AN APPROACH TO ANALYZE IRREGULARLY SAMPLED TIME SERIES WITHOUT INTERPOLATION

Deniz Eroglu
[1]Potsdam Institute for Climate Impact Research (PIK)
14473 Potsdam, Germany
[2]Department of Physics, Humboldt University
12489 Berlin Germany
eroglu@pik-potsdam.de

Jürgen Kurths
[1]Potsdam Institute for Climate Impact Research (PIK)
14473 Potsdam, Germany
[2]Department of Physics, Humboldt University
12489 Berlin Germany
[3]Institute for Complex Systems and Mathematical Biology
University of Aberdeen
Aberdeen AB24 3UE, United Kingdom
kurths@pik-potsdam.de

Abstract

In time series analysis, one of the challenges is irregularly sampled data sets, that is: their time resolution is not constant and may contain large variations of the time span between two consecutive points. Most time series analysis methods require regular sampling. Traditionally, some form of interpolation is used to deal with these irregularities. But such interpolation introduces additional information into the time series with much higher uncertainty than the real observations. Here we present a novel technique called 'the TrAnsformation Cost Time Series method (TACTS)' to deal with irregularity difficulties without degenerating the quality of the data set.

Key words

irregularly sampled time series, metric distance, recurrence plot

1 Introduction

In time series analyses, an important issue is to calculate the distance between two data patterns. This distance problem frequently occurs for example in recurrence plots (RP) [Marwan et al., 2007], the estimation of the maximum Lyapunov exponent [Michael T. Rosenstein and Luca, 1993], scale-dependent correlations [Rodó and Rodriguez-Arias, 2006], data classification, dynamic time [Sakoe and Chiba, 1978] or correlation dimension estimations [Grassberger, 1983]. When doing these kinds of analyses and the data has an equidistant time resolution the Euclidean distance is often used. However, in applications that generate data with non-regular sampling such an approach is not applicable. Such data sets include almost all paleo climate observations which can have a very erratic almost random time resolution. One way to deal with such data sets is to interpolate them. Such an interpolation will not only fill the gaps but replace real measurements with new interpolated data points close by that have regular sampling. But this is not the optimum method since the subsequent analysis will be biased [Rehfeld et al., 2012]. Moreover these interpolated values have a higher uncertainty than the measured data points they replace.

Our new approach is based on a measure introduced by Victor and Purpura [Victor and Purpura, 1997]. For the dynamics of firing neurons, Victor and Purpura [Victor and Purpura, 1997] showed that the spike time distance is a useful method that applies to irregularly sampled data sets. The basic idea of this method is a distance metric that provides information of how easily one data segment can be transformed into another data segment. To transform one segment of data into another segment three elementary operations are required: adding or deleting of a data point and moving the data point to a different time. Using associated costs for the elementary operations an optimal data transformation will be achieved if the cost of the transformation is minimised.

2 Metric distance

Consider the metric D as a mapping of two pairs of spike trains or data segments, say S_a and S_b, onto a real value. In order for D to be metric, it must satisfy three conditions since D is a generalised distance. These three conditions are;

$D(S_a, S_b) \geq 0$ (must be positive)
$D(S_a, S_b) = D(S_b, S_a)$ (must be symmetric)
$D(S_a, S_c) \leq D(S_a, S_b) + D(S_b, S_c)$ (satisfy the triangle inequality)[Victor and Purpura, 1997]

Suzuki *et al.* [Suzuki et al., 2010] have extended this method to be used for continuous marked data and we follow their approach. The continuous data set is transformed to an event time series that is very similar to spike trains, and allows events to have different amplitudes. We define the cost function in the following way:

$$\begin{aligned} p(c) = & \sum_{(\alpha,\beta)\in C}\{\lambda_0|t_a(\alpha)-t_b(\beta)| && (1)\\ & + \tfrac{1}{m}\sum_{k=1}^{m}\lambda_k|L_{a,k}(\alpha)-L_{b,k}(\beta)|\} && (2)\\ & + \lambda_S(|I|+|J|-2|C|), && (3)\end{aligned}$$

where I and J are a set of indices of the events in S_a and S_b, respectively. Note that the summation is over the pairs $(\alpha, \beta) \in C$, where C is the set of points that will be shifted in time. α and β are the αth event in S_a and βth event in S_b. The first term with the coefficient λ_0 is the cost to shift some event in time [Suzuki et al., 2010]. The second summation involves the difference $|L_{a,k}(m) - L_{b,k}(n)|$, where $L_{a,k}$ is the amplitude of the kth event in S_a. Therefore the parameter λ_k has the unit of amplitude^{-1} and the sum is over the different components of the amplitude. That is, if we are dealing with one dimensional data $m = 1$ while for a three dimensional phase space m would be three. The last terms in the cost function deal with the events not in C which have to be added or deleted. Note that $|\cdot|$ denotes the cardinality of the set and λ_S is the cost parameter for this operation. Suzuki *et al.* omitted this parameter, since they chose a cost of one for such an operation [Suzuki et al., 2010].

As we can see minimisation of the cost function, eq. (3), will depend on the choice of $\lambda_{0,k,S}$. The work of Suzuki *et al.* and Victor & Purpura [Victor and Purpura, 1997] shows that $\lambda_{0,k}$ can be optimised to give excellent results in many cases. Instead of trying to optimise $\lambda_{0,k}$ for each time series we suggest two possible new ways to choose these values and explore one of them in detail.

Note that in eq. (3) the cost to delete or add a new data point is equal to λ_S for each of these operations. On the other hand shifting and changing the amplitude is proportional to the difference in time and amplitude. In order to determine whether the deleting and adding operations are preferred, a natural choice for $\lambda_{0,k}$ should be such that every time shifting and changing cost more than $2\lambda_S$ (total cost of adding and deleting a point). Consequently $\lambda_{0,k}$ should be chosen in such a way that in average both terms give a contribution of λ_S. Therefore:

$$\lambda_0 = \frac{M}{\text{total time}} \tag{4a}$$

$$\lambda_k = \frac{M-1}{\sum_i^{M-1}|L_{a,k}-L_{b,k}|}, \tag{4b}$$

where M is the total number of events in the time series. As we can see our first possibility chooses λ_0 as the mean event frequency and λ_k as the inverse of the average amplitude difference.

The second option we are focussing on also uses $\lambda_{0,k}$ according to eqs. (4) but optimises the cost of the deleting and adding operations. That is, while $\lambda_{0,k}$ are fixed, we explore the costs for deleting and adding in the range from $\lambda_S \in [0, 4]$. If our time series consist of $n + 1$ segments of equal length, we therefore can calculate n costs for each individual transformation of the segments. Assuming that the costs are linearly independent, the central limit theorem indicates that the distribution of the costs should be a normal distribution. Especially when dealing with non-stationary data, changing λ_S such that the distribution becomes normal greatly improves the skill of our time series analysis method.

3 Recurrence plot

Recurrence plots visualise a fundamental property of dynamical systems – namely, when the system 'repeats' itself, returning to a previous state. Formally, for a set of observations $\overrightarrow{x_i}$ for $i = 1, \ldots, N$ this is defined as

$$R_{i,j}(\epsilon) = \Theta(\epsilon - ||\overrightarrow{x_i} - \overrightarrow{x_j}||)\ i, j = 1, \ldots, N \tag{5}$$

where ϵ is some threshold distance, $||\cdot||$ is some distance measure, and $\Theta(x) = 1$ if $x \geq 0$ and 0 otherwise [Marwan et al., 2007]. This method is well-suited to capture regime changes, as such an event would result in a sudden reduction in the number of recurring events. Plotting this matrix allows visual analysis of the system, and from this quantitative measures can be derived. Diagonal structures within the plot, running parallel to the main diagonal (bottom left to top right), indicate sections of the trajectory with locally similar paths. Calculating the ratio of 'long' diagonal lengths relative to the total number of diagonals gives us the measure *determinism*. This is a measure of the amount of predictability within the system, as stochastic or chaotic systems result in short diagonals. For the selected threshold distance ϵ, a histogram of diagonal lengths, $P(\epsilon, l)$, is derived and a threshold diagonal length, l_{min}, is set. Determinism is then given by

$$DET = \frac{\Sigma_{l=l_{min}}^{N} lP(\epsilon, l)}{\Sigma_{l=1}^{N} lP(\epsilon, l)} \tag{6}$$

In this analysis, the recurrence plot is derived using a Euclidean distance norm and ϵ-threshold distance is chosen adaptively to ensure a sensible density of 'ones' in the RP, fixed at 10%. In the determinism calculation, we take l_{min} to be 2. These parameters were selected to ensure a balance between stability and, particularly in the case of the threshold distance, the inclusion of

enough data points for the recurrence structure of the underlying system to be captured. The statistical significance of the DET measure can be determined using a bootstrapping method as outlined in [Marwan et al., 2013]. For details on the embedding required to transform time series data into a trajectory in phase space, see [Kantz and Schreiber, 1997].

4 Application

While recurrence plot based quantification is not directly applicable to irregularly sampled time series, we show that a recurrence plot analysis based on the transformation cost time series can be used to identify the changes in the dynamics of the system.

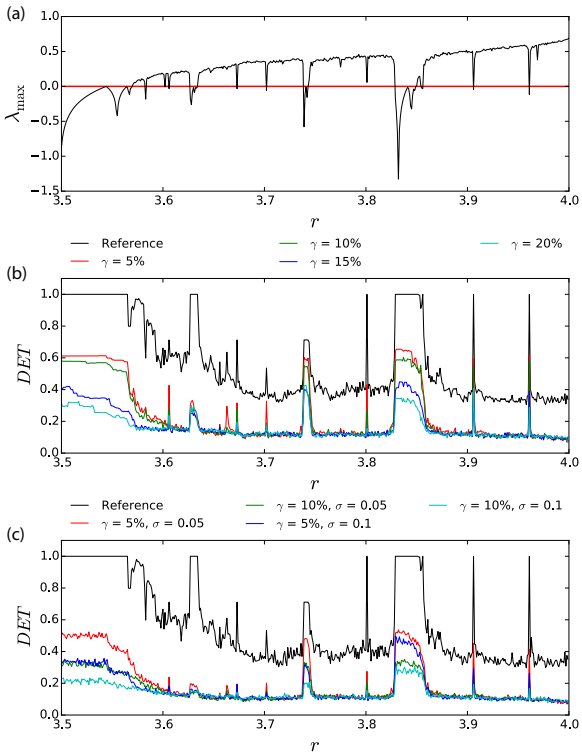

Figure 1. Recurrence quantification analyses for Logistic Map: (a) Lyapunov exponent of the Logistic Map; (b) - (c) Determinism calculated from the transformation cost time series for (b) various levels of deleting; and (c) two measurement noise levels ($\sigma = 0.05$ and 0.1) and for two different rates of irregularity ($\gamma = 5\%$ and 10%). For details see text as well as the legends in the figures.

In Fig. 1 we present the results of our investigation on the Logistic map. Panel (a) shows the Lyapunov exponent calculated from the time series. In panel (b) the determinism calculated for the time series is shown for increasingly irregular sampling. Comparison with the Lyapunov exponent shows that the determinism tracks the transitions of the dynamics for all data sets. We can clearly see that randomly deleting points leads to a distinct drop in the determinism. While this drop is most pronounced when comparing the reference line ($\gamma = 0\%$) with $\gamma = 5\%$ for higher γ values the determinism does not decrease as much. In Fig. 1 panel (c), we show that our method cannot only detect changes in the dynamics for irregular sampling but is also stable if the data is additionally compromised by measurement noise. We added Gaussian white noise ($\langle\xi\rangle = 0$ and $\langle\xi(t)\xi(t')\rangle = \delta(t-t')$) and the results shown are for a noise level of $\sigma = 0, 0.05$ and 0.1 ($\gamma = 5\%$ and 10%). Note that the σ is scaled relative to the variance of the time series. Again we clearly see that for all noise levels considered, our method is able to identify the changes in the dynamics and closely follows the Lyapunov exponent.

5 Conclusion

Our new method have demonstrated that TACTS is an useful technique even for extreme irregularities in the resolution of time series and in addition can deal under affect of high measurement noise. Promising results have shown in simple toy model. In [Ozken et al.,] and [Eroglu et al.,] we further show that our new method gives good results with time series created from the Rössler system and also higher dimensional complex paleo climate data sets.

Acknowledgements

We thank Ibrahim Ozken, Fiona McRobie, Thomas Stemler, Norbert Marwan and Karl-Heinz Wyrwoll for fruitful discussions and collaborations. DE acknowledges support by the Leibniz Association (WGL) under Grant No. SAW-2013-IZW-2

References

Eroglu, D., McRobie, F., Ozken, I., Stemler, T., Wyrwoll, K.-H., Marwan, N., and Kurths, J. A solar variability driven monsoon seesaw: switching relationships of the holocene east asian - australian summer monsoon. *(arXiv): unpublished*.

Grassberger, P. (1983). Generalized dimensions of strange attractors. *Physics Letters A*, 97(6):227–230.

Kantz, H. and Schreiber, T. (1997). *Nonlinear Time Series Analysis*. University Press, Cambridge.

Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. (2007). Recurrence Plots for the Analysis of Complex Systems. *Physics Reports*, 438(5–6):237–329.

Marwan, N., Schinkel, S., and Kurths, J. (2013). Recurrence plots 25 years later – gaining confidence in dynamical transitions. *Europhysics Letters*, 101:20007.

Michael T. Rosenstein, J. J. C. and Luca, C. J. D. (1993). A practical method for calculating largest lyapunov exponents from small data sets. *Physica D*, 65(3):117–134.

Ozken, I., Eroglu, D., Stemler, T., Marwan, N., Bagci, G. B., and Kurths, J. The transformation cost time series method for analysing irregularly sampled data. *Physical Review E: in press*.

Rehfeld, K., Marwan, N., Breitenbach, S. F. M., and Kurths, J. (2012). Late Holocene Asian summer monsoon dynamics from small but complex networks of paleoclimate data. *Climate Dynamics*, 41(1):3–19.

Rodó, X. and Rodriguez-Arias, M.-A. (2006). A new method to detect transitory signatures and local time/space variability structures in the climate system: the scale-dependent correlation analysis. *Climate Dynamics*, 27(5):441–458.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Suzuki, S., Hirata, Y., and Aihara, K. (2010). Definition of distance for marked point process data and its application to recurrence plot-based analysis of exchange tick data of foreign currencies. *International Journal of Bifurcation and Chaos*, 20(11):3699–3708.

Victor, J. D. and Purpura, K. P. (1997). Metric-space analysis of spike trains: theory, algorithms and application. *Network: Computation in Neural Systems*, 8(2):127–164.