A new method for similarity and anomaly detection in cryptocurrency markets

Nick James\textsuperscript{a}, Max Menzies\textsuperscript{b}, Jennifer Chan\textsuperscript{a}

\textsuperscript{a}School of Mathematics and Statistics, University of Sydney, NSW, Australia
\textsuperscript{b}Yau Mathematical Sciences Center, Tsinghua University, Beijing, China

Abstract

We propose a new approach using the $M_{J_1}$ semi-metric, from the more general $M_{J_p}$ class of semi-metrics \cite{1}, to detect similarity and anomalies in collections of cryptocurrencies. Since change points are signals of potential risk, we apply this metric to measure distance between change point sets, with respect to returns and variance. Such change point sets can be identified using algorithms such as the Mann-Whitney test, while the distance matrix is analysed using three approaches to detect similarity and identify clusters of similar cryptocurrencies. This aims to avoid constructing portfolios with highly similar behaviours, reducing total portfolio risk.

Keywords: Cryptocurrency, change point detection, semi-metrics, anomaly detection, volatility

1. Introduction

There is significant literature studying the variance of cryptocurrencies. Most research has focused on applying ARCH and GARCH-style models, stochastic volatility and CARR models to model the volatility of individual cryptocurrencies, such as in \cite{2, 3, 4} and others. Despite the highly non-stationary behaviour cryptocurrencies often exhibit, these models do not segment the time series into locally stationary segments. Although Phillip, et al. \cite{5} apply buffer threshold model with different forward and backward change points, they consider one cryptocurrency at a time and so they fail to compare changes across cryptocurrencies.

Hawkins \cite{6} and Hawkins et al., \cite{7, 8, 9} first introduced change point models, a developing field of algorithms designed to break time series into locally stationary segments, where each segment is governed by a separate probability distribution function. Since then, the work has been extended by several authors, most significantly Ross et al. \cite{10, 11, 12}. These methods were importantly...
consolidated by Ross [13], where software was designed to easily apply these change point algorithms to various time series. In this paper, we apply the Mann-Whitney test [13, 14] as one of the change point detection framework to detect change points in cryptocurrency log returns and variance. The inclusion of variance in change point detection is clear as the cryptocurrency market is known to be highly volatile.

James et al. [1] apply semi-metrics, a recent development of the field of metric spaces, to evaluate distances between sets of change point and determine similarity in structural breaks for a collection of time series. Metric spaces are fundamental in mathematics, and have been used to measure distances between finite sets of points for several years. Conci and Kubrusly [15] give an overview of certain such metrics and some applications, which primarily include: computational aspects, distances between fuzzy sets, and distances between images. Most pertinent to this context, they discuss the Hausdorff metric, used in many applications such as [16], [17], and its sensitivity to outliers, discussed in [18]. Then, he outlines improvements in this sensitivity that can be provided by the use of semi-metrics, which sacrifice the triangle inequality property of a metric. Semi-metrics have been used in various applications where a distance is needed, such as [19] and [20].

This paper proposes a new approach to the cryptocurrency literature, using semi-metrics, rather than classical time series models, to uncover patterns within large collections of time series. The method we apply is flexible, and may build upon any underlying criteria for change point identification. There is also flexibility to use other distance measures explored in James et al. [1] such as the MJ0.5 and MJ2 distances.

2. Methodology and data

The data we analyse is taken from Coinmarketcap. Among the 30 largest cryptocurrencies by market capitalisation, we include only those with price histories which go as far back as 01-01-2018. Twenty-two cryptocurrencies remain in our collection. We analyse the daily log returns and Parkinson range variance measures of these cryptocurrencies between 01-01-2018 and 19-11-2019, a period of 688 days. Log returns are calculated by $R_t = \log(P_t/P_{t-1})$ and the Parkinson range variance is calculated as, $\sigma^2_t = (\log(H_t) - \log(L_t))^2 / 4 \log 2$.

2.1. Generating the distance matrix

First, we apply the two phase change point detection algorithm detailed by Ross in [13] to generate sets of change points of days. This algorithm may be modified by choosing different threshold levels appropriate for the context. Then, we apply the MJ1 distance of [1] between sets of change points. The distance
measure calculates the $L^1$ norm average of all the minimum distances from a set $S$ to a set $T$ and from $T$ to $S$, that is,

$$D_1(S, T) = \frac{1}{2} \left( \frac{\sum_{t \in T} d(t, S)}{|T|} + \frac{\sum_{s \in S} d(s, T)}{|S|} \right)$$  \hspace{1cm} (1)$$

where $d(t, S)$ is the distance from a point $t$ to a set $S$.

Given $n$ time series, extract their change point sets $S_1, \ldots, S_n$. Then form the distance matrix $D$ according to $d_{ij} = D_1(S_i, S_j) \in D$.

2.2. Analysing the distance matrix

We propose three ways of analysing this matrix of distances between time series. First, one can extract, order and plot the absolute values of the eigenvalues $|\lambda_1| < \cdots < |\lambda_n|$. For both the log returns and variance time series, if many eigenvalues are relatively close to zero, relative to normalisation by the length of the total time period, one may conclude that many of the time series are highly similar.

The second approach is hierarchical clustering on the distance matrix, which allows us to see which time series are most similar based on the MJ$_1$ distance measurement. This method produces easy to interpret dendrograms, seen in Figures 1b and 2b below.

The third approach uses spectral clustering on the graph Laplacian matrix to aid in detecting groups of similarity, and anomalies. The distance matrix $D$ is transformed into an affinity matrix $A$ as follows:

$$a_{ij} = 1 - \frac{d_{ij}}{\max_{i,j} \{d_{ij} \in D\}}$$  \hspace{1cm} (2)$$

The graph Laplacian matrix is given by:

$$L = E - A,$$  \hspace{1cm} (3)$$

where $E$ is the diagonal degree matrix with $e_{ii} = \sum_j a_{ij}$.

The eigenvectors of the graph Laplacian are then clustered using a standard algorithm such as $K$-means.

3. Results and Discussion

We run two experiments, with log returns and variance for 22 cryptocurrencies. In each experiment, the Mann-Whitney test is applied to identify the number and locations of change points and the respective distance matrix is analysed with our three methods.

3.1. Similarity in log returns structural breaks

The first experiment determines the similarity between cryptocurrencies with respect to change points in the log returns series. Figure 1a demonstrates
there is a high degree of similarity in the returns of the cryptocurrency market. Approximately 16 eigenvalues are less than 200 in absolute value, small relative to the total time period of 688 days. It is clear from this plot that there are at least two outliers, with eigenvalues well outside the typical range for the time series collection.

The distance matrix’s dendrogram in Figure 1b demonstrates that there is one cluster of broad similarity, and several outliers including Tether and Monero. Within the large cluster of similar cryptocurrencies, there are sub-clusters of highly similar cryptocurrencies. There are approximately 4 sub-clusters of extreme similarity around the matrix diagonal. Both Tether and Monero are identified as dissimilar to most cryptocurrencies, however they are determined to be similar to each other. This is an important insight, highlighting that clusters of anomalous cryptocurrencies may in fact be similar to other cryptocurrencies also deemed to be dissimilar to the majority of cryptocurrencies.

The third method, spectral clustering, algorithmically shows that NEO, Chainlink, Monero and Bitcoin Cash are anomalous cryptocurrencies, with NEO and Chainlink in the same anomalous cluster. Note that anomalies may be defined differently by hierarchical and spectral clustering.

3.2. Similarity in variance structural breaks

The second experiment determines the similarity between cryptocurrencies with respect to change points in the time series’ variance.

An analysis of the distance matrix indicates more inherent similarity with respect to structural breaks in the variance than for the log returns. Figure 2a shows that all eigenvalues are less than 160, even smaller with respect to the time period of 688 days; in Figure 1a, one eigenvalue was as high as 1400. The dendrogram in Figure 2b has two key insights.

First, the scale of the dendrogram in Figure 2b is 20 times smaller than that of Figure 1b, suggesting that anomalies are far less significant when determining structural breaks in variance. Second, although the general degree
of similarity is much higher between change points in the variance, the pattern of highly similar clusters of cryptocurrencies grouping around the diagonal has disappeared. Given the 20 fold difference in scale, the proper interpretation of this is that all 22 cryptocurrencies form one cluster. Hierarchical clustering indicates essentially all 22 cryptocurrencies have very similar structural breaks in variance. Spectral clustering indicates that there are three clusters of anomalous cryptocurrencies, each of which contain one cryptocurrency. The first cluster contains Basic Attention, the second contains IOTA and the third contains Chainlink.

In Figure 3, we illustrate the distinct differences in similarity between the structural breaks of log returns and variance between two cryptocurrencies, Monero and Ripple. These were identified as highly similar with respect to variance, but reasonably dissimilar in log returns. Plots 1 and 2 in Figure 3 display a high degree of similarity in the structural breaks corresponding to variance. On the other hand, plots 3 and 4 in Figure 3 show that the log returns display significantly less similar patterns with respect to structural breaks. These results further confirm the ability of the semi-metric MJ\textsubscript{1} to identify anomalous patterns within the cryptocurrency market and provide insights for trading ideas and decision support in asset allocation decisions for portfolios of cryptocurrencies and multi-asset portfolios that contain cryptocurrencies. This analysis also highlights that the cryptocurrency market may provide exposure to highly correlated investment risk, as the variance of most cryptocurrencies exhibits pronounced synchronicity.

4. Conclusion

We propose using the MJ\textsubscript{1} semi-metric to measure similarity between 22 cryptocurrencies, according to the change points of log returns and variances. Our analysis of distance matrices suggests that the cryptocurrency market is
Figure 3: Log returns and Parkinson daily variance for Monero and Ripple. x axis represents time. Plots 1 and 2 display Parkinson daily variance measure with change points for Monero and Ripple; plots 3 and 4 display log returns with change points for Monero and Ripple. The structural breaks are more frequent, and more similar in the variance time series.
highly similar with respect to structural breaks in both log returns and variances, but anomalous behaviour is more profound in log returns than variances. Sub-clusterings according to change points of cryptocurrency returns suggest that members within each sub-cluster should not appear together in a portfolio to minimise risk. Identifying these clusters may also provide investment opportunities within the cryptocurrency market; in particular, analysing clusters holistically may provide pairs trading opportunities. Alternatively, change points of cryptocurrency variance form essentially one cluster that covers the entire market. The extraordinary similarity of the variance among the cryptocurrency market agrees with previous findings that the cryptocurrency market is highly volatile and risky.

References

[1] N. James, M. Menzies, L. Azizi, and J. Chan. Novel semi-metrics for multivariate change point analysis and anomaly detection. arXiv e-prints, page arXiv:1911.00995, 2019.

[2] P. Katsiampa. Volatility estimation for Bitcoin: A comparison of GARCH models. Economics Letters, 158:3–6, 2017.

[3] J. Chu, S. Chen, S. Nadarajah, and J. Osterrieder. GARCH modelling of cryptocurrencies. Journal of Risk and Financial Management, 4:10:17, 2017.

[4] S. K. Tan, J. S. K. Chan, and K. H. Ng. On the speculative nature of cryptocurrencies: a study on realised volatility. Finance Research Letters - to appear, 2018.

[5] A. Phillip, J. S. K. Chan, and M. S. Peiris. On long memory effects in the volatility measure of cryptocurrencies. Finance Research Letters - to appear, 2019.

[6] D. M. Hawkins. Testing a sequence of observations for a shift in location. Journal of the American Statistical Association, 72:180–186, 1977.

[7] D. M. Hawkins, P. H. Qiu, and C. W. Kang. A change-point model for a shift in variance. Journal of Quality Technology, 42:355–366, 2003.

[8] D. M. Hawkins and K. D. Zamba. Statistical process control for shifts in mean or variance using a changepoint formulation. Technometrics, 47:164–173, 2005.

[9] D. M. Hawkins and Q. Deng. A nonparametric change-point control chart. Journal of Quality Technology, 42:165–173, 2010.

[10] G. J. Ross and N. M. Adams. Sequential monitoring of a Bernoulli sequence: when the pre-change parameter is unknown. Computational Statistics, 28:463–479, 2011.
[11] G. J. Ross and N. M. Adams. Two nonparametric control charts for detecting arbitrary distribution changes. *Journal of Quality Technology*, 44:102–116, 2012.

[12] G. J. Ross, D. K. Tasoulis, and N. M. Adams. Nonparametric monitoring of data streams for changes in location and scale. *Technometrics*, 53:379–389, 2011.

[13] G. J. Ross. Parametric and nonparametric sequential change detection in R: the cpm package. *Journal of Statistical Software, Articles*, 66(3):1–20, 2015.

[14] A. N. Pettitt. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society. Series C*, 28:126–135, 1979.

[15] A. Conci and C. Kubrusly. Distances between sets - a survey. *Advances in Mathematical Sciences and Applications*, 26:1–18, 2017.

[16] M. J. Atallah, C. C. Ribeiro, and S. Lifischitz. Computing some distance functions between polygons. *Pattern Recognition*, 24:775–781, 1991.

[17] M. Barton, I. Hanniel, G. Elber, and M. S. Kim. Precise Hausdorff distance computation between polygonal meshes. *Comput. Aided Geom. Design*, 27:580–591, 2010.

[18] A. J. Baddeley. Errors in binary images and an $L^p$ version of the Hausdorff metric. *Nieuw Arch. Wisk*, 10:157–183, 1992.

[19] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: image retrieval and class representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:583–600, 2000.

[20] W. J. Rucklidge. Efficient visual recognition using Hausdorff distance. *Springer*, 1996.