Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

# New efficient algorithms for multiple change-point detection with reproducing kernels

A. Celisse [c,e,*], G. Marot [a,c], M. Pierre-Jean [a,b], G.J. Rigaill [b,d]

[a] Univ. Lille Droit et Santé, EA 2694 - CERIM, F-59000 Lille, France
[b] UMR 8071 CNRS - Université d'Evry - INRA, Laboratoire Statistique et Génome Evry, France
[c] Inria Lille Nord Europe, Équipe-projet Inria MODAL, France
[d] Institute of Plant Sciences Paris-Saclay, UMR 9213/UMR1403, CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Diderot, Sorbonne Paris-Cité, France
[e] Univ. Lille Sciences et Technologies, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France

ARTICLE INFO

*Article history:*
Received 30 August 2016
Received in revised form 4 July 2018
Accepted 6 July 2018
Available online 17 July 2018

*Keywords:*
Kernel method
Gram matrix
Nonparametric change-point detection
Model selection
Algorithms
Dynamic programming
DNA copy number
Allele B fraction

ABSTRACT

Several statistical approaches based on reproducing kernels have been proposed to detect abrupt changes arising in the full distribution of the observations and not only in the mean or variance. Some of these approaches enjoy good statistical properties (oracle inequality, consistency). Nonetheless, they have a high computational cost both in terms of time and memory. This makes their application difficult even for small and medium sample sizes ($n < 10^4$). This computational issue is addressed by first describing a new efficient procedure for kernel multiple change-point detection with an improved worst-case complexity that is quadratic in time and linear in space. It is based on an exact optimization algorithm and deals with medium size signals (up to $n \approx 10^5$). Second, a faster procedure (based on an approximate optimization algorithm) is described. It relies on a low-rank approximation to the Gram matrix and is linear in time and space. The resulting procedure can be applied to large-scale signals ($n \geq 10^6$). These two procedures (based on the exact or approximate optimization algorithms) have been implemented in R and C for various kernels. The computational and statistical performances of these new algorithms have been assessed through empirical experiments. The runtime of the new algorithms is observed to be faster than that of other considered procedures. Finally, simulations confirmed the higher statistical accuracy of kernel-based approaches to detect changes that are not only in the mean. These simulations also illustrate the flexibility of kernel-based approaches to analyze complex biological profiles made of DNA copy number and allele B frequencies.

## 1. Introduction

In this paper we consider the multiple change-point detection problem (Brodsky and Darkhovsky, 2013) where the goal is to recover abrupt changes arising in the distribution of a sequence of $n$ independent random variables $X_1, \ldots, X_n$ observed at respective time $t_1 < t_2 < \cdots < t_n$.

*State-of-the-art.* Many parametric models (Normal, Poisson,…) have been proposed (Hautaniemi et al., 2003; Rigaill et al., 2012; Cleynen and Lebarbier, 2014). These models allow detecting different types of changes: in the mean, in the

[*] Corresponding author at: Univ. Lille Sciences et Technologies, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France.
*E-mail address:* alain.celisse@math.univ-lille1.fr (A. Celisse).

https://doi.org/10.1016/j.csda.2018.07.002

variance and in both the mean and variance (see also Hautaniemi et al., 2003, Jong et al., 2003, Picard et al., 2005). Efficient algorithms and heuristics have been proposed for these models. Some of them scale in $\mathcal{O}(n\log(n))$ or even in $\mathcal{O}(n)$. In practice, these parametric approaches have proven to be successful for various application fields (see for example Hocking et al., 2013, Cleynen et al., 2014a). However one of their main drawbacks is their lack of flexibility. For instance, any change of distributional assumption requires the development of a new dedicated inference scheme.

By contrast, the recently proposed kernel change-point detection approach (Harchaoui and Cappé, 2007; Arlot et al., 2012) is more generic. It has the potential to detect any change arising in the distribution, which is not easily captured by standard parametric models. More precisely in this approach, the observations are first mapped into a Reproducing Kernel Hilbert Space (RKHS) through a kernel function (Aronszajn, 1950). The difficult problem of detecting changes in the distribution is then recast as simply detecting changes in the mean element of observations in the RKHS.

One practical limitation of this kernel-based approach is its considerable computational cost owing to the use of a $n \times n$ Gram matrix combined with a dynamic programming algorithm (Auger and Lawrence, 1989). More precisely (Harchaoui and Cappé, 2007) described a dynamic programming algorithm to recover the best segmentation from 1 to $D_{\max}$ segments. They claim that their algorithm has a $\mathcal{O}(D_{\max}n^2)$ time complexity. However, the latter is not described in full details and its straightforward implementation is not efficient. First, it requires the storage of a $n \times n$ cost matrix (personal communication with the first author of Harchaoui and Cappé (2007) who was kind enough to send us his code). Thus the algorithm has a $\mathcal{O}(n^2)$ space complexity, which is a severe limitation with nowadays sample sizes. For instance analyzing a signal of length $n = 10^5$ requires storing a $10^5 \times 10^5$ matrix of doubles, which takes 80 GB. Second, computing the cost matrix is not straightforward. In fact simply using formula (8) of Harchaoui and Cappé (2007) to compute each term of this cost matrix leads to an $\mathcal{O}(n^4)$ time complexity.

*Contributions.* The present paper contains several contributions to the computational aspects and the statistical performance of the kernel change-point procedure introduced by Arlot et al. (2012).

The first one is to describe a new algorithm to simultaneously perform the dynamic programming step of Harchaoui and Cappé (2007) and also compute the required elements of the cost matrix on the fly. On the one hand, this algorithm has a complexity of order $\mathcal{O}(D_{\max}n^2)$ in time and $\mathcal{O}(D_{\max}n)$ in space (including both the dynamic programming and the cost matrix computation). We also emphasize that this improved space complexity comes without an increased time complexity. This is a great algorithmic improvement upon the change-point detection approach described by Arlot et al. (2012) since it allows the efficient analysis of signals with up to $n = 10^5$ data-points in a matter of a few minutes on a standard laptop.

On the other hand, our approach is generic in the sense that it works for any positive semidefinite kernels. Importantly one cannot expect to exactly recover the best segmentations from 1 to $D_{\max}$ segments in less than $\mathcal{O}(D_{\max}n^2)$ without additional specific assumptions on the kernel. Indeed, computing the cost of a given segmentation has already a time complexity of order $\mathcal{O}(n^2)$.

It is also noticeable that our algorithm can be applied to other existing strategies such as the so-called ECP (Matteson and James, 2014). To be specific, we show that the *divisive clustering algorithm* it is based on and that provides an approximate solution with a complexity of order $O(n^2)$ in time and space can be replaced by our algorithm that provides the exact solution with the same time complexity but a reduced memory complexity.

Our second contribution is a new algorithm dealing with larger signals ($n > 10^5$) based on a low-rank approximation to the Gram matrix. This computational improvement is possible at the price of an approximation. It returns approximate best segmentations from 1 to $D_{\max}$ segments with a complexity of order $\mathcal{O}(D_{\max}p^2n)$ in time and $\mathcal{O}((D_{\max}+p)n)$ in space, where $p$ is the rank of the approximation.

The last contribution of the paper is the empirical assessment of the statistical performance of the KCP procedure introduced by Arlot et al. (2012). This empirical analysis is carried out in the biological context of detecting abrupt changes from a two-dimensional signal made of DNA copy numbers and allele B fractions (Lai, 2012). The assessment is done by comparing our approach to state-of-the-art alternatives on resampled real DNA copy number data (Pierre-Jean et al., 2014; Matteson and James, 2014). This illustrates the versatility of the kernel-based approach. To be specific this approach allows the detection of changes in the distribution of such complex signals without explicitly modeling the type of change we are looking for. The described procedure has been implemented in an R package called *KernSeg* (Marot et al., 2018)

The remainder of the paper is organized as follows. In Section 2, we describe our kernel-based framework and detail the connection between detecting abrupt changes in the distribution and model selection as described in Arlot et al. (2012). A slight generalization of the KCP procedure (Arlot et al., 2012) is also derived in Section 2.5 by introducing a new parameter $\ell$ encoding an additional constraint on the minimal length of any candidate segment. This turns out to be particularly useful in low signal-to-noise ratio settings. The versatility of this kernel-based framework is emphasized in Section 2.6 where it is shown how the ECP approach (Matteson and James, 2014) can be rephrased in terms of kernels. Our main algorithmic improvements are detailed and justified in Section 3. We empirically illustrate the improved runtime of our algorithm and compare it to the ones of ECP and RBS in Section 3.1.3. In Section 3.2 we detail our faster (but approximate) algorithm used to analyze larger profiles ($n > 10^5$). It is based on the combination of a low-rank approximation to the Gram matrix and the binary segmentation heuristic (Yang, 2012). An empirical comparison of the runtimes of the exact and approximate algorithms is provided in Section 3.2.3. Finally, Section 4 illustrates the statistical performance of our kernel-based change-point procedure in comparison with state-of-the-art alternatives in the context of biological signals such as DNA copy numbers and allele B fractions (Lai, 2012).