Taylor & Francis
Taylor & Francis Group

Sequential Analysis

Design Methods and Applications

ISSN: 0747-4946 (Print) 1532-4176 (Online) Journal homepage: https://www.tandfonline.com/loi/lsqa20

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

Hock Peng Chan & Tze Leung Lai

To cite this article: Hock Peng Chan & Tze Leung Lai (2013) Discussion on "Change-Points: From Sequential Detection to Biology and Back" by David Siegmund, Sequential Analysis, 32:1, 22-27, DOI: 10.1080/07474946.2013.751840

To link to this article: https://doi.org/10.1080/07474946.2013.751840

Sequential Analysis, 32: 22–27, 2013

ISSN: 0747-4946 print/1532-4176 online
DOI: 10.1080/07474946.2013.751840

Taylor & Francis
Taylor & Francis Group

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

Hock Peng Chan[1] and Tze Leung Lai[2]
[1]Department of Statistics and Applied Probability,
National University of Singapore, Singapore
[2]Department of Statistics, Stanford University, Stanford,
California, USA

Abstract: Professor Siegmund has provided a unified treatment of change-point problems in sequential detection and fixed-sample DNA/protein sequence analysis and other biological applications via boundary-crossing probabilities and likelihood-based procedures. Our discussion elaborates further on this theme and moves beyond biology to engineering and finance.

Keywords: Fault detection; Generalized likelihood ratios; Random fields; Scan statistics; Segmentation.

Subject Classifications: 62L15; 60G40; 62F12.

1. INTRODUCTION

We begin by congratulating Professor Siegmund on his very insightful and elegant paper that starts with a brief review of the early history of sequential change-point detection and moves on to "the continuously increasing scope of scientific problems having change-point or change-point-like characteristics," thereby pointing to directions "for new research, and common basic principles [that] provide the foundation for a continuously expanding theoretical framework," which has been motivated by "the diversity of applications." Our discussion will elaborate on this continuously expanding theoretical framework for change-point methodology, motivated by the application domains in which we have worked.

Received September 9, 2012, Revised October 7, 2012, Accepted October 9, 2012
Recommended by A. G. Tartakovsky
Address correspondence to H. P. Chan, Department of Statistics and Applied Probability, 6 Science Drive 2, National University of Singapore, Singapore 117546; Fax: +65 6872-3919; E-mail: stachp@nus.edu.sg

2. FAULT DETECTION IN STOCHASTIC SYSTEMS

While Page's (1954) cumulative sum (CUSUM) rule for detecting a change from a baseline density f_{θ_0} to a post-change density f_{θ_1} has a recursive formula for updating the likelihood ratio detection statistics, there are no recursive formulas for updating generalized likelihood ratio (GLR) detection statistics that do not assume the post-change parameter to be known. For fault detection in engineering systems involving multidimensional parameters that may undergo changes when faults occur, "practical implementation of the GLR algorithm is not always possible" (Basseville and Nikiforov, 1993, p. 369) because of real-time computational constraints for on-line fault detection. To reduce the computational burden of the GLR scheme, which grows to infinity with n and involves maximization of the log-likelihood over $\theta \in \Theta$ for each possible change time k between 1 and n, Willsky and Jones (1976) proposed to use a window-limited GLR scheme with a stopping rule of the form

$$N_W = \inf\left\{n > \widetilde{N} : \max_{\widetilde{M} \le n-k+1 \le M} \sup_{\theta \in \Theta}\left[\sum_{i=k}^{n} \log\left(\frac{f_\theta(X_i \mid X_1, \ldots, X_{i-1})}{f_0(X_i \mid X_1, \ldots, X_{i-1})}\right)\right] \ge c_\gamma\right\}, \quad (2.1)$$

but did not indicate how M, $\widetilde{M}$, and c_γ should be chosen. The performance of (2.1) was subsequently found to be quite sensitive to these choices. As noted in Lai (1995, Section 3.2), the Siegmund-Venkatraman (1995) paper referenced by Professor Siegmund provided an important clue to address this issue.

Siegmund and Venkatraman gave a definitive asymptotic analysis of Barnard's GLR scheme for the case in which X_i are independent $N(\theta, 1)$ random variables. In this normal case, the GLR statistics in (2.1) have a simple explicit form given in Section 1.2 of Professor Siegmund's paper. In addition, normal random walks yield explicit asymptotic formulas for the associated boundary-crossing probabilities and expected stopping times (average run length, ARL). Making use of these formulas developed by Siegmund and Venkatraman, Lai (1995) showed that in the normal case for which $\widetilde{M}$ can be chosen to be 1, if $M \sim a \log \gamma$ is chosen to achieve an ARL of γ under the baseline model f_0, then the window-limited scheme is asymptotically efficient for the detection of changes that are larger than $\sqrt{2/a}$, with an asymptotic detection delay of $(2 \log \gamma)/\theta^2$. However, for a change magnitude smaller than $\sqrt{2/a}$, the asymptotic detection delay is of order γ and the scheme is inefficient. This is resolved by the consideration of geometrically increasing window sizes $\mathcal{N} = \{[b^j M] : j = 1, 2, \ldots\}$, where $b > 1$ and $[\cdot]$ is the greatest integer function. Then, for a stopping rule of the form $N = \min(N_W, \widetilde{N}_W)$, where

$$\widetilde{N}_W = \inf\left\{n : \max_{k:n-k+1 \in \mathcal{N}} [(X_k + \cdots + X_n)^2/2(n-k+1)] \ge c_\gamma\right\},$$

a detection delay of not more than $2b \log \gamma/\theta^2$ is achieved uniformly for $|\theta| \le \sqrt{2/a}$. Lai (1998) extended these window-limited GLR detection rules to general stochastic systems and proved their asymptotic optimality under not only Lordon's criterion or the Bayesian criterion pioneered by Shiryaev (see Section 2 of Professor Siegmund's paper) but also other criteria that are more suitable for dependent time series/control systems data. To determine the threshold c_γ in complex stochastic systems, one has to use Monte Carlo simulations as tractable analytical

approximations are not available. As noted by Lai (1995, Section 3.3) and Chan and Zhang (2007), although the baseline ARL is too large to simulate, one can relate it to the probability of false alarm for these window-limited GLR detection rules via

$$E_0(T_c) \approx m/P_0(T_c \le m) \tag{2.2}$$

for large m and simulate $P_0(T_c \le m)$ by importance sampling instead.

3. FROM SEQUENTIAL TO FIXED-SAMPLE CHANGE-POINT DETECTION

The close connection between sequential and fixed-sample change-point problems is pointed out in Section 1.2 of Professor Siegmund's paper, in which Section 3 contains a number of fixed-sample problems that are related to the sequential problems in Section 2. Because of the "non-regular" nature of GLR scan statistics in fixed-sample problems, the usual χ^2-approximations for GLR statistics to test under the null hypothesis of no change-points in the sample are invalid. Chan and Lai (2003) have developed a general theory that connects the asymptotic distributions of sequential and fixed-sample change-point problems in the context of nonlinear functions of Markov random walks, motivated by the sequential detection applications in the preceding section. Consider a d-dimensional Markov random walk S_n and let $g : \mathbb{R}^d \to \mathbb{R}$ be a smooth function. Let

$$M_n = \max_{1 \le i < j \le n,\, j-i \in J_n} (j-i) g((S_j - S_i)/(j-i)),$$
$$T_c = \inf\{n : \max_{k < n : n-k \in J(c)} (n-k) g((S_n - S_k)/(n-k)) > c\},$$

where J_n and $J(c)$ are certain subsets of $\{1, 2, \dots\}$. The special case of independent and identically distributed univariate observations, $g(x) = x$ and $J_n = J(c) = \{1, 2, \dots\}$ corresponds to the CUSUM and that with $g(x) = x^2/2$ to Barnard's test. Making use of saddlepoint approximations for Markov random walks, Chan and Lai (2003) showed that there exist $q \in \{0, 1, \dots, d\}$ and positive constants r and ζ depending on g such that under the null hypothesis of no change-points,

$$\zeta (c/r)^{q/2} e^{-c/r} T_c \Rightarrow \mathrm{Exp}(1), \tag{3.1}$$

whereas M_n has a corresponding Gumbel-type limiting distribution given by

$$P\{M_n - r[\log n + (q/2) \log\log n] \le t\} \to \exp(-\zeta e^{-t/r}). \tag{3.2}$$

Using the connection $P(T_c \le n) = P(M_n > c)$ between T_c and M_n, they first proved (3.2) and then used the result to prove (3.1). Section 3.2 of Professor Siegmund's paper discusses this connection in the special case of $g(x) = x$ and relates Page's CUSUM rule for T_c in this case to the Karlin-Dembo-Kawabata statistic M_n for testing high-scoring segments in a protein sequence.

4. RANDOM FIELDS AND BOUNDARY-CROSSING PROBABILITIES

Section 3.1 of Professor Siegmund's paper discusses applications of the Hotelling-Weyl formula, for the volume of a tube around a smooth closed curve or more general manifold to fixed-sample change-point testing problems in signal detection. Integration over tubular neighborhoods of extremal manifolds also arises in extending Laplace's method for asymptotic evaluation of integrals in the derivation of (3.2) and in boundary-crossing probabilities for GLR test and detection statistics and for asymptotically Gaussian random fields; see Chan and Lai (2000, 2003) and Chan and Lai (2006), respectively.

Professor Siegmund's overshoot correction for the scan statistic in linkage analysis mentioned in his Section 3.3 provides a powerful method for boundary-crossing probability computations for a variety of biological applications. Examples include the scoring of word counts in biomolecular sequences (Chan and Zhang, 2007) and of template patterns in neural spike trains (Chan and Loh, 2007); see also Chan et al. (2008).

5. ALTERNATIVE TO THE SCAN STATISTIC: SHIRYAEV OR PAGE?

Page's CUSUM and Barnard's GLR detection rules take the maximum, over all candidate change times k, of the likelihood ratio or generalized likelihood ratio statistics. Shiryaev's detection rule involves the sum of the likelihood ratios over all candidate change times k. The test statistic in Professor Siegmund's change-point model for aligned copy number variation involves both sums and maxima, with the sums in his equation (3.4) to accumulate scores within the cohort of size N and over candidate change-points in the interval $(s, t]$, and the maximum taken over the unknown parameters δ_i to yield the statistic $Z(s, t)$ in his equation (3.5), and also over s and t in the test statistic $\max_{t\le m, m_0\le t-s\le m_1} Z(s, t)$ in his equation (3.7). Note that the scores $U_i(s, t)$ for subject i are not directly summed over $1 \le i \le N$ in his equation (3.5) but are first transformed by

$$f(U) = \log[1 - p_0 + p_0 \exp(U^2/2)]$$

in order to downweight "weak" scores, as these scores are likely to be noisy when the true proportion of signals is small and they can overwhelm the actual signals if the transformation is not applied. The choice of p_0 is left to the user, and Professor Siegmund has shown via extensive simulation studies in his 2010 and 2011 papers with Yakir, Zhang, Ji and Li the robustness of their procedure with respect to p_0 that ranges between 0.01 and 0.25 in these studies. As $p_0 \to 1$, the right-hand side of his equation (3.5) converges to a sum of untransformed GLR statistics. On the other hand, as $p_0 \to 0$, since

$$f(U) = p_0[\exp(U^2/2) - 1 + o(1)],$$

the right-hand side of his equation (3.5) is asymptotically equivalent to $p_0 N$ multiplied by the average likelihood ratio $N^{-1} \sum_{i=1}^{N} \exp[U_i^2(s, t)/2]$.

The average likelihood ratio statistic has detection properties that are closely related to the scan statistic. Chan (2009) analyzed the average likelihood ratio

statistic and found it to have slightly more detection power than the scan statistic, which is not surprising given that it was developed to maximize detection power. An added bonus is that the tail probability of the average likelihood ratio statistic is robust against the correlations among individual scores that may not be known precisely. Chan and Walther (2013) have recently shown that a properly weighted average likelihood ratio achieves asymptotic optimality in the context of multiple-scale signal detection, very much like scan statistics that are appropriately modified to deal with the multiple scales. Optimality is in the sense of asymptotic power 1 at the smallest possible detectable signal. Might there be a similar result for Professor Siegmund's setting; that is, is there a smallest value of p_0 for which his test has asymptotic power 1? This minimum value would depend on the total length m of the sequence and the number N of profiles, which are given, and signal strengths δ_i that can be queried from genetics experts.

6. MULTIPLE CHANGE-POINTS AND SEQUENTIAL SURVEILLANCE

In his Sections 3.4 and 3.5, Professor Siegmund has used the frequentist multiple change-point model of Olshen et al. and their circular binary segmentation (CBS) algorithm. Lai et al. (2008) introduced a simple stochastic segmentation model and applied the associated empirical Bayes procedure as an alternative to the CBS algorithm. Lai and Xing (2011) further refined this approach in the setting of multiparameter families and developed explicit recursive formulas for the empirical Bayes estimates of the piecewise constant parameters. They also showed how the empirical Bayes approach, with its computationally attractive recursive estimators, can be used to address the frequentist problem of segmentation and demonstrated its advantages over the CBS algorithm in terms of computational speed and segmentation accuracy. Chen et al. (2011) and Xing et al. (2012) recently applied this empirical Bayes approach to DNA and protein sequence analysis and obtained very promising results.

Lai and Xing (2013a) used the empirical Bayes approach to develop a stochastic change-point ARX-GARCH model (autoregressive model with exogenous inputs and generalized autoregressive conditional heteroskedastic errors) that provides substantial improvements of Lai and Xing's structural change model introduced as an alternative to long memory in financial time series and referenced in Professor Siegmund's paper. The empirical Bayes approach can also be readily extended to sequential surveillance that involves multiple change-points in multiple time series, as shown in the forthcoming book by Lai and Xing (2013b) on active risk management in response to the Dodd-Frank Act and the new Basel III regulations for bank supervision. While there are some similarities to sequential detection of changes in multiple sequences in Section 4 of Professor Siegmund's paper, empirical Bayes modeling is heavily used in these financial time series data.

ACKNOWLEDGMENTS

The authors acknowledge support from the National University of Singapore (grant R-155-000-120-112), and the National Science Foundation (grant DMS-1106535).

REFERENCES

Basseville, M. and Nikiforov, I. (1993). *Detection of Abrupt Changes—Theory and Application*, Englewood Cliffs: Prentice-Hall.

Chan, H. P. (2009). Detection of Spatial Clustering with Average Likelihood Ratio Test Statistics, *Annals of Statistics* 37: 3985–4010.

Chan, H. P. and Lai, T. L. (2000). Asymptotic Approximations for Error Probabilities of Sequential or Fixed Sample Size Tests in Exponential Families, *Annals of Statistics* 8: 1638–1669.

Chan, H. P. and Lai, T. L. (2003). Saddlepoint Approximations and Nonlinear Boundary Crossing Probabilities of Markov Random Walks, *Annals of Applied Probability* 13: 395–429.

Chan, H. P. and Lai, T. L. (2006). Maxima of Asymptotically Gaussian Random Fields and Moderate Deviations Approximations to Boundary Crossing Probabilities of Sums of Random Variables with Multidimensional Indices, *Annals of Probability* 34: 80–121.

Chan, H. P. and Loh, W. L. (2007). Some Theoretical Results on Neural Spike Train Probability Models, *Annals of Statistics* 35: 2691–2733.

Chan, H. P., Tu, I., and Zhang, N. R. (2008). Boundary Crossing Probability Computations in the Analysis of Scan Statistics, in *Scan Statistics–Theory and Applications*, J. Glaz, V. Pozdnyakov, and S. Wallenstein, eds., pp. 87–108, New York: Birkhauser.

Chan, H. P. and Walther, G. (2013). Detection with the Scan and the Average Likelihood Ratio, *Statistica Sinica* 23: 409–428.

Chan, H. P. and Zhang, N. R. (2007). Scan Statistics with Weighted Observations, *Journal of American Statistical Association* 102: 595–602.

Chen, H., Xing, H., and Zhang, N. R. (2011). Estimation of Parent Specific DNA Copy Numbers in Tumors Using High-Density Genotyping Arrays, *PLoS Computational Biology* 7: 1–15.

Lai, T. L. (1995). Sequential Changepoint Detection in Quality Control and Dynamical Systems (with Discussions), *Journal of Royal Statistal Society, Series B* 57: 613–658.

Lai, T. L. (1998). Information Bounds and Quick Detection of Parameter Changes in Stochastic Systems, *IEEE Transactions on Information Theory* 44: 2917–2929.

Lai, T. L. and Xing, H. (2011). A Simple Bayesian Approach to Multiple Change-Points, *Statistica Sinica* 21: 539–569.

Lai, T. L. and Xing, H. (2013a). Stochastic Change-Point ARX-GARCH Models and Their Applications to Econometric Time Series, *Statistica Sinica* in press.

Lai, T. L. and Xing, H. (2013b). *Active Risk Management: Financial Models and Statistical Methods*, New York: Chapman & Hall/CRC.

Lai, T. L., Xing, H., and Zhang, N. R. (2008). Stochastic Segmentation Models for Array-Based Comparative Genomic Hybridization Data Analysis, *Biostatistics* 9: 290–307.

Siegmund, D. O., Yakir, B., and Zhang, N. R. (2011). Detecting Simultaneous Variant Intervals in Aligned Sequences, *Annals of Applied Statistics* 5: 645–668.

Willsky, A. S. and Jones, H. L. (1976). A Generalized Likelihood Ratio Approach to Detection and Estimation of Jumps in Linear Systems, *IEEE Transactions in Automatic Control* 21: 108–112.

Xing, H., Mo, Y., Liao, W., and Zhang, M. (2012). Genomewide Localization of Protein-DNA Binding and Histone Modification by BCP with ChIP-Seq Data, *PLoS Computational Biology* 8: e1002613.

Zhang, N. R., Siegmund, D. O., Ji, H., and Li, J. Z. (2010). Detecting Simultaneous Change-Points in Multiple Sequences, *Biometrika* 97: 631–646.