For the identification of time-series models. application to arma processes.

Carole Toque

École Doctorale
d'Informatique,
Télécommunications
et Électronique de Paris

TELECOM PARIS
école nationale supérieure des télécommunications

THÈSE

présentée pour obtenir le titre de

DOCTEUR DE L'ÉCOLE NATIONALE SUPÉRIEURE
DES TÉLÉCOMMUNICATIONS

Spécialité : Informatique et Réseaux

soutenue par
Carole Toque

le 13 septembre 2006

POUR L'IDENTIFICATION DE MODÈLES FACTORIELS DE SÉRIES TEMPORELLES

APPLICATION AUX ARMA STATIONNAIRES

Jury

Président	Ludovic Lebart
Directeur	Bernard Burtschy
Rapporteurs	Jean-Pierre Indjehagopian Bernard Goldfarb (contresign. Edwin Diday)
Examinateurs	Thomas Fullerton Mohsen Hamoudia Ludovic Lebart
Invité	Dominique Ladiray

Résumé

Cette thèse est axée sur le problème de l'identification de modèles factoriels de séries temporelles et est à la rencontre des deux domaines de la Statistique, l'analyse des séries temporelles et l'analyse des données avec ses méthodes descriptives.

La première étape de notre travail a pour but d'étendre à plusieurs séries temporelles discrètes, l'étude des composantes principales de Jenkins développée dans les années 70. Notre approche adapte l'analyse en composantes principales « classique » (ou ACP) aux séries temporelles en s'inspirant de la technique Singular Spectrum Analysis (ou SSA). Un principe est déduit et est appliqué au processus multidimensionnel générateur des séries. Une matrice de covariance à structure « remarquable » est construite autour de vecteurs aléatoires décalés : elle exploite la chronologie, la stationnarité et la double dimension du processus. A l'aide de deux corollaires établis par Friedman B. dans les années 50 basés sur le produit tensoriel de matrices, et de propriétés de covariance des processus circulaires, nous approchons les éléments propres de la matrice de covariance. La forme générale des composantes principales de plusieurs séries temporelles est déduite. Dans le cas des processus « indépendants », une propriété des scores est établie et les composantes principales sont des moyennes mobiles des séries temporelles.

A partir des résultats obtenus, une méthodologie est présentée permettant de construire des modèles factoriels de référence sur des ARMA vectoriels « indépendants ». L'objectif est alors de projeter une nouvelle série dans un des modèles graphiques pour son identification et une première estimation de ses paramètres. Le travail s'effectue dans un cadre théorique, puis dans un cadre expérimental en simulant des échantillons de trajectoires AR(1) et MA(1) stationnaires, « indépendantes » et à coefficients symétriques.

Plusieurs ACP, construites sur la matrice temporelle issue de la simulation, produisent de bonnes qualités de représentation des processus qui se regroupent ou s'opposent selon leur type en préservant la propriété des scores et la symétrie dans le comportement des valeurs propres. Mais, ces modèles factoriels reflètent avant tout la variabilité des bruits de la simulation. Directement basées sur les autocorrélations, de nouvelles ACP donnent de meilleurs résultats quels que soient les échantillons. Un premier modèle factoriel de référence est retenu pour des séries à forts coefficients.

La description et la mesure d'éventuels changements structurels conduisent à introduire des oscillateurs, des fréquences et des mesures entropiques. C'est l'approche structurelle. Pour établir une possible non-linéarité entre les nombreux critères et pour augmenter la discrimination entre les séries, une analyse des correspondances multiples suivie d'une classification est élaborée sur les entropies et produit un deuxième modèle de référence avec trois classes de processus dont celle des processus à faibles coefficients.

Ce travail permet également d'en déduire une méthode d'analyse de séries temporelles qui combine à la fois, l'approche par les autocorrélations et l'approche par les entropies, avec une visualisation par des méthodes factorielles. La méthode est appliquée à des trajectoires AR(2) et MA(2) simulées et fournit deux autres modèles factoriels de référence.

Mots-clés : Identification de séries temporelles, processus multidimensionnel, stationnarité, processus circulaire, produit tensoriel, processus ARMA, modèles factoriels de référence, autocorrélation, oscillateur, entropie, analyses factorielles.

Abstract

This thesis is centered on the problem of the identification of time series models with the meeting of two fields of the Statistics, Time Series Analysis and Data Analysis with its descriptive methods.

The first stage of our work is to extend to several discrete time series the Jenkins' principal component study developed in the Seventies. Our approach adapts « classic » Principal Component Analysis (PCA) to time series while taking as a starting point the technique Singular Spectrum Analysis (SSA). A principle is deduced and applied to the multidimensional process generating series. A Toeplitz bloc covariance matrix is built around lagged random vectors: it exploits the chronology, the stationarity and the double dimension of the process. Using two corollaries based on the tensorial product of matrices and established by Friedman B. in the Fifties, like the covariance properties of a circular process, we approach the eigenvalues and the eigenvectors of the covariance matrix. The general shape of the principal components of several time series is deduced. In the case of the « independent » processes, a scores property is established and the principal components become moving averages of time series.

From the obtained results, we propose a methodology allowing to build reference factorial models on « independent » vector ARMA. The objective is then to project a new series in one of the graphic models for its identification and a first estimate of its parameters. We work within a theoretical framework, then within an experimental framework by simulating samples of stationary, « independent » with symmetrical coefficients AR(1) and MA(1) processes.

Based on simulated temporal matrices, several PCA produce good qualities of processes representation, with significant groupings and oppositions preserving the scores property and the eigenvalues symetric behavior. But above all, these factorial models reflect the variability of simulated white noises.

Directly based on autocorrelation matrices, PCA give better results whatever the samples except for some processes said « weak ». A first reference graphic model ensues with identification and estimation.

Description and measure of possible structural changes lead us to introduce oscillators, frequencies and measures of entropy. This is the structural approach. To establish non-linearity between the numerous criteria and to increase the discriminative ability between the series, classifications on MCA are built over measures of entropy and produce outstanding quality of classes' characterization. A second reference graphic model ensues with the class of « weak » processes.

This work also makes it possible to deduce a method of time series analysis which combines the usual approach by autocorrelations and a structural approach, less usual, by analysis of oscillators and theory of information, through visualization by factorial methods. The method is applied to simulated AR(2) and MA(2) processes and provides two more reference factorial models.

Key words: Identification of time series models, multidimensional process, stationarity, circular process, tensorial product, ARMA process, reference factorial models, autocorrelation, oscillators, theory of information, entropy, factorial analyses.

Remerciements

En premier lieu, je remercie vivement le professeur Bernard Burtschy de l'ENST Paris qui est à l'origine de cette thèse et qui a manifesté tout au long de ce travail confiance et intérêt. Nos fréquentes discussions m'ont permis de rebondir et de progresser pour me donner goût à la recherche.

Je tiens à remercier le directeur de recherche Ludovic Lebart de l'ENST Paris pour ses remarques constructives et pour avoir accepté de présider le jury de cette thèse.

Je remercie le professeur Jean-Pierre Indjehagopian de l'ESSEC d'avoir bien voulu être rapporteur de ce travail.
Je suis reconnaissante au directeur d'UFR Bernard Goldfarb de l'Université Paris Dauphine du temps et de l'attention consacrés pour rapporter de cette thèse.
Leurs récentes suggestions devront participer à la poursuite de ce travail de recherche.

Je remercie le département INFRES de l'ENST Paris de m'avoir donné l'opportunité de me rendre à trois congrès internationaux qui se tenaient à Sydney, Minnéapolis et Santander.
J'ai pu alors faire la connaissance du professeur Thomas Fullerton de l'Université du Texas à El Paso avec qui j'ai travaillé pour une première publication américaine dans les « 2005 American Statistical Association Proceedings », et du professeur associé Mohsen Hamoudia de l'ESDES à Lyon qui me propose d'écrire un article dans la revue « International Journal of Forecasting ». Je tiens donc tout particulièrement à remercier ces deux personnes de m'avoir soutenu et d'avoir accepté de faire partie du jury de cette thèse.
J'ai aussi fait la connaissance de l'enseignant-chercheur Heping Pan de l'Université de Ballarat en Australie avec qui je souhaiterai poursuivre mes recherches pour la partie 'Application à des données financières'.

Je remercie également Dominique Ladiray, administrateur de l'INSEE, d'avoir accepté d'être l'invité du jury de cette thèse, et des conseils en matière de présentation orale.

Je remercie enfin les professeurs Nina Golyandina de l'Université de Saint Pétersbourg et Michel Terraza de l'Université de Montpellier pour leurs conseils sur la première partie du mémoire.

Je termine en saluant chaleureusement mes proches pour qui je suis très reconnaissante d'avoir accepté les moments d'indisponibilité avec une attention toute particulière à mes deux filles Delphine et Emilie qui ont été remarquables d'autonomie.

Table des Matières

Introduction

Contexte général du travail

Deux domaines de la statistique, aux histoires longues et souvent parallèles, sont évoqués dans l'intitulé même de notre travail, « Pour l'identification de modèles factoriels de séries temporelles. Application aux ARMA stationnaires ». En effet, il s'agit d'une part de *l'analyse des données* avec la construction de modèles factoriels, et d'autre part de *l'analyse des séries temporelles* avec les incontournables processus ARMA.

Pour l'analyse des données, nous travaillons tout particulièrement avec les méthodes descriptives et de visualisation graphique qui sont les méthodes d'analyse factorielle [Thurstone31], [Hotelling33], [Lebart2000] et les méthodes de classification automatique [Sokal et al.63], [Jardine et al.71], [Benzecri73a] et [Benzecri73b], pour les distinguer des méthodes à caractère décisionnel telles que la régression et la discrimination que nous n'utilisons pas. Pour l'analyse des séries temporelles, nous travaillons le plus souvent avec les processus ARMA stationnaires [Wold38], soit dans le cadre multidimensionnel en nous aidant par exemple des travaux de [Tiao et al.81], [Granger et al.86] et [Hamilton94], soit dans le cadre unidimensionnel avec la méthodologie de [Box et al.76].

Notre travail de recherche est aussi à la rencontre des deux domaines et est axé sur la *problématique de l'identification de modèles factoriels de séries temporelles*.

Nous rappelons que la fusion de l'analyse des données et de l'analyse des séries temporelles a principalement débuté dans les années 70, avec comme cadre théorique les processus aléatoires générateurs des séries. Ce sont, par exemple, les travaux :

- [Jenkins et al.68] pour une analyse temporelle des composantes principales d'une seule série, dans le domaine du temps discret ;

- [Deville74] pour une analyse harmonique des processus aléatoires qui utilise la décomposition orthogonale d'un processus, et pour le rapprochement avec l'analyse en composantes principales ;

- [Basilevsky79] pour une étude des composantes principales d'une seule série temporelle, dans le domaine du temps continu, en application de la théorie de Karhunen-Loeve ;

- [Brillinger81] pour une analyse spectrale des composantes principales de plusieurs séries temporelles, dans le domaine des fréquences.

Mais, ces travaux restent très difficiles à appliquer dans la plupart des cas pratiques des chroniques à temps discret.

Après quelques années d'abandon, la fusion des deux domaines reprend avec surtout la recherche de modèles d'oscillateurs dans les cadres unidimensionnel et multidimensionnel des séries temporelles. Ce sont les techniques SSA (Singular Spectrum Analysis) et M-SSA (Multivariate SSA) dont l'essentiel se trouve dans l'ouvrage [Golyandina et al.2001]. Ces techniques ont été empruntées aux physiciens pour être appliquées le plus souvent en climatologie et en géophysique. Par exemple, lorsque nous nous plaçons dans le cadre multidimensionnel, les modèles d'oscillateurs étudiés ont une double dimension spatio-temporelle.

Pour la réduction des données et pour la prévision des séries temporelles, nous retenons, d'une part les travaux d'analyse factorielle dynamique avec par exemple [Doz et al.97] dont la méthode passe par la construction de modèles à composantes inobservables, et d'autre part un nombre impressionnant de papiers publiés dans le Data Mining des séries temporelles. Mais, la *chronologie des données* qui constitue la nature même des séries temporelles n'est pas toujours respectée à la différence des techniques SSA ou M-SSA.

Notre développement à côté de ces nombreux travaux est alors le suivant :

- étendre au cas des processus multidimensionnels l'étude de Jenkins citée précédemment, en nous plaçant toujours dans le domaine du temps discret qui est le plus souvent rencontré en pratique ;

- adapter l'analyse en composantes principales (ACP) aux séries temporelles en s'inspirant des techniques SSA et M-SSA allant de pair avec la notion essentielle de variables décalées et respectant ainsi la chronologie des données ;

- prendre en compte les connaissances supplémentaires pour construire des modèles factoriels de séries temporelles.

  C'est ici la première étape de notre travail de recherche qui passe ainsi du contexte d'une population au contexte théorique des processus aléatoires à temps discret en appliquant tout d'abord un principe simple déduit des techniques SSA et M-SSA.

Méthodologie générale

Pour l'identification de modèles factoriels de séries temporelles, nous proposons une méthodologie en quatre étapes :

- établir des propriétés « remarquables » des composantes principales de plusieurs séries temporelles en généralisant au cas multidimensionnel l'étude des composantes principales de Jenkins ;

- appliquer ces propriétés aux ARMA stationnaires, puis « indépendants » ou non corrélés ;

- construire les premiers modèles factoriels basés sur les résultats obtenus et sur des ARMA simulés ;

- améliorer les modèles en s'aidant de l'approche « classique » par les autocorrélations et d'une approche « moins classique » par les entropies, pour établir des modèles factoriels de référence.

Originalité et objectif double

Notre recherche a un objectif double :

- identifier et estimer les paramètres d'une nouvelle série en la projetant dans un des modèles factoriels de référence ;

- en déduire une méthode d'analyse des séries temporelles, c'est à dire une méthode d'identification de modèles de séries temporelles qui combine à la fois une approche par les autocorrélations et une approche par les entropies, avec une visualisation par des méthodes factorielles.

L'originalité de notre travail se situe alors à deux niveaux :

- dans la technique graphique d'identification et de première estimation des paramètres d'une série temporelle par projection dans un modèle factoriel de référence, ce qui la différencie de l'approche bien connue de Box et Jenkins par les corrélogrammes et par la méthode du maximum de vraisemblance ;

- dans l'apprentissage de modèles de séries temporelles par : des techniques de visualisation et les deux approches par les autocorrélations et par les entropies.

Plan de la thèse

Il en découle le plan suivant qui se compose de quatre chapitres complétés par une annexe.

Le chapitre 1 fait l'étude des composantes principales de plusieurs séries temporelles, dans le domaine du temps discret. Il se divise en cinq parties. La première partie rappelle le formalisme des processus multidimensionnels et stationnaires, et leurs propriétés de covariance. La deuxième partie est consacrée à la problématique qui est d'adapter l'ACP aux séries temporelles. Pour cela, nous réalisons la synthèse des techniques SSA et M-SSA dans le contexte d'une population pour en dégager un principe assez « simple » de construction d'une matrice de covariance autour de variables d'états ou de vecteurs d'états. La troisième partie consiste à appliquer ce principe aux processus aléatoires à temps discret, c'est à dire les processus générateurs des séries. Une matrice de covariance à structure « remarquable » est alors construite autour de vecteurs aléatoires et reflète la stationnarité ainsi que la double dimension du processus multidimensionnel. Cette matrice est à structure bloc-Toeplitz. Pour approcher ses éléments propres, nous développons dans la quatrième partie de ce chapitre une méthode qui généralise l'approche de Jenkins faite dans le cas unidimensionnel et nous utilisons des résultats établis par Friedman B. basés sur le produit tensoriel de matrices. Par ailleurs, nous écrivons les composantes principales de plusieurs séries temporelles dans le cas général. Enfin, la cinquième partie précise la forme des composantes principales de plusieurs séries temporelles dans le cas de l'indépendance des processus. Des propriétés des scores sont déduites.

Le chapitre 2 applique les résultats précédents aux ARMA vectoriels et stationnaires, dans le cas général et dans le cas « indépendant ». Il se divise en quatre parties. Les deux premières parties rappellent le formalisme des ARMA vectoriels, leurs propriétés de covariance ainsi que l'expression de leurs fonctions matricielles d'autocovariance. Nous pouvons alors approcher dans la troisième partie de ce chapitre les éléments propres de la matrice de covariance de processus ARMA quelconques et « indépendants ». Dans la quatrième partie, nous précisons la forme des composantes principales de modèles AR(1), MA(1) et ARMA « indépendants ». Nous obtenons des moyennes mobiles pour les composantes principales.

Le chapitre 3 vise à présenter les premiers modèles factoriels de référence ou de projection d'une nouvelle série pour son identification et une première estimation de ses paramètres. La première partie rappelle la méthodologie de Box et Jenkins utilisée pour identifier quelques moyennes mobiles obtenues dans le chapitre 2. La deuxième partie explique les conditions générales et particulières de la simulation de trajectoires AR(1) et MA(1) « indépendantes », stationnaires et à coefficients symétriques. Puis, les troisième et quatrième parties entament la construction des premiers modèles factoriels, pour une seule série avec la technique SSA, et pour

plusieurs séries avec des ACP temporelles sur un ou plusieurs échantillons. Les modèles obtenus sont de bonne qualité mais ne répondent pas encore à notre objectif de projection d'une nouvelle série pour son identification. Dans les cinquième et sixième parties, une approche par les autocorrélations, puis une approche par les entropies pour décrire et mesurer d'éventuels changements structurels, permettent d'améliorer les modèles factoriels. Une ACP basée directement sur les autocorrélations fournit un premier modèle factoriel de référence pour des séries à forts coefficients. Une analyse des correspondances multiples suivie d'une classification basée sur des mesures d'entropies fournit un second modèle de référence pour des séries à faibles coefficients.

Le chapitre 4 présente deux modèles factoriels de référence pour des AR(2) et MA(2) stationnaires qui engendreraient une nouvelle série projetée. Ces modèles sont obtenus en appliquant la méthode qui utilise les deux approches par les autocorrélations et par les entropies, avec des méthodes de visualisation graphique telles que l'ACP et l'analyse des correspondances multiples (ACM) suivie d'une classification.

L'annexe contient les publications d'acte dont celle dans les « JSM 2005 Proceedings » ou publications des « Joint Statistical Meetings » de 2005, ainsi que le projet informatique pour les traitements des séries brutes ou des séries d'états intégrant les calculs des entropies de Pincus et de Shannon.

Chapitre 1

1 Composantes principales de plusieurs séries temporelles

Introduction

Le cadre général de ce chapitre est celui de *l'étude des composantes principales de plusieurs séries temporelles dans le domaine du temps discret*. En effet, il pourrait s'agir d'une analyse des composantes principales dans le domaine des fréquences, mais ce travail a déjà été réalisé dans [Brillinger75] et [Brillinger81], et s'est révélé souvent très complexe à mettre en œuvre dans des cas pratiques. Il ne s'agit pas non plus d'appliquer la théorie de Karhunen-Loeve comme dans [Basilevsky et al.79] et [Burtschy87], d'ailleurs davantage adaptée au cas des séries continues et difficile à étendre au cas de plusieurs séries temporelles.

Pour l'analyse des composantes principales de plusieurs séries temporelles, nous proposons une méthode qui *adapte* la technique de l'analyse en composantes principales à la dimension « espace$\times$temps » des séries. Cette méthode s'inspire essentiellement de la technique « Singular Spectrum Analysis » ou SSA dont l'essentiel se trouve dans [Golyandina et al.2001].

Mais, qu'est-ce qu'une série temporelle ? Le terme « série temporelle » désigne à la fois la série réelle chronologique, et une suite théorique de variables aléatoires indicées par le temps qui va servir à modéliser la première. Dès lors, si nous travaillons avec plusieurs séries temporelles, nous avons recours à la notion essentielle de *processus aléatoire multidimensionnel* ou famille de vecteurs aléatoires indexés par le temps. C'est donc dans la première partie de ce chapitre que nous rappelons la définition d'un processus multidimensionnel à temps discret et ses propriétés de covariance dans le cadre stationnaire.

Comment adapter la technique de l'analyse en composantes principales aux séries temporelles ? C'est dans le contexte d'une population que nous nous plaçons avec tout d'abord la technique de l'analyse en composantes principales ou ACP « classique ». Les principes généraux de la méthode sont rappelés et nous constatons que cette technique qui exploite la structure de la matrice de covariance de k variables aléatoires à valeurs dans $\mathbb{R}^n$, n'est pas adaptée au cas de n séries

temporelles qui sont la réalisation d'un processus multidimensionnel. Deux approches pour adapter la technique de l'analyse en composantes principales aux séries temporelles sont, la technique SSA qui s'applique à une seule série temporelle et la technique M-SSA ou extension de la SSA qui s'applique à plusieurs séries temporelles. Ces deux techniques sont présentées et nous retiendrons le principe assez simple qui est de construire une matrice de covariance autour de variables ou vecteurs d'états qui sont des versions décalées de la série ou des séries temporelles. Cette étude fait l'objet de la deuxième partie de notre chapitre.

Nous travaillons ensuite avec les processus sous-jacents aux séries temporelles. Le principe cité précédemment est tout d'abord appliqué à un processus stationnaire unidimensionnel. Une première matrice de covariance est créée autour de variables aléatoires « décalées ». Puis, la méthode est étendue au cas d'un processus stationnaire multidimensionnel. Il s'agit de concevoir une matrice de covariance qui tienne compte des dépendances entre les diverses composantes des vecteurs aléatoires à des retards différents. Par ailleurs, cette matrice doit exploiter l'hypothèse de stationnarité du processus. Dès lors, les regroupements dans la matrice se font par paquets de jours ou instants successifs à l'image du cas unidimensionnel et à la différence de la technique M-SSA où les regroupements se font par série. La matrice ainsi construite est une matrice à structure bloc-Toeplitz « remarquable » et définit la dimension « temps$\times$espace » des composantes principales du processus. Ces deux matrices de covariance sont présentées dans la troisième partie de ce chapitre.

Dans la plupart des cas, les éléments propres de la matrice de covariance à structure bloc-Toeplitz ne peuvent pas être calculés directement. Pour approcher ces éléments, nous avons recours à la théorie des matrices à structure bloc-circulaire/Toeplitz et aux résultats établis dans [Friedman61] autour du concept de produit tensoriel de deux matrices. La méthode permet d'obtenir « simplement » une approximation des valeurs propres de matrices à structure bloc-Toeplitz et n'utilise pas la théorie complexe des formes Toeplitz que nous pouvons trouver dans [Grenander et al.84] et appliquée par exemple dans [Gray72] et [Gray2000]. La forme des vecteurs propres et des composantes principales des n séries temporelles, est déduite. Dans le cas unidimensionnel, nous retrouvons les résultats établis dans [Jenkins et al.68]. Ce travail fait l'objet de la quatrième partie de ce chapitre.

Le cas des séries « indépendantes » ou non-corrélées est examiné dans la cinquième partie de ce chapitre. Tout d'abord, la méthode qui utilise la matrice de covariance « temps$\times$espace » est appliquée sur les processus sous-jaçents aux séries. Les composantes principales (ou scores) des n séries temporelles coïncident avec les composantes principales (ou scores) de chacune des séries. Puis, une seconde matrice de covariance, « espace$\times$temps » et bloc-diagonale, est construite à l'image de la technique M-SSA. La propriété énoncée précédemment est retrouvée. Dans ce cas, les deux matrices de covariance « temps$\times$espace » et

« espace×temps » sont équivalentes, et la méthode qui utilise la théorie des matrices bloc-circulaires et le concept du produit tensoriel, est confortée. Cette propriété est utilisée dans les chapitres 2 et 3 de notre travail pour la construction de modèles factoriels de séries temporelles.

1.1 Processus multidimensionnel et stationnaire

Dans ce qui suit, une série temporelle $(Z_t,\ t \in \mathbb{Z})$ est la réalisation d'un processus à temps discret, unidimensionnel, noté aussi et sans ambiguïté $(Z_t,\ t \in \mathbb{Z})$. Ce processus est appelé le processus générateur de la série, elle-même appelée chronique ou échantillon ou encore trajectoire du processus aléatoire sous-jacent.

Dans le cas de plusieurs séries temporelles, il s'agit de la réalisation d'un processus multidimensionnel noté $(\underline{Z}_t,\ t \in \mathbb{Z})$ dont nous rappelons la définition et ses propriétés de covariance dans le cadre stationnaire.

Nombreux sont les ouvrages qui traitent des séries temporelles comme des réalisations de processus aléatoires. Nous retiendrons principalement [Box et al. 70], [Hannan70], [Brockwell et al.93], [Gouriéroux et al.83] et [Gouriéroux et al.95].

1.1.1 Processus multidimensionnel

Définition 1. 1

Un **processus multidimensionnel** $(\underline{Z}_t,\ t \in \mathbb{Z})$ est une famille de vecteurs aléatoires indexés par le temps et à valeurs dans $\mathbb{R}^n$.

Ce processus admet donc n composantes $Z_{it}\ (i = 1,\ldots,n)$ ou n processus réels à temps discret. Nous rappelons qu'un processus réel à temps discret est une suite de variables aléatoires réelles indicées par le temps. Dans la suite, nous supposerons que ces diverses composantes sont de carré intégrable.

Pour cela, il est nécessaire et suffisant que :

$$E\left\|\underline{Z}_t\right\|^2 = E\left[\sum_{i=1}^{n} Z_{it}^2\right] < +\infty$$

Il est alors possible de calculer l'espérance de $\underline{Z}_t$, vecteur de taille n, dont les composantes sont les espérances de chacune des coordonnées :

$$E[\underline{Z}_t] = \begin{bmatrix} E[Z_{1t}] \\ \vdots \\ E[Z_{nt}] \end{bmatrix}$$

et aussi de calculer la matrice de variance-covariance de $\underline{Z}_t$, matrice carrée de taille n, dont le terme général est la covariance entre deux coordonnées de $\underline{Z}_t$:

$$V[\underline{Z}_t] = \begin{bmatrix} V[Z_{1t}] & Cov[Z_{1t}, Z_{2t}] & \cdots & Cov[Z_{1t}, Z_{nt}] \\ Cov[Z_{1t}, Z_{2t}] & V[Z_{2t}] & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Z_{1t}, Z_{nt}] & \cdots & \cdots & V[Z_{nt}] \end{bmatrix}$$

1.1.2 Stationnarité

Si de plus, le processus sous-jacent aux n séries temporelles est supposé stationnaire, il vérifie la définition suivante.

Définition 1. 2

Un processus multidimensionnel est dit **stationnaire** d'ordre 2 (ou simplement stationnaire) si :

$$\forall t \in \mathbb{Z},\ E[\underline{Z}_t] = \underline{m} \text{ (indépendant de } t\text{)}$$

$$\forall t, l \in \mathbb{Z},\ Cov(\underline{Z}_t, \underline{Z}_{t+l}) = E(\underline{Z}_t - \underline{m})(\underline{Z}_{t+l} - \underline{m})' = \underline{\Gamma_{(n)}}(l) \text{ (indépendant de } t\text{)}$$

Si le processus $\underline{Z}_t$ est stationnaire, la matrice $V[\underline{Z}_t]$ égale à $\underline{\Gamma_{(n)}}(0)$, ne dépend pas de t.

Si les n composantes Z_{it} $(i = 1, \ldots, n)$ sont « indépendantes » ou non-corrélées et stationnaires, le processus multidimensionnel est stationnaire.

Démonstration :

Pour cela, il suffit de rappeler la définition de processus aléatoires unidimensionnels non-corrélés dits « indépendants ».

Définition 1. 3 : Processus « indépendants »

Deux processus unidimensionnels sont « indépendants » ou non-corrélés si leur fonction de covariance 'croisée' est identiquement nulle.

D'où, pour un processus multidimensionnel à n composantes Z_{it} $(i = 1,\ldots,n)$

$Cov(Z_{it}, Z_{jt+l}) = 0$ $(i \neq j)$ sont indépendantes de t

et, $E[Z_{it}]$ et $Cov(Z_{it}, Z_{it+l})$ indépendants de t pour tout i et pour tout l.

Fin démonstration.

➢ ***Un processus stationnaire multidimensionnel peut donc être caractérisé par ses autocovariances et covariances décalées.***

1.1.3 Propriétés de covariance

Définition 1. 4

Si $(\underline{Z}_t,\ t \in \mathbb{Z})$ est un processus *stationnaire*, sa **matrice de covariance décalée** au retard l ($l \in \mathbb{Z}$) s'écrit :

$$\underline{\Gamma_{(n)}}(l) = \{\gamma_{ij}(l)\} = Cov(\underline{Z}_t, \underline{Z}_{t+l})$$

ou encore

(1.1)
$$\underline{\Gamma_{(n)}}(l) = \{\gamma_{ij}(l)\} = \begin{pmatrix} \gamma_{11}(l) & \cdots & \gamma_{1n}(l) \\ \vdots & \ddots & \vdots \\ \gamma_{n1}(l) & \cdots & \gamma_{nn}(l) \end{pmatrix}$$

avec

$$\gamma_{ij}(l) = Cov(Z_{it}, Z_{jt+l})\ (i, j = 1,\ldots,n)$$

Notons que $\gamma_{ij}(l)$ n'est pas égal à $\gamma_{ji}(l)$ et donc que la matrice $\underline{\Gamma_{(n)}}(l)$ n'est pas symétrique.

Cependant, $\underline{\Gamma_{(n)}}(l)$ vérifie :

Propriété 1. 1

$$\underline{\Gamma_{(n)}}'(l) = \underline{\Gamma_{(n)}}(-l) \quad (l \in \mathbb{Z}) \tag{1.2}$$

Démonstration :

En effet

$$\underline{\Gamma_{(n)}}(-l) = Cov(\underline{Z}_t, \underline{Z}_{t-l}) = Cov(\underline{Z}_{t-l}, \underline{Z}_t)' = \underline{\Gamma_{(n)}}'(l) \quad (l \in \mathbb{Z})$$

Fin démonstration.

L'élément (i,j) de la matrice de covariance décalée est une indication sur la façon dont la série i est influencée par la série j une période de temps l après.

Il suffira par la suite d'analyser la fonction matricielle $\underline{\Gamma_{(n)}}(.)$ pour un retard l ($l = 0,1,2\cdots$).

Cas unidimensionnel :

Nous rappelons que si le processus stationnaire à temps discret est unidimensionnel, la fonction matricielle $\underline{\Gamma_{(n)}}(.)$ est la fonction d'autocovariance théorique $\gamma(.)$ définie sur $\mathbb{Z}$, **paire** et de **type positif**.

avec :

$$E\left[Z_t^{\,2}\right] < \infty, \ \forall t \in \mathbb{Z}$$

$$E\left[Z_t\right] = m, \text{ (indépendant de } t\text{)}$$

$$Cov(Z_t, Z_{t+h}) = \gamma(h), \ \forall t \in \mathbb{Z} \text{ (indépendant de } t\text{)}$$

Nous pouvons trouver les démonstrations dans [Jenkins70].

1.2 Analyses en composantes principales de séries temporelles

C'est dans le contexte d'une population que nous nous plaçons pour présenter les trois techniques qui sont : l'analyse en composantes principales ou ACP « classique », la « Singular Spectrum Analysis » ou SSA et son extension la « Multivariate » SSA ou M-SSA.

Nous constaterons les limites de l'ACP « classique » pour l'analyse des composantes principales de séries temporelles et nous verrons comment la SSA s'adapte à une seule série temporelle avant d'envisager le cas de plusieurs séries.

1.2.1 ACP « classique »

La technique d'analyse en composantes principales peut être présentée de divers points de vue. Nous pouvons trouver des exposés détaillés de la méthode dans les ouvrages de [Morrison76], [Lebart et al.2000], [Jolliffe2002], sans oublier [Pearson01] qui a entrevu les idées essentielles, puis [Hotelling33] à qui nous devons la première publication sur ce sujet.

C'est du point de vue des analystes de données que nous nous plaçons pour faire quelques rappels sur les principes généraux de la méthode.

1.2.1.1 Les principes généraux

Nous nous limitons à l'analyse en composantes principales dans l'espace $\mathbb{R}^p$ muni de la norme euclidienne et du produit scalaire associé.

Soient p variables quantitatives $Z^{(1)}$, $Z^{(2)}$,$\cdots$,$Z^{(p)}$ mesurées sur un ensemble de n individus.

Les données obtenues sont présentées sous la forme d'un tableau ou matrice $\underline{Z}$ à n lignes et p colonnes :

$$\underline{Z} = \begin{bmatrix} z_{1,1} & \cdots & z_{1,j} & \cdots & z_{1,p} \\ \vdots & & & & \vdots \\ z_{i,1} & \cdots & z_{i,j} & \cdots & z_{i,p} \\ \vdots & & & & \vdots \\ z_{n,1} & \cdots & z_{n,j} & \cdots & z_{n,p} \end{bmatrix}$$

où $z_{i,j}$ est la valeur prise par la variable $Z^{(j)}$ pour l'individu Z_i.

Le problème posé consiste à réduire les p variables initiales en un nombre plus petit de q variables « composées », ou facteurs ($q < p$), appelées encore composantes principales.

Il s'agit donc de passer de la matrice des données initiales (n individus x p variables) à une matrice réduite :

$$\underline{F} = \begin{pmatrix} f_{1,1} & \cdots & f_{1,q} \\ \vdots & f_{i,j} & \vdots \\ f_{n,1} & \cdots & f_{n,q} \end{pmatrix}$$

où l'élément général $f_{i,k}$ est la valeur du facteur k pour l'individu i et $\underline{F}$ est la matrice des 'scores'.

Ces facteurs doivent répondre aux deux conditions suivantes :

- linéarité :

$$F^{(k)} = \sum_{j=1}^{p} a_j^{(k)} Z^{(j)}$$

- indépendance :

$$cor(F^{(k)}, F^{(m)}) = 0 \ \ (k \neq m)$$

Ils doivent aussi restituer le maximum de l'information contenue dans le nuage de points, c'est à dire la quantité $I(g)$ ou inertie au point g centre de gravité du nuage.

Le centre de gravité G d'un nuage de points dans $\mathbb{R}^p$ est défini par :

$$G = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

et ses coordonnées sont données par :

$$G_k = \frac{1}{n} \sum_{i=1}^{n} z_{k,i}$$

L'inertie en un point A du nuage de points est la somme des carrés des distances des points Z_i du nuage au point A :

$$I(A) = \frac{1}{n} \sum_{i=1}^{n} (d(Z_i, A))^2$$

En prenant la distance euclidienne classique, le carré de la distance entre deux points est donné par la somme des carrés des différences des coordonnées de ces deux points :

$$d(Z_i, A)^2 = \sum_{k=1}^{p} (z_{k,i} - A_k)^2$$

D'où l'inertie au point G , centre de gravité du nuage :

$$I(G) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{p} (z_{k,i} - G_k)^2$$

Soit $\widehat{\underline{\Gamma}_{(p)}}$ la matrice de variance-covariance empirique des variables $Z^{(1)},\ Z^{(2)},\cdots,Z^{(p)}$ suivante :

$$\widehat{\underline{\Gamma}_{(p)}} = \begin{pmatrix} Var(Z^{(1)}) & Cov(Z^{(1)},Z^{(2)}) & \dots & Cov(Z^{(1)},Z^{(p)}) \\ Cov(Z^{(1)},Z^{(2)}) & Var(Z^{(2)}) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Z^{(1)},Z^{(p)}) & \cdots & \cdots & Var(Z^{(p)}) \end{pmatrix}$$

L'inertie au centre de gravité du nuage de points (ou la variance totale du nuage de points) qui s'écrit aussi :

$$I(G) = \ Var(Z^{(1)}) + Var(Z^{(2)}) + \cdots + Var(Z^{(p)})$$

est donc égale à la trace de la matrice $\widehat{\underline{\Gamma}_{(p)}}$.

Pour déterminer la première composante principale $F^{(1)}$, le problème revient à trouver une droite D, engendrée par un vecteur unitaire $\overrightarrow{V_1}$, telle que cette droite restitue le maximum de l'information initiale.

(Dans ce qui suit, nous supposons chaque variable centrée pour faire coïncider le centre du nuage de points avec l'origine et retrouver la matrice .)

Les projections des n observations sur la droite $D(\overrightarrow{V_1})$ sont données par le produit scalaire $d = \underline{Z}\overrightarrow{V_1}$ et leur inertie est donnée par la somme des carrés de ces projections, c'est à dire la forme quadratique $d'd = \overrightarrow{V_1}' \underline{Z}' \underline{Z} \overrightarrow{V_1}$.

$\overrightarrow{V_1}$ est donc choisi de façon à maximiser $\overrightarrow{V_1}' \underline{Z}' \underline{Z} \overrightarrow{V_1}$ sous la contrainte $\overrightarrow{V_1}.\overrightarrow{V_1}' = 1$.

Il s'agit là d'un problème classique d'optimisation sous contrainte résolu par la

méthode de Lagrange.

Pour cela, nous formons le Lagrangien $L(\lambda,\overrightarrow{V_1}) = \overrightarrow{V_1}^{'}\underline{Z}^{'}\underline{Z}\overrightarrow{V_1} - \lambda(\overrightarrow{V_1}.\overrightarrow{V_1}^{'} - 1)$ dont les dérivées partielles doivent s'annuler.

En dérivant par rapport à chacune des p composantes du vecteur $\overrightarrow{V_1}$ ainsi que par rapport au multiplicateur de Lagrange λ, et en posant les dérivées partielles égales à zéro, nous obtenons $2\left[\underline{Z}^{'}\underline{Z}\overrightarrow{V_1} - \lambda\overrightarrow{V_1}\right] = 0$ ou encore

$$\underline{Z}^{'}\underline{Z}\overrightarrow{V_1} = \lambda\overrightarrow{V_1} \text{ et } \overrightarrow{V_1}.\overrightarrow{V_1}^{'} = 1\,.$$

C'est l'équation des vecteurs propres et des valeurs propres de la matrice $\underline{Z}^{'}\underline{Z}$. Cette matrice est à un facteur $\frac{1}{n}$ ou $\frac{1}{n-1}$ près la matrice de variance-covariance $\underline{\widehat{\Gamma_{(p)}}}$, ou des corrélations si les variables sont en plus réduites.

$\overrightarrow{V_1}$ est donc le 'premier' vecteur propre de la matrice $\underline{\widehat{\Gamma_{(p)}}}$ correspondant à la plus grande valeur propre λ_1, et la première composante principale vérifie $F^{(1)} = \underline{Z}\overrightarrow{V_1}$.

Puis, la seconde composante principale $F^{(2)}$, orthogonale à la première, doit restituer le maximum de la variance, c'est à dire qu'il faut trouver le vecteur $\overrightarrow{V_2}$ tel que $F^{(2)} = \underline{Z}\overrightarrow{V_2}$ avec $\overrightarrow{V_2}.\overrightarrow{V_2}^{'} = 1$ et $F^{(1)}F^{(2)} = 0$. Il s'ensuit que $\overrightarrow{V_2}$ est le second vecteur propre de la matrice $\underline{\widehat{\Gamma_{(p)}}}$ pour la 'deuxième plus grande' valeur propre λ_2. Et ainsi de suite …

1.2.1.2 Quelques propriétés

En pratique, la matrice symétrique $\underline{\widehat{\Gamma_{(p)}}}$ de variance-covariance autour des p variables $Z^{(1)}$, $Z^{(2)},\cdots,Z^{(p)}$ est diagonalisée pour en déduire les valeurs propres réelles et les vecteurs propres associés qui sont orthogonaux deux à deux.

Propriété 1. 2

Les composantes des vecteurs propres $\overrightarrow{V_k}$ ($k = 1,\ldots,q$) de la matrice $\underline{\widehat{\Gamma_{(p)}}}$ sont alors les coefficients générateurs des q variables composées $F^{(k)}$ appelées aussi les q premières composantes principales des variables $Z^{(j)}$ et vérifient :

$$F^{(k)} = \sum_{j=1}^{p} a_j^{(k)} Z^{(j)} \qquad (1.3)$$

avec comme coefficients $f_{i,k}$ pour $1 \leq i \leq n$:

$$f_{i,k} = \sum_{j=1}^{p} a_j^{(k)} z_{i,j} \qquad (1.4)$$

Propriété 1. 3

Les vecteurs $\overrightarrow{V_k}$ $(k = 1,\ldots,q)$ sont portés par les axes principaux d'inertie du nuage des observations.

Propriété 1. 4

La valeur propre λ_k correspondant à la $k^{ième}$ composante principale $F^{(k)}$ représente la variance des projections du nuage de points sur $\overrightarrow{V_k}$ $(V(F^{(k)}) = \lambda_k)$.

➢ ***L'approche par l'ACP « classique » exploite la structure de la matrice de covariance autour de*** p ***variables aléatoires à valeurs dans*** $\mathbb{R}^n$ ***.***

1.2.2 Techniques SSA et M-SSA

Supposons que les données analysées par l'ACP « classique » soient des données provenant de n séries temporelles de la forme $\underline{Z} = (z_{i,j})$ avec $1 \leq i \leq n$ et $1 \leq j \leq p_t$, i représentant le numéro de la série temporelle et j le temps.

La matrice de covariance empirique $\widehat{\underline{\Gamma_{(p_t)}}}$ définit alors la **dimension temporelle** des composantes principales des n séries temporelles et la relation (1.4) s'écrit pour $1 \leq i \leq n$:

$$f_{i,k} = \sum_{j=1}^{p_t} a_j^{(k)} z_{i,j} \ (k = 1,\cdots,q_t)$$

Mais, dans l'hypothèse où les séries sont la réalisation d'un processus multidimensionnel ($\underline{Z}_t$, $t \in \mathbb{N}^*$ borné), la structure de la matrice de covariance n'est pas adaptée car elle ignore les dépendances entre les diverses composantes du

vecteur aléatoire à différents retards.

Il s'agit alors **d'adapter** la technique d'analyse en composantes principales à des séries temporelles considérées comme un processus multidimensionnel. Une première approche est la technique SSA et elle s'applique à une **seule** série temporelle.

1.2.2.1 La technique SSA

La technique SSA est aussi une méthode d'analyse de données. Cependant, dans le contexte d'une population et avec seulement une série de données, comment est-il possible de réaliser une analyse multivariée ? La technique SSA propose une solution.

Les travaux sur le sujet sont nombreux. Nous retiendrons principalement ceux de [Fraedrich86], [Broomhead et al.86a], [Vautard et al.89], [Vautard et al.92], [Elsner et al.96] et plus récemment [Golyandina et al.2001].

En pratique, la méthode est souvent utilisée en géophysique et en climatologie avec, par exemple, les études de [Ghil et al.91a], [Ghil et al.91b] et [Rasmusson et al.90] sur précisément « El Niño-Southern Oscillation index ». Bien plus que la réduction des données, la recherche de modèles d'oscillateurs à différentes échelles de temps est le but essentiel de ces travaux.

Nous rappelons ci-dessous les principes généraux de la technique.

Soit $(z_i,\ 1 \le i \le p_t)$ la seule série de données temporelles dont nous disposons.

La méthode SSA propose de transformer la suite de données de longueur p_t en une matrice de dimension $(n \times q_t)$ dont la $i^{ème}$ ligne s'écrit pour $i = 1,2,\cdots n$:

$$\widetilde{Z}_i = (z_i, z_{i+1}, \cdots, z_{i+q_t-1})$$

où

$$n = p_t - q_t + 1$$

Le paramètre q_t est appelé la « largeur de la fenêtre ». Il n'existe pas à priori de meilleur choix de q_t. Par exemple, dans [Elsner et al.96], (section 5.2), le meilleur choix de q_t se fait à postériori en comparant les spectres des valeurs propres des matrices trajectoires pour différentes « largeurs de fenêtre », et correspond à $q_t = p_t / 4$.

Soit $\underline{\widetilde{Z}}$ la matrice ainsi construite, elle est dite matrice « trajectoire » et s'écrit :

$$\underline{\widetilde{Z}} = \begin{pmatrix} z_1 & z_2 & z_3 & \cdots & z_{q_t} \\ z_2 & z_3 & z_4 & \cdots & z_{q_t+1} \\ z_3 & z_4 & z_5 & \cdots & z_{q_t+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_n & z_{n+1} & z_{n+2} & \cdots & z_{p_t} \end{pmatrix}$$

La **variable d'états** $\widetilde{Z}_i$ définit alors q_t variables $Z^{(1)}, Z^{(2)}, \cdots, Z^{(q_t)}$ qui sont des versions décalées de la série temporelle.

Une matrice de covariance empirique, notée $\underline{S_{q_t}}$, est calculée autour des q_t variables et détermine **la dimension temporelle** des composantes principales de la série.

En référence à (1.4), les coefficients $f_{i,k}$ des composantes principales s'écrivent :

pour $i = 1,2,\cdots n$ et $n = p_t - q_t + 1$

$$f_{i,k} = \sum_{j=1}^{q_t} a_j^{(k)} z_{i+j-1} \quad (k = 1,\cdots,q_t)$$

où $a_j^{(k)}$ sont les coefficients des vecteurs propres $\overrightarrow{V_k}$ q_t-dimensionnels de la matrice de covariance $\underline{S_{q_t}}$.

➢ ***La SSA est une ACP « classique » avec comme variables*** $Z^{(1)}, Z^{(2)}, \cdots, Z^{(p)}$ ***des versions décalées de la série temporelle.***

1.2.2.2 La technique M-SSA : une extension de la SSA

Nous pouvons trouver une extension de la technique SSA appliquée à plusieurs séries temporelles dans les travaux de [Broomhead et al.86b], [Kimoto et al.91] et [Plaut et al.94]. De même, c'est en climatologie que la méthode est le plus souvent utilisée dans le but de détecter des oscillations spatio-temporelles de champs climatiques à différentes échelles de temps.

Nous rappelons les résultats principaux de la méthode.

Soient n séries temporelles, chacune de longueur p_t. Les données sont alors rangées dans une matrice de dimension $(n \times p_t)$, notée aussi $\underline{Z} = (z_{i,j})$, avec $1 \leq i \leq n$ et $1 \leq j \leq p_t$.

La technique M-SSA consiste à créer une matrice plus grande $\widetilde{\underline{Z}}$ de dimension $(n^{'} \times p_t^{'})$ dont la $i^{ème}$ ligne s'écrit pour $i = 1,2,\cdots n^{'}$:

$$\widetilde{Z}_i = (z_{1,i}, z_{1,i+1}, \cdots, z_{1,i+m_t-1}, z_{2,i}, z_{2,i+1}, \cdots, z_{2,i+m_t-1}, \cdots, z_{n,i}, z_{n,i+1}, \cdots, z_{n,i+m_t-1})$$

avec

$$n^{'} = p_t - m_t + 1 \text{ et } p_t^{'} = n \times m_t .$$

Le vecteur d'états $\widetilde{Z}_i$ $n \times m_t$-dimensionnel, est alors constitué de n variables d'états m_t-dimensionnelles, avec m_t la largeur de la fenêtre qui a le même rôle que q_t dans la technique SSA.

La matrice de covariance, notée $\widetilde{\underline{S_{nm_t}}}$, construite autour de ce vecteur d'états, a la forme suivante :

(1.5)
$$\widetilde{\underline{S_{nm_t}}} = \begin{pmatrix} \underline{S_{m_t}^{(11)}} & \underline{S_{m_t}^{(12)}} & \cdots & \underline{S_{m_t}^{(1n)}} \\ \underline{S_{m_t}^{(21)}} & \underline{S_{m_t}^{(22)}} & \cdots & \underline{S_{m_t}^{(2n)}} \\ \vdots & \vdots & \ddots & \\ \underline{S_{m_t}^{(n1)}} & \underline{S_{m_t}^{(n2)}} & \cdots & \underline{S_{m_t}^{(nn)}} \end{pmatrix}$$

où les regroupements se font par « croisement » de séries, sur une largeur de fenêtre égale à m_t.

$\underline{S_{m_t}^{(kk)}}$ est la matrice carrée d'ordre m_t, de covariance à différents retards de la $k^{ième}$ série. Cette matrice $\underline{S_{m_t}^{(kk)}}$ a la même structure que la matrice de covariance de la $k^{ième}$ série dans une SSA.

Posons $\widetilde{Z}_i = (\widetilde{z_{i,kl}},\ 1 \leq i \leq n^{'})$ avec $\widetilde{z_{i,kl}} = (z_{l,i+k-1},\ 1 \leq l \leq n$ et $1 \leq k \leq m_t)$.

Par analogie au cas SSA, il vient que :

$$f_{i,kl} = \sum_{j=1}^{n \times m_t} a_j^{(kl)} \widetilde{z_{i,kl}}$$

où $a_j^{(kl)}$ sont les coefficients des vecteurs propres $\overrightarrow{V_{kl}}$ $(n \times m_t)$-dimensionnels, de la matrice de covariance $\underbrace{\widehat{S_{nm_t}}}$.

Les composantes principales des n séries temporelles ont dans ce cas **une dimension (espace $\times$ temps)**.

Si $m_t = 1$ alors la M-SSA est une ACP « classique » et si $n = 1$ alors la M-SSA est une SSA.

1.2.2.3 Un principe « simple »

➢ ***Nous retiendrons des techniques SSA et M-SSA le principe assez simple qui est de construire une matrice de covariance autour de variables d'états ou de vecteurs d'états qui sont des versions décalées d'une ou plusieurs séries temporelles.***

1.3 Matrices de covariance de processus stationnaires – Application du principe

Nous travaillions précédemment dans le contexte d'une population. Désormais, nous supposons connu et stationnaire le processus aléatoire unidimensionnel ou multidimensionnel sous-jacent à la série ou aux séries temporelles.

Dés lors, appliquer le principe énoncé ci-dessus à un processus aléatoire à temps discret revient à construire des matrices de covariance autour de variables aléatoires décalées dans le cas unidimensionnel et de vecteurs aléatoires décalés dans le cas multidimensionnel. Par exemple, une réalisation à un instant t des variables aléatoires décalées donne simplement les variables d'états définies dans la technique SSA.

Les matrices de covariance ainsi construites sont soit à structure simple Toeplitz dans le cas unidimensionnel, soit à structure bloc-Toeplitz dans le cas multidimensionnel.

La structure bloc-Toeplitz des matrices est essentielle dans notre travail, nous rappelons d'emblée sa définition.

1.3.1 Matrices Toeplitz à structure simple ou à structure bloc

Définition 1. 5

Une matrice carrée $\underline{\widehat{T_{(pn)}}}$ $(pn)\times(pn)$, est dite à structure bloc-Toeplitz si elle est constituée de matrices $n\times n$ de la forme $\underline{T_{(n)}}_{jk} = \underline{\tau}(k-j)$, $(j,k=1,\ldots,p)$, où $\underline{\tau}(.)$ est une fonction matricielle $n\times n$, c'est à dire de la forme :

$$\underline{\widehat{T_{(pn)}}} = \begin{bmatrix} \underline{\tau}(0) & \underline{\tau}(1) & \cdots & \underline{\tau}(p-1) \\ \underline{\tau}(-1) & \underline{\tau}(0) & \cdots & \underline{\tau}(p-2) \\ \underline{\tau}(-2) & \vdots & \cdots & \underline{\tau}(p-3) \\ \vdots & \vdots & \cdots & \vdots \\ \underline{\tau}(-(p-1)) & \underline{\tau}(-(p-2)) & \cdots & \underline{\tau}(0) \end{bmatrix}$$

Cas particulier :

Si $n=1$, la matrice est dite à structure simple Toeplitz et est notée $\underline{T_{(p)}}$.

1.3.2 $\underline{\Gamma_{(p)}}$ pour un processus unidimensionnel

Comme il a été rappelé précédemment, un processus à temps discret est une suite de variables aléatoires indexées par le temps.

Il vient qu'en application du principe de base de la technique SSA, les p variables choisies sont les p variables aléatoires ou p composantes successives suivantes $Z_t, Z_{t+1}, Z_{t+2}, \cdots, Z_{t+(p-1)}$, notées aussi $Z^{(1)}$, $Z^{(2)}, \cdots, Z^{(p)}$.

De plus, une réalisation à l'instant t de $Z^{(1)}$, $Z^{(2)}, \cdots, Z^{(p)}$ est :

(1.6) $(z_1(t), z_2(t), \cdots, z_p(t))$ avec $z_i(t) = Z_{t+(i-1)}(t)$ $(1 \le i \le p)$

Dans l'hypothèse où le processus est stationnaire, la matrice de covariance théorique autour de ces p variables s'écrit plus simplement :

(1.7) $$\underline{\Gamma_{(p)}} = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(p-1) \\ \gamma(1) & \gamma(0) & \ddots & \vdots \\ \vdots & & \ddots & \gamma(1) \\ \gamma(p-1) & \cdots & \gamma(1) & \gamma(0) \end{bmatrix}$$

Propriété 1. 5

$\underline{\Gamma_{(p)}}$ est une matrice à structure simple Toeplitz, symétrique et définie positive.

Démonstration :

La démonstration est immédiate, il suffit de s'en référer à la définition 1.5 et aux rappels qui ont été faits sur les propriétés de covariance d'un processus unidimensionnel.

En effet, $\underline{\Gamma_{(p)}}$ est de la forme $\underline{T_{(p)}} = \left\{\tau_{j,k}\right\}$ ($j = 1,\cdots p$ et $k = 1,\cdots p$) avec $\tau_{j,k} = \tau(k-j)$ et $\tau(.) = \gamma(.)$, une fonction paire et positive.

Fin démonstration.

1.3.3 $\widehat{\underline{\Gamma_{(pn)}}}$ pour un processus multidimensionnel

En référence à la définition 1.1, un processus multidimensionnel est une famille de vecteurs aléatoires indexés par le temps. Il vient qu'en application du principe de base, nous pouvons construire une matrice de covariance autour de p vecteurs aléatoires décalés, n -dimensionnels, de la forme $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$.

Ces vecteurs sont notés $\underline{Z}^{(1)}$, $\underline{Z}^{(2)}, \cdots, \underline{Z}^{(p)}$ avec :

$$\underline{Z}^{(1)} = (Z_{1t}, Z_{2t}, \cdots, Z_{nt})', \cdots, \underline{Z}^{(p)} = (Z_{1t+(p-1)}, Z_{2t+(p-1)}, \cdots, Z_{nt+(p-1)})'$$

Une réalisation à l'instant t de ($\underline{Z}^{(1)}$, $\underline{Z}^{(2)}, \cdots, \underline{Z}^{(p)}$) s'écrit :

(1.8) $$((z_{1,1}(t), z_{2,1}(t), \cdots, z_{n,1}(t))', \cdots, (z_{1,p}(t), z_{2,p}(t), \cdots, z_{n,p}(t))')$$

où $z_{i,j}(t) = Z_{it+(j-1)}(t)$ est une réalisation à l'instant t de la variable aléatoire $Z_{it+(j-1)}$, $i = 1,\cdots n$ et $j = 1,\cdots p$.

De plus, dans l'hypothèse où le processus sous-jacent aux séries est supposé **stationnaire**, la matrice de covariance autour des p vecteurs aléatoires $\underline{Z}^{(1)}$, $\underline{Z}^{(2)},\cdots,\underline{Z}^{(p)}$, notée $\underline{\widetilde{\Gamma_{(pn)}}}$, peut s'écrire plus simplement :

$$\begin{array}{c} \\ \underline{Z}^{(1)} \\ \\ \underline{Z}^{(2)} \\ \vdots \\ \vdots \\ \underline{Z}^{(p)} \end{array} \begin{array}{c} \begin{array}{cccccc} \underline{Z}^{(1)'} & & \underline{Z}^{(2)'} & & & \underline{Z}^{(p)'} \end{array} \\ \begin{bmatrix} \underline{\Gamma}_{(n)}(0) & \underline{\Gamma}_{(n)}(1) & \cdots & \cdots & \underline{\Gamma}_{(n)}(p-1) \\ \\ \underline{\Gamma}_{(n)}(-1) & \underline{\Gamma}_{(n)}(0) & & & \underline{\Gamma}_{(n)}(p-2) \\ \vdots & & & & \vdots \\ \vdots & & & \ddots & \vdots \\ \underline{\Gamma}_{(n)}(-(p-1)) & \underline{\Gamma}_{(n)}(-(p-2)) & & & \underline{\Gamma}_{(n)}(0) \end{bmatrix} \end{array}$$

où les regroupements se font par paquets de jours ou instants successifs comme dans le cas unidimensionnel.

$\underline{\Gamma}_{(n)}(j)$ $(0 \le j \le p-1)$ est la matrice de covariance décalée au retard j du processus multidimensionnel. Cette matrice a été définie dans la première partie de ce chapitre pour précisément caractériser un processus stationnaire multidimensionnel.

Par construction de la matrice $\underline{\widetilde{\Gamma_{(pn)}}}$, il s'ensuit la propriété essentielle suivante.

Propriété 1. 6

La matrice $\underline{\widetilde{\Gamma_{(pn)}}} = \left[\underline{\tau}(k-j)\right]$ est une matrice $(pn)\times(pn)$ à structure bloc-Toeplitz.

Démonstration :

De même, la démonstration est immédiate. Pour cela, il suffit de s'en référer à

la définition 1.5 et de poser $\underline{\tau}(.) = \underline{\Gamma_{(n)}}(.)$, la fonction de covariance matricielle définie en (1.1).

Fin démonstration.

$\widehat{\underline{\Gamma_{(pn)}}}$ s'écrit aussi en référence à (1.2) :

(1.9)
$$\begin{bmatrix} \underline{\Gamma_{(n)}}(0) & \underline{\Gamma_{(n)}}(1) & \cdots & \cdots & \underline{\Gamma_{(n)}}(p-1) \\ \underline{\Gamma_{(n)}}'(1) & \underline{\Gamma_{(n)}}(0) & & & \underline{\Gamma_{(n)}}(p-2) \\ \vdots & & & & \vdots \\ \vdots & & & \ddots & \vdots \\ \underline{\Gamma_{(n)}}'(p-1) & \underline{\Gamma_{(n)}}'(p-2) & & & \underline{\Gamma_{(n)}}(0) \end{bmatrix}$$

1.4 Composantes principales d'un processus multidimensionnel et stationnaire

La recherche des composantes principales de plusieurs séries temporelles doit maintenant passer par l'analyse des éléments propres de la matrice de covariance $\widehat{\underline{\Gamma_{(pn)}}}$.

Tout comme dans le cas des matrices à structure simple Toeplitz $(n = 1)$, le calcul direct des valeurs propres de la matrice $\widehat{\underline{\Gamma_{(pn)}}}$ est souvent infaisable.

Mais, dans un premier temps, si le processus multidimensionnel sous-jacent aux n séries temporelles est supposé circulaire, le calcul est alors simplifié. Ce résultat est ici démontré et il utilise principalement les travaux de [Williamson31], [MacDuffee33], [Friedman51], [Aitken54] et [Friedman61].

Il suffira ensuite de remarquer que si la période du processus circulaire tend vers l'infini, ce processus se rapproche alors du processus stationnaire étudié. Les propriétés d'un processus circulaire multidimensionnel fournissent des approximations aux propriétés d'un processus simplement stationnaire.

Nous en déduirons les propriétés asymptotiques des éléments propres de la matrice $\widehat{\underline{\Gamma_{(pn)}}}$ ainsi que la forme des composantes principales de plusieurs séries temporelles.

1.4.1 Approximations des éléments propres de $\widehat{\underline{\Gamma}_{(pn)}}$

Dans cette partie, la notion de structure bloc-circulaire d'une matrice est essentielle. Nous rappelons sa définition. Puis, c'est autour du concept de processus circulaire que nous allons travailler pour nous rapprocher du processus simplement stationnaire.

1.4.1.1 Matrice à structure bloc-circulaire

Définition 1. 6

Une matrice carrée $\widehat{\underline{C_{(pn)}}}$ $(pn)\times(pn)$, est dite à structure bloc-circulaire d'ordre p, si elle est constituée de matrices $n\times n$ de la forme $\underline{C_{(n)}}_{jk} = \underline{c}(k-j)$ $(j,k=1,\ldots,p)$ où $\underline{c}(.)$ est une fonction matricielle $n\times n$ de **période** p, c'est à dire de la forme :

$$\widehat{\underline{C_{(pn)}}} = \begin{bmatrix} \underline{c}(0) & \underline{c}(1) & \cdots & \underline{c}(p-1) \\ \underline{c}(p-1) & \underline{c}(0) & \cdots & \underline{c}(p-2) \\ \underline{c}(p-2) & \vdots & \cdots & \underline{c}(p-3) \\ \vdots & \vdots & \cdots & \vdots \\ \underline{c}(1) & \underline{c}(2) & \cdots & \underline{c}(0) \end{bmatrix}.$$

Cas particulier :

Si $n=1$, la matrice est dite à structure simple circulaire et est notée $\underline{C_{(p)}}$.

1.4.1.2 Processus circulaire multidimensionnel

Un processus circulaire multidimensionnel supposé stationnaire peut se caractériser de la façon suivante :

Définition 1. 7

Un processus multidimensionnel $(\underline{Z}_t,\ t\in\mathbb{Z})$ est périodique ou circulaire de période p si

$$\underline{Z}_t = \underline{Z}_{t+p}\ \ (\forall t\in\mathbb{Z})$$

Propriété 1. 7

Si de plus le processus $(\underline{Z}_t)$ est supposé stationnaire, sa fonction d'autocovariance matricielle $\underline{\Gamma}_{(n)}(.)$ vérifie :

$$\underline{\Gamma}_{(n)}(l) = \underline{\Gamma}_{(n)}{}'(p-l) \qquad (1.10)$$

Démonstration :

$$\underline{\Gamma}_{(n)}(l) = Cov(\underline{Z}_t, \underline{Z}_{t+l}) = Cov(\underline{Z}_{t+p}, \underline{Z}_{t+l}) = \underline{\Gamma}_{(n)}{}'(p-l)$$

Fin démonstration.

Il vient que la matrice de covariance de $(\underline{Z}_t)$, notée $\underline{\widehat{W}_{(pn)}}$, autour des p composantes successives $\underline{Z}_t, \underline{Z}_{t+1}, \underline{Z}_{t+2}, \cdots, \underline{Z}_{t+(p-1)}$, s'écrit d'après (1.9) et (1.10) :

$$\underline{\widehat{W}_{(pn)}} = \begin{bmatrix} \underline{\Gamma}_{(n)}(0) & \underline{\Gamma}_{(n)}(1) & \cdots & \underline{\Gamma}_{(n)}(p-1) \\ \underline{\Gamma}_{(n)}(p-1) & \underline{\Gamma}_{(n)}(0) & \cdots & \underline{\Gamma}_{(n)}(p-2) \\ \underline{\Gamma}_{(n)}(p-2) & \vdots & \cdots & \underline{\Gamma}_{(n)}(p-3) \\ \vdots & \vdots & \cdots & \vdots \\ \underline{\Gamma}_{(n)}(1) & \underline{\Gamma}_{(n)}(2) & \vdots & \underline{\Gamma}_{(n)}(0) \end{bmatrix} \qquad (1.11)$$

Démonstration :

$$\underline{\Gamma}_{(n)}{}'(l) = \underline{\Gamma}_{(n)}{}''(p-l) = \underline{\Gamma}_{(n)}(p-l) \ (l = 0, \cdots, (p-1))$$

Fin démonstration.

Enfin, en référence à la définition 1.6, nous pouvons en déduire immédiatement la propriété suivante de la matrice $\underline{\widehat{W}_{(pn)}}$.

Propriété 1. 8

$\widehat{\underline{W_{(pn)}}}$ est une matrice carrée d'ordre p, à structure bloc-circulaire constituée de matrices $n \times n$ de la forme $\underline{\Gamma_{(n)}}_{jk} = \underline{\Gamma_{(n)}}(k-j)\ (j,k=0,\ldots,p-1)$ où $\underline{\Gamma_{(n)}}(.)$ est la fonction de covariance matricielle définie en (1.1).

1.4.1.3 Eléments propres de la matrice de covariance d'un processus circulaire – Produit tensoriel de matrices

Les éléments propres d'une matrice à structure bloc-circulaire font l'objet du théorème 1.2 qui va suivre. Nous pourrons constater que ce théorème est une généralisation du théorème 1.1 bien connu dans le cadre unidimensionnel dont nous rappelons brièvement l'essentiel.

Théorème 1. 1

Soit une **matrice carrée** $\underline{C_{(p)}} = \left\{c_{j,k}\right\}$ $(j = 1,\cdots p$ et $k = 1,\cdots p)$ **circulaire d'ordre** p**,** i.e. $c_{j,k} = c(k-j)$ où $c(.)$ est une fonction de période p**,**

$$\underline{C_{(p)}} = \begin{bmatrix} c(0) & c(1) & \cdots & c(p-1) \\ c(p-1) & c(0) & \cdots & c(p-2) \\ c(p-2) & \vdots & \cdots & c(p-3) \\ \vdots & \vdots & \cdots & \vdots \\ c(1) & c(2) & \cdots & c(0) \end{bmatrix},$$

ses **valeurs propres** λ_k sont données par :

$$\text{(1.12)} \qquad \lambda_k(\underline{C_{(p)}}) = \sum_{j=0}^{p-1} c(j)\, \mathrm{e}^{-i(2\pi jk/p)} \ \ (k = 0,\ 1,\cdots p-1)$$

et ses **vecteurs propres** $\overrightarrow{V_k}$ vérifiant $\underline{C}\ \overrightarrow{V_k} = \lambda_k\ \overrightarrow{V_k}$, $k = 0,\ 1,\cdots p-1$, sont donnés par :

$$\text{(1.13)} \qquad \overrightarrow{V_k}(\underline{C_{(p)}}) = p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}; j = 0,\cdots,p-1\right]' \ \ (k = 0,\ 1,\cdots p-1)$$

Les valeurs propres de la matrice circulaire $\underline{C_{(p)}}$ *coïncident avec la transformée de Fourier discrète des séquences* $c(t)$, $t = 0,\ 1,\cdots p-1$.

Les vecteurs propres obtenus sont indépendants des coefficients de la matrice $\underline{C_{(p)}}$.

Nous nous plaçons à nouveau dans le cadre multidimensionnel pour énoncer le théorème 1.2 relatif aux éléments propres d'un processus circulaire multidimensionnel.

Théorème 1. 2

Soit une **matrice** $\widetilde{\underline{C_{(pn)}}} = \left[\underline{c}(k-j)\right]$, $(pn)\times(pn)$, à structure **bloc-circulaire d'ordre** p**,** alors ses **valeurs propres** sont données par les valeurs propres de :

$$(1.14) \qquad \sum_{j=0}^{p-1} \underline{c}(j)\,\mathrm{e}^{-i(2\pi jk/p)} \quad (k = 0,\ 1,\cdots p-1)$$

et ses **vecteurs propres** sont donnés par :

$$(1.15) \qquad p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\,\underline{u}_{lk}; j = 0,\cdots,(p-1)\right]$$
$$(k = 0,\ 1,\cdots p-1,\ l = 0,\ 1,\cdots n-1)$$

où $\underline{u}_{lk}$ sont les vecteurs propres de (1.14).

Démonstration :

Pour démontrer ce théorème, nous avons besoin de la définition du produit tensoriel de vecteurs et de matrices, de propriétés qui lui sont associées ainsi que de deux corollaires.

Définition 1. 8
ou produit tensoriel de vecteurs et espaces vectoriels linéaires.

Soient deux espaces vectoriels linéaires S_n et S_r de dimension respective n et r.

Soit $\underline{x}$ un vecteur quelconque de S_n dont les composantes sont $\xi_1, \xi_2, \cdots, \xi_n$ et soit $\underline{y}$ un vecteur quelconque de S_r dont les composantes sont $\eta_1, \eta_2, \cdots, \eta_n$.

Le produit tensoriel de $\underline{x}$ et $\underline{y}$ est défini pour être le vecteur $\underline{z}$ à nr composantes de la forme $\xi_i \eta_j$, $(i = 1,2,\cdots,n)$ et $(j = 1,2,\cdots,r)$. Il est noté : $\underline{x} \otimes \underline{y}$.

L'espace vectoriel engendré par l'ensemble des vecteurs $\underline{z}$ possibles est appelé le produit tensoriel des espaces S_n et S_r. Il est bien sûr de dimension nr et est noté $S_n \otimes S_r$.

Définition 1. 9
ou produit tensoriel de transformations linéaires et matrices.

Supposons $\underline{A}$ une transformation linéaire de S_n dans lui-même et supposons $\underline{B}$ une transformation linéaire de S_r dans lui-même, le produit tensoriel de $\underline{A}$ et $\underline{B}$ (noté $\underline{A} \otimes \underline{B}$) est la transformation linéaire $\underline{C}$ de dimension nr définie comme il suit :

si

$$\underline{x}' = \underline{A}\ \underline{x} \text{ et } \underline{y}' = \underline{B}\ \underline{y}$$

alors

(1.16) $$\underline{z}' = \underline{C}\ \underline{z} = \underline{C}(\underline{x} \otimes \underline{y}) = \underline{A}\underline{x} \otimes \underline{B}\underline{y} = \underline{x}' \otimes \underline{y}'$$

La transformation $\underline{C}$ est représentée par différentes matrices correspondant aux différents rangements possibles des nr composantes de $\underline{z}$. Les différents rangements possibles sont donnés par $\xi_k = \xi_i \eta_j$ avec k $(1 \le k \le nr)$ associé à la paire de nombres i, j $(1 \le i \le n)$ et $(1 \le j \le r)$.

Par exemple, nous choisissons pour $\underline{A} \otimes \underline{B}$, l'écriture matricielle suivante

$$\begin{pmatrix} b_{11}A & b_{12}A & \cdots & b_{1r}A \\ b_{21}A & b_{22}A & \cdots & b_{2r}A \\ \vdots & \vdots & & \vdots \\ b_{r1}A & b_{r2}A & & b_{rr}A \end{pmatrix},$$

et pour $\underline{B} \otimes \underline{A}$, l'écriture matricielle

$$\begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & & \vdots \\ a_{n1}B & a_{n2}B & & a_{nn}B \end{pmatrix}.$$

De la définition du produit tensoriel, nous pouvons en déduire facilement les propriétés suivantes.

Propriété 1. 9

(1.17) $$(\underline{A}_1 + \underline{A}_2) \otimes \underline{B} = \underline{A}_1 \otimes \underline{B} + \underline{A}_2 \otimes \underline{B},$$

où le symbole $+$ est l'addition de matrices ordinaire, et

(1.18) $$(\underline{A}_1 \otimes \underline{B}_1).(\underline{A}_2 \otimes \underline{B}_2) = (\underline{A}_1.\underline{A}_2) \otimes (\underline{B}_1.\underline{B}_2)$$

où le symbole $.$ est le produit de matrices ordinaire.

De la définition en (1.16) et des propriétés (1.17) et (1.18), il en découle les 2 corollaires suivants.

Corollaire 1. 1

Si $\underline{x}$ est un vecteur propre de $\underline{A}$ correspondant à la valeur propre λ et si $\underline{y}$ est un vecteur propre de $\underline{B}$ correspondant à la valeur propre μ alors $\underline{x} \otimes \underline{y}$ est un vecteur propre de $\underline{A} \otimes \underline{B}$ correspondant à la valeur propre $\lambda\mu$.

Démonstration :

$$(\underline{A} \otimes \underline{B})(\underline{x} \otimes \underline{y}) = \underline{A}\underline{x} \otimes \underline{B}\underline{y} = \lambda\underline{x} \otimes \mu\underline{y} = \lambda\mu(\underline{x} \otimes \underline{y})$$

Fin démonstration.

Corollaire 1. 2

Soit $\underline{C} = \underline{A}_1 \otimes \underline{B}_1 + \cdots + \underline{A}_p \otimes \underline{B}_p$.

Si le vecteur $\underline{y}$ est un vecteur propre pour toutes les matrices $\underline{B}_1, \underline{B}_2, \cdots, \underline{B}_p$, c'est à dire :

(1.19) $$\underline{B}_j\underline{y} = \mu_j\underline{y} \qquad (1 \leq j \leq p)$$

et si $\underline{x}$ est un vecteur tel que

(1.20) $$(\mu_1 \underline{A}_1 + \cdots + \mu_p \underline{A}_p)\underline{x} = \lambda \underline{x}$$

alors λ est une valeur propre de $\underline{C}$ et

$$\underline{C}(\underline{x} \otimes \underline{y}) = \lambda(\underline{x} \otimes \underline{y}) .$$

Démonstration :

$$\underline{C}(\underline{x} \otimes \underline{y}) = \sum_{j=1}^{p} (\underline{A}_j \otimes \underline{B}_j)\underline{x} \otimes \underline{y} = \sum_{j=1}^{p} (\underline{A}_j \underline{x}) \otimes (\underline{B}_j \underline{y})$$

$$= \sum_{j=1}^{p} (\mu_j \underline{A}_j \underline{x}) \otimes \underline{y} = \lambda \underline{x} \otimes \underline{y}$$

Fin démonstration.

➢ ***Les corollaires 1.1 et 1.2 fournissent une méthode « simple » pour obtenir les valeurs propres de matrices à structure bloc.***

Démonstration du théorème 1.2 :

La matrice $\underline{\widetilde{C_{(pn)}}}$ peut s'écrire comme une somme de produits tensoriels de matrices :

$$\underline{\widetilde{C_{(pn)}}} = \sum_{j=0}^{p-1} \underline{c}(j) \otimes \underline{J}^j$$

où

$$\underline{J} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \vdots \\ \vdots & & \cdots & & \vdots \\ \vdots & & \cdots & & 1 \\ 1 & & \cdots & & 0 \end{pmatrix}$$

et $\underline{c}(j)$ est une matrice carrée d'ordre n .

La matrice $\underline{J}$ est une matrice carrée de **permutation** p-dimensionnelle qui par construction a la propriété suivante $\underline{J}^p = \underline{I_{(p)}}$, c'est à dire la matrice identité de dimension p.

Les valeurs propres de $\underline{J}$ sont donc les racines $p^{ièmes}$ de l'unité égales à $\mathrm{e}^{-i(2\pi k/p)}$ pour $0 \le k \le p-1$.

Il s'ensuit que les valeurs propres de $\underline{J}^j$ $(0 \le j \le p-1)$ peuvent s'écrire $\mathrm{e}^{-i(2\pi jk/p)}$ pour $0 \le k \le p-1$.

Soit le vecteur $\underline{y} = p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}; j = 0,\cdots,(p-1)\right]'$, il est un vecteur propre commun aux matrices $\underline{J}^j$ $(0 \le j \le p-1)$ pour la valeur propre $\mu_j = \mathrm{e}^{-i(2\pi jk/p)}$ $(0 \le k \le p-1)$.

Démonstration :

Posons $\underline{y} = p^{-1/2}\tilde{\underline{y}}$ et vérifions simplement la relation $\underline{J}^j\tilde{\underline{y}} = \mu_j\tilde{\underline{y}}$.

$$\underline{J}^j\tilde{\underline{y}} = \underline{J}^j\left[\mathrm{e}^{-i(2\pi jk/p)}; j = 0,\cdots,(p-1)\right] = \begin{bmatrix} \mathrm{e}^{-i(2\pi jk/p)} \\ \mathrm{e}^{-i(2\pi(j+1)k/p)} \\ \vdots \\ \mathrm{e}^{-i(2\pi(j+(p-1))k/p)} \end{bmatrix},$$

c'est à dire exactement

$$\mu_j\tilde{\underline{y}} = \mathrm{e}^{-i(2\pi jk/p)}\left[\mathrm{e}^{-i(2\pi jk/p)}; j = 0,\cdots,(p-1)\right].$$

Quant au coefficient $p^{-1/2}$, il intervient uniquement dans la normalisation de $\tilde{\underline{y}}$ pour la norme euclidienne.

Rappel : la quantité $\sqrt{N}\left|\underline{A}\right|$ représente la norme euclidienne de la matrice $\underline{A} = \left\{a_{k,j}\right\}(N \times N)$ avec $\left|\underline{A}\right|$ la norme de Hilbert-Schmidt de $\underline{A}$ définie par :

$$\left|\underline{A}\right| = \left(N^{-1}\sum_{k=0}^{N-1}\sum_{j=0}^{N-1}\left|a_{k,j}\right|^2\right)^{1/2} = \left(N^{-1}Tr\left[\underline{A}'\underline{A}\right]\right)^{1/2} = \left(N^{-1}\sum_{k=0}^{N-1}\lambda_k\right)^{1/2}.$$

Fin démonstration.

L'hypothèse (1.19) du corollaire 1.2 est donc vérifiée avec $\underline{B}_j = \underline{J}^j$ et $\mu_j = \mathrm{e}^{-i(2\pi jk/p)}$ ($0 \le j \le p-1$ et $0 \le k \le p-1$).

Pour satisfaire l'hypothèse (1.20) du corollaire 1.2, posons $\underline{A}_j = \underline{c}(j)$.

Il vient que rechercher les valeurs propres λ de $\underline{\widetilde{C_{(pn)}}}(= \underline{C})$ revient à rechercher les valeurs propres de la matrice carrée d'ordre n suivante, notée $\underline{\Sigma_{(n)}^{(k)}}$:

$$\underline{\Sigma_{(n)}^{(k)}} = (\mu_0 \underline{A}_0 + \cdots + \mu_{p-1}\underline{A}_{p-1}) = \sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)}\, \underline{c}(j) \ (0 \le k \le p-1)$$

Les valeurs propres λ dépendent alors de deux paramètres k ($k = 0,\ 1, \cdots p-1$) et l ($l = 0,\ 1, \cdots n-1$). Nous noterons par la suite ces valeurs propres λ_{lk}.

De plus, les vecteurs propres $\underline{z}$ pour les valeurs propres λ_{lk} sont les vecteurs propres $\underline{x} \otimes \underline{y}$ avec :

- $\underline{x}$ vecteurs propres pour λ_{lk} valeurs propres de $\sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)}\, \underline{c}(j)$. Nous noterons par la suite ces vecteurs $\underline{u}_{lk}$.

- et $\underline{y} = p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}; j = 0, \cdots, (p-1)\right]$ vecteurs propres pour $\mu_j = \mathrm{e}^{-i(2\pi jk/p)}$.

Les vecteurs propres $\underline{z}$ s'écrivent donc :

$$p^{-1/2}\underline{u}_{lk} \otimes \left[\mathrm{e}^{-i(2\pi jk/p)}; j = 0, \cdots, (p-1)\right]',\ k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1$$

où $\underline{u}_{lk}$ sont les vecteurs propres pour les valeurs propres de $\sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)}\, \underline{c}(j)$.

Les vecteurs propres $\underline{z}$ sont $(p \times n)$-dimensionnels. Le produit tensoriel s'écrit alors plus simplement :

$$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\, \underline{u}_{lk}; j = 0, \cdots, (p-1)\right],\ k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1$$

Le théorème 1.2 est démontré.

Fin démonstration.

Fin démonstration du théorème 1.2.

➢ ***Le théorème 1.1 peut s'appliquer à la matrice*** $\widehat{\underline{W_{(pn)}}}$ ***définie en (1.11).***

1.4.1.4 Approximations aux propriétés des processus stationnaires

Propriété 1. 10

Si p la période du processus est finie, les propriétés d'un processus circulaire fournissent des approximations aux propriétés des processus simplement stationnaires.

Démonstration :

Il suffit de remarquer que si $p \to \infty$, le processus circulaire tend vers le processus simplement stationnaire.

Fin démonstration.

Les éléments propres de la matrice $\widehat{\underline{\Gamma_{(pn)}}}$ définie en (1.9) peuvent donc être approchés par les éléments propres de la matrice $\widehat{\underline{W_{(pn)}}}$.

Les valeurs propres de $\widehat{\underline{\Gamma_{(pn)}}}$ se rapprochent des valeurs propres de :

(1.21) $$\sum_{j=0}^{p-1} \underline{\Gamma_{(n)}}(j)\, \mathrm{e}^{-i(2\pi jk/p)} \quad (k = 0,\ 1, \cdots p-1)$$

et

les vecteurs propres de $\widehat{\underline{\Gamma_{(pn)}}}$ se rapprochent de :

(1.22) $$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\, \underline{u}_{lk}; j = 0, \cdots, p-1\right] \ (k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1)$$

où $\underline{u}_{lk}$ sont les vecteurs propres de (1.21).

Ecriture simplifiée pour la recherche des valeurs propres de $\underline{\widehat{\Gamma_{(pn)}}}$:

La relation (1.21) s'écrit aussi à l'aide de (1.1) :

$$(1.23)\quad \underline{\Sigma_{(n)}^{(k)}} = \begin{pmatrix} \sum_{j=0}^{p-1} \gamma_{11}(j)\,\mathrm{e}^{-i(2\pi jk/p)} & \cdots & \sum_{j=0}^{p-1} \gamma_{1n}(j)\,\mathrm{e}^{-i(2\pi jk/p)} \\ \vdots & \ddots & \vdots \\ \sum_{j=0}^{p-1} \gamma_{n1}(j)\,\mathrm{e}^{-i(2\pi jk/p)} & \cdots & \sum_{j=0}^{p-1} \gamma_{nn}(j)\,\mathrm{e}^{-i(2\pi jk/p)} \end{pmatrix} (k = 0,\ 1, \cdots p-1)$$

Le calcul des valeurs propres des matrices $\underline{\Sigma_{(n)}^{(k)}}$ définies en (1.23) fournit donc une expression approchée des valeurs propres de la matrice $\underline{\widehat{\Gamma_{(pn)}}}$.

Ecriture simplifiée pour la recherche des vecteurs propres de $\underline{\widehat{\Gamma_{(pn)}}}$:

Les vecteurs propres en (1.22) ($p \times n$)-dimensionnels, notés $\overrightarrow{V_{lk}}$, peuvent s'écrire comme il suit :

$$(1.24)\qquad \overrightarrow{V_{lk}} = \left[(a_{ji}^{(lk)}); 1 \le j \le p, 1 \le i \le n)\right]' = \left[(a_{11}^{(lk)} \cdots a_{1n}^{(lk)}) \cdots (a_{p1}^{(lk)} \cdots a_{pn}^{(lk)})\right]'$$

$$= \left[\overrightarrow{V_{lk}^{(1)}}' \quad \cdots \quad \overrightarrow{V_{lk}^{(p)}}'\right]$$

où une composante $\overrightarrow{V_{lk}^{(j)}}$ n -dimensionnelle se rapproche de $p^{-1/2}\,\mathrm{e}^{-i(2\pi(j-1)k/p)}\,\underline{u}_{lk}$ avec $\underline{u}_{lk}$ les vecteurs propres n -dimensionnels de (1.23).

1.4.2 Forme des composantes principales de plusieurs séries temporelles

En référence à (1.24), la $(l,k)^{ième}$ composante principale $F^{(lk)}$ des vecteurs aléatoires $\underline{Z}^{(1)}$, $\underline{Z}^{(2)}, \cdots, \underline{Z}^{(p)}$ s'écrit :

$$(1.25)\qquad F^{(lk)} = \sum_{j=1}^{p} \overrightarrow{V_{lk}^{(j)}}\,\underline{Z}^{(j)}\ ,\ k = 0,\ 1, \cdots p-1\ ,\ l = 0,\ 1, \cdots n-1$$

Et en référence à (1.8), une réalisation à l'instant t de $F^{(lk)}$ est :

$$f_{lk}(t) = \sum_{j=1}^{p}\sum_{i=1}^{n} a_{ji}^{(lk)} z_{i,j}(t) \qquad (1.26)$$

avec $z_{i,j}(t) = Z_{it+(j-1)}(t)$, une réalisation à l'instant t de la variable aléatoire $Z_{it+(j-1)}$, $i = 1,\cdots n$ et $j = 1,\cdots p$,

et $a_{ji}^{(lk)}$, les composantes des vecteurs propres $\overrightarrow{V_{lk}}$ définies en (1.24).

➢ ***Les composantes principales (scores) de plusieurs séries temporelles sont des combinaisons linéaires complexes du passé, du présent, du futur des séries temporelles.***

1.4.3 Cas unidimensionnel : des résultats retrouvés

1.4.3.1 Les éléments propres de $\underline{\Gamma_{(p)}}$

Dans le cas particulier où $n = 1$, les valeurs propres des matrices $\underline{\Sigma_{(n)}^{(k)}}$ définies en (1.23) se réduisent aux combinaisons linéaires $\sum_{j=0}^{p-1} \gamma(j)\,\mathrm{e}^{-i(2\pi jk/p)}$ ($k = 0,\ 1,\cdots p-1$) lesquelles fournissent une expression approchée des valeurs propres de la matrice $\underline{\Gamma_{(p)}}$ définie en (1.7).

Il s'ensuit les résultats bien connus établis dans [Brillinger75] et [Brillinger81] sur les approximations des valeurs propres $\lambda_k(\underline{\Gamma_{(p)}})$ de la matrice $\underline{\Gamma_{(p)}}$:

$$\sum_{j=0}^{p-1} \gamma(j)\,\mathrm{e}^{-i(2\pi jk/p)} \quad (k = 0,\ 1,\cdots p-1) \qquad (1.27)$$

et les approximations des vecteurs propres $\overrightarrow{V_k}(\underline{\Gamma_{(p)}})$:

$$p^{-1/2}\left[\mathrm{e}^{-i(2\pi kj/p)}; j = 0,\cdots,(p-1)\right]' \quad (k = 0,\ 1,\cdots p-1) \qquad (1.28)$$

Notons que dans l'ouvrage de Brillinger, l'approximation pour les valeurs propres et vecteurs propres utilise le théorème de Wielandt-Hoffman basé sur l'équivalence

asymptotique de matrices hermitiennes pour la norme euclidienne. Nous renvoyons aussi à l'ouvrage [Wilkinson65] dans lequel ce théorème est démontré.

➢ ***Les vecteurs propres*** $\overrightarrow{V_k}(\underline{\Gamma_{(p)}})$ ***sont indépendants (en limite) des coefficients du modèle unidimensionnel.***

1.4.3.2 La forme des composantes principales d'une seule série temporelle

Les q premières composantes principales $F^{(k)}$, $1 \leq k \leq q$, pour les variables aléatoires $Z^{(1)}, Z^{(2)}, \cdots, Z^{(p)}$ vérifient $F^{(k)} = \sum_{j=1}^{p} a_j^{(k)} Z^{(j)}$.

Une réalisation à l'instant t de $F^{(k)}$, la $k^{ième}$ composante principale de la série temporelle , s'écrit donc :

$$f_k(t) = \left[a_1^{(k)} \cdots a_p^{(k)}\right] \begin{bmatrix} z_1(t) \\ \vdots \\ z_p(t) \end{bmatrix} \qquad (1.29)$$

où $z_1(t), \cdots, z_p(t)$ sont les réalisations à l'instant t des variables $Z^{(1)}, Z^{(2)}, \cdots, Z^{(p)}$ et $\left[a_1^{(k)} \cdots a_p^{(k)}\right]$ sont les coordonnées en ligne du vecteur propre $\overrightarrow{V_k}$ en (1.28).

➢ ***Les composantes principales (scores) d'une série temporelle sont des moyennes mobiles de la série temporelle dont les poids sont les coordonnées des vecteurs propres associés.***

1.5 Cas des processus « indépendants »

Nous nous plaçons désormais dans le cas de n séries temporelles qui sont la réalisation de n processus « indépendants » ou non-corrélés (au sens de la définition 1.3) et stationnaires.

Deux matrices de covariance à structure bloc sont examinées et sont respectivement de dimension (temps$\times$espace) en application de la méthode précédente et (espace$\times$temps) à l'image de la technique M-SSA.

1.5.1 En passant par la matrice de covariance bloc-Toeplitz

Le premier cas de figure envisagé est une application directe de la méthode qui utilise la théorie des matrices bloc-circulaires Toeplitz et le produit tensoriel.

1.5.1.1 Les éléments propres de $\widehat{\underline{\Gamma}_{(pn)}}$

Si les n processus sont « indépendants », les matrices $\underline{\Sigma}_{(n)}^{(k)}$, $k = 0,\ 1,\cdots p-1$ définies en (1.23) sont les matrices diagonales suivantes :

$$\underline{\Sigma}_{(n)}^{(k)} = \begin{pmatrix} \sum_{j=0}^{p-1} \gamma_{11}(j)\,\mathrm{e}^{-i(2\pi jk/p)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{j=0}^{p-1} \gamma_{nn}(j)\,\mathrm{e}^{-i(2\pi jk/p)} \end{pmatrix} (k = 0,\ 1,\cdots p-1)$$

Les valeurs propres réelles de $\widehat{\underline{\Gamma}_{(pn)}}$ sont donc approximativement :

(1.30) $$\sum_{j=0}^{p-1} \gamma_{ll}(j)\,\mathrm{e}^{-i(2\pi jk/p)},\ k = 0,\ 1,\cdots p-1 \text{ et } l = 0,\cdots,n-1$$

En référence au résultat (1.27), elles coïncident avec les valeurs propres de la matrice de covariance $\underline{\Gamma}_{(p)}$ de chacun des processus.

Les vecteurs propres $\underline{u}_{lk}$ des matrices $\underline{\Sigma}_{(n)}^{(k)}$ $(k = 0,\ 1,\cdots p-1)$ sont donc les vecteurs unitaires $\underline{e}_l$ de $\mathbb{R}^n$ qui ont comme seule coordonnée non nulle celle à la $(l+1)^{ième}$ ligne égale à 1.

Les vecteurs propres $\overrightarrow{V_{lk}}$ de $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent de :

(1.31) $$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\,\underline{e}_l; j = 0,\cdots,(p-1)\right],$$
$$k = 0,\ 1,\cdots p-1,\ l = 0,\ 1,\cdots n-1$$

Reprenons (1.24) et notons désormais les approximations faites par le symbole $\sim$. L'écriture de $\overrightarrow{V_{lk}}$ est alors simplifiée :

(1.32) $$\overrightarrow{V_{lk}} = \left[(a_{ji}^{(lk)}); 1 \le j \le p, 1 \le i \le n)\right]'$$

avec $a_{ji}^{(lk)} \sim 0$ pour $i \neq l$ et $a_{jl}^{(lk)} \sim p^{-1/2}\, \mathrm{e}^{-i(2\pi(j-1)k/p)}$, $1 \leq j \leq p$

Nous pouvons constater que les coefficients $a_{jl}^{(lk)}$ pour $i = l$ coïncident avec les coefficients des vecteurs propres définis en (1.28).

➢ ***Dans le cas où les processus sont « indépendants », les vecteurs propres*** $\overrightarrow{V_{lk}}(\underline{\Gamma_{(np)}})$ ***sont indépendants (en limite) des coefficients du modèle multidimensionnel.***

1.5.1.2 La forme des composantes principales

Il s'ensuit, en référence à (1.26), (1.28) et (1.32), qu'une réalisation à l'instant t de $F^{(lk)}$ s'écrit :

$$f_{lk}(t) \sim \left[a_{1,l}^{(lk)}, a_{2,l}^{(lk)}, \cdots, a_{p,l}^{(lk)}\right]\left[z_1^{(l)}(t), z_2^{(l)}(t), \cdots, z_p^{(l)}(t)\right]', \quad (1.33)$$
$$(k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1)$$

avec $z_j^{(l)}(t) = z_{l,j}(t) = Z_{t+(j-1)}(t)$, une réalisation à l'instant t de la variable $Z^{(j)}$ ($1 \leq j \leq p$) pour le $l^{ième}$ processus,

et $a_{ji}^{(lk)} \sim 0$ pour $i \neq l$ et $a_{jl}^{(lk)} \sim p^{-1/2}\, \mathrm{e}^{-i(2\pi(j-1)k/p)}$, $1 \leq j \leq p$, coefficients indépendants des paramètres du modèle.

➢ ***Les composantes principales (scores) des n séries temporelles « indépendantes » coïncident avec les composantes principales (scores) de chacune des séries temporelles.***

1.5.2 En passant par une matrice de covariance bloc-diagonale

La seconde matrice de covariance étudiée est une matrice bloc-diagonale de dimension (espace×temps) ou ($n \times p$).

1.5.2.1 La matrice de covariance bloc-diagonale $\widehat{\underline{\Gamma_{(np)}}}$

En effet, si les n processus sont « indépendants », nous pouvons construire une matrice de covariance à structure bloc-diagonale autour de n vecteurs aléatoires $\underline{Z}^{(1)}$, $\underline{Z}^{(2)}, \cdots, \underline{Z}^{(n)}$, p-dimensionnels.

Ces vecteurs aléatoires s'écrivent :

$$\underline{Z}^{(1)} = (Z_{1t}, Z_{1t+1}, \cdots, Z_{1t+(p-1)})', \cdots, \underline{Z}^{(n)} = (Z_{nt}, Z_{nt+1}, \cdots, Z_{nt+(p-1)})'$$

Une réalisation à l'instant t de ($\underline{Z}^{(1)}$, $\underline{Z}^{(2)}, \cdots, \underline{Z}^{(n)}$) est alors :

(1.34) $$((z_{1,1}(t), z_{1,2}(t), \cdots, z_{1,p}(t))', \cdots, (z_{n,1}(t), z_{n,2}(t), \cdots, z_{n,p}(t))')$$

où $z_{j,i}(t) = Z_{jt+(i-1)}(t)$ est une réalisation à l'instant t de la variable aléatoire $Z_{jt+(i-1)}$, $i = 1, \cdots p$ et $j = 1, \cdots n$.

Soit $\underline{\widehat{\Gamma_{(np)}}}$ cette matrice, elle est de dimension ($n \times p$) et s'écrit :

$$\underline{\widehat{\Gamma_{(np)}}} = \begin{pmatrix} \underline{\Gamma_{(p)}^{(1,1)}} & \underline{0_{(p)}} & \cdots & \underline{0_{(p)}} \\ \underline{0_{(p)}} & \underline{\Gamma_{(p)}^{(2,2)}} & \cdots & \underline{0_{(p)}} \\ \vdots & \vdots & \ddots & \\ \underline{0_{(p)}} & \underline{0_{(p)}} & \cdots & \underline{\Gamma_{(p)}^{(n,n)}} \end{pmatrix}$$

où les regroupements se font par « croisement » de processus à différents retards à l'image de la technique M-SSA ,

et où $\underline{\Gamma_{(p)}^{(l,l)}}$ est la matrice de covariance définie en (1.7) du $l^{ième}$ processus.

1.5.2.2 Les éléments propres de $\underline{\widehat{\Gamma_{(np)}}}$

Pour la recherche des éléments propres de la matrice $\underline{\widehat{\Gamma_{(np)}}}$, rappelons simplement :

(a) que les valeurs propres λ_k et les vecteurs propres (à droite) $\overrightarrow{V_k}$ d'une matrice $\underline{A}$ carrée sont solutions de l'équation :

$$\underline{A}\overrightarrow{V} = \lambda\overrightarrow{V}$$

(b) et que les valeurs propres sont alors les racines de l'équation caractéristique de $\underline{A}$:

$$\det(\underline{A} - \lambda\underline{I}) = 0$$

avec $\lambda_1 \geq \lambda_2 \geq \cdots$.

Démonstration :

λ_{kl} $(l = 1,\ 2,\cdots n$ et $k = 0,\ 2,\cdots p-1)$ est valeur propre de $\underline{\widetilde{\Gamma_{(np)}}}$ est équivalent à :

$$Det\left[\underline{\widetilde{\Gamma_{(np)}}} - \lambda_{kl}\underline{I_{(np)}}\right] = 0$$

ou encore

$$Det\begin{bmatrix} \underline{\Gamma_{(p)}^{(1,1)}} - \lambda_{kl}\underline{I_{(p)}} & \cdots & \underline{0_p} \\ \vdots & \ddots & \vdots \\ \underline{0_p} & \cdots & \underline{\Gamma_{(p)}^{(n,n)}} - \lambda_{kl}\underline{I_{(p)}} \end{bmatrix} = 0$$

c'est à dire

$$Det\left[\underline{\Gamma_{(p)}^{(1,1)}} - \lambda_{kl}\underline{I_{(p)}}\right] \times \cdots \times Det\left[\underline{\Gamma_{(p)}^{(n,n)}} - \lambda_{kl}^{(l,l)}\underline{I_{(p)}}\right] = 0$$

ou

$$Det\left[\underline{\Gamma_{(p)}^{(1,1)}} - \lambda_{kl}\underline{I_{(p)}}\right] = 0 \text{ ou ... ou } Det\left[\underline{\Gamma_{(p)}^{(n,n)}} - \lambda_{kl}\underline{I_{(p)}}\right] = 0$$

Fin démonstration.

Il s'ensuit la propriété suivante.

Propriété 1. 11

Une valeur propre de $\underline{\Gamma_{(p)}^{(l,l)}}$ $(l = 1,\ 2,\cdots n)$ est valeur propre de $\underline{\widetilde{\Gamma_{(np)}}}$ et inversement une valeur propre de $\underline{\widetilde{\Gamma_{(np)}}}$ est une valeur propre d'au moins une matrice $\underline{\Gamma_{(p)}^{(l,l)}}$ $(l = 1,\ 2,\cdots n)$.

Soit $\overrightarrow{V_{kl}}$, un vecteur propre $(n \times p)$-dimensionnel de $\underline{\widetilde{\Gamma_{(np)}}}$ pour la valeur propre λ_{kl}, nous avons :

$$\underline{\widetilde{\Gamma_{(np)}}}\ \overrightarrow{V_{kl}} = \lambda_{kl}\overrightarrow{V_{kl}}$$

ou encore

$$\begin{bmatrix} \underline{\Gamma_{(p)}^{(1,1)}} & \cdots & \underline{0_p} \\ \vdots & \ddots & \vdots \\ \underline{0_p} & \cdots & \underline{\Gamma_{(p)}^{(n,n)}} \end{bmatrix} \begin{pmatrix} \overrightarrow{V_{kl}^{(1)}} \\ \vdots \\ \overrightarrow{V_{kl}^{(n)}} \end{pmatrix} = \lambda_{kl} \begin{pmatrix} \overrightarrow{V_{kl}^{(1)}} \\ \vdots \\ \overrightarrow{V_{kl}^{(n)}} \end{pmatrix}$$

c'est à dire

$$\begin{cases} \underline{\Gamma_{(p)}^{(1,1)}}\ \overrightarrow{V_{kl}^{(1)}} = \lambda_{kl} \overrightarrow{V_{kl}^{(1)}} \\ \qquad\qquad \vdots \\ \underline{\Gamma_{(p)}^{(n,n)}}\ \overrightarrow{V_{kl}^{(n)}} = \lambda_{kl} \overrightarrow{V_{kl}^{(n)}} \end{cases}$$

Il vient la propriété suivante pour les vecteurs propres de $\underline{\widehat{\Gamma_{(np)}}}$.

Propriété 1. 12

Si $\overrightarrow{V_{kl}^{(i)}}$ est vecteur propre de $\underline{\Gamma_{(p)}^{(i,i)}}$ $(i = 1,\ 2, \cdots n)$ pour λ_{kl} valeur propre *unique*, alors le vecteur propre $\overrightarrow{V_{kl}}$ de $\underline{\widehat{\Gamma_{(np)}}}$ est unique et s'écrit :

$$\overrightarrow{V_{kl}} = \left[\ \underline{0_p} \quad \cdots \quad \overrightarrow{V_{kl}^{(i)}} \quad \cdots \quad \underline{0_p}\ \right]'$$

où, en référence à (1.28), pour $k = 0,\ 1, \cdots p-1$,

(1.35) $$\overrightarrow{V_{kl}^{(i)}}(\underline{\Gamma_{(p)}^{(ii)}}) \sim p^{-1/2}\left[\mathrm{e}^{-i(2\pi kj/p)}; j = 0, \cdots, (p-1)\right]' = \begin{bmatrix} a_{i1}^{(kl)} \\ \vdots \\ a_{ip}^{(kl)} \end{bmatrix} = \begin{bmatrix} a_{1i}^{(lk)} \\ \vdots \\ a_{pi}^{(lk)} \end{bmatrix}$$

Si λ est valeur propre multiple alors les vecteurs propres ne sont pas uniques et s'écrivent comme combinaisons linéaires des vecteurs propres précédents.

1.5.2.3 La forme des composantes principales : des résultats retrouvés

En référence à (1.34) et (1.35), une réalisation à l'instant t de $F^{(kl)}$, $(k,l)^{ième}$ composante principale des vecteurs aléatoires $\underline{Z}^{(1)},\ \underline{Z}^{(2)}, \cdots, \underline{Z}^{(n)}$ est donc :

$$f_{kl}(t) = \left[a_{l1}^{(kl)}, a_{l2}^{(kl)}, \cdots, a_{lp}^{(kl)}\right]\left[z_1^{(l)}(t), z_2^{(l)}(t), \cdots, z_p^{(l)}(t)\right]' = f_{lk}(t) \quad (1.36)$$

avec $z_j^{(l)}(t) = z_{j,l}(t) = Z_{t+(j-1)}(t)$, une réalisation à l'instant t de la variable $Z^{(j)}$ ($1 \leq j \leq p$) pour le $l^{ième}$ processus,

et $a_{ji}^{(lk)} = a_{ij}^{(kl)}$ pour $1 \leq j \leq p$ et $1 \leq i \leq n$.

Nous retrouvons le résultat établi précédemment avec la méthode qui utilise la matrice de covariance à structure bloc-Toeplitz, c'est à dire :

Propriété 1. 13

Les composantes principales (scores) de n séries temporelles « indépendantes » coïncident avec les composantes principales (scores) de chacune des séries temporelles.

Dans le cas des n processus « indépendants », les deux matrices de covariance (espace$\times$temps) et (temps$\times$espace) sont équivalentes.

Chapitre 2

2 Application aux ARMA vectoriels stationnaires

Introduction

C'est avec les *modèles linéaires ARMA* (autorégressif - moyenne mobile) que l'analyse temporelle des séries chronologiques a connu un essor considérable. Ces modèles, étudiés de façon systématique dès 1938 par Wold, se sont révélés des outils particulièrement efficaces pour l'analyse et la prévision des séries temporelles, grâce aux liens qui existent entre les modèles linéaires et la forme canonique, dite *décomposition de Wold*. L'existence et l'unicité de cette décomposition ont été établies dans [Wold38] pour les *processus stationnaires à variance finie*. Un peu plus tard, dans les années 60, Wold et Cramer montrent que cette classe de processus représente assez bien de nombreux cas concrets lorsque les hypothèses de *linéarité* et de *stationnarité* sont satisfaites.

A partir des années 70, ces modèles sont popularisés avec surtout la méthodologie dite de Box et Jenkins que nous pouvons trouver dans [Box et al.76]. C'est une méthodologie itérative qui s'inscrit dans le cadre unidimensionnel et qui permet la modélisation et la prévision des séries temporelles. Elle est basée essentiellement sur des propriétés des fonctions d'autocorrélations des modèles ARMA. En pratique, il s'agit de comparer les mêmes caractéristiques empiriques de la chronique et théoriques des processus ARMA.

Pour que ces modèles soient étudiés dans le cadre multidimensionnel, il faut attendre les années 80 avec principalement les travaux de [Granger80], [Sims80], [Tiao et al.81], [Tiao et al.83], [Hannan et al.85], [Granger et al.86] et [Hamilton94]. De même que dans le cas unidimensionnel, l'objectif de ces travaux est l'identification des processus au moyen d'équations ou de systèmes d'équations aux différences stochastiques. Les relations établies deviennent des équations matricielles de récurrence plus complexes et leur utilisation en pratique est plus difficile.

Pour l'*identification des modèles ARMA linéaires et stationnaires*, nous proposons la poursuite de l'analyse avec l'étude des composantes principales de

plusieurs de ces processus. L'objectif de ce chapitre est donc la recherche de propriétés « remarquables » des composantes principales de processus ARMA.

Dans la première partie de ce chapitre, nous nous plaçons d'emblée dans le cadre stationnaire et multidimensionnel des processus ARMA. Nous définissons le concept des modèles ARMA vectoriels et stationnaires que nous illustrons par quelques exemples.

Comme nous l'avons constaté dans le chapitre précédent, un processus multidimensionnel et stationnaire est caractérisé par sa fonction matricielle de covariance. Dans la deuxième partie de ce chapitre, nous rappelons les équations matricielles de récurrence de cette fonction pour les modèles ARMA(p,q), avec tout particulièrement les modèles AR(1), MA(1) et ARMA(1,1) vectoriels.

Puis, les résultats du chapitre 1 sont appliqués directement à la matrice de covariance à structure bloc-Toeplitz des modèles ARMA vectoriels. Nous en déduisons des approximations de ses éléments propres, dans le cas général et dans le cas particulier de modèles constitués de processus « indépendants ». Cette étude fait l'objet de la troisième partie de ce chapitre.

Dans le cadre multidimensionnel général, nous ne pouvons que constater la complexité des systèmes de récurrence obtenus. Pour l'identification de propriétés des composantes principales de modèles ARMA vectoriels, nous nous plaçons désormais uniquement dans le cas de modèles vectoriels constitués de processus « indépendants ». En référence aux résultats du chapitre 1, nous sommes ramenés à l'étude des composantes principales de chacun des processus, qui sont des *moyennes mobiles* dont les poids sont les coefficients des vecteurs propres associés. Pour les processus MA($\tilde{q}$), nous pouvons établir que les composantes principales sont des MA($p+\tilde{q}-1$) si p est le nombre de variables étudiées. Pour les AR($\tilde{p}$) et plus généralement pour les ARMA stationnaires, l'identification théorique des moyennes mobiles est alors plus délicate. Cette approche analytique des composantes principales est réalisée dans la quatrième partie de ce chapitre.

Enfin, nous renvoyons au chapitre 3 pour l'identification des composantes principales de séries temporelles, dans le contexte d'une population et dans le cadre univarié ou multivarié d'un modèle constitué de processus « indépendants ». Nous constaterons que la spécification de séries temporelles par les composantes principales est possible.

2.1 Modèles ARMA vectoriels

2.1.1 ARMA vectoriels et stationnaires

Avant d'aborder le cas d'un processus ARMA multidimensionnel, nous rappelons brièvement la définition d'un processus ARMA univarié et stationnaire.

Un processus ARMA est en fait constitué d'une partie autorégressive notée AR (AutoRegressive Process) qui est une combinaison linéaire finie en t des valeurs passées du processus, et d'une partie moyenne mobile notée MA (Moving Average Process) qui est une combinaison linéaire finie en t des valeurs passées d'un bruit blanc, c'est à dire d'un processus aléatoire formé d'une suite de variables aléatoires « indépendantes » et d'espérance mathématique nulle. Plus formellement, nous écrivons la définition suivante.

Définition 2. 1

Un processus ARMA centré $(Z_t,\ t \in \mathbb{Z})$ d'ordre (p,q) a la forme suivante :

(2.1) $$\Phi(B)Z_t = \Theta(B)a_t$$

où B est l'opérateur retard tel que $BZ_t = Z_{t-1}$ et pour tout entier b, $B^b Z_t = Z_{t-b}$,

Φ et Θ sont des polynômes en B de degré respectif p et q tel que

$$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$$

où $(\phi_1, \phi_2, \cdots, \phi_p)$ et $(\theta_1, \theta_2, \cdots, \theta_q)$ sont des réels

et $(a_t,\ t \in \mathbb{Z})$ est un bruit blanc centré de variance $\sigma_a^2 (< \infty)$.

Les propriétés du bruit blanc sont les suivantes :

$$E[a_t] = 0 \quad (t \in \mathbb{Z})$$

$$V[a_t] = \sigma_a{}^2 \quad (t \in \mathbb{Z})$$

$$\mathrm{cov}[a_t, a_{t-k}] = 0 \quad (k \neq 0)$$

L'équation (2.1) s'écrit plus simplement :

(2.2) $$Z_t - \phi_1 Z_{t-1} - \cdots - \phi_p Z_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$$

Si de plus le processus est supposé stationnaire, il vérifie la propriété suivante.

Propriété 2. 1

Le processus Z_t est *stationnaire* si et seulement si les racines de $\Phi(z)$ sont toutes de module supérieur à 1.

Nous pouvons trouver la démonstration de ce résultat dans les livres de [Box et al.76] et [Priestley81].

La stationnarité ne concerne en fait que la partie autorégressive d'un processus ARMA (p,q) *.*

Nous rappelons qu'un processus AR(1) centré tel que $Z_t = \phi_1 Z_{t-1} + a_t$ est stationnaire si et seulement si

(2.3) $$|\phi_1| < 1$$

et qu'un processus AR(2) centré tel que $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t$ est stationnaire si et seulement si :

(2.4) $$\begin{aligned} \phi_2 + \phi_1 &< 1 \\ \phi_2 - \phi_1 &< 1 \\ -1 < \phi_2 &< 1 \end{aligned}$$

Nous renvoyons par exemple à l'ouvrage de [Bourbonnais et al.98] pour les démonstrations.

Nous nous plaçons désormais dans le cadre multidimensionnel avec les modèles ARMA stationnaires. La définition qui suit est une généralisation de la définition précédente

Définition 2. 2

La représentation unidimensionnelle d'un processus ARMA (p,q), qui utilise l'opérateur de retard B ($BZ_t = Z_{t-1}$), s'étend au cas multidimensionnel d'un processus $\underline{Z}_t$ n-dimensionnel et *supposé centré*, de la façon suivante :

(2.5) $$\underline{\Phi}(B)\underline{Z}_t = \underline{\Theta}(B)\underline{a}_t$$

où

$$\underline{\Phi}(B) = \underline{I} - \underline{\Phi}_1 B - \cdots \underline{\Phi}_p B^p$$

$$\underline{\Theta}(B) = \underline{I} - \underline{\Theta}_1 B - \cdots \underline{\Theta}_q B^q$$

sont des polynômes matriciels en B, les $\underline{\Phi}$ et les $\underline{\Theta}$ des matrices $(n \times n)$ et $\underline{a}_t$ est un vecteur de bruits blancs, i.e. de moyenne nulle et de matrice de covariance $\underline{\Sigma}$.

De même, les conditions de stationnarité reposent sur la partie AR ce qui se traduit par la propriété suivante.

Propriété 2. 2

Le processus $\underline{Z}_t$ est stationnaire si et seulement si les racines du polynôme déterminant de $\underline{\Phi}(z)$, noté $\det\underline{\Phi}(z)$, sont toutes de module supérieur à 1.

La démonstration de ce résultat se trouve par exemple dans le livre de [Brockwell et al. 1991].

$\underline{Z}_t$ est appelé aussi modèle ARMA(p,q) vectoriel.

Dans un premier temps, pour simplifier les notations et pour interpréter quelques uns des éléments matriciels, trois exemples de modèles bidimensionnels sont présentés.

2.1.2 1er exemple : les modèles AR(1) vectoriels

Soit le modèle AR(1) vectoriel centré $(p = 1\ ,\ q = 0)$ comme premier exemple,

$$(\underline{I} - \underline{\Phi}_1 B)\underline{Z}_t = \underline{a}_t$$

avec pour $n = 2$

$$\underline{\Phi}_1 = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix}$$

Le modèle s'écrit encore :

$$\begin{aligned} Z_{1t} &= \phi_{11} Z_{1(t-1)} + \phi_{12} Z_{2(t-1)} + a_{1t} \\ Z_{2t} &= \phi_{21} Z_{1(t-1)} + \phi_{22} Z_{2(t-1)} + a_{2t} \end{aligned} \tag{2.6}$$

Si Z_{1t} et Z_{2t} sont les « sorties » ou variables dépendantes, et $Z_{1(t-1)}$ et $Z_{2(t-1)}$ les entrées ou variables indépendantes, alors le modèle est linéaire bidimensionnel.

2.1.3 2ième exemple : les modèles MA(1) vectoriels

Soit le modèle MA(1) vectoriel centré $(p = 0\ ,\ q = 1)$ comme second exemple,

$$\underline{Z}_t = (\underline{I} - \underline{\Theta}_1 B)\underline{a}_t$$

avec pour $n = 2$

$$\underline{\Theta}_1 = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}$$

Le modèle s'écrit encore :

$$Z_{1t} = a_{1t} - \theta_{11} a_{1(t-1)} - \theta_{12} a_{2(t-1)}$$
$$Z_{2t} = a_{2t} - \theta_{21} a_{1(t-1)} - \theta_{22} a_{2(t-1)}$$

Les processus unidimensionnels Z_{it} dépendent seulement des a_{it} et des éléments de $\underline{a}_{t-1}$.

2.1.4 3ième exemple : les modèles ARMA(1,1) vectoriels

Soit le modèle ARMA(1,1) vectoriel centré $(p = 1\ ,\ q = 1)$ comme troisième exemple,

$$(\underline{I} - \underline{\Phi}_1 B)\underline{Z}_t = (\underline{I} - \underline{\Theta}_1 B)\underline{a}_t$$

avec pour $n = 2$

$$\underline{\Phi}_1 = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} \text{ et } \underline{\Theta}_1 = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}$$

Le modèle s'écrit encore :

$$Z_{1t} = \phi_{11} Z_{1(t-1)} + \phi_{12} Z_{2(t-1)} + a_{1t} - \theta_{11} a_{1(t-1)} - \theta_{12} a_{2(t-1)}$$
$$Z_{2t} = \phi_{21} Z_{1(t-1)} + \phi_{22} Z_{2(t-1)} + a_{2t} - \theta_{21} a_{1(t-1)} - \theta_{22} a_{2(t-1)}$$

Dans ce modèle mixte, Z_{1t} et Z_{2t} dépendent des éléments de $\underline{a}_{t-1}$ et du passé de Z_{1t} et Z_{2t}. Les coefficients ϕ_{ij} et θ_{ij} reflètent à chaque fois les effets du processus j sur le processus i.

2.2 Fonction d'autocovariance de modèles ARMA vectoriels et stationnaires

Avant de traiter directement des composantes principales d'un modèle ARMA vectoriel, nous rappelons dans cette partie des propriétés de sa fonction matricielle de covariance, en commençant par les modèles AR(1) puis MA(1) et ARMA(1,1), pour finalement écrire le système d'équations matricielles de récurrence d'un ARMA(p,q) vectoriel.

2.2.1 Modèles AR(1) vectoriels

Propriété 2. 3

Pour un modèle vectoriel AR(1), *stationnaire et centré*, la matrice de covariance décalée au retard l, $\underline{\Gamma}_{(n)}(l)$ ($l = 0,1,2,\cdots$), s'écrit plus simplement :

(2.7)
$$\begin{aligned} \underline{\Gamma}_{(n)}(0) &= \underline{\Sigma} + \underline{\Phi}_1\, \underline{\Gamma}_{(n)}(0)\, \underline{\Phi}_1{}' \\ \underline{\Gamma}_{(n)}(l) &= \underline{\Gamma}_{(n)}(0)\, [(\underline{\Phi}_1)^l]'\,(l = 1,2,\cdots) \end{aligned}$$

Démonstration :

En référence à (1.1) et (1.2), la fonction de covariance d'un processus stationnaire et centré vérifie :

$$\underline{\Gamma}_{(n)}(l) = \{\gamma_{ij}(l)\} = E\left[\underline{Z}_t \underline{Z}_{t+l}{}'\right] \text{ et } \underline{\Gamma}_{(n)}{}'(l) = \underline{\Gamma}_{(n)}(-l)\ (l \in \mathbb{Z})$$

Pour démontrer (2.7), prenons l'expression $\underline{Z}_t - \underline{\Phi}_1 \underline{Z}_{t-1} = \underline{a}_t$ et multiplions de chaque côté de l'équation par $\underline{Z}'_{t-l}$, nous obtenons :

$$\underline{Z}_t \underline{Z}'_{t-l} - \underline{\Phi}_1 \underline{Z}_{t-1} \underline{Z}'_{t-l} = \underline{a}_t \underline{Z}'_{t-l}\ (l = 0,1,2,\cdots)$$

En prenant les espérances mathématiques de chaque côté, nous avons :

$$\underline{\Gamma}_{(n)}(0) - \underline{\Phi}_1 \underline{\Gamma}_{(n)}(1) = \underline{\Sigma}$$

$$\underline{\Gamma}_{(n)}(-1) - \underline{\Phi}_1 \underline{\Gamma}_{(n)}(0) = \underline{0}$$

$$\underline{\Gamma}_{(n)}(-l) - \underline{\Phi}_1 \underline{\Gamma}_{(n)}(-l+1) = \underline{0}\ (l > 1)$$

sachant que $E(\underline{a}_{t-i}\underline{a}'_{t-j}) = \underline{\Sigma}$ si $i = j$, et $E(\underline{a}_{t-i}\underline{a}'_{t-j}) = \underline{0}$ sinon.

Par ailleurs, notons que $\underline{Z}_t$ peut s'écrire :

$$\begin{aligned}\underline{Z}_t &= (\underline{I} - \underline{\Phi}_1 B)^{-1}\underline{a}_t \\ &= (\underline{I} + \underline{\Phi}_1 B + \underline{\Phi}_1{}^2 B^2 + \underline{\Phi}_1{}^3 B^3 + \cdots)\underline{a}_t\end{aligned}$$

(décomposition qui se trouve par exemple dans [LutKepolh 1993])

Par conséquent : $E(\underline{a}_t\underline{Z}'_{t-l}) = \underline{\Sigma}$ si $l = 0$, $E(\underline{a}_t\underline{Z}'_{t-l}) = \underline{0}$, si $l \geq 1$.

Il s'ensuit : $\underline{\Gamma}_{(n)}(-l) = \underline{\Phi}_1\,\underline{\Gamma}_{(n)}(-(l-1))$ $(l = 1,2,\cdots)$

Pour $\underline{\Gamma}_{(n)}(0)$, il suffit d'écrire : $\underline{\Gamma}_{(n)}(0) = \underline{\Sigma} + \underline{\Phi}_1\underline{\Gamma}_{(n)}(1)$ et $\underline{\Gamma}_{(n)}(1) = \underline{\Gamma}_{(n)}(0)\,\underline{\Phi}_1{}'$.

Puis par substitution de $\underline{\Gamma}_{(n)}(1)$, nous obtenons $\underline{\Gamma}_{(n)}(0) = \underline{\Sigma} + \underline{\Phi}_1\,\underline{\Gamma}_{(n)}(0)\,\underline{\Phi}_1{}'$.

Enfin par substitution successive, nous avons :

$$\underline{\Gamma}_{(n)}(-l) = (\underline{\Phi}_1)^l\,\underline{\Gamma}_{(n)}(0) \quad (l = 1,2,\cdots)$$

c'est à dire

$$\underline{\Gamma}_{(n)}(l) = \underline{\Gamma}_{(n)}(0)[(\underline{\Phi}_1)^l]' \quad (l = 1,2,\cdots)$$

Le système d'équations matricielles en (2.7) est démontré ; il a comme point de départ $\underline{\Gamma}_{(n)}(0)$.

Fin démonstration.

Nous retrouvons des similitudes avec la fonction d'autocovariance des processus AR(1) unidimensionnels et stationnaires.

2.2.2 Modèles MA(1) vectoriels

Propriété 2. 4

Pour un modèle MA(1) vectoriel, stationnaire et centré, la matrice de covariance décalée au retard l, $\underline{\Gamma}_{(n)}(l)$ $(l = 0,1,2,\cdots)$ s'écrit plus simplement :

$$
\begin{aligned}
&\underline{\Gamma}_{(n)}(0) = \underline{\Sigma} + \underline{\Theta}_1 \, \underline{\Sigma} \, \underline{\Theta}_1^{'} \\
(2.8) \qquad &\underline{\Gamma}_{(n)}(1) = -\underline{\Sigma} \, \underline{\Theta}_1^{'} \\
&\underline{\Gamma}_{(n)}(l) = \underline{0} \ (l > 1)
\end{aligned}
$$

Démonstration :

Comme précédemment, nous utilisons les résultats en (1.1) et (1.2).

Pour démontrer (2.8), prenons l'expression $\underline{Z}_t = \underline{a}_t - \underline{\Theta}_1 \underline{a}_{t-1}$.

D'où :

$$
\begin{aligned}
&\underline{\Gamma}_{(n)}(0) = E(\underline{Z}_t \underline{Z}^{'}_t) = E\left[(\underline{a}_t - \underline{\Theta}_1 \underline{a}_{t-1})(\underline{a}_t - \underline{\Theta}_1 \underline{a}_{t-1})^{'}\right] \\
&= E(\underline{a}_t \underline{a}^{'}_t - \underline{a}_t \underline{a}^{'}_{t-1} \underline{\Theta}^{'}_1 - \underline{\Theta}_1 \underline{a}_{t-1} \underline{a}^{'}_t + \underline{\Theta}_1 \underline{a}_{t-1} \underline{a}^{'}_{t-1} \underline{\Theta}^{'}_1) \\
&= \underline{\Sigma} + \underline{\Theta}_1 \, \underline{\Sigma} \, \underline{\Theta}_1^{'}
\end{aligned}
$$

sachant que $E(\underline{a}_{t-i} \underline{a}^{'}_{t-j}) = \underline{\Sigma}$ si $i = j$, et $E(\underline{a}_{t-i} \underline{a}^{'}_{t-j}) = \underline{0}$ sinon.

De même :

$$
\begin{aligned}
&\underline{\Gamma}_{(n)}(-1) = E(\underline{Z}_t \underline{Z}^{'}_{t-1}) = E\left[(\underline{a}_t - \underline{\Theta}_1 \underline{a}_{t-1})(\underline{a}_{t-1} - \underline{\Theta}_1 \underline{a}_{t-2})^{'}\right] \\
&= E(\underline{a}_t \underline{a}^{'}_{t-1} - \underline{a}_t \underline{a}^{'}_{t-2} \underline{\Theta}^{'}_1 - \underline{\Theta}_1 \underline{a}_{t-1} \underline{a}^{'}_{t-1} + \underline{\Theta}_1 \underline{a}_{t-1} \underline{a}^{'}_{t-2} \underline{\Theta}^{'}_1) \\
&= -\underline{\Theta}_1 \underline{\Sigma}
\end{aligned}
$$

c'est à dire $\underline{\Gamma}_{(n)}(1) = -\underline{\Sigma} \, \underline{\Theta}_1^{'}$

et

$$\underline{\Gamma}_n(-l) = E(\underline{Z}_t \underline{Z}^{'}_{t-l}) = \underline{0} = \underline{\Gamma}_{(n)}(l) \ (l > 1)$$

Le système d'équations matricielles en (2.8) est démontré.

Fin démonstration.

Pour un retard plus grand que 1, les matrices de covariance sont nulles. Nous rappelons qu'il en va de même avec la fonction d'autocovariance des MA(1) unidimensionnels.

2.2.3 Modèles ARMA(1,1) vectoriels

Propriété 2. 5

Pour un modèle ARMA(1,1) vectoriel, stationnaire et centré, la matrice de covariance décalée au retard l, $\underline{\Gamma}_{(n)}(l)$ $(l = 0,1,2,\cdots)$ s'écrit plus simplement :

(2.9)
$$\begin{aligned}
&\underline{\Gamma}_{(n)}(0) = \underline{\Sigma} + \underline{\Phi}_1\, \underline{\Gamma}_{(n)}(0)\, \underline{\Phi}_1{}' + (\underline{\Theta}_1 - \underline{\Phi}_1)\underline{\Sigma}(\underline{\Theta}_1 - \underline{\Phi}_1)' - \underline{\Theta}_1\, \underline{\Sigma}\, \underline{\Theta}_1{}' \\
&\underline{\Gamma}_{(n)}(1) = \underline{\Gamma}_{(n)}(0)\underline{\Phi}_1{}' - \underline{\Sigma}\underline{\Theta}_1{}' \\
&\underline{\Gamma}_{(n)}(l) = \underline{\Gamma}_{(n)}(l-1)\underline{\Phi}_1{}' \ (l > 1)
\end{aligned}$$

Démonstration :

Comme précédemment, nous utilisons les résultats en (1.1) et (1.2).

Pour démontrer (2.9), prenons l'expression $\underline{Z}_t - \underline{\Phi}_1\underline{Z}_{t-1} = \underline{a}_t - \underline{\Theta}_1\underline{a}_{t-1}$, multiplions de chaque côté de l'équation par $\underline{Z}'_{t-l}$ et calculons l'espérance mathématique, nous obtenons :

$$E(\underline{Z}_t - \underline{\Phi}_1\underline{Z}_{t-1})\underline{Z}'_{t-l} = E(\underline{a}_t - \underline{\Theta}_1\underline{a}_{t-1})\underline{Z}'_{t-l} \ (l = 0,1,2,\cdots)$$

Par ailleurs, notons que $\underline{Z}_t$ peut s'écrire :

(2.10)
$$\begin{aligned}
\underline{Z}_t &= (\underline{I} - \underline{\Phi}_1 B)^{-1}(\underline{I} - \underline{\Theta}_1 B)\underline{a}_t \\
&= \left[\underline{I} + (\underline{\Phi}_1 - \underline{\Theta}_1)B + \underline{\Phi}_1(\underline{\Phi}_1 - \underline{\Theta}_1)B^2 + \underline{\Phi}^2{}_1(\underline{\Phi}_1 - \underline{\Theta}_1)B^3 + \cdots\right]\underline{a}_t
\end{aligned}$$

avec $(\underline{I} - \underline{\Phi}_1 B)^{-1} = (\underline{I} + \underline{\Phi}_1 B + \underline{\Phi}_1{}^2 B^2 + \underline{\Phi}_1{}^3 B^3 + \cdots)$

Par conséquent,

$$E(\underline{a}_t\underline{Z}'_{t-l}) = \underline{\Sigma}\ \underline{\Psi}'_l \ (l = 0,1,2,\cdots)$$

où

$$\begin{aligned}
\underline{\Psi}_0 &= \underline{I} \\
\underline{\Psi}_1 &= \underline{\Phi}_1 - \underline{\Theta}_1 \\
\underline{\Psi}_2 &= \underline{\Phi}_1(\underline{\Phi}_1 - \underline{\Theta}_1) \\
\underline{\Psi}_3 &= \underline{\Phi}^2{}_1(\underline{\Phi}_1 - \underline{\Theta}_1) \\
&\vdots
\end{aligned}$$

En utilisant ces résultats, nous obtenons :

$$\underline{\Gamma}_{(n)}(0) - \underline{\Phi}_1\,\underline{\Gamma}_{(n)}(1) = \underline{\Sigma} - \underline{\Theta}_1\,\underline{\Sigma}\,\underline{\Psi}_1{}' = \underline{\Sigma} - \underline{\Theta}_1\underline{\Sigma}(\underline{\Phi}_1 - \underline{\Theta}_1)'$$

$$\underline{\Gamma}_{(n)}(-1) - \underline{\Phi}_1\underline{\Gamma}_{(n)}(0) = -\underline{\Theta}_1\,\underline{\Sigma}$$

$$\underline{\Gamma}_{(n)}(-l) - \underline{\Phi}_1\,\underline{\Gamma}_{(n)}(-(l-1)) = \underline{0}\,, \quad l > 1$$

En remplaçant par ailleurs $\underline{\Gamma}_{(n)}(1)$ par $\underline{\Gamma}(0)\underline{\Phi}'_1 - \underline{\Sigma}\underline{\Theta}'_1$, les équations sur les moments matriciels sont retrouvées.

Le système d'équations matricielles en (2.9) est démontré ; il a comme point de départ $\underline{\Gamma}_{(n)}(0)$.

Fin démonstration.

Les matrices de covariance pour $l > 1$ et pour un vecteur ARMA(1,1) sont similaires à celles d'un vecteur AR(1), comme dans le cas unidimensionnel.

2.2.4 Modèles ARMA(p,q) vectoriels

Les résultats précédents peuvent s'étendre au cas d'un modèle ARMA(p,q) vectoriel, centré et stationnaire, comme défini en (2.5). Nous obtenons alors un système d'équations matricielles de récurrence que nous pouvons trouver par exemple dans [Tiao et al.81].

Propriété 2. 6

Les matrices de covariances décalées $\underline{\Gamma}_{(n)}(l)$ $(l = 0,1,2,\cdots)$ d'un modèle ARMA(p,q) vectoriel, centré et stationnaire, vérifient

$$(2.11) \qquad \underline{\Gamma}_{(n)}(l) - \sum_{i=1}^{p} \underline{\Gamma}_{(n)}(l-i)\, \underline{\Phi}_i^{'} = \begin{cases} -\sum_{i=l}^{q} \underline{\Psi}_{i-l} \underline{\Sigma}\, \underline{\Theta}_i^{'} & (l = 0, \cdots, q) \\ \underline{0} & (l > q) \end{cases}$$

où les racines $\underline{\Psi}_j$ sont obtenues à partir de la relation

$$\psi(B) = \underline{\Phi}^{-1}(B)\, \underline{\Theta}(B) = (\underline{I} + \underline{\Psi}_1 B + \cdots),$$

$\underline{\Theta}_0 = -I$ et sachant que (a) si $p < q$, $\underline{\Phi}_{p+1} = \cdots = \underline{\Phi}_m = \underline{0}$, et (b) si $q < p$, $\underline{\Theta}_{q+1} = \cdots = \underline{\Theta}_m = \underline{0}$.

Démonstration

La démonstration utilise la définition suivante d'un processus ARMA vectoriel.

Définition 2. 3

Si $(\underline{Z}_t,\ t \in \mathbb{Z})$ est un processus ARMA vectoriel satisfaisant (2.5), c'est à dire $\underline{\Phi}(B)\underline{Z}_t = \underline{\Theta}(B)\underline{a}_t$ avec un polynôme det $\underline{\Phi}(z)$ admettant des racines dont le module est plus grand que 1, alors $(\underline{Z}_t,\ t \in \mathbb{Z})$ s'écrit sous forme d'une moyenne mobile infinie

$$\underline{Z}_t = \sum_{i=0}^{\infty} \underline{\Psi}_i \underline{a}_{t-i}$$

avec

$$\underline{\Psi}_0 = \underline{I}\,,\ \sum_{i=1}^{\infty} \left\| \underline{\Psi}_i \right\| < \infty$$

et où les racines $\underline{\Psi}_i$ sont obtenues de la relation :

$$\psi(B) = \underline{\Phi}^{-1}(B)\, \underline{\Theta}(B) = (\underline{I} + \underline{\Psi}_1 B + \underline{\Psi}_2 B^2 \cdots).$$

Il vient que la covariance entre $\underline{Z}_{t-l}$ et $\underline{Z}_t - \sum_{i=1}^{p} \underline{\Phi}_i \underline{Z}_{t-i}$ est donnée par

$$Cov\left[\underline{Z}_{t-l}, \underline{Z}_t - \sum_{i=1}^{p} \underline{\Phi}_i \underline{Z}_{t-i}\right] = -\sum_{i=0}^{q} Cov[\underline{Z}_{t-l}, \underline{a}_{t-i}]\, \Theta_i^{'}$$

$$= -\sum_{i=0}^{q} Cov\left[\sum_{i=0}^{\infty} \underline{\Psi}_i \underline{a}_{t-l-i}, \underline{a}_{t-i}\right] \Theta_i^{'}$$

avec la convention $\underline{\Theta}_0 = -I$.

En référence à (1.1) et (1.2), la fonction d'autocovariance du processus $\underline{Z}_t$ vérifie donc (2.11).

Fin démonstration.

Ces relations permettent, au moins en théorie, de trouver l'expression de la fonction d'autocovariance en fonction des matrices $\underline{\Phi}_i$ et $\underline{\Theta}_i$ qui interviennent dans les polynômes autorégressif et moyenne mobile.

Ecrites pour $l = 0,1,\ldots,p$, les équations (2.11) constituent un système linéaire en $\underline{\Gamma_{(n)}}(0), \underline{\Gamma_{(n)}}(1),\ldots, \underline{\Gamma_{(n)}}(p)$ qui peut être résolu pour obtenir ces quantités. Les autres $\underline{\Gamma_{(n)}}(l)$ sont alors déduits de ces valeurs initiales en utilisant (2.11) pour $l > p$.

Les matrices $\underline{\Gamma_{(n)}}(l)$ satisfont un système de récurrence linéaire.

Dans le cas particulier d'un MA(q), nous pouvons constater que la fonction d'autocovariance s'annule à partir du rang $q+1$, c'est à dire la même propriété que pour la fonction d'autocovariance de chacun des processus composant $(\underline{Z}_t)$.

2.3 Approximation des éléments propres de $\widetilde{\underline{\Gamma_{(pn)}}}$ pour des ARMA vectoriels

Tenant compte des résultats précédents, nous pouvons approcher les éléments propres de la matrice de covariance définie en (1.9). Tout d'abord, nous travaillons avec les modèles AR(1), puis MA(1) et ARMA(1,1) vectoriels, en traitant à chaque fois du cas général et du cas particulier des modèles vectoriels constitués de processus « indépendants ». Comme cas particulier pour les ARMA(1,1) vectoriels, nous construisons un modèle constitué de n_1 processus AR(1) et n_2 processus MA(1), « indépendants » et stationnaires. Enfin, nous abordons le cas des ARMA(p,q) vectoriels qui se révèle très complexe à utiliser.

2.3.1 Modèles AR(1) vectoriels

2.3.1.1 Le cas général

Nous nous plaçons dans le cas d'un modèle AR(1) vectoriel, centré et stationnaire, où les matrices $\underline{\Phi}_i$ et $\underline{\Theta}_i$ qui interviennent dans les polynômes autorégressif et moyenne mobile sont quelconques.

Soit $(\underline{Z}_t,\ t \in \mathbb{Z})$ un modèle AR(1) vectoriel n-dimensionnel, centré et stationnaire, il s'écrit

$$(\underline{I} - \underline{\Phi}_1 B)\underline{Z}_t = \underline{a}_t$$

ou

$$\begin{aligned}
Z_{1t} &= \phi_{11}Z_{1(t-1)} + \phi_{12}Z_{2(t-1)} + \cdots + \phi_{1n}Z_{n(t-1)} + a_{1t} \\
Z_{2t} &= \phi_{21}Z_{1(t-1)} + \phi_{22}Z_{2(t-1)} + \cdots + \phi_{2n}Z_{n(t-1)} + a_{2t} \\
&\vdots \\
Z_{nt} &= \phi_{n1}Z_{1(t-1)} + \phi_{n2}Z_{2(t-1)} + \cdots + \phi_{nn}Z_{n(t-1)} + a_{nt}
\end{aligned}$$

avec

$$\underline{\Phi}_1 = \begin{pmatrix} \phi_{11} & & \phi_{n1} \\ & \ddots & \\ \phi_{n1} & & \phi_{nn} \end{pmatrix} \text{ et } \underline{\Sigma} = \begin{pmatrix} \sigma^2_{a_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_n} \end{pmatrix}$$

Les éléments propres de la matrice de covariance définie en (1.9) vérifient les propriétés asymptotiques suivantes.

Propriété 2. 7 : cas général

Pour un modèle AR(1) vectoriel, centré et stationnaire, les éléments propres de la matrice $\underline{\widehat{\Gamma_{(pn)}}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres de $\underline{\widehat{\Gamma_{(pn)}}}$ se rapprochent des valeurs propres de

$$\begin{aligned} &\underline{\Gamma_{(n)}}(0) + \sum_{j=1}^{p-1} \underline{\Gamma_n}(0)[(\underline{\Phi}_1)^j]' \, \mathrm{e}^{-i(2\pi jk/p)} \\ &= \sum_{j=0}^{p-1} \underline{\Gamma_{(n)}}(0)[(\underline{\Phi}_1)^j]' \, \mathrm{e}^{-i(2\pi jk/p)} \end{aligned} \qquad (k = 0,\ 1, \cdots p-1) \tag{2.12}$$

avec comme point de départ

$$\underline{\Gamma_{(n)}}(0) = \underline{\Sigma} + \underline{\Phi}_1 \, \underline{\Gamma_{(n)}}(0) \, \underline{\Phi}_1{}' \tag{2.13}$$

(b) et les vecteurs propres de $\widehat{\underline{\Gamma_{(pn)}}}$ se rapprochent de

$$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)} \, \underline{u}_{lk}; j = 0, \cdots, p-1\right] \ (k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1)$$

où $\underline{u}_{lk}$ sont les vecteurs propres de (2.12).

Démonstration :

La démonstration est immédiate. Pour cela, il suffit de combiner les résultats en (1.21) et (2.7) pour les valeurs propres et de s'en référer à (1.22) pour les vecteurs propres.

Fin démonstration.

2.3.1.2 Le cas des processus « indépendants »

Nous nous plaçons désormais dans le cas des processus non-corrélés. Les coefficients ϕ_{ij} sont nuls pour $i \neq j$ (le processus i n'étant pas influencé par le processus j pour $i \neq j$). Les matrices $\underline{\Phi}_1$ et $\underline{\Gamma_{(n)}}(l)$ s'écrivent alors plus simplement :

$$\underline{\Phi}_1 = \begin{pmatrix} \phi_{11} & & 0 \\ & \ddots & \\ 0 & & \phi_{nn} \end{pmatrix}$$

$$\underline{\Gamma_{(n)}}(l) = \begin{pmatrix} \gamma_{11}(l) & & 0 \\ & \ddots & \\ 0 & & \gamma_{nn}(l) \end{pmatrix} (l = 0,1,2,\cdots)$$

L'expression des valeurs propres est simplifiée avec comme point de départ (2.13), c'est à dire :

$$\underline{\Gamma_{(n)}}(0) = \underline{\Sigma} + \underline{\Phi}_1 \, \underline{\Gamma_{(n)}}(0) \, \underline{\Phi}_1^{'}$$

ou encore

$$\begin{pmatrix} \gamma_{11}(0) & & 0 \\ & \ddots & \\ 0 & & \gamma_{nn}(0) \end{pmatrix} = \begin{pmatrix} \sigma_{a_1}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{a_n}^2 \end{pmatrix} + \begin{pmatrix} \phi_{11} & & 0 \\ & \ddots & \\ 0 & & \phi_{nn} \end{pmatrix} \begin{pmatrix} \gamma_{11}(0) & & 0 \\ & \ddots & \\ 0 & & \gamma_{nn}(0) \end{pmatrix} \begin{pmatrix} \phi_{11} & & 0 \\ & \ddots & \\ 0 & & \phi_{nn} \end{pmatrix}^{'}$$

$$\begin{aligned} \gamma_{11}(0) &= \sigma_{a_1}^2 + {\phi_{11}}^2 \gamma_{11}(0) \\ &\vdots \\ \gamma_{nn}(0) &= \sigma_{a_n}^2 + {\phi_{nn}}^2 \gamma_{nn}(0) \end{aligned}$$

c'est à dire

$$\begin{aligned} \gamma_{11}(0) &= \frac{\sigma_{a_1}^2}{1 - {\phi_{11}}^2} \\ &\vdots \\ \gamma_{nn}(0) &= \frac{\sigma_{a_n}^2}{1 - {\phi_{nn}}^2} \end{aligned}$$

La matrice $\underline{\Gamma_{(n)}}(0)$ s'écrit donc :

$$\underline{\Gamma_{(n)}}(0) = \begin{pmatrix} \dfrac{\sigma_{a_1}^2}{1-\phi_{11}{}^2} & & 0 \\ & \ddots & \\ 0 & & \dfrac{\sigma_{a_n}^2}{1-\phi_{nn}{}^2} \end{pmatrix}$$

En référence à (1.23), les valeurs propres de $\widehat{\underline{\Gamma_{(pn)}}}$ peuvent donc être approchées par les valeurs propres réelles de la matrice :

(2.14)
$$\begin{pmatrix} \displaystyle\sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)} \dfrac{\sigma_{a_1}^2 \phi_{11}{}^j}{1-\phi_{11}{}^2} & & 0 \\ & \ddots & \\ 0 & & \displaystyle\sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)} \dfrac{\sigma_{a_n}^2 \phi_{nn}{}^j}{1-\phi_{nn}{}^2} \end{pmatrix} (k = 0,\ 1, \cdots p-1)$$

Pour les vecteurs propres, il suffit de reprendre les résultats établis en (1.31) et (1.32) pour des processus « indépendants » et il s'ensuit la propriété suivante.

Propriété 2. 8 : indépendance des processus

Dans le cas de n processus « indépendants » AR(1), centrés et stationnaires, les éléments propres de la matrice $\widehat{\underline{\Gamma_{(pn)}}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres réelles de la matrice $\widehat{\underline{\Gamma_{(pn)}}}$ sont approximativement :

(2.15)
$$\frac{\sigma_{a_m}^2}{1-\phi_{mm}{}^2} \sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)} \phi_{mm}{}^j \quad (k = 0,\ 1, \cdots p-1,\ m = 1, \cdots, n)$$

(b) et les vecteurs propres de $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent de :

(2.16) $$\overrightarrow{V_{lk}} \sim p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\,\underline{e}_l; j=0,\cdots,(p-1)\right]$$
$$(k=0,\,1,\cdots p-1\,,\,l=0,\,1,\cdots n-1)$$

$$\text{avec } \overrightarrow{V_{lk}} = \left[(a_{ji}^{(lk)});1\le j\le p, 1\le i\le n)\right]'$$
$$\text{où } a_{ji}^{(lk)} \sim 0 \ (i\neq l) \text{ et } a_{jl}^{(lk)} \sim p^{-1/2}\,\mathrm{e}^{-i(2\pi(j-1)k/p)} \ (1\le j\le p)$$

Nous retrouvons le résultat établi dans le chapitre 1 (dernière section) : les valeurs propres définies en (2.15) coïncident avec les valeurs propres de chacun des processus AR(1) unidimensionnels.

En effet, les valeurs propres du $m^{ième}$ processus AR(1) unidimensionnel sont approximativement :

(2.17) $$\sum_{j=0}^{p-1}\mathrm{e}^{-i(2\pi jk/p)}\,\frac{\sigma_{a_m}^2}{1-{\phi_{mm}}^2}\,{\phi_{mm}}^j\,,\ k=0,\,1,\cdots p-1$$

avec $\gamma(j)=\dfrac{\sigma_{a_m}^2}{1-{\phi_{mm}}^2}\,{\phi_{mm}}^j$ $(j=0,\,1,\cdots p-1)$, la fonction de covariance du processus AR(1) centré et stationnaire.

2.3.2 Modèles MA(1) vectoriels

2.3.2.1 Le cas général

Dans ce qui suit, le processus $(\underline{Z}_t,\ t\in\mathbb{Z})$ est un modèle MA(1) vectoriel n-dimensionnel, centré et stationnaire. Il vérifie :

$$\underline{Z}_t = (\underline{I}-\underline{\Theta}_1 B)\underline{a}_t$$

ou

$$\begin{aligned}
Z_{1t} &= a_{1t} + \theta_{11}a_{1(t-1)} + \theta_{12}a_{2(t-1)} + \cdots + \theta_{1n}a_{n(t-1)}\\
Z_{2t} &= a_{2t} + \theta_{21}a_{1(t-1)} + \theta_{22}a_{2(t-1)} + \cdots + \theta_{2n}a_{n(t-1)}\\
&\vdots\\
Z_{nt} &= a_{nt} + \theta_{n1}a_{1(t-1)} + \theta_{n2}a_{2(t-1)} + \cdots + \theta_{nn}a_{n(t-1)}
\end{aligned}$$

avec

$$\underline{\Theta}_1 = \begin{pmatrix} \theta_{11} & & \theta_{1n} \\ \vdots & \ddots & \vdots \\ \theta_{n1} & & \theta_{nn} \end{pmatrix} \quad \text{et} \quad \underline{\Sigma} = \begin{pmatrix} \sigma^2_{a_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_n} \end{pmatrix}.$$

Propriété 2. 9 : cas général

Pour un MA(1) vectoriel n-dimensionnel, centré et stationnaire, les éléments propres de la matrice $\underline{\widehat{\Gamma_{(pn)}}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres de la matrice $\underline{\widehat{\Gamma_{(pn)}}}$ se rapprochent des valeurs propres de :

$$\underline{\Gamma_{(n)}}(0) - \mathrm{e}^{-i(2\pi k/p)}\, \underline{\Sigma}\, \underline{\Theta}_1{}' \quad (k = 0,\ 1, \cdots p-1) \tag{2.18}$$

avec comme point de départ

$$\underline{\Gamma_{(n)}}(0) = \underline{\Sigma} + \underline{\Theta}_1\, \underline{\Sigma}\, \underline{\Theta}_1{}'$$

(b) et les vecteurs propres peuvent être approchés par :

$$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\, \underline{u}_{lk}; j = 0, \cdots, p-1\right] \ (k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1)$$

où $\underline{u}_{lk}$ sont les vecteurs propres de (2.18).

Démonstration :

La démonstration est immédiate. Pour cela, il suffit de combiner les résultats en (1.21) et (2.8) pour les valeurs propres et de s'en référer à (1.22) pour les vecteurs propres.

Fin démonstration.

2.3.2.2 Le cas des processus « indépendants »

Dans le cas particulier des processus non-corrélés où les coefficients θ_{ij} sont nuls pour $i \neq j$, c'est à dire que la série i n'est pas influencée par la série j pour $i \neq j$, les matrices $\underline{\Theta_1}$ et $\underline{\Gamma_{(n)}}(l)$ s'écrivent plus facilement :

$$\underline{\Theta_1} = \begin{pmatrix} \theta_{11} & & 0 \\ & \ddots & \\ 0 & & \theta_{nn} \end{pmatrix}$$

$$\underline{\Gamma_{(n)}}(l) = \begin{pmatrix} \gamma_{11}(l) & & 0 \\ & \ddots & \\ 0 & & \gamma_{nn}(l) \end{pmatrix} (l = 0,1,2,\cdots)$$

Comme $\underline{\Gamma_{(n)}}(0) = \underline{\Sigma} + \underline{\Theta_1}\,\underline{\Sigma}\,\underline{\Theta_1}^{'}$, c'est à dire

$$\begin{pmatrix} \gamma_{11}(0) & & 0 \\ & \ddots & \\ 0 & & \gamma_{k'k'}(0) \end{pmatrix} = \begin{pmatrix} \sigma^2_{a_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_{k'}} \end{pmatrix} + \begin{pmatrix} \theta_{11} & & 0 \\ & \ddots & \\ 0 & & \theta_{k'k'} \end{pmatrix} \begin{pmatrix} \sigma^2_{a_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_{k'}} \end{pmatrix} \begin{pmatrix} \theta_{11} & & 0 \\ & \ddots & \\ 0 & & \theta_{k'k'} \end{pmatrix}^{'}$$

nous obtenons :

$$\underline{\Gamma_{(n)}}(0) = \begin{pmatrix} \sigma^2_{a_1}(1+\theta_{11}{}^2) & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_{k'}}(1+\theta_{k'k'}{}^2) \end{pmatrix}$$

Par ailleurs

$$-\underline{\Sigma}\,\underline{\Theta_1}^{'} = \begin{pmatrix} -\theta_{11}\sigma^2_{a_1} & & 0 \\ & \ddots & \\ 0 & & -\theta_{kk}\sigma^2_{a_k} \end{pmatrix}$$

En référence à (2.18), les valeurs propres de $\widehat{\underline{\Gamma}_{(pn)}}$ peuvent donc être approchées par les valeurs propres réelles de la matrice :

$$(2.19)\qquad \begin{pmatrix} \sigma_{a_1}^2(1+\theta_{11}{}^2)-\theta_{11}\sigma_{a_1}^2\,\mathrm{e}^{-i(2\pi k/p)} & & 0 \\ & \ddots & \\ 0 & & \sigma_{a_n}^2(1+\theta_{nn}{}^2)-\theta_{nn}\sigma_{a_n}^2\,\mathrm{e}^{-i(2\pi k/p)} \end{pmatrix}$$

Pour les vecteurs propres, il suffit de s'en référer à (1.31) et (1.32) et il s'ensuit la propriété suivante.

Propriété 2. 10 : « indépendance » des processus

Dans le cas de n processus non-corrélés MA(1), centrés et stationnaires, les éléments propres de la matrice $\widehat{\underline{\Gamma}_{(pn)}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres *réelles* de $\widehat{\underline{\Gamma}_{(pn)}}$ sont approximativement :

$$(2.20)\qquad \sigma_{a_m}^2(1+\theta_{mm}{}^2)-\theta_{mm}\sigma_{a_m}^2\,\mathrm{e}^{-i(2\pi k/p)}\,(k=0,\,1,\cdots p-1\,,\;m=1,\cdots,n\,)$$

(b) et les vecteurs propres de $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent de :

$$(2.21)\qquad \overrightarrow{V_{lk}} \sim p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\,\underline{e}_l; j=0,\cdots,(p-1)\right]$$
$$(k=0,\,1,\cdots p-1\,,\;l=0,\,1,\cdots n-1\,)$$

$$\text{avec } \overrightarrow{V_{lk}} = \left[(a_{ji}^{(lk)});1\le j\le p, 1\le i\le n)\right]'$$
$$\text{où } a_{ji}^{(lk)} \sim 0 \;(i\neq l\,) \text{ et } a_{jl}^{(lk)} \sim p^{-1/2}\,\mathrm{e}^{-i(2\pi(j-1)k/p)} \;(1\le j\le p\,)$$

De même, les valeurs propres obtenues coïncident avec les valeurs propres de chacun des processus MA(1) unidimensionnel.

En effet, les valeurs propres du $m^{ième}$ processus MA(1) unidimensionnel sont approximativement :

(2.22) $$\sigma^2_{a_m}(1+{\theta_{mm}}^2)-\theta_{mm}\sigma^2_{a_m}\,\mathrm{e}^{-i(2\pi k/p)},\ k=0,\,1,\cdots p-1$$

avec $\gamma(0)=\sigma^2_{a_m}(1+{\theta_{mm}}^2)$, $\gamma(1)=-\theta_{mm}\sigma^2_{a_m}$ et $\gamma(j)=0$ pour $j>1$, c'est à dire la fonction de covariance du processus MA(1) centré et stationnaire.

2.3.3 Modèles ARMA(1,1) vectoriels

2.3.3.1 Le cas général

Dans ce qui suit, le processus $(\underline{Z}_t,\ t\in\mathbb{Z})$ est un modèle ARMA(1,1) vectoriel $(p=1\ ,\ q=1)$ n-dimensionnel, centré et stationnaire de la forme :

$$(\underline{I}-\underline{\Phi}_1B)\underline{Z}_t=(\underline{I}-\underline{\Theta}_1B)\underline{a}_t$$

avec

$$\underline{\Phi}_1=\begin{pmatrix}\phi_{11} & & \phi_{n1}\\ & \ddots & \\ \phi_{n1} & & \phi_{nn}\end{pmatrix},\ \underline{\Theta}_1=\begin{pmatrix}\theta_{11} & & \theta_{1n}\\ \vdots & \ddots & \vdots\\ \theta_{n1} & & \theta_{nn}\end{pmatrix}\ \text{et}\ \underline{\Sigma}=\begin{pmatrix}\sigma^2_{a_1} & & 0\\ & \ddots & \\ 0 & & \sigma^2_{a_n}\end{pmatrix}.$$

Les résultats en (2.9) peuvent s'écrire aussi :

$$\underline{\Gamma_{(n)}}(0)=\underline{\Sigma}+\underline{\Phi}_1\,\underline{\Gamma_{(n)}}(0)\,{\underline{\Phi}_1}^{'}+(\underline{\Theta}_1-\underline{\Phi}_1)\underline{\Sigma}(\underline{\Theta}_1-\underline{\Phi}_1)^{'}-\underline{\Theta}_1\,\underline{\Sigma}\,{\underline{\Theta}_1}^{'}$$
$$\underline{\Gamma_{(n)}}(1)=\underline{\Gamma_{(n)}}(0){\underline{\Phi}_1}^{'}-\underline{\Sigma}{\underline{\Theta}_1}^{'}$$
$$\underline{\Gamma_{(n)}}(l)=\underline{\Gamma_{(n)}}(0)[{\underline{\Phi}_1}^{(l-1)}]^{'}\ (l>1)$$

ou encore

(2.23)
$$\underline{\Gamma_{(n)}}(0)=\underline{\Sigma}+\underline{\Phi}_1\,\underline{\Gamma}(0)\,{\underline{\Phi}_1}^{'}+(\underline{\Theta}_1-\underline{\Phi}_1)\underline{\Sigma}(\underline{\Theta}_1-\underline{\Phi}_1)^{'}-\underline{\Theta}_1\,\underline{\Sigma}\,{\underline{\Theta}_1}^{'}$$
$$\underline{\Gamma_{(n)}}(l)=\underline{\Gamma_{(n)}}(0)[{\underline{\Phi}_1}^{l}]^{'}\ -\underline{\Sigma}{\underline{\Theta}_1}^{'}[{\underline{\Phi}_1}^{l-1}]^{'}\ (l\geq 1)$$

Il vient la propriété suivante.

Propriété 2. 11 : cas général

Pour un ARMA(1,1) vectoriel n-dimensionnel, centré et stationnaire, les éléments propres de la matrice $\widehat{\underline{\Gamma}_{(pn)}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres de la matrice $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent des valeurs propres de :

$$\underline{\Gamma}_{(n)}(0) + \sum_{j=1}^{p-1} \underline{\Gamma}_{(n)}(0)[\underline{\Phi}_1{}^j]' \, \mathrm{e}^{-i(2\pi jk/p)} - \sum_{j=1}^{p-1} \underline{\Sigma\Theta}_1{}'[\underline{\Phi}_1{}^{j-1}]' \, \mathrm{e}^{-i(2\pi jk/p)} \quad (k = 0,\ 1, \cdots p-1) \tag{2.24}$$

avec comme point de départ :

$$\underline{\Gamma}_{(n)}(0) = \underline{\Sigma} + \underline{\Phi}_1 \, \underline{\Gamma}_{(n)}(0) \, \underline{\Phi}_1{}' + (\underline{\Theta}_1 - \underline{\Phi}_1)\underline{\Sigma}(\underline{\Theta}_1 - \underline{\Phi}_1)' - \underline{\Theta}_1 \, \underline{\Sigma} \, \underline{\Theta}_1{}'$$

(b) et les vecteurs propres peuvent être approchés par :

$$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)} \, \underline{u}_{lk}; j = 0, \cdots, p-1\right] \ (k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1)$$

où $\underline{u}_{lk}$ sont les vecteurs propres de (2.24).

Démonstration :

La démonstration est immédiate. Pour cela, il suffit de combiner les résultats en (1.21) et (2.23) pour les valeurs propres et de s'en référer à (1.22) pour les vecteurs propres.

Fin démonstration.

2.3.3.2 Le cas particulier de n_1 AR(1) et n_2 MA(1) processus « indépendants »

Le modèle vectoriel est maintenant constitué de n_1 processus AR(1) et n_2 processus MA(1) non-corrélés, centrés et stationnaires où $n_1 + n_2 = n$.

Le modèle étudié peut donc être considéré comme un modèle ARMA(1,1) vectoriel avec pour matrices $\underline{\Phi}_1$, $\underline{\Theta}_1$ et $\underline{\Sigma}$, les matrices diagonales suivantes :

$$\underline{\Phi}_1 = \begin{bmatrix} \underline{\Phi_{(n_1)}} & \underline{0_{(n_1)}} \\ \underline{0_{(n_2)}} & \underline{0_{(n_2)}} \end{bmatrix}, \ \underline{\Theta}_1 = \begin{bmatrix} \underline{0_{(n_1)}} & \underline{0_{(n_1)}} \\ \underline{0_{(n_2)}} & \underline{\Theta_{(n_2)}} \end{bmatrix} \text{ et } \underline{\Sigma} = \begin{bmatrix} \underline{\Sigma_{(n_1)}} & \underline{0_{(n_1)}} \\ \underline{0_{(n_2)}} & \underline{\Sigma_{(n_2)}} \end{bmatrix}$$

où

$$\underline{\Phi_{(n_1)}} = \begin{pmatrix} \phi_{11} & & 0 \\ & \ddots & \\ 0 & & \phi_{n_1 n_1)} \end{pmatrix}, \ \underline{\Theta_{(n_2)}} = \begin{pmatrix} \theta_{(n_1+1)(n_1+1)} & & 0 \\ & \ddots & \\ 0 & & \theta_{nn} \end{pmatrix}$$

et

$$\underline{\Sigma_{(n_1)}} = \begin{pmatrix} \sigma^2_{a_1} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_{n_1}} \end{pmatrix} \text{ et } \underline{\Sigma_{(n_2)}} = \begin{pmatrix} \sigma^2_{a_{n_1+1}} & & 0 \\ & \ddots & \\ 0 & & \sigma^2_{a_n} \end{pmatrix}$$

La série i n'étant pas influencée par la série j pour $i \neq j$, la matrice des covariances décalées $\underline{\Gamma_{(n)}}(l)$ s'écrit aussi plus simplement :

$$\underline{\Gamma_{(n)}}(l) = \begin{bmatrix} \underline{\Gamma_{(n_1)}}(l) & \underline{0_{(n_1)}} \\ \underline{0_{(n_2)}} & \underline{\Gamma_{(n_2)}}(l) \end{bmatrix} \ (l = 0,1,2,\cdots)$$

avec $\underline{\Gamma_{(n_1)}}(l)$ la matrice de covariance vérifiant (2.7) et $\underline{\Gamma_{(n_2)}}(l)$ la matrice de covariance vérifiant (2.8).

$\underline{\Gamma_{(n)}}(l)$ est une matrice bloc-diagonale, il vient d'après les propriétés 1.11 et 1.12 établies dans le chapitre 1, la propriété suivante.

Propriété 2. 12 : n_1 AR(1) et n_2 MA(1) « indépendants »

Pour un modèle mixte AR(1) n_1-dimensionnel et MA(1) n_2-dimensionnel, centré et stationnaire, où les processus sont non-corrélés et $n_1 + n_2 = n$, les éléments propres de la matrice $\widehat{\underline{\Gamma}_{(pn)}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres réelles de $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent des valeurs propres :

$$(2.25) \qquad \frac{\sigma^2_{a_m}}{1-\phi_{mm}{}^2} \sum_{j=0}^{p-1} \mathrm{e}^{-i(2\pi jk/p)}\, \phi_{mm}{}^j$$

$(k = 0,\ 1, \cdots p-1)$ et $(m = 1, \cdots, n_1)$

et

$$(2.26) \quad \sigma^2_{a_m}(1 + \theta_{mm}{}^2 - \theta_{mm}\, \mathrm{e}^{-i(2\pi k/p)})$$

$(m = n_1 + 1, \cdots, n)$

(b) et les vecteurs propres de $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent de

$$(2.27) \qquad \overrightarrow{V_{lk}} \sim p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\, \underline{e}_l ; j = 0, \cdots, (p-1)\right]$$

$(k = 0,\ 1, \cdots p-1,\ l = 0,\ 1, \cdots n-1)$

avec $\overrightarrow{V_{lk}} = \left[(a_{ji}^{(lk)}); 1 \le j \le p,\ 1 \le i \le n)\right]'$

où $a_{ji}^{(lk)} \sim 0$ $(i \neq l)$ et $a_{jl}^{(lk)} \sim p^{-1/2}\, \mathrm{e}^{-i(2\pi(j-1)k/p)}$ $(1 \le j \le p)$

Démonstration :

La démonstration est immédiate. Pour les valeurs propres, il suffit de reprendre la propriété 1.11 et les résultats en (2.15) et (2.20).

Pour les vecteurs propres, nous rappelons que dans le cas où les processus sont non-corrélés, les vecteurs propres sont indépendants des paramètres du modèle. Le résultat en (2.27) est donc le même qu'en (2.16), (2.21) et (1.31).

Fin démonstration.

2.3.4 Modèles ARMA(p,q) vectoriels : cas général

C'est à partir du système d'équations matricielles de récurrence établi en (2.11) et des résultats en (1.21) et (1.22) que nous pouvons donner une expression approchée des éléments propres de la matrice de covariance $\widehat{\underline{\Gamma}_{(pn)}}$ d'un processus ARMA(p,q) . Nous obtenons après substitution la propriété suivante.

Propriété 2. 13

Pour un ARMA(p,q) vectoriel n-dimensionnel, centré et stationnaire, les éléments propres de la matrice $\widehat{\underline{\Gamma}_{(pn)}}$ de covariance autour des p vecteurs aléatoires décalés $\underline{Z}_t, \underline{Z}_{(t+1)}, \underline{Z}_{(t+2)}, \cdots, \underline{Z}_{t+(p-1)}$, ont les propriétés asymptotiques suivantes :

(a) les valeurs propres de la matrice $\widehat{\underline{\Gamma}_{(pn)}}$ se rapprochent des valeurs propres de :

$$\sum_{j=0}^{q}\left(\sum_{i=1}^{p}\underline{\Gamma}_{(n)}(j-i)\,\underline{\Phi}_i^{'}-\sum_{i=j}^{q}\underline{\Psi}_{i-j}\underline{\Sigma}\,\underline{\Theta}_i^{'}\right)\mathrm{e}^{-i(2\pi jk/p)} - \sum_{j=q+1}^{p-1}\left(\sum_{i=1}^{p}\underline{\Gamma}_{(n)}(j-i)\,\underline{\Phi}_i^{'}\right)\mathrm{e}^{-i(2\pi jk/p)} \qquad (2.28)$$

$$(k=0,\,1,\cdots p-1)$$

b) et les vecteurs propres peuvent être approchés par :

$$p^{-1/2}\left[\mathrm{e}^{-i(2\pi jk/p)}\,\underline{u}_{lk};j=0,\cdots,p-1\right]\ (k=0,\,1,\cdots p-1,\,l=0,\,1,\cdots n-1)$$

où $\underline{u}_{lk}$ sont les vecteurs propres de (2.28).

L'expression des valeurs propres reste cependant très complexe à analyser dans le cas général. C'est pour cela que nous nous ramenons au cas de modèles ARMA(p, q) vectoriels constitués de processus non-corrélés, centrés et stationnaires.

2.4 Identification des composantes principales de processus ARMA « indépendants »

Du chapitre 1, nous savons que d'une part les composantes principales d'un modèle vectoriel constitué de processus stationnaires et « indépendants » coïncident avec les composantes principales de chacun des processus (propriété 1.13), et d'autre part que chacune des composantes principales est une moyenne mobile dont les poids sont les coefficients des vecteurs propres associés (section 1.4, cas univarié). Nous proposons alors d'identifier ces moyennes mobiles pour respectivement un processus MA(1), un processus AR(1) et un processus ARMA(p, q).

2.4.1 Cas d'un MA(1) et d'un AR(1)

En référence à (1.28) et (1.29), une réalisation à l'instant t de $F^{(k)}$, la $k^{ième}$ composante principale d'un processus (Z_t) centré et stationnaire, pour les variables $Z^{(j)} = Z_{t+(j-1)}$ $(1 \le j \le p)$ s'écrit :

$$f_k(t) = \left[a_1^{(k)} \cdots a_p^{(k)}\right] \begin{bmatrix} z_1(t) \\ \vdots \\ z_p(t) \end{bmatrix}$$

où $z_1(t) = z(t), \cdots, z_p(t) = z(t+(p-1))$ sont les réalisations à l'instant t des variables $Z^{(1)}$, $Z^{(2)}, \cdots, Z^{(p)}$,

et $\left[a_1^{(k)} \cdots a_p^{(k)}\right] \sim p^{-1/2}\left[\mathrm{e}^{-i(2\pi kj/p)}; j = 0, \cdots, (p-1)\right]$ $(k = 0,\ 1, \cdots p-1)$.

Pour (Z_t) **un processus MA(1)** centré et stationnaire tel que $Z_t = a_t - \theta_1 a_{t-1}$, nous obtenons :

(2.29) $$f_k(t) = a_1^{(k)}[a(t) - \theta_1 a(t-1)] + a_2^{(k)}[a(t+1) - \theta_1 a(t)] + \cdots$$

$$\cdots + a_{p-1}^{(k)}[a(t+(p-2)) - \theta_1 a(t+(p-3))] + a_p^{(k)}[a(t+(p-1)) - \theta_1 a(t+(p-2))]$$

Posons $T = t + (p-1)$, *c'est à dire que nous travaillons désormais avec les variables retardées* $Z^{(1)-}$, $Z^{(2)-}$,$\cdots$,$Z^{(p)-}$ tel que $Z^{(j)} = Z_{T-(j-1)}$ $(1 \le j \le p)$,

(2.29) s'écrit :

$$f_k(T) = a_p^{(k)}[a(T) - \theta_1 a(T-1)] + a_{p-1}^{(k)}[a(T-1) - \theta_1 a(T-2)] + \cdots$$

$$\cdots + a_2^{(k)}[a(T-(p-2)) - \theta_1 a(T-(p-1))] + a_1^{(k)}[a(T-(p-1)) - \theta_1 a(T-p)]$$

Il vient :

$$\left[a_p^{(k)}\right]^{-1} f_k(T) = a(T) + \sum_{J=1}^{p} \beta_J^{(k)} a(T-J)$$

avec

$$\beta_J^{(k)} = (-a_{p-J+1}^{(k)}\theta_1 + a_{p-J}^{(k)}) / a_p^{(k)} \ \ (J = 1,\cdots,p) \text{ et } (a_0^{(k)} = 0).$$

Nous en déduisons la propriété suivante.

Propriété 2. 14

Soit (Z_t) un processus MA(1) centré et stationnaire tel que $Z_t = a_t + \theta_1 a_{t-1}$. Les composantes principales $\left[a_p^{(k)}\right]^{-1} F^{(k)}$ $(k = 0,\ 1,\cdots p-1)$ de (Z_t) pour les variables $Z^{(j)-} = Z_{t+(j-1)}$ $(1 \le j \le p)$ suivent un **modèle MA(** p **)**

avec comme coefficients :

(2.30) $\quad \beta_J^{(k)} = (-a_{p-J+1}^{(k)}\theta_1 + a_{p-J}^{(k)}) / a_p^{(k)} \ \ (J = 1,\cdots,p)$ et $(a_0^{(k)} = 0)$

Pour (Z_t) **un processus AR(1)** centré et stationnaire tel que $Z_t = \phi_1 Z_{t-1} + a_t$:

(2.31) $\quad f_k(t) = a_1^{(k)}[\phi_1 z(t-1) + a(t)] + a_2^{(k)}[\phi_1 z(t) + a(t+1)] + \cdots$

$$\cdots + a_{p-1}^{(k)}[\phi_1 z(t+(p-3)) + a(t+(p-2))] + a_p^{(k)}[\phi_1 z(t+(p-2)) + a(t+(p-1))]$$

De même, posons $T = t + (p-1)$,

(2.31) s'écrit :

$$f_k(T) = a_p^{(k)}[\phi_1 z(T-1) + a(T)] + a_{p-1}^{(k)}[\phi_1 z(T-2) + a(T-1)] + \cdots$$

$$\cdots + a_2^{(k)}[\phi_1 z(T-(p-1)) + a(T-(p-2))] + a_1^{(k)}[\phi_1 z(T-p) + a(T-(p-1))]$$

Il vient :

$$\left[a_p^{(k)}\right]^{-1} f_k(T) = \sum_{I=1}^{p} \alpha_I z(T-I) + a(T) + \sum_{J=1}^{p-1} \beta_J a(T-J)$$

où

$$\alpha_I = a_{p-I+1}^{(k)} \phi_1 \,/\, a_p^{(k)} \text{ et } \beta_J = a_{p-J}^{(k)} \,/\, a_p^{(k)} \ (I = 1,\cdots,p)\ (J = 1,\cdots,p-1).$$

Nous en déduisons la propriété suivante.

Propriété 2. 15

Soit (Z_t) un processus AR(1) centré et stationnaire tel que $Z_t = \phi_1 Z_{t-1} + a_t$. Les composantes principales $\left[a_p^{(k)}\right]^{-1} F^{(k)}$ $(k = 0,\ 1,\cdots p-1)$ de (Z_t) pour les variables $Z^{(j)-} = Z_{t+(j-1)}$ $(1 \le j \le p)$ sont des **moyennes mobiles du type** :

$$\text{(2.32)} \qquad \left[a_p^{(k)}\right]^{-1} f_k(t) = \sum_{I=1}^{p} \alpha_I^{(k)} z(t-I) + a(t) + \sum_{J=1}^{p-1} \beta_J^{(k)} a(t-J)$$

avec

$$\text{(2.33)} \qquad \alpha_I^{(k)} = a_{p-I+1}^{(k)} \phi_1 \,/\, a_p^{(k)}$$

$$\text{(2.34)} \quad \beta_J^{(k)} = a_{p-J}^{(k)} \,/\, a_p^{(k)} \ (I = 1,\cdots,p)\ (J = 1,\cdots,p-1)$$

2.4.2 Cas d'un processus ARMA

Les résultats précédents peuvent s'étendre au cas d'un processus ARMA($\tilde{p},\tilde{q}$) centré et stationnaire, et donnent lieu à la propriété suivante.

Propriété 2. 16

Soit (Z_t) un processus ARMA($\tilde{p},\tilde{q}$) centré et stationnaire tel que $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_{\tilde{p}} Z_{t-\tilde{p}} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_{\tilde{q}} a_{t-\tilde{q}}$. Les composantes principales $\left[a_p^{(k)}\right]^{-1} F^{(k)}$ $(k = 0,\ 1,\cdots p-1)$ de (Z_t) pour les variables $Z^{(j)-} = Z_{t-(j-1)}$ $(1 \le j \le p)$ sont :

a) si $\tilde{p} \neq 0$, des **moyennes mobiles du type** :

$$\left[a_p^{(k)}\right]^{-1} f_k(t) = \sum_{I=1}^{p+\tilde{p}-1} \alpha_I^{(k)} z(t-I) + a(t) + \sum_{J=1}^{p+\tilde{q}-1} \beta_J^{(k)} a(t-J) \tag{2.35}$$

avec

$$\alpha_I^{(k)} = \left[a_p^{(k)}\right]^{-1} \sum_{u=\max(1,I-p+1)}^{\min(\tilde{p},I)} a_{p+u-I}^{(k)} \phi_u \quad (1 \leq I \leq p+\tilde{p}-1) \text{ et } (\tilde{p} \leq p) \tag{2.36}$$

$$\beta_J^{(k)} = \left[a_p^{(k)}\right]^{-1} \left(- \sum_{u=\max(1,J-p+1)}^{\min(\tilde{q},J)} a_{p+u-J}^{(k)} \Theta_u + a_{p+\max(1,J-p+1)-(J+1)}^{(k)} \right) \tag{2.37}$$

$(1 \leq J \leq p+\tilde{q}-1)$, $(\tilde{q} \leq p)$ et $a_0^{(k)} = 0$

Nous retrouvons pour $\tilde{p} = 1$ et $\tilde{q} = 0$ la propriété 2.15.

b) si $(\tilde{p} = 0)$, des **MA(** $p + \tilde{q} - 1$ **)** avec comme coefficients, les coefficients β_J définis en (2.37). Nous retrouvons pour $\tilde{q} = 1$ la propriété 2.14.

Démonstration :

De même que dans les cas des processus AR(1) et MA(1) précédents, il suffit d'écrire :

$$\begin{aligned}
f_k(t) = {} & a_1^{(k)}[\phi_1 z(t-1) + \cdots + \phi_{\tilde{p}} z(t-\tilde{p}) + a(t) - \theta_1 a(t-1) - \cdots - \theta_{\tilde{q}} a(t-\tilde{q})] + \\
& a_2^{(k)}[\phi_1 z(t) + \cdots + \phi_{\tilde{p}} z((t+1)-\tilde{p}) + a(t+1) - \theta_1 a(t) - \cdots - \theta_{\tilde{q}} a((t+1)-\tilde{q})] \\
& + \cdots + \\
& a_{p-1}^{(k)}[\phi_1 z(t+(p-3)) + \cdots + \phi_{\tilde{p}} z(t+(p-2)-\tilde{p}) + \\
& a(t+(p-2)) - \cdots - \theta_{\tilde{q}} a(t+(p-2)-\tilde{q})] + \\
& a_p^{(k)}[\phi_1 z(t+(p-2)) + \cdots + \phi_{\tilde{p}} z(t+(p-1)-\tilde{p}) \\
& + a(t+(p-1)) - \cdots - \theta_{\tilde{q}} a(t+(p-1)-\tilde{q})]
\end{aligned} \tag{2.38}$$

Posons $T = t + (p-1)$,

(2.38) s'écrit :

$$f_k(T) = a_p^{(k)}[\phi_1 z(T-1) + \cdots + \phi_{\tilde{p}} z(T-\tilde{p}) + a(T) - \cdots - \theta_{\tilde{q}} a(T-\tilde{q})] +$$
$$a_{p-1}^{(k)}[\phi_1 z(T-2) + \cdots + \phi_{\tilde{p}} z(T-1-\tilde{p}) + a(T-1) - \cdots - \theta_{\tilde{q}} a(T-1-\tilde{q})]$$
$$+\cdots+$$
$$a_2^{(k)}[\phi_1 z(T-(p-1)) + \cdots + \phi_{\tilde{p}} z(T-(p-2+\tilde{p}))$$
$$+a(T-(p-2)) - \cdots - \theta_{\tilde{q}} a((T-(p-2+\tilde{q}))]$$
$$+a_1^{(k)}[\phi_1 z(T-p) + \cdots + \phi_{\tilde{p}} z(T-(p-1+\tilde{p}))$$
$$+a(T-(p-1)) - \cdots - \theta_{\tilde{q}} a(T-(p-1+\tilde{q}))]$$

Il vient :

$$\left[a_p^{(k)}\right]^{-1} f_k(t) = \sum_{I=1}^{p+\tilde{p}-1} \alpha_I^{(k)} z(T-I) + a(T) + \sum_{J=1}^{p+\tilde{q}-1} \beta_J^{(k)} a(T-J)$$

avec

(a) pour la partie «autorégressive» d'ordre $p+\tilde{p}-1$, les coefficients suivants sachant que $\tilde{p} \leq p$ et $\tilde{p} \neq 0$:

$$\begin{aligned}
a_p^{(k)}\alpha_1^{(k)} &= a_p^{(k)}\phi_1 \\
a_p^{(k)}\alpha_2^{(k)} &= a_p^{(k)}\phi_2 + a_{p-1}^{(k)}\phi_1 \\
&\vdots \\
a_p^{(k)}\alpha_{\tilde{p}}^{(k)} &= a_p^{(k)}\phi_{\tilde{p}} + a_{p-1}^{(k)}\phi_{\tilde{p}-1} + a_{p-2}^{(k)}\phi_{\tilde{p}-2} + \cdots + a_{p-\tilde{p}+1}^{(k)}\phi_1 \\
&\vdots \\
a_p^{(k)}\alpha_p^{(k)} &= a_{\tilde{p}}^{(k)}\phi_{\tilde{p}} + a_{\tilde{p}-1}^{(k)}\phi_{\tilde{p}-1} + a_{\tilde{p}-2}^{(k)}\phi_{\tilde{p}-2} + \cdots + a_1^{(k)}\phi_1 \\
a_p^{(k)}\alpha_{p+1}^{(k)} &= a_{\tilde{p}-1}^{(k)}\phi_{\tilde{p}} + a_{\tilde{p}-2}^{(k)}\phi_{\tilde{p}-1} + a_{\tilde{p}-3}^{(k)}\phi_{\tilde{p}-2} + \cdots + a_1^{(k)}\phi_2 \\
&\vdots \\
a_p^{(k)}\alpha_{p+(\tilde{p}-2)}^{(k)} &= a_2^{(k)}\phi_{\tilde{p}} + a_1^{(k)}\phi_{\tilde{p}-1} \\
a_p^{(k)}\alpha_{p+(\tilde{p}-1)}^{(k)} &= a_1^{(k)}\phi_{\tilde{p}}
\end{aligned}$$

c'est à dire

$$a_p^{(k)}\alpha_I^{(k)} = \sum_{u=\max(1,I-p+1)}^{\min(\tilde{p},I)} a_{p+u-I}^{(k)}\phi_u$$

$(1 \le I \le p+\tilde{p}-1)$ et $(0 < \tilde{p} \le p)$

(b) et pour la partie «moyenne mobile» d'ordre $p+\tilde{q}-1$, les coefficients suivants :

$$\begin{aligned}
&a_p^{(k)}\beta_1^{(k)} = -a_p^{(k)}\theta_1 + a_{p-1}^{(k)}\\
&a_p^{(k)}\beta_2^{(k)} = -(a_p^{(k)}\theta_2 + a_{p-1}^{(k)}\theta_1) + a_{p-2}^{(k)}\\
&\vdots\\
&a_p^{(k)}\beta_{\tilde{q}}^{(k)} = -(a_p^{(k)}\theta_{\tilde{q}} + a_{p-1}^{(k)}\theta_{\tilde{q}-1} + a_{p-2}^{(k)}\theta_{\tilde{q}-2} + \cdots + a_{p-\tilde{q}+1}^{(k)}\theta_1) + a_{p-\tilde{q}}^{(k)}\\
&a_p^{(k)}\beta_{\tilde{q}+1}^{(k)} = -(a_{p-1}^{(k)}\theta_{\tilde{q}} + a_{p-2}^{(k)}\theta_{\tilde{q}-1} + a_{p-3}^{(k)}\theta_{\tilde{q}-2} + \cdots + a_{p-\tilde{q}}^{(k)}\theta_1) + a_{p-\tilde{q}-1}^{(k)}\\
&\vdots\\
&a_p^{(k)}\beta_{p-1}^{(k)} = -(a_{\tilde{q}+1}^{(k)}\theta_{\tilde{q}} + a_{\tilde{q}}^{(k)}\theta_{\tilde{q}-1} + a_{\tilde{q}-1}^{(k)}\theta_{\tilde{q}-2}) + \cdots + a_1^{(k)}\\
&a_p^{(k)}\beta_p^{(k)} = -(a_{\tilde{q}}^{(k)}\theta_{\tilde{q}} + a_{\tilde{q}-1}^{(k)}\theta_{\tilde{q}-1} + a_{\tilde{q}-2}^{(k)}\theta_{\tilde{q}-2}) + \cdots + a_2^{(k)}\\
&\vdots\\
&a_p^{(k)}\beta_{p+(\tilde{q}-2)}^{(k)} = -a_2^{(k)}\theta_{\tilde{q}} + a_1^{(k)}\\
&a_p^{(k)}\beta_{p+(\tilde{q}-1)}^{(k)} = -a_1^{(k)}\theta_{\tilde{q}}\ (+a_0^{(k)})
\end{aligned}$$

c'est à dire

$$a_p^{(k)}\beta_J^{(k)} = \left(-\sum_{u=\max(1,J-p+1)}^{\min(\tilde{q},J)} a_{p+u-J}^{(k)}\Theta_u + a_{p+\max(1,J-p+1)-(J+1)}^{(k)}\right)$$

$(1 \le J \le p+\tilde{q}-1)$, $(\tilde{q} \le p)$ et $a_0^{(k)} = 0$

Fin démonstration.

➢ ***Pour les processus MA($\tilde{q}$), les composantes principales sont clairement des modèles MA($p+\tilde{q}-1$) si p est le nombre de variables étudiées. Dans le cas des AR($\tilde{p}$) et plus généralement des ARMA stationnaires, leur identification est plus délicate. Nous renvoyons au chapitre 3 pour leur spécification dans le contexte d'une population.***

Chapitre 3

3 Des modèles factoriels de processus ARMA – Une méthode d'analyse

Introduction

L'objectif de ce chapitre est la construction de premiers modèles factoriels de processus ARMA, pour l'*identification* et « l'*estimation* » d'une série temporelle. Ces deux étapes sont bien connues dans la méthodologie de Box et Jenkins ([Box et al.70]), mais notre approche est différente :

- elle utilise les résultats théoriques des chapitres 1 et 2 sur les composantes principales de modèles vectoriels constitués de processus « indépendants » et stationnaires ;
- elle est de nature *graphique* car il s'agit de *projeter* une série dans un modèle factoriel de référence ;
- elle a l'originalité de fournir *simultanément* l'identification et une première estimation des paramètres d'un modèle AR(1) ou MA(1) qui engendrerait la série.

Pour atteindre notre but, nous nous plaçons dans le *contexte d'une population* qui est *la réalisation* soit d'un modèle unidimensionnel, c'est à dire d'un seul processus stationnaire, soit d'un modèle multidimensionnel constitué de plusieurs *processus « indépendants » et stationnaires*.

Nous procédons alors par une longue phase d'apprentissage qui comporte essentiellement cinq étapes :

- la *simulation* d'échantillons de trajectoires AR(1) et MA(1*), « indépendantes » et stationnaires* ;
- la *spécification graphique* par les composantes principales de chacune des trajectoires simulées, avec tout d'abord la technique SSA et la *visualisation graphique* de deux types de composantes principales, puis l'*identification* des modèles dynamiques obtenus par la méthode de Box et Jenkins ;
- la *construction* de premiers modèles factoriels de séries basés sur *des ACP « temporelles »* d'un ou plusieurs échantillons ;
- le recours à l'analyse des corrélations et
- le recours à l'analyse structurelle

pour améliorer et rendre possible l'identification puis l'estimation des séries en fonction de leurs coefficients, quels que soient les échantillons.

L'analyse des corrélations utilise des éléments de la fonction d'autocorrélation, notée FAC, et des éléments de la fonction d'autocorrélation partielle, notée FAP. Plusieurs ACP sont construites sur ces critères et nous en déduisons un premier modèle factoriel de référence. L'identification et l'estimation des processus à « forts » coefficients est alors possible à l'exception des processus dits « faibles ».

L'analyse structurelle a pour but de décrire et de mesurer d'éventuels changements structurels. Pour cela, elle calcule des entropies sur des séries d'états ou des séries brutes. Les critères sont alors nombreux et pour faciliter l'interprétation des regroupements et oppositions de séries, nous construisons une analyse des correspondances multiples suivie d'une classification. Un second modèle factoriel de référence est déduit et il permet l'identification et l'estimation de trois classes de processus dont celle qui comporte les processus dits « faibles ».

Chacune des cinq étapes fait l'objet d'une partie distincte de ce chapitre auxquels s'ajoutent, en première section, des rappels sur les principes généraux de la méthodologie de Box et Jenkins.

Enfin, que ce soit pour la simulation, la spécification par la technique SSA, la méthodologie de Box et Jenkins, les analyses des corrélations et structurelle, ou les représentations graphiques, nous développons à chaque fois des procédures en S en nous aidant des ouvrages principaux [Becker et al.88], [Spector94], [Venables et al.2000], [Ferrara et al.2002], ainsi que des manuels d'utilisation du logiciel S-Plus [Mathsoft99]. Un projet est aussi développé en Fortran 90 et intégré à S-Plus pour le calcul des différentes entropies.

3.1 La méthode de Box et Jenkins : quelques rappels

La méthodologie de Box et Jenkins repose avant tout sur une modélisation de la série étudiée, stationnaire ou rendue stationnaire, par un processus de type ARMA(p,q). Cette technique comporte principalement quatre étapes : l'identification, l'estimation, la vérification, et la prévision du processus. Dans ce qui suit, nous proposons quelques rappels sur les trois premières étapes que nous utiliserons pour identifier les composantes principales des trajectoires simulées, dans un cadre univarié ou multivarié.

3.1.1 FAC et FAP de processus ARMA

Tout d'abord, l'étape d'identification d'un processus ARMA(p,q) consiste à choisir les entiers p pour la partie AR et q pour la partie MA en s'aidant d'une part de la fonction d'autocorrélation FAC, et d'autre part de la fonction d'autocorrélation partielle FAP. Nous connaissons la fonction d'autocovariance $\gamma(.)$ (chapitre 1,

section 1.2.3) ; la FAC se déduit de cette fonction en divisant par la variance du processus. Les éléments obtenus sont des corrélations. La FAP s'analyse de façon analogue en mesurant les corrélations entre Z_t et Z_{t-k} une fois retirée l'influence des variables Z_{t-k-i} pour ($i < k$). Nous rappelons leurs principales caractéristiques pour d'une part les processus AR(p) et MA(q) avec la propriété 3.1, et d'autre part les processus ARMA(p,q) avec la propriété 3.2. Pour les démonstrations, nous proposons les ouvrages [Anderson76], puis [Hamilton94].

Propriété 3. 1
ou le cas des processus MA(q) et AR(p)

Soient (Z_t) un processus stationnaire, $\rho_{(.)}$ la FAP et $\rho(.)$ la FAC.

- Si $(Z_t) \sim AR(p)$, alors $\rho_{(kk)} = 0$, si $k > p$ et la FAC se présente selon une exponentielle amortie et/ou une sinusoïde amortie.
- Si $(Z_t) \sim MA(q)$, alors $\rho(k) = 0$, si $k > p$ et la FAP se présente selon une exponentielle amortie et/ou une sinusoïde amortie.

Nous pouvons alors constater la symétrie des comportements de la FAC d'un AR(p) et de la FAP d'un MA(q) d'une part, et de la FAP d'un AR(p) et de la FAC d'un MA(q) d'autre part. Il suffit d'inverser la lecture des corrélogrammes comme les tableaux TABLEAU 3.1 et 3.2 le montrent pour les AR(1) et MA(1), puis les AR(2) et MA(2). Notons que les graphiques de la FAC nous donnent $\rho(k)$ pour $k = 0,1,2,\cdots$ alors que les graphiques de la FAP nous donnent $\rho_{(kk)}$ pour $k = 1,2,\cdots$. Il en va de même pour tous les autres corrélogrammes.

➢ ***En pratique, nous cherchons le retard k à partir duquel la FAC ou la FAP de la chronique s'annule.***

Propriété 3. 2
ou le cas des processus ARMA(p,q) d'ordres élevés

Pour un processus ARMA(p,q) stationnaire d'ordres élevés, la FAC se comporte comme celle d'un processus AR(p) après $q-p$ retards, et la FAP se comporte comme celle d'un processus MA(q) après $p-q$ retards.

Souvent, la FAC et la FAP sont complétées par des tests statistiques pour la sélection des ordres maximaux des parties AR et MA. Ces tests utilisent les théorèmes de Bartlett pour les MA(q) et de Quenouille pour les processus AR(p). Nous renvoyons aux ouvrages de [Bartlett46] et [Quenouille49] pour leur définition.

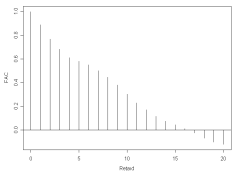
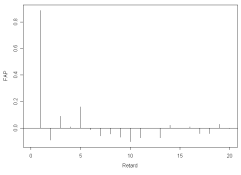
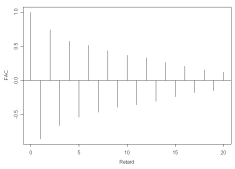
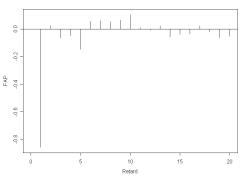
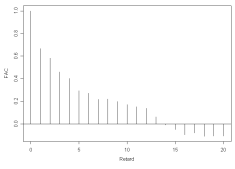
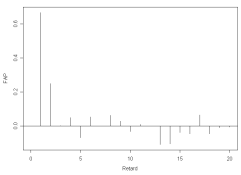
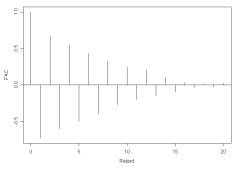
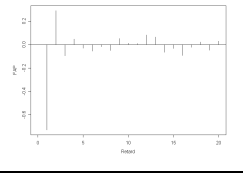
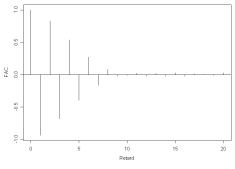
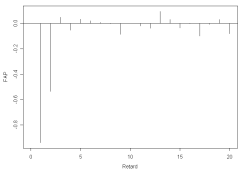
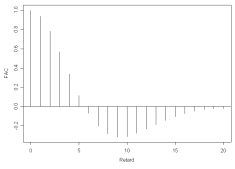
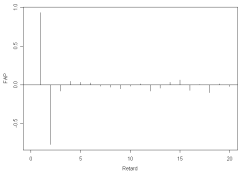
MODELES AR(1) et AR(2)	**FAC corrélogrammes**	**FAP corrélogrammes**	**FAC et FAP théoriques**
$Z_t = \phi_1 Z_{t-1} + a_t$ $0 < \phi_1 < 1$			$\rho(k) = \phi_1^{\,k} \;\; k \geq 0$ et $\rho_{(11)} = \phi_1$ $\rho_{(kk)} = 0 \;\; k > 1$
idem $-1 < \phi_1 < 0$			
$Z_t = \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + a_t$ $\phi_1 > 0$ $\phi_1^{\,2} + 4\phi_2 > 0$			$\rho(1) = \dfrac{\phi_1}{1-\phi_2}$ $\rho(2) = \phi_2 + \dfrac{\phi_1^{\,2}}{1-\phi_2}$ etc. et $\rho_{(11)} = \dfrac{\phi_1}{1-\phi_2}$ $\rho_{(22)} = \phi_2$ $\rho_{(kk)} = 0 \;\; k > 2$
idem $\phi_1 < 0$ $\phi_1^{\,2} + 4\phi_2 > 0$			
idem $\phi_1 < 0$ $\phi_1^{\,2} + 4\phi_2 < 0$			
idem $\phi_1 > 0$ $\phi_1^{\,2} + 4\phi_2 < 0$			

TABLEAU 3.1 – FAC et FAP de processus AR(1) et AR(2)

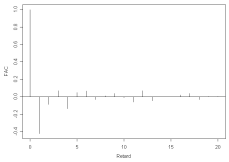
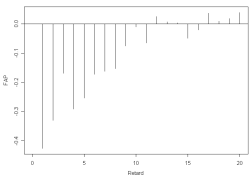
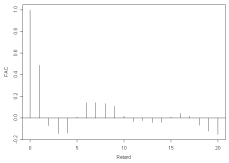
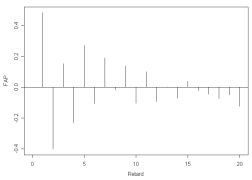
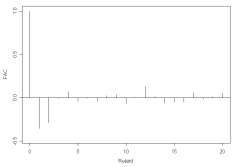
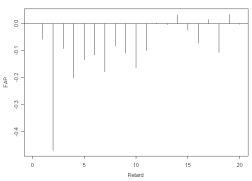
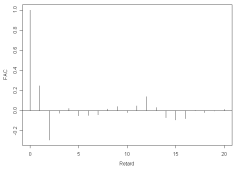
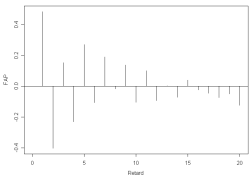
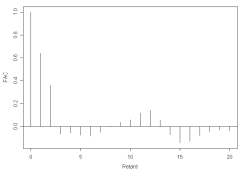
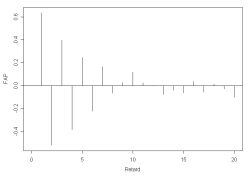
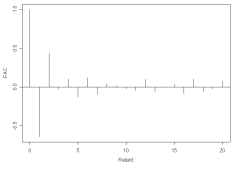
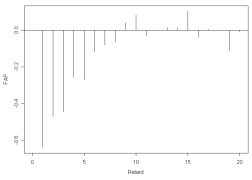
MODELES MA(1) et MA(2)	FAC corrélogrammes	FAP corrélogrammes	FAC et FAP théoriques
$Z_t = a_t - \theta_1 a_{t-1}$ $0 < \theta_1 < 1$			$\rho(1) = \frac{-\theta_1}{1+\theta_1^2}$ $\rho(k) = 0 \; k > 2$
idem $-1 < \theta_1 < 0$			$\rho_{(kk)} = \frac{-\theta_1^k(1-\theta_1^2)}{1-\theta_1^{2(k+1)}}$ $k \geq 0$
$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$ $\theta_1 > 0$ $\theta_1^2 + 4\theta_2 > 0$			$\rho(1) = \frac{-\theta_1 + \theta_1\theta_2}{1+\theta_1^2+\theta_2^2}$ $\rho(2) = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2}$ $\rho(k) = 0 \; k > 2$
idem $\theta_1 < 0$ $\theta_1^2 + 4\theta_2 > 0$			
idem $\theta_1 < 0$ $\theta_1^2 + 4\theta_2 < 0$			Ecriture complexe de la FAP
idem $\theta_1 > 0$ $\phi_1^2 + 4\phi_2 < 0$			

TABLEAU 3.2 – FAC et FAP de processus MA(1) et MA(2)

3.1.2 Estimation des paramètres et vérification

Cependant, pour ne retenir que les ordres qui ajustent correctement la série, deux autres étapes sont essentielles dans la méthode de Box et Jenkins. C'est d'une part l'estimation des coefficients qui utilise la méthode du maximum de vraisemblance, et d'autre part la comparaison des modèles avec un critère d'information suivie d'une analyse des résidus.

La méthode du maximum de vraisemblance :

Par la méthode du maximum de vraisemblance proposée dans [Box et al.70], section 7, l'estimateur $\hat{\theta}$ pour $\theta = (\mu, \sigma_a^2, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q)$ s'écrit :

$$\hat{\theta} = Arg \max_{\theta} L(\theta)$$

et correspond au paramètre qui maximise la log-vraisemblance du processus donnée par l'équation suivante :

(3.1) $$L(\theta) = -\frac{T-p}{2}\log(2\pi\sigma_a^2) - \sum_{t=p+1}^{T} \frac{a_t^2}{2\sigma_a^2}\,.$$

La vraisemblance du processus est conditionnée sur les p premières valeurs observées de la série (Z_t) de longueur T, et sur les q valeurs du bruit supposé gaussien telles que :

$$a_p = a_{p-1} = \cdots = a_{p-q+1} = 0\,.$$

En pratique, la résolution du problème de maximisation de la log-vraisemblance se fait à l'aide d'un algorithme du gradient conjugué du type Newton-Raphson qui effectue une recherche du minimum global par descente vers ce minimum à partir d'une valeur initiale.

Un critère d'information :

Pour choisir le meilleur des modèles, nous pouvons utiliser le critère d'information de Akaike introduit dans [Akaike77] et tout particulièrement adapté aux processus ARMA. Il est noté AIC (Akaike Information Criterion) et est défini de la façon suivante :

$$AIC = T\log(\hat{\sigma}_a^2) + 2(p+q)$$

où $\hat{\sigma}_a^2$ est l'estimation de la variance résiduelle.

Akaike montre que le meilleur des modèles ARMA est celui qui minimise la statistique AIC avec une variance résiduelle faible.

D'autres méthodes d'estimation des paramètres d'un processus ARMA ainsi que divers critères d'information existent et sont définis par exemple dans l'ouvrage de [Hamilton94].

L'analyse des résidus :

Par ailleurs, si le modèle est correctement spécifié, les résidus estimés doivent former une trajectoire issue d'un bruit blanc. Pour cela, les FAC et FAP sont analysées avec l'aide du test Portmanteau que nous pouvons trouver dans [Box et al.70].

Quelques particularités du logiciel S-Plus :

Le logiciel S-Plus utilise :

- un autre critère AIC (AIC_{S+}) donné dans [Brockwell et al.93],
- et la méthode du maximum de vraisemblance connue sous le nom de décomposition de l'erreur de prévision définie dans [Harvey81].

Par exemple, la valeur du critère AIC défini précédemment est facilement calculée à partir de la relation : $AIC_{S+} = -2L(\hat{\theta}) + 2(p+q)$ où $L(\theta)$ est donnée par l'équation (3.1).

Le logiciel possède :

- la fonction *arima.mle()* pour l'estimation des paramètres du modèle ARMA,
- et la fonction *arima.diag()* pour l'analyse des résidus.

La première fonction prend comme arguments le nom de la chronique et les ordres choisis après l'analyse des corrélogrammes. Elle renvoie des estimations des valeurs des paramètres, de la variance résiduelle et la valeur du critère AIC après convergence ou non de l'algorithme.

La seconde fonction prend comme arguments la liste fournie par la première.

Les procédures développées en S tiennent compte de ces particularités.

3.2 La simulation de neuf échantillons de trajectoires AR(1) et MA(1)

Comme très souvent dans la phase d'apprentissage (ou expérimentale) d'une méthode d'analyse sur les séries temporelles, il s'agit tout d'abord de simuler des échantillons de trajectoires qui sont la réalisation d'un ou plusieurs processus ARMA. Il s'ensuit des contraintes d'ordre général liées à la nature même des processus auxquelles s'ajoutent des contraintes particulières liées aux résultats des chapitres 1 et 2. Nous présentons alors l'ensemble de ces conditions.

3.2.1 Conditions générales de simulation

La simulation d'une trajectoire ARMA(p,q) du type $\Phi(B)Z_t = \Theta(B)a_t$ où Φ et Θ sont des polynômes en B de degré respectif p et q tel que

$$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q ,$$

où $(\phi_1, \phi_2, \cdots, \phi_p)$ et $(\theta_1, \theta_2, \cdots, \theta_q)$ sont des réels,

et $(a_t)_{t\in\mathbb{Z}}$ est un processus orthogonal ou bruit blanc centré de variance finie σ_a^2,

nécessite les données suivantes :
- le(s) coefficient(s) du modèle ;
- la loi du bruit et sa variance ;
- le point de départ de la série si le modèle contient une partie autorégressive.

Les coefficients du modèle sont fixés par l'expérimentateur et dans le cas d'un processus AR(1) stationnaire, ils doivent vérifier $|\phi_1| < 1$.

La loi du bruit est ici posée gaussienne et est simulée par un générateur de nombres aléatoires qui nécessite une valeur initiale de calage aléatoire.

Pour itérer la trajectoire, le point de départ a été choisi arbitrairement égal à zéro mais pour enlever l'influence de cette valeur initiale, nous avons choisi de simuler à chaque fois mille valeurs de la série pour ne retenir que les cinq cent dernières.

3.2.2 Les neuf échantillons simulés et la matrice temporelle

Dans ces conditions et en référence aux résultats des chapitres 1 et 2, nous procédons à la simulation de neuf échantillons de processus AR(1) et MA(1) qui doivent être « indépendants » et stationnaires. En effet, rappelons qu'un modèle vectoriel constitué de processus non-corrélés et stationnaires, est *stationnaire* (section 1.2.2), et que les scores du modèle coïncident alors avec les scores de chacun des processus (propriété 1.13).

Les caractéristiques d'un échantillon :

Un échantillon contient 18 processus AR(1) centrés et stationnaires et 18 processus MA(1) centrés. Ces 36 processus sont issus d'un même bruit blanc (même loi, même valeur de calage aléatoire et même variance) et sont ordonnés de la façon suivante :
- les 9 premiers, notés AR(1)ne, sont des modèles AR(1) à coefficients négatifs, ϕ_1 compris entre –0.9 et –0.1 et avec un pas de 0.1 ;
- les 9 suivants, notés MA(1)ne, sont des modèles MA(1) à coefficients

négatifs, θ_1 compris entre –0.9 et –0.1 et avec un pas de 0.1 ;
- les 9 suivants, notés AR(1)po, sont des modèles AR(1) à coefficients positifs, ϕ_1 compris entre +0.1 et +0.9 et avec un pas de 0.1 ;
- les 9 derniers, notés MA(1)po, sont des modèles MA(1) à coefficients positifs, θ_1 compris entre +0.1 et +0.9 et avec un pas de 0.1.

Les neuf échantillons :

Puis, en faisant varier la valeur de calage aléatoire et/ou la variance du bruit, nous obtenons les 9 échantillons suivants, avec pour :
- les 3 premiers échantillons : les variances de bruits (10, 50, 90), la même valeur de calage aléatoire (1023), et des séries numérotées de 1 à 108 ;
- les 3 suivants : les mêmes variances de bruits (10, 50, 90), une autre valeur de calage aléatoire (100), et des séries numérotées de 109 à 216 ;
- les 3 derniers : les mêmes variances de bruits (10, 50, 90), la valeur de calage aléatoire (50), et des séries numérotées de 217 à 324.

La matrice « temporelle » obtenue à l'issue de la simulation est de dimension (324×500) et est la réalisation d'un modèle vectoriel stationnaire constitué de processus « indépendants ».

3.3 Spécification graphique par les composantes principales

La spécification des séries temporelles par les composantes principales se fait tout d'abord trajectoire par trajectoire simulée. C'est la seconde étape de notre phase d'apprentissage. Elle utilise d'une part la technique SSA pour une visualisation graphique du comportement des deux premières composantes principales, et d'autre part la méthode de Box et Jenkins pour l'identification des modèles dynamiques obtenus.

3.3.1 Visualisation graphique par la technique SSA

Pour l'analyse des composantes principales de chacune des trajectoires, nous appliquons la technique SSA comme référencée dans le chapitre 1, section 1.3.2.1.

Précisément, chacune des chroniques de longueur $p = 500$ est transformée en une matrice 'trajectoire' de dimension $(n \times q) = (376 \times 125)$, avec $n = p - q + 1$ et $q = p/4 (= 125)$ où q est la largeur de la fenêtre choisie.

Puis, sur chacune des matrices « trajectoires », une ACP est construite et nous représentons graphiquement :
- leurs éléments propres (valeurs propres, CP1 et CP2),
- leurs deux premières composantes principales (CP1 scores et CP2 scores) et

- leur premier modèle des scores.

A la vue des nombreuses représentations, nous retenons les graphiques ‘types’ suivants.

3.3.1.1 Le cas des AR(1) négatifs

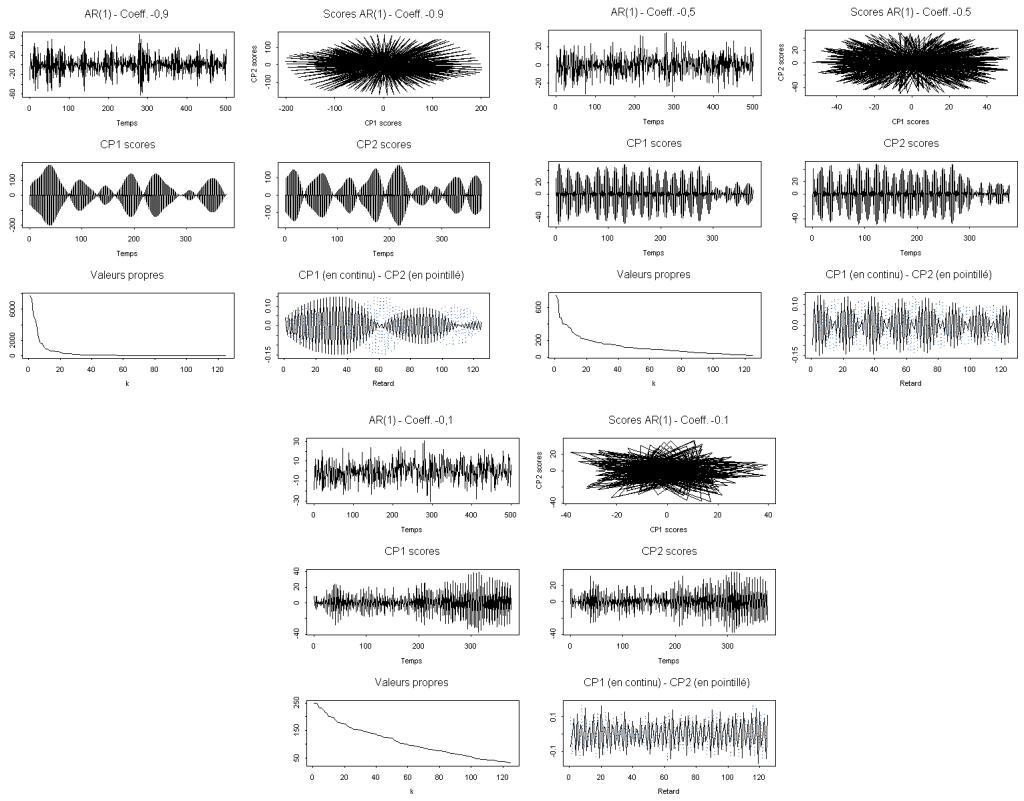


FIG. 3.1 – AR(1) à coefficients négatifs

3.3.1.2 Le cas des MA(1) positifs

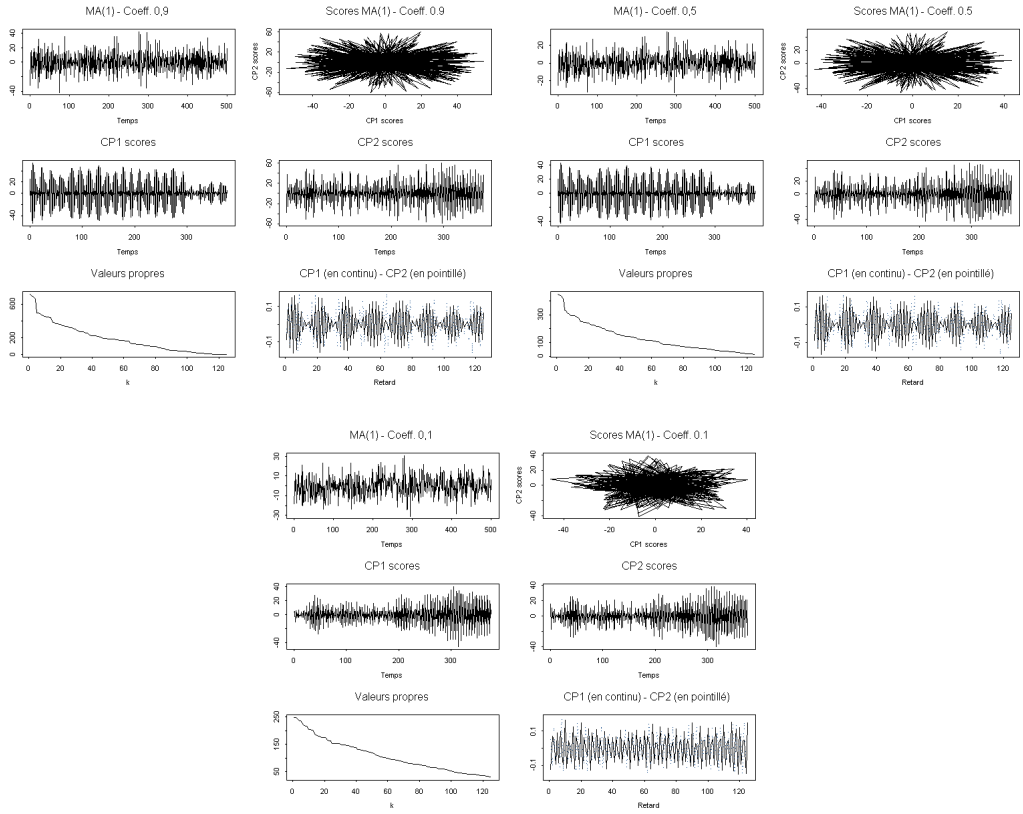
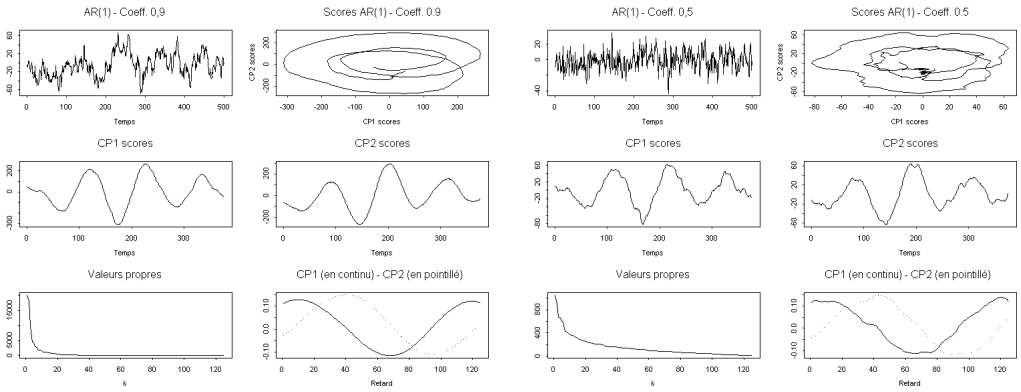


FIG. 3.2 – MA(1) à coefficients positifs

3.3.1.3 Le cas des AR(1) positifs



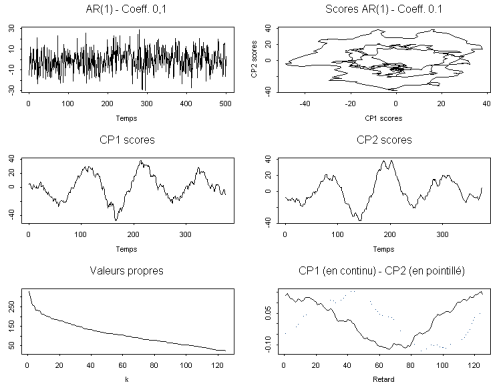


FIG. 3.3 – AR(1) à coefficients positifs

3.3.1.4 Le cas des MA(1) négatifs

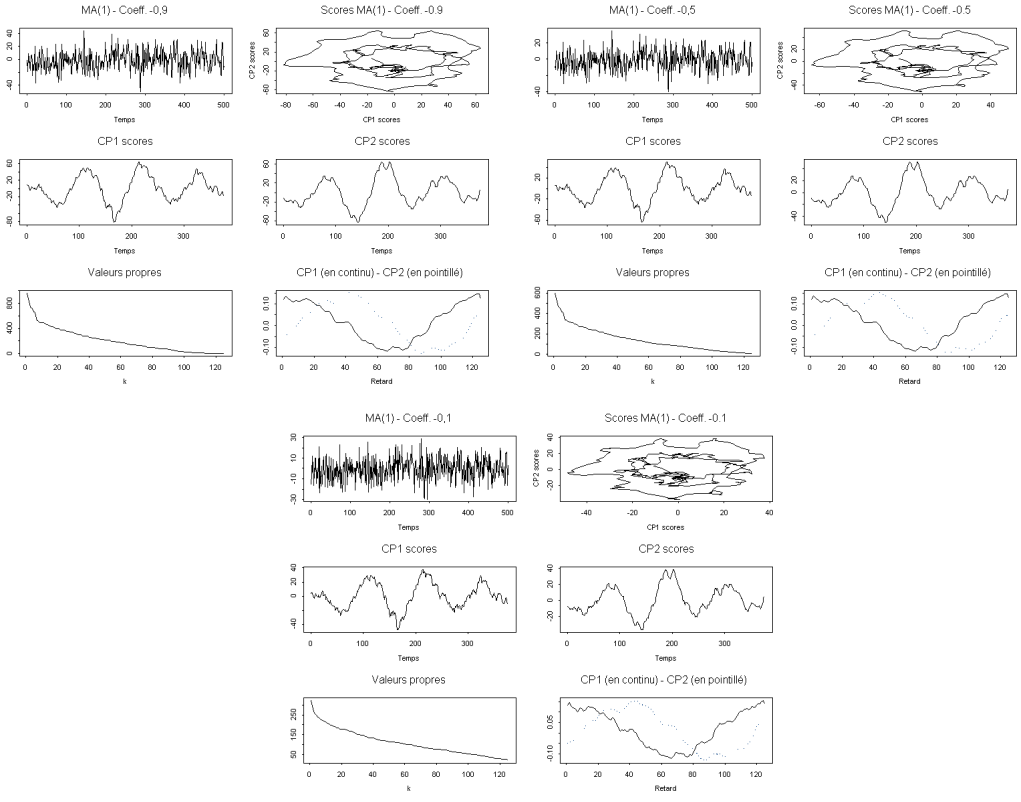


FIG. 3.4 – MA(1) à coefficients négatifs

3.3.1.5 Les premiers résultats

Deux groupes de séries se distinguent, pour deux familles de composantes principales, ou pour deux types de modèles des scores qui ont l'apparence d'attracteurs symétriques comme nous pouvons en trouver dans [Sprott 2004].

Un premier groupe correspond aux AR(1) à coefficients négatifs et aux MA(1) à coefficients positifs (FIG. 3.1 et 3.2), et un deuxième groupe correspond aux AR(1) à coefficients positifs et aux MA(1) à coefficients négatifs (FIG. 3.3 et 3.4).

Pour le premier groupe :

- le modèle des scores a la forme d'une étoile dont la *symétrie* et la complexité reflètent le comportement des première et deuxième composantes principales (CP1 et CP2). En effet, ces dernières se présentent comme des courbes sinusoïdales complexes qui se distinguent par leur période. Nous rencontrons ce type de courbe lorsqu'un phénomène oscillant se trouve perturbé par un signal sinusoïdal proche de sa fréquence.

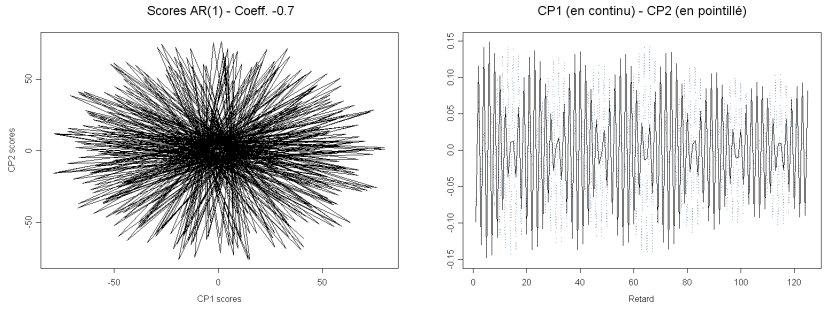


FIG. 3.5 – Modèle des scores et composantes principales

- les première et deuxième composantes principales scores (CP1 scores et CP2 scores) sont aussi des courbes complexes qui pour les MA(1)po doivent se rapprocher de modèles MA(p) d'après les résultats du chapitre 2.

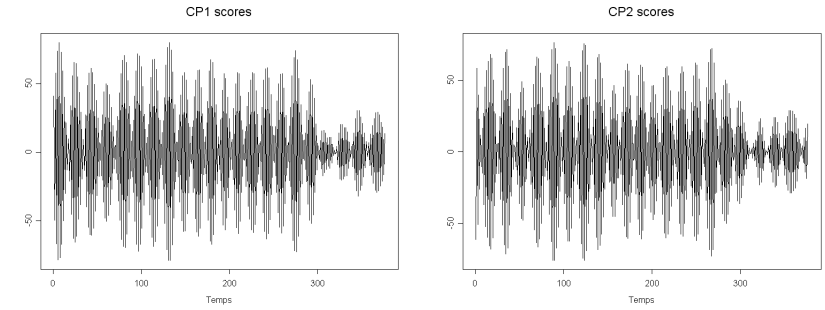


FIG. 3.6 – Composantes principales (scores)

Pour le second groupe :

- le modèle des scores a la forme d'une spirale plus ou moins déformée dont la symétrie reflète le comportement sinusoïdal alterné des première et deuxième composantes principales (CP1 et CP2) ;

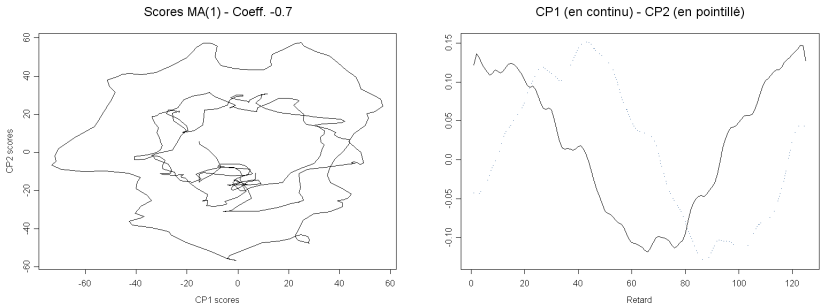


FIG. 3.7 – Modèle des scores et composantes principales

- les première et deuxième composantes principales (CP1 scores et CP2 scores) sont des courbes elles aussi moins complexes en apparence et moins « bruitées ».

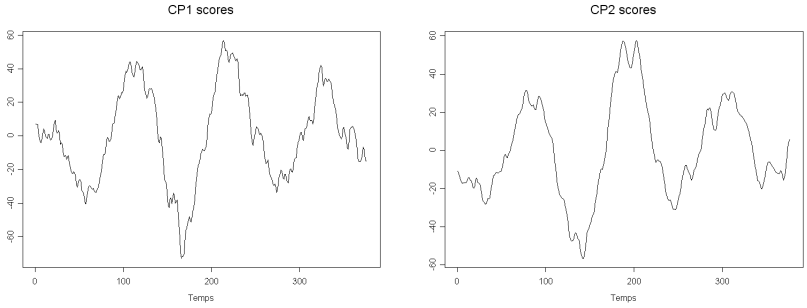


FIG. 3.8 – Composantes principales (scores)

La spécification des séries par les deux premières composantes principales est alors possible.

Remarque : cette symétrie dans le comportement des composantes principales reflète exactement celle des valeurs propres et vecteurs propres associés.

En application des résultats en 2.25 et 2.26 du chapitre 2, nous représentons graphiquement les valeurs propres *réelles* pour chacune des séries AR(1) et MA(1) du premier échantillon. C'est la figure FIG. 3.9.

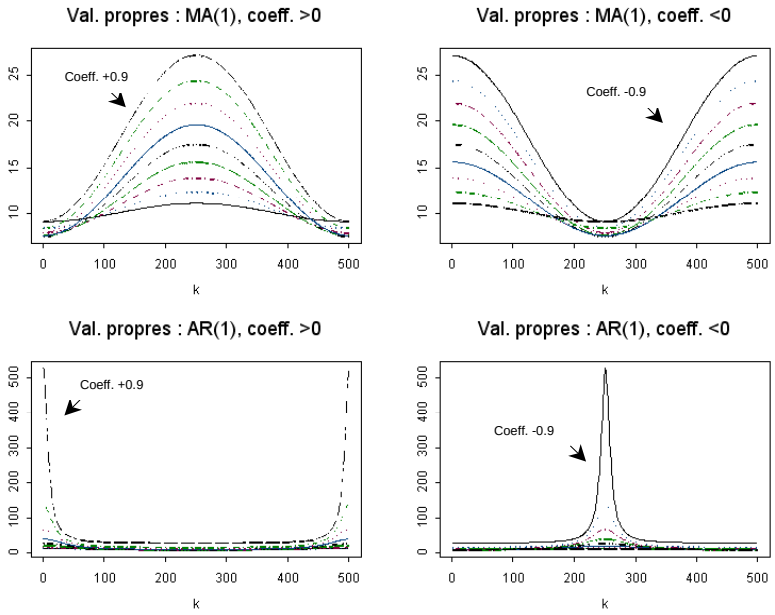


FIG. 3.9 – Valeurs propres des processus AR(1) et MA(1) pour le premier échantillon

Nous pouvons alors constater que :

(a) quels que soient les trajectoires simulées :

- les fonctions (λ_k), $(k = 0, 1, \cdots p-1)$ sont périodiques de période p $(p = 500)$;
- $\lambda_k = \lambda_{p-k}$ sont les valeurs propres doubles $(k \neq 0)$;
- λ_0, $\lambda_{\frac{p}{2}}$ sont les valeurs propres distinctes.

(b) pour le premier groupe, ou pour les processus AR(1) tel que $-1 < \phi_1 < 0$ et les MA(1) tel que $0 < \theta_1 < +1$, les valeurs propres sont ordonnées comme il suit :

- $\lambda_0 < \lambda_1 < \cdots < \lambda_{\frac{p}{2}}$.

(c) pour le second groupe, ou pour les processus AR(1) tel que $0 < \phi_1 < +1$ et les MA(1) tel que $-1 < \theta_1 < 0$, les valeurs propres vérifient :

- $\lambda_0 > \lambda_1 > \cdots > \lambda_{\frac{p}{2}}$.

(d) pour le second groupe :

- les deux premières valeurs propres sont λ_0 et λ_1 ;
- leurs vecteurs propres *réels* associés sont :

$$\overrightarrow{V_0} \sim p^{-1/2}\left[1\,;j=0,\cdots,(p-1)\right]'$$

et

$$\overrightarrow{C_1} \sim p^{-1/2}\left[\cos(2\pi j/p)\,;j=0,\cdots,(p-1)\right]'$$

ou

$$\overrightarrow{S_1} \sim p^{-1/2}\left[\sin(2\pi j/p)\,;j=0,\cdots,(p-1)\right]',$$

obtenus par combinaison linéaire de $\overrightarrow{V_k}$ et $\overrightarrow{V_{p-k}}$ ($\overrightarrow{V_k}+\overrightarrow{V_{p-k}}$ et $i(\overrightarrow{V_k}-\overrightarrow{V_{p-k}})$)

avec $\overrightarrow{V_k} \sim p^{-1/2}\left[\mathrm{e}^{-i(2\pi kj/p)};j=0,\cdots,(p-1)\right]',\ k=0,\ 1,\cdots p-1.$

(e) pour le premier groupe :

- les deux premières valeurs propres sont $\lambda_{\frac{p}{2}}$ et $\lambda_{\frac{p}{2}-1}$;
- leurs vecteurs propres *réels* associés sont :

$$\overrightarrow{V_{\frac{p}{2}}} \sim p^{-1/2}\left[(-1)^j\,;j=0,\cdots,(p-1)\right]'$$

et

$$\overrightarrow{C_{\frac{p}{2}-1}} \sim p^{-1/2}\left[(-1)^j\cos(2\pi j/p)\,;j=0,\cdots,(p-1)\right]'$$

ou

$$\overrightarrow{S_{\frac{p}{2}-1}} \sim p^{-1/2}\left[(-1)^j\sin(2\pi j/p)\,;j=0,\cdots,(p-1)\right]'.$$

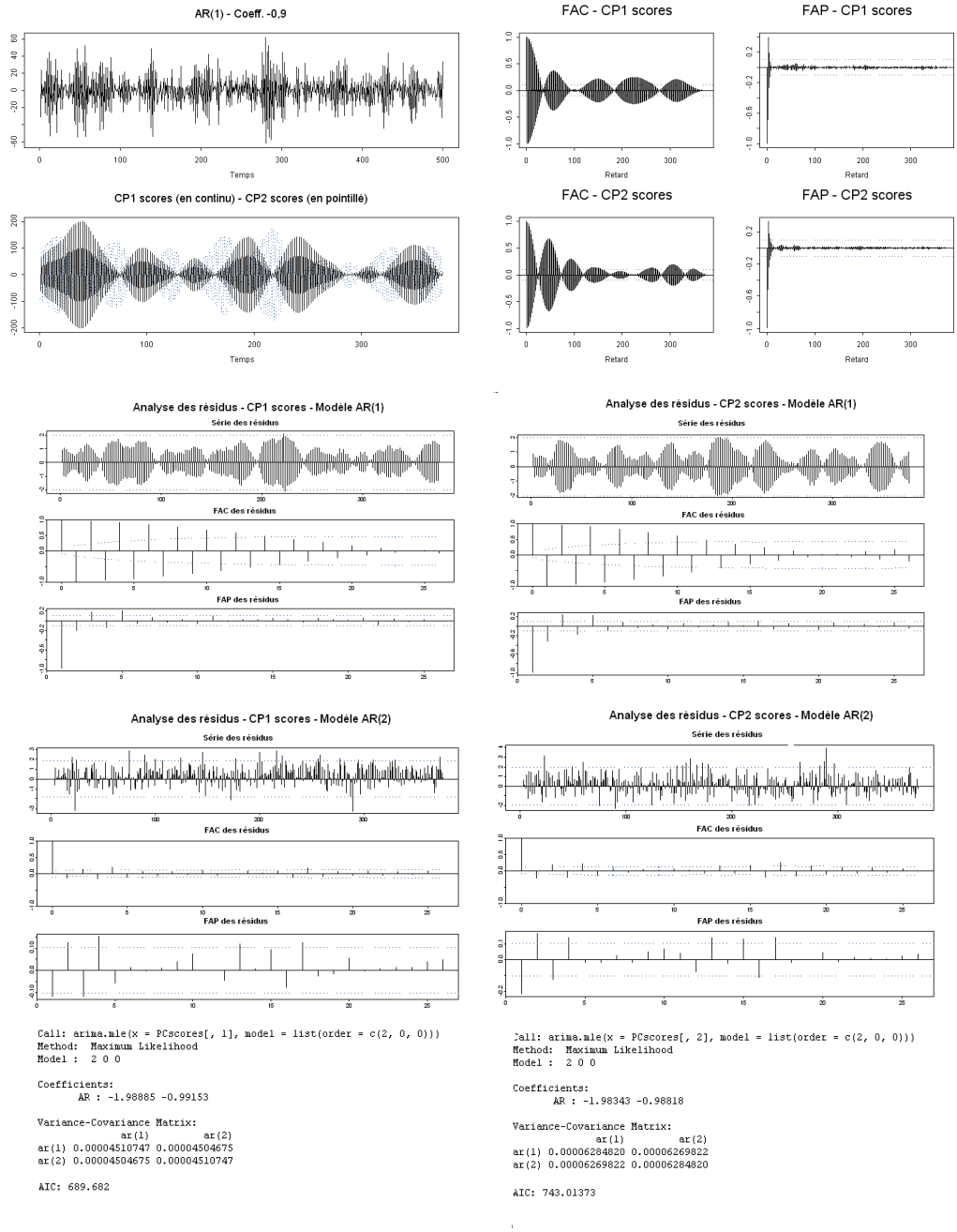
3.3.2 Identification par la méthode de Box et Jenkins

Après la visualisation graphique des composantes principales, nous proposons leur identification par la méthode de Box et Jenkins. Nous rappelons que d'après la section 2.4 du chapitre 2, les modèles doivent se rapprocher de MA(p) pour un MA(1) si p est le nombre de variables étudiées, et de moyennes mobiles à préciser pour un AR(1).

Tout d'abord, nous présentons une sélection de graphiques « diagnostics » obtenus sur des trajectoires types, suite :

- à l'analyse des corrélations,
- à l'estimation des paramètres et tests AIC,
- et à l'analyse des résidus, comme il est rappelé en section 3.1 de ce chapitre.

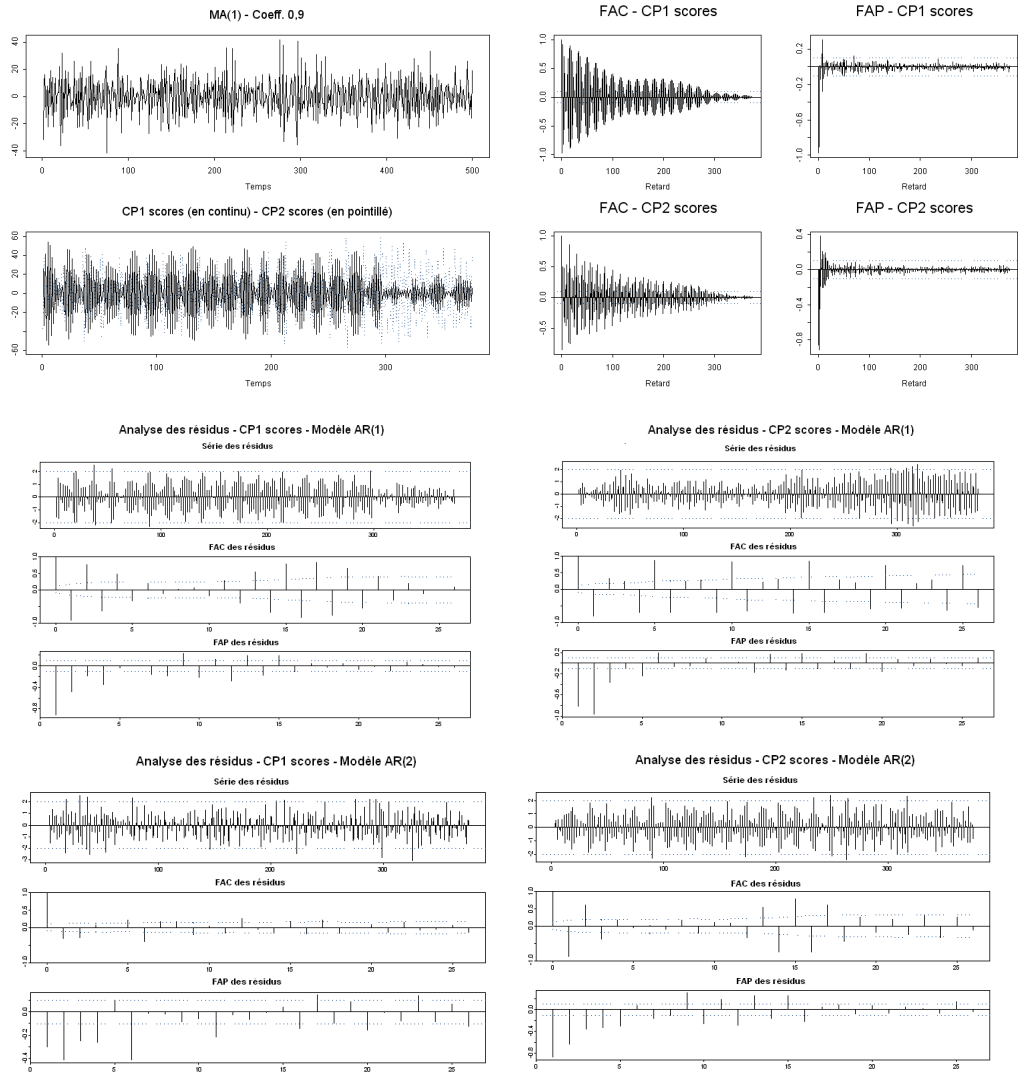
3.3.2.1 Le cas du premier groupe




```
Call: arima.mle(x = PCscores[, 1], model = list(order = c(2, 0, 0)))
Method:  Maximum Likelihood
Model :  2 0 0

Coefficients:
       AR : -1.98885 -0.99153

Variance-Covariance Matrix:
             ar(1)         ar(2)
ar(1) 0.00004510747 0.00004504675
ar(2) 0.00004504675 0.00004510747

AIC: 689.682
```

```
Call: arima.mle(x = PCscores[, 2], model = list(order = c(2, 0, 0)))
Method:  Maximum Likelihood
Model :  2 0 0

Coefficients:
       AR : -1.98343 -0.98818

Variance-Covariance Matrix:
             ar(1)         ar(2)
ar(1) 0.00006284820 0.00006269822
ar(2) 0.00006269822 0.00006284820

AIC: 743.01373
```

FIG. 3.10 – Diagnostics : AR(1) à coefficient négatif

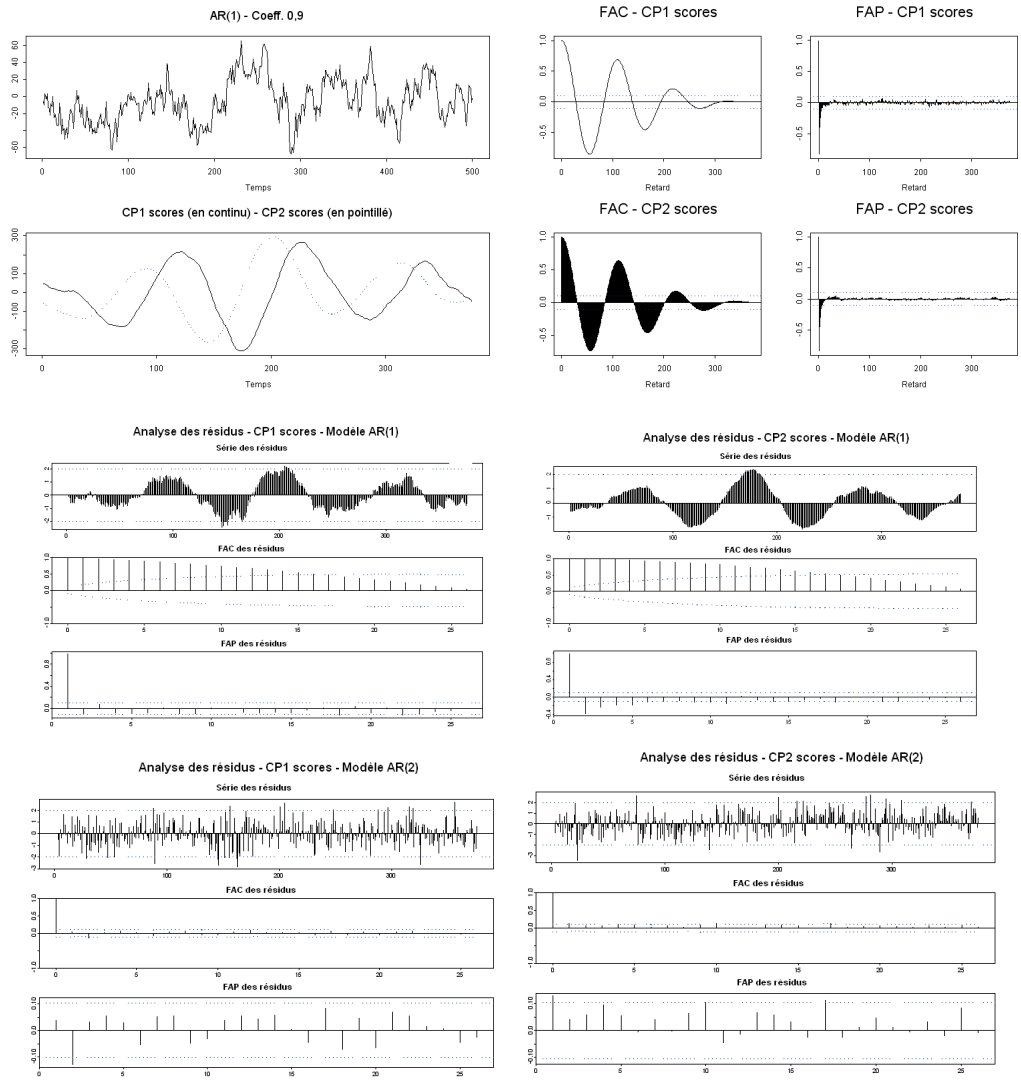



```
Call: arima.mle(x = PCscores[, 1], model = list(order = c(2, 0, 0)))
Method:  Maximum Likelihood
Model :  2 0 0

Coefficients:
       AR : -1.89941 -0.9403

Variance-Covariance Matrix:
             ar(1)        ar(2)
ar(1) 0.0003097047 0.0003031781
ar(2) 0.0003031781 0.0003097047

AIC: 1496.68951
```

```
Call: arima.mle(x = PCscores[, 2], model = list(order = c(2, 0, 0)))
Method:  Maximum Likelihood
Model :  2 0 0

Coefficients:
       AR : -1.67774 -0.94397

Variance-Covariance Matrix:
             ar(1)        ar(2)
ar(1) 0.0002912359 0.0002513507
ar(2) 0.0002513507 0.0002912359

AIC: 2170.22455
```

FIG. 3.11 – Diagnostics : MA(1) à coefficient positif

3.3.2.2 Le cas du deuxième groupe




```
Call: arima.mle(x = PCscores[, 1], model = list(order = c(2, 0, 0)))
Method:  Maximum Likelihood
Model :  2 0 0

Coefficients:
       AR : 1.95743 -0.9608

Variance-Covariance Matrix:
                ar(1)           ar(2)
ar(1)  0.0002055299 -0.0002051772
ar(2) -0.0002051772  0.0002055299

AIC: 1397.57184
```

```
Call: arima.mle(x = PCscores[, 2], model = list(order = c(2, 0, 0)))
Method:  Maximum Likelihood
Model :  2 0 0

Coefficients:
       AR : 1.99521 -0.99805

Variance-Covariance Matrix:
                ar(1)           ar(2)
ar(1)  0.00001041311 -0.00001039829
ar(2) -0.00001039829  0.00001041311

AIC: 513.74397
```

FIG. 3.12 – Diagnostics : AR(1) à coefficient positif

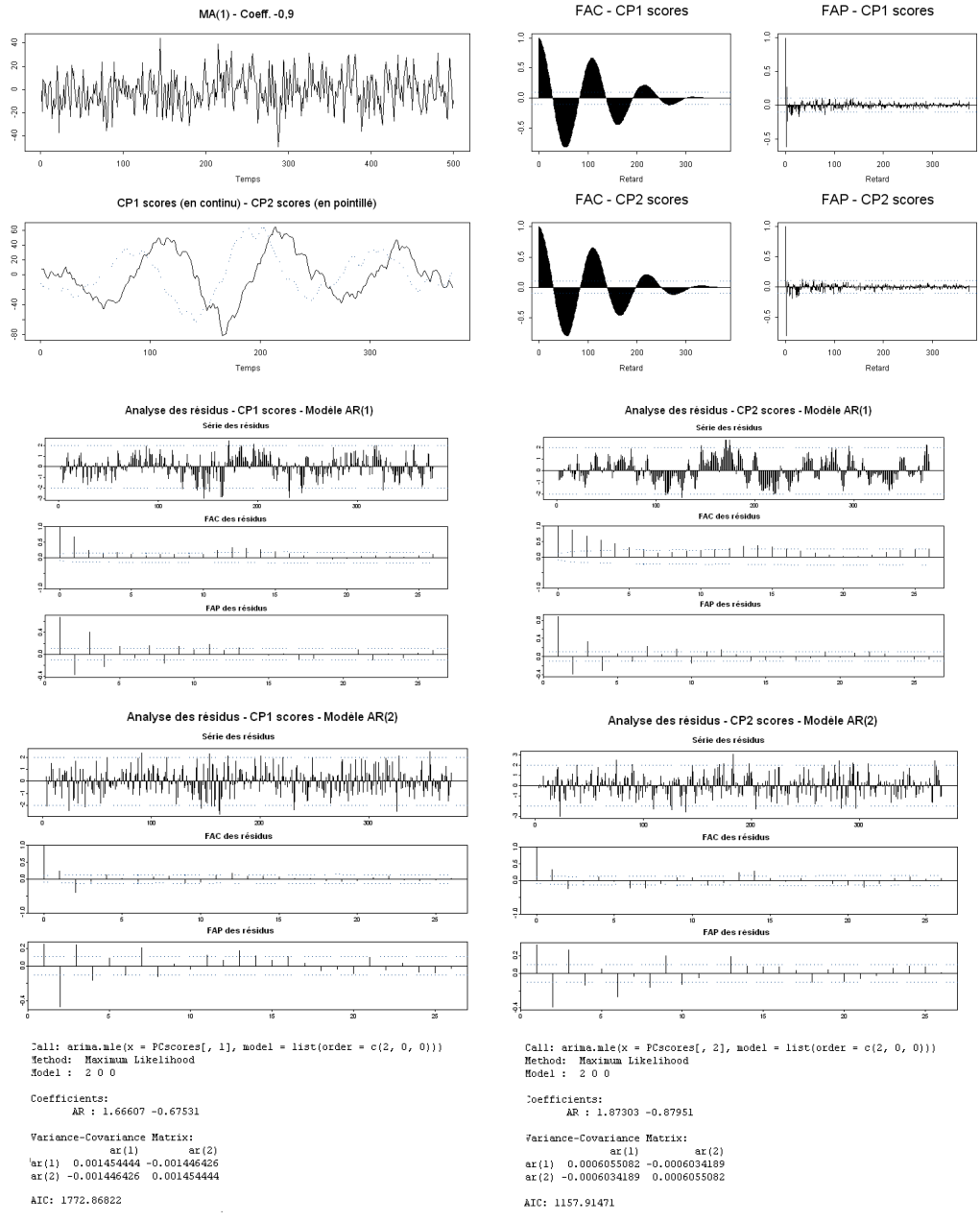


FIG. 3.13 – Diagnostics : MA(1) à coefficient négatif

3.3.2.3 Des modèles AR(2) pour les composantes principales

La stationnarité des composantes principales est acquise à la vue des fonctions FAC et FAP qui, pour chacune des trajectoires étudiées, s'annulent significativement.

Le choix des ordres maximaux p et q des parties AR et MA est aussi facilité à la lecture des corrélogrammes. Que ce soit pour le premier ou le deuxième groupe, un ou deux pics significatifs peuvent être retenus pour la FAP, alors que la FAC se comporte comme une sinusoïde amortie dont la forme est plus complexe dans le cas du premier groupe. Nous retrouvons la perturbation déjà évoquée pour les CP1 et CP2. En référence à la propriété 3.1, nous pouvons déjà retenir les ordres $p = 2$ ou $p = 1$, et $q = 0$ pour le modèle ARMA qui engendrerait les composantes principales.

Après l'estimation des paramètres, le test AIC et l'analyse des résidus, nous choisissons comme meilleurs modèles :

pour le premier groupe :

- *des processus AR(2) dont les coefficients sont tels que* $\phi_1 < 0$ et $\phi_1^{\,2} + 4\phi_2 < 0$ en référence au tableau TABLEAU 3.1*, avec* ϕ_1 *proche de -2 et* ϕ_2 *proche de –1* ;
- plutôt que des processus AR(1) avec un coefficient ϕ_1 proche de –1, pour lesquels l'analyse des résidus est moins satisfaisante.

pour le second groupe :

- *des processus AR(2) dont les coefficients sont tels que* $\phi_1 > 0$ et $\phi_1^{\,2} + 4\phi_2 < 0$ en référence au tableau TABLEAU 3.1*, avec* ϕ_1 *proche de +2 et* ϕ_2 *proche de –1* ;
- plutôt que des processus AR(1) qui tendent vers des marches aléatoires avec un coefficient ϕ_1 très proche de +1, pour lesquels aussi l'analyse des résidus est moins satisfaisante.

➢ ***Les deux premières composantes principales ou les moyennes mobiles obtenues en fin de chapitre 2 se rapprochent de modèles AR(2) avec :***

- $\phi_1 < 0$ et $\phi_1^{\,2} + 4\phi_2 < 0$ pour le premier groupe ;
- $\phi_1 > 0$ et $\phi_1^{\,2} + 4\phi_2 < 0$ pour le deuxième groupe.

3.4 ACP temporelles de processus AR(1) et MA(1)

Jusque là, la spécification par les composantes principales se faisait série par série, et nous avons pu constater que les trajectoires se regroupent ou s'opposent en fonction de leurs deux premières composantes principales.

Désormais, nous travaillons avec un ou plusieurs échantillons de trajectoires simulées, et nous construisons des ACP dites « temporelles » directement sur la matrice de données issue de la simulation.

Nous rappelons que les échantillons sont des réalisations de modèles vectoriels stationnaires constitués de processus « indépendants ». Dans ces conditions et en référence à la propriété 1.13 du chapitre 1, les premiers modèles des scores issus des ACP temporelles doivent *refléter les regroupements et oppositions* des séries rencontrés précédemment.

Ces modèles doivent aussi *fournir la base* de graphiques de projection d'une nouvelle série pour son identification et une première estimation de ses paramètres.

Pour le vérifier, quatre ACP temporelles sont construites sur un, trois et neuf échantillons des trajectoires AR(1) et MA(1) simulées :

- la première ACP est réalisée sur le premier échantillon des trajectoires numérotées de 1 à 36, avec 18 AR(1) et 18 MA(1) « indépendants » et stationnaires, et tous issus d'un même bruit blanc (même valeur initiale de calage aléatoire et même variance de bruit blanc) ;
- la seconde ACP est construite sur les trois derniers échantillons dont les trajectoires se distinguent par leurs bruits blancs (même valeur initiale de calage aléatoire mais une variance différente par échantillon (10, 50, 90) ;
- la troisième ACP basée sur les neuf échantillons, a pour objectif de représenter la variabilité des paramètres que sont la valeur initiale de calage aléatoire et la variance d'un bruit blanc posé gaussien ;
- la quatrième ACP est aussi basée sur les neuf échantillons mais nous avons extrait à titre d'exemple les neuf trajectoires AR(1) de coefficient –0.7 pour les projeter comme éléments supplémentaires ou « illustratifs » dans le modèle des scores.

3.4.1 Modèles des scores temporels

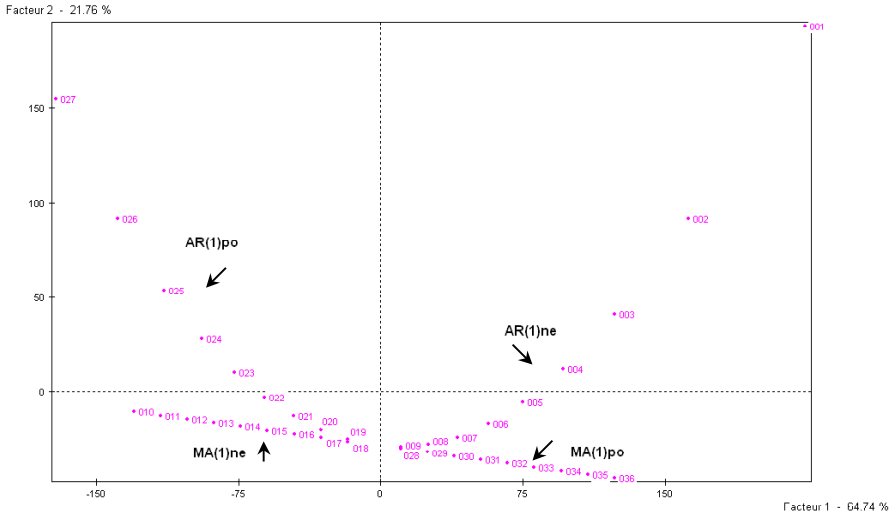


FIG. 3.14 – Modèle des scores sur le 1er échantillon

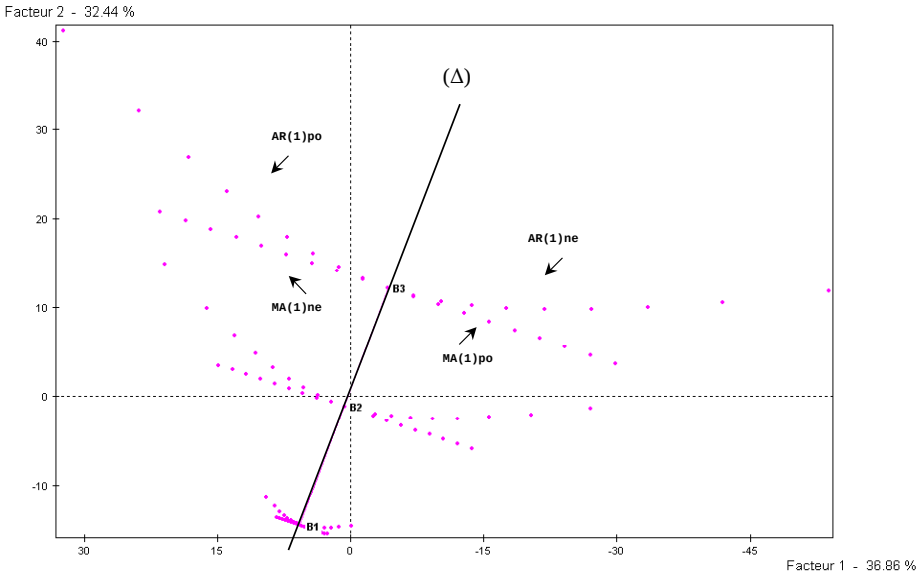


FIG. 3.15 – Modèle des scores sur les trois derniers échantillons

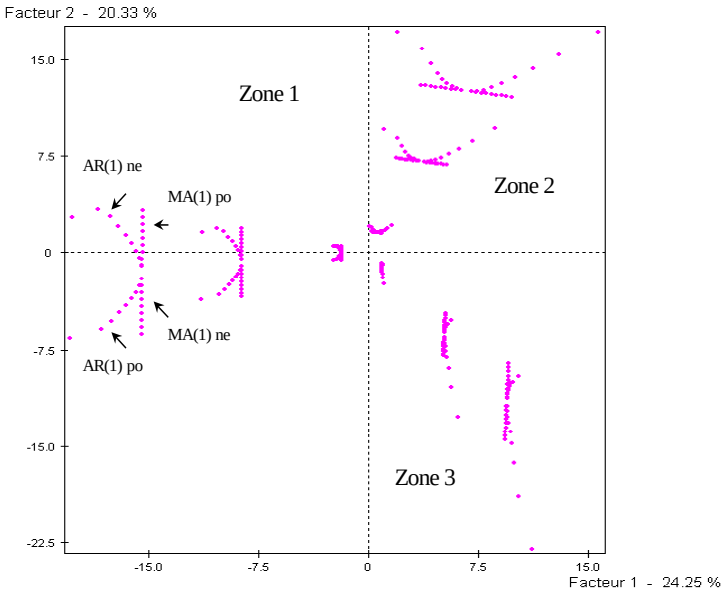


FIG. 3.16 – Modèle des scores sur les neuf échantillons

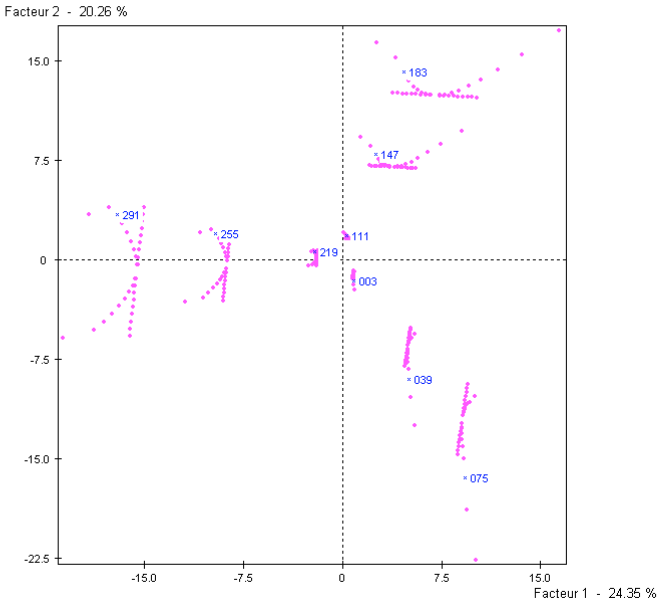


FIG. 3.17 – Modèles des scores sur les neuf échantillons dont neuf individus illustratifs

3.4.2 Analyse graphique

Les figures FIG. 3.14 et FIG. 3.15 :

Les figures FIG. 3.14 et 3.15 illustrent parfaitement la propriété 1.13 du chapitre 1. En effet, dans le plan factoriel (F1,F2) nous retrouvons les regroupements et oppositions des séries évoqués précédemment. Un axe Δ, qui n'est ni F1 ni F2, sépare d'une part les AR(1) à coefficients négatifs et les MA(1) à coefficients positifs, et d'autre part les AR(1) à coefficients positifs et les MA(1) à coefficients négatifs. Ces regroupements et oppositions des séries sont conservés grâce à la *propriété des scores* d'un modèle vectoriel constitué de processus « indépendants ».

Montrons alors que :

- pour un même échantillon, les processus MA(1) à coefficients symétriques ont même milieu et que ce milieu est le bruit blanc ;
- un même axe Δ passe par les 3 bruits blancs dont sont issus les 3 échantillons des trajectoires AR(1) et MA(1) simulées.

Démonstrations :

Remarque : les résultats à démontrer sont vrais dans tout plan factoriel (F_{k_1}, F_{k_2}).

Pour les MA(1) d'un même échantillon :

Soient λ_{k_1} et λ_{k_2} deux valeurs propres et, $\overrightarrow{V_{k_1}}$ et $\overrightarrow{V_{k_2}}$ leurs vecteurs propres associés.

En référence à la propriété 2.16 du chapitre 2, si (Z_t) est un processus MA(1), ses scores dans le plan factoriel (F_{k_1}, F_{k_2}) vérifient :

$$\left[a_p^{(k_1)}\right]^{-1} f_{k_1}(t) = a(t) + \sum_{J=1}^{p} \beta_J^{(k_1)} a(t-J)$$

avec $\beta_J^{(k_1)} = (-a_{p-J+1}^{(k_1)}\theta_1 + a_{p-J}^{(k_1)}) / a_p^{(k_1)}$ $(J = 1,\cdots,p)$ et $(a_0^{(k_1)} = 0)$

et

$$\left[a_p^{(k_2)}\right]^{-1} f_{k_2}(t) = a(t) + \sum_{J=1}^{p} \beta_J^{(k_2)} a(t-J)$$

avec $\beta_J^{(k_2)} = (-a_{p-J+1}^{(k_2)}\theta_1 + a_{p-J}^{(k_2)}) / a_p^{(k_2)}$ $(J = 1,\cdots,p)$ et $(a_0^{(k_2)} = 0)$

Il s'ensuit que les milieux de 2 processus MA(1) à coefficients symétriques $-|\theta_1|$ et $|\theta_1|$, notés $MA(1)_{ne}^{-|\theta_1|}$ et $MA(1)_{po}^{|\theta_1|}$, ont pour coordonnées :

$$\frac{f_{k_1}^{MA(1)_{po}^{|\theta_1|}}(t) + f_{k_1}^{MA(1)_{ne}^{-|\theta_1|}}(t)}{2} = \left[a_p^{(k_1)}\right] a(t) + \sum_{J=1}^{p} a_{p-J}^{(k_1)}\, a(t-J)$$

et

$$\frac{f_{k_2}^{MA(1)_{po}^{|\theta_1|}}(t) + f_{k_2}^{MA(1)_{ne}^{-|\theta_1|}}(t)}{2} = \left[a_p^{(k_2)}\right] a(t) + \sum_{J=1}^{p} a_{p-J}^{(k_2)}\, a(t-J)$$

Les scores obtenus sont indépendants de θ_1 et coïncident exactement avec ceux du bruit blanc (propriété 2.14 avec $\theta_1 = 0$). Les processus MA(1) à coefficients symétriques ont donc même milieu qui est le bruit blanc. Il vient que tous les processus MA(1) sont alignés avec le bruit blanc.

Un même axe Δ passe par les trois bruits blancs :

Soient $a_1(t)$, $a_2(t)$ et $a_3(t)$ les bruits blancs pour chacun des échantillons, et σ_1, σ_2, σ_3, leurs variances respectives (non nulles).

Comme ils sont issus d'une même valeur initiale de calage aléatoire, ils vérifient en particulier :

$$\frac{a_1(t)}{\sigma_1} = \frac{a_2(t)}{\sigma_2} = \frac{a_3(t)}{\sigma_3}.$$

Il vient que si B_1, B_2 et B_3 sont les 3 bruits blancs dans le plan factoriel (F_{k_1}, F_{k_2}), leurs coordonnées s'écrivent :

pour B_1 :

$$f_{k_1}^{B_1}(t) = \left[a_p^{(k_1)}\right] a_1(t) + \sum_{J=1}^{p} a_{p-J}^{(k_1)}\, a_1(t-J)$$

$$f_{k_2}^{B_1}(t) = \left[a_p^{(k_2)}\right] a_1(t) + \sum_{J=1}^{p} a_{p-J}^{(k_2)}\, a_1(t-J)$$

pour B_2 :

$$f_{k_1}^{B_2}(t) = \frac{\sigma_2}{\sigma_1}\left(\left[a_p^{(k_1)}\right] a_1(t) + \sum_{J=1}^{p} a_{p-J}^{(k_1)}\, a_1(t-J)\right) = \frac{\sigma_2}{\sigma_1} f_{k_1}^{B_1}(t)$$

$$f_{k_2}^{B_2}(t) = \frac{\sigma_2}{\sigma_1}\left(\left[a_p^{(k_2)}\right]a_1(t) + \sum_{J=1}^{p} a_{p-J}^{(k_2)}\, a_1(t-J)\right) = \frac{\sigma_2}{\sigma_1} f_{k_2}^{B_1}(t)\,,$$

et pour B_3 **:**

$$f_{k_1}^{B_3}(t) = \frac{\sigma_3}{\sigma_1}\left(\left[a_p^{(k_1)}\right]a_1(t) + \sum_{J=1}^{p} a_{p-J}^{(k_1)}\, a_1(t-J)\right) = \frac{\sigma_3}{\sigma_1} f_{k_1}^{B_1}(t)$$

$$f_{k_2}^{B_3}(t) = \frac{\sigma_3}{\sigma_1}\left(\left[a_p^{(k_2)}\right]a_1(t) + \sum_{J=1}^{p} a_{p-J}^{(k_2)}\, a_1(t-J)\right) = \frac{\sigma_3}{\sigma_1} f_{k_1}^{B_1}(t)\,.$$

D'où :

$$(\frac{\sigma_2}{\sigma_1} - 1) f_{k_1}^{B_1}(t) \times (\frac{\sigma_3}{\sigma_1} - 1) f_{k_2}^{B_1}(t) = (\frac{\sigma_2}{\sigma_1} - 1) f_{k_2}^{B_1}(t) \times (\frac{\sigma_3}{\sigma_1} - 1) f_{k_1}^{B_1}(t)\,,$$

c'est à dire $\det(\overrightarrow{B_1B_2}, \overrightarrow{B_1B_3}) = 0$ ou encore les 3 points B_1, B_2 et B_3 sont alignés.

Fin démonstrations.

La figure FIG. 3.16 :

Comme conséquence directe des résultats précédents, il s'ensuit l'existence de trois axes de symétrie qui chacun rassemble les échantillons issus d'un bruit blanc à même valeur de calage aléatoire : c'est la figure FIG. 3.16.

Un premier niveau de regroupement des séries se fait donc *par zone de simulation* ou par valeur initiale de calage aléatoire du bruit blanc. La Zone 1 regroupe les échantillons 7 à 9, la Zone 2 regroupe les échantillons 4 à 6, et la Zone 3 regroupe les échantillons 1 à 3.

Un deuxième niveau de regroupement des séries est celui *par échantillon dans chaque zone.* Les regroupements se font alors par variance du bruit blanc pour une même valeur de calage aléatoire.

Un troisième et dernier niveau de regroupement des séries est celui déjà rencontré dans les FIG. 3.14 et 3.15. Pour chaque échantillon, il s'agit des *regroupements et oppositions de séries* en *fonction de leurs coefficients.*

La figure FIG. 3.17 :

Puis, neuf trajectoires AR(1) de coefficient –0.7 sont projetées comme éléments supplémentaires ou « illustratifs » dans le modèle des scores en FIG. 3.17. Nous pouvons alors constater que les FIG. 3. 16 et 3.17 sont très proches l'une de l'autre. En effet, les neuf trajectoires AR(1) retrouvent la place qu'elles occupaient dans la FIG. 3.16 : *la propriété 1.13 est à nouveau vérifiée graphiquement.*

Des AR(1) pour les deux premières composantes principales du modèle multivarié :

De même que pour l'identification des moyennes mobiles ou composantes principales de chacune des trajectoires, nous avons recours à la méthodologie de Box et Jenkins pour identifier les modèles qui lient les scores pour la première valeur propre et les scores pour la seconde valeur propre.

Après l'analyse des corrélations, l'estimation des paramètres par la méthode du maximum de vraisemblance, le test AIC et l'analyse des corrélations, nous déduisons comme meilleurs modèles, des *processus AR(1)* qui pour la 3ième ACP *tendent vers des marches aléatoires*.

Pour la 1ère ACP :

- le premier axe est un AR(1) avec comme coefficient $\phi_1 \sim 0.81$;
- le deuxième axe est un AR(1) avec comme coefficient $\phi_1 \sim 0.55$.

Pour la 2ième ACP :

- le premier axe est un AR(1) avec comme coefficient $\phi_1 \sim 0.84$;
- le deuxième axe est un AR(1) avec comme coefficient $\phi_1 \sim 0.93$.

Pour la 3ième ACP :

- le premier axe est un AR(1) qui tend vers une marche aléatoire ($\phi_1 \sim 0.98$) ;
- le deuxième axe est un AR(1) qui tend vers une marche aléatoire ($\phi_1 \sim 0.99$).

Pour la dernière ACP, nous présentons les courbes CP1 scores et CP2 scores en FIG. 3.18, les corrélogrammes en FIG. 3.19, puis les estimations des coefficients, le test AIC et l'analyse des résidus en FIG. 3.20.

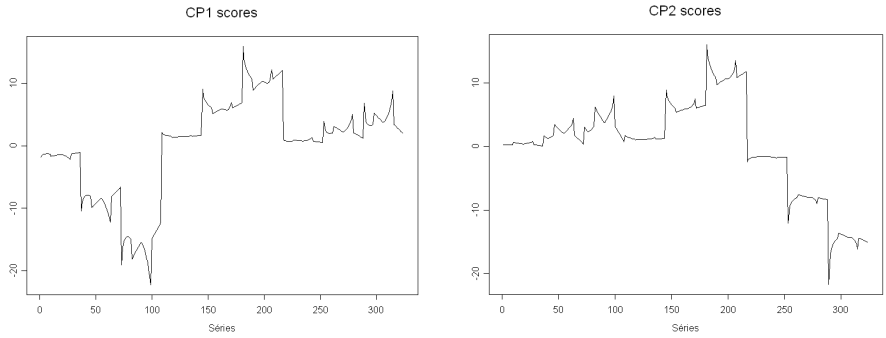


FIG. 3.18 – Séries CP1 scores et CP2 scores

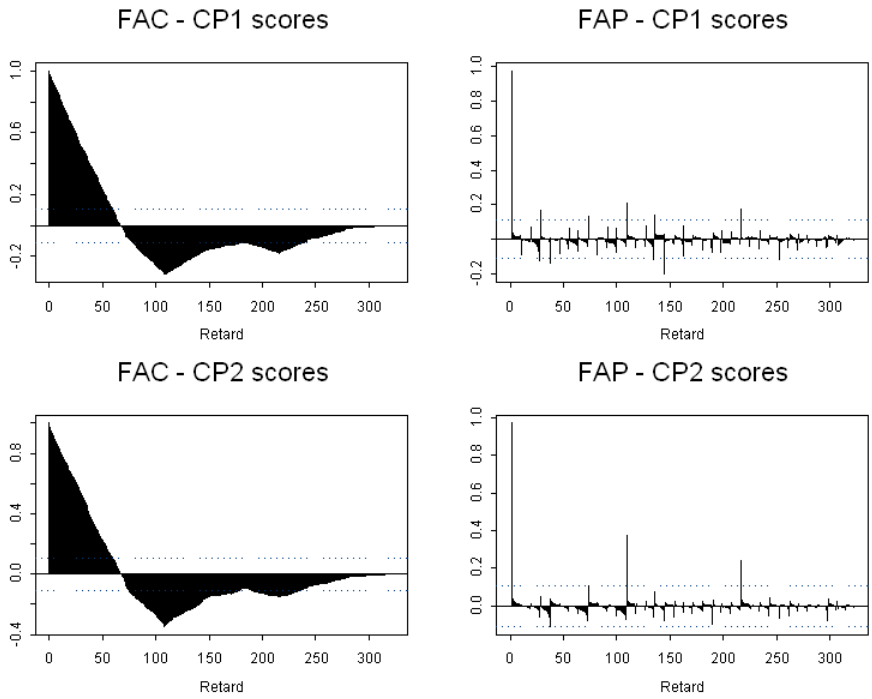
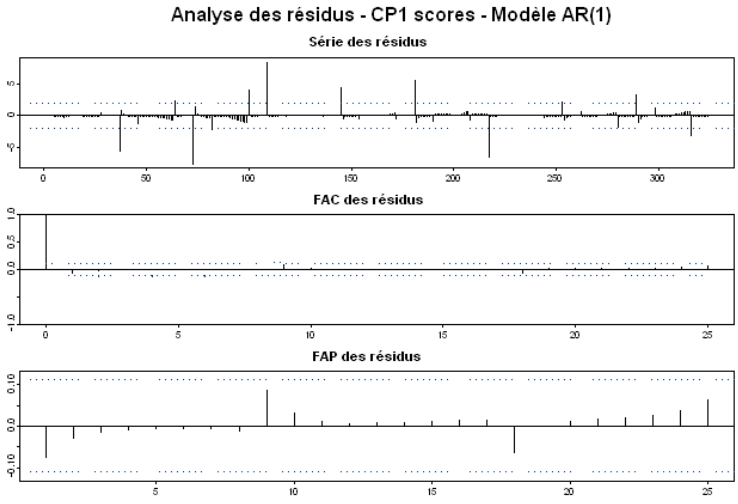
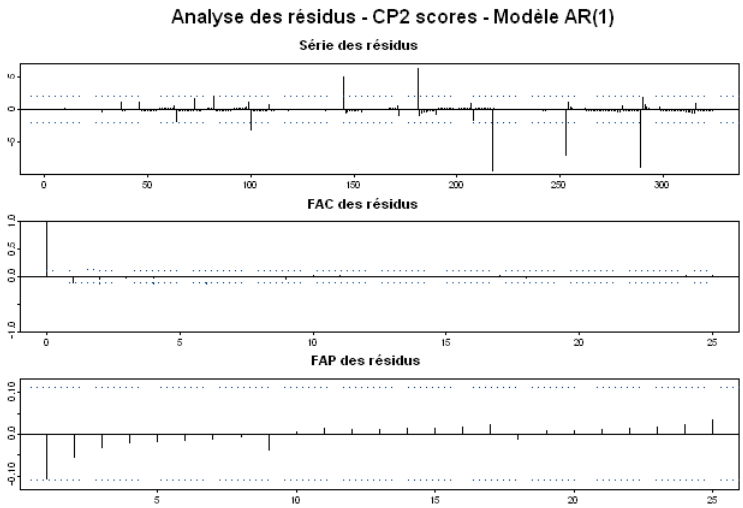


FIG. 3.19 – Corrélogrammes






```
Call: arima.mle(x = PCscores[, 1], model = list(order = c(1, 0, 0)))
Method:  Maximum Likelihood
Model :  1 0 0

Coefficients:
       AR : 0.97579

Variance-Covariance Matrix:
             ar(1)
ar(1) 0.0001481208

Optimizer has  converged
Convergence Type: relative function convergence
AIC: 1257.35494
Call: arima.mle(x = PCscores[, 2], model = list(order = c(1, 0, 0)))
Method:  Maximum Likelihood
Model :  1 0 0

Coefficients:
       AR : 0.98536

Variance-Covariance Matrix:
              ar(1)
ar(1) 0.00009001521

Optimizer has  converged
Convergence Type: relative function convergence
AIC: 1186.73436
```

FIG. 3.20 – Estimation des paramètres et analyse des résidus

Les limites de ces premiers modèles :

Les premiers modèles des scores obtenus (FIG 3.14 à 3.17) reflètent assez bien les regroupements et oppositions des séries que nous attendions, d'une part des résultats théoriques des chapitres 1 et 2, et d'autre part de la spécification graphique des composantes principales effectuée en section 3.3. Cependant, comme nous avons pu le constater en FIG. 3.16, les modèles *reflètent avant tout la variabilité des bruits de la simulation*. Le premier niveau de regroupement des séries ne se faisant pas en fonction de leurs coefficients, ces modèles ne fournissent donc pas encore la base de graphiques de projection d'une nouvelle série pour son

identification et une première estimation de ses paramètres.

Le problème est alors le suivant : est-il possible de concevoir un modèle factoriel qui rapproche les séries en fonction de leurs coefficients, quels que soient les paramètres du bruit blanc ou encore quels que soient les échantillons ?

Un bon moyen pour atténuer les bruits est le recours aux fonctions FAC et FAP qui en théorie s'annulent pour de tels processus. C'est ce qui fait l'objet de la partie suivante.

3.5 ACP sur des éléments de la FAC et FAP

C'est avec deux éléments de la FAC que nous construisons les premières ACP sur un et neuf échantillons. Ces éléments ou variables sont calculés directement sur les données issues de la simulation, et correspondent aux retards un et deux pour la FAC. Ils sont notés respectivement (C1=FAC1) et (C2=FAC2). Nous en déduisons deux cercles des corrélations « remarquables » et deux modèles des scores dont l'un d'entre eux rappelle la FIG. 3.14. Ce sont les FIG. 3.21 à 3.23.

D'autres ACP sont testées avec un nombre de retards allant jusqu'à dix pour la FAC, mais la qualité des modèles n'est pas améliorée. Nous avons alors recours à la fonction FAP pour laquelle nous considérons jusqu'à dix retards. Nous retenons finalement un seul retard significatif noté FAP2 qui correspond au second retard pour la FAP. Rappelons que pour les MA(1) comme pour les AR(1), l'élément FAP1 est égal à l'élément FAC1 (TABLEAU 3.1 et TABLEAU 3.2). Les variables étudiées sont donc (FAC1,FAC2,FAP2). Nous en déduisons trois autres modèles des scores sur respectivement un, trois et neuf échantillons à l'image des ACP temporelles. Ce sont les figures FIG. 3.24, FIG. 3.25 et FIG. 3.26. Pour illustrer quelques propriétés de la dernière ACP, nous ajoutons le cercle des corrélations en FIG. 3.27.

Enfin, un quatrième modèle des scores (ou FIG. 3.28) illustre la projection de trois séries AR(1) à coefficient -0.7, dans le modèle factoriel de référence basé sur les neuf échantillons.

3.5.1 Modèles factoriels basés sur (FAC1,FAC2)

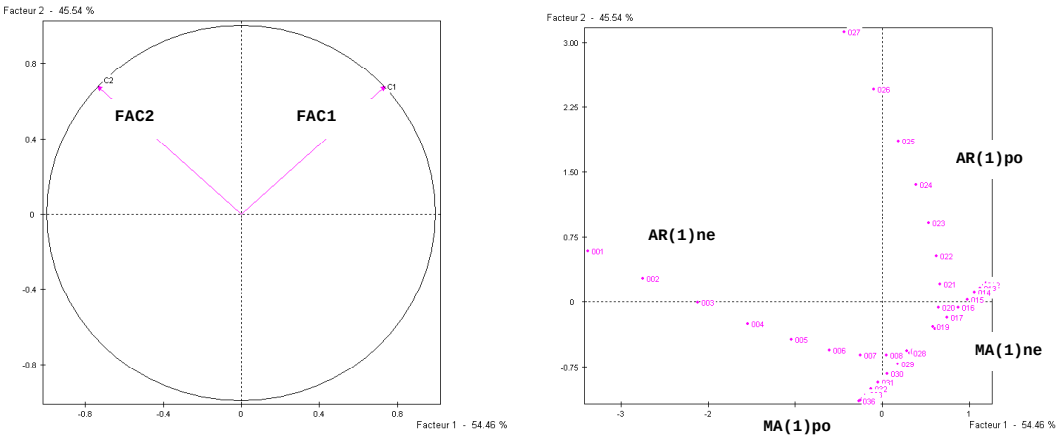


FIG. 3.21 – Modèles factoriels sur un échantillon

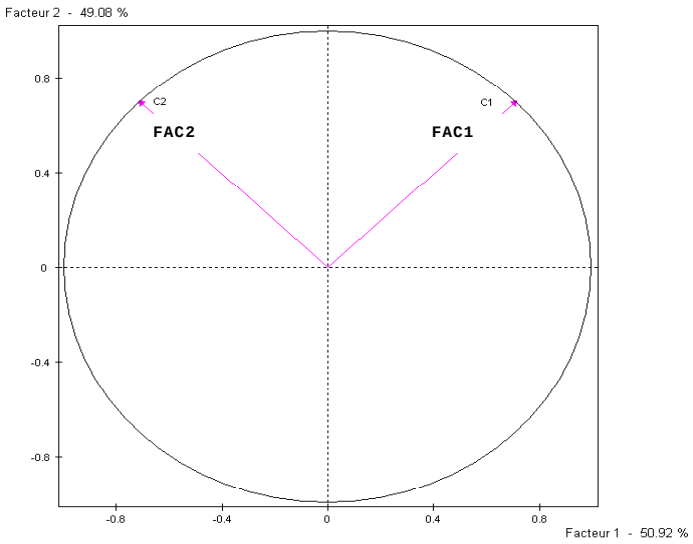


FIG. 3.22 – Cercle des corrélations sur les neuf échantillons

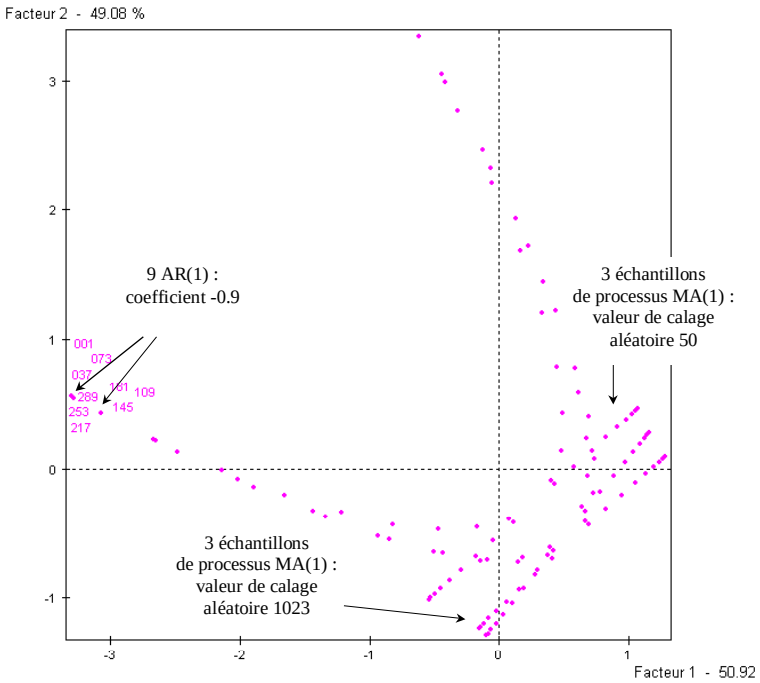


FIG. 3.23 – Modèle des scores sur les neuf échantillons

3.5.2 Modèles des scores basés sur (FAC1,FAC2,FAP2)

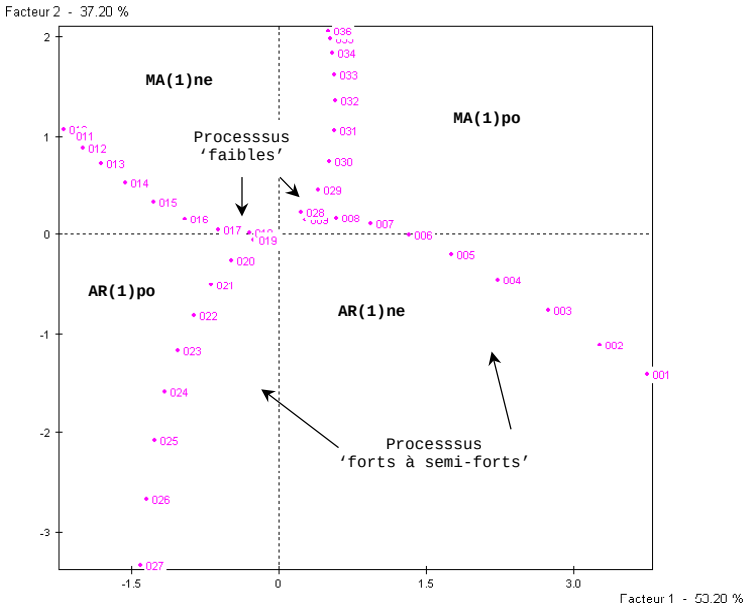


FIG. 3.24 – Modèle des scores sur un échantillon

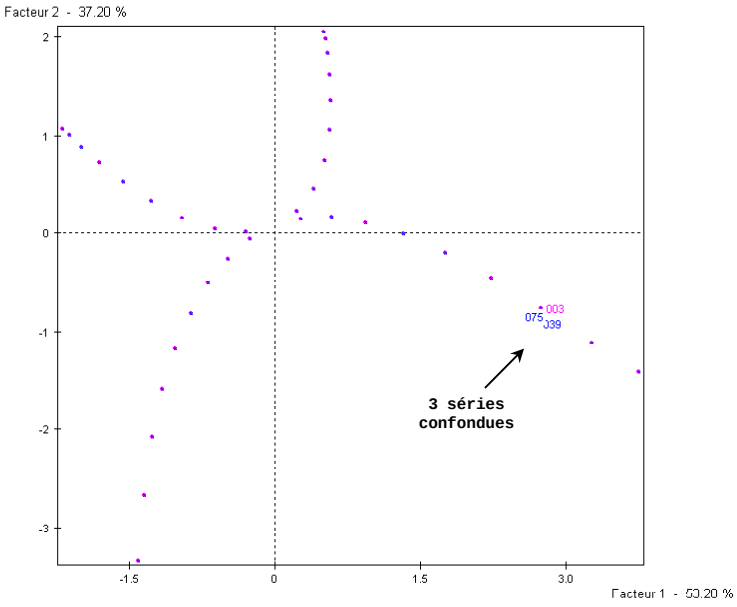


FIG. 3.25 – Modèle des scores sur trois échantillons
dont deux échantillons illustratifs (même valeur initiale de calage aléatoire)

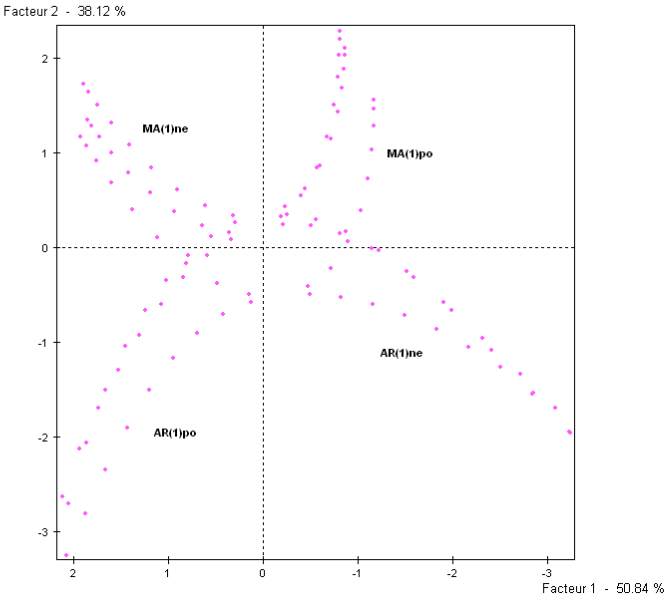


FIG. 3.26 – Modèle des scores sur les neuf échantillons

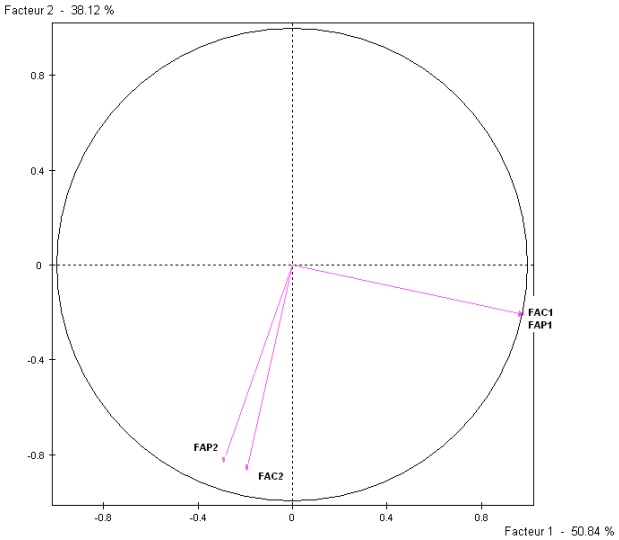


FIG . 3.27 – Cercle des corrélations sur les neuf échantillons

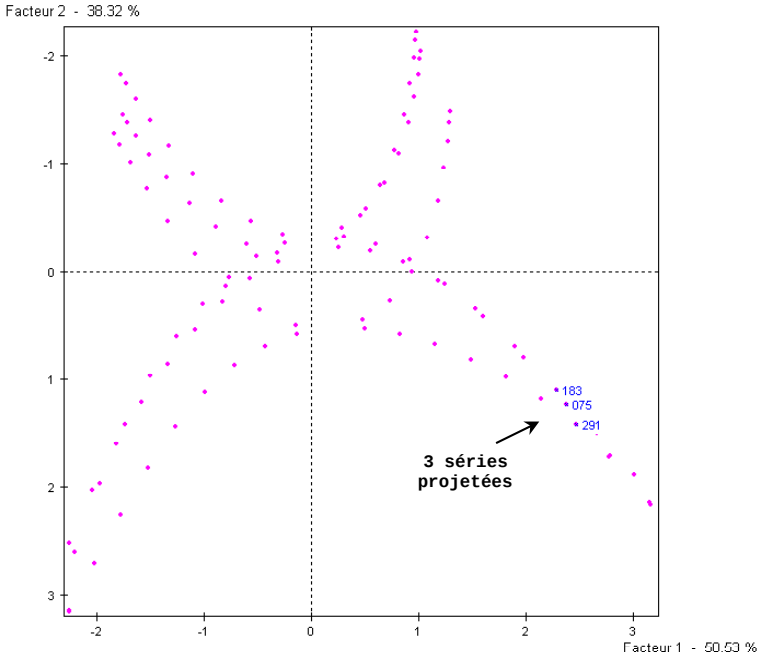


FIG. 3.28 – Modèle des scores sur les neuf échantillons dont trois individus illustratifs

3.5.3 Analyses graphiques

3.5.3.1 Le cas des modèles factoriels basés sur (FAC1,FAC2)

Des graphiques de bonne qualité et une propriété retrouvée :

Les représentations graphiques FIG. 3.21 à 3.23 sont de bonne qualité. En effet, les deux axes F1 et F2 expliquent à eux seuls 100% de l'information initiale ! Sur le cercle des corrélations, les trois variables sont représentées par deux vecteurs orthogonaux qui, par rotation orthogonale autour de l'origine, peuvent être confondus aux axes F1 et F2. Pour le démontrer, il suffit d'écrire les matrices analysées sur un ou neuf échantillons.

Démonstration :

Dans le cas d'un échantillon et en référence aux résultats des TABLEAU 3.1 et TABLEAU 3.2, la matrice théorique des éléments (FAC1,FAC2), notée $\underline{A}$, s'écrit :

$$\underline{A} = \begin{array}{c} \begin{matrix} \text{FAC1} & \text{FAC2} \end{matrix} \\ \begin{bmatrix} {}_{(1)}\phi_1 & {}_{(1)}\phi_1^{\,2} \\ {}_{(2)}\phi_1 & {}_{(2)}\phi_1^{\,2} \\ \vdots & \vdots \\ {}_{(18)}\phi_1 & {}_{(18)}\phi_1^{\,2} \\ -\dfrac{{}_{(19)}\theta_1}{1+{}_{(19)}\theta_1^{\,2}} & 0 \\ \vdots & \vdots \\ \vdots & \vdots \\ -\dfrac{{}_{(36)}\theta_1}{1+{}_{(36)}\theta_1^{\,2}} & 0 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \text{18 AR(1)} \\ \\ \\ \text{18 MA(1)} \\ \\ \\ \end{array}$$

où ${}_{(1)}\phi_1$, … ${}_{(18)}\phi_1$ sont les coefficients respectifs des processus AR(1) et ${}_{(19)}\theta_1$, … ${}_{(36)}\theta_1$, les coefficients respectifs des processus MA(1).

Les trajectoires simulées sont à coefficients symétriques :

$$\sum_{i=1}^{18} {}_{(i)}\phi_1^{\,k} = 0 \text{ et } \sum_{i=19}^{36} {}_{(i)}\theta_1^{\,k} = 0 \ (k \text{ impair}),$$

il vient que les deux vecteurs $\overrightarrow{FAC1}$ et $\overrightarrow{FAC2}$ sont orthogonaux.

Comme à l'intérieur de chaque échantillon, les trajectoires sont à coefficients symétriques, les vecteurs $\overrightarrow{FAC1}$ et $\overrightarrow{FAC2}$ restent orthogonaux pour l'ensemble des trajectoires simulées.

Fin démonstration.

Les rapprochements entre les axes avec les variables, et les séries est donc possible. Cependant, autant les variables sont très bien représentées dans le cercle des corrélations, autant il nous faut distinguer deux classes de processus :

- les processus correctement représentés dits « forts à semi-forts », dont les coefficients en valeur absolue sont compris entre 0.9 et 0.4 ;

- les processus moins bien représentés dits « faibles », dont les coefficients en valeur absolue sont compris entre 0.1 et 0.3.

A la lecture du graphique FIG 3.23, nous pouvons alors constater que :
- l'axe F1 avec la variable FAC1 oppose la classe des AR(1)ne et MA(1)po à la classe des AR(1)po et MA(1)ne, comme dans les ACP temporelles ;
- l'axe F2 avec la variable FAC2 oppose la classe des AR(1) 'forts à semi-forts' à la classe des AR(1) 'faibles' et des MA(1).

Enfin, avec l'exemple des séries AR(1) de coefficients –0.9, les scores pour les séries de même coefficient sont sur la FIG. 3.23 :
- *confondus* lorsque les séries sont issues de bruits blancs à même valeur de calage aléatoire ;
- *rapprochés* lorsque les séries sont issues de bruits blancs à valeur de calage aléatoire distincte, et d'autant plus éloignés que les processus sont des MA(1) ou des AR(1) faibles.

Premier résultat :

L'objectif qui était de rapprocher les séries en fonction de leurs coefficients est atteint pour uniquement les processus AR(1) forts à semi-forts. Il s'ensuit la recherche d'autres critères pour améliorer la représentativité des séries MA(1) et AR(1) faibles.

3.5.3.2 Le cas des modèles factoriels basés sur (FAC1,FAC2,FAP2)

Ajouter la variable FAP2 dans les ACP sur un, trois ou neuf échantillons améliore la discrimination entre la classe des AR(1) et la classe des MA(1) : nous pouvons l'observer sur les FIG. 3.24, 3.25 et 3.26.

Avec une variable supplémentaire, les axes (F1,F2) expliquent à eux seuls près de 90% de l'information initiale. Les proximités des variables du cercle des corrélations (FIG. 3.27), les proximités des variables entre elles et des variables avec les axes factoriels, nous permettent de caractériser :
- l'axe F1 avec les variables FAC1 ou FAP1 ;
- l'axe F2 avec les variables FAC2 et FAP2.

Il vient que, pour les séries les mieux représentées, c'est à dire les processus AR(1) et MA(1) forts à semi-forts :
- l'axe F1, avec les variables FAC1 ou FAP1, oppose les AR(1)ne et les MA(1)po aux AR(1)po et MA(1)ne ;
- l'axe F2, avec les variables FAC2 et FAP2, oppose la classe des AR(1) à la classe des MA(1).

Dans ce cas, les regroupements et oppositions des séries reflètent précisément le comportement *symétrique* des FAC et FAP que nous avions illustrés par des corrélogrammes dans les tableaux TABLEAU 3.1 et 3.2.

Notons que les scores pour les séries de même coefficient ne sont pas tous identiques :

- ils sont confondus lorsque les séries sont issues de bruits blancs de même valeur de calage aléatoire (par exemple FIG. 3.25) ;
- ils se rapprochent comme nous pouvons le constater sur la FIG. 3.26.

Enfin, les trois séries AR(1) de coefficient –0.7 projetées comme individus supplémentaires dans le modèle des scores en FIG. 3.28, se positionnent à l'identique de la FIG. 3.26. Nous retenons alors comme premier modèle factoriel de référence la FIG. 3.26 où il est possible de projeter une nouvelle série pour l'identification et une première estimation des paramètres d'un modèle AR(1) ou MA(1) fort à semi-fort qui l'engendrerait.

Nous avons utilisé les corrélations pour atténuer les bruits de la simulation : il n'est donc pas surprenant que l'identification des processus faibles soit délicate. Désormais, nous privilégions une approche plus qualitative qui n'utilise pas directement les moments d'ordre 2 : c'est l'analyse structurelle.

3.6 Identification structurelle

L'analyse structurelle que nous présentons a un objectif triple :
- décrire et mesurer d'éventuels changements structurels des séries temporelles ;
- établir une possible non-linéarité entre les nouveaux critères ;
- identifier des classes de processus dont la classe des processus faibles.

Pour cela, nous avons recours à deux types d'approches. La première approche utilise d'une part l'analyse technique d'oscillateurs empruntée à [Wilder78], et d'autre part la théorie de l'information que nous pouvons trouver dans [Shannon48]. Elle intègre les points de retournement tenant compte des études de [Kendall et al.76] et plus récemment de [Harding2003], et les mesures d'entropie de Shannon. La seconde approche utilise les entropies définies dans [Pincus92] appliquées par exemple dans [Richman2000] pour l'étude de séries temporelles en cardiologie.

3.6.1 Deux approches pour l'analyse structurelle

3.6.1.1 Les entropies de Shannon sur les séries d'états

Tout d'abord, chacune des trajectoires simulées est transformée par différence première des données en une série d'états ou de symboles (0) et (1). Un symbole (0) représente une baisse alors qu'un symbole (1) représente une hausse. Les séries d'états reflètent donc les points de retournement mais plus encore.

Puis, nous mesurons les fréquences des symboles et k-symboles (ou k-grammes) sur chaque série d'états pour estimer les différentes probabilités.

Cependant, comment qualifier l'information dont nous disposons sur les différentes probabilités ? Ou encore, comment qualifier l'incertitude ? C'est ici qu'intervient la théorie de l'information avec différentes entropies dont celles de Shannon, et celles conditionnelles et résiduelles provenant des travaux de [Yaglom59] et [Bavaud99].

Entropies simples de divers ordres :

Soit $P_1 = \{ p_{1,1}, p_{1,2}..., p_{1,m} \}$ les probabilités empiriques des m états ou modalités d'une série (Z_t). Comment qualifier l'incertitude ?

Une propriété souhaitable de toute mesure d'incertitude serait d'être minimale si et seulement si toutes les probabilités sont concentrées dans une seule modalité (cas déterministe : $p_{1,i} = 1,\ 1 \le i \le m$), et maximale si et seulement si toutes les

probabilités sont égales (équiprobabilité : $p_{1,1} = p_{1,2} = ... = p_{1,m} = \frac{1}{m}$).

Shannon a démontré que l'entropie $H_1(P_1)$, c'est à dire *l'entropie du premier ordre*, présente ces propriétés :

$$H_1(P_1) = -\sum_{i=1}^{m} p_{1,i} \log_2 p_{1,i}$$

Soient $P_2 = \{p_{2,1}, p_{2,2}..., p_{2,m^2}\}$ les probabilités empiriques des m^2 paires d'états d'une série $\{Z_t\}$, $P_3 = \{p_{3,1}, p_{3,2}..., p_{3,m^3}\}$ les probabilités empiriques des m^3 triplets d'états, etc.

Nous définissons de même les entropies simples d'ordre k ($k \geq 1$) par :

$$H_k(P_k) = -\sum_{i=1}^{m^k} p_{k,i} \log_2 p_{k,i}$$

Entropies conditionnelles de divers ordres :

Les entropies conditionnelles sont construites sur la base de probabilités conditionnelles qui elles-mêmes formalisent l'idée que la probabilité d'un événement puisse être conditionnée par l'information qu'un autre événement se soit réalisé.

Nous définissons ainsi des entropies conditionnelles d'ordre k notée h_k de la façon suivante :

$$h_k = H_1(P_1),\ k = 1$$

$$h_k = H_k(P_k / P_{k-1}) = H_k(P_k) - H_{k-1}(P_{k-1}),\ \text{sinon.}$$

Cette entropie h_k s'interprète comme l'incertitude moyenne sur le k ième symbole d'un k-gramme dont les $k-1$ premiers symboles sont connus.

Entropies résiduelles de divers ordres

En continuant dans cette voie, nous définissons l'entropie résiduelle d'ordre k, notée d_k par :

$$d_k = h_k - h_{k+1},\ k \geq 1$$

Cette entropie s'interprète comme la réduction moyenne d'incertitude sur un symbole selon que nous connaissons le k-gramme plutôt que le $(k-1)$-gramme.

3.6.1.2 Les entropies de Pincus sur les séries brutes

Pour la seconde approche, nous analysons directement sur les trajectoires simulées, les « irrégularités » ou « régularités » séquentielles. Pour cela, nous utilisons la décomposition suggérée par Pincus d'une série temporelle en vecteurs de même longueur l pour calculer des probabilités de vecteurs similaires sur la base d'un facteur « distance maximale ».

Soit $\left(z(1),\cdots,z(n)\right)$ la trajectoire simulée de longueur n. Nous construisons des vecteurs $v(i)$ de longueur l comme il suit :

$$\left(v(i),v(i+1),\cdots,v(i+l-1)\right)', \; i \le n-l+1$$

Nous comparons les vecteurs $v(i)$ entre eux en calculant la distance $D(i,j)$ définie comme la différence maximale entre les composantes des vecteurs $v(i)$ et $v(j)$.

Puis, pour i fixé, nous en déduisons le nombre $N^{l,r}(i)$ de vecteurs j $(j \le n-l+1)$ tel que $D(i,j) < r$ (r paramètre fixé comme le seuil de tolérance des comparaisons successives).

Nous définissons alors :

une probabilité

$$C^{l,r}(i) = \frac{N^{l,r}(i)}{n-l+1}$$

qu'un vecteur $v(j)$ soit proche d'un vecteur $v(i)$ dans un rayon de longueur r,

et la moyenne logarithmique des probabilités $C^{l,r}(i)$ $(i \le n-l+1)$ par :

$$F^{l,r} = \frac{\sum_{i=1}^{n-l+1} \log_2(C^{l,r}(i))}{n-l+1}$$

L'entropie de Pincus, notée 'ApEn', est finalement donnée par :

$$ApEn^{l,r} = F^{l,r} - F^{l+1,r}$$

'ApEn ' mesure la probabilité logarithmique que des vecteurs de longueur l proches entre eux restent proches de vecteurs de longueur $l+1$.

'ApEn' peut s'écrire aussi :

$$ApEn^{l,r} = \log_2\left[\frac{C^{l,r}}{C^{l+1,r}}\right]$$

avec $C^{l,r}$: la moyenne des $C^{l,r}(i)$.

Enfin, pour quantifier la réduction de l'incertitude, nous pouvons construire sur les séries brutes une 'ApEn' résiduelle, notée 'ApEnRes' comme il suit :

$$ApEn\,\mathrm{Re}\,s^{l,r} = ApEn^{l,r} - ApEn^{l+1,r},\ k \geq 1$$

3.6.1.3 Le projet intégré à Splus

Pour le calcul des différentes entropies, nous avons développé un projet en Fortran 90 que nous avons intégré à S-Plus (fourni en annexe). Il comprend essentiellement deux modules pour d'une part les traitements des séries brutes, et d'autre part les traitements des séries d'états obtenues après transformation des séries brutes.

Le module de traitements des séries brutes :

- il estime les entropies 'ApEn' d'ordre k, k allant de 1 à l ;
- il estime les entropies 'ApEnRes' d'ordre k allant de 1 à $l-1$.

Le module de traitements des séries d'états :

- il estime les probabilités des symboles (0) ou (1) et séquences de k-symboles (pour k quelconque)
- il estime les probabilités de transition d'ordre $l_t - 1$ pour caractériser les relations entre les k-symboles consécutifs, avec par exemple :

$$p((0,0) \to (0)) = p(0,0,0) / p(0,0)) ;$$

- il estime les entropies simples d'ordre k, k allant de 1 à l ;
- il estime les entropies conditionnelles d'ordre k, k allant de 1 à l ;
- il estime les entropies résiduelles d'ordre k, k allant de 1 à $l-1$.

Deux applications :

Tout d'abord, le projet est appliqué dans sa totalité sur le premier échantillon des trajectoires AR(1) et MA(1) simulées. Nous avons choisi comme valeurs de paramètres : $m = 2$ (ce sont les états ou modalités (0) et (1)), $l = 10$, $l_t = 3$ et $r = 1$. Nous obtenons une matrice d'entropies de dimension (36x56), dans laquelle les 56 variables sont :

- les entropies simples d'ordre 1 à 10, notées 'Shan1 à Shan10' ;
- les entropies conditionnelles d'ordre 1 à 10, notées 'Cond1 à Cond10' ;
- les entropies résiduelles d'ordre 1 à 9, notées 'Res1 à Res9' ;

- les probabilités de transition d'ordre 3, notées 'Mtrans1 à 8' ;
- les entropies simples de Pincus d'ordre 1 à 10, notées 'ApEn1 à 10' ;
- les entropies résiduelles de Pincus d'ordre 1 à 9, notées 'ApEnRes1 à 10',

et où les 36 individus correspondent aux 36 trajectoires du premier échantillon.

Les critères sont alors très nombreux. Nous construisons une première ACP qui nous permet de constater que les entropies de Shannon ou les entropies de Pincus suffisent à discriminer les séries. Nous retenons comme variables significatives les entropies simples, conditionnelles et résiduelles de Shannon.

Le nombre de variables reste encore important. Pour faciliter l'interprétation de regroupements et oppositions de séries avec les variables, pour mettre en évidence une éventuelle non-linéarité entre les critères et pour accroître la discrimination entre les séries, nous construisons en référence aux travaux [Benzecri73a] et [Benzecri73b] une analyse des correspondances multiples (ou ACM) sur les entropies. Nous souhaitons aussi un graphique de projection des variables et des séries auquel s'ajoutent des classes de processus. Nous choisissons alors comme méthode factorielle une classification sur les composantes principales d'une ACM « structurelle ».

Pour cela, nous appliquons le projet sur le premier échantillon avec pour valeurs de paramètres : $m = 2$, $l_t = 0$ et $l = 5$. Seul le module de traitement des séries d'états est pris en compte et nous obtenons une matrice d'entropies de dimension (36x14). Les 14 variables de la matrice sont :

- les entropies simples d'ordre 1 à 5, notées 'Shan1 à Shan5' ;
- les entropies conditionnelles d'ordre 1 à 5, notées 'Cond1 à Cond5' ;
- les entropies résiduelles d'ordre 1 à 4, notées 'Res1 à Res4' ;

et les 36 individus correspondent aux 36 trajectoires du premier échantillon.

Une ACM à trois modalités suivie d'une classification donne lieu au modèle factoriel en FIG. 3.29.

3.6.2 ACM structurelle suivie d'une classification - Trois classes de processus AR(1) et MA(1)

La lecture graphique (FIG. 3.29) est en effet facilitée par la méthode qui utilise d'une part l'ACM, et d'autre part la classification.

Tout d'abord, c'est dans le plan factoriel (F1,F2) que le modèle de représentation des séries, des variables, et des classes est de meilleure qualité. Il explique à lui seul près de 65% de l'information initiale, et les séries comme les variables y sont assez bien représentées (sommes de carrés de cosinus souvent proches de 0.70).

Nous pouvons aussi remarquer une non-linéarité entre les divers critères et traduire une éventuelle progression en reliant leurs modalités respectives. Nous

constatons alors une certaine cohérence des données avec la présence de nombreuses lignes polygonales régulières qui suivent trois classes de séries. La méthode et le choix du nombre des modalités semblent pertinents.

Il en résulte les rapprochements des variables et des classes de séries dont les contributions à la formation des axes sont souvent voisines de 0.75.

L'axe F1, avec surtout les entropies conditionnelles d'ordre 2 à 5 (modalités 1 et 3) et l'entropie résiduelle d'ordre 1 (modalités 3 et 1) sépare la classe des AR(1) à coefficients négatifs et des MA(1) à coefficients positifs, de la classe des MA(1) à coefficients négatifs et des AR(1) à coefficients positifs.

L'axe F2 oppose les processus forts à semi-forts du côté positif, aux processus faibles du côté négatif, avec surtout des entropies conditionnelles d'ordre 2 à 5 (modalité 2) et une entropie résiduelle d'ordre 1 (modalité 2).

Précisons que les k-grammes utilisés dans le calcul des entropies conditionnelles d'ordre 2 (modalités 1 à 3) ou des entropies résiduelles d'ordre 1 (modalités 1 à 3) sont (0,1) et (1,0), c'est à dire les *« signatures » des points de retournement*.

La Fig. 3.29 nous permet donc de caractériser aussi bien les deux classes de processus forts à semi-forts que la classe des processus faibles, par des mesures de réduction d'incertitude comme les modalités de l'entropie résiduelle d'ordre 1. Nous retenons un second modèle factoriel de référence pour les processus AR(1) et MA(1).

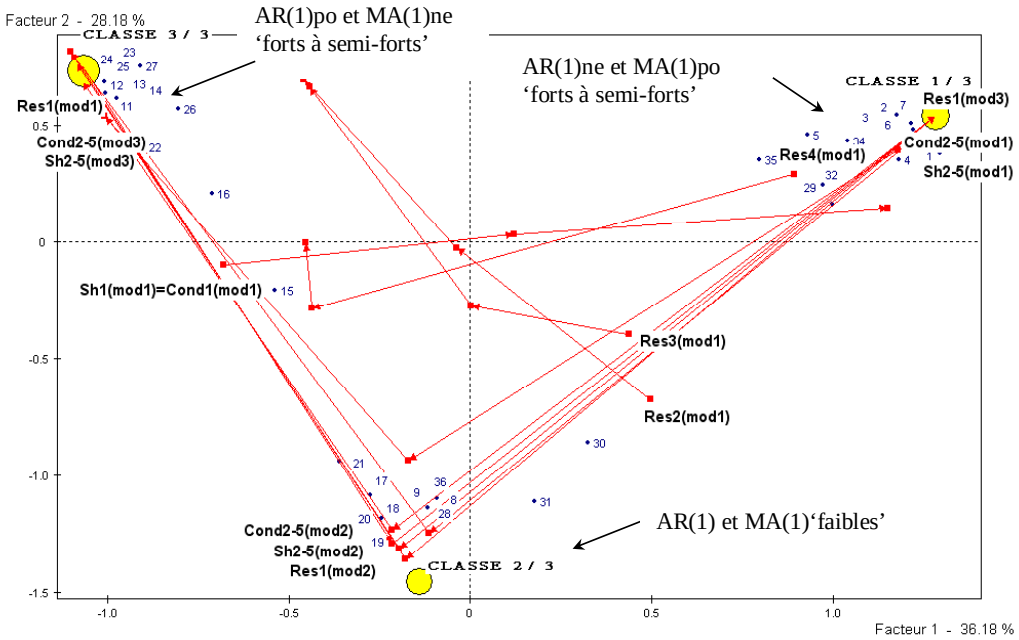


FIG. 3.29 – Modèle factoriel « structurel » sur un échantillon

Conclusion

Notre objectif qui était de construire des premiers modèles factoriels de processus ARMA pour l'*identification et « l'estimation »* d'une série temporelle est atteint pour les processus AR(1) et MA(1).

Deux modèles factoriels (FIG. 3.26 et FIG. 3.29) suffisent pour caractériser aussi bien les processus AR(1) ou MA(1) à forts coefficients que les processus AR(1) ou MA(1) à faibles coefficients. Le premier modèle (FIG. 3.26) est seulement basé sur trois variables : les deux premiers retards de la FAC et le second retard de la FAP. Le second modèle (FIG. 3.29) est lui basé sur des entropies de Shannon qui possède la signature des points de retournement.

Nous pouvons alors projeter une nouvelle série dans les deux modèles factoriels, pour une identification et une première estimation des paramètres d'un modèle AR(1) ou MA(1), forts à semi-forts ou faibles qui l'engendrerait.

Pour rendre possible la projection d'une série temporelle dans le modèle en FIG. 3.26, il suffit de calculer sur la chronique les deux premiers retards de la FAC et le deuxième retard de la FAP. Pour la projection d'une série temporelle dans le modèle factoriel en FIG. 3.29, il suffit de calculer les entropies de Shannon jusqu'à l'ordre 5 et en déduire leurs trois modalités respectives.

Enfin, l'objectif atteint de ce chapitre est double. Au delà des deux premiers modèles factoriels de référence, nous pouvons en déduire une méthode d'identification de modèles de séries temporelles qui combine à la fois :

- l'approche par les autocorrélations,
- et l'approche par les entropies,

avec une visualisation par les méthodes factorielles comme l'ACP et l'ACM suivie d'une classification.

Plusieurs directions sont alors possibles. Dans le chapitre qui suit, nous avons choisi d'appliquer la méthode à des processus AR(2) et MA(2) stationnaires.

Chapitre 4

4 Application de la méthode aux AR(2) et MA(2)

Introduction

Dans le chapitre précédent, nous avons construit deux modèles factoriels de projection d'une série temporelle, pour l'identification et une première estimation des paramètres d'un modèle AR(1) ou MA(1) qui l'engendrerait.

Désormais, nous voulons étendre la méthode qui utilise l'analyse des corrélations et l'analyse structurelle avec une visualisation par des méthodes factorielles, aux séries AR(2) et MA(2) stationnaires.

Pour cela, nous procédons en trois étapes :

- la simulation de trajectoires AR(2) et MA(2), « indépendantes » et stationnaires, et à coefficients symétriques, suivie de deux ACP temporelles ;
- l'analyse des corrélations avec la construction de modèles factoriels issus d'ACP basées sur des éléments de la FAC et de la FAP calculés directement sur la matrice de simulation ;
- l'analyse structurelle avec une ACM suivie d'une classification sur des mesures d'entropie.

Tout d'abord, les caractéristiques de la simulation sont précisées dans la première partie de ce chapitre, et le choix des coefficients tient compte des résultats des chapitres précédents. Puis, deux modèles des scores issus d'ACP temporelles sont construits, et regroupent ou opposent les trajectoires AR(2) et MA(2) provenant d'un même bruit blanc. Cependant, les regroupements des séries en fonction des coefficients ne sont pas satisfaisants. Il s'ensuit l'analyse des corrélations et l'analyse structurelle dans les deux dernières parties de ce chapitre. Nous retiendrons finalement deux modèles factoriels de référence pour les AR(2) et les MA(2) stationnaires.

4.1 Analyse temporelle sur les AR(2) et les MA(2)

4.1.1 La simulation de trajectoires AR(2) et MA(2)

La simulation des trajectoires AR(2) et MA(2) tient compte tout d'abord des conditions générales de simulation liées à la nature même des processus ARMA et énoncées en section 3.2 du chapitre 3.

Les trajectoires AR(2) doivent aussi être stationnaires. Nous rappelons ces conditions.

Soit le processus AR(2), $Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t$ où $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2$. (Z_t) est stationnaire si et seulement si :

$$\phi_2 + \phi_1 < 1$$
$$\phi_2 - \phi_1 < 1$$
$$-1 < \phi_2 < 1$$

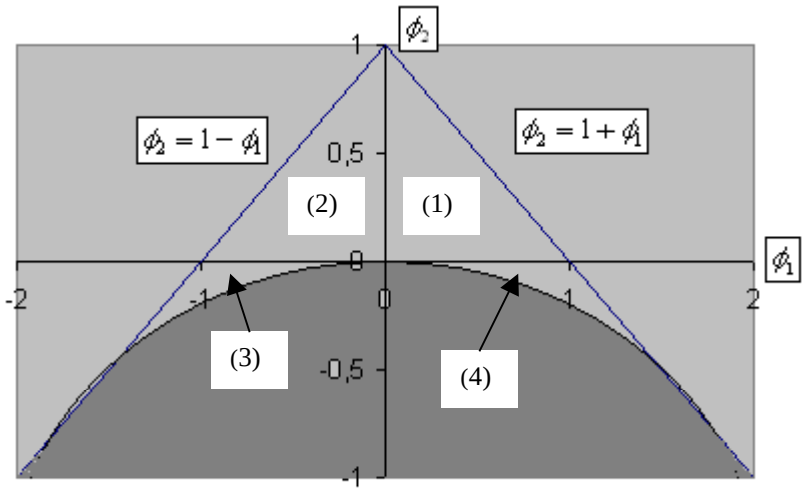


FIG. 4.1 – Les quatre zones de simulation

La figure FIG. 4.1 illustre la résolution graphique du système d'inéquations. La solution géométrique est donc l'intérieur du triangle (ABC) qui représente l'ensemble des couples (ϕ_1, ϕ_2) pour lesquels un processus AR(2) est stationnaire.

Il en découle quatre zones de simulation, c'est à dire les quatre zones de définition : (1) ou zone 1, (2) ou zone 2, (3) ou zone 3 et (4) ou zone 4 pour les coefficients (ϕ_1, ϕ_2) réels. La zone noircie correspond aux solutions complexes de l'équation : $1-\phi_1 B-\phi_2 B^2=0$.

Dans chaque zone de simulation, nous avons engendré six trajectoires issues d'un même bruit blanc (même valeur de calage aléatoire et même variance), et une ou deux trajectoires supplémentaires issues de bruits blancs à valeurs de calage aléatoire et variances distinctes. Des résultats obtenus du chapitre 3, nous rappelons qu'il est inutile de multiplier le nombre de trajectoires à simuler pour des valeurs de calage aléatoire et variances distinctes. A titre de vérification, seules quelques unes suffisent pour les projeter comme individus supplémentaires dans les modèles factoriels de référence.

Précisément, pour les AR(2) ou les MA(2), les couples (ϕ_1, ϕ_2) ou (θ_1, θ_2) sont :

- pour la zone 1 de simulation : (0.7,0.1), (0.6,0.2), (0.5,0.3), (0.4,0.4), (0.3,0.2), (0.2,0.1) et les 2 supplémentaires (0.5,0.3), (0.2,0.1) ;
- pour la zone 2 de simulation : (-0.7,0.1), (-0.6,0.2), (-0.5,0.3), (-0.4,0.4), (-0.3,0.2), (-0.2,0.1) et les 2 supplémentaires (-0.5,0.3), (-0.2,0.1) ;
- pour la zone 3 de simulation : (-1.8,-0.9), (-1.7,-0.8), (-1.6,-0.7), (-1.5,-0.6), (-1.4,-0.5), (-1.3,-0.4) et 1 supplémentaire (-1.6,-0.7) ;
- pour la zone 4 de simulation : (1.8,-0.9), (1.7,-0.8), (1.6,-0.7), (1.5,-0.6), (1.4,-0.5), (1.3,-0.4) et 1 supplémentaire (1.6,-0.7),

avec dans l'ordre des zones :

- pour les AR(2), des trajectoires numérotées de 1 à 24 sans les supplémentaires, et de 25 à 30 pour les supplémentaires ;
- pour les MA(2), des trajectoires numérotées de 31 à 54 sans les supplémentaires, et de 55 à 60 pour les supplémentaires.

La matrice « temporelle » obtenue à l'issue de la simulation est de dimension (60×500) et est la réalisation d'un modèle vectoriel stationnaire constitué de processus « indépendants ».

4.1.2 Les premiers modèles factoriels pour les AR(2) et MA(2)

Dans le but de faire apparaître des regroupements et oppositions significatifs des séries et de même que pour les AR(1) et MA(1) dans le chapitre 3, nous construisons :

- une première ACP temporelle sur les trajectoires AR(2) et MA(2) issues d'un même bruit blanc, c'est à dire celles numérotées de 1 à 24 et de 31 à 54,
- et une seconde ACP temporelle sur l'ensemble des trajectoires avec en éléments supplémentaires les trajectoires issues des autres bruits blancs, c'est à dire les trajectoires numérotées de 25 à 30 et de 55 à 60.

La première ACP donne lieu à la figure FIG 4.2, et la seconde ACP à la figure FIG. 4.3.

Pour la FIG 4.2 dans laquelle les trajectoires sont toutes issues d'un même bruit blanc, les regroupements et oppositions des séries sont « remarquables ».

Un axe de symétrie vient séparer :
- les AR(2) des zones 1 et 4 et les MA(2) des zones 2 et 3,
- des AR(2) des zones 2 et 3 et des MA(2) des zones 1 et 4.

Par ailleurs, l'axe F2 oppose :
- la classe des AR(2) zones 3 et 4,
- à la classe des MA(2) et des AR(2) zones 1 et 2.

Cependant, le modèle des scores n'est pas satisfaisant : il ne fournit pas un modèle de référence, c'est à dire un modèle de projection d'une nouvelle série pour son identification et une première estimation de ses paramètres. Pour le constater, il suffit de projeter les éléments supplémentaires ou les trajectoires 25 à 30, et 55 à 60 : c'est la FIG. 4.3. Les séries ainsi projetées ne se positionnent pas dans les zones prédéfinies par les trajectoires issues d'un même bruit blanc. Ces résultats nous rappellent ceux obtenus après une ACP temporelle sur les AR(1) et les MA(1).

Il s'ensuit logiquement l'approche par les corrélations qui a pour but :
- d'atténuer l'influence des bruits,
- et de construire un modèle dans lequel les regroupements se font quelques soient les échantillons, en fonction des coefficients ou zones de la simulation.

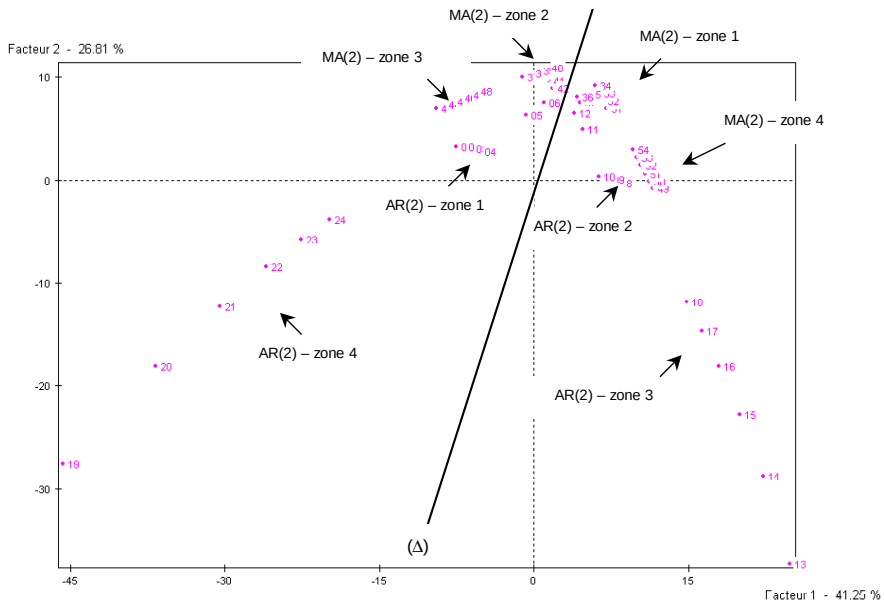


FIG. 4.2 – Modèle des scores temporel pour les AR(2) et les MA(2)

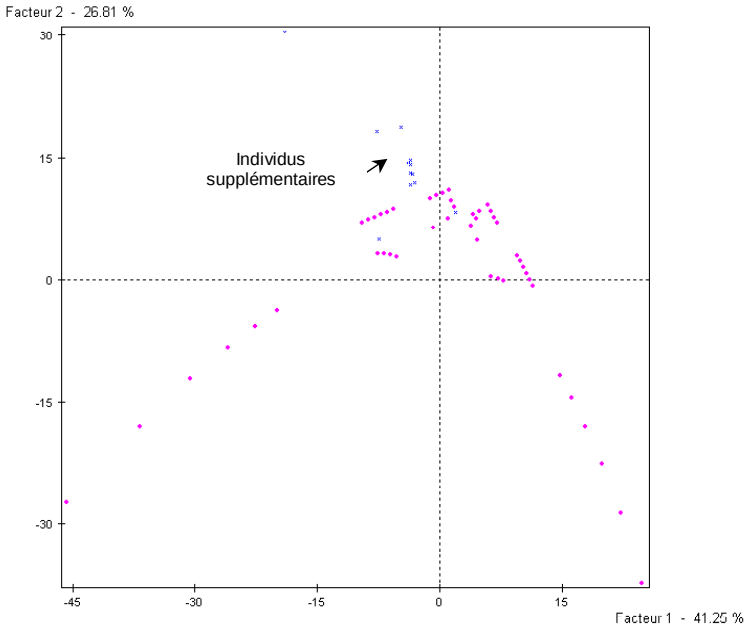


FIG. 4.3 – Modèle des scores temporel pour les AR(2) et les MA(2) avec les individus supplémentaires

4.2 Analyse des corrélations

Après plusieurs ACP construites sur des éléments de la FAC et de la FAP, nous retenons finalement les trois éléments significatifs suivants : FAC1, FAC2 et FAP2. Pour le choix des variables, il en va de même que pour les processus AR(1) et MA(1).

Dans le plan factoriel (F1,F2), deux modèles graphiques sont ici représentés. Ce sont le cercle des corrélations en figure FIG. 4.4 et le modèle des scores en figure FIG. 4.5.

Tout d'abord, les représentations sont de bonne qualité. En effet, 75% de l'information initiale est expliquée par les deux axes F1 et F2.

Sur la FIG. 4.4, l'axe F1 est confondu avec la variable FAC1, et les variables FAC2 et FAP2 sont très proches de l'axe F2. Par ailleurs, les AR(2) comme les MA(2) sont assez bien représentés dans le modèle des scores en FIG. 4.5.

Nous pouvons alors expliquer les regroupements et oppositions des séries par les axes ou par les trois variables FAC1, FAC2 et FAP2.

L'axe F2 avec les variables FAC2 et FAP2 oppose nettement et à la différence des modèles en FIG. 4.2 et FIG. 4.3 :

- la classe des AR(2),
- à celle des MA(2).

L'axe F1 avec la variable FAC1 (ou FAP1) oppose :

- les AR(2) des zones 1 et 4 et les MA(2) des zones 2 et 3,
- aux AR(2) des zones 2 et 3 et les MA(2) des zones 1 et 4.

➢ ***Notons que cette unique représentation factorielle des séries AR(2) et MA(2) résume à elle seule les propriétés des fonctions d'autocorrélation de ces processus que nous avions illustrées par l'intermédiaire de plusieurs corrélogrammes dans les TABLEAU 3.1 et TABLEAU 3.2 du chapitre 3.***

Enfin et à la différence de la FIG. 4.3, les individus supplémentaires pour des bruits blancs à valeurs de calage aléatoire distinctes, se positionnent précisément dans les zones prédéfinies par les trajectoires issues d'un même bruit blanc.

Le modèle des scores en FIG 4.5 fait donc l'objet d'un modèle factoriel de projection d'une nouvelle série pour l'identification et l'estimation d'un modèle AR(2) ou MA(2) qui l'engendrerait.

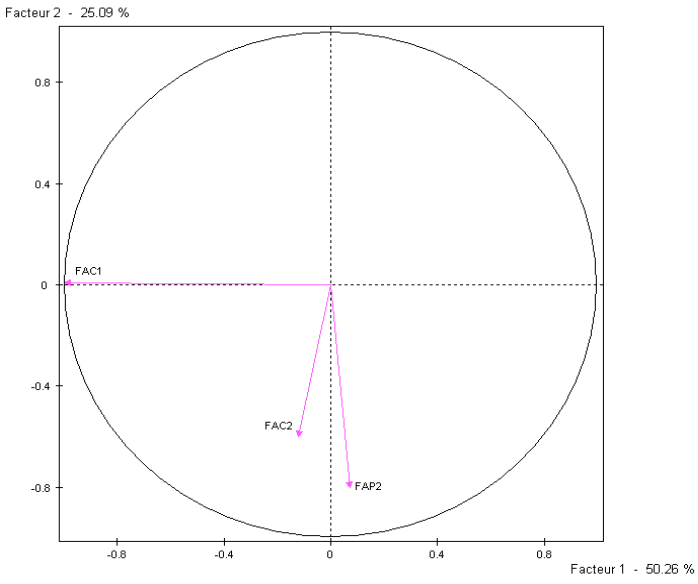


FIG. 4.4 – Cercle des corrélations

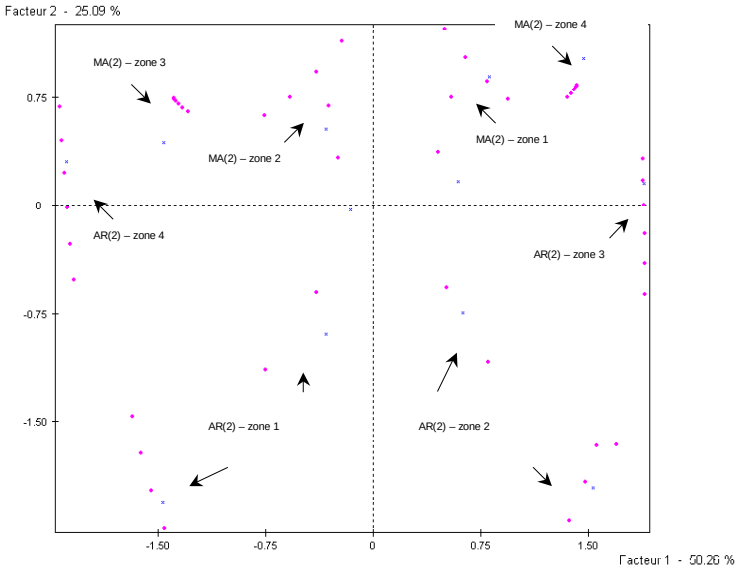


FIG. 4.5 – Modèle des scores avec les individus supplémentaires

4.3 Analyse structurelle

Le modèle obtenu en FIG. 4.5 est de bonne qualité. Cependant, est-il possible de discriminer aussi bien sinon mieux les trajectoires AR(2) et MA(2) simulées, en nous aidant d'une part de l'analyse structurelle, et d'autre part de l'ACM suivie d'une classification ? C'est l'objectif de ce travail.

De nombreuses analyses structurelles ont été effectuées sur la base des modules de traitement des séries d'états ou des séries brutes définis dans le chapitre 3. De même que pour les AR(1) et MA(1), nous retenons les entropies de Shannon de divers ordres comme variables, et une ACM à trois modalités suivie d'une classification comme technique de visualisation graphique. Le modèle obtenu est représenté en figure FIG. 4.6 et présente trois classes distinctes des séries AR(2) et MA(2) simulées.

C'est à nouveau dans le premier plan factoriel (F1,F2) que la représentation est de meilleure qualité avec près de 70% de l'information initiale expliquée.

Les deux axes factoriels sont très bien représentés par les variables Res1 et Sh2, c'est à dire l'entropie résiduelle d'ordre 1 et l'entropie simple d'ordre 2. Plus précisément :

- l'axe F1 dans sa partie négative est caractérisé par les modalités 2 des variables Res1 et Sh2 ;
- l'axe F2 dans sa partie positive est caractérisé par la modalité 1 de la variable Sh2 et par la modalité 3 de la variable Res1 ;

- l'axe F2 dans sa partie négative est caractérisé par la modalité 3 de la variable Sh2 et par la modalité 1 de la variable Res1.

Par ailleurs, trois groupes ou trois classes de séries se distinguent entre elles. Après identification des séries, il s'ensuit les rapprochements suivants avec les variables :

- la classe des MA(2) pour les zones 2 et 3 et des AR(2) pour la zone 1 est caractérisée par les variables Res1(mod1) et Sh2(mod3) ;
- la classe des MA(2) pour les zones 1 et 4 est caractérisée par les variables Res1(mod2) et Sh2(mod2) ;
- la classe des AR(2) pour les zones 2, 3 et 4 est caractérisée par les variables Res1(mod3) et Sh2(mod1).

Enfin, la projection des douze trajectoires issues de bruits blancs distincts vient conforter la validité du modèle. En effet, chacune de ces trajectoires se positionne dans le groupe qui lui correspond.

Les regroupements et oppositions des séries diffèrent de ceux obtenus par l'analyse des corrélations mais la discrimination des séries est effective et la projection d'une nouvelle série est aussi possible. Nous pouvons alors retenir un second modèle factoriel de référence pour les AR(2) et les MA(2) stationnaires.

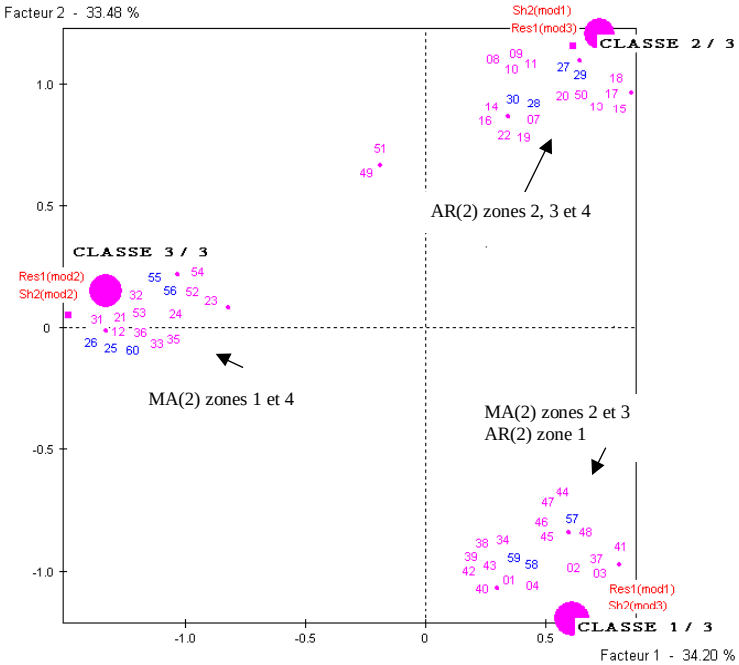


FIG. 4.6 – Modèle factoriel structurel sur les AR(2) et MA(2)

Conclusion

L'objet de la thèse consiste en l'étude des composantes principales de plusieurs séries temporelles dans le domaine du temps discret, ce qui représente le cas le plus souvent rencontré en pratique. Des propriétés des composantes principales (scores et modèles dynamiques) ont été établies, tout en respectant la *chronologie des données*, la *double dimension espace et temps*, et la *stationnarité* des processus. Le cas unidimensionnel a été précisé et généralisé au cas multidimensionnel. Puis, quelques uns des résultats obtenus ont été exploités, en se plaçant dans le contexte des processus « indépendants », pour construire des premiers modèles factoriels de référence.

Les avantages, les limites et les perspectives de ce travail de recherche sont les suivants.

Avantages

L'objectif atteint est double, sans compter l'originalité de la méthode par son aspect *graphique* qui peut être exploité en automatisation de la prévision.

Dans un premier temps, est proposé une *méthode simultanée d'identification et de première estimation* des paramètres d'un modèle AR(1) ou MA(1), AR(2) ou MA(2), qui engendre une série projetée dans un des modèles factoriels de référence.

Dans un second temps et compte tenu des résultats obtenus sur les modèles ARMA les plus souvent rencontrés, une méthode d'analyse des séries temporelles peut en être déduite. Cette méthode d'identification de modèles factoriels de séries temporelles combine à la fois l'approche par les autocorrélations et l'approche par les entropies, avec des techniques de visualisation graphique telles que l'ACP et l'ACM suivie d'une classification.

Limites

Le cadre fondamental de l'étude est celui des processus « *indépendants »* ou non-corrélés et *stationnaires* en covariance.

Dans le cas des processus non « indépendants », les pistes qui suivent donnent quelques perspectives.

Dans le cas des processus non stationnaires, pour utiliser ces méthodes, il est

entendu qu'il suffit d'une stationnarité au sens faible qui concerne plus particulièrement les processus linéaires. Heureusement, beaucoup d'entre eux peuvent être rendu stationnaires par des transformations classiques telles que la différenciation, la désaisonnalisation et la transformation logarithmique et plus généralement puissance [Box et al.64], pour obtenir la stabilité dans le temps de la moyenne et de la variance.

De même, dans le cas des processus non linéaires, quelques perspectives peuvent être dégagées.

Perspectives pratiques et théoriques

D'un point de vue pratique :

- la méthode avec ses deux approches par les autocorrélations et par les mesures d'entropie s'applique à des données opérationnelles ;

- quand l'analyse des corrélations n'est plus efficace, l'approche par les entropies qui ne nécessite pas de contraintes des moments d'ordre 2 (ou de stationnarité) pourrait prendre le relais et ouvrir ainsi une voie vers le non-linéaire, comme le constate le chapitre 3 de ce travail avec les processus dits « faibles » ;

- l'approche structurelle peut parfaitement s'appliquer à de nombreux cas pratiques tels que les données financières.

D'un point de vue purement théorique, l'étude des composantes principales peut être poursuivie dans le cas des processus non indépendants :

- en travaillant à nouveau sur la matrice de covariance à structure bloc-Toeplitz, sur ses éléments propres ou leur approximation,

- en considérant tout d'abord le cas bi-dimensionnel pour lequel la dépendance entre deux composantes est unidirectionnelle, puis bi-directionnelle ;

- pour mieux appréhender le cas général.

Bibliographie

[Aitken54] AITKEN (A. C.). – *Determinants and Matrices*. – London, Oliver and Boyd, 1954.

[Akaike77] AKAIKE (H.). – *On entropy maximisation principle. Applications of Statistics*. – Edition Krisnaiah, 27–41, North–Holland, 1977.

[Anderson76] ANDERSON (O. D.). – *Time series analysis and forecasting. The Box–Jenkins approach*. Butterworths, London and Boston, 1976.

[Bartlett46] BARTLETT (M. S.). – On the theoretical specification of sampling properties of autocorrelated time series. – Journal Royal Stat. Soc., B8, 27, 1946.

[Basilevsky et al.79] BASILEVSKY (A.), HUM (D. P. J.). – *Karhunen–Loeve analysis of historical time series with an application to plantation births in Jamaïca*. – Journal of American Statistical Association, 74, 366, 284–290, 1979.

[Bavaud99] BAVAUD (F.). – *Modèles et Données*. – L'Harmattan, Paris, 1999.

[Becker et al.88] BECKER (R. A.), CHAMBERS (J. M.), WILKS (A. R.). – *The New S Language : A programming Environment for Data Analysis and Graphics*. – Chapman and Hall, New York, 1988.

[Beghin et al.80] BEGHIN (J.M.), GOURIEROUX (C.), MONFORT (A.). – *Identification of an ARIMA process : the corner method*. – Time Series, ed. T. Anderson, North Holland, 1980.

[Benzecri73a] BENZECRI (J.P.) – *L'analyse des données, Tome 1 : La taxinomie*. – Dunod, 1973.

[Benzecri73b] BENZECRI (J.P.) – L'analyse des données, Tome 2 : L'analyse des correspondances. – Dunod, 1973.

[Bourbonnais et al.98] BOURBONNAIS (R.), TERRAZA (M.). – *Analyse des séries temporelles en économie*. – Presses universitaires de France, 1998.

[Bourdeau93] BOURDEAU (M.), TUBACH (J. P.). – *Structures phonologiques : une étude de cinq corpus de langue*. – Rapport technique, Télécom Paris, D025, 1993.

[Box et al.64] BOX (G. E. P.), COX (D. R.). – *An Analysis of Transformations.* – Journal of the Royal Statistical Society, 26, 1964.

[Box et al.70] BOX (G. E. P.), JENKINS (G. M.). – *Time Series Analysis, Forecasting and Control*. – Holden–Day, San Francisco, 1970.

[Box et al.76] BOX (G. E. P.), JENKINS (G. M.). – *Time Series Analysis, Forecasting and Control.* – Holden–Day, Second Edition, San Francisco, 1976.

[Brillinger75] BRILLINGER (D. R.). – *Time Series Data Analysis and Theory*. – Holden Day, San Francisco, 1975.

[Brillinger81] BRILLINGER (D. R.). – *Time Series Data Analysis and Theory*. – Holden Day, San Francisco, 1981.

[Brockwell et al.93] BROCKWELL (P. J.), DAVIS (R. A*.). – Time Series : Theory and Methods*. – Springer–Verlag, 2ième édition, 1993.

[Broomhead et al.86a] BROOMHEAD (D. S.), KING (G. P.). – *Extracting qualitative dynamics from experimental data.* – Physica. D. 20, 217–236, 1986.

[Broomhead et al.86b] BROOMHEAD (D. S.), KING (G. P.). – *On the qualitative analysis of experimental dynamical systems. Nonlinear Phenomena and Chaos*. – S. Sarkar, Ed., Adam Hilger, 113–144, 1986.

[Burtschy87] BURTSCHY (B.). – *Factorial Analysis of Multiple Time Series*. – Proceedings of the Business and economic section, American Statistical Association, 310–314, 1987.

[Deville74] DEVILLE (J.C.). – *Méthodes statistiques et numériques de l'analyse harmonique.* – Annales de l'INSEE, 15, jan-avril 1974.

[Elsner et al.96] ELSNER (J. B.), TSONIS (A. A.). – *Singular Spectrum Analysis. A New Tool in Time Series Analysis*. – Plenum Press, New York and London, 1996.

[Ferrara et al.2002] FERRARA (L.), GUEGAN (D.). – *Analyser les séries chronologiques avec S–Plus : une approche paramétrique*. – Presses universitaires de Rennes, 2002.

[Fraedrich86] FRAEDRICH (K.). – *Estimating the dimensions of weather and climate attractors*. – J. Atmos. Sci. 43, 419–432, 1986.

[Friedman51] FRIEDMAN (B.). – *Eigenvalues of compound matrices*. – Research Rept. No. TW–16, Math. Res. Group. New York U., New York, 1951.

[Friedman61] FRIEDMAN (B.). – *Eigenvalues of composite matrices*. – Proc. Camb. Philos. Soc. 57:37–49, 1961.

[Ghil et al.91a] GHIL (M.), MO (K. C.). – Intraseasonal oscillations in the global atmosphere, Part. : Northern Hemisphere and tropics. – J. Atmos. Sci., 48, 752–779, 1991.

[Ghil et al.91b] GHIL (M.), VAUTARD (R.). – Interdecadal oscillations and the warming trend in global temperature time series. – Nature, 350, 324–327, 1991.

[Golyandina et al.2001] GOLYANDINA (N.), NEKRUTKIN (V.*). – Analysis of Time Series Structure. SSA and Related Techniques*. – Boca Raton : Chapman and Hall, 2001.

[Gouriéroux et al.83] GOURIEROUX (C.), MONFORT (A.). – *Cours de séries temporelles*. – Economica, 1983.

[Gouriéroux et al.95] GOURIEROUX (C.), MONFORT (A.). – *Séries Temporelles et Modèles Dynamiques*. – Economica, 2ième édition,1995.

[Granger80] GRANGER (C. W. J.). – *Forecasting in business and economics*. – New York, Academic Press, 1980.

[Granger et al.86] GRANGER (C. W. J.), NEWBOLD (P.). – *Forecasting economic time series*. – Academic Press, 1986.

[Gray72] GRAY (R. M.). – *On the asymptotic eigenvalue distribution of Toeplitz matrices*. – IEEE Transf. Inform. Theory, Vol. IT–18, 725–730, 1972.

[Gray2000] GRAY (R. M.). – Toeplitz and circulant matrices : A review. Information Theory Laboratory. – Standford Univ., Stanford, CA., 2000.

[Grenander et al.84] GRENANDER (U.), SZEGO (G.). – *Toeplitz Forms and their Applications.* – Chelsea Publishing Company, New York, 1984.

[Hamilton94] HAMILTON (J.D.). – *Time Series Analysis*. – Princeton Univ. Press, 1994.

[Hannan et al.85] HANNAN (E.), KRISHNAIAH (P. R.), RAO (M. M.). – *Time series in the time domain*. – North Holland, 1985.

[Hamburger et al.51] HAMBURGER (H.), GRIMSHAW (M. E.). – *Linear Transformations in n–dimensional Vector Space*. – Cambridge : Cambridge Univ. Press, 1951.

[Harding2003] HARDING (D.). – *Using turning point information to study economic dynamics*. – The University of Melbourne, 2003.

[Harvey81] HARVEY (A. C.). – *Time Series Models.* – Wiley, New York, 1981.

[Hasselmann88] HASSELMANN (K.). – PIPs and POPs : The reduction of complex dynamical systems using principal interaction and oscillation patterns. – J. Geophys. Res., 93, 11,015–11,021, 1988.

[Hotelling33] HOTELLING (H.). – *Analysis of a complex of statistical variables into principal components*. – J. Educ. Psy. 24, 417–441, 498–520, 1933.

[Jardine et al. 71] JARDINE (N.), SIBSON (R.). – *Mathematical taxonomy*. – John Wiley & Sons, 1971.

[Jenkins et al.68] JENKINS (G. M.), WATTS (D. G.). – *Spectral analysis and its applications.* – Holden–Day, San Francisco, 1968.

[Jolliffe2002] JOLLIFFE (I. T.). – *Principal Component Analysis*. – Springer, Second edition, 2002.

[Kendall et al.76] KENDALL (M.), STUART (A.). – *The advanced theory of statistics : Volume 3 : design and analysis, and time–series.* – Charles Griffin and Co Ltd, London, 1976.

[Kim et al.99] KIM (K.–Y.), WU (Q., A.). – Comparison study of EOF techniques : analysis of nonstationary data with periodic statistics. – J. Climate, 12, 185–199, 1999.

[Kimoto et al.91] KIMOTO (M.), GHIL (M.), MO (K. C.). – *Spatial structure of the extratropical 40–day oscillation*. – Proc. 8th Conf. Atmos. Oceanic Waves and Stability, Amer. Meteor. Soc., Boston, 115–116.

[Lebart et al.2000] LEBART (L.), MORINEAU (A.), PIRON (M.). – *Statistique exploratoire multidimensionnelle*. – Dunod, 3ième édition, 2000.

[LutKepohl93] LUTKEPOHL (H.). – *Introduction to Multiple Time Series Analysis*. – Springer Verlag, 1993.

[MacDuffee33] MACDUFFEE (C. C.). – *Theory of matrices*. – Chelsea Publishing Company, New York, 1933.

[Mathsoft99] *S–PLUS 2000 User's Guide*. – Data Analysis Products Division, Seattle, WA, 1999.

[Mathsoft99] *S–PLUS 2000 Programmer's Guide*. – Data Analysis Products Division, Seattle, WA, 1999.

[Miranda et al.96] MIRANDA (M.), TILLI (P.). – *Block Toeplitz Matrices and Preconditioning*. – Calcolo, Vol. 33, 79–86, 1996.

[Morrison76] MORRISON (D. F.). – *Multivariate Statistical Methods*. – McGRAW–HILL BOOK COMPANY, 1976.

[Pearson01] PEARSON (K.). – *On lines and planes of closest fit to systems of points in space.* – Phil. Mag. 2, n°11, 559–572, 1901.

[Pincus92] PINCUS (S.M.), HUANG (W.M). – *Approximate entropy : statistical properties and applications*. – Comm. Statist., Theory Meth., 21 (11), 3061–3077, 1992.

[Plaut et al.94] PLAUT (G.), VAUTARD (R.). – Spells of Low–Frequency Oscillations and Weather Regimes in the Northern Hemisphere. – J. Atmos. Sci. 51, 210–236, 1994.

[Priestley81] PRIESTLEY (M., B.). – *Spectral analysis and time series, Vol. 1*. – Academic Press, New York, 1981.

[Quenouille49] QUENOUILLE (M. H.). – *Approximate tests of correlations in time series*. – Jour. Royal Stat. Soc., B11, 68, 1949.

[Rasmusson et al.90] RASMUSSON (E. M.), WANG (X.), ROPELEWSKI (C. F.). – *The biennal component of ENSO variability.* – J. Mar. Syst., 1, 71–96, 1990.

[Richman2000] RICHMAN (J.S.), MOORMAN (J.R.). – *Physiological time–series analysis using approximate entropy and sample entropy*. – Am. Journal Physiol. Heart Circ. Physiol. 278(6) : H2039–H2049, 2000.

[Shannon48] SHANNON (C. E.). – *A Mathematical Theory of communication*. – Bell Syst. Tech. J. 27, 379–423, 1948.

[Sims80] SIMS (C. A.). – *Macroeconomics and reality*. – Econometrica, Vol. 48, 1980.

[Sokal et al.63] SOKAL (R.R.), SNEATH (P.H.A.). – *Principles of Numerical Taxonomy*. – San Francisco, W.H. Freeman, 1963.

[Spector94] SPECTOR (P.). – *An Introduction to S and S–Plus.* – Duxbury Press, Belmont, C.A., 1994.

[Sprott2004] SPROTT (J. C.). – *Chaos and Time–Series Analysis*. – Oxford, University Press, 2004.

[Thurstone31] THURSTONE (L.). – *Multiple factor analysis*. – Psychological Review, vol. 38, 406-427, 1931.

[Tiao et al.81] TIAO (G. C.), BOX (G. E. P.). – *Modelling multiple time series with applications*. – Journal of the American Statistical Association, 76, 802–816, 1981.

[Tiao et al.83] TIAO (G. C.), TSAY (R. S.). – *Multiple time series modelling and extended sample cross correlations*. – Journal of Business and Economic Statistics, 1, 43–56, 1983.

[Tiao et al.85] TIAO (G. C.), TSAY (R. S.). – *A canonical correlation approach to modeling multivariate time series*. – American Statistical Association 1985, Proceedings of the Business and Economic Statistics Section, 112–120, 1985.

[Tsay et al.84] TSAY (R. S.), TIAO (G. C.). – *Consistent estimated sample autocorrelation function for stationary and nonstationary ARMA ARMA models.* – Journal of the American Statistical Association, Vol. 76, 1981.

[Vautard et al.89] VAUTARD (R.), GHIL (M.). – Singular spectrum analysis in nonlinear dynamics with applications to paleoclimatic time series. – Physica D., 35, 395–424, 1989.

[Vautard et al.92] VAUTARD (R.), YIOU (P.), GHIL (M.). – *Singular spectrum analysis : a toolkit for short, noisy chaotic signals*. – Physica D., 58, 95–126, 1992.

[Venables et al.2000] VENABLES (W. N.), RIPLEY (B. D.). – S Programming, Springer–Verlag, New York, 2000.

[Weare et al.82] WEARE (B. C.), NASSTROM (J. N.). – *Examples of extended empirical orthogonal function analyses*. – Mon. Wea. Rev., 110, 481–485, 1982.

[Wilder78] WILDER (G. W.). – *New Concept in Technical Trading Systems*. – Greensboro NC, Trend Research, 1978.

[Wilkinson65] WILKINSON (J. H.). – *The Algebraic Eigenvalue Problem*. – Oxford: Oxford Univ. Press, 1965.

[Williamson31] WILLIAMSON (J.). – *The latent roots of a matrix of special type*. – Bull. Amer. Math. Soc. 37, 585,1931.

[Wold38] WOLD (H.). – *A study in the analysis of stationary time series*. – Almqvist and Wiksell, Stocholm, (Second editon, 1954).

[Yaglom et al.59] YAGLOM (A. M.), YAGLOM (I. M.). – Probabilité et Information. – Dunod, Paris, 1959.

Annexe A

Projet informatique

Programme principal pour les traitements sur les séries et les calculs des entropies

Program main_entropie

```
! main permettant :
! . de lire une serie contenue dans un fichier de type Texte ;
! . de calculer les entropies de Shannon, conditionnelles et
résiduelles d'ordre 1 à m ou m-1 de la serie d'états
! . de calculer les probabilités de transition d'ordre mt-1 sur la
série d'états et les Approximate Entropy sur la série brute
! . de stocker tous ces résultats dans un même fichier de type
Texte.

use lec_ecr_serie
use traitement_entropie_shannon
use traitement_approximate_entropy

implicit none

! declaration de toutes les variables

integer*4, parameter :: ifilec = 1

character (len=len_ligne) :: cficlec
integer*4 :: lficlec

integer*4, parameter :: ifiecr = 2
character (len=len_ligne) :: cficecr
integer*4 :: lficecr

logical :: bouvert

integer*4 :: nrserie
real*4, dimension(:), allocatable :: rserie

integer*4 :: m,mt
real*4 :: r
real*4, dimension(:,:), allocatable :: mfreq
real*4, dimension(:), allocatable :: entrop
real*4, dimension(:), allocatable :: entropcond
real*4, dimension(:), allocatable :: entropres
real*4, dimension(:,:), allocatable :: mtrans
real*4, dimension(:), allocatable :: apen_mr
real*4, dimension(:), allocatable :: apenres_mr

integer*4 :: ifin, iok, ipb, ierr
character (len=len_ligne) :: ctitre
```

```
character (len=len_ligne) :: ctitre_ecr

! saisie du nom de fichier d entree contenant la serie
do
   print *, 'donnez le nom du fichier d entree contenant une serie'
   read *, cficlec
   lficlec = len_trim(cficlec)
   open (ifilec,file=cficlec(1:lficlec), &
         status='old', &
           access='sequential', &
           form='formatted', &
           action='read', &
             iostat=ierr)
   if (ierr.eq.0) then
      ! le fichier existe
        exit
   else
      ! le fichier n existe pas
      print *, 'fichier donne non trouve, recommencez'
   endif
enddo

! saisie du nom de fichier de sortie pour stocker les entropies  sur
séries d'états et ApEn sur série brute
do
   print *, 'donnez un nom de fichier de sortie pour stocker les
entropies'
   read *, cficecr
   if (cficecr(1:len_trim(cficecr)).eq.'non') then
      lficecr = 0
        exit
   endif
   lficecr = len_trim(cficecr)
   open (ifiecr,file=cficecr(1:lficecr), &
         status='replace', &
           access='sequential', &
           form='formatted', &
             iostat = ierr)
   if (ierr.eq.0) then
      ! le nom de fichier est correcte
        exit
   else
      ! le chemin du fichier n existe pas
       print *, 'Pb de nom de fichier donne, recommencez'
   endif
enddo


! saisie de m pour le calcul des entropies
```

```
print *, "donnez la valeur de m"
read *, m
if (m.le.0) then
   print *, 'm=',m,' <=0 dans main_entropie'
   print *, '==> stop'
   go to 9999
endif

! saisie de mt pour le calcul de mtrans d'ordre mt-1
print *, "donnez la valeur de mt"
read *, mt
if (mt.le.0) then
   print *, 'mt=',mt,' <=0 dans main_new'
   print *, '==> stop'
   go to 9999
endif

! saisie de r pour le calcul des ApEn
print *, "donnez la valeur de r"
read *, r
if (r.le.0) then
   print *, 'r=',r,' <=0 dans main_new'
   print *, '==> stop'
   go to 9999
endif

allocate (mfreq(m,2**m),stat=iok)
if (iok.eq.0) allocate (entrop(m),stat=iok)
if (iok.eq.0) allocate (entropcond(m),stat=iok)
if (iok.eq.0) allocate (entropres(m-1),stat=iok)
if (iok.eq.0) allocate (mtrans(2**(mt-1),2),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans main_new'
   print *, '==> stop'
   go to 9999
endif
allocate (apen_mr(m),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans main_new'
   print *, '==> stop'
   go to 9999
endif
allocate (apenres_mr(m-1),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans main_new'
   print *, '==> stop'
   go to 9999
endif
```

```
! boucle de lecture sur le contenu du fichier d entree
do

      ! decodage du bloc de tete de la serie
    call lec0_vect_serie (ifilec,ctitre,nrserie,ifin,ipb)

    if (ipb.ne.0) then
       print *, 'PB dans la routine lec0_serie appelee '
       print *, 'par main_entropie'
       print *, '==> stop'
       go to 9999
    endif

    if (ifin.ne.0) exit  ! la fin du fichier est atteinte

    print *
    print *, 'traitement de la serie :'
    print *, ctitre(1:len_trim(ctitre))

    allocate (rserie(nrserie),stat=iok)
    if (iok.ne.0) then
       print *, 'Pb d allocation dynamique dans main_entropie'
       print *, '==> stop'
       go to 9999
    endif

      ! lecture de la serie
    call lec_vect_serie (ifilec,rserie,nrserie,ifin,ipb)

    if (ipb.ne.0) then
       print *, 'PB dans la routine lec_serie appelee '
       print *, 'par main_entropie'
       print *, '==> stop'
       go to 9999
    endif

    if (ifin.ne.0) exit  ! la fin du fichier est atteinte

      ! calcul des entropies sur la série d'états
    call freq_entrop (m,mt, &
                      rserie,nrserie, &
                      mfreq,entrop,entropcond,entropres,mtrans, &
                        ipb)

    if (ipb.ne.0) then
       print *, 'PB dans la routine freq_entropappelee '
       print *, 'par main_new'
       print *, '==> stop'
```

```
      go to 9999
    endif

      ! calcul de ApEn sur la série brute

      call approximate_entropy
(m,r,rserie,nrserie,ApEn_mr,ApEnres_mr,ipb)

    if (ipb.ne.0) then
       print *, 'PB dans la routine approximate_entropy    appelee '
       print *, 'par main_new'
       print *, '==> stop'
       go to 9999
    endif


      if (lficecr.ne.0) then

       !  ecriture du bloc de tete du fichier
         ctitre_ecr = cligne_blanc
         write (ctitre_ecr,*) trim(ctitre),&
         ' entropies de Shannon d ordre 1 a',m
       call ecr0_vect_serie (ifiecr,ctitre_ecr,m,ipb)
       if (ipb.ne.0) then
          print *, 'PB dans la routine ecr0_vect_serie appelee '
          print *, 'par main_new'
          print *, '==> stop'
          go to 9999
       endif
         ctitre_ecr = cligne_blanc
         write (ctitre_ecr,*) trim(ctitre),&
         ' entropies conditionnelles d ordre 1 a',m
       call ecr0_vect_serie (ifiecr,ctitre_ecr,m,ipb)
       if (ipb.ne.0) then
          print *, 'PB dans la routine ecr0_vect_serie appelee '
          print *, 'par main_new'
          print *, '==> stop'
          go to 9999
       endif
         ctitre_ecr = cligne_blanc
         write (ctitre_ecr,*) trim(ctitre),&
         ' entropies residuelles d ordre 1 a',m-1
       call ecr0_vect_serie (ifiecr,ctitre_ecr,m-1,ipb)
       if (ipb.ne.0) then
          print *, 'PB dans la routine ecr0_vect_serie appelee '
          print *, 'par main_new'
          print *, '==> stop'
          go to 9999
       endif
```

```
          ctitre_ecr = cligne_blanc
          write (ctitre_ecr,*) trim(ctitre),&
          ' mtrans'
        call ecr0_mat_serie (ifiecr,ctitre_ecr,2**(mt-1),2,ipb)
        if (ipb.ne.0) then
           print *, 'PB dans la routine ecr0_mat_serie appelee '
           print *, 'par main_new'
           print *, '==> stop'
           go to 9999
        endif
          ctitre_ecr = cligne_blanc
          write (ctitre_ecr,*) trim(ctitre),&
          ' ApEn d ordre 1 a',m
        call ecr0_vect_serie (ifiecr,ctitre_ecr,m,ipb)
        if (ipb.ne.0) then
           print *, 'PB dans la routine ecr0_vect_serie appelee '
           print *, 'par main_new'
           print *, '==> stop'
           go to 9999
        endif
          ctitre_ecr = cligne_blanc
          write (ctitre_ecr,*) trim(ctitre),&
          ' ApEnres d ordre 1 a',m-1
        call ecr0_vect_serie (ifiecr,ctitre_ecr,m-1,ipb)
        if (ipb.ne.0) then
           print *, 'PB dans la routine ecr0_vect_serie appelee '
           print *, 'par main_new'
           print *, '==> stop'
           go to 9999
        endif

        ! ecriture des entropies de Shannon dans le fichier de sortie
          call ecr_vect_serie (ifiecr,entrop,m,ipb)
        if (ipb.ne.0) then
           print *, 'PB dans la routine ecr_vect_serie appelee '
           print *, 'par main_entropie'
           print *, '==> stop'
           go to 9999
        endif

        ! ecriture de l entropie conditionnelle dans le fichier de
sortie
        call ecr_vect_serie (ifiecr,entropcond,m,ipb)
        if (ipb.ne.0) then
           print *, 'PB dans la routine ecr_vect_serie appelee '
           print *, 'par main_entropie'
           print *, '==> stop'
           go to 9999
        endif
```

```
      ! ecriture de l entropie residuelle dans le fichier de sortie
      call ecr_vect_serie (ifiecr,entropres,m-1,ipb)
      if (ipb.ne.0) then
         print *, 'PB dans la routine ecr_vect_serie appelee '
         print *, 'par main_entropie'
         print *, '==> stop'
         go to 9999
      endif

      ! ecriture de mtrans dans le fichier de sortie
      call ecr_mat_serie (ifiecr,mtrans,2**(mt-1),2,ipb)
      if (ipb.ne.0) then
         print *, 'PB dans la routine ecr_mat_serie appelee '
         print *, 'par main_entropie'
         print *, '==> stop'
         go to 9999
      endif

        ! ecriture des ApEn dans le fichier de sortie
      call ecr_vect_serie (ifiecr,apen_mr,m,ipb)
      if (ipb.ne.0) then
         print *, 'PB dans la routine ecr_vect_serie appelee '
         print *, 'par main_entropie'
         print *, '==> stop'
         go to 9999
      endif

        ! ecriture des ApEnres dans le fichier de sortie
      call ecr_vect_serie (ifiecr,apenres_mr,m-1,ipb)
      if (ipb.ne.0) then
         print *, 'PB dans la routine ecr_vect_serie appelee '
         print *, 'par main_entropie'
         print *, '==> stop'
         go to 9999
      endif


     endif

   deallocate (rserie)

enddo

deallocate (mfreq)
deallocate (entrop)
deallocate (entropcond)
deallocate (entropres)
deallocate (mtrans)
```

```
deallocate (apen_mr)
deallocate (apenres_mr)

close (ifilec)
if (lficecr.ne.0) close (ifiecr)

print *, 'Fin nominale de main_entropie'
stop

! sortie non nominale du pgm
9999 continue
if (allocated(rserie)) deallocate (rserie)
if (allocated(mfreq)) deallocate (mfreq)
if (allocated(entrop)) deallocate (entrop)
if (allocated(entropcond)) deallocate (entropcond)
if (allocated(entropres)) deallocate (entropres)
if (allocated(mtrans)) deallocate (mtrans)
if (allocated(apen_mr)) deallocate (apen_mr)
if (allocated(apenres_mr)) deallocate (apenres_mr)
inquire (ifilec,opened=bouvert)
if (bouvert) close(ifilec)
inquire (ifiecr,opened=bouvert)
if (bouvert) close(ifiecr)

end program main_entropie
```

Module de lecture et écriture des fichiers de données

module lec_ecr_entropie

```
! Module contenant tous les spgm de lecture et d ecriture des
! fichiers de type texte.

integer*4, parameter :: len_ligne = 500
integer*4, parameter :: lformat_standard = 6
character (len=lformat_standard), parameter :: cformat_standard
='5e14.7'

contains

!*************************************************

subroutine lec0_vect_serie (ific,ctitre,nvect,ifin,ipb)

! Spgm de lecture initiale d un fichier de type texte
! contenant un vecteur.
! Ce spgm ne lit que le bloc de tete du fichier afin
! de decoder le titre contenu dans le fichier
! ainsi que sa dimension.

! entrees :
! ific : numero logique du fichier a lire

! sorties :
! ctitre : titre du fichier
! nvect : dimension du vecteur
! ifin : indice de detection de fin de fichier
! ipb : indice de mauvais fonctionnement du spgm

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: ific
character (len=len_ligne), intent(out) :: ctitre
integer*4, intent(out) :: nvect
integer*4, intent(out) :: ifin
integer*4, intent(out) :: ipb

! declarations des internes
character (len=len_ligne) :: cformat
integer*4 :: lformat
integer*4 :: ipb_nvect, ipb_format
```

```
ipb = 0

ifin = 0

! lecture du bloc de tete
read (ific,"(a)",end=1000,err=9999) ctitre
read (ific,*,end=1000,err=9999) nvect,cformat
lformat = len_trim(cformat)

! detection de dimension incorrecte
ipb_nvect = 0
if (nvect.le.0) then
   ipb_nvect = 1
endif
if (ipb_nvect.ne.0) then
   print *, 'Nvect dans le fichier est negatif ou nul'
endif

! detection de format de lecture incompatible du spgm
ipb_format = 0
if (lformat.eq.0) then
   ipb_format = 1
else
   if (lformat.ne.lformat_standard) then
      ipb_format = 1
   else
      if (cformat(1:lformat).ne.cformat_standard) then
           ipb_format = 1
        endif
   endif
endif
if (ipb_format.ne.0) then
   print *, 'Le format decrit dans le fichier est incorrect'
endif

if (ipb_nvect.ne.0  .or. ipb_format.ne.0) then
   ipb = 1
endif

return

! sortie prematuree du spgm sur detection de la fin du fichier
1000 continue
ifin = 1
return

! sortie non nominale du spgm
9999 continue
print *, 'Pb lors du read dans lec0_vect_serie'
```

```
ipb = 1
return

end subroutine lec0_vect_serie

!*****************************************************

subroutine lec0_mat_serie (ific,ctitre,nmat1,nmat2,ifin,ipb)

! Spgm de lecture initiale d un fichier de type texte
! contenant une matrice.
! Ce spgm ne lit que le bloc de tete du fichier afin
! de decoder le titre contenu dans le fichier
! ainsi que les dimensions de la matrice.

! entrees :
! ific : numero logique du fichier a lire

! sorties :
! ctitre : titre du fichier
! nmat1 : 1 ere dimension de la matrice
! nmat2 : 2 ieme dimension de la matrice
! ifin : indice de detection de fin de fichier
! ipb : indice de mauvais fonctionnement du spgm

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: ific
character (len=len_ligne), intent(out) :: ctitre
integer*4, intent(out) :: nmat1,nmat2
integer*4, intent(out) :: ifin
integer*4, intent(out) :: ipb

! declarations des internes
character (len=len_ligne) :: cformat
integer*4 :: lformat
integer*4 :: ipb_nmat, ipb_format

ipb = 0

ifin = 0

! lecture du bloc de tete
read (ific,"(a)",end=1000,err=9999) ctitre
read (ific,*,end=1000,err=9999) nmat1,nmat2,cformat
lformat = len_trim(cformat)

! detection de dimension incorrecte
```

```
ipb_nmat = 0
if (nmat1.le.0 .or. nmat2.le.0) then
   ipb_nmat = 1
endif
if (ipb_nmat.ne.0) then
   print *, 'Nmat1 ou nmat2 dans le fichier est negatif ou nul'
endif

! detection de format de lecture incompatible du spgm
ipb_format = 0
if (lformat.eq.0) then
   ipb_format = 1
else
   if (lformat.ne.lformat_standard) then
      ipb_format = 1
   else
      if (cformat(1:lformat).ne.cformat_standard) then
           ipb_format = 1
        endif
   endif
endif
if (ipb_format.ne.0) then
   print *, 'Le format decrit dans le fichier est incorrect'
endif

if (ipb_nmat.ne.0  .or. ipb_format.ne.0) then
   ipb = 1
endif

return

! sortie prematuree du spgm sur detection de la fin du fichier
1000 continue
ifin = 1
return

! sortie non nominale du spgm
9999 continue
print *, 'Pb lors du read dans lec0_mat_serie'
ipb = 1
return

end subroutine lec0_mat_serie

!************************************************************

subroutine lec_vect_serie (ific,vect,nvect,ifin,ipb)

! Spgm de lecture d un vecteur contenu dans  un fichier
```

```
! de type texte. La dimension du vecteur a lire
! doit etre definie au niveau du pgm appelant apres
! avoir appele le spgm lec0_vect_serie.

! entrees :
! ific : numero logique du fichier a lire
! nvect : dimension du vecteur

! sorties :
! vect : vecteur lu de dimension nvect
! ifin : indice de detection de fin de fichier
! ipb : indice de mauvais fonctionnement du spgm

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: ific
integer*4, intent(in) :: nvect
real*4, dimension(nvect), intent(out) :: vect
integer*4, intent(out) :: ifin
integer*4, intent(out) :: ipb

! declarations des internes
integer*4 :: ivect

ipb = 0

ifin = 0

read (ific,"(5e14.7)",end=1000,err=9999) (vect(ivect),ivect=1,nvect)

return

! sortie du spgm sur detection de fin de fichier
1000 continue
ifin = 1

! sortie non nominale de la routine
9999 continue
print *, 'Pb lors du read dans lec_vect_serie'
ipb = 1
return

end subroutine lec_vect_serie

!************************************************************

subroutine lec_mat_serie (ific,xmat,nmat1,nmat2,ifin,ipb)
```

```
! Spgm de lecture d une matrice contenu dans un fichier
! de type texte. La dimension de la matrice a lire
! doit etre definie au niveau du pgm appelant apres
! avoir appele le spgm lec0_serie.

! entrees :
! ific : numero logique du fichier a lire
! nmat1 : 1 ere dimension de la matrice
! nmat2 : 2 ieme dimension de la matrice

! sorties :
! xmat : vecteur lu de dimension nmat
! ifin : indice de detection de fin de fichier
! ipb : indice de mauvais fonctionnement du spgm

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: ific
integer*4, intent(in) :: nmat1,nmat2
real*4, dimension(nmat1,nmat2), intent(out) :: xmat
integer*4, intent(out) :: ifin
integer*4, intent(out) :: ipb

! declarations des internes
integer*4 :: imat1,imat2

ipb = 0

ifin = 0

do imat1 = 1,nmat1
    read (ific,"(5e14.7)",end=1000,err=9999) &
      (xmat(imat1,imat2),imat2=1,nmat2)
enddo
return

! sortie du spgm sur detection de fin de fichier
1000 continue
ifin = 1

! sortie non nominale de la routine
9999 continue
print *, 'Pb lors du read dans lec_mat_serie'
ipb = 1
return

end subroutine lec_mat_serie
```

```
!*************************************************************

subroutine ecr_vect_serie (ific,ctitre, &
                           vect,nvect,ipb)

! Spgm d ecriture d un vecteur dans un fichier de type texte.

! entrees :
! ific : numero logique du fichier a lire
! ctitre : titre du fichier
! vect : vecteur de dimension nvect
! nvect : dimension du vecteur

! sorties :
! ipb : indice de mauvais fonctionnement du spgm

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: ific
character (len=len_ligne), intent(in) :: ctitre
integer*4, intent(in) :: nvect
real*4, dimension(nvect), intent(in) :: vect
integer*4, intent(out) :: ipb

! declarations des internes
integer*4 :: ivect

ipb = 0

write (ific,"(a)",err=9999) adjustl(ctitre)
write (ific,"(i5,4x,'(',a,')')",err=9999) nvect,cformat_standard
write (ific,"(5e14.7)",err=9999) (vect(ivect),ivect=1,nvect)

return

! sortie non nominale du spgm
9999 continue
print *, 'Pb lors du write dans ecr_vect_serie'
ipb = 1
return

end subroutine ecr_vect_serie

!*************************************************************

subroutine ecr_mat_serie (ific,ctitre, &
                          xmat,nmat1,nmat2,ipb)
```

```
! Spgm d ecriture d une matrice dans un fichier de type texte.

! entrees :
! ific : numero logique du fichier a lire
! ctitre : titre du fichier
! xmat : matrice de dimension nmat1,nmat2
! nmat1 : 1 ere dimension de la matrice
! nmat2 : 2 ieme dimension de la matrice

! sorties :
! ipb : indice de mauvais fonctionnement du spgm

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: ific
character (len=len_ligne), intent(in) :: ctitre
integer*4, intent(in) :: nmat1,nmat2
real*4, dimension(nmat1,nmat2), intent(in) :: xmat
integer*4, intent(out) :: ipb

! declarations des internes
integer*4 :: imat1,imat2

ipb = 0

write (ific,"(a)",err=9999) adjustl(ctitre)
write (ific,"(i5,4x,i5,4x,'(',a,')')",err=9999) &
nmat1,nmat2,cformat_standard
do imat1 = 1,nmat1
    write (ific,"(5e14.7)",err=9999) &
      (xmat(imat1,imat2),imat2=1,nmat2)
enddo
return

! sortie non nominale du spgm
9999 continue
print *, 'Pb lors du write dans ecr_mat_serie'
ipb = 1
return

end subroutine ecr_mat_serie

end module lec_ecr_entropie
```

Module de traitement des séries et calcul des entropies de Shannon

module traitement_entropie_shannon

```
! Module de traitement permettant :
! 1. de transformer la série initiale 'rserie' en une série des
points de retournements
!      puis en une série des modalités 'True' pour les 'pics' ou
'False' pour les 'creux':'bserie';
! 2. de définir la matrice 'bgram' de tous les k-grammes possibles
(unimodale, bimodale, etc.) de longueur 1 à m
!      pour les 2 modalités 'True' ou 'False' ;
! 3. de calculer à partir de la matrice 'bgram', la matrice 'mfreq'
des fréquences des k-grammes rencontrés dans 'bserie';
! 4. de calculer à partir de la matrice 'mfreq', les vecteurs
'entrop','entropcond' et 'entropres'
!      des entropies (Hk), des entropies conditionnelles (hk) et
résiduellles (dk) d'ordre 1 à m ou m-1;
! 5. de calculer à partir de la matrice 'mfreq', la matrice 'mtrans'
des probabilités de transition d'ordre mt-1 ;


contains


subroutine freq_entrop (m,mt, &
                          rserie,nrserie, &

mfreq,entrop,entropcond,entropres,mtrans, &
                             ipb)

implicit none

! declarations des entrees sorties
integer*4, intent(in) :: m,mt
integer*4, intent(in) :: nrserie
real*4, dimension(nrserie), intent(in):: rserie
real*4, dimension(m,2**m), intent(out) :: mfreq
real*4, dimension(2**(mt-1),2), intent(out) :: mtrans
real*4, dimension(m), intent(out) :: entrop
real*4, dimension(m), intent(out) :: entropcond
real*4, dimension(m-1), intent(out) :: entropres
integer*4, intent(out) :: ipb

! declaration des internes
logical*1, dimension(:), allocatable :: bserie
logical*1, dimension(:,:), allocatable :: bgram
```

```
integer*4 :: i,j,iok,mmm,mm ,nrgram,nrgram_old,ncompt
logical *1 :: bcompt

ipb = 0
! detection de valeur de m incorrecte
if (m.le.0) then
   print *, 'm=',m,'<= 0 dans freq_entrop'
   ipb = 1
   return
endif
if (m.gt.nrserie) then
   print *, 'm=',m,'> nrserie=',nrserie,' dans freq_entrop'
   ipb = 1
   return
endif

allocate (bserie(nrserie-1),stat=iok)
if (iok.eq.0) allocate (bgram(m,2**m),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans freq_entrop'
   go to 9999
   return
endif

! Algo : des points de retournements à bsérie
do i= 1,nrserie-1
    if (rserie(i)-rserie(i+1).lt.0.) then
         bserie(i) = .false.
    else
       bserie(i) = .true.
      endif
enddo

! Algo : création matrice bgram
bgram = .false.
nrgram = 1
do mm =1,m
      nrgram_old = nrgram
      do i = 1,2**(mm-1)
            nrgram = nrgram+1
          do mmm = 1,mm-1
                bgram(mmm,nrgram) = bgram(mmm,nrgram-nrgram_old)
            enddo
          bgram(mm,nrgram) = .true.
      enddo
enddo

!Algo : création matrice mfreq
do mm =1,m
```

```
   do i =1,2**mm
        ncompt=0
        do j =1,nrserie-1-mm+1
           bcompt = .true.
           do mmm =1,mm
                if (bserie(mmm+j-1).NEQV.bgram(mmm,i)) then
                   bcompt = .false.
                   exit
            endif
         enddo
         if (bcompt) then
            ncompt = ncompt+1
         endif
      enddo
        if (ncompt.ne.0) then
         mfreq(mm,i) = float(ncompt)/float(nrserie-1-mm+1)
      else
           mfreq(mm,i) = 0.
      endif
   enddo
enddo

! Algo : création vecteurs entropies simples, conditionnelles
d'ordre m et résiduelles d'ordre m-1
entrop = 0.
do mm =1,m
   do i=1,2**mm
      if (mfreq(mm,i).ne.0.) then
         entrop(mm) = entrop(mm)-(mfreq(mm,i)*log(mfreq(mm,i)))
        endif
   enddo
enddo

entropcond(1)=entrop(1)
do mm =2,m
   entropcond(mm) = entrop(mm)-entrop(mm-1)
enddo

do mm =1,m-1
   entropres(mm) = entropcond(mm)-entropcond(mm+1)
enddo


! Algo : création matrice de transition d'ordre mt-1
do j = 1,2**(mt-1)
   if (mfreq(mt-1,j).eq.0.) then
      mtrans(j,1) = 0.
   else
      mtrans(j,1) = mfreq(mt,j)/mfreq(mt-1,j)
```

```
   endif
enddo
do j =1,2**(mt-1)
   if (mfreq(mt-1,j).eq.0.) then
      mtrans(j,2) = 0.
   else
      mtrans(j,2) = mfreq(mt,j+2**(mt-1))/mfreq(mt-1,j)
   endif
enddo

!Affichage vecteur bserie
!print *, 'Vecteur bserie'
!print *, (bserie(j),j= 1,nrserie-1)

!Affichage matrice bgram
!print *, 'Matrice bgram'
!do mm =1,m
!    print *, (bgram(mm,j),j= 1,2**mm)
!enddo

! Affichage matrice mfreq
!print *, 'Matrice mfreq'
!do mm =1,m
!    print *, (mfreq(mm,i),i= 1,2**mm)
!enddo

! Affichage matrice mtrans
print *, 'Matrice mtrans d ordre',mt-1
do j =1,2**(mt-1)
    print *, (mtrans(j,i),i= 1,2)
enddo

deallocate (bgram)
deallocate (bserie)
return

! sortie nom nominale du spgm
9999 ipb = 1
if (allocated(bgram)) deallocate(bgram)
if (allocated(bserie)) deallocate(bserie)

end   subroutine freq_entrop
```

end module traitement_entropie_shannon

Module de traitement pour le calcul des entropies de Pincus

module traitement_approximate_entropy

```
! ApEn er ApEn residuelles sur les séries brutes
! Dans ce module, 2 routines pour les 2 méthodes de calcul suivantes
:
!   méthode de Grassberger-Procaccia (moyenne des (Cmr(i)):
approximate_entropy
!   méthode de Pincus-Kolmogorov (moyenne des (log(Cmr(i)):
approximate2_entropy à faire !!

contains

subroutine approximate_entropy (m,r, &
                 rserie,nrserie, &
                        ApEn_mr,ApEnres_mr,ipb)


implicit none

! declarations des entrees sorties
integer*4,  intent(in):: m
integer*4, intent(in) :: nrserie
real*4,  intent(in):: r
real*4, dimension(nrserie-m+1), intent(in) :: rserie
real*4, dimension(m), intent(out) :: ApEn_mr
real*4, dimension(m-1), intent(out) :: ApEnres_mr
integer*4, intent(out) :: ipb


!declaration des internes
integer*4 :: i,j,k,iok,mm
logical*1 :: bcompt
integer*4, dimension(:), allocatable :: vcompt
real*4, dimension(:), allocatable :: C1
real*4, dimension(:), allocatable :: C2

 ipb = 0
! detection de valeur de m incorrecte
if (m.le.0) then
   print *, 'm=',m,'<= 0 dans approximate_entropy'
   ipb = 1
   return
endif
if (m.gt.nrserie) then
   print *, 'm=',m,'> nrserie=',nrserie,' dans approximate_entropy'
```

```
   ipb = 1
   return
endif


allocate (vcompt(nrserie),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans ApEn'
endif

allocate (C1(m),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans ApEn'
endif

allocate (C2(m),stat=iok)
if (iok.ne.0) then
   print *, 'Pb d allocation dynamique dans ApEn'
endif

! Algo : pour le calcul des C(m,r)=C1(m)
C1=0.
Do mm= 1,m
   Do i= 1,nrserie-mm+1
      vcompt(i)=0
      Do j =1, nrserie-mm+1
         bcompt= .true.
           Do k= 0,mm-1
              if (ABS(rserie(i+k)-rserie(k+j)).GE.r) then
                   bcompt = .false.
                   exit
                endif
           Enddo
           if (bcompt) then
              vcompt(i)=vcompt(i)+1
           endif
        Enddo
        C1(mm)=C1(mm)+float(vcompt(i))/float(nrserie-mm+1)
   Enddo
   C1(mm)=C1(mm)/float(nrserie-mm+1)
  Enddo

! Algo : pour le calcul des C(m+1,r)=C2(m)
C2=0.
Do mm= 1,m
   Do i= 1,nrserie-mm
      vcompt(i)=0
      Do j =1, nrserie-mm
         bcompt= .true.
```

```
         Do k= 0,mm
            if (ABS(rserie(k+i)-rserie(k+j)).GE.r) then
                 bcompt = .false.
                 exit
              endif
         Enddo
         if (bcompt) then
            vcompt(i)=vcompt(i)+1
         endif
      Enddo
      C2(mm)=C2(mm)+float(vcompt(i))/float(nrserie-mm)
 Enddo
 C2(mm)=C2(mm)/float(nrserie-mm)
enddo

! Calcul des ApEn(m,r) et ApEnres(m,r)

do mm= 1,m
ApEn_mr(mm) = log(C1(mm)/C2(mm))
enddo

do mm =1,m-1
ApEnres_mr(mm) = ApEn_mr(mm)-ApEn_mr(mm+1)
enddo


!print *, C1

!print *, C2

deallocate (vcompt)
deallocate (C1)
deallocate (C2)

end subroutine approximate_entropy

end module traitement_approximate_entropy
```

Annexe B

2005 ASA Proceedings

Publication Acte SFC 2004

Communication orale ISF 2006

Identification of Time-series Models: Application to ARMA Processes.

Carole Toque, Bernard Burtschy
Ecole Nationale Supérieure des Télécommunications
46, rue Barrault,
75634 PARIS CEDEX 13
FRANCE

Abstract

In identification of time series, the proposed method combines the usual approach by autocorrelations and a structural approach, less usual, by analysis of oscillators and theory of information, through visualization by factorial methods (principal component analyses PCA and multiple correspondences MCA). It supplies reference graphic models and pertinent criteria for, identification and estimation of models, and identification of classes. In this paper, the method is applied to simulated AR(1) and MA(1) processes.

Based on simulated temporal matrices, first PCA produce good quality of processes representation, with significant groupings and oppositions preserving some properties of ARMA autocorrelation functions. PCA becomes a reliable technic in the research of pertinent criteria to identify time-series models. Directly based on autocorrelation matrices, PCA give better results and it ensues a first reference graphic model with identification and estimation.

Description and measure of possible structural change lead us to introduce oscillators, frequencies and measures of entropy. This is the structural approach. To establish non-linearity between the numerous criteria and to increase the discriminative ability between the series, classifications on MCA are built over measures of entropy and produce a second reference graphic model with three classes of processes. The method with incertitude measures is justified. It must be extended to usual ARMA processes and operational data.

Keywords: identification of time series, ARMA processes, reference graphic models, autocorrelation functions, structural analysis, oscillators, theory of information, entropy, principal component analysis, multiple correspondences analysis.

1. Introduction

How, factorial technics like PCA and MCA may contribute to identify time-series models ? First, learning PCA after a significant reduction of time-series data, brings evidence on their known properties. To illustrate this step, we choose to present in this paper, a PCA said 'temporal' on simulated AR(1) and MA(1) processes. So, the scores model on the first two principal components, produces a good quality of processes representation with groupings and oppositions preserving some properties of ARMA autocorrelation functions. These results remind the correlogram properties used in the Box and Jenkins methodology, (Box and Jenkins, 1976). It ensues a second PCA directly based on autocorrelation (ACF) and partial autocorrelation (PACF) functions. The scores model is of better quality providing the first reference graphic model. Groupings and oppositions of AR(1) and MA(1) time series are independant of samples : at the same time, identification of time series and estimation of the coefficients are possible.

In the research of other pertinent criteria to identify time-series models, the resort of other technics is justified when it combines structural changes with advance (or peak) and decline (or through). In that case, the approach by the autocorrelations is inefficient and it's better to use a structural approach like the analysis of oscillators coming from the technical analysis (Wilder, 1978) and reflecting turning points (Kendall and Stuart, 1976 and Harding, 2003), finally combined with the theory of information (Shannon, 1948). A MCA on measures of entropy completed by a classification, allows to identify classes of time series said 'weak' which are distinguished from classes of processes said 'robust' and 'semi-robust'.

2. PCA to identify time-series models

In this section, we present a temporal PCA followed by a PCA based on ACF and PACF elements of simulated AR(1) and MA(1) processes. A first reference graphic model is deducted and it permits identification and estimation of time series.

2.1 Simulation of stationary AR(1) and MA(1) time series

18 AR(1) models and 18 MA(1) models, stationary with mean zero, have been generated from ϕ_1 and θ_1 coefficients between -0.9 and +0.9 values and a 0.1 step. For each model of length 500, 9 samples are computed coming from gaussian noise processes with

mean zero and distinct variances. Precisely, we choose for the first three samples (Zone 1 - Fig.1.), a same value (10) to initialize the random numbers generator and 3 distinct values of gaussian noises variances (10,50,90), for the next three, (Zone 2 - Fig.1.), an other initial value (50) and the 3 distinct variances (10,50,90), and for the last three samples (Zone 3 - Fig.1.), the initial value (1023) with the same distinct variances (10,50,90). So, the simulated temporal matrix is of dimension (324x500) : the number of rows corresponds to (9x36) AR(1) and MA(1) time series and the number of columns corresponds to 500 variables or 500 t-instants. A first PCA is directly built on this matrix.

2.2 PCA on the temporal matrix

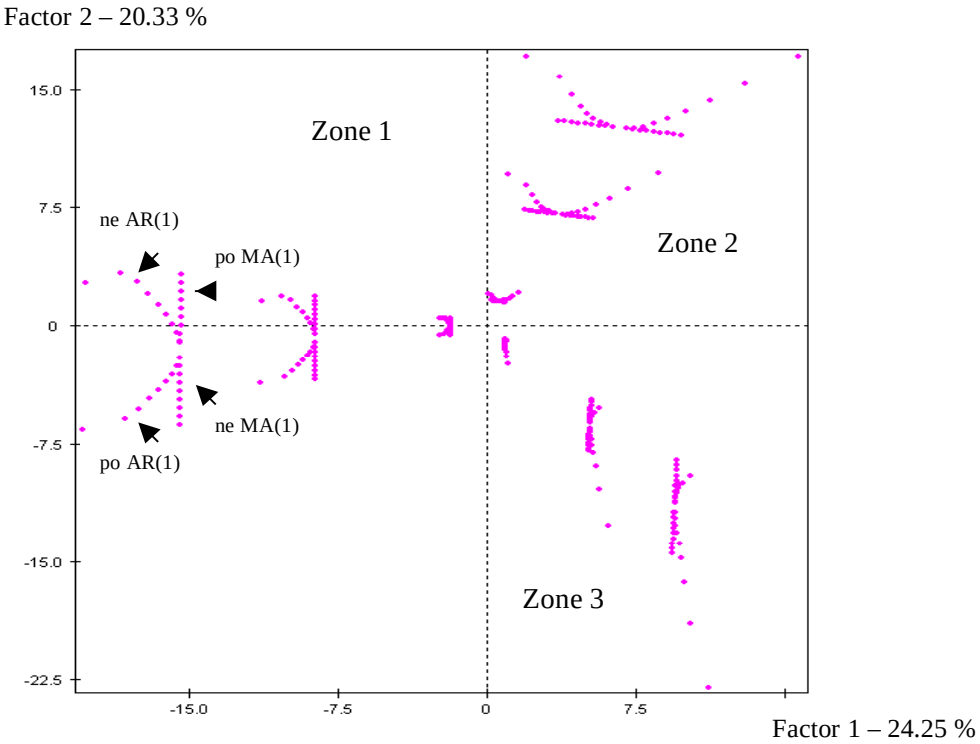


Fig. 1. The temporal scores model and the 9 samples.

The analysis of eigenvalues histogram (degressive form from the third rank) and correlations circles, showed that the best factorial visualization of the 9 samples is obtained in the (F1,F2) first two principal components plane, represented in (Fig. 1.). For 500 observed instants, the 2 axes explain near 45 % of the initial information. The number of variables is very important, so, at the first time, we limit our study to the time series of the (Fig. 1.).

Several levels of groupings and oppositions are observed for the simulated time series : groupings by zone of simulation (Zones 1 to 3), groupings by sample and, inside each sample, groupings and oppositons function of their coefficients. In each zone, we find the 3 samples with the same initial value but differentied by the gaussian noises variances. For instance, the F1 axis in its negative part, reflects the gaussian noises variances of the three (Zone 1) samples. Finally, isolating one sample whatever the zone, it appears significant groupings and oppositions of AR(1) and MA(1) processes. We can notice that, for the zone 1, the F2 axis opposes the class of AR(1) with negative coefficients (ne AR(1)) and MA(1) with positive coefficients (po MA(1)), against the class of MA(1) with negative coefficients (ne MA(1)) and AR(1) with positive coefficients (po AR(1)). We find the results of symetry of AR(1) ACF and MA(1) PACF on the one hand, and MA(1) ACF and AR(1) PACF on the other hand.

PCA is a reliable technic in the research of pertinent criteria to identify time-series models. However, the main groups reflect the 9 samples of simulated AR(1) et MA(1) time series. Is it possible to view a scores model in which the AR(1) and MA(1) processes get nearer function of their coefficients, whatever the samples ? So, we propose to build a second PCA, based on ACF and PACF elements of the simulated AR(1) and MA(1) processes.

2.3 PCA on the ACF and PACF of simulated processes

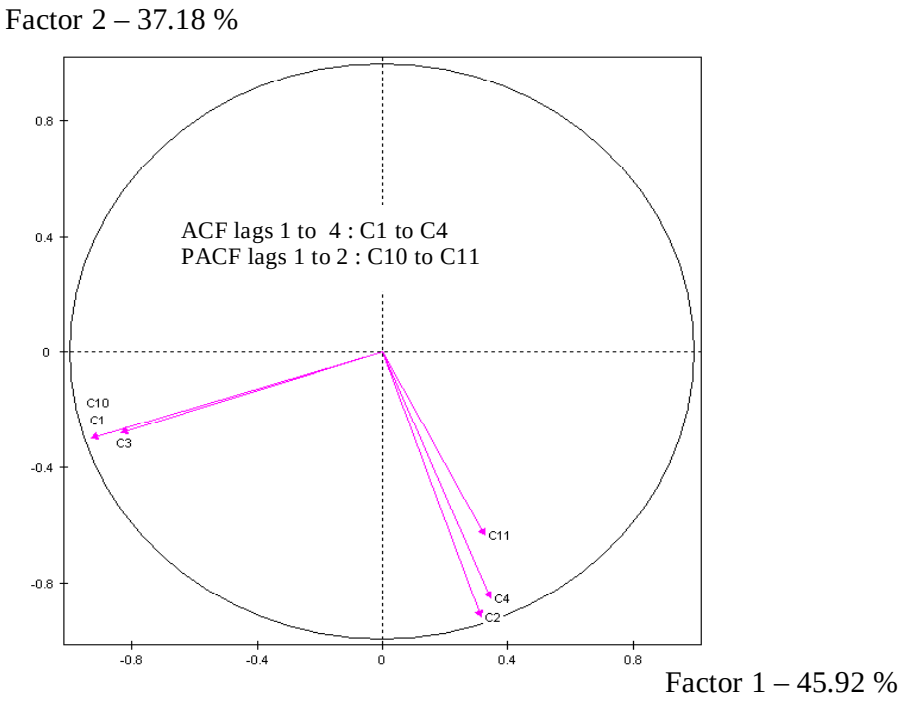


Fig. 2. The correlations circle.

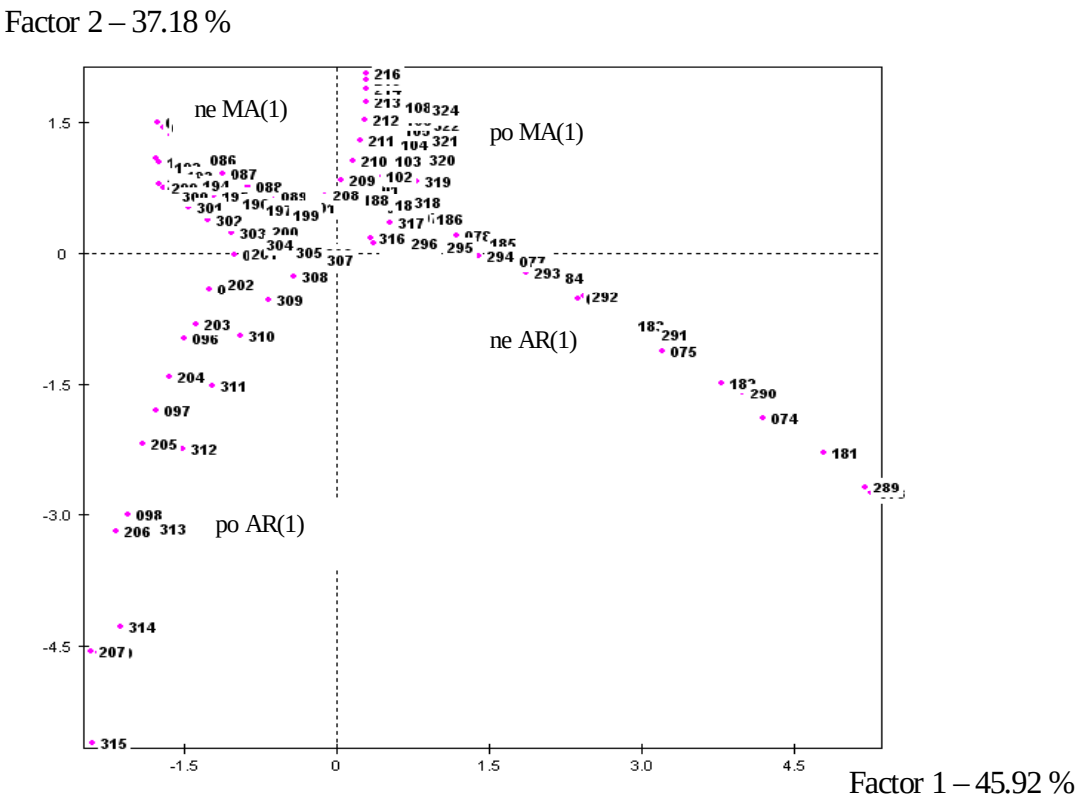


Fig. 3. The first reference scores model.

A principal component analysis, based on the first ten ACF and the first ten PACF elements, permitted to isolate the more significant lags. So, we retained the lags 1 to 4 for the ACF (denoted ACF1,ACF2, …) and the lags 1 to 2 for the PACF (denoted PACF1,PACF2, …) of the simulated time series. Therefore, a next PCA is built and it's again in the (F1,F2) factorial plane that the correlations circle (Fig. 2.) and the scores graphics (Fig. 3.) are of better quality. Proximities of variables near the correlations circle, proximities between variables and with the factorial axes, allow to characterize the F2 axis by the couple (ACF2,ACF4) and the F1 axis by the couple (ACF1,PACF1). It ensues possible interpretations of

groupings and oppositions between the time series which are made independently from the samples.

But, what about the time series representativity ? An AR(1) or an MA(1) process with a coefficient in absolute value near 1 is better represented than a process with a coefficient in absolute value near zero. Then, we distinguish 2 groups of processes : the processes said 'robust to semi-robust' with coefficients in absolute value between 0.9 and 0.4, and the processes said 'weak' with coefficients in absolute value lower than 0.4. In brief, for these 'robust to semi-robust' processes, the F2 axis with the (ACF2,ACF4) variables opposes the AR(1) to the MA(1), and the F1 axis with the (ACF1,PACF1) opposes the class of negative AR(1) and positive MA(1), against the class of negative MA(1) and positive AR(1).

Finally, the objective which was to aggregate the time series function of their coefficients whatever the samples, is effective, excepted the processes said 'weak'. The (Fig. 3.) illustrates the first reference graphic model where we can project an additional AR(1) or MA(1) time series and recognize its type and its coefficient.

In the next section, the approach is more qualitative and does not require any property of 2-order moments. Its aim is to identify classes of 'robust', 'semi-robust' and 'weak' processes. We remind that the identification of these time series is particularly delicate in the classical approach.

3. Structural identification of AR(1) and MA(1) processes

3.1 Structural analysis and measures of incertitude

To describe and measure structural changes of a time series, the initial process is changed in a series of states. By first differences of data, a series of symbols (0) or (1) is built, representing declines or advances, and reflecting turning points of the initial series (Kendall and Stuart, 1976).
Then, we measure frequencies of symbols and sequencies of k-symbols on each series of states to estimate the different probabilities. However, how to qualify the information about these probabilities ? Or, how to qualify the incertitude ?
So, we use the theory of information with measures of entropy (Shannon, 1948).

A project developped in Fortran 90 and integrated in 'Splus' allows to build the series of states, to estimate the probabilities $P_k = \{ p_{k,1}, p_{k,2}..., p_{k,m^k} \}$ of the m^k k-uplets states of a series $\{Z_t\}$, to calculate the simple entropies H_k of k-order (denoted too 'Shk') defined by $H_k(P_k) = -\sum_{i=1}^{m^k} p_{k,i} \log_2 p_{k,i}$, to calculate the conditional entropies h_k of k-order (denoted too 'Condk') defined by $h_1 = H_1(P_1)$ else $h_k = H_k(P_k / P_{k-1}) = H_k(P_k) - H_{k-1}(P_{k-1})$, and to calculate the residual entropies d_k of k-order (denoted too 'Resk') defined by $d_k = h_k - h_{k+1}$, k between 1 and q or $q-1$ for d_k.

For instances, the entropy h_k is the mean incertitude of the k^{th} symbol of a k-sequence where the first $(k-1)$ symbols are known, and a conditional entropy of 2-order takes in account the k-sequences (0,1) and (1,0) which are the expressions of turning points.

The project is applied to one sample of simulated AR(1) and MA(1) time series, with $q = 5$ to calculate frequencies and entropies. It ensues for each process : a vector of (0) and (1) and a vector of entropies of dimension (14). At last, to make easier the interpretation of time series classes, we choose to build on the entropies matrix, a classification on principal components of a 3 modalities MCA.

3.2 Classification on principal components of a structural MCA: 3 classes of AR(1) and MA(1) processes

A hierarchical ascendant classification (HAC) on a MCA, allows to illustrate graphics with variables and time series adding classes. The graphic reading (Fig. 4.) is actually facilitated in the (F1,F2) factorial plane where near 64 % of the initial information is explained and in which time series and variables are well enough represented (sums of squared cosinus near 0.70). Also, we can notice data coherence with regular and polygonal lines following 3 groups of processes and their respective modalities of entropy. The method and the choice of the three modalities seem to be pertinent.
It results that variables and classes of time series can be joined (their contributions to axes build are nearly 0.75).

The F1 axis, with especially the conditional entropies of order 2 to 5 (modalities 1 and 3) and the residual entropy of order 1 (modalities 3 and 1) separates the class of negative AR(1) and positive MA(1), against the class of negative MA(1) and positive AR(1). The F2 axis opposes on its positive part the 'robust to semi-robust' processes, against on its negative part the 'weak' processes, with principally the conditional entropies of order 2 to 5 (modality 2) and the residual entropy of order 1 (modality 2).

The (Fig. 4.) illustrates the second reference graphic model where all the classes of 'robust', 'semi-robust' and 'weak' time series are explained by the modalities of incertitude measures and reduction of incertitude.

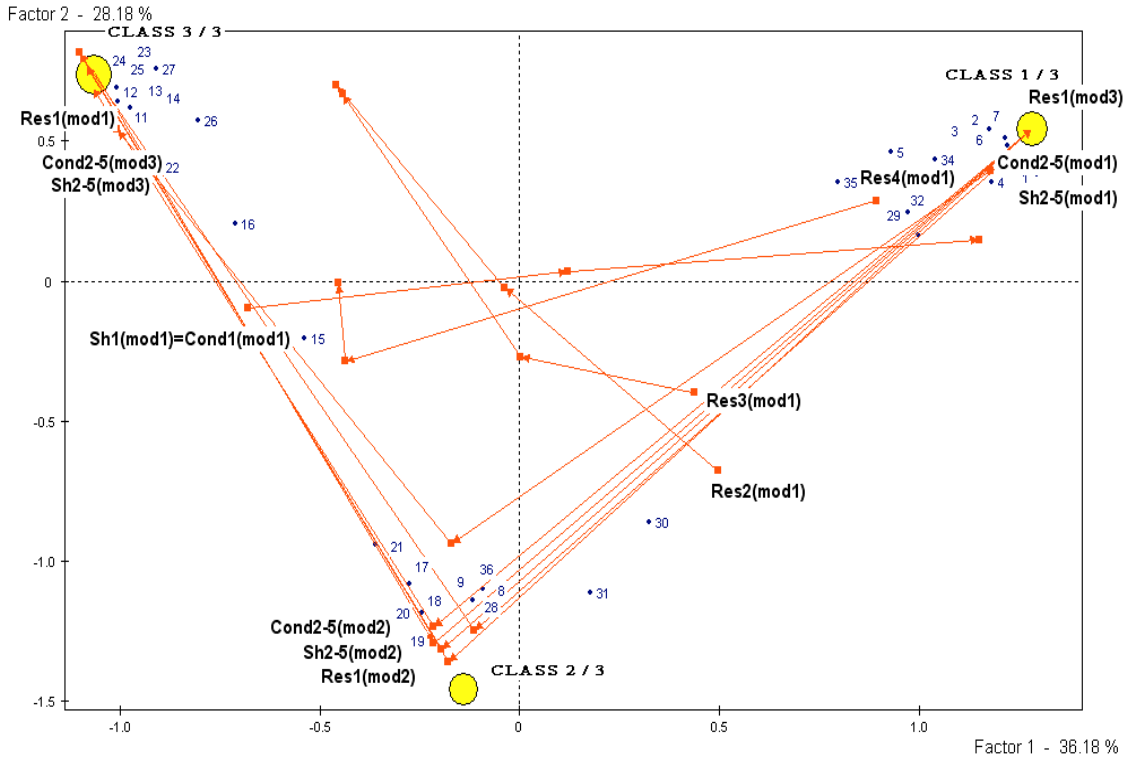


Fig. 4. The second reference scores model.

4. Conclusion

The method with the approaches by the autocorrelations and structural is justified for the stationary AR(1) and MA(1) time series. The both reference graphic models are too complementary : they permit on the one hand, the identification and the estimation of 'robust to semi-robust' processes, and on the other hand, a characterization by measures of incertitude of classes of 'robust', 'semi-robust' and 'weak' processes.

The method must be extended to usual stationary ARMA processes like the AR(2), MA(2) and ARMA(1,1). Also, if the method is adjusted on simulated processes, it's destinated, by nature, to be applied to operational data, like for instance financial data with the structural approach.

References

Box, G.E.P. and Jenkins, G.M. (1976), *"Time Series Analysis, Forecasting and Control"*, Holden Day, San Francisco.

Harding, D. (2003), "Using turning point information to study economic dynamics", The University of Melbourne.

Kendall, M. and Stuart, A. (1976), *"The advanced theory of statistics"*, Volume 3: design and analysis, and time-series,. Charles Griffin and Co Ltd, London.

Shannon, C.E. (1948), *"A Mathematical Theory of communication"*, Bell Syst. Tech. J. 27, pp379-423.

Wilder, G.W. (1978), *"New Concept in Technical Trading System"*, Greensboro NC: Trend Research.

Classification de processus ARMA basée sur l'analyse structurelle.

Carole TOQUE

Ecole Nat. Sup. des Télécommunications
46, rue Barrault,
75634 Paris Cedex 13
Email : carole.toque@educagri.fr

RÉSUMÉ.

Pour l'identification des processus ARMA, la méthode proposée combine à la fois une approche structurelle par analyse des points de retournements et de critères d'information avec des techniques de classification (classification ascendante hiérarchique (CAH)) et factorielle (analyse des correspondances multiples (ACM)). La méthode est appliquée à des séries AR(1) et MA(1) simulées.

MOTS-CLÉS :

Analyse structurelle, processus ARMA, identification de processus, théorie de l'information, entropie, mesure d'incertitude, analyse des correspondances multiples, classification hiérarchique ascendante.

1 Introduction

Pour la prévision des séries chronologiques, la méthodologie historique de Box et Jenkins [BOX 76], basée essentiellement sur l'examen des fonctions d'autocorrélation (FAC) et d'autocorrélation partielle (FAP), reste toujours d'actualité. Cependant, l'étape d'identification de la chronique échantillon à la classe des processus ARMA linéaires et stationnaires, est délicate et un peu trop restrictive. Le recours à d'autres techniques peut donc se justifier lorsque se combinent par exemple des changements structurels dont on veut mesurer la puissance (reprise (ou pic) et essoufflement (ou creux)).
On propose alors une méthode qui utilise à la fois une approche structurelle avec l'analyse des points de retournement et la théorie de l'information, et une approche par des techniques de classification.
Une classification est le résultat d'une succession de choix : le choix de la matrice des données initiales (quelles sont les variables à utiliser ?), le choix de la distance, le choix de la méthode de classification et le choix de la méthode d'agrégation pour une classification hiérarchique.
Précisément, les matrices initiales étudiées sont successivement la matrice temporelle issue de la simulation de processus AR(1) et MA(1), la matrice des points de retournements et la matrice des mesures 'entropiques' des séries simulées. Enfin, la méthode de classification retenue est la méthode hiérarchique ascendante avec le plus souvent la distance euclidienne ou la distance binaire pour la matrice des vecteurs d'état (0) ou (1), et le critère du lien complet d'agrégation, à l'exception de la CAH sur facteurs d'une ACM pour laquelle le critère de Ward est utilisé.

2 «Séparabilité» de processus ARMA en classes

2.1 Les simulations de trajectoires et la matrice initiale temporelle

18 processus AR(1) et 18 processus MA(1) centrés et stationnaires, chacun de longueur 500, ont été générés à partir des valeurs de coefficients ϕ_1 et θ_1, comprises entre -0.9 et +0.9 et avec un pas de 0.1.

Par ailleurs, nous avons fixé comme valeurs des paramètres du modèle : la variance du bruit ($\sigma_u^{\,2} = 90$) et une valeur de calage générateur aléatoire, d'un bruit dont la loi est posée gaussienne et centrée.
La matrice temporelle à analyser est donc de dimension (36x500) et les observations sont les processus AR(1) de coefficients -0.9 à -0.1 numérotés de 1 à 9 puis de coefficients +0.1 à +0.9 numérotés de 19 à 27, et les processus MA(1) de coefficients -0.9 à -0.1 numérotés de 10 à 18 puis de coefficients +0.1 à +0.9 numérotés de 28 à 36.

2.2 Classification hiérarchique sur la matrice temporelle

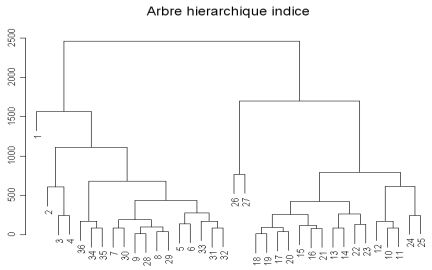


Fig. 1. Dendrogramme

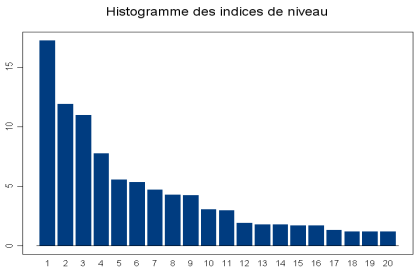


Fig. 2. Histogramme des indices de niveau

La CAH obtenue avec la distance euclidienne et le lien complet d'agrégation, est représentée par le dendrogramme ci-dessus (Fig. 1.) et est complétée par le graphique des 'indices de niveaux par ordre décroissant' (Fig. 2.) pour le choix du nombre de classes.
On peut alors constater que l'arbre hiérarchique ne présente pas de 'forts' effets de chaînage : les individus se répartissent en classes. Cette structure globale de l'arbre est déjà un bon indicateur de la 'séparabilité' des processus AR(1) et MA(1) en classes.
Pour ce qui est du choix du niveau de coupure de l'arbre, on retient une séparation en 4 classes avec, d'une part le seul processus AR(1) à 'fort' coefficient (-0.9) en module, d'autre part les 2 processus AR(1) à 'forts' coefficients positifs (+0.8 et +0.9), puis la classe des AR(1) à coefficients négatifs et des MA(1) à coefficients positifs, et la classe des MA(1) à coefficients négatifs et des AR(1) à coefficients positifs. A l'exception des deux premières classes qui isolent les processus AR(1) à très 'forts' coefficients en module, on retrouve certaines propriétés de symétrie des comportements de la FAC d'un AR(1) et de la FAP d'un MA(1), et de la FAP d'un AR(1) et de la FAC d'un MA(1).
Ces premiers résultats montrent que la classification est un outil d'aide à l'identification de processus ARMA. Cependant, il manque la description des classes pour renforcer la méthode : on propose alors de recourir à l'analyse structurelle.

3 Identification structurelle de processus ARMA

3.1 L'analyse structurelle et les mesures d'incertitude

Pour décrire et mesurer les changements structurels d'une série temporelle, on choisit de transformer la série initiale en une série de points de retournements ou série d'états. On construit par différences premières des données, une série de symboles (0) ou (1) correspondant respectivement aux 'pics' ou aux 'creux' de la série initiale (Kendall et Stuart [KEN 76]). Puis, on mesure les fréquences des symboles et séquences de k symboles sur chaque série d'états pour estimer les différentes probabilités.
Cependant, comment qualifier l'information dont on dispose sur ces probabilités ? Ou encore, comment qualifier l'incertitude ? C'est bien sûr la théorie de l'information qui intervient avec les mesures entropiques de divers ordres (Shannon [SHA 48] et, Yaglom et Yaglom [YAG 59]).

Un projet développé en Fortran 90 puis intégré à 'Splus' permet de construire les séries d'états, d'estimer les probabilités $P_k = \{ p_{k,1}, p_{k,2}..., p_{k,m^k} \}$ des m^k k-uplets d'états d'une série $\{Z_t\}$, de calculer les entropies simples H_k d'ordre k (notées aussi 'Shk') définies par $H_k(P_k) = -\sum_{i=1}^{m^k} p_{k,i} \log_2 p_{k,i}$, de calculer les entropies conditionnelles h_k d'ordre k (notées aussi 'Condk') définies par $h_1 = H_1(P_1)$ sinon $h_k = H_k(P_k / P_{k-1}) = H_k(P_k) - H_{k-1}(P_{k-1})$, et de calculer les entropies résiduelles d_k d'ordre k (notées aussi 'Resk') définies par $d_k = h_k - h_{k+1}$, pour k allant de 1 à q ou (q-1) pour d_k.

Par exemple, l'entropie d_k s'interprète comme la réduction moyenne d'incertitude sur un symbole selon qu'on connaît le k-gramme précédent plutôt que le $(k-1)$–gramme.
Enfin, le projet est appliqué aux séries AR(1) et MA(1) déjà simulées, avec comme valeur de paramètre q=5 pour les calculs des fréquences et des entropies. Il en résulte pour chaque processus : un vecteur d'états de (0) et de (1) et un vecteur des entropies de dimension (14).

3.2 Classification sur la matrice des points de retournements

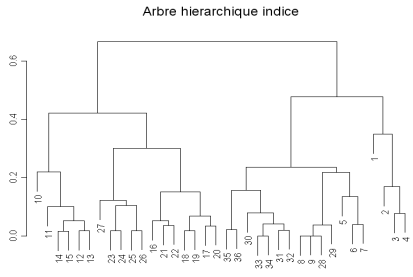


Fig. 3. Dendrogramme

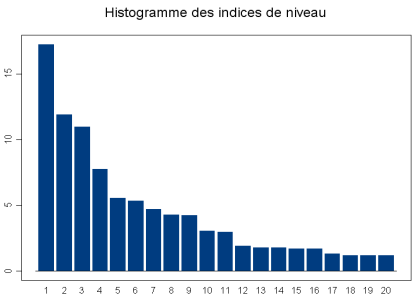


Fig. 4. Histogramme des indices de niveau

Pour cette première CAH ‘structurelle’, la matrice initiale des données est la matrice des 36 vecteurs de (0) et (1), la métrique (pour le calcul de la matrice des distances) est la métrique ‘binaire’ (la distance entre deux vecteurs lignes est le nombre d’occurrences de (01) ou de (10) divisé par le nombre de colonnes où au moins un de ces individus a un (1)) et la méthode d’agrégation est le lien complet.
L’arbre en Fig.3. présente des classes plus compactes avec moins d’effets de chaînage que dans la classification obtenue sur la matrice temporelle. Aussi, des classes de processus à coefficients ‘faibles à semiforts’ en module se différencient de classes de processus à coefficients ‘forts’ en module. En effet, pour une séparation en 6 classes, on distingue par exemple, des classes de processus dits ‘faibles à semi-forts’ (5-6-7-8-9 et 28-29) et (16-17-18 et 19-20-21-22) qui reflètent les résultats de la symétrie des comportements des fonctions d’autocorrélations, et une classe de processus dits ‘forts’ (1-2-3-4).
La méthode par l’analyse structurelle et la classification est encourageante. Pour faciliter l’interprétation des classes et pour faire apparaître des classes encore plus compactes, on choisit de construire une ACM à 3 modalités sur les vecteurs d’entropie.

3.3 Classification sur composantes principales d’une ACM ‘structurelle’

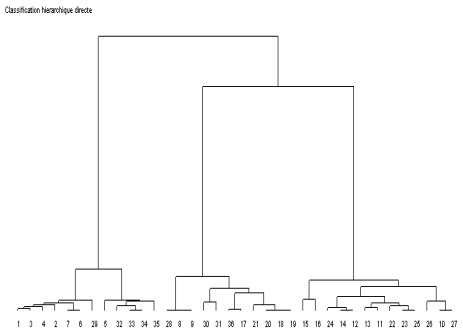


Fig. 5. Dendrogramme

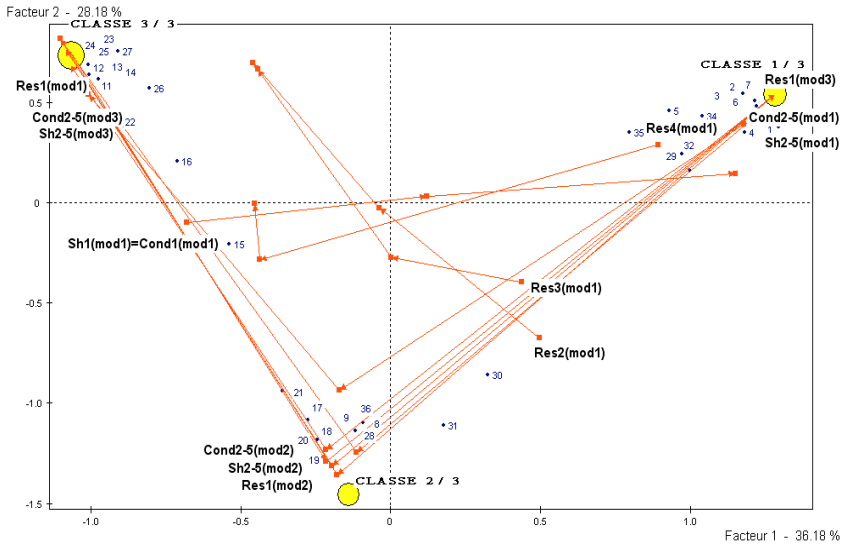


Fig. 6. Représentation des classes et trajectoires

Une classification réalisée après une ACM permet d'illustrer les graphiques de projection des variables (en nombre réduit) et des individus en y ajoutant les classes obtenues à l'issue de la CAH (Fig. 5.).
La lecture graphique (Fig. 6.) est en effet facilitée dans ce plan factoriel (F1,F2) qui explique à lui seul près de 64% de l'inertie totale et dans lequel les individus et caractères sont assez bien représentés ('sommes de carré de cosinus' souvent proches de 0.70).
On peut aussi mettre en évidence une non-linéarité entre les différents critères et traduire une éventuelle progression en reliant leurs modalités respectives.
On constate alors une certaine cohérence des données avec la présence de nombreuses lignes 'polygonales' régulières qui suivent 3 classes d'individus. La méthode et le choix du nombre des modalités semblent pertinents.
Il en résulte les rapprochements des variables et des classes d'individus dont les contributions à la formation des axes sont souvent voisines de 0.75. L'axe F1, avec surtout les entropies conditionnelles d'ordre 2 à 5 (modalités 1 et 3) et l'entropie résiduelle d'ordre 1 (modalités 3 et 1) sépare la classe des AR(1) à coefficients négatifs et des MA(1) à coefficients positifs, de la classe des MA(1) à coefficients négatifs et des AR(1) à coefficients positifs. L'axe F2 oppose du côté positif les processus dits 'forts' à 'semi-forts' aux processus dits 'faibles' du côté négatif, avec surtout des entropies conditionnelles d'ordre 2 à 5 (modalité 2) et une entropie résiduelle d'ordre 1 (modalité 2).
Cette ACM sur les entropies, complétée par une CAH, fait donc apparaître explicitement les modalités de mesures d'incertitude et de réduction d'incertitude pour caractériser aussi bien les 2 classes de processus dits 'forts' à 'semi-forts' que la classe des processus dits 'faibles'.

4 Conclusion

La méthode avec les mesures d'incertitude et la classification hiérarchique est justifiée et peut s'étendre à tous les processus ARMA.
Cependant, tout comme le choix des variables à utiliser est essentiel, le choix de la distance l'est aussi.
Par exemple, la distance 'binaire' utilisée peut être améliorée en tenant compte des k-uplets d'états des séries de points de retournement pour k>2. Ou encore, au delà des méthodes de classification hiérarchique, il est possible de construire une partition non supervisée, basée sur l'entropie calculée à partir de fréquences de k-uplets d'états (ou points de retournements) d'une série quelconque
Des résultats bien meilleurs sont alors attendus en estimant les distances entre points, non plus avec la norme L^2 comme il a été fait, mais avec la norme L^1 en vue de la prévision de processus linéaires et non linéaires.

5 Bibliographie

[BOX 76] BOX G.E.P., JENKINS G.M., *Time Series Analysis, Forecasting and Control*, Holden Day, San Francisco, 1976.

[KEN 76] KENDALL M., STUART A., *The advanced theory of statistics (Volume 3 : design and analysis, and time-series)*, Charles Griffin and Co Ltd, London, 1976.

[SHA 48] SHANNON C.E., *A Mathematical Theory of communication*, Bell Syst. Tech. J. 27, 1948, pp379-423.

[YAG 59] YAGLOM A. M., YAGLOM I. M., *Probabilité et Information*, Dunod, Paris, 1959.

Identification of Time-series models: Application to ARMA processes

Carole Toque, Bernard Burtschy

Ecole Nat. Sup. des Télécommunications, 46, rue Barrault, 75634 Paris Cedex 13. Email : carole.toque@educagri.fr

In identification of time series, the proposed method combines the usual approach by autocorrelations and a structural approach, less usual, by analysis of oscillators and theory of information, through visualisation by factorial methods (principal component analyses PCA and multiple correspondences MCA). It supplies reference graphic models and pertinent criteria for, identification and estimation of models, and identification of classes. The method was applied to simulated and usual ARMA processes that are stationary and independent.

Based on simulated temporal matrices, first PCA produce good quality of processes representation, with significant groupings and oppositions reflecting the eigenvalues symmetric behaviour. However, the main groups of series preserve the samples or the variability of simulated white noises.

Directly based on autocorrelation matrices, PCA give better results except for some processes said “weak”. Groupings and oppositions of series are function of their coefficients whatever the samples : autocorrelations reduce white noises. The first reference graphic models ensue with identification and estimation.

Description and measure of possible structural change lead us to introduce oscillators, frequencies and measures of entropy. This is the structural approach.

To establish non-linearity between the numerous criteria and to increase the discriminative ability between the processes, classifications on MCA are built over measures of entropy and produce outstanding quality of classes’ characterization. The second reference graphic models ensue with the class of “weak” processes.

Therefore, we can project a series in the reference factorial graphics to define the type and estimate the coefficients of an ARMA model. The method with the approaches by the autocorrelations and structural is justified and can be extended to operational data.