HAL open science

Contribution à l'analyse de données temporelles

Ahlame Douzal-Chouakria

▶ **To cite this version:**

Ahlame Douzal-Chouakria. Contribution à l'analyse de données temporelles. Machine Learning [stat.ML]. Université Joseph-Fourier - Grenoble I, 2012. tel-00908426

HAL Id: tel-00908426

https://theses.hal.science/tel-00908426v1

Submitted on 22 Nov 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

Mémoire de synthèse

présenté en vue de l'obtention d'une

Habilitation à diriger des recherches

Spécialité informatique

par

Ahlame Douzal

Contributions à l'analyse de données temporelles

Soutenue le 29 Novembre 2012 devant le jury composé de :

Prof. Younès Bennani, Université Paris XIII	Examinateur
Prof. Francisco Decarvalho, Université Fédérale de Pernambuco	Rapporteur
Prof. Patrick Gallinari, Université Pierre et Marie Curie	Rapporteur
Prof. Eric Gaussier, Université Joseph Fourier	Président
Prof. Mohamed Nadif, Université Paris Descartes	Rapporteur
Prof. Gilbert Saporta , CNAM Paris	Rapporteur
Prof. Marc Sebban, Université Jean Monnet	Examinateur

Remerciements

Je souhaite remercier, tout d'abord, les rapporteurs de ce mémoire, Gilbert Saporta, Mohamed Nadif, Patrick Gallinari et Francisco De Carvalho pour le temps qu'ils ont consacré à sa lecture et les commentaires pertinents qu'ils ont fournis. Je voudrai également remercier Marc Sebban et Younès Bennani pour avoir accepté de faire partie du jury de mon habilitation. Je remercie Eric Gaussier pour le temps dédié à la lecture du mémoire et les remarques judicieuses faites.

Je remercie plus généralement tous les collègues et collaborateurs avec qui j'ai échangé des idées ou collaboré.

Enfin, mes derniers remerciements vont pour ma famille.

Table des matières

Remerciements **2**

Introduction **1**

I Représentation compacte de séries temporelles 7

1 Réduction de la dimension temporelle de séries multivariées **8**
1.1 Résumé 8
1.2 Variance-covariance locale vs. temporelle 8
1.3 Segmentation d'une série multivariée et propriétés associées 10
1.4 Réduction de la dimension temporelle sous contraintes de préservation des corrélations 11
1.5 Application et résultats 13
1.6 Conclusion 15

II Définition et apprentissage de métriques à partir de données temporelles 16

1 Apprentissage de couplages pour la discrimination de séries temporelles **17**
1.1 Résumé 17
1.2 Introduction 17
1.3 Variance/covariance généralisée à des données temporelles 18
1.4 Apprentissage de couplages discriminants 20
1.5 Applications et étude comparative 23
1.6 Conclusion 28

2 Mesures de proximité intégrant la forme des séries : application à des données d'expression de gènes **29**
2.1 Résumé 29
2.2 Introduction 29
2.3 Identification des gènes exprimés au cours du cycle cellulaire 30
2.4 Une formalisation unifiée des métriques pour séries temporelles 31
2.5 Classification/catégorisation de profils d'expression de gènes 34

2.6 Étude comparative fondée sur un modèle génératif de profils d'expression périodiques . . . 34
2.7 Conclusion . . . 40

III Classification de séries temporelles 41

1 Classification/régression de séries temporelles par arbres 42
1.1 Résumé . . . 42
1.2 Introduction . . . 42
1.3 Algorithme de construction de l'arbre de classification . . . 44
1.4 Applications . . . 46
1.5 Conclusion . . . 48

Perspectives 50

Annexe 55

Introduction

Ce document présente une synthèse de mes travaux de recherche dont l'objectif est l'étude de nouvelles approches et de techniques pour l'analyse et l'extraction de connaissances à partir de données temporelles. Ces travaux s'inscrivent principalement dans les domaines de l'analyse des données et de l'apprentissage automatique.

Par données temporelles, je désigne des données numériques évoluant dans le temps, dites communément *séries temporelles*, ou des suites chronologiques de données symboliques dites *séquences temporelles*. Plus généralement, on désigne par *données de séquences* toute collection de données ordonnées selon un critère qui peut être sémantique, biologique, temporel ou autre ; c'est le cas, par exemple, des séquences de mots dans un texte ; on parle alors d'ordre syntaxique, de séquences d'acides aminés composant une chaine d'ADN ou de peptides constituant une protéine.

Les données temporelles sont omniprésentes, elles constituent la catégorie de données de séquences la plus fréquemment rencontrée dans les applications. Les données temporelles apparaissent naturellement dans des applications émergentes visant à analyser, par exemple, le comportement des utilisateurs du web, l'évolution structurelle ou informationnelle au sein de réseaux sociaux, ou des données issues de réseaux de capteurs en vue d'appréhender et de contrôler les phénomènes dynamiques sous-jacents. Les données temporelles sont également impliquées dans des applications plus classiques telles que l'analyse de signaux décrivant la progression de paramètres physiologiques en médecine (IRM fonctionnelle, EEG, ECG), les profils d'expression de gènes, les courbes de charge de consommation d'énergie ou l'évolution des cours boursiers.

Mes travaux de recherche s'inscrivent dans le cadre de processus d'apprentissage et d'analyse de données temporelles, dont l'objectif se ramène au choix fondamental de l'espace de description et du modèle de représentation des données, de la définition de métriques ou de mesures statistiques intégrant l'information d'interdépendance souvent multidimensionnelle, riche et complexe, pour la classification, la prédiction ou le contrôle des données temporelles.

Analyse de données temporelles

Représentation des données temporelles Les travaux portant sur la représentation des données temporelles foisonnent dans la littérature. On compte plusieurs approches motivées par des objectifs d'analyse divers émanant de domaines variés.
Sans être exhaustif, on distingue les approches issues de la statistique ou du traitement

du signal dont l'objectif est la projection des séries temporelles dans de nouveaux espaces définis par des descripteurs statiques. Ces projections correspondent, par exemple, à des transformations de Fourier, par ondelettes, ou à des décompositions polynomiales ([12], [67], [149], [92], [119]) des séries. Les séries temporelles ainsi décrites peuvent être analysées par des approches conventionnelles dédiées aux données statiques.

Notons les approches dites de *segmentation*, qui contrairement aux approches précédentes, préservent la dimension temporelle des données. Les objectifs des techniques de segmentation sont multiples ; ils sont utilisés, par exemple, comme moyen de réduction de la dimension temporelle des données, pour l'extraction de sous-séquences ou de motifs saillants, ou pour le filtrage de données bruitées ([66], [78], [132], [60], [59], [31]) . Les méthodes de segmentation peuvent également être utilisées en vue du codage des séries par des séquences temporelles, par des processus d'agrégation et d'abstraction et leurs analyses par des approches appliquées à des données de séquences.

Considérons, en particulier, l'espace de représentation adopté pour la prédiction des séries temporelles. Il s'appuie sur le théorème de Takens [133], énonçant la possibilité de reconstruire efficacement un système dynamique via son espace des phases. Par exemple, en segmentant une série $y(1), ..., y(N)$ en un ensemble de vecteurs $x(k) = [y(k), y(k - \tau), ..., y(k - (m-1)\tau)]$ où la dimension de l'espace de projection m correspond au nombre d'observations passées prises en compte pour la prédiction des valeurs futures à l'horizon de prédiction τ, la problème de prédiction se ramène à un problème d'estimation de la fonction $f : x(k) \rightarrow y(k+1)$.

Métriques associées aux données temporelles L'apprentissage de métriques est au cœur des techniques d'apprentissage et d'analyse de données. Les données temporelles introduisent, au niveau des métriques, une complexité supplémentaire liée à l'interdépendance des données. De nombreux travaux portant sur les mesures de proximités entre des séries temporelles ont été proposés cette dernière décennie ; ils s'articulent essentiellement autour de l'apprentissage de trois composantes :
a) les fonctions d'alignements des données de séquences intégrant, entre autres, des contraintes d'ordre temporelles, la prise en compte de variations de fréquences, ou l'insertion de sauts (Yu et al. [167], Ratanamahatana et al. [139]),
b) les fonctions de coût entre les données alignées, permettant de prendre en compte, par exemple, des variations d'amplitude, ou l'intégration du voisinage temporel des données alignées (Xie and Wiltgen [164],),
c) les fonctions de pondération des périodes d'observation, permettant de focaliser les mesures sur des caractéristiques locales ou globales des séries (Gaudin and Nicoloyannis [68], Jeong et al. [88]).

Notons également les nombreux travaux portant sur des méthodes à noyaux pour des analyses non linéaires de données de séquences souvent complexes. Une fonction noyau définit à la fois une mesure de similarité (un produit scalaire) et un espace de description souvent de grande dimension et implicite. Dans cet espace, les données projetées ne sont pas manipulées explicitement mais via leur fonction noyau. Grace aux fonctions noyaux, toutes les approches linéaires, basées sur la définition d'un produit scalaire, peuvent être mises en

œuvre pour estimer, dans l'espace de projection, des régularités et fonctions linéaires correspondant à des régularités et des fonctions non linéaires dans l'espace d'origine.

Les méthodes à noyaux [147], [32], [151] se sont révélées efficaces pour l'analyse de données structurées, telles que des images (Cuturi et al. [34], Harchaoui et al. [81]), des graphes (Shervashidze et al. [150]), du texte (Moschitti et al. [125]), ou des séquences (Cuturi et al. [35], Sonnenburg et al. [152]). Comparativement, les travaux sur les séries temporelles occupent une faible part dans la littérature sur les noyaux.

Les principales propositions de noyaux pour des séries temporelles sont obtenues par régularisations de la dynamique time warping (DTW). C'est le cas, par exemple, des travaux de Cuturi [33] et Cuturi et al. [36] qui considèrent le softmax des fonctions d'alignements assurant la propriété définie positive du noyau, ceux de Hayashi et al. [84] projetant les séries temporelles dans un espace euclidien définissant une distance entre les séries proche de celle de la DTW, ou encore la proposition de Kumara et al. [104] considérant une approche non paramétrique pour l'interpolation splines des séries temporelles, puis la définition d'un noyau sur les représentation interpolées obtenues.

Classification et prédiction de données temporelles Etant donné un espace de représentation des données et une métrique associée, la classification supervisée ou non supervisée des données temporelles repose principalement sur des paradigmes, modèles ou heuristiques similaires à ceux déployés sur des données statiques.

La prédiction constitue inversement un processus inhérent aux données évolutives et central en analyse de données temporelles. On peut organiser les nombreux travaux portant sur la prédiction de séries temporelles en deux grandes catégories. D'une part, on distingue les approches issues du domaine de la statistique fondées sur les modèles auto-régressifs (AR, ARMA, ARIMA) [16], [117], [120]. Ces méthodes reposent sur des processus convergents et peu coûteux pour des modèles à faibles ordres. Ces modèles se limitent, néanmoins, à des processus linéaires simples. L'utilisation de modèles à ordre supérieur pour la prise en compte de processus plus complexes peut s'avérer très coûteuse. Dans le même esprit, on distingue les filtres de Kalman [93] largement utilisés en automatique et en traitement du signal pour la prédiction, mais également pour le filtrage et le lissage des données. Ces modèles assurent la convergence et sont peu coûteux mais nécessitent la connaissance d'un modèle a priori, supposé linéaire et stationnaire.

En revanche, les données issues des systèmes dynamiques sont en général plus complexes, ils peuvent présenter des irrégularités, être générés par des systèmes déterministes non-linéaires ou chaotiques [94], [162].

La littérature foisonne de propositions pour la prédiction de données temporelles complexes. Citons, par exemple, les méthodes de Tong (1990) [157], Kantz et al. (2004) [94], et Fan et al. (2003) [58] permettant la prise en compte des changement de régimes au sein des séries, les méthodes auto-régressives exponentielles [120], [58] ou les modèles ARCH auto-régressifs à homoscédasticité conditionnelle [120], [17], [117], [58].

Dans une deuxième catégorie, on considère les approches issues de l'apprentissage pour la prédiction des séries temporelles. Ces approches reposent sur la description des séries dans un espace qui peut être explicite (par exemple l'espace des phases) ou implicite (dérivé par le choix d'un noyaux), visant à ramener le problème de la prédiction à un problème de classification ou plus précisément de régression. Nous citons, par exemple, les approches à base de perceptrons [154], [111], [74] ; ils présentent l'avantage de ne pas se référer à un modèle particulier, de ne pas présumer de la linéarité et stationnarité des données et peuvent être de coût raisonnable. Ils peuvent nécessiter, cependant, l'estimation souvent empirique d'un grand nombre de paramètres libres, ne garantissent pas la convergence vers une solution optimale et le processus d'apprentissage peut être assez coûteux. Enfin, un grand nombre de travaux est dédié à la prédiction de séries temporelles par séparateurs à vastes marges (SVM/SVR) [128], [127], [13]. Ces modèles ont l'avantage également de ne pas dépendre d'un modèle a priori, ils garantissent la convergence vers une solution optimale, tout en ne nécessitant que l'estimation, à faible coût, d'un petit nombre de paramètres libres. L'un des défis pour ces approches demeure l'estimation empirique des paramètres libres, pouvant engendrer des coûts très élevés en phase d'apprentissage.

Principales contributions

Mes travaux de recherche portent sur l'analyse de séries temporelles, bien que certaines de nos propositions puissent être généralisées à des données de séquences. Mes principales contributions s'articulent en trois parties : -la représentation des séries temporelles, -la définition de métriques et leur apprentissage, -ainsi que la proposition de nouvelles approches de classification dédiées aux séries temporelles.

Représentation compacte de séries temporelles (Partie I)

L'étude de différentes statistiques d'autocorrélation spatiale (Moran [122, 123], Geary [69], Getis [70]) et leurs applications à des structures de contiguïté particulières comme celle induite par les séries temporelles, offrent des propriétés intéressantes permettant d'appréhender le comportement des séries (comportement aléatoire, chaotique), d'évaluer le niveau de saillance d'un événement, ou de mesurer la dépendance entre une structure a priori (en l'occurrence temporelle) et les observations. Ces propriétés ont inspiré mes deux premiers travaux portant sur la réduction de la dimension temporelle et la proposition de mesures capturant la forme des séries. Ainsi, dans ma première contribution, résumée au chapitre 1 de la première partie, je propose une méthode de réduction de la dimension temporelle de séries multivariées par segmentation. L'approche proposée [28] vise à décrire une séries multivariée par un nombre minimal de points préservant l'information de variance/covariance de la série. Notons que, classiquement, les approches de segmentation s'adressent à des séries univariées, dont le nombre de segments est connu a priori ou estimé [82], [166], [105], [106]. Ce travail a été mis en application sur des données en anesthésie-réanimation, fournies par l'équipe PRETA du TIMC-IMAG, où la grande dimension des séries constituait une limite à leur analyse.

Définition et apprentissage de métriques pour des séries temporelles (Partie II)

- *L'apprentissage de couplages pour la discrimination de séries temporelles (Chapitre 1)*
  Mon travail de recherche le plus récent porte sur l'apprentissage de couplages, induisant une métrique, pour la discrimination de classes de séries temporelles. Il n'est pas rare dans les applications que des séries temporelles d'une même classe soient de profils globaux dissimilaires, tout en partageant une signature locale commune dans la classe et pouvant apparaître à divers instants des séries ; ou, à l'inverse, que des séries de classes distinctes ne présentent que de faible différences. Motivés par la discrimination de telles séries complexes, nous proposons une approche pour l'apprentissage de couplages discriminants visant à connecter les séries d'une même classe selon les caractéristiques communes au sein des classes et différentielles entre les classes. L'approche proposée est guidée par la minimisation de la variance intra-classe et la maximisation de la variance inter-classe [61, 64, 63]. Parmi les résultats majeurs de ce travail, nous proposons : 1) une extension des stratégies d'alignements classiques à des couplages moins contraints temporellement, 2) la prise en compte lors de l'apprentissage des couplages de la dynamique de toutes les séries intra et inter classes, 3) une extension de l'expression usuelle de la variance/covariance à un ensemble de séries temporelles, ainsi qu'à des classes de séries, 4) l'apprentissage d'une métrique locale pondérée, restreignant la comparaison des séries aux attributs discriminants. Les résultats de ce travail sont résumés dans le chapitre 1 de la partie II du document. Ce travail s'inscrit dans le cadre de la thèse de C. Frambourg en co-direction avec J. Demongeot et en collaboration avec E. Gaussier.

- *Définition de métriques intégrant la composante forme des séries (Chapitre 2)*
  Une de mes contributions, initiée en collaboration avec P. Nagabhushan (Professeur à l'université de Mysore), s'inspire des études portant sur la variance/covariance locale, globale et l'analyse de la contiguïté (Geary [69], Lebart [107], Thioulouse et al. [155]). Nous avons proposé, dans un premier temps, une mesure de proximité capturant la forme des séries, puis avons étendu celle-ci à des mesures intégrant les composantes forme et valeurs [29, 30, 53]. Nous avons proposé, dans un second temps, un cadre unifié permettant de situer les principales familles de mesures de proximité classiques et proposées. Enfin, nous nous sommes intéressés à l'étude de l'efficacité des mesures introduites pour la classification de données complexes. Pour cela, nous avons considéré des données d'expression de gènes au cours du cycle cellulaire, décrivant des profils périodiques, pouvant inclure des variations d'amplitudes, des atténuations de phases ainsi que des effets de tendances. Cette étude a été réalisée dans le cadre de la thèse de A. Diallo en co-direction avec F. Giroud, biologiste en génomique au laboratoire TIMC-IMAG [50, 51].

Classification de données temporelles (Partie III)

Sur la base des métriques étudiées dans l'axe de recherche précédant, nous nous sommes intéressés à l'extension des arbres de classification/régression à des variables prédictives temporelles. L'arbre de classification temporel proposé est fondé sur un nouveau critère de coupure pour des variables de type séries temporelles. Ce critère de coupure, guidé par

la minimisation de l'erreur de Gini, est basé sur l'apprentissage du meilleur compromis forme-valeurs de la métrique considérée ainsi que de la localisation de sous-séquences discriminantes au niveau de chaque noeud [44]. Dans la partie III, je résume les principaux algorithmes d'inférence de l'arbre de classification temporel ainsi que les expérimentations conduites. Ce travail a été réalisé en collaboration avec C. Amblard de l'équipe AMA-LIG.

Organisation du mémoire

La suite du mémoire s'articule en trois parties. Dans la première partie, je présente l'approche de réduction de la dimension temporelle basée sur le concept de variance/covariance locale. La deuxième partie est composée de deux chapitres. Dans le chapitre 1, j'introduis une nouvelle approche d'apprentissage de couplages pour la discrimination de séries temporelles complexes. Dans le chapitre 2, je présente une nouvelle mesure de proximité fondée sur les composantes forme et valeurs des séries, la situe dans un cadre plus large composé de trois familles majeurs de métriques pour des séries temporelles, puis étudie ses performances sur des données réelles complexes issues de la biologie. Dans la troisième partie du mémoire, je présente une nouvelle méthode de classification/régression par arbre pour des séries temporelles, dont les performances sont confrontées sur les trois familles de mesures. Je conclue ce mémoire par les perspectives et travaux de recherche futurs.

Première partie

Représentation compacte de séries temporelles

Chapitre 1

Réduction de la dimension temporelle de séries multivariées

1.1 Résumé

L'application des méthodes d'analyse de données à des séries temporelles se trouve vite limitée face au nombre souvent très élevé des observations composant les séries.
Ce travail propose une nouvelle approche de réduction de la dimension temporelle (i.e., nombre d'observations) de séries multivariées par segmentation. L'approche proposée est fondée sur la détection d'un nombre inconnu de points de coupures, sous la contrainte de préservation de la structure de variance/covariance (ou des corrélations) ; information fondamentale à de nombreux processus en analyse de données.
Pour ce faire, j'introduis, tout d'abord, la notion de variance/covariance locale, puis temporelle. Je définis ensuite une fonction, basée sur la variance/covariance temporelle, associant à chaque segment sa contribution à l'inertie totale de la série qu'il compose. Les propriétés d'additivité et de monotonicité de la fonction "d'inertie" permettent la proposition d'un algorithme itératif de segmentation assurant un compris entre la réduction de dimension et la préservation de la structure de variance/covariance. La méthode proposée est illustrée par une application médicale portant sur des données de monitoring en anesthésie-réanimation.

1.2 Variance-covariance locale vs. temporelle

La nature ou l'origine du recueil de données suggèrent souvent une structure a priori de l'ensemble des observations. Cette structure définit des liens entre les observations de nature diverse (temporelle, spatiale, géographique, sémantique, ...). Analyser de telles données soulève au moins deux problèmes. D'une part, la mesure de la dépendance entre les observations et la structure a priori ; d'autre part, la prise en compte de cette structure dans le processus d'analyse. Nous nous intéressons ici à la notion de variance/covariance locale, une mesure centrale, impliquée dans de nombreux travaux visant à répondre aux deux objectifs cités. Les premières traces de la notion de variance/covariance locale remontent aux travaux de J.von Neumann [158, 159] développés dans un contexte d'observations liées temporelle-

ment. Des variantes de celle-ci ont été ensuite proposées. Sans être exhaustif, citons d'autres travaux majeurs et pionniers portant sur l'analyse de données structurées. Tout d'abord les travaux de Moran [122, 123] et de Gestis [70] proposant des indices d'autocorrélation spatiale, les travaux de Geary [69] développés dans le cadre de liens de type géographique, sans oublier l'analyse locale de structures de graphes proposée par Lebart [107, 7, 6], Wartenberg [161], ou l'analyse globale proposée par Thioulouse et al. 1995 [155]. Un exposé complet de l'analyse locale est introduit dans Lebart et al. 1973 [108], et diverses applications sont présentées dans Banet et al. 1984 [7]. Remarquons que ces travaux, souvent développés dans un contexte théorique, fournissent aujourd'hui un cadre riche d'approches et d'estimateurs permettant de mener des analyses efficaces et pertinentes de données structurées, au coeur de nombreuses applications telles que l'analyse des réseaux sociaux ou l'analyse de données issues de réseaux de capteurs.

Variance/Covariance locale On note X la matrice $(n \times p)$ donnant la description de N individus par p variables $X_1, ..., X_p$, et x_{ij} la valeur prise par la variable X_j pour l'individu i. On suppose défini une structure a priori sur l'ensemble des individus (par exemple, un réseau social connectant un ensemble d'utilisateurs).
On définit la matrice de variance-covariance locale $V_L(v_{jl}^L)$ de dimension p associée aux N observations et tenant compte de la structure des individus :

$$v_{jl}^L = \frac{1}{2\,m} \sum_{i=1}^{N} \sum_{i' \in E_i} (x_{ij} - x_{i'j})(x_{il} - x_{i'l}) \tag{1.1}$$

où $E_i \subset \{1, ..., N\}$ définit l'ensemble des voisins de i et $m = \sum_{i=1}^{N} Card(E_i)$.

Il est aisé de montrer que, dans le cas particulier où toutes les observations sont liées (i.e., la structure a priori définit un graphe complet), l'expression 1.1 de la variance-covariance locale correspond à la variance/covariance classique, dont l'expression moins usitée implique les différences entre couples d'observations :

$$v_{jl}^L = \frac{1}{2\,N\,(N-1)} \sum_{i=1}^{N} \sum_{i'=1}^{N} (x_{ij} - x_{i'j})(x_{il} - x_{i'l})$$

Notons que la variance-covariance locale rapportée à la variance-covariance classique permet de mesurer la dépendance entre les observations et la structure *a priori*. Dans le cas de données indépendantes de la structure a priori, variance/covariance locale et classique sont confondues ; dans le cas contraire, la variance/covariance classique surestime la variance locale. Notons, par exemple, que l'indice de Geary n'est autre que le rapport entre ces deux variances. Une valeur de l'indice de Geary proche de 0 correspond à une situation de dépendance forte entre les observations et la structure *a priori*, autrement dit la variabilité des observations traduit la structure sous-jacente, à l'inverse une valeur de l'indice de Geary proche de 1 indique une situation d'observations indépendantes de la structure *a priori*.

Variance/Covariance temporelle Considérons dans ce qui suit le cas particulier d'une matrice X donnant la description d'une série multivariée S composée de N observations effectuées aux instants $t_1, ..., t_N$. On définit le voisinage temporel d'ordre $k \geq 1$ d'un individu i par l'ensemble des observations effectuées dans la fenêtre temporelle $[t_{i-k}, t_{i+k}]$. On se

limite dans ce qui suit à la variance/covariance temporelle $V_T(S)$ d'ordre $k = 1$ de terme général $v_{jl}^T(S)$:

$$v_{jl}^T(S) \quad = \quad \frac{1}{4\,(N-1)} \sum_{i=1}^{N-1} (x_{(i+1)j} - x_{ij})(x_{(i+1)l} - x_{il}) \tag{1.2}$$

1.3 Segmentation d'une série multivariée et propriétés associées

La technique de réduction de la dimension temporelle que nous avons proposée est fondée sur la définition d'un opérateur de segmentation, de la fonction "d'inertie" ainsi que ses propriétés d'additivité et de monotonicité.

Définition 1 *On définit $S = S_1 \oplus S_2 \oplus ... \oplus S_k$, la segmentation en k segments d'une série $S = (1, ..., N)$ composée de N observations :*

$$S_1 = (n_0, ..., n_1),\ S_2 = (n_1, ..., n_2),\ ...\ , S_k = (n_{k-1}, ..., n_k)$$

avec $n_0 = 1$, $n_k = N$, $i \in \{1, ..., k\}$ et $n_i \in \{1, .., N\}$. On note $L_i = n_i - n_{i-1} + 1$ le nombre d'observations du segment S_i.

On note $I(S) = Tr(V_T(S))$ l'inertie d'une série multivariée définie par la trace de sa matrice de variance/covariance temporelle. On note $P(S)$ l'ensemble des segments de S.

Définition 2 *On définit la fonction C_S mesurant la contribution du segment S_i de $P(S)$ à l'inertie totale de S :*

$$\begin{aligned} C_S : P(S) \quad &\longrightarrow \quad R^+ \\ S_i \quad &\longrightarrow \quad C_S(S_i) = \frac{L_i - 1}{N-1}\, Tr(V_T(S_i)) \end{aligned} \tag{1.3}$$

avec $C_S(S) = Tr(V_T(S))$.

Définition 3 *On définit la relation d'ordre $\leq$ dans $P(S)$ comme suit :*

$$\forall\ S_i, S_j \in P(S) \quad S_i \leq S_j \quad \Leftrightarrow \quad \exists\ S_l \in P(S) \text{ tel que } S_j = S_i \oplus S_l.$$

Propriété 1 *La variance/covariance temporelle de $S = S_1 \oplus S_2 \oplus ... \oplus S_k$, est égale à la moyenne pondérée des matrices de variance/covariance des segments $S_1, ..., S_k$:*

$$V_T(S) = V_T(S_1 \oplus ... \oplus S_k) \quad = \quad \sum_{i=1}^{k} \frac{L_i - 1}{N-1}\, V_T(S_i) \tag{1.4}$$

Propriété 2 *La fonction C_S est additive par rapport à l'opérateur $\oplus$:*

$$C_S(S) = C_S(S_1 \oplus ... \oplus S_k) \quad = \quad \sum_{i=1}^{k} C_S(S_i) \tag{1.5}$$

Propriété 3 *La fonction C_S est monotone croissante par rapport à l'opérateur $\oplus$.*

$$\forall\ S_i,\ S_j \in P(S)\ \ S_i \leq S_j \quad \Rightarrow \quad C_S(S_i) \leq C_S(S_j) \tag{1.6}$$

1.4 Réduction de la dimension temporelle sous contraintes de préservation des corrélations

Toute méthode visant à représenter une série temporelle par un nombre d'observations plus faible engendre une perte d'information. On s'intéresse ici à l'information de variance-covariance, et plus précisément aux corrélations entre les descripteurs de la série.
Nous proposons une méthode de segmentation visant à décrire une série multivariée par un nombre minimal d'observations préservant un taux fixé a priori des corrélations initiales.
L'approche proposée est un algorithme itératif s'articulant en trois étapes. Dans la première étape, on procède à la décomposition de la série S en segments d'inertie minimale α_{C_S}. α_{C_S} est initialisé arbitrairement, puis ajusté, à chaque itération, en fonction de l'erreur d'approximation des corrélations induite par la segmentation. Dans la seconde étape, chaque segment (sous-série multivariée) issue de la segmentation est approchée par régression linéaire multiple ; le nombre d'observations de chaque segment est alors réduit à 2. Dans la troisième étape, on évalue l'erreur d'approximation des corrélations induite par la segmentation, l'erreur d'estimation des corrélations permet de réajuster la contribution minimale α_{C_S}. On réitère les trois étapes précédentes, jusqu'à stabilisation du seuil minimal α_{C_S} et de la segmentation, assurant une réduction maximale du nombre d'observations préservant un taux α_p des corrélations initiales. Explicitons dans ce qui suit les principales étapes de l'algorithme.

Initialisation On note $\alpha_p \in [0,1]$ la perte des corrélations fixée a priori. On initialise la valeur de $\alpha_{C_S} \in [0,1]$, en pratique cette valeur est fixée à 0.01, ce qui correspond à une contribution des segments extraits à hauteur de 1% de l'inertie totale de S. Notons que le nombre de segments extraits est inversement proportionnel à α_{C_S}. On évalue la matrice $V_T(S)$, ce qui peut être effectué de manière incrémentale grâce à la propriété 1 (1.4).

Segmentation L'étape de segmentation consiste à décomposer la série $S = S_1 \oplus S_2 \oplus ... \oplus S_k$ en k segments (k inconnu) tel que chaque segment S_i ait une inertie minimale $C_S(S_i) = \alpha_{C_S}.C_S(S)$. On note $S_1 = (n_0, ..., n_1)$, $S_2 = (n_1, ..., n_2)$, ... , $S_k = (n_{k-1}, ..., n_k)$ les k segments, avec $n_0 = 1$, $n_k = N$, $i \in \{1, ..., k\}$ et $n_i \in \{1, .., N\}$.

Pour l'extraction du segment S_i, on note S_i^r le segment portant sur les r observations $n_{i-1}, ..., n_{i-1} + r$. De part la propriété de monotonicité et de croissance de C_S (Eq. 1.6), on déduit :

$$\forall r \in [0, N] \;\; S_i^{r-1} \subset S_i^r \quad \Rightarrow \quad C_S(S_i^{r-1}) \leq C_S(S_i^r)$$

Ainsi, l'extraction du segment S_i consiste à rechercher la valeur $r*$ vérifiant la condition de coupure suivante :

$$C_S(S_i^{r*}) \leq \quad \alpha_{C_S} C_S(S) \quad < C_S(S_i^{r*+1})$$

Après l'extraction de la sous-séquence $S_i = S_i^{r*}$, on procède à l'extraction des segments suivants S_{i+1}, S_{i+2}..., jusqu' à la segmentation totale de S. La Figure 1.1 illustre, pour une série (Figure du haut), le processus de segmentation fondé sur l'extraction de segments de contribution minimale α_{C_S}, la figure du bas montre, pour chacun des segments extraits, la croissance de la fonction $C_S(S_i^r)$ jusqu'à l'atteinte du seuil de coupure $\alpha_{C_S}.C_S(S)$.

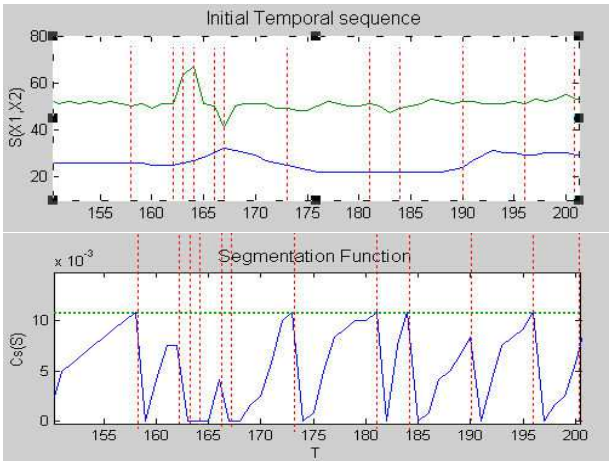

FIGURE 1.1 – Segmentation d'une série S par extraction de segments d'inertie minimale $\alpha_{C_S}.C_S(S)$

Régression linéaire multiple Pour chaque sous-série multivariée S_i obtenue on procède à une régression linéaire multiple afin de déterminer le segment approchant, au sens des moindres carrés, les L_i observations. La sous séquence S_i est décrite par les deux observations associées aux instants $t_{n_{i-1}}$ et t_{n_i} délimitant le segment approché (un plus grand nombre d'observations aurait pu être sélectionner pour caractériser le segment extrait).

Estimation de la perte de corrélation et ajustement de α_{C_S} On note S^l la série réduite de S obtenue à l'itération l. On note $Cor(S^l)$, de terme général $c_{ij}(S^l)$ et $Cor(S)$ de terme général $c_{ij}(S)$, les matrices des corrélations associées respectivement à S^l et S. On évalue la perte de corrélation suite à la réduction de la dimension temporelle de S à :

$$e(S, S^l) \quad = \quad \frac{2}{p(p-1)} \sum_{i=1}^{p} \sum_{j=i+1}^{p} \mid c_{ij}(S) - c_{ij}(S^l) \mid$$

avec $e(S, S^l) \in [0, 1]$. On ajuste le taux α_{C_S} proportionnellement à la marge d'erreur :

$$\alpha_{C_S} = \frac{\alpha_p}{e(S, S^l)}.\alpha_{C_S}$$

Cette ajustement va avoir deux effet distincts :

- Si $e(S, S^l) \leq \alpha_p$, le taux α_{C_S} est alors augmenté proportionnellement à la marge d'erreur, en vue de maximiser la réduction du nombre d'observations. Les étapes de segmentation et régression sont réitérées avec le nouveau taux α_{C_S} et ce jusqu'à la satisfaction de la condition d'arrêt :

$$e(S, S^{l*}) \leq \quad \alpha_p \quad < e(S, S^{l*+1})$$

- Si $e(S, S^l) > \alpha_p$: le taux α_{C_S} est diminué proportionnellement à la marge d'erreur, en vue de minimiser la perte de corrélations. De manière similaire, les étapes de segmentation et régression sont réitérées avec le nouveau taux α_{C_S} et ce jusqu'à la satisfaction

de la condition d'arrêt :

$$e(S, S^{l*}) \leq \quad \alpha_p \quad < e(S, S^{l*-1})$$

Dans les deux cas ci-dessus, la série réduite S^{l*} obtenue à l'itération $l*$ assure une réduction maximale du nombre d'observations apar rapport à la perte de corrélations α_p fixé a priori.

1.5 Application et résultats

L'approche proposée a été appliquée à des données de monitoring observant l'état de patients en unité intensive en anesthésie-réanimation. Les données d'un patient constituent une série temporelle multivariée (Figure 1.2) décrivant l'évolution de 11 paramètres physiologiques échantillonnés régulièrement sur 5269 instants. La réduction du nombre d'observa-

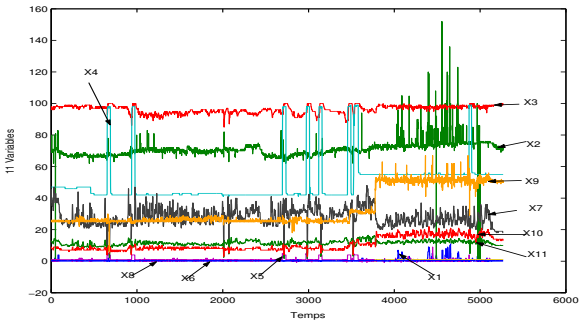

FIGURE 1.2 – Série multivariée longue décrivant l'évolution de 11 variables physiologiques, régulièrement échantillonnées sur 5269 instants.

tions de la série à un taux de préservation des corrélations à 80% ($\alpha_p = 0.2$) est illustré dans la Table 1.1. On y indique l'évolution de différents indicateurs le long des itérations de l'algorithme : le seuil de contribution minimal des segments à l'inertie totale $\alpha_{C(S)}$, le nombre de segments, la perte de corrélations induite, la valeur $\alpha_{C(S)}$ ajustée à chaque début d'itération, ainsi que le temps d'exécution.

La réduction de la série multivariée est obtenue en 5 itérations. A la première itération, l'étape de segmentation extrait des segments de contribution minimale à l'inertie totale de 20% ($\alpha_{C(S)} = 0.2$) . Suite à l'étape de régression, la perte de corrélation estimée à $59,93\%$ est largement supérieure au taux $\alpha_p = 0.2$, induisant une diminution importante de la valeur de $\alpha_{C(S)} = \frac{0.2}{0.5937}\,0.2 = 0.0674$. La deuxième itération procède à la segmentation de la série avec la nouvelle valeur ajustée de $\alpha_{C(S)} = 0.0674$. Remarquons, que le taux de diminution de $\alpha_{C(S)}$ est de plus en plus faible à mesure que la perte de corrélation s'approche du seuil fixé a priori α_p. L'algorithme converge à la 5 ème itération avec une série réduite à 72 observations et une préservation des corrélations initiales à hauteur de $100 - 18,69 = 81,69\%$.

Dans une seconde étape, on propose d'appliquer la méthode de réduction à différents niveaux de perte des corrélations α_p allant de 10% à 70%. Chaque ligne du tableau 1.2 résume les résultats obtenus.

TABLE 1.1 – Réduction du nombre d'observations de 5269 à 72 avec une préservation de 80% des corrélations initiales ($\alpha_p = 0.2$)

Nb. itération	α_{CS} initial	Nb. Obs.	Perte de correlation	α_{CS} ajusté	Temps (s) Exec.
1	0.2	10	0.5937	0.0674	38.4690
2	0.0674	10	0.4196	0.0321	34.8280
3	0.0321	26	0.2492	0.0258	37.4370
4	0.0258	52	0.3164	0.0163	34.7190
5	0.0163	72	0.1831	–	38.7500

TABLE 1.2 – Réduction de la série multivariée à différents taux de perte des corrélations (α_p allant 10% to 70%

α_p	Final length	τ_{RD} (%)	Correlation loss estimate	Runtime (s)
0.1	370	92.97	0.09580	576.8430
0.2	72	98.63	0.1831	184.2030
0.3	26	99.50	0.2840	180.1400
0.4	20	99.62	0.3433	248.1870
0.5	8	99.84	0.4257	105.4070
0.6	4	99.92	0.5973	68.4840
0.7	4	99.92	0.6591	68.7650

Pour une perte maximale, par exemple, de 10%, on obtient une série temporelle représentée par uniquement 388 observations, soit un taux de réduction de 92.63% ($\tau_{RD} = 1 - \frac{Nb.Obs.final}{Nb.Obs.initial}$), la perte effective d'information est de 9.3% soit une conservation de 90.7% des corrélations initiales, le temps d'exécution total est de 515.46 secondes. L'algorithme converge en trois itérations avec une valeur finale de $\alpha^*_{C_S}$ de 0.0013. La figure 1.3 représente les séries réduites aux seuils de perte des corrélations de 10% et 20%. Remarquons que plus

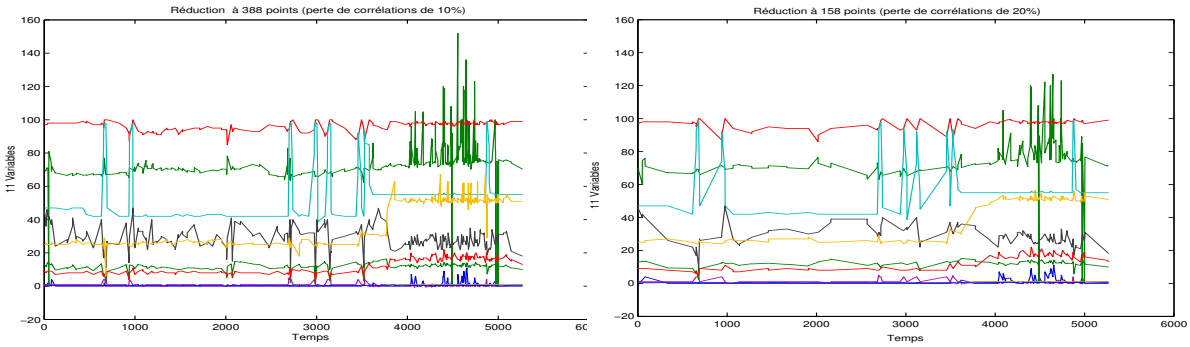

FIGURE 1.3 – $\alpha_p = 10\%$ $\alpha_p = 20\%$

α_p est élevé, plus fort sont le taux de réduction τ_{RD} et la perte de corrélations $e(S, S^{l*})$. La figure 1.4 représente la progression du rapport $\frac{\tau_{RD}}{e(S,S^{l*})}$ à différentes valeurs de α_p. Ainsi, la réduction d'une série peut être effectuée en fixant le taux de perte d'information maximale, ou en procédant à la réduction pour plusieurs valeurs de α_p puis en sélectionnant la réduction maximisant le rapport $\frac{\tau_{RD}}{e(S,S^{l*})}$. La figure 1.4 montre que la réduction qui maximise ce

rapport est obtenue pour un seuil de perte d'information de 10%, soit à un taux de réduction de dimension de 92.63% et à un taux de préservation de corrélations de 90.7%.

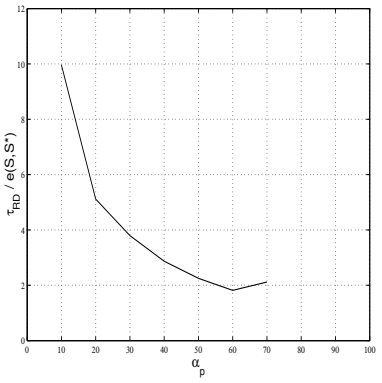

FIGURE 1.4 – Progression du rapport $\frac{\tau_{RD}}{e(S,S^{l*})}$ en fonction de α_p.

1.6 Conclusion

J'ai proposé une méthode de réduction du nombre d'observations de séries temporelles multivariées, par segmentation [28]. L'approche proposée est fondée sur la détection d'un nombre inconnu de points de coupures, sous la contrainte de préservation de la corrélation entre les descripteurs. Pour cela, j'ai introduit une fonction d'inertie dont les propriétés d'additivité et de monotonie, permettent la décomposition d'une série en un ensemble de segments saillants de contribution significative à l'inertie totale d'une série.
Ce travail a été mis en application sur des données en anesthésie-réanimation, où la grande dimension des séries constituait une limite à leur analyse. Ces données ont été fournies par A.S. Silvent de l'équipe SIC et en collaboration avec l'équipe PRETA du TIMC-IMAG.
Dans ce travail, j'ai exploré une propriété de la variance/covariance locale permettant de mesurer le niveau de saillance d'un événement dans une série. Dans le chapitre 2 de la partie II, je définis une mesurer de similarité basée sur la forme des séries, et montre que celle-ci correspond à une mesure de cross-corrélation locale associée à une structure de contiguïté en chaîne, induite par les séries temporelles.

Deuxième partie

Définition et apprentissage de métriques à partir de données temporelles

Chapitre 1

Apprentissage de couplages pour la discrimination de séries temporelles

1.1 Résumé

Il n'est pas rare dans les applications que des séries temporelles aient des profils globaux dissimilaires au sein d'une même classe, ou fortement similaires entre les classes. Pour discriminer de telles classes de séries complexes, nous proposons une nouvelle approche d'apprentissage de couplages discriminant, visant à connecter les séries selon des caractéristiques communes au sein des classes et différentielles entre les classes. Cette proposition repose sur un nouveau critère de variance/covariance dont l'objectif consiste à renforcer ou à pénaliser les liens entre les observations en fonction de la variabilité induite au sein et entre les classes. Pour cela, nous proposons une généralisation de l'expression usuelle de la variance/covariance à un ensemble de séries, puis à des classes de séries temporelles. Le couplage appris est utilisé pour la définition d'une métrique locale pondérée, limitant la comparaison de séries à des caractéristiques discriminantes. Les expérimentations menées sur plusieurs jeux de données publics et simulés rendent compte de l'efficacité des couplages appris par rapport aux couplages standards pour la discrimination de séries temporelles. Un version plus complète de ce travail est proposée dans Frambourg et al. [62].

1.2 Introduction

Les séries temporelles provenant des mêmes sources ou mesurant le même phénomène sont souvent bruitées et les événements saillants pouvant apparaître avec des délais très variables. Pour permettre la comparaison de séries temporelles en prenant en compte d'éventuels délais, de nombreuses stratégies d'alignement ont été proposées, telles que celles basées sur la Dynamic Time Warping (DTW) [145, 101, 146, 130, 112]. Notons cependant que l'alignement opéré par la DTW classique demeure une vue locale, limitée à la lumière d'un seul couple de

séries, ignorant la dynamique des séries de la même classe et des autres classes ([37], [129], [141], [79],). Ce type d'alignement affaiblit le potentiel discriminatoire de la métrique dans des applications réelles complexes.

Ainsi, il est d'intérêt en discrimination, d'une part, de rapprocher les séries de même classe selon des caractéristiques communes, d'autre part, d'éloigner les séries de classes différentes sur la base de caractéristiques différentielles. La mesure de la variance/covariance est un critère classiquement utilisé dans de nombreuses approches en analyse de données multivariées. Parmi ces approches, on note quelques travaux pionniers étendant le critère de variance/covariance pour la généralisation, entre autres, de l'analyse en composante principale, de l'analyse factorielle et de l'analyse discriminante, à des structures de graphes ([161], [7], [6], [155], et [121]). Nous proposons ici une nouvelle stratégie d'apprentissage de couplages discriminants fondée sur un critère de variance. Il s'agit de renforcer ou de pénaliser les liens en fonction de leur contribution à la variance intra et inter classe. Ce travail s'articule comme suit : dans la section 1.3, nous proposons tout d'abord une extension de l'expression standard de la variance/covariance à un ensemble, puis à une partition de séries temporelles. Ces définitions permettent d'introduire, dans la section 1.4, deux algorithmes pour l'apprentissage des couplages temporels intra et inter classe. La section 1.5 évalue la pertinence des couplages appris, en vue de la discrimination des séries, sur trois jeux de données publics et un jeu simulé. Enfin, les résultats obtenus et les perspectives sont discutés dans la section 1.5

1.3 Variance/covariance généralisée à des données temporelles

La notion de variance/covariance est bien connue dans le cadre de données numériques non structurées, tant sous sa forme matricielle qu'analytique. Nous avons généralisé ces définitions à un ensemble de séries temporelles, puis plus particulièrement à une partition de séries.

Soit $X(x_{ij})$, une matrice de données $(n \times p)$, caractérisant n observations par p variables numériques $X_1, ..., X_p$. La matrice de variance/covariance usuelle $V_{p\times p}$ s'écrit :

$$V = X^t(I_n - 1_n 1_n^t P)^t P(I - 1_n 1_n^t P)X \tag{1.1}$$

avec I_n la matrice identité, $1_n 1_n^t$ la matrice unité d'ordre n et $P(p_i)$ la matrice diagonale des poids vérifiant $\sum_{i=1}^n p_i = 1$. La variance V_j de la variable X_j est donnée par la formule suivante :

$$V_j = \sum_{i=1}^{n} p_i(x_{ij} - \overline{x_j})^2 = \sum_{i,i'} \frac{1}{2} p_i p_{i'}(x_{ij} - x_{i'j})^2 \tag{1.2}$$

où $\overline{x_j} = \sum_{i=1}^n p_i x_{ij}$ est la moyenne de X_j. Dans le cas d'observations indépendantes, il est aisé de voir dans l'expression de droite de V_j (Eq. 1.2) que l'estimation de la variance fait intervenir toutes les différences de valeurs des observations (i,i'). Ainsi, la matrice $1_n 1_n^t$ peut être vue comme une matrice d'adjacence indiquant les couples d'observations impliqués dans l'estimation de la variance.

Variance induite par un ensemble de séries temporelles Soit X la matrice de données $(nT \times p)$ décrivant n séries temporelles $S_1, ..., S_n$ par p variables $X_1, ..., X_p$ à T instants

d'observations.

$$X = \begin{array}{c} \\ S_1 \\ \\ \vdots \\ \\ S_n \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & \dots & X_p \end{array} \\ \left(\begin{array}{ccc} x_{11}^1 & \dots & x_{1p}^1 \\ \dots & \dots & \dots \\ x_{T1}^1 & \dots & x_{Tp}^1 \\ & \vdots & \\ x_{11}^n & \dots & x_{1p}^n \\ \dots & \dots & \dots \\ x_{T1}^n & \dots & x_{Tp}^n \end{array} \right) \end{array} \tag{1.3}$$

Soit $M(M^{ll'})$ la matrice $(nT \times nT)$ constituée de n^2 blocs matriciels $M^{ll'}(m_{ii'}^{ll'})$. Chaque bloc $M^{ll'}$ explicite le couplage entre S_l et $S_{l'}$, où $m_{ii'}^{ll'} \in [0,1]$ exprime l'intensité du lien entre l'observation i de S_l et l'observation i' de $S_{l'}$ avec $\sum_{i'} m_{ii'}^{ll'} = 1$.

$$\mathbf{M}^{ll'} = \begin{array}{c} \\ S_l \end{array} \begin{array}{c} S_{l'} \\ \left[\begin{array}{ccc} m_{11}^{ll'} & ... & m_{1T}^{ll'} \\ ... & m_{ii'}^{ll'} & ... \\ m_{T1}^{ll'} & ... & m_{TT}^{ll'} \end{array} \right] \end{array} \tag{1.4}$$

On note en particulier trois couplages classiques : le couplage "complet" définit par $M^{ll'} = \frac{1}{T}U$, (U matrice unité) où toutes les observations sont liées et équipondérées, le couplage "Identité" défini par $M^{ll'} = I$ (I matrice identité) où seules sont liées les observations effectuées aux mêmes instants et le couplage "DTW" dans lequel $M^{ll'}$ est défini par l'alignement obtenu par la dissimilarité usuelle *Dynamic Time Warping* (DTW) entre S_l et $S_{l'}$ ([146]).

Soit M la matrice des couplages définis entre les n séries temporelles. On définit V_M la matrice de variance-covariance induite par l'ensemble des séries et généralisant l'expression usuelle définie à l'Eq. 1.1.

$$V_M = X^t(I-M)^t P(I-M)X \tag{1.5}$$

où P est la matrice $(nT \times nT)$ diagonale des poids, avec $p_i = \frac{1}{n\,T}$ sous l'hypothèse d'équipondération des observations. Pour des raisons de clarté, les développements suivants sont donnés dans le cas de séries univariés.

Ainsi, soit x_i^l la valeur de la variable X prise par S_l $(l = 1, ..., n)$ à l'instant i $(i = 1, ..., T)$.

Définition 4 *La variance V_M de la variable X est donnée par :*

$$V_M = \sum_{l=1}^{n}\sum_{i=1}^{T} p_i (x_i^l - \sum_{l'=1}^{n}\sum_{i'=1}^{T} m_{i\,i'}^{l\,l'} x_{i'}^{l'})^2 \tag{1.6}$$

Notons que chaque valeur x_i^l est centrée par rapport au terme $\sum_{l'=1}^{n}\sum_{i'=1}^{T} m_{i\,i'}^{l\,l'} x_{i'}^{l'}$ estimant la valeur moyenne de X au voisinage de l'instant i de S_l. Le voisinage de i est défini par l'ensemble des instants i' de $S_{l'}$ ($l' = 1...n$) connectés à i avec un poids $m_{ii'}^{ll'} \neq 0$. Définissons, dans ce qui suit, la variance intra et inter classes pour un ensemble de séries temporelles partitionné en classes.

Variance induite par une partition de séries temporelles On considère à présent que l'ensemble des séries temporelles $S_1, ..., S_n$ est partitionné en K groupes, avec $y_i \in \{1, ..., K\}$ la classe d'appartenance de la série S_i. Pour une partition de séries temporelles, la variance intra évalue la dispersion induite par les séries au sein de la classe. De ce fait, l'estimation de la variance doit se limiter aux observations i de S_l et i' de $S_{l'}$, pour lesquelles $y_l = y_{l'}$ (i.e., S_l, $S_{l'}$ sont de la même classe).

Ainsi, une matrice M définissant les couplages temporels de séries d'une même classe s'écrit

$$M^{ll'} = \begin{cases} I_T & \text{si } l = l' \\ \neq 0 & \text{si } y_l = y_{l'} \\ 0 & \text{si } y_l \neq y_{l'} \end{cases} \tag{1.7}$$

avec 0 la matrice nulle ($T \times T$) et I_T la matrice identité ($T \times T$). La matrice de variance/covariance V_M introduite à l'Eq. 1.5 définit alors une extension de la variance/covariance intra classique à une classe de séries temporelles. De façon similaire, la variance inter mesure la séparabilité induite par les séries de classes différentes. Elle fait intervenir les observations i de S_l et i' de $S_{l'}$, pour $y_l \neq y_{l'}$ (i.e., S_l, $S_{l'}$ sont de classes différentes). Ainsi, une matrice M définissant les couplages temporels de séries de classes différentes s'écrit :

$$M^{ll'} = \begin{cases} I_T & \text{si } l = l' \\ 0 & \text{si } y_l = y_{l'} \\ \neq 0 & \text{si } y_l \neq y_{l'} \end{cases} \tag{1.8}$$

La matrice V_M définit une extension de la variance/covariance inter usuelle à des classes de séries temporelles. On note dans la suite M_W et M_B les matrices de couplage intra et inter fondées respectivement sur les blocs matriciels définis en 1.7 et 1.8.
Notons que les blocs nuls introduits dans la définition des matrices M_W et M_B traduisent la structure de partition des séries (i.e. classification supervisée). Ces blocs sont non nuls, en particulier, dans un contexte d'apprentissage de couplages pour la catégorisation de séries.

1.4 Apprentissage de couplages discriminants

A chaque couplage intra et inter classes de séries temporelles correspond une estimation de la variance intra et de la variance inter. En vue de la discrimination de classes de séries temporelles, notre objectif consiste à apprendre des couplages discriminants M_W et M_B, à savoir minimisant la variance intra V_{M_W} et maximisant la variance inter V_{M_B}.

L'idée général pour l'apprentissage de tels couplages discriminants consiste à évaluer de manière itérative la contribution de chaque couple (i, i') à la variance intra ou inter estimée. Les poids $m_{ii'}^{ll'}$ sont alors pénalisés pour tous les couples (i, i') dégradant le critère de discrimination. Ce processus est réitéré jusqu'à stabilisation de la variance intra ou inter induite.

Nous introduisons dans ce qui suit deux algorithmes $LearnWitAlig$ et $LearnBetAlig$ proposés respectivement pour l'apprentissage de couplages intra et inter classes. La partie 1.4 discute de la convergence du processus d'apprentissage proposé.

Apprentissage des couplages entre séries au sein d'une classe Pour la minimisation de la variance intra V_{M_W}, on fait appel à la procédure $LearnWitAlig$ (Algorithm 1). Elle prend comme paramètres d'entrée la matrice de données X, le vecteur décrivant les classes d'appartenance Y, le paramètre d'arrêt α, et la matrice d'initialisation des couplages intra MW^0 (i.e. couplage complet), fondée sur les blocs matriciels suivants :

$$Mw^0_{ll'} = \begin{cases} I_T & \text{si } l = l' \\ \frac{1}{T}U & \text{si } y_l = y_{l'} \\ 0 & \text{si } y_l \neq y_{l'} \end{cases} \tag{1.9}$$

L'algorithme $LearnWitAlig$ comporte deux phases. Dans la première phase (ligne 4 à 9), les contributions à la variance intra de chaque couple (i, i') sont évaluées. Pour cela, l'effet induit (augmentation, diminution de la variance) suite à la suppression d'un lien (i,i') est mesuré. Notons que la suppression d'un lien engendre par effet de normalisation la redistribution de son poids sur les voisins. On note $M_{W\setminus(i,i',l,l')}$ la matrice des couplages intra privée du lien (i, i') entre S_l et S'_l, c'est à dire avec $m^{ll'}_{ii'} = 0$ et $y_l = y_{l'}$. Ainsi, on définit la contribution $WC^{ll'}_{ii'}$ de (i, i') à la variance intra :

$$WC^{ll'}_{ii'} = tr(V_{M_W}) - tr(V_{M_{W\setminus(i,i',l,l')}}) \tag{1.10}$$

On note $\mathcal{E}$ l'ensemble des liens (i, i') dont la contribution $WC^{ll'}_{ii'}$ est positive (dont la suppression engendre une diminution de la variance). Ils constituent des liens à pénaliser car augmentant la variance intra.

Dans la seconde phase (lignes 10 à 19), la pénalisation des liens i, i' de $\mathcal{E}$ se traduit par une diminution du poids $m^{ll'}_{ii'}$ d'un couple sélectionné aléatoirement dans $\mathcal{E}$ proportionnellement à $WC^{ll'}_{ii'}$.

On note M^s_W la matrice des poids mise à jour à l'itération s et $V_{M^s_W}$ la variance intra induite. Tant que cette dernière n'est pas stabilisée (ligne 23), on réitère les phases 1 et 2 du procéssus de pénalisation. Dans le cas contraire, la procédure renvoie le meilleur couplage intra classe appris.

Apprentissage des couplages entre séries de classes différentes L'apprentissage des couplages pour des séries de classes différentes se fait de manière symétrique. La matrice d'initialisation M^0_B est fondée sur des couplages complet entre les séries de classes différentes, selon les blocs suivants :

$$Mb^0_{ll'} = \begin{cases} I_T & \text{si } l = l' \\ 0 & \text{si } y_l = y_{l'} \\ \frac{1}{T}U & \text{si } y_l \neq y_{l'} \end{cases} \tag{1.11}$$

On définit la contribution d'un lien (i, i') à la variance inter :

$$BC^{ll'}_{i,i'} = tr(V_{M^s_B}) - tr(V_{M^s_{B\setminus(i,i',l,l')}}) \tag{1.12}$$

Algorithm 1 *LearnWitAlig*(X, Y, M_W^0, α)

1: $s = 0$
2: Soit M_W^s la matrice de couplage initiale
3: **repeat**
4: **for all** $(S_l, S_{l'}) : y_l = y_{l'}$ and $l \neq l'$ **do**
5: **for all** $(i, i') \in [1, T] \times [1, T]$ **do**
6: {évaluation des contributions intra}
7: $WC_{i,i'}^{ll'} = tr(V_{M_W^s}) - tr(V_{M^s_{W \setminus (i,i',l,l')}})$
8: **end for**
9: **end for**
10: $\mathcal{E}$={(i, i') / $WC_{i,i'}^{ll'} > 0$ }
11: Choisir aléatoirement $(i, i') \in \mathcal{E}$
12: {pénalisation du poids (i, i')}
13: $m_{i,i'}^{ll'} = m_{i,i'}^{ll'}.(1 - \frac{WC_{i,i'}^{ll'}}{\sum_{ii',ll'} |WC_{i,i'}^{ll'}|})$
14: {normalisation de la ligne i}
15: **for all** $S_k : y_l = y_k$ and $l \neq k$ **do**
16: **for all** $r \in [1, T]$ **do**
17: $m_{i,r}^{lk} = \frac{m_{i,r}^{lk}}{\sum_{i'} m_{i,r}^{lk}}$
18: **end for**
19: **end for**
20: $s = s + 1$
21: {mise à jour de M_W}
22: **until** $\frac{tr(V_{M_W^{s-1}}) - tr(V_{M_W^s})}{tr(V_{M_W^{s-1}})} \leq \alpha$ {répéter l'apprentissage jusqu'à stabilisation de la variance intra}
23: $return(M_W^s)$

$\mathcal{E}$ est l'ensemble des liens (i, i') dont la contribution $BC_{ii'}^{ll'}$ est négative, c'est-à-dire les liens qui tendent à diminuer la variance inter. À chaque itération, un lien (i, i') est choisi aléatoirement dans $\mathcal{E}$, et son poids est pénalisé proportionnellement à sa contribution $BC_{i,i'}^{ll'}$.

$$m_{ii'}^{ll'} = m_{ii'}^{ll'}.(1 + \frac{BC_{i,i'}^{ll'}}{\sum_{ii',ll'} |BC_{i,i'}^{ll'}|}) \tag{1.13}$$

L'algorithme converge lorsque l'augmentation de la variance descend sous le seuil α.

$$\frac{tr(V_{M_W^{s-1}}) - tr(V_{M_W^s})}{tr(V_{M_W^{s-1}})} \leq \alpha \tag{1.14}$$

Convergence du processus d'apprentissage des couplages temporels Par symétrie des deux algorithmes *LearnWitAlig* et *LearnBetAlig*, notre discussion va se concentrer sur la convergence de *LearnWitAlig*.
La convergence de l'algorithme est liée à la décroissance de la variance au cours du processus d'apprentissage. Or, l'estimation de la variance V_{M_W} dépend du poids $m_{ii'}^{ll'}$ du lien (i, i'). La variance peut alors s'exprimer comme une fonction V, avec $V(m_{ii'}^{ll'}) = V_{M_W}$. Soit f définie par :

$$\begin{aligned} f : [0, 1] &\rightarrow \mathbb{R} \\ \alpha &\mapsto V_{M_W} - V(\alpha m_{ii'}^{ll'}) \end{aligned}$$

On peut montrer que la fonction f est polynomiale, avec $f(1) = 0$ et $f(0) = WC_{ii'}^{ll'} > 0$. 1 est une racine de f, et en pratique, les autres racines du polynôme sont grandes (de l'ordre de $\frac{1}{m_{ii'}^{ll'}}$). Il n'y a donc pas de zéros du polynôme entre 0 et 1. En particulier, la fonction f est monotone et de signe constant sur $[0, 1[$. Soit α_0 le facteur de pénalisation dans l'algorithme :

$$\alpha_0 = (1 - \frac{WC_{i,i'}^{ll'}}{\sum_{ii',ll'} |WC_{i,i'}^{ll'}|}) \tag{1.15}$$

$f(\alpha_0)$ est donc compris entre 0 et $WC_{ii'}^{ll'}$. En particulier, $f(\alpha_0)$ est positif. Donc, $V_{M_W} > V(\alpha_0\ m_{ii'}^{ll'})$. La pénalisation entraîne une diminution de la variance, donc la variance chute au cours du processus itératif, ce qui assure la convergence de l'algorithme *LearnWitAlig* et par symétrie, celle de *LearnBetAlig*.

1.5 Applications et étude comparative

Description des jeux de données Les algorithmes proposés pour l'apprentissage des couplages discriminants (*LearnWitAli* et *LearnBetAli*) sont appliqués à des données fréquemment utilisées en classification de séries temporelles, CBF (Cylinder-Bell-Funnel) proposé par [144], CC (Synthetic Control Chart) et TRAJ (Character trajectories) proposés par [4].

Ces données usuelles partagent des caractéristiques commmunes : chaque classe identifie un profil distinct, les classes sont facilement séparables selon le profil global des séries et, au sein d'une même classe, les séries varient dans des domaines de valeurs relativement proches. Il est évident que les séries temporelles rencontrées dans des applications réelles peuvent présenter des caractériqtiques beaucoup plus complexes. Ainsi, pour étendre le processus de validation à des données moins "simples", on introduit un jeu de données supplémentaire BME, caractérisé par des séries ayant des profils globaux distincts au sein d'une même classe.

BME comprend trois classes de séries (Figure 1.1). Dans la classe Begin, les séries partagent une signature commune caractérisée par l'apparition d'une cloche apparaissant en début de trajectoire, et peuvent diverger sur la trajectoire restante selon que la cloche principale est orientée vers le haut ou vers le bas. La classe Middle est constituée de séries partageant un comportement global similaire caractérisé par une grande cloche centrale. Les séries de la classe End partagent un événement commun caractérisé par une cloche située en fin de trajectoire, et dont le comportement global peut différer selon que la cloche principale est orientée vers le haut ou vers le bas.

Le Tableau 1.1 précise les principales caractéristiques des quatre jeux de données ci-dessus, indiquant pour chacun : l'origine des jeux de données (source=1 : simulés selon le papier d'origine, source=2 : téléchargés du site Machine Learning Repository, source=3, simulés selon les préconisation du présent papier), la taille du jeu (Taille), le nombre de classes (Nb. cla), le nombre de séries par classe (Nb. ST/cla), la longueur des séries (long ST), ainsi que la nature multivariée (Multi.) ou réelle (Reel) des séries.

Apprentissage de couplages discriminants La Figure 1.2 nous permet d'illustrer un exemple de couplage appris entre deux séries de la classe *Cylinder* du jeu CBF. D'une part, la figure de gauche visualise les liens entre une observation de S_l et toutes les observations de

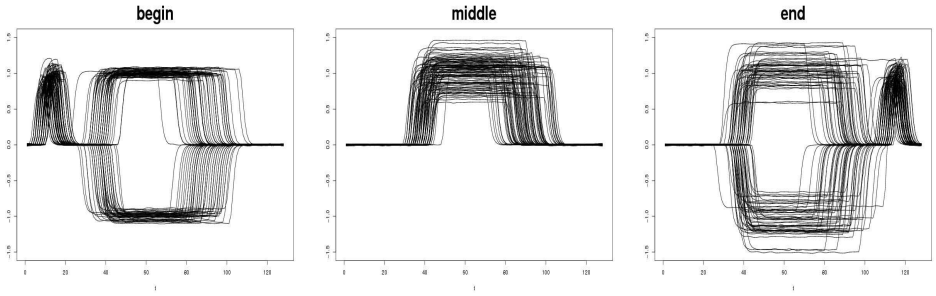

FIGURE 1.1 – Les classes du jeu BME

Nom	Source	Taille	Nb.cla	Nb. ST/cla	Long ST	Multi.	Reel
CBF	1	300	3	100	128	Oui	Non
CC	2	600	6	100	60	Non	Non
TRAJ	2	1000	20	50	20	Oui	Oui
BME	3	300	3	100	128	Non	Non

TABLE 1.1 – Description des jeux de données

$S_{l'}$. Les traits les plus épais traduisent des liens de poids forts, ici essentiellement entre l'observation i de S_l et les observations i' du plateau de $S_{l'}$. Les traits hachurés correspondent à des liens de très faibles poids ($mw_{ii'}^{ll'} \approx 0$). D'autre part, la figure de droite visualise l'intensité des connections entre les observations de S_l et $S_{l'}$. Les cellules les plus claires identifient des régions correspondant à des liens de poids forts. Par exemple, la forme rectangulaire centrale clair illustre une zone de couplage fort entre les plateaux des deux séries de la classe *Cylinder*.

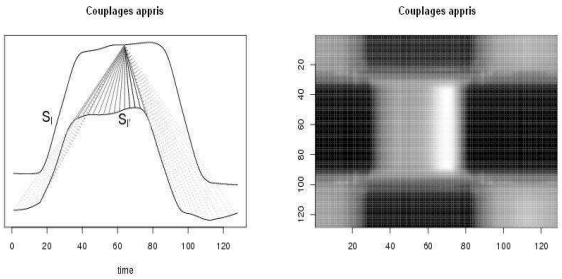

FIGURE 1.2 – Les couplages appris entre deux séries de la classe *Cylinder* (jeu CBF)

Évaluation du pouvoir discriminant des couplages M_W^* et M_B^* Pour évaluer le pouvoir discriminant des couplages appris M_W^* et M_B^*, nous estimons pour chaque jeu de données les variances intra $V_{M_W^*}$ (compacité des classes), inter $V_{M_B^*}$ (isolation des classes)

et les ratios intra/inter ρ^* (taux de discrimination) induits. L'efficacité de ces couplages est ensuite étayée par comparaison aux couplages standards : euclidien ($M^I_{ll'}$) et DTW ($M^{DTW}_{ll'}$)

Ces résultats sont résumés dans le Tableau 1.2, permettant de comparer les critères de compacité, de séparabilité et de discrimination sur l'ensemble des jeux de données, et pour les trois couplages principaux (appris, euclidien et DTW).

	Compacité			Séparabilité			Discrimination		
Jeux	$V_{M^*_W}$	$V_{M^I_W}$	$V_{M^{DTW}_W}$	$V_{M^*_B}$	$V_{M^I_B}$	$V_{M^{DTW}_B}$	ρ^*	ρ^I	ρ^{DTW}
CBF	**0.119**	1.771	0.163	**18.441**	4.844	1.004	**0.006**	0.366	0.162
CC	**1.732**	14.597	2.587	**212.339**	130.001	107.818	**0.008**	0.112	0.024
TRAJ	**0.057**	0.341	0.145	**10.830**	1.902	0.739	**0.005**	0.305	0.196
BME	**22.161**	65.955	22.734	**199.476**	109.089	35.548	**0.111**	0.605	0.640

TABLE 1.2 – Comparaison des pouvoirs discriminants des couplages appris, Euclidien, et DTW

Les Figures 1.3 et 1.4 illustrent, par exemple, pour les jeux de données CC et TRAJ, la progression des variances intra et inter durant le processus d'apprentissage. Les figures de gauche montrent la décroissance significative de la variance intra $V_{M^s_W}$ et la comparent aux variances intra $V_{M^I_W}$, $V_{M^0_W}$ et $V_{M^{DTW}_W}$ fondées respectivement sur les couplages euclidien, complet et DTW.

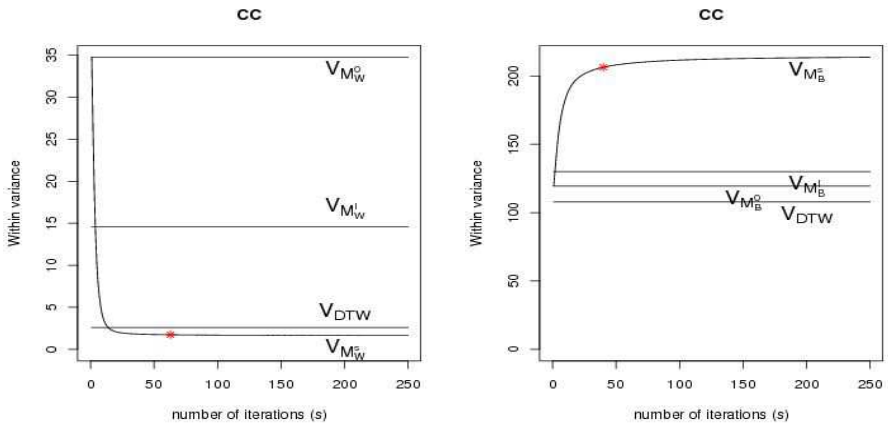

FIGURE 1.3 – Évolution des variances intra et inter classe au cours du processus d'apprentissage ($\alpha = 10^{-3}$) pour le jeu CC.

De manière similaire, nous constatons dans les Figures 1.3 droite et 1.4 droite une croissance drastique de la variance inter $V_{M^s_B}$ comparée aux variances inter $V_{M^I_B}$, $V_{M^0_B}$ et $V_{M^{DTW}_B}$. L'étoile indique les variances intra et inter apprises $V_{M^*_W}$ et $V_{M^*_B}$, retenues lors de l'apprentissage pour un seuil d'arrêt $\alpha = 10^{-3}$.

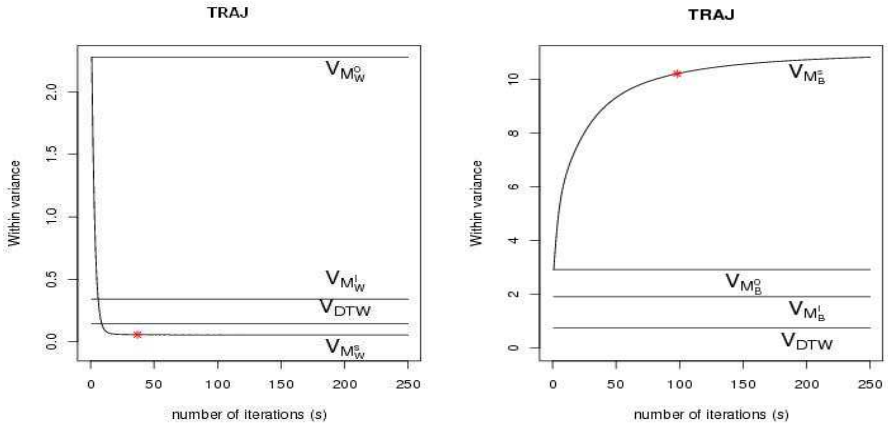

FIGURE 1.4 – Évolution des variances intra et inter classe au cours du processus d'apprentissage ($\alpha = 10^{-3}$) pour le jeu TRAJ.

Classification fondée sur les couplages appris Notre avons ensuite comment le couplage appris permet la définition d'une mesure de proximité limitant la comparaison des séries à leurs caractéristiques discriminantes. Une classification (ici 1-Kppv) est alors utilisée pour comparer les performances de la métrique fondée sur le couplage appris aux métriques classiques. Pour ces dernières, nous avons considéré la distance euclidienne et la DTW. Etant donné le couplage appris M^*, définissons l'indice de proximité D^* entre une nouvelle série S^* et une série S^l de l'échantillon d'apprentissage :

$$D^*(S^*, S^l) = \left\|(S^* - \mathbb{S}^l)\right\|_2 \tag{1.16}$$

$$\text{avec} \quad \mathbb{S}_l = Mw_{l.}\, X \tag{1.17}$$

où $Mw_{l.}$ correspond à la sous-matrice de Mw constituée de l'ensemble des blocs $Mw_{ll'}$, l' parcourant l'ensemble des séries de la classe. Ainsi, $\mathbb{S}_l$ définit le profil moyen au voisinage de S^l. En particulier, la valeur de l'observation i du profil moyen $\mathbb{S}_l$ se note $\mathbb{S}_{li} = \overline{x_{ij}^{S_l}}$ (i.e. la valeur moyenne au voisinage de i). Le Tableau 1.3 résume les taux d'erreur obtenus avec l'algorithme 1-kppv, pour la métrique fondée sur D^* ainsi que pour les métriques euclidienne (DE) et Dynamic Time Warping (DTW).

	Métrique	CBF	CC	TRAJ	BME
Taux d'erreur	$D*$	**9.2%**	**3.8%**	1.3%	**8.4%**
	DE	10.5%	4.7%	**1.2%**	17.7%
	DTW	33.2%	9.6%	1.9%	13.4%

TABLE 1.3 – Erreurs de classification (1-kppv)

Discussion et perspectives Dans un premier temps, nous discutons et analysons les comportements des processus d'apprentissage présentés dans les Figures 1.3 et 1.4. Les deux figures de gauche indiquent une décroissance qui traduit la chute de la variance intra $V_{M_W^s}$ au cours du processus itératif d'apprentissage. La valeur optimale retenue pour la variance intra est significativement inférieure à $V_{M_W^I}$ et à $V_{M_W^{DTW}}$. De même, il ressort des Figures 1.3 droite et 1.4 droite que la croissance de la variance inter $V_{M_B^s}$ est régulière au cours du processus d'apprentissage. La valeur optimale de la variance inter apprise $V_{M_B^*}$ est significativement plus forte que $V_{M_B^I}$ et $V_{M_B^{DTW}}$. Pour un seuil d'arrêt α fixé à 10^{-3}, le processus d'apprentissage converge en moins de 50 itérations pour la variance intra, entre 50 et 100 itérations pour la variance inter.
Le comportement décroissant de la variance intra (qui s'oppose au comportement croissant des variances inter) à travers l'ensemble des jeux de données révèle la pertinence de la pénalisation adoptée en vue de la maximisation de la compacité et de la séparabilité des classes. La progression drastique de la variance en début d'apprentissage et son ralentissement jusqu'à stabilisation met en lumière deux caractéristisques : la convergence du processus d'apprentissage, et sa capacité à moduler l'intensité des modifications, fortes en début d'apprentissage, quand le critère est loin de l'optimalité, et plus faibles au fil des itérations.

On remarque sur la Figure 1.4 que les valeurs obtenues pour les couplages euclidien, DTW et D^*, sont assez proches. En effet, au travers de la Figure 1.5, nous constatons que les matrices de couplages apprises pour le jeu TRAJ révèlent un alignement diagonal correspondant à un couplage euclidien, indiquant le potentiel de l'approche proposée à apprendre également des alignements classiques globaux.

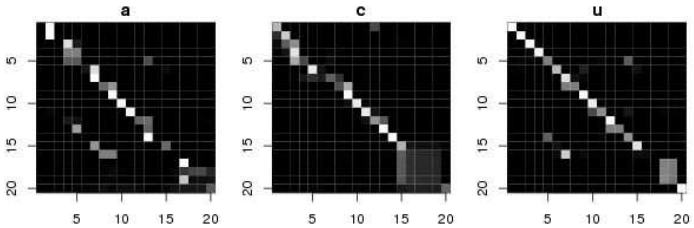

FIGURE 1.5 – Les couplages appris M_W^* pour des classes du jeu TRAJ ("a","c","u").

Le Tableau 1.2 fait état, sur tous les jeux, d'une meilleure discrimination (compacité et isolation) des classes pour M^*. Les alignements de la DTW donnent des classes plus compactes que le couplage euclidien, avec une isolation moindre et un critère de discrimination équivalent. En revanche, les couplages appris donnent le meilleur pouvoir discriminant pour tous les jeux. Le Tableau 1.3 résume les taux de classification obtenus sur l'ensemble des jeux. Ces résultats montrent l'apport de la métrique fondée sur les couplages appris en comparaison des métriques standards. DE et DTW engendrent des taux d'erreur élevés, en particulier pour des jeux complexes tels que BME. La Dynamic Time Warping donne des taux d'erreur particulièrement élevés pour le jeu CBF. En effet, pour ce jeu, la structure des classes sous-jacentes est fortement liée aux instants d'observation. Enfin, pour le jeu TRAJ, les taux de classement pour DE et D* sont équivalents puisque les couplages appris tendent vers des alignements euclidiens, tel que l'illustre la Figure 1.5.
Dans le cadre de la discrimination de séries temporelles, ce travail est une première avancée,

puisqu'il permet la prise en compte de la structure globale des séries. Cependant, les apprentissages des couplages intra et inter demeurent des processus séparés. Les perspectives de ce travail visent à étendre la proposition actuelle pour un apprentissage des structures intra et inter simultanément.

1.6 Conclusion

Ce travail a été motivé par la discrimination de séries complexes par apprentissage de couplages de séries multiples. Nous proposons une approche visant à connecter les séries d'une même classe selon les caractéristiques communes au sein des classes et différentielles entre les classes. L'approche proposée est guidée par la minimisation de la variance intra-classe et la maximisation de la variance inter-classe [61, 64, 63]. Les principaux résultats de ce travail sont : -une extension des stratégies d'alignements classiques à des couplages moins contraints temporellement, -la prise en compte lors de l'apprentissage des couplages de la dynamique de toutes les séries intra et inter classes (approche non pair-à-pair), -une extension de l'expression usuelle de la variance/covariance à un ensemble de séries temporelles, ainsi qu'à des classes de séries, -l'apprentissage d'une dissimilarité locale pondérée, restreignant la comparaison des séries aux attributs discriminants. Ce travail s'inscrit dans le cadre de la thèse de C. Frambourg en co-direction avec J. Demongeot et en collaboration avec E. Gaussier. Les perspectives de mes travaux de recherche discutent des développements futurs de ce travail.

Chapitre 2

Mesures de proximité intégrant la forme des séries : application à des données d'expression de gènes

2.1 Résumé

Le problème biologique d'intérêt porte sur la classification et la catégorisation de la dynamique des gènes au cours du cycle cellulaire. Ces profils d'expression constituent des données complexes : de comportements périodiques, ils peuvent inclure des variations d'amplitudes, des atténuations de phases au cours du cycle cellulaires ainsi que des effets de tendances. J'introduis une mesure de proximité intégrant les composantes forme et valeurs des séries, le meilleur compromis des deux composantes étant appris au cours des processus de classification ou de catégorisation. Je situe cette mesure dans un cadre unifié portant sur trois familles de métriques pour des séries temporelles. Le comparaison des performances de la mesure proposée et des mesures classiques est réalisée sur la base de données réelles et d'un modèle génératif. Le modèle considéré rend compte de phénomènes complexes tels que la désynchronisation cellulaire provoquant à la fois l'atténuation en amplitude des valeurs d'expression et des modifications de périodicité des cycles cellulaires successifs. Je présente ici un résumé du contexte biologique et des objectifs de l'application, l'approche d'analyse menée et quelques résultats ; pour un exposé complet de ces travaux Cf. Douzal-Chouakria et al. [50, 51].

2.2 Introduction

Toutes les cellules de notre corps contiennent les mêmes gènes, mais tous n'interviennent pas dans chaque cellule : les gènes sont activés ou réprimés selon les besoins. De tels gènes spécifiques définissent le modèle moléculaire lié à une fonction spécifique d'une cellule et apparaissent dans la plupart des cas comme organisés dans des réseaux de régulation moléculaire. Pour comprendre comment les cellules réalisent une telle spécialisation, il est nécessaire d'identifier quels gènes s'expriment dans chaque type de cellules (par exemple, des tissus cancéreux versus des tissus sains). La technologie des puces à ADN nous permet d'étudier

simultanément les niveaux d'expression de plusieurs milliers de gènes, au cours de processus biologiques importants, pour déterminer ceux qui sont exprimés dans un type de cellule spécifique[55]. Les techniques de catégorisation et de classification sont utilisées et se sont montrées particulièrement efficaces pour comprendre la fonction des gènes, des voies de régulation et des processus cellulaires (e.g., [115],[134],[148]). Nous distinguons au moins deux principales approches de catégorisation et de classification de profils ou de séries temporelles. D'une part, les approches paramétriques consistant à projeter les séries temporelles dans des espaces de fonctions correspondant, par exemple, aux polynômes d'un modèle ARIMA, aux transformées de Fourier, ou plus généralement aux paramètres d'un modèle approximant les séries temporelles. Des mesures standards peuvent alors être utilisées dans le nouvel espace de projection (e.g.,[8],[12],[67]). D'autre part, on distingue les approches non-paramétriques dont l'objectif est la proposition de nouvelles mesures de proximité définies dans l'espace de description initial et intégrant la dimension temporelle des données (e.g.,[3],[85],[99]). Dans le cadre des approches non-paramétriques, nous proposons d'étudier l'efficacité de quatre métriques majeures pour la catégorisation et la classification des profils temporels d'expression de gènes. Cette étude est basée sur la mise en œuvre d'un modèle périodique aléatoire pour la simulation de gènes d'expression cyclique. Ce modèle tient compte des caractéristiques principalement observées sur les profils de gènes du cycle cellulaire : l'amplitude initiale du profil, la période du profil, l'atténuation des amplitudes au cours du temps et les effets de tendance. La suite de l'article est organisée en quatre sections. La section suivante définit ce que sont les données d'expression de gènes et présente le problème biologique abordé. La section 2.4 présente les quatre principales métriques à évaluer et discute de leurs caractéristiques. La section 2.5 indique comment les mesures seront comparées au sein d'un processus de catégorisation et de classification des profils de gènes. Enfin, nous présentons les méthodes d'évaluation basées sur le modèle génératif aléatoire et discutons les résultats obtenus dans la section 2.6

2.3 Identification des gènes exprimés au cours du cycle cellulaire

Le problème biologique d'intérêt est l'analyse de la progression de l'expression des gènes durant le processus de la division cellulaire. La division cellulaire est le processus principal assurant la prolifération des cellules, et se décompose en quatre phases principales (G_1, S, G_2 et M) et trois transitions de phase (G_1/S, G_2/M et M/G_1)(Figure 2.1). Le processus de division commence à la phase G_1 pendant laquelle la cellule se prépare à la synthèse de l'ADN. Vient ensuite la phase S où l'ADN est dupliqué (c-à-d chaque chromosome est dupliqué), suivie par la phase G_2 pendant laquelle la cellule se prépare à la division. Enfin, vient la phase M où la cellule se divise en deux cellules filles. Pendant ces quatre phases, certains gènes sont actifs (fortement exprimés) à des périodes spécifiques, d'autres pas. Un des objectifs consiste à identifier les gènes fortement exprimés et caractérisant chaque phase du cycle cellulaire. Ceci fournit des informations importantes, par exemple, pour comprendre comment le traitement hormonal peut induire la prolifération cellulaire par l'activation de gènes spécifiques. Ce sont les développements de la technologie des puces à ADN qui ont permis de répondre à cet objectif. Des molécules d'ADN représentant les différents gènes sont placées sur des spots discrets régulièrement répartis en une matrice ligne/colonne. En déposant sur ces puces à ADN des brins d'ARN extraits de populations cellulaires on peut

mesurer le niveau d'expression de chaque gène au sein des populations cellulaires étudiées. En échantillonnant au cours du temps une population cellulaire initialement synchronisée, chaque gène étudié peut être décrit par son profil d'expression observé au cours du temps sur un ou plusieurs cycles de la division cellulaire.

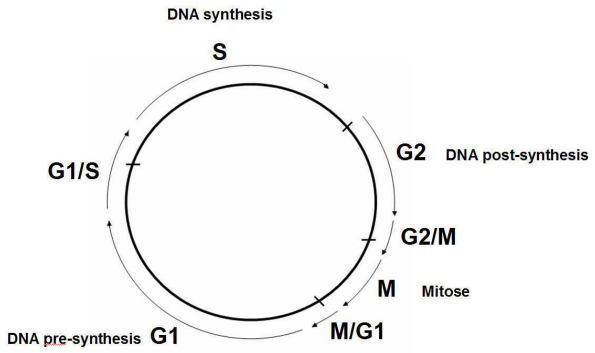

FIGURE 2.1 – Processus de la division cellulaire

2.4 Une formalisation unifiée des métriques pour séries temporelles

Nous présentons, dans un cadre unifié, trois catégories de métriques pour des séries temporelles. La première catégorie porte sur des mesures limitant la comparaison des séries à leurs valeurs, la dépendance temporelle des valeurs étant ignorée. Appartiennent à cette catégorie, entre autres, la distance euclidienne et la dynamic time warping. La deuxième catégorie de métrique s'intéresse à la comparaison des formes (dynamiques) des séries, elle inclut en particulier les coefficients de corrélation de Pearson et temporelle. Enfin, nous introduisons une troisième catégorie de mesure, permettant de couvrir les deux aspects formes et valeurs des séries.

Soit $S_1 = (u_1, ..., u_p)$ et $S_2 = (v_1, ..., v_q)$ deux séries temporelles composées de p et q observations effectuées respectivement aux instants $(t_1, ..., t_p)$ et $(t'_1, ..., t'_q)$. Un alignement r entre S_1 et S_2 est défini par la séquence de m couples d'observations :

$$((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m})),$$

avec $a_i \in \{1, .., p\}$, $b_i \in \{1, ..q\}$, vérifiant pour $i \in \{1, .., m-1\}$ les contraintes suivantes :

$$a_1 = 1,\ a_m = p,\ a_{i+1} = a_i \text{ ou } a_i + 1 \text{ et,}$$
$$b_1 = 1,\ b_m = q,\ b_{i+1} = b_i \text{ ou } b_i + 1.$$

avec $m \in [max(p,q), p+q-1]$. Soit R un sous ensemble d'alignements ainsi définis entre S_1 et S_2, et $c(r)$ une fonction coût associée à l'alignement $r \in R$ mesurant l'écart entre les valeurs couplées dans r. Les principales métriques proposées dans la littérature pour des séries temporelles peuvent être formalisées comme un problème de recherche d'alignements r dans R optimisant une fonction de coût $c(r)$:

$$dUnif_{(c,R)}(S_1, S_2) = \min_{r \in R} c(r). \tag{2.1}$$

Métriques limitées aux valeurs On note, par exemple, que pour $c(r) = \sum_{i=1}^{m} |u_{a_i} - v_{b_i}|$, $dUnif_{(c,R)}$ (Eq. 2.1) définit la dynamic time warping classique (Kruskall and Liberman 1983 [101]) :

$$d_{Dtw}(S_1, S_2) = \min_{r \in R}(\sum_{i=1}^{m} |u_{a_i} - v_{b_i}|), \tag{2.2}$$

$dUnif$ définit la distance de Fréchet [65] pour $c(r) = \max_{i=1}^{m} |u_{a_i} - v_{b_i}|$, :

$$d_F(S_1, S_2) \quad = \quad \min_{r \in R} c(r) = \min_{r \in R} \left(\max_{i=1}^{m} |u_{a_i} - v_{b_i}| \right) \tag{2.3}$$

et la distance euclidienne est obtenue pour $m = p = q$ et la fonction $c(r) = (\sum_{i=1}^{m} (u_i - v_i)^2)^{\frac{1}{2}}$ minimisée dans $R = \{r_0\}$:

$$r_0 = ((u_1, v_1), (u_2, v_2), ..., (u_m, v_m)) \tag{2.4}$$

$$d_E(S_1, S_2) = c(r_0) = (\sum_{i=1}^{m} (u_i - v_i)^2)^{\frac{1}{2}}. \tag{2.5}$$

Métriques limitées à la forme La comparaison de séries temporelles, sur la base de leur formes, a suscité ces dernières années beaucoup d'intérêt émanant d'applications assez variées telles que la reconnaissance de paroles, la conception de systèmes de contrôle, l'analyse de données microarrays et d'expression de gènes. La métrique majoritairement utilisée est le coefficient de corrélation de Pearson comme l'attestent les récents travaux de MacArthur et al. 2010 [118], Ernst et al. 2005 [57], Abraham et al. 2010 [2], Cabestaing et al. 2007 [11], et Rydell et al. 2008 [143]. L'expression du coefficient de corrélation impliquant l'ensemble des couples d'observations est :

$$Cor(S_1, S_2) = \frac{\sum_{i,i'} (u_i - u_{i'})(v_i - v_{i'})}{\sqrt{\sum_{i,i'} (u_i - u_{i'})^2} \sqrt{\sum_{i,i'} (v_i - v_{i'})^2}}. \tag{2.6}$$

L'implication des différences de valeurs entre tous les couples d'observations (c-à-d, observées à tous les couples d'instants (i, i')), fait une hypothèse d'indépendance des observations, pouvant induire une surestimation de la mesure de similarité des séries. Ainsi, comme alternative au coefficient de corrélation classique, nous avons introduit dans Douzal-Chouakria et al. [26] le coefficient de corrélation temporelle, il permet d'inclure la dépendance des observations en limitant le coefficient de corrélation classique aux différences d'ordre faible.

Par exemple, le coefficient de corrélation temporelle limité aux différences de premier ordre est défini :

$$Cort(S_1, S_2) = \frac{\sum_i (u_{i+1} - u_i)(v_{i+1} - v_i)}{\sqrt{\sum_i (u_{i+1} - u_i)^2} \sqrt{\sum_i (v_{i+1} - v_i)^2}} \tag{2.7}$$

Les deux coefficients de corrélation classique (2.6) et temporelle (2.7) présument un alignement du type r_0 (Eq. 2.4) entre des séries de même longueur m. Nous proposons une généralisation de ces expressions à un alignement donné $r = ((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m}))$ dans R :

$$Cor(S_1, S_2) = \frac{\sum_{i,i'} (u_{a_i} - u_{a'_i})(v_{b_i} - v_{b'_i})}{\sqrt{\sum_{i,i'} (u_{a_i} - u_{a'_i})^2} \sqrt{\sum_{i,i'} (v_{b_i} - v_{b'_i})^2}} \tag{2.8}$$

$$Cort(S_1, S_2) = \frac{\sum_i (u_{a_i} - u_{a_{i+1}})(v_{b_i} - v_{b_{i+1}})}{\sqrt{\sum_i (u_{a_i} - u_{a_{i+1}})^2} \sqrt{\sum_i (v_{b_i} - v_{b_{i+1}})^2}} \tag{2.9}$$

Métriques couvrant les composantes forme et valeurs Pour définir une mesure couvrant les deux aspects forme et valeurs des séries, nous avons introduit dans Douzal-Chouakria et al. (2009) [50] une fonction de coût $c_k(r)$ permettant de moduler la proximité centrée sur les valeurs en fonction de la proximité fondée sur la forme :

$$c_k(r) = \frac{2}{1 + exp(k\ Co(r))}.c(r), \quad k \geq 0 \tag{2.10}$$

avec $c(r)$ et $Co(r)$ définissant des fonctions de coûts liées, respectivement, mesurant les écarts entre les valeurs (Eqs. (2.2) et (2.5)) et les formes des séries (Eqs. (2.8), (2.9)). Etant donné la fonction $c_k(r)$, la définition de la mesure $D_k(S_1, S_2)$ alliant les deux composantes forme et valeurs est :

$$D_k(S_1, S_2) = \min_{r \in R}(\frac{2}{1 + exp(k\ Co(r))} c(r)). \tag{2.11}$$

En particulier, notons que pour $R = \{r_0\}$, $Co(r) = Cort(r)$, et $c(r) = (\sum_{i=1}^{m} (u_i - v_i)^2)^{\frac{1}{2}}$, D_k définit une extension de la distance euclidienne, noté DE_k^{cort}, prenant en compte les composantes forme et valeurs des séries :

$$\text{DE}_k^{cort}(S_1, S_2) = \frac{2}{1 + exp(k\ Cort(r_0))}\ (\sum_{i=1}^{m} (u_i - v_i)^2)^{\frac{1}{2}}.$$

Une extension similaire de la dynamic time warping est obtenue pour $Co(r) = Cort(r)$ et $c(r) = \sum_{i=1}^{m} |u_{a_i} - v_{b_i}|$,

$$\text{DTW}_k^{cort}(S_1, S_2) = \min_{r \in R}(\frac{2}{1 + exp(k\ Cort(r))}\ \sum_{i=1}^{m} |u_{a_i} - v_{b_i}|)$$

En résumé, la Table 2.1 présente, dans un cadre unifié, les différentes métriques discutées dans cette section, et quelques une de leur extensions.

Type	R	$c(r)$	$Co(r)$	Metric
Values	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	-	$d_{Dtw} = \min_{r \in R}(\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert)$
	$R = \{r_0\}$	$(\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$	-	$d_E = c(r_0) = (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$
Behavior	$R = \{r_0\}$	-	$Cor(r)$	$d_{Cor} = 1 - Cor(r_0)$
	$R = \{r_0\}$	-	$Cort(r)$	$d_{Cort} = 1 - Cort(r_0)$
	$R \subset M$	-	$Cor(r)$	$dtw_{Cor} = \min_{r \in R}(1 - Cor(r))$
	$R \subset M$	-	$Cort(r)$	$dtw_{Cort} = \min_{r \in R}(1 - Cort(r))$
	$R = \{r_0\}$	$(\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$	$Cor(r)$	$DE_k^{Cor} = \frac{2}{1 + exp(k\ Cor(r_0))}\ (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$
Val.&	$R = \{r_0\}$	$(\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$	$Cort(r)$	$DE_k^{Cort} = \frac{2}{1 + exp(k\ Cort(r_0))}\ (\sum_{i=1}^{m}(u_i - v_i)^2)^{\frac{1}{2}}$
Beh.	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	$Cor(r)$	$DTW_k^{Cor} = \min_{r \in R}\left(\frac{2}{1 + exp(k\ Cor(r))}\ \sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert\right)$
	$R \subset M$	$\sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert$	$Cort(r)$	$DTW_k^{Cort} = \min_{r \in R}\left(\frac{2}{1 + exp(k\ Cort(r))}\ \sum_{i=1}^{m} \lvert u_{a_i} - v_{b_i} \rvert\right)$

TABLE 2.1 – Définition des principales métriques pour des séries temporelles

2.5 Classification/catégorisation de profils d'expression de gènes

Des simulations fondées sur un modèle génératif sont menées pour évaluer l'efficacité des métriques introduites en section 2.4 Pour la procédure de catégorisation, nous proposons d'utiliser l'algorithme PAM (Partitioning Around Medoïds) afin de partitionner les gènes simulés en n classes (n étant le nombre de phases du cycle cellulaire ou de transitions de phases étudiées). L'algorithme PAM est préféré à l'approche classique des K-means pour plusieurs raisons. Il est plus robuste aux valeurs aberrantes qui sont nombreuses dans les données d'expression de gènes. PAM permet une analyse détaillée de la partition en fournissant des indices permettant d'apprécier la qualité des classes ainsi que celle des gènes. En effet, PAM mesure la *silhouette width* (sw) de chaque gène, un indicateur de confiance quant à l'appartenance d'un gène à une classe. Pour plus de détails sur l'algorithme PAM voir[95]. L'efficacité de chaque métrique est basée sur trois critères : *la silhouette width* moyenne d'une partition notée asw, le ratio standard $wbr = \frac{intra}{inter}$ et l'indice de Rand corrigé (RI). Pour la procédure de classification des gènes, l'algorithme 10-NN est utilisé, et les taux d'erreur de gènes mal classés sont retenus pour apprécier l'efficacité de chaque métrique.

2.6 Étude comparative fondée sur un modèle génératif de profils d'expression périodiques

Modèle génératif de profils d'expression périodiques Nous utilisons ici génératif dans son sens le plus large, celui d'un modèle pour la simulation de profils d'expression de gènes. Nous nous basons sur un modèle de régression non-linéaire proposé par [114]. Ce modèle permet de simuler des variations similaires à celles observées expérimentalement, par exemple, des atténuations dans l'amplitude de l'expression des gènes au cours des différentes phases du cycle cellulaire. La fonction sinusoïdale caractérisant la périodicité de l'expression

d'un gène g au cours de différents cycles cellulaires est définie par :

$$f(t,\theta_g) \quad = \quad a_g + b_g t + \frac{K_g}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} cos(\frac{2\pi t}{Texp(\sigma z)} + \Phi_g) exp(-\frac{z^2}{2}) dz. \qquad (2.12)$$

où $\theta_g = (K_g, T, \sigma, \Phi_g, a_g, b_g)$ est spécifique à chaque gène g. Le paramètre K_g représente son amplitude initiale, et T la durée du cycle cellulaire. Le paramètre σ contrôle le taux d'atténuation des amplitudes au cours des différents cycles, Φ_g détermine la phase de forte expression au cours du cycle. Les paramètres de tendance des profils sont contrôlés par a_g et b_g définissant, respectivement, l'ordonnée à l'origine et la pente. La Figure 2.2 illustre la progression des expressions de gènes au cours des 5 phases et transitions de phase G_1/S, S, G_2, G_2/M et M/G_1. Nous désignerons, dans la suite, indistinctement par le mot "phase", une phase ou une transition de phase.

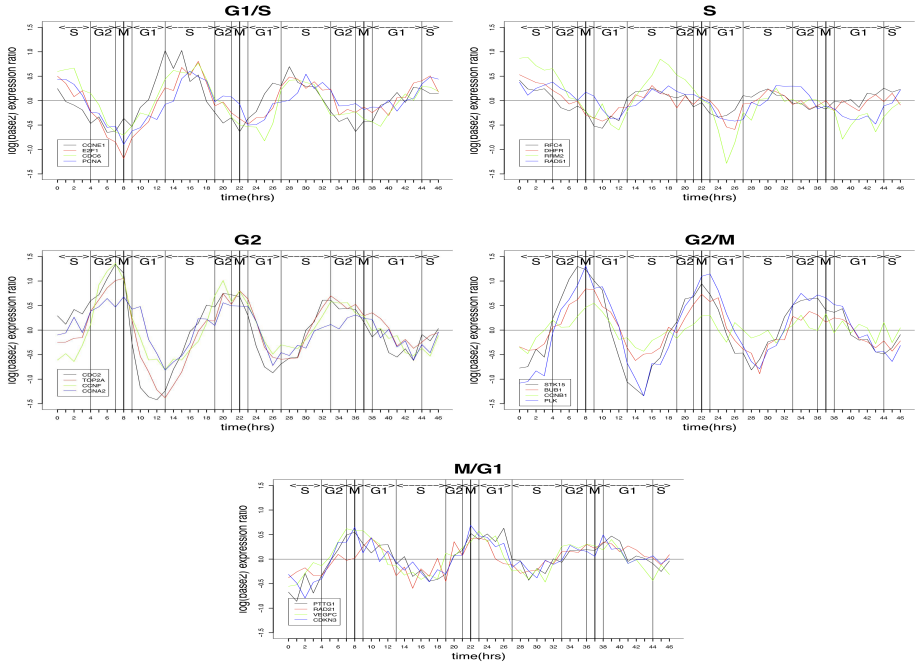

FIGURE 2.2 – Progression de l'expression des gènes durant les 5 phases G_1/S, S, G_2, G_2/M et M/G_1.

Protocole de simulation Sur la base du modèle introduit, quatre expérimentations sont menées. La première expérimentation génère des profils incluant des variations d'amplitude avec K_g évoluant dans $[0.34, 1.33]$. La seconde expérience inclut une atténuation des amplitudes avec σ variant dans $[0.054, 0.115]$. La troisième inclut des effets de tendance en modulant $b_g \in [-0.05, 0.05]$ et $a_g \in [0, 0.8]$ tout en annulant les variations d'amplitude σ. Enfin la quatrième expérimentation simule des profils d'expression, tels que que ceux observés biologiquement, incluant plusieurs variations simultanées des paramètres K_g, σ, a_g, et

b_g. L'évolution des profils est simulée sur 3 cycles cellulaires, T est fixé à 15 heures pour toutes les simulations et Φ_g prend les valeurs 0, 5.190, 3.823, 3.278 et 2.459 pour la génération, respectivement, des 5 phases G_1/S, S, G_2, G_2/M et M/G_1. Les spécifications des paramètres du modèle des quatre expériences sont résumées dans le Tableau 2.2. La Figure 2.3 illustre, pour des gènes de la phase G_1/S, les variations produites par chacune des expériences. Pour chaque expérience $j \in \{1, ..., 4\}$, 10 échantillons S_{ij} $i \in \{1, ..., 10\}$ sont simulés. Chaque échantillon est composé de 500 gènes de profils d'expression de longueur 47, avec 100 gènes pour chacune des 5 phases G_1/S, S, G_2, G_2/M et M/G_1. La comparaison des métriques est effectuée sur un total de 5000 gènes (c'est-à-dire sur 10 échantillons de 500 gènes chacun).

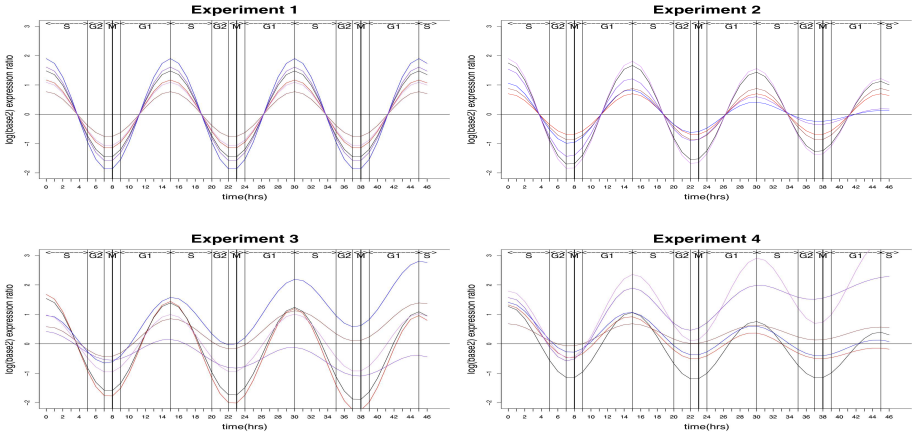

FIGURE 2.3 – Profils des gènes de la phase G_1/S suivant les quatre expériences.

Expérience	K_g	σ	b_g	a_g
1	[0.34, 1.33]	0	0	0
2	[0.34, 1.33]	[0, 0.115]	0	0
3	[0.34, 1.33]	0	[-0.05, 0.05]	[0, 0.8]
4	[0.34, 1.33]	[0, 0.115]	[-0.05, 0.05]	[0, 0.8]

TABLE 2.2 – Spécification des paramètres du modèle.

Evaluation de l'efficacité des métriques pour la catégorisation des gènes Pour chaque expérience et pour chaque métrique δ_E (Eq. (2.5)), COR (Eq. (2.6)), et CORT (Eq. (2.7)), nous partitionnons l'ensemble des profils de chaque échantillon S_{ij} en 5 classes (correspondant aux 5 phases). Par exemple, pour l'expérience j et la métrique δ_E, l'algorithme PAM est appliqué sur les 10 échantillons $S_{1j}, ..., S_{10j}$ afin d'extraire les 10 partitions $\mathcal{P}^{1j}_{\delta_E}$,..., $\mathcal{P}^{10j}_{\delta_E}$. Pour chaque partition $\mathcal{P}^{ij}_{\delta_E}$, les valeurs des trois critères asw, wbr, RI sont mesurées. Ainsi, les valeurs moyennes des critères asw, RI et wbr sur les 10 partitions $\mathcal{P}^{1j}_{\delta_E}$,..., $\mathcal{P}^{10j}_{\delta_E}$ évaluent l'efficacité de la métrique δ_E au sein de l'expérience j. Par ailleurs, une catégorisation fondée sur l'apprentissage de D_k (Eq. (2.11)) est utilisée, avec $c(r) = \delta_E$ et $Co(r) =$ CORT

pour la comparaison de profils d'expression de gènes. Elle consiste, pour chaque échantillon S_{ij}, à exécuter l'algorithme PAM pour des valeurs de k allant de 0 à 6 (avec un pas égal à 0.01). Ceci permet d'apprendre la valeur k^* produisant une partition optimale $\mathcal{P}^{ij}_{D_{k*}}$ au sens des critères asw et wbr. L'efficacité de la métrique D_k au sein de l'expérience j est mesurée au travers des valeurs moyennes des critères asw, RI et wbr sur les 10 partitions $\mathcal{P}^{1j}_{D_{k*}}$,..., $\mathcal{P}^{10j}_{D_{k*}}$ obtenues. La Figure 2.4 montre pour chaque métrique et pour chaque expérience la progression des valeurs moyennes des critères asw (en haut à gauche), wbr (en haut à droite) et RI (en bas).

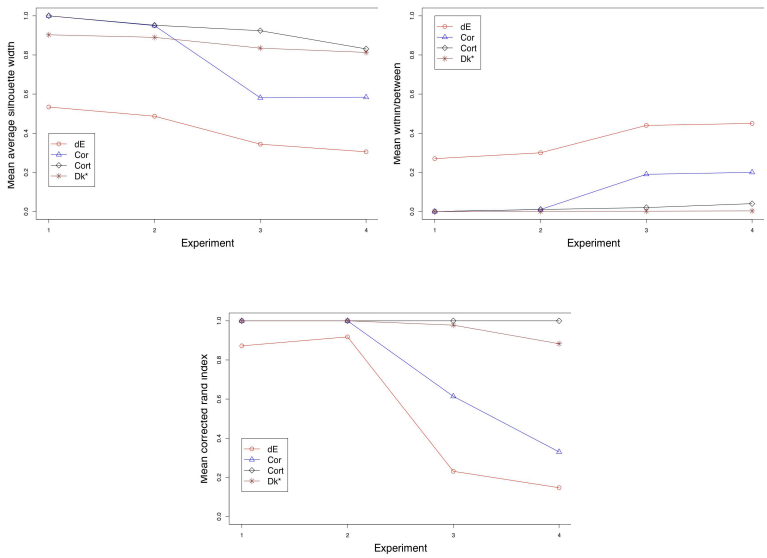

FIGURE 2.4 – Évaluation des métriques pour la catégorisation des profils d'expression simulés.

Évaluation de l'efficacité des métriques pour la classification des gènes Pour chaque expérience et pour chaque métrique δ_E, COR et CORT, l'algorithme 10-NN est appliqué pour la classification des 10 échantillons simulés. Par exemple, pour l'expérience j et la métrique δ_E, l'algorithme 10-NN est appliqué pour la classification des 10 échantillons $S_{1j}, ..., S_{10j}$; soit $\mathcal{C}^{1j}_{\delta_E}$,..., $\mathcal{C}^{10j}_{\delta_E}$ les classifications correspondantes obtenues. Pour chaque classification $\mathcal{C}^{ij}_{\delta_E}$, le taux des gènes mal classés est mesuré. Le taux d'erreur moyen à l'issue des 10 classifications est retenu pour l'évaluation de l'efficacité de la métrique δ_E dans l'expérience j. Pour l'indice de dissimilarité D_k, une classification adaptative est utilisée. Elle consiste à exécuter l'algorithme 10-NN sur l'échantillon S_{ij} avec des valeurs de k allant de 0 à 6 (avec un pas égal à 0.01). On note k^* le paramètre minimisant le taux d'erreur pour la classification de S_{ij}. De manière similaire, le taux d'erreur moyen à l'issue de la classification des 10 échantillons est retenu pour mesurer l'efficacité de la métrique D_k au sein de l'expérience j. La Figure 2.5 illustre la progression des taux d'erreur moyens pour chaque métrique et expérience menées. Enfin, le Tableau 2.3 et la Figure 2.5 résument pour chaque

expérience, la distribution des valeurs k^* $(\overline{k*}, var(k*))$.

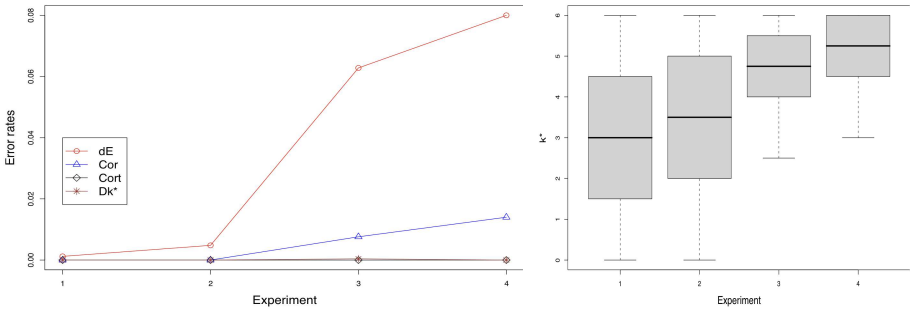

FIGURE 2.5 – Évaluation des métriques pour la classification des profils d'expression simulés (gauche). Distribution de $k*$ en classification adaptative (droite)

Adaptative	Exp1	Exp2	Exp3	Exp4
Catégorisation	(6,0)	(6,0)	(6,0)	(5.85,0.06)
Classification	(3,3.53)	(3,3.53)	(4.55,1.18)	(4.84,0.98)

TABLE 2.3 – k*(moyenne, variance)

Discussion et conclusion Nous discutons, d'abord, l'intérêt du modèle génératif pour l'évaluation des métriques pour la catégorisation et la classification des gènes exprimés au cours du cycle cellulaire. Le modèle génératif considéré permet la simulation de trajectoires périodiques incluant des variations similaires à celles observées biologiquement : varier considérablement en valeur d'un gène à un autre de la même classe (i.e. gènes exprimés au sein de la même phase du cycle cellulaire), comporter des atténuations d'amplitude, dont l'intensité peut elle varier au cours des différents cycles. Le modèle périodique aléatoire permet également d'isoler et d'étudier l'effet de chaque type de variation sur l'efficacité des métriques en catégorisation et en classification. D'autres modèles sont proposés dans la littérature pour la simulation des profils d'expression au cours du cycle cellulaire. On peut distinguer au moins deux principales techniques. D'une part, les données d'expression sont simulées à partir de modèles paramétriques ([138],[116], etc.). Ces modèles fournissent une estimation assez bonne des trajectoires exprimées, cependant les paramètres estimés demeurent difficilement interprétables biologiquement. D'autres travaux se fondent, pour la catégorisation ou la classification des gènes, sur des données d'expression réelle de gènes (dits de référence) (e.g.,[153],[163]). Ces données observées constituent souvent des jeux de données de petites taille, et la littérature ne fournit pas de consensus clair entre les biologistes quand à la catégorisation des gènes de référence. Pour les raisons citées ci-dessus, nous avons opté pour l'utilisation d'un modèle génératif pour la simulation des données d'expression.

Examinons maintenant les résultats obtenus dans le cadre de la catégorisation. Notons quelques informations complémentaires sur les critères utilisés pour l'évaluation de la qualité

des partitions. La valeur asw indique une partition de structure forte pour asw proche de 1 et faible pour $asw < 0.5$. Le critère wbr mesure la compacité des classes (variabilité au sein des classes) et leur séparabilité (variabilité entre les classes). Une bonne partition est caractérisée par une faible valeur de wbr. Enfin, l'indice de Rand corrigé (RI) permet de mesurer l'adéquation entre deux partitions. Une valeur $RI = 0$ correspond à une absence totale d'adéquation, et une valeur de $RI = 1$ traduit une adéquation totale. La Figure 2.4 montre que la catégorisation basée sur δ_E donne, pour les expériences 1 à 4, les partitions les moins fortes comparée à celles fondées sur COR, CORT, ou D_k. En effet, on observe pour les partitions fondées sur δ_E les plus faibles valeurs des critères asw et RI, et les valeurs les plus élevées pour wbr. Les valeurs moyennes des critères asw, wbr et RI de la catégorisation basée sur δ_E se dégradent (diminution des asw et RI et augmentation de wbr) de l'expérience 1 à 4, montrant l'inadéquation de la distance euclidienne face aux variations de plus en plus complexes des profils des gènes. La catégorisation basée sur COR donne, pour les expériences 1 et 2, de bonnes structures de partitions avec de très bonnes valeurs des critères asw, wbr et RI. Toutefois, cette qualité diminue de façon drastique dans les expériences 3 et 4 (Figure 2.4). Comme expliqué dans la section 3, ces résultats confirment la limite du coefficient de corrélation de $Pearson$ face aux variations de tendance. Enfin, les meilleures catégorisations et les plus fortes structures de partitions sont produites par CORT et D_k sur l'ensemble des expériences, avec des valeurs de asw variant dans $[0.8, 1]$, des valeurs de wbr autour de 0, et des valeurs de RI variant dans $[0.83, 1]$. Notons que la qualité de la catégorisation basée sur D_k est légèrement inférieure à celle fondée sur CORT, révélant des profils d'expression de gènes plus différentiables par leur formes que par leurs valeurs. Cette hypothèse est soutenue par les fortes valeurs de $k*$ (proche de 6, avec une variabilité de 0) obtenues dans la catégorisation adaptative pour les quatre expériences (Tableau 2.3).
Considérons les résultats de la classification, la Figure 2.5 (gauche) montre que, pour les expériences 1 et 2, les quatre métriques sont toutes aussi efficaces avec des taux d'erreur de classification autour 0. Toutefois, pour les expériences 3 et 4, nous notons une forte augmentation du taux d'erreur pour les classifications basées sur δ_E, une légère augmentation du taux d'erreur pour les classifications fondées sur COR, une augmentation négligeable pour D_k. Le Tableau 2.3 et la Figure 2.5 (droite) illustrent la distribution des valeurs de $k*$ dans les classifications adaptatives. Pour les expériences 1 et 2, nous notons une distribution uniforme de $k*$ dans $[0, 6]$. Ce cas se présente lorsque une bonne classification peut être aussi bien obtenue avec une métrique fondée sur la forme ou sur les valeurs. En effet, dans les deux premières expériences, la Figure 2.5 (gauche) montre que les quatre métriques sont toutes aussi efficaces pour la classification des gènes avec des taux d'erreur négligeables. Pour les expériences 3 et 4, $k*$ prend des valeurs plus élevées indiquant que les mesures fondées sur la forme (c-à-d CORT et D_k) sont plus efficaces pour la classification des profils d'expression de gènes, avec de très faibles taux d'erreur (Figure 2.5 (gauche)). En résumé, les expériences menées permettent de rendre compte de l'efficacité des mesures CORT et D_k pour la classification des profils d'expression de gènes.

2.7 Conclusion

Ce travail a été réalisé dans le cadre de la thèse de A. Diallo en co-direction avec F. Giroud, biologiste en génomique au laboratoire TIMC-IMAG [50, 51]. L'intérêt de cette étude est double. Du point de vue du domaine de l'apprentissage, nous avons défini une nouvelle mesure intégrant les caractéristiques de forme des séries, et montré que celle-ci correspond à une cross-corrélation locale définie sur une structure en chaîne. Nous avons ensuite situé cette mesure dans un cadre plus large portant sur trois familles de métriques pour les séries temporelles. Enfin, nous avons étudié ses performances sur des données réelles complexes. Du point de vue de la biologie, cette étude a permis d'extraire les principales dynamiques d'activation des gènes au cours du cycle cellulaire, de fournir un nouvel ensemble de gènes de référence (représentants des classes), complétant la faible nombre de gènes de références déterminés expérimentalement, ainsi que d'apprendre une mesure de proximité spécifique aux données d'expressions, permettant d'identifier les principales phases d'activation de nouveaux gènes. Les principales métriques définies dans ce chapitre sont utilisées pour l'évaluation d'une nouvelle approche de classification de séries temporelles par arbre, présentée dans la partie III.

Troisième partie

Classification de séries temporelles

Chapitre 1

Classification/régression de séries temporelles par arbres

1.1 Résumé

Nous nous sommes intéressés dans ce travail à l'extension de la méthode de classification/régression par arbre Breiman et al. (1984) [9] à des variables explicatives de type séries temporelles. Ce problème a suscité quelques propositions dans la littérature, discutables sur différents points. D'une part, l'utilisation de métriques standards limitant la comparaison des séries à leurs valeurs au détriment de leur dynamiques. D'autre part, les spécifications de ces mesures demeurent fixes le long de l'induction de l'arbre, quand bien même les caractéristiques des séries seraient extrêmement variables d'un noeud à l'autre de l'arbre. Enfin, l'implication de l'ensemble des observations dans la comparaison des séries, rend difficile l'atteinte de partitions optimales régies par des caractéristiques locales. Face à ces limites, nous proposons un nouveau critère de coupure fondé sur une métrique adaptative couvrant les composantes formes et valeurs des séries. Les paramètres de la métrique sont appris pour chaque noeud, afin de déterminer une partition des séries au meilleur gain d'information. L'approche proposée permet, en particulier, de localiser au niveau de chaque noeud les sous-séquences discriminantes. La méthode de classification par arbre est mise en application sur un large ensemble de données et pour plusieurs configurations de métriques. Les expérimentations menées illustrent l'efficacité de l'approche et sa performance face aux approches alternatives. Ce chapitre présente une brève synthèse de la méthode proposée et des résultats obtenus ; un exposé complet de ce travail est présenté dans Douzal-Chouakria et Amblard [44].

1.2 Introduction

La classification de séries temporelles a été au centre de nombreux travaux de recherche ces dernières années. On distingue principalement les travaux proposant de nouvelles heuristiques pour une segmentation préalable des séries, puis leur représentation par des descripteurs ad-hoc numériques, pouvant être analysés par des approches standard en classification

(Kudo et al. 1999 [102], Rodríguez et al. 2001 [142], Geurts 2001-2002 [72, 71], Geurts et al. 2005 [73], Kadous et al. 2005 [91]). Une autre catégorie d'approches est fondée sur les chaines de Markov cachées (Rabiner 1989 [137]), largement utilisée dans le domaine de la reconnaissance de formes et du traitement de signal.

Citons, en particulier, deux approches alternatives proposées par Yamada et al. (2003) [165] et Balakrishnan and Madigan (2006) [5], visant à étendre les arbres de classification à des séries temporelles. Dans Yamada et al. (2003) les auteurs proposent un critère de coupure basé sur la recherche d'une série, dite de référence, engendrant la segmentation d'un noeud au meilleur gain d'information. En effet, étant donné une série, la segmentation ou bi-partition d'un noeud est obtenue en affectant au noeud fils droit l'ensemble des séries à une distance de la série de référence inférieure d'un seuil fixé a priori ; les séries restantes sont affectées au noeud fils gauche. Une variante de ce critère consiste à rechercher un couple de séries de référence. La bi-partition est obtenue en affectant chaque série du noeud à la série de référence la plus proche. L'approche proposée par Yamada et al. (2003) [165] se réfère à la dynamic time warping pour la comparaison des séries.

Dans Balakrishnan and Madigan (2006) [5] une approche similaire est proposée, elle consiste à rechercher un couple de séries de référence pour la division de chaque noeud de l'arbre. Pour cela, les auteurs ont recours à l'algorithme des k-means pour la bi-partition de l'ensemble des séries. Les centres des classes étant assimilés aux deux séries de références. Cette approche assure une partition optimisant des critères de catégorisation, en l'occurrence de compacité et d'isolation des classes, cependant au détriment du critère d'homogénéité (gain d'information) de la division. Pour corriger cette insuffisance, plusieurs partitionnement k-means sont effectués, pour en retenir celui de gain d'information maximal. Les performances de l'arbre proposé sont étudiées pour la distance eucilidienne et la dynamic time warping.

Notons que les deux approches décrites ci-dessus, à l'image d'un grand nombre de méthodes de classification de séries temporelles, requiert l'utilisation de la distance euclidienne ou de la dynamic time warping comme mesure de proximité entre les séries temporelles. L'utilisation de ces métriques standards, en particulier dans la définition des critères de coupures, présentent plusieurs limites. D'une part, elles sont fondées sur la comparaison des valeurs des séries au détriment de l'information portant sur la dynamique ou la forme des séries. D'autre part, les spécifications de ces mesures demeurent fixes le long de l'induction de l'arbre, quand bien même les caractéristiques des séries seraient variables d'un noeud à l'autre de l'arbre. Enfin, en impliquant l'ensemble des observations dans la comparaison des séries, les métriques standards rendent difficile la détermination de partitions optimales définies par des caractéristiques locales.

Ce travail s'inscrit dans le cadre des approches basées sur les distances pour étendre les arbres de classification à des données temporelles. Nous proposons un nouveau critère de coupure caractérisé par, d'une part, une métrique adaptative couvrant les aspects formes et valeurs des séries. Les spécifications de la métrique peuvent changer d'un noeud à l'autre de l'arbre pour une meilleur division de l'ensemble des séries. D'autre part, la méthode proposée permet la localisation de segments discriminants à chaque noeud de l'arbre. Dans la Section 1.3 nous présentons les principales étapes de construction de l'arbre de classification proposé.

Enfin, dans la Section 1.4, nous illustrons quelques résultats obtenus.

1.3 Algorithme de construction de l'arbre de classification

Nous présentons, dans cette section, le nouveau critère de coupure *TSSplit* pour l'induction d'arbres de classification pour des séries temporelles. Soit $\{s_1, ..., s_N\}$ un ensemble de N séries temporelles multivariées partitionné en C classes, et $I_1, ..., I_N$ ($I_i = [1, T_i]$) les intervalles d'observations associés. Préalablement à la construction de l'arbre, les séries sont ramenées à une même longueur $I = [1, T]$, et les dissimilarités évaluées pour tous les couples de séries.

Pour bi-partitionner un noeud S, la procédure $TSSplit(S, I, \alpha)$ est utilisée avec comme paramètre en entrée : l'ensemble $S = \{s_1, ..., s_N\}$ des séries à partitionner, l'intervalle d'observation $I = [1, T]$, et le paramètre α nécéssaire à la recherche de sous-séquences discriminantes. La procédure $TSSplit(S, I, \alpha)$ (Algorithm 2), procède à un premier appel à la fonction $AdaptSplit(S, I)$ dont la fonction est de déterminer la meilleure coupure de S au sens du critère de Gini, fondée sur la métrique adaptative D_k et comparant les séries sur la base des observations de I.

Etant donné une valeur du paramètre $k \in [0, 6]$ et deux séries (l, r) de $S \times S$, une bi-partition $\sigma(l, r, k, I)$ de S est obtenue en affectant chaque séries $ts \in S$ au noeud fils gauche si $D_k(ts, l) \leq D_k(ts, r)$, ts est assignée au noeud fils droit sinon (Figure 1.1). Ainsi, pour déterminer la meilleur partition de S, $AdaptSplit(S, I)$ procède à la recherche du triplet (l, r, k) engendrant une partition minimisant l'erreur de Gini. En sortie, $AdaptSplit(S, I)$ retourne la meilleure coupure $\sigma(l^I_*, r^I_*, k^I_*, I)$ ainsi que la valeur de Gini associée $GI(\sigma(l^I_*, r^I_*, k^I_*, I))$.

Notons que la partition $\sigma(l^I_*, r^I_*, k^I_*, I)$ est obtenue en comparant les séries sur la base de tout l'intervalle d'observation I. Pour prendre en compte un partitionnement des séries sur la base de caractéristiques locales, la procédure *DichoSplit* est utilisée, elle consiste à rechercher de manière dichotomique et récursive, dans les sous intervalles gauche puis droit de I, des sous-séquences améliorant l'erreur de Gini.

Ainsi, la procédure $DichoSplit(S, \sigma(l^I_*, r^I_*, k_*, I), e_I, \alpha)$ (Algorithm 4) est appelée avec les paramètres suivants : l'ensemble des séries S, la meilleur coupure $\sigma(l^I_*, r^I_*, k^I_*, I)$ de S obtenue par comparaison des séries sur I, la valeur de Gini associée e_I, et le taux α utilisé pour la définition des bornes des sous-intervalles gauche I_L et droit I_R de I. Deux appels de *DichoSplit* vers *AdaptSplit* sont effectués pour le partitionnement de S sur la base des observations de I_L, puis de I_R.

Dans le cas où l'erreur de Gini n'est pas améliorée en comparant les séries sur la base de I_L ou I_R ($e_I \leq \min(e_{I_L}, e_{I_R})$), alors toutes les observations de I s'avèrent nécessaires à la discrimination de S. La condition d'arrêt de *DichoSplit* est atteinte, et la procédure retourne la coupure $\sigma(l^I_*, r^I_*, k^I_*, I)$. En revanche, si au moins un des intervalles I_L ou I_R améliore l'erreur de Gini, on réitère l'appel à *DichoSplit* pour poursuivre la recherche dans les sous-intervalles I_L ou I_R.

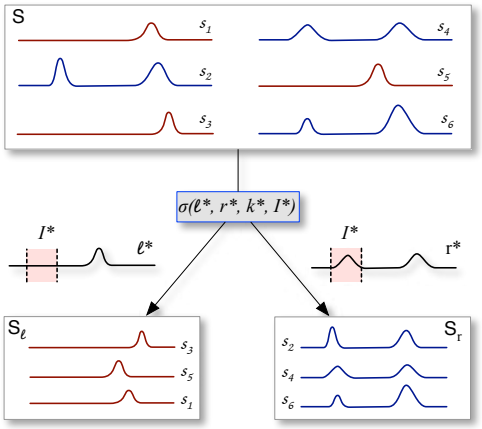

FIGURE 1.1 – Le critère de coupure adaptatif $\sigma(l_*^I, r_*^I, k_*^I, I)$

Classement de nouvelles séries Chaque noeud de l'arbre induit *TSTree* est caractérisé par la coupure optimale apprise $\sigma(l_*, r_*, k_*, I_*)$ définie par les deux séries de référence (l_*, r_*), la valeur optimale k_* de la métrique apprise D_{k_*} ainsi que l'intervalle discriminant I_*. Une nouvelle séries ts est affectée au sous-noeud de gauche si $D_{k_*}(ts, l_*) <= D_{k_*}(ts, r_*)$; elle est affectée au sous-noeud droit sinon. La dissimilarité D_{k_*} entre les séries est évaluée sur la période d'observation I_*. Enfin, la classe de ts est celle de la feuille de l'arbre à laquelle il appartient.

Algorithm 2 *TSSplit*(S, I, α)

1: $(\sigma(l_*^I, r_*^I, k_*^I, I), e_I) =$ *AdaptSplit*(S, I)
2: $(\sigma(l_*, r_*, k_*, I_*), e_{I_*})=$*DichoSplit*$(S, \sigma(l_*^I, r_*^I, k_*^I, I), e_I, \alpha)$
3: **return**$(\sigma(l_*, r_*, k_*, I_*), e_{I_*})$

Algorithm 3 *AdaptSplit*(S, I)

1: $e_* = \infty$
2: **for** k in $[0; 6]$ **do**
3: $\quad (l_k, r_k) = \arg\min_{(l,r)}(GI(\sigma(l, r, k, I)))$
4: $\quad$ **if** $GI(\sigma(l_k, r_k, k, I)) < e_*$ **then**
5: $\quad\quad e_* = GI(\sigma(l_k, r_k, k, I))$
6: $\quad\quad l_*^I = l_k,\ r_*^I = r_k,\ k_*^I = k$
7: $\quad$ **end if**
8: **end for**
9: **return**$(\sigma(l_*^I, r_*^I, k_*^I, I), e_*)$

Algorithm 4 *DichoSplit*$(S, \sigma(l_*^I, r_*^I, k_*^I, I), e_I, \alpha)$

1: $[a, b] = I$
2: $I_L = [a, a + \alpha(b - a)]$
3: $I_R = [b - \alpha(b - a), b]$
4: $(\sigma(l_*^{I_L}, r_*^{I_L}, k_*^{I_L}, I_L), e_{I_L}) = AdaptSplit(S, I_L)$
5: $(\sigma(l_*^{I_R}, r_*^{I_R}, k_*^{I_R}, I_R), e_{I_R}) = AdaptSplit(S, I_R)$
6: **if** $e_I \leq \min(e_{I_L}, e_{I_R})$ **then**
7: **return**$(\sigma(l_*^I, r_*^I, k_*^I, I), e_I)$
8: **else if** $e_{I_L} \leq e_{I_R}$ **then**
9: *DichoSplit*$(S, \sigma(l_*^{I_L}, r_*^{I_L}, k_*^{I_L}, I_L), e_{I_L}, \alpha)$
10: **else**
11: *DichoSplit*$(S, \sigma(l_*^{I_R}, r_*^{I_R}, k_*^{I_R}, I_R), e_{I_R}, \alpha)$
12: **end if**

1.4 Applications

L'approche *TSTree* de classification de séries temporelle par arbre est d'abord mise en application sur quatre jeux de données publiques, CBF (Saito 1994 [144]), CBF-TR (Geurts 2002) [72], CC (Asuncion et al. 2007 [4]), and TWO-PAT (Geurts 2002 [72]) ; utilisés comme base de validation par la majorité des approches concurrentes. Ces jeux portent sur des séries univariées de caractéristiques simples. Par exemple, les séries au sein d'une même classe sont en général de formes globales similaires, chaque classe identifie un profil global distinct, et les profils des séries de classes différentes sont aisément différentiables. Dans les applications réelles, les spécifications des séries dans et entre les classes peuvent être plus complexes. Par exemple, les événements saillants des séries peuvent apparaître avec des délais variables, des effets de tendance ou des variations d'amplitude peuvent entacher les observations (séries non stationnaires), les séries peuvent être de profils globaux dissimilaires au sein des classes tout en partageant des caractéristiques locales communes.

Ainsi pour compléter et élargir le processus de validation de l'approche proposée à des propriétés plus complexes, nous avons considéré cinq jeux de données supplémentaires portant sur trois jeux de données simulés et deux réels. Les données utilisées portent sur des séries pouvant être périodiques, multivariées, incluant des variations d'amplitude et de tendance, et ils introduisent une discrimination des classes de séries dirigée par des événements locaux et non globaux.

Les Tables 1.1 et 1.2 fournissent les principales caractéristiques et propriétés de l'ensemble des jeux de données utilisés. La Table 1.3 résume les différentes configurations de métriques étudiées pour l'induction des arbres de classification *TSSplit*. Un arbre de classification est construit pour chaque jeu de données indiqué dans la Table 1.1 et pour chaque métrique spécifiée dans Table 1.3.

Dans ce chapitre nous nous limitons à quelques résultats succincts obtenus pour un jeu de données DIGITS (Handwritten digits [4]) décrivant des classes de tracés de caractères manuscrits. Une étude comparative et une discussion très complètes des arbres induits et des performances obtenues sur l'ensemble des jeux sont présentées dans Douzal-Chouakria [44].

Name	Source	Sample size	Num. classes	Num. TS/class	TS lengths	Multi. TS.	Real data
CBF	1	300	3	100	128	No	No
CBF-TR	1	300	3	100	128	No	No
CC	2	600	6	100	60	No	No
TWO-PAT	1	400	4	100	128	No	No
LOCAL-DISC	3	300	3	100	128	No	No
CBF-RANGVAR	3	300	3	100	128	No	No
GENES	3	250	5	50	47	No	No
CHAR-TRAJ	2	400	20	20	[100-200]	Yes	Yes
DIGITS	2	220	10	22	110	Yes	Yes

TABLE 1.1 – Description des jeux de données

Name	Time delay	Range vari.	Tend. effect	Local discr.
CBF	No	No	No	No
CBF-TR	Yes	No	No	No
CC	Yes	No	No	No
TWO-PAT	Yes	No	No	No
LOCAL-DISC	Yes	Yes	No	Yes
CBF-RANGVAR	Yes	Yes	No	No
GENES	No	Yes	Yes	No
CHAR-TRAJ	Yes	No	No	No
DIGITS	Yes	No	No	Yes

TABLE 1.2 – Spécifications des séries temporelles

Chaque classe du jeu de données DIGITS identifie un caractère, dont les occurrences correspondent aux différents tracés d'un même caractère, exécutés par différentes personnes. Notons que le ductus (profil global) d'un caractère peut être très différent d'une personne à l'autre. Ainsi, le jeu de données DIGITS constitue une application réelle dans laquelle les profils globaux des séries au sein d'une même classes peuvent être dissimilaires.

La Figure 1.3 visualise l'arbre de classification induit à partir de DIGITS . Introduisant, tout d'abord, quelques éléments d'interprétation de l'arbre obtenu. Chaque noeud de l'arbre est caractérisé par le triplet $(Type, I_*, Class)$ indiquant, respectivement, le type de la métrique apprise, à savoir, si D_{k_*} est fondée essentiellement sur la forme des séries étiquetée "B" pour $k_* \geq 3$, sur les valeurs "V" pour $k_* < 3$, ou constitue un compromis forme-valeurs "BV" pour $k = 3$, l'intervalle discriminant localisé I_*, délimitant les observations base de comparaison des séries, et la classe d'appartenance de la série de référence du noeud.

La recherche dichotomique améliore significativement les performances de l'arbre. Notons que DIGITS portent sur des trajectoires de caractères de profils globaux très distincts au sein d'une même classe (plusieurs personnes transcrivent un même caractère). A partir de l'arbre donné en Figure 1.3, nous pouvons voir que la recherche dichotomique intervient à deux noeuds de l'arbre : pour la séparation des caractères 3 et 5, puis 4 et 9. Ce résultat révèle que les trajectoires des caractères 3 et 5 (resp. 4 et 9) sont très similaires sur la seconde moitié des tracés comme indiqué dans la Figure 1.2. Ainsi, la recherche dichotomique sélectionne, pour la discrimination des classes 3 et 5 (resp. 4 et 9), l'intervalle d'observation portant sur la première moitié des tracés (souligné en rouge dans la Figure 1.3) pour la comparaison des caractères via D_{k_*}

Time delay	Adap. metric	Dicho. search	Behav. cost	Metric
Yes	Yes	Yes	$Cort$	DTW_k^{cort}
	Yes	Yes	Cor	DTW_k^{cor}
	Yes	No	$Cort$	DTW_k^{cort}
	Yes	No	Cor	DTW_k^{cor}
	No	No	-	d_{Dtw}
No	Yes	Yes	$Cort$	DE_k^{cort}
	Yes	Yes	Cor	DE_k^{cor}
	Yes	No	$Cort$	DE_k^{cort}
	Yes	No	Cor	DE_k^{cor}
	No	No	-	d_E

TABLE 1.3 – Les configurations de métriques considérées dans le critère de coupure *TSSplit*

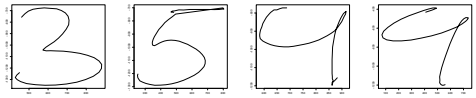

FIGURE 1.2 – Similitude de la seconde moitié des ductus des caractères 3 et 5 (resp. 4 et 9)

1.5 Conclusion

J'ai proposé, en collaboration avec C. Amblard AMA/LIG, une nouvelle approche étendant les arbres de classification à des séries temporelles. Pour cela j'ai introduit un nouveau critère de coupure caractérisé par deux points fondamentaux. D'une part, l'utilisation d'une métrique adaptative, dont les spécifications peuvent changer le long de l'induction de l'arbre, pour une meilleur bipartition des noeuds. D'autre part, par l'implication d'une recherche dichotomique permettant la localisation de sous-séquences discriminantes au niveau de chaque noeud. Les expérimentations menées sur un grand nombre de jeux de données complexes, révèlent *TSTree* comme une approche prometteuse pour la classification de séries temporelles complexes, dont les performances demeurent très compétitives face à celles des approches concurrentes.

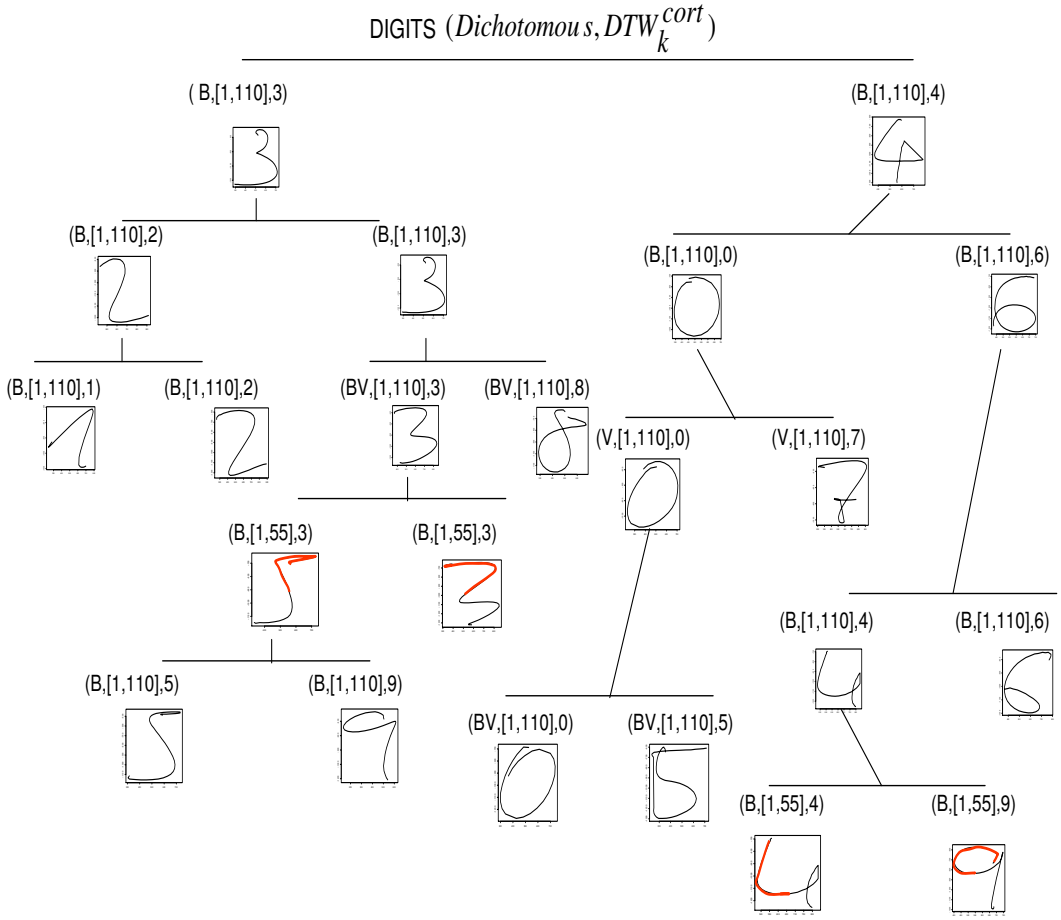

FIGURE 1.3 – Arbre de classification associé aux données DIGITS

Perspectives

L'apprentissage de couplages pour la discrimination de séries temporelles, discuté dans le chapitre 5, peut s'inscrire dans le cadre des approches d'alignement de séquences multiples largement étudiées en biologie. Notons cependant deux points distinguant principalement notre approche : elle s'adresse à un ensemble de séries réparties en classes et élargit la recherche d'alignements à des couplages en vue de révéler des caractéristiques partagées dans les classes et différentielles entre les classes.

Une littérature très dense, émanant en particulier du domaine de la biologie, est consacrée au problème d'alignement de séquences multiples [80, 14, 113, 56] et qui pose à ce jour de nombreux défis à relever [100]. Bien que les séries temporelles soient omniprésentes, l'intérêt pour les problèmes d'alignement de séries temporelles multiples est assez récent et les quelques travaux traitant de ce sujet [1, 83, 136, 135] révèlent l'importance de telles approches et les perspectives qu'elles ouvrent pour l'extraction de prototypes pertinents et la classification de séries temporelles complexes.

Dans ce qui suit, nous rappelons les principales approches d'alignement de séquences multiples et verrous sous-jacents. Nous situerons la méthode d'apprentissage de couplages de séries multiples proposée et discuterons quelques verrous à relever et direction de recherche à mener dans nos travaux futurs.

Le problème d'alignement de séquences multiples n'est pas qu'un exercice technique, il permet 1) de révéler des caractéristiques communes souvent cachées ou clairsemées à partir de plusieurs séquences, 2) de limiter la comparaison de séries multiples aux caractéristiques communes et 3) d'extraire et de caractériser, à partir des attributs partagés, des familles de séquences. En biologie, par exemple, l'alignement de séquences multiples d'ADN, ARN ou de protéines constitue un des moyens fondamentaux pour l'extraction et la représentation de connaissances biologique à partir de plusieurs séquences. La comparaison de séquences biologiques ainsi alignées permet, par exemple, de révéler un héritage génomique commun ou l'implication dans une même fonction biologique. Les approches d'alignement de séquences multiples, de complexité drastique, sont souvent basées sur des méthodes heuristiques plutôt que sur des approches d'optimisation globale visant à identifier un alignement optimal entre plusieurs séquences. On distingue principalement trois approches majeures d'alignements de séquences multiples. La première catégorie de méthodes utilisent des techniques de **programmation dynamique**, la recherche d'un alignement entre plusieurs séquences généralise l'alignement pair-à-pair en recherchant dans un hypercube de dimension n (n étant le nombre de séquences à aligner) le chemin optimisant au mieux une fonction de coût. Cependant la complexité de recherche d'un tel chemin rend ces approches non fonctionnelles dans la pratique [160, 89, 56]. Une alternative à cette démarche, procède d'abord à l'alignement

pair-à-pair des séquences, l'espace de recherche de l'alignement multiple est alors réduit au voisinage du croisement des alignements pair-à-pair obtenus [14, 113]. Les approches **progressives** [126, 86, 156, 131] permettent de contourner la complexité des techniques utilisant la programmation dynamique. Dans un premier temps, les séquences sont organisées hiérarchiquement ou en arbre, une classification hiérarchique ascendante est souvent utilisée pour cet effet. Dans un second temps, on procède progressivement à l'alignement pair-à-pair des séquences les plus proches vers les plus éloignées. Le principal inconvénient de ces approches est leur dépendance de l'ordre des séquences alignés, un mauvais alignement initial se répercute à l'ensemble des alignements multiples construits. Une troisième catégorie d'approches dites **itératives** [77, 10, 54], permet la construction d'alignements multiples de manière itérative où, contrairement aux méthodes progressives, les alignements pair-à-pair sont affinés à chaque intégration d'une nouvelle séquence. Les approches itératives permettent d'obtenir des alignements multiples plus précis que ceux construits via des approches progressives.

Notons que la plus part des approches d'alignements de séquences multiples décrites ci-dessus sont basées sur des appariements pair-à-pair **globaux** incluant toutes les observations des séquences. De nombreuses applications renforcent l'idée qu'un ensemble de séquences partagent plus souvent un ensemble de sous-segments ou motifs plutôt que l'ensemble de leurs observations, que ces motifs peuvent apparaître dans un ordre non équivalent dans les séries, rendant inapplicables les approches d'alignements classiques [126], [98], [96], [97]. Ainsi, a vu le jour de nombreuses approches fondées sur la recherche d'appariement **locaux** en vue de révéler des régions homologues caractérisant l'ensemble des séquences. Les premières approches d'alignements locaux sont fondées sur l'utilisation de fonctions de pénalisation des régions non similaires (*gap penality*). De nouvelles variantes, s'affranchissent de la pénalisation de régions non appariées en utilisant, par exemple, des matrices à pixels et des techniques de filtrage et de seuillage pour limiter la complexité des régions alignées [10], [124], ou en procédant à des alignements croisés pour de motifs pouvant apparaître dans un ordre quelconque dans des séquences de proteines [126], [98], [96], [97]. Ces approches constituent aujourd'hui les techniques les plus à la pointe pour l'alignement de séquences multiples.

L'apprentissage de couplages pour la discrimination de séries temporelles proposé au chapitre 5 est une approche itérative, fondée sur des couplages locaux entre plusieurs séries réparties en classes, afin de révéler des caractéristiques communes dans les classes et différentielles entre les classes. Les couplages sont formalisés par des matrices de poids, généralisant les matrices à pixels utilisées dans Brudno et al. (2004) [10] et Morgenstern (2004) [124], dont les valeurs sont renforcées ou pénalisées selon la similarité des régions couplées. Enfin, nous rejoignions les travaux proposées par Kelil et al. [98], [96] et [97] en couplant des saillances indépendamment de leur ordre d'apparition.

En revanche, plusieurs points distinguent l'approche proposée des approches d'alignement de séquences multiples citées ci-dessus. La méthode proposée généralise la recherche d'alignements à des couplages plus large. Elle permet l'apprentissage de couplages pouvant être locaux ou globaux sans recours aux couplages pair-à-pair. Enfin, l'approche proposée vise l'apprentissage de couplages discriminant plusieurs ensembles de séries.

Verrous et perspectives

- **Réduction de la complexité calculatoire**
  Les différentes familles d'algorithmes décrites ci-dessus pour l'alignement de séquences multiples ou le couplage de séries multiples restent aujourd'hui limitées par leur complexité calculatoire. Cette limite se traduit par des algorithmes lents et inapplicables sur de grands volumes de données.
  Ainsi, la suite du travail présenté au chapitre 5 vise à accélérer l'algorithme d'apprentissage de couplages par la réduction de sa complexité calculatoire. La complexité de notre algorithme est principalement liée à l'utilisation d'un couplage complet dans la phase d'initialisation. Dans le même esprit que l'algorithme proposé dans Brudno et al. (2004) [10], nous proposons de déployer notre algorithme en deux phases. Dans la première phase, l'idée générale consiste à segmenter chaque série temporelle afin de localiser un ensemble de sous-segments saillants de taille réduite résumant la série, qu'on appellera **segments d'ancrage** de la série. La matrice de couplage initiale connectera les couples de segments d'ancrage des diverses séries. Notons que les segments d'ancrage sont connectés indépendamment de leur ordre d'apparition car, comme discuté dans le chapitre 5, un événement saillant peut apparaître à tout moment de l'observation d'une série. Par ailleurs, comme les segments d'ancrage correspondent à des régions réduites homogènes, ils seront alignés de manière euclidienne. Si l'on suppose une moyenne de k segments d'ancrage tous de même taille $T_{\min}$ dans chaque série, ce nombre k de segments est nécessairement plus faible que $T/T_{\min}$ i.e. $k \times T_{\min} \leq T$. Les segments d'ancrage étant couplés selon un lien euclidien, le nombre d'observations couplées décroît de T^2à $k^2 T_{\min}$. En particulier, nous pouvons borner ce nombre par $T^2/T_{\min}$. Plus la longueur des segments d'ancrage augmente, plus la complexité calculatoire diminue, le cas extrême étant le cas $T_{\min} = T$, où toutes les observations sont liées selon un couplage euclidien.

- **Apprentissage de couplages pour la classification non supervisée de séries temporelles**
  La classification de séries temporelles est souvent réalisée par des algorithmes classiques basés sur des distances conventionnelles, la plus fréquentes d'entres elles étant la Dynamique Time Warping (DTW). Cependant, comme discuté au chapitre 4, l'alignement opéré par la DTW classique demeure limité face à des séries temporelles de structures complexes. Par ailleurs, rappelons qu'étant donné un couplage appris entre un ensemble de séries, une dissimilarité d_M peut en être dérivée, elle permet de comparer les séries sur la base de caractéristiques communes révélées par le couplage. Ainsi, le problème de classification non supervisée d'un ensemble de séries $S_1, ..., S_N$ en K classes peut être exprimé par le problème d'optimisation suivant :

$$\min_M \left(\sum_{k=1}^{K} \sum_{l \in C_k} d_M(S_l, c_k)\right)$$

  M étant le couplage à apprendre afin de minimiser l'écart entre les séries S_l et le profil moyen c_k de leur classe d'assignation. Ce problème peut être résolu par l'approche suivante : 1) Choisir aléatoirement K séries $c_1, ..., c_K$, 2) Apprendre le couplage M minimisant l'écart entre les séries $S_1, ..., S_N$ et les centres c_k. Notons que ce problème constitue une restriction de LearnWitAlig à K séries, 3) Assigner chaque série S_l au

centre c_k le plus proche au sens de la dissimilarité d_M induite par le couplage M en cours, 4) Estimer pour chaque classe C_k son nouveau centre, il correspond au profil moyen de c_k tenant compte de son couplage aux séries de la classe C_k, 5) Réitérer les étapes 2) à 4) jusqu'à stabilisation de la partition. La convergence de l'algorithme, l'initialisation du couplage, ainsi que la complexité calculatoire constituent des points cruciaux à étudier.

– **Factorisation de matrices pour l'analyse et l'extraction de connaissances à partir de systèmes dynamiques, en-ligne et à grande échelle**

Plusieurs spécifications rendent complexe l'analyse des données dynamiques. En particulier, le processus de génération des données peut impliquer plusieurs processus cachés, la structure des dépendances sous-jacente aux données peut être évolutive et de causes multiples, et les observations constituent de grands volumes de données, pouvant arriver selon un processus continu. Extraire des connaissances pertinentes à partir de telles données nécessite de révéler et de quantifier les éventuels facteurs impliqués dans la génération des données, de mettre en œuvre des approches subtiles d'identification des dépendances évolutives, et l'utilisation d'approches de complexité raisonnable prenant en compte des flux de données en grands volumes.

Des travaux récents montrent l'efficacité des approches par décomposition de matrices pour l'apprentissage ou la fouille de données complexes. Ces méthodes permettent en particulier de répondre à des problèmes de séparation de sources, de révéler des liens complexes entre des groupes d'individus, de variables, ou à restreindre la représentation des données à des structures compactes [109], [90], [87]. Plusieurs vues et interprétations peuvent être formulées à partir d'une matrice de données ainsi décomposée. Les données peuvent être exprimées en termes de facteurs cachés, être représentées dans un nouvel espace de dimension réduite, ou exprimées sous forme de graphe révélant les liens entre individus, facteurs latents et variables .

Nes travaux futurs visent entre autres à explorer l'analyse des données dynamiques par décomposition de matrices (factorisation de matrices non-négatives, codage sparse). En particulier, il peut être intéressant de prendre en compte l'ensemble des vues ou connaissances révélées par la décomposition pour l'enrichissement des modèles mis en œuvre. Cette étude sera mise en application pour la désagrégation non-invasive de la consommation d'énergie en vue du contrôle et de la prédiction de la consommation dans un habitat intelligent (Projet Schneider) ou pour la supervision de la consommation électrique des ménages. Une seconde application directe de ces travaux porte sur la détection précoce de thèmes émergents dans des médias sociaux tel que Twitter (thèse Cifre BestOfMedia).

[15, 29, 22, 30, 24, 21, 18, 27, 20, 25, 23, 19, 38, 42, 41, 40, 39, 45, 43, 64, 46, 48, 53, 51, 50, 49, 52, 47, 28, 44, 62, 63, 61, 76, 75, 110, 103, 140]

Annexe

1. Liste des publications à l'appui de l'HDR

Articles de revues internationales avec comité de lecture

- A. Douzal-Chouakria, C. Amblard (2012). Classification trees for time series. Pattern Recognition Journal, 45, 3, 1076-1091. Elsevier.

- A. Douzal-Chouakria, A. Diallo, F. Giroud (2010). A random-periods model for the comparison of a metrics efficiency to classify cell-cycle expressed genes. Pattern Recognition Letters. 31, 1601-1617. Elsevier.

- A. Douzal-Chouakria, A. Diallo, F. Giroud (2009). Adaptive clustering for time series : application for identifying cell cycle expressed genes. Computational Statistics and Data Analysis, 53 (4), 1414-1426. Elsevier.

- A. Douzal-Chouakria, P.N. Nagabhushan (2007). Adaptive dissimilarity index for measuring time series proximity. Advances in Data Analysis and Classification Journal. 1, 5-21, Springer Berlin / Heidelberg.

Articles de revues nationales avec comité de lecture

- A. Diallo, A. Douzal-Chouakria, F.Giroud (2008). Classification adaptative de séries temporelles : application à l'identification des gènes exprimés au cours du cycle cellulaire. Revue des Nouvelles Technologies de l'Information (RNTI-E-11), 487-498, Cépaduès.

Chapitres d'ouvrages collectif

- A. Douzal-Chouakria, A. Diallo, F. Giroud (2007). Adaptive dissimilarity index for Gene Expression Profiles Classification. In : Selected Contributions in Data Analysis and Classification, Series : Studies in Classification, Data Analysis, and Knowledge Organization, Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F. (Eds.). XIII, 483-494, Springer Berlin Heidelberg.

- A. Chouakria-Douzal and P.N. Nagabhushan (2006). Improved Fréchet Distance for Time Series. In : V. Batagelj, H.-H. Bock, A. Ferligoj, A. Ziberna (eds.) Data Science

and Classification, 13-20, Springer, ISBN : 3-540-34415-2.

Conférences internationales avec comité de lecture et publication des actes

- C. Frambourg, A. Douzal-Chouakria, E. Gaussier, J. Demongeot (2011). Learning time series dissimilarities, GFKL'2011 (conférence invité).

- A Diallo, A Douzal-Chouakria, Françoise Giroud (2009). Which Distance for the Identification and the Differentiation of cell-cycle Expressed Genes ?. 8th International Symposium on Intelligent Data Analysis. 273-284. Springer-Verlag.

- A. Douzal-Chouakria , A. Diallo, F. Giroud (2007). Adaptive clustering for time series : application for identifying cell cycle expressed genes. International Association for Statistical Computing, Statistics for Data Mining, Learning and Knowledge Extraction (IASC'07), Aveiro, Portugal.

- A. Douzal-Chouakria , N. Hammami, C. Garbay (2007). Local Factorial Analysis of Time Series. 56th Session of the International Statistical Institute (ISI'07), Lisboa, Portugal.

- G. Rizk, A. Douzal-Chouakria , C. Amblard (2007). Temporal Decision Trees. 56th Session of the International Statistical Institute (ISI'07), Lisboa, Portugal.

- K. Pradeep, P.N. Nagabhushan, A. Douzal-Chouakria (2006). WaveSim and Adaptive Transform for subsequence Time series Clustering. IEEE 9th International Conference on Information Technology (ICIT'06), 197-202.

- A. Chouakria-Douzal (2003). Compression Technique Preserving Correlations of a Multivariate Temporal Sequence. In : M.R. Berthold, H-J Lenz, E. Bradley, R. Kruse, C. Borgelt (eds.) Advances in Intelligent Data Analysis, V, 566-577, Springer, ISBN : 3-540-40813-4.

- J. Demongeot, B. Beaucamps, T. Chaperon, A. Douzal-Chouakria, T. Faraut, M. Simonet and A. Simonet (1999). Estimating joint probabilities in the context of probabilistic management of querying and integrity in a knowledge and data base. In : Conditional Independence Structures and Graphical Models, F. Matús and M. Studeny eds., Field Institute, Toronto, 21-23.

Conférences internationales avec comité de lecture à publications courtes

- F. Giroud, A. Diallo, A. Douzal-Chouakria (2007). Identification of cell cycle expressed genes. Workshop Towards Systems Biology, Grenoble.

- F. Giroud, A. Diallo, A. Douzal-Chouakria (2007). A new approach for molecular dynamic network analysis. Réaumur Meeting, Grenoble.

- F. Giroud, A. Diallo, A. Douzal-Chouakria (2007). Identification of cell cycle expressed genes : a new approach for molecular dynamic network analysis. Workshop Towards Systems Biology 2007, Grenoble, France.

- A. Douzal-Chouakria , A. Diallo and F. Giroud (2006). Adaptive dissimilarity index for genes expression profiles classification. Integrative Post Genomics Conference. Lyon.

- A. Chouakria-Douzal (2005). On the distance measure between time series. International Association for Statistical Computing (IASC'05), Chypres.

Conférences nationales avec comité de lecture et publication des actes

- A. Douzal-Chouakria, C. Amblard (2012). Adaptive split test for multivariate time series classification trees. Conférence d'Apprentissage Cap'2012.

- A. Diallo, A. Douzal-Chouakria, F. Giroud (2011). Un modèle génératif pour la comparaison de métriques en classification de profils d'expression de gènes. Conférence d'Apprentissage Cap'2011, 135-150, Edition Publibook.

- C. Frambourg, A. Douzal-Chouakria, E. Gaussier, J. Demongeot (2011). Apprentissage de couplages pour la discrimination de séries temporelles. Cap'2011, 151-166, Edition Publibook.
- A. Diallo, A. Douzal-Chouakria, F. Giroud (2009). Comparaisons et évaluation de métriques pour la classification de profils d'expression de gènes. 65-68. SFC'09. Grenoble.

- C. Frambourg, A. Douzal-Chouakria, J. Demongeot (2009). Moran and Geary indices for multivariate time series exploratory analysis. 177-180. SFC'09. Grenoble.

- A. Douzal-Chouakria (2006). Réduction de la dimension de séries temporelles multidimensionnelles par extraction des tendances locales. Extraction et Gestion des Connaissances (EGC'06), Atelier Fouille de données temporelles, Lille.

2. Animations scientifiques

Relecteur dans des revues internationales

- Pattern Recognition Letters,
- Advances in Data Analysis and Classification Journal,
- Journal of Classification,
- Computation Statistics,
- Computation Statistics and Data Analysis,
- Statistical Analysis and Data Mining,
- Artificial Intelligence Research

Membre de comités de programme

- International Conference on Advanced Computing & communication,
- International Conference of Pattern Recognition and Machine Intelligence,
- International Federation of Classification Society,
- Société Francophone de Classification

3. Co-encadrements de thèses

- sep 2011 : F. Kawala, en co-direction avec E. Gaussier. Détection et prédiction de thèmes émergents dans les médias sociaux (début : sep 2011, thèse Cifre en partenariat avec BestOfMedias, encadrement à 80%).

- oct 2008 -nov. 2012 : C. Frambourg en co-direction avec J. Demongeot (Timc-Imag). Apprentissage de couplages pour la discrimination de séries temporelles (début : oct 2008, soutenance prévue : déc 2012, financement bourse MESR, encadrement à 90%).

- oct 2005 - juin 2010 : A. Diallo en co-direction avec F. Giroud (Rfmq, Timc-Imag). La classification de séries temporelles : application à l'identification et à la différenciation de profils d'expression de gènes (début : octobre 2005, soutenue : juin 2010, financement partielle, encadrement à 80%).

4. Encadrements de Master 2R

- 2008 : C. Frambourg (M1 de l'Institut Fourier et agrégé de mathématiques), analyse exploratoire de données temporelles, stage de M2R MIMB, EDISCE.

- 2007 : G. Rizk (Ingénieur Ensimag), les arbres de décision temporels (stage de M2R MIMB, EDISCE).

- 2005 : A. Diallo (M2R Recherche Opérationnelle, Grenoble-INP), classification de données d'expression de gènes, stage de M2R MIMB, EDISCE.

- 2004 : P. Ravel (Ingénieur SupAéro), analyse de données physiologiques pour la détection de limitation de débit respiratoire, stage de M2R MIMB, EDISCE.

- 2003 : L. Collonge (TIS, Polytech Grenoble), les mesures de proximités entre séquences, stage de M2R MIMB, EDISCE.

5. Projets et collaborations

Projets scientifiques

- 2012 : Projet ANR DYNOTEP (SVSE6)
  Le projet DYNOTEP (durée : 48 mois, montant total demandé : 585K euros) est développé en partenariat avec le CEA. La coordination est assurée par L. Aubry responsable de l'équipe Odycell du CEA. Le projet porte sur l'analyse de la dynamique des cellules endocytaires. Dynotep comprend trois partenaires : les équipes Odycell et EDyp du CEA et l'équipe AMA/LIG. Je participe au projet à hauteur de 25% et assure la responsabilité locale du projet (Montant local : 106keuros), notre rôle porte l'apprentissage et l'analyse de données dynamiques.

- 2012 : Projet Schneider Innovation pour l'Efficacité Energétique
  Le projet porte sur le développement de capteurs virtuels fondés sur des solutions issues du domaine de l'apprentissage pour la prédiction de données de capteurs (régression temporelle supervisée ou semi-supervisée), en vue du contrôle de données de capteurs existants ou le remplacement de capteurs réels coûteux. Ce projet d'étude et de conseil (Montant : 10k euros) implique 3 membres de l'équipe AMA (G. Bisson (40%), E. Gaussier (20%) et moi même (40%)).

- 2012 : Projet européen BioASQ (Specific Support Action)
  Le projet BIOASQ (durée : 30 mois, Montant : 1M euros) implique 6 partenaires dont l'équipe AMA/LIG. Ce projet vise la mise en place d'un challenge multi-tâches, dont l'objectif est la construction d'un système de question/réponse pour l'évaluation d'un système de catégorisation à large échelle. La responsabilité locale du projet est assurée par E. Gaussier (Montant AMA : 175 k euros). Ma participation est à hauteur de 10%.

- 2011 : Projet BestOfMedia
  Le projet BestOfMedia (durée 36 mois, Montant : 30k euros) dont je suis responsable porte sur la détection et la prédiction de thèmes émergents dans les médias sociaux.

- 2006 - 2008 : Projet-THEMIS en épidémiologie
  Le projet dont j'ai assuré la responsabilité, porte sur une analyse exploratoire et statistique de données médicales, ainsi qu'une formation en analyse de la survie en partenariat avec l'entreprise en épidémiologie THEMIS (Montant : 10k euros).

Collaborations scientifiques

- Depuis 2009 : L. Billard, Professeur à l'université de Géorgie (USA), collaborations portant sur l'analyse exploratoire de données intervalles et temporelles, cette collabo-

ration a donné lieu à 1 publication dans une revue et une en cours de soumission.

- 2005 - 2007 : P.N. Nagabhushan, Professeur à l'université de Mysore, collaboration autour des mesures de proximités et de distance entre des séries temporelles. Cette collaboration a donné lieu à plusieurs publications dont principalement.

- 2001 - 2003 : Collaboration avec l'équipe PRETA du TIMC-IMAG, pour l'analyse de données de monitoring pour la détection de limitation de débit respiratoire chez des patients atteints de l'apnée du sommeil.

- 2005 - 2010 : F. Giroud (RFMQ du TIMC-IMAG), collaboration autour de l'analyse de données génomiques pour l'identification et la caractérisation de profils d'expression de gènes.

- 1999 - 2000 : J. Demongeot (TIMC), collaboration portant sur l'estimation de probabilités jointes dans un contexte de gestion de l'intégrité des requêtes dans une base de connaissance.

Bibliographie

[1] W.H. Abdulla, D. Chow, and G. Sin. Cross-words ref- erence template for dtw-based speech recognition sys- tems. In *Proc. TENCON*, volume 2, pages 1576–1579, 2003.

[2] Z. Abraham and P. Tan. An integrated framework for simultaneous classification and regression of time-series data. In *SIAM International Conference on Data Mining*, pages 653–664, 2010.

[3] A. Anagnostopoulos, M. Vlachos, M. Hadjieleftheriou, E.J. Keogh, and P.S. Yu. Global distance-based segmentation of trajectories. In *ACM SIGKDD*, pages 34–43., 2006.

[4] A. Asuncion and D.J. Newman. UCI, machine learning repository, 2007.

[5] S. Balakrishnan and D. Madigan. Decision trees for functional variables. In *International Conference on Data Mining*, pages 798 –802, 2006.

[6] T. A. Banet. Local and partial correspondence analysis : application to the analysis of electoral data. *Computational statistics quarterly*, 2 :89–103, 1988.

[7] T. A. Banet and L. Lebart. Local and partial principal component analysis and correspondence analysis. *Computational Statistics*, pages 113–118, 1984.

[8] Z. Bar-Joseph, G.K. Gerber, D.K. Gifford, T. Jaakkola, and I. Simon. Continuous representations of time-series gene expression data. *Journal of Computational Biology*, 10(3) :341–356, 2003.

[9] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and regression trees*. 1984.

[10] M. Brudno, M. Chapman, B. Göttgens, S. Batzoglou, and B. Morgenstern. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics*, 4(66), 2003.

[11] F. Cabestaing, T.M. Vaughan, D.J. McFarland, and J.R. Wolpaw. Classification of evoked potentials by pearsonís correlation in a brain-computer interface. *Modelling C Automatic Control (theory and applications)*, 67 :156–166, 2007.

[12] J. Caiado, N. Crato, and D. Pena. A periodogram-based metric for time series classification. *Computational Statistics and Data Analysis*, 50 :2668–2684, 2006.

[13] L. J. Cao. Support vector machine experts for time series prediction. *Neurocomputing*, 51(321–339), 2003.

[14] H. Carrillo and D.J. Lipman. The multiple sequence alignment problem in biology. *SIAM Journal of Applied Mathematics*, 48(5) :1073–1082, 1988.

[15] P. Cazes, A. Chouakria, E. Diday, and Y. Schektman. Extension de l'analyse en composantes principales à des données intervalles. *Revue de statistiques appliquées*, XLV(3) :5–24, 1997.

[16] C. Chatfield. *The analysis of time series.* Chapman and Hall, 1996.

[17] C. Chatfield. *Time-series Forecasting.* Champman and Hall/CRC, 2000.

[18] A. Chouakria. *Extension de l'analyse en composantes principales à des données de type intervalle. Thèse,Paris IX Dauphine.* PhD thesis, Paris IX Dauphine, INRIA-Rocquencourt, 1998.

[19] A. Chouakria, P. Cazes, and E. Diday. *Analysis of Symbolic Data Exploratory Methods for Extracting Statistical Information from Complex Data*, chapter Symbolic Principal component Analysis, pages 200–212. Springer-Verlag. Berlin, bock, h.h. and diday, e. edition, 2000.

[20] A. Chouakria, E. Diday, and P. Cazes. Extension de l'analyse factorielle des correspondances multiples à des données de type intervalle et de type ensemble. In *Société Francophone de Classification (SFC'95)*, Namur, 1995.

[21] A. Chouakria, E. Diday, and P. Cazes. Extension of the principal component analysis to interval data. In *New Techniques and Technologies for Statistics*, Bonn, 1995.

[22] A. Chouakria, E. Diday, and P. Cazes. Généralisation, en vue d'une acm du découpage en classes d'effectifs égaux à des variables de type intervalle. Technical report, Association de la Statistique et de ses Utilisateurs (ASU'97), Carcassonne, 1997.

[23] A. Chouakria, E. Diday, and P. Cazes. *Advances in Data Science and Classification*, chapter Vertices Principal Components Analysis with an Improved Factorial Representation, pages 397–402. ISBN : 3-540-6441-8. Springer Verlag, a. rizzi, and m. vichi, and h.h. bock edition, 1998.

[24] A. Chouakria, E. Diday, and P. Cazes. An improved factorial representation of symbolic objects. In *Knowledge Extraction from Statistical Data (KESDA'98)*, Luxembourg, 1998.

[25] A. Chouakria, R. Verde, E. Diday, and P. Cazes. Généralisation de l'analyse factorielle des correspondances multiples à des objets symboliques. In *Société Francophone de Classification (SFC'96)*, Vannes, 1996.

[26] Ahlame Douzal Chouakria and Panduranga Naidu Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1(1) :5–21, 2007.

[27] A. Chouakria-Douzal. Réduction de dimension par extraction des tendances locales composant une séquence temporelle multivariée. Technical Report RR-IMAG 1053-I, UJF/CNRS/TIMC, 2002.

[28] A. Chouakria-Douzal. Compression technique preserving correlations of a multivariate temporal sequence. In M. R. Berthold, H-J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, editors, *Advances in Intelligent Data Analysis*, volume V, pages 566–577. Springer, 2003.

[29] A. Chouakria Douzal. On the distance measure between time series. In *International Association for Statistical Computing (IASC'05)*, Cyprus, 2005.

[30] A. Chouakria-Douzal and P. Nagabhushan. Improved fréchet distance for time series. In *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 13–20. 2006.

[31] C.S.J. Chu. Time series segmentation : a sliding window approach. i. *Information Sciences*, 85(1-3) :147–173, 1995.

[32] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K. : Cambridge Univ. Press, 2000.

[33] M. Cuturi. Fast global alignment kernels. In *International Conference on Machine Learning*, 2011.

[34] M. Cuturi and K. Fukumizu. Kernels on structured objects through nested histograms. In J. Schölkopf, B. Platt and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.

[35] M. Cuturi and J.-P. Vert. The context-tree kernel for strings. *Neural Networks*, 18(8), 2005.

[36] M. Cuturi, J.-P. Vert, Øystein Birkenes, and T. Matsui. A kernel for time series based on global alignments. In *the International Conference on Acoustics, Speech and Signal Processing*, volume 11, pages 413–416, 2007.

[37] G. Das, D. Gunopulos, and H. Mannila. Finding similar time series. In *Proc. of the Principles of Knowledge Discovery and Data Mining*, pages 454–456, 1997.

[38] J. Demongeot, B. Beaucamps, T. Chaperon, A. Douzal-Chouakria, T. Faraut, M. Simonet, and A. Simonet. Estimating joint probabilities in the context of probabilistic management of querying and integrity in a knowledge database. In F. Matús and M. Studeny, editors, *Conditional Independence Structures and Graphical Models*, pages 21–23, Field Institute, Toronto, 1999.

[39] A. Diallo, A. Douzal-Chouakria, and F. Giroud. Classification adaptative de séries temporelles : application à l'identification des gènes exprimés au cours du cycle cellulaire. *Revue des Nouvelles Technologies de l'Information*, 11 :487–498, 2008.

[40] A. Diallo, A. Douzal-Chouakria, and F. Giroud. Comparaisons et évaluation de métriques pour la classification de profils d'expression de gènes. In *Société Francophone de Classification*, number 65-68, Grenoble, 2009.

[41] A. Diallo, A. Douzal-Chouakria, and F. Giroud. Which distance for the identification and the differentiation of cell-cycle expressed genes ? In *Advances in Intelligent Data Analysis VIII*, volume 5772 of *Lecture Notes in Computer Science*, pages 273–284, 2009.

[42] A. Diallo, A. Douzal-Chouakria, and F. Giroud. Un modèle génératif pour la comparaison de métriques en classification de profils d 'expression de gènes. In *Conférence d'Apprentissage automatique*, pages 135–150. Edition Publibook., 2011.

[43] A. Douzal-Chouakria. Réduction de la dimension de séries temporelles multidimensionnelles par extraction des tendances locales. In *Extraction et Gestion des Connaissances, Atelier Fouille de données temporelles*, Lille, 2006.

[44] A. Douzal-Chouakria and C. Amblard. Classification trees for time series. *Pattern Recognition*, 45(3) :1076–1091, 2012.

[45] A. Douzal-Chouakria, L. Billard, and E. Diday. Principal component analysis for interval-valued observations. *Statistical Analysis and Data Mining. Wiley*, 4(2) :229–246, 2011.

[46] A. Douzal-Chouakria, A. Diallo, and F. Giroud. Adaptive dissimilarity index for genes expression profiles classification. In *Integrative Post Genomics Conference*, Lyon, 2006.

[47] A. Douzal-Chouakria, A. Diallo, and F. Giroud. Adaptive clustering for time series. In *Statistics for Data Mining, Learning and Knowledge Extraction*, 2007.

[48] A. Douzal-Chouakria, A. Diallo, and F. Giroud. Adaptive clustering for time series : application for identifying cell cycle expressed genes. In *IASC'07, International Association for Statistical Computing, Statistics for Data Mining, Learning and Knowledge Extraction*, Aveiro, Portugal, 2007.

[49] A. Douzal-Chouakria, A. Diallo, and F. Giroud. *Adaptive dissimilarity index for Gene Expression Profiles Classification*, volume 13, pages 483–494. Brito, P. and Bertrand, P. and Cucumel, G. and De Carvalho, F., 2007.

[50] A. Douzal-Chouakria, A. Diallo, and F. Giroud. Adaptive clustering for time series : application for identifying cell cycle expressed genes. *Computational Statistics and Data Analysis*, 53(4) :1414–1426, 2009.

[51] A. Douzal-Chouakria, A. Diallo, and F. Giroud. A random-periods model for the comparison of a metrics efficiency to classify cell-cycle expressed genes. *Pattern Recognition Letters*, 31 :1601–1617, 2010.

[52] A. Douzal-Chouakria, N. Hammami, and C. Garbay. Local factorial analysis of time series. In *56th Session of the International Statistical Institute*, 2007.

[53] A. Douzal-Chouakria and P.N. Nagabhushan. Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification Journal.*, 1(1) :5–21, 2007.

[54] R.C. Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5) :1792–1797, 2004.

[55] M.B. Eisen and P.O. Brown. Dna arrays for analysis of gene expression. *Methods Enzymol*, 303 :179–205, 1999.

[56] I. Elias. Settling the intractability of multiple alignment. *J. Comput. Biol Comput Biol*, 13(7) :1323–1339, 2006.

[57] J. Ernst, GJ Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21 :159–168, 2005.

[58] J. Fan and Q. Yao. *Nonlinear Time Series : Nonparametric and Parametric Methods.* Springer, 2003.

[59] C.L. Fancourt and J.C. Principe. Competitive principal component analysis for locally stationary time series. *IEEE Transactions on Signal Processing*, 46(11) :3068–3082, 1997.

[60] L.J. Fitzgibbon, D.L. Dowe, and L. Allison. Change-point estimation using new minimum message length approximations. In *the Seventh Pacific Rim International Conference on Artificial Intelligence : Trends in Artificial Intelligence*, number 244–254, 2002.

[61] C. Frambourg, A. Douzal-Chouakria, and J. Demongeot. Moran and geary indices for multivariate time series exploratory analysis. In *Société Francophone de Classification*, pages 177–180, Grenoble, 2009.

[62] C. Frambourg, A. Douzal-Chouakria, and E. Gaussier. Learning temporal matchings for time series discrimination. *Soumission à Machine Learning*, 2012.

[63] C. Frambourg, A. Douzal-Chouakria, E. Gaussier, and J. Demongeot. Apprentissage de couplages pour la discrimination de séries temporelles. In *Conférence d'Apprentissage automatique*, pages 151–166. Edition Publibook, 2011.

[64] C. Frambourg, A. Douzal-Chouakria, Eric Gaussier, and Jacques Demongeot. Learning time series dissimilarities. In *Joint Conference of the German Classification Society (GfKl'11)*, Frankfurt, 2011.

[65] M. Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Mathematico di Palermo*, 22 :1–74, 1906.

[66] T.C. Fu, F.L. Chung, and C.M. Ng. Financial time series segmentation based on specialized binary tree representation. In *International Conference on Data Mining*, pages 3–9, 2006.

[67] L. A. Garcia-Escudero and A. Gordaliza. A proposal for robust curve clustering. *Journal of Classification*, 22 :185–201, 2005.

[68] R. Gaudin and N. Nicoloyannis. An adaptable time warping distance for time series learning. In *the 5th International Conference on Machine Learning and Applictions.*, pages 213–218, 2006.

[69] R.C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3) :115–145, 1954.

[70] A. Getis and J.K. Ord. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3) :189–206, 1992.

[71] P Geurts. Pattern extraction for time series classification. In *LNCS Principles of Data Mining and Knowledge Discovery*, pages 115–127, 2001.

[72] P. Geurts. *Contributions to decision tree induction : bias/variance tradeoff and time series classification.* PhD thesis, Department of Electrical Engineering, University of Liege, Belgium., 2002.

[73] P. Geurts and L. Wehenkel. Segment and combine approach for non-parametric time-series classification. In *PKDD*, number 478-485, 2005.

[74] A. B. Geva. Scalenet-multiscale neural-network architecture for time series prediction. *IEEE Transactions on Neural Networks*, 9(6) :1471–1482, 1998.

[75] F. Giroud, A. Diallo, and A. Douzal-Chouakria. Identification of cell cycle expressed genes : a new approach for molecular dynamic network analysis. In *Workshop Towards Systems Biology*, Grenoble, 2007.

[76] F. Giroud, A. Diallo, and A. Douzal-Chouakria. A new approach for molecular dynamic network analysis. In *Réaumur Meeting*, Grenoble, 2007.

[77] O. Gotoh. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J Mol Biol*, 264(4) :823–838, 1996.

[78] V. Guralnik and J. Srivastava. Event detection from time series data. In *the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, number 33—42, 1999.

[79] D. Gusfield. *Algorithms on strings, trees, and sequences. Computer science and computational biology.* Cambridge, UK, Cambridge University Press, 1997.

[80] D. Gusfield. *Algorithms on Strings, Trees, and Sequences : Computer Science and Computational Biology.* Cambridge University Press, 1997.

[81] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel fisher discriminant analysis. In *advances in Neural Information Processing Systems*, 2008.

[82] Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total-variation penalty. *Journal of the American Statistical Association*, pages 105 – 492, 2010.

[83] V. Hautamaki, P. Nykanen, and P. Franti. Time-series clustering by approximate prototypes. In *19th International Conference on Pattern Recognition*, 2008.

[84] A. Hayashi, Y. Mizuhara, and N. Suematsu. Embedding time series data for classification. In *Machine Learning and Data Mining in Pattern Recognition*, pages 356–365, 2005.

[85] N.E. Heckman and R.H. Zamar. Comparing the shapes of regression functions. *Biometrika*, 22 :135–144, 2000.

[86] D.G. Higgins and P.M. Sharp. Clustal : a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1) :237–244, 1988.

[87] L. Hubert, J. Meulman, and W. Heiser. Two purposes for matrix factorization : A historical appraisal. *SIAM Review*, 42(1) :68–82, 2000.

[88] Y.S. Jeong, M.K. Jeong, and O.A. Omitaomu. Weighted dynamic time warping for time series classification. *Pattern Recognition*, 44 :2231–2240, 2011.

[89] W. Just. "computational complexity of multiple sequence alignment with sp-score". *J. Comput. Biol*, 8(6) :615–623, 2001.

[90] M. Juvela, K. Lehtinen, and P. Paatero. The use of positive matrix factorization in the analysis of molecular line spectra. In *MNRAS*, volume 280, pages 616–626, 1996.

[91] M W. Kadous and C. Sammut. Classification of multivariate time series and structured data using constructive induction. *Machine Learning Journal*, 58 :179–216, 2005.

[92] Y. Kakizawa, R.H. Shumway, and N. Taniguchi. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441) :328–340, 1998.

[93] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Eng., ser.*, 82 :35–45, 1960.

[94] H. Kantz and T. Schreiber. *Nonlinear time series analysis.* Cambridge University Press, New York, NY, 2004.

[95] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data. An Introduction to Cluster Analysis.* John Wiley & Sons, New York., 1990.

[96] A. Kelil, S. Wang, and R. Brzezinski. Clustering of non-alignable protein sequences. In *BIOKDD'07*, 2007.

[97] A. Kelil, S. Wang, and R. Brzezinski. A new alignment-independent algorithm for clustering protein sequences. In *IEEE BIBE'07*, 2007.

[98] A. Kelil, S. Wang, R. Brzezinski, and A. Fleury. Cluss : Clustering of protein sequences based on a new similarity measure. *BMC Bioinformatics*, 8(286), 2007.

[99] K. Keller and K. Wittfeld. Distances of time series components by means of symbolic dynamics. *International Journal of Bifurcation Chaos*, 14 :693–704, 2004.

[100] C. Kemena and C. Notredame. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics*, 25(2455–2465), 2009.

[101] J.B. Kruskall and M. Liberman. *The symmetric time warping algorithm : From continuous to discrete. In Time Warps, String Edits and Macromolecules.* Addison-Wesley., 1983.

[102] M. Kudo, J. Toyama, and M. Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20(11) :1103–1111, 1999.

[103] R.P. Kumar, P. Nagabhushan, and A. Chouakria-Douzal. Wavesim and adaptive wavesim transform for subsequence time-series clustering. In *9th IEEE International Conference on Information Technology*, pages 197 –202, dec. 2006.

[104] K. Kumara, R. Agrawal, and C. Bhattacharyya. A large margin approach for writer independent online handwriting classification. *Pattern Recognition Letters*, 29(7) :933–937, 2008.

[105] M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, 1521(1) :33–59, 2000.

[106] M. Lavielle and G. Teyssière. Detection of multiple change-points in multiple time-series. *Lithuanian Mathematical Journal*, 46(4), 2006.

[107] L. Lebart. Analyse statistique de la contiguité. *Publication de l'ISUP*, 18 :81–112, 1969.

[108] L. Lebart and N. Tabard. Recherches sur la description automatique des données socio-économiques. Technical Report 13/1971, Rapport CREDOC, Convention de recherche CORDES, 1973.

[109] D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401 :788–791, 1999.

[110] B. Leroy, A. Chouakria, I. Herlin, and E. Diday. Approche géométrique et classification pour la reconnaissance de visage. In *Reconnaissance des Formes et Intelligence Artificielle (RFIA'96)*, pages 548–557, Rennes, 1996.

[111] H. Leung, T. Lo, and S. Wang. Prediction of noisy chaotic time series using an optimal radial basis function neural network. *IEEE Trans. Neural Networks*, 12(5) :1163–1172, 2001.

[112] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR*, 163(4) :707–710, 1966.

[113] D.J. Lipman, S.F. Altschul, and J.D. Kececioglu. A tool for multiple sequence alignment. *PNAS Proc Natl Acad Sci U S A*, 86(12) :4412–4415, 1989.

[114] D. Liu, D. M. Umbach, S. D. Peddada, Li L., P. W. Crockett, and C.R. Weinberg. A random-periods model for expression of cell-cycle genes. *Proc Natl Acad Sci USA*, 101 :7240–7245, 2004.

[115] X. Liu, S. Lee, G. Casella, and G.F. Peter. Assessing agreement of clustering methods with gene expression microarray data. *Computational Statistics and Data Analysis*, 52(12) :5356–5366, 2008.

[116] Y. Luan and H. Li. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19 :474–482, 2003.

[117] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg New York, 2007.

[118] B.D. MacArthur, A. Lachmann, I.R. Lemischka, and A. Ma'ayan. Gate : Software for the analysis and visualization of high-dimensional time series expression data,. *Bioinformatics*, 26(1) :143–144, 2010.

[119] E.A. Maharaj. Cluster of time series. *Journal of Classification*, 17 :297–314, 2000.

[120] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman. *Forecasting : Methods and Applications*. John Wiley and Sons, 1998.

[121] A. Mom. *Méthodologie statistique de la classification de réseaux de transport*. PhD thesis, U.S.T.L., Montpellier., 1988.

[122] P.A.P. Moran. The interpretation of statistical maps. *Journal Royal Statistical Society*, Series B, 10 :243–251, 1948.

[123] P.A.P. Moran. Notes on continuious stochastic phenomena. *Biometrika*, 37 :17–23, 1950.

[124] B. Morgenstern. Dialign : multiple dna and protein sequence alignment at bibiserv. *Nucl. Acids Res*, 32(2) :W33–W36, 2004.

[125] A. Moschitti and F. Zanzotto. Fast and effective kernels for relational learning from texts. In *the 24th international conference on Machine learning*, pages 649–656, 2007.

[126] D.M. Mount. *Bioinformatics : Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press : Cold Spring Harbor, NY., 2004.

[127] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *IEEE Workshop – Neural Networks for Signal Processing*, volume 7, pages 511–520, 1997.

[128] K.-R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. Predicting time series with support vector machines. In *Int. Conf. on Artificial Neural Networks*. Springer, 1997.

[129] A. Nanopoulos, R. Alcock, and Y. Manolopoulos. Feature-based classification of time-series data. *International Journal of Computer Research*, pages 49–61, 2001.

[130] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1) :31–88, 2001.

[131] C. Notredame, D.G. Higgins, and J. Heringa. T-coffee : A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1) :205–217, 2000.

[132] J.J. Oliver and C.S. Forbes. Bayesian approaches to segmenting a simple time series. In *Proceedings of the Econometric Society Australasian Meeting.*, 1997.

[133] N. H. Packard, J. P. Crutchfeld, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Physical Review Letters*, 45(9) :712–716, 1980.

[134] C. Park, J. Koo, S. Kim, I. Sohn, and J.W. Lee. Classification of gene functions using support vector machine for time-course gene expression data. *Computational Statistics and Data Analysis*, 52(5) :2578–2587, 2008.

[135] F. Petitjean and P. GanÇarski. Summarizing a set of time series by averaging : from steiner sequence to compact multiple alignment. *Theoretical Computer Science*, 414(1) :76–91, 2012.

[136] F. Petitjean, A. Ketterlin, and P. GanÇarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*, 44(3) :678–693, 2011.

[137] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[138] M. F. Ramoni, P. Sebastiani, and I.S. Kohane. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. US.A*, 99 :9121–9126, 2002.

[139] C. A. Ratanamahatana and E. Keogh. Making time-series classification more accurate using learned constraints. In *SIAM International Conference on Data Mining*, pages 11–22., 2004.

[140] G. Rizk, A. Douzal-Chouakria, and C. Amblard. Temporal decision trees. In *56th Session of the International Statistical Institute*, Lisboa, Portugal, 2007.

[141] J.J. Rodriguez and C.J. Alonso. Interval and dynamic time warping-based decision tree. In *Proc of the ACM Symposium on applied computing,*, pages 548–552, 2004.

[142] J.J. Rodriguez, C.J. Alonso, and H. Bostrom. Boosting interval-based literals. *Intelligent Data Analysis*, 5(3) :245–262, 2001.

[143] J. Rydell, M. Borga, and H. Knutsson. Robust correlation analysis with an application to functional mri. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, number 453-456, 2008.

[144] N. Saito. *Local feature extraction and its application using a library of bases.* PhD thesis, Department of Mathematics, Yale University., 1994.

[145] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1) :43–49, 1978.

[146] D. Sankoff and J.B. Kruskal. *Time warps, string edits, and macromolecules : the theory and practice of sequence comparison.* Addison-Wesley, 1983.

[147] B. Schölkopf, A. J. Smola, and C. Burges. *Advances in Kernel Methods–Support Vector Learning.* Cambridge, MA : MIT Press, 1999.

[148] L. Scrucca. Class prediction and gene selection for dna microarrays using regularized sliced inverse regression. *Computational Statistics and Data Analysis.*, 52(1) :438–451, 2007.

[149] N. Serban and L. Wasserman. CATS : Cluster after transformation and smoothing. *Journal of the American Statistical Association*, 100(471) :990–999, 2005.

[150] N. Shervashidze and K. Borgwardt. Fast subtree kernels on graphs. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

[151] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistical Computing*, 14(3) :199–222, 2004.

[152] S. Sonnenburg, K. Rieck, F. F. I Ida, and G. Rtsch. Large scale learning with string kernels. In *Large Scale Kernel Machines*, pages 73–103. MIT Press, 2007.

[153] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, and P.O. Brown. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol*, 9 :3273–3297, 1998.

[154] T. Z. Tan, C. Quek, and G. S. Ng. Brain-inspired genetic complimentary learning for stock market prediction. In *IEEE Congress on Evolutionary Computation*, volume 3, pages 2653–2660, 2005.

[155] J. Thioulouse, D. Chessel, and S. Champely. Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environmental and Ecological Statistics*, 2 :1–14, 1995.

[156] J.D. Thompson, D.G. , Higgins, and T.J. Gibson. Clustal w : improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22) :4673–4680, 1994.

[157] H. TONG. *Non-linear Time Series : A Dynamical System Approach*. Oxford University Press, 1990.

[158] J. Von Neumann. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics*, 12(4), 1941.

[159] J. Von Neumann, R.H. Kent, H.R. Bellinson, and B.I. Hart. The mean square successive difference to the variance. *The Annals of Mathematical Statistics*, pages 153–162, 1942.

[160] L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *J Comput Biol*, 1(4) :337–348, 1994.

[161] D. Wartenberg. Multivariate spatial correlation : A method for exploratory geographical analysis. *Geographical Analysis*, 17(4) :263–283, 1985.

[162] A. S. Weigend and N. A. Gershenfeld, editors. *Time series prediction : Forecasting the future and understanding the past*. Addison Wesley, 1993.

[163] M.L. Whitfield, G. Sherlock, , J.I. Murray, C.A. Ball, K.A. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, and D. Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors molecular. *Biology of the Cell*, 13 :1977–2000, 2002.

[164] Y. Xie and B. Wiltgen. Adaptive feature based dynamic time warping. *International Journal of Computer Science and Network Security*, 10(1), January 2010.

[165] Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi. Decision-tree induction from time-series data based on standard-example split test. In *In Proceedings of the 20th International Conference on Machine Learning*, pages 840–847. Morgan Kaufmann, 2003.

[166] Y. Yao and S.T. Au. Least-squares estimation of a step function. *The Indian Journal of Statistics*, 51(3) :370–381, 1989.

[167] D. Yu, X. Yu, Q. Hu, J. Liu, and A. Wu. Dynamic time warping constraint learning for large margin nearest neighbor classification. *Information Sciences*, 181 :2787–2796, 2011.