Data Imputation – Papers analysis

Eduarda T. C. Chagas

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte – Minas Gerais – Brazil

eduarda.chagas@dcc.ufmg.br

Comparing different approaches to compute Permutation Entropy with coarse time series [Traversaro et al. 2019]

Was test the new Data-Driven Method of Imputation that cope with this type of time series without modifying the essence of the Bandt and Pompe Probability Distribution Function and compare it with the Modified Permutation Entropy, a complexity measure that assumes that equal values are not from artifacts of observations but they are typical of the data generating process. The Data-Driven Method of Imputation proves to outperform the Modified Permutation Entropy.

Numerical Analysis

The Modified Permutation Entropy (MPE) and the Data Driven Method of Imputation (DDMI) are evaluated using data from simulated chaotic processes. The simulation consists in 44 time series generated by different chaotic processes in every dimension, previously analyzed by Rosso and coworkers in Rosso et al. 2013. Each original time series was truncated up to two decimal resolution, leading to a coarse version of those original time series. This yields to quantify how well each strategy estimates the actual PE – $H(\boldsymbol{D})$ – for the different chaotic processes using the PE estimation – $\hat{H}(\boldsymbol{D})$ – obtained for each methodology.

Was used the initial conditions present in [Sprott and Sprott 2003] and refer to this book for a comprehensive analysis of those maps. Each map was simulated for a length of 10^6 and then starting in the position 10^5. They were produced with IEEE 754 double precision floating point numbers. For those initial condition, as the series is deterministic, the $H(\boldsymbol{D})$, $D = \{3, 4, 5, 6\}$ are the original PE of the processes.

For every length motiv D the DDMI estimation is better that the MPE, and the variability of the Error of the DDMI is considerably lower than the MPE. The improvement in the estimation made by DDMI over the methodology of eliminating this patterns with ties can be seen as the gain of the information given by this methodology. Another interesting property to notice is how well each methodology performs for different amounts of patterns $X_D(t)$ with ties. As might be expected, the error increases with the repeated ratio, but the DDMI is always better than the MPE. It should be noted that for the highest level of repeated ratio, at least $20\,\%$ of the information is lost by eliminating this patterns and the DDMI retrieves some of this information lost improving the estimation.

An empirical evaluation of alternative methods of estimation for Permutation Entropy in time series with tied values [Traversaro et al. 2017]

This contribution has a twofold objective. The first is to introduce a new method to deal with tied values and the second is to do an exhaustive exploration of all the methods presented. Was review the different existing methodologies that treats this subject by classifying them according to their different strategies.

Extended alphabets: Modified Permutation-entropy algorithm

If equal values may represent a feature state of the system under study, mapping equal values in x_t to equal representation in a symbol π could be considered. This modified order based alphabet results in more possible symbols for each embedding dimension D so it characterizes more system states than the original PE method.

Numerical Analysis

In order to get a reproducible set of time series, all maps presented in Rosso et al. 2013 were simulated using the initial conditions presented therein. Each original time series was truncated up to one decimal resolution, leading to a coarse version of those original time series. The simulation consists in 39 time series generated by different chaotic processes, and each one was simulated for a length of $n = (5000, 10000, 30000, 90000)$.

This yields to quantify how well each strategy estimates the actual Permutation Entropy $H(\boldsymbol{P})$ comparing the result of each original time series with heir respective Permutation Entropy calculated with all the above methodologies over the coarse version. To quantify the behavior of the estimator, the Mean Square Error, defined as $E[\hat{H}(\boldsymbol{P}) - H(\boldsymbol{P})^2]$, is used.

Time Order Imputation and Bayesian Imputation consistently beat the other methodologies for every embedding dimension D and for all the repeated ratios as the MSE is the lowest in all the cases. Methodologies that use extended alphabet tend to sub estimate the entropy, specially when the embedding dimension is small. Random Imputation methodology over estimates the entropy as it create noise and this is more evident when the imputation is made over a time series with a large number of repeated values.

Real Application

We applied the presented strategies to classify using heart rate time series from a (healthy) control group and from patients suffering from Congestive Heart Failure (CHF). The data have been collected from internet data bases:

- NSR2DB Normal Sinus Rhythm RR interval database for the healthy patients, and
- CHF2DB Congestive heart failure RR interval database, for the patients suffering from CHF, from `http://www.physionet.org/cgi-bin/atm/ATM`.

For each database, 15 series were taken, each one with approximately 100000 observations. Bayesian Imputation was an optimum classifier for this application since it does a great job separating the populations.

Bandt-Pompe symbolization dynamics for time series with tied values: A data-driven approach [Traversaro et al. 2018]

The present contribution was to review the different existing methodologies for treating time series with tied values by classifying them according to their different strategies. In addition, a novel data-driven imputation is presented that proves to outperform the existing methodologies and avoid the false conclusions pointed by Zunino and co-workers.

Data-driven imputation

This methodology uses information of the actual time series to deal with ties. Was assumed that there is an underlying (unobserved) time series with no ties and that the data at hand are a corrupted version. With this in mind, the proposal estimates the true patterns from the observed ones using the information available from the corrupted data through a suitable probability distribution. It can be seen as a Data-Driven methodology.

Numerical Analysis

For evaluation of the randomness of the decimal expansion of irrational numbers, was analyze the temporal sequences of the decimal expansion of the irrational numbers π , e, and $\sqrt{2}$ with length $T = 5 \times 10^3$, 10^4, 10^5, and 10^6 of the first digits and evaluate the Bandt-Pompe PDF with embedding dimensions $D = \{3, 4, 5, 6\}$, and time lag $\tau = 1$.

With the effect of large vector effect ($D = 6$) of values in the set $\{0, 1, \ldots, 9\}$, which would probably have repetitions, was is notable that Time Ordered Shannon entropy is significantly greater than that obtained by the Data-Driven technique. The values of $H(D)[\Pi]$ evaluated with BP-PDF following the time order imputation suggest the presence of temporal complex structures in the data, for increasing values of $D \geq 4$, regardless the time series length T. The same operator, when applied to Data-Driven imputed data, does not show any kind of correlation structure, indicating that the decimal expansion of the irrational numbers are compatible with uncorrelated random behavior in agreement with the results obtained by Luque et al. 2009. Note also that similar results are obtained when using BP-PDF with complete imputation, however, in this case, we have a strong reduction of the number of reconstructed vectors, implying that longer time series are necessary in order to have zero forbidden/missing patterns, as for $D = 6$.

Chaotic maps time series

Was evaluated times series $\mathcal{X}^*$ of $T = 10^5$ iteration data (after discarding 10^5 iterations for stability) from the chaotic maps considered by Rosso et al. 2013. The coarse version (low precision) time series $\mathcal{X}$ are obtained truncating the original values x_t^* to two decimal digits.

Was noted that NFMP for the other methodologies present in general lower values than those for BP-PDF Data-Driven, and was conclude that Random Imputation and Complete should not be used. The other two techniques are competitive, but in some cases, Data-Driven outperforms Time Ordered Imputation.

Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions [Zunino et al. 2017]

In this work, was carefully study the effect that the presence of equalities has on permutation entropy estimated values when these ties are symbolized, as it is commonly done, according to their order of appearance.

Numerical Analysis

To illustrate the effect that the occurrence of a high frequency of ties has on PE estimated values, we have numerically generated an ensemble of one hundred independent sequences of $N = 1.000$ pseudorandom integer values drawn from a discrete uniform distribution on the interval $[0, i]$ with i ranging from 1 to 50 with step equal to one. The normalized PE with different embedding dimensions, $D \in \{3, 4, 5, 6\}$, and embedding delay $\tau = 1$ (consecutive data points) has been estimated.

Was is clearly observed that normalized PE values from pseudorandom discrete time series with a high frequency of occurrence of equal values are much lower than those obtained for a pseudorandom continuous time series, leading to a totally spurious identification of non-random temporal structures. Because of the way ties are ordered, the relative frequencies of some permutation patterns are overestimated in detriment of those associated with other motifs which are underestimated, and, consequently, a non-uniform ordinal pattern probability distribution is obtained. Analysis by implementing the WPE has been also carried out. Results obtained confirm that this improved ordinal permutation quantifier also suffers from this weakness.

Two simple applications

As a first application, was investigate the randomness of the decimal expansion of irrational numbers by using the PE, analyzing the ordinal pattern probability distribution of the temporal sequences obtained by picking the first 10.000 digits of the decimal expansion of several irrational numbers such as π , e, and $\sqrt{2}$.

As was is visually concluded, the different motifs are not equiprobable. Some of them are more probable than others, indicating, apparently, the presence of temporal complex structures in the data. But the results imply that the irregularly observed frequency of motifs is a totally spurious effect due to the significant number of equalities that are present in the original time series. Moreover, a surrogate analysis with shuffled realizations appears as a practical alternative to overcome this limitation of the PE.

Was finally developed an ordinal symbolic analysis for radioactive decay data. A signal of plutonium-239 activity with length $N = 10.000$ has been examined. The relative frequencies of the ordinal patterns for embedding dimension $D = 3$ and embedding delays τ between 1 and 100 was analyzed. In this case, the varied the embedding delay in order to check the behavior of the experimental data for different time scales, i.e. for different sampling times. It is concluded that, independently of the time scale, the ordinal pattern probability distribution is irregular.

Conclusions

In summary, we can see that the main analyzes were performed based on digits of the decimal expansion of irrational numbers such as π, and $\sqrt{2}$, and chaotic maps considered in [Rosso et al. 2013]. The methods present in the literature and which have their methodologies analyzed, are:

- Chronological extended alphabet;
- Rank extended alphabet;
- Complete cases;
- Time Ordered imputation;
- Random imputation;
- Bayesian/Data-driven Imputation.

Therefore, the variability of the Error (absolute error or mean square) between current PE and the obtained for each methodology, and the analysis of the number of forbidden/missing patterns (NFMP) are checked to see which one can more accurately represent the dynamics of the sequence generating process with repeated values.

Supplementary material

Characterization of chaotic maps using the permutation Bandt-Pompe probability distribution

This paper provides information about the chaotic maps and our initialization parameters used in Traversaro et al. 2017,Traversaro et al. 2018, and Traversaro et al. 2019. Was consider 27 chaotic maps described by Sprott in the appendix of his book [Sprott and Sprott 2003]. These chaotic maps are grouped as:

- noninvertible maps: (1) logistic map; (2) sine map; (3) tent map; (4) linear congruential generator; (5) cubic map; (6) Ricker's population model; (7) Gauss map; (8) Cusp map; (9) Pinchers map; (10) Spence map; (11) sinecircle map.
- dissipative maps: (12) Hénon map; (13) Lozi map; (14) delayed logistic map; (15) Tinkerbell map; (16) Burgers' map; (17) Holmes cubic map; (18) dissipative standard map; (19) Ikeda map; (20) Sinai map; (21) discrete predator-prey map.
- conservative maps: (22) Chirikov standard map; (23) H ´enon area-preserving quadratic map; (24) Arnold's cat map; (25) Gingerbreadman map; (26) chaotic web map; (27) Lorenz threedimensional chaotic map.

References

Luque, B., Lacasa, L., Ballesteros, F., and Luque, J. (2009). Horizontal visibility graphs: Exact results for random time series. *Physical Review E*, 80(4):046103.

Rosso, O. A., Olivares, F., Zunino, L., De Micco, L., Aquino, A. L., Plastino, A., and Larrondo, H. A. (2013). Characterization of chaotic maps using the permutation bandt-pompe probability distribution. *The European Physical Journal B*, 86(4):116.

Sprott, J. C. and Sprott, J. C. (2003). *Chaos and time-series analysis*, volume 69. Citeseer.

Traversaro, F., Ciarrocchi, N., Cattaneo, F. P., and Redelico, F. (2019). Comparing different approaches to compute permutation entropy with coarse time series. *Physica A: Statistical Mechanics and its Applications*, 513:635–643.

Traversaro, F., Redelico, F. O., Risk, M. R., Frery, A. C., and Rosso, O. A. (2018). Bandt-pompe symbolization dynamics for time series with tied values: A data-driven approach. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075502.

Traversaro, F., Risk, M., Rosso, O., and Redelico, F. (2017). An empirical evaluation of alternative methods of estimation for permutation entropy in time series with tied values. *arXiv preprint arXiv:1707.01517*.

Zunino, L., Olivares, F., Scholkmann, F., and Rosso, O. A. (2017). Permutation entropy based time series analysis: Equalities in the input signal can lead to false conclusions. *Physics Letters A*, 381(22):1883–1892.