Time-series classification using matrix-based methods: Application to blackhole state identification of RXTE satellite data

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 non-stochastic classification, without requiring the computation-intensive phase space. The proposed matrix-based methods are: (a) SVD-decomposition followed by topological analysis (b) PCA-based technique. The proposed methods have been applied to synthetic data, as proof of concept. The utility of the methods is illustrated on astronomy data which are 12 categories of timeseries pertaining to blackhole *GRS 1915 + 105*, obtained from RXTE satellite. Comparisons of obtained results with those in literature are also presented. The order of computational complexity using the proposed approaches is XXXXXXX of N, where N is the length of the timeseries. In contrast, CI based approaches require XXXX, while Entropy-based approaches need XXXXX. 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, non-stochastic, SVD analysis, PCA analysis**

I. Introduction

Several real-world phenomena are studied by collecting associated measurements over time, popularly called as time-series. 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 characteistics 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. For such phenomena, computing parameters such as Correlation Dimension helps in revealing the underlying dynamics. However, for stochastic timeseries the Correlation Dimension never saturates. Hence if the goal of the study is to check if the timeseries is stochastic, then such computations must be avoided.

In literature, methods that accomplish this classification can be broadly categorised as : (i) Correlation Integral based (ii) Entropy-based. Correlation Integral based approach was proposed in [**?**]. It is a computation-intensive process, since the Correlation Integral (CI) needs to be computed for different choices of Embedding dimension, which can only be approximated using the autocorrelation plot. 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 a large range of values of Embedding Dimension, making the order of computations needed greater by that factor. Entropy-based approaches utilize concepts of phase-space. This is also a computation-intensive process, with several assumptions about the charatceristics of the timeseries being considered.

The problem of stochastic vs non-stochastic classification is important for one of the challenging problems in astrophysics, which is 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 classified into 12 different temporal classes: α, β, γ, δ, λ, κ, μ, ν, ρ, ϕ, χ and θ [**?**], with their respective distinct time series. One fundamental aspect of the understanding is to determine if the black hole source is a stochastic system or a non-stochastic one. The latter one is related to the well-known turbulent nature of the system. There are several studies that utilize the Correlation Integral (CI) approach to determine the characterization of the black hole data [**?**], [**?**]. However, there can also be other approaches to understanding the same data by applying, for e.g., matrix-based methods such as Principal Component Analysis (PCA) and Singular Value Decomposition (SVD). It

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 next few paragraphs should contain the authors’ current affiliations, including current address and e-mail. For example, F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar.colostate.edu).

is useful to compare the inferences obtained using these two distinct approaches; the implications of the (dis)similarities in inferences, if any, could lead to questions about understanding the temporal dynamics of the system.

Interestingly, to quantify the properties of a black hole source, along with temporal features one has to look for spectral features as well, they together lead to the true nature of the source. If the source radiation is temperature dependent, it produces more like a blackbody radiation, namely multicolour blackbody or "diskbb" [?]. On the other hand, the temperature independent radiation consists of a power-law tail, named as "PL" [?], [?]. While the former leads the underlying accretion disc around the black hole to be geometrically thin, the latter leads to a geometrically thick disc.

In the present study, black hole states are determined by classifying the given time series, which is photon count rate as a function of time, as being either stochastic or non-stochastic. This classification is performed using classical matrix based methods, SVD and PCA. However, the novelty of the study lies in (i) quantifying temporal complexity obtained by SVD decomposition, using topological techniques and (ii) utilizing features derived from PCA for classification. Based on our analysis there are four possible black hole states [?]:

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

THESE FOUR BLACK HOLE STATES SHOULD BE BROUGHT IN A TABLE AT THE END, BASED ON OBTAINED RESULTS Utility of the proposed approach is illustrated by comparing the results with previously established methods. Following are the major contributions of this paper:

- SVD decomposition of the data matrix is used for identifying the temporal dynamics of the time series as in [?]. A plot involving the top two right singular vectors of the data matrix shows a clear distinction between stochastic and non-stochastic time series. This distinction is captured using the topological descriptor called Betti numbers [?]. This descriptor for a stochastic time series is topologically simpler than that for a non-stochastic one.
- PCA, which is a widely used approach for decorrelating features and dimensionality reduction, is utilized for characterizing a time series as stochastic vs non-stochastic. We propose a novel approach by iteratively computing eigenvalue ratios of covariance matrix for different subintervals of the time series. We derive multiple features from the eigenvalue ratios and use them to characterize the time series.

II. Related Work

Several groups have worked on distinguishing between stochastic and non-stochastic time series. The idea of utilizing Permutation Entropy (PE) to determine the complexity measure of a time series was explored in [?]. In the work reported in [?], 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 is strongly correlated with known stochastic signals (noise). The claim was that for non-stochastic signals the deviation of the parameter is relatively large as compared to that of the parameter of a stochastic signal. Another set of reported studies are based on graph theory. In the work reported in [?], the authors have utilized the horizontal visibility algorithm in order to distinguish between stochastic and non-stochastic processes. A recent work, reported in [?], 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 [?], 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 well-known functions such as polynomials, trigonometric functions and exponentials. However, the optimal choice of dictionary for a specific choice of problem remains a challenge.

In this work, we propose to utilize 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 characterize time series as stochastic vs non-stochastic. They are 1) SVD decomposition followed by 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 utilizing the optimal value of embedding dimension [?]. 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.

$$D = U\Sigma V^T. \tag{1}$$

We observe the topology of the plot E1 vs E2. For non-stochastic time series this plot is expected to show a specific pattern (attractor behavior, where the plot follows a structured trajectory leaving a 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 [?]. 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 and 0 respectively, as the E1 vs E2 plot consists of one single blob. 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 and the value of β_1 is always greater than 0 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 $(z_1, z_2 \dots z_{\lfloor \frac{n}{2} \rfloor})$ and $(z_{\lfloor \frac{n}{2} \rfloor + 1}, \dots z_n)$.
- Compute covariance matrix, C, by treating the samples in two halves as $\lfloor \frac{n}{2} \rfloor$ observations of two dimensional 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 less than a value of threshold, τ (computing optimal value is described later), 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

- **V**ariance of **E**igenvalue **R**atio (**VER**): This is the variance of the eigenvalue ratios of covariance matrices across sub-intervals in the entire time series.
- **A**rea **U**nder the **E**igenvalue **R**atio curve (**AUER**): This measure captures the area under the curve of the eigen-value ratio for the entire time series.

In order to arrive at the optimal value of τ, we observe the plot of the Silhoutte score of K-Means clustering into 2 clusters (stochastic and non-stochastic) performed on the feature set, as a function of the threshold value. The value of the threshold that results in the best Silhoutte clustering score is taken as τ.

The figure below shows the variation in the Silhoutte score of K-means clustering across various values of Threshold. For the considered time series, it is evident from the plot that a minimal value of Silhoutte score is obtained at a threshold of 7. Hence we used this threshold value to compute the PCA features.

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. The plot in Fig. **??** is used for determining the Betti descriptors. PCA-based features,

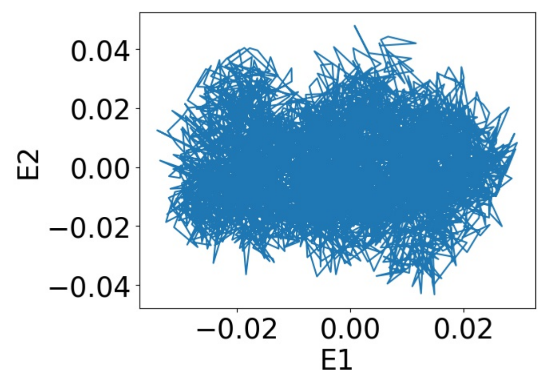

Fig. 1. Plot of Top-2 Right singular vectors for a stochastic timeseries

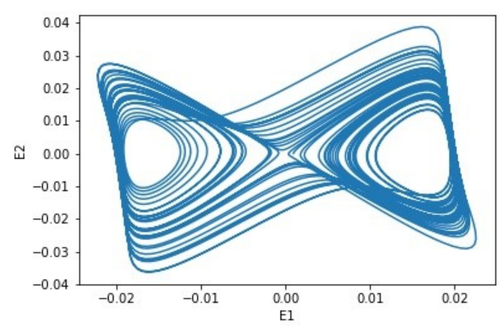

Fig. 2. Plot of Top-2 Right singular vectors for a non-stochastic timeseries

(i) VER (Variance of Eigen ratios) and (ii) AUER (Area under the curve of the eigen ratios) are computed. The scatter plot of these features for multiple realizations (total = 24) of the considered synthetic signals is shown below. The plot makes it evident that a decision boundary separating the two classes, stochastic and non-stochastic can be easily found in this feature space. Hence a linear SVM classifier is utilized. For training the SVM, features from white noise and Logistic map are utilized. For testing the trained SVM, pink noise and Lorentz system are used.

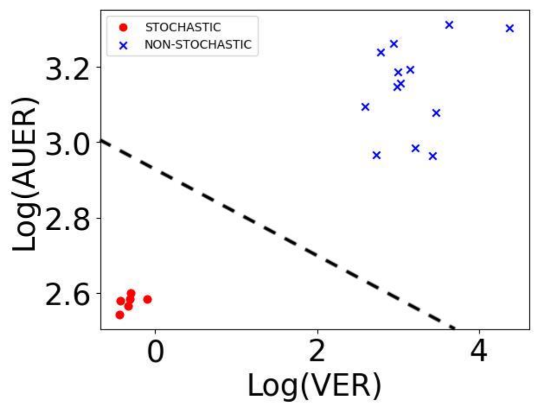

Fig. 3. Scatter plot of PCA-based features for synthetic data

IV. Results and Discussion

In this section, we present the real data used, results obtained using proposed approaches and comparison with results in literature.

A. Real Data

The proposed approaches are illustrated on the publicly available data of *GRS 1915 + 105* taken from website [?]. 12 distinct 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 [?], 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. Figure **??** shows representative E1 vs E2 plots for time series that are classified as non-stochastic and Figure **??** shows the corresponding plots for time series that are classified as stochastic. The Betti number descriptors for each of the E1 vs E2 plots are tabulated in Table II under the column *Betti descriptor*. In order to infer the label of the time series from the Betti descriptors, we use the L1-norm of β, $\|\beta\|_1$. If $\|\beta\|_1 > 1$, the time series is classified as non-stochastic, else the time series is stochastic.

TODO :Table of Betti number - Time series, Betti Descriptor, l1 noram , SVD label

C. Results of PCA based analysis

TODO : table of derived features and SVM inference - Timer series, VER, AUER, SVM label

Figures **??** and **??** show the eigenvalue ratio plots for stochastic time series shown in Figure **??** and non-stochastic time series shown in Figure **??** respectively. We compute PCA-derived features, VER and AUER. These features are input to the SVM classifier to obtain the class labels.

- VER: For a stochastic signal since the variation in the eigenvalue ratios is typically small, the computed variance across the values is 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.

D. Consolidated Results

1) Comparison of Results: Table II tabulates the computed features and the respective inferences using the proposed approaches. Comparison of our results with CI based approach [?] is also presented. The columns of the table are described below.

1) Column 1 (*Class*) gives the class of the time series [?].
2) Column 2 (*diskbb*) and column 3 (*PL*) give quantities diskbb and PL, respectively, which indicate the spectral states of the black hole [?].
3) Column 4 (CI Inference) gives the inference about the state of the time series using CI approach [?].
4) Column 5 SVD based inference.
5) Our inference using these PCA features is given in column XX.
6) Finally the last column gives if there is a match between all three inferences.

E. Identification of black hole states

We observe that SVD based analysis results in classification are consistent with CI based results for all the 12 classes of time series. However, with the PCA based approach the inference for δ time series is not consistent with the other two approaches. We observe that the PCA based features, VAR and Area, result in visible clustering as shown in Figure **??**. This could be attributed to the fact that these features take into account the entire span of time series and hence form robust feature space. According to the CI based analysis δ turns out to be in between states slim disc and GAAF [?]. However, the present analysis shows that δ falls in between ADAF and Keplerian disc.

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

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 Label	Betti Norm	SVD Label	VER	AUER	PCA Label	Match
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 CI	State by SVD	State by PCA	Match
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

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 classes of time series analysed, a mismatch is observed in the PCA based inference of only one class, while all other classes concur.