HAL open science

Robust Retrospective Multiple Change-point Estimation for Multivariate Data

Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, Olivier Cappé

To cite this version:

Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, Olivier Cappé. Robust Retrospective Multiple Change-point Estimation for Multivariate Data. Statistical Signal Processing Workshop, Jun 2011, Nice, France. pp.405–408, 10.1109/SSP.2011.5967716 . hal-00564410v2

HAL Id: hal-00564410

https://hal.science/hal-00564410v2

Submitted on 10 Feb 2011

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.

Robust Retrospective Multiple Change-point Estimation for Multivariate Data

Alexandre Lung-Yut-Fong, Céline Lévy-Leduc, Olivier Cappé

February 3, 2011

Abstract

We propose a non-parametric statistical procedure for detecting multiple change-points in multidimensional signals. The method is based on a test statistic that generalizes the well-known Kruskal-Wallis procedure to the multivariate setting. The proposed approach does not require any knowledge about the distribution of the observations and is parameter-free. It is computationally efficient thanks to the use of dynamic programming and can also be applied when the number of change-points is unknown. The method is shown through simulations to be more robust than alternatives, particularly when faced with atypical distributions (e.g., with outliers), high noise levels and/or high-dimensional data.

Keywords: Change-point estimation, multivariate data, Kruskal-Wallis test, robust statistics, joint segmentation.

1 Introduction

Retrospective multiple change-point estimation consists in partitioning a non-stationary series of observations into several contiguous stationary segments of variable durations [1]. It is particularly appropriate for analyzing a posteriori time series in which the quantity driving the behavior of the time series jumps from one level to another at random instants called change-points. Such a task, also known as temporal signal segmentation in signal processing, arises in many applications, ranging from EEG to speech processing and network intrusion detection [2, 3].

The classic setting for change-point detection and estimation is the case where the sequence of observations is univariate and the use of the least-squares criterion is considered to be adequate, for instance, in the presence of Gaussian noise [4, 5]. When there are multiple change-points to be detected, the composite least-squares criterion can be efficiently optimized using a dynamic programming recursion whose computational cost scales like the square of the number of observations [6, 7, 8]. The approach can be generalized to multivariate settings [9].

In this work, we consider methods that can be applied to high-dimensional data without requiring strong prior knowledge of the underlying distribution of the observations. To the best of our knowledge, there are very few approaches in the literature that are suitable for this task. The first option is the kernel-based approach of [10] which uses a general kernel metric to evaluate distances between multivariate observations. This approach is powerful but requires the choice of an appropriate kernel function and, furthermore, it is not robust in situations where the noise level is very significant as will be illustrated in Section 4. The other relevant reference is [11] which combines univariate rank statistics using a weighting matrix derived from an asymptotic analysis. This approach, which is based on the principle of the Mann-Whitney

Wilcoxon two-sample test, is however limited to the case where, at most, a single change-point is present in the observed signal. Our contribution with this paper is to show that the test statistic of [11] can be modified so as to deal with multiple change-points (in analogy with the way the classical Kruskal-Wallis test generalizes the Mann-Whitney Wilcoxon statistic to multiple group testing [12]. Furthermore, we show that the proposed test statistic is amenable to dynamic programming and can thus be optimized efficiently despite the combinatorial nature of the multiple change-point detection task. We will show in particular that the proposed method, termed *dynMKW* (for "dynamic programming multivariate Kruskal-Wallis"), is more efficient than greedy strategies based on repeated use of single change-point tests as suggested in [13].

When dealing with multiple change-points, one needs a criterion for estimating the number of change-points which is an instance of a model selection problem. Traditional approaches include the use of a complexity penalty that is added to the test statistic [5, 7, 14] or the use of priors on the locations of change-points when adopting the Bayesian point of view, see, e.g., [15, 16] and references therein. An original approach adopted in [17, 18] is to consider the penalty associated to the use of the ℓ_1 norm (or block-ℓ_1 norm in [18]). In this paper, we propose a simple construction which relies on the so-called "slope heuristic" and is a variant of the procedure described in [14]. Although heuristic this approach is preferable to off-the-shelf Schwarz-like penalties (AIC, BIC, etc.) which are most often found to be badly-calibrated for the type of problems considered here.

The paper is organized as follows. The statistical methodology for testing homogeneity in several groups of multivariate observations is described in Section 2. The procedure for estimating the change-points derived from this testing procedure is then described in Section 3. In Sections 4, we report the results of numerical experiments carried out on synthetic and real data.

2 Testing homogeneity within several groups of multivariate data

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n L-dimensional observed random vectors such that $\mathbf{X}_j = (X_{j,1}, \ldots, X_{j,L})'$, where A' denotes the transpose of the matrix A. We assume in the following that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent. In this section, we first consider testing the hypothesis (H_0) that K given groups, $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}, \mathbf{X}_{n_1+1}, \ldots, \mathbf{X}_{n_2}, \ldots, \mathbf{X}_{n_{K-1}+1}, \ldots, \mathbf{X}_{n_K}$, have the same distribution, where we will use the convention that $n_0 = 1$ and $n_K = n$.

For j in $\{1, \ldots, n\}$ and ℓ in $\{1, \ldots, L\}$, let $R_j^{(\ell)}$ denote the rank of $X_{j,\ell}$ among $(X_{1,\ell}, \ldots, X_{n,\ell})$ that is $R_j^{(\ell)} = \sum_{k=1}^n \mathbf{1}_{\{X_{k,\ell} \leq X_{j,\ell}\}}$. Also define $\bar{R}_k^{(\ell)}$, for k in $\{0, \ldots, K-1\}$, by

$$\bar{R}_k^{(\ell)} = (n_{k+1} - n_k)^{-1} \sum_{j=n_k+1}^{n_{k+1}} R_j^{(\ell)} .$$

We propose to use the following statistic:

$$T(n_1, \ldots, n_{K-1}) = \frac{1}{n^2} \sum_{k=0}^{K-1} (n_{k+1} - n_k) \bar{R}_k' \, \hat{\Sigma}_n^{-1} \, \bar{R}_k , \tag{1}$$

where $\bar{R}_k$ is defined by

$$\bar{R}_k = (\bar{R}_k^{(1)} - n/2, \ldots, \bar{R}_k^{(L)} - n/2)' , \tag{2}$$

and $\hat{\Sigma}_n$ is the $L \times L$ matrix whose (ℓ, ℓ')–th element is defined by

$$\hat{\Sigma}_{n,\ell\ell'} = \frac{1}{n^2} \sum_{i=1}^n \{R_i^{(\ell)} - n/2\}\{R_i^{(\ell')} - n/2\} . \tag{3}$$

Note that $\hat{\Sigma}_{n,\ell\ell'}$ can be rewritten as

$$\hat{\Sigma}_{n,\ell\ell'} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{F}_{n,\ell}(X_{i,\ell}) - 1/2\}\{\hat{F}_{n,\ell'}(X_{i,\ell'}) - 1/2\} \ , \tag{4}$$

where $\hat{F}_{n,\ell}(t) = n^{-1}\sum_{j=1}^{n}\mathbf{1}_{\{X_{j,\ell}\leq t\}}$ denotes the empirical cumulative distribution function (c.d.f.) of the ℓth coordinate $X_{1,\ell}$. The matrix $\hat{\Sigma}_n$ thus corresponds to an empirical estimate of the covariance matrix Σ with general term

$$\Sigma_{\ell\ell'} = \mathrm{Cov}\left(F_\ell(X_{1,\ell}); F_{\ell'}(X_{1,\ell'})\right) \ , \ 1 \leq \ell, \ell' \leq L \ , \tag{5}$$

F_ℓ denoting the c.d.f. of $X_{1,\ell}$ which we shall assume in the sequel to be continuous.

Observe that (1) extends to the multivariate setting the classical Kruskal-Wallis test [12] which applies to the univariate data. Indeed, when $L = 1$, (1) is equivalent to

$$T(n_1, \ldots, n_{K-1}) = \frac{12}{n^2}\sum_{k=0}^{K-1}(n_{k+1} - n_k)\left(\bar{R}_k^{(1)} - n/2\right)^2 \ , \tag{6}$$

where we have replaced $\hat{\Sigma}_{n,11}$ by $\Sigma_{11} = \mathrm{Var}(F_1(X_{1,1})) = \mathrm{Var}(U) = 1/12$, where U is a uniform random variable on $[0,1]$, F_1 being a continuous cumulative distribution function. In the case where there is only one change-point, i.e. when $K = 2$, (1) reduces to the test statistic proposed in [11] to extend the Mann-Whitney/Wilcoxon rank test to multivariate data. We state below (without proof for reasons of space) a result which shows that (1) is properly normalized in the sense that under the null hypothesis it converges, as n increases, to a fixed limiting distribution that only depends on K and L.

Theorem 1. *Assume that $(\mathbf{X}_i)_{1\leq i\leq n}$ are $\mathbb{R}^L$-valued i.i.d. random vectors such that, for all ℓ, the c.d.f. F_ℓ of $X_{1,\ell}$ is a continuous function. Assume also that for $k = 0, \ldots, K-1$, there exists λ_k in $(0,1)$ such that $(n_{k+1} - n_k)/n \to \lambda_{k+1}$, as n tends to infinity. Then, $T(n_1, \ldots, n_{K-1})$ defined in (1) satisfies*

$$T(n_1, \ldots, n_{K-1}) \xrightarrow{d} \chi^2\left((K-1)L\right) \ , \ \textit{as } n \to \infty \ , \tag{7}$$

where d denotes the convergence in distribution and $\chi^2((K-1)L)$ denotes the chi-square distribution with $(K-1)L$ degrees of freedom.

3 Detecting change-point locations

In this section, we consider applying the test statistic described in the previous section to determine the positions of the segment boundaries $n_1, \ldots, n_{K-1}$, starting with the simpler case where the number of change-points K is known.

3.1 Estimation of a known number of change-points

To estimate the change-point locations, we propose to maximize the statistic $T(n_1, \ldots, n_{K-1})$ defined in (1) with respect to the segment boundaries $n_1, \ldots, n_{K-1}$:

$$(\hat{n}_1, \ldots, \hat{n}_{K-1}) = \underset{1\leq n_1<\cdots<n_{K-1}\leq n}{\mathrm{argmax}} \ T(n_1, \ldots, n_{K-1}) \ . \tag{8}$$

Of course, direct maximization of (8) is impossible as it corresponds to a combinatorial problem whose complexity is exponential in K. We can however perform this maximization efficiently using the dynamic programming algorithm described in [6]. More precisely, using the notations

$$\Delta(n_k + 1 : n_{k+1}) = (n_{k+1} - n_k)\bar{R}_k' \, \hat{\Sigma}_n^{-1} \, \bar{R}_k \ ,$$

and

$$I_K(p) = \max_{1<n_1<\cdots<n_{K-1}<n_K=p} \sum_{k=0}^{K-1} \Delta(n_k+1:n_{k+1}) \ ,$$

we have

$$I_K(p) = \max_{n_{K-1}} \left\{ I_{K-1}(n_{K-1}) + \Delta(n_{K-1}+1:p) \right\} \ . \tag{9}$$

Thus, for solving the optimization problem (8), we proceed as follows. We start by computing the $\Delta(i:j)$ for all (i,j) such that $1 \leq i < j \leq n$. All the $I_1(E)$ are thus available for $E = 2, \ldots, n$. Then $I_2(E)$ is computed by using the recursion (9) and so on. The overall numerical complexity of the procedure is thus proportional to $K \times n^2$.

3.2 Estimation of the number of change-points

In practice the number of change-points can rarely be assumed to be known. Although it is not the main focus of this paper, we propose a heuristic algorithm to find the optimal number of change-points. Values of the statistics $I_K(n)$, for $K = 0, \ldots, K_{\max}$, are first computed using the procedure described in the previous section. The algorithm is based on the principle that in the presence of $K^\star \geq 1$ change-points, if $I_K(n)$ is plotted against K, the resulting graph can be decomposed into two distinct linear trends: a first one, for $K = 0, \ldots, K^\star$ where the criterion is growing rapidly; and a second one, for $K = K^\star, \ldots, K_{\max}$, where the criterion is barely increasing. Hence, for each possible value of K in $K = 1, \ldots, K_{\max}$, we compute least square linear regressions for both parts of the graph (before and after K); the estimated number of change-points is the value of K that yields the best fit, that is, the value for which the sum of the residual sums of squares computed on both parts of the graph is minimal (see Figure 1). The case $K = 0$ is treated separately and the procedure that has just been described is used only when the value of $T(\hat{n}_1)$ is significant, based on the asymptotic p-value for the single change-point case obtained in [11].

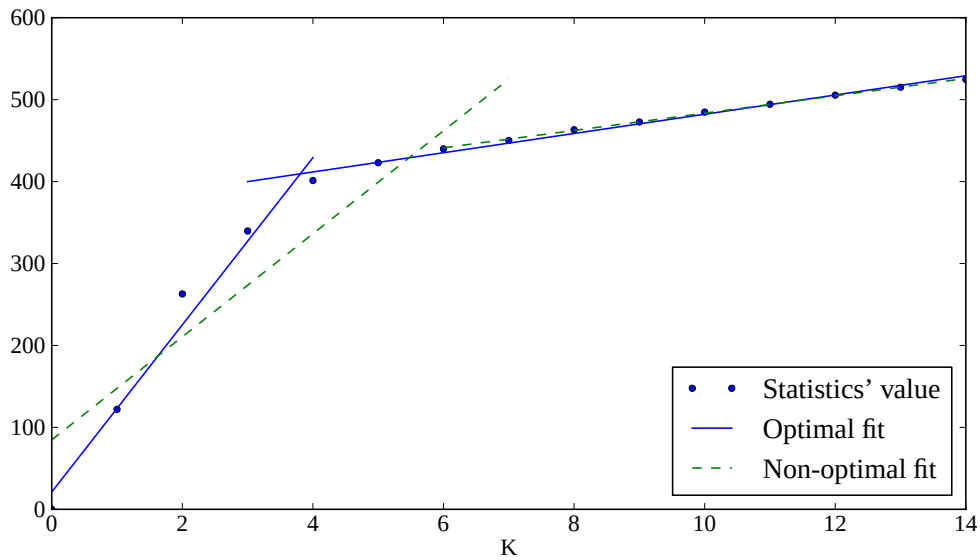

Figure 1: Determining the optimal number of change-points. Here, the actual number of change-points is $K^\star = 4$; the optimal regression is displayed in solid lines, while a non-optimal alternative (for $K = 6$) is displayed in dashed lines.

4 Numerical experiments

In this section, the performance of the *dynMKW* algorithm is assessed using simulated signals and we next consider its application to data arising from a biological context.

4.1 Simulated data

The simulation reported is this section is based on a fixed 5-dimensional piecewise constant signal of length 500 with a predefined set of four shared breakpoints. Note that this signal reproduces an important feature of many applications (see [3] as well as the example of the cancer data below) in that only a subset of its coordinates actually change simultaneously. To this baseline signal is added some moderately correlated Gaussian noise with marginal variance σ^2, see Figure 2 for an example corresponding to a marginal SNR of 16 dB, where the SNR is defined as the ratio of the jump amplitude over σ. The performance of the algorithms is assessed from 1000 Monte-Carlo replications of the data for each value of σ^2 and measured in terms of *precision* (proportion of correctly estimated changes found among the detected change-points) and *recall* (proportion of actual change-points that have been correctly estimated). We used a ± 1 sample tolerance for deciding whether a change-point has been correctly estimated but the conclusion were qualitatively the same for larger values of this tolerance.

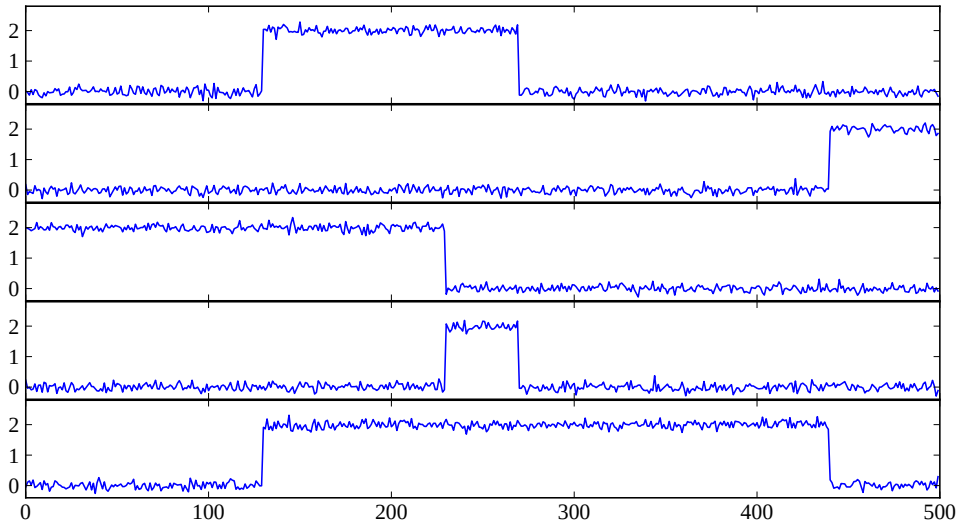

Figure 2: Baseline signal with superimposed noise at an SNR of 16 dB.

4.1.1 Known number of change-points

We compare our algorithm with two other different methods – which also rely on a dynamic programming step – for multiple change-points detection: first, with the kernel-based approach of [10] that computes "intra-segment scatter" using an isotropic Gaussian kernel function and second, the parametric method corresponding to the multivariate Gaussian assumption.

As shown in Figure 3 left, the parametric method ("Linear") yields the best results, as the noise added to the signal is indeed Gaussian. Unsurprisingly, the proposed *dynMKW* approach performs slightly worse in this scenario but is certainly preferable than the kernel test, particularly when the level of noise increases. In Figure 3 right, we repeat the experiment with

significant contamination by outliers. The *dynMKW* method demonstrates its robustness as, contrary to both alternatives, its performance barely suffer from the presence of outliers.

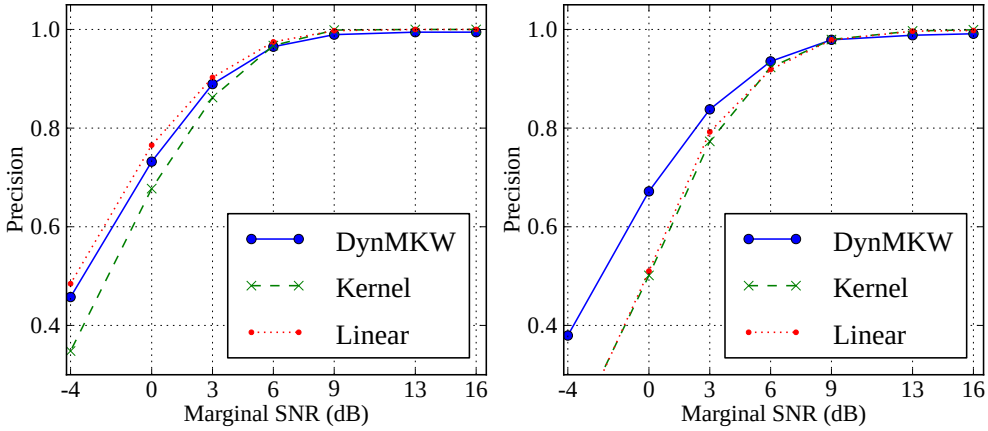

Figure 3: Precision curves of the *dynMKW*, kernel-based and linear algorithms at different levels of noise for a known number of change-points. Left: signals with no outliers; right: signals with 5% of outliers (whose variance is 10 dB higher than that of the background noise). Recall curves are not displayed as they are identical in the known number of change-points case.

4.1.2 Unknown number of change-points

Using the same signal as previously, we suppose in the experiments presented here that the number of change-points is unknown. Our algorithm, combined with the approach for estimating the number of change-points presented in Section 3.2 is compared with a greedy hierarchical method suggested by [13]: Breakpoints are detected by iteratively testing for the presence of a single change-point in the sub-segments determined in previous stages. This method ("*Vost*") is applied using (8) restricted to $K = 2$ groups as a single change-point detector, where the asymptotic p-value determined in [11] is computed to decide whether the current segment should be segmented any further (segmentation stops whenever the p-value is larger than a threshold, say 1 or 5%).

Results are shown in Figure 4 with precision and recall measures. It is observed that *dynMKW* outperforms the *Vost* procedure both in precision and recall. From a more qualitative perspective, the *Vost* procedure tends to systematically over-segment the signal while our method, when faulting, is more inclined to miss some change-points.

4.2 Bladder tumor CGH profiles

We consider the bladder cancer micro-array aCGH dataset[1] of [18]. The objective here is to jointly segment data recorded from different subjects so as to robustly detect regions of frequent deletions or additions of DNA which could be characteristic of cancer. Each of the 57 profiles provides the relative quantity of DNA for 2215 probes measured on 23 chromosomes. We ran the *dynMKW* algorithm on each chromosome separately, thus treating 23 different 9- to 57-dimensional signals (depending on the selected groups of patients at different stages of cancer) of length 50 to 200 (the number of probes varies for each chromosome). Results are shown for a group of 9 individuals corresponding to Stage T1 of a tumor. In Figure 5, the smoothed version of the whole set of probes resulting from the segmentation is displayed, while a focus is given

[1] `http://cbio.ensmp.fr/~frapaport/CGHfusedSVM/index.html`

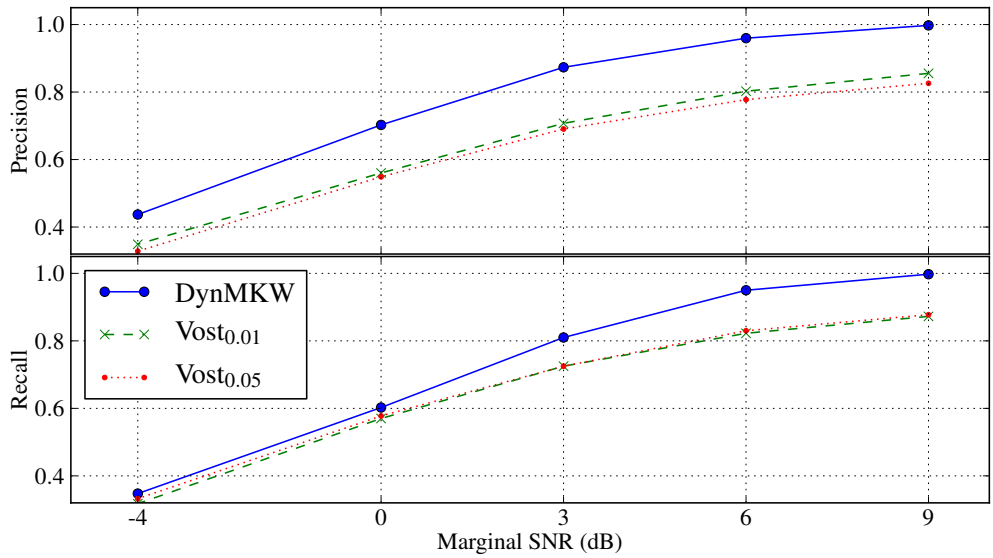

Figure 4: Precision (top) and recall (bottom) for *dynMKW* with estimates of the number of change-points and Multi-Rank with binary segmentation (Vost.) with threshold of 0.01 and 0.05.

on the 10th chromosome in Figure 6. In each case, the segmentation result is represented by a signal which is constant (and equal to the mean of the data) within the detected segments.

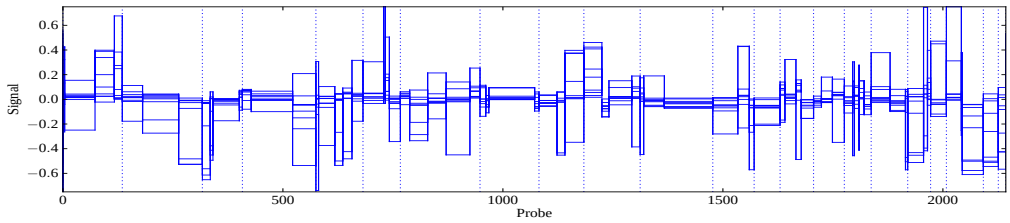

Figure 5: Superposition of the smoothed bladder tumor aCGH data for 9 individuals in Stage T1 cancer that results from the segmentation. Vertical dashed lines represent the separation between the different chromosomes.

References

[1] B. E. Brodsky and B. S. Darkhovsky, *Nonparametric Methods in Change-Point Problems*, Kluwer Academic Publisher, 1993.

[2] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Applications*, Prentice-Hall, 1993.

[3] C. Lévy-Leduc and F. Roueff, "Detection and localization of change-points in high-dimensional network traffic data," *Annals of Applied Statistics*, vol. 3, no. 2, pp. 637–662, 2009.

[4] Y.C. Yao and S. T. Au, "Least-squares estimation of a step function," *Sankhya: The Indian Journal of Statistics, Series A*,, vol. 51, no. 3, pp. 370–381, 1989.

[5] M. Lavielle and E. Moulines, "Least-squares estimation of an unknown number of shifts in a time series," *Journal of Time Series Analysis*, vol. 21, no. 1, pp. 33–59, 2000.

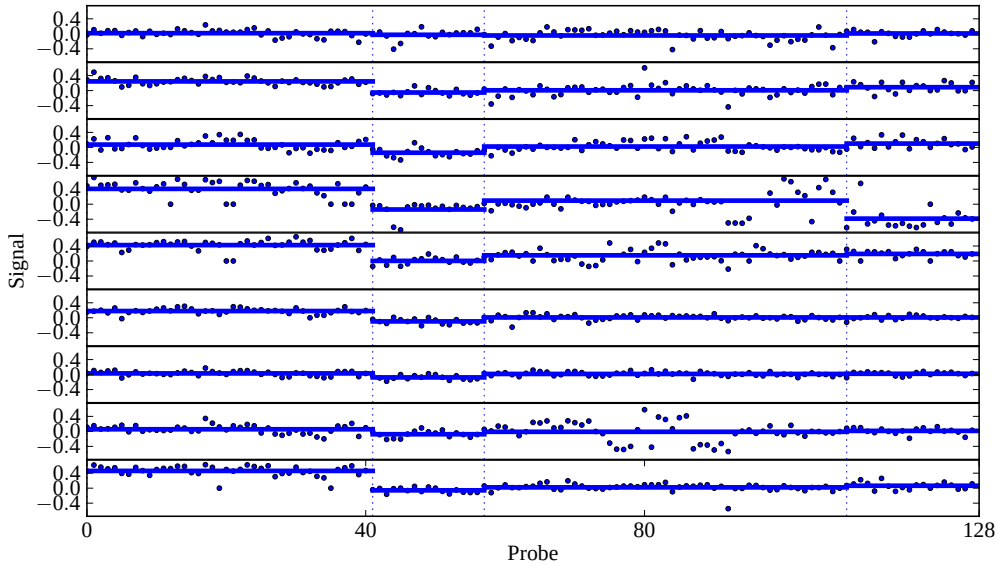

Figure 6: Data for 9 individuals in Stage T1 bladder cancer with superimposed segmentation for chromosome 10. The dashed vertical lines correspond to the estimated segment boundaries.

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

[7] E. Lebarbier, "Detecting multiple change-points in the mean of a Gaussian process by model selection," *Signal Process.*, vol. 85, pp. 717–736, 2005.

[8] 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*, vol. 12, pp. 105–108, 2005.

[9] M. S. Srivastava and K. J. Worsley, "Likelihood ratio tests for a change in the multivariate normal mean," *J. Amer. Statist. Assoc.*, vol. 81, no. 393, pp. 199–204, 1986.

[10] Z. Harchaoui and O. Cappé, "Retrospective multiple change-point estimation with kernels," in *IEEE Workshop on Statistical Signal Processing*, 2007.

[11] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, "Robust changepoint detection based on multivariate rank statistics," in *To appear in IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2011.

[12] W.H. Kruskal and W.A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[13] L. Yu. Vostrikova, "Detecting disorder in multidimensional random processes," *Soviet Math. Dokl.*, vol. 24, pp. 55–59, 1981.

[14] M. Lavielle, "Using penalized contrasts for the change-points problems," *Signal Process.*, vol. 85, no. 8, pp. 1501–1510, 2005.

[15] J.J. Ruanaidh and W.J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer, 1996.

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

[17] Z. Harchaoui and C. Lévy-Leduc, "Multiple change-point estimation with a total variation penalty," *J. Amer. Statist. Assoc.*, vol. 105, no. 492, pp. 1480–1493, 2010.

[18] J.P. Vert and K. Bleakley, "Fast detection of multiple change-points shared by many signals using group LARS," in *Advances in Neural Information Processing Systems 23*, 2010.