INTERFACE FOCUS

royalsocietypublishing.org/journal/rsfs

Research

Cite this article: Aminian M, Andrews-Polymenis H, Gupta J, Kirby M, Kvinge H, Ma X, Rosse P, Scoggin K, Threadgill D. 2019 Mathematical methods for visualization and anomaly detection in telemetry datasets. *Interface Focus* **10**: 20190086.
http://dx.doi.org/10.1098/rsfs.2019.0086

Accepted: 4 November 2019

One contribution of 9 to a theme issue 'Multi-scale dynamics of infectious diseases'.

Subject Areas:
bioinformatics, biomathematics, computational biology

Keywords:
high-dimension time series, temperature telemetry data, Radial Basis Functions, MSET, Multivariate State Estimation Technique

Author for correspondence:
Manuchehr Aminian
e-mail: manuchehr.aminian@colostate.edu

THE ROYAL SOCIETY PUBLISHING

Mathematical methods for visualization and anomaly detection in telemetry datasets

Manuchehr Aminian[1], Helene Andrews-Polymenis[2], Jyotsana Gupta[2], Michael Kirby[1], Henry Kvinge[1], Xiaofeng Ma[1], Patrick Rosse[1], Kristin Scoggin[3] and David Threadgill[3]

[1]Department of Mathematics, Colorado State University, Fort Collins, CO, USA
[2]Department of Microbial Pathogenesis and Immunology, and [3]Department of Molecular and Cellular Medicine, Texas A&M University, College Station, TX, USA

HA-P, 0000-0002-1818-1335; MK, 0000-0001-9802-9263; HK, 0000-0003-4108-1364

Recent developments in both biological data acquisition and analysis provide new opportunities for data-driven modelling of the health state of an organism. In this paper, we explore the evolution of temperature patterns generated by telemetry data collected from healthy and infected mice. We investigate several techniques to visualize and identify anomalies in temperature time series as temperature relates to the onset of infectious disease. Visualization tools such as Laplacian Eigenmaps and Multidimensional Scaling allow one to gain an understanding of a dataset as a whole. Anomaly detection tools for nonlinear time series modelling, such as Radial Basis Functions and Multivariate State Estimation Technique, allow one to build models representing a healthy state in individuals. We illustrate these methods on an experimental dataset of 306 Collaborative Cross mice challenged with *Salmonella typhimurium* and show how interruption in circadian patterns and severity of infection can be revealed directly from these time series within 3 days of the infection event.

1. Introduction

Body temperature is a basic vital sign that has long been used for evaluating the health of an individual. Though it is often treated as a single statistic, the richness of temperature as a vital sign comes when we consider it as a function of time. In most organisms, body temperature follows a 24 h circadian rhythm. Deviations in this pattern can often indicate changes in the underlying health state. After exposure to a pathogen that initiates a host immune response to infection, the signatures of the healthy temperature data may trend away from this circadian pattern. The impact of exposure to a pathogen may not be visibly apparent in the host's physical behaviour for 2–3 days or more after infection, but, as we will see, temperature time series may reveal significant anomalies even within the first 24 h.

The investigation of the temporal evolution of body temperature as a predictor of health state has attracted considerable interest. In [1], for example, various features were manually extracted from temperature time series taken from febrile critically ill patients in an intensive care unit, and these were used to predict whether the patient had sepsis. In [2], on the other hand, the complexity of the temperature time series was used to predict the outcome of critically ill patients, while in [3] the regularity of the time series (based on approximate entropy) was used to predict survival of critically ill patients. What these works together show is that sufficiently dense temperature monitoring enables one to extract fairly specific information about the biological state of the individual organism. The recent development of new machine learning tools for time-series classification and prediction (e.g. [4,5]) will facilitate the large-scale analysis of biological mechanisms related to temperature profiles and the host immune response to infection.

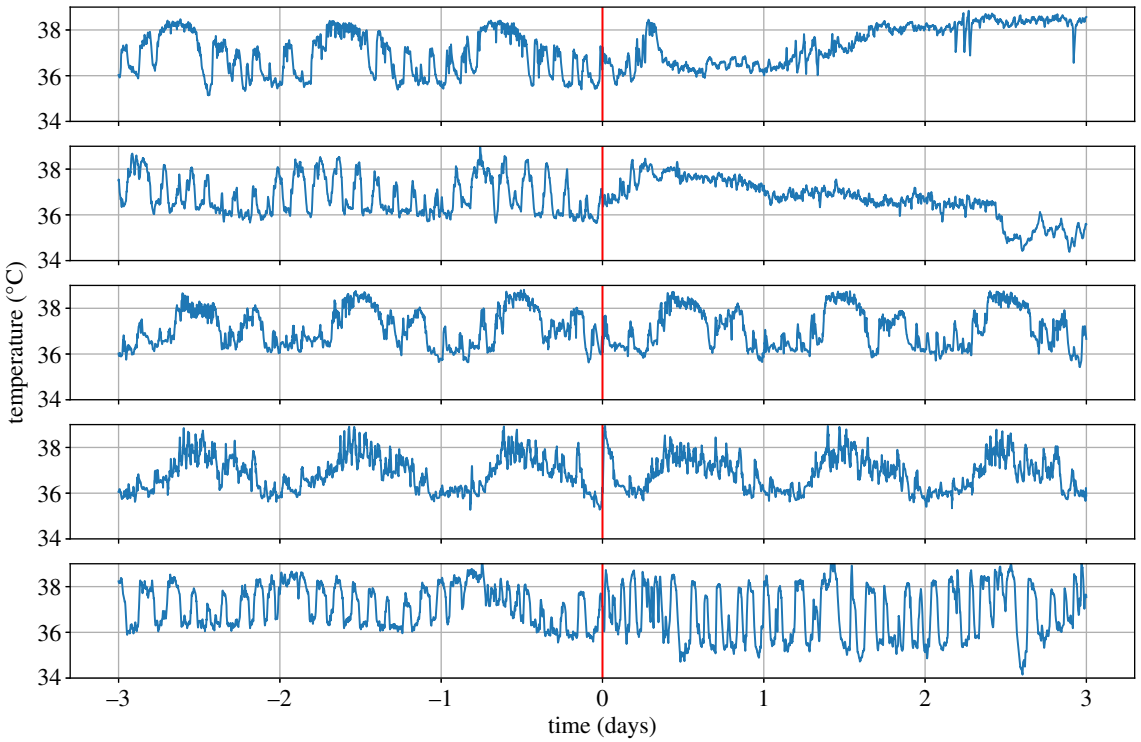

Figure 1. Temperature profiles for five different mouse prototypes aligned relative to the time of infection (red). Significant qualitative differences in patterns before and after infection are observed. (Online version in colour.)

We will explore a collection of mathematical and machine learning techniques to visualize and build models of these time series, either individually or as a collection, while using very little prior knowledge of the data. We consider several approaches to answering two fundamental questions. First, what quantitative notions of similarity between temperature time series do well in capturing relationships in the underlying health state between individuals? Stated another way, if body temperature reveals information about health, then how can we quantitatively establish similarity between two individuals whose body temperatures follow similar patterns? And does this imply a similarity in their underlying health state? Second, from a practical point of view, can we identify anomalies in temperature time series that can be characterized as a deviation from the healthy state?

We have observed that simple measures such as increased mean temperature (e.g. a fever) provide an incomplete picture here—severe interruptions to healthy behaviour may occur while the mean temperature is nearly constant. Additionally, we observe a broad spectrum of responses to infection in the data beyond changes of the mean temperature. Figure 1 illustrates some of this broad variation. The data are focused on an interval from 3 days before to 3 days after infection, aligned to the time of infection marked in red. Prior to infection, the dominant mode is approximately circadian—periodic with period 24 h—but the strength of this mode relative to oscillations on shorter time scales may vary greatly depending on the individual mouse. After infection, a variety of responses are possible: from top to bottom, elevated temperature, depressed temperature and potentially no noticeable effect in the mean temperature or even in the oscillations on various time scales. This illustrates a need for careful visualization and anomaly detection tools tailored to the individual.

From a mathematical perspective, we view a time series as having potential structure, signatures or patterns that serve to characterize the biological state. As an observable, temperature may be viewed as a projection onto one dimension of this *latent* geometry that can be recovered using time-delay embeddings [6,7]. Recent developments in machine learning and data fitting make the exploration of this structure on large time-series datasets more tractable, creating the potential for new discoveries.

This paper is structured as follows: in §2, we describe the dataset we study as well as pre-processing techniques that we employed. Our first analysis in §3 explores how visualization techniques can be brought to bear on these datasets in bulk. Laplacian Eigenmaps, a tool for identifying low-dimensional patterns in datasets, can be used to identify distinct signatures in the collection of mice post-infection. Specialized metrics and Multidimensional Scaling (MDS) can tease out associations between these time series and the severity of infection as measured by the severity of infection by *Salmonella* upon necropsy. In §4, we introduce techniques for detecting anomalies in the time series to identify interruptions in circadian patterns post-infection. Finally, we conclude in §5.

2. The data and pre-processing

The dataset we consider is a set of time series collected from 306 Collaborative Cross mice [8,9] challenged with *Salmonella typhimurium* as part of a broader study conducted by the Andrews-Polymenis and Threadgill laboratories at Texas A&M University, College Station, TX, USA, to better understand the broad range of host–pathogen dynamics in mice with *Salmonella*. As the scope of this paper is in methods of analysing time series, rather than conducting a full analysis of this experiment, we only briefly address the experimental background information here.

Several months prior to inoculation, each mouse is surgically implanted with a telemetry device which serves to measure body temperature and net movement (termed 'activity') of the mouse once per minute continuously during the experiment. The

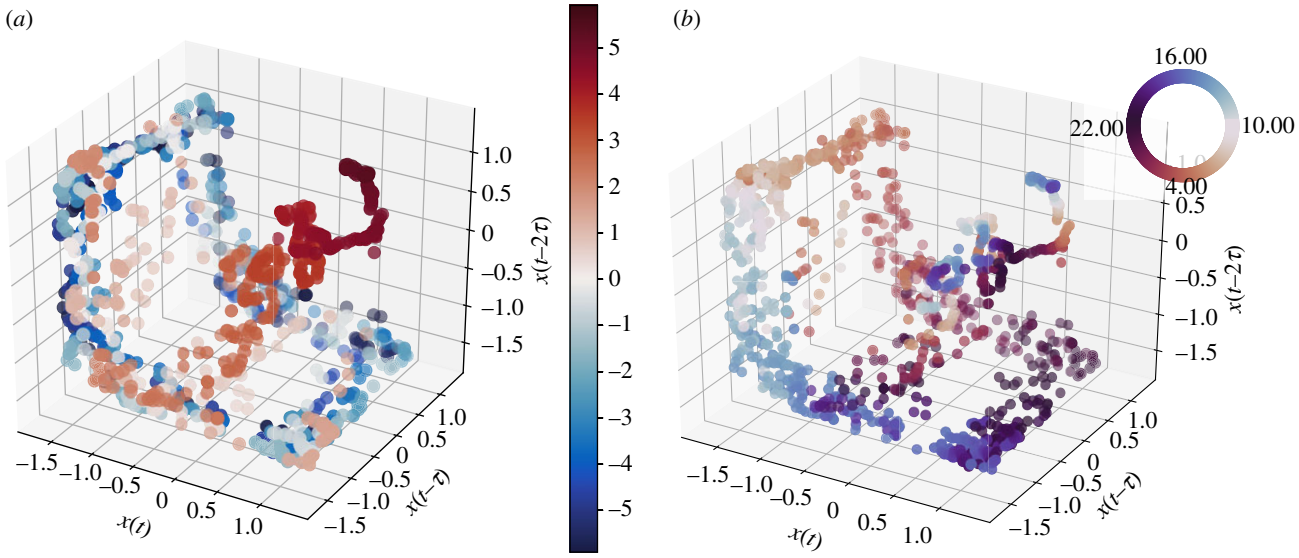

Figure 2. Time-delayed embedding of mouse CC004-021. Data are smoothed via median filtering on a 3 h window for the purposes of visualization. (*a*) Colour indicates time in days relative to time of inoculation. Observe pre-infection data (blues) traverse a periodic loop. (*b*) Same embedded data are coloured by time of day; state is predictable while healthy, with relative location dependent on time of day. Interruption of circadian patterns after infection corresponds with a departure from this loop. (Online version in colour.)

mouse is then allowed to recover from the surgery so as not to impede immune response during the experimental phase. Experiments run between approximately 14 and 28 days. Approximately 7 days of telemetry data are recorded prior to infection for all mice. Each mouse is then inoculated with a dosage of *S. typhimurium* and observation continues for the remaining 7–21 days of the experiment. While in §3 we illustrate the power of visualization tools equipped with special metrics to associate time series with severity of *Salmonella* infection, we otherwise limit our attention in this paper to the temperature time-series data in isolation and corresponding methods for visualization and anomaly detection.

2.1. Data filtering

The time series are processed in several ways throughout the paper. Time series, whether as a whole or windowed, are discarded entirely if there is excessive missingness of data, ranging from 120 to 240 min of missingness in the data. This resulted in keeping between 131 and 185 time series depending on the visualization or anomaly detection task. With the exception of §§3.2 and 4.4, we primarily use median filtering included in the `scipy` package in Python with `kernel_size=31`. In §§3.2 and 4.4, we use wavelet smoothing with Haar wavelets with provided discrete and inverse discrete wavelet transform in the `PyWavelets` package; we discard detail coefficients beyond the sixth degree, which corresponds to the filtering of time scales smaller than 64 min. We have not seen any significant differences in the results of the algorithms depending on the choice of median filtering or wavelet smoothing.

2.2. Time-delayed embedding

For the purpose of analysing scalar time series $x(t)$, we frequently use time-delayed embeddings to produce a vector representation,

$$x(t) \mapsto \hat{\mathbf{x}}(t) = (x(t), x(t-\tau), \ldots, x(t-(d-1)\tau)), \quad t \in \mathbb{R}, \ \tau > 0, \tag{2.1}$$

with positive delay parameter τ and prescribed embedding dimension d. Time-delayed embeddings have theoretical guarantees via Takens' embedding theorem [7] when an observable $x(t)$ is part of a multivariate deterministic system $\mathbf{x}(t)$. Effectively, if the time series is a scalar sample from an m-dimensional manifold, the dynamical system may be reconstructed in a Euclidean space of dimension $2m+1$. It is natural to consider $m=1$ for periodic temperature time series which may be viewed topologically as a circle.

The choice of delay τ and embedding dimension d is dependent on the nature of the phenomenon being studied and the number of observable variables. For a scalar $x(t)$ with $t \in \mathbb{R}$, a common choice is to choose τ to be the first zero of the autocorrelation,

$$\tau = \min_{s>0}\left\{ s : 0 = \int_D x(t)x(t-s)\,\mathrm{d}t \right\}, \tag{2.2}$$

for some domain of integration D. For example, with a sinusoid $x(t) = \sin(2\pi t/T)$, a calculation by hand shows $\tau = T/4$; that is, the first zero-autocorrelation time is a quarter-period. Our studies of the numerical autocorrelation in mouse time series result in a value of τ between 4 and 12 h, varying from mouse to mouse. Supposing a 24 h period for a mouse time series is typical, the quarter-period reasoning suggests that a delay of $\tau = 6$ h is reasonable for our studies below.

To give a sense of the typical behaviour, we visualize the three-dimensional time-delayed embedding of a mouse's time series in figure 2. Smoothing and sub-sampling are employed to improve visibility of the topological loop pre-infection. In figure 2*a*, shades of blue and red indicate time before or after infection, respectively. The pre-infection data navigate the loop, but eventually exit the trajectory after infection, as seen in the deep reds.

This loop indeed corresponds to the circadian pattern of the mice, as can be seen in figure 2*b*, with a robust colouring along the loop associated with the time of day.

2.3. Pre-processing of the experimental data

Here we describe the pre-processing involved and introduce notation used throughout these sections. Because of the nature of the varying length of experiments, either by early

termination due to succumbing to infection or simply by varying length of the experiments, the time series come in a variety of lengths ranging from 15 833 to 40 313 min. Very few algorithms can work with time series of different lengths in computing pairwise similarities, identifying clusters, etc., and those that do often implicitly map time series to a common latent space before computing similarities. Hence, it is natural to restrict attention to a common interval of interest. We will apply such restrictions frequently in this paper, specifying the time interval which we restrict ourselves to in each case.

3. Visualization techniques for collections of time series

In this section, we review a few tools to visualize collections of these time series in aggregate. Visualization on this scale is a critical step in the analysis of these data for identifying outliers and revealing, or at least confirming, expected mathematical structure to the data. A recurring theme we will see in this section is the presence of topological loops associated with the data, especially during the healthy phase. We explore the utility of Laplacian Eigenmaps to reveal this information on windows of the data in §3.1. In §3.2, we investigate a technique and the mathematical interpretation of constructing similarity measures between the time series to account for differences in phase between the mice on the dominant 24 h mode. Initial experiments show that this does reveal some of the underlying health state—specifically the severity of infection in the associated mice upon necropsy.

3.1. Comparing signatures with Laplacian Eigenmaps

Laplacian Eigenmaps is a nonlinear dimensionality reduction method that aims to take a high-dimensional dataset potentially living on a low-dimensional manifold and provide a representation of that data in a low-dimensional space [10]. After building a graph from neighbourhood information in the dataset $\mathbf{X}$ which contains uniformly sized data—in our case, time series restricted to a window $[a, b]$ relative to the time of infection—the graph Laplacian is used to compute a low-dimensional representation $\widetilde{\mathbf{X}}$ of the dataset that optimally preserves local neighbourhood information. To achieve this, the algorithm seeks to minimize the following objective function, in general a weighted sum over pairwise distances in the embedded space:

$$\sum_{0\le i,j\le|\mathbf{X}|-1} \|\widetilde{\mathbf{X}}_i - \widetilde{\mathbf{X}}_j\| \boldsymbol{W}_{ij}, \tag{3.1}$$

under the appropriate constraints. Note that here $\widetilde{\mathbf{X}}_i$ and $\widetilde{\mathbf{X}}_j$ are elements of $\widetilde{\mathbf{X}}$, and $\boldsymbol{W}$ is a weight matrix which depends on the variation of the algorithm. It can be seen that, with the appropriate choice of weights, the objective function incurs a heavy penalty if neighbouring points $\mathbf{X}_i$ and $\mathbf{X}_j$ are mapped far apart. Thus, this mapping attempts to preserve the pairwise distances between all points $\mathbf{X}_i$, $\mathbf{X}_j$ in $\mathbf{X}$; that is, if $\mathbf{X}_i$ is close to $\mathbf{X}_j$, then we typically expect $\widetilde{\mathbf{X}}_i$ will be close to $\widetilde{\mathbf{X}}_j$. Once we have chosen an interval of time that determines what time-series windows will be in $\mathbf{X}$, we use Euclidean distance to generate an adjacency matrix $\boldsymbol{W}$ for all the points in $\mathbf{X}$. Using a fixed number of k closest neighbours to a point $\mathbf{X}_i$ in $\mathbf{X}$ (we chose $k = 5$ in the experiments in this paper), we populate $\boldsymbol{W}_{ij}$ with a 1 if $\mathbf{X}_i$ belongs to the set of nearest neighbours of $\mathbf{X}_j$. Otherwise, $\boldsymbol{W}_{ij} = 0$. We then generate a diagonal weight matrix $\boldsymbol{D}$ with $\boldsymbol{D}_{ii} = \sum_j \boldsymbol{W}_{ij}$. This allows us to construct the graph Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, which is a symmetric, positive semidefinite matrix. Lastly, we solve the generalized eigenvector problem

$$\boldsymbol{L}\mathbf{v} = \lambda \boldsymbol{D}\mathbf{v}. \tag{3.2}$$

Let $\mathbf{v}_0, \ldots, \mathbf{v}_{|\mathbf{X}|-1}$ be the eigenvectors of equation 3.2 ordered according to their eigenvalue size

$$0 = \lambda_0 \le \lambda_1 \le \lambda_{|\mathbf{X}|-1}.$$

We leave out the eigenvector $\mathbf{v}_0$ corresponding to eigenvalue $\lambda_0 = 0$ and use the next m eigenvectors for embedding in m-dimensional Euclidean space. For our visualizations, we choose $m = 2$. For a detailed exposition of the algorithm, along with possible variations, see [10].

Now we use this tool on the dataset. To distinguish qualitative differences in the healthy and infected signatures, we work with the data on 1 day windows and study the qualitative structure of the embedded data as the window moves from before to after infection. Figure 3 illustrates the results.

Recall that, for this experiment, we use Euclidean distances to determine neighbouring points. Figure 3*a* shows the embedding in two dimensions; figure 3*b* shows the corresponding time series for seven mice selected uniformly from the embedded structure following inoculation (figure 3*a*(iii)). Panels in figure 3*a* and intervals in figure 3*b* are associated with one another and colour-coded appropriately.

The first window (green shading) starts 3 days before inoculation and end 2 days before inoculation. These are considered the healthy parts of the time series. As shown by the time series in figure 3*b*, most of the mice during this time window exhibit some sort of circadian rhythm. This is what causes the circular shape of the embedding; our studies into synthetic examples with collections of sinusoids (not shown here) have produced similar behaviour.

This circle retains its structure as the window continues forward to the time of inoculation. However, when the sampling window crosses through the time of inoculation (yellow shading), we see that the circle collapses with no obvious structure. The immediate responses to infection are varied and inconsistent. As the sampling window is moved to 2 days after infection (red shading), we observe a quite different 'V'-shaped structure in the embedding. When comparing the representative time series of mice along this line in the red highlighted regions, we see explicit differences in the behaviours; we refer to these as signatures. Mouse CC043-078, for example, has highly erratic behaviour which persisted from prior to infection. Tracing along the embedded 'V' from the top left down to the centre and back to the upper right (top to bottom in figure 3*b*), we see a range of behaviours trending generally from erratic to stable, preserved oscillations with no mean temperature elevation, to complete interruption and consistently high temperature.

3.2. Modified correlation distance and Multi-dimensional Scaling

Our studies with Laplacian Eigenmaps have shown that distinctive signatures in the response to infection can be successfully identified using visualization techniques. In this section, we investigate various notions of *dissimilarity* between pairs of time series and show how using this in conjunction with MDS [11] can reveal further structure in the data.

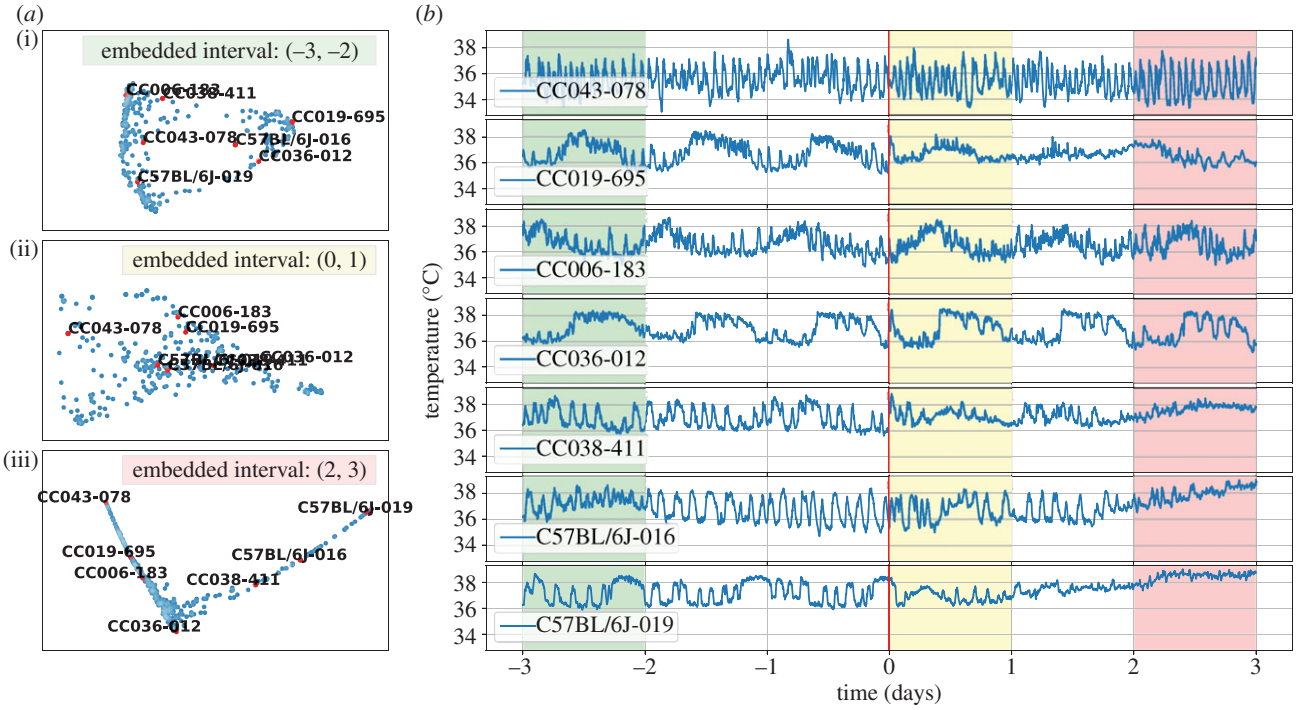

Figure 3. Laplacian Eigenmaps algorithm applied to all mouse time series truncated by a sliding 1 day window with various starting points. (*b*) Seven selected time series; visualizing region of data used in Laplacian Eigenmaps embeddings in the corresponding panels in (*a*). Note Laplacian Eigenmaps only preserves neighbourhoods in the data; the axes in the embedded space have no significance. For this implementation of the experiment, axes are immaterial for the purposes of portraying what we are trying to accomplish. (Online version in colour.)

Consider the set $\mathbf{X}$ of time series restricted to a time window $[a, b]$ relative to infection. One would often like to understand whether these can be clustered in a way such that distance between time series correlates with difference in reaction to infection. In order to do this, one needs to decide what it means for pairs of time series to either be 'close together' or 'far apart'; namely, we need to identify a distance function or, more precisely, dissimilarity $d : \mathbf{X} \times \mathbf{X} \to \mathbb{R}_{\geq 0}$ that assigns a non-negative number to each pair of time series. Because time series in $\mathbf{X}$ are simply vectors in $\mathbb{R}^m$, where m is the number of time steps from a to b, there are many known choices for dissimilarity, including: ℓ_p-distances, dynamic time warping distance and correlation distance. Each of these captures some aspect of similarity between elements of $\mathbf{X}$, but is arguably ineffective for our purposes. ℓ_p distances can have unintuitive behaviour when working with the original time series as the inputs. Dynamic time warping, which has been quite successful in applications such as speech recognition, requires many assumptions about the data which may not be appropriate for this application.

Given these considerations, here we begin with correlation distance d_c defined between pairs of time series x, $y \in \mathbf{X}$,

$$d_c(x, y) := 1 - \frac{(x - \bar{x})^{\mathrm{T}}(y - \bar{y})}{\|x - \bar{x}\|_2 \|y - \bar{y}\|_2}, \tag{3.3}$$

where $\bar{x}$ and $\bar{y}$ denote the mean of x and y, respectively. We will see shortly that d_c captures the standard circadian rhythm found in the time series of healthy mice. Since disruption of the circadian rhythm is a basic reaction to infection, one can take this as evidence that d_c is measuring properties of x and y that are also related to the response to infection.

Our approach for visualization in this subsection is to use a notion of dissimilarity followed by MDS [11]. MDS is an algorithm which attempts to represent the data in low dimensions (two or three dimensions for visualization), while preserving the distances, dissimilarities, etc., between pairs of points as best as possible. Because MDS only requires a matrix $\boldsymbol{D}$ as input which measures dissimilarity between all pairs of points, it has great utility in scenarios where visualizations of high-dimensional data may not be appropriate or one does not have a good sense of how to construct a *bona fide* metric on the space.

We are interested in techniques which can easily associate all data obeying a plain circadian rhythm in a single 'healthy' cluster, with outliers representing those mice succumbing to infection. The primary obstacle to this is the fact that correlation distance d_c does not recognize two time series with a standard circadian rhythm, but differing in phases, to be 'close'. To this end, let $H \subseteq \mathbf{X}$ be all those time series that have been determined to exhibit a 'healthy' circadian rhythm. Define

$$\widehat{H} = \{g^k \cdot \mathbf{x} \,|\, 0 \leq k \leq m, \mathbf{x} \in H\},$$

where g represents the operation of a right circular shift of the vector x; hence g^k represents a circular shift of length k. This represents all possible circular shifts of all mouse time series deemed to be healthy. Mathematically, this set $\widehat{H}$ is the orbit of H with respect to the action of C_m, the cyclic group of order m. We will revisit this notion near the end of this subsection. With this structure in place, we propose to account for phase differences in comparing such time series by using a modified dissimilarity

$$d_{qc}(x, y) = \min\left(d_c(x, y), \min_{h \in \widehat{H}} (d_c(x, h) + d_c(y, h))\right). \tag{3.4}$$

Since $\widehat{H}$ will generally be a very large set, we found that in practice $\widehat{H}$ can be reasonably approximated by the set of

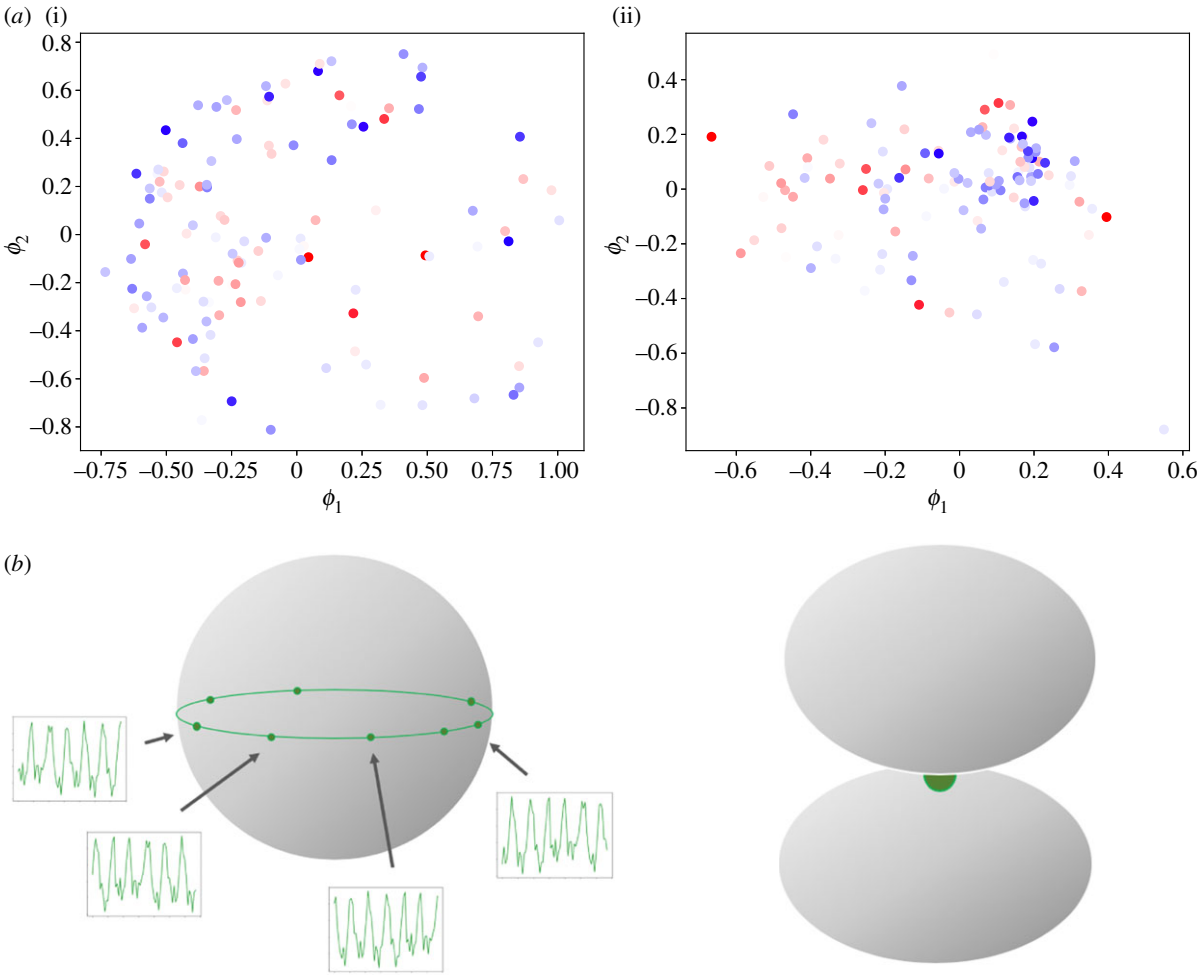

Figure 4. MDS results for two choices of dissimilarity and time windows. Row (*a*) uses correlation distance d_c (i) and the modified function d_{qc} (ii), which accounts for phase differences in the data. Axes are components of the scaled eigenvectors associated with MDS; these are different for each plot. Points are coloured by the severity of infection by *Salmonella* upon autopsy, with darker shades of blue being less colonized and red being more colonized, suggesting this dissimilarity d_{qc} could be a tool in separating health states. Mathematically, the expected effect of the metric is to *mod out* phase differences in the group structure; see the text for a discussion of the illustration in (*b*). (Online version in colour.)

all cyclic shifts of a single prototypical time series for a healthy mouse.

In figure 4*a*, we show the MDS approximation in $\mathbb{R}^2$ with respect to correlation distance d_c and this alternative d_{qc}, respectively, using a window [0, 5]; that is, starting at time of inoculation and ending 5 days after inoculation. The circular structure formed by the standard circadian rhythms collapses to a dense cluster near the top of the plot. This allows for better spatial expression of abnormal variation in temperature. Here we have coloured each time series by the severity of *Salmonella* infection in the liver measured in colony-forming units (CFUs), a measure of the severity of infection. The blue points correspond to mice with low liver CFU values while the red points correspond to mice with high liver CFU values. We see that the mice with low liver CFU values cluster together and the mice with high CFU values are generally located on the periphery of this cluster. This lends evidence to the idea that d_{qc} can account for differences in phase in temperature time series in a way that captures differing health outcomes of mice.

We conclude by remarking on the underlying mathematical structure of d_c and its relation to d_{qc}. How can we visualize the new space induced from d_{qc}? Since correlation distance is invariant with respect to changes in the magnitude of x, $y \in \mathbf{X}$, we can visualize U with correlation distance as points on the high-dimensional sphere S^m, where m is the number of time steps in each time-series window. d_c is then related to the angle between x and y. As shown in figure 4*b*, we visualize a schematic of S^m, with the orbit of all phases of a healthy time series forming a meridian on this sphere. In this illustration, the phases of a healthy circadian rhythm sit on a one-dimensional meridian of S^m. Modifying d_c to d_{qc} will have the effect of approximately collapsing this meridian to a single point. What we have now is two spheres S^m connected at a single point which corresponds to all phases of the time series sitting on the green line. This space is usually denoted by $S^m \vee S^m$. We remark that, while this understanding does not enter insofar as the data analysis, enriching the mathematical theory behind dissimilarity measures is highly valuable to give solid footing from which practitioners have an understanding of why their algorithms work. We are interested in building upon this theory further in future work.

4. Anomaly detection

In this section, we discuss approaches to anomaly detection for temperature time series. Our primary focus is on Radial Basis Function (RBF) and Multivariate State Estimation

Technique (MSET) approaches, which can build models for time series in real time as data are observed. We initially focus on two mice in the dataset: one infected with pathogen while the other is a control, or a sham inoculation consisting of a saline solution with no pathogen. The infected mouse develops a severe interruption in its circadian temperature pattern as the disease progresses while the control mouse is apparently unaffected. Representation of these two extremes allows us to analyse a range of behaviours of the algorithms. After this, we narrow focus on the online MSET algorithm and study its performance on the entire dataset in aggregate.

The RBF and MSET approaches are both well-known approaches for representing nonlinear relationships in data where little to no prior information can be assumed. Both approaches hope to represent a larger dataset in terms of a relatively small number of 'exemplars', which both aid in efficiently comparing new data with what has been seen previously and allow for easier interpretation of the resulting model, which consists of the collection of exemplars and an algorithm for representing new data with them.

Online algorithms progressively build a model as data are streamed in, as opposed to being allowed to build a single model witnessing the entire dataset at once. The algorithms considered here also have the property that they only employ novel exemplars to update the model, resulting in highly efficient real-time processing.

4.1. Anomaly detection via sparse Radial Basis Functions

The RBF approach to modelling data extends back to the 1971 paper by Hardy [12] towards an application representing topographic data using multi-quadrics, but has since been applied to a wide variety of fields; in particular, initially studied for their general utility in approximating high-dimensional data in [13,14].

RBFs may be viewed as a method of approximating a function f such that the input–output pairs $(\mathbf{x_i}, y_i)$ satisfy

$$y_i = f(\mathbf{x_i}).$$

The RBF approach seeks to well represent the data in a form

$$y_i = w_0 + \sum_{k=1}^{N_c} w_k \phi(\|\mathbf{x}_i - \mathbf{c}_k\|), \tag{4.1}$$

for some choices of basis function ϕ, norm and the number of basis functions N_c. Typically, one then optimizes the collection of weights $\{w_k,\ k = 0, \ldots,\ N_c\}$ and sometimes also the RBF centres $\mathbf{c}_k$ to minimize an error function.

There exist many algorithmic implementations of the basic mathematical representation (4.1). We focus our attention on those aimed at applications for online modelling of streaming data. For a fixed ϕ, much of the focus is on techniques for adding to or pruning the number of basis functions as new data arrive.

Following the work in [5], we determine the parameters in the RBF approximations by solving the optimization problem

$$\left.\begin{array}{ll} \underset{w,\ \epsilon}{\text{minimize}} & \|w\|_1 + C\epsilon, \\ \text{subject to} & \|\Phi w - b\|_\infty \le \epsilon, \end{array}\right\} \tag{4.2}$$

where the fitting parameter ϵ and the weight vector w are the decision variables in the optimization problem.

In [5], this optimization problem is solved as a sequential linear programming problem. The optimal solution evolves as the data are being observed one point at a time.

4.2. Anomaly detection via Multivariate State Estimation Technique

MSET is an algorithm originally developed at Argonne National Laboratory by Singer *et al.* [15] for analysis of time series related to industrial processes. The core step of MSET is to compare n new d-dimensional observation(s) $\mathbf{Y} \in \mathbb{R}^{d\times n}$ with an approximation $\hat{\mathbf{Y}}$ via a nonlinear mapping against a collection of exemplars, sometimes termed the 'memory' of m exemplars $\mathbf{X} \in \mathbb{R}^{d\times m}$,

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X} \otimes \mathbf{X})^{-1}(\mathbf{X} \otimes \mathbf{Y}). \tag{4.3}$$

If the approximation $\hat{\mathbf{Y}}$ is not close to $\mathbf{Y}$ in some sense, then $\mathbf{Y}$ is declared an anomaly and appropriate action may be taken. Conventions relating to the operator $\otimes$ vary depending on the reference (and can sometimes be contradictory within a single reference). Here we generally think of $\otimes$ operating on a pair of ordered collections of data, with the analogy to the linear least-squares projection shown in equation (4.4), so use of transposes in the general definition of $\hat{\mathbf{Y}}$ does not make sense.

The operator $\otimes$ maps matrices $\mathbf{X} \in \mathbb{R}^{d\times m}$ and $\mathbf{Y} \in \mathbb{R}^{d\times n}$ to a matrix $\mathbf{Z} \in \mathbb{R}^{m\times n}$. When the operator $\otimes$ is based on the standard inner product

$$\mathbf{X} \otimes \mathbf{Y} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}, \tag{4.4}$$

the approximation (4.3) corresponds to the linear least-squares projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbf{X}$, and $\mathbf{X} \otimes \mathbf{X}$ corresponds to the Gram matrix and contains information about the quality of the columns of $\mathbf{X}$ in representing their underlying subspace.

However, a linear representation of the data in exemplars will be ineffective when receiving the time series sequentially. Here instead we receive large collections of low-dimensional streaming data $(N > d)$, which invalidates linear approaches since with probability one the first d vectors will form a basis for $\mathbb{R}^d$ and any subsequent vector can be reconstructed exactly. Therefore, in the context of MSET, $\otimes$ is understood to be a nonlinear operator, with the primary condition being that $\mathbf{X} \otimes \mathbf{X}$ is non-singular. As discussed in [16], the original operators developed are either patented or proprietary. Various other choices of operators $\otimes$ have been proposed in the literature [16–18]. A straightforward class of operators are those which construct the output of $\mathbf{X} \otimes \mathbf{Y}$ by pair-wise comparisons over the columns of matrices $\mathbf{X} \in \mathbf{R}^{d\times m}$ and $\mathbf{Y} \in \mathbb{R}^{d\times n}$ in analogy to the dot product's relationship with matrix multiplication. Overloading notation for simplicity

$$(\mathbf{A} \otimes \mathbf{B})_{i,j} = \mathbf{A}_{\cdot,i} \otimes \mathbf{B}_{\cdot,j}; \quad i = 1, \ldots, m, \quad j = 1, \ldots, n, \tag{4.5}$$

so that similarity need only be defined between vectors in $\mathbb{R}^d$. A common choice in data analysis is to use what is termed 'cosine similarity', which is the cosine of the angle formed between two vectors,

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} \equiv \cos(\theta); \tag{4.6}$$

with $s(\mathbf{x}, \mathbf{y}) \in [-1, 1]$. In §3, we apply the closely related notion of correlation distance as a way to quantify *dis*-similarity between chunks of the scalar time series.

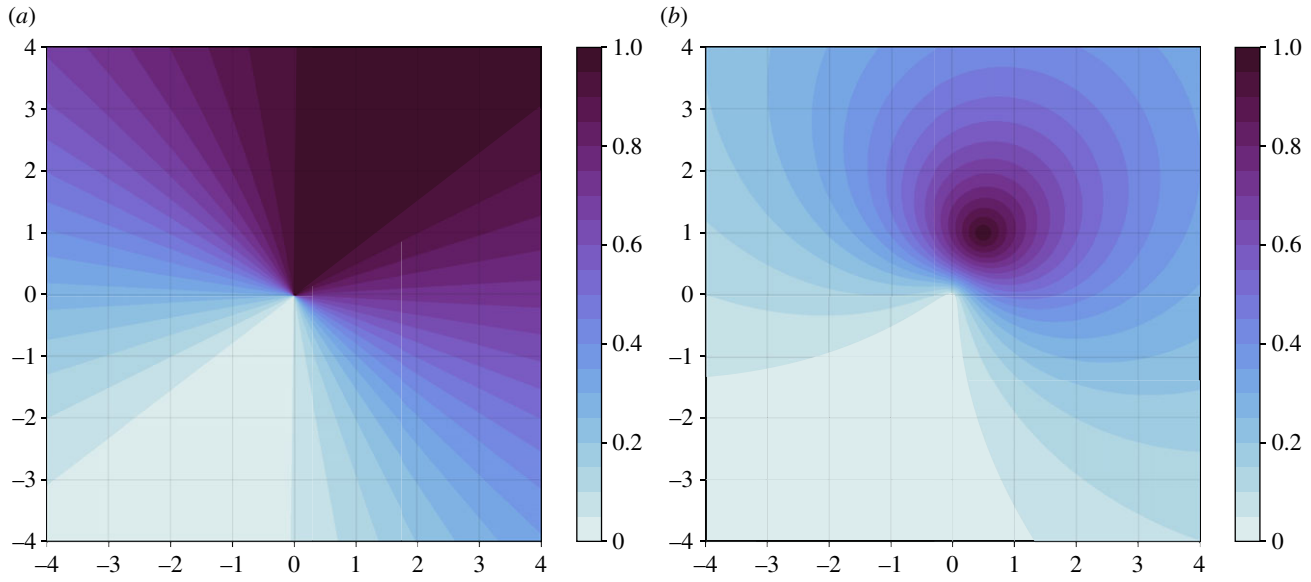

Figure 5. Visualization of [0, 1]-normalized cosine similarity (*a*) and the nonlinear similarity (4.7) values relative to a reference point (1/2, 1) (*b*). Cosine similarity is constant along rays anchored at the origin, as it is based on angles between subspaces, while (4.7) is locally radially symmetric around the reference point but has complex non-local behaviour. (Online version in colour.)

For the application of MSET, we follow Wang *et al.* [18] with the choice of similarity

$$\mathbf{x} \otimes \mathbf{y} = 1 - \frac{\|\mathbf{x} - \mathbf{y}\|_2}{\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2}. \tag{4.7}$$

In analogy with cosine similarity, $\mathbf{x} \otimes \mathbf{x} = 1$ and $\mathbf{x} \otimes \mathbf{0} = 0$. By contrast, the same is not true if $\mathbf{x} = c\mathbf{y}$, $c \neq 0$ or 1; a calculation shows $\mathbf{x} \otimes c\mathbf{x} = 0$ for $c \leq 0$ and decays asymptotically as $2/c$ for $c \to +\infty$. To better understand these similarities, we visualize both the transformed cosine similarity $1/2 + 1/2s(\mathbf{x}, \mathbf{y})$ and this nonlinear similarity in figure 5. We fix a reference point $\mathbf{y} = (1/2,\ 1)$ and study the value of $\mathbf{x} \otimes (1/2,\ 1)$ with $\mathbf{x} = (x_1,\ x_2)$ over a range of values $[-4, 4] \times [-4, 4]$. Cosine similarity is constant on rays emanating from the origin, attaining a value of 1 along the ray passing through (1/2, 1) itself and zero in the opposing direction. However, (4.7) is locally radially symmetric around the reference point and exhibits complex behaviour on a larger scale. This local radial symmetry in fact puts it in closer in relation to RBFs than subspace-based techniques.

To implement MSET for anomaly detection, a good choice of exemplars $\mathbf{X}$ is needed which well represents the nominal state. This may be possible for isolated systems with large amounts of nominal data to work with. However, using a single collection of exemplars for the memory $\mathbf{X}$ across all mouse time series is likely to be a bad idea to begin with, since this cannot account for the large variability in the dynamics of the healthy state between organisms. Stated another way, we expect that these models will succeed when each individual has its own set of exemplars.

Working with this hypothesis, in the scenario of a limited-time experiment, exemplars need to be built for each mouse in an online fashion. We propose the following online variation of MSET to do this.

(i) Choose index-valued delay τ, embedding dimension d and relative tolerance ε.
(ii) Observe the time series up to index $\tau(d-1)$. Estimate the mean value μ of the scalar time series from these data as $\hat{\mu}$ and store as a fixed entry-wise shift $\hat{\mu}\mathbf{1}$.
(iii) Initialize $\mathbf{X}$ with the first embedded vector $\mathbf{x}_{\tau(d-1)} - \hat{\mu}\mathbf{1}$.
(iv) For each subsequent embedded data point $k = \tau(d-1)+1, \ldots$, calculate the MSET approximation $\hat{\mathbf{x}}_k$ to $\mathbf{x}_k$ from (4.3) using similarity (4.7).
(v) Evaluate the relative Euclidean error

$$e_k = \frac{\|\hat{\mathbf{x}}_k - \mathbf{x}_k\|_2}{\|\mathbf{x}_k\|_2}. \tag{4.8}$$

If $e_k \geq \varepsilon$, then append $\mathbf{x}_k$ to the memory $\mathbf{X}$ and mark time point k as an anomaly (1). Otherwise, mark as nominal (0) and continue.

The choice to mean-centre the data in an online fashion (step 2) is used to increase the sensitivity of the algorithm by increasing the range of observed pairwise similarities when using (4.7), which to some degree loses sensitivity when the data are large in magnitude relative to variations. However, this may not be necessary if another operator $\otimes$ is used.

4.3. Comparison of online anomaly detectors

Here we compare the two anomaly detectors discussed in this section on the complete time series of two mice with distinct post-infection responses. One mouse, CBA-218, was inoculated with a saline solution and can be considered a control for the purposes of studying the association of temperature time series with infection. Aside from a momentary disturbance from the mock-inoculation process, their temperature patterns go unchanged for the duration of the experiment. The other mouse, CC004-021, was inoculated with *Salmonella* and its temperature is moderately affected initially in the day after infection, then severely so in the following days.

The results for online MSET and sparse RBFs are shown in figure 6. For both models, we observe an initial model-building phase in the first 2 days of the time series. After this phase, the frequency of new anomalies significantly drops off prior to infection. For the mock-infected CBA-218, the number of new anomalies after saline inoculation is very small, demonstrating the ability of each algorithm to avoid significant false positives. Shortly after the time of infection (on day 7) their circadian

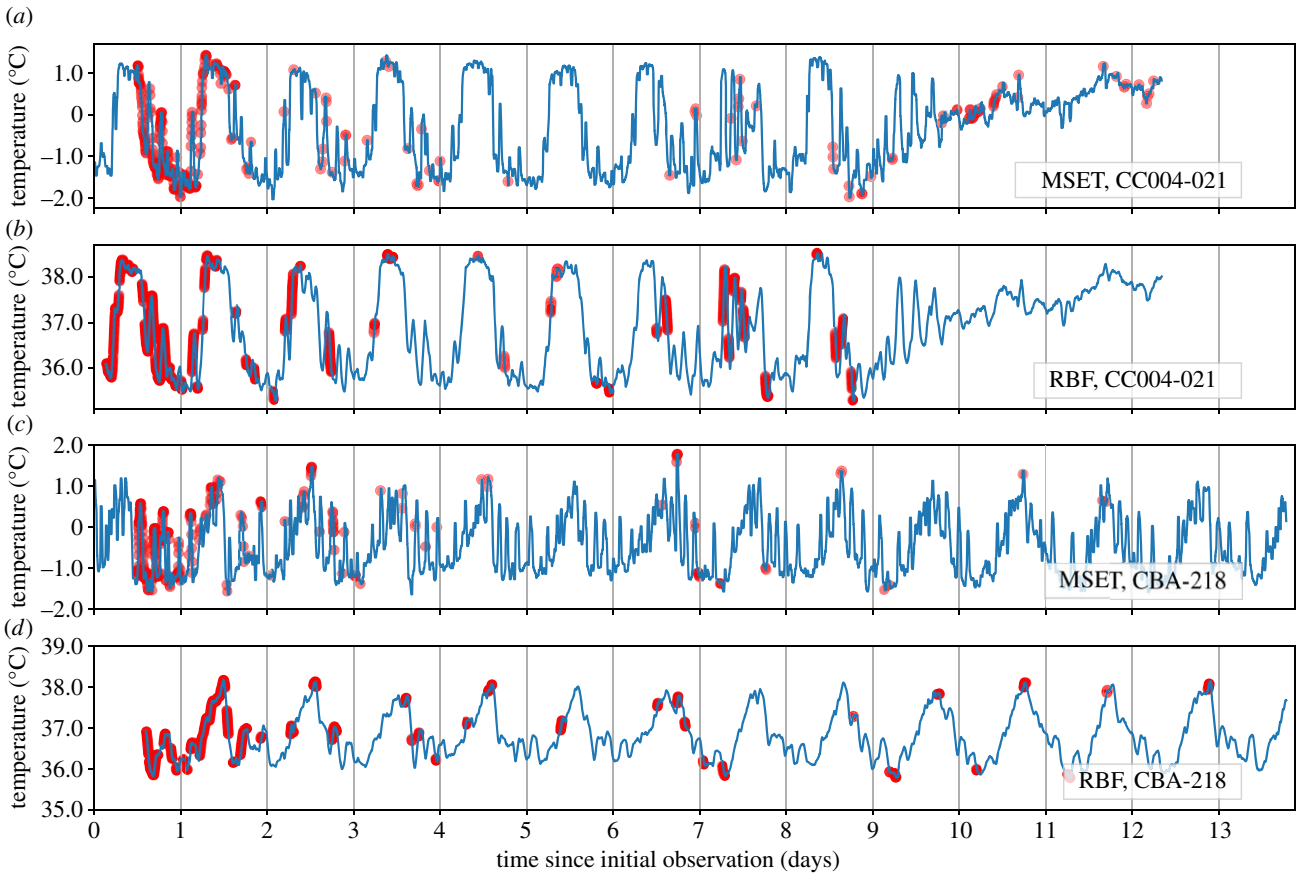

Figure 6. Comparison of online MSET and sparse RBF techniques on the same mice. For CC004-021 (*a*,*b*), RBF strongly identifies interruptions in pattern soon after infection on day 7, while MSET has a lighter scattering of identified anomalies persisting through the infected period. For CBA-218 (*c*,*d*), both algorithms identify anomalies around the time of infection, and otherwise have minimal false alarms at extreme values. Note that the RBF simulations have used slightly different smoothing and time-delayed embedding parameters, but the underlying time series are the same. (Online version in colour.)

pattern is disturbed, and the time density of new anomalies is larger than that for CBA-218 as the mouse's condition worsens. The parameters used for MSET are $\tau = 360$ (6 h), $d = 3$ and $\varepsilon = 0.05$. In the case of the sparse RBF, we define a new data point as an anomaly if more than zero iterations are needed to update the model.

For mouse CC004-021 (figure 6*a*,*b*), the sparse RBF model flags anomalies in greater density in the first noticeable interruption (by eyeball-norm) in the circadian rhythm following infection, while not having further anomalies flagged in the last 3–4 days of data; MSET by contrast flags anomalies in a more spread-out fashion as the mouse's condition worsens. For the mock-infected mouse CBA-218, both algorithms identify some anomalies around the period of infection; after which there are only occasional anomalies at the extrema of the temperature time series.

4.4. From proof of principle to practical anomaly detection

In this section, we pursue the goal of anomaly detection to its final stage for the case of the online MSET algorithm and evaluate its performance on a large portion of the data. A criterion to separate the 'learning' and anomaly detection phases of the online algorithm are defined, then a training and validation scheme are used to identify and evaluate the choice of parameter needed for this criterion in practice. Further refinement of the algorithm requires careful consideration of the time series of model errors, which we will pursue further in future work.

The algorithm described for online MSET produces a collection of anomalies, but these need to be further processed to make actionable decisions. We illustrate these considerations in figure 7. Decisions can sometimes be made by eye to separate anomalies which are part of learning the nominal pattern and those which correspond to irregular patterns, but it is important to formally quantify success or failure of algorithms using unbiased measures. To formalize our intuitive notion, we declare the end of the 'learning' phase the first time an anomaly is detected after some period of time T since the previous anomaly. Hence, our goal is to study various choices of T and consider the trade-off between excessively small and large values. For this section, we apply a formal training/validation scheme to avoid overfitting the dataset. The data consisted of 185 time series of the original 306, screened for having sufficient post-inoculation data and little or no missing data to impute. The training and validation sets for the choice of delay T have 92 and 93 data points, respectively, chosen randomly using the random number generator in `numpy` with a fixed (known) seed for reproducibility.

4.4.1. Online MSET with identification of learning phase—training.

For a threshold T small, we expect the 'training' phase to be halted quite early, and consequently there will be significant numbers of false-positive anomalies prior to inoculation.

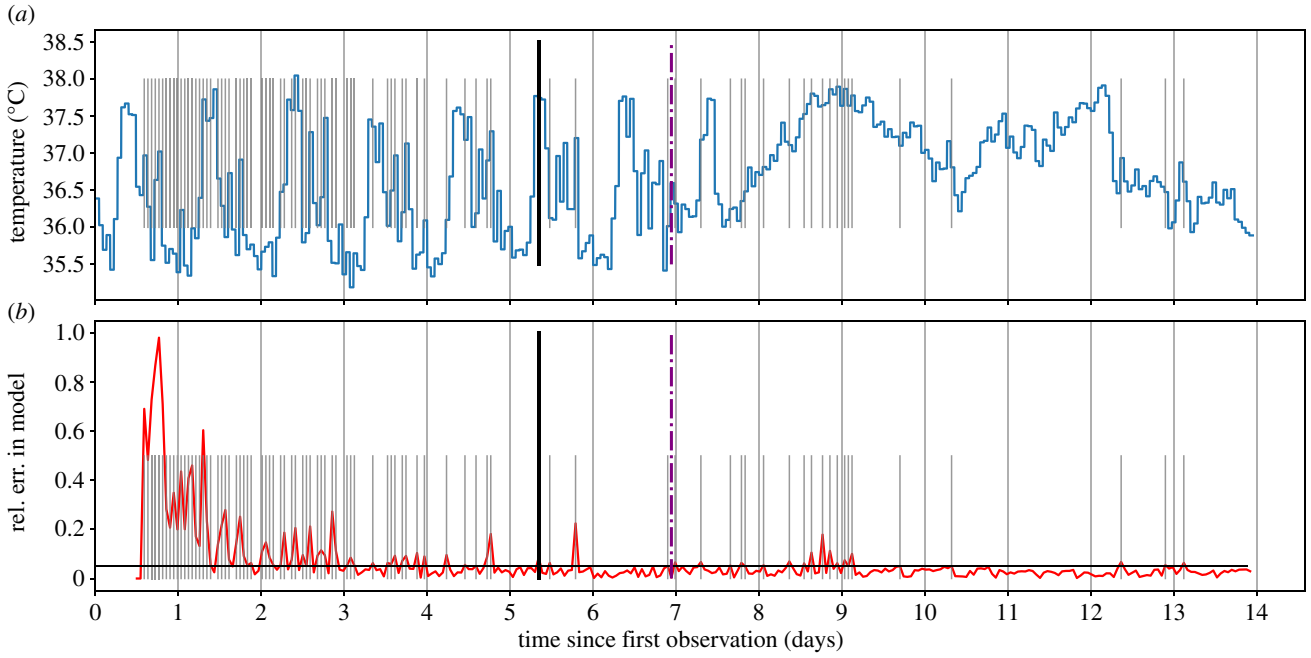

Figure 7. Example of separation of MSET anomalies into those considered part of the learning phase (*a*) and those associated with anomalies to the learned pattern (*b*). Here, the vertical black lines mark the first anomaly after the training phase. Small grey lines mark the locations of all anomalies: those to the left of the black line are considered part of training; those to the right are predicted anomalies. Dashed purple lines denote the known time of inoculation; hence anomalies prior to this are considered false positives and those to the right are true positives. (Online version in colour.)

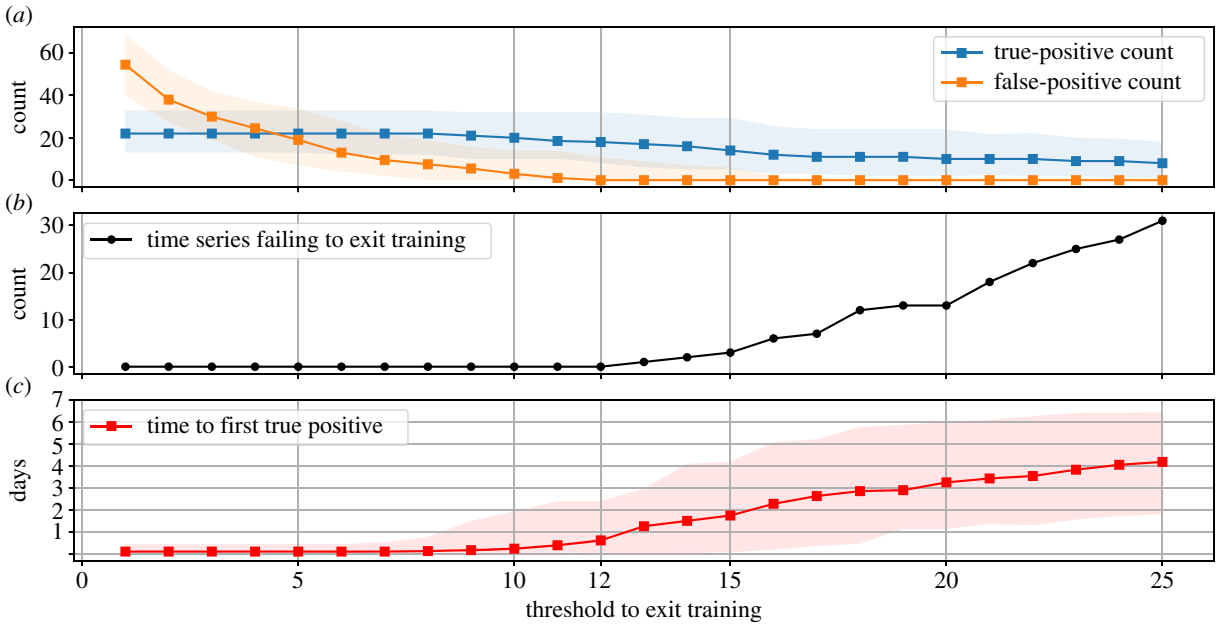

Figure 8. Study of online MSET for various choices of delay. Dotted lines in (*a*,*c*) represent median values across all time series for that value of *T*. Shaded regions represent the [10%, 90%] quantile. (*b*) The growth in time series that fail to exit the training phase owing to the excessively large gap *T* required. (Online version in colour.)

Large values of T result in a decrease in both false and true positives and an increase in the number of days after inoculation until the first anomaly. For even larger values of T, there are not sufficiently large gaps in the data to even declare an end to the learning phase.

In figure 8, we visualize how this manifests in practice on the training data. Here the data were preprocessed via Haar wavelets, keeping a small number of detail coefficients, and the threshold for MSET chosen as $\varepsilon = 0.05$. Because the wavelet smoothing results in piecewise constant time series, our parameter T is in multiples of 64 min—the displayed threshold is displayed as the number of multiples of 64. For a reasonable comparison across the data, we restrict our attention from pre-inoculation to at most 7 days after the point of inoculation. We do see excessively small thresholds to exit training result in high false-positive anomalies, while true-positive counts are relatively insensitive to this threshold in aggregate. The number of time series which fail to exit the learning stage grows steadily after $T = 15 \times 64$, as does the time the algorithm identifies its first true-positive anomaly.

Considering these results, we determined a value of $T = 12 \times 64$ (768 min, or nearly 13 h) as a threshold between

Table 1. Results for the online MSET algorithm with a 768 min separation between learning and anomaly detection phases on 92 validation time series. Note that the algorithm failed to exit the training phase on a single time series, not included in the statistics here. Data are organized by the number of false-positive anomalies identified in the anomaly detection phase. Of primary interest here is the first time anomalies are detected following infection; hence all numbers are positive and are reported in hours following infection.

false positives	0	1–5	6–10	11–15	16–20	aggregate
count	*55*	21	12	2	2	92
median first anomaly (hours)	*39.5*	1.1	1.6	4.3	4.8	15.0
10% quantile first anomaly (hours)	*7.5*	0.0	0.0	3.4	2.7	6.0
90% quantile first anomaly (hours)	*84.3*	5.3	4.3	5.1	6.9	31.0

successive anomalies to formally end the training period achieves a careful balance between minimizing false positives and not being overly stringent for the purposes of training and early detection. At this value, only a single time series in the training data fails to finish the learning phase, and the median time of the first true-positive anomaly is within 1 day after infection (the 90th percentile is within 2.5 days). Note that individual mice may still have false positives with this value.

4.4.2. Online MSET with identification of learning phase—validation.

To validate the choice of parameter $T = 768$ separating learning and anomaly detection phases, we applied the algorithm to 93 time series not involved in the decision of parameters. Online MSET with this gap time and $\varepsilon = 0.05$ was applied to all time series in this set. The algorithm failed to leave the learning phase for a single time series of these 93 because it did not achieve the threshold gap of 768 min between successive updates. For the remaining 92 time series, we summarize the results in table 1.

The primary statistic we are interested in here is the time required for the anomaly to first detect off-pattern behaviour in the data. While the patterning of anomalies is different from mouse to mouse, we summarize by reporting statistics across each group. The median anomaly time following infection for the group with no false positives was 39.5 h, with a [10%, 90%] quantile range of between 7.5 and 84.3 h after infection. The remaining groups with a range of false positives have noticeably lower times to detecting a first anomaly, but this observation should be interpreted carefully depending on the application. In our broader interests of inferring underlying health state, which is beyond the scope of this paper, it may be acceptable to ignore false positives or include them directly as an additional feature by which one could separate different modes of response to infection. For those interested in detecting interruptions and intervening for positive health outcomes, the cost of a false positive needs to be judged more carefully, i.e. this may result in healthy subjects being treated unnecessarily.

While the main goal of this study is to successfully predict anomalies in relation to *Salmonella* infection, we observe that false positives prior to inoculation time do often correspond to interruptions in prior circadian pattern. Robustly addressing this observation may require a much longer nominal observation period (which is not available with these data) or cross-time-series inference via, for example, dictionary learning methods and proper normalization across mice, which we leave for future work.

5. Conclusion

Observations of mouse temperature telemetry data suggest that the health state of an individual may be determined using techniques for modelling time series on high-dimensional domains. To this end, we have explored machine and geometry learning algorithms to characterize patterns in mouse temperature time-series data, in the direction of *post hoc* visualization, and 'online' anomaly detection.

In the direction of visualization, we have illustrated that Laplacian Eigenmaps may be used to extract prototypical signatures at the onset of infection and provide a tool to visualize the collapse of healthy circadian rhythms across the entire dataset. We also investigated the impact of using different dissimilarity measures between time series: pointwise Euclidean, correlation distance and a custom correlation distance. When one only has a sense of dissimilarity between objects, we highlight that MDS is a useful tool to embed the data to best visualize these dissimilarities in aggregate. Similar structures to those from Laplacian Eigenmaps—i.e. loops—are observed when applying such techniques to healthy circadian patterns. There is little structure when focusing attention after infection, aside from the collapse of the loop, except when one builds new dissimilarities such as to account for phase differences in the data. Such visualizations reveal some association between the resulting radial clustering of the MDS embeddings and severity of infection.

In the direction of anomaly detection, we have demonstrated the utility of both sparse RBF and MSET to identify anomalies in these time series and provide insight into quantifying the time to a time series becoming 'off-pattern' as measured by the first post-infection anomaly. Beyond a proof of concept, we report detailed results of the performance of the online variant of MSET with a simple approach to separating learning and anomaly detection phases. We can most confidently report on those data for which no false-positive anomalies are recorded. Of this subset of 55 of 93 mice, their first post-infection anomalies were at a median of 39.5 h after infection, with 80% of the data having their first anomalies between 7.5 and 84.3 h after infection. This is a reflection of the ability of the methods to identify anomalies in time series with very little prior knowledge and the changes in the data as a result of host responses against *Salmonella* infection.

Applications of the anomaly detectors could include the identification of specific time points, e.g. disease onset, or

progression to full blown disease, for further clinical study of an animal. In addition, anomaly detectors provide an automated tool for health monitoring and interventions, e.g. euthanasia in the case of severe anomalies.

Data accessibility. This article has no additional data.

Authors' contributions. M.A. generated figures 2, 5, 7, 8 and wrote large portions of the text. H.K. generated figures 4 and text in §3b. X.M. generated figures 6 and 4*a* and text in §3a. P.R. generated figures 1 and 3. H.A.-P., K.S., J.G. and D.T. generated the mouse time series data and provided feedback. M.K. proposed the analysis, developed tools and wrote portions of the paper.

Competing interests. We declare that we have no competing interests.

Funding. This paper is based on research partially supported by the National Science Foundation grant no. DMS-1513633 and DARPA contracts N66001-17-2-4020 and D17AP00004.

References

1. Drewry AM, Fuller BM, Bailey TC, Hotchkiss RS. 2013 Body temperature patterns as a predictor of hospital-acquired sepsis in afebrile adult intensive care unit patients: a case-control study. *Crit. Care* **17**, R200. (doi:10.1186/cc12894)
2. Varela M, Churruca J, Gonzalez A, Martin A, Ode J, Galdos P. 2006 Temperature curve complexity predicts survival in critically ill patients. *Am. J. Respir. Crit. Care Med.* **174**, 290–298. (doi:10.1164/rccm.200601-058OC)
3. Cuesta D, Varela M, Miró P, Galdós P, Abásolo D, Hornero R, Aboy M. 2007 Predicting survival in critical patients by use of body temperature regularity measurement based on approximate entropy. *Med. Biol. Eng. Comput.* **45**, 671–678. (doi:10.1007/s11517-007-0200-3)
4. Cui Z, Chen W, Chen Y. 2016 Multi-scale convolutional neural networks for time series classification. (http://arxiv.org/abs/quant-ph/1603.06995)
5. Ma X, Aminian M, Kirby M. 2018 Error-adaptive modeling of streaming time-series data using radial basis functions. *J. Comput. Appl. Math.* **362**, 295–308. (doi:10.1016/j.cam.2018.10.056)
6. Packard NH, Crutchfield JP, Farmer JD, Shaw RS. 1980 Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716. (doi:10.1103/PhysRevLett.45.712)
7. Takens F. 1981 *Detecting strange attractors in turbulence*. Berlin, Germany: Springer.
8. Churchill GA *et al.* 2004 The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.* **36**, 1133. (doi:10.1038/ng1104-1133)
9. Threadgill DW, Churchill GA. 2012 Ten years of the Collaborative Cross. *Genetics* **190**, 291–294. (doi:10.1534/genetics.111.138032)
10. Belkin M, Niyogi P. 2003 Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396. (doi:10.1162/089976603321780317)
11. Bibby JM, Kent JT, Mardia KV. 1979 *Multivariate analysis*. London, UK: Academic Press.
12. Hardy RL. 1971 Multiquadric equations of topography and other irregular surfaces. *J. Geophys. Res. (1896–1977)* **76**, 1905–1915. (doi:10.1029/JB076i008p01905)
13. Broomhead DS, Lowe D. 1988 Radial basis functions, multi-variable functional interpolation and adaptive networks, Technical Report, Royal Signals and Radar Establishment, Malvern, UK.
14. Moody J, Darken CJ. 1989 Fast learning in networks of locally-tuned processing units. *Neural Comput.* **1**, 281–294. (doi:10.1162/neco.1989.1.2.281)
15. Singer RM, Gross KC, Wegerich S, Henke D. 1996 MSET. Multivariate state estimation technique. See https://www.osti.gov/biblio/436752-mset-multivariate-state-estimation-technique.
16. Hines J, Garvey J, Seibert R, Usynin A. 2008 Technical review of on-line monitoring techniques for performance assessment, volume 2: theoretical issues. Technical review NUREG/CR-6895. See https://www.nrc.gov/docs/ML0814/ML081430058.pdf.
17. Thompson J, Dreisigmeyer DW, Jones T, Kirby M, Ladd J. 2010 Accurate fault prediction of BlueGene/P RAS logs via geometric reduction, pp. 8–14. See https://dl.acm.org/citation.cfm?id=1909521.
18. Wang K, Thompson J, Peterson C, Kirby M. 2015 Identity maps and their extensions on parameter spaces: applications to anomaly detection in video. In *Proc. 2015 Science and Information Conf. (SAI), London, UK, 28–30 July 2015*, pp. 345–351. New York, NY: IEEE.