Information Sciences 462 (2018) 55–75

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Estimating dynamical dimensions from noisy observations

Jack Murdoch Moore[a,*], Michael Small[a,b]

[a] Department of Mathematics and Statistics, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia
[b] Mineral Resources, Australian Resources Research Centre (ARRC), CSIRO, 26 Dick Perry Avenue, Kensington, Western Australia 6151, Australia

ARTICLE INFO

Article history:
Received 4 August 2017
Revised 4 May 2018
Accepted 3 June 2018
Available online 5 June 2018

Keywords:
Description length
Intrinsic dimension
Dynamical dimension
System classification
Invariance

ABSTRACT

Knowledge of the dynamical dimension mitigates the "curse of dimensionality" by permitting analysis in dimension lower than that of the original state vectors. The description length quantifies complexity and so allows us to use Occam's razor to estimate the dynamical dimension underlying noisily observed data. Applying our method, based on the description length, to a coarsely sampled scalar time series requires the choice of only one parameter; an embedding dimension. For the three systems considered in this study observed amid observational noise, a single choice of embedding dimension does provide reasonable estimates of the dynamical dimension. The spatial distribution of local estimates of dynamical dimension aids visualisation and provides extra insight into the geometric structure of many systems.

1. Introduction

1.1. Dynamical dimension

Dynamical or intrinsic dimension [7,8,10,28] is the dimension of the most restrictive surface which suffices to capture the information available in data. Dynamical dimension can be used as an input parameter in lossless dimension reduction [8]. Dimension reduction involves representing features using fewer coordinates than appeared in the original data, and can make analysis more reliable, less computationally challenging, and less likely to mislead [9,20,35]. Camastra [7] and Camastra and Staiano [8] surveyed the numerous methods which have so far been applied to estimation of the dynamical dimension. They considered the approaches taken in the context of dynamical systems and in the context of machine learning.

Consider a subset $\mathcal{A}$ of $\mathbb{R}^m$. The dynamical dimension of $\mathcal{A}$ is the dimension d of the lowest dimensional manifold of which $\mathcal{A}$ is a subset. From dimension invariance [21, p. 40] it follows that d does not change under invertible coordinate transformations. Hence, as Abarbanel and Kennel [2] observed, the dynamical dimension is a system invariant; a quantity independent of the coordinate system in which observations happen to be expressed.

The dynamical dimension specifies a fundamental aspect of the structure of a dynamical system. Abarbanel and Kennel [2] described the dynamical dimension as the "dimension of the local dynamics" and as the number of "active degrees of freedom". Judd [17] locally modelled a bounded attractor of a chaotic dynamical system as the topological product of (1) a Cantor-like set (that is, fractal dust; a set comprising disconnected singletons); and (2) a continuous set. Under this model,

* Corresponding author.
E-mail address: jack.moore@research.uwa.edu.au (J.M. Moore).

https://doi.org/10.1016/j.ins.2018.06.011
0020-0255/

the dynamical dimension d is the number of independent directions which these span. As a system invariant, the dynamical dimension provides a meaningful way to classify dynamical systems [1,2].

Knowledge of the dynamical dimension mitigates the "curse of dimensionality" by allowing local computations in fewer coordinates than would be required to unfold the global state space [6]. Furthermore, the dynamical dimension guides the construction of both physical and empirical models [2,6].

The machine learning community places a different emphasis on intrinsic dimension. They are less explicitly interested in deterministic evolution, and more explicitly interested in the number of directions needed to describe observations [11]. Kramer [19] identified d-dimensional representations of higher dimensional state vectors by fitting a neural network with input and output nodes comprising the training data, and with an internal layer containing only d nodes. The dynamical dimension is identified as d which leads to the best fit according to penalised least squares criteria. Potapov and Ali [28] reasoned that data of dynamical dimension d should be able to be mapped locally into $\mathbb{R}^d$ without changes in interpoint distances. Hence, they fitted a neural network mapping from the space of observations into $\mathbb{R}^d$. This fit minimises a scalar function $S(d)$ which increases in the relative magnitude of the change in the distance between each pair of points. Those authors inferred the dynamical dimension to be the least dimension d for which $S(d)$ ceased to decrease significantly with further increases in d. Amsaleg et al. [3] estimated the dynamical dimension as the maximum likelihood choice of a parameter of the generalised Pareto distribution. The rationale of this approach arises from the result that as the radius of a ball in the space of measurements tends to zero, the upper end of the distribution of distances from the centre of the ball approaches the generalised Pareto distribution [3].

We propose estimation of the dynamical dimension by comparing the efficiency of local linear models of different dimension.

1.2. Selection of local linear models

Local, as distinct from global, models, may only be valid in a proper subset of the full state space. Likelihood maximisation can identify appropriate values of a fixed number and type of parameters with which to describe a fixed neighbourhood, but subtleties are involved in applying it to choosing which parameters, and which data, to include in a model. As the number of parameters increases, so too can the likelihood. A maximum likelihood model could have a parameter for each datum, and would be useless. Analogously, in the context of locally linear fits in state space, as neighbourhood population increases more observations must be fitted and so likelihood tends to decrease.

Local least squares regression identifies optimal local linear models. As Kantz and Schreiber [18, pp. 180–183] illustrated, the affine subspace of dimension m_S which is closest, in the least square sense, to a series of state vectors is that spanned by the eigenvectors corresponding to the m_S largest eigenvalues of the covariance matrix to which the state vectors correspond.

The same affine subspace could have been identified probabilistically, by regarding the components of the deviation from an affine subspace as noise. Specifically, let us assume a Gaussian distribution for the coordinates in each of the $(m - m_S)$ directions which are perpendicular to the subspace. The affine subspace which corresponds to the maximum likelihood values of the Gaussian variables is, again, that spanned by the eigenvectors corresponding to the m_S largest eigenvalues of the covariance matrix.

Least squares regression identifies the best m_S-dimensional model for a fixed number of data, but cannot prescribe an appropriate signal dimension m_S. As the signal dimension m_S increases, fewer data must be described and the apparent goodness of fit can only increase. Non-trivial maxima of the likelihood as a function of the decomposition into signal and noise require description of all m components, rather than only m_S components, of state vectors. For example, Welling et al. [39] described directions with the highest and lowest variabilities as Gaussian variables with their own variance parameters. All other directions, those of moderate variability, are together described using a single variance parameter. The approach, called extreme component analysis, features the complication that the number of parameters available to describe data changes with the number of directions deemed extreme.

Bouveyron et al. [4] avoided that issue by describing via one variance parameter all measurements in the m_S signal directions, and using a second variance parameter for all measurements in the remaining $(m - m_S)$ directions. With this description, all decompositions with $0 < m_S < m$ employ the same number of parameters. The approach, called isotropic probabilistic principal component analysis, reliably estimated the optimal dimension of global linear models. However, in estimates of local linear models the apparent goodness of fit will tend to decrease with increasing number of data.

Description length avoids a monotonically trending dependency of likelihood on sample size. The minimum description length principle is a quantitative application of Occam's razor, which states that "the simplest explanation is probably the correct one" [26]. Given a countable set of observations, the description length of a model for those observations is the complexity of the explanation which that model provides. Specifically, an explanation is a binary (that is, comprising the digits '0' and '1') encoding of the observations, and its complexity is just the expected length of this digital sequence. The minimum description length principle [30] posits that the best model for observations provides the simplest explanation; the mean shortest code; the least description length.

To each efficient and unambiguous code on a countable set X there corresponds a probability mass function (pmf) on X for which that code would be optimal. If the code has length l: $X \to [0, \infty)$ then the optimal pmf will be p: $X \to [0, 1]$, $x \mapsto 2^{-l(x)}$. Conversely, each pmf p: $X \to [0, 1]$ on a countable set corresponds to an optimal code for that set; this code has length l: $X \to [0, \infty)$, $x \mapsto -\log_2(p(x))$. More precisely, as Rissanen [29] showed, the function l provides a strict lower limit,