Université de Rennes 1
DEA STIR
Signal — Télécommunications — Images — Radar
Option Signal

Introduction au Filtrage en Temps Discret

Filtre de Kalman
Modèles de Markov Cachés

François LeGland
IRISA / INRIA

2000–01

Table des matières

Chapitre 1

Introduction

Le filtrage consiste à estimer l'état d'un système dynamique, c'est–à–dire évoluant au cours du temps, à partir d'observations partielles, généralement bruitées.

Typiquement, on dispose d'une suite $Y_1,Y_2,\cdots,Y_n$ d'observations, obtenues après traitement préalable du signal brut recueilli au niveau des capteurs. Chaque observation Y_n est reliée à l'état inconnu X_n par une relation du type

$$Y_n = h(X_n) + V_n \ ,$$

où V_n est un *bruit*, qui modélise l'erreur d'observation.

Pour aller plus loin, il est nécessaire de définir plus précisément la notion de *bruit*. On trouvera à l'Annexe A les rappels de probabilités dont on aura besoin dans ce cours.

Tel qu'il est formulé, le problème de l'estimation de l'état inconnu X_n à partir des observations $Y_1,Y_2,\cdots,Y_n$ est en général mal–posé. Pour s'en convaincre, considérons le cas très simple où il n'y a pas de dynamique dans l'évolution de l'état du système, c'est–à–dire que $X_n \equiv x$, pour tout $n = 1,2,\cdots$, et $x \in \mathbf{R}^m$ est un paramètre inconnu. On désigne par x^0 la vraie valeur du paramètre. Pour simplifier encore la discussion, on suppose que les observations d–dimensionnelles $Y_1,Y_2,\cdots,Y_n$ dépendent linéairement du paramètre. On a donc

$$Y_n = H\,x + V_n \ ,$$

où H est une matrice $d \times m$.

- Si $m = d$, et si la matrice carrée H est inversible, alors on peut considérer l'estimateur suivant

  $$\widehat{x}_n = H^{-1}\;(\frac{1}{n}\sum_{k=1}^n Y_k) = H^{-1}\;(H\,x^0 + \frac{1}{n}\sum_{k=1}^n V_k) = x^0 + H^{-1}\;(\frac{1}{n}\sum_{k=1}^n V_k) \ .$$

  Sous l'hypothèse

  $$\frac{1}{n}\sum_{k=1}^n V_k \longrightarrow 0 \ , \tag{1.1}$$

quand le nombre n d'observations tend vers l'infini, on obtient la convergence de l'estimateur $\widehat{x}_n$ vers la vraie valeur du paramètre.

- Si $m > d$, alors le problème est en général mal–posé, même dans le cas favorable où la matrice H est de rang maximal égal à d, c'est–à–dire où la matrice carrée $H\,H^*$ est inversible. Considérons en effet le problème d'optimisation suivant

$$\min_{x \in \mathbf{R}^m} \; \{ \, \tfrac{1}{2} \sum_{k=1}^n |Y_k - H\,x|^2 \, \} \; .$$

Les conditions d'optimalité du premier ordre pour la minimisation par rapport à $x \in \mathbf{R}^m$ du critère

$$\tfrac{1}{2} \sum_{k=1}^n |Y_k - H\,x|^2 = \tfrac{1}{2} \sum_{k=1}^n |Y_k|^2 - x^*\,H^*\,(\sum_{k=1}^n Y_k) + n\,\tfrac{1}{2}\,x^*\,H^*\,H\,x \; ,$$

s'écrivent

$$\begin{aligned} & H^* \sum_{k=1}^n Y_k = n\,H^*\,H\,x \\ \Longrightarrow \quad & H^*\,(\frac{1}{n} \sum_{k=1}^n Y_k - H\,x) = 0 \\ \Longrightarrow \quad & H\,H^*\,(\frac{1}{n} \sum_{k=1}^n Y_k - H\,x) = 0 \\ \Longrightarrow \quad & H\,x = \frac{1}{n} \sum_{k=1}^n Y_k \; . \end{aligned}$$

Dans le cas précédent, où $m = d$ et la matrice H est inversible, on obtient la solution unique

$$\widehat{x}_n = H^{-1}\,(\frac{1}{n} \sum_{k=1}^n Y_k) \; .$$

Dans le cas considéré ici, il y a un nombre infini de solutions, et on peut seulement affirmer que

$$\widehat{x}_n \in \{ \, x \in \mathbf{R}^m \; : \; H\,x = \frac{1}{n} \sum_{k=1}^n Y_k \, \} \; .$$

On vérifie que

$$H\,\widehat{x}_n = \frac{1}{n} \sum_{k=1}^n Y_k = H\,x^0 + \frac{1}{n} \sum_{k=1}^n V_k \; ,$$

et à la limite quand le nombre n d'observations tend vers l'infini, on obtient sous l'hypothèse (1.1)

$$H\,\widehat{x}_n \longrightarrow H\,x^0 \; ,$$

c'est–à–dire qu'asymptotiquement, lorsque le bruit d'observation a été éliminé par moyennisation, on sait seulement que le paramètre inconnu x appartient au sous–espace affine $\mathcal{I}(x^0)$ de dimension $(m-d)$ défini par

$$\mathcal{I}(x^0) = \{\; x \in \mathbf{R}^m \; : \; H\, x = H\, x^0 \;\} \; .$$

L'existence d'un nombre infini de solutions possibles n'est donc pas liée à la présence du bruit d'observation. Elle existe même en absence de bruit d'observation, c'est–à–dire même si $V_n \equiv 0$, pour tout $n = 1, 2, \cdots$.

- Pour lever l'indétermination $x \in \mathcal{I}(x^0)$, on essaye d'utiliser des informations supplémentaires sur le paramètre inconnu x, par exemple : x est *proche* de μ, c'est–à–dire qu'on introduit une information *a priori.* On peut formaliser la prise en compte de cette information supplémentaire en considérant le problème d'optimisation suivant

$$\min_{x \in \mathbf{R}^m} \; \{\; \tfrac{1}{2} \sum_{k=1}^n |Y_k - H\, x|^2 + \tfrac{1}{2}\, (x-\mu)^*\, \Sigma^{-1}\, (x-\mu) \;\} \; ,$$

où Σ est une matrice symétrique définie positive, de dimension m. Les conditions d'optimalité du premier ordre pour la minimisation par rapport à $x \in \mathbf{R}^m$ du critère

$$\begin{aligned} &\tfrac{1}{2} \sum_{k=1}^n |Y_k - H\, x|^2 + \tfrac{1}{2}\, (x-\mu)^*\, \Sigma^{-1}\, (x-\mu) \\ &\qquad = \tfrac{1}{2} \sum_{k=1}^n |Y_k|^2 - x^*\, H^*\, (\sum_{k=1}^n Y_k) + n\, \tfrac{1}{2}\, x^*\, H^*\, H\, x \\ &\qquad\qquad + \tfrac{1}{2}\, \mu^*\, \Sigma^{-1}\, \mu - x^*\, \Sigma^{-1}\, \mu + \tfrac{1}{2}\, x^*\, \Sigma^{-1}\, x \; , \end{aligned}$$

s'écrivent

$$\begin{aligned} & H^*\, (\sum_{k=1}^n Y_k) + \Sigma^{-1}\, \mu = (n\, H^*\, H + \Sigma^{-1})\, x \\ \Longrightarrow \quad & (H^*\, H + \frac{1}{n}\, \Sigma^{-1})\, x = H^*\, (\frac{1}{n} \sum_{k=1}^n Y_k) + \frac{1}{n}\, \Sigma^{-1}\, \mu \; . \end{aligned}$$

En utilisant le résultat du Lemme 1.1 ci–dessous, avec le choix $R = I$ et $Q = n\, \Sigma$, on obtient

$$(H^*\, H + \frac{1}{n}\, \Sigma^{-1})^{-1} = n\, \Sigma - n\, \Sigma\, H^*\, (H\, \Sigma\, H^* + \frac{1}{n}\, I)^{-1}\, H\, \Sigma \; .$$

On en déduit

$$(H^*\, H + \frac{1}{n}\, \Sigma^{-1})^{-1}\, H^* = \Sigma\, H^*\, (H\, \Sigma\, H^* + \frac{1}{n}\, I)^{-1} \; ,$$

et

$$(H^* \, H + \frac{1}{n} \, \Sigma^{-1})^{-1} \, \frac{1}{n} \, \Sigma^{-1} = I - \Sigma \, H^* \, (H \, \Sigma \, H^* + \frac{1}{n} \, I)^{-1} \, H \ ,$$

ce qui donne la solution *unique* suivante

$$\widehat{x}_n = \Sigma \, H^* \, (H \, \Sigma \, H^* + \frac{1}{n} \, I)^{-1} \, (\frac{1}{n} \sum_{k=1}^n Y_k) + [\, I - \Sigma \, H^* \, (H \, \Sigma \, H^* + \frac{1}{n} \, I)^{-1} \, H \,] \, \mu \ .$$

On vérifie que

$$\begin{aligned} \widehat{x}_n \; &= \; \Sigma \, H^* \, (H \, \Sigma \, H^* + \frac{1}{n} \, I)^{-1} \, H \, x^0 + [\, I - \Sigma \, H^* \, (H \, \Sigma \, H^* + \frac{1}{n} \, I)^{-1} \, H \,] \, \mu \\ &\quad + \Sigma \, H^* \, (H \, \Sigma \, H^* + \frac{1}{n} \, I)^{-1} \, (\frac{1}{n} \sum_{k=1}^n V_k) \ , \end{aligned}$$

d'où on déduit la limite suivante

$$\widehat{x}_n \longrightarrow x^\perp = \Sigma \, H^* \, (H \, \Sigma \, H^*)^{-1} \, H \, x^0 + [\, I - \Sigma \, H^* \, (H \, \Sigma \, H^*)^{-1} \, H \,] \, \mu \ ,$$

quand le nombre n d'observations tend vers l'infini. L'inversibilité de la matrice $H \, \Sigma \, H^*$ est démontrée dans le Lemme 1.3 ci–dessous. On vérifie que

$$H \, x^\perp = H \, x^0 \ ,$$

c'est–à–dire que $x^\perp$ appartient au sous–espace affine $\mathcal{I}(x^0)$. On définit l'opérateur linéaire (matrice $m \times m$)

$$P_H = \Sigma \, H^* \, (H \, \Sigma \, H^*)^{-1} \, H \ .$$

On remarque d'abord que $H \, P_H = H$, c'est–à–dire que pour tout $x \in \mathbf{R}^m$, les points x et $P_H \, x$ appartiennent au même espace affine de dimension $(m - d)$ de $\mathbf{R}^m$. On en déduit également que P_H est un projecteur : en effet

$$P_H^2 = \Sigma \, H^* \, (H \, \Sigma \, H^*)^{-1} \, H \, P_H = \Sigma \, H^* \, (H \, \Sigma \, H^*)^{-1} \, H = P_H \ .$$

On remarque aussi que

$$P_H^* \, \Sigma^{-1} = \Sigma^{-1} \, P_H \ .$$

Au sous–espace affine $\mathcal{I}(x^0)$ est associé le sous–espace linéaire $\ker H$: en effet, deux points x' et x'' dans $\mathcal{I}(x^0)$ définissent un vecteur $u = x'' - x'$ qui appartient au noyau de H, puisque $H \, u = H \, x'' - H \, x' = 0$. Soit v un vecteur de l'image $\mathcal{R}(\Sigma \, H^*)$, c'est–à–dire un vecteur de la forme $v = \Sigma \, H^* \, \lambda$, pour un certain $\lambda \in \mathbf{R}^d$: ce vecteur est orthogonal (pour le produit scalaire associé à la matrice Σ^{-1}) à $\ker H$. En effet, pour tout $u \in \ker H$

$$v^* \, \Sigma^{-1} \, u = \lambda^* \, H \, u = 0 \ .$$

On a donc $\ker H \oplus \mathcal{R}(\Sigma\, H^*) = \mathbf{R}^m$. On remarque que pour tout $v \in \mathcal{R}(\Sigma\, H^*)$

$$P_H\, v = \Sigma\, H^*\, (H\, \Sigma\, H^*)^{-1}\, H\, \Sigma\, H^*\, \lambda = \Sigma\, H^*\, \lambda = v \ ,$$

c'est–à–dire que P_H laisse chaque vecteur du sous–espace linéaire $\mathcal{R}(\Sigma\, H^*)$ inchangé. D'autre part, pour tout $u \in \ker H$

$$P_H\, u = \Sigma\, H^*\, (H\, \Sigma\, H^*)^{-1}\, H\, u = 0 \ .$$

Il en résulte que P_H est le projecteur orthogonal (pour le produit scalaire associé à la matrice Σ^{-1}) sur le sous–espace linéaire $\mathcal{R}(\Sigma\, H^*) = (\ker H)^\perp$, et $(I - P_H)$ est le projecteur orthogonal sur le sous–espace linéaire $\ker H$. Finalement

$$x^\perp - x^0 = (I - P_H)\, (\mu - x^0) \ ,$$

c'est–à–dire que le vecteur $(x^\perp - x^0)$ est la projection orthogonale (pour le produit scalaire associé à la matrice Σ^{-1}) du vecteur $(\mu - x^0)$ sur le sous–espace linéaire $\ker H$. La valeur limite $x^\perp$ de l'estimateur, est donc le point de l'espace affine $\mathcal{I}(x^0)$ qui est le plus proche (pour le produit scalaire associé à la matrice Σ^{-1}) de l'estimateur a priori μ.

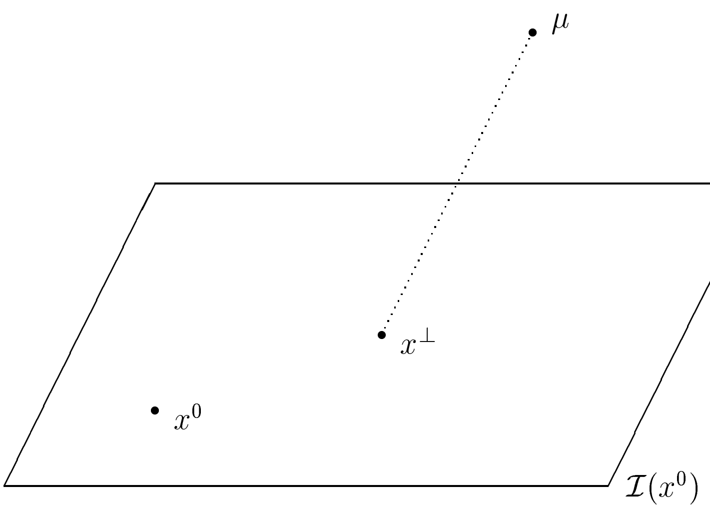


FIG. 1.1 – *Prise en compte de l'information a priori*

Lemme 1.1 *Soit Q et R deux matrices symétriques définies positives, de dimension m et d respectivement. Alors*

$$(H^*\, R^{-1}\, H + Q^{-1})^{-1} = Q - Q\, H^*\, (H\, Q\, H^* + R)^{-1}\, H\, Q \ ,$$

et de plus

$$(H^*\, R^{-1}\, H + Q^{-1})^{-1}\, H^* = Q\, H^*\, (H\, Q\, H^* + R)^{-1}\, R \ .$$

Preuve. On remarque d'abord que

$$H\,Q\,H^* + R \geq R \qquad \text{et} \qquad H^*\,R^{-1}\,H + Q^{-1} \geq Q^{-1}$$

au sens des matrices symétriques, ce qui prouve que les matrices

$$(H\,Q\,H^* + R) \qquad \text{et} \qquad (H^*\,R^{-1}\,H + Q^{-1})$$

sont inversibles. On vérifie alors que

$$\begin{aligned}[\,Q - Q\,H^*\,(H\,Q\,H^* + R)^{-1}\,H\,Q\,]\,[H^*\,R^{-1}\,H + Q^{-1}] = Q\,H^*\,R^{-1}\,H + I \\ - Q\,H^*\,(H\,Q\,H^* + R)^{-1}\,(H\,Q\,H^* + R - R)\,R^{-1}\,H \\ - Q\,H^*\,(H\,Q\,H^* + R)^{-1}\,H = I\ ,\end{aligned}$$

et d'autre part, en multipliant à droite par H^*, on obtient

$$\begin{aligned}(H^*\,R^{-1}\,H + Q^{-1})^{-1}\,H^* &= Q\,H^* - Q\,H^*\,(H\,Q\,H^* + R)^{-1}\,H\,Q\,H^* \\ &= Q\,H^* - Q\,H^*\,(H\,Q\,H^* + R)^{-1}\,(H\,Q\,H^* + R - R) \\ &= Q\,H^*\,(H\,Q\,H^* + R)^{-1}\,R\ . \qquad \Box\end{aligned}$$

Remarque 1.2 Cette formule d'inversion permet de remplacer l'inversion de la matrice $(H^*\,R^{-1}\,H + Q^{-1})$ de dimension m, par l'inversion de la matrice $(H\,Q\,H^* + R)$ de dimension d, avec en général $d \leq m$. En particulier, dans le cas où $d = 1$, la matrice H est un vecteur ligne $H = h^*$, la matrice R est un scalaire $R = r$, et la formule devient

$$(\frac{h\,h^*}{r} + Q^{-1})^{-1} = Q - \frac{Q\,h\,h^*\,Q}{r + h^*\,Q\,h}\ .$$

Lemme 1.3 *Soit Σ une matrice symétrique définie positive, de dimension m, et soit H une matrice $d \times m$, avec $d \leq m$, de rang plein égal à d. Alors la matrice $H\,\Sigma\,H^*$ est inversible.*

Preuve. Soit $u \in \mathbf{R}^d$ tel que

$$u^*\,(H\,\Sigma\,H^*)\,u = (H^*\,u)^*\,\Sigma\,(H^*\,u) = 0\ .$$

Comme Σ est inversible, alors nécessairement $H^*\,u = 0$, et comme H est de rang plein, on en déduit que $u = 0$. $\Box$

Il est donc important de disposer d'une information *a priori* sur l'état inconnu X_n, par exemple de disposer d'une équation d'état décrivant l'évolution de X_n quand n varie. On considérera deux types de modèles :

- les systèmes linéaires gaussiens,
- les chaînes de Markov à état fini.

Dans chacun de ces deux cas, il sera possible de résoudre le problème de filtrage de façon optimale, par la mise en œuvre :

- du filtre de Kalman, dans le cas des systèmes linéaires gaussiens,
- des équations forward–backward de Baum, ou de l'algorithme de Viterbi, dans le cas des chaînes de Markov à état fini.

Chapitre 2

Systèmes linéaires gaussiens

On appelle *processus aléatoire* en temps discret une famille $\{X_k\,, k \in \mathbf{N}\}$ de vecteurs aléatoires (notée $\{X_k\}$) définies sur un espace probabilisé $(\Omega,\mathcal{F},P)$ à valeurs dans $\mathbf{R}^m$. Un *processus aléatoire gaussien* est un processus aléatoire $\{X_k\}$ tel que pour tout N et tout $i_1,\ldots,i_N \in \mathbf{N}$

$$\begin{pmatrix} X_{i_1} \\ \vdots \\ X_{i_N} \end{pmatrix}$$

est un vecteur aléatoire gaussien (de dimension $m \times N$).

Deux processus aléatoires $\{X_k\}$ et $\{Y_k\}$ sont *indépendants* si pour tout N et M et $i_1,\ldots,i_N,j_1,\ldots,j_M \in \mathbf{N}$, les vecteurs aléatoires

$$\begin{pmatrix} X_{i_1} \\ \vdots \\ X_{i_N} \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} Y_{j_1} \\ \vdots \\ Y_{j_M} \end{pmatrix}$$

sont indépendants.

Un *bruit blanc* est un processus aléatoire $\{X_k\}$ tel que

$$\mathbf{E}[X_k] = 0 \;, \quad \mathbf{E}[X_k\, X_l^*] = 0 \;, \quad \text{si } k \neq l \;.$$

2.1 Systèmes totalement observés

On considère le système dynamique suivant

$$X_{k+1} = F_k\, X_k + f_k + G_k\, W_k \tag{2.1}$$

où $\{X_k\}$ et $\{W_k\}$ prennent respectivement leurs valeurs dans $\mathbf{R}^m$ et $\mathbf{R}^l$. On fait les hypothèses suivantes sur les coefficients : $F_k \in \mathbf{R}^{m\times m}$, $f_k \in \mathbf{R}^m$, $G_k \in \mathbf{R}^{m\times l}$, pour tout $k \in \mathbf{N}$. On suppose que

- *Le bruit $\{W_k\}$ est un bruit blanc gaussien de covariance Q^W_k.*
- *La condition initiale X_0 est gausienne, de moyenne $\bar{X}_0$ et de covariance Q^X_0.*
- *Le bruit $\{W_k\}$ et la condition initiale X_0 sont mutuellement indépendants.*

Proposition 2.1 *La sortie $\{X_k\}$ du système (2.1) est un processus gaussien à valeurs dans $\mathbf{R}^m$. En particulier, X_k est gaussien, de moyenne $\bar{X}_k$ et de matrice de covariance Q^X_k, avec*

$$\begin{aligned} \bar{X}_{k+1} &= F_k\,\bar{X}_k + f_k \;, \\ Q^X_{k+1} &= F_k\,Q^X_k\,F^*_k + G_k\,Q^W_k\,G^*_k \;. \end{aligned}$$

Preuve. Comme sortie d'un système linéaire à entrées gaussiennes, $\{X_k\}$ est un processus gaussien. En effet : pour tout N et $i_1,\cdots,i_N \geq 0$, il existe M tel que $\{i_1,\cdots,i_N\} \subset \{0,1,\cdots,M\}$ et on vérifie aisément qu'il existe une matrice $\mathcal{A} \in \mathbf{R}^{M\times(M+1)}$ et $\mathcal{B} \in \mathbf{R}^M$ tels que

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{pmatrix} = \mathcal{A} \begin{pmatrix} X_0 \\ W_0 \\ W_1 \\ \vdots \\ W_{M-1} \end{pmatrix} + \mathcal{B}$$

D'après les hypothèses $(X_0,W_0,W_1,\cdots,W_{M-1})$ est un vecteur gaussien, donc $(X_1,X_2,\cdots, X_M)$ est un vecteur aléatoire gaussien (comme tranformation affine d'un vecteur aléatoire gaussien).

Par ailleurs, d'après (2.1)

$$\begin{aligned} \bar{X}_{k+1} &= \mathbf{E}[X_{k+1}] \\ &= \mathbf{E}[F_k\,X_k + f_k + G_k\,W_k] \\ &= F_k\,\mathbf{E}[X_k] + f_k + G_k\,\mathbf{E}[W_k] \\ &= F_k\,\bar{X}_k + f_k \;, \end{aligned}$$

$$\begin{aligned} Q^X_{k+1} &= \mathbf{E}[(X_{k+1} - \bar{X}_{k+1})\,(X_{k+1} - \bar{X}_{k+1})^*] \\ &= \mathbf{E}[(F_k\,(X_k - \bar{X}_k) + G_k\,W_k)\,(F_k\,(X_k - \bar{X}_k) + G_k\,W_k)^*] \\ &= F_k\,\mathbf{E}[(X_k - \bar{X}_k)\,(X_k - \bar{X}_k)^*]\,F^*_k + G_k\,\mathbf{E}[W_k\,(X_k - \bar{X}_k)^*]\,F^*_k \\ &\qquad + F_k\,\mathbf{E}[(X_k - \bar{X}_k)\,W^*_k]\,G^*_k + G_k\,\mathbf{E}[W_k\,W^*_k]\,G^*_k \\ &= F_k\,Q^X_k\,F^*_k + G_k\,Q^W_k\,G^*_k \;. \end{aligned}$$

Dans cette dernière égalité on a utilisé le fait que $X_k - \bar{X}_k$ est indépendant de W_k, donc $\mathbf{E}[(X_k - \bar{X}_k)\,W^*_k] = 0$. □

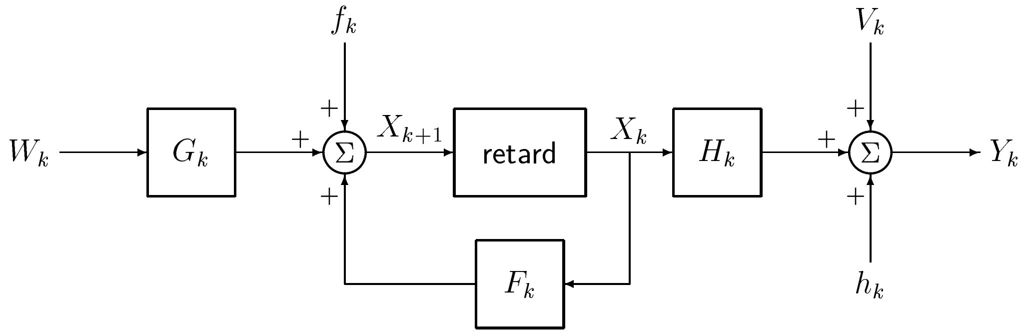


FIG. 2.1 – *Système linéaire discret avec observation partielle*

2.2 Systèmes partiellement observés

Coefficients déterministes

On considère le système dynamique suivant

$$X_{k+1} = F_k\, X_k + f_k + G_k\, W_k \;, \tag{2.2}$$

$$Y_k = H_k\, X_k + h_k + V_k \;, \tag{2.3}$$

où $\{X_k\}$, $\{Y_k\}$, $\{W_k\}$ et $\{V_k\}$ prennent respectivement leurs valeurs dans $\mathbf{R}^m$, $\mathbf{R}^d$, $\mathbf{R}^l$ et $\mathbf{R}^d$. On fait les hypothèses du paragraphe 2.1 sur les coefficients F_k, f_k, G_k, et de plus $H_k \in \mathbf{R}^{d\times m}$, $h_k \in \mathbf{R}^d$, pour tout $k \in \mathbf{N}$. On suppose que

- *Le bruit $\{W_k\}$ est un bruit blanc gaussien de covariance Q_k^W.*
- *La condition initiale X_0 est gaussienne, de moyenne $\bar{X}_0$ et de covariance Q_0^X.*
- *Le bruit d'observation $\{V_k\}$ est un bruit blanc gaussien de covariance Q_k^V.*
- *Les bruits $\{W_k\}$ et $\{V_k\}$, et la condition initiale X_0 sont mutuellement indépendants.*

Dans (2.2), (2.3), X_k représente l'état d'un système à l'instant k. On suppose que l'on ne peut pas observer directement ce système, mais que l'on dispose d'une observation Y_k qui est la somme d'un signal $H_k\, X_k + h_k$, et d'un bruit d'observation V_k (cf. le schéma 2.1).

D'après la Proposition 2.1, le processus $\{(X_k, Y_k)\}$ est un processus gaussien. En particulier, (X_k, Y_k) est un vecteur gaussien de moyenne et de matrice de covariance

$$\begin{pmatrix} \bar{X}_k \\ \bar{Y}_k \end{pmatrix} \;, \quad \begin{pmatrix} Q_k^X & Q_k^{XY} \\ Q_k^{YX} & Q_k^Y \end{pmatrix} \;,$$

respectivement, avec

$$\bar{X}_{k+1} = F_k\, \bar{X}_k + f_k \;,$$

$$
\begin{aligned}
Q^X_{k+1} &= F_k\, Q^X_k\, F^*_k + G_k\, Q^W_k\, G^*_k \ , \\
\bar{Y}_k &= H_k\, \bar{X}_k + h_k \ , \\
Q^Y_k &= H_k\, Q^X_k\, H^*_k + Q^V_k \ , \\
Q^{XY}_k &= Q^X_k\, H^*_k \ .
\end{aligned}
$$

Coefficients dépendants de l'observation

Dans le système (2.2), (2.3), les coefficients F_k, f_k, G_k, H_k et h_k, et les matrices de covariance Q^W_k, Q^V_k des bruits $\{W_k\}$, $\{V_k\}$ peuvent dépendre de l'observation $\{Y_k\}$ de la manière suivante :

- F_k, f_k, G_k et Q^W_k peuvent dépendre de $(Y_0,Y_1,\cdots,Y_k)$,
- H_k, h_k, et Q^V_k peuvent dépendre de $(Y_0,Y_1,\cdots,Y_{k-1})$.

Dans ce cas le processus $\{(X_k,Y_k)\}$ (et a fortiori $\{X_k\}$) n'est plus gaussien, mais le processus $\{X_k\}$ est *conditionnellement gaussien* par rapport à l'observation $\mathcal{Y}_{k-1} = \{Y_0,\ldots,Y_{k-1}\}$ (i.e. pour des valeurs de l'observation données, l'état du système (2.2) est gaussien).

Chapitre 3

Filtrage linéaire optimal

Le problème de filtrage (en temps discret) se présente en général de la manière suivante : on considère $\{X_k\}$, un processus (dont les caractéristiques statistiques sont connues) représentant l'état d'un système non observé. A l'instant k, on recueille une observation Y_k qui est formée d'un signal (i.e. une fonction $h(X_k)$ de l'état X_k) et d'un bruit additif V_k :

$$Y_k = h(X_k) + V_k \ .$$

Les caractéristiques statistiques du bruit de mesure $\{V_k\}$ sont également supposées connues. A l'instant k, on dispose de l'information $\mathcal{Y}_k = \{Y_0, \dots, Y_k\}$ et le but est d'obtenir *le plus d'information possible* sur l'état du système X_k (on veut, par exemple, pouvoir calculer un estimateur $\widehat{X}_k$ de X_k). Comme on le verra au paragraphe 3.1, la solution est de calculer la loi conditionnelle de X_k sachant $\mathcal{Y}_k$.

Dans le cas des systèmes décrits dans le chapitre 2, on est dans un cadre gaussien et l'évolution de cette loi conditionnelle (déterminée par sa moyenne et sa matrice de covariance) est régie par un système dynamique (le filtre de Kalman–Bucy) simple à mettre en œuvre (cf. paragraphe 3.2). Dans tous les autres cas (non linéaires), l'évolution de cette loi conditionnelle est determinée par un tout autre type de systèmes souvent impossibles à utiliser en pratique. Mais les techniques développées dans le cas linéaire peuvent s'étendre au cas non linéaire par des méthodes de linéarisation (cf. paragraphe 4). Les filtres ainsi obtenus sont souvent utilisables en pratique mais conduisent parfois à de mauvais résultats.

3.1 Critère d'estimation

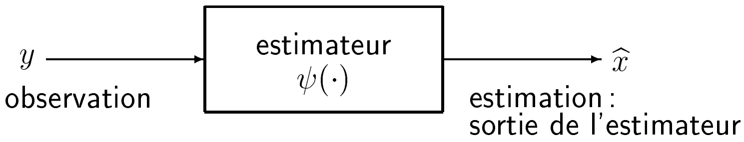
On considère le problème général suivant : étant donnés deux vecteurs aléatoires X et Y, qu'apporte le fait d'observer la réalisation $Y = y$ sur la connaissance que l'on a de X ?

Cas général

Soit X et Y deux vecteurs aléatoires de dimension m et d respectivement. Par définition, un *estimateur* de X à partir de l'observation de Y est une application mesurable

$$\mathbf{R}^d \ni y \to \psi(y) \in \mathbf{R}^m \ .$$

Par abus de langage, la variable aléatoire $\psi(Y)$ sera également notée ψ. Pour une réalisation particulière $Y = y$ de l'observation (y fixé), $\widehat{x} = \psi(y)$ est appelée une *estimation* de X.



Estimateur du minimum de variance Soit $\psi(\cdot)$ un estimateur de X sachant Y. Naturellement $\psi = \psi(Y)$ n'est pas égal à X : une mesure de l'écart entre l'estimateur et la vraie valeur est fournie par la variance de l'erreur d'estimation (ou *erreur quadratique moyenne*)

$$\mathbf{E}[\,|X - \psi(Y)|^2\,] \ . \tag{3.1}$$

L'estimateur du minimum de variance de X sachant Y est un estimateur $\widehat{X}(\cdot)$ tel que

$$\mathbf{E}[\,|X - \widehat{X}(Y)|^2\,] \leq \mathbf{E}[\,|X - \psi(Y)|^2\,]$$

pour tout autre estimateur $\psi(\cdot)$.

La Proposition 3.1 ci–dessous montre que cet estimateur est obtenu à l'aide de la densité conditionnelle $p_{X|Y=y}(x)$ de X sachant $Y = y$, définie par

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{\displaystyle\int_{\mathbf{R}^m} p_{X,Y}(x,y)\,dx} = \frac{p_{X,Y}(x,y)}{p_Y(y)} \ , \tag{3.2}$$

où $p_{X,Y}$ désigne la densité conjointe des variables aléatoires X et Y.

Proposition 3.1 *Soit X et Y des vecteurs aléatoires de dimension m et d respectivement. L'estimateur du minimum de variance de X sachant Y est la moyenne conditionnelle, i.e.*

$$\widehat{X}(y) = \mathbf{E}[X \mid Y = y] = \int_{\mathbf{R}^m} x\, p_{X|Y=y}(x)\,dx \ .$$

Preuve. Soit $\psi(\cdot)$ un estimateur quelconque.

$$\begin{aligned}\mathbf{E}[\,|X - \psi(Y)|^2\,] \;=\; & \mathbf{E}[\,|X - \widehat{X}(Y)|^2\,] + 2\,\mathbf{E}[\,(\widehat{X}(Y) - \psi(Y))^*\,(X - \widehat{X}(Y))\,] \\ & + \mathbf{E}[\,|\widehat{X}(Y) - \psi(Y)|^2\,] \ ,\end{aligned}$$

et on remarque que

$$\mathbf{E}[\,(\widehat{X}(Y) - \psi(Y))^* \,(X - \widehat{X}(Y))\,] =$$

$$= \int_{\mathbf{R}^m} \int_{\mathbf{R}^d} (\widehat{X}(y) - \psi(y))^* \,(x - \widehat{X}(y))\, p_{X,Y}(x,y)\, dx\, dy$$

$$= \int_{\mathbf{R}^d} (\widehat{X}(y) - \psi(y))^* \,\Big\{ \int_{\mathbf{R}^m} (x - \widehat{X}(y))\, p_{X|Y=y}(x)\, dx \Big\}\, p_Y(y)\, dy = 0 \ ,$$

par définition de $\widehat{X}(y)$ (on peut aussi utiliser directement le résultat de la Proposition A.4). On a donc

$$\mathbf{E}[\,|X - \psi(Y)|^2\,] \,=\, \mathbf{E}[\,|X - \widehat{X}(Y)|^2\,] + \int_{\mathbf{R}^d} |\widehat{X}(y) - \psi(y)|^2\, p_Y(y)\, dy \ ,$$

et le vecteur $\psi(y)$ qui minimise cette expression est $\psi(y) = \widehat{X}(y)$ □

Biais d'un estimateur Soit X et Y des vecteurs aléatoires et $\psi(\cdot)$ un estimateur de X sachant Y. Le biais de ψ est défini par

$$b(y) \stackrel{\triangle}{=} \mathbf{E}[X - \psi(y) \mid Y = y] \ .$$

On dit que $\psi(\cdot)$ est un estimateur sans biais si $b(y) = 0$ pour tout y. D'après la définition de la moyenne conditionnelle $\widehat{X}$ de X sachant Y, le résultat suivant est immédiat

Proposition 3.2 *La moyenne conditionnelle $\widehat{X}$ de X sachant Y est un estimateur sans biais.*

Cas gaussien

On va maintenant préciser le résultat général obtenu ci–dessus, dans le cas particulier des vecteurs aléatoires gaussiens.

Proposition 3.3 *Soit $Z = (X,Y)$ un vecteur aléatoire gaussien de dimension $m + d$, de moyenne et de matrice de covariance*

$$\bar{Z} = \begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \ , \quad Q_Z = \begin{pmatrix} Q_X & Q_{XY} \\ Q_{YX} & Q_Y \end{pmatrix} \ ,$$

respectivement. Si la matrice Q_Y est inversible, alors la densité conditionnelle $p_{X|Y=y}(x)$ du vecteur aléatoire X sachant $Y = y$, est une densité gaussienne de moyenne

$$\widehat{X}(y) = \bar{X} + Q_{XY}\, Q_Y^{-1}\, (y - \bar{Y}) \ ,$$

et de matrice de covariance

$$R = Q_X - Q_{XY}\, Q_Y^{-1}\, Q_{YX} \ .$$

Preuve. On donne une première démonstration, dans le cas où la matrice Q_Z est inversible. Par définition

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{\dfrac{1}{(\sqrt{2\pi})^{m+d}\sqrt{\det Q_Z}} \; \exp\{ -\tfrac{1}{2}(z-\bar{Z})^* Q_Z^{-1} (z-\bar{Z}) \}}{\dfrac{1}{(\sqrt{2\pi})^{d}\sqrt{\det Q_Y}} \; \exp\{ -\tfrac{1}{2}(y-\bar{Y})^* Q_Y^{-1} (y-\bar{Y}) \}} \; .$$

On utilise la formule suivante (simple à vérifier)

$$\begin{pmatrix} I & -Q_{XY}\,Q_Y^{-1} \\ 0 & I \end{pmatrix} Q_Z \begin{pmatrix} I & 0 \\ -Q_Y^{-1}\,Q_{XY}^* & I \end{pmatrix} = \begin{pmatrix} Q_X - Q_{XY}\,Q_Y^{-1}\,Q_{YX} & 0 \\ 0 & Q_Y \end{pmatrix} \; . \tag{3.3}$$

En prenant le déterminant dans (3.3), on obtient

$$1 \times \det Q_Z \times 1 = \det R \times \det Q_Y \; .$$

L'identité (3.3) implique aussi que

$$\begin{pmatrix} I & 0 \\ -Q_Y^{-1}\,Q_{XY}^* & I \end{pmatrix}^{-1} Q_Z^{-1} \begin{pmatrix} I & -Q_{XY}\,Q_Y^{-1} \\ 0 & I \end{pmatrix}^{-1} = \begin{pmatrix} R^{-1} & 0 \\ 0 & Q_Y^{-1} \end{pmatrix} \; ,$$

c'est–à–dire

$$Q_Z^{-1} = \begin{pmatrix} I & 0 \\ -Q_Y^{-1}\,Q_{XY}^* & I \end{pmatrix} \begin{pmatrix} R^{-1} & 0 \\ 0 & Q_Y^{-1} \end{pmatrix} \begin{pmatrix} I & -Q_{XY}\,Q_Y^{-1} \\ 0 & I \end{pmatrix} \; .$$

Donc

$$(z-\bar{Z})^* Q_Z^{-1} (z-\bar{Z}) = [x - \widehat{X}(y)]^* R^{-1} [x - \widehat{X}(y)] + (y-\bar{Y})^* Q_Y^{-1} (y-\bar{Y}) \; ,$$

et

$$p_{X|Y=y}(x) = \frac{1}{(\sqrt{2\pi})^{n}\sqrt{\det R}} \; \exp\{ -\tfrac{1}{2}[x - \widehat{X}(y)]^* R^{-1} [x - \widehat{X}(y)] \} \; ,$$

ce qui montre le résultat. □

Preuve (cas général). Dans le cas où la matrice Q_Z n'est pas nécessairement inversible, on montre que la fonction caractéristique de la loi conditionnelle du vecteur aléatoire X sachant $Y = y$ est égale à

$$\exp\{ i\,u^* \widehat{X}(y) - \tfrac{1}{2}\,u^* R u \} \; ,$$

c'est–à–dire que la loi conditionnelle du vecteur aléatoire X sachant $Y = y$ est une loi gaussienne de moyenne $\widehat{X}(y)$ et de matrice de covariance R. Par définition

$$\begin{aligned} \Phi_{X,Y}(u,v) = \mathbf{E}[\, e^{\,i\,u^* X + i\,v^* Y} \,] &= \mathbf{E}[\, e^{\,i\,v^* Y} \; \mathbf{E}[\, e^{\,i\,u^* X} \mid Y \,]\,] \\ &= \mathbf{E}[\, e^{\,i\,v^* Y} \; \Phi_{X|Y}(u) \,] = \int_{\mathbf{R}^d} e^{\,i\,v^* y} \; \Phi_{X|Y=y}(u) \; p_Y(y)\, dy \; , \end{aligned}$$

et on vérifie d'autre part que

$$\begin{aligned}
&\int_{\mathbf{R}^d} e^{\,i\,v^* y}\ \exp\{\ i\,u^*\widehat{X}(y) - \tfrac{1}{2}\,u^* R u\ \}\ p_Y(y)\,dy = \\
&= \exp\{\ i\,u^*\bar{X} - i\,u^*\,Q_{XY}\,Q_Y^{-1}\,\bar{Y} - \tfrac{1}{2}\,u^* R u\ \}\ \int_{\mathbf{R}^d} e^{\,i\,v^* y}\ \exp\{\ i\,u^*\,Q_{XY}\,Q_Y^{-1}\,y\ \}\ p_Y(y)\,dy \\
&= \exp\{\ i\,u^*\bar{X} - i\,u^*\,Q_{XY}\,Q_Y^{-1}\,\bar{Y} - \tfrac{1}{2}\,u^* R u\ \}\ \Phi_Y(v + Q_Y^{-1}\,Q_{YX}\,u) \\
&= \exp\{\ i\,u^*\bar{X} - i\,u^*\,Q_{XY}\,Q_Y^{-1}\,\bar{Y} - \tfrac{1}{2}\,u^* Q_X\,u + \tfrac{1}{2}\,u^* Q_{XY}\,Q_Y^{-1}\,Q_{YX}\,u \\
&\qquad + i\,(v^* + u^*\,Q_{XY}\,Q_Y^{-1})\,\bar{Y} - \tfrac{1}{2}\,(v^* + u^*\,Q_{XY}\,Q_Y^{-1})\,Q_Y\,(v + Q_Y^{-1}\,Q_{YX}\,u)\ \} \\
&= \exp\{\ i\,u^*\bar{X} + i\,v^*\bar{Y} - \tfrac{1}{2}\,u^* Q_X u - u^* Q_{XY} v - \tfrac{1}{2}\,v^* Q_Y v\ \} = \Phi_{X,Y}(u,v)\ .
\end{aligned}$$

Par injectivité de la transformé de Fourier, on obtient

$$\Phi_{X|Y=y}(u) = \exp\{\ i\,u^*\widehat{X}(y) - \tfrac{1}{2}\,u^* R u\ \}\ . \qquad \square$$

Remarque 3.4 On vérifie que

$$0 \leq R \leq Q_X$$

au sens des matrices symétriques, c'est–à–dire que l'utilisation de l'information supplémentaire $(Y = y)$, ne peut que réduire l'incertitude que l'on a sur le vecteur aléatoire X.

3.2 Filtre de Kalman–Bucy

On considère un système linéaire du type (2.2), c'est–à–dire

$$X_{k+1} = F_k\,X_k + f_k + G_k\,W_k\ , \tag{3.4}$$

$$Y_k = H_k\,X_k + h_k + V_k\ , \tag{3.5}$$

avec les hypothèses du paragraphe 2.2. A l'instant k on dispose de l'information

$$\mathcal{Y}_k \triangleq (Y_0,Y_1,\ldots,Y_k)\ .$$

L'objectif est d'estimer le vecteur aléatoire X_k à partir de $\mathcal{Y}_k$, de façon optimale et récursive. Si on adopte le critère du minimum de variance, il s'agit d'après le paragraphe 3.1 de calculer la loi conditionnelle du vecteur aléatoire X_k sachant $\mathcal{Y}_k$. Comme le cadre est gaussien, il suffit de calculer la moyenne et la matrice de covariance

$$\widehat{X}_k \triangleq \mathbf{E}[X_k \mid \mathcal{Y}_k]\ , \quad P_k \triangleq \mathbf{E}[(X_k - \widehat{X}_k)\,(X_k - \widehat{X}_k)^* \mid \mathcal{Y}_k]\ .$$

On définit également les quantités suivantes

$$\widehat{X}_k^- \triangleq \mathbf{E}[X_k \mid \mathcal{Y}_{k-1}]\ , \quad P_k^- \triangleq \mathbf{E}[(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^* \mid \mathcal{Y}_{k-1}]\ .$$

Remarque 3.5 Par définition, $\widehat{X}_k^-$ est la projection orthogonale de X_k sur le sous–espace $\mathcal{H}_{k-1}$, donc $(X_k - \widehat{X}_k^-) \perp \mathcal{H}_{k-1}$, et à cause du caractère gaussien $(X_k - \widehat{X}_k^-) \perp \mathcal{Y}_{k-1}$. De même, $\widehat{X}_k$ est la projection orthogonale de X_k sur le sous–espace $\mathcal{H}_k$, donc $(X_k - \widehat{X}_k) \perp \mathcal{H}_k$, et à cause du caractère gaussien $(X_k - \widehat{X}_k) \perp \mathcal{Y}_k$. On en déduit que

$$P_k^- = \mathbf{E}[(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^* \mid \mathcal{Y}_{k-1}] = \mathbf{E}[(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^*] \;,$$

et

$$P_k = \mathbf{E}[(X_k - \widehat{X}_k)\,(X_k - \widehat{X}_k)^* \mid \mathcal{Y}_k] = \mathbf{E}[(X_k - \widehat{X}_k)\,(X_k - \widehat{X}_k)^*] \;,$$

c'est–à–dire que les matrices de covariance P_k^- et P_k sont *déterministes.*

Supposons connue la loi conditionnelle du vecteur aléatoire X_{k-1} sachant $\mathcal{Y}_{k-1}$. Pour calculer la loi conditionnelle du vecteur aléatoire X_k sachant $\mathcal{Y}_k$, on procède en deux étapes.

- Dans l'étape de *prédiction*, on calcule la loi conditionnelle du vecteur aléatoire X_k sachant les observations passées $\mathcal{Y}_{k-1}$, ce qui est facile à partir de l'équation (3.4).
- Dans l'étape de *correction*, on utilise la nouvelle observation Y_k. En particulier, on considère la composante de l'observation Y_k qui apporte une information nouvelle par rapport aux observations passées $\mathcal{Y}_{k-1}$, c'est–à–dire

  $$I_k = Y_k - \mathbf{E}[Y_k \mid \mathcal{Y}_{k-1}] \;.$$

  D'après l'équation (3.5)

  $$I_k = Y_k - (H_k\,\mathbf{E}[X_k \mid \mathcal{Y}_{k-1}] + h_k + \mathbf{E}[V_k \mid \mathcal{Y}_{k-1}]) = Y_k - (H_k\,\widehat{X}_k^- + h_k) \;,$$

  compte tenu que V_k et $\mathcal{Y}_{k-1}$ sont indépendants.

Lemme 3.6 *Le processus $\{I_k\}$ est un processus gaussien à valeurs dans $\mathbf{R}^d$, appelé* processus d'innovation. *En particulier, I_k est un vecteur aléatoire gaussien de dimension d, de moyenne nulle et de matrice de covariance*

$$Q_k^I = H_k\,P_k^-\,H_k^* + Q_k^V \;,$$

indépendant de $\mathcal{Y}_{k-1}$.

Preuve. D'après l'équation (3.5)

$$I_k = Y_k - (H_k\,\widehat{X}_k^- + h_k) = H_k\,(X_k - \widehat{X}_k^-) + V_k \;. \qquad \square$$

Remarque 3.7 On en déduit également que

$$\mathbf{E}[(X_k - \widehat{X}_k^-)\,I_k^*] = P_k^-\,H_k^* \;.$$

Théorème 3.8 (Filtre de Kalman–Bucy) *On suppose que la matrice de covariance Q^V_k est* inversible, *pour tout $k \in \mathbf{N}$. Alors $\{\widehat{X}_k\}$ et $\{P_k\}$ sont définis par les équations suivantes*

$$\widehat{X}^-_{k+1} = F_k\,\widehat{X}_k + f_k \ , \tag{3.6}$$

$$P^-_{k+1} = F_k\,P_k\,F^*_k + G_k\,Q^W_k\,G^*_k \ , \tag{3.7}$$

et

$$\widehat{X}_k = \widehat{X}^-_k + K_k\,[Y_k - (H_k\,\widehat{X}^-_k + h_k)] \ , \tag{3.8}$$

$$P_k = [I - K_k\,H_k]\,P^-_k \tag{3.9}$$

$$K_k = P^-_k\,H^*_k\,[H_k\,P^-_k\,H^*_k + Q^V_k]^{-1} \ ,$$

avec les initialisations

$$\widehat{X}^-_0 = \bar{X}_0 = \mathbf{E}[X_0] \ , \quad P^-_0 = Q^X_0 = \mathrm{cov(X_0)} \ .$$

La matrice K_k est appelé le gain *du filtre de Kalman–Bucy.*

Remarque 3.9 La suite $\{P_k\}$ ne dépend pas des observations $\{Y_k\}$, ni des coefficients $\{f_k\}$ et $\{h_k\}$. Elle peut donc être pré–calculée, en particulier dans le cas simple où les coefficients $\{F_k\}$, $\{G_k\}$, $\{H_k\}$, $\{Q^W_k\}$ et $\{Q^V_k\}$ sont constants, c'est–à–dire où

$$F_k = F \ , \quad G_k = G \ , \quad H_k = H \ , \quad Q^W_k = Q^W \ , \quad Q^V_k = Q^V \ ,$$

pour tout $k \geq 0$.

Construction du filtre de Kalman–Bucy On procède en plusieurs étapes. Le point central est la Proposition 3.3 qui sera constamment utilisée.

$\boxed{1}$ *On montre (3.8), (3.9) pour $k = 0$*

Le vecteur aléatoire (X_0,Y_0) est gaussien, de moyenne et de matrice de covariance données par

$$\begin{pmatrix} \bar{X}_0 \\ H_0\,\bar{X}_0 + h_0 \end{pmatrix} \ , \quad \begin{pmatrix} Q^X_0 & Q^X_0\,H^*_0 \\ H_0\,Q^X_0 & H_0\,Q^X_0\,H^*_0 + Q^V_0 \end{pmatrix} \ ,$$

respectivement. D'après la Proposition 3.3, la loi de X_0 sachant $\mathcal{Y}_0$ est gaussienne, de moyenne

$$\widehat{X}_0 = \bar{X}_0 + Q^X_0\,H^*_0\,[H_0\,Q^X_0\,H^*_0 + Q^V_0]^{-1}\,[Y_0 - (H_0\,\bar{X}_0 + h_0)] \ ,$$

et de matrice de covariance

$$P_0 = Q^X_0 - Q^X_0\,H^*_0\,[H_0\,Q^X_0\,H^*_0 + Q^V_0]^{-1}\,H_0\,Q^X_0 \ .$$

$\boxed{2}$ *On montre (3.6), (3.7)*

Le vecteur aléatoire $(X_{k+1}, Y_0, \cdots, Y_k)$ est gaussien, et d'après la Proposition 3.3, la loi de X_{k+1} sachant $\mathcal{Y}_k$ est gaussienne, de moyenne $\widehat{X}^-_{k+1}$ et de matrice de covariance P^-_{k+1}. D'après l'équation

$$X_{k+1} = F_k\, X_k + f_k + G_k\, W_k \ , \tag{3.10}$$

on a

$$\begin{aligned}
\widehat{X}^-_{k+1} &= \mathbf{E}[X_{k+1} \mid \mathcal{Y}_k] \\
&= F_k\, \mathbf{E}[X_k \mid \mathcal{Y}_k] + f_k + G_k\, \mathbf{E}[W_k \mid \mathcal{Y}_k] \\
&= F_k\, \widehat{X}_k + f_k \ ,
\end{aligned}$$

compte tenu que W_k et $\mathcal{Y}_k$ sont indépendants. Par différence

$$X_{k+1} - \widehat{X}^-_{k+1} = F_k\, (X_k - \widehat{X}_k) + G_k\, W_k \ ,$$

de sorte que

$$\begin{aligned}
P^-_{k+1} &= \mathbf{E}[(X_{k+1} - \widehat{X}^-_{k+1})\, (X_{k+1} - \widehat{X}^-_{k+1})^*] \\
&= \mathbf{E}[(F_k\, (X_k - \widehat{X}_k) + G_k\, W_k)\, (F_k\, (X_k - \widehat{X}_k) + G_k\, W_k)^*] \\
&= F_k\, \mathbf{E}[(X_k - \widehat{X}_k)\, (X_k - \widehat{X}_k)^*]\, F^*_k + G_k\, \mathbf{E}[W_k\, (X_k - \widehat{X}_k)^*]\, F^*_k \\
&\quad + F_k\, \mathbf{E}[(X_k - \widehat{X}_k)\, W^*_k]\, G^*_k + G_k\, \mathbf{E}[W_k\, W^*_k]\, G^*_k \\
&= F_k\, P_k\, F^*_k + G_k\, Q^W_k\, G^*_k \ .
\end{aligned}$$

Dans cette dernière égalité on a utilisé le fait que $(X_k - \widehat{X}_k)$ est indépendant de W_k, donc $\mathbf{E}[(X_k - \widehat{X}_k)\, W^*_k] = 0$. □

$\boxed{3}$ *On montre (3.8), (3.9)*

Le vecteur aléatoire $(X_k, Y_0, \cdots, Y_k)$ est gaussien, et d'après la Proposition 3.3, la loi de X_k sachant $\mathcal{Y}_k$ est gaussienne, de moyenne $\widehat{X}_k$ et de matrice de covariance déterministe P_k. D'après le Lemme 3.6

$$\begin{aligned}
\widehat{X}_k &= \mathbf{E}[X_k \mid \mathcal{Y}_k] \\
&= \widehat{X}^-_k + \mathbf{E}[X_k - \widehat{X}^-_k \mid \mathcal{Y}_k] \\
&= \widehat{X}^-_k + \mathbf{E}[X_k - \widehat{X}^-_k \mid \mathcal{Y}_{k-1}] + \mathbf{E}[X_k - \widehat{X}^-_k \mid I_k] \\
&= \widehat{X}^-_k + \mathbf{E}[X_k - \widehat{X}^-_k \mid I_k] \ .
\end{aligned}$$

Donc

$$\begin{aligned}
X_k - \widehat{X}_k &= (X_k - \widehat{X}^-_k) - (\widehat{X}_k - \widehat{X}^-_k) \\
&= (X_k - \widehat{X}^-_k) - \mathbf{E}[X_k - \widehat{X}^-_k \mid I_k] \ .
\end{aligned}$$

Il suffit donc de calculer la moyenne et la matrice de covariance de la loi conditionnelle du vecteur aléatoire $(X_k - \widehat{X}_k^-)$ sachant I_k. Le vecteur aléatoire $(X_k - \widehat{X}_k^-, I_k)$ est un vecteur aléatoire gaussien, de moyenne nulle et de matrice de covariance

$$\begin{pmatrix} P_k^- & P_k^- \, H_k^* \\ H_k \, P_k^- & H_k \, P_k^- \, H_k^* + Q_k^V \end{pmatrix} .$$

L'hypothèse que Q_k^V est inversible entraîne que $H_k \, P_k^- \, H_k^* + Q_k^V$ est inversible. D'après la Proposition 3.3, on a immédiatement

$$\widehat{X}_k = \widehat{X}_k^- + P_k^- \, H_k^* \, [H_k \, P_k^- \, H_k^* + Q_k^V]^{-1} \, I_k \ ,$$

et

$$P_k = P_k^- - P_k^- \, H_k^* \, [H_k \, P_k^- \, H_k^* + Q_k^V]^{-1} \, H_k \, P_k^- \ ,$$

ce qui termine la démonstration. □

Chapitre 4

Extensions au cas non–linéaire

On considère un système non linéaire

$$X_{k+1} = f_k(X_k) + g_k(X_k)\, W_k \;, \tag{4.1}$$

$$Y_k = h_k(X_k) + V_k \;, \tag{4.2}$$

où $\{X_k\}$, $\{Y_k\}$, $\{W_k\}$, $\{V_k\}$ prennent respectivement leurs valeurs dans $\mathbf{R}^m$, $\mathbf{R}^d$, $\mathbf{R}^l$ et $\mathbf{R}^d$, et où les applications $f_k(\cdot)$, $g_k(\cdot)$ et $h_k(\cdot)$ sont définies sur $\mathbf{R}^m$, à valeurs dans $\mathbf{R}^m$, $\mathbf{R}^{m\times l}$ et $\mathbf{R}^d$ respectivement. On va supposer que les fonctions $f_k(\cdot)$ et $h_k(\cdot)$ sont dérivables par rapport à la variable x. $\{W_k\}$ et $\{V_k\}$ sont des bruits blancs gaussiens (de covariances respectives Q_k^W et Q_k^V) indépendants entre eux et indépendants de la condition initiale X_0 de (4.1).

Pour le système (4.1), (4.2), la plupart des propriétés obtenues au chapitre 2 ne sont plus vraies. En particulier le processus solution de (4.1), (4.2) n'est pas gaussien (ni même conditionnellement gaussien), ses moments ne peuvent pas être calculés de manière simple.

4.1 Filtre de Kalman linéarisé

On se donne $\{\bar{x}_k\}$ une suite (déterministe) dans $\mathbf{R}^m$, appelée *trajectoire nominale* (on peut prendre par exemple $\bar{x}_k$ comme une approximation de la moyenne de X_k). La méthode consiste à linéariser les fonctions f_k,g_k,h_k autour de $\bar{x}_k$, c'est–à–dire

$$\begin{aligned} f_k(x) &\simeq f_k(\bar{x}_k) + \nabla f_k(\bar{x}_k)\,(x - \bar{x}_k) \;, \\ g_k(x) &\simeq g_k(\bar{x}_k) \;, \\ h_k(x) &\simeq h_k(\bar{x}_k) + \nabla h_k(\bar{x}_k)\,(x - \bar{x}_k) \;. \end{aligned}$$

On note

$$\begin{aligned} F_k &\triangleq \nabla f_k(\bar{x}_k) \;, \quad f_k \triangleq f_k(\bar{x}_k) \;, \\ G_k &\triangleq g_k(\bar{x}_k) \;, \\ H_k &\triangleq \nabla h_k(\bar{x}_k) \;, \quad h_k \triangleq h_k(\bar{x}_k) \;. \end{aligned}$$

Le système (4.1) (4.2) est alors approché par

$$X_{k+1} = F_k\,[X_k - \bar{x}_k] + f_k + G_k\,W_k \ , \tag{4.3}$$
$$Y_k = H_k\,[X_k - \bar{x}_k] + h_k + V_k \ . \tag{4.4}$$

On applique le filtre de Kalman–Bucy, et on obtient

$$\begin{aligned}
\widehat{X}^-_{k+1} &= f_k(\bar{x}_k) + \nabla f_k(\bar{x}_k)\,(\widehat{X}_k - \bar{x}_k) \ , \\
P^-_{k+1} &= \nabla f_k(\bar{x}_k)\,P_k\,[\nabla f_k(\bar{x}_k)]^* + g_k(\bar{x}_k)\,Q^W_k\,[g_k(\bar{x}_k)]^* \ , \\[1em]
\widehat{X}_k &= \widehat{X}^-_k + K_k\,[\,Y_k - [\nabla h_k(\bar{x}_k)\,(\widehat{X}^-_k - \bar{x}_k) + h_k(\bar{x}_k)]\,] \ , \\
P_k &= [I - K_k\,\nabla h_k(\bar{x}_k)]\;P^-_k \ , \\
K_k &= P^-_k\,\nabla h_k(\bar{x}_k)^*\,[\,\nabla h_k(\bar{x}_k)\,P^-_k\,[\nabla h_k(\bar{x}_k)]^* + Q^V_k\,]^{-1} \ .
\end{aligned}$$

A la place de la première et la troisième de ces équations, on peut utiliser

$$\begin{aligned}
\widehat{X}^-_{k+1} &= f_k(\widehat{X}_k) \ , \\
\widehat{X}_k &= \widehat{X}^-_k + K_k\,[Y_k - h_k(\widehat{X}^-_k)] \ .
\end{aligned}$$

On choisit enfin l'initialisation $\widehat{X}^-_0$ et R^-_0 de telle sorte que $N(\widehat{X}^-_0, R^-_0)$ soit une bonne approximation de la loi de X_0.

Résultat 4.1 (filtre de Kalman linéarisé)

$$\begin{aligned}
\widehat{X}^-_{k+1} &= f_k(\widehat{X}_k) \ , \\
P^-_{k+1} &= F_k\,P_k\,F^*_k + G_k\,Q^W_k\,G^*_k \ , \\[1em]
\widehat{X}_k &= \widehat{X}^-_k + K_k\,[Y_k - h_k(\widehat{X}^-_k)] \ , \\
P_k &= [I - K_k\,H_k]\,P^-_k \ , \\
K_k &= P^-_k\,H^*_k\,[H_k\,P^-_k\,H^*_k + Q^V_k]^{-1} \ ,
\end{aligned}$$

avec : $F_k \triangleq \nabla f_k(\bar{x}_k)$, $G_k \triangleq g_k(\bar{x}_k)$, et $H_k \triangleq \nabla h_k(\bar{x}_k)$, où $\{\bar{x}_k\}$ est une trajectoire nominale donnée (une suite dans $\mathbf{R}^m$).

4.2 Filtre de Kalman étendu

On a vu (cf. paragraphe 3.2) que les coefficients du système linéaire peuvent dépendre des observations (jusqu'à l'instant $k-1$). Donc, au lieu d'utiliser une trajectoire nominale déterministe $\{\bar{x}_k\}$, on peut utiliser des estimateurs de X_k : $\widehat{X}_k$ pour l'étape de prédiction et $\widehat{X}_k^-$ pour l'étape de correction. Pour l'étape de prédiction (à l'instant k) on utilise la linéarisation

$$\begin{aligned} f_k(x) &\simeq f_k(\widehat{X}_k) + \nabla f_k(\widehat{X}_k)\,(x - \widehat{X}_k) \ , \\ g_k(x) &\simeq g_k(\widehat{X}_k) \ , \end{aligned}$$

pour l'étape de correction (à l'instant k) on utilise les linéarisations

$$h_k(x) \simeq h_k(\widehat{X}_k^-) + \nabla h_k(\widehat{X}_k^-)\,(x - \widehat{X}_k^-) \ .$$

Résultat 4.2 (filtre de Kalman étendu)

$$\begin{aligned} \widehat{X}_{k+1}^- &= f_k(\widehat{X}_k) \ , \\ P_{k+1}^- &= F_k\,P_k\,F_k^* + G_k\,Q_k^W\,G_k^* \ , \\ \\ \widehat{X}_k &= \widehat{X}_k^- + K_k\,[Y_k - h_k(\widehat{X}_k^-)] \ , \\ P_k &= [I - K_k\,H_k]\,P_k^- \ , \\ K_k &= P_k^-\,H_k^*\,[H_k\,P_k^-\,H_k^* + Q_k^V]^{-1} \ , \end{aligned}$$

avec : $F_k \triangleq \nabla f_k(\widehat{X}_k)$, $G_k \triangleq g_k(\widehat{X}_k)$, *et* $H_k \triangleq \nabla h_k(\widehat{X}_k^-)$.

Remarque 4.3

- On peut s'attendre à de bons résultats avec cette technique de filtrage lorsque l'on est proche d'une situation "linéaire" ou lorsque le rapport signal/bruit est grand.
- Pour vérifier si le filtre de Kalman étendu se comporte bien, on peut, en sortie, tester le processus de "pseudo–innovation"

  $$I_k \triangleq Y_k - h_k(\widehat{X}_k^-)$$

  et vérifier s'il est "proche" d'un bruit blanc.
- Le choix du système de coordonnées dans lequel on exprime le problème influence beaucoup le comportement du filtre de Kalman étendu.

Chapitre 5

Modèles de Markov cachés

Dans cette seconde partie, on se propose d'étudier à nouveau le problème de filtrage, c'est–à–dire le problème de l'estimation d'un état inconnu au vu d'observations bruitées, dans le cas où l'état inconnu est modélisé par une chaîne de Markov à temps *discret* et espace d'état *fini*.

5.1 Chaînes de Markov à état fini

On considère un espace d'état *fini* $E = \{1, \cdots, N\}$. Une suite $\{X_k\}$ de v.a. à valeurs dans E est une chaîne de Markov si la propriété suivante est vérifiée (propriété de Markov)

$$\mathbf{P}[X_{k+1} = i_{k+1} \mid X_0 = i_0, \cdots, X_k = i_k] = \mathbf{P}[X_{k+1} = i_{k+1} \mid X_k = i_k] \ ,$$

pour tout instant k et toute suite $i_0, \cdots, i_{k+1} \in E$.

Cette notion généralise la notion de système dynamique déterministe (machine à état fini, suite récurrente, ou équation différentielle ordinaire) : la distribution de probabilité de l'état futur X_{k+1} ne dépend que de l'état présent X_k.

Remarque 5.1 Les entiers $\{1, \cdots, N\}$ servent uniquement à *étiqueter* les états de la chaîne. En général, il n'y a aucun sens à effectuer des opérations *algébriques* avec ces entiers, comme par exemple de calculer un état moyen.

Définition 5.2 *Une* probabilité *sur E est un vecteur $\nu = (\nu_i)$ de dimension N, vérifiant*

$$0 \leq \nu_i \leq 1 \ , \qquad \textit{pour tout } i \in E, \textit{ et} \qquad \sum_{i \in E} \nu_i = 1 \ .$$

Une matrice markovienne *sur E est une matrice $\pi = (\pi_{i,j})$ de dimension $N \times N$, vérifiant*

$$0 \leq \pi_{i,j} \leq 1 \ , \qquad \textit{pour tout } i,j \in E, \textit{ et} \qquad \sum_{j \in E} \pi_{i,j} = 1 \ , \qquad \textit{pour tout } i \in E.$$

Il résulte de la Proposition 5.3 ci–dessous qu'une chaîne de Markov $\{X_k\}$ est entièrement caractérisée par la donnée

- de la *loi initiale* $\nu = (\nu_i)$

$$\nu_i = \mathbf{P}[X_0 = i] \ , \qquad \text{pour tout } i \in E,$$

- et de la *matrice de transition* $\pi = (\pi_{i,j})$

$$\pi_{i,j} = \mathbf{P}[X_{k+1} = j \mid X_k = i] \ , \qquad \text{pour tout } i,j \in E,$$

  qu'on suppose indépendante de l'instant k (chaîne de Markov *homogène*).

Il suffit donc d'une donnée locale (les probabilités de transition entre deux instants successifs) pour caractériser de façon globale une chaîne de Markov.

Proposition 5.3 *Soit ν une probabilité sur E, et π une matrice markovienne sur E. La* distribution de probabilité *de la chaîne de Markov $\{X_k\}$, de loi initiale ν et de matrice de transition π, est donnée par*

$$\mathbf{P}[X_0 = i_0, \cdots, X_k = i_k] = \nu_{i_0} \, \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i_k} \ ,$$

pour tout instant k, et tout $i_0, \cdots, i_k \in E$.

Preuve. On conditionne par l'évènement $\{X_0 = i_0, \cdots, X_{k-1} = i_{k-1}\}$ et on applique la propriété de Markov

$$\begin{aligned} &\mathbf{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k] = \\ &= \mathbf{P}[X_k = i_k \mid X_0 = i_0, \cdots, X_{k-1} = i_{k-1}] \, \mathbf{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}] \\ &= \mathbf{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}] \, \pi_{i_{k-1},i_k} \ . \end{aligned}$$

En itérant cette relation, on obtient le résultat annoncé. □

5.2 Modèles de Markov cachés

On considère ensuite le cas des modèles de Markov *cachés*, ou chaînes de Markov partiellement observées. Dans ce modèle, on n'observe pas directement la suite $\{X_k\}$, mais on dispose d'observations $\{Y_k\}$ à valeurs dans un espace *fini* $O = \{1, \cdots, M\}$, ou dans $\mathbf{R}^d$. On suppose que les observations sont recueillies à travers un canal *sans mémoire*, c'est–à–dire que conditionnellement aux états $\{X_k\}$, les observations $\{Y_k\}$ sont mutuellement indépendantes, et que chaque observation Y_k ne dépend que de l'état X_k au même instant. Cette propriété s'exprime de la façon suivante :

- dans le cas *fini*

$$\mathbf{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n \mid X_0 = i_0, \cdots, X_n = i_n] = \prod_{k=0}^{n} \mathbf{P}[Y_k = \ell_k \mid X_k = i_k] \;,$$

  pour tout $i_0, \cdots, i_n \in E$, et tout $\ell_0, \cdots, \ell_n \in O$,
- et dans le cas *continu*

$$\mathbf{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n \mid X_0 = i_0, \cdots, X_n = i_n] = \prod_{k=0}^{n} \mathbf{P}[Y_k \in dy_k \mid X_k = i_k] \;,$$

  pour tout $i_0, \cdots, i_n \in E$, et tout $y_0, \cdots, y_n \in \mathbf{R}^d$.

Exemple 5.4 Supposons que les observations $\{Y_k\}$ soient reliées aux états $\{X_k\}$ de la façon suivante

$$Y_k = h(X_k) + V_k \;,$$

où la suite $\{V_k\}$ est un bruit blanc gaussien de dimension d, de moyenne nulle et de matrice de covariance R, indépendant de la chaîne de Markov $\{X_k\}$.

La fonction h définie sur E à valeurs dans $\mathbf{R}^d$ est caractérisée par la donnée d'une famille $h = (h_i)$ de N vecteurs de $\mathbf{R}^d$, et on a

$$\mathbf{P}[Y_k \in dy \mid X_k = i] = \frac{1}{(\sqrt{2\pi})^d \sqrt{\det R}} \; \exp\{\, -\tfrac{1}{2}\, (y - h_i)^* \, R^{-1} \, (y - h_i) \,\} \; dy \;.$$

Conditionnellement à $\{X_0 = i_0, \cdots, X_n = i_n\}$, les vecteurs aléatoires $Y_0, \cdots, Y_n$ sont mutuellement indépendants, et chaque Y_k est un vecteur aléatoire gaussien de dimension d, de moyenne h_{i_k} et de matrice de covariance R, de sorte que la propriété de canal sans mémoire est vérifiée.

Définition 5.5 *Une* matrice markovienne *sur $E \times O$ est une matrice $b = (b_i^\ell)$ de dimension $N \times M$, vérifiant*

$$0 \leq b_i^\ell \leq 1 \;, \qquad \textit{pour tout } i \in E \textit{ et tout } \ell \in O, \textit{ et}$$

$$\sum_{\ell \in O} b_i^\ell = 1 \;, \qquad \textit{pour tout } i \in E.$$

Un noyau markovien *sur $E \times \mathbf{R}^d$ est une famille $\psi = (\psi_i)$ de N fonctions définies sur $\mathbf{R}^d$, vérifiant*

$$\psi_i(y) \geq 0 \;, \qquad \textit{pour tout } i \in E \textit{ et tout } y \in \mathbf{R}^d, \textit{ et}$$

$$\int_{\mathbf{R}^d} \psi_i(y) \, dy = 1 \;, \qquad \textit{pour tout } i \in E.$$

Il résulte de la Proposition 5.6 ci–dessous qu'un modèle de Markov caché $\{(X_k,Y_k)\}$ est entièrement caractérisé par la donnée

- de la *loi initiale* $\nu = (\nu_i)$

$$\nu_i = \mathbf{P}[X_0 = i] \ , \qquad \text{pour tout } i \in E,$$

- de la *matrice de transition* $\pi = (\pi_{i,j})$

$$\pi_{i,j} = \mathbf{P}[X_{k+1} = j \mid X_k = i] \ , \qquad \text{pour tout } i,j \in E,$$

- et dans le cas *fini*, des *probabilités d'observation* $b = (b_i^\ell)$

$$b_i^\ell = \mathbf{P}[Y_k = \ell \mid X_k = i] \ , \qquad \text{pour tout } i \in E, \text{ et tout } \ell \in O,$$

- ou dans le cas *continu*, des *densités d'observation* $\psi = (\psi_i)$

$$\psi_i(y)\, dy = \mathbf{P}[Y_k \in dy \mid X_k = i] \ , \qquad \text{pour tout } i \in E, \text{ et tout } y \in \mathbf{R}^d.$$

Il suffit donc d'une donnée locale (les probabilités de transition entre deux instants successifs, et les probabilités/densités d'observation à un instant donné) pour caractériser de façon globale un modèle de Markov caché.

Proposition 5.6 *Dans le cas* fini, *la* distribution de probabilité *du modèle de Markov caché $\{(X_k,Y_k)\}$, de loi initiale ν, de matrice de transition π, et de probabilités d'observation b, est donnée par*

$$\mathbf{P}[X_0 = i_0, \cdots, X_k = i_k, Y_0 = \ell_0, \cdots, Y_k = \ell_k] =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i_k}\ b_{i_0}^{\ell_0} \cdots b_{i_k}^{\ell_k} \ ,$$

pour tout instant k, tout $i_0, \cdots, i_k \in E$, et tout $\ell_0, \cdots, \ell_k \in O$.

Dans le cas continu, *la* distribution de probabilité *du modèle de Markov caché $\{(X_k,Y_k)\}$, de loi initiale ν, de matrice de transition π, et de densités d'observation ψ, est donnée par*

$$\mathbf{P}[X_0 = i_0, \cdots, X_k = i_k, Y_0 \in dy_0, \cdots, Y_k \in dy_k] =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i_k}\ \psi_{i_0}(y_0) \cdots \psi_{i_k}(y_k)\, dy_0 \cdots dy_k \ ,$$

pour tout instant k, tout $i_0, \cdots, i_k \in E$, et tout $y_0, \cdots, y_k \in \mathbf{R}^d$.

Preuve. On considère d'abord le cas *fini*. On utilise la formule de Bayes, et la propriété de canal sans mémoire

$$\mathbf{P}[X_0 = i_0, \cdots ,X_k = i_k, Y_0 = \ell_0, \cdots ,Y_k = \ell_k] =$$

$$= \mathbf{P}[Y_0 = \ell_0, \cdots ,Y_k = \ell_k \mid X_0 = i_0, \cdots ,X_k = i_k] \, \mathbf{P}[X_0 = i_0, \cdots ,X_k = i_k]$$

$$= \mathbf{P}[X_0 = i_0, \cdots ,X_k = i_k] \, b_{i_0}^{\ell_0} \cdots b_{i_k}^{\ell_k} \ ,$$

et on conclut en utilisant la Proposition 5.3.

Dans le cas *continu*, on procède de la même manière

$$\mathbf{P}[X_0 = i_0, \cdots ,X_k = i_k, Y_0 \in dy_0, \cdots ,Y_k \in dy_k] =$$

$$= \mathbf{P}[Y_0 \in dy_0, \cdots ,Y_k \in dy_k \mid X_0 = i_0, \cdots ,X_k = i_k] \, \mathbf{P}[X_0 = i_0, \cdots ,X_k = i_k]$$

$$= \mathbf{P}[X_0 = i_0, \cdots ,X_k = i_k] \, \psi_{i_0}(y_0) \cdots \psi_{i_k}(y_k) \, dy_0 \cdots dy_k \ ,$$

et on conclut de la même manière, en utilisant la Proposition 5.3. □

On désigne par $\boldsymbol{M} = (\nu,\pi,b)$ dans le cas *fini*, et par $\boldsymbol{M} = (\nu,\pi,\psi)$ dans le cas *continu*, les paramètres caractéristiques du modèle.

On s'intéresse aux deux problèmes suivants :

- **Evaluer** le modèle $\boldsymbol{M}$: Il s'agit de calculer *efficacement* la distribution de probabilité de la suite d'observations $(Y_0, \cdots ,Y_n)$ (ou *fonction de vraisemblance*) en fonction des paramètres du modèle. La réponse à ce problème est fournie par l'équation *forward* de Baum.
- **Estimer** l'état de la chaîne : Etant donnée une suite d'observations $(Y_0, \cdots ,Y_n)$, il s'agit d'estimer de façon récursive l'etat présent X_n (problème de *filtrage*), ou bien d'estimer un état intermédiaire X_k pour $k = 0, \cdots ,n$ (problème de *lissage*), ou encore d'estimer globalement la suite d'états $(X_0, \cdots ,X_n)$, pour un modèle donné $\boldsymbol{M}$. La réponse aux deux premiers problèmes est fournie par les équations *forward* et *backward* de Baum, qui permettent de calculer la distribution de probabilité conditionnelle de l'état X_k sachant les observations $(Y_0, \cdots ,Y_n)$. La réponse au dernier problème est fournie par un algorithme de *programmation dynamique*, l'algorithme de Viterbi, qui permet de maximiser la distribution de probabilité conditionnelle de la suite d'états $(X_0,X_1, \cdots ,X_n)$.

Chapitre 6

Equations forward / backward de Baum

On commence par présenter une première méthode pour calculer la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$.

Proposition 6.1 *La distribution de probabilité des observations* $(Y_0, \cdots, Y_n)$ *est donnée :*

- *dans le cas* fini *par*

$$\mathbf{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n] = \sum_{i_0, \cdots, i_n \in E} \nu_{i_0} \, \pi_{i_0, i_1} \cdots \pi_{i_{n-1}, i_n} \, b_{i_0}^{\ell_0} \cdots b_{i_n}^{\ell_n} \; ,$$

pour tout $\ell_0, \cdots, \ell_n \in O$,

- *et dans le cas* continu *par*

$$\mathbf{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n] =$$

$$= \sum_{i_0, \cdots, i_n \in E} \nu_{i_0} \, \pi_{i_0, i_1} \cdots \pi_{i_{n-1}, i_n} \, \psi_{i_0}(y_0) \cdots \psi_{i_n}(y_n) \, dy_0 \cdots dy_n \; ,$$

pour tout $y_0, \cdots, y_n \in \mathbf{R}^d$.

PREUVE. On considère d'abord le cas *fini*. On utilise la Proposition 5.6 pour calculer la distribution de probabilité marginale

$$\mathbf{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n] =$$

$$\begin{aligned} &= \sum_{i_0, \cdots, i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 = \ell_0, \cdots, Y_n = \ell_n] \\ &= \sum_{i_0, \cdots, i_n \in E} \nu_{i_0} \, \pi_{i_0, i_1} \cdots \pi_{i_{n-1}, i_n} \, b_{i_0}^{\ell_0} \cdots b_{i_n}^{\ell_n} \; . \end{aligned}$$

Dans le cas *continu*, on procède de la même manière

$$\mathbf{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n] =$$

$$= \sum_{i_0, \cdots, i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n]$$

$$= \sum_{i_0, \cdots, i_n \in E} \nu_{i_0} \; \pi_{i_0, i_1} \cdots \pi_{i_{n-1}, i_n} \; \psi_{i_0}(y_0) \cdots \psi_{i_n}(y_n) \, dy_0 \cdots dy_n \; . \qquad \Box$$

Remarque 6.2 Le nombre d'opérations nécessaires pour calculer la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$ à partir des formules données dans la Proposition 6.1 est considérable : pour chaque trajectoire possible $(i_0, \cdots, i_n)$ de la chaîne de Markov, il faut effectuer le produit de $2(n+1)$ termes, et il y a N^{n+1} trajectoires possibles différentes. Le nombre total d'opérations élémentaires (additions et multiplications) à effectuer est donc de l'ordre de : $2(n+1)\; N^{n+1}$. Ce nombre croît de façon *exponentielle* avec le nombre n d'observations.

6.1 Equation forward

Pour tout instant k, la distribution de probabilité jointe des observations passées $(Y_0, \cdots, Y_k)$ et de l'état présent X_k est définie :

- dans le cas *fini* par

$$\mathbf{P}[Y_0 = \ell_0, \cdots, Y_k = \ell_k, X_k = i] = \alpha_i^k[\ell_0, \cdots, \ell_k] \; ,$$

  pour tout $i \in E$, et tout $\ell_0, \cdots, \ell_k \in O$,
- et dans le cas *continu* par

$$\mathbf{P}[Y_0 \in dy_0, \cdots, Y_k \in dy_k, X_k = i] = \alpha_i^k[y_0, \cdots, y_k] \; dy_0 \cdots dy_k \; ,$$

  pour tout $i \in E$, et tout $y_0, \cdots, y_k \in \mathbf{R}^d$.

En particulier à l'instant initial $k = 0$, on a :

- dans le cas *fini*

$$\alpha_i^0[\ell] \; = \; \mathbf{P}[Y_0 = \ell, X_0 = i]$$

$$= \; \mathbf{P}[Y_0 = \ell \mid X_0 = i] \; \mathbf{P}[X_0 = i] = \nu_i \; b_i^\ell \; ,$$

  pour tout $i \in E$, et tout $\ell \in O$,

- et dans le cas *continu*

$$\alpha_i^0[y]\,dy = \mathbf{P}[Y_0 \in dy, X_0 = i]$$

$$= \mathbf{P}[Y_0 \in dy \mid X_0 = i]\;\mathbf{P}[X_0 = i] = \nu_i\;\psi_i(y)\;dy\;,$$

 pour tout $i \in E$, et tout $y \in \mathbf{R}^d$.

On définit la variable *forward* $p^k = (p_i^k)$ par

$$p_i^k = \alpha_i^k[Y_0, \cdots, Y_k]\;,$$

pour tout $i \in E$.

Notations Pour tout $\ell \in O$, on définit la matrice diagonale $N \times N$

$$B^\ell = \mathrm{diag}(b_1^\ell, \cdots, b_N^\ell)\;.$$

De même, pour tout $y \in \mathbf{R}^d$, on définit la matrice diagonale $N \times N$

$$\Psi(y) = \mathrm{diag}(\psi_1(y), \cdots, \psi_N(y))\;.$$

Remarque 6.3 Pour tout $i \in E$, la fonction b_i définie sur O à valeurs dans $\mathbf{R}$, est équivalente à la donnée du vecteur $b_i = (b_i^\ell)$ de dimension M. De même, la fonction B définie sur O et à valeurs dans l'espace des matrices $N \times N$ diagonales, est équivalente à la donnée de la famille $B = (B^\ell)$ de M matrices diagonales $N \times N$.

Théorème 6.4 *La suite $\{p^k\}$ vérifie l'équation récurrente suivante :*

- *dans le cas* fini

$$p_j^{k+1} = b_j(Y_{k+1})\;\Big[\sum_{i \in E} \pi_{i,j}\;p_i^k\Big]\;, \tag{6.1}$$

 pour tout $j \in E$, avec la condition initiale

$$p_i^0 = \nu_i\;b_i(Y_0)\;, \qquad \textit{pour tout } i \in E,$$

 ou sous forme vectorielle

$$p^{k+1} = B(Y_{k+1})\;\pi^*\;p^k\;, \qquad p^0 = B(Y_0)\;\nu\;,$$

- *et dans le cas* continu

$$p_j^{k+1} = \psi_j(Y_{k+1})\;\Big[\sum_{i \in E} \pi_{i,j}\;p_i^k\Big]\;, \tag{6.2}$$

 pour tout $j \in E$, avec la condition initiale

$$p_i^0 = \nu_i\;\psi_i(Y_0)\;, \qquad \textit{pour tout } i \in E,$$

 ou sous forme vectorielle

$$p^{k+1} = \Psi(Y_{k+1})\;\pi^*\;p^k\;, \qquad p^0 = \Psi(Y_0)\;\nu\;.$$

Preuve. On considère uniquement le cas *fini*. Par définition

$$\alpha_i^k[\ell_0,\cdots,\ell_k] = \mathbf{P}[Y_0 = \ell_0,\cdots,Y_k = \ell_k, X_k = i] =$$

$$= \sum_{i_0,\cdots,i_{k-1}\in E} \mathbf{P}[Y_0 = \ell_0,\cdots,Y_k = \ell_k, X_0 = i_0,\cdots,X_{k-1} = i_{k-1}, X_k = i]$$

$$= \sum_{i_0,\cdots,i_{k-1}\in E} \nu_{i_0}\ \pi_{i_0,i_1}\cdots\pi_{i_{k-1},i}\ b_{i_0}^{\ell_0}\cdots b_{i_{k-1}}^{\ell_{k-1}}\ b_i^{\ell_k}\ ,$$

pour tout $i \in E$, et tout $\ell_0,\cdots,\ell_k \in O$. De même

$$\mathbf{P}[Y_0 = \ell_0,\cdots,Y_k = \ell_k, Y_{k+1} = \ell_{k+1}, X_k = i, X_{k+1} = j] =$$

$$= \sum_{i_0,\cdots,i_{k-1}\in E} \mathbf{P}[Y_0 = \ell_0,\cdots,Y_k = \ell_k, Y_{k+1} = \ell_{k+1},$$

$$X_0 = i_0,\cdots,X_{k-1} = i_{k-1}, X_k = i, X_{k+1} = j]$$

$$= \sum_{i_0,\cdots,i_{k-1}\in E} \nu_{i_0}\ \pi_{i_0,i_1}\cdots\pi_{i_{k-1},i}\ \pi_{i,j}\ b_{i_0}^{\ell_0}\cdots b_{i_{k-1}}^{\ell_{k-1}}\ b_i^{\ell_k}\ b_j^{\ell_{k+1}}$$

$$= b_j(\ell_{k+1})\ \pi_{i,j}\ \alpha_i^k[\ell_0,\cdots,\ell_k]\ ,$$

pour tout $i,j \in E$, et tout $\ell_0,\cdots,\ell_k,\ell_{k+1} \in O$. En sommant pour tout $i \in E$, on obtient

$$\alpha_j^{k+1}[\ell_0,\cdots,\ell_{k+1}] = b_j(\ell_{k+1}) \sum_{i\in E} \pi_{i,j}\ \alpha_i^k[\ell_0,\cdots,\ell_k]\ ,$$

d'où le résultat. □

Remarque 6.5 La distribution de probabilité des observations $(Y_0,\cdots,Y_n)$ peut se calculer de la façon suivante :

- dans le cas *fini*

$$\mathbf{P}[Y_0 = \ell_0,\cdots,Y_n = \ell_n] = \sum_{i\in E} \alpha_i^n[\ell_0,\cdots,\ell_n]\ ,$$

  pour tout $\ell_0,\cdots,\ell_n \in O$,
- et dans le cas *continu*

$$\mathbf{P}[Y_0 \in dy_0,\cdots,Y_n \in dy_n] = \{\ \sum_{i\in E} \alpha_i^n[y_0,\cdots,y_n]\ \}\ dy_0\cdots dy_n\ ,$$

  pour tout $y_0,\cdots,y_n \in \mathbf{R}^d$.

Remarque 6.6 La variable forward permet de calculer, de façon récursive, la distribution de probabilité conditionnelle de l'état présent X_n sachant les observations $(Y_0, \cdots, Y_n)$:

$$\mathbf{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{p_i^n}{\displaystyle\sum_{j \in E} p_j^n} \;,$$

pour tout $i \in E$. En ce sens, p^n est une distribution de probabilité non–normalisée.

Remarque 6.7 Le calcul récursif de la variable forward p^n fait seulement intervenir des produits matrice / vecteur, et permet de calculer plus efficacement la distribution de probabilité des observations $(Y_0, \cdots, Y_n)$. Il suffit de $N\,(2N+1)$ opérations élémentaires (additions et multiplications) pour passer de l'instant k à l'instant $(k+1)$. Le nombre total d'opérations élémentaires à effectuer est donc de l'ordre de : $n\;N\,(2N+1) + (N-1)$. Ce nombre croît de façon *linéaire* avec le nombre n d'observations.

6.2 Equation backward

Soit n un instant final *fixé*. Pour tout instant k antérieur à n, la distribution de probabilité jointe des observations à venir $(Y_{k+1}, \cdots, Y_n)$ et de l'état présent X_k est définie :

- dans le cas *fini* par

$$\mathbf{P}[Y_{k+1} = \ell_{k+1}, \cdots, Y_n = \ell_n \mid X_k = i] = \beta_i^k[\ell_{k+1}, \cdots, \ell_n] \;,$$

  pour tout $i \in E$, et tout $\ell_{k+1}, \cdots, \ell_n \in O$,
- et dans le cas *continu* par

$$\mathbf{P}[Y_{k+1} \in dy_{k+1}, \cdots, Y_n \in dy_n \mid X_k = i] = \beta_i^k[y_{k+1}, \cdots, y_n]\; dy_{k+1} \cdots dy_n \;,$$

  pour tout $i \in E$, et tout $y_{k+1}, \cdots, y_n \in \mathbf{R}^d$.

La définition ci–dessus n'a pas de sens à l'instant final $k = n$. En revanche, pour $k = n-1$ on a :

- dans le cas *fini*

$$\begin{aligned}
\beta_i^{n-1}[\ell] &= \mathbf{P}[Y_n = \ell \mid X_{n-1} = i] \\
&= \sum_{j \in E} \mathbf{P}[Y_n = \ell, X_n = j \mid X_{n-1} = i] \\
&= \sum_{j \in E} \mathbf{P}[Y_n = \ell \mid X_n = j]\; \mathbf{P}[X_n = j \mid X_{n-1} = i] = \sum_{j \in E} \pi_{i,j}\; b_j^\ell \;,
\end{aligned}$$

  pour tout $i \in E$, et tout $\ell \in O$,

- et dans le cas *continu*

$$\begin{aligned}\beta_i^{n-1}[y]\,dy \;&=\; \mathbf{P}[Y_n \in dy \mid X_{n-1} = i] \\ &= \sum_{j \in E} \mathbf{P}[Y_n \in dy, X_n = j \mid X_{n-1} = i] \\ &= \sum_{j \in E} \mathbf{P}[Y_n \in dy \mid X_n = j]\; \mathbf{P}[X_n = j \mid X_{n-1} = i] = \sum_{j \in E} \pi_{i,j}\; \psi_j(y)\,dy\ ,\end{aligned}$$

pour tout $i \in E$, et tout $y \in \mathbf{R}^d$.

On définit la variable *backward* $v^k = (v_i^k)$ par

$$v_i^k = \beta_i^k[Y_{k+1}, \cdots, Y_n]\ ,$$

pour tout $i \in E$.

Remarque 6.8 Conditionnellement à $\{X_k = i\}$, la suite $X_{k+1}, X_{k+2}, \cdots$ est une chaîne de Markov, de loi initiale $\pi_{i,\bullet}$ (ligne i de la matrice π) — c'est–à–dire que

$$\mathbf{P}[X_{k+1} = j \mid X_k = i] = \pi_{i,j}\ , \qquad \text{pour tout } j \in E,$$

et de matrice de transition π.

Théorème 6.9 *La suite $\{v^k\}$ vérifie l'équation récurrente rétrograde suivante :*

- *dans le cas* fini

$$v_i^k = \sum_{j \in E} \pi_{i,j}\; b_j(Y_{k+1})\; v_j^{k+1}\ , \tag{6.3}$$

pour tout $i \in E$, avec la condition initiale

$$v_i^n = 1\ , \qquad \textit{pour tout } i \in E,$$

ou sous forme vectorielle

$$v^k = \pi\; B(Y_{k+1})\; v^{k+1}\ , \qquad v^n \equiv 1\ ,$$

- *et dans le cas* continu

$$v_i^k = \sum_{j \in E} \pi_{i,j}\; \psi_j(Y_{k+1})\; v_j^{k+1}\ , \tag{6.4}$$

pour tout $i \in E$, avec la condition initiale

$$v_i^n = 1\ , \qquad \textit{pour tout } i \in E,$$

ou sous forme vectorielle

$$v^k = \pi\; \Psi(Y_{k+1})\; v^{k+1}\ , \qquad v^n \equiv 1\ .$$

Preuve. On considère uniquement le cas *fini*. Avec l'initialisation proposée à l'instant final, l'équation (6.3) permet de retrouver à l'instant $k = n-1$

$$v_i^{n-1} = \sum_{j\in E} \pi_{i,j}\ b_j(Y_n) = \beta_i^{n-1}[Y_n]\ .$$

D'autre part, il résulte de la Remarque 6.8 et de la Proposition 5.6 que

$$\beta_i^k[\ell_{k+1},\cdots,\ell_n] = \mathbf{P}[Y_{k+1}=\ell_{k+1},\cdots,Y_n=\ell_n \mid X_k = i] =$$

$$= \sum_{i_{k+1},\cdots,i_n\in E} \mathbf{P}[Y_{k+1}=\ell_{k+1},\cdots,Y_n=\ell_n,X_{k+1}=i_{k+1},\cdots,X_n=i_n \mid X_k=i]$$
$$= \sum_{i_{k+1},\cdots,i_n\in E} \pi_{i,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{\ell_{k+1}}\cdots b_{i_n}^{\ell_n}\ ,$$

pour tout $i\in E$, et tout $\ell_{k+1},\cdots,\ell_n\in O$. De même

$$\mathbf{P}[Y_{k+1}=\ell_{k+1},\cdots,Y_n=\ell_n,X_{k+1}=j \mid X_k=i] =$$

$$= \sum_{i_{k+2},\cdots,i_n\in E} \mathbf{P}[Y_{k+1}=\ell_{k+1},\cdots,Y_n=\ell_n,X_{k+1}=j,X_{k+2}=i_{k+2},\cdots,X_n=i_n \mid X_k=i]$$
$$= \sum_{i_{k+2},\cdots,i_n\in E} \pi_{i,j}\ \pi_{j,i_{k+2}}\cdots\pi_{i_{n-1},i_n}\ b_j^{\ell_{k+1}}\ b_{i_{k+2}}^{\ell_{k+2}}\cdots b_{i_n}^{\ell_n}$$
$$= \pi_{i,j}\ b_j(\ell_{k+1})\ \beta_j^{k+1}[\ell_{k+2},\cdots,\ell_n]\ ,$$

pour tout $i,j\in E$, et tout $\ell_{k+1},\cdots,\ell_n\in O$. En sommant pour tout $j\in E$, on obtient

$$\beta_i^k[\ell_{k+1},\cdots,\ell_n] = \sum_{j\in E}\pi_{i,j}\ b_j(\ell_{k+1})\ \beta_j^{k+1}[\ell_{k+2},\cdots,\ell_n]\ ,$$

d'où le résultat. □

Proposition 6.10 *Les équations forward et backward sont* duales *l'une de l'autre :*

$$\sum_{i\in E} p_i^0\ v_i^0 = \sum_{i\in E} p_i^k\ v_i^k = \sum_{i\in E} p_i^n\ v_i^n\ ,$$

pour tout instant k.

Preuve. On considère uniquement le cas *fini*. En utilisant successivement l'équation backward (6.3) et l'équation forward (6.1), on obtient

$$\begin{aligned}\sum_{i\in E} p_i^k\ v_i^k &= \sum_{i\in E} p_i^k\ \Big[\sum_{j\in E}\pi_{i,j}\ v_j^{k+1}\ b_j(Y_{k+1})\Big]\\ &= \sum_{j\in E} b_j(Y_{k+1})\ \Big[\sum_{i\in E}\pi_{i,j}\ p_i^k\Big]\ v_j^{k+1} = \sum_{j\in E} p_j^{k+1}\ v_j^{k+1}\ ,\end{aligned}$$

d'où le résultat. $\square$

Proposition 6.11 *Pour tout instant k, la distribution de probabilité jointe de l'état présent X_k et des observations $(Y_0,\cdots,Y_n)$ jusqu'à l'instant final n, est donnée :*

- *dans le cas* fini *par*
$$\mathbf{P}[Y_0=\ell_0,\cdots,Y_n=\ell_n,X_k=i]=\alpha_i^k[\ell_0,\cdots,\ell_k]\ \beta_i^k[\ell_{k+1},\cdots,\ell_n]\ ,$$
pour tout $i\in E$, et tout $\ell_0,\cdots,\ell_n\in O$,
- *et dans le cas* continu *par*
$$\mathbf{P}[Y_0\in dy_0,\cdots,Y_n\in dy_n,X_k=i]=\alpha_i^k[y_0,\cdots,y_k]\ \beta_i^k[y_{k+1},\cdots,y_n]\,dy_0\cdots dy_n\ ,$$
pour tout $i\in E$, et tout $y_0,\cdots,y_n\in\mathbf{R}^d$.

Preuve. On considère uniquement le cas *fini*. Fixer l'état à l'instant k permet d'effectuer une coupure entre le passé jusqu'à l'instant $(k-1)$ et le futur à partir de l'instant $(k+1)$, de la façon suivante :

$$\mathbf{P}[Y_0=\ell_0,\cdots,Y_n=\ell_n,X_k=i]=$$

$$=\sum_{\substack{i_0,\cdots,i_{k-1}\in E\\ i_{k+1},\cdots,i_n\in E}}\mathbf{P}[Y_0=\ell_0,\cdots,Y_n=\ell_n,\ X_0=i_0,\cdots,X_{k-1}=i_{k-1},X_k=i,X_{k+1}=i_{k+1},\cdots,X_n=i_n]$$

$$=\sum_{\substack{i_0,\cdots,i_{k-1}\in E\\ i_{k+1},\cdots,i_n\in E}}\nu_{i_0}\ \pi_{i_0,i_1}\cdots\pi_{i_{k-1},i}\ \pi_{i,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_{i_0}^{\ell_0}\cdots b_{i_{k-1}}^{\ell_{k-1}}\ b_i^{\ell_k}\ b_{i_{k+1}}^{\ell_{k+1}}\cdots b_{i_n}^{\ell_n}$$

$$=\sum_{i_0,\cdots,i_{k-1}\in E}\nu_{i_0}\ \pi_{i_0,i_1}\cdots\pi_{i_{k-1},i}\ b_{i_0}^{\ell_0}\cdots b_{i_{k-1}}^{\ell_{k-1}}\ b_i^{\ell_k}\ \Big[\sum_{i_{k+1},\cdots,i_n\in E}\pi_{i,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{\ell_{k+1}}\cdots b_{i_n}^{\ell_n}\ \Big]$$

$$=\alpha_i^k[\ell_0,\cdots,\ell_k]\ \beta_i^k[\ell_{k+1},\cdots,\ell_n]\ ,$$

d'où le résultat. $\square$

Remarque 6.12 Le produit composante–par–composante des variables forward et backward permet de calculer la distribution de probabilité conditionnelle de l'état X_k à un instant intermédiaire k, sachant les observations $(Y_0,\cdots,Y_n)$ jusqu'à l'instant final n :

$$\mathbf{P}[X_k=i\mid Y_0,\cdots,Y_n]=\frac{q_i^k}{\displaystyle\sum_{j\in E}q_j^k}\ ,$$

pour tout $i \in E$, avec la définition

$$q_i^k = p_i^k \; v_i^k \;, \qquad \text{pour tout } i \in E.$$

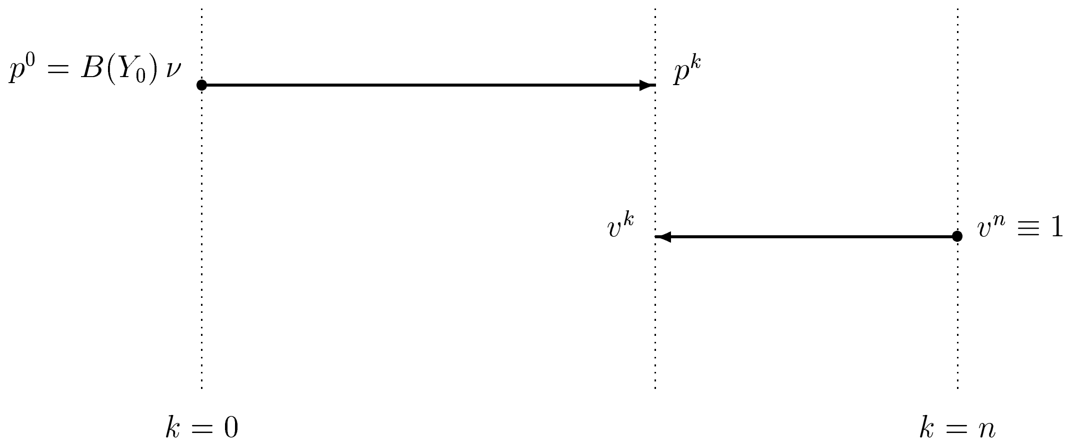


FIG. 6.1 – *Equations forward–backward*

On remarque que la constante de normalisation

$$\sum_{i \in E} q_i^k = \sum_{i \in E} p_i^k \; v_i^k = \sum_{i \in E} p_i^n = P \;,$$

ne dépend pas de l'instant k considéré.

Chapitre 7

Algorithme de Viterbi

Il résulte des Remarques 6.6 et 6.12 que les variables forward et backward étudiées au Chapitre 6 permettent de calculer la distribution de probabilité conditionnelle de l'état présent X_n, ou de l'état X_k à un instant intermédiaire, sachant les observations $(Y_0, \cdots, Y_n)$:

$$\mathbf{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{p_i^n}{\displaystyle\sum_{j \in E} p_j^n} \ ,$$

et

$$\mathbf{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{q_i^k}{\displaystyle\sum_{j \in E} q_j^k} \ ,$$

respectivement, pour tout $i \in E$.

Compte tenu que les états possibles pour la chaîne de Markov sont seulement *indexés* par les entiers $\{1, \cdots, N\}$, mais ne se prètent pas en général aux opérations *algébriques*, il n'y aurait aucun sens à utiliser ces distributions de probabilités conditionnelles pour calculer des moyennes conditionnelles. En revanche, on peut proposer l'estimateur suivant basé sur les observations $(Y_0, \cdots, Y_n)$, soit pour l'état présent

$$X_n^{\mathrm{MAP,loc}} = \operatorname*{argmax}_{i \in E} \ \mathbf{P}[X_n = i \mid Y_0, \cdots, Y_n] = \operatorname*{argmax}_{i \in E} \ p_i^n$$

soit pour l'état à un instant intermédiaire

$$X_k^{\mathrm{MAP,loc}} = \operatorname*{argmax}_{i \in E} \ \mathbf{P}[X_k = i \mid Y_0, \cdots, Y_n] = \operatorname*{argmax}_{i \in E} \ q_i^k$$

(en supposant que dans chacun des cas le maximum est atteint en un point unique). Cet estimateur est appelé *estimateur local du maximum a posteriori.*

Cependant, il peut arriver que la suite $(X_0^{\mathrm{MAP,loc}}, \cdots, X_n^{\mathrm{MAP,loc}})$ ainsi générée soit incohérente avec le modèle, dans le sens suivant : il peut arriver que l'on obtienne $X_k^{\mathrm{MAP,loc}} =$

i et $X_{k+1}^{\text{MAP, loc}} = j$ pour deux instants successifs, alors que $\pi_{i,j} = 0$ pour cette même paire (i,j), ce qui signifie que la transition de l'état i vers l'état j est *impossible* pour le modèle. Pour cette raison, on utilise plutôt un autre estimateur, appelé *estimateur global du maximum a posteriori*, ou simplement *estimateur du maximum a posteriori*, défini par

$$(X_0^{\text{MAP}}, \cdots, X_n^{\text{MAP}}) = \operatorname*{argmax}_{i_0,\cdots,i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] \ .$$

Le calcul efficace de cet estimateur est fourni par un algorithme de programmation dynamique, appelé *algorithme de Viterbi.*

Programmation dynamique

D'après la formule de Bayes, la trajectoire qui maximise la distribution de probabilité conditionnelle de $(X_0, \cdots, X_n)$ sachant les observations $(Y_0, \cdots, Y_n)$ maximise également la distribution de probabilité jointe de $(X_0, \cdots, X_n)$ et des observations $(Y_0, \cdots, Y_n)$, c'est–à–dire que :

- dans le cas *fini*

$$\operatorname*{argmax}_{i_0,\cdots,i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 = \ell_0, \cdots, Y_n = \ell_n] =$$

$$= \operatorname*{argmax}_{i_0,\cdots,i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0 = \ell_0, \cdots, Y_n = \ell_n] \ ,$$

  pour tout $\ell_0, \cdots, \ell_n \in O$,
- et dans le cas *continu*

$$\operatorname*{argmax}_{i_0,\cdots,i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n] =$$

$$= \Big\{ \operatorname*{argmax}_{i_0,\cdots,i_n \in E} \mathbf{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0 = y_0, \cdots, Y_n = y_n] \Big\} \, dy_0 \cdots dy_n \ ,$$

  pour tout $y_0, \cdots, y_n \in \mathbf{R}^d$.

Pour tout instant k, on définit :

- dans le cas *fini*

$$\max_{i_0,\cdots,i_{k-1}} \mathbf{P}[X_0 = i_0, \cdots, X_k = i_{k-1}, X_k = i, Y_0 = \ell_0, \cdots, Y_k = \ell_k] = \delta_i^k[\ell_0, \cdots, \ell_k] \ ,$$

  pour tout $i \in E$, et tout $\ell_0, \cdots, \ell_k \in O$,
- et dans le cas *continu*

$$\max_{i_0,\cdots,i_{k-1}} \mathbf{P}[X_0 = i_0, \cdots, X_k = i_{k-1}, X_k = i, Y_0 \in dy_0, \cdots, Y_k \in dy_k] =$$

$$= \delta_i^k[y_0, \cdots, y_k] \, dy_0 \cdots dy_k \ ,$$

  pour tout $i \in E$, et tout $y_0, \cdots, y_k \in \mathbf{R}^d$.

La fonction *valeur* $V^k = (V_i^k)$ est définie par

$$V_i^k = \delta_i^k[Y_0, \cdots, Y_k] \ ,$$

pour tout $i \in I$.

Théorème 7.1 *La suite $\{V^k\}$ vérifie la récurrence suivante :*

- *dans le cas* fini
$$V_j^{k+1} = b_j(Y_{k+1}) \ \big[\max_{i \in E} \pi_{i,j} \ V_i^k \ \big] \ , \tag{7.1}$$
pour tout $j \in E$, avec la condition initiale
$$V_i^0 = \nu_i \ b_i(Y_0) \ , \qquad \textit{pour tout } i \in E,$$

- *et dans le cas* continu
$$V_j^{k+1} = \psi_j(Y_{k+1}) \ \big[\max_{i \in E} \pi_{i,j} \ V_i^k \ \big] \ , \tag{7.2}$$
pour tout $j \in E$, avec la condition initiale
$$V_i^0 = \nu_i \ \psi_i(Y_0) \ , \qquad \textit{pour tout } i \in E.$$

A chaque instant k, on définit pour tout $j \in E$ l'indice

$$I_k(j) = \operatorname*{argmax}_{i \in E} \ [\pi_{i,j} \ V_i^k]$$

(en supposant que le maximum est atteint en un point unique).

Preuve. On considère uniquement le cas *fini*. Il résulte de la Proposition 5.6 que

$$\begin{aligned} &\mathbf{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, X_{k+1} = j, \\ &\qquad\qquad Y_0 = \ell_0, \cdots, Y_k = \ell_k, Y_{k+1} = \ell_{k+1}] = \\ &= \nu_{i_0} \ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i} \ \pi_{i,j} \ b_{i_0}^{\ell_0} \cdots b_i^{\ell_k} \ b_j^{\ell_{k+1}} \\ &= b_j(\ell_{k+1}) \ \pi_{i,j} \ \mathbf{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, Y_0 = \ell_0, \cdots, Y_k = \ell_k] \ , \end{aligned}$$

pour tout $i,j \in E$, tout $i_0, \cdots, i_{k-1} \in E$, et tout $\ell_0, \cdots, \ell_k, \ell_{k+1} \in O$. On en déduit que

$$\begin{aligned} &\max_{i_0, \cdots, i_{k-1} \in E} \mathbf{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, X_{k+1} = j, \\ &\qquad\qquad Y_0 = \ell_0, \cdots, Y_k = \ell_k, Y_{k+1} = \ell_{k+1}] = \\ &= b_j(\ell_{k+1}) \ \pi_{i,j} \ \delta_i^k[\ell_0, \cdots, \ell_k] \ , \end{aligned}$$

pour tout $i,j \in E$, et tout $\ell_0,\cdots,\ell_k,\ell_{k+1} \in O$. En maximisant par rapport à $i \in E$, on obtient

$$\delta_j^{k+1}[\ell_0,\cdots,\ell_{k+1}] =$$

$$= \max_{i_0,\cdots,i_{k-1},i\in E} \mathbf{P}[X_0 = i_0,\cdots,X_{k-1} = i_{k-1},X_k = i,X_{k+1} = j,$$

$$Y_0 = \ell_0,\cdots,Y_k = \ell_k,Y_{k+1} = \ell_{k+1}]$$

$$= \max_{i\in E} \max_{i_0,\cdots,i_{k-1}\in E} \mathbf{P}[X_0 = i_0,\cdots,X_{k-1} = i_{k-1},X_k = i,X_{k+1} = j,$$

$$Y_0 = \ell_0,\cdots,Y_k = \ell_k,Y_{k+1} = \ell_{k+1}]$$

$$= b_j(\ell_k) \ \left[\max_{i\in E} \pi_{i,j} \ \delta_i^k[\ell_0,\cdots,\ell_k] \ \right] \ ,$$

d'où le résultat. □

Remarque 7.2 Parmi toutes les trajectoires qui aboutissent dans l'état j à l'instant $(k+1)$, la trajectoire de plus grande probabilité est passé dans l'état

$$I_k(j) = \operatorname*{argmax}_{i\in E} \, [\pi_{i,j} \ V_i^k] \ ,$$

à l'instant précédent k (en supposant que le maximum est atteint en un point unique). En outre, on a nécessairement

$$\pi_{I_k(j),j} > 0 \ ,$$

ce qui garantit que la transition de l'état $I_k(j)$ vers l'état j est possible pour le modèle.

La trajectoire optimale est alors calculée de la façon suivante : On remarque d'abord que

- dans le cas *fini*

$$\max_{i_0,\cdots,i_n\in E} \mathbf{P}[X_0 = i_0,\cdots,X_n = i_n,Y_0 = \ell_0,\cdots,Y_n = \ell_n] =$$

$$= \max_{i\in E} \max_{i_0,\cdots,i_{n-1}\in E} \mathbf{P}[X_0 = i_0,\cdots,X_{n-1} = i_{n-1},X_n = i,Y_0 = \ell_0,\cdots,Y_n = \ell_n]$$

$$= \max_{i\in E} \ \delta_i^n[\ell_0,\cdots,\ell_n] \ ,$$

- et dans le cas *continu*

$$\max_{i_0,\cdots,i_n\in E} \mathbf{P}[X_0 = i_0,\cdots,X_n = i_n,Y_0 \in dy_0,\cdots,Y_n \in dy_n] =$$

$$= \left\{ \max_{i\in E} \ \delta_i^n[y_0,\cdots,y_n] \ \right\} \ dy_0 \cdots dy_n \ .$$

On en déduit que la trajectoire optimale aboutit dans l'état

$$X_n^{\mathrm{MAP}} = \operatorname*{argmax}_{i \in E} \; V_i^n$$

(en supposant que le maximum est atteint en un point unique), à l'instant final. De proche en proche, on en déduit que la trajectoire optimale est passée dans l'état

$$X_k^{\mathrm{MAP}} = I_k(X_{k+1}^{\mathrm{MAP}}) \ ,$$

à l'instant k.

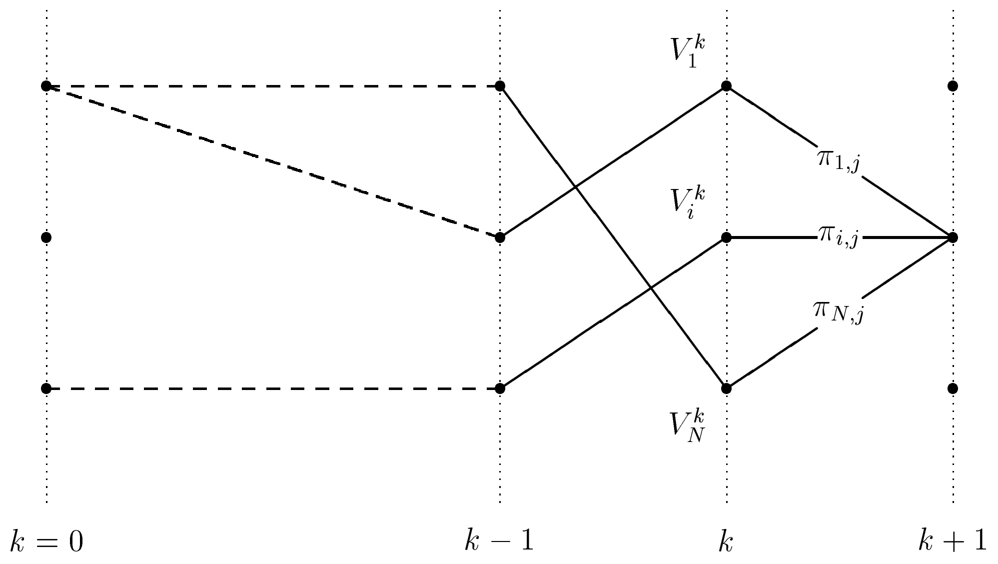


FIG. 7.1 – *Algorithme de Viterbi (programmation dynamique)*

Annexe A

Rappels de probabilités

L'objectif de la théorie des probabilités est l'étude des phénomènes aléatoires. La caractéristique d'une expérience aléatoire est que le comportement quantitatif ou qualitatif de grandeurs tentant de décrire le phénomène en question, ne peut pas être complètement prédit au vu des conditions expérimentales, mais dépend aussi du hasard.

Pour modéliser une expérience aléatoire, on se donne

- un ensemble Ω décrivant toutes les issues possibles de l'expérience, les *réalisations*,
- une collection $\mathcal{F}$ d'*évènements* possibles, qui sont des parties de Ω,
- une application $\mathbf{P}$ qui à chaque évènement A associe la *probabilité* que celui–ci se réalise.

L'évaluation des probabilités résulte

- soit d'une formulation *a priori*,
- soit de l'expérimentation statistique : on réalise un grand nombre d'expériences et on évalue le rapport N_A/N, où N_A désigne le nombre d'expériences qui ont vu l'évènement A se réaliser, et N désigne le nombre total d'expériences,
- soit du *calcul* : on utilise alors des axiomes, consistants avec la notion intuitive et expérimentale de probabilité.

Espace de probabilités

Un triplet $(\Omega,\mathcal{F},\mathbf{P})$ est appelé *espace de probabilités* si

- Ω est un ensemble de *réalisations*,
- $\mathcal{F}$ est un ensemble, appelé *tribu*, de parties de Ω, sont appelées *évènements*, vérifiant
  (i) $\Omega \in \mathcal{F}$.
  (ii) si $A \in \mathcal{F}$, alors $A^c \in \mathcal{F}$ (où par définition $A^c = \Omega \setminus A$),

(iii) si $A_n \in \mathcal{F}$ pour tout $n \in \mathbf{N}$, alors $\bigcup_{n\in\mathbf{N}} A_n \in \mathcal{F}$.

- $\mathbf{P}$ est une application, appelée mesure de probabilité (ou *probabilité*), définie sur la tribu $\mathcal{F}$ et vérifiant

  (iv) pour tout $A \in \mathcal{F}$, $P(A) \geq 0$,

  (v) $P(\Omega) = 1$,

  (vi) si $A_n \in \mathcal{F}$ pour tout $n \in \mathbf{N}$, et $A_n \cap A_m = \emptyset$ pour tout $n \neq m$, alors

$$\mathbf{P}(\bigcup_{n\in\mathbf{N}} A_n) = \sum_{n\in\mathbf{N}} \mathbf{P}(A_n) \ .$$

A partir des axiomes, on peut montrer les propriétés suivantes

(vii) pour tout $A \in \mathcal{F}$, $0 \leq \mathbf{P}(A) \leq 1$,

(viii) pour tout $A \in \mathcal{F}$, $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$,

(ix) si $A_n \in \mathcal{F}$ pour tout $n \in \mathbf{N}$, alors

$$\mathbf{P}(\bigcup_{n\in\mathbf{N}} A_n) \leq \sum_{n\in\mathbf{N}} \mathbf{P}(A_n) \ .$$

Si $\mathcal{F}_0 \subset \mathcal{F}$, on appelle tribu engendrée par $\mathcal{F}_0$ la plus petite tribu contenant tous les éléments de $\mathcal{F}_0$. Par exemple, si $\Omega = \mathbf{R}$ et $\mathcal{F}_0$ désigne l'ensemble des *intervalles ouverts* de $\mathbf{R}$, on appelle tribu *borélienne* la tribu $\mathcal{B}$ engendrée par $\mathcal{F}_0$. De même, si $\Omega = \mathbf{R}^n$ et $\mathcal{F}_0$ désigne l'ensemble des *parties ouvertes* de $\mathbf{R}^n$, on appelle tribu *borélienne* la tribu $\mathcal{B}^n$ engendrée par $\mathcal{F}_0$.

Variables aléatoires

On appelle *variable aléatoire réelle* sur $(\Omega,\mathcal{F})$, une application X définie sur Ω, à valeurs dans $\mathbf{R}$, telle que pour tout $B \in \mathcal{B}$

$$\{\omega \,:\, X(\omega) \in B\} \in \mathcal{F} \ ,$$

où $\mathcal{B}$ est la tribu borélienne sur $\mathbf{R}$.

On appelle *vecteur aléatoire* de dimension n sur $(\Omega,\mathcal{F})$, une application X définie sur Ω, à valeurs dans $\mathbf{R}^n$, telle que pour tout $B \in \mathcal{B}^n$

$$\{\omega \,:\, X(\omega) \in B\} \in \mathcal{F} \ ,$$

où $\mathcal{B}^n$ est la tribu borélienne sur $\mathbf{R}^n$.

Plus généralement, on appelle *variable aléatoire* sur $(\Omega,\mathcal{F})$ à valeurs dans un espace probabilisable $(E,\mathcal{E})$ (on dit également *application mesurable* de $(\Omega,\mathcal{F})$ dans $(E,\mathcal{E})$), une application X définie sur Ω, à valeurs dans E, telle que pour tout $B \in \mathcal{E}$

$$\{\omega \,:\, X(\omega) \in B\} \in \mathcal{F} \ .$$

Pour tout $B \in \mathcal{E}$, on utilise les notations suivantes

$$\{X \in B\} \triangleq \{\omega : X(\omega) \in B\} ,$$

et

$$\mathbf{P}(X \in B) \triangleq \mathbf{P}(\{X \in B\}) .$$

On vérifie que l'application μ_X définie sur la tribu $\mathcal{E}$ par la relation

$$\mu_X(B) \triangleq \mathbf{P}(X \in B) ,$$

pour tout $B \in \mathcal{E}$, est une mesure de probabilité sur $(E,\mathcal{E})$, appelée *loi* de X (on dit également *distribution de probabilité* de X).

Densité, densité jointe, densités marginales

Soit X un vecteur aléatoire de dimension n sur $(\Omega,\mathcal{F},\mathbf{P})$. S'il existe une fonction p_X définie sur $\mathbf{R}^n$, telle que pour tout $B \in \mathcal{B}^n$

$$\mathbf{P}(X \in B) = \mu_X(B) = \int_B p_X(x)\, dx ,$$

on dit que la loi de X est *absolument continue*, et que p_X est la *densité* de X (on dit également *densité de probabilité* de X).

Exemple A.1 [densité gaussienne] On appelle variable aléatoire gaussienne réelle, de moyenne μ et de variance σ^2, une variable aléatoire réelle dont la densité est définie par

$$p_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\{ - \frac{(x-\mu)^2}{2\sigma^2} \} .$$

Soit X (resp. Y) un vecteur aléatoire de dimension n (resp. de dimension p) sur $(\Omega,\mathcal{F},\mathbf{P})$. S'il existe une fonction $p_{X,Y}$ définie sur $\mathbf{R}^{n+p}$, telle que pour tout $B \in \mathcal{B}^{n+p}$

$$\mathbf{P}[(X,Y) \in B] = \int_B p_{X,Y}(x,y)\, dx\, dy ,$$

on dit que $p_{X,Y}$ est la *densité jointe* de X et Y.

On remarque que les *densités marginales* de $p_{X,Y}$, définies respectivement par

$$p_X(x) \triangleq \int_{\mathbf{R}^p} p_{X,Y}(x,y)\, dy , \quad \text{et} \quad p_Y(y) \triangleq \int_{\mathbf{R}^n} p_{X,Y}(x,y)\, dx ,$$

coïncident avec les densités de X et de Y. En effet, pour tout $B \in \mathcal{B}^n$

$$\begin{aligned} \mathbf{P}(X \in B) &= \mathbf{P}[(X,Y) \in B \times \mathbf{R}^p] \\ &= \int_{B \times \mathbf{R}^p} p_{X,Y}(x,y)\, dx\, dy = \int_B \Big\{ \int_{\mathbf{R}^p} p_{X,Y}(x,y)\, dy \Big\}\, dx \;, \end{aligned}$$

et de même pour tout $B \in \mathcal{B}^p$

$$\begin{aligned} \mathbf{P}(Y \in B) &= \mathbf{P}[(X,Y) \in \mathbf{R}^n \times B] \\ &= \int_{\mathbf{R}^n \times B} p_{X,Y}(x,y)\, dx\, dy = \int_B \Big\{ \int_{\mathbf{R}^n} p_{X,Y}(x,y)\, dx \Big\}\, dy \;. \end{aligned}$$

Moyenne, covariance

L'*espérance mathématique* (ou la *moyenne*) de la variable aléatoire X, notée $\mathbf{E}[X]$, est définie par

$$\mathbf{E}[X] \stackrel{\triangle}{=} \int_{\mathbf{R}^n} x\, p_X(x)\, dx \;.$$

Si $Y = g(X)$ est une fonction (mesurable) réelle de la variable aléatoire X, alors Y a pour espérance

$$\mathbf{E}[Y] = \mathbf{E}[g(X)] = \int_{\mathbf{R}} g(x)\, p_X(x)\, dx \;.$$

La *matrice de covariance* (ou simplement la *variance* dans le cas réel) est définie par

$$\mathrm{cov}(X) \stackrel{\triangle}{=} \mathbf{E}[(X - \bar{X})\; (X - \bar{X})^*] = \int_{\mathbf{R}^n} (x - \bar{X})\; (x - \bar{X})^*\, p_X(x)\, dx \;,$$

avec la notation $\bar{X} = \mathbf{E}[X]$. Il s'agit d'une matrice $n \times n$ symétrique et semi–définie positive.

Exemple A.2 Soit X une variable aléatoire gaussienne réelle, de densité

$$p_X(x) = \frac{1}{\sqrt{2\pi}\;\sigma}\; \exp\Big\{ -\frac{(x-\mu)^2}{2\sigma^2} \Big\} \;.$$

On vérifie par le calcul que $\mathbf{E}[X] = \mu$ et $\mathrm{var}(X) = \sigma^2$, ce qui justifie la terminologie employée dans l'Exemple A.1 ci–dessus.

L'opérateur d'espérance mathématique ainsi défini est linéaire : soit $\alpha,\beta \in \mathbf{R}$ et X,Y deux vecteurs aléatoires de dimension n,

$$\mathbf{E}[\alpha\, X + \beta\, Y] = \alpha\, \mathbf{E}[X] + \beta\, \mathbf{E}[Y] \;.$$

En effet

$$\begin{aligned}
\mathbf{E}[\alpha\, X + \beta\, Y] &= \int_{\mathbf{R}^n \times \mathbf{R}^n} (\alpha\, x + \beta\, y)\, p_{X,Y}(x,y)\, dx\, dy \\
&= \alpha \int_{\mathbf{R}^n} x\ \{ \int_{\mathbf{R}^n} p_{X,Y}(x,y)\, dy\ \}\ dx + \beta \int_{\mathbf{R}^n} y\ \{ \int_{\mathbf{R}^n} p_{X,Y}(x,y)\, dx\ \}\ dy \\
&= \alpha \int_{\mathbf{R}^n} x\, p_X(x)\, dx + \beta \int_{\mathbf{R}^n} y\, p_Y(y)\, dy = \alpha\, \mathbf{E}[X] + \beta\, \mathbf{E}[Y]\ .
\end{aligned}$$

Probabilité conditionnelle, indépendance

Soit $A,B \in \mathcal{F}$ deux évènements. La connaissance que l'évènement B est réalisé conduit à réévaluer la probabilité de voir l'évènement A se réaliser, de la façon suivante : on définit la *probabilité conditionnelle* de l'évènement A sachant B par la formule

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}\ , \tag{A.1}$$

pourvu que $\mathbf{P}(B) > 0$.

Cette définition est conforme à l'intuition fondée sur la notion de *fréquence relative* : on réalise un grand nombre d'expériences et on évalue le rapport $N_{A\cap B}/N_B$, où N_B désigne le nombre d'expériences qui ont vu l'évènement B se réaliser, et $N_{A\cap B}$ désigne le nombre d'expériences parmi celles–ci qui ont également vu l'évènement A se réaliser, c'est–à–dire le nombre d'expériences qui ont vu l'évènement $A \cap B$ se réaliser. Si N désigne le nombre total d'expériences, on a bien $N_{A\cap B}/N_B = N_{A\cap B}/N \cdot (N_B/N)^{-1}$, ce qui justifie la définition.

A partir de la définition, on obtient la *formule de Bayes*

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A) \cdot \mathbf{P}(A)}{\mathbf{P}(B)}\ ,$$

pourvu que $\mathbf{P}(B) > 0$. On montre aussi que, si $A_1, \cdots, A_n$ est une partition de Ω, alors

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B \mid A_i) \cdot \mathbf{P}(A_i)\ ,$$

pour tout $B \in \mathcal{F}$. On en déduit

$$\mathbf{P}(A_j \mid B) = \frac{\mathbf{P}(B \mid A_j) \cdot \mathbf{P}(A_j)}{\displaystyle\sum_{i=1}^{n} \mathbf{P}(B \mid A_i) \cdot \mathbf{P}(A_i)}\ ,$$

pour tout $B \in \mathcal{F}$.

Deux évènements $A,B \in \mathcal{F}$ sont dits *indépendants*, et on note $A \perp B$, si la connaissance que l'un de ces évènements s'est réalisé n'entraîne aucune modification de la probabilité de voir l'autre évènement se réaliser, c'est–à–dire

$$\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} = \mathbf{P}(A) \ ,$$

ou de façon plus symétrique

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B) \ .$$

Des événements $A_1, \cdots, A_n \in \mathcal{F}$ sont *mutuellement indépendants* si

$$\mathbf{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbf{P}(A_{i_1}) \cdots \mathbf{P}(A_{i_k})$$

pour tout choix $1 \leq i_1 < \cdots < i_k \leq n$. Attention : on peut avoir $A \perp B$, $B \perp C$, et $A \perp C$ mais cela n'entraîne pas que A,B,C sont mutuellement indépendants.

Soit X (resp. Y) un vecteur aléatoire de dimension n (resp. de dimension p) défini sur $(\Omega,\mathcal{F},\mathbf{P})$. On dit que les vecteurs aléatoires X et Y sont *indépendants*, et on note $X \perp Y$, si pour tout $A \in \mathcal{B}^n$, $B \in \mathcal{B}^p$, les évènements $(X \in A)$ et $(Y \in B)$ sont indépendants, c'est–à–dire

$$\mathbf{P}(X \in A\, , Y \in B) = \mathbf{P}(X \in A) \cdot \mathbf{P}(Y \in B) \ .$$

Si $p_{X,Y}$ désigne la densité jointe de (X,Y), alors pour tout $A \in \mathcal{B}^n$, $B \in \mathcal{B}^p$

$$\mathbf{P}(X \in A\, , Y \in B) = \int_{A \times B} p_{X,Y}(x,y)\, dx\, dy \ ,$$

et

$$\mathbf{P}(X \in A) \cdot \mathbf{P}(Y \in B) = \int_A p_X(x)\, dx \int_B p_Y(y)\, dy = \int_{A \times B} p_X(x)\, p_Y(y)\, dx\, dy \ .$$

Il en résulte que la propriété d'indépendance est équivalente à la propriété de *factorisation* de la densité jointe : pour (presque) tout $x \in \mathbf{R}^n$, $y \in \mathbf{R}^p$

$$p_{X,Y}(x,y) = p_X(x) \cdot p_Y(y) \ .$$

Soit f (resp. g) une fonction (mesurable) réelle définie sur $\mathbf{R}^n$ (resp. sur $\mathbf{R}^p$). On a

$$\mathbf{E}[f(X)\, g(Y)] = \int_{\mathbf{R}^n \times \mathbf{R}^p} f(x)\, g(y)\, p_{X,Y}(x,y)\, dx\, dy \ ,$$

et

$$\begin{aligned} \mathbf{E}[f(X)] \cdot \mathbf{E}[g(Y)] \ &= \ \Big\{ \int_{\mathbf{R}^n} f(x)\, p_X(x)\, dx \Big\} \Big\{ \int_{\mathbf{R}^p} g(y)\, p_Y(y)\, dy \Big\} \\ &= \int_{\mathbf{R}^n \times \mathbf{R}^p} f(x)\, g(y)\, p_X(x)\, p_Y(y)\, dx\, dy \ . \end{aligned}$$

On obtient ainsi un autre critère pour vérifier l'indépendance de deux vecteurs aléatoires : les vecteurs aléatoires X et Y, de dimension n et p respectivement, sont indépendants si et seulement si

$$\mathbf{E}[f(X)\, g(Y)] = \mathbf{E}[f(X)] \cdot \mathbf{E}[g(Y)] \ ,$$

pour toute paire f, g de fonctions (mesurables) réelles définies sur $\mathbf{R}^n$ et $\mathbf{R}^p$ respectivement.

Conditionnement par $(Y = y)$

Etant donnés deux vecteurs aléatoires X et Y définis sur $(\Omega,\mathcal{F},\mathbf{P})$, de dimension n et p respectivement, qu'apporte le fait d'observer la réalisation $Y = y$ sur la connaissance que l'on a de X ?

On aimerait utiliser la formule (A.1), c'est–à–dire écrire

$$\mathbf{P}(X \in A \mid Y = y) = \frac{\mathbf{P}(X \in A, Y = y)}{\mathbf{P}(Y = y)} \ ,$$

mais en général $\mathbf{P}(Y = y) = 0$. On introduit donc la définition suivante : s'il existe une fonction (mesurable) $\psi(\cdot)$ définie sur $\mathbf{R}^p$ telle que

$$\mathbf{P}(X \in A, Y \in B) = \int_B \psi(y)\, p_Y(y)\, dy \ ,$$

pour tout $A \in \mathcal{B}^n$, $B \in \mathcal{B}^p$, on dit que $\psi(y)$ est (une version de) la probabilité conditionnelle de l'évènement $(X \in A)$ sachant $Y = y$, et on note $\mathbf{P}(X \in A \mid Y = y)$.

Remarque A.3 Si $B \in \mathcal{B}^p$, avec $y \in B$ et $\mathbf{P}(Y \in B) > 0$, alors la formule (A.1) peut être utilisée, et donne

$$\mathbf{P}(X \in A \mid Y \in B) = \frac{\mathbf{P}(X \in A, Y \in B)}{\mathbf{P}(Y \in B)} = \frac{\displaystyle\int_B \psi(z)\, p_Y(z)\, dz}{\displaystyle\int_B p_Y(z)\, dz} \longrightarrow \psi(y) \ ,$$

quand $B \downarrow \{y\}$, c'est–à–dire quand l'ensemble B décroît vers le point y, ce qui justifie intuitivement la définition donnée plus haut.

Le calcul pratique de la probabilité conditionnelle $\mathbf{P}(X \in A \mid Y = y)$ se fait de la façon suivante : soit (X,Y) un vecteur aléatoire de dimension $(n + p)$ défini sur $(\Omega,\mathcal{F},\mathbf{P})$, et soit $p_{X,Y}$ sa densité jointe. Par définition

$$\begin{aligned} \mathbf{P}(X \in A, Y \in B) &= \int_{A \times B} p_{X,Y}(x,y)\, dy\, dx \\ &= \int_B \Big\{ \int_A p_{X,Y}(x,y)\, dx \Big\}\, dy = \int_B \Big\{ \int_A \frac{p_{X,Y}(x,y)}{p_Y(y)}\, dx \Big\}\, p_Y(y)\, dy \ , \end{aligned}$$

ce qui donne l'expression suivante

$$\mathbf{P}(X \in A \mid Y = y) = \int_A \frac{p_{X,Y}(x,y)}{p_Y(y)}\, dx \ .$$

La densité de la loi conditionnelle (ou *densité conditionnelle*) du vecteur aléatoire X sachant $Y = y$, est définie par la formule

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \ .$$

Soit $\phi(\cdot)$ une fonction (mesurable) réelle définie sur $\mathbf{R}^n$. On définit la moyenne conditionnelle de la variable aléatoire réelle $\phi(X)$ sachant $Y = y$ par

$$\mathbf{E}[\phi(X) \mid Y = y] = \int_{\mathbf{R}^n} \phi(x)\ p_{X|Y=y}(x)\, dx \ .$$

Le calcul donne

$$\begin{aligned}
\mathbf{E}[\ \phi(X)\ \mathbf{1}_{(Y \,\in\, B)}\,] &= \int_{\mathbf{R}^n \times B} \phi(x)\ p_{X,Y}(x,y)\, dy\, dx \\
&= \int_B \Big\{ \int_{\mathbf{R}^n} \phi(x)\ \frac{p_{X,Y}(x,y)}{p_Y(y)}\, dx\ \Big\}\ p_Y(y)\, dy \\
&= \int_B \Big\{ \int_{\mathbf{R}^n} \phi(x)\ p_{X|Y=y}(x)\, dx\ \Big\}\ p_Y(y)\, dy \\
&= \int_B \mathbf{E}[\phi(X) \mid Y = y]\ p_Y(y)\, dy \ ,
\end{aligned} \tag{A.2}$$

pour tout $B \in \mathcal{B}^p$, ce qui fournit une autre caractérisation de la moyenne conditionnelle.

Le résultat suivant montre que la moyenne conditionnelle sachant Y peut s'interpréter comme une projection orthogonale sur la tribu engendrée par le vecteur aléatoire Y (pour le produit scalaire $\langle \xi, \eta \rangle = \mathbf{E}[\xi\, \eta]$ défini sur l'ensemble des variables aléatoires réelles de carré intégrable).

Proposition A.4 *Soit* $\widehat{\phi}(y) = \mathbf{E}[\phi(X) \mid Y = y]$*. Alors la variable aléatoire réelle* $\widehat{\phi}(Y)$*, notée aussi* $\mathbf{E}[\phi(X) \mid Y]$*, est caractérisée par*

$$\mathbf{E}[\ [\phi(X) - \widehat{\phi}(Y)]\ \psi(Y)\ ] = 0 \ ,$$

pour toute fonction (mesurable) réelle $\psi(\cdot)$ *définie sur* $\mathbf{R}^p$*.*

Preuve. Prenons $\psi(\cdot)$ de la forme $\psi(y) = \mathbf{1}_{(y \,\in\, B)}$, où $B \in \mathcal{B}^p$. Alors, d'après (A.2)

$$\begin{aligned}
\mathbf{E}[\ \widehat{\phi}(Y)\, \psi(Y)\ ] = \int_B \widehat{\phi}(y)\ p_Y(y)\, dy \ &= \int_B \mathbf{E}[\phi(X) \mid Y = y]\ p_Y(y)\, dy \\
&= \mathbf{E}[\ \phi(X)\ \mathbf{1}_{(Y \,\in\, B)}\,] = \mathbf{E}[\ \phi(X)\, \psi(Y)\ ] \ . \qquad \square
\end{aligned}$$

Une écriture équivalente est

$$\mathbf{E}[\,\mathbf{E}[\phi(X) \mid Y]\;\psi(Y)\,] = \mathbf{E}[\,\phi(X)\,\psi(Y)\,]\ ,$$

pour toute fonction (mesurable) réelle $\psi(\cdot)$ définie sur $\mathbf{R}^p$.

On obtient en particulier

$$\mathbf{E}[\,\mathbf{E}[\phi(X) \mid Y]\,] = \mathbf{E}[\phi(X)]\ ,$$

en prenant $\psi(y) \equiv 1$. D'autres conséquences de la Proposition A.4 sont listées ci–dessous.

Corollaire A.5 *(i) Si $X = f(Y)$, alors :* $\mathbf{E}[\phi(X) \mid Y] = \phi(X)$.
(ii) Si $Y \perp X$, alors : $\mathbf{E}[\phi(X) \mid Y] = \mathbf{E}[\phi(X)]$.
(iii) Si $Z \perp (X{,}Y)$, alors : $\mathbf{E}[\,\phi(X) \mid Y{,}Z\,] = \mathbf{E}[\,\phi(X) \mid Y\,]$.

Remarque A.6 La première propriété (i) exprime que lorsque X dépend explicitement de Y, l'observation de Y permet de connaître X exactement.

La seconde propriété (ii) exprime que dans la situation opposée où les vecteurs aléatoires X et Y sont indépendants, l'observation de Y n'apprend rien de nouveau sur $\phi(X)$. La dernière propriété (iii) est une généralisation de (ii).

Preuve. On utilise systématiquement la caractérisation donnée à la Proposition A.4.

Si $X = f(Y)$, alors

$$\mathbf{E}[\,\phi(X)\,\psi(Y)\,] = \mathbf{E}[\,\phi[f(Y)]\,\psi(Y)\,]\ ,$$

d'où

$$\mathbf{E}[\phi(X) \mid Y] = \phi[f(Y)] = \phi(X)\ ,$$

ce qui prouve (i).

Si $Y \perp X$, alors

$$\mathbf{E}[\,\phi(X)\,\psi(Y)\,] = \mathbf{E}[\phi(X)]\;\mathbf{E}[\psi(Y)] = \mathbf{E}[\,\mathbf{E}[\phi(X)]\,\psi(Y)\,]\ ,$$

ce qui prouve (ii).

Si $Z \perp (X{,}Y)$, alors

$$\begin{aligned}\mathbf{E}[\,\phi(X)\,\psi(Y)\,\chi(Z)\,] &= \mathbf{E}[\,\phi(X)\,\psi(Y)\,]\;\mathbf{E}[\chi(Z)]\\ &= \mathbf{E}[\,\mathbf{E}[\phi(X) \mid Y]\;\psi(Y)\,]\;\mathbf{E}[\chi(Z)]\\ &= \mathbf{E}[\,\mathbf{E}[\phi(X) \mid Y]\;\psi(Y)\,\chi(Z)\,]\ ,\end{aligned}$$

ce qui prouve (iii). □

Finalement le résultat suivant, dont la démonstration est similaire à celle de la Proposition 3.1, montre que la moyenne conditionnelle sachant Y peut également s'interpréter comme un estimateur du minimum de variance.

Proposition A.7 *La moyenne conditionnelle $\widehat{\phi}(Y) = \mathbf{E}[\phi(X) \mid Y]$ de la variable aléatoire $\phi(X)$ sachant le vecteur aléatoire Y, est l'estimateur de $\phi(X)$ construit à partir de Y qui minimise la variance de l'erreur d'estimation, c'est–à–dire que*

$$\mathbf{E}[\,|\phi(X) - \widehat{\phi}(Y)|^2\,] \leq \mathbf{E}[\,|\phi(X) - \psi(Y)|^2\,]$$

pour tout autre estimateur $\psi(\cdot)$.

Fonction caractéristique

Soit X un vecteur aléatoire de dimension n défini sur $(\Omega,\mathcal{F},\mathbf{P})$. On appelle *fonction caractéristique* de X, la transformée de Fourier de la densité p_X, définie par

$$\Phi_X(u) = \mathbf{E}[\,e^{\,i\,u^* X}\,] = \int_{\mathbf{R}^n} e^{\,i\,u^* x}\,p_X(x)\,dx \ ,$$

pour tout $u \in \mathbf{R}^n$. Grace à la formule d'inversion, la donnée de la densité p_X est équiavelente à la donnée de la fonction caractéristique Φ_X.

Exemple A.8 Soit X une variable aléatoire gaussienne réelle, de moyenne μ et de variance σ^2. On vérifie que

$$\Phi_X(u) = \exp\left\{\ i\,u\mu - \tfrac{1}{2}\,\sigma^2 u^2\ \right\}\ .$$

Si les composantes $(X_1,\cdots,X_n)$ du vecteur aléatoire $X = (X_1,\cdots,X_n)$ sont mutuellement indépendantes, alors

$$\Phi_X(u) = \Phi_{X_1}(u_1)\cdots\Phi_{X_n}(u_n)\ ,$$

pour tout $u = (u_1,\cdots,u_n)$, ce qui fournit un nouveau critère pour vérifier l'indépendance mutuelle de vecteurs aléatoires.

Proposition A.9 *Soit X un vecteur aléatoire de dimension n défini sur $(\Omega,\mathcal{F},\mathbf{P})$. Soit A une application linéaire de $\mathbf{R}^n$ dans $\mathbf{R}^p$, c'est–à–dire une matrice $p \times n$, et soit b un vecteur de $\mathbf{R}^p$. On définit $Y = A\,X + b$, et on vérifie qu'il s'agit d'un vecteur aléatoire de dimension p, dont la fonction caractéristique vérifie*

$$\Phi_Y(u) = e^{\,i\,u^* b}\,\Phi_X(A^* u)\ ,$$

pour tout $u \in \mathbf{R}^p$.

Preuve. Par définition

$$\Phi_Y(u) \;=\; \mathbf{E}[\, e^{\,i\,u^* Y}\,] = \mathbf{E}[\, e^{\,i\,u^*(A\,X+b)}\,]$$

$$= e^{\,i\,u^* b}\,\mathbf{E}[\, e^{\,i\,u^*\,A\,X}\,] = e^{\,i\,u^* b}\,\mathbf{E}[\, e^{\,i\,(A^*\,u)^* X}\,] = e^{\,i\,u^* b}\,\Phi_X(A^* u)\ ,$$

pour tout $u \in \mathbf{R}^p$. □

Vecteurs aléatoires gaussiens

Soit X un vecteur aléatoire de dimension n défini sur $(\Omega,\mathcal{F},\mathbf{P})$. On dit que X est un vecteur aléatoire *gaussien* si toute combinaison linéaire des composantes du vecteur X est une variable aléatoire gaussienne réelle, c'est–à–dire si, pour tout $u \in \mathbf{R}^n$, la variable aléatoire réelle $u^* X$ est gaussienne.

Proposition A.10 *Soit X un vecteur aléatoire gaussien de dimension n, de moyenne μ et de matrice de covariance Q. Sa fonction caractéristique vérifie*

$$\Phi_X(u) = \exp\{\; i\,u^*\mu - \tfrac{1}{2}\,u^* Q u \;\}\ ,$$

pour tout $u \in \mathbf{R}^n$.

Preuve. Comme la variable aléatoire réelle $u^* X$ est gaussienne, sa loi est complètement caractérisée par sa moyenne

$$\mathbf{E}[u^* X] = u^*\,\mathbf{E}[X] = u^*\mu\ ,$$

et sa variance

$$\mathbf{E}[\,(u^*(X-\mu))^2\,] = \mathbf{E}[\,u^*(X-\mu)(X-\mu)^* u\,] = u^* Q u\ ,$$

qui définissent respectivement une *forme linéaire* et une *forme quadratique symétrique semi–définie positive* sur $\mathbf{R}^n$. La fonction caractéristique de la variable aléatoire gaussienne réelle $u^* X$ vérifie donc, d'après le résultat donné à l'Exemple A.8

$$\Phi_{u^* X}(\lambda) = \mathbf{E}[\, e^{\,i\,\lambda\,u^* X}\,] = \exp\{\; i\,\lambda\,u^*\mu - \tfrac{1}{2}\,\lambda^2\,u^* Q u \;\} = \Phi_X(\lambda u)\ ,$$

pour tout réel λ. En faisant $\lambda = 1$, on vérifie que la fonction caractéristique du vecteur aléatoire gaussien X vérifie

$$\Phi_X(u) = \exp\{\; i\,u^*\mu - \tfrac{1}{2}\,u^* Q u \;\}\ ,$$

pour tout $u \in \mathbf{R}^n$. □

Remarque A.11 Par définition, les composantes d'un vecteur aléatoire gaussien sont des variables aléatoires gaussiennes. Mais un vecteur aléatoire dont les composantes sont des variables aléatoires gaussiennes n'est pas nécessairement gaussien.

On énonce le résultat suivant, sans démonstration.

Proposition A.12 *Soit X un vecteur aléatoire gaussien de dimension n, de moyenne μ et de matrice de covariance Q. Si la matrice Q est* non–dégénérée *(inversible), alors la loi de X possède une densité p_X, qui vérifie*

$$p_X(x) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det Q}} \; \exp\{ -\tfrac{1}{2}(x-\mu)^* \, Q^{-1} \, (x-\mu) \} \; .$$

Proposition A.13 *Soit X un vecteur aléatoire gaussien de dimension n, de moyenne μ et de matrice de covariance Q. Soit A une application linéaire de $\mathbf{R}^n$ dans $\mathbf{R}^p$, c'est–à–dire une matrice $p \times n$, et soit b un vecteur de $\mathbf{R}^p$. Alors, le vecteur aléatoire $Y = A\,X + b$ est gaussien, de moyenne $A\,\mu + b$ et de matrice de covariance AQA^*.*

Preuve. Il suffit de montrer le caractère gaussien. En combinant les Propositions A.9 et A.10, on obtient

$$\begin{aligned}\Phi_Y(u) = e^{\,i\,u^* b}\,\Phi_X(A^* u) \;&=\; e^{\,i\,u^* b}\;\exp\{\, i\,(A^*u)^*\mu - \tfrac{1}{2}(A^*u)^*Q(A^*u) \,\} \\ &=\; \exp\{\, i\,u^*(A\mu + b) - \tfrac{1}{2}\,u^*(AQA^*)u \,\} \;,\end{aligned}$$

pour tout $u \in \mathbf{R}^p$. □

Le résultat suivant montre que deux composantes d'un vecteur aléatoire gaussien sont indépendantes, si et seulement si ces composantes sont non–corrélées (ou orthogonales).

Proposition A.14 *Soit (X,Y) un vecteur aléatoire gaussien de dimension $(n+p)$. Alors $X \perp Y$ si et seulement si*

$$Q_{X,Y} = \mathbf{E}[(X-\mu_X)\,(Y-\mu_Y)^*] = 0 \; .$$

Preuve. Si $X \perp Y$, alors il est évident que

$$Q_{X,Y} = \mathbf{E}[\,(X-\mu_X)(Y-\mu_Y)^*\,] = \mathbf{E}[X-\mu_X]\;\mathbf{E}[Y-\mu_Y]^* = 0 \; .$$

indépendamment du caractère gaussien.

Réciproquement, pour tout $u \in \mathbf{R}^n$, $v \in \mathbf{R}^p$

$$\begin{aligned}\Phi_{X,Y}(u,v) \;&=\; \exp\{\, i\,(\,u^*\;\; v^*\,)\begin{pmatrix}\mu_X \\ \mu_Y\end{pmatrix} - \tfrac{1}{2}\,(\,u^*\;\; v^*\,)\begin{pmatrix} Q_X & Q_{X,Y} \\ Q_{Y,X} & Q_Y\end{pmatrix}\begin{pmatrix}u \\ v\end{pmatrix} \} \\ &=\; \exp\{\, i\,u^*\mu_X + i\,v^*\mu_Y - \tfrac{1}{2}\,u^*Q_X u - u^*Q_{X,Y}v - \tfrac{1}{2}\,v^*Q_Y v \,\} \\ &=\; \exp\{\, i\,u^*\mu_X - \tfrac{1}{2}\,u^*Q_X u \,\}\;\exp\{\, i\,v^*\mu_Y - \tfrac{1}{2}\,v^*Q_Y v \,\}\;\exp\{\, -\,u^*Q_{X,Y}v \,\} \\ &=\; \Phi_X(u)\;\Phi_Y(v)\;\exp\{\, -\,u^*Q_{X,Y}v \,\} \; .\end{aligned}$$

Si $Q_{X,Y} = 0$, alors la fonction caractéristique se factorise : pour tout $u \in \mathbf{R}^n$, $v \in \mathbf{R}^p$

$$\Phi_{X,Y}(u,v) = \Phi_X(u)\ \Phi_Y(v)\ ,$$

c'est–à–dire que $X \perp Y$. □

Soit X et Y deux vecteurs aléatoires de dimension n et p respectivement. D'après la Proposition A.4, l'espérance conditionnelle de X sachant Y, notée $\widehat{X} = \mathbf{E}[X \mid Y]$ est la projection orthogonale du vecteur aléatoire X sur la tribu $\mathcal{Y}$ engendrée par le vecteur aléatoire Y.

Soit $X^\perp$ la projection orthogonale du vecteur aléatoire X sur l'espace vectoriel $\mathcal{H}$ engendré par les constantes et par les composantes du vecteur aléatoire Y. Evidemment $\mathcal{H} \subset \mathcal{Y}$, de sorte que

$$\mathbf{E}[\,|X - X^\perp|^2\,] \geq \mathbf{E}[\,|X - \widehat{X}|^2\,]\ .$$

Le résultat suivant montre que les deux projections coïncident dans le cas particulier des vecteurs aléatoires gaussiens.

Proposition A.15 *Soit (X,Y) un vecteur aléatoire gaussien de dimension $(n+p)$, et soit $X^\perp$ la projection orthogonale du vecteur aléatoire X sur l'espace vectoriel $\mathcal{H}$ engendré par les constantes et par les composantes du vecteur aléatoire Y. On a alors*

$$X^\perp = \mathbf{E}[X \mid Y]\ .$$

Preuve. Par définition

$$X^\perp = \alpha + A\,Y\ ,$$

où α est un vecteur de $\mathbf{R}^n$ et A est une matrice $n \times p$, et chaque composante du vecteur aléatoire $(X - X^\perp)$ est orthogonale à la constante 1, et à chacune des composantes du vecteur aléatoire Y, ce qui peut se traduire par les relations

$$\mathbf{E}[X - X^\perp] = 0\ , \tag{A.3}$$

$$\mathbf{E}[\,(X - X^\perp)\,Y^*\,] = 0\ . \tag{A.4}$$

D'autre part, le vecteur aléatoire $(X - X^\perp,Y)$ est un vecteur aléatoire gaussien de dimension $(n+p)$: en effet, pour tout $u \in \mathbf{R}^n$, $v \in \mathbf{R}^p$

$$u^*\,(X - X^\perp) + v^*\,Y = u^*\,(X - \alpha - A\,Y) + v^*\,Y = u^*\,X + (v - A^*\,u)^*\,Y\ .$$

D'après la Proposition A.14 ci–dessus, la propriété d'orthogonalité (A.4) entraîne l'indépendance des vecteurs aléatoires $(X - X^\perp)$ et Y. En utilisant (A.3), on obtient

$$\mathbf{E}[\,(X - X^\perp)\,\psi(Y)\,] = \mathbf{E}[X - X^\perp]\ \mathbf{E}[\psi(Y)] = 0\ ,$$

pour toute fonction (mesurable) réelle $\psi(\cdot)$ définie sur $\mathbf{R}^p$. Il suffit alors d'appliquer la Proposition A.4 pour conclure. □