Online Change Detection in Exponential Families with Unknown Parameters

Arnaud Dessein[*] and Arshia Cont

MuTant Project-Team (INRIA), Music Representations Group
UMR 9912 STMS (IRCAM, CNRS, UPMC)
1 place Stravinsky, 75004 Paris, France
{dessein,cont}@ircam.fr

Abstract. This paper studies online change detection in exponential families when both the parameters before and after change are unknown. We follow a standard statistical approach to sequential change detection with generalized likelihood ratio test statistics. We interpret these statistics within the framework of information geometry, hence providing a unified view of change detection for many common statistical models and corresponding distance functions. Using results from convex duality, we also derive an efficient scheme to compute the exact statistics sequentially, which allows their use in online settings where they are usually approximated for the sake of tractability. This is applied to real-world datasets of various natures, including onset detection in audio signals.

Keywords: Change detection, exponential families, generalized likelihood ratio, information geometry, onset detection, segmentation.

1 Introduction

Let us consider a time series $x_1, x_2, \ldots$ of observations that are sampled according to an unknown discrete-time stochastic process. In general terms, the problem of *change detection* is to decide whether there are changes in the distribution of the process or not. This decision is often coupled with the estimation of the times when such changes occur. These time instants are called *change points* and delimit contiguous temporal regions called *segments*.

Historically, change detection has been addressed from a statistical perspective [12, 19–21, 24, 26]. We refer to the seminal book [4] for a thorough review, and to [22,23] for up-to-date accounts. Modern approaches have also intersected machine learning, notably kernel methods [6, 8, 14], optimization techniques [15,16,29], and have provided enhanced statistical frameworks [1,11,13,28].

In online approaches, the procedure generally starts with an empty window $\vec{x} \leftarrow ()$, and processes the data incrementally. Then, for each time increment $n = 1, 2, \ldots,$ we concatenate the incoming observation x_n with the previous ones as $\vec{x} \leftarrow \vec{x} \,|\, x_n$, and attempt to detect a change. If a change is detected, then

[*] The first author was supported by a doctoral fellowship from UPMC (EDITE).

we discard the observations before the estimated change point i, and restart the procedure with an initial window $\vec{x} \leftarrow (x_{i+1}, \ldots, x_n)$. In the framework of *abrupt change detection*, it is thus usual to reduce the problem to that of finding one change point in a given window $\overline{x} = (x_1, \ldots, x_n)$.

We follow a standard approach to sequential change detection seen as a problem of multiple hypothesis testing with dominated parametric statistical models, mutually independent random variables, and test statistics based on likelihood ratios. In this context, many approaches assume known parameters before change [4, 22, 23]. This is suitable for applications such as quality control where a normal regime is known, but this is limited in many real-world applications. However, considering unknown parameters before change breaks down the computational efficiency of standard cumulative sum algorithms. Therefore, some simplifications of the exact statistics are generally made to accommodate these situations, such as learning the distribution before change on the whole window, or in a dead region at the beginning of the window where change detection is turned off, leading to approximate generalized likelihood ratio schemes.

A few specific exact generalized likelihood ratio statistics have yet been studied, notably under normality assumptions [27]. Nonetheless, normal distributions do not always model reliably the signals considered. A more general Bayesian framework for independent observations in exponential families has been proposed recently [17]. This Bayesian framework, however, relies on a geometric prior on the time between change points, which is not always well-suited for arbitrary signals. Moreover, it requires prior knowledge on the distributions of the parameters in the respective segments, which is not always available. To overcome this, we seek to formulate a generic sequential change detection with unknown parameters before and after change, but without any a priori information on the respective distributions of the change points and parameters. Our contributions in this context can be summed up as follows.

We study the generalized likelihood ratio test statistics in the light of dually flat information geometry for exponential families. We restrict the study to full minimal steep standard families. While standardness and minimality are actually unrestrictive, fullness and steepness are crucial to the existence and simplicity of maximum likelihood estimates. In this framework, we show that the generalized likelihood ratios find both statistical and geometrical grounds. It therefore provides a unifying view of change detection for many common statistical models and corresponding distance functions.

Using results from convex duality, we also derive a computationally efficient scheme for computing the exact statistics sequentially. This scheme thus addresses the shortcomings inherent to the traditional approaches based on cumulative sum statistics and on approximation heuristics for estimating the unknown parameters before change. Due to its generic nature, the proposed paradigm applies to many common statistical models. It is showcased on real-world datasets of various natures, including an evaluation for onset detection in audio signals.

For complementary information on the work presented here and further applications in audio segmentation, we refer the interested reader to [9, 10].

2 Change detection framework

2.1 Multiple hypothesis statistical decision

Let $\mathcal{S} = \{P_\xi\}_{\xi \in \Xi}$ be a dominated parametric statistical model on a measurable space $(\mathcal{X}, \mathcal{A})$, and let $X_1, \ldots, X_n$ be $n > 1$ mutually independent random variables that are distributed according to probability distributions from $\mathcal{S}$. The problem of *change detection* is to decide, on the basis of sample observations $\bar{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, whether the random variables $X_1, \ldots, X_n$ are identically distributed or not. As discussed previously, we suppose that there is at most one change point, so that the problem reduces to a statistical decision between multiple hypotheses: the null *hypothesis of no change* and the alternative *hypothesis of a change at time* i, respectively defined as

$$H_0 : \ X_1, \ldots, X_n \sim P_{\xi_0} \, ; \tag{1}$$

$$H_1^i : \ X_1, \ldots, X_i \ \sim P_{\xi_0^i} \, , \quad \text{and} \quad X_{i+1}, \ldots, X_n \sim P_{\xi_1^i} \ . \tag{2}$$

To assess the plausibility of the alternative hypotheses compared to the null hypothesis, some test statistics are needed. A standard decision rule is then applied as follows. If at least one of the statistics is above a threshold $\lambda > 0$, then we reject the null hypothesis in favor of the corresponding alternative and detect a change. Otherwise, we fail to reject the null hypothesis and no change is detected. In the case where a change is detected, the change point is estimated as the first time point where the maximum of the statistics is reached.

2.2 Generalized likelihood ratio test statistic

When both the parameters before and after change are unknown, the hypotheses are composite and we cannot use simple likelihood ratios. A common approach is to replace the unknown parameters ξ_0, ξ_0^i, ξ_1^i with their m.l. estimators $\widehat{\xi}_0, \widehat{\xi}_0^i, \widehat{\xi}_1^i : \mathcal{X}^n \to \Xi$, and define a *generalized likelihood ratio at time* i

$$\widehat{\Lambda}^i(\bar{x}) = -2 \log \frac{\prod_{j=1}^n p_{\widehat{\xi}_0(\bar{x})}(x_j)}{\prod_{j=1}^i p_{\widehat{\xi}_0^i(\bar{x})}(x_j) \prod_{j=i+1}^n p_{\widehat{\xi}_1^i(\bar{x})}(x_j)} \ . \tag{3}$$

Some approximations of the generalized likelihood ratios have been proposed to keep the simplicity and tractability of the likelihood ratios in cumulative sum schemes [4]. Most of the time, the parameters before change are assumed to be known, and are in practice estimated either on the whole window, or in a dead region at the beginning of the window where change detection is turned off. Such approximations work when the time intervals between successive changes are important so that the approximation is valid, but fail because of estimation errors as soon as changes occur too often. We argue after that we can still employ computationally efficient decision schemes based on exact generalized likelihood ratios, for the large class of exponential families, without such approximations.

3 Application to exponential families

3.1 Information geometry of exponential families

A *standard exponential family* is a parametric statistical model $\{P_\theta\}_{\theta \in \Theta \subseteq \mathbb{R}^m}$ on the Borel subsets of $\mathbb{R}^m$, which is dominated by a σ-finite measure μ, and whose respective probability densities p_θ with respect to μ can be written as $p_\theta(x) = \exp(\theta^\top x - \psi(\theta))$, where $\psi \colon \Theta \to \mathbb{R}$ is the *log-normalizer*, θ the *natural parameter*, x the *sufficient observation*. The family is called *full* if $\Theta = \mathcal{N}$, where $\mathcal{N} = \{\theta \in \mathbb{R}^m \colon \int_{\mathbb{R}^m} \exp(\theta^\top x)\, \mu(\mathrm{d}x) < +\infty\}$ is the *natural parameter space*. The family is *minimal* if $\dim \mathcal{N} = \dim \mathcal{K} = k$, where $\mathcal{K}$ is the convex support of μ.

More general exponential families can be defined as models that reduce to standard families under sufficiency, reparametrization, and proper choice of a dominating measure. Examples include Bernoulli, Dirichlet, Gaussian, Laplace, Poisson, Rayleigh, exponential, beta, gamma, categorical, multinomial models. Since the reduced standard family can be chosen minimal [3,5], we consider minimal standard families without restriction. These families possess useful properties. First, ψ is a strictly convex function with convex effective domain $\operatorname{dom} \psi = \mathcal{N}$. Moreover, ψ is smooth on $\operatorname{int} \mathcal{N}$, where its gradient is one-to-one, so that we can reparametrize the family with the *expectation parameter* $\eta(\theta) = \nabla\psi(\theta)$.

It is also convenient to require that $\lim_{n \to +\infty} \nabla\psi(\theta_n) = +\infty$ for any sequence of points $\theta_1, \theta_2, \ldots \in \operatorname{int} \mathcal{N}$ that converges to a boundary point of $\mathcal{N}$. This ensures that ψ is essentially smooth, and the family is called *steep*. Considering the framework of convex duality [25], ψ is then of Legendre type with Legendre-Fenchel conjugate φ. The conjugate φ is also of Legendre type and we have $\nabla\varphi = (\nabla\psi)^{-1}$. We further have $\nabla\psi(\operatorname{int} \mathcal{N}) = \operatorname{int} \mathcal{K}$, so that $\nabla\psi$ actually defines a homeomorphism of $\operatorname{int} \mathcal{N}$ and $\operatorname{int} \mathcal{K}$. In this context, there is existence and uniqueness of the m.l. estimator $\widehat{\theta}$ for the full model based on i.i.d. samples $\bar{x}$, which is given as the average $\frac{1}{n} \sum_{j=1}^{n} x_j$ of the sufficient observations in expectation parameters, as soon as that average lies in $\operatorname{int} \mathcal{K}$.

These notions are interpretable within the framework of information geometry [2]. In particular, a minimal standard exponential family $\mathcal{S} = \{P_\theta\}_{\theta \in \operatorname{int} \mathcal{N}}$ endowed with the well-known Fisher information metric g, is a Riemannian manifold and can be enhanced with a family of dual affine α-connections $\nabla^{(\alpha)}$. The statistical manifold $(\mathcal{S}, g)$ is a Hessian manifold since the metric g is induced by the Hessian of ψ. In addition, $(\mathcal{S}, g, \nabla^{(1)}, \nabla^{(-1)})$ is a dually flat space in which θ and η form dual affine coordinate systems. This dually flat geometry generalizes the standard self-dual Euclidean geometry, with two dual Bregman divergences B_ψ and B_φ instead of the self-dual Euclidean distance, where the *Bregman divergence* $B_\phi \colon \Xi \times \Xi \to \mathbb{R}$, generated by a smooth strictly convex function $\phi \colon \Xi \to \mathbb{R}$ on a convex open set Ξ, is defined as $B_\phi(\xi \| \xi') = \phi(\xi) - \phi(\xi') - (\xi - \xi')^\top \nabla\phi(\xi')$. Finally, these two dual divergences between parameters are linked on $\operatorname{int} \mathcal{N}$ with the Kullback-Leibler divergence between the corresponding distributions, through the relation $K(P_\theta \| P_{\theta'}) = B_\psi(\theta' \| \theta) = B_\varphi(\eta(\theta) \| \eta(\theta'))$.

3.2 Results on generalized likelihood ratios

We now derive results on exact generalized likelihood ratio statistics for full minimal steep standard exponential families.

Theorem 1. *The generalized likelihood ratio $\widehat{\varLambda}^i$ at time i verifies*

$$\frac{1}{2}\,\widehat{\varLambda}^i(\bar{x}) = i\,K\Big(P_{\widehat{\theta}_0^i(\bar{x})}\Big\|P_{\widehat{\theta}_0(\bar{x})}\Big) + (n-i)\,K\Big(P_{\widehat{\theta}_1^i(\bar{x})}\Big\|P_{\widehat{\theta}_0(\bar{x})}\Big) \ , \tag{4}$$

as soon as $\bar{x} \in \mathcal{K}_0^i \cap \mathcal{K}_1^i$, where $\mathcal{K}_0^i = \{\bar{x} \in (\mathbb{R}^m)^n \colon \frac{1}{i}\sum_{j=1}^{i} x_j \in \operatorname{int}\mathcal{K}\}$, and $\mathcal{K}_1^i = \{\bar{x} \in (\mathbb{R}^m)^n \colon \frac{1}{n-i}\sum_{j=i+1}^{n} x_j \in \operatorname{int}\mathcal{K}\}$.

Proof. Assuming the samples lie in $\mathcal{K}_0^i \cap \mathcal{K}_1^i$, the m.l. estimates over the full family do exist, belong to $\operatorname{int}\mathcal{N}$, and are given in expectation parameters by the average of the sufficient observations. The generalized likelihood ratios then read

$$\begin{aligned}\frac{1}{2}\,\widehat{\varLambda}^i(\bar{x}) = &\sum_{j=1}^{i}\Big\{\big(\widehat{\theta}_0^i(\bar{x}) - \widehat{\theta}_0(\bar{x})\big)^\top x_j - \psi(\widehat{\theta}_0^i(\bar{x})) + \psi(\widehat{\theta}_0(\bar{x}))\Big\}\\ &+ \sum_{j=i+1}^{n}\Big\{\big(\widehat{\theta}_1^i(\bar{x}) - \widehat{\theta}_0(\bar{x})\big)^\top x_j - \psi(\widehat{\theta}_1^i(\bar{x})) + \psi(\widehat{\theta}_0(\bar{x}))\Big\} \ .\end{aligned} \tag{5}$$

We add and subtract the m.l. estimates $\widehat{\theta}_0^i(\bar{x}), \widehat{\theta}_1^i(\bar{x})$, and their log-normalizers $\psi(\widehat{\theta}_0^i(\bar{x})), \psi(\widehat{\theta}_1^i(\bar{x}))$, to make Bregman divergences B_ψ appear as

$$\frac{1}{2}\,\widehat{\varLambda}^i(\bar{x}) = i\,B_\psi(\widehat{\theta}_0(\bar{x})\|\widehat{\theta}_0^i(\bar{x})) + (n-i)\,B_\psi(\widehat{\theta}_0(\bar{x})\|\widehat{\theta}_1^i(\bar{x})) \ . \tag{6}$$

The result follows by rewriting the Bregman divergences on the natural parameters as Kullback-Leibler divergences on the swapped corresponding distributions.

The statistics can be interpreted as computing the divergence between the m.l. estimates over the full family before/after change and the m.l. estimator with no change, and weighting by the number of samples before/after change. Using convex duality, we also find an alternative expression for the statistics.

Corollary 1. *The generalized likelihood ratio $\widehat{\varLambda}^i$ at time i verifies*

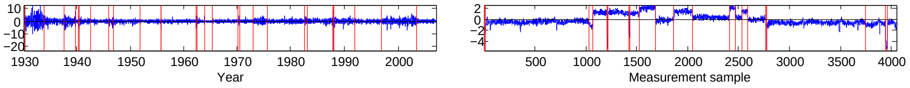
$$\frac{1}{2}\,\widehat{\varLambda}^i(\bar{x}) = i\,\varphi(\widehat{\eta}_0^i(\bar{x})) + (n-i)\,\varphi(\widehat{\eta}_1^i(\bar{x})) - n\,\varphi(\widehat{\eta}_0(\bar{x})) \ . \tag{7}$$

Proof. Rewriting the statistics with Bregman divergences B_φ leads to

$$\frac{1}{2}\,\widehat{\varLambda}^i(\bar{x}) = i\,B_\varphi(\widehat{\eta}_0^i(\bar{x})\|\widehat{\eta}_0(\bar{x})) + (n-i)\,B_\varphi(\widehat{\eta}_1^i(\bar{x})\|\widehat{\eta}_0(\bar{x})) \ . \tag{8}$$

Developing the Bregman divergences and regrouping the terms, we obtain

$$\begin{aligned}\frac{1}{2}\,\widehat{\varLambda}^i(\bar{x}) = &\ i\,\varphi(\widehat{\eta}_0^i(\bar{x})) + (n-i)\,\varphi(\widehat{\eta}_1^i(\bar{x})) - n\,\varphi(\widehat{\eta}_0(\bar{x}))\\ &- \big(i\,\widehat{\eta}_0^i(\bar{x}) + (n-i)\,\widehat{\eta}_1^i(\bar{x}) - n\,\widehat{\eta}_0(\bar{x})\big)^\top \nabla\varphi(\widehat{\eta}_0(\bar{x})) \ .\end{aligned} \tag{9}$$



(a) Daily log-return of the Dow Jones.

(b) Well-log data.

Fig. 1. Change detection in real-world datasets.

The last term vanishes since the m.l. estimate for all samples in H_0, is the barycenter of the m.l. estimates for the samples before and after change in H_1^i.

Since the m.l. estimates between successive windows are related by simple time shifts or barycentric updates in expectation parameters, the above result provides an efficient scheme for calculating the exact statistics sequentially.

4 Experimental results

4.1 Sample examples from real-world datasets

We first considered two well-known real-world time series from finance and geophysics, consisting respectively of 19344 and 4050 continuous univariate measures, namely the daily log-return of the Dow Jones and well-log data. For the first dataset, we chose univariate normal distributions to detect changes in variance mainly. For the second dataset, we chose univariate normal distributions with a fixed variance to detect different regimes related to the mean.

The results are represented in Figure 1, and show that the proposed scheme has been able to detect relevant change points regarding variance and mean respectively. The changes in the first dataset reflect financial fluctuations that can a posteriori be related to politic and economical events [1, 16]. Concerning the second dataset, changes in the mean carry geological information that is interpretable in terms of rock stratification structure [11, 13, 28].

4.2 Evaluation for onset detection in audio signals

Finally, we assessed qualitative improvements of the approach on a difficult dataset for musical onset detection with standard methodological guidelines for evaluation [18]. The audio was represented through normalized magnitude spectra with a frame size of 1024 samples and a hop size of 126 samples at a sampling rate of 12600 Hz, leading to discrete histograms of 513 dimensions, modeled with categorical distributions. We compared the proposed approach (`GLR`) to a baseline spectral flux method based on the Kullback-Leibler divergence with the very same analysis parameters (`SF`), and to a recent information-geometric approach based on a symmetrized Kullback-Leibler divergence coupled with a more elaborate representation via a filter bank on a logarithmic frequency scale (`IG`) [7].

The obtained results show that both `GLR` and `IG` largely outperform `SF`, with respective F-measures of 64.52 %, 57.72 %, 37.53 %, hence proving the relevancy of an information-geometric approach to onset detection in audio signals. The baseline SF method is actually a crude approximation of the exact GLR scheme, with search for a change point in a sliding window of two observations, and with rough estimation of the unknown parameters before change using the first observation only. The results thus confirm the benefits in using exact statistics instead of approximation heuristics, though spectral flux is still considered as a reference method in the literature. Finally, even if the sound representation considered in GLR is simplistic, it still significantly improves the results over IG. This is because IG relies on a heuristic detection procedure defined over a somewhat ad hoc geometry, whereas both the detection procedure and the geometry are tied to relevant statistical considerations in GLR.

5 Conclusion

We discussed the problem of online change detection in exponential families with unknown parameters before and after change. We considered a standard statistical approach based on generalized likelihood ratio test statistics. We interpreted these statistics in the framework of information geometry, hence providing a unified view of change detection for many statistical models and corresponding distances functions. We also discussed a tractable scheme for change detection based on exact generalized likelihood ratios and applied it to various datasets.

References

1. R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. Technical report, University of Cambridge, Cambridge, UK, 2007.
2. S.-i. Amari and H. Nagaoka. *Methods of Information Geometry.* American Mathematical Society, Providence, USA, 2000.
3. O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory.* Wiley, Chichester, UK, 1978.
4. M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application.* Prentice-Hall, Inc., Upper Saddle River, USA, 1993.
5. L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics, Hayward, USA, 1986.
6. S. Canu and A. Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7–9):714–720, Mar. 2006.
7. A. Cont, S. Dubnov, and G. Assayag. On the information geometry of audio streams with applications to similarity computing. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):837–846, May 2011.
8. F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(8):2961–2974, Aug. 2005.
9. A. Dessein. *Computational Methods of Information Geometry with Real-Time Applications in Audio Signal Processing.* PhD thesis, Université Pierre et Marie Curie, Paris, France, Dec. 2012.
10. A. Dessein and A. Cont. An information-geometric approach to real-time audio segmentation. *IEEE Signal Processing Letters*, 20(4):331–334, Apr. 2013.

11. P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006.
12. M. A. Girshick and H. Rubin. A Bayes approach to a quality control model. *The Annals of Mathematical Statistics*, 23(1):114–125, Mar. 1952.
13. Y. Guédon. Exploring the segmentation space for the assessment of multiple change-point models. Technical report, Institut National de Recherche en Informatique et en Automatique, Sophia Antipolis, France, 2008.
14. Z. Harchaoui, F. Bach, and E. Moulines. Kernel change-point analysis. In *Advances in Neural Information Processing Systems*, volume 21, pages 609–616. NIPS Foundation, La Jolla, USA, 2009.
15. Z. Harchaoui and C. Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493, Dec. 2010.
16. R. Killick, P. Fearnhead, and I. A. Eckley. Optimal detection of changepoints with a linear computational cost. Technical report, Lancaster University, Lancaster, UK, 2011.
17. T. L. Lai and H. Xing. Sequential change-point detection when the pre- and post-change parameters are unknown. *Sequential Analysis: Design Methods and Applications*, 29(2):162–175, Apr. 2010.
18. P. Leveau, L. Daudet, and G. Richard. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In *5th International Conference on Music Information Retrieval*, pages 72–75, Barcelona, Spain, Oct. 2004.
19. G. Lorden. Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, 42(6):1897–1908, Dec. 1971.
20. E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, June 1954.
21. M. Pollak and D. Siegmund. Approximations to the expected sample size of certain sequential tests. *The Annals of Statistics*, 3(6):1267–1282, Nov. 1975.
22. A. S. Polunchenko and A. G. Tartakovsky. State-of-the-art in sequential change-point detection. *Methodology and Computing in Applied Probability*, 14(3):649–684, Sept. 2012.
23. V. H. Poor and O. Hadjiliadis. *Quickest Detection.* Cambridge University Press, New York, USA, 2009.
24. S. W. Roberts. A comparison of some control charts procedures. *Technometrics*, 8(3):411–430, Aug. 1966.
25. R. T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, USA, 1970.
26. A. N. Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability and Its Applications*, 8(1):22–46, 1963.
27. D. Siegmund and E. S. Venkatraman. Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics*, 23(1):255–271, Feb. 1995.
28. R. Turner, Y. Saatci, and C. E. Rasmussen. Adaptive sequential Bayesian change point detection. In *NIPS Workshop on Temporal Segmentation*, Whistler, Canada, Dec. 2009.
29. 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*, volume 23, pages 2343–2351. NIPS Foundation, La Jolla, USA, 2010.