Résumé

Les processus informatisés actuels génèrent des quantités massives de flux de données. Dans nombre d'applications, ces flux de données sont collectées en vue de modéliser les processus. Les modèles de processus obtenus ont pour but la réalisation d'objectifs tels que l'aide à la décision, la visualisation de données, l'informatique décisionnelle, l'automatisation et le contrôle, la reconnaissance de formes et la classification, etc. La modélisation de processus sur la base de données implique cependant de faire face à d'importants défis. Outre les erreurs, les données aberrantes et le bruit, le principal défi provient de la large dimensionnalité, i.e. du nombre de variables dans chaque échantillon de données mesurées. Les échantillons forment souvent une longue séquence temporelle appelée série temporelle multivariée, où chaque échantillon est influencé par les autres. Notre objectif est de construire un modèle robuste qui garantisse la génération, la révision et la représentation de nouvelles séries temporelles multivariées cohérentes avec le processus sous-jacent.

Dans cette thèse, nous adoptons un cadre de modélisation capable d'extraire, à partir de séries temporelles multivariées, des caractéristiques correspondant à des variations - covariations dynamiques communes aux variables mesurées dans tous les échantillons. Ces caractéristiques sont appelées «points communs» et une mesure qui leur est appropriée est définie. Ce qui rend le modèle de séries temporelles multivariées polyvalent est l'hypothèse relative à l'existence de séries temporelles latentes de caractéristiques connues ou présumées et de dimensionnalité beaucoup plus faible que les séries temporelles mesurées; le résultat est le bien connu «modèle factoriel dynamique». Des variantes originales de méthodes existantes pour estimer le modèle factoriel dynamique sont développées : l'estimation est réalisée en utilisant l'équivalent du modèle factoriel dynamique au niveau du domaine de fréquence, désigné comme le «modèle factoriel spectral». Pour estimer le modèle factoriel spectral, nous nous basons sur des idées relatives à la théorie des estimations spectrales. Cette théorie est utilisée pour aboutir à une formulation probabiliste, qui fournit des estimations de probabilité maximale pour les paramètres du modèle factoriel spectral. Des paramètres de probabilité maximale sont alors développés, en plaçant notre analyse entièrement dans le domaine spectral, de façon à ce que les séries temporelles latentes transformées dynamiquement héritent au maximum des points communs.

La principale contribution de cette thèse consiste en un cadre d'apprentissage utilisant le modèle factoriel spectral. Nous désignons par apprentissage la capacité d'un modèle de processus à caractériser de façon robuste les données générées par le processus à des fins de filtrage par motif, classification et prédiction. Dans ce contexte, le modèle factoriel spectral est considéré comme ayant appris une série temporelle multivariée si la série temporelle latente, une fois dynamiquement transformée, permet d'extraire les points communs de façon fiable et maximale. Le modèle factoriel spectral

sera utilisé principalement pour deux applications d'apprentissage de séries multivariées : en premier lieu, des ensembles de données sous forme de flux venant de différents processus du monde réel doivent être classifiés; lors de cet exercice, la classification porte sur des signaux magnétoencéphalographiques obtenus chez l'homme au cours de différentes tâches physiques et cognitives; en second lieu, les points communs obtenus sont testés en demandant une prédiction fiable d'une série temporelle multivariée étant donnée l'évolution passée; les prix d'un portefeuille d'actions sont prédits dans le cadre de ce défi.

À la fois pour la modélisation et pour l'apprentissage factoriel spectral, une solution analytique aussi bien qu'une solution itérative sont développées. Tandis que la solution analytique est basée sur une approximation de rang inférieur de la fonction de densité spectrale, la solution itérative est basée, quant à elle, sur l'algorithme de maximisation des attentes. Pour l'exercice de classification des signaux magnétoencéphalographiques humains, une stratégie de comparaison des similitudes entre les points communs des différentes classes de processus de séries temporelles multivariées est développée. Pour le problème de prédiction des prix des actions, un modèle vectoriel autorégressif dont les paramètres sont enrichis avec les points communs de probabilité maximale est conçu. Dans ces deux problèmes d'apprentissage, le modèle factoriel spectral atteint des performances louables en regard d'approches concurrentes.