REPUBLIQUE ALGERIENNE DEMOCRATIQUE et POPULAIRE.
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU
Faculté des Sciences
Département de Mathématiques

MEMOIRE DE MASTER

en

Mathématiques appliquées

Option Processus aléatoires et statistique de la décision

THEME

Approche Bayésienne des tests sur les modèles autorégressifs vectoriels

Présenté par

ILLA AMADOU Oumoul Hair

Sous la direction de

Pr FELLAG Hocine

Devant le jury d'examen composé de :

Hamadouche Djamel	Professeur	U.M.M.T.O	Président
Fellag Hocine	Professeur	U.M.M.T.O	Rapporteur
Boudiba Mohand Arezki	MCA	U.M.M.T.O	Examinateur
Atil Lynda	MCB	U.M.M.T.O	Examinatrice

Soutenu le 03 / 07 / 2013

Remerciements

Au terme de ce travail, j'adresse mes sincères remerciements à ***M.Fellag Hocine*** *pour l'honneur qu'il m'a fait en assurant la direction, le suivi scientifique et technique; ainsi que la grande contribution qu'il a apporté à l'aboutissement de celui-ci sans oublier l'immense disponibilité dont il a fait preuve et cela malgré son emploi du temps chargé.*
Je noterais par la même occasion mes parents bien aimés ***M. et Mme Amadou Illa*** *pour leur incessante aide et profiterais également pour leur dédier cette modeste oeuvre .*

Mes vifs remerciements s'adressent également à ***M. HAMADOUCHE Djamel*** *pour l'honneur qu'il me fait en acceptant de présider le jury de ce mémoire, ainsi qu'à* ***M. BOUDIBA Mohand Arezki*** *et* ***Melle Atil Lynda*** *pour ce privilège qu'ils m'offrent en acceptant d'examiner ce présent travail. Je ne finirais pas sans pour autant citer tous les enseignants qui m'ont encadrée tout au long de mon cursus et particulièrement ceux du Département des Mathématiques de l'Université Mouloud Mammeri de Tizi-Ouzou, de même que les camarades étudiants et amis du Master pour la collaboration et la discussion tout au long de ce travail.*
Ensemble, nous avons compris que c'est du choc des idées que jaillit la lumière pour l'avancée de la science.

Table des matières

Introduction générale **3**

1 Modèles autorégressifs vectoriels **5**
1.1 Généralités 6
1.1.1 Processus strictement stationnaire 6
1.1.2 Processus faiblement stationnaire 6
1.1.3 Bruit blanc 7
1.1.4 Autocovariance 7
1.1.5 Autocorrélations 7
1.1.6 Opérateur retard 8
1.2 Représentation d'un modèle VAR 8
1.3 Les modèles VAR standards 10
1.3.1 Conditions de stationnarité 10
1.3.2 Représentation VMA du processus VAR 11
1.4 Les modèles VAR structurels 13
1.5 Les caractéristiques d'un processus VAR 14
1.5.1 Espérance 14
1.5.2 Fonction d'autocovariance 14
1.5.3 Fonction d'autocorrélation 15
1.6 Estimation des paramètres d'un modèle VAR 15
1.6.1 Maximum de vraisemblance 15
1.6.2 Détermination du nombre de retards p 16
1.7 Prévision des modèles VAR 18
1.7.1 Cas d'un VAR(1) 19
1.7.2 Cas d'un VAR(p) 20
1.8 Mise en oeuvre pratique avec R 20
1.9 Conclusion 23

2 L'approche Bayésienne et le Facteur de Bayes **24**
2.1 La décision de l'approche Bayésienne 24
2.1.1 Le choix bayésien 25
2.1.2 Notions de base 25
2.1.3 Théorème de Bayes 26
2.2 Une introduction à la théorie de la décision 30
2.2.1 Fonctions de coût usuelles 30
2.2.2 Fonction perte et risque 31

2.2.3 Estimateur de Bayes . . . 31
2.2.4 Admissibilité et minimaxité . . . 33
2.3 Choix des lois a priori . . . 35
2.3.1 Approche partiellement informative . . . 36
2.3.2 Approche non informative . . . 40
2.4 Méthodes de calcul Bayésien . . . 43
2.4.1 Méthode classique d'approximation . . . 44
2.4.2 Méthodes de Monte Carlo par Chaînes de Markov (MCMC) . . . 44
2.5 Avantages de l'approche Bayésienne . . . 45
2.6 Le Facteur de Bayes . . . 47
2.6.1 Facteur de Bayes . . . 47
2.6.2 Le critère de Schwarz . . . 50
2.6.3 Facteur de Bayes contre Critère de Schwarz . . . 52
2.7 Conclusion . . . 52

3 Test Bayésien sur les modèles VAR **54**
3.1 Introduction . . . 54
3.2 Modèle VAR Bayésien . . . 55
3.2.1 Lois a Priori et a posteriori des modèles VARs identifiés . . . 56
3.3 Test Bayésien . . . 59
3.3.1 Estimation de la vraisemblance marginale . . . 59
3.3.2 Comparaison du facteur de Bayes au critère de Schwarz . . . 60
3.4 Exemple numérique . . . 61
3.5 Application économique . . . 63

Conclusion générale **65**

Bibliographie **65**

Introduction générale

"La statistique est la première de science inexactes."
Anonyme.

La statistique doit être considérée comme l'interprétation d'un phénomène naturel, plutôt que son explication. En effet, l'inférence statistique s'accompagne d'une modélisation probabiliste du phénomène observé et implique nécessairement une étape de formalisation réductrice. Sans cette base probabiliste, aucune conclusion utile ne pourra être tirée.

L'objet principal de la Statistique est de faire une inférence sur la distribution probabiliste à l'origine de ce phénomène, grâce à l'observation d'un phénomène aléatoire, c'est-à-dire faire une analyse pour donner une description d'un phénomène passé, ou faire une prédiction d'un phénomène arrivant avec une nature similiaire.

la Statistique est la plupart du temps motivée par un objectif tel que:
- Une usine devrait-elle recruter plus de travailleurs?
- Un laboratoire pharmaceutique devrait-il arrêter la production d'un médicament à cause de quelques critiques?
- Un pays en faillite devrait-il prendre des prêts d'un autre pays ou pas?

Le monde qui nous entoure nous place souvent dans une perspective de décision en univers incertain, cependant, il est nécessaire de prendre des décisions en avenir risqué car, bien souvent, les informations sont insuffisantes pour lever complétement les incertitudes. Dans ce cas, ce qui intéresse un décideur c'est de savoir si, ayant adopté une décision "d", son incertitude sur un paramètre inconnu ne risque pas d'entraîner des conséquences désagréables et dans ce cas s'il importe de choisir une autre décision mieux appropriée. Nous entrons alors dans le monde de la statistique par la porte de la décision sous informations.

L'approche Bayésienne occupe une place de choix dans le monde scientifique. Ce qui nous éloigne considérablement des anciennes réticences à employer l'approche Bayésienne dûes essentiellement à la conception subjectiviste des probabilités que l'on associe à la démarche Bayésienne. C'est aujourd'hui une science mathématique dont l'objectif est de décrire ce qui s'est produit et de faire des projections quant à ce qu'il va advenir dans le futur.

Par ailleurs depuis la nuit des temps l'Homme a toujours cherché à connaître le futur, ou du moins à avoir une idée du futur, pour plusieurs raisons. Aujourd'hui encore, les

raisons socio-économiques impliquent la nécessité d'anticiper l'avenir. Mais, du coup une question se pose. Celle de savoir sur quoi s'appuyer pour prédire le futur. Une série chronologique est une série statistique ordonnée en fonction du temps. On peut dire aussi que c'est la réalisation d'un processus aléatoire indicé par le temps. On modélise un processus par la somme d'une partie déterministe et d'une partie aléatoire (modèle additif), ou par le produit d'une partie déterministe et d'une partie aléatoire (modèle multiplicatif).

Les objectifs d'étude d'un processus sont multiples. La prévision est sans doute le but le plus fréquent. Il s'agit de prévoir les valeurs futures d'une variable grâce aux valeurs observées dans le présent et le passé de cette même variable. Mais, parmi les autres objectifs de l'étude des séries temporelles, figure aussi le problème de l'estimation d'une tendance et l'évaluation de l'impact d'un évenement sur une variable. L'étude des séries temporelles est intéressante car elle peut permettre de prévoir l'évolution du phénomène observé dans le temps. Le domaine des séries chronologiques est en pleine expansion et les notions présentées ici, quoique largement utilisées, ne constituent qu'une petite part des connaissances actuelles sur le sujet.

Ainsi ce mémoire s'articule autour de trois chapitres. Le premier est consacré à l'étude des series chronologiques en général et aux modèles vectoriels autorégressifs en particulier. Nous y présenterons quelques propriétés avec une mise en oeuvre écrit en langage R appliqués à des données réelles. Le second présente l'essentiel de ce qui est nécessaire à tout décideur souhaitant utiliser l'approche Bayésienne dans sa gestion. Et quant au dernier, il est consacré à l'approche Bayesienne des tests dans le cas des modèles VAR. Nous allons, en particulier, aborder la comparaison du critère de Schwarz et du facteur de Bayes.

Dans la littérature courante des VAR, le test d'hypothèse est souvent basé sur le critère de Schwarz qui approxime bien le log du facteur de Bayes quand l'échantillon est de grande taille.

L'objectif principal est d'illustrer, pour une approche Bayesienne des VAR, la performance du facteur de Bayes pour les échantillons finis comparativement au critère de Schwarz largement utilisé à cet effet. Pour ce faire, on s'appuiera sur un article intitulé "Bayesian testing of restrictions on vector autoregressive models" publié, par Dongchu Sun & Shawn Ni, dans Journal of Statistcal Planning and Inference, en 2012 consacré entièrement à ce thème. Ce qui constitue l'essence même de notre présent travail.

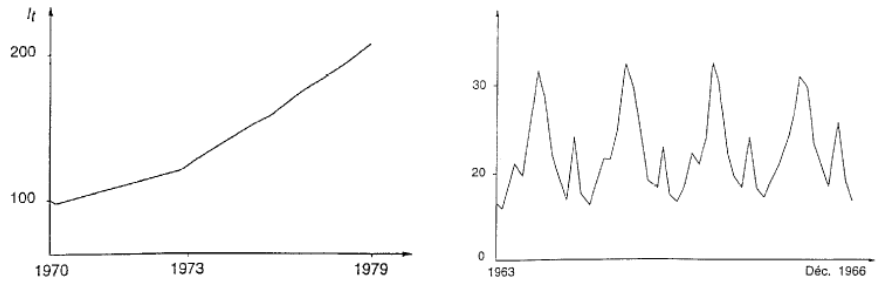
Chapitre 1

Modèles autorégressifs vectoriels

"Prévoir consiste à projeter dans l'avenir ce qu'on a perçu dans le passé."
Henri Bergson, le possible et le réel (1930)

L'étude des séries temporelles (ou séries chronologiques), correspond à l'analyse statistique d'observations régulièrement espacées dans le temps. Appelé aussi séries chronologiques, les séries temporelles occupent une place importante dans tous les domaines de l'observation ou de la collecte des données.
L'étude d'un processus aléatoire à partir d'une série chronologique a généralement deux objectifs :

- Expliquer les variations,
- Prédire les valeurs futures.



(a) Indice mensuel des prix à la consommation I_t.

(b) Trafic voyageur de la SNCF en 2ième classe

FIG. 1.1 – Exemples de series chronologiques

Les processus autorégressifs vectoriels constituent une généralisation des processus univariés. Nous parlons d'approche multivariée lorsque l'analyse et la description portent sur plusieurs variables considérées simultanément. En général il est possible de faire cette étude lorsque les variables en jeu sont en relation de dépendance et sont toutes stationnaires avec un même nombre de décalages. Ainsi, dans ce chapitre nous allons d'abord présenter des concepts généraux des processus aléatoires vectoriels qui seront utiles pour la suite du travail. Puis nous parlerons de modèles autorégressifs vectoriels.

1.1 Généralités

Considérons un vecteur X à n composantes :

$$X_t = \begin{pmatrix} X_{1,t} \\ \vdots \\ X_{n,t} \end{pmatrix}$$

Chacune des composantes $X_{j,t}$, $j = 1,...,n$, $t = 1,...,T$, est une variable aléatoire réelle. La suite $\{X_t,\ t \in \mathbb{Z}\}$ constitue un processus vectoriel à n dimension.

1.1.1 Processus strictement stationnaire

On dit que le processus discret $(...,X_{-1},X_0,X_1,...)$ est stationnaire au sens strict si les vecteurs

$$(X_{t_1},X_{t_2},X_{t_3},\ldots,X_{t_k}) \text{ et } (X_{t_1+h},X_{t_2+h},X_{t_3+h},\ldots,X_{t_k+h})$$

ont même loi $\forall k \in \mathbb{Z}$ et $\forall h \in \mathbb{R}$

Remarque: Notons que la stationarité stricte est très difficile à obtenir, en pratique, sauf dans le cas gaussien. C'est pour cela qu'on définit une stationarité moins restrictive que nous définissons ci-dessous.

1.1.2 Processus faiblement stationnaire

Un processus est dit faiblement stationnaire si :

- si les espérances sont constantes

$$\mathbb{E}(X_t) = \mu \qquad \forall t$$

et

- si les covariances sont constantes par translation dans le temps

i.e pour tout h,

$$Cov(X_t,X_{t+h}) = \sigma(h) \qquad \forall t$$

où la covariance est définie par

$$Cov(X_t,X_{t+h}) = \mathbb{E}(X_t,X_{t+h}) - \mathbb{E}(X_t)\mathbb{E}(X_{t+h})$$

Cela veut dire que les variable X_t sont dans $\mathbb{L}^2$. Une telle suite est appelée processus stationnaire du second ordre.
On remarque que, dans ce cas, que $Var(X_t) = \sigma(0)$ pour tout t. Donc la variance est constante.

En général, un processus est dit stationnaire au second ordre si ses moments d'ordre 1 et 2 sont constants dans le temps :

$$E(X_t) = \mu, \quad V(X_t) = \sigma^2 \quad Cov(X_t,X_{t-h}) = \gamma_h.$$

Remarque: La stationnarité stricte entraîne la stationnarité faible.

1.1.3 Bruit blanc

Un processus (ϵ_t) est un bruit blanc s'il constitue un échantillon non corrélé:

$$\forall t : \epsilon_t \sim F, \qquad E(\epsilon_t) = 0,$$

$$\forall t \neq s : \qquad Cov(\epsilon_t,\epsilon_s) = 0.$$

Remarque: Dans la majorité des cas, le bruit blanc est gaussien et les variables ϵ_t sont indépendantes. Cependant, dans la littérature de recherche, on trouve que ces variables peuvent suivre d'autres lois (ex. exponentielle, gamma,..) et sont même très souvent corellées (autoregressives,..etc)

1.1.4 Autocovariance

Soit X_t (notation simplifié pour $\{X_t,t \in T\}$), un processus, r et s deux instants. L'autocovariance (ou fonction d'autocovariance) de X_t pour les deux instants r et s est, par définition, la covariance des variables X_r et X_s.
L'autocovariance s'écrit γ et:

$$\gamma(r,s) = Cov(X_r,X_s)$$

1.1.5 Autocorrélations

La notion d'autocorrélation découle de la notion d'autocovariance comme la corrélation de la covariance. On étudie la "mémoire" d'un processus en calculant son **autocorrélation** de retard h noté ρ_h :

$$\rho_h = Corr(X_t,X_{t-h}) = \frac{Cov(X_t,X_{t-h})}{\sqrt{V(X_t)V(X_{t-h})}}$$

qui mesure le lien entre les valeurs du processus à deux dates distances de h. Pour un processus stationnaire, ρ_h prend une forme simple :

$$\rho_h = \frac{Cov(X_t,X_{t-h})}{V(X_t)} = \frac{\gamma_h}{\gamma_0}.$$

On peut tracer la courbe ρ_h = f(h) qui est appelée (auto-)corrélogramme.

Autocorrélation partielle

De même, on définit l'**autocorrélation partielle** de retard h comme la corrélation entre $(X_t - X_t^*)$ et $(X_{t-h} - X_{t-h}^*)$ où X_t^* désigne la régression de X_t sur les valeurs (h-1) valeurs $\{X_{t-1},X_{t-2},...,X_{t-h+1}\}$:

$$\tau_h = Corr(X_t - X_t^*,X_{t-h} - X_{t-h}^*) = \frac{Cov(X_t - X_t^*,X_{t-h} - X_{t-h}^*)}{\sqrt{V(X_t - X_t^*)V(X_{t-h}X_{t-h}^*)}}$$

avec

$$X_t^* = \sum_{k=1}^{h-1} \alpha_k X_{t-k}, \quad X_{t-h}^* = \sum_{k=1}^{h-1} \beta_k X_{t-k}$$

où les α_k et les β_k sont les coefficients des régressions. Cette quantité rend compte de l'intensité de la liaison entre X_t et X_{t-h} en supprimant les liaisons induites par variables intermédiaires $\{X_{t-1},X_{t-2},...,X_{t-h+1}\}$. On peut remarquer que pour tout processus,

$$\rho_1 = \tau_1$$

puisque qu'il n'y a aucune variable intermédiaire entre X_t et X_{t-1}.

1.1.6 Opérateur retard

La manipulation pratique ou théorique des séries temporelles se trouve considérablement simplifiée par l'usage de l'opérateur retard (*Lag operator*). On note indifféremment B (backwards) ou L (lag), l'opérateur qui fait passer de x_t à x_{t-1} :

$$B(x_t) = x_{t-1}.$$

On a :

$$B^2(x_t) = B(B(x_t)) = B(x_{t-1}) = x_{t-2}.$$

Si on applique h fois cet opérateur, on décale le processus de h unités de temps.

$$B(B(...B(x_t)...)) = B^h(x_t) = x_{t-h}.$$

1.2 Représentation d'un modèle VAR

Depuis les premiers travaux de Sim, les techniques économétriques sur les modèles VAR (Vector Autoregressive) ont connus des avancées considérables. Leur popularité est due à leur caractère flexible et leur facilité d'utilisation pour produire des modèles ayant des caractéristiques descriptives utiles.

Considérons deux processus stationnaires $y_{1,t}$ et $y_{2,t}$. Chaque variable est fonction de ses propres valeurs passées mais aussi des valeurs passées et présentes des autres variables. Le modèle VAR d'ordre p noté $VAR(p)$ décrivant ces deux variables s'écrit:

$$y_{1,t} = a_1 + \sum_{i=1}^{p} b_{1,i} y_{1,t-i} + \sum_{i=1}^{p} c_{1,i} y_{2,t-i} - d_1 y_{2,t} + \epsilon_{1,t} \tag{1.1}$$

$$y_{2,t} = a_2 + \sum_{i=1}^{p} b_{2,i} y_{1,t-i} + \sum_{i=1}^{p} c_{2,i} y_{2,t-i} - d_2 y_{1,t} + \epsilon_{1,t} \qquad (1.2)$$

Où $\epsilon_{1,t}$ et $\epsilon_{2,t}$ sont deux bruits blancs non corrélés de variances respectives respectives σ_1^2 et σ_2^2.
Sous forme matricielle, ce processus $VAR(p)$ s'écrit :

$$BY_t = \Phi_0 + \sum_{i=1}^{p} \Phi_i Y_{t-i} + \epsilon_t \qquad (1.3)$$

avec :

$$B = \begin{pmatrix} 1 & d_1 \\ d_2 & 1 \end{pmatrix} \qquad \Phi_0 = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \qquad Y_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix}$$

$$\epsilon_t = \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \quad \Phi_i = \begin{pmatrix} b_{1,i} & c_{1,i} \\ b_{2,i} & c_{2,i} \end{pmatrix} \quad \forall i \in [1,p]$$

Ce système initial donné par les équations (1.1) et (1.2), ou par la définition matricielle (1.3) est qualifié de représentation structurelle.
On constate que dans cette représentation le niveau $y_{2,t}$ a un effet immédiat sur $y_{1,t}$ et vice versa. L'estimation des paramètres de ce modèle suppose donc estimer $(4*(p+1)+2+2)$ paramètres.
C'est pourquoi on travaille généralement à partir de la forme réduite du modèle VAR. Ce modèle obtenu en multipliant les deux membres de (1.3) par B^{-1}, s'écrit alors sous la forme :

$$Y_t = \bar{\Phi}_0 + \sum_{i=1}^{p} \bar{\Phi}_i Y_{t-i} + v_t \qquad (1.4)$$

avec :

$$\begin{aligned} \bar{\Phi}_i &= B^{-1}\Phi_i \quad \forall i \in [0,p] \\ v_t &= B^{-1}\epsilon_t \quad \forall t \in \mathbb{Z} \end{aligned}$$

Ce qui peut s'écrire sous la forme :

$$y_{1,t} = \bar{a}_1 + \sum_{i=1}^{p} \bar{b}_{1,i} y_{1,t-i} + \sum_{i=1}^{p} \bar{c}_{1,i} y_{2,t-i} + v_{1,t} \qquad (1.5)$$

$$y_{2,t} = \bar{a}_2 + \sum_{i=1}^{p} \bar{b}_{2,i} y_{1,t-i} + \sum_{i=1}^{p} \bar{c}_{2,i} y_{2,t-i} + v_{2,t} \qquad (1.6)$$

On qualifie la représentation donnée par (1.4), ou par les équations (1.5) et (1.6) de processus $VAR(p)$ standard.
Dans cette représentation on constate que le niveau de $y_{2,t}$ ne dépend plus directement de $y_{1,t}$, mais seulement des valeurs passées de $y_{2,t}$ et de $y_{1,t}$, et de l'innovation $v_{2,t}$. On distingue deux types des modèles VAR : Les modèles VAR standard et structurel.

1.3 Les modèles VAR standards

La modélisation VAR standard consiste à modéliser un vecteur de variables stationnaires à partir de sa propre histoire et où donc chaque variable est expliquée par le passé de l'ensemble des autres variables, cette forme est caractérisée par :

- les variables à modéliser sont toutes stationnaires.
- les variables à modéliser sont toutes potentiellement endogènes.
- le nombre de décalage associé à chaque équation est identique.

Définition 1.3.1. *Un processus vectoriel* $\{X_t,\ t \in \mathbb{Z}\}$*, de dimension n, admet une représentation* VAR *d'ordre p, notée* $VAR(p)$ *si :*

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \epsilon_t \tag{1.7}$$

où c désigne un vecteur de constantes de dimension n, et $\{\epsilon_t,\ t \in \mathbb{Z}\}$ *est un bruit blanc vectoriel de matrice de covariance* Σ*.* Φ_1,$\Phi 2, \ldots$,Φ_p *représentent les paramètres du modèle, matrices réelles satisfaisant* $\Phi_0 = I_n$ *et* $\Phi_p \neq 0$*. Nous pouvons réécrire l'équation précédente avec l'opérateur de retard L comme suit :*

$$(I_n - \Phi_1 L - \Phi_2 L^2 - \Phi_p L^p) X_t = c + \epsilon_t$$

ou encore

$$\Phi(L) X_t = c + \epsilon_t$$

où $\phi(L)$ *est le polynôme matriciel de retard de degré p, donné par :*

$$\Phi(L) = I_n - \sum_{i=1}^{p} \Phi_i L^i \tag{1.8}$$

1.3.1 Conditions de stationnarité

La notion de stationnarité joue un rôle fondamental dans les applications économétriques, plus particulièrement dans l'étude de méthodes qui trouveraient des applications dans l'élaboration d'un système de prévision.

Le théorème suivant, appelé couramment théorème de Wold, donne les conditions de stationnarité d'un processus $VAR(p)$

Théorème 1.3.1. *Brockwell et Davis (1991).*
Le processus $VAR(p)$ *représenté par l'équation* (1.8) *est stationnaire si :*

$$\det[\Phi(z)] \neq 0, \text{ pour tout } z \in \mathbb{C} \text{ tel que } |z| \leq 1.$$

Dans ce cas, le processus $VAR(p)$ *admet la représentation linéaire suivante :*

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} = \Psi(L)\epsilon_t,$$

où les matrices Ψ_i sont déterminées par :

$$\Psi(L) = \sum_{i=0}^{\infty} \psi_i L^i = \Phi^{-1}(L),$$

et les coefficients $\{\psi_i\}$ sont absolument sommables, c'est-à-dire que :

$$\sum_{i=0}^{\infty} ||\psi_i|| < \infty,$$

avec $\|.\|$ correspondant à la norme d'une matrice en général euclidienne ou spectrale.

1.3.2 Représentation VMA du processus VAR

Comme en univarié, on constate donc que tout processus stationnaire admet une représentation moyenne mobile infinie.

Proposition 1.3.1. *(Brockwell and Davis, 1991)*
Tout processus $\{X_t,\, t \in \mathbb{Z}\}$, de dimension n, stationnaire, satisfaisant une représentation $VAR(p)$ admet une représentation moyenne mobile convergente définie par :

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} = \mu + \Psi(L)\epsilon_t$$

avec $\mu = E(X_t) = \Phi(L)^{-1}c$, où ϵ_t est un bruit blanc vectoriel et où la séquence des matrices de dimension $(n \times n)$, $\{\psi_i\}_{i=0}^{\infty}$ satisfait $\psi_0 = I_n$ et est absolument sommable au sens où les éléments $\psi_{j,k}^i$ de ψ_i satisfont la condition :

$$\sum_{s=0}^{\infty} |(\psi_{j,k}^i)^s| < \infty \quad \forall i \geq 1,\ \ \forall (j,k) \in \mathbb{N}^2.$$

IL est possible de déterminer de façon générale la forme des matrices de l'opérateur polynômial associée à la représentation $VMA(\infty)$.

Proposition 1.3.2 (Christophe HURLIN, 1998). *Le polynôme matriciel $\Psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ associé à la reprèsentation $VMA(\infty)$ d'un processus stationnaire $\{X_t,\ t \in \mathbb{Z}\}$ telle que*

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \epsilon_t$$

satisfait la récurrence suivante :

$$\begin{cases} \psi_0 = I_n \\ \psi_1 - \psi_0\Phi_1 = 0 \\ \psi_2 - \psi_1\Phi_1 - \psi_0\Phi_1 = 0 \\ \vdots \\ \psi_i - \sum_{j=1}^{i} \psi_{i-j}\Phi_j = 0 \end{cases}$$

où $\Phi_j = 0$ pour $j > p$.

Preuve. *Une façon simple d'obtenir la représentation $VMA(\infty)$ d'un processus $VAR(p)$ consiste à identifier les polynômes matriciels $\Phi(L)^{-1}$ et $\Psi(L)$.*
En effet, supposons que le processus soit centré ($c = 0$), on a :

$$X_t = \Phi(L)^{-1}\epsilon_t = \Psi(L)\epsilon_t \Leftarrow : \Phi(L)\Psi(L) = I_n$$

Cette égalité peut se réécrire sous la forme :

$$\lim_{k\to\infty}[(I_n - \Phi_1 L - \Phi_2 L^2 - ... - \Phi_p L^p)(I_n - \Psi_1 L - \Psi_2 L^2 - ... - \Psi_p L^p - ... - \Psi_k L^k)] = I_n$$

Ainsi par identification des termes de même ordre, on montre que les matrices de l'opérateur polynômial associées à la forme $VMA(\infty)$ satisfont bien à une équation de récurrence correspondant à celle de la proposition précédente.

Soit un processus bivarié stationnaire $\{Y_t,\, t \in \mathbb{Z}\}$ admettant une représentation $VAR(1)$ telle que :

$$\Phi(L)Y_t = c + \epsilon_t$$

avec $\epsilon_t \sim N(0,\Sigma)$, $c = (3,1)'$ et

$$\Phi(L) = \Phi_0 + \Phi_1 L = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} -0.2 & -0.7 \\ -0.3 & -0.4 \end{pmatrix} L = \begin{pmatrix} 1 - 0.2L & -0.7L \\ -0.3L & 1 - 0.4L \end{pmatrix}$$

Par application du théorème de Wold, on sait que ce processus peut être représenté sous une forme $VMA(\infty)$ telle que :

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} = \mu + \Psi(L)\epsilon_t$$

Immédiatement, on montre que

$$\mu = E(Y_t) = \begin{pmatrix} 0.8 & -0.7 \\ -0.3 & 0.6 \end{pmatrix}^{-1} \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 9.25 \\ 6.29 \end{pmatrix}$$

Par définition, on a $\Phi(L)\Psi(L) = I_2$, ce qui peut se récrire sous la forme :

$$\lim_{k\to\infty}[(I_2 - \Phi_1 L)(\Psi_0 - \Psi_1 L - \Psi_2 L^2 - ... - \Psi_p L^p - ... - \Psi_k L^k)] = I_2$$

Par identification des membres de même terme, on montre que :

$$\begin{aligned} \Psi_0 &= I_2 \\ \Psi_1 &= \Phi_1 = \begin{pmatrix} -0.2 & -0.7 \\ -0.3 & -0.4 \end{pmatrix} \\ \Psi_2 &= \Phi_1\Psi_1 = \Phi_1^2 = \begin{pmatrix} -0.2 & -0.7 \\ -0.3 & -0.4 \end{pmatrix}^2 \end{aligned}$$

et de façon générale, on a :

$$\Psi_n = \Phi_1\Psi_{n-1} = \Phi_1^n = \begin{pmatrix} -0.2 & -0.7 \\ -0.3 & -0.4 \end{pmatrix}^n \quad \forall n \geq 1$$

On retrouve ainsi la formule générale que l'on avait établit précédemment :

$$Y_t = \mu + \Psi(L)\epsilon_t = \mu + \sum_{i=0}^{\infty} \Phi_1^i \epsilon_{t-i}$$

ainsi on a :

$$Y_t = \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} = \begin{pmatrix} 9.25 \\ 6.29 \end{pmatrix} + \sum_{i=0}^{\infty} \begin{pmatrix} -0.2 & -0.7 \\ -0.3 & -0.4 \end{pmatrix}^i \begin{pmatrix} \epsilon_{1,t-i} \\ \epsilon_{2,t-i} \end{pmatrix}$$

1.4 Les modèles VAR structurels

Le processus VAR structurel est une représentation utile dérivée du VAR standard. Soit w_t le vecteur des chocs structurels. Il s'agit de chocs interprétables économiquement. On suppose ainsi que l'économie, représentée par un vecteur de séries observables $X_t = (X_{1t}, ..., X_{nt})'$ à chaque date t, résulte de la combinaison dynamique de n chocs structurels passés $(w_{1s}, ..., w_{ns}),\ s \leq t$.
La représentation VAR structurel se deduit de celle de VAR standard en supposant que le vecteur des innovations standards ϵ_t est une combinaison linéaire des innovations structurelles w_t de la même date :

$$\epsilon_t = Pw_t$$

où P est une matrice de passage (inversible et de dimension $n \times n$) qui doit être estimée. Si l'on part de la représentation standard :

$$X_t = \sum_{i=1}^{p} \Phi_i X_{t-i} + \epsilon_t$$

et que l'on prémultiplie les deux membres par la matrice $\hat{P}^{-1}$ ($\hat{P}$ étant un estimateur de P) :

$$\hat{P}^{-1} X_t = \hat{P}^{-1} \sum_{i=1}^{p} \Phi_i X_{t-i} + \hat{P}^{-1} \epsilon_t$$

on déduit l'expresssion du processus VAR structurel :

$$X_t = \sum_{i=0}^{p} \Psi X_{t-i} + w_t$$

avec : $w_t = \hat{P}^{-1}\epsilon_t$, $\Psi_0 = I - \hat{P}^{-1}$ et $\Psi_i = \hat{P}^{-1}\Phi_i$ pour $1 \leq i \leq p$.
On constate que l'estimation du modèle VAR structurel est acquise dès que la matrice P a été estimée. De même, dès que cette matrice P est estimée, l'identification des chocs est realisée puisqu'il est alors possible de passer des chocs estimés aux chocs structurels (interprétables économiquement) par :

$$\hat{w}_t = \hat{P}^{-1}\hat{\epsilon}_t$$

1.5 Les caractéristiques d'un processus VAR

Considérons un processus $VAR(p)$:

$$X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \epsilon_t$$

1.5.1 Espérance

on a :

$$\mathbb{E}(X_t) = \mathbb{E}(X_t = c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p} + \epsilon_t)$$

Le processus étant stationnaire, on a : $\mathbb{E}(X_t) = \mathbb{E}(X_{t-1}) = ... = \mathbb{E}(X_{t-p})$.
On peut donc écrire (sachant que $\mathbb{E}(\epsilon_t) = 0$):

$$\mathbb{E}(X_t) = c + \Phi_1 \mathbb{E}(X_t) + \Phi_2 \mathbb{E}(X_t) + ... + \Phi_p \mathbb{E}(X_t)$$

D'où :

$$\mathbb{E}(X_t) = (I - \Phi_1 - \Phi_2 - ... - \Phi_p)^{-1} c = \mu$$

1.5.2 Fonction d'autocovariance

Commençons par illustrer le cas d'un $VAR(1)$.
Soit un processus centré $(c = 0)$: $X_t = \Phi_1 X_{t-1} + \epsilon_t$
La fonction d'autocovariance Γ est donnée par :

$$\Gamma(0) = \mathbb{E}(X_t X_t^{'}) = \mathbb{E}(\Phi_1 X_{t-1} X_t^{'} + \epsilon_t X_t^{'})$$

Or,

$$\mathbb{E}(\epsilon_t X_t^{'}) = \mathbb{E}[\epsilon_t (\Phi_1 X_{t-1} + \epsilon_t)^{'}] = \Phi_1 \mathbb{E}(\epsilon_t X_{t-1}^{'}) + \mathbb{E}(\epsilon_t \epsilon_t^{'})$$

Et, comme $\epsilon_t \sim BB(0, \Sigma)$, on a :

$$\mathbb{E}(\epsilon_t X_{t-1}^{'}) = 0$$

et par suite

$$\mathbb{E}(\epsilon_t X_t^{'}) = \mathbb{E}(\epsilon_t \epsilon_t^{'}) = \Sigma$$

D'où

$$\Gamma(0) = \phi_1 \mathbb{E}(X_{t-1} X_t^{'}) + \Sigma$$

En remarquant que $\mathbb{E}(X_{t-1} X_t^{'}) = \Gamma(1)$, on en déduit :

$$\Gamma(0) = \Phi_1 \Gamma(1)^{'} + \Sigma$$

On calcule la matrice d'autocovariance d'ordre 1 :

$$\begin{aligned} \Gamma(1) &= \mathbb{E}(X_t X_{t-1}^{'}) = \mathbb{E}[(\Phi_1 X_{t-1} + \epsilon_t) X_{t-})^{'}] = \Phi_1 \mathbb{E}(X_{t-1} X_{t-1}^{'}) \\ &= \Phi_1 \Gamma(0) \end{aligned}$$

Pour un ordre h on en déduit la formule de récurrence suivante :

$$\Gamma(h) = \Phi_1 \Gamma(h-1) \quad \forall h \geq 1$$

Cas d'un VAR(p)
Pour un $VAR(p)$, cette récurrence se généralise par :

$$\Gamma(h) = \Phi_1 \Gamma(h-1) + \Phi_2 \Gamma(h-2) + ... + \Phi_p \Gamma(h-p).$$

1.5.3 Fonction d'autocorrélation

La fonction d'autocorrélation de retard h d'un processus $VAR(p)$ de moyenne $\mathbb{E}(X_t) = \mu$ et de matrice de covariance $\Gamma(h)$, notée $R(h) = (\rho_{ij}(h))$ est définie par :

$$\rho_{ij}(h) = \frac{\gamma_{ij}(h)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}$$

ou encore sous forme matricielle

$$R(h) = D^{-1}\Gamma(h)D^{-1} \ \ \forall h \in \mathbb{Z}$$

où

$$D^{-1} = diag\left(\frac{1}{\sqrt{\gamma_{11}(0)}},...,\frac{1}{\sqrt{\gamma_{nn}(0)}}\right)$$

1.6 Estimation des paramètres d'un modèle VAR

Tous comme les processus AR univariés plusieurs méthodes d'estimations sont envisageables pour les processus VAR. Le modèle peut être convénablement estimé soit par mco (moindres carrés ordinaires), soit par maximum de vraisemblance.

1.6.1 Maximum de vraisemblance

Considérons un processus $\{X_t,\, t \in \mathbb{Z}\}$ satisfaisant la représentation $VAR(p)$ suivante :

$$\Phi(L)X_t = X_t - \Phi_1 X_{t-1} - \Phi_2 X_{t-2} - ... - \Phi_p X_{t-p} = c + \epsilon_t$$

On suppose que les innovations ϵ_t sont $iid\ \mathcal{N}(0, \Sigma)$ et que l'on dispose de $T + p$ observations du processus X_t. On cherche à déterminer la vraisemblance conditionnelle de X_t en fonction des réalisations passées X_{t-i}, $\forall i \in [1, p]$.
Par définition, la distribution conditionnelle de X_t s'écrit :

$$D(X_t/X_{t-1},X_{t-2},...,X_{t-p}) \sim \mathcal{N}(c + \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + ... + \Phi_p X_{t-p}, \Sigma)$$

Afin de simplifier les calculs, posons :

$$\bar{X}_t = \begin{pmatrix} 1 \\ X_{t-1} \\ \vdots \\ X_{t-p} \end{pmatrix} \qquad \Pi' = \begin{pmatrix} c \\ \Phi_1 \\ \vdots \\ \Phi_p \end{pmatrix}$$

On a alors

$$D(X_t/X_{t-1},X_{t-2},...,X_{t-p}) \sim \mathcal{N}(\Pi'\bar{X}_t, \Sigma)$$

Soit $\Theta = (\Pi', \Sigma)$ le vecteur des paramètres à estimer alors la densité conditionnelle de X_t s'écrit :

$$f(X_t/X_{t-1},X_{t-2},...,X_{t-p};\, \Theta) = \frac{1}{(\sqrt{2})^n\sqrt{\det \Sigma}} \times \exp\left[-\frac{1}{2}(X_t - \Pi'\bar{X}_t)'\Sigma^{-1}(X_t - \Pi'\bar{X}_t)\right]$$

On déduit donc la vraisemblance de l'échantillon $\{X_t\}_{t=1}^T$ conditionnellement aux valeurs initiales qui est :

$$f(X_t,X_{t-1},...,X_1/X_{t-1},X_{t-2},...,X_{t-p};\,\Theta) = \prod_{t=1}^{T} f(X_t/X_{t-1},X_{t-2},...,X_{t-p};\,\Theta)$$

Il s'en suit que la log-vraisemblance du processus VAR notée $\mathcal{L}(\Theta)$ est donnée par :

$$\begin{aligned}\mathcal{L}(\Theta) &= \sum_{t=1}^{T} \log[f(X_t/X_{t-1},X_{t-2},...,X_{t-p};\,\Theta)] \\ &= -\frac{nT}{2}\log 2\pi - \frac{T}{2}\log\det\Sigma - \frac{1}{2}\sum_{t=1}^{T}\left[(X_t - \Pi'\bar{X}_t)'\Sigma^{-1}(X_t - \Pi'\bar{X}_t)\right]\end{aligned}$$

On maximise ensuite cette expression afin d'obtenir les estimateurs convergents des paramètres Π' et de la matrice de variance covariance des innovations Σ

1.6.2 Détermination du nombre de retards p

La détermination du nombre de retards optimal pour un $VAR(p)$ peut être faite en utilisant toutes les méthodes de comparaison de modèles étudiées au cas univaríe. Tout au plus, on pourrait raisonnablement dire que le nombre de décalage à retenir est celui qui permet de modéliser adéquatement le vecteur X et d'aboutir à un vecteur d'innovations qui soit un bruit blanc.
De ce fait, trois méthodes sont plus généralement retenues dans la littérature, qui sont :

- Une méthode basée sur l'examen des propriétés statistiques des innovations du VAR.
- Une méthode basée sur les tests de nullité sur les paramètres associés au dernier décalage.
- Une méthode basée sur les critères d'informations qui est d'ailleurs la plus utilisée en pratique.

Estimation de l'ordre p à partir de l'examen de résidus

Cette méthode consiste à vérifier la blancheur des résidus du modèle VAR successivement estimés pour $p = 1,2,3,...$
L'appellation "blancheur des résidus" est choisie ici par abus de langage voulant dire le caractère "bruit blanc" à considerer par son aspect technique et non littéraire. En partant de VAR minimal ($p = 1$) il suffit d'arrêter la procédure au nombre p pour lequel les résidus sont de type bruit blanc. Cette blancheur des bruits blancs peut être examinée à partir des différents tests de non-autocorrélation (Ljung-Box, Box-Pierce,...), de normalité (Jarque Bera) et d'abscence d'hétérocédasticité conditionnelle etc.

Estimation de nombre de retards p à partir de tests de rapport de vraisemblance

On peut effectuer des tests sur l'ordre p du VAR. Considérons le test suivant :
H_0 : $\Phi_{p+1} = 0$: processus $VAR(p)$

$H_1 : \Phi_{p+1} \neq 0$: processus $VAR(p+1)$
La matrice d'information de Fisher est difficile à calculer, ce qui explique l'utilisation du test du rapport de maximum de vraisemblance. La technique consiste à estimer un modèle contraint $VAR(p)$ et un modèle non contraint $VAR(p+1)$ et à effectuer le rapport des log-vraisemblances. Rappelons que la log-vraisemblance d'un processus VAR s'écrit :

$$\mathcal{L}(\Theta) = -\frac{nT}{2}\log 2\pi - \frac{T}{2}\log\det\Sigma - \frac{1}{2}\sum_{t=1}^{T}\epsilon_t'\Sigma^{-1}\epsilon_t$$

$\sum_{t=1}^{T}\epsilon_t'\Sigma^{-1}\epsilon_t$ est un scalaire, on a donc, en notant Tr la trace :

$$\begin{aligned}\sum_{t=1}^{T}\epsilon_t'\Sigma^{-1}\epsilon_t &= Tr\left(\sum_{t=1}^{T}\epsilon_t'\Sigma^{-1}\epsilon_t\right) = Tr\left(\Sigma^{-1}\sum_{t=1}^{T}\epsilon_t'\epsilon_t\right)\\ &= Tr\left(T\Sigma^{-1}\frac{1}{T}\sum_{t=1}^{T}\epsilon_t'\epsilon_t\right)\\ &= Tr(T\Sigma^{-1}\Sigma) = Tr(TI_n) = nT\end{aligned}$$

Soient $\log L^c$ la log-vraisemblance estimée du modèle contraint :

$$\log L^c = -\frac{nT}{2}\log 2\pi - \frac{T}{2}\log\det\hat{\Sigma}^c - \frac{1}{2}nT$$

et $\log L^{nc}$ la log-vraisemblance estimée du modèle non contraint :

$$\log L^{nc} = -\frac{nT}{2}\log 2\pi - \frac{T}{2}\log\det\hat{\Sigma}^{nc} - \frac{1}{2}nT$$

où $\hat{\Sigma}^c$ (respectivement $\hat{\Sigma}^{nc}$) désigne l'estimateur de la matrice de variance covariance des résidus du modèle contraint (respectivement non contraint).
On calcule la statistique de test $\xi = T \times RMV$ où RMV désigne le rapport du maximum de vraisemblance :

$$\xi = T\log\left(\frac{\det\hat{\Sigma}^c}{\det\hat{\Sigma}^{nc}}\right)$$

sous l'hypothèse nulle, cette statistique suit une loi de Khi-deux à r degrés de liberté où r désigne le nombre de contraintes.

Estimation d'ordre p de retards à partir des critères d'information

En régression linéaire comme en modélisation de séries temporelles, on est amené à choisir entre différents modèles, emboîtés ou non. Il est classique d'utiliser pour cela un critère d'information.

Critères d'information

Différents critères sont fournis par les logiciels de statistique : AIC (Akaike's Information Criterion), SC(Schwarz's Criterion)... Ils sont Basés sur la notion de vraisemblance.

En principe le modèle pour lequel le critère a la plus faible valeur est le mieux adapté. Cependant le critère AIC favorise les modèles sur-paramétrés; il est donc à utiliser avec précaution. On lui préférera le critère SC qui a des meilleurs propriétés. De façon générale, il vaut mieux éviter d'employer un quelconque critère "en aveugle"; il s'agit plutôt d'un *outil complémentaire* de sélection. Etant donné plusieurs modèles et leurs estimations, et un critère étant choisi, on retient le modèle pour lequel le critère est minimum. Nous donnons quelques détails sur deux de ces critères.

- **AIC (Critère d'information d'Akaike).**L'AIC (Akaike's Information Criterion) est :
$$AIC(k) = -2ln(L) + 2k,$$
où L est la fonction de vraisemblance évaluée en les estimations par maximum de vraisemblance (MV) des paramètres, et k est le nombres de paramètres estimés. On rencontre quelque fois une autre expression de l'AIC: (-2ln(L) + 2k)/T où T est le nombre d'observations.
- **SC (Critère de Schwarz).** Le SC (Schwarz's Criterion) ou BIC (Bayesian Information Criterion) est :
$$SC(k) = -2ln(L) + 2kln(T).$$
Si les erreurs sont normalement distribuées, il prend la forme
$$SC(k) = Tln(\widehat{\sigma}^2) + kln(T).$$

On voit que ces critères (AIC et SC) sont formés de deux termes :

- Le terme -2ln(L) qui est d'autant plus faible que l'ajustement par maximum de vraisemblance est bon;
- Le terme 2k ou 2kln(T) qui pénalise cette faible valeur en l'augmentant par une fonction croissante de nombre de paramètres estimés.

Afin de déterminer l'ordre p du processus VAR on peut également utiliser des critères d'information. Ainsi, on estime un certain nombre de modèles VAR pour un ordre p allant de 0 à h où h est le retard maximun. On retient le retard p qui minimise les critères AIC, SC et Hannan-Quinn(HQ) définis comme suit :

$$\begin{aligned} AIC &= \log\det\hat{\Sigma} + \frac{2n^2p}{T} \\ SC &= \log\det\hat{\Sigma} + n^2p\frac{\log T}{T} \\ HQ &= \log\det\hat{\Sigma} + n^2p\frac{2\log(\log T)}{T} \end{aligned}$$

où n est le nombre de variables du système, T est le nombre d'observations et $\hat{\Sigma}$ est un estimateur de la matrice de variance-covariance des résidus.

1.7 Prévision des modèles VAR

La prévision de X à l'horizon h est traditionnellement définie comme l'espérance conditionnelle de X_{t+h} calculée en utilisant l'information disponible au temps t.

1.7.1 Cas d'un VAR(1)

Considérons un modèle VAR d'ordre 1 : $X_t = c + \Phi_1 X_{t-1} + \epsilon_t$ et supposons que l'on dispose de T réalisations $(X_1, X_2, ..., X_T)$.
La prévision à l'horizon $T+1$ notée $\hat{X}_{T+1}$ est définie par :

$$\hat{X}_{T+1} = \mathbb{E}(X_{T+1}/X_T, X_{T-1}, ..., X_1) = \hat{c} + \hat{\Phi}_1 X_T$$

qui est la conséquence des propriétés classiques de l'espérance conditionnelle dans les modèles de regression.
A l'horizon $T+2$ on a :

$$\hat{X}_{T+2} = \mathbb{E}(X_{T+2}/X_T, X_{T-1}, ..., X_1) = \hat{c} + \hat{\Phi}_1 X_{T+1} = (I + \hat{\Phi}_1)\hat{c} + \hat{\Phi}_1^2 X_T$$

A l'horizon $T+h$ on en déduit par récurrence que :

$$\hat{X}_{T+h} = \mathbb{E}(X_{T+h}/X_T, X_{T-1}, ..., X_1) = (I + \hat{\Phi}_1 + \hat{\Phi}_1^2 + ... + \hat{\Phi}_1^{h-1})\hat{c} + \hat{\Phi}_1^h X_T$$

Il est intéressant de calculer la limite de cette espérance quand h tend vers l'infini.
Si le polynôme $\Phi(L)$ est inversible, Φ^h converge vers zéro quand $h \to \infty$ et on a :

$$\lim_{h\to\infty} \mathbb{E}(X_{T+h}/X_T, X_{T-1}, ..., X_1) = \Phi(1)c = \mu$$

avec

$$\Phi(1) = \sum_{h=1}^{\infty} \Phi^{h-1}$$

Quand h tend vers l'infini, la prévision $\hat{X}_{T+h}$ tend vers la moyenne du processus.
Erreur de prévision

L'erreur de prévision est définie par :

$$X_{T+h} - \hat{X}_{T+h} = X_{T+h} - \mathbb{E}(X_{T+h}/X_T, X_{T-1}, ..., X_1)$$

En utilisant la forme VMA du processus c'est-à-dire

$$X_T = \mu + \sum_{i=0}^{\infty} \Phi_i \epsilon_{T-i}$$

on aura :

$$\begin{aligned} X_{T+h} - \hat{X}_{T+h} &= X_{T+h} - \mathbb{E}(X_{T+h}/X_T, X_{T-1}, ..., X_1) \\ &= X_{T+h} - \mathbb{E}(X_{T+h}/\epsilon_T, \epsilon_{T-1}, ..., \epsilon_1) \\ &= \sum_{i=0}^{\infty} \Phi_i \epsilon_{T+h-i} - \sum_{i=h}^{\infty} \Phi_i \epsilon_{T+h-i} \\ &= \sum_{i=0}^{h-1} \Phi_i \epsilon_{T+h-i} \end{aligned}$$

Par définition des bruits blancs, cette erreur de prévision a une espérance nulle. La matrice de variance covariance de cette erreur est donc :

$$\begin{aligned} &\mathbb{E}\left[(X_{T+h} - \hat{X}_{T+h})(X_{T+h} - \hat{X}_{T+h})'/X_T, X_{T-1}, ..., X_1\right] \\ &= \mathbb{E}\left[(\epsilon_{T+h} + \Phi_1\epsilon_{T+h-1} + ... + \Phi_1^{h-1}\epsilon_{T+1})(\epsilon_{T+h} + \Phi_1\epsilon_{T+h-1} + ... + \Phi_1^{h-1}\epsilon_{T+1})'\right] \\ &= \Sigma + \textstyle\sum_{i=1}^{h-1} \Phi_i \Sigma \Phi_i' \end{aligned}$$

1.7.2 Cas d'un VAR(p)

La variance de l'erreur de prévision s'obtient très facilement à partir de la représentation $VMA(\infty)$ d'un VAR d'ordre p quelconque. En effet, si X_t est un processus stationnaire, alors on peut l'écrire sous la forme :

$$X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} = \Psi(L)\epsilon_t$$

Dès lors, l'erreur de prévision s'écrit sous la forme :

$$\begin{aligned} X_{T+h} - \hat{X}_{T+h} &= X_{T+h} - \mathbb{E}(X_{T+h}/X_T, X_{T-1}, ..., X_1) \\ &= X_{T+h} - \mathbb{E}(X_{T+h}/\epsilon_T, \epsilon_{T-1}, ..., \epsilon_1) \\ &= \sum_{i=0}^{\infty} \psi_i \epsilon_{T+h-i} - \sum_{i=h}^{\infty} \psi_i \epsilon_{T+h-i} \\ &= \sum_{i=0}^{h-1} \psi_i \epsilon_{T+h-i} \end{aligned}$$

Par définition des bruits blancs, cette erreur de prévision a une espérance nulle. La matrice de variance covariance de cette erreur est donc :

$$\begin{aligned} &\mathbb{E}\left[(X_{T+h} - \hat{X}_{T+h})(X_{T+h} - \hat{X}_{T+h})'/X_T, X_{T-1}, ..., X_1\right] \\ &= \mathbb{E}\left[(\epsilon_{T+h} + \psi_1\epsilon_{T+h-1} + ... + \psi_1^{h-1}\epsilon_{T+1})(\epsilon_{T+h} + \psi_1\epsilon_{T+h-1} + ... + \psi_1^{h-1}\epsilon_{T+1})'\right] \\ &= \Sigma + \textstyle\sum_{i=1}^{h-1} \psi_i \Sigma \psi_i' \end{aligned}$$

1.8 Mise en oeuvre pratique avec R

Nous disposons des données macro-économiques du Canada. Les auteurs ont étudiés le marché du travail canadien. Ils ont utilisé la série suivante: Productivité de travail définie comme différence de notation entre le PIB et l'emploi ("prod"), la notation de l'emploi ("e"), le taux de chomage ("U") et salaires réels ("rw") et elle se trouve dans le package **vars**. Les données sont prise de la base de données et des envergures d'OCDE du premier trimestre 1980 jusqu'au quatrième trimestre 2004. (voir R Package vars, Bernhard Pfaff, Kronberg im Taunus)

```
>library(vars)
>data(Canada)
>summary(Canada)
```

```
 e               prod             rw               U
 Min.   :928.6   Min.   :401.3   Min.   :386.1   Min.   : 6.700
 1st Qu.:935.4   1st Qu.:404.8   1st Qu.:423.9   1st Qu.: 7.782
 Median :946.0   Median :406.5   Median :444.4   Median : 9.450
 Mean   :944.3   Mean   :407.8   Mean   :440.8   Mean   : 9.321
 3rd Qu.:950.0   3rd Qu.:410.7   3rd Qu.:461.1   3rd Qu.:10.607
```

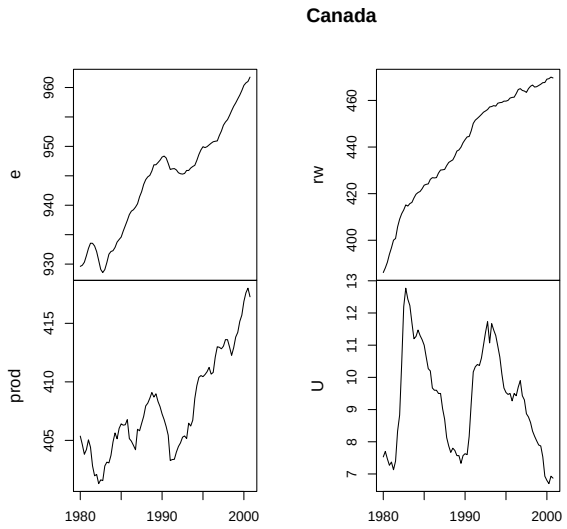


FIG. 1.2 – Série chronologique canadienne de marché du travail

```
 Max.   :961.8   Max.   :418.0   Max.   :470.0   Max.   :12.770
plot(Canada, nc = 2, xlab = "")
```

Dans une étape suivante, les auteurs ont déterminé une longueur optimale de retard pour une variété sans restriction.

```
> VARselect(Canada, lag.max = 8, type = "both")

$selection AIC(n)  HQ(n)  SC(n) FPE(n)
     3      2      1      3

$criteria
                  1            2            3            4            5
AIC(n) -6.272579064 -6.636669705 -6.771176872 -6.634609210
-6.398132246 HQ(n)  -5.978429449 -6.146420347 -6.084827770
-5.752160366 -5.319583658 SC(n)  -5.536558009 -5.409967947
-5.053794411 -4.426546046 -3.699388378 FPE(n)  0.001889842
0.001319462  0.001166019  0.001363175  0.001782055
                  6            7           8
AIC(n) -6.307704843 -6.070727259 -6.06159685 HQ(n)  -5.033056512
-4.599979185 -4.39474903 SC(n)  -3.118280272 -2.390621985
-1.89081087 FPE(n)  0.002044202  0.002768551  0.00306012
```

Selon l'AIC le nombre optimal de retard p=3, tandis que le critère de HQ indique p=2 et celui de SC indique une longueur optimale de retard de p=1.

```
> Canada <- Canada[, c("prod", "e", "U", "rw")]
> p1ct <- VAR(Canada, p = 1, type = "both")
> p1ct

VAR Estimation Results:
=======================

Estimated coefficients for equation prod:
=========================================
Call: prod = prod.l1 + e.l1 + U.l1 + rw.l1 + const + trend

    prod.l1        e.l1        U.l1       rw.l1       const       trend
 0.96313671  0.01291155  0.21108918 -0.03909399 16.24340747  0.04613085


Estimated coefficients for equation e:
======================================
Call: e = prod.l1 + e.l1 + U.l1 + rw.l1 + const + trend

      prod.l1          e.l1          U.l1         rw.l1         const
   0.19465028    1.23892283    0.62301475   -0.06776277 -278.76121138
        trend
  -0.04066045


Estimated coefficients for equation U:
======================================
Call: U = prod.l1 + e.l1 + U.l1 + rw.l1 + const + trend

     prod.l1         e.l1         U.l1        rw.l1        const        trend
 -0.12319201  -0.24844234   0.39158002   0.06580819 259.98200967   0.03451663


Estimated coefficients for equation rw:
=======================================
Call: rw = prod.l1 + e.l1 + U.l1 + rw.l1 + const + trend

     prod.l1         e.l1         U.l1        rw.l1        const        trend
 -0.22308744  -0.05104397  -0.36863956   0.94890946 163.02453066   0.07142229
plot(p1ct, names = "e")
```

Le graphique résultant est montré sur la figure 1.2. Ces tests de diagnostiques ont conduit au modèle VAR(1).

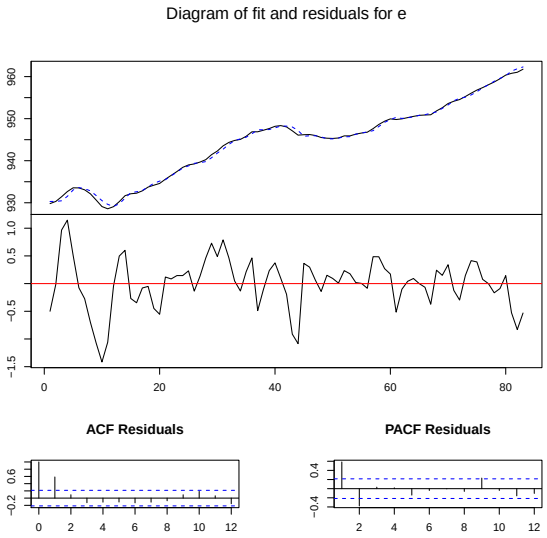


FIG. 1.3 – Corrélogramme résiduels pour le modèle VAR(1) adapté à la série chronologique canadienne

1.9 Conclusion

Dans ce chapitre nous avons exposé quelques concepts fondamentaux des séries chronologiques en général et des processus vectoriels autorégressif en particulier. Nous avons mis l'accent sur les mises en oeuvres moyennant le logiciel R aussi que des exemples pratiques illustratifs.

Chapitre 2

L'approche Bayésienne et le Facteur de Bayes

"La theorie, c'est quand on sait tout et que rien ne fonctionne.
La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi."
Albert EINSTEIN.

Dans ce chapitre sur la statistique Bayésienne, nous nous attachons à démontrer qu'il s'agit d'une approche cohérente et surtout pratique pour résoudre les problèmes d'inférence statistique. Notre objectif est de démontrer que cette approche est moderne, adaptée oux outils informatiques de simulation et apte à repondre aux problèmes de modélisation les plus avancés dans toutes les disciplines, plutôt que de l'ancrer sur ses querelles du passé. D'abord, nous présenterons les fondements de l'inference bayésienne. Puis nous insisterons sur le facteur de Bayes qui est le point clé de notre travail.

2.1 La décision de l'approche Bayésienne

Le troisième millénaire sera, dit-on, celui de l'information. Aussi, la statistique y sera-t-elle appelée à jouer un rôle important et le paradigme Bayésien plus que tout autre, puisqu'il offre un cadre de raisonnement bien adapté à l'intégration des opinions et des faits de toutes provenances qui interviennent dans la gestion des risques et la prise de décision en contexte d'incertitude.

De la collecte des données à la prévision, l'analyse statistique pose plusieurs défis. L'élaboration du modèle représente sans doute la phase la plus délicate de l'exercice, car elle doit répondre a un double impératif de réalisme et de parcimonie. Hormis quelques cas de figure, une démarche Bayésienne n'est envigeable qu'à charge de disposer d'outils efficaces pour la quantification et la mise à jour de l'information.

Nous présenterons donc dans cette partie un apercus des fondements de la théorie bayésienne.

2.1.1 Le choix bayésien

La statistique est un art interdisciplinaire de la quantification sous incertitudes utilisé par les physiciens, les économistes, les ingénieurs, les géographes, les biologistes, les assureurs, les psychologues,les météorologues etc. bref tous les praticiens soucieux de bâtir, sur des fondations solides, un pont entre théorie et données expérimentales.

Depuis un siècle, la statistique s'est considérablement développée, initiant une révolution dans les modes de pensée, car elle porte un langage de representation du monde et des ses incertitudes. C'est aujourd'hui une science mathématique dont l'objectif est de décrire ce qui s'est produit et de faire des projections quant à ce qu'il peut advenir dans le futur. Parfois, la situation peut être simplement décrite par quelques représentations graphiques d'analyse élémentaire des données. Bien souvent, le problème est beaucoup plus compliqué car des multiples facteurs d'influences doivent être pris en compte. Schématiquement, on construit deux ensembles avec ces facteurs, on ne sait, ou ne veut pas, représenter leur effet pertubateur au cas par cas et, plus grossière par ses caractéristiques statistiques générales.

Dans tous les cas, l'étude de la variabilité est au centre des débats: il s'agit d'abord de caractériser l'influence des facteurs identifiés et ensuite de représenter et d'évaluer le bruit résiduel dû à ces autres facteurs non pris en compte dans l'analyse de façon explicite. Dans une telle situation, le statisticien classique utilise à la fois un raisonnement déterministe par l'absurde, afin de proposer des valeurs acceptables pour les paramètres décrivant les effets des facteurs explicatifs et un raisonnement probabiliste, pour traduire la variabilté des résultats observés due au bruit. Ce mode de pensée s'appuie sur l'hypothèse de la réalité objective des paramètres ainsi que sur l'interpretation de la probabilité comme limite des fréquences des resultats observés. C'est cette conception, dite fréquentiste, qui est générale. Par contre, le statisticien Bayésien utilise le même cadre de pensée pour traiter par le pari probabiliste l'interaction de ces deux niveaux d'incertitudes: ignorance quant aux valeurs possibles des paramètres et aléa des bruits entachant les résultats expérimentaux.

Choisir la piste Bayésienne paraîtra à certains inutilement trop sophistiqué si on se limite aux modèles élémentaires (binomial; normal, etc.), pour ces cas d'écoles simples, l'approche fréquentiste est facile (nombreux logiciels), et offre au praticien des résultats souvent trés proches de ceux que donnerait une analyse Bayésienne avec une distribution a priori peu informative. Mais pour peu que l'analyste souhaite prendre à bras le corps des problèmes plus proche de son réel quotidien, apparaissent variables multiples, données manquantes, effets aléatoires, grandeurs latentes..., la structure des modèles de la vie scientifique moderne se présente sous forme *où* des couches successives de conditionnement s'êmboitent... et pour lesquels l'approche Bayésienne affirme sa véritable pertinence.

2.1.2 Notions de base

Un système est caractérisé par un certain nombre de variables définissant son état. Les valeurs de ces variables, les mesurandes, sont obtenues par le biais d'un système de mesure. Les résultats de la mesure, que nous appelerons aussi observations, ne permettent qu'une

estimation de l'état, car elles interviennent dans le processus des phénomènes de nature aléatoire non maîtrisés par l'observateur. On obtient une correspondance entre l'état et l'observation qui est de nature statistique, associant à un état fixé une répartition des observations (gaussienne, poissonnienne ou autre). Le but de la mesure est alors d'inverser cette relation, en ce sens que l'observateur ayant obtenu un résultat, doit en inférer l'état.

La relation donnant la repartition des observations pour un état donné se nomme *le modèle*. Elle est objective, extérieure à l'observateur, car elle n'est dépendante que de l'objet mesuré et de l'instrument utilisé. En répétant l'experience en face du même état, l'observateur verra ses résultats se distribuent selon le modèle. On trouve une probabilité au sens fréquentiel.

Comme un état peut donner des résultats de mesure differents, une observation peut trés bien correspondre à plusieurs états. Quel est le bon? Ou mieux, quelle répartition peut-on imaginer sur les états ayant en main cette observation? L'incertitude passe du domaine des résultats dans celui des états. Cette incertitude est subjective, en ce sens qu'elle est propre à l'observateur.

En statistique classique, le modèle est donné par une fonction dite de vraisemblance $\mathcal{L}(\theta|x)$, *où* θ désigne l'état du système et x l'observation. Cette fonction est normalisée et considérée comme densité de probabilité.

2.1.3 Théorème de Bayes

Soient A et B deux événements aléatoires. La probabilité de B, conditionnellement à la réalisation de A, est par définition exprimée (si la probabilité de l'événement A est non-nulle) par la relation suivante:

$$P[B|A] = \frac{P[B,A]}{P[A]}$$

où $P[B,A]$ est la probabilité que les deux événements A et B aient lieu simultanément.

Thomas Bayes, pieux serviteur de Dieu[1] [1] et grand mathématicien, n'est pas passé à l'histoire pour la découverte de la formule qui porte son nom, mais pour son interprétation et son application à un problème d'estimation.

Le problème du Rev.Bayes (Bernier et al.,2000) est le suivant: une balle est lancée sur une table parfaitement horizontale et s'arrête à un certain point P_0. Ensuite, une deuxième balle est lancée n fois et on s'intéresse au nombre de fois x, qu'elle s'est arrêtée à la droite de la première (appelons ces événements "*succés*").Comment estimer la probabilté de succés θ?

Bayes imagine que θ est une variable aléatoire définie sur l'intervalle [0,1] et décrit avec

1. Il est auteur en 1731 de l'essai:"Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Governement is Happiness of His Créatures"

une distribution de probabilité donnée π_θ sa connaissance concernant cette variable, préablement à l'observation des données.

En utilisant la formule de Bayes, il est possible d'écrire l'expression de la distribution de probabilté de θ, conditionnellement à l'observation des données x:

$$\pi(\theta|x) = \frac{\pi(\theta)P(x|\theta)}{M(x)} = \frac{P(x|\theta)\pi(\theta)}{\int_\theta P(x|\theta)\pi(\theta)}$$

Les différents termes de cette formule peuvent être interprétés de la manière suivante:

$\pi(\theta)$: est la distribution de probabilité a priori du paramètre inconu θ. L'appelation *a priori* exprime le fait qu'elle a été établie préalablement à l'observation des données x. Elle peut être issue de l'opinion personnelle de statisticien, ou établie sur la base de l'analyse d'autres données similaires ou de l'avis d'expert. Des exemples de mise en oeuvre de ces méthodes sont fournis par Kadane et Wolfson (1998), O'Hagan (1998), Garthwaite et O'Hagan (2000) et Parent et Prevost (2003). En revanche, son indépendance des données est obligatoire pour éviter d'utiliser deux fois la même information (Berry, 1996).

$P(x|\theta)$: est la probabilté des observations conditionnnellement à la valeur θ du paramètre de modèle statistique qu'on utilise pour leur description. Il s'agit de la vraisemblance des données, sous le modèle paramétré par θ.

$\pi(\theta|x)$: est la distribution de probabilité a posteriori du paramètre du modèle, sur la base de la connaissance a priori et de l'information apportée par les données. L'appellation *a posteriori* vient du fait que, logiquement, elle suit l'observation des données.

Le dénominateur, indépendant de θ, est uniquement une constante de normalisation.

Le passage de la distribution a priori à la distribution a posteriori des paramètres du modèle statistique, exprimé par la formule Bayes, peut être alors interprété comme une mise à jour de la connaissance, sur la base des observations (figure 2.1).

Cette lecture de la formule de Bayes est à la base de l'origine de la distinction entre les statisticien dits fréquentistes et les Bayésiens.

La différence fondamentale entre les deux approches est que, pour les fréquentistes, toute technique statistique (inférence, test, choix de modèle...) doit être fondée uniquement sur les données et aucune information externe à l'échantillon observé ne peut être introduite dans les calculs.

Exemple 2.1. *(Jean-Michel Marin, Christian P.Robert)*
Dans le cadre d'une observation normale de moyenne inconnue θ, $x \sim \mathcal{N}_1(\theta,1)$ [2], *la loi a posteriori associée à la loi a priori* $\theta \sim \mathcal{N}_1(0,10)$ *est*

$$\pi(\theta|x) \propto \exp -\frac{1}{2}\{10^2 + (\theta - x^2)\}$$

2. Le vecteur aléatoire X à valeurs dans $\mathbb{R}^p$ distribué suivant une loi normale d'espérance μ et de

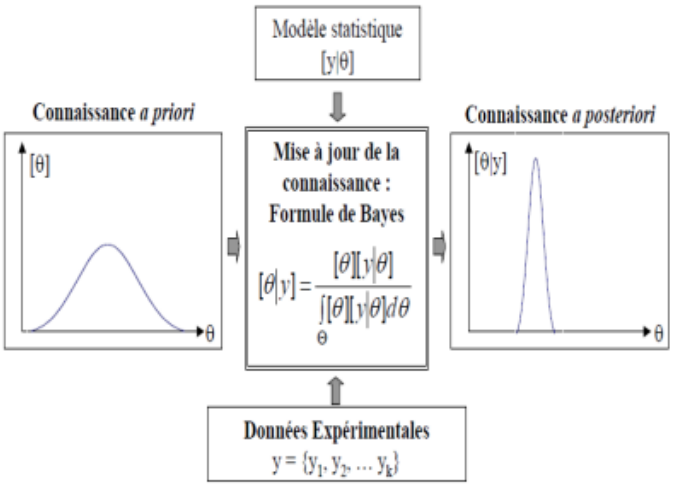


FIG. 2.1 – Mise à jour de la connaissance de la formule de Bayes

ce qui équivaut à la loi $\theta|x \sim \mathcal{N}(\frac{10x}{11},\frac{10}{11})$. *L'espérance a posteriori de* θ *est donc* $\frac{10x}{11}$.

Exemple 2.2. *Revenons au problème du Rev. Bayes. Si* x *est le nombre de succés observés et* θ *la probabilité de succés, le modèle naturel pour décrire le phénomène est le modèle binomial. C'est à dire que, conditionnellement à* θ *la probabilité d'observer* x *succés s'écrit:*

$$P(x|\theta) = C_n^\theta \theta^x (1-\theta)^{n-x}$$

Concernant la distribution a priori de θ, *un bon choix est une loi de la famille Bêta. Ces lois, à deux paramètres* α *et* β, *sont définies dans l'intervalle [0,1] et, selon les valeurs de* α *et* β *peuvent avoir des formes trés différentes. Cette souplesse rend la famille Bêta trés adaptée à la formalisation de la connaissance préliminaire sur une variable bornée. En vertu de ce choix, l'expression de la loi a priori de* θ *est:*

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

D'aprés la formule de Bayes la loi a posteriori $\pi(\theta|x)$, *est proportionnelle au produit entre loi a priori et la vraisemblance:*

$$\pi(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

est alors encore une loi Bêta de paramètres: $x+\alpha$ *et* $n-x+\beta$

---

structure de covariance Σ,$\mathcal{N}_p(\mu,\Sigma)$, admet comme densité de probabilté:

$$f_X(X|\mu,\Sigma) \propto \exp[-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)],$$

où A^T désigne la transposée de la matrice A.

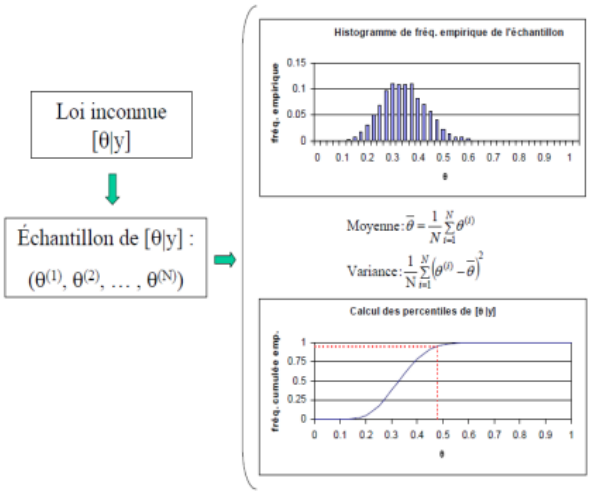


FIG. 2.2 – Approximation de la loi a posteriori avec des échantillons aléatoires

Dans ce cas l'expression du numérateur de la formule de Bayes nous a permis directement de reconnaître que la loi a priori et la loi a posteriori appartiennent à la même famille (Bêta). On exprime cette circonstance heureuse en disant que les lois Bêta et Binomiale sont conjugués.

Malheureusement, sauf quelques autres cas d'école, dans la pratique courante de la modélisation, la propriété de conjugaison n'est pas vérifié et alors le calcul du dénominateur de la formule de Bayes s'impose.

Or, le cacul de cette intégrale, souvent multidimensionnelle, est infaisable dans la plupart des cas. C'est pour cette raison que l'approche Bayésienne a été longtemps mise à l'écart, à l'avantage de l'approche classique qui offre, elle, des solutions simples à de nombreuxs problèmes statistiques.

La découverte et la possibilité de mettre en oeuvre des algorithmes de simulation capable d'obtenir des tirages aléatoires dans la loi a posteriori des paramètres a libéré les Bayésiens des fardeaux du calcul intégral et a permis l'estimation de modèles à structures complexes. L'expression mathématique de la loi a posteriori restera inconnue à jamais mais, avec des échantillons aléatoires de cette loi de taille significative, on peut en calculer empiriquement la moyenne, la variance, les percentiles et toutes les autres grandeurs statistiques qui la décrivent, ce qui est d'ailleurs plus intéressant en pratique que d'avoir la formule mathématique de la loi jointe des paramètres du modèle (figure 2.2)

2.2 Une introduction à la théorie de la décision

2.2.1 Fonctions de coût usuelles

Perte quadratique

Introduit par Légende (1805) et Gauss (1810), ce coût est sans contexte le critère d'évaluation le plus commun. Fondant sa validité sur l'ambiguïté de la notion d'érreur dans un contexte statitique (soit erreur de mesure, soit variation aléatoire), il a aussi donné lieu à des nombreuses critiques, la plus fréquente étant sans doute le fait que le coût quadratique

$$\mathrm{L}(\theta,\delta) = (\theta - \delta)^2$$

pénalise trop fortement les grandes erreurs.
Le coût quadratique est particulièrement intéressant lorsque l'espace des paramètres est borné et le choix d'un coût plus subjectif.

L'erreur de coût absolu

Une solution alternative au coût quadratique en dimension une est d'utiliser le coût absolu,

$$L(\theta,d) = |\theta - d|$$

déja considéré par Laplace (1773) ou, plus généralement, une fonction linéaire par morceaux

$$L_{k_1,k_2}(\theta,d) = \begin{cases} k_2(\theta - d) & \text{si } \theta > d, \\ k_1(d - \theta) & \text{sinon.} \end{cases}$$

Le coût 0-1

Ce coût est surtout utilisé dans l'approche classique des tests d'hypothèse, proposée par Neyman et Pearson. Plus généralement, c'est un exemple typique d'un coût non quantitatif. En effet, pour ce coût, la pénalité associée à un estimateur δ est 0 si la réponse est correcte et 1 sinon.

Exemple 2.3. *Soit le test de $H_0 : \theta \in \Theta_0$ contre H_1 : sinon. Alors D=$\{0,1\}$, où 1 représente l'acceptation de H_1 et 0 son rejet, pour la fonction de coùt 0-1, qui vaut*

$$L(\theta,d) = \begin{cases} 1 - d & \textit{si } \theta \in \Theta_0, \\ d & \textit{sinon.} \end{cases}$$

Le risque associé est

$$R(\theta,\delta) = \mathbb{E}_\theta[L(\theta,\delta(x))] = \begin{cases} P_\theta(\delta(x) = 0) & \textit{si } \theta \in \Theta_0, \\ P_\Theta(\delta(x) = 1) & \textit{sinon.} \end{cases}$$

ce qui donne les erreurs de première et deuxième espèce qui sous-tendent la Théorie de Neyman-Pearson.

2.2.2 Fonction perte et risque

Pour le modèle $X \in \{\chi,\beta,\{P_\theta,\theta \in \Theta\}\}$, on définit D l'ensemble des décisions possibles. c'est-à-dire l'ensemble des fonctions de Θ dans $g(\Theta)$ o*ù* g dépend du contexte:

- si le but est d'estimer θ alors D = Θ
- pour un test, D=$\{0,1\}$

La fonction perte est une fonction mesurable de $(\Theta \times D)$ à valeurs positives: L: $\Theta \times D \rightarrow \mathbb{R}_+$. Elle est définie selon le problème étudié et constitue l'armature du problème statistique.

Définition 2.2.1. *Risque fréquentiste*
Pour (θ,δ), le risque fréquentiste est défini par

$$R(\theta,\delta) = E_\theta[L(\theta,\delta(x))] = \int_X L(\theta,\delta(x))f(x|\theta)dx$$

C'est une fonction de θ et ne définit donc pas un ordre total sur D et ne permet donc pas de comparer toutes décisions et estimateurs. Il n'existe donc pas de meilleur estimateur dans un sens absolu. Ainsi, l'approche fréquentiste restreint l'espace d'estimation en préférant la classe des estimateurs sans biais dans laquelle il existe des estimateurs de risque uniformément minimal; l'école Bayésienne ne perd pas en généralité en définissant un risque a posteriori. L'idée est d'intégrer sur l'espace des paramètres pour pallier cette difficulté.

Définition 2.2.2. *Risque a posteriori*
Une fois donnée la loi a priori et la fonction perte, le risque a posteriori est défini par:

$$\rho(\pi,\delta|x) = E^\pi[L(\theta,\delta)|x] = \int_\Theta R(\theta,\delta)d\pi(\theta)$$

2.2.3 Estimateur de Bayes

Soit une fonction de coût $L(\theta,\delta)$, et une loi de probabilté a priori (ou une loi impropre) π, pour trouver l'estimateur de Bayes $\delta^\pi(x)$, on applique la règle suivante:

$$\delta^\pi(x) = \min_\delta \mathbb{E}^\pi[L(\theta,\delta)/x]$$

L'estimateur $\delta^\pi(x)$ sera déterminé analytiquement ou numériquement ceci dépendra de la fonction perte, sa nature et complexité.

Généralement, les solutions associées à des coûts classiques sont formellement connues et correspondents aux caractéristiques usuelles d'une distribution (moyenne, mediane, fractiles etc...)

Définition 2.2.3. *Estimateur Bayésien*
Un estimateur Bayésien est un estimateur vérifiant:

$$r(\pi,\delta^\pi) = \inf_{\delta \in D} r(\pi,\delta) < \infty$$

La valeur $r(\pi,\delta^\pi)$ est alors appelée risques de Bayes.

Proposition 2.2.1. *(Christian P.Robert 2006)*
L'estimateur de Bayes δ^π associé à la loi a priori π et le coût quadratique est la moyenne a posteriori

$$\delta^\pi(x) = \mathbb{E}^\pi[\theta|x] = \frac{\int_\theta \theta f(x|\theta)\pi(\theta)d\theta}{\int_\theta f(x|\theta)\pi(\theta)d\theta}$$

Corollaire 1. *(Christian P.Robert 2006)*
Quand $\Theta \in \mathbb{R}^p$, l'estimateur de Bayes δ^π associé à π et au coût quadratique,

$$L(\theta,\delta) = (\theta - \delta)^t Q(\theta - \delta)$$

est la moyenne a posteriori, $\delta^\pi = \mathbb{E}^\pi[\theta|x]$, pour toute matrice Q $(p \times p)$ symétrique définie positive

Le tableau ci-dessous représente quelques estimateurs de Bayes du paramètre θ sous coût quadratique pour les lois a priori conjuguées des familles exponentielles usuelles.

Loi de x	Loi conjuguée	Moyenne a posteriori
Normale $N(\theta,\sigma^2)$	$N(\mu,\tau^2)$	$\frac{\mu\sigma^2+\tau^2x}{\sigma^2}$
Poisson $P(\theta)$	$\Gamma(\alpha,\beta)$	$\frac{\alpha+x}{\beta+1}$
Gamma $\Gamma(\nu,\theta)$	$\Gamma(\alpha,\beta)$	$\frac{\alpha+\nu}{\beta+x}$
Binomiale $B(n,\theta)$	$Be(\alpha,\beta)$	$\frac{\alpha+x}{\alpha+\beta+n}$
Binomiale Négative $Neg(m,\theta)$	$Be(\alpha,\beta)$	$\frac{\alpha+n}{\alpha+\beta+x+n}$
Multinomiale $M_k(\theta_1,...,\theta_k)$	$D(\alpha_1,...,\alpha_k)$	$\frac{\alpha_i+x_i}{(\sum_j \alpha_j)+n}$
Normale $N(\mu,1/\theta)$	$\Gamma(\alpha,\beta)$	$(\frac{\alpha+1}{\beta+(\mu-x)})^2$

Tab1.1-Quelques estimateurs de Bayes usuels

Proposition 2.2.2. *(Christian P. Robert 2006)*
L'estimateur de Bayes associé à la loi a priori π et à la fonction de coût linéaire par morceaux est le fractile $(k_1/(k_1+k_2))$ de $\pi(\theta|x)$.
En particulier, si $k_1 = k_2$, dans le coût absolu, l'estimateur de Bayes est la médiane a posteriori, qui est l'estimateur obtenu par Laplace

Proposition 2.2.3. *(Christian P. Robert 2006)*
L'estimateur de Bayes associé à π et au coût 0-1 est

$$\delta^\pi(x) = \begin{cases} 1 & si\ P(\theta \in \Theta_0|x) > P(\theta \nsubseteq \Theta_0|x) \\ 0 & sinon \end{cases}$$

donc $\delta^\pi(x)$ vaut 1 si et seulement si $P(\theta \in \Theta_0|x) > 1/2$.

2.2.4 Admissibilité et minimaxité

Définition 2.2.4. *Estimateur randomisé*
Pour le modèle $X \in \{\chi,\beta,\{P_\theta,\theta \in \Theta\}\}$, un ensemble de décisions D, on définit D^* comme l'ensemble des probabilités sur D. $\delta^* \in D^*$ est appelé estimateur randomisé.
L'idée à l'origine de cette notion est de rendre D convexe pour pouvoir maximiser facilement.

Théorème 2.2.1. *(Christian P. Robert 2006)*
Pour toute distribution a priori π sur Θ, le risque de Bayes pour l'ensemble des estimateurs randomisés est le même que celui pour l'ensemble des estimateurs non randomisés, soit

$$\inf_{\delta \in D} r(\pi,\delta) = \inf_{\delta^* \in D^*} r(\pi,\delta^*) = r(\pi)$$

Minimaxité

Le critère de minimaxité apparaît comme une assurance contre le pire, car il vise à minimiser le coût moyen dans le cas le moins favorable. Il représente aussi un éffort fréquentiste pour éviter de recourir au paradigme Bayésien, tout en engendrant un ordre (faible) sur D^*

Définition 2.2.5. *On appelle risque minimax*

$$\overline{R} = \inf_{\delta \in D^*} \sup_{\theta} \mathbb{E}_\theta[L(\theta,\delta(x))]$$

et estimateur minimax tout estimateur δ_0 tel que

$$\overline{R} = \sup_{\theta} R(\theta,\delta_0)$$

L'estimateur minimax correspond au point de vue de faire le mieux dans le pire des cas, c'est-à-dire à s'assurer contre la pire. Il est utile dans des cadres complexes mais trop conservateur dans certains cas où le pire est trés probable. Il peut être judicieux de voir l'estimation comme un jeu entre le statisticien (choix de δ) et la Nature (choix de θ), l'estimateur minimax rejoint alors celle de la Théorie des Jeux.

Règle minimax et Stratégie maximin

Une difficulté importante liée à la notion de minimaxité est que les estimateurs minimax n'existent pas nécessairement. En particulier, il existe une stratégie minimax quand Θ est fini et la fonction de coût est continue. Plus généralement , Brown (1976) (voir aussi Le Cam, 1986, et Strasser, 1985) considère l'espace de décision D comme plongé dans un autre espace de manière telle que l'ensemble des fonctions de risque sur D est compact dans ce grand espace. Dans cette perspective et sous des hypothèses supplémentaires,

il est alors possible de construire des estimateurs minimax lorsque la fonction coût est continue.

Théorème 2.2.2. *(Christian P. Robert 2006)*
Si $D \subset \mathbb{R}^k$ est convexe et compact et si $L(\theta,d)$ est continue et convexe en tant que fonction de d, pour chaque $\theta \in \Theta$, alors, il existe un estimateur minimax non randomisé.

Lemme 2.2.1. *(Rousseau 2009)*
Le risque de Bayes est toujours plus petit que le risque minimax,

$$\underline{R} = \sup_{\pi} r(\pi) = \sup_{\pi} \inf_{\delta \in D} r(\pi,\delta) \leq \overline{R} = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta)$$

La première valeur est dite *risque maximin* et une distribution π^* telle que $r(\pi^*) = \underline{R}$ est appelée *distribution a priori la moins favorable*, quand de telles distributions existent. En général, la borne supérieure $r(\pi^*)$ est atteinte plutôt par une distribution impropre pouvant s'exprimer comme une limite de distribution a priori propre π_n. Mais ce phénomène n'empêche pas nécessairement la construction d'estimateurs minimax. Quand elles existent, les distributions les moins favorables sont celles qui ont le risque de Bayes le plus grand, donc aussi les moins intéressantes en terme de coût lorsqu'elles ne sont pas suggérées par l'information a priori disponible. Le résultat ci-dessus est assez logique au sens *où* l'information a priori ne peut qu'améloirer l'erreur d'estimation, même dans le pire des cas.

Définition 2.2.6. *Un problème d'estimation est dit admettre une valeur si $\underline{R} = \overline{R}$, c'est-à-dire quand*

$$\sup_{\pi} \inf_{\delta \in D} r(\pi,\delta) = \inf_{\delta \in D^*} \sup_{\theta} R(\theta,\delta)$$

Quand le problème admet une valeur, certains estimateurs minimax sont des estimateurs de Bayes correspondant aux lois a priori les moins favorables. Cependant, ils peuvent être randomisés. Par conséquent le principe minimax ne fournit pas toujours des estimateurs acceptables.

Lemme 2.2.2. *(Christian P. Robert 2006)*
Si δ_0 est un estimateur de Bayes pour π_0 et si $R(\theta,\delta_0) \leq r(\pi_0)$ pour tout θ dans le support de π_0, δ_0 est minimax et π_0 est la distribution la moins favorable.

Admissibilité

Définition 2.2.7. *Estimateur admissible Soit $X \in \{\chi,\beta,\{P_0,\theta \in \Theta\}\}$ un modèle paramétrique et L une fonction de perte sur $\Theta \times D$ où D est l'ensemble des décisions. On dit que $\delta \in \Theta$ est inadmissible si et seulement si $\exists \delta_0 \in D$, $\forall \theta \in \Theta$, $R(\theta,\delta) \geq R(\theta,\delta_0)$ et $\exists \theta_0 \in \Theta$, $R(\theta_0,\delta) \geq R(\theta_0,\delta_0)$. Dans le cas contraire, δ est admissible.*

Proposition 2.2.4. *(Christian P. Robert 2006)*
S'il existe un unique estimateur minimax, cet estimateur est admissible

Notons que la reciproque de ce résultat est fausse, car il peut exister plusieurs estimateurs minimax admissibles. Par exemple, dans le cas $N_p(\theta,I_p)$, il existe des estimateurs de Bayes réguliers minimax pour $p \geq 5$. Quand la fonction de coût L est absolument convexe (en d), la caractérisation suivante est aussi possible.

Proposition 2.2.5. *(Rousseau 2009)*
Si δ_0 est admissible de risque constant, δ_0 est l'unique estimateur minimax.

Théorème 2.2.3. *Estimateurs Bayésiens admissibles (Rousseau 2009)*
Si l'estimateur Bayésien δ^π associé à une fonction perte L et une loi a priori π est unique, alors il est admissible

Proposition 2.2.6. *(Christian P. Robert 2006)*
Si un estimateur de Bayes, δ^π, associé à une loi a priori (propre ou impropre) π, est tel que le risque de Bayes,

$$r(\pi) = \int_\Theta R(\theta,\delta^\pi \pi(\theta) d\theta$$

soit fini, δ^π est admissible

Définition 2.2.8. *π-admissibilité*
Un estimateur δ_0 est π-admissible si et seulement si

$$\forall(\delta,\theta), R(\theta,\delta) \leq R(\theta,\delta_0) : \pi(\{\theta \in \Theta, R(\theta,\delta) < R(\theta,\delta_0\}) = 0$$

Proposition 2.2.7. *(Christian P. Robert 2006)*
Tout estimateur Bayésien tel que $r(\pi) < \infty$ est π-admissible

Théorème 2.2.4. *Continuité et π-admissibilité, (Rousseau 2009)*
Si $\pi > 0$ sur Θ, $r(\pi) < \infty$ pour une fonction perte L donnée, si δ^π estimateur Bayésien correspondant existe et si $\theta \mapsto R(\theta,\delta)$ est continu, alors δ^π est admissible

2.3 Choix des lois a priori

Le point le plus signifiant dans l'analyse Bayésienne est le choix de la loi a priori, sa détermination est donc l'étape la plus importante dans la mise en oeuvre de cette inférence.
En statistique Bayésienne, on considère, en plus des données récoltées dans le cadre d'une expérience, un a priori sur le paramètre θ que l'on cherche à estimer. C'est le terme $\pi(\theta)$ Cela peut permettre d'inclure, dans les résultats précédents, formels ou non. En pratique,

toute la difficulté consiste à estimer de manière correcte nos a priori. Avec beaucoup d'observations, le comportement asymptotique peut guider ce choix mais sinon il est nécessaire de le justifier avec précision.

2.3.1 Approche partiellement informative

Quand on dispose de peu d'information a priori, ou quand l'information dont on dispose est trop vague, alors souvent le statisticien ne peut faire une construction subjective complète de l'a priori.
De telles situations peuvent obliger le statisticien à avoir recours à des méthodes d'estimation fréquentiste comme: Estimateur du maximum de vraisemblance, estimateurs sans biais optimaux, etc.

Cependant, tout en gardant à l'esprit les fondements bayésiens des critères fréquentistes d'optimalité, il parait donc préferable de suivre l'approche bayésienne, en utilisant un a priori dit objectif, c'est-à-dire construit à partir du modèle d'échantillonnage.
Lorsque aucune information a priori n'est disponible, ces a priori sont dits *non informatifs*.

Entropie maximale

Si l'on possède des informations partielles du type $\mathbb{E}^{\pi}[\eth\neg(\theta)] = \mu_k$ o*ù* pour chaque k = 1,...,n, g_k est une fonction donnée.
Pour $\theta \in \{1,...,n\}$ et $\pi(\theta) = (\pi_1,...,\pi_n)$ tel que $\pi_i > 0$ et $\sum_{i=1}^{n} \_i = 1$, l'entropie de la loi est définie par

$$Ent(\pi) = -\sum_{i=1}^{n} \pi_i \log(\pi_i) \leq -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = \log n$$

Ce dernier terme correspond à une répartition uniforme. Pour la masse de Dirac $\delta(j)$ (telle que $\pi_j = 1$ et $\forall i \neq j$, $\pi_i = 0$), $Ent(\delta(j)) = 0$ ce qui correspond à l'intuition puisqu'alors il n'y a plus d'incertitude et l'information est totale. Une entropie petite s'interprète comme une loi concentrée et informative. La maximisation de l'entropie sous les contraintes permet de chercher la loi qui apporte le moins d'information. Le principe à la base de cette méthode est donc de chercher à calculer:

$$\arg\max_{\pi} Ent(\pi) \textit{sous la contrainte } \mathbb{E}^{\pi}[\eth\neg(\theta)] = \mu_k$$

La solution de ce problème est alors donnée par:

$$\pi^* \propto e^{\sum_{k=1}^{n} \lambda_k g_k(\theta)}$$

o*ù* les λ_k sont les multiplicateurs de Lagrange associés. Dans la pratique, on determine ces valeurs λ à partir des contraintes (systèmes d'équations) comme l'indique l'exemple à suivre.

Exemple 2.4. *Un cas dénombrable*
Ici, $\Theta = \mathbb{N}$ *et* $\mathbb{E}^{\pi}[\theta] = x > 1$, *c'est-à-dire qu'ici* $g(\theta) = \theta$ *et* $\mu = x$. *On sait que* $\pi^* \propto e^{\lambda\theta}$ *et* λ *est déterminé par:*

$$\frac{\sum_{\theta\in\mathbb{N}} \theta e^{\lambda\theta}}{\sum_{\theta\in\mathbb{N}} e^{\lambda\theta}} = x$$

Cela conduit à résoudre:

$$\frac{x}{1-e^{\lambda}} = \frac{1}{e^{\lambda}} \frac{e^{\lambda}}{(1-e^{\lambda})^2}$$

$$e^{\lambda} = \frac{x-1}{x}$$

Par exemple si $x = \frac{12}{11}$ *alors* $\lambda = -\log(12)$. *En continu, il n'est pas possible de définir l'entropie comme ci-dessus puisqu'on ne peut dénombrer les états (pas de mesure de comptage) en l'absence de mesure de référence. Dans le cas continu, on définit alors l'équivalent de l'entropie par rapport à une mesure* π_0*:*

$$Ent(\pi|\pi_0) = \int_{\Theta} \pi(\theta) \log(\frac{\pi(\theta)}{\pi(\theta_0)}) d\theta$$

C'est en fait la divergence de Kullback. Dans l'idée π_0 *est la plus proche de la repartitition uniforme. L'objectif est donc de maximiser* $Ent(\pi|\pi_0)$ *sous les contraintes* $\mathbb{E}^{\pi}[g_k(\theta)] = \mu_k$. *Là encore, la solution générale est connue:*

$$\pi^*(\theta) \propto e^{\sum_{k=1}^{n} \lambda_k g_k(\theta)} \pi_0(\theta)$$

Lois a priori conjuguées

Ce type de lois a priori est utilisé quand l'information a priori disponible sur le modèle est trop vague ou peu faible. Dans ce cas l'analyste regarde la forme de la fonction de vraisemblance et choisit une famille de lois qui se marie bien avec elle.

L'avantage des familles conjuguées est avant tout la simplicité des calculs. Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs. L'intérêt principal du caractère conjugué se manifeste quand $\mathcal{F}$ est paramétrée.

Effectivement le passage de distribution a priori à la distribution a posteriori n'est dans ce cas qu'une mise à jour des paramètres correspondants. Et par conséquence, les distributions a posteriori sont toujours calculables dans ce cas.

L'approche a priori conjuguée, introduite par Raiffa et Schlaifer (1961), peut être considérée comme un point de départ pour l'élaboration de distributions a priori fondées sur une information a priori limitée. On considère une variable x suivant une fonction de densité

$f(x|\theta)$

Définition 2.3.1. *Famille conjuguée*
Une famille $\mathcal{F}$ *de distributions de probabilité sur* Θ *est dite conjuguée (ou fermée par échantillonnage) par une fonction de vraisemblance* $f(x|\theta)$ *si, pour tout* $\pi \in \mathcal{F}$, *la distribution a posteriori* $\pi(.|x)$ *appartient également à* $\mathcal{F}$.

Un exemple trivial d'une famille conjuguée est l'ensemble $\mathcal{F}_0$ *de toutes les lois de probabilité sur* Θ. *L'avantage des familles conjuguées est avant tout de simplifier les calculs.*

Avant l'essor du calcul numérique, ces familles étaient pratiquement les seules qui permettaient de faire aboutir des calculs.

les lois a priori conjuguées sont généralement associées à un type particullier de lois d'échantillonnage qui permet toujours leur obtention; il est même caractéristique des lois a priori conjuguées comme nous le verrons ci-dessous. Ces lois constituent ce qu'on appelle des familles exponentielles

Définition 2.3.2. *Familles exponentielles*
Soient μ une mesure σ-finie sur χ, Θ l'espace des paramètres, C et h des fonctions respectivement de χ et Θ dans $\mathbb{R}_+$, et $\mathcal{R}$ et T des fonctions de Θ et χ dans $\mathbb{R}^k$. La famille des distributions de densité (par rapport à μ)

$$f(x|\theta) = C(\theta)h(x)\exp\{\mathcal{R}(\theta).T(x)\}$$

est dite famille exponentielle de dimension k. Dans le cas particulier où $\Theta \subset \mathbb{R}^k$ et

$$f(x|\theta) = C(\theta)h(x)\exp\{\theta.x\}$$

la famille est dite naturelle.

D'un point de vue analytique, les familles exponentielles ont certaines caractéristiques intéressantes. Il existe une statistique exhaustive de dimension constante, en effet, si $x_1,...,x_n \sim f(x|\theta)$, avec f satisfaisant

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \in \mathbb{R}^k$$

est exhaustive pour tout n. La réciproque de ce résultat a été aussi établie par Koopman (1936) et Pitman (1936).

Théorème 2.3.1. *(Lemme de Pitman-Koopman)*
Si une famille de lois f(.|x) à support constant est telle que, à partir d'une taille d'échantillon suffisament grande, il existe une statistique exhaustive de taille fixe, la famille est exponentielle.

Exemple 2.5. *Soit $x \sim \mathcal{N}_p(\theta,\sigma^2 I_p)$ alors,*

$$f(x|\theta) = \frac{1}{\sigma^p}\frac{1}{(2p)^{p/2}}\exp\{-\sum_{i=1}^{p}(x_i-\theta_i)^2/2\sigma^2\}$$

$$= C(\theta,\sigma)h(x)\exp\{x.(\theta/\sigma^2) + \|x\|^2(-1/2\sigma^2)\}$$

et la distribution normale appartient à une famille exponentielle de paramètres naturels θ/σ^2 et $-1/2\sigma^2$. De la même façon, si $x_1,...,x_n \sim \mathcal{N}_p(\theta,\sigma^2 I_p)$, la distribution jointe satisfait

$$f(x_1,...,x_n) = C^{'}(\theta,\sigma)h^{'}(x_1,...,x_n) \times \exp\{n\overline{x}.(\theta/\sigma^2) + \sum_{i=1}^{n}\|x_i-\overline{x}\|^2(-1/2\sigma^2)\}$$

et la statistique $\overline{x}$, $\sum_{i=1}\|x_i-\overline{x}\|^2$ est exhaustive pour tout $n \geq 2$

Définition 2.3.3. *Soit $f(x|\theta) = C(\theta)h(x)\exp\{\theta.x\}$, une famille exponentielle naturelle. L'espace naturel des paramètres est*

$$N = \{\theta; \int_{\chi} e^{\theta.x} h(x) d\mu(x) < +\infty\}$$

La famille est dite régulière si N est un ensemble ouvert et minimale si $dim(N) = dim(K) = k$, où K est la clôture de l'enveloppe convexe du support de μ.

Remarque 2.3.1. *Les familles exponentielles naturelles peuvent aussi être réécrites sous la forme*

$$f(x|\theta) = h(x)e^{\theta.x-\psi(\theta)}$$

et $\psi(\theta)$ est dite fonction cumulante des moments.

Lois conjuguées des familles exponentielles

Soit

$$f(x|\theta) = h(x)e^{\theta.x-\psi(\theta)},$$

loi générique d'une famille exponentielle. Cette loi admet alors une famille conjuguée.

Proposition 2.3.1. *Une famille conjuguée pour $f(x|\theta)$ est donnée par*

$$\pi(\theta|\mu,\lambda) = K(\mu,\lambda)e^{\theta.\mu-\lambda\psi(\theta)}$$

où $K(\mu,\lambda)$ est la constante de normalisation de la densité.
La loi a posteriori correspondante est $\pi(\theta|\mu + x,\lambda + 1)$.

Remarque 2.3.2. *Seuls les modèles à structure exponentielle admettent une famille conjuguée.*

Exemple 2.6. *Le tableau suivant contient quelques lois a priori conjuguées usuelles.*

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
$N(\theta,\sigma^2)$	$N(\mu,\tau^2)$	$N(\varrho(\sigma^2\mu + \tau^2 x),\varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
$P(\theta)$	$\Gamma(\alpha,\beta)$	$\Gamma(\alpha + x,\beta + 1)$
$\Gamma(\nu,\theta)$	$\Gamma(\alpha,\beta)$	$\Gamma(\alpha + \nu,\beta + x)$
$B(n,\theta)$	$Be(\alpha,\beta)$	$Be(\alpha + x,\beta + n + x)$
$Neg(m,\theta)$	$Be(\alpha,\beta)$	$Be(\alpha + m,\beta + x)$
$M_k(\theta_1,...,\theta_k)$	$D(\alpha_1,...,\alpha_k)$	$D(\alpha_1 + x_1,...,\alpha_k + x_k)$
$N(\mu,1/\theta)$	$\Gamma(\alpha,\beta)$	$\Gamma(\alpha + 0.5,\beta + (\mu - x^2)/2)$

Tab1.2-Lois a priori conjuguées usuelles.

Lois a priori subjectives

Précisons tout d'abord que cette démarche n'est pas forcément facile dans la pratique. L'idée est d'utiliser les données antérieures. Par exemple, dans un cadre paramètrique, cela revient à choisir une valeur particulière du paramètre.

Dans un cas concret, il peut être judicieux de baser son raisonnement sur les dires d'experts, notamment à l'aide de questionnaires. Il est alors nécessaire de veiller à ce que les questions soient compréhensibles, par exemple en prenant comme base les quantiles plutôt que les moments. Pour plusieurs experts, il peut être utile de pondérer leurs reponses et d'utiliser des modèles hiérarchiques.

Ainsi, la difficulté ici n'est pas mathématique mais plus psychométrique pour réduire les biais sur les réponses fournies. Nous allons nous concentrer sur le second aspect de la détermination.

2.3.2 Approche non informative

Lorsque aucune information a priori n'est disponible, le choix de la loi a priori est analytique, puisqu'elles donnent des expressions exactes pour quelques quantités a posteriori. Dans de telles situations, il est impossible de justifier le choix d'une loi a priori sur des bases subjectives. Plutôt que de revenir aux alternatives classiques, comme l'estimation par maximum de vraisemblance, ou d'utiliser les données pour approcher ces hyperparamètres, comme dans une analyse Bayésienne empirique, il est préférable de faire appel à des techniques bayésiennes, ne serait-ce que parce qu'elles sont à la base des critères classiques d'optimalité. Dans un tel cas, ces lois a priori particulières doivent être construites à partir de la distribution d'échantillonnage, puisque c'est la seule information disponible. Pour des raisons évidentes, de telles lois sont dites *non informatives*.

Les lois de probabilités non informatives nous amènent souvent à des résultats qui sont des mesures et non des probabilités qu'on appelle des lois impropre c'est-à-dire:

$$\int_{\mathbb{R}} \pi(\theta) d\theta = +\infty$$

l'ensemble des lois a priori impropres constituent un prolongement des lois a priori propres. En effet, elles permettent une bonne description du manque d'information a priori.

Voici quelques approches pour déterminer des lois non informatives:

Lois a priori de Laplace

Laplace fut le premier à utiliser des techniques non informatives puisque, bien que ne disposant pas d'information, il munit ces paramètres d'une loi a priori qui prends en compte son ignorance en donnant la même vraisemblance à chaque valeur du paramètre, soit donc en utilisant une loi uniforme. Son raisonnement, appelé plus tard principe de la raison insuffisante, se fondait sur l'équiprobabilité des événements élémentaires.

Trois critiques ont été plus tard avancées sur ce choix. Premièrement, les lois résultantes sont impropres quand l'espace des paramètres n'est pas compact et certains statisticiens se refusent à utiliser de telles lois, car elles mènent à des difficultés comme le paradoxe de marginalisation.

Deuxièmement, le principe des événements équiprobables de Laplace n'est pas cohérent en termes de partitionnement si: $\Theta = \{\theta_1,\theta_2\}$, la règle de Laplace donne $\pi(\theta_1) = \pi(\theta_2) = 1/2$ mais, si la définition de Θ est plus détaillé, avec $\Theta = \{\theta_1,\theta_2,\theta_3\}$, la règle de Laplace mène à $\pi(\theta_1) = 1/3$, ce qui évidemment n'est pas cohérent avec la première formulation, cette cohérence n'est pas un problème important: il peut être évacué en argumentant que le niveau de partionnement doit être fixé à un certain stade de l'analyse et que l'introduction d'un degré plus fin dans le partitionnement modifie le problème d'inférence.

La troisième critique est plus fondamentale, car elle concerne le problème de l'invariance par reparamétrisation. Si on passe de $\theta \in \Theta$ à $\eta = g(\theta)$ par une transformation bijective g, l'information a priori reste totalement inéxistante et ne devrait pas être modifiée. Cependant, si $\pi(\theta) = 1$, la loi a priori sur η est :

$$\pi^*(\eta) = |\frac{d}{d\eta} g^{-1}(\eta)|$$

Loi a priori de Jeffreys

Les lois a priori non informatives de Jeffreys (dans le cas unidimensionnel) sont définies par

$$p(\theta) = I^{\frac{1}{2}}(\theta)$$

où $I(\theta)$ est la quantité de l'information de Fischer

$$I(\theta) = \mathbb{E}_\theta\{\frac{\partial \log p(x|\theta)}{\partial \theta}\}^2,$$

ce qui, sous certaines conditions de régularité, est égale à

$$I(\theta) = -\mathbb{E}_\theta\{\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2}\}.$$

Dans le cas où $\theta \in \mathbb{R}^k$, on définit la matrice de l'information de Fisher qui a pour éléments

$$I_{ij}(\theta) = -\mathbb{E}_\theta\{\frac{\partial^2 \log p(x|\theta)}{\partial \theta_i \partial \theta_j}\}^2, \qquad i,j = 1,...,k,$$

et la loi non informative de Jeffreys est alors

$$p(\theta) = |I(\theta)|^{\frac{1}{2}}.$$

Notons que, si $p(x|\theta)$ définit une famille exponentielle,

$$p(x|\theta) = h(x)\exp[\theta x - \psi(\theta)],$$

on aura $I(\theta) = \nabla\nabla^t\psi(\theta)$, et

$$p(\theta) = (\prod_{i=1}^{k} \frac{\partial^2\psi(\theta)}{\partial\theta^2})^{\frac{1}{2}}.$$

Exemple 2.7. *Soit* $x \sim \mathcal{N}(\mu,\sigma^2)$ *avec* $\theta = (\mu,\sigma^2)$ *inconnu. Dans ce cas,*

$$I(\theta) = -\mathbb{E}_\theta\{\begin{pmatrix} 1/\sigma^2 & 2(x-\mu)/\sigma^3 \\ 2(x-\mu)/\sigma^3 & 3(x-\mu)^2/\sigma^4 - 1/\sigma^2 \end{pmatrix}\} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix},$$

et la loi non informative associée est

$$P(\theta) = \pi(\mu,\sigma^2) \propto 1/\sigma^2.$$

Deux critiques ont étés faite à cette approche :

- Contradiction avec le principe de vraisemblance, puisque l'information de Fisher dépend des facteurs de proportionalité dans la vraisemblance.
- L'extention multidimensionnelle peut parfois conduire à des incohérences (paradoxe de marginalisation de Stein)

Loi de référence

Cette téchinique a été mis au point par Bernardo (1979), c'est une modification de l'approche de Jeffreys dans le cas unidimensionnel, appelée approche de la loi de référence.

La différence qui caractérise cette méthode est que la loi a priori résultante par la méthode de référence ne dépend pas seulement de la loi d'échantillonnage, mais aussi du problème inférentiel considéré.

Quand $x \sim f(x/\theta)$ et $\theta = (\theta_1,\theta_2)$, où θ_1 est le paramètre d'intérêt, la loi de référence est obtenue en définissant d'abord $\pi(\theta_2/\theta_1)$ comme la loi de Jeffreys associée à $f(x/\theta)$ pour θ_1 fixé, puis en calculant la loi marginale

$$\widetilde{f}(x/\theta_1) = \int f(x/\theta_1,\theta_2)\pi(\theta_2/\theta_1)d\theta_2$$

et la loi de Jeffreys $\pi(\theta_1)$ associée à $\widetilde{f}(x/\theta_1)$.

Loi a priori de concordance (matching priors)

Le but est de trouver une loi a priori cernant le paramètre θ qui se rapproche le plus possible de la méthode de choix fréquentiste, cela revient à faire en sorte que le tirage x n'influence pas le résultat.
On appelle dans un premier temps des exemples d'une région de confiance ou α-crédible. Elle peut être par exemple un intervalle unilatéral $\{\theta \leq \theta_d^{(x)}\}$ ou bien bilatéral $\{\theta_{\alpha,1} \leq \theta \leq \theta_{\alpha,2}\}$. Il peut s'agir aussi de région HPD, $\theta \in \mathcal{C}_\alpha^\pi$ avec par exemple $\{\log(\widehat{\theta}) - \log(\theta) \leq h_\alpha\}$ tel que $P^\pi(\theta \in \mathcal{C}|x) = 1 - \alpha$.

On cherche π tel que $\forall\theta$; $P_\theta(\theta \in \mathcal{C}) = 1-\alpha$, appelé la parfaite concordance. C'est en général impossible. On va chercher r_n le plus petit possible tel que :

$$\forall\theta \in \Theta; \forall\alpha \in]0{,}1[, P_\theta(\theta \in \mathcal{C}) = 1-\alpha+o(r_n)$$

La loi a priori est alors dite concordante à l'ordre r_n.

Lois a priori invariante impropres

Définition 2.3.4. *Une loi impropre M sur Θ est invariante par transformation sur le paramètre h de Θ dans Θ si M est identique à son image par h définie par Moh^{-1}.*

Remarque 2.3.3. *Si M est caractérisée par sa densité π et h est bijective bidérivable, la densité de Moh^{-1} est égale à $\|\rho^{-1}\|Moh^{-1}$, la condition d'invariance s'exprime par l'égalité :*

$$\pi = \|\partial h^{-1}\|\pi oh^{-1}$$

où $\|\partial h^{-1}\|$ est la valeur absolue du déterminant de la matrice des dérivées partielles.
Exemple 2.8. *(Berger et yang 1995)*
Dans le modèle AR(1), en prennant la transformation :

$$h : \rho \mapsto h(\rho) = \frac{1}{\rho}, |\rho| > 1$$

et la loi a priori :

$$\begin{cases} 1/(2\pi\sqrt{(1-\rho^2)}) & si\ |\rho| < 1; \\ 1/(2\pi|\rho|\sqrt{(\rho^2-1)}) & si\ |\rho| > 1. \end{cases}$$

Dans ce cas, la condition d'invariance par la transformation h sur le paramètre ρ s'exprime par l'égalité

$$\pi(h(p)) = \pi(\rho)|\frac{\partial h(\rho)}{\partial\rho}|^{-1}$$

Exemple 2.9. *(Le choix Bayésien 2002)*
La famille de lois $f(x-\theta)$ est invariante par translation, car y = x - x_0 a une loi de la même famille pour tout x_0, $f(y-(\theta-x_0))$, θ est alors dit paramètre de position et une exigence d'invariance est que la loi a priori soit invariante par translation, donc satisfasse

$$\pi(\theta) = \pi(\theta-\theta_0)$$

pour tout θ_0. La solution est $\pi(\theta) = c$ la loi uniforme sur Θ.

2.4 Méthodes de calcul Bayésien

La simplicité ultime de l'approche Bayésienne est que, pour une fonction coût L et une loi a priori π données, l'estimation Bayésienne associée à une observation x est la décision (habituellement unique)d minimisant le coût a posteriori

$$L(\pi,d|x) = \int_\Theta L(\theta,d)\pi(\theta|x)d\theta. \qquad 1.1$$

Cependant dans la pratique, minimiser 1.1 peut être rendu difficile pour deux raisons :

- Le calcul explicite de la loi posteriori, $\pi(\theta|x)$, peut être impossible; et
- même si $\pi(\theta|x)$, est connu, cela n'implique pas nécessairement que minimiser 1.1 soit facile; en effet, lorsque l'intégration analytique est impossible, la minimisation numérique nécessite parfois du temps de calcul considérable, en particulier lorsque Θ et $\mathcal{D}$ sont de grandes dimensions.

2.4.1 Méthode classique d'approximation

A partir de la simple méthode de Simpson, plusieurs approches ont été conçues en mathématiques appliquées pour l'approximation numérique d'intégrales. Par exemple, la quadrature polynômial est censée approcher les intégrales liées à des distributions proches de la loi normale. L'approximation de base est donnée par

$$\int_{-\infty}^{+\infty} e^{-t^2/2} f(t)dt \approx \sum_{i=1}^{n} w_i f(t_i)$$

où

$$w_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(t_i)]^2}$$

et t_i est le i-ième zéro du n-ième polynôme d'Hermite $H_n(t)$

D'autres approximations d'intégrales reliées à la méthode précédente sont disponibles, qui reposent sur différentes bases orthonormales classiques (voir Abramowitz et Stegun, 1964)mais ces méthodes requièrent généralement des hypothèses de régularité sur la fonction f, ainsi que des études préliminaires pour déterminer quelle base est la plus adéquate et à quel point cette approximation est précise. Par exemple, des transformations du modèle peuvent être nécessaires pour mettre en pratique l'approximation d'Hermite (voir Naylor et Smith, 1982 et Hills et Smith 1992).

Remarque 2.4.1. *Quelle que soit la méthode d'intégration numérique utilisée, sa précision diminue dramatiquement lorsque la dimension de Θ augmente. De façon plus spécifique, l'erreur associée aux méthodes numériques se comporte comme une puissance de la dimension de Θ. En pratique, une règle empirique est que la plupart des méthodes standard ne devraient pas être utilisées pour l'intégration en dimension supérieure à 4. En effet, la taille de la partie de l'espace non pertinente pour le calcul d'une intégration donnée augmente considérablement avec la dimension de l'espace. Ce problème est appelé fléau de la dimension.*

2.4.2 Méthodes de Monte Carlo par Chaînes de Markov (MCMC)

Les méthodes de Monte-Carlo par chaîne de Markov (MCMC) génèrent une suite de variables aléatoires (θ^1,...,θ^n,...) et, hormis la première à laquelle on donne une valeur arbitraire, chacune d'entres elles dépend uniquement de celle qui la précède, les calculs sont ensuite poursuivis en appliquant à cette séquence une loi des grands nombres pour les chaînes markoviennes ergodiques de forme identique.

Algorithme Metropolis-Hasting

Pour $\theta^{(0)}$ est une valeur initiale, on définit par récurrence les valeurs de $\theta^{(t)}$. A l'étape t, à partir de $\theta^{(t-1)}$, $\theta^{(t)}$ est construit en tirant un $\theta^{'}$ à l'aide d'une distribution de probabilité instrumentale : $\theta^{'} \sim q(.|\theta^{(t-1)})$.$\theta^{(t)}$ est alors donné par :

$$\theta^{(t)} = \begin{cases} \theta^{'} & \text{avec une probabilité} \alpha(\theta^{'},\theta^{(t-1)}) \\ \theta^{(t-1)} & \text{avec une probabilité } 1-\alpha(\theta^{'},\theta^{(t-1)}) \end{cases}$$

où $\alpha(\theta^{'},\theta^{(t-1)}) = \min(\frac{\pi(\theta^{'})q(\theta^{(t-1)}|\theta^{'})}{\pi(\theta^{(t-1)}q(\theta^{'}|\theta^{(t-1)}))},1)$.
La loi de densité $\pi(\theta)$ est souvent appelée *loi cible* ou loi objet, tandis que la loi de densité $q(.|\theta)$ est dite *loi de proposition*. Une propriété stupéfiante de cet algorithme est d'autoriser un nombre infini de lois de proposition produisant toute une chaîne de Markov convergeant vers la loi d'intérêt.

Remarque 2.4.2. *Notons qu'il est possible suivant cette construction de rester au même endroit aprés une itération. On peut alors montrer en écrivant la condition de balance, que pour ce choix de α, on obtient une chaîne de Markov de loi stationnaire π.*
Cette chaîne de Markov est ergodique si et seulement si $(\theta^{(t)})_t$ est irréductible et apériodique.

L'echantillonnage de Gibbs

Le second groupe de méthodes de Monte Carlo par chaînes de Markov (MCMC) est encore appelé échantillonnage de Gibbs. Plus intuitif pour certains praticiens, il ne demande pas de mettre en place une fonction d'exploration de l'espace des états de la nature. De plus, algorithme de Gibbs pour l'estimation et construction de modèle par conditionnement probabiliste. De fait, les méthodes de Gibbs utilisent plus complètement que ne le font les méthodes de Metropolis-Hasting, les structures conditionnelles des modèles.
Pour $\theta = (\theta_1,...,\theta_p)$, on veut simuler $\pi(\theta)$ à partir de $\pi_i(\theta_i|\theta_{(-i)}) = \pi_i(\theta_i|\theta_j, j \neq i)$ pour tout i. On initialise avec $\theta^{(0)}$ et à l'instant t, on écrit :

$$(\theta_1^{(t)}|\theta^{(t-1)}) \sim \pi_1(\theta_1^{(t)}|\theta_{(-1)}^{(t-1)})$$

$$(\theta_2^{(t)}|\theta^{(t-1)},\theta_1^{(t)}) \sim \pi_2(\theta_2^{(t)}|\theta_1^{(t)},\theta_3^{(t-1)},...,\theta_p^{(t-1)})$$

$$(\theta_p^{(t)}|\theta^{(t-1)},\theta_{(-p)}^{(t)}) \sim \pi_p(\theta_p^{(t)}|\theta_{(-p)}^{(t)})$$

Dans le cas où une telle loi π existe, $\theta^{(t)}$ issu de cet algorithme est une chaîne de Markov ergodique de loi stationnaire π.

2.5 Avantages de l'approche Bayésienne

– La vision Bayésienne des probabilités, nous semble plus cohérente que les théories fréquentistes qui interprètent la probabilité comme une proportion limite déterminée

sur la base d'une séquence infinie d'expériences (Von Mises et Geiringer, 1964). D'ailleurs on associe souvent une probabilité à des événements qui par définition ne sont pas répétables,(par exemple l'argumentaire de Hartigan (1983) autour de l'assertion, qui n'est heureusement plus d'actualité : *"Je pense que la probabilité qu'un conflit nucléaire éclate entre Etas Unis et Union Soviétique avant l'année 2000 est de 0.05"* L'interprétation Bayésienne de la probabilité (De Finetti, 1937, 1974) est associée à une notion de pari rationnel : la probabilité attribuée à un événement est définie par les conditions auxquelles un individu rationnel est prêt à parier sur la réalisation de tel événement. La rationnalité de l'individu, est nécessaire pour éviter que la définition de la probabilité soit arbitraire et est décrite par certaines règles de comportement, face à l'incertitude (Savage, 1972).

– La question de choisir une interprétation Bayésienne ou fréquentiste de la probabilité n'est pas uniquement philosophique. Il est possible de combiner les informations objectives, apportées par les données, avec des informations extérieures à l'échantillon observé et la formule de Bayes est l'instrument qui rend possible ce couplage dans un cadre chérent basé uniquement sur le calcul des probabilités (Bernardo et Smith, 1994). Le rôle de la formule de Bayes comme trait d'union entre la vision du modèlisateur et l'evidence des résultats expérimentaux est souligné, entre autres, par Box et Tiao (1973), Berry et al.(1996), Press et Tanur (2000).
Cette possibilité d'intégrer des éléments extérieurs aux échantillon est trés adaptée à la pratique technique. Dans le monde réel, le statisticien est normalement confronté à des données peu répresentatives ou incohérentes mais en revanche il dispose de l'avis technique des experts qui , sur la base de leur expérience et savoir-faire, sont capables de donner des informations complémentaires de grande utilité, qu'il serait dommage de ne pas prendre en compte (Cullen et Frey, 1999), (Bernier et al., 2000), (Perreault, 2000).

– L'analyse Bayésienne fournit des résultats d'interprétation plus directe que ceux de la statistique classique. L'exemple le plus flagrant est la définition de l'intervalle de confiance du paramètre θ d'un modèle. Pour les Bayésiens, qui préfèrent parler plutôt d'intervalle de crédibilité[3] [3], il s'agit de l'intervalle qui contient le paramètre avec la probabilité donnée. Par exemple l'intervalle de crédibilité a posteriori à 95% est typiquement celui délimite inférieurement par le percentile d'ordre 2.5% et supérieurement par le percentile d'ordre 97.2%. Dans l'inférence classique cette assertion n'est plus vraie parce que le paramètre (inconnu) du modèle n'est pas une variable aléatoire mais une grandeur constante. L'interprétation correcte de l'intervalle de confiance est que, si on imagine l'ensemble des échantillons aléatoires pouvant être obtenus à partir du modèle, paramétré par 0.95% des intervalles de confiance calculés (sur la base des différents échantillons) contiennent la vraie valeur du paramètre. L'interprétation Bayésienne, décidément plus naturelle, est d'ailleurs celle de la plupart des particiens qui font de l'inférence Bayésienne ... sans le savoir (Lecoutre et Poitevineau, 1996).

– Les résultats de l'inférence Bayésienne sont plus riche que les estimateurs fournis par les techniques classiques d'inférences (Berger, 1985). Les techniques Bayésiennes

3. Autre terminologies d'usage moins courant : Region à la plus grande probabilité ou HDR, (Lee, 1997), Intervale Bayésien de confiance (Lindley 1965)

permettent d'obtenir la loi jointe des paramètres du modèle et donc de prendre en compte simultanément l'effet de l'incertitude globale sur l'ensemble des paramètres inconnus sur les prévisions futures du comportement du système étudié et sur les décisions suggérées par ce comportement (Krzysztofowicz, 1983).

– Enfin, les techniques d'inférence par les méthodes MCMC, relativement faciles à mettre en oeuvre, sont trés adaptées à l'estimation de modèles complexes à plusieurs paramètres ou à structure hiérarchique. L'estimation de ces modèles avec la technique usuelle du Maximum de Vraisemblance se révèle parfois nettement plus compliqué. L'argument de la commodité opérationnelle qui a longtemps joué contre l'approche Bayésienne commence aujourd'hui à peser dans le sens opposé dans la vieille querelle entre fréquentistes et Bayésiens (Robert, 1992).

2.6 Le Facteur de Bayes

2.6.1 Facteur de Bayes

Bien que, d'un point de vue decisionnel, le *facteur de Bayes* ne soit qu'une tranformation bijective de la probabilité a posteriori, il a fini par être considéré comme réponse en soi en théorie des tests bayésiens, sous l'implusion de Jeffreys (1939).

Définition 2.6.1. *Le facteur de Bayes est le rapport des probabilités a posteriori des hypothèses nulle et alternative sur le rapport des probabilités a priori des ces mêmes hypothèses, soit*

$$B_{01}^{\pi}(x) = \frac{P(\theta \in \Theta_0)}{P(\theta \in \Theta_1)} \Big/ \frac{\pi(\theta \in \Theta_0)}{\pi(\theta \in \Theta_1)}$$

Ce rapport évalue la modification de la vraisemblance de l'ensemble Θ_0 par rapport à celle de l'ensemble Θ_1 due à l'observation et peut se comparer naturellement à 1, bien qu'une échelle de comparaison exacte doive être fondée sur une fonction de cout. Dans le cas particulier *où* $\Theta_0 = \theta_0$ et $\Theta_1 = \theta_1$, le facteur de Bayes se simplifie et devient le *rapport de vraisemblance* classique

$$B_{01}^{\pi}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}.$$

En général, le facteur de Bayes dépend de l'information a priori, mais il est souvent proposé comme réponse bayésienne "*objective*", car il élimine partiellement l'influence du modèle a priori et souligne le rôle des observations. De fait, il peut être perçu comme un rapport de vraisemblance bayesien, car, si π_0 est la loi a priori sous H_0 et π_1, la loi a priori sous H_1 $B_{01}^{\pi}(x)$ peut s'ecrire

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x|\theta_0)\pi_0(\theta)d\theta}{\int_{\Theta_1} f(x|\theta_0)\pi_1(\theta)d\theta}$$

ce qui revient donc à remplacer les vraisemblances par des marginales sous les deux hypothèses. Le facteur de Bayes est un critère de sélection de modèles, comme il est un outil

pour comparer la crédibilité de deux hypothèses.

– **Cas de sélection de modèles.** Lorsqu'on est appelé à faire une selection entre deux modèles $\mathcal{M}_1$ et $\mathcal{M}_2$ par l'approche bayésienne, on affecte à chacun des deux modèles des probabilités a priori $P(\mathcal{M}_i)$,$i = 1,2$ vérifiant

$$P(\mathcal{M}_1) + P(\mathcal{M}_2) = 1.$$

Du théorème de Bayes, on peut alors exprimer les probabilités a posteriori

$$P(\mathcal{M}_i) = \frac{P(X/\mathcal{M}_i)P(\mathcal{M}_i)}{m(X)} i = 1,2.$$

où X = $(X_1,...X_n)$ est le n-échantillon, et m(X) = $P(X/\mathcal{M}_1)P(\mathcal{M}_1)+P(X/\mathcal{M}_2)P(\mathcal{M}_2)$, représente la probabilité marginale des données (qu'on suppose nulle) avec $P(\mathcal{M}_2/X) = 1 - P(\mathcal{M}_1/X)$. On peut écrire

$$\frac{P(\mathcal{M}_1/X)}{P(\mathcal{M}_2/X)} = \frac{P(X/\mathcal{M}_1)}{P(X/\mathcal{M}_2)} \times \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)} \qquad (1)$$

où

$$B_{12} = \frac{P(X/\mathcal{M}_1)}{P(X/\mathcal{M}_2)}$$

répresente le facteur de Bayes en faveur du modèle $\mathcal{M}_1$ contre $\mathcal{M}_2$.

Le rapport $\frac{P(\mathcal{M}_1/X)}{P(\mathcal{M}_2/X)}$ est dit quotient d'enjeux a posteriori en faveur du modèle $\mathcal{M}_1$ contre $\mathcal{M}_2$. dans la littérature anglaise il est dit "posterior odds ratio".

le rapport $\frac{P(X/\mathcal{M}_1)}{P(X/\mathcal{M}_2)}$ est dit quotient d'enjeux a priori en faveur de $\mathcal{M}_1$ contre $\mathcal{M}_2$. Dans la littérature anglaise, il est dit "prior odds ratio".

On peut alors exprimer la relation (1) comme suit:
Quotient d'enjeux a posteriori = Facteur de Bayes × Quotient d'enjeux a priori

Remarque 2.6.1. *Si on accorde le même poids a priori pour les deux hypothèses i.e*

$$P(\mathcal{M}_1) = P(\mathcal{M}_2) = \frac{1}{2}.$$

Le facteur de Bayes s'écrit

$$B_{12} = \frac{P(\mathcal{M}_1/X)}{P(\mathcal{M}_2/X)}.$$

Dans cette situation, il correspond exactement au quotient d'enjeux a posteriori.

– **Cas de test d'hypothèses.** On considère le test d'hypothèses composites:

$$H_0 : \theta \in \overline{\Theta}_0 contre H_1 : \theta \in \overline{\Theta}_1$$

La loi a priori est donnée par le mélange de lois

$$\pi(\theta) = P(H_0)\pi_0(\theta) + (1 - P(H_0)\pi_1(\theta))$$

où $\pi_i(\theta) = P(\theta/H_i)$ est la probabilité a priori de θ sous l'hypothèse H_i et $P(H_1) = 1 - P(H_0)$.
La fonction de vraisemblance sous H_i s'écrit:

$$L(X/H_i) = \int_{\theta \in \overline{\Theta}_i} L(X/H_i,\theta)\pi_i(\theta)d\theta, \qquad i = 0,1$$

Le facteur de Bayes de l'hypothèse H_0 en faveur de l'hypothèse H_1 s'exprime par:

$$B_{01} = \frac{L(X/H_0)}{L(X/H_1)}$$

Comme nous l'avons déja indiqué, le facteur de Bayes donne un indicateur objectif de l'hypothèse $H_0 : \theta \in \overline{\Theta}_0$. Malheureusement, l'utilisation d'une loi impropre rend impossible le calcul de ce facteur. En effet, les probabilités a posteriori des hypothèses nulle et alternative ne sont pas définies pour une telle loi. La résolution générale de cette incompatibilité entre les tests bayésiens et lois a priori impropre reste un problème ouvert, même si plusieurs solutions partielles ont été déja proposées, via la définition de "pseudo-facteur" (le facteur de Bayes fractionné, voir Conigliani et O'Hagan, 2000 et le facteur de Bayes intrisèque, voir O'Hagan, 1995, 1997)

Remarque 2.6.2. *1. Lorsque les deux hypothèses sont équiprobables, le facteur de Bayes correspond au rapport des probabilités a posteriori de H_0 sur H_1.*

$$B_{01} = \frac{P(H_0/X)}{P(H_1/X)}$$

2. Dans le cas d'un test d'une hypothèse simple contre une hypothèse composite suivant: $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ où $P(H_0) = P(H_1) = \frac{1}{2}$ et $\pi_0(\theta)$ est la masse de Dirac au point θ_0
Le facteur de Bayes s'écrit

$$B_{01} = \frac{L(X/\theta_0)}{\int_{\theta \neq \theta_0} L(X/\theta)\pi_1(\theta)d\theta}$$

3. Le facteur de Bayes est une quantité non borné, il peut prendre des valeurs entre 0 et ∞

Interpretation

A la suite de Jeffreys (1939) et de Good (1952), le facteur de Bayes est désormais un outil à part entière (voir, par exemple, Kass et Raftery, 1995, pour une revue détaillé). En particulier, Jeffreys (1939) a développé une échelle "*absolue*" pour évaluer le degré de certitude en faveur ou au détriment de H_0 apporté par les données, *en l'absence d'un cadre décisionnel véritable.*
L'échelle de Jeffreys est la suivante:

1. si $log_{10}(B^{\pi}_{10})$ varie entre 0 et 0.5, la certitude que H_0 est $faible$
2. si elle est entre 0.5 et 1, cette certitude est *substantielle*
3. si elle est entre 1 et 2, elle est $forte$ et

4. si elle est au-dessus de 2, elle est *décisive*,

avec la même échelle en faveur de H_0 pour les valeurs négatives. Bien entendu, cette graduation du facteur de Bayes donne quelques indications sur le degré de certitude, mais les limites précises séparant une catégorie d'une autre sont conventionnelles et peuvent être changées de façon arbitraire, comme l'ont illustré Kass et Raftery (1995).

Exemple 2.10. *(Suite de l'Example 2.1) Si l'on teste* $H_0 := \theta = 0$ *contre* $H_1 := \theta \neq 0$, *on compare le modèle* $\mathcal{M}_1$ *où* $x \sim \mathcal{N}_1(0,1)$ *au modèle* $\mathcal{M}_2$ *où* $x \sim \mathcal{N}_1(\theta,1)$ *et* $\theta \sim \mathcal{N}_1(0,10)$. *Le facteur de Bayes* $B_{12}(x)$ *est donc le rapport des dénsités marginales*

$$B_{12}(x) = \frac{(1/\sqrt{2\pi})\exp(-x^2/2)}{(1/\sqrt{2\pi})\int_{\mathbb{R}} \frac{1}{\sqrt{20\pi}}\exp(\frac{-\theta^2}{20})\exp(\frac{-(x-\theta)^2}{2})d\theta} = \frac{\exp(-x^2/2)}{\sqrt{1/11}\exp(-x^2/2)} = \sqrt{11}\exp(-10x^2/22)$$

Le maximum de $B_{12}(x)$ *est atteint pour x=0 et est favorable [au sense de l'echelle ci-dessus] à* $\mathcal{M}_1$ *puisqu'il prend la valeur* $\sqrt{11}$, *de logarithme égal à 1.2. De plus* $\log_{10} B_{12}(x)$ *vaut 0 pour x=1.62, et -1 pour x=2.19. On peut remarquer la difference avec les bornes classiques, puisque x=1.62 correspond presque à un niveau de significativité de 0.1 et x=2.19 à un niveau de significativité de 0.01. On rejettera donc plus difficilement l'hypothèse nulle* $H_0 : \theta = 0$ *en utilisant une procédure bayésienne.*

Signalons que la différence notée dans cet exemple tient à une interprétation radicalement différente de l'erreur de mesure: dans l'approche classique des tests, l'erreur traditionnellement de 5% correspond à la probabilité de rejeter à tort l'hypothèse nulle et découle donc d'un choix implicite d'une fonction de coût asymétrique. Dans l'approche bayésienne, le facteur de Bayes compare les deux probabilités $\mathbb{P}(\mathcal{M}_1|x)$ et $\mathbb{P}(\mathcal{M}_2|x)$, et ne conclut que si elles différent suffisamment.
Il est à noter qu'il existe d'autres critères bayésiens de comparaison de modèles tels que le critère d'information bayésien noté BIC introduit par Schwarz appelé aussi critère de Schwarz (voir Schwarz, 1978)

2.6.2 Le critère de Schwarz

Pour les modèles réguliers $\mathcal{M}_1 \subset \mathcal{M}_2$, le rapport de vraisemblance entre $\mathcal{M}_2$ et $\mathcal{M}_1$ est approximativement distribué selon une loi $\chi^2_{p_2-p_1}$,

$$-2\log\lambda_n \approx \chi^2_{p_2-p_1}$$

où p est la dimension de Θ et en supposant que $\mathcal{M}_1$ est le vrai modèle (Gouriéroux et Monfort, 1996, et Lehmann et Casella, 1998). On a

$$P(\mathcal{M}_2 choisi|\mathcal{M}_1) = P(\lambda_n < c|\mathcal{M}_1) \approx P(\chi^2_{p_2-p_1} > -2\log(c)) > 0.$$

Donc d'un point de vue fréquentiste, un critère dépendant seulement du rapport de vraisemblance ne converge pas vers une réponse certaine sous $\mathcal{M}_1$ (mais il converge sous $\mathcal{M}_2$). C'est la raison pour laquelle on ajoute des facteurs de pénalisation au rapport de vraisemblance pour compenser ce biais, comme dans le cas du critère d'Akaike (1983),

$$-2\log\lambda_n - \alpha(p_2 - p_1).$$

Pour $\alpha = \log 2$, on retrouve l'approximation obtenue par une procédure d'Aitkin (1991) dans laquelle l'auteur utilise les données deux fois, une première fois pour construire un (pseudo) a priori propre en utilisant la distribution a posteriori, puis une seconde fois pour calculer le facteur de Bayes comme si la distribution a priori était exacte.

Le développement de Laplace donne une approximation d'intégrale,

$$\int_\Theta \exp\{nh(\theta)\}d\theta = \exp\{nh(\widehat{\theta})\}(2\pi)^{p/2}n^{-p/2}|H^{-1}(\widehat{\theta})| + O(n^{-1}),$$

où p est la dimension de Θ, $\widehat{\theta}$ le point où h atteint son maximum et H la matrice hessienne de h. En développant à la fois le numérateur et le dénominateur du facteur de Bayes grâce à cette approximation, on obtient:

$$B_{12}^\pi \simeq \frac{L_{1,n(\widehat{\theta}_{1,n})}}{L_{2,n(\widehat{\theta}_{2,n})}}|\frac{H_1^{-1}(\widehat{\theta}_{1,n})}{H_2^{-1}(\widehat{\theta}_{2,n})}|^{1/2}(\frac{n}{2\pi})^{(p_2-p_1)/2},$$

avec p_1 et p_2 dimensions de Θ_1 et Θ, $L_{1,n}$ et $L_{2,n}$ fonctions de vraisemblance calculées sur n observations, et $\widehat{\theta}_{1,n}$ et $\widehat{\theta}_{2,n}$ maximums respectifs de L_1 et L_2. D'où:

$$\log(B_{12}^\pi) \simeq \log\lambda_n + \frac{p_2-p_1}{2}\log(n) + K(\widehat{\theta}_{1,n},\widehat{\theta}_{2,n}),$$

en notant λ_n le rapport de vraisemblance usuel pour la comparaison de $\mathcal{M}_1$ et $\mathcal{M}_2$,

$$\lambda_n = L_{1,n}(\widehat{\theta}_{1,n})/L_{2,n}(\widehat{\theta}_{2,n}),$$

et $K(\widehat{\theta}_{1,n},\widehat{\theta}_{2,n})$ le terme restant.

Cette approximation est à l'origine du *critère de Schwarz* (Schwarz, 1978): pour $\mathcal{M}_1 \subset \mathcal{M}_2$, le facteur de Bayes est approché par

$$S = -\log\lambda_n - \frac{p_2-p_1}{2}\log(n)$$

Le critère de Schwarz, également appelé BIC (pour *Bayes Information Criterion*), est donc une première approximation à l'ordre 1 du facteur de Bayes, comme le décrivent Kass et Raftery (1995).

Exemple 2.11. *Un exemple cité dans la plupart des ouvrages portant sur l'estimation des mélanges est celui des données galactiques. D'abord abordé par Roeder (1992), il a ensuite été analysé par, entre autres, Chib (1995), Escobar et West (1995), Phillips et Smith (1996), Richardson et Green (1997), Roeder er Wasserman (1997) et Robert et Mengersen (1999). Il consiste en l'observation de quatre-vingt-deux vitesses de galaxies. Pour des raisons liées à l'Astrophysique, cet ensemble peut être modélisé par un mélange de distributions normales dont le nombre de composantes k est inconnu. les modèles esn concurrence sont donc*

$$\mathcal{M}_i : n_j \sim \sum_{l=1}^{i} p_{l_i}\mathcal{N}(\mu_{l_i},\sigma_{l_i}^2),$$

Aprés maintes étapes, on décompose le critère de Schwarz en

$$S = \log\{L_{2,n}(\widehat{\theta}_{2,n})/L_{1,n}(\widehat{\theta}_{2,n})\} - \frac{p_2 - p_1}{2}\log(n) = \log L_{2,n}(\widehat{\theta}_{2,n}) - \frac{p_2}{2}\log(n) - \log L_{1,n}(\widehat{\theta}_{1,n}) + \frac{p_1}{2}\log(n).$$

La partie relative au modèle $\mathcal{M}_i$ est donc

$$S = \log L_{i,n}(\widehat{\theta}_{i,n}) - \frac{p_i}{2}\log(n)$$

Si $\mathcal{M}_k$ est assoccié à la composante k du modèle, p_k = 3k - 1. Pour les données de vitesses de galaxies, Raftery (1996) obtient

$$S_1 = -271.8, S_2 = -249.7, S_3 = -256.7, S_4 = -263.6,$$

en utilisant l'algorithme EM (Espérance Maximisation) pour obtenir des approximations des estimateurs de maximum de vraisemblance $\widehat{\theta}_{i,n}$ pour $k > 1$. On en déduit que, selon le critère de Schwarz, il faut préférer le modèle à deux composantes aux autres.

2.6.3 Facteur de Bayes contre Critère de Schwarz

Les facteurs de Bayes offrent une manière d'evaluer l'evidence en faveur d'une hypothèse nulle, ils offrent aussi une méthode d'incorporer l'information externe dans l'evaluation de l'evidence au sujet d'une hypothèse. Les facteurs de Bayes sont généraux et sont utiles pour guider un procédé évolutionnaire de modélisation. les facteurs de Bayes ont plusieurs forces dont l'essentielle est leur base logique pleine qui offre une grande flexibilité. Malgré ces avantages, les facteurs de Bayes sont sensibles aux prétentions dans le modèle paramètrique et le choix du prior.

Quant au critère de Schwarz (ou BIC), il donne une approximation brute au logarithme du facteur de Bayes. Il est facile à employer et n'exige pas l'evaluation des priors distributions.
Néanmoins, la pertinence de ce critère dans le contexte bayésien est contestable pour deux raisons

1. L'influence de l'hypothèse a priori disparaît;
2. Cette approximation n'est acceptable que pour les modèles réguliers.

2.7 Conclusion

La pertinence et l'éfficacité de l'approche Bayésienne, comme guide du raisonnement scientifique face à l'incertitude, sont reconnus depuis longtemps (Finetti, 1937; Savage, 1954; etc...). La mise en oeuvre des principes Bayésiens, en dehors du cas d'école, s'est longtemps heurtée aux difficultés pratiques de calcul. Les moyens informatiques, dont on disposait avant les années 1990, étaient insuffisamment puissants. Les problèmes réels, avec leurs dimensions et leurs complexités importantes, faisaient alors la part belle aux méthodes statistiques classiques. La situation, depuis lors, a subi une véritable révolution. On doit ce nouveau paysage scientifique au développement de nouveaux outils de

calcul : les méthodes MCMC (simulations Monte Carlo par Chaînes de Markov) et à l'amélioration des anciens (échantillonnage pondéré ou importance sampling et méthodes des particules), et à leur relance par la puissance nouvelle de la micro-informatique. Les fondements conceptuels des ces méthodes de calcul sont solidaires des modes de raisonnements conditionnels de la modélisation Bayésienne et le paradigme Bayésien apparaît comme une démarche rationnelle, efficace et solidement intégrée du programme complet : modélisation → calcul → décision.

Chapitre 3

Test Bayésien sur les modèles VAR

"La connaissance populaire a besoin d'une manière,
la science d'une méthode,
c'est-à-dire d'un ensemble de procédés reposant
sur des principes de raison."
E.Kant

3.1 Introduction

Pour combiner nos deux chapitres précédents nous proposons un résumé de l'article intitulé "Bayesian testing of restrictions on vector autoregressive models" (Dongchu Sun & Shawn Ni, Journal of statistical planning and inference, 2012).
Les vecteurs autoregressifs (VAR) sont devenus un important outil pour l'analyse des séries chronologiques dans plusieurs domaines. les auteurs ont défini le VAR Y_t comme suit:

$$Y_t^{'} = C + \sum_{j=1}^{L} Y_{t-j}^{'} B_j + \epsilon_t^{'} \tag{1}$$

avec C: un vecteur colonne inconnu, B_j une matrice inconnu $p \times p$ et ϵ_t sont des bruits blancs i.i.d $\sim \mathcal{N}_p(0,\Sigma)$, et la matrice de variance-covariance Σ est inconnue $p \times p$ et definie positive.

L'application des modèles VAR exige souvent des restrictions sur les coefficients de régression $\phi = (C^{'},B_1^{'},...,B_L^{'})$ et la matrice de variance-covariance Σ dont la décomposition de Cholesky est

$$\Sigma = \Psi^{-1'}\Psi^{-1} \tag{2}$$

avec Ψ une matrice triangulaire supérieure.

L'analyse Bayésienne combine l'information venant de l'échantillon et du prior pour former la distribution a posteriori de l'échantillon fini des paramètres. Dans cette étude, ils proposent un ensemble de lois a priori sur les composantes de Ψ. Toutes les lois a posteriori conditionnelles sont de distributions standard et usuelles. Et cela a deux conséquences

avantageuses:

- Premièrement, il a été possible de calculer la loi a posteriori Ψ et Φ en utilisant l'algorithme classique de MCMC (Markov Chain Monte Carlo) de Metropolis
- En second lieu, la loi a posteriori conditionnelle dans une distribution standard permet de calculer la vraisemblance marginale et le facteur de Bayes par une approche suggerée par Chib (1995).

Dans la littérature courante des VAR, le test d'hypothèse est souvent basé sur le critère de Schwarz, qui approxime bien le log du facteur de Bayes quand l'échantillon est de grande taille.
La loi a priori qu'ils ont utilisée dans leur article est similaire à celle utilisée par Daniels et Pourahmadi (2002) pour les données longitudinales même si leur usage diffère de leur méthode.

L'objectif principal de leur étude est d'illustrer pour les utilisateurs bayésiens du VAR, la performance du facteur de Bayes quand l'échantillon est fini comparativement au critère de Schwarz.

Dans leur article, ils se sont focalisés sur la selection des modèles VAR basés sur la décomposition de Cholesky pour deux raisons:

- D'abord, une telle structure accorde une agréable forme analytique de la loi a posteriori conditionnelle
- Ensuite, et plus important, il assure que les modèles restreints sont globalement bien identifiés

Nous donnons ci-aprés une présentation sommaire de ce travail.

3.2 Modèle VAR Bayésien

Soit le modèle suivant appelé par les auteurs "modèle VAR Bayésien"

$$Y = X\Phi + \epsilon \tag{3}$$

où $X_t = (1, Y'_{t-1}, ..., Y'_{t-L})$

$$Y = \begin{pmatrix} Y'_1 \\ . \\ . \\ . \\ Y'_T \end{pmatrix} \quad X = \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_T \end{pmatrix} \quad \Phi = \begin{pmatrix} c \\ B_1 \\ . \\ . \\ . \\ B_L \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ . \\ . \\ . \\ \epsilon_T \end{pmatrix} \tag{4}$$

avec Y et ϵ des matrices $t \times p$, Φ est une $(1 + L_p) \times p$, X_t est un $1 \times (1 + L_p)$ vecteur ligne, and X est une $T \times (1 + L_p)$ matrice des observations.

La fonction de vraisemblance de (Φ,Σ) est

$$l(\Phi,\Sigma) = f(Y/\Phi,\Sigma) \propto |\Sigma|^{-T/2} etr\{-\frac{1}{2}\Sigma^{-1}S(\Phi)\} \tag{5}$$

avec etr(A) = exp(tr(A)) et $S(\Phi) = (Y - X\Phi)'(Y - X\Phi)$

L'estimateur de (Φ,Σ) généralement utilisé est le MLE (Estimateur du maximum de vraisemblance)

$$(\widehat{\Phi}_M,\widehat{\Sigma}_M) = ((X'X)^{-1}X'Y,S(\widehat{\Phi})/T) \tag{6}$$

avec T la taille de l'échantillon assez grande.

3.2.1 Lois a Priori et a posteriori des modèles VARs identifiés

Paramétrisation identifiée simplement(Just-identified)

Ils ont pris la loi a priori suivante

$$\phi \sim \mathcal{N}(\phi_0,\Xi_0) \tag{7}$$

Ψ est donné en (2) tout comme comme Ψ_p et on définit $\psi_{p-1,p}$ comme le vecteur (p-1)×1 représentant la dernière colonne de la matrice Ψ_p à laquelle on enlève Ψ_{pp}. En utilisant cette notation de façon itérative on obtient

$$\begin{aligned}
\Psi_1 &= \psi_{11} \\
\Psi_2 &= \begin{pmatrix} \Psi_{11} & \psi_{12} \\ 0 & \psi_{22} \end{pmatrix} \\
\dots &= \dots \\
\Psi_p &= \begin{pmatrix} \Psi_{p-1} & \psi_{p-1,p} \\ 0 & \psi_{pp} \end{pmatrix}
\end{aligned} \tag{8}$$

Les lois a priori pour les élements extra-diagonaux sont supposées être indépendantes de lois normales multivariées.

$$\Psi_{i-1,i} = \mathcal{N}(0,\Omega_{i-1}^{-1}) \qquad i = 1,...,p \tag{9}$$

Ils considèrent la loi a priori pour les éléments diagonaux car la loi gamma est jugée plus globale et flexible que, par exemple, le χ_2. Avec cette loi, Eaton et Olkin (1987) ont obtenu un estimateur Bayésien qui est un cas particulier de ce présent travail.

Toutes les distributions conditionnelles leurs seront utiles pour les simulations par les méthodes MCMC. Notons que la fonction de vraisemblance en (5) peut être réecrite comme suit

$$l(\phi,\Sigma) \propto |\Sigma|^{-T/2} exp[-\frac{1}{2}(\phi - \widehat{\phi}_M)'\{\Sigma^{-1} \otimes (X'X)\}(\phi - \widehat{\phi}_M) - \frac{1}{2} tr\{\Sigma^{-1} S_M\}] \tag{10}$$

où $\widehat{\phi}_M = vec(\widehat{\Phi}_M)$, $S_M = S(\widehat{\Phi}_M)$ qui est donné en (6) et $\otimes$ le produit tensoriel .

Combinant (10) et (7) ils trouvent que la densité a posteriori conditionnelle de ϕ donné par $(\Psi_p; Y)$ est

$$[\phi|\Psi_p; Y] =\propto exp\{-\frac{1}{2}(\phi - \widehat{\phi}_M)'\{\Sigma^{-1} \otimes (X'X)\}(\phi - \widehat{\phi}_M) - \frac{1}{2}(\phi - \widehat{\phi}_0)'\Xi^{-1}(\phi - \widehat{\phi}_0)\}$$

Dans la suite de leur travail ils ont noté la distribution conditionnelle par (.|.) et la densité conditionnelle par [.|.]

Théorème 3.2.1. *La distribution a posteriori conditionnelle de ϕ sachant $(\Psi_p; Y)$ est*

$$(\phi|\Psi_p; Y) \sim \mathcal{N}(\widehat{\phi},\widehat{\Xi}), \tag{11}$$

où

$$\widehat{\Xi} = \{\Sigma^{-1} \otimes (X'X) + \Xi_0^{-1}\}^{-1} \tag{12}$$

$$\widehat{\phi} = \widehat{\Xi}\{\Sigma^{-1} \otimes (X'X)\widehat{\phi}_M + \Xi_0^{-1}\phi_0\} \tag{13}$$

Théorème 3.2.2.

1. *On se donne $(\phi,\psi_{11},...,\psi_{pp}; Y)$. $\psi_{i-1,i}$ i=2,...,p sont mutuellement indépendants et $(\psi_{i-1,i}|\phi,\psi_{11},...,\psi_{pp}; Y)$ depend seulement de (ϕ,ψ_{ii},S_i)*
   (On utilise $S=S_p$ pour représenter la matrice de covariance résiduelle du VAR $S(\Phi)$ et S_i la partie supérieure gauche i par i block de $S(\Phi)$). Alors,

$$(\psi_{i-1,i}|\psi_{ii}; S_i) \propto \mathcal{N}(h_i(S_{i-1} + \Omega_{i-1})^{-1}) \tag{14}$$

   avec

$$h_i = -\psi_{ii}(S_{i-1} + \Omega_{i-1})^{-1} S_{i-1,i} \tag{15}$$

2. *Ayant (Φ,Y), les lois a posteriori conditionelles de $\psi_{11}^2,...,\psi_{pp}^2$ sont indépendantes et*

$$(\psi_{ii}^2|\phi; Y) \sim gamma(a_i + \frac{1}{2}T; B_i) \tag{16}$$

   avec

$$B_i = \begin{cases} b_1 + \frac{1}{2}s_{11}, & si\ i = 1 \\ b_i + \frac{1}{2}(s_{ii} - s'_{i-1,i}(S_{i-1} + \Omega_{i-1})^{-1} s_{i-1,i}), & si\ i = 2,...,p \end{cases} \tag{17}$$

Paramétrisation Sur-identifiée (Over-identified)

Cette loi a priori proposée précedemment a la caractéristique souhaitée. Les lois a porteriori de la matrice Ψ peuvent être déduites en même temps grace au lemme suivant:

Lemme 3.2.1. Consédirons un vecteur $\eta = (\eta_1,...,\eta_n)'$ avec m de ces éléments à savoir $\eta_{k_1},...,\eta_{k_n}$ sont nuls. Soit $\Theta = (\theta_{ij})$ une matrice symétrique $n \times n$. Définissons $\widetilde{I}_{nm}$ comme le reste de la matrice I_n aprés suppression des m lignes correspondant aux entrées nulles dans η. On note le vecteur $\widetilde{\eta}$ de dimension (nxm)

$$\widetilde{\eta} = \widetilde{I}_{nm}\eta$$

et la matrice $\widetilde{\Theta}$ de dimension $(n-m) \times (n-m)$

$$\widetilde{\Theta} = \widetilde{I}_{nm}\Theta\widetilde{I}'_{nm}$$

Alors

$$\eta'\Theta\eta = \widetilde{\eta}'\widehat{\Theta}\widetilde{\eta}$$

Comme dans le cas précedent, les auteurs ont choisi des lois a priori indépendantes, à savoir (7) pour Φ et (9) pour $(\psi_{11},...,\psi_{pp})$. De plus ils supposent que la loi a priori normale indépendante pour $\widetilde{\psi}_{i-1,i}$ est

$$\widetilde{\psi}_{i-1,i} \sim^{indep} \mathcal{N}_{m_{i-1}}(0,\widetilde{\Omega}_{i-1}) \tag{18}$$

avec i=2,...p Alors, on a le théorème suivant

Théorème 3.2.3.

1. *La distribution a posteriori conditionnelle de ϕ donnée par $(\Psi_p; Y)$ est la même qu'en (11)*
2. *Les lois a posteriori conditionnelles de $\widetilde{\psi}_{i-1,i}$ i=2,...,p sont indépendantes et $(\widetilde{\psi}_{i-1,i}|\phi,\psi_{11},...,\psi_{pp};$ dépendent seulement de $(\phi,\psi_{ii}; S_i)$:*

$$(\widetilde{\psi}_{i-1,i}|\psi_{ii},S_i) \sim \mathcal{N}(\widetilde{h}_i(\widetilde{S}_{i-1} + \widetilde{\Omega}_{i-1})^{-1}) \tag{19}$$

avec

$$\widetilde{h}_i = -\psi_{ii}(\widetilde{S}_{i-1} + \widetilde{\Omega}_{i-1})^{-1}\widetilde{s}_{i-1,i} \tag{20}$$

3. *Les lois a posteriori conditionnelles de $\psi_{11}^2,...,\psi_{pp}^2$ sont indépendantes et $(\psi_{ii}^2|\phi; Y)$ suivent une gamma $(a_i + \frac{1}{2}T,\widetilde{B}_i)$ où*

$$\widetilde{B}_i = \begin{cases} b_1 + \frac{1}{2}s_{11}, & si\ i = 1 \\ b_i + \frac{1}{2}(s_{ii} - \widetilde{s}'_{i-1,i}(\widetilde{S}_{i-1} + \widetilde{\Omega}_{i-1})^{-1}\widetilde{s}_{i-1,i}), & si\ i = 2,...,p \end{cases} \tag{21}$$

Aprés l'utilisation des théorémes précedents et du lemme, ils proposent une loi a priori sur le VAR restreint et déduisent des résultats sur la loi a posteriori.

Ils ont employé l'algorithme qui suit pour calculer la loi a posteriori (11) pour la loi a priori proposée.

Algorithme

1^{ere} étape: Simuler Σ_k

1. Calculer $S = S(\widehat{\phi}_{k-1})$
2. Pour i=1,...,p Simuler $\xi_i \sim^{indep} \Gamma(a_i + \frac{1}{2}T, B_i)$ où est donné en (17)
3. Simuler les éléments extra-diagonaux de $\Psi_{i-1,i}$ (14)
4. Calculer $\Sigma_k = \Psi_{pk}^{-1'}\Psi_{pk}^{-1}$ où Ψ_{pk} matrice triangulaire supérieure avec les $\Psi_{ii} = \sqrt{\xi_i}$ et les premiers éléments i-1 de la i^{ieme} colonne $\Psi_{i-1,i}$ (1.c)

2^{eme} étape: Simuler $\phi_k \sim \mathcal{N}(\widehat{\phi}_k, \widehat{\Xi}_k)$ où

$$\Xi_k = [(\Psi_{pk}^{'}\Psi_{pk}) \otimes (X^{'}X) + \Xi_0^{-1}]^{-1}$$

$$\widehat{\phi}_k = \widehat{\Xi}_k[(\Psi_{pk}^{'}\Psi_{pk}) \otimes (X^{'}X)\widehat{\phi}_M + \Xi_0^{-1}\phi_0]$$

3.3 Test Bayésien

Dans cette section, on présente une discussion du test d'hypothèse Bayésien. La méthode choisie est basée sur le facteur de Bayes du modèle 1 contre le modèle 2, lequel est le rapport de vraisemblances marginales

$$B_{12} = \frac{m^1(Y)}{m^2(Y)}$$

3.3.1 Estimation de la vraisemblance marginale

Dans cette partie, ils ont estimé la vraisemblance marginale pour le VAR identifié. Ils supposent que le modèle 1 est le VAR identifié simplement. Sous les lois a priori (7) et(9), bien que les distributions a posteriori conditionnelles soient standards, la loi a posteriori conjointe ne l'est pas.
En fait, la loi a posteriori conjointe de (ϕ,Ψ) est proportionnelle à

$$l(\phi,\Sigma)exp[-\frac{1}{2}(\phi - \phi_0)^{'}\Xi_0^1(\phi - \phi_0)]\pi(\Psi)$$

avec $l(\phi,\Sigma)$ est donné en (10) et $\pi(\Psi)$ est la densité a priori de Ψ. La loi a posteriori marginale de Ψ est proportionnelle à

$$\pi(\Psi)|\Psi|^T|\Psi\Psi^{'}\otimes(X^{'}X)+\Xi_0^{-1}|^{\frac{T}{2}}exp\{-\frac{1}{2}tr\{(\Psi\Psi^{'}\otimes(X^{'}X))^{-1}+\Xi_0^{-1}\}^{-1}(\widehat{\phi}_M-\phi_0)(\widehat{\phi}_M-\phi_0)^{'}\}.$$

La loi a posteriori ci-dessous de Ψ ne suit pas une distribution standard.

Pour estimer cette vraisemblance, les auteurs ont utilisé la méthode proposée par Chib (1995),qui calcule la vraisemblance sans passer par une intégration numérique intensive.

Comme suggéré par Chib (1995), prendre le logarithme de l'identité de la règle de Bayes conduit à

$$\log\{m(Y)\} = \log\{L(Y/\theta^*)\} + \log\{\pi(\theta^*)\} - \log\{\pi(\theta^*/Y)\}$$

avec m(Y) la vraisemblance marginale, $\pi(\theta)$ la loi a priori et $\pi(\theta/Y)$ la loi a posteriori. Dans ce contexte, $\theta = (\phi,\Psi)$. Ils ont supposé ne pas connaître la constante de normalisation de la loi a posteriori $\pi(\theta/Y)$ mais ils connaissent (Ψ/Y) et la loi a posteriori conditionnelle $(\phi/\Psi;Y)$. Ils ont pu alors réecrire le logarithme de la loi a posteriori conjointe de (ϕ^*,Ψ^*) comme suit

$$\log\{\pi(\phi^*,\Psi^*|Y)\} = \log\{\pi(\phi^*|\Psi^*,Y)\} + \log\{\pi(\Psi^*|Y)\} \tag{22}$$

Le premier terme a droite de (22) est simple car la loi a posteriori conditionnelle $\pi(\phi|\Psi^*,Y)$ est normale. Quant au second terme a gauche, il n'est pas standard mais on peut le calculer en utilisant la rélation suivante:

$$\pi(\Psi^*|Y) = \int \pi(\Psi^*|\phi,Y)\pi(\phi|Y)d\phi.$$

Le calcul pour le modèle alternatif (Sur-identifié) est fait de façon analogue.

3.3.2 Comparaison du facteur de Bayes au critère de Schwarz

Le critère de Schwarz est très largement utilisé par les chercheurs dans les problèmes d'identification de modèles, en particulier, quand le facteur de Bayes est difficile à obtenir. Kass et Raftery (1995) l'expliquent d'ailleurs très bien. Ce qui constitue une différence entre ces deux critères est que le facteur de Bayes est un rapport de deux vraisemblances moyennes pondérées par les probabilités a priori et que le critère de Schwarz est plutôt basé sur le rapport des vraisemblances maximales.
Supposons que le modèle 1 est sur-identifié (over-identified) et que le MLE est $\theta_1^\star$. On suppose que les q éléments extra-diagonaux de Ψ sont nuls et que le modèle 2 (de MLE $\theta_2^\star$)est simplement identifié (just-identified).Alors, le critère de Schwarz est donné par

$$S_{12} = log\left\{l(Y/\theta_1^\star)\right\} - log\left\{l(Y/\theta_2^\star)\right\} + \frac{q}{2}log(T)$$

Si la taille de l'échantillon T est assez grande, le critère de Schwarz S_{12} approxime bien le logarithme du facteur de Bayes B_{12} en ce sens que

$$\frac{S_{12}}{\log B_{12}}_{T\to\infty} \to 1$$

Le critère AIC est souvent utilisé comme une alternative du critère de Schwarz. Pour le cas ci-dessus, l'AIC est alors

$$AIC_{12} = log\left\{l(Y/\theta_1^\star)\right\} - log\left\{l(Y/\theta_2^\star)\right\} + q$$

Dans ce qui suit, les performances du facteur de Bayes et du critère de Schwarz sont présentées et comparées dans le cas des échantillons finis.

3.4 Exemple numérique

Pour mettre en application tout ce qui a été dit un peu plus haut, Sun et Ni (2002) proposent comme application des exemples numériques des estimations Bayésiennes sur le prior proposé .
Les paramètres de la loi a priori sont comme suit: la moyenne a priori ϕ_0 est supposée prendre sa vraie valeur donnée plus haut. Sa covariance Σ_0 est égale à la matrice identité. La loi a priori du vecteur formé par les éléments extra-diagonaux de Ψ est supposée "dispersées" contrairement aux cas du modèle sur-identifié avec une précision Ω supposée 0.1 fois égale à l'identité $\mathbf{I}$. Les hyperparamètres de la loi a priori gamma des éléments diagonaux de ψ sont $a_i = 0.1$ et $b_i = 0.1$ pour $i = 1,2,\ldots,p$
Exemple 3.1. On considère un VAR avec une seule restriction sur un élément de la matrice Ψ contraint à être égal à zéro:

$$\Psi = \begin{pmatrix} 1 & 0.5 & 05 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & -0.5 & -0.5 \\ -0.5 & 1.25 & 0.25 \\ -0.5 & 0.25 & 1.25 \end{pmatrix}$$

Le VAR généré est d'ordre 1. Le terme constant est supposé égal à zéro. La matrice des coefficients de regression est la matrice identité. Ceci veut dire que le VAR est une marcje aléatoire. En utilisant 1000 repetitions et une chaine de Markov de longueur 10 500, les simulations MCMC convergent rapidement même quand le nombre de répétitions a été réduit à 5000.
Dans les résultats on note l'indice M pour le MLE et CM pour le MLE avec contraintes. Aussi, on note 1 pour la moyenne a posteriori de Ψ avec contraintes (sur-identifiée) et 2 pour le cas sans contraintes (simplement identifié).
Bien que la contrainte soit assignée à Ψ, on donne les moyennes des fréquences sur Σ car la vraisemblance est donnée par $\Sigma = \Psi^{-1'}.\Psi^{-1}$. La probabilité est déterminée par $\Sigma = \Psi^{-1'}\Psi^{-1}$. La moyenne du MLEs de Σ sur les 1000 echantillons est

$$\widehat{\mathbb{E}}_0(\widehat{\Sigma}_M) = \begin{pmatrix} 0.848 & -0.419 & -0.422 \\ -0.419 & 1.052 & 0.209 \\ -0.422 & 0.209 & 1.073 \end{pmatrix}$$

et celle du MLEs contraint est

$$\widehat{\mathbb{E}}_0(\widehat{\Sigma}_{CM}) = \begin{pmatrix} 0.922 & -0.456 & -0.459 \\ -0.456 & 1.144 & 0.228 \\ -0.459 & 0.228 & 1.166 \end{pmatrix}$$

Pour tester le modèle sur-identifié (modèle 1), on considère une alternative dans laquelle les coefficients de regression Φ sont les mêmes que plus haut mais la matrice Ψ est identifiée

simplement. Les fréquences moyennes de la loi a posteriori sous les deux modèles sont

$$\widehat{\mathbb{E}}_0(\widehat{\Sigma}_1) = \begin{pmatrix} 0.965 & -0.477 & -0.480 \\ -0.477 & 1.137 & 0.238 \\ -0.480 & 0.280 & 1.160 \end{pmatrix}$$

et celle du MLEs contraint est

$$\widehat{\mathbb{E}}_0(\widehat{\Sigma}_2) = \begin{pmatrix} 0.964 & -0.476 & -0.480 \\ -0.476 & 1.216 & 0.239 \\ -0.480 & 0.239 & 1.263 \end{pmatrix}$$

Les fonctions coût choisies sont l'entropie pour Σ et la quadratique pour Φ. Ils trouvent alors que le MLE sous contrainte domine le MLE sans contraintes.
Pour le test d'hypothèse, 1000 echantillons sont générés et le facteur de Bayes est calculé en utilisant la méthode suggérée par Chib (1995). Les auteurs ont calculé la vraisemblance marginale, la loi a priori et la loi a posteriori du MLE. Est calculé aussi, le facteur de Bayes B_{12} avec le modèle 1 servant les données générant le modèle VAR sur-identifié et le modèle 2 identifié simplement.
Un facteur de bayes plus grand que 1 ou un logarithme du facteur de Bayes positif suggère que les données sont en faveur du modèle1 contre le modèle2.

Les resultats du test basé sur le facteur de Bayes sont comparés avec d'autre basés sur le critère de Schwarz. Le logarithme du facteur de bayes est positif dans 980 sur 1000. Mais, il est plus grand que 3 dans 265 echantillons; tandis que le critère de Schwarz est positif dans 878 sur 1000 echantillons et aucun n'est plus grand que 3.
Pour examiner la performance des méthodes testées, ils ont simulé 1000 autres échantilons en utilisant cette fois le modèle 2 identifié simplement comme données de génération du modèle, avec Ψ definie comme suit:

$$\Psi = \begin{pmatrix} 1 & 0.5 & 05 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Dans 882 sur 1000 echantillons, le logarithme du facteur de Bayes B_{12} est négatif et 684 d'entre eux sont plus petit que -3, supportant le Modèle 2 contre le Modèle 1. Le critère de Schwarz est négatif dans 900 echantillons et est plus petit que -3 dans 608 echantillons. Quand le modèle est identifié simplement est le modèle des données générées, les amplitudes du facteur de Bayes sont plus grandes comparées à celles qui correspondent au ca où c'est l'autre modèle sur-identifié qui représente le modèle des données générées.

Avec le modèle 1 ou le 2 comme modèles de données générées ayant l'autre comme modèle alternatif, les test basés sur le facteur de bayes ne sont pas trés décisifs. Ceci peut

se produire car dans l'exemple précedent la différence entre le MLE contraint, le MLE sans contrainte et les estimations bayésiennes n'est pas substancielle, puiqu'il n'y a que trois variables dans le VAR et une seule contrainte.
Dans l'exemple qui va suivre, ils ont considéré un VAR avec six variables et 10 contraintes sur la matrice Ψ.

Exemple 3.2. Considérons un VAR avec six variables.

$$\Psi = \begin{pmatrix} 1 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & -0.5 & -0.5 & -0.5 & -0.5 & -0.5 \\ -0.5 & 1.25 & 0.25 & 0.25 & 0.25 & 0.25 \\ -0.5 & 0.25 & 1.25 & 0.25 & 0.25 & 0.25 \\ -0.5 & 0.25 & 0.25 & 1.25 & 0.25 & 0.25 \\ -0.5 & 0.25 & 0.25 & 0.25 & 1.25 & 0.25 \\ -0.5 & 0.25 & 0.25 & 0.25 & 0.25 & 1.25 \end{pmatrix}$$

Aprés avoir suivi la même démarche que dans l'exemple précédent, les auteurs trouvent que : L'hypothèse testée basée sur le facteur de Bayes montre que pour 880 sur 1000 echantillons, le logarithme du facteur de Bayes est plus petit que -3, indiquant que les données sont en faveur du modèle 2.

La performance relative du facteur de Bayes contre le critère de Schwarz depend de la taille de l'échantillon. Quand la taille de l'échantillon varie de 30 à 50, le facteur de Bayes devient très différent du critère de Schwarz. Avec un échantillon reduit, le critère de Schwarz devient moins précis que le logarithme du facteur de Bayes et moins efficace quand au choix du modèle adéquat. En résumé, ces exemples numériques montrent que la performance relative des tests dépends de la taille de l'échantillon et de la similitude des modèles en compétition.

3.5 Application économique

Dans ce qui suit, ils ont reproduis une application proposée sur données réelles portant sur la croissance de l'emploi de 3 états et dans 3 secteurs aux états unis d'amérique. Les données sont fournies par la Banque fédérale de Saint Louis et données en pourcentages. Elles sont mesuelles allant de janvier 1982 à décembre 2012 et contiennent 252 observations. L'ordre du VAR est égal à 1. Ainsi la matrice de corrélation est calculée pour observer la nature des corrélations dans la croissance de l'emploi. Six modèles sont en compétition. Le premier (Modèle 1) correspond aux effets macro économiques et aux effets sectoriels. Le second (Modèle 2) inclue les effets internes aux états. Le troisième (Modèle 3) ajoute au modèle 2 les effets sectoriels. Le quatrième (Modèle 4) ajoute au modèle 3 les effets intersectoriels. Le modèle 5 est obtenu en ajoutant au quatrième les

effets intersectoriels nationaux. Enfin, le sixième (Modèle 6) inclue les effets industriels inter-états. L'analyse diffère de celle qui existe dans la littérature sur deux aspects:

- D'abord on utilise les VARs identifiés pour désagréger les données de la croissance de l'emploi dans trois secteurs (manufacture, construction et services) et dans trois états (Illinois, Missouri et Arkansas). Le modèle est moins restrictif que celui utilisé dans la littérature car il permette à chacune des séries d'être affectées.
- Ensuite l'approche Bayésienne est appliquée comme développée dans l'article de ces auteurs décrit plus haut pour l'inférence et le test d'hypothèse.

Le facteur de Bayes et les critères de Schwarz et AIC pour les modèles 1 à 5 contre le modèle 6. Les résulats sont la table suivante

test	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$
$logB_{i6}$	-211.43	-202.61	-188.87	-181.75	-2.57
S_{i6}	56.64	56.92	73.20	46.68	8.48
AIC_{i6}	-40.42	-20.73	11.43	4.33	-2.11

Le facteur de Bayes recommande le modèle 6 identifié simplement. Par contre, le critère de Schwarz suggère que tous les modèles sur-identifiés sont les meilleurs et recommande le modèle 3. Quant au critère AIC, il suggère de favoriser les modèles plus larges que ceux recommandés par le critère de Schwarz. Cette étude montre que même si le critère de Schwarz converge asymptotiquement vers le logarithme du facteur de bayes, il reste que ce dernier est plus approprié pour les échantillons finis comme l'ont suggéré les exemples numériques ci-dessus.

Conclusion générale

"L'important n'est pas de dire tout ce que l'on sait
mais plutôt de savoir tout ce que l'on dit."
Ecrivain contemporain

L'objectif de ce mémoire a été de présenter l'inférence Bayésienne des tests sur les modèles VAR. Pour ce faire, nous nous sommes, tout au long de ce travail, consacrés aux modèles VAR moyennant l'approche Bayésienne. Nous avons mis l'accent sur les derniers développements de recherche portant sur l'inférence Bayeseinne dans les échantillons finis des modèles VAR.

Les tests éffectués sur les modèles autorégréssifs vectoriels montrent que, quand la taille de l'échantillon est petite, le facteur de Bayes est plus éfficace pour choisir le modèle adéquat que le critère de Schwarz largement utilisé dans la littérature. Le facteur de Bayes reste encore utile mais des tentatives sont toujours en cours pour l'améliorer. C'est en ce

sens que ce travail ouvre des perspectives intéressantes dans cette thématique. On peut, à cet effet, s'inspirer des travaux de Chib et Jeliaznov (2001) portant sur les tests Bayesiens sur VAR restreints ou de l'approche stochastique de George et al (2008). Enfin, l'approche par les fonctions coût autres que l'entropie et quadratique peut être une bonne piste.

Bibliographie

[1] Amadou Moussa Mahaman Laouali:*Analyse des séries chronologiques multivariées et applications* Mémoire de master U.M.M.T.O, 2012.

[2] Anderson T.W: *An introduction to Multivariate Statistical Analysis.* Wiley, 1984.

[3] Belkacem Nadia: *Modélisation d'incertitude appliqués au problème de management de l'eau* U.M.M.T.O 2012.

[4] Box G.E.P., Tiao G.C.: *Bayesian inference in statistical analysis.* Addison-Wesley, (1973).

[5] Brockwell et Davis: *Introduction to time series and forecasting.* Second Edition, Springer, 1991.

[6] Christian P.Robert : *Le choix bayésien.* Springer-Verlag France, Paris, 2006.

[7] Chib,S. (1995).*Marginal likelihood from the Gibbs output.*Journal of the American Statistical Association. 90,1313Ű1321.

[8] Chib,S.,Jeliazkov,I.(2001). Marginal likelihood from the MetropolisŰHastings output. *Journal of the American Statistical Association* 96(453), 270Ű281.

[9] Christophe HURLIN: *Econométrie appliquée: séries temporelles.* Cours tronc commun, UFR Economie appliquée, 2001.

[10] Daniels M.J.,Pourahmadi,M.,2002.*Bayesian analysis of covariance matrices and dynamic models for longitudinal data.* Biometrika 89(3),553Ű566.

[11] Dongchu Sun, Shawn Ni:*Bayesian testing of restrictions on vector autoregressive models.* Journal of Statistical Planning and Inference, 2012.

[12] Eaton, M.L. et Olkin, I. (1987) *Best equivariant estimators of a cholesky decomposition.* Annals of stat. 15, 1639-1650

[13] Eric Parent, Jacques Bernier: *Le raisonnement bayésien :Modélisation et inférence.* Edition springer, 2007.

[14] George, E.,Sun,D.,Ni,S.(2008).Stochastic search models election for restricted VAR models. *Journal of Econometrics* 142,553Ű580.

[15] Guillaume Chevillon: *Pratique des séries temporelles.* OFCE et Université d'Oxford, 2004.

[16] Kass,R.E.,Raftery,A.E. (1995). Bayes factors. *Journal of the American Statistical* Association 90,773Ű795.

[17] Hamilton J.D. *Time series analysis* Princeton University Press, 1994.

[18] Jayanta, K. Ghosh, Mohan Delampady and Tapas Samanta: *An introduction to Bayesian analysis: Theory and methods.* Springer Science and Business Media, LLC, USA, 2006.

[19] Jean-Jaques Boreux, Eric Parent et Jacques Bernier: *Pratique du calcul bayésien.* Springer-Verlag France,Paris, 2010.

[20] Jean-Michel Marin, Christian P. Robert:*Les bases de la statistique bayésienne.* INSEE, Paris

[21] Judith Rousseau. *Statistique Bayésienne: Notes de cours.* ENSAE ParisTech troisieme année 2009-2010.

[22] J.M.Bernardo, A.F.M.Smith: *Bayesian theory.* Editions Wiley series in probability and statistics, 2000.

[23] J.Roudier: *Une application de la théorie de la décision statistique bayésienne.* Revue de statistique appliquées, tome 22, nř4, 1974, p.3-28.

[24] J.Ulmo: *La décision statistique dans le cadre bayésien.* Revue statistique appliquées, tome 19, nř3, 1971, p.27-66.

[25] Larbi.L *Sur la décision statistique dans le contexte Bayésien.* Mémoire de master en statistique, U.M.M.T.O, 2011.

[26] Lütkelpohl, H. *New introduction to multiple time series analysis.* Springer-verlag-Berlin, 2005.

[27] Paul S.P Cowpertwait,Andrew V. Metcalfe: *Introductory Time Series with R* Springer Science+Business Media, LLC 2009.

[28] Rainer von Sachs et Sébastien Van Bellegem: *Séries chronologiques* Belgium, 2005.

[29] Sandrine, L et Valérie, M (2002). Économétrie des séries temporelles macroéconomiques et financières. *Économica. Edts*