Adaptive Detection of Variance Change Point

Santosh Srivastava * Ritwik Chaudhuri † Ankur Narang ‡ Maya Gupta §
Sudhanshu Singh ¶

Abstract
The task of finding variance change points has been the focus of considerable research in sequential data analysis. In spite of empirical success of many change point algorithms, there are several unresolved issues: (a) use various probabilistic modeling assumptions in one form and another, (b) fail when there are multiple change points, especially when a dominant change point masks other change points, (c) check each point is a change point or not, thus increase computation extensively. We present a novel offline algorithm which uses a dynamic mode decomposition based data-driven dynamical system and local adaptive window to iteratively detect variance change points. We propose a variance descriptor function which is used for guiding the focus-of-attention of change points. For detecting change points, it is used for generating regions of interest and providing coarse information, which automatically governs the location and the size of window to detect change points at different scales. The proposed algorithm is completely data driven, does not require a probabilistic model, and detects multiple variance change points accurately and efficiently on many time series.

Key Words: Time series, Change point, Dynamic mode decomposition, Peak finding, Scale space.

1. Introduction

We consider the problem of iteratively detecting and handling variance change in sequential data. Given a fixed sample size of a sequential data, detecting all its variance change points accurately is a challenging problem and has important applications in many areas such as healthcare, oil & gas, economics, business analytics and others. In these application areas, it is always desirable to search for the causes and sources of change points, so that such change point behavior can be properly analyzed and better understood; and if desired post detection prescriptive recommendation could be taken. Therefore, the task of finding change points has been the focus of considerable research in sequential data analysis.

Market fluctuation and change in company's policies can cause abrupt variance change and occasional bursts of activities in stock price [9], [16]. Sensor data from oil and natural gas may have abrupt changes in the variable of interest because of presence of oil and gas underneath the sea surface or anomalies in large equipment operation [11]. Similarly, change point detection could be characterized as presence of cancerous cell in a micro environment of the tissue sample and thus could be use as a useful diagnostic tool [18]. Further, implantation of new devices and sensors can affect the nature and accuracy of the measurements and can give rise to variance change points [3], [4].

In workplace platform of the call center, call volume (number of calls per day) of million of customers from hundreds of companies are monitored and analyzed around the clock and throughout the year, see Fig. 1. In call volume data, change points arise either due to known business calender like Thanksgiving, Christmas, and New Year or external

*IBM Research, Vasant Kunj, New Delhi, India 110070
†IBM Research, Vasant Kunj, New Delhi, India 110070
‡Sagacito Tech, Nehru Place, New Delhi, India 110019
§Google Research Labs, Mountain View, San Francisco, USA 94043
¶IBM Research, Vasant Kunj, New Delhi, India 110070

events like movement in economy, merging and splitting of businesses. For example, estimating a change point and its impact due to an announcement of new iphone in the market by Apple is a very difficult problem, and its inaccurate estimation generally costs hundreds of million of dollars to a company. Accurate detection of variance change point in call volume data is very useful as it provides insights into the business events, market fluctuations, maturation, and instabilities of the business and sales; and thus it helps the call center to generate good forecast, better capacity plan and seat allocation plan to handle customer's call without financial risks Fig. 1.

In addition to call volume, another important feature that governs the dynamics of a call center is the average handle time (AHT). AHT is the average time taken by an agent to answer a call from a customer, starting from initiation of the call to completion. Therefore the estimation of AHT is a very important metric because the number of agents required to answer all the calls is calculated based on AHT. In the work presented here, we are given AHT of the agents in a time series form and goal of the analysis is to find the variance change points within the data. It is important to determine the time instances when the variance changes occur in AHT. If the variation in AHT increases consistently from the previously seen pattern, that means agents are taking more time to answer calls, and also within the same time period some other calls are answered in relatively small time. For some calls, AHT is so high, it may lead to a situation where a large number of calls are kept in a queue and more often than not important calls may get dropped without being answered. Hence when the variation within AHT is large, it is important to have enough number of agents deployed to answer all the calls. Only after the variance change points are detected in a time series, one can schedule enough agents to answer calls within the specified time.

Often time series exhibit abrupt changes in variable(s) of interest such as sudden change in mean or variance or both, or occasional bursts of activities and outliers [9], [16]. Sometime, they undergo such a drastic change at a point that their behavior before or after this point is completely independent. Sequential data in which each observation is assumed to be made up of distinct components: trend, season, cycle, regression, intervention, and error, might have variance changes in all these components independently or semi-dependently. It is important to model such behavioral changes explicitly for robust estimation procedure and for extracting outlier information.

Variance change point affects larger fraction of the time series as compared to additive and innovational change points. Additive change point represents sudden "pulse" which die out immediately, while innovational one represents sudden "pulse" followed by gradual decay to the original time series. If a large variance change in time series is ignored, then the time series in high variance region would be highly erratic and therefore the standard change point algorithms would detect several additive and innovational change points in this region [16]. On the other hand, if this large variance change is accurately detected and properly handled, then only few or no change points would be left. Thus, one large variance change point is equivalent to detecting several others additive and innovational change points. Therefore ignoring variance change in the data could have a catastrophic affects, could lead to wrong statistical results, can have a detrimental impact on forecasting accuracy, and can even affect the model specification.

1.1 Challenges & Limitations in Prior Art:

Accurate detection of change points can be challenging due to the presence of inherent noise, volatility, and haphazardness. Most of the existing change point algorithms perform well when the time series has few relative change points, however difficulties in accurate

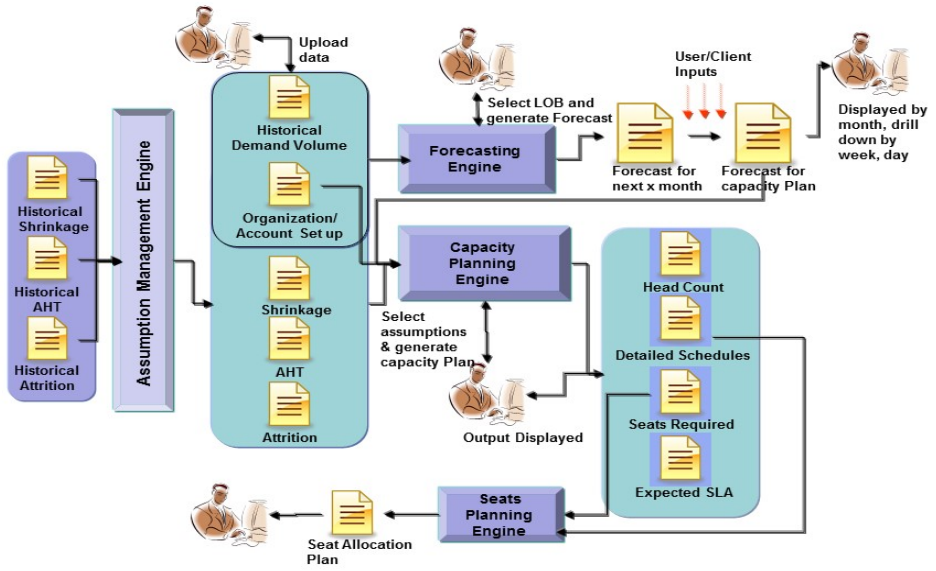

Figure 1: Workforce analytics of call center. The system leverages standard assets and power by back-end analytics of forecasting, capacity planning, and assumption management. Forecasting engine develops good quality medium to long-range forecasts of call volume and average handle time (AHT), and performs season extraction. Capacity planning engine develops analytics to improve scheduling efficiency index for both front and back office environment and prescribes optimal mix of flexible workforce. Assumption management engine develops prediction of head counts and attrition for robust and optimal planning of future workforce.

detection of change points arise due to masking effects, when time series has multiple variance change points that occurs in patches. Detection becomes most difficult when these patches overlap significantly, which models many realistic phenomenons. Before one market fluctuation or change in company policy decays out, another effect enters into the system. Further, many existing algorithms use generative models for sequential data that could make the change points detection methods suffer from several pitfalls: (a) To use parametric distribution, harsh simplifying assumptions usually need to be made e.g. the time series is stationary and error is iid and follows a Gaussian distribution, the *form* of the probability model remains the same but the parameters change over a prolonged period of time etc.. , and (b) Log likelihood can have many local maxima or the maximum is on or near the boundary. Change points methods based on autoregressive moving average (ARMA) model and likelihood ratio test have similar problems, and do not fit a good model in general, and therefore are not able to detect multiple variance change points in the data [16].

Inclán and Tiao proposed a method to detect multiple variance change points using normalized cumulative sums of squares of a sequence of independent random variables [7]. They establish two facts about cumulative sum of squares: (a) It is a function of likelihood ratio test and F statistics to test equality of variances change, (b) When there is a sudden variance change the cumulative sum of squares exhibits an abnormal pattern with high probability and therefore its plot represents a better picture of variance change. The success of their algorithm was largely dependent on: potential change point, impact factor

of the detected influential data point, order of the variance change points, and threshold value, and thus was not able to capture successfully all the variance change points present in the data. One reason their algorithm fails, is that the cumulative sum of squares is based on the assumption of uncorrelated random variables, which doesn't hold true in many realistic sequential datasets.

1.2 Key Contributions:

We make the following contributions:

1. We present a novel and efficient algorithm for variance change point detection which doesn't require any modeling assumption or any knowledge of underlying governing equation of the time series process. The proposed algorithm, DVCPD (Dynamic Mode Decomposition based Variance Change Point Detection), is completely data driven. It is based on DMD (Dynamic Mode Decomposition) [14] [15] and uses a local adaptable window and sequential hypothesis testing to iteratively detect variance change point(s).

    - The location and size of the window is automatically governed by the acceptance and rejection of the hypothesis. Thus, the variance change is detected locally as well as globally unlike the well-known methods (Tsay 1988) [16] or (Inclán and Tiao 1994) [7] where the variance change is detected on the whole dataset.

    - DVCPD addresses the masking problem by normalizing the data in a suitable way so that the presence of other variance changes, that might have been hidden due to masking, could be accurately detected. We use a suitable normalizing method using the notion of a base model. A base model is the largest subsection of the time series where variance remained constant. The reason for considering such a base model is to capture the behavior of the time series during which the variance remained constant over the longest time span.

    - Using real-world uni-variate datasets (even with multiple change points) we demonstrate the superior accuracy as well as performance of our approach as compared to well-known prior-art.

1.3 Organization

The remainder of this paper is as follows: In Section 2, we give a complete review of the DMD theory and algorithm which consider the time series as dynamical system. In Section 3 we demonstrate how DMD and local adaptable window can be combined to give a model free and robust variance change point detection algorithm (DVCPD). In Section **??**, we show extension of DVCPD to multivariate data. In Section 4 we provide empirical analysis on real-world univariate and multi-variate datasets, demonstrating the ability of our algorithm to effectively detect variance change points. In the conclusions Section 5 we summarize the merits, limitations, and extensions of the DMD method for prediction and change point detection, and highlight some open problems and directions for future research.

2. Dynamic mode decomposition (DMD) Theory

DMD is a data-driven dynamical system [14] [15] [17] [8] that works by extracting the relevant information from a sequence of data. Similar to the Arnoldi algorithm, it fits

a high-degree matrix polynomial to the data sequence. An inter-snapshot linear map is identified that acts as a low dimensional approximation of the system dynamics. This linear map is computed by processing the data sequence, generated by a nonlinear process, and represents the optimal linear operator (in a least- squares sense) that describes the evolution of the dynamics over a small time interval. The eigenvalues and eigenvectors of this map then capture the principal dynamics contained in the snapshot basis.

DMD is an attractive choice for high dimensional time series as well as multiple input-output response of a highly complex non-linear dynamical system. Unlike state space methods [3], auto regression integrated moving average (ARIMA) [2], and other predictive models that are based on various assumptions of random or Gaussian process, DMD is completely model free for non-stationary and non-linear multi-variate time series. It has an elegant formulation, has relatively low time complexity and can be applied to big data with ease.

The representation of a nonlinear process by a linear sample-to-sample map is closely linked to the concept of a Koopman operator, an analysis tool for dynamical systems. This type of spectral analysis of nonlinear processes provides the mathematical foundation of DMD and has recently been applied to complex fluid flows [12]. Consider data points collected at a given time, with a total of $T+1$ samplings in time. Let $x_t \in \mathbb{R}^J$ be a vector of the J data points collected at time $t = 1, 2, ..., T, T+1$. The data can be grouped into matrices as follows:

$$\begin{aligned} X &= [x_1\ x_2\ x_3\ ...\ x_T] \\ Y &= [x_2\ x_3\ x_4\ ...\ x_{T+1}] \end{aligned}$$

The Koopman operator A maps the data at time t to the data at time $t+1$ such that $x_{t+1} = Ax_t$. The DMD algorithm will estimate the Koopman operator A that best represents the data in Krylov matrix $K(A, x_1, T) \in \mathbb{R}^{J\times T}$ [5] defined by:

$$K \stackrel{\text{def}}{=} K(A, x_1, T) = [x_1\ Ax_1\ A^2x_1\ ...\ A^{T-1}x_1].$$

If $c \in \mathbb{R}^T$ solves the linear system $Kc = -x_{T+1} = -A^T x_1$, then it follows that [5]

$$AK = KC, \tag{1}$$

where C has the form

$$C = \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_1 \\ 1 & 0 & \cdots & 0 & -c_2 \\ 0 & 1 & \cdots & 0 & -c_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_T \end{bmatrix}.$$

The matrix $C \in \mathbb{R}^{T\times T}$ is said to be a companion matrix, whose eigenvalues are given by the characteristic polynomial

$$\det(\lambda I - C) = c_1 + c_2\lambda + \cdots + c_T\lambda^{T-1} + \lambda^T.$$

Thus, DMD algorithm connects the estimation of Koopman operator A to the companion matrix methods and sparseness of C. It follows from (1) and singular value decomposition (SVD) of $X = U\Sigma V^T$,

$$\begin{aligned} AX &= XC \\ AU\Sigma V^T &= U\Sigma V^T C \\ AU\Sigma &= U\Sigma V^T CV \\ AU &= U\left(\Sigma V^T\right) C \left(V\Sigma^{-1}\right) \\ AU &= U\tilde{C}, \end{aligned} \tag{2}$$

where $\tilde{C} = \left(V\Sigma^{-1}\right)^{-1} C \left(V\Sigma^{-1}\right) \in \mathbb{R}^{r\times r}$ is mathematically similar to the companion matrix C. Thus, in practice DMD method is the idea that because $AU = U\tilde{C}$, eigenvalues of the matrix $\tilde{C}$ approximate the eigenvalues of the Koopman operator A. However, the eigenvectors of matrix $\tilde{C}$ also approximate the eigenvectors of A. Apply eigen-decomposition of $\tilde{C} = W\Omega W^{-1}$ to

$$\begin{aligned} AU &= U\tilde{C} \\ AU &= UW\Omega W^{-1} \\ A\left(UW\right) &= \left(UW\right)\Omega \\ A\Phi &= \Phi\Omega, \end{aligned} \tag{3}$$

where $\Phi = UW$. The j^{th} DMD mode or the j^{th} eigenvector of the Koopman operator A is the j^{th} column of the matrix Φ,

$$\phi_j = Uw_j. \tag{4}$$

The overall time complexity of DMD s dominated by thin SVD and hence is $O(J.T^2 + T^3)$. Since the window for DMD computation, n, moves over the whole time series it much less than total number of time stamps. Further, parallel and online implementation of thin SVD result in analysis in real-time.

2.1 DMD Prediction

The DMD prediction of the data x_{DMD} at a lead time l after the data vector x_T was collected is given by

$$\begin{aligned} x_{DMD}(l) &= A^l x_T \\ &= \left(\Phi\Lambda\Phi^{-1}\right)^l x_T \\ &= \Phi\Lambda^l\Phi^{-1}x_T \\ &= \Phi\Lambda^l b, \qquad (5) \\ &= \sum_{j=1}^{r} \lambda^l b_j \phi_j \qquad (6) \end{aligned}$$

where $\Lambda = \text{diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_r\right) \in \mathbb{R}^{r\times r}$ is the diagonal matrix of eigenvalues of A, ϕ_j is the j^{th} column vector of the matrix Φ, and the vector b solves the over-determined system

$$\Phi b = x_T \tag{7}$$

whose least square solution is given by

$$b = \left(\Phi^T\Phi\right)^{-1}\Phi^T x_T. \tag{8}$$

2.2 Connection of DMD to Fourier Transform

In this section we show that under the assumption $x_{T+1} = x_1$, DMD eigenvalues and modes becomes the eigenvalues and eigenvectors of the well known Fourier transformation matrix. The assumption of $x_{T+1} = x_1$ can easily be seen when time series data has periodic season with period T. In order to prove the result we need a Lemma [5].

Lemma 1. *Let $\mathcal{D}_T \in \mathbb{R}^{T\times T}$ is the downshift permutation matrix that pushes the components of a vector down one notch with wraparound*

$$\mathcal{D}_T = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix},$$

then $V = F_T$, $V^{-1}\mathcal{D}_T V = \Lambda = diag\,(\lambda_1, \ldots, \lambda_T)$, where F_T is a $T \times T$ discrete Fourier transform matrix, and

$$\lambda_{j+1} = \bar{\omega}^j = cos\left(\frac{2\pi j}{T}\right) + i\ sin\left(\frac{2\pi j}{T}\right) \tag{9}$$

for $j = 0, \ldots, T-1$.

Theorem 2. *If $x_{T+1} = x_1$, then DMD eigenvalues are given by 9.*

Proof.

$$\begin{aligned} AX &= [x_2|x_3|x_4|...|x_{T+1}] \\ &= [x_2|x_3|x_4|...|x_1] \\ &= X\mathcal{D}_T \\ &= XF_T\Lambda F_T^{-1} \\ A\,(XF_T) &= (XF_T)\,\Lambda \\ A\Phi &= \Phi\Lambda, \end{aligned}$$

where $\Phi = XF_T$ is the matrix of DMD modes. The j^{th} DMD mode for $1 \le j \le T$ could be obtained by comparing j^{th} column of $\Phi = XF_T$,

$$\begin{aligned} \phi_j &= XF_T(:, j) \\ &= X \begin{bmatrix} 1 \\ \omega^{j-1} \\ \omega^{2(j-1)} \\ \vdots \\ \omega^{(T-1)(j-1)} \end{bmatrix} \\ \phi_j &= \sum_{k=1}^{T} \omega^{(k-1)(j-1)}\, x_k. \end{aligned} \tag{10}$$

Note from the expression (10) when $j = 1$ the first DMD mode becomes $\phi_1 = \sum_{k=1}^{T} x_k$, proportional to the mean. □

2.3 Convolutive Dynamic Mode Decomposition

The standard DMD serves as basis for the development of convolutive DMD (CDMD). The simplest form of the CDMD is defined as

$$Y = \sum_{p=0}^{P-1} A_p X_p + E, \tag{11}$$

where $E \in \mathbb{R}^{I\times T}$ is random error matrix, $Y \in \mathbb{R}^{I\times T}$ is a output data matrix, $A_p \in \mathbb{R}^{I\times J}$ is a set of unknown matrices. X_p's can be defined in various ways. $X = X_0 \in \mathbb{R}^{J\times T}$ is the input data matrix, and X_p means columns of X are shifted to the right p spots, while the columns shifted into the matrix from the outside are set to zero. This shift represents a time delay or lag$p$ in the input data matrix and is performed by horizontal shift matrix operator such that $X_p = XT_p$. The columns shifted into the matrix represent the prior data before the reference starting time, and if no prior information is available we could set it to zero. When $p = 0$, CDMD reduces to the standard DMD and has a closed form solution $A_0 = YX_0^\dagger$, where $X_0^\dagger = X_0^T\left(X_0X_0^T\right)^{-1} \in \mathbb{R}^{T\times J}$ is the pseudo inverse of X_0.

$$\begin{aligned} & D\left(Y \parallel \sum_p A_pX_p\right) \\ = \;& \frac{1}{2}\|Y - \sum_{p=0}^{P-1} A_pX_p\|_F^2 \\ = \;& \frac{1}{2}\text{tr}\left[\left(Y - \sum_p A_pX_p\right)^T\left(Y - \sum_p A_pX_p\right)\right] \\ \overset{(a)}{=} \;& \frac{1}{2}\text{tr}\left[\left(Y - \sum_p A_pX_p\right)\left(Y - \sum_p A_pX_p\right)^T\right], \end{aligned}$$

(a) follows from $\text{tr}(A^TA) = \text{tr}(AA^T)$. Using the identity $\nabla_A\text{tr}(AA^T) = 2A$ in calculation of gradient $\nabla_{A_p}D$ with respect to A_p, and setting it to zero we obtain

$$-YX_p^T + \sum_{j=0}^{P-1} A_jX_jX_p^T = 0, \tag{12}$$

for $p = 0, 1, \ldots, P-1$. If $P \leq 2$ one can obtain a closed form solution. Analytical solution provides insight about the matrix A_p. For example, taking $P = 2$ (12) become

$$-YX_0^T + A_0X_0X_0^T + A_1X_1X_0^T = 0 \tag{13}$$
$$-YX_1^T + A_0X_0X_1^T + A_1X_1X_1^T = 0. \tag{14}$$

Using (14) and definition of pseudo inverse $X_1^\dagger$, A_1 is rewritten as $A_1 = (Y - A_0X_0)X_1^\dagger$, and substituting it in (13) gives

$$\begin{aligned} A_0 &= Y(I - X_1^\dagger X_1)X_0^T\left(X_0X_0^T - X_0X_1^\dagger X_1X_0^T\right)^{-1}, \\ &= Y(I - P_1)X_0^T\left(X_0X_0^T - X_0P_1X_0^T\right)^{-1}, \end{aligned}$$

where $P_i = X_i^\dagger X_i$ is an orthogonal projection matrix. Similarly, substituting $A_0 = (Y - A_1X_1)X_0^\dagger$ from (13) into (14), we get

$$A_1 = Y(I - P_0)X_1^T\left(X_1X_1^T - X_1P_0X_1^T\right)^{-1}$$

When $P \geq 3$ finding an analytical solution using elimination technique becomes intractable. In order to find an optimal solution an iterative updates are necessary. From

equation (12), iterative update of A_p is obtain as follow:

$$\begin{aligned} A_p X_p X_p^T &= \left(Y - \sum_{j \neq p}^{p} A_j X_j \right) X_p^T, \\ &\overset{(a)}{=} \left(A_p X_p + Y - \sum_{p=1}^{P} A_p X_p \right) X_p^T, \\ &\overset{(b)}{=} (A_p X_p + Y - \hat{Y}) X_p^T, \\ &= A_p X_p X_p^T + (Y - \hat{Y}) X_p^T, \\ A_p^{(k+1)} &\overset{(c)}{=} A_p^{(k)} + (Y - \hat{Y}) X_p^{\dagger}, \end{aligned} \tag{15}$$

in (a) we add and subtract $A_p X_p$, in (b) we used $\hat{Y} = \sum_{p=1}^{P} A_p^{(k)} X_p$ the current estimate of Y, and in (c) $X_p^{\dagger}$ represent the pseudo inverse of X_p.

2.4 DMD Algorithm

Algorithm 1 DMD Algorithm

1: **procedure** DMD(data $= (x_1, x_2, ..., x_T)$)
2: Arrange data into **matrices** $X = [x_1|x_2|...|x_{T-1}]$ and $Y = [x_2|x_3|...|x_T]$
3: Compute **(reduced) SVD** of X, $X = U\Sigma V^T$
4: Define matrix, $\tilde{A} \triangleq U^T Y V \Sigma^{-1}$
5: Compute **eigenpairs** of $\tilde{A}$, writing $\tilde{A}w = \lambda w$.
6: Each non-zero eigenvalue λ is a **DMD eigenvalue**
7: **DMD mode** corresponding to λ is : $\varphi = \frac{1}{\lambda} Y V \Sigma^{-1} w$
8: **end procedure**

3. DVCPD Algorithm

This section provides an overview the design for the DVCPD algorithm as well as details on the phases of the algorithm (in separate sub-sections for sake of clarity) along with sample run of the main procedure used in the algorithm (Algorithm, Fig. 2).

3.1 Design Overview:

We designed the variance change point detection algorithm so that it can accurately detect local and global change points in the given time series. For this, we consider varying window sizes. Varying window sizes are supposed to describe the local state of the underlying process. We start with smaller window sizes to detect true change point locally and then increase the window size unto a significant proportion of the whole time series so we can detect change points globally. Thus, using adaptive window sizes we can detect change points both locally and globally. Further, we re-normalise the time series to cover the cases where some change points may be hard to discover due as they are in the vicinity of other change points. In the experimental section, we show that our algorithm accurately detects the variance change points as compared to typical algorithms [16] [7].

3.2 Model Construction:

We fit DMD model on the given time series, $\{X_1, \ldots, X_T\}$, using Algorithm 1. (In general, we could also use other models as well such as: (a) the best possible $ARIMA(p, d, q)$ model and estimate the exact shocks $a = \{a_1, a_2, \ldots, a_T\}$ using maximum likelihood method (Box, Jenkins, and Reinsel, 2008) [2], or (b) State space model with a represented as estimated smoothed observation disturbance or smoothed state disturbance (Durbin and Koopman, 2012) [3].)

3.3 Iterations with Adaptable Window Sizes:

This section describes the procedure AdaptWindowIter. We consider a subsection of the data by constructing a starting window of size $[\alpha T]$ in each iteration, where $\alpha \in \mathcal{P} = \{\alpha_0, \alpha_0 + \beta, \alpha_0 + 2\beta \ldots, \alpha_{max}\}$, and $[x]$ is the largest integer smaller than equal to x, and we search for the variance change point in this window. The reason behind searching variance change in a window is to detect true change point locally. If the data has a variance change within a small section then our algorithm should be able to capture it by taking windows of smaller sizes. In order for our algorithm to detect variance change globally we take the starting window size (of the last iteration) as large as α_{max} of the entire data. Hence, by adaptably changing the window size we can capture variance change points locally as well as globally.

We use the method cumulative sum of squares ((Inclán and Tiao, 1994) [7]) on the shock $\{a_1, a_2, \ldots, a_T\}$, starting with $\alpha = \alpha_0$, to detect the potential variance change point k. After determining the value of k we split the time series observation X into two subsections $\{X_1, X_2, \ldots, X_{k-1}\}$ and $\{X_k, X_2, \ldots, X_{[\alpha T]}\}$ and apply the Wald test statistics W^* to decided if the k^{th} point is a true variance change point or not. If the k^{th} is a variance change point, we move our window to the next section of data or otherwise increase the starting size of the window by $[\alpha T]$ and check for the potential change point and the true variance change point again. If two variance change points detected are within 10 data points of each other we choose the one which has smaller P-value. We iterate the algorithm for different values of α till $\alpha = \alpha_{max}$ as shown in Fig. 2.

3.4 Sample Run of AdaptWindowIter Procedure:

The execution of DVCPD algorithm (Fig. 2) can be visualized using the grid plot in Fig. 2. This grid plot is obtained by reading the results of each iteration of DVCPD on the IBM Stock Price dataset.

The first row is the first iteration of the algorithm, with an initial window size of 10% of the total length of the IBM dataset. Similarly, the subsequent rows describe subsequent iterations. In the grid plot shown, the algorithm runs with $\alpha_0 = 10\%$ to $\alpha_{max} = 30\%$, as the lower and upper bounds on the starting window sizes respectively, with a step of $\beta = 2\%$ across any two iterations. Hence there are total 11 rows in the plot. Each row describes how the algorithm progresses in each iteration. This is done by reading information from the individual smaller square boxes that comprises the grid plot. Each square box has two levels of text:

(a) The lower text, which appears as a pair of numbers describes the window. The number prior to the comma is the starting position of the window, while the number after the comma is the size of the window. In the first row, the closest integer value to 10% of the total dataset length is 36, while the window starts from 0 in every iteration. Hence the lower level of text in the first element of the first row is 0,36.

Algorithm 2 DVCPD: Variance change point detection

```
1: while no variance change is detected do
2:     DMD model construction (Algorithm, Fig. 1)
3:     procedure ADAPTWINDOWITER(data, residual)
4:         for α = α_0, (α_0 + 2 * β),..., α_max do
5:             while All data is read do
6:                 Select window(start, size)
7:                 Find Impact point(window) k
8:                 Wald test(window, k)
9:                 if rejected then variance CP
10:                    move window to next section
11:                else Increase window size by [αT]
12:                end if
13:            end while
14:        end for
15:        Determine split points and base model
16:    end procedure
17:    Time series ←Normalization
18: end while
19: Change points in previous iteration are true variance changes
```

(b) The upper level of text in the square box is the impact point observed in the window using the cumulative sum of squares technique.

As described by DVCPD Algorithm (line 8), this impact point is verified by Wald test statistics to check if the point is a true variance change point. The colour of every element describes if it was accepted by Wald test as a true variance change point. The green colour represents that these points are true variance change points, while the points in the blue colour elements were marginally rejected and points in the red colour elements were rejected by a significant margin.

Consider the first iteration of algorithm as represented in the first row of the grid plot (Fig. 2). The window increased from 36 (in the first square box) to 72 (in the second square box), while the starting position remained the same as the impact point in the first element was rejected to be a true variance change point. Similarly, the window size increases to 108 and then to 144 and 180, because the impact point fails to qualify as a true change point while the starting position of the window remains the same (0). Finally, for window (0, 216), a true change point is obtained at point 180. If the impact point qualifies as a true change point, then the starting position shifts while the size of the window remains the same.

subsectionChange Points Selection: Let m_i is the number of variance change points detected when Algorithm 2 was started by taking $\alpha = \alpha_0 + \beta(i-1)$, $i = 1, 2, \ldots, \gamma$, where, $\gamma = (\frac{\alpha_{max} - \alpha_0}{\beta} + 1)$. The number of variance change points present in the data is defined as $m =$ mode of $m_1, m_2, \ldots, m_\gamma$. This definition is intuitive since we are applying our algorithm for each α from α_0 to α_{max}, the variance change points that were detected for some value $\alpha = \alpha_0$ should also be detected for other values of α which are close to α_0 if it has a significantly large impact factor. So, we are taking the number of variance change points to be that size which has occurred most number of times. Finally the true $i^{th}, 1 \leq i \leq m$ variance change point is the mode of $k_i^1, k_i^2, \ldots, k_i^\gamma$. This definition is also intuitive because if a variance change point is a true variance change point it should get

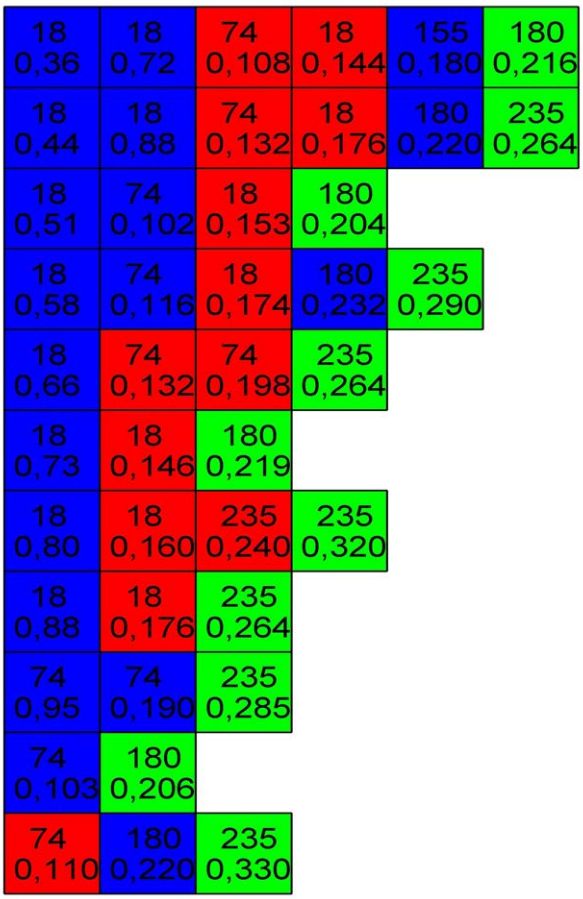

Figure 2: A sample run of the first iteration of the Algorithm 2 on the IBM time series data. Each row corresponds to different value of p. Different colors represent different range of values of Wald test statistics T for testing hypothesis $\theta = \sigma_2^2 - \sigma_1^2 = 0$ against $\theta \neq 0$: green means $T \geq 8$ and reject the hypothesis, blue means $3 \leq T < 6.6$, and red means $T \leq 3$. The pair of numbers at the bottom of each box represent the start position and the size of the window, while number on the top describes the impact point observed in the window using the cumulative sum of squares. From the above figure, reading information from the green square boxes, 235 is the variance change point in the first iteration.

detected for most values of α. By using this method let m detected variance change points are denoted by $\{k_1, k_2, \ldots, k_m\}$ such that $k_1 < k_2 < \ldots < k_m$. Then the entire original data set $\{X_1, X_2, \ldots, X_T\}$ can divided in $m + 1$ different sections such as

- subsection 1: $\{X_1, X_2, \ldots, X_{k_1-1}\}$,
- subsection 2: $\{X_{k_1}, X_{k_1+1}, \ldots, X_{k_2-1}\}$,
- .....,
- subsection $m + 1$: $\{X_{k_m}, X_{k_m+1}, \ldots, X_T\}$.

3.5 Base Model Selection & Normalization:

The variances of the data set in two adjacent subsections are significantly different from each other. The subsection with largest number of data points is taken as the base model.

It is essential to determine a base model because one needs to normalize the entire dataset with respect to the base model. The main purpose of normalization is to remove the effect of already detected variance change point and re-run the above proposed algorithm on the updated data to capture new variance change points which could have been present in the data but remained hidden due to the effect of an already detected variance change point which has a very large impact factor.

Now we describe how the data is normalized to obtain the updated data. Let assume the process in the subsection i has the largest number of data points and $i \in \{2, 3, \ldots, m\}$. Let V_b be the variance of the process $\{X_{k_{i-1}}, X_{k_{i-1}+1}, \ldots, X_{k_i-1}\}$ in subsection i. Let V_l be the variance of the process in the subsection $i-1$ with data points $\{X_{k_{i-2}}, X_{k_{i-2}+1}, \ldots, X_{k_{i-1}-1}\}$ and let Let V_r be the variance of the process in the subsection $i+1$ with data points $\{X_{k_i}, X_{k_i+1}, \ldots, X_{k_{i+1}-1}\}$. Then the updated new process $X_t^*, t = 1, 2, \ldots T$ is obtained by setting:

$$X_t^* = \begin{cases} \bar{X} + \sqrt{\frac{V_b}{V_l}}(X_t - \bar{X}) & \text{if } 1 \le t < k_{i-1} \\ X_t & \text{if } k_{i-1} \le t < k_i \\ \bar{X} + \sqrt{\frac{V_b}{V_r}}(X_t - \bar{X}) & \text{if } k_i \le t \le T \end{cases}$$

where, $\bar{X}$ is the average of all T data points in the original data set. Observe that due to this normalization variances in subsection $i-1$, i and $i+1$ of that updated data $\{X_1^*, X_2^*, \ldots, X_T^*\}$ are all same and in particular the variance is equal to V_b.

3.6 Final Change Points Selection:

The next iteration in the outer while loop in Algorithm (Fig. 2) is executed again on the updated data $\{X_1^*, X_2^*, \ldots, X_T^*\}$. This leads to further change points, re-normalization and updates to the data using the new base model. This process is repeated until no variance change point is detected. Let us assume that the entire process is repeated S times and $\mathcal{A}^I$ is the collection of true variance change points detected in the Ith repetition. Then the collection of all true variance change points is given as

$$\mathcal{K} = \bigcup_{I=1}^{s} \mathcal{A}^I. \tag{16}$$

4. Experimental Results & Analysis

In this section, we present the experimental results on variance change point detection using both real data and simulated data. The results on both univariate and multi-variate datasets have been presented below.

4.1 Analysis on Univariate Data

We performed time series modelling of uni-variate data using DMD as well as ARIMA and State Space Model (Kalman Filter) to compare and demonstrate the quality of results obtained by DMD (Algo. 1) based time series modelling and prediction. The errors obtained from these models are then fed into the variance change point detection algorithm, DVCPD to obtain the change points. The obtained change points are then compared against the change points obtained using well-known algorithms: (Tsay 1988) [16] and (Inclan & Tiao 1994) [7]. In *DVCPD* algorithm, we took the value of $\alpha_0 = 0.1$, $\beta = 0.01$, and $\alpha_{max} = 0.3$. Other combinations of these parameters also give similar results, hence those results obtained have been omitted for brevity and clarity. We used the following uni-variate datasets:

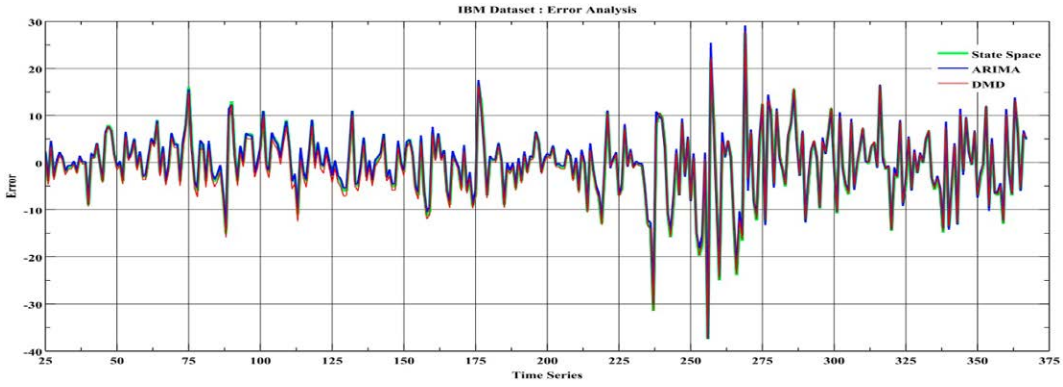

Figure 3: IBM Stock price: Modeling Error of DMD vs ARIMA vs State Space

- **IBM Stock Price**: Variations in first difference of the closing IBM stock price from 1961 to 1962 (also used in (Tsay 1988) [16]) with total 369 points.
- **Nile data** [1]: Yearly water level in Nile river from 622 AD to 1284 AD, total 663 points in this series.
- **Dow Jones Returns** 1972 **to** 1975: Daily returns of Dow Jones Industrial Average from 1972 to 1975 (also used in (Ryan 2007) [1]) with total 1093 points.
- **Average Handling Time**: Average handling time of staff members from 2009 to 2012 at a large call center

4.1.1 IBM stock price

The data set considered here is the first difference of the IBM stock closing price from May 17, 1961 to November 2, 1962 as reported by (Box, Jenkins, and Reinsel, 2008) [2]. This paper identified an ARIMA(0, 1, 1) as the best model for this series. However, they found that some evidence of possibility of inadequacy of the ARIMA(0, 1, 1) might be in part by change in variance. Therefore, this data set is extensively studied by various authors to illustrate their theory of variance change (Inclán and Tiao, 1994) [7], (Tsay, 1988) [16], and (Nhu, Martin, and Raftery, 1996) [9].

Model Accuracy Comparison: Fig. 3 presents the modeling error (measured as absolute difference in values predicted by the model and the actual value) for DMD, ARIMA and State Space modelling algorithms. It shows that DMD modelling error is very similar to the ARIMA and State Space Models. The change points detected with these models (using DVCPD, Fig. 2) are all same, demonstrating the modelling fidelity of DMD.

Change Points Detected: The change points detected with these models (using DVCPD, Fig. 2) are all same. Fig. 4 presents the results from multiple iterations in DVCPD algorithm and the corresponding change points detected. DVCPD detects 236, 279, and 181 points as the variance change points in the first iteration. Since the largest section on which variance remained stable is from data point 1 to data point 235, this section is taken as the base model. With respect to the base model the difference data is normalized to get the updated data. The plot of the normalized data is given in Fig. 4. The algorithm is then applied on the updated data, and 279^{th} point is found as the variance change point. And finally, once again on the re-normalized data the algorithm detected 180^{th} as the variance change point.

[1] http://lib.stat.cmu.edu/S/beran

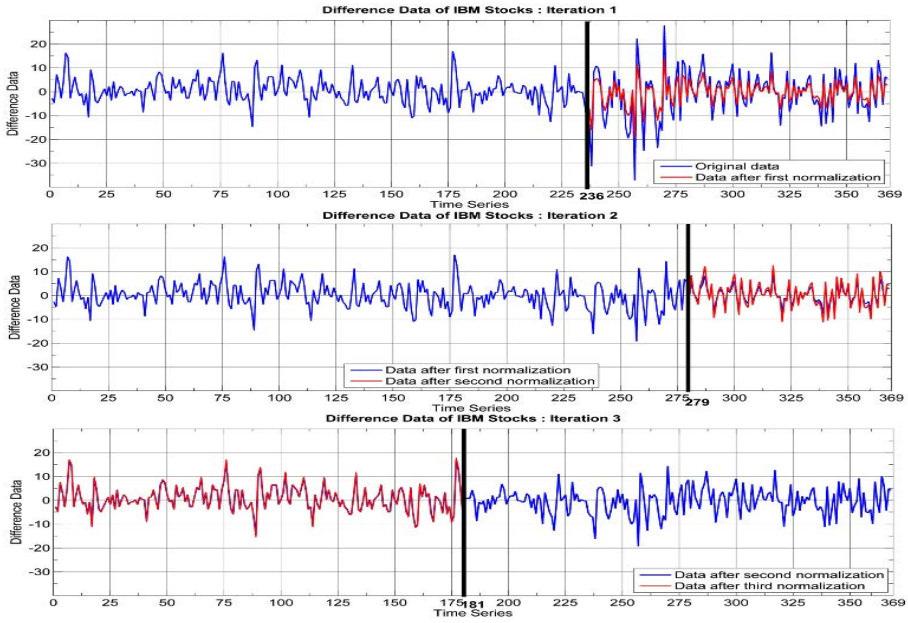

Figure 4: IBM Stock Price: Analysis. Variance change points as thick vertical lines

Superior Detection Over Prior Art: In contrast, prior algorithms are not able to detect all these change points. (Inclan and Tiao 1994) [7] found the change points as 235^{th} and the 279^{th} data points. (Tsay 1988) [16] detects variance change at only the 237^{th} point. Wichern et. al. (Wichern, Miller and Hsu, 1974) [6] finds the variance change points as 180 and 235. Thus, DVCPD detects more change points that well-known prior change point detection algorithms.

Superior Runtime: The complete DVCPD takes $49.59s$ when DMD is used for model construction (Sec. 3) while with ARIMA it takes $240s$ and with State Space it takes $50.31s$. The DMD model construction step over all iterations is faster ($1.59s$) as compared to ARIMA ($192s$) and State Space ($2.31s$) which results in faster DVCPD. This demonstrates the superior performance of our approach.

4.1.2 Nile data

We consider the classic example of change point data set, the minimum water levels of the Nile river during the AD 622-1284, measured at the island of Roda, near Cairo, Egypt. Several authors have reported evidence supporting a change point in this data around the year AD 722 (Ray and Tsay, 2002). The domain knowledge suggests the reason for this change point is the implantation of new device, nilometer, in the year AD 715, which affected the nature and accuracy of the water level measurements.

Model Accuracy Comparison: Fig. 5 presents the modeling error (measured as absolute difference in values predicted by the model and the actual value) for DMD, ARIMA and State Space algorithms. It shows that the DMD modelling error is very similar to the ARIMA and State Space Models though at some points DMD error is higher and at other points State space modelling error is higher as compared to ARIMA. The change points detected with these models (using DVCPD, Fig. 2) are very similar.

Change Points Detected: Fig. 6 presents the results from multiple iterations of DVCPD and the corresponding change points detected on Nile data. DVCPD detects 720 and 805 points as the variance change points. Since the largest section on which variance remained stable is from data point 805 onwards, this section is taken as the base model. With re-

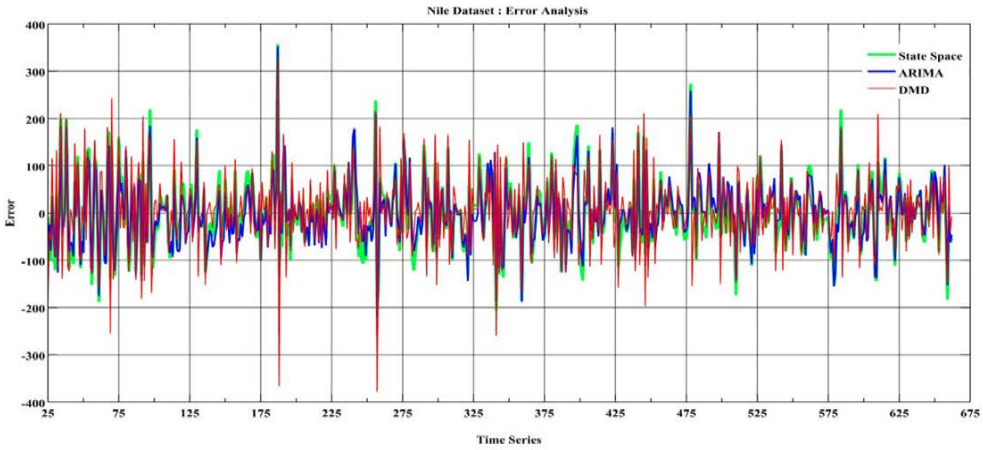

Figure 5: Nile Data: Modeling Error of DMD vs ARIMA vs State Space

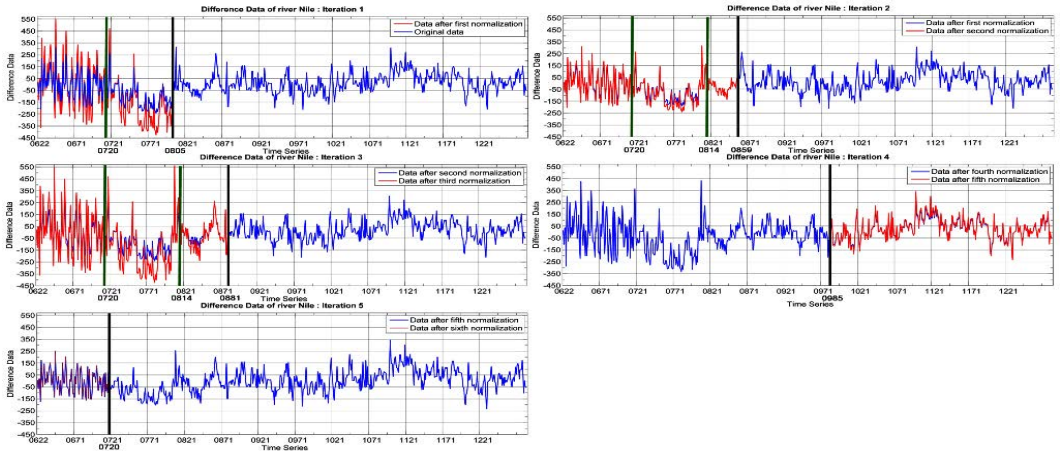

Figure 6: Nile dataset: Analysis

spect to the base model the difference data is normalized to get the updated data. The plot of the normalized data is given in Fig. 6. The algorithm is then applied on the updated data, and 881, 814 and 720 are found as the variance change points. And finally, on the re-normalized data (which starts looking like white noise) the algorithm detected 720 again as the variance change point. A change point, which is a variance change point, around AD 722 is clearly visible and agrees with previous result which uses sequential Bayesian one step ahead prediction in the presence of change points (Garnett and et. al. 2010) [4], (Saatci, Turner, and Rasmussen, 2010) [13].

Superior Detection Over Prior Art: We need to fix this paragraph. It got repetition from the IBM result! In contrast, prior algorithms are not able to detect all these change points. Inclan and Tiao found the change points as 235^{th} and the 279^{th} data points. [16] detects variance change at only the 237^{th} point. Wichern et. al. (Wichern, Miller and Hsu, 1974) [6] finds the variance change points as 180 and 235. Thus, DVCPD detects more change points that well-known prior change point detection algorithms.

Superior Runtime: The complete DVCPD takes $88.7s$ when DMD is used for model construction (Sec. 3) while with ARIMA it takes $421s$ and with State Space it takes $93.6s$. The DMD model construction step over all iterations is faster ($3.7s$) as compared to ARIMA ($336s$) and State Space ($8.61s$) which results in faster DVCPD. This demonstrates the superior performance of our approach.

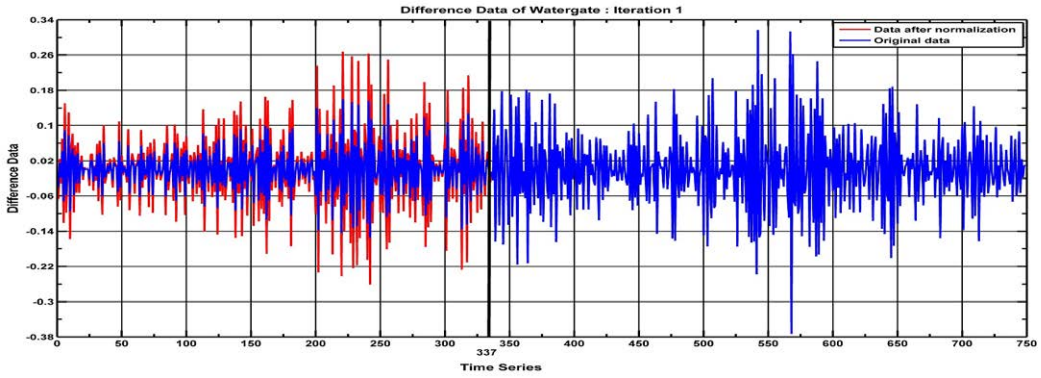

Figure 7: Dow Jones Returns: Analysis

4.1.3 Dow Jones Returns 1972 - 1975

Here, we consider the time series data of daily returns of the Dow Jones Industrial Average from July 3, 1972 to June 30, 1975. Several major events including the OPEC oil embargo occurred during this time period that had potential macroeconomic effects. The errors from DMD, ARIMA and State Space are very similar and hence lead to very similar variance change points (Figure in supplementary material). Results of multiple variance change point detection and normalization in different iterations, using DVCPD, are shown in the Fig. 7. A single variance change point is detected at point 337, which represents Oct.22, 1973. The oil embargo was put in effect in October 1973 and our algorithm is able to capture it correctly.

4.1.4 Average Handling Time

Here, we consider the time series data of daily average handling time (AHT) of staff members from 01-08-2009 to 09-30-2012 at a call center. If the staffs are newly employed they take more time to answer customer's query and their daily averaging handling time fluctuate significantly. While an experienced staff members answer each customer more efficiently and thus their daily average handling time are relatively smooth. It was expected this data would show variance change points as daily average handling time fluctuate among staff members depending upon skill and experience level of the staff, the day and the nature of the business and the customer itself. We first extracted long term trend and weekly and yearly seasons from this time series using elastic smooth season fitting algorithm (Li and Moore, 2008) [10]. Fig. 8 shows the difference between the modeled data and actual. The errors from DMD, ARIMA and State Space are very similar and hence lead to very similar variance change points using DVCPD (Fig. 2).

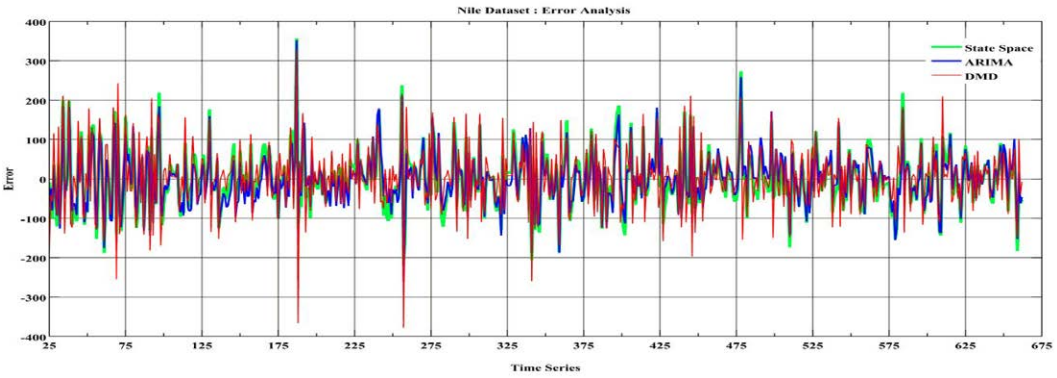

Figure 8: AHT Data: Modeling Error of DMD vs ARIMA vs State Space

Results of multiple variance change point detection and normalization in different iterations, using DVCPD, are shown in the Fig. 9. In the first iteration, the variance change points obtained are: 56, 276 and 786. In the second iteration no new change point is obtained, while in the third iteration the new change points obtained are: 355 and 983.

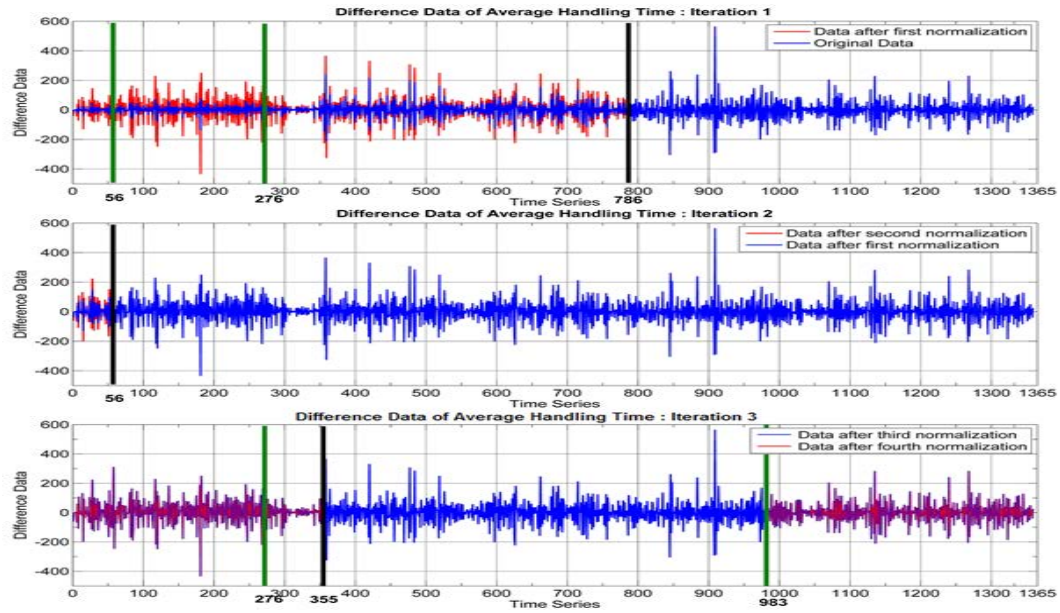

Figure 9: AHT Data: Variance change points and normalized series in different iteration on the time series of average handling time. Variance change points are marked with the thick vertical lines.

5. Conclusions & Future Work

We proposed the design DVCPD which elegantly captures the variance change locally and globally and uses normalisation to capture hidden change points. By using DMD, it provides model-free approach that makes it flexible while still being accurate. On multiple un-variate data sets, empirical results show the efficacy of the DVCPD algorithm. It detects variance change points that are similar or better than prior approaches [16] [7] and owing to DMD provides superior performance over ARIMA and State Space.

Many existing change point algorithms check each point is a change point or not, therefore they become computationally intensive if the time series is large. Further, high dimen-

sional time series make the computational performance suffer most. We noted that most variance changes do not arise instantaneously, they enter into the time series at some and their impacts get maximized some time point later. Using this insight, if it is possible to derive probability distribution of change points over time indices, then one could significantly decrease the computation time by allowing algorithm to adaptively check change points only in the regions of high probability.

References

[1] R. Adams and M. D.J.C. Bayesian online change point detection. *arXiv:0710.3742*, 2007.

[2] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: forecasting and control*. Wiley Series in Probability and Statistics, 4^{th} Edition, 2008.

[3] J. Durbin and S. J. Koopman. *Time series analysis by state state space methods*. Oxford University Press, 2^{nd} Edition, 2012.

[4] R. Garnett, M. A. Osborne, S. Reece, A. Rogers, and S. J. Roberts. Sequential bayesian prediction in the presence of changepoints and faults. *The Computer Journal*, pages 1–17, 2010.

[5] G. H. Golub and C. F. V. Loan. *Matrix computations*. The John Hopkins University Press, 4^{th} Edition, 2013.

[6] D. A. Hsu, R. B. Miller, and D. W. Wichern. On the stable paretian behavior of stock market prices. *Journal of the American Statistical Association*, Vol 69:108–113, 1974.

[7] C. Inclán and G. C. Tiao. Use of cumulative sums of squares for retrospective detection of change of variance. *Journal of the American Statistical Association*, Vol 89(427):913–923, 1994.

[8] N. Kutz. *Data-driven modelling and scientific computation: methods for complex systems and big data*. Oxford University Press, 1^{st} Edition, 2013.

[9] N. Le, R. Martin, and A. Raftery. Modeling flat stretches, burst, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, 91(436):1504–1515, 1996.

[10] J. Li and A. W. Moore. Forecasting web page views: Method and observations. *Journal of Machine Learning Research*, 9:2217–2250, 2008.

[11] L. Marti, N. Sanchez-Pi, J. M. Molina, and A. C. B. Garcia. Anomaly detection based on sensor data in petroleum industry applications. *Sensors*, 15:2774 – 2797, January 2015.

[12] C. Rowley, I. Mezic, S. Bagheri, S. Brunton, and J. N. Kutz. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

[13] Y. Saatci, R. Turner, and C. E. Rasmussen. Gaussian process change point models. *Proceedings of the* 27^{th} *International Conference on Machine Learning, Haifa, Israel*, 2010.

[14] P. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

[15] P. Schmid. Application of the dynamic mode decomposition to experimental data. *Experimental Fluids*, 50:1123–1130, 2011.

[16] R. Tsay. Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, 7:1–20, 1988.

[17] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. Generalizing dynamic mode decomposition to a larger class of datasets. *Journal of Computational Dynamics (submitted)*, 2013.

[18] Y. Wang, G. Sun, Z. Ji, C. Xing, and Y. Liang. Weighted change-point method for detecting differential gene expression in breast cancer microarray data. *PloS one*, 7(1), 2012.