Self-dissimilarity as a high dimensional complexity measure

David H. Wolpert and William Macready[1]
[1]*NASA Ames Research Center, Moffett Field, CA 94035,*
{dhw@email.arc.nasa.gov,wgm@email.arc.nasa.gov

For many systems characterized as "complex" the patterns exhibited on different scales differ markedly from one another. For example the biomass distribution in a human body "looks very different" depending on the scale at which one examines it. Conversely, the patterns at different scales in "simple" systems (e.g., gases, mountains, crystals) vary little from one scale to another. Accordingly, the degrees of self-*dis*similarity between the patterns of a system at various scales constitute a complexity "signature" of that system. Here we present a novel quantification of self-dissimilarity. This signature can, if desired, incorporate a novel information-theoretic measure of the distance between probability distributions that we derive here. Whatever distance measure is chosen, our quantification of self-dissimilarity can be measured for many kinds of real-world data. This allows comparisons of the complexity signatures of wholly different kinds of systems (e.g., systems involving information density in a digital computer vs. species densities in a rain-forest vs. capital density in an economy, *etc.*). Moreover, in contrast to many other suggested complexity measures, evaluating the self-dissimilarity of a system does not require one to already have a model of the system. These facts may allow self-dissimilarity signatures to be used as the underlying observational variables of an eventual overarching theory relating all complex systems. To illustrate self-dissimilarity we present several numerical experiments. In particular, we show that underlying structure of the logistic map is picked out by the self-dissimilarity signature of time series' produced by that map.

I. INTRODUCTION

The search for a measure quantifying the intuitive notion of the "complexity" of systems has a long history [1, 6]. One striking aspect of this search is that for almost all systems commonly characterized as complex, the spatio-temporal patterns exhibited on different scales differ markedly from one another. Conversely, for systems commonly characterized as simple the patterns are quite similar.

The human body is a familiar example of such self-dissimilarity; as one changes the scale of the spatio-temporal microscope with which one observes the body, the pattern one sees varies tremendously. The (out of equilibrium) terrestial climate system is another excellent illustration, having very different dynamic processes operating at all spatiotemporal scales, and typically being viewed as quite complex. Complex human artifacts also share this property, as anyone familiar with large-scale engineering projects will attest.

Conversely, the patterns at different scales in "simple" systems like gases and crystals do not vary significantly from one another. Similarly, once it has fossilized a dead organism is static across time, i.e., completely self-similar along the time axis. What relatively little spatio-temporal complexity it still possesses is purely spatial, a relic of its complex past. Based on such examples, we argue that it is the self-*similar* aspects of simple systems, as revealed by allometric scaling, scaling analysis of networks, etc. [7], that reflects their inherently simple nature. Such self-similarity means that the pattern across all scales can be encoded in a short description for simple systems, unlike the pattern for complex systems.

More generally, even if one could find a system commonly viewed as complex that was clearly self-similar in all important regards, it is hard to see how the same system wouldn't be considered even more "complex" if it were self-dissimilar. Indeed, it is hard to imagine a system that is highly self-dissimilar in both space and time that wouldn't be considered complex.

Accordingly, it is the self-*dissimilarity* (SD) between the patterns at various scales that constitutes the complexity "signature" of a system [11]. Intuitively, such a signature tells us how the information and its processing [2] at one scale in a system is related to that at the other scales. Highly different information processing at different scales means the system is efficient at encoding as much processing into its dynamics as possible. In contrast, having little difference between the various scales, i.e., high redundancy, is often associated with robustness.

The simplest version of such a signature is to reduce all of the patterns to a single number measuring their aggregate self-dissimilarity. This would be analogous to conventional measures which quantify a system's "complexity" as a single number [12]. We can use richer signatures however. One is the symmetric matrix of the dissimilarity values between all pairs of patterns at different scales. More generally, say we have a dissimilarity measure that can be used to quantify how "spread out" a set of more than two patterns is. Then we can measure the spread of triples of scale-indexed patterns, quadruples, etc. In such a situation the signature could be a tensor, (e.g., a real number for each possible triple of patterns), not just a matrix.

SD signatures may exploit model-based understanding about the system generating a data set of spatio-temporal patterns (for example, to statistically extend that data set). However they are functions of such a data set rather than of any model of the underlying system. So in contrast to some other suggested complexity measures, with SD one does not need to understand a system and then express that understanding in a formal model in order to measure its complexity. This is important if one's complexity measure is to serve as a fundamental observational variable used to gain understanding of particular complex systems, rather than as a post-hoc characterizer of such understanding.

Indeed, one application of SD is to (in)validate models of the system that generated a dataset, by comparing the SD signature of that dataset to the signature of data generated by simulations based on those models. Model-independence also means that the SD complexity measure can be applied to a broad range of (data sets associated with) systems found in nature, thereby potentially allowing us to compare the processes underlying those types of systems. Such comparisons need not involve formal models. For example, SD signatures can be viewed as machine learning "features" synopsizing a dataset [3]. Given multiple such datasets the associated SD signatures/features can be clustered. This may reveal relationships between the systems underlying those datasets. We can do this even when the underlying systems live in wholly different kinds of spaces, since we are characterizing (the datasets generated by) those systems with their SD signatures, and those signatures all live in the same space (e.g., real-valued matrices). In this way clustering can generate a taxonomy of "kinds of systems" that share the same complexity character. SD signatures can also serve as supervised learning predictor variables for extrapolating a dataset (e.g., into the future, as in non-linear time-series prediction). In all this, SD signatures are "complexity-based" analogues of traditional measures used for these purposes, e.g., power spectra.

The first formalization of SD appeared in [11]. This paper begins by motivating a new formalization. We then present several examples of that formalization. Next we present a discussion of information theoretic measures of dissimilarity between probability distributions, an important issue of SD analysis. We end by illustrating SD analysis with several computer experiments [13]. We end by discussing some of the broader context of the notion of self-dissimilarity

Intuitively, a number quantifying the self-simmilarity of a system (e.g., its fractal dimension) is akin to the first moment of a distribution, with the self-dissimilarity being analogous to the higher-order moments. Our central thesis is that just as the higher higher-order moments capture much that is important concerning a distribution, the self-dissimilarity signature of a system — that which is not captured in the self-simmilarity value — capture much that is important concerning that system. We do not claim that self-dissimilarity captures all that is important in complex systems. We only suggest that self-dissimilarity is an important component of complexity, one with the novel advantage that it can actually be evaluated for many different types of real-world systems, as demonstrated by the quantification of it that we present below.

Furthmore, we do not claim that our particular quantification of it is the best way to measure self-dissimilarity. There are many other quantifications that bear investigating, e.g., Fourier decompositions, wavelet-based analysis, multi-fractals, etc. They seem to be ill-suited in various ways for our purposes. For example, conventional Fourier decompositions and wavelet analysis does not directly compare patterns at two scales; rather they quantify "what is left over" at the finer scale after the coarser scale is accounted for. But it may well turn out that some variant of them is the best quantification of self-dissimilarity.

II. FORMALIZATION OF SELF-DISSIMILARITY

There are two fundamental steps in our approach to constructing the SD signature of a dataset.

The first step is to quantify the scale-dependent patterns in the dataset. We want to do this in a way that treats all scales equally (rather than taking the pattern at one scale to be what's "left over" after fitting the pattern at another scale to a data set, for example). We also want to minimize the *a priori* structure and associated statistical artifacts introduced in the quantification of the patterns. Accordingly, we wish to avoid the use of arbitrary bases, and work with entire probability distributions rather than low-dimensional synopses of such distributions.

The second fundamental step in forming a SD signature is numerically comparing the scale-dependent patterns, which for us means comparing probability distributions. We illustrate these steps in turn.

A. Generation of scale-indexed distributions

1. Let q^* be the element in a space Q_0 whose self-dissimilarity interests us. Usually q^* will be a data set, although the following holds more generally.

2. Typically there is a set of transformations of q^* that we wish our SD measure to ignore. For example, we might want the measure to give the same value when applied both to an image and to a slight translation of that image. We start by applying those transformations to q^*, thereby generating a set of elements of Q_0. Taken as a whole, that set is "cleansed" of what we wish to ignore. Formally, we quantify such an invariance we wish to ignore with a function g that maps any $q_0 \in Q_0$ to the set of all elements of Q_0 related by our invariance to that q_0. We will work with the entire set $g(q^*)$ rather than (as is often done) a lower-dimensional synopsis of it. In this way we avoid introducing

statistical artifacts and the issue of how to choose the synopsizing function.

3. In the next step we apply a series of scale-indexed transformations to the elements in $g(q^*)$ (e.g., magnifications to different powers). The choice of transformations will depend on the precise domain at hand. Intuitively, the scale-indexed sets produced by these transformations are the "patterns" at the various scales. They reflect what one is likely to see if the original q^* were "examined at that scale", and if no attention were paid to the transformations we wish to ignore.

   We write this set of transformations as the θ-indexed set $W_\theta : Q_0 \mapsto Q_1$ (θ is the generalized notion of "scale"). So formally, the second step of our procedure is the application of W_θ to the elements in the set $g(q^*)$ for many different θ values. After this step we have a θ-indexed collection of subsets of Q_1. Each such collection constitutes a pattern of ponts at the associated scale.

   Note that we again work with full distributions rather than synopses of them. This allows us to avoid spatial averaging or similar operations in the W_θ, and thereby avoid limiting the types of Q_0 on which SD may be applied, and to avoid introducing statistical biases.

4. The pattern at each scale, a set, is a probability distribution p^θ. At this point in the procedure we may elect to use machine learning and available prior knowledge [3] to transform p^θ, e.g., by smoothing it. Such a step, which we use in our experiments reported below, can often help us in the subsequent quantification of the dissimilarities between the scales' patterns. More generally, if one wishes to introduce model-based structure into the analysis, it can be done through this kind of transformation [14].

B. Quantifying dissimilarity among multiple probability distributions

Applying the preceding analysis to a q^* will give us a collection of distributions, $\{p^\theta\}$, one such distribution for each value of θ. All those distributions are dfined over the same space, Q_1. It is this collection as a whole that characterizes the system's self-dissimilarity.

Note that different domains Q_0 will have different spaces Q_1. So to be able to use SD analysis to relate many different domains, we need to distill each domain's collection $\{p^\theta\}$, consisting of many distributions over the associated Q_1, into values in some common tractable space Q^*. In fact, often there is too much information in a collection of distributions over Q_1 values for them to be a useful way of analyzing a system; even when just analyzing a single system by itself, without comparing it to other systems, often we will want to distill its collection down to a tractable characterization, i.e., to an element of Q^*.

Now what we are ultimately interested in concerning any such collection is the dissimilarity of the distributions comprising it. So a natural choice for Q^* is one or more real numbers measuring how "spread out" the distributions in any particular collection are. Accordingly, we need a measure ρ quantifying how spread out an arbitrary collectionof distributions is.

We want to use ρ both to quantify the aggregate self-dissimilarity of an entire collection, and to quantify the dissimilarity between any pair of distributions from the collection. More generally, we would like to be able to use ρ to quantify the dissimilarity relating any n-tuple of distributions from the collection. Ideally then, such a measure ρ of how "spread out" a collections of distributions is should:

1. Obey the usual properties of a metric when it takes two arguments, i.e., obey the usual properties when comparing two distributions each of which is just a single delta function. More generally it should obey the extension of those properties appropriate for when there are more than two arguments, and in particular the extensions for when those arguments are sets of multiple distributions) [10];

2. Be finite even for the delta-function distributions commonly formed from small data sets;

3. Be quickly calculable even for large spaces;

4. Have a natural interpretation in terms of the total amount of information stored in its (probability distribution) arguments.

Until recently, perhaps the measure best satisfying these desiderata was the Jensen-Shannon (JS) distance [2], i.e., the entropy of the average of the distributions minus the average of their entropies. However this measure fails to satisfy 1. In Section IV we present an alternative, which like JS distance obeys 3 and 4, and may be better suited to SD analysis. (It is unclear whether it obeys 1.)

Recent work has uncovered many multi-argument versions of distance, called **multimetrics** [10]. These obey 1 through 2 by construction, and many of them obey 3 as well. These are what we actually use in our experiments. However the multimetrics uncovered to date do not obey 4.

III. EXAMPLES

To ground the discussion we now present some examples of the foregoing:

Example 1: Q_0 is the space of real-valued functions over a Euclidean space X, e.g., a space of images over

$x \in X$. If we wish our measure to ignore a set of translations over X then $g(q_0)$ is that set of translations of image q_0. Thus if $q^* = f(x)$ then $g(q^*)$ is the set $\{f(x-x_1), f(x-x_2), \cdots\}$ where x_i are translation vectors. Each W_θ is magnification by θ followed by windowing about the origin so that only the local structure of the image around x_i is considered. So if T is an operator which truncates an image $f(x)$ to a window around the origin then $W_\theta(g(q_0)) = \{T[f(\frac{x-x_1}{\theta})], T[f(\frac{x-x_2}{\theta})], \cdots\}$. Accordingly each $q_1^{\theta,i} \equiv T[f(\frac{x-x_i}{\theta})]$ is a real-valued function over a subspace of X.

If we discretize X into n bins, we can convert each such function into an element of $\mathbb{R}^n$. In this way each scale's (finite) set of functions gets converted into a (finite) set of Euclidean vectors.

While multimetrics generalize to distances between objects which are not probability densities, to apply the JS or Kullback-Leibler (KL) distance [2] to our scale-indexed sets of vectors we need to convert them to probabilities. Say the range of the functions over X making up Q_0 were finite rather than all of $\mathbb{R}$ (e.g., if the range were a finite set of possible numeric readings on an observation apparatus.) In this case our "vectors" would be fixed-length strings over a finite alphabet (see Ex. 2). In this case we could convert each set of "vectors" to a probability simply by setting that probability to be uniform over the elements of the set and zero off it. For real-valued vectors this is typically ill-behaved (the support of the distribution has measure 0). So we must run a density-estimation algorithm to convert each finite set of vectors in $\mathbb{R}^n$ into a smooth probability density across $\mathbb{R}^n$.

In real-world images made via a noisy observation apparatus typically there is intrinsic blurring that biases nearby points in the image to have similar intensities. We can choose our density estimation algorithm to reflect that effect directly. For example, with a Bayesian density estimation algorithm, we can build the blurring operator into the likelihood function. Alternatively, we can capture such blurring effects directly in the set of transformations g, by by expanding that set to include localized blurring transformations.

However they are produced, we need a way to convert our set of density functions (one for each scale) into a SD signature. The simplest approach is to form the symmetric matrix of all pairwise comparisons whose i, j element is the multimetric (or JS distance or what have you) between the probability of θ_i and that of θ_j.

All of this can be naturally extended to "images" that are not real-valued functions, but instead take on values in some other space (e.g., of symbols, or of matrices). For example, an element of Q_0 could be the positions of particles of various types in $\mathbb{R}^3$.

Note that q^* may itself be generated from an observational windowing process. This may be accounted for in a likelihood model $P(D|q_0)$ which smooths intensities and admits Gaussian noise.

Example 2: This example is a variant of Ex. 1, but is meant to convey the generality of what "scale" might mean. We have the same Q_0 and g as in Ex. 1. However say we are not interested in comparing a q^* to a scaled version of itself. Instead, each θ represents a set of n vectors $\{v_i(\theta) \in X\}$. Then have $W_\theta(q_0)$ be the n-vector "stencil" $(q_0(v_1(\theta)), q_0(v_2(\theta)), \cdots q_0(v_n(\theta)))$. Then we could have ρ be any distance measure over sets of vectors in $Q_1 = \mathbb{R}^n$, as discussed in Ex. 1. (The difference with Ex. 1 is that here we arrived at those vectors without any binning.)

As an example, we could have stencils consist of two points, with $v_1 = 0$ for all θ, and then have $v_2 = \theta a$, where the vector a is the same for all θ. In this example W_θ isolates a pair of points separated by a multiple θ of the vector a. So our self-dissimilarity measure quantifies how the patterns of pairs of points in f separated by θa change as one varies θ. Another possibility is to have $v_1 = R_\theta(a)$, where $R_\theta(\cdot)$ is rotation by θ. In this case our measure quantifies how the patterns of pairs of points changes as one rotates the space.

Another important extension is where $n > 2$, so that we aren't just looking at pairs of points. In particular, say X is N-dimensional, and have $v_i = \theta a_i \; \forall i$, where each a_i is a vector in X, a_1 equaling 0 and θ being the scale, as usual. Then we might want to have the distances between any pair of points in a scale's stencil, $|\theta a_i - \theta a_j|$, be a constant times θ, independent of i and j. This would ensure there is no "cross-talk" between scales; all distances in a scale's stencil are identical. To obey this desideratum requires that the underlying stencil $\{a_i\}$ be a tetrahedron, of at most $N + 1$ points.

Example 3: This example is the same as Ex. 2, except that X is an M-dimensional infinite lattice rather than a Euclidean space, and g and the W_θ are modified appropriately. For instance, we could have $M = 1$ and have symbolic-valued functions f, so that an element of q_0 is a symbolic time series. Take $n = 2$, with $v_1 = 0$, and $v_2 = \theta$, a now being an integer. Since the range of f is now a finite set of symbols rather than the reals, we do not need to do any binning or even density estimation; each $W_\theta(g(q^*))$ is a histogram, i.e., it is already a probability distribution.

Since distributions now are simply vectors in a Euclicean space, we can measure their dissimilarity with something as unsophisticated as L_2 distance. Alternatively, as before, we can compare scales by using JS distance for ρ. In this case our SD measure is an information-theoretic quantification of how time-lagged samples of the time-series q_0 differ from each other as one changes the lag size.

Having $n > 2$ allows even more nuanced versions of this quantification. Furthermore, other choices of ρ (described below) allow it take more than two sets at once as arguments. In this case, ρ takes an entire set of time-lagged samples, running over many time lags, and measures how "spread out" the members of that full set are.

These measures complement more conventional information-theoretic approaches to measuring how the time-lagged character of q_0 varies with lag size. A typical such approach would evaluate the mutual information between the symbol at a random point in q_0 and the symbol θ away, and see how that changes with θ. Such an approach compares singletons: it sees how the distribution of symbols at a single point are related to the distribution of symbols at the single time-lagged version of that point. These new measures instead allow us to compare distributions of n-tuples to one another.

Example 4: This is a dramatically different example to show that self-dissimilarity can be measured for quite different kinds of objects. Let Q_0 be a space of networks, i.e., undirected graphs with labeled nodes. Have $g(q_0)$ be the set of relabelings of the nodes of network q_0. Such relabelings are what we want the SD analysis to ignore.

Have each W_θ run a decimation algorithm on q_0, with θ parameterizing the precise algorithm used. Each such algorithm iteratively grows outward from some fixed starting (θ-independent) node a, tagging some nodes which it passes over, and removing other nodes it passes over. Changing θ changes parameters of the algorithm, e.g., changes which iterations are the ones at which nodes are removed. Intuitively, each algorithm W_θ demagnifies the network by decimation, and then windows it. Different W_θ demagnify by different amounts.

More precisely, at the start of each iteration t, there is a subset of all the nodes that are labeled the "current" nodes for t. Another subset of nodes, perhaps overlapping those current at t, constitutes the "tagged" nodes. During the iteration, for each current node i, a set of non-tagged nodes $S_t(i)$ is chosen based on i. For example, this could be done by looking at all non-tagged nodes within a certain number of links of i. Then a subset of the nodes in $S_t(i)$ is removed, with compensating links added as needed. The remaining nodes are added to the set of tagged nodes, and a subset of them are added to a set of nodes that will be current for iteration $t+1$. Then the process repeats.

At the earliest iteration at which the number of tagged nodes is at least N, the iterations stop, and all remaining nodes in q_0 are removed. Some fixed rule is then used for removing any excess nodes to ensure that the final net has exactly N nodes. (Typically N is far smaller than the number of nodes in q_0.) ρ can then be any algorithm for measuring distance between sets of identically-sized networks.

IV. DISSIMILARITY OF PROBABILITY DISTRIBUTIONS

In the experiments presented below, we use one of the multimetrics discussed in [10]. However other measures could be used, and in particular it is worth briefly discussing measures derived from information-theoretic arguments concerning the distance between probability distributions.

The most commonly used way to define a distance between two distributions is their KL distance. This is the infinite limit log-likelihood of generating data from one distribution but mis-attributing it to the other distributions. Unfortunately, the KL distance between two distributions is infinite if either distribution has points at which it is identically zero; violates the triangle inequality; is not even a symmetric argument of its two arguments. (It is non-negative though, equaling zero iff its two arguments are identical.)

Some proposals have been made for overcoming some of these shortcomings. In particular, the JS distance between two distributions does not blow up and is symmetric. However it violates the triangle inequality [4, 9]. A more important problem for us is that it is not clear that JS distance is the proper information-theoretic measure for SD analysis. To illustrate this it helps to consider an alternative information-theoretic measure for distance between probability distributions, by modifying the type of reasoning originally employed by Shannon.

Say we have a set of K distributions $\{\pi^i\}$. (For us that set is generated by application of g and the members of $\{W_\theta\}$, as discussed above.) Intuitively, our alternative to JS distance quantifies how much information there is in the knowledge of whether a particular x was generated from one member of $\{\pi^i\}$ or another. To do this we subtract two terms, each being an average over all possible K-tuples of x values, $(x_1, x_2, \ldots, x_K)$.

The summand of the first average is the Shannon information in $(x_1, x_2, \ldots, x_K)$ when that K-tuple is produced by simultaneously sampling each of the K distributions, so that each x_i is a sample of the associated π^i. The summand of the second average is the information in $(x_1, x_2, \ldots, x_K)$ according to the "background" version of the joint distribution, in which all information about which distribution generated which x is averaged out. Intuitively, the difference in these averages tells us how much information there is in the labels of which distribution generates which x:

$$\rho(\{\pi\}) = -\sum_{x_1, x_2, \cdots} \prod \pi^i(x_i) \ln\left[\frac{\sum_P \prod_k \pi^k(Px_k)}{\prod_k \pi^k(x_k)}\right] \quad (1)$$

where the $\sum_P$ notation means a sum over all permutations of the $\{x_j\}$ that rearranges them as the $P\{x_j\}$, and the sum is over all such permutations.

Being a KL distance, this ρ equals 0 when all the distributions are equal, and is never negative. It is not yet known though if it is a full-blown multimetric.

V. EXPERIMENTS

We illustrate the SD framework with two simple sets of computational experiments. The datasets (i.e., the q_0's) in all the experiments are functions over either one-dimensional or two-dimensional finite lattices. The SD

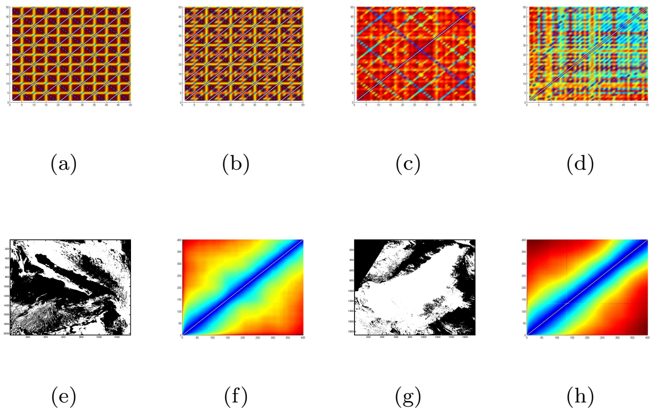

FIG. 1: Self-dissimilarity signatures of binary datasets. Blue indicates low dissimilarity (high similarity), and red indicates high dissimilarity (low similarity): (a) the repeating sequence 1111100000, (b) the repeating sequence 1111111000, (c) a quasi-periodic sequence, (d) the cantor set. For each of these datasets the aggregate self-dissimilarity of the associated scale-indexed set of distributions are 15.5. 13.9, 50.3, and 2.4 respectively. All signatures were obtained using a window of length 9. The signatures (f) and (h) are from the satellite images (e) and (g) over Baja California and Greenland respectively. A 3x3 window was used for these two-dimensional images.

analyses we employed were special cases of Ex. 3, using a square observation "window" of width w to specify the W_θ.

In our first experiments our datasets were binary-valued (i.e., each q_0 was a map from a lattice into $\mathbb{B}$). Accordingly, the task of estimating each scale's probability density, p^θ, simplifies to estimating the probability of sequences of w bits. For small w this can be done using frequency counts (cf. Ex. 3.). We then used a modified bounding box multimetric[10]:

$$\rho(p^{\theta_1}, p^{\theta_2}, \cdots) = -1 + \sum_i \max(p_i^{\theta_1}, p_i^{\theta_2}, \cdots) \qquad (2)$$

where p_i^θ is the i'th component of the w-dimensional Euclidean vector p^θ. Note that being a multimetric, this measure can be used to give both the aggregate self-dissimilarity of all distributions $\{p^\theta\}$ as well as the distance between any two of the distributions.

The pairwise (matrix) SD signatures of six datasets are presented in Fig. 1. The integrals were all evaluated by Monte Carlo importance sampling. The periodicity of the underlying data in 1(a),(b) is reflected in the repeating nature of the SD signature. The quasiperiodic dataset, 1(c) shows hints of periodicity in its signature, and significantly greater overall structure. The fractal-like object 1(d) shows little overall structure (beyond that arising from finite-data-size artifacts). 1(e),(g) show results for satellite images which have been thresholded to binary values.

Note that the aggregate self-dissimilarity values are far from zero. To a degree this is due to edge effects and other statistical artifacts. However even without such artifacts one would expect a non-zero aggregate self-dissimilarity for the periodic data. The reason is that our window width is on the order of the period of the system. An analogy is a crystal examined with a window on the scale of the lattice spacing. While such a crystal would appear roughly self-similar for windows extending over a large number of lattice cells, when the window width is of the order of the lattice spacing, the pattern *does* change as one magnifies the data. Examined for window widths around the lattice spacing, the system is not self-similar.

Clustering of our 6 datasets is done by finding the partitions of (a), (b), (c), (d), (f), (h) which minimize the total intra-group multimetric distance. For 2 clusters the optimal grouping is [(a)(b)(c)(f)(h)] and [(d)]; for 3 clusters the best grouping is [(a)(b)(c)], [(d)], and [(f)(h)]; for 4 clusters the best grouping is [(a)(b)(c)], [(d)], [(f)], and [(h)]; and for 5 clusters the best grouping is [(a)], [(b)(c)], [(d)], [(f)], and [(h)].

We also provide results for the time series generated by the logistic map $x_{t+1} = rx_t(1 - x_t)$, where as usual r is a parameter varying from 0 to 4 and $0 \leq x_t \leq 1$ [15].

We iterated the map 2000 times before collecting data, to ensure that the data is taken from the attractor. For each r-dependent time series on the attractor we generate a self-dissimilarity signature by taking g to be possible initial conditions x_0, and W_θ to be a decimation and windowing, as in Ex. 3. W_θ acts on a real-valued vector $\mathbf{x} = [x_1, x_2, \cdots]$ to return a vector of length 3 whose components are $x_1, x_{1+\theta}, x_{1+2\theta}$ where the allowed values for θ are the positive integers. g and W_θ produce points in $\mathbb{R}^3$. Note that in these experiments each p^θ is a probability density function over $\mathbb{R}^3$. We estimated each such p^θ by centering a zero mean spherical Gaussian on every vector in the associated $W_\theta[g(q_0)]$, with an overall covariance determined by cross validation. We again used a modified bounding box multimetric [10] of Eq. (2) modified for continuous probability densities. The resulting integral was evaluated by Monte Carlo importance sampling.

The aggregate complexity results are presented as the solid red line of 2. The results confirm what we would like to see in a complexity measure. The measure peaks at the accumulation point and is low for small r (where there is a fixed point) and large r (where the time series is random). Additional structure is seen for $r > 3.57$, paralleling the complexity seen in the bifurcation diagram of the logistic map.

To investigate the effects of noise on the SD measure we contaminated all time series the zero mean Gaussian noise having standard deviation of 0.001, and applied the same algorithm. The resulting aggregate complexity measure is plotted as the black dashed line of 2. The major features of the aggregate SD measure are preserved but with some blurring of fine detail.

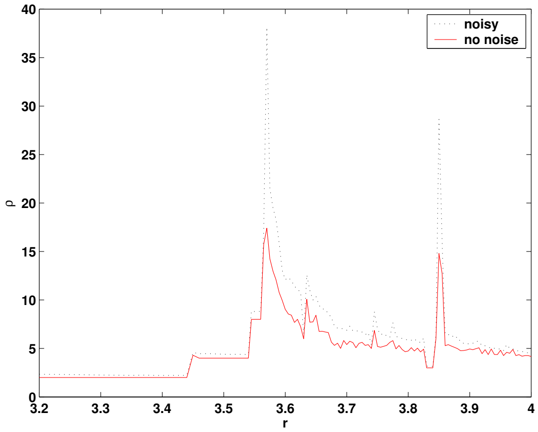

FIG. 2: Aggregate SD complexity measure as a function of r (red line) for the time series generated from the logistic map $x_{t+1} = rx_t(1 - x_t)$. The dashed black line corresponds to a noisy version of the data where zero mean Gaussian noise has been added.

VI. DISCUSSION

Most systems commonly viewed as complex/interesting have been constructed by an evolutionary process (*e.g.* life, culture, intelligence). If we assume that there is some selective advantage in such systems for maximizing the amount of information processing within the system's volume, then we are led to consider systems which are able to process information in many different ways on many spatio-temporal scales, with those different processes all communicating with one another. By exploiting different scales to run different information processing, such systems are in a certain sense maximally dense with respect to how much information processing they achieve in a given volume. Systems processing information similarly on different scales, or even worse not exploiting different scales at all, are inefficient in their information-processing capabilities.

Despite these potential benefits of multi-scale information processing, explicitly constructing a system which engages in such processing is not trivial. Even specifying the necessary dynamical conditions (e.g., a Hamiltonian) for a system to support multi-scale information processing appears difficult. Tellingly, it is also difficult to explicitly construct a physical system that engages in what most researchers would consider "life-like" behavior, or one that engages in "intelligent" behavior; our hypothesis is that this is not a coincidence, but reflects the fact that such systems engage in multi-scale information processing.

In this paper, rather than try to construct systems that engage in multi-scale information processing, we merely assume that nature has stumbled upon ways to do so. Our present goal is only to determine how to recognize and quantify such multi-scale information processing in the first place, and then to measure such processing in real-world systems. Future work would integrate such real-world self-dissimilarity data into a theoretical framework.

In this regard, note that to make maximal use of the different compuational processes at different scales, presumably there must be efficient communication between those processes. Such inter-scale communication is common in systems usually viewed as complex. For example, typically the effects of large scale occurrences (like broken bones in organisms) propagate to the smallest levels (stimulating bone cell growth) in complex systems. Similarly, slight changes at small scales (the bankruptcy of a firm, or the mutation of a gene) can have marked large-scale (industry-wide, or body-wide) effects.

It is generally agreed that any "intelligent" organism has a huge amount of adaptively formed extra-genetic information-processing concerning the outside world, in its brain. The information-processing in the brain of such an organism is tightly and extensively coupled to the information processing of the outside world. Now that outside world is physically a scale up from the organism. So an "intelligent" organism is a subsystem of the world that has the same type of computational coupling with the larger scales of the world that living, complex organisms have between the various scales within their own bodies.

On the other hand, for a complex system to be stable it must be robust with respect to changes in its environment. This suggests that the effects of random perturbations on a particular scale should be isolated to one or a few scales lest the full system be prone to collapse. To this extent scales must be insulated from each other. Accordingly, as a function of the noise inherent in an environment, there may be very precise and constrained ways in which scales can interact in robust systems. If so it would be hoped that when applied to real-world complex systems a self-dissimilarity measure would uncover such a modularity of multi-scale information processing.

Puzzles like how to decide whether a system "is alive" are rendered moot if approached from the perspective of self-dissimilariy. We argue that such puzzles arise from trying to squeeze physical phenomena into pre-existing theoretical models (e.g., for models concerning "life" one must identify the atomic units of the physical system, define what is meant for them to reproduce, *etc.*). We instead view life as a characteristic signature of a system's self-dissimilarity over a range of spatio-temporal scales. In this view life is more than a yes/no bit, and even more than a real number signifying a degree—it is an entire signature.

It may be possible to use such self-dissimilarity signatures to compare entirely different kinds of systems. This would allos us to address questions like the following: How does a modern economy's complexity signature compare to that of the organelles inside a prokaryotic cell? What naturally occurring ecology is most like that of a modern city? What one is most like that of the charge densities moving across the internet? Can cultures be distinguished according to their self-dissimilarity measure?

By concentrating on self-dissimilarity signatures we can compare systems over different regions of scales, thereby investigating how the complexity character itself changes as one varies the scale. This allows us to address questions like: For what range of scales is the associated self-dissimilarity signature of a transportation system most like the signature of the current densities inside a computer? How much is the self-dissimilarity signature of the mass density of the astronomy-scale universe like that of an ideal gas when examined on mesoscopic scales, *etc.*?

VII. ACKNOWLEDGEMENTS

We would like to thank Chris Henze for stimulating discussion.

[1] Badii, R., and A. Politi, *Complexity: Hierarchical Structures and Scaling in Physics*, Cambridge university Press (1997).

[2] Cover, T., and J. Thomas, *Elements of Information Theory*, Wiley-Interscience New York (1991).

[3] Duda, R. O., P. E. Hart, and D. G. Stork, *Pattern Classification (2nd ed.)*, Wiley and Sons (2000).

[4] Fuglede, Bent, and Flemming Topsoe, "Jensen-shannon divergence and hilbert space embedding", Submitted to ISIT2004 (2004).

[5] Gray, A., and A. Moore, "Rapid evaluation of multiple density models", *Artificial Intelligence and Statistics* (C. M. Bishop and B. J. Frey eds.), (2003).

[6] Lloyd, S., "Physical measures of complexity", *1989 Lectures in Complex Systems* (E. Jen ed.), Addison Wesley (1990).

[7] Stanley, M. H. R., L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Les chhorn, P. Maass, M. A. Salinger, and H. E. Stanley, "Scaling behaviour in the growth of companies", *Nature* **379** (1996), 804–806.

[8] Strogatz, S. H., *Nonlinear Dynamics and Chaos : With Applications in Physics, Biology, Chemistry, and Engineering*, Perseus Press (1994).

[9] Topsoe, Flemming, "Inequalities for the jensen-shannon divergence", unpublished.

[10] Wolpert, David H., "Metrics for sets of more than two points", *Proceedings of the International Conference on Complex Systems, 2004*, (2004), in press, short version of paper.

[11] Wolpert, David H., and William Macready, "Self-dissimilarity: An empirically observable complexity measure", *Unifying Themes in Complex Systems*, New England Complex Systems Institute (2000), 626–643.

[12] Some measures go even further and try to gainfully characterize a system with a single bit, as for example in various formal definitions of whether the system is "living/dead".

[13] Here we concentrate on SD analysis of spatial data arrays and the logistic map, but it can also be applied to data set types ranging from symbolic dynamics processes to systems that are not "spatio-temporal" in the conventional sense, like networks.

[14] Machine learning may also arise in that often we do not know q^* directly, but instead have a data set together with a likelihood function giving the likelihood of generating the observed data D. Ultimately, we are interested in the expected value of our self-dissimilarity signature given that data.

[15] The logistic map is a well-studied dynamical system exhibiting a period doubling route to chaos. The first two period doublings occur at $r = 3$ and $r = 1 + \sqrt{6}$.