Covariance Changepoint Detection for Time Series Data

Sabrina Reis
University of Oregon
Eugene, Oregon, USA
sreis@uoregon.edu

Weng-Keen Wong
Oregon State University
Corvallis, USA
wongwe@eecs.oregonstate.edu

ABSTRACT

This paper investigates the ability of the covariance changepoint detection (CCPD) method introduced by Galeano and Peña to detect and explain changes in covariance [2]. The algorithm is tested on two case studies that use time series data from air quality index (AQI) sensors that experience positive covariance shifts due to wildfire outbreaks. These experiments demonstrate that, when applied to lower-dimensional problems, the CCPD algorithm provides highly accurate and timely covariance changepoint detection. However, the high computational complexity of the algorithm demands further research to increase its efficiency. Until then, the CCPD method is only a viable alternative to existing univariate methods in lower-dimensional settings where variance changes are more effectively explained by detecting underlying changes in covariance.

CCS CONCEPTS

• **Mathematics of computing → Probabilistic algorithms**; **Multivariate statistics**.

KEYWORDS

changepoint detection, covariance, correlation, anomaly detection, time series

ACM Reference Format:
Sabrina Reis and Weng-Keen Wong. 2022. Covariance Changepoint Detection for Time Series Data. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The identification of anomalies is a fundamental part of event detection, which seeks to determine if events of interest have taken place based on patterns observed in data. Because anomalies often occur when there is a change in the relationship between data and the environment from which it emerges, anomalies commonly mark changepoints at which the distribution of the data changes in response to an outside event. Due to the connection between anomalies and events of interest, event detection may be performed by tracking these changepoints in a process known as changepoint detection. When working with time series data, or sequentially-ordered data that is collected over a period of time, changepoints signify specific timesteps when events of interest occur.

The approach used to detect changepoints in time series data varies based on problem constraints and properties of the data. The first main problem constraint that guides the development of a changepoint detection algorithm is the determination of the need for an online or offline solution. An online algorithm processes data points as they occur and attempts to detect changepoints in real time, while an offline algorithm has access to the entire time series when detecting changepoints. Another key constraint is the choice between minimizing false positives and minimizing time to detection. Certain problems demand a high degree of certainty before alerting that a changepoint has been found, as in the case of fraud detection, while others are more concerned with prompt detection, as with real-time earthquake detection. Because achieving a high degree of certainty typically requires more data to verify that an anomaly is truly a changepoint, minimizing false positives is at odds with minimizing time to detection.

Along with these two constraints imposed by the nature of the problem, certain properties of the input data, such as its dimensionality and volume, determine the characteristics of the algorithm used for changepoint detection. Dimensionality involves the number of variables tracked in the time series data; univariate methods work with time series that involve only one variable, while multivariate methods work with time series that involve two or more variables. An additional concern is the volume of the data; if expected to work efficiently on both low and high volumes of data, the changepoint detection algorithms must be scalable, meaning that it must be able to process the data efficiently regardless of the input volume. Dimensionality and data volume are often related since multivariate time series contain many more data points than univariate time series with the same number of timesteps. As a result, scalability is commonly a key concern when working with multivariate data due to its association with high data volume.

While univariate changepoint detection methods for time series data have been widely studied, multivariate methods have received less attention, as the dimensionality and volume of multivariate data introduce scalability challenges that constrain the applicability of multivariate detectors. However, the ability of univariate methods to fully explain changes in the data is limited for certain applications, as Galeano and Peña found in their 1997 paper comparing variance changepoint detection methods to their proposed covariance changepoint detection (CCPD) method, an offline multivariate detector [2]. This paper examines the utility of their proposed CCPD method in detecting and explaining shifts in multivariate data through a case study that uses air quality index (AQI) data to track wildfire activity.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

2 RELATED WORK

Much of the work on offline variance changepoint detection methods build off of an approach introduced by Incl`an and Tiao in 1994 [3]. Their method iterates over the data, constructing a running cumulative sum of the square of each data point encountered so far; the sum is then centered and normalized so that it has approximate bounds of -1 and 1 [3]. The value of the statistic at each timestep is then compared against a pre-determined threshold that is set based on the number of timesteps in the data and the desired level of confidence needed to mark the timestep as a changepoint [3]. In cases of homogeneous variance, the cumulative sum of squares statistic oscillates around 0 and does not exceed the threshold, whereas in cases of variance changes, the statistic increases past the threshold. Changepoints are identified by marking the timesteps where the statistic exceeds the threshold.

Many later papers borrowed the iterative approach featured in the Incl`an and Tiao paper. However, their cumulative sum of squares approach implicitly assumes that a change in variance would be accompanied by a change in the mean, which is not always true. The 1997 paper put forth by Chen and Gupta addressed this shortcoming with a changepoint detection method that detects variance changes even when the mean remains constant [1]. Their method uses the Schwarz Information Criterion (SIC), a heuristic used to select between models based on the amount of variation that the model is able to explain and the number of parameters used. Unexplained variation in the dependent variable results in a high SIC, meaning that a lower SIC indicates a better model. In the event of a variance change, the SIC for a model fitted to the entire time series is higher than the SIC for the subsequence of the time series preceding the changepoint, as the variance change experienced in the time series produces more unexplained variation in the dependent variable and, accordingly, a higher SIC. It follows that the subsequence that precedes the changepoint has a minimal SIC since there is no variance change and therefore minimal unexplained variation. Chen and Gupta leverage these properties of the SIC in the instance of a variance change to conclude that, if the SIC for the entire time series is significantly greater than the minimal SIC based on the desired confidence level, then the time series must experience a variance change [1]. This SIC procedure makes few assumptions about the context of the variance change within the data, making it a flexible method for identifying variance changes in a variety of contexts.

While Chen and Gupta managed to expand the applicability of variance changepoint detection algorithms with their approach, their univariate method naturally does not account for cases in which variance changes are driven by underlying changes in covariance. In these instances, a covariance changepoint detector is better suited to describing shifts in the distribution of the data than a variance detector. However, multivariate approaches to variance changepoint detection were seldom studied until Galeano and Peña proposed a covariance changepoint detection algorithm in 2005. While testing their algorithm, they found that multiple variance changes can often be "explained by a single covariance change," illustrating that a covariance detector may provide more insight than a univariate detector when attempting to explain shifts in data [2]. Given the potential for Galeano and Peña's CCPD algorithm to more effectively describe and draw conclusions about variance-related changes in data, this paper tests the performance and explanatory power of their proposed CCPD method using real-world case studies on wildfire data.

3 METHODS

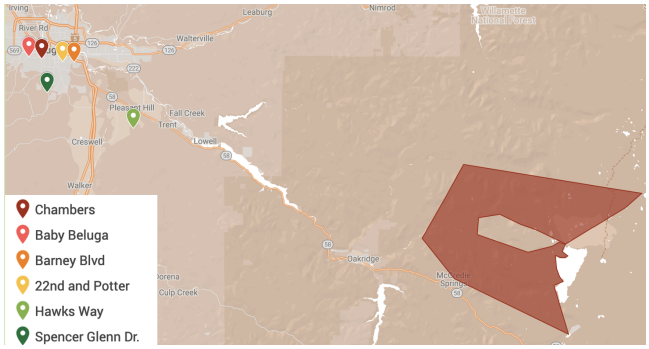

Figure 1: Locations of the sensors used in the Cedar Creek Fire case study and a polygon showing the fire perimeter.

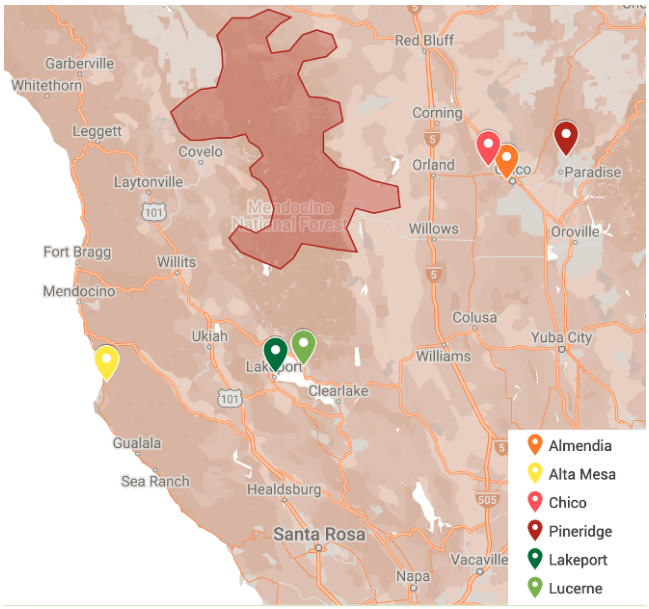

Figure 2: Locations of the sensors used in the August Complex Fire case study and a polygon showing the fire perimeter.

3.1 Covariance Detection Algorithm

To detect covariance changes, the Galeano and Peña algorithm monitors the covariance matrix, which stores the covariance values between variables, for specific changes at each timestep. Note that a covariance matrix containing observations for n variables has n^2 entries. As a result, the time and space required to compute the algorithm grows exponentially due its reliance on the covariance matrix

for changepoint detection. In the event of a positive covariance change, the covariance values between different variables increase, causing the off-diagonal entries in the covariance matrix to increase. Consequently, we may test for the existence of a changepoint by looking for evidence of an increase in the off-diagonals [2].

To assess the level of evidence for a changepoint, the algorithm performs hypothesis testing, with the null hypothesis corresponding to a lack of change in the off-diagonals and the alternative hypothesis corresponding to an increase in the off-diagonals. If a changepoint exists, it must occur at the timestep where the alternative hypothesis is most likely due to the connection between an increase in the off-diagonals and a positive covariance shift. To locate this timestep, we apply the likelihood ratio test (LRT) to the time series to compare the likelihood of the two hypotheses. The LRT statistic reaches its maximum value at the timestep where the likelihood of the alternative hypothesis is highest; because this timestep demonstrates the most evidence of a covariance change, it is subsequently marked as a changepoint.

3.2 Case Study Formation

The CCPD algorithm was then applied to two case studies that used time series data sourced from the Purple Air sensor network to capture wildfire activity [1]. The first case study used AQI data from six randomly selected sensors in the vicinity of the Cedar Creek Fire and included readings from June 2022 to October 2022. Similarly, the second case study used AQI data from six randomly selected sensors around the August Complex fire with readings from May 2020 to November 2020. The locations of sensors used in the case studies are shown in Figures 1 and 2 and the time series for the sensors are plotted in Figures 3 and 5. The goal of each case study was to test whether the algorithm could detect the positive covariance shift among the sensors as their AQI readings simultaneously increased due to the wildfire smoke.

4 RESULTS

In the Cedar Creek case study, the maximum LRT value occurred at timestep 9,702, which corresponds to September 10, 2022. According to the Lane Regional Air Monitoring Program (LRAPA), the AQI in the Eugene area exceeded the acceptable AQI threshold of 50 on September 10, 2022 [2] Based on the LRAPA data, the timestep of the maximum LRT value aligns with the real-world shift in AQI readings around Eugene. From this, we conclude that the CCPD algorithm successfully identified the changepoint using the positive covariance shift among the sensors. Underscoring the performance of the algorithm is the speed with which it picked up on the covariance shift, illustrating that the CCPD procedure minimizes time to detection without sacrificing detection accuracy.

The algorithm also delivered noteworthy results with the data from the August Complex case study, where the maximum LRT value took place at timestep 10,529 on August 18, 2020. The separate fires that merged to form the August Complex ignited on August 16 and August 17, suggesting that the maximum LRT value is consistent with the positive covariance shift caused by the increase in AQI readings across sensors following the fire outbreaks. These results further demonstrate the ability of the CCPD algorithm to detect changepoints accurately.

[1] purpleair.com
[2] https://www.lrapa.org/.

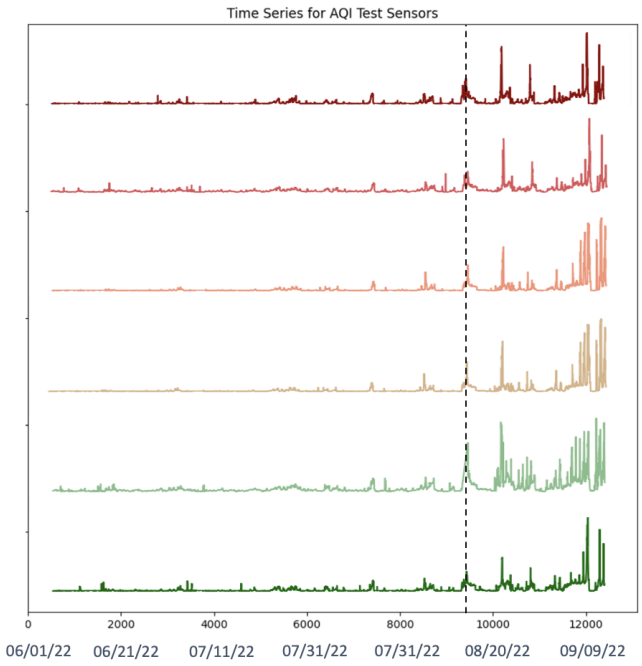

Figure 3: Time series for Cedar Creek sensors with dashed line marking the detected changepoint at timestep 9,702. Note that time series color corresponds to sensor marker color in Figure 1.

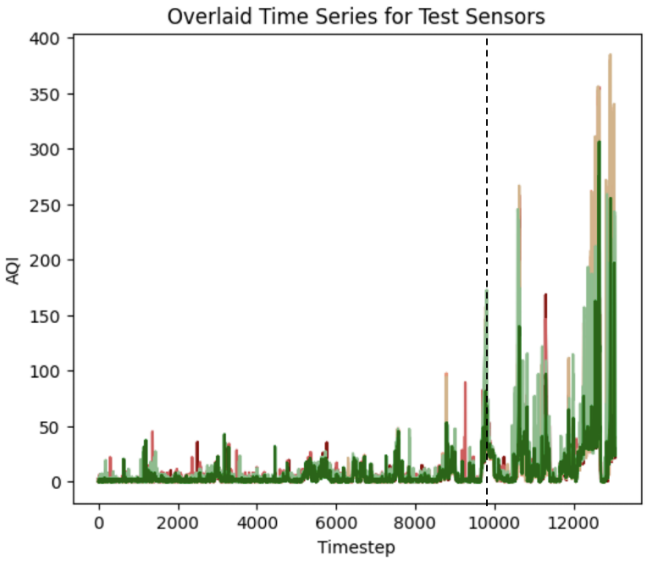

Figure 4: An alternative view of the six Cedar Creek fire sensors created by superimposing the time series.

Though the detection delay of one to two days initially implies worse performance than in the Cedar Creek case study, an examination of the time series data in Figure 5 reveals a more complicated result. We can plainly observe that the Lakeport and Lucerne sensors, shown in dark green and light green, began to spike later than the other sensors. The delay in detection may therefore be partially attributed to the time series data available, as the sensors evidently experienced a delay in picking up the fire. An additional reason for this delay lies in the design of the CCPD method. Because the detector is looking for the strongest possible evidence of a positive covariance shift through the maximum LRT value, the changepoint is marked as August 18 when the readings from every sensor signal a positive covariance change, rather than marked earlier when only a subset of the six sensors indicate a positive covariance change. By marking the changepoint as the timestep with the most evidence for a positive covariance shift, the CCPD algorithm guards against a false positive, which is appropriate for an offline algorithm. Such an outcome would not be suitable for an online algorithm, however, as the underlying logic would wrongly prioritize the minimization of false positives over reducing time to detection. For this paper's offline application, the one to two day delay represents a satisfactory level of expedience in detecting the covariance change and ensures a high degree of accuracy.

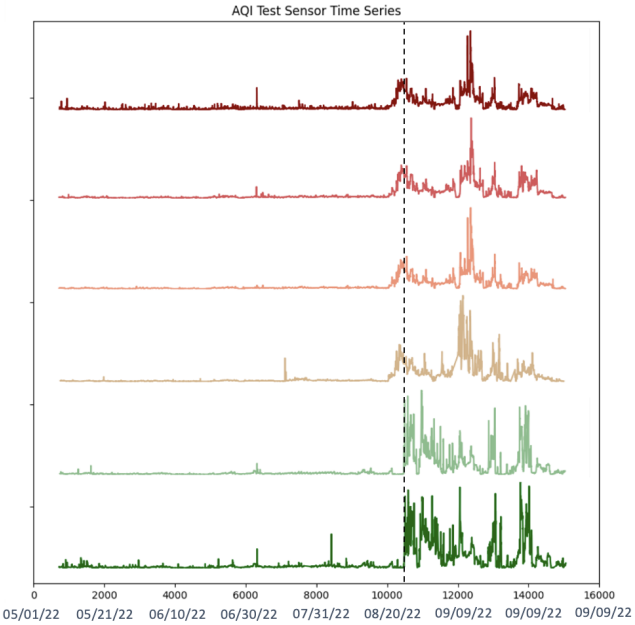

Figure 5: Time series for six August Complex sensors with dashed line marking the detected changepoint at timestep 10,529. Note that time series color corresponds to sensor marker color in Figure 2.

5 DISCUSSION AND CONCLUSION

As the results of the case studies indicate, the CCPD algorithm proposed by Galeano and Peña detects changes in covariance quickly and accurately, providing a compelling alternative to univariate

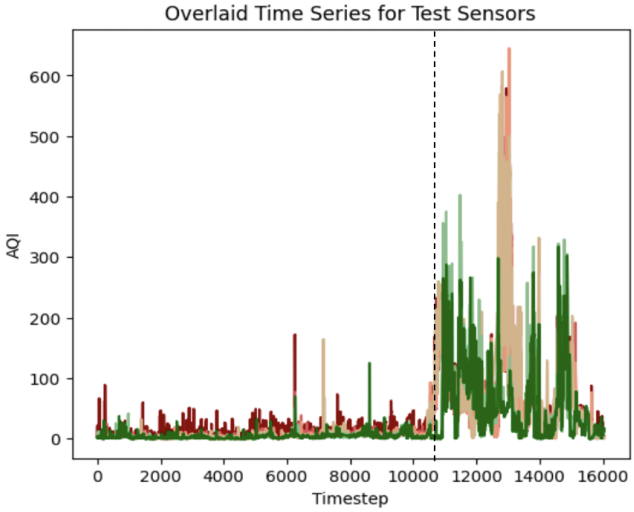

Figure 6: An alternative view of the six August Complex fire sensors created by superimposing the time series.

methods when faced with variance changes driven by underlying covariance shifts in lower-dimensional settings. However, the n^2 computational complexity of the algorithm prevents it from working well on higher-dimensional examples. More specifically, the property of the algorithm that makes it advantageous in certain settings–namely, its ability to leverage data from multiple time series to detect changepoints–does not scale. This limitation constrains the applicability of the CCPD procedure to low-dimensional settings in which variance changes are better explained through changes in covariance. To overcome this constraint, we may approximate the most expensive computations involved in monitoring changes in the covariance matrix to decrease the computational complexity of the algorithm. Literature on approximating the matrix determinant is of particular interest since the determinant is the most expensive operation in the CCPD algorithm, with a lower complexity bound that is approximately exponential. By exploring methods to reduce the time complexity of the determinant, future work may further develop the utility of CCPD algorithms as an alternative to univariate approaches.

ACKNOWLEDGMENTS

I have enormous gratitude for Weng-Keen and his patient mentorship as I completed this project. Thank you for devoting so much time and energy to guiding me as a researcher. Another massive thank you to Joe Sventek for drawing on his often-amusing career experiences to provide direction on the project and my future in research. I sincerely appreciate the opportunity to complete a capstone project and believe that I am a better researcher because of this experience.

REFERENCES

[1] Jie Chen and A. K. Gupta. 1997. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92, 438, 739–747. Retrieved Dec. 15, 2022 from http://www.jstor.org/stable/2965722.

[2] Pedro Galeano and Daniel Pena. 2005. Covariance changes detection in multivariate time series. *Journal of Statistical Planning and Inference*. DOI: 10.1016/j.jspi.2005.09.003.

[3] Carla Inclán and George C. Tiao. 1994. Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89, 427, 913–923. eprint: https://doi.org/10.1080/01621459.1994.10476824. DOI: 10.1080/01621459.1994.10476824.

Received 16 December 2022