Finding universal relations using statistical data analysis

Praveen Manoharan[*] and Kostas D. Kokkotas[†]
Theoretical Astrophysics, IAAT, University of Tübingen, 72076 Tübingen, Germany

(Received 13 January 2023; accepted 21 April 2024; published 20 May 2024)

We present applications of statistical data analysis methods from both bivariate and multivariate statistics to find suitable sets of neutron star features that can be leveraged for accurate and equation of state independent—or universal—relations. To this end, we investigate the ability of various correlation measures such as distance correlation and mutual information in identifying universally related pairs of neutron star features. We also evaluate relations produced by methods of multivariate statistics such as principal component analysis to assess their suitability for producing universal relations with multiple independent variables. As part of our analyses, we also put forward multiple entirely novel relations, including a multivariate relation for the f-mode frequency of neutron stars with a reduced average relative error of 0.010, compared to an error of 0.015 of existing, bivariate relations.

DOI: 10.1103/PhysRevD.109.103033

I. INTRODUCTION

The successful detection of gravitational waves from binary neutron star (BNS) mergers through the LIGO-Virgo detectors [1,2] has opened a new avenue into probing and understanding the structure of neutron stars and will hopefully allow us to eventually uncover their true equation of state (EoS).

Important tools for this task are EoS independent—or (approximately) universal—relations that allow for the inference of neutron star bulk parameters through information extracted from gravitational waves. Inspired by early work on such universal relations for isolated neutron stars [3–8], the last five years have also given rise to universal relations for BNSs [9–11]; they relate features of the premerger neutron stars to the early postmerger remnant, primarily relying on numerical relativity simulations.

Following our own recent work on universal relations for BNSs using perturbative calculations [12,13], we found that, with the increasing number of features and amount of data that theoretical computations are able to produce, the traditional method of relying on physical intuition to find universal relations might not always uncover all possible or the best universal relations for a given scenario: instead, an automated approach fueled by statistical data analysis might yield better results in finding highly correlated features, and the best functional form to relate them with. A recent work by Soldateschi *et al.* [14] demonstrated the application of principal component analysis (PCA) to the construction of universal relations with multiple independent variables.

In this paper, we present applications of statistical data analysis methods from both bi- and multivariate statistics to find suitable sets of neutron star features that can be leveraged for accurate and EoS independent relations. To this end, we first analyze the effectiveness of four different correlation measures—Pearson correlation, distance correlation [15], mutual information [16], and maximal information [17]—in identifying pairs of features amenable to universal relations. We find that the conventional wisdom that Pearson correlation only detects linearly correlated features also applies to the use case of finding bivariate universal relations for neutron stars. Furthermore, we also find that mutual information based features are more suited for finding nonlinear correlation between features, making them more useful for this application.

In a second step, inspired by [14], we investigate the application of principal component analysis (PCA) in constructing multivariate universal relations, i.e., relations with multiple independent variables. We find this method suitable for constructing universal relations that combine several features of a neutron star to predict a target feature. Among others, we find the an entirely novel relation between the average density $\tilde{\rho} = \sqrt{M/R^3}$, compactness $C = M/R$ and the f-mode frequency ω_f of a neutron star of the form,

$$\omega_f = 0.00017\hat{F}^2 + 0.006\hat{F} + 0.003 \tag{1}$$

with

$$\hat{F} = 6.911\frac{\tilde{\rho}}{0.04} - 1.716\frac{C\tilde{\rho}}{0.01}. \tag{2}$$

Since the factor $\hat{F}$ is approximately proportional to the factor $(1 - C)$, this relation could be considered a first

[*]praveen.manoharan@uni-tuebingen.de
[†]kostas.kokkotas@uni-tuebingen.de

order, relativistic correction to the original relation between $\tilde{\rho}$ and ω_f derived by Andersson and Kokkotas [3,4], which was inspired from Newtonian gravity. In particular, it can be considered a step towards the well-known general relativistic universal relation between the f-mode frequency ω_f and the compactness C put forward by Tsui and Leung [5].

We perform our analyses using two different datasets from the literature [12,18], exemplifying the generalizability of the methods discussed in this work. The results in this work present a first step towards a automated, statistical data analysis driven effort towards the identification and construction of universal relations for neutron stars (and other objects of astrophysical interest). In a time where the amounts of theoretical model data for astrophysical objects is drastically increasing, we expect having such robust and automated methods available as tools will have a tremendous effect on the quality and quantity of universal relations that will become available in the future.

A. Outline

We begin by introducing the two datasets that we will base our analyses on in Sec. II. We then introduce the bivariate approach to finding universal relations in Sec. III, and discuss the found relations and the implications for the effectiveness of the analyzed correlation measures in Sec. IV.

In Sec. V, we introduce the multivariate approach based on PCA for finding universal relations, before we discuss some exemplary universal relations we were able to construct in Sec. VI. We finally conclude our work and give an outlook into potential future directions in Sec. VII.

Note that, unless stated otherwise, we will assume geometrized units in which $G = c = 1$ throughout this paper.

II. NEUTRON STAR DATA

In this work we consider nonrotating neutron stars from a wide range of equations of state. We here give a brief description of the origin and shape of the datasets we utilize for our analysis. For a detailed treatment of the computation of the neutron star models we refer to the original work [12,18].

A. Datasets

For our analyses, we utilize two different datasets that were used in previous publications; dataset A contains a subset of around 58 nonrotating neutron star models contained in the dataset originally put forward in [19] for the study of rotating and nonrotating neutron stars. This subset covers five different EoSs.

Dataset B contains a subset of 126 nonrotating neutron star models contained in the dataset originally put forward in [18] for the study of f- and g-mode frequencies of nonrotating neutron stars. This subset covers a wider range of 15 EoSs.

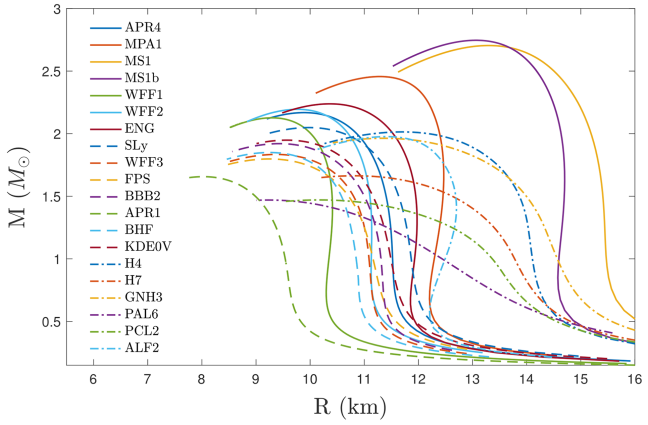

FIG. 1. Mass-radius relations taken from Fig. 1 in [18].

Both datasets contain models of nonrotating stars of different EoSs, providing the values of a wide range of parameters of these neutron stars. There is some overlap in the parameter space considered within each dataset, but both datasets were generated independently as part of different research projects.

While dataset A only covers a subset of the EoSs considered in dataset B, it contains some additional features of nonrotating neutron stars that we can include in our analysis. For a comprehensive discussion of the EoSs covered in each dataset we refer to each respective publication. For an overview, we show in Fig. 1 the mass-radius relations of a wider range of EoSs, taken from [18], of which the EoSs considered in this work are a subset.

The main purpose of utilizing two different datasets is that it allows us to investigate in how far our qualitative observations regarding, e.g., the relative performance of each correlation measure, generalize to different data. To this end, we treat each dataset independently, and do not merge the data to obtain one larger dataset. By observing the same behavior independently in both datasets increases the confidence that the observations made here also generalize to other data.

B. Neutron star features

The features considered in our analysis are obtained through either the direct integration of the TOV equations, or through first-order perturbation of the nonrotating neutron star models. The formal description on how these features are obtained are presented in the previous publications that introduced this data [12,18]. We here summarize the properties of these features. Table I gives an overview of all the features mentioned here.

The first group of features is comprised of macroscopic equilibrium features of the computed neutron star models. In a first step, this includes the gravitational mass M (typically normalized $\bar{M} = M/M_\odot$, where $M_\odot$ is the solar

TABLE I. Neutron star features considered in this paper. The last column indicates whether these features are available in datasets A or B.

Name	Symbol	Dataset
Gravitational mass	$\bar{M} = M/M_\odot$	A, B
Radius	R	A, B
Square root of average density	$\tilde{\rho} = \sqrt{M/R^3}$	A, B
Compactness	$C = M/R$	A, B
Moment of inertia	$\bar{I} = I/M^3$	A
Effective compactness	$\eta = \sqrt{M^3/I}$	A
f-mode frequency	$\omega_f = 2\pi f_2$	A, B
g-mode frequency	$\omega_{g_1} = 2\pi f_{g_1}$	B
Tidal deformability	$\bar{\lambda} = \frac{\lambda}{M^5}$	A, B

mass), the radius R and the compactness $C = M/R$. In a second step, we here also consider other neutron star features that have been identified in the literature as useful in the construction of universal relations. This includes the square root of the average density $\tilde{\rho} = \sqrt{M/R^3}$, the moment of inertia I (typically normalized $\bar{I} = I/M^3$) and effective compactness $\eta = \sqrt{M^3/I}$ of the neutron star.

All of these equilibrium features we try to correlate to various perturbative features that are computed using linear perturbations; this includes the tidal deformability λ (typically normalized $\bar{\lambda} = \lambda/M^5$), the (angular) f-mode frequency ω_f and the (angular) g-mode frequency ω_{g_1} (we here only consider the first g-mode frequency for brevity, but keep the given notation to go along with the notation presented in [18]). To keep in line with a commonly used notion in the literature [3,4], we will denote relations involving the latter as astroseismological relations.

III. BIVARIATE CORRELATION ANALYSIS

The simplest universal relations try to directly relate two different features of neutron stars, i.e., they are bivariate relations. We believe that by evaluating the correlation between different features, we can automate finding such bivariate relations to a high degree. The main issue, however, is identifying which correlation measure is best suited to the task of finding universal relations (for neutron stars).

In this section, we first discuss the concept of linear correlation and the corresponding linear correlation measure (Pearson correlation). We then introduce three additional nonlinear measures of relation that will allow us to find universally related features that are not uncovered by linear correlation.

A. Linear correlation

Throughout this paper, we differentiate between linearly and nonlinearly related features. On a basic level, we will use these two terms to describe the structure that is visually apparent in the scatter plot of a given feature pair; we consider two features to be linearly related if their functional relation can be well-approximated by a linear function, and non-linearly related if they show any other kind of functional relation that is not represented by a linear function.

The degree of linear relation between these features is more formally quantified by their linear correlation, otherwise known as Pearson correlation. Assume that the values of the two considered features are given by two random variables X and Y. Then the Pearson correlation coefficient ρ of X and Y is given by

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}, \tag{3}$$

where, given the means $\bar{X}$ and $\bar{Y}$ of the random variables, $\mathrm{cov}(X, Y)$ is the covariance of the two random variables given by

$$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \bar{X})(Y - \bar{Y})], \tag{4}$$

with $\mathbb{E}$ being the expected value of a random variable, and σ_X and σ_Y their standard deviations given by

$$\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{5}$$

Note that, by definition, $-1 \leq \rho \leq 1$, and we generally consider the absolute value $|\rho|$ to quantify the degree of linear relation; a value $|\rho| \sim 1$ indicates a perfect linear linear relation between the random variables, whereas $|\rho| \sim 0$ indicates no linear relation.

By considering features pairs with high correlation value ρ, the Pearson correlation coefficient can be used to find universally related feature pairs that show a sufficiently strong linear relation. Typically, the threshold for the Pearson correlation coefficient has to be chosen to be high in order to avoid too many false positives, and as such only feature pairs that show a high degree of linear relation will be identified as universally related. The exact choice of this threshold depends highly on the use case, and we will explore this issue further in the following sections.

B. Measures of nonlinear relation

We denote the relation of any feature pair that shows some type of universal relation, but that is not classified as linearly related by using the Pearson correlation coefficient, as nonlinear. Our hypothesis for this work is that by utilizing suitable measures that quantify some type of (not necessarily linear) relation between two random variables, we will be able to identify such nonlinearly related features.

1. Distance correlation

The first measure that we consider is distance correlation (DistCor), which was specifically introduced as a generalization of Pearson correlation to identify pairs of random variables that show any kind of linear or nonlinear relations. The distance correlation [15] dCor of two random variables X and Y is defined similarly to the Pearson correlation by

$$\mathrm{dCor}(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dVar}(X)\mathrm{dVar}(Y)}}, \tag{6}$$

where, by definition, we have that $0 \leq \mathrm{dCor} \leq 1$, and $\mathrm{dCor} = 0$ if and only if X and Y are statistically independent.

For distance correlation, the standard notions of covariance and standard deviation are replaced by sample distance covariance dCov and distance standard deviation dVar. Similar to covariance and standard deviation, which are computed based the distance of each sample from the means of the random variables, dCov and dVar denote quantities that are instead based on the *pairwise* distance of all samples to each other as well as the sample means. As their definitions are slightly longer, and do not necessarily provide critical insight required for the rest of the paper, we refer to the original publication [15] for their full definitions.

2. Mutual information

The second measure that we consider is mutual information (MI), which is an information-theoretic quantity that measures how much we can learn about one random variable Y by having knowledge of another random variable X (or vice versa), and is zero exactly when the two distribution are independent (i.e., knowledge about X does not tell us anything about Y). As a quantity, it measures how many bits can be saved if we try to binary encode Y while assuming knowledge of X (in contrast to binary encoding Y on its own without any further knowledge).

The mutual information [16] $I(X;Y)$ of two random variables X and Y is given by

$$I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}, \tag{7}$$

where P_{XY} is the joint probability distribution of X and Y given by

$$P_{XY}(x,y) = Pr[X = x, Y = y], \tag{8}$$

and P_X and P_Y are the marginal distributions given by

$$P_X(x) = \sum_y P_{X,Y}(x,y). \tag{9}$$

By definition, we only have that $0 \leq I(X;Y)$, i.e., this measure is not normalized to a range between 0 and 1. While there exist variations that transform mutual information into a metric within this domain, we do not strictly require these properties for our use case. As such, we decided to remain with this basic definition. The notion of maximal information discussed below presents an extension that is normalized to 1.

Technically, the above definition for mutual information is for discrete variables, while our use case is centered around continuous random variables. However, in practice, the data vectors we use are discrete, and computational methods can be used to estimate the continuous sample distribution from the actual, discrete sample vectors. In this paper, we rely on the `MUTUAL_INFO_REGRESSION` method implemented in the SKLEARN Python package.

3. Maximal information

The measure of maximal information (MaxI) [17] is a direct extension of mutual information to continuous variables. It is based on the binning-based method to estimate the mutual information where continuous random variables are discretized by transforming them into histograms of some fixed bin size. Instead of arbitrarily choosing this bin size, leading to varying degrees of accuracy in your mutual information estimate, maximal information computes the mutual information for a series of histograms of varying bin sizes and finally chooses the binning that maximizes the mutual information. That is, the maximal information coefficient (MIC) of two random variables X, Y is given by

$$\mathrm{MIC}(X;Y) = \max_B \frac{I(X;Y)}{\log_2(\min(B_X, B_Y))}, \tag{10}$$

where B is the total number of used bins (typically with some upper bound, cf. [17]), and B_X and B_Y are the number of bins used for X and Y, respectively. By definition, $0 \leq \mathrm{MIC}(X;Y) \leq 1$, where a value of 0 indicates statistical independence of the random variables, whereas a value of 1 indicates a strong relation. To compute the maximal information between two vectors, we will utilize the MINEPY package for Python [20].

C. Comparison of measures

As mentioned above, the main issue with the more prominent Pearson correlation measure is that it only identifies linearly related features. While we can adjust to this to some degree by computing some function values of our features (i.e., computing some polynomials or exponential function on the features values), this can become fairly cumbersome in practice. In recent years,

especially with the advent of big data and the necessity of finding nonlinear correlations in various applications, the other above mentioned correlation measures have been developed [15,17]. The main idea behind them is that instead of looking for a global, linear correlation, they instead approximate global correlation by finding local (linear correlation), i.e., correlation of data points that are in close proximity, and generalize it over the whole dataset. This applies to both distance correlation, which to some degree generalizes the Pearson correlation in such a manner, and maximal information, which directly generalizes the measure of mutual information.

A similar comparison has already been performed in the past by Clark [21]. They find that, in particular for nonlinear relations, distance correlation and mutual/maximal information outperform Pearson correlation in identifying correlated variables. Our purpose for this work is to verify that the same observations can also be made for the use case of finding universal relations in neutron star model data, and evaluate which correlation measure indeed performs best for this use case.

D. General methodology

Our general approach to evaluating the different correlation measures introduced above, and also for later automatically finding bivariate universal relations, is the following:

(1) Obtain neutron star model data with features $F_1, \ldots, F_n$ from theoretical/numerical computations.
(2) Compute the pairwise correlation of all feature pairs using on of the above correlation measures. This provides us with the correlation matrix $\mathbf{M}$, where the entry $\mathbf{M}_{i,j}$ is the correlation between features F_i and F_j.
(3) Specify a correlation threshold τ above which we will consider feature pairs correlated, i.e., find all entries in $\mathbf{M}$ with

$$\mathbf{M}_{i,j} \geq \tau. \tag{11}$$

This threshold will depend on the correlation measure used, and finding the best value for it is something we want to achieve here, but might need to be further explored in future work.
(4) For each selected feature pair, choose a suitable model. Here, model denotes the expected functional relation between the two selected features. This can be, e.g., a linear, polynomial, exponential model, etc. Model selection is a notoriously difficult task in data analysis, and we will here simply choose to evaluate a number of preset templates for the functional relations, and choose the one with the best fit after the following step.
(5) Fit the model to the given data to determine the coefficients of the best fit for the universal relation.

IV. BIVARIATE UNIVERSAL RELATIONS

In the following, we inspect the universal relations found by the correlation measures we discussed in the previous section. For each relation, we will also indicate the correlation value obtained by each respective measure. This will allow us to inspect in which cases each of the correlation measures succeed or fail in correctly identifying features that are suited for universal relations.

Since the features we correlate cover very different ranges of values, we will evaluate the quality of each proposed universal relation through the average relative error $\bar{e}$ given by

$$\bar{e} = \frac{1}{n}\sum_i \frac{|\hat{y}_i - y_i|}{|y_i|}, \tag{12}$$

where $\hat{y}_i$ is the value predicted by the universal relation, and y_i the actual data point.

In some cases, our automated approach will find an exponential relation between two feature that we are analyzing. We find that by instead fitting for the logarithm of the target feature we achieve better universality. In such cases, after performing the correlation analysis on the regular features, we therefore manually fit a polynomial relation between the logarithm of the target feature and the independent feature. Note that the correlation values, however, will still be given between the regular features, and not after applying the logarithm, as this is how the features are fed into the automatic method described in Sec. III.

As discussed in Sec. II A, we derive relations independently for both datasets to demonstrate to some degree that our approach generalizes across different data. For each relation, we will indicate form which dataset it specifically was derived. In most cases, the choice of data set for a given relation was predicated by the features available within each dataset (cf. Table I).

A table summarizing all universal relations presented in this section can be found in the Conclusion (Sec. VII).

A. Tidal deformability relations

In Fig. 2 we show a universal relation between the normalized tidal deformability $\bar{\lambda}$ and the normalized moment of inertia $\bar{I}$, derived from dataset A. This relation was also previously put forward by Yagi and Yunes [8] as part of their *I-Love-Q* relations. The best fit for this relation is given by the function,

$$\bar{I} = 0.019 \log \bar{\lambda}^2 - 0.076 \log \bar{\lambda} + 0.334. \tag{13}$$

This relation achieves an average relative error of 0.020.

In Fig. 3 we show a universal relation between the effective compactness η and the logarithm of the normalized tidal deformability $\log \bar{\lambda}$, derived from dataset A.

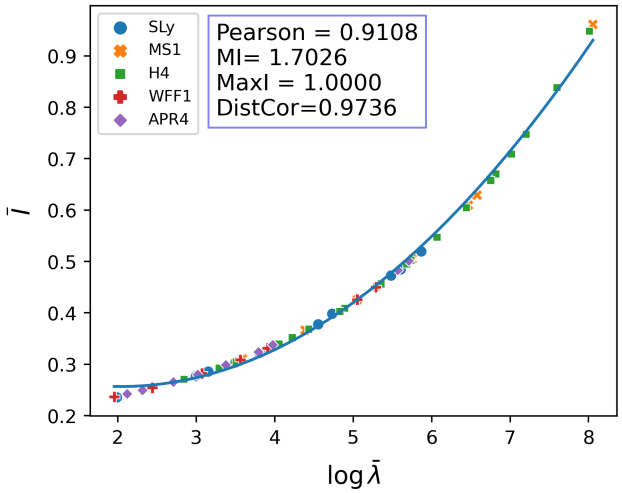

FIG. 2. Universal relation between the normalized tidal deformability $\bar{\lambda}$ and normalized moment of inertia $\bar{I}$, derived from dataset A and as given by Eq. (13) (cf. Yagi and Yunes [8]).

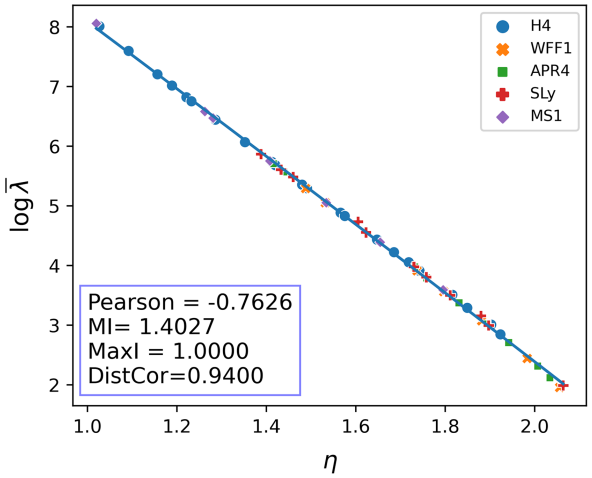

FIG. 3. Universal relation between normalized tidal deformability $\bar{\lambda}$ and effective compactness η, derived from dataset A and as given by Eq. (14). The correlation values are given for the correlation between η and $\bar{\lambda}$.

A similar relation was also previously proposed by us in the context of a binary neutron star merger connecting the premerger binary tidal deformability to the postmerger effective compactness [12].

This is a case in which the automated approach yields an exponential relation between η and $\bar{\lambda}$, and as discussed above, we manually fit a polynomial relation for $\log \bar{\lambda}$, yielding the relation,

$$\log \bar{\lambda} = -0.093\eta^2 - 5.425\eta + 13.604. \tag{14}$$

This relation achieves an average relative error of 0.008. In this case, the originally exponential relation between the two features causes the Pearson correlation coefficient in particular to give a comparatively low correlation value of −0.763. In comparison, the other correlation measures still assign a fairly high correlation value.

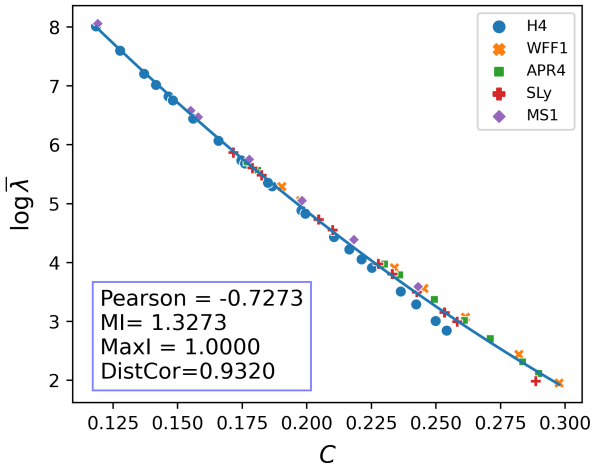

FIG. 4. Universal relation between the logarithm of the tidal deformability $\log \bar{\lambda}$ and compactness C, derived from dataset A and as given by Eq. (16). The correlation values are given for the correlation between C and $\bar{\lambda}$.

In Fig. 4 we show a universal relation between the compactness C and the logarithm of the normalized tidal deformability $\log \bar{\lambda}$, derived from dataset A. Such a relation was put forward previously by Jiang and Yagi [22], and follows directly from the definition of $\bar{\lambda}$ in terms of the tidal Love number k_2, i.e.,

$$\bar{\lambda} = \frac{\lambda}{M^5} = \frac{2}{3} k_2 \frac{R^5}{M^5} = \frac{2}{3} k_2 C^5. \tag{15}$$

The automatic approach again finds an exponential relation between the features C and $\bar{\lambda}$, and as before, we find that fitting for $\log \bar{\lambda}$ instead yields the more accurate, universal relations. The manual fit yields the relation

$$\log \bar{\lambda} = 46.123C^2 - 53.045C + 13.633. \tag{16}$$

This relation achieves a relative error of 0.020. As before, the regular features have a nonlinear, exponential relation for which the Pearson correlation measure again assigns a comparatively low correlation value of −0.727, while the other correlation measures still assign high correlation values.

B. Astroseismological relations

We here present some of the astroseismological, universal relations we were able to find for the f-mode and g-mode oscillation frequencies.

In Fig. 5 we show a universal relation between the compactness C and the normalized f-mode frequency $\bar{M}\omega_f$, derived from dataset B. This relation was previously

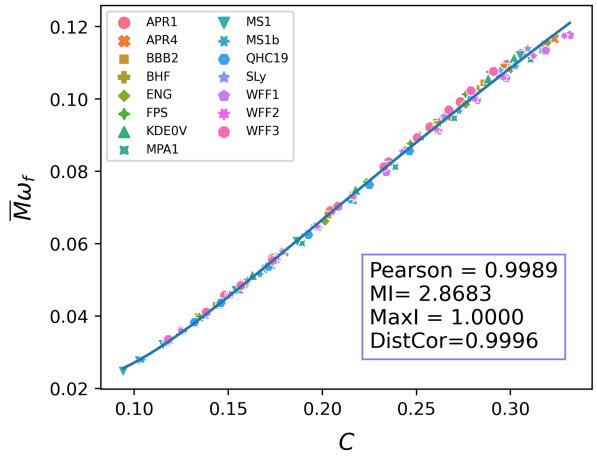

FIG. 5. Astroseismological relation between the normalized f-mode frequency $\bar{M}\omega_f$ and the compactness C, derived from dataset B and as given by Eq. (17) (cf. Tsui and Leung [5]).

put forward by Tsui and Leung [5]. The best fit for this relation is given by the function

$$\bar{M}\omega_f = 0.042 \log C^2 + 0.222 \log C + 0.315. \quad (17)$$

This relation achieves an average relative error of 0.011. While the optimal fit is given by a logarithmic relation, visually the relation can still be fit fairly well by a linear function. As expected, in this case even the Pearson correlation coefficient assigns a high value of 0.999, and the other correlation measures also identify a strong relation between these two features.

In Fig. 6 we show a universal relation between the normalized moment of inertia $\bar{I}$ and the normalized f-mode frequency $\bar{M}\omega_f$, derived from dataset A. This relation follows straight-forwardly by combining the relation by Tsui and Leung [23] between the f-mode and compactness

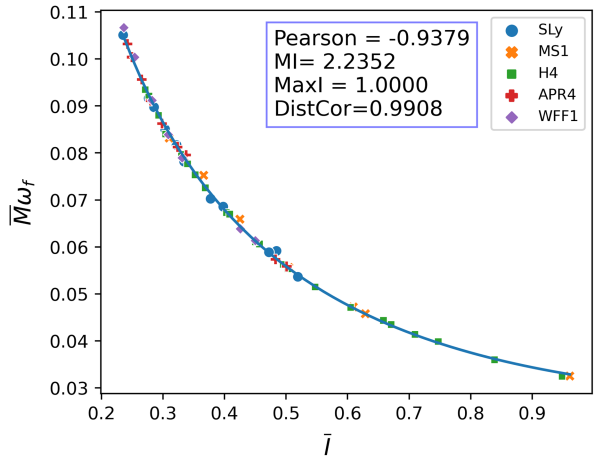

FIG. 6. Astroseismological relation between the normalized f-mode frequency $\bar{M}\omega_f$ and the normalized moment of inertia $\bar{I}$, derived from dataset A and as given by Eq. (18).

C, with the understanding that the compactness C and effective compactness η can often be used interchangeably in such general relativistic relations. However, to our knowledge, this is the first time that this relation is presented explicitly.

The best fit for this relation is given by the function,

$$\bar{M}\omega_f = 0.021 \log \bar{I}^2 - 0.020 \log \bar{I} + 0.032. \quad (18)$$

This relation achieves an average relative error of 0.007. The best fit is given by a logarithmic relation, and visually the relation also does not seem to allow a good fit by a linear function. Still, in this case, the Pearson correlation still applies a comparatively high value of −0.938. The remaining measures also identify a strong relation between the features.

Figure 7 shows a universal relation between the normalized tidal deformability $\bar{\lambda}$ and the normalized f-mode frequency $\bar{M}\omega_f$, derived from dataset B. This relation was also previously put forward by Chan *et al.* [7]. The best fit for this relation is given by the function,

$$\bar{M}\omega_f = 0.0003 \log \bar{\lambda}^2 - 0.015 \log \bar{\lambda} + 0.127. \quad (19)$$

This relation achieves an average relative error of 0.014. The highly nonlinear, logarithmic relation between these features causes the Pearson correlation coefficient to fail to detect the correlation between these features, assigning a value of −0.612 and even the distance correlation assigns a comparatively small correlation value of 0.911, compared to the previous relations.

Figure 8 shows a universal relation between the effective compactness η and the normalized f-mode frequency $\bar{M}\omega_f$, derived from dataset A. This relation was also previously

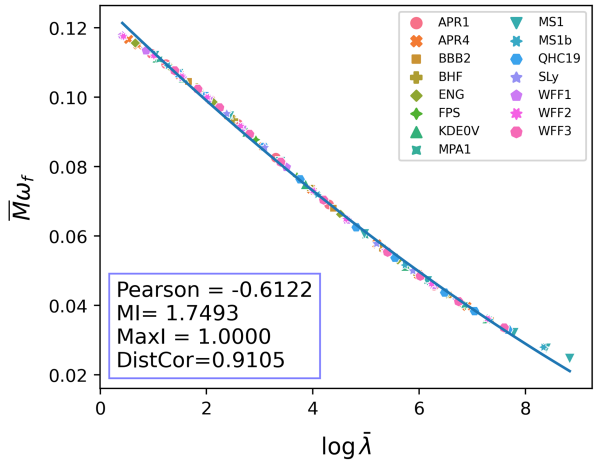

FIG. 7. Astroseismological relation between the normalized f-mode frequency $\bar{M}\omega_f$ and normalized tidal deformability $\bar{\lambda}$, derived from dataset B and as given by Eq. (19) (cf. Chan *et al.* [7]).

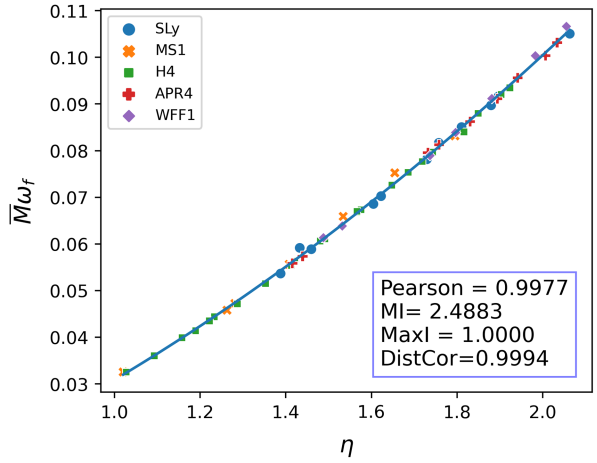

FIG. 8. Astroseismological relation between the normalized f-mode frequency $\bar{M}\omega_f$ and the effective compactness η, derived from dataset A and as given by Eq. (20) (cf. Lau *et al.* [6] and Krüger and Kokkotas [19]).

put forward in [6,19,24]. The best fit for this relation is given by the function,

$$\bar{M}\omega_f = 0.015\eta^2 + 0.025\eta - 0.009. \tag{20}$$

This relation achieves an average relative error of 0.007. Visually, this relation again appears to also allow a good fit through a linear function, which is reflected by all correlation measures (including Pearson correlation) assigning a high correlation value.

Figure 9 shows a universal relation between the average density $\tilde{\rho}$ and the f-mode frequency ω_f, derived from dataset B. This relation was also previously put forward by Andersson and Kokkotas [3,4] and Benhar *et al.* [25]. The best fit for this relation is given by the function

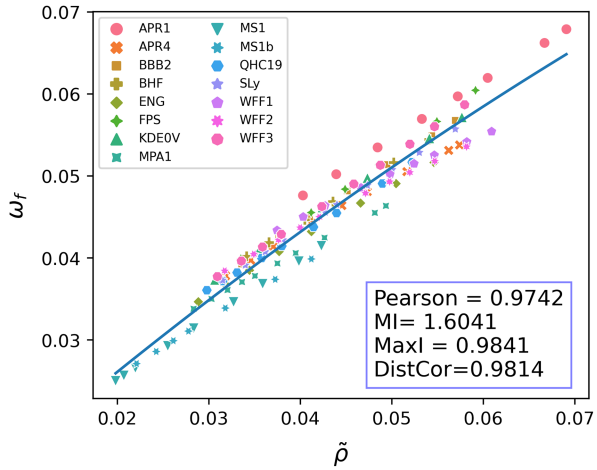

FIG. 9. Astroseismological relation between the f-mode frequency ω_f and the square-root of the average density $\tilde{\rho}$, derived from dataset B and as given by Eq. (21) (cf. Andersson and Kokkotas [3,4]).

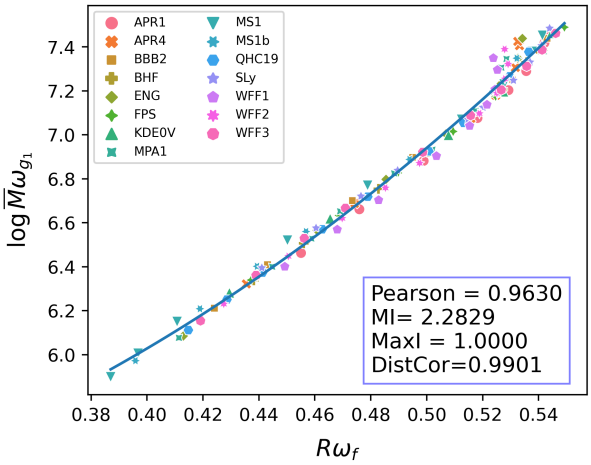

FIG. 10. Astroseismological relation between the normalized f-mode frequency $R\omega_f$ and the logarithm of the normalized g-mode frequency $\log \bar{M}\omega_{g_1}$, derived from dataset B and as given by Eq. (22) (cf. Kuan *et al.* [18]). The correlation values are given for the correlation between $R\omega_f$ and $\bar{M}\omega_{g_1}$.

$$\omega_f = -2.199\tilde{\rho}^2 + 0.985\tilde{\rho} + 0.007. \tag{21}$$

This relation achieves an average relative error of 0.035. Again, the fact that this relation appears to be mostly linear is reflected in the fact that all correlation measures assign a fairly high correlation value to these two features.

Figure 10 shows a universal relation between the normalized f-mode frequency $R\omega_f$ and the logarithm of the normalized g-mode frequency $\log \bar{M}\omega_{g_1}$, derived from dataset B. This relation was also previously put forward by Kuan *et al.* [18]. As was the case for the relations in Eqs. (14) and (16), the automatic method finds an exponential relation between the features $R\omega_f$ and $\bar{M}\omega_{g_1}$. As before, we find that manually fitting the relation for the logarithm $\log \bar{M}\omega_{g_1}$ gives the more accurate relation, yielding

$$\log \bar{M}\omega_{g_1} = 16.052(R\omega_f)^2 - 5.323R\omega_f + 5.58908468. \tag{22}$$

This relation achieves an average relative error of 0.004. Even though the automatic method finds an exponential function to be the best fit between the original features, visually it is apparent that the relation could, to some degree, also be fit by a linear function. As such, even the Pearson correlation coefficient achieves a fairly high correlation value, however notably lower than the other correlation measures.

C. Quantitative comparison of correlation measures

We can perform a more quantitative analysis and comparison of the four different correlation measures by considering some specific performance measures

commonly used in statistics. To define these, we first introduce a few notions for binary classifiers. We define them here in terms of our use case of identifying universally related neutron star features; A *true positive* is a pair of features that is universally related, and also identified as such by a given correlation measure. The number of true positives is denoted by TP.

A *false positive* is a pair of features that is *not* universally related, but classified as such. The number of false positives is denoted by FP.

A *false negative* is a pair of features that is universally related, but not classified as such. The number of false negatives is denoted by FN.

A *true negative* is a pair of features that is not universally related, and also not classified as such. The number number of true negatives is denoted by TN.

Given these notions, we can now define performance measures that quantify how well our classifiers correctly label pairs of features. Recall, or *true positive rate* (TPR) is the rate at which the classifier correctly labels universally related pairs of features as universally related. It is given by

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{23}$$

Precision, or positive predictive value (PPV), is the rate of pairs of features classified as universally related that are in fact universally related. It is given by

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{24}$$

Finally, the fallout, or false positive rate (FPR), is the rate at which not related pairs of features are classified as being universally related. It is given by

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{25}$$

We can now compute the precision, recall and fallout for each correlation measure at a given classification threshold τ, and compare how these performance measures develop with τ. Ideally, we would like to achieve high recall, while keeping precision high, and fallout low.

Typically, one considers the Precision/Recall and ROC curves for a better understanding on how these quantities evolve with each other. The Precision/Recall curves plot the maximum precision achieved by a classifier for a required recall, and allow us to understand how accurate a positive prediction (i.e., classification as universal relation) is, given a specific correlation measure and classification threshold. We show the Precision/Recall curves for each correlation measure, and one combined plot, in Fig. 11.

The ROC (or receiver operating characteristic) curve plots the recall against the fallout of the classifier. This plot allows us to better understand how many incorrectly classified universal relations we should expect for a given recall requirement. The ROC curves for each correlation measure applied to each of the two datasets considered in this work can be found in Fig. 12.

As we can see in all figures, the standard Pearson correlation measure is outperformed by the other metrics significantly for most of the recall range. The distance correlation measures, in turn, is also outperformed by the mutual information based measures. Both the kernel-based mutual information measure, and the maximal information measure show high precision and low fallout for high recall values, identifying them as the preferred measures for the task of identifying universally related features.

Note that the above analyses were performed by manually labeling all feature pairs in our limited dataset as

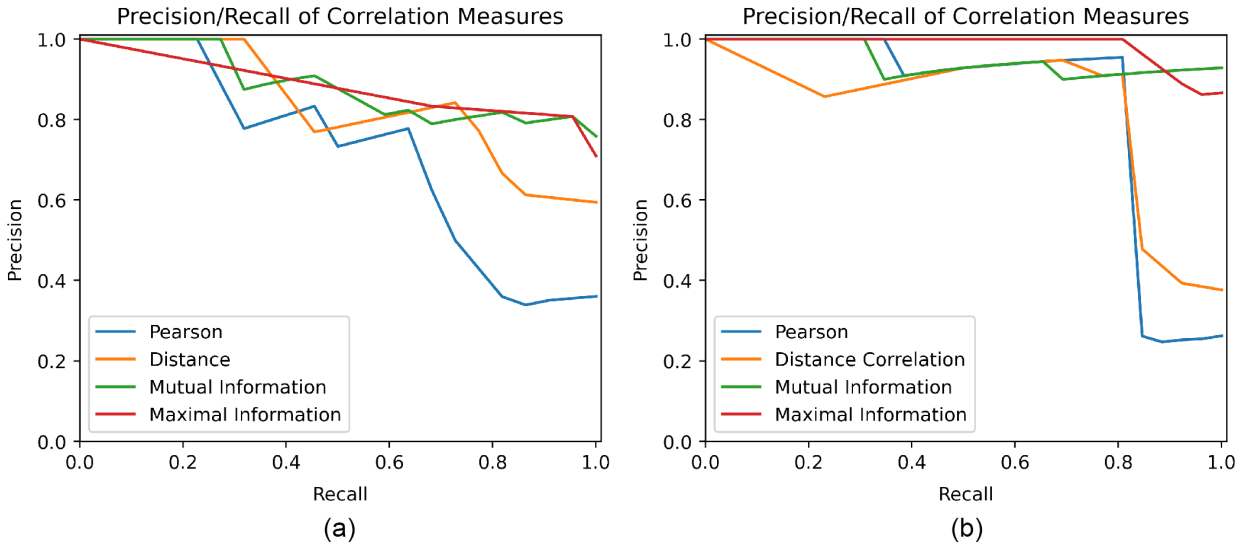

FIG. 11. Precision/Recall curves for each correlation measure. (a) Precision/Recall curves for dataset A. (b) Precision/Recall curves for dataset B.

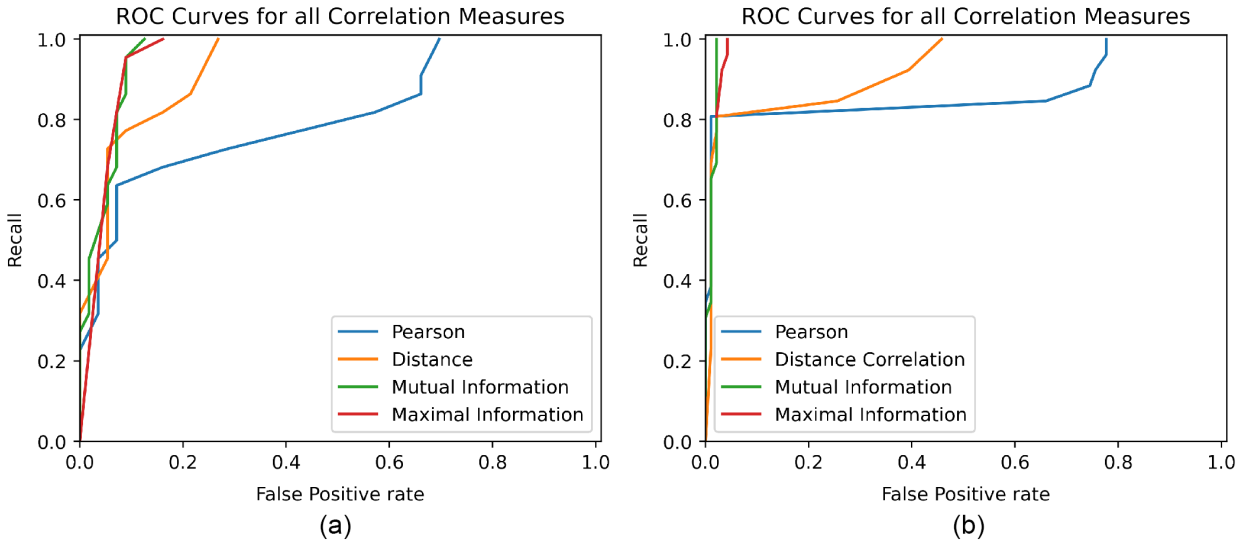

FIG. 12. ROC (Receiver operating characteristic) curves for each correlation measure. (a) ROC curves for dataset A. (b) ROC curves for dataset B.

either being universally related or not in order to obtain the true/false positive/negative counts. As such, the exact values for each performance measure will most likely vary with different datasets and labels. However, the difference in behavior of each correlation measure appears to be significant enough to warrant the conclusion drawn above.

D. Ideal classification thresholds

Closer inspection of both the Precision/Recall and the ROC curves can also provide hints for an ideal classification threshold that optimizes the trade-off between recall and precision/false positive rate. This can be done by, e.g., considering thresholds at which precision or false positive rate show drastic changes. This threshold will, however, depend on whether one prioritizes higher recall over precision/false positive rates, or vice versa. For our use case of finding universally related features, we would naturally prefer achieving a higher recall value as false positives can usually be discarded after a simple visual inspection of the corresponding plots. As an example, in Tables II and III, we list the classification thresholds manually selected from the data points of the respective Precision/Recall and ROC curves.

Determining a generally optimal classification threshold, however, will require further detailed analysis that is out of the scope of this work as the optimal threshold can often depend on the exact application and data that is used.

V. MULTIVARIATE CORRELATION ANALYSIS

Until now, we have only considered the functional relation between two features, and tried to find such pairs of features that allow for universal relations across different equations of state. A straightforward extension then, of course, is to look for multivariate relations, i.e., such relations where one predicted/target features is described in terms of a function that depends on more than one explanatory feature.

The field of high-dimensional data analysis is a widely studied field that in particular gained a lot of notoriety in recent times due to the advent of the big data paradigm. While many different approaches, theories and methods have been developed to deal with high dimensionality in data, we will here consider one very prominent method; principal component analysis (PCA). PCA is a dimensionality reduction and feature extraction technique that has been used to great success in various data analysis use

TABLE II. Classification thresholds τ and corresponding performance measures derived from the Precision/Recall [cf. Fig. 11(a)] and ROC curves [cf. Fig. 12(a)] for dataset A.

Measure	τ	Recall	Precision	FPR
Pearson	0.909	0.682	0.625	0.161
DistCor	0.939	0.864	0.613	0.214
MI	1.188	0.955	0.808	0.089
MaxI	0.960	0.955	0.808	0.089

TABLE III. Classification thresholds τ and corresponding performance measures derived from the Precision/Recall [cf. Fig. 11(b)] and ROC curves [cf. Fig. 12(b)] for dataset B.

Measure	τ	Recall	Precision	FPR
Pearson	0.939	0.808	0.955	0.011
DistCor	0.960	0.808	0.913	0.021
MI	1.333	1.000	0.929	0.021
MaxI	0.949	1.000	0.867	0.043

cases [26]. Recently, Soldateschi *et al.* [14] utilized PCA to construct multivariate universal relations for magnetized neutron stars. Here, we will investigate how we can apply PCA in general to identify potential universal relations, and evaluate how well this approach performs on our own data.

A. Finding multivariate correlation using PCA

The main purpose of PCA is to identify the principal directions in which a given dataset varies the most; the principal components $\mathbf{A}$ of a dataset consisting of a set of features $\mathbf{F}$ (i.e., the dimensions of the dataset's underlying vector space) are a sequence of vectors (called principal component) $\mathbf{A_i}$ in the space span by $\mathbf{F}$ which:

(1) are orthogonal to all previous principal components $\mathbf{A_0}, \ldots, \mathbf{A_{i-1}}$; and
(2) show in the direction of the line that best fits the data set (using least squares regression).

As a consequence, the principal component A_0 shows the direction that maximizes the variance within the dataset, while each subsequent principal component covers less and less variance of the dataset. One can then choose the m first principal components as the basis vectors of a lower-dimensional vector space into which the dataset can be projected while retaining most of the variance (i.e., information) within the dataset.

While the general use case of PCA does not directly match our goal of constructing universal relations, we can make use of the properties of the principal components to potentially find multivariate universal relations. Note that each principal component $\mathbf{A_i}$ represents a linear combination of the features $\mathbf{F}$ that assigning a weight $a_{i,j}$ to each feature F_j, which, in this context, is also called the *loading* of F_j in $\mathbf{A_i}$. After computing all principal components, we try to identify those that have a proportionally large loading for our target feature F, if any such component exists; usually, if there are no strong correlations within our data that lead to a large variance for F, all principal components will show a comparatively small loading for F. However, in the case of a principal component that has a large loading for F, we might be able to leverage it to construct a universal relation. By projecting the considered features onto the identified principal component and solving for F, we potentially obtain a first-order multivariate universal relations.

Soldateschi *et al.* [14] claim that such universal relations should be found using the last principal component. In the following we will also investigate if this claim is true.

B. General methodology

We now describe the general methodology we follow for finding multivariate universal relations using PCA:

(1) Select a number of explanatory variables $F_1, \ldots, F_n$ and a target feature F.
(2) Perform PCA on the feature set

$$\mathbf{F} = \{F_1, \ldots, F_n, F\}. \tag{26}$$

(3) For each principal component $\mathbf{A_i}$, solve the equation

$$\begin{aligned} \mathbf{A_i} \cdot \mathbf{F} &= \sum_{j=1}^{n} a_{i,j} F_j + a_{i,n+1} F \\ \Rightarrow \qquad F &= -\left(\sum_{j=1}^{n} \frac{a_{i,j}}{a_{i,n+1}} F_j \right) \\ \Rightarrow \qquad F &= \hat{a}_1 F_1 + \cdots + \hat{a}_n F_n \end{aligned} \tag{27}$$

with

$$\hat{a}_j = -\frac{a_{i,j}}{a_{i,n+1}}, \tag{28}$$

where we denote the right hand side as the new combined feature $\hat{F}$

$$\hat{F} = \hat{a}_1 F_1 + \cdots + \hat{a}_n F_n. \tag{29}$$

(4) Evaluate whether there exists a strong correlation between F and a combined feature $\hat{F}$ using bivariate correlation analysis.
(5) If a strong correlation is found, choose a suitable model and fit it for the relation between F and $\hat{F}$.

In contrast to the bivariate case, this approach cannot be fully automated yet. A lot of guesswork is involved in identifying the principal components from which we can derive suitable combined features. The most straightforward approach for this task is to simply construct the combined feature for all principal components and then perform a bivariate correlation analysis of the target feature F with each found combined feature.

Also, this method will not always yield universal relations: sometimes, there will be no principal component that will suitably explain the variance in the target feature F. This might happen in cases where (a) F simply does not present much variance across the whole dataset, or (b) there exist many colinearities within the selected set of features $\mathbf{F}$. We discuss some cases where the method described above does not yield a universal relation in Appendixes B and C.

VI. MULTIVARIATE UNIVERSAL RELATIONS

We present the results of using PCA to find multivariate universal relations for neutron stars as described in the previous section. A table summarizing all universal relations presented in this section can be found in the Conclusion (Sec. VII).

TABLE IV. Loadings of features in each principal component obtained from performing PCA on the feature set $\mathbf{F} = \{M, R, C, \bar{\lambda}\}$ on dataset A.

Component	M	R	C	$\bar{\lambda}$
0	−0.488	0.419	−0.571	0.511
1	0.596	0.793	−0.034	−0.119
2	0.322	−0.097	0.412	0.847
3	0.550	−0.431	−0.710	0.086

A. Multivariate universal relations for tidal deformability

We here consider the case where we want to construct a universal relation for the normalized tidal deformability $\bar{\lambda}$, using the features M, R, and C. To this end, we perform the principal component analysis on all four features using dataset A (cf. Sec. II). The resulting principal components are given in Table IV by means of the loading of each feature within the principal components. A visual representation of the combined feature obtained from each principal component is shown in Appendix A.

As we can see in Table IV, the target feature $\bar{\lambda}$ has the largest loading for principal component 2, with component 0 also showing a relatively large loading of $\bar{\lambda}$. Performing the bivariate correlation analysis of the combined features derived from each principal component with the target feature $\bar{\lambda}$ shows that the best correlation is actually given by principal component 0. However, through the visual inspection of the combined features, as depicted in Appendix A, we can see that component 2 could also be leveraged for a universal relation, albeit with a larger error. For the remainder of the text we will focus on the relation with the lesser error, induced by component 0.

Since the automated approach finds an exponential relation between $\bar{\lambda}$ and the combined feature, and we again fit for $\log\bar{\lambda}$ to obtain a more accurate relation. Through our manual fit, we obtain the following universal relation for the normalized tidal deformability,

$$\log\bar{\lambda} = -0.635\hat{F} + 7.399 \tag{30}$$

with

$$\hat{F} = 3.391\frac{M}{M_\odot} - 5.241\frac{R}{10\ \mathrm{km}} + 4.768\frac{C}{0.2}. \tag{31}$$

This relation is presented in Fig. 13 and achieves an average relative error of 0.023. Compared to the bivariate relation between the tidal deformability and compactness we presented in Fig. 4, we essentially introduce a linear order correction involving the radius and the mass. While the overall relative error is approximately the same as for the bivariate relation, the multivariate relation remains entirely linear in all independent variables, reducing its sensitivity to potential estimation errors for these quantities.

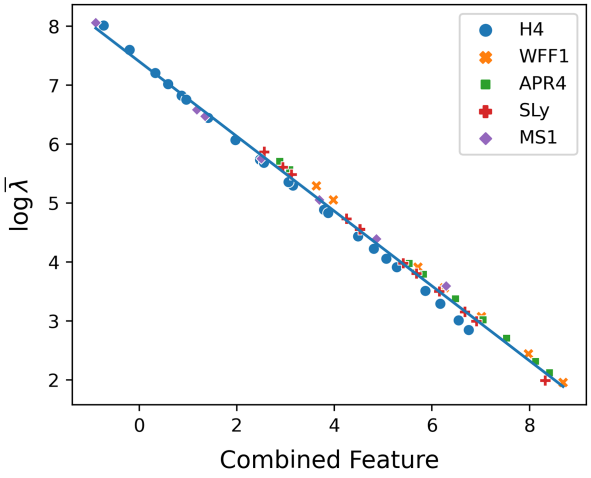

FIG. 13. Universal relation between logarithm of the normalized tidal deformability $\log\bar{\lambda}$ and the combined feature given in Eq. (31), derived from principal component 0 (cf. Table IV), using dataset A [12,19].

B. Relation with dataset B

We also perform the same analysis using the data by Kuan *et al.* [18]. The principal components obtained from the PCA are listed in Table V. The principal components show a similar behavior to the previous examples using dataset A, however we can observe some slight differences caused by the different equations of state used in the dataset.

As before, after performing the bivariate correlation analysis on the combined features derived from each principal component, we find that the combined feature derived from principal component 0 shows the best universality. Leveraging this component, we obtain the universal relation

$$\log\bar{\lambda} = -0.939\hat{F} + 6.521 \tag{32}$$

this time with the combined feature

$$\hat{F} = 2.249\frac{M}{M_\odot} - 4.316\frac{R}{10\ \mathrm{km}} + 3.533\frac{C}{0.2}. \tag{33}$$

TABLE V. Loadings of features in each principal component obtained from performing PCA on the feature set $\mathbf{F} = \{M, R, C, \bar{\lambda}\}$ on dataset B.

Component	M	R	C	$\bar{\lambda}$
0	−0.514	0.348	−0.593	0.513
1	0.555	0.799	0.145	0.182
2	0.127	−0.352	0.406	0.834
3	−0.642	0.342	0.681	−0.089

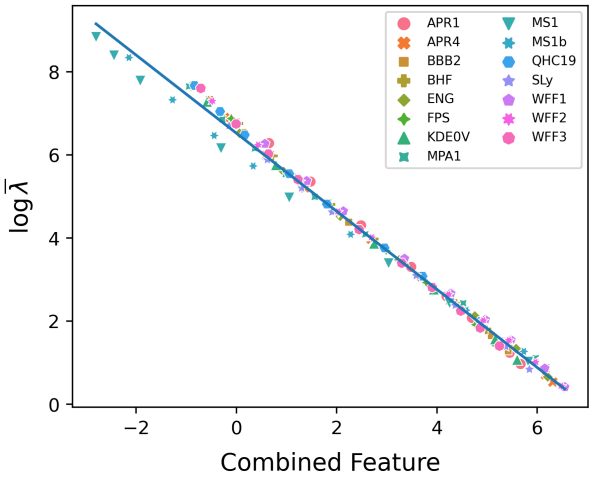

FIG. 14. Universal relation between the logarithm of the normalized tidal deformability $\log \bar{\lambda}$ and the combined feature given in Eq. (33), derived from principal component 0 (cf. Table V), using dataset B [18].

The resulting best fit is presented in Fig. 14. It achieves an average relative error of 0.043, which is slightly higher than what we achieved for dataset A. We suspect this is caused by some of the outlying neutron star models that are introduced by the larger configuration space considered in dataset B.

However, the fact remains that our approach for the multivariate correlation analysis yields the same form for the universal relation independent of which dataset is used. This is indicative of this approach further generalizing well for different datasets, and that the results presented here are not dependent on the underlying data used for the analysis.

C. Multivariate astroseismological relations

Andersson and Kokkotas [3,4] previously proposed a universal relation linking the average density $\tilde{\rho}$ to the f-mode frequency of a neutron star. We here attempt to apply the same method as above to potentially find corrections to their original astroseismological relation that improve its universality. To this end, we perform the principal component analysis on the features ω_f, M, C, and $\tilde{\rho}$, aiming at finding corrections in terms of M and C for the universal relation.

The best relation is found for the combined feature derived from the fourth principal component found through PCA performed in the feature set $\mathbf{F} = \{M, C, \tilde{\rho}, \omega_f\}$. The best fit for the relation between ω_f and this combined feature is shown in Fig. 15. The best fit shows a quadratic universal relation for the f-mode frequency of the form,

$$\omega_f = -0.00033\hat{F}^2 + 0.013\hat{F} - 0.023 \tag{34}$$

with

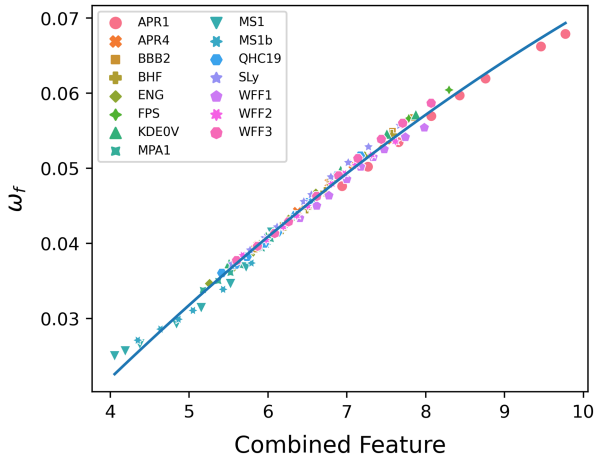

FIG. 15. Universal relation for the f-mode frequency ω_f using the combined feature of M, $\tilde{\rho}$, and C given in Eq. (35), obtained from the PCA on dataset B [18].

$$\hat{F} = 2.980\frac{M}{M_\odot} + 10.231\frac{\tilde{\rho}}{0.04} - 8.398\frac{C}{0.2}. \tag{35}$$

This relation achieves an average relative error of 0.015. When compared to the old relation shown in Fig. 9, we can clearly observe an improved universality, which is also reflected in the average relative error that is reduced by half. We thus achieve a significant improvement over the existing relation by using our multivariate approach.

D. Improved astroseismological relations for the f-mode frequency

We next consider another variation on the astroseismological relation we inspected above. This time, instead of introducing mass and compactness as independent variables, we instead only introduce the product $C\tilde{\rho}$ of compactness and average density as a new independent variable. Our goal now is therefore to find a universal relation for ω_f using the average density $\tilde{\rho}$ and $C\tilde{\rho}$.

In this case, the best relation is found for the combined feature derived from the third principal component found through the PCA performed in the feature set $\mathbf{F} = \{\tilde{\rho}, C\tilde{\rho}, \omega_f\}$. The best fit for the relation between ω_f and this combined feature is shown in Fig. 16. The best fit shows a quadratic universal relation for the f-mode frequency of the form,

$$\omega_f = 0.0002\hat{F}^2 + 0.006\hat{F} + 0.003 \tag{36}$$

with

$$\hat{F} = 6.911\frac{\tilde{\rho}}{0.04} - 1.716\frac{C\tilde{\rho}}{0.01}. \tag{37}$$

When compared to the relation shown in the section above (cf. Fig. 15) we observe an improved universality;

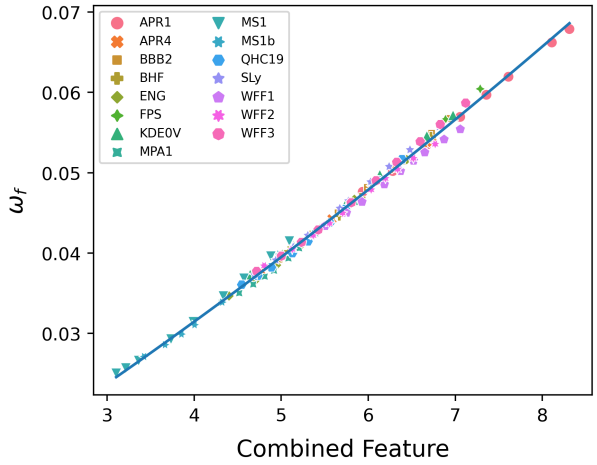

FIG. 16. Universal relation for the f-mode frequency ω_f using the combined feature of $\tilde{\rho}$ and $C\tilde{\rho}$ given in Eq. (37), obtained from the PCA on dataset B [18].

the previous relation has an average relative error of 0.015, whereas the relation with the new combined feature achieves an error of 0.010.

Considering that the original relation put forward by Andersson and Kokkotas [3,4] was inspired by Newtonian gravity, the additional factor in $C\tilde{\rho}$ could be considered a first order correction to account for general relativity, since

$$\hat{F} = 172.773\tilde{\rho} - 171.650C\tilde{\rho} \approx 172\tilde{\rho}(1 - C). \quad (38)$$

Essentially, this new relation is a stepping-stone between the relation by Andersson and Kokkotas [3,4], and other general relativistic universal relations, such as the one between the f-mode frequency ω_f and the compactness C put forward by Tsui and Leung [5].

E. Discussion of results

As we have demonstrated above, we can utilize the principal components obtained from PCA to construct multivariate universal relations for neutron stars. Since the relations we construct are, for now, first-order relations, this approach is also suited for finding first-order corrections to existing universal relations, allowing an improvement of the accuracy of the universal relations.

Despite these positive results, our approach here has only been descriptive: while we provide a methodology that can yield multivariate universal relations, the formal reasons for why this approach works is still not fully clear. Gaining further understanding of the mathematical underpinnings of this approach can allow us to further improve its output, but also better understand its limits.

For instance, our findings do not agree with the observations made in [14]; they claim that the best relations would follow from the last principal component obtained through the principal component analysis. In our findings, however, the best universal relations can appear from any of the principal components. It is therefore our belief that further analysis of the PCA method and the structure of its principal components is necessary to obtain a more rigorous understanding of this approach. Until then, the PCA approach should only be used to generate potential candidate relations that have to be further analyzed for their accuracy.

We have also seen that it is not always the principal component with the largest loading for our target feature that will induce the best universal relation: in the example discussed in Sec. VI A, we had two principal components with relatively large loadings for the target feature, but ultimately the combined feature derived from the principal component with the second largest loading offered the relation with the smallest error. A test of all possible combined features using the bivariate analysis therefore remains necessary until we potentially devise alternative criteria for deciding which principal component should be used to induce a universal relation.

Finally, in Appendixes B and C, we show some cases where our approach will not yield any universal relations. Sometimes this is caused by the data used, as, ultimately, not all feature combinations will be amenable to universal relations. Furthermore, specific properties of the used data, such as the existence of strong collinearities with the target feature, can also hinder our approach from producing universal relations. We currently can only provide superficial reasons for why our approach does not perform well in such situations, and we hope to obtain a more rigorous understanding through future work.

VII. CONCLUSION AND FUTURE DIRECTIONS

In this work, we discussed the potential of approaching the task of constructing universal relations for neutron stars from a statistical data analysis point of view. Instead of relying on physical intuition, our goal was to approach neutron star data using statistical methods only and thus enable a more automated approach to finding universal relations.

In a first step, we investigated the suitability of four different correlations measures for identifying pairs of features amenable to bivariate universal relations. We found that the usual Pearson correlation measure will have difficulties with nonlinear relations between features, which has also been observed in the past in the statistical data analysis literature for more general use cases [21]. Using generalized correlation measures that were explicitly constructed to detect non-linear correlations proved more useful; overall, mutual information and maximal information both performed best in finding universally related features, and while distance correlation did not perform as well as the aforementioned ones, it still outperformed Pearson correlation for our use case.

In a second step, we also approached the problem of constructing multivariate universal relations. Inspired by an

TABLE VI. List of all universal relations presented in this work.

Type	Features	Form	Average relative error	Equation	Figure	Reference
Bivariate	$\bar{\lambda}$, $\bar{I}$	$\bar{I} = 0.019 \log \bar{\lambda}^2 - 0.076 \log \bar{\lambda} + 0.334$	0.020	(13)	2	[8]
	$\bar{\lambda}$, η	$\log \bar{\lambda} = -0.093\eta^2 - 5.425\eta + 13.604$	0.008	(14)	3	[12]
	$\bar{\lambda}$, C	$\log \bar{\lambda} = 46.123C^2 - 53.045C + 13.633$	0.020	(16)	4	[22]
	$\bar{M}\omega_f$, C	$\bar{M}\omega_f = 0.042 \log C^2 + 0.222 \log C + 0.315$	0.011	(17)	5	[5]
	$\bar{M}\omega_f$, $\bar{I}$	$\bar{M}\omega_f = 0.021 \log \bar{I}^2 - 0.020 \log \bar{I} + 0.032$	0.007	(18)	6	…
	$\bar{M}\omega_f$, $\bar{\lambda}$	$\bar{M}\omega_f = 0.0003 \log \bar{\lambda}^2 - 0.015 \log \bar{\lambda} + 0.127$	0.014	(19)	7	[7]
	$\bar{M}\omega_f$, η	$\bar{M}\omega_f = 0.015\eta^2 + 0.025\eta - 0.009$	0.007	(20)	8	[6,19,24]
	ω_f, $\tilde{\rho}$	$\omega_f = -2.199\tilde{\rho}^2 + 0.985\tilde{\rho} + 0.007$	0.035	(21)	9	[3,4,25]
	$\bar{M}\omega_{g_1}$, $R\omega_f$	$\log \bar{M}\omega_{g_1} = 16.052(R\omega_f)^2 - 5.323R\omega_f + 5.589$	0.004	(22)	10	[18]
Multivariate	$\bar{\lambda}$, M, R, C	$\log \bar{\lambda} = -0.6345\hat{F} + 7.399$ $\hat{F} = 3.391 \frac{M}{M_\odot} - 5.241 \frac{R}{10km} + 4.768 \frac{C}{0.2}$	0.023	(30)	13	…
	ω_f, M, $\tilde{\rho}$, C	$\omega_f = -0.00033\hat{F}^2 + 0.013\hat{F} - 0.023$ $\hat{F} = 2.980 \frac{M}{M_\odot} + 10.231 \frac{\tilde{\rho}}{0.04} - 8.398 \frac{C}{0.2}$	0.015	(34)	15	…
	ω_f, $\tilde{\rho}$, $C\tilde{\rho}$	$\omega_f = 0.0002\hat{F}^2 + 0.006\hat{F} + 0.003$ $\hat{F} = 6.911 \frac{\tilde{\rho}}{0.04} - 1.716 \frac{C\tilde{\rho}}{0.01}$	0.010	(36)	16	…

idea presented in [14], we used the principal components found through PCA to construct a new combined feature that we then related to a initially selected target feature. While this approach is not yet fully automated and requires manual considerations in some steps, our results show that this approach can yield highly accurate, multivariate universal relations. Our approach works particularly well when we try to find first-order corrections to previously known bivariate relations. For instance, we were able to construct an entirely novel universal relation that allows us to relate the f-mode frequency to the average density and compactness of the neutron stars, significantly improving the error of the relation compared to existing bivariate relations.

In Table VI we give an overview of all universal relations presented in this paper. For each relation, we indicate which features are connected through these relations, their form, and the average relative error achieved through our best fits. We also give references to all corresponding equations and figures in this paper. Finally, if a relation was already presented previously in a different work, we also give a reference to that work.

In a time where theoretical model data for various (astro-)physical objects becomes more widely available, finding useful data analysis tools for the specific use cases that we are interested in will be an important direction of work that will later enable more comprehensive data exploration. The methods discussed in this paper present a first step in this direction.

For future work, a straightforward extension is the application of the presented methods to even more and different neutron star data. While we have only considered nonrotating neutron stars in this paper, the presented methods should easily apply to other configurations including rotation or magnetic fields. Furthermore, gaining deeper understanding on why and under which constraints the PCA approach will work well can allow us to, in the future, reduce the amount of manual intervention that is still required right now.

APPENDIX A: COMBINED FEATURES FROM MULTIVARIATE CORRELATION ANALYSIS

We here show in Fig. 17 a visual representation of correlating the combined features we obtained in Sec. VI A with the target feature $\bar{\lambda}$. From this figure we can clearly see the strong correlation of the combined feature obtained from both principal components 0 and 2 with $\bar{\lambda}$. While the purely visual inspection already points towards principal component 0 allowing for the smaller error, a precise analysis using the bivariate correlation method was ultimately necessary to decide which component induces the universal relation with the least error. However, a similar visual analysis can and should be performed to assist in any attempt to construct universal relations using multivariate data analysis.

APPENDIX B: THE SPECIAL CASE WITH STRONG COLLINEARITY

Unfortunately, the approach using multivariate statistical analysis we described in this work (cf. Sec. V) does not always produce conclusive results: in cases where there exist strong correlations between features, the conditions we formulated in Sec. V B will not necessarily or sufficiently lead to the construction of universal relations.

For instance, let us consider the case where we want to predict the compactness C given the features $\bar{M}\omega$ and η. The principal component analysis leads to the loadings

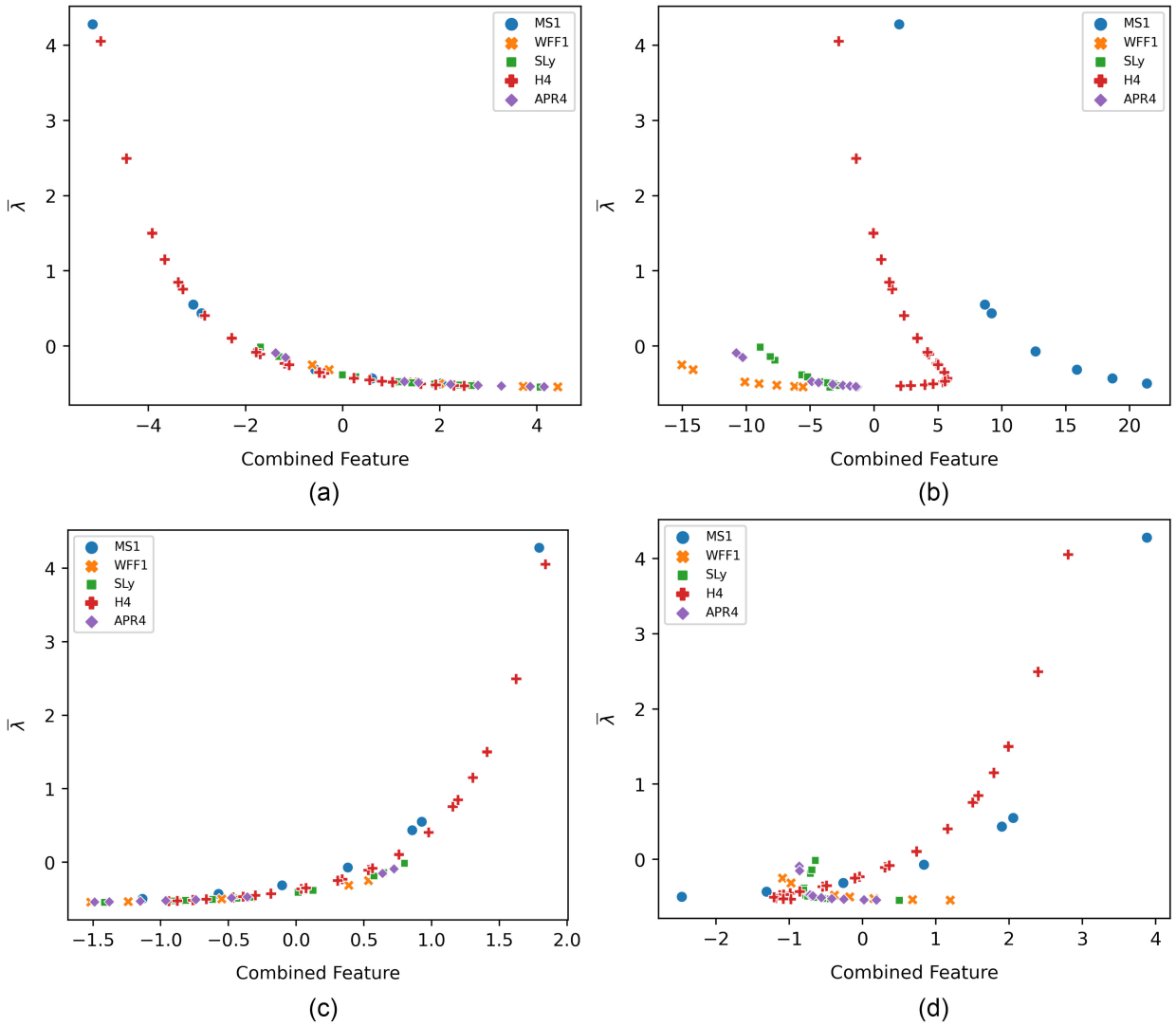

FIG. 17. The combined feature derived from each principal component solved for the target feature λ after performing PCA on the feature set $\mathbf{F} = \{M, R, C, \lambda\}$. The corresponding loadings are given in Table IV. (a) Principal component 0. (b) Principal component 1. (c) Principal component 2. (d) Principal component 3.

given in Table VII, and the associated combined features shown in Fig. 18. As we can see, each corresponding combined feature is strongly correlated to C, however inspection of the loading does not necessarily yield any specific principal component for which C has a significantly larger contribution. As such, not finding principal component with a proportionally large loading for our target feature does necessarily imply that no potential universal relation exists.

TABLE VII. Loadings of features in each principal component shown in Fig. 18.

Component	η	$\bar{M}\omega$	C
0	−0.577	0.578	−0.577
1	−0.687	−0.038	0.725
2	−0.441	0.815	−0.375

APPENDIX C: COUNTEREXAMPLE FOR MULTIVARIATE CORRELATION ANALYSIS

We now attempt to construct a universal relation for the unnormalized tidal deformablity λ, using the features M, ρ_c, and $\bar{I}$. We again apply the principal component analysis on all four features. The resulting principal components are shown in Fig. 19. The loadings of each feature corresponding to each principle component are given in Table VIII.

As we can clearly see here, none of the combined features derived from the principal components are well correlated with λ. This is also reflected in the loadings:

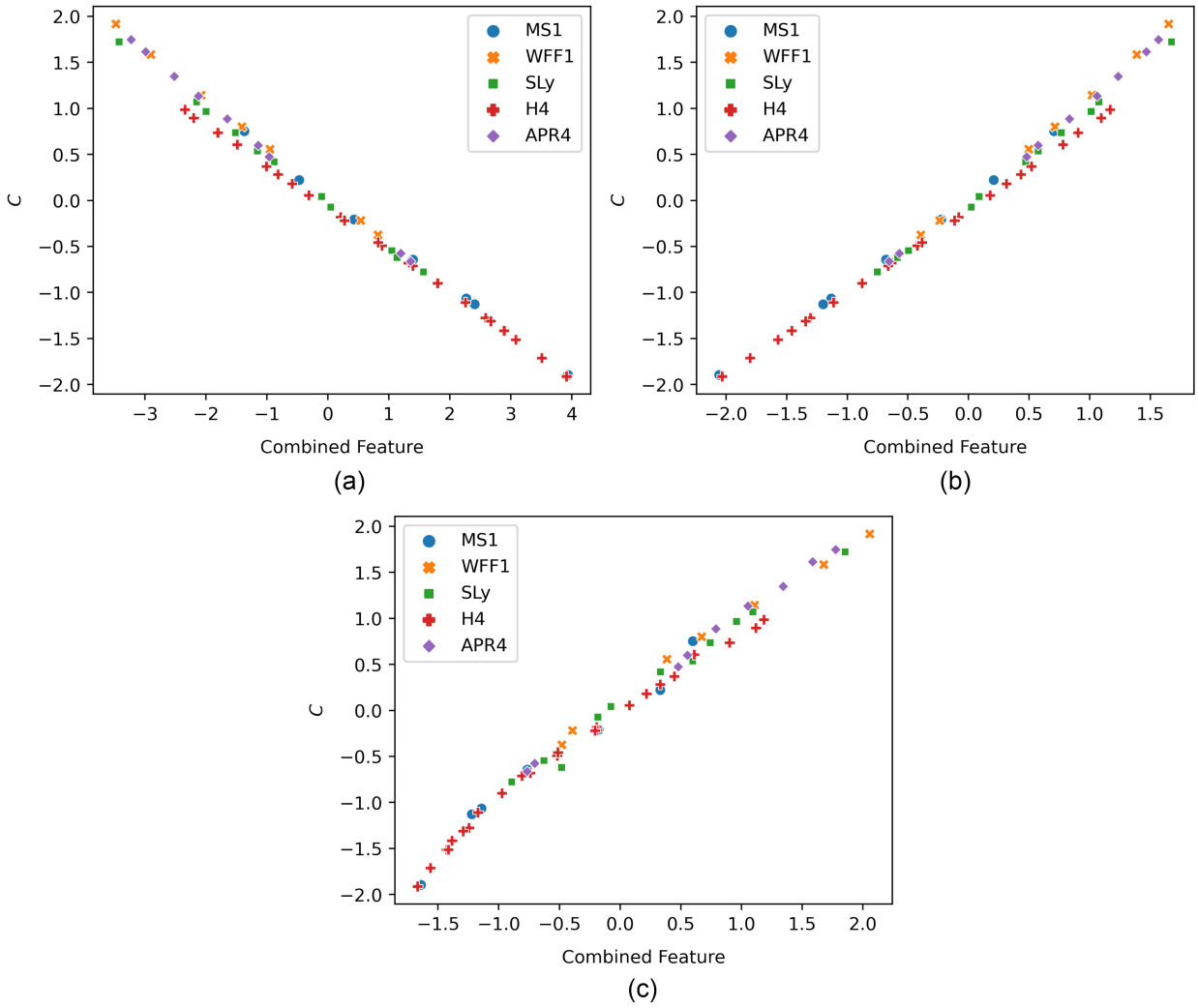

FIG. 18. The combined feature derived from each principal component solved for the target feature C after performing PCA on the feature set $\mathbf{F} = \{\bar{M}\omega_f, \eta, C\}$. The corresponding loadings are given in Table VII. (a) Principal component 0. (b) Principal component 1. (c) Principal component 2.

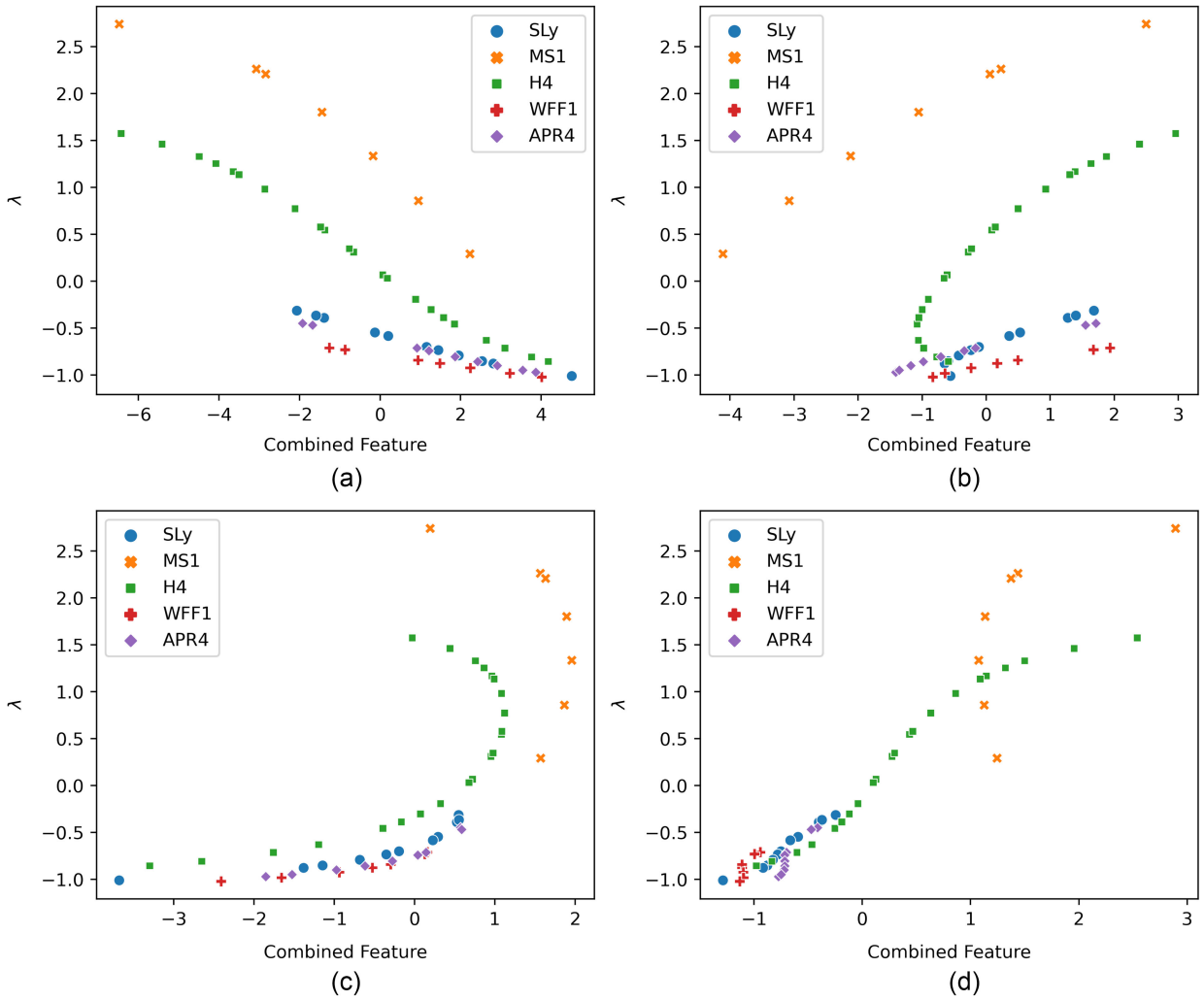

FIG. 19. The combined feature derived from each principal component solved for the target feature λ after performing PCA on the feature set $\mathbf{F} = \{M, \rho_c, \bar{I}, \lambda\}$. The corresponding loadings are given in Table VIII. (a) Principal component 0. (b) Principal component 1. (c) Principal component 2. (d) Principal component 3.

there is no principal component for which the feature λ shows a significantly higher contribution than the other features.

However, through bivariate analysis, we were previously able to find the well-known I-Love [8] relation between the normalized tidal deformability $\bar{\lambda}$ and $\bar{I}$ (cf. Fig. 2). This shows that typically employed normalizations can therefore also not necessarily be overcome by simply employing the PCA approach.

TABLE VIII. Loadings of features in each principal component shown in Fig. 19.

Component	ρ_c	M	$\bar{I}$	λ
0	−0.512	−0.448	0.539	0.497
1	−0.299	0.778	−0.132	0.537
2	0.790	0.090	0.441	0.417
3	−0.158	0.432	0.705	−0.540

[1] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **119**, 161101 (2017).
[2] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Astrophys. J. Lett. **892**, L3 (2020).
[3] N. Andersson and K. D. Kokkotas, Phys. Rev. Lett. **77**, 4134 (1996).
[4] N. Andersson and K. D. Kokkotas, Mon. Not. R. Astron. Soc. **299**, 1059 (1998).
[5] L. K. Tsui and P. T. Leung, Phys. Rev. Lett. **95**, 151101 (2005).
[6] H. Lau, P. Leung, and L. Lin, Astrophys. J. **714**, 1234 (2010).
[7] T. Chan, Y.-H. Sham, P. Leung, and L.-M. Lin, Phys. Rev. D **90**, 124023 (2014).
[8] K. Yagi and N. Yunes, Science **341**, 365 (2013).
[9] S. Bernuzzi, T. Dietrich, and A. Nagar, Phys. Rev. Lett. **115**, 091101 (2015).
[10] L. Rezzolla and K. Takami, Phys. Rev. D **93**, 124051 (2016).
[11] K. Kiuchi, K. Kawaguchi, K. Kyutoku, Y. Sekiguchi, and M. Shibata, Phys. Rev. D **101**, 084006 (2020).
[12] P. Manoharan, C. J. Krüger, and K. D. Kokkotas, Phys. Rev. D **104**, 023005 (2021).
[13] C. J. Krüger, K. D. Kokkotas, P. Manoharan, and S. H. Völkel, Front. Astron. Space Sci. **8**, 166 (2021).
[14] J. Soldateschi, N. Bucciantini, and L. Del Zanna, Astron. Astrophys. **654**, A162 (2021).
[15] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, Ann. Stat. **35**, 2769 (2007).
[16] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (The University of Illinois Press, Urbana, IL, 1949).
[17] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, Science **334**, 1518 (2011).
[18] H.-J. Kuan, A. G. Suvorov, and K. D. Kokkotas, Mon. Not. R. Astron. Soc. **506**, 2985 (2021).
[19] C. J. Krüger and K. D. Kokkotas, Phys. Rev. Lett. **125**, 111106 (2020).
[20] D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman, and C. Furlanello, Bioinformatics **29**, 407 (2012).
[21] M. Clark, A comparison of correlation measures, https://m-clark.github.io/docs/CorrelationComparison.pdf (2013).
[22] N. Jiang and K. Yagi, Phys. Rev. D **101**, 124006 (2020).
[23] L. K. Tsui and P. T. Leung, Mon. Not. R. Astron. Soc. **357**, 1029 (2005).
[24] C. Chirenti, G. H. de Souza, and W. Kastaun, Phys. Rev. D **91**, 044034 (2015).
[25] O. Benhar, V. Ferrari, and L. Gualtieri, Phys. Rev. D **70**, 124015 (2004).
[26] T. Barnett and R. Preisendorfer, Mon. Weather Rev. **115**, 1825 (1987).