HAL open science

Détection de ruptures dans la moyenne d'un processus hétéroscédastique par validation-croisée

Sylvain Arlot, Alain Celisse

To cite this version:

Sylvain Arlot, Alain Celisse. Détection de ruptures dans la moyenne d'un processus hétéroscédastique par validation-croisée. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386677

HAL Id: inria-00386677

https://inria.hal.science/inria-00386677

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection de ruptures dans la moyenne d'un processus hétéroscédastique par validation-croisée

Sylvain Arlot[1] & Alain Celisse[2]

[1] *CNRS ; Willow Project-Team*
Laboratoire d'Informatique de l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
45, rue d'Ulm, 75230 Paris, France

[2] *UMR 518 AgroParisTech/INRA MIA,*
AgroParisTech
16 rue Claude Bernard, F-75231 Paris Cedex 05, France

Résumé

Dans ce travail, la sélection de modèle est utilisée afin de détecter des changements abrupts dans la moyenne d'un signal hétéroscédastique, sans connaissance *a priori* du type de bruit. À la différence de la plupart des méthodes actuelles qui s'effondrent dans un cadre hétéroscédastic, nous proposons une nouvelle famille d'algorithmes basés sur le rééchantillonnage, et en particulier la validation-croisée, qui demeurent performants y compris dans un tel cadre. Les phénomènes sous-jacents sont expliqués par des résultats théoriques récents, tandis qu'une vaste étude de simulations garantit d'un point de vue empirique les performances des algorithmes proposés. Enfin, une application aux données de puce CGH souligne l'intérêt de méthodes robustes à l'hétéroscédasticité, tout en illustrant l'applicabilité des algorithmes proposés à des données réelles.
Mots clés: détection de ruptures, sélection de modèles, validation-croisée, données hétéroscédastiques, segmentation de profils CGH.

Abstract

The main concern of this work is the problem of detecting abrupt changes in the mean of a heteroscedastic signal by model selection, without knowledge on the variations of the noise. Whereas most existing methods are not robust to heteroscedasticity, a new family of algorithms is proposed showing that resampling methods —in particular cross-validation— can be successful in this framework. These facts are supported by an extensive simulation study, together with recent theoretical results. Finally, an application to comparative genomic hybridization data is provided, showing that robustness to heteroscedasticity can indeed be required for their analysis.
Keywords: change-point detection, model selection, cross-validation, heteroscedastic data, CGH profile segmentation.

1 Introduction

The change-point detection deals with a stochastic process the distribution of which abruptly changes at some unknown instants. The purpose is then to recover the location of these changes.

Due to a wide range of applications, from voice recognition to biology and CGH (Comparative Genomic Hybridization) data analysis (Picard (2005)), the literature about change-point detection is very abundant (see Basseville and Nikiforov (1993)).

Assume we are given n points $(t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathbb{R}$, which are n successive observations of a signal Y_i at point t_i, and satisfying

$$Y_i = s(t_i) + \sigma_i \epsilon_i, \quad (\epsilon_i)_i \ \ i.i.d.\,, \ \mathbb{E}\left[\epsilon_i\right] = 0, \ \mathrm{Var}(\epsilon_i) = 1 \quad \text{and} \quad (\sigma_i)_i \in (\mathbb{R}_+)^n\,, \tag{1}$$

where s denotes the unknown regression function, $s(x) = \mathbb{E}\left[Y \mid X = x\right]$.

In this work, we focus on the problem of detecting changes in the mean s of the signal, assuming s is piecewise-constant, or more generally that s is smooth except at a few large jumps. Besides, we consider general heteroscedastic data, that is σ_i can strongly depend on i and this dependence is unknown. This is not usual in the change-point literature. Indeed, as pointed out by Lavielle (2005), three main problems can be distinguished: detect changes in the mean assuming the variance is constant, detect changes in the variance assuming the mean is constant, and detect changes in the whole distribution (including changes in the mean and changes in the variance). The present framework is a generalization of the first problem, meaning that we only aim at detecting changes in the mean whatever the variations of the variance, which can be seen as a nuisance parameter.

2 Statistical framework

2.1 Notations

Let $(t_1, Y_1), \ldots, (t_n, Y_n) \in [0,1] \times \mathbb{R}$ denote n random variables $Z_i = (t_i, Y_i)$ which are n successive observations of a signal Y at points $(t_i)_{1 \le i \le n}$, and $Z_{1,n} = (Z_1, \ldots, Z_n)$. Moreover, let us assume that the following model holds

$$Y_i = s(t_i) + \sigma(t_i)\epsilon_i, \quad (\epsilon_i)_i \ \ i.i.d.\,, \ \mathbb{E}\left[\epsilon_i\right] = 0, \ \mathrm{Var}(\epsilon_i) = 1, \tag{2}$$

where the t_is are deterministic points in $[0,1]$ and $\sigma : \ [0,1] \to \mathbb{R}_+$. Set $P_n = n^{-1} \sum_{i=1}^n \delta_{(t_i, Y_i)}$, the empirical measure of $Z_1, \ldots, Z_n$ and P, the distribution of a "new observation" $Z = (t, Y)$. Then,

$$P = P_n^t \otimes P_{Y|t}$$

where $P_n^t = n^{-1} \sum_{i=1}^n \delta_{t_i}$ and $P_{Y|t}$ denotes the conditional distribution of Y given t.

Set γ, the quadratic contrast such that for any *predictor* $f : \ [0,1] \to \mathbb{R}$, the prediction error is denoted by $P\gamma(f) = \mathbb{E}_{Z \sim P}\left[\gamma(f; Z)\right]$, where $\gamma(f; z) = (f(x) - y)^2$ with $z = (x, y)$. As we know, the prediction error reaches its minimum for $f = s$. Thus, the loss of f is defined by

$$\ell\left(s, f\right) = P\gamma(f) - P\gamma(s) \ \ \geq 0 \ \ .$$

2.2 Collection of models

Our strategy relies on the model selection point of view, like Lebarbier (2005). The sets (named *models*) of piecewise constant functions we use in the following are defined from partitions of

$[0,1]$. For a given set of indices of cardinality $2^n - 1$, let $\mathcal{I} = \{I(m) \mid m \in \mathcal{M}_n\}$ denote the set of all the partitions $I(m)$ of $[0,1]$ built from the subdivision $t_0 = 0 < t_1 < \ldots < t_n = 1$. For each $I(m) \in \mathcal{I}$, we define the model S_m as the linear space of piecewise constant functions built from $I(m) = (I_\lambda)_{\lambda \in \Lambda(m)}$, where $\Lambda(m)$ denotes a set of indices associated with m and $I_\lambda = [t_k, t_l)$, $0 \leq k < \ell \leq n$. Set D_m the dimension of the model S_m and $\mathcal{D} = \{D_m \mid m \in \mathcal{M}_n\}$. A model S_m is then defined by a list of $D_m - 1$ breakpoints.
In the sequel, we often consider the set of models with dimension $1 \leq D \leq n$, which is denoted by $\mathcal{M}(D) = \{m \mid D_m = D\}$.

Given a model S_m and some observations (represented by the empirical measure P_n), the least-square estimator is denoted by

$$\widehat{s}_m = \widehat{s}_m(P_n) = \mathrm{ERM}(S_m, P_n) := \mathrm{Argmin}_{t \in S_m} \{P_n \gamma(t)\} \quad .$$

The notation ERM stresses that we use an algorithm (the minimization of the empirical risk) which takes in input a model and some data and outputs an estimator $\widehat{s}_m$. The empirical risk minimizer $\widehat{s}_m$ associated with S_m is also called *segmentation*.

Our choice of the best segmentation is made by use of model selection in order to design an algorithm A: $P_n \mapsto A(P_n) = \widehat{m}(P_n)$, such that $\widetilde{s}(P_n) = \mathrm{ERM}\left(S_{\widehat{m}(P_n)}, P_n\right)$. The typical optimality criterion we have in mind is an oracle inequality, with high probability for instance:

$$\ell\left(s, \widetilde{s}(P_n)\right) \leq C \inf_{m \in \mathcal{M}_n} \{\ell\left(s, \widehat{s}_m(P_n)\right) + R(n,m)\} \tag{3}$$

where C is close to 1 and $R(n,m)$ is a remainder term.

3 Breakpoint locations

In Lavielle (2005) and Lebarbier (2005), a usual strategy to find the best segmentation is to dissociate the search for the best segmentation with a given number of breakpoints from the choice of this number of breakpoints. In the sequel, we consider the number K of breakpoints as known and we are looking for the change-point locations.

3.1 Empirical risk minimization deficiency

Lavielle (2004) explains the empirical risk minimization (ERM) is a reliable means to get the best segmentation with K change-points. We therefore perform the ERM algorithm over $\mathcal{M}(D)$, where $D = K - 1$.

Algorithm 1.

$$\widehat{m}(D) = \mathrm{Argmin}_{m \in \mathcal{M}(D)} P_n \gamma(\widehat{s}_m) \;\; = \;\; ERM\left(\widetilde{S}_D, P_n\right),$$

where γ denotes the least-square contrast and $\widetilde{S}_D := \cup_{m \in \mathcal{M}(D)} S_m$.

At this stage, the effective computation of $\widehat{m}(D)$ is performed thanks to dynamic programming (Bellman and Dreyfus (1962)), which can be computed with a complexity of order $\mathcal{O}\left(n^2\right)$.

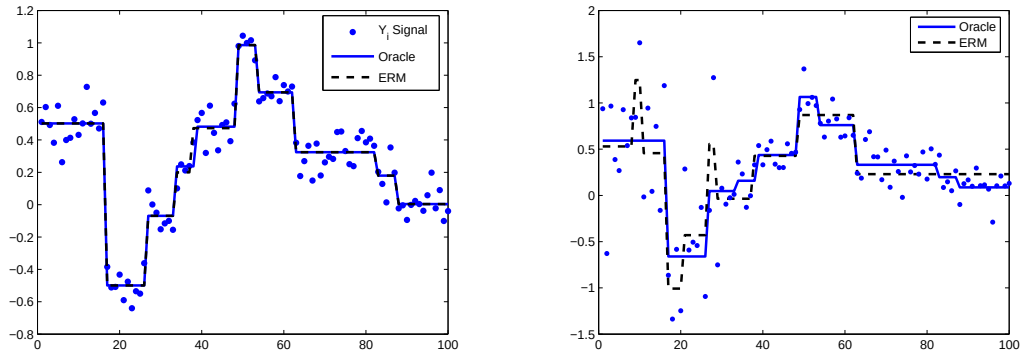

Figure 1: Segmentation provided by the ERM algorithm compared with the oracle in the homoscedastic setup (left panel) and with heteroscedastic data (Right panel).

From our simulation setting (Celisse Ph.D. (2008)), Figure 1 displays segmentations yielded by the ERM algorithm with either homoscedastic (left panel), or heteroscedastic (right panel) data.
Under homoscedasticity, ERM provides segmentations close to the oracle (the best estimator ($\widehat{s}_m$) we can get, knowing s), whereas with heteroscedastic data, segmentations have breakpoints in noisy regions, where we have only a poor confidence in their locations. This can be explained by some considerations on the expectation of the optimized criterion.

Lemma 1.
Homoscedastic:

$$\mathbb{E}\left[P_n\gamma\left(\widehat{s}_m\right)\right] = \ell\left(s, s_m\right) - \sigma^2\frac{D_m}{n} + \sigma^2. \tag{4}$$

Heteroscedastic:

$$\mathbb{E}\left[P_n\gamma\left(\widehat{s}_m\right)\right] = \ell\left(s, s_m\right) - \frac{1}{n}\sum_{\lambda}\left(\sigma_\lambda^r\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2, \tag{5}$$

where s_m denotes the orthogonal projection of s onto S_m, and $\left(\sigma_\lambda^r\right)^2 := 1/n_\lambda \sum_{i=1}^n \sigma_i^2 \mathbb{1}_{I_\lambda}(t_i)$ with $n_\lambda = \mathrm{Card}\left(\{k \mid t_k \in I_\lambda\}\right)$.

For the sake of completeness, we also mention expressions of the risk of $\widehat{s}_m$ for both homoscedastic (6) and heteroscedastic (7) data:

$$\mathbb{E}\left[\|Y - \widehat{s}_m\|_n^2\right] = \ell\left(s, s_m\right) + \sigma^2\frac{D_m}{n} + \sigma^2, \tag{6}$$

$$\mathbb{E}\left[\|Y - \widehat{s}_m\|_n^2\right] = \ell\left(s, s_m\right) + \frac{1}{n}\sum_{\lambda}\left(\sigma_\lambda^r\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2. \tag{7}$$

The expectation in (4) classically breaks down into a bias-term ($\ell\left(s, s_m\right)$) and a variance-term, up to some constants with respect to the models. This expression is similar to that of the risk

(6), except the negativity of the variance-term. However since all the models we compare share the same dimension, these are only distinguished on the basis of their bias-terms. This explains why with homoscedastic data, ERM provides nearly optimal segmentations on average.

With heteroscedastic data, the variance-term in (5) differs from (4) in that $\sum_\lambda (\sigma_\lambda^r)^2$ takes into account the variance spreading along the intervals I_λ of the original partition $I(m)$. Every models with the same dimension do not have the same variance-terms, which entails that both the bias- and the variance-terms contribute to the choice of the best segmentation, unlike what happens in the homoscedastic setting. As a consequence, for reasonable models with similar bias-terms, the ERM algorithm rather puts breakpoints in noisy regions in order to minimize $-\sum_\lambda (\sigma_\lambda^r)^2$ in the variance-term, which is supported by the right panel of Figure 1.

3.2 Cross-validation

The deficiency of the ERM algorithm can also be interpreted as an elementary consequence of the fact that the empirical risk only takes into account how well an estimator fits the data, without paying attention to the variance (heterogeneity). We now explore a possible way to overcome this trouble, using leave-one-out cross-validation (Loo) (see Stone (1974)), to get an estimator of the risk of $\widehat{s}_m$, which does include an estimator of the variance-term.

Since Loo estimates the risk of each $\widehat{s}_m$, we suggest to replace the ERM by the following algorithm.

Algorithm 2.

$$\widehat{m}_{\text{Loo}}(D) := \text{Argmin}_{m|D_m=D} \widehat{R}_1(ERM(S_m)).$$

Figure 2 displays the segmentation yielded by Loo in the same setting as the right panel of Figure 1. This segmentation has no breakpoint in the noisy region where ERM segmentation used to overfit. Cross-validation actually seems to overcome this overfitting phenomenon. This

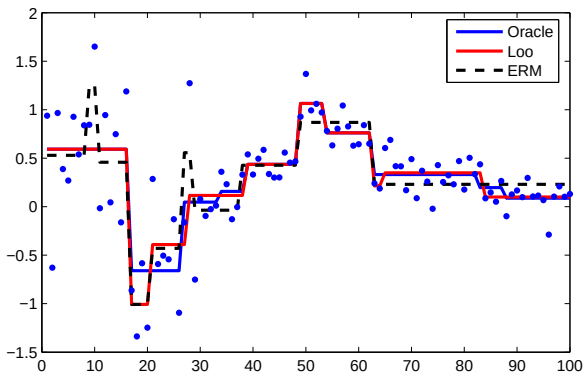

Figure 2: Segmentations provided by ERM and Loo compared with the oracle in a heteroscedastic setting.

phenomenon is justified by theoretical results in Celisse's Ph.D. (2008), where other resampling algorithms are also studied.

4 Choice of the number of breakpoints

In Celisse (2008), it is similarly argued that cross-validation improves upon penalized criteria like that of Birgé and Massart (2001) (derived from a gaussian homoscedastic framework) with heteroscedastic data. We also provide a new interpretation of this usual penalty and illustrate the better behaviour of cross-validation algorithms in order to choose the actually unknown number of segments.

Finally, we present a new family of resampling-based algorithms, which are both computationally tractable and robust to heteroscedasticity. From the observed signal, these algorithms first outperform upon ERM in the choice of the best segmentation for each dimension, and enable to select the number of breakpoints in a reliable way.
The behaviour of these algorithms is extensively assessed (Celisse (2008)) in a wide simulation study as well as by comparison with well understood algorithms on real CGH microarray data (Picard (2005)).

Bibliographie

[1] Basseville, M. and Nikiforov, N. (1993) The detection of abrupt changes – Theory and Applications. *Prentice and Hall, Information and System Sciences Series.*
[2] Bellman, R. E. and Dreyfus, S. E. (1962) Aplied dynamic programming. *Princeton.*
[3] Birgé, L. and Massart, P. (2001) Gaussian model selection. *Journal of the European Mathematical Society*, 3, 203–268.
[4] Celisse, A. (2008) Model selection via cross-validation in density estimation, regression, and change-point detection, Ph.D. Université Paris-Sud 11.
[5] Lavielle, M. (2005) Using penalized contrasts for the change-point problem. *Signal processing*, 85, 1501–1510.
[6] Lebarbier, E. (2005) Detecting mutiple change-points in the mean of a Gaussian process by model selection. *Signal processing*, 85, 717–736.
[7] Picard, F. (2005) Process segmentation/clustering Application to the analysis of array CGH data, Ph.D., Université Paris-Sud 11.
[8] Stone, M. (1974) Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36, 111–147.