Grid-based partitioning for comparing attractors

T. L. Carroll* and J. M. Byers
US Naval Research Lab, Washington, DC 20375, USA
(Received 16 December 2015; published 13 April 2016)

Stationary dynamical systems have invariant measures (or densities) that are characteristic of the particular dynamical system. We develop a method to characterize this density by partitioning the attractor into the smallest regions in phase space that contain information about the structure of the attractor. To accomplish this, we develop a statistic that tells us if we get more information about our data by dividing a set of data points into partitions rather than just lumping all the points together. We use this method to show that not only can we detect small changes in an attractor from a circuit experiment, but we can also distinguish between a large set of numerically generated chaotic attractors designed by Sprott. These comparisons are not limited to chaotic attractors—they should work for signals from any finite-dimensional dynamical system.

DOI: 10.1103/PhysRevE.93.042206

I. INTRODUCTION

The description and analysis of chaotic attractors is an evolving field. Much of this work is concerned with modeling and prediction of dynamical systems [1–5]. Finding a model of the actual vector field is probably the most efficient way to describe a dynamical system, but finding a model without knowing the proper functional form is computationally difficult and sensitive to noise. If the goal is only to characterize, but not to predict, then studies of the geometry of the attractor may be useful. Such techniques have been used for many years [6–11]. More recently, graph theory has been used to characterize attractors as networks [12], although with the exception of networks based on recurrences, it is not clear what the physical significance of the network is.

The shape of a chaotic attractor is well defined and not subject to prediction errors, but there are few tools for describing this shape [13,14]. Diks *et al.* [15] compared delay vector distributions by convolving the individual embedded vectors with Gaussians to create a smooth probability distribution. Attractors were compared by taking the difference between densities.

Our work is similar in spirit to Diks *et al.*, but we use partitioning methods that speed up the computation. Diks *et al.* were limited to comparing attractors of a few hundred points, while we have compared attractors of 100 000 points. In this work, we describe the shape of a chaotic attractor by partitioning the attractor into regions of different densities. It is well known that the probability measure of a dynamical system reflects its long-term behavior in phase space [10,16]. For some dynamical systems, such as complex electronic circuits, or driven structures, it is not mathematically tractable to generate a model for the system [17,18], so characterizing the attractor without including details of the dynamics is all that is possible. Partitioning the attractor into regions of different density reduces the size of the data set, which could be a useful first step for graph theory calculations, which require one to find the distance between each point in the data set and all other points [12,19]. In the work described here, the attractor is partitioned into regions that each contain the same amount of information. By information, we mean that we can distinguish that the distribution of points in some local region is not best described by a uniform distribution. We develop a statistic to make this comparison. We then use common statistical methods to compare densities in corresponding regions of different attractors. In previous work, it was established that density can be used to compare attractors [20]. The current method extends the range of application of density-based methods to attractors that are not similar to each other, and it does away with an arbitrary parameter by using the data itself to determine how to create a histogram for the attractor.

There are other methods for characterizing attractors, such as fractal dimension, Lyapunov exponents, linking numbers, etc. [21]. These methods are commonly used because in theory they are invariant under orientation preserving diffeomorphisms, so that a change in the embedded variable or the embedding method should not change the measurement. In practice, there are well-known problems when applying these standard methods to real data; see, for example, Ref. [22]. Real data is finite and limited in resolution by digitization. In order to understand the topology of data, after embedding the data we must establish some sort of metric. This metric is necessary to understand the topology of the data—without it, we can not distinguish where the attractor is and where it may have holes or cavities. Changing embedding parameters will change this metric—this is true for any measurement on data, so no data analysis is completely invariant under changes in the embedding parameters. There is still the outstanding question of how to choose embedding dimension and delay; we don't solve that problem here, but once our data is embedded, further analysis depends only at what level of information we want to partition our data. There are no other arbitrary thresholds to be determined.

In order to show the flexibility of this data-partitioning method, we use two different partitioning examples below. First, we show how data partitioning may be used to distinguish between the 19 different Sprott attractors [23]. Second, we track the small deviation from linearity in an experiment using an operational amplifier.

II. DENSITY IN PHASE SPACE

We begin by embedding a time series s into a d-dimensional phase space using the method of delays [21].

*Thomas.Carroll@nrl.navy.mil

For each point in s, a vector $\mathbf{s}(i)$ is defined as $\mathbf{s}(i) = \{s(i),s(i+\tau),\ldots,s[i+(d-1)\tau]\}$. The embedding dimension d and the delay τ may be found by any one of a number of standard methods [21].

A histogram of the embedded attractor in the phase space is then created. The phase space is divided into partitions, and the partition locations and sizes are recorded, as well as the number of points in each partition. The partitions may have different sizes.

A. Determining partitions

We choose to subdivide, or partition, the attractor based on whether subdividing a part of the attractor yields more information than not subdividing. We measure information by counting the number of attractor points m_k that fall into each of our K partitions, or subdivisions. We compare the counts $m_k, k = 1,\ldots,K$ to the number of counts we would expect if the points on the attractor were distributed uniformly over the region we are subdividing. If we can't distinguish between the observed values of m_k and what we would expect from a uniform distribution, then we don't subdivide. If, on the other hand, the set of counts m_k differs from what we would expect from a uniform distribution, then we gain information about the attractor by subdividing, so we proceed with the subdivision.

Partitioning the attractor might seem to be losing spatial information, since we are grouping points into bins. Because of the information criteria for choosing partition size however, we subdivide the attractor until the set of points in each partition can not be distinguished from a uniform distribution over the same space, so we are finding the minimum size partitions that contain information about the attractor.

In most cases, the final partitions found using this information criteria will not be all the same size—some regions of the attractor contain structure at smaller length scales than other parts. The size of the final partitions will also depend on the number of data points, as more data will allow us to better see small-scale variations.

We will refer to the type of partitioning just described as top-down partitioning. Top-down partitioning is fast, but it divides the phase space by a factor of two each time, so the final partitions may miss some structure in the data. There will be situations described later in this paper where it may be better to start with some small region of the attractor containing only a few points and expand. In this bottom-up approach, the phase space is first partitioned using the top-down approach. The smallest resulting partition is taken as the smallest meaningful length scale on the attractor. The phase space is then divided into N_b bins along each axis, where N_b = (axis length)/(minimum length scale). Using the information criterion defined below [Eq. (6)], individual bins may be combined into larger partitions. Bottom-up partitioning is slower than top-down partitioning, but because the initial partitions are located where there is data, and the partitions are expanded in small increments, bottom-up partitioning can better reflect the structure of the data.

B. Information criteria

Based on the number of points m_k found in each partition, we want to estimate the probability π_k of finding a point in each of the K partitions. We will then use the Kullback-Leibler divergence [24] to find the difference between the probability distribution given by the set of π_k's and a distribution that is uniform over all K partitions.

Initially we don't know anything about the probabilities π_k. If the vector of counts is $\mathbf{m} = [m_1,m_2,\ldots]$, while the vector of probabilities is $\pi = [\pi_1,\pi_2,\ldots]$, then the probability of finding m things in k partitions when the probability for each partition is π_k is given by a multinomial distribution

$$\mathrm{Mult}(\mathbf{m}|\pi) = M!\prod_{k=1}^{K}\frac{(\pi_k)^{m_k}}{m_k!}\ \text{s.t.}\ \sum_{k=1}^{K}m_k = M \text{ and } \sum_{k=1}^{K}\pi_k = 1. \tag{1}$$

The conjugate prior for the multinomial distribution is the Dirichlet distribution [25]

$$p(\pi|\alpha) = \mathrm{Dir}(\pi|\alpha) = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\prod_k \Gamma(\alpha_k)}\prod_k \pi_k^{\alpha_k-1}, \tag{2}$$

where Γ is the γ function and the α_k's are adjustable parameters. The Dirichlet distribution is sometimes referred to as a probability of a probability. For a given set of parameters α, the Dirichlet distribution gives us the prior probability of a particular set of π's. We use the maximum entropy prior distribution, for which all α_k's $= 1/2$.

The probability of seeing a particular set of π_k's based on the observed m_k's may then be found from Bayes' theorem

$$\mathrm{Dir}(\pi|\alpha) \propto \mathrm{Mult}(\mathbf{m}|\pi)\mathrm{Dir}(\pi|\alpha'). \tag{3}$$

The probability distribution in Eq. (3) can be compared to a constant distribution over the same part of phase space. The density ρ_0 is equal to $\sum_{k=1}^{K} m_k / \sum_{k=1}^{K} V_k$, where V_k is the volume of an individual partition. The Kullback-Leibler divergence [24] is used to compare probability distributions. For two probability distributions p and q, the Kullback-Leibler divergence is

$$D_{KL}(p||q) \equiv \sum_{k=1}^{K} p(k)\ln\left(\frac{p(k)}{q(k)}\right). \tag{4}$$

The Kullback-Leibler divergence is described as the number of bits needed to encode the probability distribution p using samples from q. As an example, if p was the alphabet and q was a binary code, then $D_{KL}(p||q)$ would be the number of bits needed to encode the alphabet.

The Appendix shows how Eqs. (2)–(4) may be combined to give an analytic result that gives the Kullback-Leibler divergence between the probabilities estimated from Eq. (3) and a distribution with a constant density given by ρ_0. The divergence is

$$\begin{aligned} &D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)] \\ &= \frac{1}{\ln 2}\sum_{k=1}^{K}\left[(m_k-\rho_0 V_k)\psi\left(m_k+\frac{1}{2}\right) - \ln\Gamma\left(m_k+\frac{1}{2}\right)\right. \\ &\quad \left. + \ln\Gamma\left(\rho_0 V_k+\frac{1}{2}\right)\right]. \end{aligned} \tag{5}$$

Equation (5) represents the amount of information we gain by dividing the data into K partitions, rather than just considering the data to be uniformly distributed over the same volume.

We also need to add a partitioning penalty to the information difference function. Dividing the data into K partitions creates information; we could partition the entire phase space so finely that no partition contained more than one point, in which case $D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)]$ would be large. Specifying K partitions requires $\log_2 K$ bits for each partition; for example, four partitions could be specified by $k = 0,1,2,3$, or in binary, 00, 01, 10, 11, so four partitions require two bits for each partition. A penalty function $L(\Theta) = K\log_2 K$ assigns a cost to partitioning the data. The final information criterion is then

$$R(\mathbf{X},\Theta) = \frac{D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)] - L(\Theta)}{K}. \quad (6)$$

The units of $R(\mathbf{X},\Theta)$ are bits/partition. We may set a reasonable threshold: if $R(\mathbf{X},\Theta) > 1$ bit, then partitioning the data into K partitions gives more information than treating the data as a constant distribution over the same volume.

III. IDENTIFYING SPROTT ATTRACTORS

Sprott [23] found a family of 19 different chaotic attractors defined by three-dimensional ODE's with one or two quadratic nonlinearities. This group of attractors is a useful test set for our attractor comparison methods.

Each set of ODE's for the Sprott attractors was integrated using a fourth-order Runge-Kutta integrator with a time step of 0.01. The integrator output was decimated by keeping every 50th point to produce a time series. Time series of 20000 points were embedded in a three-dimensional space with an embedding delay of two points.

As an example, the Sprott C attractor was described by the differential equations

$$\begin{aligned} \frac{dx}{dt} &= yz \\ \frac{dy}{dt} &= x - y \\ \frac{dz}{dt} &= 1 - x^2. \end{aligned} \quad (7)$$

Figure 1 is a plot of the embedded attractor for the Sprott C system.

The Sprott attractors were partitioned by initially dividing the phase space into two bins/axis, for a total of eight bins. The information criterion $R(\mathbf{X},\Theta)$ [Eq. (6)] was found by counting the number of points in each of the eight bins (the m_k values). For the initial division, $R(\mathbf{X},\Theta)$ was much greater than one bit/partition, so each of the eight bins was further partitioned into eight bins. The initial set of eight bins can be denoted as (1),(2),(3), . . . ,(8). At the next level, bin (1) is divided into bins (1,1),(1,2),(1,3), . . . ,(1,8). In order to determine whether a further subdivision is required, $R(\mathbf{X},\Theta)$ is computed using the number of points in the bins (1,1),(1,2),(1,3), . . . (1,8). If $R(\mathbf{X},\Theta) > 1$ bit, each of the bins at this level are again subdivided, and the process continues until $R(\mathbf{X},\Theta) < 1$ bit. In the same manner, all the other top level bins are also subdivided. The final bins may have different sizes.

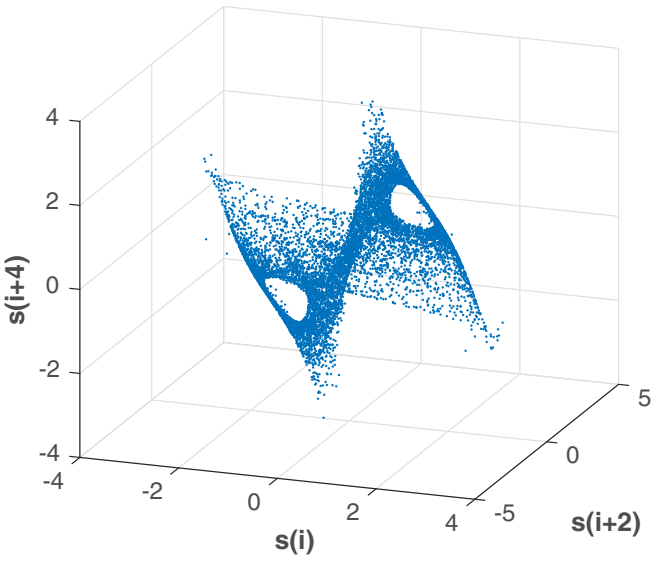

FIG. 1. Embedded time series signal for the Sprott C attractor with an embedding delay of 2.

The result of top-down partitioning for the Sprott C attractor is shown in Fig. 2. The partitioning yielded 2427 total bins of different sizes. The density of the bin with the highest density is $\rho_{\max}$. Figure 2(a) shows all the bins for the partitioned attractor whose density is $> 0.1\rho_{\max}$, Fig. 2(b) shows bins with densities from $10^{-1}\rho_{\max}$ to $10^{-2}\rho_{\max}$, Fig. 2(c) shows bins with densities from $10^{-2}\rho_{\max}$ to $10^{-4}\rho_{\max}$, and Fig. 2(d) shows bins with densities $< 10^{-4}\rho_{\max}$.

A. Comparing densities

Once attractors have been partitioned, they can be compared by comparing densities at the same locations in their respective phase spaces. The Kullback-Leibler divergence [Eq. (5)] can be used to for this comparison, but there can be situations where one attractor has a finite density at a particular location while the other attractor has zero density. As a result, the Kullback-Leibler divergence can not be used to compare densities for such a location. To avoid this problem, we use the Jensen-Shannon divergence [26] to compare attractors. The Jensen-Shannon divergence is a symmetrized version of the Kullback-Leibler divergence:

$$D_{JS}(p\|q) = \sum_{k=1}^{K} \frac{1}{2}\left[\log\left(\frac{p(k)}{0.5[p(k)+q(k)]}\right)p(k) + \log\left(\frac{q(k)}{0.5[p(k)+q(k)]}\right)q(k)\right]. \quad (8)$$

We use the Jensen-Shannon divergence only to compare attractors, so using the Jensen-Shannon divergence here does not affect the derivation of the information criterion in Eq. (5). Why not use the Jensen-Shannon divergence to derive the information criterion? As described in Eq. (5), the Kullback-Leibler divergence is described as the number of bits needed to encode the probability distribution p using samples from q. The Jensen-Shannon divergence does not have such a simple interpretation.

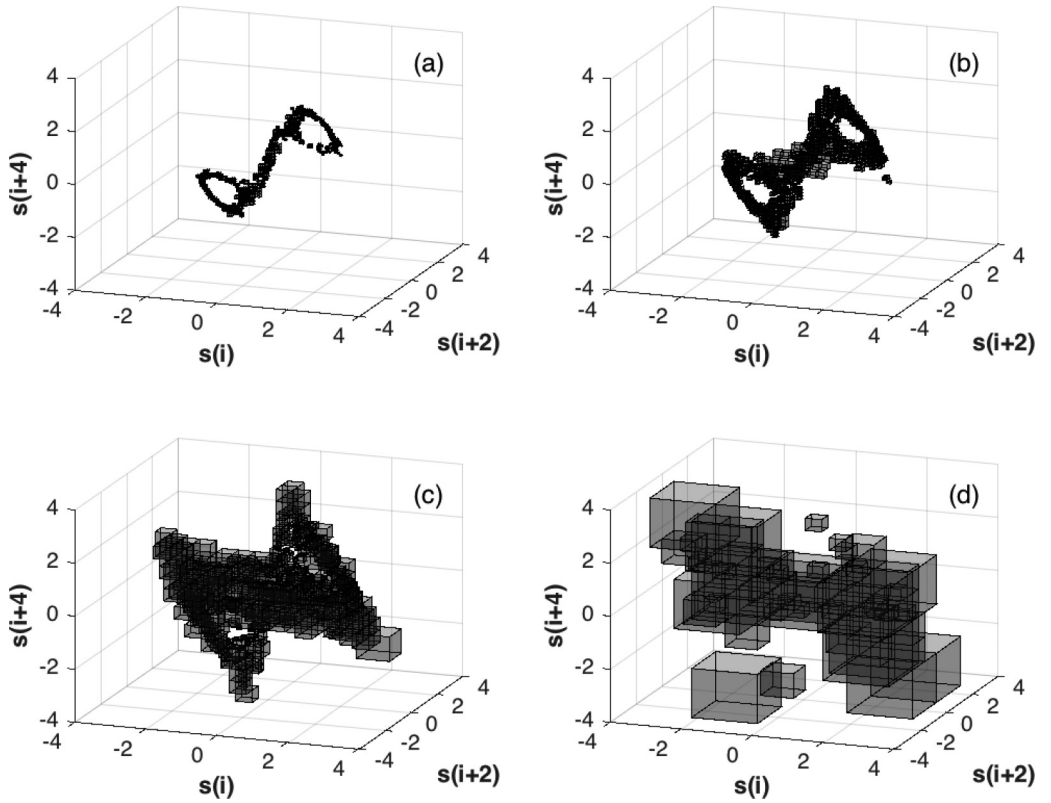

FIG. 2. Histogram bins found by top-down partitioning of the Sprott C attractor. The density of the bin with the highest density is $\rho_{\max}$. (a) shows all the bins for the partitioned attractor whose density is $> 0.1\rho_{\max}$, (b) shows bins with densities from $10^{-1}\rho_{\max}$ to $10^{-2}\rho_{\max}$, (c) shows bins with densities from $10^{-2}\rho_{\max}$ to $10^{-4}\rho_{\max}$, and (d) shows bins with densities $< 10^{-4}\rho_{\max}$.

B. Distinguishing attractors

We want to build up statistics on how well we can distinguish the 19 Sprott attractors, so for each attractor we generate a time series of 200 000 points and divide each time series into ten parts of 20000 points each. We embed the 20000 point times series in three dimensions with an embedding delay of two points, and apply the top-down partitioning method to create partitions to divide the phase space into local regions in which the attractor density appears constant; Fig. 2 shows an example of these regions. The embedded Sprott attractors are denoted $\mathbf{S}(i,j)$, where $i = 1,\ldots,19$ indicated the particular Sprott system and $j = 1,\ldots,10$ indicates the part of the time series.

We choose a partition (or bin) on one attractor $\mathbf{S}(i_1,j_1)$,$i_1 = 1,\ldots,19, j_1 = 1,\ldots,5$ and look for partitions on a different attractor $\mathbf{S}(i_2,j_2)$,$i_2 = 1,\ldots,19, j_2 = 6,\ldots,10$ that overlap. From the densities in these two overlapping regions we calculate the Jensen-Shannon divergence [Eq. (8)]. The Jensen-Shannon divergence compares probabilities, so the density in each partition is multiplied by the volume by which the two partitions overlap. There may be more than one partition on $\mathbf{S}(i_2,j_2)$ that overlaps with the partition on $\mathbf{S}(i_1,j_1)$. It is also possible that no partition on $\mathbf{S}(i_2,j_2)$ overlaps with the chosen partition on $\mathbf{S}(i_1,j_1)$, but the Jensen-Shannon divergence still gives a result.

Once the Jensen-Shannon divergence has been calculated, a different partition on $\mathbf{S}(i_1,j_1)$ is chosen, the comparison is repeated and the result is summed with the previous value. The process is continued for all the partitions on $\mathbf{S}(i_1,j_1)$.

The number of errors in identification n_e is given by

$$n_e = \sum_{i_1=1}^{19}\sum_{j_1=1}^{5}\sum_{j_2=1}^{5} H(i_1,j_1,j_2), \tag{9}$$

where $H(i_1,j_1,j_2)$ is defined as

$$\begin{aligned}&\text{if}\\ &\quad \min\left[D_{JS}(\mathbf{S}(i_1,j_1)\|\mathbf{S}(i_2,j_2)\right] i_2 = 1,\ldots,19)\\ &\quad < D_{JS}[\mathbf{S}(i_1,j_1)\|\mathbf{S}(i_1,j_2)]\\ &\quad H(i_1,j_1,j_2) = 1\\ &\text{else}\\ &\quad H(i_1,j_1,j_2) = 0.\end{aligned} \tag{10}$$

The error fraction is n_e divided by the total number of comparisons. For the 19 Sprott attractors with no noise, the error rate in correctly distinguishing the attractors is 0.002.

C. Sprott C vs Sprott D comparison

As an example, we show some of the details of the comparison between the Sprott C and D systems. The Sprott D system was described by

$$\begin{aligned}\frac{dx}{dt} &= -y\\ \frac{dy}{dt} &= x + z\\ \frac{dz}{dt} &= xz + 3y^2.\end{aligned} \tag{11}$$

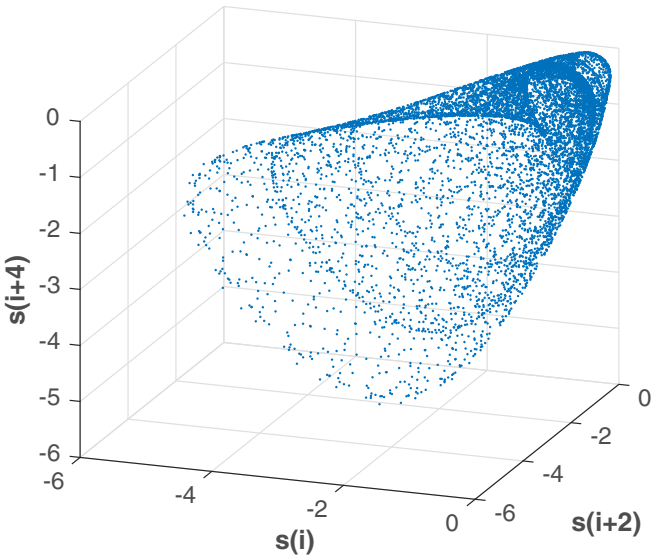

FIG. 3. Embedded time series signal for the Sprott D attractor with an embedding delay of 2.

Figure 3 shows the Sprott D attractor, embedded from a time series of the x signal with a delay of 2. The Sprott C and D attractors have very different shapes, so they could not be compared using the method of Ref. [20]. We choose the unknown signal to be a signal from the Sprott C system, and compare to references from either the C or D system. We take each cluster from the reference and compare to all the other clusters in the signal. Figure 4 shows for each reference cluster the total overlap area with all the clusters in the signal. Figure 4 shows that the overlap between system C and a reference from system C is larger than system C and a reference from system D, so it is not surprising that the Jensen-Shannon divergence between C and C is smaller than the divergence between C and D. For many reference clusters,

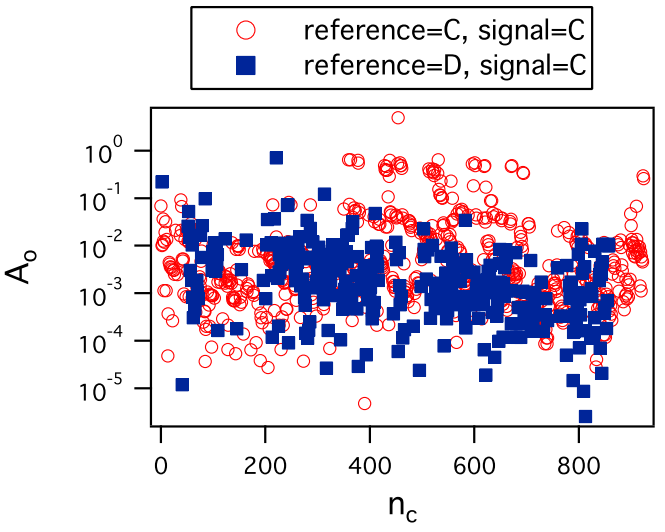

FIG. 4. For each cluster in the reference, total overlap A_o with all the clusters in the signal from system C. The cluster index is n_c. The average overlap area when reference is system C is 0.03, while the average when D is the reference is 0.003.

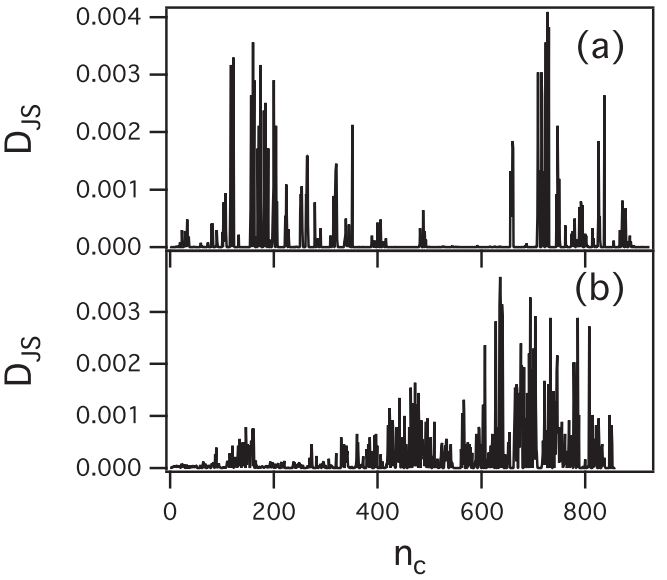

FIG. 5. Jensen-Shannon divergence D_{JS} [Eq. (8)] for each reference cluster and all of the other clusters from the signal from system C. in (a) the reference is from system C, while in (b), the reference is from system D. The sum of the Jensen-Shannon divergence for all reference clusters is 0.14 when the reference is C, while the same sum is 0.21 when the reference is D. A smaller value of the Jensen-Shannon divergence indicates a better match.

the overlap area is 0, which is why the Jensen-Shannon divergence is used in place of the Kulback-Leibler divergence. Figure 5 shows the Jensen-Shannon divergence between each cluster in the reference and all the clusters in the signal. When we compare five realizations of the Sprott C system to five instances of a reference the same system, the mean Jensen-Shannon divergence is 0.03 ± 0.01. When comparing the Sprott C to references from the Sprott D system, the mean Jensen-Shannon divergence is 0.21 ± 0.01. Likewise, when comparing five realizations of the Sprott D system to five instances of a reference the same system, the mean Jensen-Shannon divergence is 0.03 ± 0.01. When comparing the Sprott D to references from the Sprott C system, the mean Jensen-Shannon divergence is 0.23 ± 0.01.

D. Noise considerations

Rarely in the real world do we have access to a noise-free signal, so the attractor density partitioning method must also be robust to added noise. When noise is added to a signal and the result is normalized, the amplitude of the actual signal is reduced. In order for the densities such as that shown in Fig. 2 to properly overlap, the density for the noisy signal must be rescaled so that the actual signal covers the same region of phase space as the noise-free signal. It is complicated to calculate the size of this rescaling, however, as it depends on the relative statistics of the noise and the signal. In order to avoid this complication, we add noise with the same amplitude and spectrum to the original noise-free signal

For this noise study, bandpass filtered noise with the same amplitude and spectrum was added to both $\mathbf{S}(i_1, j_1)$ and

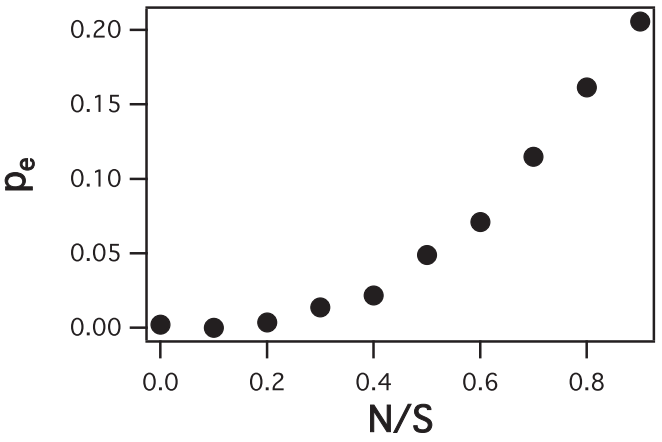

FIG. 6. Probability of error p_e in identifying the 19 Sprott systems when additive noise is present. N/S is the noise level divided by the signal level.

$\mathbf{S}(i_2, j_2)$. The noise spectrum occupied the same frequency range as the spectra of the Sprott systems.

Figure 6 shows the error rate for identifying the Sprott attractors when noise was added to all signals. Figure 6 shows a slight drop in the probability of error for small noise levels. Beyond that, the probability of error increases to about 20% when the noise is as large as the signal.

IV. EXPERIMENT: DETECTING NONLINEARITY IN OP AMPS

In the previous section, we showed that density partitioning could distinguish between attractors that were very different from each other. In this section, we use the same method to detect very small changes in a circuit experiment.

Operational amplifiers (op amps) are widely used devices that are generally assumed to be linear. Like all active electronic devices, however, op amps are based on semiconductors whose behavior is not linear. Attempts are made in amplifier design to minimize nonlinearity, but in the experiment in this section we show that we can still detect that op amps are not linear.

The experimental circuit is shown in Fig. 7. To create the signal driving signal V_0, a series of sinusoids with different

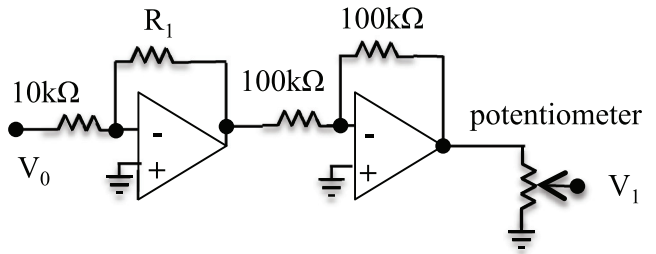

FIG. 7. Op amp circuit used in the experiment. The op amps are type OP-07. The resistor R_1 could be changed to change the gain of the circuit. The potentiometer was used to maintain the peak to peak amplitude of the output signal V_1 at a constant value of 2 V. The driving signal V_0 was sine wave with a chaotic frequency modulation.

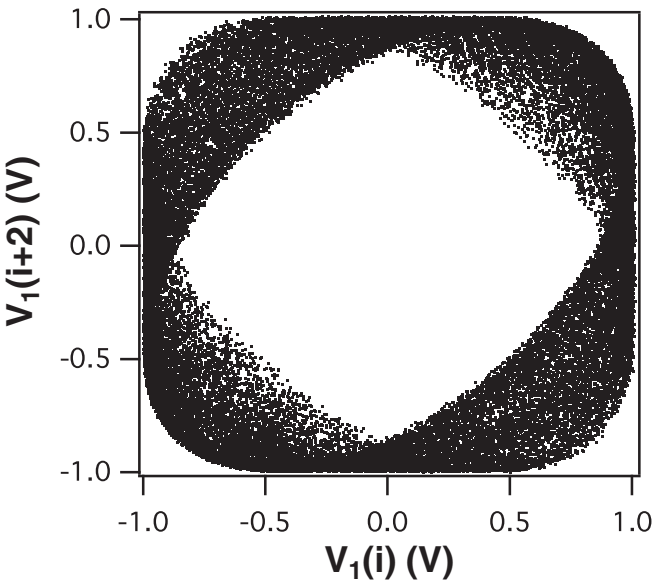

FIG. 8. Output signal $V_1(i)$ from the op amp experiment, embedded in two dimensions with a delay of two points.

frequencies were concatenated so that they matched in phase. The center frequency of each sinusoid was 10 kHz, while the frequency deviation was determined by a signal derived from the shift map. The process of creating the signal V_0 was

$$\begin{aligned} x_{n+1} &= 2.1x_n \mod 1 \\ T_n &= 100 + \beta(x_n - 0.5) \\ V_0(i, \ldots, i + T_n) &= \sin(2\pi\tau i/T_n). \end{aligned} \tag{12}$$

The time step $\tau = 10^{-6}$s, while the frequency deviation factor $\beta = 40$, resulting in a bandwidth of 3 kHz. The gain of the circuit was changed by changing the resistor R_1. The goal of this experiment was to measure changes in the op amp circuit, not in the digitizer, so the potentiometer was used to maintain the peak to peak amplitude of the output signal V_1 at 2 V, independent of the op amp circuit gain.

The output signal V_1 was digitized at a rate of 100 000 points/sec. Dimension estimates show that V_1 is two dimensional [27]. Figure 8 shows V_1 embedded in two dimensions with a embedding delay of 2. The circuit gain for Fig. 8 was 1.0.

A. Density partitioning

The top-down partitioning approach was applied to the embedded circuit data, as described previously in Eqs. (2)–(8), but the top-down method had difficulty in detecting changes in the circuit experiment. The top-down method appears to create overly large partitions in two dimensions. Figure 9 shows the partitions found for the circuit attractor of Fig. 8, using a total of 100 000 points. The top-down partitioning in Fig. 9 does find a large number of partitions (881 partitions found), but the partitions are very uniform, washing out fine scale variations in the actual data density. Figure 9 shows little density variation over large regions of the attractor.

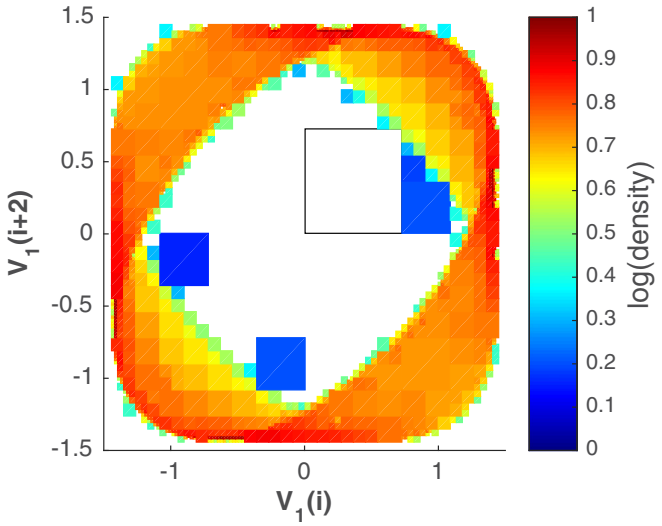

FIG. 9. Partitions for the V_1 signal from the op amp circuit (gain = 1) using the top-down partitioning method. A total of 881 partitions were found.

It appears that in two dimensions, top-down partitioning does not capture fine scale variations in the data. As mentioned above, the initial partitions in top-down partitioning are not related to the actual data, and the partition size is halved at each iteration, so some structure in the data may be missed. The attractor is obviously nonuniform on the largest length scales, and we hope to see structure on small length scales, but in two dimensions there appears to be an intermediate length scale on which the attractor can not be distinguished from a uniform distribution. In order to see small length scale variations in the op amp circuit attractor, we use the bottom-up partitioning method. While bottom-up partitioning is slower than top-down partitioning, its speed is not too bad for two-dimensional data, and bottom-up partitioning is better at capturing structure in the data at small length scales.

In order to speed up the bottom-up clustering method, the top-down method is first applied to the op amp data, and the minimum bin size found from the top-down data is used as the length scale for binning the op amp data. The bottom-up partitioning will be accelerated because when expanding the size of the region for partitioning, the region can be expanded by one set of bins in each direction, so points may be added to the partition in groups, rather than just one at a time. The top-down partitioning divides partitions by a factor of 2 along each axis for each level of partitioning; for the circuit data, the maximum number of divisions by 2 was 9, so the data histogram should have $2^9 = 512$ bins along each axis. The smallest length scale on which the top-down partitioning detected structure in the attractor was 1/512 of the full attractor size.

B. Data histogram

Each of the d dimensional points is sorted into a bin by assigning a bin number. If the maximum and minimum values of s are $s_{\max}$ and $s_{\min}$, and there are N_b bins along each dimension of the phase space, then the bin number k_b for a point $s(i)$ is

$$k_b = \sum_{j=1}^{d} [(s(i) - s_{\min})/(s_{\max} - s_{\min})] N_b^{j-1}. \quad (13)$$

For the circuit data, $d = 2$ and $N_b = 512$. In order to save memory, only the bins containing points are counted.

C. Bottom-up partitioning

Most bins contain only a small number of points, so the bottom-up partitioning method may be used to combine bins. To initiate this procedure, an initial bin k_1 and its neighboring bins $k_2, \ldots, k_4$ are chosen. The bin k_1 is chosen from among the filled bins. The neighboring bins are $k_2 = k_1 + 1, k_3 = k_1 + N_b, k_4 = k_1 + N_b + 1$ (for the two-dimensional case). Any points in these bins are divided into four partitions, and the information criteria $R(\mathbf{X}, \Theta)$ of Eq. (6) is calculated. If $R(\mathbf{X}, \Theta) > 1$ bit, then these bins are retained as partitions; otherwise, the number of points is increased by considering the next level of bins.

At each level of expansion l_e, the bins searched for points are

$$k_1 + i_1 + i_2 N_b \quad i_1, i_2 = 0, \ldots, l_e - 1. \quad (14)$$

The points found in these bins are divided into four partitions and $R(\mathbf{X}, \Theta)$ is calculated based on these four partitions. Expansion continues until $R(\mathbf{X}, \Theta) > 1$ bit or $l_e/N_b > 0.1$. The second requirement prevents overly large partitions in regions of the attractor that contain few or no points.

In order to prevent partitions from overlapping, expansion is also stopped if a bin is encountered that is already part of a partition. Because the partitions are arranged on a grid, this requirement may mean that some points are not assigned to a partition, but in practice, the points that are missed are in low-density parts of the attractor, so missing these points did not have a large effect on attractor comparisons.

The bottom-up partitioning method could also be used for embedding dimensions higher than 2, but the requirement for searching nearby histogram bins made the bottom-up method slower than the top-down method for dimensions greater than 2.

Figure 10 shows the results of using bottom-up partitioning on the op amp circuit data. The bottom-up partitioning yielded 431 total partitions. More importantly, the partitions better reflected the structure of the attractor.

D. Detecting nonlinearity

Time series of 1×10^6 points at a digitization rate of 100 000 points/s were obtained for the op amp circuit with gains of 1.0, 1.1, 1.2, 1.3, 1.5, 1.6, 1.8, and 2.0. Each time series was divided into 10 parts of 100 000 points each. Each section of the time series was embedded in two dimensions with a delay of two points to create $\mathbf{S}(i, j)$ $i = 1 \ldots 8$, $j = 1, \ldots, 10$, where the index i referred to the gain and j referred to the particular 100 000 point section of the time series. Bottom up density partitioning was performed on all attractors. The density partitioned attractors for $i = 1, \ldots, 8$, $j = 6, \ldots, 10$ were compared to the density partitioned attractors for a gain of 1, or $i = 1$, $j = 1, \ldots, 5$ using the Jensen-Shannon divergence

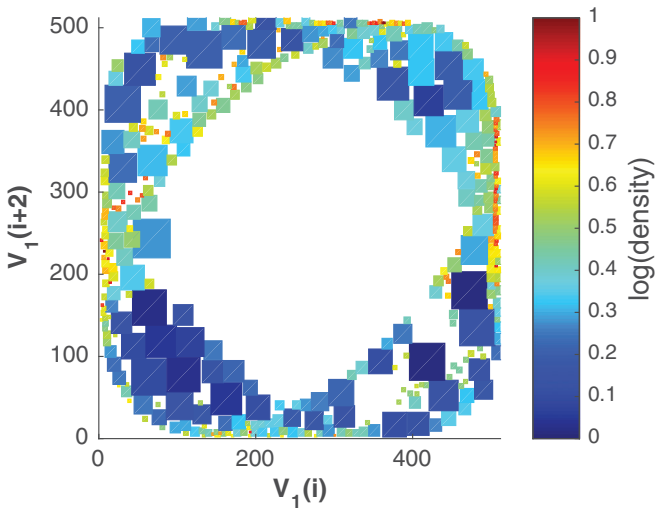

FIG. 10. Partitions for the V_1 signal from the op amp circuit (gain = 1) using the bottom-up partitioning method. A total of 431 partitions were found.

as defined in Secs. III A and III B above. For each gain level therefore there were a total of 25 comparisons. Figure 11 shows the Jensen-Shannon divergence $D_{JS}[\mathbf{S}(i_1, j_1)\|\mathbf{S}(i_2, j_2)]$ found by comparing density partitioned attractors for different gains to density partitioned attractors for gain = 1. The error bars show the standard deviation for each comparison. Figure 11 shows that the bottom-up partitioning detects a monotonically increasing difference in the op amp output as gain increases from 1–2. Op amps are usually treated as linear devices, so this plot shows that the density partitioning method is sensitive to small differences as well as large differences in attractors.

Figure 11 also shows the derivative difference Δ between attractors calculated for the same data by the method of

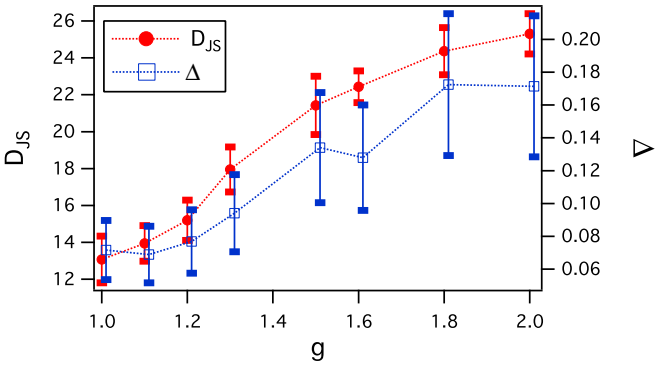

FIG. 11. The left axis is Jensen-Shannon divergence D_{JS} as a function of gain g found by comparing attractors from the op amp circuit experiment to attractors for $g = 1$. The attractors were partitioned using the bottom-up density partitioning method. The error bars were the standard deviations found by performing 25 comparisons at each gain level. The right axis shows the difference Δ between op amp attractors found by using the derivative difference method of Ref. [27]. The curve for the density method is slightly offset along the gain axis so that the error bars for both curves are visible. The two methods yield similar results, but the variance for the density method is smaller.

Ref. [27]. The derivative difference method was also sensitive to small differences between attractors, but the derivative difference method required that the two attractors remain close together in phase space, so the method was less general. As can be seen in Fig. 11, the variance of the derivative difference method was smaller.

Standard methods for detecting nonlinearity include driving a component with a sum of two sine waves with incommensurate frequencies and measuring the difference in the power spectrum at the sum and difference frequencies. Such measurements require a very stable frequency sources; the computer-generated sine waves we used for this experiment were not stable enough to detect nonlinearity in the op amp experiment. The phase-space methods can detect nonlinearity with a relatively unstable signal source.

V. CONCLUSIONS

We have shown a method to estimate the invariant measure of a dynamical system by partitioning an embedded signal from the system into regions containing approximately equal information. While this method can detect small changes in an attractor, as shown in a circuit experiment, it is not limited to attractors that are similar to each other, as shown by the ability to distinguish between the 19 different Sprott attractors. The way in which the partitioning is done depends on the embedding dimension, but for dimensions of 2 or 3, the partitioning is relatively fast to compute.

These partitioning methods partition the data into regions that can not be distinguished from constant densities, so they provide a way to eliminate redundant information from the description of an attractor. As such, they may provide a useful first step in networking or graph theory calculations for systems with many data points. Many of these calculations require the calculation of an affinity matrix, which compares each point in a data set to every other point. Removing redundant information by partitioning will reduce the number of data points, reducing the size of the affinity matrix.

APPENDIX: DERIVING THE KL DIVERGENCE IN THE PARTITION DECISION CRITERIA

Deriving a closed-form solution for the Kullback-Leibler (KL) divergence between Dirichlet distributions with concentration parameter vectors α and α' requires a few nontrivial steps. The Dirichlet distribution is a probability distribution, $\mathrm{Dir}(\pi|\alpha)$, over a $k - ary$ exclusive probability,

$$\mathrm{Dir}(\pi|\alpha) = \frac{1}{\mathbf{Z}(\alpha)} \prod_{k=1}^{K} (\pi_k)^{\alpha_k - 1},$$

$$\text{where } \mathbf{Z}(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\alpha_0)} \text{ such that } \sum_{k=1}^{K} \pi_k = 1, \quad \text{(A1)}$$

where $\alpha_0 \doteq \sum_{k=1}^{K} \alpha_k$. The KL divergence has the form

$$D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)] = \int_{K-\mathrm{simplex}} d\pi \,\mathrm{Dir}(\pi|\alpha') \ln\left(\frac{\mathrm{Dir}(\pi|\alpha')}{\mathrm{Dir}(\pi|\alpha)}\right). \tag{A2}$$

The technical issue that needs to be addressed is simplifying the domain of the integral over the K simplex while obeying the sum rule on the probability p. The integration domain can be extended over R_+^K if the sum rule is enforced by using a δ function

$$\int\limits_{K-\mathrm{simplex}} d\pi = \int_0^\infty \int_0^\infty \ldots \int_0^\infty \prod_{k=1}^{K} d\pi_k \delta\left(1 - \sum_{k=1}^{K} \pi_k\right), \pi_k \in [0,1]. \tag{A3}$$

The bounds are simpler but the d function still makes the integral complicated. However, by substituting the Fourier transform of the d function and performing a change of variables ($k = ik$):

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk e^{-ikx} \rightarrow \delta(x) = \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa e^{\kappa x}. \tag{A4}$$

This adds an extra dimension to the integral but the boundary is now straightforward,

$$\int\limits_{K-\mathrm{simplex}} d\pi = \frac{1}{2\pi i}\int_{-\infty}^{\infty} d\kappa e^{\kappa} \int_0^\infty d\pi_1 e^{-\kappa\pi_1} \int_0^\infty d\pi_2 e^{-\kappa\pi_2} \ldots \int_0^\infty d\pi_K e^{-\kappa\pi_K}. \tag{A5}$$

The KL divergence can be written in terms of this new integration domain as

$$D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)] = \frac{1}{Z(\alpha')}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa e^{\kappa} \prod_{k=1}^{K} \int_0^\infty d\pi_k e^{-\kappa\pi_k} (\pi_k)^{\alpha'_k - 1} \left(\ln\frac{Z(\alpha)}{Z(\alpha')} + \sum_{j=1}^{K} (\alpha'_j - \alpha_j)\ln\pi_j\right) \tag{A6}$$

$$= \frac{1}{Z(\alpha')}\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa e^{\kappa} \left(\ln\frac{Z(\alpha)}{Z(\alpha')} + \sum_{j=1}^{K} (\alpha'_j - \alpha_j)\frac{\int_0^\infty d\pi_j \ln\pi_j e^{-\kappa\pi_j}(\pi_{k'})^{\alpha'_j - 1}}{\int_0^\infty d\pi_j e^{-\kappa\pi_j}(\pi_j)^{\alpha'_j - 1}}\right)$$

$$\times \left(\prod_{k=1}^{K} \int_0^\infty d\pi_k e^{-\kappa\pi_k}(\pi_k)^{\alpha'_k - 1}\right). \tag{A7}$$

There are few definite integrals that are useful in further simplifying this expression:

$$l\int_0^\infty d\pi_k e^{-\kappa\pi_k}(\pi_k)^{\alpha'_k - 1} = \kappa^{-\alpha'_k}\Gamma(\alpha'_k) \tag{A8}$$

$$\int_0^\infty d\pi_j \ln\pi_j e^{-\kappa\pi_j}(\pi_j)^{\alpha'_j - 1} = \kappa^{-\alpha'_j}\Gamma(\alpha'_j)[\psi(\alpha'_j) - \ln\kappa]. \tag{A9}$$

After some algebraic manipulation and then using these integrals,

$$D_{KL} = \frac{1}{Z(\vec{\alpha}')}\left(\prod_{k=1}^{K}\Gamma(\alpha'_k)\right)\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa\kappa^{-\alpha'_0}e^{\kappa}\left(\ln\frac{Z(\vec{\alpha})}{Z(\vec{\alpha}')} + \sum_{j=1}^{K}\frac{\alpha'_j - \alpha_j}{\Gamma(\alpha'_j)}\kappa^{\alpha'_j}\int_0^\infty d\pi_j \ln\pi_j e^{-\kappa\pi_j}(\pi_j)^{\alpha'_j - 1}\right) \tag{A10}$$

$$= \Gamma(\alpha'_0)\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa\kappa^{-\alpha'_0}e^{\kappa}\left(\ln\frac{Z(\vec{\alpha})}{Z(\vec{\alpha}')} + \sum_{k=1}^{K}(\alpha'_k - \alpha_k)\cdot\psi(\alpha'_k) - (\alpha'_0 - \alpha_0)\ln\kappa\right)(\mathrm{Relabel}: j \rightarrow k) \tag{A11}$$

$$= \Gamma(\alpha'_0)\left(\ln\frac{Z(\vec{\alpha})}{Z(\vec{\alpha}')} + \sum_{k=1}^{K}(\alpha'_k - \alpha_k)\psi(\alpha'_k)\right)\left(\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa\kappa^{-\alpha'_0}e^{\kappa}\right) - \Gamma(\alpha'_0)(\alpha'_0 - \alpha_0)\left(\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa\kappa^{-\alpha'_0}e^{\kappa}\ln\kappa\right). \tag{A12}$$

To evaluate the remaining integrals, one can note these are known inverse Laplace transforms,

$$l\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa\kappa^{-\nu}e^{\kappa t} = \frac{t^{\nu-1}}{\Gamma(\nu)} \tag{A13}$$

$$\frac{1}{2\pi i}\int_{-i\infty}^{i\infty} d\kappa\kappa^{-\nu}e^{\kappa t}\ln\kappa = t^{\nu-1}\frac{\psi(\nu) - \ln t}{\Gamma(\nu)} \text{for } \nu > 0, t > 0. \tag{A14}$$

Finally, we obtain the general form for the KL divergence,

$$D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)] = \ln\frac{Z(\alpha)}{Z(\alpha')} + \sum_{k=1}^{K}(\alpha'_k - \alpha_k)\psi(\alpha'_k) - (\alpha'_0 - \alpha_0)\psi(\alpha'_0). \tag{A15}$$

The specific KL divergence used in the information partition criterion can be derived from this expression by substituting $\alpha_k = \frac{M}{K} + \frac{1}{2}$ and $\alpha'_k = m_k + \frac{1}{2}$ while noting that $\alpha'_0 = \alpha_0$:

$$D_{KL}[\mathrm{Dir}(\pi|\alpha')||\mathrm{Dir}(\pi|\alpha)] = \sum_{k=1}^{K}\left[\left(m_k - \frac{M}{K}\right)\psi\left(m_k + \frac{1}{2}\right) - \ln\Gamma\left(m_k + \frac{1}{2}\right) + \ln\Gamma\left(\frac{M}{K} + \frac{1}{2}\right)\right]. \tag{A16}$$

[1] J. D. Farmer and J. J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).
[2] G. Sugihara and R. M. May, Nature (London) **344**, 734 (1990).
[3] M. Casdagli, Physica D **35**, 335 (1989).
[4] B. R. Hunt, E. J. Kostelich, and I. Szunyogh, Physica D **230**, 112 (2007).
[5] R. Brown, N. F. Rulkov, and E. R. Tracy, Phys. Rev. E **49**, 3784 (1994).
[6] G. P. King, R. Jones, and D. S. Broomhead, Nucl. Phys. B, Proc. Suppl. **2**, 379 (1987).
[7] T. Buzug and G. Pfister, Phys. Rev. A **45**, 7073 (1992).
[8] P. Grassberger and I. Procaccia, Physica D **9**, 189 (1983).
[9] N. B. Tufillaro, R. Holzner, L. Flepp, E. Brun, M. Finardi, and R. Badii, Phys. Rev. A **44**, R4786 (1991).
[10] H. Suetani, K. Soejima, R. Matsuoka, U. Parlitz, and H. Hata, Phys. Rev. E **86**, 036209 (2012).
[11] C. Diks, W. R. van Zwet, F. Takens, and J. DeGoede, Phys. Rev. E **53**, 2169 (1996).
[12] J.-P. Eckmann and D. Ruelle, Rev. Mod. Phys. **57**, 617 (1985).
[13] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).
[14] R. V. Donner, M. Small, J. F. Donges, N. Marwan, Y. Zou, R. X. Xiang, and J. Kurths, Int. J. Bifurcation Chaos **21**, 1019 (2011).
[15] J. D. Farmer, Zeitschrift fur Naturforschung A (Astrophysik, Physik und Physikalische Chemie) **37A**, 1304 (1982).
[16] J. Wood, D. E. Root, and N. B. Tufillaro, IEEE Trans. Microwave Theory Tech. **52**, 2274 (2004).
[17] M. D. Todd, K. Erickson, L. Chang, K. Lee, and J. M. Nichols, Chaos **14**, 387 (2004).
[18] U. von Luxburg, Stat. Comput. **17**, 395 (2007).
[19] E. Bradley and H. Kantz, Chaos **25**, 097610 (2015).
[20] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsmring, Rev. Mod. Phys. **65**, 1331 (1993).
[21] J. C. Sprott, Phys. Rev. E **50**, R647(R) (1994).
[22] T. L. Carroll, Chaos **25**, 013111 (2015).
[23] S. Kullback and R. A. Leibler, Ann. Mat. Stat. **22**, 79 (1951).
[24] T. M. Cover and J. M. Thomas, *Elements of Information Theory* (Wiley, New York, 2006).
[25] M. L. Menendez, J. A. Pardo, L. Pardo, and M. C. Pardo, J. Franklin Inst. **334B**, 307 (1997).
[26] L. M. Pecora, L. Moniz, J. Nichols, and T. L. Carroll, Chaos **17**, 013110 (2007).
[27] T. L. Carroll, Chaos **21**, 023128 (2011).