# DETECTING CHANGES IN DATA WITH LONG-RANGE DEPENDENCE

Jonathan J. Wylie

Department of Mathematics,
City University of Hong Kong,
Tat Chee Avenue,
Kowloon,
Hong Kong.
e-mail: mawylie@cityu.edu.hk

**Abstract**

There are a number of important applications in which one must identify behavioral changes in data. We will study the classical change-point problem in which one must estimate the location of a change in behavior from a given data set. In the case in which the statistical errors are independent, the problem has been widely studied and there exists a wide literature that spans a number of distinct fields. On the other hand, for cases in which there is dependence in the statistical errors, the standard techniques that can be applied in the independent case are not appropriate and there are much fewer results in the literature. In recent years, the importance of effects associated with dependence in data has been more completely understood and it has become increasingly clear that dependence can lead to fundamentally different behavior than is observed in problems with independent data. This is true in a number of applications relating to financial data in which dependence can play a critical role in determining the price dynamics and hedging strategies. It is also of crucial importance in telecommunications data. In this work, we will discuss the role that dependence plays when attempting to identify a single change point in a continuous process via discrete observations and consider how increasing the frequency of the observations affects the accuracy of the detection process.

There has been extensive work on the change-point problem and authors from a number of fields have made important contributions. This has led to a number of varied techniques being applied such as linear-model based approaches and nonparametric approaches. We refer the reader to Basseville and Nikiforov (1993) for a comprehensive review of the subject.

It is probably true to say that most of the works have focused on the case of uncorrelated processes or independent data. In particular, Carlstein (1988) studied sequences in which

the distribution is stationary on both sides of the jump and using a class of nonparametric estimators that are based on cumulative sums of empirical distributions on either side of a proposed change-point. A very general class of estimators was proposed by Dumbgen (1991) who obtained an $O_p(n^{-1})$ rate of convergence, where $n$ is the number of data points.

A method based on wavelets has been proposed by Wang (1995) to estimate the location of a discontinuity of a process in which the mean varies continuously on either side of the discontinuity. Wavelets were also used by Wong *et al.* (2001) who considered jump detection in a heteroscedastic autoregressive model. In this case, he proved the consistency the proposed estimator.

Heavy correlations in the noise make the problem significantly more challenging. One of the few results for long-range dependent sequences was obtained by Ben Hariz *et al.* (2007) who considered data with long-range dependence where correlations decay to zero algebraically or faster. In this case they found that a natural family of estimators (similar to that proposed by Dumbgen) achieve the $O_p(n^{-1})$ rate of convergence.

However, as far as we are aware, previous works cannot be applied to the case in which there is a finite interval and the sampling frequency tends to infinity. This is because the noise in the processes in previous works have had either independent increments, weakly dependent structure, or at least with correlations structures that tend to zero.

When any continuous process is sampled at sufficiently high frequency it is clear that correlations between neighboring points must become important. Therefore, the work presented here is of importance to any application in which one must detect a change in behavior in a continuous random process. In this case, we show that the traditional estimators (based on cumulative sums) and wavelet-based estimators may not even be consistent. We propose a different class of estimators that use local information and show that these estimators dramatically outperform cumulative-sum-based estimators and wavelet-based estimators. In particular, we focus on the case of Gaussian data to derive concrete bounds on the probability of predicting an incorrect change-point location.

We consider a model for our underlying process of the following form

$$Y_t = \delta \mathbf{1}_{\{\mathbf{t}>\theta\}} + \mathbf{X_t},\ \mathbf{t} \in [\mathbf{0},\mathbf{T}], \tag{0.1}$$

where $\delta$ is the size of the jump and $\theta \in (0,T)$ is the location of the jump. In particular, we make the assumption that $(X_t)_{0\le t\le T}$ has a constant mean. Without loss of generality, we take the mean to be zero. We further assume that the process has the following property (that is typically used to ensure that a process is continuous)

$$\exists \alpha > 0, C > 0, \text{ such that for } s,t \in [0,T], \quad \mathbf{E}\left(\mathbf{X_s} - \mathbf{X_t}\right)^{\mathbf{2}} \le \mathbf{C}\left|\mathbf{s}-\mathbf{t}\right|^{\alpha}. \tag{0.2}$$

In any application, one can only observe the process at a finite number of points. Without loss of generality, we will set $T = 1$. Furthermore, we will restrict our attention to the case of equally spaced observations of the following form $t_1 = T/n, ..., t_k = kT/n, ..., t_n = T$. Using the observations $(Y_{t_i})_{1\le i\le n}$ , the task that we aim to accomplish is to estimate within which interval $[\hat{\theta}, \hat{\theta} + 1/n[$ the discontinuity in the process lies. We will adopt a localized version of the cumulative-sum estimator of the following form

$$\hat{\theta} = \frac{1}{n} \min\left(\arg\max_{1\le k<n} \{|U_k|\}\right), \tag{0.3}$$

with

$$U_k = \frac{1}{k-k_l+1}\sum_{i=k_l}^{k} Y_i - \frac{1}{k_u-k}\sum_{i=k+1}^{k_u} Y_i,\ k = 1,..,n-1, \tag{0.4}$$

where $k_l = \max(1, k-L+1)$ and $k_u = \min(k+L, N)$. Here, $L$ is a window for a mean comparison between two subsequences of size $L$, one ends at $k$ and the second starts from $k+1$. The parameter $L$ can take values ranging from 1 to $n$.

We will show that this estimator dramatically outperforms traditional cumulative-sum estimators, and show that if the information used by the estimator is sufficiently localized, that exponentially fast convergence can be achieved.

**Key Words:** Change-point detection, Correlations, Long-range dependence.

# References

[1] Basseville M., Nikiforov I.V., Detection of Abrupt Changes: Theory and Application. Englewood Cliffs, NJ: Prentice Hall, 1993.

[2] Ben Hariz S., Wylie J.J., Zhang Q., Optimal rate of convergence for nonparametric change-point estimators for non-stationary sequences, Ann. Stat 35 (2007a) 1802–1826.

[3] Carlstein E., Nonparametric change-point estimation, Ann. Statist. 16 (1988) 188–197.

[4] Dumbgen L., The asymptotic behavior of some nonparametric change-point estimators, Ann. Statist. 19 (1991) 1471–1495.

[5] Wang Y., Jump and sharp cusp detection by wavelets. Biometrika 82 (2) (1995) 385–397.

[6] Wong H., Ip W., Li Y., Detection of jumps by wavelets in a heteroscedastic autoregressive model, Statist. Probab. Lett. 52 (4) (2001) 365–372.