Author for correspondence:
Nick S. Jones
e-mail: nick.jones@imperial.ac.uk

Highly comparative time-series analysis: the empirical structure of time series and their methods

Ben D. Fulcher[1], Max A. Little[1] and Nick S. Jones[1,2]

[1]Department of Physics, University of Oxford, Oxford, UK
[2]Department of Mathematics, Imperial College London, London, UK

The process of collecting and organizing sets of observations represents a common theme throughout the history of science. However, despite the ubiquity of scientists measuring, recording and analysing the dynamics of different processes, an extensive organization of scientific time-series data and analysis methods has never been performed. Addressing this, annotated collections of over 35 000 real-world and model-generated time series, and over 9000 time-series analysis algorithms are analysed in this work. We introduce reduced representations of both time series, in terms of their properties measured by diverse scientific methods, and of time-series analysis methods, in terms of their behaviour on empirical time series, and use them to organize these interdisciplinary resources. This new approach to comparing across diverse scientific data and methods allows us to organize time-series datasets automatically according to their properties, retrieve alternatives to particular analysis methods developed in other scientific disciplines and automate the selection of useful methods for time-series classification and regression tasks. The broad scientific utility of these tools is demonstrated on datasets of electroencephalograms, self-affine time series, heartbeat intervals, speech signals and others, in each case contributing novel analysis techniques to the existing literature. Highly comparative techniques that compare across an interdisciplinary literature can thus be used to guide more focused research in time-series analysis for applications across the scientific disciplines.

1. Introduction

Time series, measurements of a quantity taken over time, are fundamental data objects studied across the scientific disciplines, including measurements of stock prices in finance, ion fluxes in astrophysics, atmospheric air temperatures in meteorology and human heartbeats in medicine. In order to understand the structure in these signals and the mechanisms that underlie them, scientists have developed a large variety of techniques: methods based on fluctuation analysis are frequently used in physics [1], generalized autoregressive conditional heteroskedasticity (GARCH) models are common in economics [2] and entropy measures such as sample entropy are popular in medical time-series analysis [3], for example. But are these methods that have been developed in different disciplinary contexts summarizing time series in unique and useful ways, or is there something to be learned by synthesizing and comparing them? In this paper, we address this question by assembling extensive annotated libraries of both time-series data and methods for time-series analysis. We use each to organize the other: methods are characterized by their behaviour across a wide variety of different time series, and time series are characterized by the outputs of diverse scientific methods applied to them. The structure that emerges from these collections of data and methods provides a new, unified platform for understanding the interdisciplinary time-series analysis literature.

Although a strong theoretical foundation underlies many time-series analysis methods, their great number and interdisciplinary diversity makes it very

difficult to determine how methods developed in different disciplines relate to one another, and for scientists to select appropriate methods for their finite, noisy data. It is similarly difficult to determine how time series studied in one scientific discipline compare to those studied in other disciplines or to the dynamics defined by theoretical time-series models. The automated, data-driven techniques introduced in this work address these practical difficulties by allowing diverse methods and data from across the sciences to be compared in a unified representation. The high-throughput analysis of the DNA microarray in biology is now a standard complement to traditional, more focused research efforts; we envisage the comparative techniques introduced in this work providing a similarly guiding role for scientific time-series analysis. Previous comparisons of time-series analysis methods have been performed only in specific disciplinary contexts and on a small scale, and attempts to organize large time-series datasets have typically involved time series of a fixed length and measured from a single system [4,5]. This work, therefore, represents the first scientific comparison of its kind and is unprecedented in both its scale and interdisciplinary breadth.

The paper is structured as follows. In §2, we describe our extensive, annotated libraries of scientific time-series data and time-series analysis methods, and introduce our computational framework. Then, in §3, we show how these diverse collections of both methods (§3.1) and data (§3.2) can be organized meaningfully: by representing methods using their behaviour on the data, and representing time series by their measured properties. Using this new ability to structure diverse collections of time series and their methods, we introduce a range of useful techniques, including the ability to connect specific pieces of time-series data to similar real-world and model-generated time series, and to link specific time-series analysis methods to a range of alternatives from across the literature. The highly comparative analysis techniques are demonstrated using scientific case studies in §4. The diverse range of scientific methods in our library is used to contribute new insights into existing time-series analysis problems: uncovering meaningful structure in time-series datasets, and selecting relevant methods for general time-series classification and regression tasks automatically. We contrast our approach, which compares across a wide, interdisciplinary time-series analysis literature, to that of conventional studies that focus on small sets of manually selected techniques with minimal comparison to alternatives. Our main conclusions are summarized in §5.

2. Framework

We assembled annotated libraries of: (i) 38 190 univariate time series measured from diverse real-world systems and generated from a variety of synthetic model systems and (ii) 9613 time-series analysis algorithms developed in a range of scientific literatures as well as new analysis methods developed by us in this work. Over 20 000 real-world time series were acquired, primarily from publicly available databases, which have been selected to encompass a representative sample of the types of signals measured in scientific disciplines: from meteorology (e.g. temperature, air pressure, rainfall and river flow), medicine (e.g. heartbeat intervals, electrocardiograms and electroencephalograms (EEGs)), audio (e.g. human speech, music, sound effects and animal sounds), astrophysics (e.g. solar radio flux and interplanetary magnetic field), finance (e.g. exchange rates and stock prices) and others. In addition, we generated over 10 000 synthetic time series, including outputs from a range of nonlinear and chaotic maps, various dynamical systems/flows, correlated noise models and other stochastic processes. Although it is clearly unfeasible to produce exhaustive libraries of all types of scientific time-series data, we have attempted to be as comprehensive and even-handed as possible. However, emphasis was inevitably given to time series in publicly available repositories and outputs from time-series models that were relatively simple to implement. A list of time series used in this work, including descriptions and references, is in the electronic supplementary material, Time Series List.

Methods for time-series analysis take on a variety of forms, from simple summary statistics to statistical model fits. We implemented each such method as an algorithm: an *operation* that summarizes an input time series with a single real number. Our library of over 9000 such operations quantifies a wide range of time-series properties, including basic statistics of the distribution (e.g. location, spread, Gaussianity and outlier properties), linear correlations (e.g. autocorrelations and features of the power spectrum), stationarity (e.g. StatAv, sliding window measures and unit root tests, prediction errors), information theoretic and entropy measures (e.g. auto-mutual information, approximate entropy and Lempel–Ziv complexity), methods from the physical nonlinear time-series analysis literature (e.g. correlation dimension, Lyapunov exponent estimates and surrogate data analysis), linear and nonlinear model fits (e.g. goodness of fit and parameter values from autoregressive moving average, GARCH, Gaussian process and state space models) and others (e.g. wavelet methods, properties of networks derived from time series, etc.). A large component of this work involved implementing existing time-series analysis methods (including publicly available packages and toolboxes) in the form of operations. In some cases, it was necessary to formulate new types of operations that appropriately summarized the outputs of existing methods, and in other cases we developed new, qualitatively different types of operations that are introduced for the first time in this work (see the electronic supplementary material, §S1.1.2). The collection is inevitably incomplete, and indeed this operational framework is more suited to those methods that can be automated than more subtle types of analysis that require delicate hands-on experimentation by a time-series analysis expert. However, in the process of collecting and implementing these methods, we made a concerted effort to incorporate and appropriately automate as many distinct types of scientific time-series analysis methods as possible. The result is sufficiently comprehensive to achieve the range of successful results reported in this work. Note that although we list over 9000 operations, this number includes cases for which a single method is repeated for multiple parameter values (e.g. calculating the autocorrelation function at 40 different time lags constitutes 40 different operations despite simply varying a single parameter of a single method); the number of conceptually distinct methods is significantly lower (one estimate arrives at approx. 1000 unique operations, cf. electronic supplementary material, §S1.1.2). A full list of operations developed for this work, including references and descriptions, is in the electronic supplementary material, Operation List.
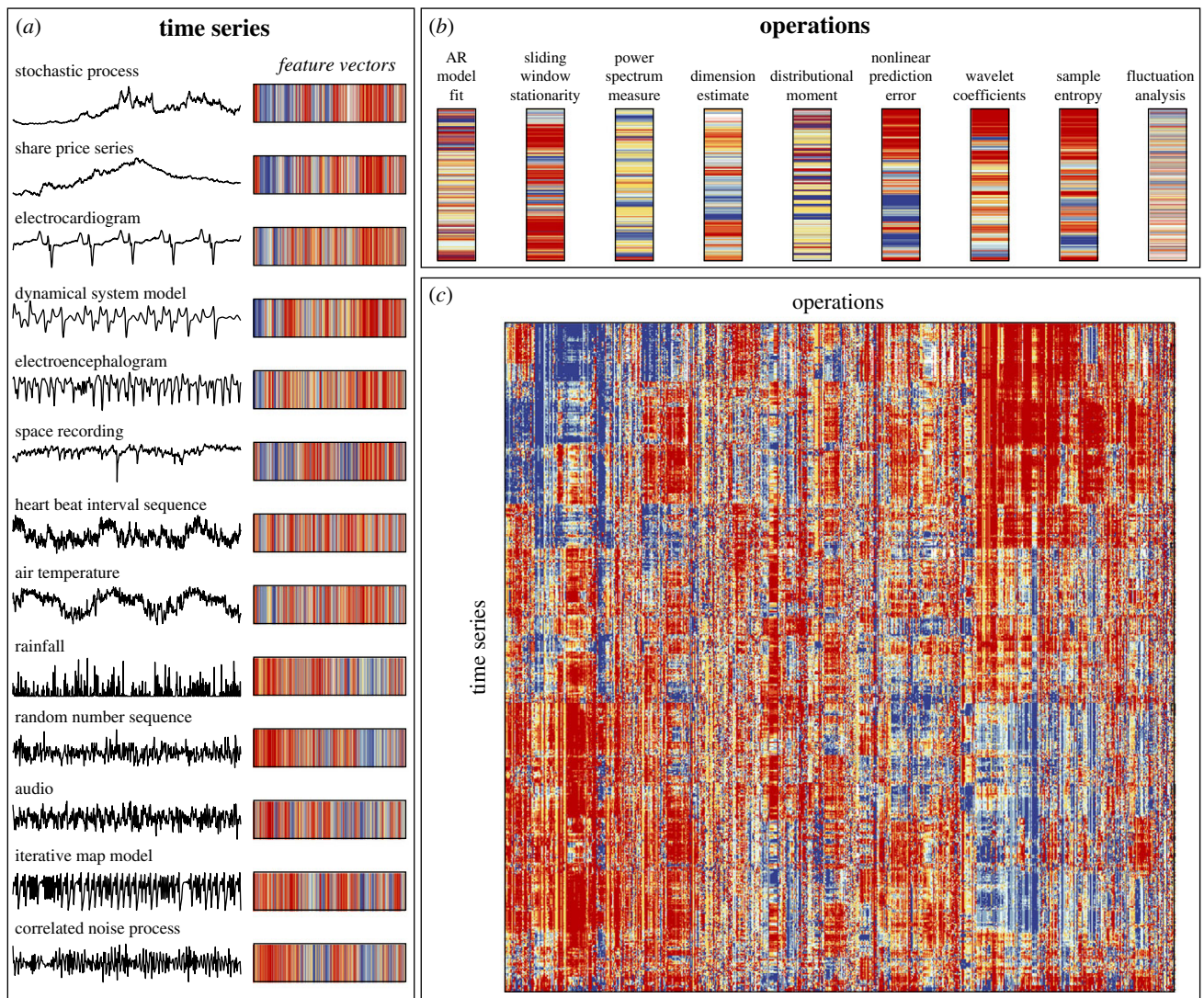
Figure 1. A unified representation of time series and their analysis methods. A fundamental component of this work, which involves applying large numbers of time-series analysis methods, or operations, to large time-series datasets, can be visualized as a data matrix, shown in (c), where each column represents an operation and each row represents a time series. Elements of the matrix contain the (normalized) results of applying each operation to each time series, and are visualized using colour: from blue (low operation outputs) to red (high operation outputs). We note the similarity of this time-series data matrix to the DNA microarray used in biology, that simultaneously measures the expression levels under multiple conditions (our operations) of large numbers of genes (our time series). Rows and columns have been reordered using linkage clustering to place similar time series and similar operations close to one another (cf. electronic supplementary material, §S1.3.2). Thus, time series (rows of the data matrix) are represented as feature vectors containing a large number of informative properties, illustrated in (a), and operations (columns of the data matrix) are represented as feature vectors containing their outputs across a set of time series, illustrated in (b). In this figure, we use an interdisciplinary set of 875 time series and 8651 well-behaved operations (cf. electronic supplementary material, §1.1.5); the matrix in (c) has been resized to approximately square proportions for the purpose of visualization. The rich structure in this data matrix encodes relationships between different scientific time series and diverse types of analysis methods, relationships that are investigated in detail in this work. (Online version in colour.)

A fundamental component of this work involves analysing the result of applying a large set of operations to a large set of time series. This computation can be visualized as a data matrix with time series as rows and operations as columns, as shown in figure 1c. Each element of the matrix, D_{ij}, contains the output of an operation, F_j, applied to a time series, $\mathbf{x}_i$, so that $D_{ij} = F_j(\mathbf{x}_i)$. Correspondingly, time series are represented as feature vectors containing measurements of an extensive range of their properties (figure 1a), and operations are represented as feature vectors containing their outputs across a time-series dataset (figure 1b). In order to allow operations with different ranges and distributions of outputs to be compared meaningfully, we applied an outlier-robust sigmoidal normalizing transformation to the outputs of each operation, as described in the electronic supplementary material, §S1.2.2. Time series are compared using Euclidean distances calculated

between their feature vectors, and operations are compared using correlation-based distances measured between their outputs (either linear correlation-based distances to capture linear relationships or normalized mutual information-based distances to capture potentially nonlinear relationships, cf. electronic supplementary material, §S1.3.1). Thus, time series are judged as similar that have many similar properties, and operations are judged as similar that have highly correlated outputs across a time-series dataset. Note that when operations do not output a real number or are inappropriate (e.g. it is not appropriate to fit a positive-only distribution to non-positive data), these outputs are referred to as 'special values' and are treated as missing elements of the data matrix that can be filtered out (cf. electronic supplementary material, §S1.1.5).

The rich structure in the data matrix shown in figure 1c, which combines the results of applying a wide range of
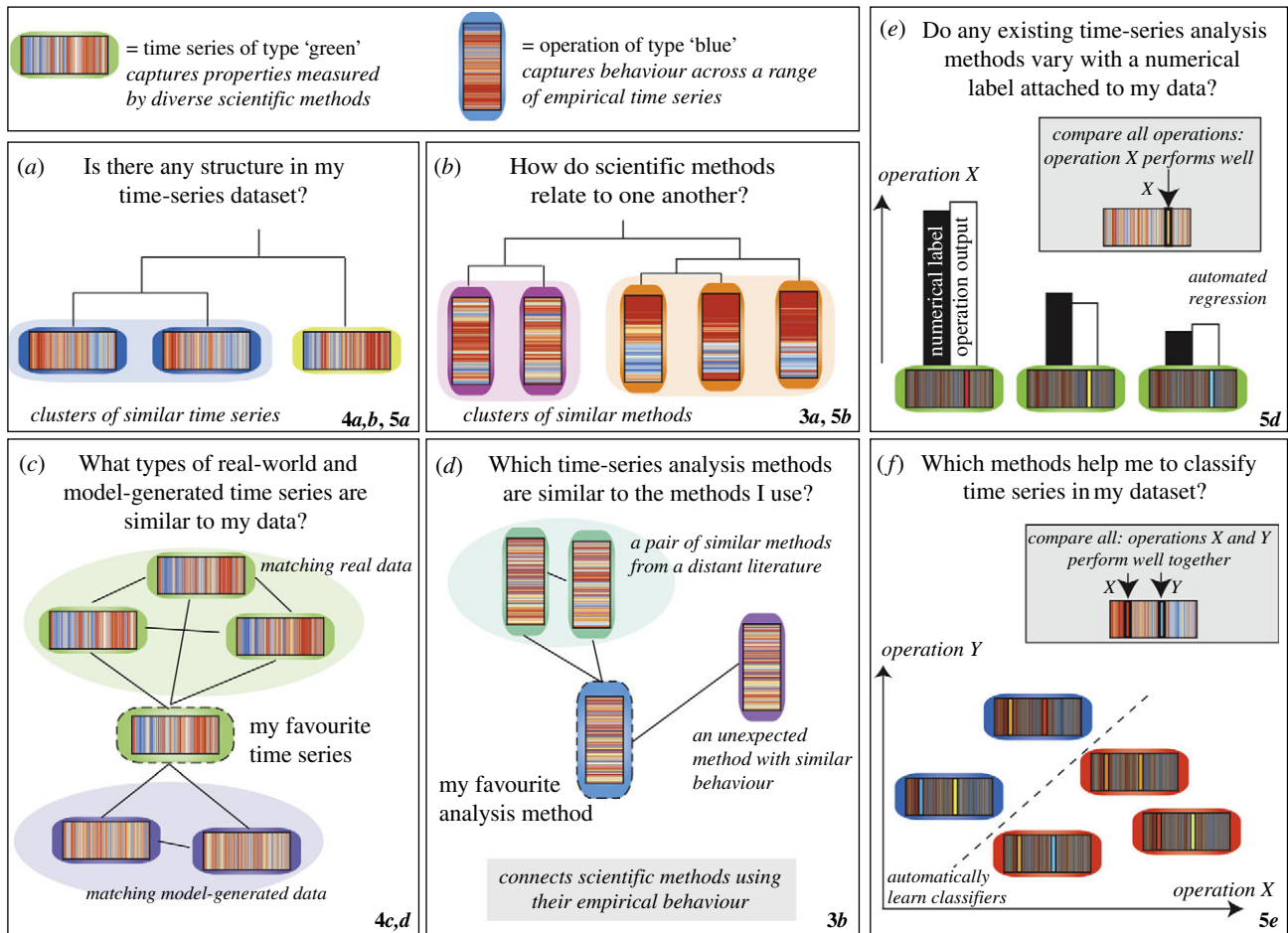
Figure 2. Key scientific questions that can be addressed by representing time series by their properties (measured by many types of analysis methods) and operations by their behaviour (across many types of time-series data). We show that this representation facilitates a range of versatile techniques for addressing scientific time-series analysis problems, which are illustrated schematically in this figure. The representations of time series (rows of the data matrix, figure 1a) and operations (columns of the data matrix, figure 1b) serve as empirical fingerprints, and are shown in the top panel. Coloured borders are used to label different classes of time series and operations, and other figures in this paper that explicitly demonstrate each technique are given in the bottom right-hand corner of each panel. (a) Time-series datasets can be organized automatically, revealing the structure in a given dataset (cf. figures 4a,b and 5a). (b) Collections of scientific methods can be organized automatically, highlighting relationships between methods developed in different fields (cf. figures 3a and 5b). (c) Real-world and model-generated data with similar properties to a specific time-series target can be identified (cf. figure 4c,d). (d) Given a specific operation, alternatives from across science can be retrieved (cf. figure 3b). (e) Regression: the behaviour of operations in our library can be compared to find operations that vary with a target characteristic assigned to time series in a dataset (cf. figure 5d). (f) Classification: operations can be selected based on their classification performance to build useful classifiers and gain insights into the differences between classes of labelled time-series datasets (cf. figure 5e).

scientific time-series analysis methods to a diverse set of time series, thus encapsulates interesting relationships between different ways of measuring structure in time series (columns), and relationships between data generated by different types of systems (rows). For example, redundancy across analysis methods is indicated by a set of adjacent columns in the data matrix that display similar patterns (these operations exhibit similar behaviour across a great variety of different time series). In this work, we introduce a range of simple techniques for extracting this kind of interesting and scientifically meaningful structure from data matrices, as illustrated schematically in figure 2. In particular, we show that representing time series in terms of their measured properties, and analysis methods in terms of their behaviour on empirical data, can form a useful basis for answering the types of questions depicted in figure 2, including the ability to structure collections of data (figure 2a) and methods (figure 2b), find matches to particular time series (figure 2c) or methods (figure 2d), and perform time-series regression (figure 2e) or classification (figure 2f) automatically. Our approach is

unusual in that it is completely automated and uses no domain knowledge about the time series or any information about the theoretical assumptions underlying the analysis methods: we simply use the empirical behaviour of methods and time series as a platform for comparison. Using numerous examples from across science, we demonstrate that our libraries of time-series data and their analysis methods are sufficiently comprehensive for this approach to yield novel and scientifically meaningful results for a range of applications.

3. Empirical structure

In this section, we investigate the empirical structure of our annotated libraries of time series and their methods using the types of analysis depicted schematically in figure 2a−d. First, in §3.1, we show that the operations (columns of the data matrix in figure 1c) can be organized in a meaningful way using their behaviour on a set of 875 different time series. Then, in §3.2, we perform a similar treatment for
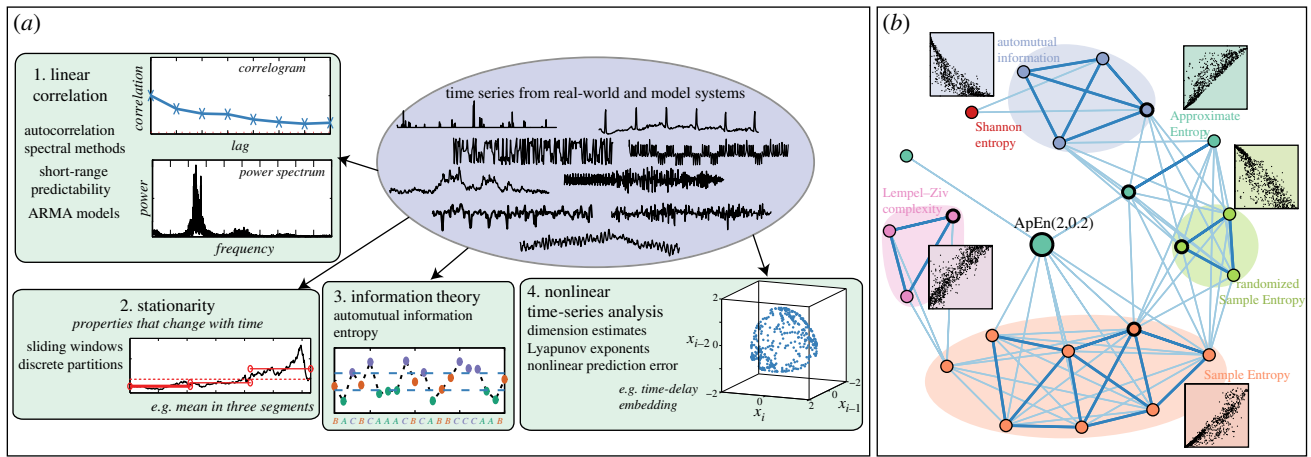
Figure 3. Structure in a library of 8651 time-series analysis operations. (*a*) A summary of the four main classes of operations in our library, as determined by a *k*-medoids clustering, reflects a crude but intuitive overview of the time-series analysis literature. (*b*) A network representation of the operations in our library that are most similar to the approximate entropy algorithm, ApEn(2,0.2) [7], which were retrieved from our library automatically. Each node in the network represents an operation and links encode distances between them (computed using a normalized mutual information-based distance metric, cf. electronic supplementary material, §S1.3.1). Annotated scatter plots show the outputs of ApEn(2,0.2) (horizontal axis) against a representative member of each shaded community (indicated by a heavily outlined node, vertical axis). Similar pictures can be produced by targeting any given operation in our library, thereby connecting different time-series analysis methods that nevertheless display similar behaviour across empirical time series. (Online version in colour.)

time series (rows of the data matrix in figure 1*c*). We find that by judging operations and time series using a form of empirical fingerprint of their behaviour facilitates a useful means of comparing them.

3.1. Empirical structure of time-series analysis methods

First, we analyse the structure in our library of time-series analysis operations when applied to a representative inter-disciplinary set of 875 real-world and model-generated time series (a selection of time series in this set is illustrated in figure 1*a*). This carefully controlled set of 875 time series was selected to encompass the different types of scientific time series as evenhandedly as possible (for example, the full time-series library contains a large number of ECGs and rainfall time series; by using a controlled set, we avoid emphasizing the behaviour of operations on these particular classes of time series that happen to be more numerous in our library, cf. electronic supplementary material, §S2.1). Operations that returned less than 20 per cent special-valued outputs on this set of time series were analysed: a collection of 8651 operations. In this section, we show that the behaviour of operations on these 875 diverse scientific time series becomes a useful form of empirical fingerprint for them that groups similar types of operations and distinguishes very different types of operations.

We used clustering [6] to uncover structure in these 8651 operations automatically (as in figure 2*b*). Clustering can be performed at different resolutions to produce different overviews of the time-series analysis literature. For example, clustering the operations into four groups using *k*-medoids clustering [6] is depicted in figure 3*a*, and reflects a crude but intuitive summary of the types of methods that scientists have developed to study time series. Although highly simplified, the result shows that we can begin to organize an interdisciplinary methodological literature in an automatic, data-driven way.

Next, we ask a more subtle question: how many different types of operations are required to provide a good summary of the rich diversity of behaviours exhibited by the full set of 8651 scientific time-series analysis operations? Clustering at finer levels allows us to probe this, and to construct reduced sets of operations that efficiently approximate the behaviours in the full library by eliminating methodological redundancy. The quality of such reduced sets of operations is quantified using the *residual variance* measure, $1 - R^2$, where R is the linear correlation coefficient measured between the set of distances between time series in a reduced space and those in the full space [8]. Using *k*-medoids clustering for a range of cluster numbers, *k*, we found that a reduced set of 200 operations provides a very good approximation to the full set of 8651 operations, with a residual variance of just 0.05 (cf. electronic supplementary material, §S2.2). These 200 operations thus form a concise summary of the different behaviours of time-series analysis methods applied to scientific time series, and draw on techniques developed in a range of disciplines, including autocorrelation, auto-mutual information, stationarity, entropy, long-range scaling, correlation dimension, wavelet transforms, linear and nonlinear model fits, and measures from the power spectrum (cf. electronic supplementary material, 200 Operations). Constraints on the dynamics observed in empirical time series thus allows us to exploit redundancy in scientific methods acting on them. In §3.2, this reduced set of operations is used to organize our time-series library.

As well as analysing the overall structure in our library of operations, it is also informative to explore the local structure surrounding a given target operation (as depicted in figure 2*d*). In this way, the behaviour of any given operation can be contextualized with respect to alternatives from across the time-series analysis literature, including methods developed in unfamiliar fields or in the distant past. An example is shown in figure 3*b* for the Approximate Entropy algorithm, ApEn(2,0.2), a 'regularity' measure that has been applied widely [7]. In figure 3*b*, nodes in the network are the most similar operations to ApEn(2,0.2) in our library, and links indicate their similarities (using a mutual information-based distance metric to capture correlations across the set of 875 time series, cf. electronic supplementary material, §S1.3.1). The network contains different communities of similar operations, including methods based on Sample Entropy,

Lempel–Ziv complexity, auto-mutual information, Shannon entropy and other approximate entropies. Inset scatter plots show the relationships between ApEn(2,0.2) and the other operations across different types of scientific time series. We thus discover that, even when mixed with 8650 different operations, related families of entropy measures are retrieved automatically and organized meaningfully by comparing their behaviour across a controlled set of 875 time series. Similar pictures can be produced straightforwardly for any operation in our library, as is done for fluctuation analysis and singular spectrum analysis, for example, in the electronic supplementary material, §S2.4.

We thus find that representing time-series analysis methods using an empirical fingerprint of their behaviour across 875 different time series provides a useful means of comparing and structuring a methodological literature. Across the scientific disciplines, there exists a vast number of time-series analysis methods, but no framework with which to judge whether progress is really being made through the continual development of new types of methods. By comparing their empirical behaviour, the techniques demonstrated above can be used to connect new methods to alternatives developed in other fields in a way that encourages interdisciplinary collaboration on the development of novel methods for time-series analysis that do not simply reproduce the behaviour of existing methods.

3.2. Empirical structure of time series

Above, we studied the structure in our library of scientific operations using their behaviour on empirical time series; in this section, we analyse a collection of 24 577 time series in a similar way. Time series are compared using the set of 200 representative operations formed above to measure their properties, providing an empirical means of comparing them and addressing useful scientific questions (as depicted in figure 2a,c). We restricted our clustering analysis to this set of 24 577 time series, which is an appropriate subset of the full library of 38 190 time series that removes very short time series (of less than 1000 samples) and filters over-represented classes of time series (as explained in the electronic supplementary material, §S3.2). Finding structure in a set of time series of different lengths and measured in different ways from different systems is a major challenge [5]; we show that this empirical fingerprint of 200 diverse time-series analysis operations facilitates a meaningful comparison of scientific time series. First, we formed 2000 clusters of time series from our library (using complete linkage clustering, cf. electronic supplementary material, §S3.2). Despite the size of our library and the diversity of time series contained in it, most clusters grouped time series measured from the same system; some examples are shown in figure 4a (more examples are in the electronic supplementary material, §S3.2.1). As the clustering is done by comparing a wide range of time-series properties, clusters appear to group time series according to their dynamics, even when they have different lengths. Some clusters contained time series generated by different systems, such as the cluster illustrated in figure 4b that contains time series generated from three different iterative maps: the *Cubic Map*, the *Sine Map* and the *Asymmetric Logistic Map* [9]. Although this cluster contains time series of different lengths and generated from different maps, it groups outputs from each map with parameters that specify a similar recurrence relationship, as shown in the inset plot of figure 4b. This cluster therefore

distinguishes a distinct and meaningful class of dynamical behaviour from a large library of time series (as do many other clusters, cf. electronic supplementary material, §S3.2.2). Conventional measures of time-series similarity are often based on distances measured between the time-series values themselves and are hence restricted to sets of time series of a fixed length [5]. Representing time series by a diverse range of their properties is clearly a powerful alternative that captures important dynamical behaviour in general collections of time series.

Our reduced representation also allows us to retrieve a local neighbourhood of time series with similar properties to any given target time series (note that this search is done across the full library of 38 190 time series). This automated procedure can be used to relate real-world time series to similar, model-generated time series in a way that suggests suitable families of models for understanding real-world systems (as depicted in figure 2c). An example is shown in figure 4c, where real-world and model-generated matches to a series of opening share prices for Oxford Instruments are plotted as a network. Real-world matches are other opening share price series, and model-generated matches are outputs from stochastic differential equations (SDEs). Indeed, the form of SDE models suggested by this set of matches, e.g. $dX_t = \mu X_t dt + \sigma X_t dW_t$, is the same as geometric Brownian motion, which is used in financial modelling [10] (where X_t represents the time series, W_t denotes a Wiener process and other variables are parameters). Matches to SDE models with slightly different forms, e.g. $dX_t = a(b - X_t)dt + \sigma\sqrt{X_t}dW_t$, suggest that other models with particular parameter values can also reproduce many properties of the target share price series that are not unique to the geometric Brownian motion model. Although this is an extremely crude alternative to conventional time-series modelling, it nevertheless allows real-world time series to be linked to relevant model systems in a completely automated and data-driven way. Analogous insights were gained for rainfall patterns, astrophysical recordings, human speech recordings and others in the electronic supplementary material, §S3.3.

Applying the same method in reverse, we also targeted time series generated by models and retrieved real-world time series with similar dynamics. For example, in figure 4d we targeted a time series generated by a stochastic sine map model, which has a fixed probability of additive uniformly distributed noise at each time step that can switch the system between two stable limit cycles [11]. As expected, the closest model-generated matches were other stochastic sine map time series generated using the same parameters as the target (not shown here, cf. electronic supplementary material, §S3.3.2), whereas real-world matches were meteorological processes that exhibit the same qualitative 'noisy switching' dynamics (figure 4d). Thus, the dynamics specified by the stochastic sine map model were explicitly linked to that of relevant real-world meteorological processes automatically, suggesting that this type of stochastic switching mechanism may capture some of the properties of these meteorological systems. Using this simple, general method for connecting real-world and model dynamics, we obtained similar results using noise-corrupted sine waves and self-affine time series in the electronic supplementary material, §S3.3.2.

We have introduced large, annotated collections of time series and their methods from across the scientific disciplines
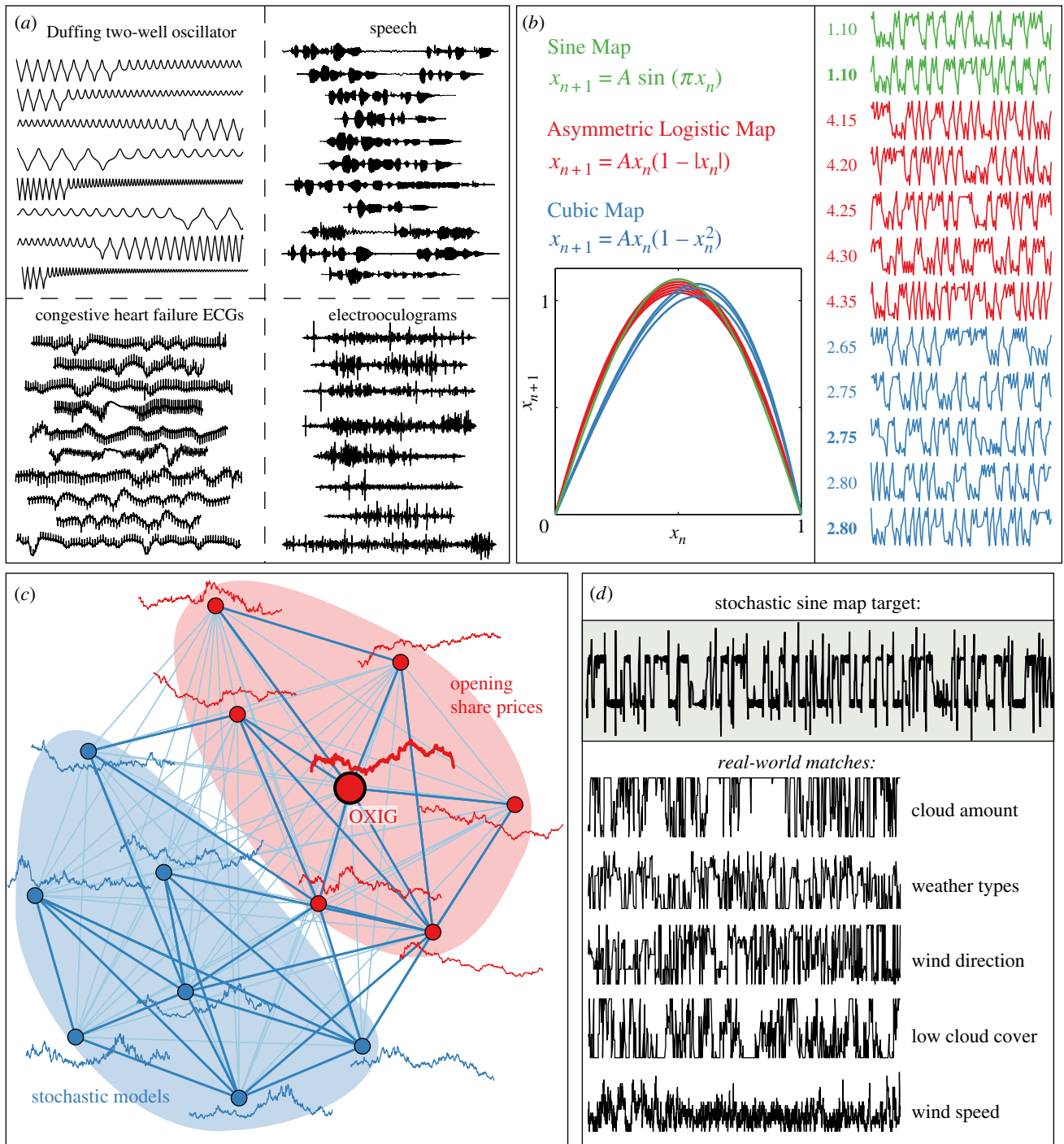
Figure 4. Structure in a large, diverse library of scientific time series. Representing time series using an empirical fingerprint containing 200 of their measured properties, a diverse collection of 24 577 time series was clustered into 2000 groups. (*a*) Most clusters formed in this way are homogeneous groups of time series of a given real-world or model system; four such examples are shown. (*b*) A time-series cluster is plotted that contains time series generated by three different iterative maps with parameters that specify a common recurrence relationship. Time-series segments of 150 samples are plotted and labelled with the parameter *A* of the map that generated them (bold labels indicate 5000-sample time series, the others are 1000 samples long). Recurrence relationships, $x_{n+1}(x_n)$, are plotted and are similar for all time series. (*c*) An opening share price series for Oxford Instruments (labelled OXIG) is targeted; the most similar real-world time series are opening share prices of other stocks (red nodes), and the most similar model-generated time series are from SDEs (blue nodes). Links in the network represent similarities between the time series judged using Euclidean distances, *d*, between their normalized feature vectors (darker, thicker links have $d < 3$, cf. electronic supplementary material, §S1.3.1). (*d*) The most similar real-world time series to a stochastic sine map target are meteorological processes that display qualitatively similar 'noisy switching' dynamics.

and used the behaviour of each applied to the other to analyse structure in them. Although representing time series and their methods in terms of their empirical behaviour might seem like an unusual idea that might not yield particularly meaningful results, we found that it indeed provides a powerful, data-driven means of investigating relationships between them. In particular, representing time series using

200 operations and representing operations using 875 time series appear to be sufficient to organize them meaningfully. As illustrated in figure 2*a*–*d*, the methods we introduce exploit this unified representation to connect individual pieces of time-series data to other time series with similar properties, and particular time-series analysis methods to alternatives with similar behaviour. The examples shown
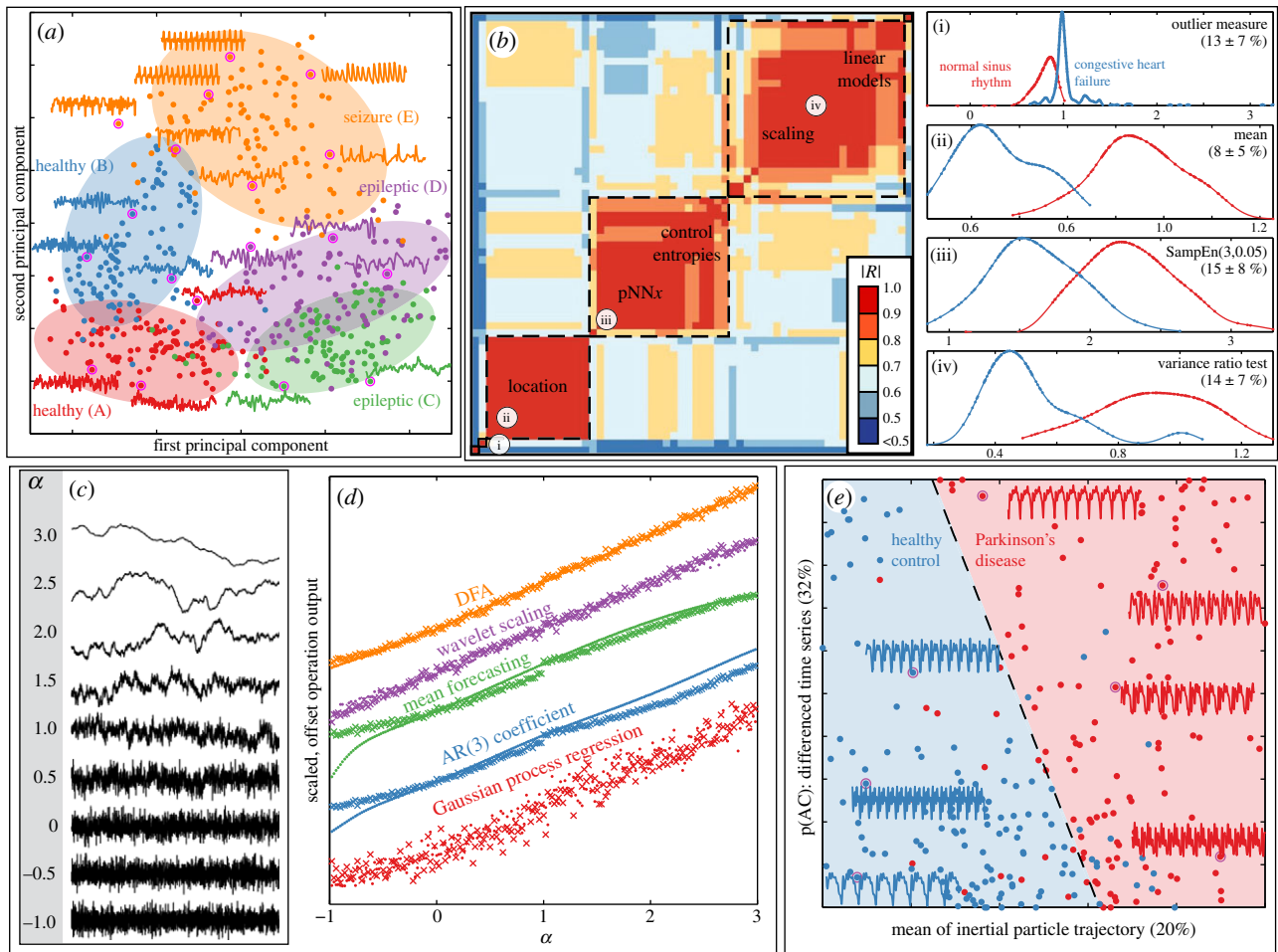
Figure 5. Highly comparative techniques for time-series analysis tasks. We draw on our full library of time-series analysis methods to: (a) structure datasets in meaningful ways, and retrieve and organize useful operations for (b,e) classification and (c,d) regression tasks. (a) Five classes of EEG signals are structured meaningfully in a two-dimensional principal components space of our library of operations. (b) Pairwise linear correlation coefficients measured between the 60 most successful operations for classifying congestive heart failure and normal sinus rhythm RR interval series. Clustering reveals that most operations are organized into one of three groups (indicated by dashed boxes). Distributions of selected operations, labelled (i), (ii), (iii) and (iv) in the correlation matrix, are plotted for the two classes, including the mean $\pm$ s.d. of their 10-fold cross-validation misclassification rate using a linear (threshold) classifier. (c) Segments of self-affine time series generated with scaling exponents in the range $-1 \leq \alpha \leq 3$. (d) Five selected operations with outputs that vary approximately linearly with α for self-affine time series generated by the Fourier filtering method (dots) and the random midpoint displacement method (crosses). (e) A linear classifier that combines the outputs of two different operations distinguishes Parkinsonian speech recordings (red) from those of healthy controls (blue) with a mean 10-fold cross-validation misclassification rate of 14%; the mean cross-validation misclassification rate of each individual operation is given in parentheses. Some sample segments of time traces are annotated to the plot.

here and in the comprehensive electronic supplementary material demonstrate the general applicability of these techniques to a broad range of time-series analysis problems.

4. Applications

In this section, we show how our library of operations can be used to provide new insights into the analysis of specific time-series datasets. We use the types of analysis demonstrated above, including organizing sets of time series (depicted in figure 2a) and methods (figure 2b), and also introduce new techniques for selecting useful operations for classification and regression tasks automatically (figure 2e,f). The broad utility of our highly comparative approach for specific scientific applications is shown using a wide range of case studies. Despite comparing across thousands of operations, we find that the selection of useful operations can be done in a statistically controlled manner using a multiple hypothesis testing framework described in the electronic

supplementary material, §S1.3.6. A detailed analysis of each case study, including information about the dataset, statistical tests, interpretations of selected operations and further comparisons to the existing literature, can be found in the electronic supplementary material, §S4 (time-series regression) and §S5 (time-series classification).

4.1. Electroencephalogram recordings

In the first case study, we show how the collective behaviour of our library of operations can be used to organize different types of EEG time series (cf. figure 2a), and we assess whether progress is being made in a literature concerned with the classification of epileptic seizure EEGs (cf. figure 2f). The dataset contains 100 EEG signals from each of five classes: set A (healthy, eyes open), set B (healthy, eyes closed), set C (epileptic, not seizure, recorded from opposite hemisphere of the brain to epileptogenic zone), set D (epileptic, not seizure, recorded within the epileptogenic zone) and set E (seizure) [12]. The two-dimensional principal components

projection of the dataset in the space of well-behaved operations is shown in figure 5a. This representation uses no knowledge of the class labels attached to the data, but simply organizes the time series according to their measured properties using our large and diverse library of operations. The dataset is structured in a way that is consistent with its known class structure: seizure time series (set E) are particularly distinguished, and the two healthy (sets A and B) and the two epileptic (sets C and D) sets are close in this space. Drawing on a large literature of time-series analysis operations using lower-dimensional representations can evidently uncover useful structure in time-series datasets, and indeed the same approach was found to be informative for a range of other systems (including periodic and noisy signals in the electronic supplementary material, §S5.1.2; seismic signals in §S5.2.3; emotional speech recordings in §S5.3.3; and RR intervals in §S5.6.2).

This dataset has previously been used to build classifiers that distinguish healthy EEGs from seizures using sets A and E. For example, one study reported a support vector machine classifier using operations derived from the discrete wavelet transform with a classification accuracy of at least 98.75 per cent using re-sampled subsegments of this dataset [13]. However, as shown in figure 5a, these two types of signals are separated automatically in the two-dimensional principal components space of operations, indicating that this task is relatively straightforward when exploiting a wide range of time-series analysis methods. Indeed, 172 different operations in our library each *individually* discriminated between healthy EEGs and seizures with a 10-fold cross-validation linear classification rate exceeding 95 per cent, with eight single operations exceeding 98.75 per cent (see the electronic supplementary material, §S5.4.3). These most successful operations are derived from diverse areas of the time-series analysis literature, and provide interpretable insights into this classification problem: e.g. revealing that EEGs recorded during seizures have lower entropy, lower correlation dimension estimates, lower long-range scaling exponents and distributions with a greater spread than those from healthy patients. In contrast to existing research that tends to focus on developing increasingly complicated classifiers for this problem (see the electronic supplementary material, §5.4.3, for details), the comparative analysis performed here selects simple and interpretable methods for the task automatically. Without performing such a comparison, it is difficult to assess whether methodological progress is being made in a given literature, or whether many simpler alternative methods may outperform the current state of the art.

4.2. Heart rate variability

We applied our highly comparative techniques to the problem of distinguishing 'normal sinus rhythm' and 'congestive heart failure' heartbeat (RR) interval series. Many existing studies have analysed RR interval data, each reporting the usefulness of a particular method, or a small set of methods [14]. By contrast, we show how a range of useful methods for this task can be selected from our library and organized in a way that synthesizes a large and disparate literature on the subject, as well as identifying promising new types of analysis techniques. Our dataset contains 105 recordings from each class, with lengths ranging from 800 to 19 900 samples (cf. electronic supplementary material, §S5.6.1). Although the analysis could

have been performed on a dataset of time series of a fixed length, here we show how informative operations can be retrieved even in the case where the time-series recordings are of very different lengths.

The 60 most successful operations for this task (those with a mean 10-fold classification rate exceeding 85%) were retrieved from our library and are represented as a clustered pairwise similarity matrix in figure 5b, where colour represents the linear correlation coefficient measured between the normalized outputs of all pairs of operations. Most operations cluster into three main groups of behaviour: measures of location (e.g. mean, median), entropy/complexity estimates and pNNx measures (the probability that successive RR intervals differ by more than x) [14] and linear correlation-based methods (e.g. autoregressive model fits and power spectrum scaling). Two other methods, in the bottom-left corner of the matrix in figure 5b, display relatively unique behaviour on the dataset: a discrete wavelet transform-based operation and an outlier-adjusted autocorrelation measure. Example distributions from each of these families of successful operations are plotted in panels of figure 5b: (i) an outlier measure returns the ratio of lag-3 autocorrelations of the time series before and after removing 10 per cent of outliers, (ii) the mean, (iii) the Sample Entropy [3], SampEn(3,0.05), and (iv) a variance ratio hypothesis test developed in the economics literature [15]. Despite being selected in a completely automated way, each selected operation provides an understanding of the dataset by contributing an interpretable measure of structural difference between the two classes of RR intervals. For example, the well-known results that normal sinus rhythm series tend to have longer interbeat intervals (higher mean, figure 5b(ii)) and greater entropy (higher sample entropy, figure 5b(iii)) than congestive heart failure series.

This case study demonstrates the ability of our highly comparative approach to select and also organize useful methods for time-series classification tasks automatically, using their empirical behaviour. The result provides interpretable insights into the differences between the known classes of time series. This case study represents another example of how an interdisciplinary methodological literature can be structured in a purely data-driven way (cf. figure 2b). However, unlike the treatment applied to general time-series analysis operations in §3.1, here additional knowledge (in the form of class labels assigned to the data) is used to guide the selection of useful methods, which are then organized according to their behaviour on the data. The resulting synthesis highlights similarities between different methods and distinguishes the novelty of others, and could be used to assess new contributions to the literature. For example, a new scientific paper could introduce the variance ratio hypothesis test [15]—a method originally developed in the economics literature—as a novel measure for analysing heart rate variability data. However, as shown figure 5b, the operation behaves like a range of simple linear model-based methods on this dataset, i.e. it is immediately clear that it is not actually measuring a new property of these time series, but is simply reproducing the behaviour of these existing operations. On the other hand, a new operation that we devised and introduce in this work, based on the impact of outliers on time-series autocorrelation properties, is distinguished as both unique and useful for this dataset (detailed information about all selected operations is in the electronic supplementary material, §S5.6.3). Referencing our comparative framework in this way

can thus help to ensure that new methods actually represent new contributions to a given analysis literature.

4.3. Self-affine time series

Fluctuation analysis has attracted substantial attention in the statistical physics literature, and has been used to provide evidence for scale-invariance in a variety of real-world processes. But, do these conventional methods outperform alternative approaches, or do simpler, faster methods exist that display comparable (or even superior) performance? We addressed this question by generating a synthetic dataset of self-affine time series and then searching our library (in a statistically controlled way, cf. electronic supplementary material, §S1.3.6) for operations that vary linearly with their known scaling exponents (as depicted schematically in figure 2f). Self-affine time series can be characterized by a single scaling exponent, α, according to $S(f) \propto f^{-\alpha}$, where S is the power spectral density as a function of frequency, f [16]. Time series of 5000 samples were generated with scaling exponents uniformly distributed in the range $-1 \leq \alpha \leq 3$ by two different methods: (i) 199 time series generated using the *Fourier filtering method* and (ii) 199 time series generated using the *random midpoint displacement method* [16] (cf. electronic supplementary material, §S4.3), as shown in figure 5c. In order to accurately characterize the scaling exponent, α, of these time series, operations must combine local and global information to capture the self-similarity of the time series across multiple time scales.

A selection of five operations with the strongest linear correlations to α are plotted in figure 5d. As expected, many fluctuation analysis-based operations perform well on this task (cf. electronic supplementary material, §S4.3), including the scaling exponent estimated using detrended fluctuation analysis (DFA) [1], and a wavelet-based alternative (labelled 'DFA' and 'wavelet scaling' in figure 5d, respectively). However, other types of operations that are not based on fluctuation analysis also exhibit strong linear correlations with α: the autocorrelation of residuals from a local mean forecaster (labelled 'mean forecaster' in figure 5d), the first-order coefficient of an autoregressive AR(3) model fitted to the time series (labelled 'AR(3) coefficient') and the mean log hyperparameter of the squared exponential length scale from a Gaussian process regression on local segments of the time series (labelled 'Gaussian process regression').

Thus, as well as confirming the utility of methods based on fluctuation analysis to estimate the scaling exponent of self-affine time series, we also discovered a surprising selection of other useful methods that capture the known variation in α by combining local and global structure in various ways. Many of these alternative methods require significantly less computational effort and can be updated iteratively, making them more suited to real-time applications than the slightly more accurate but also more computationally intensive fluctuation-based methods. Our highly comparative approach to time-series regression, which selects relevant scientific operations based on their empirical performance, can thus be used to find fast approximations to traditional methods for real applications involving finite, noisy time series. This general approach could also be used to select methods that help predict important diagnostic quantities assigned to physiological recordings, such as predicting the stage of sleep from an EEG recording or the arterial pH of a baby from its foetal heart rate time series recorded during labour. Additional examples are presented in the electronic supplementary material, where we successfully selected interpretable estimators of the variance of white noise added to periodic signals (see the electronic supplementary material, §S4.1) and the Lyapunov exponent of logistic map time series (see the electronic supplementary material, §S4.2). The success of our highly comparative approach relies on our library of operations being sufficiently comprehensive to produce useful and informative results.

4.4. Parkinsonian speech

This final case study involves the particularly challenging task of distinguishing stationary phonemes recorded from patients with Parkinson's disease from those of healthy controls [17]. Rather than implementing a specific set of standard speech analysis techniques, the structure in the data was used to select appropriate operations from our general library. Using the procedure illustrated in figure 2f, we found that our operation library is rich enough to construct useful classifiers for this difficult task. The dataset contains 127 speech recordings from Parkinsonian patients and 127 speech recordings from healthy controls, with recording lengths ranging from 50 ms to 1.25 s (cf. electronic supplementary material, §S5.5).

We found that many of the most successful operations in our library for classifying Parkinsonian speech were closely related to existing measures from speech analysis (e.g. the 'jitter' summary statistic [17]), and we also identified some new techniques (cf. electronic supplementary material, §S5.5.2). However, rather than restricting ourselves to individual operations, we also used forward feature selection to construct classifiers that combine pairs of complementary operations to achieve improved out-of-sample classification performance (using appropriate partitions of data into training and test sets, as described in the electronic supplementary material, §S1.3.5). An example two-feature linear discriminant classifier is shown in figure 5e, and combines a new operation introduced by us in this work, which simulates an inertial particle that experiences an attractive force to the time series, with another operation that constructs a symbolic string from incremental differences of the time series using a three-letter alphabet ('A', 'B', 'C'), and returns the probability of the word 'AC'. The two operations complement one another and the resulting classifier is simple, interpretable and has a mean 10-fold cross-validation misclassification rate of 14 per cent, which is comparable with the state of the art in this literature (that typically treats unbalanced datasets of fixed-length time series [17]; cf. electronic supplementary material, §S5.5.2). Using our method, multi-feature classifiers for time-series classification are constructed automatically, require no domain knowledge of the mechanisms underlying the time series, and can be extended to classifiers containing three or more operations straightforwardly (although for this dataset, we found minimal out-of-sample improvement in classification rate on adding more than two features, cf. electronic supplementary material, §S5.5.3).

To further demonstrate the applicability of our methods to diverse types of time-series recordings, we analysed two additional datasets in the electronic supplementary material, where an unknown seismic recording was classified as an explosion rather than earthquake, consistent with previous studies (see the electronic supplementary material, §S5.2), and competitive multi-feature classifiers were constructed

for distinguishing seven different classes of emotional content in human speech recordings (see the electronic supplementary material, §S5.3). Thus, despite using extremely simple statistical learning techniques (including linear classifiers, forward feature selection, linkage and k-medoids clustering and principal components analysis), we are able to contribute meaningful results to a wide variety of scientific time-series analysis problems. These simple methods have the advantage of being transparent and producing readily interpretable results to demonstrate our approach; more sophisticated methods should yield improved results and can be explored in future work. On several occasions, we also found that new operations developed by us were among the most useful operations for various analysis tasks, including the outlier autocorrelation measure selected for classifying heartbeat interval data in §4.2 and the particle trajectory-based operation for Parkinsonian speech in §4.4. This highlights the benefit of being creative in the development of new operations for our library, as useful operations are selected based on their performance on real datasets. The results of this section demonstrate that our library of operations is sufficiently comprehensive to be broadly useful in guiding the selection of methods for scientific time-series analysis tasks.

5. Conclusions

In summary, we have shown that interesting and scientifically meaningful structure can be discovered automatically in extensive annotated collections of time-series data and time-series analysis methods from a wide selection of theoretical and empirical literatures. Relationships are determined by comparing empirical behaviour: the outputs of the methods applied to data, and the properties of the data as measured by the methods. Representing time series and operations in this way turns out to provide a powerful framework for organizing general collections of time series and operations. By unifying a previously disjoint interdisciplinary literature, we thus motivate a complementary and highly comparative approach to time-series analysis that provides insights into the properties of time series studied in science and the techniques that scientists have developed to study them.

Using our framework, data analysts can now readily ask new types of questions of their time-series data and methods, as depicted in figure 2. In §3.1, we demonstrated how time-series analysis operations can be represented using their outputs across a controlled set of 875 real and model-generated time series from across the sciences. This form of empirical fingerprinting allowed us to structure a large and diverse library of scientific methods in a meaningful way. The result provides time-series analysts with a new means of placing the methods they use to analyse their data in a wider scientific context. For example, data analysts can now investigate relationships between the set of methods familiar to them and extensive libraries of alternative methods that may have been developed in different disciplinary contexts. Furthermore, new methods for time-series analysis can now be compared with the existing literature (as represented in our library) to check for redundancy, and hence help ensure that they really constitute advances and do not simply reproduce existing behaviour. In §3.2, using the output of 200 diverse operations as a form of empirical fingerprint for a time series allowed us to identify meaningful

clusters from a large and diverse library, and to retrieve meaningful matches to a given target time series. For example, on receiving a new dataset, a time-series analyst may wish to understand what key structures exist in it, or what other types of real-world and model-generated time series have similar types of dynamics: this can now be achieved in an automated, data-driven fashion. By structuring datasets and connecting them to relevant types of real-world and model systems, the results provide an interdisciplinary context for the problem that can be used to guide a more focused analysis.

As well as providing useful tools for understanding the structure in large collections of time series and their methods, in §4, we showed how additional knowledge about particular datasets can be incorporated to automate the selection of useful analysis methods. Because each operation provides an interpretable measure of some kind of structure in the time series, selecting operations in this way yields insights into the most informative properties of the data and how they vary across the labelled classes. The operations are selected according to their behaviour, and the result connects a range of ways of thinking about structure in time series to a given analysis problem. Using EEG seizure data (§4.1), we showed how difficult it can be to assess existing studies that quote classification rates without comparison with alternative methods and demonstrated how our library of operations can be used to perform this comparison. In §4.2, we retrieved and organized operations that help to distinguish healthy and congestive heartbeat intervals by comparing the behaviour of all operations on the data. Each operation provides an interpretable measure of structural difference between the two classes, and the organization reveals which groups of operations exhibit similar behaviour on the dataset and which are unique (figure 5b). These results suggest that only a reduced number of operations need to be measured to characterize and classify the heartbeat interval data, information that could be used by a domain expert to guide the selection of a reduced set of unique analysis operations. Further investigation of the selected operations by a domain expert could also provide new insights into the dynamical structure of heartbeat intervals for healthy patients and for various pathologies. A similar approach was used to select operations that most accurately predict the scaling exponent, α, of self-affine time series in §4.3, and we also demonstrated how forward feature selection can be used to construct interpretable multifeature classifiers for identifying Parkinsonian speech recordings. The success of our approach in each case relied on our library of operations being sufficiently comprehensive. Thus, although far from complete, our library appears to contain enough diversity to provide good classification performance and contribute useful insights into a range of time-series analysis problems. The results could be improved further in the future by refining and growing the library with methods contributed by time-series analysis experts.

We now revisit the analogy to the high-throughput analysis of the DNA microarray in biology, which is used routinely to guide more focused research efforts in that field. While time-series analysis has traditionally focused on the use of specific methods and models motivated by domain knowledge, the highly comparative techniques introduced in this work represent a similarly powerful complement to this approach. As vast quantities of time-series data continue to be recorded, and new methods for their analysis are developed, this highly comparative approach will allow us to

make sense of this large and complex resource and contribute to directing progress in an inherently interdisciplinary field. The results are wide-reaching, from the diagnosis of pathologies in medical recordings to the detection of anomalies in an industrial process or on an assembly line. We hope that our work, while stimulating theory, adds to the experimental and empirical aspect of the study of time-series data. The Matlab source code for all operations used in this work can be obtained from the authors via the website http://www. comp-engine.org/timeseries/, and instructions of how our framework can be applied to new datasets are summarized in the electronic supplementary material, §S1.4.

References

1. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simons M, Stanley HE. 1995 Statistical properties of DNA sequences. *Physica A* **221**, 180–192. (doi:10.1016/0378-4371(95)00247-5)

2. Bollerslev T. 1986 Generalized autoregressive conditional heteroskedasticity. *J. Econ.* **31**, 307–327. (doi:10.1016/0304-4076(86)90063-1)

3. Richman JS, Moorman JR. 2000 Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **278**, H2039–H2049.

4. Liao TW. 2005 Clustering of time series data—a survey. *Pattern Recognit.* **38**, 1857–1874. (doi:10.1016/j.patcog.2005.01.025)

5. Wang X, Mueen A, Ding H, Trajcevski G, Scheuermann P, Keogh E. 2012 Experimental comparison of representation methods and distance measures for time series data. *Data Min. Knowl. Discov.* **26**, 275–309. (doi:10.1007/s10618-012-0250-5)

6. Hastie T, Tibshirani R, Friedman J. 2009 *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn. Berlin, Germany: Springer.

7. Pincus SM, Gladstone IM, Ehrenkranz RA. 1991 A regularity statistic for medical data analysis. *J. Clin. Monit. Comp.* **7**, 335–345. (doi:10.1007/BF01619355)

8. Tenenbaum JB, de Silva V, Langford JC. 2000 A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323. (doi:10.1126/science.290.5500.2319)

9. Sprott JC. 2003 *Chaos and time-series analysis*. New York, NY: Oxford University Press.

10. Samuelson P. 1965 Rational theory of warrant pricing. *Ind. Manage. Rev.* **6**, 13–31.

11. Freitas US, Letellier C, Aguirre LA. 2009 Failure in distinguishing colored noise from chaos using the 'noise titration' technique. *Phys. Rev. E* **79**, 035201(R). (doi:10.1103/PhysRevE.79.035201)

12. Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE. 2001 Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Phys. Rev. E* **64**, 061907. (doi:10.1103/PhysRevE.64.061907)

13. Subasi A, Ismail Gursoy M. 2010 EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Syst. Appl.* **37**, 8659–8666. (doi:10.1016/j.eswa.2010.06.065)

14. Malik M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, Schwartz PJ. 1996 Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Eur. Heart J.* **17**, 354–381. (doi:10.1093/oxfordjournals.eurheartj.a014868)

15. Cecchetti SG, Lam P-s. 1994 Variance-ratio tests: small-sample properties with an application to international output data. *J. Bus. Econ. Stat.* **12**, 177–186. (doi:10.1080/07350015.1994.10510006)

16. Peitgen H-O, Saupe D, Fisher Y, McGuire M, Voss RF, Barnsley MF, Devaney RL, Mandelbrot BB. 1988 *The science of fractal images*. Berlin, Germany: Springer.

17. Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. 2009 Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Trans. Biomed. Eng.* **56**, 1015–1022. (doi:10.1109/TBME.2008.2005954)