Modèle linéaire mixte avec segmentation : application à la détection de changements dans les dates de vendanges

Emilie Lebarbier, Franck Picard, Eva Budinska, Stéphane Robin

Modèle linéaire mixte avec segmentation : application à la détection de changements dans les dates de vendanges

Emilie Lebarbier[1] & Franck Picard[2] & Eva Budinskà[3] & Stéphane Robin[1]

1. UMR AgroParisTech / INRA MIA 518, 16 rue Claude Bernard, F - 75 231 Paris
2. UMR CNRS-8071/INRA-1152/Université d'Évry, 91000 Évry
3.Centre of Biostatistics and Analysis, Faculty of Science and Faculty of Medicine, Masaryk University, Brno

Résumé

Nous nous intéressons à la détection de changements dans les dates de vendanges de plusieurs stations qui seraient dûs à des changements de pratiques et non à des changements climatiques. Ces séries sont analysées simultanément à l'aide d'un modèle linéaire mixte avec ruptures qui permet de prendre en compte à la fois des covariables et des corrélations entre series. Pour obtenir les paramètres du maximum de vraisemblance, nous utilisons un algorithme EM et proposons un nouvel algorithme de programmation dynamique pour l'étape de segmentation. Cependant, se pose la question du choix du nombre de segments. Ici nous généralisons trois critères de sélection de modèles, qui avaient été proposés dans le cas de la segmentation d'une serie, à la segmentation jointe de plusieurs series. Nous comparons ces critères par une étude de simulation.

***Mots-clés**:* Segmentation; Modèle linéaire mixte; Processus gaussien multivarié; Programmation Dynamique; EM algorithm.

Abstract

The problem is to detect abrupt changes in the harvest dates of grapes which are not explained by changes in climate but by technical or practical changes. We consider the joint segmentation of multiple series. We use a mixed linear model with breakpoints to account for both covariates and correlations between signals. We propose an estimation algorithm based on EM which involves a new dynamic programming strategy for the segmentation step. Moreover the joint segmentation raises the model selection issue. To this end, we generalize three penalized criteria, proposed in the univariate segmentation case, to the multiple series case. These criteria are compared in a simulation study.

Nous nous intéressons à la détection de changements dans les dates de vendanges de plusieurs stations qui seraient dus à des changements de pratiques ou de techniques.

Les vendanges ayant lieu plus tôt les années chaudes, nous souhaitons distinguer les changements spécifiques des stations de ceux qui seraient dus aux variations climatiques, et qui donc affecteraient toutes les séries aux mêmes périodes. Afin de prendre en compte ces évènements climatiques, nous cherchons à analyser les séries simultanément. Nous proposons un modèle linéaire mixte avec segmentation qui permet de prendre en compte à la fois des covariables, soit un effet climat, et des corrélations entre series à une date donnée, soit un éventuel effet année. De plus la segmentation est ici spécifique à chacune des séries, et non commune. On note Y_{mt} la date de vendange de l'année t pour la station m and x_{mt} la temperature, le modèle est:

$$\forall t \in I_k^m \;\; Y_{mt} = \mu_{mk} + bx_{mt} + U_t + E_{mt}. \tag{1}$$

où I_k^m représente le $k$ème segment de la série m, μ_{mk} sa moyenne et où les erreurs E_{mt} sont supposées i.i.d. de distribution Gaussienne centrée et de variance σ_0^2, et les effets aléatoires U_t indépendants de distribution Gausssienne centrée de variance σ_u^2. Ce modèle s'écrit sous la forme général suivante:

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{U} + \mathbf{E},$$

où $\mathbf{T}, \mathbf{X}$ et $\mathbf{Z}$ sont respectivement les matrices d'incidence des ruptures, des paramètres constants et des effets aléatoires. $\mathbf{U}$ et $\mathbf{E}$ sont supposés indépendants de distribution Gaussienne centrée avec comme matrices de variance-covariance respectives $\mathbf{G}$ et $\mathbf{R}$. Comparée aux modèles linéaires mixtes classiques, la matrice d'indicidence $\mathbf{T}$ est ici inconnue, puisqu'elle représente la position des instants de ruptures, et doit être estimée.

Les paramètres du modèle sont estimés par maximum de vraisemblance en utilisant l'algorithme EM [3], dont son utilisation est aujourd'hui bien connu pour l'estimation des paramètres de modèles linéaires mixtes (cf par exemple [7]). Une des étapes de maximisation consiste en l'estimation de la localisation des ruptures. Pour obtenir la solution optimale, l'algorithme classiquement utilisé est la programmation dynamique (cf [1], [2], [6]) Même si cet algorithme permet déjà de réduire la complexité algorithme, il ne peut être utilisé dès lors que le nombre ou la taille des series devient trop grand. Nous proposons ici un nouvel algorithme basé sur deux étapes de l'algorithme de programmation dynamique pour obtenir la solution optimale.

Cette procédure permet d'obtenir la meilleure segmentation de toutes les séries en K segments. Cependant, se pose la question du choix de ce nombre de segments. Ici nous généralisons trois critères de sélection de modèles ([4], [5], [8]), qui avaient été proposés dans le cas de la segmentation d'une serie, à la segmentation jointe de plusieurs series. Nous comparons la performance de ces critères par une étude de simulation.

La méthode proposée est appliquée à la détection de changements dans les dates de vendanges de 10 stations françaises.

Bibliographie

[1] J. Bai and P. Perron (2003), Computation and analysis of multiple structural change models, J. Appl. Econ., 18, 1–22.
[2] H. Caussinus and O. Mestre (2004), Detection and correction of artificial shifts in climate series, JRSS-C, 53(3), 405–425.
[3] A. P. Dempster and N. M. Laird and D. B. Rubin (1977), Maximum Likelihood from Incomplete Data via the EM algorithm, Journal of the Royal Statistical Society Series B, 39, 1–38.
[4] M. Lavielle (2005), Using penalized contrasts for the change-point problem, Signal Processing, 85(8), 1501–1510.
[5] E. Lebarbier (2005), Detecting Multiple Change-Points in the Mean of Gaussian Process by Model Selection, Signal Processing, 85, 717–736.
[6] Picard, F. and Robin, S. and Lavielle, M. and Vaisse, C. and Daudin, J.-J. (2005), A statistical approach for array CGH data analysis, BMC Bioinformatics, 27(6), 1.
[7] D.A. van Dyk (2000), Fitting mixed-effects models using efficient EM-type algorithms, Jour. Comp. and Graph. Statistics, 9, 78–98.
[8]Zhang, N. R. and Siegmund, D. O. (2007), A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data, Biometrics, 63(1), 22–32.