Université de Rennes 1
Master Recherche STI

Introduction au Filtrage en Temps Discret

Filtre de Kalman
Filtrage Particulaire
Modèles de Markov Cachés

François LeGland
IRISA / INRIA

2005–06

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.

Exemple : Navigation d'un véhicule sous–marin autonome On considère le problème de la navigation (c'est–à–dire de la détermination de la position et si possible de la vitesse à chaque instant) d'un véhicule d'exploration sous–marin autonome.

Dans un environnement structuré, on peut utiliser la reconnaissance en temps réel de points caractéristiques dont la position est disponible dans une base de données, sur une carte, etc. On considère ici le cas où l'environnement n'est pas structuré, et où le système de navigation utilise un réseau d'antennes à base longue, et éventuellement un capteur d'immersion. La configuration du réseau consiste en une unité hydrophone / projecteur à bord du véhicule, et un ensemble de quatre transpondeurs sous–marins déposés au fond avant le début de la mission, et dont les positions sont supposées connues. Le projecteur du véhicule interroge les quatre transpondeurs, chacun desquels émet une impulsion acoustique dès qu'il reçoit l'impulsion d'interrogation du véhicule. La durée de la transmission aller–retour entre le véhicule et un transpondeur donné fournit une mesure de la distance entre le véhicule et ce transpondeur. Le capteur d'immersion mesure la hauteur de la colonne d'eau au dessus du véhicule.

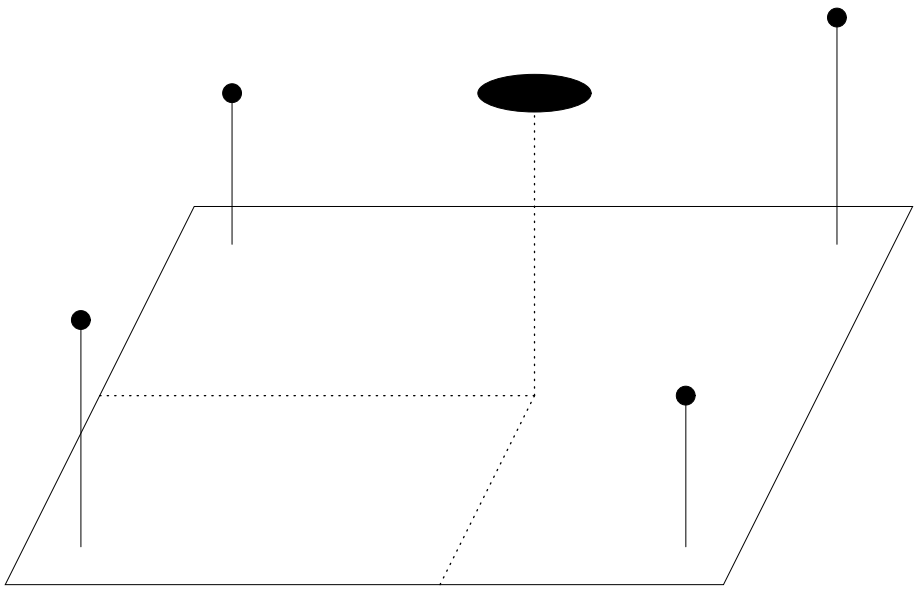

FIG. 1.1 – Navigation d'un véhicule sous–marin autonome

À partir des quatre mesures de distance, la position du véhicule est déterminé de façon unique (pourvu que les quatre transpondeurs ne soient pas situés dans un même plan) par l'intersection de quatre sphères, centrées chacune à la position d'un transpondeur différent, et de rayon égal à la distance entre le véhicule et ce transpondeur.

Dans la pratique, ces mesures de distance sont entachées d'erreur, et il peut se produire que le problème d'intersection n'ait pas de solution. Il peut aussi arriver qu'un ou plusieurs des transpondeurs soit défaillant, c'est–à–dire ne fournisse pas de mesure pendant quelque temps, voire de façon définitive, ce qui rend la triangulation impossible. Enfin il peut arriver que les mesures de distance fournies par les transpondeurs ne puissent pas être considérées comme synchrones, c'est–à–dire que les distances mesurées entre le véhicule et une paire de transpondeurs correspondent à deux dates légèrement différentes : le véhicule s'étant déplacé entre ces deux dates, la procédure de triangulation elle–même est entachée d'erreur.

Ces différents problèmes (erreurs de mesure, défaillance des capteurs, asynchronisme, etc.) sont résolus en introduisant un modèle a priori pour l'évolution du véhicule.

1.1 Importance de l'information a priori

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 $x \in \mathbb{R}^m$ 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

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

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

$$H^* \sum_{k=1}^{n} Y_k = n\, H^*\, H\, x$$

$$\Longrightarrow \quad H^* \,(\frac{1}{n} \sum_{k=1}^{n} Y_k - H\, x) = 0$$

$$\Longrightarrow \quad H\, H^* \,(\frac{1}{n} \sum_{k=1}^{n} Y_k - H\, x) = 0$$

$$\Longrightarrow \quad H\, x = \frac{1}{n} \sum_{k=1}^{n} Y_k \ .$$

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

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

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

$$\mathcal{I}(x^0) = \left\{ \ x \in \mathbb{R}^m \ : \ H\, x = H\, x^0 \ \right\} \ .$$

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

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

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

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

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

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

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 $x \in \mathbb{R}^m$, les points x et $P_H\,x$ appartiennent au même espace affine de dimension $(m-d)$ de $\mathbb{R}^m$. 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 $\lambda \in \mathbb{R}^d$: 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 $\ker H \oplus \mathcal{R}(\Sigma\,H^*) = \mathbb{R}^m$. 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.

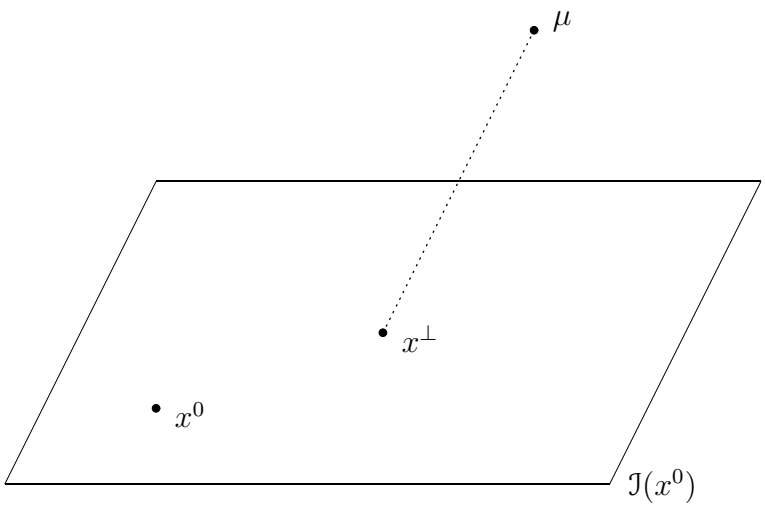

FIG. 1.2 – Prise en compte de l'information a priori

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

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

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 $u \in \mathbb{R}^d$ 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 $u = 0$. □

1.2 Prise en compte de l'information a priori

Dans de nombreux cas, la prise en compte de l'information a priori peut se ramener au problème statique 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

$$y \in \mathbb{R}^d \longmapsto \psi(y) \in \mathbb{R}^m \ .$$

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*)

$$\mathbb{E}[\,|X - \psi(Y)|^2\,] \ . \tag{1.2}$$

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

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

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

La Proposition 1.4 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

$$p_{X|Y=y}(x) = \frac{p_{X,Y}(x,y)}{\displaystyle\int_{\mathbb{R}^m} p_{X,Y}(x,y)\,dx} = \frac{p_{X,Y}(x,y)}{p_Y(y)}\ , \tag{1.3}$$

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

Proposition 1.4 *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.*

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

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

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

et on remarque que

$$\begin{aligned}&\mathbb{E}[\,(\widehat{X}(Y) - \psi(Y))^*\,(X - \widehat{X}(Y))\,] = \\ &\quad= \int_{\mathbb{R}^m}\int_{\mathbb{R}^d} (\widehat{X}(y) - \psi(y))^*\,(x - \widehat{X}(y))\; p_{X,Y}(x,y)\,dx\,dy \\ &\quad= \int_{\mathbb{R}^d} (\widehat{X}(y) - \psi(y))^*\,\Big\{\int_{\mathbb{R}^m} (x - \widehat{X}(y))\; p_{X|Y=y}(x)\,dx\Big\}\; p_Y(y)\,dy = 0\ ,\end{aligned}$$

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

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

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

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

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

Cas gaussien

Dans le cas particulier des vecteurs aléatoires gaussiens, le résultat général obtenu ci–dessus peut être précisé de la façon suivante.

Proposition 1.6 *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

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

Remarque 1.7 On vérifie aisément 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. En outre, la matrice R ne dépend pas de y, et peut donc être calculée avant même de disposer de la valeur prise par l'observation Y.

Remarque 1.8 Soit $\widehat{X} = \widehat{X}(Y)$ l'estimateur du minimum de variance de X sachant Y. Compte tenu que

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

dépend de façon affine du vecteur aléatoire Y, on en déduit que $(X, \widehat{X}, Y)$ est un vecteur aléatoire gaussien, comme transformation affine du vecteur aléatoire gaussien $Z = (X, Y)$.

PREUVE. On donne une première démonstration, dans le cas où la matrice Q_Z est inversible. Dans ce cas, les lois des vecteurs aléatoires gaussiens Y et Z ont chacune une densité, et par définition

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

On utilise la formule suivante (simple à vérifier)

$$\begin{pmatrix} I & -Q_{XY}\,Q_Y^{-1} \\ 0 & I \end{pmatrix}\ Q_Z\ \begin{pmatrix} I & 0 \\ -Q_Y^{-1}\,Q_{XY}^* & I \end{pmatrix} = \begin{pmatrix} Q_X - Q_{XY}\,Q_Y^{-1}\,Q_{YX} & 0 \\ 0 & Q_Y \end{pmatrix}\ . \tag{1.4}$$

En prenant le déterminant dans (1.4), on obtient

$$\det Q_Z = \det R\ \det Q_Y\ .$$

L'identité (1.4) 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

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

Compte tenu que

$$\begin{pmatrix} I & -Q_{XY}\,Q_Y^{-1} \\ 0 & I \end{pmatrix}\ \begin{pmatrix} x-\bar{X} \\ y-\bar{Y} \end{pmatrix} = \begin{pmatrix} (x-\bar{X}) - Q_{XY}\,Q_Y^{-1}\ (y-\bar{Y}) \\ y-\bar{Y} \end{pmatrix} = \begin{pmatrix} x-\widehat{X}(y) \\ y-\bar{Y} \end{pmatrix}\ ,$$

on obtient

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

et

$$p_{X|Y=y}(x) = \frac{1}{(\sqrt{2\pi})^{m}\,\sqrt{\det R}}\ \exp\left\{\ -\tfrac{1}{2}\,(x-\widehat{X}(y))^*\,R^{-1}\,(x-\widehat{X}(y))\ \right\}\ ,$$

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

$$\exp\left\{\ i\,u^*\,\widehat{X}(y) - \tfrac{1}{2}\,u^*Ru\ \right\}\ ,$$

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

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

et on vérifie d'autre part que

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

Par injectivité de la transformé de Fourier, on obtient

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

Conclusion

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 espace d'état fini,

et dans chacun de ces deux cas, il sera possible de résoudre exactement 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.

Ces deux cas peuvent être vus comme des cas particuliers de modèles beaucoup plus généraux :

- les chaînes de Markov à espace d'état quelconque (fini, dénombrable, continu, hybride, etc.),

et dans ce cas il ne sera pas possible de résoudre exactement le problème de filtrage de façon optimale, qui s'exprime pourtant très simplement en termes de flots de Feynman–Kac, et il faudra avoir recours à la mise en œuvre de méthodes de résolution approchées, en l'occurrence :

- de filtres particulaires, c'est–à–dire de méthodes de Monte Carlo avec interaction.

Chapitre 2

Systèmes linéaires gaussiens

On appelle *processus aléatoire* en temps discret une famille $\{X_k\,,\, k \in \mathbb{N}\}$ de vecteurs aléatoires (notée $\{X_k\}$) définis sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ à valeurs dans $\mathbb{R}^m$. Un *processus aléatoire gaussien* est un processus aléatoire $\{X_k\}$ tel que pour tout $n \in \mathbb{N}$ le vecteur $(X_0, \cdots, X_n)$ est un vecteur aléatoire gaussien (de dimension $(n+1) \times m$). Deux processus aléatoires $\{X_k\}$ et $\{X'_k\}$ sont *indépendants* si pour tout $n, n' \in \mathbb{N}$, les vecteurs aléatoires $(X_0, \cdots, X_n)$ et $(X'_0, \cdots, X'_n)$ sont indépendants. Un *bruit blanc* est un processus aléatoire $\{X_k\}$ tel que

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

2.1 Équations d'état

On considère le système dynamique suivant

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

où $\{X_k\}$ et $\{W_k\}$ prennent respectivement leurs valeurs dans $\mathbb{R}^m$ et $\mathbb{R}^p$. On fait les hypothèses suivantes sur les coefficients : $F_k \in \mathbb{R}^{m\times m}$, $f_k \in \mathbb{R}^m$, $G_k \in \mathbb{R}^{m\times p}$, pour tout $k \in \mathbb{N}$. 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 $\{W_k\}$ 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 $\mathbb{R}^m$. En particulier, X_k est gaussien, de moyenne $\bar{X}_k$ et de matrice de covariance Q_k^X, avec*

$$\begin{aligned} \bar{X}_k &= F_k\, \bar{X}_{k-1} + f_k\ , \\ Q_k^X &= F_k\, Q_{k-1}^X\, F_k^* + G_k\, Q_k^W\, G_k^*\ . \end{aligned}$$

PREUVE. Comme sortie d'un système linéaire à entrées gaussiennes, $\{X_k\}$ est un processus gaussien. En effet, pour tout $n \in \mathbb{N}$, il existe une matrice $A \in \mathbb{R}^{((n+1)\,m)\times(m+(n+1)\,p)}$ et un vecteur $b \in \mathbb{R}^{(n+1)\,m}$ tels que

$$\begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_n \end{pmatrix} = A \begin{pmatrix} X_0 \\ W_1 \\ \vdots \\ W_n \end{pmatrix} + b \ .$$

D'après les hypothèses $(X_0, W_1, \cdots, W_n)$ est un vecteur gaussien, donc $(X_0, X_1, \cdots, X_n)$ est un vecteur aléatoire gaussien (comme transformation affine d'un vecteur aléatoire gaussien).

Par ailleurs, d'après (2.1)

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

$$\begin{aligned} Q_k^X &= \mathbb{E}[(X_k - \bar{X}_k)\,(X_k - \bar{X}_k)^*] \\ &= \mathbb{E}[(F_k\,(X_{k-1} - \bar{X}_{k-1}) + G_k\, W_k)\,(F_k\,(X_{k-1} - \bar{X}_{k-1}) + G_k\, W_k)^*] \\ &= F_k\, \mathbb{E}[(X_{k-1} - \bar{X}_{k-1})\,(X_{k-1} - \bar{X}_{k-1})^*]\, F_k^* + G_k\, \mathbb{E}[W_k\,(X_{k-1} - \bar{X}_{k-1})^*]\, F_k^* \\ &\quad + F_k\, \mathbb{E}[(X_{k-1} - \bar{X}_{k-1})\, W_k^*]\, G_k^* + G_k\, \mathbb{E}[W_k\, W_k^*]\, G_k^* \\ &= F_k\, Q_{k-1}^X\, F_k^* + G_k\, Q_k^W\, G_k^* \ . \end{aligned}$$

Dans cette dernière égalité, on a utilisé le fait que $(X_{k-1} - \bar{X}_{k-1})$ est indépendant de W_k, donc $\mathbb{E}[(X_{k-1} - \bar{X}_{k-1})\, W_k^*] = 0$. □

2.2 Équations d'état et d'observation

On considère le système dynamique suivant

$$X_k = F_k\, X_{k-1} + f_k + G_k\, W_k \ , \tag{2.2}$$
$$Y_k = H_k\, X_k + h_k + V_k \ , \tag{2.3}$$

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

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

$$\begin{aligned} \bar{X}_k &= F_k\, \bar{X}_{k-1} + f_k \ , \\ Q_k^X &= F_k\, Q_{k-1}^X\, F_k^* + G_k\, Q_k^W\, G_k^* \ , \\ \bar{Y}_k &= H_k\, \bar{X}_k + h_k \ , \\ Q_k^Y &= H_k\, Q_k^X\, H_k^* + Q_k^V \ , \\ Q_k^{XY} &= Q_k^X\, H_k^* \ . \end{aligned}$$

Remarque 2.2 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_k^W, Q_k^V 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_k^W peuvent dépendre de $(Y_0, Y_1, \cdots, Y_k)$,
- H_k, h_k, et Q_k^V peuvent dépendre de $(Y_0, Y_1, \cdots, Y_{k-1})$.

Dans ce cas le processus $\{(X_k, Y_k)\}$ (et a fortiori le processus $\{X_k\}$) n'est plus gaussien, mais le processus $\{X_k\}$ est *conditionnellement gaussien* par rapport au processus $\{Y_k\}$ (i.e. pour des valeurs de l'observation $Y_{0:k-1} = (Y_0, \ldots, Y_{k-1})$ données, l'état X_k du système (2.2) est gaussien).

Chapitre 3

Filtre de Kalman, et extensions

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 $Y_{0:k} = (Y_0, \cdots, Y_k)$ 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 1.2, la solution est de calculer la loi conditionnelle de X_k sachant $Y_{0:k}$.

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 3.1). 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 3.2). Les filtres ainsi obtenus sont souvent utilisables en pratique mais conduisent parfois à de mauvais résultats.

3.1 Filtre de Kalman

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

$$X_k \ = \ F_k \, X_{k-1} + f_k + G_k \, W_k \ , \tag{3.1}$$

$$Y_k \ = \ H_k \, X_k + h_k + V_k \ , \tag{3.2}$$

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

$$Y_{0:k} \triangleq (Y_0, Y_1, \cdots, Y_k) \ .$$

L'objectif est d'estimer le vecteur aléatoire X_k à partir de $Y_{0:k}$, de façon optimale et récursive. Si on adopte le critère du minimum de variance, il s'agit d'après le paragraphe 1.2 de calculer la loi conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$. Comme le cadre est gaussien, il suffit de calculer la moyenne et la matrice de covariance

$$\widehat{X}_k \triangleq \mathbb{E}[X_k \mid Y_{0:k}] \qquad \text{et} \qquad P_k \triangleq \mathbb{E}[(X_k - \widehat{X}_k)\,(X_k - \widehat{X}_k)^* \mid Y_{0:k}] \ .$$

On définit également les quantités suivantes

$$\widehat{X}_k^- \triangleq \mathbb{E}[X_k \mid Y_{0:k-1}] \qquad \text{et} \qquad P_k^- \triangleq \mathbb{E}[(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^* \mid Y_{0:k-1}] \ .$$

D'après la remarque 1.7, les matrices de covariances conditionnelles P_k et P_k^- ne dépendent pas des observations, c'est–à–dire que

$$P_k \triangleq \mathbb{E}[(X_k - \widehat{X}_k)\,(X_k - \widehat{X}_k)^*] \qquad \text{et} \qquad P_k^- \triangleq \mathbb{E}[(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^*] \ .$$

Supposons connue la loi conditionnelle du vecteur aléatoire X_{k-1} sachant $Y_{0:k-1}$. Pour calculer la loi conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$, on procède en deux étapes.

- Dans l'étape de *prédiction*, on calcule la loi conditionnelle du vecteur aléatoire X_k sachant les observations passées $Y_{0:k-1}$, ce qui est facile à partir de l'équation (3.1).
- 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 $Y_{0:k-1}$, c'est–à–dire

  $$I_k = Y_k - \mathbb{E}[Y_k \mid Y_{0:k-1}] \ .$$

  D'après l'équation (3.2)

  $$I_k = Y_k - (H_k\,\mathbb{E}[X_k \mid Y_{0:k-1}] + h_k + \mathbb{E}[V_k \mid Y_{0:k-1}]) = Y_k - (H_k\,\widehat{X}_k^- + h_k) \ ,$$

  compte tenu que V_k et $Y_{0:k-1}$ sont indépendants.

Lemme 3.1 *Le processus $\{I_k\}$ est un processus gaussien à valeurs dans $\mathbb{R}^d$, 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 $Y_{0:k-1}$.

PREUVE. D'après la Remarque 1.8, l'observation prédite $\mathbb{E}[Y_k \mid Y_{0:k-1}]$ dépend de façon affine des observations passées $(Y_0, Y_1, \cdots, Y_{k-1})$, et donc l'innovation I_k dépend de façon affine des observations $(Y_0, Y_1, \cdots, Y_k)$. On en déduit que $(I_0, I_1, \cdots, I_n)$ est un vecteur aléatoire gaussien (comme transformation affine d'un vecteur aléatoire gaussien).

D'après l'équation (3.2)

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

On en déduit que

$$\begin{aligned} Q_k^I &= \mathbb{E}[I_k\,I_k^*] \\ &= \mathbb{E}[(H_k\,(X_k - \widehat{X}_k^-) + V_k)\;(H_k\,(X_k - \widehat{X}_k^-) + V_k)^*] \\ &= H_k\,\mathbb{E}[(X_k - \widehat{X}_k^-)\;(X_k - \widehat{X}_k^-)^*]\;H_k^* + \mathbb{E}[V_k\,(X_k - \widehat{X}_k^-)^*]\;H_k^* \\ &\quad + H_k\,\mathbb{E}[(X_k - \widehat{X}_k^-)\,V_k^*] + \mathbb{E}[V_k\,V_k^*] \\ &= H_k\;P_k^-\;H_k^* + Q_k^V\ . \end{aligned}$$

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

Remarque 3.2 On en déduit également que

$$\begin{aligned} \mathbb{E}[(X_k - \widehat{X}_k^-)\;I_k^*] &= \mathbb{E}[(X_k - \widehat{X}_k^-)\;(H_k\,(X_k - \widehat{X}_k^-) + V_k)^*] \\ &= \mathbb{E}[(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^*]\;H_k^* + \mathbb{E}[(X_k - \widehat{X}_k^-)\,V_k^*] \\ &= P_k^-\;H_k^*\ . \end{aligned}$$

Théorème 3.3 (Filtre de Kalman–Bucy) *On suppose que la matrice de covariance Q_k^V est* inversible, *pour tout $k \in \mathbb{N}$. Alors $\{\widehat{X}_k\}$ et $\{P_k\}$ sont définis par les équations suivantes*

$$\widehat{X}_k^- = F_k\,\widehat{X}_{k-1} + f_k\ , \tag{3.3}$$

$$P_k^- = F_k\,P_{k-1}\,F_k^* + G_k\,Q_k^W\,G_k^*\ , \tag{3.4}$$

et

$$\widehat{X}_k = \widehat{X}_k^- + K_k\,[Y_k - (H_k\,\widehat{X}_k^- + h_k)]\ , \tag{3.5}$$

$$P_k = [I - K_k\,H_k]\;P_k^-\ , \tag{3.6}$$

où la matrice

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

est appelée gain *de Kalman, et avec les initialisations*

$$\widehat{X}_0^- = \bar{X}_0 = \mathbb{E}[X_0]\ , \quad P_0^- = Q_0^X = \mathrm{cov(X_0)}\ .$$

Remarque 3.4 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 $\{F_k\}$, $\{G_k\}$, $\{H_k\}$, $\{Q_k^W\}$ et $\{Q_k^V\}$ sont constants, c'est–à–dire où

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

pour tout $k \geq 0$.

Construction du filtre de Kalman–Bucy On procède en plusieurs étapes. Le point central est la Proposition 1.6 qui sera constamment utilisée.

$\widehat{X}_0$ et P_0 en fonction de $\widehat{X}_0^-$ et P_0^-

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

$$\begin{pmatrix} \bar{X}_0 \\ H_0\,\bar{X}_0 + h_0 \end{pmatrix} \ , \quad \begin{pmatrix} Q_0^X & Q_0^X\,H_0^* \\ H_0\,Q_0^X & H_0\,Q_0^X\,H_0^* + Q_0^V \end{pmatrix} \ ,$$

respectivement. D'après la Proposition 1.6, la loi de X_0 sachant Y_0 est gaussienne, de moyenne

$$\widehat{X}_0 = \bar{X}_0 + Q_0^X\,H_0^*\,[H_0\,Q_0^X\,H_0^* + Q_0^V]^{-1}\,[Y_0 - (H_0\,\bar{X}_0 + h_0)] \ ,$$

et de matrice de covariance

$$P_0 = Q_0^X - Q_0^X\,H_0^*\,[H_0\,Q_0^X\,H_0^* + Q_0^V]^{-1}\,H_0\,Q_0^X \ .$$

$\widehat{X}_k^-$ et P_k^- en fonction de $\widehat{X}_{k-1}$ et P_{k-1}

Le vecteur aléatoire $(X_k, Y_0, \cdots, Y_{k-1})$ est gaussien, et d'après la Proposition 1.6, la loi de X_k sachant $Y_{0:k-1}$ est gaussienne, de moyenne $\widehat{X}_k^-$ et de matrice de covariance P_k^-. D'après l'équation

$$X_k = F_k\,X_{k-1} + f_k + G_k\,W_k \ , \tag{3.7}$$

on a

$$\begin{aligned} \widehat{X}_k^- &= \mathbb{E}[X_k \mid Y_{0:k-1}] \\ &= F_k\,\mathbb{E}[X_{k-1} \mid Y_{0:k-1}] + f_k + G_k\,\mathbb{E}[W_k \mid Y_{0:k-1}] \\ &= F_k\,\widehat{X}_{k-1} + f_k \ , \end{aligned}$$

compte tenu que W_k et Y_{k-1} sont indépendants. Par différence

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

de sorte que

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

Dans cette dernière égalité, on a utilisé le fait que $(X_{k-1} - \widehat{X}_{k-1})$ est indépendant de W_k, donc $\mathbb{E}[(X_{k-1} - \widehat{X}_{k-1})\,W_k^*] = 0$. □

$\widehat{X}_k$ et P_k en fonction de $\widehat{X}_k^-$ et P_k^-

Le vecteur aléatoire $(X_k, Y_0, \cdots, Y_k)$ est gaussien, et d'après la Proposition 1.6, la loi de X_k sachant $Y_{0:k}$ est gaussienne, de moyenne $\widehat{X}_k$ et de matrice de covariance déterministe P_k. D'après le Lemme 3.1

$$\begin{aligned} \widehat{X}_k &= \mathbb{E}[X_k \mid Y_{0:k}] \\ &= \widehat{X}_k^- + \mathbb{E}[X_k - \widehat{X}_k^- \mid Y_{0:k}] \\ &= \widehat{X}_k^- + \mathbb{E}[X_k - \widehat{X}_k^- \mid Y_{0:k-1}, I_k] \\ &= \widehat{X}_k^- + \mathbb{E}[X_k - \widehat{X}_k^- \mid I_k]\ . \end{aligned}$$

Par différence

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

de sorte que

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

Il suffit donc de calculer la moyenne conditionnelle et la matrice de covariance 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 1.6, 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. □

3.2 Extensions au cas non–linéaire

On considère un système non linéaire

$$X_k = f_k(X_{k-1}) + g_k(X_{k-1})\, W_k \ , \tag{3.8}$$
$$Y_k = h_k(X_k) + V_k \ , \tag{3.9}$$

où $\{X_k\}$, $\{Y_k\}$, $\{W_k\}$, $\{V_k\}$ prennent respectivement leurs valeurs dans $\mathbb{R}^m$, $\mathbb{R}^d$, $\mathbb{R}^p$ et $\mathbb{R}^d$, et où les fonctions f_k, g_k et h_k sont définies sur $\mathbb{R}^m$, à valeurs dans $\mathbb{R}^m$, $\mathbb{R}^{m\times p}$ et $\mathbb{R}^d$ respectivement. On suppose que les fonctions f_k et h_k sont dérivables. $\{W_k\}$ 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 (3.8).

Pour le système (3.8), (3.9), la plupart des propriétés obtenues au chapitre 2 ne sont plus vraies. En particulier le processus solution de (3.8), (3.9) n'est pas gaussien (ni même conditionnellement gaussien), ses moments ne peuvent pas être calculés de manière simple.

Filtre de Kalman linéarisé

On se donne $\{\bar{x}_k\}$ une suite (déterministe) dans $\mathbb{R}^m$, 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 et g_k autour de $\bar{x}_{k-1}$, c'est–à–dire

$$f_k(x) \simeq f_k(\bar{x}_{k-1}) + f_k'(\bar{x}_{k-1})\,(x - \bar{x}_{k-1}) \qquad \text{et} \qquad g_k(x) \simeq g_k(\bar{x}_{k-1}) \ ,$$

et la fonction h_k autour de $\bar{x}_k$, c'est–à–dire

$$h_k(x) \simeq h_k(\bar{x}_k) + h_k'(\bar{x}_k)\,(x - \bar{x}_k) \ .$$

Le système (3.8) (3.9) est alors remplacé par

$$X_k = F_k\,(X_{k-1} - \bar{x}_{k-1}) + f_k + G_k\, W_k \ , \tag{3.10}$$
$$Y_k = H_k\,(X_k - \bar{x}_k) + h_k + V_k \ , \tag{3.11}$$

avec $F_k \stackrel{\triangle}{=} f_k'(\bar{x}_{k-1})$, $f_k \stackrel{\triangle}{=} f_k(\bar{x}_{k-1})$, $G_k \stackrel{\triangle}{=} g_k(\bar{x}_{k-1})$, $H_k \stackrel{\triangle}{=} h_k'(\bar{x}_k)$ et $h_k \stackrel{\triangle}{=} h_k(\bar{x}_k)$. On applique alors le filtre de Kalman–Bucy à ce nouveau système, et on obtient exactement

$$\begin{aligned}
\widehat{X}_k^- &= f_k(\bar{x}_{k-1}) + f_k'(\bar{x}_{k-1})\,(\widehat{X}_{k-1} - \bar{x}_{k-1}) \ , \\
P_k^- &= f_k'(\bar{x}_{k-1})\, P_{k-1}\, [f_k'(\bar{x}_{k-1})]^* + g_k(\bar{x}_{k-1})\, Q_k^W\, [g_k(\bar{x}_{k-1})]^* \ ,
\end{aligned}$$

$$\begin{aligned}
\widehat{X}_k &= \widehat{X}_k^- + K_k\,[\,Y_k - [h_k'(\bar{x}_k)\,(\widehat{X}_k^- - \bar{x}_k) + h_k(\bar{x}_k)]\,] \ , \\
P_k &= [I - K_k\, h_k'(\bar{x}_k)]\, P_k^- \ , \\
K_k &= P_k^-\, h_k'(\bar{x}_k)^*\,[\,h_k'(\bar{x}_k)\, P_k^-\, [h_k'(\bar{x}_k)]^* + Q_k^V\,]^{-1} \ .
\end{aligned}$$

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

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

On choisit enfin l'initialisation $\widehat{X}_0^-$ et R_0^- de telle sorte que $\mathcal{N}(\widehat{X}_0^-, R_0^-)$ soit une bonne approximation de la loi de X_0.

Résultat 3.5 (filtre de Kalman linéarisé)

$$\begin{aligned} \widehat{X}_k^- &= f_k(\widehat{X}_{k-1}) \ , \\ P_k^- &= F_k \, P_{k-1} \, F_k^* + G_k \, Q_k^W \, G_k^* \ , \end{aligned}$$

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

avec $F_k \stackrel{\triangle}{=} f_k'(\bar{x}_{k-1})$, $G_k \stackrel{\triangle}{=} g_k(\bar{x}_{k-1})$, *et* $H_k \stackrel{\triangle}{=} h_k'(\bar{x}_k)$, *où* $\{\bar{x}_k\}$ *est une trajectoire nominale donnée (une suite dans* $\mathbb{R}^m$*).*

Filtre de Kalman étendu

On a vu (cf. paragraphe 3.1) 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 l'estimateur courant. La méthode consiste à linéariser les fonctions f_k et g_k autour de $\widehat{X}_{k-1}$, c'est–à–dire

$$f_k(x) \simeq f_k(\widehat{X}_{k-1}) + f_k'(\widehat{X}_{k-1}) \, (x - \widehat{X}_{k-1}) \qquad \text{et} \qquad g_k(x) \simeq g_k(\widehat{X}_{k-1}) \ ,$$

et à linéariser la fonction h_k autour de $\widehat{X}_k^-$, c'est–à–dire

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

Le système (3.8) (3.9) est alors remplacé par

$$X_k = F_k \, (X_{k-1} - \widehat{X}_{k-1}) + f_k + G_k \, W_k \ , \tag{3.12}$$

$$Y_k = H_k \, (X_k - \widehat{X}_k^-) + h_k + V_k \ , \tag{3.13}$$

avec $F_k \stackrel{\triangle}{=} f_k'(\widehat{X}_{k-1})$, $f_k \stackrel{\triangle}{=} f_k(\widehat{X}_{k-1})$, $G_k \stackrel{\triangle}{=} g_k(\widehat{X}_{k-1})$, $H_k \stackrel{\triangle}{=} h_k'(\widehat{X}_k^-)$ et $h_k \stackrel{\triangle}{=} h_k(\widehat{X}_k^-)$. On applique alors le filtre de Kalman–Bucy à ce nouveau système, et on obtient exactement le résultat suivant.

Résultat 3.6 (filtre de Kalman étendu)

$$\begin{aligned}
\widehat{X}_k^- &= f_k(\widehat{X}_{k-1}) \ , \\
P_k^- &= F_k \, P_{k-1} \, F_k^* + G_k \, Q_k^W \, G_k^* \ ,
\end{aligned}$$

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

avec $F_k \stackrel{\triangle}{=} f_k'(\widehat{X}_{k-1})$, $G_k \stackrel{\triangle}{=} g_k(\widehat{X}_{k-1})$, *et* $H_k \stackrel{\triangle}{=} h_k'(\widehat{X}_k^-)$.

Remarque 3.7

- 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"

$$I_k \stackrel{\triangle}{=} Y_k - h_k(\widehat{X}_k^-)$$

  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 4

Systèmes non–linéaires non–gaussiens, et extensions

On considère un système non–linéaire

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

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

plus général que le système (3.8), (3.9), et où $\{X_k\}$, $\{Y_k\}$, $\{W_k\}$, $\{V_k\}$ prennent respectivement leurs valeurs dans $\mathbb{R}^m$, $\mathbb{R}^d$, $\mathbb{R}^p$ et $\mathbb{R}^d$, et où les fonctions f_k et h_k sont définies sur $\mathbb{R}^m \times \mathbb{R}^p$ et $\mathbb{R}^m$, à valeurs dans $\mathbb{R}^m$ et $\mathbb{R}^d$ respectivement. On ne suppose pas que les fonctions f_k et h_k sont dérivables. $\{W_k\}$ et $\{V_k\}$ sont des bruits blancs, pas nécessairement gaussiens, indépendants entre eux et indépendants de la condition initiale X_0 de (4.1). On suppose que pour tout instant k

- il est facile de *simuler* un vecteur aléatoire selon la loi $p_k^W(dw)$ de W_k,
- la loi du vecteur aléatoire V_k admet une densité $q_k^V(v)$ qu'il est facile d'*évaluer* pour tout $v \in \mathbb{R}^d$.

4.1 Équations d'état

Proposition 4.1 *La suite $\{X_k\}$ est une chaîne de Markov à valeurs dans $\mathbb{R}^m$, c'est–à–dire que la loi conditionnelle par rapport au passé*

$$\mathbb{P}[X_k \in dx' \mid X_0, \cdots, X_{k-1}] = \mathbb{P}[X_k \in dx' \mid X_{k-1}] \;,$$

ne dépend que du passé immédiat, avec le noyau de probabilités de transition

$$\mathbb{P}[X_k \in dx' \mid X_{k-1} = x] = Q_k(x, dx') \;,$$

défini par

$$Q_k\,\phi(x) = \mathbb{E}[\,\phi(X_k) \mid X_{k-1} = x] = \int_{\mathbb{R}^p} \phi(f_k(x,w))\; p_k^W(dw) \;,$$

pour toute fonction test ϕ mesurable bornée, définie sur $\mathbb{R}^m$.

PREUVE. Compte tenu que W_k est indépendant de $(X_0, \cdots, X_{k-1})$, on a

$$\mathbb{E}[\,\phi(X_k) \mid X_0, \cdots, X_{k-1}] = \mathbb{E}[\,\phi(f_k(X_{k-1}, W_k)) \mid X_0, \cdots, X_{k-1}]$$

$$= \int_{\mathbb{R}^p} \phi(f_k(X_{k-1}, w))\, p_k^W(dw)\ ,$$

pour toute fonction ϕ mesurable bornée définie sur $\mathbb{R}^m$. Clairement, le résultat ne dépend que de X_{k-1}, c'est–à–dire que

$$\mathbb{E}[\,\phi(X_k) \mid X_0, \cdots, X_{k-1}] = \mathbb{E}[\,\phi(X_k) \mid X_{k-1}]\ ,$$

et

$$\mathbb{E}[\,\phi(X_k) \mid X_{k-1} = x] = \int_{\mathbb{R}^p} \phi(f_k(x, w))\, p_k^W(dw)\ . \qquad \square$$

Remarque 4.2 Si $f_k(x, w) = b_k(x) + w$, et si la loi $p_k^W(dw)$ de W_k admet une densité encore notée $p_k^W(w)$, c'est–à–dire si $p_k^W(dw) = p_k^W(w)\, dw$, alors

$$Q_k(x, dx') = p_k^W(x' - b_k(x))\, dx'$$

c'est–à–dire que le noyau $Q_k(x, dx')$ admet une densité. En effet, le changement de variable $x' = b_k(x) + w$ donne immédiatement

$$Q_k\,\phi(x) = \int_{\mathbb{R}^m} \phi(b_k(x) + w)\, p_k^W(w)\, dw = \int_{\mathbb{R}^m} \phi(x')\, p_k^W(x' - b_k(x))\, dx'\ ,$$

pour toute fonction test ϕ mesurable bornée, définie sur $\mathbb{R}^m$.

Remarque 4.3 En général, le noyau $Q_k(x, dx')$ n'admet pas de densité. En effet, conditionnellement à $X_{k-1} = x$, le vecteur aléatoire X_k appartient nécessairement au sous–ensemble

$$\mathcal{M}(x) = \{x' \in \mathbb{R}^m \ : \text{il existe } w \in \mathbb{R}^p \text{ tel que } x' = f_k(x, w)\}\ ,$$

et dans le cas où $p < m$ ce sous ensemble $\mathcal{M}(x)$ est généralement, sous certaines hypothèses de régularité, une sous–variété différentielle de dimension p dans l'espace $\mathbb{R}^m$. Il ne peut donc pas y avoir de densité pour la loi $Q_k(x, dx')$ du vecteur aléatoire X_k.

4.2 Équations d'état et d'observation

Proposition 4.4 *La suite $\{Y_k\}$ vérifie l'hypothèse de canal sans mémoire, c'est–à–dire que*

- *conditionnellement aux états cachés $X_0, \cdots, X_n$ les observations $Y_0, \cdots, Y_n$ sont mutuellement indépendantes,*

- *pour tout $k = 0, \cdots, n$, la loi conditionnelle de Y_k sachant $X_0, \cdots, X_n$ ne dépend que de X_k, avec les probabilités d'*émission

$$\mathbb{P}[Y_k \in dy \mid X_k = x] = q_k^V(y - h_k(x))\, dy \ ,$$

et on définit la fonction de vraisemblance

$$\Psi_k(x) = q_k^V(Y_k - h_k(x)) \ ,$$

qui mesure l'adéquation d'un état quelconque $x \in \mathbb{R}^m$ avec l'observation Y_k.

En d'autres termes

$$\mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n \mid X_0, \cdots, X_n] = \prod_{k=0}^{n} \mathbb{P}[Y_k \in dy_k \mid X_k] \ .$$

Preuve. Pour toute famille $g_0, \cdots, g_n$ de fonctions mesurables bornées définies sur $\mathbb{R}^d$, et compte tenu que les vecteurs aléatoires $V_0, \cdots, V_n$ sont mutuellement indépendants et indépendants des vecteurs aléatoires $X_0, \cdots, X_n$, on a

$$\begin{aligned}
&\mathbb{E}[g_0(Y_0) \cdots g_n(Y_n) \mid X_0, \cdots, X_n] \\
&= \mathbb{E}[g_0(h_0(X_0) + V_0) \cdots g_n(h_n(X_n) + V_n) \mid X_0, \cdots, X_n] \\
&= \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} g_0(h_0(X_0) + v_0) \cdots g_n(h_n(X_n) + v_n)\, \mathbb{P}[V_0 \in dv_0, \cdots, V_n \in dv_n] \\
&= \prod_{k=0}^{n} \int_{\mathbb{R}^d} g_k(h_k(X_k) + v)\, \mathbb{P}[V_k \in dv] \\
&= \prod_{k=0}^{n} \int_{\mathbb{R}^d} g_k(h_k(X_k) + v)\, q_k^V(v)\, dv \\
&= \prod_{k=0}^{n} \int_{\mathbb{R}^d} g_k(y) \underbrace{q_k^V(y - h_k(X_k))\, dy}_{\mathbb{P}[Y_k \in dy \mid X_k]} = \prod_{k=0}^{n} \mathbb{E}[g_k(Y_k) \mid X_k] \ . \qquad \square
\end{aligned}$$

Extension : Modèles de Markov cachés

Plus généralement, on peut aussi considérer un modèle de Markov caché où les états cachés $\{X_k\}$ forment une chaîne de Markov à valeurs dans un espace E, de noyaux de transition

$$\mathbb{P}[X_k \in dx' \mid X_{k-1} = x] = Q_k(x, dx') \ ,$$

et de loi initiale

$$\mathbb{P}[X_0 \in dx] = \eta_0(dx) \ ,$$

et où les observations $\{Y_k\}$ vérifient l'hypothèse de *canal sans mémoire*, c'est–à–dire que

- conditionnellement aux états cachés $X_0, \cdots, X_n$ les observations $Y_0, \cdots, Y_n$ sont mutuellement indépendantes,
- pour tout $k = 0, \cdots, n$, la loi conditionnelle de Y_k sachant $X_0, \cdots, X_n$ ne dépend que de X_k, avec la probabilité d'*émission*

  $$\mathbb{P}[Y_k \in dy \mid X_k = x] = g_k(x, y)\, dy \ ,$$

  et on définit la *fonction de vraisemblance*

  $$\Psi_k(x) = g_k(x, Y_k) \ ,$$

  qui mesure l'adéquation d'un état quelconque $x \in \mathbb{R}^m$ avec l'observation Y_k.

On suppose en outre que pour tout instant k

- il est facile de *simuler* pour tout $x \in E$, un vecteur aléatoire selon la loi $Q_k(x, dx')$,
- il est facile d'*évaluer* pour tout $x \in E$, la fonction de vraisemblance $\Psi_k(x)$.

Chapitre 5

Filtre bayésien optimal

L'objectif de ce chapître est d'établir les équations du filtre non–linéaire optimal, pour les systèmes non–linéaires et non–gaussiens, ou plus généralement les équations du filtre bayésien optimal, pour les modèles de Markov cachés. Il s'agit donc de calculer la loi conditionnelle de la variable aléatoire X_k sachant $Y_{0:k}$, et la loi conditionnelle de la variable aléatoire X_k sachant $Y_{0:k-1}$, définies par

$$\mu_k(dx) = \mathbb{P}[X_k \in dx \mid Y_{0:k}] \qquad \text{et} \qquad \mu_k^-(dx) = \mathbb{P}[X_k \in dx \mid Y_{0:k-1}] \ ,$$

respectivement.

5.1 Flots de Feynman–Kac

D'après la formule de Bayes, et d'après la propriété de canal sans mémoire

$$\mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n]$$

$$= \mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n \mid X_0 = x_0, \cdots, X_n = x_n] \, \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n]$$

$$= \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \, \prod_{k=0}^{n} g_k(x_k, y_k) \, dy_0 \cdots dy_n \ .$$

En intégrant par rapport aux variables $x_0, \cdots, x_n$, on obtient la loi jointe des observations $(Y_0, \cdots, Y_n)$, c'est–à–dire

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

$$= \int_E \cdots \int_E \prod_{k=0}^{n} g_k(x_k, y_k) \, \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \, dy_0 \cdots dy_n$$

$$= \mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k) \,] \, dy_0 \cdots dy_n \ .$$

D'après la formule de Bayes, il vient

$$\mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n]$$

$$= \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \prod_{k=0}^{n} g_k(x_k, y_k)\, dy_0 \cdots dy_n$$

$$= \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n \mid Y_0 = y_0, \cdots, Y_n = y_n]\ \mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n]$$

$$= \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n \mid Y_0 = y_0, \cdots, Y_n = y_n]\ \mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k)\,]\, dy_0 \cdots dy_n\ ,$$

et on obtient

$$\prod_{k=0}^{n} g_k(x_k, y_k)\ \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n]$$

$$= \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n \mid Y_0 = y_0, \cdots, Y_n = y_n]\ \mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k)\,]\ ,$$

pour toute suite $(y_0, \cdots, y_n)$ d'observations. Pour toute fonction test f_n définie sur l'espace produit E^{n+1}

$$\mathbb{E}[\, f_n(X_0, \cdots, X_n) \prod_{k=0}^{n} g_k(X_k, y_k)\,]$$

$$= \int_E \cdots \int_E f_n(x_0, \cdots, x_n) \prod_{k=0}^{n} g_k(x_k, y_k)\ \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n]$$

$$= \int_E \cdots \int_E f_n(x_0, \cdots, x_n)$$

$$\mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n \mid Y_0 = y_0, \cdots, Y_n = y_n]\ \mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k)\,]$$

$$= \mathbb{E}[\, f_n(X_0, \cdots, X_n) \mid Y_0 = y_0, \cdots, Y_n = y_n]\ \mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k)\,]\ ,$$

et on en déduit que

$$\mathbb{E}[f_n(X_0, \cdots, X_n) \mid Y_0 = y_0, \cdots, Y_n = y_n] = \frac{\mathbb{E}[f_n(X_0, \cdots, X_n) \prod_{k=0}^{n} g_k(X_k, y_k)\,]}{\mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k)\,]}\ .$$

Comme cette identité est vérifié pour toute suite $(y_0, \cdots, y_n)$ d'observations, on a finalement

$$\mathbb{E}[f_n(X_0, \cdots, X_n) \mid Y_0, \cdots, Y_n] = \frac{\mathbb{E}[f_n(X_0, \cdots, X_n) \prod_{k=0}^{n} \Psi_k(X_k)]}{\mathbb{E}[\prod_{k=0}^{n} \Psi_k(X_k)]} ,$$

où l'espérance porte seulement sur les états cachés successifs $(X_0, \cdots, X_n)$: les fonctions de vraisemblance $\Psi_0(x), \cdots, \Psi_n(x)$ dépendent implicitement des observations $(Y_0, \cdots, Y_n)$, mais celles–ci sont considérées comme fixées dans l'expression ci–dessus. Si la fonction test $f_n(x_0, \cdots, x_n)$ ne dépend que de x_n, c'est–à–dire si $f_n(x_0, \cdots, x_n) = \phi(x_n)$, alors

$$\langle \mu_n, \phi \rangle = \mathbb{E}[\phi(X_n) \mid Y_0, \cdots, Y_n] = \frac{\mathbb{E}[\phi(X_n) \prod_{k=0}^{n} \Psi_k(X_k)]}{\mathbb{E}[\prod_{k=0}^{n} \Psi_k(X_k)]} = \frac{\langle \gamma_n, \phi \rangle}{\langle \gamma_n, 1 \rangle} ,$$

où la mesure positive (non–normalisée) $\gamma_n(dx)$ est définie par

$$\langle \gamma_n, \phi \rangle = \mathbb{E}[\phi(X_n) \prod_{k=0}^{n} \Psi_k(X_k)] .$$

De la même manière

$$\langle \mu_n^-, \phi \rangle = \mathbb{E}[\phi(X_n) \mid Y_0, \cdots, Y_{n-1}] = \frac{\mathbb{E}[\phi(X_n) \prod_{k=0}^{n-1} \Psi_k(X_k)]}{\mathbb{E}[\prod_{k=0}^{n-1} \Psi_k(X_k)]} = \frac{\langle \gamma_n^-, \phi \rangle}{\langle \gamma_n^-, 1 \rangle} ,$$

où la mesure positive (non–normalisée) $\gamma_n^-(dx)$ est définie par

$$\langle \gamma_n^-, \phi \rangle = \mathbb{E}[\phi(X_n) \prod_{k=0}^{n-1} \Psi_k(X_k)] .$$

5.2 Équation du filtre bayésien optimal

Pour obtenir une équation récurrente permettant d'esprimer μ_k en fonction de μ_{k-1}, il suffit donc d'une équation récurrente permettant d'esprimer γ_k en fonction de γ_{k-1}, puis de normaliser.

Théorème 5.1 (Filtre bayésien optimal) *La suite* $\{\mu_k\}$ *vérifie l'équation récurrente suivante*

$$\mu_{k-1} \xrightarrow{\text{prédiction}} \mu_k^- = \mu_{k-1}\, Q_k \xrightarrow{\text{correction}} \mu_k = \Psi_k \cdot \mu_k^- \ ,$$

où par définition

$$(\mu_{k-1}\, Q_k)(dx') = \int_E \mu_{k-1}(dx)\, Q_k(x, dx')$$

désigne l'action du noyau markovien $Q_k(x, dx')$ *sur la distribution de probabilité* $\mu_{k-1}(dx)$*, et où*

$$(\Psi_k \cdot \mu_k^-)(dx') = \frac{\Psi_k(x')\, \mu_k^-(dx')}{\langle \mu_k^-, \Psi_k \rangle} \ ,$$

désigne le produit projectif de la distribution de probabilité a priori $\mu_k^-(dx')$ *et de la fonction de vraisemblance* $\Psi_k(x')$*.*

μ_n en fonction de μ_n^-

On a

$$\begin{aligned} \langle \gamma_n, \phi \rangle &= \mathbb{E}[\phi(X_n) \prod_{k=0}^{n} \Psi_k(X_k)\,] \\ &= \mathbb{E}[\phi(X_n)\, \Psi_n(X_n) \prod_{k=0}^{n-1} \Psi_k(X_k)\,] = \langle \gamma_n^-, \Psi_n\, \phi \rangle = \langle \Psi_n\, \gamma_n^-, \phi \rangle \ , \end{aligned}$$

pour toute fonction test ϕ définie sur E, où la dernière égalité exprime simplement que

$$\langle \gamma_n^-, \Psi_n\, \phi \rangle = \int_E [\Psi_n(x)\, \phi(x)]\, \gamma_n^-(dx) = \int_E \phi(x)\, [\Psi_n(x)\, \gamma_n^-(dx)] = \langle \Psi_n\, \gamma_n^-, \phi \rangle \ .$$

Comme la fonction test ϕ est quelconque, on en déduit que

$$\gamma_n(dx) = \Psi_n(x)\, \gamma_n^-(dx) \ ,$$

et en normalisant, on obtient

$$\mu_n(dx) = \frac{\gamma_n(dx)}{\langle \gamma_n, 1 \rangle} = \frac{\Psi_n(x)\, \gamma_n^-(dx)}{\langle \gamma_n^-, \Psi_n \rangle} = \frac{\Psi_n(x)\, \mu_n^-(dx)}{\langle \mu_n^-, \Psi_n \rangle} \ ,$$

où la dernière égalité est obtenue en divisant numérateur et dénominateur par la constante de normalisation $\langle \gamma_n^-, 1 \rangle$.

μ_n^- en fonction de μ_{n-1}

On remarque immédiatement que

$$\langle \gamma_n^-, 1 \rangle = \mathbb{E}[\prod_{k=0}^{n-1} \Psi_k(X_k)\,] = \langle \gamma_{n-1}, 1 \rangle \ ,$$

c'est–à–dire que la constante de normalisation est conservée. En utilisant la propriété de Markov, on a

$$\begin{aligned}
\langle \gamma_n^-, \phi \rangle &= \mathbb{E}[\phi(X_n) \prod_{k=0}^{n-1} \Psi_k(X_k)\,] \\
&= \mathbb{E}[\,\mathbb{E}[\phi(X_n) \prod_{k=0}^{n-1} \Psi_k(X_k) \mid X_{0:n-1}]\,] \\
&= \mathbb{E}[\,\mathbb{E}[\phi(X_n) \mid X_{0:n-1}] \prod_{k=0}^{n-1} \Psi_k(X_k)\,] \\
&= \mathbb{E}[\,\mathbb{E}[\phi(X_n) \mid X_{n-1}] \prod_{k=0}^{n-1} \Psi_k(X_k)\,] \\
&= \mathbb{E}[Q_n\,\phi(X_{n-1}) \prod_{k=0}^{n-1} \Psi_k(X_k)] = \langle \gamma_{n-1}, Q_n\,\phi \rangle = \langle \gamma_{n-1}\,Q_n, \phi \rangle \ ,
\end{aligned}$$

pour toute fonction test ϕ définie sur E, où la dernière égalité exprime simplement que

$$\begin{aligned}
\langle \gamma_{n-1}, Q_n\,\phi \rangle &= \int_E (Q_n\,\phi)(x)\,\gamma_{n-1}(dx) \\
&= \int_E [\int_E Q_n(x, dx')\,\phi(x')\,]\,\gamma_{n-1}(dx) = \int_E [\int_E \gamma_{n-1}(dx)\,Q_n(x, dx')\,]\,\phi(x') \\
&= \int_E (\gamma_{n-1}\,Q_n)(dx')\,\phi(x') = \langle \gamma_{n-1}\,Q_n, \phi \rangle \ .
\end{aligned}$$

Comme la fonction test ϕ est quelconque, on en déduit que

$$\gamma_n^-(dx') = (\gamma_{n-1}\,Q_n)(dx') \ ,$$

et en normalisant, on obtient

$$\mu_n^-(dx') = \frac{\gamma_n^-(dx')}{\langle \gamma_n^-, 1 \rangle} = \frac{(\gamma_{n-1}\,Q_n)(dx')}{\langle \gamma_{n-1}, 1 \rangle} = (\mu_{n-1}\,Q_n)(dx') \ .$$

L'équation du filtre bayésien optimal a été obtenue très simplement, mais il est en général impossible de la résoudre, sauf dans le cas particulier des systèmes linéaires gaussiens, où elle se ramène aux équations du filtre de Kalman–Bucy, présentées au Chapitre 3. Il faut donc avoir recours à une approximation numérique, et on présente ci–dessous une approximation de type Monte Carlo, appelée filtre particulaire, qui a connu un développement spectaculaire au cours des dernières années, et qui est maintenant largement répendu, en particulier dans les applications en localisation, navigation ou poursuite de mobiles, aussi bien dans le domaine militaire (aéronef, sous–marin, bâtiment de surface, missile, drone, etc.), que dans le domaine civil, avec des applications en robotique mobile ou en communications sans–fil.

5.3 Approximation particulaire

L'idée du filtrage particulaire consiste à chercher une approximation de la distribution de probabilité conditionnelle $\mu_k(dx)$ sous la forme d'une combinaison linéaire pondérée de masses de Dirac, appelées *particules*, de la forme

$$\mu_k \approx \mu_k^N = \sum_{i=1}^N w_k^i \, \delta_{\xi_k^i} \qquad \text{avec} \qquad \sum_{i=1}^N w_k^i = 1 \ ,$$

où les *positions* $\{\xi_k^i \, , \, i = 1, \cdots, N\}$ des particules sont des éléments de l'espace d'état E, et où les *poids* $\{w_k^i \, , \, i = 1, \cdots, N\}$ des particules sont des nombres compris entre 0 et 1. Cette approximation est complètement caractérisée par la donnée du système de particules $\Sigma_k = \{\xi_k^i, w_k^i \, , \, i = 1, \cdots, N\}$, et l'algorithme est complètement décrit par le mécanisme qui permet de construire Σ_k à partir de Σ_{k-1}. On rappelle que la suite $\{\mu_k\}$ vérifie l'équation récurrente

$$\mu_{k-1} \xrightarrow{\text{prédiction}} \mu_k^- = \mu_{k-1} \, Q_k \xrightarrow{\text{correction}} \mu_k = \Psi_k \cdot \mu_k^- \ ,$$

d'après le Théorème 5.1. Si on applique le noyau markovien $Q_k(x, dx')$ à l'approximation

$$\mu_{k-1}^N = \sum_{i=1}^N w_{k-1}^i \, \delta_{\xi_{k-1}^i} \ ,$$

on obtient exactement

$$(\mu_{k-1}^N \, Q_k)(dx') = \sum_{i=1}^N w_{k-1}^i \, Q_k(\xi_{k-1}^i, dx') \ ,$$

qui est un mélange de lois, peu pratique à manipuler, et qu'on décide de remplacer par la loi empirique

$$\mu_k^{-,N} = \frac{1}{N} \, \sum_{i=1}^N \delta_{\xi_k^i} \ ,$$

associée à un N–échantillon $\{\xi_k^i\,,\, i=1,\cdots,N\}$ de variables aléatoires ayant précisément la loi commune $(\mu_{k-1}^N\, Q_k)(dx')$. Générer un échantillon selon un mélange de lois est très simple, et peut être réalisé de la façon suivante : indépendemment pour tout $i=1,\cdots,N$

(i) on génère un indice τ_{k-1}^i appartenant à l'ensemble $\{1,\cdots,N\}$ selon la loi discrète $(w_{k-1}^1,\cdots,w_{k-1}^N)$, c'est–à–dire que

$$\mathbb{P}[\tau_{k-1}^i = j] = w_{k-1}^j\ , \qquad \text{pour tout } j=1,\cdots,N$$

et on pose $\widehat{\xi}_{k-1}^i = \xi_{k-1}^{\tau_{k-1}^i}$, c'est–à–dire qu'on fait le choix de la particule correspondante dans la population Σ_{k-1},

(ii) on génère une variable aléatoire ξ_k^i selon la loi $Q_k(\widehat{\xi}_{k-1}^i, dx')$, ce qui est facile par hypothèse.

On applique ensuite la formule de Bayes à l'approximation $\mu_k^{-,N}(dx')$, et on obtient exactement

$$\Psi_k \cdot \mu_k^{-,N} = \sum_{i=1}^N \frac{\Psi_k(\xi_k^i)}{\displaystyle\sum_{j=1}^N \Psi_k(\xi_k^j)}\ \delta_{\xi_k^i} = \sum_{i=1}^N w_k^i\ \delta_{\xi_k^i}\ ,$$

avec

$$w_k^i = \frac{\Psi_k(\xi_k^i)}{\displaystyle\sum_{j=1}^N \Psi_k(\xi_k^j)} \qquad \text{pour tout } i=1,\cdots,N.$$

En résumé, cet algorithme, appelé filtre particulaire *bootstrap*, peut être décrit de la façon suivante.

passage de Σ_{k-1} à Σ_k indépendemment pour tout $i=1,\cdots,N$

- générer un indice $\tau_{k-1}^i \sim (w_{k-1}^1,\cdots,w_{k-1}^N)$ et poser $\widehat{\xi}_{k-1}^i = \xi_{k-1}^{\tau_{k-1}^i}$,
- générer la nouvelle position $\xi_k^i \sim Q_k(\widehat{\xi}_{k-1}^i, dx')$,
- calculer le nouveau poids $w_k^i \propto \Psi_k(\xi_k^i)$.

Il s'agit d'une approximation numérique, très simple à mettre en œuvre puisqu'il suffit de savoir simuler des transitions indépendantes de la chaîne de Markov, et qui converge vers le filtre optimal lorsque le nombre N de particules utilisées pour les calculs tend vers l'infini. L'étape essentielle dans l'algorithme est l'étape de rééchantillonage, qui sélectionne les particules ayant une forte vraisemblance, et concentre ainsi automatiquement la puissance de calcul disponible dans les régions d'intérêt de l'espace d'état E.

Chapitre 6

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*.

6.1 Chaînes de Markov à état fini

On considère un espace d'état *fini* E à N éléments. 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)

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

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

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 présent X_k ne dépend que de l'état immédiatement passé X_{k-1}.

Définition 6.1 *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 6.2 ci–dessous qu'une chaîne de Markov $\{X_k\}$ est entièrement caractérisée par la donnée

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

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

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

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

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 6.2 *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*

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

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

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k] =$$

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

$$= \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}] \, \pi_{i_{k-1},i_k} \ .$$

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

6.2 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 $\mathbb{R}^d$. 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*

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

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

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

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

Exemple 6.3 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 $\mathbb{R}^d$ est caractérisée par la donnée d'une famille $h = (h_i)$ de N vecteurs de $\mathbb{R}^d$, et on a

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

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

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

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

Il résulte de la Proposition 6.5 ci–dessous qu'un modèle de Markov caché $\{(X_k, Y_k)\}$ est entièrement caractérisé par la donnée

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

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

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

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

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

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

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

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

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 6.5 *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*

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

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

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*

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

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

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

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

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

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

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

et on conclut en utilisant la Proposition 6.2.

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

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

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

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

et on conclut de la même manière, en utilisant la Proposition 6.2. □

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 7

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

- *dans le cas* fini *par*

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

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

- *et dans le cas* continu *par*

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

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

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

Preuve. On considère d'abord le cas *fini*. On utilise la Proposition 6.5 pour calculer la distribution de probabilité marginale

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

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

$$= \sum_{i_0, \cdots, i_n \in E} \nu_{i_0} \, \pi_{i_0, i_1} \cdots \pi_{i_{n-1}, i_n} \, b_{i_0}^{\ell_0} \cdots b_{i_n}^{\ell_n} \ .$$

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

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

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

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

Remarque 7.2 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 7.1 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.

7.1 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

$$\mathbb{P}[Y_0 = \ell_0, \cdots, Y_k = \ell_k, X_k = i] = \alpha_k^i[\ell_0, \cdots, \ell_k] \ ,$$

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

$$\mathbb{P}[Y_0 \in dy_0, \cdots, Y_k \in dy_k, X_k = i] = \alpha_k^i[y_0, \cdots, y_k]\ dy_0 \cdots dy_k \ ,$$

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

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

- dans le cas *fini*

$$\alpha_0^i[\ell] \ = \ \mathbb{P}[Y_0 = \ell, X_0 = i]$$

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

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

$$\alpha_0^i[y]\ dy \ = \ \mathbb{P}[Y_0 \in dy, X_0 = i]$$

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

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

On définit la variable *forward* $p_k = (p_k^i)$ par

$$p_k^i = \alpha_k^i[Y_0, \cdots, Y_k] \ ,$$

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 $y \in \mathbb{R}^d$, on définit la matrice diagonale $N \times N$

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

Remarque 7.3 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 7.4 *La suite* $\{p_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* fini

$$p_{k+1}^j = b_j(Y_{k+1}) \ \Big[\sum_{i \in E} \pi_{i,j} \ p_k^i \Big] \ , \tag{7.1}$$

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

$$p_0^i = \nu_i \ b_i(Y_0) \ , \qquad \textit{pour tout } i \in E,$$

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

$$p_{k+1}^j = \psi_j(Y_{k+1}) \ \Big[\sum_{i \in E} \pi_{i,j} \ p_k^i \Big] \ , \tag{7.2}$$

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

$$p_0^i = \nu_i \ \psi_i(Y_0) \ , \qquad \textit{pour tout } i \in E,$$

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

$$\alpha_k^i[\ell_0,\cdots,\ell_k] = \mathbb{P}[Y_0=\ell_0,\cdots,Y_k=\ell_k,X_k=i] =$$

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

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

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

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

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

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

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

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

$$\alpha_{k+1}^j[\ell_0,\cdots,\ell_{k+1}] = b_j^{\ell_{k+1}}\ \sum_{i\in E}\pi_{i,j}\ \alpha_k^i[\ell_0,\cdots,\ell_k]\ ,$$

d'où le résultat. □

Remarque 7.5 La distribution de probabilité des observations $(Y_0,\cdots,Y_n)$ peut se calculer de la façon suivante :

- dans le cas *fini*

$$\mathbb{P}[Y_0=\ell_0,\cdots,Y_n=\ell_n] = \sum_{i\in E}\alpha_n^i[\ell_0,\cdots,\ell_n]\ ,$$

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

$$\mathbb{P}[Y_0\in dy_0,\cdots,Y_n\in dy_n] = \Big\{\sum_{i\in E}\alpha_n^i[y_0,\cdots,y_n]\Big\}\ dy_0\cdots dy_n\ ,$$

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

Remarque 7.6 La variable forward permet de calculer, de façon récursive, la distribution de probabilité conditionnelle de l'état présent X_n sachant les observations $(Y_0, \cdots, Y_n)$:

$$\mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{p_n^i}{\sum_{j \in E} p_n^j} \ ,$$

pour tout $i \in E$ (en ce sens, p_n est une distribution de probabilité non–normalisée), et la constante de normalisation

$$L_n = \sum_{j \in E} p_n^j \ ,$$

s'interprète comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$.

Remarque 7.7 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.

7.2 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

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

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

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

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

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*

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

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

- et dans le cas *continu*

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

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

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 7.8 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

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

et de matrice de transition π.

Théorème 7.9 *La suite $\{v_k\}$ vérifie l'équation récurrente rétrograde suivante :*

- *dans le cas* fini

$$v^i_k = \sum_{j\in E} \pi_{i,j}\ b_j(Y_{k+1})\ v^j_{k+1}\ , \tag{7.3}$$

  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

$$v^i_k = \sum_{j\in E} \pi_{i,j}\ \psi_j(Y_{k+1})\ v^j_{k+1}\ , \tag{7.4}$$

  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 (7.3) permet de retrouver à l'instant $k = n - 1$

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

D'autre part, il résulte de la Remarque 7.8 et de la Proposition 6.5 que

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

$$= \sum_{i_{k+1}, \cdots, i_n \in E} \mathbb{P}[Y_{k+1} = \ell_{k+1}, \cdots, Y_n = \ell_n, X_{k+1} = i_{k+1}, \cdots, X_n = i_n \mid X_k = i]$$

$$= \sum_{i_{k+1}, \cdots, i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{\ell_{k+1}} \cdots b_{i_n}^{\ell_n}\ ,$$

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

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

$$= \sum_{i_{k+2}, \cdots, i_n \in E} \mathbb{P}[Y_{k+1} = \ell_{k+1}, \cdots, Y_n = \ell_n, X_{k+1} = j, X_{k+2} = i_{k+2}, \cdots, X_n = i_n \mid X_k = i]$$

$$= \sum_{i_{k+2}, \cdots, i_n \in E} \pi_{i,j}\ \pi_{j,i_{k+2}} \cdots \pi_{i_{n-1},i_n}\ b_j^{\ell_{k+1}}\ b_{i_{k+2}}^{\ell_{k+2}} \cdots b_{i_n}^{\ell_n}$$

$$= \pi_{i,j}\ b_j^{\ell_{k+1}}\ \beta_{k+1}^j[\ell_{k+2}, \cdots, \ell_n]\ ,$$

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

$$\beta_k^i[\ell_{k+1}, \cdots, \ell_n] = \sum_{j \in E} \pi_{i,j}\ b_j^{\ell_{k+1}}\ \beta_{k+1}^j[\ell_{k+2}, \cdots, \ell_n]\ ,$$

d'où le résultat. □

Proposition 7.10 *Les équations forward et backward sont* duales *l'une de l'autre :*

$$\sum_{i \in E} p_0^i\ v_0^i = \sum_{i \in E} p_k^i\ v_k^i = \sum_{i \in E} p_n^i\ ,$$

pour tout instant k.

PREUVE. On considère uniquement le cas *fini*. En utilisant successivement l'équation backward (7.3) et l'équation forward (7.1), on obtient

$$\sum_{i \in E} p_k^i\ v_k^i\ = \sum_{i \in E} p_k^i\ \Big[\ \sum_{j \in E} \pi_{i,j}\ b_j(Y_{k+1})\ v_{k+1}^j\ \Big]$$

$$= \sum_{j \in E} b_j(Y_{k+1})\ \Big[\ \sum_{i \in E} \pi_{i,j}\ p_k^i\ \Big]\ v_{k+1}^j = \sum_{j \in E} p_{k+1}^j\ v_{k+1}^j\ ,$$

d'où le résultat. □

Proposition 7.11 *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*

$$\mathbb{P}[Y_0 = \ell_0, \cdots, Y_n = \ell_n, X_k = i] = \alpha_k^i[\ell_0, \cdots, \ell_k]\ \beta_k^i[\ell_{k+1}, \cdots, \ell_n]\ ,$$

  pour tout $i \in E$, et tout $\ell_0, \cdots, \ell_n \in O$,

- *et dans le cas* continu *par*

$$\mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n, X_k = i] = \alpha_k^i[y_0, \cdots, y_k]\ \beta_k^i[y_{k+1}, \cdots, y_n]\, dy_0 \cdots dy_n\ ,$$

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

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 :

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

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

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

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

$$= \alpha_k^i[\ell_0, \cdots, \ell_k]\ \beta_k^i[\ell_{k+1}, \cdots, \ell_n]\ ,$$

d'où le résultat. □

Remarque 7.12 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 :

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{q_k^i}{\sum_{j \in E} q_k^j}\ ,$$

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

$$q_k^i = p_k^i \; v_k^i \; , \qquad \text{pour tout } i \in E.$$

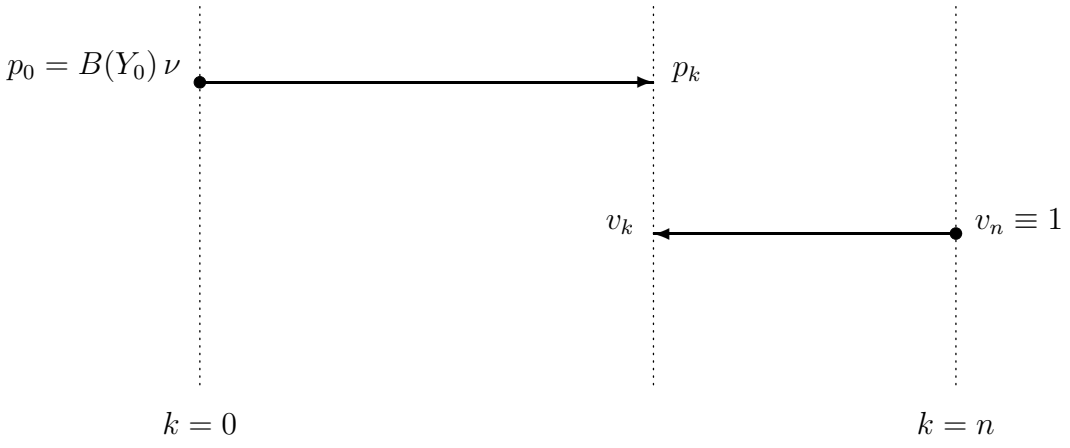

FIG. 7.1 – Equations forward–backward

On remarque que la constante de normalisation

$$\sum_{i \in E} q_k^i = \sum_{i \in E} p_k^i \; v_k^i = \sum_{i \in E} p_n^i = L_n \; ,$$

ne dépend pas de l'instant k considéré, et s'interprète comme la vraisemblance du modèle sachant les observations $(Y_0, \cdots, Y_n)$.

Chapitre 8

Algorithme de Viterbi

Il résulte des Remarques 7.6 et 7.12 que les variables forward et backward étudiées au Chapitre 7 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 $(Y_0, \cdots, Y_n)$:

$$\mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{p_n^i}{\sum_{j \in E} p_n^j} \ ,$$

et

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{q_k^i}{\sum_{j \in E} q_k^j} \ ,$$

respectivement, pour tout $i \in E$.

Compte tenu que les états possibles pour la chaîne de Markov ne se prêtent 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

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

soit pour l'état à un instant intermédiaire

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

(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 $X_{k+1}^{\mathrm{MAP,\,loc}} = j$ 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

$$(X_0^{\mathrm{MAP}}, \cdots, X_n^{\mathrm{MAP}}) = \underset{i_0,\cdots,i_n \in E}{\operatorname{argmax}} \ \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] \ .$$

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*

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

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

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

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

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

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

Pour tout instant k, on définit :

- dans le cas *fini*

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

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

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

$$= \delta_k^i[y_0, \cdots, y_k] \, dy_0 \cdots dy_k \ ,$$

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

La fonction *valeur* $V_k = (V_k^i)$ est définie par

$$V_k^i = \delta_k^i[Y_0, \cdots, Y_k] \ ,$$

pour tout $i \in I$.

Théorème 8.1 *La suite* $\{V_k\}$ *vérifie la récurrence suivante :*

- *dans le cas* fini

$$V_{k+1}^j = b_j(Y_{k+1}) \ \big[\max_{i \in E} \pi_{i,j} \ V_k^i \ \big] \ , \tag{8.1}$$

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

$$V_0^i = \nu_i \ b_i(Y_0) \ , \qquad \textit{pour tout } i \in E,$$

- *et dans le cas* continu

$$V_{k+1}^j = \psi_j(Y_{k+1}) \ \big[\max_{i \in E} \pi_{i,j} \ V_k^i \ \big] \ , \tag{8.2}$$

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

$$V_0^i = \nu_i \ \psi_i(Y_0) \ , \qquad \textit{pour tout } i \in E.$$

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

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

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

Preuve. On considère uniquement le cas *fini*. Il résulte de la Proposition 6.5 que

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

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

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

$$= b_j(\ell_{k+1}) \ \pi_{i,j} \ \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, Y_0 = \ell_0, \cdots, Y_k = \ell_k] \ ,$$

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

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

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

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

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

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

d'où le résultat. □

Remarque 8.2 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*

$$\begin{aligned}
&\max_{i_0,\cdots,i_n\in E} \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 = \ell_0, \cdots, Y_n = \ell_n] = \\
&= \max_{i\in E} \max_{i_0,\cdots,i_{n-1}\in E} \mathbb{P}[X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_n = i, Y_0 = \ell_0, \cdots, Y_n = \ell_n] \\
&= \max_{i\in E} \ \delta^i_n[\ell_0, \cdots, \ell_n] \ ,
\end{aligned}$$

- et dans le cas *continu*

$$\begin{aligned}
&\max_{i_0,\cdots,i_n\in E} \mathbb{P}[X_0 = i_0, \cdots, X_n = i_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n] = \\
&= \Big\{ \max_{i\in E} \ \delta^i_n[y_0, \cdots, y_n] \Big\} \ dy_0 \cdots dy_n \ .
\end{aligned}$$

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

$$X_n^{\mathrm{MAP}} = \underset{i \in E}{\operatorname{argmax}}\ V_n^i$$

(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.

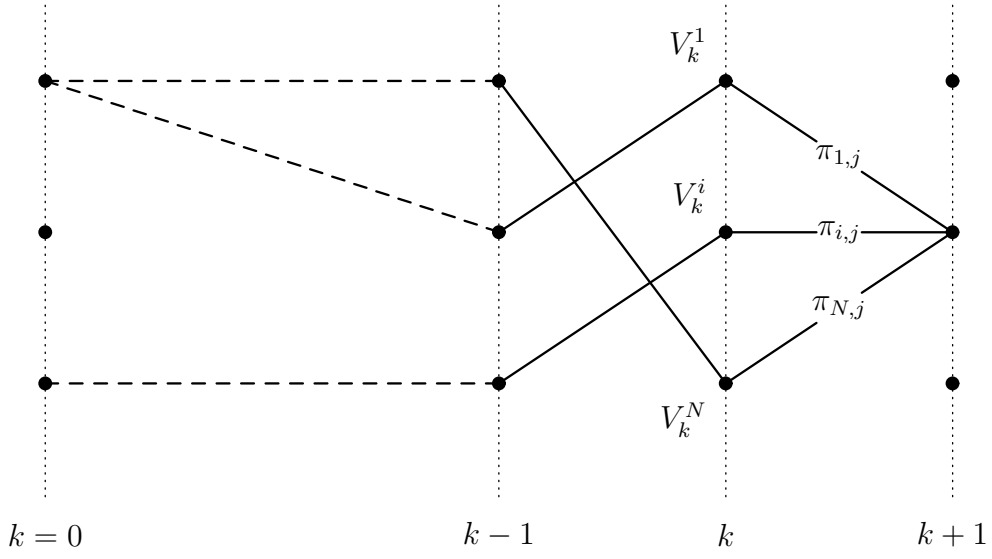

FIG. 8.1 – Algorithme de Viterbi (programmation dynamique)

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 $\mathbb{P}$ 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 $(\Omega, \mathcal{F}, \mathbb{P})$ 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 $n \in \mathbb{N}$, alors $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$.

- $\mathbb{P}$ 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 $n \in \mathbb{N}$, et $A_n \cap A_m = \emptyset$ pour tout $n \neq m$, alors

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

A partir des axiomes, on peut montrer les propriétés suivantes

- (vii) pour tout $A \in \mathcal{F}$, $0 \leq \mathbb{P}(A) \leq 1$,
- (viii) pour tout $A \in \mathcal{F}$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$,
- (ix) si $A_n \in \mathcal{F}$ pour tout $n \in \mathbb{N}$, alors

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

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 $\Omega = \mathbb{R}^n$ et $\mathcal{F}_0$ désigne l'ensemble des *parties ouvertes* de $\mathbb{R}^n$, 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 $\mathbb{R}^n$, 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 $\mathbb{R}^n$.

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

$$\{X \in B\} \stackrel{\triangle}{=} \{\omega \,:\, X(\omega) \in B\} \ ,$$

et

$$\mathbb{P}(X \in B) \stackrel{\triangle}{=} \mathbb{P}(\{X \in B\}) \ .$$

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

$$\mu_X(B) \stackrel{\triangle}{=} \mathbb{P}(X \in B) \ ,$$

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 $(\Omega, \mathcal{F}, \mathbb{P})$. S'il existe une fonction p_X définie sur $\mathbb{R}^n$, telle que pour tout $B \in \mathcal{B}^n$

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

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

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

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

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

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

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

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

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

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

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

Moyenne, covariance

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

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

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

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

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

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

avec la notation $\bar{X} = \mathbb{E}[X]$. 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 $\mathbb{E}[X] = \mu$ 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,

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

En effet

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

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

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

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

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*

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

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

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

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

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

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

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

ou de façon plus symétrique

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

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

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

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 $(\Omega, \mathcal{F}, \mathbb{P})$. 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

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

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

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

et

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

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

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

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

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

et

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

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

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

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

Conditionnement par $(Y = y)$

Etant donnés deux vecteurs aléatoires X et Y définis sur $(\Omega, \mathcal{F}, \mathbb{P})$, 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

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

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

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

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 $\mathbb{P}(X \in A \mid Y = y)$.

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

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

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

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

ce qui donne l'expression suivante

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

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 $\mathbb{R}^n$. On définit la moyenne conditionnelle de la variable aléatoire réelle $\phi(X)$ sachant $Y = y$ par

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

Le calcul donne

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

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 $\langle \xi, \eta \rangle = \mathbb{E}[\xi\, \eta]$ défini sur l'ensemble des variables aléatoires réelles de carré intégrable).

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

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

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

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)

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

Une écriture équivalente est

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

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

On obtient en particulier

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

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 :* $\mathbb{E}[\phi(X) \mid Y] = \phi(X)$.
(ii) Si $Y \perp X$, alors : $\mathbb{E}[\phi(X) \mid Y] = \mathbb{E}[\phi(X)]$.
(iii) Si $Z \perp (X, Y)$, alors : $\mathbb{E}[\,\phi(X) \mid Y, Z\,] = \mathbb{E}[\,\phi(X) \mid Y\,]$.

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

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

d'où

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

ce qui prouve (i).

Si $Y \perp X$, alors

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

ce qui prouve (ii).

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

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

ce qui prouve (iii). □

Finalement le résultat suivant, dont la démonstration est similaire à celle de la Proposition 1.4, montre que la moyenne conditionnelle sachant Y peut également s'interpréter comme un estimateur du minimum de variance.

Proposition A.7 *La moyenne conditionnelle* $\widehat{\phi}(Y) = \mathbb{E}[\phi(X) \mid Y]$ *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*

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

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

Fonction caractéristique

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

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

pour tout $u \in \mathbb{R}^n$. 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* $(\Omega, \mathcal{F}, \mathbb{P})$*. Soit* A *une application linéaire de* $\mathbb{R}^n$ *dans* $\mathbb{R}^p$*, c'est–à–dire une matrice* $p \times n$*, et soit* b *un vecteur de* $\mathbb{R}^p$*. 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 $u \in \mathbb{R}^p$*.*

Preuve. Par définition

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

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

pour tout $u \in \mathbb{R}^p$. □

Vecteurs aléatoires gaussiens

Soit X un vecteur aléatoire de dimension n défini sur $(\Omega, \mathcal{F}, \mathbb{P})$. 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 $u \in \mathbb{R}^n$, 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*

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

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

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

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

et sa variance

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

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

$$\Phi_{u^*X}(\lambda) = \mathbb{E}[\, e^{\,i\,\lambda\, u^*X}\,] = \exp\left\{ \, i\,\lambda\, u^*\mu - \tfrac{1}{2}\,\lambda^2\, u^*Qu \, \right\} = \Phi_X(\lambda u) \ ,$$

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

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

pour tout $u \in \mathbb{R}^n$. □

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*

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

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 $\mathbb{R}^n$ dans $\mathbb{R}^p$, c'est–à–dire une matrice $p \times n$, et soit b un vecteur de $\mathbb{R}^p$. 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

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

$$=\; \exp\Big\{\; i\,u^*(A\mu + b) - \tfrac{1}{2}\,u^*(AQA^*)u\;\Big\}\;\;,$$

pour tout $u \in \mathbb{R}^p$. □

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*

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

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

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

indépendamment du caractère gaussien.

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

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

$$=\; \exp\Big\{\; i\,u^*\mu_X + i\,v^*\mu_Y - \tfrac{1}{2}\,u^*Q_Xu - u^*Q_{X,Y}v - \tfrac{1}{2}\,v^*Q_Yv\;\Big\}$$

$$=\; \exp\Big\{\; i\,u^*\mu_X - \tfrac{1}{2}\,u^*Q_Xu\;\Big\}\;\exp\Big\{\; i\,v^*\mu_Y - \tfrac{1}{2}\,v^*Q_Yv\;\Big\}\;\exp\Big\{\; -\,u^*Q_{X,Y}v\;\Big\}$$

$$=\; \Phi_X(u)\;\Phi_Y(v)\;\exp\Big\{\; -\,u^*Q_{X,Y}v\;\Big\}\;\;.$$

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

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

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 $\widehat{X} = \mathbb{E}[X \mid Y]$ 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

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

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*

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

Preuve. Par définition

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

où α est un vecteur de $\mathbb{R}^n$ 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

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

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

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

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

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

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

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