RANK-BASED MULTIPLE CHANGE-POINT DETECTION IN MULTIVARIATE TIME SERIES

F. Harlé[*†], *F. Chatelain*[†], *C. Gouy-Pailler*[*], *S. Achard*[†]

[*] CEA, LIST, LADIS, 91191 Gif-sur-Yvette CEDEX, France
[†] University of Grenoble, GIPSA-Lab, 11 rue des Mathématiques, 38402 St Martin d'Hères, France

ABSTRACT

In this paper, we propose a Bayesian approach for multivariate time series segmentation. A robust non-parametric test, based on rank statistics, is derived in a Bayesian framework to yield robust distribution-independent segmentations of piecewise constant multivariate time series for which mutual dependencies are unknown. By modelling rank-test p-values, a pseudo-likelihood is proposed to favour change-points detection for significant p-values. A vague prior is chosen for dependency structure between time series, and a MCMC method is applied to the resulting posterior distribution. The Gibbs sampling strategy makes the method computationally efficient. The algorithm is illustrated on simulated and real signals in two practical settings. It is demonstrated that change-points are robustly detected and localized, through implicit dependency structure learning or explicit structural prior introduction.

***Index Terms*—** Rank statistics, joint segmentation, dependency structure learning, Bayesian inference, MCMC methods, Gibbs sampling

1. INTRODUCTION

Detecting change-points in multivariate data is a classical but major issue in many fields, like bioinformatics, industrial monitoring, finance, or genomics. The literature in signal processing suggests different approaches for the multiple change-point detection problem. Existing monovariate methods (as presented in [1–3]) can be extended to the multivariate case. A fused lasso latent feature model has been proposed, for example in genomics [4]. Other approaches introduce a hierarchical Bayesian model [5]. Our framework is the off-line joint segmentation of piecewise constant time series: the events are abrupt changes in the distribution of the observations, their number and localizations are unknown.

We consider a matrix $\boldsymbol{X}$ of K time series of N points: $x_{j,i}$ is the observation of sensor j at time i, and is a change-point if a change occurs at $t \in]i, i+1]$, then $x_{j,i}$ is the last point of a segment. The nature of these points is represented by an indicator variable $r_{j,i}$, element of the matrix $\boldsymbol{R}$:

$$r_{j,i} = \begin{cases} 1 & \text{if } x_{j,i} \text{ is a change-point,} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

with the convention that $r_{j,1} = r_{j,N} = 1, \forall j \in \{1, ..., K\}$. $\boldsymbol{R}$ denotes the segmentation of $\boldsymbol{X}$ by the change-points in each signal, it is the parameter to estimate. Another parameter of interest is the implicit dependency structure that express the links between the signals, assuming that if a change-point appears in a signal j, it occurs simultaneously on the signals that depend of j. Then the probabilities that a change-point is shared across several signals is represented in the vector $\boldsymbol{P}$. Following the Bayes' rule, we get the posterior for $\boldsymbol{R}$ and $\boldsymbol{P}$:

$$f(\boldsymbol{R}, \boldsymbol{P}|\boldsymbol{X}) \propto L_*(\boldsymbol{X}|\boldsymbol{R}) f(\boldsymbol{R}|\boldsymbol{P}) f(\boldsymbol{P}). \quad (2)$$

We propose to combine a non-parametric test to Bayesian inference to express this statistical model. We chose the Wilcoxon signed-rank test, that is often presented as robust to several non-normal distributions, and makes the model free of strong prior. Its power is discussed in [6], the test is shown to be optimum for various alternatives in hypothesis testing.

This paper is organised as follows. The methodology is described in section 2. First, we explain how the Wilcoxon test is applied to define the pseudo-likelihood $L_*(\boldsymbol{X}|\boldsymbol{R})$. Then the choice of priors for $\boldsymbol{R}$ and $\boldsymbol{P}$ is detailed to model the dependency structure between time series. Finally a marginalized posterior density for the parameter $\boldsymbol{R}$ is expressed, and the procedure to estimate the maximum *a posteriori* (MAP), with a Markov Chain Monte Carlo (MCMC) method and a Gibbs sampling strategy, is given. In section 3, we present the results of the application on synthetic and real data. The discussion about the model is in section 4.

2. BERNOULLI DETECTOR MODEL

The parameters to estimate are the change-points configurations at each time index $i \in \{1, ..., N\}$: they are modelled by the vectors of indicators defined in (1): $R_i = (r_{1,i}, ..., r_{K,i})^T$, for signals 1 to K, where configuration $R_i \in \{0, 1\}^K$. This leads to consider the parameter $\boldsymbol{R} = (R_1, ..., R_N) \in \{0, 1\}^{K \times N}$, which is inferred within a Bayesian framework.

Here the data term is derived from the rank statistics, and a prior on $\boldsymbol{P}$ is chosen to infer the dependencies between signals. All observations in $\boldsymbol{X}$ are assumed to be mutually independent, but have similar statistical properties within a segment, like the median. In the following parts, we derive the joint probability distribution for $\boldsymbol{R}$, $\boldsymbol{P}$ and other hyperparameters given the data $\boldsymbol{X}$.

2.1. Change-point model

Each change-point probability is computed by the Wilcoxon rank sum (*aka* Mann-Whitney) test. In [7], the authors present a statistic inspired by this test on multivariate time series to detect multiple change-points, but they are supposed to occur simultaneously on all time series, assuming a fully connected structure. In our model, the Wilcoxon test is applied on each observation $x_{j,i}$, for a given segmentation in $\boldsymbol{R}$, between the two samples defined by the segments $s_1 = \{x_{j,i^-+1}, ..., x_{j,i}\}$, of length n_1, and $s_2 = \{x_{j,i+1}, ..., x_{j,i^+}\}$, of length n_2, where i^- and i^+ denotes the previous and next change positions respectively in signal j. The Wilcoxon statistic is defined from the sum of the ranks, denoted R_1 and R_2 for the observations of s_1 and s_2 respectively, in the global segment $s = (s_1, s_2)$:

$$U = \min(U_1, U_2) \text{ with } U_k = n_k n_l + \frac{n_k(n_k+1)}{2} - R_k, \quad (3)$$

where $(k, l) = (1, 2)$ or $(2, 1)$. For small n_1 and n_2, the resulting p-values are tabulated, otherwise they are computed from a normal approximation of a standardized value of U. It yields a statistical test to reject the null hypothesis H_0 that the difference between the observations of each segment is symmetrically distributed around zero; this is for instance the case when the samples of both segments are identically distributed.

The key-point of our approach is to define the observation model on the p-values statistics derived from the Wilcoxon test and considered as random variables, rather than on the samples $\boldsymbol{X}$ or the statistic U. A classical result (see [8] for instance) is that the p-value is uniformly distributed on $[0, 1]$ under the null hypothesis H_0, *i.e.* when $x_{j,i}$ is not a change-point. We have to specify now a model for the alternative hypothesis H_1, *i.e.* when $x_{j,i}$ is a change-point. In this case, the p-value can be viewed as an outlier with respect to its null distribution. Several authors proposed some distributions or approximations for some alternatives hypotheses, see for instance [9–12]. To be precise, the alternative hypothesis must support the largest values of the test statistic U, *i.e.* the smallest p-values. This leads to consider families of distributions whose densities are decreasing in p. Inspired by [11], we assume that the p-value is distributed under H_1 according to a Beta $\mathcal{B}(\gamma, 1)$ distribution where the parameter γ is in $[0, 1]$. This is a specific case of a general class of distribution for the choice of the distribution of p-values under the alternative hypothesis, for which a lower bound on Bayes factor is given [11]. It yields the following density:

$$f(p_{j,i}|\boldsymbol{R}) = \begin{cases} \mathbb{1}_{[0,1]}(p_{j,i}) & \text{if } r_{j,i} = 0 \ (H_0), \\ \gamma p_{j,i}^{\gamma-1} \mathbb{1}_{[0,1]}(p_{j,i}) & \text{if } r_{j,i} = 1 \ (H_1). \end{cases} \quad (4)$$

Note that when $\gamma = 1$, the alternative distribution reduces to the uniform distribution, *i.e.* the null distribution. Another example of use of a Beta-Uniform model can be found in [12]. In this work, we propose to calibrate the alternative distribution with respect to a frequentist significance level. Then, the Beta parameter γ is defined after the acceptance level $\alpha \in [0, e^{-1}]$ (for brevity reason, the notation $\gamma(\alpha)$ is omitted) such that γ is the unique solution on $[0, 1]$ of the following equation

$$f(\alpha|r = 1) = 1 \quad \Leftrightarrow \quad \gamma\alpha^{\gamma-1} = 1. \quad (5)$$

This calibration is discussed in more details in section 2.3.

Finally, the Wilcoxon test is applied on each observation $x_{j,i}$, and following (4), the data term for our model is formed by the inference function

$$L_*(\boldsymbol{X}|\boldsymbol{R}) = \prod_{j=1}^{K}\prod_{i=1}^{N} f(p_{j,i}|\boldsymbol{R}) = \prod_{j=1}^{K}\prod_{i=1}^{N} \left(\gamma p_{j,i}^{\gamma-1}\right)^{r_{j,i}}. \quad (6)$$

It is important to note that this inference function is a composite marginal likelihood based on the univariate distributions of the $(p_{j,i})_{1\le j\le K, 1\le i\le N}$. In particular, the dependencies between the $(p_{j,i})_{1\le i\le N}$ in signal j are not taken into account. As a consequence, this is not a proper likelihood as defined in [13] and the coverage probabilities induced by this model should differ from the real ones. However $L_*(\boldsymbol{X}|\boldsymbol{R})$ depends on the data through the marginal distributions of the p-values of the Wilcoxon statistics and is calibrated up to a significance level α. The impact this level on the detection will be discussed in a future work. As a consequence, this inference function acts like a data term, and can be used as a pseudo-likelihood in the Bayesian framework. We refer to this model as the Bernoulli detector model. It differs from classical Bernoulli Gaussian models where the likelihood is directly derived from the parametric Gaussian assumption of the samples.

2.2. Prior on configurations R_i

The vectors $(R_i)_{1\le i\le N}$ are assumed to be *a priori* independent. Thus the prior distribution can be written:

$$f(\boldsymbol{R}) = \prod_{i=1}^{N} f(R_i). \quad (7)$$

Following the approach presented in [5], where vague prior on the dependency structure is chosen, the vector $\boldsymbol{P} = (P_\epsilon)_{\epsilon\in\mathcal{E}}$ represents the dependency structure. Here P_ϵ denotes the probability of having $R_i = \epsilon$, where ϵ is a $K \times 1$ vector

of zeros and ones, called a configuration, and $\mathcal{E}$ is the subset of $\{0,1\}^K$ of all configurations for R_i. This notation means that if the probability P_ϵ is high, ϵ is more likely to appear in $\boldsymbol{R}$, then changes tend to be simultaneous in all signals j such that $\epsilon(j) = 1$. Then (7) becomes

$$f(\boldsymbol{R}|\boldsymbol{P}) = \prod_{\epsilon\in\mathcal{E}} P_\epsilon^{S_\epsilon(\boldsymbol{R})} \tag{8}$$

where $S_\epsilon(\boldsymbol{R})$ is the number of times that ϵ is found in $\boldsymbol{R}$. The parameter $\boldsymbol{P}$ follows a Dirichlet distribution:

$$\boldsymbol{P}|d \sim \mathcal{D}_L(d) \tag{9}$$

with the hyperparameter vector $d = (d_\epsilon)_{\epsilon\in\mathcal{E}}$ and where L is the cardinal of $\mathcal{E}$. As in [5], all the d_ϵ are set to the same deterministic value $d_\epsilon \equiv d = 1$, so distribution (9) is uniform.

2.3. Posterior distribution

The unnormalised posterior distribution (2) of the change-points indicators and the hyperparameters expresses as follows:

$$f(\boldsymbol{R},\boldsymbol{P}|\boldsymbol{X}) \propto \left(\prod_{j=1}^{K}\prod_{i=1}^{N}(\gamma p_{j,i}^{\gamma-1})^{r_{j,i}}\right)\left(\prod_{\epsilon\in\mathcal{E}} P_\epsilon^{S_\epsilon(\boldsymbol{R})+d_\epsilon-1}\right). \tag{10}$$

Based on this posterior, it is now possible to express some properties of the classical Bayesian estimates of a change-point for a given location (j,i) under some simple hypothesis. We assume in the following proposition that there is no other change-point in the signals. We denote as $\boldsymbol{R}_0^{\setminus(j,i)}$ the void configuration event such that $r_{l,k} = 0$ for all $(l,k) \neq (j,i)$, and ϵ_0 and ϵ_1 stand for the configurations of the column R_j when $r_{j,i} = 0$ and $r_{j,i} = 1$ respectively.

Proposition 2.1 *MAP and MMSE estimators given $\boldsymbol{R}_0^{\setminus(j,i)}$, P_{ϵ_0} and P_{ϵ_1}. Under the previous hypothesis, if $P_{\epsilon_0} = P_{\epsilon_1}$ and for a chosen significance level α:*

- *the conditional MAP estimate is 1 iff the p-value is lower than α, according to* (4) *and* (5)*,*
- *the conditional MMSE estimate is larger than $1/2$ iff the p-value is lower than α.*

The proofs are directly derived from the definition of these estimators and the expression of the posterior distribution (10). These properties illustrate the influence of the significance level α chosen in (5) to calibrate the distribution (4) under H_1 for the single change-point problem. If the priors on the configurations are equivalent then the presence of a change-point is favoured when the support against the null hypothesis is significant for the level α.

Finally, based on (10), one can see that the posterior of the hyperparameter $\boldsymbol{P}$ reduces to a Dirichlet distribution

$$\boldsymbol{P}|\boldsymbol{R},d \sim \mathcal{D}_L(S_\epsilon(\boldsymbol{R}) + d_\epsilon), \tag{11}$$

in agreement with the equation (22) of [5]. Thus, the P_ϵ can be easily integrated out in (10) as nuisance parameters, yielding the following marginalized posterior

$$f(\boldsymbol{R}|\boldsymbol{X}) \propto \left(\prod_{j=1}^{K}\prod_{i=1}^{N}(\gamma p_{j,i}^{\gamma-1})^{r_{j,i}}\right) \times \frac{\prod_{\epsilon\in\mathcal{E}}\Gamma(S_\epsilon(\boldsymbol{R})+d_\epsilon)}{\Gamma(N+L)}. \tag{12}$$

2.4. Algorithm

The problem of estimating $\boldsymbol{R}$ is solved by a MCMC method, where samples are drawn according to distribution (12). At each step, all vectors R_i are simulated, following a Gibbs sampling strategy. At each modification of $r_{j,i}$, the segmentation changes and the p-values of the previous and the next change-point in the signal j have to be updated. However an approximation is done the reduce the number of steps in the algorithm, and only the current p-value $p_{j,i}$ is computed for the new segmentation. For brevity reason, the empirical validation is not presented here. The main steps of the algorithm are detailed below in Alg. 1, for M MCMC iterations.

Algorithm 1: Bernoulli detector

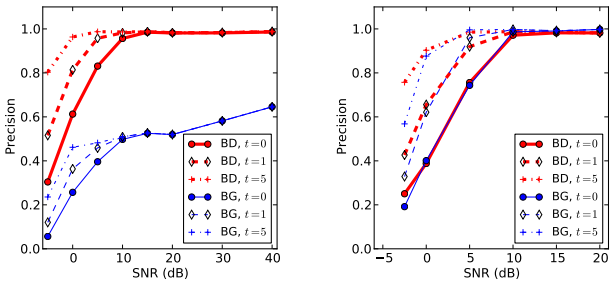
require $\mathcal{E} = \{\epsilon_0, ..., \epsilon_l, ..., \epsilon_L\} \subset \{0,1\}^K$, α
initialize $R^{(0)}$, $S_\epsilon(\boldsymbol{R}^{(0)})$
for $m \leftarrow 1$ **to** M **do**
  initialize the index set $I = \{1, \ldots, N\}$
  while $I \neq \emptyset$ **do**
    pick randomly i in I
    compute $(p_{j,i}^{(m)})_{1\leq j\leq K}$
    sample $R_i^{(m)}$ according to (12)
    remove i from I
  Optional step: sample $\boldsymbol{P}^{(m)}$ from its posterior (11)
return R

3. APPLICATION ON DATA

3.1. Simulation

At first, the impact of the noise level and the presence of outliers is studied. Two signals of $N = 100$ points are generated, with a change-point at $t = 50$, that defines two segments in each signal. In the first case, the observations on each segment k follow a non-standardized Student's t-distribution with a heavy tail, whose parameters are (ν, μ_k, σ) where $\nu = 3.0$, and $\sigma^2 = \nu/(\nu-2)$. In the second case, the distribution is normal $\mathcal{N}(\mu_k, \sigma)$. The SNR is defined as $\mathrm{SNR} = 10\log\frac{(\mu_0-\mu_1)^2}{\sigma^2}$. To validate the non-parametric approach of our model, we compare it with a classical Bernoulli Gaussian model. The performances are computed on the MAP estimator in terms of precision, that is the proportion of true change-points found in the detected ones, with a tolerance of $\pm t$ points in time. The results are shown in figure 1. It appears

in 1(a) that the Bernoulli detector model has much less false positives than the Bernoulli Gaussian, as expected because of robustness of the rank-test to distributions with heavy tail in the non-parametric pseudo-likelihood. The generic nature our model is pointed out by the fact that the precision values are the same in the normal case (1(b)). It is equivalent then to the Bernoulli Gaussian model, and slightly better for small SNR. Due to the restricted length of the paper, the recall values are not plotted, however, we found results in agreement to precision values: the recall values are converging to one with high SNR and the Bernoulli detector model presents better recall performance for a Student noise.



(a) Student's t-distribution (b) Normal distribution

Fig. 1. Precision for several SNR, of Bernoulli detector model (BD, red) and Bernoulli Gaussian model (BG, blue), with tolerance t.

In the second test we consider five signals following Normal distributions, classified in two independent groups: group 1 with signals 1 and 2 (3 segments), and group 2 with signals 3, 4 and 5 (4 segments). The change-points are the same for all signals within a group. The SNR defined previously is 14.0 for signals 1 to 4 and -6.0 for signal 5, where change-points are not visible (figure 2). The data are processed independently (results in blue in figure 2) and jointly with a non-informative prior on dependency structure (in red in figure 2). Despite a higher computational cost, the advantage of the of joint estimation is obvious in the results: the dependency structure of both groups is perfectly learned, and all change-points are precisely found, especially in signal 5, whereas the estimation fails with independent processing.

3.2. Household electric power consumption data

In this section two principled applications of the proposed algorithm are illustrated using measurements of household power consumption [14]. This real dataset consists of four time series. One of them depicts the global electrical energy consumption in the house, while the others are devoted to the measurements of power demand by specific devices.

Learning structure from data. In a first setting, the algorithm is applied with non-informative priors on the dependency structure between the four signals, namely all configurations in $\{0,1\}^K$ are considered. Thanks to the Bayesian framework, the Gibbs sampling process can be swimmingly

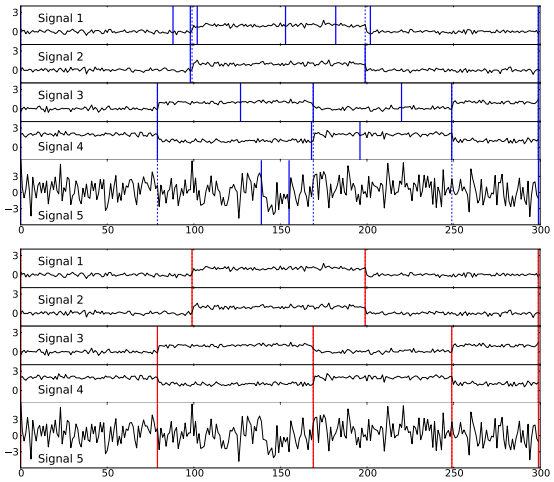


Fig. 2. At top, independent MAP estimation of each signal. At bottom, joint estimation with a non-informative prior on the change-points structure. Real change-points are represented in dashed lines.

supplemented to yield *a posteriori* probabilities on the configurations. Resulting histograms of the links between the global household energy consumption, denoted 1, and the three sub-metering measurements (respectively denoted 2, 3 and 4), plus the link between 3 and 4, are depicted in figure 3. Unsurprisingly these figures confirm that, when a change-point is observed in signals 2, 3 or 4, it is likely to be observed in signal 1, whereas events appear independently between sub-metering signals. Therefore this setting amounts to building a weighted graph of the dependencies between time series by focusing on small-scale events such as change-points. Visual results on a small portion of signals are represented in figure 4(a).

Using a priori data structure to alleviate data processing. In a second setting the introduction of informative priors on the structure of the dependencies is considered. It is indeed known that strong additive relationships exist between signal 1 and signals 2, 3 and 4 respectively. The change-point detection algorithm is thus modified by restricting $\mathcal{E}$ such that change-points in signals 2, 3 or 4 must occur simultaneously in signal 1. Consequences of such an approach are two-fold. First computations are made significantly faster thanks to the sharp decrease of the number of considered configurations. Second spurious or missing change-points can be filtered out with this simple approach. Visual results are represented in figure 4(b).

4. DISCUSSION AND CONCLUSION

An innovative change-point detection algorithm has been proposed. Its strength relies mainly on two characteristics. First the algorithm is built upon robust foundations, namely rank-based statistics. This confers the approach with a stable behaviour for various signal distributions, a worthy robustness

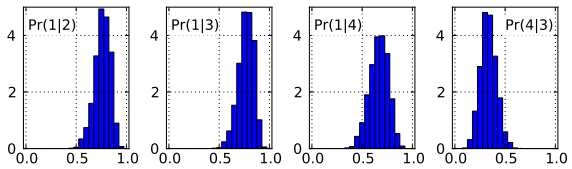
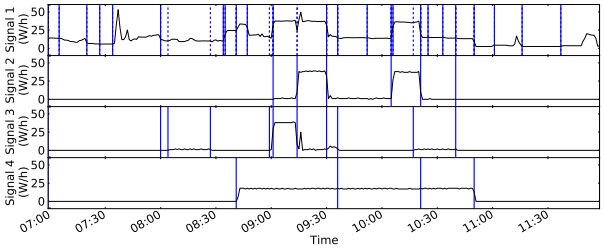
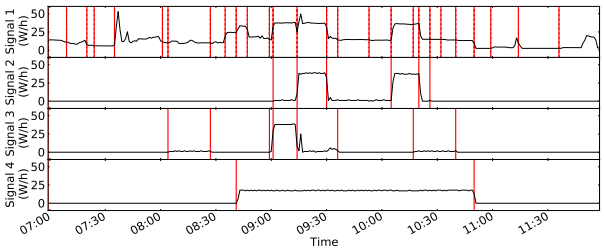


Fig. 3. Posterior conditional probabilities $\Pr(i|j)$ to have a change-point in signal i when there is a change-point in signal j, for non-informative prior.



(a) Non-informative prior



(b) Informative prior

Fig. 4. Change-point detection in real data (MAP). At top (a), change-points found for signals 2, 3 and 4 but not for global signal 1 are represented in dashed lines.

to noise outliers and a strong efficiency in small samples problems. Second its surrounding Bayesian formulation provides the algorithm with awesome flexibility regarding the problems at hand. While classical multivariate approaches focus on detecting simultaneous change-points across time series, the presented framework offers a wide range of possible usages, from joint independent detections across time series, to simultaneous change-points detection through the introduction of priors.

Two main settings have been presented to illustrate the use of the algorithm. On the one hand it has been demonstrated that the approach can be successfully applied to infer dependency structure from change-points occurrences. This constitutes an innovative way of estimating the dependencies between signals, and will be explored in future work. This algorithm therefore provides a tool to analyse multivariate time series behaviour through the observation of small-scale events in the signal. On the other hand, it has been demonstrated that priors on the dependency structure can be favourably introduced to improve change-point detections, and also reduce the expensive computational cost, due to MCMC sampling and configuration testing.

REFERENCES

[1] M. Basseville and I.V. Nikiforov, *Detection of abrupt changes: theory and application*, Prentice-Hall information and system sciences series. Prentice Hall, 1993.

[2] Z. Harchaoui, F. Bach, O. Cappé, and E. Moulines, "Kernel-based methods for hypothesis testing: A unified view," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 87–97, June 2013.

[3] C. Zou, Y. Liu, P. Qin, and Z. Wang, "Empirical likelihood ratio test for the change-point problem," *Statistics & Probability Letters*, vol. 77, no. 4, pp. 374–382, 2007.

[4] G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani, "A fused lasso latent feature model for analyzing multi-sample aCGH data," *Biostatistics*, vol. 12, no. 4, pp. 776–791, 2011.

[5] N. Dobigeon, J.-Y. Tourneret, and M. Davy, "Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a bayesian sampling approach," *Signal Processing, IEEE Transactions on*, vol. 55, no. 4, pp. 1251–1263, 2007.

[6] E. L. Lehmann, "The power of rank tests," *The Annals of Mathematical Statistics*, pp. 23–43, 1953.

[7] A. Lung-Yut-Fong, C. Levy-Leduc, and O. Cappé, "Homogeneity and change-point detection tests for multivariate data using rank statistics," 2011.

[8] H. Sackrowitz and E. Samuel-Cahn, "P values as random variables-expected p values," *The American Statistician*, vol. 53, no. 4, pp. 326–331, 1999.

[9] R. R. Bahadur, "Simultaneous comparison of the optimum and sign tests of a normal mean," in *Contributions to Probability and Statistic*, I. Olkin et al, Ed., pp. 77–88. Stanford University Press, Stanford, 1960.

[10] B. J. Becker, "Small-sample accuracy of approximate distributions of functions of observed probabilities from t tests," *Journal of Educational Statistics*, vol. 16, 1991.

[11] T. Sellke, M. J. Bayarri, and J. O. Berger, "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, vol. 55, no. 1, pp. 62–71, 2001.

[12] D. B. Allison, G. L. Gadbury, M. Heo, J. R. Fernández, C.-K. Lee, T. A. Prolla, and R. Weindruch, "A mixture model approach for the analysis of microarray gene expression data," *Computational Statistics & Data Analysis*, vol. 39, no. 1, pp. 1–20, 2002.

[13] J. F. Monahan and D. D. Boos, "Proper likelihoods for bayesian analysis," *Biometrika*, vol. 79, no. 2, 1992.

[14] K. Bache and M. Lichman, "UCI machine learning repository," 2013.