

Time and Again:

Time Series Mining via Recurrence Quantification Analysis

Stephan Spiegel^{1(✉)} and Norbert Marwan²

¹ DAI Lab, Berlin Institute of Technology, Berlin, Germany
spiegel@dai-lab.de

² Potsdam Institute for Climate Impact Research, Potsdam, Germany
marwan@pik-potsdam.de

Abstract. Recurrence quantification analysis (RQA) was developed in order to quantify differently appearing recurrence plots (RPs) based on their small-scale structures, which generally indicate the number and duration of recurrences in a dynamical system. Although RQA measures are traditionally employed in analyzing complex systems and identifying transitions, recent work has shown that they can also be used for pairwise dissimilarity comparisons of time series. We explain why RQA is not only a modern method for nonlinear data analysis but also is a very promising technique for various time series mining tasks.

Keywords: Time series mining · Recurrence quantification analysis

1 Introduction and Background

A recurrence plot (RP) is an advanced technique of nonlinear data analysis [3]. Technically speaking, a recurrence plot R visualizes those times when the trajectory x of a dynamical system visits roughly the same phase space [3]: $R_{i,j} = \Theta(\epsilon - \|x_i - x_j\|)$, where ϵ is the similarity threshold, $\|\cdot\|$ a norm, $\Theta(\cdot)$ the unit step function, and $i, j = 1 \dots N$ is the number of states. In addition, a cross recurrence plot (CRP) shows all those times at which a state $x_i \in \mathbb{R}^m$ in one dynamical system co-occurs $y_j \in \mathbb{R}^m$ in a second dynamical system [3]: $R_{i,j} = \Theta(\epsilon - \|x_i - y_j\|)$, where the dimension m of both systems must be the same, but the number of states can be different.

The recurrence quantification analysis (RQA) is a method of nonlinear data analysis which quantifies the number and duration of recurrences of a dynamical system presented by its state space trajectory [3]. RQA measures are derived from RP structures and can be employed to study the dynamics, transitions, or synchronization of complex systems [3, 4]. The determinism measure (DET^μ), which is the fraction of recurrence points that form diagonal lines of minimum length μ , has e.g. been successfully applied to detect dynamical transitions [4].

2 Recent Trends and Advances

In time series mining, many algorithms are based on analogical reasoning or pairwise dissimilarity comparisons of (sub)sequences [13]. In general, the distance

between time series needs to be carefully defined in order to reflect the underlying dissimilarity of the data, where the choice of distance measure usually depends on the invariance required by the domain [1].

Recent work [9–12] has introduced novel time series distance measures that use recurrence quantification analysis (RQA) techniques. The main idea [9] is to pairwise compare time series by (i) computing a cross recurrence plot (CRP) that reveals all times at which roughly the same states co-occur and, subsequently, (ii) quantifying the number and length of all diagonal line structures that indicate similar subsequences. Figure 1(a-b) shows a toy example, where a labeled time series is compared to two unlabeled data stream segments using CRPs as well as corresponding RQA measures.

It has been shown [9, 11] that traditional RQA measures, such as the average diagonal line length and the determinism, can be used to compare time series that exhibit similar segments or subsequences at arbitrary positions. Time series with such an order invariance [9] can, for instance, be found in automotive engineering [11], where vehicular sensors observe driving behavior patterns in their natural occurring order and the recorded car drives are compared according to the co-occurrence of these patterns. Although the recurrence plot-based distance [11] was originally developed to determine characteristic driving profiles [12], this approach can be used to find representatives in arbitrary sets of single- or multi-dimensional time series of variable length [10].

In addition, it has been proposed to employ video compression algorithms for measuring the dissimilarity between un-thresholded recurrence plots and accordingly the time series that generated them [8]. This approach relies on the underlying assumption that video compression algorithms are able to detect similar structures in images or recurrence plots, which correspond to time series patterns. The result [8] show that the compression distance of recurrence plots works especially well for time series that represent shapes. A follow-up study [5] compared the performance of various MPEG video compression algorithms and furthermore introduced a compression distance for cross recurrence plots. Figure 1(c) contrasts two un-thresholded recurrence plots, which reveal structural dissimilarities between the examined time series.

Although recurrence plots have been adopted by the data mining community [2, 5, 8–12], their computation and quantification generally involve operations with quadratic time and space complexity. Hence, recent work [7, 14] has introduced approximate RQA measures, which exhibit significantly lower complexity while maintaining high accuracy. Most important, these novel approximations [7, 14] enable us to efficiently use recurrence quantification analysis for relatively long time series and fast time series streams. Figure 1(d) illustrates the fast computation of the approximate determinism ($aDET$) [7], which allows us, for example, to filter or identify time series segments with a certain behavior in an online fashion. The approximation of various RQA measures, such as laminarity and determinism, is explained at full length in a recent publication [14]

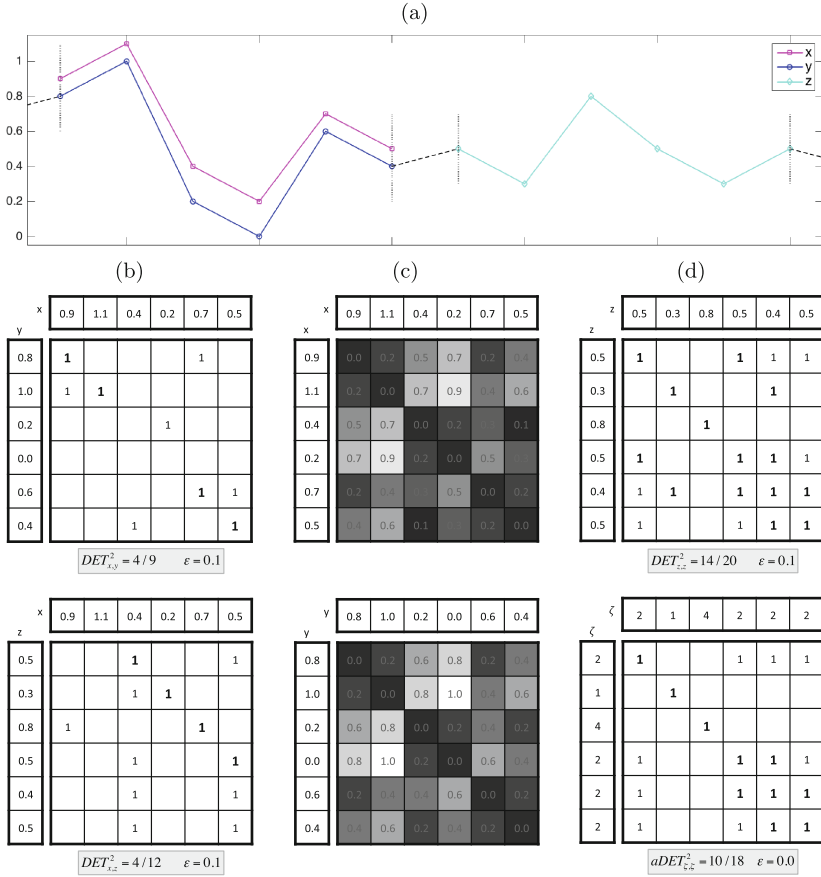


Fig. 1. Recurrence plot-based distances: (a) illustrates a time series mining scenario that assumes a labeled sequence x and a data stream with unlabeled segments y and z . In case (b) we compare time series x with segment y and z to assign labels. (b) shows two cross recurrence plots that indicate similar states ($\epsilon = 0.1$) for time series pairs (x, y) and (x, z) , where recurrence points are represented by ‘1’ entries and diagonal line structures are highlighted in bold font. According to the determinism, $DET_{x,y}^2 = 4/9 > 4/12 = DET_{x,z}^2$, the pair (x, y) is more similar than (x, z) [11], meaning x and y might be from the same class. (c) shows another way to determine the pairwise dissimilarity of time series. In this case (c) we create un-thresholded recurrence plots ($\epsilon = 0$), which facilitate pairwise comparisons by means of image processing and video compression algorithms [5, 8]. The images in (c) resemble each other in structure since time series x and y have a similar shape. In case (d) we compute the approximate determinism to assess the ‘complexity’ of our sample data stream at time interval z and to filter/identify ‘ir-relevant’ segments with a certain (nonlinear) behavior. (d) illustrates the recurrence plot of segment z and its discretized version $\zeta = \lfloor \frac{z}{2\epsilon} \rfloor$. In our example (d) we achieve a fairly reasonable approximation of the determinism, $DET_{z,z}^2 = 14/20 \approx 10/18 = aDET_{\zeta,\zeta}^2$. Although the discretization step introduces some rounding errors, it allows us to approximate all traditional RQA measures in an efficient way without even creating and quantifying the RP [7, 14].

3 Conclusion and Open Problems

Recurrence quantification analysis (RQA) is a method of nonlinear data analysis for the investigation of dynamical systems, which has its origin in theoretical physics [3,4]. Recently, RQA was adopted by the data mining community in order to: (i) define novel time series distance measures [5,8,11] and (ii) process massive data streams by means of approximate measures [7,14].

Although RQA has been successfully applied to data mining problems from engineering [12] and climatology [6,14], there exist open problems which prevent its widespread acceptance by the time series fraternity. The main problem with traditional RQA is that it excludes curved structures, which prevents us from comparing time series with local scaling or warping invariance [1]. This issue might be addressed by feeding un-thresholded RPs [5,8] into convolutional neural networks. In the case of the recently introduced approximate RQA [7,14], it is necessary to investigate time series representations and discretization techniques that enable us to bound the approximation error.

References

1. Batista, G., Keogh, E.J., Tataw, O.M., De Souza, V.M.A.: CID: an efficient complexity-invariant distance for time series. *Data Min. Knowl. Disc.* **28**, 634–669 (2014)
2. Gaebler, J., Spiegel, S., Albayrak, S.: MatArcs: an exploratory data analysis of recurring patterns in multivariate time series. In: *Proceedings of ECML-PKDD* (2012)
3. Marwan, N., Romano, M.C., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. *Phys. Rep.* **438**, 237–329 (2007)
4. Marwan, N., Schinkel, S., Kurths, J.: Recurrence plots 25 years later - gaining confidence in dynamical transitions. *Europhys. Lett.* **101**, 20007 (2013)
5. Michael, T., Spiegel, S., Albayrak, S.: Time series classification using compressed recurrence plots. In: *Proceedings of ECML-PKDD* (2015)
6. Rawald, T., Sips, M., Marwan, N., Dransch, D.: Fast computation of recurrences in long time series. In: Marwan, N., Riley, M., Giuliani, A., Webber Jr., C.L. (eds.) *Translational Recurrences. Springer Proceedings in Mathematics and Statistics*, pp. 17–29. Springer, Switzerland (2014)
7. Schultz, D., Spiegel, S., Marwan, N., Albayrak, S.: Approximation of diagonal line based measures in recurrence quantification analysis. *Phys. Lett. A* **379**, 997–1011 (2015)
8. Silva, D.F., De Souza, V.M.A., Batista, G.: Time series classification using compression distance of recurrence plots. In: *Proceedings of ICDM* (2013)
9. Spiegel, S., Albayrak, S.: An order-invariant time series distance measure - position on recent developments in time series analysis. In: *Proceedings of KDIR* (2012)
10. Spiegel, S., Schultz, D., Albayrak, S.: BestTime: finding representatives in time series datasets. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) *ECML PKDD 2014, Part III. LNCS*, vol. 8726, pp. 477–480. Springer, Heidelberg (2014)
11. Spiegel, S., Jain, B.J., Albayrak, S.: A recurrence plot-based distance measures. In: Marwan, N., Riley, M., Giuliani, A., Webber Jr., C.L. (eds.) *Translational Recurrences. Springer Proceedings in Mathematics and Statistics*, pp. 1–15. Springer, Switzerland (2014)

12. Spiegel, S.: Discovery of driving behavior patterns. In: Hopfgartner, F. (ed.) *Smart Information Services - Computational Intelligence for Real-Life Applications*, pp. 315–343. Springer, Switzerland (2015)
13. Spiegel, S.: Time series distance measures: segmentation, classification and clustering of temporal data. Technische Universitaet Berlin (2015)
14. Spiegel, S., Schultz, D., Marwan, N.: Approximate recurrence quantification analysis in best code of practice. In: Webber Jr., C.L., Ioana, C., Marwan, N. (eds.) *Recurrence Plots and Their Quantifications: Expanding Horizons. Springer Proceedings in Physics*, pp. 113–136. Springer, Switzerland (2016)