Heuristique de pente pour les modèles de détection de ruptures multiples

Yann Guédon [1]

[1] *CIRAD, UMR AGAP et Inria, Virtual Plants, Montpellier, guedon@cirad.fr*

Résumé. En ce qui concerne la détection de ruptures multiples, la sélection du nombre de ruptures a fait l'objet ces dernières années de nombreux travaux. Mais les approches proposées sont soit dédiées à un modèle particulier (par exemple modèle gaussien de changement sur la moyenne) soit donnent des résultats peu satisfaisants sur des séquences de taille petite ou moyenne. Nous proposons ici d'appliquer l'heuristique de pente, un critère non-asymptotique de vraisemblance pénalisée récemment proposé, pour sélectionner le nombre de ruptures. Nous appliquons en particulier la méthode d'estimation de la pente dirigée par les données, le point clé étant de définir la forme de la pénalité. L'approche proposée est illustrée sur deux jeux de données de référence pour les modèles de détection de ruptures multiples.

Mots-clés. Détection de ruptures multiples, estimation de la pente dirigée par les données, modèle à structure latente, sélection de modèles.

Abstract. With regard to the retrospective multiple change-point detection problem, much effort has been devoted in recent years to the selection of the number of change points. But, the proposed approaches are either dedicated to specific models (e.g. Gaussian change in the mean model) or give unsatisfactory results for short or medium length sequences. We propose to apply the slope heuristic, a recently proposed non-asymptotic penalized likelihood criterion, for selecting the number of change points. We in particular apply the data-driven slope estimation method, the key point being to define a relevant penalty shape. The proposed approach is illustrated using two benchmark data sets.

Keywords. Data-driven slope estimation, latent structure model, model selection, multiple change-point detection.

1 Introduction

L'heuristique de pente a été introduite par Birgé et Massart (2001) comme un nouveau critère non-asymptotique de vraisemblance pénalisée pour la sélection de modèles. Ils ont montré qu'il existe une pénalité minimale telle que la dimension des modèles (et le risque de l'estimateur associé) sélectionnés avec une pénalité plus petite devient très grande. De plus, ils ont montré qu'en prenant une pénalité égale à 2 fois la pénalité minimale permettait de sélectionner un modèle proche du meilleur modèle possible (ou modèle

oracle) en termes de risque d'estimateur. Cette approche a récemment été popularisée par l'introduction par Baudry *et al.* (2012) de la méthode d'estimation de la pente dirigée par les données qui est une méthode pratique pour implémenter les heuristiques de pente. Dans le cadre de l'estimation par maximum de vraisemblance, cette méthode est basée sur la relation linéaire attendue entre la forme de la pénalité (une fonction de la dimension du modèle) et la log-vraisemblance maximisée pour des modèles sur-paramétrés. Nous nous intéressons ici à l'application des heuristiques de pente pour sélectionner le nombre de ruptures dans des modèles de détection de ruptures multiples.

2 Définition de la fonction de log-vraisemblance et de la forme de la pénalité pour des modèles de détection de ruptures multiples

Pour les modèles de détection de ruptures multiples, les deux fonctions de log-vraisemblance possibles sont :

- $\log f(\mathbf{s}^*, \mathbf{x}; J)$, log-vraisemblance de la segmentation la plus probable $\mathbf{s}^*$ en J segments (le nombre de ruptures est donc égal à $J-1$) de la séquence observée $\mathbf{x}$. Lebarbier (2005) a utilisé cette fonction de log-vraisemblance pour définir une heuristique de pente pour des modèles gaussiens de changement sur la moyenne.

- $\log f(\mathbf{x}; J)$, log-vraisemblance de toutes les segmentations possibles en J segments de la séquence observée $\mathbf{x}$ avec $f(\mathbf{x}; J) = \sum_{\mathbf{s}} f(\mathbf{s}, \mathbf{x}; J)$.

Ces log-vraisemblances peuvent être calculées exactement pour $K = 2, \ldots, J$ par un algorithme de programmation dynamique dans le cas de $\log f(\mathbf{s}^*, \mathbf{x}; J)$ et par un algorithme de lissage dans le cas de $\log f(\mathbf{x}; J)$ (Guédon, 2013). L'estimation de la pente repose sur des log-vraisemblances maximisées pour des modèles sur-paramétrés et, comme l'a montré Guédon (2013, 2015), la segmentation la plus probable a souvent peu de sens pour des modèles sur-paramétrés où des segmentations très différentes ont des probabilités *a posteriori* voisines. Par ailleurs, la log-vraisemblance de la segmentation la plus probable fluctue fortement fonction de J alors que la log-vraisemblance de toutes les segmentations possibles est au contraire très lisse. Nous allons nous focaliser sur la log-vraisemblance de toutes les segmentations possibles $\log f(\mathbf{x}; J)$ pour définir une heuristique de pente de manière cohérente avec notre vue des modèles de détection de ruptures multiples comme des modèles à structure latente (Guédon, 2013; 2015). Ceci diffère du point du vue non-bayésien classique sur ces modèles où les ruptures sont vues comme des paramètres fixes à estimer (Lebarbier, 2005). Une fois les ruptures estimées, il n'y a bien entendu plus de structure latente. Ce point de vue est par contre homogène avec le point de vue bayésien où ces modèles sont toujours vus comme des modèles à structure latente. Un modèle de

détection de ruptures multiples sera donc vu comme un modèle empirique hiérarchique dans le cas non-bayésien et comme un modèle bayésien hiérarchique dans le cas bayésien.

Nous avons étudié sur de nombreux jeux de données le comportement de la log-vraisemblance de toutes les segmentations possibles $\log f(\mathbf{x}; J)$ sur la plage de valeurs de J correspondant à des modèles sur-paramétrés et nous avons remarqué qu'elle était concave de manière marquée si $J < T$ (par exemple si $10 < T/J < 100$), où T est la longueur de la séquence, mais beaucoup moins si $J \ll T$.

Pour appliquer les heuristiques de pente, il est necéssaire que (Baudry *et al.*, 2012) :
(C1) La log-vraisemblance soit croissante fonction de J.
(C2) La forme de la pénalité $\text{pen}_{\text{shape}}(J)$ soit croissante fonction de J.

À ces deux conditions standards, nous ajoutons les deux conditions suivantes spécifiques aux modèles de détection de ruptures multiples :
(C3) La forme de la pénalité $\text{pen}_{\text{shape}}(J)$ dépend de la longueur T de la séquence. Ajouter un segment pour disons $J = T/10$ induit une augmentation plus faible de la forme de la pénalite qu'ajouter un segment pour $J \ll T$.
(C4) La différenciation au premier ordre de la forme de la pénalité $\text{pen}_{\text{shape}}(J) - \text{pen}_{\text{shape}}(J-1)$ est décroissante fonction de J. Cette décroissance n'est pas une fonction linéaire de J.

Pour definir la forme de la pénalité, notre point de départ a été $\text{pen}_{\text{shape}}(J) = \log n_J$ où $n_J = \binom{T-1}{J-1}$ est le nombre de segmentations possibles en J segments. Considérons le cas limite où toutes les segmentations sont équiprobables pour J fixé, alors $\log f(\mathbf{x}; J) = \log \gamma_J + \log n_J$. En fait pour des modèles sur-paramétrés comme l'a montré Guédon (2013, 2015) sur différents exemples, $\log f(\mathbf{x}; J)$ se décompose en une part structurelle correspondant aux vrais ruptures et un bruit qui croit avec J. La fonction $\log f(\mathbf{x}; J) = \alpha + \beta \log n_J$ traduit ce comportement où β quantifie la balance entre le poids des vrais ruptures et le bruit dans $\log f(\mathbf{x}; J)$.

Pour respecter le caractère monotone de $\text{pen}_{\text{shape}}(J)$ fonction de J, nous proposons finalement

$$\text{pen}_{\text{shape}}(J) = \log \left\{ \frac{T^{J-1}}{(J-1)!} \right\},$$

avec

$$\text{pen}_{\text{shape}}(J) - \text{pen}_{\text{shape}}(J-1) = \log \left(\frac{T}{J-1} \right),$$

dont le comportement est proche de

$$\log n_J - \log n_{J-1} = \log \left\{ \frac{T-(J-1)}{J-1} \right\},$$

pour $J \ll T$.

3 Illustration sur deux jeux de données de référence

L'heuristique de pente proposée est illustrée sur deux jeux de données de référence correspondant à différents modèles de segment (Poisson et gaussien) et des longueurs de séquences contrastées. L'heuristique de pente est comparée avec le critère ICL "exact"proposé par Rigaill *et al.* (2012).

3.1 Catastrophes dans les mines de charbon en Grande-Bretagne

Les données sont les dates de 191 catastrophes minières entre 1851 et 1962 résumées par des comptages annuels durant la période d'étude de 112 ans. Nous supposons que le nombre de catastrophes par année suit une loi de Poisson et que le paramètre de cette loi est constant par morceaux au cours du temps.

La différenciation au 1er ordre de la log-vraisemblance de toutes les segmentations possibles $\log f(\mathbf{x}; J) - \log f(\mathbf{x}; J-1)$ est strictement décroissante fonction du nombre de segments J (Figure 1a). Cette log-vraisemblance sur la plage de valeurs de J correspondant à des modèles sur-paramétrés est donc concave de manière marquée et le nombre de segments J n'est certainement pas une forme de pénalité adaptée. Cette figure illustre aussi le fait que la log-vraisemblance de la segmentation la plus probable $\log f(\mathbf{s}^*, \mathbf{x}; J)$ fluctue fortement fonction de J alors que la log-vraisemblance de toutes les segmentations possibles $\log f(\mathbf{x}; J)$ est au contraire très lisse. La différence de log-vraisemblances sur la différence de formes de pénalité

$$\frac{\log f(\mathbf{x}; J) - \log f(\mathbf{x}; J-1)}{\text{pen}_{\text{shape}}(J) - \text{pen}_{\text{shape}}(J-1)}$$

est au contraire quasi-constante à partir de 8 segments (Figure 1b). Ceci démontre de manière empirique que la log-vraisemblence de toutes les segmentations possibles est bien dans ce cas là une fonction linéaire de la forme de la pénalité pour des modèles sur-paramétrés.

Les pentes ont été estimées sur la plage $J = 6, ..., 20$ et nous avons obtenu un écart-type résiduel de 1.05 avec la forme de pénalité naïve J au lieu de 0.04 avec la forme de pénalité proposée. Le critère ICL exact ainsi que l'heuristique de pente avec la forme de pénalité proposée sélectionne 2 segments alors que l'heuristique de pente avec la forme de pénalité naïve sélectionne 3 segments et met du poids sur 4 segments (Table 1) ce qui est incohérent avec les différentes approches de validation illustrées dans Guédon (2013, 2015).

3.2 Données de forage

Les données consistent en 4050 mesures de la réponse magnétique nucléaire de roches forées. Le signal est grossièrement constant par morceaux où chaque segment correspond

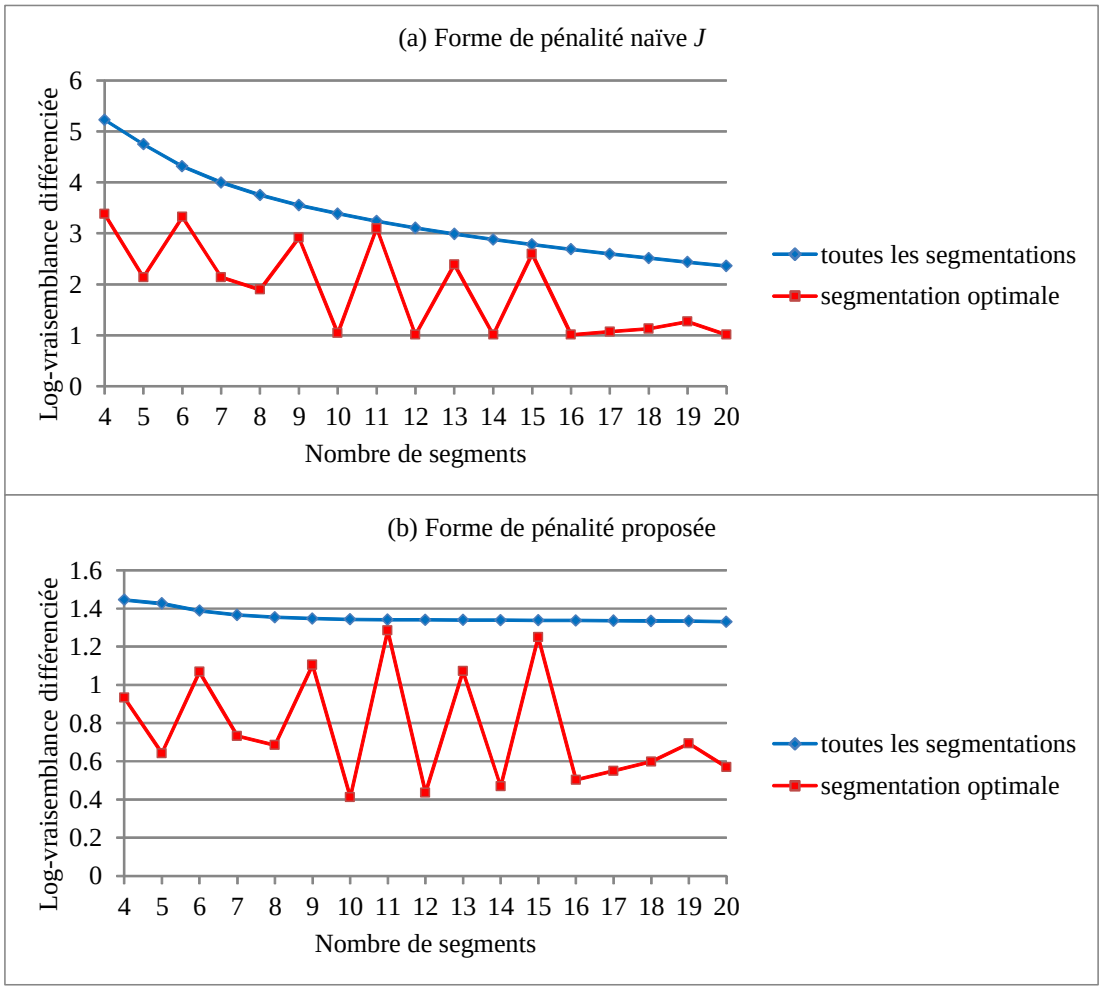

Figure 1: Log-vraisemblances différenciées : (a) forme de pénalité naïve J; (b) forme de pénalité proposée.

à un type de roche ayant des propriétés physiques constantes. Une rupture dans le signal correspond à un changement de roche. Nous avons estimé des modèles gaussiens de changement sur la moyenne et la variance à partir de ce signal échantillonné.

La pente a été estimée sur la plage $J = 30, ..., 80$. Le critère ICL exact sélectionne 18 ou 19 segments alors que l'heuristique de pente avec la forme de pénalité proposée sélectionne 16 segments (Table 2). Le modèle à 16 segments sélectionné par l'heuristique de pente est beaucoup plus cohérent avec l'analyse de l'espace de segmentations latent présentée dans Guédon (2013) que les modèles à 18 ou 19 segments sélectionnés par le critère ICL exact.

Table 1: Catastrophes dans les mines de charbon en Grande-Bretagne : Comparaison entre le critère ICL exact et l'heuristique de pente (SH) avec J comme forme de pénalité ($\text{pen}_{\text{shape}}0$) et la forme de pénalité proposée ($\text{pen}_{\text{shape}}1$). Les valeurs des critères ainsi que les probabilités *a posteriori* des modèles $P(\mathcal{M}_J|\mathbf{x})$ sont données pour chaque J.

J	1	2	3	4	5	6
ICL_J	-413.14	-358.7	-362.31	-369.34	-375.74	-382.01
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.855	0.141	0.004	0	0
SH_J ($\text{pen}_{\text{shape}}0$)	-419.68	-358.81	-357.04	-358.55	-361.01	-364.34
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.2	0.49	0.23	0.07	0.01
SH_J ($\text{pen}_{\text{shape}}1$)	-407.72	-360.19	-368.05	-377	-385.38	-393.42
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.98	0.02	0	0	0

Table 2: Données de forage : Comparaison entre le critère ICL exact et l'heuristique de pente (SH) avec la forme de pénalité proposée. Les valeurs des critères ainsi que les probabilités *a posteriori* des modèles $P(\mathcal{M}_J|\mathbf{x})$ sont données pour chaque J.

J	15	16	17	18	19	20	21
ICL_J	-69355.4	-69330	-69316.3	-69309.7	-69309.6	-69313.3	-69312.6
$P(\mathcal{M}_J\|\mathbf{x})$	0	0	0.014	0.378	0.403	0.063	0.088
SH_J	-69479.6	-69461.8	-69466	-69478.6	-69482.8	-69495.4	-69500.8
$P(\mathcal{M}_J\|\mathbf{x})$	0	0.89	0.11	0	0	0	0

Bibliographie

[1] Baudry, J.-P., Maugis, C. et Michel, B. (2012), Slope heuristics: overview and implementation, *Statistics and Computing*, 22(2), 455–470.

[2] Birgé, L. et Massart, P. (2001), Gaussian model selection, *Journal of the European Mathematical Society*, 3, 203–268.

[3] Guédon, Y. (2013), Exploring the latent segmentation space for the assessment of multiple change-point models, *Computational Statistics* 28(6), 2641–2678.

[4] Guédon, Y. (2015), Segmentation uncertainty in multiple change-point models, *Statistics and Computing*, 25(2), 303–320.

[5] Lebarbier, E. (2005), Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing*, 85(4), 717–736.

[6] Rigaill, G., Lebarbier, E. et Robin, S. (2012), Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and Computing*, 22(4), 917–929.