Distribution de l'espace des segmentations

Guillem Rigaill, Stéphane Robin, Emilie Lebarbier

Distribution de l'espace des segmentations

Guillem Rigaill[1,2], Stéphane Robin[1] & Emilie Lebarbier [1]

1. UMR AgroParisTech / INRA MIA 518, 16 rue Claude Bernard F - 75 231 Paris
2. Institut Curie. Département de Transfert. Laboratoire de Signalisation. Quadrilatère historique, Porte 13 - 1 rue Claude Vellefaux - Hôpital Saint-Louis - 75 010 Paris

Mots-clés : Statistique des processus - séries temporelles, Biologie - Génôme

Résumé

La segmentation, ou détection de ruptures multiples, est un problème rencontré dans de nombreux domaines, par exemple en météorologie, économétrie ou génétique moléculaire. De manière générale, on observe un signal bruité, affecté par des changements brutaux appelés ruptures. Les méthodes classiques ont comme objectif de trouver la position de ces ruptures et d'inférer la vraie valeur du signal. Calculer la probabilité *a posteriori* d'une segmentation ou celle d'une rupture est techniquement difficile car le nombre de segmentations possibles est grand. Toutefois, ces informations pourraient permettre de mieux comprendre la structure des données.
Nous proposons un algorithme efficace pour l'exploration exhaustive de l'espace des segmentations. Cet algorithme nous permet de calculer exactement la probabilité *a posteriori* d'une segmentation. Il devient alors possible de calculer la probabilité d'une rupture ou d'un segment, ainsi que l'entropie de l'espace des segmentations. De plus, à partir de ces quantités, nous dérivons des critères BIC et ICL pour sélectionner le nombre de segments. Dans le cadre de la régression, nous obtenons un estimateur de l'espérance du signal à chaque position.

Abstract

The multiple change-point detection problem occurs in many areas, for example meteorology, econometry or molecular genetics. In all these cases, a noisy signal affected by abrupt changes is observed. Classical methods aim at finding an optimal segmentation of the signal and recovering the breakpoint positions. Computing the posterior probability of a segmentation and of breakpoints is technically difficult due to the very high number of possible segmentations. However, these quantities would bring valuable insights on the structure of the data.
We propose an efficient algorithm to exhaustively explore the segmentation space. Using this algorithm, it is possible to recover the posterior probability of a given segmentation, breakpoint or segment. Moreover, it allows the computation of the segmentation space entropy. Using these quantities, we derive some BIC and ICL criteria to select the number of segments. In the context of regression, we obtain a Bayesian averaged estimate of the signal where all possible segmentations are weighted by their posterior probability.

Introduction. La segmentation, ou détection de ruptures multiples, est un problème rencontré dans de nombreux domaines comme la météorologie, l'économétrie ou la génétique moléculaire. De manière générale, on observe un signal bruité $\{Y_t\}_{t\in[1..n]}$ affecté par des ruptures ou points de cassure. Il est difficile de calculer des quantités comme la probabilité *a posteriori* d'avoir une rupture à la position t. Ce calcule nécessite en effet le parcours des C_{n-1}^{j-1} segmentations en j morceaux. Toutefois, la connaissance de ces quantités permettrait une meilleure compréhension de la structure des données. En particulier, on peut penser qu'elle permettrait de mieux sélectionner le nombre de segments.

Nous proposons un algorithme efficace de parcours de l'espace des segmentations qui, dans un cadre bayésien, nous permet d'obtenir la probabilité *a posteriori* d'une segmentation. Connaissant celle-ci, il devient alors possible de calculer la probabilité *a posteriori* d'une rupture ou d'un segment. Dans le cadre de l'estimation du signal, cela nous permet de moyenner l'ensemble des segmentations envisageables en les pondérant par leur probabilité *a posteriori*. Enfin, nous dérivons des critères BIC et ICL dans le cadre de la sélection de modèle.

Définition. Dans la suite, nous appellerons $\mathcal{M}_j$ l'ensemble des segmentations en j segments. Il y en a C_{n-1}^{j-1}. Une segmentation particulière de $\mathcal{M}_j$ sera notée m et r représentera un des j segments de m.

Modèles. Les modèles que nous considérons sont de la forme générale :

$$\text{Si } t \in r \qquad Y_t \sim P(\theta_r) \qquad \text{où les } \{Y_t\}_t \text{ sont indépendants}$$

Où P est une loi de probabilité quelconque dont les paramètres θ_r sont spécifiques à chaque segment r. Deux cas particuliers sont le modèle normal hétéroscédastique (1) et le modèle de Poisson (2) :

$$\text{Si } t \in r \qquad Y_t \sim \mathcal{N}(\mu_r, \sigma_r^2) \tag{1}$$
$$\text{Si } t \in r \qquad Y_t \sim \mathcal{P}(\lambda_r) \tag{2}$$

.

Probabilité d'une segmentation. Nous proposons une méthodologie pour calculer exactement la probabilité *a posteriori* d'une segmentation. Pour cela, nous utilisons un cadre bayésien. Nous notons les probabilités suivantes :

$P(m)$ probabilité *a priori* de la segmentation m ;

$P(\theta_m|m)$ distribution des paramètres, étant donnée m ;

$P(Y|\theta_m, m)$ distribution des données, étant donnés m et θ_m.

Les probabilités *a posteriori* que l'on souhaite calculer sont :

$$P(m|Y) = \frac{P(m)}{P(Y)}.P(Y|m) = \frac{P(m)}{P(Y)}\int_\Theta P(Y|\theta_m, m)P(\theta_m|m)d\theta_m$$
$$P(m|Y,j) = \frac{P(m)}{P(Y\cap j)}.P(Y|m) = \frac{P(m)}{P(Y\cap j)}\int_\Theta P(Y|\theta_m, m)P(\theta_m|m)d\theta_m$$

Le calcul de $P(Y|m)$ est possible soit en intégrant directement $\int_\Theta P(Y|\theta_m, m)P(\theta_m|m)d\theta_m$, soit en utilisant l'approximation de Laplace quand l'intégration n'est pas possible.

Calcul de $\mathbf{P(Y \cap j)}$ et parcours de l'espace des segmentations. $P(Y\cap j)$ se calcule comme une somme sur toutes les segmentations de dimension j :

$$P(Y \cap j) = \sum_{m' \in \mathcal{M}_j} P(Y|m)P(m)$$

Le calcul naïf est impossible car le nombre de segmentations envisageables est grand : C_{n-1}^{j-1}. Guédon [3] propose un algorithme de type "backward-forward" pour parcourir l'espace des segmentations. Nous montrons que le calcule de $P(Y \cap j)$ se ramène en fait à calculer le terme à la première ligne et à la dernière colonne d'une matrice $\mathbf{F}$ élevée à la puissance j :

$$P(Y \cap j) = (\mathbf{F}^j)_{1(n+1)}$$

où $\mathbf{F}$ est une matrice carrée à $(n+1)$ lignes.

Caractéristique du signal. Une fois les probabilités $P(m|Y)$ et $P(m|Y,j)$ calculées, il est possible d'obtenir :

- la probabilité *a posteriori* d'un point de rupture ;
- la probabilité *a posteriori* d'un segment ;
- dans un objectif de régression, un estimateur de l'espérance du signal à chaque position.

Sélection de modèle. Dans le cadre de la sélection de modèle, nous dérivons un critère BIC. Par ailleurs, connaissant $P(m|Y,j)$, il est aussi possible de calculer l'entropie de l'espace des segmentations en j morceaux : $\mathcal{H}(j)$.

$$\mathcal{H}(j) = \sum_{m \in \mathcal{M}_j} P(m|Y)\log P(m|Y)$$

Cette entropie est petite si la probabilité est concentrée sur quelques segmentations. Elle est grande si, au contraire, la probabilité est dispersée sur de nombreuses segmentations. La connaissance de l'entropie nous permet de calculer un critère ICL qui favorise les dimensions d'entropie faible.

Bibliographie

[1] Picard F., Robin S., Lavielle M., Vaisse C., Daudin J.J. (2005), A statistical approach for array CGH data analysis, BMC Bioinformatics, 11, 6-27.
[2] 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, 22-32.
[3]Guédon Y. (2008), Exploring the segmentation space for multiple change-point models, Rapport de recherche 6619, INRIA, Number 6619 - 2008.