.

Notes de cours de statistiques paramétriques

MAT4513

TELECOM SudParis

Jean-Pierre Delmas

– 31 décembre 2024

Table des matières

References

Jean-Pierre Delmas, *Introduction aux Probabilités, Applications aux télécommunications avec exercices et problèmes commentés*, Ellipses, Paris, France, 2000.

Allan Gut, *Probability : a graduate course* Springer Text in Statistics, 2005.

Olivier Cappé, Maurice Charbit, Michel Grosnowski, Eric Moulines, *Traitement statistique du signal et de l'image*, polycopie ENST, Février 2002.

Erich L Lehmann, *Elements of large sample theory* Springer Text in Statistics, 2004.

Erich L Lehmann and George Casella *Theory of point estimation* Spinger Text in Statistics, Second edition, 1998.

1 Estimateur par intervalle de confiance ou région de confiance

1.1 Problématique et définition

Une *expérience aléatoire* est décrite mathématiquement par la donnée d'un ensemble $\mathcal{X}$ de résultats possibles de cette expérience. Toute observation possible, appelée aussi issue de cette expérience est un point $\mathbf{x} \in \mathcal{X}$ de cet ensemble muni d'une loi de probabilité P. Dans chaque problème considéré, on dispose d'une certaine connaissance a priori sur cette loi de probabilité P. En particulier, en traitement statistique du signal et des images (et dans d'autres disciplines appliquées), on déduit la forme de cette loi de probabilité des modèles physiques (ou d'approximations raisonnables de ces modèles physiques) qui indiquent comment ces observations $\mathbf{x}$ ont été engendrées.

Dans de nombreux modèles statistiques, cette loi de probabilité P appartient à une famille de loi de probabilité $\mathcal{P}$ et un paramètre $\theta \in \Theta$ apparaît de façon naturelle en association avec $\mathcal{P}$. Si pour toute valeur θ de ce paramètre, on ne peut associer qu'une seule probabilité P de $\mathcal{P}$, on dit que le modèle est *paramétrique* et l'on peut utiliser ce paramètre θ pour indexer la famille $\mathcal{P}$.

$$\mathcal{P} = \{P_\theta(\mathbf{x}), \theta \in \Theta\}.$$

On supposera de plus que l'application $\theta \mapsto P_\theta(\mathbf{x})$ est injective. C'est à dire que

$$P_{\theta_1}(\mathbf{x}) = P_{\theta_2}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \Rightarrow \theta_1 = \theta_2.$$

Sans cette condition essentielle appelé *condition d'identifiabilité*, la problématique de l'estimation paramétrique n'a pas de sens.

Dans le cas où θ ne caractérise pas complètement la probabilité P, on dit que le modèle est *semi-paramétrique*. Dans le cas où $\mathbf{x}$ représente un échantillon $(x_1, x_2, ..., x_n)$ de variables aléatoires x_k indépendantes de même loi de probabilité, θ peut représenter par exemple dans ce cadre semi-paramétrique la moyenne, la médiane et/ou la variance des variables aléatoires x_k. Dans ce chapitre, nous nous intéresserons essentiellement à l'estimation de paramètre dans le cadre paramétrique avec seulement quelques exemples dans le cadre semi-paramétrique.

Dans un cours précédent de statistique, nous avons estimé un paramètre inconnu θ d'un modèle paramétrique $\mathcal{P} = \{P_\theta(\mathbf{x}), \theta \in \Theta\}$ par une valeur unique $\widehat{\theta}(\mathbf{x})$ (où $\widehat{\theta}_n(\mathbf{x}_n)$ pour rappeler la taille n de l'observation $\mathbf{x}$) appelée *estimation ponctuelle*. Si l'estimateur $\widehat{\theta}(.)$ possède de bonnes propriétés (sans biais, variance minimale, efficacité,...), on peut s'attendre à ce que $\widehat{\theta}(\mathbf{x})$ soit proche de la vraie valeur θ. Cependant il est très peu probable que $\widehat{\theta}(\mathbf{x})$ soit exactement égal à θ. En particulier si la loi de probabilité de $\widehat{\theta}(\mathbf{x})$ est continue $P(\widehat{\theta}(\mathbf{x}) = \theta) = 0$.

Par conséquent, plutôt que d'estimer θ par la seule valeur $\widehat{\theta}(\mathbf{x})$, il semble raisonnable de proposer un ensemble de valeurs vraisemblables pour θ qu'il est intuitif de prendre proches de $\widehat{\theta}(\mathbf{x})$. Cet ensemble de valeurs est appelé *intervalle de confiance* quand θ est monodimensionnel ou *région de confiance* quand θ est multidimensionnel. Il s'agit d'une estimation ensembliste et dire que toutes les valeurs de cet ensemble sont vraisemblables pour θ, c'est dire qu'il y a une forte probabilité que θ appartienne à cet ensemble.

1.1.1 Intervalle ou région de confiance

De façon précise, nous avons la définition suivante :

Définition : Un intervalle ou région de confiance est une domaine aléatoire $\Delta^\alpha(\mathbf{x})$ tel que

$$P[\Delta^\alpha(\mathbf{x}) \ni \theta] = 1 - \alpha. \tag{1.1}$$

On dit que l'ensemble aléatoire $\Delta^\alpha(\mathbf{x})$ qui dépend de l'observation $\mathbf{x}$ est un *intervalle ou région de confiance* de *niveau ou degré de confiance* $1 - \alpha$. α est lui appelé *seuil de confiance* ou *niveau de*

signification. Les valeurs usuelles de α sont 10%, 5% ou 1%.

Remarque 1.1 : Toutes les définitions et méthodes que nous allons exposer dans ce chapitre s'étendent à des intervalles ou régions de confiance pour toute fonction $\phi(\theta)$, donc en particulier à des composantes particulière θ_k d'un paramètre multidimensionnel $\boldsymbol{\theta} \in \mathbb{R}^q$ du modèle paramétrique $\mathcal{P} = \{P_\theta(\mathbf{x}), \boldsymbol{\theta} \in \Theta\}$. Ce dernier cas est particulièrement utile en présence de *paramètres de nuisance*, c'est-à-dire dans le cas où $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ où $\boldsymbol{\theta}_1 = \phi(\boldsymbol{\theta})$ représente des paramètres d'intérêt et $\boldsymbol{\theta}_2$ des paramètres qu'on ne cherchera pas à estimer appelés paramètres de nuisance.

Remarque 1.2 : Les intervalles de confiance suscitent souvent des erreurs d'interprétation et des abus de langage. C'est la raison pour laquelle, nous avons évité d'écrire (1.1) sous la forme plus traditionnelle :

$$P[\theta \in \Delta^\alpha(\mathbf{x})] = 1 - \alpha,$$

qui laisserait comprendre que θ est aléatoire. Alors que θ est une quantité inconnue mais non aléatoire (on pourrait dire purement inconnue) et c'est l'intervalle ou la région de confiance $\Delta^\alpha(\mathbf{x})$ qui est aléatoire. Il est donc incorrect de dire par exemple que θ a 95% de chances d'être compris entre 33 et 35. Il est plus correct de dire que nous avons une confiance de 95% pour que θ soit compris entre 33 et 35.

Le problème consiste donc à trouver un procédé pour déterminer un intervalle ou région de confiance pour θ. Il semble intuitif de proposer un intervalle de confiance centré sur un estimateur ponctuel performant $\widehat{\theta}(.)$, c'est à dire de la forme

$$\Delta^\alpha(\mathbf{x}) = [\widehat{\theta}(\mathbf{x}) - \epsilon, \widehat{\theta}(\mathbf{x}) + \epsilon].$$

Il resterait alors à déterminer ϵ telle que :

$$P[\theta \in \Delta^\alpha(\mathbf{x})] = P[\widehat{\theta}(\mathbf{x}) - \epsilon \leq \theta \leq \widehat{\theta}(\mathbf{x}) + \epsilon] = P[|\widehat{\theta}(\mathbf{x}) - \theta| \leq \epsilon] = 1 - \alpha. \tag{1.2}$$

Mais cette démarche n'aboutit que dans des situations très spécifiques car α est en général fixé par avance (donc ne doit pas dépendre de θ qui est inconnu). Par ailleurs, ϵ ne doit pas non plus dépendre de θ pour que l'intervalle de confiance puisse être utilisé. En conséquence on ne peut déterminer un ϵ vérifiant (1.2) que si la loi de probabilité de $\widehat{\theta}(\mathbf{x}) - \theta$ ne dépende pas de θ, ce qui est tout à fait exceptionnel.

Notons dès maintenant, qu'un compromis devra être adopté entre le *degré de confiance* $1 - \alpha$ et la *précision* ϵ souhaitée pour cet intervalle ou région de confiance. En effet, plus on exigera un degré de confiance élevé $1 - \alpha$, moins on veut prendre le risque de se tromper en disant que θ est dans l'intervalle ou région, donc plus l'intervalle ou la région de confiance va s'agrandir. A la limite, on ne prend aucun risque ($\alpha = 0$) en proposant un intervalle de confiance $\mathbb{R}$ ou région $\mathbb{R}^q$ tout entier !

1.1.2 Exemple élémentaire

Considérons une observation $\mathbf{x} = (x_1, ..., x_n)$ constituée de n variables aléatoires $(x_1, ..., x_n)$ indépendantes de loi de probabilité gaussienne de moyenne inconnue θ et de variance connue σ_0^2. Dans ce cas la moyenne empirique $\frac{1}{n}\sum_{k=1}^n x_k$ est de loi de probabilité gaussienne de moyenne θ et de variance $\frac{\sigma_0^2}{n}$. Par suite la variable aléatoire

$$h(\mathbf{x}, \theta) \stackrel{\text{def}}{=} \left(\frac{1}{n}\sum_{k=1}^n x_k\right) - \theta$$

est de loi de probabilité gaussienne $\mathcal{N}(0, \frac{\sigma_0^2}{n})$ qui ne dépend pas de θ. Cela nous permet d'écrire, qu'il existe pour tout $\alpha \in]0,1[$, une valeur u_α telle que :

$$P\left[\left|\frac{\left(\frac{1}{n}\sum_{k=1}^n x_k\right) - \theta}{\frac{\sigma_0}{\sqrt{n}}}\right| \leq u_\alpha\right] = P\left(\theta \in \left[\frac{1}{n}\sum_{k=1}^n x_k - \frac{\sigma_0}{\sqrt{n}}\ u_\alpha, \frac{1}{n}\sum_{k=1}^n x_k + \frac{\sigma_0}{\sqrt{n}}\ u_\alpha\right]\right) = 1-\alpha.$$

D'où l'intervalle de confiance :

$$\Delta^\alpha(\mathbf{x}) = \left[\frac{1}{n}\sum_{k=1}^n x_k - \frac{\sigma_0}{\sqrt{n}}\ u_\alpha; \frac{1}{n}\sum_{k=1}^n x_k + \frac{\sigma_0}{\sqrt{n}}\ u_\alpha\right].$$

Cet exemple élémentaire, va nous permettre de donner une méthodologie pour obtenir des intervalles ou régions de confiance.

1.1.3 Fonction pivotale

Pour trouver un intervalle ou région de confiance, la principale méthode consiste à trouver, comme dans l'exemple élémentaire précédent, une fonction $h(\mathbf{x}, \theta)$ dite *fonction pivotale*

$$(\mathbf{x}, \theta) \in \mathbb{R}^n \times \Theta \longmapsto h(\mathbf{x}, \theta) \in \mathbb{R},$$

dont la loi de probabilité ne dépend pas de θ et est non dégénérée (en pratique à densité de probabilité). Cette fonction s'écrira souvent à l'aide d'une estimation ponctuelle particuliere $\widehat{\theta}(\mathbf{x})$ ou d'une statistique exhaustive $s(\mathbf{x})$, c'est à dire sous la forme $k[\widehat{\theta}(\mathbf{x}), \theta]$ ou $k[s(\mathbf{x}), \theta]$ comme nous en verrons des exemples par la suite.

Cette fonction pivotale $h(\mathbf{x}, \theta)$ permet en général de trouver des régions $\Sigma(\alpha)$ de $\mathbb{R}$ qui ne dépendent pas de θ telle que :

$$P[h(\mathbf{x}, \theta) \in \Sigma(\alpha)] = 1 - \alpha. \tag{1.3}$$

Cette relation permet généralement de déduire *par inversion*, pour tout $\alpha > 0$ et $\mathbf{x} \in \mathcal{X}$, un domaine $\Delta^\alpha(\mathbf{x}) = \{\theta \in \Theta\ ; h(\mathbf{x}, \theta) \in \Sigma(\alpha)\}$ tel que :

$$\theta \in \Delta^\alpha(\mathbf{x}) \Leftrightarrow h(\mathbf{x}, \theta) \in \Sigma(\alpha).$$

Ce procédé permet de déterminer des intervalles ou régions de confiance lorsqu'on saura mettre en évidence une fonction pivotale. Très souvent la fonction $\theta \in \Theta \longmapsto h(\mathbf{x}, \theta) \in \mathbb{R}$ sera monotone, ce qui facilitera le calcul de la région $\Delta^\alpha(\mathbf{x})$ à partir de $\Sigma(\alpha)$.

Remarque 1.3 : En général la région $\Sigma(\alpha)$ solution de (1.3) n'est pas unique. Pour minimiser la longueur de l'intervalle ou le volume de la région de confiance $\Delta^\alpha(\mathbf{x})$, on aura souvent intérêt à minimiser la longueur de $\Sigma(\alpha)$ bien que cette minimisation n'assure pas toujours la minimisation de la longueur ou du volume de $\Delta^\alpha(\mathbf{x})$ pour α et $\mathbf{x}$ fixés.

Pour cela on dispose de la propriété suivante : Si $f_H(h)$ désigne la densité de probabilité de la variable aléatoire $h(\mathbf{x}, \theta)$, le domaine $\Sigma(\alpha)$ de longueur minimale satisfaisant la contrainte (1.3) est donné par l'ensemble $\Sigma(\alpha)$ des valeurs de h solutions de l'équation

$$f_H(h) \geq a(\alpha) \ \text{ où le seuil } \ a(\alpha) \ \text{ est ajusté pour que } \ \int_{\Sigma(\alpha)} f_H(h)dh = 1-\alpha. \tag{1.4}$$

Ainsi, dans le cas particulier où $f_H(h)$ est une fonction paire, le domaine $\Sigma(\alpha)$ sera symétrique par rapport à l'origine.

Preuve : Si $\Sigma(\alpha)$ vérifie (1.4) et si $\Sigma'(\alpha)$ est un autre intervalle tel que $P[h(\mathbf{x},\theta) \in \Sigma'(\alpha)] = 1-\alpha$, nous avons :

$$P[h(\mathbf{x},\theta) \in \Sigma(\alpha) \cap \bar{\Sigma}'(\alpha)] = P[h(\mathbf{x},\theta) \in \Sigma'(\alpha) \cap \bar{\Sigma}(\alpha)] = 1 - \alpha - P[h(\mathbf{x},\theta) \in \Sigma(\alpha) \cap \Sigma'(\alpha)]$$

avec

$$P[h(\mathbf{x},\theta) \in \Sigma(\alpha) \cap \bar{\Sigma}'(\alpha)] = \int_{\Sigma(\alpha)\cap\bar{\Sigma}'(\alpha)} f_H(h)dh \geq a(\alpha)\text{long}(\Sigma(\alpha) \cap \bar{\Sigma}'(\alpha))$$

et

$$P[h(\mathbf{x},\theta) \in \Sigma'(\alpha) \cap \bar{\Sigma}(\alpha)] = \int_{\Sigma'(\alpha)\cap\bar{\Sigma}(\alpha)} f_H(h)dh \leq a(\alpha)\text{long}(\Sigma'(\alpha) \cap \bar{\Sigma}(\alpha))$$

d'où

$$\text{long}(\Sigma(\alpha) \cap \bar{\Sigma}'(\alpha)) \leq \text{long}(\Sigma'(\alpha) \cap \bar{\Sigma}(\alpha)).$$

Ce qui implique : $\text{long}(\Sigma(\alpha) \cap \bar{\Sigma}'(\alpha)) + \text{long}(\Sigma(\alpha) \cap \Sigma'(\alpha)) \leq \text{long}(\Sigma'(\alpha) \cap \bar{\Sigma}(\alpha)) + \text{long}(\Sigma(\alpha) \cap \Sigma'(\alpha))$, d'où $\text{long}(\Sigma(\alpha)) \leq \text{long}(\Sigma'(\alpha))$. ■

Résultat de probabilités 1.1 : Dans tout ce qui suit nous aurons besoin de la notion de *quantile d'une loi de probabilité* d'une variable aléatoire scalaire :

Pour une variable aléatoire X à fonction de répartition $F_X(x)$ strictement monotone (c'est à dire continue), on appelle *quantile d'ordre* $p \in]0,1[$, la valeur q_p telle que

$$F_X(q_p) \stackrel{\text{def}}{=} P(X \leq q_p) = p. \tag{1.5}$$

Ainsi la *médiane* d'une variable aléatoire est son quantile d'ordre $1/2$. On parle quelquefois de *quartile, décile* et *centile* pour respectivement

$$p \in \{\frac{1}{4}, \frac{2}{4}, \frac{3}{4}\}, \quad p \in \{\frac{1}{10}, \frac{2}{10}, ..., \frac{9}{10}\} \quad \text{et} \quad p \in \{\frac{1}{100}, \frac{2}{100}, ..., \frac{99}{100}\}.$$

Par exemple pour la loi gaussienne centrée réduite $\mathcal{N}(0,1)$ qui satisfait :

$$P(|X| \leq 1,98) \approx 0.95 \quad \text{et} \quad P(|X| \leq 2,58) \approx 0.99,$$

le quantile d'ordre 0,025 est -1,98 ≈ -2 et celui d'ordre 0,975 est de 1,98 ≈ 2 et celui d'ordre 0,005 est -2,58 ≈ -2,6 et celui d'ordre 0,995 est de 2,58 ≈ 2,6.

On utilisera par ailleurs dans ce chapitre les quantiles de lois de probabilité gaussienne centrée réduite $\mathcal{N}(0,1)$, de Student à n degrés de liberté et de Khi deux à n degrés de liberté, et l'on utilisera les notations respectives

$$P[|U| \geq u_\alpha] = \alpha, \quad P[|T| \geq t_{n,\alpha}] = \alpha \quad \text{et} \quad P[Z \geq z_{n,\alpha}] = \alpha.$$

Si la fonction de répartition $F_X(x)$ n'est pas strictement monotone (par exemple pour une variable aléatoire discrète prenant les valeurs $(x_i)_{i\in\mathbb{N}}$, pour laquelle $F_X(x)$ est en "marche d'escalier"), on appellera quantile d'ordre $p \in]0,1[$, la valeur q_p définie par

$$q_p = \sup\{x_i; F_X(x_i) \leq p\}.$$

On aura ainsi en particulier $F_X(q_p) \leq p$.

Remarque 1.4 : L'intervalle ou la région de confiance $\Delta^\alpha(\mathbf{x})$ solution de (1.1) *n'est pas unique.* Dans certaines applications, on ne cherchera pas un intervalle $\Delta^\alpha(\mathbf{x})$ de longueur minimale, mais plutôt, un intervalle de la forme $]-\infty, a_1(\mathbf{x},\alpha)]$ où $]a_2(\mathbf{x},\alpha), +\infty]$ appelé *intervalle unilatéral* par opposition à un intervalle de la forme $]a_1(\mathbf{x},\alpha), a_2(\mathbf{x},\alpha)]$ appelé *intervalle bilatéral.* Ces divers choix dépendent du contexte dans lequel intervient la notion d'intervalle de confiance. Ainsi si θ est une résistance de rupture d'un métal, on veut assurer une résistance minimale, soit un intervalle de confiance de la forme $\theta > a_2(\mathbf{x},\alpha)$.

Par contre si θ est une concentration de poluant, on ne veut pas dépasser un seuil de toxicité : $\theta < a_1(\mathbf{x}, \alpha)$.

Dans la suite de ce chapitre, nous allons illustrer cette approche par fonction pivotale dans des cas particuliers.

1.2 Intervalle de confiance pour les paramètres de la loi normale

Considérons ici un modèle d'échantillonnage $\mathbf{x} = (x_1, ..., x_n)$ de taille n. Les variables aléatoires X_1,...,X_n sont indépendantes de mêmes distributions gaussiennes $\mathcal{N}(m, \sigma^2)$. Pour construire des intervalles de confiance pour les paramètres m et σ^2 nous aurons besoin des résultats suivants de probabilité :

Résultat de probabilités 1.2 :

1. Si X_1,... , X_n sont indépendantes de loi gaussiennes centrées réduites $\mathcal{N}(0, 1)$, alors

$$\chi^2 \stackrel{\text{def}}{=} \sum_{k=1}^{n} X_k^2 \text{ est de } \textit{loi de Khi-deux} \text{ à } n \text{ degrés de liberté (notée } \chi_n^2).$$

2. Si X est de loi $\mathcal{N}(0, 1)$ et Y de loi χ_n^2 et de plus si X et Y sont indépendantes, alors

$$T \stackrel{\text{def}}{=} \frac{X}{\sqrt{\frac{Y}{n}}} \text{ est de } \textit{loi de Student} \text{ à } n \text{ degrés de liberté (qui est une loi de densité de probabilité paire).}$$

3. Si X_1,... , X_n sont indépendantes de loi gaussiennes $\mathcal{N}(m, \sigma^2)$, alors

$$\text{la } \textit{moyenne empirique } \bar{X}_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} X_k \text{ est de loi gaussienne } \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

et si S_n^2 désigne la *variance empirique* $S_n^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2 = \frac{1}{n}\sum_{k=1}^{n} X_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} X_k\right)^2$ alors

$$\frac{nS_n^2}{\sigma^2} \text{ est de loi de Khi-deux à } n-1 \text{ degrés de liberté,} \tag{1.6}$$

de plus

$$\text{les variables aléatoires } \bar{X}_n \text{ et } S_n^2 \text{ sont indépendantes.} \tag{1.7}$$

1.2.1 Intervalle de confiance pour la moyenne

Deux cas sont à distinguer : où bien σ^2 est connu, alors le seul paramètre inconnu du modèle paramétrique est m, où bien σ^2 est aussi inconnu et le paramètre inconnu du modèle paramétrique est $\boldsymbol{\theta} = (m, \sigma^2)$ et nous nous intéressons seulement ici à $m = \phi(\boldsymbol{\theta})$.

Cas σ^2 connu : Ici $\bar{X}_n - m$ est de loi gaussienne $\mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$. Il s'agit donc d'une fonction pivotale pour le paramètre m. Par suite :

$$U \stackrel{\text{def}}{=} \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \text{ est de loi gaussienne } \mathcal{N}(0, 1),$$

d'où

$$P\left[\left|\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| \leq u_\alpha\right] = 1 - \alpha \text{ où } u_\alpha \text{ désigne la valeur telle que } P(|U| \geq u_\alpha) = \alpha.$$

Par suite nous obtenons l'intervalle de confiance :

$$m \in \left[\bar{X}_n - \frac{\sigma}{\sqrt{n}}\ u_\alpha, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\ u_\alpha \right] \text{ où } \bar{X}_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} x_k.$$

Remarque 1.5 : Nous remarquons sur ce premier exemple que la largeur $\frac{2\sigma}{\sqrt{n}}\ u_\alpha$ est une fonction à la fois décroissante de la taille n de l'observation et de α. Ceci est intuitif, car plus n croit, plus l'échantillon apporte de l'information et plus l'incertitude sur le paramètre diminue. De même, plus α croît, plus u_α décroît et moins cet intervalle de confiance a de chance de contenir la valeur de θ. Cette remarque est tout à fait générale pour tout intervalle de confiance.

Cas σ^2 inconnu : D'après les propriétés 2 et 3 précédemment rappelées ci dessus,

$$T \stackrel{\text{def}}{=} \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \frac{1}{\sqrt{\frac{nS_n^2}{(n-1)\sigma^2}}} = \frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n-1}}}$$

est de loi de Student à $n-1$ degrés de liberté. Par suite $\frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n-1}}}$ est une fonction pivotale pour le paramètre m et nous obtenons l'intervalle de confiance :

$$m \in \left[\bar{X}_n - \frac{S_n}{\sqrt{n-1}}\ t_{n-1,\alpha}, \bar{X}_n + \frac{S_n}{\sqrt{n-1}}\ t_{n-1,\alpha} \right] \text{ où } \bar{X}_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n} x_k \text{ et } S_n^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{X}_n)^2 \quad (1.8)$$

et où $t_{n-1,\alpha}$ désigne la valeur telle que $P(|T| \geq t_{n-1,\alpha}) = \alpha$.

On préfère quelquefois remplacer S_n^2 par l'estimateur sans biais de la variance σ^2

$$S'^2_n \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{X}_n)^2 = \frac{n}{n-1}S_n^2.$$

Dans ce cas, l'intervalle de confiance précédent s'écrit sous la forme :

$$m \in \left[\bar{X}_n - \frac{S'_n}{\sqrt{n}}\ t_{n-1,\alpha}, \bar{X}_n + \frac{S'_n}{\sqrt{n}}\ t_{n-1,\alpha} \right] \text{ où } S'^2_n \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{X}_n)^2.$$

1.2.2 Intervalle de confiance pour la variance

Ici aussi deux cas sont à distinguer : où bien m est connu, alors le seul paramètre inconnu du modèle paramétrique est σ^2, où bien m est aussi inconnu et le paramètre inconnu du modèle paramétrique est $\boldsymbol{\theta} = (m, \sigma^2)$ et nous nous intéressons seulement ici à $\sigma^2 = \phi(\boldsymbol{\theta})$.

Cas m connu : Ici $\sum_{k=1}^{n}\left(\frac{x_k - m}{\sigma}\right)^2 = \frac{nS''^2_n}{\sigma^2}$ (où $S''^2_n \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - m)^2$) est d'après la propriété 1 précédente de loi de Khi-deux à n degrés de liberté. Par suite $Z \stackrel{\text{def}}{=} \frac{nS''^2_n}{\sigma^2}$ est une fonction pivotale pour σ^2. Il existe donc des intervalles $[a, b]$ tels que :

$$P\left(a \leq \frac{nS''^2_n}{\sigma^2} \leq b\right) = P\left(\frac{nS''^2_n}{b} \leq \sigma^2 \leq \frac{nS''^2_n}{a}\right) = 1 - \alpha.$$

Ici un intervalle $[a, b]$ de longueur minimale pour $\Sigma(\alpha)$ n'entraînera pas un intervalle de longueur minimale pour $\Delta^\alpha(\mathbf{x})$. Il est d'usage ici de ne plus prendre l'intervalle $[a, b]$ de longueur minimale mais de

procéder plutôt à un équilibrage des risques, c'est à dire à prendre a et b tels que

$$P\left(\frac{nS"_n^2}{\sigma^2} \leq a\right) = P\left(\frac{nS"_n^2}{\sigma^2} \geq b\right) = \frac{\alpha}{2}.$$

Par suite si $z_{n,\alpha}$ désigne la valeur telle que $P(Z \geq z_{n,\alpha}) = \alpha$, alors $a = z_{n,1-\frac{\alpha}{2}}$ et $b = z_{n,\frac{\alpha}{2}}$ (en fait quantile d'ordre respectivement de $\frac{\alpha}{2}$ et $1-\frac{\alpha}{2}$) et nous obtenons alors l'intervalle de confiance suivant :

$$\sigma^2 \in \left[\frac{nS"_n^2}{z_{n,\frac{\alpha}{2}}}, \frac{nS"_n^2}{z_{n,1-\frac{\alpha}{2}}}\right] \text{ où } S"_n^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - m)^2.$$

Cas m inconnu : Ici d'après la propriété 3, $\frac{nS_n^2}{\sigma^2}$ est loi de Khi-deux à $n-1$ degrés de liberté. Par suite $Z \stackrel{\text{def}}{=} \frac{nS_n^2}{\sigma^2}$ est une fonction pivotale pour σ^2. Nous pouvons reprendre la démarche précédente et nous obtenons alors l'intervalle de confiance :

$$\sigma^2 \in \left[\frac{nS_n^2}{z_{n-1,\frac{\alpha}{2}}}, \frac{nS_n^2}{z_{n-1,1-\frac{\alpha}{2}}}\right] \text{ où } S_n^2 \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{X}_n)^2. \tag{1.9}$$

Remarque 1.6 : Ces deux intervalles de confiance pour σ^2 ne sont pas de la forme $[\widehat{\theta} - \epsilon, \widehat{\theta} + \epsilon]$ avec $\widehat{\theta} = S"_n^2$ ou $\widehat{\theta} = S_n^2$, mais de la forme

$$[\epsilon_1\widehat{\theta}, \epsilon_2\widehat{\theta}] \text{ où } \epsilon_1 < \epsilon_2,$$

avec pour les valeurs usuelles de α et n, $\epsilon_1 < 1$ et $\epsilon_2 > 1$. En général la structure de l'intervalle de confiance est imposé par le choix de la fonction pivotale qui notons le n'est pas unique.

Remarque 1.7 : Notons que l'intervalle de confiance pour le paramètre σ^2 est légèrement plus étroit lorsque m est connu que lorsque m est inconnu en étant quasi identique pour les grandes valeurs de n. A chaque fois qu'un paramètre est connu, nous avons intérêt a en tenir compte dans le choix de la fonction pivotale.

1.2.3 Région de confiance pour (m, σ^2)

Dans le cas où m et σ^2 sont inconnus et où l'on s'intéresse à la fois à m et σ^2, $\boldsymbol{\theta} = (m, \sigma^2)$. Dans ce cas la variable aléatoire

$$h(\mathbf{x}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right) - m}{\frac{\sigma}{\sqrt{n}}}$$

est de loi $\mathcal{N}(0,1)$ et donc une fonction pivotale avec :

$$P[|h(\mathbf{x}, \boldsymbol{\theta})| \leq u_\alpha] = 1 - \alpha.$$

Cependant la région de confiance associée

$$\Delta^\alpha(\mathbf{x}) = \left\{(m, \sigma^2); \left((\frac{1}{n}\sum_{k=1}^{n} x_k) - m\right)^2 - \frac{\sigma^2}{n}u_\alpha^2 \leq 0\right\}$$

est dans $\mathbb{R}^2$, la région intérieure à la parabole $\sigma^2 = \frac{n(m-\bar{X}_n)^2}{u_\alpha^2}$. Cette région non bornée de $\mathbb{R}^2$ n'a pas d'intérêt pratique. Nous lui préférerons des régions de confiance sous forme de rectangles où d'ellipses que nous obtiendrons respectivement sous forme de région de confiance par excès (1.12) ou de région de confiance asymptotique (1.33).

1.3 Intervalle de confiance pour une proportion

Le problème connu sous le nom d'intervalle de confiance d'une proportion est en fait le problème de la détermination d'un intervalle de confiance pour le paramètre $p \in]0,1[$ de la *loi de probabilité de Bernoulli* au vu d'un échantillon de variables aléatoires indépendantes $x_1, .., x_n$ de cette loi. Nous donnerons deux exemples de ce problème dans le contexte des sondages d'opinion et dans celui de l'estimation d'un taux d'erreur en communications numériques dans le paragraphe **??**.

Dans ce cadre nous savons que l'estimateur

$$\widehat{p}(\mathbf{x}) \stackrel{\text{def}}{=} \bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} x_k,$$

est sans biais de variance minimale et de plus efficace. Nous savons également que la statistique

$$s(\mathbf{x}) = \sum_{k=1}^{n} x_k$$

est une statistique exhaustive (la distribution de $(x_1, .., x_n)$ conditionnelle à la connaissance de $s(\mathbf{x})$ ne dépend plus du paramètre p). Par suite toute fonction pivotale du paramètre p devrait être une fonction de $s(\mathbf{x})$ ou de $\bar{X}_n$. Mais la loi de probabilité de $s(\mathbf{x})$ est une loi Binomiale de paramètre p et n, i.e.,

$$P[s(\mathbf{x}) = k] = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{pour} \quad k \in \{0, ..., n\}$$

qui n'est pas facile à manipuler.

Remarque 1.8 : De plus, puisque $s(\mathbf{x})$ est une variable aléatoire discrète, tout intervalle de confiance $\Delta^\alpha(\mathbf{x})$ pour le paramètre p (construit à partir de $s(\mathbf{x})$) ne peut qu'être discret (au sens que l'on peut compter ces intervalles). Par suite il ne peut exister d'intervalle de confiance $\Delta^\alpha(\mathbf{x})$ tel que l'égalité (1.1) soit exacte pour n'importe quelle valeur du degré de confiance $1-\alpha$ imposé.

Par suite il n'existe pas ici de fonction pivotale au sens général exact (1.3) pour toute valeur de α. Nous ne pourrons exhiber en pratique que des intervalles de confiance approchés (par excès ou asymptotiques) pour le paramètre p que l'on décrira dans le paragraphe **??**.

Cette remarque s'applique aussi à toute observation $\mathbf{x}$ de distribution de probabilité discrète car $P[\Delta^\alpha(\mathbf{x}) \ni \theta]$ ne peut prendre alors que des valeurs discrètes.

1.4 Intervalles de confiance de quantiles d'un échantillon

Nous considérons ici un échantillon $\mathbf{x} = (x_1, x_2, ..., x_n)$ où $(x_i)_{i=1,..,n}$ sont indépendants et de même loi de probabilité P inconnue de fonction de répartition $F_X(x)$ continue strictement monotone croissante, dont on s'intéresse aux quantiles q_p défini par (1.5). Dans ce cadre, on définit à partir de $\mathbf{x}$, la *statistique d'ordre* $(x_{(1)}, x_{(2)}, ..., x_{(n)})$ qui représente les valeurs $(x_i)_{i=1,..,n}$ ordonnées dans l'ordre croissant $((x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}))$. Puisque les quantiles ne caractérisent pas la loi de probabilité P, il s'agit ici d'estimation semi-paramétrique. Pour obtenir directement des intervalles de confiance pour les paramètres q_p de P, on utilisera le résultat suivant :

Résultat de probabilités 1.3 : $\forall (i,j)$ tel que $1 \leq i < j \leq n$, on a la propriété suivante :

$$P(x_{(i)} \leq q_p \leq x_{(j)}) = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k} \quad \text{avec} \quad F_X(q_p) = p,$$

Preuve : Puisque :

$$\{x_{(i)} \leq q_p\} = (\{x_{(i)} \leq q_p\} \cap \{x_{(j)} \geq q_p\}) \cup \{x_{(j)} < q_p\}$$

qui permet d'écrire puisque $\{x_{(i)} \leq q_p\}$ signifie qu'au moins i valeurs de $(x_k)_{k=1,...,n}$ sont inférieurs ou égaux à q_p :

$$\begin{aligned} P(x_{(i)} \leq q_p \leq x_{(j)}) &= P(x_{(i)} \leq q_p) - P(x_{(j)} < q_p) = F_{x_{(i)}}(q_p) - F_{x_{(j)}}(q_p) \\ &= \sum_{k=i}^{n} \binom{n}{k} [F(q_p)]^k [1 - F(q_p)]^{n-k} - \sum_{k=j}^{n} \binom{n}{k} [F(q_p)]^k [1 - F(q_p)]^{n-k} \\ &= \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

■

Par suite, si pour le degré de confiance $1 - \alpha$, il existe deux entiers i et j tels que

$$\sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k} = 1 - \alpha$$

avec $1 \leq i < j \leq n$, nous obtenons l'intervalle de confiance pour le paramètre q_p :

$$q_p \in [x_{(i)}, x_{(j)}]. \tag{1.10}$$

Dans la pratique, on cherchera pour α fixé, le couple (i, j) telle que la somme $\sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}$ soit la plus proche possible de $1 - \alpha$ et nous aurons seulement :

$$P\left(q_p \in [x_{(i)}, x_{(j)}]\right) \approx 1 - \alpha,$$

qui sera un intervalle de confiance approché, dont on verra d'autres exemples dans la section 1.5.

Exemple 1.1 : Si l'on s'intéresse à la médiane, on a $p = 1/2$. On cherchera alors i et j tels que $\frac{1}{2^n} \sum_{k=i}^{j-1} \binom{n}{k}$ soit la plus proche possible de $1 - \alpha$. Ainsi pour $n = 10$, on a $\frac{1}{2^{10}} \sum_{k=3}^{7} \binom{10}{k} \approx 0,89$. On en déduit que $[x_{(3)}, x_{(8)}]$ est un intervalle de confiance approché de la médiane $q_{1/2}$ avec un degré de confiance de 0,89.

1.5 Régions et intervalles de confiance approchés

Il est assez rare que l'on puisse obtenir des fonctions pivotales nous permettant d'en dériver des intervalles ou régions de confiance exacts. On se contente en pratique soit d'intervalles ou de régions de confiance approchés, soit dans le cadre d'une séquence d'observations $\mathbf{x}_n$, d'*intervalles ou régions de confiance asymptotiques* obtenus à l'aide de *fonctions asymptotiquement pivotales.*

1.5.1 Régions et intervalles de confiance par excès

Définition : Un intervalle ou région de confiance par excès est une domaine aléatoire $\Delta^\alpha(\mathbf{x})$ tel que

$$P[\Delta^\alpha(\mathbf{x}) \ni \theta] \geq 1 - \alpha. \tag{1.11}$$

L'intérêt de cette notion est de fournir des approximations de régions ou d'intervalles de confiance quand on ne sait pas exhiber de fonctions pivotales, mais également de fournir une région de confiance approchée sous forme de pavé pour les paramètres $\boldsymbol{\theta} = (\theta_1, .., \theta_q)$ multidimensionnels à partir d'intervalles de confiance de leurs composantes scalaires $\theta_1, ...\theta_q$.

Exemple 1.2 : Revenons ici à l'intervalle de confiance d'une proportion de la section 1.3. Nous pouvons obtenir facilement un intervalle de confiance par excès du paramètre p à l'aide de *l'inégalité de Bienaymé-*

Tchebychev appliquée à l'estimateur $\widehat{p}(\mathbf{x}) = \bar{X}_n$, d'espérance mathématique p et de variance $\frac{p(1-p)}{n}$ car

$$P[|\bar{X}_n - p| \leq a] \geq 1 - \frac{\mathrm{var}(\bar{X}_n)}{a^2} = 1 - \frac{p(1-p)}{na^2} \geq 1 - \frac{1}{4na^2},$$

donne l'intervalle de confiance par excès de niveau de confiance α suivant :

$$p \in \left[\bar{X}_n - \frac{1}{2\sqrt{\alpha n}}, \bar{X}_n + \frac{1}{2\sqrt{\alpha n}}\right].$$

Propriété : Supposons dans le cadre d'un paramètre multidimensionnel $\boldsymbol{\theta} \in \mathbb{R}^q$ que $\Delta_i^{\alpha_i}(\mathbf{x})$ soit un intervalle de confiance exact ou par excès de niveau de confiance $1 - \alpha_i$ pour θ_i pour tout $i \in \{1, .., q\}$, c'est à dire :

$$P[\Delta_i^{\alpha_i}(\mathbf{x}) \ni \theta_i] \geq 1 - \alpha_i.$$

Alors si $\Delta(\mathbf{x})$ désigne le produit cartésien des q intervalles $\Delta_i^{\alpha}(\mathbf{x})$, nous avons

$$\begin{aligned} P[\Delta(\mathbf{x}) \ni \boldsymbol{\theta}] &= P[\cap_{i=1}^q(\theta_i \in \Delta_i^{\alpha_i}(\mathbf{x}))] = 1 - P[\cup_{i=1}^q(\theta_i \notin \Delta_i^{\alpha_i}(\mathbf{x})] \\ &\geq 1 - \sum_{i=1}^q P[\theta_i \notin \Delta_i^{\alpha_i}(\mathbf{x})] \geq 1 - \sum_{i=1}^q \alpha_i, \end{aligned}$$

où nous avons utilisé l'égalité de Morgan $\overline{\cap_i A_i} = \cup_i(\bar{A}_i)$ pour la deuxième égalité et la propriété suivante $P(\cup_i A_i) \leq \sum_i P(A_i)$ dans la première inégalité.

Ainsi $\Delta(\mathbf{x})$ est un pavé de confiance par excès pour le paramètre $\boldsymbol{\theta}$ de niveau de confiance $1 - \sum_{i=1}^q \alpha_i$. Nous voyons ici que pour un seuil de confiance fixé $\alpha = \sum_{i=1}^q \alpha_i$, des compromis seront à effectuer dans le choix de chacun des seuils α_i selon la précision exigée pour chacune des composantes θ_i de $\boldsymbol{\theta}$.

Exemple 1.3 : Revenons à l'échantillon gaussien de la section 1.2. Les intervalles de confiance (1.8) et (1.9) nous permettent d'obtenir d'après la propriété précédente le pavé de confiance par excès de niveau de confiance $1 - (\alpha_1 + \alpha_2)$ du paramètre $\boldsymbol{\theta} = (m, \sigma^2)$.

$$(m, \sigma^2) \in \left[\bar{X}_n - \frac{S_n}{\sqrt{n-1}}\, t_{n-1,\alpha_1}, \bar{X}_n + \frac{S_n}{\sqrt{n-1}}\, t_{n-1,\alpha_1}\right] \times \left[\frac{nS_n^2}{z_{n-1,\frac{\alpha_2}{2}}}, \frac{nS_n^2}{z_{n-1,1-\frac{\alpha_2}{2}}}\right]. \tag{1.12}$$

1.5.2 Régions et intervalles de confiance asymptotiques (généralités)

Définition : Un intervalle ou région de confiance asymptotique de niveau de confiance $1 - \alpha$ est une suite de domaines aléatoires $\Delta^{\alpha}(\mathbf{x}_n)$ tel que

$$\lim_{n \to \infty} P[\Delta^{\alpha}(\mathbf{x}_n) \ni \theta] = 1 - \alpha. \tag{1.13}$$

En pratique, on aboutit à l'approximation suivante :

$$P[\Delta^{\alpha}(\mathbf{x}_n) \ni \theta] \approx 1 - \alpha \quad \text{pour} \quad n \gg 1.$$

Cet intervalle ou région de confiance asymptotique pourra être obtenu à l'aide d'une fonction dite asymptotiquement pivotale définie par une suite de fonctions $h_n(\mathbf{x}_n, \theta)$

$$(\mathbf{x}_n, \theta) \in \mathbb{R}^n \times \Theta \longmapsto h_n(\mathbf{x}_n, \theta) \in \mathbb{R}$$

qui converge en loi vers une loi de probabilité qui ne dépend pas de θ et est non dégénérée (en pratique à densité de probabilité).

1.5.3 Intervalles de confiance asymptotiques pour une proportion

L'application du théorème central limite nous donne la fonction asymptotiquement pivotale suivante

$$U_n \stackrel{\text{def}}{=} \frac{\bar{X}_n - \mathrm{E}(\bar{X}_n)}{\sqrt{\mathrm{var}(\bar{X}_n)}} = \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \text{ converge en loi vers la loi de probabilité gaussienne } \mathcal{N}(0,1). \quad (1.14)$$

Par suite si u_α désigne la valeur telle que $P(|U| \geq u_\alpha) = \alpha$ où U est de loi gaussienne $\mathcal{N}(0,1)$, nous avons :

$$\lim_{n\to\infty} P\left[\left|\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right| < u_\alpha\right] = 1 - \alpha.$$

Or

$$\left|\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq u_\alpha \Leftrightarrow \frac{(\bar{X}_n - p)^2}{\frac{p(1-p)}{n}} \leq u_\alpha^2 \Leftrightarrow p^2(n + u_\alpha^2) - p(2n\bar{X}_n + u_\alpha^2) + n\bar{X}_n^2 < 0.$$

Ce binome en p sera négatif entre ses racines :

$$\frac{\bar{X}_n + \frac{u_\alpha^2}{2n} \pm u_\alpha\sqrt{\frac{u_\alpha^2}{4n^2} + \frac{\bar{X}_n(1-\bar{X}_n)}{n}}}{1 + \frac{u_\alpha^2}{n}}.$$

Par suite, nous en déduisons l'intervalle de confiance asymptotique de niveau de confiance $1 - \alpha$:

$$p \in \left[\frac{\bar{X}_n + \frac{u_\alpha^2}{2n} - \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ u_\alpha}{1 + \frac{u_\alpha^2}{n}}, \frac{\bar{X}_n + \frac{u_\alpha^2}{2n} + \sqrt{\frac{u_\alpha^2}{4n^2} + \frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ u_\alpha}{1 + \frac{u_\alpha^2}{n}}\right]. \quad (1.15)$$

Pour les valeurs usuelles de α et pour les grandes valeurs de n nous avons $u_\alpha^2/n \ll 1$. De plus rappelons que l'approximation issue de la convergence en loi (1.14) est valide dans les conditions

$$n > 30, np > 5 \text{ et } n(1-p) > 5$$

Sous toutes ces conditions, l'*intervalle de confiance asymptotique approché* donné par (1.15) donne l'intervalle de confiance approché de degré de confiance $1 - \alpha$ classique

$$p \in \left[\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ u_\alpha, \bar{X}_n + \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\ u_\alpha\right]. \quad (1.16)$$

Remarque 1.9 : Cet intervalle de confiance (1.16) symétrique par rapport à l'estimateur ponctuel $\widehat{p}(\mathbf{x}_n) = \bar{X}_n$ de p a une demi largeur de $\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}u_\alpha$ et une demi largeur relative de

$$\sqrt{\frac{(1-\bar{X}_n)}{n\bar{X}_n}}u_\alpha, \quad (1.17)$$

qu'on peut ainsi appeler respectivement *précision absolue* et *précision relative* de l'estimateur ponctuel $\bar{X}_n$. Nous voyons en particulier que la précision relative peut être très mauvaise pour p proche de 0 si n n'est pas assez élevé.

Nous retrouvons ici les remarques générales précédentes au sujet du compromis entre degré de confiance et précision ainsi que de l'amélioration apportée par l'augmentation de la taille n de l'échantillon. Ici il faut multiplier n par 4 pour voir les précisions absolues et relatives s'améliorer d'un facteur 2.

Exemple 1.4 : Une valeur typique de n de l'ordre de 1000 est utilisée dans les sondages d'opinion binaires. Dans ce cas pour la valeur classique d'un degré de confiance de $1-\alpha = 95\%$ associé à $u_\alpha \approx 2$ et pour p proche de $1/2$, nous obtenons à partir de (1.16), l'intervalle de confiance approché suivant :

$$P\{[\bar{X}_n - 0,032, \bar{X}_n + 0,032] \ni p\} \approx 95\%.$$

Ainsi pour un tel effectif et un tel degré de confiance, on ne peut rien conclure quant au résultat d'une élection pour un échantillon qui fournirait $\bar{X}_n \in [0,468; 0,532]$.

En pratique, les sondages d'opinion, qu'ils soient à vocation politique ou marketing n'utilisent pas un échantillon prélevé au hasard dans la population globale. Ils utilisent une méthode dite des *quotas*, qui tient compte d'une connaissance a priori sur l'échantillon. Cette méthode est moins coûteuse et de plus, permet de diminuer quelque peu la variance de la fréquence relative $\bar{X}_n$; ce qui améliore la précision de l'estimation de p. Nous n'aborderons pas cet aspect car sa modélisation est beaucoup plus longue.

Exemple 1.5 : En communications numériques, on cherche à estimer une probabilité d'erreur p à l'aide d'un taux d'erreur $\bar{X}_n$ mesuré par l'observation de n symboles. Puisque dans ce cas le paramètre à estimer vérifie $p \ll 1$, on a une précision relative sur p pour un degré de confiance $1-\alpha = 95\%$ associé à $u_\alpha \approx 2$ donnée par (1.17) de

$$\sqrt{\frac{(1-\bar{X}_n)}{n\bar{X}_n}} u_\alpha \approx \frac{2}{\sqrt{n\bar{X}_n}}.$$

or $n\bar{X}_n$ représente le nombre d'erreurs observées sur n symboles.

Il faudra donc par exemple attendre l'apparition de 400 erreurs pour avoir une précision relative de 10% seulement ! C'est suffisant en pratique, car seul l'ordre de grandeur de la probabilité d'erreur est utile. Exiger une précision relative de 1% avec ce même degré de confiance de 95% nécessiterait d'attendre l'apparition de 4.10^4 erreurs (soit pour des probabilités d'erreur de l'ordre de 10^{-7} des temps d'observation importants malgré des débits de l'ordre de 10^9 bits/s) !

Remarque 1.10 : Nous pouvons démontrer que l'intervalle de confiance (1.16) est en fait un intervalle de confiance asymptotique exact au sens de la définition (1.13) en utilisant les résultats suivants de probabilité.

Résultat de probabilités 1.4 :

1. *Théorème de Slutsky* : Soit $(U_n)_{n=1,..}$, une suite de variables aléatoires qui converge en loi vers U et $(V_n)_{n=1,..}$ une suite de variables aléatoires qui converge en probabilité vers une constante c. Alors la suite de variables aléatoires $(U_n, V_n)_{n=1,..}$ converge en loi vers la variable aléatoire (U, c). De plus pour toute fonction continue $u, v \mapsto g(u, v)$, en appliquant le résultat de probabilités 1.4.2 (c) ci dessous, la suite de variables aléatoires $g(U_n, V_n)$ converge en loi vers la variable aléatoire $g(U, c)$.
2. *Théorèmes de continuité* : Pour toute fonction continue $x \mapsto g(x)$ et toute suite $(X_n)_{n=1,..}$ de variables aléatoires, nous avons les implications :
   (a) Si X_n converge presque sûrement vers X, alors $g(X_n)$ converge presque sûrement vers $g(X)$,
   (b) Si X_n converge en probabilité vers X, alors $g(X_n)$ converge en probabilité vers $g(X)$,
   (c) Si X_n converge en loi vers X, alors $g(X_n)$ converge en loi vers $g(X)$.

   La condition de continuité de la fonction $x \mapsto g(x)$ est inutilement forte. Si l'ensemble des points de discontinuité de $g(.)$ est noté D_g et si $P(X \in D_g) = 0$, alors la suite de variables aléatoires $g(X_n)$ hérite du même mode de convergence de la suite X_n. Ainsi par exemple, si X est presque sûrement non nulle, alors la suite de variables aléatoires $1/X_n$ converge vers $1/X$ de la même manière.

3. Si la suite déterministe $c_n > 0$ vérifie $\lim_{n\to\infty} c_n = +\infty$ tel que

$$c_n(X_n - a) \text{ converge en loi vers } X,$$

alors la suite de variables aléatoires X_n converge en probabilité vers a.

Par application de la loi faible des grands nombres, la suite de variables aléatoires $\bar{X}_n$ converge en probabilité vers $\mathrm{E}(\bar{X}_n) = p$. Par suite grâce à la propriété précédente 2(b),

$$\frac{\sqrt{p(1-p)}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \to 1 \text{ en probabilité.} \tag{1.18}$$

Par suite par application du théorème de Slutsky avec $g(u,v) = uv$ où

$$U_n = \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \text{ et } V_n = \frac{\sqrt{p(1-p)}}{\sqrt{\bar{X}_n(1-\bar{X}_n)}},$$

nous avons d'après le théorème central limite (1.14) : U_n converge en loi vers U de loi $\mathcal{N}(0,1)$ et par ailleurs V_n converge en probabilité vers la constante 1 d'après (1.18). Par suite :

$$\frac{\bar{X}_n - p}{\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}} \text{ converge en loi vers la loi gaussienne } \mathcal{N}(0,1). \tag{1.19}$$

et l'intervalle de confiance (1.16) est un intervalle de confiance asymptotique exact de degré de confiance $1-\alpha$.

1.5.4 Intervalles de confiance asymptotiques pour la moyenne et la variance d'un échantillon

Nous allons démontrer ici qu'il est possible d'exhiber des fonctions pivotales asymptotiques et donc des intervalles de confiance asymptotiques dans le cadre d'observations $\mathbf{x}_n$ constituées de n variables aléatoires $x_1, ..., x_n$ indépendantes équidistribuées de moyenne $m = \mathrm{E}(x_k)$ et de variance $\sigma^2 = \mathrm{var}(x_k)$ sans connaître la loi de probabilité de cette distribution commune. Nous sommes donc ici dans un cadre semi-paramétrique où le paramètre d'intérêt sera soit m, soit σ^2.

En effet par application du théorème central limite à

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} x_k \text{ et } S"^2_n = \frac{1}{n}\sum_{k=1}^{n}(x_k - m)^2$$

dans lequel nous supposons de plus que $\mathrm{E}(x_k^4)$ existe, nous obtenons respectivement :

$$\frac{\bar{X}_n - \mathrm{E}(\bar{X}_n)}{\sqrt{\mathrm{var}(\bar{X}_n)}} \text{ et } \frac{S"^2_n - \mathrm{E}(S"^2_n)}{\sqrt{\mathrm{var}(S"^2_n)}} \text{ convergent en loi vers la loi gaussienne } \mathcal{N}(0,1) \tag{1.20}$$

avec

$$\frac{\bar{X}_n - \mathrm{E}(\bar{X}_n)}{\sqrt{\mathrm{var}(\bar{X}_n)}} = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \text{ et } \frac{S"^2_n - \mathrm{E}(S"^2_n)}{\sqrt{\mathrm{var}(S"^2_n)}} = \frac{S"^2_n - \sigma^2}{\sqrt{\frac{\mu_4 - \sigma^4}{n}}} \text{ avec } \mu_4 \stackrel{\text{def}}{=} \mathrm{E}[(x_k - m)^4].$$

Puis par application du théorème de Slutsky avec

$$g(u,v) = uv \quad \text{où} \quad U_n = \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \quad \text{et} \quad V_n = \frac{\sigma}{S_n}$$

et de la propriété 2(b) sachant que $S_n^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{X}_n)^2$ converge en probabilité vers la constante σ^2, nous obtenons que

$$\frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \quad \text{converge en loi vers la loi gaussienne } \mathcal{N}(0,1). \tag{1.21}$$

De même par plusieurs applications du théorème de Slutsky et de la propriété 2(b) sachant que $\mu_4(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^4$ converge en probabilité vers la constante μ_4, on obtiendrait aussi :

$$\frac{S'^2_n - \sigma^2}{\sqrt{\frac{\mu_4(\mathbf{x}) - S'^4_n}{n}}} \quad \text{converge en loi vers la loi gaussienne } \mathcal{N}(0,1) \tag{1.22}$$

avec $S'^2_n \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{X}_n)^2$. Par suite les deux fonctions asymptotiquement pivotales (1.21) et (1.22) permettent d'en déduire les deux intervalles de confiance asymptotiques suivants avec degré de confiance $1-\alpha$:

$$m \in \left[\bar{X}_n - \frac{S_n}{\sqrt{n}}\, u_\alpha, \bar{X}_n + \frac{S_n}{\sqrt{n}}\, u_\alpha\right], \tag{1.23}$$

$$\sigma^2 \in \left[S'^2_n - \frac{\sqrt{\mu_4(\mathbf{x}_n) - S'^4_n}}{\sqrt{n}}\, u_\alpha, S'^2_n + \frac{\sqrt{\mu_4(\mathbf{x}_n) - S'^4_n}}{\sqrt{n}}\, u_\alpha\right], \tag{1.24}$$

dans lequels u_α désigne la valeur telle que $P(|U| \geq u_\alpha) = \alpha$ où U est de loi gaussienne $\mathcal{N}(0,1)$.

1.6 Régions et intervalles de confiance asymptotiques construits à partir d'une suite d'estimateurs ponctuels

1.6.1 Estimateurs asymptotiquement gaussiens

Nous verrons dans les chapitres 2 et 3 que de très nombreuses suites d'estimateurs associés à des observations $\mathbf{x}$ de taille n (que nous noterons $\mathbf{x}_n$), $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ de $\boldsymbol{\theta} \in \mathbb{R}^q$ issus de la *méthode du maximum de vraisemblance* ou de la *méthode des moments* satisfont la propriété asymptotique suivante :

Il existe une suite $c_n > 0$, telle que $\lim_{n\to\infty} c_n = +\infty$, appelé *vitesse de convergence* (qui est dans la plupart des situations $c_n = \sqrt{n}$), telle que

$$c_n\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}\right) \quad \text{converge en loi vers la loi gaussienne } \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})). \tag{1.25}$$

où $\boldsymbol{\theta} \mapsto \boldsymbol{\Sigma}(\boldsymbol{\theta})$ est une fonction continue et où la matrice $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ dite *matrice de covariance de la loi asymptotique* de la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$, est inversible pour $\boldsymbol{\theta} \in \Theta$.

Cette relation (1.25) implique d'après le rappel de probabilité 1.4.3, que la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ est une suite d'estimateurs faiblement consistante de $\boldsymbol{\theta}$. Par suite en appliquant le résultat de probabilité 1.4.2b, la suite matricielle $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n))$ sera une suite de variables matricielles aléatoires qui converge en probabilité vers la matrice constante $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. (appelée suite d'estimateurs faiblement consistante de $\boldsymbol{\Sigma}(\boldsymbol{\theta})$). Cette dernière propriété va nous permettre de construire une fonction asymptotiquement pivotale et par suite des intervalles ou régions de confiance asymptotique pour le paramètre $\boldsymbol{\theta}$.

1.6.2 Intervalles ou régions de confiance asymptotique

Distinguons pour cela les cas scalaires et multidimensionnels pour $\boldsymbol{\theta}$. Dans le cas scalaire, la suite de variables aléatoires $\frac{c_n(\widehat{\theta}_n(\mathbf{x}_n)-\theta)}{\sigma(\theta)}$, (avec $\sigma^2(\theta) \stackrel{\text{def}}{=} \boldsymbol{\Sigma}(\boldsymbol{\theta})$) converge en loi vers la loi de probabilité $\mathcal{N}(0,1)$. Puis par application du théorème de Slutsky avec

$$g(u,v) = uv \quad \text{où} \quad U_n = \frac{c_n(\widehat{\theta}_n(\mathbf{x}_n) - \theta)}{\sigma(\theta)} \quad \text{et} \quad V_n = \frac{\sigma(\theta)}{\sigma(\widehat{\theta}_n(\mathbf{x}_n))}$$

où $\sigma(\widehat{\theta}_n(\mathbf{x}_n))$ joue le rôle de $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n))$ dans le cas scalaire. Nous en déduisons

$$\frac{c_n(\widehat{\theta}_n(\mathbf{x}_n) - \theta)}{\sigma(\widehat{\theta}(\mathbf{x}_n))} \quad \text{converge en loi vers la loi gaussienne } \mathcal{N}(0,1). \tag{1.26}$$

Cette fonction asymptotiquement pivotale $\frac{c_n(\widehat{\theta}_n(\mathbf{x}_n)-\theta)}{\sigma(\widehat{\theta}_n(\mathbf{x}_n))}$ nous permet alors de donner l'intervalle de confiance asymptotique de niveau de confiance $1-\alpha$:

$$\theta \in \left[\widehat{\theta}_n(\mathbf{x}_n) - \frac{\sigma(\widehat{\theta}_n(\mathbf{x}_n))}{c_n}\ u_\alpha, \widehat{\theta}_n(\mathbf{x}_n) + \frac{\sigma(\widehat{\theta}_n(\mathbf{x}_n))}{c_n}\ u_\alpha \right]. \tag{1.27}$$

où u_α désigne la valeur telle que $P(|U| \geq u_\alpha) = \alpha$ où U est de loi gaussienne $\mathcal{N}(0,1)$.

Dans le cas multidimensionnel $\boldsymbol{\theta} \in \mathbb{R}^q$, nous avons besoin du résultat de probabilité suivant :

Résultat de probabilités 1.5 : Si $\mathbf{x} \in \mathbb{R}^q$ est une variable aléatoire mutidimensionnelle de loi de probabilité $\mathcal{N}(\mathbf{0}, \mathbf{C})$ où $\mathbf{C}$ est inversible, alors la variable aléatoire scalaire $\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}$ est de loi de Khi-deux à q degrés de liberté.

Par application du résultat de probabilité 1.4.2b, qui implique que la suite de matrices aléatoires $\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n))$ converge en probabilité vers la matrice constante $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$, le théorème de Slutsky appliqué à

$$g(\mathbf{u}, \mathbf{V}) = \mathbf{u}^T\mathbf{V}\mathbf{u} \quad \text{où} \quad U_n = c_n(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}) \quad \text{et} \quad V_n = \boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)),$$

implique que

$$g(U_n, V_n) = c_n^2(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)) \left(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}\right)$$

converge en loi vers la variable aléatoire $g(U, \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) = U^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})U$ où U est de loi $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. Par suite d'après le résultat de probabilité 1.5, nous en déduisons :

$$c_n^2 \left(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}\right)^T \boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)) \left(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}\right) \quad \text{converge en loi vers la loi de Khi-deux à } q \text{ degrés de liberté.} \tag{1.28}$$

Cette fonction asymptotiquement pivotale donne la région de confiance asymptotique $\boldsymbol{\Delta}^\alpha(\mathbf{x}_n)$ suivante pour le degré de confiance $1-\alpha$.

$$\boldsymbol{\Delta}^\alpha(\mathbf{x}_n) = \{\boldsymbol{\theta} \in \mathbb{R}^q; c_n^2(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n))(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}) \leq z_{q,\alpha}\} \tag{1.29}$$

où $z_{q,\alpha}$ désigne la valeur telle que $P(Z \leq z_{q,\alpha}) = 1-\alpha$ avec Z variable aléatoire de loi de Khi-deux à q degrés de liberté. Notons que $\boldsymbol{\Delta}^\alpha(\mathbf{x}_n)$ est ici un ellipsoïde centré sur l'estimateur ponctuel $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ de $\boldsymbol{\theta}$ appelé *ellipsoïde asymptotique de confiance.*

1.6.3 Ellipsoïde asymptotique de confiance

Le volume de l'ellipsoïde asymptotique de confiance $\mathbf{\Delta}^{\alpha}(\mathbf{x}_n) \in \mathbb{R}^q$ (1.29) est donné par

$$\text{vol}[\mathbf{\Delta}^{\alpha}(\mathbf{x}_n)] = \frac{\pi^{q/2}[c_n^{-2} z_{q,\alpha}]^{q/2} \det^{1/2}(\mathbf{\Sigma}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)))}{\Gamma(\frac{q}{2}+1)}$$

et par suite : vol$[\mathbf{\Delta}^{\alpha}(\mathbf{x}_n)]$ est équivalent en probabilité [1] à $\frac{\pi^{q/2}[c_n^{-2} z_{q,\alpha}]^{q/2} \det^{1/2}(\mathbf{\Sigma}(\boldsymbol{\theta}))}{\Gamma(\frac{q}{2}+1)}$.

Si deux estimateurs ponctuels $\widehat{\boldsymbol{\theta}}_n^{(1)}(\mathbf{x}_n)$ et $\widehat{\boldsymbol{\theta}}_n^{(2)}(\mathbf{x}_n)$ sont asymptotiquement gaussiens de même vitesse de convergence c_n, leur comparaison, se ramène alors à la comparaison des matrices de covariance $\mathbf{\Sigma}^{(1)}(\boldsymbol{\theta})$ et $\mathbf{\Sigma}^{(2)}(\boldsymbol{\theta})$ des lois asymptotiques associées.

En particulier si $\det[\mathbf{\Sigma}^{(1)}(\boldsymbol{\theta})] \leq \det[\mathbf{\Sigma}^{(2)}(\boldsymbol{\theta})]$, le volume de l'ellipsoïde asymptotique de confiance associé à la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n^{(1)}(\mathbf{x}_n)$ est plus petit que celui associé à la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n^{(2)}(\mathbf{x}_n)$ et l'on préférera donc l'ellipsoïde asymptotique de confiance associé à la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n^{(1)}(\mathbf{x}_n)$.

Notons que si nous avions la condition plus forte $\mathbf{\Sigma}^{(1)}(\boldsymbol{\theta}) \leq \mathbf{\Sigma}^{(2)}(\boldsymbol{\theta})$, nous ne pouvons pas affirmer que l'ellipsoïde asymptotique de confiance associé à la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n^{(1)}(\mathbf{x}_n)$ soit contenu dans celui associé à la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n^{(2)}(\mathbf{x}_n)$ car ces deux ellipsoïdes sont centrés sur des points $\widehat{\boldsymbol{\theta}}_n^{(1)}(\mathbf{x}_n)$ et $\widehat{\boldsymbol{\theta}}_n^{(2)}(\mathbf{x}_n)$ différents.

1.6.4 Cas particuliers de l'estimateur du maximum de vraisemblance

Lorsque $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ est un estimateur du maximum de vraisemblance et que l'observation $\mathbf{x}_n$ est constituée de n variables aléatoires $x_1, ...x_n$ indépendantes et équidistribuées, sous certaines conditions techniques (vérifiées par exemple par toutes les lois de probabilité de la famille exponentielle (voir paragraphe 2.3.7)), la convergence en loi (1.25) est assurée avec la vitesse de convergence $c_n = \sqrt{n}$ et la matrice de convariance $\mathbf{\Sigma}(\boldsymbol{\theta}) = \mathbf{I}_1^{-1}(\boldsymbol{\theta})$:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}) \text{ converge en loi vers la loi gaussienne } \mathcal{N}(\mathbf{0}, \mathbf{I}_1^{-1}(\boldsymbol{\theta})). \tag{1.30}$$

où $\mathbf{I}_1(\boldsymbol{\theta})$ désigne la *matrice d'information de Fisher* associée à une seule observation x_i sur le paramètre $\boldsymbol{\theta}$. On dit alors que l'*estimateur* $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ *est asymptotiquement gaussien et efficace*. L'intervalle de confiance asymptotique (1.27) et l'ellispoïde de confiance asymptotique (1.29) deviennent respectivement :

$$\theta \in \left[\widehat{\theta}_n(\mathbf{x}_n) - \frac{\sqrt{I_1^{-1}(\widehat{\theta}_n(\mathbf{x}_n))}}{\sqrt{n}}\, u_\alpha, \widehat{\theta}_n(\mathbf{x}_n) + \frac{\sqrt{I_1^{-1}(\widehat{\theta}_n(\mathbf{x}_n))}}{\sqrt{n}}\, u_\alpha \right] \tag{1.31}$$

et

$$\mathbf{\Delta}^{\alpha}(\mathbf{x}_n) = \{\boldsymbol{\theta} \in \mathbb{R}^q; n(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta})^T \mathbf{I}_1(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n))(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}) \leq z_{q,\alpha}\}. \tag{1.32}$$

Exemple 1.6 : Revenons au cas de l'estimation d'une proportion p à l'aide d'un échantillon $(x_1, ..., x_n)$ de variables aléatoires indépendantes de loi de Bernoulli de paramètre p. Nous savons que les estimateurs issus de la méthode du maximum de vraisemblance et de la méthode des moments coincident et sont donnés par $\widehat{\boldsymbol{\theta}}_n = \bar{X}_n$. Cet estimateur est sans biais, efficace et asymptotiquement gaussien d'information de Fisher associée à X_i donnée par $I_1(p) = \frac{1}{p(1-p)}$. Par suite l'intervalle de confiance asymptotique (1.31)

1. Deux suites de variables alatoires U_n et V_n sont équivalentes en probabilité si la suite U_n/V_n converge en probabilité vers 1.

redonne l'intervalle de confiance asymptotique précédent (1.16).

Exemple 1.7 : Revenons à l'échantillon gaussien de la section 1.2 où $\boldsymbol{\theta} = \begin{bmatrix} m \\ \sigma^2 \end{bmatrix}$. Il est aisé de démontrer que l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ est ici donné par $\begin{bmatrix} \bar{X}_n \\ S_n^2 \end{bmatrix}$ et que la matrice d'information de Fisher associée est ici $\mathbf{I}_1(\boldsymbol{\theta}) = \frac{1}{\sigma^2} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2\sigma^2} \end{bmatrix}$. Par suite un estimateur consistant de $\mathbf{I}_1(\boldsymbol{\theta})$ est donné par $\mathbf{I}_1(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)) = \frac{1}{S_n^2} \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2S_n^2} \end{bmatrix}$ et donc l'ellipsoïde asymptotique de confiance (1.32) pour le paramètre $\boldsymbol{\theta}$ est ci donné par

$$\boldsymbol{\Delta}^{\alpha}(\mathbf{x}_n) = \{\boldsymbol{\theta} \in \mathbb{R}^2; \frac{n}{S_n^2}(\bar{X}_n - m)^2 + \frac{n}{2S_n^4}(S_n^2 - \sigma^2)^2 \leq z_{2,\alpha}\}. \tag{1.33}$$

Exemple 1.8 : L'estimateur du maximum de vraisemblance ne vérifie pas nécessairement la convergence en loi (1.25) comme dans l'exemple d'un échantillon de variables aléatoires $x_1, ...x_n$ indépendantes et équidistribuées de loi uniforme sur $[0, \theta]$. En effet dans ce cas l'estimateur du maximum de vraisemblance est donné par

$$\widehat{\theta}_n(\mathbf{x}_n) = \sup(x_1, ..., x_n)$$

qui satisfait $\widehat{\theta}_n(\mathbf{x}_n) \leq \theta$. Dans ce cas élémentaire, la distribution exacte de l'estimateur $\widehat{\theta}_n(\mathbf{x}_n)$ est disponible car

$$P(\sup(x_1, ..., x_n) \leq u) = P(\cap_{k=1}^n (x_k \leq u) = \prod_{k=1}^n P(x_k \leq u).$$

Sa fonction de répartition est donnée par

$$F_{\widehat{\theta}_n(\mathbf{x}_n)}(u) = P(\widehat{\theta}_n(\mathbf{x}_n) \leq u) = \left(\frac{u}{\theta}\right)^n \mathbb{1}_{[0,\theta]}(u) + \mathbb{1}_{]\theta,+\infty[}(u). \tag{1.34}$$

Par suite, cherchons une suite c_n telle que $c_n(\widehat{\theta}_n(\mathbf{x}_n) - \theta)$ converge en loi vers une variable aléatoire non dégénérée. A partir de (1.34), on obtient :

$$\begin{aligned} P\left(c_n(\widehat{\theta}_n(\mathbf{x}_n) - \theta) \leq y\right) &= P\left(\widehat{\theta}_n(\mathbf{x}_n) \leq \theta + \frac{y}{c_n}\right) \\ &= \left(1 + \frac{y}{\theta c_n}\right)^n \mathbb{1}_{[-\theta c_n, 0]}(y) + \mathbb{1}_{]0,+\infty[}(y) \end{aligned}$$

et nous voyons que la seule valeur de la suite c_n telle que $\lim_{n\to\infty} \left(1 + \frac{y}{\theta c_n}\right)^n$ existe et soit différente de 0 et 1 est la valeur $c_n = n$ pour laquelle $\lim_{n\to\infty} \left(1 + \frac{y}{n\theta}\right)^n = e^{\frac{y}{\theta}}$ et nous obtenons donc :

$$n(\widehat{\theta}_n(\mathbf{x}_n) - \theta) \text{ converge en loi vers la loi de fonction de répartition } e^{\frac{y}{\theta}} \mathbb{1}_{]-\infty,0]}(y) + \mathbb{1}_{]0,+\infty[}(y), \tag{1.35}$$

c'est à dire de loi de probabilité exponentielle négative de paramètre $1/\theta$ de densité de probabilité

$$\frac{1}{\theta} e^{\frac{u}{\theta}} \mathbb{1}_{]-\infty,0[}(u).$$

Mais comme cette loi de probabilité limite dépend de θ, $n(\widehat{\theta}_n(\mathbf{x}_n) - \theta)$ n'est pas une fonction asymptoti-

quement pivotale, cependant le résultat de probabilité 1.4(2-c) nous permet d'affirmer que la fonction

$$\frac{n}{\theta}\left(\widehat{\theta}_n(\mathbf{x}_n)-\theta\right)$$

est elle, une fonction asymptotiquement pivotale car sa distribution limite est une loi de probabilité exponentielle négative de paramètre 1 (loi de Weibull de paramètre 1). Il s'agit ici d'un cas particulier du *théorème de valeurs extrêmes*. Par suite en particulier :

$$\lim_{n\to\infty} P\left[-a \leq \frac{n}{\theta}(\widehat{\theta}_n(\mathbf{x}_n)-\theta) \leq 0\right] = 1-e^{-a},$$

soit après inversion puisque

$$P[-a \leq \frac{n}{\theta}(\widehat{\theta}_n(\mathbf{x}_n)-\theta) \leq 0] = P[\widehat{\theta}_n(\mathbf{x}_n) \leq \theta \leq \widehat{\theta}_n(\mathbf{x}_n)\left(1-\frac{a}{n}\right)^{-1}],$$

nous obtenons l'intervalle de confiance asymptotique de niveau de confiance $1-\alpha$:

$$\theta \in \left[\widehat{\theta}_n(\mathbf{x}_n), \widehat{\theta}_n(\mathbf{x}_n)\left(1+\frac{\ln\alpha}{n}\right)^{-1}\right]. \tag{1.36}$$

Notons que contrairement aux cas habituels, cet intervalle de confiance asymptotique est monolatéral, car ici $\widehat{\theta}_n(\mathbf{x}_n) \leq \theta$.

Si l'on était parti de l'estimation $\widehat{\theta}'_n(\mathbf{x}_n) = 2\bar{X}_n$ issu de la méthode des moments, nous obtiendrions par le théorème centrale limite avec $\mathrm{E}(x_i) = \frac{\theta}{2}$ et $\mathrm{var}(x_i) = \frac{\theta^2}{12}$:

$$\frac{\widehat{\theta}'_n(\mathbf{x}_n)-\theta}{\frac{\theta}{\sqrt{3n}}} \text{ converge en loi vers la loi gaussienne } \mathcal{N}(0,1).$$

Ce qui nous donnerait, en appliquant à nouveau le théorème de Slutsky du résultat de probabilité 4.4.1, le nouvel intervalle de confiance asymptotique de niveau de confiance $1-\alpha$:

$$\theta \in \left[\widehat{\theta}'_n(\mathbf{x}_n) - \frac{\widehat{\theta}'_n(\mathbf{x}_n)}{\sqrt{3n}}\ u_\alpha, \widehat{\theta}'_n(\mathbf{x}_n) + \frac{\widehat{\theta}'_n(\mathbf{x}_n)}{\sqrt{3n}}\ u_\alpha\right].$$

C'est à dire :

$$\theta \in \left[2\bar{X}_n - \frac{2\bar{X}_n}{\sqrt{3n}}\ u_\alpha, 2\bar{X}_n + \frac{2\bar{X}_n}{\sqrt{3n}}\ u_\alpha\right]. \tag{1.37}$$

Notons enfin qu'en appliquant le résultat (1.23) obtenu dans le cadre semi paramétrique où $\theta = 2\mathrm{E}(X_k) = 2m$, on obtiendrait un intervalle de confiance asymptotique de niveau de confiance $1-\alpha$ différent du précédent [2] :

$$\theta \in \left[2\bar{X}_n - \frac{2S_n}{\sqrt{n}}\ u_\alpha, 2\bar{X}_n + \frac{2S_n}{\sqrt{n}}\ u_\alpha\right]. \tag{1.38}$$

Remarque 1.11 : L'exemple 1.8 est exceptionnel car l'estimateur du maximum de vraisemblance permet ici d'accéder à un intervalle de confiance non asymptotique grâce à sa fonction de répartition, sans utilisation de fonction pivotale. En effet d'après la relation (1.34), pour $u \in [0,\theta]$:

$$P[\widehat{\theta}_n(\mathbf{x}_n) \leq u] = \left(\frac{u}{\theta}\right)^n = \alpha$$

2. Ceci est un exemple de la non unicité des intervalles de confiance asymptotique (ou non asymptotique) qui ici centrés sur la même valeur $\widehat{\theta}'_n(\mathbf{x}_n)$ ont des largeurs différentes pour chaque valeur de l'observation $\mathbf{x}_n$ tout en ayant asymptotiquement la même probabilité de contenir le paramètre θ.

implique pour tout que $\alpha \in]0,1[$, $P[\widehat{\theta}_n(\mathbf{x}_n) \leq \theta\alpha^{1/n}] = \alpha \iff P[\theta\alpha^{1/n} \leq \widehat{\theta}_n(\mathbf{x}_n)] = 1-\alpha$ et donc puisque $\widehat{\theta}_n(\mathbf{x}_n) \leq \theta$

$$P[\widehat{\theta}_n(\mathbf{x}_n) \leq \theta \leq \widehat{\theta}_n(\mathbf{x}_n)\alpha^{-1/n}] = 1-\alpha.$$

Soit l'intervalle de confiance non asymptotique de niveau de confiance $1-\alpha$:

$$\theta \in [\widehat{\theta}_n(\mathbf{x}_n), \widehat{\theta}_n(\mathbf{x}_n)\alpha^{-1/n}]. \tag{1.39}$$

Remarque 1.12 : Il est important de noter qu'en pratique, les intervalles ou régions de confiance asymptotiques peuvent conduire à des approximations satisfaisantes

$$P[\Delta^{\alpha}(\mathbf{x}_n) \ni \theta] \approx 1-\alpha$$

pour des valeurs en général assez faibles de n (de l'ordre de 100) pour des observations $(x_1, ...x_n)$ scalaires. Par contre dans le cas d'observations $(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathbb{R}^r$ multidimensionnelles de dimension r élevée, ces approximations seront en général inutilisables car exigeant de trop grandes valeurs de n. Il faudra avoir alors recours à des intervalles ou régions de confiance asymptotiques en n et r sous la contrainte $n/r = c$ où c est une constante. C'est l'objet d'un cours de statistiques en grande dimension.

1.7 Points essentiels du chapitre estimateur par intervalle de confiance ou région de confiance

- La définition (1.1) d'un *intervalle* ou *région de confiance* et du *niveau de confiance* associé. Comprendre que cet intervalle ou région de confiance est aléatoire et n'est pas unique.
- La définition d'une *fonction pivotale* donnée en section 1.1.3. Comprendre qu'il n'y a pas de méthodes générales pour trouver de telles fonctions pivotales, qui n'existent que dans des cas particuliers (ex des paramètres m et σ^2 dans l'observation $(x_1, ..., x_n)$ où $(x_k)_{k=1,..,n}$ sont indépendant identiquement distribués (i.i.d.) de loi gaussienne $\mathcal{N}(m, \sigma^2)$.
- On peut construire un parallépipède de confiance en excès $\Delta^{\alpha}(\mathbf{x})$ de niveau de confiance $1-\alpha$ pour un paramètre $\boldsymbol{\theta} = (\theta_1, .., \theta_q)$ multidimensionnel à partir d'intervalles de confiance $\Delta_i^{\alpha_i}(\mathbf{x})$ sur chaque composante θ_i, avec possibilité d'ajuster les degrés de confiance $1-\alpha_i$ tels que $\sum_{i=1}^{q} \alpha_i = \alpha$.
- Dans le cas particulier d'une observation $(x_1, ..., x_n)$ où $(x_k)_{k=1,..,n}$ sont i.i.d. de loi de Bernoulli, il n'existe pas de fonction pivotale au sens strict (1.1), mais on peut exhiber une fonction asymptotiquement pivotale qui permet de déduire un intervalle de confiance approchée (1.16) qui est aussi un intervalle de confiance asymptotique.
- On peut en général construire à partir d'une suite d'estimateurs ponctuels $\widehat{\theta}_n(\mathbf{x}_n)$ qui satisfont la propriété asymptotique (1.25), en particulier issues de la méthode de maximum de vraisemblance ou de la méthode des moments, une fonction *asymptotiquement pivotale* et par suite des intervalles (1.27) ou région de confiance asymptotiques (1.29) ; cette dernière étant sous la forme de l'intérieur d'une ellipse est appelée *ellipsoïde asymptotique de confiance*.
- Dans le cas particulier d'une observation $(x_1, ..., x_n)$ où $(x_k)_{k=1,..,n}$ sont i.i.d. de loi gaussienne $\mathcal{N}(m, \sigma^2)$ et de la suite d'estimateurs ponctuels $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) = (\frac{1}{n}\sum_{k=1}^{n} x_k,\ \frac{1}{n}\sum_{k=1}^{n}[x_k - (\frac{1}{n}\sum_{k=1}^{n} x_k)]^2)$ issue de la méthode du maximum de vraisemblance, nous obtenons un ellipsoïde asymptotique de confiance pour le paramètre $\boldsymbol{\theta} = (m, \sigma^2)$.

2 Estimateur du maximum de vraisemblance (MV)

2.1 Introduction

L'estimateur du maximum de vraisemblance s'appuie sur l'intuition qu'il est fondé d'estimer θ par la valeur du paramètre qui maximise la probabilité de ce qui est observé. Cela conduit dans le cas d'un modèle discret (i.e., l'ensemble $\mathcal{X}$ des observations $\mathbf{x}$ est fini ou dénombrable), que si nous avons observé $\mathbf{X} = \mathbf{x}$, alors nous estimerons θ par la valeur du paramètre qui maximise la fonction $\theta \in \Theta \mapsto P(\mathbf{X} = \mathbf{x}; \theta)$. L'estimateur du maximum de vraisemblance $\widehat{\theta} = \arg\max_{\theta\in\Theta} P(\mathbf{X} = \mathbf{x}; \theta)$ représente dans ce cas élémentaire, l'estimateur du maximum de vraisemblance de θ.

Dans le cas d'un modèle continu, nous avons $P(\mathbf{X} = \mathbf{x}; \theta) = 0$ pour tout θ et tout $\mathbf{x}$. En considérant cette fois des intervalles (cas scalaire) ou des pavés (cas multidimensionnel) autour de $\mathbf{x}$, l'estimateur du maximum de vraisemblance de θ sera obtenu par la recherche du maximum de la fonction $\theta \in \Theta \mapsto f_{\mathbf{X}}(\mathbf{x}; \theta)$.

Nous supposerons ici que la loi de probabilité du modèle paramétrique $\mathcal{P} = \{P_\theta(\mathbf{x}), \theta \in \Theta\}$ est décrite par la densité de probabilité de l'observation $\mathbf{x} \in \mathcal{X}$ au sens de Radon Nikodym par rapport à une mesure μ qui ne dépend pas de θ. En pratique μ sera soit la mesure de comptage, soit la mesure de Lebesgue. On notera cette densité de probabilité par la notation commune $p(\mathbf{x}; \theta)$ qui jouera le rôle usuel des probabilités $P(\mathbf{X} = \mathbf{x}; \theta)$ et des densités de probabilité $f_{\mathbf{X}}(\mathbf{x}; \theta)$ dans les cas de loi de probabilité respectivement discrète ou continue.

Nous obtenons alors la définition générale suivante :

Définition : Pour tout modèle paramétrique $\mathcal{P} = \{P_\theta(\mathbf{x}), \theta \in \Theta\}$ décrit par la densité de probabilité $p(\mathbf{x}; \theta)$, pour toute observation $\mathbf{x}$ fixée, on appelle *vraisemblance* (et qui est notée $L_{\mathbf{x}}(\theta)$), la fonction $\theta \in \Theta \mapsto p(\mathbf{x}; \theta)$.

On appelle *estimateur du maximum de vraisemblance* (MV) de θ, tout estimateur $\widehat{\theta}_{\mathrm{MV}}(\mathbf{x})$ qui, pour chaque observation $\mathbf{x}$ est égal à un argument du maximum de la fonction $\theta \in \Theta \mapsto p(\mathbf{x}; \theta)$. Cet estimateur $\widehat{\theta}_{\mathrm{MV}}(\mathbf{x})$ est alors défini par : $\forall \mathbf{x} \in \mathcal{X}$ et $\forall \theta \in \Theta$:

$$p(\mathbf{x}; \widehat{\theta}_{\mathrm{MV}}(\mathbf{x})) \geq p(\mathbf{x}; \theta), \tag{2.1}$$

ou

$$\widehat{\theta}_{\mathrm{MV}}(\mathbf{x}) = \arg\max_{\theta\in\Theta} p(\mathbf{x}; \theta). \tag{2.2}$$

Comme de nombreuses lois de probabilité appartiennent à la famille exponentielle (2.8) et que le logarithme de la densité de probabilité intervient dans la définition de l'information de Fisher, nous utiliserons avantageusement la *log-vraisemblance* (notée $l_{\mathbf{x}}(\boldsymbol{\theta})$) définie par la fonction

$$\theta \in \Theta \mapsto \ln p(\mathbf{x}; \theta)$$

à la place de la vraisemblance, dont les maxima sont identiques par monotonie de la fonction logarithme.

Notons que la motivation intuitive de la méthode du maximum de vraisemblance ne suffit pas, bien entendu à justifier tout l'intérêt de cette méthode. Le but de ce chapitre est de montrer que bien que ne satisfaisant aucune condition d'optimalité (cet estimateur n'est pas nécessairement unique, de biais nul, efficace, de variance minimale, admissible ou MinMax,...), cet estimateur possède sous des conditions techniques assez générales de "bonnes" propriétés asymptotiques dans le cadre d'une séquence $\mathbf{x}_n = (x_1, ..., x_n)$ d'observations.

Remarque 2.1 : Cet estimateur du maximum de vraisemblance trouve son fondement théorique dans le cadre particulier où l'observation $\mathbf{x}$ est constituée d'observations $x_1, ..., x_n$ discrètes indépendantes de même loi de probabilité, comme l'estimateur qui minimise une "distance" entre une loi de probabilité

empirique $P_{\mathbf{x}}$: loi de probabilité d'une variable aléatoire discrète qui prend comme valeurs, les observations $x_1, ..., x_n$ avec les probabilités $1/n$, et la loi de probabilité paramétrée P_θ d'une variable aléatoire associée au modèle paramétrique $\{P_\theta(\mathbf{x}), \theta \in \Theta\}$ à composantes indépendantes de même loi de probabilité. Avant d'aller plus loin, cette "distance" appelée "distance" de Kullback nécessite d'être définie :

On appelle "distance" de Kullback entre les lois de probabilités P_1 et P_2, de densités de probabilité respectives f_1 et f_2 (au sens de Radon Nikodym) par rapport à une mesure commune μ :

$$d(P_1, P_2) \stackrel{\text{def}}{=} \int \ln\left(\frac{f_1(x)}{f_2(x)}\right) f_1(x) d\mu = -\int \ln[f_2(x)] f_1(x) d\mu + \int \ln[f_1(x)] f_1(x) d\mu.$$

Ce n'est pas une distance au sens usuel, car bien que l'on démontre que $d(P_1, P_2) \geq 0$ et $d(P_1, P_2) = 0 \Leftrightarrow P_1 = P_2$, ce n'est pas une fonction symétrique de P_1 et P_2. Mais cette "distance" appelée aussi *dissemblance de Kullback*, caractérise de façon pertinente l'écart ou la dissemblance entre P_1 et P_2.

Avec cette "distance" où ici μ est la mesure de comptage

$$d(P_{\mathbf{x}}, P_\theta) = -\int \ln[p_\theta(x)] P_{\mathbf{x}} d\mu + \int \ln[P_{\mathbf{x}}] P_{\mathbf{x}} d\mu.$$

La minimisation de $d(P_{\mathbf{x}}, P_\theta)$ par rapport à θ est équivalente à la maximisation par rapport à θ de

$$\int \ln[p_\theta(x)] P_{\mathbf{x}} d\mu = \frac{1}{n} \sum_{k=1}^{n} \ln[p_\theta(x_k)] = \frac{1}{n} \ln\left[\prod_{k=1}^{n} p_\theta(x_k)\right] = \frac{1}{n} \ln[p(\mathbf{x}; \theta)].$$

2.2 Exemple élémentaire

Considérons une observation $\mathbf{x}$ constituée de n variables aléatoires $(x_1, ..., x_n)$ indépendantes de loi de Bernoulli de paramètre inconnu θ. Nous avons :

$$p(\mathbf{x}; \theta) = \prod_{k=1}^{n} p(x_k; \theta) = \prod_{k=1}^{n} \theta^{x_k} (1-\theta)^{1-x_k} = \theta^{\sum_{k=1}^n x_k} (1-\theta)^{n - \sum_{k=1}^n x_k}.$$

Par suite

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = \frac{\sum_{k=1}^n x_k}{\theta} - \frac{n - \sum_{k=1}^n x_k}{1-\theta} = 0 \quad \text{pour} \quad \theta = \frac{\sum_{k=1}^n x_k}{n}$$

et pour cette valeur de θ, nous avons :

$$\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}\Big|_{\theta = \frac{\sum_{k=1}^n x_k}{n}} = -\frac{\sum_{k=1}^n x_k}{\theta^2} + \frac{n - \sum_{k=1}^n x_k}{(1-\theta)^2}\Big|_{\theta = \frac{\sum_{k=1}^n x_k}{n}} = -n^2 \left(\frac{1}{\sum_{k=1}^n x_k} + \frac{1}{n - \sum_{k=1}^n x_k}\right) < 0.$$

Par suite

$$\widehat{\theta}_{\text{MV}}(\mathbf{x}) = \frac{\sum_{k=1}^n x_k}{n}$$

est maximum unique de la vraisemblance $p(\mathbf{x}; \theta)$. Il s'agit de l'estimateur du maximum de vraisemblance du paramètre θ.

2.3 Propriétés élémentaires

2.3.1 Existence, contraintes

Il peut arriver que l'équation (2.1) n'ait pas de solution dans Θ. Cette situation est quelque peu pathologique. Si la vraisemblance $\boldsymbol{\theta} \in \Theta \mapsto p(\mathbf{x}; \boldsymbol{\theta})$ est une fonction suffisamment régulière de $\boldsymbol{\theta}$ et si le maximum de cette fonction est atteint en un point intérieur de Θ et non pas sur son bord, alors

l'estimateur du maximum de vraisemblance est une des solutions des *équations de vraisemblance* :

$$\frac{\partial p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_k} = 0 \text{ pour } k = 1, ..., q \text{ et } \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q. \tag{2.3}$$

Notons qu'il est quelquefois possible d'obtenir l'estimateur du maximum de vraisemblance sans passer par les équations de vraisemblance comme le montrent les exemples 2.1 et 2.2.

Exemple 2.1 : Considérons le cas où l'observation $\mathbf{x}$ est un échantillon de variables aléatoires $(x_k)_{k=1,...,n}$ indépendantes identiquement distribuées de loi uniforme sur $[0, \theta]$. Dans ce cas,

$$p(\mathbf{x};\theta) = \frac{1}{\theta^n} \prod_{k=1}^{n} \mathbb{1}_{[0,\theta]}(x_k) = \frac{1}{\theta^n} \mathbb{1}_{[\sup(x_1,...,x_n),+\infty[}(\theta).$$

Par suite

$$\widehat{\theta}_{\mathrm{MV}}(\mathbf{x}) = \arg\max_{\theta>0} p(\mathbf{x}_n;\theta) = \sup(x_1, ..., x_n).$$

Exemple 2.2 : Considérons le cas où $\mathbf{x} = (x_1, ..., x_k, ..., x_n)$ avec $x_k = ae^{i(2\pi kf+\phi)} + e_k$ où $(e_k)_{k=1,...,n}$ sont des variables aléatoires complexes indépendantes et où $\mathrm{Re}(e_k)$ et $\mathrm{Im}(e_k)$ sont indépendantes centrées de loi de probabilité gaussienne de même variance $\sigma^2/2$. Nous avons ici :

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{\pi^n \sigma^{2n}} e^{-\frac{1}{\sigma^2}\sum_{k=1}^{n} |x_k - ae^{i(2\pi kf+\phi)}|^2} \quad \text{avec} \quad \boldsymbol{\theta} = (a, \phi, f, \sigma^2).$$

La vraisemblance logarithmique $\ln p(\mathbf{x};\boldsymbol{\theta})$ s'exprime à l'aide de la *transformée de Fourier* de la suite x_k $X(f) \stackrel{\text{def}}{=} \sum_{k=1}^{n} x_k e^{-i2\pi kf}$:

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = -n \ln \sigma^2 + \frac{1}{\sigma^2}\left(2a\mathrm{Re}[X(f)e^{-i\phi}] - na^2 - \sum_{k=1}^{n} |x_k|^2\right) - n \ln \pi. \tag{2.4}$$

Pour maximiser cette fonction (2.4) par rapport à (a, ϕ, f, σ^2), nous utiliserons la propriété : $\max_{x,y} g(x, y) = \max_x [\max_y g(x, y)]$. Ainsi pour σ^2, a et f fixés, $\ln p(\mathbf{x};\boldsymbol{\theta})$ est maximum pour

$$\widehat{\phi}_{\mathrm{MV}}(\mathbf{x}) = \mathrm{Arg}[X(\widehat{f}_{\mathrm{MV}}(\mathbf{x}))].$$

Puis pour σ^2 et f fixés, $\ln p(\mathbf{x};\boldsymbol{\theta})$ est maximum pour

$$\widehat{a}_{\mathrm{MV}}(\mathbf{x}) = \frac{1}{n}|X(\widehat{f}_{\mathrm{MV}}(\mathbf{x})|.$$

Pour σ^2 fixé, $\ln p(\mathbf{x};\boldsymbol{\theta})$ est maximum pour

$$\widehat{f}_{\mathrm{MV}}(\mathbf{x}) = \arg\max |X(f)|.$$

Enfin $\ln p(\mathbf{x};\boldsymbol{\theta})$ est maximum pour

$$\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}) = \frac{1}{n}\left(\sum_{k=1}^{n} |x_k|^2 - n\widehat{a}^2_{\mathrm{MV}}(\mathbf{x})\right).$$

Nous remarquons que nous avons utilisé la dérivation que pour les maximisations par rapport à a et σ^2.

Il est par ailleurs assez facile d'énoncer des conditions suffisantes qui garantissent l'existence d'au

moins une solution de (2.1). C'est par exemple le cas où la fonction $\theta \in \Theta \mapsto p(\mathbf{x};\theta)$ est continue et où Θ est un sous ensemble compact de $\mathbb{R}^q$, car dans ce cas $p(\mathbf{x};\theta)$ est bornée et atteint en particulier son maximum dans Θ.

Remarque 2.2 : Le sous ensemble Θ de $\mathbb{R}^q$ est quelquefois défini par un ensemble de contraintes du type $g_i(\boldsymbol{\theta}) = 0$, $i = 1, ..., m$ et/ou $h_i(\boldsymbol{\theta}) \leq 0$, $i = 1, ..., p$. La recherche du maximum de vraisemblance se ramène alors au problème d'optimisation de la fonction $\boldsymbol{\theta} \in \mathbb{R}^q \mapsto p(\mathbf{x};\boldsymbol{\theta})$ sous les contraintes

$$g_i(\boldsymbol{\theta}) = 0, i = 1, ..., m, \quad \text{et/ou} \quad h_i(\boldsymbol{\theta}) \leq 0, i = 1, ..., p. \tag{2.5}$$

Un cas particulier à ce problème intervient lorsque la fonction $\boldsymbol{\theta} \in \mathbb{R}^q \mapsto -\ln p(\mathbf{x};\boldsymbol{\theta})$ est convexe avec des fonctions $g_i(\boldsymbol{\theta})$ linéaires et des fonctions $h_i(\boldsymbol{\theta})$ convexes. Ce problème d'optimisation est alors dénommé *problème d'optimisation convexe* et ainsi il bénéficie de la propriété importante que tout *maximum local* de la fonction de vraisemblance est un *maximum global* : propriété fondamentale dans la recherche par des moyens numériques du maximum de vraisemblance. De plus si la fonction $\boldsymbol{\theta} \in \mathbb{R}^q \mapsto -\ln p(\mathbf{x};\boldsymbol{\theta})$ est strictement convexe et différentiable sur Θ, il n'existe qu'un maximum local unique de la fonction de vraisemblance qui est ainsi maximum global.

Exemple 2.3 : Considérons une observation $\mathbf{x}$ constituée de variables aléatoires indépendantes identiquement distribuées $x_1, ..., x_n$ d'un mélange de q lois de probabilité de densités connues $p_i(x)$, $i = 1, ..., q$, de densité de probabilité

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{k=1}^{n} \left[\sum_{i=1}^{q} \theta_i p_i(x_k) \right] \quad \text{où} \quad \boldsymbol{\theta} = (\theta_1, ..., \theta_q) \quad \text{est le vecteur de probabilité du mélange.}$$

La fonction de vraisemblance $p(\mathbf{x};\boldsymbol{\theta})$ est donc une fonction polynomiale de degré global n en $(\theta_1, ..., \theta_q)$. Sa maximisation de façon analytique n'est donc pas possible et sa maximisation par des moyens numériques apparaît très délicate. Mais si nous considérons la log-vraisemblance $\ln p(\mathbf{x};\boldsymbol{\theta})$, nous constatons que la fonction

$$\boldsymbol{\theta} \in \mathbb{R}^q \mapsto -\ln p(\mathbf{x};\boldsymbol{\theta}) = -\sum_{k=1}^{n} \left[\ln \left(\sum_{i=1}^{q} \theta_i p_i(x_k) \right) \right],$$

est convexe (comme somme des fonctions convexes $-\ln\left(\sum_{i=1}^{q} \theta_i p_i(x_k)\right)$) et que le vecteur $\boldsymbol{\theta} \in \mathbb{R}^q$ doit suivre les contraintes :

$$\sum_{i=1}^{q} \theta_i = 1 \quad \text{et} \quad \theta_i \leq 1, \quad -\theta_i \leq 0 \quad \text{pour} \quad i = 1, .., q,$$

qui sont de la forme (2.5) où la fonction d'égalité est linéaire et les fonctions des inégalités sont convexes. Par suite, nous avons affaire à un problème d'optimisation convexe où toute recherche numérique du maximum global de la log-vraisemblance se ramène à la recherche d'un maximum local ; ce qui est beaucoup plus simple !

Exemple 2.4 : Dans certains cas exceptionnels, la recherche du maximum de vraisemblance sous contrainte, dérivée sous forme analytique est possible grâce à la technique des *multiplicateurs de Lagrange*. Ainsi considérons une observation $\mathbf{x}$ constituée de variables aléatoires indépendantes identiquement distribuées $x_1, ..., x_n$ prenant q valeurs possibles $\{a_1, ..., a_q\}$ de probabilités respectives $(\theta_1, ..., \theta_q)$. Le vecteur de paramètre $\boldsymbol{\theta}$ satisfait alors la contrainte $\sum_{i=1}^{q} \theta_i = 1$. La fonction de vraisemblance est alors :

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{k=1}^{n} \left[\prod_{i=1}^{q} \theta_i^{\mathbb{1}_{x_k = a_i}} \right],$$

où $\mathbb{1}_A = 1$ si l'événement A est réalisé et $\mathbb{1}_A = 0$ sinon. la log-vraisemblance est alors :

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = \sum_{k=1}^{n}\left[\sum_{i=1}^{q}\mathbb{1}_{x_k=a_i}\ln(\theta_i)\right] = \sum_{i=1}^{q}\ln(\theta_i)\left[\sum_{k=1}^{n}\mathbb{1}_{x_k=a_i}\right] = \sum_{i=1}^{q} n_i \ln(\theta_i),$$

où $n_i \stackrel{\text{def}}{=} \sum_{k=1}^{n}\mathbb{1}_{x_k=a_i}$ désigne le nombre de termes de l'observation $\mathbf{x}$ de taille n prenant la valeur a_i. La technique des multiplicateurs de Lagrange consiste à considérer la maximisation de la fonction

$$(\boldsymbol{\theta},\lambda) \in \mathbb{R}^q \times \mathbb{R} \mapsto \ln p(\mathbf{x};\boldsymbol{\theta}) + \lambda\left(\sum_{i=1}^{q}\theta_i - 1\right).$$

Ce qui donne les équations $\frac{n_i}{\theta_i} + \lambda = 0$ et $\sum_{i=1}^{q}\theta_i - 1 = 0$, de solutions : $\lambda = -n$ et $\theta_i = \frac{n_i}{n}$. L'estimateur du maximum de vraisemblance est donc :

$$\widehat{\boldsymbol{\theta}}_{\text{MV}}(\mathbf{x}) = \left(\frac{n_1}{n},...,\frac{n_q}{n}\right).$$

C'est à dire les fréquences relatives empiriques des différentes valeurs possibles a_i, $i = 1,..,q$ dans l'échantillon. Dans le cas particulier où $q = 2$, nous retrouvons l'échantillon de Bernoulli où l'estimateur du maximum de vraisemblance de θ_1 est la fréquence relative

$$\widehat{\theta}_{\text{MV}}(\mathbf{x}) = \frac{\sum_{k=1}^{n}\mathbb{1}_{x_k=a_1}}{n} = \frac{n_1}{n}.$$

2.3.2 Non unicité

L'estimateur du maximum de vraisemblance n'est pas nécessairement unique comme on peut s'en rendre compte avec l'exemple 2.5 et il est très difficile de trouver des conditions simples permettant d'en garantir l'unicité.

Exemple 2.5 : Considérons le cas où l'observation $\mathbf{x}_n$ est constituée de variables aléatoires $(x_k)_{k=1,...,n}$ indépendantes identiquement distribuées de loi uniforme sur $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$. Dans ce cas,

$$p(\mathbf{x}_n;\theta) = \prod_{k=1}^{n}\mathbb{1}_{[\theta-\frac{1}{2},\theta+\frac{1}{2}]}(x_k) = \mathbb{1}_{[\sup(x_1,...,x_n)-\frac{1}{2},\min(x_1,...,x_n)+\frac{1}{2}]}(\theta).$$

Par suite la vraisemblance est constante dans l'intervalle $[\sup(x_1,...,x_n) - \frac{1}{2}, \min(x_1,...,x_n) + \frac{1}{2}]$ et toute valeur prise dans cet intervalle est un estimateur du maximum de vraisemblance.

Considérons par exemple les deux estimateurs suivants

$$\widehat{\theta}^{(1)}_{\text{MV}}(\mathbf{x}_n) \stackrel{\text{def}}{=} \sup(x_1,...,x_n) - \frac{1}{2} \quad \text{et} \quad \widehat{\theta}^{(2)}_{\text{MV}}(\mathbf{x}_n) \stackrel{\text{def}}{=} \min(x_1,...,x_n) + \frac{1}{2}.$$

Il est facile de démontrer à partir des lois de probabilité des variables aléatoires $\sup(x_1,...,x_n)$ et $\min(x_1,...,x_n)$ que

$$\text{E}[\widehat{\theta}^{(1)}_{\text{MV}}(\mathbf{x}_n) - \theta]^2 = \text{E}[\widehat{\theta}^{(2)}_{\text{MV}}(\mathbf{x}_n) - \theta]^2 = \frac{2}{(n+1)(n+2)}.$$

Par suite les séquences d'estimateurs $\widehat{\theta}^{(1)}_{\text{MV}}$ et $\widehat{\theta}^{(2)}_{\text{MV}}$ sont consistantes en moyenne quadratique et donc aussi faiblement consistantes. Pour tout estimateur du maximum de vraisemblance

$$\widehat{\theta}_{\text{MV}}(\mathbf{x}_n) = a\widehat{\theta}^{(1)}_{\text{MV}}(\mathbf{x}_n) + (1-a)\widehat{\theta}^{(2)}_{\text{MV}}(\mathbf{x}_n) \quad \text{où} \quad a \in [0,1],$$

nous aurions

$$\widehat{\theta}^{(1)}_{\rm MV}(\mathbf{x}_n) \leq \widehat{\theta}_{\rm MV}(\mathbf{x}_n) \leq \widehat{\theta}^{(2)}_{\rm MV}(\mathbf{x}_n) \;\text{ et }\; |\widehat{\theta}_{\rm MV}(\mathbf{x}_n) - \theta| \leq |\widehat{\theta}^{(1)}_{\rm MV}(\mathbf{x}_n) - \theta| + |\widehat{\theta}^{(2)}_{\rm MV}(\mathbf{x}_n) - \theta|.$$

Par conséquent la suite $\widehat{\theta}_{\rm MV}(\mathbf{x}_n)$ serait aussi faiblement consistante et son risque quadratique vérifierait

$$\mathrm{E}[\widehat{\theta}_{\rm MV}(\mathbf{x}_n) - \theta]^2 \leq 2(\mathrm{E}[\widehat{\theta}^{(1)}_{\rm MV}(\mathbf{x}_n) - \theta]^2 + \mathrm{E}[\widehat{\theta}^{(2)}_{\rm MV}(\mathbf{x}_n) - \theta]^2).$$

La suite $\widehat{\theta}_{\rm MV}(\mathbf{x}_n)$ serait donc aussi consistante en moyenne quadratique. Nous remarquons que tous les estimateurs du maximum de vraisemblance n'ont pas tous le même risque quadratique. Parmi ces estimateurs, on démontrerait par exemple que l'estimateur (associé à $a = 1/2$)

$$\widehat{\theta}^{(3)}_{\rm MV}(\mathbf{x}_n) \stackrel{\rm def}{=} \frac{1}{2}\left(\sup(x_1, ..., x_n) + \min(x_1, ..., x_n)\right)$$

qui satisferait $\mathrm{E}[\widehat{\theta}^{(3)}_{\rm MV}(\mathbf{x}_n) - \theta]^2 = \frac{1}{2(n+1)(n+2)}$ minimiserait $\mathrm{E}[\widehat{\theta}_{\rm MV}(\mathbf{x}_n) - \theta]^2$.

2.3.3 Invariance

Nous allons montrer que si $\widehat{\theta}_{\rm MV}(\mathbf{x})$ est l'estimateur du maximum de vraisemblance du paramètre $\theta \in \Theta$, alors $g[\widehat{\theta}_{\rm MV}(\mathbf{x})]$ est l'estimateur du maximum de vraisemblance du paramètre $\phi = g(\theta) \in \Phi \stackrel{\rm def}{=} g(\Theta)$. Pour démontrer cette propriété, nous devons d'abord définir la vraisemblance du nouveau paramètre ϕ appelée *vraisemblance induite* de ϕ :

$$L_{\mathbf{x}}(\phi) \stackrel{\rm def}{=} \max_{\theta = g^{-1}(\phi)} L_{\mathbf{x}}(\theta). \tag{2.6}$$

Preuve : Soit $\widehat{\phi} \stackrel{\rm def}{=} g[\widehat{\theta}_{\rm MV}(\mathbf{x})]$. Puisque les sous ensembles $g^{-1}(\phi)$ de Θ pour $\phi \in \Phi$ forment une partition de Θ, $\widehat{\theta}_{\rm MV}(\mathbf{x})$ appartient à $g^{-1}(\widehat{\phi})$ et donc $\widehat{\phi}_{\rm MV}(\mathbf{x}) = g(\widehat{\theta}_{\rm MV}(\mathbf{x}))$. ■

Cette propriété justifie l'écriture

$$g[\widehat{\theta}_{\rm MV}(\mathbf{x})] = \widehat{[g(\theta)]}_{\rm MV}(\mathbf{x}). \tag{2.7}$$

Ce théorème trivial lorsque g est bijective (où il s'agit simplement d'une reparamétrisation) présente un grand intérêt pratique. Il prouve que l'estimateur de vraisemblance est *intrinsèque* dans le sens qu'il ne dépend pas du paramétrage particulier de la famille de loi de probabilité que l'on considère.

Exemple 2.6 : Si l'observation $\mathbf{x}$ est constituée de variables aléatoires $(x_k)_{k=1,...,n}$ indépendantes de Bernoulli de paramètre θ, il est souvent intéressant de s'intéresser au paramètre $\sigma^2 = \mathrm{var}(x_k) = \theta(1-\theta)$. Puisque

$$\widehat{\theta}_{\rm MV}(\mathbf{x}) = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad \text{alors} \quad \widehat{\sigma^2}_{\rm MV}(\mathbf{x}) = \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)\left(1 - \frac{1}{n}\sum_{k=1}^{n} x_k\right).$$

Remarque 2.3 : Dans le cas particulier où $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ où $\boldsymbol{\theta}_1$ et $\boldsymbol{\theta}_2$ sont respectivement les parties *paramètre utile* et *paramètre de nuisance* de $\boldsymbol{\theta}$, la vraisemblance induite de $\boldsymbol{\theta}_1$ est simplement la *vraisemblance concentrée*

$$L_{\mathbf{x}}(\boldsymbol{\theta}_1) = \max_{\boldsymbol{\theta}_2} L_{\mathbf{x}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

et le principe d'invariance se réduit à la propriété classique

$$\arg\max_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2} L_{\mathbf{x}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \left(\arg\max_{\boldsymbol{\theta}_1}(\max_{\boldsymbol{\theta}_2} L_{\mathbf{x}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)), \arg\max_{\boldsymbol{\theta}_2}(\max_{\boldsymbol{\theta}_1} L_{\mathbf{x}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2))\right)$$

2.3.4 Biais

L'estimateur du maximum de vraisemblance n'est sans biais que dans des cas très particuliers comme par exemple dans la famille exponentielle (2.8) de lois de probabilités que nous verrons dans le paragraphe 2.3.7 et de plus pour des paramétrages bien particuliers.

Notons que dans les cas très particuliers où l'espérance mathématique de l'estimateur est proportionnelle au paramètre, un simple changement d'échelle permet d'obtenir un estimateur sans biais. Ainsi dans l'exemple 2.1,

$$\mathrm{E}(\sup(x_1, ..., x_n)) = \frac{n}{n+1}\theta < \theta$$

impliquera que l'estimateur $\frac{n+1}{n}\sup(x_1, ..., x_n)$ sera non biaisé.

2.3.5 Relation avec l'efficacité

S'il existe un estimateur $\widehat{\theta}(\mathbf{x})$ efficace (estimateur sans biais dont la variance atteint la borne de Cramer-Rao) du paramètre θ, alors il est aussi l'unique estimateur du maximum de vraisemblance de θ.

Preuve : La preuve s'appuit sur la démonstration de l'inégalité de Cramer Rao : $\mathrm{var}[\widehat{\theta}(\mathbf{x})] \geq I_{\mathbf{x}}^{-1}(\theta)$ où

$$I_{\mathbf{x}}(\theta) = -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{x}, \theta)}{\partial \theta^2}\right] = \mathrm{E}\left[\left(\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right]$$

est l'information de Fisher de $\mathbf{x}$ sur le paramètre θ, à l'aide de l'inégalité de Cauchy Schwarz. Nous avons en effet en cas d'égalité pour θ scalaire (l'extension au cas multidimensionel est immédiat)

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = I_{\mathbf{x}}(\theta)[t(\mathbf{x}) - \theta].$$

Par suite $\theta = t(\mathbf{x})$ maximise $\ln p(\mathbf{x}; \theta)$ car $\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}_{|\theta=\widehat{\theta}(\mathbf{x})} = 0$ et $\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} > 0$ [resp. < 0] si $\theta < \widehat{\theta}(\mathbf{x})$ [resp. $\theta > \widehat{\theta}(\mathbf{x})$]. ■

2.3.6 Relation avec l'exhaustivité

S'il existe une statistique exhaustive $s(\mathbf{x})$, alors $\widehat{\theta}_{\mathrm{MV}}(\mathbf{x})$ est une fonction de $s(\mathbf{x})$.

Preuve : En effet d'après le critère d'exhaustivité de Neymann-Fisher, on a dans ce cas $p(\mathbf{x}; \theta) = g[s(\mathbf{x}), \theta]h(\mathbf{x})$. Par suite :

$$\max_{\theta} p(\mathbf{x}; \theta) \Leftrightarrow \max_{\theta} g[s(\mathbf{x}), \theta]h(\mathbf{x}) \Leftrightarrow \max_{\theta} g[s(\mathbf{x}), \theta].$$

D'où $\widehat{\theta}_{\mathrm{MV}}(\mathbf{x})$ est une fonction de $s(\mathbf{x})$. ■

Exemple 2.7 : Considérons le cas où l'observation $\mathbf{x}$ est constituée de variables aléatoires $(x_k)_{k=1,...,n}$ indépendantes identiquement distribuées de loi uniforme sur $[\theta, 2\theta]$. Dans ce cas,

$$p(\mathbf{x}; \theta) = \frac{1}{\theta^n}\prod_{k=1}^{n} \mathbb{1}_{[\theta,2\theta]}(x_k) = \frac{1}{\theta^n}\mathbb{1}_{[\frac{1}{2}\sup(x_1,...,x_n),\inf(x_1,...,x_n)]}(\theta)$$

et ainsi grâce au critère d'exhaustivité de Neymann-Fisher, $s(\mathbf{x}) = [\inf(x_1, ..., x_n), \sup(x_1, ..., x_n)]$ est une statistique exhaustive. On vérifie alors que $\widehat{\theta}_{\mathrm{MV}}(\mathbf{x}) = \frac{1}{2}\sup(x_1, ..., x_n)$ est bien une fonction de $s(\mathbf{x})$.

Exemple 2.8 : Considérons le cas où l'observation $\mathbf{X}$ est constituée de variables aléatoires multidimensionnelles $(\mathbf{x}_k)_{k=1,...,n}$ où $\mathbf{x}_k \in \mathbb{R}^r$, indépendantes identiquement distribuées de loi de probabilité

gaussienne $\mathcal{N}(\mathbf{m}(\boldsymbol{\theta}), \mathbf{C}(\boldsymbol{\theta}))$. Alors :

$$p(\mathbf{x};\boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k;\boldsymbol{\theta}) \text{ avec } \; p(x_k;\boldsymbol{\theta}) = \frac{1}{(2\pi)^{r/2}\det^{1/2}[\mathbf{C}(\boldsymbol{\theta})]} e^{-\frac{1}{2}(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))^T\mathbf{C}^{-1}(\boldsymbol{\theta})(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))}$$

soit $p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{nr/2}\det^{n/2}[\mathbf{C}(\boldsymbol{\theta})]} e^{-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))^T\mathbf{C}^{-1}(\boldsymbol{\theta})(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))}$ avec :

$$\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))^T\mathbf{C}^{-1}(\boldsymbol{\theta})(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta})) = \mathrm{tr}\left[\mathbf{C}^{-1}(\boldsymbol{\theta})\left(\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))^T\right)\right]$$

où

$$\sum_{k=1}^{n}(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))(\mathbf{x}_k-\mathbf{m}(\boldsymbol{\theta}))^T = n\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^T - \mathbf{m}(\boldsymbol{\theta})(\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k)^T - (\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k)\mathbf{m}^T(\boldsymbol{\theta}) + \mathbf{m}(\boldsymbol{\theta})\mathbf{m}^T(\boldsymbol{\theta})\right).$$

Dans ce cas, là aussi par le critère d'exhaustivité de Neymann-Fisher,

$$s(\mathbf{X}) = \left[\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k, \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^T\right]$$

est une statistique exhaustive, de même d'ailleurs que

$$\left[\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k, \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k-\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\right)\left(\mathbf{x}_k-\frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\right)^T\right]$$

(*vecteur moyenne empirique et matrice de covariance empirique*). L'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{X})$ sera une fonction de $s(\mathbf{X})$, dit *estimateur du second-ordre.*

2.3.7 Cas particulier de la famille exponentielle

Définition : Une loi de probabilité appartient au modèle exponentiel s'il existe une paramétrisation de la densité de probabilité de l'observation $\mathbf{x}$ (au sens de Radon Nikodym par rapport à la mesure μ de Lebesgue ou de comptage) qui s'écrit sous la forme :

$$p(\mathbf{x};\boldsymbol{\theta}) = e^{\boldsymbol{\theta}^T\mathbf{t}(\mathbf{x})-\phi(\boldsymbol{\theta})}h(\mathbf{x}) \;\text{ avec } \boldsymbol{\theta}\in\Theta\subset\mathbb{R}^q \text{ et } \mathbf{t}(\mathbf{x})\in\mathbb{R}^q, \tag{2.8}$$

et où la fonction $\boldsymbol{\theta}\mapsto\phi(\boldsymbol{\theta})\in\mathbb{R}$ satisfait la normalisation $e^{\phi(\boldsymbol{\theta})} = \int e^{\boldsymbol{\theta}^T\mathbf{t}(\mathbf{x})}h(\mathbf{x})d\mu$ où l'intégrale précédente est supposée exister et dérivable deux fois par rapport à $\boldsymbol{\theta}$.

Propriétés : $\mathbf{t}(\mathbf{x})$ est l'estimateur du maximum de vraisemblance [3] unique du paramètre

$$\boldsymbol{\psi}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial\phi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \in \Psi\subset\mathbb{R}^q \;\text{ où }\; \Psi \stackrel{\text{def}}{=} \boldsymbol{\psi}(\Theta).$$

De plus cet estimateur est

- *sans biais*
- de variance $\frac{\partial^2\phi(\theta)}{\partial\theta^2}$ (cas $q=1$) ou de matrice de covariance $\frac{\partial^2\phi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$ (dans le cas $q>1$, qui est par suite définie positive)

3. sous la condition que l'application $\boldsymbol{\theta}\in\Theta\mapsto\boldsymbol{\psi}(\boldsymbol{\theta})\in\Psi$ soit bijective. Lorsque cette application est seulement injective, il s'agit de la vraisemblance induite de $\boldsymbol{\psi}$ définie en (2.6).

- *efficace.*

Preuve : Par dérivation sous le signe somme de la normalisation $e^{\phi(\boldsymbol{\theta})} = \int e^{\boldsymbol{\theta}^T \mathbf{t}(\mathbf{x})} h(\mathbf{x}) d\mu$, on déduit :

$$\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} e^{\phi(\boldsymbol{\theta})} = \int \mathbf{t}(\mathbf{x}) e^{\boldsymbol{\theta}^T \mathbf{t}(\mathbf{x})} h(\mathbf{x}) d\mu = e^{\phi(\boldsymbol{\theta})} \int \mathbf{t}(\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}) d\mu.$$

Par suite, d'après le théorème de transfert

$$\mathrm{E}[\mathbf{t}(\mathbf{x})] = \int \mathbf{t}(\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}) d\mu = \frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Puis en dérivant une deuxième fois, on obtient :

$$\left[\left(\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T + \left(\frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) \right] e^{\phi(\boldsymbol{\theta})} = \int \mathbf{t}(\mathbf{x}) \mathbf{t}^T(\mathbf{x}) e^{\boldsymbol{\theta}^T \mathbf{t}(\mathbf{x})} h(\mathbf{x}) d\mu = e^{\phi(\boldsymbol{\theta})} \int \mathbf{t}(\mathbf{x}) \mathbf{t}^T(\mathbf{x}) p(\mathbf{x}; \boldsymbol{\theta}) d\mu.$$

Ce qui donne :

$$\left(\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T + \left(\frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) = \mathrm{E}[\mathbf{t}(\mathbf{x}) \mathbf{t}^T(\mathbf{x})]$$

et donc :

$$\mathrm{cov}[\mathbf{t}(\mathbf{x})] \stackrel{\mathrm{def}}{=} \mathrm{E}[\mathbf{t}(\mathbf{x}) \mathbf{t}^T(\mathbf{x})] - \mathrm{E}[\mathbf{t}(\mathbf{x})] \mathrm{E}[\mathbf{t}^T(\mathbf{x})] = \frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

Ainsi l'estimateur $\mathbf{t}(\mathbf{x})$ est un estimateur sans biais du paramètre $\boldsymbol{\psi}(\boldsymbol{\theta})$ de matrice de covariance $\frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$.
Démontrons qu'il s'agit de l'estimateur du maximum de vraisemblance unique du paramètre $\boldsymbol{\psi}(\boldsymbol{\theta})$. Puisque

$$\ln p(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^T \mathbf{t}(\mathbf{x}) - \phi(\boldsymbol{\theta}) + \ln(h(\mathbf{x})), \tag{2.9}$$

$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{t}(\mathbf{x}) - \frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ où la fonction $\boldsymbol{\theta} \mapsto \phi(\boldsymbol{\theta})$ est strictement convexe car $\frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ est définie positive. Par suite $p(\mathbf{x}; \boldsymbol{\theta})$ est maximum pour $\boldsymbol{\theta}$ vérifiant $\frac{\partial \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{t}(\mathbf{x})$ et d'après la définition de la vraisemblance induite de $\boldsymbol{\psi}$ définie en (2.6), celle-ci est bien maximum pour $\boldsymbol{\psi} = \mathbf{t}(\mathbf{x})$ et ce maximum est unique.

Démontrons que cet estimateur $\mathbf{t}(\mathbf{x})$ sans biais du paramètre $\boldsymbol{\psi}$ atteint la borne de Cramer-Rao $\mathrm{CRB}(\boldsymbol{\psi})$. L'information de Fisher apporté par $\mathbf{x}$ sur le paramètre $\boldsymbol{\theta}$ est donnée par

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right] = \mathrm{E}\left[\left(\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \right],$$

d'après (2.9). Par suite à partir de $\mathrm{CRB}(\boldsymbol{\theta}) = \mathbf{I}^{-1}(\boldsymbol{\theta})$, on déduit pour le paramètre $\boldsymbol{\psi}$:

$$\mathrm{CRB}(\boldsymbol{\psi}) = \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \mathbf{I}^{-1}(\boldsymbol{\theta}) \left(\frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\theta}} \right)^T = \frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \left(\frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1} \frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{\partial^2 \phi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \mathrm{cov}[\mathbf{t}(\mathbf{x})].$$

■

Exemple 2.9 : De nombreuses lois de probabilité appartiennent à la famille exponentielle qui comprend par exemple les lois de probabilités de Bernoulli, binomiale, de Poisson, géométrique, exponentielle, gamma, gaussienne... Notons que si des observations $x_1, ..., x_n$ sont indépendantes équidistribuées dont la loi de probabilité appartient à la famille exponentielle, il en est de même de l'observation $\mathbf{x}_n = (x_1, ..., x_n)$, car nous avons alors :

$$p(\mathbf{x}_n; \boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k; \boldsymbol{\theta}) = \prod_{k=1}^{n} \left[e^{\boldsymbol{\theta}^T \mathbf{t}(x_k) - \phi(\boldsymbol{\theta})} h(x_k) \right] = e^{\boldsymbol{\theta}^T \left[\sum_{k=1}^{n} \mathbf{t}(x_k) \right] - n\phi(\boldsymbol{\theta})} \prod_{k=1}^{n} h(x_k).$$

En particulier pour les lois de probabilités de Bernoulli (p), exponentielle (λ) et gaussienne (m, σ^2),

nous avons respectivement :

$$\begin{aligned}
p(\mathbf{x}_n;p) &= \prod_{k=1}^{n} p^{x_k}(1-p)^{1-x_k} = p^{\sum_{k=1}^n x_k}(1-p)^{n-\sum_{k=1}^n x_k} = e^{\left(n\ln\frac{p}{1-p}\right)\frac{1}{n}\sum_{k=1}^n x_k} e^{n\ln(1-p)} \\
p(\mathbf{x}_n;\lambda) &= \lambda^n e^{-\lambda\sum_{k=1}^n x_k}\prod_{k=1}^{n}\mathbb{1}_{[0,+\infty]}(x_k) = e^{-n\lambda\left(\frac{1}{n}\sum_{k=1}^n x_k\right)} e^{n\ln\lambda}\prod_{k=1}^{n}\mathbb{1}_{[0,+\infty]}(x_k) \\
p(\mathbf{x}_n;m,\sigma^2) &= \frac{1}{\sigma^n}\frac{1}{(\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{k=1}^n (x_k-m)^2} = e^{(\frac{nm}{\sigma^2},-\frac{n}{2\sigma^2})(\frac{1}{n}\sum_{k=1}^n x_k,\frac{1}{n}\sum_{k=1}^n x_k^2)^T} e^{-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}nm^2},
\end{aligned}$$

qui s'écrivent toutes sous la structure (2.8) avec le nouveau paramètre θ défini respectivement par

$$\theta = n\ln\frac{p}{1-p},\quad \theta = -n\lambda \ \text{ et } \ \boldsymbol{\theta} = \begin{bmatrix} \frac{nm}{\sigma^2} \\ -\frac{n}{2\sigma^2} \end{bmatrix},$$

les statistiques $t(\mathbf{x}_n)$:

$$t(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k, t(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k \ \text{ et } \ \mathbf{t}(\mathbf{x}_n) = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n x_k \\ \frac{1}{n}\sum_{k=1}^n x_k^2 \end{bmatrix}$$

et les fonctions $\phi(\theta)$:

$$\phi(\theta) = -n\ln(1-p),\quad \phi(\theta) = -n\ln\lambda \ \text{ et } \ \phi(\boldsymbol{\theta}) = \frac{n}{2}\ln 2\pi + \frac{n}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}nm^2.$$

Par suite les statistiques $t(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k$, $t(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k$ et $\mathbf{t}(\mathbf{x}_n) = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n x_k \\ \frac{1}{n}\sum_{k=1}^n x_k^2 \end{bmatrix}$ sont des estimateurs du maximum de vraisemblance sans biais et efficace des paramètres respectifs

$$\begin{aligned}
\psi &= \frac{d\phi}{d\theta} = \frac{d\phi}{dp}\frac{dp}{d\theta} = \left(\frac{n}{1-p}\right)\left(\frac{n}{p}+\frac{n}{1-p}\right)^{-1} = p \\
\psi &= \frac{d\phi}{d\theta} = \frac{d\phi}{d\lambda}\frac{d\lambda}{d\theta} = \frac{1}{\lambda} \\
\boldsymbol{\psi} &= \begin{bmatrix} \frac{\partial\phi}{\partial d\theta_1} \\ \frac{\partial\phi}{\partial d\theta_2} \end{bmatrix} = \begin{bmatrix} -\frac{\theta_1}{2\theta_2} \\ -\frac{n}{2\theta_2}+\frac{\theta_1^2}{4\theta_2^2} \end{bmatrix} = \begin{bmatrix} m \\ \sigma^2+m^2 \end{bmatrix}.
\end{aligned}$$

Nous obtenons ainsi par application de la propriété d'invariance (2.7) de l'estimateur du maximum de vraisemblance, les estimateurs du maximum de vraisemblance respectifs :

$$\widehat{p}_{\mathrm{MV}}(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k \tag{2.10}$$

$$\widehat{\lambda}_{\mathrm{MV}}(\mathbf{x}_n) = \frac{1}{\frac{1}{n}\sum_{k=1}^n x_k} \tag{2.11}$$

$$\begin{bmatrix} \widehat{m}_{\mathrm{MV}}(\mathbf{x}_n) \\ \widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n x_k \\ \frac{1}{n}\sum_{k=1}^n \left(x_k - \frac{1}{n}\sum_{k=1}^n x_k\right)^2 \end{bmatrix}. \tag{2.12}$$

Remarque 2.4 : Ces résultats peuvent bien entendu s'obtenir de façon élémentaire par résolution directe des équations de vraisemblance (2.3). Mais notons que le choix du paramétrage de la loi de probabilité du modèle exponentiel est essentiel pour l'obtention d'estimateurs efficaces. Ainsi les estimateurs du

maximum de vraisemblance $\widehat{\lambda}_{\mathrm{MV}}(\mathbf{x}_n)$ de λ et $\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)$ de σ^2 ne sont plus sans biais (en particulier $\mathrm{E}(\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x})) = \frac{n-1}{n}\sigma^2$).

De plus, on peut démontrer que la famille exponentielle est le seul ensemble de loi de probabilité qui permettent d'obtenir des estimateurs efficaces. Comme cette famille exponentielle de loi de probabilité est très spécifique, nous voyons qu'en général il n'existera pas d'estimateur sans biais de variance/covariance atteignant la borne de Cramer-Rao). En pratique ces propriétés seront seulement approchées. Leur étude sera menée en considérant des séquences $\widehat{\theta}_{\mathrm{MV}}(\mathbf{x}_n)$ d'estimateurs MV construites à partir d'une séquence $\mathbf{x}_n$ d'observations.

2.4 Consistance et normalité asymptotique

2.4.1 Exemple élémentaire

Considérons le cas d'une observation $\mathbf{x}_n = (x_1, ..., x_n)$ constituée de variables aléatoires indépendantes équidistribuées de loi de probabilité gaussienne $\mathcal{N}(m, \sigma^2)$, donc de paramètre $\boldsymbol{\theta} = (m, \sigma^2)^T$. Etudions les propriétés asymptotiques de la suite d'estimateurs $(\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n), \widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n))$ de (2.12).

Puisque

$$\mathrm{E}(\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n)) = m \quad \text{et} \quad \mathrm{E}(\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)) = \frac{n-1}{n}\sigma^2,$$

$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ est un estimateur biaisé mais *asymptotiquement sans biais*. Plus précisément, le biais tend vers 0 quand $n \to \infty$ avec

$$\mathrm{E}(\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)) - \boldsymbol{\theta} = -\frac{1}{n}\begin{bmatrix} 0 \\ \sigma^2 \end{bmatrix}.$$

Puisque

$$\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{X}_n \quad \text{et} \quad \widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2 = S_n^2,$$

la loi forte des grands nombres nous assure que ces suites d'estimateurs sont *fortement consistantes* (convergence presque sure vers le paramètre).

Ces suites d'estimateurs sont aussi *consistantes en moyenne quadratique* car des calculs élémentaires (pour le calcul de $\mathrm{E}[\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) - \sigma^2]^2$, on pourra utiliser la normalisation de la variable aléatoire $X = m + \sigma X_u$ de loi $\mathcal{N}(m, \sigma^2)$ qui donne $\mathrm{E}(X_u^4) = 3$) montrent que

$$\mathrm{E}[\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n) - m]^2 = \frac{\sigma^2}{n} \quad \text{et} \quad \mathrm{E}[\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) - \sigma^2]^2 = \frac{(2n-1)\sigma^4}{n^2}.$$

Considérons maintenant la loi de probabilité asymptotique de la suite de variables aléatoires

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n) = \begin{bmatrix} \widehat{m}_{\mathrm{MV}}(\mathbf{x}_n) \\ \widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) \end{bmatrix}.$$

D'après le résultat de probabilité 1.2 (1.7), les variables aléatoires $\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n)$ et $\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)$ sont indépendantes car $\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n) = \bar{X}_n$ et $\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) = S_n^2$. Puisque la loi de probabilité de $\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n)$ est gaussienne de moyenne m et de variance $\frac{\sigma^2}{n}$, nous avons naturellement

$$\sqrt{n}\,(\widehat{m}_{\mathrm{MV}}(\mathbf{x}_n) - m) \quad \text{est de loi gaussienne } \mathcal{N}(0, \sigma^2),$$

ce qui est un cas particulier de la convergence en loi. Puis sachant que $\frac{n\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)}{\sigma^2}$ d'après le résultat de probabilité 2 (1.6) est de loi de probabilité de Khi-deux à $n-1$ degrés de liberté, nous obtenons à l'aide

du théorème central limite classique appliqué à une somme de $n-1$ carrés de variables aléatoires U_k indépendantes de loi gaussienne $\mathcal{N}(0,1)$ (avec $\mathrm{E}(U_k^2)=1$ et $\mathrm{E}(U_k^4)=3$, donc $\mathrm{var}(U_k^2)=2$),

$$\frac{\frac{n\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)}{\sigma^2}-(n-1)}{\sqrt{2(n-1)}} \text{ converge en loi vers la loi gaussienne } \mathcal{N}(0,1).$$

Puis à l'aide de 2 applications du théorème de Slutsky rappelé au résultat de probabilité 1.4 où V_n sont des suites déterministes, nous obtenons successivement : $\frac{\frac{n\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)}{\sigma^2}-(n-1)}{\sqrt{2n}}$, $\frac{\frac{n\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)}{\sigma^2}-n}{\sqrt{2n}}$ convergent en loi vers la loi gaussienne $\mathcal{N}(0,1)$, puis

$$\sqrt{n}\left(\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)-\sigma^2\right) \text{ converge en loi vers la loi gaussienne } \mathcal{N}(0,2\sigma^4). \tag{2.13}$$

Ce qui permet d'obtenir la convergence en loi suivante d'après l'exemple 1.7, qui s'étendra à la plupart des échantillons $\mathbf{x}_n$ constitué de variables aléatoires $x_1,...,x_n$ indépendantes équidistribuées au paragraphe 2.4.2 :

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)-\boldsymbol{\theta}\right) \text{ converge en loi vers la loi gaussienne } \mathcal{N}\left(\mathbf{0},\mathbf{I}_1^{-1}(\boldsymbol{\theta})\right)$$

avec ici

$$\mathbf{I}_1(\boldsymbol{\theta})=\begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^4}\end{bmatrix}.$$

Remarque 2.5 : Notons que ce résultat (2.13) aurait pu être obtenu directement au moyen d'un théorème classique dit *théorème de continuité* que nous rappelons :

Résultat de probabilités 2.1 : Si $\mathbf{x}_n$ est une suite de variables aléatoires scalaires ou multidimensionnelles d'espérance $\mathbf{m}$ et de matrice de covariance $\mathbf{C}$, si $v_n>0$ est une suite déterministe satisfaisant

$$\lim_{n\to\infty} v_n=+\infty$$

(appelée *vitesse de convergence*) satisfaisant :

$$v_n\left(\mathbf{x}_n-\mathbf{m}\right) \text{ converge en loi vers la loi gaussienne } \mathcal{N}(\mathbf{0},\mathbf{C})$$

et si $\mathbf{y}_n=\mathbf{g}(\mathbf{x}_n)$ où $\mathbf{g}$ est une application de $\mathbb{R}^p$ dans $\mathbb{R}^q$ différentiable dans un voisinage de $\mathbf{m}$ dont la matrice de l'application différentiable au point $\mathbf{m}$ est notée $\mathbf{D}_g(\mathbf{m})$ (quelquefois aussi $\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}_{|\mathbf{x}=\mathbf{m}}$) avec $\mathbf{D}_g(\mathbf{m})\neq\mathbf{0}$, c'est à dire :

$$\mathbf{g}(\mathbf{m}+\delta\mathbf{m})=\mathbf{g}(\mathbf{m})+\mathbf{D}_g(\mathbf{m})\delta\mathbf{m}+o(||\delta\mathbf{m}||),$$

alors nous avons :

$$v_n\left(\mathbf{g}(\mathbf{x}_n)-\mathbf{g}(\mathbf{m})\right) \text{ converge en loi vers la loi gaussienne } \mathcal{N}\left(\mathbf{0},\mathbf{D}_g(\mathbf{m})\mathbf{C}\mathbf{D}_g^T(\mathbf{m})\right). \tag{2.14}$$

Ici $\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n)=g\left(\frac{1}{n}\sum_{k=1}^n x_k,\frac{1}{n}\sum_{k=1}^n x_k^2\right)$ avec $g(u,v)=v-u^2$ et le théorème limite central classique appliqué à la suite $\begin{bmatrix}\frac{1}{n}\sum_{k=1}^n x_k\\ \frac{1}{n}\sum_{k=1}^n x_k^2\end{bmatrix}$ donne :

$$\sqrt{n}\begin{bmatrix}\frac{1}{n}\sum_{k=1}^n x_k-m\\ \frac{1}{n}\sum_{k=1}^n x_k^2-(\sigma^2+m^2)\end{bmatrix} \text{ converge en loi vers la loi gaussienne } \mathcal{N}\left(\mathbf{0},\begin{bmatrix}\sigma^2 & 2\sigma^2 m\\ 2\sigma^2 m & 2\sigma^4+4\sigma^2 m^2\end{bmatrix}\right)$$

car

$$\mathrm{var}(x_k) = \sigma^2, \;\; \mathrm{var}(x_k^2) = \mathrm{E}(x_k^4) - [\mathrm{E}(x_k^2)]^2 = 2\sigma^4 + 4\sigma^2 m^2 \;\; \text{et} \;\; \mathrm{cov}(x_k, x_k^2) = \mathrm{E}(x_k^3) - \mathrm{E}(x_k)\mathrm{E}(x_k^2) = 2\sigma^2 m.$$

L'application du théorème de continuité (2.14) précédent redonne le résultat (2.13) car ici : $\mathbf{D}_g(\mathbf{m}) = [-2m, 1]$.

2.4.2 Modèle indépendant identiquement distribué général

Nous avons pu démontrer sur l'exemple élémentaire précédent que la suite d'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ jouissait des propriétés asymptotiques suivantes :

- Biais asymptotique en $1/n$:

$$\mathrm{biais}(\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)) \stackrel{\mathrm{def}}{=} \mathrm{E}(\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)) - \boldsymbol{\theta} = O\left(\frac{1}{n}\right),$$

- Consistance forte : convergence presque sûre de $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ vers $\boldsymbol{\theta}$ pour tout $\boldsymbol{\theta} \in \Theta$,
- Loi de probabilité asymptotique gaussienne avec une vitesse en $\sqrt{n}$:

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n) - \boldsymbol{\theta}) \;\; \text{converge en loi vers une loi gaussienne} \;\; \mathcal{N}(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})),$$

- Efficacité asymptotique : $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{I}_1^{-1}(\boldsymbol{\theta})$ pour tout $\boldsymbol{\theta} \in \Theta$, où $\mathbf{I}_1(\boldsymbol{\theta})$ désigne la matrice d'information de Fisher associée à une seule observation x_k.

Considérons le modèle plus général où l'observation $\mathbf{x}_n = (x_1, ..., x_n)$ est constituée de variables aléatoires x_k indépendantes et identiquement distribuées (i.i.d.). Sous certaines hypothèses techniques sur l'ensemble $\Theta \in \mathbb{R}^q$ des paramètres et sur la vraisemblance $p(x_k; \boldsymbol{\theta})$ de x_k, il a été démontré que pour tout $\boldsymbol{\theta} \in \Theta$, il existe une solution $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ des équations de vraisemblances (2.3) qui est un estimateur du paramètre $\boldsymbol{\theta}$ qui satisfaisait les 4 propriétés précédentes.

2.4.3 Modèle général

Dans le cas d'une séquence de modèles paramétriques $\{P_\theta(\mathbf{x}_n), \boldsymbol{\theta} \in \Theta\}$ générale, nous ne disposons pas de résultats similaires et les différentes propriétés asymptotiques devront être démontrées au cas par cas. Cependant dans l'immense majorité des cas rencontrés en pratique satisfaisant certaines propriétés de régularité, il existera une solution $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ des équations de vraisemblances (2.3) qui sera fortement consistante, asymptotiquement sans biais, asymptotiquement gaussienne au sens général suivant :

Il existe une suite $v_n > 0$ déterministe monotone croissante satisfaisant $\lim_{n\to\infty} v_n = \infty$ telle que

$$v_n(\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n) - \boldsymbol{\theta}) \;\; \text{converge en loi vers une loi gaussienne} \;\; \mathcal{N}(0, \mathbf{C}(\boldsymbol{\theta})). \tag{2.15}$$

Pour définir l'efficacité asymptotique, il faut définir d'abord le concept de *matrice d'information de Fisher asymptotique normalisée* $\mathbf{J}_0(\boldsymbol{\theta})$. Pour cela si $\mathbf{I}_n(\boldsymbol{\theta})$ désigne la matrice d'information de Fisher associée à $\mathbf{x}_n$ et si la suite v_n satisfait que la limite de $v_n^{-2}\mathbf{I}_n(\boldsymbol{\theta})$ existe, alors $\mathbf{J}_0(\boldsymbol{\theta})$ est définie comme cette limite :

$$\mathbf{J}_0(\boldsymbol{\theta}) = \lim_{n\to\infty} v_n^{-2}\mathbf{I}_n(\boldsymbol{\theta}).$$

Et dans ce cadre, si la matrice de covariance $\mathbf{C}(\boldsymbol{\theta})$ de la loi asymptotique (2.15) vérifie :

$$\mathbf{C}(\boldsymbol{\theta}) = \mathbf{J}_0^{-1}(\boldsymbol{\theta}),$$

l'estimateur $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ est dit asymptotiquement efficace. Cette définition étend la définition précédente définie dans paragraphe 2.4.2 dans le cadre i.i.d. pour lequel :

$$\mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{I}_1(\boldsymbol{\theta}), \quad \text{soit} \quad v_n = \sqrt{n} \quad \text{et} \quad \mathbf{J}_0(\boldsymbol{\theta}) = \mathbf{I}_1(\boldsymbol{\theta}).$$

Lorsque la loi de probabilité du modèle paramétrique $\mathcal{P} = \{P_{\boldsymbol{\theta}}(\mathbf{x}), \theta \in \Theta\}$ est non linéaire, l'efficacité asymptotique peut prendre des sens plus généraux comme dans l'exemple suivant :

Exemple 2.10 : Les propriétés asymptotiques de l'estimateur du maximum de vraisemblance du paramètre

$$\boldsymbol{\theta} = (a, \phi, f, \sigma^2)$$

décrit dans l'exemple 2.2 ont été abondamment étudiées dans la littérature. On y a démontré que $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ est fortement consistant et consistant en moyenne quadratique, asymptotiquement sans biais et asymptotiquement efficace au sens suivant :

$$\begin{bmatrix} n^{1/2}(\widehat{a}_{\mathrm{MV}}(\mathbf{x}_n) - a) \\ n^{1/2}(\widehat{\phi}_{\mathrm{MV}}(\mathbf{x}_n) - \phi) \\ n^{3/2}(\widehat{f}_{\mathrm{MV}}(\mathbf{x}_n) - f) \\ n^{1/2}(\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) - \sigma^2) \end{bmatrix} \text{ converge en loi vers la loi gaussienne } \mathcal{N}\left(\mathbf{0}, \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{4}{a^2} & -\frac{3}{\pi a^2} & 0 \\ 0 & -\frac{3}{\pi a^2} & \frac{3}{\pi^2 a^2} & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}\right),$$

et la matrice d'information de Fisher $\mathbf{I}_n(\boldsymbol{\theta})$ associée à l'observation $\mathbf{x}_n = (x_1, ..., x_n)$ est donnée par :

$$\mathbf{I}_n(\boldsymbol{\theta}) = \frac{n}{\sigma^2} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & a^2 & \pi(n-1)a^2 & 0 \\ 0 & \pi(n-1)a^2 & \frac{2\pi^2(n-1)(2n-1)}{3} a^2 & 0 \\ 0 & 0 & 0 & \frac{1}{\sigma^2} \end{bmatrix} \sim \frac{1}{\sigma^2} \begin{bmatrix} \frac{1}{n} & 0 & 0 & 0 \\ 0 & \frac{4}{na^2} & -\frac{3}{n^2\pi a^2} & 0 \\ 0 & -\frac{3}{n^2\pi a^2} & \frac{3}{n^3\pi^2 a^2} & 0 \\ 0 & 0 & 0 & \frac{\sigma^2}{n} \end{bmatrix}^{-1} \text{ pour } n \to \infty.$$

Puisque les lois de probabilités marginales issues de la loi de probabilité gaussienne sont aussi gaussiennes, on en déduit les convergences en loi suivantes :

$$\sqrt{n}(\widehat{a}_{\mathrm{MV}}(\mathbf{x}_n) - a) \to_{\mathrm{Loi}} \mathcal{N}\left(0, \sigma^2\right), \quad \sqrt{n}(\widehat{\phi}_{\mathrm{MV}}(\mathbf{x}_n) - \phi) \to_{\mathrm{Loi}} \mathcal{N}\left(0, \frac{4\sigma^2}{a^2}\right),$$

$$n^{3/2}(\widehat{f}_{\mathrm{MV}}(\mathbf{x}_n) - f) \to_{\mathrm{Loi}} \mathcal{N}\left(0, \frac{3\sigma^2}{\pi^2 a^2}\right), \quad \sqrt{n}(\widehat{\sigma^2}_{\mathrm{MV}}(\mathbf{x}_n) - \sigma^4) \to_{\mathrm{Loi}} \mathcal{N}\left(0, \sigma^4\right).$$

Cet exemple montre que contrairement au cas des échantillons de variables aléatoires indépendantes et identiquement distribuées, on peut obtenir des vitesses de convergence supérieures à $\sqrt{n}$. Notons qu'ici les observations $x_1, ..., x_n$ sont bien indépendantes mais *non identiquement distribuées.*

2.5 Calcul numérique de l'estimateur MV par des méthodes de type Newton

Dans un grand nombre de situations pratiques, les équations de vraisemblance (2.3) n'admettent pas de solutions explicites. La détermination de l'estimateur du maximum de vraisemblance doit alors être effectuée à l'aide de méthodes numériques de maximisation.

A priori, toutes les méthodes classiques de recherche de maximum de fonctions peuvent être utilisées à cet effet : méthode de gradient, de Newton-Raphson,... D'autres solutions, plus liées au problème statistique lui-même peuvent aussi être employées comme la *méthode du score* (méthode de Newton-Raphson adaptée) et la méthode EM.

En général, ces méthodes ne convergent vers le maximum global de la vraisemblance que si la valeur initiale de l'algorithme est suffisamment proche de l'optimum ou si ce problème d'optimisation est un

problème convexe (auquel cas tout maximum local de la vraisemblance en est un maximum global). Dans le cas contraire, l'algorithme convergera vers un maximum local ou vers un point selle. La difficulté principale de ces algorithmes itératifs est donc de déterminer si l'algorithme a convergé et dans le cas où il a convergé, de savoir si la valeur trouvée correspond à un maximum global de la vraisemblance.

Il est important, de remarquer qu'à la différence des problèmes de maximisation classiques, la vraisemblance, fonction du paramètre, n'est pas complètement définie a priori car elle dépend de l'observation $\mathbf{x}$.

Dans la méthode classique de Newton-Raphson [4] appliquée à la log-vraisemblance $l_{\mathbf{x}}(\boldsymbol{\theta}) = \ln p(\mathbf{x};\boldsymbol{\theta})$, les itérations sont les suivantes :

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}_{|\boldsymbol{\theta}=\boldsymbol{\theta}_i} \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}_{|\boldsymbol{\theta}=\boldsymbol{\theta}_i}. \tag{2.16}$$

Si la log-vraisemblance $l_{\mathbf{x}}(\boldsymbol{\theta})$ est une fonction concave de $\boldsymbol{\theta}$, la séquence $\boldsymbol{\theta}_i$ converge vers $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x})$ (en une seule itération d'ailleurs si $l_{\mathbf{x}}(\boldsymbol{\theta})$ est une fonction quadratique de $\boldsymbol{\theta}$). La convergence est dans ce cas quadratique (donc jugée rapide par rapport à la convergence de l'algorithme du gradient) :

$$\|\boldsymbol{\theta}_{i+1} - \widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x})\| \leq c\|\boldsymbol{\theta}_i - \widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x})\|^2, \quad \text{où} \quad 0 < c < 1.$$

Mais en général $l_{\mathbf{x}}(\boldsymbol{\theta})$ n'est pas concave et la convergence vers $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x})$ dépendra fortement de la condition initiale $\boldsymbol{\theta}_0$ de l'algorithme.

La méthode dite du *score* est obtenue à partir de la méthode de Newton-Raphson par remplacement [5] du Hessien de la log-vraisemblance par l'opposé de la matrice d'information de Fisher.

$$\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \curvearrowright \mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right] = -\mathbf{I}_{\mathbf{x}}(\boldsymbol{\theta}),$$

donnant les itérations :

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \mathbf{I}^{-1}_{\mathbf{x}}(\boldsymbol{\theta})_{|\boldsymbol{\theta}=\boldsymbol{\theta}_i} \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}_{|\boldsymbol{\theta}=\boldsymbol{\theta}_i}. \tag{2.17}$$

2.6 Calcul numérique de l'estimateur MV par la méthode EM

2.6.1 Introduction

Certains problèmes d'estimation paramétriques à *données cachées* ou *variables latentes* simples nécessitent des maximisations de la vraisemblance $L_{\mathbf{x}}(\boldsymbol{\theta})$ très complexes. Citons en quatre exemples :

1. Mélange de q lois de probabilité continues $(f_j(x_k;\boldsymbol{\phi}_j))_{j=1,...,q}$ de n variables aléatoires $x_1, ..., x_n$ i.i.d. de vraisemblance

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{k=1}^{n}\left(\sum_{j=1}^{q} p_j f_j(x_k;\boldsymbol{\phi}_j)\right) \quad \text{où} \quad \boldsymbol{\theta} = (p_1, ..., p_q, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_q) \quad \text{avec} \quad \sum_{j=1}^{q} p_j = 1.$$

Par exemple, mélange de deux distributions gaussiennes que nous examinerons en détail dans la

4. Obtenue par recherche du zéro de $\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ au voisinage de $\boldsymbol{\theta}_i$ par le développement $\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}_{|\boldsymbol{\theta}=\boldsymbol{\theta}_i} + \left[\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]_{|\boldsymbol{\theta}=\boldsymbol{\theta}_i} (\boldsymbol{\theta} - \boldsymbol{\theta}_i) + o((\boldsymbol{\theta} - \boldsymbol{\theta}_i))$.

5. Ce remplacement peut être interprété comme une approximation qui peut être justifiée dans le cadre d'observations $\mathbf{x}_n = (x_1, ..., x_n)$ i.i.d. par la loi des grands nombres car : $\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2 \ln p(x_k;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \approx \mathrm{E}\left(\frac{\partial^2 \ln p(x_1;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right) = \mathbf{I}_1(\boldsymbol{\theta}) = \frac{1}{n}\mathbf{I}_{\mathbf{x}}(\boldsymbol{\theta})$.

partie 2.6.5 :

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{k=1}^{n} \left(p_1 \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x_k-m_1)^2}{2\sigma_1^2}} + (1-p_1)\frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x_k-m_2)^2}{2\sigma_2^2}} \right) \quad \text{où} \quad \boldsymbol{\theta} = (p_1, m_1, m_2, \sigma_1^2, \sigma_2^2).$$

2. Modèle stochastique dynamique : on considère l'observation $\mathbf{x} = (x_1, .., x_n)$ générée par les deux équations suivantes pour $k = 1, .., n$

$$\begin{aligned} x_k &= g_k(z_k, v_k) \qquad \text{équation d'observation} \\ z_k &= h_k(z_{k-1}, u_k) \quad \text{équation d'évolution} \quad \Rightarrow z_k \text{ est un processsus de Markov,} \end{aligned}$$

où $(v_1, ..., v_n, u_1, .., u_n, z_0)$ sont des variables aléatoires mutuellement indépendantes. Les fonctions linéaires ou non linéaires g_k et h_k, pour $k = 1, .., n$ et les distributions de $(v_1, ..., v_n, u_1, .., u_n, z_0)$ sont paramétrées par $\boldsymbol{\theta}$. La fonction de vraisemblance est alors en général, une fonction sous forme intégrale très complexe.

3. Fusion de bits (détection en sortie de K canaux binaires symétriques à erreurs indépendantes)

$$x_{k,t} = b_t \oplus n_{k,t}, \text{ avec } \; k = 1, ..., K \text{ et } \; t = 1, ..., T$$

où $b_1, .., b_t, ..., b_T$ sont des bits $\{0, 1\}$ équiprobables, $n_{k,t}$ représente la variable indicatrice $\{0, 1\}$ d'erreur sur le bit b_t à la sortie du canal de transmission k, $\oplus$ représente la somme modulo 2 (ou le ou exclusif) et où $(b_t, n_{k,t})_{k=1,...,K,\ t=1,...,T}$ sont indépendants avec $P(n_{k,t} = 1) = p_k$. Ici, l'observation est $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_t, .., \mathbf{x}_T)$ avec $\mathbf{x}_t \stackrel{\text{def}}{=} (x_{1,t}, .x_{k,t}.., x_{K,t})$ et

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{x}_t; \boldsymbol{\theta}) \;\; \text{où } \boldsymbol{\theta} = (p_1, ..., p_K)$$

avec

$$p(\mathbf{x}_t; \boldsymbol{\theta}) = \frac{1}{2}p(\mathbf{x}_t/b_t = 1) + \frac{1}{2}p(\mathbf{x}_t/b_t = 0) = \frac{1}{2}\prod_{k=1}^{K} p_k^{1-x_{k,t}}(1-p_k)^{x_{k,t}} + \frac{1}{2}\prod_{k=1}^{K} p_k^{x_{k,t}}(1-p_k)^{1-x_{k,t}}.$$

4. Somme de q sinusoïdes bruitées par du bruit gaussien e_k centré i.i.d. de loi de probabilité gaussienne de variance σ^2.

$$x_k = \sum_{j=1}^{q} a_j \cos(2\pi k f_j + \phi_j) + e_k, \quad k = 1, ..., n,$$

pour laquelle

$$L_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{k=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(x_k - \sum_{j=1}^{q} a_j \cos(2\pi k f_j + \phi_j)\right)^2}{2\sigma^2}} \text{où} \;\; \boldsymbol{\theta} = (a_1, ..., a_q, f_1, ..., f_q, \phi_1, ..., \phi_q, \sigma^2).$$

Nous constatons que ces quatre expressions de la vraisemblance $L_{\mathbf{x}}(\boldsymbol{\theta})$ se simplifieraient si *des données cachées étaient observables* et si les *données observées* $\mathbf{x}$ étaient remplacées par des *données complètes* $\mathbf{y}$ que nous définissons dans les quatre exemples précédents par :

1. Pour chaque k, l'indice $w_k \in \{1, ..., q\}$, (pour $q = 2$, on préfère $w_k \in \{0, 1\}$) du mélange observé j est une donnée cachée. D'où les données cachées :

$$\text{ensemble des indices de mélange } \mathbf{w} = (w_1, ..., w_n)$$

2. Le processus de Markov $\mathbf{z} = (z_0, z_1, ..., z_n)$ constitue les données cachées.
3. Données cachées :
$$\text{bits émis } \mathbf{b} = (b_1, ..., b_T),$$
4. Données cachées artificielles :
$$y_{j,k} = a_j \cos(2\pi k f_j + \phi_j) + e_{j,k}$$

où $(e_{j,k})_{j=1,...,q,k=1,...,n}$ sont des variables aléatoires indépendantes de loi de probabilité gaussienne $\mathcal{N}\left(0, \frac{\sigma^2}{q}\right)$ avec $\sum_{j=1}^{q} e_{j,k} = e_k$, ce qui implique $x_k = \sum_{j=1}^{q} y_{k,j}$ pour $k = 1, ..., n$.

Dans les trois premiers exemples, les données complètes $\mathbf{y}$ sont obtenues en ajoutant aux données observées $\mathbf{x}$, des *données cachées* respectives $\mathbf{w}$, $\mathbf{z}$ et $\mathbf{b}$. Au contraire, dans le quatrième exemple, ces données complètes sont artificielles, en particulier quant au choix de la répartition du bruit $e_{j,k}$ entre les q sinusoïdes. Nous obtenons les ensembles de données complètes :

1. $\mathbf{y} = (\mathbf{x}, \mathbf{w})$,
2. $\mathbf{y} = (\mathbf{x}, \mathbf{z})$,
3. $\mathbf{y} = (\mathbf{X}, \mathbf{b})$ où $\mathbf{X} \stackrel{\text{def}}{=} (\mathbf{x}_1, ..., \mathbf{x}_T)$,
4. $\mathbf{Y} = (\mathbf{y}_1, ..., \mathbf{y}_q)$ avec $\mathbf{y}_j \stackrel{\text{def}}{=} (y_{j,1}, ..y_{j,k}, ..., y_{j,n})^T$.

Dans chacun de ces quatre exemples, nous pouvons reconstituer les données observées $\mathbf{x}$ à partir des données complètes $\mathbf{y}$. Nous avons :

$$\underbrace{\mathbf{y}}_{\text{données complètes}} \in \mathcal{Y} \mapsto \underbrace{\mathbf{x}}_{\text{données observées}} = g(\mathbf{y}) \in \mathcal{X} \tag{2.18}$$

où g est une fonction non injective.

Les données observées sont aussi appelées dans ce contexte *données incomplètes*. Mais d'une façon plus générale, la méthodologie de l'algorithme EM s'étend au cas où

La loi conditionnelle $p(\mathbf{x}/\mathbf{y})$ ne dépend pas du paramètre $\boldsymbol{\theta}$.

2.6.2 Méthode EM

La méthode EM consiste à remplacer la log-vraisemblance [6] $l_{\mathbf{x}}(\boldsymbol{\theta}) = \ln p(\mathbf{x}; \boldsymbol{\theta})$ des données observées par celles des données complètes $\ln p(\mathbf{y}; \boldsymbol{\theta})$. Mais comme les données complètes $\mathbf{y}$ ne sont pas toutes observables, on remplace cette vraisemblance $\ln p(\mathbf{y}; \boldsymbol{\theta})$ par son approximation $\mathrm{E}[\ln p(\mathbf{y}; \boldsymbol{\theta})/\mathbf{x}]$ espérance conditionnelle sachant l'observation $\mathbf{x}$ qui devient alors une fonction de $\mathbf{x}$ et de $\boldsymbol{\theta}$. Mais comme, nous avons besoin de connaître $\boldsymbol{\theta}$ pour calculer cette espérance conditionnelle, la méthode EM consiste à utiliser la procédure itérative suivante :

1. Initialisation : choix de $\boldsymbol{\theta}_0$
2. **Etape Expectation**
$$\text{Calcul de } \; Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i) \stackrel{\text{def}}{=} \mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y}; \boldsymbol{\theta})/\mathbf{x}] = \int_{\mathcal{Y}} \ln p(\mathbf{y}; \boldsymbol{\theta}) p(\mathbf{y}; \boldsymbol{\theta}_i/\mathbf{x}) d\mathbf{y}. \tag{2.19}$$
3. **Etape Maximization**
$$\text{Recherche de maximum } \; \boldsymbol{\theta}_{i+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i). \tag{2.20}$$

6. La log-vraisemblance est préférée à la vraisemblance pour des simplifications de calcul dans le cadre de la famille exponentielle et pour des raisons plus conceptuelles liées à l'information de Fisher.

2.6.3 Méthode EM dans le contexte bayesien

La méthodologie EM s'étend dans le contexte bayésien pour la recherche de l'*estimateur du maximum a postériori* (dit estimateur MAP)

$$\widehat{\boldsymbol{\theta}}_{\mathrm{MAP}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}/\mathbf{x}).$$

L'étape d'*Expectation* devient

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\boldsymbol{\theta}/\mathbf{y})/\mathbf{x}] &= \mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y}/\boldsymbol{\theta})/\mathbf{x}] + \ln p(\boldsymbol{\theta}) - \mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y})/\mathbf{x}] \\ &= \underbrace{\mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y}/\boldsymbol{\theta})/\mathbf{x}]}_{Q(\boldsymbol{\theta},\boldsymbol{\theta}_i)} + \ln p(\boldsymbol{\theta}) + \mathrm{Cte}, \end{aligned} \tag{2.21}$$

en utilisant la relation de Bayes : $p(\boldsymbol{\theta}/\mathbf{y}) = \frac{p(\mathbf{y}/\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$ et en remarquant que dans (2.21), $\boldsymbol{\theta}$ joue le rôle d'un paramètre fixe et que $\mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y})/\mathbf{x}]$ ne dépend pas de $\boldsymbol{\theta}$ car dans le contexte bayesien, $p(\mathbf{y})$ représente la loi marginale de $\mathbf{y}$.

Puis l'étape de *Maximization* devient

$$\boldsymbol{\theta}_{i+1} = \arg\max_{\boldsymbol{\theta}} \left(Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i) + \ln p(\boldsymbol{\theta}) \right).$$

2.6.4 Propriétés élémentaires de la méthode EM

La plus importante propriété de l'algorithme EM est sous des hypothèses techniques très générales (dont le support $\mathcal{Y}$ de la distribution de la variable aléatoire $\mathbf{y}$ ne dépend pas de $\boldsymbol{\theta}$) que la suite des valeurs $\boldsymbol{\theta}_i$ augmente la vraisemblance à chaque itération :

$$L_{\mathbf{x}}(\boldsymbol{\theta}_{i+1}) \geq L_{\mathbf{x}}(\boldsymbol{\theta}_i). \tag{2.22}$$

Plus généralement, la monotonie (2.22) est assurée si seulement $Q(\boldsymbol{\theta}_{i+1}, \boldsymbol{\theta}_i) \geq Q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i)$. Ce qui n'impose pas que $\boldsymbol{\theta}_{i+1}$ maximise la fonction $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i)$, mais seulement augmente la valeur de cette fonction par rapport à $\boldsymbol{\theta}_i$. Cette propriété est importante quand il n'existe pas de solution exacte à cette maximisation.

Preuve :

$$\begin{aligned} \ln p(\mathbf{x};\boldsymbol{\theta}) &= \ln \int_{\mathcal{Y}} p(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) d\mathbf{y}, \text{ d'après la formule des probabilités totales} \\ &= \ln \int_{\mathcal{Y}} \frac{p(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}_i)} p(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}_i) d\mathbf{y} \\ &= \ln \mathrm{E}_{\boldsymbol{\theta}_i}\left[\frac{p(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}_i)} / \mathbf{x} \right] \\ &\geq \mathrm{E}_{\boldsymbol{\theta}_i}\left[\ln \frac{p(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y}/\mathbf{x};\boldsymbol{\theta}_i)} / \mathbf{x} \right], \text{ d'après l'inégalité de Jansen car la fonction ln est concave} \\ &= \mathrm{E}_{\boldsymbol{\theta}_i}\left[\ln \frac{p(\mathbf{y};\boldsymbol{\theta}) p(\mathbf{x}/\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{y};\boldsymbol{\theta}_i) p(\mathbf{x}/\mathbf{y};\boldsymbol{\theta}_i)/p(\mathbf{x};\boldsymbol{\theta}_i)} / \mathbf{x} \right], \text{ par la règle de Bayes} \\ &= \mathrm{E}_{\boldsymbol{\theta}_i}\left[\ln \frac{p(\mathbf{y};\boldsymbol{\theta}) p(\mathbf{x}/\mathbf{y})}{p(\mathbf{y};\boldsymbol{\theta}_i) p(\mathbf{x}/\mathbf{y})/p(\mathbf{x};\boldsymbol{\theta}_i)} / \mathbf{x} \right], \text{ parce que la densité de probabilité } p(\mathbf{x}/\mathbf{y}) \text{ ne dépend pas de } \boldsymbol{\theta} \\ &= \mathrm{E}_{\boldsymbol{\theta}_i}\left[\ln \frac{p(\mathbf{y};\boldsymbol{\theta}) p(\mathbf{x};\boldsymbol{\theta}_i)}{p(\mathbf{y};\boldsymbol{\theta}_i)} / \mathbf{x} \right] \\ &= \mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y};\boldsymbol{\theta})/\mathbf{x}] - \mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y};\boldsymbol{\theta}_i)/\mathbf{x}] + \ln p(\mathbf{x};\boldsymbol{\theta}_i) \\ &= Q(\boldsymbol{\theta},\boldsymbol{\theta}_i) - Q(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i) + \ln p(\mathbf{x};\boldsymbol{\theta}_i). \end{aligned}$$

soit : $\ln p(\mathbf{x};\boldsymbol{\theta}) \geq \ln p(\mathbf{x};\boldsymbol{\theta}_i) + [Q(\boldsymbol{\theta},\boldsymbol{\theta}_i) - Q(\boldsymbol{\theta}_i,\boldsymbol{\theta}_i)] \quad \Rightarrow \quad \ln p(\mathbf{x};\boldsymbol{\theta}_{i+1}) \geq \ln p(\mathbf{x};\boldsymbol{\theta}_i).$ ■

De plus, sous certaines conditions de régularité, la suite $\boldsymbol{\theta}_i$ converge vers un point stationnaire (c'est à dire ici un maximum local ou un point selle) de la vraisemblance $L_{\mathbf{x}}(\boldsymbol{\theta})$. Mais cette monotonie et cette convergence vers un point stationnaire n'assure pas la convergence de la suite $\boldsymbol{\theta}_i$ vers l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x})$. La convergence de la suite $\boldsymbol{\theta}_i$ dépendant en général de la condition initiale $\boldsymbol{\theta}_0$, son choix au voisinage de $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x})$ est souvent crucial.

2.6.5 Application au mélange de deux distributions gaussiennes

Nous allons détailler ici les 2 étapes *Expectation* et *Maximization* de l'algorithme EM en revenant quelquefois pour des raisons pédagogiques aux notations usuelles de probabilités et de densité de probabilité pour les variables aléatoires respectivement discrètes et continues. Nous avons ici : [7]

$\mathbf{y} = (\mathbf{x}, \mathbf{w})$ avec $\mathbf{x} = (x_1, .., x_n)$ et $\mathbf{w} = (w_1, .., w_n)$, où $w_k \in \{0, 1\}$ indicatrice de la classe 1 de x_k

$$f_1(x_k) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x_k - m_1)^2}{2\sigma_1^2}} \text{ et } f_2(x_k) = \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{(x_k - m_2)^2}{2\sigma_2^2}}$$

représentent les densités de probabilité de l'observation x_k conditionnelles respectivement à $w_k = 1$ et $w_k = 0$. Ici :

$$\boldsymbol{\theta} = (p_1, m_1, m_2, \sigma_1^2, \sigma_2^2).$$

$$\begin{aligned} p(\mathbf{y};\boldsymbol{\theta}) &= \prod_{k=1}^{n} p(y_k;\boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k, w_k;\boldsymbol{\theta}) = \prod_{k=1}^{n} p(x_k/w_k;\boldsymbol{\theta}) p(w_k;\boldsymbol{\theta}) \\ &= \prod_{k=1}^{n} \left([p_1 f_1(x_k)]^{w_k} [(1-p_1) f_2(x_k)]^{1-w_k} \right), \end{aligned}$$

car $p(x_k/w_k;\boldsymbol{\theta}) p(w_k;\boldsymbol{\theta}) = \begin{cases} p_1 f_1(x_k) & \text{si } w_k = 1 \\ (1-p_1) f_2(x_k) & \text{si } w_k = 0 \end{cases}$.

Par suite :

$$\ln p(\mathbf{y};\boldsymbol{\theta}) = \sum_{k=1}^{n} w_k \left(\ln p_1 + \ln f_1(x_k) \right) + (1 - w_k) \left(\ln(1-p_1) + \ln f_2(x_k) \right).$$

En appliquant l'étape d'*Expectation*. On obtient :

$$\mathrm{E}_{\boldsymbol{\theta}_i}[\ln p(\mathbf{y};\boldsymbol{\theta})/\mathbf{x}] = \sum_{k=1}^{n} \mathrm{E}_{\boldsymbol{\theta}_i}[w_k/\mathbf{x}] \left(\ln p_1 + \ln f_1(x_k) \right) + (1 - \mathrm{E}_{\boldsymbol{\theta}_i}[w_k/\mathbf{x}]) \left(\ln(1-p_1) + \ln f_2(x_k) \right),$$

où nous avons appliqué le résultat suivant de probabilité :

Résultat de probabilité 2.2 : Pour toute variable aléatoire $g(X, Y)$:

$$\mathrm{E}[g(X,Y)/X = x] = \mathrm{E}[g(x,Y)/X = x].$$

7. Notons qu'ici nous avons utilisé pour faciliter la compréhension la double notation : densité de probabilité au sens de Radon Nikodym, et densité de probabilité et probabilités pour des variables aléatoires respectivement continues et discrètes.

Or

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}_i}[w_k/\mathbf{x}] &= P(w_k = 1; \boldsymbol{\theta}_i/\mathbf{x}) = \frac{p(w_k = 1, \mathbf{x}; \boldsymbol{\theta}_i)}{p(\mathbf{x}; \boldsymbol{\theta}_i)} \\
&= \frac{p(x_1; \boldsymbol{\theta}_i)..p(x_{k-1}; \boldsymbol{\theta}_i)p(w_k = 1, x_k; \boldsymbol{\theta}_i)p(x_{k+1}; \boldsymbol{\theta}_i)..p(x_n; \boldsymbol{\theta}_i)}{p(x_1; \boldsymbol{\theta}_i)..p(x_k; \boldsymbol{\theta}_i)..p(x_n; \boldsymbol{\theta}_i)} \qquad (2.23) \\
&= \frac{p(w_k = 1, x_k; \boldsymbol{\theta}_i)}{p(x_k; \boldsymbol{\theta}_i)} = P(w_k = 1; \boldsymbol{\theta}_i/x_k) \\
&= \frac{P_{\boldsymbol{\theta}_i}(w_k = 1) f_1(x_k; \boldsymbol{\theta}_i)}{f_{X_k}(x_k; \boldsymbol{\theta}_i)} \qquad (2.24) \\
&= \frac{p_{1,i} f_1(x_k; \boldsymbol{\theta}_i)}{p_{1,i} f_1(x_k; \boldsymbol{\theta}_i) + (1 - p_{1,i}) f_2(x_k; \boldsymbol{\theta}_i)} \stackrel{\text{def}}{=} p_1^{x_k,i}, \qquad (2.25)
\end{aligned}$$

où nous avons appliqué en (2.23), l'indépendance des variables aléatoires $\{x_1, (w_k, x_k), .., x_n\}$ et avons utilisé le résultat suivant de probabilité en (2.24) et (2.25)

Résultat de probabilité 2.3 : La *relation de Bayes* et la *formule des probabilités totales* s'étendent dans le cas où X et Y sont des variables aléatoires respectivement discrètes (prenant les valeurs $\{a_i\}_{i \in I}$) et continues de densité de probabilité $f_Y(y)$

$$P(X = a_i / Y = y) = \frac{P(X = a_i) f_{Y/X=a_i}(y)}{f_Y(y)}$$

$$f_Y(y) = \sum_{i \in I} P(X = a_i) f_{Y/X=a_i}(y).$$

Par suite, nous obtenons

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i) = \sum_{k=1}^{n} p_1^{x_k,i} \left(\ln p_1 + \ln f_1(x_k)\right) + (1 - p_1^{x_k,i}) \left(\ln(1 - p_1) + \ln f_2(x_k)\right). \qquad (2.26)$$

Puis nous pouvons mener l'étape de *Maximization*. $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i)$ est maximisée par rapport à $\boldsymbol{\theta}$ pour

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i)}{\partial p_1} = \sum_{k=1}^{n} p_1^{x_k,i} \left(\frac{1}{p_1}\right) - \sum_{k=1}^{n} (1 - p_1^{x_k,i}) \left(\frac{1}{1 - p_1}\right) = 0,$$

soit pour

$$p_{1,i+1} = \frac{1}{n} \sum_{k=1}^{n} p_1^{x_k,i}, \qquad (2.27)$$

où $p_1^{x_k,i}$ est donnée par (2.25). $p_{1,i+1}$ peut s'interpréter comme la *moyenne empirique des probabilités a postériori à l'itération i de la classe 1* à partir des observations x_k. De même pour $m_j, j = 1, 2$:

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i)}{\partial m_j} = \sum_{k=1}^{n} p_1^{x_k,i} \frac{1}{f_1(x_k)} \frac{\partial f_1(x_k)}{\partial m_j} + \sum_{k=1}^{n} (1 - p_1^{x_k,i}) \frac{1}{f_2(x_k)} \frac{\partial f_2(x_k)}{\partial m_j} = 0,$$

avec $\frac{1}{f_j(x_k)} \frac{\partial f_j(x_k)}{\partial m_j} = \frac{x_k - m_j}{\sigma_j^2}, j = 1, 2$. Soit

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i)}{\partial m_j} = \sum_{k=1}^{n} p_j^{x_k,i} \left(\frac{x_k - m_j}{\sigma_j^2}\right) = 0, \quad \text{avec} \quad p_2^{x_k,i} \stackrel{\text{def}}{=} (1 - p_1^{x_k,i}),$$

soit pour

$$m_{j,i+1} = \frac{\sum_{k=1}^{n} p_j^{x_k,i} x_k}{\sum_{k=1}^{n} p_j^{x_k,i}}, j = 1, 2, \tag{2.28}$$

qui peut s'interpréter comme la *moyenne empirique des espérances mathématiques a postériori à l'itération i de la classe j*, $j = 1, 2$. De même pour $\sigma_j^2, j = 1, 2$:

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}_i)}{\partial \sigma_j^2} = \sum_{k=1}^{n} p_1^{x_k,i} \frac{1}{f_1(x_k)} \frac{\partial f_1(x_k)}{\partial \sigma_j^2} + \sum_{k=1}^{n} (1 - p_1^{x_k,i}) \frac{1}{f_2(x_k)} \frac{\partial f_2(x_k)}{\partial \sigma_j^2} = 0$$

avec $\frac{1}{f_j(x_k)} \frac{\partial f_j(x_k)}{\partial \sigma_j^2} = -\frac{1}{2\sigma_j^2} + \frac{(x_k - m_j)^2}{2\sigma_j^4}, j = 1, 2$, qui donne

$$\sigma_{j,i+1}^2 = \frac{\sum_{k=1}^{n} p_j^{x_k,i} (x_k - m_{j,i+1})^2}{\sum_{k=1}^{n} p_j^{x_k,i}}, j = 1, 2, \tag{2.29}$$

où $\sigma_{j,i}^2$ peut s'interpréter comme la *moyenne empirique des variances a postériori à l'itération i de la classe j*, $j = 1, 2$.

Dans cet exemple, les deux équations (2.19) et (2.20) ont des solutions analytiques explicites (2.26) et (2.27),(2.28),(2.29). Ceci s'étend à des mélanges d'un nombre quelconques q de loi de probabilité de la famille exponentielle (bien entendu aussi bien scalaire que multidimensionnelle).

Remarque 2.6 : Dans le cas où les lois de probabilité du mélange de q lois sont parfaitement connues, seules les probabilités $(p_j)_{j=1,..,q}$ du mélange de loi sont des paramètres à estimer. Alors l'algorithme EM se résume aux itérations suivantes :

1.
$$\boldsymbol{\theta}_0 = (p_{1,0}, ..., p_{q,0}),$$

2.
$$p_j^{x_k,i} = \frac{p_{j,i} f_j(x_k)}{\sum_{\ell=1}^{q} p_{\ell,i} f_\ell(x_k)}, j = 1, ..., q$$

3.
$$p_{j,i+1} = \frac{1}{n} \sum_{k=1}^{n} p_j^{x_k,i}, j = 1, ..., q.$$

Remarque 2.7 : L'initialisation de l'algorithme EM peut se faire dans ce dernier cas par simple résolution du système d'équations linéaires de la méthode des q premiers moments du mélange (voir le paragraphe 3.2.2 :

$$\sum_{j=1}^{q} p_{j,0} \mu_{\ell,j} = \frac{1}{n} \sum_{k=1}^{n} x_k^\ell, \ \ell = 0, .., q - 1$$

où $\mu_{\ell,j} \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} x^\ell f_j(x) dx$.

Remarque 2.8 : L'initialisation de l'algorithme EM dans le cadre général du mélange de q loi de variables aléatoires i.i.d. scalaires ou multidimensionnelles se fait en général par une méthode empirique de résolution du *k-means clustering* qui consiste à regrouper les n observations $x_1, ..., x_k, ..x_n$ en q classes $(S_j)_{j=1,..,q}$ de façon à minimiser [8] la distance interne à chaque classe entre chacun de ses éléments et une valeur m_j,

8. Il existe dans la littérature *machine learning* des algorithmes heuristiques pour approcher cette minimisation.

j=1,...,q. Les q classes S_j sont obtenues par la minimisation

$$\min_{S_1,..,S_q,m_1,..,m_q} \sum_{j=1}^{q} \sum_{x_k \in S_j} (x_k - m_j)^2.$$

Les centres de gravité $m_{j,0} = \frac{1}{\text{card}(S_j)} \sum_{x_k \in S_j} x_k$ de chaque classe S_j obtenue, les dispersions moyennes $\sigma^2_{j,0} = \frac{1}{\text{card}(S_j)} \sum_{x_k \in S_j} (x_k - m_j)^2$ dans chaque classe et les proportions $p_{j,0} = \frac{\text{card}(S_j)}{n}$ de chaque classe S_j fournissent des initialisations respectives des paramètres $m_1,..,m_q$, $\sigma^2_1,..,\sigma^2_q$, $p_1,..,p_q$.

2.7 Points essentiels du chapitre estimateur du maximum de vraisemblance (MV)

- La définition (2.38) de l'estimateur du maximum de vraisemblance (MV) qui dans le cas où $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ est défini par un ensemble de contraintes se réduit à un problème d'optimisation sous contrainte de la fonction de vraisemblance $p(\mathbf{x}; \boldsymbol{\theta})$.
- Pour une observation fixée $\mathbf{x}$, l'estimateur MV :
  - n'existe pas nécessairement ;
  - n'est pas nécessairement unique ;
  - n'est pas nécessairement sans biais ;
  - se confond avec l'estimateur efficace si celui-ci existe ;
  - est une fonction de toute statistique exhaustive.
- Le principe d'invariance à l'aide de la vraisemblance induite (2.6)
- La famille exponentielle (2.42) de loi de probabilité permet la construction d'estimateurs MV sans biais dont la variance/covariance atteint la borne de Cramer Rao (estimateur dit *efficace*) pour une paramétrisation particulière.
- Dans le cas d'une séquence d'observations $\mathbf{x}_n = (x_1, ..., x_n)$ où $(x_k)_{k=1,...,n}$ sont indépendantes et identiquement distribuées, sous certaines conditions techniques sur l'ensemble Θ et sur la fonction de vraisemblance $p(\mathbf{x}; \boldsymbol{\theta})$, il existe une solution $\widehat{\boldsymbol{\theta}}_{\mathrm{MV}}(\mathbf{x}_n)$ des équations de vraisemblance (2.3) qui satisfait les propriétés suivantes :
  - Biais asymptotique en $1/n$;
  - Convergence presque sure vers $\boldsymbol{\theta}$;
  - Distribution asymptotique gaussienne avec une vitesse de convergence en $\sqrt{n}$;
  - Efficacité asymptotique.
- Dans un grand nombre de cas pratiques où l'observation $\mathbf{x} = (x_1, ..., x_n)$ est consituée de variables aléatoires indépendantes et équidistribuées, le calcul de l'estimateur du maximum de vraisemblance est effectué par des méthodes itératives comme la *méthode du score* dérivée de la *méthode de Newton-Raphson*.
- Certains modèles paramétriques peuvent s'interpréter à l'aide de *données cachées* ou *variables latentes*. Dans ce cadre, il existe une méthode dite EM pour *Expectation/Maximisation* qui est une approche itérative qui augmente la vraisemblance à chaque itération. Sous certaines conditions de régularité du modèle paramétrique, cette suite $\boldsymbol{\theta}_i$ converge vers un maximum local de la fonction de vraisemblance. D'où l'importance du choix de la condition initiale.

3 Estimateur par méthode de substitution (ou des moments)

3.1 Introduction

La méthode du maximum de vraisemblance, en dépit de son attrait théorique du à ses bonnes propriétés asymptotiques est souvent difficile à implémenter. Des solutions analytiques sont rarement possibles et les algorithmes disponibles peuvent n'exhiber que des maximums locaux de la fonction de vraisemblance.

L'idee de base de la méthode de substitution est de baser la construction de l'estimateur sur une statistique $\mathbf{s}_n(\mathbf{x}_n)$ de dimension fixée indépendamment du nombre n d'observations $\mathbf{x}_n \stackrel{\text{def}}{=} (x_1, .., x_k, .., x_n)$, qui peut ne pas être exhaustive tout en conservant encore la plus grande quantité d'information sur le paramètre $\boldsymbol{\theta} \in \Theta$. Cette méthode doit son succès à sa grande facilité d'utilisation. Dans tous les cas et indépendamment de la complexité du modèle paramétrique sous-jacent $\mathcal{P} = \{P_{\boldsymbol{\theta}}(\mathbf{x}_n), \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m\}$, elle permet d'obtenir des estimateurs calculables.

Cette méthode s'appuie essentiellement sur une suite de statistiques $\mathbf{s}_n(\mathbf{x}_n) \in \mathbb{R}^q$ de l'observation $\mathbf{x}_n$ qui converge de façon *consistante (faible, forte ou en moyenne quadratique)* vers une constante qui dépend de la distribution sous-jacente de l'observation $\mathbf{x}_n$, donc qui est une fonction $\mathbf{s}(\boldsymbol{\theta})$ du paramètre. [9]

Dans le cas d'observations indépendantes identiquement distribuées (i.i.d.) scalaires $(x_k)_{k=1,...,n}$ ou multidimensionnelles $(\mathbf{x}_k)_{k=1,...,n}$ avec $\mathbf{x}_k = (x_{k,1}, ..., x_{k,r})^T \in \mathbb{R}^r$, les fonctions $\mathbf{s}(\boldsymbol{\theta})$ les plus souvent utilisées sont des vecteurs de *moments*, de composantes

$$\mathrm{E}(x_1^\ell), \ell = 1, ..., q \quad \text{(cas scalaire)}$$

et

$$\mathrm{E}(x_{1,j}^\ell), \ell = 1, ..., q', j = 1, ..., r \quad \text{où} \quad q = q'r \quad \text{(cas multidimensionnel)}.$$

Dans le cas d'observations i.i.d. $(\mathbf{x}_k)_{k=1,...,n}$ multidimensionnelles, on pourra n'utiliser dans certaines applications (par exemple dans le cas où la distribution de $(\mathbf{x}_k)_{k=1,...n}$ est gaussienne) que les moments du premier et second ordre

$$\mathbf{s}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{m}(\boldsymbol{\theta}) \\ \mathrm{vec}(\mathbf{R}(\boldsymbol{\theta})) \end{bmatrix} \quad \text{où} \quad \mathbf{m}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathrm{E}(\mathbf{x}_k) \quad \text{et} \quad \mathbf{R}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathrm{E}(\mathbf{x}_k \mathbf{x}_k^T).$$

Pour cette raison la méthode de substitution est souvent aussi appelé *méthode des moments*. On préfère aussi quelquefois remplacer ces vecteurs de moments par des vecteurs de *cumulants* (dont le plus simple est la variance).

Dans ce cadre d'observations i.i.d. scalaires ou multidimensionnelles, on pourra aussi utiliser des vecteurs de *moments généralisés*, de composantes

$$\mathrm{E}(h_\ell(x_1)), \ell = 1, ..., q \quad \text{(cas scalaire)}$$

ou

$$\mathrm{E}(h_\ell(x_{1,j})), \ell = 1, ..., q', j = 1, ..., r \text{ où } \quad q = q'r \quad \text{(cas multidimensionnel)}$$

avec par exemple :

$$h_\ell(x) = e^{iu_\ell x} \quad \text{ou encore} \quad h_\ell(x) = \mathbb{1}_{C_\ell}(x) \quad \text{avec} \quad C_\ell =]-\infty, c_\ell] \quad \text{ou} \quad C_\ell =]c_\ell, c_{\ell+1}].$$

Cette méthode des moments peut s'étendre dans le cas d'observations $(x_k)_{k=1,...,n}$ non i.i.d. Par exemple, dans le cas d'observations $(x_k)_{k=1,...,n}$ stationnaires centrées du second ordre, on pourra uti-

9. Notons que dans le cadre d'un modèle semi-paramétrique, cette suite $\mathbf{s}_n(\mathbf{x}_n)$ pourra aussi converger vers un fonction $\mathbf{s}(\boldsymbol{\theta})$ du paramètre.

liser un ensemble fini de q covariances

$$\{\mathrm{E}(x_1^2), \mathrm{E}(x_1 x_2), ..., \mathrm{E}(x_1 x_q)\}.$$

L'idée de base de cette méthode consiste alors à choisir un couple $[\mathbf{s}(\boldsymbol{\theta}), \mathbf{s}_n(\mathbf{x}_n)]$ telle que $\mathbf{s}(\boldsymbol{\theta})$ caractérise $\boldsymbol{\theta}$ et à "égaler" l'expression de

$$\mathbf{s}(\boldsymbol{\theta}) = \begin{bmatrix} \mathrm{E}(h_1(x_1)) \\ \vdots \\ \mathrm{E}(h_q(x_1)) \end{bmatrix}$$

avec la statistique consistante

$$\mathbf{s}_n(\mathbf{x}_n) = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^n h_1(x_k) \\ \vdots \\ \frac{1}{n}\sum_{k=1}^n h_q(x_k) \end{bmatrix},$$

dont on déduit par "inversion" de la fonction $\mathbf{s}(\boldsymbol{\theta})$ différentes estimées possibles $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ de $\boldsymbol{\theta}$, car cette "inversion" n'est en général pas unique comme cela sera illustré dans les exemples des sections 3.2.2 et 3.2.3.

Puisqu'en general il y a perte d'information en passant de l'observation $(x_k)_{k=1,...,n}$ à la statistique $\mathbf{s}_n(\mathbf{x}_n)$ constituée d'estimées empiriques de moments ou de moments généralisés, la méthode de substitution fournit en general des estimateurs qui ne sont ni efficaces, ni même asymptotiquement efficaces. Mais nous verrons que la consistance des estimateurs des moments se transmet aux estimateurs $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$.

3.2 Exemples introductifs

3.2.1 Estimation du paramètre d'un échantillon i.i.d. de loi exponentielle

Considérons une observation $\mathbf{x}_n$ constituée de variables aléatoires i.i.d. $(x_1, ..., x_n)$ de loi de probabilité exponentielle, i.e., de densité de probabilité

$$p(x;\theta) = \theta e^{-\theta x} \mathbb{1}_{[0,+\infty[}(x) \quad \text{où} \quad \theta \in]0, +\infty[.$$

Un calcul simple donne pour $s(\theta) = \mathrm{E}(x_k)$

$$s(\theta) = \frac{1}{\theta} \in]0, +\infty[.$$

D'autre part, d'après la loi forte des grands nombres, la suite de statistiques

$$s_n(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k \in]0, +\infty[$$

converge presque sûrement vers le moment $s(\theta) = \frac{1}{\theta}$. Elle converge d'ailleurs aussi en moyenne quadratique. La méthode de substitution nous donne ainsi par inversion immédiate l'estimateur :

$$\widehat{\theta}_n(\mathbf{x}_n) = \frac{1}{\frac{1}{n}\sum_{k=1}^n x_k}.$$

Une question simple survient : a-t-on intérêt à utiliser plutôt un vecteur $\mathbf{s}_n(\mathbf{x}_n)$ de statistiques de composantes $\left(\frac{1}{n}\sum_{k=1}^n x_k^\ell\right)_{\ell=1,..,q}$ qui converge aussi presque sûrement par la loi des grands nombres vers le vecteur $\mathbf{s}(\boldsymbol{\theta})$ de composantes

$$\mathrm{E}(x_k^\ell) = \frac{\ell\,!}{\theta^\ell},$$

plutôt que la simple statistique du premier ordre $s_n(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k$. La réponse est non puisque la statistique $s_n(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k$ est exhaustive. [10]

3.2.2 Estimation de la proportion d'un échantillon i.i.d. de loi de mélanges

Considérons une observation $\mathbf{x}_n$ constituée de variables aléatoires i.i.d. $(x_1, ..., x_n)$ d'un mélange de deux lois de probabilité connues $p_1(x)$ et $p_2(x)$, i.e., de densité de probabilité

$$p(x;\theta) = \theta p_1(x) + (1-\theta)p_2(x) \quad \text{où} \quad \theta \in [0,1].$$

Ici, contrairement à l'exemple précédent, il n'existe pas d'après le critère de factorisation de Neymann-Fisher de statistique exhaustive. Donc il y aura perte d'information en passant de l'échantillon $(x_1, ..., x_n)$ à n'importe quelle statistique $s_n(\mathbf{x}_n)$ ou vecteur de statistiques $\mathbf{s}_n(\mathbf{x}_n)$ de dimension $q < n$. Nous aurons donc intérêt à prendre en compte une statistique à plusieurs composantes pour limiter la perte d'information. Nous pouvons choisir en particulier :

$$\mathbf{s}_n(\mathbf{x}_n) = \begin{bmatrix} s_{n,1}(\mathbf{x}_n) \\ . \\ s_{n,q}(\mathbf{x}_n) \end{bmatrix} \text{ avec } \quad s_{n,\ell}(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n x_k^\ell \text{ ou } \quad s_{n,\ell}(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{C_\ell}(x_k) \quad \ell = 1, ..., q$$

qui d'après la loi forte des grands nombres converge presque sûrement vers le vecteur des moments :

$$\mathbf{s}(\theta) = \begin{bmatrix} s_1(\theta) \\ . \\ s_q(\theta) \end{bmatrix} \text{ avec } \quad s_\ell(\theta) = \mathrm{E}(x^\ell) \quad \text{ou} \quad s_\ell(\theta) = P(x \in C_\ell), \quad \ell = 1, ..., q$$

car $\mathrm{E}(\mathbb{1}_{C_\ell}(x)) = P(x \in C_\ell)$. $\mathbf{s}(\theta)$ satisfait, d'après le théorème de transfert, la relation :

$$\mathbf{s}(\theta) = \theta\boldsymbol{\mu}_1 + (1-\theta)\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2 + \theta(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{3.1}$$

avec $\boldsymbol{\mu}_i = (\mu_i^1, ..., \mu_i^q)^T$, $i = 1, 2$ avec $\mu_i^\ell = \int x^\ell p_i(x)dx$ (moment d'ordre ℓ de la loi i du mélange) ou $\mu_i^\ell = \int_{C_\ell} p_i(x)dx$ (en particulier, probabilité que la composante i du mélange appartienne à l'intervalle $[c_{\ell-1}, c_\ell[$ avec $c_0 = -\infty$ et $c_q = +\infty$).

Contrairement à la relation $s(\theta) = \frac{1}{\theta}$ précédente, la relation (3.1) présente si $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, une infinité de fonctions g définies sur $\mathbb{R}^q$ qui vérifient

$$\mathbf{s} = \mathbf{s}(\theta) \in \mathbf{s}(\Theta) \subset \mathbb{R}^q \mapsto \theta = g(\mathbf{s}) \in [0,1].$$

On pourrait en particulier, par exemple si $\mu_1^\ell \neq \mu_2^\ell$ ne considérer que la statistique $s_{n,\ell}(\mathbf{x}_n)$ associée au seul moment $s_\ell(\theta)$ et l'on obtiendrait

$$\theta = g(\mathbf{s}) \stackrel{\text{def}}{=} \frac{s_\ell - \mu_2^\ell}{\mu_1^\ell - \mu_2^\ell} \tag{3.2}$$

qui donnerait l'estimateur :

$$\widehat{\theta}_n(\mathbf{x}_n) = \frac{s_{n,\ell}(\mathbf{x}_n) - \mu_2^\ell}{\mu_1^\ell - \mu_2^\ell}.$$

10. D'après le critère de factorisation de Neymann-Fisher, une statistique $\mathbf{s}_n(\mathbf{x}_n)$ est exhaustive si et seulement si $p(\mathbf{x}_n;\theta)$ peut se factoriser sous la forme $g[\mathbf{s}_n(\mathbf{x}_n), \boldsymbol{\theta}]h(\mathbf{x}_n)$; ce qui est réalisé ici car : $p(\mathbf{x}_n;\theta) = \theta^n e^{-\theta\sum_{k=1}^n x_k}\prod_{k=1}^n \mathbb{1}_{[0,+\infty[}(x_k)$.

Pour des pondérations α_ℓ quelconques satisfaisant $\sum_{\ell=1}^q \alpha_\ell = 1$, si $\mu_1^\ell \neq \mu_2^\ell$ pour $\ell = 1, ..., q$:

$$\theta = g(\mathbf{s}) \stackrel{\text{def}}{=} \sum_{\ell=1}^{q} \alpha_\ell \frac{s_\ell - \mu_2^\ell}{\mu_1^\ell - \mu_2^\ell}, \tag{3.3}$$

qui utiliserait par contre toutes les composantes de $\mathbf{s}(\mathbf{x}_n)$, la relation (3.3) donnerait la famille d'estimateurs

$$\widehat{\theta}_n(\mathbf{x}_n) = \sum_{\ell=1}^{q} \alpha_\ell \frac{\left(\frac{1}{n}\sum_{k=1}^n x_k^\ell\right) - \mu_2^\ell}{\mu_1^\ell - \mu_2^\ell} \quad \text{ou} \quad \widehat{\theta}_n(\mathbf{x}_n) = \sum_{\ell=1}^{q} \alpha_\ell \frac{\left(\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{C_\ell}(x_k)\right) - \mu_2^\ell}{\mu_1^\ell - \mu_2^\ell}.$$

Mais bien d'autres constructions seraient possibles. Par exemple l'expression (3.1) donne pour $\mathbf{a} \in \mathbb{R}^q$ quelconque, $\mathbf{a}^T\mathbf{s}(\theta) = \theta\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \mathbf{a}^T\boldsymbol{\mu}_2$, dont on déduit :

$$\theta = g(\mathbf{s}) \stackrel{\text{def}}{=} \frac{\mathbf{a}^T(\mathbf{s} - \boldsymbol{\mu}_2)}{\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)},$$

donnerait la nouvelle famille d'estimateurs :

$$\widehat{\theta}_n(\mathbf{x}_n) = \frac{\mathbf{a}^T(\mathbf{s}_n(\mathbf{x}_n) - \boldsymbol{\mu}_2)}{\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}.$$

Face à cette infinie d'estimateurs possibles, le problème de leur comparaison reste à traiter.

3.2.3 Estimation du paramètre dans un modèle AR d'ordre 1

Considérons une observation $(x_1, ..., x_n)$ extraite d'un processus $(x_k)_{k=...,1,...,n,...}$ stationnaire centré (appelé modèle causal AR (Auto Regressive) du premier ordre) satisfaisant la relation suivante

$$x_k = \theta x_{k-1} + u_k,$$

où $\theta \in]-1, 1[$ et où $(u_k)_{k=..,-1,0,+1,...}$ est une suite de variables aléatoires centrées i.i.d. de variance σ^2. Ce processus satisfait aussi la relation

$$x_k = \sum_{\ell=0}^{+\infty} \theta^\ell u_{k-\ell}.$$

Nous supposerons ici que cette variance σ^2 est égale à l'unité. Le seul paramètre inconnu de ce modèle semi-paramétrique est donc $\theta \in]-1, 1[$. Dans cet exemple aucune statistique n'est exhaustive. Nous avons

$$\mathrm{E}(x_k x_{k+\ell}) = \frac{\sigma^2\theta^{|\ell|}}{1-\theta^2} \quad \text{avec} \quad \sigma^2 = 1.$$

Par suite en prenant le seul moment $s_1(\theta) = \mathrm{E}(x_k x_{k+1})$ associé à la statistique $s_{n,1}(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^{n-1} x_k x_{k+1}$, nous obtenons l'inversion unique :

$$s_1 \mapsto \theta = g(s_1) = \frac{-1 + \sqrt{1 + 4s_1^2}}{2s_1},$$

racine unique en θ de module strictement inférieur à 1 de l'équation

$$s_1 = \frac{\theta}{1-\theta^2} \Leftrightarrow s_1\theta^2 + \theta - s_1 = 0 \quad \text{où} \quad \mathrm{sign}(\theta) = \mathrm{sign}(s_1).$$

Par suite nous obtenons l'estimateur :

$$\widehat{\theta}_n^{(1)}(\mathbf{x}_n) = \frac{-1 + \sqrt{1 + 4\left(\frac{\sum_{k=1}^{n-1} x_k x_{k+1}}{n}\right)^2}}{2\frac{\sum_{k=1}^{n-1} x_k x_{k+1}}{n}}.$$

Par contre en utilisant le vecteur constitué de deux moments

$$\mathbf{s}(\theta) = \begin{bmatrix} s_0(\theta) \\ s_1(\theta) \end{bmatrix} = \begin{bmatrix} \mathrm{E}(x_k^2) \\ \mathrm{E}(x_k x_{k+1}) \end{bmatrix},$$

il existe une infinité de fonctions g telles que $\mathbf{s} \mapsto \theta = g(\mathbf{s})$. Par exemple les fonctions g définies par :

$$g_2(\mathbf{s}) = \frac{s_1}{s_0}, \quad \text{et} \quad g_3(\mathbf{s}) = s_1\left(1 - \frac{s_1^2}{s_0^2}\right),$$

aboutissent aux estimateurs :

$$\widehat{\theta}_n^{(2)}(\mathbf{x}_n) = \frac{\sum_{k=1}^{n-1} x_k x_{k+1}}{\sum_{k=1}^{n} x_k^2} \quad \text{et} \quad \widehat{\theta}_n^{(3)}(\mathbf{x}_n) = \left(\frac{1}{n}\sum_{k=1}^{n-1} x_k x_{k+1}\right)\left(1 - \frac{(\sum_{k=1}^{n-1} x_k x_{k+1})^2}{(\sum_{k=1}^{n} x_k^2)^2}\right).$$

Dans le cas général où σ^2 serait aussi un paramètre inconnu, nous voyons que cette dernière inversion, qui est toujours possible conduit au même estimateur.

Remarque 3.1 : Il est important de noter que la méthode de substitution peut dans certains cas ne pas nécessiter la connaissance explicite de la loi de probabilité de l'observation $\mathbf{x}_n$. C'est le cas du modèle semi-paramétriques du dernier exemple. De ce fait la méthode de substitution peut s'appliquer dans certains cas dans un cadre *semi-paramétrique*.

Dans ces trois examples, une forme analytique a pu être donnée à des solutions de l'"inversion"

$$\mathbf{s} = \mathbf{s}(\theta) \mapsto \theta = g(\mathbf{s}),$$

pour $\mathbf{s} \in \cup_{n=1}^{\infty}\{\mathbf{s}_n(\mathbf{x}_n); \mathbf{x}_n \in \mathcal{X}\}$. Malheureusement ces situations sont exceptionnelles, en particulier pour un paramètre $\boldsymbol{\theta}$ multidimensionnel où $\boldsymbol{\theta}$ sera solution implicite d'équations (par exemple argument du maximum d'une fonction, zéros d'une fonction...). Pour traiter de toutes ces situations nous allons aborder une méthodologie générale.

3.3 Problématique générale

On considère un modèle statistique paramétrique dans lequel le paramètre est $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^m$ et une suite de statistiques $\mathbf{s}_n(\mathbf{x}_n)$ de dimension fixée q ($\mathbf{s}_n(\mathbf{x}_n) \in \mathbb{R}^q$) qui converge vers une constante en au moins un des sens habituels (*en probabilité, en moyenne quadratique ou presque sûrement*). Puisque cette convergence dépend de la distribution sous-jacente [11], cette constante définit une fonction $\boldsymbol{\theta} \in \Theta \mapsto \mathbf{s}(\boldsymbol{\theta}) \in \mathbf{s}(\Theta) \subset \mathbb{R}^q$. Soit :

$$\mathbf{s}_n(\mathbf{x}_n) \to \mathbf{s}(\boldsymbol{\theta}) \quad \text{pour} \quad n \to \infty.$$

Nous supposons que q est assez grand pour que $\mathbf{s}(.)$ soit bijective, i.e., réalise une bijection de Θ dans $\mathbf{s}(\Theta)$ (image de Θ par $\mathbf{s}$). Par suite en pratique $q \geq m$.

11. Cette méthode de substitution s'étend aux modèles semi-paramétriques (c'est à dire où $\boldsymbol{\theta}$ ne caractérise pas la distribution de $\mathbf{x}_n$) si cette constante est une fonction de $\boldsymbol{\theta}$.

Cette condition est souvent appelée *condition d'identifiabilité*

$$\mathbf{s}(\boldsymbol{\theta}) = \mathbf{s}(\boldsymbol{\theta}') \quad \Rightarrow \quad \boldsymbol{\theta} = \boldsymbol{\theta}'.$$

Nous supposerons que cette application $\boldsymbol{\theta} \mapsto \mathbf{s}(\boldsymbol{\theta})$ est différentiable dans un voisinage de $\boldsymbol{\theta} \in \Theta$

$$\mathbf{s}(\boldsymbol{\theta} + \delta\boldsymbol{\theta}) = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{S}(\boldsymbol{\theta})\delta\boldsymbol{\theta} + o(\|\delta\boldsymbol{\theta}\|) \tag{3.4}$$

et que la matrice $\mathbf{S}(\boldsymbol{\theta})$ (de dimension $q \times m$) de cette différentielle (appelé souvent *Jacobien* de cette application) est de rang plein colonne pour tout $\boldsymbol{\theta} \in \Theta$.

$$[\mathbf{S}(\boldsymbol{\theta})]_{i,j} = \frac{\partial s_i(\boldsymbol{\theta})}{\partial \theta_j} \quad \text{avec} \quad \mathbf{s}(\boldsymbol{\theta}) = [s_1(\boldsymbol{\theta}), ..., s_q(\boldsymbol{\theta})]^T \quad \text{et} \quad \boldsymbol{\theta} = (\theta_1, ..., \theta_m)^T.$$

Appelons $\mathcal{S}$ l'ensemble des valeurs prises par la suite de statistiques $\mathbf{s}_n(\mathbf{x}_n)$ et par $\mathbf{s}(\boldsymbol{\theta})$, soit

$$\mathcal{S} = \cup_{n=1}^{\infty} \{\mathbf{s}_n(\mathbf{x}_n); \mathbf{x}_n \in \mathcal{X}\} \cup \mathbf{s}(\Theta).$$

Dans ce cadre, un estimateur par méthode de substitution de $\boldsymbol{\theta}$ sera défini par toute fonction g définie sur $\mathcal{S}$ qui satisfait la condition

$$g[\mathbf{s}(\boldsymbol{\theta})] = \boldsymbol{\theta}, \quad \forall\, \boldsymbol{\theta} \in \Theta \tag{3.5}$$

et qui est continue en tout point de $\mathbf{s}(\boldsymbol{\theta}) \in \mathbf{s}(\Theta)$. Autrement dit, un estimateur $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ par méthode de substitution de $\boldsymbol{\theta}$ sera défini par toute extension g à l'ensemble $\mathcal{S}$ de l'application définie sur $\mathbf{s}(\Theta)$ qui associe $\boldsymbol{\theta}$ à $\mathbf{s}(\boldsymbol{\theta})$.

$$\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) = g[\mathbf{s}_n(\mathbf{x}_n)]. \tag{3.6}$$

Cette extension étant bien entendue non unique, à chacune d'elles, nous obtiendrons un estimateur par méthode de substitution différent. Notons que cette méthode ne nécessite pas que cette fonction g soit définie de façon explicite, comme nous le verrons dans le paragraphe 3.5. Dans ces situations on utilisera plutôt le terme d'algorithme dans la littérature traitement statistique de l'information.

Remarque 3.2 : Il existe des situations où l'on ne s'interesse par à $\boldsymbol{\theta} \in \mathbb{R}^m$ mais plutôt à $\phi(\boldsymbol{\theta}) \in \mathbb{R}^p$ avec $p \leq m$ (en particulier en présence de paramètres de nuisance). On définira aussi des estimateurs par méthode de substitution dans ce cadre par toute fonction g définie sur $\mathcal{S}$ et continue en tout point de $\mathbf{s}(\boldsymbol{\theta}) \in \mathbf{s}(\Theta)$ telle que :

$$g[\mathbf{s}(\boldsymbol{\theta})] = \phi(\boldsymbol{\theta}), \quad \forall\, \boldsymbol{\theta} \in \Theta.$$

Un exemple est donné au paragraphe 3.2.3 où le paramètre général est (θ, σ^2) et où l'on ne s'intéresserait qu'au paramètre utile θ.

Dans les situations où il n'existe pas de statistique exhaustive, il se pose deux questions :

- choix des moments $\mathbf{s}(\boldsymbol{\theta})$ associés aux choix des statistiques $\mathbf{s}_n(\mathbf{x}_n)$ consistantes ;
- choix de l'application ou algorithme g.

La réponse à ces deux questions est donnée par la résolution du compromis complexité et mise en oeuvre du calcul de g versus performance de l'estimateur obtenu. Ce deuxième point est l'objet du paragraphe suivant.

3.4 Performances asymptotiques

3.4.1 Consistance de l'estimateur

Par la propriété de continuité des fonctions g, on démontre que la consistance faible ou forte de la suite des statistiques $\mathbf{s}_n(\mathbf{x}_n)$ se transfère aux suites $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ d'estimateurs construite par la méthode de

substitution. Par contre il n'y a pas de résultats généraux pour la consistance en moyenne quadratique.

Dans le cadre d'observations i.i.d. $(x_1, ..., x_n)$, les suites de statistiques $\mathbf{s}_n(\mathbf{x}_n)$ construites à partir de moments ou de moments généralisés sont d'après les lois faibles et fortes des grands nombres [12] respectivement faiblement et fortement consistantes. Par suite, les suites $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ sont aussi fortement et faiblement consistantes.

Dans le cadre d'observations $(x_1, ..., x_n)$ non i.i.d., il n'y a pas de résultats généraux de consistance pour les suites de moments empiriques $\mathbf{s}_n(\mathbf{x}_n)$. Cependant dans le cas de processus linéaires

$$x_k = m + \sum_{\ell=-\infty}^{+\infty} \alpha_\ell u_{k-\ell},$$

où $(u_k)_{k=...,-1,0,+1,...}$ est une suite i.i.d. centrée et où la suite α_k satisfait certaines conditions techniques, la consistance faible est assurée pour les statistiques construites à partir du moment empirique du premier ordre

$$\frac{1}{n}\sum_{k=1}^{n} x_k$$

sous certaines techniques du moment du second ordre de la variable aléatoire u_k, et construites à partir d'un vecteur de moment empirique du second ordre

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2, ..., \frac{1}{n}\sum_{k=1}^{n-q+1} x_k x_{k-1+q}$$

sous certaines conditions techniques sur moment du troisième ordre de la variable aléatoire u_k. Et ainsi dans ce dernier cas, la suite $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ serait aussi faiblement consistante.

3.4.2 Normalité asymptotique

Dans le cadre d'observations i.i.d. $(x_1, ..., x_n)$, les suites de statistiques $\mathbf{s}_n(\mathbf{x}_n)$ construites à partir du vecteur de moments $\mathrm{E}(x_1^\ell)$, $\ell = 1, ..., q$ ou de vecteur de moments généralisés $\mathrm{E}(h_\ell(x_1))$, $\ell = 1, ..., q$ sous l'hypothèse que les moments d'ordre deux de x_1^ℓ ou de $h_\ell(x_1)$, $\ell = 1, ..., q$ existent, permettent d'appliquer le théorème central classique. Par suite cette suite $\mathbf{s}_n(\mathbf{x}_n)$ est asymptotiquement gaussiennes au sens suivant :

$$\sqrt{n}\left(\mathbf{s}_n(\mathbf{x}_n) - \mathbf{s}(\boldsymbol{\theta})\right) \quad \text{converge en loi vers la loi gaussienne} \quad \mathcal{N}\left(\mathbf{0}; \mathbf{C}_s(\boldsymbol{\theta})\right)$$

où $\mathbf{C}_s(\boldsymbol{\theta})$ désigne la matrice de covariance du vecteur aléatoire $(x_1^1, ..., x_1^q)^T$ où $(h_1(x_1), ..., h_q(x_1))^T$.

Si de plus l'application g définie en (3.5) est non seulement continue en tout point de $\mathbf{s}(\boldsymbol{\theta}) \in \mathbf{s}(\Theta)$, mais de plus différentiable en tout point de $\mathbf{s}(\boldsymbol{\theta}) \in \mathbf{s}(\Theta)$, de matrice $\mathbf{D}_g(\boldsymbol{\theta}) \neq \mathbf{0}$:

$$g[\mathbf{s}(\boldsymbol{\theta}) + \delta\mathbf{s}] = g[\mathbf{s}(\boldsymbol{\theta})] + \mathbf{D}_g(\boldsymbol{\theta})\delta\mathbf{s} + o(\|\delta\mathbf{s}\|), \tag{3.7}$$

l'application du résultat de probabilité 2.1 (2.14) nous permet de déduire que les suites d'estimateurs par la méthode de substitution sont asymptotiquement gaussiennes au sens suivant :

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) - \boldsymbol{\theta}\right) \quad \text{converge en loi vers la loi gaussienne} \quad \mathcal{N}\left(\mathbf{0}; \mathbf{D}_g(\boldsymbol{\theta})\mathbf{C}_s(\boldsymbol{\theta})\mathbf{D}_g^T(\boldsymbol{\theta})\right). \tag{3.8}$$

La matrice de covariance $\mathbf{D}_g(\boldsymbol{\theta})\mathbf{C}_s(\boldsymbol{\theta})\mathbf{D}_g^T(\boldsymbol{\theta})$ de la loi asymptotique est appelée *matrice de covariance asymptotique* de la suite d'estimateurs $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$.

Sous des hypothèses suffisantes supplémentaires (malheureusement qu'il est difficile de prouver en

12. Sous certaines conditions d'existence de ces moments.

pratique) nous avons de plus :

$$\mathrm{cov}[\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)] = \frac{1}{n}\mathbf{D}_g(\boldsymbol{\theta})\mathbf{C}_s(\boldsymbol{\theta})\mathbf{D}_g^T(\boldsymbol{\theta}) + o\left(\frac{1}{n}\right). \tag{3.9}$$

De plus si la fonction g est deux fois différentiable en tout point de $\mathbf{s}(\boldsymbol{\theta}) \in \mathbf{s}(\Theta)$, i.e.,

$$g_i[\mathbf{s}(\boldsymbol{\theta}) + \delta\mathbf{s}] = g_i[\mathbf{s}(\boldsymbol{\theta})] + [\mathbf{D}_g(\boldsymbol{\theta})]_{i,.}\delta\mathbf{s} + \frac{1}{2}(\delta\mathbf{s})^T\mathbf{H}_{g_i}(\boldsymbol{\theta})\delta\mathbf{s} + o(\|\delta\mathbf{s}\|^2), \quad i = 1, ...m,$$

nous avons accès sous des conditions techniques supplémentaires (difficiles aussi à prouver) à une expression suivante du biais asymptotique :

$$\mathrm{E}[\widehat{\boldsymbol{\theta}}_{n,i}(\mathbf{x}_n)] = \boldsymbol{\theta}_i + \frac{1}{2n}\mathrm{tr}[\mathbf{H}_{g_i}(\boldsymbol{\theta})\mathbf{C}_s(\boldsymbol{\theta})] + o\left(\frac{1}{n}\right), \quad i = 1, ...m. \tag{3.10}$$

Exemple 3.1 : Reprenons l'exemple développé au paragraphe 3.2.1 où $g(s) = \frac{1}{s}$ et par suite $\frac{dg}{ds} = -\frac{1}{s^2}$ et $\frac{d^2g}{ds^2} = \frac{2}{s^3}$. Par suite par simple application de (3.8), nous avons puisque $\mathrm{var}(x_1) = \frac{1}{\theta^2}$ et donc $\left(\frac{dg}{ds}\right)\mathrm{var}(x_1)\left(\frac{dg}{ds}\right) = (-\theta)^2\frac{1}{\theta^2}(-\theta)^2 = \theta^2$

$$\sqrt{n}(\widehat{\theta}_n(\mathbf{x}_n) - \theta) \text{ converge en loi vers la loi gaussienne } \mathcal{N}\left(0; \theta^2\right).$$

Puis par application de (3.9) et (3.10) avec $\frac{d^2g}{ds^2}\mathrm{var}(x_1) = 2\theta^3\frac{1}{\theta^2}$ nous obtenons :

$$\mathrm{cov}[\widehat{\theta}_n(\mathbf{x}_n)] = \frac{\theta^2}{n} + o\left(\frac{1}{n}\right) \text{ et } \mathrm{E}[\widehat{\theta}_n(\mathbf{x}_n)] = \theta + \frac{\theta}{n} + o\left(\frac{1}{n}\right).$$

Dans le cadre d'observations $(x_1, ..., x_n)$ non i.i.d., il n'y a pas de résultats généraux de normalité asymptotique. Cependant dans le cadre où ces observations sont extraites de processus linéaires

$$x_k = m + \sum_{\ell=-\infty}^{+\infty} \alpha_\ell u_{k-\ell} \tag{3.11}$$

où $(u_k)_{k=...,-1,0,+1,...}$ est une suite i.i.d. centré de variance σ^2, nous avons un grand nombre de résultats pour les suites de statistiques du premier et second ordre, parmi lesquels :

Résultat 3.1 : Sous les hypothèses $\sum_{k=-\infty}^{+\infty}|\alpha_k| < \infty$ et $\sum_{k=-\infty}^{+\infty}\alpha_k \neq 0$, la suite $s_n(\mathbf{x}_n) = \frac{1}{n}\sum_{k=1}^{n} x_k$ est asymptotiquement gaussienne au sens suivant :

$$\sqrt{n}\,(s_n(\mathbf{x}_n) - m) \text{ converge en loi vers la loi gaussienne } \mathcal{N}\left(0; S_x(0)\right), \tag{3.12}$$

où $S_x(f)$ appelé *spectre du processus stationnaire* x_k (3.11) est défini par

$$S_x(f) = \sum_{k=-\infty}^{\infty} r_x(k)e^{-i2\pi kf} \tag{3.13}$$

avec $r_x(k) \stackrel{\text{def}}{=} \mathrm{E}[(x_n - m)(x_{n+k} - m)]$.

Résultat 3.2 : Pour $m = 0$ et pour $(u_n)_{n=...,-1,0,+1,...}$ suite i.i.d. de loi de probabilité gaussienne $\mathcal{N}(0, \sigma^2)$, sous les hypothèses $\sum_{k=-\infty}^{+\infty} |\alpha_k| < \infty$ et $\sum_{k=-\infty}^{+\infty} |k|\alpha_k^2 < \infty$, alors la suite de statistiques

$$\mathbf{s}_n(\mathbf{x}_n) = \left(\frac{1}{n} \sum_{k=1}^{n} x_k^2, ..., \frac{1}{n} \sum_{k=1}^{n-q+1} x_k x_{k-1+q} \right)^T \in \mathbb{R}^q$$

est asymptotiquement gaussienne au sens suivant :

$$\sqrt{n}\,(\mathbf{s}_n(\mathbf{x}_n) - \mathbf{s}) \quad \text{converge en loi vers la loi gaussienne} \quad \mathcal{N}\,(\mathbf{0}; \mathbf{C}_s)\,, \tag{3.14}$$

où $\mathbf{s} = (\mathrm{E}(x_1^2), \mathrm{E}(x_1 x_2), ..., \mathrm{E}(x_1 x_q))^T$ et

$$[\mathbf{C}_s]_{k,l} = \int_0^1 S_x^2(f)[e^{i2\pi(k-l)f} + e^{i2\pi(k+l)f}]df \quad \text{dite } \textit{formule de Bartlett.} \tag{3.15}$$

Ces deux résultats permettent d'en déduire la normalité asymptotique de tout estimateur par la méthode de substitution construit à partir des moments du premier et second ordre de ces processus, en utilisant le résultat de probabilité 2.1.

Remarque 3.3 : Dans le cas plus général où $(u_k)_{k=...,-1,0,+1,...}$ est une suite i.i.d. centré de variance σ^2 de loi de probabilité non nécessairement gaussienne, le résultat 3.2 précédent s'étend si le moment d'ordre quatre de u_n existe. Dans ce cas s'ajoute au terme de (3.15), un deuxième terme qui est une somme de cumulants d'ordre quatre de u_k (qui est nul dans le cas où u_n est de loi gaussienne).

Exemple 3.2 : Considérons une observation $(x_1, ..., x_n)$ extraite d'un processus MA (Moving Average) du premier ordre qui est un processus stationnaire

$$x_k = \theta + u_k + a u_{k-1},$$

où $(u_k)_{k=...,-1,0,+1,...}$ est une suite i.i.d. de variance σ_u^2. Puisque $r_x(0) = (1 + a^2)\sigma_u^2$, $r_x(\pm 1) = a\sigma_u^2$ et $r_x(k) = 0$ pour $k \neq \{-1, 0, +1\}$, le spectre défini en (3.13) est alors

$$S_x(f) = (1 + a^2 + 2a \cos 2\pi f)\sigma_u^2$$

et donc $S_x(0) = (1 + a)^2\sigma_u^2$, l'application de (3.12) donne pour $\sum_{k=-\infty}^{+\infty} \alpha_k = 1 + a \neq 0$, soit $a \neq -1$

$$\sqrt{n}(\widehat{\theta}_n(\mathbf{x}_n) - \theta) \quad \text{converge en loi vers la loi gaussienne} \quad \mathcal{N}\left(0; (1 + a)^2\sigma_u^2\right). \tag{3.16}$$

Pour comparer ce résultat avec celui de la suite i.i.d. $(x_1, ..., x_n)$ de moyenne θ et de variance σ^2, nous devons donner à σ_u^2 la valeur $\sigma_u^2 = \frac{\sigma^2}{1+a^2}$. Par suite (3.16) devient :

$$\sqrt{n}(\widehat{\theta}_n(\mathbf{x}_n) - \theta) \quad \text{converge en loi vers la loi gaussienne} \quad \mathcal{N}\left(0; \left(1 + \frac{2a}{1 + a^2}\right)\sigma^2\right)$$

avec $(1 + \frac{2a}{1+a^2}) \in]0, 2]$ pour $a \in]-\infty, +\infty[$. La corrélation entre deux termes successifs de x_k introduit par $a \neq 0$ peut augmenter ou diminuer la variance asymptotique de l'estimateur *moyenne empirique* selon le signe de a.

Remarquons que dans le cas particulier où u_k est de loi gaussienne $\mathcal{N}(0, \sigma_u^2)$, la loi de probabilité de l'estimateur $\widehat{\theta}_n(\mathbf{x}_n)$ est directement accessible $\forall a$. En effet dans ce cas, le vecteur $\mathbf{x}_n$ est de loi gaussienne

et par suite $\widehat{\theta}_n(\mathbf{x}_n)$ l'est aussi avec la variance :

$$\begin{aligned}\mathrm{var}(\widehat{\theta}_n(\mathbf{x}_n)) = \frac{1}{n^2}\left(\sum_{k=1}^{n}\mathrm{var}(x_k) + \sum_{1\leq k\neq l\leq n}\mathrm{cov}(x_k,x_l)\right) &= \frac{1}{n^2}\left(n(1+a^2)\sigma_u^2 + 2(n-1)a\sigma_u^2\right)\\ &= \frac{1}{n}\left((1+a)^2\sigma_u^2 - \frac{2a}{n}\sigma_u^2\right),\end{aligned}$$

car $\mathrm{cov}(x_k,x_l) = a\sigma_u^2$ pour $|k-l|=1$ et 0 ailleurs. Par suite l'estimateur

$$\widehat{\theta}_n(\mathbf{x}_n) \text{ est de loi gaussienne } \mathcal{N}\left(\theta; \frac{(1+a)^2\sigma_u^2 - \frac{2a}{n}\sigma_u^2}{n}\right) \quad \forall a. \tag{3.17}$$

Soit

$$\sqrt{n}(\widehat{\theta}_n(\mathbf{x}_n) - \theta) \text{ est de loi gaussienne } \mathcal{N}\left(0; (1+a)^2\sigma_u^2 - \frac{2a}{n}\sigma_u^2\right),$$

qui converge bien en loi vers la loi gaussienne $\mathcal{N}\left(0;(1+a)^2\sigma_u^2\right)$ pour $a\neq -1$ en cohérence avec (3.16). En outre pour $a=-1$, (3.17) implique que $n(\widehat{\theta}_n(\mathbf{x}_n)-\theta)$ est de loi gaussienne $\mathcal{N}(0;2\sigma_u^2)$.

Exemple 3.3 : En reprenant l'exemple illustratif 3.2.3 où ici le processus u_n est de loi de probabilité gaussienne centrée de variance unité, l'expression $S_x(f) = \frac{1}{|1-\theta e^{-i2\pi f}|^2}$ issu de (3.13) donne par la formule de Bartlett (3.15)

$$\mathbf{C}_s(\boldsymbol{\theta}) = \frac{1}{(1-\theta^2)^3}\begin{bmatrix} 2(1+\theta^2) & 4\theta \\ 4\theta & 1+4\theta^2-\theta^4 \end{bmatrix},$$

qui permet d'en déduire par la relation (3.8) et par le calcul des différentielles des deux applications g associées à $s_0(\theta)$ et $\begin{bmatrix} s_0(\theta) \\ s_1(\theta)\end{bmatrix}$:

$$\mathbf{D}_g(\theta) = -\frac{(1-\theta^2)^2}{1+\theta^2} \text{ et } \mathbf{D}_g(\theta) = -(1-\theta^2)[-\theta,1],$$

les variances asymptotiques des deux estimateurs $\widehat{\theta}_n^{(1)}(\mathbf{x}_n)$ et $\widehat{\theta}_n^{(2)}(\mathbf{x}_n)$ donnés respectivement par :

$$v_1 = \frac{(1-\theta^2)(1+4\theta^2-\theta^4)}{(1+\theta^2)^2} \text{ et } v_2 = \frac{1}{(1-\theta^2)}[-\theta,1]\begin{bmatrix} 2(1+\theta^2) & 4\theta \\ 4\theta & 1+4\theta^2-\theta^4 \end{bmatrix}\begin{bmatrix} -\theta \\ 1\end{bmatrix} = 1-\theta^2.$$

Il est aisé de voir que $v_2 < v_1$ avec égalité si et seulement $\theta = 0$. Ce qui est assez intuitif car l'estimateur $\widehat{\theta}_n^{(2)}(\mathbf{x}_n)$ est construit à partir des deux moments $(\mathrm{E}(x_k^2), \mathrm{E}(x_k x_{k+1}))$ contrairement à $\widehat{\theta}_n^{(1)}(\mathbf{x}_n)$ qui n'est construit que sur le seul moment $\mathrm{E}(x_k^2)$. C'est un résultat général qui fera l'objet de la remarque 3.6.

3.5 Estimateur à variance asymptotique minimale

Etant donné la non unicité de l'application g qui associe à la suite de statistiques $\mathbf{s}_n(\mathbf{x}_n)$, l'estimateur $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ (3.6) une question s'impose. Existe t'il de façon similaire à la borne de Cramer Rao et à l'existence d'estimateur asymptotiquement efficace, une borne inférieure de la matrice de covariance asymptotique associée à une statistique $\mathbf{s}_n(\mathbf{x}_n)$? Et si oui, existe t'il une fonction g telle que la matrice de covariance asymptotique associée atteint cette borne inférieure ?

3.5.1 Borne inférieure de la covariance asymptotique

Nous allons démontrer par le résultat suivant qu'une telle borne inférieure existe.

Résultat 3.3 : Tous les estimateurs issus de la méthode de substitution construit à partir de la suite de statistiques $\mathbf{s}_n(\mathbf{x}_n)$ ont une covariance asymptotique bornée inférieurement [13] par la matrice $[\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1}$ qui ne dépend que de la statistique $\mathbf{s}_n(\mathbf{x}_n)$ à travers sa matrice de covariance asymptotique $\mathbf{C}_s(\boldsymbol{\theta})$ supposée inversible [14] et le Jacobien $\mathbf{S}(\boldsymbol{\theta})$.

$$\mathbf{D}_g(\boldsymbol{\theta})\mathbf{C}_s(\boldsymbol{\theta})\mathbf{D}_g^T(\boldsymbol{\theta}) \geq [\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1}. \tag{3.18}$$

Preuve :
A partir de (3.5) et des applications différentiables g (3.7) et $\mathbf{s}$ (3.4), nous en déduisons

$$\begin{aligned} g[\mathbf{s}(\boldsymbol{\theta}+\delta\boldsymbol{\theta})] &= \boldsymbol{\theta}+\delta\boldsymbol{\theta}, \quad \forall\ \boldsymbol{\theta}+\delta\boldsymbol{\theta} \in \Theta \\ &= g[\mathbf{s}(\boldsymbol{\theta}) + \mathbf{S}(\boldsymbol{\theta})\delta\boldsymbol{\theta} + o(\|\delta\boldsymbol{\theta}\|] \\ &= g[\mathbf{s}(\boldsymbol{\theta})] + \mathbf{D}_g(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})\delta\boldsymbol{\theta} + o(\|\delta\boldsymbol{\theta}\|) = \boldsymbol{\theta} + \mathbf{D}_g(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})\delta\boldsymbol{\theta} + o(\|\delta\boldsymbol{\theta}\|]. \end{aligned}$$

Par suite les fonctions g sont contraintes à satisfaire la relation

$$\mathbf{D}_g(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta}) = \mathbf{I}_m \ \text{ où } \mathbf{D}_g(\boldsymbol{\theta}) \text{ et } \mathbf{S}(\boldsymbol{\theta}) \text{ sont des matrices } m\times q \text{ et } q\times m \text{ avec } q \geq m. \tag{3.19}$$

Puis à partir de cette contrainte, nous avons :

$$\begin{aligned} &\mathbf{D}_g(\boldsymbol{\theta})\mathbf{C}_s(\boldsymbol{\theta})\mathbf{D}_g^T(\boldsymbol{\theta}) - [\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1} \\ &= \left[\mathbf{D}_g(\boldsymbol{\theta}) - [\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1}\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\right]\mathbf{C}_s(\boldsymbol{\theta})\left[\mathbf{D}_g(\boldsymbol{\theta}) - [\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1}\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\right]^T \geq 0. \end{aligned}$$

■

Remarque 3.4 : Le cas où l'"inversion" de l'application $\boldsymbol{\theta} \mapsto \mathbf{s}(\boldsymbol{\theta})$ est unique (par exemple dans l'exemple illustratif 3.2.1 correspond au cas $q = m$, $\mathbf{S}(\boldsymbol{\theta})$ et $\mathbf{D}_g(\boldsymbol{\theta})$ inversible avec $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{D}_g^{-1}(\boldsymbol{\theta})$ et l'estimateur par substitution g atteint naturellement la borne (3.18).

Remarque 3.5 : Cette borne inférieure $[\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1}$ appelée borne AMV (*asymptotically minimum variance*) joue un role similaire à la borne de Cramer Rao où tout estimateur sans biais construit à partir de l'observation $\mathbf{x}_n = (x_1, ..., x_n)$ est remplacé par toute suite d'estimateurs consistante construit à partir d'une statistique $\mathbf{s}_n(\mathbf{x}_n)$ issue de $\mathbf{x}_n$. En outre, elle présente l'avantage de toujours pouvoir en obtenir une formule analytique (malheureusement rarement interprétable). Au contraire, l'expression analytique

$$\mathrm{cov}[\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)] \geq \mathbf{I}_{\mathbf{x}_n}^{-1}(\boldsymbol{\theta}),$$

où

$$\mathbf{I}_{\mathbf{x}_n}(\boldsymbol{\theta}) = -\mathrm{E}\left[\frac{\partial^2 \ln p(\mathbf{x}_n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right] = \mathrm{E}\left[\left(\frac{\partial \ln p(\mathbf{x}_n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln p(\mathbf{x}_n;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^T\right],$$

de la borne de Cramer Rao n'est guère accessible en dehors du cadre de la famille exponentielle (voir paragraphe 2.3.7).

Remarque 3.6 : D'après le principe d'inclusion, si une statistique $\mathbf{s}_n^{(2)}(\mathbf{x}_n)$ provient d'une statistique

13. Au sens de matrices symétriques réelles $m \times m$, i.e., $\mathbf{A} \geq \mathbf{B} \Leftrightarrow \mathbf{u}^T\mathbf{A}\mathbf{u} \geq \mathbf{u}^T\mathbf{B}\mathbf{u}$, $\forall \mathbf{u} \in \mathbb{R}^m$.
14. Cette condition est équivalente à ce que les différentes composantes du vecteur $\mathbf{s}_n(\mathbf{x}_n)$ soient des variables aléatoires linéairement indépendantes.

$\mathbf{s}_n^{(1)}(\mathbf{x}_n)$ par ajout de composantes supplémentaires : $\mathbf{s}_n^{(2)}(\mathbf{x}_n) = \begin{bmatrix} \mathbf{s}_n^{(1)}(\mathbf{x}_n) \\ \times \end{bmatrix}$), la borne AMV associée à l'estimateur par la méthode de substitution sera inférieure, i.e.,

$$[\mathbf{S}_2^T(\boldsymbol{\theta})\mathbf{C}_{s_2}^{-1}(\boldsymbol{\theta})\mathbf{S}_2(\boldsymbol{\theta})]^{-1} \leq [\mathbf{S}_1^T(\boldsymbol{\theta})\mathbf{C}_{s_1}^{-1}(\boldsymbol{\theta})\mathbf{S}_1(\boldsymbol{\theta})]^{-1}.$$

Exemple 3.4 : En reprenant à nouveau l'exemple illustratif 3.2.3, l'estimateur $\widehat{\theta}_n^{(1)}(\mathbf{x}_n)$ fonction de la seule statistique scalaire $\frac{1}{n}\sum_{k=1}^{n-1} x_k x_{k+1}$ a une variance asymptotique qui atteint la borne (3.18) car l'estimateur $\widehat{\theta}_n^{(1)}(\mathbf{x}_n)$ est issu d'un "inversion" unique (voir remarque 3.4).

Quant à l'estimateur $\widehat{\theta}_n^{(2)}(\mathbf{x}_n)$, nous avons :

$$\mathbf{C}_s^{-1}(\boldsymbol{\theta}) = \frac{1}{2}\begin{bmatrix} 1+4\theta^2-\theta^4 & -4\theta \\ -4\theta & 2(1+\theta^2) \end{bmatrix}.$$

Le Jacobien $\mathbf{S}(\boldsymbol{\theta})$ de $\mathbf{s}(\boldsymbol{\theta})$ et donné par :

$$\mathbf{S}(\boldsymbol{\theta}) = \frac{1}{(1-\theta^2)^2}\begin{bmatrix} -2\theta \\ -(1+\theta^2) \end{bmatrix}.$$

Par suite nous obtenons $[\mathbf{S}^T(\boldsymbol{\theta})\mathbf{C}_s^{-1}(\boldsymbol{\theta})\mathbf{S}(\boldsymbol{\theta})]^{-1} = 1-\theta^2$ qui est égal à la variance asymptotique v_2 obtenue dans l'exemple 3.3. Par suite l'estimateur $\widehat{\theta}_n^{(2)}(\mathbf{x}_n)$ a une variance asymptotique minimum parmi tous les estimateurs issus de la méthode de substitution construits à partir du couple de statistiques $\left(\frac{1}{n}\sum_{k=1}^{n} x_k^2, \frac{1}{n}\sum_{k=1}^{n-1} x_k x_{k+1}\right)$.

3.5.2 Estimateur minimum de distance

Nous allons voir maintenant qu'il existe toujours au moins un estimateur issu de la méthode de substitution dont la covariance asymptotique atteint cette borne inférieure (3.18). Par un simple calcul de perturbation permettant de démontrer que la matrice $\mathbf{D}_g(\boldsymbol{\theta})$ (définie en (3.7)) de l'application différentiable qui associe $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ à la statistique $\mathbf{s}_n(\mathbf{x}_n)$ satisfait l'égalité (3.18), il est aisé de démontrer les deux résultats suivants :

Résultat 3.4 : Tout estimateur, argument de la minimisation du critère *moindre carré pondéré* suivant a une covariance asymptotique qui atteint la borne inférieure (3.18)

$$\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) = \arg\min_{\boldsymbol{\alpha}} \left[\mathbf{s}_n(\mathbf{x}_n) - \mathbf{s}(\boldsymbol{\alpha})\right]^T \mathbf{C}_s^{-1}(\boldsymbol{\alpha}) \left[\mathbf{s}_n(\mathbf{x}_n) - \mathbf{s}(\boldsymbol{\alpha})\right]. \tag{3.20}$$

De plus, pour rendre cette minimisation moins complexe, nous disposons de ce deuxième résultat :

Résultat 3.5 : Tout estimateur, argument de la minimisation du critère moindre carré pondéré suivant a une covariance asymptotique qui atteint la borne inférieure (3.18)

$$\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) = \arg\min_{\boldsymbol{\alpha}} \left[\mathbf{s}_n(\mathbf{x}_n) - \mathbf{s}(\boldsymbol{\alpha})\right]^T \mathbf{W}(\mathbf{x}_n) \left[\mathbf{s}_n(\mathbf{x}_n) - \mathbf{s}(\boldsymbol{\alpha})\right]. \tag{3.21}$$

où $\mathbf{W}(\mathbf{x}_n)$ désigne une estimée faiblement consistante de $\mathbf{C}_s^{-1}(\boldsymbol{\theta})$.

3.6 Points essentiels du chapitre estimateur par méthode de substitution (ou de moments)

- Cette méthode s'applique dans le cadre d'une observation $\mathbf{x}_n = (x_1, ..., x_n)$ scalaires ou multidimensionnelles où $(x_k)_{k=1,...,n}$ sont soit indépendantes identiquement distribuées (i.i.d), soit extraite d'une suite stationnaire $(x_k)_{...,1,...,n,...}$. Elle repose sur le choix d'un couple $[\mathbf{s}(\boldsymbol{\theta}), \mathbf{s}_n(\mathbf{x}_n)]$ tel que :
  - $\mathbf{s}(\boldsymbol{\theta})$ est constitué de *moments* (ex : $\mathrm{E}(x_k^\ell)$), ou *moments genéralisés* (ex : $\mathrm{E}(h_\ell(x_k))$)
  - $\mathbf{s}_n(\mathbf{x}_n)$ est constitué de moments empiriques (ex : $\frac{1}{n}\sum_{k=1}^n x_k^\ell$) ou moments genéralisés empiriques (ex : $\frac{1}{n}\sum_{k=1}^n h_\ell(x_k)$) associés.
  - $\mathbf{s}(\boldsymbol{\theta})$ caractérise $\boldsymbol{\theta}$, i.e., l'application $\boldsymbol{\theta} \in \Theta \mapsto \mathbf{s}(\boldsymbol{\theta})$ est injective.
  - La suite $\mathbf{s}_n(\mathbf{x}_n)$ de statistiques converge vers $\mathbf{s}(\boldsymbol{\theta})$ dans au moins un des sens *presque sûre (consistance forte), en probabilité (consistance faible) ou en moyenne quadratique.*
- A partir de ce choix d'un couple $[\mathbf{s}(\boldsymbol{\theta}), \mathbf{s}_n(\mathbf{x}_n)]$, un estimateur par méthode de substitution (ou de moments) peut être défini comme une extension g de l'application $\mathbf{s}(\boldsymbol{\theta}) \mapsto \boldsymbol{\theta}$ pour des valeurs de $\mathbf{s}_n(\mathbf{x}_n)$ qui n'appartiennent pas nécessairement à l'ensemble $\{\mathbf{s}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$: $\mathbf{s}_n(\mathbf{x}_n) \mapsto \widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n) = g(\mathbf{s}_n(\mathbf{x}_n))$. Il est donc en général non unique.
- La méthode de substitution fournit en général des estimateurs qui ne sont ni efficaces, ni asymptotiquement efficaces. Mais elle produit des suites d'estimateurs $\widehat{\boldsymbol{\theta}}_n(\mathbf{x}_n)$ consistantes et asymptotiquement gaussiens dont on sait en général calculer la matrice de covariance asymptotique $\mathbf{C}(\boldsymbol{\theta})$.
- A partir du choix d'un couple $[\mathbf{s}(\boldsymbol{\theta}), \mathbf{s}_n(\mathbf{x}_n)]$, on peut exhiber parmi les différentes méthodes des moments possibles, l'estimateur qui minimise (au sens des matrices définies positives) la matrice de covariance asymptotique $\mathbf{C}(\boldsymbol{\theta})$. C'est un estimateur qui est solution de la minimisation par rapport à $\boldsymbol{\theta}$ d'une distance euclidienne entre $\mathbf{s}(\boldsymbol{\theta})$ et $\mathbf{s}_n(\mathbf{x}_n)$.