Nonlin. Processes Geophys. Discuss., 1, 371–401, 2014
www.nonlin-processes-geophys-discuss.net/1/371/2014/
doi:10.5194/npgd-1-371-2014


Nonlinear Processes in Geophysics
Discussions
Open Access



# Instability and change detection in exponential families and generalized linear models, with a study of Atlantic tropical storms

**Y. Lu[1] and S. Chatterjee[2]**

[1]Amazon.com, Seattle, Washington, USA
[2]University of Minnesota, School of Statistics, Minneapolis, Minnesota, USA

Received: 10 February 2014 – Accepted: 21 February 2014 – Published: 21 March 2014

Correspondence to: S. Chatterjee (chatterjee@stat.umn.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union & American Geophysical Union.

## Abstract

Exponential family statistical distributions, including the well-known Normal, Binomial, Poisson, and exponential distributions, are overwhelmingly used in data analysis. In the presence of covariates, an exponential family distributional assumption for the response random variables results in a generalized linear model. However, it is rarely ensured that the parameters of the assumed distributions are stable through the entire duration of data collection process. A failure of stability leads to nonsmoothness and nonlinearity in the physical processes that drive the data under. In this paper, we propose testing for stability of parameters of exponential family distributions and generalized linear models. A rejection of the hypothesis of stable parameters leads to change detection. We derive the related likelihood ratio test statistic. We compare the performance of this test statistic to the popular Normal distributional assumption dependent cumulative sum (Gaussian-CUSUM) statistic in change detection problems. We study Atlantic tropical storms using the techniques developed here, to understand whether the nature of these tropical storms has remained stable over the last few decades.

## 1 Introduction

One important way in which nonlinear structures may be present in data related to many physical and natural phenomena is by structural breaks and changes. Generally, elicitation of the time and nature of such breaks with statistical guarantees involves change detection techniques like the *cumulative sum* (CUSUM), or the *exponentially weighted moving average* (EWMA).

The standard framework for applying such change detection techniques requires assuming that the order in which the sampled observations arrive is known, with the question of interest being whether the data generating process has remained stable over time. The observations are assumed to follow a known Gaussian distribution, and





NPGD
1, 371–401, 2014



Y. Lu and S. Chatterjee

are monitored for a potential change to a different, but still known, Gaussian distribution. Statistical guarantees are typically expressed in terms of expected *run length*, i.e., how long it takes on average for a true change to be detected, when there is a control for the expected length of time before false signaling occurs.

These Normality-based sequential monitoring and stability detection techniques originated from industrial process control (Page, 1954), although they have far ranging applications nowadays. Examples of such applications are in the fields of health care monitoring (Steiner et al., 1999), detection of genetic mutation (Krawczak et al., 1999), credit card and financial fraud detection (Bolton and Hand, 2002), and insider trading in stock markets (Meulbroek, 1992), detect jamming attacks in wireless networks (Chen et al., 2007).

Note that in many modern applications, the assumption of Normality is not tenable. In this paper, we discuss change detection in general exponential family, and in regression models including generalized linear models like logistic regression and log-linear regression. Our main finding in this paper is that while at a mathematical level, the *general form* of the CUSUM statistic remains the same as that of the traditional, Normality-based CUSUM statistic, the *component terms* that make up the general form changes. We present several mathematical results concerning the different kinds of CUSUM statistics that may result, depending on the probabilistic structure under consideration. A natural question here is on the performance of the Normality-based CUSUM statistic, when the probability models do not satisfy the Gaussian assumptions. We study this issue, and present mathematical results, simulation studies and discussions about when and how the Gaussian-CUSUM may yield high quality results. Finally, we discuss properties of Atlantic tropical storms, and use the techniques developed in the rest of this paper to study structural changes in the fundamental physical properties for which we have data records for such storms.

In order to generalize the scope of statistical change detection tools, in this paper we propose a variant of the sequential industrial monitoring framework, by considering the stability of the data generation process as a problem of detecting the *time of*





NPGD
1, 371–401, 2014
Instability detection
Y. Lu and S. Chatterjee

*the distributional change*. That is, we conduct a hypothesis test, and under the null hypothesis, the data generation process remains stable through the entire sampling time $t = 1, \ldots, n$. Under the alternative hypothesis, the distribution of the individual observations remain stable up to an unknown point of time $\tau \leq n$ and then it changes to another distribution. There are several advantages to this *testing for distributional stability* (TDS hereafter) approach. First, the sequential process monitoring statistics like CUSUM are obtained as a special case, so there is no loss of generality. Second, the TDS approach is applicable to non-sequential framework, so it may be applied to a finite sample data record also. Third, the TDS approach is applicable in a much broader framework compared to just temporally ordered observations, and can be generalized to any kind of partition of the data. Fourth, multiple changes of distribution of the observations is also easily incorporated in the TDS framework. The two generalizations, that of extending TDS to any partitioning of the data and that of using multiple change times, can be easily visualized in this hypothesis testing framework, but we do not pursue them here. However, we briefly comment on these generalizations in Sect. 3 below. We call the proposed testing procedure the *exponential family CUSUM* (or EF-CUSUM in short), while the statistic obtained under Gaussian framework is called normal-CUSUM or Gaussian-CUSUM.

Simulation studies show that in most situations, EF-CUSUM method performs better than Gaussian-CUSUM. The EF-CUSUM has a shorter average run length, smaller variation of run length and shorter maximum run length compared with Gaussian-CUSUM. Moreover, smaller shifts can be detected more quickly by EF-CUSUM than by Gaussian-CUSUM, which is a big advantage of using EF-CUSUM. Under some circumstances the Gaussian-CUSUM approximates the EF-CUSUM well, we discuss this issue below. It is also important to note that whether the change point $\tau$ is at the beginning, in the middle or at the end, the EF-CUSUM generally outperforms the Gaussian-CUSUM, so the unknown parameter $\tau$ plays little role in our analysis. Finally, in the case of a large parameter shift, the exponential family CUSUM and the Gaussian-





NPGD
1, 371–401, 2014

CUSUM perform similarly. This is not unusual, and even visual and ad hoc techniques suffice for many cases of large changes.

We also extend our study to that of parameter change in the generalized linear model. In this context, Brown et al. (1975), and Jandhyala and MacNeill (1991) discussed general linear model, Lee et al. (2004), Chihwa and Ross (1995), and Ploberger et al. (1989) focused on detecting linear model with different types of error terms. In this paper we propose methodology for detecting change in regression coefficients in the generalized linear model setting and the EF-CUSUM scheme associated with it.

Our case study for illustrating our instability and change detection techniques is based on Atlantic tropical storm data. There are several studies in recent times on whether, and how, the properties of these storms have changed with climate change. Note that such storms can do immense harm to human and other living beings and to property, consequently a change in their patterns is of interest. Apart from being of current interest, the presence of some amount of evidence for change in the literature is helpful for evaluating whether our proposed methods can detect known instabilities. We study the yearly number of such storms, as well as the joint relationship between pressure and windspeed. We detect changes compatible with known facts. Interestingly, we find that although windspeeds and central pressure values of Atlantic hurricanes have changed, they have changed in-sync, that is, their mutual relationship has remained stable over time. This lends credence that our methodology might be able to detect true changes and discard false signals well, since large scale energy balance relationships (as that between pressure and windspeed) are not expected to change.

The paper goes as follows: Section 2 is literature review. Section 3 deals with EF-CUSUM statistic derivation. Multivariate Gaussian-CUSUM is discussed as well, with covariance matrix either singular or positive definite. A few examples are given as to how to derive CUSUM statistic, and Tables 1 and 2 are provided for the convenience of readers. Section 4 talks about change detection in the generalized linear model setting.

Section 5 contains simulation studies, the results are summarized in Tables 3–5 and Fig. 1. The data analysis for Atlantic tropical storms is provided in section 6. Conclusion and future work go into Sect. 7.

## 2 Literature review

In this section we provide a partial list of techniques for change detection. As mentioned earlier, some of these originated in industrial quality context, and related methods include Shewhart control charts Shewhart (1931), EWMA control charts Roberts (1966) and CUSUM Page (1954).

In the context of the CUSUM statistic, which originated from Page (1954) and Page (1955), Lorden (1971) proposed an asymptotic optimality using the minimax criterion. Later, Moustakides (1986) also established that under Lorden's criterion, when the data is independently and identically distributed with known distributions before and after the change, the CUSUM procedure was indeed optimal. Additionally, Ritov (1990) showed that CUSUM was Bayesian optimal under Lorden's measure, and Pollak (1987) derived asymptotic expression for average run length.

The CUSUM technique has been extended to better suit practical needs, including Shu et al. (2010) on adaptive CUSUM, Hawkins (1992) on robust average run length with Winsorization, Liu et al. (2006) proposed transformation of exponential data into approximately normal distribution and compared transformed CUSUM with existing CUSUM procedures. Also, Yashchin (1993) proposed transforming serially correlated observations (such as ARMA) into independent, identically distributed sequences while keeping average run length roughly the same. In other directions, Lucas and Saccucci (1990) compared the average run length properties of EWMA with CUSUM, Atienza et al. (2000) proved that the CUSUM scheme that utilized BCUSUM mask was uniformly most powerful and compared it with other existing CUSUM procedures, MacEachern et al. (2007) developed robust CUSUM by modifying the likelihood function, Albers and Kallenberg (2009) proposed CUMIN charts for grouped data and compared

CUMIN with CUSUM and Shewhart charts, Chatterjee and Qiu (2009) proposed CUSUM control charts with control limits estimated using bootstrapping when the distribution was unknown, Steiner et al. (1999) used simultaneous CUSUM control charts to monitor correlated bivariate outcomes in the field of medical research, Crosier (1988) proposed vector CUSUM and Hotelling $T^2$ based CUSUM when dealing with multivariate case and compared them to Shewhart scheme, Lucas (1982) proposed Shewhart-CUSUM scheme to draw advantages of both methods for quick detection of mean change in the normal distribution setting, and Morais and Pacheco (2006) extended the approach to binomial data.

Some researchers have treated special cases in the EF-CUSUM family, including Hawkins and Olwell (1997) on detecting known location and shape change in inverse gamma distribution, Hawkins and Zamba (2005) on change point detection in unknown mean and variance for normal distribution, Watkins et al. (2008) used negative binomial CUSUM to study outbreaks of Ross River virus disease and compared it to Early Aberration Reporting System (EARS) CUSUM algorithms, Wu et al. (2008) studied large shifts in fraction non-conforming in Poisson CUSUM chart, Lucas (1985) improved the Poisson CUSUM with FIR and introduced two-in-a-row rule to robust CUSUM. Khan (1979) showed optimality of CUSUM for exponential distribution by calculating ARL using Wald's approximation. Healy (1987) discussed shift in mean and covariance for multivariate normal distribution using CUSUM, Alwan (2000) proposed transformation to normality to deal with EF-CUSUM chart, Severo and Gama (2010) discussed using Kalman Filter and CUSUM to detect residual mean and variance in the regression model, and Qiu and Hawkins (2001) used rank-based CUSUM procedure to deal with multivariate measurements without normality assumption.

## 3 Distributional stability in exponential families

Let the data be the random sample $\{X_1, ..., X_n\}$, where we know $X_1$ is observed first, then $X_2$ is observed, and so on. We assume that $X_1, ..., X_\tau$ are identically and independently distributed following an exponential family distribution with probability density or mass function given by

$$p(x;\theta,\delta) = \exp\left\{a(\phi)^{-1}(x\theta - b(\theta)) + c(x,\phi)\right\}.$$

Here the parameters are $\theta$, which is of the same dimensionality as each of the data-points, and $\phi$.

We assume that $X_{\tau+1},\ldots$ are identically and independently distributed from another exponential family distribution, with probability density function given by

$$p(x;\theta+\delta_1,\phi+\delta_2) = \exp\left\{a(\phi+\delta_2)^{-1}(x(\theta+\delta_1) - b(\theta+\delta_1)) + c(x,\phi+\delta_2)\right\}.$$

Here $\tau$ is a fixed but unknown parameter denoting the time of change from one distribution to another, and $0 < \tau < \infty$. In the *testing for distributional stability* (TDS) framework we adopt in this paper, our interest is in testing the null hypothesis $H_0 : \tau \geq n$ against the alternative hypothesis $H_1 : \tau < n$. We consider all parameter values, other than $\tau$ as known constants.

Assuming some, or all, of these parameters as unknown is an easy extension but requires additional technical conditions and assumptions, and we do not consider that case in this paper. Note that the time-ordering of the observations is not an integral part to our methodology. Also, multiple change-points may be allowed. For the former, we would assume that there is some permutation of the data, say $X_{\sigma_1},\ldots,X_{\sigma_n}$ such that $X_{\sigma_1},\ldots,X_{\sigma_\tau}$ are independent and identically distributed with some exponential family distribution with parameters $\theta$ and $\phi$, while $X_{\sigma_{\tau+1}},\ldots$ independent and identically distributed with the same distribution with a different set of parameter values. Also, multiple change-points $\tau_1,\ldots,\tau_k$ can be easily accommodated in the above framework, and both the null and alternative hypothesis made more complex. In other words, we can extend our study to the case where, for some permutation of the indices, the data may be partitioned into $k_0$ segments under the null and $k_1$ segments under the alternative. Here, each segment of data is a set of independent, identically distributed





NPGD

1, 371–401, 2014

exponential family random variables with its own distinct set of parameters. Our current problem may be thought of as the special case where $\sigma_i = i$ for $i = 1, \ldots, n$, $k_0 = 1$, and $k_1 = 2$. Extensions like those described above may lead to new approaches for solving several problems in applied statistics. However, in the interest of clarity of presentation, and to keep this paper at a reasonable length, we do not pursue such extensions here.

In our first result below, we obtain the test statistic for the hypothesis test described above. We adopt the convention that $\sum_{i=a}^{b} Y_i = 0$ whenever $a > b$, for any sequence of (possibly random) reals $\{Y_i\}$.

**Theorem 3.1.** *Let*

$$Y_i = a(\phi+\delta_2)^{-1}\left(X_i(\theta+\delta_1) - b(\theta+\delta_1)\right) + c(X_i, \phi+\delta_2) - a(\phi)^{-1}\left(X_i\theta - b(\theta)\right) - c(X_i, \phi),$$

*for $i = 1, \ldots, n$, and further define $S_k = \sum_{i=1}^{k} Y_i$, adopting the convention that $S_0 = 0$.*

*The likelihood ratio test statistic for testing the null hypothesis $H_0 : \tau \geq n$ against the alternative hypothesis $H_1 : \tau < n$ is given by $T_n = S_n - \min_{0 \leq k < n} S_k$, and the null hypothesis is rejected if $T_n \geq L$ for some constant $L$.*

We omit the proof of this and several other Theorems in the interest of brevity. These proofs are available from the authors.

Note that the test statistic $T_n$ may be written recursively as $T_n = \max\{0, T_{n-1} + Y_n\}$, with $T_0 = 0$. This form is reminiscent of the the celebrated CUSUM statistic. In view of this, we call $T_n$ the *exponential family CUSUM* statistic. We obtain the classical CUSUM statistic as a special case in Corollary 3 below.

**Remark 3.1**

Recall that we reject the null hypothesis $H_0 : \tau \geq n$ if $T_n \geq L$ for some $L$. The standard method for choosing $L$ in the hypothesis testing paradigm is by controlling the probability of Type-1 error at some pre-determined level $\alpha$. However, in the sequential statistics literature, the comparable technique is to control the expectation of the *run length* under the null hypothesis. The run length $R$ is defined as the number of





NPGD
1, 371–401, 2014

observations gathered before a decision is reached on the rejection or acceptance of the null hypothesis. In the present framework, we have $R = \inf\{n : S_n - \min_{0\le k<n} S_k = T_n \ge L\}$. The value of $L$ may be obtained by fixing the value of $\mathbb{E}R(=\text{ARL})$ assuming $\tau = \infty$, at a pre-determined value $\text{ARL}_0$. The notation ARL stands for *average run length*. For ease in comparison with existing procedures for change detection, we will report expected run length $\mathbb{E}\max(R-\tau, 0)$ under the alternative as a measure of power, the way it is done in sequential statistics literature. The probability of Type-1 (Type-2) error and the expected run length under the null (alternative) hypothesis are related, though the relation is generally not easy to obtain.

Note that $T_n \ge 0$ almost surely, hence a non-trivial test is obtained only when $L$ is strictly positive. Our next result shows that this relation is fairly easy to ensure in practice.

**Theorem 3.2** $\mathbb{E}_{\tau=\infty}R\,(=\text{ARL}_0) \ge 1$ *if and only if the critical value L is positive.*

PROOF OF THEOREM 3 *The necessity part*: if $L \le 0$, since $R = \inf\{n : S_n - \min_{0\le k<n} S_k \ge L\}$, we have $S_0 - \min_{0\le k<0} S_k = 0 \ge L$. Hence we have $R = 0$ almost surely, and therefore $\mathbb{E}_{\tau=\infty}(R) = 0$, which is contradictory to $\text{ARL}_0 \ge 1$.

*The sufficiency part*: if $L > 0$, then $R$ cannot be zero because $S_0 - \min_{0\le k<0} S_k = 0 < L$, hence $R$ is at least 1. Therefore $\text{ARL}_0 \ge 1$.

We now state some special cases of Theorem 3, which are of interest. Our first such result deals with the case where the observations are Normally distributed. We use the notation $\overset{\text{i.i.d.}}{=}$ for independent and identically distributed.

**Corollary 3.1**

Suppose $X_1,\ldots,X_\tau \overset{\text{i.i.d.}}{=} N(\mu,\sigma_1^2)$ and $X_{\tau+1},\ldots,X_n \overset{\text{i.i.d.}}{=} N\left(\mu+\delta_1,\sigma_2^2\right)$. For testing the null hypothesis $H_0 : \tau \ge n$ against the alternative $H_1 : 0 \le \tau < n$, the likelihood ratio statistic





NPGD
1, 371–401, 2014

Instability detection

Y. Lu and S. Chatterjee

is given by $C_n = S_n - \min_{0\le k<n} S_k$, where $S_k = \sum_{i=1}^{k} Y_i$ and

$$Y_i = \log(\sigma_1) + \frac{1}{2}\sigma_1^{-2}(X_i-\mu)^2 - \log(\sigma_2) - \frac{1}{2}\sigma_2^{-2}(X_i-\mu-\delta_1)^2 .$$

We omit the proof of this Corollary, which follows easily from Theorem 3. In the very special case where $\sigma_1 = \sigma_2 = 1, \mu = 0$, we obtain $Y_i = \left(X_i - \delta/2\right)$, and hence obtain $S_n - \min_{0\le k<n} S_k = C_n = \max\{0, C_{n-1} + X_i - \delta/2\}$, with $C_0 = 0$. This expression is that of the classical Gaussian-CUSUM, where the factor $\delta/2$ is often called the *allowance constant*.

The statistic $C_n$ defined as $C_n = \max\{0, C_{n-1} + X_i - \delta/2\}$ (with $C_0 = 0$) is often used as a default statistic for change detection. Our result above shows that this statistic may also be obtained in a non-sequential framework, however, the assumption of Normal distribution seems unavoidable. Since $C_n$ is used for change detection in non-Normal data also, it is of interest to know under what circumstance it may obtain reasonable accuracy and precision with change detection. Our next theorem describes the conditions under which using $C_n$ as a statistic may be a reasonable procedure.

**Theorem 3.3**

*Consider the framework of Theorem 3. In addition, assume that the third derivative of $b(\cdot)$ at $\theta_0$ is zero, i.e., $b'''(\theta_0) = 0$, that $\delta_1$ is small and $\delta_2 = 0$.*

*Under these assumptions, the difference between the Normality-based CUSUM $C_n$ and the exponential family CUSUM $T_n$ is as follows: $|C_n - T_n| = o(n\delta_1)$.*

Remark In the case of binomial distribution with parameter $p$, the natural parameter is $\theta = \log\left((1-p)^{-1}p\right)$, and $b(\theta) = n\log\left(1+\exp\{\theta\}\right)$, $\phi$ is taken as a constant. Also $b'''(\theta) = \left(1+\exp\{\theta\}\right)^{-4}\{n\exp\{\theta\}\left(1+\exp\{\theta\}\right)\left(1-\exp\{\theta\}\right)\}$, $b'''(\theta_0) = 0$ iff $\theta_0 = 0$. In that case, $p = \frac{1}{2}$. To conclude, when $p = \frac{1}{2}$, a change from $p \to p + \delta_1$ using Gaussian-CUSUM $\tilde{y}$ and exponential family CUSUM $y$ yield similar performance in the sense





NPGD
1, 371–401, 2014

Instability detection

Y. Lu and S. Chatterjee

that $|\tilde{y} - y| = o(\delta_1)$.

**Corollary 3.2**
For the same detection problem as above, under the condition of $b'''(\theta_0) = b''''(\theta_0) = 0$, $\delta_1$ is small and $\delta_2 = 0$, we get an even stronger result $|\tilde{y} - y| = o(\delta_1^2)$.

**Example 3.0.1**
*Change from $N_p(\mu, \Sigma_1)$ to $N_p(\mu + \delta, \Sigma_2)$*

The CUSUM for multivariate normal distribution is somewhat more complicated and therefore we divide this problem into the following cases based on the nature of the variance-covariance matrix. In all the cases listed below, the test statistic is $C_n = S_n - \min_{0 \le k < n} S_k$, where $S_k = \sum_{i=1}^{k} Y_i$ and $Y_i$ depends from one case to another.

1. $\Sigma_1 = \Sigma_2 = \Sigma$, where $\Sigma$ is positive definite.

   Based on the following density function: $f(x|\mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\}$ it is straightforward to derive the CUSUM statistic based on $Y_i = (x_i - \mu - \frac{1}{2}\delta)'\Sigma^{-1}\delta$. If we let $p = 1$, we are back to the univariate normal situation.

2. $\Sigma_1 = \Sigma_2 = \Sigma$, where $\Sigma$ is a singular.

   Assume rank $(\Sigma) = r, r < p$. By linear algebra, there exists an orthogonal matrix $Q_{p*p}$, such that $Q\Sigma Q' = \Lambda$, where $\Lambda$ is $\begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}_{p*p}$.





NPGD
1, 371–401, 2014



Y. Lu and S. Chatterjee

Here $\lambda_i > 0, i = 1,2...r$. So $Z = QX \sim N_p(Q\mu, \Lambda)$. Let **P** be the matrix $\begin{pmatrix} 1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & 1 \end{pmatrix}_{r*p}$.

Let $K = PZ \sim N_r(PQ\mu, \tilde{\Lambda})$, where $\tilde{\Lambda}$ is $\begin{pmatrix} \lambda_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \lambda_r \end{pmatrix}_{r*r}$. So the problem is reduced to a change of $N_r(PQ\mu, \tilde{\Sigma})$ to $N_r(PQ(\mu+\delta), \tilde{\Sigma})$, and we are back to case 1. The CUSUM statistic is based on $Y_i = (x_i - \mu - \frac{1}{2}\delta)'(PQ)'\tilde{\Sigma}^{-1}PQ\delta$.

3. $\Sigma_1 \neq \Sigma_2$, where $\Sigma_1$, $\Sigma_2$ are both positive definite.

   Following previous discussion, the CUSUM statistic is based on $Y_i = \frac{1}{2}\log(|\Sigma_1|^{-1}|\Sigma_2|) + \frac{1}{2}(x_i - \mu - \delta)'\Sigma_2^{-1}(x_i - \mu - \delta) - \frac{1}{2}(x_i - \mu)'\Sigma_1^{-1}(x_i - \mu)$.

4. $\Sigma_1 \neq \Sigma_2$, where $\Sigma_1$ and $\Sigma_2$ are both singular.

   Based on discussion of case 2, our CUSUM statistic is based on $Y_i = (\frac{r_2}{2} - \frac{r_1}{2})\log(2\pi) + \frac{1}{2}\log(|\tilde{\Lambda_1}|^{-1}|\tilde{\Lambda_2}|) - \frac{1}{2}(P_1Q_1(x_i - \mu))'\tilde{\Lambda_1}^{-1}(P_1Q_1(x_i - \mu)) + \frac{1}{2}(P_2Q_2(x_i - \mu - \delta))'\tilde{\Lambda_2}^{-1}(P_2Q_2(x_i - \mu - \delta))$. Here $P_1, Q_1, P_2, Q_2$ are such that $P_1Q_1\Sigma_1Q_1'P_1' = \tilde{\Lambda}_1$, $P_2Q_2\Sigma_2Q_2'P_2' = \tilde{\Lambda}_2$, and rank ( $\tilde{\Lambda}_1$ ) = rank ( $\Sigma_1$), rank( $\tilde{\Lambda}_2$ ) = rank ( $\Sigma_2$ ), $\tilde{\Lambda}_1$, $\tilde{\Lambda}_2$ are $r_1 \times r_1$ and $r_2 \times r_2$ diagonal matrix.

5. $\Sigma_1 \neq \Sigma_2$, where $\Sigma_1$ is positive definite, $\Sigma_2$ is singular.

   In this case we have $Y_i = \frac{r_2 - p}{2}\log(2\pi) + \frac{1}{2}\log(|\tilde{\Lambda}_1|^{-1}|\tilde{\Lambda}_2|) + \frac{1}{2}(P_2Q_2(x_i - \mu - \delta))'\tilde{\Lambda_2}^{-1}(P_2Q_2(x_i - \mu - \delta))$
   $- \frac{1}{2}(x_i - \mu)\Sigma_1^{-1}(x_i - \mu)$, where $P_2Q_2\Sigma_2Q_2'P_2' = \tilde{\Lambda}_2$, rank ( $\tilde{\Lambda}_2$ ) = rank ( $\Sigma_2$), $\tilde{\Lambda}_2$ is $r_2 \times r_2$ diagonal matrix.





NPGD
1, 371–401, 2014

## 4 Generalized linear model and CUSUM

In this section, we consider data of the form $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$. Here, the $y_i$'s are the responses, and the $\mathbf{x}_i$'s are covariates that are considered to be fixed constant vectors. We assume that $y_i's$ come from the distribution $p(y_i|\theta_i) = \exp\{a(\phi)^{-1}(y_i\theta_i - b(\theta_i)) + c(y_i, \phi)\}$, where $\theta_i = \mathbf{x}_i'\beta$ is the canonical parameter under stable distributional regime and $a(\phi) > 0$ is a dispersion parameter. Our main result below generalizes the main result of the previous section, and presents change detection test statistic for generalized linear models:

**Theorem 4.1**
*Assume that $(y_1, \mathbf{x}_1), \ldots, (y_\tau, \mathbf{x}_\tau)$, the true model is $\theta_i = \mathbf{x}_i'\beta$, and for $(y_{\tau+1}, \mathbf{x}_{\tau+1}), \ldots, (y_n, \mathbf{x}_n)$, the true model is $\theta_i = \mathbf{x}_i'(\beta + \delta)$, where $\beta$, $\delta$ is known. For the hypothesis testing $H_0 : \tau \geq n$ vs. $H_1 : 0 \leq \tau < n$. If we denote $z_i = y_i \mathbf{x}_i'\delta - b(\mathbf{x}_i'(\beta + \delta)) + b(\mathbf{x}_i'\beta)$ and $S_k = \sum_{i=1}^{k} z_i$, then the rejection region is $S_n - \min_{0 \leq k < n} S_k \geq L$.*

## 5 Simulation study

In this section, we discuss a simulation study on the change of parameter(s) for binomial, exponential, gamma and poisson distributions, and compare the EF-CUSUM statistic with the Gaussian-CUSUM statistic, under the constraint that the mean and the standard deviation of both distributions are equal. Based on the exponential family density $f(x; \theta, \phi) = \exp\{a(\phi)^{-1}(x\theta - b(\theta)) + c(x, \phi)\}$, it is easy to calculate $E(X) = b'(\theta)$, and $\mathrm{var}(X) = b''(\theta)a(\phi)$. When there is change in parameter from $\theta$ to $\theta + \delta_1$ and from $\phi$ to $\phi + \delta_2$, we have $E(X) = b'(\theta + \delta_1)$ and $\mathrm{var}(X) = b''(\theta + \delta_1)a(\phi + \delta_2)$. So the corresponding Gaussian assumption-based setting is a change from $N(b'(\theta), b''(\theta)a(\phi))$ to $N(b'(\theta + \delta_1), b''(\theta + \delta_1)a(\phi + \delta_2))$.





NPGD
1, 371–401, 2014

The simulation procedure can be described as follows: first, we control false alarms by carefully choosing $L$ under the null distribution. We simulate $\{x_n\}_{n=1}^{T=2000} \overset{\text{i.i.d.}}{=} f(x|\theta)$ for 2500 times. Here $T$ is fixed at 2000 for illustration. The density $f(x|\theta)$ is a distribution belonging to the appropriate exponential family. Define $R = \inf\{i : S_i - \min_{0\leq k<i} S_k \geq L\}$, where $S_k = \sum_{i=1}^{k} y_k$ is EF-CUSUM statistic as we derived. For a fixed $L$, and for each simulation, we can compute a value of $R$. Its expectation $E(R)$ can be computed based on these 2500 simulations. Fixing $\text{ARL}_0 = 200$, we obtain $L$ such that $\frac{|E_0(R)-200|}{200}$ is minimized. Since $E_0(R)$ is an increasing function of L, with values ranging from 0 to $\infty$, such an $L$ exists, and is unique.

Second, we compute $E((R-\tau)^+)$ under the alternative distribution. Let $\tau$ be an unknown parameter. Again we simulate $x_1, \ldots, x_\tau \overset{\text{i.i.d.}}{=} f(x|\theta)$ and $x_{\tau+1}, \ldots, x_T \overset{\text{i.i.d.}}{=} f(x|\theta+\delta)$ for 2500 times, where $\delta$ is known. For each $\tau = 0, 1, \ldots, 100$, use the $L$ in the first step and compute $R$ for 2500 times to get the mean, median, standard deviation, and maximum of $(R(\tau))$. Finally, we repeat the same procedure for the normal case, compute $E_1(R(\tau))$ (and other summary statistics) for the Gaussian-CUSUM and compare it with $E_1(R(\tau))$ for the exponential family CUSUM.

From the simulation results in Tables 3–5 and Fig. 1, one key finding is that in most cases, EF-CUSUM statistic performs better than Gaussian-CUSUM statistic except for one occasion when the underlying distribution is exponential distribution. Here performance is based on the mean of the run length after the change time $\tau$ until a signal occurs. Also note that for small shift in parameter, exponential CUSUM has a considerable advantage over the Gaussian-CUSUM, while for large shift in parameter, EF-CUSUM still works better than Gaussian-CUSUM, but not significantly different.

We also discover that $E_1(R(\tau))$ does not vary a lot with $\tau$ changing from 0 to 100 for a particular distribution in the exponential family. Particularly, for $\tau$ close to 0 or close to 100, $E_1(R(\tau))$ is still quite stable. In the table, we showed the $E_1(R(99))$ as a representative of the performance for the statistic. In addition, the median, standard deviation and maximum of average run length tell the same story as the mean.





NPGD
1, 371–401, 2014

## 6 Hurricane data analysis

We now discuss a case study of Atlantic tropical storms, for which data is available for every six hours from its inception till finish. For each storm, the following information is recorded: date and time, hurricane identity, hurricane name, position in latitude and longitude, maximum sustained winds in knots, and central pressure in millibars.

We present our results from three studies on Atlantic hurricanes here. Each of these studies are carried out on two data sets: a longer series from 1851 to 2008 and a shorter series from 1951 to 2008. The expectation-maximization algorithm was used for missing data segments in the longer series when required, this problem does not arise in the shorter series. Please see Fig. 2 for time series plots of the data we consider here.

First, we consider the problem of testing for distributional stability for the yearly number of hurricanes between 1851 and 2008. This yearly data is modeled as Poisson($\hat{\mu}$), and a potential change to Poisson($\hat{\mu}+\delta$) is studied. We assume that any potential change point occurred after 1900, and use the data previous to it for estimating parameters. We estimate $\hat{\mu}=7.54$, and fix $\delta=c\hat{\sigma}$, where $c$ is predetermined as $\frac{1}{4}$, $\frac{1}{2}$ and 1, and $\hat{\sigma}=2.75$ is the estimated standard deviation. Note that $\sigma \approx \mu^{\frac{1}{2}}$ because for the Poisson distribution, the mean equals the variance. Then we create the Poisson CUSUM statistic as given in Table 1. We get $L$ based on $E_0(R)=200$, and search for the first n that makes $S_n - \min_{0\le k<n} S_k \ge L$ with the hurricane data.

In view of the fact that the data from the 19th century and the first half of the 20th century may not be entirely reliable, we repeated the above analysis on detecting change for the Atlantic tropical storms from year 1951–2008. We assume that the potential change could only occur after 1970. For detecting potential change Poisson($\hat{\mu}$) to Poisson($\hat{\mu}+\delta$), we now have $\hat{\mu}=9.8$, and $\delta=c\hat{\sigma}$, where $c$ is predetermined as $\frac{1}{4}$, $\frac{1}{2}$ and 1, and $\hat{\sigma}=2.97$.

The second study has two parts. For the data from 1851-2008, we model the maximum sustained winds and maximum central pressure as $N_2(\hat{\mu}, \hat{\Sigma})$, and study potential change to $N_2(\hat{\mu}+\delta, \hat{\Sigma})$. We estimate the mean $\hat{\mu}$ and variance-covariance matrix $\hat{\mathbf{\Sigma}}$ based on the first 50 observations. Here $\hat{\mu} = \begin{pmatrix} 104.8 \\ 982.99 \end{pmatrix}$, and $\hat{\Sigma} = \begin{pmatrix} \hat{\sigma_{11}} & \hat{\sigma_{12}} \\ \hat{\sigma_{21}} & \hat{\sigma_{22}} \end{pmatrix}$ $= \begin{pmatrix} 199.96 & -20.66 \\ -20.66 & 367.56 \end{pmatrix}$. Let $\delta = \begin{pmatrix} c\hat{\sigma_{11}} \\ c\hat{\sigma_{22}} \end{pmatrix}$, where $c$ is predetermined as $\frac{1}{4}$, $\frac{1}{2}$ and 1.

In a variation of the second study, we consider maximum sustained wind speed and *minimum* central pressure as $N_2(\hat{\mu}, \hat{\Sigma})$ and study potential change to $N_2(\hat{\mu}+\delta, \hat{\Sigma})$. Here $\hat{\mu} = \begin{pmatrix} 129.5 \\ 937.6 \end{pmatrix}$, and $\hat{\mathbf{\Sigma}} = \begin{pmatrix} \hat{\sigma_{11}} & \hat{\sigma_{12}} \\ \hat{\sigma_{21}} & \hat{\sigma_{22}} \end{pmatrix} = \begin{pmatrix} 376.05 & -220.47 \\ -220.47 & 237.41 \end{pmatrix}$. Let $\delta = \begin{pmatrix} c\hat{\sigma_{11}} \\ c\hat{\sigma_{22}} \end{pmatrix}$, where $c$ is predetermined as $\frac{1}{4}$, $\frac{1}{2}$ and 1.

The results are summarized in Tables 6 and 7. We discover that the number of hurricanes had a significant increase around year 1933–1936, and the strength of the hurricanes had a sharp increase around the year 1923–1924. This is consistent with the historical records. In history, the 1924 hurricane Cuba was the earliest officially classified Category 5 Atlantic hurricane on the Saffir-Simpson scale, and it became the strongest hurricane on record to hit the country; 1928 Okeechobee hurricane was the second recorded hurricane to reach Category 5 status on the Saffir-Simpson Hurricane Scale in the Atlantic basin after the 1924 Cuba hurricane. The 1933 Atlantic hurricane season was the second most active Atlantic hurricane season on record with 21 storms; The 1936 season was fairly active, with 17 tropical cyclones including a tropical depression. From the analysis of the shorter series, we detect that the year 2000–2001 saw an increase in the number of hurricanes. According to National Hurricane Center, the 2001 Atlantic hurricane season produced 17 tropical storms and hurricanes.

In the third study, we consider the relationship between the number of hurricanes $Y$, the maximum sustained winds $X_1$ and maximum (minimum) central pressure for data





NPGD
1, 371–401, 2014

between 1851 and 2008 (1951–2008) $X_2$. We model $Y$ as Poisson($\lambda$), where $\theta = \log\lambda$, $p(y,\theta) = \exp\{y\theta - e^{\theta} - \log y!\}$ and use the canonical link $\theta = (1, X)'\beta$.

For the 1851–2008 data, we take the first 50 observations, and get $\hat{\beta} = (-4.99, 0.01, 0.006)'$. We also estimate the bivariate mean and covariance as $\hat{\mu} = (104.8, 982.99)'$ and $\hat{\Sigma} = \begin{pmatrix} 199.96 & -20.66 \\ -20.66 & 367.56 \end{pmatrix}$. Secondly, we select $\delta = c\hat{\beta}$, where $c = \frac{1}{4}, \frac{1}{2}, 1$. Next we search for $L$, assuming $\mathrm{ARL}_0 = 200$. To implement this, we simulate the bivariate series $X$ using $\hat{\mu}$ and $\hat{\Sigma}$. Based on equation $\log(\hat{\lambda}) = (1, X)'\hat{\beta}$, we get $\hat{\lambda}$, and we can simulate $Y$ from Poisson ( $\hat{\lambda}$). Construct the CUSUM statistic and the stopping rule $S_n - \min_{0 \le k < n} S_k \ge L$ to satisfy $\mathrm{ARL}_0 = 200$. Finally, we fit the stopping rule to the real data and discover the signal. Results shows that there is no significant change in terms of $\beta$, which means the way how the maximum sustained winds and maximum central pressure of a hurricane relates to the number of hurricanes has not changed over the past 158 years.

For the 1951–2008 data, we take the first 20 observations, and get $\hat{\beta} = (3.08, 0.003, -0.0016)'$. We also estimate the bivariate mean and covariance as $\hat{\mu} = \begin{pmatrix} 129.5 \\ 937.6 \end{pmatrix}$, and $\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 376.05 & -220.47 \\ -220.47 & 237.41 \end{pmatrix}$. Secondly, we select $\delta = c\hat{\beta}$, where $c = \frac{1}{4}, \frac{1}{2}, 1$. Results shows that there is no significant change in terms of $\beta$, which means the way how the maximum sustained winds and minimum central pressure of a hurricane relate to the number of hurricanes has not changed over the past 58 years. Thus, the third part of our study shows broad physical relations between windspeeds and pressures have not changed, which is to be expected.

## 7 Conclusion and future work

The exponential family CUSUM generally performs better than the Gaussian-CUSUM. In practice, in situations where the data do not follow normal distribution, we should consider the appropriate distribution for modeling the data and choose the





NPGD
1, 371–401, 2014



Y. Lu and S. Chatterjee

corresponding CUSUM statistic to effectively detect the change in parameter(s) if there is any. Further details for the mathematical proofs, simulation studies, and our analysis of Atlantic tropical storms record are available from the authors.

In general, optimality results for our proposed methods should follow along lines similar to those established by Moustakides (1986) and related works, but this requires a separate proof. There are other situations of interest in geophysical studies where an exponential family model may not be appropriate. Examples include extremes, cases where the parameter is a boundary point of the support of the random variable, and mixtures of distributions. Our future work will consist of stability detection for such cases.

The presence of temporal dependence is typically not problematic; our likelihood-based schemes generalize easily to standard time series frameworks, but additional mathematical technicalities cannot be avoided. In addition, cases where the observations are not temporally ordered, or when there are multiple break points, need suitable generalizations and mathematical treatment. Note that there is a relationship between the number of structural breaks in the distribution of a data sequence, the size of such breaks, and the probabilities of true/false inference from hypothesis testing. Establishing the limits of our proposed methodology along these lines is a future work to accomplish.

*Acknowledgements.* This research is partially supported by the National Science Foundation under grant # IIS-1029711 and # SES-0851705, and by grants from the Institute on the Environment (IonE), and College of Liberal Arts, University of Minnesota.

## References

Albers, W. and Kallenberg, W. C. M.: CUMIN Charts, Metrika, 70, 111–130, 2009. 376

Alwan, L. C.: Designing an Effective EF-CUSUM Chart without the Use of Nomographs, Commun. Stat. A-Theor., 29, 2879–2893, 2000. 377

Atienza, O. O., Tang, L. C., and Ang, B. W.: A Uniform Most Powerful Cumulative Sum Scheme Based on Symmetry, J. Roy. Stat. Soc. D-Sta., 49, 209–217, 2000. 376

Bolton, R. J. and Hand, D. J.: Statistical fraud detection: A review, Stat. Sci., 17, 235–255, 2002. 373

Brown, R. L., Durbin, J., and Evans, J. M.: Techniques for Testing the Constancy of Regression Relationships over Time, J. Roy. Stat. SOC. B Met., 37, 149–192, 1975. 375

Chihwa, K. and Ross, S. L.: A Cusum Test in the Linear Regression Model with Serially Correlated Disturbances, Economet. Rev., 14, 331–346, 1995. 375

Chatterjee, S. and Qiu, P.: Distribution-Free Cumulative Sum Control Charts Using Bootstrap-Based Control Limits, Ann. Appl. Stat., 3, 349–369, 2009. 377

Chen, W., Chen, D., Sun, G., and Zhang, Y.: Defending Against Jamming Attacks in Wireless Local Area Networks, Lect. Notes Comput. Sc., 4610, 519–528, 2007. 373

Crosier, R. B.: Multivariate Generalizations of Cumulative Sum Quality-Control Schemes, Technometrics, 30, 291–303, 1988. 377

Hawkins, D. M.: Evaluation of Average Run Lengths of Cumulative Sum Charts for an Arbitrary Data Distributions, Commun. Stat. Simulat., 21, 1001–1020, 1992. 376

Hawkins, D. M. and Olwell, D. H.: Inverse Gaussian Cumulative Sum Control Charts for Location and Shape, J. Roy. Stat. Soc. D-Sta., 46, 323–335, 1997. 377

Hawkins, D. M. and Zamba, K. D.: A Change-point Model for a Shift in Variance, J. Qual. Technol., 37, 21–31, 2005. 377

Healy, J. D.: A Note on Multivariate CUSUM Procedures, Technometrics, 29, 409–412, 1987. 377

Jandhyala, V. K. and MacNeill, I. B.: Tests for Parameter Changes At Unknown Times in Linear Regression Models, J. Stat. Plan. Infer., 27, 291–316, 1991. 375

Khan, R. A.: A Sequence Detection Procedure and the Related Cusum Procedure, The Indian Journal of Statistics, Series B, 40, 146–162, 1979. 377

Krawczak, M., Ball, E., Fenton, I., Stenson, P., Abeysinghe, S., Thomas, N., and Cooper, D. N.: Human Gene Mutation Database: A biomedical Information and Research Resource, Hum. Mutat., 15, 45–51, 1999. 373

Lee, S., Tokutsu, Y., and Maekawa, K.: The CUSUM Test for Parameter Change in Regression Models with ARCH Errors, Journal of the Japanese Statistical Society, 34, 173–188, 2004. 375

Liu, J. Y., Xie, M., and Goh, T. N.: CUSUM Chart with Transformed Exponential Data, Commun. Stat. A-Theor., 35, 1829–1843, 2006. 376

Lorden, G.: Procedures for Reacting to a Change in Distribution, Ann. Stat., 42, 1897–1908, 1971. 376

Lucas, J. M.: Combined Shewhart-CUSUM Quality Control Schemes, J. Qual. Technol., 14, 51–59, 1982. 377

Lucas, J. M.: Counted Data CUSUM's, Technometrics, 27, 129–144, 1985. 377

Lucas, J. M. and Saccucci, M. S.: Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements, Technometrics, 32, 1–12, 1990. 376

MacEachern, S. N., Rao, Y., and Wu, C.: A Robust-Likelihood Cumulative Sum Chart, J. Am. Stat. Assoc., 102, 1440–1447, 2007. 376

Meulbroek, L. K.: An empirical analysis of illegal insider trading, J. Financ., 47, 1661–1699, 1992. 373

Morais, M. C. and Pacheco, A.: Combined CUSUM-Shewhart Schemes for Binomial Data, Economic Quality Control, 21, 43–57, 2006. 377

Moustakides, G. V.: Optimal Stopping Times for Detecting Changes in Distribution, Ann. Stat., 14, 1379–1387, 1986. 376, 389

Page, E. S.: Continuous inspection schemes, Biometrika, 41, 100–115, 1954. 373, 376

Page, E. S.: A Test for a Change in a Parameter Occurring at an Unknown Point, Biometrika, 41, 100–115, 1955. 376

Pollak, M.: Average Run Lengths of an Optimal Method of Detecting a Change in Distribution, Ann. Stat., 15, 749–779, 1987. 376

Qiu, P. and Hawkins, D.: A Rank Based Multivariate CUSUM Procedure, Technometrics, 43, 120–132, 2001. 377

Ploberger, W., Kramer, W., and Alt, R.: A Modification of the CUSUM Test in the Linear Regression model with Lagged Dependent Variables, Empir. Econ., 2, 65–75, 1989. 375

Ritov, Y.: Decision Theoretic Optimality of the CUSUM Procedure, Ann. Stat., 18, 1464–1469, 1990. 376

Roberts, S. W.: A Comaprison of Some Control Chart Procedures, Technometrics, 8, 411–430, 1966. 376

Severo, M. and Gama, J.: Change Detection with Kalman Filter and CUSUM, Lect. Notes Comput. Sc., 6202/2010, 148–162, 2010. 377





NPGD

1, 371–401, 2014

Instability detection

Y. Lu and S. Chatterjee

Shewhart, W. A.: Economic Control of Quality of Manufactured Product, Van Nostrand, Princeton, NJ, USA, 1931. 376

Shu, L., Yeung, H., and Jiang, W.: An Adaptive CUSUM Procedure for Signaling Process Variance Changes of Unknown Sizes, J. Qual. Technol., 42, 69–85, 2010. 376

Steiner, S. H., Cook, R. J., and Farewell, V. T.: Monitoring Paired Binary Surgical Outcomes Using Cumulative Sum Chart, Stat. Med., 18, 69–86, 1999. 373, 377

Watkins, R. E., Eagleson, S., Veenendaal, B., Wright, G., and Plant, A. J.: Applying Cusum-Based Methods for the Detection of outbreaks of Ross River virus disease in Western Australia, BMC Medical Informatics and Decision Making, 8, 37, doi:10.1186/1472-6947-8-37, 2008. 377

Wu, Z., Jiao, J., and Liu, Y.: A binomial CUSUM chart for detecting large shifts in fraction nonconforming, J. Appl. Stat., 35, 1267–1276, 2008. 377

Yashchin, E.: Performance of CUSUM Control Schemes for Serially Correlated Observations, Technometrics, 35, 37–52, 1993. 376

**Table 1.** Exponential Family CUSUM: Binomial, Exponential, Gamma and Multivariate Normal distributions.

| Type of Distribution | Density Function | EF-CUSUM based on |
|---|---|---|
| Binomial($n$, $p$): $p \to p+\delta$ | $\binom{n}{k} p^x (1-p)^{n-x}$ | $x \log\left(\frac{p+\delta}{p}\right) + (N-x)\log\left(\frac{1-p-\delta}{1-p}\right)$ |
| Poisson($\lambda$): $\lambda \to \lambda+\delta$ | $\frac{\lambda^x e^{-\lambda}}{\lambda!}$ | $x \log\frac{\lambda+\delta}{\lambda} - \delta$ |
| Gamma($\alpha$, $\beta$): $\alpha \to \alpha+\delta_1, \beta \to \beta+\delta_2$ | $\frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$ | $\frac{\delta_2}{\beta(\beta+\delta_2)} x + \delta_1 \log\frac{x}{\beta+\delta_2} - \alpha\log\frac{\beta+\delta_2}{\beta} - \log\frac{\Gamma(\alpha+\delta_1)}{\Gamma(\alpha)}$ |
| Multivariate normal: $N_p(\mu,\Sigma) \to N_p(\mu+\delta,\Sigma)$ $\Sigma$ is positive definite | $\frac{1}{(2\pi)^{\frac{p}{2}}\lvert\Sigma\rvert^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}$ | $\left(x-\mu-\frac{1}{2}\delta\right)'\Sigma^{-1}\delta$ |

**Table 2.** CUSUM Statistic for Normal Distribution: the first row is more general with both mean and variance change. The rest three rows are special cases of the first one.

| Distribution | CUSUM statistic |
|---|---|
| $N\left(\mu,\sigma_1^2\right) \to N\left(\mu+\delta_1,\sigma_2^2\right)$ | $\log\sigma_1 + \frac{1}{2}\sigma_1^{-2}(x_i-\mu)^2 - \log\sigma_2 - \frac{1}{2}\sigma_2^{-2}(x_i-\mu-\delta_1)^2$ |
| $N\left(\mu,\sigma^2\right) \to N\left(\mu+\delta,\sigma^2\right)$ | $\sigma^{-2}\left(x_i-\mu-\frac{1}{2}\delta_1\right)\delta_1 \propto \left(x_i-\mu-\frac{1}{2}\delta_1\right)\delta_1$ |
| $N\left(\mu,\sigma_1^2\right) \to N\left(\mu,\sigma_2^2\right)$ | $\log\left(\sigma_2^{-1}\sigma_1\right) + \frac{1}{2}\sigma_1^{-2}\sigma_2^{-2}\left(\sigma_2^2-\sigma_1^2\right)(x_i-\mu)^2$ |
| $N(\theta,\theta^2) \to N\left(\theta+\delta_1,(\theta+\delta_1)^2\right)$ | $\log\left((\theta+\delta_1)^{-1}\theta\right) + \frac{1}{2}\theta^{-2}(x_i-\theta)^2 - \frac{1}{2}(\theta+\delta_1)^{-2}(x_i-\theta-\delta_1)^2$ |





NPGD

1, 371–401, 2014

**Table 3.** Simulated binomial distribution changes: the three rows describe change from binomial(5, 0.95) to binomial(5, 0.90), from binomial(15, 0.95) to binomial(15, 0.90) and from binomial(5, 0.95) to binomial(5, 0.94) respectively. Here $\tau$ is fixed at 99 for illustration.

| Method | Mean | Median | Std. Deviation | Max |
|---|---|---|---|---|
| EF-CUSUM | 18.45902 | 15 | 14.36446 | 124 |
| Gaussian-CUSUM | 21.51974 | 16 | 17.84431 | 136 |
| EF-CUSUM | 9.049310 | 7 | 6.562318 | 68 |
| Gaussian-CUSUM | 10.678476 | 8 | 8.520224 | 65 |
| EF-CUSUM | 79.83406 | 59 | 71.60187 | 477 |
| Gaussian-CUSUM | 85.44692 | 62 | 79.26974 | 551 |





NPGD
1, 371–401, 2014



Y. Lu and S. Chatterjee

**Table 4.** Simulated Poisson distribution changes: the four rows describe change from Poisson(3) to Poisson(3.1), from Poisson(3) to Poisson(2.9), from Poisson(4) to Poisson(7) and from Poisson(4) to Poisson(1) respectively. Here $\tau$ is fixed at 99 for illustration.

| Method | Mean | Median | Std. Deviation | Max |
|---|---|---|---|---|
| EF-CUSUM | 101.51383 | 74 | 95.09682 | 730 |
| Gaussian-CUSUM | 111.9772 | 80 | 106.36463 | 755 |
| EF-CUSUM | 93.05656 | 68 | 86.17123 | 546 |
| Gaussian-CUSUM | 98.06841 | 71 | 90.09476 | 634 |
| EF-CUSUM | 4.480411 | 4 | 2.815292 | 22 |
| Gaussian-CUSUM | 4.587708 | 4 | 3.149742 | 22 |
| EF-CUSUM | 2.858086 | 3 | 1.232671 | 11 |
| Gaussian-CUSUM | 3.085586 | 3 | 1.159972 | 11 |





NPGD
1, 371–401, 2014

**Table 5.** Simulated Gamma distribution changes: the five rows describe change from Gamma(1, 2) to Gamma(1.5, 2.5), from Gamma(1, 2) to Gamma(1.5, 2.5), from Gamma(3, 4) to Gamma(3.5, 3.5), from Gamma(3, 4) to Gamma(3.5, 4.5) and from Gamma(10, 10) to Gamma(17, 18) respectively. Here $\tau$ is fixed at 99 for illustration.

| Method | Mean | Median | Std. Deviation | Max |
|---|---|---|---|---|
| EF-CUSUM | 9.923995 | 8 | 6.839445 | 53 |
| Gaussian-CUSUM | 15.64539 | 12 | 12.99359 | 165 |
| EF-CUSUM | 28.95583 | 25 | 19.30729 | 152 |
| Gaussian-CUSUM | 35.11858 | 29 | 27.01831 | 248 |
| EF-CUSUM | 70.26987 | 55 | 58.85137 | 437 |
| Gaussian-CUSUM | 75.67127 | 60 | 61.48744 | 417 |
| EF-CUSUM | 16.24952 | 13 | 11.87364 | 130 |
| Gaussian-CUSUM | 21.89900 | 17 | 18.15123 | 142 |
| EF-CUSUM | 1.063716 | 1 | 0.2497610 | 3 |
| Gaussian-CUSUM | 1.069783 | 1 | 0.2795194 | 3 |





NPGD
1, 371–401, 2014

**Table 6.** Atlantic hurricane data from 1851 to 2008 are used to detect any mean change in hurricane characteristics. Here $c$ is the magnitude representing the number of standard deviation from the mean. Result shows that the number of hurricane had a significant increase around 1933–1936, and strength of the hurricane increased around 1923–1924.

| Distribution | $c = \frac{1}{4}$ | $c = \frac{1}{2}$ | $c = 1$ |
|---|---|---|---|
| Poisson | 1936 | 1933 | 1933 |
| Bivariate Normal | 1924 | 1923 | 1924 |





NPGD
1, 371–401, 2014



Y. Lu and S. Chatterjee

**Table 7.** Atlantic hurricane data from 1951 to 2008 are used to detect any mean change in hurricane characteristics. Here $c$ is the magnitude representing the number of standard deviation from the mean. Result shows that the number of hurricane had a significant increase around the year of 2000, and strength of the hurricane has not changed.

| Distribution | $c = \frac{1}{4}$ | $c = \frac{1}{2}$ | $c = 1$ |
|---|---|---|---|
| Poisson | 2001 | 2001 | 2000 |
| Bivariate Normal | 2008 | 2008 | 2008 |





NPGD
1, 371–401, 2014

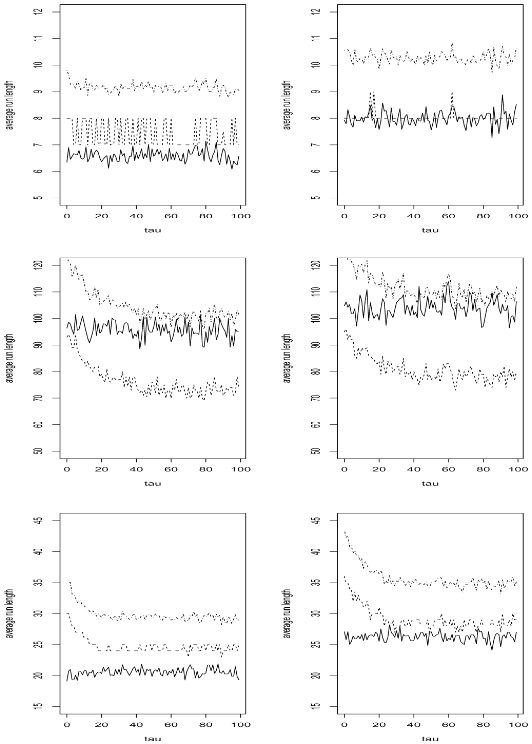


**Fig. 1.** Performance Comparison: exponential Family CUSUM with Gaussian-CUSUM. Dotdash, dashed and solid line stand for mean, median and standard deviation. The top panel describes run length comparison from Binomial(15, 0.95) to Binomial(15, 0.90), the middle panel describes run length comparison from Poisson(3) to Poisson(3.1), the bottom panel describes run length comparison from Gamma(1, 2) to Gamma(1.5, 1.5). Due to length limitation of the graphs, we here do not include the MAX line.





NPGD
1, 371–401, 2014



Y. Lu and S. Chatterjee

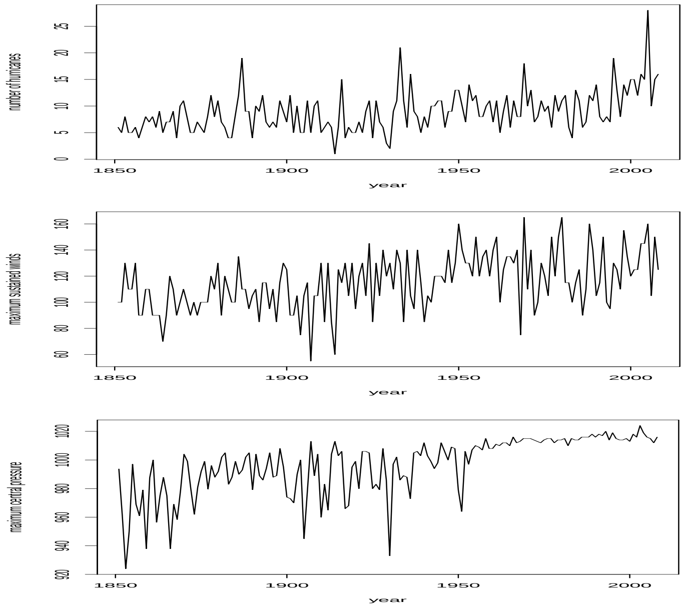


**Fig. 2.** The three panels describes time series plots for the number of hurricanes, the maximum sustained winds and the maximum central pressure across hurricanes from year 1851 to 2008.





NPGD

1, 371–401, 2014

Instability detection

Y. Lu and S. Chatterjee

Title Page
Abstract
Introduction
Conclusions
References
Tables
Figures
◂
▸
Back
Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion