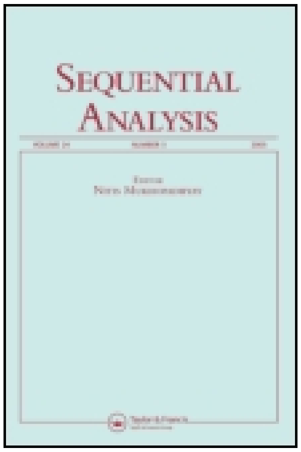
This article was downloaded by: [Michael Baron]
On: 02 February 2013, At: 21:20
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Sequential Analysis: Design Methods and Applications
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsqa20

Discussion on "Change-Points: From Sequential Detection to Biology and Back" by David Siegmund

Michael Baron [a]

[a] Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas, USA

To cite this article: Michael Baron (2013): Discussion on "Change-Points: From Sequential Detection to Biology and Back" by David Siegmund, Sequential Analysis: Design Methods and Applications, 32:1, 15-18

To link to this article: http://dx.doi.org/10.1080/07474946.2013.751837

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Sequential Analysis, 32: 15–18, 2013
Copyright © Taylor & Francis Group, LLC
ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474946.2013.751837

Discussion on "Change-Points: From Sequential Detection to Biology and Back" by David Siegmund

Michael Baron
Department of Mathematical Sciences, University of Texas at Dallas,
Richardson, Texas, USA

Downloaded by [Michael Baron] at 21:20 02 February 2013

Abstract: Motivated by Professor Siegmund's article and his other works in the area of change-point analysis and its biomedical applications, we discuss approaches to change-point–related problems that involve parameters of unknown dimension.

Keywords: Change point; CUSUM; Error rate; Multiple sequences; Unknown dimension.

Subject Classifications: 62L10; 62L15; 62P10.

1. INTRODUCTION

In March 1993, the *Applied Change Point Conference* hosted by my graduate school at the University of Maryland Baltimore County (UMBC) discussed the latest 20th century applications of change-point analysis to diverse fields, from system reliability to baseball (Ahsanullah et al., 1995). It is amazing how quickly this range of applications developed and expanded to include rather complex (21st century) models arising in biology and medicine that are presented in Professor Siegmund's inspiring and thought-provoking paper.

Statistical analysis of complicated stochastic models often meets a challenge when it involves estimation of parameters of *unknown dimensions*. Without a proper penalty for overfitting (such as, e.g., Chen and Gupta (1997) or Harchaoui and Lévy-Leduc (2010)), conventional methods tend to overestimate the number of unknown parameters. The choice of a penalty is typically subjective; however, it has a major effect on the estimation results.

Here we discuss two unknown dimension problems that appear in the context of change-point estimation in genetic mapping. One of them is estimation of multiple

Received July 28, 2012, Revised October 5, 2012, Accepted October 6, 2012
Recommended by A. G. Tartakovsky

Address correspondence to M. Baron, Department of Mathematical Sciences, University of Texas at Dallas, FO-35, 800 W. Campbell Rd., Richardson, TX 75080, USA; Fax: +1 (972) 883-6622; E-mail: mbaron@ utdallas.edu

change-points, for example, in the case of multiple variant intervals, where the number of variant intervals is *a priori* unknown. The other problem is about multiple channels such as DNA sequences, an unknown portion of which experience a change-point.

2. MULTIPLE CHANGE POINTS

In DNA sequencing as well as quality control, psychology, economics, finance, and other areas, the observed process may experience several change-points. In this case, the maximum likelihood routines (Fu and Curnow, 1990) including the Vierbi algorithm (Rabiner, 1989) tend to detect too many false change-points. Indeed, the likelihood may only increase when the parameter space is expanded. On the other hand, the binary segmentation scheme (Vostrikova, 1981) tends to miss some of the change-points because it assumes only one change-point at each step. It is tempting to impose constraints on the dispersion of change-points and to use restricted maximum likelihood estimation methods under deterministic constraints or Bayesian computation (Chib, 1998) under a prior distribution. These constraints turn out to have a strong impact on the estimation result and, typically, the procedure will detect as many change-points as it is allowed.

Thus, the best option seems to be detecting the multiple change-points sequentially, or one at a time, as mentioned by Professor Siegmund. No wrong assumptions have to be taken, each change-point has a chance to be detected, and the nuisance parameters are estimated using the corresponding blocks of data between the estimated change points.

Noticeably, these parameters are likely to be estimated from *contaminated data*. Initially detected change points are not estimated accurately and, thus, the pre- and post-change parameters may be estimated from subsamples that inadvertently overlap with the adjacent inter-change-point segments. For this reason, it is even possible to miss a change-point without ever detecting it (Baron, 2000). A solution is to refine the set of estimated change-points and nuisance parameters iteratively. During this process, similar adjacent segments may be merged, if a change-point separating them is not found to be significant at some step. In a reverse situation, a segment may be divided, generating a new change-point, if such a split is found to create two significantly different distributions. Hopefully, a more accurately estimated set of multiple change-points is computed at each iteration, and less contamination appears in the estimation of nuisance parameters (Baron, 2004). This, however, cannot be guaranteed, as long as the underlying distributions have overlapping supports.

3. MULTIPLE SEQUENCES WITH POSSIBLE CHANGE POINTS

The problem of change-point detection in multiple DNA sequences is certainly important and interesting but it also contains a challenge. Any of N sequences may have a variant interval that marks two change-points, τ_1 and τ_2. Then, in the presence of nuisance parameters, it is not clear which sequences should be modeled with one parameter (no change-point) and which ones with two parameters (one before τ_1 and after τ_2 and another between τ_1 and τ_2). If one allows the maximum likelihood method to choose between these two options, it should always support the latter, simply because the maximum over a larger set is higher.

Downloaded by [Michael Baron] at 21:20 02 February 2013

Different approaches have been proposed for this problem. Tartakovsky et al. (2003) and Tartakovsky and Veeravalli (2004) allowed the occurrence of a change-point in exactly one of N sequences. More generally, Professor Siegmund assumes a known fraction p_0 of sequences with a change-point. In the joint likelihood, p_0 plays the role of a prior probability for each sequence to have a change-point.

The asymptotic behavior of the resulting change-point estimator is interesting. It is generally known that no consistent change-point estimator exists, in the current setting, where the change-point is defined as the index of the first data point from a new distribution (e.g., Hinkley, 1970). On the other hand, when subsamples before and after the change-point are sufficiently large to allow consistent estimation of each nuisance parameter, the number of change-points should be estimated consistently, so no change remains unnoticed. Intuitively, since large data sets should dominate over the prior distribution, the fraction of sequences with detected change-points should converge to the true proportion. Perhaps for this reason the procedure is robust with respect to the choice of p_0, as mentioned by Professor Siegmund. Then, if the limiting proportion of detected change-points differs from p_0, does it contradict the initial choice of p_0?

Seeking an alternative method that is not based on arbitrary assumptions, one might recall that the cumulative sum (CUSUM) procedure itself was derived by Page (1954) as a result of a sequence of likelihood ratio tests. Page's CUSUM algorithm reports a change-point the first time when the no-change null hypothesis is rejected. Furthermore, as seen from Theorem 4.7 of Shiryaev (1978), the Bayes stopping rule for change-point detection under the geometric prior can also be obtained from a sequence of Bayes tests as the moment of the first rejection of the no-change hypothesis. The same can be stated about its limiting case, the Shiryaev-Roberts procedure (Roberts, 1966), that can be obtained from a sequence of Bayes tests under the non informative generalized prior.

Applying Page's idea to multiple sequences, one may test *multiple* hypotheses of no change in each sequence and define the stopping time to be the first moment of (at least one) rejection. For the set of N parallel sequences, there are N hypotheses, $H_0^{(n)} : \mu_{t,n} = \mu_{0,n}$ for all $t \in [1, m]$ vs. $H_A^{(n)} : \exists \tau_1, \tau_2 \in [1, m]$ and $\delta_n \neq 0$ such that $\mu_{t,n} = \mu_{0,n} + \delta_n I_{\tau_1 < t \leq \tau_2}$, for $n = 1, \ldots, N$.

Two directions can be considered here.

A simpler problem arises when one is interested in the time of changes only. Then it suffices to test one *composite hypothesis* $H_0 = \bigcap H_0^{(n)}$; that is, no change in any sequence so far, vs. $H_A = \bigcup H_A^{(n)}$; that is, a change has occurred in the distribution of at least one sequence. In fact, the maximum log-likelihood ratio proposed in Tartakovsky et al. (2003) can be viewed as a test statistic for this H_0 yielding significance level $N\alpha$ when each individual Wald's stopping boundary is based on level α.

As briefly mentioned by Professor Siegmund, one may also be interested in finding *which* sequences experienced change-points. A composite hypothesis H_0 will no longer serve this purpose. Instead, N individual hypotheses $H_0^{(n)}$ are to be tested sequentially until a decision is made on the occurrence of a change in each of them. The needed multiple testing methodology is well developed for non-sequential data, although only a few algorithms are designed for sequential experiments. The methods of Bartroff and Lai (2010) or De and Baron (2012a,b) can be used to test $\{H_0^{(n)}, n = 1, \ldots, N\}$ controlling the family-wise error across N sequences. Application of such sequential multiple testing will detect, with a certain false alarm

Downloaded by [Michael Baron] at 21:20 02 February 2013

rate, a change and identify, with a certain family-wise error rate, the sequences that are affected by this change. In practice, for large N, it may suffice to determine only a portion of sequences that experienced a change.

ACKNOWLEDGMENTS

The author greatly appreciates his research funding by the National Science Foundation (grant DMS 1007775) and the National Security Agency (grant H98230-11-1-0147).

REFERENCES

Ahsanullah, M., Rukhin, A. L., and Sinha, B. (1995). *Applied Change Point Problems in Statistics*, New York: Nova Science Publishers.

Baron, M. (2000). Nonparametric Adaptive Change-Point Estimation and On-Line Detection, *Sequential Analysis* 19: 1–23.

Baron, M. (2004). Sequential Methods for Multistate Processes, in *Applications of Sequential Methodologies*, N. Mukhopadhyay, S. Datta, and S. Chattopadhyay, eds., pp. 55–73, New York: Marcel Dekker.

Bartroff, J. and Lai, T.-L. (2010). Multistage Tests of Multiple Hypotheses, *Communications in Statistics - Theory and Methods* 39: 1597–1607.

Chen, J. and Gupta, A. K. (1997). Testing and Locating Variance Changepoints with Application to Stock Prices, *Journal of American Statistical Association* 92: 739–747.

Chib, S. (1998). Estimation and Comparison of Multiple Change-Point Models, *Journal of Econometrics* 86: 221–241.

De, S. and Baron, M. (2012a). Sequential Bonferroni Methods for Multiple Hypothesis Testing with Strong Control of Familywise Error Rates I and II, *Sequential Analysis* 31: 238–262.

De, S. and Baron, M. (2012b). Step-Up and Step-Down Methods for Testing Multiple Hypotheses in Sequential Experiments, *Journal of Statistical Planning and Inference* 142: 2059–2070.

Fu, Y.-X. and Curnow, R. N. (1990). Maximum Likelihood Estimation of Multiple Change Points, *Biometrika* 77: 563–573.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple Change-Point Estimation with a Total Variation Penalty, *Journal of American Statistical Association* 105: 1480–1493.

Hinkley, D. V. (1970). Inference about the Change-Point in a Sequence of Random Variables, *Biometrika* 57: 1–17.

Page, E. S. (1954). Continuous Inspection Schemes, *Biometrika* 41: 100–115.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *IEEE Proceedings* 77: 257–285.

Roberts, S. W. (1966). A Comparison of Some Control Chart Procedures, *Technometrics* 8: 411–430.

Shiryaev, A. N. (1978). *Optimal Stopping Rules*, New York: Springer-Verlag.

Tartakovsky, A. G., Li, X. R., and Yaralov, G. (2003). Sequential Detection of Targets in Multichannel Systems, *IEEE Transactions on Information Theory* 49: 425–445.

Tartakovsky, A. G. and Veeravalli, V. V. (2004). Change-Point Detection in Multichannel and Distributed Systems with Applications, in *Applications of Sequential Methodologies*, N. Mukhopadhyay, S. Datta, and S. Chattopadhyay, eds., pp. 339–370, New York: Marcel Dekker.

Vostrikova, L. J. (1981). Detecting "Disorder" in Multidimensional Random Processes, *Soviet Mathematics Doklady* 24: 55–59.

Downloaded by [Michael Baron] at 21:20 02 February 2013