RETROSPECTIVE MUTIPLE CHANGE-POINT ESTIMATION WITH KERNELS

Zaïd Harchaoui & Olivier Cappé

LTCI, Télécom Paris & CNRS
46 rue Barrault, 75013 Paris, France
(e-mail: `zharchao, cappe` at `enst.fr`)

ABSTRACT

This contribution proposes an extension of the classic dynamic programming algorithm for detecting jumps in noisily observed piecewise-constant signals. The proposed algorithm operates (virtually) in a reproducing kernel Hilbert space through the use of an arbitrary kernel mapping. The resulting approach provides a computationally efficient an versatile tool for segmenting complex signals whose structure is not appropriately captured by standard parametric models.

1. INTRODUCTION

Segmentation tasks are pervasive in various fields, ranging from audio [13] to EEG segmentation [5]. The goal is to segment the signal into several homogeneous segments of variable durations, in which some quantity remains approximately constant over time. This issue was addressed in a large literature, mainly from a Bayesian point of view (see [10] and references therein), where almost as many parametric models as definitions of intra-segment homogeneity were developed. Although such models often proved successful in practice, they require extensive data modelling knowledge and sophisticated numerical methods for training. These limitations invite the use of kernel-based methods, which already demonstrated good performances in real experiments for other unsupervised learning tasks [19, 20]. A kernel-based segmentation method would offer the nonparametric flexibility of kernel-based approaches, while remaining relatively easy to train.

Indeed, recently kernel-based approaches were proposed for online segmentation (sequential change-point detection), and showed beyond state-of-the-art performances for segmentation of audio signals [9]. However online segmentation approaches, though computationally attractive, require a fine tuning of the sliding window size. Moreover, as opposed to retrospective approaches, they do not take into account the whole signal at once for change-point estimation. Offline segmentation of piecewise-constant signals observed in Gaussian white noise can be efficiently performed in polynomial time, for a fixed number of segments, using the dynamic programming approach originally proposed in [11, 2]. This algorithm however relies on specific parametric modelling assumptions which limit its practical usefulness. We will show how kernels can be incorporated into this algorithm by using a kernel-based measure of intra-segment homogeneity. Hence, our approach widens the range of possible applications *via* the kernel trick, while keeping the simplicity of the original algorithm. This allows a great versatility in modelling complex piecewise-stationary signals. We refer to this approach as KCpE which stands for Kernel Change-point Estimation.

In Section 2, we introduce the signal model as well as the kernel-based variance-like criterion (which is referred to as the intra-segment scatter in this context). In Section 3, we describe the resulting dynamic programming algorithm. In Section 4, we briefly discuss the issues of selecting the number of change-points and the kernel. Finally, we present in Section 5 experimental on both artificial and real datasets.

2. MULTIPLE CHANGE-POINT ESTIMATION

2.1. Model

Offline segmentation or retrospective change-point estimation is the problem of partitioning a sequence of vector-valued observations $(Y_t)_{t=1,\ldots,n}$ into, say, K segments

$$(Y_t)_{t=1,\ldots,n} = \bigcup_{k=0,\ldots,K-1} (Y_t)_{t=\tau_k+1,\ldots,\tau_{k+1}} \tag{1}$$

where each segment $[\tau_k + 1, \ldots, \tau_{k+1}]$ is considered homogeneous (and $\tau_0 = 0, \tau_K = n$).

Following the principle at the core of several recent development in machine learning [19, 20], the observations are mapped in an abstract space, namely a reproducing kernel Hilbert space (rkhs) $\mathcal{H}$ associated with a reproducing kernel $k(\cdot,\cdot)$ and a feature map $\Phi(\cdot)$. For notational convenience, we denote by $Y_t^\Phi = \Phi(Y_t)$ the image in the feature space through the Aronsjazn map $\Phi(Y) = k(Y,\cdot)$ of a given observation Y_t. We assume that, *in the rkhs*, the sequence $(Y_t^\phi)_{t=1,\ldots,n}$ is such that

$$Y_t^\phi = m_k + \epsilon_t \quad \text{with } \tau_k + 1 \le t \le \tau_{k+1} \tag{2}$$

where m_k are elements of $\mathcal{H}$ and ϵ_t is an isotropic white $\mathcal{H}$-noise [3]. Our goal is to compute estimates $\hat{\tau}_1, \dots, \hat{\tau}_{K-1}$ of the true change-point instants $\tau_1, \dots, \tau_{K-1}$. Note that because the nuisance parameters $(m_k)_{k=0,\dots,K-1}$ are abstract quantities, the proposed algorithm does not rely on their direct estimation.

2.2. Intra-segment scatter

2.2.1. *Abstract definitions*

To deal with statistics in infinite-dimensional spaces, we introduce respectively the mean element and the covariance operator [3] [12]. We denote by $\mathbb{E}$ the expectation with respect to the distribution of the random variable Y and defined the mean operator m_Φ on $\mathcal{H}$, for all $f \in \mathcal{H}$, as:

$$\langle m_\Phi, f \rangle_{\mathcal{H}} = \mathbb{E}[f(Y)] = \mathbb{E}\langle Y^\Phi, f \rangle_{\mathcal{H}} \tag{3}$$

Using the above definition of the mean, the covariance operator Σ_Φ on $\mathcal{H}$ is defined, for all $f \in \mathcal{H}$, as:

$$\langle f, \Sigma_\Phi f \rangle_{\mathcal{H}} = \mathbb{E}[\langle f, Y^\Phi - m_\Phi \rangle_{\mathcal{H}} \langle Y^\Phi - m_\Phi, f \rangle_{\mathcal{H}}] \tag{4}$$

By analogy with the intra-cluster scatter matrix in multivariate clustering, we define the **average scatter** $S \in \mathbb{R}$ of Y as:

$$S(Y) = \mathbb{E}[\langle Y^\Phi - m_\Phi, \Sigma_\Phi^{-1}(Y^\Phi - m_\Phi) \rangle_{\mathcal{H}}] \tag{5}$$

2.2.2. *Empirical counterparts*

The empirical mean over $(Y_s)_{s=1,\dots,d}$ is given by:

$$\hat{m}_{\mathbf{Y}} = \frac{1}{d} \sum_{s=1}^{d} Y_s^\Phi = \frac{1}{d} \sum_{s=1}^{d} k(Y_s, \cdot) \tag{6}$$

Assuming isotropic unit-variance $\Sigma_\Phi \equiv \mathrm{Id}$ in feature space, the empirical average scatter of $(Y_s)_{s=1,\dots,d}$ is then given by:

$$\begin{aligned} \hat{S}(Y_1, \dots, Y_d) &= \frac{1}{d} \sum_{s=1}^{d} \langle Y_s^\Phi - \hat{m}_\Phi, Y_s^\Phi - \hat{m}_\Phi \rangle_{\mathcal{H}} \\ &= \frac{1}{d} \sum_{s=1}^{d} \left\| k(Y_s, \cdot) - \frac{1}{d} \sum_{r=1}^{d} k(Y_r, \cdot) \right\|_{\mathcal{H}}^2 \end{aligned} \tag{7}$$

In the following sections, we consider the unnormalized quantities $\hat{V}(Y_1, \dots, Y_d) = d\hat{S}(Y_1, \dots, Y_d)$, which we simply call scatter from now on. By expanding the norm (7), the scatter can be expressed without explicitly determining $\hat{m}_\Phi$ [19]:

$$\begin{aligned} \hat{V}(Y_{t+1}, \dots, Y_{t+d}) = &\sum_{s=1}^{d} \langle k(Y_s, \cdot), k(Y_s, \cdot) \rangle \\ &- \frac{1}{d} \sum_{r=1}^{d} \sum_{s=1}^{d} \langle k(Y_r, \cdot), k(Y_s, \cdot) \rangle \end{aligned} \tag{8}$$

2.3. Minimum overall scatter

In order to estimate the change-point instants, we look for the sequence $\hat{\tau}_1, \dots, \hat{\tau}_{K-1}$ minimizing the sum of intra-segment scatter. The objective can then be formulated as:

$$\underset{\tau_1, \dots, \tau_{K-1}}{\text{Minimize}} \sum_{k=0}^{K-1} \hat{V}(Y_{\tau_k+1}, \dots, Y_{\tau_{k+1}}) \tag{9}$$

Fortunately all quantities involved in the above objective are easily computable from the data Gram matrix as discussed below.

3. COMPUTATIONAL ASPECTS

3.1. Scatter computation

Let us denote by $\mathbf{K}_{[t+1,t+d]} = [k(Y_i, Y_j)]_{i,j=t+1,\dots,t+d}$ the block of the Gram matrix $\mathbf{K}$ corresponding to the segment $[t+1, t+d]$. According to (8), the intra-segment scatter is simply given by $\hat{V}(Y_{t+1}, \dots, Y_{t+d}) = \text{trace}(\mathbf{K}_{[t+1,\dots,t+d]}) - \frac{1}{d} \mathbf{1}_d^\top \mathbf{K}_{[t+1,\dots,t+d]} \mathbf{1}_d$.

3.2. Optimality equation

Because the objective function in (9) is additive, it can be minimized using a dynamic programming algorithm of complexity $\mathcal{O}(Kn^2)$ as described in [16]. We mention that a slightly different arrangement of the recursions was recently presented in [15]. In the approach of [15], estimation and model selection are handled simultaneously by using a global penalized criterion which include both the sum of intra-segment scatters and a penalty term for the number of segments. The recursion then determines the best segmentation of all sub-portions of the signal that start at $t = 1$ and end at $t = 2, \dots, n$. In contrast, the recursion described below, following [15], determines the optimal segmentations of the whole signal for all number of segments from 2 to K. If needed, model selection can then be handled separately, typically by using a complexity penalty to select the optimal number of segments. Note that since this latter aspect is not the main focus of the paper, we choose to illustrate the performance of the proposed method only when K is fixed (see section 5).

Let $I_k(t)$ denote the minimal value of the objective on the portion $[1, t]$ of the signal assuming k segments:

$$I_k(t) = \min_{\substack{\tau_1 < \dots < \tau_{k-1} \\ \tau_k = t}} \sum_{j=0}^{k-1} \hat{V}(Y_{\tau_j+1}, \dots, Y_{\tau_{j+1}}) \tag{10}$$

The dynamic programming recursion exploits the observation that

$$\begin{aligned} I_k(t) &= \min_{\substack{\tau_{k-1} \\ \tau_k = t}} \min_{\tau_1 < \cdots < \tau_{k-2}} \sum_{j=0}^{k-1} \hat{V}(Y_{\tau_j+1}, \ldots, Y_{\tau_{j+1}}) \\ &= \min_{\tau_{k-1}} \Big(I_{k-1}(\tau_{k-1}) + \hat{V}(Y_{\tau_{k-1}}, \ldots, Y_t) \Big) \end{aligned} \qquad (11)$$

Thus a dynamic programming algorithm allows to compute the minima $I_k(t)$ of the criterion for each $k = 1, \ldots, K$ and for all $t = 1, \ldots, n$ by forward recursion *via* the optimality condition above.

3.3. Backward recursions

Let $\tau_k(t)$ denote a minimizer of $I_{k+1}(t)$. Once the change-point instants estimates $\hat{\tau}_K, \ldots, \hat{\tau}_{K-l+1}$ are found, the next estimate $\hat{\tau}_{K-l}$ is computed through $\hat{\tau}_{K_l} = \tau_{K-l}\big(\hat{\tau}_{K_l+1}\big)$.

4. FURTHER ISSUES

4.1. Choosing the number of change-points

Determining the number of change-points is a difficult but well-studied problem which is usually dealt with using a penalized version of the least-square criterion [18, 17], with a penalty $C_n k$ proportional to the number of change-points k. As discussed in section 3, the penalized criterion can even be minimized directly by dynamic programming using the approach of [15]. Note however that the calibration of C_n is usually based on asymptotic arguments (obtained by letting n tend to infinity) which may be questionable for moderate length signals. In addition, the scatter criterion considered here corresponds to a variance estimate, after mapping into an infinite-dimensional space (the rkhs), which could have an impact on the appropriate form of C_n.

Another popular approach consists in using a Bayesian model by introducing a prior distribution on the number of change-points and a conditional prior on change-point locations. Interestingly, in many cases of interest the fully Bayesian inference (which includes marginalization with respect to the segment parameters rather than simply maximization) remains feasible [1, 10]. The approach proposed here being in essence nonparametric, it is seems more difficult to build on these Bayesian approaches.

Finally, in several practical applications of interest the number of change-points is known or fixed by exogenous information rather than determined directly from the signal. We consider several examples of this situation in the next section.

4.2. Kernel selection

The kernel will generally be selected using some prior knowledge on the structure of the signals to be analyzed. The kernel parameter (the so-called “bandwidth” parameter in the examples to be discussed below) can be determined in supervised mode, as is usually done in clustering, by selecting the kernel which yields the change-point estimates closest to the target, known, change-points. In the sequel, we shall use so-called isotropic Gaussian kernels $k(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2)$. Furthermore, since kernel design issues are not the primary focus of this contribution, we have chosen to the tune the bandwidth parameter σ^2 simply using the plugin rule-of-thumb $\sigma = 1.06\, \hat{s}\, n^{-1/5}$ ($\hat{s}$ being the empirical variance) classically used for density estimation with a Gaussian Parzen window [14].

5. EXPERIMENTAL RESULTS

5.1. Islands dataset

To illustrate the potential of the method, we consider a simulated two-dimensional signal built from the so-called *islands* distributions. The signal consists of 200 observations, whose eight segments have variable durations and are sampled alternatively from one of the semi-rings displayed on Figure 1. KCpE is used with a Gaussian kernel (purposefully selected with a slightly suboptimal bandwidth) and is compared to the raw dynamic programming algorithm (DP) which uses the weighted norm associated with the actual two-dimensional covariance matrix of the data. The true change-point locations are depicted in black, while DP's and KCpE's change-point estimates are displayed in green and red, respectively. As one can notice from Figure 2, KCpE gives fairly accurate estimates for change-point locations, whereas a standard multivariate approach fails to retrieve segments boundaries.

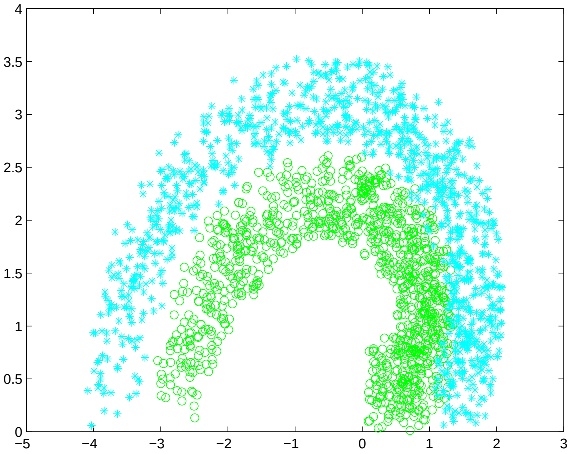

Fig. 1. Islands dataset

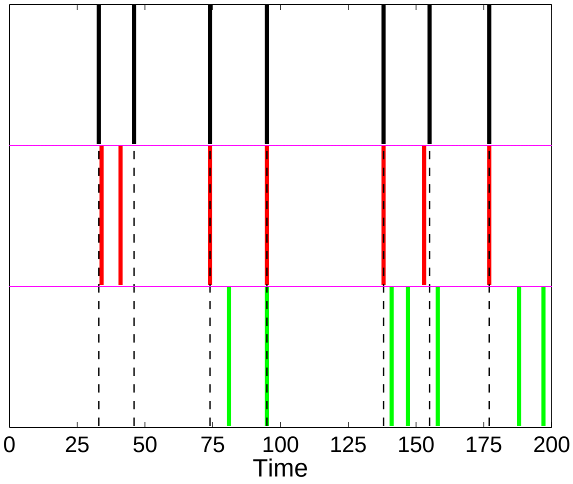

Fig. 2. Estimated change-points on the islands datasets, with, from top to bottom, the actual change-point locations, change-point estimates for KCpE and rDP

5.2. Brain-Computer Interface data

Data arising from Brain-Computer Interface (BCI) experiments naturally exhibit temporal structure, which could serve as a benchmark for investigating the relevance of our algorithm on real-world data. We considered a dataset proposed in BCI competition III [6] acquired during 4 non-feedback sessions on 3 normal subjects. Each subject was asked to perform the following tasks: imagination of repetitive self-paced left then right hand movements and generation of words beginning with the same random letter. The subject performed a given task for about 15 seconds and then switched randomly to another prescribed task. Input data consist of 96-dimensional vectors of features. The challenge originally cast the task into a online supervised multi-class classification framework, and provided 4 labelled samples for each subject.

We chose to perform temporal segmentation in a completely unsupervised fashion. In order to evaluate the performance of our algorithm from a retrospective point of view, we provide results in terms of classification accuracy. Each segment is considered as a sample of a given class. Classification accuracy then corresponds to the proportion of correctly assigned points at the end of the segmentation process.

Results averaged over the four sessions are provided in the table below. A Gaussian kernel tuned as described in Section 4.2 was fed into our algorithm for subsequent segmentation. Comparison (yet unfair since KCpE works in an unsupervised setting) with results obtained by a supervised muti-class (one-versus-one) Support Vector Machine are also provided. Performance of KCpE is competitive, especially given that KCpE was used as a blind segmentation algorithm, in contrast to the SVM which was trained on one session and tested on the three remaining, as required in the original design of the benchmark [6]. In this particular case, the algorithm that is trained from the data (SVM) but ignores the temporal homogeneity of EEG signals performs slightly worse than the unsupervised KCpE approach.

	Subject 1	Subject 2	Subject 3
KCpE	79%	74%	61%
SVM	76%	69%	60%

Table 1. Average classification accuracy for each subject

5.3. Music signals

For the purpose of indexation, a music piece may be temporally segmented into several sections highlighting its structure through dynamic, tonal or timbral characteristics. A standard approach [13], working in an online setting, runs a window-limited change detection algorithm along the signal, which raises an alarm when crossing a boundary between two sections. However one might rather consider a retrospective approach, where the signal is taken as a whole for subsequent change-point detection instead of being locally scanned through a window of limited width.

We investigated the performance of KCpE on a database of 100 full-length "pop music" signals, whose manual segmentation is available. Results for the Kernel Change Detection algorithm (KCD) as described in [8] are also displayed. As in [13], scores output by KCD were detrended and normalized. The decision threshold was chosen so as to output the correct known number of segment boundaries. In order to evaluate the performance of our algorithm, as in Section 5.2 we provide results in terms of classification accuracy. Table 2 suggests that the proposed method is indeed quite competitive in this context.

KCpE	KCD
72±3.2%	68±3.8%

Table 2. Average classification accuracy

Moreover, we observed in experiments that when the correct number of boundaries is overestimated (oversegmentation regime), our method is less sensitive to artefacts than KCD. Typically, KCD is prone to suggest spurious segments when the threshold is below the range of values yielding the correct number of segments. Conversely, when the number of change-points is underestimated, the proposed method will typically tend to ignore short epidemic changes [7], i.e. changes producing very short segments whose duration is negligible when compared to the length of the whole signal, due to the use of a global optimality criterion.

6. CONCLUSION

We proposed an efficient kernel-based nonparametric approach for retrospective multiple change-point estimation. Experimental results of blind segmentation on both BCI data and audio music signals are very promising. Statistical consistency results as developed in [17, 4] would be helpful in providing a firmer theoretical ground and providing some insights on the form of penalization appropriate in the case where the number of change-points is unknown.

7. REFERENCES

[1] D. Barry and J. Hartigan. Product partition models for change point problems. *Annals of Statistics*, 20:260–279, 1992.

[2] R. Bellman. On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, 4(6), 1961.

[3] D. Bosq. *Linear Processes in Function Spaces: Theory and Applications*. Springer, 2000.

[4] L. Boysen, A. Kempe, A. Munk, V. Liebscher, and O. Wittich. Consistencies and rates of convergence of jump penalized least squares estimators. *Annals of Statistics*, In revision.

[5] B. Brodsky and B. Darkhovsky. *Non-parametric statistical diagnosis: problems and methods*. Kluwer Academic Publishers, 2000.

[6] S. Chiappa and J. d. R. Millan. Data set V mental imagery, multi-class, 2004. `http://ida.first.fraunhofer.de/projects/bci/competition_iii/`.

[7] M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley, 1998.

[8] M. Davy, F. Désobry, and S. Canu. Estimation of minimum measure sets in reproducing kernel hilbert spaces and applications. IEEE ICASSP, 2006.

[9] F. Désobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, August 2005.

[10] P. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16:203–213, 2006.

[11] W. D. Fisher. On grouping for maximum homogeneity. *Journal of the American Statistical Society*, 53:789–798, 1958.

[12] K. Fukumizu, F. Bach, and A. Gretton. Statistical convergence of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(8), 2007.

[13] O. Gillet, S. Essid, and G. Richard. On the correlation of automatic audio and visual segmentation of music videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2007.

[14] W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis (2nd ed.)*. Springer, 2007.

[15] B. Jackson, J. Scargle, D. Barnes, A. Sundararajan, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12:105–108, 2005.

[16] S. M. Kay. *Fundamentals of statistical signal processing: detection theory*. Prentice-Hall, Inc., 1993.

[17] M. Lavielle and E. Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000.

[18] C.-B. L. Lee. Estimating the number of change-points in a sequence of independent random variables. *Statistics and Probability Letters*, 25:241–248, 1995.

[19] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[20] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.