

Estimation - Détection, Notes de cours (provisoires)

L.Deneire, emprunts à Eric Thierry

Année 2007-2008

Chapitre 1

Introduction

Ce cours d'Estimation-Détection est destiné aux étudiants du Master STIC/Signaux et Communications ainsi qu'aux étudiants ingénieur en Mathématiques Appliquées et Modélisation de l'EPU / UNSA. Le cours est très largement inspiré du cours d'Eric Thierry, donné les années précédentes, et des livres de Kay et Scharf [Kay93b, Kay93a, Sch91]. La structure du cours suivra le livre de Kay.

Chapitre 2

Estimation déterministe : Introduction

2.1 Le problème d'estimation

On observe N valeurs $\mathbf{x} = (x_1, \dots, x_N)$ provenant de tirage indépendants d'une variable aléatoire X . Cette v.a. a pour densité de probabilité $f(x; \theta)$ où θ représente un ensemble de paramètres inconnus considérés comme étant déterministes (valeur fixée). On notera $\mathbf{X} = (X_1, \dots, X_N)$ le vecteur aléatoire dont une réalisation est \mathbf{x} .

Une estimation de θ est une fonction mesurable $T(\mathbf{x})$ des observations. La valeur de l'estimée dépend de la réalisation \mathbf{x} . On appellera donc un estimateur, la v.a. $T(\mathbf{X})$. Des exemples d'estimateur sont :

- $T(\mathbf{x}) = \frac{x_1 + \dots + x_N}{N}$,
- $T(\mathbf{x}) = x_1$,
- $T(\mathbf{x}) = \max(x_1, \dots, x_N)$.

2.2 Exemple simple

Nous allons utiliser un exemple extrêmement simple : l'estimation d'une constante dans du bruit (par exemple la détermination d'une tension continue (DC : Direct Current) dans du bruit). Une étape cruciale dans l'estimation est la modélisation des données. Par exemple, pour $N = 1$, on peut modéliser les données à l'aide de la densité de probabilité (PDF : Probability Density Function) par :

$$p(x_1; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (x_1 - \theta)^2 \right] \quad (2.1)$$

où on a considéré simplement le modèle suivant :

$$x_n = A + w_n, n = 1 \dots N \quad (2.2)$$

avec A la constante à estimer (et donc le paramètre θ "vrai" vaut A) et $w_n; n = 1 \dots N$ sont des variables aléatoires gaussiennes de moyenne nulle et de variance σ^2 , et indépendantes entre elles (on dit que w_i sont des variables i.i.d. : indépendantes et identiquement distribuées).

Un autre exemple simple est le cas d'une rampe : $x_n = A + B.n + w_n; n = 1 \dots N$ la PDF s'écrit alors, en considérant $\theta = [AB]^T$

$$p(\mathbf{x}; \theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp \left[-\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - A - Bn)^2 \right] \quad (2.3)$$

2.3 Qualité d'un estimateur

2.3.1 Le biais

On définit le biais de la façon suivante :

$$B(T) = E[T(\mathbf{X})] - \theta$$

Un estimateur est sans biais si $E[T(\mathbf{X})] = \theta$. Un estimateur sera *asymptotiquement non biaisé* si $\lim_{n \rightarrow \infty} B(T) = 0$.

2.3.2 L'erreur quadratique moyenne - MSE : Mean Squared Error

Elle permet de définir la qualité de l'estimateur au second degré.

$$MSE(T) = E[(T - \theta)^2] = \text{Var}[T] - B(T)^2$$

Si $\lim_{n \rightarrow \infty} MSE(T) = 0$ alors l'estimateur est *asymptotiquement consistant*.

2.3.3 Efficacité

Supposons que les deux estimateurs T_1 et T_2 soient non biaisés alors on dira que T_1 est plus efficace que T_2 si

$$\text{Var}[T_1] < \text{Var}[T_2]$$

.

2.3.4 Exemples

On se base toujours sur l'estimation d'une moyenne dans du bruit. L'estimateur naturel est $\hat{\theta} = T(\mathbf{x}) = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_n$. Cette estimateur est non biaisé et asymptotiquement consistant. Supposons maintenant que le paramètre inconnu soit la variance $\theta = \sigma^2$. L'estimateur naturel est $\hat{\theta} = T(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (x_n - \bar{x})^2$. Cette estimateur est biaisé et asymptotiquement non biaisé. En effet on a :

$$\begin{aligned} \sum_{n=1}^N (x_n - A)^2 &= \sum_{n=1}^N (x_n - \bar{x} + \bar{x} - A)^2 = \sum_{n=1}^N (x_n - \bar{x})^2 + \sum_{n=1}^N (\bar{x} - A)^2 \\ E[S^2] &= \sum_{n=1}^N (\text{Var}[x_i] - \text{Var}[\bar{x}]) = \sigma^2(1 - 1/N) \end{aligned}$$

Un simple changement d'échelle permet de définir un estimateur non biaisé

$$\tilde{S}^2 = \frac{N}{N-1} S^2$$

2.3.5 Loi de Poisson

Dans l'exemple précédent le modèle probabiliste n'était défini que par les deux premiers moments des observations. On donne maintenant la loi des observations ¹ :

$$p(x; \theta) = \frac{\exp(-\theta)\theta^x}{x!} \text{ avec } x = 0, 1, 2, \dots$$

¹ $E(X) = \theta \quad \text{Var}(X) = \theta$

Comparons les estimateurs $T_1(\mathbf{x}) = \bar{x}$ et $T_2(\mathbf{x}) = x_2$. On a :

$$\begin{aligned} E[T_1] &= E[T_2] = \theta \\ \text{Var}[T_1] &= \theta/n \text{ et } \text{Var}[T_2] = \theta \end{aligned}$$

Les deux estimateurs sont sans biais mais l'estimateur T_1 est plus efficace que l'estimateur T_2 .

2.4 Etude d'un estimateur

2.4.1 Le modèle probabiliste et l'estimateur

On fait N observations d'une variable aléatoire discrète suivant une loi de Bernoulli $Pr[X = 1] = \theta$ et $Pr[X = 0] = 1 - \theta$. Le paramètre que l'on veut estimer est θ . On notera la vraie valeur du paramètre : p .

On utilise l'estimateur $\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n$.

2.4.2 Motivation

L'exemple choisi va nous permettre de conduire les calculs sans trop de difficultés. Le problème qui nous intéresse est de déterminer la qualité d'un estimateur. Pour cela plusieurs approches sont possibles par ordre de difficultés décroissantes :

- Déterminer la densité de probabilité de l'estimateur pour une taille n donnée. Les calculs sont en général très durs à effectuer.
- Déterminer la moyenne et la variance de l'estimateur pour une taille n donnée.
- Déterminer la loi asymptotique de l'estimateur. En général c'est faisable lorsque l'on peut utiliser le théorème central limite qui dit que la densité asymptotique est gaussienne. Il reste ensuite à calculer la moyenne et la variance asymptotique de l'estimateur. Un plus est de pouvoir calculer la vitesse de convergence de la densité de l'estimateur vers la loi normale.
- Déterminer la moyenne et la variance asymptotique de l'estimateur sans pouvoir montrer que la loi asymptotique est normale. Cela va permettre de déterminer des intervalles de confiance.
- Déterminer empiriquement la loi et les moments de l'estimateur à partir de tirage de Monte-Carlo. Même si on a pu conduire les calculs analytiques précédents cette étape est nécessaire pour les valider.

2.4.3 Loi de l'estimateur

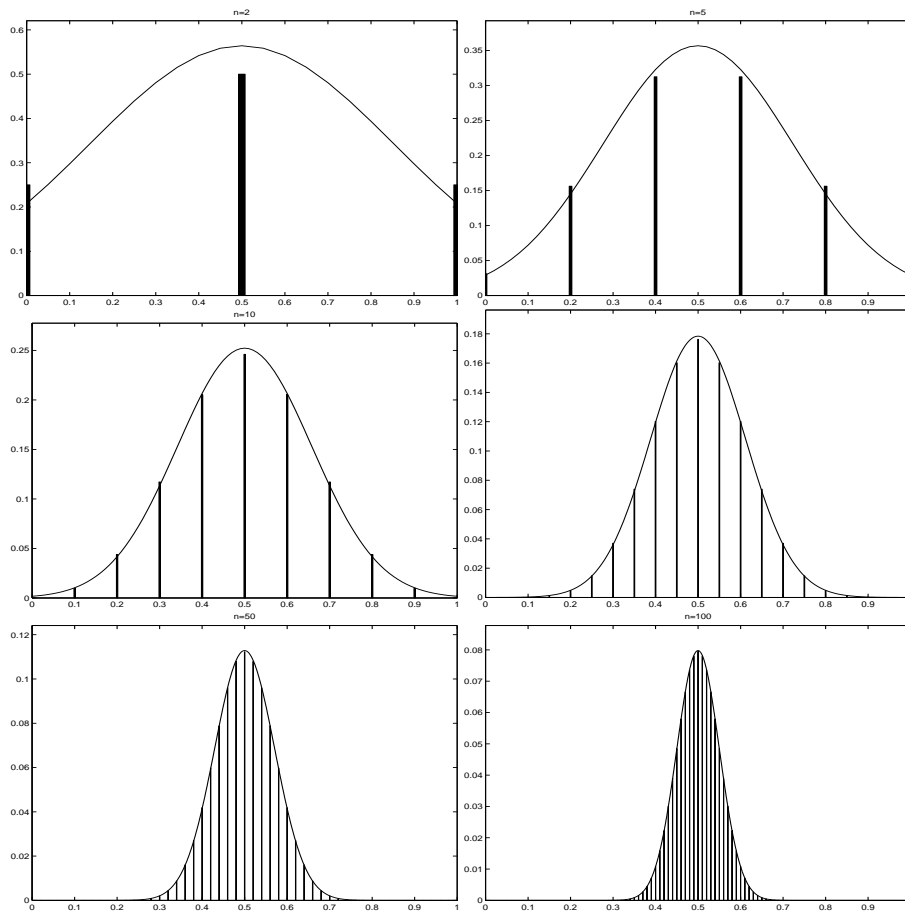
Il est facile de voir que $N\bar{x}$ suit une **loi Binomiale** $B(N,k)$ k étant le nombre de 1 dans la séquence observée

$$Pr[N\bar{x} = k] = C_k^N p^k (1-p)^{N-k}$$

2.4.4 Moyenne et variance de l'estimateur

On peut montrer (approximation de Moivre-Laplace) que si $Np(1-p) \gg 1$ [cf Papoulis p 49] alors

$$C_k^N p^k (1-p)^{N-k} \simeq \frac{1}{\sqrt{2\pi Np(1-p)}} \exp\left(-\frac{(k - Np)^2}{2Np(1-p)}\right)$$



Cela revient à dire qu'asymptotiquement (c'est-à-dire lorsque $N \rightarrow \infty$) on peut approximer la densité de l'estimateur par une densité gaussienne de moyenne p et de variance $p(1-p)/N$.

L'estimateur utilisé précédemment n'est pas le seule estimateur possible, on pourrait en effet utiliser un estimateur de la forme

$$T_2(\mathbf{X}) = a_1X_1 + a_2X_2 + \dots + a_NX_N$$

avec $\sum_{n=1}^N a_n = 1$. Cette contrainte assure un biais nul. Maintenant pour comparer les deux estimateurs nous allons comparer leur variance, on va donc faire une étude au second ordre. En fait on va montrer qu'aucun estimateur de biais nul ne peut avoir une variance inférieure à la variance de T_1 . Cette borne inférieure porte le nom de **borne de Cramer-Rao**.

Chapitre 3

Estimateur non-biaisé à variance minimale (MVUE : Minimum Variance Unbiased Estimator)

3.1 Introduction

Un estimateur MVU est un estimateur dont la moyenne est égale à la vraie valeur du paramètre sur tout l'intervalle de définition de celui-ci (par exemple $a < \theta < b$) :

$$E\{\hat{\theta}\} = \theta \quad a < \theta < b.$$

Dans l'exemple de la détermination de la valeur de A ci-dessus, on peut aisément vérifier que $\hat{A} = A$.

3.2 Le critère de variance minimale

On peut légitimement chercher à trouver l'estimateur de variance minimale, fut-il non biaisé. Si on utilise le critère de variance minimale cité ci-dessus, cela nous amène à minimiser

$$\text{mse}[\hat{\theta}] = E\{(\hat{\theta} - \theta)^2\} = \text{var}(\hat{\theta}) - B^2(\theta),$$

qui indique que l'erreur quadratique moyenne est composée d'une erreur due à la variance de l'estimateur et d'une erreur due au biais. En reprenant l'exemple d'une composante continue dans du bruit, on peut considérer l'estimateur modifié :

$$\hat{A} = a \frac{1}{N} \sum_{n=1}^N x_n$$

pour une constante a à déterminer. On a alors

$$\text{mse}(\hat{A}) = \frac{a^2 \sigma^2}{N} + (a - 1)^2 A^2.$$

, en dérivant par rapport à a et en annulant la dérivée, on obtient aisément

$$a_{opt} = \frac{A^2}{A^2 + \sigma^2/N}.$$

On voit donc que l'on obtient une valeur de a qui dépend du paramètre, et l'estimateur MSE n'est pas réalisable. De manière générale, l'obtention des estimateurs à variance minimale (biaisés et non-biaisés) n'est pas triviale, et on doit, pour caractériser la performance d'un estimateur quelconque, le comparer à une borne inférieure pertinente : c'est le cas de la borne de Cramer-Rao.

3.3 Borne de Cramer-Rao

Soit Ξ l'ensemble de toutes les réalisations possibles de N observations de la v.a X soit $\mathbf{x} = (x_1, \dots, x_N) \in \Xi$. On supposera que Ξ ne dépend pas de θ . Soit $\tilde{T}[\mathbf{X}]$ un estimateur non biaisé de θ .

On a les résultats suivants :

$$\begin{aligned} \sum_{\mathbf{x} \in \Xi} Pr[\mathbf{X} = \mathbf{x}] = 1 &\Rightarrow \frac{\partial \sum_{\mathbf{x} \in \Xi} Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta} = 0 \\ \sum_{\mathbf{x} \in \Xi} \frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta} Pr[\mathbf{X} = \mathbf{x}] = 0 &\Rightarrow E\left[\frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta}\right] = 0 \\ \Rightarrow E\left[\theta \frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta}\right] &= 0 \end{aligned}$$

et

$$\begin{aligned} E[\tilde{T}(\mathbf{x})] = \theta &\Rightarrow \sum_{\mathbf{x} \in \Xi} \tilde{T}[\mathbf{x}] Pr[\mathbf{X} = \mathbf{x}] = \theta \Rightarrow \frac{\partial \sum_{\mathbf{x} \in \Xi} \tilde{T}[\mathbf{x}] Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta} = 1 \\ \sum_{\mathbf{x} \in \Xi} \tilde{T}[\mathbf{x}] \frac{\partial Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta} = 1 &\Rightarrow E[\tilde{T}[\mathbf{X}] \frac{\partial \ln(Pr[\mathbf{X}])}{\partial \theta}] = 1 \end{aligned}$$

En soustrayant les deux expressions précédentes, on obtient :

$$E[(\tilde{T}[\mathbf{X}] - \theta) \frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta}] = 1$$

En utilisant l'inégalité de Cauchy-Schwartz on montre que :

$$E[(\tilde{T}[\mathbf{X}] - \theta)^2] E\left[\left(\frac{\partial \ln(Pr[\mathbf{X}])}{\partial \theta}\right)^2\right] \geq 1$$

ce qui donne finalement :

$$\text{Var}[\tilde{T}[\mathbf{X}]] \geq \frac{1}{E\left[\left(\frac{\partial \ln(Pr[\mathbf{X}])}{\partial \theta}\right)^2\right]}$$

Ainsi la variance d'un estimateur non biaisé de θ ne peut pas être inférieure à une valeur appelée borne de Cramer-Rao. Le calcul de cette borne est indépendant de la forme de l'estimateur, il dépend uniquement du modèle probabiliste adopté. Dans le cas d'une variable aléatoire discrète suivant une loi de Bernouilli, on a :

$$\begin{aligned} Pr[\mathbf{X} = \mathbf{x}] = \theta^{\sum_{i=1}^n X_i} (1 - \theta)^{n - \sum_{i=1}^n X_i} &\Rightarrow \ln(Pr[\mathbf{X}]) = \sum_{i=1}^n X_i \ln(\theta) + n - \left(\sum_{i=1}^n X_i\right) \ln(1 - \theta) \\ \frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta} = \frac{\sum_{i=1}^n (X_i - \theta)}{\theta(1 - \theta)} &\Rightarrow E\left[\left(\frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta}\right)^2\right] = \frac{n}{\theta(1 - \theta)} \end{aligned}$$

Remarque :

$$E\left[\left(\frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta}\right)^2\right] = E\left[\left(\frac{\frac{\partial^2 Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta^2}}{Pr[\mathbf{X} = \mathbf{x}]} - \frac{\left(\frac{\partial Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta}\right)^2}{Pr[\mathbf{X} = \mathbf{x}]^2}\right)\right]$$

On a $E\left[\frac{\partial^2 Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta^2}\right] = 0$ donc $E\left[\left(\frac{\partial \ln(Pr[\mathbf{X} = \mathbf{x}])}{\partial \theta}\right)^2\right] = -E\left[\frac{\left(\frac{\partial Pr[\mathbf{X} = \mathbf{x}]}{\partial \theta}\right)^2}{Pr[\mathbf{X} = \mathbf{x}]^2}\right]$

L'estimateur proposé atteint donc la borne pour tout N , il est efficace. On ne peut pas trouver un autre estimateur non biaisé ayant une variance inférieure. Dans ce sens on peut dire que l'on ne peut pas trouver mieux.

Il faut noter que la valeur de la borne diminue lorsque N augmente. Donc plus on fait d'observations et plus l'estimateur est "précis", nous allons clarifier ce point dans la suite.

Théorème 1 (Borne de Cramer-Rao - paramètre scalaire) *On suppose que la PDF $p(\mathbf{x}; \theta)$ satisfait la condition de régularité :*

$$E\left\{\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta}\right\} = 0 \quad \forall \theta$$

où l'espérance est prise par rapport à $p(\mathbf{x}; \theta)$. Alors, la variance de n'importe quel estimateur non biaisé $\hat{\theta}$ satisfait

$$\text{var}(\hat{\theta}) \geq \frac{1}{-E\left\{\frac{\partial^2 \ln p(\mathbf{x}; \theta)}{\partial \theta^2}\right\}} \quad (3.1)$$

où la dérivée est évaluée à la vraie valeur du paramètre et l'espérance est prise par rapport à $p(\mathbf{x}; \theta)$. De plus, un estimateur non biaisé atteint la borne si et seulement si

$$\frac{\partial \ln p(\mathbf{x}; \theta)}{\partial \theta} = I(\theta)(g(\mathbf{x}) - \theta) \quad (3.2)$$

pour une fonction g et I . cet estimateur, qui est l'estimateur MVU, est donné par $\hat{\theta} = g(\mathbf{x})$ et sa variance vaut $1/I(\theta)$.

On notera que $I(\theta)$ est appelée l'information de Fisher. En effet, plus l'information $I(\theta)$ est grande, plus la variance est faible (ce qui est cohérent ...).

Théorème 2 (Borne de Cramer-Rao - Paramètre vectoriel) *On peut étendre le théorème précédent au cas d'un paramètre vectoriel $\boldsymbol{\theta} = [\theta_1 \theta_2 \dots \theta_p]^T$.*

On suppose que $p(\mathbf{x}; \boldsymbol{\theta})$ satisfait les conditions de régularité :

$$E\left\{\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} = 0 \quad \forall \boldsymbol{\theta}$$

où l'espérance est prise par rapport à $p(\mathbf{x}; \boldsymbol{\theta})$. Alors, la matrice de covariance de tout estimateur non biaisé satisfait :

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq 0 \quad (3.3)$$

où ≥ 0 signifie que la matrice est positive semi-définie. En particulier, soit un estimateur non biaisé $\hat{\boldsymbol{\theta}}$, la variance d'un élément vecteur de paramètres est bornée par l'élément diagonal correspondant de l'inverse de la matrice d'information de Fisher :

$$\text{var}(\hat{\theta}_i) \geq [\mathbf{I}^{-1}(\boldsymbol{\theta})]_{ii} \quad (3.4)$$

où $\mathbf{I}(\boldsymbol{\theta})$ est la matrice d'information de Fisher de dimension $p \times p$. Elle est définie par :

$$[\mathbf{I}(\boldsymbol{\theta})]_{ij} = -E \frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \quad (3.5)$$

pour $i = 1, 2, \dots, p; j = 1, 2, \dots, p$. De plus, un estimateur atteint la borne si et seulement si :

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \quad (3.6)$$

pour une fonction \mathbf{g} et une matrice \mathbf{I} . Cet estimateur, qui est l'estimateur MVU, est donné par $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$ et sa matrice de covariance vaut $\mathbf{I}^{-1}(\boldsymbol{\theta})$.

3.4 Statistique suffisante

Une statistique $T(\mathbf{X})$ est suffisante pour le paramètre θ si la loi de \mathbf{X} conditionnelle à $T(\mathbf{X}) = t$ ne dépend pas de θ .

Exemple : Loi de Bernoulli

$$P[X_i = 1] = \theta \quad P[X_i = 0] = 1 - \theta \quad P[\mathbf{X}_1^n = \mathbf{x}_1^n] = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

On prend comme statistique : $T(\mathbf{X}_1^n) = \sum_{i=1}^n X_i$ et $P[T(\mathbf{X}_1^n) = t] = C_t^n \theta^t (1 - \theta)^{n-t}$.

On a alors

$$P[\mathbf{X}_1^n / T(\mathbf{X}_1^n) = t] = \frac{\theta^t (1 - \theta)^{n-t}}{C_t^n \theta^t (1 - \theta)^{n-t}} = \frac{1}{C_t^n}$$

La statistique est suffisante.

Théorème 3 Pour une loi régulière, une statistique $T(\mathbf{X})$ est suffisante pour θ ssi il existe une fonction $g(t, \theta)$ et une fonction $h(\mathbf{X})$ telles que

$$f(\mathbf{X}, \theta) = g(T(\mathbf{X}), \theta) h(\mathbf{X})$$

Chapitre 4

Modèle linéaire et meilleur estimateur linéaire non biaisé (BLUE)

4.1 Introduction

Bon nombre de problèmes peuvent être raisonnablement représentés par un modèle linéaire. De plus, dans ce cas, l'estimateur efficace est très facile à déterminer, et il est utile de donner brièvement les résultats principaux.

4.2 le modèle linéaire

Le modèle le plus simple peut s'écrire sous la forme

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$$

, où \mathbf{w} est un bruit blanc (de matrice de covariance $\sigma^2\mathbf{C}$ et de moyenne nulle. Dans ce cas, l'estimateur MVU est donné par :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} \quad (4.1)$$

et sa covariance est donnée par

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2 (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \quad (4.2)$$

De plus, dans le cas linéaire, l'estimateur MVU est efficace. Cet estimateur étant une transformée linéaire d'un v.a. gaussienne, il a la forme $\hat{\boldsymbol{\theta}} \simeq \mathcal{N}(\boldsymbol{\theta}, \sigma^2 (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1})$.

4.3 BLUE : Best Linear Unbiased Estimators

La recherche d'un estimateur MVU n'est en général pas triviale, et on recourt alors à des estimateurs sous-optimaux. Le BLUE est un de ces estimateurs, qui a l'avantage de ne requérir que les deux premiers moments de la PDF des observations. Si on peut assurer que sa performance est acceptable (en la comparant à la borne de Cramer-Rao par exemple), il peut s'avérer un outil intéressant.

4.3.1 l'estimateur BLUE

Un estimateur linéaire est nécessairement de la forme (on se place dans le cas vectoriel) :

$$\hat{\theta}_i = \sum_{n=1}^N a_{in} x_n \quad i = 1, 2, \dots, p \quad (4.3)$$

soit, sous la forme matricielle :

$$\hat{\boldsymbol{\theta}} = \mathbf{A} \mathbf{x}$$

avec la contrainte de non biais suivante :

$$\mathbf{E} \{ \hat{\boldsymbol{\theta}} \} = \mathbf{A} \mathbf{E} \{ \mathbf{x} \} = \boldsymbol{\theta}$$

De l'équation précédente on voit qu'il faut

$$\mathbf{E} \{ \mathbf{x} \} = \mathbf{H} \boldsymbol{\theta}$$

et que

$$\mathbf{A} \mathbf{H} = \mathbf{I}$$

En définissant $\mathbf{a}_i = [a_{i1} a_{i2} \dots a_{iN}]^T$, tel que $\hat{\theta}_i = \mathbf{a}_i^T \mathbf{x}$ et la matrice $\mathbf{A} = [\mathbf{a}_1^T \mathbf{a}_2^T \dots \mathbf{a}_p^T]^T$. En notant également $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \dots \mathbf{h}_p]$, la contrainte de non biais devient simplement

$$\mathbf{a}_i^T \mathbf{h}_j = \delta_{ij}$$

La variance de l'estimateur est donnée par

$$\text{var}(\hat{\theta}_i) = \mathbf{a}_i^T \mathbf{C} \mathbf{a}_i \quad \mathbf{C} \text{ est la matrice de covariance des observations}$$

On trouve donc l'estimateur en minimisant sa variance, et on obtient :

$$\hat{\boldsymbol{\theta}} = (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} \quad (4.4)$$

et sa covariance est donnée par

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \sigma^2 (\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H})^{-1} \quad (4.5)$$

ce qui est l'estimateur MVU dans le cas d'un modèle linéaire à bruit gaussien. Donc, si les données sont réellement gaussiennes, l'estimateur BLUE est efficace.

Chapitre 5

Estimateur au maximum de vraisemblance (MLE)

C'est une des méthodes les plus utilisées pour contruire des estimateurs car les estimateurs ainsi obtenu possèdent les propriétés suivantes (sous des conditions de régularité) :

- Ils sont asymptotiquement non biaisés.
- Ils atteignent asymptotiquement la borne de Cramer-Rao.
- Leur distribution asymptotique est normale.

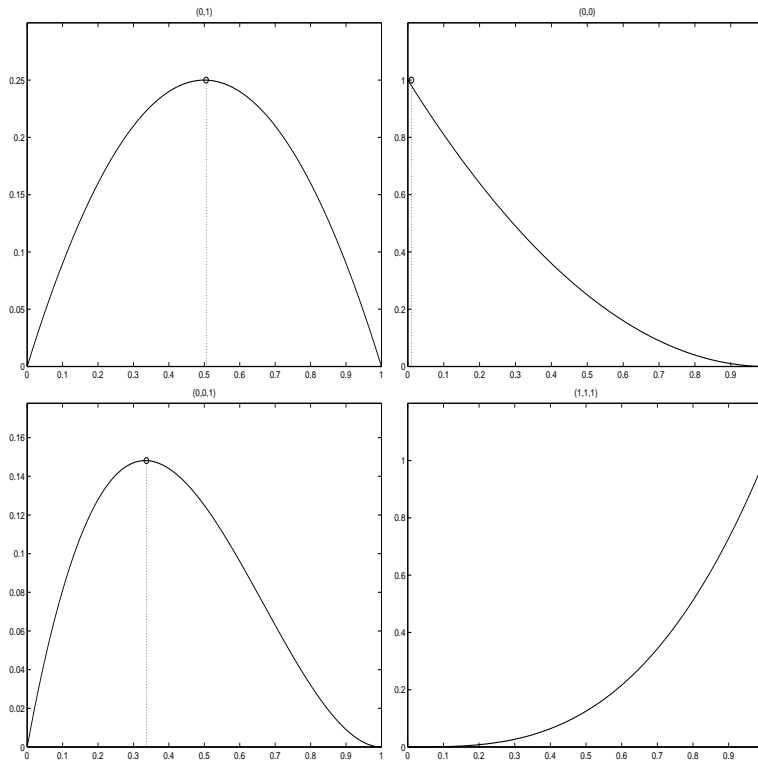
5.1 Introduction

Etudions le problème qui consiste à estimer le probabilité d'apparition d'un 1 pour une v.a de Bernouilli. On va noter : $P(X = 1) = p$ et $P(x = 0) = 1 - p = q$. On cherche la valeur la plus vraisemblable après avoir effectué les observations. On sait que le paramètre que l'on recherche satisfait la contrainte $0 \leq p \leq 1$. Supposons que l'on effectue deux observations $\mathbf{x} = (0, 1)$. Comme on a observé les deux valeurs possibles, il semble naturel de dire que $p = 0.5$. Si on observe $\mathbf{x} = (0, 0)$ on va dire que $p = 0$. Supposons maintenant que l'on fasse 3 observations : $\mathbf{x} = (0, 0, 1)$ on va dire que $p = 1/3$ si on observe $\mathbf{x} = (1, 1, 1)$ on va dire que $p = 1$. Essayons de formaliser notre raisonnement intuitif. On a fait des observations $\mathbf{x} = (x_1, \dots, x_n)$ dont on connait les valeurs. On sait quelle est la probabilité de faire cette observation. Cette probabilité dépend du paramètre p que l'on ne connaît pas. On va donc dire que la valeur la plus vraisemblable du paramètre est celle qui maximise cette probabilité.

On notera la fonction de vraisemblance : $L(\theta) = f(\mathbf{x}; \theta)$. Il faut bien remarquer que la variable de la fonction de vraisemblance est θ et que l'estimateur du maximum de vraisemblance correspond à la valeur de θ qui maximise cette fonction. Dans un premier temps on va supposer que la fonction de vraisemblance ne possède qu'un seul maximum.

Reprenons l'exemple de départ, on voit que :

- Si $\mathbf{x} = (0, 1)$ alors $L(\theta/\mathbf{x}) = p(1-p)$ et le maximum correspond à $p = 0.5$. On peut remarquer que dans ce cas la dérivée seconde est négative ce qui caractérise bien un maximum de la fonction.
- Si $\mathbf{x} = (0, 0)$ alors $L(\theta/\mathbf{x}) = (1-p)^2$ et le maximum correspond à $p = 0$. Ici l'extremum est sur un des bords de la région de recherche.
- Si $\mathbf{x} = (0, 0, 1)$ alors $L(\theta/\mathbf{x}) = p(1-p)^2$ et le maximum correspond à $p = 1/3$.
- Si $\mathbf{x} = (1, 1, 1)$ alors $L(\theta/\mathbf{x}) = p^3$ et le maximum correspond à $p = 1$.



Une autre approche plus rigoureuse consiste à partir de la mesure de Kullback-Leibler

$$KL(f(\mathbf{X}, \theta), f(\mathbf{X}, \bar{\theta})) = E_{\bar{\theta}}[\ln(f(\mathbf{X}, \bar{\theta})/f(\mathbf{X}, \theta))]$$

Cette mesure possède les propriétés suivantes :

1. $KL(f(\mathbf{X}, \bar{\theta}), f(\mathbf{X}, \bar{\theta})) = 0$
2. $KL(f(\mathbf{X}, \theta), f(\mathbf{X}, \bar{\theta})) \geq 0$

Donc la distance de Kullback-Leibler est minimale lorsque $\theta = \bar{\theta}$.

Pour faire le lien avec la méthode du maximum de vraisemblance, il faut remarquer que

$$KL(f(\mathbf{X}, \theta), f(\mathbf{X}, \bar{\theta})) = -E_{\bar{\theta}}[\ln(f(\mathbf{X}, \theta))] + E_{\bar{\theta}}[\ln(f(\mathbf{X}, \bar{\theta}))]$$

Le second terme à droite de l'égalité correspond à l'opposé de l'entropie de la loi $f(\mathbf{X}, \bar{\theta})$ et ne dépend pas de θ . Si on veut minimiser la mesure de Kullback-Leibler il faut maximiser $E_{\bar{\theta}}[\ln(f(\mathbf{X}, \theta))]$, le maximum étant atteint lorsque $\theta = \bar{\theta}$. Si la quantité que l'on veut maximiser, qui est une espérance mathématique, est estimée à partir des observations en utilisant la moyenne arithmétique alors on obtient la méthode du maximum de vraisemblance :

1. $L(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(\mathbf{x}_i, \theta)$
2. Rechercher θ qui maximise $L(\theta)$

5.2 Estimateur du maximum de vraisemblance

On observe n réalisations indépendantes $\mathbf{x} = (x_1, \dots, x_n)$ d'une variable aléatoire X ayant comme densité de probabilité $f(x; \theta)$. La fonction de vraisemblance est définie comme étant une fonction $L(\theta)$ telle que :

$$L(\theta) \propto f(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

On a remplacé dans la définition la relation d'égalité par une relation de proportionnalité de façon à éliminer tous les facteurs multiplicatifs ne dépendant pas du paramètre à estimer θ .

L'estimateur du **maximum de vraisemblance** correspond à la valeur $\hat{\theta}$ qui maximise la fonction $L(\theta)$. Le maximum de vraisemblance correspond à la valeur du paramètre θ qui maximise la probabilité d'apparition des observations (x_1, \dots, x_n) .

En général, il est plus commode de travailler avec la log-vraisemblance

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^n \ln(f(x_i; \theta))$$

5.2.1 Exemples

Loi exponentielle

Supposons que $X_1, \dots, X_n \sim \text{Exp}(\theta)$ donc

$$f(x_i; \theta) = \theta \exp(-\theta x_i) \quad x_i \text{ et } \theta > 0$$

. On obtient :

$$L(\theta) = \prod_{i=1}^n \theta \exp(-\theta x_i) = \theta^n \exp(-\theta \sum x_i)$$

$$l(\theta) = n \ln(\theta) - \theta \sum_{i=1}^n x_i$$

$$\frac{dl(\theta)}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^n x_i = 0 \Rightarrow \boxed{\hat{\theta} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}}$$

ce qui correspond à un maximum car :

$$\frac{d^2 l(\theta)}{d\theta^2} = \frac{-n}{\theta^2} < 0$$

Loi Normale

Supposons que $X_1, \dots, X_n \sim N(\theta, \sigma)$ (σ connu) donc

$$f(x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x_i - \theta)^2 / 2\sigma^2)$$

On obtient :

$$L(\theta) \sim \prod_{i=1}^n \exp(-(x_i - \theta)^2 / 2\sigma^2) = \exp(-\sum_{i=1}^n (x_i - \theta)^2 / 2\sigma^2)$$

$$l(\theta) = -\sum_{i=1}^n (x_i - \theta)^2 / 2\sigma^2$$

$$\frac{dl(\theta)}{d\theta} = 2 \sum_{i=1}^n (x_i - \theta) / 2\sigma^2 = 0 \Rightarrow \boxed{\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}}$$

ce qui correspond à un maximum car :

$$\frac{d^2 l(\theta)}{d\theta^2} = \frac{-n}{2\sigma^2} < 0$$

Loi uniforme

Supposons que $X_1, \dots, X_n \sim U(\theta, 1 + \theta)$ donc

$$f(x_i; \theta) = \begin{cases} 1 & \text{si } x_i \in [\theta, 1 + \theta] \\ 0 & \text{sinon} \end{cases}$$

on obtient sachant que $\max(x_1, \dots, x_n) - \min(x_1, \dots, x_n) \leq 1$:

$$\begin{cases} L(\theta) = 1 & \text{si } \theta \in [\max(x_1, \dots, x_n) - 1, \min(x_1, \dots, x_n)] \\ L(\theta) = 0 & \text{sinon} \end{cases}$$

La fonction de vraisemblance est maximale en une infinité de points, toutes les valeurs de θ comprise entre la valeur minimale et la valeur maximale des observations est un estimateur possible. Il n'y a donc pas unicité de l'estimateur du maximum de vraisemblance.

Regression

Dans les exemples précédents les observations X_i étaient indépendantes et identiquement distribuées. Ici les observations seront toujours indépendantes mais pas identiquement distribuées, les moyennes seront différentes et elles dépendent d'une variable auxiliaire u_i . Supposons que $X_i \sim N(\theta u_i, \sigma)$ et que les variables u_i $i = 1, \dots, n$ et σ soient connus. Alors

$$f(x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x_i - \theta u_i)^2 / 2\sigma^2)$$

On obtient :

$$L(\theta) \sim \prod_{i=1}^n \exp(-(x_i - \theta u_i)^2 / 2\sigma^2) = \exp(-\sum_{i=1}^n (x_i - \theta u_i)^2 / 2\sigma^2)$$

$$l(\theta) = -\sum_{i=1}^n (x_i - \theta u_i)^2 / 2\sigma^2$$

$$\frac{dl(\theta)}{d\theta} = 2 \sum_{i=1}^n u_i (x_i - \theta u_i) / 2\sigma^2 = 0 \Rightarrow \hat{\theta} = \frac{\sum_{i=1}^n x_i u_i}{\sum_{i=1}^n u_i^2}$$

ce qui correspond à un maximum car :

$$\frac{d^2l(\theta)}{d\theta^2} = \frac{-\sum_{i=1}^n u_i^2}{2\sigma^2} < 0$$

Il est facile de calculer la moyenne et la variance de l'estimateur. On trouve

$$E[\hat{\theta}] = \theta \text{ et } \text{Var}[\hat{\theta}] = \sigma^2 / \sum_i u_i^2$$

Série Temporelle

Dans cet exemple les observations X_i ne sont plus indépendantes. Supposons que $X_1 \sim N(0, 1)$ et pour $i = 1, \dots, n$

$$X_i / (X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \sim N(\theta x_{i-1}, 1)$$

C'est un processus de Markov la v.a X_i ne dépend que de v.a. précédente X_{i-1} . On peut donc écrire :

$$\begin{aligned} f(x_1, \dots, x_n; \theta) &= f(x_1/\theta) f(x_2/x_1, \theta) \dots f(x_n/x_{n-1}, \theta) \\ f(x_1; \theta) &= \frac{1}{\sqrt{2\pi}} \exp(-x_1^2/2) \\ f(x_i/x_{i-1}; \theta) &= \frac{1}{\sqrt{2\pi}} \exp(-(x_i - \theta x_{i-1})^2/2) \end{aligned}$$

La fonction de vraisemblance s'écrit :

$$\begin{aligned} L(\theta) &= f(x_1; \theta) \prod_{i=2}^n f(x_i/x_{i-1}; \theta) \\ L(\theta) &\sim \exp(-x_1^2/2) \prod_{i=2}^n \exp(-(x_i - \theta x_{i-1})^2/2) \\ L(\theta) &\sim \exp(-x_1^2/2) \exp\left(-\sum_{i=2}^n (x_i - \theta x_{i-1})^2/2\right) \end{aligned}$$

On obtient donc :

$$\begin{aligned} l(\theta) &\sim -x_1^2 - \sum_{i=2}^n (x_i - \theta x_{i-1})^2 \\ \frac{dl(\theta)}{d\theta} &\sim \sum_{i=2}^n x_{i-1}(x_i - \theta x_{i-1}) = 0 \Rightarrow \boxed{\hat{\theta} = \frac{\sum_{i=2}^n x_{i-1} x_i}{\sum_{i=2}^n x_{i-1}^2}} \end{aligned}$$

ce qui correspond à un maximum car :

$$\frac{d^2l(\theta)}{d\theta^2} = -\sum_{i=2}^n x_{i-1}^2 < 0$$

5.3 Invariance de l'estimateur MV

Si $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ alors $g(\hat{\theta})$ correspond à l'estimateur du maximum de vraisemblance de $g(\theta)$.

Supposons dans un premier temps que la fonction $g(\cdot)$ est bijective. Alors le résultat est évident car si le maximum de $L(\theta)$ se trouve en $\hat{\theta}$ alors le maximum de $L(g(\theta))$ sera en $L(g(\hat{\theta}))$.

Si la fonction $g(\cdot)$ n'est pas bijective alors l'équation $g^{-1}(y)$ peut avoir plusieurs solutions. Pour contourner ce problème on va regrouper les solutions ayant le même antécédent et prendre le maximum sur ces solutions. Soit $G^{-1}(y) = \{\theta : g(\theta) = y\}$ l'ensemble des antécédents de y et notons $L(y)$ la valeur maximale atteinte par $L(\theta)|_{\theta \in G^{-1}(y)}$ soit :

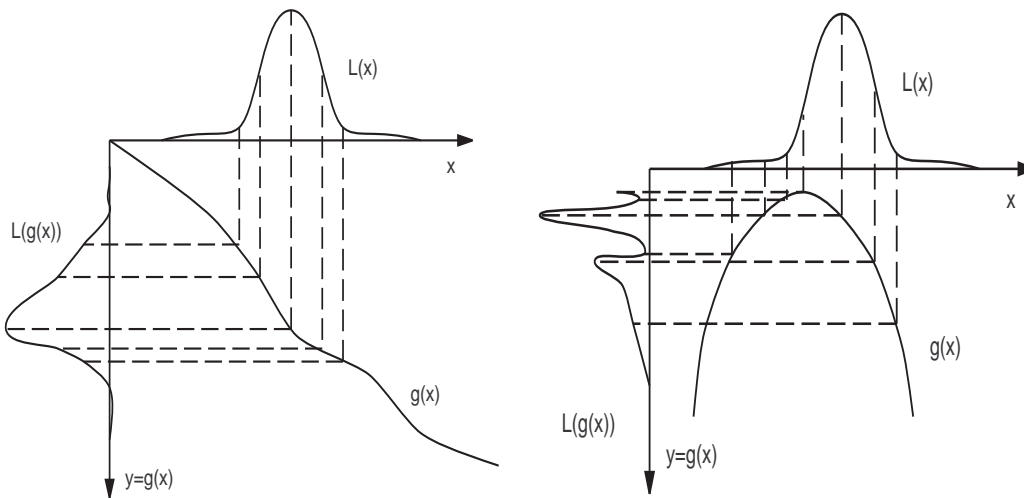
$$L(y) = \max_{\theta \in G^{-1}(y)} L(\theta)$$

alors le maximum de $L(y)$ est atteinte pour \hat{y} ce qui correspond à

$$L(\hat{y}) = \max_y \max_{\theta \in G^{-1}(y)} L(\theta)$$

les deux maximisations de droites donnent :

$$\boxed{\hat{y} = g(\hat{\theta})}$$



5.3.1 Exemple 1

1. Soit (X_1, \dots, X_n) un ensemble de v.a iid avec

$$X_i \sim \frac{1}{\theta} \exp(-x/\theta) \text{ pour } x \geq 0 \text{ et } \theta > 0$$

On a $E[X_i] = \theta$ et $\text{Var}[X_i] = \theta^2$.

Il est facile de calculer l'estimateur du maximum de vraisemblance de θ on trouve :

$$\hat{\theta}_{MV} = \frac{1}{n} \sum_{i=1}^n x_i$$

donc $E[\hat{\theta}_{MV}] = \theta$ et $\text{Var}[\hat{\theta}_{MV}] = \theta^2/n$. L'estimateur est non biaisé.

Maintenant on change la paramétrisation et on cherche l'estimateur de $\phi = \theta^2$ donc ici $g(\theta) = \theta^2$ mais comme $\theta \geq 0$ la relation est bijective. L'estimateur du maximum de vraisemblance de ϕ sera :

$$\hat{\phi}_{MV} = \hat{\theta}_{MV}^2$$

et il est facile de voir que

$$E[\hat{\phi}_{MV}] = \theta^2 \left(1 + \frac{1}{n}\right)$$

l'estimateur est donc biaisé mais asymptotiquement non biaisé.

2. Soit (X_1, \dots, X_n) un ensemble de v.a iid avec

$$X_i \sim N(\theta, \sigma^2)$$

σ^2 ayant une valeur connue. L'estimateur du maximum de vraisemblance de θ est $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ qui est non biaisé et de variance σ^2/n .

Supposons maintenant que l'on cherche à estimer $\theta^2 = g(\theta)$ alors on va prendre comme estimateur

$$\hat{\theta}^2 = \bar{x}^2$$

Il est facile de voir que cet estimateur est biaisé $E[\hat{\theta}^2] = \theta^2 + \sigma^2/n$.

Chapitre 6

Estimateur du maximum de vraisemblance multidimensionnelle

6.1 Vraisemblance

Supposons que $\mathbf{x}_1^n = (\mathbf{x}_1, \dots, \mathbf{x}_n)^t$ sont des réalisations **iid** d'une v.a $\mathbf{X}^{p \times 1}$ de densité de probabilité $f(\mathbf{x}; \theta)$ o $\theta^{m \times 1} = (\theta_1, \dots, \theta_m)^t$ est un vecteur de paramètres. On définit la fonction de vraisemblance de la façon suivante :

$$L(\theta; \mathbf{x}_1^n) = L(\theta) = \prod_{i=1}^n f(\mathbf{x}_i; \theta)$$

De même la fonction de log-vraisemblance est définie de la façon suivante :

$$l(\theta; \mathbf{x}_1^n) = l(\theta) = \sum_{i=1}^n \ln(f(\mathbf{x}_i; \theta))$$

Pour un ensemble de valeurs de $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ fixé la fonction de vraisemblance est une fonction de θ .

Mais pour tenir compte des différentes réalisations possibles, il faut considérer la fonction de vraisemblance comme une v.a. que l'on notera $L(\theta; \mathbf{X})$ ou pour la log-vraisemblance $l(\theta; \mathbf{X})$.

6.1.1 Exemple

Supposons que $\mathbf{X} \sim N_p(\theta, \Sigma)$ avec $\Sigma^{p \times p}$ connue alors on a :

$$L(\theta) = |2\pi\Sigma|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \theta)^t \Sigma^{-1} (\mathbf{x}_i - \theta)\right)$$

et

$$l(\theta) = -\frac{n}{2} \ln(|2\pi\Sigma|) - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \theta)^t \Sigma^{-1} (\mathbf{x}_i - \theta)$$

Pour simplifier l'expression de la log-vraisemblance notons que :

$$(\mathbf{x}_i - \theta)^t \Sigma^{-1} (\mathbf{x}_i - \theta) = (\mathbf{x}_i - \bar{\mathbf{x}})^t \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) + (\bar{\mathbf{x}} - \theta)^t \Sigma^{-1} (\bar{\mathbf{x}} - \theta) + 2(\bar{\mathbf{x}} - \theta)^t \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

ce qui donne :

$$\sum_{i=1}^n (\mathbf{x}_i - \theta)^t \Sigma^{-1} (\mathbf{x}_i - \theta) = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^t \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) + n(\bar{\mathbf{x}} - \theta)^t \Sigma^{-1} (\bar{\mathbf{x}} - \theta)$$

Comme $(\mathbf{x}_i - \bar{\mathbf{x}})^t \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ est un scalaire on peut écrire :

$$(\mathbf{x}_i - \bar{\mathbf{x}})^t \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) = \text{tr}[\Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t]$$

En posant $nS = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t$ on obtient l'expression :

$$l(\theta) = -\frac{n}{2} \ln(|2\pi\Sigma|) - \frac{n}{2} \text{tr}[\Sigma^{-1}S] - \frac{n}{2} (\bar{\mathbf{x}} - \theta)^t \Sigma^{-1} (\bar{\mathbf{x}} - \theta)$$

Si $\Sigma = I$ alors :

$$l(\theta) = -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \text{tr}[S] - \frac{n}{2} (\bar{\mathbf{x}} - \theta)^t (\bar{\mathbf{x}} - \theta)$$

6.2 Caractéristiques des estimateurs

Soit un estimateur $T(\mathbf{x}_1^n)^{m \times 1}$ une fonction mesurable des v.a \mathbf{X}_1^n . Alors on définit le biais de l'estimateur comme

$$B(T) = E[T - \theta]$$

et l'erreur quadratique moyenne comme

$$\begin{aligned} C^{m \times m} &= E[(T - \theta)(T - \theta)^t] = E[(T - E[T] + E[T] - \theta)(T - E[T] + E[T] - \theta)^t] \\ &= E[(T - E[T])(T - E[T])^t] + B(T)B(T)^t = \text{Covariance} + \text{biais} \end{aligned}$$

6.3 La fonction score

On définit la fonction score comme étant le gradient de la log-vraisemblance :

$$\mathbf{u}(\theta) = \mathbf{u}(\theta, \mathbf{x}_1^n) = \nabla_{\theta} l(\theta; \mathbf{x}_1^n) = (\partial l(\theta) / \partial \theta_1, \dots, \partial l(\theta) / \partial \theta_m)^t$$

La fonction score est une fonction mesurable par rapport à \mathbf{X}_1^n aussi c'est une variable aléatoire que l'on notera $\mathbf{U}(\theta)$.

La covariance de cette v.a est la matrice d'information de **Fisher** que l'on notera $I(\theta)$

$$I(\theta) = E[\mathbf{U}(\theta)\mathbf{U}(\theta)^t]$$

6.4 Propriétés de la fonction score

On a besoin d'un certain nombre de propriétés de régularité sur la loi des observations pour faire le calcul. On supposera que toutes les conditions nécessaires sont vérifiées.

Théorème 4 On a

$$E[\mathbf{U}(\theta)] = 0$$

Démonstration 1

$$\begin{aligned} E[\mathbf{U}(\theta)] &= \int \nabla_{\theta} \ln(f(\mathbf{x}; \theta)) f(\mathbf{x}; \theta) d\mathbf{x} \\ &= \int \nabla_{\theta} f(\mathbf{x}; \theta) d\mathbf{x} = \nabla_{\theta} \int f(\mathbf{x}; \theta) d\mathbf{x} = 0 \end{aligned}$$

Théorème 5

$$I(\theta) = E[\nabla_{\theta} \ln(f(\mathbf{x}; \theta)) \nabla_{\theta^t} \ln(f(\mathbf{x}; \theta))] = -E[\nabla_{\theta} \nabla_{\theta^t} \ln(f(\mathbf{x}; \theta))]$$

$$I_{ij} = E[\nabla_{\theta_i} \ln(f(\mathbf{x}; \theta)) \nabla_{\theta_j} \ln(f(\mathbf{x}; \theta))] = -E[\nabla_{\theta_i} \nabla_{\theta_j} \ln(f(\mathbf{x}; \theta))]$$

Démonstration 2 On peut écrire

$$\begin{aligned} E[\nabla_{\theta} \nabla_{\theta^t} \ln(f(\mathbf{x}; \theta))] &= E\left[\frac{f(\mathbf{x}; \theta) \nabla_{\theta} \nabla_{\theta^t} f(\mathbf{x}; \theta) - \nabla_{\theta} f(\mathbf{x}; \theta) \nabla_{\theta^t} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)^2}\right] \\ &= E\left[\frac{\nabla_{\theta} \nabla_{\theta^t} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} - \nabla_{\theta} \ln(f(\mathbf{x}; \theta)) \nabla_{\theta^t} \ln(f(\mathbf{x}; \theta))\right] \\ &= -E[\nabla_{\theta} \ln(f(\mathbf{x}; \theta)) \nabla_{\theta^t} \ln(f(\mathbf{x}; \theta))] \end{aligned}$$

On a utilisé les résultats suivants :

- $\nabla_{\theta} \ln(f(\mathbf{x}; \theta)) = \nabla_{\theta} f(\mathbf{x}; \theta) / f(\mathbf{x}; \theta)$
- $E[\nabla_{\theta} \mathbf{U}(\theta)^t] = E[\nabla_{\theta} \nabla_{\theta^t} f(\mathbf{x}; \theta) / f(\mathbf{x}; \theta)]$

On notera J la matrice d'information stochastique

$$J = \mathbf{u}(\theta) \mathbf{u}(\theta)^t \text{ ou } J_{ij} = \nabla_{\theta_i} \nabla_{\theta_j} \ln(f(\mathbf{x}; \theta))$$

Théorème 6 Si $\mathbf{T}(\mathbf{x}_1^n)$ est un estimateur non biaisé de θ alors on a

$$E[\mathbf{U}(\theta)(\mathbf{T} - \theta)^t] = I^{m \times m}$$

Démonstration 3 Comme $E[(\mathbf{T} - \theta)^t] = 0^t$ alors

$$\begin{aligned} \int (\mathbf{T} - \theta)^t f(\mathbf{x}, \theta) d\mathbf{x} &= 0^t \text{ alors } \nabla_{\theta} \left(\int (\mathbf{T} - \theta)^t f(\mathbf{x}, \theta) d\mathbf{x} \right) = \int \nabla_{\theta} ((\mathbf{T} - \theta)^t f(\mathbf{x}, \theta)) d\mathbf{x} \\ &= \int \nabla_{\theta} (f(\mathbf{x}, \theta)) (\mathbf{T} - \theta)^t d\mathbf{x} - \int f(\mathbf{x}, \theta) I d\mathbf{x} = \nabla_{\theta} 0^t = 0^{m \times m} \Rightarrow \int \nabla_{\theta} (f(\mathbf{x}, \theta)) (\mathbf{T} - \theta)^t d\mathbf{x} = I \\ \Rightarrow \int \nabla_{\theta} \ln(f(\mathbf{x}, \theta)) (\mathbf{T} - \theta)^t f(\mathbf{x}, \theta) d\mathbf{x} &= I \Rightarrow E[\mathbf{U}(\theta)(\mathbf{T} - \theta)^t] = I^{m \times m} \end{aligned}$$

6.5 Borne de Cramer-Rao

Théorème 7 Si $\mathbf{T}(\mathbf{x}_1^n)$ est un estimateur non biaisé de θ alors sa matrice de covariance vérifie la condition suivante

$$C = E[(\mathbf{T} - \theta)(\mathbf{T} - \theta)^t] \geq I^{-1}$$

ce qui signifie que $C - I^{-1}$ est une matrice non-définie négative.

Démonstration 4 Soit le vecteur

$$\begin{pmatrix} (\mathbf{T} - \theta) \\ \mathbf{U}(\theta) \end{pmatrix}$$

alors

$$Q = E\left[\begin{pmatrix} (\mathbf{T} - \theta) \\ \mathbf{U}(\theta) \end{pmatrix} \begin{pmatrix} (\mathbf{T} - \theta)^t & \mathbf{U}(\theta)^t \end{pmatrix}\right] = \begin{pmatrix} C & I \\ 0 & I(\theta) \end{pmatrix}$$

Comme Q est une matrice non-définie négative symétrique on peut la diagonaliser

$$\begin{pmatrix} I & -I(\theta)^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} C & I \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ -I(\theta)^{-1} & I \end{pmatrix} = \begin{pmatrix} C - I(\theta)^{-1} & 0 \\ 0 & I(\theta) \end{pmatrix}$$

Comme Q est définie non-négative il en va de même pour $C - I(\theta)^{-1}$

Remarques :

1. L'estimateur du maximum de vraisemblance est solution de l'équation $\mathbf{u}(\hat{\theta}_{MV}; \mathbf{x}_1^n) = 0$
2. On a pour tout θ^*

$$E_{\theta}[l(\theta^*) - l(\theta)] \leq 0$$

Démonstration 5

$$E_{\theta}[\ln \frac{f(\mathbf{X}_1^n; \theta^*)}{f(\mathbf{X}_1^n; \theta)}] \leq \ln E_{\theta}[\frac{f(\mathbf{X}_1^n; \theta^*)}{f(\mathbf{X}_1^n; \theta)}] = \ln(1) = 0$$

On a l'égalité lorsque

$$l(\theta^*) = l(\theta)$$

On a utilisé l'inégalité de Jensen : $E[g(\mathbf{X})] \leq g(E[\mathbf{X}])$ si $g(\cdot)$ est une fonction concave comme par exemple la fonction $\log(\cdot)$.

Donc la fonction de vraisemblance est maximum en θ paramètre correspondant à la densité de probabilité des observations.

6.5.1 Exemples**Estimation de la moyenne d'une loi normale**

Supposons que $\mathbf{X} \sim \mathcal{N}(\theta, I)$ alors on a :

$$\begin{aligned} l(\theta) &= -\frac{np}{2} \ln(2\pi) - \frac{n}{2} \text{tr}(S) - \frac{n}{2} (\bar{\mathbf{x}} - \theta)^t (\bar{\mathbf{x}} - \theta) \\ \text{donc } \mathbf{u}(\theta) &= n(\bar{\mathbf{x}} - \theta) \text{ et } \hat{\theta}_{MV} = \bar{\mathbf{x}} \text{ et } I(\theta) = nI \end{aligned}$$

6.6 Influence de la paramétrisation sur la Borne

Supposons que $\nu = g(\theta)$ alors d'après la propriété d'invariance du maximum de vraisemblance on a $\hat{\nu}_{MV} = g(\hat{\theta}_{MV})$. On va supposer que la fonction g est continue et inversible. On a dans

$$l(\nu) = l(\theta = g^{-1}(\nu)) \text{ et } l(\theta) = l(\nu = g(\theta))$$

mais en généralisation la formule

$$\frac{\partial f(g(x))}{\partial x} = \frac{\partial g(x)}{\partial x} \frac{\partial f(g(x))}{\partial g(x)}$$

on obtient

$$\nabla_{\nu_i} l(\nu) = \sum_{j=1}^m \nabla_{\nu_i} \theta_j \nabla_{\theta_j} l(\theta) = \nabla_{\nu_i} \theta^t \nabla_{\theta} l(\theta)$$

de même on a

$$\nabla_{\theta_i} l(\theta) = \sum_{j=1}^m \nabla_{\theta_i} \nu_j \nabla_{\nu_j} l(\nu) = \nabla_{\theta_i} \nu^t \nabla_{\nu} l(\nu)$$

L'écriture sous forme matricielle est

$$\nabla_{\nu} l(\nu) = \begin{pmatrix} \nabla_{\nu_1} \theta^t \nabla_{\theta} l(\theta) \\ \vdots \\ \nabla_{\nu_m} \theta^t \nabla_{\theta} l(\theta) \end{pmatrix} = G \nabla_{\theta} l(\theta)$$

On a de même

$$\nabla_{\theta} l(\theta) = H \nabla_{\nu} l(\nu)$$

avec $GH = HG = I$. Donc

$\begin{aligned} I(\nu) &= GI(\theta)G^t && \text{avec } \theta = g^{-1}(\nu) \\ I(\theta) &= HI(\nu)H^t && \text{avec } \nu = g(\theta) \end{aligned}$

6.6.1 Exemple

Soit

$$\mathbf{x}_i = \mathbf{s}(\theta) + \epsilon_i$$

donc

$$X_i \sim N(\mathbf{s}(\theta)^{p \times 1}, \sigma^2 I^{p \times p})$$

En posant $\nu = \mathbf{s}(\theta)$ on a vu que

$$I(\nu) = \frac{n}{\sigma^2} I^{p \times p}$$

On va calculer $I(\theta)$:

$$I(\theta) = HI(\nu)H^t = \frac{n}{\sigma^2} HH^t$$

avec

$$H = \begin{pmatrix} \nabla_{\theta_1} \mathbf{s}(\theta)^t \\ \vdots \\ \nabla_{\theta_m} \mathbf{s}(\theta)^t \end{pmatrix}$$

Si θ est un paramètre scalaire et

$$\mathbf{s}(\theta) = \begin{pmatrix} s_1(\theta) \\ \vdots \\ s_p(\theta) \end{pmatrix}$$

alors

$$HH^t = \sum_{i=1}^p \left(\frac{\partial s_i(\theta)}{\partial \theta} \right)^2$$

et

$$I(\theta) = \frac{1}{\sigma^2 \sum_{i=1}^p \left(\frac{\partial s_i(\theta)}{\partial \theta} \right)^2}$$

6.7 Détermination numérique du maximum de vraisemblance

Soit θ la valeur exacte du paramètre que l'on recherche et soit $\hat{\theta}_n$ un estimateur de ce paramètre. On peut supposer que la log-vraisemblance est quadratique au alentour de θ ce qui revient à dire que l'on peut écrire

$$l(\theta) = l(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^t \nabla_{\theta} l(\hat{\theta}_n) + \frac{1}{2} (\theta - \hat{\theta}_n)^t \nabla_{\theta} (\nabla_{\theta^t} l(\hat{\theta}_n)) (\theta - \hat{\theta}_n)$$

Le maximum de vraisemblance est une des solutions de l'équation

$$\nabla_{\theta} l(\theta) = 0$$

ce qui donne en utilisant l'approximation précédente

$$\mathbf{u}(\theta) = \mathbf{u}(\hat{\theta}_n) - J(\hat{\theta}_n)(\theta - \hat{\theta}_n) = 0$$

Ceci suggère l'algorithme récursif

$$\hat{\theta}_{n+1} = \hat{\theta}_n + J^{-1}(\hat{\theta}_n)u(\hat{\theta}_n)$$

On peut remplacer dans cette algorithme $J(\hat{\theta}_n)$ par $I(\hat{\theta}_n)$.

6.8 Exemples

6.8.1 Modèle linéaire généralisé

L'abréviation anglo-saxonne est GLM pour "Generalized Linear Model". L'observation suit le modèle suivant :

$$\mathbf{Y}^{n \times 1} = X^{n \times k} \mathbf{b}^{k \times 1} + \epsilon^{n \times 1} \text{ et } \mathbf{Y} \sim N(X\mathbf{b}, \sigma^2 I^{n \times n})$$

Ici $\theta^{(k+1) \times 1} = \begin{pmatrix} \mathbf{b}^{k \times 1} \\ \sigma^2 \end{pmatrix}$ et la fonction de log-vraisemblance est

$$l(\theta) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - X\mathbf{b})^t (\mathbf{y} - X\mathbf{b})$$

Calculons le gradient de la log-vraisemblance :

$$\begin{aligned} \nabla_{\mathbf{b}} l(\theta) &= \frac{-1}{2\sigma^2} \nabla_{\mathbf{b}} (\mathbf{y}\mathbf{y}^t - 2\mathbf{b}^t X\mathbf{y} + \mathbf{b}^t X^t X\mathbf{b}) = \frac{-1}{2\sigma^2} (-2X^t \mathbf{y} + 2X^t X\mathbf{b}) \\ \nabla_{\sigma^2} l(\theta) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - X\mathbf{b})^t (\mathbf{y} - X\mathbf{b}) \end{aligned}$$

La résolution de $\nabla_{\theta} l(\theta) = 0$ donne comme solutions

$$\hat{\mathbf{b}} = (X^t X)^{-1} X^t \mathbf{y} \text{ et } \hat{\sigma}^2 = \frac{(\mathbf{y} - X\hat{\mathbf{b}})^t (\mathbf{y} - X\hat{\mathbf{b}})}{n}$$

6.9 Loi normale multidimensionnelle

On a vu que la loi multidimensionnelle s'écrivait sous la forme :

$$f(\mathbf{x}_1^n; \theta) = (2\pi)^{np} \det(\Sigma(\theta))^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}(\theta))^t \Sigma(\theta)^{-1} (\mathbf{x}_i - \mathbf{m}(\theta))\right)$$

avec $\theta^{m \times 1}$, $\Sigma(\theta)^{p \times p}$ et $\mathbf{m}(\theta)^{p \times 1}$. On peut montrer que la fonction score $\mathbf{u}(\theta) = (\mathbf{u}_1(\theta) \dots \mathbf{u}_p(\theta))^t$ s'écrit :

$$\mathbf{u}_k(\theta) = -\frac{1}{2} \text{tr}(\Sigma^{-1} \nabla_{\theta_k} \Sigma) + \frac{n}{2} \text{tr}(\Sigma^{-1} \nabla_{\theta_k} (\Sigma \Sigma^{-1} S)) - \frac{n}{2} \text{tr}(\Sigma^{-1} \nabla_{\theta_k} S)$$

avec :

$$S = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}(\theta)) (\mathbf{x}_i - \mathbf{m}(\theta))^t$$

De même on peut montrer que la matrice d'information de Fisher s'écrit :

$$I_{ij} = \frac{n}{2} \text{tr}(\Sigma^{-1} \nabla_{\theta_i} (\Sigma) \Sigma^{-1} \nabla_{\theta_j} \Sigma) + n \nabla_{\theta_i} \mathbf{m}^t \Sigma^{-1} \nabla_{\theta_j} \mathbf{m}^t$$

6.9.1 Cas Σ connue

Dans ce cas là on a :

$$\begin{aligned}
 \mathbf{u}_k(\theta) &= -\frac{n}{2} \text{tr}(\Sigma^{-1} \nabla_{\theta_k} S) = \text{tr}(\Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}(\theta)) \nabla_{\theta_k} \mathbf{m}(\theta)^t) \\
 &= \nabla_{\theta_k} \mathbf{m}(\theta)^t \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}(\theta)) \\
 \text{donc } \mathbf{u}(\theta) &= (u_1(\theta), \dots, u_p(\theta))^t \\
 &= \begin{pmatrix} \nabla_{\theta_1} \mathbf{m}(\theta)^t \\ \vdots \\ \nabla_{\theta_m} \mathbf{m}(\theta)^t \end{pmatrix} \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}(\theta)) \\
 &= \nabla_{\theta} \mathbf{m}(\theta) \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{m}(\theta))
 \end{aligned}$$

et

$$\begin{aligned}
 I(\theta) = E[\mathbf{u}(\theta) \mathbf{u}(\theta)] &= E[\nabla_{\theta} \mathbf{m}(\theta) \Sigma^{-1} \sum_{i=1}^n (\mathbf{X}_i - \mathbf{m}(\theta)) \sum_{i=1}^n (\mathbf{X}_i - \mathbf{m}(\theta))^t \Sigma^{-1} \nabla_{\theta} (\mathbf{m}(\theta))^t] \\
 &= \nabla_{\theta} \mathbf{m}(\theta) \Sigma^{-1} n \Sigma \Sigma^{-1} \nabla_{\theta} (\mathbf{m}(\theta))^t \\
 I(\theta) &= n \nabla_{\theta} \mathbf{m}(\theta) \Sigma^{-1} \nabla_{\theta} (\mathbf{m}(\theta))^t
 \end{aligned}$$

6.9.2 Σ connue et modèle linéaire

Dans ce cas on

$$\mathbf{m}(\theta)^{p \times 1} = H^{p \times m} \theta^{m \times 1} \quad \text{et} \quad \nabla_{\theta^t} \mathbf{m}(\theta) = H^t$$

alors la fonction score s'écrit

$$\mathbf{u}(\theta) = H^t \Sigma^{-1} \sum_{i=1}^n (x_i - H\theta)$$

Le maximum de vraisemblance est la solution de l'équation $\mathbf{u}(\theta) = 0$ soit

$$\begin{aligned}
 H^t \Sigma^{-1} \sum_{i=1}^n x_i - n H^t \Sigma^{-1} H \theta &= 0 \\
 \hat{\theta}_{MV} = \frac{1}{n} (H^t \Sigma^{-1} H)^{-1} H^t \Sigma^{-1} \sum_{i=1}^n x_i &= (H^t \Sigma^{-1} H)^{-1} H^t \Sigma^{-1} \bar{x} \\
 \text{avec } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i &
 \end{aligned}$$

La matrice d'information de Fisher vaut :

$$I(\theta) = n \frac{\partial \mathbf{m}(\theta)}{\partial \theta^t} \Sigma^{-1} \frac{\partial \mathbf{m}(\theta)}{\partial \theta} = n H^t \Sigma^{-1} H$$

On a

$$E[\hat{\theta}_{MV}] = \theta \quad \text{car} \quad E[\bar{x}] = H\theta$$

et

$$E[(\hat{\theta}_{MV} - \theta)(\hat{\theta}_{MV} - \theta)^t] = \frac{1}{n}(H^t \Sigma^{-1} H)^{-1}$$

ce qui montre que l'estimateur atteint la borne de Cramer-Rao pour toutes les valeurs de n (et pas uniquement de façon asymptotique) ce qui est remarquable.

Chapitre 7

Estimation : approche Bayésienne

Dans l'approche Bayésienne, les paramètres sont considérés comme étant des variables aléatoires.

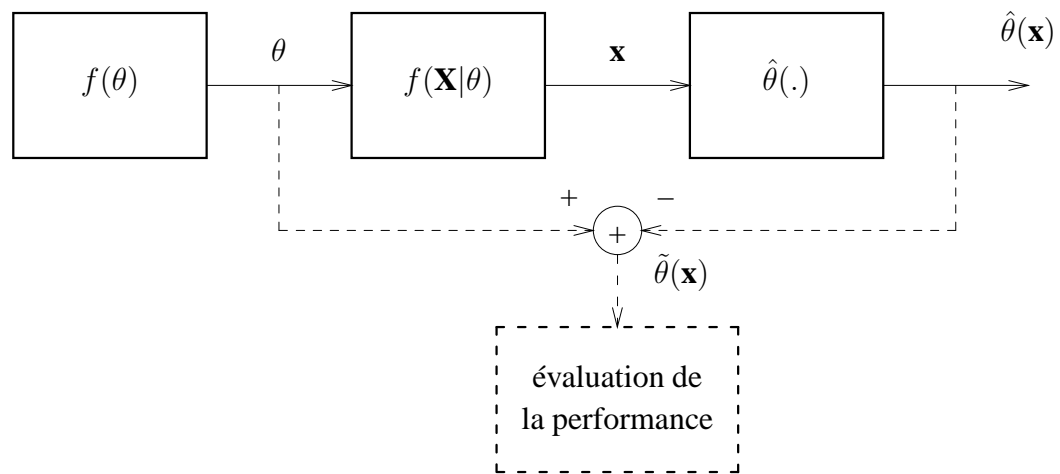


FIG. 7.1: relation entre paramètres, observations et estimées

Chapitre 8

Théorie de la détection

8.1 Test d'hypothèses

8.1.1 Introduction

Dans un test d'hypothèses, à partir d'un vecteur d'observations $\mathbf{x} = [x_1 \cdots x_N]^T$, on doit décider quelle hypothèse est la plus probable, étant donné ces observations.

Exemple 8.1 *Test d'hypothèse binaire*

Soit $x_n \simeq \mathcal{N}(\theta, 1)$, où θ est une inconnue, que l'on suspecte avoir les valeurs $\theta = -1$ ou $\theta = 1$:

$$\mathcal{H}_0 : \theta = -1$$

$$\mathcal{H}_1 : \theta = 1$$

◁

En général, il y a $M \geq 2$ hypothèses $\mathcal{H}_1, \dots, \mathcal{H}_M$ (par exemple, un système de reconnaissance de chiffres pour les codes postaux aura 10 hypothèses : les chiffres de 0 à 9).

Une **règle de décision** est utilisée pour décider entre M hypothèses $\mathcal{H}_1, \dots, \mathcal{H}_M$. Cette règle de décision est un mapping de

$$h : \mathbb{R}^N \rightarrow \{\mathcal{H}_1, \dots, \mathcal{H}_M\}$$

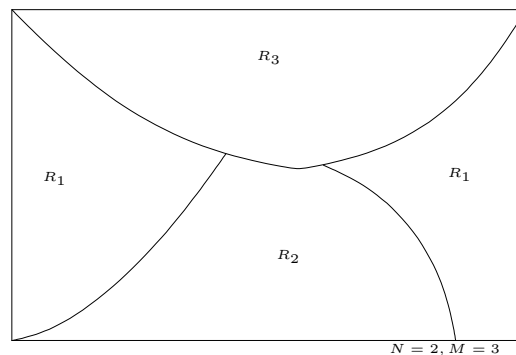
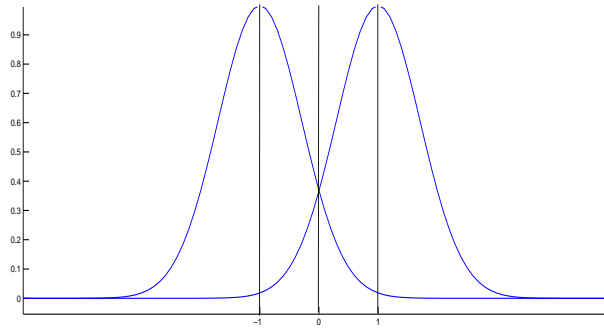


FIG. 8.1: Exemple de partition de l'espace en 3 régions de décisions

Exemple 8.2 *Test binaire avec deux observations*

Soit $\mathbf{x} = [x_1 \ x_2]^T$ où $x_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ et le test

$$\begin{aligned} \mathcal{H}_0 : \theta &= -1 \\ \mathcal{H}_1 : \theta &= 1 \end{aligned}$$



Un test intuitivement OK est :

$$h(\mathbf{x}) = \begin{cases} \mathcal{H}_0 & \text{si } \bar{x} \leq 0 \\ \mathcal{H}_1 & \text{si } \bar{x} > 0 \end{cases}$$

$$\text{où } \bar{x} = \frac{1}{2}(x_1 + x_2)$$

◁

8.1.2 Terminologie

Hypothèse simple Si une hypothèse \mathcal{H} spécifie une distribution unique pour \mathbf{x} , on dit que \mathcal{H} est une hypothèse simple.

Exemple 8.3 hypothèse simple

On veut détecter la présence d'une sinusoïde de fréquence connue f_o et de phase ϕ inconnue :

$$s_n = \cos(2\pi f_o n + \phi), \quad n = 0, 1, \dots, N-1$$

soit $w_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, σ^2 connu, L'hypothèse

$$\mathcal{H}_0 : \mathbf{x} = \mathbf{w}$$

est une hypothèse simple. ◁

Hypothèse composite Si une hypothèse \mathcal{H} spécifie une classe de distributions possibles pour \mathbf{x} , on dit que \mathcal{H} est une hypothèse composite.

Exemple 8.4 hypothèse composite

Dans l'exemple précédent, l'hypothèse

$$\mathcal{H}_1 : \mathbf{x} = \mathbf{s} + \mathbf{w}$$

est une hypothèse composite (étant donné que ϕ est inconnu). ◁

Hypothèse nulle On parle en général d'hypothèse nulle si un événement intéressant n'a pas eu lieu (dans l'exemple précédent, \mathcal{H}_0 est une hypothèse nulle, puisqu'il correspond à la non détection d'une sinusoïde).

Test unilatéral

$$\begin{aligned}\mathcal{H}_0 &: \theta \leq \theta_o \\ \mathcal{H}_1 &: \theta > \theta_o\end{aligned} \quad \theta \in \mathbb{R}$$

Test biatéral

$$\begin{aligned}\mathcal{H}_0 &: \theta_o \leq \theta \leq \theta_1 \\ \mathcal{H}_1 &: \theta < \theta_o \text{ ou } \theta > \theta_1\end{aligned} \quad \theta \in \mathbb{R}$$

test de seuillage

$$\begin{aligned}\mathcal{H}_0 &: \Gamma(\mathbf{x}) < \gamma \\ \mathcal{H}_1 &: \Gamma(\mathbf{x}) > \gamma\end{aligned} \quad \text{où } \Gamma(\mathbf{x}) \text{ est une statistique scalaire}$$

8.2 Détection bayésienne

On considère un test d'hypothèses binaire, avec des hypothèses simples :

$$\begin{aligned}\mathcal{H}_0 &: \mathbf{X} \sim f_0(\mathbf{x}) = f(\mathbf{x}|\mathcal{H}_0) \\ \mathcal{H}_1 &: \mathbf{X} \sim f_1(\mathbf{x}) = f(\mathbf{x}|\mathcal{H}_1)\end{aligned}$$

On suppose qu'à chaque observation \mathbf{x} correspond exactement une des deux hypothèses. On considère ces hypothèses comme étant des variables aléatoires :

$$\begin{aligned}p_0 &:= \text{la probabilité que } \mathcal{H}_0 \text{ est vraie} \\ p_1 &:= \text{la probabilité que } \mathcal{H}_1 \text{ est vraie}\end{aligned}$$

avec $p_0 + p_1 = 1$.

8.2.1 Risque Bayésien

La question posée ici est la mesure de la performance de la règle de décision. En supposant qu'on a deux régions de décision R_0 et R_1 , telles que

$$\begin{aligned}\mathbf{x} \in R_0 &\Leftrightarrow \mathcal{H}_0 \text{ est vraie} \\ \mathbf{x} \in R_1 &\Leftrightarrow \mathcal{H}_1 \text{ est vraie}\end{aligned}$$

Il y a quatre situations possibles :

Décision	Vérité	
	\mathcal{H}_0	\mathcal{H}_1
$\mathbf{x} \in R_0$	Réjection ($P_{00} = P_R$)	détection Manquée ($P_{01} = P_M$)
$\mathbf{x} \in R_1$	Fausse alarme ($P_{10} = P_F$)	Détection ($P_{11} = P_D$)

On a $P_{00} + P_{10} = 1$ et $P_{01} + P_{11} = 1$.

A chaque décision on associe un coût : $C_{i,j}$ est le coût associé au choix de l'hypothèse \mathcal{H}_i dans le cas où c'est l'hypothèse \mathcal{H}_j qui est la bonne. Logiquement, on devrait avoir :

$$C_{i,i} < C_{i,j}, \quad i \neq j$$

On définit un coût moyen sous hypothèse \mathcal{H}_i :

$$\bar{C}_0 = E[C_{\cdot,0}] = \int C_{\cdot,0} f(\mathbf{x}|\mathcal{H}_0) d\mathbf{x} = C_{00}P_{00} + C_{10}P_{10}$$

de même on a :

$$\bar{C}_1 = E[C_{\cdot,1}] = \int C_{\cdot,1} f(\mathbf{x}|\mathcal{H}_1) d\mathbf{x} = C_{01}P_{01} + C_{11}P_{11}$$

On définit le risque bayésien moyen par :

$$p_0 \bar{C}_0 + p_1 \bar{C}_1 = \sum_{i,j=0}^1 C_{i,j} p_j P(\text{declarer } \mathcal{H}_i | \mathcal{H}_j \text{ vrai})$$

Exemple 8.5

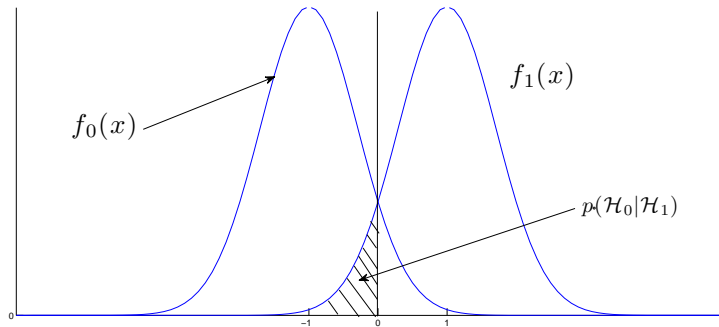
Soit une v.a. scalaire et les hypothèses :

$$\mathcal{H}_0 : \mathbf{X} \sim \mathcal{N}(-1, 1)$$

$$\mathcal{H}_1 : \mathbf{X} \sim \mathcal{N}(1, 1)$$

On choisit les régions de décisions $R_0 = (-\infty, 0]$ et $R_1 = (0, \infty)$, alors,

$$P(\mathcal{H}_0 | \mathcal{H}_1) = P(\mathbf{x} \in R_0 | \mathcal{H}_1) = \int_{-\infty}^0 f_1(\mathbf{x})$$



◁

8.2.2 Test Bayésien

Etablissement du test

Le test de Bayes correspond au test qui pour une probabilité a priori donnée minimise le risque bayésien moyen, qui peut être réécrit comme :

$$\begin{aligned} \bar{C} = & \int_{R_0} C_{00} p_0 f_0(\mathbf{x}) + C_{01} p_1 f_1(\mathbf{x}) d\mathbf{x} \\ & + \int_{R_1} C_{10} p_0 f_0(\mathbf{x}) + C_{11} p_1 f_1(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (8.1)$$

Pour minimiser cette valeur il faut choisir les régions R_i de façon à ce que la somme des deux intégrales dans (8.1) soit la plus petite possible. De plus, en se souvenant que $\{R_0, R_1\}$ forment une partition de \mathbb{R}^N , \mathbf{x} fait partie d'une et d'une seule région R_i . Il s'ensuit que l'on choisira $\mathbf{x} \in R_i$ si l'intégrand est le plus petit :

$$\begin{aligned} \mathbf{x} \in R_0 & \Leftrightarrow C_{00} p_0 f_0(\mathbf{x}) + C_{01} p_1 f_1(\mathbf{x}) < C_{10} p_0 f_0(\mathbf{x}) + C_{11} p_1 f_1(\mathbf{x}) \\ & \Leftrightarrow \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} < \frac{p_0 C_{10} - C_{00}}{p_1 C_{01} - C_{11}} \end{aligned}$$

On exprime le détecteur bayésien comme :

$$\boxed{\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{p_0 C_{10} - C_{00}}{p_1 C_{01} - C_{11}}} \quad (8.2)$$

Test du rapport de vraisemblance

Le test bayésien est un cas particulier du test du rapport de vraisemblance (LRT : Likelihood Ratio Test). Un LRT a la forme

$$\Lambda(\mathbf{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta$$

où

$$\Lambda(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$$

est un rapport de vraisemblance et $\eta > 0$ est un seuil.

Détecteur bayésien : cas discret

Si f_0 et f_1 sont des masses et si \mathbf{X} est défini dans le domaine \mathcal{X} , alors le risque bayésien moyen s'écrit :

$$\begin{aligned} \bar{C} &= \sum_{i,j} C_{ij} p_j P(\mathcal{H}_i | \mathcal{H}_j) \\ &= \sum_{\mathbf{x} \in \mathcal{X} \cap R_0} (C_{00} p_0 f_0(\mathbf{x}) + C_{01} p_1 f_1(\mathbf{x})) \\ &\quad + \sum_{\mathbf{x} \in \mathcal{X} \cap R_1} (C_{10} p_0 f_0(\mathbf{x}) + C_{11} p_1 f_1(\mathbf{x})), \end{aligned}$$

Et on obtient la même expression du détecteur bayésien (8.2).

Détecteur au minimum de probabilité d'erreur

Un cas particulier important du détecteur bayésien est celui où $C_{00} = C_{11} = 1$ et $C_{01} = C_{10} = 0$. Dans ce cas, le risque bayésien moyen est égal à $\bar{C} = p_0 P_{10} + p_1 P_{01} = P_E$. On voit donc que le détecteur au minimum de probabilité d'erreur est un cas particulier du détecteur bayésien.

Exemple 8.6 *Détection bayésienne d'une constante dans du bruit*

On considère le problème de la détection d'une constante de valeur $A > 0$ dans un bruit additif Gaussien :

$$\begin{aligned}\mathcal{H}_0 : X_i &= W_i & i = 1, \dots, N \\ \mathcal{H}_1 : X_i &= A + W_i & i = 1, \dots, N\end{aligned}$$

où $W_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ et A, σ^2 supposés connus.

Pour trouver le détecteur bayésien, on écrit l'expression du rapport de vraisemblance :

$$\begin{aligned}\Lambda(\mathbf{x}) &= \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} = \frac{\prod_{n=1}^N f_1(x_n)}{\prod_{n=1}^N f_0(x_n)} \\ &= \frac{\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n - A)^2}{2\sigma^2}}}{\prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x_n^2}{2\sigma^2}}} \\ &= \frac{e^{-\frac{1}{\sigma^2} \sum_{n=1}^N (x_n - A)^2}}{e^{-\frac{1}{\sigma^2} \sum_{n=1}^N x_n^2}} \\ &= e^{\frac{A}{\sigma^2} \sum_{n=1}^N x_n - \frac{NA^2}{2\sigma^2}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta\end{aligned}$$

En utilisant la transformation (monotone) par le logarithme :

$$\begin{aligned}& e^{\frac{A}{\sigma^2} \sum_{n=1}^N x_n - \frac{NA^2}{2\sigma^2}} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta \\ \Leftrightarrow & \frac{A}{\sigma^2} \sum_{n=1}^N x_n - \frac{NA^2}{2\sigma^2} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \ln(\eta) \\ \Leftrightarrow & \sum_{n=1}^N x_n \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{NA} \ln(\eta) + \frac{a}{2} \equiv \gamma\end{aligned}$$

Le détecteur bayésien se réduit donc à un test de seuillage de la valeur de la moyenne des observations.

D'autre part, si $p_0 = p_1 = 0.5$ et les coûts "classiques" ($C_{ij} = 1 - \delta_{ij}$), alors $\eta = 1$ et $\gamma = A/2$: on décide \mathcal{H}_1 si la moyenne est supérieur à $A/2$. On notera également que le détecteur bayésien est un détecteur, linéaire, i.e. obtenu par seuillage d'une fonction linéaire des observations. De ce fait, la frontière entre les deux régions est un hyperplan.

◁

8.2.3 Calcul de la probabilité d'erreur

Un indicateur de performance de test est la probabilité d'erreur, qui peut s'exprimer par :

$$P_E = p_0 P_{01} + p_1 P_{10} = p_0 \int_{R_0} f_1(\mathbf{x}) d\mathbf{x} + p_1 \int_{R_1} f_0(\mathbf{x}) d\mathbf{x}$$

Pour évaluer cette expression en toute généraliser, il faudrait recourir à l'évaluation d'intégrales multidimensionnelles, ce qui n'est jamais aisé. C'est pourquoi on recourt à l'utilisation de statis-

tiques simples et à des transformations monotones pour calculer la probabilité d'erreur. Dans le cas d'un courant continu dans du bruit par exemple, la statistique était $t = 1/N \sum_n x_n$. L'intérêt de ces statistiques est qu'elles sont souvent unidimensionnelles et ont des distributions connues (ou facile à déterminer).

Pour calculer la probabilité d'erreur, il est utile d'introduire la fonction $Q(\gamma)$:

DÉfinition 8.1 La fonction Q

Soit $W \sim \mathcal{N}(0, 1)$, alors, on définit :

$$Q(\gamma) \equiv P(x \geq \gamma) = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad (8.3)$$

Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $P(x \geq \gamma) = Q\left(\frac{\gamma - \mu}{\sigma}\right)$ (aisément vérifiable par changement de variables).

Q est une fonction de $\mathbb{R} \rightarrow (0, 1)$ et est monotone décroissante, et admet donc un inverse. En matlab :

$$\begin{aligned} Q(\gamma) &= 1/2 \left(1 - \operatorname{erf}\left(\frac{\gamma}{\sqrt{2}}\right)\right) \\ Q^{-1}(\gamma) &= \sqrt{2} \operatorname{erfinv}(1 - 2\alpha) \end{aligned}$$

Exemple 8.7 Détection bayésienne d'une constante dans du bruit : suite

Le test bayésien peut s'écrire :

$$\sum_{n=1}^N x_n \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{\sigma^2}{NA} \ln(\eta) + \frac{A}{2}$$

La statistique t peut s'écrire comme une v.a. T avec :

$$\begin{aligned} \mathcal{H}_0 : T &\sim \mathcal{N}\left(0, \frac{\sigma^2}{N}\right) \\ \mathcal{H}_1 : T &\sim \mathcal{N}\left(A, \frac{\sigma^2}{N}\right) \end{aligned}$$

Sous l'hypothèse \mathcal{H}_0 , $P(\mathcal{H}_1|\mathcal{H}_0) = P(T > \gamma|\mathcal{H}_0) = Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right)$; sous l'hypothèse \mathcal{H}_1 , $P(\mathcal{H}_0|\mathcal{H}_1) = P(T < \gamma|\mathcal{H}_1) = Q\left(\frac{\gamma - A}{\sigma/\sqrt{N}}\right)$. On a donc que :

$$\begin{aligned} P_E &= p_0 P(\mathcal{H}_1|\mathcal{H}_0) + p_1 P(\mathcal{H}_0|\mathcal{H}_1) \\ &= p_0 Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right) + p_1 Q\left(\frac{\gamma - A}{\sigma/\sqrt{N}}\right) \end{aligned}$$

En se souvenant que $\gamma = A/2$ pour $p_0 = p_1 = 1/2$, on obtient

$$P_E = Q\left(\frac{A\sqrt{N}}{2\sigma}\right) = Q(\sqrt{\text{SNR}}),$$

où on a exprimé le rapport signal sur bruit par $\text{SNR} = \frac{A^2 N}{\sigma^2}$. Cette relation confirme l'intuition qu'on a une probabilité d'erreur plus faible pour un rapport signal/bruit plus grand.

◁

8.3 Détecteur MAP

8.3.1 Etablissement du détecteur MAP

Pour dériver le détecteur MAP, on maximiser P_C , ce qui est équivalent à minimiser P_E :

$$\begin{aligned} P_C &= P(\mathcal{H}_0, \mathcal{H}_0) + P(\mathcal{H}_1, \mathcal{H}_1) \\ &= p_0 \int_{R_0} f_0(\mathbf{x}) d\mathbf{x} + p_1 \int_{R_1} f_1(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Toujours en tenant compte que R_i forment une partition de l'espace, on choisira :

$$\mathbf{x} \in R_i \Leftrightarrow p_i f_i(\mathbf{x}) \text{ est maximal}$$

Par la règle de Bayes, on obtien $P(\mathcal{H}_i|\mathbf{x}) = \frac{P(\mathcal{H}_i)f(\mathbf{x}|\mathcal{H}_i)}{f(\mathbf{x})} = \frac{p_i f_i(\mathbf{x})}{f(\mathbf{x})}$. La probabilité **a posteriori** $P(\mathcal{H}_i|\mathbf{x})$ est maximale quand $p_i f_i(\mathbf{x})$ est maximale (en effet, $f(\mathbf{x})$ est indépendant de i). Il s'ensuit que maximiser P_C revient à utiliser un détecteur MAP (Maximum A posteriori Probability) :

$$\mathbf{x} \in R_i \Leftrightarrow p_i f_i(\mathbf{x}) \text{ est maximal}$$

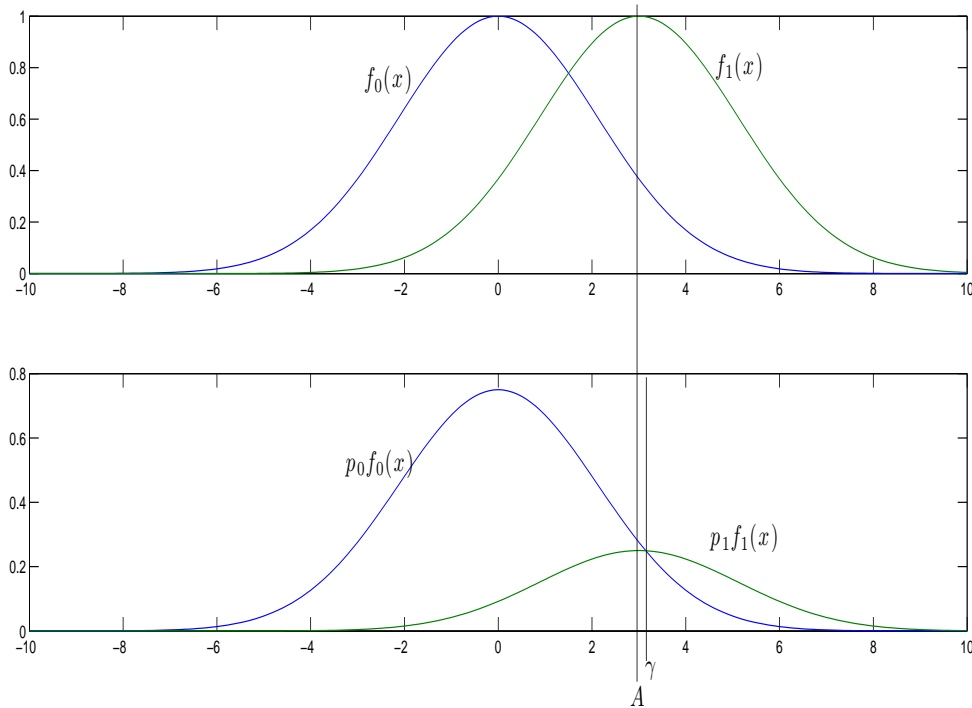


FIG. 8.2: Exemple de Test LRT appliqué au détecteur MAP

8.3.2 Cas d'hypothèses multiples

Dans le cas d'hypothèses multiples

$$\mathcal{H}_i : \mathbf{X} \sim f_i(\mathbf{x}), i = 1 \dots M,$$

$P_E = 1 - P_C = 1 - \sum_i \int_{R_i} p_i f_i(\mathbf{x}) d\mathbf{x}$, le détecteur MAP est optimal :

$$\mathbf{x} \in R_i \Leftrightarrow p_i f_i(\mathbf{x}) \text{ est maximal}$$

8.3.3 Remarques

Le détecteur MAP est un cas particulier de l'estimateur MAP où $\theta = \mathcal{H}_0$ ou $\theta = \mathcal{H}_1$, avec p_0 et p_1 déterminent la densité a priori. En général, on peut dire que le problème de la détection est

un cas particulier du problème d'estimation avec des paramètres pouvant prendre un nombre fini de valeurs ;

8.4 Test de Neyman-Pearson

Dans le détecteur Bayésien, on suppose une connaissance a priori. Cette hypothèse, acceptable dans le cas de signaux de communication par exemple (les "0" et "1" sont envoyés avec une probabilité 1/2), ne l'est pas dans beaucoup de cas de détection (détection d'un objet par un radar par exemple).

Le test de Neyman-Pearson va considérer le cas où on n'a aucune connaissance a priori.

En préalable, on remarquera que, dans un test d'hypothèse binaire, quoique l'on ai quatre probabilités $P_{ij}, i, j \in 1, 2$, on a en fait que deux degrés de liberté (car $P_D + P_M = 1$ et $P_F + P_R = 1$). De plus, P_D et P_F n'impliquent pas les probabilités a priori que les hypothèses \mathcal{H}_i soient réalisées. Il est donc raisonnable de formuler un critère de détection en fonction de P_D et P_F .

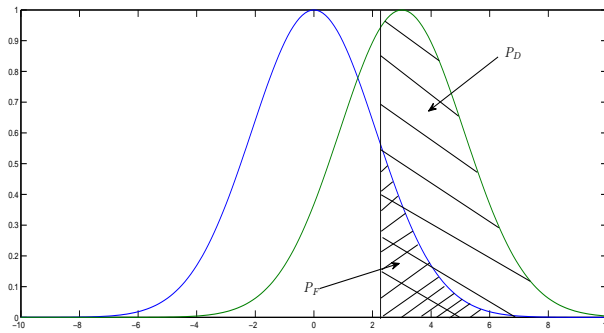


FIG. 8.3: Illustration des Probabilités de détection et de fausse alarme pour test de seuillage

En considérant la règle de seuillage $x \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\leq}} \gamma$, si γ augmente, on a une diminution de la probabilité de fausse alarme, ce qui est désirable, mais une diminution de la probabilité de détection, ce qui n'est pas désirable. Idéalement, on voudrait $P_D = 1$ et $P_F = 0$, ce qui n'est possible que si les densités f_0 et f_1 ont des supports disjoints.

Le critère d'optimisation de Neyman-Pearson est le suivant : parmi tous les détecteurs ayant une probabilité de fausse alarme valant α on choisit le détecteur qui possède la plus grande probabilité de détection :

$$\max_{s.c. P_F \leq \alpha} P_D \quad (8.4)$$

En introduisant la terminologie suivante : P_D est la puissance du test et P_F est la taille du test, on dira que ce détecteur est le plus puissant pour une taille α .

Lemme de Neyman-Pearson Soit $\alpha \in [0, 1]$, le détecteur de Neyman-Pearson est

$$\Lambda(\mathbf{x}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\leq}} \eta$$

où $\Lambda(\mathbf{x}) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ et η est choisi tel que

$$P_F = \int_{\Lambda(\mathbf{x}) > \eta} f_0(\mathbf{x}) d\mathbf{x} = \alpha$$

Remarque : On suppose qu'il est possible de trouver η tel que $P_F = \alpha$, ce qui n'est pas toujours garanti, comme dans le cas de problèmes discrets.

Preuve

Le problème d'optimisation sous contraintes (8.4) peut être résolu en utilisant un multiplicateur de Lagrange. On cherche à trouver

$$\begin{aligned} & \max_{R_1} (P_D + \lambda(P_F - \alpha)) \\ &= \max_{R_1} \int_{R_1} f_1(\mathbf{x}) d\mathbf{x} + \lambda \left(\int_{R_1} f_0(\mathbf{x}) d\mathbf{x} - \alpha \right) \\ &= -\lambda\alpha + \int_{R_1} f_1(\mathbf{x}) + \lambda f_0(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Il suffit donc, pour maximiser l'intégrale, de choisir la région R_1 telle que l'intégrand est positif et la règle est donc du type rapport de vraisemblance avec $\eta = -\lambda$:

$$\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} -\lambda$$

D'autre part, on a que $P_F = \int_{-\lambda}^{\infty} f_0(\mathbf{X}) d\mathbf{x}$, et il suffit donc de déterminer $\eta = -\lambda$ tel que $P_F = \alpha$

Exemple 8.8 Détermination d'une constante dans du bruit

On rappelle les hypothèses :

$$\begin{aligned} \mathcal{H}_0 : \mathbf{X} &\sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N) \\ \mathcal{H}_1 : \mathbf{X} &\sim \mathcal{N}(A\mathbf{1}, \sigma^2 \mathbf{I}_N) \end{aligned}$$

Pour le cas bayésien, on a déterminé que

$$\begin{aligned} \Lambda(\mathbf{x}) &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta \\ \Leftrightarrow \frac{1}{N} \sum_{n=1}^N x_n (\equiv t) &\underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma (\equiv \frac{\sigma^2}{NA} \ln(\eta) + \frac{A}{2}) \end{aligned}$$

Avec $T \sim \mathcal{N}(0, \frac{\sigma^2}{N})$ si \mathcal{H}_0
 $T \sim \mathcal{N}(A, \frac{\sigma^2}{N})$ si \mathcal{H}_1

Détermination de P_F et P_D

$$P_F = \text{Prob}(t > \gamma | \mathcal{H}_0) = Q\left(\frac{\gamma}{\sigma/\sqrt{N}}\right) \leq \alpha$$

$$P_D = \text{Prob}(t > \gamma | \mathcal{H}_1) = Q\left(\frac{\gamma - A}{\sigma/\sqrt{N}}\right)$$

Détermination du seuil Pour déterminer le seuil, on fixe $P_F = \alpha$

$$\Rightarrow \gamma = \frac{\sigma}{\sqrt{N}} Q^{-1}(\alpha)$$

$$\Rightarrow P_D = Q\left(Q^{-1}(P_F) - \underbrace{A\sqrt{N}/\sigma}_{=\sqrt{SNR}}\right)$$

◁

8.5 Caractéristique Opératoire du Récepteur

Une des figures importantes pour caractériser la qualité d'un détecteur est la caractéristique opératoire du récepteur (COR), qui est un graphique représentant la probabilité de détection en fonction de la probabilité de fausse alarme (figure 8.4).

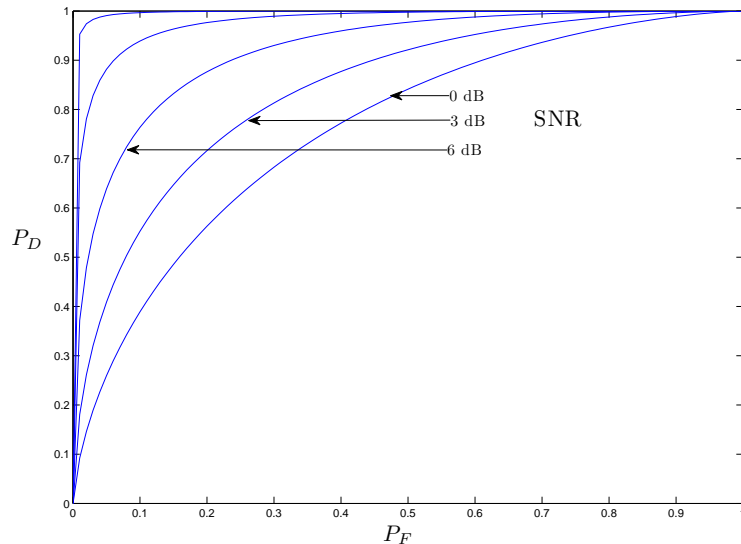


FIG. 8.4: Caractéristique Opératoire du Récepteur pour la détermination d'une constante dans du bruit avec un SNR variable

Sans les démontrer, les principales propriétés d'une COR basée sur un rapport de vraisemblance sont :

1. la courbe est concave ;
2. la courbe est située au-dessus de la droite $P_D = P_F$;
3. la pente de la courbe au point $(P_F(\eta), P_D(\eta))$ est égale à η c.à.d $\left(\frac{dP_F}{dP_D} = \eta \right)$.

8.6 Application au canal symétrique binaire

Dans cette première application, on va s'intéresser à l'application des décisions bayésiennes et de Neyman-Pearson au canal symétrique binaire. Dans un premier temps, on posera le problème et on dérivera la règle de décision "à la main", ensuite, on verra que la méthode générale bayésienne donne (heureusement) les mêmes résultats, qu'on étendra au cas d'un canal avec codage par répétition. Enfin, on appliquera la détection de Neyman-Pearson pour le canal avec codage.

8.6.1 Introduction

Un canal symétrique binaire transmet un bit (0 ou 1) correctement avec probabilité $1 - \theta$ et en erreur avec probabilité θ .

En notant x le bit transmis et y le bit reçu, et avec l'addition binaire, on peut écrire :

$$y = x + w \quad \text{où } w \sim \text{Bernouilli}(\theta).$$

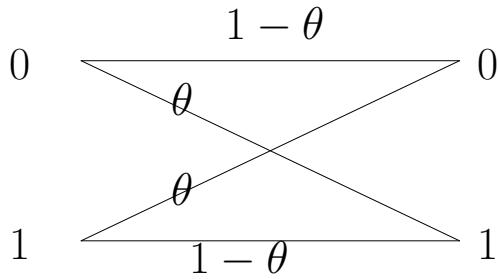


FIG. 8.5: Canal symétrique binaire : un modèle très utilisé en communications numériques

L'objectif est de concevoir un détecteur qui détermine le bit transmis en observant le bit reçu. Le test d'hypothèse correspondant peut s'écrire :

$$\begin{aligned}\mathcal{H}_0 : y &= 0 + w \\ \mathcal{H}_1 : y &= 1 + w\end{aligned}$$

ou, de manière équivalente :

$$\begin{aligned}\mathcal{H}_0 : y &\sim \text{Bernouilli}(\theta) \\ \mathcal{H}_1 : y &\sim \text{Bernouilli}(1 - \theta)\end{aligned}$$

En supposant, sans perte de généralité, que $\theta < 1/2$ et les probabilités a priori p_0 et p_1 , la question posée est de trouver l'estimée \hat{x} qui minimise la probabilité d'erreur $P(\hat{x} \neq x)$. Le problème étant extrêmement simple, on peut écrire les 4 possibilités de détection :

$$\begin{array}{cccc} 1. & 0 \rightarrow 0 & 2. & 0 \rightarrow 0 & 3. & 0 \rightarrow 1 & 4. & 0 \rightarrow 1 \\ & 1 \rightarrow 1 & & 1 \rightarrow 0 & & 1 \rightarrow 1 & & 1 \rightarrow 0 \end{array}$$

La probabilité d'erreur s'écrit $P_E = p_1 P(\mathcal{H}_0 | \mathcal{H}_1) + p_0 P(\mathcal{H}_1 | \mathcal{H}_0)$, soit, pour chaque cas :

$$\begin{aligned} 1. & : p_1 \theta + p_0 \theta = \theta \\ 2. & : p_1 \\ 3. & : p_0 \\ 4. & : p_1 (1 - \theta) + p_0 (1 - \theta) = 1 - \theta \end{aligned}$$

Comme $\theta < 1/2$, le cas 4 n'est jamais la meilleure décision, on a donc 3 cas :

$$\begin{aligned} 1. & \text{ Si } \theta < p_0, p_1 & \hat{x} &= y \\ 2. & \text{ Si } p_1 < \theta & \hat{x} &= 0 \\ 3. & \text{ Si } p_0 > \theta & \hat{x} &= 1 \end{aligned}$$

En général, les problèmes étant un peu moins simples que le simple canal symétrique binaire, il est quasi-impossible d'énumérer toutes les règles

Chapitre 9

Annexe : Variables aléatoires.

9.1 Rappels de théorie des probabilités.

9.1.1 Événements, expériences, axiomes de probabilité.

Un expérience déterminée comporte un ensemble S de tous ses résultats possibles E , appelés événements associés à cette expérience. A chaque événement associé à l'expérience considérée, on fait correspondre un nombre réel P_E appelé «probabilité de l'événement E » défini par les axiomes suivants :

1. $P_E \geq 0$
2. $P_I = 1$ un événement certain a une probabilité égale à 1
3. Si E_1 et E_2 sont mutuellement exclusifs, i.e. si $E_1 \cap E_2 = \emptyset$
 $P_{E_1 \cup E_2} = P_{E_1} + P_{E_2}$ (axiome de la somme)
4. $P_{E_1 \cap E_2} = P_{E_1} \cdot P_{E_2|E_1}$ (axiome du produit) où $P_{E_2|E_1}$ est la probabilité que E_2 survienne dans l'hypothèse où E_1 arrive.

On peut déduire de ces axiomes que P_E est toujours compris entre 0 et 1 et que la probabilité de l'événement impossible est nulle (la réciproque n'étant pas nécessairement vraie).

9.1.2 Indépendance statistique.

Deux événements E_1 et E_2 sont statistiquement indépendants si

$$P_{E_1 \cap E_2} = P_{E_1} \cdot P_{E_2} \quad (9.1)$$

$E_1 \dots E_n$ sont n événements statistiquement indépendants si

$$P_{E_i \cap E_j} = P_{E_i} \cdot P_{E_j} \quad (9.2)$$

$$P_{E_i \cap E_j \cap \dots \cap E_k} = P_{E_i} \cdot P_{E_j} \cdot \dots \cdot P_{E_k} \quad (9.3)$$

9.1.3 Lois de composition

1. $P_{\bar{E}} = 1 - P_E$ où \bar{E} est le complément de E
2. $P_{E_1 \cup E_2} = P_{E_1} + P_{E_2} - P_{E_1 \cap E_2}$
3. $P_{E_1 \cap E_2 \cap \dots \cap E_n} = P_{E_1} \cdot P_{E_2|E_1} \cdot P_{E_3|E_2 \cap E_1} \cdot \dots \cdot P_{E_n|E_{n-1} \cap \dots \cap E_2 \cap E_1}$

4. Soient E_1, \dots, E_n n événements indépendants.

$$P_{E_1 \cup E_2 \cup \dots \cup E_n} = 1 - P_{\bar{E}_1} \cdot P_{\bar{E}_2} \cdot \dots \cdot P_{\bar{E}_n}$$

5. Soient E_1, \dots, E_n n événements indépendants.

$$P_{E_1 \cup E_2 \cup \dots \cup E_n} = 1 - P_{E_1} \cdot P_{E_2} \cdot \dots \cdot P_{E_n}$$

9.1.4 Probabilités a posteriori

Soient H_1, \dots, H_n , un ensemble d'événements mutuellement exclusifs ($H_i \cap H_j = \emptyset$) tels que $H_1 \cup H_2 \cup \dots \cup H_n = I$ et non indépendants de E . Les H_i sont appelés hypothèses et peuvent être des causes de l'événement E .

$$P_E = \sum_{i=1}^n P_{H_i} \cdot P_{E|H_i} \quad (9.4)$$

Et nous avons la formule de Bayes :

$$P_{H_i|E} = \frac{P_{E \cap H_i}}{P_E} = \frac{P_{H_i} \cdot P_{E|H_i}}{P_E} \quad (9.5)$$

où

- $P_{H_i|E}$ est la probabilité a posteriori ;
- P_{H_i} est la probabilité a priori ;
- $P_{E|H_i}$ est la probabilité conditionnelle.

9.2 Variables aléatoires.

La classe S des événements associés à une expérience peut toujours être décrite à l'aide d'un ensemble d'événements mutuellement exclusifs appelés *événements élémentaires* de cette expérience.

Tout événement E consiste en la réunion d'un certain nombre d'événements élémentaires et sa probabilité est la somme des probabilités de ceux-ci.

Une **variable aléatoire** est définie par correspondance biunivoque avec un ensemble d'événements élémentaires et est caractérisée par la distribution de probabilité de celui-ci. Elle peut être à une ou plusieurs dimensions, discrète ou continue.

9.2.1 Fonction de répartition.

On considère une variable aléatoire X dont le domaine de définition sera pris systématiquement $[-\infty, \infty]$, bien que l'arrivée dans certains sous-domaines puisse être impossible. Cette variable aléatoire est entièrement définie par sa *fonction de répartition* (en anglais : *c.d.f. : cumulative distribution function*)

$$F(x) = P\{X \leq x\} \quad (9.6)$$

Cette fonction possède les propriétés suivantes :

1. $F(-\infty) = 0$; $\lim_{x \rightarrow \infty} F(x) = 1$
2. $F(b) - F(a) = P\{a < X \leq b\}$
3. $F(x)$ est une fonction monotone non décroissante.

La variable aléatoire X est dite *discrète* si la fonction $F(x)$ est une fonction en escalier, c'est-à-dire de la forme :

$$F(x) = \sum_i P_i u(x - x_i) \quad ; \quad p_i > 0 \quad ; \quad \sum_i P_i = 1 \quad (9.7)$$

où $u()$ est la fonction échelon. Une telle variable ne peut prendre que les valeurs x_i et ce, avec les probabilités p_i .

La variable X est dite *continue* si la fonction de répartition $F(x)$ est continue.

Dans le cas général, la fonction $F(x)$ peut contenir des discontinuités par saut brusque positif.

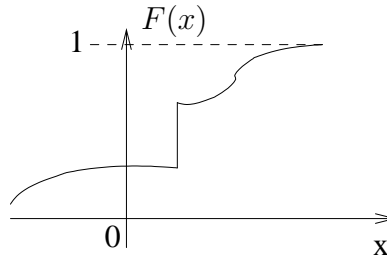


FIG. 9.1: Fonction de Répartition

9.2.2 Densité de probabilité.

Pour une fonction de répartition continue, on définit la densité de probabilité de la variable aléatoire (*p.d.f* : *probability density function*)

$$p(x) = \frac{dF}{dx} \quad (9.8)$$

On déduit aisément la signification de $p(x)$ de la propriété :

$$P\{a < X \leq b\} = F(b) - F(a) = \int_a^b p(x) dx \quad (9.9)$$

d'où $\int_{-\infty}^{\infty} p(x) dx = 1$ et, en particulier,

$$P\{x < X \leq x + dx\} = p(x) dx \quad (9.10)$$

ce qui illustre bien la dénomination de densité : plus $p(x)$ est grande, plus la probabilité que la variable tombe au voisinage de x est grande.

A condition d'admettre l'introduction de fonctions impulsions de Dirac, la notion de densité de probabilité peut être étendue aux variables aléatoires non continues, par exemple, pour une variable discrète :

$$p(x) = \sum_i P_i \delta(x - x_i) \quad (9.11)$$

9.2.3 Moments d'une variable aléatoire.

L'opérateur *espérance mathématique* $E\{f(X)\}$ fait correspondre à une fonction donnée f de la variable aléatoire X un nombre, et cela par la définition :

$$E\{f(X)\} = \int_{-\infty}^{\infty} f(x)p(x)dx \quad (9.12)$$

C'est une moyenne pondérée de la fonction f , la fonction de poids étant la densité de probabilité.

Dans le cas d'une variable aléatoire discrète, on a :

$$E\{f(x)\} = \sum_i P_i f(x_i) \quad (9.13)$$

Les moments de la variable aléatoire X sont les espérances mathématiques des puissances X^n

$$m_n = E\{X^n\} = \int_{-\infty}^{\infty} x^n p(x) dx \quad x \text{ continu} \quad (9.14)$$

$$m_n = E\{X^n\} = \sum_i x_i^n P_i \quad x \text{ discret} \quad (9.15)$$

En particulier, on distingue le moment du premier ordre m_1 qui est l'espérance mathématique de la variable elle-même :

$$m_1 = E\{X\} = \int_{-\infty}^{\infty} xp(x)dx \quad x \text{ continu} \quad (9.16)$$

$$m_1 = E\{X\} = \sum_i x_i P_i \quad x \text{ discret} \quad (9.17)$$

que l'on appelle *moyenne* ou *valeur la plus probable*. La moyenne donne la valeur de X autour de laquelle les résultats d'essais devraient se disperser. ¹Une variable aléatoire est dite *centrée* si sa moyenne est nulle. On travaille souvent avec les variables aléatoires centrées ($X - m_1$). Les moments d'ordre supérieur à 1 de la nouvelle variable, notés $\mu_n = E\{(x - m_1)^n\}$ sont souvent plus utiles que les moments m_n .

$$\text{On a } \mu_n = \sum_{k=0}^n (-1)^k C_n^k m_{n-k} m_1^k.$$

En particulier le moment centré d'ordre deux est la *variance* :

$$\mu_2 = \sigma_x^2 = E\{(x - m_1)^2\} = m_2 - m_1^2$$

La racine carrée de la variance σ_x est appelée *dispersion* ou *écart-type*. Elle donne une mesure de la dispersion vraisemblable de résultats d'essais autour de la moyenne, ainsi que le montre l'inégalité de Bienaymé-Tchebychef :

$$p\{|X - m_1| > \alpha\} < \left(\frac{\sigma_x}{\alpha}\right)^2 \quad (9.18)$$

Il peut être utile de considérer la variable centrée et réduite $\frac{X - m_1}{\sigma}$ dont la variance est l'unité.

9.2.4 Variables réelles à plusieurs dimensions.

On peut étendre les notions vues précédemment à une variable aléatoire à n dimensions $X = (X_1, X_2, \dots, X_n)$. On définit la fonction de répartition (fonction scalaire) :

¹Il ne faut pas la confondre avec la *médiane* qui est la valeur $x_{1/2}$ telle que $F(x_{1/2}) = \frac{1}{2}$ et à laquelle elle n'est pas toujours égale.

$$F(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\} \quad (9.19)$$

La densité de probabilité est alors définie par :

$$p(x_1, x_2, \dots, x_n) = \frac{\partial^n F(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_n} \quad (9.20)$$

L'extension des notions précitées est immédiate.

On exprime les lois marginales par

$$F_{X_1}(x_1) = F_{X_1 \dots X_n}(x_1, \infty, \dots, \infty) \quad (9.21)$$

et par conséquent :

$$p_{X_1}(x_1) = \frac{\partial F_{X_1}(x_1)}{\partial x_1} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p_{X_1 \dots X_n}(x_1, x_2, \dots, x_n) dx_2 \dots dx_n \quad (9.22)$$

Notons également l'introduction possible de lois conditionnelle du type :

$$\begin{aligned} & F_{X_1, \dots, X_k | X_{k+1}, \dots, X_n}(x_1, \dots, x_k | x_{k+1}, \dots, x_n) \\ &= P\{X_1 \leq x_1, \dots, X_k \leq x_k | X_{k+1} = x_{k+1}, \dots, X_n = x_n\} \end{aligned} \quad (9.23)$$

$$\begin{aligned} & p_{X_1, \dots, X_k | X_{k+1}, \dots, X_n}(x_1, \dots, x_k | x_{k+1}, \dots, x_n) \\ &= \frac{\partial^k F_{X_1, \dots, X_k | X_{k+1}, \dots, X_n}(x_1, x_2, \dots, x_n)}{\partial x_1 \partial x_2 \dots \partial x_k} \end{aligned} \quad (9.24)$$

Et on déduit la généralisation de la formule de Bayes :

$$p_{X_1, \dots, X_k | X_{k+1}, \dots, X_n} = \frac{p_{X_1, \dots, X_n}}{p_{X_1, \dots, X_k}} \quad (9.25)$$

Les moments et moyennes sont alors définis gr, ce à l'opérateur espérance mathématique :

$$E\{f(X_1, \dots, X_n)\} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) p(x_1, \dots, x_n) dx_1 \dots dx_n \quad (9.26)$$

Les moments des deux premiers ordres sont alors :

$$m_{10} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 p(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} x_1 p_{X_1}(x_1) dx_1 \quad (9.27)$$

$$m_{01} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 p(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} x_2 p_{X_2}(x_2) dx_2 \quad (9.28)$$

$$m_{20} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1^2 p(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} x_1^2 p_{X_1}(x_1) dx_1 \quad (9.29)$$

$$m_{02} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2^2 p(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{\infty} x_2^2 p_{X_2}(x_2) dx_2 \quad (9.30)$$

$$m_{11} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 p(x_1, x_2) dx_1 dx_2 \quad (9.31)$$

On considère souvent les moments centrés d'ordre 2, dont les variances :

$$\sigma_1^2 = E\{(X_1 - m_{10})^2\} = m_{20} - m_{10}^2 \quad (9.32)$$

$$\sigma_2^2 = E\{(X_2 - m_{01})^2\} = m_{02} - m_{01}^2 \quad (9.33)$$

la *corrélation mutuelle* (cross-correlation)

$$r_{12} = E\{(X_1)(X_2)\} = m_{11} \quad (9.34)$$

l'*autocorrélation*

$$r_{11} = E\{(X_1)(X_1)\} = m_{20} = \sigma_1^2 + m_{10}^2 \quad (9.35)$$

et la *covariance mutuelle*

$$\mu_{12} = E\{(X_1 - m_{10})(X_2 - m_{01})\} = m_{11} - m_{10}.m_{01} \quad (9.36)$$

On peut aisément étendre cette théorie des moments aux variables à plus de deux dimensions. En particulier, on utilise couramment les moments du premier ordre ou moyennes, permettant de centrer les variables. On forme avec les moments centrés d'ordre deux une matrice de covariance ($n \times n$) qui est symétrique et dont les éléments de la diagonale principale sont les variances :

$$C = \begin{bmatrix} \sigma_1^2 & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \sigma_2^2 & \dots & \mu_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n1} & \mu_{n2} & \dots & \sigma_n^2 \end{bmatrix} \quad (9.37)$$

9.2.5 Fonction caractéristique

La fonction caractéristique de la variable aléatoire à n dimension X est définie par :

$$\psi_x(q_1, \dots, q_n) = E\{e^{j \sum_{k=1}^n q_k X_k}\} = \int_{-\infty}^{\infty} e^{jq_1 x_1} dx_1 \int_{-\infty}^{\infty} e^{jq_2 x_2} dx_2 \dots \int_{-\infty}^{\infty} e^{jq_n x_n} p(x_1, \dots, x_n) dx_n \quad (9.38)$$

On peut aussi définir les fonctions caractéristiques marginales et conditionnelles.

Si tous les moments sont finis, on dispose du développement en série :

$$\psi_x(q_1, \dots, q_n) = 1 + \frac{(j)^k}{k!} \sum_{i_1 + \dots + i_n = k} E\{X_1^{i_1} \dots X_n^{i_n}\} q_1^{i_1} \dots q_n^{i_n} \quad (9.39)$$

En particulier, pour $n = 2$:

$$\psi_{x_1, x_2}(q_1, q_2) = 1 + j(m_{10}q_1 + m_{01}q_2) + \frac{j^2}{2!}(m_{20}q_1^2 + 2m_{11}q_1q_2 + m_{02}q_2^2) + \dots \quad (9.40)$$

et l'on a, sous réserve de différentiabilité :

$$E\{X_1^i X_2^k\} = \frac{1}{(j)^{i+k}} \left\{ \frac{\partial^{i+k} \psi_{x_1 x_2}}{\partial q_1^i \partial q_2^k} \right\}_{q_1=q_2=0} \quad (9.41)$$

9.3 Extension aux variables complexes

9.3.1 Fonction de répartition et densité de probabilité

Une variable aléatoire complexe X étant l'ensemble des nombres réels ordonnés (A,B) , sa fonction de répartition et sa densité de probabilité sont définies comme étant celles de la variable bidimensionnelle (A,B) . De même, pour une variable complexe à plusieurs dimensions, on devra considérer la variable réelle de dimension double constituée des parties réelles et imaginaires.

9.3.2 Moments

Espérance mathématique de la variable ou moyenne.

Soit $x = A + jB$ une variable aléatoire complexe à une ou à plusieurs dimensions. La moyenne de X est définie par

$$E\{X\} = m_x = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (a + jb)p_{AB}(a, b) da db \quad (9.42)$$

Comme on a, par exemple :

$$m_A = \int_{-\infty}^{\infty} a p_A(a) da \quad ; \quad \int_{-\infty}^{\infty} p(a, b) db = p(a) \quad (9.43)$$

on a aussi :

$$m_x = m_A + jm_B \quad (9.44)$$

Comme précédemment, la variable $(X - m_X)$ est dite *centrée*.

Moments du second ordre.

On considère une variable aléatoire complexe à une dimension $X = A + jB$ de moyenne m_x . Par définition, sa variance est :

$$\sigma_x^2 = E\{|X - m_x|^2\} \quad (9.45)$$

On voit que l'on a veillé à ce que la variance soit un nombre *réel et positif*.

Comme $|X - m_x|^2 = (A - m_A)^2 + (B - m_B)^2$, on a

$$\sigma_x^2 = \sigma_A^2 + \sigma_B^2 = E\{|X|^2\} - |m_X|^2 \quad (9.46)$$

Considérons maintenant une *variable complexe à n dimensions* $\mathbf{X} = \mathbf{A} + j\mathbf{B}$. Par définition, sa matrice $(n \times n)$ de covariance est

$$\mathbf{C}_{\mathbf{X}} \equiv E\{(\mathbf{X} - m_{\mathbf{X}})(\mathbf{X} - m_{\mathbf{X}})^H\} \quad (9.47)$$

où le H signifie la transposée hermitienne (conjugué complexe du transposé). Comme $(\mathbf{X} - m_{\mathbf{X}})(\mathbf{X} - m_{\mathbf{X}})^H = [\mathbf{A} - m_{\mathbf{A}} + j(\mathbf{B} - m_{\mathbf{B}})][\mathbf{A}^T - m_{\mathbf{A}}^T - j(\mathbf{B}^T - m_{\mathbf{B}}^T)]$, on a

$$\mathbf{C}_{\mathbf{X}} = \mathbf{C}_{\mathbf{A}} + \mathbf{C}_{\mathbf{B}} + j(\mathbf{C}_{\mathbf{BA}} - \mathbf{C}_{\mathbf{AB}}) \quad (9.48)$$

De la propriété $\mathbf{C}_{\mathbf{BA}} = \mathbf{C}_{\mathbf{AB}}^T$ vue pour les variables réelles résulte la propriété

$$\mathbf{C}_{\mathbf{X}} = \mathbf{C}_{\mathbf{X}}^H \quad (9.49)$$

C'est-à-dire que la *matrice de covariance est hermitienne*. De même, elle est *définie non négative*.

On adapte d'une manière analogue la définition de la matrice de covariance mutuelle de deux variables à plusieurs dimensions

$$\begin{aligned} \mathbf{C}_{\mathbf{XY}} = E\{(\mathbf{X} - m_{\mathbf{X}})(\mathbf{Y} - m_{\mathbf{Y}})^H\} & \quad \mathbf{X}(n \times 1) \\ & \quad \mathbf{Y}(m \times 1) \\ \mathbf{C}_{\mathbf{XY}} & \quad (n \times m) \end{aligned} \quad (9.50)$$

La démonstration de la propriété :

$$\mathbf{C}_{\mathbf{YX}} = \mathbf{C}_{\mathbf{XY}}^H \quad (9.51)$$

est analogue à celle de l'hermiticité de $\mathbf{C}_{\mathbf{X}}$.

9.4 Quelques lois de probabilité importantes

9.4.1 Loi à deux valeurs

Densité de probabilité.

La variable réelle à une ou plusieurs dimensions X peut prendre les valeurs a et b avec des probabilités respectives p_a et p_b telles que p_a et $p_b = 1$:

$$p(x) = p_a \delta(x - a) + p_b \delta(x - b) \quad (9.52)$$

Rappelons qu'en coordonnées cartésiennes à n dimensions

$$\sigma(\mathbf{x} - \mathbf{a}) = \delta(x_1 - a_1) \delta(x_2 - a_2) \dots \delta(x_n - a_n) \quad (9.53)$$

Moments.

$$m_1 = E\{X\} = a.p_a + b.p_b \quad (9.54)$$

Plus généralement : $m_n = E\{X^n\} = a^n p_a + b^n p_b$; (pour une dimension)

$$\sigma_x^2 = p_a p_b |b - a|^2; \text{ (pour une dimension)} \quad (9.55)$$

$$\mathbf{C}_{\mathbf{X}} = p_a p_b (\mathbf{b} - \mathbf{a})(\mathbf{b} - \mathbf{a})^H; \text{ (pour plusieurs dimensions)} \quad (9.56)$$

Fonction caractéristique

$$\psi_x(q) = p_a e^{jaq} + p_b e^{jbq}; \text{ (pour une dimension)} \quad (9.57)$$

9.4.2 Loi binomiale

Densité de probabilité.

On fait n essais indépendants relatifs à un événement E ayant une probabilité p . La variable aléatoire X "nombre d'arrivées de l'événement au cours des n essais" peut prendre les valeurs $0, 1, \dots, n$. La densité de probabilité de la variable discrète X est :

$$p(x) = \sum_{k=0}^n p_k \delta(x - k) \quad (9.58)$$

avec

$$p_k = C_n^k p^k (1-p)^{n-k} \quad (9.59)$$

Cette loi est unimodale, le mode étant le plus grand entier inférieur ou égal à $[(n+1)p]$.

Moments.

$$\begin{aligned} m_1 = E\{X\} &= np & \mu_2 = \sigma_x^2 &= np(1-p) \\ m_2 = E\{X^2\} &= Mp(1-p) + n^2p^2 & \mu_3 &= np(1-p)(1-2p) \\ & & \mu_4 &= 3n^2p^2(1-p)^2 + np(1-p)(1-6p+6p^2) \end{aligned} \quad (9.60)$$

Fonction caractéristique

$$\psi_x(q) = (1-p + pe^{jq})^n \quad (9.61)$$

Propriétés.

1. Stabilité² : si X_1 et X_2 suivent des lois binomiales de paramètres respectifs (n_1, p) et (n_2, p) – le même p – alors $X_1 + X_2$ suit une loi binomiale de paramètres $(n_1 + n_2, p)$.
2. Si $n \rightarrow \infty$, X est asymptotiquement normale de moyenne np et de variance $\sigma^2 = np(1-p)$.
3. Si $n \rightarrow \infty$ et simultanément $p \rightarrow 0$ de telle manière que $np \rightarrow \lambda$, on obtient la loi de Poisson. On peut approcher la loi de Poisson dès que $p < 0,1$ et $np > 1$.

9.4.3 Loi de Poisson

Densité de Probabilité.

La loi de Poisson :

$$p(x) = \sum_{k=0}^{\infty} p_k \delta(x - k) \quad \text{avec} \quad p_k = \frac{\lambda^k e^{-\lambda}}{k!} \quad (9.62)$$

est relative à une variable réelle à une dimension ne pouvant prendre que les valeurs entières positives (Fig. 9.2).

Interprétation :

²Une loi de probabilité est dite stable si la somme de variables indépendantes suivant une loi de ce type obéit elle aussi à une loi de ce type.

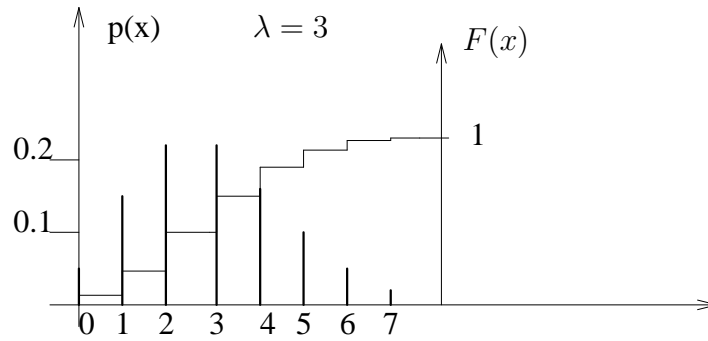


FIG. 9.2: Loi de Poisson

Exemple : un nombre $n \gg 1$ de téléviseurs de même marque et de même âge sont en service dans une ville. La probabilité que l'un quelconque de ces appareils tombe en panne demain est $p \ll 1$. La variable $X =$ nombre d'appareils qui tomberont en panne demain est une loi de Poisson de paramètre $\lambda = np$.

b) Moments

$$m_1 = E\{X\} = \lambda$$

$$\sigma_x^2 = \lambda$$

Fonction caractéristique

$$\psi_x(q) = e^{-\lambda(e^{jq}-1)} \quad (9.63)$$

Propriété de stabilité.

La somme de deux variables aléatoires obéissant à des lois de Poisson de paramètres λ_1 et λ_2 suit une loi de Poisson de paramètre $\lambda_1 + \lambda_2$.

Exemple : le nombre total de téléviseurs de tous âges et marques tombant en panne demain dans telle ville suit une loi de Poisson.

Loi de Poisson à plusieurs dimensions.

$$p(x_1, \dots, x_n) = \sum e^{-(\lambda_1 + \dots + \lambda_n)} \frac{\lambda_1^{k_1} \dots \lambda_n^{k_n}}{k_1! \dots k_n!} \delta(x_1 - k_1) \dots \delta(x_n - k_n) \quad (9.64)$$

c'est-à-dire que les composantes x_1, \dots, x_n sont indépendantes, mais suivent chacune une loi de Poisson.

9.4.4 Loi uniforme

Densité de probabilité.

X étant une variable réelle à une dimension

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{ailleurs} \end{cases} \quad (9.65)$$

c'est-à-dire que X arrive au hasard entre a et b . On dit que X est une *variable de chance*. La fonction de répartition est :

$$F(X) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases}$$

Moments.

$$\begin{aligned} m_1 = E\{X\} &= \frac{a+b}{2} & \sigma_x^2 &= \frac{(b-a)^2}{12} \\ m_n &= \frac{1}{n+1} \sum_{k=0}^n b^k a^{n-k} \end{aligned} \quad (9.66)$$

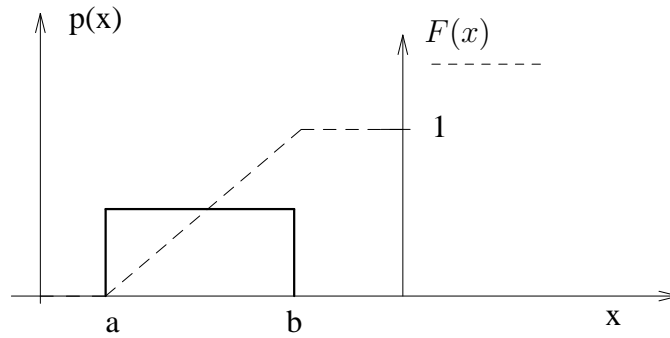


FIG. 9.3: Loi uniforme

Fonction caractéristique

$$\phi_x(q) = \frac{1}{j(b-a)q} (e^{jqb} - e^{jq a}) \quad (9.67)$$

9.4.5 Loi de Gauss ou loi normale

Variable à une dimension

Densité de probabilité. La variable aléatoire scalaire réelle X est dite gaussienne ou normale si sa densité de probabilité est de la forme

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} \quad (9.68)$$

Cette loi est unimodale et symétrique autour de m . (Fig. 9.4)

En utilisant la fonction

$$\operatorname{erf}(z) = -\operatorname{erf}(-z) = \sqrt{\frac{2}{\pi}} \int_0^z e^{-\frac{\xi^2}{2}} d\xi; z \geq 0 \quad (9.69)$$

on peut écrire la fonction de répartition sous la forme :

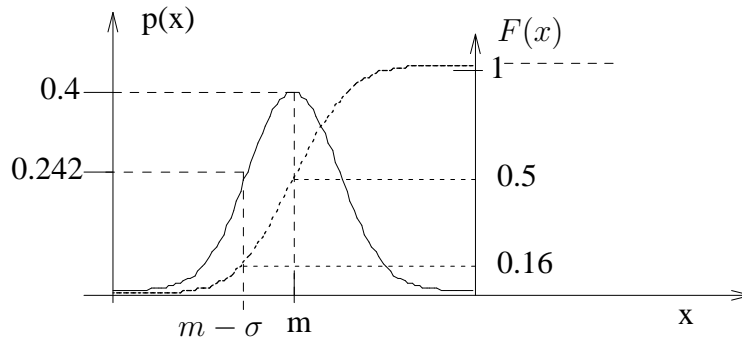


FIG. 9.4: Loi Gaussienne

$$F(x) = \frac{1}{2} \left[1 + \operatorname{erf} \frac{x - m}{\sigma \sqrt{2}} \right] \quad (9.70)$$

Moments.

$$\begin{aligned} m_1 = E\{X\} &= m \\ m_2 = E\{X^2\} &= \sigma^2 + m^2; & \mu_2 &= \sigma^2 \\ & & \mu_3 &= 0 \\ & & \mu_4 &= 3\sigma^4 \\ & & \dots & \\ & & \mu_{2n-1} &= 0 \\ \mu_{2n} &= (2n - 1)!! \sigma^{2n} \end{aligned} \quad (9.71)$$

Fonction caractéristique

$$\psi_x(q) = e^{jm q} e^{-\frac{\sigma^2 q^2}{2}} \quad (9.72)$$

Propriétés

- Stabilité : si X_1 et X_2 sont des variables gaussiennes de paramètres (m_1, σ_1) et (m_2, σ_2) , respectivement, $X_1 + X_2$ est gaussienne de paramètres $(m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.
- On a vu qu'en général, la connaissance des moments de tous les ordres est nécessaire pour définir une variable aléatoire. *Dans le cas de la loi de Gauss, les deux premiers moments suffisent.*
- Si X est normale, $Y = aX + b$ l'est aussi, a et b étant certains.

Variable à plusieurs dimensions.

Densité de probabilité. La variable aléatoire réelle vectorielle \mathbf{X} ($n \times 1$) est dite normale si

$$p(x_1, \dots, x_n) = \sqrt{\frac{\operatorname{Det} \mathbf{C}^{-1}}{(2\pi)^n}} e^{-\frac{1}{2}(\mathbf{x} - \mathbf{m})^H \mathbf{C}^{-1}(\mathbf{x} - \mathbf{m})} \quad (9.73)$$

où \mathbf{m} (vecteur $n \times 1$) est l'espérance mathématique de \mathbf{X}

Propriétés.

1. Une variable vectorielle normale est entièrement définie par sa moyenne et sa matrice de covariance.
2. Les densités de probabilité conditionnelles et marginales sont toutes gaussiennes.

NB : L'inverse n'est pas toujours vrai. Par exemple si X_1 est normale ($m = 0, \sigma = 1$) et si

$$\begin{aligned} X_2 = & X_1 \text{ avec une probabilité } 1/2 \\ & -X_1 \text{ avec une probabilité } 1/2 \end{aligned} \quad (9.74)$$

X_1 et X_2 sont normales, mais la variable bidimensionnelle (X_1, X_2) ne l'est pas.

3. La condition nécessaire et suffisante pour que les composantes X_1, \dots, X_n d'une distribution normale à n dimensions soient indépendantes est qu'elles ne soient pas corrélées (C diagonale). *L'indépendance statistique et l'indépendance en moyenne sont donc équivalentes pour la loi normale.*
4. Toute transformation linéaire sur des variables gaussiennes conserve le caractère normal. En particulier, il existe une transformation linéaire qui rend les nouvelles variables indépendantes.

Exemple : $n = 2$, variables centrées

$$\begin{aligned} m = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}; \quad \mathbf{C} = \begin{bmatrix} \sigma_1^2 & \mu \\ \mu & \sigma_2^2 \end{bmatrix}; \quad \mathbf{C}^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 - \mu^2} \begin{bmatrix} \sigma_1^2 & -\mu \\ -\mu & \sigma_2^2 \end{bmatrix}; \\ \text{Det } \mathbf{C}^{-1} = (\sigma_1^2 \sigma_2^2 - \mu^2)^{-2}; \\ p(x_1, x_2) = \frac{1}{2\pi \sqrt{\sigma_1^2 \sigma_2^2 - \mu^2}} e^{\frac{1}{2} \cdot \frac{\sigma_2^2 x_1^2 - 2\mu x_1 x_2 + \sigma_1^2 x_2^2}{\sigma_1^2 \sigma_2^2 - \mu^2}} \end{aligned} \quad (9.75)$$

9.4.6 Loi de Rayleigh

Densité de probabilité.

$$p(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} & \text{si } x \geq 0 \end{cases} \quad (9.76)$$

Fonction de répartition :

$$F(x) = 1 - e^{-\frac{x^2}{2\sigma^2}} \text{ pour } x \geq 0 \quad (9.77)$$

Moments

$$\begin{aligned} m_1 &= \sigma \sqrt{\frac{\pi}{2}} \\ &\dots \quad \mu_2 = \sigma_x^2 = 2\sigma^2 \\ m_n &= (2\sigma^2)^{\frac{n}{2}} \mathbf{C}\left(1 + \frac{n}{2}\right) \end{aligned} \quad (9.78)$$

Fonction Caractéristique

$$\psi_x(q) = 1 + \sqrt{\frac{\pi}{2}} p e^{\frac{p^2}{2}} \left[1 - \operatorname{erf} \frac{p}{\sqrt{2}} \right] \quad (9.79)$$

avec $p = \frac{jg}{\sigma}$

Interprétation

Si X_1 et X_2 sont deux variables normales centrées de variance σ^2 , la variable $X = \sqrt{X_1^2 + X_2^2}$ suit la loi de Rayleigh.

Exemple : tir sur une cible avec des dispersions égales suivant les axes vertical et horizontal ; le rayon du point d'impact suit la loi de Rayleigh. **Autre Exemple : Canal caractérisé par des chemins multiples.** Dans le cas d'un canal physique réel, les réflexions, de quelque type qu'elles soient, font emprunter plusieurs chemins au signal. L'expression du signal reçu devient alors :

$$r(t) = \sum_{i=0}^L \alpha_i(t) e^{-j2\pi(f_c \tau_i(t) + \phi_i(t))} \delta(\tau - \tau_i(t)) \quad (9.80)$$

où $\alpha_i(t)$ sont les amplitudes des chemins i et $\tau_i(t)$ les délais. Dans ce cas, on peut exprimer le signal équivalent passe-bas par :³

$$r(t) = \sum_n \alpha_n(t) e^{-j\omega_c \tau_n(t)} u[t - \tau_n(t)] \quad (9.81)$$

La réponse impulsionnelle du canal équivalent passe-bas a donc l'expression :

$$C(\tau; t) = \sum_n \alpha_n(t) e^{-j\omega_c \tau_n(t)} \delta[\tau - \tau_n(t)] \quad (9.82)$$

Dans le cas où $u(t) = 1$, le signal reçu devient :

$$r(t) = \sum_n \alpha_n(t) e^{-j\theta_n(t)} \quad (9.83)$$

On voit clairement que cette somme, en fonction des angles $\theta_n(t)$, peut engendrer des affaiblissements importants par interférences destructives.⁴ On parle dans ce cas de canal à évanouissement (*fading channel*). En général, on peut approximer $C(\tau; t)$ par un processus gaussien complexe de moyenne nulle. De par ce qui a été dit ci-dessus, on déduit immédiatement que l'enveloppe $|C(\tau; t)|$ suit une loi de Rayleigh. On parle dans ce cas de **canal de Rayleigh**. De plus, si des réflecteurs ou obstacles fixes existent, $C(\tau; t)$ n'est plus à moyenne nulle et on a un **canal de Rice**.

³En bref, le signal équivalent passe-bas correspond au signal passe-bande en ce sens que son spectre a la même allure, mais qu'il est translaté pour avoir une «fréquence centrale» nulle. On a alors, si $s(t)$ est le signal émis passe-bande, $u(t)$, le signal émis équivalent passe-bas sera relié à $s(t)$ par : $s(t) = \operatorname{Re}[u(t)e^{j\omega_c t}]$.

⁴Un cas simple est celui où on a deux chemins de même amplitude et de phase opposée

9.4.7 Loi de Rice

Densité de probabilité.

Soit $Y = X_1^2 + X_2^2$ où X_1 et X_2 sont des variables gaussiennes centrées et indépendantes de moyenne m_1 et m_2 et de variance σ^2 . On note $s^2 = m_1^2 + m_2^2$. La densité de probabilité de Y vaut :

$$p(y) = \frac{1}{2\sigma^2} e^{-\frac{s^2+y}{2\sigma^2}} \cdot I_0\left(\sqrt{y} \frac{s}{\sigma^2}\right); \quad y \geq 0 \quad (9.84)$$

En définissant $R = \sqrt{Y}$, on obtient :

$$p(r) = \frac{r}{\sigma^2} e^{-\frac{s^2+r^2}{2\sigma^2}} \cdot I_0\left(\frac{rs}{\sigma^2}\right); \quad r \geq 0 \quad (9.85)$$

Qui est la densité de probabilité d'une variable aléatoire Ricienne.

Fonction de répartition :

$$F(r) = 1 - e^{-\left(\frac{s^2}{\sigma^2} + \frac{r^2}{\sigma^2}\right)/2} \sum_{k=0}^{\infty} \left(\frac{s}{r}\right)^k I_k\left(\frac{rs}{\sigma^2}\right) \quad (9.86)$$

où $I_k()$ sont les fonctions de Bessel d'ordre k .

Chapitre 10

Annexe : Fonctions aléatoires

10.1 Notion de fonction aléatoire.

Le calcul des probabilités traite de variables aléatoires qui ne dépendent pas, du moins explicitement, du temps ou d'autres paramètres tels que les coordonnées spatiales ; on traite de variables et non de fonctions.

On peut introduire la notion de fonction aléatoire comme étant une fonction du temps et d'un certain nombre de variables aléatoires, e.g.

$$X(t) = A \sin(\omega t + \phi) \quad (10.1)$$

où A , ω et ϕ sont des variables aléatoires. En multipliant le nombre de paramètres aléatoires, on peut arriver à définir de cette manière des fonctions aléatoires très générales, par exemple :

$$X(t) = \sum_{k=1}^n A_k \sin(\omega_k t + \phi_k) \quad (10.2)$$

Si l'on considère l'ensemble des fonctions définies de cette manière, on appelle une **réalisation** de la fonction aléatoire $X(t)$, une fonction $X_r(t)$ où une épreuve a été faite sur les paramètres.

Une autre manière de définir la notion de fonction aléatoire est de dire qu'elle représente un processus où le hasard peut intervenir à tout moment. On arrive alors naturellement à la définition suivante :

Une fonction aléatoire (réelle ou complexe, à une ou plusieurs dimensions) de l'argument t est, pour toute valeur de t , une variable aléatoire (réelle ou complexe, à une ou plusieurs dimensions)

L'argument t sera considéré comme une variable réelle à une dimension, généralement le temps. On peut étendre l'étude à des arguments à plusieurs dimensions, par exemple les coordonnées spatiales.

Il est évident que la théorie à établir ne devra pas seulement décrire la variable aléatoire $X(t_1)$, mais aussi les interdépendances qui existent entre les variables aléatoires $X(t_1), X(t_2), \dots$ pour divers instants.

10.2 Fonctions de répartition et densités de probabilité

Soit une fonction aléatoire scalaire $X(t)$, on peut caractériser la variable aléatoire $X_{t_1} = X(t_1)$ par sa fonction de répartition ou sa densité de probabilité $p_{X_{t_1}}(x_1, t_1)$, fonction des deux

variables x_1 et t_1 . Mais ce n'est pas suffisant pour déterminer la fonction aléatoire, car on ne sait rien de l'interdépendance des valeurs prises à des instants différents.

Si l'on considère n valeurs de l'argument t_1, t_2, \dots, t_n , on obtient la variable aléatoire à n dimensions $[X_{t_1}, \dots, X_{t_n}]$ que l'on doit caractériser par la densité de probabilité

$$p_{X_{t_1} \dots X_{t_n}}(x_1, t_1; x_2, t_2; \dots; x_n, t_n) \quad (10.3)$$

appelée *densité de probabilité du n^{eme} ordre*.

La connaissance de la densité de probabilité du n^{eme} ordre fournit automatiquement celle des densités d'ordre inférieur à n , car ce sont des densités marginales pouvant se calculer par la formule :

$$\begin{aligned} & p_{X_{t_1} \dots X_{t_k}}(x_1, t_1; x_2, t_2; \dots; x_n, t_k) \\ = & \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} p_{X_{t_1} \dots X_{t_n}}(x_1, t_1; x_2, t_2; \dots; x_n, t_n) dx_{k+1} \dots dx_n \end{aligned} \quad (10.4)$$

On peut aussi faire intervenir les densités de probabilité conditionnelle et écrire :

$$\begin{aligned} & p_{X_{t_1} \dots X_{t_n}}(x_1, t_1; x_2, t_2; \dots; x_n, t_n) \\ = & p_{X_{t_1}}(x_1, t_1) p_{X_{t_2}}(x_2, t_2 | x_1, t_1) \dots p_{X_{t_n}}(x_n, t_n | x_1, t_1; x_2, t_2; \dots; x_{n-1}, t_{n-1}) \end{aligned} \quad (10.5)$$

On admettra qu'une fonction aléatoire est complètement définie par sa densité de probabilité d'ordre $n \rightarrow \infty$. Il suffit donc, en principe de procéder aux extensions de la théorie des variables aléatoires pour un nombre de variables infini.

Ces notions peuvent être extrapolées sans difficulté aux fonctions aléatoires multidimensionnelles ou complexes.

Il est certain que l'on ne connaÓtra que très rarement les densités de probabilité de tous ordres, mais on verra que de nombreux problèmes relatifs à la transmission de l'énergie peuvent être résolus si l'on connaÓt la densité de probabilité d'ordre 2. Cette connaissance est d'ailleurs suffisante pour caractériser complètement les fonctions aléatoires gaussiennes.

10.3 Classification des fonctions aléatoires selon leurs propriétés statistiques.

10.3.1 Fonction aléatoire à valeurs indépendantes.

La fonction aléatoire $X(t)$ est dite à valeurs indépendantes si, pour tout ensemble de valeurs (t_1, \dots, t_n) , n quelconque, différentes de l'argument, les variables aléatoires X_{t_1}, \dots, X_{t_n} sont indépendantes. On a alors :

$$p_{X_{t_1} \dots X_{t_n}}(x_1, t_1; x_2, t_2; \dots; x_n, t_n) = p_{X_{t_1}}(x_1, t_1) p_{X_{t_2}}(x_2, t_2) \dots p_{X_{t_n}}(x_n, t_n) \quad (10.6)$$

Une telle fonction aléatoire est entièrement définie par sa première densité de probabilité. A tout instant, le futur est indépendant du présent et du passé, car on peut écrire :

$$p_{X_{t_n}}(x_n, t_n | x_1, t_1; x_2, t_2; \dots; x_{n-1}, t_{n-1}) = p_{X_{t_n}}(x_n, t_n) \quad (10.7)$$

10.3.2 Fonction aléatoire à valeurs non corrélées ou orthogonales.

La fonction aléatoire $X(t)$ est dite à valeurs non corrélées ou orthogonales si, pour tout couple de valeurs différentes de l'argument (t_1, t_2) , les variables aléatoires X_{t_1} et X_{t_2} sont non corrélées (orthogonales), c'est-à-dire :

$$\text{non corrélation} \quad \text{Cov}[X_{t_1}, X_{t_2}] = 0 \quad (10.8)$$

$$\text{orthogonalité} \quad E\{X_{t_1}\} = E\{X_{t_2}\} = 0 \text{ et } \text{Cov}[X_{t_1}, X_{t_2}] = 0 \quad (10.9)$$

La notion de non-corrélation est un affaiblissement de celle d'indépendance et est surtout utile dans la théorie du second ordre ¹. Les deux notions se confondent pour des fonctions aléatoires normales.

10.3.3 Fonction aléatoire additive.

C'est une fonction aléatoire à accroissements indépendants : pour tout ensemble de n couples de valeurs différentes (t'_k, t''_k) , $k = 1, \dots, n$ (n quelconque), les accroissements $\Delta_k = X_{t''_k} - X_{t'_k}$ sont des variables aléatoires indépendantes.

10.3.4 Fonction aléatoire gaussienne

La fonction aléatoire $X(t)$ est normale ou gaussienne si, pour tout ensemble de valeurs (t_1, \dots, t_n) de l'argument (n quelconque), la variable aléatoire vectorielle à n dimensions $\mathbf{X} = [X_{t_1} \dots X_{t_n}]^T$ est normale, c'est-à-dire :

$$p_{X_{t_1} \dots X_{t_n}}(x_1, t_1; x_2, t_2; \dots; x_n, t_n) = \sqrt{\frac{\text{Det} \mathbf{C}^{-1}}{(2\pi)^n}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^H \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})} \quad (10.10)$$

où

$$\mathbf{m} = E\{\mathbf{X}\} \quad (10.11)$$

10.4 Moments d'une fonction aléatoire

10.4.1 Moyenne ou espérance mathématique

Soit une fonction aléatoire réelle ou complexe, scalaire ou vectorielle $X(t)$. La moyenne ou espérance mathématique de $X(t)$ est une fonction certaine réelle ou complexe, scalaire ou vectorielle de même dimension que $X(t)$, égale à tout instant t à la moyenne de la variable aléatoire $X(t)$.

On note :

$$\boxed{m_x(t) \text{ ou } E\{X(t)\}} \quad (10.12)$$

Il s'agit d'une **moyenne d'ensemble** définie par

$$m_x(t) = \int_{-\infty}^{\infty} x p_X(x, t) dx \quad (X \text{ scalaire réelle}) \quad (10.13)$$

¹i.e. si on étudie uniquement les densités de probabilité d'ordre un et deux

où le temps n'intervient que comme variable muette. Par exemple, si l'on étudie la tension de bruit à la sortie d'un amplificateur, on peut s'imaginer que l'on dispose de n amplificateurs macroscopiquement identiques et que l'on fasse à l'instant t la moyenne des tensions de sortie des amplificateurs. Pour $n \rightarrow \infty$, cette moyenne tend vers $m_x(t)$, en probabilité.

10.4.2 Variance. Covariance.

Covariance d'une fonction aléatoire.

Soit une fonction aléatoire réelle ou complexe, scalaire ou vectorielle $X(t)$ de dimension $n \times 1$. La matrice de covariance de $X(t)$ est une fonction certaine réelle ou complexe, scalaire ou matricielle (de dimension $n \times n$) fonction de t et t' égale pour tout couple (t, t') à la covariance des variables aléatoires $X(t)$ et $X(t')$.

$$C_x(t, t') = E\{[X(t) - m_x(t)][X(t') - m_x(t')]^H\} \quad (10.14)$$

Variance d'une fonction aléatoire.

Il s'agit de la covariance pour $t = t'$

$$\sigma_x^2(t) = C_x(t, t) \quad (10.15)$$

Covariance mutuelle de deux fonctions aléatoires.

Soient deux fonctions aléatoires $X(t)$ et $Y(t)$ de dimensions $(n \times 1)$ et $(m \times 1)$, respectivement. Leur matrice de covariance mutuelle est, par définition, la matrice $(n \times m)$ fonction de t et t'

$$C_{XY}(t, t') = E\{[X(t) - m_x(t)][Y(t') - m_y(t')]^H\} \quad (10.16)$$

10.4.3 Propriétés des variances et covariances

Fonctions aléatoires scalaires, réelles ou complexes.

$$\sigma_x^2(t) = C_X(t, t) \geq 0 \quad (10.17)$$

$$C_X(t, t') = C_X^*(t', t) \quad (10.18)$$

$$|C_X(t, t')| \leq \sqrt{\sigma_x^2(t)\sigma_x^2(t')} \quad (10.19)$$

Ceci est une conséquence immédiate des propriétés d'hermiticité et de non négative définition de la matrice de covariance d'une variable aléatoire vectorielle appliquées à la variable à deux dimensions $\begin{bmatrix} X(t) \\ X(t') \end{bmatrix}$.

La fonction de covariance est définie non négative, c'est-à-dire que, pour toute fonction continue $\xi(t)$, on a

$$\int_{\tau} \int_{\tau} \xi^*(t) C_x(t, t') \xi(t') dt dt' \geq 0 \quad (\text{réel !}) \quad (10.20)$$

Fonctions aléatoires vectorielles, réelles ou complexes.

On peut assez aisément étendre aux fonctions aléatoires vectorielle les propriétés qui viennent d'être démontrées pour les fonctions scalaires : les variances, covariances et covariances mutuelles sont des matrices telles que :

$$\mathbf{C}_{\mathbf{X}}(t, t') = \mathbf{C}_{\mathbf{X}}(t', t)^H \quad ; \quad \sigma_{\mathbf{x}}^2(t) = \sigma_{\mathbf{x}}^{2H}(t) \quad (10.21)$$

$$|[\mathbf{C}_{\mathbf{X}}(t, t')]_{ij}| \leq \sqrt{[\mathbf{C}_{\mathbf{X}}(t, t)]_{ii} [\mathbf{C}_{\mathbf{X}}(t', t')]_{jj}} \quad (10.22)$$

$$\mathbf{C}_{\mathbf{X}\mathbf{Y}}(t, t') = \mathbf{C}_{\mathbf{Y}\mathbf{X}}^H(t', t) \quad (10.23)$$

Si $\xi(t)$ est un vecteur à n dimensions :

$$\int_{\tau} \int_{\tau} \xi^H(t) \mathbf{C}_{\mathbf{X}}(t, t') \xi(t') dt dt' \geq 0 \quad (10.24)$$

10.4.4 Fonctions aléatoires non corrélées ou orthogonales.

Pour une fonction vectorielle, la non corrélation revient à la nullité de la matrice de covariance ($\mathbf{C}_{\mathbf{X}}(t, t')$) pour $t \neq t'$. On aura non corrélation de deux fonctions aléatoires $\mathbf{X}(t)$ et $\mathbf{Y}(t)$ si la matrice de covariance mutuelle $\mathbf{C}_{\mathbf{X}\mathbf{Y}}(t, t')$ est identiquement nulle, même pour $t = t'$ et orthogonalité si, en outre, $\mathbf{m}_{\mathbf{X}} = \mathbf{m}_{\mathbf{Y}} = \mathbf{0}$.

10.5 Stationnarité et Ergodisme

10.5.1 Stationnarité.

Il est courant, dans la pratique, que les propriétés statistiques d'une fonction aléatoire ne semblent pas évoluer au cours du temps ; c'est souvent le cas pour une tension de bruit observée à l'oscilloscope. Mathématiquement, on exprime cette permanence en introduisant les concepts de stationnarité. Cette notion est cependant réservée aux fonctions aléatoires s'étendant du domaine $-\infty < t < \infty$, de sorte que, rigoureusement, de telles fonctions n'existent pas. Cependant, si les caractéristiques statistiques d'une fonction aléatoire ne se modifient pas sur la durée d'observation, on peut toujours imaginer qu'il en est de même sur tout le passé et le futur.

Stationnarité stricte.

Une fonction aléatoire est dite strictement stationnaire si toutes ses densités de probabilité ne dépendent pas du choix de l'origine des temps, c'est-à-dire ne changent pas lorsqu'on remplace t par $t + t_0$, t_0 quelconque.

Ceci implique que la première densité de probabilité est indépendante de t et que la $n^{\text{ème}}$ densité de probabilité (n quelconque) ne dépend des t_i que par les seules différences $t_2 - t_1, t_3 - t_1, \dots, t_n - t_1$ (en prenant t_1 comme référence).

Cette notion n'étant guère utilisable en pratique, on introduit la stationnarité faible.

Stationnarité faible (au sens large).

Définition.

- La fonction aléatoire réelle ou complexe, scalaire ou vectorielle $X(t)$ est dite faiblement stationnaire si son espérance mathématique est indépendante de t et si sa covariance ne dépend de t et t' que par la différence $t - t'$.

$$m_x(t) = m \quad \text{constante} \quad (10.25)$$

$$\text{Cov}[X(t)] = C_X(\tau) \quad \text{avec } \tau = t - t' \quad (10.26)$$

- Deux fonctions aléatoires $X(t)$ et $Y(t)$ sont dites mutuellement stationnaires au sens faible si elles sont chacune faiblement stationnaires et si en outre leur covariance mutuelle ne dépend que de $\tau = t - t'$.

Propriétés des variances et covariances. Les propriétés générales des variances et covariances se transposent de la manière suivante pour les fonctions aléatoires faiblement stationnaires :

1. $\sigma_x^2 = C_X(0)$ est une constante (scalaire ou matrice).
2. Hermiticité : $C_X(\tau) = C_X(-\tau)^H$; $C_{XY}(\tau) = C_{YX}(-\tau)^H$
3. Positive-définition

- Pour une fonction scalaire :

$$\begin{aligned} \sigma_x^2 = C_X(0) &\geq 0 \quad (\text{nécessairement réel}) \\ |C_X(\tau)| &\leq C_X(0) \end{aligned}$$

Pour une fonction vectorielle, outre cette propriété valable pour chaque composante, i.e. pour les éléments diagonaux de la matrice de covariance comparés à ceux de la matrice de variance, on a

$$|[C_X(\tau)]_{ij}| \leq \sqrt{[\sigma_x^2]_{ii} [\sigma_x^2]_{jj}} \quad (10.27)$$

- La matrice $\mathfrak{S}_X(\omega) = \mathcal{F}\{C_X(\tau)\}$, dont les éléments sont les transformées de Fourier de ceux de la matrice de covariance, est définie non négative. En particulier, pour une fonction scalaire réelle ou complexe : $\mathfrak{S}_X(\omega) \geq 0$ et pour deux fonctions scalaires réelles ou complexes

$$|S_{XY}(\omega)| = |S_{YX}^*(\omega)| \leq \sqrt{S_X(\omega) S_Y(\omega)}$$

Cas des fonctions aléatoires périodiques. Une fonction aléatoire $X(t)$ est dite périodique, de période T , si son espérance mathématique est périodique et si sa fonction de covariance $C_x(t, t')$ est périodique en t et en t' .

Il en est ainsi si toutes les réalisations possibles de la fonction le sont. Inversement, si $m_x(t)$ et $C_X(t, t')$ sont périodiques, $X(t)$ et $X(t - T)$ sont égales avec une probabilité unité.

Bien sur, une fonction aléatoire périodique n'est pas nécessairement stationnaire au sens faible, mais pour la rendre telle, il suffit de la décaler sur l'axe t d'une quantité t_0 qui est une variable de chance sur le domaine $[0, T]$, c'est-à-dire de rendre l'origine de la période purement aléatoire.

10.5.2 Ergodisme.

Dans la pratique, on dispose souvent d'une seule réalisation de la fonction aléatoire à l'étude. On imagine difficilement devoir construire un millier d'amplificateurs pour déterminer les caractéristiques statistiques de ceux-ci. Dès lors, on est confronté à la question suivante : dans quelle

mesure peut-on déduire de la réalisation dont on dispose (la seule et unique) certaines caractéristiques statistiques de la fonction aléatoire, par exemple des moments qui sont des moyennes faites sur l'ensemble des réalisations ? La théorie de l'ergodisme tente de répondre à cette question, principalement pour des fonctions aléatoires.

En bref, cette théorie essaiera de rapprocher les moyennes d'ensemble (moyenne des variables aléatoires X_{t_i} des moyennes temporelles (notées $\langle X(t_0) \rangle$)

Typiquement, on considère une variable aléatoire définie à partir d'une fonction aléatoire, généralement une fonction des valeurs prises par la fonction aléatoire à divers instants :

$$\eta = f[X_{t_1}, \dots, X_{t_n}] \quad (10.28)$$

Par exemple X_{t_1} ou le produit $X_{t_1} X_{t_2}$. En faisant l'hypothèse que $X(t)$ est stationnaire, on peut espérer qu'en faisant "glisser" sur l'axe des temps l'échantillon $X_r(t)$ disponible, i.e. en considérant $X_r(t + t_0)$, et ce pour tout t_0 , on obtienne un ensemble de fonctions présentant les mêmes propriétés que l'ensemble des réalisations de la fonction aléatoire $X(t)$.

Nous obtenons alors, pour η :

$$\eta(t_0) = f[X_{t_0+t_1}, \dots, X_{t_0+t_n}] \quad (10.29)$$

et on peut calculer des moyennes de $\eta(t_0)$ sur l'ensemble des fonctions glissées :

$$\langle \eta \rangle \equiv \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \eta(t_0) dt_0 \quad (10.30)$$

dont on peut espérer qu'elle tende vers la moyenne d'ensemble $E\{f\} = E\{\eta\}$.

Evidemment, η étant une variable aléatoire, $\eta(t_0)$ est une fonction aléatoire ; pour obtenir la moyenne temporelle $\langle \eta \rangle$, il faut intégrer et prendre un passage à la limite.

Comme nous étudions surtout la théorie du second ordre, nous nous intéresserons uniquement à la possibilité de déterminer $m_x(t)$ et $C_X(t, t')$ dans le cadre cohérent d'hypothèses :

– **convergence en moyenne quadratique :**

- On dit que la suite de fonctions aléatoires $X_n(t)$ définies sur $t \in \tau$ converge en moyenne quadratique pour $n \rightarrow \infty$ vers la fonction aléatoire $X(t)$ définie sur τ , et l'on écrit :

$$l.i.m. \lim_{n \rightarrow \infty} X_n(t) = X(t) \quad (10.31)$$

si, pour tout $t \in \tau$, la variable aléatoire $X_n(t)$ converge en moyenne vers $X(t)$, i.e. si :

$$\lim_{n \rightarrow \infty} E\{|X_n(t) - X(t)|^2\} = 0 \quad (10.32)$$

- On dit que la fonction aléatoire $X(t)$ converge en moyenne quadratique pour $t \rightarrow t_0$ vers la variable aléatoire X_0 , et l'on écrit :

$$l.i.m. \lim_{n \rightarrow \infty} X(t) = X_0 \quad (10.33)$$

si

$$\lim_{t \rightarrow t_0} E\{|X(t) - X_0|^2\} = 0 \quad (10.34)$$

– **intégrale en moyenne quadratique**

- **Définition** : L'intégrale en moyenne quadratique de la fonction aléatoire $X(t)$ est, quand elle existe, la variable aléatoire :

$$\int_a^b X(t) dt = \underset{n \rightarrow \infty}{l.i.m.} \sum_{k=1}^n X(t'_{k,n})(t_{k,n} - t_{k-1,n}) \quad (10.35)$$

avec

$$a = t_{0,n} < t_{1,n} < \dots < t_{n,n} = b \text{ partition de (a,b)}$$

$$(t_{k,n} - t_{k-1,n}) \rightarrow 0 \text{ quand } n \rightarrow \infty$$

$$t'_{k,n} \text{ point quelconque de } (t_{k-1,n}, t_{k,n})$$

- **Théorème** : La condition nécessaire et suffisante d'existence de $\int_a^b X(t) dt$ est l'existence des intégrales certaines :

$$\int_a^b m_x(t) dt \quad ; \quad \int_a^b \int_a^b C_X(t, t') dt dt' \quad (10.36)$$

- stationnarité faible.

Moyennes temporelles.

Moyenne temporelle d'une fonction aléatoire $X(t)$ C'est, quand elle existe,

$$\langle X(t_0) \rangle = \underset{T \rightarrow \infty}{l.i.m.} \left[Y(T, t_0) \equiv \frac{1}{T} \int_{t_0}^{t_0+T} X(t) dt \right] \quad (10.37)$$

On notera qu'en effectuant un changement de variable :

$$Y(T, t_0) = \frac{1}{T} \int_0^T X(t + t_0) dt \quad (10.38)$$

$Y(T, t_0)$ est une variable aléatoire dont la limite $\langle X(t_0) \rangle$ est en général aléatoire et dépendante de t_0 , ayant les mêmes dimensions que $X(t)$.

Valeur quadratique moyenne (temporelle) d'une fonction aléatoire $X(t)$. C'est, quand elle existe :

$$\langle |X(t_0)|^2 \rangle = \underset{T \rightarrow \infty}{l.i.m.} \left[Y(T, t_0) \equiv \frac{1}{T} \int_{t_0}^{t_0+T} |X(t)|^2 dt \right] \quad (10.39)$$

C'est une variable aléatoire scalaire positive, dépendant de t_0 . Cette notion est surtout utilisée pour une fonction aléatoire scalaire, auquel cas elle prend souvent la signification d'énergie moyenne de la réalisation $X(t)$, à un facteur constant près.

Fonction de corrélation d'une fonction aléatoire $X(t) < X(t_0) >$ étant supposé existant, c'est, quand elle existe :

$$R_X(\tau, t_0) = \underset{T \rightarrow \infty}{l.i.m.} \left[Y(T, t_0) \equiv \frac{1}{T} \int_{t_0}^{t_0+T} [X(t+\tau) - \langle X(t_0+\tau) \rangle][X(t) - \langle X(t_0) \rangle]^H dt \right] \quad (10.40)$$

C'est en général une matrice aléatoire ($n \times n$) si $X(t)$ est un vecteur ($n \times 1$), fonction aléatoire de τ et de t_0 .

Fonction de corrélation mutuelle de deux fonctions aléatoires $X(t)$ et $Y(t)$ On suppose $\langle X(t_0) \rangle$ et $\langle Y(t_0) \rangle$ existants. C'est, quand elle existe :

$$R_{XY}(\tau, t_0) = \underset{T \rightarrow \infty}{l.i.m.} \left[Y(T, t_0) \equiv \frac{1}{T} \int_{t_0}^{t_0+T} [X(t+\tau) - \langle X(t_0+\tau) \rangle][Y(t) - \langle Y(t_0) \rangle]^H dt \right] \quad (10.41)$$

C'est en général une matrice aléatoire ($n \times m$).

Ergodisme au sens large d'une fonction aléatoire.

Définition : une fonction aléatoire (non nécessairement stationnaire) est ergodique au sens large si $\langle X(t_0) \rangle$ existe et est une variable aléatoire indépendante de t_0 .

Théorème : si $X(t)$ est faiblement stationnaire et intégrable en moyenne quadratique sur tout intervalle fini, elle est ergodique au sens large.

Ergodisme au sens strict d'une fonction aléatoire.

Définition : une fonction aléatoire (non nécessairement stationnaire) est ergodique au sens strict si

1. $\langle X(t_0) \rangle$ existe et est un nombre certain indépendant de t_0 . (on le note $\langle X \rangle$);
2. $\langle X \rangle = \lim_{t \rightarrow \infty} m_x(t)$

Cette limite étant supposée existante. Ce n'est pas toujours le cas, même lorsque $m_x(t)$ est fini. Par exemple $X(t) = A \sin(\omega t)$ où A est aléatoire, a une espérance mathématique $m_x(t) = m_A \sin(\omega t)$.

Ergodisme relativement à la fonction de corrélation.

Définition : une fonction aléatoire scalaire faiblement stationnaire $X(t)$ est ergodique relativement à sa fonction de corrélation si :

1. elle est strictement ergodique ;
2. sa fonction de corrélation $R_X(\tau, t_0)$ existe, est certaine et indépendante de t_0 . On la note $R_X(\tau)$;
3. $R_X(\tau) = C_X(\tau)$. La fonction de corrélation est égale à la fonction de covariance

Conclusions.

Dans les applications, on suppose fréquemment qu'il y a ergodisme relativement à la fonction de corrélation, bien que souvent on ne puisse le démontrer du fait d'une connaissance insuffisante du modèle mathématique de la fonction aléatoire stationnaire. Cette *hypothèse ergodique* revient à admettre l'égalité des moyennes d'ensemble et des moyennes temporelles jusqu'au deuxième ordre, ces dernières étant calculées sur la réalisation dont on dispose.

$$\begin{aligned} m_x &= \langle X \rangle \\ C_X(\tau) &= R_X(\tau) \\ \sigma_x^2 &= C_X(0) = R_X(0) = \langle |X - m_x|^2 \rangle \end{aligned}$$

En particulier, l'égalité des fonctions de corrélation (facile à estimer) et de covariance (plus difficile) est d'une importance capitale pour la détermination de la bande passante (transformée de Fourier de la covariance) et des propriétés d'indépendance et d'orthogonalité des signaux.

Bibliographie

- [Kay93a] Steven Kay. *Fundamentals of statistical Signal Processing : Detection Theory*. Prentice Hall, 1993.
- [Kay93b] Steven Kay. *Fundamentals of statistical Signal Processing : Estimation Theory*. Prentice Hall, 1993.
- [Sch91] Louis L. Scharf. “*Statistical Signal Processing, Detection, Estimation, and Time Series Analysis*”. Addison-Wesley, 1991.