

Segmentation de séries temporelles avec composantes de variance

Christian DERQUENNE¹

Les séries temporelles se décomposent généralement en plusieurs types d'évolution : tendance, saisonnalité, volatilité et bruit. Elles peuvent être plus ou moins régulières selon le domaine d'application. Les changements de comportements qui caractérisent principalement ces séries sont de plusieurs types : pics, sauts en niveau, en tendance, en variabilité.

La modélisation de ces séries est donc très délicate et demande beaucoup d'expérience. Il peut alors être intéressant de détecter des ruptures de comportement pour la construction de sous-modèles, la stationnarisation de la série, la construction de courbes symboliques pour la classification de courbes. De nombreuses méthodes de segmentation dans [2,3,4] ont été et sont développées pour répondre à ces différentes problématiques en économie, en finance, en séquençage humain, en météorologie, en management de l'énergie, etc. La plupart de ces méthodes reposent sur l'utilisation de la programmation dynamique pour diminuer drastiquement le nombre de segmentations possible.

Ces méthodes de détection de points de rupture ont pour vocation de résoudre trois problèmes [3] : (i) la détection de changement de la moyenne, avec une variance constante, (ii) la détection de changement de variance avec une moyenne constante et (iii) la détection de changements dans l'ensemble de la distribution du phénomène étudié. Nous avons introduit une méthode [1], qui permet non seulement de réduire la complexité par rapport à d'autres méthodes, mais surtout de proposer des solutions de segmentation de la série contenant des segments croissants, décroissants, constants et des dispersions différents.

Notre méthode est originale dans son approche car elle propose, par étapes successives, une aide à la décision pour la segmentation des données. Elle contient deux phases principales : la préparation des données offrant une première segmentation des données et la modélisation des segments à l'aide d'un modèle linéaire gaussien hétéroscédastique par adaptations de simplification successives. Chacune de ses deux phases est répétée un certain nombre de fois en fonction du degré de lissage appliqué aux données.

Cette méthode a été testée sur de nombreuses séries et a fourni des résultats encourageants à la fois sur des données simulées afin de juger de la qualité de reconstitution de la série, en termes de détection et de modélisation des segments, mais a surtout été appliquée sur des données réelles, notamment dans le domaine de la formation des prix de marché de l'énergie. Cependant pour l'ensemble des méthodes de segmentation qu'elles soient fondées sur une approche par programmation dynamique ou sur une approche exploratoire comme la nôtre, il s'avère que la qualité de la segmentation peut faire défaut lors de la détection de segments contigus quand les niveaux (constants ou pentes) sont proches statistiquement mais ont des variances différentes. Dans ce cas un seul segment sera détecté, alors qu'il y en a deux structurellement.

Par conséquent, nous proposons dans cet article, une nouvelle méthode améliorant la précédente. Cette nouvelle approche contient trois phases. La première consiste à établir une transformation adéquate des données afin d'obtenir une nouvelle série caractérisant l'évolution temporelle de la dispersion des observations. Cette phase repose sur deux théorèmes pour estimer la dispersion.

La deuxième phase revient à segmenter cette nouvelle série avec le même principe que la méthode [1] pour obtenir des segments de dispersion, enfin la troisième phase applique à

¹ EDF R&D –1, avenue du Général de Gaulle – 92141 Clamart Cedex

nouveau [1] mais en tenant compte de la distribution des segments de dispersion, notamment lors de la construction du modèle linéaire hétéroscédastique.

Afin de tester notre approche, nous avons alors réalisé une étude comparative avec des algorithmes de programmation dynamique proposés en [3]. Sur des exemples simulés, cette nouvelle approche permet à la fois d'améliorer l'ancienne méthode, mais aussi de montrer qu'elle est plus performante que des approches par programmation dynamique dans la majorité des cas. Cette méthode est notamment très intéressante pour des applications dans lesquelles les signaux changent de processus. Enfin, nous proposons des voies futures de recherche consistant à généraliser les deux théorèmes introduits dans cet article.

[1] Derquenne, C. (2011), An Explanatory Segmentation Method for Time Series, in *Proceedings of Compstat 2010*, Y. Lechevallier and G. Saporta (eds.), 1st Edition, pp. 935-942.

[2] Guédon, Y. (2008), Exploring the segmentation space for the assessment of multiple change-point models. Institut National de Recherche en Informatique et en Automatique, *Cahier de recherche 6619*.

[3] Lavielle, M. (2009), Detection of Changes using a Penalized Contrast (the DCPC algorithm), http://www.math.u-psud.fr/~lavielle/programmes/_lavielle.html.

[4] Perron, P. and Kejriwal, M. (2006), Testing for Multiple Structural Changes in Cointegrated Regression Models. Boston University, C22.