Physica D 291 (2015) 62–73

Contents lists available at ScienceDirect

Physica D

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

Dimensionality reduction of collective motion by principal manifolds

Kelum Gajamannage [a], Sachit Butail [b], Maurizio Porfiri [b], Erik M. Bollt [a,*]

[a] Department of Mathematics, Clarkson University, Potsdam, NY-13699, USA

[b] Department of Mechanical and Aerospace Engineering, New York University Polytechnic School of Engineering, Brooklyn, NY-11201, USA

HIGHLIGHTS

- We present an algorithm to find principal manifolds of high-dimensional datasets.
- We illustrate the approach using the standard swiss-roll dataset.
- The presented algorithm rejects noise gracefully and avoids sudden failures.
- Compared to Isomap, we show improved results on data reduction of collective motion.
- Performance with respect to smoothing, data density, and noise is analyzed.

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form
22 September 2014
Accepted 24 September 2014
Available online 2 October 2014
Communicated by J. Garnier

Keywords:
Dimensionality reduction
Algorithm
Collective behavior
Dynamical systems

ABSTRACT

While the existence of low-dimensional embedding manifolds has been shown in patterns of collective motion, the current battery of nonlinear dimensionality reduction methods is not amenable to the analysis of such manifolds. This is mainly due to the necessary spectral decomposition step, which limits control over the mapping from the original high-dimensional space to the embedding space. Here, we propose an alternative approach that demands a two-dimensional embedding which topologically summarizes the high-dimensional data. In this sense, our approach is closely related to the construction of one-dimensional principal curves that minimize orthogonal error to data points subject to smoothness constraints. Specifically, we construct a two-dimensional principal manifold directly in the high-dimensional space using cubic smoothing splines, and define the embedding coordinates in terms of geodesic distances. Thus, the mapping from the high-dimensional data to the manifold is defined in terms of local coordinates. Through representative examples, we show that compared to existing nonlinear dimensionality reduction methods, the principal manifold retains the original structure even in noisy and sparse datasets. The principal manifold finding algorithm is applied to configurations obtained from a dynamical system of multiple agents simulating a complex maneuver called predator mobbing, and the resulting two-dimensional embedding is compared with that of a well-established nonlinear dimensionality reduction method.

1. Introduction

With advancements in data collection and video recording methods, high-volume datasets of animal groups, such as fish schools [1,2], bird flocks [3,4], and insect and bacterial swarms [5,6], are now ubiquitous. However, analyzing these datasets is still a nontrivial task, even when individual trajectories of all members are available. A desirable step that may ease the experimenter's task of locating events of interest is to identify coarse observables [7–9] and behavioral measures [10] as the group navigates through space. In this context, Nonlinear Dimensionality Reduction (NDR) offers a large set of tools to infer properties of such complex multi-agent dynamical systems.

Traditional Dimensionality Reduction (DR) methods based on linear techniques, such as Principal Components Analysis (PCA), have been shown to possess limited accuracy when input data is nonlinear and complex [11]. DR entails finding the axes of maximum variability [12] or retaining the distances between points [13]. Multi Dimensional Scaling (MDS) with Euclidean metric is another DR method, which attains low-dimensional representation by retaining the pairwise distance of points in low dimensional representations [13]. However, Euclidean distance calculates the shortest distance between two points on a manifold

* Corresponding author. Tel.: +315 268 2307.
E-mail addresses: dineshhk@clarkson.edu (K. Gajamannage), sb4304@nyu.edu (S. Butail), mporfiri@nyu.edu (M. Porfiri), bolltem@clarkson.edu (E.M. Bollt).

http://dx.doi.org/10.1016/j.physd.2014.09.009
0167-2789/

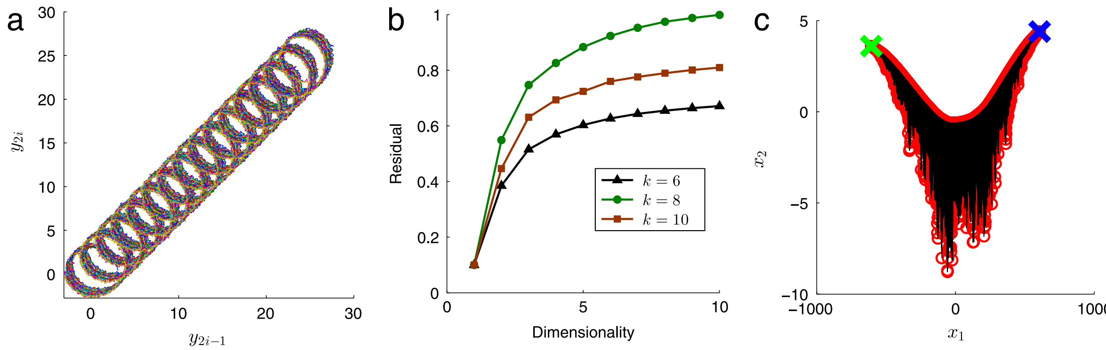

Fig. C.1. Using Isomap to create a two-dimensional embedding of a simulation of collective behavior. (a) Predator mobbing of twenty agents moving on a translating circular trajectory on a plane (enclosing a predator moving at constant speed at a 45° angle), axes y_{2i-1}, y_{2i} generally represent coordinates for the ith agent. (b) Scaled residual variance of candidate low-dimensional embeddings produced by Isomap using different nearest neighbor values k (green-circle, brown-square, and black-triangle), and (c) two-dimensional representation of the data for five nearest neighbors (black-triangle). Green and blue crosses mark the start and end points of the trajectory. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

instead of the genuine manifold distance, which may lead to difficulty in inferring low-dimensional embeddings. The isometric mapping algorithm (Isomap) resolves the problem associated with MDS by preserving the pairwise geodesic distance between points [14]; it has recently been used to analyze group properties in collective behavior, such as the level of coordination and fragmentation [15–19]. Within Isomap, however, short-circuiting [20] created by faulty connections in the neighborhood graph, manifold non-convexity [21,22] and holes in the data [23] can degrade the faithfulness of the reconstructed embedding manifold.

Diffusion maps [24] have also been shown to successfully identify coarse observables in collective phenomena [25] that would otherwise require hit-and-trial guesswork [26]. Beyond Isomap and diffusion maps, the potential of other NDR methods to study collective behavior is largely untested. These include, Kernel PCA (KPCA) which requires the computation of the eigenvectors of the kernel matrix instead of the eigenvectors of the covariance matrix of the data [27]; Local Linear Embedding (LLE) that embeds high-dimensional data through global minimization of local linear reconstruction errors [11]; Hessian LLE (HLLE) that minimizes the curviness of the higher dimensional manifold by assuming that the low-dimensional embedding is locally isometric [28]; and Laplacian Eigenmaps (LE) that perform a weighted minimization (instead of global minimization as in LLE) of the distance between each point and its given nearest neighbors to embed high dimensional data [29].

Iterative NDR approaches have also been recently developed in order to bypass spectral decomposition which is common in most of NDR methods [30]. Curvilinear Component Analysis (CCA) employs a self-organized neural network to perform two tasks, namely, vector quantization of submanifolds in the input space and nonlinear projection of quantized vectors onto a low dimensional space [31]. This method minimizes the distance between the input and output spaces. Manifold Sculpting (MS) transforms data by balancing two opposing heuristics: first, scaling information out of unwanted dimensions, and second, preserving local structure in the data. MS is robust to sampling issues, and iteratively reduces the dimensionality by using a cost function that simulates the relationship among points in a local neighborhoods [30]. The Local Spline Embedding (LSE) is another NDR technique that embeds the data points using splines that map each local coordinate into a global coordinate of the underlying manifold by minimizing the reconstruction error of the objective function [32]. This method reduces the dimensionality by solving an eigenvalue problem while the local geometry is exploited by the tangential projection of data. LSE assumes that the data is not only unaffected by noise or outliers, but also, sampled from a smooth manifold, which ensures the existence of a smooth low dimensional embedding.

Due to the global perspective of all these methods, none of them provide sufficient control over the mapping from the original high-dimensional dataset to the low-dimensional representation, limiting the analysis in the embedding space. In other words, the low-dimensional coordinates are not immediately perceived as useful, whereby one must correlate the axes of the embedding manifold with selected functions of known observables to deduce their physical meaning [26,14]. In this context, a desirable feature of DR that we emphasize here is the regularity in the spatial structure and range of points on the embedding space, despite the presence of noise.

With regard to datasets of collective behavior, nonlinear methods have limited use for detailed analysis at the level of the embedding space. This is primarily because the majority of these methods collapse the data onto a lower dimensional space, whose coordinates are not guaranteed to be linear functions of known system variables [33]. In an idealized simulation of predator induced mobbing [34], a form of collective behavior where a group of animals crowd around a moving predator, two degrees of freedom are obvious, namely, the translation of the group and the rotation about the predator (center of the translating circle). This two-dimensional manifold is not immediately perceived by Isomap, even for the idealized scenario presented in Fig. C.1, where a group of twenty individuals rotate about a predator moving at a constant speed about a line bisecting the first quadrant. Specifically, the algorithm is unable to locate a distinct elbow in the residual variance vs. dimensionality curve, not withstanding substantial tweaking of the parameters—the inferred dimensionality is always 1 (Fig. C.1b). For a two-dimensional embedding (Fig. C.1c), visual comparison of the relative scale of the axes indicates that the horizontal axis represents a greater translation than the vertical axis. It is likely that the horizontal axis captures the motion of the group along the translating circle. The vertical axis could instead be associated with (i) motion about the center of the circle, or (ii) noise, which is independent and identically distributed at each time step. The ambiguity in determining the meaning of such direction indicates a drawback of Isomap in providing meaningful interpretations of the low-dimensional coordinates.

An alternative approach to DR, one that does not require heavy matrix computations or orthogonalization, involves working directly on raw data in the high-dimensional space [35,36]. We propose here a method for DR that relies on geodesic rather than Euclidean distance and emphasizes manifold regularity. Our approach is based on a spline representation of the data that allows us to control the expected manifold regularity. Typically, this entails conditioning the data so that the lower dimensions are revealed. For example, in [35], raw data is successively clustered through a