Proposition de sujet de thèse "inter-ED 509/548" Segmentation et catégorisation temporelle sur données économiques historiques

Cécile BASTIDON (LEAD, ED509) et Sébastien PARIS (LIS, ED548)

1 Sujet

Approche dominante de l'identification de régimes homogènes. L'analyse et l'interprétation de données économiques s'appuient souvent au préalable sur l'identification de périodes homogènes spécifiques à l'intérieur de (longues) séries temporelles. On cherchera par exemple à détecter, horodater et potentiellement caractériser des segments/régimes temporels particuliers (en terme de modèle de processus), mais également à déterminer la présence de phénomènes transitoires, notamment de crises.

L'estimation de périodes temporelles est un problème complexe dans la mesure ou ni leur nombre, ni leur durée, ni leurs caractéristiques ne sont *a priori* connus avec certitude. En macroéconomie et macroéconomie financière, l'approche dominante mobilise des modèles à changement de régimes markoviens. Ces modèles présentent l'avantage de bien se prêter à l'identification d'un petit nombre de régimes aux caractéristiques stables, aisément interprétables comme phases du cycle macroéconomique et financier (voir, par exemple, pour une revue de littérature élargie dans le domaine de la finance empirique, [Gui11]). En contrepartie, les modèles markoviens supposent l'invariance dans le temps des phases du cycle et se prêtent mal à l'estimation sur séries multivariées. Leur application à l'Histoire soulève aussi le problème de l'indépendance des probabilités de transitions successives, *a priori* inappropriée aux dynamiques persistantes de type dépendance au sentier (voir, par exemple, [Dav85]).

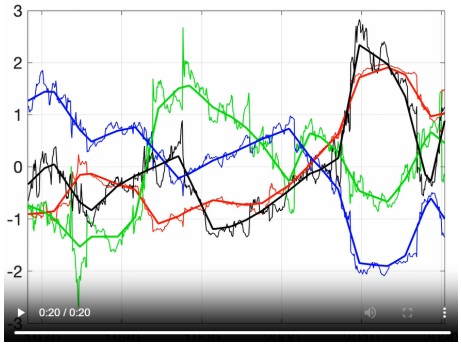

Figure 1: Segmentation multivariée des quatre premiers moments de la matrice de corrélations de séries temporelles longues de prix d'actions 1960-2022 [ABB+22]

Approche par la segmentation multivariée. Une alternative mobilisée dans le domaine de la physique statistique pour des applications aux Sciences sociales consiste à segmenter des séries temporelles multivariées, soit directement [ABB+22] soit via leurs corrélations [PK04] ou les graphes dynamiques qui leur sont associés [BPJ+20]. Ces méthodes ont en général en commun l'identification du "modèle vrai" par le débruitage des données disponibles. Le principe consiste, après avoir choisi en cohérence avec la littérature les caractéristiques générales des régimes (stables, ou simplement linéaires), à poser des segments jointifs en minimisant l'écart entre données disponibles et signal vrai sous contrainte de minimiser le nombre de points de changement, la plupart du temps conjoints sur l'ensemble des composantes du signal (Figure 1).

Comparativement à l'approche dominante, ces méthodes de segmentation ne supposent pas que les régimes associés aux phases du cycle économique et financier soient invariants dans le temps, et sont applicables à des données hautement multivariées.

Objectifs du projet. Avec l'avènement de l'apprentissage statistique/intelligence artificielle dans cette dernière décénie, il est maintenant possible de solutionner ce genre de problématique à partir de modèles appris sur des corpus de séries temporelles. Un gros avantange de ces approches est qu'il n'est plus nécessaire d'imposer des a priori forts sur les régimes (et leurs caractéristiques), permettant de capturer une plus grande diversités intra-régime en terme de processus. Par exemple dans [GR23] avec l'algorithme *PrecTime* la segmentation temporelle est réalisée à l'aide de combinaison de CNN (Convolutional Neural Networks) et LSTM (Long Short-Term Memory) sur des signaux de machineries industrielles.

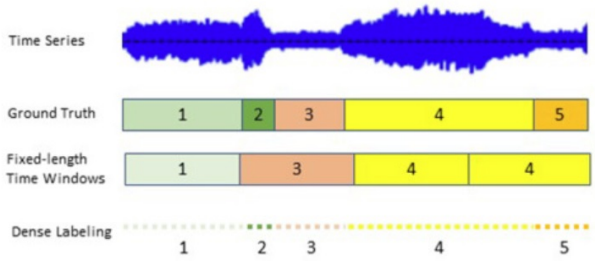

Figure 2: Prectime identifiant les différents segments temporels grâce au *dense labeling*

Les nouvelles topologies de réseaux de Neurones à base de Transformers permettent de capturer des dépendances *spatio-temporelles* de façon beaucoup plus efficaces que les réseaux récurrents (RNN) classiques. Dans [WZZ+23], une revue complète des algorithmes IA à base de *Transformer* dédiés à la détection/classification de séries temporelles est présentée. Dans [SKDD22], pour une tâche de NILM (Non Intrusive Load Monitoring), non seulement l'identification temporelles des régimes est faite (à base de transformers) mais également la reconstruction/débruitage des signaux.

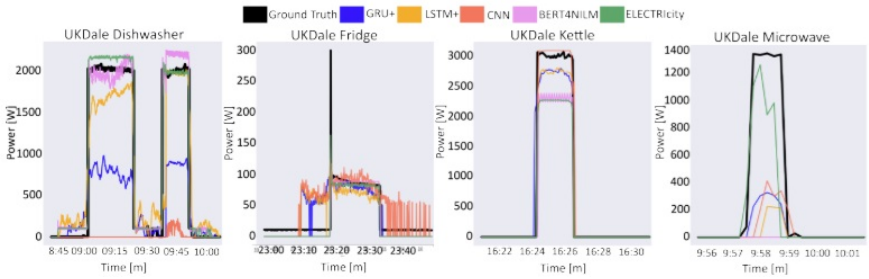

Figure 3: ELECTRICITY permet l'identification et le débruitage de séries temporelles

Enfin, un dernier exemple d'application des Transformers dans la prédiction des vibrations mécaniques transmises dans une structure est ([ZGZ+24]), où là encore une topololgie complexe à base de transformer permet une prédiction robuste de la série temporelle à partir de l'onde de choc initiale.

L'idée générale de cette proposition de thèse est de coupler/hybrider les approches traditionnelles en segmentation de séries temporelles économiques avec ces nouvelles approches IA. Par exemple, la partie détection/horadatage pourrait être réalisée *via* IA, permettant de mieux initialiser des modèles fins d'identification de régimes temporels rencontrés, en particulier, en histoire économiquee. Un avantage critique de cette approche est de ne requérir aucune hypothèse préalable sur la forme des non-linéarités éventuelles des données observées.

Dans le domaine de l'Histoire économique, l'existence de bases de données fondées sur l'expertise identifiant les phases des cycles économiques et financiers constitue une labellisation des données immédiatement mobilisable. L'application à des séries longues est ainsi réalisable, avec en particulier

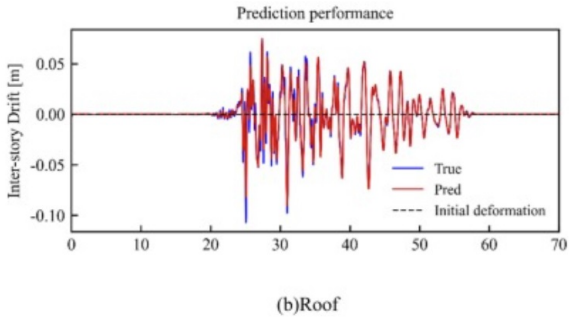

Figure 4: Algorithme TLM pour la prédiction d'ondes de choc

pour objet final l'identification de configurations originales précédant habituellement les crises macroéconomiques et financières.

2 Profil requis du candidat

La/le candidat(e) devra avoir une solide formation en mathématiques, statistiques, optimisation idéalement issu(e) d'un Master 2 en systèmes complexes/mathématique ou modélisation macroéconomique. Le/la candidat devra également avoir une expérience dans la programmation en language Python/Matlab.

References

[ABB+22] Patrice Abry, Cécile Bastidon, Pierre Borgnat, Pablo Jensen, Antoine Parent, and Barbara Pascal. Detecting global financial crises over history: A multivariate nonlinear denoising strategy. *World Economic History Conference Paris 2022*, 2022.

[BPJ+20] Cécile Bastidon, Antoine Parent, Pablo Jensen, Patrice Abry, and Pierre Borgnat. Graph-based era segmentation of international financial integration. *Physica A: Statistical Mechanics and its Applications*, 539:122877, 2020.

[Dav85] Paul A David. Clio and the economics of qwerty. *The American economic review*, 75(2):332–337, 1985.

[GR23] Stefan Gaugel and Manfred Reichert. Prectime: A deep learning architecture for precise time series segmentation in industrial manufacturing operations. *Engineering Applications of Artificial Intelligence*, 122:106078, 2023.

[Gui11] Massimo Guidolin. Markov switching models in empirical finance. In *Missing data methods: Time-series methods and applications*, pages 1–86. Emerald Group Publishing Limited, 2011.

[PK04] Szilard Pafka and Imre Kondor. Estimated correlation matrices and portfolio optimization. *Physica A: statistical mechanics and its applications*, 343:623–634, 2004.

[SKDD22] Stavros Sykiotis, Maria Kaselimi, Anastasios Doulamis, and Nikolaos Doulamis. Electricity: An efficient transformer for non-intrusive load monitoring. *Sensors*, 22(8), 2022.

[WZZ+23] Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey, 2023.

[ZGZ+24] Qingyu Zhang, Maozu Guo, Lingling Zhao, Yang Li, Xinxin Zhang, and Miao Han. Transformer-based structural seismic response prediction. *Structures*, 61:105929, 2024.