Université de Rennes 1
Master EEEA — Spécialité SISEA
(Signal, Image, Systèmes Embarqués, Automatique)

Introduction au Filtrage en Temps Discret

Filtrage de Kalman et Modèles de Markov Cachés

François Le Gland
INRIA Rennes et IRMAR

http://people.rennes.inria.fr/Francois.Le_Gland/rennes-1/

2019–20

Table des matières

1 Introduction (estimation) 1

1.1 Importance de l'information a priori, via un exemple 1

1.2 Estimation bayésienne 6

2 Filtrage de Kalman 13

2.1 Systèmes linéaires gaussiens 13

2.2 Filtre de Kalman 15

2.3 Lisseur de Kalman 20

3 Extensions aux systèmes non–linéaires 29

3.1 Filtre de Kalman linéarisé, filtre de Kalman étendu 30

3.2 Filtre de Kalman *unscented* 32

4 Systèmes non–linéaires non–gaussiens, et extensions 39

4.1 Équation d'état (modèle a priori) 40

4.2 Équation d'observation (modèle de capteur) 41

5 Filtre bayésien optimal 43

5.1 Représentation probabiliste 43

5.2 Équation récurrente 45

5.3 Approximation particulaire 47

6 Introduction (classification) 51

7 Modèles de Markov cachés 55

7.1 Chaînes de Markov à état fini 55

7.2 Modèles de Markov cachés 56

8 Equations forward / backward de Baum **61**

8.1 Equation forward 63

8.2 Equation backward 67

9 Algorithme de Viterbi **75**

10 Formules de re–estimation de Baum–Welch **81**

A Rappels de probabilités **89**

Chapitre 1

Introduction (estimation)

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 $(Y_0, Y_1, \cdots, Y_n)$ d'observations, obtenues après traitement préalable du signal brut recueilli au niveau des capteurs. Chaque observation Y_k est reliée à l'état inconnu X_k par une relation probabiliste du type

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

par exemple

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

avec un *bruit* additif V_k indépendant de X_k, qui modélise l'erreur d'observation.

Tel qu'il est formulé, le problème de l'estimation de l'état inconnu X_n à partir des observations $(Y_0, Y_1, \cdots, Y_n)$ est en général mal–posé, et il est utile d'introduire un modèle *a priori* qui donne une description probabiliste de la suite $(X_0, X_1, \cdots, X_n)$.

1.1 Importance de l'information a priori, via un exemple

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, et $x \in \mathbb{R}^m$ est un paramètre inconnu. On désigne par x_{true} 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

$$Y_k = H\, x + V_k \ ,$$

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

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

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. Considérons en effet le problème d'optimisation suivant

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

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

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

s'écrivent

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

compte tenu que la matrice H est de rang plein. 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

$$H\,\widehat{x}_n = \frac{1}{n}\,\sum_{k=1}^{n} Y_k = H\,x_{\rm true} + \frac{1}{n}\,\sum_{k=1}^{n} V_k\ ,$$

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

$$H\,\widehat{x}_n \longrightarrow H\,x_{\rm true}\ ,$$

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 $\mathcal{I}(x_{\rm true})$ de dimension $(m-d)$ défini par

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

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 $x \in \mathcal{I}(x_{\rm true})$, 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

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

s'écrivent

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

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 que

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

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

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

d'où on déduit la limite suivante

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

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

$$H\, x^\perp = H\, x_{\rm true} \ ,$$

c'est–à–dire que $x^\perp$ appartient au sous–espace affine $\mathcal{I}(x_{\rm true})$, et on peut montrer qu'il s'agit du point projeté orthogonal (pour le produit scalaire associé à la matrice Σ^{-1}) du point μ sur le sous–espace affine $\mathcal{I}(x_{\rm true})$, solution du problème d'optimisation

$$\min_{x \in \mathcal{I}(x_{\rm true})} \ \left\{ \ \tfrac{1}{2}\,(x-\mu)^*\,\Sigma^{-1}\,(x-\mu) \ \right\} \ ,$$

c'est–à–dire du problème d'optimisation

$$\min_{x \in \mathbb{R}^m} \ \left\{ \ \tfrac{1}{2}\,(x-\mu)^*\,\Sigma^{-1}\,(x-\mu) \ \right\} \qquad \text{sous la contrainte} \qquad H\, x = H\, x_{\rm true} \ .$$

En effet, on peut formaliser la prise en compte de cette contrainte en introduisant le multiplicateur de Lagrange $u \in \mathbb{R}^d$ et en considérant le problème d'optimisation

$$\min_{x \in \mathbb{R}^m} \ \left\{ \ \tfrac{1}{2}\,(x-\mu)^*\,\Sigma^{-1}\,(x-\mu) + u^*\, H(x - x_{\rm true}) \ \right\} \ .$$

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

$$\begin{aligned} &\tfrac{1}{2}\,(x-\mu)^*\,\Sigma^{-1}\,(x-\mu) + u^*\, H(x - x_{\rm true}) \\ &\qquad = \tfrac{1}{2}\, x^*\,\Sigma^{-1}\, x - x^*\,\Sigma^{-1}\,\mu + \tfrac{1}{2}\,\mu^*\,\Sigma^{-1}\,\mu + x^*\, H^*\, u - x^*_{\rm true}\, H^*\, u \ , \end{aligned}$$

s'écrivent

$$\Sigma^{-1}\, x = \Sigma^{-1}\,\mu - H^*\, u \quad \Longrightarrow \quad x = \mu - \Sigma\, H^*\, u \ ,$$

et on obtient le multiplicateur de Lagrange en exprimant que la solution doit vérifier la contraite, soit

$$H\, x = H\,\mu - H\,\Sigma\, H^*\, u = H\, x_{\rm true} \quad \Longrightarrow \quad u = (H\,\Sigma\, H^*)^{-1}\, H\,(\mu - x_{\rm true}) \ ,$$

ce qui donne bien la solution

$$\begin{aligned} x^\perp \ &= \ \mu - \Sigma\, H^*\, u \\ &= \ \mu - \Sigma\, H^*\,(H\,\Sigma\, H^*)^{-1}\, H\,(\mu - x_{\rm true}) \\ &= \ \Sigma\, H^*\,(H\,\Sigma\, H^*)^{-1}\, H\, x_{\rm true} + [I - \Sigma\, H^*\,(H\,\Sigma\, H^*)^{-1}\, H]\,\mu \ . \end{aligned}$$

En d'autres termes, l'accumulation des observations permet d'apprendre le sous–espace affine $\mathcal{I}(x_{\rm true})$, et l'information a priori permet de choisir un point particulier dans ce sous–espace.

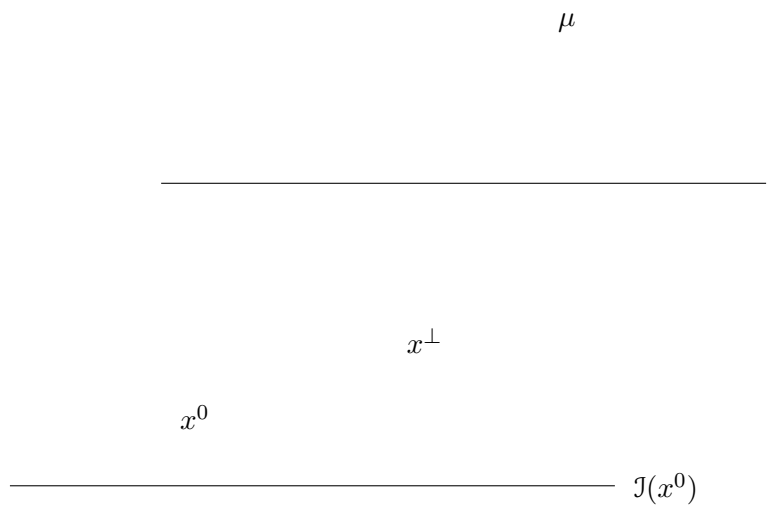

Figure 1.1: Prise en compte de l'information a priori

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

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

et de plus

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

Preuve. On remarque d'abord que

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

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

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

sont inversibles. On vérifie alors que

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

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

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

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 $u=0$. $\square$

1.2 Estimation bayésienne

Dans de nombreux cas, la prise en compte de l'information a priori peut se ramener au problème statique suivant : étant donnés deux variables aléatoires X et Y, qu'apporte le fait d'observer la réalisation $Y=y$ sur la connaissance que l'on a de X ?

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 un vecteur aléatoire $\psi(Y)$ de dimension m, où ψ est une application mesurable définie sur $\mathbb{R}^d$ à valeurs dans $\mathbb{R}^m$. 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 l'*erreur quadratique moyenne*

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

L'estimateur du minimum d'erreur quadratique moyenne (MMSE, pour *minimum mean square error*) de X sachant Y est un estimateur $\widehat{X}(Y)$ tel que

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

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

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 $p_{X,Y}(x,y)$ 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 MMSE 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 $\psi(Y)$ un estimateur quelconque.

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

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

et on remarque que

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

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

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

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 $\psi(y) = \widehat{X}(y)$ □

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

Proposition 1.5 *Soit $Z = (X,Y)$ un vecteur aléatoire gaussien de dimension $m + d$, de moyenne $\bar{Z} = (\bar{X},\bar{Y})$ et de matrice de covariance*

$$Q_Z = \begin{pmatrix} Q_X & Q_{XY} \\ \\ Q_{YX} & Q_Y \end{pmatrix} \ .$$

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 1.6 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. La majoration $R \leq Q_X$ est évidente, et la minoration $R \geq 0$ résulte de l'identité

$$\begin{aligned} 0 \ \leq & \begin{pmatrix} u^* & -(Q_Y^{-1}\, Q_{YX}\, u)^* \end{pmatrix} \begin{pmatrix} Q_X & Q_{XY} \\ \\ Q_{YX} & Q_Y \end{pmatrix} \begin{pmatrix} u \\ \\ -Q_Y^{-1}\, Q_{YX}\, u \end{pmatrix} \\ = & \begin{pmatrix} u^* & -(Q_Y^{-1}\, Q_{YX}\, u)^* \end{pmatrix} \begin{pmatrix} (Q_X - Q_{XY}\, Q_Y^{-1}\, Q_{YX})\, u \\ \\ 0 \end{pmatrix} = u^*\, R\, u \ , \end{aligned}$$

pour tout vecteur $u \in \mathbb{R}^m$, ce qui permet de conclure que $R \geq 0$. 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 1.7 Soit $\widehat{X} = \widehat{X}(Y)$ l'estimateur MMSE 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 $Z = (X, Y)$.

Remarque 1.8 Si $Y = (Y', Y'')$ où les composantes Y' et Y'' sont indépendantes, alors

$$\bar{Z} = \begin{pmatrix} \bar{X} \\ \\ \bar{Y}' \\ \\ \bar{Y}'' \end{pmatrix} \qquad \text{et} \qquad Q_Z = \begin{pmatrix} Q_X & Q_{XY'} & Q_{XY''} \\ \\ Q_{Y'X} & Q_{Y'} & 0 \\ \\ Q_{Y''X} & 0 & Q_{Y''} \end{pmatrix} \ ,$$

et si les matrices $Q_{Y'}$ et $Q_{Y''}$ sont inversibles, alors la distribution de probabilité conditionnelle du vecteur aléatoire X sachant $Y = y$, avec $y = (y', y'')$, est une distribution de probabilité gaussienne de moyenne

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

et de matrice de covariance

$$\begin{aligned} R &= Q_X - Q_{XY}\, Q_Y^{-1}\, Q_{YX} \\ &= Q_X - \begin{pmatrix} Q_{XY'} & Q_{XY''} \end{pmatrix} \begin{pmatrix} Q_{Y'}^{-1} & 0 \\ 0 & Q_{Y''}^{-1} \end{pmatrix} \begin{pmatrix} Q_{Y'X} \\ Q_{Y''X} \end{pmatrix} \\ &= Q_X - Q_{XY'}\, Q_{Y'}^{-1}\, Q_{Y'X} - Q_{XY''}\, Q_{Y''}^{-1}\, Q_{Y''X} \ . \end{aligned}$$

Exemple 1.9 Soit X et V deux vecteurs aléatoires gaussiens indépendants, de moyenne $\bar{X}$ et 0, et de matrice de covariance Q_X et Q_V, respectivement, et on pose $Y = H\,X + V$. Le vecteur aléatoire $Z = (X, Y)$ est alors gaussien, comme transformatiuon affine du vecteur aléatoire gaussien (X, V), de moyenne $\bar{Z} = (\bar{X}, H\,\bar{X})$ et de matrice de covariance

$$Q_Z = \begin{pmatrix} Q_X & Q_X\, H^* \\ H\, Q_X & H\, Q_X\, H^* + Q_V \end{pmatrix} \ .$$

En effet, par différence

$$Y - H\,\bar{X} = H\,(X - \bar{X}) + V \ ,$$

de sorte que

$$\begin{aligned} Q_{XY} &= \mathbb{E}[\,(X - \bar{X})\,(H\,(X - \bar{X}) + V)^*\,] \\ &= \mathbb{E}[\,(X - \bar{X})\,(X - \bar{X}))^*\,]\,H^* + \mathbb{E}[\,(X - \bar{X})\,V^*\,] \\ &= Q_X\, H^* \ , \end{aligned}$$

et

$$\begin{aligned} Q_Y &= \mathbb{E}[\,(H\,(X - \bar{X}) + V)\,(H\,(X - \bar{X}) + V)^*\,] \\ &= H\,\mathbb{E}[\,(X - \bar{X})\,(X - \bar{X}))^*\,]\,H^* + \mathbb{E}[\,V\,V^*\,] \\ &\qquad + H\,\mathbb{E}[\,(X - \bar{X})\,V^*\,] + \mathbb{E}[\,V\,(X - \bar{X})^*\,]\,H^* \\ &= H\,Q_X\,H^* + Q_V \ . \end{aligned}$$

Pour établir ces deux identités, on a utilisé dans la dernière égalité le fait que $(X - \bar{X})$ est indépendant de V, donc $\mathbb{E}[\,(X - \bar{X})\,V^*] = 0$. Si la matrice Q_V est inversible, alors a fortiori la matrice $Q_Y = H\,Q_X\,H^* + Q_V$ est inversible, et il découle de la Proposition 1.5 que la densité conditionnelle $p_{X|Y}(x)$ du vecteur aléatoire X sachant Y, est une densité gaussienne de moyenne

$$\widehat{X}(Y) = \bar{X} + Q_X\, H^*\,(H\,Q_X\,H^* + Q_V)^{-1}\,(Y - H\,\bar{X}) \ ,$$

et de matrice de covariance déterministe

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

Si en outre la matrice Q_X est inversible, alors il découle du Lemme 1.1 d'inversion matricielle que la matrice R est inversible, et

$$R^{-1} = H^*\, Q_V^{-1}\, H + Q_X^{-1}\ .$$

Preuve de la Proposition 1.5. 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 est égale à

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

c'est–à–dire que la loi conditionnelle du vecteur aléatoire X sachant Y est une loi gaussienne de moyenne $\widehat{X}$ et de matrice de covariance R. On vérifie que

$$\begin{aligned}
&\mathbb{E}[\,\exp\{i\, v^* Y\}\ \exp\{i\, u^* \widehat{X} - \tfrac{1}{2}\, u^* R\, u\}\,] \\
&\quad = \exp\{i\, u^* \bar{X} - i\, u^*\, Q_{XY}\, Q_Y^{-1}\, \bar{Y} - \tfrac{1}{2}\, u^* R\, u\}\ \mathbb{E}[\,\exp\{i\, v^* Y\}\ \exp\{i\, u^*\, Q_{XY}\, Q_Y^{-1}\, Y\}\,] \\
&\quad = \exp\{i\, u^* \bar{X} - i\, u^*\, Q_{XY}\, Q_Y^{-1}\, \bar{Y} - \tfrac{1}{2}\, u^* R\, u\}\ \Phi_Y(v + Q_Y^{-1}\, Q_{YX}\, u) \\
&\quad = \exp\{i\, u^* \bar{X} - i\, u^*\, Q_{XY}\, Q_Y^{-1}\, \bar{Y} - \tfrac{1}{2}\, u^* Q_X\, u + \tfrac{1}{2}\, u^* Q_{XY}\, Q_Y^{-1}\, Q_{YX}\, u \\
&\qquad\qquad + i\,(v^* + u^*\, Q_{XY}\, Q_Y^{-1})\, \bar{Y} - \tfrac{1}{2}\,(v^* + u^*\, Q_{XY}\, Q_Y^{-1})\, Q_Y\, (v + Q_Y^{-1}\, Q_{YX}\, u)\} \\
&\quad = \exp\{i\, u^* \bar{X} + i\, v^* \bar{Y} - \tfrac{1}{2}\, u^* Q_X\, u - u^* Q_{XY}\, v - \tfrac{1}{2}\, v^* Q_Y\, v\} \\
&\quad = \Phi_{X,Y}(u, v) \\
&\quad = \mathbb{E}[\,\exp\{i\, v^* Y\}\ \exp\{i\, u^* X\}\,]\ ,
\end{aligned}$$

et compte tenu que $v \in \mathbb{R}^d$ est arbitraire, on obtient

$$\mathbb{E}[\,\exp\{i\, u^* X\} \mid Y] = \exp\{i\, u^* \widehat{X} - \tfrac{1}{2}\, u^* R\, u\}\ . \qquad \square$$

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.

L'objectif de cette première partie du cours est de fournir des algorithmes efficaces de calcul des distributions de probabilité conditionnelles

$$\mathbb{P}[X_n \in dx \mid Y_0, Y_1, \cdots, Y_n] \ ,$$

dans le cas particulier des systèmes linéaires gaussiens. Dans ce cas, il sera possible de résoudre exactement le problème de filtrage de façon optimale, par la mise en œuvre du filtre (et du lisseur) de Kalman.

Ce cas peut être vu comme un cas particulier de modèles beaucoup plus généraux, comme les systèmes non–linéaires à bruits non–gaussiens, ou les modèles de Markov cachés à espace d'état général. Dans ce cas, il ne sera pas possible de résoudre exactement le problème de filtrage de façon optimale, et il faudra avoir recours à la mise en œuvre de méthodes de résolution approchées, par exemple de filtres particulaires.

Chapitre 2

Filtrage de Kalman

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

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

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). On a vu à la Section 1.2 que la solution consiste à calculer la distribution de probabilité conditionnelle de la variable aléatoire X_k sachant $Y_{0:k}$.

Dans le cas des systèmes décrits à la Section 2.1, le cadre est gaussien et l'évolution de cette distribution de probabilité conditionnelle (déterminée par sa moyenne et sa matrice de covariance) est régie par les équations du filtre de Kalman, présentées à la Section 2.2 et très simples à mettre en œuvre. Dans tous les autres cas, par exemple dans le cas des systèmes non–linéaires avec des bruits non gaussiens, ou dans le cas de modèles encore plus généraux qui seront introduits au Chapitre 4, l'évolution de cette distribution de probabilité conditionnelle est determinée par un tout autre type d'équations, qui seront décrites au Chapitre 5 et dont la mise–en–oeuvre pratique sera présentée à la Section 5.3. Les techniques développées dans le cas linéaire peuvent parfois s'étendre au cas non linéaire par des méthodes de linéarisation, présentées à la Section 3.1. Les filtres ainsi obtenus sont très souvent utilisés en pratique mais ont parfois des performances peu satisfaisantes.

2.1 Systèmes linéaires gaussiens

On considère une suite d'états cachés $\{X_k\}$ à valeurs dans $\mathbb{R}^m$, vérifiant

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

et une suite d'observations $\{Y_k\}$ à valeurs dans $\mathbb{R}^d$, vérifiant

$$Y_k = H_k \, X_k + h_k + V_k \ , \tag{2.2}$$

et on suppose que

- la condition initiale X_0 est gaussienne, de moyenne $\bar{X}_0$ et de matrice de covariance Q_0^X,
- la suite $\{W_k\}$ est un bruit blanc gaussien, de matrice de covariance Q_k^W,
- la suite $\{V_k\}$ est un bruit blanc gaussien, de matrice de covariance Q_k^V,
- les suites $\{W_k\}$ et $\{V_k\}$ et la condition initiale X_0 sont mutuellement indépendants.

La signification du modèle (2.1) est la suivante

- même si l'état $X_{k-1} = x$ est connu exactement à l'instant $(k-1)$, on peut seulement dire que l'état X_k à l'instant k est incertain, et distribué comme un vecteur aléatoire gaussien, de moyenne $F_k\, x + f_k$ et de matrice de covariance Q_k^W,
- si l'état X_{k-1} est incertain à l'instant $(k-1)$, et distribué comme un vecteur aléatoire gaussien, de moyenne $\bar{X}_{k-1}$ et de matrice de covariance Q_{k-1}^X, alors cette incertitude se propage à l'instant k : même en absence de bruit, c'est–à–dire même si $G_k = 0$, l'état X_k à l'instant k est incertain, et distribué comme un vecteur aléatoire gaussien, de moyenne $F_k\, \bar{X}_{k-1} + f_k$ et de matrice de covariance $F_k\, Q_{k-1}^X\, F_k^*$.

Proposition 2.1 *La suite $\{Z_k = (X_k, Y_k)\}$ est une suite gaussienne à valeurs dans $\mathbb{R}^{m+d}$.*

Preuve. Comme sortie d'un système linéaire à entrées gaussiennes, la suite $\{Z_k\}$ est un processus aléatoire gaussien. En effet, pour tout instant n, le vecteur aléatoire $(Z_0, Z_1, \cdots, Z_n)$ peut s'exprimer comme transformation affine du vecteur aléatoire $(X_0, W_1, \cdots, W_n, V_0, V_1, \cdots, V_n)$ qui par hypothèse est un vecteur aléatoire gaussien, donc le vecteur aléatoire $(Z_0, Z_1, \cdots, Z_n)$ est gaussien, comme transformation affine d'un vecteur aléatoire gaussien. □

Remarque 2.2 Si les coefficients dépendent des observations passées, on parle de système *conditionnellement* linéaire gaussien : on considère ainsi une suite d'états cachés $\{X_k\}$ à valeurs dans $\mathbb{R}^m$, vérifiant

$$X_k = F_k(Y_{0:k-1})\, X_{k-1} + f_k(Y_{0:k-1}) + G_k(Y_{0:k-1})\, W_k \ ,$$

où la suite $\{W_k\}$ prend ses valeurs dans $\mathbb{R}^p$, et une suite d'observations $\{Y_k\}$ à valeurs dans $\mathbb{R}^d$, vérifiant

$$Y_k = H_k(Y_{0:k-1})\, X_k + h_k(Y_{0:k-1}) + V_k \ ,$$

et on suppose que

- la condition initiale X_0 est gaussienne, de moyenne $\bar{X}_0$ et de matrice de covariance Q_0^X,
- la suite $\{W_k\}$ est un bruit blanc gaussien, de matrice de covariance identité,
- la suite $\{V_k\}$ est un bruit blanc gaussien, de matrice de covariance Q_k^V,

- les suites $\{W_k\}$ et $\{V_k\}$ et la condition initiale X_0 sont mutuellement indépendants.

Dans ce cas, la suite $\{Z_k = (X_k, Y_k)\}$ n'est en général pas une suite gaussienne, mais on peut vérifier que conditionnellement à $Y_{0:k-1}$

- le vecteur aléatoire $W_k^{\mathrm{CLG}} = G_k(Y_{0:k-1})\, W_k$ est gaussien centré, de matrice de covariance conditionnelle $Q_k^W(Y_{0:k-1}) = G_k(Y_{0:k-1})\, G_k^*(Y_{0:k-1})$,
- le couple (X_k, Y_k) forme conjointement un vecteur aléatoire gaussien.

2.2 Filtre de Kalman

On considère un système linéaire du type (2.1) (2.2), c'est–à–dire

$$X_k \;=\; F_k\, X_{k-1} + f_k + W_k \;, \tag{2.3}$$

$$Y_k \;=\; H_k\, X_k + h_k + V_k \;, \tag{2.4}$$

avec les hypothèses faites à la Section 2.1. A l'instant k, on dispose de l'information

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

L'objectif est d'estimer de façon optimale et récursive le vecteur aléatoire X_k à partir de $Y_{0:k}$. Si on adopte le critère du minimum de variance, il s'agit d'après la Section 1.2 de calculer la distribution de probabilité 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

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

On définit également les quantités suivantes

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

D'après la Remarque 1.6, les matrices de covariances conditionnelles P_k et P_k^- ne dépendent pas des observations, c'est–à–dire que

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

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

- dans l'étape de *prédiction*, on calcule la distribution de probabilité conditionnelle du vecteur aléatoire X_k sachant les observations passées $Y_{0:k-1}$, ce qui est facile à partir de (2.3),

- dans l'étape de *correction*, on utilise la nouvelle observation Y_k, et 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

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

et d'après (2.4), on a

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

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

Remarque 2.3 Par définition, toute fonction des variables $(Y_0, \cdots, Y_{k-1}, Y_k)$ peut s'exprimer en fonction des variables $(Y_0, \cdots, Y_{k-1}, I_k)$, et réciproquement. On en déduit que $(Y_{0:k-1}, I_k)$ contient exactement la même information que $Y_{0:k}$.

Lemme 2.4 *Le processus $\{I_k\}$ est un processus gaussien à valeurs dans $\mathbb{R}^d$, appelé* processus d'innovation. *En particulier, le vecteur aléatoire I_k est gaussien, de moyenne nulle et de matrice de covariance*

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

et indépendant de $Y_{0:k-1}$. Plus généralement, le vecteur aléatoire $(X_k - \widehat{X}_k^-, I_k)$ est gaussien, de moyenne nulle et de matrice de covariance

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

et indépendant de $Y_{0:k-1}$.

Preuve. D'après la Remarque 1.7, l'observation prédite $\mathbb{E}[Y_k \mid Y_{0:k-1}]$ dépend de façon affine des observations passées $(Y_0, Y_1, \cdots, Y_{k-1})$, de sorte que l'innovation I_k dépend de façon affine des observations $(Y_0, Y_1, \cdots, Y_k)$. On en déduit que le vecteur aléatoire $(I_0, I_1, \cdots, I_k)$ est gaussien, comme transformation affine d'un vecteur aléatoire gaussien.

Toujours d'après la Remarque 1.7, l'état prédit $\widehat{X}_k^- = \mathbb{E}[X_k \mid Y_{0:k-1}]$ dépend de façon affine des observations passées $(Y_0, \cdots, Y_{k-1})$, de sorte que le vecteur aléatoire $(Y_0, \cdots, Y_{k-1}, X_k - \widehat{X}_k^-, I_k)$ dépend de façon affine du vecteur $(Y_0, Y_1, \cdots, Y_k, X_k)$ formé de l'état courant X_k et des observations $(Y_0, Y_1, \cdots, Y_k)$. On en déduit que le vecteur aléatoire $(Y_0, \cdots, Y_{k-1}, X_k - \widehat{X}_k^-, I_k)$ est gaussien, et donc a fortiori le vecteur aléatoire $(X_k - \widehat{X}_k^-, I_k)$ est gaussien, comme transformation affine d'un vecteur aléatoire gaussien. Compte tenu que

$$\mathbb{E}[X_k - \widehat{X}_k^- \mid Y_{0:k-1}] = 0 \qquad \text{et} \qquad \mathbb{E}[I_k \mid Y_{0:k-1}] = 0 \ ,$$

par définition, on en déduit que le vecteur aléatoire $(X_k - \widehat{X}_k^-, I_k)$ est indépendant de $Y_{0:k-1}$. D'après l'équation (2.4), on a

$$I_k = Y_k - (H_k \, \widehat{X}_k^- + h_k) = H_k \, (X_k - \widehat{X}_k^-) + V_k \ , \tag{2.5}$$

et on en déduit que

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

Dans cette dernière égalité, on a utilisé le fait que $(X_k - \widehat{X}_k^-)$ est indépendant de V_k, donc $\mathbb{E}[(X_k - \widehat{X}_k^-)\,V_k^*] = 0$. On déduit également de (2.5) que

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

Dans cette dernière égalité, on a de nouveau 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 2.5 Si la matrice de covariance Q_k^V est inversible, alors a fortiori la matrice de covariance $Q_k^I = H_k\,P_k^-\,H_k^* + Q_k^V$ est inversible, pour tout instant k.

Remarque 2.6 Compte tenu que la distribution de probabilité conditionnelle du vecteur aléatoire Y_k sachant $Y_{0:k-1}$ est gaussienne, de moyenne $H_k\,\widehat{X}_k^- + h_k$ et de matrice de covariance Q_k^I, et pourvu que la matrice Q_k^I soit inversible, on obtient l'expression suivante

$$\begin{aligned} L_n &= \prod_{k=0}^{n} \exp\{-\tfrac{1}{2}\,(Y_k - (H_k\,\widehat{X}_k^- + h_k))^*\,(Q_k^I)^{-1}\,(Y_k - (H_k\,\widehat{X}_k^- + h_k))\,\} \\ &= \prod_{k=0}^{n} \exp\{-\tfrac{1}{2}\,I_k^*\,(Q_k^I)^{-1}\,I_k\,\}\ , \end{aligned}$$

pour la vraisemblance du modèle, à une constante multiplicative près.

Théorème 2.7 (Filtre de Kalman) *On suppose que la matrice de covariance Q_k^V est* inversible*, pour tout instant k. Alors les suites $\{\widehat{X}_k\}$ et $\{P_k\}$ vérifient les équations récurrentes suivantes*

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

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

et

$$\widehat{X}_k \;=\; \widehat{X}_k^- + K_k\,(Y_k - (H_k\,\widehat{X}_k^- + h_k))\;,$$

$$P_k \;=\; (I - K_k\,H_k)\;P_k^-\;,$$

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*

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

Remarque 2.8 Au vu de l'expression développée

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

on vérifie aisément que $P_k \leq P_k^-$, c'est–à–dire que la matrice de covariance de l'erreur de filtrage est plus petite (au sens des matrices symétriques) que la matrice de covariance de l'erreur de prédiction, pour tout instant k.

Remarque 2.9 On vérifie que la suite $\{P_k\}$ ne dépend pas des observations : elle peut donc être pré–calculée, en particulier dans le cas simple où les coefficients $F_k = F$, $H_k = H$, $Q_k^W = Q_W$ et $Q_k^V = Q_V$ sont constants.

Remarque 2.10 Si les coefficients F_k, f_k et Q_k^W, et les coefficients H_k et h_k dépendent des observations passées $Y_{0:k-1}$, on a indiqué à la Remarque 2.2 que conditionnellement à $Y_{0:k-1}$ le couple (X_k, Y_k) forme conjointement un vecteur aléatoire gaussien, et on peut vérifier que la distribution de probabilité conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$ est gaussienne, de moyenne $\widehat{X}_k$ et de matrice de covariance P_k données par les équations du Théorème 2.7 avec des coefficients dépendant des observations.

Preuve. On procède en plusieurs étapes. Le point central est la Proposition 1.5 qui sera constamment utilisée.

Expression de $\widehat{X}_0$ 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

$$\begin{pmatrix} \widehat{X}_0^- \\ \\ H_0\,\widehat{X}_0^- + h_0 \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} P_0^- & P_0^-\;H_0^* \\ \\ H_0\,P_0^- & H_0\,P_0^-\;H_0^* + Q_0^V \end{pmatrix}\;,$$

respectivement. D'après la Proposition 1.5, la distribution de probabilité conditionnelle du vecteur aléatoire X_0 sachant Y_0 est gaussienne, de moyenne

$$\widehat{X}_0 = \widehat{X}_0^- + P_0^-\;H_0^*\;[H_0\,P_0^-\;H_0^* + Q_0^V]^{-1}\;(Y_0 - (H_0\,\widehat{X}_0^- + h_0))\;,$$

et de matrice de covariance

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

Expression de $\widehat{X}_k^-$ 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 1.5, la distribution de probabilité conditionnelle du vecteur aléatoire 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 (2.3), c'est–à–dire

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

on a

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

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

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

de sorte que

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

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

Expression de $\widehat{X}_k$ 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 1.5, la distribution de probabilité conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$ est gaussienne, de moyenne $\widehat{X}_k$ et de matrice de covariance déterministe P_k. D'après la Remarque 2.3, on a

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

compte tenu que les vecteurs aléatoires $(X_k - \widehat{X}_k^-)$ et I_k sont indépendants de $Y_{0:k-1}$, d'après le Lemme 2.4. Par différence

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

de sorte que

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

Pour calculer la moyenne conditionnelle et la matrice de covariance conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$, 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. En d'autres termes, pour estimer l'état caché X_k au vu des observations $Y_{0:k}$ il suffit d'estimer de quelle quantité, exprimée en fonction de l'écart I_k constaté entre la nouvelle observation et l'observation prédite, corriger l'estimation prédite $\widehat{X}_k^-$. C'est de cette propriété que découle la forme récursive du filtre de Kalman. D'après le Lemme 2.4, le vecteur aléatoire $(X_k - \widehat{X}_k^-, I_k)$ est gaussien, de moyenne nulle et de matrice de covariance

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

Si la matrice Q_k^V est inversible, alors a fortiori la matrice $Q_k^I = H_k\,P_k^-\,H_k^* + Q_k^V$ est inversible, et d'après la Proposition 1.5 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. □

2.3 Lisseur de Kalman

On dispose désormais de l'information

$$Y_{0:n} = (Y_0, Y_1, \cdots, Y_n) \ ,$$

et l'objectif est d'estimer de façon optimale le vecteur aléatoire X_k à partir de $Y_{0:n}$, pour un instant k intermédiaire entre l'instant initial 0 et l'instant final n. Si on adopte le critère du minimum de variance, il s'agit d'après la Section 1.2 de calculer la distribution de probabilité conditionnelle du vecteur aléatoire X_k sachant $Y_{0:n}$. Comme le cadre est gaussien, il suffit de calculer la moyenne et la matrice de covariance

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

et clairement, $\widehat{X}_n^n = \widehat{X}_n$ et $P_n^n = P_n$ pour $k = n$. D'après la Remarque 1.6, la matrice de covariance conditionnelle P_k^n ne dépend pas des observations, c'est–à–dire que

$$P_k^n = \mathbb{E}[(X_k - \widehat{X}_k^n)\,(X_k - \widehat{X}_k^n)^*] \ .$$

Théorème 2.11 (Lisseur de Kalman (formulation de Rauch–Tung–Striebel)) *On suppose que les matrices de covariance Q_k^W et Q_k^V sont* inversibles*, pour tout instant k. Alors les suites $\{\widehat{X}_k^n\}$ et $\{P_k^n\}$ vérifient les équations récurrentes rétrogrades suivantes*

$$\widehat{X}_{k-1}^n \;=\; \widehat{X}_{k-1} + L_k\,(\widehat{X}_k^n - \widehat{X}_k^-)\;,$$

$$P_{k-1}^n \;=\; P_{k-1} + L_k\,(P_k^n - P_k^-)\,L_k^*\;,$$

avec la matrice de gain

$$L_k = P_{k-1}\,F_k^*\,(P_k^-)^{-1}\;,$$

et avec les initialisations

$$\widehat{X}_n^n = \widehat{X}_n \qquad et \qquad P_n^n = P_n\;.$$

Remarque 2.12 Au vu de l'expression développée

$$P_{k-1} - L_k\,P_k^-\,L_k^* = P_{k-1} - P_{k-1}\,F_k^*\,[F_k\,P_{k-1}\,F_k^* + Q_k^W]^{-1}\,F_k\,P_{k-1}\;,$$

on vérifie que la matrice $P_{k-1} - L_k\,P_k^-\,L_k^*$ est semi–définie positive, pour tout instant k. On en déduit par récurrence arrière que la matrice P_k^n (telle qu'elle est définie par l'équation rétrograde de l'énoncé) est semi–définie positive, pour tout instant k. Par définition, $P_n^n = P_n$, c'est–à–dire que la relation est vraie au rank $k = n$. Si la relation est vraie au rang k, c'est–à–dire si la matrice P_k^n est semi–définie positive, alors nécessairement la matrice

$$P_{k-1}^n = P_{k-1} + L_k\,(P_k^n - P_k^-)\,L_k^* = (P_{k-1} - L_k\,P_k^-\,L_k^*) + L_k\,P_k^n\,L_k^*\;,$$

aussi est semi–définie positive, c'est–à–dire que la relation est vraie au rang $(k-1)$.

Remarque 2.13 On vérifie par récurrence arrière que $P_k^n \leq P_k$, c'est–à–dire que la matrice de covariance de l'erreur de lissage est plus petite (au sens des matrices symétriques) que la matrice de covariance de l'erreur de filtrage, pour tout instant k. Par définition $P_n^n = P_n$, c'est–à–dire que la relation est vraie au rank $k = n$. Si la relation est vraie au rang k, c'est–à–dire si $P_k^n \leq P_k$, alors nécessairement $P_k^n \leq P_k^-$ compte tenu que $P_k \leq P_k^-$ d'après la Remarque 2.8. En d'autres termes, la différence $(P_k^n - P_k^-)$ est semi–définie négative, de sorte que la différence

$$P_{k-1}^n - P_{k-1} = L_k\,(P_k^n - P_k^-)\,L_k^*\;,$$

aussi est semi–définie négative. En d'autres termes, $P_{k-1}^n \leq P_{k-1}$, c'est–à–dire que la relation est vraie au rang $(k-1)$.

Preuve. On remarque que le vecteur aléatoire Y_k peut s'exprimer comme transformation affine du vecteur aléatoire (X_k, V_k), et donc a fortiori comme transformation affine du vecteur aléatoire $(Y_{0:k-1}, X_k - \widehat{X}_k^-, V_k)$. De même, le vecteur aléatoire Y_{k+p} peut s'exprimer comme transformation affine du vecteur aléatoire (X_{k+p}, V_{k+p}), et par transitivité comme transformation affine du vecteur aléatoire $(X_k, W_{k+1}, \cdots, W_{k+p}, V_{k+p})$, et donc a fortiori comme transformation affine du vecteur aléatoire $(Y_{0:k-1}, X_k - \widehat{X}_k^-, W_{k+1}, \cdots, W_{k+p}, V_{k+p})$. On en déduit que le vecteur aléatoire $Y_{0:n} = (Y_{0:k-1}, Y_k, \cdots, Y_n)$ peut s'exprimer comme transformation affine du vecteur

aléatoire $(Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n})$ où $Z_{k+1:n} = (W_{k+1}, \cdots, W_n, V_k, V_{k+1}, \cdots, V_n)$ par définition. Les vecteurs aléatoires $Y_{0:k-1}$, $X_k - \widehat{X}_k^-$ et $Z_{k+1:n}$ sont mutuellement indépendants, et il résulte de la Remarque 1.8 que

$$\begin{aligned} U_{k-1}^n &= \mathbb{E}[X_{k-1} \mid Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n}] \\ &= \widehat{X}_{k-1} + \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n}] \\ &= \widehat{X}_{k-1} + \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid Y_{0:k-1}] + \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid X_k - \widehat{X}_k^-] \\ &\qquad + \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid Z_{k+1:n}] \\ &= \widehat{X}_{k-1} + \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid X_k - \widehat{X}_k^-] \ , \end{aligned}$$

compte tenu que $\mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid Y_{0:k-1}] = 0$ par définition, et où on a utilisé dans la dernière égalité le fait que $(X_{k-1} - \widehat{X}_{k-1})$ est indépendant de $Z_{k+1:n}$, donc $\mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid Z_{k+1:n}] = 0$. Par différence

$$\begin{aligned} X_{k-1} - U_{k-1}^n &= (X_{k-1} - \widehat{X}_{k-1}) - (U_{k-1}^n - \widehat{X}_{k-1}) \\ &= (X_{k-1} - \widehat{X}_{k-1}) - \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid X_k - \widehat{X}_k^-] \ , \end{aligned}$$

de sorte que

$$\begin{aligned} &\mathbb{E}[(X_{k-1} - U_{k-1}^n)\,(X_{k-1} - U_{k-1}^n)^*] \\ &\quad = \mathbb{E}[\,((X_{k-1} - \widehat{X}_{k-1}) - \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid X_k - \widehat{X}_k^-]) \\ &\qquad ((X_{k-1} - \widehat{X}_{k-1}) - \mathbb{E}[X_{k-1} - \widehat{X}_{k-1} \mid X_k - \widehat{X}_k^-])^*\,] \ . \end{aligned}$$

Pour calculer la moyenne conditionnelle et la matrice de covariance conditionnelle du vecteur aléatoire X_{k-1} sachant $(Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n})$, il suffit donc de calculer la moyenne conditionnelle et la matrice de covariance conditionnelle du vecteur aléatoire $(X_{k-1} - \widehat{X}_{k-1})$ sachant $(X_k - \widehat{X}_k^-)$. D'après la Remarque 1.7, l'état estimé $\widehat{X}_{k-1} = \mathbb{E}[X_{k-1} \mid Y_{0:k-1}]$ et l'état prédit $\widehat{X}_k^- = \mathbb{E}[X_k \mid Y_{0:k-1}]$ dépendent de façon affine des observations passées $(Y_0, \cdots, Y_{k-1})$, de sorte que le vecteur aléatoire $(X_{k-1} - \widehat{X}_{k-1}, X_k - \widehat{X}_k^-)$ dépend de façon affine du vecteur aléatoire $(Y_0, \cdots, Y_{k-1}, X_{k-1}, X_k)$. On en déduit que le vecteur aléatoire $(X_{k-1} - \widehat{X}_{k-1}, X_k - \widehat{X}_k^-)$ est gaussien, comme transformation affine d'un vecteur aléatoire gaussien. Par différence

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

de sorte que

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

Dans cette dernière égalité, on a utilisé le fait que $(X_{k-1} - \widehat{X}_{k-1})$ et W_k sont indépendants, donc $\mathbb{E}[(X_{k-1} - \widehat{X}_{k-1})\,W_k^*] = 0$. On en déduit que le vecteur aléatoire gaussien $(X_{k-1} - \widehat{X}_{k-1}, X_k - \widehat{X}_k^-)$ est de moyenne nulle et de matrice de covariance

$$\begin{pmatrix} P_{k-1} & P_{k-1}\,F_k^* \\ F_k\,P_{k-1} & P_k^- \end{pmatrix}\;.$$

Par hypothèse la matrice P_k^- est inversible, et d'après la Proposition 1.5 on a immédiatement

$$U_{k-1}^n = \widehat{X}_{k-1} + P_{k-1}\,F_k^*\,(P_k^-)^{-1}\,(X_k - \widehat{X}_k^-) = \widehat{X}_{k-1} + L_k\,(X_k - \widehat{X}_k^-)\;,$$

et

$$\mathbb{E}[(X_{k-1} - U_{k-1}^n)\,(X_{k-1} - U_{k-1}^n)^*] = P_{k-1} - P_{k-1}\,F_k^*\,(P_k^-)^{-1}\,F_k\,P_{k-1} = P_{k-1} - L_k\,P_k^-\,L_k^*\;.$$

On rappelle que $(Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n})$ contient davantage d'information que $Y_{0:n}$, de sorte que

$$\widehat{X}_{k-1}^n = \mathbb{E}[X_{k-1} \mid Y_{0:n}] = \mathbb{E}[U_{k-1}^n \mid Y_{0:n}] = \widehat{X}_{k-1} + L_k\,(\widehat{X}_k^n - \widehat{X}_k^-)\;.$$

Par différence

$$X_{k-1} - \widehat{X}_{k-1}^n = (X_{k-1} - U_{k-1}^n) + (U_{k-1}^n - \widehat{X}_{k-1}^n) \qquad \text{et} \qquad U_{k-1}^n - \widehat{X}_{k-1}^n = L_k\,(X_k - \widehat{X}_k^n)\;,$$

de sorte que

$$\begin{aligned}
P_{k-1}^n \;&=\; \mathbb{E}[\,(X_{k-1} - \widehat{X}_{k-1}^n)\,(X_{k-1} - \widehat{X}_{k-1}^n)^*\,] \\
&=\; \mathbb{E}[\,((X_{k-1} - U_{k-1}^n) + (U_{k-1}^n - \widehat{X}_{k-1}^n))\,((X_{k-1} - U_{k-1}^n) + (U_{k-1}^n - \widehat{X}_{k-1}^n))^*\,] \\
&=\; \mathbb{E}[\,(X_{k-1} - U_{k-1}^n)\,(X_{k-1} - U_{k-1}^n)^*\,] + \mathbb{E}[\,(U_{k-1}^n - \widehat{X}_{k-1}^n)\,(U_{k-1}^n - \widehat{X}_{k-1}^n)^*\,] \\
&\qquad + \mathbb{E}[\,(U_{k-1}^n - \widehat{X}_{k-1}^n)\,(X_{k-1} - U_{k-1}^n)^*\,] + \mathbb{E}[\,(X_{k-1} - U_{k-1}^n)\,(U_{k-1}^n - \widehat{X}_{k-1}^n)^*\,] \\
&=\; (P_{k-1} - L_k\,P_k^-\,L_k^*) + L_k\,P_k^n\,L_k^*\;.
\end{aligned}$$

Dans cette dernière égalité, on a utilisé le fait que

- $(U_{k-1}^n - \widehat{X}_{k-1}^n)$ dépend de $(Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n})$,
- et $\mathbb{E}[X_{k-1} - U_{k-1}^n \mid Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n}] = 0$ par définition,

donc $\mathbb{E}[\,(X_{k-1} - U_{k-1}^n)\,(U_{k-1}^n - \widehat{X}_{k-1}^n)^*\,] = 0$. □

Proposition 2.14 *La matrice de corrélation C_k^n entre les erreurs de lissage $(X_{k-1} - \widehat{X}_{k-1}^n)$ et $(X_k - \widehat{X}_k^n)$ à deux instants successifs vérifie la relation suivante*

$$C_k^n = \mathbb{E}[\,(X_{k-1} - \widehat{X}_{k-1}^n)\,(X_k - \widehat{X}_k^n)^*\,] = L_k\,P_k^n\ .$$

Preuve. On rappelle que

$$X_{k-1} - \widehat{X}_{k-1}^n = (X_{k-1} - U_{k-1}^n) + L_k\,(X_k - \widehat{X}_k^n)\ ,$$

de sorte que

$$\begin{aligned} C_k^n &= \mathbb{E}[\,(X_{k-1} - \widehat{X}_{k-1}^n)\,(X_k - \widehat{X}_k^n)^*\,] \\ &= \mathbb{E}[\,(X_{k-1} - U_{k-1}^n)\,(X_k - \widehat{X}_k^n)^*\,] + L_k\,\mathbb{E}[\,(X_k - \widehat{X}_k^n)\,(X_k - \widehat{X}_k^n)^*\,] \\ &= L_k\,P_k^n\ . \end{aligned}$$

Dans cette dernière égalité, on a utilisé le fait que

- $(X_k - \widehat{X}_k^n) = (X_k - \widehat{X}_k^-) + (\widehat{X}_k^- - \widehat{X}_{k-1}^n)$ dépend de $(Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n})$,
- et $\mathbb{E}[X_{k-1} - U_{k-1}^n \mid Y_{0:k-1}, X_k - \widehat{X}_k^-, Z_{k+1:n}] = 0$ par définition,

donc $\mathbb{E}[\,(X_{k-1} - U_{k-1}^n)\,(X_k - \widehat{X}_k^n)^*\,] = 0$. □

Il existe plusieurs formulations équivalentes pour le lissage de Kalman, et on présente ci–dessous une formulation alternative, qui ne fait pas l'hypothèse que la matrice de covariance Q_k^W est inversible, et qui n'utilise pas l'inverse de la matrice de covariance P_k^-.

Pour tout $k = 1, \cdots, n$, on introduit les variables

$$r_{k-1} = F_k^*\,(P_k^-)^{-1}\,(\widehat{X}_k^n - \widehat{X}_k^-) \qquad \text{et} \qquad \Pi_{k-1} = -F_k^*\,(P_k^-)^{-1}\,(P_k^n - P_k^-)\,(P_k^-)^{-1}\,F_k\ ,$$

et on pose $r_n = 0$ et $\Pi_n = 0$ par convention. On rappelle que la différence $(P_k^n - P_k^-)$ est semi–définie négative, de sorte que la matrice Π_{k-1} est semi–définie positive. Clairement

$$\begin{aligned} \widehat{X}_k^n &= \widehat{X}_k + L_{k+1}\,(\widehat{X}_{k+1}^n - \widehat{X}_{k+1}^-) \\ &= \widehat{X}_k + P_k\,F_{k+1}^*\,(P_{k+1}^-)^{-1}\,(\widehat{X}_{k+1}^n - \widehat{X}_{k+1}^-) \\ &= \widehat{X}_k + P_k\,r_k\ , \end{aligned}$$

et de même

$$\begin{aligned} P_k^n &= P_k + L_{k+1}\,(P_{k+1}^n - P_{k+1}^-)\,L_{k+1}^* \\ &= P_k + P_k\,F_{k+1}^*\,(P_{k+1}^-)^{-1}\,(P_{k+1}^n - P_{k+1}^-)\,(P_{k+1}^-)^{-1}\,F_{k+1}\,P_k \\ &= P_k - P_k\,\Pi_k\,P_k\ , \end{aligned}$$

de sorte que le lisseur de Kalman $\widehat{X}_k^n$ et la matrice de covariance d'erreur de lissage P_k^n s'expriment comme

$$\widehat{X}_k^n = \widehat{X}_k + P_k\,r_k \qquad \text{et} \qquad P_k^n = P_k - P_k\,\Pi_k\,P_k\ , \tag{2.6}$$

en fonction du filtre de Kalman $\widehat{X}_k$, de la matrice de covariance d'erreur de filtrage P_k, et des variables r_k et Π_k, respectivement. On pose

$$\Xi_k = [H_k\,P_k^-\,H_k^* + Q_k^V]^{-1} \qquad \text{de sorte que} \qquad K_k = P_k^-\,H_k^*\,\Xi_k\ ,$$

pour tout $k = 0, 1, \cdots, n$.

Théorème 2.15 (Lisseur de Kalman (formulation de Fraser–Potter)) *On suppose que la matrice de covariance Q_k^V est* inversible*, pour tout instant k. Alors les suites $\{\widehat{X}_k^n\}$ et $\{P_k^n\}$ sont données par les expressions suivantes*

$$\widehat{X}_k^n = \widehat{X}_k + P_k\,r_k \qquad \textit{et} \qquad P_k^n = P_k - P_k\,\Pi_k\,P_k\ ,$$

où les suites $\{r_k\}$ et $\{\Pi_k\}$ vérifient les équations récurrentes rétrogrades suivantes

$$\begin{aligned} r_k^- &= (I - K_k\,H_k)^*\,r_k + H_k^*\,\Xi_k\,(Y_k - (H_k\,\widehat{X}_k^- + h_k))\ , \\ \Pi_k^- &= (I - K_k\,H_k)^*\,\Pi_k\,(I - K_k\,H_k) + H_k^*\,\Xi_k\,H_k\ , \end{aligned}$$

et

$$r_{k-1} = F_k^*\,r_k^- \qquad \textit{et} \qquad \Pi_{k-1} = F_k^*\,\Pi_k^-\,F_k\ ,$$

avec les initialisations

$$r_n = 0 \qquad \textit{et} \qquad \Pi_n = 0\ .$$

Preuve. On rappelle que

$$P_k = (I - K_k\,H_k)\,P_k^- = P_k^-\,(I - K_k\,H_k)^*\ ,$$

de sorte que

$$P_k\,(P_k^-)^{-1} = I - K_k\,H_k \qquad \text{et} \qquad (P_k^-)^{-1}\,P_k = (I - K_k\,H_k)^*\ , \tag{2.7}$$

et par définition

$$K_k = P_k^-\,H_k^*\,\Xi_k\ ,$$

de sorte que

$$(P_k^-)^{-1}\, K_k = H_k^*\, \Xi_k \ . \tag{2.8}$$

D'après l'étape de correction du filtre de Kalman, on a

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

de sorte que

$$\widehat{X}_k^n - \widehat{X}_k^- = \widehat{X}_k - \widehat{X}_k^- + P_k\, r_k = K_k\, (Y_k - (H_k\, \widehat{X}_k^- + h_k)) + P_k\, r_k \ ,$$

et en reportant cette expression dans la définition de la variable r_{k-1}, on obtient

$$\begin{aligned} r_{k-1} &= F_k^*\, (P_k^-)^{-1}\, (\widehat{X}_k^n - \widehat{X}_k^-) \\ &= F_k^*\, (P_k^-)^{-1}\, [\, P_k\, r_k + K_k\, (Y_k - (H_k\, \widehat{X}_k^- + h_k))\,] \\ &= F_k^*\, [\, (I - K_k\, H_k)^*\, r_k + H_k^*\, \Xi_k\, (Y_k - (H_k\, \widehat{X}_k^- + h_k))\,] \ , \end{aligned}$$

compte tenu des identités (2.7) et (2.8). D'après l'étape de correction du filtre de Kalman, on a

$$P_k = P_k^- - P_k^-\, H_k^*\, \Xi_k\, H_k\, P_k^- \ ,$$

de sorte que

$$P_k^n - P_k^- = P_k - P_k^- - P_k\, \Pi_k\, P_k = -P_k^-\, H_k^*\, \Xi_k\, H_k\, P_k^- - P_k\, \Pi_k\, P_k \ ,$$

et en reportant cette expression dans la définition de la variable Π_{k-1}, on obtient

$$\begin{aligned} \Pi_{k-1} &= -F_k^*\, (P_k^-)^{-1}\, (P_k^n - P_k^-)\, (P_k^-)^{-1}\, F_k \\ &= F_k^*\, (P_k^-)^{-1}\, [\, P_k\, \Pi_k\, P_k + P_k^-\, H_k^*\, \Xi_k\, H_k\, P_k^-\,]\, (P_k^-)^{-1}\, F_k \\ &= F_k^*\, [\, (I - K_k\, H_k)^*\, \Pi_k\, (I - K_k\, H_k) + H_k^*\, \Xi_k\, H_k\,]\, F_k \ , \end{aligned}$$

compte tenu de l'identité (2.8). □

Les deux formulations partagent la même phase aller, qui comprend le calcul du filtre de Kalman $\widehat{X}_k$ et de la matrice de covariance d'erreur de filtrage P_k. Une condition nécessaire pour cette phase aller est l'inversibilité de la matrice de covariance $Q_k^I = H_k\, P_k^-\, H_k^* + Q_k^V$ de dimension $d \times d$, et une condition suffisante est l'inversibilité de la matrice de covariance Q_k^V, une donnée du problème. Le calcul de la matrice inverse n'est pas nécessaire, mais la résolution de systèmes linéaires de dimension d de la forme $Q_k^I\, y = b$ est requise, et passe par exemple par la décomposition de Cholesky de la matrice Q_k^I.

Dans la formulation de Rauch–Tung–Striebel, qui fait l'objet du Théorème 2.11, une condition nécessaire pour la phase retour est l'inversibilité de la matrice de covariance $P_k^- = F_k\, P_{k-1}\, F_k^* + Q_k^W$ de dimension $m \times m$, et une condition suffisante est l'inversibilité de la matrice de covariance Q_k^W, une donnée du problème. Le calcul de la matrice inverse n'est pas nécessaire, mais la

résolution de systèmes linéaires de dimension m de la forme $P_k^- \, x = b$ est requise, et passe par exemple par la décomposition de Cholesky de la matrice P_k^-. L'équation récurrente rétrograde pour le calcul du lisseur $\widehat{X}_{k-1}^n$ utilise les valeurs numériques du filtre $\widehat{X}_{k-1}$ et de la matrice de covariance d'erreur de filtrage P_{k-1} (à partir desquelles il est facile de reconstruire les valeurs numériques du prédicteur $\widehat{X}_k^-$ et de la matrice de covariance d'erreur de prédiction P_k^-). Ces valeurs numériques sont calculées dans la phase aller, et doivent donc être conservées en mémoire pour être utilisées dans la phase retour. En revanche, cette équation récurrente rétrograde pour le calcul du lisseur $\widehat{X}_{k-1}^n$ n'utilise ni la valeur numérique de l'observation Y_k ni celle de l'innovation $I_k = Y_k - (H_k \, \widehat{X}_k^- + h_k)$.

Dans la formulation de Fraser–Potter, qui fait l'objet du Théorème 2.15, il n'y a pas de condition nécessaire d'inversibilité pour la phase retour qui ne soit pas déjà nécessaire pour la phase aller. Les expressions (2.6) pour le lisseur $\widehat{X}_k^n$ et pour la matrice de covariance d'erreur de lissage P_k^n utilisent les valeurs numériques du filtre $\widehat{X}_k$ et de la matrice de covariance d'erreur de filtrage P_k. Ces valeurs numériques sont calculées dans la phase aller, et doivent donc être conservées en mémoire pour être utilisées dans la phase retour. L'équation récurrente rétrograde pour le calcul de la variable r_k utilise la valeur numérique de l'observation Y_k ou de manière équivalente celle de l'innovation $I_k = Y_k - (H_k \, \widehat{X}_k^- + h_k)$. Ces valeurs numériques sont calculées dans la phase aller, et doivent donc être conservées en mémoire pour être utilisées dans la phase retour.

En conclusion :

- les deux formulations requièrent dans la phase aller une même condition d'inversibilité et l'inversion de systèmes linéaires de dimension d,
- la formulation de Rauch–Tung–Striebel requiert dans la phase retour une condition d'inversibilité supplémentaire et l'inversion de systèmes linéaires de dimension m, tandis que la formulation de Fraser–Potter ne requiert aucune condition d'inversibilité supplémentaire,
- les deux formulations utilisent dans la phase retour les valeurs numériques du filtre et de la matrice de covariance d'erreur de filtrage — ces valeurs numériques sont calculées dans la phase aller, et doivent donc être conservées en mémoire pour être utilisées dans la phase retour,
- la formulation de Fraser–Potter utilise dans la phase retour la valeur numérique de l'observation ou de manière équivalente celle de l'innovation, tandis que la formulation de Rauch–Tung–Striebel n'utilise aucune de ces valeurs numériques — ces valeurs numériques sont calculées dans la phase aller, et doivent donc être conservées en mémoire pour être utilisées dans la phase retour.

Remarque 2.16 Il est également possible d'obtenir une équation récurrente pour le lisseur, dans le sens direct (et pas dans le sens rétrograde) et autonome (ne faisant pas intervenir ni le filtre ni la matrice de covariance de l'erreur de filtrage). Par différence, on obtient

$$\begin{aligned} \widehat{X}_k^n - F_k \, \widehat{X}_{k-1}^n \; &= \; \widehat{X}_k + P_k \, r_k - F_k \, (\widehat{X}_{k-1} + P_{k-1} \, r_{k-1}) \\ &= \; (\widehat{X}_k - F_k \, \widehat{X}_{k-1}) + (P_k \, r_k - F_k \, P_{k-1} \, r_{k-1}) \; . \end{aligned}$$

D'après l'étape de correction du filtre de Kalman, on a

$$\widehat{X}_k - F_k\,\widehat{X}_{k-1} = \widehat{X}_k - \widehat{X}_k^- = K_k\,(Y_k - (H_k\,\widehat{X}_k^- + h_k))\ ,$$

et on remarque que

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

compte tenu des identités

$$P_k^-\,(I - K_k\,H_k)^* = P_k \qquad \text{et} \qquad P_k^-\,H_k^*\,\Xi_k = K_k\ ,$$

de sorte que

$$(\widehat{X}_k - F_k\,\widehat{X}_{k-1}) + (P_k\,r_k - P_k^-\,r_k^-) = 0\ .$$

D'autre part

$$\begin{aligned} P_k\,r_k - F_k\,P_{k-1}\,r_{k-1} \ &= \ P_k\,r_k - F_k\,P_{k-1}\,F_k^*\,r_k^- \\ &= \ P_k\,r_k - (P_k^- - Q_k^W)\,r_k^- \\ &= \ (P_k\,r_k - P_k^-\,r_k^-) + Q_k^W\,r_k^-\ . \end{aligned}$$

On en déduit que

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

c'est–à–dire qu'on obtient une équation récurrente, dans le sens direct, et faisant seulement intervenir la variable r_k^-.

Chapitre 3

Extensions aux systèmes non–linéaires

On considère une suite d'états cachés $\{X_k\}$ à valeurs dans $\mathbb{R}^m$, vérifiant

$$X_k = b_k(X_{k-1}) + \sigma_k(X_{k-1})\, W_k \ , \tag{3.1}$$

où $\{W_k\}$ prend ses valeurs dans $\mathbb{R}^p$, et une suite d'observations $\{Y_k\}$ à valeurs dans $\mathbb{R}^d$, vérifiant

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

et on suppose que

- la condition initiale X_0 est gaussienne, de moyenne $\bar{X}_0$ et de matrice de covariance Q_0^X,
- la suite $\{W_k\}$ est un bruit blanc gaussien, de matrice de covariance identité,
- la suite $\{V_k\}$ est un bruit blanc gaussien, de matrice de covariance Q_k^V inversible,
- les suites $\{W_k\}$ et $\{V_k\}$ et la condition initiale X_0 sont mutuellement indépendants.

La signification du modèle (3.1) est la suivante

- même si l'état $X_{k-1} = x$ est connu exactement à l'instant $(k-1)$, on peut seulement dire que l'état X_k à l'instant k est incertain, et distribué comme un vecteur aléatoire gaussien, de moyenne $b_k(x)$ et de matrice de covariance $\sigma_k(x)\, \sigma_k^*(x)$.

La plupart des propriétés obtenues à la Section 2.1 ne sont pas vraies pour le système décrit par les équations (3.1) et (3.2). En particulier, le processus $\{Z_k = (X_k, Y_k)\}$ n'est pas gaussien (ni même conditionnellement gaussien), et les moments conditionnels de X_k sachant $Y_{0:k}$ ne peuvent pas être calculés de manière simple. Deux approches pragmatiques sont présentées dans ce chapitre, qui permettent d'obtenir des estimateurs sous–optimaux, c'est–à–dire qui n'atteignent pas nécessairement le minimum de l'erreur quadratique moyenne, mais qui sont néanmoins très largement utilisés en pratique. La première approche présentée à la Section 3.1 repose sur des

techniques de linéarisation, et donne lieu au filtre de Kalman linéarisé et au filtre de Kalman étendu. La deuxième approche présentée à la Section 3.2 repose sur des techniques d'approximation gaussienne et de quadrature numérique, et donne lieu au filtre de Kalman dit *unscented*. Dans les chapitres suivants, on abandonnera ce point de vue de lineéarisation ou d'approximation gaussienne, et on s'attachera d'abord à caractériser la distribution de probabilité conditionnelle de l'état caché sachant les observations, soit par une représentation probabiliste, soit par une équation récurrente dans l'espace des distributions de probabilité, et on proposera ensuite des approximations numériques reposant sur méthodes de simulation de type Monte Carlo.

3.1 Filtre de Kalman linéarisé, filtre de Kalman étendu

On considère le système non linéaire

$$\begin{aligned} X_k &= b_k(X_{k-1}) + \sigma_k(X_{k-1})\, W_k \ , \\ Y_k &= h_k(X_k) + V_k \ , \end{aligned} \tag{3.3}$$

et on suppose que les fonctions b_k et h_k sont dérivables. En linéarisant le système (3.3) autour d'une suite déterministe donnée, ou bien autour de l'estimateur courant, on peut obtenir des algorithmes sous–optimaux, qui sont décrits ci–dessous.

Filtre de Kalman linéarisé

On se donne une suite (déterministe) $\{\bar{x}_k\}$ à valeurs 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 b_k et σ_k autour de $\bar{x}_{k-1}$, c'est–à–dire

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

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 non–linéaire (3.3) est alors remplacé par le système linéaire gaussien

$$X_k = F_k^{\mathrm{L}}\, X_{k-1} + f_k^{\mathrm{L}} + W_k^{\mathrm{L}} \ ,$$

$$Y_k = H_k^{\mathrm{L}}\, X_k + h_k^{\mathrm{L}} + V_k \ ,$$

avec

$$F_k^{\mathrm{L}} = b_k'(\bar{x}_{k-1}) \qquad \text{et} \qquad f_k^{\mathrm{L}} = -b_k'(\bar{x}_{k-1})\ \bar{x}_{k-1} + b_k(\bar{x}_{k-1}) \ ,$$

et avec

$$H_k^{\mathrm{L}} = h_k'(\bar{x}_k) \qquad \text{et} \qquad h_k^{\mathrm{L}} = -h_k'(\bar{x}_k)\ \bar{x}_k + h_k(\bar{x}_k) \ .$$

Ici, le vecteur aléatoire $W_k^{\mathrm{L}} = \sigma_k(\bar{x}_{k-1})\, W_k$ est gaussien, centré et de matrice de covariance $Q_k^{\mathrm{L}} = \sigma_k(\bar{x}_{k-1})\, \sigma_k^*(\bar{x}_{k-1})$. On applique alors exactement le filtre de Kalman à ce nouveau système, et on obtient l'algorithme sous–optimal suivant

$$\widehat{X}_k^- \;=\; b_k(\bar{x}_{k-1}) + b_k'(\bar{x}_{k-1})\,(\widehat{X}_{k-1} - \bar{x}_{k-1})\;,$$

$$P_k^- \;=\; b_k'(\bar{x}_{k-1})\, P_{k-1}\, (b_k'(\bar{x}_{k-1}))^* + \sigma_k(\bar{x}_{k-1})\, \sigma_k^*(\bar{x}_{k-1})\;,$$

et

$$\widehat{X}_k \;=\; \widehat{X}_k^- + K_k\,(Y_k - (h_k(\bar{x}_k) + h_k'(\bar{x}_k)\,(\widehat{X}_k^- - \bar{x}_k)))\;,$$

$$P_k \;=\; (I - K_k\, h_k'(\bar{x}_k))\; P_k^-\;,$$

avec la matrice de gain

$$K_k = P_k^-\; (h_k'(\bar{x}_k))^*\; [\, h_k'(\bar{x}_k)\, P_k^-\; (h_k'(\bar{x}_k))^* + Q_k^V\,]^{-1}\;.$$

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

$$\widehat{X}_k^- \;=\; b_k(\widehat{X}_{k-1})\;,$$

$$\widehat{X}_k \;=\; \widehat{X}_k^- + K_k\,(Y_k - h_k(\widehat{X}_k^-))\;.$$

On choisit l'initialisation $\widehat{X}_0^-$ et P_0^- de telle sorte que $\mathcal{N}(\widehat{X}_0^-, P_0^-)$ soit une bonne approximation de la distribution de probabilité du vecteur aléatoire X_0.

Filtre de Kalman étendu

Au lieu de linéariser autour d'une trajectoire nominale déterministe $\{\bar{x}_k\}$, on peut utiliser l'estimateur courant. La méthode consiste à linéariser les fonctions b_k et σ_k autour de $\widehat{X}_{k-1}$, c'est–à–dire

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

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 non–linéaire (3.3) est alors remplacé par le système *conditionnellement* linéaire gaussien

$$X_k \;=\; F_k^{\mathrm{L}}\, X_{k-1} + f_k^{\mathrm{L}} + W_k^{\mathrm{L}}\;,$$

$$Y_k \;=\; H_k^{\mathrm{L}}\, X_k + h_k^{\mathrm{L}} + V_k\;,$$

avec

$$F_k^{\mathrm{L}} = b_k'(\widehat{X}_{k-1}) \qquad \text{et} \qquad f_k^{\mathrm{L}} = -b_k'(\widehat{X}_{k-1})\; \widehat{X}_{k-1} + b_k(\widehat{X}_{k-1})\;,$$

et avec

$$H_k^{\mathrm{L}} = h_k'(\widehat{X}_k^-) \qquad \text{et} \qquad h_k^{\mathrm{L}} = -h_k'(\widehat{X}_k^-)\,\widehat{X}_k^- + h_k(\widehat{X}_k^-)\ ,$$

et on remarque que

$$F_k^{\mathrm{L}}\,\widehat{X}_{k-1} + f_k^{\mathrm{L}} = b_k(\widehat{X}_{k-1}) \qquad \text{et} \qquad H_k^{\mathrm{L}}\,\widehat{X}_k^- + h_k^{\mathrm{L}} = h_k(\widehat{X}_k^-)\ .$$

Conditionnellement à $Y_{0:k-1}$, le vecteur aléatoire $W_k^{\mathrm{L}} = \sigma_k(\widehat{X}_{k-1})\,W_k$ est gaussien, centré et de matrice de covariance conditionnelle $Q_k^{\mathrm{L}} = \sigma_k(\widehat{X}_{k-1})\,\sigma_k^*(\widehat{X}_{k-1})$. On remarque que les coefficients F_k^{L}, f_k^{L} et Q_k^{L}, et les coefficients H_k^{L} et h_k^{L} dépendent des observations passées $Y_{0:k-1}$. On applique alors exactement le filtre de Kalman à ce nouveau système, et au vu de la Remarque 2.10 on obtient l'algorithme sous–optimal suivant

$$\widehat{X}_k^- \ = \ b_k(\widehat{X}_{k-1})\ ,$$

$$P_k^- \ = \ b_k'(\widehat{X}_{k-1})\,P_{k-1}\,(b_k'(\widehat{X}_{k-1}))^* + \sigma_k(\widehat{X}_{k-1})\,\sigma_k^*(\widehat{X}_{k-1})\ ,$$

et

$$\widehat{X}_k \ = \ \widehat{X}_k^- + K_k\,(Y_k - h_k(\widehat{X}_k^-))\ ,$$

$$P_k \ = \ (I - K_k\,h_k'(\widehat{X}_k^-))\,P_k^-\ ,$$

avec la matrice de gain

$$K_k = P_k^-\,(h_k'(\widehat{X}_k^-))^*\,[h_k'(\widehat{X}_k^-)\,P_k^-\,(h_k'(\widehat{X}_k^-))^* + Q_k^V]^{-1}\ .$$

On choisit l'initialisation $\widehat{X}_0^-$ et P_0^- de telle sorte que $\mathcal{N}(\widehat{X}_0^-, P_0^-)$ soit une bonne approximation de la distribution de probabilité du vecteur aléatoire X_0.

Remarque 3.1 Dans cet algorithme, la suite $\{P_k\}$ dépend des observations, et ne peut donc pas être pré–calculée.

3.2 Filtre de Kalman *unscented*

On considère à nouveau le système non linéaire (3.3), c'est–à–dire

$$X_k \ = \ b_k(X_{k-1}) + \sigma_k(X_{k-1})\,W_k\ ,$$

$$Y_k \ = \ h_k(X_k) + V_k\ ,$$

et on ne suppose plus que les fonctions b_k et h_k sont dérivables, mais on suppose que les fonctions b_k, h_k et σ_k et certaines fonctions associées, peuvent être intégrées par rapport à certaines distributions de probabilité gaussiennes.

Au lieu de s'appuyer sur une linéarisation des fonctions autour de l'estimateur courant, on se propose ici

- de remplacer les différentes distributions de probabilité conditionnelles par des distributions de probabilité gaussiennes ayant même moyenne et même matrice de covariance,
- d'utiliser des formules de quadrature, développées initialement pour le calcul numérique d'intégrales, pour approcher ces moyennes et ces matrices de covariance conditionnelles.

Le premier point peut s'interpréter comme une projection, au sens de la distance de Kullback–Leibler, sur la famille des distributions de probabilité gaussiennes.

▶ Le calcul des deux premiers moments (moyenne et matrice de covariance) de la distribution de probabilité conditionnelle $\mu_k^-(dx) = \mathbb{P}[X_k \in dx \mid Y_{0:k-1}]$, c'est–à–dire le calcul de la moyenne conditionnelle et de la matrice de covariance conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k-1}$, est facile. Par définition

$$\begin{aligned}\widehat{X}_k^- &= \mathbb{E}[X_k \mid Y_{0:k-1}] \\ &= \mathbb{E}[b_k(X_{k-1}) \mid Y_{0:k-1}] + \mathbb{E}[\sigma_k(X_{k-1})\, W_k \mid Y_{0:k-1}] \\ &= \int_{\mathbb{R}^m} b_k(x)\, \mu_{k-1}(dx)\ ,\end{aligned}$$

compte tenu que

$$\begin{aligned}\mathbb{E}[\sigma_k(X_{k-1})\, W_k \mid Y_{0:k-1}] &= \mathbb{E}[\,\mathbb{E}[\sigma_k(X_{k-1})\, W_k \mid X_{k-1}, Y_{0:k-1}] \mid Y_{0:k-1}] \\ &= \mathbb{E}[\sigma_k(X_{k-1})\, \mathbb{E}[W_k \mid X_{k-1}, Y_{0:k-1}] \mid Y_{0:k-1}] = 0\ ,\end{aligned}$$

où on a utilisé dans la dernière égalité l'indépendance de $(Y_0, \cdots, Y_{k-1}, X_{k-1})$ et de W_k, donc $\mathbb{E}[W_k \mid X_{k-1}, Y_{0:k-1}] = 0$. Par différence

$$X_k - \widehat{X}_k^- = (b_k(X_{k-1}) - \widehat{X}_k^-) + \sigma_k(X_{k-1})\, W_k\ ,$$

de sorte que

$$\begin{aligned}P_k^- &= \mathbb{E}[\,(X_k - \widehat{X}_k^-)\,(X_k - \widehat{X}_k^-)^* \mid Y_{0:k-1}] \\ &= \mathbb{E}[\,((b_k(X_{k-1}) - \widehat{X}_k^-) + \sigma_k(X_{k-1})\, W_k)\,((b_k(X_{k-1}) - \widehat{X}_k^-) + \sigma_k(X_{k-1})\, W_k)^* \mid Y_{0:k-1}] \\ &= \mathbb{E}[\,(b_k(X_{k-1}) - \widehat{X}_k^-)\,(b_k(X_{k-1}) - \widehat{X}_k^-)^* \mid Y_{0:k-1}] \\ &\quad + \mathbb{E}[\sigma_k(X_{k-1})\, W_k\,(b_k(X_{k-1}) - \widehat{X}_k^-)^* \mid Y_{0:k-1}] \\ &\quad + \mathbb{E}[\,(b_k(X_{k-1}) - \widehat{X}_k^-)\, W_k^*\, \sigma_k^*(X_{k-1}) \mid Y_{0:k-1}] \\ &\quad + \mathbb{E}[\sigma_k(X_{k-1})\, W_k\, W_k^*\, \sigma_k^*(X_{k-1}) \mid Y_{0:k-1}] \\ &= \int_{\mathbb{R}^m} (b_k(x) - \widehat{X}_k^-)\,(b_k(x) - \widehat{X}_k^-)^*\, \mu_{k-1}(dx) + \int_{\mathbb{R}^m} \sigma_k(x)\, \sigma_k^*(x)\, \mu_{k-1}(dx)\ ,\end{aligned}$$

compte tenu que

$$\begin{aligned}
&\mathbb{E}[\sigma_k(X_{k-1})\, W_k\, W_k^*\, \sigma_k^*(X_{k-1}) \mid Y_{0:k-1}] \\
&\quad = \mathbb{E}[\,\mathbb{E}[\sigma_k(X_{k-1})\, W_k\, W_k^*\, \sigma_k^*(X_{k-1}) \mid X_{k-1}, Y_{0:k-1}] \mid Y_{0:k-1}] \\
&\quad = \mathbb{E}[\,\sigma_k(X_{k-1})\, \mathbb{E}[W_k\, W_k^* \mid X_{k-1}, Y_{0:k-1}]\, \sigma_k^*(X_{k-1}) \mid Y_{0:k-1}] \\
&\quad = \mathbb{E}[\,\sigma_k(X_{k-1})\, \sigma_k^*(X_{k-1}) \mid Y_{0:k-1}]\ ,
\end{aligned}$$

où on a utilisé dans la dernière égalité l'indépendance de $(Y_0, \cdots, Y_{k-1}, X_{k-1})$ et de W_k, donc $\mathbb{E}[W_k\, W_k^* \mid X_{k-1}, Y_{0:k-1}] = I$, et compte tenu que

$$\begin{aligned}
&\mathbb{E}[\sigma_k(X_{k-1})\, W_k\, (b_k(X_{k-1}) - \widehat{X}_k^-)^* \mid Y_{0:k-1}] \\
&\quad = \mathbb{E}[\,\mathbb{E}[\sigma_k(X_{k-1})\, W_k\, (b_k(X_{k-1}) - \widehat{X}_k^-)^* \mid X_{k-1}, Y_{0:k-1}] \mid Y_{0:k-1}] \\
&\quad = \mathbb{E}[\sigma_k(X_{k-1})\, \mathbb{E}[W_k \mid X_{k-1}, Y_{0:k-1}]\, (b_k(X_{k-1}) - \widehat{X}_k^-)^* \mid Y_{0:k-1}] = 0\ ,
\end{aligned}$$

où on a encore utilisé dans la dernière égalité l'indépendance de $(Y_0, \cdots, Y_{k-1}, X_{k-1})$ et de W_k, donc $\mathbb{E}[W_k \mid X_{k-1}, Y_{0:k-1}] = 0$.

▶ En revanche, le calcul des deux premiers moments (moyenne et matrice de covariance) de la distribution de probabilité conditionnelle $\mu_k(dx) = \mathbb{P}[X_k \in dx \mid Y_{0:k}]$, c'est–à–dire le calcul de la moyenne conditionnelle et de la matrice de covariance conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$, n'est pas immédiat, et on commence par le calcul des deux premiers moments (moyenne et matrice de covariance) de la distribution de probabilité conditionnelle jointe du vecteur aléatoire (X_k, Y_k) sachant $Y_{0:k-1}$, qui est plus facile. On rappelle que

$$\widehat{X}_k^- = \int_{\mathbb{R}^m} b_k(x)\, \mu_{k-1}(dx)\ ,$$

a déjà été obtenu plus haut, et par définition

$$\begin{aligned}
\widehat{Y}_k^- &= \mathbb{E}[Y_k \mid Y_{0:k-1}] \\
&= \mathbb{E}[h_k(X_k) \mid Y_{0:k-1}] + \mathbb{E}[V_k \mid Y_{0:k-1}] \\
&= \int_{\mathbb{R}^m} h_k(x)\, \mu_k^-(dx)\ .
\end{aligned}$$

On rappelle que

$$P_k^- = \int (b_k(x) - \widehat{X}_k^-)\, (b_k(x) - \widehat{X}_k^-)^*\, \mu_{k-1}(dx) + \int_{\mathbb{R}^m} \sigma_k(x)\, \sigma_k^*(x)\, \mu_{k-1}(dx)\ ,$$

a déjà été obtenu plus haut, et par différence

$$Y_k - \widehat{Y}_k^- = (h_k(X_k) - \widehat{Y}_k^-) + V_k\ ,$$

de sorte que

$$\begin{aligned}
\Xi_k &= \mathbb{E}[\,(Y_k - \widehat{Y}_k^-)\,(Y_k - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] \\
&= \mathbb{E}[\,((h_k(X_k) - \widehat{Y}_k^-) + V_k)\,((h_k(X_k) - \widehat{Y}_k^-) + V_k)^* \mid Y_{0:k-1}] \\
&= \mathbb{E}[\,(h_k(X_k) - \widehat{Y}_k^-)\,(h_k(X_k) - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] + \mathbb{E}[V_k\,V_k^* \mid Y_{0:k-1}] \\
&\qquad + \mathbb{E}[\,(h_k(X_k) - \widehat{Y}_k^-)\,V_k^* \mid Y_{0:k-1}] \\
&\qquad + \mathbb{E}[V_k\,(h_k(X_k) - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] \\
&= \int_{\mathbb{R}^m} (h_k(x) - \widehat{Y}_k^-)\,(h_k(x) - \widehat{Y}_k^-)^*\,\mu_k^-(dx) + Q_k^V\ ,
\end{aligned}$$

compte tenu que

$$\begin{aligned}
&\mathbb{E}[V_k\,(h_k(X_k) - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] \\
&\qquad = \mathbb{E}[\,\mathbb{E}[V_k\,(h_k(X_k) - \widehat{Y}_k^-)^* \mid X_k, Y_{0:k-1}] \mid Y_{0:k-1}] \\
&\qquad = \mathbb{E}[\,\mathbb{E}[V_k \mid X_k, Y_{0:k-1}]\,(h_k(X_k) - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] = 0\ .
\end{aligned}$$

où on a utilisé dans la dernière égalité l'indépendance de $(Y_0, \cdots, Y_{k-1}, X_k)$ et de V_k, donc $\mathbb{E}[V_k \mid X_k, Y_{0:k-1}] = 0$, et

$$\begin{aligned}
C_k &= \mathbb{E}[\,(X_k - \widehat{X}_k^-)\,(Y_k - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] \\
&= \mathbb{E}[\,(X_k - \widehat{X}_k^-)\,(h_k(X_k) - \widehat{Y}_k^-)^* \mid Y_{0:k-1}] + \mathbb{E}[\,(X_k - \widehat{X}_k^-)\,V_k^* \mid Y_{0:k-1}] \\
&= \int_{\mathbb{R}^m} (x - \widehat{X}_k^-)\,(h_k(x) - \widehat{Y}_k^-)^*\,\mu_k^-(dx)\ .
\end{aligned}$$

On remplace la distribution de probabilité conditionnelle jointe du vecteur aléatoire (X_k, Y_k) sachant $Y_{0:k-1}$ par la distribution de probabilité gaussienne de moyenne et de matrice de covariance

$$\begin{pmatrix} \widehat{X}_k^- \\ \\ \widehat{Y}_k^- \end{pmatrix} \qquad \text{et} \qquad \begin{pmatrix} P_k^- & C_k \\ \\ C_k^* & \Xi_k \end{pmatrix}\ ,$$

respectivement. Si la matrice Q_k^V est inversible, alors a fortiori la matrice Ξ_k est inversible, et d'après la Proposition 1.5 on obtient immédiatement les approximations suivantes

$$\widehat{X}_k = \widehat{X}_k^- + C_k\,\Xi_k^{-1}\,(Y_k - \widehat{Y}_k^-) \qquad \text{et} \qquad P_k = P_k^- - C_k\,\Xi_k^{-1}\,C_k^*\ ,$$

pour les deux premiers moments de la distribution de probabilité conditionnelle μ_k, c'est–à–dire pour la moyenne conditionnelle et de la matrice de covariance conditionnelle du vecteur aléatoire X_k sachant $Y_{0:k}$.

Ces équations ne sont pas fermées, c'est–à–dire que les moments $\widehat{X}_k^-$ et P_k^- ne s'expriment pas en fonction des moments $\widehat{X}_{k-1}$ et P_{k-1} seulement, mais en fonction de toute la distribution de probabilité conditionnelle μ_{k-1}, et de même, les moments $\widehat{X}_k$ et P_k ne s'expriment pas en fonction des moments $\widehat{X}_k^-$ et P_k^- seulement, mais en fonction de toute la distribution de probabilité conditionnelle μ_k^-. Pour fermer ces équations, on adopte le principe de projection énoncé plus haut.

▶ On remplace la distribution de probabilité conditionnelle μ_{k-1} par la distribution de probabilité gaussienne de moyenne $\widehat{X}_{k-1}$ et de matrice de covariance $P_{k-1} = S_{k-1}\, S_{k-1}^*$, et en effectuant le changement de variable $x = \widehat{X}_{k-1} + S_{k-1}\, u$, on obtient les approximations

$$\widehat{X}_k^- \approx \int \widehat{b}_k(u)\ \exp\{-\tfrac{1}{2}\,|u|^2\}\ \frac{du}{(2\pi)^{m/2}}\ ,$$

et

$$P_k^- \ \approx\ \int (\widehat{b}_k(u) - \widehat{X}_k^-)\,(\widehat{b}_k(u) - \widehat{X}_k^-)^*\ \exp\{-\tfrac{1}{2}\,|u|^2\}\ \frac{du}{(2\pi)^{m/2}}$$
$$+ \int \widehat{\sigma}_k(u)\,\widehat{\sigma}_k^*(u)\ \exp\{-\tfrac{1}{2}\,|u|^2\}\ \frac{du}{(2\pi)^{m/2}}$$

où par définition

$$\widehat{b}_k(u) = b_k(\widehat{X}_{k-1} + S_{k-1}\, u) \qquad \text{et} \qquad \widehat{\sigma}_k(u) = \sigma_k(\widehat{X}_{k-1} + S_{k-1}\, u)\ .$$

▶ De même, on remplace la distribution de probabilité conditionnelle μ_k^- par la distribution de probabilité gaussienne de moyenne $\widehat{X}_k^-$ et de matrice de covariance $P_k^- = S_k^-\ (S_k^-)^*$, et en effectuant le changement de variable $x = \widehat{X}_k^- + S_k^-\, u$, on obtient les approximations

$$\widehat{Y}_k^- \approx \int_{\mathbb{R}^m} \widehat{h}_k(u)\ \exp\{-\tfrac{1}{2}\,|u|^2\}\ \frac{du}{(2\pi)^{m/2}}\ ,$$

et

$$\Xi_k \approx \int_{\mathbb{R}^m} (\widehat{h}_k(u) - \widehat{Y}_k^-)\,(\widehat{h}_k(u) - \widehat{Y}_k^-)^*\ \exp\{-\tfrac{1}{2}\,|u|^2\}\ \frac{du}{(2\pi)^{m/2}} + Q_k^V\ ,$$

et

$$C_k \approx S_k^-\ \int_{\mathbb{R}^m} u\,(\widehat{h}_k(u) - \widehat{Y}_k^-)^*\ \exp\{-\tfrac{1}{2}\,|u|^2\}\ \frac{du}{(2\pi)^{m/2}}\ ,$$

où par définition

$$\widehat{h}_k(u) = h_k(\widehat{X}_k^- + S_k^-\, u)\ .$$

Il reste donc à calculer les intégrales des fonctions non–linéaires

$$\widehat{b}_k(u),\ \widehat{b}_k(u)\,\widehat{b}_k^*(u),\ \widehat{\sigma}_k(u)\,\widehat{\sigma}_k^*(u),\ \widehat{h}_k(u),\ u\,\widehat{h}_k^*(u) \text{ et } \widehat{h}_k(u)\,\widehat{h}_k^*(u)\ ,$$

par rapport à la densité gaussienne réduite centrée.

Remarque 3.2 Si on suppose que les fonctions b_k et h_k sont dérivables, et qu'on utilise un développement limité au premier ordre au voisinage de $u = 0$ dans les intégrales ci–dessus, on retrouve les équations du filtre de Kalman étendu. L'idée ici est de *ne pas linéariser*, et de calculer les intégrales en utilisant des formules de quadrature numérique.

On introduit les formules de quadrature suivantes, reposant sur la notion de σ–points. En dimension m, la densité de probabilité gaussienne centrée réduite (de matrice de covariance identité) est représentée par $2m+1$ points de quadrature $(u_{-m}, \cdots, u_m)$ appelés σ–points, et définis par

$$u_0 = 0 \ , \qquad u_i = e_i \ \sqrt{m+\kappa} \qquad \text{et} \qquad u_{-i} = -u_i \ ,$$

où e_i désigne le i–ème vecteur de base, affectés des poids

$$w_0 = \frac{\kappa}{m+\kappa} \qquad \text{et} \qquad w_{-i} = w_i = \frac{1}{2\,(m+\kappa)} \ , \tag{3.4}$$

pour tout $i = 1, \cdots, m$ (d'autres choix de σ–points sont possibles). On vérifie que

$$\sum_{i=-m}^{+m} w_i = 1 \ , \qquad \sum_{i=-m}^{+m} w_i \ u_i = 0 \qquad \text{et} \qquad \sum_{i=-m}^{+m} w_i \ u_i \ u_i^* = \sum_{i=1}^{m} e_i \ e_i^* = I \ ,$$

c'est–à–dire que les deux premiers moments sont pris en compte exactement. Plus généralement

$$\int_{\mathbb{R}^m} \phi(u) \ \exp\{-\tfrac{1}{2}\,|u|^2\} \ \frac{du}{(2\pi)^{m/2}} \approx \sum_{i=-m}^{+m} w_i \ \phi(u_i) \ ,$$

et un changement de variable évident donne aussitôt

$$\int_{\mathbb{R}^m} \phi(\mu + \Sigma^{1/2}\,u) \ \exp\{-\tfrac{1}{2}\,|u|^2\} \ \frac{du}{(2\pi)^{m/2}} \approx \sum_{i=-m}^{+m} w_i \ \phi(\mu + \Sigma^{1/2}\,u_i) \ ,$$

pour toute fonction ϕ définie sur $\mathbb{R}^m$, c'est–à–dire que les σ–points $(x_{-m}, \cdots, x_m)$ associés à la distribution de probabilité gaussienne de vecteur moyenne μ et de matrice de covariance Σ, sont définis par la relation $x_i = \mu + \Sigma^{1/2}\,u_i$, soit

$$x_0 = \mu \ , \qquad x_i = \mu + \Sigma^{1/2}\,e_i \ \sqrt{m+\kappa} \qquad \text{et} \qquad x_{-i} = \mu - \Sigma^{1/2}\,e_i \ \sqrt{m+\kappa} \ ,$$

pour tout $i = 1, \cdots, m$. On vérifie que

$$\sum_{i=-m}^{+m} w_i \ x_i = \mu \qquad \text{et} \qquad \sum_{i=-m}^{+m} w_i \ (x_i - \mu) \ (x_i - \mu)^* = \sum_{i=1}^{m} \Sigma^{1/2}\,e_i \ (\Sigma^{1/2}\,e_i)^* = \Sigma \ ,$$

c'est–à–dire que les deux premiers moments sont pris en compte exactement. Plus généralement encore, soit X un vecteur aléatoire gaussien de vecteur moyenne μ et de matrice de covariance Σ, et soit T une transformation non–linéaire définie sur $\mathbb{R}^m$. Clairement

$$\int \phi(T(\mu + \Sigma^{1/2}\,u)) \ \exp\{-\tfrac{1}{2}\,|u|^2\} \ \frac{du}{(2\pi)^{m/2}} \approx \sum_{i=-m}^{+m} w_i \ \phi(T(x_i)) \ ,$$

pour toute fonction ϕ définie sur $\mathbb{R}^m$, c'est–à–dire que les σ–points $(x'_{-m}, \cdots, x'_m)$ associés au vecteur aléatoire transformé $X' = T(X)$, sont simplement obtenus par la relation $x'_i = T(x_i)$ à partir des σ–points $(x_{-m}, \cdots, x_m)$ associés au vecteur aléatoire X, soit

$$x'_0 = T(\mu) \ , \qquad x'_i = T(\mu + \Sigma^{1/2}\, e_i\, \sqrt{m+\kappa}) \qquad \text{et} \qquad x'_{-i} = T(\mu - \Sigma^{1/2}\, e_i\, \sqrt{m+\kappa}) \ ,$$

pour tout $i = 1, \cdots, m$.

Avec ces formules de quadrature, on obtient l'algorithme de filtrage sous–optimal suivant.

Expression de $\widehat{X}_k^-$ et P_k^- en fonction de $\widehat{X}_{k-1}$ et $P_{k-1} = S_{k-1}\, S_{k-1}^*$:

On introduit les σ–points

$$x_0 = \widehat{X}_{k-1} \ , \qquad x_i = \widehat{X}_{k-1} + S_{k-1}\, e_i\, \sqrt{m+\kappa} \qquad \text{et} \qquad x_{-i} = \widehat{X}_{k-1} - S_{k-1}\, e_i\, \sqrt{m+\kappa} \ ,$$

affectés des poids (3.4) pour tout $i = 1, \cdots, m$, et on définit le vecteur moyenne

$$\widehat{X}_k^- = \sum_{i=-m}^{+m} w_i\, b_k(x_i) \ ,$$

et la matrice de covariance

$$P_k^- = \sum_{i=-m}^{+m} w_i\, (b_k(x_i) - \widehat{X}_k^-)\, (b_k(x_i) - \widehat{X}_k^-)^* + \sum_{i=-m}^{+m} w_i\, \sigma_k(x_i)\, \sigma_k^*(x_i) = S_k^-\, (S_k^-)^* \ .$$

Expression de $\widehat{X}_k$ et P_k en fonction de $\widehat{X}_k^-$ et $P_k^- = S_k^-\, (S_k^-)^*$:

On introduit les σ–points

$$x_0 = \widehat{X}_k^- \ , \qquad x_i = \widehat{X}_k^- + S_k^-\, e_i\, \sqrt{m+\kappa} \qquad \text{et} \qquad x_{-i} = \widehat{X}_k^- - S_k^-\, e_i\, \sqrt{m+\kappa} \ ,$$

affectés des poids (3.4) pour tout $i = 1, \cdots, m$, on définit le vecteur moyenne

$$\widehat{Y}_k^- = \sum_{i=-m}^{+m} w_i\, h_k(x_i) \ ,$$

la matrice de covariance

$$\Xi_k = \sum_{i=-m}^{+m} w_i\, (h_k(x_i) - \widehat{Y}_k^-)\, (h_k(x_i) - \widehat{Y}_k^-)^* + Q_k^V \ ,$$

et la matrice de corrélation

$$C_k = \sum_{i=-m}^{+m} w_i\, (x_i - \widehat{X}_k^-)\, (h_k(x_i) - \widehat{Y}_k^-)^* \ ,$$

et on pose

$$\widehat{X}_k = \widehat{X}_k^- + C_k\, \Xi_k^{-1}\, (Y_k - \widehat{Y}_k^-) \qquad \text{et} \qquad P_k = P_k^- - C_k\, \Xi_k^{-1}\, C_k^* = S_k\, S_k^* \ .$$

Chapitre 4

Systèmes non–linéaires non–gaussiens, et extensions

On considère une suite d'états cachés $\{X_k\}$ à valeurs dans $\mathbb{R}^m$, vérifiant

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

où $\{X_k\}$ et $\{W_k\}$ prennent respectivement leurs valeurs dans $\mathbb{R}^m$ et $\mathbb{R}^p$, et une suite d'observations $\{Y_k\}$ à valeurs dans $\mathbb{R}^d$, vérifiant

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

et on suppose que

- la condition initiale X_0 n'est pas nécessairement gaussienne,
- la suite $\{W_k\}$ est un bruit blanc, pas nécessairement gaussien,
- la suite $\{V_k\}$ est un bruit blanc, pas nécessairement gaussien,
- les suites $\{W_k\}$ et $\{V_k\}$ et la condition initiale X_0 sont mutuellement indépendants.

On ne suppose pas que les fonctions f_k et h_k sont dérivables. On suppose en revanche que

- il est facile de *simuler* un vecteur aléatoire selon la loi $\eta_0(dx)$ de X_0,
- 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 Équation d'état (modèle a priori)

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

$$\begin{aligned}\mathbb{E}[\,\phi(X_k) \mid X_0, \cdots, X_{k-1}] \ &= \ \mathbb{E}[\,\phi(f_k(X_{k-1}, W_k)) \mid X_0, \cdots, X_{k-1}] \\ &= \int_{\mathbb{R}^p} \phi(f_k(X_{k-1}, w))\; p_k^W(dw) \ ,\end{aligned}$$

pour toute fonction test 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 Équation d'observation (modèle de capteur)

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 la probabilité d'*émission

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

définie par

$$g_k(x, y) = q_k^V(y - h_k(x)) \ ,$$

et on définit la fonction de vraisemblance

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

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

En d'autres termes

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

Preuve. Pour toute famille $\phi_0, \cdots, \phi_n$ 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

$$\begin{aligned}
&\mathbb{E}[\phi_0(Y_0) \cdots \phi_n(Y_n) \mid X_0, \cdots, X_n] \\
&\quad = \mathbb{E}[\phi_0(h_0(X_0) + V_0) \cdots \phi_n(h_n(X_n) + V_n) \mid X_0, \cdots, X_n] \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \phi_0(h_0(X_0) + v_0) \cdots \phi_n(h_n(X_n) + v_n)\ \mathbb{P}[V_0 \in dv_0, \cdots, V_n \in dv_n] \\
&\quad = \prod_{k=0}^{n} \int_{\mathbb{R}^d} \phi_k(h_k(X_k) + v)\ \mathbb{P}[V_k \in dv] \ ,
\end{aligned}$$

et de même

$$\begin{aligned}
\mathbb{E}[\phi_k(Y_k) \mid X_k] \ &= \ \mathbb{E}[\phi_k(h_k(X_k) + V_k) \mid X_k] \\
&= \int_{\mathbb{R}^d} \phi_k(h_k(X_k) + v)\ \mathbb{P}[V_k \in dv] \\
&= \int_{\mathbb{R}^d} \phi_k(h_k(X_k) + v)\ q_k^V(v)\, dv = \int_{\mathbb{R}^d} \phi_k(y)\ q_k^V(y - h_k(X_k))\, dy \ ,
\end{aligned}$$

de sorte que

$$\mathbb{E}[\phi_0(Y_0)\cdots\phi_n(Y_n) \mid X_0,\cdots,X_n] = \prod_{k=0}^{n} \mathbb{E}[\phi_k(Y_k) \mid X_k] \ ,$$

et

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

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*

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

  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 $g_k(x)$.

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 Représentation probabiliste

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

$$\begin{aligned}
&\mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n, Y_0 \in dy_0, \cdots, Y_n \in dy_n] \\
&\quad = \mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n \mid X_0 = x_0, \cdots, X_n = x_n] \, \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \\
&\quad = \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \, \prod_{k=0}^{n} g_k(x_k, y_k) \, dy_0 \cdots dy_n \ .
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}[Y_0 \in dy_0, \cdots, Y_n \in dy_n] \\
&\quad = \int_E \cdots \int_E \prod_{k=0}^{n} g_k(x_k, y_k) \, \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \, dy_0 \cdots dy_n \\
&\quad = \mathbb{E}[\prod_{k=0}^{n} g_k(X_k, y_k) \,] \, dy_0 \cdots dy_n \ .
\end{aligned}$$

D'après la formule de Bayes, il vient

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

et on obtient

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

pour toute suite $(y_0, \cdots, y_n)$ d'observations. Pour toute fonction test ϕ définie sur E

$$\begin{aligned}
&\mathbb{E}[\, \phi(X_n) \prod_{k=0}^{n} g_k(X_k, y_k)\,] \\
&\quad = \int_E \cdots \int_E \phi(x_n) \prod_{k=0}^{n} g_k(x_k, y_k)\, \mathbb{P}[X_0 \in dx_0, \cdots, X_n \in dx_n] \\
&\quad = \int_E \phi(x_n)\, \mathbb{P}[X_n \in dx_n \mid Y_0 = y_0, \cdots, Y_n = y_n]\, \mathbb{E}[\, \prod_{k=0}^{n} g_k(X_k, y_k)\,] \\
&\quad = \mathbb{E}[\, \phi(X_n) \mid Y_0 = y_0, \cdots, Y_n = y_n]\, \mathbb{E}[\, \prod_{k=0}^{n} g_k(X_k, y_k)\,] \ ,
\end{aligned}$$

et on en déduit que

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

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

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

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

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

et où l'espérance porte seulement sur les états cachés successifs $(X_0, \cdots, X_n)$: les fonctions de vraisemblance $g_0(x), \cdots, g_n(x)$ dépendent implicitement des observations $(Y_0, \cdots, Y_n)$, mais celles–ci sont considérées comme fixées dans l'expression ci–dessus. De la même manière

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

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

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

5.2 Équation récurrente

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*

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

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

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

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

Remarque 5.2 Une autre manière de caractériser les distributions de probabilité $\mu_{k-1} Q_k$ et $g_k \cdot \mu_k^-$ est de décrire leur action sur des fonctions test : pour toute fonction test ϕ mesurable

bornée définie sur $\mathbb{R}^m$

$$\begin{aligned}\langle \mu_{k-1}\, Q_k, \phi\rangle &= \int_E \mu_{k-1}\, Q_k(dx')\ \phi(x') \\ &= \int_E [\ \int_E \mu_{k-1}(dx)\, Q_k(x,dx')\ ]\ \phi(x') = \int_E \mu_{k-1}(dx)\ [\ \int_E Q_k(x,dx')\, \phi(x')\ ] \\ &= \int_E \mu_{k-1}(dx)\ Q_k\, \phi(x) = \langle \mu_{k-1}, Q_k\, \phi\rangle\ ,\end{aligned}$$

et

$$\langle g_k \cdot \mu_k^-, \phi\rangle = \int_E g_k \cdot \mu_k^-(dx')\, \phi(x') = \frac{\displaystyle\int_E \mu_k^-(dx')\, g_k(x')\, \phi(x')}{\displaystyle\int_E \mu_k^-(dx')\, g_k(x')} = \frac{\langle \mu_k^-, g_k\, \phi\rangle}{\langle \mu_k^-, g_k\rangle}\ .$$

Expression de μ_n^- en fonction de μ_{n-1} :

On remarque immédiatement que

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

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

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

pour toute fonction test mesurable bornée ϕ définie sur E. En normalisant, on obtient

$$\langle \mu_n^-, \phi\rangle = \frac{\langle \gamma_n^-, \phi\rangle}{\langle \gamma_n^-, 1\rangle} = \frac{\langle \gamma_{n-1}, Q_n\, \phi\rangle}{\langle \gamma_{n-1}, 1\rangle} = \langle \mu_{n-1}, Q_n\, \phi\rangle = \langle \mu_{n-1}\, Q_n, \phi\rangle\ ,$$

et comme la fonction test ϕ est quelconque, on en déduit que

$$\mu_n^- = \mu_{n-1}\, Q_n\ .$$

Expression de μ_n en fonction de μ_n^- :

On a

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

pour toute fonction test mesurable bornée ϕ définie sur E. En normalisant, on obtient

$$\langle \mu_n, \phi \rangle = \frac{\langle \gamma_n, \phi \rangle}{\langle \gamma_n, 1 \rangle} = \frac{\langle \gamma_n^-, g_n\, \phi \rangle}{\langle \gamma_n^-, g_n \rangle} = \frac{\langle \mu_n^-, g_n\, \phi \rangle}{\langle \mu_n^-, g_n \rangle} = \langle g_n \cdot \mu_n^-, \phi \rangle \ ,$$

où l'avant–dernière égalité est obtenue en divisant numérateur et dénominateur par la constante de normalisation $\langle \gamma_n^-, 1 \rangle$, et comme la fonction test ϕ est quelconque, on en déduit que

$$\mu_n = g_n \cdot \mu_n^- \ . \qquad \square$$

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, présentées au Chapitre 2. 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 répandu, 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

On rappelle que la suite $\{\mu_k\}$ vérifie l'équation récurrente

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

d'après le Théorème 5.1. L'idée du filtrage particulaire consiste à chercher une approximation des distributions de probabilité conditionnelle $\eta_k(dx)$ et $\mu_k(dx)$ sous la forme de combinaisons linéaires (éventuellement pondérées) de masses de Dirac, appelées *particules*, de la forme

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

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}. 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 fini* de distributions de probabilité, peu pratique à manipuler, et qu'on décide de remplacer par la distribution de probabilité empirique

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

associée à un N–échantillon $\{\xi_k^i , i = 1, \cdots , N\}$ de variables aléatoires. On peut par exemple décider de générer un échantillon ayant précisément pour loi commune $\mu_{k-1}^N Q_k(dx')$, compte tenu que 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 $\{1, \cdots , N\}$ et selon la loi discrète définie par les poids $(w_{k-1}^1, \cdots , w_{k-1}^N)$, 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.

Plus généralement, on peut remplacer l'étape (i) par

(i') on sélectionne un individu $\widehat{\xi}_{k-1}^i$ au sein de la population Σ_{k-1} et en fonction des poids,

ce qui peut se réaliser de nombreuses manières différentes, qui ne sont pas précisées ici. On applique ensuite la formule de Bayes à l'approximation $\eta_k^N(dx')$, et on obtient exactement

$$g_k \cdot \eta_k^N = \sum_{i=1}^N \frac{g_k(\xi_k^i)}{\sum_{j=1}^N g_k(\xi_k^j)} \, \delta_{\xi_k^i} = \sum_{i=1}^N w_k^i \, \delta_{\xi_k^i} ,$$

avec

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

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 $i = 1, \cdots , N$

- on sélectionne un individu $\widehat{\xi}_{k-1}^i$ au sein de la population Σ_{k-1} et en fonction des poids,

- on génère la nouvelle position $\xi_k^i \sim Q_k(\widehat{\xi}_{k-1}^i, dx')$,
- on calcule le nouveau poids $w_k^i \propto g_k(\xi_k^i)$.

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

Introduction (classification)

La classification consiste à décider parmi un nombre *fini* d'hypothèses, décrites par un ensemble *fini* E, au vu d'observations généralement bruitées, dont la distribution dépend de l'hypothèse vraie. Ainsi, si l'hypothèse $X = i$ est vraie pour $i \in E$, alors l'observation Y recueillie a pour distribution

$$\mathbb{P}[Y \in dy \mid X = i] = g_i(y)\, dy \qquad \text{ou bien} \qquad \mathbb{P}[Y = \ell \mid X = i] = b_i^\ell \ ,$$

selon qu'il s'agit d'une observation *numérique* à valeurs dans $\mathbb{R}^d$, ou bien d'une observation *symbolique* à valeurs dans un autre ensemble *fini* O.

On considère ici le problème où l'hypothèse vraie varie au cours du temps, et on souhaite décider, de manière récursive si possible, parmi un nombre *fini* d'hypothèses, au vu d'une suite d'observations généralement bruitées.

Typiquement, on dispose d'une suite $(Y_0, Y_1, \cdots, Y_n)$ d'observations, où chaque observation Y_k est reliée à l'hypothèse X_k par une relation probabiliste (supposée indépendante de l'instant considéré) de la forme

$$\mathbb{P}[Y_k \in dy \mid X_k = i] = g_i(y)\, dy \qquad \text{ou bien} \qquad \mathbb{P}[Y_k = \ell \mid X_k = i] = b_i^\ell \ ,$$

par exemple

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

avec un bruit additif V_k indépendant de X_k.

Tel qu'il est formulé, le problème de décision, vu aussi comme le problème d'estimation de l'hypothèse X_n, à partir des observations $(Y_0, Y_1, \cdots, Y_n)$ est en général mal–posé, et il est utile d'introduire un modèle *a priori* qui donne une description probabiliste de la suite $(X_0, X_1, \cdots, X_n)$. On considérera en particulier le cas où la suite des hypothèses et des observations forme un modèle de Markov caché.

Dans de nombreux cas, la prise en compte de l'information a priori peut se ramener au problème statique suivant : étant donnés deux variables aléatoires X et Y, qu'apporte le fait d'observer la réalisation $Y = y$ sur la connaissance que l'on a de X ?

On suppose que la variable cachée X prend ses valeurs dans un ensemble *fini* E et que la variable observée Y prend ses valeurs dans un ensemble quelconque F. Par définition, un *estimateur* de X à partir de l'observation de Y est un élément aléatoire $I(Y)$ dans E, où I est une application mesurable définie sur F à valeurs dans E, c'est–à–dire une règle de décision, ou un *classifieur*, qui fait pour toute observation le choix d'un élément de E. Naturellement $I(Y)$ n'est pas égal à X : une mesure de l'écart entre l'estimateur et la vraie valeur est fournie par la probabilité d'erreur

$$\mathbb{P}[I(Y) \neq X] \ . \tag{6.1}$$

L'estimateur du minimum de la probabilité d'erreur (MPE, pour *minimum probability of error*) de X sachant Y est un estimateur $X_*(Y)$ tel que

$$\mathbb{P}[X_*(Y) \neq X] \leq \mathbb{P}[I(Y) \neq X] \ ,$$

pour tout autre estimateur $I(Y)$. La Proposition 6.1 ci–dessous montre que cet estimateur est obtenu à l'aide de la distribution de probabilité conditionnelle de X sachant $Y = y$, définie à partir de la distribution de probabilité jointe de (X, Y) par la décomposition

$$\mathbb{P}[X = i, Y \in dy] = \mathbb{P}[X = i \mid Y = y] \, \mathbb{P}[Y \in dy] \ . \tag{6.2}$$

Proposition 6.1 *Soit X et Y deux variables aléatoires à valeurs dans l'ensemble fini E et dans F respectivement. L'estimateur MPE de X sachant Y est le maximum a posteriori, i.e.*

$$X_*(y) = \operatorname*{argmax}_{i \in E} \mathbb{P}[X = i \mid Y = y] \ .$$

Preuve. Pour tout classifieur I, on a

$$\begin{aligned}
\mathbb{P}[I(Y) = X \mid Y = y] &= \sum_{i \in E} \mathbb{P}[I(Y) = X, X = i \mid Y = y] \\
&= \sum_{i \in E} \mathbb{P}[I(Y) = X \mid X = i, Y = y] \, \mathbb{P}[X = i \mid Y = y] \\
&= \sum_{i \in E} 1_{(I(y) = i)} \, \mathbb{P}[X = i \mid Y = y] \ ,
\end{aligned}$$

pour tout $y \in F$, de sorte que

$$\begin{aligned}
&\mathbb{P}[X_*(Y) \neq X \mid Y = y] - \mathbb{P}[I(Y) \neq X \mid Y = y] \\
&\quad = \sum_{i \in E} [\, 1_{(I(y) = i)} - 1_{(X_*(y) = i)} \,] \, \mathbb{P}[X = i \mid Y = y] \\
&\quad = \sum_{i \in E} [\, 1_{(I(y) = i)} - 1_{(X_*(y) = i)} \,] \, [\, \mathbb{P}[X = i \mid Y = y] - p_*(y) \,] \ ,
\end{aligned}$$

où

$$p_*(y) = \max_{i \in E} \mathbb{P}[X = i \mid Y = y] \ .$$

Par définition

$$\mathbb{P}[X = i \mid Y = y] - p_*(y) \leq 0 \ ,$$

pour tout $i \in E$, avec égalité pour $i = X_*(y)$, tandis que

$$1_{(I(y)=i)} - 1_{(X_*(y)=i)} = 1_{(I(y)=i)} \geq 0 \ ,$$

pour tout $i \neq X_*(y)$. On en déduit que

$$\mathbb{P}[X_*(Y) \neq X \mid Y = y] - \mathbb{P}[I(Y) \neq X \mid Y = y] \leq 0 \ ,$$

pour tout $y \in F$, de sorte que

$$\begin{aligned} &\mathbb{P}[X_*(Y) \neq X] - \mathbb{P}[I(Y) \neq X] \\ &\quad = \int_F [\, \mathbb{P}[X_*(Y) \neq X \mid Y = y] - \mathbb{P}[I(Y) \neq X \mid Y = y] \,] \, \mathbb{P}[Y \in dy] \leq 0 \ , \end{aligned}$$

avec égalité pour $I = X_*$. □

L'objectif de cette deuxième partie du cours est de fournir des algorithmes efficaces de calcul des probabilités conditionnelles

$$\mathbb{P}[X_n = i \mid Y_0, Y_1, \cdots, Y_n] \ ,$$

dans le cas particulier où la suite des hypothèses et des observations forme un modèle de Markov caché.

Chapitre 7

Modèles de Markov cachés

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

7.1 Chaînes de Markov à état fini

On considère un espace d'état *fini* E. 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}.

Il résulte de la Proposition 7.1 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 7.1 *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. □

7.2 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, 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 *symbolique*

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

  pour tout $i_0, \cdots, i_n \in E$, et tout $\ell_0, \cdots, \ell_n \in O$,

- et dans le cas *numérique*

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

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

Exemple 7.2 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 inversible, 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 *finie* $h = (h_i)$ de vecteurs de $\mathbb{R}^d$, et on a

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

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.

Il résulte de la Proposition 7.3 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* $\pi = (\pi_{i,j})$

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

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

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

- ou dans le cas *numérique*, des *densités d'émission* $g = (g_i)$

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

Les probabilités / densités d'émissions sont rangées dans les matrices diagonales

$$B^\ell = \mathrm{diag}(b_i^\ell) \quad \text{pour tout } \ell \in O \qquad \text{et} \qquad G(y) = \mathrm{diag}(g_i(y)) \quad \text{pour tout } y \in \mathbb{R}^d.$$

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

Proposition 7.3 *Dans le cas* symbolique, *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'émission* 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 numérique*, 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'émission g, est donnée par*

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

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

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 *symbolique*. 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 7.1.

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

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

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

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

et on conclut de la même manière, en utilisant la Proposition 7.1. □

On désigne par $\boldsymbol{M} = (\nu, \pi, b)$ dans le cas *symbolique*, et par $\boldsymbol{M} = (\nu, \pi, g)$ dans le cas *numérique*, les paramètres caractéristiques du modèle, et on s'intéresse aux trois 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 $(X_0, X_1, \cdots, X_n)$.

- **Identifier** le modèle $\boldsymbol{M}$: Etant donnée une suite d'observations $(Y_0, \cdots, Y_n)$, il s'agit de calculer l'estimateur du *maximum de vraisemblance* pour les paramètres inconnus du modèle. La réponse à ce problème est fournie par les *formules de re–estimation* de Baum–Welch, qui définissent un algorithme itératif pour maximiser la fonction de vraisemblance.

Chapitre 8

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

- *dans le cas* symbolique *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* numérique *par*

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

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

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

Preuve. On considère d'abord le cas *symbolique*. On utilise la Proposition 7.3 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 *numérique*, on procède de la même manière

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

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

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

Remarque 8.2 Cette méthode fournit une première expression pour la distribution de probabilité conditionnelle de la suite des états $(X_0, \cdots, X_n)$ sachant les observations $(Y_0, \cdots, Y_n)$:

- dans le cas *symbolique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] = \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\displaystyle\sum_{j_0,\cdots,j_n \in E} \nu_{j_0}\ \pi_{j_0,j_1} \cdots \pi_{j_{n-1},j_n}\ b_{j_0}^{Y_0} \cdots b_{j_n}^{Y_n}}\ ,$$

- et dans le cas *numérique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] =$$

$$= \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)}{\displaystyle\sum_{j_0,\cdots,j_n \in E} \nu_{j_0}\ \pi_{j_0,j_1} \cdots \pi_{j_{n-1},j_n}\ g_{j_0}(Y_0) \cdots g_{j_n}(Y_n)}\ ,$$

et pour la vraisemblance du modèle (obtenue en utilisant la suite des observations $(Y_0, \cdots, Y_n)$ à la place des variables muettes) :

- dans le cas *symbolique*

$$L_n = \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}\ ,$$

- et dans le cas *numérique*

$$L_n = \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)\ .$$

On en déduit les expressions suivantes pour les distributions non–normalisées :

- dans le cas *symbolique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\ L_n = \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}\ ,$$

- et dans le cas *numérique*

$$\mathbb{P}[X_0 = i_0, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\ L_n = \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)\ .$$

Remarque 8.3 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 8.1 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 $|E|^{n+1}$ trajectoires possibles différentes. Le nombre total d'opérations élémentaires (additions et multiplications) à effectuer est donc de l'ordre de : $2(n+1)\ |E|^{n+1}$. Ce nombre croît *exponentiellement* avec le nombre n d'observations.

8.1 Equation forward

On introduit la variable *forward* $p_k = (p_k^i)$ vue comme un vecteur–ligne, et définie par

$$p_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_k]\ L_k \qquad \text{pour tout } i \in E.$$

Remarque 8.4 La variable forward permet de calculer la distribution de probabilité conditionnelle de l'état présent X_n sachant les observations $(Y_0, \cdots, Y_n)$:

$$\mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\ p_n^i \qquad \text{pour tout } i \in E,$$

(en ce sens, p_n est une distribution de probabilité non–normalisée), et la constante de normalisation

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

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

Théorème 8.5 *La suite* $\{p_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* symbolique

$$p_k^j = \Big[\sum_{i \in E} p_{k-1}^i\ \pi_{i,j}\ \Big]\ b_j^{Y_k} \qquad \textit{pour tout } j \in E, \tag{8.1}$$

  avec la condition initiale : $p_0^i = \nu_i\ b_i^{Y_0}$ *pour tout* $i \in E$,

- *et dans le cas* numérique

$$p_k^j = \Big[\sum_{i \in E} p_{k-1}^i\ \pi_{i,j}\ \Big]\ g_j(Y_k) \qquad \textit{pour tout } j \in E, \tag{8.2}$$

  avec la condition initiale : $p_0^i = \nu_i\ g_i(Y_0)$ *pour tout* $i \in E$.

Remarque 8.6 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable forward vue comme un vecteur–ligne, ce qui donne

- dans le cas *symbolique*

$$p_k = p_{k-1}\ \pi\ B^{Y_k} \qquad \text{et} \qquad p_0 = \nu\ B^{Y_0}\ ,$$

- et dans le cas *numérique*

$$p_k = p_{k-1}\ \pi\ G(Y_k) \qquad \text{et} \qquad p_0 = \nu\ G(Y_0)\ ,$$

avec les matrices *diagonales* définies par

$$B^\ell = \mathrm{diag}(b_i^\ell) \qquad \text{pour tout } \ell \in O,$$

et

$$G(y) = \mathrm{diag}(g_i(y)) \qquad \text{pour tout } y \in \mathbb{R}^d,$$

respectivement.

Preuve. On considère uniquement le cas *symbolique*. Il résulte de la Remarque 8.2 que

$$\begin{aligned}
&\mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k = \\
&\quad = \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i}\ \pi_{i,j}\ b_{i_0}^{Y_0} \cdots b_i^{Y_{k-1}}\ b_j^{Y_k} \\
&\quad = \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k}\ ,
\end{aligned}$$

pour tout $i, j \in E$ et tout $i_0, \cdots, i_{k-2} \in E$. En sommant par rapport à $i_0, \cdots, i_{k-2} \in E$, on obtient

$$\begin{aligned}
\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k \ &= \ \mathbb{P}[X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k} \\
&= \ p_{k-1}^i\ \pi_{i,j}\ b_j^{Y_k}\ ,
\end{aligned}$$

pour tout $i, j \in E$. En sommant ensuite par rapport à $i \in E$, on obtient

$$p_k^j = \sum_{i \in E} [p_{k-1}^i\ \pi_{i,j}]\ b_j^{Y_k}\ ,$$

pour tout $j \in E$, d'où le résultat. □

Remarque 8.7 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 $|E|\,(2|E| + 1)$ 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 : $n\ |E|\,(2|E| + 1) + (|E| - 1)$. Ce nombre croît seulement *linéairement* avec le nombre n d'observations.

Au lieu de résoudre d'abord l'équation forward pour la version non–normalisée de la distribution conditionnelle, définie par

$$p_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_k]\ L_k \qquad \text{pour tout } i \in E,$$

et d'en déduire ensuite la constante de normalisation (vraisemblance) et la version normalisée de la distribution conditionnelle (filtre), définies par

$$L_k = \sum_{i \in E} p_k^i \qquad \text{et} \qquad \bar{p}_k^i = \frac{p_k^i}{\sum_{j \in E} p_k^j} = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_k] \qquad \text{pour tout } i \in E,$$

respectivement, il est plus efficace, d'un point de vue *numérique*, de propager directement le filtre.

Proposition 8.8 *La suite* $\{\bar{p}_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* symbolique

$$\bar{p}_k^j = \frac{1}{c_k}\ \Big[\ \sum_{i \in E} \bar{p}_{k-1}^i\ \pi_{i,j}\ \Big]\ b_j^{Y_k} \qquad \textit{pour tout } j \in E,$$

  avec la condition initiale : $\bar{p}_0^i = \dfrac{1}{c_0}\ \nu_i\ b_i^{Y_0}$ *pour tout* $i \in E$*, où les constantes de normalisation sont définies par*

$$c_k = \sum_{i,j \in E} \bar{p}_{k-1}^i\ \pi_{i,j}\ b_j^{Y_k} \qquad \textit{et} \qquad c_0 = \sum_{i \in E} \nu_i\ b_i^{Y_0}\ ,$$

- *et dans le cas* numérique

$$\bar{p}_k^j = \frac{1}{c_k}\ \Big[\ \sum_{i \in E} \bar{p}_{k-1}^i\ \pi_{i,j}\ \Big]\ g_j(Y_k) \qquad \textit{pour tout } j \in E,$$

  avec la condition initiale : $\bar{p}_0^i = \dfrac{1}{c_0}\ \nu_i\ g_i(Y_0)$ *pour tout* $i \in E$*, où les constantes de normalisation sont définies par*

$$c_k = \sum_{i,j \in E} \bar{p}_{k-1}^i\ \pi_{i,j}\ g_j(Y_k) \qquad \textit{et} \qquad c_0 = \sum_{i \in E} \nu_i\ g_i(Y_0)\ .$$

Remarque 8.9 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable forward normalisée vue comme un vecteur–ligne, ce qui donne

- dans le cas *symbolique*

$$\bar{p}_k = \frac{1}{c_k}\ \bar{p}_{k-1}\ \pi\ B^{Y_k} \qquad \text{et} \qquad \bar{p}_0 = \frac{1}{c_0}\ \nu\ B^{Y_0}\ ,$$

  où les constantes de normalisation sont définies par

$$c_k = \bar{p}_{k-1}\ \pi\ b^{Y_k} \qquad \text{et} \qquad c_0 = \nu\ b^{Y_0}\ ,$$

- et dans le cas *numérique*

$$\bar{p}_k = \frac{1}{c_k}\, \bar{p}_{k-1}\, \pi\, G(Y_k) \qquad \text{et} \qquad \bar{p}_0 = \frac{1}{c_0}\, \nu\, G(Y_0)\ ,$$

où les constantes de normalisation sont définies par

$$c_k = \bar{p}_{k-1}\, \pi\, g(Y_k) \qquad \text{et} \qquad c_0 = \nu\, g(Y_0)\ .$$

PREUVE. On considère uniquement le cas *symbolique* : en utilisant l'équation forward (8.1), on obtient

$$\bar{p}_k^j = \frac{1}{L_k}\, p_k^j = \frac{1}{L_k}\, \Big[\sum_{i\in E} p_{k-1}^i\, \pi_{i,j} \Big]\, b_j^{Y_k} = \frac{L_{k-1}}{L_k}\, \Big[\sum_{i\in E} \bar{p}_{k-1}^i\, \pi_{i,j} \Big]\, b_j^{Y_k}\ ,$$

pour tout $j \in E$, et nécessairement

$$\frac{L_k}{L_{k-1}} = \sum_{i,j\in E} \bar{p}_{k-1}^i\, \pi_{i,j}\, b_j^{Y_k} = c_k\ ,$$

et en utilisant la condition initiale de l'équation forward (8.1), on obtient

$$\bar{p}_0^i = \frac{1}{L_0}\, p_0^i = \frac{1}{L_0}\, \nu_i\, b_i^{Y_0}\ ,$$

pour tout $i \in E$, et nécessairement

$$L_0 = \sum_{i\in E} \nu_i\, b_i^{Y_0} = c_0\ . \qquad \square$$

Remarque 8.10 La suite $\{\log L_k\}$ vérifie l'équation récurrente suivante, valide dans le cas *symbolique* et dans le cas *numérique*

$$\log L_k = \log L_{k-1} + \log c_k\ ,$$

avec la condition initiale

$$\log L_0 = \log c_0\ ,$$

et en itérant cette relation, on obtient

$$\log L_n = \sum_{k=0}^{n} \log c_k\ .$$

8.2 Equation backward

Pour tout instant intermédiaire k, antérieur à l'instant final n, on définit

$$q_k^i = \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n]\ L_n \qquad \text{pour tout } i \in E.$$

Remarque 8.11 Cette variable permet de calculer la distribution de probabilité conditionnelle de l'état présent X_k sachant toutes les observations $(Y_0, \cdots, Y_n)$:

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\ q_k^i \qquad \text{pour tout } i \in E,$$

et la constante de normalisation

$$L_n = \sum_{i \in E} q_k^i\ ,$$

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

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 $(k+1)$: dans le cas *symbolique*, il résulte en effet de la Remarque 8.2 que

$$\begin{aligned}
q_k^i \ &= \ \mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n]\ L_n \\
&= \sum_{\substack{i_0, \cdots, i_{k-1} \in E \\ i_{k+1}, \cdots, i_n \in E}} \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i, \\
&\qquad\qquad X_{k+1} = i_{k+1}, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n]\ L_n \\
&= \sum_{\substack{i_0, \cdots, i_{k-1} \in E \\ i_{k+1}, \cdots, i_n \in E}} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i}\ \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\ b_i^{Y_k}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \\
&= \sum_{i_{k+1}, \cdots, i_n \in E} \Big[\sum_{i_0, \cdots, i_{k-1} \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-1},i}\ b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\ b_i^{Y_k}\ \Big] \\
&\qquad\qquad \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \\
&= \sum_{i_{k+1}, \cdots, i_n \in E} p_k^i\ \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \\
&= p_k^i\ \Big[\sum_{i_{k+1}, \cdots, i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n}\ \Big]\ ,
\end{aligned}$$

et une expression similaire peut être obtenue dans le cas *numérique*, ce qui justifie d'introduire la variable *backward* $v_k = (v_k^i)$ vue comme un vecteur–colonne, et définie :

- dans le cas *symbolique* par

$$v_k^i = \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}} \cdots b_{i_n}^{Y_n} \qquad \text{pour tout } i \in E,$$

  et en particulier : $v_{n-1}^i = \sum_{j \in E} \pi_{i,j}\ b_j^{Y_n}$ pour tout $i \in E$,

- et dans le cas *numérique* par

$$v_k^i = \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ g_{i_{k+1}}(Y_{k+1}) \cdots g_{i_n}(Y_n) \qquad \text{pour tout } i \in E,$$

  et en particulier : $v_{n-1}^i = \sum_{j \in E} \pi_{i,j}\ g_j(Y_n)$ pour tout $i \in E$.

Remarque 8.12 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. On déduit alors de la Proposition 8.1 que la distribution de probabilité des observations $(Y_{k+1}, \cdots, Y_n)$ sachant $(X_k = i)$ est donnée :

- dans le cas *symbolique* par

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

  pour tout $\ell_{k+1}, \cdots, \ell_n \in O$,

- et dans le cas *numérique* par

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

$$= \sum_{i_{k+1},\cdots,i_n \in E} \pi_{i,i_{k+1}} \cdots \pi_{i_{n-1},i_n}\ g_{i_{k+1}}(y_{k+1}) \cdots g_{i_n}(y_n)\ dy_{k+1} \cdots dy_n\ ,$$

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

ce qui permet d'interpréter la variable *backward* comme la vraisemblance du modèle issu de l'état $X_k = i$ à l'instant k (obtenue en utilisant la suite des observations $(Y_{k+1}, \cdots, Y_n)$ à la place des variables muettes).

Théorème 8.13 *La suite* $\{v_k\}$ *vérifie l'équation récurrente rétrograde suivante :*

- *dans le cas* symbolique

$$v_{k-1}^i = \sum_{j\in E} \pi_{i,j}\ b_j^{Y_k}\ v_k^j \qquad \textit{pour tout } i \in E, \tag{8.3}$$

  avec la condition initiale : $v_n^i = 1$ *pour tout* $i \in E$,

- *et dans le cas* numérique

$$v_{k-1}^i = \sum_{j\in E} \pi_{i,j}\ g_j(Y_k)\ v_k^j \qquad \textit{pour tout } i \in E, \tag{8.4}$$

  avec la condition initiale : $v_n^i = 1$ *pour tout* $i \in E$.

Remarque 8.14 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable backward vue comme un vecteur–colonne, ce qui donne

- dans le cas *symbolique*

$$v_{k-1} = \pi\ B^{Y_k}\ v_k \qquad \text{et} \qquad v_n \equiv 1\ ,$$

- et dans le cas *numérique*

$$v_{k-1} = \pi\ G(Y_k)\ v_k \qquad \text{et} \qquad v_n \equiv 1\ .$$

Preuve. On considère uniquement le cas *symbolique*. Avec l'initialisation proposée à l'instant n, l'équation (8.3) permet de retrouver à l'instant $(n-1)$

$$v_{n-1}^i = \sum_{j\in E} \pi_{i,j}\ b_j^{Y_n} \qquad \text{pour tout } i \in E.$$

Par définition

$$\begin{aligned}
v_{k-1}^i &= \sum_{i_k,\cdots,i_n\in E} \pi_{i,i_k}\cdots\pi_{i_{n-1},i_n}\ b_{i_k}^{Y_k}\cdots b_{i_n}^{Y_n} \\
&= \sum_{j\in E}\ \sum_{i_{k+1},\cdots,i_n\in E} \pi_{i,j}\ \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_j^{Y_k}\ b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n} \\
&= \sum_{j\in E} \pi_{i,j}\ b_j^{Y_k}\ \Big[\ \sum_{i_{k+1},\cdots,i_n\in E} \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\ b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n}\ \Big] \\
&= \sum_{j\in E} \pi_{i,j}\ b_j^{Y_k}\ v_k^j\ ,
\end{aligned}$$

pour tout $i \in E$, d'où le résultat. □

Proposition 8.15 *Les équations forward et backward sont* duales *l'une de l'autre :*

$$\sum_{i\in E} p_0^i \; v_0^i = \sum_{i\in E} p_k^i \; v_k^i = \sum_{i\in E} p_n^i = L_n \;, \tag{8.5}$$

pour tout instant k.

Preuve. On considère uniquement le cas *symbolique*. En utilisant successivement l'équation forward (8.1) et l'équation backward (8.3), on obtient

$$\begin{aligned}\sum_{j\in E} p_k^j \; v_k^j \; &= \; \sum_{j\in E} \Big[\sum_{i\in E} p_{k-1}^i \; \pi_{i,j} \Big] \; b_j^{Y_k} \; v_k^j \\ &= \sum_{i\in E} p_{k-1}^i \; \Big[\sum_{j\in E} \pi_{i,j} \; b_j^{Y_k} \; v_k^j \Big] = \sum_{i\in E} p_{k-1}^i \; v_{k-1}^i \;,\end{aligned}$$

d'où le résultat. □

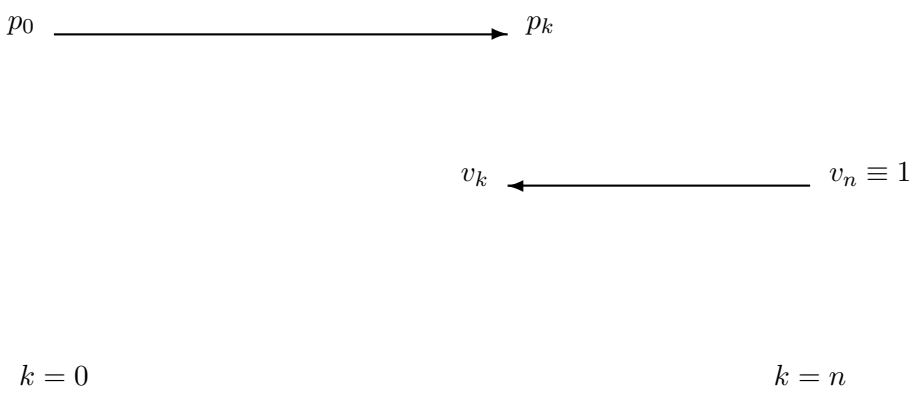

Figure 8.1 : Equations forward / backward

Proposition 8.16 *La distribution de probabilité de la* transition (X_{k-1}, X_k) *à un instant intermédiaire sachant les observations* $(Y_0, \cdots, Y_n)$ *jusqu'à l'instant final est donnée :*

- *dans le cas* symbolique *par*

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{L_n} \; p_{k-1}^i \; \pi_{i,j} \; b_j^{Y_k} \; v_k^j \qquad \textit{pour tout } i, j \in E,$$

- *et dans le cas* numérique *par*

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{L_n} \; p_{k-1}^i \; \pi_{i,j} \; g_j(Y_k) \; v_k^j \qquad \textit{pour tout } i, j \in E.$$

En sommant pour tout $j \in E$ et en utilisant l'équation backward, ou bien en sommant pour tout $i \in E$ et en utilisant l'équation forward, on retrouve le résultat suivant, en terme du produit composante–par–composante des variables forward et backward.

Corollaire 8.17 *La distribution de probabilité de l'état X_k à un instant intermédiaire sachant les observations $(Y_0, \cdots, Y_n)$ jusqu'à l'instant final est donnée par :*

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\, q_k^i \qquad \textit{pour tout } i \in E,$$

avec la définition

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

Remarque 8.18 On vérifie que les constantes de normalisation

$$\sum_{i,j \in E} p_{k-1}^i\, \pi_{i,j}\, b_j^{Y_k}\, v_k^j = \sum_{j \in E} \Big[\sum_{i \in E} p_{k-1}^i\, \pi_{i,j} \Big]\; b_j^{Y_k}\, v_k^j = \sum_{j \in E} p_k^j\, v_k^j = L_n \ ,$$

et

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

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

Preuve de la Proposition 8.16. On considère uniquement le cas *symbolique*. Fixer la transition entre les instants $(k-1)$ et k permet d'effectuer une coupure entre le passé jusqu'à l'instant $(k-2)$ et le futur à partir de l'instant $(k+1)$: il résulte en effet de la Remarque 8.2 que

$$\begin{aligned}
&\mathbb{P}[X_{k-1}=i, X_k=j \mid Y_0,\cdots,Y_n]\; L_n = \\
&= \sum_{\substack{i_0,\cdots,i_{k-2}\in E\\ i_{k+1},\cdots,i_n\in E}} \mathbb{P}[X_0=i_0,\cdots,X_{k-2}=i_{k-2},X_{k-1}=i, \\
&\qquad\qquad X_k=j,X_{k+1}=i_{k+1},\cdots,X_n=i_n \mid Y_0,\cdots,Y_n]\; L_n \\
&= \sum_{\substack{i_0,\cdots,i_{k-2}\in E\\ i_{k+1},\cdots,i_n\in E}} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; \pi_{i,j}\; \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; b_j^{Y_k}\; b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n} \\
&= \sum_{i_0,\cdots,i_{k-2}\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; \pi_{i,j}\; b_j^{Y_k} \\
&\qquad\qquad \Big[\sum_{i_{k+1},\cdots,i_n\in E} \pi_{j,i_{k+1}}\cdots\pi_{i_{n-1},i_n}\; b_{i_{k+1}}^{Y_{k+1}}\cdots b_{i_n}^{Y_n}\Big] \\
&= \sum_{i_0,\cdots,i_{k-2}\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j \\
&= \Big[\sum_{i_0,\cdots,i_{k-2}\in E} \nu_{i_0}\; \pi_{i_0,i_1}\cdots\pi_{i_{k-2},i}\; b_{i_0}^{Y_0}\cdots b_{i_{k-2}}^{Y_{k-2}}\; b_i^{Y_{k-1}}\Big]\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j \\
&= p_{k-1}^i\; \pi_{i,j}\; b_j^{Y_k}\; v_k^j\;,
\end{aligned}$$

d'où le résultat. □

Au lieu de résoudre d'abord l'équation forward et l'équation backward séparément, et d'en déduire successivement la version non–normalisée de la distribution conditionnelle, définie par

$$q_k^i = p_k^i\; v_k^i = \mathbb{P}[X_k=i \mid Y_0,\cdots,Y_n]\; L_n \qquad \text{pour tout } i\in E,$$

puis la version normalisée de la distribution conditionnelle (lisseur), définie par

$$\bar{q}_k^i = \frac{q_k^i}{\sum_{j\in E} q_k^j} = \frac{p_k^i\; v_k^i}{\sum_{j\in E} p_k^j\; v_k^j} = \frac{\bar{p}_k^i\; v_k^i}{\sum_{j\in E} \bar{p}_k^j\; v_k^j} = \mathbb{P}[X_k=i \mid Y_0,\cdots,Y_n] \qquad \text{pour tout } i\in E,$$

il est plus efficace, d'un point de vue *numérique*, de propager directement le filtre, comme dans la Proposition 8.8, puis de propager la variable $\bar{v}_k = (\bar{v}_k^i)$ définie par

$$\bar{v}_k^i = \frac{v_k^i}{\sum_{j\in E} \bar{p}_k^j\; v_k^j} \qquad \text{pour tout } i\in E.$$

Remarque 8.19 Avec cette normalisation de la variable backward, la distribution de probabilité conditionnelle de l'état X_k sachant les observations $(Y_0, \cdots, Y_n)$ s'exprime comme

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \bar{p}_k^i \; \bar{v}_k^i = \bar{q}_k^i \qquad \text{pour tout } i \in E.$$

Proposition 8.20 *La suite $\{\bar{v}_k\}$ vérifie l'équation récurrente rétrograde suivante :*

- *dans le cas* symbolique

$$\bar{v}_{k-1}^i = \frac{1}{c_k} \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; \bar{v}_k^j \qquad \textit{pour tout } i \in E,$$

  avec la condition initiale : $\bar{v}_n^i = 1$ pour tout $i \in E$,

- *et dans le cas* numérique

$$\bar{v}_{k-1}^i = \frac{1}{c_k} \sum_{j \in E} \pi_{i,j} \; g_j(Y_k) \; \bar{v}_k^j \qquad \textit{pour tout } i \in E,$$

  avec la condition initiale : $\bar{v}_n^i = 1$ pour tout $i \in E$,

où les constantes de normalisation sont celles introduites dans l'énoncé de la Proposition 8.8.

Remarque 8.21 Ce résultat énoncé composante–par–composante peut être aussi formulé pour la variable backward normalisée vue comme un vecteur–colonne, ce qui donne

- dans le cas *symbolique*

$$\bar{v}_{k-1} = \frac{1}{c_k} \; \pi \; B^{Y_k} \; \bar{v}_k \qquad \text{et} \qquad \bar{v}_n \equiv 1 \; ,$$

- et dans le cas *numérique*

$$\bar{v}_{k-1} = \frac{1}{c_k} \; \pi \; G(Y_k) \; \bar{v}_k \qquad \text{et} \qquad \bar{v}_n \equiv 1 \; ,$$

où les constantes de normalisation sont celles introduites à la Remarque 8.9.

Preuve. On considère uniquement le cas *symbolique* : en utilisant la relation (8.5), on remarque que

$$\bar{v}_k^i = \frac{v_k^i}{\sum_{j \in E} \bar{p}_k^j \; v_k^j} = \sum_{j \in E} p_k^j \; \frac{v_k^i}{\sum_{j \in E} p_k^j \; v_k^j} = \frac{L_k}{L_n} \; v_k^i \; ,$$

et en utilisant l'équation backward (8.13), on obtient

$$\bar{v}_{k-1}^i = \frac{L_{k-1}}{L_n} \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; v_k^j = \frac{L_{k-1}}{L_n} \; \frac{L_n}{L_k} \sum_{j \in E} \pi_{i,j} \; b_j^{Y_k} \; \bar{v}_k^j \; ,$$

pour tout $i \in E$, d'où le résultat compte tenu que $\dfrac{L_k}{L_{k-1}} = c_k$. □

Remarque 8.22 On remarque que

$$\frac{1}{L_n}\, p_{k-1}^i\, v_k^j = \frac{L_{k-1}}{L_n}\, \bar{p}_{k-1}^i\, \frac{L_n}{L_k}\, \bar{v}_k^j = \frac{1}{c_k}\, \bar{p}_{k-1}^i\, \bar{v}_k^j \qquad \text{pour tout } i,j \in E,$$

et en reportant cette identité dans les expressions obtenues à la Proposition 8.16, on vérifie que la distribution de probabilité conditionnelle de la transition (X_{k-1}, X_k) sachant les observations $(Y_0, \cdots, Y_n)$ s'exprime

- dans le cas *symbolique*, comme

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{c_k}\, \bar{p}_{k-1}^i\, \pi_{i,j}\, b_j^{Y_k}\, \bar{v}_k^j\ ,$$

  pour tout $i, j \in E$,

- et dans le cas *numérique*, comme

$$\mathbb{P}[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] = \frac{1}{c_k}\, \bar{p}_{k-1}^i\, \pi_{i,j}\, g_j(Y_k)\, \bar{v}_k^j\ ,$$

  pour tout $i, j \in E$,

et les constantes de normalisation sont celles introduites dans l'énoncé de la Proposition 8.8.

Chapitre 9

Algorithme de Viterbi

Il résulte de la Remarque 8.4 et du Corollaire 8.17 que les variables forward et backward étudiées au Chapitre 8 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 $(Y_0, \cdots, Y_n)$, définies par

$$\mathbb{P}[X_n = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\, p_n^i \qquad \text{pour tout } i \in E,$$

et

$$\mathbb{P}[X_k = i \mid Y_0, \cdots, Y_n] = \frac{1}{L_n}\, q_k^i \qquad \text{pour tout } i \in E,$$

respectivement, où la constante de normalisation

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

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

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. D'après la Proposition 6.1, on peut proposer en revanche l'estimateur du *maximum a posteriori*, qui minimise la probabilité de l'erreur d'estimation sachant les observations $(Y_0, \cdots, Y_n)$, défini pour l'état présent par

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

et pour l'état à un instant intermédiaire par

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

(en supposant que dans chacun des cas le maximum est atteint en un point unique).

Cependant, il peut arriver que la suite $(X_0^{\mathrm{LMAP}}, \cdots, X_n^{\mathrm{LMAP}})$ ainsi générée soit incohérente avec le modèle, dans le sens suivant : il peut arriver que l'on obtienne $X_{k-1}^{\mathrm{LMAP}} = i$ et $X_k^{\mathrm{LMAP}} = j$ 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 *trajectoriel* du *maximum a posteriori*, défini par

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

et qui minimise la probabilité de l'erreur d'estimation de la suite des états cachés sachant les observations $(Y_0, \cdots, Y_n)$. Il n'est pas possible de calculer cette probabilité pour chacune des $|E|^{n+1}$ suites possibles, ni d'effectuer la maximisation de manière exhaustive. Le calcul efficace de cet estimateur est fourni par un algorithme de programmation dynamique, appelé *algorithme de Viterbi*, qui exploite la remarque suivante.

Remarque 9.1 Si (x_1^*, x_2^*) atteint le maximum de la fonction $f(x_1, x_2)$ définie sur l'ensemble produit $E_1 \times E_2$, alors nécessairement x_1^* et x_2^* atteignent respectivement le maximum des fonctions

$$h(x_1) = f(x_1, x_2^*) \qquad \text{et} \qquad g(x_2) = \max_{x_1 \in E_1} f(x_1, x_2) \ ,$$

définies sur les ensembles E_1 et E_2. Clairement

$$h(x_1^*) = f(x_1^*, x_2^*) \geq f(x_1, x_2^*) = h(x_1) \ ,$$

pour tout $x_1 \in E_1$, et d'autre part

$$g(x_2^*) = \max_{x_1 \in E_1} f(x_1, x_2^*) \geq f(x_1^*, x_2^*) \geq f(x_1, x_2) \ ,$$

pour tout $(x_1, x_2) \in E_1 \times E_2$, et comme la majoration est valide pour tout $x_1 \in E_1$, alors elle reste valide pour le maximum, c'est–à–dire que

$$g(x_2^*) \geq \max_{x_1 \in E_1} f(x_1, x_2) = g(x_2) \ ,$$

pour tout $x_2 \in E_2$.

Si la suite $(i_0^*, \cdots, i_k^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_k) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_k = i_k \mid Y_0, \cdots, Y_k] \ ,$$

alors nécessairement, d'après la Remarque 9.1, i_k^* atteint le maximum de la fonction

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

ce qui justifie d'introduire la fonction *valeur* $V_k = (V_k^i)$ définie par

$$V_k^i = \max_{i_0,\cdots,i_{k-1} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i \mid Y_0, \cdots, Y_k] \, L_k \ ,$$

pour tout $i \in E$.

Théorème 9.2 *La suite* $\{V_k\}$ *vérifie l'équation récurrente suivante :*

- *dans le cas* symbolique

$$V_k^j = \max_{i \in E} \left[V_{k-1}^i \, \pi_{i,j} \right] b_j^{Y_k} \qquad \textit{pour tout } j \in E, \tag{9.1}$$

  avec la condition initiale : $V_0^i = \nu_i \, b_i^{Y_0}$ *pour tout* $i \in E$,

- *et dans le cas* numérique

$$V_k^j = \max_{i \in E} \left[V_{k-1}^i \, \pi_{i,j} \right] g_j(Y_k) \qquad \textit{pour tout } j \in E, \tag{9.2}$$

  avec la condition initiale : $V_0^i = \nu_i \, g_i(Y_0)$ *pour tout* $i \in E$.

La suite $\{V_k\}$ *est instrumentale, et permet de définir à chaque instant* k *l'indice*

$$I_{k-1}(j) = \operatorname*{argmax}_{i \in E} \left[V_{k-1}^i \, \pi_{i,j} \right] \qquad \textit{pour tout } j \in E,$$

qui peut s'interpréter comme un pointeur vers un état à l'instant précédent $(k-1)$ *(en supposant que le maximum est atteint en un point unique).*

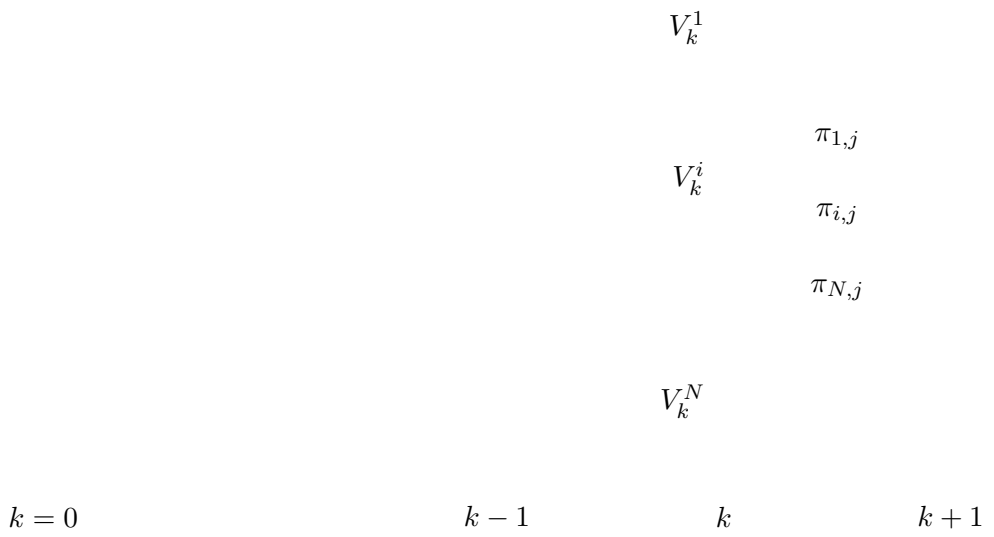

FIGURE 9.1 : Algorithme de Viterbi (programmation dynamique)

PREUVE. On considère uniquement le cas *symbolique*. Il résulte de la Remarque 8.2 que

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i}\ \pi_{i,j}\ b_{i_0}^{Y_0} \cdots b_i^{Y_{k-1}}\ b_j^{Y_k}$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k}\ ,$$

pour tout $i, j \in E$ et tout $i_0, \cdots, i_{k-2} \in E$. En maximisant par rapport à $i_0, \cdots, i_{k-2} \in E$, on obtient

$$\max_{i_0,\cdots,i_{k-2} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ L_k =$$

$$= \max_{i_0,\cdots,i_{k-2} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}\ \pi_{i,j}\ b_j^{Y_k}$$

$$= V_{k-1}^i\ \pi_{i,j}\ b_j^{Y_k}\ ,$$

pour tout $i, j \in E$. En maximisant ensuite par rapport à $i \in E$, on obtient

$$V_k^j\ =\ \max_{i \in E} [V_{k-1}^i\ \pi_{i,j}]\ b_j^{Y_k}\ ,$$

pour tout $j \in E$, d'où le résultat. □

Remarque 9.3 Soit $j \in E$ fixé. Si la suite $(i_0^*, \cdots, i_{k-1}^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = j \mid Y_0, \cdots, Y_k]\ ,$$

alors nécessairement, d'après la Remarque 9.1, i_{k-1}^* atteint le maximum de la fonction

$$i \longmapsto \max_{i_0,\cdots,i_{k-2} \in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_k]\ ,$$

c'est–à–dire que i_{k-1}^* atteint le maximum de la fonction

$$i \longmapsto V_{k-1}^i\ \pi_{i,j}\ .$$

En d'autres termes, parmi toutes les suites qui aboutissent dans l'état $j \in E$ à l'instant k, la suite de plus grande probabilité, conditionnellement aux observations $(Y_0, \cdots, Y_k)$, est nécessairement passée dans l'état

$$I_{k-1}(j) = \operatorname*{argmax}_{i \in E} [V_{k-1}^i\ \pi_{i,j}]\ ,$$

à l'instant précédent $(k-1)$ (en supposant que le maximum est atteint en un point unique). Ce calcul permet de pré–positionner à tout instant k et pour tout état $j \in E$ un pointeur vers un état $I_{k-1}(j)$ à l'instant précédent $(k-1)$. En outre, on a nécessairement

$$\pi_{I_{k-1}(j),j} > 0\ ,$$

ce qui garantit que la transition de l'état $I_{k-1}(j)$ vers l'état j est possible pour le modèle.

On obtient alors un algorithme efficace pour le calcul de la suite optimale, c'est–à–dire de l'estimateur *trajectoriel* du *maximum a posteriori*.

Théorème 9.4 *La suite* $\{X_k^{\mathrm{MAP}}\}$ *vérifie l'équation récurrente rétrograde suivante :*

$$X_{k-1}^{\mathrm{MAP}} = I_{k-1}(X_k^{\mathrm{MAP}}) \ ,$$

avec la condition initiale

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

Preuve. Si la suite $(i_0^*, \cdots, i_n^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_n) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{n-1} = i_{n-1}, X_n = i_n \mid Y_0, \cdots, Y_n] \ ,$$

alors nécessairement, d'après la Remarque 9.1, i_n^* atteint le maximum de la fonction

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

c'est–à–dire que

$$i_n^* = \underset{i\in E}{\mathrm{argmax}}\, V_n^i \ ,$$

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

Si la suite $(i_0^*, \cdots, i_{k-1}^*, i_k^*, \cdots, i_n^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}, i_k, \cdots, i_n)$$

$$\longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k, \cdots, X_n = i_n \mid Y_0, \cdots, Y_n] \ ,$$

alors nécessairement, d'après la Remarque 9.1, la suite $(i_0^*, \cdots, i_{k-1}^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k^*, \cdots, X_n = i_n^* \mid Y_0, \cdots, Y_n] \ .$$

Dans le cas *symbolique*, il résulte de la Remarque 8.2 que

$$\mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1}, X_k = i_k^*, \cdots, X_n = i_n^* \mid Y_0, \cdots, Y_n]\ L_n =$$

$$= \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{k-2},i_{k-1}}\ \pi_{i_{k-1},i_k^*} \cdots \pi_{i_{n-1}^*,i_n^*}\ b_{i_0}^{Y_0} \cdots b_{i_{k-1}}^{Y_{k-1}}\ b_{i_k^*}^{Y_k} \cdots b_{i_n^*}^{Y_n}$$

$$= \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i_{k-1} \mid Y_0, \cdots, Y_{k-1}]\ L_{k-1}$$

$$\pi_{i_{k-1},i_k^*} \cdots \pi_{i_{n-1}^*,i_n^*}\ b_{i_k^*}^{Y_k} \cdots b_{i_n^*}^{Y_n} \ ,$$

c'est–à–dire que la suite $(i_0^*, \cdots, i_{k-1}^*)$ atteint le maximum de la fonction

$$(i_0, \cdots, i_{k-1}) \longmapsto \mathbb{P}[X_0 = i_0, \cdots, X_{k-1} = i_{k-1} \mid Y_0, \cdots, Y_{k-1}]\ \pi_{i_{k-1},i_k^*} \ ,$$

et nécessairement, d'après la Remarque 9.1, i^*_{k-1} atteint le maximum de la fonction

$$i \longmapsto \max_{i_0,\cdots,i_{k-2}\in E} \mathbb{P}[X_0 = i_0, \cdots, X_{k-2} = i_{k-2}, X_{k-1} = i \mid Y_0, \cdots, Y_{k-1}]\ \pi_{i,i^*_k}\ ,$$

c'est–à–dire que

$$i^*_{k-1} = \operatorname*{argmax}_{i\in E} [V^i_{k-1}\ \pi_{i,i^*_k}] = I_{k-1}(i^*_k)\ ,$$

en supposant que le maximum est atteint en un point unique. □

Chapitre 10

Formules de re–estimation de Baum–Welch

Dans les chapitres précédent, l'accent a porté sur l'estimation d'un état caché ou de la suite des états cachés successifs, à partir d'une suite d'observations et pour un modèle connu a priori et 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 *symbolique*, des *probabilités d'émission* $b = (b_i^\ell)$

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

- ou dans le cas *numérique*, des *densités d'émission* $g = (g_i)$

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

  par exemple des densités gaussiennes caractérisées par la donnée d'une famille *finie* $h = (h_i)$ de vecteurs de $\mathbb{R}^d$, et d'une famille *finie* $R = (R_i)$ de matrices de covariance inversibles, c'est–à–dire

$$g_i(y) = g(h_i, R_i, y) = \frac{1}{\sqrt{\det(2\pi\, R_i)}}\ \exp\{-\tfrac{1}{2}\,(y - h_i)^*\, R_i^{-1}\,(y - h_i)\}\ ,$$

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

L'objectif ici est d'estimer les paramètres caractéristiques du modèle, à partir d'une suite d'observations, et le point de vue adopté est celui de l'estimation par maximum de vraisemblance.

Dans le cas *symbolique*, il résulte de la Remarque 8.2 que la fonction de vraisemblance du modèle $\boldsymbol{M} = (\nu, \pi, b)$ admet l'expression suivante

$$L_n = \sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}\ ,$$

et on se propose d'étudier un algorithme itératif pour maximiser la fonction de vraisemblance L_n par rapport aux paramètres (ν, π, b) du modèle. Soit $\boldsymbol{M'} = (\nu', \pi', b')$ un autre modèle, pour lequel on a déja évalué la fonction de vraisemblance

$$L'_n = \sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}\ ,$$

par exemple en terme des solutions $\{p'_k\}$ et $\{v'_k\}$ des équations forward / backward de Baum pour le modèle $\boldsymbol{M'}$. D'après la Remarque 8.2, le rapport de vraisemblance entre le modèle $\boldsymbol{M}$ et le modèle $\boldsymbol{M'}$ peut s'écrire

$$\begin{aligned}
\frac{L_n}{L'_n} &= \frac{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}} \\
&= \frac{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}\ [\ \frac{\nu_{i_0}\ \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\ b_{i_0}^{Y_0} \cdots b_{i_n}^{Y_n}}{\nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}}\ ]}{\displaystyle\sum_{i_0,\cdots,i_n \in E} \nu'_{i_0}\ \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\ b_{i_0}'^{Y_0} \cdots b_{i_n}'^{Y_n}} \\
&= \mathbb{E}'[\ \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}}\ |\ Y_0, \cdots, Y_n]\ ,
\end{aligned}$$

et compte tenu que la fonction $x \mapsto \log x$ est concave, le logarithme du rapport de vraisemblance est minoré par

$$\begin{aligned}
\log \frac{L_n}{L'_n} &= \log \mathbb{E}'[\ \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}}\ |\ Y_0, \cdots, Y_n] \\
&\geq \mathbb{E}'[\log \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}}\ |\ Y_0, \cdots, Y_n] = Q_n\ ,
\end{aligned}$$

qui s'annule quand le modèle $\boldsymbol{M}$ coïncide avec le modèle $\boldsymbol{M'}$. Maximiser Q_n par rapport aux paramètres (ν, π, b) du modèle $\boldsymbol{M}$ garantit donc que la vraisemblance du modèle qui atteint le maximum de Q_n sera supérieure à la vraisemblance L'_n du modèle courant $\boldsymbol{M'}$. Les formules de re–estimation de Baum–Welch permettent de trouver explicitement les paramètres du nouveau modèle en fonction des paramètres (ν', π', b') du modèle courant $\boldsymbol{M'}$. En répétant cette procédure, on construit une suite de modèles de vraisemblance croissante, et idéalement cette suite converge vers un modèle qui atteint le maximum de la fonction de vraisemblance.

Théorème 10.1 *Dans le cas* symbolique*, l'algorithme itératif pour l'estimation par maximum de vraisemblance des paramètres du modèle au vu des observations* $(Y_0, \cdots, Y_n)$*, est donné par les formules explicites de re–estimation*

$$\nu_i = \vec{p}_0'^{\,i}\ \vec{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\ \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\ \vec{p}_{k-1}'^{\,i}\ b_j'^{Y_k}\ \vec{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \vec{p}_{k-1}'^{\,i}\ \vec{v}_{k-1}'^{\,i}} \qquad \textit{et} \qquad b_i^{\ell} = \frac{\displaystyle\sum_{k=0}^{n} 1(Y_k = \ell)\ \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}\ ,$$

pour tout $i, j \in E$*, et tout* $\ell \in O$*, où les deux suites* $\{\vec{p}'_k\}$ *et* $\{\vec{v}'_k\}$ *sont les solutions normalisées des équations forward et backward respectivement pour les valeurs* (ν', π', b') *des paramètres.*

Remarque 10.2 Concrètement, si $\boldsymbol{M}_{s-1} = (\nu_{s-1}, \pi_{s-1}, b_{s-1})$ désigne le modèle courant à l'étape $(s-1)$ de l'algorithme, alors

- pour les valeurs $(\nu', \pi', b') = (\nu_{s-1}, \pi_{s-1}, b_{s-1})$ des paramètres, on calcule les solutions normalisées $\{\vec{p}'_k\}$ et $\{\vec{v}'_k\}$ des équations forward et backward définies aux Propositions 8.8 et 8.20 respectivement,
- on calcule les paramètres $(\nu_s, \pi_s, b_s) = (\nu, \pi, b)$ grâce aux formules de re–estimation du Théorème 10.1,

ce qui définit le nouveau modèle $\boldsymbol{M}_s = (\nu_s, \pi_s, b_s)$ à l'étape s de l'algorithme.

Preuve. On remarque que

$$\begin{aligned}
Q_n \ &= \ \mathbb{E}'[\log \frac{\nu_{X_0}\ \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\ b_{X_0}^{Y_0} \cdots b_{X_n}^{Y_n}}{\nu'_{X_0}\ \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\ b_{X_0}'^{Y_0} \cdots b_{X_n}'^{Y_n}} \mid Y_0, \cdots, Y_n] \\
&= \ \mathbb{E}'[\log \nu_{X_0} + \sum_{k=1}^{n} \log \pi_{X_{k-1},X_k} + \sum_{k=0}^{n} \log b_{X_k}^{Y_k} \mid Y_0, \cdots, Y_n] + \text{cste} \\
&= \ \sum_{i \in E} \mathbb{P}'[X_0 = i \mid Y_0, \cdots, Y_n]\ \log \nu_i \\
&\quad + \sum_{i,j \in E} \Big\{ \sum_{k=1}^{n} \mathbb{P}'[X_{k-1} = i, X_k = j \mid Y_0, \cdots, Y_n] \Big\}\ \log \pi_{i,j} \\
&\quad + \sum_{i \in E} \sum_{\ell \in O} \Big\{ \sum_{k=0}^{n} 1(Y_k = \ell)\ \mathbb{P}'[X_k = i \mid Y_0, \cdots, Y_n] \Big\}\ \log b_i^{\ell} + \text{cste}\ ,
\end{aligned}$$

et en utilisant les expressions obtenues aux Remarques 8.19 et 8.22, on obtient

$$\begin{aligned} Q_n \;=\; & \sum_{i\in E} \bar{p}_0'^{\,i}\; \bar{v}_0'^{\,i}\; \log \nu_i \\ & + \sum_{i,j\in E} \Big\{ \sum_{k=1}^{n} \frac{1}{c'_k}\; \bar{p}_{k-1}'^{\,i}\; \pi'_{i,j}\; b_j'^{\,Y_k}\; \bar{v}_k'^{\,j} \Big\}\; \log \pi_{i,j} \\ & + \sum_{i\in E} \sum_{\ell\in O} \Big\{ \sum_{k=0}^{n} 1(Y_k=\ell)\; \bar{p}_k'^{\,i}\; \bar{v}_k'^{\,i} \Big\}\; \log b_i^{\ell} + \text{cste}\ . \end{aligned}$$

La maximisation par rapport aux paramètres (ν, π, b) sous les contraintes d'égalité

$$\sum_{i\in E} \nu_i = 1\ , \qquad \sum_{j\in E} \pi_{i,j} = 1 \qquad \text{et} \qquad \sum_{\ell\in O} b_i^{\ell} = 1 \qquad \text{pour tout } i\in E,$$

est explicite, et on obtient les formules de re–estimation

$$\nu_i = \bar{p}_0'^{\,i}\; \bar{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\; \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\; \bar{p}_{k-1}'^{\,i}\; b_j'^{\,Y_k}\; \bar{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \bar{p}_{k-1}'^{\,i}\; \bar{v}_{k-1}'^{\,i}} \qquad \text{et} \qquad b_i^{\ell} = \frac{\displaystyle\sum_{k=0}^{n} 1(Y_k=\ell)\; \bar{p}_k'^{\,i}\; \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \bar{p}_k'^{\,i}\; \bar{v}_k'^{\,i}}\ ,$$

pour tout $i, j \in E$, et tout $\ell \in O$. □

Dans le cas *numérique*, il résulte de la Remarque 8.2 que la fonction de vraisemblance du modèle $\boldsymbol{M} = (\nu, \pi, h, R)$ avec des densités d'émission gaussiennes, admet l'expression suivante

$$L_n = \sum_{i_0,\cdots,i_n\in E} \nu_{i_0}\; \pi_{i_0,i_1} \cdots \pi_{i_{n-1},i_n}\; g_{i_0}(Y_0) \cdots g_{i_n}(Y_n)\ ,$$

et on se propose d'étudier un algorithme itératif pour maximiser la fonction de vraisemblance L_n par rapport aux paramètres (ν, π, h, R) du modèle. Soit $\boldsymbol{M'} = (\nu', \pi', h', R')$ un autre modèle, pour lequel on a déja évalué la fonction de vraisemblance

$$L'_n = \sum_{i_0,\cdots,i_n\in E} \nu'_{i_0}\; \pi'_{i_0,i_1} \cdots \pi'_{i_{n-1},i_n}\; g'_{i_0}(Y_0) \cdots g'_{i_n}(Y_n)\ ,$$

par exemple en terme des solutions $\{p'_k\}$ et $\{v'_k\}$ des équations forward / backward de Baum pour le modèle $\boldsymbol{M'}$. En procédant comme dans le cas *symbolique*, le (logarithme du) rapport de vraisemblance entre le modèle $\boldsymbol{M}$ et le modèle $\boldsymbol{M'}$ est minoré par

$$\log \frac{L_n}{L'_n} \geq \mathbb{E}'[\log \frac{\nu_{X_0}\; \pi_{X_0,X_1} \cdots \pi_{X_{n-1},X_n}\; g_{X_0}(Y_0) \cdots g_{X_n}(Y_n)}{\nu'_{X_0}\; \pi'_{X_0,X_1} \cdots \pi'_{X_{n-1},X_n}\; g'_{X_0}(Y_0) \cdots g'_{X_n}(Y_n)} \mid Y_0, \cdots, Y_n] = Q_n\ ,$$

qui s'annule quand le modèle $\boldsymbol{M}$ coïncide avec le modèle $\boldsymbol{M'}$. Maximiser Q_n par rapport aux paramètres (ν, π, h, R) du modèle $\boldsymbol{M}$ garantit donc que la vraisemblance du modèle qui atteint le maximum de Q_n sera supérieure à la vraisemblance L'_n du modèle courant $\boldsymbol{M'}$. Les formules de

re–estimation de Baum–Welch permettent de trouver explicitement les paramètres du nouveau modèle en fonction des paramètres (ν', π', h', R') du modèle courant $\boldsymbol{M}'$. En répétant cette procédure, on construit une suite de modèles de vraisemblance croissante, et idéalement cette suite converge vers un modèle qui atteint le maximum de la fonction de vraisemblance.

Théorème 10.3 *Dans le cas* numérique *avec des densités d'émission gaussiennes, l'algorithme itératif pour l'estimation par maximum de vraisemblance des paramètres du modèle au vu des observations $(Y_0, \cdots, Y_n)$, est donné par les formules explicites de re–estimation*

$$\nu_i = \vec{p}_0'^{\,i}\ \vec{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\ \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\ \vec{p}_{k-1}'^{\,i}\ g'_j(Y_k)\ \vec{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \vec{p}_{k-1}'^{\,i}\ \vec{v}_{k-1}'^{\,i}}\ ,$$

$$h_i = \frac{\displaystyle\sum_{k=0}^{n} Y_k\ \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}} \qquad \textit{et} \qquad R_i = \frac{\displaystyle\sum_{k=0}^{n} (Y_k - h_i)\,(Y_k - h_i)^*\ \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \vec{p}_k'^{\,i}\ \vec{v}_k'^{\,i}}\ ,$$

pour tout $i, j \in E$, où les deux suites $\{\vec{p}'_k\}$ et $\{\vec{v}'_k\}$ sont les solutions normalisées des équations forward et backward respectivement pour les valeurs (ν', π', h', R') des paramètres.

Remarque 10.4 Concrètement, si $\boldsymbol{M}_{s-1} = (\nu_{s-1}, \pi_{s-1}, h_{s-1}, R_{s-1})$ désigne le modèle courant à l'étape $(s-1)$ de l'algorithme, alors

- pour les valeurs $(\nu', \pi', h', R') = (\nu_{s-1}, \pi_{s-1}, h_{s-1}, R_{s-1})$ des paramètres, on calcule les solutions normalisée $\{\vec{p}'_k\}$ et $\{\vec{v}'_k\}$ des équations forward et backward définies aux Propositions 8.8 et 8.20 respectivement,
- on calcule les paramètres $(\nu_s, \pi_s, h_s, R_s) = (\nu, \pi, h, R)$ grâce aux formules de re–estimation du Théorème 10.3,

ce qui définit le nouveau modèle $\boldsymbol{M}_s = (\nu_s, \pi_s, h_s, R_s)$ à l'étape s de l'algorithme.

PREUVE. On remarque que

$$\begin{aligned} Q_n &= \mathbb{E}'[\log \frac{\nu_{X_0}\ \pi_{X_0,X_1}\cdots\pi_{X_{n-1},X_n}\ g_{X_0}(Y_0)\cdots g_{X_n}(Y_n)}{\nu'_{X_0}\ \pi'_{X_0,X_1}\cdots\pi'_{X_{n-1},X_n}\ g'_{X_0}(Y_0)\cdots g'_{X_n}(Y_n)} \mid Y_0,\cdots,Y_n] \\ &= \mathbb{E}'[\log \nu_{X_0} + \sum_{k=1}^{n} \log \pi_{X_{k-1},X_k} + \sum_{k=0}^{n} \log g_{X_k}(Y_k) \mid Y_0,\cdots,Y_n] + \text{cste} \\ &= \sum_{i\in E} \mathbb{P}'[X_0 = i \mid Y_0,\cdots,Y_n]\ \log \nu_i \\ &\quad + \sum_{i,j\in E} \Big\{ \sum_{k=1}^{n} \mathbb{P}'[X_{k-1} = i, X_k = j \mid Y_0,\cdots,Y_n] \Big\}\ \log \pi_{i,j} \\ &\quad + \sum_{i\in E} \Big\{ \sum_{k=0}^{n} \mathbb{P}'[X_k = i \mid Y_0,\cdots,Y_n]\ \log g_i(Y_k) \Big\} + \text{cste}\ , \end{aligned}$$

et aussi que

$$\begin{aligned} \log g_i(y) &= -\tfrac{1}{2}\ \log\det R_i - \tfrac{1}{2}\ (y-h_i)^*\ R_i^{-1}\ (y-h_i) + \text{cste} \\ &= \tfrac{1}{2}\ \log\det M_i - \tfrac{1}{2}\ \text{trace}[\,(y-h_i)\,(y-h_i)^*\ M_i\,] + \text{cste}\ , \end{aligned}$$

avec $M_i = R_i^{-1}$ pour tout $i \in E$, et tout $y \in \mathbb{R}^d$, et en utilisant les expressions obtenues aux Remarques 8.19 et 8.22, on obtient

$$\begin{aligned} Q_n &= \sum_{i\in E} \bar{p}_0'^{\,i}\ \bar{v}_0'^{\,i}\ \log \nu_i \\ &\quad + \sum_{i,j\in E} \Big\{ \sum_{k=1}^{n} \frac{1}{c'_k}\ \bar{p}_{k-1}'^{\,i}\ \pi'_{i,j}\ g'_j(Y_k)\ \bar{v}_k'^{\,j}\ \Big\}\ \log \pi_{i,j} \\ &\quad + \tfrac{1}{2}\ \sum_{i\in E} \Big\{ \sum_{k=0}^{n} \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}\ \Big\}\ \log\det M_i \\ &\quad - \tfrac{1}{2}\ \sum_{i\in E} \text{trace}[\ \Big\{ \sum_{k=0}^{n} \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}\ (Y_k - h_i)\ (Y_k - h_i)^*\ \Big\}\ M_i\,] + \text{cste}\ . \end{aligned}$$

On rappelle que la dérivée dans la direction D de l'application

$$M \longmapsto a\ \log\det M - \text{trace}(A\,M)\ ,$$

définie sur l'ensemble des matrices inversibles, est égale à

$$a\,\text{trace}(R\,D) - \text{trace}(A\,D) = \text{trace}[\,(a\,R - A)\,D]\ ,$$

où $R = M^{-1}$ par définition. La maximisation par rapport aux paramètres (ν, π, h, R) sous les contraintes d'égalité

$$\sum_{i\in E} \nu_i = 1 \qquad \text{et} \qquad \sum_{j\in E} \pi_{i,j} = 1 \qquad \text{pour tout } i \in E,$$

est explicite, et on obtient les formules de re–estimation

$$\nu_i = \bar{p}_0'^{\,i}\ \bar{v}_0'^{\,i}\ , \qquad \pi_{i,j} = \pi'_{i,j}\ \frac{\displaystyle\sum_{k=1}^{n} \frac{1}{c'_k}\ \bar{p}_{k-1}'^{\,i}\ g'_j(Y_k)\ \bar{v}_k'^{\,j}}{\displaystyle\sum_{k=1}^{n} \bar{p}_{k-1}'^{\,i}\ \bar{v}_{k-1}'^{\,i}}\ ,$$

$$h_i = \frac{\displaystyle\sum_{k=0}^{n} Y_k\ \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}} \qquad \text{et} \qquad R_i = \frac{\displaystyle\sum_{k=0}^{n} (Y_k - h_i)\,(Y_k - h_i)^*\ \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}}{\displaystyle\sum_{k=0}^{n} \bar{p}_k'^{\,i}\ \bar{v}_k'^{\,i}}\ ,$$

pour tout $i, j \in E$. □

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, 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 $\Omega = \mathbb{R}$ et $\mathcal{F}_0$ désigne l'ensemble des *intervalles ouverts* de $\mathbb{R}$, 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 $\mathbb{R}$, telle que pour tout $B \in \mathcal{B}$

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

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

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

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

et

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

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

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

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

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

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

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

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

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

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

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

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 $\alpha, \beta \in \mathbb{R}$ 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

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

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

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

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

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

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 $\psi(y) = 1_{(y\,\in\,B)}$, où $B \in \mathcal{B}^p$. Alors, d'après (A.2)

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

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 l'estimateur du minimum d'erreur quadratique moyenne.

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 l'erreur quadratique moyenne, 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 $\psi(Y)$.

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

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

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*

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

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

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

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

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

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

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

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.

Exemple A.12 Soit X et ε deux variables aléatoires indépendantes, avec X une variable aléatoire gaussienne réduite centrée et ε une variable de Bernoulli à valeurs ± 1 et probabilité $\frac{1}{2}$.

On vérifie d'abord que le vecteur aléatoire $Y = \varepsilon\,X$ est gaussien. En effet, compte tenu que X et ε sont indépendantes, on a

$$\mathbb{E}[\,\exp\{i\,u\,Y\}\mid \varepsilon = +1] = \mathbb{E}[\,\exp\{i\,u\,X\}] = \exp\{-\tfrac{1}{2}\,u^2\}\;,$$

et

$$\mathbb{E}[\,\exp\{i\,u\,Y\}\mid \varepsilon = -1] = \mathbb{E}[\,\exp\{-i\,u\,X\}] = \exp\{-\tfrac{1}{2}\,u^2\}\;,$$

de sorte que

$$\begin{aligned}
\Phi_Y(u) \;&=\; \mathbb{E}[\,\exp\{i\,u\,Y\}] \\
&=\; \mathbb{E}[\,\exp\{i\,u\,Y\}\mid \varepsilon = +1]\;\mathbb{P}[\varepsilon = +1] + \mathbb{E}[\,\exp\{i\,u\,Y\}\mid \varepsilon = -1]\;\mathbb{P}[\varepsilon = -1] \\
&=\; \exp\{-\tfrac{1}{2}\,u^2\}\;,
\end{aligned}$$

pour tout réel u. On reconnait la fonction caractéristique d'une variable aléatoire gaussienne réduite centrée.

Pourtant, bien que chacune des deux variables aléatoires X et Y soit gaussienne, le vecteur aléatoire (X, Y) *n'est pas* gaussien. ou de manière équivalente, quelque soit λ et μ deux réels non–nuls le vecteur aléatoire $Z = \lambda\,X + \mu\,Y = (\lambda + \varepsilon\,\mu)\,X$ n'est pas gaussien. En effet, compte tenu que X et ε sont indépendantes, on a

$$\mathbb{E}[\,\exp\{i\,u\,Z\}\mid \varepsilon = +1] = \mathbb{E}[\,\exp\{i\,u\,(\lambda+\mu)\,X\}] = \exp\{-\tfrac{1}{2}\,(\lambda+\mu)^2\,u^2\}\;,$$

et

$$\mathbb{E}[\,\exp\{i\,u\,Z\}\mid \varepsilon = -1] = \mathbb{E}[\,\exp\{-i\,u\,x(\lambda-\mu)\,X\}] = \exp\{-\tfrac{1}{2}\,(\lambda-\mu)^2\,u^2\}\;,$$

de sorte que

$$\begin{aligned}\Phi_Z(u) &= \mathbb{E}[\exp\{i\,u\,Z\}] \\ &= \mathbb{E}[\exp\{i\,u\,Z\} \mid \varepsilon = +1]\,\mathbb{P}[\varepsilon = +1] + \mathbb{E}[\exp\{i\,u\,Z\} \mid \varepsilon = -1]\,\mathbb{P}[\varepsilon = -1] \\ &= \tfrac{1}{2}\,\exp\{-\tfrac{1}{2}\,(\lambda+\mu)^2\,u^2\} + \tfrac{1}{2}\,\exp\{-\tfrac{1}{2}\,(\lambda-\mu)^2\,u^2\}\ ,\end{aligned}$$

pour tout réel u. On reconnait la fonction caractéristique d'un mélange de deux distributions de probabilité gaussiennes centrées et de variance $(\lambda+\mu)^2$ et $(\lambda-\mu)^2$ respectivement. Sauf si $\lambda\,\mu = 0$, ces deux variances sont distinctes, de sorte que la distribution de probabilité de la variable aléatoire Z *n'est pas* gaussienne.

On énonce le résultat suivant, sans démonstration.

Proposition A.13 *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*

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

Proposition A.14 *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

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

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

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

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

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

indépendamment du caractère gaussien.

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

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

Si $Q_{XY} = 0$, 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.16 *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.15 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. □