Time-series classification by matrix-based methods: Application to blackhole state identification

First A. Author, Fellow, IEEE, Second B. Author, and Third C. Author, Jr., Member, IEEE

Abstract—Across diverse domains such as medicine, weather, finance, agriculture, astronomy, etc., it is required to deal with timeseries of measurements. Classification of timeseries as stochastic (noise-like) or non-stochastic (which has a well-defined structure), helps understand the underlying phenomenon. The methods used to accomplish this classification are either: (i) Correlation Integral (CI)-based or (ii) Entropy-based approaches, both of which are computationally expensive. In this work, we propose two matrix-based methods to achieve stochastic vs nonstochastic classification, without requiring computation-intensive parameters such as Correlation Dimension. The proposed matrixbased methods are: (a) SVD-decomposition followed by topological analysis (using Betti number descriptors) (b) PCA-based features followed by SVM classification. SVD decomposition studies the temporal dynamics while PCA-based approach analyses the timeseries like a stack of data with no regard for temporal ordering, making the two approaches complementary. The proposed methods have been applied to synthetic data, which are 48 distinct realisations of white noise, pink noise (stochastic timeseries), Logistic map and Lorentz system (nonstochastic timeseries), as proof of concept. The utility of the proposed methods is illustrated on astronomy data which are 12 categories of timeseries pertaining to blackhole GRS 1915 + 105, obtained from RXTE satellite. Comparison of obtained results with those in literature are also presented. It is found that among the proposed matrix based methods, SVD analysis concurs with CI based analysis on all 12 categories of time series utilized. However, the inference using PCA based approach illustrates that one class among the 12 turns out to be inconsistent with the other approaches. Investigation into these (in)consistencies is expected to have long standing implications in astrophysics and otherwise.

Index Terms—Timeseries classification, stochastic, nonstochastic, SVD analysis, PCA analysis

I. Introduction

Several real-world phenomena are studied by collecting associated measurements over time, called as timeseries. Timeseries classification as stochastic (noise-like) or non-stochastic (which has a well-defined structure), is the first step in understanding the underlying physical phenomenon. Standard stochastic signals such as white noise, pink noise, etc. exhibit characteristics such as nearly zero auto-correlation coefficients for all possible values of lags and a power spectral density that decays with frequency. The rate of decay determines the kind of noise. On the other hand, standard non-stochastic signals such as Logistic map (at growth rate = 4), Lorenz system result in timeseries that exhibit a well-defined structure, such as having a certain number of fixed points. The trajectory of

This paragraph of the first footnote will contain the date on which you submitted your paper for review. It will also contain support information, including sponsor and financial support acknowledgment. For example, "This work was supported in part by the U.S. Department of Commerce under Grant BS123456."

the vectors taken pair-wise, also exhibits a distinct pattern, called "attractor behavior".

Reported methods that accomplish stochastic vs nonstochastic classification can be broadly divided as: (i) Correlation Integral (CI) based (ii) Entropy-based. CI based approach was proposed in [1]. It is a computation-intensive process, since the CI measure needs to be computed for different choices of embedding dimension and also for different ranges of neighbourhood radius. Besides, it is well-known that this value of correlation dimension does not saturate for a stochastic time series. Hence to establish if the considered timeseries is stochastic, this computation needs to be repeated for diverse values of embedding dimension, making the order of computations needed greater by that factor. Entropy-based approaches [15], [3], [2] utilize concepts of phase-space reconstruction, approximate entropy and recurrence plots. This approach also requires computations of expensive parameters such as Lyapunov exponent.

The problem of stochastic vs non-stochastic classification is important for one of the challenging problems in astrophysics, which could lead to the understanding of black holes. As a black hole cannot be seen directly, to identify it, one has to look for its environment forming a disc-like structure by the infalling matter called accretion disc. In this work, we focus on the black hole source GRS 1915+105, which presents several intriguing facets. It has been divided into 12 different temporal categories: α , β , γ , δ , λ , κ , μ , ν , ρ , ϕ , χ and θ [4], with their respective distinct timeseries. One fundamental aspect of the understanding is to determine if the black hole source is stochastic or non-stochastic (implying turbulent nature of the system). There are studies reported that utilize CI-based approach to determine the characterization of this specific black hole data [5], [6]. However, in this work, we propose to utilize matrix-based methods, Singular Value Decomposition (SVD) and Principal Component Analysis (PCA), to understand the same data. It is useful to compare the inferences obtained using these distinct approaches; the implications of the (dis)similarities in inferences, if any, could lead to better understanding of the temporal dynamics of the system.

It is widely known that the true nature of the source is understood by studying both temporal and spectral features. If the source radiation is temperature dependent, it is called multicolour blackbody or "diskbb" [7]. On the other hand, the temperature independent radiation consists of power-law tail ("PL") [8], [9]. The difference in their implications lies in the fact that the former leads the underlying accretion disc around the black hole to be geometrically thin, while the latter leads to a geometrically thick disc. Studies in literature combine these observations into four possible black hole states [10]:

- 1) Non-stochastic and diskbb: Keplerian disc [7].
- 2) Non-stochastic and PL: Advection Dominated Accretion Flow (ADAF) [9].
- 3) Stochastic and diskbb: Slim disc [11].
- 4) Stochastic and PL: General Advective Accretion Flow (GAAF) [8], [12].

The state identification of the blackhole is critically impacted by the inference if it is labelled stochastic or non-stochastic. The inferences obtained using various approaches can help in localizing the discrepancies in the understanding of the blackhole.

The contributions of this paper for timeseries classification as stochastic or non-stochastic, are :

- SVD decomposition of the data matrix followed by topological analysis of the plot of right singular vectors.
- Novel application of PCA, to derive PCA-based features, followed by SVM classification.

II. RELATED WORK

Several groups have worked on distinguishing between stochastic and non-stochastic time series. In CI-based approach, one constructs an $K \times M$ matrix where each row contains K consecutive samples taken from the time series with a time-shift introduced between successive rows. Each column of the matrix is considered as an observation vector in M-dimensional space. Correlation sum is then computed for various values of radius and embedding dimension,M, as explained in [?]. This method is known to be computationally expensive- $O(N^2M)$. Among entropy based approaches a recent study explored the idea of utilizing Permutation Entropy (PE) to determine the complexity measure of a time series [14]. In the work reported in [15], PE was used to parameterize a given time series followed by classification using Neural Network. The paper explored the idea of utilizing PE of a time series to determine if it strongly resembles known stochastic signals such as pink noise. The claim was that for nonstochastic signals the deviation of the parameter is relatively large as compared to that of the parameter of a stochastic signal. The drawbacks of this approach are that the order of computations required is $O(N^3)$ besides assuming prior knowledge on the length of ordinal sequences.

Apart from CI based and entropy based approaches there are also works that utilize graph-based methods and dictionary based methods. In the work reported in [16], the authors have utilized the horizontal visibility algorithm in order to distinguish between stochastic and non-stochastic processes. A recent work, reported in [17], mapped time series into graphs and computed various topological properties, which they called *NetF*, capturing measures such as centrality, distance, connectivity etc. PCA was applied on the *NetF* feature matrix and clustering was performed on the principal components. In the approach outlined in [18], the authors combined the idea of sparsity and machine learning with non-linear dynamical systems, in order to determine the governing dynamics. Sparse regression was used to determine the fewest terms in the equations that govern the dynamics of the phenomenon. The user-defined dictionary of basis functions consists of wellknown functions such as polynomials, trigonometric functions and exponentials. However, the optimal choice of dictionary for a specific choice of problem remains a challenge.

2

In this work, we propose to utilize computationally simple classical matrix based methods which do not require any assumptions about the underlying phenomenon.

III. PROPOSED METHOD

In this work, we propose two different matrix based approaches to classify time series as stochastic vs non-stochastic. They are 1) SVD decomposition followed by topological analysis (using Betti number descriptors) and 2) PCA derived features followed by SVM classification. Proof of concept on synthetic signals is also presented.

A. SVD based approach

In this approach, we form uncorrelated observation vectors from the raw time series data by choosing an appropriate value of embedding dimension [6] using autocorrelation plot. Data matrix, D, is formed with each row as the time shifted version of the original time series. The time shift is chosen to be large enough so that each column can be viewed as a different observation vector of the same time-evolving phenomenon. Temporal dynamics is understood by utilizing the right singular vectors of the SVD decomposition of D as given in equation (1) below. We consider the top two right singular vectors, E1 and E2, for our analysis.

$$D = U\Sigma V^T. (1)$$

We observe the topology of the plot E1 vs E2. For nonstochastic time series this plot is expected to show a specific pattern (attractor behavior, where the plot follows a structured trajectory leaving well-defined voids). On the other hand, E1 Vs E2 plot for a stochastic time series, appears as a single blob without any voids. The toplogy of the E1 vs E2 plot is captured using Betti numbers [13]. Betti number descriptor for a d-dimensional manifold is a vector of d integers which is represented as $\beta = (\beta_0, \beta_1 \dots \beta_{d-1})$. Here β_0 is the number of blobs (connected components) and β_k represents number of k-dimensional holes for k > 0. The E1 vs E2 plots are 2-D manifolds, which are described by $\beta = (\beta_0, \beta_1)$. For a stochastic time series the values of β_0 and β_1 are expected to be 1 (single connected component) and 0 (no voids) respectively. Hence the L1-norm of a stochastic time series will be 1. However, for a non-stochastic time series, we observe that the value of β_0 can be greater than 1 (can have one or more connected components) and the value of β_1 is always greater than 0 (presence of well defined voids) due to the attractor behavior. Hence the L1-norm of non-stochastic time series will always be greater than 1. In this work, we utilize the L1-norm of the E1 Vs E2 plot of a given time series to classify it as stochastic or non-stochastic.

B. PCA Based approach

PCA decomposition is carried out to infer if the given time sereis possesses a dominant orientation or not. This is computed by hierarchally splitting the time series into two halves, and computing the covariance matrix of this split observations. The eigenvalues of this 2×2 covariance matrix will show one of the signatures: If the data indeed show any dominant direction (as in non-stochastic time series), then the larger eigenvalue will be significantly greater than the other. This will lead to a large eigen value ratio. On the other hand, if the data do not show any dominant direction (as in stochastic time series), then the two eigenvalues of the covariance matrix will be comparable. This will lead to small eigen value ratio. This observation is utilized in devising features for stochastic Vs non-stochastic classification. The steps are outlined as below.

For a time series consisting of n values $z_1, z_2 \dots z_n$.

- Split the series into two halves $(z_1, z_2 \dots z_{\left \lfloor \frac{n}{2} \right \rfloor})$ and
- $(z_{\lfloor \frac{n}{2} \rfloor + 1}, \dots z_n).$ Compute covariance matrix, C, by treating the samples in two halves as $\left|\frac{n}{2}\right|$ observations of 2-D vectors.
- Find eigenvalues of C, λ_1 and λ_2 ; the eigenvalue ratio is computed as λ_1/λ_2 where $\lambda_1 > \lambda_2$ (eigenvalues of a covariance matrix are real).

If the eigenvalue ratio for an interval is greater than a value of threshold, τ (computing optimal value of τ is described in subsection III-B1 below), the interval is further split into two sub-intervals of equal size. Subsequently, the eigenvalue ratio for each sub-interval is computed. The process is repeated as long as the length of the sub-interval is greater than a predefined number of samples (here taken as 100). For a fixed value of τ , the following features are derived

- Variance of Eigenvalue Ratio (VER): This is the variance of the eigenvalue ratios of covariance matrices across subintervals in the entire time series.
- Area Under the Eigenvalue Ratio curve (AUER): This measure captures the area under the curve of the eigenvalue ratio for the entire time series.

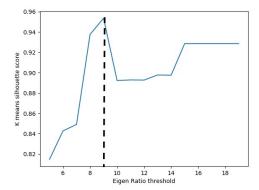


Fig. 1. Plot of Silhoutte score vs eigenratio threshold. Maximum Silhoutte score value, which indicates the best clustering, is obtained at a threshold of

1) Computing optimal value of threshold τ : In order to arrive at the optimal value of τ , we observe the plot of the Silhoutte score of K-Means clustering, with K=2 (stochastic and non-stochastic), performed using the above feature set, as a function of the threshold value. The value of the threshold

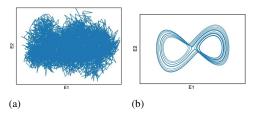


Fig. 2. Comparison of E1 Vs E2 plots for: (a) white noise (stochastic) (b) Lorentz system (non-stochastic)

that results in the best Silhoutte clustering score is taken as τ . This process is illustrated in the Silhoutte score plot shown in Fig1. For the time series considered here for illustration, it is evident from the plot that the best clustering is obtained at threshold value 9, resulting in the maximum value of Silhoutte score. Hence we use the corresponding $\tau = 9$ to arrive at the optimal hierarchical splitting and subsequent computing of the devised PCA-based features, VER and AUER.

C. Proof of Concept on Synthetic Data

The proposed approaches have been applied to standard synthetic signals. For stochastic class of signals, white noise and pink noise are considered; for non-stochastic class of signals, Lorenz system and Logistic map (for growth rate = 4) are considered.

SVD-Decomposition based technique: The SVD decomposition of the data is computed, followed by the plot of the top two right singular vectors. This plot is utilized to determine the Betti descriptors. The plot in Fig. ?? corresponds to the E1 vs E2 plot for a realization of white noise, which is known to be stochastic. The plot shows one single blob implying Betti descriptor of (1 0), which has L1-norm of 1. As discussed before, L1-norm of 1 implies that the timeseries is labelled as stochastic. On the other hand, the plot in Fig. ?? corresponds to the E1 vs E2 plot of a realization of Lorentz system, which is known to be non-stochastic. The plot shows two distinct voids, implying Betti descriptor to be (1 2), which has L1-norm of 3, implying that the timeseries is labelled as non-stochastic. This inference mechanism has been utilized on real data described in section IV. The order of computations needed is mainly determinied by the complextity of SVD Decomposition (For M x N matrix = $O(M^2N)$, which is linear in N).

PCA-Decomposition based technique :PCA-based features (i) VER and (ii) AUER are computed for the considered synthetic signals which are 24 different realizations of white noise and logistic map (growth rate = 4). The scatter plot of these features is shown Fig 3 below. We observe the following:

- VER: For a stochastic signal since the variation in the eigenvalue ratios is typically small, the computed variance across the values is also small. On the other hand, for a non-stochastic signal, since the eigenvalue ratios occupy diverse values, VER is typically high.
- AUER: For a stochastic time series, since the eigenvalue ratios are small across the entire span, the value of AUER is also small. However, for a non-stochastic signal, the eigenvalue ratios remain high for longer time intervals. Hence the value of AUER is significantly higher.

The plot makes it evident that in this feature space the two classes of timeseries, stochastic and non-stochastic, become linearly separable. Hence a linear SVM classifier is utilized. For training the SVM, computed features from white noise and Logistic map are utilized. For validating the trained SVM, 12 realizations of pink noise and Lorentz system are used. The classification on all 12 realizations of pink noise, yields the label stochastic, and the classification label for Lorentz system is obtained as non-stochastic, leading to perfect validation accuracy. This trained SVM is used to classify real data as described in section IV

The order of computations needed can be split into computations for covariance matrix (which involves computing X^TX where X is of dimensions $\left\lfloor \frac{N}{2} \right\rfloor \times 2:O(N)$), computing eigen values of 2×2 matrix which takes constant time and finally number of iterations (For N-length timeseries = assuming maximum number of splits, (log N)). Since final SVM 2-class classification step also takes constant time, the overall time complexity of PCA based approach is O(NloqN)

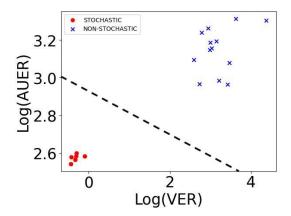


Fig. 3. Scatter plot of PCA-based features for synthetic data

IV. RESULTS AND DISCUSSION

In this section, we present the real data used, results obtained using proposed approaches and comparison with results in literature. All codes and results are available at the link https://github.com/sunilvengalil/ts_analysis_pca_eig

A. Real Data

The proposed approaches are illustrated on the publicly available data of $GRS\ 1915 + 105$ taken from website [19]. 12 categories of time series are utilized from the available data. All these time series are re-sampled with a sampling interval of 0.1 second. These datasets were also used in the work reported in [10], where the authors use CI based approaches, leading us to be able to compare our obtained results with theirs.

B. Results of SVD based analysis

From SVD decomposition of the data matrix, we plot the top 2 right singular vectors (E1 vs E2) to understand the temporal dynamics for each time series. For classification, we use the L1-norm of the Betti descriptor, β , $\|\beta\|_1$. If $\|\beta\|_1 = 1$, the

time series is classified as stochastic, else the time series is non-stochastic. The obtained values of L1-norm, along with the classification labels have been tabulated in Table I.

C. Results of PCA based analysis

PCA-derived features, VER and AUER, are computed for each of the timeseries. These features are input to the SVM classifier to obtain the class labels. The obtained feature values along with the classification labels have been tabulated in Table I. Comparison of results using proposed approaches with CI based approach [10] is shown in Table I.

D. Comparison of Blackhole state Inference

Comparison of Blackhole state Inferencere using proposed approaches as against CI based approach [10] is shown in Table II. We observe that SVD based analysis results in classification are consistent with CI based results for all the 12 categories of time series. However, with the PCA based approach the inference for δ time series is not consistent with the other two approaches. According to the CI based analysis δ turns out to be in between states slim disc and GAAF [10]. However, the present analysis shows that δ falls in between ADAF and Keplerian disc.

E. SVD Vs PCA approach

SVD decomposition is utilized to study the temporal dynamics in the timeseries considered. This is in contrast to the PCA approach which does not consider the temporal ordering of data. Hence each of the approaches follows a perspective that is complementary to the other. Utilizing them both together to obtain inferences that concur makes the study reliable. In case of non-concurrence, it is clear that the problem would need further investigations.

V. CONCLUSION

Exploring different techniques in order to have a conclusive inference for black hole systems turns out to be indispensable. We explore two different classical matrix based techniques to identify states of GRS 1915+105 black hole using the time series obtained from RXTE satellite data. Based on our analysis, we are able to identify two extreme temporal dynamical classes of accretion around black holes. In the first approach we extend SVD decomposition to understand temporal dynamics, by adding topological descriptors, to classify time series as stochastic vs non-stochastic. In yet another approach, a novel application of PCA to characterize the time series is proposed. We compare inferences of the CI based approach with those obtained using the proposed matrix based methods. Of the 12 categories of time series analysed, a mismatch is observed in the PCA based inference of only one class, while all other classes concur. Since SVD approach studies temporal dynamics, and PCA analysis is agnostic to temporal ordering, the two approaches together are required for unambiguous inference.

TABLE I

Timeseries: Comparison between CI based label and inference using proposed approaches. The mismatched time series class, δ , is shown in bold. (Here NS stands for non-stochastic and S stands for stochastic)

Class	CI	Betti	SVD	VER	AUER	PCA	Match
	Label	Norm	Label			Label	
β	NS	4	NS	483	43	NS	Yes
θ	NS	5	NS	778	58	NS	Yes
λ	NS	4	NS	6782	314	NS	Yes
κ	NS	4	NS	5199	144	NS	Yes
μ	NS	2	NS	51	12	NS	Yes
ν	NS	7	NS	32	16	NS	Yes
α	NS	6	NS	1.9	27.7	NS	Yes
ρ	NS	2	NS	147	35	NS	Yes
δ	S	1	S	9.7	26.2	NS	No
ϕ	S	1	S	0.5	15	S	Yes
γ	S	1	S	1	16	S	Yes
χ	S	1	S	0.25	6.05	S	Yes

TABLE II
BLACKHOLE STATE INFERENCE COMPARISON ACROSS CI-BASED AND PROPOSED APPROACHES:

Name	diskbb	PL	State by	State by	State by	Match
			CI	SVD	PCA	
β	46	52	ADAF	ADAF	ADAF	Yes
θ	11	88	ADAF	ADAF	ADAF	Yes
λ	54	46	Keplerian	Keplerian	Keplerian	Yes
κ	59	51	Keplerian	Keplerian	Keplerian	Yes
μ	56	41	Keplerian	Keplerian	Keplerian	Yes
ν	28	72	ADAF	ADAF	ADAF	Yes
α	23	77	ADAF	ADAF	ADAF	Yes
ρ	28	72	ADAF	ADAF	ADAF	Yes
δ	48	50	ADAF	ADAF	GAAF	No
ϕ	50	34	Slimdisc	Slimdisc	Slimdisc	Yes
γ	60	31	Slimdisc	Slimdisc	Slimdisc	Yes
χ	09	89	GAAF	GAAF	GAAF	Yes

REFERENCES

- [1] Procacia, I. "Measuring the strangeness of strange attractors." Physica. D 9.1-2 (1983): 189-208.
- [2] Rostaghi, Mostafa, and Hamed Azami. "Dispersion entropy: A measure for time-series analysis." IEEE Signal Processing Letters 23.5 (2016): 610-614.
- [3] Kirichenko, L. O., and A. Yu Habacheva. "Comparative analysis of the complexity of chaotic and stochastic time series." Radio Electronics Computer Science Control Follow journal 2 (31) (2014): 126-134.
- [4] Belloni, T., et al. "A model-independent analysis of the variability of GRS 1915+ 105." arXiv preprint astro-ph/0001103 (2000).
- [5] Mukhopadhyay, Banibrata. "Chaotic behavior of micro quasar GRS 1915+ 105." AIP Conference Proceedings. Vol. 714. No. 1. American Institute of Physics, 2004.
- [6] Misra, Ranjeev, et al. "The nonlinear behavior of the black hole system grs 1915+105." The Astrophysical Journal 643.2 (2006): 1114.
- [7] Shakura, Ni I., and Rashid Alievich Sunyaev. "Black holes in binary systems. Observational appearance." Astronomy and Astrophysics 24 (1973): 337-355.
- [8] Chakrabarti, Sandip K., and Lev G. Titarchuk. "Spectral properties of accretion disks around galactic and extragalactic black holes." arXiv preprint astro-ph/9510005 (1995).
- [9] Narayan, Ramesh, and Insu Yi. "Advection-dominated accretion: Underfed black holes and neutron stars." arXiv preprint astro-ph/9411059
- [10] Adegoke, Oluwashina, et al. "Correlating non-linear properties with spectral states of RXTE data: possible observational evidences for four different accretion modes around compact objects." Monthly Notices of the Royal Astronomical Society 476.2 (2018): 1581-1595.
- [11] Abramowicz, M. A., et al. "Slim accretion disks." The Astrophysical Journal 332 (1988): 646-658.
- [12] Rajesh, S. R., and Banibrata Mukhopadhyay. "Two-temperature accretion around rotating black holes: a description of the general advective flow paradigm in the presence of various cooling processes to explain low

- to high luminous sources." Monthly Notices of the Royal Astronomical Society 402.2 (2010): 961-984.
- [13] Naitzat, Gregory, Andrey Zhitnikov, and Lek-Heng Lim. "Topology of Deep Neural Networks." J. Mach. Learn. Res. 21.184 (2020): 1-40.
- [14] Bandt, Christoph, and Bernd Pompe. "Permutation entropy: a natural complexity measure for time series." Physical review letters 88.17 (2002): 174102.
- [15] Boaretto, B. R. R., et al. "Discriminating chaotic and stochastic time series using permutation entropy and artificial neural networks." Scientific reports 11.1 (2021): 1-10.
- [16] Lacasa, Lucas, and Raul Toral. "Description of stochastic and chaotic series using visibility graphs." Physical Review E 82.3 (2010): 036120.
- [17] Silva, Vanessa Freitas, et al. "Novel features for time series analysis: a complex networks approach." Data Mining and Knowledge Discovery (2022): 1-40.
- [18] Brunton, Steven L., Joshua L. Proctor, and J. Nathan Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems." Proceedings of the national academy of sciences 113.15 (2016): 3932-3937.
- [19] RXTE Public Data https://heasarc.gsfc.nasa.gov/docs/xte/xte_public. html