait de retrouver des modèles classiques en les associant à une information particulière. Ces modèles trouvent ainsi une nouvelle justification a posteriori de leur efficacité et leurs expressions se comprennent en termes de choix d'observables (les modèles ne donnent plus l'impression d'être "parachutés"). Les caractéristiques inhérentes à chacun de ces modèles pourront alors être vues, non plus comme des hypothèses (d'indépendance par exemple) à la base de leur définition, mais comme des conséquences du type d'information auquel ils correspondent. Le PME fournit donc ici une "relecture" de ces modèles qui nous paraît éclairer leur nature.

6.1 Relecture des modèles log-linéaires

Il s'agit d'illustrer ici le traitement des tableaux de contingence multidimensionnels par application du PME. Selon la manière dont aura été résumé ce tableau, l'expression de l'état de Gibbs et les caractéristiques du modèle obtenu seront différentes.

Nous nous limitons ici au cas de données issues de l'observation de trois variables (X_1, X_2, X_3), à valeurs dans un espace discret $\mathcal{X}$ de taille $H \times K \times R$. Il s'agit donc de choisir un modèle donnant l'expression des probabilités conjointes :

$$p_{hkr} = P(X_1 = h, X_2 = k, X_3 = r),$$

où $h \in \{1, \ldots, H\}$, $k \in \{1, \ldots, K\}$, $r \in \{1, \ldots, R\}$.

On note f_{hkr} la proportion d'individus parmi les N observations croisant les modalités h de X_1, k de X_2, r de X_3. On note d'autre part les proportions marginales de la manière suivante :

$$\left\{\begin{array}{l} f_{h..} = \sum_{kr} f_{hkr} \\ f_{.k.} = \sum_{hr} f_{hkr} \\ f_{..r} = \sum_{hk} f_{hkr} \end{array}\right. \qquad \left\{\begin{array}{l} f_{.kr} = \sum_{h} f_{hkr} \\ f_{h.r} = \sum_{k} f_{hkr} \\ f_{hk.} = \sum_{r} f_{hkr} \end{array}\right.$$

On désigne les lois marginales des variables et les lois conjointes des variables deux à deux par les expressions suivantes :

$$\left\{\begin{array}{l} p_{h..} = \sum_{kr} p_{hkr} \\ p_{.k.} = \sum_{hr} p_{hkr} \\ p_{..r} = \sum_{hk} p_{hkr} \end{array}\right. \qquad \left\{\begin{array}{l} p_{.kr} = \sum_{h} p_{hkr} \\ p_{h.r} = \sum_{k} p_{hkr} \\ p_{hk.} = \sum_{r} p_{hkr} \end{array}\right.$$

6.1.1 Résumé des données par les lois marginales empiriques

Nous allons considérer dans un premier temps que le tableau de données est résumé par les lois marginales de chacune des variables. On étudie donc la modélisation de tableaux de contingence de dimension $H \times K \times R$, à marges fixées par celles du tableau dont on dispose. Les niveaux de contraintes empiriques utilisés sont les composantes des lois empiriques de chacune des variables. Les observables correspondant à ces niveaux sont donc les indicatrices de chaque modalité possible de chacune des trois variables.

La recherche de la loi de maximum d'entropie est résumée dans le tableau I.3, p. 48.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$\mathbf{1}_{\{X_1=h\}}$	$f_{h..}$	λ_h, $h \in \{1,\ldots,H\}$
$\mathbf{1}_{\{X_2=k\}}$	$f_{.k.}$	ν_k, $k \in \{1,\ldots,K\}$
$\mathbf{1}_{\{X_3=r\}}$	$f_{..r}$	α_r, $r \in \{1,\ldots,R\}$

État de Gibbs

$$p^*_{hkr} = \frac{1}{Z}\exp(\lambda_h + \nu_k + \alpha_r)$$

Équations du PME

$$f_{h..} = p^*_{h..}$$
$$f_{.k.} = p^*_{.k.}$$
$$f_{..r} = p^*_{..r}$$

Constante de normalisation

$$Z = \sum_{hkr}\exp(\lambda_h + \nu_k + \alpha_r)$$

TAB. I.3 – *Résumé par les lois marginales*

En identifiant les lois marginales imposées par les contraintes à celles issues de l'expression de la loi de maximum d'entropie, on obtient finalement l'expression du logarithme de la loi conjointe :

$$\ln p^*_{hkr} = \ln f_{h..} + \ln f_{.k.} + \ln f_{..r}\ ,$$

pour chaque triplet (h,k,r) dans $\mathcal{X}$.

On retrouve donc l'expression d'un modèle *log-linéaire avec hypothèse d'indépendance complète* dont les paramètres (qui sont ici toutes les probabilités p^*_{hkr}) sont estimés par maximum de vraisemblance ([Christensen 90]).

Dans l'optique du PME, cette propriété d'indépendance n'est plus l'hypothèse à la base de l'écriture du modèle, mais la conséquence d'avoir résumé le tableau de données par chaque loi marginale.

6.1.2 Résumé des données par la loi marginale empirique d'une variable et la loi conjointe empiriques des deux autres

On considère à présent que le tableau de données est résumé par la loi marginale empirique de X_1 et par la loi conjointe empirique des deux autres variables (X_2, X_3). Le tableau I.4, p. 49, résume la recherche de la loi de maximum d'entropie.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$\mathbf{1}_{\{X_1=h\}}$	$f_{h..}$	$\lambda_h,\ h \in \{1,\dots,H\}$
$\mathbf{1}_{\{X_2=k,X_3=r\}}$	$f_{.kr}$	$\nu_{kr},\ k \in \{1,\dots,K\}$, $r \in \{1,\dots,R\}$
État de Gibbs $p^*_{hkr} = \frac{1}{Z}\exp(\lambda_h + \nu_{kr})$		
Équations du PME $f_{h..} = p^*_{h..}$; $f_{.kr} = p^*_{.kr}$		
Constante de normalisation $Z = \sum_{hkr} \exp(\lambda_h + \nu_{kr})$		

TAB. I.4 – *Résumé par la loi marginale de X_1 et la loi conjointe de (X_2, X_3)*

En identifiant les lois marginales imposées par les contraintes à celles issues de l'expression de la loi de maximum d'entropie, on obtient finalement :

$$\ln p^*_{hkr} = \ln f_{h..} + \ln f_{.kr} \ ,$$

pour chaque triplet (h, k, r) dans $\mathcal{X}$.

On retrouve à nouveau l'expression d'un modèle log-linéaire. Mais l'indépendance se situe ici entre la variable X_1 et le couple (X_2, X_3). Cette indépendance est là aussi une conséquence du résumé des données définissant l'information dans l'application du PME, et non l'hypothèse qui détermine la forme du modèle.

Notons de plus que dans le modèle ainsi obtenu, les variables X_1 et X_3 sont indépendantes conditionnellement à X_2 et les variables X_1 et X_2 sont indépendantes conditionnellement à X_3.

6.1.3 Résumé des données par deux lois conjointes

Un autre type de résumé du tableau peut être constitué par les lois conjointes empiriques du couple (X_1, X_2) d'une part et de celle du couple (X_1, X_3) d'autre part.

L'obtention de la loi de maximum d'entropie se résume alors par le tableau I.5, p. 50.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$\mathbf{1}_{\{X_1=h,X_2=k\}}$ $\mathbf{1}_{\{X_1=h,X_3=r\}}$	$f_{hk.}$ $f_{.kr}$	λ_{hk}, $h \in \{1,\ldots,H\}$ ν_{hr}, $k \in \{1,\ldots,K\}$ $r \in \{1,\ldots,R\}$
État de Gibbs $p^*_{hkr} = \dfrac{1}{Z}\exp(\lambda_{hk} + \nu_{hr})$		
Équations du PME $f_{hk.} = p^*_{hk.}$ $f_{h.r} = p^*_{h.r}$		
Constante de normalisation $Z = \sum_{hkr} \exp(\lambda_{hk} + \nu_{hr})$		

TAB. I.5 – *Résumé par les lois conjointes de* (X_1, X_2) *et* (X_1, X_3)

En identifiant les lois marginales imposées par les contraintes à celles issues de l'expression de la loi de maximum d'entropie, on obtient finalement :

$$\ln p^*_{hkr} = \ln f_{hk.} + \ln f_{h.r} - \ln f_{h..}\,,$$

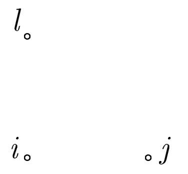
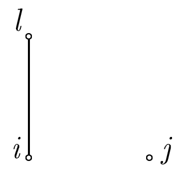
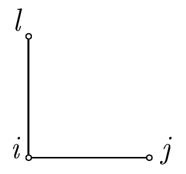
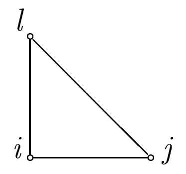
pour chaque triplet (h, k, r) dans $\mathcal{X}$.

On retrouve alors l'expression d'un *modèle log-linéaire avec hypothèse d'indépendance conditionnelle.* Il y a en effet indépendance de X_2 et X_3 *conditionnellement* à X_1. C'est ici la seule relation d'indépendance que l'on peut observer avec un tel résumé des données.

6.1.4 Représentation schématisée et généralisation

Pour chaque modèle, nous pouvons représenter les dépendances entre les trois variables mises en jeu à l'aide d'un graphe composé de trois nœuds qui représentent les variables et d'un ensemble d'arêtes reliant ces nœuds ([Whittaker 89]). Ces arêtes représentent les relations de dépendance au sens où deux sommets (i, j) ne sont pas reliés dans le graphe si et seulement si il y a indépendance des variables correspondantes X_i et X_j conditionnellement à X_l, $(i, j, l) \in \{1, 2, 3\}^3$.

Ce type de représentation permet de visualiser rapidement les différentes relations d'indépendances conditionnelles imposées par l'expression du modèle. C'est ce que nous indiquons dans le tableau I.6, p. 51, qui associe sur chaque colonne une caractéristique d'indépendance, la représentation du graphe d'indépendance correspondant et les caractéristiques d'indépendances conditionnelles que cela implique.

Caractéristique d'indépendance principale			
Indépendance mutuelle	$X_j \perp\!\!\!\perp (X_i, X_l)$	$X_j \perp\!\!\!\perp X_l \mid X_i$	Aucune indépendance
Graphe d'indépendance associé			
			
Caractéristiques d'indépendances conditionnelles			
$X_j \perp\!\!\!\perp X_i \mid X_l$ $X_i \perp\!\!\!\perp X_l \mid X_j$ $X_l \perp\!\!\!\perp X_i \mid X_j$	$X_j \perp\!\!\!\perp X_i \mid X_l$ $X_j \perp\!\!\!\perp X_l \mid X_i$	$X_j \perp\!\!\!\perp X_l \mid X_i$	Aucune

TAB. I.6 – *Représentations graphiques des indépendances conditionnelles*

Chacun de ces graphes correspond à un modèle log-linéaire donné. Nous venons de montrer qu'il est possible de relire chacun de ces modèles comme un modèle de maximum d'entropie correspondant à une information empirique particulière. Nous pouvons généraliser les calculs précédents à tous les résumés possibles des données en terme de lois marginales empiriques associées ou non à des lois conjointes empiriques. Les arêtes du graphe schématisant un modèle peuvent alors être vues comme une représentation des lois conjointes empiriques disponibles ou choisies comme résumé des données.

Le tableau I.7, p. 52, associe à chaque résumé possible l'expression du modèle obtenu et sa caractéristique d'indépendance principale.

Dans la relecture de ces modèles log-linéaires par le PME, les propriétés d'indépendances sont une conséquence de l'information empirique utilisée, et non les hypothèses à la base du modèle.

▷ **Remarque**

L'idée de retrouver l'expression d'un modèle log-linéaire comme une conséquence du choix du résumé des données se retrouve dans [Collombier 82]. L'auteur définit un *critère d'exhaustivité* pour formuler des hypothèses de non-interaction dans l'analyse d'une table de contingence. En ajoutant l'hypothèse que tous les échantillons issus de la table de contingence qui respectent certains effectifs marginaux sont équiprobables, il caractérise ainsi les modèles log-linéaires par rapport au résumé exhaustif utilisé. Mais cette approche n'associe aucune propriété de concentration au modèle obtenu.

Résumé des données : lois empiriques de	**Modèle obtenu : expression de** $\ln p^*_{hkm}$	**Indépendance principale**
X_1, X_2, X_3	$\ln f_{h..} + \ln f_{.k.} + \ln f_{..m}$	Mutuelle
X_1, (X_2, X_3)	$\ln f_{h..} + \ln f_{.km}$	$X_1 \perp\!\!\!\perp (X_2, X_3)$
X_2, (X_1, X_3)	$\ln f_{.k.} + \ln f_{h.m}$	$X_2 \perp\!\!\!\perp (X_1, X_3)$
X_3, (X_1, X_2)	$\ln f_{..m} + \ln m_{hk.}$	$X_3 \perp\!\!\!\perp (X_1, X_2)$
(X_1, X_3), (X_2, X_3)	$\ln f_{h.m} + \ln f_{.km} - \ln f_{..m}$	$X_1 \perp\!\!\!\perp X_2 \vert X_3$
(X_1, X_2), (X_2, X_3)	$\ln f_{hk.} + \ln f_{.km} - \ln f_{.k.}$	$X_1 \perp\!\!\!\perp X_3 \vert X_2$
(X_1, X_2), (X_1, X_3)	$\ln f_{hk.} + \ln f_{h.m} - \ln f_{h..}$	$X_2 \perp\!\!\!\perp X_3 \vert X_1$
(X_1, X_2), (X_1, X_3),(X_2, X_3)	$\ln(f_{hk.}.f_{h.m}f_{.km}) - \ln(f_{h..}.f_{.k.}.f_{..m})$	Aucune

TAB. I.7 – *Relecture des modèles log-linéaires par le PME*

6.2 Relecture des modèles graphiques gaussiens

Dans le même ordre d'idée, nous allons nous intéresser à présent aux modèles pouvant être associés à un *vecteur aléatoire réel*. Pour chaque type d'information considéré nous préciserons les conditions d'indépendances associées au modèle obtenu par application du PME.

Considérons le cas de J variables $(X_1, \ldots, X_J)$ à valeurs dans $\mathbf{R}^J$. Une réalisation de ce vecteur est notée $\boldsymbol{x} = (x_1, \ldots, x_J)$.

L'information considérée ici est constituée de deux familles d'observables :

– d'une part, K observables qui lient des paires de variables :

$$\varphi_{kk'}(\boldsymbol{x}) = x_k x_{k'} \ , \ (k, k') \in \{1, \ldots, J\}^2.$$

On notera $\mathcal{K}$ l'ensemble des couples d'indices qui apparaissent dans ces observables.

– d'autre part, l'ensemble des moments d'ordre un et deux de chaque variable mise en jeu :

$$\gamma_j(\boldsymbol{x}) = x_j \ , \ \delta_j(\boldsymbol{x}) = x_j^2 \ , \ j \in \{1, \ldots, J\}.$$

L'information empirique utilisée pour appliquer le PME est notée $\overline{\mathcal{I}}$:

$$\overline{\mathcal{I}} = \{(\mathcal{K}, \overline{\mathcal{K}}), (x_1, \overline{x_1}), (x_1^2, \overline{x_1^2}), \ldots, (x_J, \overline{x_J}), (x_J^2, \overline{x_J^2})\},$$

où $\overline{\mathcal{K}}$ désigne l'ensemble des moyennes empiriques $\overline{x_k x_{k'}}$, $(k, k') \in \mathcal{K}$.

Le tableau I.8, p. 53, résume la recherche de la loi de maximum d'entropie.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$x_k x_{k'}$	$\overline{x_k x_{k'}}$	$\lambda_{kk'}$, $(k, k') \in \mathcal{K}$
x_j	$\overline{x_j}$	α_j, $j \in \{1, \ldots, J\}$
x_j^2	$\overline{x_j^2}$	ν_j, $j \in \{1, \ldots, J\}$

État de Gibbs

$$f^*(\boldsymbol{x}) = \frac{1}{Z} \exp(\sum_{k,k'} \lambda_{k,k'} x_k x_{k'} + \sum_j \alpha_j x_j + \sum_j \nu_j x_j^2)$$

Équations du PME

$$E_{P^*}(X_k X_{k'}) = \overline{x_k x_{k'}}$$
$$E_{P^*}(X_j) = \overline{x_j}$$
$$E_{P^*}(X_j^2) = \overline{x_j^2}$$

Constante de normalisation

$$Z = \int \exp(\sum_{k,k'} \lambda_{k,k'} \varphi_{k,k'}(\boldsymbol{x}) + \sum_j \alpha_j x_j + \sum_j \nu_j x_j^2)$$

TAB. I.8 – *Relecture des modèles gaussiens*

La densité de la loi de maximum d'entropie s'écrit de la manière suivante :

$$f^*(\boldsymbol{x}) \propto \exp(\sum_{k,k'} \lambda_{k,k'} x_k x_{k'}) \exp(\sum_j \alpha_j x_j + \nu_j x_j^2) \ , \ \boldsymbol{x} \in \mathcal{X}.$$

L'information considérée aboutit donc à la loi d'un vecteur gaussien, telle qu'il y ait indépendance de deux composantes X_k et $X_{k'}$ conditionnellement à toutes les autres lorsque (k, k') n'appartient pas à $\mathcal{K}$.

Nous retrouvons donc la logique de l'exemple précédent puisque le modèle obtenu peut être représenté par un graphe $\mathcal{G} = (\{1, \ldots, J\}, \mathcal{K})$ ([Whittaker 89]). Les propriétés d'indépendance entre variables ne sont plus à la base de la construction du modèle, mais sont une conséquence de l'utilisation de certaines moyennes empiriques $\overline{x_k x_{k'}}$ (qui étaient disponibles ou que l'on a jugées essentielles pour résumer les données).

6.3 Prise en compte d'une information a priori : le cas gaussien

Nous avons considéré jusqu'à présent des applications du PME pour lesquelles les valeurs des niveaux de contraintes étaient exclusivement de nature empirique. Dans cet exemple, nous allons étudier la possibilité d'inclure dans l'information des connaissances a priori sur certains paramètres de la loi.

Nous étudierons à titre illustratif un couple de variables réelles $Z = (X, Y)$ en recherchant un modèle qui permette d'étudier leur interaction.

6.3.1 Famille d'observables minimale

On dispose d'un jeu de données $\underline{z} = (\underline{x}, \underline{y}) = ((x_1, y_1), \ldots, (x_N, y_N))$ du couple (X, Y). En supposant que les couples d'observations sont indépendants et de même nature on dispose d'un échantillon de (X, Y).

Une approche classique de l'étude de l'interaction entre ces deux variables consiste à modéliser la loi du couple (X, Y) à l'aide de lois gaussiennes de la manière suivante :

loi marginale de X : $\mathcal{N}(m, \sigma^2)$, (on note la densité f_X).

loi de Y conditionnellement à $X = x$: $\mathcal{N}(ax + b, \tau^2)$, (on note la densité f_Y^x).

La densité f du couple s'écrit alors :

$$f(x, y) = \frac{1}{2\pi\sigma\tau} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \exp\left(-\frac{(y-(ax+b))^2}{2\tau^2}\right), \tag{I.8}$$

où $(m, a, b, \sigma^2, \tau^2)$ est l'ensemble des 5 paramètres réels choisis.

Une autre paramétrisation de cette densité permet de retrouver l'expression d'un état de Gibbs :

$$f(x, y) = \frac{1}{Z} \exp(\lambda_1 x + \lambda_2 x^2 + \lambda_3 y + \lambda_4 y^2 + \lambda_5 xy),$$

où, par simple identification, les réels $(\lambda_1, \ldots, \lambda_5)$ s'expriment explicitement en fonction des paramètres des lois gaussiennes, comme nous le préciserons par la suite.

Ainsi, toute loi dont la densité s'exprime sous la forme I.8 peut être relue comme une *loi de maximum d'entropie* correspondant à une information mettant en jeu les observables x , x^2 , y , y^2 et xy.

On peut considérer que cette famille d'observables est minimale pour étudier un tel couple. En effet, les deux premiers moments sont les observables que l'on a décidé de toujours utiliser dans le cas de variables réelles (comme nous l'avons vu dans le cas univarié). L'observable xy est l'observable le plus simple permettant d'obtenir un modèle de maximum d'entropie avec interaction.

Ce modèle est donc bien le modèle le plus parcimonieux permettant d'étudier la liaison entre deux variables réelles.

6.3.2 Équations du PME

Nous venons de déterminer les observables permettant d'atteindre la loi I.8 par application du PME. Dans le cas où ces observables sont associés à des niveaux de contraintes estimés sur le jeu de données, les équations du PME sont exactement celles issues de la maximisation de la vraisemblance définie à partir de cette densité (proposition 6, p. 38).

Mais nous allons ici nous écarter de cette situation, en faisant intervenir dans l'information des niveaux de contraintes fixés en dehors de toute utilisation du jeu de données. Nous nous intéresserons ainsi à la manière dont le PME peut prendre spécifiquement en compte une information *a priori*. Cette situation peut se rencontrer quand certains niveaux de contraintes sont fixés par des considérations théoriques (de symétrie par exemple), obtenus à partir d'estimations subjectives d'un expert, ou estimés à partir d'échantillons antérieurs.

L'information considérée a la forme générale suivante :

$$\mathcal{I} = \{(x, c_1), (x^2, c_2), (y, c_3), (y^2, c_4), (xy, c_5)\},$$

où les niveaux de contraintes $(c_1, \ldots, c_5)$ sont des réels quelconques.

La famille de paramètres à laquelle nous nous intéressons ici n'est pas constituée par les multiplicateurs de la loi de maximum d'entropie, mais par les paramètres caractérisant la loi marginale de X et par ceux des lois conditionnelles de Y sachant X.

Le tableau I.9, p. 56, présente les différentes relations entre les multiplicateurs, les niveaux de contraintes et ces paramètres d'intérêt $(m, \sigma^2, a, b, \tau^2)$.

Expressions des paramètres $(m, a, b, \sigma^2, \tau^2)$

Nous ne nous intéressons pas ici à l'expression des multiplicateurs en fonction des niveaux de contraintes comme nous le faisions dans les exemples précédents, mais à celle des paramètres (m, σ^2) d'une part, et (a, b, τ^2) d'autre part.

Observables	**Multiplicateurs**	**Niveaux de contraintes**
x	$\lambda_1 = -\frac{ab}{\tau^2} + \frac{m}{\sigma^2}$	$c_1 = E_{P^*}(X) = m$
x^2	$\lambda_2 = -\frac{1}{2}(\frac{a^2}{\tau^2} + \frac{1}{\sigma^2})$	$c_2 = E_{P^*}(X^2) = \sigma^2 + m^2$
y	$\lambda_3 = \frac{b}{\tau^2}$	$c_3 = E_{P^*}(Y) = am + b$
y^2	$\lambda_4 = -\frac{1}{2\tau^2}$	$c_4 = E_{P^*}(Y^2) = \tau^2 + a^2(\sigma^2 + m^2) +2abm + b^2$
xy	$\lambda_5 = \frac{a}{\tau^2}$	$c_5 = E_{P^*}(XY) = a(\sigma^2 + m^2) + bm$

TAB. I.9 – *Relations entre les trois familles de paramètres du modèle:* $(\lambda_1, \ldots, \lambda_5)$, $(c_1, \ldots, c_5)$, *et* $(m, a, b, \sigma^2, \tau^2)$

On note $(m^*, a^*, b^*, \sigma^{*2}, \tau^{*2})$ les valeurs des paramètres qui vérifient les équations du PME:

$$\begin{cases} m^* &= c_1 \\ \sigma^{*2} &= c_2 - c_1^2 \end{cases} \qquad \text{et} \qquad \begin{cases} a^* &= \dfrac{c_5 - c_1 c_3}{c_2 - c_1^2} \\ b^* &= c_3 - c_1 \left(\dfrac{c_5 - c_1 c_3}{c_2 - c_1^2} \right) \\ \tau^{*2} &= (c_4 - c_3^2) - \dfrac{(c_5 - c_1 c_3)^2}{c_2 - c_1^2}. \end{cases}$$

Ces expressions sont *toujours* les mêmes, que les niveaux de contraintes prennent des valeurs calculées à partir de l'échantillon ou des valeurs fixées a priori. Les valeurs $(m^*, a^*, b^*, \sigma^{*2}, \tau^{*2})$ sont obtenues en remplaçant simplement les niveaux de contraintes par leurs valeurs. Si tous les niveaux $(c_1, \ldots, c_5)$ sont connus a priori, il n'y a pas à utiliser de données pour obtenir les valeurs des paramètres. Dans le cas contraire, nous utiliserons les moyennes empiriques des observables calculées sur un jeu de données pour estimer les niveaux de contraintes inconnus. Aucun principe d'estimation n'est à la base de l'écriture de ces équations. On désignera néammoins par le terme d'**estimateurs du maximum d'entropie** (EME) les valeurs des paramètres issues de la résolution des équations.

Dans l'optique habituelle de l'étude d'un modèle de type I.8, cette forme de modèle étant choisie, des estimations des paramètres sont fournies par maximisation de la vraisemblance associée à l'échantillon, pour l'ensemble des valeurs non connues a priori parmi certains des paramètres $(m, \sigma^2, a, b, \tau^2)$. On note $L(\underline{z}, m, \sigma^2, a, b, \tau^2)$ la vraisemblance de l'échantillon $\underline{z}$ quand les cinq paramètres doivent être estimés. Rappelons que la log-vraisemblance a alors la forme suivante :

$$\begin{aligned}\ln L(\underline{z}, m, \sigma^2, a, b, \tau^2) &= -N\ln(2\pi) - N\Bigg(\ln(\tau\sigma) + \frac{m^2}{2\sigma^2} + \frac{b^2}{2\tau^2} + (\frac{ab}{\tau^2} - \frac{m}{\sigma^2})\overline{x} \\ &\quad -\frac{b}{\tau^2}\overline{y} + (\frac{a^2}{2\tau^2} + \frac{1}{2\sigma^2})\overline{x^2} + \frac{1}{2\tau^2}\overline{y^2} - \frac{a}{\tau^2}\overline{xy}\Bigg). \qquad \text{(I.9)}\end{aligned}$$

Il s'agit dans ce cas d'annuler la dérivé de cette fonction selon chacun des cinq paramètres. Pour toute connaissance a priori, on modifie l'expression de la vraisemblance. Ainsi, si la moyenne de X est connue $(m = m_0)$, on utilisera une autre vraisemblance, notée $L(\underline{z}, \sigma^2, a, b, \tau^2)$, dont on annulera les dérivés selon les quatre paramètres inconnus.

Information empirique

Dans le cas standard où aucune connaissance théorique ou subjective sur la loi du couple n'est disponible, la modélisation par PME se base uniquement sur les niveaux de contraintes empiriques. Nous sommes alors dans le cadre d'application de la proposition 6. Les EME obtenues à partir des équations du PME s'expriment alors en fonction des moyennes empiriques des observables de la manière suivante :

$$\begin{cases} m^* &= \overline{x} \\ \sigma^{*2} &= s_x^2 \end{cases} \qquad \text{et} \qquad \begin{cases} a^* &= \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{s_x^2} \\ b^* &= \overline{y} - \overline{x}\left(\dfrac{\overline{xy} - \overline{x}\,\overline{y}}{s_x^2}\right) \\ \tau^{*2} &= s_y^2 - \dfrac{(\overline{xy} - \overline{x}\,\overline{y})^2}{s_x^2} \end{cases}$$

On retrouve effectivement les estimations issues de la maximisation de la log-vraisemblance I.9, pour l'échantillon $\underline{z}$ associée à un modèle de la forme I.8.

6.3.3 Information partiellement empirique : divergence entre équations du PME et estimateurs de la MMV

Il s'agit à présent de déterminer si des informations a priori sur une partie des paramètres $(m, a, b, \sigma^2, \tau^2)$ sont prises en compte de la même manière dans les équations du PME associées aux observables (x, x^2, y, y^2, xy) et dans la maximisation de la vraisemblance correspondant à un modèle du type I.8.

Loi de X connue a priori

Considérons à présent une situation de modélisation dans laquelle l'espérance et la variance de X sont connues a priori. Les paramètres m et σ^2 n'ont plus à être déterminés. Cela peut être le cas par exemple quand cette variable représente un phénomène très général (taille, poids, ...) qui a été étudié sur des échantillons de très grandes tailles, ou si des estimations ont été obtenues par une méta-analyse (pour laquelle des échantillons issus de différentes études ont été rassemblés).

L'application du PME se base alors sur l'information suivante, qui met en jeu à la fois des niveaux de contraintes empiriques et des niveaux fixés a priori :

$$\mathcal{I} = \{(x, m), (x^2, m^2 + \sigma^2), (y, \overline{y}), (y^2, \overline{y^2}), (xy, \overline{xy})\},$$

où $m \in \mathbf{R}$, $\sigma^2 \in \mathbf{R}^+$ sont fixés.

Il reste donc à déterminer les paramètres (a, b, τ^2) des densités conditionnelles de Y sachant $X = x$, $x \in \mathbf{R}$. Les équations du PME fournissent les estimations suivantes :

$$\left\{\begin{array}{lcl} a^* & = & \dfrac{\overline{xy} - m\overline{y}}{\sigma^2} \\ b^* & = & \overline{y} - m\left(\dfrac{\overline{xy} - m\overline{y}}{\sigma^2}\right) \\ \tau^{*2} & = & s_y^2 - \dfrac{(\overline{xy} - m\overline{y})^2}{\sigma^2}. \end{array}\right.$$

Ces expressions prennent donc en compte *de manière explicite* les connaissances a priori sur la loi de X.

Remarquons qu'il est alors possible d'assimiler ce cas à celui où X suit une loi normale centrée et réduite (il suffit d'effectuer la modélisation sur les données transformées). L'expression des trois paramètres à partir des équations du PME est alors :

$$\left\{\begin{array}{lcl} a^* & = & \overline{xy} \\ b^* & = & \overline{y} \\ \tau^{*2} & = & s_y^2 - (\overline{xy})^2. \end{array}\right.$$

Dans l'approche habituelle du modèle I.8, la vraisemblance de l'échantillon porte sur ces seuls paramètres (a, b, τ^2). Elle est alors notée $L(\underline{z}, a, b, \tau^2)$. Les estimations obtenues par sa maximisation utilisent à nouveau les moments empiriques de X, et ne font pas apparaître les connaissances de m et σ^2 (pour toutes valeurs de m et σ^2, même rapportées à 0 et 1) :

$$\left\{\begin{array}{lcl} \hat{a} & = & \dfrac{\overline{xy} - \overline{x}\,\overline{y}}{s_x^2} \\ \hat{b} & = & \overline{y} - \overline{x}\left(\dfrac{\overline{xy} - \overline{x}\,\overline{y}}{s_x^2}\right) \\ \hat{\tau}^2 & = & s_y^2 - \dfrac{(\overline{xy} - \overline{x}\,\overline{y})^2}{s_x^2}. \end{array}\right.$$

Ainsi, dans l'estimation des paramètres inconnus d'un modèle de forme I.8 par maximisation de la vraisemblance, les équations ne sont pas modifiées par la connaissance des paramètres de la loi de la variable X.

Nous verrons par la suite si la différence entre les valeurs obtenues par PME et les valeurs obtenues par MMV est notable en terme de variance dans un cas particulier.

Propriétés des estimateurs du paramètre a selon l'information a priori

L'information permettant de relire le modèle I.8 est toujours fondée sur les observables x, x^2, y, y^2, xy. Mais, suivant les valeurs des paramètres fixées a priori, il ne sera pas nécessaire d'estimer tous les niveaux de contraintes. Ainsi, dans le cas où la loi de X est connue et le paramètre b fixé, il n'est plus nécessaire de calculer la moyenne $\overline{y}$. En effet, le niveau de contrainte de l'observable Y est $(am+b)$. Sa valeur est donc également connue a priori. L'utilisation de $\overline{y}$ pourrait alors engendrer des résultats contradictoires avec les connaissances imposées a priori.

Les EME (a^*, b^*, τ^{*2}), solutions des équations, et les estimations $(\hat{a}, \hat{b}, \hat{\tau}^2)$, obtenues par maximum de vraisemblance, définissent deux séries d'estimateurs des paramètres pour chaque type d'information a priori possible.

Nous résumons dans le tableau I.10, p. 60, les différences existant entre ces deux familles d'estimateurs, pour le paramètre a selon le type de connaissance a priori utilisée.

▷ **Remarques**

Dans le cas où la loi de X et l'espérance de Y sont connues (troisième ligne du tableau I.10, p. 60), nous pouvons établir une comparaison précise entre PME et MMV en étudiant les variances des estimateurs A^* et $\hat{A}$ du paramètre a. On peut toujours se ramener au cas où X suit une loi centrée et réduite ($m = 0$, $\sigma^2 = 1$) et Y est de moyenne nulle ($b = 0$).

La variance de Y s'exprime alors en fonction du paramètre a et de la variance conditionnelle : $\mathrm{var}(Y) = \tau^2 + a^2$. Le coefficient de corrélation ρ s'écrit dans ce cas :

$$\rho^2 = \frac{a^2}{\mathrm{var}(Y)}.$$

On peut alors établir le rapport des variances des deux estimateurs en fonction du coefficient de corrélation :

$$\begin{aligned}
\frac{\mathrm{var}(A^*)}{\mathrm{var}(\hat{A})} &= \left(\frac{N-2}{N}\right)(1 + 2\frac{a^2}{\tau^2}) \\
&= \left(\frac{N-2}{N}\right)\left(\frac{\mathrm{var}(Y)+a^2}{\mathrm{var}(Y)-a^2}\right) \\
&= \left(\frac{N-2}{N}\right)\left(\frac{1+\rho^2}{1-\rho^2}\right).
\end{aligned}$$

Connaissances a priori	**Observables et niveaux**	**Estimateurs de a** EMV	EME
Aucune	$\overline{\mathcal{I}}$	$\hat{A} = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X^2}$	$A^* = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X^2}$
$m = 0$ $\sigma^2 (= 1)$	$(x, 0)$ (x^2, σ^2) $(y, \overline{y})$ $(y^2, \overline{y^2})$ $(xy, \overline{xy})$	$\hat{A} = \dfrac{\overline{XY} - \overline{X}\,\overline{Y}}{S_X^2}$ $E(\hat{A}) = a$ $\mathrm{var}(\hat{A}) = \dfrac{1}{N-3}\dfrac{\tau^2}{\sigma^2}$	$A^* = \dfrac{\overline{XY}}{\sigma^2}$ $E(A^*) = a$ $\mathrm{var}(A^*) = \dfrac{\tau^2 + b^2}{N\sigma^2} + \dfrac{2a^2}{N}$
$m = 0$ $\sigma^2 (= 1)$ $b = 0$	$(x, 0)$ (x^2, σ^2) $(y, 0)$ $(y^2, \overline{y^2})$ $(xy, \overline{xy})$	$\hat{A} = \dfrac{\overline{XY}}{\overline{X^2}}$ $E(\hat{A}) = a$ $\mathrm{var}(\hat{A}) = \dfrac{1}{N-2}\dfrac{\tau^2}{\sigma^2}$	$A^* = \dfrac{\overline{XY}}{\sigma^2}$ $E(A^*) = a$ $\mathrm{var}(A^*) = \dfrac{\tau^2}{N\sigma^2} + \dfrac{2a^2}{N}$
$m = 0$ $\sigma^2 = 1$ $b = 0$ $\mathrm{var}(Y) = \sigma'^2 = 1$	$(x, 0)$ $(x^2, 1)$ $(y, 0)$ (y^2, σ'^2) $(xy, \overline{xy})$	$\hat{a}$ solution de : $a^3 - a^2\overline{xy}$ $+a(\overline{x^2} + \overline{y^2} - 1)$ $-\overline{xy} = 0$	$A^* = \overline{XY}$ $E(A^*) = a$ $\mathrm{var}(A^*) = \dfrac{1 + a^2}{N}$

TAB. I.10 – *Comparaison de l'EMV de a pour une connaissance a priori donnée et de l'EME associé à l'information correspondante*

Cela permet d'établir la proposition suivante :

Proposition 7 *Avec les notations précédentes :*

Si X suit une loi $\mathcal{N}(0,1)$ et si $E(Y)=0$, la variance de l'estimateur de maximum d'entropie du paramètre a est inférieure à celle de l'estimateur du maximum de vraisemblance, pour des tailles d'échantillon inférieures à $1+(1/\rho^2)$, si $\rho \neq 0$:

$$\mathrm{var}(A^*) \leq \mathrm{var}(\hat{A}) \;\Leftrightarrow\; N \leq 1 + \frac{1}{\rho^2}\,.$$

Si $\rho = 0$, sous les mêmes conditions on a $\mathrm{var}(A^*) \leq \mathrm{var}(\hat{A})$, pour tout N.

On obtient par exemple les résultats suivants, pour différentes valeurs du coefficient de corrélation ρ :

ρ	1	0.5	0.3	0.25	0.2	0.15	0.12	0.1
$N \leq$	2	5	12	17	26	45	70	101

Pour des tailles d'échantillons élevées, les estimations de a par $\hat{A}$ où A^* deviennent équivalentes (on retrouve le fait que pour un grand échantillon, il devient quasiment équivalent de travailler à partir des moments connus a priori ou de leurs estimations empiriques).

Quelles que soient les connaissances a priori, les conditions de validité de la borne de Cramer-Rao pour le paramètre a sont vérifiées. Cette borne s'exprime en fonction de l'espérance de X, des variances τ^2 et σ^2 et de la taille de l'échantillon. Ainsi la variance de tout estimateur sans biais $\tilde{A}$ de a vérifie l'inégalité suivante :

$$\mathrm{var}(\tilde{A}) \geq \frac{\tau^2}{N(m^2+\sigma^2)}.$$

Dans le cas où la loi de X est connue a priori (on se rapporte toujours à une loi centrée-réduite), il apparaît que cette borne n'est jamais atteinte par la variance de $\hat{A}$.

Sous cette même condition, la variance de l'estimateur A^* n'atteint jamais la borne de Cramer-Rao, hormis dans le cas où $\overline{xy}=0$, ce qui correspond à l'indépendance des deux variables.

Dans le cas où les lois marginales de X et de Y sont connues (quatrième ligne du tableau I.10, p. 60), on peut toujours se rapporter à des lois centrées et réduites. Le paramètre a est alors exactement le coefficient de corrélation ρ et reste le seul paramètre à estimer.

La vraisemblance associée dans ce cas au modèle I.8 est alors notée $L(\underline{z},\rho)$ et s'écrit :

$$\ln L(\underline{z},\rho) = -N\ln(2\pi) - \frac{N}{2}\ln(1-\rho^2) - \frac{N}{2(1-\rho^2)}(\overline{x^2}+\overline{y^2}) + \frac{N\rho}{1-\rho^2}\overline{xy}\,.$$

L'estimateur du maximum de vraisemblance du coefficient de corrélation est alors solution d'une équation du troisième degré.

A partir des cinq mêmes observables x, x^2, y, y^2, xy, l'estimateur A^* est alors $\overline{XY}$.

Notons que dans le cas d'échantillons de grandes tailles, $\overline{x^2}$ et $\overline{y^2}$ seront très proches de 1. Les solutions de l'équation du troisième degré en ρ seront alors voisines de $-i$, i et $\overline{xy}$.

L'exemple que nous venons de présenter fournit donc une illustration des différences qui peuvent exister pour des *petits* échantillons dans l'estimation des paramètres, entre une approche classique par maximisation de la vraisemblance pour un modèle donné et une relecture par le PME de ce même modèle.

6.4 Cas mixte : un exemple simple

Les exemples précédents mettaient en jeu des variables toutes discrètes ou toutes réelles. Nous allons à présent considérer le cas de l'étude d'une variable T *discrète ordonnée* à valeurs dans $\{1, \ldots, k\}$ et d'une variable Y *réelle positive.*

L'information la plus parcimonieuse pour obtenir une loi du couple (T, Y) par application du PME ne fait apparaître que l'observable liant ces deux variables :

$$\mathcal{I} = \{(TY, a), a \in \mathbf{R}\}.$$

6.4.1 Loi de maximum d'entropie

La proposition 3 généralise l'utilisation des états de Gibbs pour les lois conditionnelles de la variable Y, dans le cas où la loi de T est fixée par l'information. Ce n'est pas le cas ici.

On note P^* la loi conjointe qui se décompose en loi marginale de T et lois conditionnelles de Y prenant les formes suivantes :

Loi marginale de T : $p_t^* = \dfrac{\tilde{k}}{t}$, où $\tilde{k} = \dfrac{1}{\sum_{t=1}^{k} 1/t}$.

Loi conditionnelle de Y sachant $T = t$: $f_Y^{*t}(y) = \dfrac{t}{a} \exp\left(-\dfrac{1}{a} ty\right)$.

Ces expressions proviennent de la maximisation du lagrangien associé au problème de maximisation de S sous la contrainte $E(TY) = a$, sans se soucier des justifications de l'application d'une telle méthode. On maximise l'expression suivante en $f_Y^t(y)$ pour t fixé, puis en p_t pour t fixé après avoir calculé $S(f_Y^t(y))$:

$$\mathrm{Lg} \quad = \quad S(P) + \nu(a - E(TY)) + \alpha(1 - \sum_t p_t) + \sum_t \beta_t \left(1 - \int_0^\infty f_Y^t(y)dy\right)$$

$$\begin{aligned} &= -\sum_t p_t \int_0^\infty f_Y^t(y) \ln f_Y^t(y) - \sum_t p_t \ln p_t \\ &\quad +\nu \left(a - \sum_t t p_t \int_0^\infty y f_Y^t(y) dy \right) + \alpha(1 - \sum_t p_t) + \sum_t \beta_t \left(1 - \int_0^\infty f_Y^t(y) dy \right) \end{aligned}$$

La proposition suivante établit le fait que cette expression de la loi conjointe est effectivement la loi de maximum d'entropie recherchée, par une démonstration du même type que celles des propositions 2 et 3.

Proposition 8 *Avec les notations précédentes :*

La loi P^* est l'unique loi du couple (T, Y) qui réalise le maximum de l'entropie S parmi toutes les lois P vérifiant la contrainte $E_P(TY) = a$.

Démonstration

Rappelons que toute loi $P = (P_T, (f_Y^t)_{t \in \{1,\ldots,k\}})$ dans $\mathcal{M}_\mathcal{I}$ vérifie la contrainte suivante :

$$a = \sum_t t p_t \int_y y f_Y^t(y) dy.$$

L'information $I(P) = -S(P)$ d'une telle loi peut se décomposer comme suit :

$$\begin{aligned} I(P) &= E_{P_T}(I(f_Y^T)) + I(P_T) \\ &= E_{P_T} \left(E_Y^T \left(\ln \frac{f_Y^T}{f_Y^{*T}} \right) + E_Y^T \left(\ln f_Y^{*T} \right) \right) + E_{P_T} \left(\ln \frac{p_T}{p_T^*} \right) + E_{P_T}(\ln p_T^*) \end{aligned}$$

En utilisant la contrainte sur P et les expressions de P^* et f_Y^{*t}, on obtient :

$$\begin{aligned} E_{P_T} \left(E_Y^T \left(\ln f_Y^{*T} \right) \right) + E_{P_T}(\ln p_T^*) &= \sum_t p_t \int_y \left(\ln \left(\frac{t}{a} \right) - \frac{ty}{a} \right) f_Y^t(y) dy + \sum_t p_t \ln \left(\frac{\tilde{k}}{t} \right) \\ &= \ln \frac{\tilde{k}}{a} - 1. \end{aligned}$$

Minimiser $I(P)$ revient donc à minimiser uniquement la somme suivante :

$$E_{P_T} \left(E_Y^T \left(\ln \frac{f_Y^T}{f_Y^{*T}} \right) \right) + E_{P_T} \left(\ln \frac{p_T}{p_T^*} \right).$$

La proposition 2 assure que le minimum de $E_Y^T \left(\ln(f_Y^T / f_Y^{*T}) \right)$ est atteint pour $f_Y^t = f_Y^{*t}$, $t \in \{1, \ldots, k\}$, et celui de $E_{P_T}(\ln(p_T/p_T^*))$ est atteint pour $P_T = P_T^*$.

Le minimum d'une somme ne peut pas être inférieur à la somme des minima. Le minimum de $I(P)$ est donc atteint pour $P_T = P_T^*$ (minimisation du deuxième terme) et $f_Y^t = f_Y^{*t}$ (minimisation du premier terme).

La loi P^* est donc bien l'unique loi de maximum d'entropie pour l'information considérée.

□

6.4.2 Remarques

• Si on recherche la loi conjointe de (T, Y) à partir de la même information, mais à loi de T fixée, la proposition 3 peut alors s'appliquer. Quelle que soit la loi marginale de T, on retrouve comme lois conditionnelles de Y sachant $T = t$ les lois f_Y^{*t}, $t \in \{1, \ldots, k\}$, de la proposition précédente.

L'entropie d'une loi conjointe P qui se décompose sous la forme $(P_T, (f_Y^{*t})_{t\in\{1,\ldots,k\}})$ s'écrit alors :

$$S(P) = \ln a + 1 - \sum_t p_t \ln(tp_t),$$

pour toute loi discrète P_T de T. Si on maximise cette expression sous l'unique contrainte de normalité de P_T, on obtient comme loi marginale de T la loi P_T^*.

• On pourrait chercher une loi répondant à la même unique contrainte mais qui rendrait les deux variables *indépendantes.*

La loi équirépartie sur $\{1, \ldots, k\}$ est la loi de maximum d'entropie de T quand aucune contrainte autre que la contrainte de normalisation n'est prise en compte.

Sous l'hypothèse d'indépendance des deux variables, la contrainte $E(TY) = a$ devient :

$$E(Y) = \frac{2a}{k+1}\,.$$

L'expression de la loi marginale de maximum d'entropie de Y est alors donnée par la proposition 2 (il y a bien ici un état de Gibbs satisfaisant à la contrainte précédente et le multiplicateur associé s'exprime explicitement en fonction du niveau de contrainte) :

$$f_Y^*(y) = \frac{2a}{k+1} \exp\left(-\left(\frac{k+1}{2a}\right) y\right)\,.$$

A partir de la même information, et sous l'hypothèse d'indépendance des deux variables, la loi de maximum d'entropie obtenue n'est plus la loi P^* de la proposition 8.

En fait, il est aisé de montrer que si aucun observable ne dépend à la fois de T et de Y, la loi de maximum d'entropie du couple (T, Y) s'exprime comme le produit de la loi de maximum d'entropie pour T et de la loi de maximum d'entropie pour Y. L'exemple présenté ici illustre le fait que si une loi $P_{T,Y}$ dans $\mathcal{M}(\mathcal{I})$ vérifie l'indépendance des deux variables, cette loi n'est pas forcément la loi de maximum d'entropie.

Cela contredit certaines visions du PME selon lesquelles ce principe conduirait à choisir un modèle rendant les variables mises en jeu les plus indépendantes possibles (*"In a nutshell, what the principle of maximum entropy seems to do is to generate hypotheses that assume as much statistical independence as possible in the world"*, [Good 66]).

7 Conclusion

Dans cette présentation, nous avons mis l'accent sur l'utilisation du PME dans la recherche d'un modèle paramétré et sur le traitement des paramètres de la loi obtenue. L'utilisation d'observables pour formaliser la connaissance du phénomène étudié permet d'appréhender la modélisation statistique et les modèles courants dans une optique différente des présentations habituelles.

Nous nous sommes restreints dans ce texte à des observables polynomiaux. Mais les propositions principales 2 et 6, portant sur la forme des lois de maximum d'entropie et sur l'équivalence entre PME et MMV au sens des estimations des paramètres, restent vraies pour tous les types d'observables.

Ainsi, la proposition 2 montre qu'une loi gamma est une loi de maximum d'entropie pour les observables x et $\ln x$, $x \in \mathbf{R}^{+*}$ et qu'une loi beta est la loi de maximum d'entropie correspondant aux observables $\ln x$ et $\ln(1-x)$, où $x \in]0,1[$.

Nous avons considéré dans ce travail des espaces munis des mesures de comptage ou de Lebesgue comme mesures de référence. Mais les propositions présentées peuvent se généraliser à l'utilisation de l'entropie de Kullback, dont la définition se base sur d'autres mesures de référence sur l'espace de travail.

Ainsi, la loi binomiale est une loi de maximum d'entropie pour une mesure de référence γ sur $\{1,\ldots,k\}$ telle que $\gamma(l) = C_k^l$, à partir de l'observable x. De même, la loi de Poisson est la loi de maximum d'entropie obtenue à partir de l'observable x, pour une mesure de référence γ sur $\mathbf{N}$ telle que $\gamma(l) = (1/l!)$. On peut retrouver également les lois multinomiales et binomiales négatives.

Il s'avère en fait que la majorité des modèles couramment utilisés peuvent être retrouvés par application du PME, pour des observables en général assez simples. Nous allons détailler dans le chapitre suivant comment le PME permet d'obtenir une relecture de différents modèles de discrimination.

Bibliographie

[Aczel 66] J. Aczel. *Lectures on functional equations and their applications*, Academic press, 1966.

[Aczel, Forte & Ng 74] J. Aczel, B. Forte & C. T. Ng. "Why the Shannon and Hartley entropies are *natural*", *Adv. Appl. Prob.*, **6**, 131-146, 1974.

[Akaike 77] H. Akaike. "On entropy maximisation principle", in *Application of statistics*, edited by P.R. Krishniah, North-Holland, 27-42, 1977.

[Armatte 88] M. Armatte. "La construction des notions d'estimation et de vraisemblance chez Ronald A. Fisher", *Statistique et analyse des données*, **13 (2)**, 65-92, 1988.

[Azencott 80] R. Azencott. "Grandes déviations : théorèmes à la Cramer-Chernoff et petites perturbations de systèmes dynamiques", *Cours de l'école d'été de probabilités de Saint-Flour 1978, VIII.* Lecture notes in mathematics, **774**, Springer.

[Berger *et al.* 96] A.L. Berger, V.J Della Pietra & S.A. Della Pietra. "A maximum entropy approach to natural language processing", *Computational linguistics*, **22**, 39-71, 1966.

[Mc Bride 68] W.J. Mc Bride. "A natural law", *Proc. IEEE*, **56**, 1713-1715, 1968.

[Campbell 70] L.L. Campbell. "Equivalence of Gauss's principle and minimum discrimination information estimation of probabilities", *Annals of Mathematical Statistics*, **41 (3)**, 1011-1015, 1970.

[Choong *et al.* 96] P.L. Choong, C.J.S. deSilva, H.J.S. Dawkins & G.F. Sterrett. "Entropy maximisation networks: an application to breast cancer prognosis", *IEEE transactions on neural networks*, **7**, 568-577, 1996.

[Christensen 90] R. Christensen. *Log-linear models*, Springer-Verlag, 1990.

[Collombier 82] D. Collombier. "Sur la formulation des hypothèses de non-interaction pour l'analyse des tables de contingence", *Metron*, **40 (1-2)**, 145-160, 1982.

[Csiszár 75] I. Csiszár. "I-divergence of probability distributions and minimization problems", *the Annals of Probability*, **3 (1)**, 146-158, 1975.

[Csiszár 85] I. Csiszár. "An extended maximum entropy principle and a bayesian justification (with discussion)", in *Bayesian Statistics 2*, 83-89, edited by J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith, North-Holland, 1985.

[Dowson & Wragg 73] D.C. Dowson & A. Wragg. "Maximum entropy distributions having prescribed first and second moments", *IEEE transactions on information theory*, **19**, 689-230, 1973.

[Dutta 66] M. Dutta. "On maximum (information-theoritic) entropy estimation", *Sankhya*, **28A**, 319-328, 1966.

[Edwards 72] A.W.F Edwards. *Likelihood : an account of the statistical concept of likelihood and its application to scientific inference*, Cambridge University Press, 1972.

[Ellis 85] R.S. Ellis. *Entropy, large deviations and statistical mechanics*, Springer, 1985.

[Geyer & Thompson 92] C.J. Geyer & E.A. Thompson. "Constrained Monte Carlo maximum likelihood for dependant data (with discussion)", *Journal of the Royal Statistical Society - Series B*, **54**, 657-699, 1992.

[Gokhale 74] D.V. Gokhale. "Maximum entropy characterizations of some distributions", in *A modern course on statistical distributions in scientific work. Vol. 3: Characterizations and applications*, NATO ASI series, serie C-Mathematical and Physical Sciences, **17**, 299-304, edited by G.P. Patil, S. Kotz & J.K. Ord, D. Reidel, 1975.

[Golan *et al.* 96] A. Golan, G. Judge, J.M. Perloff. "A maximum entropy approach to recovering information from multinomial response data", *Journal of the American Statistical Association*, **91 (434)**, 841-853, 1996.

[Good 63] I.J. Good. "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables", *Annals of Mathematical Statistics*, **34**, 911-934, 1963.

[Good 66] I.J. Good. "How to estimate probabilities", *J. Inst. Maths Appli*, **2**, 364-383, 1966.

[Gourieroux & Monfort 89] C. Gourieroux & A. Monfort. *Statistique et modèles économétriques, vol. 1 : notions générales, estimation, prévision, algorithmes*, Economica, 1989.

[Guiasu 77] S. Guiasu .*Information theory with applications*, Mc Graw-Hill, 1977.

[Jaynes 82] E.T. Jaynes. "On the rational of maximum-entropy methods", *Proc. IEEE*, **70**, 939-1952, 1982.

[Kapur 89] J.N. Kapur. *Maximum entropy models in science and engineering*, J. Wiley & sons, 1989.

[Kapur & Kesavan 87] J.N. Kapur & H.K. Kesavan. *Generalised maximum entropy principle (with application)*, Sanford education press Waterloo, Canada, 1987.

[Khinchin 57] A.I. Khinchin. *Mathematical foundations of information theory*, Dover, 1957.

[Kullback 54] S. Kullback. "Certain inequalities in information theory and the Cramer-Rao inequality", *Annals of Mathematical Statistics*, **25**, 745-751, 1954.

[Kullback 59] S. Kullback. *Information theory and statistics*, J.Wiley & sons, 1959

[Kullback & Kharait 65] S. Kullback. & M.A. Kharait. "A note on minimum discrimination information", *Annals of Mathematical Statistics*, **37**, 279-280, 1965.

[Lisman & Van Zuylen 72] J.H.C. Lisman & M.C.A. Van Zuylen. "Note on the generation of most probable frequency distributions", *Statistica Neerlandica*, **26**, 19-23, 1972.

[Robert 90] C. Robert. "An entropy concentration theorem : applications in artificial intelligence and descriptive statistics", *Journal of Applied Probabilities*, **27**, 303-313, 1990.

[Robert 91] C. Robert. *Modèles statistiques pour l'I.A. L'exemple du diagnostic médical*, Masson, 1991.

[Sanov 57] I.N. Sanov. "On the probability of large deviations of random variables", *Mat. Sbornik (Moscow)*, **42 (1)**, 11-44, 1957 (traduit dans *N.C. Inst. of Statist. Mimeo Series 192, 1958*).

[Seidenfeld 87] T. Seidenfeld. "Entropy and uncertainty", in *foundations of statistical inference*, **2**, edited by I.B. MacNeill, G.J. Umphrey, D. Reidel, 259-287, 1987.

[Shannon 48] C.E. Shannon. "A mathematical theory of communication", *Bell system Tech. J.*, **27**, 623-659, 1948.

[Shannon & Weaver 57] C.E. Shannon & W. Weaver. *The mathematical theory of comunication*, Univ. III. Press of Chicago, 1957.

[Tassi 89] P. Tassi. *Méthodes statistiques*, Economica, 1989.

[Tribus 79] M.Tribus. "Thirty years of information theory", in *Maximum entropy formalism*, edited by R.D. Levine and M.Tribus, MIT Press, 1979.

[Whittaker 89] Whittaker J. *Graphical models in applied multivariate statistics*, J. Wiley & sons, 1989.

Chapitre II

Le principe de maximum d'entropie : un nouveau regard sur les modèles de discrimination classiques

A Deux optiques de modélisation

Nous chercherons dans ce chapitre à confronter le principe de maximum d'entropie au problème d'une modélisation se situant dans le cadre d'une population caractérisée par deux types de variables :

- une série de variables *explicatives* (ou *covariables*), notées $\mathbf{X} = (X_1, \ldots, X_J)$, à valeurs dans $\mathcal{X}$, qui décrivent les caractéristiques des individus,

- une variable *à expliquer* (ou *variable de groupe*), noté Y, à valeurs dans $\mathcal{Y} = \{0, \ldots, K-1\}$, qui représente la partition de la population en différents groupes (ou classes). Nous considérerons le plus souvent une population constituée de deux groupes ($Y \in \{0, 1\}$).

Le but d'une telle modélisation est en particulier de pouvoir construire une règle d'affectation à l'un de ces groupes, de nouveaux individus dont on connait les valeurs $\boldsymbol{x}$ des covariables.

L'ensemble des techniques statistiques qui ont été développées dans le but de répondre à cette problématique sont regroupées sous le terme d'**analyse discriminante**. Dans cette modélisation, il s'agit en fait de mettre l'accent sur les caractéristiques conditionnelles des variables. Ainsi, deux démarches sont possibles :

• soit on décide de s'intéresser à la distribution des covariables dans chaque groupe (notée $P(\mathbf{X} = \boldsymbol{x}/Y = y)$, $\boldsymbol{x} \in \mathcal{X}$, $y \in \mathcal{Y}$, où P désigne indifféremment une densité ou une probabilité). Pour établir une règle de décision, il est alors nécessaire de pouvoir lui associer la loi marginale de la variable de groupe ;

• soit on décide de s'intéresser à la distribution de la variable de groupe sachant la valeur des covariables ($P(Y = y/\mathbf{X} = \boldsymbol{x})$, $y \in \mathcal{Y}$, $\boldsymbol{x} \in \mathcal{X}$).

Ces démarches sont parfois désignées respectivement par les termes de "sampling paradigm" et "diagnostic paradigm" (cf. [Dawid 76] qui définit ces deux concepts et qui discute des avantages de se placer dans le cadre du "diagnostic paradigm" pour appliquer les méthodes statistiques au problème du diagnostic médical).

Nous chercherons à appliquer le PME en nous plaçant successivement dans le cadre de chacune de ces démarches :

• si on utilise d'une part des observables traduisant des connaissances sur la distribution non conditionnelle des covariables (moments, interactions, ...), d'autre part des observables liant covariables et variable de groupe, et enfin les observables permettant de rechercher le modèle à loi de Y fixée par l'information, on obtient un modèle sous la forme $(P^*(\mathbf{X}/Y), P^*_Y)$. L'étude de ces familles d'observables fait l'objet de la partie **B** de ce texte (*"obtenir les caractéristiques conditionnelles des covariables"*) ;

• si on utilise d'une part des observables traduisant des connaissances sur la distribution de la variable de groupe, d'autre part des observables liant covariables et variable de groupe, et enfin les observables permettant de rechercher le modèle à loi de $\boldsymbol{X}$ fixée par l'information, on obtient un modèle sous la forme $(P^*(Y/\mathbf{X}), P^*_X)$. L'étude de ces familles d'observables fait l'objet de la partie **C** de ce texte (*"relecture de la procédure de régression logistique par le PME"*).

Nous reprendrons dans cette utilisation du PME les choix proposés dans le chapitre (1) (*"Choix d'observables et nature des niveaux de contraintes dans l'application du PME"*). Ainsi, pour ne pas avoir à justifier l'utilisation de mesures de référence autres que celles de Lebesgue ou de comptage, nous utiliserons l'entropie de Shannon. D'autre part, nous chercherons toujours à considérer en premier lieu la famille d'observables la plus simple possible en nombre et en nature, pour la rendre ensuite plus complexe et étudier alors les modifications apportées au modèle. Nous retrouvons dans ce choix les principes de simplicité et de parcimonie exposés précédemment.

Nous étudierons le comportement de diverses familles d'observables qui traduisent toutes une situation de discrimination, mais qui diffèrent suivant le type d'information disponible. Nous verrons comment la validité de ces informations dépend du type d'échantillonnage (se reporter aux définitions présentées dans le tableau II.2, p. 75) qui a permis de constituer le jeu de données utilisé, et nous verrons les conséquences des modifications éventuelles de l'information. Nous étudierons en particulier les modèles de maximum d'entropie obtenus dans le cadre du "sampling paradigm" à partir d'informations ne fixant pas la loi de la variable de groupe.

Pour chacune de ces démarches, nous mettrons l'accent sur le déroulement de l'application du PME, sur la justification et l'interprétation des observables utilisés, sur l'influence de la nature des niveaux de contraintes sur le calcul des paramètres, et enfin sur les liens entre ces modèles obtenus par PME et l'approche habituelle des modèles courants d'analyse discriminante.

En effet, les caractéristiques d'un modèle retrouvé par application du PME (indépendances entre variables, formes des lois marginales ou conditionnelles, ...) ne sont plus alors des hypothèses à la base de sa définition, mais les conséquences des observables utilisés. Si une famille d'observables aboutit à l'expression d'un modèle classique, celui-ci trouve ainsi une justification au travers du PME, et il peut être confronté à d'autres modèles en comparant les familles d'observables correspondant à chacun.

Les principales notations que nous utiliserons dans ce chapitre sont indiquées dans le tableau II.1, p. 74.

Groupes de la population
G_y désigne le groupe de la population caractérisé par $Y = y$, $y \in \{0, \ldots, K-1\}$.
π_y désigne la proportion d'individus de la population appartenant à G_y, $y \in \{0, \ldots, K-1\}$.
Configurations des covariables discrètes
Dans le cas de covariables discrètes, on note $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M)$ les configurations possibles du vecteur des covariables, où $\boldsymbol{x}^m = (x_1^m, \ldots, x_J^m)$, et $M = \text{card}(\mathcal{X})$.
Pour une covariable X_j, $j \in \{1, \ldots, J\}$, on note $(x_j^1, \ldots, x_j^{M_j})$ les M_j modalités qu'elle peut prendre.
Lois marginales et lois conditionnelles
Toute loi conjointe P des variables $(\boldsymbol{X}, Y)$ est notée $p_{xy} = P(\boldsymbol{X} = \boldsymbol{x}, Y = y)$, $\boldsymbol{x} \in \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^M\}$, $y \in \{0, 1\}$ (on considérera le plus souvent une variable de groupe binaire).
La loi de $\boldsymbol{X}$ est notée $P_X(\boldsymbol{x}) = P_X(\boldsymbol{X} = \boldsymbol{x})$, et celle de chaque X_j est notée $P_{X_j}(x) = P_{X_j}(X_j = x)$, $j \in \{1, \ldots, J\}$.
On note (p_0, p_1) les probabilités marginales de Y issues de la distribution P.
Pour une valeur donnée $y \in \{0, 1\}$ de Y, on note Q_X^y la loi de $\boldsymbol{X}$ conditionnellement à $Y = y$, et $Q_X^y(\boldsymbol{x}) = Q_X^y(\boldsymbol{X} = \boldsymbol{x})$ (cette loi ne se définit à partir de P que dans les cas où p_y est différent de 0). La loi conditionnelle de chaque covariable est notée $Q_{X_j}^y(x) = Q_{X_j}^y(X_j = x)$, $j \in \{1, \ldots, J\}$. Si ces lois admettent des densités celles-ci sont notées f_X^y et $f_{X_j}^y$.
Échantillon
On note un jeu de données $(\underline{\boldsymbol{x}}, \underline{y}) = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$, où $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$ est une observation de $\boldsymbol{X} = (X_1, \ldots, X_J)$ et y_i une observation de Y pour le même individu, avec $i \in \{1, \ldots, N\}$, N étant la taille de l'échantillon.
On note $N_y(\boldsymbol{x})$ le nombre d'individus observés dans le groupe G_y qui présentent $\boldsymbol{x}$ comme valeur de $\boldsymbol{X}$, $\boldsymbol{x} \in \mathcal{X}$. On note $N(\boldsymbol{x}) = \sum_y N_y(\boldsymbol{x})$ et $N_y = \sum_x N_y(\boldsymbol{x})$ les effectifs marginaux.

TAB. II.1 – *Notations*

Mélange	**Rétrospectif**	**Prospectif** (covariables discrètes)
Tirage tout aléatoire	Tirage à loi de Y fixée On fixe dans le protocole d'échantillonnage le nombre d'individus de l'échantillon pris dans chaque groupe : N_y fixés, $\forall y \in \mathcal{Y}$	Tirage à loi de $\boldsymbol{X}$ fixée On fixe dans le protocole d'échantillonnage le nombre d'individus de l'échantillon présentant chaque configuration possible en $\boldsymbol{X}$: $\sum_y N_y(\boldsymbol{x}^m)$ fixés, $\forall \boldsymbol{x}^m \in \mathcal{X}$

Tab. II.2 – *Description des protocoles d'échantillonnage*

B Obtenir les caractéristiques conditionnelles des covariables

Notre but est ici d'étudier des informations de plus en plus complexes, mais qui répondent toutes à la volonté d'obtenir un modèle qui traduit l'existence d'une influence de l'appartenance aux groupes sur la distribution de covariables $\mathbf{X}$. Nous préciserons suivant les cas s'il s'agit de variables discrètes ou continues.

1 Points de vue habituels

Nous proposons dans un premier temps un rappel succint sur la forme des principaux modèles d'analyse discriminante et des vraisemblances qui leur sont associées suivant le type d'échantillonnage.

1.1 Principaux modèles d'analyse discriminante

Dans le cadre du "sampling paradigm", l'approche probabiliste classique des modèles d'analyse discriminante consiste à poser des hypothèses sur les densités conditionnelles de ces covariables dans chaque groupe de la population (cf. [Celeux 90], [Das Gupta 73] pour un panorama et un historique des méthodes d'analyses discriminantes).

• Dans le cas de covariables continues à valeurs dans $\mathcal{X} \subseteq \mathbf{R}^J$, on peut par exemple décider de poser des hypothèses de normalité sur les covariables, à l'intérieur de chaque groupe de la population. Selon que la matrice de variance-covariance conditionnelle des covariables est commune aux deux groupes ou non, on parle alors d'**analyse discriminante linéaire (ADL)** ou d'**analyse discriminante quadratique (ADQ)** ([Celeux 90, McLachlan 92, Ripley 96]).

• Dans le cas de covariables discrètes, il existe deux modèles élémentaires pour lesquels on estime directement les probabilités conditionnelles de chaque configuration des covariables à partir d'un jeu de données. Mais on peut également décider de poser certaines hypothèses sur les caractéristiques conditionnelles des covariables ([Celeux & Nakache 94, Goldstein & Dillon 78, McLachlan 92]).

Dans le **modèle multinomial complet** (ou modèle empirique), on suppose que les probabilités conditionnelles des covariables sont des lois multinomiales sur $\mathcal{X}$. L'estimateur du maximum de vraisemblance de $P(\mathbf{X} = \boldsymbol{x}/Y = y)$ est alors la fréquence des observations dans le groupe G_y présentant la configuration $\boldsymbol{x}$. On a $\widehat{P}(\mathbf{X} = \boldsymbol{x}/Y = y) = N_y(\boldsymbol{x})/N_y$, où $N_y(\boldsymbol{x})$ désigne le nombre d'individus observés dans le groupe G_y qui présentent la configuration $\boldsymbol{x}$, et N_y désigne le nombre total d'individus observés dans le groupe G_y.

Dans le **modèle d'indépendance conditionnelle (MIC)** (ou modèle du premier ordre), on suppose qu'il y a indépendance entre les covariables, conditionnellement à l'appartenance aux groupes. La probabilité estimée par maximum de vraisemblance $\widehat{P}(\mathbf{X} = \boldsymbol{x}/Y = y)$ sera alors égale à $\prod_{j=1}^{J} N_y(x_j)/N_y$, où $N_y(x_j)$désigne le nombre d'individus observés dans le groupe G_y qui présente la valeur x_j pour la variable X_j. Nous désignerons par le terme de MIC l'association de la propriété d'indépendance conditionnelle *et* la forme des estimations des probabilités $P(\mathbf{X} = \boldsymbol{x}/Y = y)$ précédente. Il se peut par contre qu'un modèle possède une propriété d'indépendance conditionnelle des covariables avec d'autres formes des probabilités conditionnelles ([Mc Cullagh 80]).

La connaissance des probabilités $(\pi_0, \ldots, \pi_{K-1})$ d'appartenance aux groupes est ensuite nécessaire pour construire une **règle de décision**, dans une optique de classement de nouveaux individus. Il s'agit en effet de construire une frontière de discrimination en affectant tout nouvel individu dont on connait les valeurs $\boldsymbol{x}$ des covariables, au groupe le plus probable a posteriori. Cela revient par exemple dans le cas de deux groupes à affecter l'individu au groupe G_0 dans le cas suivant :

$$C_{10} f_X^1(\boldsymbol{x})\pi_1 < C_{01} f_X^0(\boldsymbol{x})\pi_0 \ ,$$

où $C_{yy'}$ désigne le coût de mauvais classement d'un individu de G_y dans $G_{y'}$, et f_X^y désigne la loi de probabilité ou la densité de $\mathbf{X}$ dans le groupe G_y, $(y, y') \in \{0, 1\}^2$, $y \neq y'$.

Les probabilités $(\pi_0, \ldots, \pi_{K-1})$ sont en général mal connues. Si cela est possible, on utilise alors un échantillon d'apprentissage de taille $N = \sum_y N_y$, où N_y désigne le nombre d'individus observés dans le groupe G_y, $y \in \{0, \ldots, K-1\}$. Deux approches sont alors classiquement utilisées :

- Si l'échantillon d'apprentissage est issu d'un schéma de **mélange**, les effectifs de groupe observés sont des réalisations d'une loi multinomiale de paramètres $(N, \pi_0, \ldots, \pi_{K-1})$. On peut alors choisir de prendre les fréquences empiriques N_y/N, $y \in \mathcal{Y}$ comme approximation des probabilités a priori π_y. Ces fréquences sont des estimateurs sans biais des proportions et sont en fait les estimateurs du maximum de vraisemblance.

- Si l'échantillon est issu d'un schéma **rétrospectif**, les effectifs ne sont plus des réalisations aléatoires et l'approche précédente n'est plus valide. L'habitude consiste dans ce cas à établir une stratégie de classement à partir d'une loi de Y équirépartie, ou à supposer implicitement que $C_{10}\pi_1 = C_{01}\pi_0$.

On utilise classiquement la méthode du maximum de vraisemblance pour estimer les différents paramètres des lois conditionnelles des covariables.

1.2 Vraisemblances des modèles d'analyse discriminante

Les lois conditionnelles des covariables peuvent mettre en jeu un ensemble de paramètres, dont certains dépendront du groupe d'appartenance (notés Θ_y, $y \in \mathcal{Y}$) et d'autres non (notés Θ). On note $\widetilde{\Theta} = \{\Theta, (\Theta_y)_{y\in\mathcal{Y}}\}$ l'ensemble des paramètres.

Après avoir choisi les formes des lois conditionnelles, on écrit la vraisemblance des paramètres pour l'échantillon, et sa maximisation fournit les valeurs estimées des paramètres : $\widehat{\widetilde{\Theta}} = (\widehat{\Theta}, (\widehat{\Theta}_y)_{y\in\mathcal{Y}})$.

La fonction de vraisemblance n'a pas la même expression selon que l'échantillon $(\underline{\boldsymbol{x}}, \underline{y}) = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$ a été choisi de manière totalement aléatoire ou en fixant a priori le nombre d'individus issus de chaque groupe de la population ([Ripley 96]).

– Dans le cas d'un échantillon de **mélange**, le vecteur des observations indépendantes et de même loi conjointe est bien $(\underline{\boldsymbol{x}}, \underline{y})$. La vraisemblance (notée $L(\widetilde{\Theta}, \underline{\boldsymbol{x}}, \underline{y})$ dans ce cas) s'écrit alors à partir de la loi conjointe $f_{\widetilde{\Theta}}$ de $(\boldsymbol{X}, Y)$ (où $f_{\widetilde{\Theta}}$ désigne une loi de probabilité ou une densité) :

$$L(\widetilde{\Theta}, \underline{\boldsymbol{x}}, \underline{y}) = \prod_{i=1}^{N} f_{\widetilde{\Theta}}(\boldsymbol{x}_i, y_i). \tag{II.1}$$

– Dans le cas d'un échantillon **rétrospectif**, on ne peut plus considérer le vecteur $(\underline{\boldsymbol{x}}, \underline{y})$ comme un échantillon. Seuls les tirages qui ont été faits dans chaque groupe de la population sont bien des échantillons (puisqu'on entend par échantillon un vecteur d'observations indépendantes et de même loi). La vraisemblance (notée $L_{\underline{y}}(\widetilde{\Theta}, \underline{\boldsymbol{x}})$) doit donc s'écrire à partir des densités conditionnelles $f^y_{\widetilde{\Theta}}$ de $\boldsymbol{X}$ sachant $Y = y$, $y \in \mathcal{Y}$ (les densités conditionnelles peuvent dépendre des paramètres Θ et/ou $(\Theta_y)_{y\in\mathcal{Y}}$.):

$$L_{\underline{y}}(\widetilde{\Theta}, \underline{\boldsymbol{x}}) = \prod_{y\in\mathcal{Y}} \prod_{i=1}^{N_y} f^y_{\widetilde{\Theta}}(\boldsymbol{x}_i). \tag{II.2}$$

Il apparait immédiatement que ces deux vraisemblances sont liées par la relation suivante :

$$L(\widetilde{\Theta}, \underline{\boldsymbol{x}}, \underline{y}) = L_{\underline{y}}(\widetilde{\Theta}, \underline{\boldsymbol{x}}) \times \prod_{y\in\mathcal{Y}} \pi_y^{N_y} . \tag{II.3}$$

Cette relation est toujours vérifiée, que le nombre N_y, $y \in \mathcal{Y}$, d'individus de l'échantillon issus de chaque groupe de la population soit aléatoire ou fixé par le schéma d'échantillonnage.

Ainsi, dans le cas d'un échantillon rétrospectif, on maximise directement la vraisemblance $L_{\underline{y}}(\widetilde{\Theta}, \underline{\boldsymbol{x}})$. Mais dans le cas d'un échantillon de mélange, on doit maximiser $L(\widetilde{\Theta}, \underline{\boldsymbol{x}}, \underline{y})$, qui fait apparaître la loi marginale de Y dans son expression. Deux situations sont alors possibles : soit on connait les probabilités a priori d'appartenance aux groupes G_y, $y \in \mathcal{Y}$ et l'expression $\prod_{y \in \mathcal{Y}} \pi_y^{N_y}$ est une constante ; soit ces valeurs sont inconnues. Dans ce dernier cas, les probabilités $(\pi_y)_{y \in \mathcal{Y}}$, sont intégrées dans le vecteur $\widetilde{\Theta}$ en tant que paramètres à part entière et on maximise la vraisemblance $L(\widetilde{\Theta}, \underline{\boldsymbol{x}}, \underline{y})$ pour en obtenir des estimations ([McLachlan & Basford 88, McLachlan 92, Celeux & Nakache 94, Ripley 96]).

Les valeurs estimées alors obtenues (notées $\widehat{\pi}_y$) sont les fréquences empiriques N_y/N, $y \in \mathcal{Y}$. On maximise ensuite la vraisemblance II.1, en fixant chaque π_y à la valeur $\widehat{\pi}_y$.

En adoptant cette solution, on suppose donc implicitement qu'aucun paramètre $\widetilde{\Theta}$ n'apparaît dans la loi de Y (c'est-à-dire que la loi marginale en Y n'apporte aucune information sur les paramètres $\widetilde{\Theta}$). C'est cette hypothèse qui permet d'écrire l'égalité suivante :

$$\frac{\partial \ln L_{\underline{y}}(\widetilde{\Theta}, \underline{\boldsymbol{x}})}{\partial \widetilde{\Theta}} = \frac{\partial \ln L(\widetilde{\Theta}, \underline{\boldsymbol{x}}, \underline{y})}{\partial \widetilde{\Theta}} . \tag{II.4}$$

On utilisera alors les mêmes équations d'estimation des paramètres des lois conditionnelles de $\boldsymbol{X}$ sachant $Y = y$, $y \in \mathcal{Y}$, quel que soit le type de l'échantillon.

2 Analyse discriminante par application du PME

Nous allons à présent appliquer le PME selon les principes de simplicité et de parcimonie présentés au chapitre (1). Ces mêmes principes nous avaient permis de relire divers modèles classiques (modèles log-linéaires, modèles graphiques, ...). Il s'agit ici de les adapter à la recherche d'un modèle de discrimination, dans le cas où l'accent doit être mis sur les lois conditionnelles des covariables dans chaque groupe.

Nous ne considérons pour l'instant que le cas d'une variable de groupe Y binaire, à valeurs dans $\{0, 1\}$ (nous verrons que le choix de ces valeurs aura certaines conséquences). On désigne toujours par G_0 le groupe constitué des individus tels que $Y = 0$, et G_1 celui constitué des individus tels que $Y = 1$.

2.1 Famille d'observables élémentaire

• Observables "d'interactions"

Si on ne dispose dans l'information d'aucun observable de la forme $\psi(\boldsymbol{x})y$, où $\psi(\boldsymbol{x})$ est une fonction portant sur au moins une covariable, le modèle obtenu présentera (s'il existe) une propriété d'indépendance des covariables et de la variable de groupe. La connaissance des valeurs des covariables ne sera alors d'aucun intérêt pour prédire le groupe d'appartenance (et inversement). Pour que l'information convienne à la problématique étudiée ici, elle doit donc faire apparaître au moins certains des observables simples $x_j y$, $j \in \{1, \ldots, J\}$.

Pour rester dans un cadre général, nous étudierons des informations les mettant tous en jeu, ce qui impliquera la dépendance dans le modèle de chaque covariable avec la variable Y.

• **Observables portant uniquement sur $\boldsymbol{X}$**

Selon les choix proposés au chapitre (1) de présentation du PME, l'information minimale pour l'étude d'une variable X à valeurs dans $\mathbf{R}$, est constituée par les observables x et x^2 qui correspondent à la connaissance des deux premiers moments (connus a priori ou estimés). Notre but étant ici l'étude de la distribution des covariables, toute information doit au minimum se baser sur ces observables pour chacune d'entre elles. (Nous verrons les conséquences que peuvent avoir l'absence d'un de ces observables.)

Si les covariables mises en jeu sont de type discret, nous devrons considérer au minimum les observables x_j, $j \in \{1, \ldots, J\}$.

• **Observables portant uniquement sur** Y

Il existe deux possibilités concernant les informations disponibles sur la variable de groupe Y :

- soit l'information contient des observables dont les niveaux de contraintes fixent directement la loi de Y. Dans le cas de deux groupes, pour une variable Y codée sur $\{0, 1\}$, l'élément de connaissance (y, d), $d \in]0, 1[$ correspond à ce cas de figure. La loi de maximum d'entropie doit vérifier l'ensemble des contraintes imposées par l'information. Ainsi, la loi marginale P_Y^* de Y de maximum d'entropie est intégralement déterminée : $p_1^* = d$. Il reste à écrire les lois conditionnelles de maximum d'entropie ;
- soit l'application du PME aboutit à une modélisation des lois conditionnelles *et* de cette loi marginale sans qu'il y ait d'observables fixant directement la loi de Y dans l'information. Nous reviendrons, dans le cadre d'une variable Y à valeurs dans $\{0, 1\}$, sur l'étude d'une information ne contenant pas l'observable y.

Nous noterons $\mathcal{I}_J^c$ l'information élémentaire ainsi définie dans le cas de covariables continues :

$$\mathcal{I}_J^c = \{(x_j, a_j), (x_j^2, b_j), (x_j y, c_j), j \in \{1, \ldots, J\}, (y, d)\}, \tag{II.5}$$

et $\mathcal{I}_J^d$ l'information élémentaire définie dans le cas de covariables discrètes :

$$\mathcal{I}_J^d = \{(x_j, a_j), (x_j y, c_j), j \in \{1, \ldots, J\}, (y, d)\}, \tag{II.6}$$

où $(a_j, c_j) \in \mathbf{R}^2$, $b_j \in \mathbf{R}^+$, $j \in \{1, \ldots, J\}$, et $d \in]0, 1[$, désignent les niveaux de contraintes. Nous ne nous interrogerons pas sur la nature de ces niveaux, car seule la forme des lois obtenues nous intéresse ici.

La loi de maximum d'entropie que nous voulons déterminer doit répondre aux contraintes associées à cette information a priori. Dans cette application du PME, nous recherchons donc *parmi toutes les lois possèdant une loi en Y fixée par l'information*, la loi de maximum d'entropie répondant à certaines contraintes minimales qui concernent les liens entre covariables et variable explicative.

Les valeurs des paramètres des lois de maximum d'entropie de $(\boldsymbol{X}, Y)$ sont déterminées en résolvant le système d'équations du PME, pour des valeurs données des niveaux de contraintes. Comme nous l'avons vu au chapitre (1), ces équations proviennent de l'écriture des contraintes associées aux éléments de connaissance.

Nous pouvons ici écrire ces contraintes à partir de la loi marginale discrète de Y et des lois conditionnelles des covariables (quand celles-ci admettent une densité par rapport à la mesure de Lebesgue dans le cas de covariables continues). On note m_{jy} et σ_{jy}^2 respectivement la moyenne et la variance conditionnelle de X_j sachant $Y = y$, $j \in \{1, \ldots, J\}$, $y \in \{0, 1\}$. On peut considérer ces paramètres comme faisant partie des paramètres "usuels" des lois conditionnelles (les multiplicateurs étant les paramètres "naturels" des lois de maximum d'entropie). Les contraintes s'écrivent alors de la manière suivante :

$$\left\{ \begin{array}{lllll} a_j & = & E(X_j) & = & p_0 \displaystyle\int_x x_j f_X^0(\boldsymbol{x}) d\boldsymbol{x} + p_1 \int_x x_j f_X^1(\boldsymbol{x}) d\boldsymbol{x} \\ & & & = & p_0 m_{j0} + p_1 m_{j1} \\ b_j & = & E(X_j^2) & = & p_0 \displaystyle\int_x x_j^2 f_X^0(\boldsymbol{x}) d\boldsymbol{x} + p_1 \int_x x_j^2 f_X^1(\boldsymbol{x}) d\boldsymbol{x} \\ & & & = & p_0 (m_{j0}^2 + \sigma_{j0}^2) + p_1 (m_{j1}^2 + \sigma_{j1}^2) \\ c_j & = & E(X_j Y) & = & p_1 \displaystyle\int_x x_j f_X^1(\boldsymbol{x}) d\boldsymbol{x} \\ & & & = & p_1 m_{j1} \\ d & = & p_1 \, . & & \end{array} \right.$$

La notation $f_X^y(\boldsymbol{x})$ désigne la densité en $\boldsymbol{x}$, par rapport à la mesure de Lebesgue, de $\boldsymbol{X}$ dans le groupe G_y ou la probabilité de $\boldsymbol{X}$ en $\boldsymbol{x}$. Le signe $\int$ désigne l'intégrale par rapport à la mesure de Lebesgue ou l'intégrale par rapport à la mesure de comptage dans le cas fini ou dénombrable.

▷ **Remarques**

Grâce à la nature binaire de Y, l'observable $x_j y$ traduit bien en termes de contrainte l'existence d'une influence de l'appartenance au groupe sur l'espérance de chaque variable X_j, $j \in \{1, \ldots, J\}$. Dans le cas d'une population présentant plus de deux groupes, la variable Y devient le numéro du groupe d'appartenance. Si cette numérotation est arbitraire, tout calcul faisant apparaître les différentes valeurs de cette variable n'a plus de signification concrète. Il faudra dans ce cas trouver une alternative aux observables $x_j y$, $j \in \{1, \ldots, J\}$ pour des valeurs y appartenant à un ensemble $\{0, \ldots, K-1\}$. La variable de groupe Y pourra être remplacée par $K-1$ variables indicatrices (ou variables *"design"*), qui définiront chacune J observables d'interactions (nous reviendrons sur ce point au paragraphe 2.3.3, p. 93).

Si les covariables sont toutes de nature discrètes *ordinales*, il est possible d'utiliser tout observable de type $\varphi(\boldsymbol{x})$ (qui fixe par exemple l'espérance ou des moments des covariables). Par contre, dans le cas de covariables discrètes non-ordinales dont le codage a été choisi de manière arbitraire (covariables codées dans $\{0, 1\}$ par exemple), on ne peut pas utiliser de tels observables. Dans ce cas, le traitement par PME de telles variables doit être vu comme le traitement d'un tableau de contingence multidimensionnel présentant les données recueillies.

2.2 Éléments de connaissance supplémentaires

• Autres observables de la forme $\psi(\boldsymbol{x})$

Si la connaissance du phénomène ou si l'étude descriptive globale de réalisations de celui-ci suggère qu'il faut considérer certaines interactions entre covariables, nous pouvons ajouter à l'information précédente des observables du type $x_j x_k$, où $(j, k) \in \mathcal{J}$, $\mathcal{J} \subseteq \{1, \ldots, J\}^2$. Pour rester dans un cadre général, nous considérerons le cas où l'ensemble des interactions possibles d'ordre deux est pris en considération.

L'application du PME se basera alors sur l'information suivante (dans le cas de covariables continues) :

$$\mathcal{I}_J^{c'} = \{(x_j, a_j), (x_j^2, b_j), (x_j y, c_j), (x_j x_k, b_{jk}), (j, k) \in \{1, \ldots, J\}^2, j \neq k, (y, d)\}, \tag{II.7}$$

où $b_{jk} \in \mathbf{R}$.

Les contraintes correspondant aux observables ainsi ajoutés peuvent également s'écrire à partir de la loi discrète marginale et des lois conditionnelles :

$$b_{jk} = E(X_j X_k) = p_0 \int_x x_j x_k f_X^0(\boldsymbol{x}) d\boldsymbol{x} + p_1 \int_x x_j x_k f_X^1(\boldsymbol{x}) d\boldsymbol{x},$$

pour $(j, k) \in \{1, \ldots, J\}^2$, $j \neq k$.

De même, on pourrait considérer des informations faisant apparaître tout autre observables de la forme $\psi(\boldsymbol{x})$ où ψ est une fonction de $\mathcal{X}$ dans $\mathbf{R}$ (moments d'ordre r, $r > 2$, transformation logarithmique, fonctions trigonométriques, ...). L'existence d'une loi de maximum d'entropie dépendra évidemment de la forme des observables utilisés.

• **Autres observables de la forme $\psi(\boldsymbol{x})y$**

Il se peut également que le modélisateur ait la possibilité ou l'obligation, au vu des données ou de connaissances a priori sur le phénomène, d'exprimer l'influence du groupe sur certaines caractéristiques de la distribution des covariables. Nous pouvons considérer par exemple l'ajout d'informations sur l'influence du groupe sur la dispersion de ces covariables et/ou les interactions entre celles-ci (on pourrait également traduire une influence du groupe sur toute autre fonction des covariables).

Nous définissons les observables voulus en restant dans la logique qui a été adoptée pour définir les informations élémentaires II.5 et II.6. Les observables les plus simples que l'on peut définir sont donc les suivants :

- Existence d'une influence du groupe sur la dispersion de X_j : on utilisera l'observable $x_j^2 y$.
- Existence d'une influence du groupe sur l'interaction entre X_j et X_k : on utilisera l'observable $x_j x_k y$.

Pour rester dans un cadre général, nous utiliserons là encore l'ensemble des observables $x_j^2 y$, $j \in \{1, \ldots, J\}$ et/ou l'ensemble des observables $x_j x_k y$, $(j,k) \in \{1, \ldots, J\}^2$, $j \neq k$.

2.3 Lois conditionnelles de maximum d'entropie

Si toutes les variables considérées sont à valeurs dans des espaces discrets, il s'agit en fait d'exprimer les lois conditionnelles de $\boldsymbol{X}$ sachant $Y = y$ à partir de la loi conjointe obtenue pour une information donnée. Une application du PME dans un but de discrimination revient dans ce cas à obtenir le modèle de maximum d'entropie correspondant à un certain résumé, sous forme d'un tableau de contingence multidimensionnel, des données disponibles.

Par contre, nous pouvons être confrontés au problème de la recherche d'une loi de maximum d'entropie pour un ensemble de variables mélant variables continues ($\boldsymbol{X}$) et variable discrète (Y). Mais l'information élémentaire II.5 ou toute information qui soit une extension de celle-ci, rentre dans le cadre de la proposition (3) du premier chapitre, car elles permettent toutes de rechercher le modèle de maximum d'entropie *à loi de Y fixée*. Cette proposition permet alors d'exprimer directement la forme des lois conditionnelles des covariables continues. Les densités de celles-ci dans chacun des groupes se définissent alors uniquement à partir des observables portant sur ces covariables et éventuellement de ceux liant covariables et variable de groupe.

2.3.1 Covariables discrètes

Une analyse discriminante portant sur des covariables discrètes *non-ordinales* revient à étudier les différents traitements possibles, par application du PME, d'un tableau de contingence multidimensionnel résumant les données disponibles.

On reprend les notations présentées dans le tableau II.1, p. 74.

Observables issus des tableaux de contingence multidimensionnels

On peut décider de construire un modèle par groupe en exploitant les données sous forme de plusieurs tableaux de contingence présentant les observations des covariables recueillies dans chaque groupe. Cela revient à étudier l'application du PME sur un ensemble de variables discrètes, pour une population donnée, sans but de discrimination. Ce cas a été traité dans le premier chapitre, entre autre à propos de la relecture des modèles log-linéaires par PME. Il y a autant de modèles possibles que d'observables que l'on peut définir à partir de variables discrètes.

Si on cherche à exploiter les données de manière globale à l'aide d'un unique tableau de contingence croisant covariables et variable de groupe, il existe deux situations selon que l'on considère l'information apportée par les covariables dans leur ensemble ou par chacune d'entre elles séparément.

• On peut utiliser un tableau croisant les modalités de la variable de groupe et les différentes configurations du vecteur des variables explicatives. Le tableau de contingence utilisé est alors le suivant :

Configurations en $\boldsymbol{X}$	Effectifs dans le groupe G_0	G_1	$\sum_{y\in\{0,1\}}$
$\boldsymbol{x}^1$	$N_0(\boldsymbol{x}^1)$	$N_1(\boldsymbol{x}^1)$	$N(\boldsymbol{x}^1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{x}^M$	$N_0(\boldsymbol{x}^M)$	$N_1(\boldsymbol{x}^M)$	$N(\boldsymbol{x}^M)$
$\sum_1^M$	N_0	N_1	N

Dans ce cas, le résumé "minimal" que l'on peut tirer de ce tableau est constitué par la loi marginale en $\boldsymbol{X}$, celle en Y, et par la loi de $\boldsymbol{X}$ dans un des groupes de la population. En effet, si on n'inclut pas cette dernière connaissance, aucune interaction entre covariables et variable de groupe n'apparaîtra dans le modèle puiqu'on obtient dans ce cas un modèle log-linéaire avec hypothèse d'indépendance entre $\boldsymbol{X}$ et Y (cf. chapitre (1)).

La recherche de la loi de maximum d'entropie est résumé dans le tableau II.3, p. 85.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$\mathbf{1}_{\{\mathbf{X}=\boldsymbol{x}^m\}}$	$N(\boldsymbol{x}^m)/N$	$\lambda_m,\ m \in \{1,\dots,M\}$
$y\mathbf{1}_{\{\mathbf{X}=\boldsymbol{x}^m\}}$	$N_1(\boldsymbol{x}^m)/N$	$\lambda'_m,\ m \in \{1,\dots,M\}$
y	N_y/N	γ

État de Gibbs

$$P^*_{x^m y} = \frac{1}{Z}\exp(\sum_m (\lambda_m + \lambda'_m y)\boldsymbol{x} + \gamma y)$$

Équations du PME

$$N(\boldsymbol{x}^m)/N = P^*_X(\boldsymbol{x}^m)$$
$$N_1(\boldsymbol{x}^m)/N = p^*_1 Q^{*1}_X(\boldsymbol{x}^m)$$
$$N_1/N = p^*_1$$

Constante de normalisation

$$Z = \exp(\sum_m (\lambda_m + \lambda'_m y)\boldsymbol{x} + \gamma y)$$

TAB. II.3 – *Résumé par la loi de* $\mathbf{X}$*, celle de* Y *et celle de* $\mathbf{X}$ *dans le groupe* G_1

La connaissance de la loi de $\mathbf{X}$ permet d'obtenir l'expression des probabilités conditionnelles de $\mathbf{X}$ sachant $Y = 0$:

$$Q^{*0}_X(x^m) = P^*(\mathbf{X} = \boldsymbol{x}^m / Y = 0) = N_0(\boldsymbol{x}^m)/N_0\,.$$

Ainsi, on retrouve à partir de cette information le **modèle multinomial complet** (MMC), avec estimation des paramètres (les probabilités conditionnelles) par maximum de vraisemblance. Il y a ici $2(M-1)$ niveaux de contraintes à fixer pour les observables mettant en jeu $\mathbf{X}$, et un niveau pour fixer la loi de Y.

Dans le cas de plus de deux groupes (Y à valeurs dans $\{0,\dots,K-1\}$), on retrouve ce même modèle en considérant les observables $\mathbf{1}_{\{\mathbf{X}=\boldsymbol{x}^m, Y=y\}}$, pour toute valeur de $y \in \mathcal{Y}$, et toute modalité $\boldsymbol{x}^m$, $m \in \{1,\dots,M\}$.

• On peut également utiliser un tableau croisant les modalités de *chaque* covariable avec la variable de groupe. Chaque variable X_j, $j \in \{1,\dots,J\}$ peut prendre M_j modalités notées $x_j^1,\dots,x_j^{M_j}$. Le tableau utilisé est le suivant :

Modalités		Effectifs dans le groupe		
		G_0	G_1	$\sum_{y\in\{0,1\}}$
X_1	x_1^1	$N_0(x_1^1)$	$N_1(x_1^1)$	$N(x_1^1)$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_1^{M_1}$	$N_0(x_1^{M_1})$	$N_1(x_1^{M_1})$	$N(x_1^{M_1})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_J	x_J^1	$N_0(x_J^1)$	$N_1(x_J^1)$	$N(x_J^1)$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$x_J^{M_J}$	$N_0(x_J^{M_J})$	$N_1(x_J^{M_J})$	$N(x_J^{M_J})$
$\sum_1^J$		N_0	N_1	N

Dans la même logique que précédemment, il convient donc d'extraire de ce tableau les lois marginales de chaque covariable, et la loi de chaque covariable dans un des groupes de la population (autrement, on obtiendrait un modèle présentant une indépendance entre covariables et variable de groupe). Le tableau II.4, p. 87, résume la recherche du modèle de maximum d'entropie dans ce cas.

La connaissance des lois de maximum d'entropie de chaque variable X_j, $j \in \{1, \ldots, J\}$ permet d'exprimer les probabilités de X_j conditionnellement à $Y = 0$.

On obtient donc à partir de cette information l'expression d'un **modèle d'indépendance conditionnel**, pour des covariables discrètes, avec estimation des probabilités par maximum de vraisemblance. Il y a ici $2((M_1 + \ldots + M_J) - 1)$ niveaux de contraintes à fixer pour les observables portant sur $\boldsymbol{X}$ et un niveau pour fixer la loi de Y.

Dans le cas de covariables binaires, codées dans $\{0, 1\}$, cette information revient à considérer les observables $(x_j, x_j y, y)_{j\in\{1,\ldots,J\}}$. Il y a dans ce cas $2(2J - 1)$ niveaux de contraintes à fixer pour les observables concernant $\boldsymbol{X}$ (contre $2(2^J - 1)$ si on utilise l'information précédente).

$\triangleright$ **Remarque**

Pour chacun de ces tableaux de contingence, les informations présentées sont minimales au sens où elles suffisent à obtenir un modèle donnant des expressions des lois conditionnelles en $\boldsymbol{X}$ différentes suivant le groupe d'appartenance. Mais elles impliquent le choix d'un grand nombre de niveaux de contraintes. Le MMC correspond à une information qui traite les covariables globalement, tandis que le MIC traite chaque covariable séparément.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$\mathbf{1}_{\{X_j=x_j^m\}}$	$N(x_j^m)/N$	$\lambda_{mj},\ m \in \{1,\dots,M_j\},\ j \in \{1,\dots,J\}$
$y\mathbf{1}_{\{X_j=x_j^m\}}$	$N_1(x_j^m)/N$	$\lambda'_{mj},\ m \in \{1,\dots,M_j\},\ j \in \{1,\dots,J\}$
y	N_y/N	γ

État de Gibbs

$$p^*_{x^m y} = \frac{1}{Z}\exp(\sum_j(\sum_{m=1}^{M_j}(\lambda_{mj}+\lambda'_{mj}y)x_j)+\gamma y)$$

Équations du PME

$$N(x_j^m)/N = P^*_{X_j}(x_j^m)$$
$$N_1(x_j^m)/N = p_1^* Q^{*1}_{X_j}(x_j^m)$$
$$N_1/N = p_1^*$$

Constante de normalisation

$$Z = \exp(\sum_j(\sum_{m=1}^{M_j}(\lambda_{mj}+\lambda'_{mj}y)x_j)+\gamma y)$$

TAB. II.4 – *Résumé par la loi de chaque X_j, celle de Y et celle de chaque X_j dans le groupe G_1, $j \in \{1,\dots,J\}$*

Observables définis comme des fonctions des covariables

Si le codage des covariables $(X_1,\dots,X_J)$ n'est pas arbitraire (variables ordinales), il est possible de considérer des observables du type $\varphi(\boldsymbol{x})$ et $\varphi(\boldsymbol{x})y$. Nous avons défini comme information élémentaire l'information II.6 qui met en jeu les observables $(x_j, x_j y)_{j\in\{1,\dots,J\}}$. La loi conjointe et la loi conditionnelle des covariables de maximum d'entropie alors obtenues sont les suivantes :

$$\begin{cases} P^*_{xy} &= \dfrac{1}{Z}\exp\left(\sum_j(\lambda_j+\lambda'_j)x_j\right) \\ Q^{*y}_X(\boldsymbol{x}) &= \dfrac{1}{Z_y}\exp\left(\sum_j(\lambda_j+\lambda'_j)x_j\right) \text{ , avec } y \text{ fixé} \\ Q^{*y}_{X_j}(x_j) &= \dfrac{1}{Z_{jy}}\exp\left((\lambda_j+\lambda'_j)x_j\right) \text{ , avec } y \text{ fixé,} \end{cases}$$

où $(\lambda_j,\lambda'_j) \in \mathbf{R}^2$ sont les multiplicateurs associés aux observables. Z et $(Z_y)_y$ sont des constantes de normalisation.

Il apparaît donc que ce modèle présente une **propriété d'indépendance conditionnelle des covariables**, dont les paramètres sont les multiplicateurs (λ_j, λ'_j) et non les probabilités $Q^{*y}_{X_j}(x_j)$, $j \in \{1, \ldots, J\}$. Les valeurs des paramètres obtenus par résolution des équations du PME n'ont aucune raison d'aboutir à des valeurs de $Q^{*y}_{X_j}(x_j)$ égales aux fréquences empiriques.

On peut généraliser cette conclusion à toute information mettant en jeu des observables qui portent uniquement sur chaque covariable (moments, logarithmes, ...) et des observables qui portent uniquement sur chaque covariable associée à la variable de groupe : $\varphi_{r_j}(x_j)$, $r_j \in \{1, \ldots, R_j\}$ et $\varphi'_{r'_j}(x_j, y)$, $r'_j \in \{1, \ldots, R'_j\}$, $j \in \{1, \ldots, J\}$.

Si on décide d'utiliser le modèle de maximum d'entropie ainsi obtenu dans le cadre du "sampling paradigm", on retrouve l'expression d'un modèle possédant une propriété d'indépendance conditionnelle des covariables (car il n'y a dans l'information aucun observable liant les covariables entre elles). Les multiplicateurs des lois ainsi obtenues sont calculés à partir des contraintes imposées par l'information (nous verrons par la suite dans quel cas les équations obtenues sont celles du maximum de vraisemblance pour un modèle provenant de la même structure).

Si on décide d'utiliser ce modèle de maximum d'entropie dans le cadre du "diagnostic paradigm", on obtient un modèle dont l'expression des probabilités d'appartenance aux groupes, sachant les covariables, ne dépend que des observables en x_j, $j \in \{1, \ldots, J\}$ et y (nous reviendrons sur ce point dans la partie **C**).

2.3.2 Covariables continues

• Rappel : lois de formes exponentielles et lois normales

Considérons un vecteur $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$ qui admet pour densité la fonction suivante :

$$f(\boldsymbol{z}) \propto \exp\left(\sum_k \left(\lambda_k z_k + \gamma_k z_k^2 + \sum_{j \neq k} \gamma_{jk} z_j z_k\right)\right) \;,\; \boldsymbol{z} \in \mathbf{R}.$$

Cette densité est en fait celle d'une loi normale $\mathcal{N}(m, \Sigma)$, dont les paramètres s'expriment de la manière suivante :

$$m = -\frac{1}{2}\Gamma^{-1}\Lambda \qquad \text{et} \qquad \Sigma = -\frac{1}{2}\Gamma^{-1}\,,$$

$$\text{où } \Lambda' = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_K \end{pmatrix} \qquad \text{et} \qquad \Gamma = \begin{pmatrix} \gamma_1 & \frac{\gamma_{12}}{2} & \cdots & \frac{\gamma_{1K}}{2} \\ \frac{\gamma_{12}}{2} & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{\gamma_{1(K-1)}}{2} \\ \frac{\gamma_{1K}}{2} & \cdots & \frac{\gamma_{1(K-1)}}{2} & \gamma_K \end{pmatrix}.$$

Cette propriété nous permettra d'exprimer les lois conditionnelles obtenues par PME sous la forme plus classique de lois normales.

- **À partir d'une information élémentaire**

La nature des observables $(x_j, x_j^2, x_j y)_{j\in\{1,\ldots,J\}}$ permet dans ce cas d'écrire une densité conditionnelle de $\mathbf{X}$ dans chacun des groupes de la population, vérifiant les contraintes imposées par cette information. À y fixé, $y \in \{0,1\}$, l'expression suivante définit bien en effet une densité :

$$f_X^{*y}(\boldsymbol{x}) = Z_y \exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2\right)\right), \tag{II.8}$$

où $Z_y^{-1} = \int_x \exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2\right)\right) d\boldsymbol{x}$ est la constante de normalisation, et $(\lambda_j, \lambda'_j, \gamma_j) \in \mathbf{R}^3$, $j \in \{1, \ldots, J\}$, les multiplicateurs associés aux observables.

Pour la loi ainsi obtenue, les covariables $(X_1, \ldots, X_J)$ sont indépendantes conditionnellement à Y. Il apparaît d'autre part immédiatement que chacune de ces variables suit une loi normale $\mathcal{N}(-\lambda_j/(2\gamma_j), -1/(2\gamma_j))$ dans le groupe G_0 $(Y = 0)$ et une loi normale $\mathcal{N}(-(\lambda_j + \lambda'_j)/(2\gamma_j), -1/(2\gamma_j))$ dans le groupe G_1 $(Y = 1)$.

La densité conjointe de $\mathbf{X}$ conditionnellement à $Y = y$ est donc celle d'une **loi normale multivariée** de moyenne m_y, dont l'expression dépend du conditionnement, et de matrice de variance-covariance Σ, indépendante de celui-ci.

Ces paramètres au sens usuel s'écrivent matriciellement en fonction des multiplicateurs de la manière suivante :

$$m_y = -\frac{1}{2}\Gamma^{-1}\Lambda_y \qquad \text{et} \qquad \Sigma = -\frac{1}{2}\Gamma^{-1},$$

$$\text{où } \Lambda'_y = (\lambda_1 + y\lambda'_1, \ldots, \lambda_J + y\lambda'_J)\ ,\ y \in \{0,1\}\ ,$$

$$\text{et } \Gamma = \begin{pmatrix} \gamma_1 & 0 & \ldots & 0 \\ 0 & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \gamma_J \end{pmatrix}.$$

Nous retrouvons donc à partir de cette information minimale l'expression des lois conditionnelles correspondant aux hypothèses définissant l'**analyse discriminante linéaire (ADL), avec matrice de variance-covariance diagonale**.

- **À partir d'informations plus complexes**

L'ajout d'observables du type $\psi(\boldsymbol{x})$ et/ou $\psi(\boldsymbol{x})y$, aux observables de l'information $\mathcal{I}_J^c$ permet toujours de rester dans le cadre de la proposition précédemment utilisée, à la condition que l'expression $\exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2\right) + \nu\psi(\boldsymbol{x})\right)$ (ou $\exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2\right) + \nu\psi(\boldsymbol{x}) + \nu'\psi(\boldsymbol{x})y\right)$) soit intégrable pour toute valeur fixée de $y \in \{0,1\}$, ce qui assure l'existence d'une loi de maximum d'entropie.

Par exemple, à partir de l'information $\mathcal{I}_J^c \bigcup \{(x_j x_k, b_{jk}), (j,k) \in \{1,\ldots,J\}^2, j \neq k\}$, la nature des observables permet d'écrire une densité conditionnelle de $\boldsymbol{X}$ sachant $Y = y$, $y \in \{0,1\}$, vérifiant les contraintes imposées par cette information :

$$f_X^{*y}(\boldsymbol{x}) = Z_y \exp\left(\sum_j (\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2 + \sum_{k,k\neq j} \gamma_{jk} x_j x_k)\right), \tag{II.9}$$

où $Z_y^{-1} = \int_x \exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j^2 + \gamma_j x_j y + \sum_{k\neq j} \gamma_{jk} x_j x_k\right)\right) d\boldsymbol{x}$ est la constante de normalisation, et $(\lambda_j, \lambda'_j, \gamma_j, \gamma_{jk}) \in \mathbf{R}^4$, $(j,k) \in \{1,\ldots,J\}^2$, les multiplicateurs associés aux observables.

Cette densité correspond là aussi à une loi normale multivariée, dont l'expression de la moyenne dépend à nouveau du conditionnement en Y, et dont la matrice de variance-covariance Σ est commune aux deux groupes. Mais, pour cette information, Σ est une matrice *pleine*. Ces paramètres s'expriment en fonction des multiplicateurs de la manière suivante :

$$m_y = -\frac{1}{2}\Gamma^{-1}\Lambda_y \qquad \text{et} \qquad \Sigma = -\frac{1}{2}\Gamma^{-1}\,,$$

$$\text{où } \Gamma = \begin{pmatrix} \gamma_1 & \frac{\gamma_{12}}{2} & \cdots & \frac{\gamma_{1J}}{2} \\ \frac{\gamma_{12}}{2} & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{\gamma_{1(J-1)}}{2} \\ \frac{\gamma_{1J}}{2} & \cdots & \frac{\gamma_{1(J-1)}}{2} & \gamma_J \end{pmatrix}.$$

Nous retrouvons donc à partir de cette deuxième information plus complexe, puisqu'elle prend en compte des informations sur les interactions entre covariables, l'expression des lois conditionnelles correspondant aux hypothèses de l'**analyse discriminante linéaire classique, avec matrice de variance-covariance pleine**.

Les tableaux II.5, p. 91, et II.6, p. 92, récapitulent les différentes lois conditionnelles de maximum d'entropie obtenues et leur correspondance aux modèles les plus courants d'analyse discriminante. L'expression des matrices de variance-covariance correspondant à chaque cas n'est pas explicitée en fonction des multiplicateurs ou des niveaux de contraintes. Seules les matrices Γ_y, $y \in \{0,1\}$, qui donnent directement l'expression des matrices de variance-covariance ($\Sigma_y = -1/2\Gamma_y^{-1}$) sont indiquées précisément.

Observables $j \in \{1 \dots J\}$ $k \in \{1 \dots J\}$ $k \neq j$	**Densité de maximum d'entropie** $f_X^y(\boldsymbol{x})$, $y \in \{0, 1\}$ **Expression de** Λ_y **et** Γ_y **où :** $m_y = -\frac{1}{2}\Gamma_y^{-1}\Lambda_y$ et $\Sigma_y = -\frac{1}{2}\Gamma_y^{-1}$ **Type d'analyse discriminante obtenue**
$(\mathcal{I}_J^c)$ x_j x_j^2 $x_j y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2\right)\right)$ $\Lambda_y = \begin{pmatrix} \lambda_1 + \lambda'_1 y \\ \vdots \\ \lambda_J + \lambda'_J y \end{pmatrix}$ $\Gamma = \begin{pmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \gamma_J \end{pmatrix}$ **ADLinéaire, Σ diagonale**
$(\mathcal{I}_J^c)$ x_j x_j^2 $x_j y$ Ajout : $x_j^2 y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left(\lambda_j x_j + \lambda'_j x_j y + \gamma_j x_j^2\right)\right)$ $f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda'_j y) x_j + (\gamma_j + \gamma'_j y) x_j^2\right)\right)$ $\Lambda = \begin{pmatrix} \lambda_1 + \lambda'_1 y \\ \vdots \\ \lambda_J + \lambda'_J y \end{pmatrix}$ $\Gamma_y = \begin{pmatrix} \gamma_1 + \gamma'_1 y & 0 & \dots & 0 \\ 0 & \gamma_2 + \gamma'_2 y & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \gamma_J + \gamma'_J y \end{pmatrix}$ **ADQuadratique, Σ_y diagonales**
x_j x_j^2 $x_j y$ Ajout : $x_j x_k y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda'_j y) x_j + \sum_k \gamma'_{jk} x_j x_k y + \gamma_j x_j^2\right)\right)$ $\Lambda = \begin{pmatrix} \lambda_1 + \lambda'_1 y \\ \vdots \\ \lambda_J + \lambda'_J y \end{pmatrix}$ $\Gamma_y = \begin{pmatrix} \gamma_1 & \gamma'_{12} y/2 & \dots & \gamma'_{1J} y/2 \\ \gamma'_{12} y/2 & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma'_{1(J\text{-}1)} y/2 \\ \gamma'_{1J} y/2 & \dots & \gamma'_{1(J\text{-}1)} y/2 & \gamma_J \end{pmatrix}$ **ADQuadratique, Σ_0 diagonale et Σ_1 pleine**
x_j x_j^2 $x_j y$ Ajout : $x_j x_k y$ Ajout : $x_j^2 y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda'_j y) x_j + \sum_k \gamma'_{jk} x_j x_k y + (\gamma_j + \gamma'_j y) x_j^2\right)\right)$ $\Lambda = \begin{pmatrix} \lambda_1 + \lambda'_1 y \\ \vdots \\ \lambda_J + \lambda'_J y \end{pmatrix}$ $\Gamma_y = \begin{pmatrix} \gamma_1 + \gamma'_1 y & \gamma'_{12} y/2 & \dots & \gamma'_{1J} y/2 \\ \gamma'_{12} y/2 & \gamma_2 + \gamma'_2 y & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma'_{1(J\text{-}1)} y/2 \\ \gamma'_{1J} y/2 & \dots & \gamma'_{1(J\text{-}1)} y & \gamma_J + \gamma'_J y/2 \end{pmatrix}$ **ADQuadratique, Σ_0 diagonale et Σ_1 pleine**

TAB. II.5 – *Classification de modèles de discrimination issus du PME, pour des familles d'observables de complexité croissante, sur la base de la famille d'observables* $(x_j, x_j^2, x_j y, y)$, $j \in \{1, \dots, J\}$

Observables $j \in \{1 \dots J\}$ $k \in \{1 \dots J\}$ $k \neq j$	**Densité de maximum d'entropie** $f_X^y(\boldsymbol{x})$, $y \in \{0,1\}$ **Expression de** Λ_y et Γ_y **où :** $m_y = -\frac{1}{2}\Gamma_y^{-1}\Lambda_y$ et $\Sigma_y = -\frac{1}{2}\Gamma_y^{-1}$ **Type d'analyse discriminante obtenue**
$(\mathcal{I}_J^{c\prime})$ x_j x_j^2 $x_j x_k$ $x_j y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda_j' y)x_j + \sum_k \gamma_{jk} x_j x_k + \gamma_j x_j^2\right)\right)$ $\Lambda_y = \begin{pmatrix} \lambda_1 + \lambda_1' y \\ \vdots \\ \lambda_J + \lambda_J' y \end{pmatrix} \qquad \Gamma = \begin{pmatrix} \gamma_1 & \gamma_{12}/2 & \dots & \gamma_{1J}/2 \\ \gamma_{12}/2 & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma_{1(J\text{-}1)}/2 \\ \gamma_{1J}/2 & \dots & \gamma_{1(J\text{-}1)}/2 & \gamma_J \end{pmatrix}$ **ADLinéaire,** Σ **pleine**
x_j x_j^2 $x_j y$ $x_j x_k$ Ajout : $x_j^2 y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda_j' y)x_j + \sum_k \gamma_{jk} x_j x_k + (\gamma_j + \gamma_j')x_j^2\right)\right)$ $\Lambda = \begin{pmatrix} \lambda_1 + \lambda_1' y \\ \vdots \\ \lambda_J + \lambda_J' y \end{pmatrix} \qquad \Gamma_y = \begin{pmatrix} \gamma_1 + \gamma_1' y & \gamma_{12}/2 & \dots & \gamma_{1J}/2 \\ \gamma_{12}/2 & \gamma_2 + \gamma_2' y & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma_{1(J\text{-}1)}/2 \\ \gamma_{1J}/2 & \dots & \gamma_{1(J\text{-}1)}/2 & \gamma_J + \gamma_J' y \end{pmatrix}$ **ADQuadratique,** Σ_y **pleines**
x_j x_j^2 $x_j x_k$ $x_j y$ Ajout : $x_j x_k y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda_j' y)x_j + \sum_k (\gamma_{jk} + \gamma_{jk}')x_j x_k + \gamma_j x_j^2\right)\right)$ $\Lambda = \begin{pmatrix} \lambda_1 + \lambda_1' y \\ \vdots \\ \lambda_J + \lambda_J' y \end{pmatrix} \quad \Gamma_y = \begin{pmatrix} \gamma_1 & \dots & \dots & (\gamma_{1J} + \gamma_{1J}' y)/2 \\ (\gamma_{12} + \gamma_{12}' y)/2 & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & (\gamma_{1(J\text{-}1)} + \gamma_{1(J\text{-}1)}' y)/2 \\ (\gamma_{1J} + \gamma_{1J}' y)/2 & \dots & \dots & \gamma_J \end{pmatrix}$ **ADQuadratique,** Σ_y **pleines**
x_j x_j^2 $x_j x_k$ $x_j y$ Ajout : $x_j x_k y$ Ajout : $x_j^2 y$ y	$f_X^y(\boldsymbol{x}) \propto \exp\left(\sum_j \left((\lambda_j + \lambda_j' y)x_j + \sum_k (\gamma_{jk} + \gamma_{jk}')x_j x_k + (\gamma_j + \gamma_j')x_j^2\right)\right)$ $\Lambda = \begin{pmatrix} \lambda_1 + \lambda_1' y \\ \vdots \\ \lambda_J + \lambda_J' y \end{pmatrix} \qquad \Gamma_y = \begin{pmatrix} \gamma_1 + \gamma_1' y & \dots & (\gamma_{1J} + \gamma_{1J}' y)/2 \\ (\gamma_{12} + \gamma_{12}' y)/2 & \ddots & \vdots \\ \vdots & \ddots & (\gamma_{1(J\text{-}1)} + \gamma_{1(J\text{-}1)}' y)/2 \\ (\gamma_{1J} + \gamma_{1J}' y)/2 & \dots & \gamma_J + \gamma_J' y \end{pmatrix}$ **ADQuadratique,** Σ_y **pleines**

TAB. II.6 – *Classification de modèles de discrimination issus du PME, pour des familles d'observables de complexité croissante, sur la base de la famille d'observables* $(x_j, x_j^2, x_j x_k, x_j y, y)$, $(j,k) \in \{1, \dots, J\}$, $j \neq k$

▷ **Remarques**

C'est la nature des observables en $\boldsymbol{x}$ qui aboutit pour ces informations à des lois conditionnelles normales. (Rappelons en effet que dans le cas d'une unique variable à valeurs dans $\mathbf{R}$, la famille d'observables (x, x^2) aboutit à une modélisation par une loi normale.)

Les matrices de variance-covariance conditionnelles Σ_y, $y \in \{0,1\}$ (ou la matrice Σ) ne sont plus diagonales si les observables $x_j x_k$ et/ou $x_j x_k y$, $(j,k) \in \{1,\ldots,J\}$, $j \neq k$, sont présents dans l'information. En fait, il suffit d'un seul observable "d'interaction" de ce type pour obtenir une matrice non-diagonale.

Le choix de coder Y en $(0,1)$ peut amener à obtenir une matrice Σ_y, $y \in \{0,1\}$, qui soit pleine dans un groupe et diagonale dans l'autre. Si on utilise un observable du type $\psi(\boldsymbol{x})y$ *seul*, l'espérance de la fonction $\psi(\boldsymbol{X})$ des covariables sera non-nulle dans le groupe $Y=1$ et sera nulle dans le groupe $Y=0$. Ainsi, la seule manière de faire apparaître des interactions dans les deux groupes consiste à introduire en plus des observables $x_j x_k y$, les observables $x_j x_k$ (ou $x_j x_k(1-y)$).

La diversité des familles d'observables qu'il est possible de considérer correspond à une grande diversité de modèles de maximum d'entropie pouvant répondre à une situation de discrimination. Cette vision des modèles de discrimination en terme d'observables permet de détailler et de justifier la construction de modèles pour lesquels on peut faire le choix de faire apparaître une interaction entre variables dans un groupe mais pas dans l'autre (interaction entre consommation de tabac et consommation d'alcool dans un groupe de malades mais pas d'interaction à prendre en compte dans le groupe des non-malades).

2.3.3 Population à plus de deux groupes

Considérons à présent le cas d'une population constituée d'un nombre connu K de groupes. Le codage de l'appartenance à un groupe se fait à l'aide d'une variable Y à valeurs dans $\{0,\ldots,K-1\}$ (pour rester cohérent avec le cas binaire). Il est alors possible de transformer celle-ci en $K-1$ variables indicatrices (variables *"design"* ou *"dummy"*), notées $(D_0,\ldots,D_{K-2})$, toutes binaires codées dans $\{0,1\}$ ([Celeux & Nakache 94]).

Pour exprimer l'appartenance à un groupe G_y avec $y \in \{0,\ldots,K-2\}$, on pose alors que $D_y = 1$ et toutes les autres variables prennent la valeur 0. Pour exprimer l'appartenance au groupe $y = K-1$, on code toutes les variables D_y à 0. Par exemple, dans le cas d'une population à trois groupes , on utilise deux variables indicatrices, D_0 et D_1 :

Valeur de D_0	1	0	0
Valeur de D_1	0	1	0
Groupe d'appartenance	G_0	G_1	G_2

Les observables ne seront alors plus définis à partir de Y, mais à partir du vecteur $\boldsymbol{D} = (D_0, \dots, D_{K-2})$ des variables indicatrices , et la loi cherchée sera celle du vecteur de taille $J + K - 1$, $(X_1, \dots, X_J, D_0, \dots, D_{K-2})$.

Ainsi, une information élémentaire traduisant une analyse discriminante sur des covariables continues se base sur les observables suivants :

- x_j et x_j^2, $j \in \{1, \dots, J\}$, qui constituent l'information minimale sur chaque covariable ;
- $x_j d_y$, $j \in \{1, \dots, J\}$ et $y \in \{0, \dots, K-2\}$, qui sont les observables d'interaction ;
- d_y, $y \in \{0, \dots, K-2\}$, qui permettent de fixer la loi de $\boldsymbol{D}$, et donc celle de Y.

Pour chaque configuration possible des variables indicatrices, on obtient alors comme loi conditionnelle de $\boldsymbol{X}$ sachant $\boldsymbol{D} = \boldsymbol{d}$ la loi suivante :

$$f_X^{d*}(\boldsymbol{x}) \propto \exp\left(\sum_{j=1}^{J}\left(\lambda_j + \sum_{y=0}^{K-2}\lambda'_{jy}d_y\right)x_j + \gamma_j x_j^2\right),$$

où $(\lambda_j, \lambda_{jy}, \gamma_j)$, $j \in \{1, \dots, J\}$, $y \in \{0, \dots, K-2\}$, désignent les multiplicateurs associés aux $J + K - 1$ observables.

• La densité de $\boldsymbol{X}$ sachant l'appartenance à un groupe G_y, $y \in \{0, \dots, K-2\}$ est donc celle d'une loi normale $\mathcal{N}(m_y, \Sigma)$ pour laquelle on a :

$$m_y = -\frac{1}{2}\Gamma^{-1}\Lambda_y \qquad \text{et} \qquad \Sigma = -\frac{1}{2}\Gamma^{-1},$$

$$\text{où } \Lambda'_y = (\lambda_1 + \lambda'_{1y}, \dots, \lambda_J + \lambda'_{Jy}),$$

$$\text{et } \Gamma = \begin{pmatrix} \gamma_1 & 0 & \dots & 0 \\ 0 & \gamma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \gamma_J \end{pmatrix}.$$

• La densité de $\boldsymbol{X}$ sachant l'appartenance au dernier groupe G_{K-1} est celle d'une loi normale $\mathcal{N}(m_{K-1}, \Sigma)$ pour laquelle on a :

$$m_y = -\frac{1}{2}\Gamma^{-1}\Lambda_{K-1} \qquad \text{et} \qquad \Sigma = -\frac{1}{2}\Gamma^{-1},$$

$$\text{où } \Lambda'_{K-1} = (\lambda_1, \dots, \lambda_J).$$

Les multiplicateurs des observables basés sur les variables indicatrices ne rentre donc plus en jeu dans l'expression de cette loi conditionnelle.

Si les niveaux de contraintes sont fixés aux niveaux empiriques, on retient donc des données les résumés suivants :

$$\begin{cases} \overline{d_y} & = & N(y)/N \\ \overline{x_j d_y} & = & 1/N \sum_i x_{ji} d_{yi} & = & N_y(x_j)/N \,, \end{cases}$$

pour $y \in \{0, \ldots, K-2\}$ et $j \in \{1, \ldots, J\}$, ainsi que les deux premiers moments empiriques de chaque covariable.

▷ **Remarque**

Avec ce type de codage, un des groupes de la population (ici le groupe G_{K-1}) prend un statut particulier de groupe de référence. En effet, les lois de $\boldsymbol{X}$ dans ce groupe ne dépendront pas des multiplicateurs correspondant à des observables qui portent à la fois sur les covariables et les variables *"design"*.

2.4 Équations du PME en discrimination et estimateurs du maximum de vraisemblance

Nous avons établi dans le chapitre (1) l'équivalence entre estimateurs du maximum de vraisemblance et valeurs obtenues par la résolution des équations du PME quand l'information ne présente que des niveaux de contraintes empiriques.

Dans le cadre des modèles d'analyse discriminante, la méthode du maximum de vraisemblance (MMV) s'applique aux lois conditionnelles, à partir d'une vraisemblance définie pour des données provenant d'un échantillonnage rétrospectif (cf. paragraphe 1.2, p. 78). Nous allons à présent préciser l'équivalence entre les équations du PME correspondant à une information qui permet d'aboutir à un modèle d'analyse discriminante et les équations de la MMV appliquée à un tel modèle.

Nous considérons ici le cas général d'une population constituée de K groupes, pour une variable de groupe à valeurs dans $\{0, \ldots, K-1\}$. Nous nous baserons sur une information formée à partir d'observables qui sont des fonctions *quelconques* de la variable de groupe et/ou des covariables, celles-ci étant indifféremment à valeurs dans un espace discret ou continu. Chaque observable est associé à son niveau empirique.

On note $\overline{\mathcal{I}}$ cette information, qui contient également des observables permettant de chercher le modèle de maximum d'entropie à loi de Y fixée à la loi empirique :

$$\overline{\mathcal{I}}_M = \Big\{(\varphi_m(\boldsymbol{x}, y), \overline{\varphi_m}), m \in \{1, \ldots, M\}, (\mathbf{1}_{\{Y=y\}}, N_y/N), y \in \mathcal{Y}\Big\}\,, \tag{II.10}$$

où $\overline{\varphi_m}$ désigne la moyenne empirique de l'observable $\varphi_m(\boldsymbol{x}, y)$, calculée sur un jeu de données $(\underline{\boldsymbol{x}}, \underline{y}) = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$, avec $N = \sum_y N_y$, et $\mathcal{Y} = \{0, \ldots, K-1\}$.

On suppose que la nature des observables permet d'obtenir un modèle de maximum d'entropie.

On note $\mathcal{S}_M = \{(\mathcal{X}, \mathcal{Y}), \mathcal{B}(\mathcal{X}) \times \mathcal{B}(\mathcal{Y}), ((f_X^y)_{y \in \mathcal{Y}}, p_y)\}$, la structure statistique dont les lois de $\boldsymbol{X}$ conditionnelles à $Y = y$, $y \in \mathcal{Y}$, sont des densités de la forme suivante :

$$f_X^y(\boldsymbol{x}) = Z_y \exp(\sum_{m=1}^{M} \nu_m \varphi_m(\boldsymbol{x}, y)) \ , \ \boldsymbol{x} \in \mathcal{X}. \tag{II.11}$$

Nous pouvons établir la proposition suivante :

Proposition 9 *Les équations du PME associées à l'information suivante :*

$$\overline{\mathcal{I}}_M = \left\{(\varphi_m(\boldsymbol{x}, y), \overline{\varphi_m}), M \in \{1, \ldots, M\}, (\boldsymbol{1}_{\{Y=y\}}, N_y/N), y \in \mathcal{Y}\right\} \ ,$$

sont les équations du maximum de vraisemblance pour le vecteur des paramètres d'une loi P_ν de la structure statistique $\mathcal{S}_M$, en considérant $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ comme un échantillon de cette loi.

Démonstration

Les équations du PME portant sur les observables $\varphi_m(\boldsymbol{x}, y)$ de $\overline{\mathcal{I}}_M$ s'écrivent de la manière suivante :

$$\begin{aligned} E_{P^*}(\varphi_m(\boldsymbol{X}, Y)) &= \overline{\varphi_m} \\ &= \frac{1}{N} \sum_{i=1}^{N} \varphi_m(\boldsymbol{x}_i, y_i) \\ &= \frac{1}{N} \sum_{y} \sum_{i=1}^{N_y} \varphi_m(\boldsymbol{x}_i, y) \ , m \in \{1, \ldots, M\} \ , \end{aligned}$$

soit, en exprimant ces espérances à partir de la loi marginale en Y (fixée par l'information à la loi empirique) et des lois conditionnelles en $\boldsymbol{X}$:

$$\sum_{y} \frac{N_y}{N} E_{P^*}(\varphi_m(\boldsymbol{X}, y)/Y = y) = \overline{\varphi_m} \ , \ m \in \{1, \ldots, M\} \ . \tag{II.12}$$

Considérons à présent l'estimation par maximum de vraisemblance des paramètres de la densité d'une loi issue de la structure statistique $\mathcal{S}_M$. Nous avons vu (cf. II.3) que pour tout type d'echantillon, l'application de la MMV pour obtenir des estimations des paramètres d'une telle loi s'établissait sur la base de la vraisemblance "rétrospective" $L_{\underline{y}}(\nu, \underline{\boldsymbol{x}})$, après avoir fixé la loi de Y à la loi empirique. Cette log-vraisemblance s'écrit ici de la manière suivante :

$$\ln L_{\underline{y}}(\nu, \underline{\boldsymbol{x}}) \ = \ \sum_{y} N_y \ln Z_y + \sum_{y} \sum_{i=1}^{N_y} (\sum_{m} \nu_m \varphi_m(\boldsymbol{x}_i, y)).$$

La constante de chaque densité f_X^y s'exprime en fonction des paramètre $(\nu_1, \ldots, \nu_M)$ de la manière suivante :

$$\ln Z_y = -\ln\left(\int_{\boldsymbol{x}} \exp\left(\sum_m \nu_m \varphi_m(\boldsymbol{x}, y)\right) d\boldsymbol{x}\right).$$

On a donc pour tout $y \in \mathcal{Y}$:

$$\frac{\partial \ln Z_y}{\partial \nu_m} = -E(\varphi_m(\mathbf{X}, y)/Y = y).$$

La dérivée de la log-vraisemblance selon chaque paramètre ν_m, à y fixé, a donc la forme suivante :

$$\begin{aligned} \frac{\partial L_{\underline{y}}(\nu, \underline{\boldsymbol{x}})}{\partial \nu_m} &= \sum_y N_y \frac{\partial \ln Z_y}{\partial \nu_m} + \sum_y \sum_{i=1}^{N_y} \varphi_m(\boldsymbol{x}_i, y) \\ &= -\sum_y N_y E(\varphi_m(\mathbf{X}, y)/Y = y) + N\overline{\varphi_m}\ , \ m \in \{1, \ldots, M\}\ . \end{aligned}$$

Les équations du maximum de vraisemblance pour la loi considérée sont donc :

$$\sum_y \sum_{i=1}^{N_y} \frac{N_y}{N} E(\varphi_m(\mathbf{X}, y)/Y = y) = \overline{\varphi_m}\ , \ m \in \{1, \ldots, M\}\ .$$

□

Ainsi, les équations du PME et les **équations du maximum de vraisemblance** obtenues à partir d'une vraisemblance rétrospective, pour une loi issue de la structure $\mathcal{S}_M$, sont les mêmes, à la condition que l'information fixe la loi marginale de Y à la loi empirique (quel que soit le type de l'échantillon).

▷ **Remarque**

Si la loi de maximum d'entropie de Y n'est pas la loi empirique, rien n'assure que les équations du PME soient les mêmes que les équations de la MMV pour une loi de même structure que la loi de maximum d'entropie. Mais nous avons vu au chapitre (1) les risques qui pouvaient apparaître en mélant niveaux empiriques et niveaux fixés a priori.

2.5 Relecture de modèles classiques

Dans le cadre de covariables continues, nous venons donc de voir que les modèles de discrimination classique d'ADL et d'ADQ sont naturellement associés à certaines familles d'observables. Les propriétés des lois conditionnelles caractérisant ces méthodes d'analyse discriminante peuvent alors être vues comme des *conséquences* des observables pris en compte, et non plus comme des *hypothèses* à la base du modèle.

Ainsi, en considérant la relecture par le PME de l'ADL, on peut voir dans l'égalité des matrices de variance-covariance conditionnellles non pas une hypothèse de modélisation forte, mais la conséquence de n'avoir introduit aucune information à propos de l'influence du groupe sur la dispersion des variables explicatives.

De même, l'utilisation d'une ADL basée sur une matrice de variance-covariance Σ diagonale correspond à une information très parcimonieuse, au sens où elle ne prend en compte que les deux premiers moments de chaque variable. Utiliser une matrice pleine ne vient plus du choix de ne *pas* faire d'hypothèses d'indépendance entre les variables, mais correspond à la possibilité (ou la volonté au vu des données) de prendre en compte des informations supplémentaires sur les interactions entre celles-ci.

Nous ne nous sommes pas interrogés jusqu'à présent sur la nature des niveaux de contraintes, car seule la forme des observables influence la forme des lois obtenues. Par contre, la nature des niveaux de contraintes rentre en jeu dans la détermination des multiplicateurs. Ces paramètres “naturels” des lois de maximum d'entropie sont entièrement déterminés par la résolution des équations du PME. Mais nous avons vu qu'il existait pour les informations considérées ici des relations simples entre multiplicateurs et paramètres au sens usuel (moyennes, variances, covariances, ...). Ce sont ces paramètres usuels que nous chercherons à déterminer.

En restant dans la logique d'application du PME présentée au chapitre (1), nous ne considérerons que des niveaux de contraintes déterminés à partir d'un jeu de données. Nous allons à présent détailler l'influence que peut avoir la nature du schéma d'échantillonnage ayant permis de récolter ces données sur ces niveaux de contraintes.

3 Discrimination par PME et nature de l'échantillon

Dans l'approche habituelle des modèles d'analyse discriminante, le problème de la nature de l'échantillon ne se pose qu'au moment de l'établissement de la règle de décision ([Celeux 90, McLachlan 92, Titterington *et al.* 96]). L'hypothèse que la loi de groupe n'apporte aucune information sur les paramètres des lois conditionnelles assure en effet que l'estimation des paramètres par maximum de vraisemblance se base sur les mêmes équations, quel que soit le type de l'échantillon disponible (cf. la relation II.4).

Nous nous plaçons ici dans le cadre d'une population caractérisée par des covariables à valeurs dans $\mathcal{X} \subseteq \mathbf{R}^J$.

3.1 Utilisation d'une information empirique

Dans l'approche de l'analyse discriminante par PME, et dans le cas où les niveaux de contraintes sont empiriques, la nature de l'échantillon doit être prise en compte avant même la construction du modèle, dès le choix des observables. En effet, il convient de s'interroger sur la pertinence de l'utilisation de niveaux de contraintes qui tirent du jeu de données une information globale. Comme nous avons fait le choix d'utiliser uniquement des niveaux de contraintes empiriques si tous ne pouvaient pas être choisis a priori, il devient impossible de conserver dans l'information un observable dont le niveau de contrainte est inutilisable.

Nous nous placerons ici dans le cadre de l'information $\mathcal{I}_J^c$, pour des niveaux de contraintes tous estimés empiriquement, à partir d'un jeu de données de taille N noté $(\underline{\boldsymbol{x}}, \underline{y}) = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$. En désignant par $\overline{\psi}$ la moyenne empirique de l'observable ψ, et par N_y le nombre d'individus observés dans le groupe G_y, $y \in \{0, 1\}$, l'information étudiée est donc la suivante :

$$\overline{\mathcal{I}_J^c} = \left\{(x_j, \overline{x_j}), (x_j^2, \overline{x_j^2}), (x_j y, \overline{x_j y}), j \in \{1, \ldots, J\}, (y, \overline{y})\right\},$$

où, pour $j \in \{1, \ldots, J\}$:

$$\left\{\begin{array}{ccccc} \overline{x_j} & = & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_{ij} & & \\ \overline{x_j^2} & = & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_{ij}^2 & & \\ \overline{x_j y} & = & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_{ij} y & = & \dfrac{N_1}{N}\overline{x_j}^1 \\ \overline{y} & = & \dfrac{N_y}{N} & & \end{array}\right.$$

en notant $\overline{x_j}^y$ la moyenne des observations de X_j dans le groupe G_y, $y \in \{0, 1\}$.

Rappelons qu'il existe une loi de maximum d'entropie correspondant à cette information. Cette loi possède la même expression quelle que soit la nature des niveaux de contraintes (lois conditionnelles de $\boldsymbol{X}$ normales de même matrice de variance-covariance, et loi de Y fixée par l'information).

Les covariables étant à valeurs dans un sous-ensemble de $\mathbf{R}^J$, nous ne considérerons pas le cas d'un échantillon prospectif (pour lequel les fréquences des observations en $\boldsymbol{X}$ seraient fixées a priori).

3.2 Échantillon rétrospectif : utilisation d'informations conditionnelles

Un échantillon rétrospectif est en fait constitué de deux échantillons dont chacun apporte des informations aléatoires sur le comportement des covariables dans le groupe dont il est issu. Tout résumé “conditionnel” des données à l'intérieur de chaque groupe est donc une image du comportement global à l'intérieur de ce groupe. On peut par contre s'interroger sur le lien entre un paramètre global et l'image qui en est donnée par son estimation à partir d'un tel échantillon.

Ainsi, la moyenne $\overline{x_j}$, $j \in \{1, \dots, J\}$, est donnée par : $\overline{x_j} = N_0/N\overline{x_j}^0 + N_1/N\overline{x_j}^1$, où N_0/N et N_1/N sont des valeurs fixées a priori dans le schéma d'échantillonnage. Cette moyenne n'est donc pas une image de $E(X_j)$ si les proportions N_y/N, $y \in \{0, 1\}$, sont très éloignées de celles que l'on aurait obtenues à partir d'un échantillon de mélange.

Ce problème peut se poser pour tout observable $\psi(\boldsymbol{x})$ ou $\psi(\boldsymbol{x})y$ quand on leur associe leur niveau de contrainte empirique. En fait, seule l'utilisation d'informations conditionnelles n'implique pas de problème d'interprétation.

Il apparaît donc que, dans le cas d'un tel échantillon, la démarche adéquate serait de construire un modèle de maximum d'entropie décrivant le comportement des covariables dans chaque groupe.

Par exemple, on peut considérer une information minimale (basée sur les deux premiers moments de chaque variable X_j, $j \in \{1, \dots, J\}$), et calculer les niveaux empiriques pour chacun des échantillons. On obtient alors un modèle par groupe, en l'occurence deux lois normales, de moyenne et matrice de variance-covariance différentes. Le schéma suivant résume cette démarche :

On choisit l'information

$$\mathcal{I} = \{(x_j, \overline{x_j}), (x_j^2, \overline{x_j^2}), j \in \{1, \dots, J\}\}$$

$$\Downarrow$$

Loi de $\boldsymbol{X}$ sachant $Y = y$, $y \in \{0, 1\}$:
$\mathcal{N}(m_y, \Sigma_y)$

Modélisation de $\boldsymbol{X}$

$\swarrow$ $\searrow$

Dans le groupe G_0 :	Dans le groupe G_1 :
$\mathcal{N}(\overline{\boldsymbol{x}}^0, \overline{\Sigma}^0)$	$\mathcal{N}(\overline{\boldsymbol{x}}^1, \overline{\Sigma}^1)$

(Où $\overline{\boldsymbol{x}}^y$ et $\overline{\Sigma}^y$ désignent la moyenne et la matrice de variance-covariance empiriques des observations en $\boldsymbol{X}$ dans le groupe G_y, $y \in \{0, 1\}$.)

Cette solution ne rentre pas en contradiction avec le but initial qui est la caractérisation de la distribution de variables explicatives conditionnellement à l'appartenance aux groupes de la population. Ainsi, on ne peut donner comme but au PME que celui de fournir des lois conditionnelles de ces covariables à partir de l'information disponible. Si on se trouve dans l'obligation d'utiliser un échantillon pour déterminer les niveaux de contraintes, et si cet échantillon ne fournit pas une image aléatoire globale de la répartition de la population, les seules informations disponibles ne pourront être que conditionnelles.

S'il s'agit de plus de déterminer une règle d'affectation de nouveaux individus, le problème reste le même que dans l'approche classique : la loi empirique de Y issue d'un échantillon rétrospectif n'est pas la réalisation d'un tirage aléatoire. Il faut donc choisir une stratégie de classement, par exemple l'utilisation d'une loi équirépartie pour Y.

3.3 Échantillon de mélange : peut-on simplifier l'information?

Dans le cas d'un échantillon de mélange, tout niveau de contrainte empirique $\overline{\psi}$ ou $\overline{\psi y}$ est bien une image aléatoire de l'espérance $E(\psi(\mathbf{X}))$ ou $E(\psi(\mathbf{X})Y)$.

Dans le cas d'un échantillon de petite taille, il pourrait être intéressant de réduire le nombre de niveaux de contraintes pour éviter de trop en “demander” à l'échantillon. L'information présentant le minimum d'observables simples parmi celles présentées dans les tableaux II.5, p. 91, et II.6, p. 92, est l'information $\mathcal{I}_J^c$ (définie à partir des observables $(x_j, x_j^2, x_j y)_{j\in\{1,\ldots,J\}}$ et y). Les simplifications possibles sont les suivantes :

suppression des observables x_j^2, $j \in \{1,\ldots,J\}$: il n'y a plus alors de lois de maximum d'entropie car l'expression $\exp\left(\sum_j(\lambda_j + \lambda'_j y)x_j\right)$ n'est pas intégrable sur $\mathbf{R}^J$ (mais le sera sur un produit d'intervalles bornés) ;

suppression des observables x_j, $j \in \{1,\ldots,J\}$: la loi conditionnelle de maximum d'entropie de $\mathbf{X}$ dans le groupe G_0 est alors une loi normale de moyenne le vecteur nul $(0,\ldots,0)$. Le PME “choisit” alors une loi symétrique en $(0,\ldots,0)$ dans ce groupe car il n'y a plus d'information de tendance centrale parmi les éléments de connaissance ;

suppression des observables x_j et $x_j y$, $j \in \{1,\ldots,J\}$: dans ce cas, les deux lois conditionnelles obtenues par PME seront des lois normales multidimensionnelles centrées et de mêmes matrice de variance-covariance. Il y a là aussi un manque d'information minimale sur les tendances centrales des covariables, quel que soit le groupe, ce qui aboutit à obtenir deux mêmes lois centrées. Les informations sur la dispersion des covariables sont globales et impliquent donc une même matrice de variance-covariance ;

suppression de l'observable y : en supprimant cet observable, on est confronté au problème de la recherche d'un modèle de maximum d'entropie pour des variables de nature différente, sans que la loi des variables discrètes soit fixée par l'information. On sort alors du cadre de la proposition (3) du chapitre (1).

Nous allons cependant détailler cette dernière situation. On peut en effet se demander s'il est tout de même possible, dans ce cas, d'obtenir un modèle de maximum d'entropie, c'est-à-dire des lois conditionnelles des covariables et une loi marginale de Y répondant aux contraintes.

4 Discrimination par PME sans fixer la loi de groupe

Nous commencerons par présenter une proposition établissant l'existence et l'expression d'un tel modèle. Mais son application pour une information élémentaire fera apparaître un paradoxe flagrant au niveau de la frontière de discrimination alors obtenue.

4.1 Proposition préliminaire

Nous nous plaçons dans le cas général d'une population à K groupes, pour une information mettant en jeu les deux types d'observables suivant :

- d'une part des observables $\psi_m(\boldsymbol{x})$, $m \in \{1, \ldots, M\}$, portant uniquement sur les covariables continues $\mathbf{X} = (X_1, \ldots, X_J)$ à valeurs dans $\mathcal{X} \subseteq \mathbf{R}^J$,
- d'autre part des observables $\varphi_{m'}(\boldsymbol{x})y$, $r \in \{1, \ldots, M'\}$, liant des fonctions des covariables (les variables elles-mêmes, leurs moments, des interactions, ...) à la variable de groupe Y.

On note $\mathcal{I}$ l'information définie à partir de ces observables :

$$\mathcal{I} = \{(\psi_m(\boldsymbol{x}), a_m)_{m \in \{1,\ldots,M\}}, (\varphi_{m'}(\boldsymbol{x})y, b_{m'})_{m' \in \{1,\ldots,M'\}}\}, \tag{II.13}$$

où $(a_1, \ldots, a_M) \in \mathbf{R}^M$ et $(b_1, \ldots, b_{M'}) \in \mathbf{R}^{M'}$ sont les niveaux de contraintes associés aux observables.

On note P^* la loi conjointe qui se décompose en loi marginale de Y et lois conditionnelles de $\mathbf{X}$ dans chaque groupe prenant les formes suivantes :

Loi conditionnelle de X sachant $Y = y$, $y \in \mathcal{Y}$: nous supposons que pour toute valeur y fixée, les observables $(\psi_m)_{m \in \{1,\ldots,M\}}$ et $(\varphi_{m'}(\boldsymbol{x})y)_{m' \in \{1,\ldots,M'\}}$ permettent d'écrire une densité sur $\mathcal{X}$ vérifiant les contraintes qui leur sont associées.

On note cette densité f_X^{*y}, et on a :

$$f_X^{*y}(x) = Z_y \exp\left(\sum_m \lambda_m \psi_m(\boldsymbol{x}) + \sum_{m'} \lambda'_{m'} \varphi_{m'}(\boldsymbol{x}) y\right) \ , \ x \in \mathcal{X}, \tag{II.14}$$

où $(\lambda_1, \ldots, \lambda_M) \in \mathbf{R}^M$ et $(\lambda'_1, \ldots, \lambda'_{M'}) \in \mathbf{R}^{M'}$ sont les multiplicateurs associés aux observables et $Z_y^{-1} = \int_x \exp\left(\sum_m \lambda_m \psi_m(\boldsymbol{x}) + \sum_{m'} \lambda'_{m'} \varphi_{m'}(\boldsymbol{x}) y\right) d\boldsymbol{x}$ est la constante de normalisation.

Loi marginale de Y : $p_y^* = \dfrac{Z}{Z_y}$, où $Z^{-1} = \sum_y \dfrac{1}{Z_y}$ est la constante de normalisation de cette loi.

Proposition 10 *Avec les notations précédentes :*
La loi P^* est l'unique loi conjointe de $(\mathbf{X}, Y)$ qui réalise le maximum de l'entropie S parmi toutes les lois vérifiant les contraintes imposées par l'information $\mathcal{I}$.

Démonstration

Rappelons que pour toute loi P, décomposée sous la forme d'une loi marginale discrète et de lois conditionnelles continues $(P_Y, (f_X^y)_{y \in \mathcal{Y}})$, qui vérifie les contraintes associées à l'information $\mathcal{I}$, on peut écrire les deux séries d'équations suivantes :

$$\begin{aligned} a_m &= \textstyle\sum_y p_y \int_x \psi_m(\boldsymbol{x}) f_X^y(\boldsymbol{x}) d\boldsymbol{x} \quad , \ m \in \{1, \ldots, M\} \\ \text{et} \quad b_{m'} &= \textstyle\sum_y y p_y \int_x \varphi_{m'}(\boldsymbol{x}) f_X^y(\boldsymbol{x}) d\boldsymbol{x} \quad , \ m' \in \{1, \ldots, M'\}. \end{aligned}$$

Pour toute loi de ce type, l'information $I(P) = -S(P)$ peut se décomposer de la manière suivante :

$$\begin{aligned} I(P) &= E_{P_Y}(I(f_X^Y)) + I(P_Y) \\ &= E_{P_Y}\left(E_X^Y\left(\ln \frac{f_X^Y}{f_X^{*Y}}\right) + E_X^Y\left(\ln f_X^{*Y}\right)\right) + E_{P_Y}\left(\ln \frac{p_Y}{p_Y^*}\right) + E_{P_Y}(\ln p_Y^*). \end{aligned}$$

En utilisant les contraintes sur P et les expressions de P_Y^* et f_X^{*y}, on obtient :

$$\begin{aligned} E_{P_Y}\left(E_X^Y\left(\ln f_X^{*Y}\right)\right) + E_{P_Y}(\ln p_Y^*) &= \sum_y p_y \ln\left(\frac{Z}{Z_y}\right) + \\ &\sum_y p_y \int_x \left(\ln Z_y + \sum_m \lambda_m \psi_m(\boldsymbol{x}) + \sum_{m'} \lambda'_{m'} \varphi_{m'}(\boldsymbol{x}) y\right) f_X^y(\boldsymbol{x}) d\boldsymbol{x} \\ &= \sum_y p_y \ln Z_y + \sum_m \lambda_m a_m + \sum_{m'} \lambda'_{m'} b_{m'} + \ln Z \\ &\quad - \sum_y p_y \ln Z_y \\ &= \sum_m \lambda_m a_m + \sum_{m'} \lambda'_{m'} b_{m'} + \ln Z \ . \end{aligned}$$

Cette dernière expression étant une constante, minimiser $I(P)$ revient donc à minimiser uniquement la somme suivante :

$$E_{P_Y}\left(E_X^Y\left(\ln\frac{f_X^Y}{f_X^{*Y}}\right)\right)+E_{P_Y}\left(\ln\frac{p_Y}{p_Y^*}\right).$$

La proposition (2) du chapitre (1) assure que le minimum de $E_X^Y\left(\ln(f_X^Y/f_X^{*Y})\right)$ est atteint pour $f_Y^y = f_X^{*y}$, $y \in \mathcal{Y}$, et que celui de $E_{P_Y}\left(\ln(p_Y/p_Y^*)\right)$ est atteint pour $P_Y = P_Y^*$.

Le minimum d'une somme ne peut pas être inférieur à la somme des minimas. Le minimum de $I(P)$ est donc atteint pour la loi P^* définie précédemment.

□

Il est donc possible d'obtenir une loi de maximum d'entropie à partir de l'information $\mathcal{I} = \{(\psi_m(\boldsymbol{x}), a_m)_{m\in\{1,\dots,M\}}, (\varphi_{m'}(\boldsymbol{x})y, b_{m'})_{m'\in\{1,\dots,M'\}}\}$, si la nature des fonctions ψ_m et $\varphi_{m'}$ permet d'écrire les densités conditionnelles pour toute valeur y fixée de Y.

Cette information *ne fixe pas la loi de* Y, mais on obtient une expression de celle-ci, à partir de la constante de normalisation des densités conditionnelles de $\boldsymbol{X}$.

4.2 Utilisation d'une information élémentaire

Les densités conditionnelles II.14 ont la même forme exponentielle que celles qu'on aurait obtenues à partir de la même information $\mathcal{I}$ augmentée d'observables fixant la loi de Y. Seules les valeurs des multiplicateurs sont modifiées, car les équations du PME sont différentes selon que l'on travaille à loi de Y fixée ou non.

Nous pouvons à présent appliquer la propriété précédente aux informations des tableaux II.5, p. 91, et II.6, p. 92, desquelles on aura oté l'observable y. Les lois conditionnelles alors obtenues ont donc la même forme, que cet observable soit présent ou non. Il reste à exprimer la loi marginale de Y de maximum d'entropie.

Nous allons détailler l'expression des probabilités d'appartenance aux groupes obtenues par PME pour l'information élémentaire suivante :

$$\widetilde{\mathcal{I}_J^c} = \{(x_j, a_j), (x_j^2, b_j), (x_j y, c_j), j \in \{1,\dots,J\}\}. \tag{II.15}$$

Les lois de maximum d'entropie conditionnelles sont donc dans ce cas des lois normales multidimensionnelles $\mathcal{N}(m_y, \Sigma)$, où Σ est une matrice diagonale.

Les probabilités marginales en y, $y \in \{0, 1\}$, s'écrivent à partir des constantes de normalisation des densités f_X^{*y}, exprimées sous la forme d'une combinaison linéaire des observables en $\boldsymbol{x}$ et en $(\boldsymbol{x}, y)$:

$$p_y^* = Z \int_x \exp\left(\sum_j \left((\lambda_j + \lambda'_j y)x_j + \gamma_j x_j^2\right)\right), y \in \{0,1\},$$

où Z est la constante de normalisation de P_Y^*.

En utilisant le lien entre multiplicateurs et paramétrage usuel des lois conditionnelles exprimées comme des lois normales multidimensionnelles $\mathcal{N}(m_y, \Sigma)$, $y \in \{0,1\}$ (cf. 2.3.2, p. 88), cette expression se simplifie comme suit :

$$\begin{aligned} p_y^* &= Z \exp\left(-\frac{1}{4} m'_y \Sigma^{-1} m_y\right) \int_x \exp\left(-\frac{1}{2}(\boldsymbol{x} - \Lambda_y)' \Gamma^{-1} (\boldsymbol{x} - \Lambda_y)\right) d\boldsymbol{x} \\ &= Z \sqrt{2\pi}^J \sqrt{|\Sigma|} \exp\left(-\frac{1}{4} \Lambda'_y \Gamma^{-1} \Lambda_y\right) . \end{aligned}$$

L'égalité de la matrice de variance-covariance pour les deux lois conditionnelles permet de simplifier l'expression de la loi marginale obtenue à partir de l'information II.15. Les deux probabilités d'appartenance aux groupes obtenues par application du PME peuvent s'exprimer soit en faisant intervenir l'ensemble des multiplicateurs associés aux contraintes, soit en faisant apparaître le paramétrage usuel des lois conditionnelles :

$$\text{soit} \begin{cases} p_0^* = \dfrac{1}{1 + \exp\left(-1/4\left(\Lambda'_1 \Gamma^{-1} \Lambda_1 - \Lambda'_0 \Gamma^{-1} \Lambda_0\right)\right)} \\ p_1^* = \dfrac{1}{1 + \exp\left(1/4\left(\Lambda'_1 \Gamma^{-1} \Lambda_1 - \Lambda'_0 \Gamma^{-1} \Lambda_0\right)\right)}, \end{cases} \tag{II.16}$$

$$\text{soit} \begin{cases} p_0^* = \dfrac{1}{1 + \exp\left(1/2\,(m_1 - m_0)'\,\Sigma^{-1}\,(m_1 + m_0)\right)} \\ p_1^* = \dfrac{1}{1 + \exp\left(-1/2\,(m_1 - m_0)'\,\Sigma^{-1}\,(m_1 + m_0)\right)}. \end{cases} \tag{II.17}$$

4.3 Comportement des équations du PME : cas d'une covariable

Le cas simple d'une unique variable explicative X à valeurs dans $\mathcal{X} \subseteq \mathbf{R}$ illustrera notre propos. Nous nous limiterons à l'étude de l'information suivante :

$$\widetilde{\mathcal{I}_1} = \{(x, a), (x^2, b), (xy, c)\}\ , \text{ où } (a, b, c) \in \mathbf{R} \times \mathbf{R}^+ \times \mathbf{R}. \tag{II.18}$$

À partir de cette information, nous avons vu que l'on obtient comme loi de maximum d'entropie de X sachant $Y = y$ une loi normale $\mathcal{N}(m_y, \sigma^2)$, $y \in \{0,1\}$

(loi II.8). Les contraintes s'écrivent en exprimant les espérances en fonction des paramètres des lois conditionnelles et des probabilités a priori d'appartenance aux groupes :

$$\left\{\begin{array}{lllll} a & = & E(X) & = & \pi_0 m_0 + \pi_1 m_1 \\ b & = & E(X^2) & = & \pi_0(m_0^2+\sigma_0^2) + \pi_1(m_1^2+\sigma_1^2) \\ c & = & E(XY) & = & \pi_1 m_1 \, . \end{array}\right. \tag{II.19}$$

On résoud ici ce système pour obtenir les valeurs m_0^*, m_1^*, σ^{*2} (pour cette loi de maximum d'entropie, on a $\sigma_0^{*2} = \sigma_1^{*2} = \sigma^{*2}$) et p_1^* (et non les valeurs des multiplicateurs). On utilise l'expression obtenue à partir des équations II.17 des probabilités p_0^* et p_1^* :

$$\left\{\begin{array}{lll} p_0^* & = & \dfrac{1}{1+\exp\left(\dfrac{(m_1^{*2}-m_0^{*2})}{2\sigma^{*2}}\right)} \\ p_1^* & = & \dfrac{1}{1+\exp\left(-\dfrac{(m_1^{*2}-m_0^{*2})}{2\sigma^{*2}}\right)} \, . \end{array}\right. \tag{II.20}$$

Notons que l'on a d'une part $p_1^* m_1^* = c$ et d'autre part $p_0^* m_0^* = a - c$ (première et troisième équation du système II.19). Cela permet d'obtenir l'égalité suivante :

$$\exp\left(\frac{m_1^{*2}-m_0^{*2}}{2\sigma^{*2}}\right) = \frac{m_0^*}{a-c} - 1 = \frac{c}{m_1^*-c} \, .$$

On peut donc écrire m_0^* en fonction de m_1^* et des niveaux de contraintes :

$$m_0^* = \frac{m_1^*(a-c)}{m_1^*-c}.$$

En utilisant la deuxième équation du système II.19, on peut à présent exprimer σ^{*2} en fonction de m_1^* et des niveaux de contraintes :

$$\sigma^{*2} = b - cm_1^* - m_1^*\frac{(a-c)^2}{m_1^*-c} \, .$$

En injectant ces deux expressions dans $p_1^* m_1^* = c$, on obtient finalement le système simplifié suivant :

$$\left\{\begin{array}{lll} m_1^* & = & c\left(1+\exp\left(\dfrac{m_1^{*2}(a-m_1^*)(m_1^*+a-2c)}{2(m_1^*-c)\left((b-cm_1)(m_1^*-c)-m_1^*(a-c)^2\right)}\right)\right) \\ m_0^* & = & \dfrac{m_1^*(a-c)}{m_1^*-c} \\ \sigma^{*2} & = & b-cm_1^*-m_1^*\dfrac{(a-c)^2}{m_1^*-c} \, . \end{array}\right. \tag{II.21}$$

Il y a donc deux étapes pour obtenir les valeurs des paramètres des lois conditionnelles de X sachant $Y = y$, $y \in \{0, 1\}$:

- on résoud une équation du type $f_{a,b,c}(z) = z$ sur $\mathbf{R}$ pour obtenir la valeur de m_1^* (où $f_{a,b,c}$ est une fonction dont l'expression dépend des niveaux de contraintes a, b et c),
- on injecte la valeur m_1^* obtenue dans les deux autres équations du système II.21 pour obtenir les valeurs m_0^* et σ^{*2}.

• **Cas particulier :** $a = 0$

Le cas particulier simple $a = 0$ permet de détailler le mécanisme de ces équations de maximum d'entropie qui se situent dans un cadre *non linéaire*, mais qu'il est néammoins possible de résoudre.

Le système d'équations II.21 devient alors :

$$\left\{\begin{array}{lcl} m_1^* & = & c\left(1+\exp\left(\dfrac{-m_1^{*3}(m_1^*-2c)}{2(m_1^*-c)\left(b(m_1-c)-cm_1^{*2}\right)}\right)\right) \\ m_0^* & = & \dfrac{cm_1^*}{c-m_1^*} \\ \sigma^2 & = & b-\dfrac{cm_1^{*2}}{m_1^*-c}\,. \end{array}\right. \tag{II.22}$$

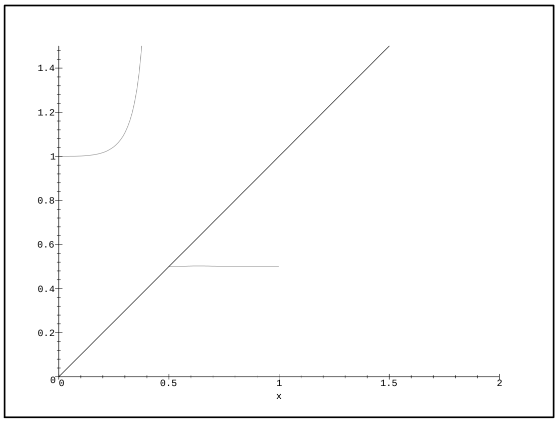
La fonction de $\mathbf{R}$ dans $\mathbf{R}$ dont il faut déterminer le point fixe (s'il existe) pour obtenir l'expression des paramètres des lois de maximum d'entropie a donc la forme suivante :

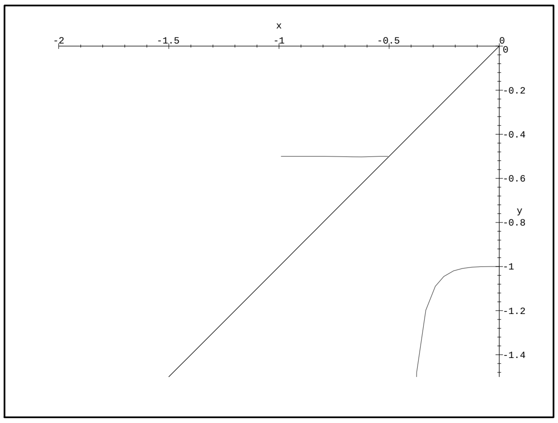
$$f_{0,b,c}(z) = c\left(1+\exp\left(\frac{-z^3(z-2c)}{2(z-c)\left(b(z-c)-cz^2\right)}\right)\right). \tag{II.23}$$

Il apparaît que $m_1^* = 2c$ en est une solution évidente. Nous n'étudierons pas ici l'unicité de cette solution, mais les figures II.1, p. 108, et II.2, p. 109, présentent quelques fonctions $f_{0,b,c}$ et font apparaître graphiquement que cette valeur est l'unique point fixe.

Le reste du système II.22 permet d'obtenir les différents paramètres des lois conditionnelles et les valeurs des probabilités d'appartenance aux groupes :

$$\left\{\begin{array}{lcl} m_1^* & = & 2c \\ m_0^* & = & -2c \\ \sigma^{2*} & = & 1-4c^2 \end{array}\right. \qquad \text{et} \qquad p_0^* = p_1^* = 1/2.$$





		0.5		0.64		1		2.36	
$f'_{0,1,0.5}$	$+$	‖	$+$		$-$	‖	$-$		$+$
$f_{0,1,0.5}$	$0.5 \nearrow +\infty$	‖	$0.5 \nearrow$	0.502	$\searrow 0.5$	‖	$+\infty \searrow$	90	$\nearrow +\infty$

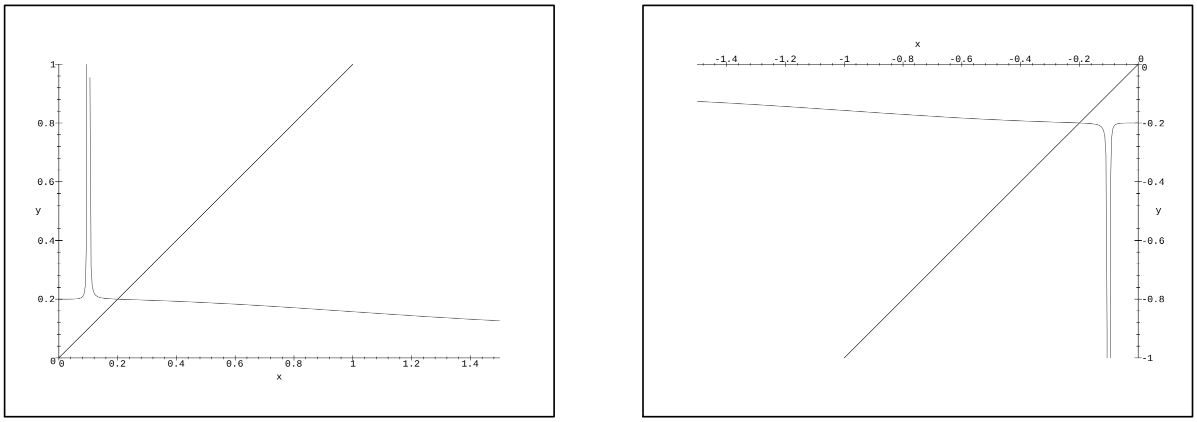
		-1.64		-1		-0.609		-0.5	
$f'_{0,1,-0.5}$	$+$		$-$	‖	$-$		$+$	‖	$+$
$f_{0,1,-0.5}$	$-\infty \nearrow$	-211.64	$\searrow -\infty$	‖	$-0.5 \searrow$	-0.502	$\nearrow -0.5$	‖	$-\infty \nearrow -0.5$

Les valeurs 0.64 *et* 2.36 *sont les solutions de l'équation* $2z^2 - 6z + 3 = 0$ *qui donne le signe de* $f'_{0,1,0.5}$

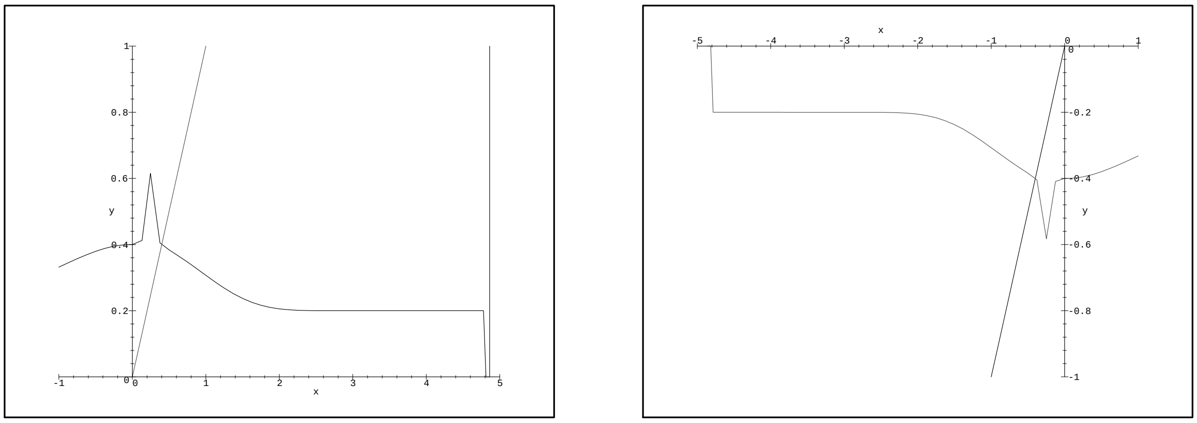
Les valeurs -1.64 *et* -0.609 *sont les solutions de l'équation* $4z^2 + 9z + 4 = 0$ *qui donne le signe de* $f'_{0,1,-0.5}$

FIG. II.1 – *Visualisation des points fixes des fonctions* $f_{0,1,0.5}$ *et* $f_{0,1,-0.5}$

$c = \pm 0.1$ *: $f_{0,1,c}$ non-définies en* ± 0.1, ± 0.101 *et* ± 9.89



$c = \pm 0.2$ *: $f_{0,1,c}$ non-définies en* ± 0.2, ± 0.209 *et* ± 4.791



$c = \pm 0.3$ *: $f_{0,1,c}$ non-définies en* ± 0.3, ± 0.33 *et* ± 3

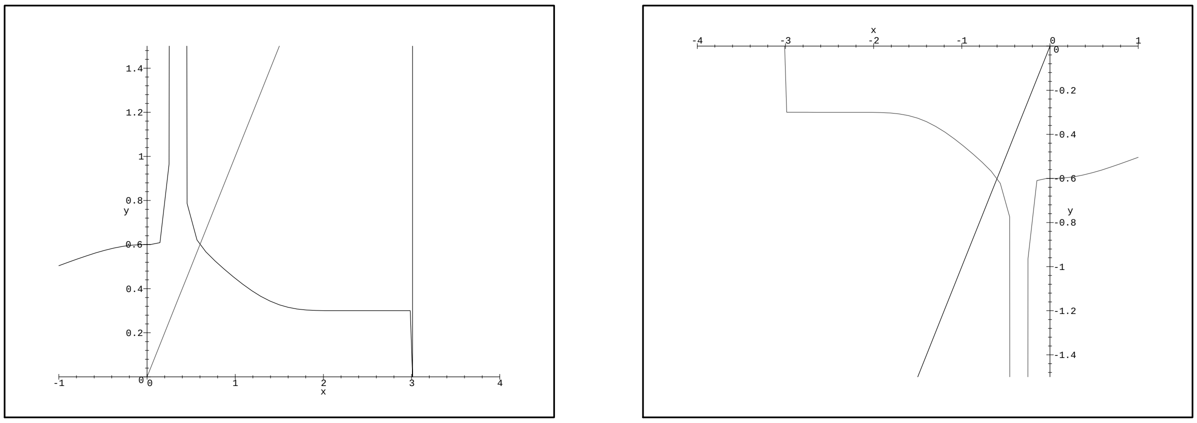


FIG. II.2 – *Visualisation de quelques points fixes de fonctions $f_{0,1,c}(z)$, pour $c < 1/4$*

La condition de positivité de σ^{*2} impose la condition d'existence d'une loi de maximum d'entropie. Il s'agit ici d'une condition sur le niveau de contrainte c qui doit vérifier : $c^2 \leq 1/4$. D'autre part, $2c$ ne peut pas être solution de II.23 si cette valeur n'appartient pas au domaine de définition de la fonction $f_{0,b,c}$. C'est le cas quand $c = 0$ et quand $1 - 4c^2 = 0$. Il n'y a donc de solution aux équations du PME que si $c \in]\text{-}1/2, 1/2[- \{0\}$.

Dans ce cas particulier ($a = 0$ et $c \in]\text{-}1/2, 1/2[- \{0\}$), la solution de maximum d'entropie amène, si elle existe, à utiliser la loi équirépartie comme loi de la variable de groupe Y.

▷ **Remarque**

Ce cas particulier ($a = 0$) ne doit pas être confondu avec la recherche d'un modèle de maximum d'entropie sur des données préalablement centrées. En effet, après un tel changement, on ne travaillerait plus sur les variables d'origine (X, Y) mais sur des variables (X', Y) à valeurs dans un nouvel espace $\mathcal{X}' \times \{0, 1\}$ et pour lesquelles il faudrait écrire un nouveau système de contraintes.

• **Sous quelles conditions obtient-on une loi de Y équirépartie?**

L'équation $f_{a,b,c}(2c) = 2c$ est vérifiée sous la condition $a(a - 2c) = 0$, sans aucune condition sur le niveau de contrainte b. On obtiendra donc comme loi de Y issue du PME la loi équirépartie dans les deux conditions suivantes :

– soit $a = 0$, ce qui implique $\left\{ \begin{array}{lcl} m_1^* & = & 2c \\ m_0^* & = & -2c \\ \sigma^{2*} & = & b - 4c^2 \end{array} \right.$

(c'est-à-dire une population constituée de deux groupes de même variance, et de moyennes opposées : c'est le cadre d'une ADL qui devrait utiliser une frontière de discrimination d'équation $x = 0$) ;

– soit $a = 2c$, ce qui implique $\left\{ \begin{array}{lcl} m_1^* & = & 2c = a \\ m_0^* & = & 2c = a \\ \sigma^{2*} & = & b - 4c^2 \end{array} \right.$

(c'est-à-dire une population constituée de deux groupes de même variance, et de même moyenne : c'est le cas trivial où aucune analyse discriminante n'est possible, puisque les deux groupes sont tout simplement confondus).

La valeur de σ^{*2} alors obtenue est identique dans les deux cas et fournit la condition d'existence d'une loi de maximum d'entropie : une telle loi n'existe que si c appartient à l'intervalle $\left]-\sqrt{b}/2, \sqrt{b}/2\right[$.

Réciproquement, on peut constater que la probabilité p_1^* ne sera égale à $1/2$ que dans le cas où les paramètres obtenus par résolution des équations du PME II.19 vérifient soit $m_0^{*2} = m_1^{*2}$ exactement, soit $m_1^{*2} - m_0^{*2} \ll \sigma^{*2}$.

En utilisant l'expression de m_0^* en fonction de m_1^*, il apparaît que la condition $m_0^{*2} = m_1^{*2}$ est réalisée dans deux cas :

- soit $m_1^* = a$, ce qui est réalisé quand $a = 2c$;
- soit $m_1^* = 2c - a$, ce qui est réalisé quand $a = 0$ (on retrouve alors $m_1^* = 2c$).

Ainsi, la loi de Y obtenue par maximum d'entropie ne sera la loi équirépartie que dans les deux cas suivant : $a = 2c$ ou $a = 0$, sans autre condition sur le niveau de contrainte correspondant à l'observable x^2. Pour ces deux niveaux de contraintes, le PME "choisit" comme loi de groupe la loi qui ne privilégie aucun des deux groupes de la population.

$\triangleright$ **Remarque**

Ces deux situations illustrent le fait qu'une loi de mélange de deux lois normales, notée $\mathcal{Q} = \pi_0\mathcal{N}(\mu_0, \sigma^2) + \pi_1\mathcal{N}(\mu_1, \sigma^2)$, n'est pas forcément la loi de maximum d'entropie correspondant à la famille d'observables (x, x^2, xy). Autrement dit, la loi de PME issue de l'information $\mathcal{I} = \{(x, \overline{x}), (x^2, \overline{x^2}), (xy, \overline{xy})\}$ ne converge pas vers $\mathcal{Q}$ quand la taille de l'échantillon N tend vers l'infini, pour des données issues d'un échantillon de mélange.

• **Observations sur le cas général ($(a, c) \in \mathbf{R}^2$, $b \in \mathbf{R}^+$)**

Nous n'étudierons pas en détail le système II.21, p. 106, mais nous nous contenterons de présenter ici quelques remarques sur le comportement des équations du PME qui correspondent à l'information II.18, en particulier au travers de la capacité que pourraient avoir ces équations à "retrouver" les proportions des groupes dans la population.

Remarquons seulement qu'il est possible d'obtenir les conditions d'existence d'une loi de maximum d'entropie en considérant la condition de positivité de σ^{*2}.

Pour illustrer ce comportement, nous pouvons résoudre le système II.21 pour des niveaux de contraintes calculés à partir de valeurs données a priori des paramètres m_0, m_1, σ^2 dans la population et de différentes probabilité a priori $\pi_1 \in]0, 1[$. Ces essais peuvent être vus comme une illustration de situations où on ne conserverait d'une série de réalisations des variables X et Y que des valeurs très bien estimées de $E(X)$, $E(X^2)$ et $E(XY)$ (ce qui revient à une connaissance de la moyenne dans le groupe G_1 puisque l'on suppose π_1 connue).

Rappelons que la probabilité p_1^* obtenue par PME peut s'écrire en fonction des paramètres obtenus par résolution du système II.21 sous la forme développée suivante :

$$p_1^* = \frac{1}{1 + \exp\left(\frac{1}{2}\frac{(m_1^* + m_0^*)}{\sigma^*}\frac{(m_0^* - m_1^*)}{\sigma^*}\right)} .$$

Nous appellerons *différence réduite* la différence entre les moyennes rapportée à l'écart type ($d_r = (m_0 - m_1)/\sigma$), et *somme réduite* la somme des moyennes rapportée à l'écart type ($s_r = (m_0 + m_1)/\sigma$).

La figure II.3 présente un exemple du comportement des probabilités p_1^*, pour une somme réduite fixée et différentes valeurs de différences réduites (une différence réduite importante traduit que la population présente deux groupes bien séparés au sens de la moyenne), pour une population dont la variance globale est faible. On fait varier la différence réduite entre 0 et la valeur de s_r.

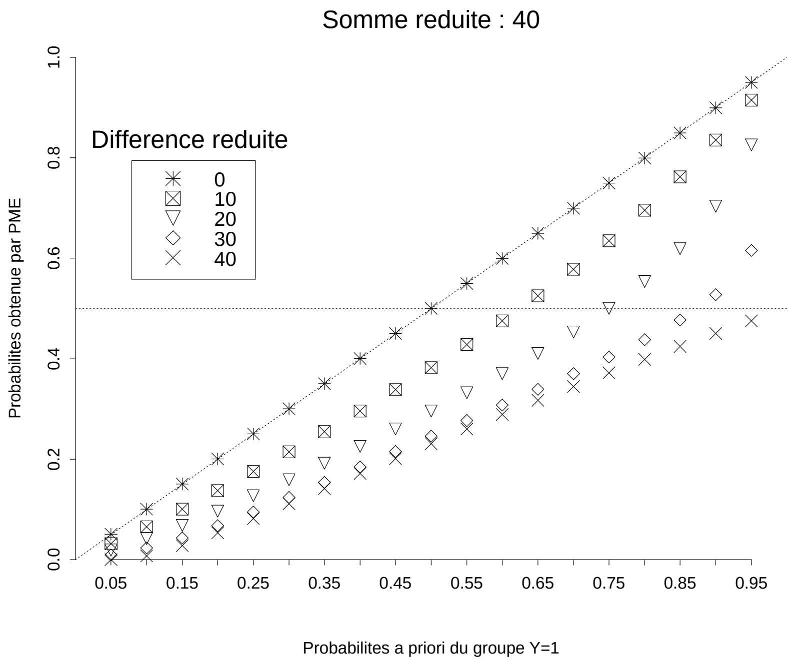


FIG. II.3 – *Comparaison des probabilités p_1^* obtenues par PME et des probabilités d'origine π_1, population de variance globale faible ($\sigma^2 = 0.07$)*

Une différence réduite négative revient à intervertir les deux groupes G_1 et G_0 (les courbes obtenues seraient donc les symétriques par rapport à la première bissectrice de celles tracées ici). D'autre part, les courbes correspondant à une différence réduite supérieure à s_r se situent entre les courbes tracées pour $d_r = 0$ et $d_r = s_r$.

La figure II.4 présente un exemple du comportement des probabilités p_1^*, pour une somme réduite fixée et une différence réduite qui varie entre 0 et s_r, pour une population dont la variance globale est très grande. Les probabilités p_1^* sont alors toujours très proches de la valeur 0.5, quelles que soient les valeurs a priori, et pour toute différence réduite. Cette constatation est en adéquation avec les expressions II.20 qui font apparaître des exponentielles en $1/\sigma^{*2}$.

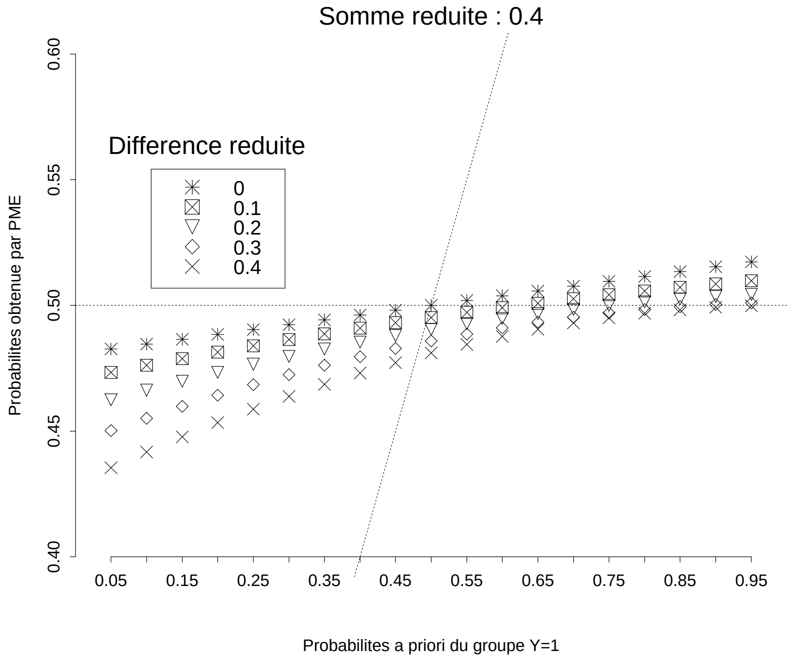


FIG. II.4 – *Comparaison des probabilités p_1^* obtenues par PME et des probabilités d'origine π_1, population de variance globale élevée($\sigma^2 = 700$)*

Ces quelques exemples illustrent que les équations du PME ne peuvent pas retrouver en général les valeurs d'origine des paramètres usuels, ni par conséquent les probabilités d'appartenance aux groupes (d'autant plus pour une population présentant une variance commune aux deux groupes élevée).

4.4 Frontières de discrimination et modèles de maximum d'entropie

Quand on étudie une population constituée de différents groupes au travers des lois conditionnelles des covariables, il est nécessaire de construire une frontière de discrimination pour classer de nouveaux individus. Un tel classement consiste généralement à affecter cet individu au groupe le plus probable a posteriori, c'est-à-dire après utilisation des valeurs observées $\boldsymbol{x}$ des covariables.

L'équation générale de toute frontière de discrimination s'écrit alors de la manière suivante, en considérant des coûts de mauvais classements égaux (cf. 1.1, p. 76) :

$$\ln \frac{f_X^0(x)}{f_X^1(x)} = \ln \frac{\pi_1}{\pi_0}\,.$$

Dans l'optique habituelle des modèles d'analyse discriminante, les formes des densités (ou des lois de probabilité) f_X^y, $y \in \{0,1\}$, proviennent du choix du modélisateur. Les proportions π_y sont en général mal connues et leur détermination peut s'avérer impossible si l'échantillon disponible est rétrospectif.

La règle d'affectation d'un individu au groupe le plus probable est indépendante de la manière dont les lois conditionnelles et marginales ont été obtenues. Nous pouvons donc déterminer la frontière de discrimination correspondant aux lois issues de l'information générale II.13 :

$$\mathcal{I} = \{(\psi_m(\boldsymbol{x}), a_m)_{m\in\{1,\dots,M\}}, (\varphi_r(\boldsymbol{x})y, b_r)_{r\in\{1,\dots,R\}}\}.$$

Rappelons que dans ce cas le PME permet d'obtenir une loi de Y de maximum d'entropie, pour laquelle $p_y^* = Z/Z_y$, où Z_y est la constante de normalisation de la loi de $\boldsymbol{X}$ sachant $Y=y$, $y \in \{0,1\}$. En reprenant la forme des lois conditionnelles en $\boldsymbol{X}$ II.14 et en prenant des coûts de mauvais classements égaux, on aboutit à la frontière suivante :

$$\ln \frac{f_X^{*0}(x)}{f_X^{*1}(x)} = \ln \frac{\pi_1^*}{\pi_0^*} \qquad \Rightarrow \qquad \sum_r \lambda'_r \varphi_r(\boldsymbol{x}) = 0\,. \tag{II.24}$$

Si on utilise les lois conditionnelles et la loi de groupe de maximum d'entropie obtenues à partir d'informations ne fixant pas la loi de Y, mais traduisant bien une situation de discrimination, la frontière de discrimination dépendra *uniquement* de la forme des fonctions $\varphi_r(\boldsymbol{x})$ qui définissent les observables "d'interaction" $\varphi_r(\boldsymbol{x})y$, $r \in \{1,\dots,R\}$.

L'équation précédente peut s'appliquer aux informations qui permettent de retrouver les hypothèses de l'ADL ou de l'ADQ mais qui ne fixent pas la loi de Y. Nous obtenons par exemple les frontières suivantes :

Type d'analyse discriminante relue	**Frontière "PME"**
ADL (observables : x_j, x_j^2, x_jy, $j \in \{1,\dots,J\}$)	$\sum_j \lambda'_j x_j = 0$
ADQ (observables : x_j, x_j^2, x_jy, x_j^2y, $j \in \{1,\dots,J\}$)	$\sum_j(\lambda'_j x_j + \gamma'_j x_j^2) = 0$

Les frontières ainsi obtenues ont donc toutes la particularité d'être des hyperplans passant par le point $(0,\dots,0)$. Le non-sens de ce résultat est encore plus flagrant dans le cas d'une unique variable, puisqu'on obtient alors comme frontière la droite $x = 0$, pour tous niveaux de contraintes tels qu'il existe une loi de maximum d'entropie.

On introduit en fait une symétrie autour de l'origine au niveau de la frontière de discrimination quand on ne considère que les observables $(x_j, x_j^2, x_j y)$ (ou $(x_j, x_j^2, x_j y, x_j^2 y)$) qui ne permettent pas de travailler à loi de Y fixée. On peut rapprocher cette constatation du comportement du PME qui choisit une loi normale *centrée* quand on impose le seul observable x^2 dans l'étude d'une variable à valeurs dans $\mathbf{R}$ (la loi obtenue sera symétrique autour de 0 si les observables sont uniquement des fonctions symétriques de x). C'est le même comportement qui se fait jour dans la construction des modèles de discrimination, mais au niveau de la frontière de discrimination.

Remarquons que si on ajoute des observables $(x_j - \alpha_j)y$, on obtiendra une frontière de discrimination passant par le point $(\alpha_1, \ldots, \alpha_J)$.

5 Conclusion

Cette étude de différentes familles d'observables dans un même but de discrimination montre bien que le choix des observables demande toujours à être discuté, entre autre par rapport aux propriétés mathématiques des fonctions des covariables qu'ils mettent en jeu. Mettre de coté certains observables pour simplifier l'information peut amener à induire des symétries dans la loi de maximum d'entropie (comme c'est le cas ici au niveau des frontières de discrimination quand on ne considère pas les observables fixant la loi de Y). On peut être également amené à s'interroger sur le choix des observables au travers de l'origine des niveaux de contraintes que l'on utilisera ensuite pour déterminer les paramètres des lois de maximum d'entropie. Le choix d'un observable pourra être mis en cause si l'utilisation du niveau de contrainte empirique associé n'est pas validée.

C Obtenir les probabilités conditionnelles d'appartenance aux groupes

Nous allons à présent nous placer dans le cadre du "diagnostic paradigm", pour lequel on cherche à caractériser non plus les lois conditionnelles des covariables, mais les probabilités d'appartenance aux groupes sachant la configuration des covariables. Notre étude se basera sur le modèle de régression logistique, dont la définition et les conditions d'application induisent de nombreuses réflexions sur le traitement de la vraisemblance dans le contexte d'une analyse discriminante. Nous verrons comment la relecture d'un tel modèle par le PME permet d'éclairer sa nature et son utilisation.

1 Points de vue habituels

Dans l'étude des liens entre une variable discrète Y (variable à expliquer ou variable de réponse) et une série de variables explicatives $\boldsymbol{X} = (X_1, \ldots, X_J)$ (covariables ou prédicteurs) de nature quelconque, une approche possible consiste à exprimer la probabilité d'appartenance à chaque groupe défini par Y sachant les valeurs des covariables par une fonction simple dont les paramètres auront une interprétation claire ([Cox 66, Dawid 76]). Ce choix conduit à définir différents modèles tels que ceux de régression linéaire, régression probit, ..., selon la nature de la fonction liant probabilité conditionnelle et fonction linéaire des mesures des variables $\boldsymbol{X}$. Ainsi, une des solutions est de mesurer l'effet des variables explicatives sur une échelle logistique (cf. schéma II.7, p. 117, et [Balakrishnan 92], chap. 1, pour un historique de la régression logistique). Le modèle utilisé sera alors construit sur l'hypothèse de la linéarité du logit des probabilités conditionnelles (la fonction logit est définie sur $\mathbf{R} - \{1\}$ par : logit $(x) = \ln(x/(1-x))$. C'est au travers de ce modèle qu'a été plus particulièrement étudiée l'influence du schéma d'échantillonnage sur l'estimation par maximum de vraisemblance des paramètres impliqués dans sa définition.

Dans le cas le plus courant de discrimination entre deux groupes pour lequel la variable Y prend ses valeurs dans $\{0, 1\}$, l'hypothèse la plus généralement utilisée pour définir le modèle est la suivante :

$$\text{logit } \pi(\boldsymbol{x}) = \ln \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \beta' \boldsymbol{x} \ , \text{ où } \beta = (\beta_0, \ldots, \beta_J) \in \mathbf{R}^{J+1} \ ,$$

en utilisant la notation $\pi(\boldsymbol{x}) = P(Y = 1/\boldsymbol{X} = \boldsymbol{x})$, pour un vecteur observé $\boldsymbol{x} = (x_0, \ldots, x_J)$ des prédicteurs (en posant $x_0 = 1$ pour simplifier l'écriture de la formule).

Ce type de modèle est particulièrement apprécié dans le domaine médical pour la facilité avec laquelle on peut l'adapter aux données, puisqu'il permet de traiter

1838	**Verhulst 1838, 1845** Modélisation de la taille d'une population humaine par une fonction dont la forme provient de certaines hypothèses posées sur l'accroissement de cette population dans le temps (l'accroissement du logarithme de la taille est une fonction du temps qui s'exprime comme une constante moins une fonction qui augmente avec la taille de la population, et la résolution de cette équation différentielle aboutit à une courbe sigmoïde). Utilisation du terme *logistique* (*"logistic response curve"*) pour désigner cette courbe.
1920	**Pearl & Reed** Ajustement d'une courbe logistique à l'évolution de la population des USA entre 1790 et 1910. Les auteurs précisent que cette courbe ne peut être utilisée pour des prévisions que si la population ne subit pas de phénomènes perturbatoires importants.
1929	**Reed & Berkson** Reprise de cette idée dans le cadre d'étude de phénomènes physico-chimiques. L'appelation *modèle logistique* est "adoptée", mais il s'agit toujours de l'ajustement (par les moindres carrés) d'une courbe théorique à la courbe des observations.
1940-50	**Berkson 44, 51; Wilson & Worcester 43**, etc Application en biologie. Polémique entre partisans du probit et du logit. Berkson préfére la courbe logistique car elle permet de rester dans le cadre de la famille exponentielle, d'aboutir à des estimateurs asymptotiquement efficaces et suffisants (*"...there are reasons for believing it to be easier to handle statistically"*) et qu'elle correspond au comportement d'un plus grand nombre de phénomènes physico-chimiques (utiliser la courbe normal-intégrée revient à faire des hypothèses supplémentaires sur le comportement de chaque individu).
1955-70	**Cox 58, 66**, etc Application du modèle logistique à l'étude de données médicales, dans le but d'étudier la dépendance entre la probabilité de succès d'un évènement (maladie, mort, ...) et un vecteur de variables explicatives. L'approche choisie consiste à lier cette probabilité et les variables par une fonction empirique simple, dont les paramètres s'interprètent clairement. Il existe différents choix possibles (l'inverse de la fonction de répartition d'une loi normale correspondra à un modèle probit, l'identité à un régression linéaire, et la courbe logistique à un modèle logistique). Cox préfère le modèle logistique pour sa simplicité (dans l'analyse statistique, les calculs et l'interprétation des paramètres).
1975-85	**Anderson 82, 84; Day-Kerridge 67; Prentice & Pike 79**, etc Étude des différents problèmes posés par l'estimation des paramètres d'un modèle logistique.

TAB. II.7 – *Chronologie de l'utilisation de la courbe logistique puis du modèle logistique* de la même manière des variables explicatives discrètes ou continues, et pour l'interprétation simple des paramètres en terme d'odds-ratios. De plus, une des propriétés souvent mise en avant pour illustrer sa simplicité d'application réside dans le fait que, quel que soit le protocole d'échantillonnage, les estimateurs du maximum de vraisemblance des paramètres s'obtiennent toujours en utilisant une vraisemblance *conditionnelle* pour laquelle la distribution de la variable Y est définie à $\boldsymbol{X}$ fixé ([Bouyer 91, Tomassone, Daudin, Danzart & Masson 88]).

Nous verrons de quelle manière la vraisemblance globale peut être décomposée, grâce aux propriétés de la fonction logit, pour faire apparaître cette partie conditionnelle correspondant en fait à un type d'échantillonnage particulier (échantillon prospectif). Dans les premiers articles traitant du problème de l'influence du schéma d'échantillonnage sur la recherche des estimateurs du maximum de vraisemblance, les auteurs ne se basent que sur cette partie conditionnelle. Nous déterminerons quelles hypothèses impliquent leurs arguments ou leurs démonstrations, et comment certaines approches du problème amènent en fait à définir une classe de méthodes logistiques plutôt qu'un unique modèle logistique.

L'étude d'une loi de la famille exponentielle nous permettra d'établir que l'équivalence des équations auxquelles répondent ces estimateurs des paramètres du modèle n'est qu'asymptotique.

2 Influence du schéma d'échantillonnage en régression logistique

L'influence du schéma d'échantillonnage dans l'utilisation de la régression logistique se traduit au niveau de la vraisemblance des paramètres du logit. Nous allons détailler les conséquences de cette influence sur la maximisation de la vraisemblance ainsi que les différentes interprétations et solutions proposées.

2.1 Expression d'un modèle logistique

Nous reprenons ici le formalisme d'Anderson ([Anderson 82]) qui propose une approche logistique de la discrimination en se basant sur une hypothèse de linéarité du rapport des vraisemblances.

Nous ne considèrerons ici que des discriminations logistiques à deux groupes (Y binaire à valeurs dans $\{0, 1\}$) pour des variables explicatives $\boldsymbol{X} = (X_1, \dots, X_J)$ *binaires* uniquement.

- On note P la loi conjointe des variables, P_X et P_Y les lois marginales.
- On note $L(\boldsymbol{X} = \boldsymbol{x}/Y = y)$, $y \in \{0, 1\}$ la vraisemblance de l'observation $\boldsymbol{x}$ pour

laquelle on sait que Y vaut y (rappelons que cette valeur est proportionnelle à la probabilité $P(\mathbf{X} = \boldsymbol{x}/Y = y)$).

Le modèle se caractérise alors par l'hypothèse suivante :

$$\ln \frac{L(\mathbf{X} = \boldsymbol{x}/Y = 1)}{L(\mathbf{X} = \boldsymbol{x}/Y = 0)} = \beta_0' + \sum_{j=1}^{J} \beta_j x_j \ , \tag{II.25}$$

où $(\beta_0', \beta_1, \ldots, \beta_J) \in \mathbf{R}^{J+1}$, pour un vecteur observé quelconque $\boldsymbol{x}$ de $\mathbf{X}$.

L'expression du logit des probabilités conditionnelles s'obtient alors par application de la formule de Bayes :

$$\begin{aligned}
\ln \frac{L(\mathbf{X} = \boldsymbol{x}/Y = 1)}{L(\mathbf{X} = \boldsymbol{x}/Y = 0)} &= \ln \frac{P(\mathbf{X} = \boldsymbol{x}/Y = 1)}{P(\mathbf{X} = \boldsymbol{x}/Y = 0)} \\
&= \ln \left(\frac{P(Y = 1/\boldsymbol{x}) P_X(\boldsymbol{x})}{P_Y(1)} \right) \left(\frac{P_Y(0)}{P(Y = 0/\boldsymbol{x}) P_X(\boldsymbol{x})} \right) \\
&= \ln \frac{P(Y = 1/\boldsymbol{x})}{P(Y = 0/\boldsymbol{x})} + \ln \frac{P_Y(0)}{P_Y(1)} \\
&= \ln \frac{P(Y = 1/\boldsymbol{x})}{P(Y = 0/\boldsymbol{x})} + \ln \frac{\pi_0}{\pi_1} \\
&= \ln \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} + \ln \frac{\pi_0}{\pi_1} \\
&= \text{logit } \pi(\boldsymbol{x}) + \ln \frac{\pi_0}{\pi_1} \ ,
\end{aligned}$$

où $\pi(\boldsymbol{x}) = P(Y = 1/\mathbf{X} = \boldsymbol{x})$, $\pi_0 = P_Y(0)$ et $\pi_1 = P_Y(1)$.

Cette hypothèse permet donc de caractériser le logit des probabilités conditionnelles de la manière suivante :

$$\begin{aligned}
\text{logit } \pi(x) &= \ln \frac{L(\mathbf{X} = \boldsymbol{x}/Y = 1)}{L(\mathbf{X} = \boldsymbol{x}/Y = 0)} - \ln \frac{\pi_0}{\pi_1} \\
&= \beta_0' + \sum_{j=1}^{J} \beta_j x_j + \ln \frac{\pi_1}{\pi_0} \\
&= \beta' \boldsymbol{x} \ ,
\end{aligned} \tag{II.26}$$

où $\beta = (\beta_0, \ldots, \beta_J)$, avec $\beta_0 = \beta_0' + \ln(\pi_1/\pi_0)$, et en posant $x_0 = 1$.

Une formulation équivalente permet de retrouver l'expression directe des probabilités conditionnelles :

$$\begin{cases} \pi(\boldsymbol{x}) & = & P(Y=1/\mathbf{X}=\boldsymbol{x}) & = & \dfrac{e^{\beta'\boldsymbol{x}}}{1+e^{\beta'\boldsymbol{x}}} \\ 1-\pi(\boldsymbol{x}) & = & P(Y=0/\mathbf{X}=\boldsymbol{x}) & = & \dfrac{1}{1+e^{\beta'\boldsymbol{x}}}\ , \ \forall \boldsymbol{x} \in \{0,1\}^{J+1}\ . \end{cases}$$

▷ **Remarques**

Cette approche par la donnée de l'expression du rapport des vraisemblances permet de ne pas imposer d'hypothèses directement sur la forme des distributions. Elle permet d'autre part d'utiliser pour l'estimation des paramètres les mêmes procédures de maximisation de la vraisemblance, que les variables explicatives soient qualitatives ou quantitatives. De plus, de nombreuses lois telles que des distributions normales avec des matrices de covariances égales, des distributions multivariées discrètes suivant un modèle log-linéaire avec des termes d'interactions égaux, ... répondent à l'équation II.25 (on se rapportera à [Anderson 82] pour une liste plus détaillée).

L'expression II.26 sera par la suite dénommée **propriété du logit linéaire**. Lorsque l'on parle de modèle de régression logistique de paramètre β, on englobe donc une famille de modèles possèdant cette propriété. Il ne s'agira donc pas dans l'étude de cette famille de modèles d'obtenir les estimateurs des paramètres de la loi conjointe, mais ceux des paramètres du logit des probabilités conditionnelles.

Signalons enfin qu'il est possible de définir un modèle de régression polynomial qui correspond de la même manière à une famille de modèles vérifiant la propriété d'avoir un logit qui est un polynôme des observations en X.

2.2 Décomposition de la vraisemblance

Il s'agit d'utiliser une série de N observations $(\underline{\boldsymbol{x}}, \underline{y}) = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N))$ des variables pour estimer les paramètres $(\beta_0, \ldots, \beta_J)$ du logit.

La vraisemblance a une expression différente suivant le type d'échantillonnage dont sont issues les données utilisées pour l'estimation. La nature binaire des variables permet en effet de travailler à $\mathbf{X}$ fixé (échantillon **prospectif**), à Y fixé (échantillon **rétrospectif**) ou de manière totalement aléatoire (échantillon de **mélange**).

2.2.1 Modification de la propriété logistique

Selon le type d'échantillonnage, il se peut que le fait pour un individu quelconque de la population d'être sélectionné dépende de la valeur de la variable réponse Y pour cet individu.

La propriété logistique ne portera alors plus sur la distribution de Y conditionnellement aux valeurs observées de $\boldsymbol{X}$, mais sur la probabilité conditionnelle à $\boldsymbol{X}$ *et* au fait d'appartenir ou non à l'échantillon, comme nous allons le détailler à présent.

Pour traduire cette influence du schéma d'échantillonnage sur le calcul de la vraisemblance, nous introduisons une variable T à valeurs dans $\{0, 1\}$ indiquant si un individu de la population appartient ($T = 1$) ou non ($T = 0$) à l'échantillon ([Hosmer & Lemeshow 89]).

Un échantillon rétrospectif ou prospectif n'est pas une image aléatoire de la population de base, mais la combinaison d'échantillons (c'est-à-dire de valeurs indépendantes et de même loi) de chaque groupe de cette population défini par les configurations en Y ou $\boldsymbol{X}$. La variable T permettra d'utiliser un tel échantillon comme un véritable échantillon de la population qui sera caractérisée alors par les variables $\boldsymbol{X}$, Y et T. Selon les conditions d'échantillonnage, T devra ou non être prise en compte.

• Tirage totalement aléatoire (échantillon de mélange)

Il n'est pas dans ce cas nécessaire de considérer T car elle est indépendante des variables mises en jeu.

Ainsi, on a : $P(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1) = P(Y = 1/\boldsymbol{X} = \boldsymbol{x})$. Il n'y a donc pas lieu de modifier l'écriture de la propriété logistique.

• Tirage à $\boldsymbol{X}$ contrôlé et Y aléatoire (échantillon prospectif)

Cette situation (dite aussi "conditionnelle" ou "étude de cohortes") se rencontre assez fréquemment dans les études de données médicales où des cohortes de la population, définies à partir des prédicteurs $(X_1, \ldots, X_L)$, sont suivies au cours du temps et pour lesquelles l'incidence de la maladie est évaluée à différentes périodes.

Dans un cadre prospectif total, les fréquences de toutes les configurations sont fixées, de manière égale ou non. Dans la pratique, il arrive plus fréquemment que ne soient fixées que les fréquences de certaines configurations.

Il n'est en fait pas nécessaire de modifier l'hypothèse logistique car, conditionnellement à $\boldsymbol{X}$, T et Y sont indépendantes. Nous pouvons "oublier" cette variable, car cette indépendance conditionnelle permet d'écrire : $P(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1) = P(Y = 1/\boldsymbol{X} = \boldsymbol{x})$.

L'hypothèse logistique porte sur l'étude de Y conditionnellement aux covariables. De manière intuitive, il est donc logique que le caractère contrôlé ou non de ces variables dans l'échantillon n'intervienne pas dans l'étude des paramètres.

• Tirage à Y contrôlé et $\boldsymbol{X}$ aléatoire (échantillon rétrospectif)

On travaille dans ce cas sur un groupe de "témoins" (individus pour lesquels $Y = 0$) et un groupe de "cas" (individus pour lesquels $Y = 1$). On mesure a posteriori les valeurs des variables explicatives pour chaque individu.

Il est alors nécessaire de travailler sur la probabilité conditionnelle en $\boldsymbol{X}$ *et* T, puisque cette dernière variable et Y ne sont pas indépendantes. Nous allons détailler à présent de quelle manière l'écriture de la propriété logistique est modifiée par l'introduction de cette variable T.

Dans ce dernier cas, il s'agit donc de lier $P(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1)$ et la propriété II.26 posée sur $P(Y = 1/\boldsymbol{X} = \boldsymbol{x})$.

En appliquant deux fois la formule de Bayes, nous obtenons le résultat suivant :

$$P(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1) = \frac{P(Y = 1/\boldsymbol{X} = \boldsymbol{x})\gamma_1(\boldsymbol{x})}{P(Y = 1/\boldsymbol{X} = \boldsymbol{x})\gamma_1(\boldsymbol{x}) + P(Y = 0/\boldsymbol{X} = \boldsymbol{x})\gamma_0(\boldsymbol{x})},$$

où $\gamma_y(\boldsymbol{x}) = P(T = 1/\boldsymbol{X} = \boldsymbol{x}, Y = y)$, $y \in \{0, 1\}$ désignent les taux de sondage utilisés pour le tirage de l'échantillon.

Dans le dernier cas de figure où Y est contrôlé, les taux de sondage sont indépendants de $\boldsymbol{X}$:

$$\gamma_y(\boldsymbol{x}) = \gamma_y = P(T = 1/Y = y),\ \forall\, y \in \{0, 1\}.$$

On pourra donc considérer finalement la relation suivante :

$$P(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1) = \frac{\gamma_1 \pi(\boldsymbol{x})}{\gamma_0(1 - \pi(\boldsymbol{x})) + \gamma_1 \pi(\boldsymbol{x})} = \widetilde{\pi}(\boldsymbol{x}) . \tag{II.27}$$

En utilisant la propriété logistique posée sur $\pi(\boldsymbol{x})$, on obtient donc l'expression correspondante de $\widetilde{\pi}(\boldsymbol{x})$ en fonction des paramètres β et des taux de sondages γ_0, γ_1 :

$$\begin{aligned} \text{logit } \widetilde{\pi}(\boldsymbol{x}) &= \ln \frac{\gamma_1 \pi(\boldsymbol{x})}{\gamma_0(1 - \pi(\boldsymbol{x}))} \\ &= \text{logit } \pi(\boldsymbol{x}) + \ln \frac{\gamma_1}{\gamma_0} \\ &= \widetilde{\beta}_0 + \sum_{j=1}^{J} \beta_j x_j , \end{aligned} \tag{II.28}$$

où $\widetilde{\beta}_0 = \beta_0 + \ln(\gamma_1/\gamma_0)$ est le paramètre constant corrigé par les taux de sondage.

On obtient à nouveau une fonction linéaire des valeurs observées $\boldsymbol{x}$, avec le même vecteur des paramètres β, *hormis pour la constante* β_0 à laquelle il est nécessaire d'ajouter un coefficient de correction dépendant des taux de sondage.

Le schéma d'échantillonnage à l'origine des données n'est donc pas une considération superficielle puisque, dans le cas où la loi observée de Y est fixée a priori, il sera nécessaire de corriger l'estimation de la constante $\widetilde{\beta}_0$ fournie par les logiciels. Cette correction peut être source de problème car ces taux de sondage dépendent de la loi a priori de Y qui n'est pas toujours facilement estimable.

En effet, pour $y \in \{0, 1\}$, on a :

$$\begin{aligned} \gamma_y &= P(T = 1/Y = y) \\ &= \frac{P(T = 1)N_y/N}{\pi_y}, \end{aligned} \tag{II.29}$$

où N est la taille de l'échantillon et N_y est le nombre imposé d'individus provenant du groupe G_y, $y \in \{0, 1\}$ (la loi de Y au sein de l'échantillon est connue : $P(Y = y/T = 1) = N_y/N$).

Le paramètre constant du logit pour un tel schéma d'échantillonnage s'écrit donc en fonction des probabilités a priori d'appartenance aux groupes définis par Y de la manière suivante :

$$\begin{aligned} \widetilde{\beta}_0 &= \beta_0 + \ln(\gamma_1/\gamma_0) \\ &= \beta_0 + \ln(N_1/N_0) + \ln(\pi_0/\pi_1) . \end{aligned} \tag{II.30}$$

2.2.2 Vraisemblance suivant les protocoles d'échantillonnage

Précisons maintenant les différentes écritures de la log-vraisemblance selon le type de protocole à l'origine des données :

• Échantillonnage prospectif

Pour ce type de données, la log-vraisemblance (notée $\ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y})$) s'écrit de la manière suivante :

$$\begin{aligned} \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) &= \sum_{i=1}^{N} \ln[P(Y = 1/\boldsymbol{x}_i)^{y_i} P(Y = 0/\boldsymbol{x}_i)^{1-y_i}] \\ &= \sum_i y_i \ln(\pi(\boldsymbol{x}_i)) + \sum_i (1 - y_i) \ln(1 - \pi(\boldsymbol{x}_i)) \\ &= \sum_i y_i \ln\left(\frac{\pi(\boldsymbol{x}_i)}{(1 - \pi(\boldsymbol{x}_i))}\right) + \sum_i \ln(1 - \pi(\boldsymbol{x}_i)) . \end{aligned} \tag{II.31}$$

• Échantillonnage de mélange

Pour ce type de données, la log-vraisemblance (notée $\ln L(\beta, \underline{\boldsymbol{x}}, \underline{y})$) se décompose de la manière suivante :

$$\ln L(\beta, \underline{\boldsymbol{x}}, \underline{y}) = \sum_{i=1}^{N} \ln P(\mathbf{X} = \boldsymbol{x}_i, Y = y_i)$$

$$
\begin{aligned}
&= \sum_i \ln\left(P(Y = y_i/\boldsymbol{X} = \boldsymbol{x}_i)P_X(\boldsymbol{x}_i)\right) \\
&= \sum_i \ln\Big(P(Y = 1/\boldsymbol{X} = \boldsymbol{x}_i)^{y_i}P(Y = 0/\boldsymbol{X} = \boldsymbol{x}_i)^{1-y_i}P_X(\boldsymbol{x}_i)\Big) \\
&= \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) + \sum_i \ln\left(P_X(\boldsymbol{x}_i)\right) \ . \qquad \text{(II.32)}
\end{aligned}
$$

Il est donc possible de faire apparaître dans cette expression la vraisemblance conditionnelle en $\boldsymbol{X}$ définie précédemment, en l'associant à un terme ne dépendant que de la loi des variables explicatives. Si une famille de paramètres autres que ceux apparaissant dans le logit intervient dans la modélisation de la loi conjointe P, on ne pourra pas obtenir d'information sur eux à partir de la seule vraisemblance $\ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y})$ qui apparaît dans cette décomposition.

• **Échantillonnage rétrospectif**

Nous avons vu que pour ce type de données, la vraisemblance devait se baser sur la distribution conditionnelle $P(\boldsymbol{X}/Y,T)$ et non pas simplement sur $P(\boldsymbol{X}/Y)$ comme pour les types d'échantillonnage précédents (prospectif et mélange) où le fait d'appartenir à l'échantillon est indépendant des valeurs de Y.

Pour des données telles que les nombres N_0 et N_1 d'observations dans chaque groupe sont fixés a priori, la log-vraisemblance (notée $\ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}})$) s'écrit de la manière suivante :

$$
\begin{aligned}
\ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}}) &= \sum_{i=1}^{N} \ln P(\boldsymbol{X} = \boldsymbol{x}_i/Y = y_i, T = 1) \\
&= \sum_{i=1}^{N} \ln\Big(P(\boldsymbol{X} = \boldsymbol{x}_i/Y = 1, T = 1)^{y_i}P(\boldsymbol{X} = \boldsymbol{x}_i/Y = 0, T = 1)^{1-y_i}\Big) \ .
\end{aligned}
$$

Nous avons vu qu'il était possible de définir $\widetilde{\pi}(\boldsymbol{x}) = P(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1)$ à partir de la propriété logistique vérifiée par $\pi(\boldsymbol{x})$, en modifiant uniquement le paramètre constant. D'autre part, du fait que T et $\boldsymbol{X}$ sont indépendants, on a :

$$
\left\{
\begin{aligned}
P(\boldsymbol{X} = \boldsymbol{x}_i/Y = 1, T = 1) &= \frac{\widetilde{\pi}(\boldsymbol{x})P_X(\boldsymbol{x}_i)}{P(Y = 1/T = 1)} \\
P(\boldsymbol{X} = \boldsymbol{x}_i/Y = 0, T = 1) &= \frac{(1 - \widetilde{\pi}(\boldsymbol{x}))P_X(\boldsymbol{x}_i)}{P(Y = 0/T = 1)} \ .
\end{aligned}
\right.
$$

Posons $\ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) = \sum_i \ln\left(\widetilde{\pi}(\boldsymbol{x}_i)^{y_i}\widetilde{\pi}(\boldsymbol{x}_i)^{1-y_i}\right)$, qui définit une log-vraisemblance de type prospective construite à partir de ces données rétrospectives.

L'expression finale de la log-vraisemblance dans le cas rétrospectif est donc la suivante :

$$\begin{aligned}
\ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}}) &= \sum_{i=1}^{N} \ln \left(\frac{\widetilde{\pi}(\boldsymbol{x}) P_X(\boldsymbol{x}_i)}{P(Y = y_i/T = 1)} \right)^{y_i} \left(\frac{(1 - \widetilde{\pi}(\boldsymbol{x})) P_X(\boldsymbol{x}_i)}{P(Y = y_i/T = 1)} \right)^{1-y_i} \\
&= \sum_{i=1}^{N} \ln \left(\frac{\widetilde{\pi}(\boldsymbol{x}) P_X(\boldsymbol{x}_i)}{P(Y = 1/T = 1)} \right)^{y_i} \left(\frac{(1 - \widetilde{\pi}(\boldsymbol{x})) P_X(\boldsymbol{x}_i)}{P(Y = 0/T = 1)} \right)^{1-y_i} \\
&= \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) + \sum_{i=1}^{N} \ln P_X(\boldsymbol{x}_i) \\
&\quad - N_1 \ln P(Y = 1/T = 1) - N_0 \ln P(Y = 0/T = 1) \\
&= \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) + \sum_{i=1}^{N} \ln(P_X(\boldsymbol{x}_i)) - N_1 \ln \frac{N_1}{N} - N_0 \ln \frac{N_0}{N} . \quad \text{(II.33)}
\end{aligned}$$

(P désigne en fait ici la loi conjointe des variables ($\boldsymbol{X}$,Y,T).)

L'expression de cette vraisemblance implique donc là aussi une vraisemblance de type prospective (dans le sens où elle se définit grâce aux probabilités conditionnelles en $\boldsymbol{X}$), bien que les données soient issues d'un schéma rétrospectif. La loi marginale des covariables doit à nouveau être prise en compte.

Par la suite nous désignerons l'expression II.31, p. 123, de la vraisemblance correspondant à des données issues d'un schéma prospectif, pour une propriété logistique donnée, par le terme de **log-vraisemblance prospective**. Nous venons de voir qu'il est possible de faire apparaître cette vraisemblance prospective dans l'expression des vraisemblances pour des échantillons de type mélange et rétrospectif.

2.3 Estimateurs du maximum de vraisemblance

Bien que nous ayons constaté les différences d'expressions des vraisemblances selon les trois types d'échantillons, les différents logiciels permettant d'estimer les paramètres d'un modèle possèdant la propriété du logit linéaire se basent tous sur les équations obtenues par maximisation d'une vraisemblance prospective. Le type de protocole à l'origine des données utilisées ne rentre alors pas en ligne de compte dans cette procédure d'estimation des paramètres.

2.3.1 La vraisemblance prospective comme base de l'estimation

D'une manière générale et sans se préoccuper pour l'instant de la justification de cette manière de procéder, il s'agit donc de maximiser l'expression suivante :

$$\ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) = \sum_{i=1}^{N} y_i \ln \left(\frac{\pi(\boldsymbol{x}_i)}{(1 - \pi(\boldsymbol{x}_i))} \right) + \sum_{i=1}^{N} \ln(1 - \pi(\boldsymbol{x}_i)) \ ,$$

sous l'hypothèse que :

$$\pi(\boldsymbol{x}) = \frac{e^{\beta'\boldsymbol{x}}}{1+e^{\beta'\boldsymbol{x}}} \ , \ \forall \boldsymbol{x} \in \{0,1\}^{J+1} \, .$$

Ce qui permet d'écrire :

$$\ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) = \sum_i y_i(\beta'\boldsymbol{x}_i) - \sum_i \ln(1+e^{\beta'\boldsymbol{x}_i}) \, .$$

Le fait d'annuler les dérivées de la log-vraisemblance par rapport à chaque paramètre s'écrit alors :

$$\frac{\partial \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y})}{\partial \beta_j} = \sum_i y_i x_{ij} - \sum_i \frac{x_{ij} e^{\beta'\boldsymbol{x}_i}}{1+e^{\beta'\boldsymbol{x}_i}} = 0 \ , \ \forall j \in \{0, \ldots, J\} \, ,$$

en posant toujours $x_{0i} = 1$, $\forall i \in \{1, \ldots, N\}$.

Chaque paramètre estimé $\widehat{\beta}_j$, $j \in \{0, \ldots, J\}$ répond donc à l'équation suivante :

$$\sum_i y_i x_{ij} = \sum_i \frac{x_{ij} e^{\widehat{\beta}'\boldsymbol{x}_i}}{1+e^{\widehat{\beta}'\boldsymbol{x}_i}} \, ,$$

ce qui peut s'écrire sous la forme :

$$\sum_i y_i x_{ij} = \sum_i x_{ij} \widehat{\pi}(\boldsymbol{x}_i) \ , \ j \in \{0, \ldots, J\} \, , \tag{II.34}$$

où $\widehat{\pi}(\boldsymbol{x})$ est l'estimation de π au point $\boldsymbol{x}$ à partir des valeurs $(\widehat{\beta}_0, \ldots, \widehat{\beta}_J)$.

Les équations obtenues ne sont pas linéaires en β. Leur résolution nécessite donc l'emploi d'algorithmes itératifs (type Newton-Raphson). Mais nous verrons que dans le cas particulier de lois conjointes répondant à une hypothèse d'indépendance conditionnelle, il est possible d'utiliser des équations permettant d'obtenir des estimations directes des paramètres.

2.3.2 La procédure de régression logistique

Malgré la différence dans l'expression de la vraisemblance selon le protocole d'échantillonnage à l'origine des données, il est admis dans toutes les présentations de la régression logistique que l'estimation des paramètres peut se baser uniquement sur cette partie conditionnelle. Nous allons voir dans la section suivante les justifications qui ont été apportées à cette pratique. Tous les arguments avancés reposent sur "l'oubli" de la partie non conditionnelle de la vraisemblance dans la recherche des estimateurs. Mais, selon la manière dont est traité cet oubli, différentes optiques peuvent être envisagées.

Ainsi, nous nous appliquerons dans ce texte à distinguer les deux optiques suivantes :

- d'une part les modèles répondant à la **propriété logistique** ("modèles à logit linéaire") pour lesquelles l'écriture de la vraisemblance dépend du type de l'échantillon. Nous verrons alors comment les auteurs traitent la partie non conditionnelle des vraisemblances des échantillons rétrospectifs ou tout aléatoires ;

- d'autre part la **procédure de régression logistique** où l'on estime les paramètres du logit des probabilités conditionnelles à partir d'une vraisemblance prospective.

2.3.3 Dépendance de la loi des covariables vis-à-vis des paramètres

La maximisation de la vraisemblance correspond simplement à l'annulation des dérivées selon chaque paramètre. On peut donc formaliser le fait d'utiliser la vraisemblance prospective quel que soit le schéma d'échantillonnage de la manière suivante :

$$\frac{\partial \ln L(\beta, \underline{\boldsymbol{x}}, \underline{y})}{\partial \beta_j} = \frac{\partial \ln L_{\underline{y}}(\beta, \underline{\boldsymbol{x}})}{\partial \beta_j} = \frac{\partial \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y})}{\partial \beta_j}\ , \ \forall j \in \{0, \dots, J\}\ ,$$

où l'échantillon $(\underline{\boldsymbol{x}}, \underline{y})$ provient d'un des trois schémas définis précédemment.

Ces égalités sous-entendent la nullité de la dérivée de la partie non conditionnelle des vraisemblances issues des schémas rétrospectif et de mélange. Dans les deux cas, c'est la loi marginale des covariables qui apparaît. On agit donc comme si les équations suivantes étaient toujours vérifiées, ou toujours acceptables :

$$\sum_i \frac{\partial \ln(P_X(\boldsymbol{x}_i))}{\partial \beta_j} = 0\ , \ \forall j \in \{0, \dots, J\}\ .$$

Nous allons voir dans le contre-exemple suivant ce qu'il en est pour une loi de $(\boldsymbol{X}, Y)$ particulière de la famille exponentielle.

• Un contre exemple

Considérons une loi conjointe des variables mises en jeu du type suivant :

$$P(\boldsymbol{X} = \boldsymbol{x}, Y = y) = \exp(\mu_0 + \beta' \boldsymbol{x} y + \delta_\nu(\boldsymbol{x}))\ ,$$

pour un vecteur observé $\boldsymbol{x} = (x_0, \dots, x_J)$, $(x_0 = 1)$, où μ_0 est une constante de normalisation et δ_ν une fonction faisant apparaître une famille de paramètres ν *totalement indépendante* des paramètres $\beta = (\beta_0, \dots, \beta_J)$.

Cette distribution possède effectivement la propriété logistique pour le vecteur des paramètres β :

$$\text{logit } \pi(\boldsymbol{x}) = \beta' \boldsymbol{x} \; .$$

Pour ne pas avoir à considérer de maximisation sous contraintes, précisons le lien entre la constante de normalisation et les autres paramètres :

$$\mu_0 = -\ln \left[\sum_{\boldsymbol{x}} e^{\delta_\nu(\boldsymbol{x})} (1 + e^{\beta' \boldsymbol{x}}) \right] = -\psi(\nu, \beta) \; .$$

Nous pouvons détailler les lois marginales et y faire apparaître les paramètres du logit :

$$\begin{cases} \textbf{loi de } \boldsymbol{X} & : \quad P_X(\boldsymbol{x}) = (1 + e^{\beta' \boldsymbol{x}}) \exp(\mu_0 + \delta_\nu(\boldsymbol{x})) \; , \\ \textbf{loi de } Y & : \quad P_Y(y) = \sum_{\boldsymbol{x}} \exp(\mu_0 + \beta' \boldsymbol{x} y + \delta_\nu(\boldsymbol{x})) \; . \end{cases}$$

Il est utile de préciser la dérivée de la constante de normalisation, selon chaque composante du vecteur des paramètres du logit :

$$\text{- pour } \beta_j \neq \beta_0 : \qquad \begin{aligned} \frac{\partial \psi(\nu, \beta)}{\partial \beta_j} &= \frac{\sum_{\boldsymbol{x}} x_j \exp(\delta_\nu(\boldsymbol{x}) + \beta' \boldsymbol{x})}{\sum_{\boldsymbol{x}} e^{\delta_\nu(\boldsymbol{x})} (1 + e^{\beta' \boldsymbol{x}})} \\ &= \sum_{\boldsymbol{x}} x_j \exp(\mu_0 + \beta' \boldsymbol{x} + \delta_\nu(\boldsymbol{x})) \\ &= E(X_j Y) \; ; \end{aligned}$$

$$\text{- pour } \beta_0 : \qquad \begin{aligned} \frac{\partial \psi(\nu, \beta)}{\partial \beta_0} &= \frac{\sum_{\boldsymbol{x}} \exp(\delta_\nu(\boldsymbol{x}) + \beta' \boldsymbol{x})}{\sum_{\boldsymbol{x}} e^{\delta_\nu(\boldsymbol{x})} (1 + e^{\beta' \boldsymbol{x}})} \\ &= \sum_{\boldsymbol{x}} \exp(\mu_0 + \beta' \boldsymbol{x} + \delta_\nu(x)) \\ &= P_Y(1) \; . \end{aligned}$$

(Remarquons qu'apparaissent dans ces dérivées des sommes sur toutes les valeurs possibles de X.)

Nous pouvons donc décrire le comportement de la partie non conditionnelle de la log-vraisemblance. En effet, pour une observation $\boldsymbol{x}_i$ quelconque issue d'un échantillon $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, on a :

$$\begin{cases} \dfrac{\partial \ln P_X(\boldsymbol{x}_i)}{\partial \beta_j} &= -E(X_j Y) + x_{ij} P(Y = 1/X = \boldsymbol{x}_i) \; , \; j \neq 0 \\ \dfrac{\partial \ln P_X(\boldsymbol{x}_i)}{\partial \beta_0} &= -P_Y(1) + P(Y = 1/X = \boldsymbol{x}_i) \; . \end{cases}$$

Soit, en sommant sur tout l'échantillon :

$$\begin{cases} \dfrac{\partial \sum_i \ln P_X(\boldsymbol{x}_i)}{\partial \beta_j} & = & -NE(X_jY) + \sum_i x_{ij}\pi(\boldsymbol{x}_i) \ , \ j \neq 0 \\ \dfrac{\partial \sum_i \ln P_X(\boldsymbol{x}_i)}{\partial \beta_0} & = & -NP_Y(1) + \sum_i \pi(\boldsymbol{x}_i) \ . \end{cases} \tag{II.35}$$

Pour un échantillon donné d'un type quelconque, ces deux termes ne sont a priori pas nuls, ni même nuls au point de $\mathbf{R}^{J+1}$ qui maximise la vraisemblance prospective.

On ne peut donc pas affirmer que la loi marginale de X peut en général être négligée dans la maximisation de la log-vraisemblance, pour des échantillons autres que prospectifs.

C'est pourquoi nous allons nous interroger sur le bien-fondé de l'utilisation d'une vraisemblance prospective pour estimer les paramètres du logit, *quel que soit le type de l'échantillon utilisé.* Cette indépendance de la méthode d'estimation des paramètres vis-à-vis du protocole d'échantillonnage est pourtant une propriété reconnue et particulièrement mise en avant dans les descriptions de la régression logistique. Ainsi, nous passerons en revue les principales argumentations et démonstrations qui sont à la base de cette manière de procéder.

Mais, auparavant, nous pouvons compléter l'étude du contre exemple en illustrant le comportement asymptotique de la maximisation de la partie non conditionnelle de la vraisemblance dans la recherche des estimateurs des paramètres.

• Comportement asymptotique

Nous avons vu qu'il est possible de faire apparaître l'expression de la vraisemblance correspondant à un échantillon de type prospectif dans l'écriture des vraisemblances pour des échantillons de mélange ou rétrospectifs :

$$\begin{aligned} \ln L(\beta, \underline{\boldsymbol{x}}, \underline{y}) &= \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) + \sum_i \ln(P_X(\boldsymbol{x}_i)) \ (\text{mélange}) \\ \ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}}) &= \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) + \sum_i \ln(P_X(\boldsymbol{x}_i)) - N_0 \ln \frac{N_0}{N} - N_1 \ln \frac{N_1}{N} \ (\text{rétrospectif}) \ . \end{aligned}$$

Nous pouvons assimiler les vecteurs β et $\widetilde{\beta}$ puisqu'ils ne diffèrent que pour leur premier élément. Les deux termes de droite de ces équations ne diffèrent que par une constante (N, N_0, N_1 sont fixés par le protocole d'échantillonnage, s'il est de type rétrospectif).

Il nous suffit donc d'étudier le comportement de la maximisation de $\sum_i \ln(P_X(\boldsymbol{x}_i))$, quand la taille de l'échantillon s'accroît, pour en tirer des conclusions sur l'influence de cette partie non conditionnelle sur l'écriture des équations auxquelles doivent répondre les estimateurs.

Nous avons détaillé dans le paragraphe précédent la dérivée de $\sum_i \ln(P_X(\boldsymbol{x}_i))$ selon les différents paramètres (cf. II.35) :

$$\left\{ \begin{array}{lcl} \dfrac{\partial \sum_i \ln P_X(\boldsymbol{x}_i)}{\partial \beta_j} & = & -NE(X_jY) + \sum_i \boldsymbol{x}_{ij}\pi(\boldsymbol{x}_i) \ , \ j \neq 0 \\ \dfrac{\partial \sum_i \ln P_X(\boldsymbol{x}_i)}{\partial \beta_0} & = & -NP_Y(1) + \sum_i \pi(\boldsymbol{x}_i) \ . \end{array} \right.$$

Nous pouvons utiliser les lemmes 2 et 3 pour caractériser le comportement asymptotique de ces équations.

Lemme 2

$$1/N \sum_{i=1}^{N} x_{ij}\pi(\boldsymbol{x}_i) \xrightarrow{ps} E(X_jY) \ , \ \forall j \in \{1, \ldots, J\} \ .$$

Démonstration

En utilisant le fait que Y est binaire, on peut en effet écrire :

$$\begin{array}{lcl} \sum_i x_{ij}\pi(\boldsymbol{x}_i) & = & \sum_i x_{ij} P(Y = 1/\mathbf{X} = \boldsymbol{x}_i) \\ & = & \sum_i x_{ij} E(Y/\mathbf{X} = \boldsymbol{x}_i) \ , \ \forall j \in \{1, \ldots, J\} \ . \end{array}$$

On peut alors utiliser la loi forte des grands nombres :

$$\begin{array}{lcl} 1/N \sum_i x_{ij} E(Y/\mathbf{X} = \boldsymbol{x}_i) & \xrightarrow{ps} & E_X(X_j E(Y/X)) \\ & & = E_X(E(X_jY/X)) \\ & & = E(X_jY) \ . \end{array}$$

□

Lemme 3

$$1/N \sum_{i=1}^{N} \pi(\boldsymbol{x}_i) \xrightarrow{ps} P_Y(1) \ .$$

Démonstration

On a :

$$\sum_i \pi(\boldsymbol{x}_i) = \sum_i P(Y = 1/\mathbf{X} = \boldsymbol{x}_i) = \sum_i E(Y/\mathbf{X} = \boldsymbol{x}_i) \ .$$

On utilise la loi forte des grands nombres :

$$\begin{aligned} 1/N \sum_i E(Y/\boldsymbol{X} = \boldsymbol{x}_i) &\xrightarrow{ps} E_X(E(Y/X)) \\ &= E(Y) \\ &= P_Y(1) . \end{aligned}$$

□

Ainsi, on peut établir une équivalence asymptotique entre les maximisations :

$$\left\{ \begin{array}{lcl} \dfrac{1}{N} \left(\dfrac{\partial \ln L(\beta, \underline{\boldsymbol{x}}, \underline{y})}{\partial \beta_j} \right) & \xrightarrow{ps} & \dfrac{1}{N} \left(\dfrac{\partial \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y})}{\partial \beta_j} \right) \\ \dfrac{1}{N} \left(\dfrac{\partial \ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}})}{\partial \beta_j} \right) & \xrightarrow{ps} & \dfrac{1}{N} \left(\dfrac{\partial \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y})}{\partial \beta_j} \right) . \end{array} \right.$$

Nous venons donc de démontrer que pour cette loi de la famille exponentielle, il est possible de laisser de coté l'information sur les paramètres du logit contenue dans la loi marginale des covariables, mais de manière *asymptotique*.

▷ **Exemple**

La figure II.5, p. 132, présente un exemple de ce comportement asymptotique, pour un modèle construit pour 3 variables explicatives (X_1, X_2, X_3) binaires. Le modèle utilisé pour les simulations présente une propriété du logit linéaire :

$$\text{logit } \pi(\boldsymbol{x}) = 1 + 0.05x_1 - 2x_2 + 6x_3 .$$

Nous avons simulé, pour chaque type d'échantillonnage, 30 échantillons de taille N d'une loi conjointe possédant cette propriété, mais ne présentant pas de propriété d'indépendance conditionnelle des covariables. Nou savons utilisé ici la loi conjointe suivante :

$$\ln P(\boldsymbol{X} = \boldsymbol{x}, Y = y) = y + 0.05x_1y - 2x_2y + 6x_3y - x_1 + x_2x_3 + 5x_1x_2 .$$

Ce modèle correspond alors à une loi marginale de Y particulièrement déséquilibrée puisque l'on a $P(Y = 0) = 0.98$ et $P(Y = 1) = 0.02$. Une telle loi correspond par exemple dans le domaine médical à une maladie ou une déformation génétique rares, pour l'étude desquelles la régression logistique est une méthode très souvent utilisée.

Ces 30 simulations ont été faites pour des tailles d'échantillon N assez faibles ($N = 30$) jusqu'à des échantillons de très grande taille ($N = 10000$). Les simulations et estimations ont été produites à partir du logiciel SPLUS (version 3.4), pour un nombre maximal d'itérations de 50. Pour chaque type et chaque taille d'échantillon, nous avons produit des simulations jusqu'à en obtenir 30 pour lesquelles on obtenait une estimation des paramètres avec ce nombre maximal d'itérations.

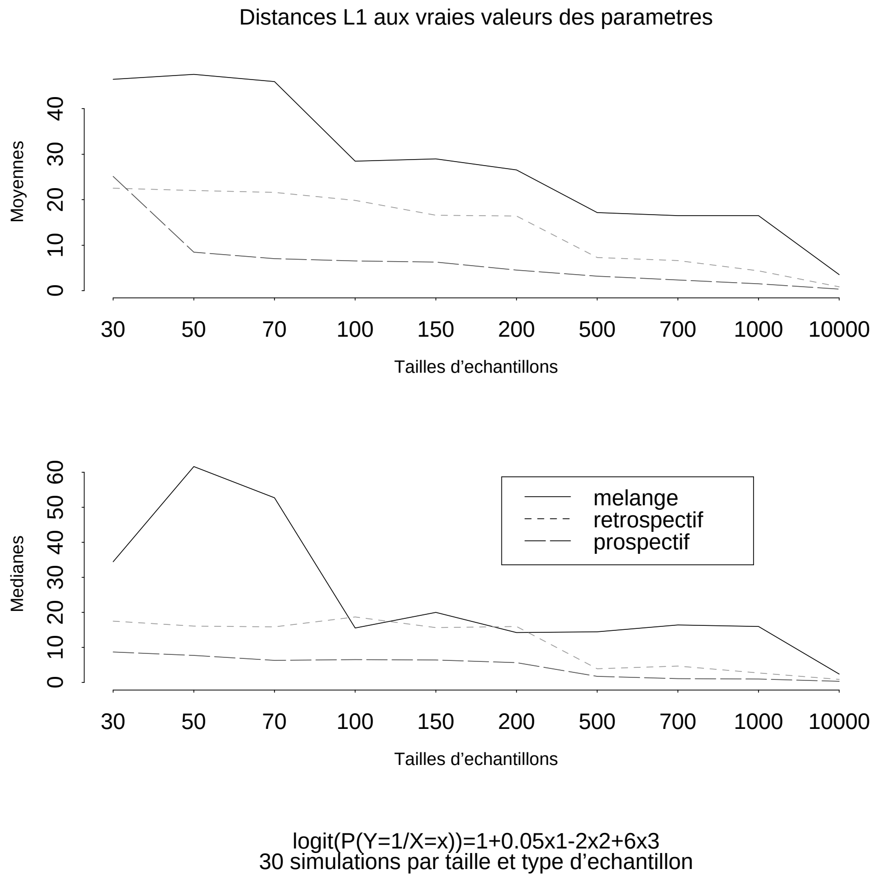


FIG. II.5 – *Comportements asymptotiques comparés des trois types d'échantillonnage pour la moyenne et médiane des distances L_1 des estimations par maximum de vraisemblance aux vraies valeurs d'un modèle à logit linéaire*

On représente sur ce graphe l'évolution des moyennes et des médianes des 30 distances L_1 des vecteurs des paramètres estimés au vecteur des vraies valeurs $(1, 0.05, -2, 6)$.

Dans le cas des petits échantillons, il apparaît que les plus mauvaises estimations correspondent à l'échantillonnage de mélange. On remarque de plus que pour ce type d'échantillonnage, beaucoup de simulations sont rejetées avant d'obtenir 30 vecteurs de paramètres estimés. Cela vient du type de loi simulée : dans le cas d'un tirage tout aléatoire, il est fort probable de ne pas obtenir d'individus du groupe $Y = 1$, ou très peu. Cela explique le nombre important de rejets de telles simulations dans le cas de petits échantillons, ainsi que la mauvaise qualité des estimations menées à terme.

L'ordre de grandeur des premières moyennes et médianes est important mais correspond en fait à lordre de grandeurs des estimations obtenues pour ce type de modèles ([Hosmer & Lemeshow 89], chap. 3).

On remarque de plus que les meilleures estimations sont obtenues pour l'échantillonnage prospectif, qui est le type d'échantillonnage correspondant naturellement à la vraisemblance maximisée.

2.4 Schéma de mélange et schéma prospectif : une argumentation historique

Nous avons vu avec l'équation II.32 que la log-vraisemblance correspondant à des données choisies de manière totalement aléatoire peut être reliée à l'expression que l'on aurait utilisée si cet échantillon avait été de type prospectif :

$$\ln L(\beta, \underline{\boldsymbol{x}}, \underline{y}) = \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) + \sum_i \ln(P_X(\boldsymbol{x}_i)) \,.$$

2.4.1 Day & Kerridge (1967)

Nous résumons ici l'approche de l'article “A general likelihood discriminant”, de Day & Kerridge paru en 1967 ([Day & Kerridge 67]). Dans cet article, les auteurs recherchent un modèle assez général de discrimination entre deux populations et cherchent pour cela à poser le moins d'hypothèses possibles sur les distributions des variables explicatives (dont le type n'est pas précisé) dans chacun de ces groupes d'individus.

Ainsi, ils se proposent d'étudier des lois qui vérifient la propriété logistique et qui sont de la forme suivante :

$$f_y(\boldsymbol{x}) = \alpha_y \phi(\boldsymbol{x}) \exp\left(-\frac{1}{2}(\boldsymbol{x} - m_y)' W^{-1} (\boldsymbol{x} - m_y)\right) \ , \ y \in \{0, 1\} \,, \qquad \text{(II.36)}$$

où ϕ est une fonction intégrale positive quelconque, W une matrice symétrique définie positive, m_0 et m_1 deux vecteurs de $\mathbf{R}^J$, α_0 et α_1 des constantes de normalisation.

Les différentes possibilités pour la fonction ϕ permettent d'englober dans cette écriture divers modèles tels que le modèle de Fisher ($\boldsymbol{X}$ suit une loi normale et ϕ est la fonction identité), celui de Warner & al (W est la matrice identité et $\phi(\boldsymbol{x})$ vaut 1 si $\boldsymbol{x}$ est un vecteur de 0 et de 1, 0 sinon), et d'autres situations fréquemment rencontrées en pratique, en particulier dans le domaine médical.

L'intérêt de cette approche est qu'aucune forme particulière de ϕ n'est requise et que la méthode s'adaptera à toute loi du type II.36.

Il est possible de retrouver l'expression de la propriété logistique sous forme de linéarité du logit en posant :

$$\left\{\begin{array}{ccl} \beta_0 & = & -\dfrac{1}{2}(m_0 - m_1)'W^{-1}(m_0 + m_1) + \ln\left(\dfrac{P_Y(0)\alpha_0}{P_Y(1)\alpha_1}\right) \\ (\beta_1, \ldots, \beta_L) & = & W^{-1}(m_0 - m_1)\,. \end{array}\right.$$

On a bien alors logit $\pi(\boldsymbol{x}) = \beta'\boldsymbol{x}$.

Deux cas de figures peuvent alors se poser :

- **soit ϕ est de nature parfaitement connue :** l'estimation des paramètres m_0, m_1 et W peut se faire par maximisation de la vraisemblance. Les estimations des paramètres β s'obtiennent alors comme fonctions de ces estimations.

- **soit ϕ est inconnue ou partiellement connue :** la vraisemblance contient donc une fonction de nuisance (ϕ). La décomposition de la vraisemblance est alors nécessaire, en faisant apparaître la partie conditionnelle et une partie ne dépendant que de la loi du vecteur $\boldsymbol{X}$ (c'est à dire de ϕ, W, α_0, α_1, m_0, m_1).

Nous trouvons alors dans cet article de référence la première justification du choix de ne baser la recherche des estimateurs que sur la partie conditionnelle de la vraisemblance. Les arguments utilisés (cf. l'extrait de l'article ci-dessous) par les auteurs ne reposent que sur leur "*examen*" de la partie non conditionnelle qui ne contiendrait pas d'information (ou une information négligeable) sur les paramètres. La nature de cette information et la manière de l'extraire de la vraisemblance ne sont pas précisées et aucun argument autre que des considérations subjectives n'étayent ce propos.

"Since $P(H_1)$, $P(H_2)$ and $\phi(X)$ may be chosen freely, we may make the first factor achieve its maximum (which occurs when the $P(X_r)$ are equal for all r) for any values of Λ_1, Λ_2, and A. The maximum does not depend on these quantities,

while the rest of the likelihood can be maximized by varying only Λ_1, Λ_2, and A. We therefore ignore the first product and estimate B and c maximising

$$L = \prod_{r \in S_1} \frac{exp(X_r' B + c)}{1 + \exp(X_r' B + c)} \times \prod_{r \in S_2} \frac{1}{1 + \exp(X_r' B + c)} .$$

This argument rests on the assumption that $\phi(X)$ can be chosen with complete freedom at every point X_r in the observed sample. Even if some information about the form of $\phi(X)$ is available, the procedure remains valid, (though not necessarily fully efficient) since we may observe that L is in fact a likelihood function, composed of the conditional frequencies $P(H_i/X_r)$ and as such maximum likelihood may properly be applied to it. Our examination of the unconditional likelihood function suggests that the loss of information, if any, which this conditional inference involves, is likely to be small unless information about ϕ is very precise."

2.4.2 Anderson (1982)

Nous présentons ici une partie d'un texte de J.A. Anderson ([Anderson 82]) présentant les propriétés de la régression logistique. Ce texte est chronologiquement éloigné du premier article présenté, mais il résume la manière dont les arguments exposés précédemment ont été systématiquement repris pour justifier l'utilisation de la vraisemblance conditionnelle dans le cadre d'un schéma de mélange, que le modèle ait été défini grâce à des hypothèses sur les lois des variables ou sur le rapport des vraisemblances.

L'auteur se base ici sur la propriété II.25 de linéarité du rapport des vraisemblances :

$$\ln \frac{L(\mathbf{X} = \boldsymbol{x}/Y = 1)}{L(\mathbf{X} = \boldsymbol{x}/Y = 0)} = \beta_0' + \sum_{j=1}^{J} \beta_j x_j ,$$

les covariables pouvant être continues *ou* discrètes.

Il reprend alors strictement l'argumentation précédente, sans autre justification : (dans l'extrait suivant la notation H_s, $s \in \{1, 2\}$, correspond aux groupes G_y, $y \in \{0, 1\}$ de la population et $n(\boldsymbol{x})$ est le nombre d'observations du vecteur $\boldsymbol{x}$ dans l'échantillon)

" Now the functional form of the likelihood ratio has been assumed in [II.25] but specifically no further assumptions have been made. This implies that $L(\boldsymbol{x})$, the marginal distribution of x contains no information about the β_j's. Even if there is further information about the form of $L(\boldsymbol{x}/H_s)$ (s=1,2), the extra information about the β_j's in $\prod\{L(\boldsymbol{x})\}^{n(\boldsymbol{x})}$ is likely to be small compared with that contained in [$L_{\underline{\boldsymbol{x}}}$]. Day and Kerridge (1967) give further details of the above argument."

2.4.3 Conclusions

Dans cette approche, on considère donc qu'en décomposant la log-vraisemblance sous la forme $\ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) + \sum_i \ln(P_X(\boldsymbol{x}_i))$, la partie non conditionnelle dépendant de la loi marginale de $\boldsymbol{X}$ ne fait pas intervenir (ou alors de manière négligeable) les paramètres du modèle. Les arguments utilisés ne constituent pas à notre avis une démonstration rigoureuse.

Nous avons d'ailleurs détaillé ce qu'il en était pour une loi d'une forme particulière appartenant à la famille exponentielle (cf. 2.3.3, p. 127). La nullité de la dérivée du reste de la vraisemblance n'est pas vérifiée pour cette loi.

L'étude des différences de traitement entre échantillon prospectif et échantillon de mélange a été largement moins traitée que celle du rapport entre échantillons prospectif et rétrospectif. En effet, ce dernier schéma d'échantillonnage représente le cas le plus fréquent dans la pratique (en particulier dans le domaine médical, où ce schéma s'adapte mieux aux données à traiter).

Nous allons voir dans la section suivante à quels types de traitements des paramètres fait aboutir l'expression de la vraisemblance dans le cas d'un échantillon rétrospectif.

2.5 Schéma rétrospectif et schéma prospectif : deux lectures de la régression logistique

Le protocole d'échantillonnage rétrospectif est un peu plus délicat à manier que le schéma de mélange car il nécessite l'introduction d'une variable T représentant l'appartenance ou non à l'échantillon traité, comme nous l'avons vu au paragraphe 2.2.2, p. 123.

Mais la log-vraisemblance issue d'un échantillon de ce type et celle définie pour ces mêmes données, comme si elles étaient issues d'un schéma prospectif, peuvent à nouveau être reliées :

$$\ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}}) = \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) + \sum_{i=1}^{N} \ln(P_X(\boldsymbol{x}_i)) - N_1 \ln \frac{N_1}{N} - N_0 \ln \frac{N_0}{N} .$$

Rappelons que $\ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y})$ correspond à la log-vraisemblance prospective pour une propriété logistique basée sur les paramètres $\beta = (\widetilde{\beta}_0, \beta_1, \ldots, \beta_L)$, avec $\widetilde{\beta}_0 = \beta_0 + \ln(\gamma_1/\gamma_0)$, où γ_0, γ_1 sont les taux de sondage.

2.5.1 Une approche par reparamétrisation a posteriori

Une première approche du problème consiste à chercher à utiliser la vraisemblance globale, mais en ne maximisant effectivement que la partie conditionnelle pour la recherche de l'estimateur de β. Nous allons voir comment est alors traitée la loi marginale des covariables.

• **Anderson (1972)**

Dans ce premier texte ([Anderson 72]), l'auteur a cherché à adapter la procédure de régression logistique à des données de type rétrospectif, alors que le modèle avait été défini pour des échantillons de mélange ou prospectifs. Ce problème peut se placer en particulier dans l'optique d'utiliser une règle de décision construite à partir de données rétrospectives pour affecter à un groupe de nouveaux individus provenant de protocoles de ces deux derniers types.

Nous résumons ici son propos en le rapportant au cas particulier d'une discrimination à deux groupes (Y binaire). Le modèle étudié se base sur l'hypothèse que la probabilité d'appartenance au groupe G_1 est donnée par :

$$P(Y = 1/\mathbf{X} = \boldsymbol{x}) = \frac{e^{\beta' \boldsymbol{x}}}{1 + e^{\beta' \boldsymbol{x}}} \, , \; \boldsymbol{x} \in \{0,1\}^{J+1} \, , \text{ où } x_0 = 1 \, .$$

Les covariables prennent dans un premier temps leurs valeurs dans un espace discret, mais l'auteur propose une adaptation de la démonstration pour le cas continu ou mixte.

Il n'y a pas dans cette démonstration d'utilisation d'une variable T indicatrice de l'appartenance à l'échantillon. Ainsi, la log-vraisemblance à maximiser s'écrit :

$$\ln L_{\underline{y}}(\beta, \underline{\boldsymbol{x}}) = \ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) + \sum_{i=1}^{N} \ln \left(P_X(\boldsymbol{x}_i)\right) - \alpha \, ,$$

où α est une constante dépendant des probabilités a priori π_0, π_1 d'appartenance aux deux groupes et du nombre d'individus N_0, N_1 de chaque groupe, fixés par le protocole (il n'y a pas ici de modification de la propriété logistique, le paramètre $\widetilde{\beta}_0$ n'est donc pas utilisé).

L'auteur fait le choix d'une maximisation sous contraintes. Les contraintes à considérer sont alors la contrainte sur la loi de $\mathbf{X}$ et celle sur la loi de Y, qui se réduit en fait à la donnée de π_1 puisque Y est binaire :

$$\left\{ \begin{array}{lcl} \sum_{\boldsymbol{x}} P_X(\boldsymbol{x}) & = & 1 \\ P_Y(1) & = & \sum_{\boldsymbol{x}} P(Y = 1/\mathbf{X} = \boldsymbol{x}) P_X(\boldsymbol{x}) = \pi_1 \, . \end{array} \right.$$

Une première étape de la démonstration consiste alors à montrer que la maximisation peut se faire indépendemment des valeurs π_0, π_1.

Ainsi, pour d'autres probabilités a priori π'_0 et π'_1, les paramètres du logit sont les mêmes, hormis pour la constante β_0 à laquelle il faut ajouter un coefficient dépendant de ces nouvelles proportions.

En effet, cette autre répartition de la population de base correspond à une autre loi marginale en $\mathbf{X}$ (notée P'_X) et à d'autres probabilités conditionnelles (notées $P'(Y = y/\mathbf{X} = \boldsymbol{x})$, $y \in \{0,1\}$). Mais la vraisemblance se définit toujours de la même manière :

$$L(\boldsymbol{x}/Y = y) = \frac{P(Y = y/\mathbf{X} = \boldsymbol{x})P_X(\boldsymbol{x})}{\pi_y} = \frac{P'(Y = y/\mathbf{X} = \boldsymbol{x})P'_X(\boldsymbol{x})}{\pi'_y}\ ,\ y \in \{0,1\}\ .$$

À partir de la *même* hypothèse logistique (c'est-à-dire à partir du même vecteur de paramètres $(\beta_0, \ldots, \beta_J)$), on obtient donc une écriture de $P'(Y/\mathbf{X})$ qui est la même que celle de $P(Y/\mathbf{X})$ avec un paramètre constant qui s'écrit alors $\beta_0 + \ln\left(\dfrac{\pi'_0\pi_1}{\pi_0\pi'_1}\right)$.

Cela n'implique pas forcément d'obtenir la même expression de la loi P_X du vecteur des prédicteurs, mais la recherche des paramètres peut se faire en prenant des valeurs quelconques des probabilités a priori d'appartenance aux groupes. Ainsi, il est possible d'effectuer la maximisation en supposant que $\pi_y = N_y/N$, $y \in \{0,1\}$. Le paramètre constant devra alors être ajusté en lui ajoutant le coefficient $\ln(\pi_1/\pi_0) + \ln(N_0/N_1)$.

On retrouve bien la correction qu'il avait été nécéssaire d'introduire lors de l'utilisation de la variable T (cf. paragraphe 2.2, p. 120). Corriger la constante β_0 revient donc là aussi à utiliser le fait que la loi de Y est fixée dans l'échantillon.

Dans une deuxième étape, l'auteur utilise alors la méthode des multiplicateurs de Lagrange pour obtenir les équations auxquelles doivent répondre les paramètres $(\beta_0, \ldots, \beta_J)$ *et* les probabilités marginales $P_X(\boldsymbol{x})$.

Le lagrangien Lg s'écrit :

$$Lg = \ln L_{\underline{y}}(\beta, \underline{\boldsymbol{x}}) + \lambda\left(\sum_{\boldsymbol{x}} P_X(\boldsymbol{x}) - 1\right) + \mu\left(\sum_{\boldsymbol{x}} P(Y = 1/\mathbf{X} = \boldsymbol{x})P_X(\boldsymbol{x}) - \pi_1\right)\ .$$

On utilise des proportions a priori π^* égales aux proportions empiriques N_y/N, $y \in \{0,1\}$. En annulant les dérivées de Lg, on obtient finalement les estimateurs suivant des probabilités $P_X(\boldsymbol{x})$:

$$\widehat{P}_X(\boldsymbol{x}) = \frac{n(\boldsymbol{x})}{N}\ ,\ \text{où } n(\boldsymbol{x}) = \text{card}\{\boldsymbol{x}_i = \boldsymbol{x}, i \in \{1, \ldots, N\}\}\ ,$$

et une série d'équations pour les estimateurs $\widehat{\beta}_j$ des paramètres β_j, $j \in \{0, \ldots, J\}$:

$$\sum_i y_i x_{ij} = \sum_i x_{ij}\widehat{\pi}(\boldsymbol{x}_i)\ ,\ \forall \boldsymbol{x} \in \{0,1\}^{J+1}\ ,$$

où $\widehat{\pi}(\boldsymbol{x})$ est l'estimation de π au point $\boldsymbol{x}$ à partir des valeurs $(\widehat{\beta}_0, \ldots, \widehat{\beta}_J)$.

Ces équations étant les mêmes que celles obtenues à partir de la vraisemblance d'un échantillon prospectif, l'auteur en conclut qu'il y une relation forte entre les deux schémas d'échantillonnage (*"there is a close relationship between the two estimation systems"*).

En fait, dans la dérivation du lagrangien, l'auteur ne fait pas intervenir les dérivées de la loi marginale de $\boldsymbol{X}$ en fonction des paramètres. Cela sous-entend donc que, comme pour l'argumentation précédente, il considère que cette loi ne contient pas d'informations sur les paramètres du modèle. Mais, contrairement aux articles précédents, il fournit une estimation de la loi marginale en $\boldsymbol{X}$, qui devient donc *a posteriori* un paramètre à part entière du modèle.

Dans l'optique de cette article, on voit donc apparaître une première solution permettant d'utiliser la log-vraisemblance prospective sur des données rétrospectives : il s'agit d'une **reparamétrisation du modèle**.

À partir d'une propriété logistique que l'on pourrait qualifier de "prospective" (puisqu'elle porte sur la loi conditionnelle en $\boldsymbol{X}$), l'auteur cherche à utiliser la log-vraisemblance que l'on aurait définie si les données avaient été justement issues d'un schéma prospectif. Cela aboutit, pour pouvoir mettre de coté "l'information" apportée par la loi marginale en $\boldsymbol{X}$ (comme dans l'étude d'un échantillon de mélange), à considérer toutes les probabilités $P_X(\boldsymbol{x})$ comme des paramètres. Ainsi, l'estimation des paramètres $(\beta_0, \ldots, \beta_J)$ est associée à la donnée d'une estimation de la loi marginale de $\boldsymbol{X}$. Cette distribution devient donc, a posteriori, non plus une fonction de nuisance, mais un paramètre du modèle.

On peut donc dire que cette méthode permet d'obtenir les estimations des paramètres d'un modèle répondant à une hypothèse logistique donnée, et dont on fournit une estimation de la loi marginale des covariables, sans avoir spécifié a priori les paramètres dont elle dépendait.

• **Anderson (1982)**

Nous présentons une autre approche du problème par le même auteur car il s'est basé pour celle-ci sur une propriété de linéarité du rapport des vraisemblances, dont nous avons vu qu'elle constituait la définition la plus générale du modèle du point de vue du traitement des covariables.

Dans cette autre partie du texte d'Anderson ([Anderson 82]) présenté dans le chapitre précédent, l'auteur suppose que la vraisemblance $L(\boldsymbol{X} = \boldsymbol{x}/Y = 0)$ peut s'exprimer comme une fonction $f(\boldsymbol{x})$ (f est une fonction qui dans le cas de données discrètes pourra par exemple être une distribution multinomiale).

On se place donc sous l'hypothèse suivante :

$$\ln \frac{L(\mathbf{X} = \boldsymbol{x}/Y = 1)}{L(\mathbf{X} = \boldsymbol{x}/Y = 0)} = \beta_0' + \sum_{j=1}^{J} \beta_j x_j \; .$$

L'expression de la log-vraisemblance pour un échantillon rétrospectif devient alors :

$$\ln L_{\underline{y}}(\beta, \underline{\boldsymbol{x}}) = \sum_{\boldsymbol{x}} N_1(\boldsymbol{x})(\beta_0' + \beta'\boldsymbol{x}) + \sum_{\boldsymbol{x}} N(\boldsymbol{x}) \ln f(\boldsymbol{x}) \; ,$$

où $N_1(\boldsymbol{x})$ représente le nombre d'observations $\boldsymbol{x}_i$ égales à $\boldsymbol{x}$ avec $y_i = 1$, $i \in \{1, \ldots, N\}$, et $N(\boldsymbol{x})$ le nombre d'observations $\boldsymbol{x}_i$ égales à $\boldsymbol{x}$ dans tout l'échantillon, pour une configuration $\boldsymbol{x}$ de $\mathbf{X}$ donnée.

À partir de cette vraisemblance, l'auteur va donc estimer les paramètres $(\beta_0, \ldots, \beta_J)$ du logit ainsi que la fonction f, qui devient un paramètre à part entière.

Il s'agit là aussi d'une maximisation sous contraintes :

$$\begin{cases} \sum_{\boldsymbol{x}} f(\boldsymbol{x}) & = \; 1 \\ \sum_{\boldsymbol{x}} f(\boldsymbol{x}) e^{\beta_0' + \beta'\boldsymbol{x}} & = \; 1 \; . \end{cases}$$

L'auteur adapte alors un résultat obtenu pour un échantillon composé d'une part d'individus de chaque groupe de la population en proportions connues et d'autre part d'individus de toute la population en proportions inconnues ([Anderson 79]). Le cas considéré revient donc à utiliser ce résultat en prenant le nombre d'individus provenant de la population entière en proportions inconnues égal à 0. On note N_y le nombre d'individus de l'échantillon provenant de chaque groupe $Y = y$, $y \in \{0, 1\}$.

Dans la démonstration du résultat général, la maximisation du lagrangien a été faite en admettant implicitement que les dérivées des vraisemblances $L(\boldsymbol{x}/Y = y)$, $y \in \{0, 1\}$, par rapport aux paramètres étaient nulles.

On obtient alors l'estimateur $\widehat{f}$ de la fonction f par maximisation sous contraintes de la vraisemblance de l'échantillon :

$$\widehat{f}(\boldsymbol{x}) = \frac{N(\boldsymbol{x})}{N_1 e^{\beta_0' + \beta'\boldsymbol{x}} + N_0} \; , \; \forall \; \boldsymbol{x} \; .$$

En remplaçant f par cet estimateur dans la log-vraisemblance, il s'agit alors de maximiser une expression notée $\ln L'_{\underline{\boldsymbol{x}}}(\beta, \underline{y})$

$$\ln L'_{\underline{\boldsymbol{x}}}(\beta, \underline{y}) = \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) - N_1 \ln N_1 - N_0 \ln N_0 + \sum_{\boldsymbol{x}} N(\boldsymbol{x}) \ln N(\boldsymbol{x}) \; ,$$

où $\ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y})$ est la log-vraisemblance définie pour un modèle basé sur la propriété suivante :

$$P(Y = 1/\mathbf{X} = \boldsymbol{x}) = e^{\beta_0' + \ln(N_1/N_0) + \beta'\boldsymbol{x}} \; ,$$

On retrouve bien la propriété énoncée au paragraphe 2.2.2, p. 123, sur l'ajustement à apporter au paramètre constant dans le cadre rétrospectif.

Après cette transformation de la log-vraisemblance, l'auteur conclut que les estimateurs des paramètres $(\beta_0, \ldots, \beta_J)$ du rapport des vraisemblances s'obtiennent donc en maximisant uniquement $\ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y})$.

Outre le fait que cette démonstration soit peu claire dans son déroulement, elle suppose de manière sous-jacente que chaque vraisemblance conditionnelle aux valeurs de Y ne contient aucune information sur les paramètres.

• **Prentice & Pyke (1979)**

Les auteurs étudient ici les liens qui existent entre deux définitions de modèles possèdant la propriété du logit linéaire. Ainsi, ils admettent deux choix de modélisation possibles, pour des covariables d'un type quelconque.

Le modèle prospectif est basé sur la propriété logistique décrivant les probabilités conditionnelles $P(Y = y/\boldsymbol{X} = \boldsymbol{x})$, $y \in \{0, 1\}$ en fonction des observations $\boldsymbol{x}$:

$$\begin{cases} \pi(\boldsymbol{x}) & = & P(Y = 1/\boldsymbol{X} = \boldsymbol{x}) & = & \dfrac{e^{\beta'\boldsymbol{x}}}{1 + e^{\beta'\boldsymbol{x}}} \\ 1 - \pi(\boldsymbol{x}) & = & P(Y = 0/\boldsymbol{X} = \boldsymbol{x}) & = & \dfrac{1}{1 + e^{\beta'\boldsymbol{x}}} . \end{cases}$$

Le modèle rétrospectif est basé sur la relation entre les probabilités conditionnelles en Y et les observations $\boldsymbol{x}$. Il s'agit en fait d'une reparamétrisation du modèle précédent :

$$\begin{cases} P(\boldsymbol{X} = \boldsymbol{x}/Y = 1) & = & c_1 e^{\gamma(\boldsymbol{x}) + \beta'\boldsymbol{x}} \\ P(\boldsymbol{X} = \boldsymbol{x}/Y = 0) & = & c_0 e^{\gamma(\boldsymbol{x})} , \end{cases} \tag{II.37}$$

où $\gamma(\boldsymbol{x}) = \ln \dfrac{P(\boldsymbol{X} = \boldsymbol{x}/Y = 0)}{P(\boldsymbol{X} = \boldsymbol{x}_0/Y = 0)}$, $\boldsymbol{x}_0$ étant un état "standard" des individus.

Les constantes de normalisation c_0 et c_1 dépendent donc des paramètres β et de la fonction γ : $c_y = c_y(\gamma, \beta) = P(\boldsymbol{X} = \boldsymbol{x}_0/Y = y)e^{-l\beta'\boldsymbol{x}_0}$, $y \in \{0, 1\}$.

Sur la base d'un modèle *prospectif*, un protocole d'échantillonnage *rétrospectif* implique d'ajuster l'expression des probabilités $P(\boldsymbol{X}/Y)$ par rapport aux données, et non pas celle des probabilités $P(Y/\boldsymbol{X})$.

Les auteurs utilisent cette reparamétrisation car ils estiment qu'elle clarifie le problème de l'estimation par maximisation de la vraisemblance prospective. Ils n'utilisent pas dans cette approche la variable T indicatrice d'appartenance à l'échantillon, mais il est possible de rapporter l'intégralité de leur démonstration au formalisme correspondant à l'utilisation de cette variable supplémentaire.

Ainsi, les auteurs définissent une fonction sur l'ensemble des configurations en $\boldsymbol{X}$ en utilisant la reparamétrisation II.37, qui est nécessaire puisque la définition de la vraisemblance se fait sur la base des probabilités conditionnelles en Y. Cette fonction q se trouve être la distribution du vecteur $\boldsymbol{X}$, conditionnellement au fait d'appartenir à l'échantillon :

$$q(\boldsymbol{x}) = P_X(\boldsymbol{X} = \boldsymbol{x}/T = 1) = e^{\gamma(\boldsymbol{x})} \sum_{0,1} \frac{N_y}{N} c_y e^{y\beta'\boldsymbol{x}} .$$

Cette fonction permet d'écrire les probabilités $P(\boldsymbol{X}/Y)$ sous la forme :

$$P(\boldsymbol{X} = \boldsymbol{x}/Y = y) = q(\boldsymbol{x}) \frac{c_y e^{y\beta'\boldsymbol{x}}}{\sum_{0,1} \frac{N_y}{N} c_y e^{y\beta'\boldsymbol{x}}} , \; \forall\, y \in \{0,1\} .$$

En posant $\delta_y = \ln(c_y N_y/N)$, on obtient :

$$P(\boldsymbol{X} = \boldsymbol{x}/Y = y) = q(\boldsymbol{x}) \left(\frac{e^{y\beta'\boldsymbol{x}} + \delta_y}{\sum_{0,1} e^{y\beta'\boldsymbol{x}} + \delta_y} \right) \frac{N_y}{N} , \; \forall\, y \in \{0,1\} .$$

Ce qui correspond bien au formalisme imposé par l'utilisation de T, car on peut relier les expressions utilisées en 2.2.2, p. 123, et les fonctions définies dans cet article :

$$\begin{cases} P_X(\boldsymbol{X} = \boldsymbol{x}/T = 1) & = & q(\boldsymbol{x}) \\ P(Y = y/T = 1) & = & \dfrac{N_y}{N} \\ P^*(Y = y/\boldsymbol{X} = \boldsymbol{x}) & = & \dfrac{e^{y\beta'\boldsymbol{x} + \delta_y}}{\sum_{0,1} e^{y\beta'\boldsymbol{x} + \delta_y}} , \end{cases}$$

où $P^*(Y = y/\boldsymbol{X} = \boldsymbol{x})$ désigne l'expression de cette probabilité conditionnelle sous l'hypothèse d'un modèle prospectif, mais avec un paramètre constant β_0^* modifié de la manière suivante :

$$\beta_0^* = \beta_0 + \delta_1 - \delta_0 .$$

Or, le facteur correctif $\delta_1 - \delta_0$ est effectivement celui introduit au paragraphe 2.2.2, p. 123, grâce aux probabilités a priori. En effet, on a :

$$\begin{aligned} \delta_1 - \delta_0 &= \ln(c_1 N_1/N) - \ln(c_0 N_0/N) \\ &= \ln(c_1/c_0) + \ln(N_1/N_0) . \end{aligned}$$

En utilisant le fait que $P(Y = y/\boldsymbol{X} = \boldsymbol{x}_0)$ répond à l'hypothèse du modèle prospectif, et par application des formules de Bayes, on peut simplifier $\ln(c_1/c_0)$

pour ne plus avoir à considérer d'état particulier $\boldsymbol{x}_0$:

$$\begin{aligned}
\ln(c_1/c_0) &= \ln \frac{P(\mathbf{X}=\boldsymbol{x}_0/Y=1)e^{-\beta'\boldsymbol{x}_0}}{P(\mathbf{X}=\boldsymbol{x}_0/Y=0)} \\
&= \ln \frac{1}{e^{\beta'\boldsymbol{x}_0}} + \ln \frac{P(\mathbf{X}=\boldsymbol{x}_0/Y=1)}{P(\mathbf{X}=\boldsymbol{x}_0/Y=0)} \\
&= \ln \left(\left(\frac{1}{1+e^{\beta'\boldsymbol{x}_0}} \right) \left(\frac{1+e^{\beta'\boldsymbol{x}_0}}{e^{\beta'\boldsymbol{x}_0}} \right) \right) + \ln \frac{P(\mathbf{X}=\boldsymbol{x}_0/Y=1)}{P(\mathbf{X}=\boldsymbol{x}_0/Y=0)} \\
&= \ln \frac{P(Y=0/\mathbf{X}=\boldsymbol{x}_0)}{P(Y=1/\mathbf{X}=\boldsymbol{x}_0)} + \ln \frac{P(\mathbf{X}=\boldsymbol{x}_0/Y=1)}{P(\mathbf{X}=\boldsymbol{x}_0/Y=0)} \\
&= \ln \frac{P_Y(0)}{P_Y(1)} \\
&= \ln \frac{\pi_0}{\pi_1} \,.
\end{aligned}$$

Le fait de définir ces deux modèles par reparamétrisation, et l'utilisation de cette transformation dans l'utilisation sur des données rétrospectives d'un modèle prospectif, revient donc au même que l'introduction de la variable T.

On retrouve donc bien la décomposition II.33 de la vraisemblance telle que nous l'avions présentée :

$$\ln L_{\underline{y}}(\widetilde{\beta}, \underline{\boldsymbol{x}}) = \ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y}) + \sum_{i=1}^{N} \ln(P_X(\boldsymbol{x}_i)) - N_1 \ln \frac{N_1}{N} - N_0 \ln \frac{N_0}{N} \,.$$

Les auteurs utilisent alors l'information contenue dans la log-vraisemblance de la même manière que Anderson dans son premier texte ([Anderson 72]) : les équations des estimateurs de $(\beta_0, \ldots, \beta_J)$ sont obtenues à partir de $\ln L_{\underline{\boldsymbol{x}}}(\widetilde{\beta}, \underline{y})$, mais une estimation de la loi marginale des $\mathbf{X}$ est aussi fournie sous la forme de la loi empirique du vecteur.

Une remarque de l'article permet de mettre l'accent sur le flou de ces démonstrations : "*note that if the prospective model were applied to the case-control data, as if the sampling were prospective, the likelihood equations would be exactly [the same]...*", (s'ajoute à cette remarque le fait que le paramètre constant n'aurait pas été modifié).

Ainsi, les auteurs considèrent qu'exploiter la partie conditionnelle de la log-vraisemblance revient exactement au même que d'exploiter l'intégralité de la log-vraisemblance correspondant à un échantillon rétrospectif. L'unique différence est la correction à apporter au paramètre constant. Ils en concluent qu'un modèle prospectif peut être appliqué à un échantillon rétrospectif ("*The maximum likelihood estimators of the odds ratio parameter* $\widehat{\beta}_j$, $j \in \{1, \ldots, J\}$ *can therefore be obtained* *from such an application of [II.34]"*). Mais ils ne prennent absolument pas en compte la perte d'information sur le vecteur des paramètres β.

La loi $P_X(\mathbf{X} = \boldsymbol{x}/T = 1) = q(\boldsymbol{x})$ peut de plus dépendre d'une autre famille de paramètres γ. Les estimateurs des paramètres du logit $(\beta_0, \ldots, \beta_J)$ seront les mêmes, quelle que soit cette nouvelle paramétrisation (dont les auteurs n'indiquent pas qu'elle est indépendante de β). Ils en concluent que des hypothèses ou des connaissances sur la loi $P_X(\mathbf{X}/T = 1)$ ne peuvent pas influencer l'estimation des paramètres, *tout du moins* pour des échantillons de grandes tailles.

2.5.2 Définition d'une procédure de régression logistique

Une solution pour éviter les problèmes posés par les différentes définitions de la vraisemblance, est de décider qu'une procédure de régression logistique est en fait toujours basée sur la maximisation d'une vraisemblance prospective, *quel que soit le type de l'échantillon utilisé.*

• Efron (1975)

L'article présenté maintenant ne traite pas vraiment de l'utilisation de la vraisemblance sous l'hypothèse d'un modèle à logit linéaire, mais de la comparaison de deux procédures de classement en terme d'efficacité relative.

On suppose ici que les covariables suivent une loi normale de paramètres (μ_y, Σ) avec la probabilité π_y, selon le groupe de la population, $y \in \{0, 1\}$. À partir de cette hypothèse il s'agit d'estimer les paramètres d'une fonction linéaire $\lambda(\boldsymbol{x}) = \beta' \boldsymbol{x}$, $\beta = (\beta_0, \ldots, \beta_J)$ (en posant comme toujours $x_0 = 1$), où :

$$\left\{ \begin{array}{ccl} \beta_0 & = & \ln \dfrac{\pi_1}{\pi_0} - \dfrac{1}{2} \left(\mu_1' \Sigma^{-1} \mu_1 - \mu_0' \Sigma^{-1} \mu_0 \right) \\ (\beta_1, \ldots, \beta_J) & = & (\mu_1 - \mu_0)' \Sigma^{-1} . \end{array} \right.$$

Le choix de l'affectation d'un nouvel individu à l'un des deux groupes de la population se fait sur la base de la valeur de cette fonction linéaire dite **fonction de Fisher**. Un individu dont l'observation des covariables est $\boldsymbol{x}$ est affecté à la population 1 si $\lambda(\boldsymbol{x}) > 0$, et à la population 0 sinon.

Il s'agit donc d'estimer les paramètres β de cette fonction. L'auteur met en parallèle deux méthodes :

La procédure de discrimination normale : on suppose que l'échantillon traité $(\underline{\boldsymbol{x}}, \underline{y})$ est tel que les observations $\boldsymbol{x}_i$ conditionnellement à y_i proviennent d'une loi normale $\mathcal{N}(\mu_{y_i}, \Sigma)$ et que les couples $(\boldsymbol{x}_i, y_i)$ sont indépendants, pour tout $i \in \{1, \ldots, N\}$.

Dans ce cas, les estimations des paramètres (Σ, π_y, μ_y), $y \in \{0,1\}$, s'obtiennent par maximisation de la vraisemblance de manière globale (sans avoir à la décomposer sous la forme II.33), ce qui permet d'obtenir :

$$\begin{cases} \widehat{\pi}_y &= N_y/N \\ \widehat{\mu}_y &= \overline{\boldsymbol{x}}_y \\ \widehat{\Sigma} &= \sum_{y\in\{0,1\}} \sum_{i=1} N_y(\boldsymbol{x}_i - \overline{\boldsymbol{x}}^y)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_y)' , \end{cases}$$

où $\overline{\boldsymbol{x}}^y$ est le vecteur des moyennes des observations de $\boldsymbol{X}$ dans le groupe G_y, $y \in \{0,1\}$. Les paramètres estimés $\widehat{\beta}$ sont obtenus comme fonctions de ces estimations.

La procédure de régression logistique : il n'y a plus pour cette autre méthode d'hypothèse particulière sur la densité conditionnelle de $\boldsymbol{X}$.

À partir de l'expression de la fonction de Fisher, qui est a posteriori le logarithme de l'odd-ratio de la population du groupe G_1 contre celle du groupe G_0, il est possible de retrouver l'écriture habituelle des probabilités $P(Y/\boldsymbol{X})$ correspondant à un modèle logistique :

$$\begin{cases} \pi(\boldsymbol{x}) &= P(Y=1/\boldsymbol{X}=\boldsymbol{x}) &= \dfrac{e^{\beta'\boldsymbol{x}}}{1+e^{\beta'\boldsymbol{x}}} \\ 1-\pi(\boldsymbol{x}) &= P(Y=0/\boldsymbol{X}=\boldsymbol{x}) &= \dfrac{1}{1+e^{\beta'\boldsymbol{x}}} , \ \forall \boldsymbol{x} \in \{0,1\}^{J+1} . \end{cases}$$

Pour un échantillon tel que les observations $(y_1, \ldots, y_N)$ sont indépendantes conditionnellement aux valeurs de $\boldsymbol{X}$, l'estimation $\widetilde{\beta}$ des paramètres β est obtenue directement, sur la base de la vraisemblance prospective $\ln L_{\underline{\boldsymbol{x}}}(\beta, \underline{y})$, sans considérations particulières sur le protocole d'échantillonnage à l'origine des données.

Une étude comparée des deux méthodes montre alors, grâce à une mesure de l'efficacité relative, que la procédure de régression logistique est moins efficace que celle de discrimination normale. D'un point de vue théorique, cela s'explique par le fait que cette deuxième méthode exploite l'intégralité de l'information contenue dans la vraisemblance. Par contre, l'utilisation répandue de la régression logistique vient du fait que les hypothèses sous-jacentes sont moins restrictives.

Bien qu'il ne base l'estimation des paramètres que sur la vraisemblance prospective, l'auteur signale tout de même que différents types d'échantillonnage sont possibles. Ainsi, dans un cadre rétrospectif, la procédure de discrimination normale reste totalement inchangée. Par contre, le traitement de la deuxième procédure est plus elliptique. Il admet finalement que le plus simple est alors *"d'ignorer"* le fait que l'échantillon est de nature rétrospective (le cas de mélange n'est pas abordé).

2.5.3 Conclusion

Des différentes approches que nous venons de détailler, nous devons tirer comme conclusion principale qu'il est nécessaire de différencier *propriété du logit linéaire* (et par extension polynomial) d'une part et *procédure de régression logistique* d'autre part. Cette procédure correspond à l'obtention des estimateurs des paramètres du logit d'un des modèles possèdant une propriété de linéarité donnée de ce logit. On se base pour cela sur une vraisemblance prospective (dans le sens où elle se définit à $\boldsymbol{X}$ contrôlé).

On ne s'intéresse donc pas avec cette procédure au paramétrage de la loi conjointe, mais à celui d'une fonction de l'espérance conditionnelle de la variable à expliquer. Nous allons voir à présent comment le principe de maximum d'entropie va nous permettre d'éclairer cette procédure de régression logistique sur la base d'observables constitués par un résumé simple des données et traduisant une situation de régression.

3 Relecture de la procédure de régression logistique par le principe de maximum d'entropie

3.1 Introduction

Nous allons voir comment la vision de cette procédure de régression logistique par le PME permet d'éclairer la manière dont celle-ci exploite les données. En nous basant sur notre logique d'utilisation du PME, nous expliquerons d'une part pourquoi tous les types de variables explicatives sont traités de la même manière et d'autre part pourquoi on ne différencie pas les protocoles d'échantillonnage dans les calculs d'estimation des paramètres du modèle.

La caractérisation d'un modèle à propriété du logit linéaire (ou polynomial) concerne une propriété des probabilités conditionnelles en $\boldsymbol{X}$, vecteur des variables explicatives.

Nous avons détaillé précédemment la nécessité de justifier l'équivalence des protocoles d'échantillonnage dans la détermination classique des équations des paramètres de ce logit par maximisation de la vraisemblance.

Nous ferons apparaître dans ce texte que les équations utilisées sont en fait des équations issues d'une application du PME où le choix des observables, la caractérisation du modèle et la détermination des paramètres se font sans préciser ce qui a pu être fixé a priori dans l'échantillon.

Nous considèrerons en effet une famille d'observables caractérisant une situation de régression où l'on cherche à prévoir une variable discrète Y à partir d'un vecteur de covariables $\boldsymbol{X}$ de type quelconque. Nous exploiterons les données disponibles de la manière la plus synthétique possible.

Ainsi, un résumé intuitif se base *au minimum* sur la loi empirique du vecteur des covariables d'une part et de celle de la variable de groupe d'autre part. En effet, dans le cas d'une variable explicative discrète, on résume naturellement les données par les marges du tableau de contingence.

Dans le cas qui nous importe, nous ne cherchons pas à étudier la loi de $\boldsymbol{X}$, aussi nous fixerons complètement la loi de ce vecteur par la *distribution empirique*. Pour pouvoir travailler systématiquement sur le même résumé des données, l'étude des covariables continues sera assimilée au cas discret.

L'information apportée par ces lois empiriques doit être complétée par une série d'observables traduisant le fait que nous cherchons à déterminer les lois d'appartenance aux groupes conditionnellement aux covariables. Ainsi, les liaisons entre au moins chaque covariable et la variable de groupe doivent intervenir dans l'information a priori sur laquelle se base la recherche du modèle.

3.2 Observables utilisés et espace de travail

Dans le cadre le plus général, $\boldsymbol{X} = (X_1, \ldots, X_J)$ le vecteur des covariables est à valeurs dans un espace $\mathcal{X}$ et Y la variable de groupe à valeurs dans un espace discret $\mathcal{Y}$. Les covariables sont toutes à support discret ordonné ou à support continu (elles peuvent avoir subi un traitement préalable pour que ce soit effectivement le cas). Nous nous attacherons au cas où Y est une variable binaire à valeurs dans $\{0, 1\}$.

Plutôt que de se placer dans l'espace de toutes les valeurs possibles du vecteur des variables explicatives $\boldsymbol{X}$, nous considèrerons le sous-espace $\mathcal{X}'$ de $\mathcal{X}$ qui représente l'ensemble des configurations en $\boldsymbol{X}$ *effectivement observées* pour les données dont on dispose.

D'une part, dans le cas où $\mathcal{X}$ est discret, cela permettra de définir sans problème de division par 0 la fréquence empirique d'appartenance à chaque groupe, conditionnellement aux valeurs possibles des configurations en $\boldsymbol{X}$ auxquelles on se limite. D'autre part, dans le cas où $\mathcal{X}$ est continu, on assimilera ainsi cette situation à la précédente. La manière de résumer l'information apportée par les données pourra alors se baser sur le même type d'observables puisque l'on se rapportera toujours à un espace fini.

La mesure de référence associée à cet espace de travail sera la mesure de comptage comme dans les applications précédentes du PME.

Que les covariables soient à valeurs dans un espace fini ou dans un espace continu, la recherche de la loi d'entropie maximum se fera donc à partir d'une même logique d'application du PME et pourra se baser sur le même type de résumé des données.

Pour faire de la régression logistique, nous appliquerons le PME dans le but de rechercher une loi sur $\mathcal{X}' \times \{0,1\}$ qui vérifie certaines contraintes basées sur un choix particulier de synthèse des données.

Ainsi, la recherche de la loi maximisant l'entropie se fera en fixant la loi de $\boldsymbol{X}$ et celle de Y grâce aux *distributions empiriques.* Cela assure en effet que le modèle obtenu soit vraiment en adéquation avec l'information apportée par l'échantillon. Une autre série d'observables devra de plus répondre au but poursuivi qui est la détermination des lois de Y *conditionnellement* au vecteur $\boldsymbol{X}$. Il faut donc au minimum faire apparaître chaque liaison $x_j y$, $j \in \{1,\ldots,J\}$ dans l'information a priori.

Nous avons à notre disposition des données qui proviennent d'un échantillonnage issu d'un protocole dont nous ne précisons pas le type (prospectif, rétrospectif, mélange).

• **Les observables $\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}}$, $\boldsymbol{x} \in \mathcal{X}'$,** constitués des indicatrices de chaque configuration observée du vecteur $\boldsymbol{X}$, permettent d'imposer de travailler à loi de $\boldsymbol{X}$ fixée. La contrainte associée à $\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}}$ est en effet :

$$f_x = E(\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}}) = P_X(\boldsymbol{x}) ,$$

où f_x désigne le niveau de contrainte empirique :

$$f_x = \frac{N(\boldsymbol{x})}{N},$$

$N(\boldsymbol{x})$ étant le nombre d'observations de la configuration $\boldsymbol{x}$, $\boldsymbol{x} \in \mathcal{X}'$.

• **L'observable** y fixe dans le modèle à obtenir la loi de cette variable, car la contrainte associée s'écrit pour une loi P quelconque :

$$\overline{y} = E(Y) = P_Y(1),$$

où $\overline{y}$ est la fréquence empirique du groupe G_1 dans le jeu de données :

$$\overline{y} = N_1/N.$$

• **Les observables $x_j y$, $j \in \{1,\ldots,J\}$,** sont les observables que nous avons convenu de considérer pour obtenir une loi de maximum d'entropie où les covariables et la variable de groupe ne sont pas indépendantes.

Si le modèle *minimal* obtenu ne convient pas, une étape supplémentaire consistera bien sûr en l'ajout des observables $x_i x_j y$, $(i,j) \in \{1,\ldots,J\}^2$, puis des observables $x_k x_i x_j y$, $(i,j,k) \in \{1,\ldots,J\}^3$, ...

Rappelons que les contraintes associées s'écrivent pour une loi P quelconque, soit en faisant apparaître les probabilités conjointes, soit en faisant apparaître les probabilités conditionnelles de Y sachant $\boldsymbol{X}$ *et* la loi marginale de $\boldsymbol{X}$:

$$\begin{aligned}\overline{x_j y} &= E(X_j Y) \\ &= \sum_{x,y} x_j y P(\boldsymbol{X}=\boldsymbol{x}, Y=y) \\ &= \sum_{x,y} x_j y P_X(\boldsymbol{x}) P(Y=y/\boldsymbol{X}=\boldsymbol{x}) \ , \ j \in \{1,\ldots,J\} \ , \qquad \text{(II.38)}\end{aligned}$$

où le niveau de contrainte $\overline{x_j y}$ est la moyenne empirique de $X_j Y$:

$$\overline{x_j y} = \frac{1}{N}\sum_{i=1}^{N} x_{ij} y_i \ , \ j \in \{1,\ldots,J\} \ .$$

(Les variables ont éventuellement subi un traitement préliminaire assurant que ces moyennes ont un sens sur l'espace $\mathcal{X}'$.)

Nous pouvons remarquer que les niveaux de contraintes sont toujours obtenus de la même manière, alors que le nombre de configurations observées en $\boldsymbol{X}$ ou en Y peut être fixé a priori par le protocole d'échantillonnage. Nous verrons à la fin de ce texte comment prendre en compte l'erreur éventuellement commise dans ces estimations.

L'information a priori est donc constituée ici par les couples observables et niveaux de contraintes estimés suivants :

$$\overline{\mathcal{I}} = \{(y, N_y/N), (x_j y, \overline{x_j y})_{j\in\{1,\ldots,J\}}, (\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}}, N(\boldsymbol{x})/N)_{x\in\mathcal{X}'}\} \ . \qquad \text{(II.39)}$$

Dans cette application du PME, nous recherchons donc *parmi toutes les lois possèdant des lois en* $\boldsymbol{X}$ *et en* Y *fixées par l'échantillon*, la loi de maximum d'entropie répondant à certaines contraintes minimales qui concernent les liens entre covariables et variable explicative.

3.3 Loi issue de l'application du PME

Nous nous sommes placés dans le cadre de la recherche de lois de maximum d'entropie pour un vecteur de variables toutes discrètes. La forme de la loi conjointe de maximum d'entropie est donc facilement déterminée. Mais nous voulons exprimer le modèle au travers de la forme du logit des probabilités de Y conditionnellement aux configurations des covariables.

3.3.1 Expression des probabilités conditionnelles d'appartenance aux groupes

Nous sommes donc dans le cadre de la détermination de la loi conjointe d'un vecteur $(\boldsymbol{X}, Y)$ de variables toutes discrètes, qui maximise l'entropie suivante (entropie de Shannon) :

$$S(P) = -\sum_{xy} p_{xy} \ln p_{xy} \,. \tag{II.40}$$

En reprenant les notations du paragraphe 2.3.1, p. 84, nous pouvons résumer la recherche de la loi de maximum d'entropie grâce au tableau II.8.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
$\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}}$	$N(\boldsymbol{x})/N$	$\gamma_x,\ \boldsymbol{x} \in \mathcal{X}'$
y	N_y/N	λ_0
$x_j y$	$\overline{x_j y}$	$\lambda_j,\ j \in \{1, \ldots, J\}$

État de Gibbs

$$p^*_{xy} = \frac{1}{Z} \exp\left(\sum_x \gamma_x \mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}} + \lambda_0 y + \sum_j \lambda_j x_j y \right)$$

Équations du PME

$$P^*_X(\boldsymbol{x}) = N(\boldsymbol{x})/N$$

$$p^*_y = N_y/N$$

$$\sum_{x,y} x_j y p^*_{xy} = N_1/N\overline{x_j}^{1}$$

Constante de normalisation

$$Z = \exp\left(\sum_x \gamma_x \mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}} + \lambda_0 y + \sum_j \lambda_j x_j y \right)$$

TAB. II.8 – *Loi conjointe de maximum d'entropie obtenue pour l'information* $\overline{\mathcal{I}}$

Nous cherchons ici à exprimer le modèle obtenu en mettant l'accent sur les probabilités conditionnelles d'appartenance aux groupes. L'état de Gibbs qui réalise le maximum de S correspond aux probabilités conditionnelles suivantes :

$$\left\{ \begin{array}{rclcl} P^*(Y=1/\boldsymbol{X}=\boldsymbol{x}) & = & \pi^*(\boldsymbol{x}) & = & \dfrac{\exp\left(\sum_j \lambda_j x_j + \lambda_0\right)}{1+\exp\left(\sum_j \lambda_j x_j + \lambda_0\right)} \\ P^*(Y=0/\boldsymbol{X}=\boldsymbol{x}) & = & 1-\pi^*(\boldsymbol{x}) & = & \dfrac{1}{1+\exp\left(\sum_j \lambda_j x_j + \lambda_0\right)} \,. \end{array} \right.$$

Le logit de ces probabilités conditionnelles s'écrit donc de la manière suivante, pour toute observation $\boldsymbol{x}$ dans $\mathcal{X}'$:

$$\text{logit } \pi^*(\boldsymbol{x}) = \ln\left(\frac{\pi^*(\boldsymbol{x})}{1-\pi^*(\boldsymbol{x})} \right) = \sum_j \lambda_j x_j + \lambda_0 \,.$$

En notant $x_0 = 1$, cette expression devient :

$$\text{logit } \pi^*(\boldsymbol{x}) = \sum_{j=0}^{J} \lambda_j x_j \, . \tag{II.41}$$

Le modèle obtenu est donc bien un **modèle à propriété du logit linéaire**, présentant des lois marginales en $\boldsymbol{X}$ et Y égales aux lois empiriques.

3.3.2 Équations du PME

On obtient les équations qui permettent de déterminer les multiplicateurs à partir des contraintes II.38 exprimées grâce aux probabilités conditionnelles en Y, en utilisant la contrainte qui fixe la loi de $\boldsymbol{X}$ à la loi empirique :

$$\frac{1}{N}\sum_{i=1}^{N} x_{ij} y_i = \sum_x x_j \frac{N(x)}{N} \left(\frac{\exp\left(\sum_j \lambda_j x_j\right)}{1 + \exp\left(\sum_j \lambda_j x_j\right)} \right) \, , \; j \in \{0, \ldots, J\} \, .$$

(La contrainte associée à l'observable y correspond au cas $j = 0$.)

En homogénéisant les indices, on obtient finalement les équations suivantes auxquelles répondent les paramètres $(\lambda_0, \ldots, \lambda_J)$:

$$\sum_i x_{ij} y_i = \sum_i x_{ij} \left(\frac{\exp\left(\sum_j \lambda_j x_j\right)}{1 + \exp\left(\sum_j \lambda_j x_j\right)} \right) \, , \; j \in \{0, \ldots, J\} \, . \tag{II.42}$$

Nous retrouvons donc les équations des estimateurs des paramètres d'un modèle à propriété du logit linéaire obtenues par maximisation d'une vraisemblance prospective.

▷ Remarques

L'observable y permet de faire apparaître un paramètre constant dans le logit des probabilités conditionnelles. Si on recherche un modèle de maximum d'entropie dans un but de discrimination en se plaçant dans le cadre du "diagnostic paradigm", sans fixer la loi de Y, il n'y aura donc pas de paramètre constant dans le logit des probabilités conditionnelles.

Cette constatation est à rapprocher des incohérences observées sur les frontières de discrimination obtenues dans le cadre du "sampling paradigm" à partir des modèles de maximum d'entropie correspondant à des information ne contenant pas l'observable y.

Si la loi des covariables n'est pas fixée par l'information, les équations du PME ne seront plus les mêmes que celles du maximum de vraisemblance. Dans le cadre des modèles exprimés sous la forme des lois conditionnelles des covariables, l'équivalence entre équations du PME et équations du maximum de vraisemblance était assurée à la condition de fixer la loi de Y à la loi empirique.

3.3.3 Informations sur les covariables et linéarité du logit

On peut remarquer que la forme des observables portant *uniquement* sur les covariables n'entrent pas en jeu dans la détermination de la forme du logit des probabilités conditionnelles d'appartenance aux groupes. Considèrons par exemple l'information de forme générale suivante :

$$\mathcal{I}' = \left\{ (\varphi_r(\boldsymbol{x}), a_r)_{r\in\{1,\dots,R\}}, (y, b), (x_j y, c_j)_{j\in\{1,\dots,J\}} \right\} , \tag{II.43}$$

où $\varphi_r(\boldsymbol{x})$ désigne un observable fonction des covariables uniquement et (a_r, b, c_j), $r \in \{1, \dots, R\}$, $j \in \{1, \dots, J\}$, les niveaux de contraintes associés (estimés ou fixés a priori). La loi conjointe de maximum d'entropie a alors l'expression suivante :

$$P^*(\boldsymbol{X} = \boldsymbol{x}, Y = y) = \frac{1}{Z} \exp\left(\sum_j \lambda_j x_j y + \lambda_0 y + \sum_r \gamma_r \varphi_r(\boldsymbol{x}) \right) .$$

Le logit qui correspond à cette loi est donc là encore linéaire (en notant $x_0 = 1$) :

$$\text{logit } \pi^*(\boldsymbol{x}) = \sum_{j=0}^{J} \lambda_j x_j .$$

Si on ne s'intéresse qu'à la forme du logit des probabilités conditionnelles en Y, seuls les observables du type $x_j y$ (et plus généralement $\varphi(\boldsymbol{x})y$) rentrent en jeu dans sa détermnination. Les multiplicateurs qui étaient associés aux observables $\varphi_r(\boldsymbol{x})$, $r \in \{1, \dots, R\}$, dans la loi conjointe ne rentrent en effet plus en ligne de compte dans l'expression du logit. Par contre, pour déterminer les valeurs des paramètres de ce logit, il faut résoudre l'ensemble des équations du PME, qui font apparaître l'ensemble des multiplicateurs.

▷ **Remarque**

L'information $\overline{\mathcal{I}}$ avait été définie à partir de covariables prenant leurs valeurs dans un espace discret représentant uniquement les valeurs effectivement observées. Nous venons de voir que pour des covariables à valeurs dans un espace quelconque $\mathcal{X}$, l'utilisation d'un logit des probabilités conditionnelles en Y linéaire correspond en fait à l'utilisation d'un modèle de maximum d'entropie issu d'une information du type II.43 basée sur les observables $(x_j y)_{j\in\{1,\dots,J\}}$, y, et une famille d'observables portant sur les covariables. Les informations concernant celles-ci sont “oubliées” dans l'expression du logit, mais pas dans les équations à résoudre pour déterminer des paramètres.

3.4 Influence du protocole d'échantillonnage

Dans l'optique d'application du PME que nous avons choisie, la loi de maximum d'entropie est toujours déterminée à lois de $\boldsymbol{X}$ et de Y fixées par les lois empiriques.

Selon le type de protocole à l'origine des données le nombre de configurations observées en $\boldsymbol{X}$ ou en Y peut être fixé a priori, mais l'écriture du modèle et les équations II.41 auxquelles doivent répondre les paramètres du logit seront toujours celles déterminées précédemment. Le type du protocole n'influence pas l'écriture de l'entropie de la loi conjointe, ni sa maximisation, car la famille d'observables a été choisie en dehors de toute considération sur ce protocole.

L'origine des données n'est pas à prendre en compte au niveau de la définition des contraintes, ni de l'estimation des niveaux de contraintes, mais au moment de *l'interprétation des paramètres du logit obtenu.*

On note T la variable indiquant si un individu appartient ou non à l'échantillon. Nous allons détailler le rapport entre la loi conjointe de $\boldsymbol{X}$ et Y obtenue précédemment (notée P^*_{XY}) et la loi de ces mêmes variables conditionnellement à $\{T = 1\}$ (notée $P^*_{XY/T=1}$) répondant à la même information a priori II.39.

• **Dans le cas d'un échantillon de mélange :** $\boldsymbol{X}$, Y, T sont indépendantes. La loi conditionnelle $P^*_{XY/T=1}$ est exactement la loi conjointe P^*_{XY}.

• **Dans le cas d'un échantillon prospectif :** l'entropie $S(P_{XY/T=1})$ se décompose comme la somme de $S(P_{X/T=1})$ et de $E_X(S(P_{Y/X}))$.

La loi $P^*_{X/T=1}$ issue de $P^*_{XY/T=1}$ est fixée par la loi empirique en $\boldsymbol{X}$. La maximisation de $E_X(S(P_{Y/X}))$ conduit à l'expression des lois conditionnelles que nous avons détaillée précédemment. On obtient la même expression de la loi d'entropie maximum conditionnellement à $\{T = 1\}$ sous la forme de sa décomposition $(P^*_{X/T=1}, (P^*_{Y/X=x})_{x\in\mathcal{X}'})$.

• **Dans le cas d'un échantillon rétrospectif :** l'entropie $S(P_{XY/T=1})$ se décompose comme la somme de $S(P_X)$ et de $E_X(S(P_{Y/X,T=1}))$ (car T et Y ne sont pas indépendantes dans ce type d'échantillonnage).

Les contraintes correspondantes s'écrivent aussi sur la base des lois conditionnelles à $\boldsymbol{X}$ et $\{T = 1\}$. L'expression de $P^*(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1)$, pour une configuration donnée $\boldsymbol{x}$, $\boldsymbol{x} \in \mathcal{X}'$, s'écrit donc sous la forme déterminée précédemment.

On utilise le résultat classique qui fait intervenir T grâce aux taux de sondages pour retrouver l'expression de $P^*(Y = 1/\boldsymbol{X} = \boldsymbol{x})$:

$$P^*(Y = 1/\boldsymbol{X} = \boldsymbol{x}, T = 1) = \frac{P^*(Y = 1/\boldsymbol{X} = \boldsymbol{x})\gamma_1(\boldsymbol{x})}{P^*(Y = 1/\boldsymbol{X} = \boldsymbol{x})\gamma_1(\boldsymbol{x}) + P^*(Y = 0/\boldsymbol{X} = \boldsymbol{x})\gamma_0(\boldsymbol{x})},$$

où γ_y, $y \in \{0, 1\}$, désignent les taux de sondage utilisés pour le tirage de l'échantillon.

On obtient donc le logit de la loi conditionnelle cherchée :

$$\operatorname{logit} \pi^*(\boldsymbol{x}) = \sum_{j=0}^{J} \lambda_j x_j - \ln \frac{\gamma_1}{\gamma_0} .$$

On retrouve donc le fait qu'une correction dépendant des taux de sondages doit être effectuée sur le paramètre constant dans le cas d'un échantillon rétrospectif.

3.5 Conclusion

Dans la méthode classique utilisée pour obtenir les estimations des paramètres d'un modèle possèdant une propriété logistique, il est nécessaire d'introduire un principe de maximisation d'une vraisemblance prospective. Ce principe qui peut apparaître très artificiel et ad hoc, impose de plus de montrer qu'il se ramène asymptotiquement au principe de maximum de vraisemblance classique.

Dans la vision que nous proposons, la méthode de régression logistique ne relève que des principes standards d'application du PME qui sont exposés dans le chapitre de présentation de ce principe. Le PME permet en effet de retrouver l'expression de ce modèle en restant cohérent par rapport à une certaine logique d'utilisation des données.

Les équations associées à la méthode de régression logistique sont des équations issues du PME pour une famille d'observables définie sans avoir à se préoccuper du type d'échantillonnage.

Il n'est donc plus nécessaire ici de montrer que les trois types de protocole sont équivalents au sens de ces équations. Mais pour faire de la prévision sur la variable Y, il est nécessaire de connaître la nature du protocole à l'origine des données. En effet, si la répartition de cette variable est fixée a priori par le protocole et n'est pas une image de la population ciblée, il est nécessaire de rajouter de l'information aux données issues de l'échantillon. Cela se traduit par une correction du paramètre constant basée sur l'introduction des taux de sondage dans le cas d'un échantillonnage de type rétrospectif.

Bibliographie

[Anderson 72] J.A. Anderson. “Separate Sample Logistic Discrimination”, *Biometrika*, **59**, 19-35, 1972.

[Anderson 79] J.A. Anderson. “Multivariate Logistic Coumpounds”, *Biometrika*, **66, 1**, 17-26, 1979.

[Anderson 82] J.A. Anderson. “Logistic Discrimination”, in *Classification, pattern recognition and reduction of dimensionality*, *Handbook of Statistics*, **2**, edited by P. Krishnaiah, L. Kanal, North-Holland, 169-191, 1982.

[Anderson 84] T.W. Anderson. *An introduction to multivariate statistical analysis*, 2nd ed., J. Wiley & sons, 1984.

[Balakrishnan 92] N. Balakrishnan (ed.). *Handbook of the logistic distribution*, Marcel Dekker, 1992.

[Berkson 44] J.Berkson. “Application of the logistic function to bio-assay”, *Journal of the American Statistical Association*, **37**, 357-365, 1944.

[Berkson 51] J. Berkson. “Why I prefer logits to probits”, *Biometrics*, 327-339, 1951.

[Bouyer 91] J. Bouyer. “La régression logistique en épidémiologie”, *Revue d'epidémiologie et santé publique*, **39**, 79-87, 183-196, 1991.

[Celeux 90] G. Celeux (ed.). *Analyse discriminante sur variables continues*, INRIA, Collection didactique, 1990.

[Celeux & Nakache 94] G. Celeux & J.-P. Nakache (eds.). *Analyse discriminante sur variables qualitatives*, Polytechnica, 1994.

[Cox 58] D.R. Cox. “The regression analysis of binary sequences”, *Journal of the Royal Statistical Society - Series B*, **20**, 2, 215-242, 1958.

[Cox 66] D.R. Cox. “Some Procedures Connected with the Logistic Qualitative Response Curve”, in *Research Papers in Statistics*, edited by Festschr. J. Neyman, 55-71, 1966.

[Das Gupta 73] S. Das Gupta. “Theories and methods in classification : a review”, in *Discriminant analysis and applications*, edited by T. Cacoullos, Academic Press, 77-137, 1973.

[Day & Kerridge 67] N.E. Day & D.F. Kerridge. “A General Likelihood Discriminant”, *Biometrics*, **23**, 313-323, 1967.

[Dawid 76] A.P. Dawid. “Properties of diagnostic data distributions”, *Biometrics*, **32**, 647-658, 1976.

[Efron 75] B. Efron. “The efficiency of Logistic Regression Compared to Normal Discriminant Analysis”, *Journal of the American Statistical Association*, **70**, 892-898, 1975.

[Goldstein & Dillon 78] M. Goldstein & W.R. Dillon. *Discrete discriminant analysis*, J. Wiley & sons, 1978.

[Hosmer & Lemeshow 89] D.H. Hosmer & S. Lemeshow. *Applied logistic regression*, Wiley, 1989.

[Kim 96] H.-J. Kim. “On a constrained optimal rule for classification with unknown prior individual group membership”, *Journal of Multivariate Analysis*, **59**, 166-186, 1996.

[Mc Cullagh 80] P. Mc Cullagh. “Regression models for ordinal data”, *Journal of the Royal Statistical Society - Series B*, **42**, 109-142, 1980.

[McLachlan 92] G. McLachlan. *Discriminant analysis and statistical pattern recognition*, J. Wiley & sons, 1992.

[McLachlan & Basford 88] G. McLachlan & K.E. Basford. *Mixture models : inference and applications to clustering*, Marcel Dekker, 1988.

[Pearl & Reed 20] R. Pearl & L.J. Reed, *Proc.Nat.Acad.Sci.*, 6, 275-288, 1920.

[Pearl *et al.* 40] R. Pearl, L.J. Reed, J.F. Fish. “The logistic curve and the census count of 1940”, *Science*, **92**, 486-488, 1940

[Prentice & Pyke 79] R.L Prentice & R. Pyke. “Logistic Disease Incidence Models and Case-control Studies”, *Biometrika*, **66, 3**, 403-411, 1979.

[Reed & Berkson 29] L.J. Reed & J. Berkson. “The application of the logistic function to experimental data”, *Journal of physical chemistry*, **33**, 760, 1929.

[Ripley 96] B.D. Ripley. *Pattern recognition and neural networks*, Cambridge University Press, 1996.

[Titterington *et al.* 96] D.M. Titterington, A.F.M. Smith, U.E. Makov. *Statistical analysis of finite mixture distributions*, J. Wiley & sons, 1985.

[Tomassone, Daudin, Danzart & Masson 88] R. Tomassone, J.-J. Daudin, M. Danzart & J.-P. Masson, *Discrimination et classement*, Masson, 1988.

[Verhulst 1838] P.J. Verhulst. "Notice sur la loi que la population suit dans son accroissement", *Corr. Math. Phys.*, **10**, 113-121, 1838.

[Verhulst 1845] P.J. Verhulst. "Recherche mathématique sur la loi d'accroissement de la population", *Acad. Bruxelles*, **18**, 1-38, 1845.

[Wilson & Worcester 43] E.B. Wilson & J. Worcester. "The determination of L.D.50 and its sampling error in bio-assay III", *Proceedings of the National Academy of Sciences*, **29**, 114, 1943.

Chapitre III

Un exemple de construction de modèles de maximum d'entropie

1 Introduction

Nous nous sommes intéressés dans ce travail à la modélisation des durées de séjour des patients du service de gériatrie du CHU de Grenoble. La variable X mesurée est le nombre de nuits passées dans ce service. Si X prend la valeur n, $n \geq 1$, cela signifie que la durée de séjour exprimée en heures est comprise entre $24n-12$ et $24n+12$. La base de données utilisée correspond aux patients hospitalisés entre le premier janvier 1994 et le 31 décembre 1995 (2292 patients enregistrés successivement). Ces durées sont un bon reflet de la charge de travail et des besoins financiers d'un service car elles doivent permettre de mesurer l'occupation des lits.

Les responsables de ce service ont souhaité une modélisation portant sur les durées de moins de 4 mois. On ne met alors de coté que 2.5 % des valeurs qui s'étendent en fait sur une échelle de 120 à 622 jours de manière très sporadique. La figure III.1 présente l'histogramme des durées de séjour de moins de 120 jours, ainsi que les fréquences d'apparition de chaque durée observée et quelques caractéristiques de la distribution.

C'est sur cet échantillon de taille 2234, pour des valeurs de X qui s'étendent de 1 à 120 jours, que nous travaillerons.

Cette distribution empirique a une forme bien spécifique, en particulier pour les valeurs de moins de 30 jours. Aucune symétrie n'apparaît, et on peut remarquer deux modes relatifs aux valeurs 1 (4.65% des personnes ne restent qu'une journée) et 14 (5.15% des personnes restent deux semaines). De plus, le troisième quartile vaut 28 alors que les valeurs s'étendent jusqu'à 120.

Le caractère irrégulier marqué du début de cette distribution est le reflet de la nature particulière de ce service hospitalier. Le premier mode correspond principalement aux patients qui doivent subir un examen à jeun et qui arrivent la veille à l'hôpital. Des durées de 1 ou 2 jours pourraient être qualifiées d'hospitalisations "anormales", car ce service est plutôt destiné à des hospitalisations plus ou moins longues et non à de simples examens. De même les individus restant entre 3 et 5 jours sont généralement des personnes réorientées vers d'autres services. La forme de cette distribution pourrait aussi s'expliquer par certaines politiques de sorties des patients (peu de sorties avant le week-end, peu de sorties dans la semaine d'arrivée, difficulté de réorienter les patients vers d'autres services plus spécialisés, ...).

Aucune loi univariée simple ne convient donc à première vue pour modéliser ces durées (l'utilisation d'une loi exponentielle ou log-normale en particulier ne convient pas). Des essais de modélisation par des mélanges de lois exponentielles ont été tentés, mais les résultats ne sont pas entièrement satisfaisants (cf. [Delhumeau 97]).

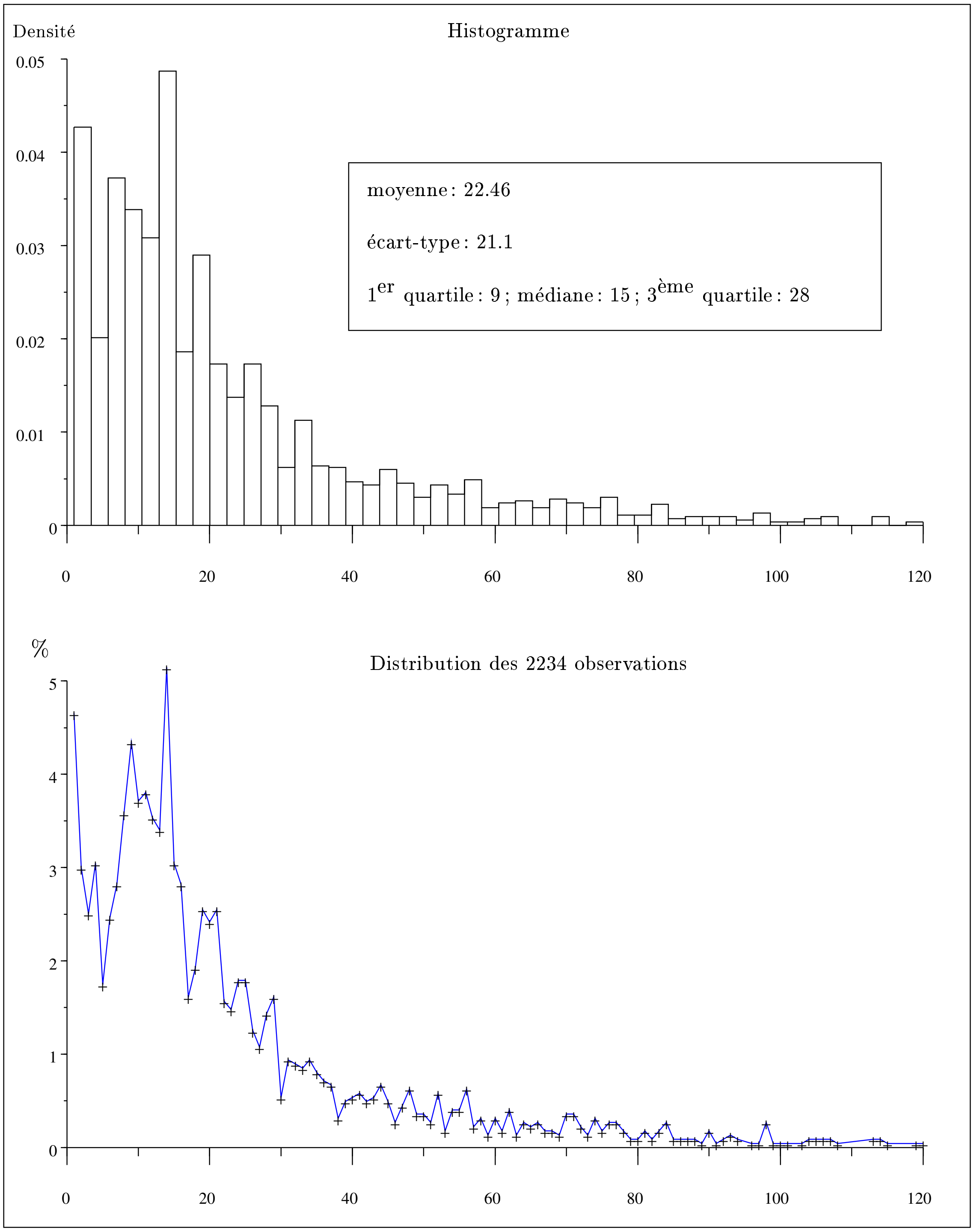


FIG. III.1 – *Distribution empirique des durées de séjour de moins de 120 jours*

C'est donc sur cette distribution que nous allons tester la modélisation par PME. En partant d'observables simples, nous observerons l'évolution des lois de maximum d'entropie quand on modifie les observables utilisés et nous chercherons à déterminer la famille d'observables qui permet d'obtenir un modèle adéquat.

Pour tester graphiquement la cohérence des modèles obtenus par rapport à cette distribution, nous pourrons comparer les densités des modèles et celle-ci. Mais les responsables de ce service nous ont de plus demandé un type de validation précis. Il s'agit de comparer les proportions empiriques et théoriques d'intervalles de durées prédéfinies. Ces intervalles sont les suivants :

- **durées de 1 à 2 jours :** cela correspond aux hospitalisations "anormales" ;
- **durées de 3 à 5 jours :** cela correspond aux réorientations ;
- **durées de 6 à 30 jours, par intervalles de 5 jours :** le traitement évolue à peu près de 5 en 5 jours durant cette période ;
- **durées de 31 à 120 jours, par intervalles de 10 jours :** (avec regroupement des deux derniers intervalles) le traitement évolue de moins en moins pour ces longs séjours.

Cela nous permettra de construire des tests d'adéquation entre les fréquences empiriques de ces 15 intervalles et les fréquences calculées à partir des différents modèles.

Dans un premier temps, nous détaillerons les différents types d'informations considérées et la forme des lois de maximum d'entropie correspondantes. Les différents modèles seront déterminés en utilisant l'ensemble des 2234 données. Mis à part le choix de l'espace de travail, nous ne disposons d'aucune information a priori sur le comportement de la variable X. Nous ne travaillerons donc qu'avec des niveaux de contraintes estimés à partir des données. Nous commenterons ensuite le comportement de la statistique de test choisie pour comparer fréquences empiriques et estimées en fonction du type et de la taille de l'information.

Une fois choisi pour chaque famille le modèle jugé le plus adéquat, au sens de la comparaison précédente, celui-ci sera validé par *half-sampling* en utilisant ce même test d'adéquation. Pour tester la stabilité des meilleurs modèles obtenus, nous utiliserons donc deux échantillons de tailles égales, tirés au hasard à partir de l'ensemble des observations. Pour chaque modèle construit sur chaque moitié des observations (c'est-à-dire à partir des niveaux de contraintes calculés sur ce sous-échantillon), nous utiliserons l'autre sous-échantillon pour construire les fréquences empiriques des intervalles prédéfinis et établir à nouveau un test d'adéquation. Le sous-échantillon ayant servi à construire le modèle et le sous-échantillon test étant différents, les *p-values* des tests d'adéquation seront calculées dans ce cas à partir d'un Khi-2 à 14 ddl.

2 Outil et méthode de modélisation

Une fois choisie une famille d'observables, la recherche d'une loi de maximum d'entropie est en fait la résolution d'un système d'équations non linéaires (se reporter au chapitre 4.2, p. 29, sur les équations du PME).

2.1 Modèles continus d'une variable discrète

Le logiciel SCILAB nous a fourni l'outil informatique de résolution de telles équations. Les programmes utilisés sont basés sur une routine provenant de la bibliothèque fortran MINPACK d'algorithmes d'optimisation pour la résolution de systèmes non linéaires et de problèmes de minimisation. L'algorithme utilise une adaptation de la méthode de Powell (*Powell hybrid method*[1]) (cf. [Powell 70]). Les programmes imposent de préciser un seuil tel que l'algorithme stoppe la recherche des solutions lorsqu'il estime que les différences entre les solutions du système et les valeurs obtenues sont au plus égales à ce seuil. Les résultats ont été obtenus avec un seuil de 10^{-4}. Les solutions obtenues ont été éditées et réutilisées avec un format de 8 chiffres par nombre. L'adéquation des solutions obtenues avec les contraintes a été vérifiée en utilisant le logiciel MAPLE.

Pour pouvoir utiliser de manière relativement simple cette routine, nous avons construit les modèles sur un support continu et non sur le support discret $\{1, \dots, 120\}$. Les données correspondent en fait à une variable de nature continue. Les observations s'étendent en effet sur une échelle très large et les mesures pourraient être faite de manière beaucoup plus précise (en demi journées, ou même en heures de présence dans le service).

Ce passage du discret au continu pour revenir au discret, nécessite de définir l'intervalle sur lequel les modèles seront construits. Chaque valeur discrète n sera en fait associée à un intervalle de longueur ϵ donnée $[n-\epsilon, n+\epsilon]$, et l'intégrale de la densité obtenue sur cet intervalle sera une approximation de la probabilité de cette valeur discrète : $p_n = \int_{n-\epsilon}^{n+\epsilon} f^*(t)dt$, $n \in \{1, \dots, 120\}$, où f^* est la densité obtenue par application du PME pour une information donnée.

Nous avons choisi de prendre des intervalles tous d'amplitude 1 autour des valeurs discrètes. Les différents modèles ont donc été construits sur l'intervalle $[0.5, 120.5]$.

▷ Remarque

D'autres essais ont été effectués, en construisant les modèles sur l'intervalle $[1, 120]$. Cela impose d'approximer p_1 par $\int_0^{0.5} f^*(t)dt$ et p_{120} par $\int_{119.5}^{120} f^*(t)dt$ (c'est-à-dire par des intervalles deux fois moins grands que ceux utilisés pour estimer p_n,

1. On pourra se reporter au site du centre CHPC d'informatique de l'université d'Utah pour plus de détails (*http://www.usi.utah.edu/documentation/software/math/minpack/hybrd1.f.*) ou au site *http://www.netlib.org/minpack/ex/* pour une documentation et des exemples

$n \in \{2, \ldots, 119\}$). Il est apparu que la forme du début des densités et les résultats des tests d'adéquation sont très sensibles au choix de cet intervalle. Les résultats sont en fait bien meilleurs pour les modèles que nous présentons ici.

2.2 Familles d'observables testées

Puisqu'il s'agit d'une variable continue, les familles d'observables les plus simples que nous pouvons utiliser sont celles basées sur les observables x^k, jusqu'à un ordre K, $K \in \mathbf{N}$. Les niveaux de contraintes associés sont alors les K **moments empiriques successifs** $\overline{x^k}$**,** $k \in \{1, \ldots, K\}$.

Mais nous pouvons modéliser non pas les données elles-mêmes, mais une transformation de celles-ci. Nous utiliserons ici les transformations classiques utilisant soit les **inverses** des données, soit les **logarithmes** des données. Nous utiliserons alors à nouveau des informations basées sur les moments successifs de ces données transformées. Cette transformation permet de travailler sur un intervalle plus réduit ($[1/120.5, 1/0.5]$ ou $[\ln(0.5), \ln(120.5)]$), à partir de moments empiriques variant sur une moins grande échelle.

Nous présenterons également les résultats obtenus à partir d'une information mélant des observables successifs x^k et des observables inverses $1/x^m$. Les niveaux de contraintes associés sont les K **moments empiriques successifs** $\overline{x^k}$ et les K **moments successifs de l'inverse des données** $\overline{1/x^m}$, $k \in \{1, \ldots, K\}$, $m \in \{1, \ldots, M\}$. Nous pourrons ainsi vérifier si l'utilisation d'observables autres que les observables élémentaires précédents peut permettre "d'attraper" le décrochement du début de la distribution. Ainsi, nous testerons d'une part l'association entre des moments successifs et l'unique observable $1/x$, et d'autre part l'association entre K moments successifs et les K moments successifs de même ordre de l'inverse des données, ce qui permettrait de rééquilibrer chaque observable x^k par son inverse.

▷ **Remarque**

Dans le cadre de l'utilisation des moments successifs, on peut s'attendre à ce que des informations de grande taille ajustent de mieux en mieux les données. De telles informations reviennent en effet à rechercher la fonction exponentielle d'un polynôme de degré élevé qui ajuste une série de points. Il peut donc être intéressant d'observer le comportement d'informations mélant observables x^k et observables particuliers (ici, le ou les observables inverses) pour vérifier si cette association permet d'obtenir des modèles plus parcimonieux.

Nous allons à présent détailler la forme des densités des lois de maximum d'entropie obtenues à partir de chacune de ces familles.

2.2.1 Utilisation des moments empiriques successifs

• Utilisation des données non transformées

Le tableau III.1 résume la recherche de la loi de maximum d'entropie obtenue à partir de l'information suivante :

$$\overline{\mathcal{I}}_K = \left\{(x, \overline{x}), (x^2, \overline{x^2}), \ldots, (x^K, \overline{x^K})\right\}. \tag{III.1}$$

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
x^k	$\overline{x^k} = 1/2234 \sum_i x_i^k$	$\lambda_k,\ k \in \{1, \ldots, K\}$

État de Gibbs

$$f^*(x) = \frac{1}{Z} \exp\left(\sum_1^K \lambda_k x^k\right),\ x \in [1, 120]$$

Équations du PME

$$\frac{1}{Z} \int_{0.5}^{120.5} x^l \exp\left(\sum_1^K \lambda_k x^k\right) dx = \overline{x^l},\ l \in \{1, \ldots, K\}$$

Constante de normalisation

$$Z = \int_{0.5}^{120.5} \exp\left(\sum_1^K \lambda_k x^k\right) dx$$

TAB. III.1 – *Loi de maximum d'entropie obtenue à partir des moments successifs*

Il faut alors résoudre $K+1$ équations pour obtenir la constante de normalisation et les multiplicateurs $(\lambda_1, \ldots, \lambda_K)$ de la loi de maximum d'entropie.

Rappelons que la fréquence théorique de toute valeur discrète n comprise entre 1 et 120 est obtenue par l'intégration de la densité f^* sur l'intervalle $[n - 0.5, n + 0.5]$ entourant cette valeur.

• Utilisation des données transformées

Si on note Y la variable $\ln(X)$ et T la variable $1/X$, le tableau III.1, p. 165, résume aussi la recherche de la loi de maximum d'entropie pour les données transformées par logarithme ou inverse, en remplaçant x par $y = \ln(x)$ ou $t = 1/x$ dans les notations.

Les contraintes deviennent selon la transformation :

$$\begin{cases} \overline{t^k} &= 1/2234 \sum_i t_i^k &= 1/2234 \sum_i (1/x_i)^k \\ \overline{y^k} &= 1/2234 \sum_i y_i^k &= 1/2234 \sum_i (\ln(x_i))^k \ . \end{cases}$$

La densité de la loi de maximum d'entropie de T construite sur l'intervalle $[1/120.5, 1/0.5]$ ($\simeq [0.0083, 2]$) et celle de Y construite sur l'intervalle $[\ln(0.5), \ln(120.5)]$ ($\simeq [-0.69, 4.79]$), sont alors les suivantes :

$$\begin{cases} f_{inv}^*(t) = \dfrac{1}{Z_{inv}} \exp\left(\sum_1^K \lambda'_k t^k\right) & \text{, où } Z_{inv} = \displaystyle\int_{1/120.5}^{1/0.5} \exp\left(\sum_1^K \lambda'_k t^k\right) \ , \\ & \quad \text{avec } t \in [1/120.5, 1/0.5] \\ f_{ln}^*(y) = \dfrac{1}{Z_{ln}} \exp\left(\sum_1^K \lambda''_k y^k\right) & \text{, où } Z_{ln} = \displaystyle\int_{\ln(0.5)}^{\ln(120.5)} \exp\left(\sum_1^K \lambda''_k y^k\right) \ , \\ & \quad \text{avec } y \in [\ln(0.5), \ln(120.5)] \ . \end{cases}$$

La fréquence théorique p_n^* de toute valeur discrète n de 1 à 120 est alors obtenue par l'intégration de ces densités selon les formules :

$$p_n^* = \int_{1/(n+0.5)}^{1/(n-0.5)} f_{inv}^*(t)dt \qquad \text{ou} \qquad \int_{\ln(n-0.5)}^{\ln(n+0.5)} f_{ln}^*(t)dt \ .$$

2.2.2 Utilisation des observables "inverses"

Pour une information basée sur les K premiers moments et sur les M premiers moments de l'inverse des durées de séjour, la recherche de la loi de maximum d'entropie peut se résumer par le tableau III.2, p. 167. On note $\overline{\mathcal{J}}_{K+M}$ l'information correspondante :

$$\overline{\mathcal{J}}_{K+M} = \left\{(x, \overline{x}), \ldots, (x^K, \overline{x^K}), (1/x, \overline{1/x}), \ldots, (1/x^M, \overline{1/x^M})\right\} . \tag{III.2}$$

Il faut alors résoudre $K + M + 1$ équations pour obtenir la constante de normalisation, les multiplicateurs $(\lambda_1, \ldots, \lambda_K)$ et les multiplicateurs $(\mu_1, \ldots, \mu_M)$ de la loi de maximum d'entropie.

Nous considérerons deux types de familles :

- les familles de taille $K + 1$ composées des K moments successifs, associés à l'unique observable "inverse" $1/x$;
- les familles de taille $2K$ composées des observables successifs $(x^1, \ldots, x^K)$, associés aux observables inverses de mêmes rangs $(1/x^1, \ldots, 1/x^K)$.

Observables	**Niveaux de contraintes empiriques**	**Multiplicateurs**
x^k	$\overline{x^k}$	$\lambda_k,\ k \in \{1,\ldots,K\}$
$1/x^m$	$\overline{1/x^m} = 1/2234 \sum_i (1/x_i^m)$	$\mu_m,\ m \in \{1,\ldots,M\}$

État de Gibbs

$$f^*(x) = \frac{1}{Z} \exp\left(\sum_1^K \lambda_k x^k + \sum_1^M \mu_m / x^m\right)$$

Équations du PME

$$\frac{1}{Z}\int_{0.5}^{120.5} x^l \exp\left(\sum_1^K \lambda_k x^k + \sum_1^M \mu_m / x^m\right) dx = \overline{x^l},\ l \in \{1,\ldots,K\}$$
$$\frac{1}{Z}\int_{0.5}^{120.5} 1/x^l \exp\left(\sum_1^K \lambda_k x^k + \sum_1^M \mu_m / x^m\right) dx = \overline{1/x^l},\ l \in \{1,\ldots,K\}$$

Constante de normalisation

$$Z = \int_{0.5}^{120.5} \exp\left(\sum_1^K \lambda_k x^k + \sum_1^M \mu_m / x^m\right) dx$$

TAB. III.2 – *Loi de maximum d'entropie obtenue à partir des moments successifs et des observables inverses*

2.3 Adéquation et stabilité des modèles

Le groupement des observations en classes suggéré par les médecins nous permettra de tester l'adéquation du modèle aux données par un test du Khi-2 au moment de l'application de la méthode du *half-sampling*. Mais nous pouvons avant cela calculer pour chaque modèle la statistique de test suivante :

$$\mathcal{T} = \sum_{g=1}^{G} \left(\frac{(N_g - N_g^*)^2}{N_g^*}\right), \tag{III.3}$$

où N_g représente l'effectif empirique de la classe g et N_g^* l'effectif de cette même classe estimé par le modèle de densité f^*, $g \in \{1,\ldots,G\}$ (G désigne le nombre d'intervalles utilisés). C'est l'évolution de cette statistique que nous étudierons dans le cadre de chaque type de famille d'observables pour une taille d'information K croissante.

Pour détailler le comportement de certains modèles nous pourrons de plus calculer un intervalle de la *p-value* correspondant à cette statistique, à l'aide de deux lois du Khi-2 (cf. [Saporta 90, p. 335]) : un Khi-2 à G-1 degré de liberté et un Khi-2 à $G-1-K$ ddl, qui serait la loi à employer si le PME avait été appliqué à la variable discrétisée sur ces G intervalles. L'utilisation des niveaux de contraintes empiriques assure en effet l'équivalence entre résolution des équations du PME et application de la méthode du maximum de vraisemblance à une loi de même structure que f^*.

Nous choisirons alors le modèle le plus adéquat (s'il y en a un) au vu de l'évolution de cette statistique de test.

La dernière classe qui regroupe les individus avec des durées de séjour de 111 à 120 jours, ne compte que 7 observations. Lors de l'emploi du *half-sampling*, il y a donc un risque d'obtenir des effectifs de moins de 5 individus pour cette classe. Pour éviter ce problème, cette classe et l'avant-dernière seront rassemblées (18 individus restent entre 101 et 120 jours). Nous travaillerons donc avec $G = 15$ intervalles pré-définis.

2.4 Équations du PME et dualité

L'application présentée dans ce chapitre met en évidence le problème de la recherche d'outils algorithmiques adaptés à la résolution des équations du PME. Les multiplicateurs des modèles que nous avons obtenus ont été déterminés par résolution directe des équations du PME. Mais il semblerait que du point de vue algorithmique, et pour minimiser les risques de problèmes numériques, il soit plus judicieux d'explorer plus avant l'utilisation de méthodes développées dans le cadre des problèmes d'optimisation.

Dans ce travail, nous avons eu la démarche suivante :

- le problème initial considéré est le problème (P) de maximisation sous contrainte

$$(P) \; : \; \max\{S(P)\, ,\; P \in \mathcal{M}_{\mathcal{I}}\}\, ,$$

  qui peut s'écrire plus classiquement sous la forme d'un "programme convexe" (P') de dimension infini :

$$(P') \; : \; \min\{-S(P)\, ,\; P \in \mathcal{M}_{\mathcal{I}}\}\, ;$$

- nous avons montré que s'il existe un état de Gibbs dans $\mathcal{M}_{\mathcal{I}}$, alors celui-ci est l'unique solution de (P'). D'un point de vue pratique, cela ramène la résolution de (P) à celle des équations du PME :

$$E_{P^*}(\varphi_m) = a_m\, ,\; m \in \{0, \ldots, M\}\, , \tag{III.4}$$

  où P^* est un état de Gibbs dont la densité est exprimée sous la forme :

$$f^*(\boldsymbol{x}) = \exp\left(-\psi(\lambda) + \sum_m \lambda_m a_m\right)\, .$$

Considérons à présent la fonction D sur $\mathbf{R}^{M+1}$ de la forme suivante :

$$D(\lambda) = -\psi(\lambda) + \sum_m \lambda_m a_m \ ,$$

où $\psi(\lambda) = \int_{\boldsymbol{x}} \exp(-\sum_m \lambda_m \varphi_m(\boldsymbol{x}))d\boldsymbol{x}$.

Soit Λ le domaine de définition convexe de D. On a alors :

$$\left\{ \begin{array}{lclcl} \dfrac{\partial D}{\partial \lambda_m} & = & a_m - \dfrac{\partial \psi}{\partial \lambda_m} & = & a_m - E_{P^*}(\varphi_m) \\ \dfrac{\partial^2 D}{\partial \lambda_m \partial \lambda_l} & = & \mathrm{cov}_{P^*}(\varphi_m, \varphi_l) & & \end{array} \right.$$

Les équations III.4 sont donc celles de l'unique solution, si elle existe, du problème d'optimisation suivant dans Λ :

$$(P'') \ : \ \max\{D(\lambda) \ , \ \lambda \in \Lambda\} \ .$$

Ce problème est alors sans contrainte et de dimension finie.

Ce dernier problème (P'') est en fait le problème dual, au sens de la théorie de la dualité, de (P) ([Ben-Tal & Charnes 79, Ben-Tal *et al.* 88, Lemaréchal 93, Rockafellar 70]).

Les algorithmes permettant d'obtenir les valeurs des multiplicateurs d'une loi de maximum d'entropie, à partir de niveaux de contraintes de nature quelconque, sont donc à rechercher parmi les algorithmes construits dans le domaine de l'optimisation et de l'analyse convexe ([Decarreau *et al.*, Iusem & Teboulle 92, Zhu 92]).

Remarquons de plus que dans le cas où les niveaux $(a_m)_{m \in \{1,\dots,M\}}$ sont les moyennes empiriques des observables, on peut montrer que la fonction D est à une constante près (l'inverse de la taille de l'échantillon), la log-vraisemblance des paramètres d'une loi de même structure que P^*.

3 Résultats : modélisation à partir de l'échantillon complet

Nous allons à présent détailler les résultats obtenus à partir des 2234 données, en présentant les graphes des densités des lois obtenues, puis l'évolution des multiplicateurs, et enfin les résultats des tests précédemment décrits.

3.1 Densité des lois de maximum d'entropie obtenues

Les figures III.2, p. 172, à III.6, p. 176, présentent l'évolution des densités des lois de maximum d'entropie obtenues pour des informations de tailles croissantes. Elles permettent une première analyse visuelle de l'amélioration de l'ajustement des densités de maximum d'entropie à la distribution empirique.

• **Utilisation des moments successifs** (figure III.2, p. 172)

Ce n'est qu'à partir de 8 moments (c'est-à-dire pour un modèle à 8 paramètres plus une constante de normalisation) que les lois de maximum d'entropie commencent à retrouver le comportement particulier du début de la distribution. La queue de distribution semble être de mieux en mieux ajustée à partir de 7 moments.

• **Utilisation de l'observable inverse** (figure III.3, p. 173)

Le "décrochement" du début de la distribution apparaît à présent dès l'utilisation de 5 observables, c'est-à-dire pour une information mettant en jeu 4 moments et l'observable $1/x$. Ce phénomène s'accentue quand on augmente le nombre de moments associés à l'observable inverse. L'observable inverse semble donc permettre de retrouver plus rapidement les particularités de la distribution. Il nécessite l'utilisation de 4 moments au lieu des 8 précédents (soit 5 multiplicateurs au lieu de 8 précédemment).

• **Utilisation des moments successifs et des moments inverses successifs** (figure III.4, p. 174)

La forme de la distribution est retrouvée ici pour l'information de taille 6 (soit 6 multiplicateurs plus une constante de normalisation). Par contre, le début de la densité obtenue présente une décroissance beaucoup plus importante que dans les cas précédents, ce qui pourrait éventuellement perturber l'estimation des fréquences du premier intervalle prédéfini (durées de 1 ou 2 jours).

• **Utilisation des moments successifs sur les inverses des données** (figure III.5, p. 175)

L'utilisation d'une modélisation sur un espace continu pose problème pour cette variable transformée, comme le montre la forme de la distribution empirique. On réussit néammoins à partir de 6 observables à retrouver dans le modèle le pic correspondant aux individus restant un jour. Mais la recherche des solutions des équations du PME prend beaucoup plus de temps que pour les autres familles dès que la taille de l'information dépasse 4 et aucun essai n'a permis d'obtenir un modèle de maximum d'entropie avec 8 observables pour des problèmes d'instabilité numérique. Devant une telle distribution empirique, il serait préférable soit de ne travailler que sur la première partie des données dans $]0, 0.5]$ en laissant de coté la queue de distribution qui est beaucoup trop écartée des autres valeurs, soit de revenir à une étude sur un espace discret.

• **Utilisation des moments successifs sur les logarithmes des données** (figure III.6, p. 176)

On peut considérer que la forme de la distribution empirique commence à apparaître dès l'utilisation de 3 moments successifs (courbe décentrée vers la droite, sans symétrie). À partir de 6 moments, la courbe est très proche de l'histogramme.

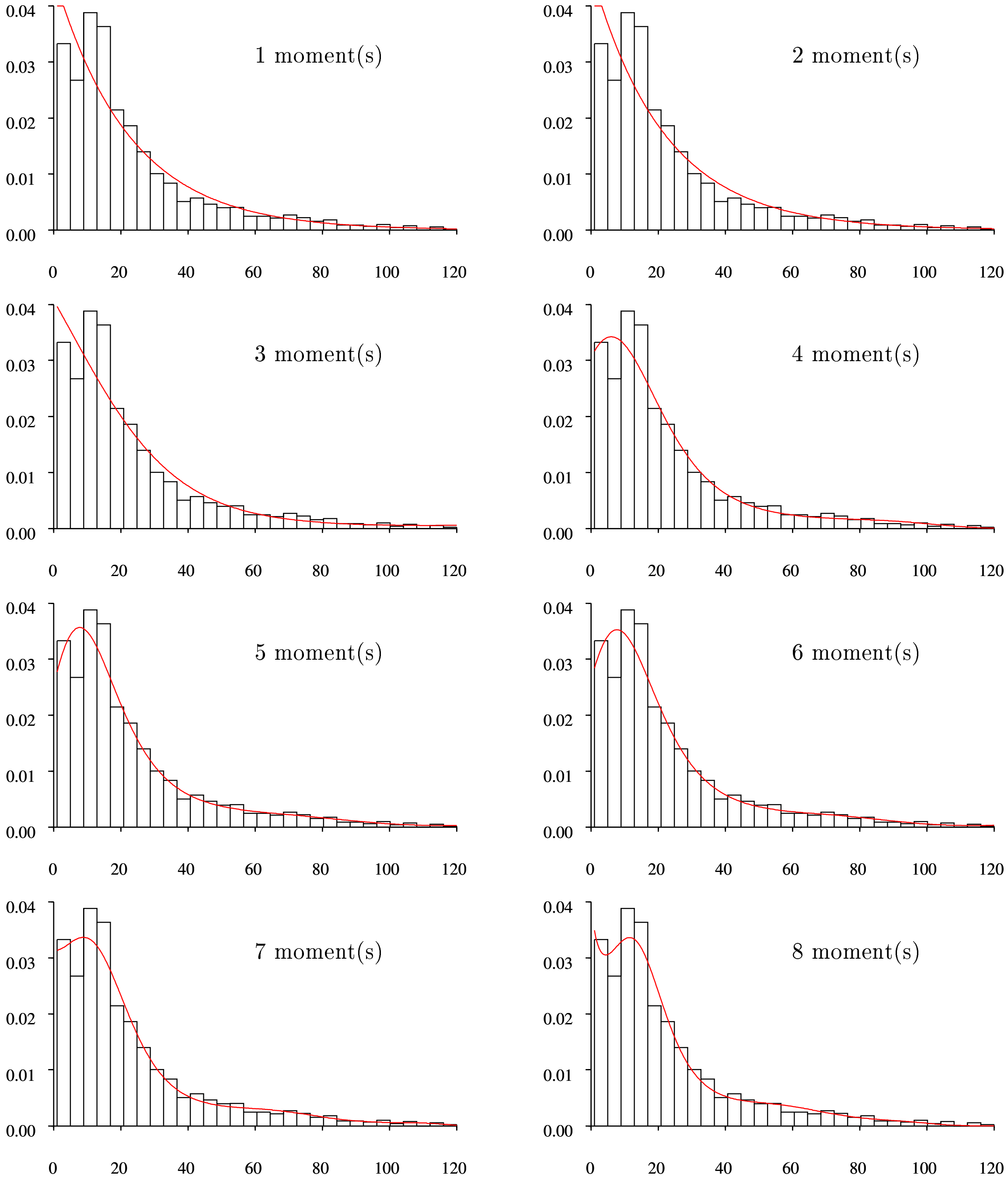


FIG. III.2 – *Lois de maximum d'entropie des durées de séjour construites sur* $[0.5, 120.5]$ *à partir des moments*

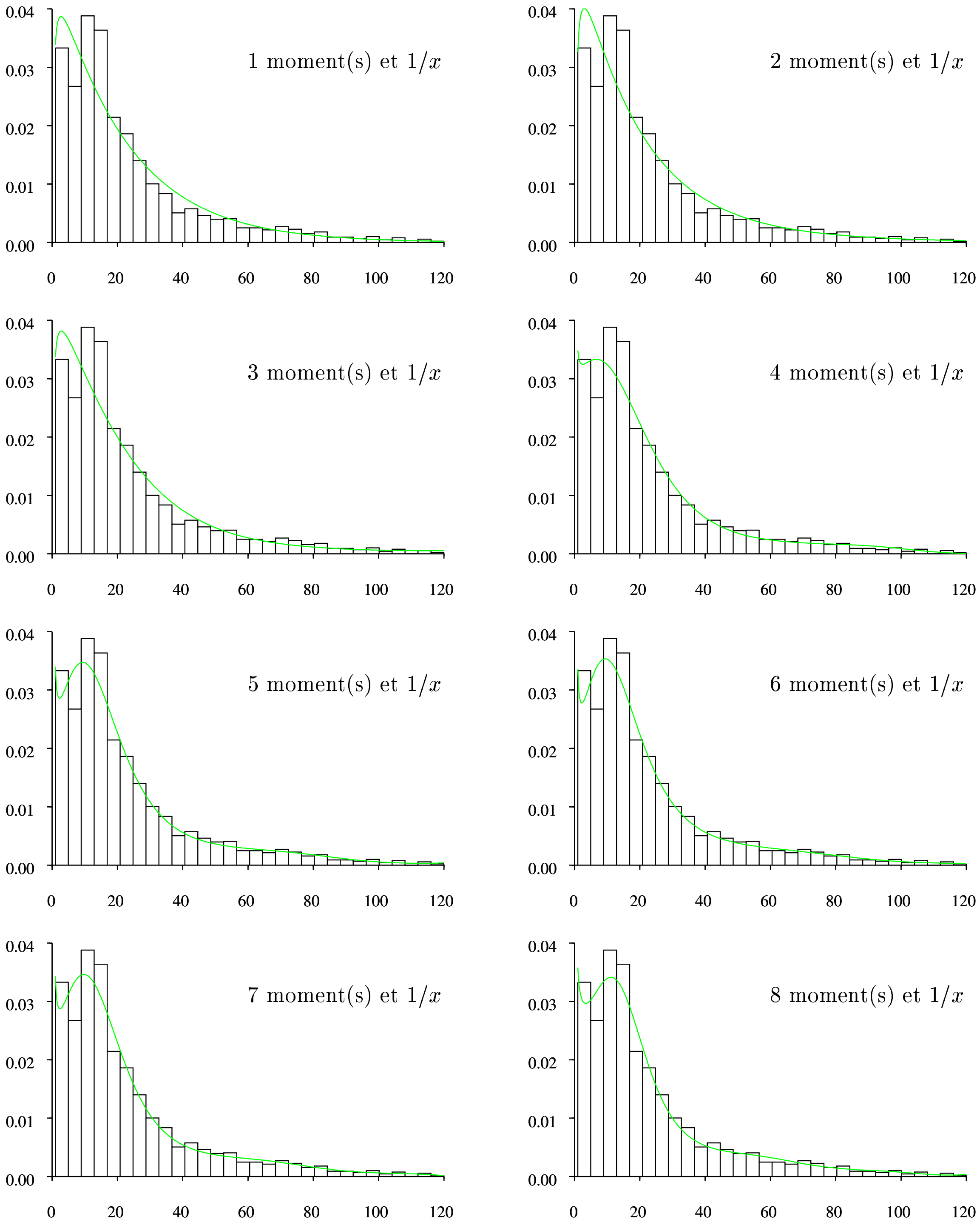


FIG. III.3 – *Lois de maximum d'entropie des durées de séjour construites sur* $[0.5, 120.5]$ *à partir des moments et de l'observable* $1/x$

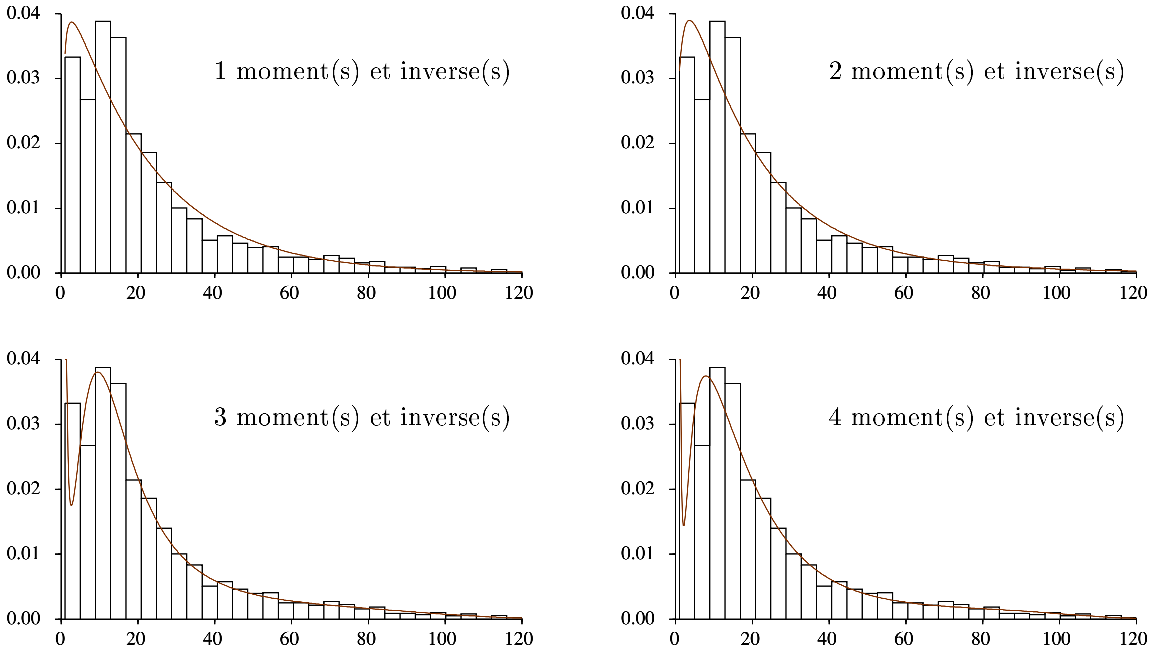


FIG. III.4 – *Lois de maximum d'entropie des durées de séjour construites sur* $[0.5, 120.5]$ *à partir des moments et des moments inverses*

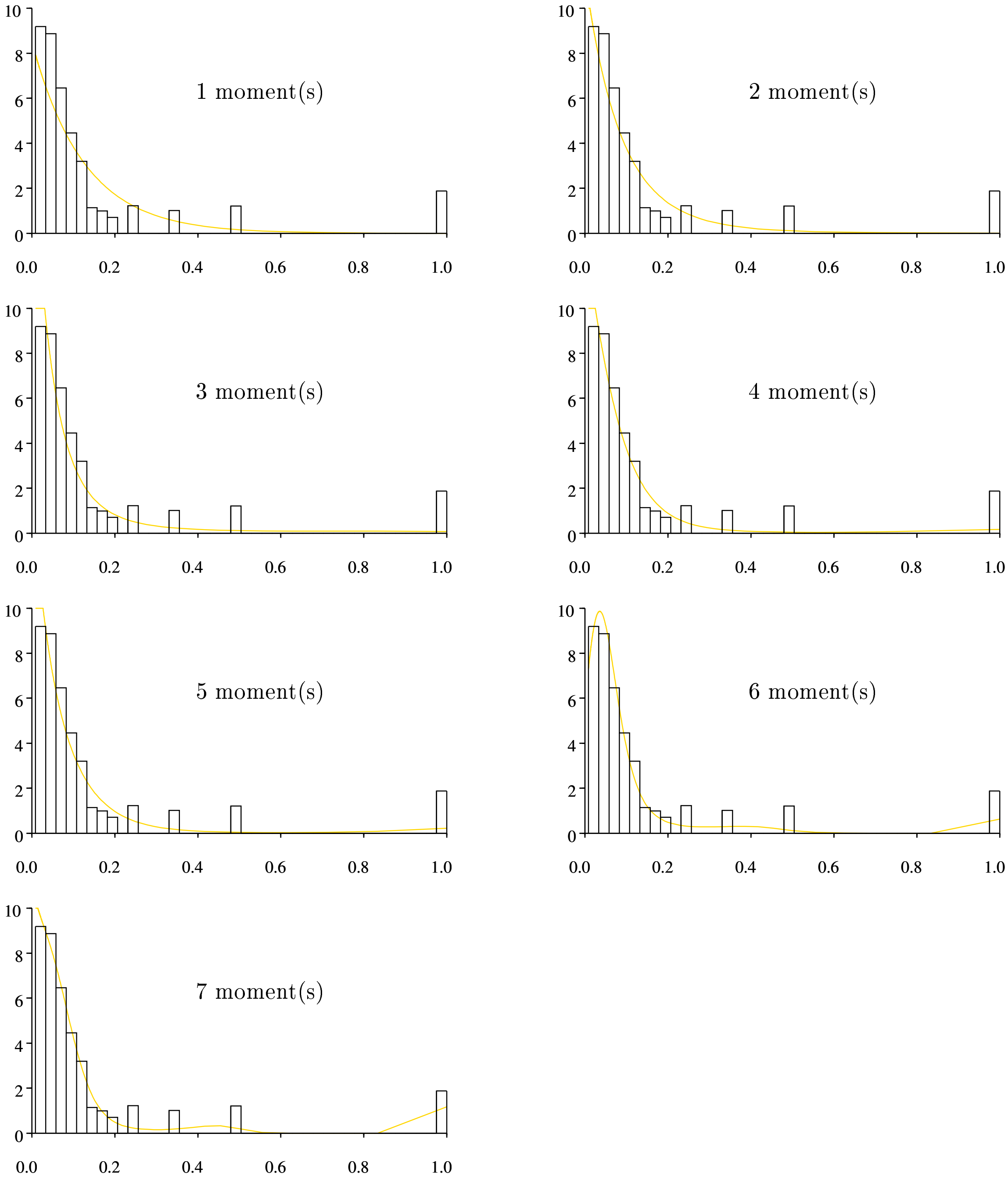


FIG. III.5 – *Lois de maximum d'entropie des inverses des durées de séjour de moins de 120 jours*

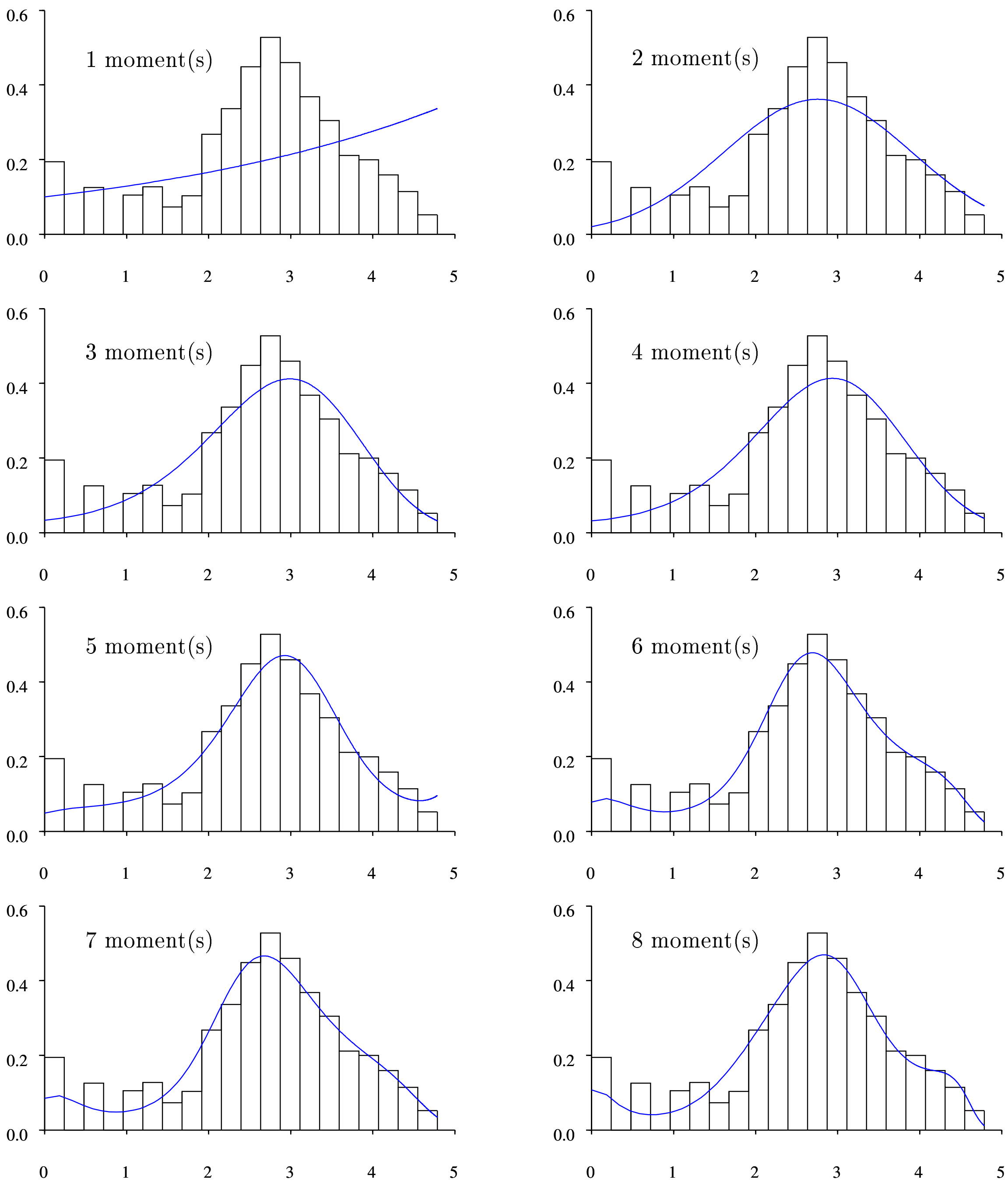


FIG. III.6 – *Lois de maximum d'entropie des logarithmes des durées de séjour de moins de 120 jours*

3.1.1 Évolution des multiplicateurs

Avant d'étudier les résultats des tests, nous présentons dans les tableaux suivants les valeurs des multiplicateurs des différents modèles obtenus pour chaque type de famille d'observables.

• Utilisation des moments successifs

Pour résoudre des problèmes de convergence de l'algorithme, les modèles ont en fait été construits pour la variable $X' = X/100$, ce qui permet de travailler sur un intervalle plus réduit et des niveaux de contraintes moins élevés. Le tableau III.3 présente les valeurs des multiplicateurs des lois de maximum d'entropie obtenues pour cette variable X' en fonction de la taille K de l'information.

	Multiplicateurs des observables							
K	x	x^2	x^3	x^4	x^5	x^6	x^7	x^8
1	-4.44							
2	-4.58	0.19						
3	-2.54	-5.78	4.14					
4	4.06	-38.84	57.46	-26.23				
5	9.66	-80.25	169.91	-150.23	47.22			
6	8.21	-65.72	112.48	-46.73	-38.22	26.2		
7	-0.62	50.04	-512.16	1591.05	-2244.32	1492.21	-380.15	
8	-14.64	282.12	-2136.04	7336.16	-13350	13370	-6969	1477

TAB. III.3 – *Multiplicateurs des lois de maximum d'entropie de la variable X' obtenues à partir des moments successifs*

Aucun comportement régulier n'apparaît donc pour ces multiplicateurs. Lorsque l'on reprend les multiplicateurs des lois de la variable d'origine X, il apparaît de plus qu'à partir d'une certaine taille d'information ($K \geq 4$), de nombreux multiplicateurs de la loi de la variable initiale X présentent des valeurs très petites (inférieures à 10^{-3}) relativement aux multiplicateurs des observables correspondant aux moments de rang inférieur à K .

• Utilisation de l'observable inverse

Pour ces familles d'observables, les modèles ont aussi été construits pour la variable $X' = X/100$. Le tableau III.4 présente les valeurs des multiplicateurs en fonction de la taille de l'information.

Ce sont ici les multiplicateurs des observables de rang supérieur à 3 qui deviennent très faibles, à partir de l'information de taille 4 (3 moments successifs et l'observable inverse). Par rapport aux multiplicateurs des observables de type x^k, $k \in \{1, \ldots, K\}$, les multiplicateurs de l'observable inverse présentent par contre des valeurs beaucoup plus élevées. Mais il n'apparaît pas non plus pour cette information de comportement régulier.

	Multiplicateurs des observables							
K	x	x^2	x^3	x^4	x^5	x^6	x^7	$1/x$
2	-4.61							-0.0034
3	-5.33	0.85						-0.0047
4	-3.59	-3.7	2.99					-0.003
5	5.53	-43.92	64.08	-29.09				0.0022
6	16.19	-112.68	238.27	-213.49	68.37			0.006
7	19.76	-143.53	351.15	-407.89	224.18	-46.79		0.0069
8	12.97	-68.38	-20.1	515.52	-977.27	732.39	-198.45	0.0056

TAB. III.4 – *Multiplicateurs des lois de maximum d'entropie de la variable X' obtenues à partir des moments successifs et de l'observable inverse*

• Utilisation des moments successifs et des moments inverses successifs

Le tableau III.5, p. 178, présente les valeurs (multipliées par 10^3) des multiplicateurs pour les modèles obtenus cette fois-ci directement pour la variable X.

	Multiplicateurs ($\lambda_k.10^3$) des observables							
K	x	x^2	x^3	x^4	$1/x$	$1/x^2$	$1/x^3$	$1/x^4$
2	-46.07				-338.28			
4	-56.42	0.11			-901.7	306		
6	-202.46	2.42	-0.01		-21593.98	36684	-17400	
8	-9.25	-2.11	0.03	-0	5776.76	-45407.96	78927.2	-37859

TAB. III.5 – *Multiplicateurs ($\lambda_k * 10^3$) des lois de maximum d'entropie obtenues à partir des moments successifs et des moments inverses successifs*

La remarque faite pour les multiplicateurs de l'observable $1/x$ pour les familles précédentes se retrouve ici pour tous les observables correspondant à des moments inverses : les multiplicateurs associés aux observables inverses successifs présentent des valeurs bien plus élevées que celles des observables x^k successifs.

• Utilisation des moments successifs sur les inverses des données

Le tableau III.6 présente les valeurs des multiplicateurs pour les modèles obtenus cette fois-ci directement pour la variable $T = 1/X$ (il n'a pas été nécessaire ici de passer par la modélisation de $T/100$).

Malgré la transformation des données, il n'y a toujours pas ici d'évolution régulière des multiplicateurs quand la taille de l'information augmente.

	Multiplicateurs des observables						
K	t	t^2	t^3	t^4	t^5	t^6	t^7
1	-7.95						
2	-11.56	4.43					
3	-20.15	25.67	-10.91				
4	-104.03	-2836.45	72511.82	-380884.29			
5	-16.51	25.4	-83.7	137.21	-66.63		
6	33.42	-611.45	2906.55	-5998.86	5541.91	-1873.78	
7	-8.94	140.84	-2438.92	11882.86	-24351.17	22141.25	-7368.16

TAB. III.6 – *Multiplicateurs des lois de maximum d'entropie des inverses des durées de séjour, obtenues à partir des moments successifs*

• Utilisation des moments successifs sur les logarithmes des données

Les valeurs des multiplicateurs des lois de maximum d'entropie de la variable X sont directement présentés dans le tableau III.7.

	Multiplicateurs des observables							
K	y	y^2	y^3	y^4	y^5	y^6	y^7	y^8
1	0.25							
2	2.09	-0.38						
3	0.73	0.35	-0.11					
4	0.55	0.61	-0.21	0.01				
5	1.11	-1.7	1.5	-0.45	0.04			
6	1.67	-7.15	7.94	-3.44	0.66	-0.05		
7	1.69	-8.71	10.71	-5.38	1.31	-0.15	0.01	
8	1.22	-14.61	26.77	-21.56	9.45	-2.35	0.31	-0.02

TAB. III.7 – *Multiplicateurs des lois de maximum d'entropie des logarithmes des durées de séjour, obtenues à partir des moments successifs*

Il n'y a pas pour ces informations de multiplicateurs aux valeurs négligeables (bien que les multiplicateurs des observables de plus grand rang soient relativement plus faible que les autres). De plus, les valeurs absolues des multiplicateurs de chaque observable y^k augmentent assez régulièrement avec la taille de l'information à partir du moment où ils interviennent dans les contraintes.

3.1.2 Estimation des moments n'intervenant pas dans les contraintes

Toutes les informations que nous avons testé dans cette étude utilisent au moins les moments successifs de la variable modélisée. Aussi, il peut être intéressant d'étudier les différences entre les moments empiriques n'intervenant pas dans les contraintes et leurs valeurs estimées grâce au modèle.

- Pour une famille utilisant les K premiers observables x^k, on calculera les moments d'ordre $K+1$ et $K+2$. On procédera de même pour les modèles de l'inverse ou du logarithme des durées de séjour.

- Pour une famille utilisant ces K observables et l'observable $1/x$, on calculera les moments d'ordre $K+1$ et $K+2$ et les moments de $1/X^2$ et $1/X^3$.

- Pour une famille utilisant ces K observables successifs et les K premiers observables inverses $1/x^k$, on calculera les moments de rang $K+1$ et $K+2$ ainsi que les moments de $1/X^{K+1}$ et $1/X^{K+2}$.

Pour un observable φ, on mesure l'écart entre son moment empirique $\overline{\varphi}$ et le moment estimé $E^*(\varphi)$ par l'écart relatif $\left(\overline{\varphi} - E^*(\varphi)\right)/\overline{\varphi}$. Les figures III.7, p. 181, et III.8, p. 182, présentent l'évolution de ces écarts relatifs quand on augmente la taille de l'information.

Il faut noter que les moments de rang supérieur à 3 ou 4 selon les types d'informations, présentent des valeurs très élevées. L'écart relatif utilisé est sensible aux variations du dénominateur. Cela explique qu'à partir d'un certain rang les écarts relatifs soient tous quasiment nuls bien que l'écart brut entre valeurs empiriques et estimées soit très élevé. De même dans le cas de l'estimation des moments inverses. On obtient des écarts relatifs très forts en valeur absolue alors que l'écart brut est faible, car les moments inverses présentent des valeurs très petites.

Ces graphes illustrent néammoins la stabilisation des écarts relatifs à partir d'une certaine taille d'information, hormis pour l'estimation du moment de $1/X^2$ quand on utilise l'information basée sur les moments successifs et sur l'observable $1/x$. C'est par contre le cas à partir de 5 observables pour le premier type d'information (moments successifs). Mais les écarts bruts sont en fait très grands (ils sont de l'ordre de 10^7 pour $K=5$ en estimant les moments d'ordre $K+1$).

Cette stabilisation apparaît pour 6 observables dans le cas des familles basées sur les moments successifs et le moment inverse, mais les écarts sont encore de l'ordre de 10^8 pour les moments. Les écarts sont par contre très petits pour l'estimation des moments inverses d'ordre 2 et 3.

Les écarts bruts obtenus à partir des modèles utilisant moments successifs et moments inverses successifs sont également grands, mais néammoins plus faibles que ceux provenant de la famille d'observables mélant les moments successifs et le moment inverse.

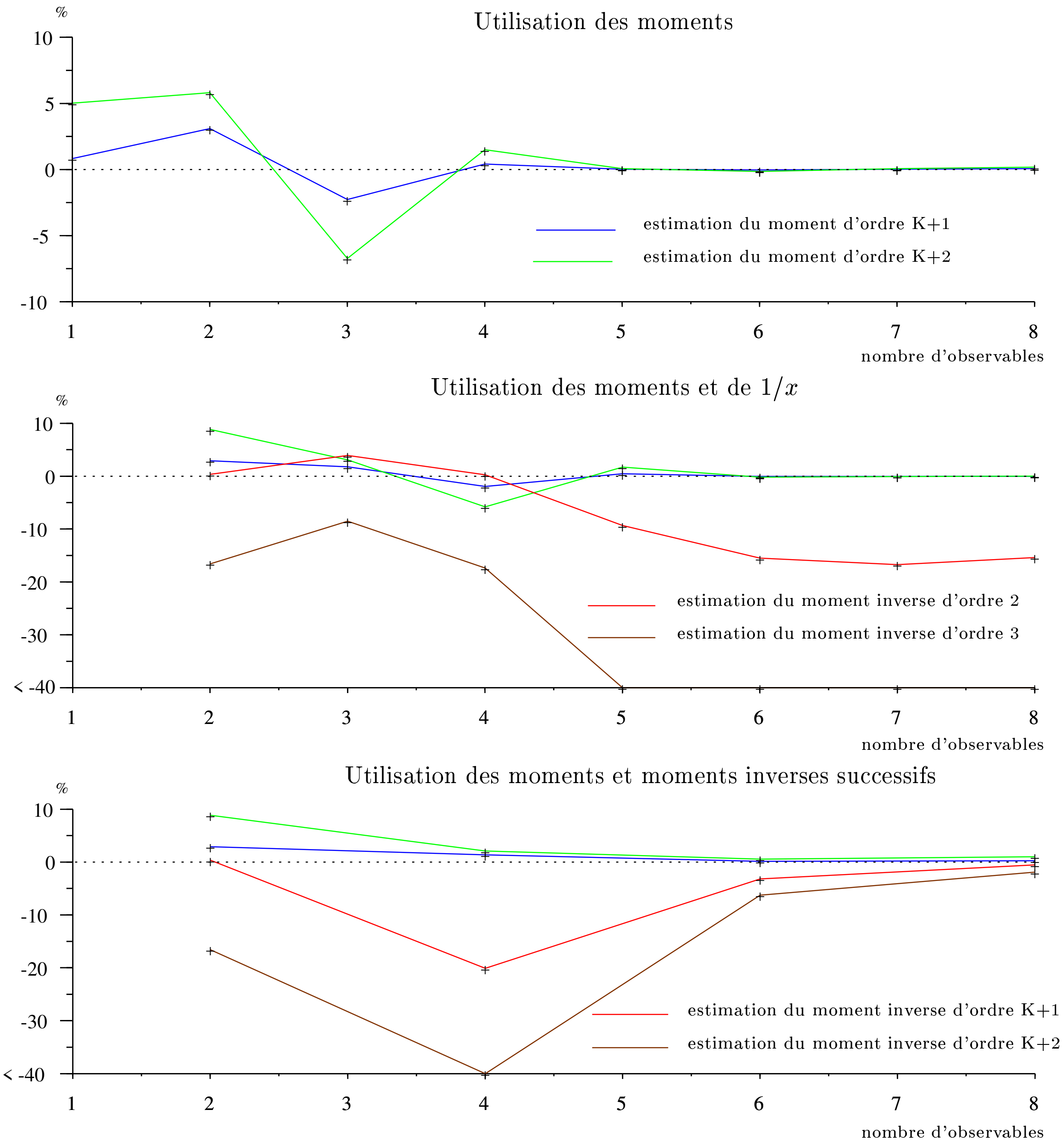


FIG. III.7 – *Pour chaque loi basée sur K observables pour les données non transformées : pourcentage d'écart entre les estimations des moments et moments inverses d'ordre $K+1$, $K+2$ et les moments empiriques correspondant*

Les écarts bruts obtenus à partir des modèles construits sur les logarithmes des données sont en fait les moins élevés (ils sont tous inférieurs à 10), mais les moments des logarithmes dépasse 300 à partir de l'ordre 6.

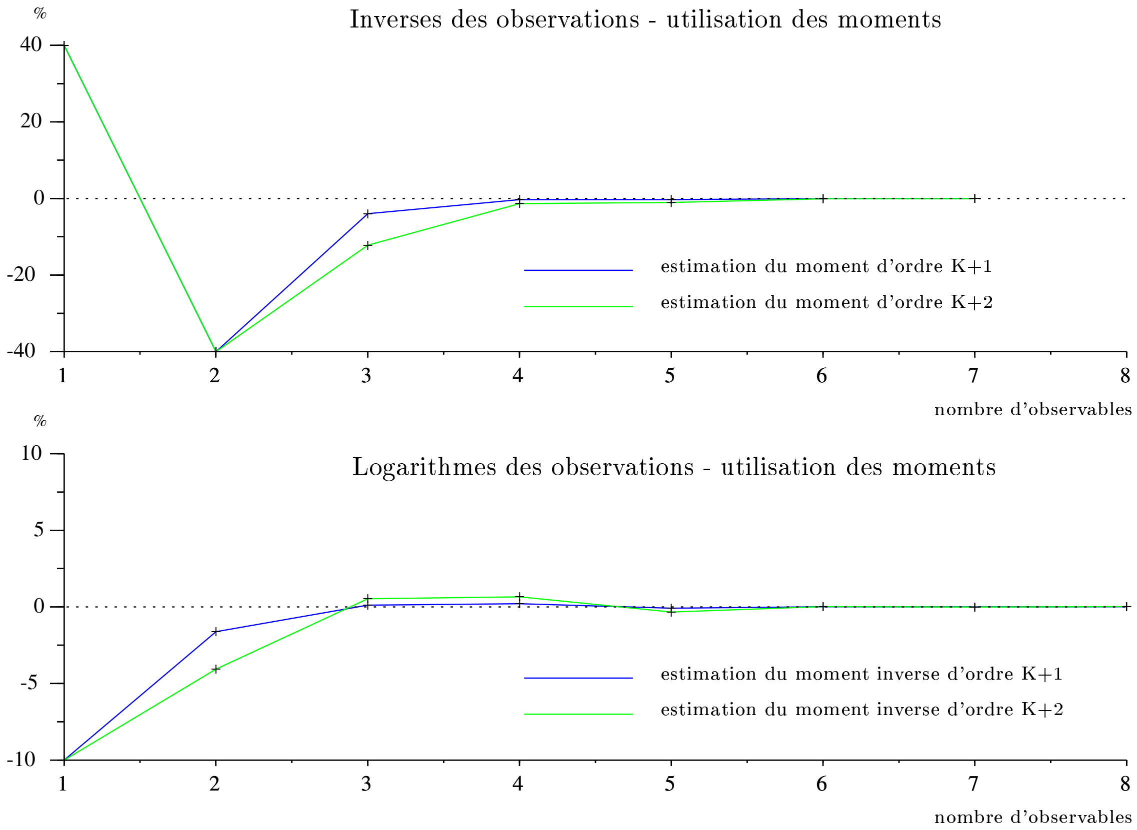


FIG. III.8 – *Pour chaque loi basée sur* K *observables pour les données transformées : pourcentage d'écart entre les estimations des moments et moments inverses d'ordre* $K+1$, $K+2$ *et les moments empiriques correspondant*

Nous allons à présent utiliser la statistique $\mathcal{T}$ (III.3, p. 167) mesurant l'adéquation aux fréquences des intervalles prédéfinis pour pouvoir évaluer la qualité de ces différents modèles.

3.2 Tests d'adéquation

La statistique de test $\mathcal{T}$ III.3 nous permet d'évaluer la différence entre les fréquences empiriques des 15 intervalles prédéfinis et les fréquences de ces mêmes intervalles estimées par les différents modèles.

Le tableau III.8, p. 183, présente les valeurs de la statistique de test selon le type de famille d'observables utilisée et sa taille.

Mis à part pour les modèles construits sur l'inverse de la variable X, les valeurs obtenues possèdent toutes la même particularité, au sein d'une même famille d'observables, de partir de valeurs élevées qui vont diminuer jusqu'à atteindre un palier.

Nombre d' observables	Famille d'observables utilisée				
	x^k	x^k et $1/x$	x^k et $1/x^k$	t^k ($T = 1/X$)	y^k ($Y = \ln(X)$)
1	135.27			487.40	1065.98
2	134.47	124.67	124.67	248.93	148.73
3	120.16	112.89		132.00	83.75
4	62.2	112.34	104.24	62.60	71.85
5	**48.84**	59.49		96.56	67.6
6	48.44	**30.23**	**13.24**	28.55	**11.78**
7	44.41	26.99		39.66	14.45
8	59.44	29.24	26.01		17.72

TAB. III.8 – *Valeurs des statistiques de tests*

Ce palier apparaît à partir de 5 observables pour la première famille constituée des moments successifs, et à partir de 6 observables pour les autres familles. Les résultats correspondant à la modélisation de l'inverse de la variable X sont par contre très instables. Pour 6 observables la statistique obtenue est meilleure que celle obtenue pour le même nombre d'observables avec 5 moments successifs et l'observable inverse, mais le comportement des statistiques ne suit pas avant et après cette taille le même comportement que pour les autres informations. Nous mettrons donc de coté cette modélisation dont l'évolution paraît déséquilibrée.

Mais même à partir de cette "stabilisation" des statistiques de test, la famille d'observables élémentaires constituée des moments successifs ne permet pas d'obtenir une bonne adéquation (les *p-values* sont quasi nulles pour ces informations, que le degré de liberté utilisé soit 14 ou $14 - K$).

Les valeurs obtenues à partir des moments successifs associés à l'observable inverse sont un peu meilleures (c'est-à-dire plus faibles) que les résultats obtenus avec la famille élémentaire. Le meilleur résultat dans le cadre de ces informations est obtenu pour 7 observables (mais la *p-value* est alors comprise selon notre encadrement entre 0 et 0.02). Les familles basées sur les moments successifs et les moments successifs inverses présentent des statistiques de tests encore inférieures aux précédentes, et permettent d'obtenir un bon résultat pour une information comprenant 6 observables.

Les meilleurs résultats, c'est-à-dire les Khi-2 les plus faibles, sont obtenus pour les modèles construits sur les logarithmes des données, à partir de leurs moments successifs. Le meilleur résultat au sein de ces familles est obtenu pour 6 observables.

Aucun modèle obtenu à partir des moments successifs ne peut donc être retenu si on utilise cette statistiques de test comme critère. Nous avons écarté l'utilisation des modèles construits sur l'inverse des données à cause de l'instabilité de la statistique $\mathcal{T}$. Nous retiendrons les trois autres familles et au sein de celles-ci les modèles correspondant au début du palier de stabilisation de ces statistiques (quelle que soit la fourchette des *p-values*). Les trois modèles retenus correspondent donc aux familles à 6 observables qui permettent d'obtenir des *p-values* comprises dans les fourchettes suivantes :

	Famille d'observables utilisée ($K = 6$)		
Khi-2	x^k et $1/x$	x^k et $1/x^k$	y^k, $Y = \ln(X)$
8 ddl	0.00	0.10	0.16
14 ddl	0.01	0.51	0.62

Les densités de ces trois modèles sont présentées dans la figure III.9, avec une échelle faisant apparaître les valeurs maximales prises par ces fonctions et après un changement de variable ramenant la densité f^*_{ln} à une fonction de x.

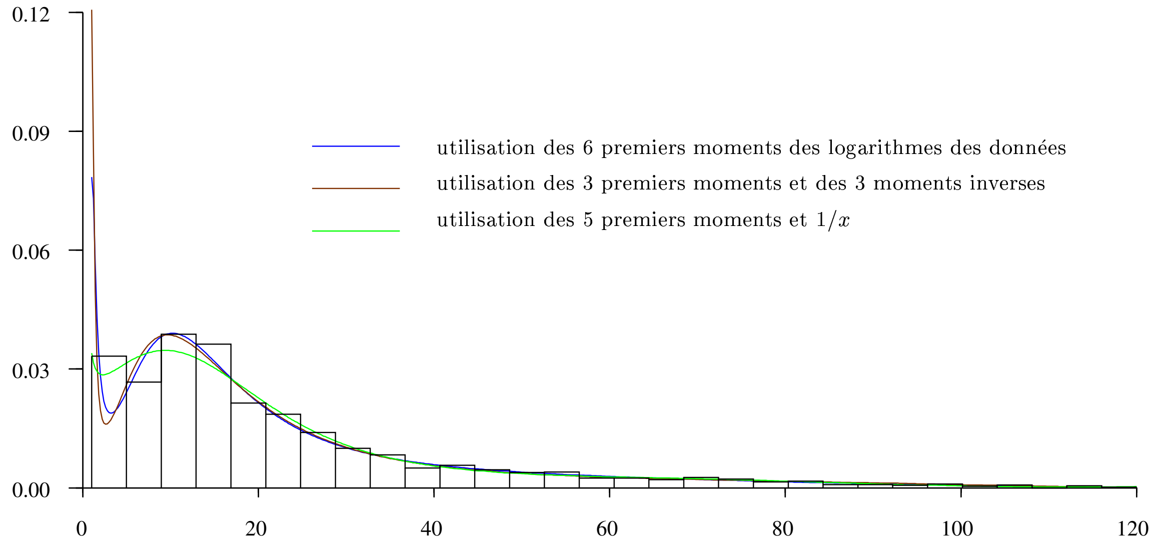


FIG. III.9 – *Comparaison des densités des trois modèles retenus*

Les différences entre ces trois modèles apparaissent dans le début de la forme des densités. Le modèle correspondant à l'utilisation des 5 premiers moments et de l'observable $1/x$ semble le moins bien approcher la forme de la distribution empirique. Les deux autres courbes sont très proches l'une de l'autre, mais l'utilisation des moments et moments inverses aboutit à une courbe qui présente une décroissance beaucoup plus forte et qui part de plus grandes valeurs que la densité construite sur les logarithmes des données.

La figure III.10, p. 185, permet de mieux visualiser le comportement des Khi-2. On y indique la valeur 24 qui est le quantile d'ordre 95 d'un Khi-2 à 14 ddl. Toute valeur inférieure à celle-ci correspond donc à une p-value supérieure à 0.05 pour un Khi-2 à 14 ddl.

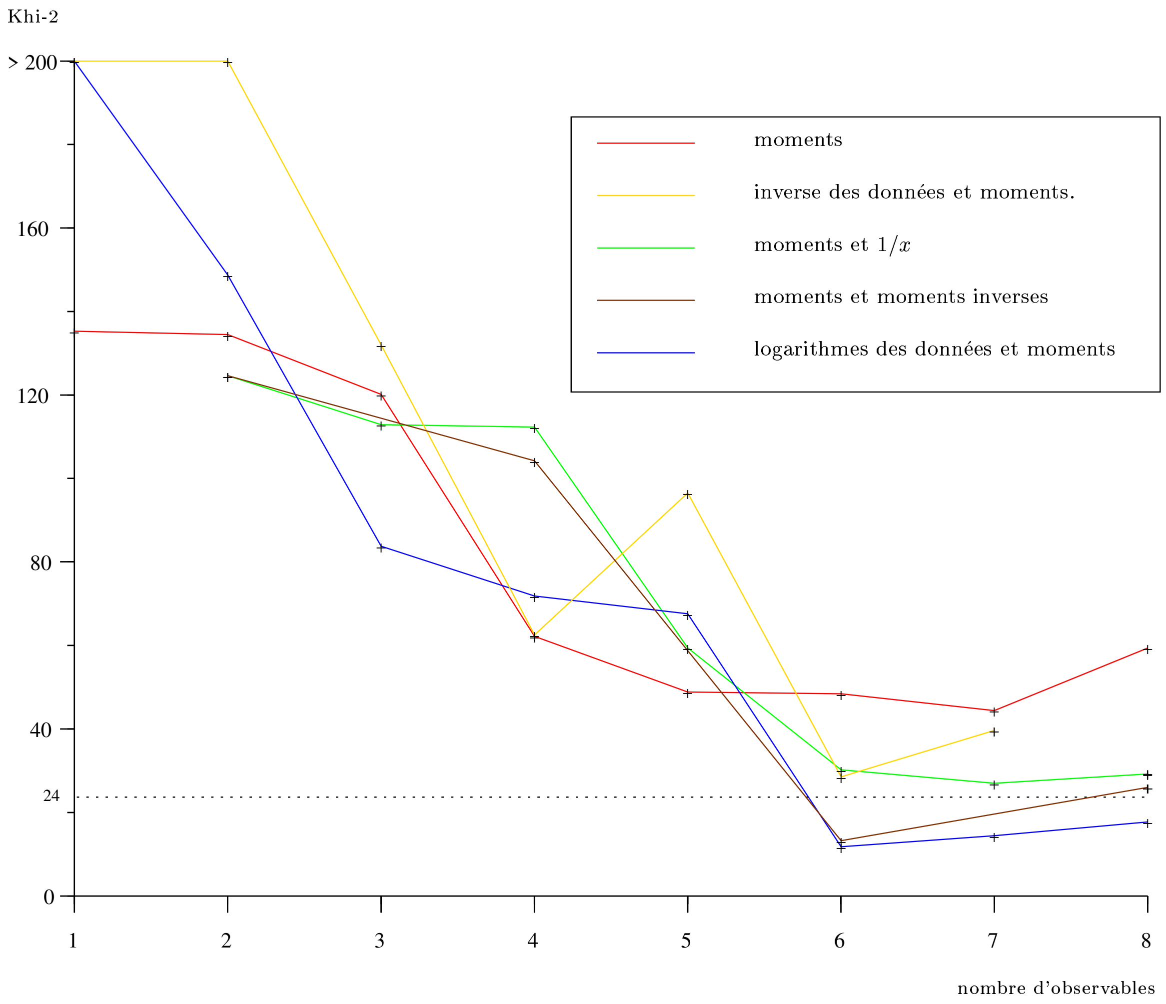


FIG. III.10 – *Évolution des statistiques des tests d'adéquation en fonction du type d'observables utilisés et de leur nombre*

▷ Remarque

Les fréquences empiriques et celles obtenues à partir des modèles correspondant aux différentes informations composées de 6 observables sont présentées dans le tableau III.9, p. 186.

Fréquences empiriques	**Famille d'observables utilisée** x^k	x^k et $1/x$	x^k et $1/x^k$	t^k	y^k
171	131	146	176	181	170
163	221	202	150	143	164
378	390	380	400	391	356
423	353	362	383	398	407
253	279	288	282	284	301
205	203	206	195	189	205
132	143	141	136	130	141
160	173	168	172	167	175
108	99	97	103	106	100
77	69	70	71	75	64
52	55	57	53	56	44
47	44	45	41	41	31
29	31	31	30	29	24
18	18	17	20	19	18
18	16	16	16	17	27

TAB. III.9 – *Fréquences empiriques et estimées des 15 intervalles prédéfinis*

4 Stabilité des modèles retenus

Pour tester la stabilité des meilleurs modèles obtenus (au sens du test d'adéquation utilisé), nous utiliserons la méthode du *half-sampling*. Nous noterons E1 et E2 les deux échantillons de taille 1117 tirés au hasard à partir des données.

Pour chaque modèle construit sur une moitié des observations, nous utilisons l'autre moitié pour construire les fréquences empiriques des intervalles prédéfinis et établir à nouveau un test d'adéquation. Dans ce cas, la statistique $\mathcal{T}$ suit approximativement un Khi-2 à 14 ddl puisque les sous-échantillon servant à l'estimation et au test sont indépendants. Pour chaque famille utilisée, on indiquera également en rappel la statistique et la *p-value* correspondante pour un Khi-2 à 14 ddl obtenues à partir de l'ensemble des données ($E1 + E2$).

• Pour le modèle obtenu à partir des 5 moments successifs et de l'observable $1/x$, on obtient les résultats suivants :

Échantillon d'estimation	E1	E2	E1+E2
Échantillon test	E2	E1	E1+E2
Khi2	79.79	23.08	30.23
***p-value* (14 ddl)**	0.00	0.06	0.01

• Pour le modèle obtenu à partir des 3 moments successifs et des 3 moments inverses successifs, on obtient les résultats suivants :

Échantillon d'estimation	E1	E2	E1+E2
Échantillon test	E2	E1	E1+E2
Khi2	18.87	16.35	13.24
***p-value* (14 ddl)**	0.17	0.29	0.51

• Pour le modèle obtenu à partir des logarithmes des données et de leurs 6 moments successifs, on obtient les résultats suivants :

Échantillon d'estimation	E1	E2	E1+E2
Échantillon test	E2	E1	E1+E2
Khi2	15.91	18.80	11.78
***p-value* (14 ddl)**	0.32	0.17	0.62

On constate donc une très grande stabilité, au sens du test du Khi-2 que nous avons utilisé, du modèle construit à partir des logarithmes des observations. Par contre, le modèle obtenu à partir de la famille utilisant l'observable inverse est à rejeter. Bien que la *p-value* reste supérieure à 0.05 dans le cas de la construction du modèle à partir de E1, il y a une nette différence avec les résultats globaux et surtout avec les résultats utilisant l'autre échantillon comme base d'estimation. Cette instabilité n'est pas étonnante étant donné la croissance importante des variances des X^k pour k croissant. Ainsi, le moment empirique d'ordre 6 calculé sur l'échantillon E1 vaut 3.9×10^{10}, et celui sur E2 vaut 2.62×10^{10}, soit une différence de 1.3×10^{10}. En comparaison, le moment d'ordre 6 des logarithmes des données calculé sur E1 vaut 1282.4, et vaut 1232.18 quand on utilise E2, soit une différence de 50.22.

Par contre, l'ajout des observables inverses successifs permet de retrouver une meilleure stabilité du modèle lors de l'utilisation des deux échantillons.

Le choix se situe donc entre le modèle obtenu sur le logarithme de X à partir des moments successifs, et celui obtenu à partir des moments successifs associés aux moments inverses de même rang pour la variable non transformée. Le premier modèle est de forme classique et marche ici particulièrement bien. La densité de ce modèle a l'expression suivante :

$$\begin{aligned} f^*_{ln}(y) &= 0.07844184\,\exp(1.6672990y - 7.1506763y^2 + 7.9424139y^3 \\ &\quad -3.44404480y^4 + 0.6550231y^5 - 0.04608337y^6)\ ,\ y \in [\ln(0.5), \ln(120.5)]. \end{aligned}$$

Le deuxième modèle a également de bonnes performances, mais présente une forme vraiment particulière :

$$\begin{aligned} f^*(x) &= 1.39537650\,\exp(-0.20246146x + 0.0024231x^2 - 1.131294.10^{-5}x^3 \\ &\quad -21.593980/x + 36.684/x^2 - 17.4/x^3)\ ,\ x \in [0.5, 120.5]. \end{aligned}$$

Nous avons renouvelé la construction et le test de modèles construits par "half-sampling" pour ces deux familles d'observables. Le tableau III.10 présente les résultats obtenus pour 10 échantillonnages (les résultats croisés pour chaque couple d'échantillons (E_1, E_2) sont séparés par une ligne horizontale, et chaque ligne présente les p-values obtenues pour chaque famille, à partir du même échantillon).

Échantillon d'estimation/de test E_1/E_2 E_2/E_1	
Logarithmes des données, y^k, $Y = \ln(X)$	Données non-transformées x^k et $1/x^k$
0.67	**0.44**
0.061	**0.067**
0.52	**0.25**
0.38	**0.56**
0.48	**0.70**
0.19	0.033
0.32	**0.17**
0.17	**0.29**
0.26	**0.28**
0.093	**0.059**
0.071	0.031
0.022	**0.081**
0.023	0.061
0.044	0.013
0.009	0.011
0.038	0.01
0.0085	0.0032
0.169	**0.12**
0.001	0.0075
0.11	**0.29**

TAB. III.10 – *p-values obtenues pour 10 essais de "half-sampling", pour les deux familles d'observables retenues*

Sur ces 10 essais, il semble donc que les deux familles d'observables aient des comportements très similaires.

▷ **Remarques**

Les calculs effectués à partir de la densité f^*_{ln} sont extrêmement sensibles aux arrondis des multiplicateurs si on ne prend pas en compte l'ordre de puissance des observables (en utilisant des arrondis à deux chiffres après la virgule pour tous les multiplicateurs, on obtient une espérance de 0.961, alors que la valeur obtenue à partir des multiplicateurs exacts est de 2.664, pour une moyenne empirique de 2.665).

Les densités obtenues à partir des échantillons E1 ou E2 ont les formes suivantes :

- pour les lois sur $Y = \ln(X)$:

$$
\begin{aligned}
f_{ln}^{1*}(y) &= 0.08877160 \exp(1.7193904y - 7.6824666y^2 + 8.4467772y^3 \\
&\quad -3.6316647y^4 + 0.68508127y^5 - 0.047765237y^6) \\
f_{ln}^{2*}(y) &= 0.06855724 \exp(1.6220081y - 6.6357946y^2 + 7.4691185y^3 \\
&\quad -3.2759361y^4 + 0.62992999y^5 - 0.044846296y^6) \; ;
\end{aligned}
$$

- pour les lois sur X :

$$
\begin{aligned}
f^{1*}(x) &= 1.5039988 \exp(-0.20509718x + 0.002406884x^2 - 1.077384.10^{-5}x^3 \\
&\quad -22.24093/x + 38.202/x^2 - 18.20/x^3) \\
f^{2*}(x) &= 1.3369017 \exp(-0.20239026x + 0.0025049144 * x^2 - 1.2344461.10^{-5}x^3 \\
&\quad -21.06745/x + 35.290/x^2 - 16.60/x^3)
\end{aligned}
$$

5 Conclusion

La modélisation par PME d'une variable dont la distribution empirique a une forme aussi particulière présente plusieurs avantages. Par rapport à une modélisation par estimation fonctionnelle, elle permet d'obtenir une densité paramétrée qui pourrait être utilisée dans d'autres études. Cela pourrait être le cas ici pour l'étude comparée des durées de séjour de ce service et de celles d'autres hôpitaux. Il existe en effet des données en plus grand nombre dans ce domaine, pour des services de gériatrie d'hôpitaux londoniens, à partir desquelles la structure des durées de séjour a été largement étudiée ([Millard & Mc Clean 93]). De telles comparaisons sont plus aisées si on dispose de modèles paramétrés, et le type d'observables adéquats ou les tailles des informations utilisées dans chaque cas peut servir de base de comparaison.

Par rapport à d'autres modèles paramétrés, ceux obtenus par PME sont de plus clairement interprétables, même si les paramètres en eux-mêmes sont difficilement lisibles. Le meilleur modèle retenu ici a été obtenu, après le changement de variable logarithmique, en fixant la valeur des 6 premiers moments. Son utilisation paraît préférable à celui correspondant à la famille qui mèle moments successifs et moments inverses qui semble plus instable. De plus, l'utilisation d'observables de type

$1/x^k$ s'explique difficilement. On peut par contre relire le modèle construit sur les logarithmes des données en terme de résumé des données qu'il a fallu retenir pour obtenir la loi la plus adaptée. Choisir une transformation adaptée des observations et utiliser des observables élémentaires en nombre suffisant se justifie beaucoup plus facilement. Cela donne de plus l'idée d'une technique de modélisation qu'il serait intéressant de tester sur d'autres applications.

La recherche des modèles de maximum d'entropie a d'autre part la particularité d'être très dépendante de la méthode de résolution des équations du PME et donc des moyens de calculs (puissance des machines, vitesse de l'algorithme, choix des valeurs initiales, conditions d'arrêt, ...). Dans le cas des modèles usuels, l'utilisation des moments empiriques assure l'équivalence avec la méthode du maximum de vraisemblance qui a été implémentée dans de nombreux logiciels. Par contre, pour des modèles originaux, on se retrouve face au problème de la résolution d'un système d'équations non linéaires. Ce type de problème est par contre largement étudié sous l'angle de la méthode des moments avec critère de maximisation d'entropie ([Vorobyev 65, Lawrence & Papanicolaou 97]), mais dans une autre optique que celle d'une modélisation statistique.

Bibliographie

[Ben-Tal *et al.* 88] A. Ben-Tal, J.-M. Borwein & M. Teboulle "A dual approach to multidimensional $L_p$spectral estimation problems", *SIAM J. control optimization*, **26**, 985-996, 1988.

[Ben-Tal & Charnes 79] A. Ben-Tal & A. Charnes "A dual optimization framework for some problems of information theory and statistics", *Problems of control and information theory*, **8**, 387-401, 1975.

[Decarreau *et al.*] A. Decarreau, D. Hilhorst, C. Lemaréchal, J. Navaza "Dual methods in entropy maximization. Application to some problems in cristallography", *SIAM J. optim.*, **2**, 2, 173-197, 199.

[Delhumeau 97] C. Delhumeau *Confrontation de la classification administrative hospitalière des durées de séjour en gériatrie avec une classification issue de la modélisation statistique : mémoire de stage de DEA à l'INRIA Rhône-Alpes et au CHU de Grenoble*, 1997.

[Iusem & Teboulle 92] A. N. Iusem & M. Teboulle "A primal-dual iterative algorithm for a maximum likelihood estimation problem", *Comput. stat. data anal.*, **14**, 4, 443-456, 1992.

[Lawrence & Papanicolaou 97] R. Lawrence Mead & N. Papanicolaou "Maximum entropy in the problem of moments", *Journal of Mathematical Physics*, **25**, 8, 2404-2417, 1997.

[Lemaréchal 93] C. Lemaréchal "Duality without tears", in *Maximum entropy and bayesian methods* , edited by A. Mohammad-Djafari & G. Demoment, Kluwer academic publishers, 127-132, 1993.

[Millard & Mc Clean 93] Ph. Millard & S. Mc Clean "Patterns of length of stay after admission in geriatric medicine : an event history approach", *The Statistician*, **42**, 263-274, 1993.

[Powell 70] M.J.D. Powell "A hybrid method for non-linear equations", in *Numerical methods for non-linear algebraic equations*, edited by P. Rabinowitz, Gordon and Breach, 1970.

[Rockafellar 70] R.-T. Rockafellar *Convex analysis*, Princeton University Press, 1970.

[Saporta 90, p. 335] G. Saporta *Probabilités, analyse des données et statistique*, Technip 1990.

[Vorobyev 65] V. Vorobyev *Method of moments in applied mathematics*, eds. Gordon & Breach, 1965.

[Zhu 92] J. Zhu "A path following algorithm for a class of convex programming problems", *Z. oper. res.*, **36**, 4, 359-377, 1992.

Conclusion générale

En physique statistique, le principe de maximum d'entropie est lié à des lois telle que la loi de conservation de l'énergie pour un gaz parfait. De telles lois imposent le choix des observables et des contraintes décrivant le système. Mais dans le domaine de la statistique appliquée à la médecine ou à la sociologie par exemple, le modélisateur ne dispose pas de telles lois qui le guideraient dans le choix des observables. Ainsi, les étapes de l'application du principe de maximum d'entropie en inférence statistique peuvent paraître subjectives :

- pourquoi maximise-t-on l'entropie de Kullback plutôt qu'une autre fonctionnelle?
- sur quoi se base le choix de la mesure de référence?
- comment choisir les observables?
- comment déterminer les niveaux de contraintes?

Pour répondre à la première question, nous nous sommes référés dans ce travail à la propriété de concentration de l'entropie de Kullback. Pour simplifier, on peut dire que cette propriété, qui n'est vérifiée que pour cette fonctionnelle, permet d'affirmer que les modèles envisageables vérifiant les contraintes d'une information donnée sont indistinguables de la loi maximisant l'entropie.

En ce qui concerne la mesure de référence, nous nous sommes restreints aux mesures de Lebesgue ou de comptage, c'est-à-dire à l'utilisation de l'entropie de Shannon. Remarquons néammoins que les divers exemples traités montrent que des moments d'ordre 1 et 2 constituent une information standard pour l'étude de variables continues : le modèle obtenu est alors le même que celui issu de la maximisation de l'entropie de Kullback avec comme mesure de référence un vecteur gaussien. L'utilisation de l'entropie de Shannon nous limite a priori à des situations statistiques relativement simples et une extension de ce travail pourrait donc se baser sur l'étude de situations où d'autres mesures de référence s'imposent ou sont envisageables.

Les relectures des modèles classiques que nous avons détaillées montrent que l'utilisation d'observables élémentaires permet de traiter efficacement un grand nombre de situations de base. En effet, les observables constitués des moments d'ordre 1 et 2 pour les variables continues, d'indicatrices pour les variables discrètes et des produits $x_i y_j$, où x_i désigne une composante continue et y_j une composante discrète, apparaissent comme des observables standard. Ces observables suffisent à caractériser les modèles log-linéaires, les modèles graphiques gaussiens, les modèles d'analyse discriminante linéaire et quadratique, les modèles d'indépendance conditionnelle et la procédure de régression logistique, et à les présenter comme des applications directes du principe de maximum d'entropie.

Cette relecture des modèles classiques repose sur le théorème qui montre que s'il existe un état de Gibbs qui satisfait aux contraintes, alors celui-ci est l'unique état de maximum d'entropie. Ces états constituent une classe de modèles dont la manipulation est particulièrement aisée. Nous pensons que, dans les situations pratiques, il doit être presque toujours possible de se ramener à leur utilisation. Prenons par exemple le cas de la modélisation d'un vecteur de variables discrètes, pour lequel il existe une contrainte d'impossibilité d'une configuration $\boldsymbol{x}_0$. L'élément de connaissance $(\mathbf{1}_{\{\boldsymbol{x}_0\}}, 0)$ devrait donc apparaître dans l'information a priori. Dans ce cas, il n'existe pas d'état de Gibbs satisfaisant aux contraintes. Mais en transformant l'élément de connaissance précédent en $(\mathbf{1}_{\{\boldsymbol{x}_0\}}, \epsilon)$, où $\epsilon > 0$, il se peut qu'un état de Gibbs les vérifie. Une telle procédure consiste donc à dire qu'un évènement physiquement impossible sera de probabilité faible, mais non nulle. Mais la transformation d'un évènement impossible en évènement hautement improbable auquel aboutit ce choix n'est pas plus choquante que la modélisation de la taille d'un individu ou du taux de cholestérol par une loi normale sur $\mathbf{R}$, alors que de telles quantités ne peuvent pas être négatives.

Toute loi obtenue par application du principe de maximum d'entropie met en jeu deux familles de paramètres : les niveaux de contraintes et les multiplicateurs. Le choix d'utiliser des niveaux empiriques conduit à retrouver les équations du maximum de vraisemblance comme équations à résoudre pour obtenir les valeurs de ces multiplicateurs. À ce propos, on peut remarquer qu'une justification usuelle de l'estimation de paramètres par application de la méthode de maximum de vraisemblance repose sur la qualité des estimateurs obtenus et sur l'argumentation suivante : soit $\mathcal{S} = \{\mathcal{X}, \mathcal{A}, P_\theta, \theta \in \Theta\}$ une structure statistique sur l'ensemble mesuré $(\mathcal{X}, \mathcal{A})$, et $(X_n)_{n\in\mathbf{N}}$ une suite de variables aléatoires indépendantes issues de simulations de $P_{\theta_0} \in \mathcal{S}$. En notant $L_n(X_1, \ldots, X_n, \theta)$ la vraisemblance de $(X_1, \ldots, X_n)$ pour la loi P_θ, la loi forte des grands nombres assure que :

$$\frac{1}{n}L_n(X_1, \ldots, X_n, \theta) \longrightarrow E_{\theta_0}(\ln L_1(X_1, \theta)) \text{ (convergence } P_{\theta_0} \text{ presque sûre).}$$

Or, cette espérance est, à un facteur constant près, l'entropie de Kullback de P_θ avec comme mesure de référence P_{θ_0}. Le fait que cette entropie soit maximum pour $\theta = \theta_0$ explique l'intérêt de chosir l'estimateur du maximum de vraisemblance comme estimateur de θ.

Le lien entre cette convergence et le choix de $\mathcal{S}$ n'est en général pas explicité. Ainsi, même dans la justification la plus classique de la méthode d'estimation utilisant le principe de vraisemblance, le principe de maximum d'entropie semble postérieur au choix de la structure statistique.

Mais, bien que cette convergence établisse un lien entre maximum d'entropie et maximum de vraisemblance pour l'obtention des paramètres, la propriété de concentration des lois de maximum d'entropie ne se lit pas directement sur ce passage à la limite. Dans notre présentation, le choix des observables détermine entièrement la structure statistique et le calcul des paramètres est explicitement lié au choix de ces observables. En effet, les multiplicateurs sont obtenus en résolvant les équations que constituent les contraintes: si on travaille à espérances fixées aux moyennes empiriques, les multiplicateurs obtenus sont alors les paramètres qui maximisent la vraisemblance.

La relecture des modèles classiques apporte un autre éclairage sur leur nature dont nous espérons qu'il peut améliorer la compréhension de certains d'entre eux. Ainsi, pour la procédure de régression logistique, il nous semble plus clair de la présenter en terme d'information plutôt que de poser une hypothèse de linéarité du logit des probabilités conditionnelles, qui impose de plus des considérations ambiguës sur l'absence d'influence du schéma d'échantillonnage sur l'estimation des paramètres.

Après cette relecture de modèles classiques, nous nous sommes intéressés à la modélisation d'une variable (la durée de séjour dans un service de gériatrie) dont la répartition n'évoque aucune loi usuelle simple. L'application du principe de maximum d'entropie a permis de déterminer différents modèles exponentiels parmi lesquels nous avons sélectionné un modèle à 6 paramètres. La densité de ce modèle s'obtient plus simplement que par application des formules de lissage ou des techniques de noyaux. Par rapport à d'autres approximations fonctionnelles, et bien que la densité n'ait pas une forme directement interprétable, la manière dont elle a été obtenue est parfaitement lisible. Cette loi a en effet été déterminée à espérances fixées et a l'avantage de "concentrer" les lois satisfaisant aux contraintes correspondantes, qui sont de plus de nature élémentaire puisqu'il s'agit d'observables polynomiaux et de niveaux de contraintes empiriques.

Il est apparu, au travers de la relecture des modèles usuels et de cette application, que les observables les plus efficaces (de type polynomial) sont en fait ceux qui sont utilisés le plus souvent en physique statistique. Les références sont abondantes sur le sujet de la résolution des équations de maximum d'entropie dans le domaine physique, et divers logiciels proposent des algorithmes de résolution de telles équations. Une suite de ce travail résiderait donc dans l'adaptation et l'étude comparée des différents algorithmes de résolution des équations permettant d'obtenir les multiplicateurs des modèles correspondant à ce type d'observables. Il reste donc un grand travail de prise de connaissance des outils existant dans d'autres domaines que l'inférence statistique, ainsi que d'adaptation et de comparaison selon que l'on manipule un vecteur de variables continues, discrètes, bornées, ...

D'autre part, nous nous sommes restreints dans ce travail à des contraintes linéaires. Il serait bien sûr intéréssant d'aborder d'autres types de contraintes.

Liste des tableaux

Table des figures

Résumé
La question dont est issu ce travail est la suivante : le principe de maximum d'entropie (PME) qui est à la base de la physique statistique s'applique-t-il aux situations élémentaires et classiques de la statistique? Après avoir formalisé et justifié le PME lorsque l'entropie utilisée est l'entropie de Shannon et que l'information a priori est traduite par des contraintes linéaires, nous montrons que le problème de maximisation posé admet comme solution unique un état de Gibbs dont les paramètres satisfont à des équations qui se révèlent être le plus souvent des équations du maximum de vraisemblance. Nous proposons ensuite une relecture de modèles classiques qui s'avèrent être des applications simples du PME (modèles log-linéaires, modèles graphiques gaussiens, modèles d'analyse discriminante linéaire et quadratique, modèles a logit linéaire). Cette relecture, outre son intérêt didactique, permet de justifier l'utilisation de ces modèles par des propriétés de concentration de l'entropie. Enfin, nous appliquons le PME dans une situation issue du domaine de la médecine (modélisation de la durée de séjour dans un service de gériatrie), afin d'illustrer comment, en pratique, il est possible de résoudre les équations du PME. Cela nous amène à considérer le problème dual (au sens de la théorie de la dualité) du problème de maximisation sous contraintes qui est à la base de la définition du PME.
Mots-clés : Principe de maximum d'entropie ; états de Gibbs ; vraisemblance ; régression logistique ; analyse discriminante ; information empirique.

Aspects of the Maximum Entropy Principle in Statistical Modeling

Abstract
This thesis presents some aspects of the Maximum Entropy Principle (MEP) as a statistical modeling principle. We justify the MEP through a concentration theorem, without any subjectivity as regards the entropy concept. In this approach to statistical modeling, the first step is not the choice of a model but the formalization of information concerning the phenomenon studied. The constraints are defined by choosing some functions of the variables which represent the phenomenon, as well as the mean levels of these functions. We link the usual reading of exponential structure models precisely with their reading as Maximum Entropy models. For instance, if only empirical constraints are used, the Maximum Entropy equations are the Maximum Likelihood equations. We also use the MEP in a didactic point of view in order to analyse classical classification models. This allows a fresh understanding into the nature of these models (especially for the logistic regression method). Finally, we present an original model not corresponding to any classical model, built from an empirical distribution. Thus we illustrate the mechanism of MEP application.
Keywords : Maximum entropy principle ; Gibbs states ; likelihood ; logistic regression ; discriminant analysis ; empirical constraints.

Discipline : Mathématiques appliquées.

Laboratoire d'accueil : INRIA Rhône-Alpes ; projet IS2, 655 avenue de l'Europe, 38330, Montbonnot St Martin.